ATMIC NETWORKS
YOUR GUIDE TO GLOBAL CERTIFIATION

"Please note that these files may not be up to date. However, the questions will help you understand the exam format and typical question patterns."

www.atmicnetworks .com

Warning: Keep connected with our support team for latest updates

Question: 1

Your company built a TensorFlow neural-network model with a large number of neurons and layers. The model fits well for the training dat

a. However, when tested against new data, it performs poorly. What method can you employ to address this?

A. Threading
B. Serialization
C. Dropout Methods
D. Dimensionality Reduction

Answer: C

Explanation:
Reference: https://medium.com/mlreview/a-simple-deep-learning-model-for-stock-price-prediction- using-tensorflow-30505541d877

Question: 2

You are building a model to make clothing recommendations. You know a user's fashion preference is likely to change over time, so you build a data pipeline to stream new data back to the model as it becomes available. How should you use this data to train the model?

A. Continuously retrain the model on just the new data.
B. Continuously retrain the model on a combination of existing data and the new data.

C. Train on the existing data while using the new data as your test set.

D. Train on the new data while using the existing data as your test set.

Answer: C

Explanation:

https://cloud.google.com/automl-tables/docs/prepare

Question: 3

You designed a database for patient records as a pilot project to cover a few hundred patients in three clinics. Your design used a single database table to represent all patients and their visits, and you used self-joins to generate reports. The server resource utilization was at 50%. Since then, the scope of the project has expanded. The database must now store 100 times more patient records. You can no longer run the reports, because they either take too long or they encounter errors with
insufficient compute resources. How should you adjust the database design?

A. Add capacity (memory and disk space) to the database server by the order of 200.

B. Shard the tables into smaller ones based on date ranges, and only generate reports with prespecified date ranges.

C. Normalize the master patient-record table into the patient table and the visits table, and create other necessary tables to avoid self-join.

D. Partition the table into smaller tables, with one for each clinic. Run queries against the smaller table pairs, and use unions for consolidated reports.

Answer: C

Explanation:

Question: 4

You create an important report for your large team in Google Data Studio 360. The report uses Google BigQuery as its data source. You notice that visualizations are not showing data that is less than 1 hour old. What should you do?

A. Disable caching by editing the report settings.

B. Disable caching in BigQuery by editing table details.

C. Refresh your browser tab showing the visualizations.

D. Clear your browser history for the past hour then reload the tab showing the virtualizations.

Answer: A

Explanation:

Reference: https://support.google.com/datastudio/answer/7020039?hl=en

Question: 5

An external customer provides you with a daily dump of data from their database. The data flows into Google Cloud Storage GCS as comma-separated values (CSV) files. You want to analyze this data in Google BigQuery, but the data could have rows that are formatted incorrectly or corrupted. How should you build this pipeline?

A. Use federated data sources, and check data in the SQL query.

B. Enable BigQuery monitoring in Google Stackdriver and create an alert.

C. Import the data into BigQuery using the gcloud CLI and set max_bad_records to 0.

D. Run a Google Cloud Dataflow batch pipeline to import the data into BigQuery, and push errors to another dead-letter table for analysis.

Answer: D

Explanation:

Question: 6

Your weather app queries a database every 15 minutes to get the current temperature. The frontend is powered by Google App Engine and server millions of users. How should you design the frontend to respond to a database failure?

A. Issue a command to restart the database servers.

B. Retry the query with exponential backoff, up to a cap of 15 minutes.

C. Retry the query every second until it comes back online to minimize staleness of data.

D. Reduce the query frequency to once every hour until the database comes back online.

Answer: B

Explanation:

https://cloud.google.com/sql/docs/mysql/manage-connections#backoff

Question: 7

You are creating a model to predict housing prices. Due to budget constraints, you must run it on a single resource-constrained virtual machine. Which learning algorithm should you use?

A. Linear regression

B. Logistic classification

C. Recurrent neural network

D. Feedforward neural network

Answer: A

Explanation:

Question: 8

You are building new real-time data warehouse for your company and will use Google BigQuery streaming inserts. There is no guarantee that data will only be sent in once but you do have a unique ID for each row of data and an event timestamp. You want to ensure that duplicates are not included while interactively querying dat

a. Which query type should you use?

A. Include ORDER BY DESK on timestamp column and LIMIT to 1.

B. Use GROUP BY on the unique ID column and timestamp column and SUM on the values.

C. Use the LAG window function with PARTITION by unique ID along with WHERE LAG IS NOT NULL.

D. Use the ROW_NUMBER window function with PARTITION by unique ID along with WHERE row equals 1.

Answer: D

Explanation:

https://cloud.google.com/bigquery/docs/reference/standard-sql/analytic-function-concepts

Question: 9

Your company is using WHILECARD tables to query data across multiple tables with similar names. The SQL statement is currently failing with the following error:

Syntax error : Expected end of statement but got "-" at [4:11]

SELECT age

```
FROM
bigquery-public-data.noaa_gsod.gsod
WHERE
age != 99
AND_TABLE_SUFFIX = '1929'
ORDER BY
```

age DESC

Which table name will make the SQL statement work correctly?

A. ‘bigquery-public-data.noaa_gsod.gsod‘

B. bigquery-public-data.noaa_gsod.gsod*

C. ‘bigquery-public-data.noaa_gsod.gsod’*

D. ‘bigquery-public-data.noaa_gsod.gsod*`

Answer: D

Explanation:

Question: 10

Your company is in a highly regulated industry. One of your requirements is to ensure individual users have access only to the minimum amount of information required to do their jobs. You want to enforce this requirement with Google BigQuery. Which three approaches can you take? (Choose three.)

A. Disable writes to certain tables.

B. Restrict access to tables by role.

C. Ensure that the data is encrypted at all times.

D. Restrict BigQuery API access to approved users.

E. Segregate data across multiple tables or databases.

F. Use Google Stackdriver Audit Logging to determine policy violations.

Answer: B,D,F

Explanation:

Question: 11

You are designing a basket abandonment system for an ecommerce company. The system will send a message to a user based on these rules:

No interaction by the user on the site for 1 hour

Has added more than $30 worth of products to the basket

Has not completed a transaction

You use Google Cloud Dataflow to process the data and decide if a message should be sent. How should you design the pipeline?

A. Use a fixed-time window with a duration of 60 minutes.

B. Use a sliding time window with a duration of 60 minutes.

C. Use a session window with a gap time duration of 60 minutes.

D. Use a global window with a time based trigger with a delay of 60 minutes.

Answer: C

Explanation:

Question: 12

Your company handles data processing for a number of different clients. Each client prefers to use their own suite of analytics tools, with some allowing direct query access via Google BigQuery. You need to secure the data so that

clients cannot see each other's dat

a. You want to ensure appropriate access to the data. Which three steps should you take? (Choose three.)

A. Load data into different partitions.

B. Load data into a different dataset for each client.

C. Put each client's BigQuery dataset into a different table.

D. Restrict a client's dataset to approved users.

E. Only allow a service account to access the datasets.

F. Use the appropriate identity and access management (IAM) roles for each client's users.

Answer: B,D,F

Explanation:

Question: 13

You want to process payment transactions in a point-of-sale application that will run on Google Cloud Platform. Your user base could grow exponentially, but you do not want to manage infrastructure scaling.

Which Google database service should you use?

A. Cloud SQL

B. BigQuery

C. Cloud Bigtable

D. Cloud Datastore

Answer: A

Explanation:

Question: 14

You want to use a database of information about tissue samples to classify future tissue samples as either normal or mutated. You are evaluating an unsupervised anomaly detection method for classifying the tissue samples. Which two characteristic support this method? (Choose two.)

A. There are very few occurrences of mutations relative to normal samples.

B. There are roughly equal occurrences of both normal and mutated samples in the database.

C. You expect future mutations to have different features from the mutated samples in the database.

D. You expect future mutations to have similar features to the mutated samples in the database.

E. You already have labels for which samples are mutated and which are normal in the database.

Answer: AD

Explanation:

Unsupervised anomaly detection techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal by looking for instances that seem to fit least to the remainder of the data set.

https://en.wikipedia.org/wiki/Anomaly_detection

Question: 15

You need to store and analyze social media postings in Google BigQuery at a rate of 10,000 messages per minute in near real-time. Initially, design the application to use streaming inserts for individual postings. Your application also performs data aggregations right after the streaming inserts. You discover that the queries after streaming inserts do not exhibit strong consistency, and reports from the queries might miss in-flight dat
a. How can you adjust your application design?

A. Re-write the application to load accumulated data every 2 minutes.

B. Convert the streaming insert code to batch load for individual messages.

C. Load the original message to Google Cloud SQL, and export the table every hour to BigQuery via streaming inserts.

D. Estimate the average latency for data availability after streaming inserts, and always run queries after waiting twice as long.

Answer: D

Explanation:

The data is first comes to buffer and then written to Storage. If we are running queries in buffer we will face above mentioned issues. If we wait for the bigquery to write the data to storage then we won't face the issue. So We need to wait till it's written tio storage

Question: 16

Your startup has never implemented a formal security policy. Currently, everyone in the company has access to the datasets stored in Google BigQuery. Teams have freedom to use the service as they see fit, and they have not documented their use cases. You have been asked to secure the data warehouse. You need to discover what everyone is doing. What should you do first?

A. Use Google Stackdriver Audit Logs to review data access.

B. Get the identity and access management IIAM) policy of each table

C. Use Stackdriver Monitoring to see the usage of BigQuery query slots.

D. Use the Google Cloud Billing API to see what account the warehouse is being billed to.

Answer: A

Explanation:

Question: 17

Your company is migrating their 30-node Apache Hadoop cluster to the cloud. They want to re-use Hadoop jobs they have already created and minimize the management of the cluster as much as possible. They also want to be able to persist data beyond the life of the cluster. What should you do?

A. Create a Google Cloud Dataflow job to process the data.

B. Create a Google Cloud Dataproc cluster that uses persistent disks for HDFS.

C. Create a Hadoop cluster on Google Compute Engine that uses persistent disks.

D. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.

E. Create a Hadoop cluster on Google Compute Engine that uses Local SSD disks.

Answer: D

Explanation:

Question: 18

Business owners at your company have given you a database of bank transactions. Each row contains the user ID, transaction type, transaction location, and transaction amount. They ask you to investigate what type of machine learning can be applied to the dat

a. Which three machine learning applications can you use? (Choose three.)

A. Supervised learning to determine which transactions are most likely to be fraudulent.

B. Unsupervised learning to determine which transactions are most likely to be fraudulent.

C. Clustering to divide the transactions into N categories based on feature similarity.

D. Supervised learning to predict the location of a transaction.

E. Reinforcement learning to predict the location of a transaction.

F. Unsupervised learning to predict the location of a transaction.

Answer: B,C,D

Explanation:

Question: 19

Your company's on-premises Apache Hadoop servers are approaching end-of-life, and IT has decided to migrate the cluster to Google Cloud Dataproc. A like-for-like migration of the cluster would require 50 TB of Google Persistent Disk per node. The CIO is concerned about the cost of using that much block storage. You want to minimize the storage cost of the migration. What should you do?

A. Put the data into Google Cloud Storage.

B. Use preemptible virtual machines (VMs) for the Cloud Dataproc cluster.

C. Tune the Cloud Dataproc cluster so that there is just enough disk for all data.

D. Migrate some of the cold data into Google Cloud Storage, and keep only the hot data in Persistent Disk.

Answer: B

Explanation:

Reference:

Question: 20

You work for a car manufacturer and have set up a data pipeline using Google Cloud Pub/Sub to capture anomalous sensor events. You are using a push subscription in Cloud Pub/Sub that calls a custom HTTPS endpoint that you have created to take action of these anomalous events as they occur. Your custom HTTPS endpoint keeps getting an inordinate amount of duplicate messages. What is the most likely cause of these duplicate messages?

A. The message body for the sensor event is too large.

B. Your custom endpoint has an out-of-date SSL certificate.

C. The Cloud Pub/Sub topic has too many messages published to it.

D. Your custom endpoint is not acknowledging messages within the acknowledgement deadline.

Answer: B

Explanation:

Question: 21

Your company uses a proprietary system to send inventory data every 6 hours to a data ingestion service in the cloud. Transmitted data includes a payload of several fields and the timestamp of the transmission. If there are any concerns about a transmission, the system re-transmits the dat

a. How should you deduplicate the data most efficiency?

A. Assign global unique identifiers (GUID) to each data entry.

B. Compute the hash value of each data entry, and compare it with all historical data.

C. Store each data entry as the primary key in a separate database and apply an index.

D. Maintain a database table to store the hash value and other metadata for each data entry.

Answer: D

Explanation:

Question: 22

Your company has hired a new data scientist who wants to perform complicated analyses across very large datasets

stored in Google Cloud Storage and in a Cassandra cluster on Google Compute Engine. The scientist primarily wants to create labelled data sets for machine learning projects, along with some visualization tasks. She reports that her laptop is not powerful enough to perform her tasks and it is slowing her down. You want to help her perform her tasks. What should you do?

A. Run a local version of Jupiter on the laptop.

B. Grant the user access to Google Cloud Shell.

C. Host a visualization tool on a VM on Google Compute Engine.

D. Deploy Google Cloud Datalab to a virtual machine (VM) on Google Compute Engine.

Answer: B

Explanation:

Question: 23

You are deploying 10,000 new Internet of Things devices to collect temperature data in your warehouses globally. You need to process, store and analyze these very large datasets in real time. What should you do?

A. Send the data to Google Cloud Datastore and then export to BigQuery.

B. Send the data to Google Cloud Pub/Sub, stream Cloud Pub/Sub to Google Cloud Dataflow, and store the data in Google BigQuery.

C. Send the data to Cloud Storage and then spin up an Apache Hadoop cluster as needed in Google Cloud Dataproc whenever analysis is required.

D. Export logs in batch to Google Cloud Storage and then spin up a Google Cloud SQL instance, import the data from Cloud Storage, and run an analysis as needed.

Answer: B

Explanation:

Question: 24

You have spent a few days loading data from comma-separated values (CSV) files into the Google BigQuery table CLICK_STREAM. The column DT stores the epoch time of click events. For convenience, you chose a simple schema where every field is treated as the STRING type. Now, you want to compute web session durations of users who visit your site, and you want to change its data type to the TIMESTAMP. You want to minimize the migration effort without making future queries computationally expensive. What should you do?

A. Delete the table CLICK_STREAM, and then re-create it such that the column DT is of the TIMESTAMP type. Reload the data.

B. Add a column TS of the TIMESTAMP type to the table CLICK_STREAM, and populate the numeric values from the column TS for each row. Reference: the column TS instead of the column DT from now on.

C. Create a view CLICK_STREAM_V, where strings from the column DT are cast into TIMESTAMP values. Reference: the view CLICK_STREAM_V instead of the table CLICK_STREAM from now on.

D. Add two columns to the table CLICK STREAM: TS of the TIMESTAMP type and IS_NEW of the BOOLEAN type. Reload all data in append mode. For each appended row, set the value of IS_NEW to true. For future queries, reference the column TS instead of the column DT, with the WHERE clause ensuring that the value of IS_NEW must be true.

E. Construct a query to return every row of the table CLICK_STREAM, while using the built-in function to cast strings from the column DT into TIMESTAMP values. Run the query into a destination table NEW_CLICK_STREAM, in which the column TS is the TIMESTAMP type. Reference: the table NEW_CLICK_STREAM instead of the table CLICK_STREAM from now on. In the future, new data is loaded into the table NEW_CLICK_STREAM.

Answer: D

Explanation:

Question: 25

You want to use Google Stackdriver Logging to monitor Google BigQuery usage. You need an instant notification to be sent to your monitoring tool when new data is appended to a certain table using an insert job, but you do not want to receive notifications for other tables. What should you do?

A. Make a call to the Stackdriver API to list all logs, and apply an advanced filter.

B. In the Stackdriver logging admin interface, and enable a log sink export to BigQuery.

C. In the Stackdriver logging admin interface, enable a log sink export to Google Cloud Pub/Sub, and subscribe to the topic from your monitoring tool.

D. Using the Stackdriver API, create a project sink with advanced log filter to export to Pub/Sub, and subscribe to the topic from your monitoring tool.

Answer: B

Explanation:

Question: 26

You are working on a sensitive project involving private user dat

a. You have set up a project on Google Cloud Platform to house your work internally. An external consultant is going to assist with coding a complex transformation in a Google Cloud Dataflow pipeline for your project. How should you maintain users' privacy?

A. Grant the consultant the Viewer role on the project.

B. Grant the consultant the Cloud Dataflow Developer role on the project.

C. Create a service account and allow the consultant to log on with it.

D. Create an anonymized sample of the data for the consultant to work with in a different project.

Answer: C

Explanation:

Question: 27

You are building a model to predict whether or not it will rain on a given day. You have thousands of input features and want to see if you can improve training speed by removing some features while having a minimum effect on

model accuracy. What can you do?

A. Eliminate features that are highly correlated to the output labels.

B. Combine highly co-dependent features into one representative feature.

C. Instead of feeding in each feature individually, average their values in batches of 3.

D. Remove the features that have null values for more than 50% of the training records.

Answer: B

Explanation:

Question: 28

Your company is performing data preprocessing for a learning algorithm in Google Cloud Dataflow. Numerous data logs are being are being generated during this step, and the team wants to analyze them. Due to the dynamic nature of the campaign, the data is growing exponentially every hour.

The data scientists have written the following code to read the data for a new key features in the

logs.

BigQueryIO.Read

.named("ReadLogData")

.from("clouddataflow-readonly:samples.log_data")

You want to improve the performance of this data read. What should you do?

A. Specify the TableReference: object in the code.

B. Use .fromQuery operation to read specific fields from the table.

C. Use of both the Google BigQuery TableSchema and TableFieldSchema classes.

D. Call a transform that returns TableRow objects, where each element in the PCollexction represents a single row in the table.

Answer: D

Explanation:

Question: 29

Your company is streaming real-time sensor data from their factory floor into Bigtable and they have noticed extremely poor performance. How should the row key be redesigned to improve Bigtable performance on queries that populate real-time dashboards?

A. Use a row key of the form <timestamp>.

B. Use a row key of the form <sensorid>.

C. Use a row key of the form <timestamp>#<sensorid>.

D. Use a row key of the form >#<sensorid>#<timestamp>.

Answer: A

Explanation:

Question: 30

Your company's customer and order databases are often under heavy load. This makes performing analytics against them difficult without harming operations. The databases are in a MySQL cluster, with nightly backups taken using mysqldump. You want to perform analytics with minimal impact on operations. What should you do?

A. Add a node to the MySQL cluster and build an OLAP cube there.

B. Use an ETL tool to load the data from MySQL into Google BigQuery.

C. Connect an on-premises Apache Hadoop cluster to MySQL and perform ETL.

D. Mount the backups to Google Cloud SQL, and then process the data using Google Cloud Dataproc.

Answer: C

Explanation:

Question: 31

You have Google Cloud Dataflow streaming pipeline running with a Google Cloud Pub/Sub subscription as the source. You need to make an update to the code that will make the new Cloud Dataflow pipeline incompatible with the current version. You do not want to lose any data when making this update. What should you do?

A. Update the current pipeline and use the drain flag.

B. Update the current pipeline and provide the transform mapping JSON object.

C. Create a new pipeline that has the same Cloud Pub/Sub subscription and cancel the old pipeline.

D. Create a new pipeline that has a new Cloud Pub/Sub subscription and cancel the old pipeline.

Answer: D

Explanation:

Question: 32

Your company is running their first dynamic campaign, serving different offers by analyzing real-time data during the holiday season. The data scientists are collecting terabytes of data that rapidly grows every hour during their 30-day campaign. They are using Google Cloud Dataflow to preprocess the data and collect the feature (signals) data that is needed for the machine learning model in Google Cloud Bigtable. The team is observing suboptimal performance with reads and writes of their initial load of 10 TB of dat

a. They want to improve this performance while minimizing cost. What should they do?

A. Redefine the schema by evenly distributing reads and writes across the row space of the table.

B. The performance issue should be resolved over time as the site of the BigDate cluster is increased.

C. Redesign the schema to use a single row key to identify values that need to be updated frequently in the cluster.

D. Redesign the schema to use row keys based on numeric IDs that increase sequentially per user viewing the offers.

Answer: A

Explanation:

Question: 33

Your software uses a simple JSON format for all messages. These messages are published to Google Cloud Pub/Sub, then processed with Google Cloud Dataflow to create a real-time dashboard for the CFO. During testing, you notice that some messages are missing in the dashboard. You check the logs, and all messages are being published to Cloud Pub/Sub successfully. What should you do next?

A. Check the dashboard application to see if it is not displaying correctly.

B. Run a fixed dataset through the Cloud Dataflow pipeline and analyze the output.

C. Use Google Stackdriver Monitoring on Cloud Pub/Sub to find the missing messages.

D. Switch Cloud Dataflow to pull messages from Cloud Pub/Sub instead of Cloud Pub/Sub pushing messages to Cloud Dataflow.

Answer: B

Explanation:

Topic 2, Flowlogistic Case Study

Company Overview

Flowlogistic is a leading logistics and supply chain provider. They help businesses throughout the world manage their resources and transport them to their final destination. The company has grown rapidly, expanding their offerings to include rail, truck, aircraft, and oceanic shipping.

Company Background

The company started as a regional trucking company, and then expanded into other logistics market. Because they have not updated their infrastructure, managing and tracking orders and shipments has become a bottleneck. To improve operations, Flowlogistic developed proprietary technology for tracking shipments in real time at the parcel level. However, they are unable to deploy it because their technology stack, based on Apache Kafka, cannot support the processing volume. In addition, Flowlogistic wants to further analyze their orders and shipments to determine how best to deploy their resources.

Solution Concept

Flowlogistic wants to implement two concepts using the cloud:

Use their proprietary technology in a real-time inventory-tracking system that indicates the location of their loads

Perform analytics on all their orders and shipment logs, which contain both structured and unstructured data, to determine how best to deploy resources, which markets to expand info. They also want to use predictive analytics to learn earlier when a shipment will be delayed.

Existing Technical Environment

Flowlogistic architecture resides in a single data center:

Databases

8 physical servers in 2 clusters

SQL Server – user data, inventory, static data

3 physical servers

Cassandra – metadata, tracking messages

10 Kafka servers – tracking message aggregation and batch insert

Application servers – customer front end, middleware for order/customs

60 virtual machines across 20 physical servers

Tomcat – Java services

Nginx – static content

Batch servers

Storage appliances

iSCSI for virtual machine (VM) hosts

Fibre Channel storage area network (FC SAN) – SQL server storage

Network-attached storage (NAS) image storage, logs, backups

Apache Hadoop /Spark servers

Core Data Lake

Data analysis workloads

20 miscellaneous servers

Jenkins, monitoring, bastion hosts,

Business Requirements

Build a reliable and reproducible environment with scaled panty of production.

Aggregate data in a centralized Data Lake for analysis

Use historical data to perform predictive analytics on future shipments

Accurately track every shipment worldwide using proprietary technology

Improve business agility and speed of innovation through rapid provisioning of new resources

Analyze and optimize architecture for performance in the cloud

Migrate fully to the cloud if all other requirements are met

Technical Requirements

Handle both streaming and batch data

Migrate existing Hadoop workloads

Ensure architecture is scalable and elastic to meet the changing demands of the company.

Use managed services whenever possible

Encrypt data flight and at rest

Connect a VPN between the production data center and cloud environment

SEO Statement

We have grown so quickly that our inability to upgrade our infrastructure is really hampering further growth and efficiency. We are efficient at moving shipments around the world, but we are inefficient at moving data around.

We need to organize our information so we can more easily understand where our customers are and what they are shipping.

CTO Statement

IT has never been a priority for us, so as our data has grown, we have not invested enough in our technology. I have a good staff to manage IT, but they are so busy managing our infrastructure that I cannot get them to do the things that really matter, such as organizing our data, building the analytics, and figuring out how to implement the CFO' s tracking technology.

CFO Statement

Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries. Knowing where out shipments are at all times has a direct correlation to our bottom line and profitability. Additionally, I don't want to commit capital to building out a server environment.

Question: 34

Flowlogistic wants to use Google BigQuery as their primary analysis system, but they still have Apache Hadoop and Spark workloads that they cannot move to BigQuery. Flowlogistic does not know how to store the data that is common to both workloads. What should they do?

A. Store the common data in BigQuery as partitioned tables.

B. Store the common data in BigQuery and expose authorized views.

C. Store the common data encoded as Avro in Google Cloud Storage.

D. Store he common data in the HDFS storage for a Google Cloud Dataproc cluster.

Answer: B

Explanation:

Question: 35

Flowlogistic's management has determined that the current Apache Kafka servers cannot handle the data volume for their real-time inventory tracking system. You need to build a new system on Google Cloud Platform (GCP) that will feed the proprietary tracking software. The system must be able to ingest data from a variety of global sources, process and query in real-time, and store the data reliably. Which combination of GCP products should you choose?

A. Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage

B. Cloud Pub/Sub, Cloud Dataflow, and Local SSD

C. Cloud Pub/Sub, Cloud SQL, and Cloud Storage

D. Cloud Load Balancing, Cloud Dataflow, and Cloud Storage

Answer: C

Explanation:

Question: 36

Flowlogistic's CEO wants to gain rapid insight into their customer base so his sales team can be better informed in the field. This team is not very technical, so they've purchased a visualization tool to simplify the creation of BigQuery reports. However, they've been overwhelmed by all the data in the table, and are spending a lot of money on queries trying to find the data they need. You want to solve their problem in the most cost-effective way. What should you do?

A. Export the data into a Google Sheet for virtualization.

B. Create an additional table with only the necessary columns.

C. Create a view on the table to present to the virtualization tool.

D. Create identity and access management (IAM) roles on the appropriate columns, so only they appear in a query.

Answer: C

Explanation:

Question: 37

Flowlogistic is rolling out their real-time inventory tracking system. The tracking devices will all send package-tracking messages, which will now go to a single Google Cloud Pub/Sub topic instead of the Apache Kafka cluster. A subscriber application will then process the messages for real-time reporting and store them in Google BigQuery for historical analysis. You want to ensure the package data can be analyzed over time.

Which approach should you take?

A. Attach the timestamp on each message in the Cloud Pub/Sub subscriber application as they are received.

B. Attach the timestamp and Package ID on the outbound message from each publisher device as they are sent to Clod Pub/Sub.

C. Use the NOW () function in BigQuery to record the event's time.

D. Use the automatically generated timestamp from Cloud Pub/Sub to order the data.

Answer: B

Explanation:

Topic 3, MJTelco Case Study

Company Overview

MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.

Company Background

Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.

Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept

MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:

Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.

Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.

MJTelco will also use three separate operating environments – development/test, staging, and production – to meet the needs of running experiments, deploying new features, and serving production customers.

Business Requirements

Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community.

Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.

Provide reliable and timely access to data for analysis from distributed research workers

Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.

Technical Requirements

Ensure secure and efficient transport and storage of telemetry data

Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.

Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day

Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.

CEO Statement

Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.

CTO Statement

Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.

CFO Statement

The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.

Question: 38

MJTelco's Google Cloud Dataflow pipeline is now ready to start receiving data from the 50,000 installations. You want to allow Cloud Dataflow to scale its compute power up as required. Which Cloud Dataflow pipeline configuration setting should you update?

A. The zone

B. The number of workers

C. The disk size per worker

D. The maximum number of workers

Answer: A

Explanation:

Question: 39

You need to compose visualizations for operations teams with the following requirements:

Which approach meets the requirements?

A. Load the data into Google Sheets, use formulas to calculate a metric, and use filters/sorting to show only suboptimal links in a table.

B. Load the data into Google BigQuery tables, write Google Apps Script that queries the data, calculates the metric, and shows only suboptimal rows in a table in Google Sheets.

C. Load the data into Google Cloud Datastore tables, write a Google App Engine Application that queries all rows, applies a function to derive the metric, and then renders results in a table using the Google charts and visualization API.

D. Load the data into Google BigQuery tables, write a Google Data Studio 360 report that connects to your data, calculates a metric, and then uses a filter expression to show only suboptimal rows in a table.

Answer: C

Explanation:

Question: 40

You create a new report for your large team in Google Data Studio 360. The report uses Google BigQuery as its data source. It is company policy to ensure employees can view only the data associated with their region, so you create and populate a table for each region. You need to enforce the regional access policy to the data.

Which two actions should you take? (Choose two.)

A. Ensure all the tables are included in global dataset.

B. Ensure each table is included in a dataset for a region.

C. Adjust the settings for each table to allow a related region-based security group view access.

D. Adjust the settings for each view to allow a related region-based security group view access.

E. Adjust the settings for each dataset to allow a related region-based security group view access.

Answer: B,D

Explanation:

Question: 41

MJTelco needs you to create a schema in Google Bigtable that will allow for the historical analysis of the last 2 years of records. Each record that comes in is sent every 15 minutes, and contains a unique identifier of the device and a data record. The most common query is for all the data for a given device for a given day. Which schema should you use?

A. Rowkey: date#device_idColumn data: data_point

B. Rowkey: dateColumn data: device_id, data_point

C. Rowkey: device_idColumn data: date, data_point

D. Rowkey: data_pointColumn data: device_id, date

E. Rowkey: date#data_pointColumn data: device_id

Answer: D

Explanation:

Question: 42

MJTelco is building a custom interface to share dat

a. They have these requirements:

They need to do aggregations over their petabyte-scale datasets.

They need to scan specific time range rows with a very fast response time (milliseconds).

Which combination of Google Cloud Platform products should you recommend?

A. Cloud Datastore and Cloud Bigtable

B. Cloud Bigtable and Cloud SQL

C. BigQuery and Cloud Bigtable

D. BigQuery and Cloud Storage

Answer: C

Explanation:

Question: 43

You need to compose visualization for operations teams with the following requirements:

Telemetry must include data from all 50,000 installations for the most recent 6 weeks (sampling once every minute)

The report must not be more than 3 hours delayed from live data.

The actionable report should only show suboptimal links.

Most suboptimal links should be sorted to the top.

Suboptimal links can be grouped and filtered by regional geography.

User response time to load the report must be <5 seconds.

You create a data source to store the last 6 weeks of data, and create visualizations that allow viewers to see multiple date ranges, distinct geographic regions, and unique installation types. You always show the latest data without any changes to your visualizations. You want to avoid creating and updating new visualizations each month. What should you do?

A. Look through the current data and compose a series of charts and tables, one for each possible combination of criteria.

B. Look through the current data and compose a small set of generalized charts and tables bound to criteria filters that allow value selection.

C. Export the data to a spreadsheet, compose a series of charts and tables, one for each possible combination of criteria, and spread them across multiple tabs.

D. Load the data into relational database tables, write a Google App Engine application that queries all rows, summarizes the data across each criteria, and then renders results using the Google Charts and visualization API.

Answer: B

Explanation:

Question: 44

Given the record streams MJTelco is interested in ingesting per day, they are concerned about the cost of Google BigQuery increasing. MJTelco asks you to provide a design solution. They require a single large data table called

tracking_table. Additionally, they want to minimize the cost of daily queries while performing fine-grained analysis of each day's events. They also want to use streaming ingestion. What should you do?

A. Create a table called tracking_table and include a DATE column.

B. Create a partitioned table called tracking_table and include a TIMESTAMP column.

C. Create sharded tables for each day following the pattern tracking_table_YYYYMMDD.

D. Create a table called tracking_table with a TIMESTAMP column to represent the day.

Answer: B

Explanation:

Topic 4, Main Questions Set B

Question: 45

Your company has recently grown rapidly and now ingesting data at a significantly higher rate than it was previously. You manage the daily batch MapReduce analytics jobs in Apache Hadoop. However, the recent increase in data has meant the batch jobs are falling behind. You were asked to recommend ways the development team could increase the responsiveness of the analytics without increasing costs. What should you recommend they do?

A. Rewrite the job in Pig.

B. Rewrite the job in Apache Spark.

C. Increase the size of the Hadoop cluster.

D. Decrease the size of the Hadoop cluster but also rewrite the job in Hive.

Answer: A

Explanation:

Question: 46

You work for a large fast food restaurant chain with over 400,000 employees. You store employee information in Google BigQuery in a Users table consisting of a FirstName field and a LastName field. A member of IT is building an application and asks you to modify the schema and data in BigQuery so the application can query a FullName field consisting of the value of the FirstName field concatenated with a space, followed by the value of the LastName field for each employee. How can you make that data available while minimizing cost?

A. Create a view in BigQuery that concatenates the FirstName and LastName field values to produce the FullName.

B. Add a new column called FullName to the Users table. Run an UPDATE statement that updates the FullName column for each user with the concatenation of the FirstName and LastName values.

C. Create a Google Cloud Dataflow job that queries BigQuery for the entire Users table, concatenates the FirstName value and LastName value for each user, and loads the proper values for FirstName, LastName, and FullName into a new table in BigQuery.

D. Use BigQuery to export the data for the table to a CSV file. Create a Google Cloud Dataproc job to process the CSV file and output a new CSV file containing the proper values for FirstName, LastName and FullName. Run a BigQuery load job to load the new CSV file into BigQuery.

Answer: C

Explanation:

Question: 47

You are deploying a new storage system for your mobile application, which is a media streaming service. You decide the best fit is Google Cloud Datastore. You have entities with multiple properties, some of which can take on multiple values. For example, in the entity 'Movie' the property 'actors' and the property 'tags' have multiple values but the property 'date released' does not. A typical query would ask for all movies with actor=<actorname> ordered by date_released or all movies with tag=Comedy ordered by date_released. How should you avoid a combinatorial explosion in the number of indexes?

A. Manually configure the index in your index config as follows

```
Indexes: -kind: Movie
    Properties:
    -name: actors
    name: date_released
-kind: Movie
    Properties:
    -name: tags
    name: date_released
```

B. Manually configure the index hi your hidex config as follows:

```
Indexes:
    -kind: Movie
        Properties:
        -name: actors
        -name: tags
-name: date_published
```

C. Set the following in your entity options: exclude_from_indexes = ‘actors, tags’

D. Set the following in your entity options: exclude_from indexes = ‘date_published’

A. Option A

B. Option B.

C. Option C

D. Option D

Answer: A

Explanation:

Question: 48

You work for a manufacturing plant that batches application log files together into a single log file once a day at 2:00 AM. You have written a Google Cloud Dataflow job to process that log file. You need to make sure the log file in processed once per day as inexpensively as possible. What should you do?

A. Change the processing job to use Google Cloud Dataproc instead.

B. Manually start the Cloud Dataflow job each morning when you get into the office.

C. Create a cron job with Google App Engine Cron Service to run the Cloud Dataflow job.

D. Configure the Cloud Dataflow job as a streaming job so that it processes the log data immediately.

Answer: C

Explanation:

Question: 49

You work for an economic consulting firm that helps companies identify economic trends as they happen. As part of your analysis, you use Google BigQuery to correlate customer data with the average prices of the 100 most common goods sold, including bread, gasoline, milk, and others. The average prices of these goods are updated every 30 minutes. You want to make sure this data stays up to date so you can combine it with other data in BigQuery as cheaply as possible. What should you do?

A. Load the data every 30 minutes into a new partitioned table in BigQuery.

B. Store and update the data in a regional Google Cloud Storage bucket and create a federated data source in BigQuery

C. Store the data in Google Cloud Datastore. Use Google Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Cloud Datastore

D. Store the data in a file in a regional Google Cloud Storage bucket. Use Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Google Cloud Storage.

Answer: A

Explanation:

Question: 50

You are designing the database schema for a machine learning-based food ordering service that will predict what users want to eat. Here is some of the information you need to store:

The user profile: What the user likes and doesn't like to eat

The user account information: Name, address, preferred meal times

The order information: When orders are made, from where, to whom

The database will be used to store all the transactional data of the product. You want to optimize the data schem

a. Which Google Cloud Platform product should you use?

A. BigQuery

B. Cloud SQL

C. Cloud Bigtable

D. Cloud Datastore

Answer: A

Explanation:

Question: 51

Your company is loading comma-separated values (CSV) files into Google BigQuery. The data is fully imported successfully; however, the imported data is not matching byte-to-byte to the source file. What is the most likely cause of this problem?

A. The CSV data loaded in BigQuery is not flagged as CSV.

B. The CSV data has invalid rows that were skipped on import.

C. The CSV data loaded in BigQuery is not using BigQuery's default encoding.

D. The CSV data has not gone through an ETL phase before loading into BigQuery.

Answer: B

Explanation:

Question: 52

Your company produces 20,000 files every hour. Each data file is formatted as a comma separated values (CSV) file that is less than 4 KB. All files must be ingested on Google Cloud Platform before they can be processed. Your company site has a 200 ms latency to Google Cloud, and your Internet connection bandwidth is limited as 50 Mbps. You currently deploy a secure FTP (SFTP) server on a virtual machine in Google Compute Engine as the data ingestion point. A local SFTP client runs on a dedicated machine to transmit the CSV files as is. The goal is to make reports with data from the previous day available to the executives by 10:00 a.m. each day. This design is barely able to keep up with the current volume, even though the bandwidth utilization is rather low.

You are told that due to seasonality, your company expects the number of files to double for the next three months. Which two actions should you take? (choose two.)

A. Introduce data compression for each file to increase the rate file of file transfer.

B. Contact your internet service provider (ISP) to increase your maximum bandwidth to at least 100 Mbps.

C. Redesign the data ingestion process to use gsutil tool to send the CSV files to a storage bucket in parallel.

D. Assemble 1,000 files into a tape archive (TAR) file. Transmit the TAR files instead, and disassemble the CSV files in the cloud upon receiving them.

E. Create an S3-compatible storage endpoint in your network, and use Google Cloud Storage Transfer Service to transfer on-premices data to the designated storage bucket.

Answer: C,E

Explanation:

Question: 53

You are choosing a NoSQL database to handle telemetry data submitted from millions of Internet-of- Things (IoT) devices. The volume of data is growing at 100 TB per year, and each data entry has about 100 attributes. The data processing pipeline does not require atomicity, consistency, isolation, and durability (ACID). However, high availability and low latency are required.

You need to analyze the data by querying against individual fields. Which three databases meet your requirements? (Choose three.)

A. Redis
B. HBase
C. MySQL
D. MongoDB
E. Cassandra
F. HDFS with Hive

Answer: B,D,F

Explanation:

Topic 5, Practice Questions

Question: 54

Suppose you have a table that includes a nested column called "city" inside a column called "person", but when you try to submit the following query in BigQuery, it gives you an error.

```
SELECT person FROM `project1.example.table1` WHERE city = "London"
```

How would you correct the error?

A. Add ", UNNEST(person)" before the WHERE clause.

B. Change "person" to "person.city".

C. Change "person" to "city.person".

D. Add ", UNNEST(city)" before the WHERE clause.

Answer: A

Explanation:

To access the person.city column, you need to "UNNEST(person)" and JOIN it to table1 using a comma.

Reference:

https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql#nested_repeated_results

Question: 55

What are two of the benefits of using denormalized data structures in BigQuery?

A. Reduces the amount of data processed, reduces the amount of storage required

B. Increases query speed, makes queries simpler

C. Reduces the amount of storage required, increases query speed

D. Reduces the amount of data processed, increases query speed

Answer: B

Explanation:

Denormalization increases query speed for tables with billions of rows because BigQuery's performance degrades when doing JOINs on large tables, but with a denormalized data

structure, you don't have to use JOINs, since all of the data has been combined into one table.

Denormalization also makes queries simpler because you do not have to use JOIN clauses.

Denormalization increases the amount of data processed and the amount of storage required because it creates redundant data.

Reference:

https://cloud.google.com/solutions/bigquery-data-warehouse#denormalizing_data

Question: 56

Which of these statements about exporting data from BigQuery is false?

A. To export more than 1 GB of data, you need to put a wildcard in the destination filename.

B. The only supported export destination is Google Cloud Storage.

C. Data can only be exported in JSON or Avro format.

D. The only compression option available is GZIP.

Answer: C

Explanation:

Data can be exported in CSV, JSON, or Avro format. If you are exporting nested or repeated data, then CSV format is not supported.

Reference: https://cloud.google.com/bigquery/docs/exporting-data

Question: 57

What are all of the BigQuery operations that Google charges for?

A. Storage, queries, and streaming inserts

B. Storage, queries, and loading data from a file

C. Storage, queries, and exporting data

D. Queries and streaming inserts

Answer: A

Explanation:

Google charges for storage, queries, and streaming inserts. Loading data from a file and exporting data are free operations.

Reference: https://cloud.google.com/bigquery/pricing

Question: 58

Which of the following is not possible using primitive roles?

A. Give a user viewer access to BigQuery and owner access to Google Compute Engine instances.

B. Give UserA owner access and UserB editor access for all datasets in a project.

C. Give a user access to view all datasets in a project, but not run queries on them.

D. Give GroupA owner access and GroupB editor access for all datasets in a project.

Answer: C

Explanation:

Primitive roles can be used to give owner, editor, or viewer access to a user or group, but they can't be used to separate data access permissions from job-running permissions.

Reference: https://cloud.google.com/bigquery/docs/access-control#primitive_iam_roles

Question: 59

Which of these statements about BigQuery caching is true?

A. By default, a query's results are not cached.

B. BigQuery caches query results for 48 hours.

C. Query results are cached even if you specify a destination table.

D. There is no charge for a query that retrieves its results from cache.

Answer: D

Explanation:

When query results are retrieved from a cached results table, you are not charged for the query.

BigQuery caches query results for 24 hours, not 48 hours.

Query results are not cached if you specify a destination table.

A query's results are always cached except under certain conditions, such as if you specify a destination table.

Reference: https://cloud.google.com/bigquery/querying-data#query-caching

Question: 60

Which of these sources can you not load data into BigQuery from?

A. File upload

B. Google Drive

C. Google Cloud Storage

D. Google Cloud SQL

Answer: D

Explanation:

You can load data into BigQuery from a file upload, Google Cloud Storage, Google Drive, or Google Cloud Bigtable. It is not possible to load data into BigQuery directly from Google Cloud SQL. One way to get data from Cloud SQL to BigQuery would be to export data from Cloud SQL to Cloud Storage and then load it from there.

Reference: https://cloud.google.com/bigquery/loading-data

Question: 61

Which of the following statements about Legacy SQL and Standard SQL is not true?

A. Standard SQL is the preferred query language for BigQuery.

B. If you write a query in Legacy SQL, it might generate an error if you try to run it with Standard SQL.

C. One difference between the two query languages is how you specify fully-qualified table names (i.e. table names that include their associated project name).

D. You need to set a query language for each dataset and the default is Standard SQL.

Answer: D

Explanation:

You do not set a query language for each dataset. It is set each time you run a query and the default query language is Legacy SQL.

Standard SQL has been the preferred query language since BigQuery 2.0 was released.

In legacy SQL, to query a table with a project-qualified name, you use a colon, :, as a separator. In standard SQL, you use a period, ., instead.

Due to the differences in syntax between the two query languages (such as with project-qualified table names), if you write a query in Legacy SQL, it might generate an error if you try to run it with Standard SQL.

Reference:

https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql

Question: 62

How would you query specific partitions in a BigQuery table?

A. Use the DAY column in the WHERE clause

B. Use the EXTRACT(DAY) clause

C. Use the PARTITIONTIME pseudo-column in the WHERE clause

D. Use DATE BETWEEN in the WHERE clause

Answer: C

Explanation:

Partitioned tables include a pseudo column named _PARTITIONTIME that contains a date-based timestamp for data loaded into the table. To limit a query to particular partitions (such as Jan 1st and 2nd of 2017), use a clause similar to this:

WHERE _PARTITIONTIME BETWEEN TIMESTAMP('2017-01-01') AND TIMESTAMP('2017-01-02')

Reference: https://cloud.google.com/bigquery/docs/partitioned- tables#the_partitiontime_pseudo_column

Question: 63

Which SQL keyword can be used to reduce the number of columns processed by BigQuery?

A. BETWEEN

B. WHERE

C. SELECT

D. LIMIT

Answer: C

Explanation:

SELECT allows you to query specific columns rather than the whole table.

LIMIT, BETWEEN, and WHERE clauses will not reduce the number of columns processed by BigQuery.

Reference: https://cloud.google.com/bigquery/launch-checklist#architecture_design_and_development_checklist

Question: 64

To give a user read permission for only the first three columns of a table, which access control method would you use?

A. Primitive role

B. Predefined role

C. Authorized view

D. It's not possible to give access to only the first three columns of a table.

Answer: C

Explanation:

An authorized view allows you to share query results with particular users and groups without giving them read access to the underlying tables. Authorized views can only be created in a dataset that does not contain the tables queried by the view.

When you create an authorized view, you use the view's SQL query to restrict access to only the rows and columns you want the users to see.

Reference: https://cloud.google.com/bigquery/docs/views#authorized-views

Question: 65

What are two methods that can be used to denormalize tables in BigQuery?

A. 1) Split table into multiple tables; 2) Use a partitioned table

B. 1) Join tables into one table; 2) Use nested repeated fields

C. 1) Use a partitioned table; 2) Join tables into one table

D. 1) Use nested repeated fields; 2) Use a partitioned table

Answer: B

Explanation:

The conventional method of denormalizing data involves simply writing a fact, along with all its dimensions, into a flat table structure. For example, if you are dealing with sales transactions, you would write each individual fact to a record, along with the accompanying dimensions such as order and customer information.

The other method for denormalizing data takes advantage of BigQuery's native support for nested and repeated structures in JSON or Avro input data. Expressing records using nested and repeated structures can provide a more natural representation of the underlying data. In the case of the sales order, the outer part of a JSON structure would contain the order and customer information, and the inner part of the structure would contain the individual line items of the order, which would be represented as nested, repeated elements.

Reference: https://cloud.google.com/solutions/bigquery-data-warehouse#denormalizing_data

Question: 66

Which of these is not a supported method of putting data into a partitioned table?

A. If you have existing data in a separate file for each day, then create a partitioned table and upload each file into the appropriate partition.

B. Run a query to get the records for a specific day from an existing table and for the destination table, specify a partitioned table ending with the day in the format "$YYYYMMDD".

C. Create a partitioned table and stream new records to it every day.

D. Use ORDER BY to put a table's rows into chronological order and then change the table's type to "Partitioned".

Answer: D

Explanation:

You cannot change an existing table into a partitioned table. You must create a partitioned table from scratch. Then you can either stream data into it every day and the data will automatically be put in the right partition, or you can load data into a specific partition by using "$YYYYMMDD" at the end of the table name.

Reference: https://cloud.google.com/bigquery/docs/partitioned-tables

Question: 67

Which of these operations can you perform from the BigQuery Web UI?

A. Upload a file in SQL format.

B. Load data with nested and repeated fields.

C. Upload a 20 MB file.

D. Upload multiple files using a wildcard.

Answer: B

Explanation:

You can load data with nested and repeated fields using the Web UI.

You cannot use the Web UI to:

- Upload a file greater than 10 MB in size
- Upload multiple files at the same time
- Upload a file in SQL format

All three of the above operations can be performed using the "bq" command.

Reference: https://cloud.google.com/bigquery/loading-data

Question: 68

Which methods can be used to reduce the number of rows processed by BigQuery?

A. Splitting tables into multiple tables; putting data in partitions

B. Splitting tables into multiple tables; putting data in partitions; using the LIMIT clause

C. Putting data in partitions; using the LIMIT clause

D. Splitting tables into multiple tables; using the LIMIT clause

Answer: A

Explanation:

If you split a table into multiple tables (such as one table for each day), then you can limit your query to the data in specific tables (such as for particular days). A better method is to use a partitioned table, as long as your data can be separated by the day.

If you use the LIMIT clause, BigQuery will still process the entire table.

Reference: https://cloud.google.com/bigquery/docs/partitioned-tables

Question: 69

Why do you need to split a machine learning dataset into training data and test data?

A. So you can try two different sets of features

B. To make sure your model is generalized for more than just the training data

C. To allow you to create unit tests in your code

D. So you can use one dataset for a wide model and one for a deep model

Answer: B

Explanation:

The flaw with evaluating a predictive model on training data is that it does not inform you on how well the model has generalized to new unseen data. A model that is selected for its accuracy on the training dataset rather than its accuracy on an unseen test dataset is very likely to have lower accuracy on an unseen test dataset. The reason is that the model is not as generalized. It has specialized to the structure in the training dataset. This is called overfitting.

Reference: https://machinelearningmastery.com/a-simple-intuition-for-overfitting/

Question: 70

Which of these numbers are adjusted by a neural network as it learns from a training dataset (select 2 answers)?

A. Weights

B. Biases

C. Continuous features

D. Input values

Answer: AB

Explanation:

A neural network is a simple mechanism that's implemented with basic math. The only difference between the traditional programming model and a neural network is that you let the computer determine the parameters (weights and bias) by learning from training datasets.

Reference: https://cloud.google.com/blog/big-data/2016/07/understanding-neural-networks-with- tensorflow-playground

Question: 71

The CUSTOM tier for Cloud Machine Learning Engine allows you to specify the number of which types of cluster nodes?

A. Workers

B. Masters, workers, and parameter servers

C. Workers and parameter servers

D. Parameter servers

Answer: C

Explanation:

The CUSTOM tier is not a set tier, but rather enables you to use your own cluster specification. When you use this tier, set values to configure your processing cluster according to these guidelines:

You must set TrainingInput.masterType to specify the type of machine to use for your master node.

You may set TrainingInput.workerCount to specify the number of workers to use.

You may set TrainingInput.parameterServerCount to specify the number of parameter servers to use.

You can specify the type of machine for the master node, but you can't specify more than one master node.

Reference: https://cloud.google.com/ml-engine/docs/training- overview#job_configuration_parameters

Question: 72

Which software libraries are supported by Cloud Machine Learning Engine?

