Google Professional Data Engineer.exam.25q

Number: Google Data Engineer
Passing Score: 800
Time Limit: 120 min

Google Certified Professional

Google Certified Professional – Data Engineer

Exam A

QUESTION 1
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

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 2
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 data. 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.

Correct Answer: D

Section: (none)
Explanation

Explanation/Reference:
QUESTION 3
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*`

Correct Answer: B
Section: (none)
Explanation

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

QUESTION 4
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.

Correct Answer: BDF

Section: (none)

Explanation

Explanation/Reference:

QUESTION 5

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.

Correct Answer: D

Section: (none)

Explanation

Explanation/Reference:

QUESTION 6

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 data. 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.

Correct Answer: BDF

Section: (none)

Explanation

Explanation/Reference:

QUESTION 7

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

Correct Answer: A

Section: (none)

Explanation

Explanation/Reference:

QUESTION 8

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.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 9
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 data. 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.

Correct Answer: BCE
Section: (none)
Explanation

Explanation/Reference:

QUESTION 10
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-forlike 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.

Correct Answer: B
Section: (none)
Explanation

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

QUESTION 11
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.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 12
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 data. 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.

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:

QUESTION 13
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.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 14
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.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 15
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.

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:

QUESTION 16
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.

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:

QUESTION 17
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 data. 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.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 18

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.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
QUESTION 19
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.
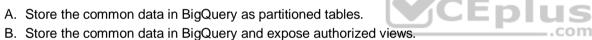
CTO Statement
IT has never been a priority for us, so as our data has grown, we have not invested enough in our technology. I have a good staff to manage IT, but they are so busy managing our infrastructure that I cannot get them to do the things that really matter, such as organizing our data, building 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.

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.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 20
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.

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

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 21
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.

You need to compose visualizations for operations teams with the following requirements:
▪ The report must include telemetry data from all 50,000 installations for the most resent 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.

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.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
QUESTION 22
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.

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.

Correct Answer: BD
Section: (none)
Explanation

Explanation/Reference:

QUESTION 23
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.

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_id` Column data: `data_point`

B. Rowkey: `date`
   Column data: `device_id, data_point`

C. Rowkey: `device_id`
   Column data: `date, data_point`

D. Rowkey: `data_point`
   Column data: `device_id, date`

E. Rowkey: `date#data_point` Column data: `device_id`

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:

QUESTION 24
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.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 25
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.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference: