

## Professional-Data-Engineer Dumps

### Google Professional Data Engineer Exam

<https://www.certleader.com/Professional-Data-Engineer-dumps.html>



**NEW QUESTION 1**

- (Exam Topic 1)

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.

**Answer: C**

**NEW QUESTION 2**

- (Exam Topic 1)

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.

**Answer: BC**

**NEW QUESTION 3**

- (Exam Topic 1)

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

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

**Answer: A**

**NEW QUESTION 4**

- (Exam Topic 1)

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

- A. Send the data to Google Cloud Datastore and then export to BigQuery.
- B. Send the data to Google Cloud Pub/Sub, stream Cloud Pub/Sub to Google Cloud Dataflow, and store the data in Google BigQuery.
- C. Send the data to Cloud Storage and then spin up an Apache Hadoop cluster as needed in Google Cloud Dataproc whenever analysis is required.
- D. Export logs in batch to Google Cloud Storage and then spin up a Google Cloud SQL instance, import the data from Cloud Storage, and run an analysis as needed.

**Answer: B**

**NEW QUESTION 5**

- (Exam Topic 1)

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.

**Answer: BCE**

**NEW QUESTION 6**

- (Exam Topic 1)

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

- A. The message body for the sensor event is too large.
- B. Your custom endpoint has an out-of-date SSL certificate.
- C. The Cloud Pub/Sub topic has too many messages published to it.

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

**Answer: B**

#### NEW QUESTION 7

- (Exam Topic 1)

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

- A. Disable writes to certain tables.
- B. Restrict access to tables by role.
- C. Ensure that the data is encrypted at all times.
- D. Restrict BigQuery API access to approved users.
- E. Segregate data across multiple tables or databases.
- F. Use Google Stackdriver Audit Logging to determine policy violations.

**Answer: BDF**

#### NEW QUESTION 8

- (Exam Topic 1)

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.

**Answer: C**

#### NEW QUESTION 9

- (Exam Topic 1)

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

- A. Run a local version of Jupiter on the laptop.
- B. Grant the user access to Google Cloud Shell.
- C. Host a visualization tool on a VM on Google Compute Engine.
- D. Deploy Google Cloud Datalab to a virtual machine (VM) on Google Compute Engine.

**Answer: B**

#### NEW QUESTION 10

- (Exam Topic 1)

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.

**Answer: D**

#### NEW QUESTION 10

- (Exam Topic 1)

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

- A. Update the current pipeline and use the drain flag.
- B. Update the current pipeline and provide the transform mapping JSON object.
- C. Create a new pipeline that has the same Cloud Pub/Sub subscription and cancel the old pipeline.
- D. Create a new pipeline that has a new Cloud Pub/Sub subscription and cancel the old pipeline.

**Answer: D**

#### NEW QUESTION 13

- (Exam Topic 1)

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

```
# Syntax error : Expected end of statement but got "-" at [4:11] SELECT age
FROM
bigquery-public-data.noaa_gsod.gsod WHERE
age != 99
```

AND\_TABLE\_SUFFIX = '1929' ORDER BY  
age DESC

Which table name will make the SQL statement work correctly?

- A. 'bigquery-public-data.noaa\_gsod.gsod'
- B. bigquery-public-data.noaa\_gsod.gsod\*
- C. 'bigquery-public-data.noaa\_gsod.gsod'\*
- D. 'bigquery-public-data.noaa\_gsod.gsod\*'

**Answer: D**

#### NEW QUESTION 17

- (Exam Topic 1)

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.

**Answer: A**

#### NEW QUESTION 22

- (Exam Topic 2)

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.

**Answer: B**

#### NEW QUESTION 25

- (Exam Topic 2)

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

- A. Export the data into a Google Sheet for virtualization.
- B. Create an additional table with only the necessary columns.
- C. Create a view on the table to present to the virtualization tool.
- D. Create identity and access management (IAM) roles on the appropriate columns, so only they appear in a query.

**Answer: C**

#### NEW QUESTION 26

- (Exam Topic 2)

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

- A. Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage
- B. Cloud Pub/Sub, Cloud Dataflow, and Local SSD
- C. Cloud Pub/Sub, Cloud SQL, and Cloud Storage
- D. Cloud Load Balancing, Cloud Dataflow, and Cloud Storage

**Answer: C**

#### NEW QUESTION 27

- (Exam Topic 3)

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

- A. Rowkey: date#device\_idColumn data: data\_point
- B. Rowkey: dateColumn data: device\_id, data\_point
- C. Rowkey: device\_idColumn data: date, data\_point
- D. Rowkey: data\_pointColumn data: device\_id, date
- E. Rowkey: date#data\_pointColumn data: device\_id

**Answer: D**

**NEW QUESTION 31**

- (Exam Topic 3)

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

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

- A. Ensure all the tables are included in global dataset.
- B. Ensure each table is included in a dataset for a region.
- C. Adjust the settings for each table to allow a related region-based security group view access.
- D. Adjust the settings for each view to allow a related region-based security group view access.
- E. Adjust the settings for each dataset to allow a related region-based security group view access.

**Answer: BD**

**NEW QUESTION 32**

- (Exam Topic 4)

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.
- C. Run an UPDATE statement that updates the FullName column for each user with the concatenation of the FirstName and LastName values.
- D. 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.
- E. Use BigQuery to export the data for the table to a CSV file.
- F. 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.
- G. Run a BigQuery load job to load the new CSV file into BigQuery.

**Answer: C**

**NEW QUESTION 37**

- (Exam Topic 4)

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

- A. Load the data every 30 minutes into a new partitioned table in BigQuery.
- B. Store and update the data in a regional Google Cloud Storage bucket and create a federated data source in BigQuery.
- C. Store the data in Google Cloud Datastore.
- D. Use Google Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Cloud Datastore.
- E. Store the data in a file in a regional Google Cloud Storage bucket.
- F. Use Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Google Cloud Storage.

**Answer: A**

**NEW QUESTION 41**

- (Exam Topic 4)

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

- A. Change the processing job to use Google Cloud Dataproc instead.
- B. Manually start the Cloud Dataflow job each morning when you get into the office.
- C. Create a cron job with Google App Engine Cron Service to run the Cloud Dataflow job.
- D. Configure the Cloud Dataflow job as a streaming job so that it processes the log data immediately.

**Answer: C**

**NEW QUESTION 43**

- (Exam Topic 4)

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

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

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

B. Manually configure the index in your index config as follows:

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

C. Set the following in your entity options: `exclude_from_indexes = 'actors, tags'`

D. Set the following in your entity options: `exclude_from_indexes = 'date_published'`

- A. Option A
- B. Option B.
- C. Option C
- D. Option D

**Answer: A**

#### NEW QUESTION 47

- (Exam Topic 5)

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

- A. `categorical_column_with_vocabulary_list`
- B. `categorical_column_with_hash_bucket`
- C. `categorical_column_with_unknown_values`
- D. `sparse_column_with_keys`

**Answer: B**

#### Explanation:

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

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

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

#### NEW QUESTION 52

- (Exam Topic 5)

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

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

**Answer: C**

#### Explanation:

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

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

#### NEW QUESTION 54

- (Exam Topic 5)

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

- A. Storage, queries, and streaming inserts
- B. Storage, queries, and loading data from a file
- C. Storage, queries, and exporting data
- D. Queries and streaming inserts

**Answer:** A

**Explanation:**

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

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

**NEW QUESTION 59**

- (Exam Topic 5)

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

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

**Answer:** B

**Explanation:**

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

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

Alternatively, you can write

a Cloud Dataflow or Hadoop MapReduce job that copies the data from one instance to another. Reference: <https://cloud.google.com/bigtable/docs/choosing-ssd-hdd->

**NEW QUESTION 63**

- (Exam Topic 5)

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

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

**Answer:** A

**Explanation:**

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

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

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

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

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

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

**NEW QUESTION 68**

- (Exam Topic 5)

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

- A. Splitting tables into multiple tables; putting data in partitions
- B. Splitting tables into multiple tables; putting data in partitions; using the LIMIT clause
- C. Putting data in partitions; using the LIMIT clause
- D. Splitting tables into multiple tables; using the LIMIT clause

**Answer:** A

**Explanation:**

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

If you use the LIMIT clause, BigQuery will still process the entire table. Reference: <https://cloud.google.com/bigquery/docs/partitioned-tables>

**NEW QUESTION 70**

- (Exam Topic 5)

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

- A. Preemptible workers cannot use persistent disk.
- B. Preemptible workers cannot store data.

- C. If a preemptible worker is reclaimed, then a replacement worker must be added manually.
- D. A Dataproc cluster cannot have only preemptible workers.

**Answer:** BD

**Explanation:**

The following rules will apply when you use preemptible workers with a Cloud Dataproc cluster: Processing only—Since preemptibles can be reclaimed at any time, preemptible workers do not store data.

Preemptibles added to a Cloud Dataproc cluster only function as processing nodes.

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

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

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

<https://cloud.google.com/dataproc/docs/concepts/preemptible-vm>

**NEW QUESTION 75**

- (Exam Topic 5)

When running a pipeline that has a BigQuery source, on your local machine, you continue to get permission denied errors. What could be the reason for that?

- A. Your gcloud does not have access to the BigQuery resources
- B. BigQuery cannot be accessed from local machines
- C. You are missing gcloud on your machine
- D. Pipelines cannot be run locally

**Answer:** A

**Explanation:**

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

Reference:

<https://cloud.google.com/dataflow/java-sdk/JavaDoc/com/google/cloud/dataflow/sdk/runners/DirectPipelineRun>

**NEW QUESTION 80**

- (Exam Topic 5)

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

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

**Answer:** D

**Explanation:**

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

Reference:

<https://cloud.google.com/dataflow/java-sdk/JavaDoc/com/google/cloud/dataflow/sdk/transforms/windowing/Tri>

**NEW QUESTION 82**

- (Exam Topic 5)

Does Dataflow process batch data pipelines or streaming data pipelines?

- A. Only Batch Data Pipelines
- B. Both Batch and Streaming Data Pipelines
- C. Only Streaming Data Pipelines
- D. None of the above

**Answer:** B

**Explanation:**

Dataflow is a unified processing model, and can execute both streaming and batch data pipelines Reference: <https://cloud.google.com/dataflow/>

**NEW QUESTION 87**

- (Exam Topic 5)

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

- A. Standard SQL is the preferred query language for BigQuery.
- B. If you write a query in Legacy SQL, it might generate an error if you try to run it with Standard SQL.
- C. One difference between the two query languages is how you specify fully-qualified table names (i. table names that include their associated project name).
- D. You need to set a query language for each dataset and the default is Standard SQL.

**Answer:** D

**Explanation:**

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

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

In legacy SQL, to query a table with a project-qualified name, you use a colon, :, as a separator. In standard SQL, you use a period, ., instead. Due to the differences in syntax between the two query languages (such as with project-qualified table names), if you write a query in Legacy SQL, it might generate an error if you try to run it with Standard SQL.

Reference:

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

**NEW QUESTION 90**

- (Exam Topic 5)

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

- A. Trigger based on element size in bytes
- B. Trigger that is a combination of other triggers
- C. Trigger based on element count
- D. Trigger based on time

**Answer:** A

**Explanation:**

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

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

**NEW QUESTION 95**

- (Exam Topic 5)

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

- A. gcloud ml-engine local train
- B. gcloud ml-engine jobs submit training
- C. gcloud ml-engine jobs submit training local
- D. You can't run a TensorFlow program on your own computer using Cloud ML Engine .

**Answer:** A

**Explanation:**

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

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

This is especially useful in the case of testing distributed models, as it allows you to validate that you are

properly interacting with the Cloud ML Engine cluster configuration. Reference: <https://cloud.google.com/sdk/gcloud/reference/ml-engine/local/train>

**NEW QUESTION 98**

- (Exam Topic 5)

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

- A. Use K-means Clustering to detect faces in the pixels.
- B. Use feature engineering to add features for eyes, noses, and mouths to the input data.
- C. Use deep learning by creating a neural network with multiple hidden layers to automatically detect features of faces.
- D. Build a neural network with an input layer of pixels, a hidden layer, and an output layer with two categories.

**Answer:** C

**Explanation:**

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

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

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

Feature engineering is difficult to perform on raw image data.