A. Theano and TensorFlow

B. Theano and Torch

C. TensorFlow

D. TensorFlow and Torch

Answer: C

Explanation:

Cloud ML Engine mainly does two things:

Enables you to train machine learning models at scale by running TensorFlow training applications in the cloud.

Hosts those trained models for you in the cloud so that you can use them to get predictions about new data.

Reference: https://cloud.google.com/ml-engine/docs/technical-overview#what_it_does

Question: 73

Which TensorFlow function can you use to configure a categorical column if you don't know all of the possible values for that column?

A. categorical_column_with_vocabulary_list

B. categorical_column_with_hash_bucket

C. categorical_column_with_unknown_values

D. sparse_column_with_keys

Answer: B

Explanation:

If you know the set of all possible feature values of a column and there are only a few of them, you can use categorical_column_with_vocabulary_list. Each key in the list will get assigned an autoincremental ID starting from 0.

What if we don't know the set of possible values in advance? Not a problem. We can use categorical_column_with_hash_bucket instead. What will happen is that each possible value in the feature column occupation will be hashed to an integer ID as we encounter them in training.

Reference: https://www.tensorflow.org/tutorials/wide

Question: 74

Which of the following statements about the Wide & Deep Learning model are true? (Select 2 answers.)

A. The wide model is used for memorization, while the deep model is used for generalization.

B. A good use for the wide and deep model is a recommender system.

C. The wide model is used for generalization, while the deep model is used for memorization.

D. A good use for the wide and deep model is a small-scale linear regression problem.

Answer: AB

Explanation:

Can we teach computers to learn like humans do, by combining the power of memorization and generalization? It's not an easy question to answer, but by jointly training a wide linear model (for memorization) alongside a deep neural network (for generalization), one can combine the strengths of both to bring us one step closer. At Google, we call it Wide & Deep Learning. It's useful for generic large-scale regression and classification problems with sparse inputs (categorical features with a large number of possible feature values), such as recommender systems, search, and ranking problems.

Reference: https://research.googleblog.com/2016/06/wide-deep-learning-better-together- with.html

Question: 75

To run a TensorFlow training job on your own computer using Cloud Machine Learning Engine, what would your command start with?

A. gcloud ml-engine local train

B. gcloud ml-engine jobs submit training

C. gcloud ml-engine jobs submit training local

D. You can't run a TensorFlow program on your own computer using Cloud ML Engine .

Answer: A

Explanation:

gcloud ml-engine local train - run a Cloud ML Engine training job locally

This command runs the specified module in an environment similar to that of a live Cloud ML Engine Training Job.

This is especially useful in the case of testing distributed models, as it allows you to validate that you are properly interacting with the Cloud ML Engine cluster configuration.

Reference: https://cloud.google.com/sdk/gcloud/reference/ml-engine/local/train

Question: 76

If you want to create a machine learning model that predicts the price of a particular stock based on its recent price history, what type of estimator should you use?

A. Unsupervised learning

B. Regressor

C. Classifier

D. Clustering estimator

Answer: B

Explanation:

Regression is the supervised learning task for modeling and predicting continuous, numeric variables. Examples include predicting real-estate prices, stock price movements, or student test scores.

Classification is the supervised learning task for modeling and predicting categorical variables. Examples include predicting employee churn, email spam, financial fraud, or student letter grades.

Clustering is an unsupervised learning task for finding natural groupings of observations (i.e. clusters) based on the inherent structure within your dataset. Examples include customer segmentation, grouping similar items in e-commerce, and social network analysis.

Reference: https://elitedatascience.com/machine-learning-algorithms

Question: 77

Suppose you have a dataset of images that are each labeled as to whether or not they contain a human face. To create a neural network that recognizes human faces in images using this labeled dataset, what approach would likely be the most effective?

A. Use K-means Clustering to detect faces in the pixels.

B. Use feature engineering to add features for eyes, noses, and mouths to the input data.

C. Use deep learning by creating a neural network with multiple hidden layers to automatically detect features of faces.

D. Build a neural network with an input layer of pixels, a hidden layer, and an output layer with two categories.

Answer: C

Explanation:

Traditional machine learning relies on shallow nets, composed of one input and one output layer, and at most one hidden layer in between. More than three layers (including input and output) qualifies as “deep” learning. So deep is a strictly defined, technical term that means more than one hidden layer.

In deep-learning networks, each layer of nodes trains on a distinct set of features based on the previous layer’s output. The further you advance into the neural net, the more complex the features your nodes can recognize, since they aggregate and recombine features from the

previous layer.

A neural network with only one hidden layer would be unable to automatically recognize high-level features of faces, such as eyes, because it wouldn't be able to "build" these features using previous hidden layers that detect low-level features, such as lines.

Feature engineering is difficult to perform on raw image data.

K-means Clustering is an unsupervised learning method used to categorize unlabeled data.

Reference: https://deeplearning4j.org/neuralnet-overview

Question: 78

What are two of the characteristics of using online prediction rather than batch prediction?

A. It is optimized to handle a high volume of data instances in a job and to run more complex models.

B. Predictions are returned in the response message.

C. Predictions are written to output files in a Cloud Storage location that you specify.

D. It is optimized to minimize the latency of serving predictions.

Answer: BD

Explanation:

Online prediction

.Optimized to minimize the latency of serving predictions.

.Predictions returned in the response message.

Batch prediction

.Optimized to handle a high volume of instances in a job and to run more complex models.

.Predictions written to output files in a Cloud Storage location that you specify.

Reference: https://cloud.google.com/ml-engine/docs/prediction-overview#online_prediction_versus_batch_prediction

Question: 79

Which of these are examples of a value in a sparse vector? (Select 2 answers.)

A. [0, 5, 0, 0, 0, 0]

B. [0, 0, 0, 1, 0, 0, 1]

C. [0, 1]

D. [1, 0, 0, 0, 0, 0, 0]

Answer: CD

Explanation:

Categorical features in linear models are typically translated into a sparse vector in which each possible value has a

corresponding index or id. For example, if there are only three possible eye colors you can represent 'eye_color' as a length 3 vector: 'brown' would become [1, 0, 0], 'blue' would become [0, 1, 0] and 'green' would become [0, 0, 1]. These vectors are called "sparse" because they may be very long, with many zeros, when the set of possible values is very large (such as all English words).

[0, 0, 0, 1, 0, 0, 1] is not a sparse vector because it has two 1s in it. A sparse vector contains only a single 1.

[0, 5, 0, 0, 0, 0] is not a sparse vector because it has a 5 in it. Sparse vectors only contain 0s and 1s.

Reference: https://www.tensorflow.org/tutorials/linear#feature_columns_and_transformations

Question: 80

How can you get a neural network to learn about relationships between categories in a categorical feature?

A. Create a multi-hot column
B. Create a one-hot column
C. Create a hash bucket
D. Create an embedding column

Answer: D

Explanation:

There are two problems with one-hot encoding. First, it has high dimensionality, meaning that instead of having just one value, like a continuous feature, it has many values, or dimensions. This makes computation more time-consuming, especially if a feature has a very large number of categories. The second problem is that it doesn't encode any relationships between the categories. They are completely independent from each other, so the network has no way of knowing which ones are similar to each other.

Both of these problems can be solved by representing a categorical feature with an embedding

column. The idea is that each category has a smaller vector with, let's say, 5 values in it. But unlike a one-hot vector, the values are not usually 0. The values are weights, similar to the weights that are used for basic features in a neural network. The difference is that each category has a set of weights (5 of them in this case).

You can think of each value in the embedding vector as a feature of the category. So, if two

categories are very similar to each other, then their embedding vectors should be very similar too.

Reference: https://cloudacademy.com/google/introduction-to-google-cloud-machine-learning- engine-course/a-wide-and-deep-model.html

Question: 81

If a dataset contains rows with individual people and columns for year of birth, country, and income, how many of the columns are continuous and how many are categorical?

A. 1 continuous and 2 categorical

B. 3 categorical

C. 3 continuous

D. 2 continuous and 1 categorical

Answer: D

Explanation:

The columns can be grouped into two types—categorical and continuous columns:

A column is called categorical if its value can only be one of the categories in a finite set. For example, the native country of a person (U.S., India, Japan, etc.) or the education level (high school, college, etc.) are categorical columns.

A column is called continuous if its value can be any numerical value in a continuous range. For example, the capital gain of a person (e.g. $14,084) is a continuous column.

Year of birth and income are continuous columns. Country is a categorical column.

You could use bucketization to turn year of birth and/or income into categorical features, but the raw columns are continuous.

Reference: https://www.tensorflow.org/tutorials/wide#reading_the_census_data

Question: 82

Which of the following are examples of hyperparameters? (Select 2 answers.)

A. Number of hidden layers

B. Number of nodes in each hidden layer

C. Biases

D. Weights

Answer: AB

Explanation:

If model parameters are variables that get adjusted by training with existing data, your hyperparameters are the variables about the training process itself. For example, part of setting up a deep neural network is deciding how many "hidden" layers of nodes to use between the input layer and the output layer, as well as how many nodes each layer should use. These variables are not directly related to the training data at all. They are configuration variables. Another difference is that parameters change during a training job, while the hyperparameters are usually constant during a job.

Weights and biases are variables that get adjusted during the training process, so they are not hyperparameters.

Reference: https://cloud.google.com/ml-engine/docs/hyperparameter-tuning-overview

Question: 83

Which of the following are feature engineering techniques? (Select 2 answers)

A. Hidden feature layers

B. Feature prioritization

C. Crossed feature columns

D. Bucketization of a continuous feature

Answer: CD

Explanation:

Selecting and crafting the right set of feature columns is key to learning an effective model.

Bucketization is a process of dividing the entire range of a continuous feature into a set of consecutive bins/buckets, and then converting the original numerical feature into a bucket ID (as a categorical feature) depending on which bucket that value falls into.

Using each base feature column separately may not be enough to explain the data. To learn the differences between different feature combinations, we can add crossed feature columns to the model.

Reference:

https://www.tensorflow.org/tutorials/wide#selecting_and_engineering_features_for_the_model

Question: 84

You want to use a BigQuery table as a data sink. In which writing mode(s) can you use BigQuery as a sink?

A. Both batch and streaming

B. BigQuery cannot be used as a sink

C. Only batch

D. Only streaming

Answer: A

Explanation:

When you apply a BigQueryIO.Write transform in batch mode to write to a single table, Dataflow invokes a BigQuery load job. When you apply a BigQueryIO.Write transform in streaming mode or in batch mode using a function to specify the destination table, Dataflow uses BigQuery's streaming inserts

Reference: https://cloud.google.com/dataflow/model/bigquery-io

Question: 85

You have a job that you want to cancel. It is a streaming pipeline, and you want to ensure that any data that is in-flight is processed and written to the output. Which of the following commands can you use on the Dataflow monitoring console to stop the pipeline job?

A. Cancel

B. Drain

C. Stop

D. Finish

Answer: B

Explanation:

Using the Drain option to stop your job tells the Dataflow service to finish your job in its current state.
Your job will immediately stop ingesting new data from input sources, but the Dataflow

service will preserve any existing resources (such as worker instances) to finish processing and writing any buffered data in your pipeline.

Reference: https://cloud.google.com/dataflow/pipelines/stopping-a-pipeline

Question: 86

When running a pipeline that has a BigQuery source, on your local machine, you continue to get permission denied

errors. What could be the reason for that?

A. Your gcloud does not have access to the BigQuery resources

B. BigQuery cannot be accessed from local machines

C. You are missing gcloud on your machine

D. Pipelines cannot be run locally

Answer: A

Explanation:

When reading from a Dataflow source or writing to a Dataflow sink using DirectPipelineRunner, the Cloud Platform account that you configured with the gcloud executable will need access to the corresponding source/sink

Reference: https://cloud.google.com/dataflow/java-sdk/JavaDoc/com/google/cloud/dataflow/sdk/runners/DirectPipelineRunner

Question: 87

What Dataflow concept determines when a Window's contents should be output based on certain criteria being met?

A. Sessions
B. OutputCriteria
C. Windows
D. Triggers

Answer: D

Explanation:

Triggers control when the elements for a specific key and window are output. As elements arrive, they are put into one or more windows by a Window transform and its associated WindowFn, and then passed to the associated Trigger to determine if the Windows contents should be output.

Reference: https://cloud.google.com/dataflow/java-sdk/JavaDoc/com/google/cloud/dataflow/sdk/transforms/windowing/Trigger

Question: 88

Which of the following is NOT one of the three main types of triggers that Dataflow supports?

A. Trigger based on element size in bytes

B. Trigger that is a combination of other triggers

C. Trigger based on element count

D. Trigger based on time

Answer: A

Explanation:

There are three major kinds of triggers that Dataflow supports: 1. Time-based triggers 2. Data-driven triggers. You can set a trigger to emit results from a window when that window has received a certain number of data elements. 3. Composite triggers. These triggers combine multiple time-based or data-driven triggers in some logical way

Reference: https://cloud.google.com/dataflow/model/triggers

Question: 89

Which Java SDK class can you use to run your Dataflow programs locally?

A. LocalRunner

B. DirectPipelineRunner

C. MachineRunner

D. LocalPipelineRunner

Answer: B

Explanation:

DirectPipelineRunner allows you to execute operations in the pipeline directly, without any optimization.

Useful for small local execution and tests

Reference: https://cloud.google.com/dataflow/java-sdk/JavaDoc/com/google/cloud/dataflow/sdk/runners/DirectPipelineRunner

Question: 90

The Dataflow SDKs have been recently transitioned into which Apache service?

A. Apache Spark

B. Apache Hadoop

C. Apache Kafka

D. Apache Beam

Answer: D

Explanation:

Dataflow SDKs are being transitioned to Apache Beam, as per the latest Google directive

Reference: https://cloud.google.com/dataflow/docs/

Question: 91

The for Cloud Bigtable makes it possible to use Cloud Bigtable in a Cloud Dataflow pipeline.

A. Cloud Dataflow connector

B. DataFlow SDK

C. BiqQuery API

D. BigQuery Data Transfer Service

Answer: A

Explanation:

The Cloud Dataflow connector for Cloud Bigtable makes it possible to use Cloud Bigtable in a Cloud Dataflow pipeline. You can use the connector for both batch and streaming operations.

Reference: https://cloud.google.com/bigtable/docs/dataflow-hbase

Question: 92

Does Dataflow process batch data pipelines or streaming data pipelines?

A. Only Batch Data Pipelines

B. Both Batch and Streaming Data Pipelines

C. Only Streaming Data Pipelines

D. None of the above

Answer: B

Explanation:

Dataflow is a unified processing model, and can execute both streaming and batch data pipelines

Reference: https://cloud.google.com/dataflow/

Question: 93

You are planning to use Google's Dataflow SDK to analyze customer data such as displayed below. Your project requirement is to extract only the customer name from the data source and then write to an output PCollection.

Tom,555 X street

Tim,553 Y street

Sam, 111 Z street

Which operation is best suited for the above data processing requirement?

A. ParDo

B. Sink API

C. Source API

D. Data extraction

Answer: A

Explanation:

In Google Cloud dataflow SDK, you can use the ParDo to extract only a customer name of each element in your PCollection.

Reference: https://cloud.google.com/dataflow/model/par-do

Question: 94

Which Cloud Dataflow / Beam feature should you use to aggregate data in an unbounded data source every hour based on the time when the data entered the pipeline?

A. An hourly watermark

B. An event time trigger

C. The with Allowed Lateness method

D. A processing time trigger

Answer: D

Explanation:

When collecting and grouping data into windows, Beam uses triggers to determine when to emit the aggregated results of each window.

Processing time triggers. These triggers operate on the processing time – the time when the data element is processed at any given stage in the pipeline.

Event time triggers. These triggers operate on the event time, as indicated by the timestamp on each data element. Beam's default trigger is event time-based.

Reference: https://beam.apache.org/documentation/programming-guide/#triggers

Question: 95

Which of the following is NOT true about Dataflow pipelines?

A. Dataflow pipelines are tied to Dataflow, and cannot be run on any other runner

B. Dataflow pipelines can consume data from other Google Cloud services

C. Dataflow pipelines can be programmed in Java

D. Dataflow pipelines use a unified programming model, so can work both with streaming and batch data sources

Answer: A

Explanation:

Dataflow pipelines can also run on alternate runtimes like Spark and Flink, as they are built using the Apache Beam SDKs

Reference: https://cloud.google.com/dataflow/

Question: 96

You are developing a software application using Google's Dataflow SDK, and want to use conditional, for loops and other complex programming structures to create a branching pipeline. Which component will be used for the data processing operation?

A. PCollection
B. Transform
C. Pipeline
D. Sink API

Answer: B

Explanation:

In Google Cloud, the Dataflow SDK provides a transform component. It is responsible for the data processing operation. You can use conditional, for loops, and other complex programming structure to create a branching pipeline.

Reference: https://cloud.google.com/dataflow/model/programming-model

Question: 97

Which of the following IAM roles does your Compute Engine account require to be able to run pipeline jobs?

A. dataflow.worker

B. dataflow.compute

C. dataflow.developer

D. dataflow.viewer

Answer: A

Explanation:

The dataflow.worker role provides the permissions necessary for a Compute Engine service account to execute work units for a Dataflow pipeline

Reference: https://cloud.google.com/dataflow/access-control

Question: 98

Which of the following is not true about Dataflow pipelines?

A. Pipelines are a set of operations

B. Pipelines represent a data processing job

C. Pipelines represent a directed graph of steps

D. Pipelines can share data between instances

Answer: D

Explanation:

The data and transforms in a pipeline are unique to, and owned by, that pipeline. While your program can create multiple pipelines, pipelines cannot share data or transforms

Reference: https://cloud.google.com/dataflow/model/pipelines

Question: 99

By default, which of the following windowing behavior does Dataflow apply to unbounded data sets?

A. Windows at every 100 MB of data

B. Single, Global Window

C. Windows at every 1 minute

D. Windows at every 10 minutes

Answer: B

Explanation:

Dataflow's default windowing behavior is to assign all elements of a PCollection to a single, global window, even for unbounded PCollections

Reference: https://cloud.google.com/dataflow/model/pcollection

Question: 100

Which of the following job types are supported by Cloud Dataproc (select 3 answers)?

A. Hive

B. Pig

C. YARN

D. Spark

Answer: ABD

Explanation:

Cloud Dataproc provides out-of-the box and end-to-end support for many of the most popular job types, including Spark, Spark SQL, PySpark, MapReduce, Hive, and Pig jobs.

Reference: https://cloud.google.com/dataproc/docs/resources/faq#what_type_of_jobs_can_i_run

Question: 101

What are the minimum permissions needed for a service account used with Google Dataproc?

A. Execute to Google Cloud Storage; write to Google Cloud Logging

B. Write to Google Cloud Storage; read to Google Cloud Logging

C. Execute to Google Cloud Storage; execute to Google Cloud Logging

D. Read and write to Google Cloud Storage; write to Google Cloud Logging

Answer: D

Explanation:

Service accounts authenticate applications running on your virtual machine instances to other Google Cloud Platform services. For example, if you write an application that reads and writes files on Google Cloud Storage, it must first authenticate to the Google Cloud Storage API. At a minimum, service accounts used with Cloud Dataproc need permissions to read and write to Google Cloud Storage, and to write to Google Cloud Logging.

Reference: https://cloud.google.com/dataproc/docs/concepts/service-accounts#important_notes

Question: 102

Which role must be assigned to a service account used by the virtual machines in a Dataproc cluster so they can execute jobs?

A. Dataproc Worker

B. Dataproc Viewer

C. Dataproc Runner

D. Dataproc Editor

Answer: A

Explanation:

Service accounts used with Cloud Dataproc must have Dataproc/Dataproc Worker role (or have all the permissions granted by Dataproc Worker role).

Reference: https://cloud.google.com/dataproc/docs/concepts/service-accounts#important_notes

Question: 103

When creating a new Cloud Dataproc cluster with the projects.regions.clusters.create operation, these four values are required: project, region, name, and .

A. zone

B. node

C. label

D. type

Answer: A

Explanation:

At a minimum, you must specify four values when creating a new cluster with the projects.regions.clusters.create operation:

The project in which the cluster will be created

The region to use

The name of the cluster

The zone in which the cluster will be created

You can specify many more details beyond these minimum requirements. For example, you can also specify the number of workers, whether preemptible compute should be used, and the network settings.

Reference: https://cloud.google.com/dataproc/docs/tutorials/python-library-example#create_a_new_cloud_dataproc_cluste

Question: 104

Which Google Cloud Platform service is an alternative to Hadoop with Hive?

A. Cloud Dataflow

B. Cloud Bigtable

C. BigQuery

D. Cloud Datastore

Answer: C

Explanation:

Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data summarization, query, and analysis.

Google BigQuery is an enterprise data warehouse.

Reference: https://en.wikipedia.org/wiki/Apache_Hive

Question: 105

Which of these rules apply when you add preemptible workers to a Dataproc cluster (select 2 answers)?

A. Preemptible workers cannot use persistent disk.

B. Preemptible workers cannot store data.

C. If a preemptible worker is reclaimed, then a replacement worker must be added manually.

D. A Dataproc cluster cannot have only preemptible workers.

Answer: BD

Explanation:

The following rules will apply when you use preemptible workers with a Cloud Dataproc cluster:

. Processing only—Since preemptibles can be reclaimed at any time, preemptible workers do not store data. Preemptibles added to a Cloud Dataproc cluster only function as processing nodes.

. No preemptible-only clusters—To ensure clusters do not lose all workers, Cloud Dataproc cannot create preemptible-only clusters.

. Persistent disk size—As a default, all preemptible workers are created with the smaller of 100GB or the primary worker boot disk size. This disk space is used for local caching of data and is not available through HDFS.

The managed group automatically re-adds workers lost due to reclamation as capacity permits.

Reference: https://cloud.google.com/dataproc/docs/concepts/preemptible-vms

Question: 106

When using Cloud Dataproc clusters, you can access the YARN web interface by configuring a browser to connect through a proxy.

A. HTTPS

B. VPN

C. SOCKS

D. HTTP

Answer: C

Explanation:

When using Cloud Dataproc clusters, configure your browser to use the SOCKS proxy. The SOCKS proxy routes data intended for the Cloud Dataproc cluster through an SSH tunnel.

Reference: https://cloud.google.com/dataproc/docs/concepts/cluster-web-interfaces#interfaces

Question: 107

Cloud Dataproc is a managed Apache Hadoop and Apache service.

A. Blaze

B. Spark

C. Fire

D. Ignite

Answer: B

Explanation:

Cloud Dataproc is a managed Apache Spark and Apache Hadoop service that lets you use open source data tools for batch processing, querying, streaming, and machine learning.

Reference: https://cloud.google.com/dataproc/docs/

Question: 108

Which action can a Cloud Dataproc Viewer perform?

A. Submit a job.

B. Create a cluster.

C. Delete a cluster.

D. List the jobs.

Answer: D

Explanation:

A Cloud Dataproc Viewer is limited in its actions based on its role. A viewer can only list clusters, get cluster details, list jobs, get job details, list operations, and get operation details.

Reference:
https://cloud.google.com/dataproc/docs/concepts/iam#iam_roles_and_cloud_dataproc_operations_summary

Question: 109

Dataproc clusters contain many configuration files. To update these files, you will need to use the -properties option. The format for the option is: file_prefix:property= .

A. details
B. value
C. null
D. id

Answer: B

Explanation:

To make updating files and properties easy, the --properties command uses a special format to specify the configuration file and the property and value within the file that should be updated. The formatting is as follows: file_prefix:property=value.

Reference: https://cloud.google.com/dataproc/docs/concepts/cluster-properties#formatting

Question: 110

Scaling a Cloud Dataproc cluster typically involves .

A. increasing or decreasing the number of worker nodes

B. increasing or decreasing the number of master nodes

C. moving memory to run more applications on a single node

D. deleting applications from unused nodes periodically

Answer: A

Explanation:

After creating a Cloud Dataproc cluster, you can scale the cluster by increasing or decreasing the number of worker nodes in the cluster at any time, even when jobs are running on the cluster. Cloud Dataproc clusters are typically scaled to:

1) increase the number of workers to make a job run faster

2) decrease the number of workers to save money

3) increase the number of nodes to expand available Hadoop Distributed Filesystem (HDFS) storage

Reference: https://cloud.google.com/dataproc/docs/concepts/scaling-clusters

Question: 111

Cloud Dataproc charges you only for what you really use with billing.

A. month-by-month

B. minute-by-minute

C. week-by-week

D. hour-by-hour

Answer: B

Explanation:

One of the advantages of Cloud Dataproc is its low cost. Dataproc charges for what you really use with minute-by-minute billing and a low, ten-minute-minimum billing period.

Reference: https://cloud.google.com/dataproc/docs/concepts/overview

Question: 112

The YARN ResourceManager and the HDFS NameNode interfaces are available on a Cloud Dataproc cluster .

A. application node

B. conditional node

C. master node

D. worker node

Answer: C

Explanation:

The YARN ResourceManager and the HDFS NameNode interfaces are available on a Cloud Dataproc cluster master node. The cluster master-host-name is the name of your Cloud Dataproc cluster followed by an -m suffix—for example, if your cluster is named "my-cluster", the master-host-name would be "my-cluster-m".

Reference: https://cloud.google.com/dataproc/docs/concepts/cluster-web-interfaces#interfaces

Question: 113

Which of these is NOT a way to customize the software on Dataproc cluster instances?

A. Set initialization actions

B. Modify configuration files using cluster properties

C. Configure the cluster using Cloud Deployment Manager

D. Log into the master node and make changes from there

Answer: C

Explanation:

You can access the master node of the cluster by clicking the SSH button next to it in the Cloud Console.

You can easily use the --properties option of the dataproc command in the Google Cloud SDK to modify many common configuration files when creating a cluster.

When creating a Cloud Dataproc cluster, you can specify initialization actions in executables and/or scripts that Cloud Dataproc will run on all nodes in your Cloud Dataproc cluster immediately after the cluster is set up.

[https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/init- actions]

Reference: https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/cluster- properties

Question: 114

In order to securely transfer web traffic data from your computer's web browser to the Cloud Dataproc cluster you should use a(n) .

A. VPN connection

B. Special browser

C. SSH tunnel

D. FTP connection

Answer: C

Explanation:

To connect to the web interfaces, it is recommended to use an SSH tunnel to create a secure connection to the master node.

Reference: https://cloud.google.com/dataproc/docs/concepts/cluster-web-interfaces#connecting_to_the_web_interfaces

Question: 115

All Google Cloud Bigtable client requests go through a front-end server they are sent to a Cloud Bigtable node.

A. before

B. after

C. only if

D. once

Answer: A

Explanation:

In a Cloud Bigtable architecture all client requests go through a front-end server before they are sent to a Cloud Bigtable node.

The nodes are organized into a Cloud Bigtable cluster, which belongs to a Cloud Bigtable instance, which is a container for the cluster. Each node in the cluster handles a subset of the requests to the cluster.

When additional nodes are added to a cluster, you can increase the number of simultaneous requests that the cluster can handle, as well as the maximum throughput for the entire cluster.

Reference: https://cloud.google.com/bigtable/docs/overview

Question: 116

What is the general recommendation when designing your row keys for a Cloud Bigtable schema?

A. Include multiple time series values within the row key

B. Keep the row keep as an 8 bit integer

C. Keep your row key reasonably short

D. Keep your row key as long as the field permits

Answer: C

Explanation:

A general guide is to, keep your row keys reasonably short. Long row keys take up additional memory and storage and increase the time it takes to get responses from the Cloud Bigtable server.

Reference: https://cloud.google.com/bigtable/docs/schema-design#row-keys

Question: 117

Which of the following statements is NOT true regarding Bigtable access roles?

A. Using IAM roles, you cannot give a user access to only one table in a project, rather than all tables in a project.

B. To give a user access to only one table in a project, grant the user the Bigtable Editor role for that table.

C. You can configure access control only at the project level.

D. To give a user access to only one table in a project, you must configure access through your application.

Answer: B

Explanation:

For Cloud Bigtable, you can configure access control at the project level. For example, you can grant the ability to:

Read from, but not write to, any table within the project.

Read from and write to any table within the project, but not manage instances.

Read from and write to any table within the project, and manage instances.

Reference: https://cloud.google.com/bigtable/docs/access-control

Question: 118

For the best possible performance, what is the recommended zone for your Compute Engine instance and Cloud Bigtable instance?

A. Have the Compute Engine instance in the furthest zone from the Cloud Bigtable instance.

B. Have both the Compute Engine instance and the Cloud Bigtable instance to be in different zones.

C. Have both the Compute Engine instance and the Cloud Bigtable instance to be in the same zone.

D. Have the Cloud Bigtable instance to be in the same zone as all of the consumers of your data.

Answer: C

Explanation:

It is recommended to create your Compute Engine instance in the same zone as your Cloud Bigtable instance for the best possible performance,

If it's not possible to create a instance in the same zone, you should create your instance in another zone within the same region. For example, if your Cloud Bigtable instance is located in us-central1-b, you could create your instance in us-central1-f. This change may result in several milliseconds of additional latency for each Cloud Bigtable request.

It is recommended to avoid creating your Compute Engine instance in a different region from your Cloud Bigtable instance, which can add hundreds of milliseconds of latency to each Cloud Bigtable request.

Reference: https://cloud.google.com/bigtable/docs/creating-compute-instance

Question: 119

Which row keys are likely to cause a disproportionate number of reads and/or writes on a particular node in a Bigtable cluster (select 2 answers)?

A. A sequential numeric ID

B. A timestamp followed by a stock symbol

C. A non-sequential numeric ID

D. A stock symbol followed by a timestamp

Answer: AB

Explanation:

...using a timestamp as the first element of a row key can cause a variety of problems.

In brief, when a row key for a time series includes a timestamp, all of your writes will target a single node; fill that node; and then move onto the next node in the cluster, resulting in hotspotting.

Suppose your system assigns a numeric ID to each of your application's users. You might be tempted to use the user's numeric ID as the row key for your table. However, since new users are more likely to be active users, this approach is likely to push most of your traffic to a small number of nodes.
[https://cloud.google.com/bigtable/docs/schema-design]

Reference: https://cloud.google.com/bigtable/docs/schema-design-time-series#ensure_that_your_row_key_avoids_hotspotting

Question: 120

When a Cloud Bigtable node fails, _____ is lost.

A. all data
B. no data
C. the last transaction
D. the time dimension

Answer: B

Explanation:

A Cloud Bigtable table is sharded into blocks of contiguous rows, called tablets, to help balance the workload of queries. Tablets are stored on Colossus, Google's file system, in SSTable format. Each tablet is associated with a specific Cloud Bigtable node.

Data is never stored in Cloud Bigtable nodes themselves; each node has pointers to a set of tablets that are stored on Colossus. As a result:

Rebalancing tablets from one node to another is very fast, because the actual data is not copied.

Cloud Bigtable simply updates the pointers for each node.

Recovery from the failure of a Cloud Bigtable node is very fast, because only metadata needs to be migrated to the replacement node.

When a Cloud Bigtable node fails, no data is lost

Reference: https://cloud.google.com/bigtable/docs/overview

Question: 121

Which is not a valid reason for poor Cloud Bigtable performance?

A. The workload isn't appropriate for Cloud Bigtable.
B. The table's schema is not designed correctly.
C. The Cloud Bigtable cluster has too many nodes.
D. There are issues with the network connection.

Answer: C

Explanation:

The Cloud Bigtable cluster doesn't have enough nodes. If your Cloud Bigtable cluster is overloaded, adding more nodes can improve performance. Use the monitoring tools to check whether the cluster is overloaded.

Reference: https://cloud.google.com/bigtable/docs/performance

Question: 122

Which is the preferred method to use to avoid hotspotting in time series data in Bigtable?

A. Field promotion

B. Randomization

C. Salting

D. Hashing

Answer: A

Explanation:

By default, prefer field promotion. Field promotion avoids hotspotting in almost all cases, and it tends to make it easier to design a row key that facilitates queries.

Reference: https://cloud.google.com/bigtable/docs/schema-design-time-series#ensure_that_your_row_key_avoids_hotspotting

Question: 123

When you design a Google Cloud Bigtable schema it is recommended that you

A. Avoid schema designs that are based on NoSQL concepts

B. Create schema designs that are based on a relational database design

C. Avoid schema designs that require atomicity across rows

D. Create schema designs that require atomicity across rows

Answer: C

Explanation:

All operations are atomic at the row level. For example, if you update two rows in a table, it's possible that one row will be updated successfully and the other update will fail. Avoid schema designs that require atomicity across rows.

Reference: https://cloud.google.com/bigtable/docs/schema-design#row-keys

Question: 124

Which of the following is NOT a valid use case to select HDD (hard disk drives) as the storage for Google Cloud Bigtable?

A. You expect to store at least 10 TB of data.

B. You will mostly run batch workloads with scans and writes, rather than frequently executing random reads of a small number of rows.

C. You need to integrate with Google BigQuery.

D. You will not use the data to back a user-facing or latency-sensitive application.

Answer: C

Explanation:

For example, if you plan to store extensive historical data for a large number of remote-sensing devices and then use the data to generate daily reports, the cost savings for HDD storage may justify the performance tradeoff. On the other hand, if you plan to use the data to display a real-time dashboard, it probably would not make sense to use HDD storage—reads would be much more frequent in this case, and reads are much slower with HDD storage.

Reference: https://cloud.google.com/bigtable/docs/choosing-ssd-hdd

Question: 125

Cloud Bigtable is Google's Big Data database service.

A. Relational

B. mySQL

C. NoSQL

D. SQL Server

Answer: C

Explanation:

Cloud Bigtable is Google's NoSQL Big Data database service. It is the same database that Google uses for services, such as Search, Analytics, Maps, and Gmail.

It is used for requirements that are low latency and high throughput including Internet of Things (IoT), user analytics, and financial data analysis.

Reference: https://cloud.google.com/bigtable/

Question: 126

When you store data in Cloud Bigtable, what is the recommended minimum amount of stored data?

A. 500 TB

B. 1 GB

C. 1 TB

D. 500 GB

Answer: C

Explanation:

Cloud Bigtable is not a relational database. It does not support SQL queries, joins, or multi-row transactions.

It is not a good solution for less than 1 TB of data.

Reference:
https://cloud.google.com/bigtable/docs/overview#title_short_and_other_storage_options

Question: 127

If you're running a performance test that depends upon Cloud Bigtable, all the choices except one below are recommended steps. Which is NOT a recommended step to follow?

A. Do not use a production instance.

B. Run your test for at least 10 minutes.

C. Before you test, run a heavy pre-test for several minutes.

D. Use at least 300 GB of data.

Answer: A

Explanation:

If you're running a performance test that depends upon Cloud Bigtable, be sure to follow these steps as you plan and execute your test:

Use a production instance. A development instance will not give you an accurate sense of how a production instance performs under load.

Use at least 300 GB of data. Cloud Bigtable performs best with 1 TB or more of data. However, 300 GB of data is enough to provide reasonable results in a performance test on a 3-node cluster. On larger clusters, use 100 GB of data per node.

Before you test, run a heavy pre-test for several minutes. This step gives Cloud Bigtable a chance to balance data across your nodes based on the access patterns it observes.

Run your test for at least 10 minutes. This step lets Cloud Bigtable further optimize your data, and it helps ensure that you will test reads from disk as well as cached reads from memory.

Reference: https://cloud.google.com/bigtable/docs/performance

Question: 128

Cloud Bigtable is a recommended option for storing very large amounts of

A. multi-keyed data with very high latency

B. multi-keyed data with very low latency

C. single-keyed data with very low latency

D. single-keyed data with very high latency

Answer: C

Explanation:

Cloud Bigtable is a sparsely populated table that can scale to billions of rows and thousands of columns, allowing you to store terabytes or even petabytes of data. A single value in each row is indexed; this value is known as the row key. Cloud Bigtable is ideal for storing very large amounts of single-keyed data with very low latency. It supports high read and write throughput at low latency, and it is an ideal data source for MapReduce operations.

Reference: https://cloud.google.com/bigtable/docs/overview

Question: 129

Google Cloud Bigtable indexes a single value in each row. This value is called the

A. primary key

B. unique key

C. row key

D. master key

Answer: C

Explanation:

Cloud Bigtable is a sparsely populated table that can scale to billions of rows and thousands of columns, allowing you to store terabytes or even petabytes of data. A single value in each row is indexed; this value is known as the row key.

Reference: https://cloud.google.com/bigtable/docs/overview

Question: 130

What is the HBase Shell for Cloud Bigtable?

A. The HBase shell is a GUI based interface that performs administrative tasks, such as creating and deleting tables.

B. The HBase shell is a command-line tool that performs administrative tasks, such as creating and deleting tables.

C. The HBase shell is a hypervisor based shell that performs administrative tasks, such as creating and deleting new virtualized instances.

D. The HBase shell is a command-line tool that performs only user account management functions to grant access to Cloud Bigtable instances.

Answer: B

Explanation:

The HBase shell is a command-line tool that performs administrative tasks, such as creating and deleting tables. The Cloud Bigtable HBase client for Java makes it possible to use the HBase shell to connect to Cloud Bigtable.

Reference: https://cloud.google.com/bigtable/docs/installing-hbase-shell

Question: 131

What is the recommended action to do in order to switch between SSD and HDD storage for your Google Cloud Bigtable instance?

A. create a third instance and sync the data from the two storage types via batch jobs

B. export the data from the existing instance and import the data into a new instance

C. run parallel instances where one is HDD and the other is SDD

D. the selection is final and you must resume using the same storage type

Answer: B

Explanation:

When you create a Cloud Bigtable instance and cluster, your choice of SSD or HDD storage for the cluster is permanent. You cannot use the Google Cloud Platform Console to change the type of storage that is used for the cluster.

If you need to convert an existing HDD cluster to SSD, or vice-versa, you can export the data from the existing instance and import the data into a new instance. Alternatively, you can write

a Cloud Dataflow or Hadoop MapReduce job that copies the data from one instance to another.

Reference: https://cloud.google.com/bigtable/docs/choosing-ssd-hdd–

Topic 6, Main Questions Set C

Question: 132

You are training a spam classifier. You notice that you are overfitting the training dat

a. Which three actions can you take to resolve this problem? (Choose three.)

A. Get more training examples

B. Reduce the number of training examples

C. Use a smaller set of features

D. Use a larger set of features

E. Increase the regularization parameters

F. Decrease the regularization parameters

Answer: ADF

Explanation:

Question: 133

You are implementing security best practices on your data pipeline. Currently, you are manually executing jobs as the Project Owner. You want to automate these jobs by taking nightly batch files containing non-public information from Google Cloud Storage, processing them with a Spark Scala job on a Google Cloud Dataproc cluster, and depositing the results into Google BigQuery.

How should you securely run this workload?

A. Restrict the Google Cloud Storage bucket so only you can see the files

B. Grant the Project Owner role to a service account, and run the job with it

C. Use a service account with the ability to read the batch files and to write to BigQuery

D. Use a user account with the Project Viewer role on the Cloud Dataproc cluster to read the batch files and write to BigQuery

Answer: B

Explanation:

Question: 134

You are using Google BigQuery as your data warehouse. Your users report that the following simple query is running very slowly, no matter when they run the query:

SELECT country, state, city FROM [myproject:mydataset.mytable] GROUP BY country

You check the query plan for the query and see the following output in the Read section of Stage:1:

What is the most likely cause of the delay for this query?

A. Users are running too many concurrent queries in the system

B. The [myproject:mydataset.mytable] table has too many partitions

C. Either the state or the city columns in the [myproject:mydataset.mytable] table have too many NULL values

D. Most rows in the [myproject:mydataset.mytable] table have the same value in the country column, causing data skew

Answer: A

Explanation:

Question: 135

Your globally distributed auction application allows users to bid on items. Occasionally, users place identical bids at nearly identical times, and different application servers process those bids. Each bid event contains the item, amount, user, and timestamp. You want to collate those bid events into a single location in real time to determine which user bid first. What should you do?

A. Create a file on a shared file and have the application servers write all bid events to that file. Process the file with Apache Hadoop to identify which user bid first.

B. Have each application server write the bid events to Cloud Pub/Sub as they occur. Push the events from Cloud Pub/Sub to a custom endpoint that writes the bid event information into Cloud SQL.

C. Set up a MySQL database for each application server to write bid events into. Periodically query each of those distributed MySQL databases and update a master MySQL database with bid event information.

D. Have each application server write the bid events to Google Cloud Pub/Sub as they occur. Use a pull subscription to pull the bid events using Google Cloud Dataflow. Give the bid for each item to the user in the bid event that is processed first.

Answer: C

Explanation:

Question: 136

Your organization has been collecting and analyzing data in Google BigQuery for 6 months. The majority of the data analyzed is placed in a time-partitioned table named events_partitioned. To reduce the cost of queries, your organization created a view called events, which queries only the last 14 days of dat

a. The view is described in legacy SQL. Next month, existing applications will be connecting to BigQuery to read the events data via an ODBC connection. You need to ensure the applications can connect. Which two actions should you take? (Choose two.)

A. Create a new view over events using standard SQL

B. Create a new partitioned table using a standard SQL query

C. Create a new view over events_partitioned using standard SQL

D. Create a service account for the ODBC connection to use for authentication

E. Create a Google Cloud Identity and Access Management (Cloud IAM) role for the ODBC connection and shared "events"

Answer: AE

Explanation:

Question: 137

You have enabled the free integration between Firebase Analytics and Google BigQuery. Firebase now automatically creates a new table daily in BigQuery in the format app_events_YYYYMMDD. You want to query all of the tables for the past 30 days in legacy SQL. What should you do?

A. Use the TABLE_DATE_RANGE function

B. Use the WHERE_PARTITIONTIME pseudo column

C. Use WHERE date BETWEEN YYYY-MM-DD AND YYYY-MM-DD

D. Use SELECT IF.(date >= YYYY-MM-DD AND date <= YYYY-MM-DD

Answer: A

Explanation:

Reference: https://cloud.google.com/blog/products/gcp/using-bigquery-and-firebase-analytics-to- understandyour-mobile-app?hl=am

Question: 138

Your company is currently setting up data pipelines for their campaign. For all the Google Cloud Pub/Sub streaming data, one of the important business requirements is to be able to periodically identify the inputs and their timings during their campaign. Engineers have decided to use windowing and transformation in Google Cloud Dataflow for this purpose. However, when testing this feature, they find that the Cloud Dataflow job fails for the all streaming insert. What is the most likely cause of this problem?

A. They have not assigned the timestamp, which causes the job to fail

B. They have not set the triggers to accommodate the data coming in late, which causes the job to fail

C. They have not applied a global windowing function, which causes the job to fail when the pipeline is created

D. They have not applied a non-global windowing function, which causes the job to fail when the pipeline is created

Answer: C

Explanation:

Question: 139

You architect a system to analyze seismic dat

a. Your extract, transform, and load (ETL) process runs as a series of MapReduce jobs on an Apache Hadoop cluster. The ETL process takes days to process a data set because some steps are computationally expensive. Then you discover that a sensor calibration step has been omitted. How should you change your ETL process to carry out sensor calibration systematically in the future?

A. Modify the transformMapReduce jobs to apply sensor calibration before they do anything else.

B. Introduce a new MapReduce job to apply sensor calibration to raw data, and ensure all other MapReduce jobs are chained after this.

C. Add sensor calibration data to the output of the ETL process, and document that all users need to apply sensor

calibration themselves.

D. Develop an algorithm through simulation to predict variance of data output from the last MapReduce job based on calibration factors, and apply the correction to all data.

Answer: A

Explanation:

Question: 140

An online retailer has built their current application on Google App Engine. A new initiative at the company mandates that they extend their application to allow their customers to transact directly via the application.

They need to manage their shopping transactions and analyze combined data from multiple datasets using a business intelligence (BI) tool. They want to use only a single database for this purpose.

Which Google Cloud database should they choose?

A. BigQuery

B. Cloud SQL

C. Cloud BigTable

D. Cloud Datastore

Answer: C

Explanation:

Reference: https://cloud.google.com/solutions/business-intelligence/

Question: 141

You launched a new gaming app almost three years ago. You have been uploading log files from the previous day to a separate Google BigQuery table with the table name format LOGS_yyyymmdd. You have been using table wildcard functions to generate daily and monthly reports for all time ranges. Recently, you discovered that some queries that cover long date ranges are exceeding the limit of 1,000 tables and failing. How can you resolve this issue?

A. Convert all daily log tables into date-partitioned tables

B. Convert the sharded tables into a single partitioned table

C. Enable query caching so you can cache data from previous months

D. Create separate views to cover each month, and query from these views

Answer: A

Explanation:

Question: 142

Your analytics team wants to build a simple statistical model to determine which customers are most likely to work with your company again, based on a few different metrics. They want to run the model on Apache Spark, using data housed in Google Cloud Storage, and you have recommended using Google Cloud Dataproc to execute this job. Testing has shown that this workload can run in approximately 30 minutes on a 15-node cluster, outputting the results into Google BigQuery. The plan is to run this workload weekly. How should you optimize the cluster for cost?

A. Migrate the workload to Google Cloud Dataflow

B. Use pre-emptible virtual machines (VMs) for the cluster

C. Use a higher-memory node so that the job runs faster

D. Use SSDs on the worker nodes so that the job can run faster

Answer: A

Explanation:

Question: 143

Your company receives both batch- and stream-based event dat

a. You want to process the data using Google Cloud Dataflow over a predictable time period. However, you realize that in some instances data can arrive late or out of order. How should you design your Cloud Dataflow pipeline to handle data that is late or out of order?

A. Set a single global window to capture all the data.

B. Set sliding windows to capture all the lagged data.

C. Use watermarks and timestamps to capture the lagged data.

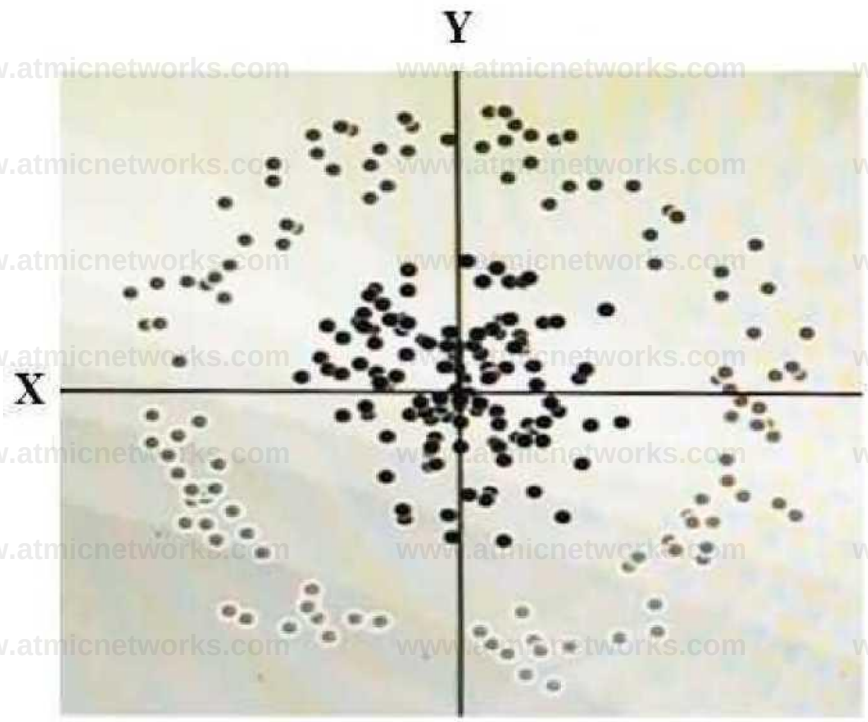
D. Ensure every datasource type (stream or batch) has a timestamp, and use the timestamps to define the logic for lagged data.

Answer: B

Explanation:

Question: 144

You have some data, which is shown in the graphic below. The two dimensions are X and Y, and the shade of each dot represents what class it is. You want to classify this data accurately using a linear algorithm.



To do this you need to add a synthetic feature. What should the value of that feature be?

A. X^2+Y^2

B. X^2

C. Y^2

D. cos(X)

Answer: D

Explanation:

Question: 145

You are integrating one of your internal IT applications and Google BigQuery, so users can query BigQuery from the application's interface. You do not want individual users to authenticate to BigQuery and you do not want to give them access to the dataset. You need to securely access BigQuery from your IT application.

What should you do?

A. Create groups for your users and give those groups access to the dataset

B. Integrate with a single sign-on (SSO) platform, and pass each user's credentials along with the query request

C. Create a service account and grant dataset access to that account. Use the service account's private key to access the dataset

D. Create a dummy user and grant dataset access to that user. Store the username and password for that user in a file on the files system, and use those credentials to access the BigQuery dataset

Answer: C

Explanation:

Question: 146

You set up a streaming data insert into a Redis cluster via a Kafka cluster. Both clusters are running on Compute Engine instances. You need to encrypt data at rest with encryption keys that you can create, rotate, and destroy as needed. What should you do?

A. Create a dedicated service account, and use encryption at rest to reference your data stored in your Compute Engine cluster instances as part of your API service calls.

B. Create encryption keys in Cloud Key Management Service. Use those keys to encrypt your data in all of the

Compute Engine cluster instances.

C. Create encryption keys locally. Upload your encryption keys to Cloud Key Management Service. Use those keys to encrypt your data in all of the Compute Engine cluster instances.

D. Create encryption keys in Cloud Key Management Service. Reference: those keys in your API service calls when accessing the data in your Compute Engine cluster instances.

Answer: C

Explanation:

Question: 147

You are developing an application that uses a recommendation engine on Google Cloud. Your solution should display new videos to customers based on past views. Your solution needs to generate labels for the entities in videos that the customer has viewed. Your design must be able to provide very fast filtering suggestions based on data from other customer preferences on several TB of dat

a. What should you do?

A. Build and train a complex classification model with Spark MLlib to generate labels and filter the results. Deploy the models using Cloud Dataproc. Call the model from your application.

B. Build and train a classification model with Spark MLlib to generate labels. Build and train a second classification model with Spark MLlib to filter results to match customer preferences. Deploy the models using Cloud Dataproc. Call the models from your application.

C. Build an application that calls the Cloud Video Intelligence API to generate labels. Store data in Cloud Bigtable, and filter the predicted labels to match the user's viewing history to generate preferences.

D. Build an application that calls the Cloud Video Intelligence API to generate labels. Store data in Cloud SQL, and join and filter the predicted labels to match the user's viewing history to generate preferences.

Answer: C

Explanation:

Question: 148

You are selecting services to write and transform JSON messages from Cloud Pub/Sub to BigQuery for a data pipeline on Google Cloud. You want to minimize service costs. You also want to monitor and accommodate input data volume that will vary in size with minimal manual intervention. What should you do?

A. Use Cloud Dataproc to run your transformations. Monitor CPU utilization for the cluster. Resize the number of worker nodes in your cluster via the command line.

B. Use Cloud Dataproc to run your transformations. Use the diagnose command to generate an operational output archive. Locate the bottleneck and adjust cluster resources.

C. Use Cloud Dataflow to run your transformations. Monitor the job system lag with Stackdriver. Use the default autoscaling setting for worker instances.

D. Use Cloud Dataflow to run your transformations. Monitor the total execution time for a sampling of jobs. Configure the job to use non-default Compute Engine machine types when needed.

Answer: B

Explanation:

Question: 149

Your infrastructure includes a set of YouTube channels. You have been tasked with creating a process for sending the YouTube channel data to Google Cloud for analysis. You want to design a solution that allows your world-wide marketing teams to perform ANSI SQL and other types of analysis on up-to- date YouTube channels log dat

a. How should you set up the log data transfer into Google Cloud?

A. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Multi-Regional storage bucket as a final destination.

B. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Regional bucket as a final

destination.

C. Use BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage Multi- Regional storage bucket as a final destination.

D. Use BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage Regional storage bucket as a final destination.

Answer: B

Explanation:

Question: 150

You are designing storage for very large text files for a data pipeline on Google Cloud. You want to support ANSI SQL queries. You also want to support compression and parallel load from the input locations using Google recommended practices. What should you do?

A. Transform text files to compressed Avro using Cloud Dataflow. Use BigQuery for storage and query.

B. Transform text files to compressed Avro using Cloud Dataflow. Use Cloud Storage and BigQuery permanent linked tables for query.

C. Compress text files to gzip using the Grid Computing Tools. Use BigQuery for storage and query.

D. Compress text files to gzip using the Grid Computing Tools. Use Cloud Storage, and then import into Cloud Bigtable for query.

Answer: D

Explanation:

Question: 151

You are developing an application on Google Cloud that will automatically generate subject labels for users' blog posts. You are under competitive pressure to add this feature quickly, and you have no additional developer resources. No one on your team has experience with machine learning. What should you do?

A. Call the Cloud Natural Language API from your application. Process the generated Entity Analysis as labels.

B. Call the Cloud Natural Language API from your application. Process the generated Sentiment Analysis as labels.

C. Build and train a text classification model using TensorFlow. Deploy the model using Cloud Machine Learning Engine. Call the model from your application and process the results as labels.

D. Build and train a text classification model using TensorFlow. Deploy the model using a Kubernetes Engine cluster. Call the model from your application and process the results as labels.

Answer: B

Explanation:

Question: 152

You are designing storage for 20 TB of text files as part of deploying a data pipeline on Google Cloud. Your input data is in CSV format. You want to minimize the cost of querying aggregate values for multiple users who will query the data in Cloud Storage with multiple engines. Which storage service and schema design should you use?

A. Use Cloud Bigtable for storage. Install the HBase shell on a Compute Engine instance to query the Cloud Bigtable data.

B. Use Cloud Bigtable for storage. Link as permanent tables in BigQuery for query.

C. Use Cloud Storage for storage. Link as permanent tables in BigQuery for query.

D. Use Cloud Storage for storage. Link as temporary tables in BigQuery for query.

Answer: A

Explanation:

Question: 153

You are designing storage for two relational tables that are part of a 10-TB database on Google Cloud.

You want to support transactions that scale horizontally. You also want to optimize data for range queries on nonkey columns. What should you do?

A. Use Cloud SQL for storage. Add secondary indexes to support query patterns.

B. Use Cloud SQL for storage. Use Cloud Dataflow to transform data to support query patterns.

C. Use Cloud Spanner for storage. Add secondary indexes to support query patterns.

D. Use Cloud Spanner for storage. Use Cloud Dataflow to transform data to support query patterns.

Answer: D

Explanation:

Reference: https://cloud.google.com/solutions/data-lifecycle-cloud-platform

Question: 154

Your financial services company is moving to cloud technology and wants to store 50 TB of financial timeseries data in the cloud. This data is updated frequently and new data will be streaming in all the time. Your company also wants to move their existing Apache Hadoop jobs to the cloud to get insights into this data.

Which product should they use to store the data?

A. Cloud Bigtable

B. Google BigQuery

C. Google Cloud Storage

D. Google Cloud Datastore

Answer: A

Explanation:

Reference: https://cloud.google.com/bigtable/docs/schema-design-time-series

Question: 155

An organization maintains a Google BigQuery dataset that contains tables with user-level dat

A. They want to expose aggregates of this data to other Google Cloud projects, while still controlling access to the user-level data. Additionally, they need to minimize their overall storage cost and ensure the analysis cost for other projects is assigned to those projects. What should they do?

A. Create and share an authorized view that provides the aggregate results.

B. Create and share a new dataset and view that provides the aggregate results.

C. Create and share a new dataset and table that contains the aggregate results.

D. Create dataViewer Identity and Access Management (IAM) roles on the dataset to enable sharing.

Answer: D

Explanation:

Reference: https://cloud.google.com/bigquery/docs/access-control

Question: 156

Government regulations in your industry mandate that you have to maintain an auditable record of access to certain types of dat

A. Assuming that all expiring logs will be archived correctly, where should you store data that is subject to that mandate?

A. Encrypted on Cloud Storage with user-supplied encryption keys. A separate decryption key will be given to each

authorized user.

B. In a BigQuery dataset that is viewable only by authorized personnel, with the Data Access log used to provide the auditability.

C. In Cloud SQL, with separate database user names to each user. The Cloud SQL Admin activity logs will be used to provide the auditability.

D. In a bucket on Cloud Storage that is accessible only by an AppEngine service that collects user information and logs the access before providing a link to the bucket.

Answer: B

Explanation:

Question: 157

Your neural network model is taking days to train. You want to increase the training speed. What can you do?

A. Subsample your test dataset.

B. Subsample your training dataset.

C. Increase the number of input features to your model.

D. Increase the number of layers in your neural network.

Answer: D

Explanation:

Reference: https://towardsdatascience.com/how-to-increase-the-accuracy-of-a-neural-network- 9f5d1c6f407d

Question: 158

You are responsible for writing your company's ETL pipelines to run on an Apache Hadoop cluster. The pipeline will require some checkpointing and splitting pipelines. Which method should you use to write the

pipelines?

A. PigLatin using Pig
B. HiveQL using Hive
C. Java using MapReduce
D. Python using MapReduce

Answer: D

Explanation:

Question: 159

Your company maintains a hybrid deployment with GCP, where analytics are performed on your anonymized customer dat

a. The data are imported to Cloud Storage from your data center through parallel uploads to a data transfer server running on GCP. Management informs you that the daily transfers take too long and have

asked you to fix the problem. You want to maximize transfer speeds. Which action should you take?

A. Increase the CPU size on your server.
B. Increase the size of the Google Persistent Disk on your server.
C. Increase your network bandwidth from your datacenter to GCP.
D. Increase your network bandwidth from Compute Engine to Cloud Storage.

Answer: C

Explanation:

Question: 160

After migrating ETL jobs to run on BigQuery, you need to verify that the output of the migrated jobs is the same as the

output of the original. You've loaded a table containing the output of the original job and want to compare the contents with output from the migrated job to show that they are

identical. The tables do not contain a primary key column that would enable you to join them together for comparison.

What should you do?

A. Select random samples from the tables using the RAND() function and compare the samples.

B. Select random samples from the tables using the HASH() function and compare the samples.

C. Use a Dataproc cluster and the BigQuery Hadoop connector to read the data from each table and calculate a hash from non-timestamp columns of the table after sorting. Compare the hashes of each table.

D. Create stratified random samples using the OVER() function and compare equivalent samples from each table.

Answer: B

Explanation:

Question: 161

You are a head of BI at a large enterprise company with multiple business units that each have different priorities and budgets. You use on-demand pricing for BigQuery with a quota of 2K concurrent on-demand slots per project. Users at your organization sometimes don't get slots to execute their query and you need to correct this. You'd like to avoid introducing new projects to your account.

What should you do?

A. Convert your batch BQ queries into interactive BQ queries.

B. Create an additional project to overcome the 2K on-demand per-project quota.

C. Switch to flat-rate pricing and establish a hierarchical priority model for your projects.

D. Increase the amount of concurrent slots per project at the Quotas page at the Cloud Console.

Answer: C

Explanation:

Reference: https://cloud.google.com/blog/products/gcp/busting-12-myths-about-bigquery

Question: 162

You have an Apache Kafka Cluster on-prem with topics containing web application logs. You need to replicate the data to Google Cloud for analysis in BigQuery and Cloud Storage. The preferred replication method is mirroring to avoid deployment of Kafka Connect plugins.

What should you do?

A. Deploy a Kafka cluster on GCE VM Instances. Configure your on-prem cluster to mirror your topics to the cluster running in GCE. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.

B. Deploy a Kafka cluster on GCE VM Instances with the PubSub Kafka connector configured as a Sink connector. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.

C. Deploy the PubSub Kafka connector to your on-prem Kafka cluster and configure PubSub as a Source connector. Use a Dataflow job to read fron PubSub and write to GCS.

D. Deploy the PubSub Kafka connector to your on-prem Kafka cluster and configure PubSub as a Sink connector. Use a Dataflow job to read fron PubSub and write to GCS.

Answer: A

Explanation:

Question: 163

You've migrated a Hadoop job from an on-prem cluster to dataproc and GCS. Your Spark job is a complicated analytical workload that consists of many shuffing operations and initial data are parquet files (on average 200-400 MB size each). You see some degradation in performance after the migration to Dataproc, so you'd like to optimize for it.

You need to keep in mind that your organization is very cost-sensitive, so you'd like to continue using Dataproc on preemptibles (with 2 non-preemptible workers only) for this workload.

What should you do?

A. Increase the size of your parquet files to ensure them to be 1 GB minimum.

B. Switch to TFRecords formats (appr. 200MB per file) instead of parquet files.

C. Switch from HDDs to SSDs, copy initial data from GCS to HDFS, run the Spark job and copy results back to GCS.

D. Switch from HDDs to SSDs, override the preemptible VMs configuration to increase the boot disk size.

Answer: C

Explanation:

To optimize the performance of a complex Spark job on Dataproc that heavily relies on shufling operations, and given the cost constraints of using preemptible VMs, switching from HDDs to SSDs and using HDFS as an intermediate storage layer can significantly improve performance. Here's why option C is the best choice:

Explanation:

Performance of SSDs:

SSDs provide much faster read and write speeds compared to HDDs, which is crucial for performance-intensive operations like shufling in Spark jobs.

Using SSDs can reduce I/O bottlenecks during the shufle phase of your Spark job, improving overall

job performance.

Intermediate Storage with HDFS:

Copying data from Google Cloud Storage (GCS) to HDFS for intermediate storage can reduce latency compared to reading directly from GCS.

HDFS provides better locality and faster data access within the Dataproc cluster, which can significantly improve the efficiency of shufling and other I/O operations.

Cost Considerations:

Although SSDs are more expensive than HDDs, the performance improvement for shufle-heavy workloads can justify the

cost, especially if the improved performance reduces the overall runtime and thereby the cost of using preemptible VMs.

Using preemptible VMs with SSDs for this workload balances the cost and performance trade-offs effectively.

Question: 164

Your team is responsible for developing and maintaining ETLs in your company. One of your Dataflow jobs is failing because of some errors in the input data, and you need to improve reliability of the pipeline (incl. being able to reprocess all failing data).

What should you do?

A. Add a filtering step to skip these types of errors in the future, extract erroneous rows from logs.

B. Add a try… catch block to your DoFn that transforms the data, extract erroneous rows from logs.

C. Add a try… catch block to your DoFn that transforms the data, write erroneous rows to PubSub directly from the DoFn.

D. Add a try… catch block to your DoFn that transforms the data, use a sideOutput to create a PCollection that can be stored to PubSub later.

Answer: C

Explanation:

Question: 165

You're training a model to predict housing prices based on an available dataset with real estate properties. Your plan is to train a fully connected neural net, and you've discovered that the dataset contains latitude and longtitude of the property. Real estate professionals have told you that the location of the property is highly influential on price, so you'd like to engineer a feature that incorporates this physical dependency.

What should you do?

A. Provide latitude and longtitude as input vectors to your neural net.

B. Create a numeric column from a feature cross of latitude and longtitude.

C. Create a feature cross of latitude and longtitude, bucketize at the minute level and use L1 regularization during optimization.

D. Create a feature cross of latitude and longtitude, bucketize it at the minute level and use L2 regularization during optimization.

Answer: B

Explanation:

Reference: https://cloud.google.com/bigquery/docs/gis-dataa

To engineer a feature that incorporates the physical dependency of location on housing prices for a neural network, creating a numeric column from a feature cross of latitude and longitude is the most effective approach. Here's why option B is the best choice:

Explanation:

Feature Crosses:
Feature crosses combine multiple features into a single feature that captures the interaction between them. For location data, a feature cross of latitude and longitude can capture spatial dependencies that affect housing prices.

This approach allows the neural network to learn complex patterns related to geographic location more effectively than using raw latitude and longitude values.

Numerical Representation:

Converting the feature cross into a numeric column simplifies the input for the neural network and can improve the model's ability to learn from the data.

This method ensures that the model can leverage the combined information from both latitude and longitude in a meaningful way.

Model Training:

Using a numeric column for the feature cross helps in regularizing the model and prevents overfitting, which is crucial for

achieving good generalization on unseen data.

Question: 166

You are deploying MariaDB SQL databases on GCE VM Instances and need to configure monitoring and alerting. You want to collect metrics including network connections, disk IO and replication status from MariaDB with minimal development effort and use StackDriver for dashboards and alerts.

What should you do?

A. Install the OpenCensus Agent and create a custom metric collection application with a StackDriver exporter.

B. Place the MariaDB instances in an Instance Group with a Health Check.

C. Install the StackDriver Logging Agent and configure fluentd in_tail plugin to read MariaDB logs.

D. Install the StackDriver Agent and configure the MySQL plugin.

Answer: C

Explanation:

Question: 167

You work for a bank. You have a labelled dataset that contains information on already granted loan application and whether these applications have been defaulted. You have been asked to train a model to predict default rates for credit applicants.

What should you do?

A. Increase the size of the dataset by collecting additional data.

B. Train a linear regression to predict a credit default risk score.

C. Remove the bias from the data and collect applications that have been declined loans.

D. Match loan applicants with their social profiles to enable feature engineering.

Answer: B

Explanation:

Question: 168

You need to migrate a 2TB relational database to Google Cloud Platform. You do not have the resources to significantly refactor the application that uses this database and cost to operate is of primary concern.

Which service do you select for storing and serving your data?

A. Cloud Spanner

B. Cloud Bigtable

C. Cloud Firestore

D. Cloud SQL

Answer: D

Explanation:

Question: 169

You're using Bigtable for a real-time application, and you have a heavy load that is a mix of read and writes. You've recently identified an additional use case and need to perform hourly an analytical job to calculate certain statistics across the whole database. You need to ensure both the reliability of your production application as well as the analytical workload.

What should you do?

A. Export Bigtable dump to GCS and run your analytical job on top of the exported files.

B. Add a second cluster to an existing instance with a multi-cluster routing, use live-traffic app profile for your regular workload and batch-analytics profile for the analytics workload.

C. Add a second cluster to an existing instance with a single-cluster routing, use live-traffic app profile for your regular workload and batch-analytics profile for the analytics workload.

D. Increase the size of your existing cluster twice and execute your analytics workload on your new resized cluster.

Answer: B

Explanation:

Question: 170

You are designing an Apache Beam pipeline to enrich data from Cloud Pub/Sub with static reference data from BigQuery. The reference data is small enough to fit in memory on a single worker. The pipeline should write enriched results to

BigQuery for analysis. Which job type and transforms should this pipeline use?

A. Batch job, PubSubIO, side-inputs

B. Streaming job, PubSubIO, JdbcIO, side-outputs

C. Streaming job, PubSubIO, BigQueryIO, side-inputs

D. Streaming job, PubSubIO, BigQueryIO, side-outputs

Answer: C

Explanation:

Question: 171

You have a data pipeline that writes data to Cloud Bigtable using well-designed row keys. You want to monitor your pipeline to determine when to increase the size of you Cloud Bigtable cluster. Which two actions can you take to accomplish this? Choose 2 answers.

A. Review Key Visualizer metrics. Increase the size of the Cloud Bigtable cluster when the Read pressure index is above 100.

B. Review Key Visualizer metrics. Increase the size of the Cloud Bigtable cluster when the Write pressure index is above 100.

C. Monitor the latency of write operations. Increase the size of the Cloud Bigtable cluster when there is a sustained increase in write latency.

D. Monitor storage utilization. Increase the size of the Cloud Bigtable cluster when utilization increases above 70% of max capacity.

E. Monitor latency of read operations. Increase the size of the Cloud Bigtable cluster of read operations take longer than 100 ms.

Answer: A,C

Explanation:

Question: 172

You want to analyze hundreds of thousands of social media posts daily at the lowest cost and with the fewest steps.

You have the following requirements:

You will batch-load the posts once per day and run them through the Cloud Natural Language API.

You will extract topics and sentiment from the posts.

You must store the raw posts for archiving and reprocessing.

You will create dashboards to be shared with people both inside and outside your organization.

You need to store both the data extracted from the API to perform analysis as well as the raw social media posts for historical archiving. What should you do?

A. Store the social media posts and the data extracted from the API in BigQuery.

B. Store the social media posts and the data extracted from the API in Cloud SQL.

C. Store the raw social media posts in Cloud Storage, and write the data extracted from the API into BigQuery.

D. Feed to social media posts into the API directly from the source, and write the extracted data from the API into BigQuery.

Answer: D

Explanation:

Question: 173

You store historic data in Cloud Storage. You need to perform analytics on the historic dat

a. You want to use a solution to detect invalid data entries and perform data transformations that will not require programming or knowledge of SQL.

What should you do?

A. Use Cloud Dataflow with Beam to detect errors and perform transformations.

B. Use Cloud Dataprep with recipes to detect errors and perform transformations.

C. Use Cloud Dataproc with a Hadoop job to detect errors and perform transformations.

D. Use federated tables in BigQuery with queries to detect errors and perform transformations.

Answer: B

Explanation:

Question: 174

Your company needs to upload their historic data to Cloud Storage. The security rules don't allow access from external IPs to their on-premises resources. After an initial upload, they will add new data from existing on-premises applications every day. What should they do?

A. Execute gsutil rsync from the on-premises servers.

B. Use Cloud Dataflow and write the data to Cloud Storage.

C. Write a job template in Cloud Dataproc to perform the data transfer.

D. Install an FTP server on a Compute Engine VM to receive the files and move them to Cloud Storage.

Answer: B

Explanation:

Question: 175

You have a query that filters a BigQuery table using a WHERE clause on timestamp and ID columns. By using bq query – -dry_run you learn that the query triggers a full scan of the table, even though the filter on timestamp and ID select a tiny fraction of the overall dat

a. You want to reduce the amount of data scanned by BigQuery with minimal changes to existing SQL queries. What should you do?

A. Create a separate table for each ID.

B. Use the LIMIT keyword to reduce the number of rows returned.

C. Recreate the table with a partitioning column and clustering column.

D. Use the bq query - -maximum_bytes_billed flag to restrict the number of bytes billed.

Answer: C

Explanation:

Question: 176

You have a requirement to insert minute-resolution data from 50,000 sensors into a BigQuery table.

You expect significant growth in data volume and need the data to be available within 1 minute of ingestion for real-time analysis of aggregated trends. What should you do?

A. Use bq load to load a batch of sensor data every 60 seconds.

B. Use a Cloud Dataflow pipeline to stream data into the BigQuery table.

C. Use the INSERT statement to insert a batch of data every 60 seconds.

D. Use the MERGE statement to apply updates in batch every 60 seconds.

Answer: B

Explanation:

Question: 177

You need to copy millions of sensitive patient records from a relational database to BigQuery. The total size of the database is 10 TB. You need to design a solution that is secure and time-efficient. What should you do?

A. Export the records from the database as an Avro file. Upload the file to GCS using gsutil, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.

B. Export the records from the database as an Avro file. Copy the file onto a Transfer Appliance and send it to Google, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.

C. Export the records from the database into a CSV file. Create a public URL for the CSV file, and then use Storage Transfer Service to move the file to Cloud Storage. Load the CSV file into BigQuery using the BigQuery web UI in the GCP Console.

D. Export the records from the database as an Avro file. Create a public URL for the Avro file, and then use Storage Transfer Service to move the file to Cloud Storage. Load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.

Answer: A

Explanation:

Question: 178

You need to create a near real-time inventory dashboard that reads the main inventory tables in your BigQuery data warehouse. Historical inventory data is stored as inventory balances by item and location. You have several thousand updates to inventory every hour. You want to maximize performance of the dashboard and ensure that the data is accurate. What should you do?

A. Leverage BigQuery UPDATE statements to update the inventory balances as they are changing.

B. Partition the inventory balance table by item to reduce the amount of data scanned with each inventory update.

C. Use the BigQuery streaming the stream changes into a daily inventory movement table. Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.

D. Use the BigQuery bulk loader to batch load inventory changes into a daily inventory movement table. Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.

Answer: A

Explanation:

Question: 179

You have a data stored in BigQuery. The data in the BigQuery dataset must be highly available. You need to define a storage, backup, and recovery strategy of this data that minimizes cost. How should you configure the BigQuery table?

A. Set the BigQuery dataset to be regional. In the event of an emergency, use a point-in-time snapshot to recover the data.

B. Set the BigQuery dataset to be regional. Create a scheduled query to make copies of the data to tables suffixed with the time of the backup. In the event of an emergency, use the backup copy of the table.

C. Set the BigQuery dataset to be multi-regional. In the event of an emergency, use a point-in-time snapshot to recover the data.

D. Set the BigQuery dataset to be multi-regional. Create a scheduled query to make copies of the data to tables suffixed with the time of the backup. In the event of an emergency, use the backup copy of the table.

Answer: B

Explanation:

Question: 180

You used Cloud Dataprep to create a recipe on a sample of data in a BigQuery table. You want to reuse this recipe on a daily upload of data with the same schema, after the load job with variable execution time completes. What

should you do?

A. Create a cron schedule in Cloud Dataprep.

B. Create an App Engine cron job to schedule the execution of the Cloud Dataprep job.

C. Export the recipe as a Cloud Dataprep template, and create a job in Cloud Scheduler.

D. Export the Cloud Dataprep job as a Cloud Dataflow template, and incorporate it into a Cloud Composer job.

Answer: D

Explanation:

Question: 181

You want to automate execution of a multi-step data pipeline running on Google Cloud. The pipeline includes Cloud Dataproc and Cloud Dataflow jobs that have multiple dependencies on each other.

You want to use managed services where possible, and the pipeline will run every day. Which tool should you use?

A. cron

B. Cloud Composer

C. Cloud Scheduler

D. Workflow Templates on Cloud Dataproc

Answer: B

Explanation:

Question: 182

You are managing a Cloud Dataproc cluster. You need to make a job run faster while minimizing costs, without losing

work in progress on your clusters. What should you do?

A. Increase the cluster size with more non-preemptible workers.

B. Increase the cluster size with preemptible worker nodes, and configure them to forcefully decommission.

C. Increase the cluster size with preemptible worker nodes, and use Cloud Stackdriver to trigger a script to preserve work.

D. Increase the cluster size with preemptible worker nodes, and configure them to use graceful decommissioning.

Answer: D

Explanation:

Reference: https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/flex

Question: 183

You work for a shipping company that uses handheld scanners to read shipping labels. Your company has strict data privacy standards that require scanners to only transmit recipients' personally identifiable information (PII) to analytics systems, which violates user privacy rules. You want to quickly build a scalable solution using cloud-native managed services to prevent exposure of PII to the analytics systems. What should you do?

A. Create an authorized view in BigQuery to restrict access to tables with sensitive data.

B. Install a third-party data validation tool on Compute Engine virtual machines to check the incoming data for sensitive information.

C. Use Stackdriver logging to analyze the data passed through the total pipeline to identify transactions that may contain sensitive information.

D. Build a Cloud Function that reads the topics and makes a call to the Cloud Data Loss Prevention API. Use the tagging and confidence levels to either pass or quarantine the data in a bucket for review.

Answer: D

Explanation:

Question: 184

You have developed three data processing jobs. One executes a Cloud Dataflow pipeline that transforms data uploaded to Cloud Storage and writes results to BigQuery. The second ingests data from on-premises servers and uploads it to Cloud Storage. The third is a Cloud Dataflow pipeline that gets information from third-party data providers and uploads the information to Cloud Storage. You need to be able to schedule and monitor the execution of these three workflows and manually execute them when needed. What should you do?

A. Create a Direct Acyclic Graph in Cloud Composer to schedule and monitor the jobs.

B. Use Stackdriver Monitoring and set up an alert with a Webhook notification to trigger the jobs.

C. Develop an App Engine application to schedule and request the status of the jobs using GCP API calls.

D. Set up cron jobs in a Compute Engine instance to schedule and monitor the pipelines using GCP API calls.

Answer: D

Explanation:

Question: 185

You have Cloud Functions written in Node.js that pull messages from Cloud Pub/Sub and send the data to BigQuery. You observe that the message processing rate on the Pub/Sub topic is orders of magnitude higher than anticipated, but there is no error logged in Stackdriver Log Viewer. What are the two most likely causes of this problem? Choose 2 answers.

A. Publisher throughput quota is too small.

B. Total outstanding messages exceed the 10-MB maximum.

C. Error handling in the subscriber code is not handling run-time errors properly.

D. The subscriber code cannot keep up with the messages.

E. The subscriber code does not acknowledge the messages that it pulls.

Answer: C,D

Explanation:

Question: 186

You are creating a new pipeline in Google Cloud to stream IoT data from Cloud Pub/Sub through Cloud Dataflow to BigQuery. While previewing the data, you notice that roughly 2% of the data appears to be corrupt. You need to modify the Cloud Dataflow pipeline to filter out this corrupt dat

a. What should you do?

A. Add a SideInput that returns a Boolean if the element is corrupt.

B. Add a ParDo transform in Cloud Dataflow to discard corrupt elements.

C. Add a Partition transform in Cloud Dataflow to separate valid data from corrupt data.

D. Add a GroupByKey transform in Cloud Dataflow to group all of the valid data together and discard the rest.

Answer: B

Explanation:

Question: 187

You have historical data covering the last three years in BigQuery and a data pipeline that delivers new data to BigQuery daily. You have noticed that when the Data Science team runs a query filtered on a date column and limited to 30–90 days of data, the query scans the entire table. You also noticed that your bill is increasing more quickly than you expected. You want to resolve the issue as cost-effectively as possible while maintaining the ability to conduct SQL queries. What should you do?

A. Re-create the tables using DDL. Partition the tables by a column containing a TIMESTAMP or DATE Type.

B. Recommend that the Data Science team export the table to a CSV file on Cloud Storage and use Cloud Datalab to explore the data by reading the files directly.

C. Modify your pipeline to maintain the last 30–90 days of data in one table and the longer history in a different table to minimize full table scans over the entire history.

D. Write an Apache Beam pipeline that creates a BigQuery table per day. Recommend that the Data Science team use wildcards on the table name suffixes to select the data they need.

Answer: C

Explanation:

Question: 188

You operate a logistics company, and you want to improve event delivery reliability for vehicle-based sensors. You operate small data centers around the world to capture these events, but leased lines that provide connectivity from your event collection infrastructure to your event processing infrastructure are unreliable, with unpredictable latency. You want to address this issue in the most cost-effective way. What should you do?

A. Deploy small Kafka clusters in your data centers to buffer events.

B. Have the data acquisition devices publish data to Cloud Pub/Sub.

C. Establish a Cloud Interconnect between all remote data centers and Google.

D. Write a Cloud Dataflow pipeline that aggregates all data in session windows.

Answer: B

Explanation:

Question: 189

You are a retailer that wants to integrate your online sales capabilities with different in-home assistants, such as Google Home. You need to interpret customer voice commands and issue an order to the backend systems. Which solutions should you choose?

A. Cloud Speech-to-Text API

B. Cloud Natural Language API

C. Dialogflow Enterprise Edition

D. Cloud AutoML Natural Language

Answer: C

Explanation:

Question: 190

Your company has a hybrid cloud initiative. You have a complex data pipeline that moves data between cloud provider services and leverages services from each of the cloud providers. Which cloud-native service should you use to orchestrate the entire pipeline?

A. Cloud Dataflow

B. Cloud Composer

C. Cloud Dataprep

D. Cloud Dataproc

Answer: D

Explanation:

Question: 191

You use a dataset in BigQuery for analysis. You want to provide third-party companies with access to the same dataset. You need to keep the costs of data sharing low and ensure that the data is current. Which solution should you choose?

A. Create an authorized view on the BigQuery table to control data access, and provide third-party companies with access to that view.

B. Use Cloud Scheduler to export the data on a regular basis to Cloud Storage, and provide third- party companies with access to the bucket.

C. Create a separate dataset in BigQuery that contains the relevant data to share, and provide third- party companies with access to the new dataset.

D. Create a Cloud Dataflow job that reads the data in frequent time intervals, and writes it to the relevant BigQuery dataset or Cloud Storage bucket for third-party companies to use.

Answer: B

Explanation:

Question: 192

A shipping company has live package-tracking data that is sent to an Apache Kafka stream in real time. This is then loaded into BigQuery. Analysts in your company want to query the tracking data in BigQuery to analyze geospatial trends in the lifecycle of a package. The table was originally created with ingest-date partitioning. Over time, the query processing time has increased. You need to implement a change that would improve query performance in BigQuery. What should you do?

A. Implement clustering in BigQuery on the ingest date column.

B. Implement clustering in BigQuery on the package-tracking ID column.

C. Tier older data onto Cloud Storage files, and leverage extended tables.

D. Re-create the table using data partitioning on the package delivery date.

Answer: A

Explanation:

Question: 193

You are designing a data processing pipeline. The pipeline must be able to scale automatically as load increases. Messages must be processed at least once, and must be ordered within windows of 1 hour. How should you design the solution?

A. Use Apache Kafka for message ingestion and use Cloud Dataproc for streaming analysis.

B. Use Apache Kafka for message ingestion and use Cloud Dataflow for streaming analysis.

C. Use Cloud Pub/Sub for message ingestion and Cloud Dataproc for streaming analysis.

D. Use Cloud Pub/Sub for message ingestion and Cloud Dataflow for streaming analysis.

Answer: D

Explanation:

Question: 194

You need to set access to BigQuery for different departments within your company. Your solution should comply with the following requirements:

Each department should have access only to their data.

Each department will have one or more leads who need to be able to create and update tables and provide them to their team.

Each department has data analysts who need to be able to query but not modify data.

How should you set access to the data in BigQuery?

A. Create a dataset for each department. Assign the department leads the role of OWNER, and assign the data analysts the role of WRITER on their dataset.

B. Create a dataset for each department. Assign the department leads the role of WRITER, and assign

the data analysts the role of READER on their dataset.

C. Create a table for each department. Assign the department leads the role of Owner, and assign the data analysts the role of Editor on the project the table is in.

D. Create a table for each department. Assign the department leads the role of Editor, and assign the data analysts the role of Viewer on the project the table is in.

Answer: D

Explanation:

Question: 195

You operate a database that stores stock trades and an application that retrieves average stock price for a given company over an adjustable window of time. The data is stored in Cloud Bigtable where the datetime of the stock trade is the beginning of the row key. Your application has thousands of concurrent users, and you notice that performance is starting to degrade as more stocks are added. What should you do to improve the performance of your application?

A. Change the row key syntax in your Cloud Bigtable table to begin with the stock symbol.

B. Change the row key syntax in your Cloud Bigtable table to begin with a random number per second.

C. Change the data pipeline to use BigQuery for storing stock trades, and update your application.

D. Use Cloud Dataflow to write summary of each day's stock trades to an Avro file on Cloud Storage. Update your application to read from Cloud Storage and Cloud Bigtable to compute the responses.

Answer: A

Explanation:

Question: 196

You are operating a Cloud Dataflow streaming pipeline. The pipeline aggregates events from a Cloud Pub/Sub subscription source, within a window, and sinks the resulting aggregation to a Cloud Storage bucket. The source has

consistent throughput. You want to monitor an alert on behavior of the pipeline with Cloud Stackdriver to ensure that it is processing dat

a. Which Stackdriver alerts should you create?

A. An alert based on a decrease of subscription/num_undelivered_messages for the source and a rate of change increase of instance/storage/used_bytes for the destination

B. An alert based on an increase of subscription/num_undelivered_messages for the source and a rate of change decrease of instance/storage/used_bytes for the destination

C. An alert based on a decrease of instance/storage/used_bytes for the source and a rate of change increase of subscription/num_undelivered_messages for the destination

D. An alert based on an increase of instance/storage/used_bytes for the source and a rate of change decrease of subscription/num_undelivered_messages for the destination

Answer: B

Explanation:

Question: 197

You currently have a single on-premises Kafka cluster in a data center in the us-east region that is responsible for ingesting messages from IoT devices globally. Because large parts of globe have poor internet connectivity, messages sometimes batch at the edge, come in all at once, and cause a spike in load on your Kafka cluster. This is becoming difficult to manage and prohibitively expensive. What is the Google-recommended cloud native architecture for this scenario?

A. Edge TPUs as sensor devices for storing and transmitting the messages.

B. Cloud Dataflow connected to the Kafka cluster to scale the processing of incoming messages.

C. An IoT gateway connected to Cloud Pub/Sub, with Cloud Dataflow to read and process the messages from Cloud Pub/Sub.

D. A Kafka cluster virtualized on Compute Engine in us-east with Cloud Load Balancing to connect to the devices

around the world.

Answer: C

Explanation:

Question: 198

You decided to use Cloud Datastore to ingest vehicle telemetry data in real time. You want to build a storage system that will account for the long-term data growth, while keeping the costs low. You also want to create snapshots of the data periodically, so that you can make a point-in-time (PIT) recovery, or clone a copy of the data for Cloud Datastore in a different environment. You want to archive these snapshots for a long time. Which two methods can accomplish this? Choose 2 answers.

A. Use managed export, and store the data in a Cloud Storage bucket using Nearline or Coldline class.

B. Use managed exportm, and then import to Cloud Datastore in a separate project under a unique namespace reserved for that export.

C. Use managed export, and then import the data into a BigQuery table created just for that export, and delete temporary export files.

D. Write an application that uses Cloud Datastore client libraries to read all the entities. Treat each entity as a BigQuery table row via BigQuery streaming insert. Assign an export timestamp for each export, and attach it as an extra column for each row. Make sure that the BigQuery table is partitioned using the export timestamp column.

E. Write an application that uses Cloud Datastore client libraries to read all the entities. Format the exported data into a JSON file. Apply compression before storing the data in Cloud Source Repositories.

Answer: C,E

Explanation:

Question: 199

You need to create a data pipeline that copies time-series transaction data so that it can be queried from within BigQuery by your data science team for analysis. Every hour, thousands of transactions are updated with a new status. The size of the intitial dataset is 1.5 PB, and it will grow by 3 TB per day. The data is heavily structured, and your data science team will build machine learning models based on this dat

a. You want to maximize performance and usability for your data science team. Which two strategies should you adopt? Choose 2 answers.

A. Denormalize the data as must as possible.

B. Preserve the structure of the data as much as possible.

C. Use BigQuery UPDATE to further reduce the size of the dataset.

D. Develop a data pipeline where status updates are appended to BigQuery instead of updated.

E. Copy a daily snapshot of transaction data to Cloud Storage and store it as an Avro file. Use BigQuery's support for external data sources to query.

Answer: A, E

Explanation:

Question: 200

You are designing a cloud-native historical data processing system to meet the following conditions:

The data being analyzed is in CSV, Avro, and PDF formats and will be accessed by multiple analysis tools including Cloud Dataproc, BigQuery, and Compute Engine.

A streaming data pipeline stores new data daily.

Peformance is not a factor in the solution.

The solution design should maximize availability.

How should you design data storage for this solution?

A. Create a Cloud Dataproc cluster with high availability. Store the data in HDFS, and peform analysis as needed.

B. Store the data in BigQuery. Access the data using the BigQuery Connector or Cloud Dataproc and Compute Engine.

C. Store the data in a regional Cloud Storage bucket. Aceess the bucket directly using Cloud Dataproc, BigQuery, and Compute Engine.

D. Store the data in a multi-regional Cloud Storage bucket. Access the data directly using Cloud Dataproc, BigQuery, and Compute Engine.

Answer: D

Explanation:

Question: 201

You have a petabyte of analytics data and need to design a storage and processing platform for it. You must be able to perform data warehouse-style analytics on the data in Google Cloud and expose the dataset as files for batch analysis tools in other cloud providers. What should you do?

A. Store and process the entire dataset in BigQuery.

B. Store and process the entire dataset in Cloud Bigtable.

C. Store the full dataset in BigQuery, and store a compressed copy of the data in a Cloud Storage bucket.

D. Store the warm data as files in Cloud Storage, and store the active data in BigQuery. Keep this ratio as 80% warm and 20% active.

Answer: C

Explanation:

Question: 202

You work for a manufacturing company that sources up to 750 different components, each from a different supplier. You've collected a labeled dataset that has on average 1000 examples for each unique component. Your team wants to implement an app to help warehouse workers recognize incoming components based on a photo of the component. You want to implement the first working version of this app (as Proof-Of-Concept) within a few working days. What should you do?

A. Use Cloud Vision AutoML with the existing dataset.

B. Use Cloud Vision AutoML, but reduce your dataset twice.

C. Use Cloud Vision API by providing custom labels as recognition hints.

D. Train your own image recognition model leveraging transfer learning techniques.

Answer: A

Explanation:

Question: 203

You are working on a niche product in the image recognition domain. Your team has developed a model that is dominated by custom C++ TensorFlow ops your team has implemented. These ops are used inside your main training loop and are performing bulky matrix multiplications. It currently takes up to several days to train a model. You want to decrease this time significantly and keep the cost low by using an accelerator on Google Cloud. What should you do?

A. Use Cloud TPUs without any additional adjustment to your code.

B. Use Cloud TPUs after implementing GPU kernel support for your customs ops.

C. Use Cloud GPUs after implementing GPU kernel support for your customs ops.

D. Stay on CPUs, and increase the size of the cluster you're training your model on.

Answer: B

Explanation:

Question: 204

You work on a regression problem in a natural language processing domain, and you have 100M labeled exmaples in your dataset. You have randomly shufled your data and split your dataset into train and test samples (in a 90/10 ratio). After you trained the neural network and evaluated your model on a test set, you discover that the root-mean-squared error (RMSE) of your model is twice as high on the train set as on the test set. How should you improve the performance of your model?

A. Increase the share of the test sample in the train-test split.

B. Try to collect more data and increase the size of your dataset.

C. Try out regularization techniques (e.g., dropout of batch normalization) to avoid overfitting.

D. Increase the complexity of your model by, e.g., introducing an additional layer or increase sizing the size of vocabularies or n-grams used.

Answer: D

Explanation:

Question: 205

You use BigQuery as your centralized analytics platform. New data is loaded every day, and an ETL

pipeline modifies the original data and prepares it for the final users. This ETL pipeline is regularly modified and can generate errors, but sometimes the errors are detected only after 2 weeks. You need to provide a method to recover from these errors, and your backups should be optimized for storage costs. How should you organize your data in BigQuery and store your backups?

A. Organize your data in a single table, export, and compress and store the BigQuery data in Cloud Storage.

B. Organize your data in separate tables for each month, and export, compress, and store the data in Cloud Storage.

C. Organize your data in separate tables for each month, and duplicate your data on a separate dataset in BigQuery.

D. Organize your data in separate tables for each month, and use snapshot decorators to restore the table to a time prior to the corruption.

Answer: D

Explanation:

Question: 206

The marketing team at your organization provides regular updates of a segment of your customer dataset. The marketing team has given you a CSV with 1 million records that must be updated in BigQuery. When you use the UPDATE statement in BigQuery, you receive a quotaExceeded error. What should you do?

A. Reduce the number of records updated each day to stay within the BigQuery UPDATE DML statement limit.

B. Increase the BigQuery UPDATE DML statement limit in the Quota management section of the Google Cloud Platform Console.

C. Split the source CSV file into smaller CSV files in Cloud Storage to reduce the number of BigQuery UPDATE DML statements per BigQuery job.

D. Import the new records from the CSV file into a new BigQuery table. Create a BigQuery job that merges the new records with the existing records and writes the results to a new BigQuery table.

Answer: D

Explanation:

Question: 207

As your organization expands its usage of GCP, many teams have started to create their own projects. Projects are further multiplied to accommodate different stages of deployments and target audiences. Each project requires unique

access control configurations. The central IT team needs to have access to all projects. Furthermore, data from Cloud Storage buckets and BigQuery datasets must be shared for use in other projects in an ad hoc way. You want to simplify access control management by minimizing the number of policies. Which two steps should you take? Choose 2 answers.

A. Use Cloud Deployment Manager to automate access provision.

B. Introduce resource hierarchy to leverage access control policy inheritance.

C. Create distinct groups for various teams, and specify groups in Cloud IAM policies.

D. Only use service accounts when sharing data for Cloud Storage buckets and BigQuery datasets.

E. For each Cloud Storage bucket or BigQuery dataset, decide which projects need access. Find all the active members who have access to these projects, and create a Cloud IAM policy to grant access to all these users.

Answer: A,C

Explanation:

Question: 208

Your United States-based company has created an application for assessing and responding to user actions. The primary table's data volume grows by 250,000 records per second. Many third parties
use your application's APIs to build the functionality into their own frontend applications. Your application's APIs should comply with the following requirements:

Single global endpoint

ANSI SQL support

Consistent access to the most up-to-date data

What should you do?

A. Implement BigQuery with no region selected for storage or processing.

B. Implement Cloud Spanner with the leader in North America and read-only replicas in Asia and Europe.

C. Implement Cloud SQL for PostgreSQL with the master in Norht America and read replicas in Asia and Europe.

D. Implement Cloud Bigtable with the primary cluster in North America and secondary clusters in Asia and Europe.

Answer: B

Explanation:

Question: 209

A data scientist has created a BigQuery ML model and asks you to create an ML pipeline to serve predictions. You have a REST API application with the requirement to serve predictions for an individual user ID with latency under 100 milliseconds. You use the following query to generate predictions: SELECT predicted_label, user_id FROM ML.PREDICT (MODEL 'dataset.model', table user_features). How should you create the ML pipeline?

A. Add a WHERE clause to the query, and grant the BigQuery Data Viewer role to the application service account.

B. Create an Authorized View with the provided query. Share the dataset that contains the view with the application service account.

C. Create a Cloud Dataflow pipeline using BigQueryIO to read results from the query. Grant the Dataflow Worker role to the application service account.

D. Create a Cloud Dataflow pipeline using BigQueryIO to read predictions for all users from the query. Write the results to Cloud Bigtable using BigtableIO. Grant the Bigtable Reader role to the application service account so that the application can read predictions for individual users from Cloud Bigtable.

Answer: D

Explanation:

Question: 210

You are building an application to share financial market data with consumers, who will receive data feeds. Data is

collected from the markets in real time. Consumers will receive the data in the following ways:

Real-time event stream

ANSI SQL access to real-time stream and historical data

Batch historical exports

Which solution should you use?

A. Cloud Dataflow, Cloud SQL, Cloud Spanner

B. Cloud Pub/Sub, Cloud Storage, BigQuery

C. Cloud Dataproc, Cloud Dataflow, BigQuery

D. Cloud Pub/Sub, Cloud Dataproc, Cloud SQL

Answer: A

Explanation:

Question: 211

You are building a new application that you need to collect data from in a scalable way. Data arrives continuously from the application throughout the day, and you expect to generate approximately 150 GB of JSON data per day by the end of the year. Your requirements are:

Decoupling producer from consumer

Space and cost-efficient storage of the raw ingested data, which is to be stored indefinitely

Near real-time SQL query

Maintain at least 2 years of historical data, which will be queried with SQ

Which pipeline should you use to meet these requirements?

A. Create an application that provides an API. Write a tool to poll the API and write data to Cloud Storage as gzipped JSON files.

B. Create an application that writes to a Cloud SQL database to store the data. Set up periodic exports of the database to write to Cloud Storage and load into BigQuery.

C. Create an application that publishes events to Cloud Pub/Sub, and create Spark jobs on Cloud Dataproc to convert the JSON data to Avro format, stored on HDFS on Persistent Disk.

D. Create an application that publishes events to Cloud Pub/Sub, and create a Cloud Dataflow pipeline that transforms the JSON event payloads to Avro, writing the data to Cloud Storage and BigQuery.

Answer: A

Explanation:

Question: 212

You are running a pipeline in Cloud Dataflow that receives messages from a Cloud Pub/Sub topic and writes the results to a BigQuery dataset in the EU. Currently, your pipeline is located in europe-west4 and has a maximum of 3 workers, instance type n1-standard-1. You notice that during peak periods, your pipeline is struggling to process records in a timely fashion, when all 3 workers are at maximum CPU utilization. Which two actions can you take to increase performance of your pipeline? (Choose two.)

A. Increase the number of max workers

B. Use a larger instance type for your Cloud Dataflow workers

C. Change the zone of your Cloud Dataflow pipeline to run in us-central1

D. Create a temporary table in Cloud Bigtable that will act as a buffer for new data. Create a new step in your pipeline to write to this table first, and then create a new pipeline to write from Cloud Bigtable to BigQuery

E. Create a temporary table in Cloud Spanner that will act as a buffer for new data. Create a new step in your pipeline to write to this table first, and then create a new pipeline to write from Cloud Spanner to BigQuery

Answer: A, B

Explanation:

Question: 213

You have a data pipeline with a Cloud Dataflow job that aggregates and writes time series metrics to Cloud Bigtable. This data feeds a dashboard used by thousands of users across the organization. You need to support additional concurrent users and reduce the amount of time required to write the dat

a. Which two actions should you take? (Choose two.)

A. Configure your Cloud Dataflow pipeline to use local execution

B. Increase the maximum number of Cloud Dataflow workers by setting maxNumWorkers in PipelineOptions

C. Increase the number of nodes in the Cloud Bigtable cluster

D. Modify your Cloud Dataflow pipeline to use the Flatten transform before writing to Cloud Bigtable

E. Modify your Cloud Dataflow pipeline to use the CoGroupByKey transform before writing to Cloud Bigtable

Answer: B, C

Explanation:

Reference:

Question: 214

You have several Spark jobs that run on a Cloud Dataproc cluster on a schedule. Some of the jobs run in sequence, and some of the jobs run concurrently. You need to automate this process. What should you do?

A. Create a Cloud Dataproc Workflow Template

B. Create an initialization action to execute the jobs

C. Create a Directed Acyclic Graph in Cloud Composer

D. Create a Bash script that uses the Cloud SDK to create a cluster, execute jobs, and then tear down the cluster

Answer: C

Explanation:

Reference:

Question: 215

You are building a new data pipeline to share data between two different types of applications: jobs generators and job runners. Your solution must scale to accommodate increases in usage and must accommodate the addition of new applications without negatively affecting the performance of existing ones. What should you do?

A. Create an API using App Engine to receive and send messages to the applications

B. Use a Cloud Pub/Sub topic to publish jobs, and use subscriptions to execute them

C. Create a table on Cloud SQL, and insert and delete rows with the job information

D. Create a table on Cloud Spanner, and insert and delete rows with the job information

Answer: A

Explanation:

Reference:

Question: 216

You need to create a new transaction table in Cloud Spanner that stores product sales dat

a. You are deciding what to use as a primary key. From a performance perspective, which strategy should you choose?

A. The current epoch time

B. A concatenation of the product name and the current epoch time

C. A random universally unique identifier number (version 4 UUID)

D. The original order identification number from the sales system, which is a monotonically increasing integer

Answer: C

Explanation:

Reference:

Question: 217

Data Analysts in your company have the Cloud IAM Owner role assigned to them in their projects to allow them to work with multiple GCP products in their projects. Your organization requires that all BigQuery data access logs be retained for 6 months. You need to ensure that only audit personnel in your company can access the data access logs for all projects. What should you do?

A. Enable data access logs in each Data Analyst's project. Restrict access to Stackdriver Logging via Cloud IAM roles.

B. Export the data access logs via a project-level export sink to a Cloud Storage bucket in the Data Analysts' projects. Restrict access to the Cloud Storage bucket.

C. Export the data access logs via a project-level export sink to a Cloud Storage bucket in a newly created projects for audit logs. Restrict access to the project with the exported logs.

