

Google.Professional Data Engineer .vJun-2024.by.Enis.155q

Number: Professional Data Engineer  
Passing Score: 800  
Time Limit: 120  
File Version: 6.6

Exam Code: Professional Data Engineer  
Exam Name: Professional Data Engineer on Google Cloud Platform



## Exam A

### QUESTION 1

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:**

### QUESTION 2

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:**

### QUESTION 3

You are testing a Dataflow pipeline to ingest and transform text files. The files are compressed gzip, errors are written to a dead-letter queue, and you are using SidelInputs to join data. You noticed that the pipeline is taking longer to complete than expected, what should you do to expedite the Dataflow job?

- A. Switch to compressed Avro files
- B. Reduce the batch size
- C. Retry records that throw an error
- D. Use CoGroupByKey instead of the SidelInput

**Correct Answer: B**

**Section:**

### QUESTION 4

You are administering a BigQuery dataset that uses a customer-managed encryption key (CMEK). You need to share the dataset with a partner organization that does not have access to your CMEK. What should you do?

- A. Create an authorized view that contains the CMEK to decrypt the data when accessed.



- B. Provide the partner organization a copy of your CMEKs to decrypt the data.
- C. Copy the tables you need to share to a dataset without CMEKs Create an Analytics Hub listing for this dataset.
- D. Export the tables to parquet files to a Cloud Storage bucket and grant the storageinsights. viewer role on the bucket to the partner organization.

**Correct Answer: C**

**Section:**

**Explanation:**

If you want to share a BigQuery dataset that uses a customer-managed encryption key (CMEK) with a partner organization that does not have access to your CMEK, you cannot use an authorized view or provide them a copy of your CMEK, because these options would violate the security and privacy of your data. Instead, you can copy the tables you need to share to a dataset without CMEKs, and then create an Analytics Hub listing for this dataset. Analytics Hub is a service that allows you to securely share and discover data assets across your organization and with external partners. By creating an Analytics Hub listing, you can grant the partner organization access to the copied dataset without CMEKs, and also control the level of access and the duration of the sharing. Reference:

Customer-managed Cloud KMS keys

[Authorized views]

[Analytics Hub overview]

[Creating an Analytics Hub listing]

#### QUESTION 5

You are designing a data mesh on Google Cloud with multiple distinct data engineering teams building data products. The typical data curation design pattern consists of landing files in Cloud Storage, transforming raw data in Cloud Storage and BigQuery datasets. and storing the final curated data product in BigQuery datasets You need to configure Dataplex to ensure that each team can access only the assets needed to build their data products. You also need to ensure that teams can easily share the curated data product. What should you do?

- A. 1 Create a single Dataplex virtual lake and create a single zone to contain landing, raw. and curated data. 2 Provide each data engineering team access to the virtual lake.
- B. 1 Create a single Dataplex virtual lake and create a single zone to contain landing, raw. and curated data. 2 Build separate assets for each data product within the zone. 3. Assign permissions to the data engineering teams at the zone level.
- C. 1 Create a Dataplex virtual lake for each data product, and create a single zone to contain landing, raw, and curated data. 2. Provide the data engineering teams with full access to the virtual lake assigned to their data product.
- D. 1 Create a Dataplex virtual lake for each data product, and create multiple zones for landing, raw. and curated data. 2. Provide the data engineering teams with full access to the virtual lake assigned to their data product.

**Correct Answer: D**

**Section:**

**Explanation:**

This option is the best way to configure Dataplex for a data mesh architecture, as it allows each data engineering team to have full ownership and control over their data products, while also enabling easy discovery and sharing of the curated data across the organization<sup>12</sup>.By creating a Dataplex virtual lake for each data product, you can isolate the data assets and resources for each domain, and avoid conflicts and dependencies between different teams<sup>3</sup>.By creating multiple zones for landing, raw, and curated data, you can enforce different security and governance policies for each stage of the data curation process, and ensure that only authorized users can access the data assets<sup>45</sup>. By providing the data engineering teams with full access to the virtual lake assigned to their data product, you can empower them to manage and monitor their data products, and leverage the Dataplex features such as tagging, quality, and lineage.

Option A is not suitable, as it creates a single point of failure and a bottleneck for the data mesh, and does not allow for fine-grained access control and governance for different data products<sup>2</sup>.Option B is also not suitable, as it does not isolate the data assets and resources for each data product, and assigns permissions at the zone level, which may not reflect the different roles and responsibilities of the data engineering teams<sup>34</sup>.Option C is better than option A and B, but it does not create multiple zones for landing, raw, and curated data, which may compromise the security and quality of the data products<sup>5</sup>.Reference:

1: Building a data mesh on Google Cloud using BigQuery and Dataplex | Google Cloud Blog

2: Data Mesh - 7 Effective Practices to Get Started - Confluent

3: Best practices | Dataplex | Google Cloud

4: Secure your lake | Dataplex | Google Cloud

5: Zones | Dataplex | Google Cloud

[6]: Managing a Data Mesh with Dataplex -- ROI Training

#### QUESTION 6

You are on the data governance team and are implementing security requirements to deploy resources. You need to ensure that resources are limited to only the europe-west 3 region You want to follow Google-recommended practices What should you do?

- A. Deploy resources with Terraform and implement a variable validation rule to ensure that the region is set to the europe-west3 region for all resources.
- B. Set the constraints/gcp.resourceLocations organization policy constraint to in:eu-locations.
- C. Create a Cloud Function to monitor all resources created and automatically destroy the ones created outside the europe-west3 region.
- D. Set the constraints/gcp.resourceLocations organization policy constraint to in: europe-west3-locations.

**Correct Answer: D**

**Section:**

**Explanation:**

To ensure that resources are limited to only the europe-west3 region, you should set the organization policy constraint constraints/gcp.resourceLocations to in:europe-west3-locations. This policy restricts the deployment of resources to the specified locations, which in this case is the europe-west3 region. By setting this policy, you enforce location compliance across your Google Cloud resources, aligning with the best practices for data governance and regulatory compliance.

Professional Data Engineer Certification Exam Guide | Learn - Google Cloud1.

Preparing for Google Cloud Certification: Cloud Data Engineer2.

Professional Data Engineer Certification | Learn | Google Cloud3.

3:Professional Data Engineer Certification | Learn | Google Cloud2:Preparing for Google Cloud Certification: Cloud Data Engineer1:Professional Data Engineer Certification Exam Guide | Learn - Google Cloud

### QUESTION 7

You have a BigQuery table that contains customer data, including sensitive information such as names and addresses. You need to share the customer data with your data analytics and consumer support teams securely. The data analytics team needs to access the data of all the customers, but must not be able to access the sensitive data. The consumer support team needs access to all data columns, but must not be able to access customers that no longer have active contracts. You enforced these requirements by using an authorized dataset and policy tags After implementing these steps, the data analytics team reports that they still have access to the sensitive columns. You need to ensure that the data analytics team does not have access to restricted data What should you do?

Choose 2 answers

- A. Create two separate authorized datasets; one for the data analytics team and another for the consumer support team.
- B. Ensure that the data analytics team members do not have the Data Catalog Fine-Grained Reader role for the policy tags.
- C. Enforce access control in the policy tag taxonomy.
- D. Remove the bigquery.dataViewer role from the data analytics team on the authorized datasets.
- E. Replace the authorized dataset with an authorized view Use row-level security and apply filter\_ expression to limit data access.

**Correct Answer: B, C**

**Section:**

**Explanation:**

To ensure that the data analytics team does not have access to sensitive columns, you should:

B) Ensure that the data analytics team members do not have the Data Catalog Fine-Grained Reader role for the policy tags.This role allows users to read metadata for data assets that have policy tags applied, which could include sensitive information.

C) Enforce access control in the policy tag taxonomy.By setting access control at the policy tag level, you can restrict access to specific columns within a dataset, ensuring that only authorized users can view sensitive data.

### QUESTION 8

You are building a streaming Dataflow pipeline that ingests noise level data from hundreds of sensors placed near construction sites across a city. The sensors measure noise level every ten seconds, and send that data to the pipeline when levels reach above 70 dBA. You need to detect the average noise level from a sensor when data is received for a duration of more than 30 minutes, but the window ends when no data has been received for 15 minutes What should you do?

- A. Use session windows with a 30-minute gap duration.
- B. Use tumbling windows with a 15-minute window and a fifteen-minute. withAllowedLateness operator.
- C. Use session windows with a 15-minute gap duration.
- D. Use hopping windows with a 15-minute window, and a thirty-minute period.

**Correct Answer: B**

**Section:**

**Explanation:**

Session windows are dynamic windows that group elements based on the periods of activity. They are useful for streaming data that is irregularly distributed with respect to time. In this case, the noise level data from the sensors is only sent when it exceeds a certain threshold, and the duration of the noise events may vary. Therefore, session windows can capture the average noise level for each sensor during the periods of high noise, and end the window when there is no data for a specified gap duration. The gap duration should be 15 minutes, as the requirement is to end the window when no data has been received for 15 minutes. A 30-minute gap duration would be too long and may miss some noise events that are shorter than 30 minutes. Tumbling windows and hopping windows are fixed windows that group elements based on a fixed time interval. They are not suitable for this use case, as they may split or overlap the noise events from the sensors, and do not account for the periods of inactivity. Reference:

Windowing concepts

Session windows

Windowing in Dataflow

#### **QUESTION 9**

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

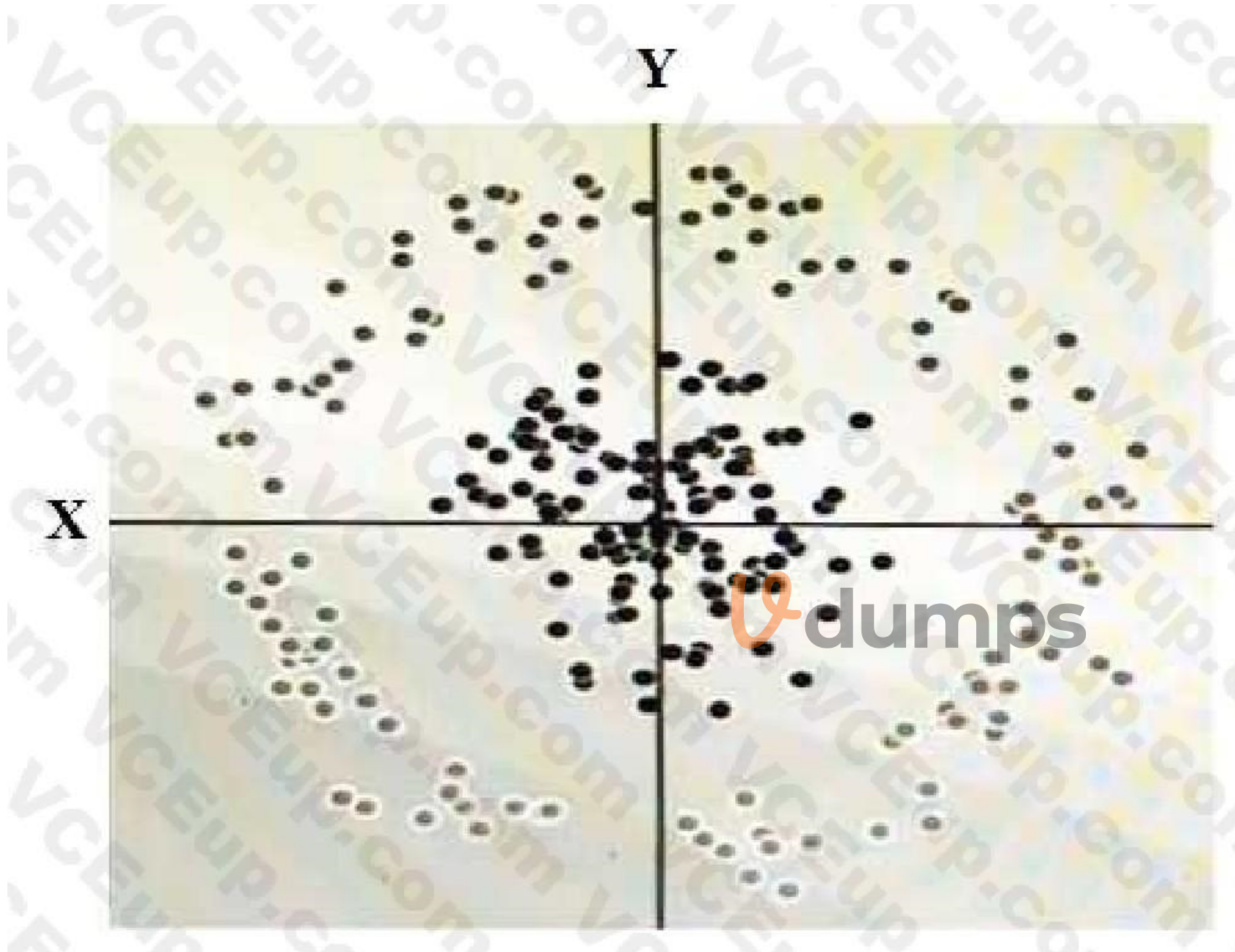
**Correct Answer: B**

**Section:**

#### **QUESTION 10**

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

The logo for 'Vdumps' is centered on the page. It features a stylized orange 'V' followed by the word 'dumps' in a grey, lowercase, sans-serif font.



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

- A.  $X^2+Y^2$
- B.  $X^2$
- C.  $Y^2$
- D.  $\cos(X)$

**Correct Answer: D**

**Section:**

**QUESTION 11**

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

What should you do?

- A. Create groups for your users and give those groups access to the dataset
- B. Integrate with a single sign-on (SSO) platform, and pass each user's credentials along with the query request
- C. Create a service account and grant dataset access to that account. Use the service account's private key to access the dataset
- D. Create a dummy user and grant dataset access to that user. Store the username and password for that user in a file on the files system, and use those credentials to access the BigQuery dataset

**Correct Answer: C**

**Section:**

**QUESTION 12**

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

What should you do?

- A. Create a dedicated service account, and use encryption at rest to reference your data stored in your Compute Engine cluster instances as part of your API service calls.
- B. Create encryption keys in Cloud Key Management Service. Use those keys to encrypt your data in all of the Compute Engine cluster instances.
- C. Create encryption keys locally. Upload your encryption keys to Cloud Key Management Service. Use those keys to encrypt your data in all of the Compute Engine cluster instances.
- D. Create encryption keys in Cloud Key Management Service. Reference those keys in your API service calls when accessing the data in your Compute Engine cluster instances.

**Correct Answer: C**

**Section:**

**QUESTION 13**

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

- A. Build and train a complex classification model with Spark MLlib to generate labels and filter the results. Deploy the models using Cloud Dataproc. Call the model from your application.
- B. Build and train a classification model with Spark MLlib to generate labels. Build and train a second classification model with Spark MLlib to filter results to match customer preferences. Deploy the models using Cloud Dataproc. Call the models from your application.
- C. Build an application that calls the Cloud Video Intelligence API to generate labels. Store data in Cloud Bigtable, and filter the predicted labels to match the user's viewing history to generate preferences.
- D. Build an application that calls the Cloud Video Intelligence API to generate labels. Store data in Cloud SQL, and join and filter the predicted labels to match the user's viewing history to generate preferences.

**Correct Answer: C**

**Section:**

**QUESTION 14**

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

- A. Disable caching by editing the report settings.
- B. Disable caching in BigQuery by editing table details.

- C. Refresh your browser tab showing the visualizations.
- D. Clear your browser history for the past hour then reload the tab showing the virtualizations.

**Correct Answer: A**

**Section:**

**Explanation:**

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

#### QUESTION 15

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

- A. Use federated data sources, and check data in the SQL query.
- B. Enable BigQuery monitoring in Google Stackdriver and create an alert.
- C. Import the data into BigQuery using the gcloud CLI and set max\_bad\_records to 0.
- D. Run a Google Cloud Dataflow batch pipeline to import the data into BigQuery, and push errors to another dead-letter table for analysis.

**Correct Answer: D**

**Section:**

#### QUESTION 16

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

- A. Issue a command to restart the database servers.
- B. Retry the query with exponential backoff, up to a cap of 15 minutes.
- C. Retry the query every second until it comes back online to minimize staleness of data.
- D. Reduce the query frequency to once every hour until the database comes back online.



**Correct Answer: B**

**Section:**

**Explanation:**

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

#### QUESTION 17

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:**

#### QUESTION 18

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:**

**Explanation:**

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

#### QUESTION 19

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: D**

**Section:**



#### QUESTION 20

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: B, D, F**

**Section:**

#### QUESTION 21

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: C**

**Section:**

#### QUESTION 22

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: B, D, F**

**Section:**

#### QUESTION 23

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:**

#### QUESTION 24

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 characteristics 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: A, D**

**Section:**

**Explanation:**

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

[https://en.wikipedia.org/wiki/Anomaly\\_detection](https://en.wikipedia.org/wiki/Anomaly_detection)

**QUESTION 25**

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: D**

**Section:**

**Explanation:**

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

**QUESTION 26**

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: A**

**Section:**

**QUESTION 27**

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: D**

**Section:**

**QUESTION 28**

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: B, C, D**

**Section:**

#### QUESTION 29

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

- A. Put the data into Google Cloud Storage.
- B. Use preemptible virtual machines (VMs) for the Cloud Dataproc cluster.
- C. Tune the Cloud Dataproc cluster so that there is just enough disk for all data.
- D. Migrate some of the cold data into Google Cloud Storage, and keep only the hot data in Persistent Disk.

**Correct Answer: B**

**Section:**

#### QUESTION 30

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 on 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:**

#### QUESTION 31

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 efficiently?

- 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:**



**QUESTION 32**

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:**

**QUESTION 33**

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:**

**QUESTION 34**

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:**

**QUESTION 35**

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

- A. Make a call to the Stackdriver API to list all logs, and apply an advanced filter.
- B. In the Stackdriver logging admin interface, and enable a log sink export to BigQuery.

- C. In the Stackdriver logging admin interface, enable a log sink export to Google Cloud Pub/Sub, and subscribe to the topic from your monitoring tool.
- D. Using the Stackdriver API, create a project sink with advanced log filter to export to Pub/Sub, and subscribe to the topic from your monitoring tool.

**Correct Answer: B**

**Section:**

#### QUESTION 36

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:**

#### QUESTION 37

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:**

#### QUESTION 38

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:**

**QUESTION 39**

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

- A. Use a row key of the form <timestamp>.
- B. Use a row key of the form <sensorid>.
- C. Use a row key of the form <timestamp>#<sensorid>.
- D. Use a row key of the form >#<sensorid>#<timestamp>.