K- means Clustering is an unsupervised learning method used to categorize unlabeled data. Reference: <https://deeplearning4j.org/neuralnet-overview>

**NEW QUESTION 102**

- (Exam Topic 5)

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

- A. Dataproc Worker
- B. Dataproc Viewer
- C. Dataproc Runner
- D. Dataproc Editor

**Answer:** A

**Explanation:**

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

Reference: [https://cloud.google.com/dataproc/docs/concepts/service-accounts#important\\_notes](https://cloud.google.com/dataproc/docs/concepts/service-accounts#important_notes)

**NEW QUESTION 103**

- (Exam Topic 5)

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

- A. Reduces the amount of data processed, reduces the amount of storage required
- B. Increases query speed, makes queries simpler
- C. Reduces the amount of storage required, increases query speed
- D. Reduces the amount of data processed, increases query speed

**Answer: B**

**Explanation:**

Denormalization increases query speed for tables with billions of rows because BigQuery's performance degrades when doing JOINS on large tables, but with a denormalized data structure, you don't have to use JOINS, since all of the data has been combined into one table. Denormalization also makes queries simpler because you do not have to use JOIN clauses.

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

Reference:

[https://cloud.google.com/solutions/bigquery-data-warehouse#denormalizing\\_data](https://cloud.google.com/solutions/bigquery-data-warehouse#denormalizing_data)

**NEW QUESTION 108**

- (Exam Topic 5)

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

- A. File upload
- B. Google Drive
- C. Google Cloud Storage
- D. Google Cloud SQL

**Answer: D**

**Explanation:**

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

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

**NEW QUESTION 113**

- (Exam Topic 5)

Which of the following is not true about Dataflow pipelines?

- A. Pipelines are a set of operations
- B. Pipelines represent a data processing job
- C. Pipelines represent a directed graph of steps
- D. Pipelines can share data between instances

**Answer: D**

**Explanation:**

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

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

**NEW QUESTION 117**

- (Exam Topic 5)

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

- A. LocalRunner
- B. DirectPipelineRunner
- C. MachineRunner
- D. LocalPipelineRunner

**Answer: B**

**Explanation:**

DirectPipelineRunner allows you to execute operations in the pipeline directly, without any optimization. Useful for small local execution and tests

Reference:

<https://cloud.google.com/dataflow/java-sdk/JavaDoc/com/google/cloud/dataflow/sdk/runners/DirectPipelineRun>

**NEW QUESTION 120**

- (Exam Topic 5)

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

**Answer:**

C

**Explanation:**

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

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

**NEW QUESTION 122**

- (Exam Topic 5)

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

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

**Answer: C**

**Explanation:**

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

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

**NEW QUESTION 123**

- (Exam Topic 5)

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

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

**Answer: C**

**Explanation:**

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

Reference:

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

**NEW QUESTION 125**

- (Exam Topic 5)

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

- A. Hidden feature layers
- B. Feature prioritization
- C. Crossed feature columns
- D. Bucketization of a continuous feature

**Answer: CD**

**Explanation:**

Selecting and crafting the right set of feature columns is key to learning an effective model. Bucketization is a process of dividing the entire range of a continuous feature into a set of consecutive

bins/buckets, and then converting the original numerical feature into a bucket ID (as a categorical feature) depending on which bucket that value falls into.

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

Reference: [https://www.tensorflow.org/tutorials/wide#selecting\\_and\\_engineering\\_features\\_for\\_the\\_model](https://www.tensorflow.org/tutorials/wide#selecting_and_engineering_features_for_the_model)

**NEW QUESTION 127**

- (Exam Topic 5)

You have a job that you want to cancel. It is a streaming pipeline, and you want to ensure that any data that is in-flight is processed and written to the output.

Which of the following commands can you use on the Dataflow monitoring console to stop the pipeline job?

- A. Cancel
- B. Drain
- C. Stop
- D. Finish

**Answer: B**

**Explanation:**

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

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

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

**NEW QUESTION 130**

- (Exam Topic 5)

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

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

**Answer: C**

**Explanation:**

The YARN ResourceManager and the HDFS NameNode interfaces are available on a Cloud Dataproc cluster master node. The cluster master-host-name is the name of your Cloud Dataproc cluster followed by an -m suffix—for example, if your cluster is named "my-cluster", the master-host-name would be "my-cluster-m". Reference: <https://cloud.google.com/dataproc/docs/concepts/cluster-web-interfaces#interfaces>

**NEW QUESTION 132**

- (Exam Topic 5)

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

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

**Answer: C**

**Explanation:**

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

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

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

**NEW QUESTION 137**

- (Exam Topic 5)

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

- A. If you have existing data in a separate file for each day, then create a partitioned table and upload each file into the appropriate partition.
- B. Run a query to get the records for a specific day from an existing table and for the destination table, specify a partitioned table ending with the day in the format "\$YYYYMMDD".
- C. Create a partitioned table and stream new records to it every day.
- D. Use ORDER BY to put a table's rows into chronological order and then change the table's type to "Partitioned".