D. Export the data access logs via an aggregated export sink to a Cloud Storage bucket in a newly created project for audit logs. Restrict access to the project that contains the exported logs.

Answer: D

Explanation:

Question: 218

Each analytics team in your organization is running BigQuery jobs in their own projects. You want to enable each team to monitor slot usage within their projects. What should you do?

A. Create a Stackdriver Monitoring dashboard based on the BigQuery metric query/scanned_bytes

B. Create a Stackdriver Monitoring dashboard based on the BigQuery metric slots/allocated_for_project

C. Create a log export for each project, capture the BigQuery job execution logs, create a custom metric based on the totalSlotMs, and create a Stackdriver Monitoring dashboard based on the custom metric

D. Create an aggregated log export at the organization level, capture the BigQuery job execution logs, create a custom metric based on the totalSlotMs, and create a Stackdriver Monitoring dashboard based on the custom metric

Answer: D

Explanation:

Question: 219

You are operating a streaming Cloud Dataflow pipeline. Your engineers have a new version of the pipeline with a different windowing algorithm and triggering strategy. You want to update the running pipeline with the new version. You want to ensure that no data is lost during the update. What should you do?

A. Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to the existing job name

B. Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to a new unique job name

C. Stop the Cloud Dataflow pipeline with the Cancel option. Create a new Cloud Dataflow job with the updated code

D. Stop the Cloud Dataflow pipeline with the Drain option. Create a new Cloud Dataflow job with the updated code

Answer: A

Explanation:

Reference:

Question: 220

You need to move 2 PB of historical data from an on-premises storage appliance to Cloud Storage within six months, and your outbound network capacity is constrained to 20 Mb/sec. How should you migrate this data to Cloud Storage?

A. Use Transfer Appliance to copy the data to Cloud Storage

B. Use gsutil cp –J to compress the content being uploaded to Cloud Storage

C. Create a private URL for the historical data, and then use Storage Transfer Service to copy the data to Cloud Storage

D. Use trickle or ionice along with gsutil cp to limit the amount of bandwidth gsutil utilizes to less than 20 Mb/sec so it does not interfere with the production traffic

Answer: A

Explanation:

Question: 221

You receive data files in CSV format monthly from a third party. You need to cleanse this data, but every third month the schema of the files changes. Your requirements for implementing these transformations include:

Executing the transformations on a schedule

Enabling non-developer analysts to modify transformations

Providing a graphical tool for designing transformations

What should you do?

A. Use Cloud Dataprep to build and maintain the transformation recipes, and execute them on a scheduled basis

B. Load each month's CSV data into BigQuery, and write a SQL query to transform the data to a standard schema. Merge the transformed tables together with a SQL query

C. Help the analysts write a Cloud Dataflow pipeline in Python to perform the transformation. The Python code should be stored in a revision control system and modified as the incoming data's schema changes

D. Use Apache Spark on Cloud Dataproc to infer the schema of the CSV file before creating a Dataframe. Then implement the transformations in Spark SQL before writing the data out to Cloud Storage and loading into BigQuery

Answer: A

Explanation:

you can use dataprep for continuously changing target schema

In general, a target consists of the set of information required to define the expected data in a dataset. Often referred to as a "schema," this target schema information can include:

Names of columns

Order of columns

Column data types

Data type format

Example rows of data

A dataset associated with a target is expected to conform to the requirements of the schema. Where there are differences between target schema and dataset schema, a validation indicator (or schema tag) is displayed.

https://cloud.google.com/dataprep/docs/html/Overview-of-RapidTarget_136155049

Question: 222

You want to migrate an on-premises Hadoop system to Cloud Dataproc. Hive is the primary tool in use, and the data format is Optimized Row Columnar (ORC). All ORC files have been successfully copied to a Cloud Storage bucket. You need to replicate some data to the cluster's local Hadoop Distributed File System (HDFS) to maximize performance. What are two ways to start using Hive in Cloud Dataproc? (Choose two.)

A. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to HDFS. Mount the Hive tables locally.

B. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to any node of the Dataproc cluster. Mount the Hive tables locally.

C. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to the master node of the Dataproc cluster. Then run the Hadoop utility to copy them do HDFS. Mount the Hive tables from HDFS.

D. Leverage Cloud Storage connector for Hadoop to mount the ORC files as external Hive tables. Replicate external Hive tables to the native ones.

E. Load the ORC files into BigQuery. Leverage BigQuery connector for Hadoop to mount the BigQuery tables as external Hive tables. Replicate external Hive tables to the native ones.

Answer: B,C

Explanation:

Question: 223

You are implementing several batch jobs that must be executed on a schedule. These jobs have many interdependent steps that must be executed in a specific order. Portions of the jobs involve executing shell scripts, running Hadoop jobs, and running queries in BigQuery. The jobs are expected to run for many minutes up to several hours. If the steps fail, they must be retried a fixed number of times. Which service should you use to manage the execution of these jobs?

A. Cloud Scheduler

B. Cloud Dataflow

C. Cloud Functions

D. Cloud Composer

Answer: A

Explanation:

Question: 224

You work for a shipping company that has distribution centers where packages move on delivery lines to route them properly. The company wants to add cameras to the delivery lines to detect and track any visual damage to the packages in transit. You need to create a way to automate the detection of damaged packages and flag them for human review in real time while the packages are in transit. Which solution should you choose?

A. Use BigQuery machine learning to be able to train the model at scale, so you can analyze the packages in batches.

B. Train an AutoML model on your corpus of images, and build an API around that model to integrate with the package tracking applications.

C. Use the Cloud Vision API to detect for damage, and raise an alert through Cloud Functions.
Integrate the package tracking applications with this function.

D. Use TensorFlow to create a model that is trained on your corpus of images. Create a Python notebook in Cloud Datalab that uses this model so you can analyze for damaged packages.

Answer: A

Explanation:

Question: 225

You are migrating your data warehouse to BigQuery. You have migrated all of your data into tables in a dataset.

Multiple users from your organization will be using the dat

a. They should only see certain tables based on their team membership. How should you set user permissions?

A. Assign the users/groups data viewer access at the table level for each table

B. Create SQL views for each team in the same dataset in which the data resides, and assign the users/groups data viewer access to the SQL views

C. Create authorized views for each team in the same dataset in which the data resides, and assign the users/groups data viewer access to the authorized views

D. Create authorized views for each team in datasets created for each team. Assign the authorized views data viewer access to the dataset in which the data resides. Assign the users/groups data viewer access to the datasets in which the authorized views reside

Answer: A

Explanation:

Question: 226

You want to build a managed Hadoop system as your data lake. The data transformation process is composed of a series of Hadoop jobs executed in sequence. To accomplish the design of separating storage from compute, you decided to use the Cloud Storage connector to store all input data, output data, and intermediary dat

a. However, you noticed that one Hadoop job runs very slowly with Cloud Dataproc, when compared with the on-premises bare-metal Hadoop environment (8-core nodes with 100-GB RAM). Analysis shows that this particular Hadoop job is disk I/O intensive. You want to resolve the issue. What should you do?

A. Allocate sufficient memory to the Hadoop cluster, so that the intermediary data of that particular Hadoop job can be held in memory

B. Allocate sufficient persistent disk space to the Hadoop cluster, and store the intermediate data of that particular Hadoop job on native HDFS

C. Allocate more CPU cores of the virtual machine instances of the Hadoop cluster so that the networking bandwidth for each instance can scale up

D. Allocate additional network interface card (NIC), and configure link aggregation in the operating system to use the combined throughput when working with Cloud Storage

Answer: A

Explanation:

Question: 227

You work for an advertising company, and you've developed a Spark ML model to predict clickthrough rates at advertisement blocks. You've been developing everything at your on-premises data center, and now your company is migrating to Google Cloud. Your data center will be migrated to BigQuery. You periodically retrain your Spark ML models, so you need to migrate existing training pipelines to Google Cloud. What should you do?

A. Use Cloud ML Engine for training existing Spark ML models

B. Rewrite your models on TensorFlow, and start using Cloud ML Engine

C. Use Cloud Dataproc for training existing Spark ML models, but start reading data directly from BigQuery

D. Spin up a Spark cluster on Compute Engine, and train Spark ML models on the data exported from BigQuery

Answer: C

Explanation:

https://cloud.google.com/dataproc/docs/tutorials/bigquery-sparkml

Question: 228

You work for a global shipping company. You want to train a model on 40 TB of data to predict which ships in each geographic region are likely to cause delivery delays on any given day. The model will be based on multiple attributes collected from multiple sources. Telemetry data, including location in GeoJSON format, will be pulled from each ship and loaded every hour. You want to have a dashboard that shows how many and which ships are likely to cause delays within a region. You want to use a storage solution that has native functionality for prediction and geospatial processing. Which storage solution should you use?

A. BigQuery

B. Cloud Bigtable

C. Cloud Datastore

D. Cloud SQL for PostgreSQL

Answer: A

Explanation:

Question: 229

You operate an IoT pipeline built around Apache Kafka that normally receives around 5000 messages per second. You want to use Google Cloud Platform to create an alert as soon as the moving average over 1 hour drops below 4000 messages per second. What should you do?

A. Consume the stream of data in Cloud Dataflow using Kafka IO. Set a sliding time window of 1 hour every 5 minutes. Compute the average when the window closes, and send an alert if the average is less than 4000 messages.

B. Consume the stream of data in Cloud Dataflow using Kafka IO. Set a fixed time window of 1 hour. Compute the average when the window closes, and send an alert if the average is less than 4000 messages.

C. Use Kafka Connect to link your Kafka message queue to Cloud Pub/Sub. Use a Cloud Dataflow template to write your messages from Cloud Pub/Sub to Cloud Bigtable. Use Cloud Scheduler to run a script every hour that counts the number of rows created in Cloud Bigtable in the last hour. If that number falls below 4000, send an alert.

D. Use Kafka Connect to link your Kafka message queue to Cloud Pub/Sub. Use a Cloud Dataflow template to write your messages from Cloud Pub/Sub to BigQuery. Use Cloud Scheduler to run a script every five minutes that counts the number of rows created in BigQuery in the last hour. If that number falls below 4000, send an alert.

Answer: C

Explanation:

Question: 230

You plan to deploy Cloud SQL using MySQL. You need to ensure high availability in the event of a zone failure. What should you do?

A. Create a Cloud SQL instance in one zone, and create a failover replica in another zone within the same region.

B. Create a Cloud SQL instance in one zone, and create a read replica in another zone within the same region.

C. Create a Cloud SQL instance in one zone, and configure an external read replica in a zone in a different region.

D. Create a Cloud SQL instance in a region, and configure automatic backup to a Cloud Storage bucket in the same region.

Answer: C

Explanation:

Question: 231

Your company is selecting a system to centralize data ingestion and delivery. You are considering messaging and data integration systems to address the requirements. The key requirements are:

The ability to seek to a particular offset in a topic, possibly back to the start of all data ever captured

Support for publish/subscribe semantics on hundreds of topics

Retain per-key ordering

Which system should you choose?

A. Apache Kafka

B. Cloud Storage

C. Cloud Pub/Sub

D. Firebase Cloud Messaging

Answer: A

Explanation:

Question: 232

You are planning to migrate your current on-premises Apache Hadoop deployment to the cloud. You need to ensure that the deployment is as fault-tolerant and cost-effective as possible for long- running batch jobs. You want to use a managed service. What should you do?

A. Deploy a Cloud Dataproc cluster. Use a standard persistent disk and 50% preemptible workers. Store data in Cloud Storage, and change references in scripts from hdfs:// to gs://

B. Deploy a Cloud Dataproc cluster. Use an SSD persistent disk and 50% preemptible workers. Store data in Cloud Storage, and change references in scripts from hdfs:// to gs://

C. Install Hadoop and Spark on a 10-node Compute Engine instance group with standard instances. Install the Cloud Storage connector, and store the data in Cloud Storage. Change references in scripts from hdfs:// to gs://

D. Install Hadoop and Spark on a 10-node Compute Engine instance group with preemptible instances. Store data in HDFS. Change references in scripts from hdfs:// to gs://

Answer: A

Explanation:

Question: 233

Your team is working on a binary classification problem. You have trained a support vector machine (SVM) classifier with default parameters, and received an area under the Curve (AUC) of 0.87 on the validation set. You want to increase the AUC of the model. What should you do?

A. Perform hyperparameter tuning

B. Train a classifier with deep neural networks, because neural networks would always beat SVMs

C. Deploy the model and measure the real-world AUC; it's always higher because of generalization

D. Scale predictions you get out of the model (tune a scaling factor as a hyperparameter) in order to get the highest AUC

Answer: A

Explanation:

https://towardsdatascience.com/understanding-hyperparameters-and-its-optimisation-techniques- f0debba07568

Question: 234

You need to deploy additional dependencies to all of a Cloud Dataproc cluster at startup using an existing initialization action. Company security policies require that Cloud Dataproc nodes do not have access to the Internet so public initialization actions cannot fetch resources. What should you do?

A. Deploy the Cloud SQL Proxy on the Cloud Dataproc master

B. Use an SSH tunnel to give the Cloud Dataproc cluster access to the Internet

C. Copy all dependencies to a Cloud Storage bucket within your VPC security perimeter

D. Use Resource Manager to add the service account used by the Cloud Dataproc cluster to the Network User role

Answer: C

Explanation:

Question: 235

You need to choose a database for a new project that has the following requirements:

Fully managed

Able to automatically scale up

Transactionally consistent

Able to scale up to 6 TB

Able to be queried using SQL

Which database do you choose?

A. Cloud SQL

B. Cloud Bigtable

C. Cloud Spanner

D. Cloud Datastore

Answer: C

Explanation:

Question: 236

You work for a mid-sized enterprise that needs to move its operational system transaction data from an on-premises database to GCP. The database is about 20 TB in size. Which database should you choose?

A. Cloud SQL

B. Cloud Bigtable

C. Cloud Spanner

D. Cloud Datastore

Answer: A

Explanation:

Question: 237

You need to choose a database to store time series CPU and memory usage for millions of computers. You need to store this data in one-second interval samples. Analysts will be performing real-time, ad hoc analytics against the database. You want to avoid being charged for every query executed and ensure that the schema design will allow for future growth of the dataset. Which database and data model should you choose?

A. Create a table in BigQuery, and append the new samples for CPU and memory to the table

B. Create a wide table in BigQuery, create a column for the sample value at each second, and update the row with the interval for each second

C. Create a narrow table in Cloud Bigtable with a row key that combines the Computer Engine computer identifier with the sample time at each second

D. Create a wide table in Cloud Bigtable with a row key that combines the computer identifier with the sample time at each minute, and combine the values for each second as column data.

Answer: C

Explanation:

A tall and narrow table has a small number of events per row, which could be just one event, whereas a short and wide table has a large number of events per row. As explained in a moment, tall and narrow tables are best suited for time-series data. For time series, you should generally use tall and narrow tables. This is for two reasons: Storing one event per row makes it easier to run queries against your data. Storing many events per row makes it more likely that the total row size will exceed the recommended maximum (see Rows can be big but are not infinite).

https://cloud.google.com/bigtable/docs/schema-design-time-series#patterns_for_row_key_design

Question: 238

You want to archive data in Cloud Storage. Because some data is very sensitive, you want to use the "Trust No One" (TNO) approach to encrypt your data to prevent the cloud provider staff from decrypting your dat

a. What should you do?

A. Use gcloud kms keys create to create a symmetric key. Then use gcloud kms encrypt to encrypt each archival file with the key and unique additional authenticated data (AAD). Use gsutil cp to upload each encrypted file to the Cloud Storage

bucket, and keep the AAD outside of Google Cloud.

B. Use gcloud kms keys create to create a symmetric key. Then use gcloud kms encrypt to encrypt each archival file with the key. Use gsutil cp to upload each encrypted file to the Cloud Storage bucket. Manually destroy the key previously used for encryption, and rotate the key once and rotate the key once.

C. Specify customer-supplied encryption key (CSEK) in the .boto configuration file. Use gsutil cp to upload each archival file to the Cloud Storage bucket. Save the CSEK in Cloud Memorystore as permanent storage of the secret.

D. Specify customer-supplied encryption key (CSEK) in the .boto configuration file. Use gsutil cp to upload each archival file to the Cloud Storage bucket. Save the CSEK in a different project that only the security team can access.

Answer: B

Explanation:

Question: 239

You have data pipelines running on BigQuery, Cloud Dataflow, and Cloud Dataproc. You need to perform health checks and monitor their behavior, and then notify the team managing the pipelines if they fail. You also need to be able to work across multiple projects. Your preference is to use managed products of features of the platform. What should you do?

A. Export the information to Cloud Stackdriver, and set up an Alerting policy

B. Run a Virtual Machine in Compute Engine with Airflow, and export the information to Stackdriver

C. Export the logs to BigQuery, and set up App Engine to read that information and send emails if you find a failure in the logs

D. Develop an App Engine application to consume logs using GCP API calls, and send emails if you find a failure in the logs

Answer: B

Explanation:

Question: 240

You work for a large financial institution that is planning to use Dialogflow to create a chatbot for the company's mobile app You have reviewed old chat logs and lagged each conversation for intent based on each customer's stated intention for contacting customer service About 70% of customer requests are simple requests that are solved within 10 intents The remaining 30% of inquiries require much longer, more complicated requests Which intents should you automate first?

A. Automate the 10 intents that cover 70% of the requests so that live agents can handle more complicated requests

B. Automate the more complicated requests first because those require more of the agents' time

C. Automate a blend of the shortest and longest intents to be representative of all intents

D. Automate intents in places where common words such as "payment" appear only once so the software isn't confused

Answer: A

Explanation:

Question: 241

You want to rebuild your batch pipeline for structured data on Google Cloud You are using PySpark to conduct data transformations at scale, but your pipelines are taking over twelve hours to run. To expedite development and pipeline run time, you want to use a serverless tool and SQL syntax You have already moved your raw data into Cloud Storage How should you build the pipeline on Google Cloud while meeting speed and processing requirements?

A. Convert your PySpark commands into SparkSQL queries to transform the data; and then run your pipeline on Dataproc to write the data into BigQuery

B. Ingest your data into Cloud SQL, convert your PySpark commands into SparkSQL queries to transform the data, and then use federated queries from BigQuery for machine learning.

C. Ingest your data into BigQuery from Cloud Storage, convert your PySpark commands into BigQuery SQL queries to transform the data, and then write the transformations to a new table

D. Use Apache Beam Python SDK to build the transformation pipelines, and write the data into BigQuery

Answer: A

Explanation:

Question: 242

You are building a teal-lime prediction engine that streams files, which may contain Pll (personal identifiable information) data, into Cloud Storage and eventually into BigQuery You want to ensure that the sensitive data is masked but still maintains referential Integrity, because names and emails are often used as join keys How should you use the Cloud Data Loss Prevention API (DLP API) to ensure that the Pll data is not accessible by unauthorized individuals?

A. Create a pseudonym by replacing the Pll data with cryptogenic tokens, and store the non- tokenized data in a locked-down button.

B. Redact all Pll data, and store a version of the unredacted data in a locked-down bucket

C. Scan every table in BigQuery, and mask the data it finds that has Pll

D. Create a pseudonym by replacing Pll data with a cryptographic format-preserving token

Answer: A

Explanation:

Question: 243

Your company is implementing a data warehouse using BigQuery, and you have been tasked with designing the data model You move your on-premises sales data warehouse with a star data schema to BigQuery but notice performance issues when querying the data of the past 30 days Based on Google's recommended practices, what should you do to speed up the query without increasing storage costs?

A. Denormalize the data

B. Shard the data by customer ID

C. Materialize the dimensional data in views

D. Partition the data by transaction date

Answer: C

Explanation:

Question: 244

You are using Cloud Bigtable to persist and serve stock market data for each of the major indices. To serve the trading application, you need to access only the most recent stock prices that are streaming in How should you design your row key and tables to ensure that you can access the data with the most simple query?

A. Create one unique table for all of the indices, and then use the index and timestamp as the row key design

B. Create one unique table for all of the indices, and then use a reverse timestamp as the row key design.

C. For each index, have a separate table and use a timestamp as the row key design

D. For each index, have a separate table and use a reverse timestamp as the row key design

Answer: A

Explanation:

Question: 245

You are testing a Dataflow pipeline to ingest and transform text files. The files are compressed gzip, errors are written to a dead-letter queue, and you are using SideInputs to join data You noticed that the pipeline is taking longer to complete than expected, what should you do to expedite the Dataflow job?

A. Switch to compressed Avro files

B. Reduce the batch size

C. Retry records that throw an error

D. Use CoGroupByKey instead of the SideInput

Answer: B

Explanation:

Question: 246

You are building a report-only data warehouse where the data is streamed into BigQuery via the streaming API Following Google's best practices, you have both a staging and a production table for the data How should you design your data loading to ensure that there is only one master dataset without affecting performance on either the ingestion or reporting pieces?

A. Have a staging table that is an append-only model, and then update the production table every three hours with the changes written to staging

B. Have a staging table that is an append-only model, and then update the production table every ninety minutes with the changes written to staging

C. Have a staging table that moves the staged data over to the production table and deletes the contents of the staging table every three hours

D. Have a staging table that moves the staged data over to the production table and deletes the contents of the staging table every thirty minutes

Answer: D

Explanation:

Question: 247

You are migrating your data warehouse to Google Cloud and decommissioning your on-premises data center Because this is a priority for your company, you know that bandwidth will be made available for the initial data load to the cloud. The files being transferred are not large in number, but each file is 90 GB Additionally, you want your transactional systems to continually update the warehouse on Google Cloud in real time What tools should you use to migrate the data and ensure that it continues to write to your warehouse?

A. Storage Transfer Service for the migration, Pub/Sub and Cloud Data Fusion for the real-time updates

B. BigQuery Data Transfer Service for the migration, Pub/Sub and Dataproc for the real-time updates

C. gsutil for the migration; Pub/Sub and Dataflow for the real-time updates

D. gsutil for both the migration and the real-time updates

Answer: A

Explanation:

Question: 248

You need (o give new website users a globally unique identifier (GUID) using a service that takes in data points and returns a GUID This data is sourced from both internal and external systems via HTTP calls that you will make via microservices within your pipeline There will be tens of thousands of messages per second and that can be multithreaded, and you worry about the backpressure on the system How should you design your pipeline to minimize that backpressure?

A. Call out to the service via HTTP

B. Create the pipeline statically in the class definition

C. Create a new object in the startBundle method of DoFn

D. Batch the job into ten-second increments

Answer: A

Explanation:

Question: 249

You are migrating an application that tracks library books and information about each book, such as author or year published, from an on-premises data warehouse to BigQuery In your current relational database, the author information is kept in a separate table and joined to the book information on a common key Based on Google's recommended practice for schema design, how would you structure the data to ensure optimal speed of queries about the author of each book that has been borrowed?

A. Keep the schema the same, maintain the different tables for the book and each of the attributes, and query as you are doing today

B. Create a table that is wide and includes a column for each attribute, including the author's first name, last name, date of birth, etc

C. Create a table that includes information about the books and authors, but nest the author fields inside the author column

D. Keep the schema the same, create a view that joins all of the tables, and always query the view

Answer: C

Explanation:

Question: 250

You have uploaded 5 years of log data to Cloud Storage A user reported that some data points in the log data are outside of their expected ranges, which indicates errors You need to address this issue and be able to run the process again in the future while keeping the original data for compliance reasons. What should you do?

A. Import the data from Cloud Storage into BigQuery Create a new BigQuery table, and skip the rows with errors.

B. Create a Compute Engine instance and create a new copy of the data in Cloud Storage Skip the rows with errors

C. Create a Cloud Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to a new dataset in Cloud Storage

D. Create a Cloud Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to the same dataset in Cloud Storage

Answer: D

Explanation:

Question: 251

An aerospace company uses a proprietary data format to store its night dat

a. You need to connect this new data source to BigQuery and stream the data into BigQuery. You want to efficiency import the data into BigQuery where consuming as few resources as possible. What should you do?

A. Use a standard Dataflow pipeline to store the raw data m BigQuery and then transform the format later when the data is used

B. Write a she script that triggers a Cloud Function that performs periodic ETL batch jobs on the new data source

C. Use Apache Hive to write a Dataproc job that streams the data into BigQuery in CSV format

D. Use an Apache Beam custom connector to write a Dataflow pipeline that streams the data into BigQuery in Avro format

Answer: D

Explanation:

Question: 252

An aerospace company uses a proprietary data format to store its night dat

a. You need to connect this new data source to BigQuery and stream the data into BigQuery. You want to efficiency import the data into BigQuery where consuming as few resources as possible. What should you do?

A. Use a standard Dataflow pipeline to store the raw data in BigQuery and then transform the format later when the data is used.

B. Write a shell script that triggers a Cloud Function that performs periodic ETL batch jobs on the new data source

C. Use Apache Hive to write a Dataproc job that streams the data into BigQuery in CSV format

D. Use an Apache Beam custom connector to write a Dataflow pipeline that streams the data into BigQuery in Avro format

Answer: D

Explanation:

Question: 253

You are using BigQuery and Data Studio to design a customer-facing dashboard that displays large quantities of aggregated dat

a. You expect a high volume of concurrent users. You need to optimize tie dashboard to provide quick visualizations with minimal latency. What should you do?

A. Use BigQuery BI Engine with materialized views

B. Use BigQuery BI Engine with streaming data.

C. Use BigQuery Bl Engine with authorized views

D. Use BigQuery Bl Engine with logical reviews

Answer: B

Explanation:

Question: 254

You need ads data to serve Al models and historical data tor analytics longtail and outlier data points need to be identified You want to cleanse the data n near-reel time before running it through Al models What should you do?

A. Use BigQuery to ingest prepare and then analyze the data and then run queries to create views

B. Use Cloud Storage as a data warehouse shell scripts tor processing, and BigQuery to create views tor desired datasets

C. Use Dataflow to identity longtail and outber data points programmatically with BigQuery as a sink

D. Use Cloud Composer to identify longtail and outlier data points, and then output a usable dataset to BigQuery

Answer: A

Explanation:

Question: 255

The Development and External teams nave the project viewer Identity and Access Management (1AM) role m a folder named Visualization. You want the Development Team to be able to read data from both Cloud Storage and BigQuery, but the External Team should only be able to read data from BigQuery. What should you do?

A. Remove Cloud Storage IAM permissions to the External Team on the acme-raw-data project

B. Create Virtual Private Cloud (VPC) firewall rules on the acme-raw-data protect that deny all Ingress traffic from the External Team CIDR range

C. Create a VPC Service Controls perimeter containing both protects and BigQuery as a restricted API Add the External Team users to the perimeter s Access Level

D. Create a VPC Service Controls perimeter containing both protects and Cloud Storage as a restricted API. Add the Development Team users to the perimeter's Access Level

Answer: C

Explanation:

Question: 256

A TensorFlow machine learning model on Compute Engine virtual machines (n2-standard -32) takes two days to complete framing. The model has custom TensorFlow operations that must run partially on a CPU You want to reduce the training time in a cost-effective manner. What should you do?

A. Change the VM type to n2-highmem-32

B. Change the VM type to e2 standard-32

C. Train the model using a VM with a GPU hardware accelerator

D. Train the model using a VM with a TPU hardware accelerator

Answer: C

Explanation:

Question: 257

An online brokerage company requires a high volume trade processing architecture. You need to create a secure queuing system that triggers jobs. The jobs will run in Google Cloud and cat the company's Python API to execute trades. You need to efficiently implement a solution. What should you do?

A. Use Cloud Composer to subscribe to a Pub/Sub tope and can the Python API.

B. Use a Pub/Sub push subscription to trigger a Cloud Function to pass the data to tie Python API.

C. Write an application that makes a queue in a NoSQL database

D. Write an application hosted on a Compute Engine instance that makes a push subscription to the Pub/Sub topic

Answer: C

Explanation:

Question: 258

You are designing a pipeline that publishes application events to a Pub/Sub topic. You need to aggregate events across hourly intervals before loading the results to BigQuery for analysis. Your solution must be scalable so it can process and load large volumes of events to BigQuery. What should you do?

A. Create a streaming Dataflow job to continually read from the Pub/Sub topic and perform the necessary aggregations using tumbling windows

B. Schedule a batch Dataflow job to run hourly, pulling all available messages from the Pub-Sub topic and performing the necessary aggregations

C. Schedule a Cloud Function to run hourly, pulling all avertable messages from the Pub/Sub topic and performing the necessary aggregations

D. Create a Cloud Function to perform the necessary data processing that executes using the Pub/Sub trigger every time a new message is published to the topic.

Answer: A

Explanation:

Question: 259

Your company is migrating its on-premises data warehousing solution to BigQuery. The existing data warehouse uses trigger-based change data capture (CDC) to apply daily updates from transactional database sources Your company wants to use BigQuery to improve its handling of CDC and to optimize the performance of the data warehouse Source system changes must be available for query m near-real time using tog-based CDC streams You need to ensure that changes in the BigQuery reporting table are available with minimal latency and reduced overhead. What should you do? Choose 2 answers

A. Perform a DML INSERT UPDATE, or DELETE to replicate each CDC record in the reporting table m real time.

B. Periodically DELETE outdated records from the reporting table

Periodically use a DML MERGE to simultaneously perform DML INSERT. UPDATE, and DELETE operations in the reporting table

C. Insert each new CDC record and corresponding operation type into a staging table in real time

D. Insert each new CDC record and corresponding operation type into the reporting table in real time and use a materialized view to expose only the current version of each unique record.

Answer: B, D

Explanation:

Question: 260

You've migrated a Hadoop job from an on-premises cluster to Dataproc and Good Storage. Your Spark job is a complex analytical workload fiat consists of many shufling operations, and initial data are parquet toes (on average 200-400 MB size each) You see some degradation in performance after the migration to Dataproc so you'd like to optimize for it. Your organization is very cost-sensitive so you'd Idee to continue using Dataproc on preemptibles (with 2 non-preemptibles workers only) for this workload. What should you do?

A. Switch from HODs to SSDs override the preemptible VMs configuration to increase the boot disk size

B. Increase the see of your parquet files to ensure them to be 1 GB minimum

C. Switch to TFRecords format (appr 200 MB per We) instead of parquet files

D. Switch from HDDs to SSDs. copy initial data from Cloud Storage to Hadoop Distributed File System (HDFS) run the Spark job and copy results back to Cloud Storage

Answer: A

Explanation:

Question: 261

Your company currently runs a large on-premises cluster using Spark Hive and Hadoop Distributed File System (HDFS) in a colocation facility. The duster is designed to support peak usage on the system, however, many jobs are batch n nature, and usage of the cluster fluctuates quite dramatically.

Your company is eager to move to the cloud to reduce the overhead associated with on-premises infrastructure and maintenance and to benefit from the cost savings. They are also hoping to modernize their existing infrastructure to use more servers offerings m order to take advantage of the cloud Because of the tuning of their contract renewal with the colocation facility they have only 2 months for their initial migration How should you recommend they approach thee upcoming migration strategy so they can maximize their cost savings in the cloud will still executing the migration in time?

A. Migrate the workloads to Dataproc plus HOPS, modernize later

B. Migrate the workloads to Dataproc plus Cloud Storage modernize later

C. Migrate the Spark workload to Dataproc plus HDFS, and modernize the Hive workload for BigQuery

D. Modernize the Spark workload for Dataflow and the Hive workload for BigQuery

Answer: D

Explanation:

Question: 262

You are collecting loT sensor data from millions of devices across the world and storing the data in BigQuery. Your access pattern is based on recent data tittered by location_id and device_version with the following query:

```
SELECT
    MAX(temperature)
FROM
    acme_iot_data.sensors
WHERE
    create_date > DATE_SUB(CURRENT_DATE(), INTERVAL 7 day)
    AND location_id = "SW1W9TQ"
    AND device_version = "202007z3"
```

You want to optimize your queries for cost and performance. How should you structure your data?

A. Partition table data by create_date, location_id and device_version

B. Partition table data by create_date cluster table data by tocation_id and device_version

C. Cluster table data by create_date location_id and device_version

D. Cluster table data by create_date, partition by location and device_version

Answer: C

Explanation:

Question: 263

You want to optimize your queries for cost and performance. How should you structure your data?

A. Partition table data by create_date, location_id and device_version

B. Partition table data by create_date cluster table data by location_Id and device_version

C. Cluster table data by create_date location_id and device_version

D. Cluster table data by create_date partition by locationed and device_version

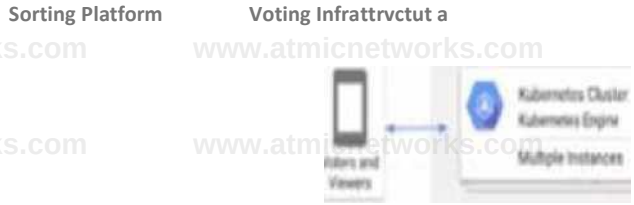
Answer: B

Explanation:

Question: 264

A live TV show asks viewers to cast votes using their mobile phones. The event generates a large volume of data during a 3 minute period. You are in charge of the Voting restructure* and must ensure that the platform can handle the load and Hal all votes are processed. You must display partial results write voting is open. After voting doses you need to count the votes exactly once white optimizing cost. What should you do?

OtetOhMf*



A. Create a Memorystore instance with a high availability (HA) configuration

B. Write votes to a Pub Sub tope and have Cloud Functions subscribe to it and write voles to BigQuery

C. Write votes to a Pub/Sub tope and toad into both Bigtable and BigQuery via a Dataflow pipeline Query Bigtable for real-time results and BigQuery for later analysis Shutdown the Bigtable instance when voting concludes

D Create a Cloud SQL for PostgreSQL database with high availability (HA) configuration and multiple read replicas

Answer: C

Explanation:

Question: 265

You are updating the code for a subscriber to a Put/Sub feed. You are concerned that upon deployment the subscriber may erroneously acknowledge messages, leading to message loss. You subscriber is not set up to retain acknowledged messages. What should you do to ensure that you can recover from errors after deployment?

A. Use Cloud Build for your deployment if an error occurs after deployment, use a Seek operation to locate a tmestamp logged by Cloud Build at the start of the deployment

B. Create a Pub/Sub snapshot before deploying new subscriber code. Use a Seek operation to redeliver messages that became available after the snapshot was created

C. Set up the Pub/Sub emulator on your local machine Validate the behavior of your new subscriber togs before deploying it to production

D. Enable dead-lettering on the Pub/Sub topic to capture messages that aren't successful acknowledged if an error occurs after deployment, re-deliver any messages captured by the deadletter queue

Answer: B

Explanation:

Question: 266

Government regulations in the banking industry mandate the protection of client's personally identifiable information (PII). Your company requires PII to be access controlled encrypted and compliant with major data protection standards In addition to using Cloud Data Loss Prevention (Cloud DIP) you want to follow Google-recommended practices and use service accounts to control access to PII. What should you do?

A. Assign the required identity and Access Management (IAM) roles to every employee, and create a single service

account to access protect resources

B. Use one service account to access a Cloud SQL database and use separate service accounts for each human user

C. Use Cloud Storage to comply with major data protection standards. Use one service account shared by all users

D. Use Cloud Storage to comply with major data protection standards. Use multiple service accounts attached to IAM groups to grant the appropriate access to each group

Answer: D

Explanation:

Question: 267

You are migrating a table to BigQuery and are deeding on the data model. Your table stores information related to purchases made across several store locations and includes information like the time of the transaction, items purchased, the store ID and the city and state in which the store is located You frequently query this table to see how many of each item were sold over the past 30 days and to look at purchasing trends by state city and individual store. You want to model this table to minimize query time and cost. What should you do?

A. Partition by transaction time; cluster by state first, then city then store ID

B. Partition by transaction tome cluster by store ID first, then city, then stale

C. Top-level cluster by stale first, then city then store

D. Top-level cluster by store ID first, then city then state.

Answer: C

Explanation:

Question: 268

You are working on a linear regression model on BigQuery ML to predict a customer's likelihood of purchasing your company's products. Your model uses a city name variable as a key predictive component in order to train and serve the

model your data must be organized in columns. You want to prepare your data using the least amount of coding while maintaining the predictable variables. What should you do?

A. Use SQL in BigQuery to transform the stale column using a one-hot encoding method, and make each city a column with binary values.

B. Create a new view with BigQuery that does not include a column which city information.

C. Cloud Data Fusion to assign each city to a region that is labeled as 1, 2 3, 4, or 5, and then use that number to represent the city in the model.

D. Use TensorFlow to create a categorical variable with a vocabulary list. Create the vocabulary file and upload that as part of your model to BigQuery ML.

Answer: C

Explanation:

Question: 269

Your new customer has requested daily reports that show their net consumption of Google Cloud compute resources and who used the resources. You need to quickly and efficiently generate these daily reports. What should you do?

A. Do daily exports of Cloud Logging data to BigQuery. Create views filtering by project, log type, resource, and user.

B. Filter data in Cloud Logging by project, resource, and user; then export the data in CSV format.

C. Filter data in Cloud Logging by project, log type, resource, and user, then import the data into BigQuery.

D. Export Cloud Logging data to Cloud Storage in CSV format. Cleanse the data using Dataprep, filtering by project, resource, and user.

Answer: B

Explanation:

https://cloud.google.com/logging/docs/view/logs-explorer-interface?cloudshell=true

Question: 270

You have a BigQuery table that ingests data directly from a Pub/Sub subscription. The ingested data is encrypted with a Google-managed encryption key. You need to meet a new organization policy that requires you to use keys from a centralized Cloud Key Management Service (Cloud KMS) project to encrypt data at rest. What should you do?

A. Create a new BigOuory table by using customer-managed encryption keys (CMEK), and migrate the data from the old BigQuery table.

B. Create a new BigOuery table and Pub/Sub topic by using customer-managed encryption keys (CMEK), and migrate the data from the old Bigauery table.

C. Create a new Pub/Sub topic with CMEK and use the existing BigQuery table by using Google- managed encryption key.

D. Use Cloud KMS encryption key with Dataflow to ingest the existing Pub/Sub subscription to the existing BigQuery table.

Answer: A

Explanation:

To use CMEK for BigQuery, you need to create a key ring and a key in Cloud KMS, and then specify the key resource name when creating or updating a BigQuery table. You cannot change the encryption type of an existing table, so you need to create a new table with CMEK and copy the data from the old table with Google-managed encryption key.

Reference:

Customer-managed Cloud KMS keys | BigQuery | Google Cloud

Creating and managing encryption keys | Cloud KMS Documentation | Google Cloud

Question: 271

You are designing a fault-tolerant architecture to store data in a regional BigOuery dataset. You need to ensure that your application is able to recover from a corruption event in your tables that occurred within the past seven days. You want to adopt managed services with the lowest RPO and most costeffective solution. What should you do?

A. Export the data from BigQuery into a new table that excludes the corrupted data.

B. Migrate your data to multi-region BigQuery buckets.

C. Access historical data by using time travel in BigQuery.

D. Create a BigQuery table snapshot on a daily basis.

Answer: C

Explanation:

Time travel is a feature of BigQuery that allows you to query and recover data from any point within the past seven days. You can use the FOR SYSTEM_TIME AS OF clause in your SQL query to specify the timestamp of the data you want to access. This way, you can restore your tables to a previous state before the corruption event occurred. Time travel is automatically enabled for all datasets and does not incur any additional cost or storage.

Reference:

Data retention with time travel and fail-safe | BigQuery | Google Cloud

BigQuery Time Travel: How to access Historical Data? | Easy Steps

Question: 272

You are developing an Apache Beam pipeline to extract data from a Cloud SQL instance by using JdbclO. You have two projects running in Google Cloud. The pipeline will be deployed and executed on Dataflow in Project A. The Cloud SQL instance is running jn Project B and does not have a public IP address. After deploying the pipeline, you noticed that the pipeline failed to extract data from the Cloud SQL instance due to connection failure. You verified that VPC Service

Controls and shared VPC are not in use in these projects. You want to resolve this error while ensuring that the data does not go through the public internet. What should you do?

A. Set up VPC Network Peering between Project A and Project B. Add a firewall rule to allow the peered subnet range to access all instances on the network.

B. Turn off the external IP addresses on the Dataflow worker. Enable Cloud NAT in Project A.

C. Set up VPC Network Peering between Project A and Project B. Create a Compute Engine instance without external IP address in Project B on the peered subnet to serve as a proxy server to the Cloud SQL database.

D. Add the external IP addresses of the Dataflow worker as authorized networks in the Cloud SOL instance.

Answer: C

Explanation:

Option A is incorrect because VPC Network Peering alone does not enable connectivity to Cloud SQL instances with private IP addresses. You also need to configure private services access and allocate an IP address range for the service producer network1.

Option B is incorrect because Cloud NAT does not support Cloud SQL instances with private IP addresses. Cloud NAT only provides outbound connectivity for resources that do not have public IP addresses, such as VMs, GKE clusters, and serverless instances2.

Option C is correct because it allows you to use a Compute Engine instance as a proxy server to connect to the Cloud SQL database over the peered network. The proxy server does not need an external IP address because it can communicate with the Dataflow workers and the Cloud SQL instance using internal IP addresses. You need to install the Cloud SQL Auth proxy on the proxy server and configure it to use a service account that has the Cloud SQL Client role.

Option D is incorrect because it requires you to assign public IP addresses to the Dataflow workers, which exposes the data to the public internet. This violates the requirement of ensuring that the data does not go through the public internet. Moreover, adding authorized networks does not work for Cloud SQL instances with private IP addresses.

Question: 273

You are on the data governance team and are implementing security requirements to deploy resources. You need to ensure that resources are limited to only the europe-west 3 region You want to follow Google-recommended practices What should you do?

A. Deploy resources with Terraform and implement a variable validation rule to ensure that the region is set to the europe-west3 region for all resources.

B. Set the constraints/gcp. resourceLocations organization policy constraint to in:eu-locations.

C. Create a Cloud Function to monitor all resources created and automatically destroy the ones created outside the europe-west3 region.

D. Set the constraints/gcp. resourceLocations organization policy constraint to in: europe-west3- locations.

Answer: D

Explanation:

To ensure that resources are limited to only the europe-west3 region, you should set the organization policy constraint constraints/gcp.resourceLocations to in:europe-west3-locations. This policy restricts the deployment of resources to the specified locations, which in this case is the europe-west3 region. By setting this policy, you enforce location compliance across your Google Cloud resources, aligning with the best practices for data governance and regulatory compliance.

Reference:

Professional Data Engineer Certification Exam Guide | Learn - Google Cloud1.

Preparing for Google Cloud Certification: Cloud Data Engineer2.

Professional Data Engineer Certification | Learn | Google Cloud3.

3: Professional Data Engineer Certification | Learn | Google Cloud 2: Preparing for Google Cloud Certification: Cloud Data Engineer 1: Professional Data Engineer Certification Exam Guide | Learn - Google Cloud

Question: 274

You have a BigQuery table that contains customer data, including sensitive information such as names and addresses. You need to share the customer data with your data analytics and consumer support teams securely. The data analytics team needs to access the data of all the customers, but must not be able to access the sensitive dat

a. The consumer support team needs access to all data columns, but must not be able to access customers that no longer have active contracts. You enforced these requirements by using an authorized dataset and policy tags After implementing these steps, the data analytics team reports that they still have access to the sensitive columns. You need to ensure that the data analytics team does not have access to restricted data What should you do?

Choose 2 answers

A. Create two separate authorized datasets; one for the data analytics team and another for the consumer support team.

B. Ensure that the data analytics team members do not have the Data Catalog Fine-Grained Reader role for the policy tags.

C. Enforce access control in the policy tag taxonomy.

D. Remove the bigquery. dataViewer role from the data analytics team on the authorized datasets.

E. Replace the authorized dataset with an authorized view Use row-level security and apply filter_ expression to limit data access.

Answer: B, C

Explanation:

To ensure that the data analytics team does not have access to sensitive columns, you should:

B . Ensure that the data analytics team members do not have the Data Catalog Fine-Grained Reader role for the policy tags. This role allows users to read metadata for data assets that have policy tags applied, which could include sensitive information.

C . Enforce access control in the policy tag taxonomy. By setting access control at the policy tag level, you can restrict access to specific columns within a dataset, ensuring that only authorized users can view sensitive data.

Question: 275

You are building a streaming Dataflow pipeline that ingests noise level data from hundreds of sensors placed near construction sites across a city. The sensors measure noise level every ten seconds, and send that data to the pipeline

when levels reach above 70 dB

A. You need to detect the average noise level from a sensor when data is received for a duration of more than 30 minutes, but the window ends when no data has been received for 15 minutes What should you do?

A. Use session windows with a 30-mmute gap duration.

B. Use tumbling windows with a 15-mmute window and a fifteen-minute. withAllowedLateness operator.

C. Use session windows with a 15-minute gap duration.

D. Use hopping windows with a 15-mmute window, and a thirty-minute period.

Answer: B

Explanation:

Session windows are dynamic windows that group elements based on the periods of activity. They are useful for streaming data that is irregularly distributed with respect to time. In this case, the noise level data from the sensors is only sent when it exceeds a certain threshold, and the duration of the noise events may vary. Therefore, session windows can capture the average noise level for each sensor during the periods of high noise, and end the window when there is no data for a specified gap duration. The gap duration should be 15 minutes, as the requirement is to end the window when

no data has been received for 15 minutes. A 30-minute gap duration would be too long and may miss some noise events that are shorter than 30 minutes. Tumbling windows and hopping windows are fixed windows that group elements based on a fixed time interval. They are not suitable for this use case, as they may split or overlap the noise events from the sensors, and do not account for the periods of inactivity. Reference:

Windowing concepts

Session windows

Windowing in Dataflow

Question: 276

You maintain ETL pipelines. You notice that a streaming pipeline running on Dataflow is taking a long time to process incoming data, which causes output delays. You also noticed that the pipeline graph was automatically optimized by Dataflow and merged into one step. You want to identify where the potential bottleneck is occurring. What should you do?

A. Insert a Reshufle operation after each processing step, and monitor the execution details in the Dataflow console.

B. Log debug information in each ParDo function, and analyze the logs at execution time.

C. Insert output sinks after each key processing step, and observe the writing throughput of each block.

D. Verify that the Dataflow service accounts have appropriate permissions to write the processed data to the output sinks

Answer: A

Explanation:

A Reshufle operation is a way to force Dataflow to split the pipeline into multiple stages, which can help isolate the performance of each step and identify bottlenecks. By monitoring the execution details in the Dataflow console, you can see the time, CPU, memory, and disk usage of each stage, as well as the number of elements and bytes processed. This can help you diagnose where the pipeline is slowing down and optimize it accordingly. Reference:

1: Reshufling your data

2: Monitoring pipeline performance using the Dataflow monitoring interface

3: Optimizing pipeline performance

Question: 277

Your organization is modernizing their IT services and migrating to Google Cloud. You need to organize the data that will be stored in Cloud Storage and BigQuery. You need to enable a data mesh approach to share the data between sales, product design, and marketing departments What should you do?

A. 1Create a project for storage of the data for your organization.

2 Create a central Cloud Storage bucket with three folders to store the files for each department.

3 . Create a central BigQuery dataset with tables prefixed with the department name.

4 Give viewer rights for the storage project for the users of your departments.

B. 1Create a project for storage of the data for each of your departments.

2 Enable each department to create Cloud Storage buckets and BigQuery datasets.

3 . Create user groups for authorized readers for each bucket and dataset.

4 Enable the IT team to administer the user groups to add or remove users as the departments' request.

C. 1 Create multiple projects for storage of the data for each of your departments' applications.

2 Enable each department to create Cloud Storage buckets and BigQuery datasets.

3 . Publish the data that each department shared in Analytics Hub.

4 Enable all departments to discover and subscribe to the data they need in Analytics Hub.

D. 1 Create multiple projects for storage of the data for each of your departments' applications.

2 Enable each department to create Cloud Storage buckets and BigQuery datasets.

3 In Dataplex, map each department to a data lake and the Cloud Storage buckets, and map the BigQuery datasets to zones.

4 Enable each department to own and share the data of their data lakes.

Answer: C

Explanation:

Implementing a data mesh approach involves treating data as a product and enabling decentralized data ownership and architecture. The steps outlined in option C support this approach by creating separate projects for each department, which aligns with the principle of domain-oriented decentralized data ownership. By allowing departments to create their own Cloud Storage buckets and BigQuery datasets, it promotes autonomy and self-service. Publishing the data in Analytics Hub facilitates data sharing and discovery across departments, enabling a collaborative environment where data can be easily accessed and utilized by different parts of the organization.

Reference:

Architecture and functions in a data mesh - Google Cloud

Professional Data Engineer Certification Exam Guide | Learn - Google Cloud

Build a Data Mesh with Dataplex | Google Cloud Skills Boost

Question: 278

You have a network of 1000 sensors. The sensors generate time series data: one metric per sensor per second, along with a timestamp. You already have 1 TB of data, and expect the data to grow by 1 GB every day You need to access this data

in two ways. The first access pattern requires retrieving the metric from one specific sensor stored at a specific timestamp, with a median single-digit millisecond latency. The second access pattern requires running complex analytic queries on the data, including joins, once a day. How should you store this data?

A. Store your data in Bigtable Concatenate the sensor ID and timestamp and use it as the row key Perform an export to BigQuery every day.

B. Store your data in BigQuery Concatenate the sensor ID and timestamp. and use it as the primary key.

C. Store your data in Bigtable Concatenate the sensor ID and metric, and use it as the row key Perform an export to BigQuery every day.