**Correct Answer: A**

**Section:**

**QUESTION 40**

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:**

**QUESTION 41**

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:**

**QUESTION 42**

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 Bigtable 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:**

**QUESTION 43**

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:**

**QUESTION 44**

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 ñ user data, inventory, static data

3 physical servers

Cassandra ñ metadata, tracking messages

10 Kafka servers ñ tracking message aggregation and batch insert

Application servers ñ customer front end, middleware for order/customs 60 virtual machines across 20 physical servers Tomcat ñ Java services Nginx ñ static content Batch servers Storage appliances iSCSI for virtual machine (VM) hosts Fibre Channel storage area network (FC SAN) ñ 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 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:**

#### QUESTION 45

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 ñ user data, inventory, static data

3 physical servers

Cassandra ñ metadata, tracking messages

10 Kafka servers ñ tracking message aggregation and batch insert

Application servers ñ customer front end, middleware for order/customs 60 virtual machines across 20 physical servers Tomcat ñ Java services Nginx ñ static content Batch servers Storage appliances iSCSI for virtual machine (VM) hosts Fibre Channel storage area network (FC SAN) ñ 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 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:**

#### QUESTION 46

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 ñ user data, inventory, static data

3 physical servers

Cassandra ñ metadata, tracking messages

10 Kafka servers ñ tracking message aggregation and batch insert

Application servers ñ customer front end, middleware for order/customs 60 virtual machines across 20 physical servers Tomcat ñ Java services Nginx ñ static content Batch servers Storage appliances iSCSI for virtual machine (VM) hosts Fibre Channel storage area network (FC SAN) ñ 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 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:**

#### QUESTION 47

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 ñ user data, inventory, static data

3 physical servers

Cassandra ñ metadata, tracking messages

10 Kafka servers ñ tracking message aggregation and batch insert

Application servers ñ customer front end, middleware for order/customs 60 virtual machines across 20 physical servers Tomcat ñ Java services Nginx ñ static content Batch servers Storage appliances iSCSI for virtual machine (VM) hosts Fibre Channel storage area network (FC SAN) ñ 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 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:**

**Explanation:**

Topic 3, MJTelco Case Study



#### QUESTION 48

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

- A. The zone
- B. The number of workers
- C. The disk size per worker
- D. The maximum number of workers

**Correct Answer: A**

**Section:**

#### QUESTION 49

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 ñ development/test, staging, and production ñ 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:

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:**

#### QUESTION 50

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 ñ development/test, staging, and production ñ 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: B, D**

**Section:**

#### QUESTION 51

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 ñ development/test, staging, and production ñ 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\_idColumn data: data\_point
- B. Rowkey: dateColumn data: device\_id, data\_point
- C. Rowkey: device\_idColumn data: date, data\_point
- D. Rowkey: data\_pointColumn data: device\_id, date
- E. Rowkey: date#data\_pointColumn data: device\_id

**Correct Answer: D**

**Section:**

### QUESTION 52

#### 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 ñ development/test, staging, and production ñ 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:  
They need to do aggregations over their petabyte-scale datasets.  
They need to scan specific time range rows with a very fast response time (milliseconds).  
Which combination of Google Cloud Platform products should you recommend?

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

**Correct Answer: C**

**Section:**

### QUESTION 53

#### 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 ñ development/test, staging, and production ñ 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:**

#### QUESTION 54

##### 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 ñ development/test, staging, and production ñ 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:**

**Explanation:**

Topic 4, Main Questions Set B

#### QUESTION 55

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:**



#### QUESTION 56

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:**

#### QUESTION 57

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

- A. HTTPS
- B. VPN
- C. SOCKS

D. HTTP

**Correct Answer: C**

**Section:**

**Explanation:**

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

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

**QUESTION 58**

Cloud Dataproc is a managed Apache Hadoop and Apache \_\_\_\_\_ service.

- A. Blaze
- B. Spark
- C. Fire
- D. Ignite

**Correct Answer: B**

**Section:**

**Explanation:**

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

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

**QUESTION 59**

Which action can a Cloud Dataproc Viewer perform?

- A. Submit a job.
- B. Create a cluster.
- C. Delete a cluster.
- D. List the jobs.

**Correct Answer: D**

**Section:**

**Explanation:**

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

Reference:

[https://cloud.google.com/dataproc/docs/concepts/iam#iam\\_roles\\_and\\_cloud\\_dataproc\\_operations\\_summary](https://cloud.google.com/dataproc/docs/concepts/iam#iam_roles_and_cloud_dataproc_operations_summary)

**QUESTION 60**

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

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

**Correct Answer: B**

**Section:**

**Explanation:**



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

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

#### QUESTION 61

Scaling a Cloud Dataproc cluster typically involves \_\_\_\_.

- A. increasing or decreasing the number of worker nodes
- B. increasing or decreasing the number of master nodes
- C. moving memory to run more applications on a single node
- D. deleting applications from unused nodes periodically

**Correct Answer: A**

**Section:**

**Explanation:**

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

- 1) increase the number of workers to make a job run faster
- 2) decrease the number of workers to save money
- 3) increase the number of nodes to expand available Hadoop Distributed Filesystem (HDFS) storage

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

#### QUESTION 62

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

- A. month-by-month
- B. minute-by-minute
- C. week-by-week
- D. hour-by-hour

**Correct Answer: B**

**Section:**

**Explanation:**

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

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

#### QUESTION 63

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

- A. application node
- B. conditional node
- C. master node
- D. worker node

**Correct Answer: C**

**Section:**

**Explanation:**

The YARN ResourceManager and the HDFS NameNode interfaces are available on a Cloud Dataproc cluster master node. The cluster master-host-name is the name of your Cloud Dataproc cluster followed by an -m suffix for



example, if your cluster is named "my-cluster", the master-host-name would be "my-cluster-m".  
Reference: <https://cloud.google.com/dataproc/docs/concepts/cluster-web-interfaces#interfaces>

#### QUESTION 64

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

- A. Set initialization actions
- B. Modify configuration files using cluster properties
- C. Configure the cluster using Cloud Deployment Manager
- D. Log into the master node and make changes from there

**Correct Answer: C**

**Section:**

**Explanation:**

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

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

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

[<https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/initactions>]

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

#### QUESTION 65

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

- A. VPN connection
- B. Special browser
- C. SSH tunnel
- D. FTP connection



**Correct Answer: C**

**Section:**

**Explanation:**

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

Reference: [https://cloud.google.com/dataproc/docs/concepts/cluster-webinterfaces#connecting\\_to\\_the\\_web\\_interfaces](https://cloud.google.com/dataproc/docs/concepts/cluster-webinterfaces#connecting_to_the_web_interfaces)

#### QUESTION 66

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

- A. before
- B. after
- C. only if
- D. once

**Correct Answer: A**

**Section:**

**Explanation:**

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

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

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

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

#### QUESTION 67

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

- A. Include multiple time series values within the row key
- B. Keep the row key as an 8 bit integer
- C. Keep your row key reasonably short
- D. Keep your row key as long as the field permits

**Correct Answer: C**

**Section:**

**Explanation:**

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

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

#### QUESTION 68

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

- A. Using IAM roles, you cannot give a user access to only one table in a project, rather than all tables in a project.
- B. To give a user access to only one table in a project, grant the user the Bigtable Editor role for that table.
- C. You can configure access control only at the project level.
- D. To give a user access to only one table in a project, you must configure access through your application.

**Correct Answer: B**

**Section:**

**Explanation:**

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

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

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

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

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

#### QUESTION 69

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

- A. Have the Compute Engine instance in the furthest zone from the Cloud Bigtable instance.
- B. Have both the Compute Engine instance and the Cloud Bigtable instance to be in different zones.
- C. Have both the Compute Engine instance and the Cloud Bigtable instance to be in the same zone.
- D. Have the Cloud Bigtable instance to be in the same zone as all of the consumers of your data.

**Correct Answer: C**

**Section:**

**Explanation:**

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

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

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

#### QUESTION 70

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

- A. A sequential numeric ID
- B. A timestamp followed by a stock symbol
- C. A non-sequential numeric ID
- D. A stock symbol followed by a timestamp

**Correct Answer: A, B**

**Section:**

**Explanation:**

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

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

Suppose your system assigns a numeric ID to each of your application's users. You might be tempted to use the user's numeric ID as the row key for your table. However, since new users are more likely to be active users, this approach is likely to push most of your traffic to a small number of nodes.

[<https://cloud.google.com/bigtable/docs/schema-design>]

Reference: [https://cloud.google.com/bigtable/docs/schema-design-timeseries#ensure\\_that\\_your\\_row\\_key\\_avoids\\_hotspotting](https://cloud.google.com/bigtable/docs/schema-design-timeseries#ensure_that_your_row_key_avoids_hotspotting)

#### QUESTION 71

When a Cloud Bigtable node fails, \_\_\_\_\_ is lost.

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

**Correct Answer: B**

**Section:**

**Explanation:**

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

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

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

Cloud Bigtable simply updates the pointers for each node.

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

When a Cloud Bigtable node fails, no data is lost

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

#### QUESTION 72

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

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





**Correct Answer: C**

**Section:**

**Explanation:**

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

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

#### QUESTION 73

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

- A. Field promotion
- B. Randomization
- C. Salting
- D. Hashing

**Correct Answer: A**

**Section:**

**Explanation:**

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

Reference: [https://cloud.google.com/bigtable/docs/schema-design-timeseries#ensure\\_that\\_your\\_row\\_key\\_avoids\\_hotspotting](https://cloud.google.com/bigtable/docs/schema-design-timeseries#ensure_that_your_row_key_avoids_hotspotting)

#### QUESTION 74

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

- A. Avoid schema designs that are based on NoSQL concepts
- B. Create schema designs that are based on a relational database design
- C. Avoid schema designs that require atomicity across rows
- D. Create schema designs that require atomicity across rows



**Correct Answer: C**

**Section:**

**Explanation:**

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

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

#### QUESTION 75

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

- A. You expect to store at least 10 TB of data.
- B. You will mostly run batch workloads with scans and writes, rather than frequently executing random reads of a small number of rows.
- C. You need to integrate with Google BigQuery.
- D. You will not use the data to back a user-facing or latency-sensitive application.

**Correct Answer: C**

**Section:**

**Explanation:**

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

storage.

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

#### QUESTION 76

Cloud Bigtable is Google's \_\_\_\_\_ Big Data database service.

- A. Relational
- B. mySQL
- C. NoSQL
- D. SQL Server

**Correct Answer: C**

**Section:**

**Explanation:**

Cloud Bigtable is Google's NoSQL Big Data database service. It is the same database that Google uses for services, such as Search, Analytics, Maps, and Gmail. It is used for requirements that are low latency and high throughput including Internet of Things (IoT), user analytics, and financial data analysis.

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

#### QUESTION 77

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

- A. 500 TB
- B. 1 GB
- C. 1 TB
- D. 500 GB

**Correct Answer: C**

**Section:**

**Explanation:**

Cloud Bigtable is not a relational database. It does not support SQL queries, joins, or multi-row transactions. It is not a good solution for less than 1 TB of data.

Reference:

[https://cloud.google.com/bigtable/docs/overview#title\\_short\\_and\\_other\\_storage\\_options](https://cloud.google.com/bigtable/docs/overview#title_short_and_other_storage_options)

#### QUESTION 78

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

- A. Do not use a production instance.
- B. Run your test for at least 10 minutes.
- C. Before you test, run a heavy pre-test for several minutes.
- D. Use at least 300 GB of data.

