

Google Certified Professional.67q

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

Google Certified Professional



Website: <https://vceplus.com>

VCE to PDF Converter: <https://vceplus.com/vce-to-pdf/>

Facebook: <https://www.facebook.com/VCE.For.All.VN/>

Twitter : https://twitter.com/VCE_Plus

<https://www.vceplus.com/>

Google Certified Professional – Data Engineer

Exam A

QUESTION 1

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 2

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?



<https://www.vceplus.com/>

- 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 3

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 4

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 5

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.

Correct Answer: BC

Section: (none)

Explanation

Explanation/Reference:



QUESTION 6

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 data. 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.

Correct Answer: A

Section: (none)

Explanation

Explanation/Reference:

QUESTION 7

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 (IAM) 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.

Correct Answer: C

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 are working on a sensitive project involving private user data. 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.

Correct Answer: C

Section: (none)

Explanation

Explanation/Reference:

QUESTION 12

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.

Correct Answer: B

Section: (none)

Explanation

Explanation/Reference:

QUESTION 13

Your company is performing data preprocessing for a learning algorithm in Google Cloud Dataflow. Numerous data logs 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 `PCollection` represents a single row in the table.

Correct Answer: D

Section: (none)

Explanation

Explanation/Reference:

QUESTION 14

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?



<https://www.vceplus.com/>

- 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>.`

Correct Answer: A

Section: (none)

Explanation

Explanation/Reference:

QUESTION 15

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.

Correct Answer: C

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 size of the BigData 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 into. 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
- 10 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 parity 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 our 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 the 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 into. 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
- 10 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 parity 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 our 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**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 into. 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
 - 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
- 10 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 party 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 our 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 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.

Correct Answer: C

Section: (none)

Explanation

Explanation/Reference:

QUESTION 22

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 into. 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
- 10 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 our 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 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 Cloud 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.

Correct Answer: B

Section: (none)

Explanation

Explanation/Reference:

QUESTION 23

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 schema. Which Google Cloud Platform product should you use?

- A. BigQuery
- B. Cloud SQL
- C. Cloud Bigtable
- D. Cloud Datastore

Correct Answer: A

Section: (none)

Explanation

Explanation/Reference:



QUESTION 24

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.

Correct Answer: B

Section: (none)

Explanation

Explanation/Reference:

QUESTION 25

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-premises data to the designated storage bucket.

Correct Answer: CE

Section: (none)

Explanation

Explanation/Reference:

QUESTION 26

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

Correct Answer: BDF

Section: (none)

Explanation

Explanation/Reference:

QUESTION 27

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

Correct Answer: ADF

Section: (none)

Explanation

Explanation/Reference:

QUESTION 28

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

Correct Answer: B

Section: (none)

Explanation

Explanation/Reference:

QUESTION 29

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

Correct Answer: A

Section: (none)

Explanation

Explanation/Reference:

QUESTION 30

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.

Correct Answer: C

Section: (none)

Explanation

Explanation/Reference:

QUESTION 31

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 data. 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”

Correct Answer: AE

Section: (none)

Explanation

Explanation/Reference:

QUESTION 32

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`

Correct Answer: A

Section: (none)

Explanation

Explanation/Reference:

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

QUESTION 33

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?



<https://www.vceplus.com/>

- 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

Correct Answer: C

Section: (none)

Explanation

Explanation/Reference:

QUESTION 34

You architect a system to analyze seismic data. 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.

Correct Answer: A

Section: (none)

Explanation

Explanation/Reference:

QUESTION 35

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

Correct Answer: C

Section: (none)

Explanation



Explanation/Reference:

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

QUESTION 36

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

Correct Answer: A

Section: (none)

Explanation

Explanation/Reference:

QUESTION 37

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

Correct Answer: A

Section: (none)

Explanation

Explanation/Reference:



QUESTION 38

Your company receives both batch- and stream-based event data. 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.