D. Store your data in BigQuery. Use the metric as a primary key.

Answer: A

Explanation:

To store your data in a way that meets both access patterns, you should:

A . Store your data in Bigtable Concatenate the sensor ID and timestamp and use it as the row key Perform an export to BigQuery every day. This option allows you to leverage the high performance and scalability of Bigtable for low-latency point queries on sensor data, as well as the powerful analytics capabilities of BigQuery for complex queries on large datasets. By using the sensor ID and timestamp as the row key, you can ensure that your data is sorted and distributed evenly across Bigtable nodes, and that you can easily retrieve the metric for a specific sensor and time. By performing an export to BigQuery every day, you can transfer your data to a columnar storage format that is optimized for analytical queries, and take advantage of BigQuery's features such as partitioning, clustering, and caching.

B . Store your data in BigQuery Concatenate the sensor ID and timestamp. and use it as the primary key. This option is not optimal because BigQuery is not designed for low-latency point queries, and using a concatenated primary key may result in poor performance and high costs. BigQuery does not support primary keys natively, and you would have to use a unique constraint or a hash function to enforce uniqueness. Moreover, BigQuery charges by the amount of data scanned, so using a long and complex primary key may increase the query cost and complexity.

C . Store your data in Bigtable Concatenate the sensor ID and metric, and use it as the row key Perform an export to BigQuery every day. This option is not optimal because using the sensor ID and metric as the row key may result in data skew and hotspots in Bigtable, as some sensors may generate more metrics than others, or some metrics may be more common than others. This may affect the performance and availability of Bigtable, as well as the efficiency of the export to BigQuery.

D . Store your data in BigQuery. Use the metric as a primary key. This option is not optimal because using the metric as a

primary key may result in data duplication and inconsistency in BigQuery, as multiple sensors may generate the same metric at different times, or the same sensor may generate different metrics at the same time. This may affect the accuracy and reliability of your analytical queries, as well as the query cost and complexity.

Question: 279

You have a variety of files in Cloud Storage that your data science team wants to use in their models

Currently, users do not have a method to explore, cleanse, and validate the data in Cloud Storage.
You are looking for a low code solution that can be used by your data science team to quickly cleanse and explore data within Cloud Storage. What should you do?

A. Load the data into BigQuery and use SQL to transform the data as necessary Provide the data science team access to staging tables to explore the raw data.

B. Provide the data science team access to Dataflow to create a pipeline to prepare and validate the raw data and load data into BigQuery for data exploration.

C. Provide the data science team access to Dataprep to prepare, validate, and explore the data within Cloud Storage.

D. Create an external table in BigQuery and use SQL to transform the data as necessary Provide the data science team access to the external tables to explore the raw data.

Answer: C

Explanation:

Dataprep is a low code, serverless, and fully managed service that allows users to visually explore, cleanse, and validate data in Cloud Storage. It also provides features such as data profiling, data quality, data transformation, and data lineage. Dataprep is integrated with BigQuery, so users can easily export the prepared data to BigQuery for further analysis or modeling. Dataprep is a suitable solution for the data science team to quickly and easily work with the data in Cloud Storage, without having to write code or manage infrastructure. The other options are not as suitable as Dataprep for this use case, because they either require more coding, more infrastructure management, or more data movement. Loading the data into BigQuery, either directly or through Dataflow, would incur additional costs and latency, and may not provide the same level of data exploration and validation as Dataprep. Creating an external table in BigQuery would allow users to query the data in Cloud Storage, but would not provide the same level of data cleansing and transformation as Dataprep. Reference:

Dataprep overview

Dataprep features

Dataprep and BigQuery integration

Question: 280

You store and analyze your relational data in BigQuery on Google Cloud with all data that resides in US regions. You also have a variety of object stores across Microsoft Azure and Amazon Web Services (AWS), also in US regions. You want to query all your data in BigQuery daily with as little movement of data as possible. What should you do?

A. Load files from AWS and Azure to Cloud Storage with Cloud Shell gautil rsync arguments.

B. Create a Dataflow pipeline to ingest files from Azure and AWS to BigQuery.

C. Use the BigQuery Omni functionality and BigLake tables to query files in Azure and AWS.

D. Use BigQuery Data Transfer Service to load files from Azure and AWS into BigQuery.

Answer: B

Explanation:

BigQuery Omni is a multi-cloud analytics solution that lets you use the BigQuery interface to analyze data stored in other public clouds, such as AWS and Azure, without moving or copying the data. BigLake tables are a type of external table that let you query structured data in external data stores with access delegation. By using BigQuery Omni and BigLake tables, you can query data in AWS and Azure object stores directly from BigQuery, with minimal data movement and consistent performance. Reference:

1: Introduction to BigLake tables

2: Deep dive on how BigLake accelerates query performance

3: BigQuery Omni and BigLake (Analytics Data Federation on GCP)

Question: 281

Your business users need a way to clean and prepare data before using the data for analysis. Your business users are less technically savvy and prefer to work with graphical user interfaces to define their transformations. After the data has been transformed, the business users want to perform their analysis directly in a spreadsheet. You need to recommend a solution that they can use. What should you do?

A. Use Dataprep to clean the data, and write the results to BigQuery Analyze the data by using Connected Sheets.

B. Use Dataprep to clean the data, and write the results to BigQuery Analyze the data by using Looker Studio.

C. Use Dataflow to clean the data, and write the results to BigQuery. Analyze the data by using Connected Sheets.

D. Use Dataflow to clean the data, and write the results to BigQuery. Analyze the data by using Looker Studio.

Answer: A

Explanation:

For business users who are less technically savvy and prefer graphical user interfaces, Dataprep is an ideal tool for cleaning and preparing data, as it offers a user-friendly interface for defining data transformations without the need for coding. Once the data is cleaned and prepared, writing the results to BigQuery allows for the storage and management of large datasets. Analyzing the data using Connected Sheets enables business users to work within the familiar environment of a spreadsheet, leveraging the power of BigQuery directly within Google Sheets. This solution aligns with the needs of the users and follows Google's recommended practices for data cleaning, preparation, and analysis.

Reference:

Connected Sheets | Google Sheets | Google for Developers

Professional Data Engineer Certification Exam Guide | Learn - Google Cloud

Engineer Data in Google Cloud | Google Cloud Skills Boost - Qwiklabs

Question: 282

You are developing a model to identify the factors that lead to sales conversions for your customers.
You have completed processing your dat

a. You want to continue through the model development lifecycle. What should you do next?

A. Use your model to run predictions on fresh customer input data.

B. Test and evaluate your model on your curated data to determine how well the model performs.

C. Monitor your model performance, and make any adjustments needed.

D. Delineate what data will be used for testing and what will be used for training the model.

Answer: D

Explanation:

After processing your data, the next step in the model development lifecycle is to test and evaluate your model on the curated data. This is crucial to determine the performance of the model and to understand how well it can predict sales conversions for your customers. The evaluation phase involves using various metrics and techniques to assess the accuracy, precision, recall, and other relevant performance indicators of the model. It helps in identifying any issues or areas for improvement before deploying the model in a production environment. Reference:: The information provided here is verified by the Google Professional Data Engineer Certification Exam Guide and related resources, which outline the steps and best practices in the model development lifecycle

Question: 283

You need to modernize your existing on-premises data strategy. Your organization currently uses.

- Apache Hadoop clusters for processing multiple large data sets, including on-premises Hadoop Distributed File System (HDFS) for data replication.
- Apache Airflow to orchestrate hundreds of ETL pipelines with thousands of job steps.

You need to set up a new architecture in Google Cloud that can handle your Hadoop workloads and requires minimal changes to your existing orchestration processes. What should you do?

A. Use Dataproc to migrate Hadoop clusters to Google Cloud, and Cloud Storage to handle any HDFS use cases Convert your ETL pipelines to Dataflow.

B. Use Bigtable for your large workloads, with connections to Cloud Storage to handle any HDFS use cases Orchestrate your pipelines with Cloud Composer.

C. Use Dataproc to migrate your Hadoop clusters to Google Cloud, and Cloud Storage to handle any HDFS use cases. Use Cloud Data Fusion to visually design and deploy your ETL pipelines.

D. Use Dataproc to migrate Hadoop clusters to Google Cloud, and Cloud Storage to handle any HDFS use cases. Orchestrate your pipelines with Cloud Composer..

Answer: D

Explanation:

Dataproc is a fully managed service that allows you to run Apache Hadoop and Spark workloads on Google Cloud. It is compatible with the open source ecosystem, so you can migrate your existing Hadoop clusters to Dataproc with minimal changes. Cloud Storage is a scalable, durable, and costeffective object storage service that can replace HDFS for storing and accessing data. Cloud Storage offers interoperability with Hadoop through connectors, so you can use it as a data source or sink for your Dataproc jobs. Cloud Composer is a fully managed service that allows you to create, schedule, and monitor workflows using Apache Airflow. It is integrated with Google Cloud services, such as Dataproc, BigQuery, Dataflow, and Pub/Sub, so you can orchestrate your ETL pipelines across different platforms. Cloud Composer is compatible with your existing Airflow code, so you can migrate your existing orchestration processes to Cloud Composer with minimal changes.

The other options are not as suitable as Dataproc and Cloud Composer for this use case, because they either require more changes to your existing code, or do not meet your requirements. Dataflow is a fully managed service that allows you to create and run scalable data processing pipelines using Apache Beam. However, Dataflow is not compatible with your existing Hadoop code, so you would need to rewrite your ETL pipelines using Beam. Bigtable is a fully managed NoSQL database service that can handle large and complex data sets. However, Bigtable is not compatible with your existing Hadoop code, so you would need to rewrite your queries and applications using Bigtable APIs. Cloud Data Fusion is a fully managed service that allows you to visually design and deploy data integration pipelines using a graphical interface. However, Cloud Data Fusion is not compatible with your existing Airflow code, so you would need to recreate your orchestration processes using Cloud Data Fusion UI. Reference:

Dataproc overview

Cloud Storage connector for Hadoop

Cloud Composer overview

Question: 284

You are running a Dataflow streaming pipeline, with Streaming Engine and Horizontal Autoscaling enabled. You have set the maximum number of workers to 1000. The input of your pipeline is Pub/Sub messages with notifications from Cloud Storage One of the pipeline transforms reads CSV files and emits an element for every CSV line. The Job performance is low. the pipeline is using only 10 workers, and you notice that the autoscaler is not spinning up additional workers. What should you do to improve performance?

A. Use Dataflow Prime, and enable Right Fitting to increase the worker resources.

B. Update the job to increase the maximum number of workers.

C. Enable Vertical Autoscaling to let the pipeline use larger workers.

D. Change the pipeline code, and introduce a Reshufle step to prevent fusion.

Answer: A

Explanation:

Fusion is an optimization technique that Dataflow applies to merge multiple transforms into a single stage. This reduces the overhead of shufling data between stages, but it can also limit the parallelism and scalability of the pipeline. By introducing a Reshufle step, you can force Dataflow to split the pipeline into multiple stages, which can increase the number of workers that can process the data in parallel. Reshufle also adds randomness to the data distribution, which can help balance the workload across workers and avoid hot keys or skewed data.

Reference:

1: Streaming pipelines

2: Batch vs Streaming Performance in Google Cloud Dataflow

3: Deploy Dataflow pipelines

4: How Distributed Shufle improves scalability and performance in Cloud Dataflow pipelines

5: Managing costs for Dataflow batch and streaming data processing

Question: 285

You are designing a Dataflow pipeline for a batch processing job. You want to mitigate multiple zonal failures at job submission time. What should you do?

A. Specify a worker region by using the —region flag.

B. Set the pipeline staging location as a regional Cloud Storage bucket.

C. Submit duplicate pipelines in two different zones by using the —zone flag.

D. Create an Eventarc trigger to resubmit the job in case of zonal failure when submitting the job.

Answer: B

Explanation:

By specifying a worker region, you can run your Dataflow pipeline in a multi-zone or multi-region configuration, which provides higher availability and resilience in case of zonal failures1. The — region flag allows you to specify the regional endpoint for your pipeline, which determines the location of the Dataflow service and the default location of the Compute Engine resources1. If you do not specify a zone by using the —zone flag, Dataflow automatically selects a zone

within the region for your job workers1. This option is recommended over submitting duplicate pipelines in two different zones, which would incur additional costs and complexity. Setting the pipeline staging location as a regional Cloud Storage bucket does not affect the availability of your pipeline, as the staging location only stores the pipeline code and dependencies2. Creating an Eventarc trigger to resubmit the job in case of zonal failure is not a reliable solution, as it depends on the availability of the Eventarc service and the zonal resources at the time of resubmission. Reference:

1: Pipeline troubleshooting and debugging | Cloud Dataflow | Google Cloud

3: Regional endpoints | Cloud Dataflow | Google Cloud

Question: 286

You are designing the architecture to process your data from Cloud Storage to BigQuery by using Dataflow. The network team provided you with the Shared VPC network and subnetwork to be used by your pipelines. You need to enable the deployment of the pipeline on the Shared VPC network. What should you do?

A. Assign the compute. networkUser role to the Dataflow service agent.

B. Assign the compute.networkUser role to the service account that executes the Dataflow pipeline.

C. Assign the dataflow, admin role to the Dataflow service agent.

D. Assign the dataflow, admin role to the service account that executes the Dataflow pipeline.

Answer: B

Explanation:

To use a Shared VPC network for a Dataflow pipeline, you need to specify the subnetwork parameter with the full URL of the subnetwork, and grant the service account that executes the pipeline the compute.networkUser role in the host project. This role allows the service account to use the subnetworks in the Shared VPC network. The Dataflow service agent does not need this role, as it only creates and manages the resources for the pipeline, but does not execute it. The dataflow.admin role is not related to the network access, but to the permissions to create and delete Dataflow jobs and resources. Reference:

Specify a network and subnetwork | Cloud Dataflow | Google Cloud

How to config dataflow Pipeline to use a Shared VPC?

Question: 287

You are building an ELT solution in BigQuery by using Dataform. You need to perform uniqueness and null value checks on your final tables. What should you do to efficiently integrate these checks into your pipeline?

A. Build Dataform assertions into your code

B. Write a Spark-based stored procedure.

C. Build BigQuery user-defined functions (UDFs).

D. Create Dataplex data quality tasks.

Answer: A

Explanation:

Dataform assertions are data quality tests that find rows that violate one or more rules specified in the query. If the query returns any rows, the assertion fails. Dataform runs assertions every time it updates your SQL workflow and alerts you if any assertions fail. You can create assertions for all Dataform table types: tables, incremental tables, views, and materialized views. You can add built-in assertions to the config block of a table, such as nonNull and rowConditions, or create manual assertions with SQLX for advanced use cases. Dataform automatically creates views in BigQuery that contain the results of compiled assertion queries, which you can inspect to debug failing assertions. Dataform assertions are an efficient way to integrate data quality checks into your ELT solution in BigQuery by using Dataform. Reference: Test tables with assertions | Dataform | Google Cloud, Test data quality with assertions | Dataform, Data quality tests and documenting datasets | Dataform, Data quality testing with SQL assertions | Dataform

Question: 288

You are designing a real-time system for a ride hailing app that identifies areas with high demand for rides to effectively reroute available drivers to meet the demand. The system ingests data from multiple sources to Pub/Sub. processes the data, and stores the results for visualization and analysis in real-time dashboards. The data sources include driver location updates every 5 seconds and app- based booking events from riders. The data processing involves real-time aggregation of supply and demand data for the last 30 seconds, every 2 seconds, and storing the results in a low-latency system for visualization. What should you do?

A. Group the data by using a tumbling window in a Dataflow pipeline, and write the aggregated data to Memorystore

B. Group the data by using a hopping window in a Dataflow pipeline, and write the aggregated data to Memorystore

C. Group the data by using a session window in a Dataflow pipeline, and write the aggregated data to BigQuery.

D. Group the data by using a hopping window in a Dataflow pipeline, and write the aggregated data to BigQuery.

Answer: B

Explanation:
A hopping window is a type of sliding window that advances by a fixed period of time, producing overlapping windows. This is suitable for the scenario where the system needs to aggregate data for the last 30 seconds, every 2 seconds, and provide real-time updates. A Dataflow pipeline can implement the hopping window logic using Apache Beam, and process both streaming and batch data sources. Memorystore is a low-latency, in-memory data store that can serve the aggregated data to the visualization layer. BigQuery is not a good choice for this scenario, as it is not optimized for low-latency queries and frequent updates.

Question: 289

You orchestrate ETL pipelines by using Cloud Composer One of the tasks in the Apache Airflow directed acyclic graph (DAG) relies on a third-party service. You want to be notified when the task does not succeed. What should you do?

A. Configure a Cloud Monitoring alert on the sla_missed metric associated with the task at risk to trigger a notification.

B. Assign a function with notification logic to the sla_miss_callback parameter for the operator responsible for the task at risk.

C. Assign a function with notification logic to the on_retry_callback parameter for the operator responsible for the task at risk.

D. Assign a function with notification logic to the on_failure_callback parameter for the operator responsible for the task at risk.

Answer: D

Explanation:

By assigning a function with notification logic to the on_failure_callback parameter, you can customize the action that is taken when a task fails in your DAG1. For example, you can send an email, a Slack message, or a PagerDuty alert

to notify yourself or your team about the task failure2. This option is more flexible and reliable than configuring a Cloud Monitoring alert on the sla_missed metric, which only triggers when a task misses its scheduled deadline3. The sla_miss_callback parameter is also related to the sla_missed metric, and it is executed when the task instance has not succeeded and the time is past the task's scheduled execution date plus its sla4. The on_retry_callback parameter is executed before a task is retried4. These options are not suitable for notifying when a task does not succeed, as they depend on the task's schedule and retry settings, which may not reflect the actual task completion status. Reference:

1: Callbacks | Cloud Composer | Google Cloud

2: How to Send an Email on Task Failure in Airflow - Astronomer

3: Monitoring SLA misses | Cloud Composer | Google Cloud

4: BaseOperator | Apache Airflow Documentation

Question: 290

You have a streaming pipeline that ingests data from Pub/Sub in production. You need to update this streaming pipeline with improved business logic. You need to ensure that the updated pipeline reprocesses the previous two days of delivered Pub/Sub messages. What should you do?

Choose 2 answers

A. Use Pub/Sub Seek with a timestamp.

B. Use the Pub/Sub subscription clear-retry-policy flag.

C. Create a new Pub/Sub subscription two days before the deployment.

D. Use the Pub/Sub subscription retain-asked-messages flag.

E. Use Pub/Sub Snapshot capture two days before the deployment.

Answer: A, E

Explanation:

To update a streaming pipeline with improved business logic and reprocess the previous two days of delivered Pub/Sub messages, you should use Pub/Sub Seek with a timestamp and Pub/Sub Snapshot capture two days before the deployment. Pub/Sub Seek allows you to replay or purge messages in a subscription based on a time or a snapshot. Pub/Sub Snapshot allows you to capture the state of a subscription at a given point in time and replay messages from

that point. By using these features, you can ensure that the updated pipeline can process the messages that were delivered in the past two days without losing any data. Reference:

Pub/Sub Seek

Pub/Sub Snapshot

Question: 291

You stream order data by using a Dataflow pipeline, and write the aggregated result to Memorystore. You provisioned a Memorystore for Redis instance with Basic Tier. 4 GB capacity, which is used by 40 clients for read-only access. You are expecting the number of read-only clients to increase significantly to a few hundred and you need to be able to support the demand. You want to ensure that read and write access availability is not impacted, and any changes you make can be deployed quickly. What should you do?

A. Create multiple new Memorystore for Redis instances with Basic Tier (4 GB capacity) Modify the Dataflow pipeline and new clients to use all instances

B. Create a new Memorystore for Redis instance with Standard Tier Set capacity to 4 GB and read replica to No read replicas (high availability only). Delete the old instance.

C. Create a new Memorystore for Memcached instance Set a minimum of three nodes, and memory per node to 4 GB. Modify the Dataflow pipeline and all clients to use the Memcached instance Delete the old instance.

D. Create a new Memorystore for Redis instance with Standard Tier Set capacity to 5 GB and create multiple read replicas Delete the old instance.

Answer: D

Explanation:

The Basic Tier of Memorystore for Redis provides a standalone Redis instance that is not replicated and does not support read replicas. This means that it cannot scale horizontally to handle more read requests, and it does not provide high availability or automatic failover. If the number of read-only clients increases significantly, the Basic Tier instance may not be able to handle the demand and may impact the read and write access availability. Therefore, option A is not a good solution, as it would require creating multiple Basic Tier instances and modifying the Dataflow pipeline and the clients to distribute the load among them. This would increase the complexity and the management overhead of the solution.

The Standard Tier of Memorystore for Redis provides a highly available Redis instance that supports replication and read

replicas. Replication ensures that the data is backed up in another zone and can fail over automatically in case of a primary node failure. Read replicas allow scaling the read throughput by adding up to five replicas to an instance and using them for read-only queries. The Standard Tier also supports in-transit encryption and maintenance windows. Therefore, option D is the best solution, as it would create a new Standard Tier instance with a higher capacity (5 GB) and multiple read replicas to handle the increased demand. The old instance can be deleted after migrating the data to the new instance.

Option B is not a good solution, as it would create a new Standard Tier instance with the same capacity (4 GB) and no read replicas. This would not improve the read throughput or the availability of the solution.

Option C is not a good solution, as it would create a new Memorystore for Memcached instance, which is a different service that uses a different protocol and data model than Redis. This would require changing the code of the Dataflow pipeline and the clients to use the Memcached protocol and data structures, which would take more time and effort than migrating to a new Redis instance. Reference: Redis tier capabilities | Memorystore for Redis | Google Cloud, Pricing | Memorystore for Redis | Google Cloud, What is Memorystore? | Google Cloud Blog, Working with GCP Memorystore - Simple Talk - Redgate Software

Question: 292

You have an Oracle database deployed in a VM as part of a Virtual Private Cloud (VPC) network. You want to replicate and continuously synchronize 50 tables to BigQuery. You want to minimize the need to manage infrastructure. What should you do?

A. Create a Datastream service from Oracle to BigQuery, use a private connectivity configuration to the same VPC network, and a connection profile to BigQuery.

B. Create a Pub/Sub subscription to write to BigQuery directly Deploy the Debezium Oracle connector to capture changes in the Oracle database, and sink to the Pub/Sub topic.

C. Deploy Apache Kafka in the same VPC network, use Kafka Connect Oracle Change Data Capture (CDC), and Dataflow to stream the Kafka topic to BigQuery.

D. Deploy Apache Kafka in the same VPC network, use Kafka Connect Oracle change data capture (CDC), and the Kafka Connect Google BigQuery Sink Connector.

Answer: A

Explanation:

Datastream is a serverless, scalable, and reliable service that enables you to stream data changes from Oracle and MySQL databases to Google Cloud services such as BigQuery, Cloud SQL, Google Cloud Storage, and Cloud Pub/Sub. Datastream

captures and streams database changes using change data capture (CDC) technology. Datastream supports private connectivity to the source and destination systems using VPC networks. Datastream also provides a connection profile to BigQuery, which simplifies the configuration and management of the data replication. Reference:

Datastream overview

Creating a Datastream stream

Using Datastream with BigQuery

Question: 293

You want to schedule a number of sequential load and transformation jobs Data files will be added to a Cloud Storage bucket by an upstream process There is no fixed schedule for when the new data arrives Next, a Dataproc job is triggered to perform some transformations and write the data to BigQuery. You then need to run additional transformation jobs in BigQuery The transformation jobs are different for every table These jobs might take hours to complete You need to determine the most efficient and maintainable workflow to process hundreds of tables and provide the freshest data to your end users. What should you do?

A. 1Create an Apache Airflow directed acyclic graph (DAG) in Cloud Composer with sequential tasks by using the Cloud Storage. Dataproc. and BigQuery operators

2 Use a single shared DAG for all tables that need to go through the pipeline

3 Schedule the DAG to run hourly

B. 1 Create an Apache Airflow directed acyclic graph (DAG) in Cloud Composer with sequential tasks by using the Dataproc and BigQuery operators.

2 Create a separate DAG for each table that needs to go through the pipeline

3 Use a Cloud Storage object trigger to launch a Cloud Function that triggers the DAG

C. 1 Create an Apache Airflow directed acyclic graph (DAG) in Cloud Composer with sequential tasks by using the Cloud Storage, Dataproc. and BigQuery operators

2 Create a separate DAG for each table that needs to go through the pipeline

3 Schedule the DAGs to run hourly

D. 1 Create an Apache Airflow directed acyclic graph (DAG) in Cloud Composer with sequential tasks by using the Dataproc and BigQuery operators

2 Use a single shared DAG for all tables that need to go through the pipeline.

3 Use a Cloud Storage object trigger to launch a Cloud Function that triggers the DAG

Answer: B

Explanation:

This option is the most efficient and maintainable workflow for your use case, as it allows you to process each table independently and trigger the DAGs only when new data arrives in the Cloud Storage bucket. By using the Dataproc and BigQuery operators, you can easily orchestrate the load and transformation jobs for each table, and leverage the scalability and performance of these services12. By creating a separate DAG for each table, you can customize the transformation logic and parameters for each table, and avoid the complexity and overhead of a single shared DAG3. By using a Cloud Storage object trigger, you can launch a Cloud Function that triggers the DAG for the corresponding table, ensuring that the data is processed as soon as possible and reducing the idle time and cost of running the DAGs on a fixed schedule4 .

Option A is not efficient, as it runs the DAG hourly regardless of the data arrival, and it uses a single shared DAG for all tables, which makes it harder to maintain and debug. Option C is also not efficient, as it runs the DAGs hourly and does not leverage the Cloud Storage object trigger. Option D is not maintainable, as it uses a single shared DAG for all tables, and it does not use the Cloud Storage operator, which can simplify the data ingestion from the bucket. Reference:

1: Dataproc Operator | Cloud Composer | Google Cloud

2: BigQuery Operator | Cloud Composer | Google Cloud

3: Choose Workflows or Cloud Composer for service orchestration | Workflows | Google Cloud

4: Cloud Storage Object Trigger | Cloud Functions Documentation | Google Cloud

[5] : Triggering DAGs | Cloud Composer | Google Cloud

[6] : Cloud Storage Operator | Cloud Composer | Google Cloud

Question: 294

You have a table that contains millions of rows of sales data, partitioned by date Various applications and users query this data many times a minute. The query requires aggregating values by using avg. max. and sum, and does not require joining to other tables. The required aggregations are only computed over the past year of data, though you need to retain full historical data in the base tables You want to ensure that the query results always include the latest data from the tables, while also reducing computation cost, maintenance overhead, and duration. What should you do?

A. Create a materialized view to aggregate the base table data Configure a partition expiration on the base table to retain only the last one year of partitions.

B. Create a materialized view to aggregate the base table data include a filter clause to specify the last one year of partitions.

C. Create a new table that aggregates the base table data include a filter clause to specify the last year of partitions. Set up a scheduled query to recreate the new table every hour.

D. Create a view to aggregate the base table data Include a filter clause to specify the last year of partitions.

Answer: C

Explanation:

A materialized view is a database object that contains the results of a query, which can be updated periodically. It can improve the performance and efficiency of queries that involve aggregations, joins, or filters. By creating a materialized view to aggregate the base table data and include a filter clause to specify the last one year of partitions, you can ensure that the query results always include the latest data from the tables, while also reducing computation cost, maintenance overhead, and duration. The materialized view will automatically refresh when the base table data changes, and will only use the partitions that match the filter clause. Option A is incorrect because it will delete the historical data from the base table, which is not desired. Option C is incorrect because it will create a redundant table that needs to be updated manually by a scheduled query, which is more complex and costly than using a materialized view. Option D is incorrect because a view does not store any data, but only references the base table data, which means it will not reduce the computation cost or duration of the query. Reference:

Materialized views, ML models in data warehouse - Google Cloud

Data Engineering with Google Cloud Platform - Packt Subscription

Question: 295

Your organization has two Google Cloud projects, project A and project B. In project A, you have a Pub/Sub topic that receives data from confidential sources. Only the resources in project A should be able to access the data in that topic. You want to ensure that project B and any future project cannot access data in the project A topic. What should you do?

A. Configure VPC Service Controls in the organization with a perimeter around the VPC of project A.

B. Add firewall rules in project A so only traffic from the VPC in project A is permitted.

C. Configure VPC Service Controls in the organization with a perimeter around project A.

D. Use Identity and Access Management conditions to ensure that only users and service accounts in project A can access resources in project.

Answer: C

Explanation:

Identity and Access Management (IAM) is the recommended way to control access to Pub/Sub resources, such as topics and subscriptions. IAM allows you to grant roles and permissions to users and service accounts at the project level or the individual resource level. You can also use IAM conditions to specify additional attributes for granting or denying access, such as time, date, or origin. By using IAM conditions, you can ensure that only the resources in project A can access the data in the project A topic, regardless of the network configuration or the VPC Service Controls. You can also prevent project B and any future project from accessing the data in the project A topic by not granting them any roles or permissions on the topic.

Option A is not a good solution, as VPC Service Controls are designed to prevent data exfiltration from Google Cloud resources to the public internet, not to control access between Google Cloud projects. VPC Service Controls create a perimeter around the resources of one or more projects, and restrict the communication with resources outside the perimeter. However, VPC Service Controls do not apply to Pub/Sub, as Pub/Sub is not associated with any specific IP address or VPC network. Therefore, configuring VPC Service Controls with a perimeter around the VPC of project A would not prevent project B or any future project from accessing the data in the project A topic, if they have the necessary IAM roles and permissions.

Option B is not a good solution, as firewall rules are used to control the ingress and egress traffic to and from the VPC network of a project. Firewall rules do not apply to Pub/Sub, as Pub/Sub is not

associated with any specific IP address or VPC network. Therefore, adding firewall rules in project A to only permit traffic from the VPC in project A would not prevent project B or any future project from accessing the data in the project A topic, if they have the necessary IAM roles and permissions.

Option C is not a good solution, as VPC Service Controls are designed to prevent data exfiltration from Google Cloud resources to the public internet, not to control access between Google Cloud projects. VPC Service Controls create a perimeter around the resources of one or more projects, and restrict the communication with resources outside the perimeter. However, VPC Service Controls do not apply to Pub/Sub, as Pub/Sub is not associated with any specific IP address or VPC network. Therefore, configuring VPC Service Controls with a perimeter around project A would not prevent project B or any future project from accessing the data in the project A topic, if they have the necessary IAM roles and permissions. Reference: Access control with IAM | Cloud Pub/Sub Documentation | Google Cloud, [Using IAM Conditions | Cloud IAM Documentation | Google Cloud], [VPC Service Controls overview | Google Cloud], [Using VPC Service Controls | Google Cloud], [Pub/Sub tier capabilities | Memorystore for Redis | Google Cloud].

Question: 296

You are administering a BigQuery dataset that uses a customer-managed encryption key (CMEK). You need to share the dataset with a partner organization that does not have access to your CMEK. What should you do?

A. Create an authorized view that contains the CMEK to decrypt the data when accessed.

B. Provide the partner organization a copy of your CMEKs to decrypt the data.

C. Copy the tables you need to share to a dataset without CMEKs Create an Analytics Hub listing for this dataset.

D. Export the tables to parquet files to a Cloud Storage bucket and grant the storageinsights. viewer role on the bucket to the partner organization.

Answer: C

Explanation:

If you want to share a BigQuery dataset that uses a customer-managed encryption key (CMEK) with a partner organization that does not have access to your CMEK, you cannot use an authorized view or provide them a copy of your CMEK, because these options would violate the security and privacy of

your data. Instead, you can copy the tables you need to share to a dataset without CMEKs, and then create an Analytics Hub listing for this dataset. Analytics Hub is a service that allows you to securely share and discover data assets across your organization and with external partners. By creating an Analytics Hub listing, you can grant the partner organization access to the copied dataset without CMEKs, and also control the level of access and the duration of the sharing.

Reference:

Customer-managed Cloud KMS keys

[Authorized views]

[Analytics Hub overview]

[Creating an Analytics Hub listing]

Question: 297

You are designing a data mesh on Google Cloud with multiple distinct data engineering teams building data products. The

typical data curation design pattern consists of landing files in Cloud Storage, transforming raw data in Cloud Storage and BigQuery datasets. and storing the final curated data product in BigQuery datasets You need to configure Dataplex to ensure that each team can access only the assets needed to build their data products. You also need to ensure that teams can easily share the curated data product. What should you do?

A. 1 Create a single Dataplex virtual lake and create a single zone to contain landing, raw. and curated data.

2 Provide each data engineering team access to the virtual lake.

B. 1 Create a single Dataplex virtual lake and create a single zone to contain landing, raw. and curated data. 2 Build separate assets for each data product within the zone.

3. Assign permissions to the data engineering teams at the zone level.

C. 1 Create a Dataplex virtual lake for each data product, and create a single zone to contain landing, raw, and curated data.

2. Provide the data engineering teams with full access to the virtual lake assigned to their data product.

D. 1 Create a Dataplex virtual lake for each data product, and create multiple zones for landing, raw. and curated data.

2. Provide the data engineering teams with full access to the virtual lake assigned to their data

product.

Answer: D

Explanation:

This option is the best way to configure Dataplex for a data mesh architecture, as it allows each data engineering team to have full ownership and control over their data products, while also enabling easy discovery and sharing of the curated data across the organization12. By creating a Dataplex virtual lake for each data product, you can isolate the data assets and resources for each domain, and avoid conflicts and dependencies between different teams3. By creating multiple zones for landing, raw, and curated data, you can enforce different security and governance policies for each stage of the data curation process, and ensure that only authorized users can access the data assets45. By providing the data engineering teams with full access to the virtual lake assigned to their data product, you can empower them to manage and monitor their data products, and leverage the Dataplex features such as tagging, quality, and lineage.

Option A is not suitable, as it creates a single point of failure and a bottleneck for the data mesh, and does not allow for fine-grained access control and governance for different data products2. Option B is also not suitable, as it does not isolate the data assets and resources for each data product, and assigns permissions at the zone level, which may not reflect the different roles and responsibilities of the data engineering teams34. Option C is better than option A and B,

but it does not create multiple zones for landing, raw, and curated data, which may compromise the security and quality of the data products5. Reference:

1: Building a data mesh on Google Cloud using BigQuery and Dataplex | Google Cloud Blog

2: Data Mesh - 7 Effective Practices to Get Started - Confluent

3: Best practices | Dataplex | Google Cloud

4: Secure your lake | Dataplex | Google Cloud

5: Zones | Dataplex | Google Cloud

6:]: Managing a Data Mesh with Dataplex – ROI Training

Question: 298

You work for an airline and you need to store weather data in a BigQuery table Weather data will be used as input to a machine learning model. The model only uses the last 30 days of weather dat

a. You want to avoid storing unnecessary data and minimize costs. What should you do?

A. Create a BigQuery table where each record has an ingestion timestamp Run a scheduled query to delete all the rows with an ingestion timestamp older than 30 days.

B. Create a BigQuery table partitioned by ingestion time Set up partition expiration to 30 days.

C. Create a BigQuery table partitioned by datetime value of the weather date Set up partition expiration to 30 days.

D. Create a BigQuery table with a datetime column for the day the weather data refers to. Run a scheduled query to delete rows with a datetime value older than 30 days.

Answer: B

Explanation:

Partitioning a table by ingestion time means that the data is divided into partitions based on the time when the data was loaded into the table. This allows you to delete or archive old data by setting a partition expiration policy. You can specify the number of days to keep the data in each partition, and BigQuery automatically deletes the data when it expires. This way, you can avoid storing unnecessary data and minimize costs.

Question: 299

You have terabytes of customer behavioral data streaming from Google Analytics into BigQuery daily Your customers' information, such as their preferences, is hosted on a Cloud SQL for MySQL database Your CRM database is hosted on a Cloud SQL for PostgreSQL instance. The marketing team wants to use your customers' information from the two databases and the customer behavioral data to create marketing campaigns for yearly active customers. You need to ensure that the marketing team can run the campaigns over 100 times a day on typical days and up to 300 during sales. At the same time you want to keep the load on the Cloud SQL databases to a minimum. What should you do?

A. Create BigQuery connections to both Cloud SQL databases Use BigQuery federated queries on the two databases and the Google Analytics data on BigQuery to run these queries.

B. Create streams in Datastream to replicate the required tables from both Cloud SQL databases to BigQuery for these queries.

C. Create a Dataproc cluster with Trino to establish connections to both Cloud SQL databases and BigQuery, to execute the queries.

D. Create a job on Apache Spark with Dataproc Serverless to query both Cloud SQL databases and the Google Analytics data on BigQuery for these queries.

Answer: B

Explanation:

Datastream is a serverless Change Data Capture (CDC) and replication service that allows you to stream data changes from Oracle and MySQL databases to Google Cloud services such as BigQuery, Cloud Storage, Cloud SQL, and Pub/Sub. Datastream captures and delivers database changes in realtime, with minimal impact on the source database performance. Datastream also preserves the schema and data types of the source database, and automatically creates and updates the corresponding tables in BigQuery.

By using Datastream, you can replicate the required tables from both Cloud SQL databases to BigQuery, and keep them in sync with the source databases. This way, you can reduce the load on the Cloud SQL databases, as the marketing team can run their queries on the BigQuery tables instead of the Cloud SQL tables. You can also leverage the scalability and performance of BigQuery to query the customer behavioral data from Google Analytics and the customer information from the replicated tables. You can run the queries as frequently as needed, without worrying about the impact on the Cloud SQL databases.

Option A is not a good solution, as BigQuery federated queries allow you to query external data sources such as Cloud SQL databases, but they do not reduce the load on the source databases. In fact, federated queries may increase the load

on the source databases, as they need to execute the query statements on the external data sources and return the results to BigQuery. Federated queries also have some limitations, such as data type mappings, quotas, and performance issues.

Option C is not a good solution, as creating a Dataproc cluster with Trino would require more resources and management overhead than using Datastream. Trino is a distributed SQL query engine that can connect to multiple data sources, such as Cloud SQL and BigQuery, and execute queries across them. However, Trino requires a Dataproc cluster to run, which means you need to provision, configure, and monitor the cluster nodes. You also need to install and configure the Trino connector for Cloud SQL and BigQuery, and write the queries in Trino SQL dialect. Moreover, Trino does not replicate or sync the data from Cloud SQL to BigQuery, so the load on the Cloud SQL databases would still be high.

Option D is not a good solution, as creating a job on Apache Spark with Dataproc Serverless would require more coding and processing power than using Datastream. Apache Spark is a distributed data

processing framework that can read and write data from various sources, such as Cloud SQL and BigQuery, and perform complex transformations and analytics on them. Dataproc Serverless is a serverless Spark service that allows you to run Spark jobs without managing clusters. However, Spark requires you to write code in Python, Scala, Java, or R, and use the Spark connector for Cloud SQL and BigQuery to access the data sources. Spark also does not replicate or sync the data from Cloud SQL to BigQuery, so the load on the Cloud SQL databases would still be high.
Reference: Datastream overview | Datastream | Google Cloud, Datastream concepts | Datastream | Google Cloud, Datastream quickstart | Datastream | Google Cloud, Introduction to federated queries | BigQuery | Google Cloud, Trino overview | Dataproc Documentation | Google Cloud, Dataproc Serverless overview | Dataproc Documentation | Google Cloud, Apache Spark overview | Dataproc Documentation | Google Cloud.

Question: 300

You are running a streaming pipeline with Dataflow and are using hopping windows to group the data as the data arrives. You noticed that some data is arriving late but is not being marked as late data, which is resulting in inaccurate aggregations downstream. You need to find a solution that allows you to capture the late data in the appropriate window. What should you do?

A. Change your windowing function to session windows to define your windows based on certain activity.

B. Change your windowing function to tumbling windows to avoid overlapping window periods.

C. Expand your hopping window so that the late data has more time to arrive within the grouping.

D. Use watermarks to define the expected data arrival window Allow late data as it arrives.

Answer: D

Explanation:

Watermarks are a way of tracking the progress of time in a streaming pipeline. They are used to determine when a window can be closed and the results emitted. Watermarks can be either eventtime based or processing-time based. Event-time watermarks track the progress of time based on the timestamps of the data elements, while processing-time watermarks track the progress of time based on the system clock. Event-time watermarks are more accurate, but they require the data source to provide reliable timestamps. Processing-time watermarks are simpler, but they can be affected by system delays or backlogs.

By using watermarks, you can define the expected data arrival window for each windowing function. You can also specify how to handle late data, which is data that arrives after the watermark has passed. You can either discard late data, or allow late data and update the results as new data arrives. Allowing late data requires you to use triggers to control when the results are emitted.

In this case, using watermarks and allowing late data is the best solution to capture the late data in the appropriate window. Changing the windowing function to session windows or tumbling windows will not solve the problem of late data, as they still rely on watermarks to determine when to close the windows. Expanding the hopping window might reduce the amount of late data, but it will also change the semantics of the windowing function and the results.

Reference:

Streaming pipelines | Cloud Dataflow | Google Cloud

Windowing | Apache Beam

Question: 301

You work for a large ecommerce company. You are using Pub/Sub to ingest the clickstream data to Google Cloud for analytics. You observe that when a new subscriber connects to an existing topic to analyze data, they are unable to subscribe to older data for an upcoming yearly sale event in two months, you need a solution that, once implemented, will enable any new subscriber to read the last 30 days of dat

a. What should you do?

A. Create a new topic, and publish the last 30 days of data each time a new subscriber connects to an existing topic.

B. Set the topic retention policy to 30 days.

C. Set the subscriber retention policy to 30 days.

D. Ask the source system to re-push the data to Pub/Sub, and subscribe to it.

Answer: B

Explanation:

By setting the topic retention policy to 30 days, you can ensure that any new subscriber can access the messages that were published to the topic within the last 30 days1. This feature allows you to replay previously acknowledged messages or initialize new subscribers with historical data2. You can configure the topic retention policy by using the Cloud Console, the gcloud command-line tool, or the Pub/Sub API1.

Option A is not efficient, as it requires creating a new topic and duplicating the data for each new subscriber, which would increase the storage costs and complexity. Option C is not effective, as it only affects the unacknowledged messages in a subscription, and does not allow new subscribers to access older data3. Option D is not feasible, as it depends on the source system's ability and willingness to re-push the data, and it may cause data duplication or inconsistency. Reference:

1: Create a topic | Cloud Pub/Sub Documentation | Google Cloud

2: Replay and purge messages with seek | Cloud Pub/Sub Documentation | Google Cloud

3: When is a PubSub Subscription considered to be inactive?

Question: 302

You use a dataset in BigQuery for analysis. You want to provide third-party companies with access to the same dataset. You need to keep the costs of data sharing low and ensure that the data is current. What should you do?

A. Use Analytics Hub to control data access, and provide third party companies with access to the dataset

B. Create a Dataflow job that reads the data in frequent time intervals and writes it to the relevant BigQuery dataset or Cloud Storage bucket for third-party companies to use.

C. Use Cloud Scheduler to export the data on a regular basis to Cloud Storage, and provide third- party companies with access to the bucket.

D. Create a separate dataset in BigQuery that contains the relevant data to share, and provide third- party companies with access to the new dataset.

Answer: A

Explanation:

Analytics Hub is a service that allows you to securely share and discover data assets across your organization and with external partners. You can use Analytics Hub to create and manage data assets, such as BigQuery datasets, views, and queries, and control who can access them. You can also browse and use data assets that others have shared with you. By

using Analytics Hub, you can keep the costs of data sharing low and ensure that the data is current, as the data assets are not copied or moved, but rather referenced from their original sources.

Question: 303

You are troubleshooting your Dataflow pipeline that processes data from Cloud Storage to BigQuery.
You have discovered that the Dataflow worker nodes cannot communicate with one another Your networking team relies on Google Cloud network tags to define firewall rules You need to identify the issue while following Google-recommended networking security practices. What should you do?

A. Determine whether your Dataflow pipeline has a custom network tag set.

B. Determine whether there is a firewall rule set to allow traffic on TCP ports 12345 and 12346 for the Dataflow network tag.

C. Determine whether your Dataflow pipeline is deployed with the external IP address option enabled.

D. Determine whether there is a firewall rule set to allow traffic on TCP ports 12345 and 12346 on the subnet used by Dataflow workers.

Answer: D

Explanation:

Dataflow worker nodes need to communicate with each other and with the Dataflow service on TCP ports 12345 and 12346. These ports are used for data shufling and streaming engine communication. By default, Dataflow assigns a network tag called dataflow to the worker nodes, and creates a firewall rule that allows traffic on these ports for the dataflow network tag. However, if you use a custom network tag for your Dataflow pipeline, you need to create a firewall rule that allows traffic on these ports for your custom network tag. Otherwise, the worker nodes will not be able to communicate with each other and the Dataflow service, and the pipeline will fail.

Therefore, the best way to identify the issue is to determine whether there is a firewall rule set to allow traffic on TCP ports 12345 and 12346 for the Dataflow network tag. If there is no such firewall

rule, or if the firewall rule does not match the network tag used by your Dataflow pipeline, you need to create or update the firewall rule accordingly.

Option A is not a good solution, as determining whether your Dataflow pipeline has a custom network tag set does not tell you whether there is a firewall rule that allows traffic on the required ports for that network tag. You need to

check the firewall rule as well.

Option C is not a good solution, as determining whether your Dataflow pipeline is deployed with the external IP address option enabled does not tell you whether there is a firewall rule that allows traffic on the required ports for the Dataflow network tag. The external IP address option determines whether the worker nodes can access resources on the public internet, but it does not affect the internal communication between the worker nodes and the Dataflow service.

Option D is not a good solution, as determining whether there is a firewall rule set to allow traffic on TCP ports 12345 and 12346 on the subnet used by Dataflow workers does not tell you whether the firewall rule applies to the Dataflow network tag. The firewall rule should be based on the network tag, not the subnet, as the network tag is more specific and secure. Reference: Dataflow network tags | Cloud Dataflow | Google Cloud, Dataflow firewall rules | Cloud Dataflow | Google Cloud, Dataflow network configuration | Cloud Dataflow | Google Cloud, Dataflow Streaming Engine | Cloud Dataflow | Google Cloud.

Question: 304

You are configuring networking for a Dataflow job. The data pipeline uses custom container images with the libraries that are required for the transformation logic preinstalled. The data pipeline reads the data from Cloud Storage and writes the data to BigQuery. You need to ensure cost-effective and secure communication between the pipeline and Google APIs and services. What should you do?

A. Leave external IP addresses assigned to worker VMs while enforcing firewall rules.

B. Disable external IP addresses and establish a Private Service Connect endpoint IP address.

C. Disable external IP addresses from worker VMs and enable Private Google Access.

D. Enable Cloud NAT to provide outbound internet connectivity while enforcing firewall rules.

Answer: C

Explanation:

Private Google Access allows VMs without external IP addresses to communicate with Google APIs and services over internal routes. This reduces the cost and increases the security of the data pipeline. Custom container images can be stored in Container Registry, which supports Private Google Access. Dataflow supports Private Google Access for both batch and streaming jobs. Reference:

Private Google Access overview

Using Private Google Access and Cloud NAT

Using custom containers with Dataflow

Question: 305

You work for a large ecommerce company. You store your customers order data in Bigtable. You have a garbage collection policy set to delete the data after 30 days and the number of versions is set to 1. When the data analysts run a query to report total customer spending, the analysts sometimes see customer data that is older than 30 days. You need to ensure that the analysts do not see customer data older than 30 days while minimizing cost and overhead. What should you do?

A. Set the expiring values of the column families to 30 days and set the number of versions to 2.

B. Use a timestamp range filter in the query to fetch the customer's data for a specific range.

C. Set the expiring values of the column families to 29 days and keep the number of versions to 1.

D. Schedule a job daily to scan the data in the table and delete data older than 30 days.

Answer: B

Explanation:

By using a timestamp range filter in the query, you can ensure that the analysts only see the customer data that is within the desired time range, regardless of the garbage collection policy1. This option is the most cost-effective and simple way to avoid fetching data that is marked for deletion by garbage collection, as it does not require changing the existing policy or creating additional jobs. You can use the Bigtable client libraries or the cbt CLI to apply a timestamp range filter to your read requests2.

Option A is not effective, as it increases the number of versions to 2, which may cause more data to

be retained and increase the storage costs. Option C is not reliable, as it reduces the expiring values to 29 days, which may not match the actual data arrival and usage patterns. Option D is not efficient, as it requires scheduling a job daily to scan and delete the data, which may incur additional overhead and complexity. Moreover, none of these options guarantee that the data older than 30 days will be immediately deleted, as garbage collection is an asynchronous process that can take up to a week to remove the data3. Reference:

1: Filters | Cloud Bigtable Documentation | Google Cloud

2: Read data | Cloud Bigtable Documentation | Google Cloud

3: Garbage collection overview | Cloud Bigtable Documentation | Google Cloud

Question: 306

You have 100 GB of data stored in a BigQuery table. This data is outdated and will only be accessed one or two times a year for analytics with SQL. For backup purposes, you want to store this data to be immutable for 3 years. You want to minimize storage costs. What should you do?

A. 1 Create a BigQuery table clone.

2. Query the clone when you need to perform analytics.

B. 1 Create a BigQuery table snapshot.

2 Restore the snapshot when you need to perform analytics.

C. 1. Perform a BigQuery export to a Cloud Storage bucket with archive storage class.

2 Enable versionmg on the bucket.

3 . Create a BigQuery external table on the exported files.

D. 1 Perform a BigQuery export to a Cloud Storage bucket with archive storage class.

2 Set a locked retention policy on the bucket.

3 . Create a BigQuery external table on the exported files.