**Correct Answer: A**

**Section:**

**Explanation:**

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

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

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



Before you test, run a heavy pre-test for several minutes. This step gives Cloud Bigtable a chance to balance data across your nodes based on the access patterns it observes. Run your test for at least 10 minutes. This step lets Cloud Bigtable further optimize your data, and it helps ensure that you will test reads from disk as well as cached reads from memory. Reference: <https://cloud.google.com/bigtable/docs/performance>

#### QUESTION 79

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

- A. multi-keyed data with very high latency
- B. multi-keyed data with very low latency
- C. single-keyed data with very low latency
- D. single-keyed data with very high latency

**Correct Answer: C**

**Section:**

**Explanation:**

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

Cloud Bigtable is ideal for storing very large amounts of single-keyed data with very low latency. It supports high read and write throughput at low latency, and it is an ideal data source for MapReduce operations.

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

#### QUESTION 80

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

- A. primary key
- B. unique key
- C. row key
- D. master key



**Correct Answer: C**

**Section:**

**Explanation:**

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

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

#### QUESTION 81

What is the HBase Shell for Cloud Bigtable?

- A. The HBase shell is a GUI based interface that performs administrative tasks, such as creating and deleting tables.
- B. The HBase shell is a command-line tool that performs administrative tasks, such as creating and deleting tables.
- C. The HBase shell is a hypervisor based shell that performs administrative tasks, such as creating and deleting new virtualized instances.
- D. The HBase shell is a command-line tool that performs only user account management functions to grant access to Cloud Bigtable instances.

**Correct Answer: B**

**Section:**

**Explanation:**

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

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

### QUESTION 82

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

- A. create a third instance and sync the data from the two storage types via batch jobs
- B. export the data from the existing instance and import the data into a new instance
- C. run parallel instances where one is HDD and the other is SDD
- D. the selection is final and you must resume using the same storage type

**Correct Answer: B**

**Section:**

**Explanation:**

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

If you need to convert an existing HDD cluster to SSD, or vice-versa, you can export the data from the existing instance and import the data into a new instance. Alternatively, you can write a Cloud Dataflow or Hadoop MapReduce job that copies the data from one instance to another.

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

Topic 6, Main Questions Set C

### QUESTION 83

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: A, D, F**

**Section:**

### QUESTION 84

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:**


### QUESTION 85

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:**

#### QUESTION 86

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:**

#### QUESTION 87

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: A, E**

**Section:**

#### QUESTION 88

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:**

**Explanation:**

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

#### QUESTION 89

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

- A. They have not assigned the timestamp, which causes the job to fail
- B. They have not set the triggers to accommodate the data coming in late, which causes the job to fail
- C. They have not applied a global windowing function, which causes the job to fail when the pipeline is created
- D. They have not applied a non-global windowing function, which causes the job to fail when the pipeline is created

**Correct Answer: C**

**Section:**

#### QUESTION 90

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:**

#### QUESTION 91

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:**

**Explanation:**

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

**QUESTION 92**

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

- A. Use Cloud Dataproc to run your transformations. Monitor CPU utilization for the cluster. Resize the number of worker nodes in your cluster via the command line.
- B. Use Cloud Dataproc to run your transformations. Use the diagnose command to generate an operational output archive. Locate the bottleneck and adjust cluster resources.
- C. Use Cloud Dataflow to run your transformations. Monitor the job system lag with Stackdriver. Use the default autoscaling setting for worker instances.
- D. Use Cloud Dataflow to run your transformations. Monitor the total execution time for a sampling of jobs. Configure the job to use non-default Compute Engine machine types when needed.

**Correct Answer: B**

**Section:**

**QUESTION 93**

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

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

**Correct Answer: B**

**Section:**

**QUESTION 94**

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

- A. Transform text files to compressed Avro using Cloud Dataflow. Use BigQuery for storage and query.
- B. Transform text files to compressed Avro using Cloud Dataflow. Use Cloud Storage and BigQuery permanent linked tables for query.
- C. Compress text files to gzip using the Grid Computing Tools. Use BigQuery for storage and query.
- D. Compress text files to gzip using the Grid Computing Tools. Use Cloud Storage, and then import into Cloud Bigtable for query.

**Correct Answer: D**

**Section:**

**QUESTION 95**

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

- A. Call the Cloud Natural Language API from your application. Process the generated Entity Analysis as labels.
- B. Call the Cloud Natural Language API from your application. Process the generated Sentiment Analysis as labels.
- C. Build and train a text classification model using TensorFlow. Deploy the model using Cloud Machine Learning Engine. Call the model from your application and process the results as labels.
- D. Build and train a text classification model using TensorFlow. Deploy the model using a Kubernetes Engine cluster. Call the model from your application and process the results as labels.

**Correct Answer: B**

**Section:**

**QUESTION 96**

You are designing storage for 20 TB of text files as part of deploying a data pipeline on Google Cloud.

Your input data is in CSV format. You want to minimize the cost of querying aggregate values for multiple users who will query the data in Cloud Storage with multiple engines. Which storage service and schema design should you use?

- A. Use Cloud Bigtable for storage. Install the HBase shell on a Compute Engine instance to query the Cloud Bigtable data.
- B. Use Cloud Bigtable for storage. Link as permanent tables in BigQuery for query.
- C. Use Cloud Storage for storage. Link as permanent tables in BigQuery for query.
- D. Use Cloud Storage for storage. Link as temporary tables in BigQuery for query.

**Correct Answer: A**

**Section:**

**QUESTION 97**

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

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

- A. Use Cloud SQL for storage. Add secondary indexes to support query patterns.
- B. Use Cloud SQL for storage. Use Cloud Dataflow to transform data to support query patterns.
- C. Use Cloud Spanner for storage. Add secondary indexes to support query patterns.
- D. Use Cloud Spanner for storage. Use Cloud Dataflow to transform data to support query patterns.

**Correct Answer: D**

**Section:**

**Explanation:**

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

**QUESTION 98**

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

Which product should they use to store the data?

- A. Cloud Bigtable
- B. Google BigQuery
- C. Google Cloud Storage
- D. Google Cloud Datastore

**Correct Answer: A**

**Section:**

**Explanation:**

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

**QUESTION 99**

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



- A. They want to expose aggregates of this data to other Google Cloud projects, while still controlling access to the user-level data. Additionally, they need to minimize their overall storage cost and ensure the analysis cost for other projects is assigned to those projects. What should they do?
- B. Create and share an authorized view that provides the aggregate results.
- C. Create and share a new dataset and view that provides the aggregate results.
- D. Create and share a new dataset and table that contains the aggregate results.
- E. Create dataViewer Identity and Access Management (IAM) roles on the dataset to enable sharing.

**Correct Answer: D**

**Section:**

**Explanation:**

Reference: <https://cloud.google.com/bigquery/docs/access-control>

#### QUESTION 100

Government regulations in your industry mandate that you have to maintain an auditable record of access to certain types of dat

- A. Assuming that all expiring logs will be archived correctly, where should you store data that is subject to that mandate?
- B. Encrypted on Cloud Storage with user-supplied encryption keys. A separate decryption key will be given to each authorized user.
- C. In a BigQuery dataset that is viewable only by authorized personnel, with the Data Access log used to provide the auditability.
- D. In Cloud SQL, with separate database user names to each user. The Cloud SQL Admin activity logs will be used to provide the auditability.
- E. In a bucket on Cloud Storage that is accessible only by an AppEngine service that collects user information and logs the access before providing a link to the bucket.

**Correct Answer: B**

**Section:**

#### QUESTION 101

Your neural network model is taking days to train. You want to increase the training speed. What can you do?

- A. Subsample your test dataset.
- B. Subsample your training dataset.
- C. Increase the number of input features to your model.
- D. Increase the number of layers in your neural network.

**Correct Answer: D**

**Section:**

**Explanation:**

Reference: <https://towardsdatascience.com/how-to-increase-the-accuracy-of-a-neural-network-9f5d1c6f407d>

#### QUESTION 102

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:**

#### QUESTION 103

You operate a database that stores stock trades and an application that retrieves average stock price for a given company over an adjustable window of time. The data is stored in Cloud Bigtable where the datetime of the stock trade is the beginning of the row key. Your application has thousands of concurrent users, and you notice that performance is starting to degrade as more stocks are added. What should you do to improve the performance of your application?

- A. Change the row key syntax in your Cloud Bigtable table to begin with the stock symbol.
- B. Change the row key syntax in your Cloud Bigtable table to begin with a random number per second.
- C. Change the data pipeline to use BigQuery for storing stock trades, and update your application.
- D. Use Cloud Dataflow to write summary of each day's stock trades to an Avro file on Cloud Storage. Update your application to read from Cloud Storage and Cloud Bigtable to compute the responses.

**Correct Answer: A**

**Section:**

#### QUESTION 104

You are operating a Cloud Dataflow streaming pipeline. The pipeline aggregates events from a Cloud Pub/Sub subscription source, within a window, and sinks the resulting aggregation to a Cloud Storage bucket. The source has consistent throughput. You want to monitor an alert on behavior of the pipeline with Cloud Stackdriver to ensure that it is processing data. Which Stackdriver alerts should you create?

- A. An alert based on a decrease of subscription/num\_undelivered\_messages for the source and a rate of change increase of instance/storage/used\_bytes for the destination
- B. An alert based on an increase of subscription/num\_undelivered\_messages for the source and a rate of change decrease of instance/storage/used\_bytes for the destination
- C. An alert based on a decrease of instance/storage/used\_bytes for the source and a rate of change increase of subscription/num\_undelivered\_messages for the destination
- D. An alert based on an increase of instance/storage/used\_bytes for the source and a rate of change decrease of subscription/num\_undelivered\_messages for the destination

**Correct Answer: B**

**Section:**

#### QUESTION 105

You currently have a single on-premises Kafka cluster in a data center in the us-east region that is responsible for ingesting messages from IoT devices globally. Because large parts of globe have poor internet connectivity, messages sometimes batch at the edge, come in all at once, and cause a spike in load on your Kafka cluster. This is becoming difficult to manage and prohibitively expensive. What is the Google-recommended cloud native architecture for this scenario?