**Answer: D**

**Explanation:**

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

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

**NEW QUESTION 140**

- (Exam Topic 5)

Which action can a Cloud Dataproc Viewer perform?

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

**Answer: D**

**Explanation:**

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

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

**NEW QUESTION 144**

- (Exam Topic 5)

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

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

**Answer: B**

**Explanation:**

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

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

**NEW QUESTION 148**

- (Exam Topic 5)

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

- A. The wide model is used for memorization, while the deep model is used for generalization.
- B. A good use for the wide and deep model is a recommender system.
- C. The wide model is used for generalization, while the deep model is used for memorization.
- D. A good use for the wide and deep model is a small-scale linear regression problem.

**Answer:** AB

**Explanation:**

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

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

**NEW QUESTION 153**

- (Exam Topic 5)

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

- A. zone
- B. node
- C. label
- D. type

**Answer:** A

**Explanation:**

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

The project in which the cluster will be created

The region to use

The name of the cluster

The zone in which the cluster will be created

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

Reference:

[https://cloud.google.com/dataproc/docs/tutorials/python-library-example#create\\_a\\_new\\_cloud\\_dataproc\\_cluste](https://cloud.google.com/dataproc/docs/tutorials/python-library-example#create_a_new_cloud_dataproc_cluste)

**NEW QUESTION 158**

- (Exam Topic 5)

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

Tom,555 X street Tim,553 Y street Sam, 111 Z street

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

- A. ParDo
- B. Sink API
- C. Source API
- D. Data extraction

**Answer:** A

**Explanation:**

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

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

**NEW QUESTION 160**

- (Exam Topic 5)

Which software libraries are supported by Cloud Machine Learning Engine?

- A. Theano and TensorFlow
- B. Theano and Torch
- C. TensorFlow
- D. TensorFlow and Torch

**Answer:** C

**Explanation:**

Cloud ML Engine mainly does two things:

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

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

Reference: [https://cloud.google.com/ml-engine/docs/technical-overview#what\\_it\\_does](https://cloud.google.com/ml-engine/docs/technical-overview#what_it_does)

**NEW QUESTION 164**

- (Exam Topic 5)

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

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

How would you correct the error?

- A. Add ", UNNEST(person)" before the WHERE clause.
- B. Change "person" to "person.city".
- C. Change "person" to "city.person".
- D. Add ", UNNEST(city)" before the WHERE clause.

**Answer:** A

**Explanation:**

To access the person.city column, you need to "UNNEST(person)" and JOIN it to table1 using a comma. Reference: [https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql#nested\\_repeated\\_resu](https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql#nested_repeated_resu)

**NEW QUESTION 167**

- (Exam Topic 5)

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

- A. Weights
- B. Biases
- C. Continuous features
- D. Input values

**Answer:** AB

**Explanation:**

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

Reference:

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

**NEW QUESTION 171**

- (Exam Topic 5)

Which of the following is not possible using primitive roles?

- A. Give a user viewer access to BigQuery and owner access to Google Compute Engine instances.
- B. Give UserA owner access and UserB editor access for all datasets in a project.
- C. Give a user access to view all datasets in a project, but not run queries on them.
- D. Give GroupA owner access and GroupB editor access for all datasets in a project.

**Answer:** C

**Explanation:**

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

Reference: [https://cloud.google.com/bigquery/docs/access-control#primitive\\_iam\\_roles](https://cloud.google.com/bigquery/docs/access-control#primitive_iam_roles)

**NEW QUESTION 172**

- (Exam Topic 6)

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

- A. Use the TABLE\_DATE\_RANGE function
- B. Use the WHERE\_PARTITIONTIME pseudo column
- C. Use WHERE date BETWEEN YYYY-MM-DD AND YYYY-MM-DD
- D. Use SELECT IF.(date >= YYYY-MM-DD AND date <= YYYY-MM-DD)

**Answer:** A

**Explanation:**

Reference:

<https://cloud.google.com/blog/products/gcp/using-bigquery-and-firebase-analytics-to-understandyour-mobile-ap>

**NEW QUESTION 177**

- (Exam Topic 6)

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

**Answer:** DE

**NEW QUESTION 180**

- (Exam Topic 6)

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.