Answer: D

Explanation:

This option will allow you to store the data in a low-cost storage option, as the archive storage class has the lowest price per GB among the Cloud Storage classes. It will also ensure that the data is immutable for 3 years, as the locked retention policy prevents the deletion or overwriting of the data until the retention period expires. You can still query the data using SQL by creating a BigQuery external table that references the exported files in the Cloud Storage bucket. Option A is incorrect because creating a BigQuery table clone will not reduce the storage costs, as the clone will have the same size and storage class as the original table. Option B is incorrect because creating a BigQuery table snapshot will also not reduce the storage costs, as the snapshot will have the same size and storage class as the original table. Option C is incorrect because enabling versioning on the bucket will not make the data immutable, as the versions can still be deleted or overwritten by anyone with the appropriate permissions. It will also increase the storage costs, as each version of the file will be charged separately. Reference:

Exporting table data | BigQuery | Google Cloud

Storage classes | Cloud Storage | Google Cloud

Retention policies and retention periods | Cloud Storage | Google Cloud

Federated queries | BigQuery | Google Cloud

Question: 307

You have designed an Apache Beam processing pipeline that reads from a Pub/Sub topic. The topic has a message retention duration of one day, and writes to a Cloud Storage bucket. You need to select a bucket location and processing strategy to prevent data loss in case of a regional outage with an RPO of 15 minutes. What should you do?

A. 1 Use a regional Cloud Storage bucket

2 Monitor Dataflow metrics with Cloud Monitoring to determine when an outage occurs

3 Seek the subscription back in time by one day to recover the acknowledged messages

4 Start the Dataflow job in a secondary region and write in a bucket in the same region

B. 1 Use a multi-regional Cloud Storage bucket

2 Monitor Dataflow metrics with Cloud Monitoring to determine when an outage occurs

3 Seek the subscription back in time by 60 minutes to recover the acknowledged messages

4 Start the Dataflow job in a secondary region

C. 1. Use a dual-region Cloud Storage bucket.

2 . Monitor Dataflow metrics with Cloud Monitoring to determine when an outage occurs

3 Seek the subscription back in time by 15 minutes to recover the acknowledged messages

4 Start the Dataflow job in a secondary region

D. 1. Use a dual-region Cloud Storage bucket with turbo replication enabled

2 Monitor Dataflow metrics with Cloud Monitoring to determine when an outage occurs

3 Seek the subscription back in time by 60 minutes to recover the acknowledged messages

4 Start the Dataflow job in a secondary region.

Answer: C

Explanation:

A dual-region Cloud Storage bucket is a type of bucket that stores data redundantly across two regions within the same continent. This provides higher availability and durability than a regional bucket, which stores data in a single region. A dual-region bucket also provides lower latency and higher throughput than a multi-regional bucket, which stores data across multiple regions within a continent or across continents. A dual-region bucket with turbo replication enabled is a premium option that offers even faster replication across regions, but it is more expensive and not necessary for this scenario.

By using a dual-region Cloud Storage bucket, you can ensure that your data is protected from regional outages, and that you can access it from either region with low latency and high performance. You can also monitor the Dataflow metrics with Cloud Monitoring to determine when an outage occurs, and seek the subscription back in time by 15 minutes to recover the acknowledged messages. Seeking a subscription allows you to replay the messages from a Pub/Sub topic that were published within the message retention duration, which is one day in this case. By seeking the subscription back in time by 15 minutes, you can meet the RPO of 15 minutes, which means the maximum amount of data loss that is acceptable for your business. You can then start the Dataflow job in a secondary region and write to the same dual-region bucket, which will resume the processing of the messages and prevent data loss.

Option A is not a good solution, as using a regional Cloud Storage bucket does not provide any redundancy or protection from regional outages. If the region where the bucket is located experiences an outage, you will not be able to access your data or write new data to the bucket. Seeking the subscription back in time by one day is also unnecessary and inefficient, as it will replay all the messages from the past day, even though you only need to recover the messages from the past 15 minutes.

Option B is not a good solution, as using a multi-regional Cloud Storage bucket does not provide the best performance or cost-efficiency for this scenario. A multi-regional bucket stores data across multiple regions within a continent or across continents, which provides higher availability and durability than a dual-region bucket, but also higher latency and lower throughput. A multi-regional bucket is more suitable for serving data to a global audience, not for processing data with Dataflow within a single continent. Seeking the subscription back in time by 60 minutes is also unnecessary and inefficient, as it will replay more messages than needed to meet the RPO of 15 minutes.

Option D is not a good solution, as using a dual-region Cloud Storage bucket with turbo replication enabled does not provide any additional benefit for this scenario, but only increases the cost. Turbo replication is a premium option that offers faster replication across regions, but it is not required to meet the RPO of 15 minutes. Seeking the subscription back in time by 60 minutes is also unnecessary and inefficient, as it will replay more messages than needed to meet the RPO of 15 minutes. Reference: Storage locations | Cloud Storage | Google Cloud, Dataflow metrics | Cloud Dataflow | Google Cloud, Seeking a subscription | Cloud Pub/Sub | Google Cloud, Recovery point objective (RPO) | Acronis.

Question: 308

You need to look at BigQuery data from a specific table multiple times a day. The underlying table you are querying is several petabytes in size, but you want to filter your data and provide simple aggregations to downstream users. You want to run queries faster and get up-to-date insights quicker. What should you do?

A. Run a scheduled query to pull the necessary data at specific intervals daily.

B. Create a materialized view based off of the query being run.

C. Use a cached query to accelerate time to results.

D. Limit the query columns being pulled in the final result.

Answer: A

Explanation:

Materialized views are precomputed views that periodically cache the results of a query for increased performance and efficiency. BigQuery leverages precomputed results from materialized views and whenever possible reads only changes from the base tables to compute up-to-date results. Materialized views can significantly improve the performance of workloads that have the characteristic of common and repeated queries. Materialized views can also optimize queries with high computation cost and small dataset results, such as filtering and aggregating large tables. Materialized views are refreshed automatically when the base tables change, so they always return fresh data. Materialized views can also be used by the BigQuery optimizer to process queries to the base tables, if any part of the query can be resolved by querying the materialized view. Reference:

Introduction to materialized views

Create materialized views

BigQuery Materialized View Simplified: Steps to Create and 3 Best Practices

Materialized view in Bigquery

Question: 309

You have data located in BigQuery that is used to generate reports for your company. You have noticed some weekly executive report fields do not correspond to format according to company standards for example, report errors include

different telephone formats and different country code identifiers. This is a frequent issue, so you need to create a recurring job to normalize the dat

a. You want a quick solution that requires no coding What should you do?

A. Use Cloud Data Fusion and Wrangler to normalize the data, and set up a recurring job.

B. Use BigQuery and GoogleSQL to normalize the data, and schedule recurring quenes in BigQuery.

C. Create a Spark job and submit it to Dataproc Serverless.

D. Use Dataflow SQL to create a job that normalizes the data, and that after the first run of the job, schedule the pipeline to execute recurrently.

Answer: A

Explanation:

Cloud Data Fusion is a fully managed, cloud-native data integration service that allows you to build and manage data pipelines with a graphical interface. Wrangler is a feature of Cloud Data Fusion that enables you to interactively explore, clean, and transform data using a spreadsheet-like UI. You can use Wrangler to normalize the data in BigQuery by applying various directives, such as parsing, formatting, replacing, and validating data. You can also preview the results and export the wrangled data to BigQuery or other destinations. You can then set up a recurring job in Cloud Data Fusion to run the Wrangler pipeline on a schedule, such as weekly or daily. This way, you can create a quick and code-free solution to normalize the data for your reports. Reference:

Cloud Data Fusion overview

Wrangler overview

Wrangle data from BigQuery

[Scheduling pipelines]

Question: 310

You are migrating a large number of files from a public HTTPS endpoint to Cloud Storage. The files are protected from unauthorized access using signed URLs. You created a TSV file that contains the list of object URLs and started a transfer job by using Storage Transfer Service. You notice that the job has run for a long time and eventually failed Checking the logs of the transfer job reveals that the job was running fine until one point, and then it failed due to HTTP 403 errors on the remaining files You verified that there were no changes to the source system You need to fix the problem to resume the migration process. What should you do?

A. Set up Cloud Storage FUSE, and mount the Cloud Storage bucket on a Compute Engine Instance Remove the completed files from the TSV file Use a shell script to iterate through the TSV file and download the remaining URLs to the FUSE mount point.

B. Update the file checksums in the TSV file from using MD5 to SHA256. Remove the completed files from the TSV file and rerun the Storage Transfer Service job.

C. Renew the TLS certificate of the HTTPS endpoint Remove the completed files from the TSV file and rerun the Storage Transfer Service job.

D. Create a new TSV file for the remaining files by generating signed URLs with a longer validity period. Split the TSV file into multiple smaller files and submit them as separate Storage Transfer Service jobs in parallel.

Answer: D

Explanation:

A signed URL is a URL that provides limited permission and time to access a resource on a web server. It is often used to grant temporary access to protected files without requiring authentication. Storage Transfer Service is a service that allows you to transfer data from external sources, such as HTTPS endpoints, to Cloud Storage buckets. You can use a TSV file to specify the list of URLs to transfer. In this scenario, the most likely cause of the HTTP 403 errors is that the signed URLs have expired before the transfer job could complete. This could happen if the signed URLs have a short validity period or the transfer job takes a long time due to the large number of files or network latency. To fix the problem, you need to create a new TSV file for the remaining files by generating new signed URLs with a longer validity period. This will ensure that the URLs do not expire before the transfer job finishes. You can use the Cloud Storage tools or your own program to generate signed URLs. Additionally, you can split the TSV file into multiple smaller files and submit them as separate Storage Transfer Service jobs in parallel. This will speed up the transfer process and reduce the risk of errors. Reference:

Signed URLs | Cloud Storage Documentation

V4 signing process with Cloud Storage tools

V4 signing process with your own program

Using a URL list file

What Is a 403 Forbidden Error (and How Can I Fix It)?

Question: 311

You want to store your team's shared tables in a single dataset to make data easily accessible to various analysts. You want to make this data readable but unmodifiable by analysts. At the same time, you want to provide the analysts with

individual workspaces in the same project, where they can create and store tables for their own use, without the tables being accessible by other analysts. What should you do?

A. Give analysts the BigQuery Data Viewer role at the project level Create one other dataset, and give the analysts the BigQuery Data Editor role on that dataset.

B. Give analysts the BigQuery Data Viewer role at the project level Create a dataset for each analyst, and give each analyst the BigQuery Data Editor role at the project level.

C. Give analysts the BigQuery Data Viewer role on the shared dataset. Create a dataset for each analyst, and give each analyst the BigQuery Data Editor role at the dataset level for their assigned dataset

D. Give analysts the BigQuery Data Viewer role on the shared dataset Create one other dataset and give the analysts the BigQuery Data Editor role on that dataset.

Answer: C

Explanation:

The BigQuery Data Viewer role allows users to read data and metadata from tables and views, but not to modify or delete them. By giving analysts this role on the shared dataset, you can ensure that they can access the data for analysis, but not change it. The BigQuery Data Editor role allows users to create, update, and delete tables and views, as well as read and write data. By giving analysts this role at the dataset level for their assigned dataset, you can provide them with individual workspaces where they can store their own tables and views, without affecting the shared dataset or other analysts' datasets. This way, you can achieve both data protection and data isolation for your team. Reference:

BigQuery IAM roles and permissions
Basic roles and permissions

Question: 312

Your company's data platform ingests CSV file dumps of booking and user profile data from upstream sources into Cloud Storage. The data analyst team wants to join these datasets on the email field available in both the datasets to perform analysis. However, personally identifiable information (PII) should not be accessible to the analysts. You need to de-identify the email field in both the datasets before loading them into BigQuery for analysts. What should you do?

A. 1. Create a pipeline to de-identify the email field by using recordTransformations in Cloud Data Loss Prevention (Cloud DLP) with masking as the de-identification transformations type.

2. Load the booking and user profile data into a BigQuery table.

B. 1. Create a pipeline to de-identify the email field by using recordTransformations in Cloud DLP with format-preserving encryption with FFX as the de-identification transformation type.

2. Load the booking and user profile data into a BigQuery table.

C. 1. Load the CSV files from Cloud Storage into a BigQuery table, and enable dynamic data masking.

2. Create a policy tag with the email mask as the data masking rule.

3. Assign the policy to the email field in both tables. A

4. Assign the Identity and Access Management bigquerydatapolicy.maskedReader role for the BigQuery tables to the analysts.

D. 1. Load the CSV files from Cloud Storage into a BigQuery table, and enable dynamic data masking.

2. Create a policy tag with the default masking value as the data masking rule.

3. Assign the policy to the email field in both tables.

4. Assign the Identity and Access Management bigquerydatapolicy.maskedReader role for the BigQuery tables to the analysts

Answer: B

Explanation:

Cloud DLP is a service that helps you discover, classify, and protect your sensitive data. It supports various de-identification techniques, such as masking, redaction, tokenization, and encryption. Format-preserving encryption (FPE) with FFX is a technique that encrypts sensitive data while preserving its original format and length. This allows you to join the encrypted data on the same field without revealing the actual values. FPE with FFX also supports partial encryption, which means you can encrypt only a portion of the data, such as the domain name of an email address. By using Cloud DLP to de-identify the email field with FPE with FFX, you can ensure that the analysts can join the booking and user profile data on the email field without accessing the PII. You can create a pipeline to de-identify the email field by using recordTransformations in Cloud DLP, which allows you to specify the fields and the de-identification transformations to apply to them. You can then load the de-identified data into a BigQuery table for analysis. Reference:

De-identify sensitive data | Cloud Data Loss Prevention Documentation

Format-preserving encryption with FFX | Cloud Data Loss Prevention Documentation

De-identify and re-identify data with the Cloud DLP API

De-identify data in a pipeline

Question: 313

You are creating a data model in BigQuery that will hold retail transaction dat

a. Your two largest tables, sales_transation_header and sales_transation_line. have a tightly coupled immutable relationship. These tables are rarely modified after load and are frequently joined when queried. You need to model the sales_transation_header and sales_transation_line tables to improve the performance of data analytics queries. What should you do?

A. Create a sal es_transaction table that Stores the sales_tran3action_header and sales_transaction_line data as a JSON data type.

B. Create a sale3_transaction table that holds the sales_transaction_header information as rows and the sales_transaction_line rows as nested and repeated fields.

C. Create a sale_transaction table that holds the sales_transaction_header and sales_transaction_line information as rows, duplicating the sales_transaction_header data for each line.

D. Create separate sales_transation_header and sales_transation_line tables and. when querying, specify the sales transition line first in the WHERE clause.

Answer: B

Explanation:

BigQuery supports nested and repeated fields, which are complex data types that can represent hierarchical and one-to-many relationships within a single table. By using nested and repeated fields, you can denormalize your data model and reduce the number of joins required for your queries. This can improve the performance and efficiency of your data analytics queries, as joins can be expensive and require shufling data across nodes. Nested and repeated fields also preserve the data integrity and avoid data duplication. In this scenario, the sales_transaction_header and sales_transaction_line tables have a tightly coupled immutable relationship, meaning that each header row corresponds to one or more line rows, and the data is rarely modified after load. Therefore, it makes sense to create a single sales_transaction table that holds the sales_transaction_header information as rows and the sales_transaction_line rows as nested and repeated fields. This way, you can query the sales transaction data without joining two tables, and use dot notation or array functions to access the

nested and repeated fields. For example, the sales_transaction table could have the following schema:

Table

Field name	Type	Mode
id	INTEGER	NULLABLE
order_time	TIMESTAMP	NULLABLE
customer_id	INTEGER	NULLABLE
line_items	RECORD	REPEATED
line_items.sku	STRING	NULLABLE
line_items.quantity	INTEGER	NULLABLE
line_items.price	FLOAT	NULLABLE

To query the total amount of each order, you could use the following SQL statement:

SQL

```
SELECT id, SUM(line_items.quantity * line_items.price) AS total_amount
FROM sales_transaction
GROUP BY id;
```

AI-generated code. Review and use carefully. More info on FAQ.

Reference:

Use nested and repeated fields

BigQuery explained: Working with joins, nested & repeated data

Arrays in BigQuery — How to improve query performance and optimise storage

Question: 314

You have created an external table for Apache Hive partitioned data that resides in a Cloud Storage bucket, which contains a large number of files. You notice that queries against this table are slow. You want to improve the performance of these queries What should you do?

A. Migrate the Hive partitioned data objects to a multi-region Cloud Storage bucket.

B. Create an individual external table for each Hive partition by using a common table name prefix Use wildcard table queries to reference the partitioned data.

C. Change the storage class of the Hive partitioned data objects from Coldline to Standard.

D. Upgrade the external table to a BigLake table Enable metadata caching for the table.

Answer: D

Explanation:

BigLake is a Google Cloud service that allows you to query structured data in external data stores such as Cloud Storage, Amazon S3, and Azure Blob Storage with access delegation and governance. BigLake tables extend the capabilities of BigQuery to data lakes and enable a flexible, open lakehouse architecture. By upgrading an external table to a BigLake table, you can improve the performance of your queries by leveraging the BigQuery storage API, which supports data format conversion, predicate pushdown, column projection, and metadata caching. Metadata caching reduces the number of requests to the external data store and speeds up query execution. To upgrade an external table to a BigLake table, you can use the ALTER TABLE statement with the SET OPTIONS clause and specify the enable_metadata_caching option as true. For example:

SQL

```
ALTER TABLE hive_partitioned_data
SET OPTIONS (
 enable_metadata_caching=true
);
```

AI-generated code. Review and use carefully. More info on FAQ.

Reference:

Introduction to BigLake tables

Upgrade an external table to BigLake

BigQuery storage API

Question: 315

Your chemical company needs to manually check documentation for customer order. You use a pull subscription in Pub/Sub so that sales agents get details from the order. You must ensure that you do not process orders twice with different sales agents and that you do not add more complexity to this workflow. What should you do?

A. Create a transactional database that monitors the pending messages.

B. Create a new Pub/Sub push subscription to monitor the orders processed in the agent's system.

C. Use Pub/Sub exactly-once delivery in your pull subscription.

D. Use a Deduphcate PTransform in Dataflow before sending the messages to the sales agents.

Answer: C

Explanation:

Pub/Sub exactly-once delivery is a feature that guarantees that subscriptions do not receive duplicate deliveries of messages based on a Pub/Sub-defined unique message ID. This feature is only supported by the pull subscription type, which is what you are using in this scenario. By enabling exactly-once delivery, you can ensure that each order is processed only once by a sales agent, and that no order is lost or duplicated. This also simplifies your workflow, as you do not need to create a separate database or subscription to monitor the pending or processed messages. Reference:

Exactly-once delivery | Cloud Pub/Sub Documentation
Cloud Pub/Sub Exactly-once Delivery feature is now Generally Available (GA)

Question: 316

You are part of a healthcare organization where data is organized and managed by respective data owners in various storage services. As a result of this decentralized ecosystem, discovering and managing data has become difficult You need to quickly identify and implement a cost-optimized solution to assist your organization with the following

- Data management and discovery
- Data lineage tracking
- Data quality validation

How should you build the solution?

A. Use BigLake to convert the current solution into a data lake architecture.

B. Build a new data discovery tool on Google Kubernetes Engine that helps with new source onboarding and data lineage tracking.

C. Use BigQuery to track data lineage, and use Dataprep to manage data and perform data quality validation.

D. Use Dataplex to manage data, track data lineage, and perform data quality validation.

Answer: D

Explanation:

Dataplex is a Google Cloud service that provides a unified data fabric for data lakes and data warehouses. It enables data governance, management, and discovery across multiple data domains, zones, and assets. Dataplex also supports data lineage tracking, which shows the origin and transformation of data over time. Dataplex also integrates with Dataprep, a data preparation and quality tool that allows users to clean, enrich, and transform data using a visual interface. Dataprep can also monitor data quality and detect anomalies using machine learning. Therefore, Dataplex is the most suitable solution for the given scenario, as it meets all the requirements of data management and discovery, data lineage tracking, and data quality validation. Reference:

Dataplex overview

Automate data governance, extend your data fabric with Dataplex-BigLake integration

Dataprep documentation

Question: 317

Your team is building a data lake platform on Google Cloud. As a part of the data foundation design, you are planning to store all the raw data in Cloud Storage You are expecting to ingest approximately 25 GB of data a day and your billing department is worried about the increasing cost of storing old

dat

a. The current business requirements are:

- The old data can be deleted anytime
- You plan to use the visualization layer for current and historical reporting
- The old data should be available instantly when accessed
- There should not be any charges for data retrieval.

What should you do to optimize for cost?

A. Create the bucket with the Autoclass storage class feature.

B. Create an Object Lifecycle Management policy to modify the storage class for data older than 30 days to nearline, 90 days to coldline. and 365 days to archive storage class. Delete old data as needed.

C. Create an Object Lifecycle Management policy to modify the storage class for data older than 30 days to coldline, 90 days to nearline. and 365 days to archive storage class Delete old data as needed.

D. Create an Object Lifecycle Management policy to modify the storage class for data older than 30 days to nearlme. 45 days to coldline. and 60 days to archive storage class Delete old data as needed.

Answer: B

Explanation:

- Autoclass automatically moves objects between storage classes without impacting performance or availability, nor incurring retrieval costs. - It continuously optimizes storage costs based on access patterns without the need to set specific lifecycle management policies.

Question: 318

You are designing a data warehouse in BigQuery to analyze sales data for a telecommunication service provider. You need to create a data model for customers, products, and subscriptions All customers, products, and subscriptions can be updated monthly, but you must maintain a historical record of all dat

a. You plan to use the visualization layer for current and historical reporting. You need to ensure that

the data model is simple, easy-to-use. and cost-effective. What should you do?

A. Create a normalized model with tables for each entity. Use snapshots before updates to track historical data

B. Create a normalized model with tables for each entity. Keep all input files in a Cloud Storage bucket to track historical data

C. Create a denormalized model with nested and repeated fields Update the table and use snapshots to track historical data

D. Create a denormalized, append-only model with nested and repeated fields Use the ingestion timestamp to track historical data.

Answer: D

Explanation:

- A denormalized, append-only model simplifies query complexity by eliminating the need for joins. - Adding data with an

ingestion timestamp allows for easy retrieval of both current and historical states. - Instead of updating records, new records are appended, which maintains historical information without the need to create separate snapshots.

Question: 319

You work for a large real estate firm and are preparing 6 TB of home sales data lo be used for machine learning You will use SOL to transform the data and use BigQuery ML lo create a machine learning model. You plan to use the model for predictions against a raw dataset that has not been transformed. How should you set up your workflow in order to prevent skew at prediction time?

A. When creating your model, use BigQuerys TRANSFORM clause to define preprocessing stops. At prediction time, use BigQuery"s ML. EVALUATE clause without specifying any transformations on the raw input data.

B. When creating your model, use BigQuery's TRANSFORM clause to define preprocessing steps Before requesting predictions, use a saved query to transform your raw input data, and then use ML. EVALUATE

C. Use a BigOuery to define your preprocessing logic. When creating your model, use the view as

your model training data. At prediction lime, use BigQuery's ML EVALUATE clause without specifying any transformations on the raw input data.

D. Preprocess all data using Dataflow. At prediction time, use BigOuery"s ML. EVALUATE clause without specifying any further transformations on the input data.

Answer: A

Explanation:

https://cloud.google.com/bigquery-ml/docs/bigqueryml-transform Using the TRANSFORM clause, you can specify all preprocessing during model creation. The preprocessing is automatically applied during the prediction and evaluation phases of machine learning

Question: 320

You have a data pipeline with a Dataflow job that aggregates and writes time series metrics to Bigtable. You notice that

data is slow to update in Bigtable. This data feeds a dashboard used by thousands of users across the organization. You need to support additional concurrent users and reduce the amount of time required to write the dat

a. What should you do?

Choose 2 answers

A. Configure your Dataflow pipeline to use local execution.

B. Modify your Dataflow pipeline lo use the Flatten transform before writing to Bigtable.

C. Modify your Dataflow pipeline to use the CoGrcupByKey transform before writing to Bigtable.

D. Increase the maximum number of Dataflow workers by setting maxNumWorkers in PipelineOptions.

E. Increase the number of nodes in the Bigtable cluster.

Answer: D, E

Explanation:

https://cloud.google.com/bigtable/docs/performance#performance-write-throughput

https://cloud.google.com/dataflow/docs/reference/pipeline-options

Question: 321

Your startup has a web application that currently serves customers out of a single region in Asi

a. You are targeting funding that will allow your startup lo serve customers globally. Your current goal is to optimize for cost, and your post-funding goat is to optimize for global presence and performance. You must use a native JDBC driver. What should you do?

A. Use Cloud Spanner to configure a single region instance initially. and then configure multi-region C oud Spanner instances after securing funding.

B. Use a Cloud SQL for PostgreSQL highly available instance first, and 8»gtable with US. Europe, and Asia replication alter securing funding

C. Use a Cloud SQL for PostgreSQL zonal instance first and Bigtable with US. Europe, and Asia after securing funding.

D. Use a Cloud SOL for PostgreSQL zonal instance first, and Cloud SOL for PostgreSQL with highly available configuration after securing funding.

Answer: A

Explanation:

https://cloud.google.com/spanner/docs/instance-configurations#tradeoffs_regional_versus_multi-region_configurations

Question: 322

You are designing a system that requires an ACID-compliant database. You must ensure that the system requires minimal human intervention in case of a failure. What should you do?

A. Configure a Cloud SQL for MySQL instance with point-in-time recovery enabled.

B. Configure a Cloud SQL for PostgreSQL instance with high availability enabled.

C. Configure a Bigtable instance with more than one cluster.

D. Configure a BJgQuery table with a multi-region configuration.

Answer: B

Explanation:

The best option to meet the ACID compliance and minimal human intervention requirements is to configure a Cloud SQL for PostgreSQL instance with high availability enabled. Key reasons: Cloud SQL for PostgreSQL provides full ACID compliance, unlike Bigtable which provides only atomicity and consistency guarantees. Enabling high availability removes the need for manual failover as Cloud SQL will automatically failover to a standby replica if the leader instance goes down. Point-in-time recovery in MySQL requires manual intervention to restore data if needed. BigQuery does not provide transactional guarantees required for an ACID database. Therefore, a Cloud SQL for PostgreSQL instance with high availability meets the ACID and minimal intervention requirements best. The automatic failover will ensure availability and uptime without administrative effort.

Question: 323

You need to migrate a Redis database from an on-premises data center to a Memorystore for Redis instance. You want to follow Google-recommended practices and perform the migration for minimal cost. time, and effort. What should

you do?

A. Make a secondary instance of the Redis database on a Compute Engine instance, and then perform a live cutover.

B. Write a shell script to migrate the Redis data, and create a new Memorystore for Redis instance.

C. Create a Dataflow job to road the Redis database from the on-premises data center. and write the data to a Memorystore for Redis instance

D. Make an RDB backup of the Redis database, use the gsutil utility to copy the RDB file into a Cloud Storage bucket, and then import the RDB tile into the Memorystore for Redis instance.

Answer: D

Explanation:

The import and export feature uses the native RDB snapshot feature of Redis to import data into or export data out of a Memorystore for Redis instance. The use of the native RDB format prevents lockin and makes it very easy to move data within Google Cloud or outside of Google Cloud. Import and export uses Cloud Storage buckets to store RDB files. Reference: https://cloud.google.com/memorystore/docs/redis/import-export-overview

Question: 324

You want to create a machine learning model using BigQuery ML and create an endpoint foe hosting the model using Vertex Al. This will enable the processing of continuous streaming data in near-real time from multiple vendors. The data may contain invalid values. What should you do?

A. Create a new BigOuery dataset and use streaming inserts to land the data from multiple vendors. Configure your BigQuery ML model to use the "ingestion' dataset as the training data.

B. Use BigQuery streaming inserts to land the data from multiple vendors whore your BigQuery dataset ML model is deployed.

C. Create a Pub'Sub topic and send all vendor data to it Connect a Cloud Function to the topic to process the data and store it in BigQuery.

D. Create a Pub/Sub topic and send all vendor data to it Use Dataflow to process and sanitize the Pub/Sub data and stream it to BigQuery.

Answer: D

Explanation:

Dataflow provides a scalable and flexible way to process and clean the incoming data in real-time before loading it into BigQuery.

Question: 325

You have a data processing application that runs on Google Kubernetes Engine (GKE). Containers need to be launched with their latest available configurations from a container registry. Your GKE nodes need to have GPUs. local SSDs, and 8 Gbps bandwidth. You want to efficiently provision the data processing infrastructure and manage the deployment process. What should you do?

A. Use Compute Engi.no startup scriots to pull container Images, and use gloud commands to provision the infrastructure.

B. Use GKE to autoscale containers, and use gloud commands to provision the infrastructure.

C. Use Cloud Build to schedule a job using Terraform build to provision the infrastructure and launch with the most current container images.

D. Use Dataflow to provision the data pipeline, and use Cloud Scheduler to run the job.

Answer: C

Explanation:

https://cloud.google.com/architecture/managing-infrastructure-as-code

Question: 326

You issue a new batch job to Dataflow. The job starts successfully, processes a few elements, and then suddenly fails and shuts down. You navigate to the Dataflow monitoring interface where you find errors related to a particular DoFn in your pipeline. What is the most likely cause of the errors?

A. Exceptions in worker code

B. Job validation

C. Graph or pipeline construction

D. Insufficient permissions

Answer: A

Explanation:

https://cloud.google.com/dataflow/docs/guides/troubleshooting-your- pipeline#detect_an_exception_in_worker_code
While your job is running, you might encounter errors or exceptions in your worker code. These errors generally mean that the DoFns in your pipeline code have generated unhandled exceptions, which result in failed tasks in your Dataflow job. Exceptions in user code (for example, your DoFn instances) are reported in the Dataflow monitoring interface.

Question: 327

You are developing a new deep teaming model that predicts a customer's likelihood to buy on your ecommerce site. Alter running an evaluation of the model against both the original training data and new test data, you find that your model is overfitting the dat

a. You want to improve the accuracy of the model when predicting new data. What should you do?

A. Increase the size of the training dataset, and increase the number of input features.

B. Increase the size of the training dataset, and decrease the number of input features.

C. Reduce the size of the training dataset, and increase the number of input features.

D. Reduce the size of the training dataset, and decrease the number of input features.

Answer: B

Explanation:

https://machinelearningmastery.com/impact-of-dataset-size-on-deep-learning-model-skill-and-performance-estimates/

Question: 328

You are implementing workflow pipeline scheduling using open source-based tools and Google Kubernetes Engine (GKE). You want to use a Google managed service to simplify and automate the

task. You also want to accommodate Shared VPC networking considerations. What should you do?

A. Use Dataflow for your workflow pipelines. Use Cloud Run triggers for scheduling.

B. Use Dataflow for your workflow pipelines. Use shell scripts to schedule workflows.

C. Use Cloud Composer in a Shared VPC configuration. Place the Cloud Composer resources in the host project.

D. Use Cloud Composer in a Shared VPC configuration. Place the Cloud Composer resources in the service project.

Answer: D

Explanation:

Shared VPC requires that you designate a host project to which networks and subnetworks belong and a service project, which is attached to the host project. When Cloud Composer participates in a Shared VPC, the Cloud Composer environment is in the service project. Reference: https://cloud.google.com/composer/docs/how-to/managing/configuring-shared-vpc

Question: 329

You are implementing a chatbot to help an online retailer streamline their customer service. The chatbot must be able to respond to both text and voice inquiries. You are looking for a low-code or no-code option, and you want to be able to easily train the chatbot to provide answers to keywords. What should you do?

A. Use the Speech-to-Text API to build a Python application in App Engine.

B. Use the Speech-to-Text API to build a Python application in a Compute Engine instance.

C. Use Dialogflow for simple queries and the Speech-to-Text API for complex queries.

D. Use Dialogflow to implement the chatbot. defining the intents based on the most common queries collected.

Answer: D

Explanation:

Dialogflow is a conversational AI platform that allows for easy implementation of chatbots without needing to code. It has built-in integration for both text and voice input via APIs like Cloud Speech-to- Text. Defining intents and entity types allows you to map common queries and keywords to responses. This would provide a low/no-code way to quickly build and iteratively improve the chatbot capabilities.

https://cloud.google.com/dialogflow/docs Dialogflow is a natural language understanding platform that makes it easy to design and integrate a conversational user interface into your mobile app, web application, device, bot, interactive voice response system, and so on. Using Dialogflow, you can provide new and engaging ways for users to interact with your product. Dialogflow can analyze multiple types of input from your customers, including text or audio inputs (like from a phone or voice recording). It can also respond to your customers in a couple of ways, either through text or with synthetic speech.

Question: 330

You are loading CSV files from Cloud Storage to BigQuery. The files have known data quality issues, including mismatched data types, such as STRINGS and INT64s in the same column, and inconsistent formatting of values such as phone numbers or addresses. You need to create the data pipeline to maintain data quality and perform the required cleansing and transformation. What should you do?

A. Use Data Fusion to transform the data before loading it into BigQuery.

B. Load the CSV files into a staging table with the desired schema, perform the transformations with SQL. and then write the results to the final destination table.

C. Create a table with the desired schema, toad the CSV files into the table, and perform the transformations in place using SQL.

D. Use Data Fusion to convert the CSV files lo a self-describing data formal, such as AVRO. before loading the data to BigOuery.

Answer: A

Explanation:

Data Fusion's advantages:

Visual interface: Offers a user-friendly interface for designing data pipelines without extensive coding, making it accessible to a wider range of users.

Built-in transformations: Includes a wide range of pre-built transformations to handle common data quality issues, such as:

Data type conversions

Data cleansing (e.g., removing invalid characters, correcting formatting)

Data validation (e.g., checking for missing values, enforcing constraints)

Data enrichment (e.g., adding derived fields, joining with other datasets)

Custom transformations: Allows for custom transformations using SQL or Java code for more complex cleaning tasks.

Scalability: Can handle large datasets efficiently, making it suitable for processing CSV files with potential data quality issues.

Integration with BigQuery: Integrates seamlessly with BigQuery, allowing for direct loading of transformed data.

Question: 331

Your organization uses a multi-cloud data storage strategy, storing data in Cloud Storage, and data in Amazon Web Services' (AWS) S3 storage buckets. All data resides in US regions. You want to query up-to-date data by using BigQuery. regardless of which cloud the data is stored in. You need to allow users to query the tables from BigQuery without giving direct access to the data in the storage buckets What should you do?

A. Set up a BigQuery Omni connection to the AWS S3 bucket data Create BigLake tables over the Cloud Storage and S3 data and query the data using BigQuery directly.

B. Set up a BigQuery Omni connection to the AWS S3 bucket data. Create external tables over the Cloud Storage and S3 data and query the data using BigQuery directly.

C. Use the Storage Transfer Service to copy data from the AWS S3 buckets to Cloud Storage buckets Create BigLake tables over the Cloud Storage data and query the data using BigQuery directly.

D. Use the Storage Transfer Service to copy data from the AWS S3 buckets to Cloud Storage buckets Create external tables over the Cloud Storage data and query the data using BigQuery directly

Answer: B

Explanation:

BigQuery Omni enables you to run BigQuery analytics directly on data stored in AWS S3 buckets without having to move or copy the data. This provides several benefits:

Reduced Data Movement Costs: Eliminates the need to egress data from AWS, potentially saving significant costs.

Real-Time Analytics: Allows you to query data in AWS S3 in real-time, providing up-to-date insights.

Simplified Architecture: Reduces the complexity of managing data pipelines and ETL processes.

Here's a breakdown of the steps involved in using BigQuery Omni:

Set up a BigQuery Omni connection: This involves configuring the connection between your Google Cloud project and your AWS S3 bucket. This connection establishes the secure link for BigQuery to access the data in AWS S3.

Create external tables: BigQuery external tables are a way to query data residing in external storage systems, such as AWS S3, without having to import the data into BigQuery. This enables you to directly query the data in its original location.

Query the data using BigQuery: Once the external tables are created, you can use standard SQL queries to analyze the data stored in both Cloud Storage and AWS S3, just as if it were native BigQuery data.

Why other options are not suitable:

Option A: BigLake tables are designed for storing large volumes of structured data within BigQuery itself, not for directly querying data in external storage systems.

Option C and D: While the Storage Transfer Service is useful for moving data between cloud providers, it introduces unnecessary data movement and latency if the goal is to simply query the data in its original location.

Key Points:

BigQuery Omni extends BigQuery's capabilities to analyze data stored in other cloud providers, such as AWS.

External tables provide a way to query data in external storage systems without having to import it into BigQuery.

By leveraging BigQuery Omni and external tables, you can efficiently and cost-effectively query data stored in multiple cloud environments using a single tool, BigQuery.

Question: 332

You have thousands of Apache Spark jobs running in your on-premises Apache Hadoop cluster. You want to migrate the jobs to Google Cloud. You want to use managed services to run your jobs instead of maintaining a long-lived Hadoop cluster yourself. You have a tight timeline and want to keep code changes to a minimum. What should you do?

A. Copy your data to Compute Engine disks. Manage and run your jobs directly on those instances.

B. Move your data to Cloud Storage. Run your jobs on Dataproc.

C. Move your data to BigQuery. Convert your Spark scripts to a SQL-based processing approach.

D. Rewrite your jobs in Apache Beam. Run your jobs in Dataflow.

Answer: B

Explanation:

Dataproc's Compatibility with Apache Spark: Dataproc is a managed service for running Hadoop and Spark clusters on Google Cloud. This means it is designed to seamlessly run Apache Spark jobs with minimal code changes. Your existing Spark jobs should run on Dataproc with little to no modification.

Cloud Storage as a Scalable Data Lake: Cloud Storage provides a highly scalable and durable storage solution for your data. It's designed to handle large volumes of data that Spark jobs typically process.

Minimizing Operational Overhead: By using Dataproc, you eliminate the need to manage and maintain a Hadoop cluster yourself. Google Cloud handles the infrastructure, allowing you to focus on your data processing tasks.

Tight Timeline and Minimal Code Changes: This option directly addresses the requirements of the question. It offers a quick and easy way to migrate your Spark jobs to Google Cloud with minimal disruption to your existing codebase.

Why other options are not suitable:

A . Copy your data to Compute Engine disks. Manage and run your jobs directly on those instances: This option requires you to manage the underlying infrastructure yourself, which contradicts the requirement of using managed services.

C . Move your data to BigQuery. Convert your Spark scripts to a SQL-based processing approach: While BigQuery is a powerful data warehouse, converting Spark scripts to SQL would require substantial code changes and might not be feasible within a tight timeline.

D . Rewrite your jobs in Apache Beam. Run your jobs in Dataflow: Rewriting jobs in Apache Beam would be a significant

undertaking and not suitable for a quick migration with minimal code changes.

Question: 333

You work for a farming company. You have one BigQuery table named sensors, which is about 500 MB and contains the list of your 5000 sensors, with columns for id, name, and location. This table is updated every hour. Each sensor generates one metric every 30 seconds along with a timestamp. which you want to store in BigQuery. You want to run an analytical query on the data once a week for monitoring purposes. You also want to minimize costs. What data model should you use?

A. 1. Create a retries column in the sensor? table.

2. Set record type and repeated mode for the metrics column.

3. Use an UPDATE statement every 30 seconds to add new metrics.

B. 1. Create a metrics column in the sensors table.

2. Set RECORD type and REPEATED mode for the metrics column.

3. Use an INSERT statement every 30 seconds to add new metrics.

C. 1. Create a metrics table partitioned by timestamp.

2. Create a sensorld column in the metrics table, that points to the id column in the sensors table.

3. Use an IHSEW statement every 30 seconds to append new metrics to the metrics table.

4. Join the two tables, if needed, when running the analytical query.

D. 1. Create a metrics table partitioned by timestamp.

2. Create a sensor Id column in the metrics table, that points to the _d column in the sensors table.

3. Use an UPDATE statement every 30 seconds to append new metrics to the metrics table.

4. Join the two tables, if needed, when running the analytical query.

Answer: C

Explanation:

For a farming company with a sensor data table updated every 30 seconds, the goal is to minimize costs while facilitating

weekly analytical queries. The best data model will effectively manage data storage, update frequency, and query performance.

Partitioned Metrics Table:

Creating a metrics table partitioned by timestamp optimizes query performance and storage costs.

Partitioning by timestamp allows for efficient querying, especially for time-based analyses.

Sensor ID Reference:

Including a sensor_id column in the metrics table that points to the id column in the sensors table ensures data normalization.

This structure avoids redundancy and maintains a clear relationship between sensors and their metrics.

Using INSERT Statements:

Using INSERT statements to append new metrics every 30 seconds is efficient and cost-effective.

INSERT operations are more suitable than UPDATE operations for adding new data entries, especially at high frequencies.

Joining Tables for Analysis:

When running analytical queries, joining the partitioned metrics table with the sensors table as needed provides a comprehensive view of the data.

This approach leverages BigQuery's powerful JOIN capabilities while keeping the data model normalized and efficient.

Google Data Engineer Reference:

BigQuery Partitioned Tables

BigQuery Best Practices

Efficient Data Partitioning

BigQuery Data Modeling

Using this data model, the farming company can manage its sensor data effectively, minimize costs, and perform weekly analytical queries with high efficiency.

Question: 334

Your infrastructure team has set up an interconnect link between Google Cloud and the on-premises network. You are designing a high-throughput streaming pipeline to ingest data in streaming from an Apache Kafka cluster hosted on-premises. You want to store the data in BigQuery, with as minima! latency as possible. What should you do?

A. Use a proxy host in the VPC in Google Cloud connecting to Kafka. Write a Dataflow pipeline, read data from the proxy host, and write the data to BigQuery.

B. Setup a Kafka Connect bridge between Kafka and Pub/Sub. Use a Google-provided Dataflow template to read the data from Pub/Sub, and write the data to BigQuery.

C. Setup a Kafka Connect bridge between Kafka and Pub/Sub. Write a Dataflow pipeline, read the data from Pub/Sub, and write the data to

BigQuery.

D. Use Dataflow, write a pipeline that reads the data from Kafka, and writes the data to BigQuery.

Answer: C

Explanation:

Here's a detailed breakdown of why this solution is optimal and why others fall short:

Why Option C is the Best Solution:

Kafka Connect Bridge: This bridge acts as a reliable and scalable conduit between your on-premises Kafka cluster and Google Cloud's Pub/Sub messaging service. It handles the complexities of securely transferring data over the interconnect link.

Pub/Sub as a Buffer: Pub/Sub serves as a highly scalable buffer, decoupling the Kafka producer from the Dataflow consumer. This is crucial for handling fluctuations in message volume and ensuring smooth data flow even during spikes.

Custom Dataflow Pipeline: Writing a custom Dataflow pipeline gives you the flexibility to implement any necessary transformations or enrichments to the data before it's written to BigQuery. This is often required in real-world streaming scenarios.

Minimal Latency: By using Pub/Sub as a buffer and Dataflow for efficient processing, you minimize the latency between the data being produced in Kafka and being available for querying in BigQuery.

Why Other Options Are Not Ideal:

Option A: Using a proxy host introduces an additional point of failure and can create a bottleneck, especially with high-throughput streaming.

Option B: While Google-provided Dataflow templates can be helpful, they might lack the customization needed for specific transformations or handling complex data structures.

Option D: Dataflow doesn't natively connect to on-premises Kafka clusters. Directly reading from Kafka would require complex networking configurations and could lead to performance issues.

Additional Considerations:

Schema Management: Ensure that the schema of the data being produced in Kafka is compatible with the schema expected in BigQuery. Consider using tools like Schema Registry for schema evolution management.

Monitoring: Set up robust monitoring and alerting to detect any issues in the pipeline, such as message backlogs or processing errors.

By following Option C, you leverage the strengths of Kafka Connect, Pub/Sub, and Dataflow to create a high-throughput, low-latency streaming pipeline that seamlessly integrates your on-premises Kafka data with BigQuery.

Question: 335

You need to connect multiple applications with dynamic public IP addresses to a Cloud SQL instance.
You configured users with strong passwords and enforced the SSL connection to your Cloud SOL instance. You want to use Cloud SQL public IP and ensure that you have secured connections. What

should you do?

A. Add all application networks to Authorized Network and regularly update them.

B. Add CIDR 0.0.0.0/0 network to Authorized Network. Use Identity and Access Management (1AM) to add users.

C. Leave the Authorized Network empty. Use Cloud SQL Auth proxy on all applications.

D. Add CIDR 0.0.0.0/0 network to Authorized Network. Use Cloud SOL Auth proxy on all applications.

Answer: C

Explanation:

To securely connect multiple applications with dynamic public IP addresses to a Cloud SQL instance using public IP, the Cloud SQL Auth proxy is the best solution. This proxy provides secure, authorized connections to Cloud SQL instances without the need to configure authorized networks or deal with IP whitelisting complexities.

Cloud SQL Auth Proxy:

The Cloud SQL Auth proxy provides secure, encrypted connections to Cloud SQL.

It uses IAM permissions and SSL to authenticate and encrypt the connection, ensuring data security in transit.

By using the proxy, you avoid the need to constantly update authorized networks as the proxy handles dynamic IP addresses seamlessly.

Authorized Network Configuration:

Leaving the authorized network empty means no IP addresses are explicitly whitelisted, relying solely on the Auth proxy for secure connections.

This approach simplifies network management and enhances security by not exposing the Cloud SQL instance to public IP ranges.

Dynamic IP Handling:

Applications with dynamic IP addresses can securely connect through the proxy without the need to modify authorized networks.

The proxy authenticates connections using IAM, making it ideal for environments where application

IPs change frequently.

Google Data Engineer Reference:

Using Cloud SQL Auth Proxy

Cloud SQL Security Overview

Setting up the Cloud SQL Auth Proxy

By using the Cloud SQL Auth proxy, you ensure secure, authorized connections for applications with dynamic public IPs without the need for complex network configurations.

Question: 336

You are creating the CI'CD cycle for the code of the directed acyclic graphs (DAGs) running in Cloud Composer. Your team has two Cloud Composer instances: one instance for development and another instance for production. Your team is using a Git repository to maintain and develop the code of the DAGs. You want to deploy the DAGs automatically to Cloud Composer when a certain tag is pushed to the Git repository. What should you do?

A. 1. Use Cloud Build to build a container and the Kubemetes Pod Operator to deploy the code of the DAG to the Google Kubernetes

Engine (GKE) cluster of the development instance for testing.

2. If the tests pass, copy the code to the Cloud Storage bucket of the production instance.

B. 1 Use Cloud Build to copy the code of the DAG to the Cloud Storage bucket of the development instance for DAG testing.

2. If the tests pass, use Cloud Build to build a container with the code of the DAG and the KubernetesPodOperator to deploy the container to the Google Kubernetes Engine (GKE) cluster of the production instance.

C. 1 Use Cloud Build to build a container with the code of the DAG and the KubernetesPodOperator to deploy the code to the Google Kubernetes Engine (GKE) cluster of the development instance for testing.

2. If the tests pass, use the KubernetesPodOperator to deploy the container to the GKE cluster of the production instance.

D. 1 Use Cloud Build to copy the code of the DAG to the Cloud Storage bucket of the development instance for DAG testing.

2. If the tests pass, use Cloud Build to copy the code to the bucket of the production instance.

Answer: C

Explanation:

Question: 337

You migrated your on-premises Apache Hadoop Distributed File System (HDFS) data lake to Cloud Storage. The data scientist team needs to process the data by using Apache Spark and SQL. Security policies need to be enforced at the column level. You need a cost-effective solution that can scale into a data mesh. What should you do?

A. 1. Deploy a long-living Dalaproc cluster with Apache Hive and Ranger enabled.

2. Configure Ranger for column level security.

3. Process with Dataproc Spark or Hive SQL.

B. 1. Define a BigLake table.

2. Create a taxonomy of policy tags in Data Catalog.
3. Add policy lags to columns.
4. Process with the Spark-BigQuery connector or BigQuery SOL.

C. 1. Load the data to BigQuery tables.
2. Create a taxonomy of policy tags in Data Catalog.
3. Add policy tags to columns.
4. Procoss with the Spark-BigQuery connector or BigQuery SQL.

D. 1 Apply an Identity and Access Management (IAM) policy at the file level in Cloud Storage
2. Define a BigQuery external table for SQL processing.
3. Use Dataproc Spark to process the Cloud Storage files.

Answer: D

Explanation:

For automating the CI/CD pipeline of DAGs running in Cloud Composer, the following approach ensures that DAGs are tested and deployed in a streamlined and efficient manner.

Use Cloud Build for Development Instance Testing:

Use Cloud Build to automate the process of copying the DAG code to the Cloud Storage bucket of the development instance.

This triggers Cloud Composer to automatically pick up and test the new DAGs in the development environment.

Testing and Validation:

Ensure that the DAGs run successfully in the development environment.

Validate the functionality and correctness of the DAGs before promoting them to production.

Deploy to Production:

If the DAGs pass all tests in the development environment, use Cloud Build to copy the tested DAG code to the Cloud Storage bucket of the production instance.

This ensures that only validated and tested DAGs are deployed to production, maintaining the stability and reliability of the production environment.

Simplicity and Reliability:

This approach leverages Cloud Build's capabilities for automation and integrates seamlessly with Cloud Composer's reliance on Cloud Storage for DAG storage.

By using Cloud Storage for both development and production deployments, the process remains simple and robust.

Google Data Engineer Reference:

Cloud Composer Documentation

Using Cloud Build

Deploying DAGs to Cloud Composer

Automating DAG Deployment with Cloud Build

By implementing this CI/CD pipeline, you ensure that DAGs are thoroughly tested in the

development environment before being automatically deployed to the production environment, maintaining high quality and reliability.

Question: 338

You are administering shared BigQuery datasets that contain views used by multiple teams in your organization. The marketing team is concerned about the variability of their monthly BigQuery analytics spend using the on-demand billing model. You need to help the marketing team establish a consistent BigQuery analytics spend each month.

What should you do?

A. Create a BigQuery Standard pay-as-you go reservation with a baseline of 0 slots and autoscaling set to 500 for the marketing team, and bill them back accordingly.