- A. Edge TPUs as sensor devices for storing and transmitting the messages.
- B. Cloud Dataflow connected to the Kafka cluster to scale the processing of incoming messages.
- C. An IoT gateway connected to Cloud Pub/Sub, with Cloud Dataflow to read and process the messages from Cloud Pub/Sub.
- D. A Kafka cluster virtualized on Compute Engine in us-east with Cloud Load Balancing to connect to the devices around the world.

**Correct Answer: C**

**Section:**

#### QUESTION 106

You decided to use Cloud Datastore to ingest vehicle telemetry data in real time. You want to build a storage system that will account for the long-term data growth, while keeping the costs low. You also want to create snapshots of the data periodically, so that you can make a point-in-time (PIT) recovery, or clone a copy of the data for Cloud Datastore in a different environment. You want to archive these snapshots for a long time. Which two methods can accomplish this?

Choose 2 answers.

- A. Use managed export, and store the data in a Cloud Storage bucket using Nearline or Coldline class.
- B. Use managed export, and then import to Cloud Datastore in a separate project under a unique namespace reserved for that export.
- C. Use managed export, and then import the data into a BigQuery table created just for that export, and delete temporary export files.
- D. Write an application that uses Cloud Datastore client libraries to read all the entities. Treat each entity as a BigQuery table row via BigQuery streaming insert. Assign an export timestamp for each export, and attach it as an extra column for each row. Make sure that the BigQuery table is partitioned using the export timestamp column.
- E. Write an application that uses Cloud Datastore client libraries to read all the entities. Format the exported data into a JSON file. Apply compression before storing the data in Cloud Source Repositories.

**Correct Answer: C, E**

**Section:**

#### QUESTION 107

You need to create a data pipeline that copies time-series transaction data so that it can be queried from within BigQuery by your data science team for analysis. Every hour, thousands of transactions are updated with a new status. The size of the initial dataset is 1.5 PB, and it will grow by 3 TB per day. The data is heavily structured, and your data science team will build machine learning models based on this data. You want to maximize performance and usability for your data science team. Which two strategies should you adopt? Choose 2 answers.

- A. Denormalize the data as much as possible.
- B. Preserve the structure of the data as much as possible.
- C. Use BigQuery UPDATE to further reduce the size of the dataset.
- D. Develop a data pipeline where status updates are appended to BigQuery instead of updated.
- E. Copy a daily snapshot of transaction data to Cloud Storage and store it as an Avro file. Use BigQuery's support for external data sources to query.

**Correct Answer: A, E**

**Section:**

#### QUESTION 108

You are designing a cloud-native historical data processing system to meet the following conditions:

The data being analyzed is in CSV, Avro, and PDF formats and will be accessed by multiple analysis tools including Cloud Dataproc, BigQuery, and Compute Engine.

A streaming data pipeline stores new data daily.

Performance is not a factor in the solution.

The solution design should maximize availability.

How should you design data storage for this solution?

- A. Create a Cloud Dataproc cluster with high availability. Store the data in HDFS, and perform analysis as needed.
- B. Store the data in BigQuery. Access the data using the BigQuery Connector or Cloud Dataproc and Compute Engine.
- C. Store the data in a regional Cloud Storage bucket. Access the bucket directly using Cloud Dataproc, BigQuery, and Compute Engine.
- D. Store the data in a multi-regional Cloud Storage bucket. Access the data directly using Cloud Dataproc, BigQuery, and Compute Engine.

**Correct Answer: D**

**Section:**

#### QUESTION 109

You have a petabyte of analytics data and need to design a storage and processing platform for it. You must be able to perform data warehouse-style analytics on the data in Google Cloud and expose the dataset as files for batch analysis tools in other cloud providers. What should you do?

- A. Store and process the entire dataset in BigQuery.

- B. Store and process the entire dataset in Cloud Bigtable.
- C. Store the full dataset in BigQuery, and store a compressed copy of the data in a Cloud Storage bucket.
- D. Store the warm data as files in Cloud Storage, and store the active data in BigQuery. Keep this ratio as 80% warm and 20% active.

**Correct Answer: C**

**Section:**

#### QUESTION 110

You work for a manufacturing company that sources up to 750 different components, each from a different supplier. You've collected a labeled dataset that has on average 1000 examples for each unique component. Your team wants to implement an app to help warehouse workers recognize incoming components based on a photo of the component. You want to implement the first working version of this app (as Proof-Of-Concept) within a few working days. What should you do?

- A. Use Cloud Vision AutoML with the existing dataset.
- B. Use Cloud Vision AutoML, but reduce your dataset twice.
- C. Use Cloud Vision API by providing custom labels as recognition hints.
- D. Train your own image recognition model leveraging transfer learning techniques.

**Correct Answer: A**

**Section:**

#### QUESTION 111

You are working on a niche product in the image recognition domain. Your team has developed a model that is dominated by custom C++ TensorFlow ops your team has implemented. These ops are used inside your main training loop and are performing bulky matrix multiplications. It currently takes up to several days to train a model. You want to decrease this time significantly and keep the cost low by using an accelerator on Google Cloud. What should you do?

- A. Use Cloud TPUs without any additional adjustment to your code.
- B. Use Cloud TPUs after implementing GPU kernel support for your custom ops.
- C. Use Cloud GPUs after implementing GPU kernel support for your custom ops.
- D. Stay on CPUs, and increase the size of the cluster you're training your model on.

**Correct Answer: B**

**Section:**

#### QUESTION 112

You work on a regression problem in a natural language processing domain, and you have 100M labeled examples in your dataset. You have randomly shuffled your data and split your dataset into train and test samples (in a 90/10 ratio).

After you trained the neural network and evaluated your model on a test set, you discover that the root-mean-squared error (RMSE) of your model is twice as high on the train set as on the test set. How should you improve the performance of your model?

- A. Increase the share of the test sample in the train-test split.
- B. Try to collect more data and increase the size of your dataset.
- C. Try out regularization techniques (e.g., dropout or batch normalization) to avoid overfitting.
- D. Increase the complexity of your model by, e.g., introducing an additional layer or increase the size of vocabularies or n-grams used.

**Correct Answer: D**

**Section:**

**QUESTION 113**

You use BigQuery as your centralized analytics platform. New data is loaded every day, and an ETL pipeline modifies the original data and prepares it for the final users. This ETL pipeline is regularly modified and can generate errors, but sometimes the errors are detected only after 2 weeks. You need to provide a method to recover from these errors, and your backups should be optimized for storage costs. How should you organize your data in BigQuery and store your backups?

- A. Organize your data in a single table, export, and compress and store the BigQuery data in Cloud Storage.
- B. Organize your data in separate tables for each month, and export, compress, and store the data in Cloud Storage.
- C. Organize your data in separate tables for each month, and duplicate your data on a separate dataset in BigQuery.
- D. Organize your data in separate tables for each month, and use snapshot decorators to restore the table to a time prior to the corruption.

**Correct Answer: D**

**Section:**

**QUESTION 114**

The marketing team at your organization provides regular updates of a segment of your customer dataset. The marketing team has given you a CSV with 1 million records that must be updated in BigQuery. When you use the UPDATE statement in BigQuery, you receive a quotaExceeded error.

What should you do?

- A. Reduce the number of records updated each day to stay within the BigQuery UPDATE DML statement limit.
- B. Increase the BigQuery UPDATE DML statement limit in the Quota management section of the Google Cloud Platform Console.
- C. Split the source CSV file into smaller CSV files in Cloud Storage to reduce the number of BigQuery UPDATE DML statements per BigQuery job.
- D. Import the new records from the CSV file into a new BigQuery table. Create a BigQuery job that merges the new records with the existing records and writes the results to a new BigQuery table.

**Correct Answer: D**

**Section:**

**QUESTION 115**

As your organization expands its usage of GCP, many teams have started to create their own projects.

Projects are further multiplied to accommodate different stages of deployments and target audiences. Each project requires unique access control configurations. The central IT team needs to have access to all projects.

Furthermore, data from Cloud Storage buckets and BigQuery datasets must be shared for use in other projects in an ad hoc way. You want to simplify access control management by minimizing the number of policies. Which two steps should you take?

Choose 2 answers.

- A. Use Cloud Deployment Manager to automate access provision.
- B. Introduce resource hierarchy to leverage access control policy inheritance.
- C. Create distinct groups for various teams, and specify groups in Cloud IAM policies.
- D. Only use service accounts when sharing data for Cloud Storage buckets and BigQuery datasets.
- E. For each Cloud Storage bucket or BigQuery dataset, decide which projects need access. Find all the active members who have access to these projects, and create a Cloud IAM policy to grant access to all these users.

**Correct Answer: A, C**

**Section:**

**QUESTION 116**

Your United States-based company has created an application for assessing and responding to user actions. The primary table's data volume grows by 250,000 records per second. Many third parties use your application's APIs to build the functionality into their own frontend applications. Your application's APIs should comply with the following requirements:

Single global endpoint

ANSI SQL support

Consistent access to the most up-to-date data

What should you do?

- A. Implement BigQuery with no region selected for storage or processing.
- B. Implement Cloud Spanner with the leader in North America and read-only replicas in Asia and Europe.
- C. Implement Cloud SQL for PostgreSQL with the master in North America and read replicas in Asia and Europe.
- D. Implement Cloud Bigtable with the primary cluster in North America and secondary clusters in Asia and Europe.

**Correct Answer: B**

**Section:**

#### QUESTION 117

A data scientist has created a BigQuery ML model and asks you to create an ML pipeline to serve predictions. You have a REST API application with the requirement to serve predictions for an individual user ID with latency under 100 milliseconds. You use the following query to generate predictions: `SELECT predicted_label, user_id FROM ML.PREDICT (MODEL 'dataset.model', table user_features)`. How should you create the ML pipeline?

- A. Add a WHERE clause to the query, and grant the BigQuery Data Viewer role to the application service account.
- B. Create an Authorized View with the provided query. Share the dataset that contains the view with the application service account.
- C. Create a Cloud Dataflow pipeline using BigQueryIO to read results from the query. Grant the Dataflow Worker role to the application service account.
- D. Create a Cloud Dataflow pipeline using BigQueryIO to read predictions for all users from the query.