**Answer:** D

**NEW QUESTION 181**

- (Exam Topic 6)

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 the globe have poor internet connectivity, messages sometimes batch at the edge, come in all at once, and cause a spike in load on your Kafka cluster. This is becoming difficult to manage and prohibitively expensive. What is the Google-recommended cloud native architecture for this scenario?

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

**Answer:** C

**NEW QUESTION 185**

- (Exam Topic 6)

A shipping company has live package-tracking data that is sent to an Apache Kafka stream in real time. This is then loaded into BigQuery. Analysts in your company want to query the tracking data in BigQuery to analyze geospatial trends in the lifecycle of a package. The table was originally created with ingest-date partitioning. Over time, the query processing time has increased. You need to implement a change that would improve query performance in BigQuery. What should you do?

- A. Implement clustering in BigQuery on the ingest date column.
- B. Implement clustering in BigQuery on the package-tracking ID column.
- C. Tier older data onto Cloud Storage files, and leverage extended tables.
- D. Re-create the table using data partitioning on the package delivery date.

**Answer:** A

**NEW QUESTION 186**

- (Exam Topic 6)

You have a requirement to insert minute-resolution data from 50,000 sensors into a BigQuery table. You expect significant growth in data volume and need the data to be available within 1 minute of ingestion for real-time analysis of aggregated trends. What should you do?

- A. Use bq load to load a batch of sensor data every 60 seconds.
- B. Use a Cloud Dataflow pipeline to stream data into the BigQuery table.
- C. Use the INSERT statement to insert a batch of data every 60 seconds.
- D. Use the MERGE statement to apply updates in batch every 60 seconds.

**Answer:** C

**NEW QUESTION 191**

- (Exam Topic 6)

You're training a model to predict housing prices based on an available dataset with real estate properties. Your plan is to train a fully connected neural net, and you've discovered that the dataset contains latitude and longitude of the property. Real estate professionals have told you that the location of the property is highly influential on price, so you'd like to engineer a feature that incorporates this physical dependency. What should you do?

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

**Answer:** B

**Explanation:**

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

**NEW QUESTION 194**

- (Exam Topic 6)

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.

**Answer: B**

#### NEW QUESTION 199

- (Exam Topic 6)

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

- Real-time event stream
- ANSI SQL access to real-time stream and historical data
- Batch historical exports

Which solution should you use?

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

**Answer: A**

#### NEW QUESTION 200

- (Exam Topic 6)

You are a retailer that wants to integrate your online sales capabilities with different in-home assistants, such as Google Home. You need to interpret customer voice commands and issue an order to the backend systems. Which solutions should you choose?

- A. Cloud Speech-to-Text API
- B. Cloud Natural Language API
- C. Dialogflow Enterprise Edition
- D. Cloud AutoML Natural Language

**Answer: D**

#### NEW QUESTION 205

- (Exam Topic 6)

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

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

**Answer: A**

#### Explanation:

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

#### NEW QUESTION 206

- (Exam Topic 6)

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

**Answer: C**

#### NEW QUESTION 208

- (Exam Topic 6)

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"

**Answer:** AE

#### NEW QUESTION 212

- (Exam Topic 6)

You operate a logistics company, and you want to improve event delivery reliability for vehicle-based sensors. You operate small data centers around the world to capture these events, but leased lines that provide connectivity from your event collection infrastructure to your event processing infrastructure are unreliable, with unpredictable latency. You want to address this issue in the most cost-effective way. What should you do?

- A. Deploy small Kafka clusters in your data centers to buffer events.
- B. Have the data acquisition devices publish data to Cloud Pub/Sub.
- C. Establish a Cloud Interconnect between all remote data centers and Google.
- D. Write a Cloud Dataflow pipeline that aggregates all data in session windows.

**Answer:** A

#### NEW QUESTION 216

- (Exam Topic 6)

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

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

**Answer:** A

#### NEW QUESTION 220

- (Exam Topic 6)

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

- Fully managed
  - Able to automatically scale up
  - Transactionally consistent
  - Able to scale up to 6 TB
  - Able to be queried using SQL
- Which database do you choose?

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

**Answer:** C

#### NEW QUESTION 225

- (Exam Topic 6)

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?