B. Create a BigQuery reservation with a baseline of 500 slots with no autoscaling for the marketing team, and bill them back accordingly.

C. Establish a BigQuery quota for the marketing team, and limit the maximum number of bytes scanned each day.

D. Create a BigQuery Enterprise reservation with a baseline of 250 slots and autoscaling set to 500 for the marketing team, and bill them back accordingly.

Answer: B

Explanation:

To help the marketing team establish a consistent BigQuery analytics spend each month, you can use BigQuery reservations to allocate dedicated slots for their queries. This provides predictable costs by reserving a fixed amount of compute resources.

BigQuery Reservations:

BigQuery Reservations allow you to purchase dedicated query processing capacity in the form of slots.

By reserving slots, you can control costs and ensure that the marketing team has the necessary resources for their queries without unexpected increases in spending.

Baseline Slots:

Setting a baseline of 500 slots without autoscaling ensures a consistent allocation of resources.

This provides a predictable monthly cost, as the marketing team will be billed for the reserved slots regardless of actual usage.

Billing Back:

The marketing team's usage can be billed back based on the fixed reservation cost, ensuring budget predictability.

This approach avoids the variability associated with on-demand billing, where costs can fluctuate based on query volume and complexity.

No Autoscaling:

By not enabling autoscaling, you prevent additional costs from being incurred due to temporary increases in query demand.

This fixed reservation ensures that the marketing team only uses the allocated 500 slots, maintaining a consistent monthly spend.

Google Data Engineer Reference:

BigQuery Reservations Documentation

BigQuery Slot Reservations

Managing BigQuery Costs

Using a fixed reservation of 500 slots provides the marketing team with predictable costs and the necessary resources for their queries without unexpected billing variability.

Question: 339

You need to create a SQL pipeline. The pipeline runs an aggregate SOL transformation on a BigQuery table every two hours and appends the result to another existing BigQuery table. You need to configure the pipeline to retry if errors occur. You want the pipeline to send an email notification after three consecutive failures. What should you do?

A. Create a BigQuery scheduled query to run the SOL transformation with schedule options that repeats every two hours, and enable email notifications.

B. Use the BigQueryUpsertTableOperator in Cloud Composer, set the retry parameter to three, and set the email_on_failure parameter to true.

C. Use the BigQuerylnsertJobOperator in Cloud Composer, set the retry parameter to three, and set the email_on_failure parameter to true.

D. Create a BigQuery scheduled query to run the SQL transformation with schedule options that repeats every two hours, and enable notification to Pub/Sub topic. Use Pub/Sub and Cloud Functions to send an email after three tailed executions.

Answer: D

Explanation:

To create a robust and resilient SQL pipeline in BigQuery that handles retries and failure notifications, consider the following:

Explanation:

BigQuery Scheduled Queries: This feature allows you to schedule recurring queries in BigQuery. It is a straightforward way to run SQL transformations on a regular basis without requiring extensive setup.

Error Handling and Retries: While BigQuery Scheduled Queries can run at specified intervals, they don't natively support complex retry logic or failure notifications directly. This is where additional Google Cloud services like Pub/Sub and Cloud Functions come into play.

Pub/Sub for Notifications: By configuring a BigQuery scheduled query to publish messages to a Pub/Sub topic upon failure, you can create a decoupled and scalable notification system.

Cloud Functions: Cloud Functions can subscribe to the Pub/Sub topic and implement logic to count consecutive failures. After detecting three consecutive failures, the Cloud Function can then send an email notification using a service like SendGrid or Gmail API.

Implementation Steps:

Set up a BigQuery Scheduled Query:

Create a scheduled query in BigQuery to run your SQL transformation every two hours.

Configure the scheduled query to publish a notification to a Pub/Sub topic in case of a failure.

Create a Pub/Sub Topic:

Create a Pub/Sub topic that will receive messages from the scheduled query.

Develop a Cloud Function:

Write a Cloud Function that subscribes to the Pub/Sub topic.

Implement logic in the Cloud Function to track failure messages. If three consecutive failure messages are detected, the function sends an email notification.

Reference: Links:

BigQuery Scheduled Queries

Pub/Sub Documentation

Cloud Functions Documentation

SendGrid Email API

Gmail API

Question: 340

You are planning to load some of your existing on-premises data into BigQuery on Google Cloud. You want to either stream or batch-load data, depending on your use case. Additionally, you want to mask some sensitive data before loading into BigQuery. You need to do this in a programmatic way while keeping costs to a minimum. What should you do?

A. Use the BigQuery Data Transfer Service to schedule your migration. After the data is populated in BigQuery. use the connection to the Cloud Data Loss Prevention {Cloud DLP} API to de-identify the necessary data.

B. Create your pipeline with Dataflow through the Apache Beam SDK for Python, customizing separate options within your code for streaming.

batch processing, and Cloud DLP Select BigQuery as your data sink.

C. Use Cloud Data Fusion to design your pipeline, use the Cloud DLP plug-in to de-identify data within your pipeline, and then move the data

into BigQuery.

D. Set up Datastream to replicate your on-premise data on BigQuery.

Answer: B

Explanation:

To load on-premises data into BigQuery while masking sensitive data, we need a solution that offers flexibility for both streaming and batch processing, as well as data masking capabilities. Here's a detailed explanation of why option B is the best choice:

Explanation:

Apache Beam and Dataflow:

Apache Beam SDK provides a unified programming model for both batch and stream data processing.

Google Cloud Dataflow is a fully managed service for executing Apache Beam pipelines, offering scalability and ease of use.

Customization for Different Use Cases:

By using the Apache Beam SDK, you can write custom pipelines that can handle both streaming and batch processing within the same framework.

This allows you to switch between streaming and batch modes based on your use case without changing the core logic of your data pipeline.

Data Masking with Cloud DLP:

Google Cloud Data Loss Prevention (DLP) API can be integrated into your Apache Beam pipeline to de-identify and mask sensitive data programmatically before loading it into BigQuery.

This ensures that sensitive data is handled securely and complies with privacy requirements.

Cost Efficiency:

Using Dataflow can be cost-effective because it is a fully managed service, reducing the operational overhead associated with managing your own infrastructure.

The pay-as-you-go model ensures you only pay for the resources you consume, which can help keep costs under control.

Implementation Steps:

Set up Apache Beam Pipeline:

Write a pipeline using the Apache Beam SDK for Python that reads data from your on-premises storage.

Add transformations for data processing, including the integration with Cloud DLP for data masking.

Configure Dataflow:

Deploy the Apache Beam pipeline on Google Cloud Dataflow.

Customize the pipeline options for both streaming and batch use cases.

Load Data into BigQuery:

Set BigQuery as the sink for your data in the Apache Beam pipeline.

Ensure the processed and masked data is loaded into the appropriate BigQuery tables.

Reference: Links:

Apache Beam Documentation

Google Cloud Dataflow Documentation

Google Cloud DLP Documentation

BigQuery Documentation

Question: 341

Your company operates in three domains: airlines, hotels, and ride-hailing services. Each domain has two teams: analytics and data science, which create data assets in BigQuery with the help of a central data platform team. However, as each domain is evolving rapidly, the central data platform team is becoming a bottleneck. This is causing delays in deriving insights from data, and resulting in stale data when pipelines are not kept up to date. You need to design a data mesh architecture by using Dataplex to eliminate the bottleneck. What should you do?

A. 1. Create one lake for each team. Inside each lake, create one zone for each domain.

2. Attach each of the BigQuery datasets created by the individual teams as assets to the respective zone.

3. Have the central data platform team manage all zones' data assets.

B. 1 Create one lake for each team. Inside each lake, create one zone for each domain.

2. Attach each to the BigQuory datasets created by the individual teams as assets to the respective zone.

3. Direct each domain to manage their own zone's data assets.

C. 1 Create one lake for each domain. Inside each lake, create one zone for each team.

2. Attach each of the BigQuery datasets created by the individual teams as assets to the respective zone.

3. Direct each domain to manage their own lake's data assets.

D. 1 Create one lake for each domain. Inside each lake, create one zone for each team.

2. Attach each of the BigQuery datasets created by the individual teams as assets to the respective zone.

3. Have the central data platform team manage all lakes' data assets.

Answer: B

Explanation:

To design a data mesh architecture using Dataplex to eliminate bottlenecks caused by a central data platform team, consider the following:

Explanation:

Data Mesh Architecture:

Data mesh promotes a decentralized approach where domain teams manage their own data pipelines and assets, increasing agility and reducing bottlenecks.

Dataplex Lakes and Zones:

Lakes in Dataplex are logical containers for managing data at scale, and zones are subdivisions within lakes for organizing data based on domains, teams, or other criteria.

Domain and Team Management:

By creating a lake for each team and zones for each domain, each team can independently manage their data assets without relying on the central data platform team.

This setup aligns with the principles of data mesh, promoting ownership and reducing delays in data processing and insights.

Implementation Steps:

Create Lakes and Zones:

Create separate lakes in Dataplex for each team (analytics and data science).

Within each lake, create zones for the different domains (airlines, hotels, ride-hailing).

Attach BigQuery Datasets:

Attach the BigQuery datasets created by the respective teams as assets to their corresponding zones.

Decentralized Management:

Allow each domain to manage their own zone's data assets, providing them with the autonomy to update and maintain their pipelines without depending on the central team.

Reference: Links:

Dataplex Documentation

BigQuery Documentation

Data Mesh Principles

Question: 342

You have important legal hold documents in a Cloud Storage bucket. You need to ensure that these documents are not deleted or modified. What should you do?

A. Set a retention policy. Lock the retention policy.

B. Set a retention policy. Set the default storage class to Archive for long-term digital preservation.

C. Enable the Object Versioning feature. Add a lifecycle rule.

D. Enable the Object Versioning feature. Create a copy in a bucket in a different region.

Answer: A

Explanation:

To ensure that important legal hold documents in a Cloud Storage bucket are not deleted or modified, the most effective method is to set and lock a retention policy. Here's why this is the best choice:

Explanation:

Retention Policy:

A retention policy defines a retention period during which objects in the bucket cannot be deleted or modified. This ensures data immutability.

Once a retention policy is set and locked, it cannot be removed or reduced, providing strong protection against accidental or malicious deletions.

Locking the Retention Policy:

Locking a retention policy ensures that the retention period cannot be changed. This action is permanent and guarantees that the specified retention period will be enforced.

Steps to Implement:

Set the Retention Policy:

Define a retention period for the bucket to ensure that all objects are protected for the required duration.

Lock the Retention Policy:

Lock the retention policy to prevent any modifications, ensuring the immutability of the documents.

Reference: Links:

Cloud Storage Retention Policy Documentation

How to Set a Retention Policy

Question: 343

You are running your BigQuery project in the on-demand billing model and are executing a change data capture (CDC) process that ingests dat

a. The CDC process loads 1 GB of data every 10 minutes into a temporary table, and then performs a merge into a 10 TB target table. This process is very scan intensive and you want to explore options to enable a predictable cost model. You need to create a BigQuery reservation based on utilization information gathered from BigQuery Monitoring and apply the reservation to the CDC process. What should you do?

A. Create a BigQuery reservation for the job.

B. Create a BigQuery reservation for the service account running the job.

C. Create a BigQuery reservation for the dataset.

D. Create a BigQuery reservation for the project.

Answer: D

Explanation:

https://cloud.google.com/blog/products/data-analytics/manage-bigquery-costs-with-custom- quotas.

Here's why creating a BigQuery reservation for the project is the most suitable solution:

Project-Level Reservation: BigQuery reservations are applied at the project level. This means that the reserved slots (processing capacity) are shared across all jobs and queries running within that project. Since your CDC process is a significant contributor to your BigQuery usage, reserving slots for the entire project ensures that your CDC process always has access to the necessary resources, regardless of other activities in the project.

Predictable Cost Model: Reservations provide a fixed, predictable cost model. Instead of paying the on-demand price for each query, you pay a fixed monthly fee for the reserved slots. This eliminates the variability of costs associated with on-demand billing, making it easier to budget and forecast your BigQuery expenses.

BigQuery Monitoring: You can use BigQuery Monitoring to analyze the historical usage patterns of your CDC process and other queries within your project. This information helps you determine the appropriate amount of slots to reserve, ensuring that you have enough capacity to handle your workload while optimizing costs.

Why other options are not suitable:

A . Create a BigQuery reservation for the job: BigQuery does not support reservations at the individual job level. Reservations are applied at the project or assignment level.

B . Create a BigQuery reservation for the service account running the job: While you can create reservations for assignments (groups of users or service accounts), it's less efficient than a projectlevel reservation in this scenario. A project-level reservation covers all jobs within the project, regardless of the service account used.

C . Create a BigQuery reservation for the dataset: BigQuery does not support reservations at the dataset level.

By creating a BigQuery reservation for your project based on your utilization analysis, you can achieve a predictable cost model while ensuring that your CDC process and other queries have the necessary resources to run smoothly.

Question: 344

A web server sends click events to a Pub/Sub topic as messages. The web server includes an event Timestamp attribute in the messages, which is the time when the click occurred. You have a Dataflow streaming job that reads from this Pub/Sub topic through a subscription, applies some transformations, and writes the result to another Pub/Sub topic for use by the advertising department. The advertising department needs to receive each message within 30 seconds of the corresponding click occurrence, but they report receiving the messages late. Your Dataflow job's system lag is about 5 seconds, and the data freshness is about 40 seconds. Inspecting a few messages show no more than 1 second lag between their event Timestamp and publish Time. What is the problem and what should you do?

A. The advertising department is causing delays when consuming the messages. Work with the advertising department to fix this.

B. Messages in your Dataflow job are processed in less than 30 seconds, but your job cannot keep up with the backlog in the Pub/Sub

subscription. Optimize your job or increase the number of workers to fix this.

C. The web server is not pushing messages fast enough to Pub/Sub. Work with the web server team to fix this.

D. Messages in your Dataflow job are taking more than 30 seconds to process. Optimize your job or increase the number of workers to fix this.

Answer: B

Explanation:

To ensure that the advertising department receives messages within 30 seconds of the click occurrence, and given the current system lag and data freshness metrics, the issue likely lies in the

processing capacity of the Dataflow job. Here's why option B is the best choice:

Explanation:

System Lag and Data Freshness:

The system lag of 5 seconds indicates that Dataflow itself is processing messages relatively quickly.

However, the data freshness of 40 seconds suggests a significant delay before processing begins, indicating a backlog.

Backlog in Pub/Sub Subscription:

A backlog occurs when the rate of incoming messages exceeds the rate at which the Dataflow job can process them, causing delays.

Optimizing the Dataflow Job:

To handle the incoming message rate, the Dataflow job needs to be optimized or scaled up by increasing the number of workers, ensuring it can keep up with the message inflow.

Steps to Implement:

Analyze the Dataflow Job:

Inspect the Dataflow job metrics to identify bottlenecks and inefficiencies.

Optimize Processing Logic:

Optimize the transformations and operations within the Dataflow pipeline to improve processing efficiency.

Increase Number of Workers:

Scale the Dataflow job by increasing the number of workers to handle the higher load, reducing the backlog.

Reference: Links:

Dataflow Monitoring

Scaling Dataflow Jobs

Question: 345

You have a BigQuery dataset named "customers". All tables will be tagged by using a Data Catalog tag template named "gdpr". The template contains one mandatory field, "has sensitive data~. with a

boolean value. All employees must be able to do a simple search and find tables in the dataset that have either true or false in the "has sensitive data" field. However, only the Human Resources (HR) group should be able to see the data inside the tables for which "hass-ensitive-data" is true. You give the all employees group the bigquery.metadataViewer and bigquery.connectionUser roles on the dataset. You want to minimize configuration overhead. What should you do next?

A. Create the "gdpr" tag template with private visibility. Assign the bigquery -dataViewer role to the HR group on the tables that contain sensitive data.

B. Create the ~gdpr" tag template with private visibility. Assign the datacatalog. tagTemplateViewer role on this tag to the all employees

group, and assign the bigquery.dataViewer role to the HR group on the tables that contain sensitive data.

C. Create the "gdpr" tag template with public visibility. Assign the bigquery. dataViewer role to the HR group on the tables that contain sensitive data.

D. Create the "gdpr" tag template with public visibility. Assign the datacatalog. tagTemplateViewer role on this tag to the all employees.

group, and assign the bijquery.dataViewer role to the HR group on the tables that contain sensitive data.

Answer: D

Explanation:

To ensure that all employees can search and find tables with GDPR tags while restricting data access to sensitive tables only to the HR group, follow these steps:

Explanation:

Data Catalog Tag Template:

Use Data Catalog to create a tag template named "gdpr" with a boolean field "has sensitive data".

Set the visibility to public so all employees can see the tags.

Roles and Permissions:

Assign the datacatalog.tagTemplateViewer role to the all employees group. This role allows users to

view the tags and search for tables based on the "has sensitive data" field.

Assign the bigquery.dataViewer role to the HR group specifically on tables that contain sensitive data. This ensures only HR can access the actual data in these tables.

Steps to Implement:

Create the GDPR Tag Template:

Define the tag template in Data Catalog with the necessary fields and set visibility to public.

Assign Roles:

Grant the datacatalog.tagTemplateViewer role to the all employees group for visibility into the tags.

Grant the bigquery.dataViewer role to the HR group on tables marked as having sensitive data.

Reference: Links:

Data Catalog Documentation

Managing Access Control in BigQuery

IAM Roles in Data Catalog

Question: 346

You are architecting a data transformation solution for BigQuery. Your developers are proficient with SOL and want to use the ELT development technique. In addition, your developers need an intuitive coding environment and the ability to manage SQL as code. You need to identify a solution for your developers to build these pipelines. What should you do?

A. Use Cloud Composer to load data and run SQL pipelines by using the BigQuery job operators.

B. Use Dataflow jobs to read data from Pub/Sub, transform the data, and load the data to BigQuery.

C. Use Dataform to build, manage, and schedule SQL pipelines.

D. Use Data Fusion to build and execute ETL pipelines

Answer: C

Explanation:

To architect a data transformation solution for BigQuery that aligns with the ELT development technique and provides an intuitive coding environment for SQL-proficient developers, Dataform is an optimal choice. Here's why:

Explanation:

ELT Development Technique:

ELT (Extract, Load, Transform) is a process where data is first extracted and loaded into a data warehouse, and then transformed using SQL queries. This is different from ETL, where data is transformed before being loaded into the data warehouse.

BigQuery supports ELT, allowing developers to write SQL transformations directly in the data warehouse.

Dataform:

Dataform is a development environment designed specifically for data transformations in BigQuery and other SQL-based warehouses.

It provides tools for managing SQL as code, including version control and collaborative development.

Dataform integrates well with existing development workflows and supports scheduling and managing SQL-based data pipelines.

Intuitive Coding Environment:

Dataform offers an intuitive and user-friendly interface for writing and managing SQL queries.

It includes features like SQLX, a SQL dialect that extends standard SQL with features for modularity and reusability, which simplifies the development of complex transformation logic.

Managing SQL as Code:

Dataform supports version control systems like Git, enabling developers to manage their SQL transformations as code.

This allows for better collaboration, code reviews, and version tracking.

Reference: Links:

Dataform Documentation

BigQuery Documentation

Managing ELT Pipelines with Dataform

Question: 347

You recently deployed several data processing jobs into your Cloud Composer 2 environment. You notice that some tasks are failing in Apache Airflow. On the monitoring dashboard, you see an increase in the total workers' memory usage, and there were worker pod evictions. You need to resolve these errors. What should you do?

Choose 2 answers

A. Increase the directed acyclic graph (DAG) file parsing interval.

B. Increase the memory available to the Airflow workers.

C. Increase the maximum number of workers and reduce worker concurrency.

D. Increase the memory available to the Airflow triggerer.

E. Increase the Cloud Composer 2 environment size from medium to large.

Answer: BC

Explanation:

To resolve issues related to increased memory usage and worker pod evictions in your Cloud Composer 2 environment, the following steps are recommended:

Explanation:

Increase Memory Available to Airflow Workers:

By increasing the memory allocated to Airflow workers, you can handle more memory-intensive tasks, reducing the likelihood of pod evictions due to memory limits.

Increase Maximum Number of Workers and Reduce Worker Concurrency:

Increasing the number of workers allows the workload to be distributed across more pods, preventing any single pod from becoming overwhelmed.

Reducing worker concurrency limits the number of tasks that each worker can handle simultaneously, thereby lowering the memory consumption per worker.

Steps to Implement:

Increase Worker Memory:

Modify the configuration settings in Cloud Composer to allocate more memory to Airflow workers.

This can be done through the environment configuration settings.

Adjust Worker and Concurrency Settings:

Increase the maximum number of workers in the Cloud Composer environment settings.

Reduce the concurrency setting for Airflow workers to ensure that each worker handles fewer tasks at a time, thus consuming less memory per worker.

Reference: Links:

Cloud Composer Worker Configuration

Scaling Airflow Workers

Question: 348

You want to encrypt the customer data stored in BigQuery. You need to implement for-user cryptodeletion on data

stored in your tables. You want to adopt native features in Google Cloud to avoid custom solutions. What should you do?

A. Create a customer-managed encryption key (CMEK) in Cloud KMS. Associate the key to the table while creating the table.

B. Create a customer-managed encryption key (CMEK) in Cloud KMS. Use the key to encrypt data before storing in BigQuery.

C. Implement Authenticated Encryption with Associated Data (AEAD) BigQuery functions while storing your data in BigQuery.

D. Encrypt your data during ingestion by using a cryptographic library supported by your ETL pipeline.

Answer: A

Explanation:

To implement for-user crypto-deletion and ensure that customer data stored in BigQuery is encrypted, using native Google Cloud features, the best approach is to use Customer-Managed Encryption Keys (CMEK) with Cloud Key Management Service (KMS). Here's why:

Explanation:

Customer-Managed Encryption Keys (CMEK):

CMEK allows you to manage your own encryption keys using Cloud KMS. These keys provide additional control over data access and encryption management.

Associating a CMEK with a BigQuery table ensures that data is encrypted with a key you manage.

For-User Crypto-Deletion:

For-user crypto-deletion can be achieved by disabling or destroying the CMEK. Once the key is disabled or destroyed, the data encrypted with that key cannot be decrypted, effectively rendering it unreadable.

Native Integration:

Using CMEK with BigQuery is a native feature, avoiding the need for custom encryption solutions. This simplifies the management and implementation of encryption and decryption processes.

Steps to Implement:

Create a CMEK in Cloud KMS:

Set up a new customer-managed encryption key in Cloud KMS.

Associate the CMEK with BigQuery Tables:

When creating a new table in BigQuery, specify the CMEK to be used for encryption.

This can be done through the BigQuery console, CLI, or API.

Reference: Links:

BigQuery and CMEK

Cloud KMS Documentation

Encrypting Data in BigQuery

Question: 349

You are designing a data mesh on Google Cloud by using Dataplex to manage data in BigQuery and Cloud Storage. You want to simplify data asset permissions. You are creating a customer virtual lake with two user groups:

- Data engineers, which require lull data lake access
- Analytic users, which require access to curated data

You need to assign access rights to these two groups. What should you do?

A. 1. Grant the dataplex.dataOwner role to the data engineer group on the customer data lake.

2. Grant the dataplex.dataReader role to the analytic user group on the customer curated zone.

B. 1. Grant the dataplex.dataReader role to the data engineer group on the customer data lake.

2. Grant the dataplex.dataOwner to the analytic user group on the customer curated zone.

C. 1. Grant the bigquery.dataownex role on BigQuery datasets and the storage.objectcreator role on Cloud Storage buckets to data engineers.

2. Grant the bigquery.dataViewer role on BigQuery datasets and the storage.objectViewer role on Cloud Storage buckets to analytic users.

D. 1. Grant the bigquery.dataViewer role on BigQuery datasets and the storage.objectviewer role on Cloud Storage buckets to data engineers.

2. Grant the bigquery.dataOwner role on BigQuery datasets and the storage.objectEditor role on Cloud Storage

buckets to analytic users.

Answer: A

Explanation:

When designing a data mesh on Google Cloud using Dataplex to manage data in BigQuery and Cloud Storage, it is essential to simplify data asset permissions while ensuring that each user group has the appropriate access levels. Here's why option A is the best choice:

Explanation:

Data Engineer Group:

Data engineers require full access to the data lake to manage and operate data assets comprehensively. Granting the dataplex.dataOwner role to the data engineer group on the customer data lake ensures they have the necessary permissions to create, modify, and delete data assets within the lake.

Analytic User Group:

Analytic users need access to curated data but do not require full control over all data assets. Granting the dataplex.dataReader role to the analytic user group on the customer curated zone provides read-only access to the curated data, enabling them to analyze the data without the ability to modify or delete it.

Steps to Implement:

Grant Data Engineer Permissions:

Assign the dataplex.dataOwner role to the data engineer group on the customer data lake to ensure full access and management capabilities.

Grant Analytic User Permissions:

Assign the dataplex.dataReader role to the analytic user group on the customer curated zone to provide read-only access to curated data.

Reference: Links:

Dataplex IAM Roles and Permissions

Managing Access in Dataplex

Question: 350

Your car factory is pushing machine measurements as messages into a Pub/Sub topic in your Google Cloud project. A Dataflow streaming job. that you wrote with the Apache Beam SDK, reads these messages, sends acknowledgment lo Pub/Sub. applies some custom business logic in a Doffs instance, and writes the result to BigQuery. You want to ensure that if your business logic fails on a message, the message will be sent to a Pub/Sub topic that you want to monitor for alerting purposes. What should you do?

A. Use an exception handling block in your Data Flow's Doffs code to push the messages that failed to be transformed through a side output and to a new Pub/Sub topic. Use Cloud Monitoring to monitor the topic/num_jnacked_messages_by_region metric on this new topic.

B. Enable retaining of acknowledged messages in your Pub/Sub pull subscription. Use Cloud Monitoring to monitor the subscription/num_retained_acked_messages metric on this subscription.

C. Enable dead lettering in your Pub/Sub pull subscription, and specify a new Pub/Sub topic as the dead letter topic. Use Cloud Monitoring to monitor the subscription/dead_letter_message_count metric on your pull subscription.

D. Create a snapshot of your Pub/Sub pull subscription. Use Cloud Monitoring to monitor the snapshot/numessages metric on this snapshot.

Answer: C

Explanation:

To ensure that messages failing to process in your Dataflow job are sent to a Pub/Sub topic for monitoring and alerting, the best approach is to use Pub/Sub's dead-letter topic feature. Here's why option C is the best choice:

Explanation:

Dead-Letter Topic:

Pub/Sub's dead-letter topic feature allows messages that fail to be processed successfully to be redirected to a specified topic. This ensures that these messages are not lost and can be reviewed for debugging and alerting purposes.

Monitoring and Alerting:

By specifying a new Pub/Sub topic as the dead-letter topic, you can use Cloud Monitoring to track metrics such as subscription/dead_letter_message_count, providing visibility into the number of failed messages.

This allows you to set up alerts based on these metrics to notify the appropriate teams when failures occur.

Steps to Implement:

Enable Dead-Letter Topic:

Configure your Pub/Sub pull subscription to enable dead lettering and specify the new Pub/Sub topic for dead-letter messages.

Set Up Monitoring:

Use Cloud Monitoring to monitor the subscription/dead_letter_message_count metric on your pull subscription.

Configure alerts based on this metric to notify the team of any processing failures.

Reference: Links:

Pub/Sub Dead Letter Policy

Cloud Monitoring with Pub/Sub

Question: 351

You are migrating your on-premises data warehouse to BigQuery. One of the upstream data sources resides on a MySQL database that runs in your on-premises data center with no public IP addresses. You want to ensure that the data ingestion into BigQuery is done securely and does not go through the public internet. What should you do?

A. Update your existing on-premises ETL tool to write to BigQuery by using the BigQuery Open Database Connectivity (ODBC) driver. Set up the proxy parameter in the Simba. googlebigqueryodbc. ini tile to point to your data center's NAT gateway.

B. Use Datastream to replicate data from your on-premises MySQL database to BigQuery. Gather Datastream public IP addresses of the Google Cloud region that will be used to set up the stream. Add those IP addresses to the firewall allowlist of your on-premises data center.

Use IP Allovlisting as the connectivity method and Server-only as the encryption type when setting up the connection profile in Datastream.

C. Use Datastream to replicate data from your on-premises MySQL database to BigQuery. Use Forward-SSH tunnel as the connectivity method to establish a secure tunnel between Datastream and your on-premises MySQL database

through a tunnel server in your on-premises data center. Use None as the encryption type when setting up the connection profile in Datastream.

D. Use Datastream to replicate data from your on-premises MySQL database to BigQuery. Set up Cloud Interconnect between your on- premises data center and Google Cloud. Use Private connectivity as the connectivity method and allocate an IP address range within your VPC network to the Datastream connectivity configuration. Use Server-only as the encryption type when setting up the connection profile in Datastream.

Answer: D

Explanation:

To securely ingest data from an on-premises MySQL database into BigQuery without routing through the public internet, using Datastream with Private connectivity over Cloud Interconnect is the best approach. Here's why:

Explanation:

Datastream for Data Replication:

Datastream provides a managed service for data replication from various sources, including onpremises databases, to Google Cloud services like BigQuery.

Cloud Interconnect:

Cloud Interconnect establishes a private connection between your on-premises data center and Google Cloud, ensuring that data transfer occurs over a secure, private network rather than the public internet.

Private Connectivity:

Using Private connectivity with Datastream leverages the established Cloud Interconnect to securely connect your on-premises MySQL database with Google Cloud. This method ensures that the data does not traverse the public internet.

Encryption:

Using Server-only encryption ensures that data is encrypted in transit between Datastream and BigQuery, adding an extra layer of security.

Steps to Implement:

Set Up Cloud Interconnect:

Establish a Cloud Interconnect between your on-premises data center and Google Cloud to create a private connection.

Configure Datastream:

Set up Datastream to use Private connectivity as the connection method and allocate an IP address range within your

VPC network.

Use Server-only encryption to ensure secure data transfer.

Create Connection Profile:

Create a connection profile in Datastream to define the connection parameters, including the use of Cloud Interconnect and Private connectivity.

Reference: Links:

Datastream Documentation

Cloud Interconnect Documentation

Setting Up Private Connectivity in Datastream

Question: 352

The data analyst team at your company uses BigQuery for ad-hoc queries and scheduled SQL pipelines in a Google Cloud project with a slot reservation of 2000 slots. However, with the recent introduction of hundreds of new non time-sensitive SQL pipelines, the team is encountering frequent quota errors. You examine the logs and notice that approximately 1500 queries are being triggered concurrently during peak time. You need to resolve the concurrency issue. What should you do?

A. Update SQL pipelines and ad-hoc queries to run as interactive query jobs.

B. Increase the slot capacity of the project with baseline as 0 and maximum reservation size as 3000.

C. Update SOL pipelines to run as a batch query, and run ad-hoc queries as interactive query jobs.

D. Increase the slot capacity of the project with baseline as 2000 and maximum reservation size as 3000.

Answer: C

Explanation:

To resolve the concurrency issue in BigQuery caused by the introduction of hundreds of non-timesensitive SQL pipelines, the best approach is to differentiate the types of queries based on their urgency and resource requirements. Here's why option C is the best choice:

Explanation:

SQL Pipelines as Batch Queries:

Batch queries in BigQuery are designed for non-time-sensitive operations. They run in a lower priority queue and do not consume slots immediately, which helps to reduce the overall slot consumption during peak times.

By converting non-time-sensitive SQL pipelines to batch queries, you can significantly alleviate the pressure on slot reservations.

Ad-Hoc Queries as Interactive Queries:

Interactive queries are prioritized to run immediately and are suitable for ad-hoc analysis where users expect quick results.

Running ad-hoc queries as interactive jobs ensures that analysts can get their results without delay, improving productivity and user satisfaction.

Concurrency Management:

This approach helps balance the workload by leveraging BigQuery's ability to handle different types of queries efficiently, reducing the likelihood of encountering quota errors due to slot exhaustion.

Steps to Implement:

Identify Non-Time-Sensitive Pipelines:

Review and identify SQL pipelines that are not time-critical and can be executed as batch jobs.

Update Pipelines to Batch Queries:

Modify these pipelines to run as batch queries. This can be done by setting the priority of the query job to BATCH.

Ensure Ad-Hoc Queries are Interactive:

Ensure that all ad-hoc queries are submitted as interactive jobs, allowing them to run with higher priority and immediate slot allocation.

Reference: Links:

BigQuery Batch Queries

BigQuery Slot Allocation and Management

Question: 353

You have several different unstructured data sources, within your on-premises data center as well as in the cloud. The data is in various formats, such as Apache Parquet and CSV. You want to centralize this data in Cloud Storage. You need

to set up an object sink for your data that allows you to use your own encryption keys. You want to use a GUI-based solution. What should you do?

A. Use Cloud Data Fusion to move files into Cloud Storage.

B. Use Storage Transfer Service to move files into Cloud Storage.

C. Use Dataflow to move files into Cloud Storage.

D. Use BigQuery Data Transfer Service to move files into BigQuery.

Answer: A

Explanation:

To centralize unstructured data from various sources into Cloud Storage using a GUI-based solution while allowing the use of your own encryption keys, Cloud Data Fusion is the most suitable option. Here's why:

Explanation:

Cloud Data Fusion:

Cloud Data Fusion is a fully managed, cloud-native data integration service that helps in building and managing ETL pipelines with a visual interface.

It supports a wide range of data sources and formats, including Apache Parquet and CSV, and provides a user-friendly GUI for pipeline creation and management.

Custom Encryption Keys:

Cloud Data Fusion allows the use of customer-managed encryption keys (CMEK) for data encryption, ensuring that your data is securely stored according to your encryption policies.

Centralizing Data:

Cloud Data Fusion simplifies the process of moving data from on-premises and cloud sources into Cloud Storage, providing a centralized repository for your unstructured data.

Steps to Implement:

Set Up Cloud Data Fusion:

Deploy a Cloud Data Fusion instance and configure it to connect to your various data sources.

Create ETL Pipelines:

Use the GUI to create data pipelines that extract data from your sources and load it into Cloud Storage. Configure the pipelines to use your custom encryption keys.

Run and Monitor Pipelines:

Execute the pipelines and monitor their performance and data movement through the Cloud Data Fusion dashboard.

Reference: Links:

Cloud Data Fusion Documentation

Using Customer-Managed Encryption Keys (CMEK)

Question: 354

You created an analytics environment on Google Cloud so that your data scientist team can explore data without impacting the on-premises Apache Hadoop solution. The data in the on-premises Hadoop Distributed File System (HDFS) cluster is in Optimized Row Columnar (ORC) formatted files with multiple columns of Hive partitioning. The data scientist team needs to be able to explore the data in a similar way as they used the on-premises HDFS cluster with SQL on the Hive query engine. You need to choose the most cost-effective storage and processing solution. What should you do?

A. Import the ORC files lo Bigtable tables for the data scientist team.

B. Import the ORC files to BigOuery tables for the data scientist team.

C. Copy the ORC files on Cloud Storage, then deploy a Dataproc cluster for the data scientist team.

D. Copy the ORC files on Cloud Storage, then create external BigQuery tables for the data scientist team.

Answer: D

Explanation:

Explore ORC formatted files with Hive partitioning.

Mimic the SQL on Hive query engine experience.

Cost-effective storage and processing.

Avoid impacting the on-premises Hadoop solution.

Let's analyze the options:

Option A (Import to Bigtable):Bigtable is a NoSQL database, not suited for SQL-based exploration of ORC files or Hive-style partitioning directly. This would require significant data transformation and a different query paradigm. Not cost-effective for this use case.

Option B (Import to BigQuery native tables):Importing data into BigQuery native storage is an option. BigQuery can load ORC files. This provides excellent query performance. However, it involves an ETL step (importing) and storage costs for the datawithin BigQuery, which might be higher than keeping it in its original format on Cloud Storage if query patterns are exploratory and not extremely frequent on all data.

Option C (Copy to Cloud Storage, deploy Dataproc):Dataproc allows you to run Hadoop/Spark (and thus Hive) clusters on Google Cloud. This would provide a very similar experience ("SQL on the Hive query engine"). However, running a persistent Dataproc cluster incurs compute costs for the cluster nodes, even when not actively querying. While ephemeral clusters are possible, it adds operational overhead for exploratory queries. Storage on Cloud Storage is cost-effective.

Option D (Copy to Cloud Storage, create external BigQuery tables):This is often the most costeffective and straightforward solution for this scenario.

Cost-effective Storage:Cloud Storage is a low-cost option for storing files like ORC.

SQL Interface:BigQuery provides a familiar SQL interface.

External Tables:BigQuery can query data directly from Cloud Storage (including ORC files) using external tables. This avoids the need to load data into BigQuery's managed storage, saving on storage costs and ETL effort.

Hive Partitioning:BigQuery external tables support Hive partitioning layouts. When you define the external table, you can specify the partitioning scheme, and BigQuery will use partition pruning to scan only relevant partitions, improving performance and reducing costs for queries that filter on partition keys. This directly mimics the Hive experience.

Processing Cost:You only pay for the data scanned by BigQuery queries, which aligns with exploratory analysis.

Comparing D with B: External tables are generally more cost-effective for storage and initial setup if the data is already in ORC and an ETL process into BigQuery native storage is to be avoided. Query performance might be slightly less than native tables but is often sufficient for exploration, especially with partitioning. Comparing D with C: BigQuery external tables are serverless, meaning no cluster to

manage or pay for when idle. Dataproc requires managing and paying for a cluster. For exploration, the serverless nature of BigQuery is usually more cost-effective.

Therefore, copying ORC files to Cloud Storage and using BigQuery external tables is the most costeffective solution that meets all requirements.

Reference:

Google Cloud Documentation: BigQuery > External data sources > Querying Cloud Storage data. "You can query data in Cloud Storage by using external tables or federated queries. External tables are tables that read data directly from

files in Cloud Storage."

Google Cloud Documentation: BigQuery > External data sources > Supported formats and compression types. ORC is a supported format.

Google Cloud Documentation: BigQuery > Creating and using tables > Creating external tables. "External tables let you query data stored in Cloud Storage as if it were a standardBigQuery table. You can use external tables to query data in various formats, including... ORC..."

Google Cloud Documentation: BigQuery > Creating and using tables > Querying partitioned external tables. "You can create an external table that is partitioned on Hive partitioning keys. When you query a Hive partitioned external table, BigQuery performs partition pruning to skip reading unnecessary partitions." This directly addresses the "Hive partitioning" and "explore data in a similar way" requirements.

Google Cloud Blog: "Choosing the right data processing option on GCP: BigQuery vs. Dataproc" (and similar articles) often highlight BigQuery external tables as a cost-effective way to query data in place on Cloud Storage, especially for data lake scenarios.

Question: 355

You are migrating your on-premises data warehouse to BigQuery. As part of the migration, you want to facilitate cross-team collaboration to get the most value out of the organization's dat

a. You need to design an architecture that would allow teams within the organization to securely publish, discover, and subscribe to read-only data in a self-service manner. You need to minimize costs while also maximizing data freshness What should you do?

A. Create authorized datasets to publish shared data in the subscribing team's project.

B. Create a new dataset for sharing in each individual team's project. Grant the subscribing team the bigquery. dataViewer role on the

dataset.

C. Use BigQuery Data Transfer Service to copy datasets to a centralized BigQuery project for sharing.

D. Use Analytics Hub to facilitate data sharing.

Answer: C

Explanation:

To provide a cost-effective storage and processing solution that allows data scientists to explore data similarly to using the on-premises HDFS cluster with SQL on the Hive query engine, deploying a Dataproc cluster is the best choice. Here's why:

Explanation:

Compatibility with Hive:

Dataproc is a fully managed Apache Spark and Hadoop service that provides native support for Hive, making it easy for data scientists to run SQL queries on the data as they would in an on-premises Hadoop environment.

This ensures that the transition to Google Cloud is smooth, with minimal changes required in the workflow.

Cost-Effective Storage:

Storing the ORC files in Cloud Storage is cost-effective and scalable, providing a reliable and durable storage solution that integrates seamlessly with Dataproc.

Cloud Storage allows you to store large datasets at a lower cost compared to other storage options.

Hive Integration:

Dataproc supports running Hive directly, which is essential for data scientists familiar with SQL on the Hive query engine.

This setup enables the use of existing Hive queries and scripts without significant modifications.

Steps to Implement:

Copy ORC Files to Cloud Storage:

Transfer the ORC files from the on-premises HDFS cluster to Cloud Storage, ensuring they are organized in a similar directory structure.

Deploy Dataproc Cluster:

Set up a Dataproc cluster configured to run Hive. Ensure that the cluster has access to the ORC files stored in Cloud Storage.

Configure Hive:

Configure Hive on Dataproc to read from the ORC files in Cloud Storage. This can be done by setting up external tables in Hive that point to the Cloud Storage location.

Provide Access to Data Scientists:

Grant the data scientist team access to the Dataproc cluster and the necessary permissions to interact with the Hive tables.

Reference: Links:

Dataproc Documentation

Hive on Dataproc

Google Cloud Storage Documentation

Question: 356

You have one BigQuery dataset which includes customers' street addresses. You want to retrieve all occurrences of street addresses from the dataset. What should you do?

A. Create a deep inspection job on each table in your dataset with Cloud Data Loss Prevention and create an inspection template that includes the STREET_ADDRESS infoType.

B. Create a de-identification job in Cloud Data Loss Prevention and use the masking transformation.

C. Write a SQL query in BigQuery by using REGEXP_CONTAINS on all tables in your dataset to find rows where the word "street" appears.

D. Create a discovery scan configuration on your organization with Cloud Data Loss Prevention and create an inspection template that includes the STREET_ADDRESS infoType.

Answer: A

Explanation:

To retrieve all occurrences of street addresses from a BigQuery dataset, the most effective and comprehensive method is to use Cloud Data Loss Prevention (DLP). Here's why option A is the best choice:

Explanation:

Cloud Data Loss Prevention (DLP):

Cloud DLP is designed to discover, classify, and protect sensitive information. It includes pre-defined infoTypes for various kinds of sensitive data, including street addresses.

Using Cloud DLP ensures thorough and accurate detection of street addresses based on advanced pattern recognition and contextual analysis.

Deep Inspection Job:

A deep inspection job allows you to scan entire tables for sensitive information.

By creating an inspection template that includes the STREET_ADDRESS infoType, you can ensure that all instances of street addresses are detected across your dataset.

Scalability and Accuracy:

Cloud DLP is scalable and can handle large datasets efficiently.

It provides a high level of accuracy in identifying sensitive data, reducing the risk of missing any occurrences.

Steps to Implement:

Set Up Cloud DLP:

Enable the Cloud DLP API in your Google Cloud project.

Create an Inspection Template:

Create an inspection template in Cloud DLP that includes the STREET_ADDRESS infoType.

Run Deep Inspection Jobs:

Create and run a deep inspection job for each table in your dataset using the inspection template.

Review the inspection job results to retrieve all occurrences of street addresses.

Reference: Links:

Cloud DLP Documentation

Creating Inspection Jobs

Question: 357

You are administering a BigQuery on-demand environment. Your business intelligence tool is submitting hundreds of queries each day that aggregate a large (50 TB) sales history fact table at the day and month levels. These queries have a slow response time and are exceeding cost expectations. You need to decrease response time, lower query costs, and minimize maintenance. What should you do?

A. Build materialized views on top of the sales table to aggregate data at the day and month level.

B. Build authorized views on top of the sales table to aggregate data at the day and month level.

C. Enable BI Engine and add your sales table as a preferred table.

D. Create a scheduled query to build sales day and sales month aggregate tables on an hourly basis.

Answer: A

Explanation:

To improve response times and reduce costs for frequent queries aggregating a large sales history fact table, materialized views are a highly effective solution. Here's why option A is the best choice:

Explanation:

Materialized Views:

Materialized views store the results of a query physically and update them periodically, offering faster query responses for frequently accessed data.

They are designed to improve performance for repetitive and expensive aggregation queries by precomputing the results.

Efficiency and Cost Reduction:

By building materialized views at the day and month level, you significantly reduce the computation required for each query, leading to faster response times and lower query costs.

Materialized views also reduce the need for on-demand query execution, which can be costly when

dealing with large datasets.

Minimized Maintenance:

Materialized views in BigQuery are managed automatically, with updates handled by the system, reducing the maintenance burden on your team.

Steps to Implement:

Identify Aggregation Queries:

Analyze the existing queries to identify common aggregation patterns at the day and month levels.

Create Materialized Views:

Create materialized views in BigQuery for the identified aggregation patterns. For example

CREATE MATERIALIZED VIEW project.dataset.sales_daily_summary AS

```
SELECT
  DATE(transaction_time) AS day,
SUM(amount) AS total_sales
FROM
  project.dataset.sales
GROUP BY
  day;

CREATE MATERIALIZED VIEW project.dataset.sales_monthly_summary AS
SELECT
  EXTRACT(YEAR FROM transaction_time) AS year,
  EXTRACT(MONTH FROM transaction_time) AS month,
  SUM(amount) AS total_sales
FROM
  project.dataset.sales
GROUP BY
  year, month;
```

Query Using Materialized Views:

Update existing queries to use the materialized views instead of directly querying the base table.

Reference: Links:

BigQuery Materialized Views

Optimizing Query Performance

Question: 358

You have a Standard Tier Memorystore for Redis instance deployed in a production environment. You need to simulate a Redis instance failover in the most accurate disaster recovery situation, and ensure that the failover has no impact on production dat

a. What should you do?

A. Create a Standard Tier Memorystore for Redis instance in a development environment. Initiate a manual failover by using the force-data-loss data protection mode.

B. Initiate a manual tailover by using the limited-data-loss data protection mode to the Memorystore for Redis instance in the production environment.

C. Increase one replica to Redis instance in production environment. Initiate a manual failover by using the force-data-loss data protection mode.

D. Create a Standard Tier Memorystore for Redis instance in the development environment. Initiate a manual failover by using the limited-data-loss data protection mode.

Answer: D

Explanation:

To simulate a Redis instance failover in a production-like environment without impacting production data, the best approach is to use a development environment. Here's why option D is the best choice:

Explanation:

Standard Tier Memorystore for Redis:

The Standard Tier provides high availability and automatic failover capabilities. It's suitable for testing failover scenarios in a controlled environment.

Development Environment:

Using a development environment ensures that any potential data loss or impact from the failover simulation does not affect production data, maintaining the integrity and availability of the production system.

Limited-Data-Loss Mode:

The limited-data-loss mode for manual failover ensures that data loss is minimized during the failover process, making it a realistic simulation of a production failover scenario.

Steps to Implement:

Create a Development Environment:

Set up a development environment with a Standard Tier Memorystore for Redis instance that mirrors the configuration of your production instance.

Initiate Manual Failover:

Initiate a manual failover using the limited-data-loss data protection mode to simulate a failover scenario:

gcloud redis instances failover INSTANCE_ID --data-protection-mode=limited-data-loss

Verify Failover:

Monitor and verify the failover process to ensure it behaves as expected, simulating the disaster recovery scenario accurately.

Reference: Links:

Memorystore for Redis Documentation

Manual Failover in Memorystore

Question: 359

You are planning to use Cloud Storage as pad of your data lake solution. The Cloud Storage bucket will contain objects ingested from external systems. Each object will be ingested once, and the access patterns of individual objects will be random. You want to minimize the cost of storing and retrieving these objects. You want to ensure that any cost optimization efforts are transparent to the users and applications. What should you do?

A. Create a Cloud Storage bucket with Autoclass enabled.

B. Create a Cloud Storage bucket with an Object Lifecycle Management policy to transition objects from Standard to Coldline storage class if an object age reaches 30 days.

C. Create a Cloud Storage bucket with an Object Lifecycle Management policy to transition objects from Standard to Coldline storage class if an object is not live.

D. Create two Cloud Storage buckets. Use the Standard storage class for the first bucket, and use the Coldline storage class for the second bucket. Migrate objects from the first bucket to the second bucket after 30 days.

Answer: A

Explanation:

To minimize the cost of storing and retrieving objects in a Cloud Storage bucket while ensuring that cost optimization efforts are transparent to the users and applications, enabling Autoclass is the best approach. Here's why:

Explanation:

Autoclass Feature:

Autoclass automatically transitions objects between different storage classes (Standard, Nearline, Coldline, and Archive) based on their access patterns.

It ensures that frequently accessed data is kept in lower-latency, higher-cost storage classes and infrequently accessed data is moved to higher-latency, lower-cost storage classes.

Cost Optimization:

Autoclass optimizes storage costs by automatically moving objects to the most cost-effective storage class based on actual usage patterns, without manual intervention.

This feature ensures that objects are stored in the most economical class appropriate for their access frequency, reducing storage costs over time.

Transparency to Users:

The transition of objects between storage classes is handled automatically by Cloud Storage, making the process transparent to users and applications.

Users and applications interact with the objects in the same way, regardless of the underlying storage class, ensuring seamless access.

Steps to Implement:

Create a Cloud Storage Bucket:

When creating a new Cloud Storage bucket, enable the Autoclass feature.

Configure Autoclass:

Autoclass configuration is typically a straightforward process in the Google Cloud Console, where you enable it during bucket creation.

Monitor and Adjust:

Monitor the storage and access patterns through the Google Cloud Console to ensure that Autoclass is optimizing costs as expected.

Reference: Links:

Google Cloud Storage Autoclass

Question: 360

You have an upstream process that writes data to Cloud Storage. This data is then read by an Apache Spark job that runs on Dataproc. These jobs are run in the us-central1 region, but the data could be stored anywhere in the United States. You need to have a recovery process in place in case of a catastrophic single region failure. You need an approach with a maximum of 15 minutes of data loss (RPO=15 mins). You want to ensure that there is minimal latency when reading the dat

a. What should you do?