Write the results to Cloud Bigtable using BigtableIO. Grant the Bigtable Reader role to the application service account so that the application can read predictions for individual users from Cloud Bigtable.

**Correct Answer: D**

**Section:**

#### QUESTION 118

You are building an application to share financial market data with consumers, who will receive data feeds. Data is collected from the markets in real time. Consumers will receive the data in the following ways:

Real-time event stream

ANSI SQL access to real-time stream and historical data

Batch historical exports

Which solution should you use?

- A. Cloud Dataflow, Cloud SQL, Cloud Spanner
- B. Cloud Pub/Sub, Cloud Storage, BigQuery
- C. Cloud Dataproc, Cloud Dataflow, BigQuery
- D. Cloud Pub/Sub, Cloud Dataproc, Cloud SQL

**Correct Answer: A**

**Section:**

#### QUESTION 119

You are building a new application that you need to collect data from in a scalable way. Data arrives continuously from the application throughout the day, and you expect to generate approximately 150 GB of JSON data per day by the end of the year. Your requirements are:

Decoupling producer from consumer

Space and cost-efficient storage of the raw ingested data, which is to be stored indefinitely Near real-time SQL query Maintain at least 2 years of historical data, which will be queried with SQL Which pipeline should you use to meet these requirements?

- A. Create an application that provides an API. Write a tool to poll the API and write data to Cloud Storage as gzipped JSON files.
- B. Create an application that writes to a Cloud SQL database to store the data. Set up periodic exports of the database to write to Cloud Storage and load into BigQuery.

- C. Create an application that publishes events to Cloud Pub/Sub, and create Spark jobs on Cloud Dataproc to convert the JSON data to Avro format, stored on HDFS on Persistent Disk.
- D. Create an application that publishes events to Cloud Pub/Sub, and create a Cloud Dataflow pipeline that transforms the JSON event payloads to Avro, writing the data to Cloud Storage and BigQuery.

**Correct Answer: A**

**Section:**

#### QUESTION 120

You are running a pipeline in Cloud Dataflow that receives messages from a Cloud Pub/Sub topic and writes the results to a BigQuery dataset in the EU. Currently, your pipeline is located in europe-west4 and has a maximum of 3 workers, instance type n1-standard-1. You notice that during peak periods, your pipeline is struggling to process records in a timely fashion, when all 3 workers are at maximum CPU utilization. Which two actions can you take to increase performance of your pipeline? (Choose two.)

- A. Increase the number of max workers
- B. Use a larger instance type for your Cloud Dataflow workers
- C. Change the zone of your Cloud Dataflow pipeline to run in us-central1
- D. Create a temporary table in Cloud Bigtable that will act as a buffer for new data. Create a new step in your pipeline to write to this table first, and then create a new pipeline to write from Cloud Bigtable to BigQuery
- E. Create a temporary table in Cloud Spanner that will act as a buffer for new data. Create a new step in your pipeline to write to this table first, and then create a new pipeline to write from Cloud Spanner to BigQuery

**Correct Answer: A, B**

**Section:**

#### QUESTION 121

You have a data pipeline with a Cloud Dataflow job that aggregates and writes time series metrics to Cloud Bigtable. This data feeds a dashboard used by thousands of users across the organization. You need to support additional concurrent users and reduce the amount of time required to write the data. Which two actions should you take? (Choose two.)

- A. Configure your Cloud Dataflow pipeline to use local execution
- B. Increase the maximum number of Cloud Dataflow workers by setting maxNumWorkers in PipelineOptions
- C. Increase the number of nodes in the Cloud Bigtable cluster
- D. Modify your Cloud Dataflow pipeline to use the Flatten transform before writing to Cloud Bigtable
- E. Modify your Cloud Dataflow pipeline to use the CoGroupByKey transform before writing to Cloud Bigtable

**Correct Answer: B, C**

**Section:**

#### QUESTION 122

You have several Spark jobs that run on a Cloud Dataproc cluster on a schedule. Some of the jobs run in sequence, and some of the jobs run concurrently. You need to automate this process. What should you do?

- A. Create a Cloud Dataproc Workflow Template
- B. Create an initialization action to execute the jobs
- C. Create a Directed Acyclic Graph in Cloud Composer
- D. Create a Bash script that uses the Cloud SDK to create a cluster, execute jobs, and then tear down the cluster

**Correct Answer: C**

**Section:**

#### QUESTION 123

You are building a new data pipeline to share data between two different types of applications: jobs generators and job runners. Your solution must scale to accommodate increases in usage and must accommodate the addition of new applications without negatively affecting the performance of existing ones. What should you do?

- A. Create an API using App Engine to receive and send messages to the applications
- B. Use a Cloud Pub/Sub topic to publish jobs, and use subscriptions to execute them
- C. Create a table on Cloud SQL, and insert and delete rows with the job information
- D. Create a table on Cloud Spanner, and insert and delete rows with the job information

**Correct Answer: A**

**Section:**

#### QUESTION 124

You need to create a new transaction table in Cloud Spanner that stores product sales data. You are deciding what to use as a primary key. From a performance perspective, which strategy should you choose?

- A. The current epoch time
- B. A concatenation of the product name and the current epoch time
- C. A random universally unique identifier number (version 4 UUID)
- D. The original order identification number from the sales system, which is a monotonically increasing integer

**Correct Answer: C**

**Section:**

#### QUESTION 125

Data Analysts in your company have the Cloud IAM Owner role assigned to them in their projects to allow them to work with multiple GCP products in their projects. Your organization requires that all BigQuery data access logs be retained for 6 months. You need to ensure that only audit personnel in your company can access the data access logs for all projects. What should you do?

- A. Enable data access logs in each Data Analyst's project. Restrict access to Stackdriver Logging via Cloud IAM roles.
- B. Export the data access logs via a project-level export sink to a Cloud Storage bucket in the Data Analysts' projects. Restrict access to the Cloud Storage bucket.
- C. Export the data access logs via a project-level export sink to a Cloud Storage bucket in a newly created project for audit logs. Restrict access to the project with the exported logs.
- D. Export the data access logs via an aggregated export sink to a Cloud Storage bucket in a newly created project for audit logs. Restrict access to the project that contains the exported logs.

**Correct Answer: D**

**Section:**

#### QUESTION 126

Each analytics team in your organization is running BigQuery jobs in their own projects. You want to enable each team to monitor slot usage within their projects. What should you do?

- A. Create a Stackdriver Monitoring dashboard based on the BigQuery metric query/scanned\_bytes
- B. Create a Stackdriver Monitoring dashboard based on the BigQuery metric slots/allocated\_for\_project
- C. Create a log export for each project, capture the BigQuery job execution logs, create a custom metric based on the totalSlotMs, and create a Stackdriver Monitoring dashboard based on the custom metric
- D. Create an aggregated log export at the organization level, capture the BigQuery job execution logs, create a custom metric based on the totalSlotMs, and create a Stackdriver Monitoring dashboard based on the custom metric

**Correct Answer: D**

**Section:**

#### QUESTION 127

You are operating a streaming Cloud Dataflow pipeline. Your engineers have a new version of the pipeline with a different windowing algorithm and triggering strategy. You want to update the running pipeline with the new version. You want to ensure that no data is lost during the update.



What should you do?

- A. Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to the existing job name
- B. Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to a new unique job name
- C. Stop the Cloud Dataflow pipeline with the Cancel option. Create a new Cloud Dataflow job with the updated code
- D. Stop the Cloud Dataflow pipeline with the Drain option. Create a new Cloud Dataflow job with the updated code

**Correct Answer: A**

**Section:**

#### QUESTION 128

You need to move 2 PB of historical data from an on-premises storage appliance to Cloud Storage within six months, and your outbound network capacity is constrained to 20 Mb/sec. How should you migrate this data to Cloud Storage?

- A. Use Transfer Appliance to copy the data to Cloud Storage
- B. Use gsutil cp ñJ to compress the content being uploaded to Cloud Storage
- C. Create a private URL for the historical data, and then use Storage Transfer Service to copy the data to Cloud Storage
- D. Use trickle or ionice along with gsutil cp to limit the amount of bandwidth gsutil utilizes to less than 20 Mb/sec so it does not interfere with the production traffic

**Correct Answer: A**

**Section:**

#### QUESTION 129

You receive data files in CSV format monthly from a third party. You need to cleanse this data, but every third month the schema of the files changes. Your requirements for implementing these transformations include:

Executing the transformations on a schedule  
Enabling non-developer analysts to modify transformations  
Providing a graphical tool for designing transformations  
What should you do?

- A. Use Cloud Dataprep to build and maintain the transformation recipes, and execute them on a scheduled basis
- B. Load each month's CSV data into BigQuery, and write a SQL query to transform the data to a standard schema. Merge the transformed tables together with a SQL query
- C. Help the analysts write a Cloud Dataflow pipeline in Python to perform the transformation. The Python code should be stored in a revision control system and modified as the incoming data's schema changes
- D. Use Apache Spark on Cloud Dataproc to infer the schema of the CSV file before creating a Dataframe. Then implement the transformations in Spark SQL before writing the data out to Cloud Storage and loading into BigQuery

**Correct Answer: A**

**Section:**

**Explanation:**

Names of columns  
Order of columns  
Column data types  
Data type format  
Example rows of data

A dataset associated with a target is expected to conform to the requirements of the schema. Where there are differences between target schema and dataset schema, a validation indicator (or schema tag) is displayed.

[https://cloud.google.com/dataprep/docs/html/Overview-of-RapidTarget\\_136155049](https://cloud.google.com/dataprep/docs/html/Overview-of-RapidTarget_136155049)

#### QUESTION 130

These primary tool in use, and the data format is Optimized Row Columnar (ORC). All ORC files have been successfully copied to a Cloud Storage bucket. You need to replicate some data to the cluster's local Hadoop

Distributed File

System (HDFS) to maximize performance. What are two ways to start using Hive in Cloud Dataproc? (Choose two.)