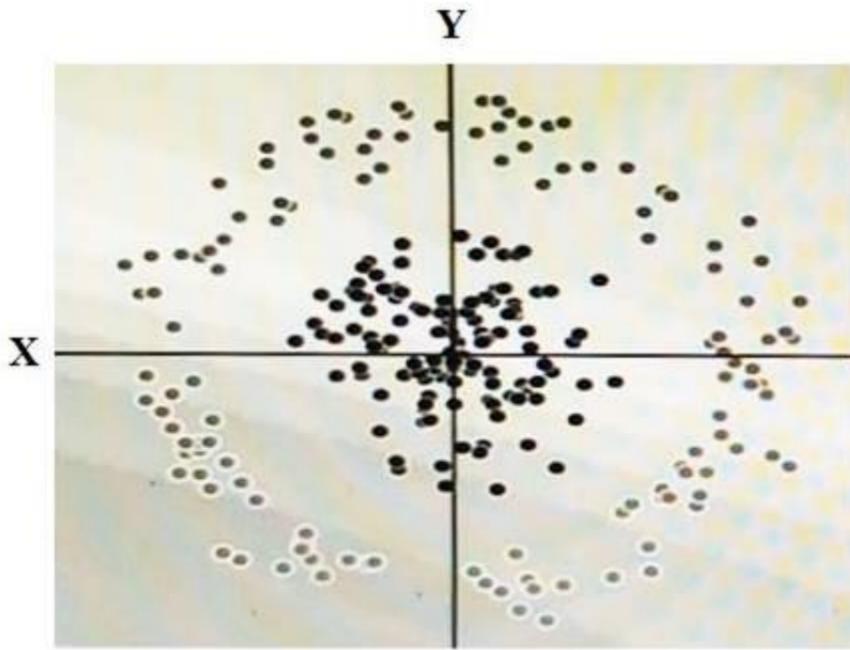
- A. Create a dedicated service account, and use encryption at rest to reference your data stored in your Compute Engine cluster instances as part of your API service calls.
- B. Create encryption keys in Cloud Key Management Service
- C. Use those keys to encrypt your data in all of the Compute Engine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

**Answer:** C

#### NEW QUESTION 228

- (Exam Topic 6)

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



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

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

**Answer:** D

**NEW QUESTION 231**

- (Exam Topic 6)

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
- D. Use the service account's private key to access the dataset
- E. Create a dummy user and grant dataset access to that user
- F. Store the username and password for that user in a file on the files system, and use those credentials to access the BigQuery dataset

**Answer:** C

**NEW QUESTION 233**

- (Exam Topic 6)

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

**Answer:** B

**NEW QUESTION 238**

- (Exam Topic 6)

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

**Answer:** D

**NEW QUESTION 242**

- (Exam Topic 6)

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

**Answer:** A

**NEW QUESTION 247**

- (Exam Topic 6)

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

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

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



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

- A. Users are running too many concurrent queries in the system
- B. The [myproject:mydataset.mytable] table has too many partitions
- C. Either the state or the city columns in the [myproject:mydataset.mytable] table have too many NULL values
- D. Most rows in the [myproject:mydataset.mytable] table have the same value in the country column, causing data skew

**Answer:** A

**NEW QUESTION 251**

- (Exam Topic 6)

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

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

**Answer:** D

**NEW QUESTION 256**

- (Exam Topic 6)

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

**Answer:** B

**NEW QUESTION 259**

- (Exam Topic 6)

You used Cloud Dataprep to create a recipe on a sample of data in a BigQuery table. You want to reuse this recipe on a daily upload of data with the same schema, after the load job with variable execution time completes. What should you do?

- A. Create a cron schedule in Cloud Dataprep.
- B. Create an App Engine cron job to schedule the execution of the Cloud Dataprep job.
- C. Export the recipe as a Cloud Dataprep template, and create a job in Cloud Scheduler.
- D. Export the Cloud Dataprep job as a Cloud Dataflow template, and incorporate it into a Cloud Composer job.

**Answer:** C

**NEW QUESTION 262**

- (Exam Topic 6)

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

**Answer: C**

**NEW QUESTION 265**

- (Exam Topic 6)

You have a query that filters a BigQuery table using a WHERE clause on timestamp and ID columns. By using bq query --dry\_run you learn that the query triggers a full scan of the table, even though the filter on timestamp and ID select a tiny fraction of the overall data. You want to reduce the amount of data scanned by BigQuery with minimal changes to existing SQL queries. What should you do?