A. 1. Create a dual-region Cloud Storage bucket in the us-central1 and us-south1 regions.

2. Enable turbo replication.

3. Run the Dataproc cluster in a zone in the us-central1 region, reading from the bucket in the us- south1 region.

4. In case of a regional failure, redeploy your Dataproc duster to the us-south1 region and continue reading from the same bucket.

B. 1. Create a dual-region Cloud Storage bucket in the us-central1 and us-south1 regions.

2. Enable turbo replication.

3. Run the Dataproc cluster in a zone in the us-central1 region, reading from the bucket in the same region.

4. In case of a regional failure, redeploy the Dataproc clusters to the us-south1 region and read from the same bucket.

C. 1. Create a Cloud Storage bucket in the US multi-region.

2. Run the Dataproc cluster in a zone in the ua-central1 region, reading data from the US multiregion bucket.

3. In case of a regional failure, redeploy the Dataproc cluster to the us-central2 region and continue reading from the same bucket.

D. 1. Create two regional Cloud Storage buckets, one in the us-central1 region and one in the us- south1 region.

2. Have the upstream process write data to the us-central1 bucket. Use the Storage Transfer Service to copy data hourly from the us-central1 bucket to the us-south1 bucket.

3. Run the Dataproc cluster in a zone in the us-central1 region, reading from the bucket in that region.

4. In case of regional failure, redeploy your Dataproc clusters to the us-south1 region and read from the bucket

in that region instead.

Answer: B

Explanation:

To ensure data recovery with minimal data loss and low latency in case of a single region failure, the best approach is to use a dual-region bucket with turbo replication. Here's why option B is the best choice:

Explanation:

Dual-Region Bucket:

A dual-region bucket provides geo-redundancy by replicating data across two regions, ensuring high availability and resilience against regional failures.

The chosen regions (us-central1 and us-south1) provide geographic diversity within the United States.

Turbo Replication:

Turbo replication ensures that data is replicated between the two regions within 15 minutes, meeting the Recovery Point Objective (RPO) of 15 minutes.

This minimizes data loss in case of a regional failure.

Running Dataproc Cluster:

Running the Dataproc cluster in the same region as the primary data storage (us-central1) ensures minimal latency for normal operations.

In case of a regional failure, redeploying the Dataproc cluster to the secondary region (us-south1) ensures continuity with minimal data loss.

Steps to Implement:

Create a Dual-Region Bucket:

Set up a dual-region bucket in the Google Cloud Console, selecting us-central1 and us-south1 regions.

Enable turbo replication to ensure rapid data replication between the regions.

Deploy Dataproc Cluster:

Deploy the Dataproc cluster in the us-central1 region to read data from the bucket located in the same region for optimal performance.

Set Up Failover Plan:

Plan for redeployment of the Dataproc cluster to the us-south1 region in case of a failure in the us- central1 region.

Ensure that the failover process is well-documented and tested to minimize downtime and data loss.

Reference: Links:

Google Cloud Storage Dual-Region

Turbo Replication in Google Cloud Storage

Dataproc Documentation

Question: 361

Different teams in your organization store customer and performance data in BigOuery. Each team needs to keep full control of their collected data, be able to query data within their projects, and be able to exchange their data with other teams. You need to implement an organization-wide solution, while minimizing operational tasks and costs. What should you do?

A. Create a BigQuery scheduled query to replicate all customer data into team projects.

B. Enable each team to create materialized views of the data they need to access in their projects.

C. Ask each team to publish their data in Analytics Hub. Direct the other teams to subscribe to them.

D. Ask each team to create authorized views of their data. Grant the biquery. jobUser role to each team.

Answer: C

Explanation:

To enable different teams to manage their own data while allowing data exchange across the organization, using Analytics Hub is the best approach. Here's why option C is the best choice:

Explanation:

Analytics Hub:

Analytics Hub allows teams to publish their data as data exchanges, making it easy for other teams to discover and subscribe to the data they need.

This approach maintains each team's control over their data while facilitating easy and secure data sharing across the

organization.

Data Publishing and Subscribing:

Teams can publish datasets they control, allowing them to manage access and updates independently.

Other teams can subscribe to these published datasets, ensuring they have access to the latest data without duplicating efforts.

Minimized Operational Tasks and Costs:

This method reduces the need for complex replication or data synchronization processes, minimizing operational overhead.

By centralizing data sharing through Analytics Hub, it also reduces storage costs associated with duplicating large datasets.

Steps to Implement:

Set Up Analytics Hub:

Enable Analytics Hub in your Google Cloud project.

Provide training to teams on how to publish and subscribe to data exchanges.

Publish Data:

Each team publishes their datasets in Analytics Hub, configuring access controls and metadata as needed.

Subscribe to Data:

Teams that need access to data from other teams can subscribe to the relevant data exchanges, ensuring they always have up-to-date data.

Reference: Links:

Analytics Hub Documentation

Publishing Data in Analytics Hub

Subscribing to Data in Analytics Hub

Question: 362

You are deploying a batch pipeline in Dataflow. This pipeline reads data from Cloud Storage, transforms the data, and then writes the data into BigQuory. The security team has enabled an organizational constraint in Google Cloud, requiring

all Compute Engine instances to use only internal IP addresses and no external IP addresses. What should you do?

A. Ensure that the firewall rules allow access to Cloud Storage and BigQuery. Use Dataflow with only internal IPs.

B. Ensure that your workers have network tags to access Cloud Storage and BigQuery. Use Dataflow with only internal IP addresses.

C. Create a VPC Service Controls perimeter that contains the VPC network and add Dataflow. Cloud Storage, and BigQuery as allowed services in the perimeter. Use Dataflow with only internal IP addresses.

D. Ensure that Private Google Access is enabled in the subnetwork. Use Dataflow with only internal IP addresses.

Answer: D

Explanation:

To deploy a batch pipeline in Dataflow that adheres to the organizational constraint of using only internal IP addresses, ensuring Private Google Access is the most effective solution. Here's why option D is the best choice:

Explanation:

Private Google Access:

Private Google Access allows resources in a VPC network that do not have external IP addresses to access Google APIs and services through internal IP addresses.

This ensures compliance with the organizational constraint of using only internal IPs while allowing Dataflow to access Cloud Storage and BigQuery.

Dataflow with Internal IPs:

Dataflow can be configured to use only internal IP addresses for its worker nodes, ensuring that no external IP addresses are assigned.

This configuration ensures secure and compliant communication between Dataflow, Cloud Storage, and BigQuery.

Firewall and Network Configuration:

Enabling Private Google Access requires ensuring the correct firewall rules and network configurations to allow

internal traffic to Google Cloud services.

Steps to Implement:

Enable Private Google Access:

Enable Private Google Access on the subnetwork used by the Dataflow pipeline

gcloud compute networks subnets update [SUBNET_NAME] \

- -region [REGION] \
- -enable-private-ip-google-access

Configure Dataflow:

Configure the Dataflow job to use only internal IP addresses

gcloud dataflow jobs run [JOB_NAME] \

- -region [REGION] \
- -network [VPC_NETWORK] \
- -subnetwork [SUBNETWORK] \
- -no-use-public-ips

Verify Access:

Ensure that firewall rules allow the necessary traffic from the Dataflow workers to Cloud Storage and BigQuery using internal IPs.

Reference: Links:

Private Google Access Documentation

Configuring Dataflow to Use Internal IPs

VPC Firewall Rules

Question: 363

You currently use a SQL-based tool to visualize your data stored in BigQuery The data visualizations require the use of outer joins and analytic functions. Visualizations must be based on data that is no

less than 4 hours old. Business users are complaining that the visualizations are too slow to generate. You want to improve the performance of the visualization queries while minimizing the maintenance overhead of the data preparation pipeline. What should you do?

A. Create materialized views with the allow_non_incremental_definition option set to true for the visualization queries. Specify the max_3taleness parameter to 4 hours and the enable_refresh parameter to true. Reference: the materialized views in the data visualization tool.

B. Create views for the visualization queries. Reference: the views in the data visualization tool.

C. Create materialized views for the visualization queries. Use the incremental updates capability of BigQuery materialized views to handle

changed data automatically. Reference: the materialized views in the data visualization tool.

D. Create a Cloud Function instance to export the visualization query results as parquet files to a Cloud Storage bucket. Use Cloud Scheduler

to trigger the Cloud Function every 4 hours. Reference: the parquet files in the data visualization tool.

Answer: C

Explanation:

To improve the performance of visualization queries while minimizing maintenance overhead, using materialized views is the most effective solution. Here's why option C is the best choice:

Explanation:

Materialized Views:

Materialized views store the results of a query physically, allowing for faster access compared to regular views which execute the query each time it is accessed.

They can be automatically refreshed to reflect changes in the underlying data.

Incremental Updates:

The incremental updates capability of BigQuery materialized views ensures that only the changed data is processed during refresh operations, significantly improving performance and reducing computation costs.

This feature helps maintain up-to-date data in the materialized view with minimal processing time, which is crucial for data that needs to be no less than 4 hours old.

Performance and Maintenance:

By using materialized views, you can pre-compute and store the results of complex queries involving outer joins and analytic functions, resulting in faster query performance for data visualizations.

This approach also reduces the maintenance overhead, as BigQuery handles the incremental updates and refreshes automatically.

Steps to Implement:

Create Materialized Views:

Define materialized views for the visualization queries with the necessary configurations

```
CREATE MATERIALIZED VIEW project.dataset.view_name
AS
SELECT ...
FROM ...
WHERE ...
```

Enable Incremental Updates:

Ensure that the materialized views are set up to handle incremental updates automatically.

Reference: in Visualization Tool:

Update the data visualization tool to reference the materialized views instead of running the original queries directly.

Reference: Links:

BigQuery Materialized Views

Optimizing Query Performance

Question: 364

Your company's customer_order table in BigOuery stores the order history for 10 million customers, with a table size of 10 PB. You need to create a dashboard for the support team to view the order

history. The dashboard has two filters, countryname and username. Both are string data types in the BigQuery table. When a filter is applied, the dashboard fetches the order history from the table and displays the query results. However, the dashboard is slow to show the results when applying the filters to the following query:

```
SELECT date, order, status FROM customer^order
WHERE country - *<ccuntry_name>* AMI username = *<usernan®>*
```

How should you redesign the BigQuery table to support faster access?

A. Cluster the table by country field, and partition by username field.

B. Partition the table by country and username fields.

C. Cluster the table by country and username fields

D. Partition the table by _PARTITIONTIME.

Answer: C

Explanation:

To improve the performance of querying a large BigQuery table with filters on countryname and username, clustering the table by these fields is the most effective approach. Here's why option C is the best choice:

Explanation:

Clustering in BigQuery:

Clustering organizes data based on the values in specified columns. This can significantly improve query performance by reducing the amount of data scanned during query execution.

Clustering by countryname and username means that data is physically sorted and stored together based on these fields, allowing BigQuery to quickly locate and read only the relevant data for queries using these filters.

Filter Efficiency:

With the table clustered by countryname and username, queries that filter on these columns can benefit from efficient data retrieval, reducing the amount of data processed and speeding up query

execution.

This directly addresses the performance issue of the dashboard queries that apply filters on these fields.

Steps to Implement:

Redesign the Table:

Create a new table with clustering on countryname and username:

```
CREATE TABLE project.dataset.new_table

CLUSTER BY countryname, username AS

SELECT * FROM project.dataset.customer_order;
```

Migrate Data:

Transfer the existing data from the original table to the new clustered table.

Update Queries:

Modify the dashboard queries to reference the new clustered table.

Reference: Links:

BigQuery Clustering Documentation

Optimizing Query Performance

Question: 365

One of your encryption keys stored in Cloud Key Management Service (Cloud KMS) was exposed. You need to re-encrypt all of your CMEK-protected Cloud Storage data that used that key. and then delete the compromised key. You also want to reduce the risk of objects getting written without customer-managed encryption key (CMEK protection in the future. What should you do?

A. Rotate the Cloud KMS key version. Continue to use the same Cloud Storage bucket.

B. Create a new Cloud KMS key. Set the default CMEK key on the existing Cloud Storage bucket to the new one.

C. Create a new Cloud KMS key. Create a new Cloud Storage bucket. Copy all objects from the old bucket to the new one bucket while specifying the new Cloud KMS key in the copy command.

D. Create a new Cloud KMS key. Create a new Cloud Storage bucket configured to use the new key as the default CMEK key. Copy all objects from the old bucket to the new bucket without specifying a key.

Answer: C

Explanation:

To re-encrypt all of your CMEK-protected Cloud Storage data after a key has been exposed, and to ensure future writes are protected with a new key, creating a new Cloud KMS key and a new Cloud Storage bucket is the best approach. Here's why option C is the best choice:

Explanation:

Re-encryption of Data:

By creating a new Cloud Storage bucket and copying all objects from the old bucket to the new bucket while specifying the new Cloud KMS key, you ensure that all data is re-encrypted with the new key.

This process effectively re-encrypts the data, removing any dependency on the compromised key.

Ensuring CMEK Protection:

Creating a new bucket and setting the new CMEK as the default ensures that all future objects written to the bucket are automatically protected with the new key.

This reduces the risk of objects being written without CMEK protection.

Deletion of Compromised Key:

Once the data has been copied and re-encrypted, the old key can be safely deleted from Cloud KMS, eliminating the risk associated with the compromised key.

Steps to Implement:

Create a New Cloud KMS Key:

Create a new encryption key in Cloud KMS to replace the compromised key.

Create a New Cloud Storage Bucket:

Create a new Cloud Storage bucket and set the default CMEK to the new key.

Copy and Re-encrypt Data:

Use the gsutil tool to copy data from the old bucket to the new bucket while specifying the new CMEK key:

```
gsutil -o "GSUtil:gs_json_api_version=2" cp -r gs://old-bucket/* gs://new-bucket/
```

Delete the Old Key:

After ensuring all data is copied and re-encrypted, delete the compromised key from Cloud KMS.

Reference: Links:

Cloud KMS Documentation

Cloud Storage Encryption

Re-encrypting Data in Cloud Storage

Question: 366

You are designing the architecture of your application to store data in Cloud Storage. Your application consists of pipelines that read data from a Cloud Storage bucket that contains raw data, and write the data to a second bucket after processing. You want to design an architecture with Cloud Storage resources that are capable of being resilient if a Google Cloud regional failure occurs. You want to minimize the recovery point objective (RPO) if a failure occurs, with no impact on applications that use the stored dat

a. What should you do?

A. Adopt two regional Cloud Storage buckets, and update your application to write the output on both buckets.

B. Adopt multi-regional Cloud Storage buckets in your architecture.

C. Adopt two regional Cloud Storage buckets, and create a daily task to copy from one bucket to the other.

D. Adopt a dual-region Cloud Storage bucket, and enable turbo replication in your architecture.

Answer: D

Explanation:

To ensure resilience and minimize the recovery point objective (RPO) with no impact on applications, using a dual-region bucket with turbo replication is the best approach. Here's why option D is the best choice:

Explanation:

Dual-Region Buckets:

Dual-region buckets store data redundantly across two distinct geographic regions, providing high availability and durability.

This setup ensures that data remains available even if one region experiences a failure.

Turbo Replication:

Turbo replication ensures that data is replicated between the two regions within 15 minutes, aligning with the requirement to minimize the recovery point objective (RPO).

This feature provides near real-time replication, significantly reducing the risk of data loss.

No Impact on Applications:

Applications continue to access the dual-region bucket without any changes, ensuring seamless operation even during a regional failure.

The dual-region setup transparently handles failover, providing uninterrupted access to data.

Steps to Implement:

Create a Dual-Region Bucket:

Create a dual-region Cloud Storage bucket in the Google Cloud Console, selecting appropriate regions (e.g., us-central1 and us-east1).

Enable Turbo Replication:

Enable turbo replication to ensure rapid data replication between the selected regions.

Configure Applications:

Ensure that applications read and write to the dual-region bucket, benefiting from its high availability and durability.

Test Failover:

Simulate a regional failure to verify that the dual-region bucket and turbo replication meet the required RPO and ensure data resilience.

Reference: Links:

Google Cloud Storage Dual-Region

Turbo Replication in Google Cloud Storage

Question: 367

You are using Workflows to call an API that returns a 1 KB JSON response, apply some complex business logic on this response, wait for the logic to complete, and then perform a load from a Cloud Storage file to BigQuery. The Workflows standard library does not have sufficient capabilities to perform your complex logic, and you want to use Python's standard library instead. You want to optimize your workflow for simplicity and speed of execution. What should you do?

A. Invoke a Cloud Function instance that uses Python to apply the logic on your JSON file.

B. Invoke a subworkflow in Workflows to apply the logic on your JSON file.

C. Create a Cloud Composer environment and run the logic in Cloud Composer.

D. Create a Dataproc cluster, and use PySpark to apply the logic on your JSON file.

Answer: A

Explanation:

Question: 368

You are using BigQuery with a regional dataset that includes a table with the daily sales volumes. This table is updated multiple times per day. You need to protect your sales table in case of regional failures with a recovery point objective (RPO) of less than 24 hours, while keeping costs to a minimum. What should you do?

A. Schedule a daily BigQuery snapshot of the table.

B. Schedule a daily export of the table to a Cloud Storage dual or multi-region bucket.

C. Schedule a daily copy of the dataset to a backup region.

D. Modify ETL job to load the data into both the current and another backup region.

Answer: A

Explanation:

To apply complex business logic on a JSON response using Python's standard library within a Workflow, invoking a Cloud Function is the most efficient and straightforward approach. Here's why option A is the best choice:

Explanation:

Cloud Functions:

Cloud Functions provide a lightweight, serverless execution environment for running code in response to events. They support Python and can easily integrate with Workflows.

This approach ensures simplicity and speed of execution, as Cloud Functions can be invoked directly from a Workflow and handle the complex logic required.

Flexibility and Simplicity:

Using Cloud Functions allows you to leverage Python's extensive standard library and ecosystem, making it easier to implement and maintain the complex business logic.

Cloud Functions abstract the underlying infrastructure, allowing you to focus on the application logic without worrying about server management.

Performance:

Cloud Functions are optimized for fast execution and can handle the processing of the JSON response efficiently.

They are designed to scale automatically based on demand, ensuring that your workflow remains performant.

Steps to Implement:

Write the Cloud Function:

Develop a Cloud Function in Python that processes the JSON response and applies the necessary business logic.

Deploy the function to Google Cloud.

Invoke Cloud Function from Workflow:

Modify your Workflow to call the Cloud Function using an HTTP request or Google Cloud Function connector.

steps:

- callCloudFunction:

call: http.post

args:

url: https://REGION-PROJECT_ID.cloudfunctions.net/FUNCTION_NAME

body:

key: value

Process Results:

Handle the response from the Cloud Function and proceed with the next steps in the Workflow, such as loading data into BigQuery.

Reference: Links:

Google Cloud Functions Documentation

Using Workflows with Cloud Functions

Workflows Standard Library

Question: 369

You have two projects where you run BigQuery jobs:

- One project runs production jobs that have strict completion time SLAs. These are high priority jobs that must have the required compute resources available when needed. These jobs generally never go below a 300 slot utilization, but occasionally spike up an additional 500 slots.

- The other project is for users to run ad-hoc analytical queries. This project generally never uses more than 200 slots at a time. You want these ad-hoc queries to be billed based on how much data users scan rather than by slot capacity.

You need to ensure that both projects have the appropriate compute resources available. What should you do?

A. Create a single Enterprise Edition reservation for both projects. Set a baseline of 300 slots. Enable autoscaling up to 700 slots.

B. Create two reservations, one for each of the projects. For the SLA project, use an Enterprise Edition with a baseline of 300 slots and enable autoscaling up to 500 slots. For the ad-hoc project, configure on-demand billing.

C. Create two Enterprise Edition reservations, one for each of the projects. For the SLA project, set a baseline of 300 slots and enable

autoscaling up to 500 slots. For the ad-hoc project, set a reservation baseline of 0 slots and set the ignore_idle_slot3 flag to False.

D. Create two Enterprise Edition reservations, one for each of the projects. For the SLA project, set a baseline of 800 slots. For the ad-hoc

project, enable autoscaling up to 200 slots.

Answer: B

Explanation:

To ensure that both production jobs with strict SLAs and ad-hoc queries have appropriate compute resources available while adhering to cost efficiency, setting up separate reservations and billing models for each project is the best approach. Here's why option B is the best choice:

Explanation:

Separate Reservations for SLA and Ad-hoc Projects:

Creating two separate reservations allows for dedicated resource management tailored to the needs of each project.

The production project requires guaranteed slots with the ability to scale up as needed, while the ad- hoc project benefits from on-demand billing based on data scanned.

Enterprise Edition Reservation for SLA Project:

Setting a baseline of 300 slots ensures that the SLA project has the minimum required resources.

Enabling autoscaling up to 500 additional slots allows the project to handle occasional spikes in

workload without compromising on SLAs.

On-Demand Billing for Ad-hoc Project:

Using on-demand billing for the ad-hoc project ensures cost efficiency, as users are billed based on the amount of data scanned rather than reserved slot capacity.

This model suits the less predictable and often lower-utilization nature of ad-hoc queries.

Steps to Implement:

Set Up Enterprise Edition Reservation for SLA Project:

Create a reservation with a baseline of 300 slots.

Enable autoscaling to allow up to an additional 500 slots as needed.

Configure On-Demand Billing for Ad-hoc Project:

Ensure that the ad-hoc project is set up to use on-demand billing, which charges based on data scanned by the queries.

Monitor and Adjust:

Continuously monitor the usage and performance of both projects to ensure that the configurations meet the needs and make adjustments as necessary.

Reference: Links:

BigQuery Slot Reservations

BigQuery On-Demand Pricing

Question: 370

You are a BigQuery admin supporting a team of data consumers who run ad hoc queries and downstream reporting in tools such as Looker. All data and users are combined under a single organizational project. You recently noticed some slowness in query results and want to troubleshoot where the slowdowns are occurring. You think that there might be some job queuing or slot contention occurring as users run jobs, which slows down access to results. You need to investigate the query job information and determine where performance is being affected. What should you do?

A. Use Cloud Monitoring to view BigQuery metrics and set up alerts that let you know when a certain percentage of slots were used.

B. Use slot reservations for your project to ensure that you have enough query processing capacity and are able to allocate available slots to the slower queries.

C. Use Cloud Logging to determine if any users or downstream consumers are changing or deleting access grants on tagged resources.

D. Use available administrative resource charts to determine how slots are being used and how jobs are performing over time. Run a query on the INFORMATION_SCHEMA to review query performance.

Answer: D

Explanation:

To troubleshoot query performance issues related to job queuing or slot contention in BigQuery, using administrative resource charts along with querying the INFORMATION_SCHEMA is the best approach. Here's why option D is the best choice:

Explanation:

Administrative Resource Charts:

BigQuery provides detailed resource charts that show slot usage and job performance over time. These charts help identify patterns of slot contention and peak usage times.

INFORMATION_SCHEMA Queries:

The INFORMATION_SCHEMA tables in BigQuery provide detailed metadata about query jobs, including execution times, slots consumed, and other performance metrics.

Running queries on INFORMATION_SCHEMA allows you to pinpoint specific jobs causing contention and analyze their performance characteristics.

Comprehensive Analysis:

Combining administrative resource charts with detailed queries on INFORMATION_SCHEMA provides a holistic view of the system's performance.

This approach enables you to identify and address the root causes of performance issues, whether they are due to slot contention, inefficient queries, or other factors.

Steps to Implement:

Access Administrative Resource Charts:

Use the Google Cloud Console to view BigQuery's administrative resource charts. These charts provide insights into slot utilization and job performance metrics over time.

Run INFORMATION_SCHEMA Queries:

Execute queries on BigQuery's INFORMATION_SCHEMA to gather detailed information about job performance. For example:

```
SELECT
  creation_time,
  job_id,
  user_email,
  query,
  total_slot_ms / 1000 AS slot_seconds,
  total_bytes_processed / (1024 * 1024 * 1024) AS processed_gb,
  total_bytes_billed / (1024 * 1024 * 1024) AS billed_gb
FROM
  `region-us`.INFORMATION_SCHEMA.JOBS_BY_PROJECT
```

```
WHERE
  creation_time > TIMESTAMP_SUB(CURRENT_TIMESTAMP(), INTERVAL 1 DAY)
  AND state = 'DONE'
ORDER BY
  slot_seconds DESC
LIMIT 100;
```

Analyze and Optimize:

Use the information gathered to identify bottlenecks, optimize queries, and adjust resource allocations as needed to improve performance.

Reference: Links:

Monitoring BigQuery Slots

BigQuery INFORMATION_SCHEMA

BigQuery Performance Best Practices

Question: 371

You are designing a messaging system by using Pub/Sub to process clickstream data with an event- driven consumer app that relies on a push subscription. You need to configure the messaging system that is reliable enough to handle temporary downtime of the consumer app. You also need the messaging system to store the input messages that cannot be consumed by the subscriber. The system needs to retry failed messages gradually, avoiding overloading the consumer app, and store the failed messages after a maximum of 10 retries in a topic. How should you configure the Pub/Sub subscription?

A. Increase the acknowledgement deadline to 10 minutes.

B. Use immediate redelivery as the subscription retry policy, and configure dead lettering to a different topic with maximum delivery attempts set to 10.

C. Use exponential backoff as the subscription retry policy, and configure dead lettering to the same source topic with maximum delivery attempts set to 10.

D. Use exponential backoff as the subscription retry policy, and configure dead lettering to a different topic with maximum delivery attempts set to 10.

Answer: D

Explanation:

Question: 372

You are preparing an organization-wide dataset. You need to preprocess customer data stored in a restricted bucket in Cloud Storage. The data will be used to create consumer analyses. You need to follow data privacy requirements, including protecting certain sensitive data elements, while also retaining all of the data for potential future use cases. What should you do?

A. Use Dataflow and the Cloud Data Loss Prevention API to mask sensitive data. Write the processed data in BigQuery.

B. Use the Cloud Data Loss Prevention API and Dataflow to detect and remove sensitive fields from the data in Cloud Storage. Write the filtered data in BigQuery.

C. Use Dataflow and Cloud KMS to encrypt sensitive fields and write the encrypted data in BigQuery. Share the encryption key by following the principle of least privilege.

D. Use customer-managed encryption keys (CMEK) to directly encrypt the data in Cloud Storage. Use federated queries from BigQuery. Share the encryption key by following the principle of least privilege.

Answer: A

Question: 373

You migrated a data backend for an application that serves 10 PB of historical product data for analytics. Only the last known state for a product, which is about 10 GB of data, needs to be served through an API to the other applications. You need to choose a cost-effective persistent storage solution that can accommodate the analytics requirements and the API performance of up to 1000 queries per second (QPS) with less than 1 second latency. What should you do?

A. 1. Store the historical data in BigQuery for analytics.

2. In a Cloud SQL table, store the last state of the product after every product change.

3. Serve the last state data directly from Cloud SQL to the API.

B. 1. Store the historical data in Cloud SQL for analytics.

2. In a separate table, store the last state of the product after every product change.

3. Serve the last state data directly from Cloud SQL to the API.

C. 1. Store the products as a collection in Firestore with each product having a set of historical changes.

2. Use simple and compound queries for analytics.

3. Serve the last state data directly from Firestore to the API.

D. 1. Store the historical data in BigQuery for analytics.

2. Use a materialized view to precompute the last state of a product.

3. Serve the last state data directly from BigQuery to the API.

Answer: D

Explanation:

Question: 374

You want to migrate an Apache Spark 3 batch job from on-premises to Google Cloud. You need to minimally change the job so that the job reads from Cloud Storage and writes the result to BigQuery. Your job is optimized for Spark, where each executor has 8 vCPU and 16 GB memory, and you want to be able to choose similar settings. You want to minimize installation and management effort to run your job. What should you do?

A. Execute the job in a new Dataproc cluster.

B. Execute as a Dataproc Serverless job.

C. Execute the job as part of a deployment in a new Google Kubernetes Engine cluster.

D. Execute the job from a new Compute Engine VM.

Answer: A

Explanation:

Question: 375

You currently have transactional data stored on-premises in a PostgreSQL database. To modernize your data environment, you want to run transactional workloads and support analytics needs with a single database. You need to move to Google Cloud without changing database management systems, and minimize cost and complexity.

What should you do?

A. Migrate your workloads to AlloyDB for PostgreSQL.

B. Migrate to BigQuery to optimize analytics.

C. Migrate and modernize your database with Cloud Spanner.

D. Migrate your PostgreSQL database to Cloud SQL for PostgreSQL.

Answer: A

Explanation:

The key requirements are:

On-premises PostgreSQL database.

Run transactional workloads AND support analytics needs with a single database.

Move to Google Cloud without changing database management systems (i.e., remain PostgreSQL-compatible).

Minimize cost and complexity.

AlloyDB for PostgreSQL (Option A) is the best fit for these requirements.

PostgreSQL-Compatible: AlloyDB is fully PostgreSQL-compatible, meaning minimal to no application changes are required ("without changing database management systems").

Transactional and Analytical Workloads: AlloyDB is designed to handle demanding transactional workloads while also providing significantly faster analytical query performance compared to standard PostgreSQL. It achieves this through its intelligent, database-optimized storage layer and

columnar engine integration. This addresses the "single database" for both needs.

Cost and Complexity: As a managed service, it reduces operational complexity. Its performance benefits for both OLTP and OLAP can lead to better cost-efficiency by handling mixed workloads effectively on a single system.

Let's analyze why other options are less suitable:

B (Migrate to BigQuery): BigQuery is an analytical data warehouse, not designed for transactional workloads. This violates the "single database" for both types of workloads and "without changing database management systems" (as BigQuery is not PostgreSQL).

C (Migrate to Cloud Spanner): Cloud Spanner is a globally distributed, horizontally scalable relational database. While excellent for high-availability transactional workloads, it has its own SQL dialect (ANSI 2011 with extensions, not fully PostgreSQL wire-compatible without tools like PGAdapter, which adds complexity) and a different architecture. This would involve more significant changes than moving to a PostgreSQL-compatible system. The requirement was "without changing database management systems."

D (Migrate to Cloud SQL for PostgreSQL): Cloud SQL for PostgreSQL is a fully managed PostgreSQL service. It's excellent for transactional workloads and simpler analytical queries. However, for more demanding analytical needs on the same database instance, AlloyDB is specifically optimized to provide superior performance due to its architectural enhancements (like the columnar engine). If the analytical needs are significant, AlloyDB offers a better converged experience. While Cloud SQL is PostgreSQL-compatible, AlloyDB is positioned for superior performance on mixed workloads.

Reference:

Google Cloud Documentation: AlloyDB for PostgreSQL > Overview. "AlloyDB for PostgreSQL is a fully managed, PostgreSQL-compatible database service for your most demandingtransactional and analytical workloads... AlloyDB offers full PostgreSQL compatibility, so you can migrate your existing PostgreSQL applications with no code changes."

Google Cloud Documentation: AlloyDB for PostgreSQL > Key benefits. Highlights include "Industryleading performance: ...up to 100x faster analytical queries than standard PostgreSQL." and "Support for transactional and analytical workloads: AlloyDB is designed to efficiently handle both transactional and analytical queries, allowing you to use a single database for a wide range of applications."

Question: 376

You designed a data warehouse in BigQuery to analyze sales data. You want a self-serving, low- maintenance, and cost-effective solution to share the sales dataset to other business units in your organization. What should you do?

A. Enable the other business units' projects to access the authorized views of the sales dataset.

B. Use the BigQuery Data Transfer Service to create a schedule that copies the sales dataset to the other business units' projects.

C. Create an Analytics Hub private exchange, and publish the sales dataset.

D. Create and share views with the users in the other business units.

Answer: C

Explanation:

The key requirements for sharing the sales dataset are:

Self-serving for other business units.

Low-maintenance.

Cost-effective.

Sharing with other business units (implying potentially different projects).

Analytics Hub (Option C) is designed precisely for this purpose of sharing data assets (like datasets) in a governed, discoverable, and self-service manner across an organization and even externally.

Self-Serving: Consumers (other business units) can browse available datasets in an exchange and subscribe to them. This makes it easy for them to discover and access the data they need without manual intervention from the data provider for each request.

Low-Maintenance for Provider: Once a dataset is published as a "listing" in Analytics Hub, the provider doesn't need to manage individual access requests or data copying for each new consumer project that subscribes. Updates to the source dataset are reflected for subscribers.

Cost-Effective:No Data Duplication for Sharing: When a dataset is shared via Analytics Hub, subscribers query the data directly from the provider's project (unless the provider explicitly opts for a replicated dataset model, which is less common for internal sharing where live access is preferred). This avoids storage costs associated with duplicating large

datasets in multiple projects.

Query Costs: Query costs are typically borne by the subscriber's project.

Governed Sharing: Analytics Hub provides a centralized way to manage and audit data sharing.

Let's analyze why other options are less suitable:

A (Enable access to authorized views): Authorized views are a good way to share specific slices or aggregations of data without exposing the underlying tables. However, managing authorizations for potentially many views across many business units/projects can become less "self-serving" and more "low-maintenance" than a dedicated data exchange platform. Discoverability is also less centralized.

B (BigQuery Data Transfer Service to copy): This creates data copies, which increases storage costs and can lead to data staleness if the copy schedule isn't frequent enough. It's not "low-maintenance" as it requires managing DTS jobs and storage for copies. It's generally not the most cost-effective way to share for querying.

D (Create and share views with users): Similar to authorized views, but sharing directly with individual users can be a permissions management challenge at scale compared to project-level or group-level subscriptions facilitated by Analytics Hub. It lacks the "exchange" concept for discovery and self-service subscription by business units/projects.

Reference:

Google Cloud Documentation: Analytics Hub > Overview. "Analytics Hub is a platform that lets you create and manage exchanges of data assets efficiently and securely... Data providers can publish listings that reference shared datasets. Subscribers can view these listings and thensubscribe to them. When a subscriber subscribes to a listing, Analytics Hub creates a linked dataset in the subscriber's project that references the shared dataset."

Google Cloud Documentation: Analytics Hub > Key benefits. "Simplified data sharing: Providers share data once, and subscribers access it in their own projects without data movement... Cost efficiency: Subscribers pay for queries run against shared data, not for storing the data." This aligns with selfserving, low-maintenance, and cost-effective sharing.

Question: 377

Your organization stores highly personal data in BigQuery and needs to comply with strict data privacy regulations. You need to ensure that sensitive data values are rendered unreadable whenever an employee leaves the organization. What should you do?

A. Use dynamic data masking and revoke viewer permissions when employees leave the organization.

B. Use column-level access controls with policy tags and revoke viewer permissions when employees leave the

organization.

C. Use AEAD functions and delete keys when employees leave the organization.

D. Use customer-managed encryption keys (CMEK) and delete keys when employees leave the organization.

Answer: C

Explanation:

Comprehensive and Detailed

The core requirement is to make specific data values permanently unreadable, a concept often referred to as cryptographic erasure or "crypto-shredding."

Option C is the correct answer because it achieves this at a granular, field-level. BigQuery's AEAD (Authenticated Encryption with Associated Data) functions allow you to encrypt individual field values within your tables using a keyset from Cloud KMS. You can encrypt the sensitive data, store the encrypted value in BigQuery, and when an employee leaves, you simply destroy the specific KMS key version used for their data. Once the key is destroyed, the encrypted data is permanently unrecoverable, thus rendering it unreadable.

Option A and B are incorrect because data masking and column-level access controls are about controlling access to the data, not about making the underlying data itself unreadable. If an administrator with sufficient privileges were to query the table, the data would still be there in its original form. Revoking permissions prevents a specific user from seeing the data, but it doesn't erase it.

Option D is incorrect because Customer-Managed Encryption Keys (CMEK) operate at the entire table level. Deleting a CMEK key would make the entire table unreadable, which is far too broad and disruptive. The goal is to target specific sensitive values, not destroy whole tables.

Reference (Google Cloud Documentation Concepts):

This solution leverages field-level encryption within BigQuery. The official documentation for "Protecting data with Cloud KMS keys" states that BigQuery's AEAD functions (AEAD.ENCRYPT and AEAD.DECRYPT) are the intended mechanism for this purpose. This approach provides fine-grained control, allowing you to manage the lifecycle of encrypted data by managing the lifecycle of the encryption keys in Cloud Key Management Service (Cloud KMS). The act of destroying a key version is the mechanism for crypto-shredding.

Question: 378

Your company needs to ingest and transform streaming data from IoT devices and store it for analysis. The data is sensitive and requires encryption with your own key in transit and at rest. The volume of data is expected to fluctuate significantly throughout the day. You need to identify a solution that is managed and elastic. What should you do?

A. Write data directly into BigQuery by using the Storage Write API, and process it in BigQuery by using SQL functions, selecting a Google-managed encryption key for each service.

B. Publish data to Pub/Sub, process it with Dataflow and store it in Cloud SQL, selecting your key from Cloud HSM for each service.

C. Publish data to Pub/Sub, process it with Dataflow and store it in BigQuery, selecting your key from Cloud KMS for each service.

D. Write data directly into Cloud Storage, process it with Dataproc, and store it in BigQuery, selecting a customer-managed encryption key (CMEK) for each service.

Answer: C

Explanation:

Comprehensive and Detailed

This question describes a classic, scalable streaming analytics architecture on Google Cloud.

Option C is the correct answer as it combines the best-in-class managed services for each part of the pipeline.

Pub/Sub is a fully managed, highly scalable messaging service perfect for ingesting fluctuating volumes of streaming IoT data.

Dataflow is a fully managed, serverless service for stream and batch processing that automatically scales resources up and down to handle fluctuating data volumes.

BigQuery is a serverless, highly scalable data warehouse optimized for analytics.

Cloud KMS is the standard Google Cloud service for creating and managing your own cryptographic keys, which are then used to enable Customer-Managed Encryption Keys (CMEK) across services like Pub/Sub, Dataflow, and BigQuery, satisfying the security requirement.

Option A is incorrect because it specifies using a Google-managed encryption key, which violates the requirement for encryption with "your own key."

Option B is incorrect because Cloud SQL is a relational database (OLTP), not an analytical data warehouse (OLAP), making it unsuitable for storing and analyzing large volumes of streaming data. BigQuery is the appropriate choice.

Option D is incorrect because Dataproc is a managed Hadoop/Spark service. While powerful, it is less "managed" and serverless than Dataflow, as you still need to provision and manage clusters. For a fully elastic and managed solution, Dataflow is the preferred choice.

Reference (Google Cloud Documentation Concepts):

This architecture is a canonical pattern for streaming analytics on Google Cloud. The "Streaming analytics" solution guide frequently highlights the Pub/Sub -> Dataflow -> BigQuery pattern. Each of these services supports Customer-Managed

Encryption Keys (CMEK) using keys from Cloud KMS, ensuring data is protected at rest with customer-controlled keys, which aligns with the principle of customer control over data security.

Question: 379

Your company is planning to migrate a large on-premises data warehouse to BigQuery. The data is currently stored in a proprietary, vendor-specific format. You need to perform a batch migration of this data to BigQuery. What should you do?

A. Use the bq command-line tool to load the data directly from the on-premises data warehouse.

B. Use the BigQuery Data Transfer Service.

C. Export the data to CSV files, upload the files to Cloud Storage, then load the files into BigQuery.

D. Use Datastream to replicate the data in real time.

Answer: C

Explanation:

Comprehensive and Detailed

The challenge here is dealing with a "proprietary, vendor-specific format" for a one-time batch migration.

Option C is the correct answer because it represents the most universal and reliable pattern for migration from any source system. By first exporting the data into a standard, interoperable format like CSV (or preferably, a self-describing format like Avro or Parquet), you decouple the process from the proprietary source. These standard files can then be easily uploaded to Cloud Storage (the recommended staging area for BigQuery loads) and loaded into BigQuery in a highly performant and parallelized manner.

Option A is incorrect because the bq command-line tool cannot connect directly to an on-premises data warehouse to pull data. It loads data from files or streams.

Option B is incorrect because the BigQuery Data Transfer Service (DTS) has connectors for specific, common data sources (like Teradata, Redshift, S3). It is unlikely to have a connector for a generic "proprietary, vendor-specific format."

Option D is incorrect because Datastream is a Change Data Capture (CDC) service designed for realtime replication of databases, not for a large-scale, one-time batch migration of a data warehouse.

Reference (Google Cloud Documentation Concepts):Google Cloud's "Data warehouse migration to BigQuery" guide outlines several migration strategies. For batch data transfer, the recommended path is Extract, Transfer, Load (ETL). This involves extracting data from the source into files (in formats like CSV, Avro, Parquet), transferring those files to Cloud Storage, and then loading them into BigQuery. This approach is recommended for its reliability and compatibility with any source system

Question: 380

You need to analyze user clickstream data to personalize content recommendations. The data arrives continuously and needs to be processed with low latency, including transformations such as sessionization (grouping clicks by user within a time window) and aggregation of user activity. You need to identify a scalable solution to handle millions of events each second and be resilient to late- arriving data. What should you do?

A. Use Firebase Realtime Database for ingestion and storage, and Cloud Run functions for processing and analytics.

B. Use Cloud Storage for ingestion, Dataproc with Apache Spark for batch processing, and BigQuery for storage and analytics.

C. Use Pub/Sub for ingestion, Dataflow with Apache Beam for processing, and BigQuery for storage and analytics.

D. Use Cloud Data Fusion for ingestion and transformation, and Cloud SQL for storage and analytics.

Answer: C

Explanation:

Comprehensive and Detailed

This question requires a solution that excels at large-scale, stateful stream processing with sophisticated windowing and handling of out-of-order data.

Option C is the correct answer because this architecture is perfectly suited for the requirements.

Pub/Sub is the global, scalable ingestion service for continuous event data.

Dataflow, with the Apache Beam programming model, is specifically designed for complex stream processing. It has powerful, built-in support for different windowing strategies (including session windows for sessionization) and sophisticated triggers for handling late-arriving data. Its serverless nature ensures it scales to handle millions of events.

BigQuery is the ideal sink for the processed data, enabling large-scale analytics for the recommendation engine.

Option A is incorrect as Firebase and Cloud Run are more suited for application backends and are not designed for complex, stateful data processing pipelines at this scale.

Option B is incorrect because it describes a batch processing pattern. Using Cloud Storage for ingestion and Dataproc for batch processing would introduce high latency, failing the "low latency" requirement.

Option D is incorrect because Cloud Data Fusion is primarily a batch-oriented ETL/ELT tool, and Cloud SQL is not an analytical data warehouse capable of handling this scale of data for analytics.

Reference (Google Cloud Documentation Concepts):

This is another example of the canonical Pub/Sub -> Dataflow -> BigQuery streaming analytics pattern. The Apache Beam Programming Guide (which is the foundation for Dataflow) extensively covers concepts like Windowing (specifically

SessionWindows) and Triggers for handling late data. These features are critical for accurately processing real-world event streams like clickstream data and are core strengths of Dataflow.

Question: 381

Your organization stores employee information in a BigQuery dataset. Your human resources (HR) admin team requires full access to the data, but the HR analyst team needs to conduct salary analysis without being able to access personally identifiable information (PII). You want to ensure that users have the correct level of access for their role managed through Dataplex, while reducing data duplication. What should you do?

A. Create an authorized view to limit data access based on a user role.

B. Create and assign policy tags based on user role to the PII columns in BigQuery.

C. Create a new dataset for salary analysis and use data masking to obfuscate all fields related to an individual.

D. Create a new dataset and use Cloud Data Loss Prevention (Cloud DLP) to mask PII in the BigQuery table.

Answer: B

Explanation:

Comprehensive and Detailed

The requirements are role-based access, masking specific columns (PII) for one role, and avoiding data duplication.

Option B is the correct answer because it directly addresses all requirements using Google Cloud's modern data governance tools. You use Data Catalog (governed by Dataplex) to create a taxonomy and policy tags (e.g., a "PII" tag). You apply this tag to the sensitive columns in your BigQuery table. Then, using IAM, you can grant the HR Analyst role access to the table but apply a data masking policy to the "PII" tag for that role. When an analyst queries the table, BigQuery dynamically masks the tagged columns for them at query time. The HR Admin role is granted full access without the masking policy. This provides fine-grained security without creating any copies of the data.

Option A is a valid but older approach. An authorized view could work, but it requires manually maintaining the view's SQL definition. Policy tags are more scalable and manageable, especially as the number of tables and policies grows.

Options C and D are incorrect because they both involve creating a new, separate dataset with masked data. This explicitly violates the requirement to "reduce data duplication."

Reference (Google Cloud Documentation Concepts):

This is the primary use case for column-level security in BigQuery. The official documentation shows how to "Restrict access with BigQuery column-level security" by creating a taxonomy and policy tags in Data Catalog. You then grant IAM permissions on these policy tags (roles/datacatalog.categoryFineGrainedReader) to different groups of users. For users who should see masked data, you do not grant them this role on the sensitive tags, and a data masking rule (e.g., nullification, hashing) can be applied. This entire governance framework can be managed centrally via Dataplex.

Question: 382

Your team runs a complex analytical query daily that processes terabytes of data. Recently, after running for 20 minutes, the query fails with a "Resources exceeded" error. You need to resolve this issue. What should you do?

A. Increase your project's BigQuery API request quota.

B. Increase the maximum table size limit.

C. Analyze the SQL syntax for errors.

D. Move from BigQuery on-demand to slot reservations.

Answer: D

Explanation:

Comprehensive and Detailed The error message "Resources exceeded" in BigQuery indicates that the query's execution plan is too complex or requires more computational resources (slots) than are available to it in the on-demand, fair-share pool.

Option D is the correct answer. BigQuery's on-demand pricing model uses a massive, shared pool of processing units called slots. While this pool is large, a single query cannot monopolize it, and there are limits to prevent runaway jobs. For consistently complex, high-resource queries, the solution is to switch to capacity-based pricing by purchasing slot reservations (e.g., using BigQuery editions). This provides your project with a dedicated, guaranteed amount of processing capacity, ensuring your complex queries have the resources they need to complete successfully.

Option A is incorrect because API request quotas relate to the number of API calls (e.g., how many jobs you can submit per minute), not the computational resources allocated to a single running query.

Option B is incorrect because table size limits are not related to query execution resources.

Option C is incorrect because while a syntax error would cause a query to fail, it would do so immediately with a syntax error message, not after 20 minutes with a "Resources exceeded" error. While optimizing the query is a good practice, the most direct way to solve a resource limit issue is to provide more resources.

Reference (Google Cloud Documentation Concepts):The Google Cloud documentation on "BigQuery pricing" explains the two main models: on-demand pricing and capacity-based pricing (editions). The "Resources exceeded" error is a known limitation of the on-demand model for extremely demanding queries. The documentation on "Introduction to slots" and "Reservations" explicitly presents purchasing dedicated slots as the solution for gaining more predictable and higher query performance for demanding workloads.

Question: 383

You are designing BigQuery tables for large volumes of clickstream event data. Your data analyst team will most frequently query by specific event date ranges and filter by the user ID UUID. You want to optimize table structure for query cost and performance. What should you do?

A. Partition the table by the event date column and cluster the table by the user ID column.

B. Partition the table by the user ID column and cluster the table by the event date column.

C. Create an ingestion-time partitioned table and cluster it by the user ID column.

D. Cluster the table by both the event date and the user ID columns.

Answer: A

Explanation:

Comprehensive and Detailed

This question is about applying the two primary optimization techniques in BigQuery: partitioning and clustering.

Partitioning divides a table into smaller segments based on a date, timestamp, or integer column. When you filter a query on the partition column, BigQuery performs "partition pruning," meaning it only scans the data in the relevant partitions. Since queries frequently filter by "specific event date ranges," partitioning by the event date column is the ideal strategy to reduce the amount of data scanned, which lowers cost and improves performance.

Clustering sorts the data within each partition based on the values in one or more columns. When you filter on a clustered column, BigQuery can use the sorted order to avoid scanning all the data within the relevant partitions. Since queries also filter by user ID, clustering by the user ID column will further improve performance and can reduce costs for those queries.

Conclusion: Option A is the correct answer because it correctly applies both techniques according to best practices: partition on the range-filtered column (event date) and cluster on the point- lookup/high-cardinality filtered column (user ID).

Option B is incorrect because you cannot partition a BigQuery table by a string column like a UUID.

Option C is incorrect because ingestion-time partitioning is less precise. Partitioning directly on the event date column is more effective for queries that filter on the event date itself.

Option D is incorrect because while clustering is helpful, partitioning is the more critical optimization for date range queries and provides the biggest cost savings. The combination of both is optimal.

Reference (Google Cloud Documentation Concepts):

The Google Cloud documentation for "Introduction to partitioned tables" and "Introduction to clustered tables" provides clear guidance. The best practice is to partition by a date or timestamp column that is commonly used as a filter to prune data. It then recommends to cluster by columns that are frequently used in WHERE clauses for filtering or in JOIN clauses. The combination of partitioning and clustering is a powerful optimization strategy.

Question: 384

You are building a data pipeline on Google Cloud. You need to prepare data using a casual method for a machine-learning process. You want to support a logistic regression model. You also need to monitor and adjust for null values, which must remain real-valued and cannot be removed. What should you do?

A. Use Cloud Dataprep to find null values in sample source data. Convert all nulls to 'none' using a Cloud Dataproc job.

B. Use Cloud Dataprep to find null values in sample source data. Convert all nulls to 0 using a Cloud Dataprep job.

C. Use Cloud Dataflow to find null values in sample source data. Convert all nulls to 'none' using a Cloud Dataprep job.

D. Use Cloud Dataflow to find null values in sample source data. Convert all nulls to using a custom script.

Answer:C