- A. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to HDFS. Mount the Hive tables locally.
- B. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to any node of the Dataproc cluster. Mount the Hive tables locally.
- C. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to the master node of the Dataproc cluster. Then run the Hadoop utility to copy them to HDFS. Mount the Hive tables from HDFS.
- D. Leverage Cloud Storage connector for Hadoop to mount the ORC files as external Hive tables. Replicate external Hive tables to the native ones.
- E. Load the ORC files into BigQuery. Leverage BigQuery connector for Hadoop to mount the BigQuery tables as external Hive tables. Replicate external Hive tables to the native ones.

**Correct Answer: B, C**

**Section:**

#### QUESTION 131

You are working on a linear regression model on BigQuery ML to predict a customer's likelihood of purchasing your company's products. Your model uses a city name variable as a key predictive component in order to train and serve the model your data must be organized in columns. You want to prepare your data using the least amount of coding while maintaining the predictable variables.

What should you do?

- A. Use SQL in BigQuery to transform the stale column using a one-hot encoding method, and make each city a column with binary values.
- B. Create a new view with BigQuery that does not include a column which city information.
- C. Cloud Data Fusion to assign each city to a region that is labeled as 1, 2, 3, 4, or 5, and then use that number to represent the city in the model.
- D. Use TensorFlow to create a categorical variable with a vocabulary list. Create the vocabulary file and upload that as part of your model to BigQuery ML.

**Correct Answer: C**

**Section:**

#### QUESTION 132

You are implementing several batch jobs that must be executed on a schedule. These jobs have many interdependent steps that must be executed in a specific order. Portions of the jobs involve executing shell scripts, running Hadoop jobs, and running queries in BigQuery. The jobs are expected to run for many minutes up to several hours. If the steps fail, they must be retried a fixed number of times.

Which service should you use to manage the execution of these jobs?

- A. Cloud Scheduler
- B. Cloud Dataflow
- C. Cloud Functions
- D. Cloud Composer

**Correct Answer: A**

**Section:**

#### QUESTION 133

You work for a shipping company that has distribution centers where packages move on delivery lines to route them properly. The company wants to add cameras to the delivery lines to detect and track any visual damage to the packages in transit. You need to create a way to automate the detection of damaged packages and flag them for human review in real time while the packages are in transit. Which solution should you choose?

- A. Use BigQuery machine learning to be able to train the model at scale, so you can analyze the packages in batches.
- B. Train an AutoML model on your corpus of images, and build an API around that model to integrate with the package tracking applications.
- C. Use the Cloud Vision API to detect for damage, and raise an alert through Cloud Functions. Integrate the package tracking applications with this function.

D. Use TensorFlow to create a model that is trained on your corpus of images. Create a Python notebook in Cloud Datalab that uses this model so you can analyze for damaged packages.

**Correct Answer: A**

**Section:**

#### QUESTION 134

You are migrating your data warehouse to BigQuery. You have migrated all of your data into tables in a dataset. Multiple users from your organization will be using the data. They should only see certain tables based on their team membership. How should you set user permissions?

- A. Assign the users/groups data viewer access at the table level for each table
- B. Create SQL views for each team in the same dataset in which the data resides, and assign the users/groups data viewer access to the SQL views
- C. Create authorized views for each team in the same dataset in which the data resides, and assign the users/groups data viewer access to the authorized views
- D. Create authorized views for each team in datasets created for each team. Assign the authorized views data viewer access to the dataset in which the data resides. Assign the users/groups data viewer access to the datasets in which the authorized views reside

**Correct Answer: A**

**Section:**

#### QUESTION 135

You want to build a managed Hadoop system as your data lake. The data transformation process is composed of a series of Hadoop jobs executed in sequence. To accomplish the design of separating storage from compute, you decided to use the Cloud Storage connector to store all input data, output data, and intermediary data. However, you noticed that one Hadoop job runs very slowly with Cloud Dataproc, when compared with the on-premises bare-metal Hadoop environment (8-core nodes with 100-GB RAM). Analysis shows that this particular Hadoop job is disk I/O intensive. You want to resolve the issue. What should you do?

- A. Allocate sufficient memory to the Hadoop cluster, so that the intermediary data of that particular Hadoop job can be held in memory
- B. Allocate sufficient persistent disk space to the Hadoop cluster, and store the intermediate data of that particular Hadoop job on native HDFS
- C. Allocate more CPU cores of the virtual machine instances of the Hadoop cluster so that the networking bandwidth for each instance can scale up
- D. Allocate additional network interface card (NIC), and configure link aggregation in the operating system to use the combined throughput when working with Cloud Storage

**Correct Answer: A**

**Section:**

#### QUESTION 136

You work for an advertising company, and you've developed a Spark ML model to predict click-through rates at advertisement blocks. You've been developing everything at your on-premises data center, and now your company is migrating to Google Cloud. Your data center will be migrated to BigQuery. You periodically retrain your Spark ML models, so you need to migrate existing training pipelines to Google Cloud. What should you do?

- A. Use Cloud ML Engine for training existing Spark ML models
- B. Rewrite your models on TensorFlow, and start using Cloud ML Engine
- C. Use Cloud Dataproc for training existing Spark ML models, but start reading data directly from BigQuery
- D. Spin up a Spark cluster on Compute Engine, and train Spark ML models on the data exported from BigQuery

**Correct Answer: C**

**Section:**

**Explanation:**

<https://cloud.google.com/dataproc/docs/tutorials/bigquery-sparkml>

#### QUESTION 137

You work for a global shipping company. You want to train a model on 40 TB of data to predict which ships in each geographic region are likely to cause delivery delays on any given day. The model will be based on multiple attributes collected from multiple sources. Telemetry data, including location in GeoJSON format, will be pulled from each ship and loaded every hour. You want to have a dashboard that shows how many and which ships are

likely to cause delays within a region. You want to use a storage solution that has native functionality for prediction and geospatial processing. Which storage solution should you use?

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

**Correct Answer: A**  
**Section:**

#### QUESTION 138

You operate an IoT pipeline built around Apache Kafka that normally receives around 5000 messages per second. You want to use Google Cloud Platform to create an alert as soon as the moving average over 1 hour drops below 4000 messages per second. What should you do?

- A. Consume the stream of data in Cloud Dataflow using Kafka IO. Set a sliding time window of 1 hour every 5 minutes. Compute the average when the window closes, and send an alert if the average is less than 4000 messages.
- B. Consume the stream of data in Cloud Dataflow using Kafka IO. Set a fixed time window of 1 hour. Compute the average when the window closes, and send an alert if the average is less than 4000 messages.
- C. Use Kafka Connect to link your Kafka message queue to Cloud Pub/Sub. Use a Cloud Dataflow template to write your messages from Cloud Pub/Sub to Cloud Bigtable. Use Cloud Scheduler to run a script every hour that counts the number of rows created in Cloud Bigtable in the last hour. If that number falls below 4000, send an alert.
- D. Use Kafka Connect to link your Kafka message queue to Cloud Pub/Sub. Use a Cloud Dataflow template to write your messages from Cloud Pub/Sub to BigQuery. Use Cloud Scheduler to run a script every five minutes that counts the number of rows created in BigQuery in the last hour. If that number falls below 4000, send an alert.

**Correct Answer: C**  
**Section:**



#### QUESTION 139

You plan to deploy Cloud SQL using MySQL. You need to ensure high availability in the event of a zone failure. What should you do?

- A. Create a Cloud SQL instance in one zone, and create a failover replica in another zone within the same region.
- B. Create a Cloud SQL instance in one zone, and create a read replica in another zone within the same region.
- C. Create a Cloud SQL instance in one zone, and configure an external read replica in a zone in a different region.
- D. Create a Cloud SQL instance in a region, and configure automatic backup to a Cloud Storage bucket in the same region.

**Correct Answer: C**  
**Section:**

#### QUESTION 140

Your company is selecting a system to centralize data ingestion and delivery. You are considering messaging and data integration systems to address the requirements. The key requirements are:  
The ability to seek to a particular offset in a topic, possibly back to the start of all data ever captured  
Support for publish/subscribe semantics on hundreds of topics  
Retain per-key ordering  
Which system should you choose?

- A. Apache Kafka
- B. Cloud Storage
- C. Cloud Pub/Sub
- D. Firebase Cloud Messaging

**Correct Answer: A**

**Section:**

**QUESTION 141**

You are planning to migrate your current on-premises Apache Hadoop deployment to the cloud. You need to ensure that the deployment is as fault-tolerant and cost-effective as possible for longrunning batch jobs. You want to use a managed service. What should you do?

- A. Deploy a Cloud Dataproc cluster. Use a standard persistent disk and 50% preemptible workers. Store data in Cloud Storage, and change references in scripts from hdfs:// to gs://
- B. Deploy a Cloud Dataproc cluster. Use an SSD persistent disk and 50% preemptible workers. Store data in Cloud Storage, and change references in scripts from hdfs:// to gs://
- C. Install Hadoop and Spark on a 10-node Compute Engine instance group with standard instances. Install the Cloud Storage connector, and store the data in Cloud Storage. Change references in scripts from hdfs:// to gs://
- D. Install Hadoop and Spark on a 10-node Compute Engine instance group with preemptible instances. Store data in HDFS. Change references in scripts from hdfs:// to gs://

**Correct Answer: A**

**Section:**

**QUESTION 142**

Your team is working on a binary classification problem. You have trained a support vector machine (SVM) classifier with default parameters, and received an area under the Curve (AUC) of 0.87 on the validation set. You want to increase the AUC of the model. What should you do?

- A. Perform hyperparameter tuning
- B. Train a classifier with deep neural networks, because neural networks would always beat SVMs
- C. Deploy the model and measure the real-world AUC; it's always higher because of generalization
- D. Scale predictions you get out of the model (tune a scaling factor as a hyperparameter) in order to get the highest AUC

**Correct Answer: A**

**Section:**

**Explanation:**

<https://towardsdatascience.com/understanding-hyperparameters-and-its-optimisation-techniquesf0debba07568>

**QUESTION 143**

You need to deploy additional dependencies to all of a Cloud Dataproc cluster at startup using an existing initialization action. Company security policies require that Cloud Dataproc nodes do not have access to the Internet so public initialization actions cannot fetch resources. What should you do?