Correct Answer: B

Section: (none)

Explanation

Explanation/Reference:

QUESTION 39

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

Correct Answer: D

Section: (none)

Explanation

Explanation/Reference:

QUESTION 40

Your company maintains a hybrid deployment with GCP, where analytics are performed on your anonymized customer data. 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.

Correct Answer: C

Section: (none)

Explanation

Explanation/Reference:

QUESTION 41

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 is building a custom interface to share data. They have these requirements:

1. They need to do aggregations over their petabyte-scale datasets.
2. 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

Correct Answer: C

Section: (none)

Explanation

Explanation/Reference:

QUESTION 42**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 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.

Correct Answer: B

Section: (none)

Explanation

Explanation/Reference:

QUESTION 43

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.

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.

Correct Answer: B

Section: (none)

Explanation

Explanation/Reference:

QUESTION 44**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 into. 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
- 10 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 capacity 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 in 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 our 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
- E. Cloud Dataflow, Cloud SQL, and Cloud Storage

Correct Answer: C

Section: (none)

Explanation

Explanation/Reference:

QUESTION 45

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.

Correct Answer: B

Section: (none)

Explanation

Explanation/Reference:

QUESTION 46

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.

Correct Answer: C

Section: (none)

Explanation

Explanation/Reference:

Reference <https://cloud.google.com/blog/products/gcp/busting-12-myths-about-bigquery>

QUESTION 47

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 from 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 from PubSub and write to GCS.

Correct Answer: A

Section: (none)

Explanation

Explanation/Reference:

QUESTION 48

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 shuffling 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.

Correct Answer: C

Section: (none)

Explanation

Explanation/Reference:

QUESTION 49

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.

Correct Answer: C

Section: (none)

Explanation

Explanation/Reference:

QUESTION 50

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 longitude 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 longitude as input vectors to your neural net.
- B. Create a numeric column from a feature cross of latitude and longitude.
- C. Create a feature cross of latitude and longitude, bucketize at the minute level and use L1 regularization during optimization.
- D. Create a feature cross of latitude and longitude, bucketize it at the minute level and use L2 regularization during optimization.

Correct Answer: B

Section: (none)

Explanation

Explanation/Reference:

Reference <https://cloud.google.com/bigquery/docs/gis-data>

QUESTION 51

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.

Correct Answer: C

Section: (none)

Explanation

Explanation/Reference:

QUESTION 52

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.

Correct Answer: B

Section: (none)

Explanation

Explanation/Reference:

QUESTION 53

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



Correct Answer: D

Section: (none)

Explanation

Explanation/Reference:

QUESTION 54

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.

Correct Answer: A

Section: (none)

Explanation

Explanation/Reference:

QUESTION 55

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.

Correct Answer: A

Section: (none)

Explanation

Explanation/Reference:

QUESTION 56

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.

Correct Answer: B

Section: (none)

Explanation

Explanation/Reference:

QUESTION 57

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.

Correct Answer: C

Section: (none)

Explanation

Explanation/Reference:

QUESTION 58

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.

Correct Answer: D

Section: (none)

Explanation

Explanation/Reference:

Reference <https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/flex>

QUESTION 59

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.

Correct Answer: A

Section: (none)

Explanation

Explanation/Reference:



QUESTION 60

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.

Correct Answer: D

Section: (none)

Explanation

Explanation/Reference:

QUESTION 61

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.

Correct Answer: CD

Section: (none)

Explanation

Explanation/Reference:

QUESTION 62

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.

Correct Answer: C

Section: (none)

Explanation

Explanation/Reference:

QUESTION 63

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.

Correct Answer: A

Section: (none)

Explanation

Explanation/Reference:

QUESTION 64

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

Correct Answer: D

Section: (none)

Explanation

Explanation/Reference:

QUESTION 65

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.

Correct Answer: B

Section: (none)

Explanation

Explanation/Reference:

QUESTION 66

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.

Correct Answer: A

Section: (none)

Explanation

Explanation/Reference:

QUESTION 67

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.

Correct Answer: D

Section: (none)

Explanation

Explanation/Reference:



<https://www.vceplus.com/>