- A. Deploy the Cloud SQL Proxy on the Cloud Dataproc master
- B. Use an SSH tunnel to give the Cloud Dataproc cluster access to the Internet
- C. Copy all dependencies to a Cloud Storage bucket within your VPC security perimeter
- D. Use Resource Manager to add the service account used by the Cloud Dataproc cluster to the Network User role

**Correct Answer: D**

**Section:**

**QUESTION 144**

You need to choose a database for a new project that has the following requirements:

Fully managed

Able to automatically scale up

Transactionally consistent

Able to scale up to 6 TB

Able to be queried using SQL  
Which database do you choose?

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

**Correct Answer: C**  
**Section:**

**QUESTION 145**

You work for a mid-sized enterprise that needs to move its operational system transaction data from an on-premises database to GCP. The database is about 20 TB in size. Which database should you choose?

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

**Correct Answer: A**  
**Section:**

**QUESTION 146**

You need to choose a database to store time series CPU and memory usage for millions of computers. You need to store this data in one-second interval samples. Analysts will be performing real-time, ad hoc analytics against the database.

You want to avoid being charged for every query executed and ensure that the schema design will allow for future growth of the dataset. Which database and data model should you choose?

- A. Create a table in BigQuery, and append the new samples for CPU and memory to the table
- B. Create a wide table in BigQuery, create a column for the sample value at each second, and update the row with the interval for each second
- C. Create a narrow table in Cloud Bigtable with a row key that combines the Computer Engine computer identifier with the sample time at each second
- D. Create a wide table in Cloud Bigtable with a row key that combines the computer identifier with the sample time at each minute, and combine the values for each second as column data.

**Correct Answer: C**  
**Section:**

**Explanation:**

A tall and narrow table has a small number of events per row, which could be just one event, whereas a short and wide table has a large number of events per row. As explained in a moment, tall and narrow tables are best suited for time-series data. For time series, you should generally use tall and narrow tables. This is for two reasons: Storing one event per row makes it easier to run queries against your data. Storing many events per row makes it more likely that the total row size will exceed the recommended maximum (see Rows can be big but are not infinite).

[https://cloud.google.com/bigtable/docs/schema-design-time-series#patterns\\_for\\_row\\_key\\_design](https://cloud.google.com/bigtable/docs/schema-design-time-series#patterns_for_row_key_design)

**QUESTION 147**

You want to archive data in Cloud Storage. Because some data is very sensitive, you want to use the "Trust No One" (TNO) approach to encrypt your data to prevent the cloud provider staff from decrypting your data. What should you do?

- A. Use `gcloud kms keys create` to create a symmetric key. Then use `gcloud kms encrypt` to encrypt each archival file with the key and unique additional authenticated data (AAD). Use `gsutil cp` to upload each encrypted file to the Cloud Storage bucket, and keep the AAD outside of Google Cloud.
- B. Use `gcloud kms keys create` to create a symmetric key. Then use `gcloud kms encrypt` to encrypt each archival file with the key. Use `gsutil cp` to upload each encrypted file to the Cloud Storage bucket. Manually destroy the

key previously used for encryption, and rotate the key once and rotate the key once.

- C. Specify customer-supplied encryption key (CSEK) in the .boto configuration file. Use gsutil cp to upload each archival file to the Cloud Storage bucket. Save the CSEK in Cloud Memorystore as permanent storage of the secret.
- D. Specify customer-supplied encryption key (CSEK) in the .boto configuration file. Use gsutil cp to upload each archival file to the Cloud Storage bucket. Save the CSEK in a different project that only the security team can access.

**Correct Answer: B**

**Section:**

#### QUESTION 148

You have data pipelines running on BigQuery, Cloud Dataflow, and Cloud Dataproc. You need to perform health checks and monitor their behavior, and then notify the team managing the pipelines if they fail. You also need to be able to work across multiple projects. Your preference is to use managed products or features of the platform. What should you do?

- A. Export the information to Cloud Stackdriver, and set up an Alerting policy
- B. Run a Virtual Machine in Compute Engine with Airflow, and export the information to Stackdriver
- C. Export the logs to BigQuery, and set up App Engine to read that information and send emails if you find a failure in the logs
- D. Develop an App Engine application to consume logs using GCP API calls, and send emails if you find a failure in the logs

**Correct Answer: B**

**Section:**

#### QUESTION 149

You work for a large financial institution that is planning to use Dialogflow to create a chatbot for the company's mobile app. You have reviewed old chat logs and lagged each conversation for intent based on each customer's stated intention for contacting customer service. About 70% of customer requests are simple requests that are solved within 10 intents. The remaining 30% of inquiries require much longer, more complicated requests. Which intents should you automate first?

- A. Automate the 10 intents that cover 70% of the requests so that live agents can handle more complicated requests
- B. Automate the more complicated requests first because those require more of the agents' time
- C. Automate a blend of the shortest and longest intents to be representative of all intents
- D. Automate intents in places where common words such as "payment" appear only once so the software isn't confused

**Correct Answer: A**

**Section:**

#### QUESTION 150

You want to rebuild your batch pipeline for structured data on Google Cloud. You are using PySpark to conduct data transformations at scale, but your pipelines are taking over twelve hours to run. To expedite development and pipeline run time, you want to use a serverless tool and SQL syntax. You have already moved your raw data into Cloud Storage. How should you build the pipeline on Google Cloud while meeting speed and processing requirements?

- A. Convert your PySpark commands into SparkSQL queries to transform the data; and then run your pipeline on Dataproc to write the data into BigQuery
- B. Ingest your data into Cloud SQL, convert your PySpark commands into SparkSQL queries to transform the data, and then use federated queries from BigQuery for machine learning.
- C. Ingest your data into BigQuery from Cloud Storage, convert your PySpark commands into BigQuery SQL queries to transform the data, and then write the transformations to a new table
- D. Use Apache Beam Python SDK to build the transformation pipelines, and write the data into BigQuery

**Correct Answer: A**

**Section:**

**QUESTION 151**

You are building a teal-time prediction engine that streams files, which may contain PII (personal identifiable information) data, into Cloud Storage and eventually into BigQuery. You want to ensure that the sensitive data is masked but still maintains referential integrity, because names and emails are often used as join keys. How should you use the Cloud Data Loss Prevention API (DLP API) to ensure that the PII data is not accessible by unauthorized individuals?

- A. Create a pseudonym by replacing the PII data with cryptogenic tokens, and store the nontokenized data in a locked-down bucket.
- B. Redact all PII data, and store a version of the unredacted data in a locked-down bucket.
- C. Scan every table in BigQuery, and mask the data it finds that has PII.
- D. Create a pseudonym by replacing PII data with a cryptographic format-preserving token.

**Correct Answer: A**

**Section:**

**QUESTION 152**

Your company is implementing a data warehouse using BigQuery, and you have been tasked with designing the data model. You move your on-premises sales data warehouse with a star data schema to BigQuery but notice performance issues when querying the data of the past 30 days. Based on Google's recommended practices, what should you do to speed up the query without increasing storage costs?

- A. Denormalize the data.
- B. Shard the data by customer ID.
- C. Materialize the dimensional data in views.
- D. Partition the data by transaction date.

**Correct Answer: C**

**Section:**

**QUESTION 153**

You are using Cloud Bigtable to persist and serve stock market data for each of the major indices. To serve the trading application, you need to access only the most recent stock prices that are streaming in. How should you design your row key and tables to ensure that you can access the data with the most simple query?

- A. Create one unique table for all of the indices, and then use the index and timestamp as the row key design.
- B. Create one unique table for all of the indices, and then use a reverse timestamp as the row key design.
- C. For each index, have a separate table and use a timestamp as the row key design.
- D. For each index, have a separate table and use a reverse timestamp as the row key design.

**Correct Answer: A**

**Section:**

**QUESTION 154**

You have a BigQuery table that ingests data directly from a Pub/Sub subscription. The ingested data is encrypted with a Google-managed encryption key. You need to meet a new organization policy that requires you to use keys from a centralized Cloud Key Management Service (Cloud KMS) project to encrypt data at rest. What should you do?

- A. Create a new BigQuery table by using customer-managed encryption keys (CMEK), and migrate the data from the old BigQuery table.
- B. Create a new BigQuery table and Pub/Sub topic by using customer-managed encryption keys (CMEK), and migrate the data from the old BigQuery table.
- C. Create a new Pub/Sub topic with CMEK and use the existing BigQuery table by using Google-managed encryption key.
- D. Use Cloud KMS encryption key with Dataflow to ingest the existing Pub/Sub subscription to the existing BigQuery table.

**Correct Answer: A**





**Section:****Explanation:**

To use CMEK for BigQuery, you need to create a key ring and a key in Cloud KMS, and then specify the key resource name when creating or updating a BigQuery table. You cannot change the encryption type of an existing table, so you need to create a new table with CMEK and copy the data from the old table with Google-managed encryption key.

[Customer-managed Cloud KMS keys | BigQuery | Google Cloud](#)

[Creating and managing encryption keys | Cloud KMS Documentation | Google Cloud](#)

**QUESTION 155**

You are designing a fault-tolerant architecture to store data in a regional BigQuery dataset. You need to ensure that your application is able to recover from a corruption event in your tables that occurred within the past seven days. You want to adopt managed services with the lowest RPO and most cost-effective solution. What should you do?

- A. Export the data from BigQuery into a new table that excludes the corrupted data.
- B. Migrate your data to multi-region BigQuery buckets.
- C. Access historical data by using time travel in BigQuery.
- D. Create a BigQuery table snapshot on a daily basis.

**Correct Answer: C**

**Section:****Explanation:**

Time travel is a feature of BigQuery that allows you to query and recover data from any point within the past seven days. You can use the FOR SYSTEM\_TIME AS OF clause in your SQL query to specify the timestamp of the data you want to access. This way, you can restore your tables to a previous state before the corruption event occurred. Time travel is automatically enabled for all datasets and does not incur any additional cost or storage.

[Data retention with time travel and fail-safe | BigQuery | Google Cloud](#)

[BigQuery Time Travel: How to access Historical Data? | Easy Steps](#)