- A. Create a separate table for each ID.
- B. Use the LIMIT keyword to reduce the number of rows returned.
- C. Recreate the table with a partitioning column and clustering column.
- D. Use the bq query --maximum\_bytes\_billed flag to restrict the number of bytes billed.

**Answer: B**

**NEW QUESTION 270**

- (Exam Topic 6)

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
- B. Restrict access to Stackdriver Logging via Cloud IAM roles.
- C. Export the data access logs via a project-level export sink to a Cloud Storage bucket in the Data Analysts' project
- D. Restrict access to the Cloud Storage bucket.
- E. Export the data access logs via a project-level export sink to a Cloud Storage bucket in a newly created project for audit log
- F. Restrict access to the project with the exported logs.
- G. Export the data access logs via an aggregated export sink to a Cloud Storage bucket in a newly created project for audit log
- H. Restrict access to the project that contains the exported logs.

**Answer: D**

**NEW QUESTION 274**

- (Exam Topic 6)

Your company is selecting a system to centralize data ingestion and delivery. You are considering messaging and data integration systems to address the requirements. The key requirements are:

- The ability to seek to a particular offset in a topic, possibly back to the start of all data ever captured
- Support for publish/subscribe semantics on hundreds of topics
- Retain per-key ordering Which system should you choose?

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

**Answer: A**

**NEW QUESTION 275**

- (Exam Topic 6)

You need to migrate a 2TB relational database to Google Cloud Platform. You do not have the resources to significantly refactor the application that uses this database and cost to operate is of primary concern.

Which service do you select for storing and serving your data?

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

**Answer: D**

**NEW QUESTION 280**

- (Exam Topic 6)

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

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

**Answer:** A

**NEW QUESTION 282**

- (Exam Topic 6)

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

**Answer:** B

**NEW QUESTION 283**

- (Exam Topic 6)

You work for a bank. You have a labelled dataset that contains information on already granted loan application and whether these applications have been defaulted. You have been asked to train a model to predict default rates for credit applicants. What should you do?

- A. Increase the size of the dataset by collecting additional data.
- B. Train a linear regression to predict a credit default risk score.
- C. Remove the bias from the data and collect applications that have been declined loans.
- D. Match loan applicants with their social profiles to enable feature engineering.

**Answer:** B

**NEW QUESTION 287**

- (Exam Topic 6)

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

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

**Answer:** A

**NEW QUESTION 289**

- (Exam Topic 6)

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 all streaming inserts. What is the most likely cause of this problem?

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

**Answer:** C

**NEW QUESTION 291**

- (Exam Topic 6)

An organization maintains a Google BigQuery dataset that contains tables with user-level data. They want to expose aggregates of this data to other Google Cloud projects, while still controlling access to the user-level data. Additionally, they need to minimize their overall storage cost and ensure the analysis cost for other projects is assigned to those projects. What should they do?

- A. Create and share an authorized view that provides the aggregate results.
- B. Create and share a new dataset and view that provides the aggregate results.
- C. Create and share a new dataset and table that contains the aggregate results.
- D. Create dataViewer Identity and Access Management (IAM) roles on the dataset to enable sharing.

**Answer:** D

**Explanation:**

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

**NEW QUESTION 294**

- (Exam Topic 6)

Your company receives both batch- and stream-based event data. You want to process the data using Google Cloud Dataflow over a predictable time period. However, you realize that in some instances data can arrive late or out of order. How should you design your Cloud Dataflow pipeline to handle data that is late or out of order?

- A. Set a single global window to capture all the data.
- B. Set sliding windows to capture all the lagged data.
- C. Use watermarks and timestamps to capture the lagged data.
- D. Ensure every datasource type (stream or batch) has a timestamp, and use the timestamps to define the logic for lagged data.

**Answer: B**

**NEW QUESTION 297**

- (Exam Topic 6)

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

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

**Answer: A**

**NEW QUESTION 299**

- (Exam Topic 6)

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.
- B. Store the data in HDFS, and perform analysis as needed.
- C. Store the data in BigQuery.
- D. Access the data using the BigQuery Connector or Cloud Dataproc and Compute Engine.
- E. Store the data in a regional Cloud Storage bucket.
- F. Access the bucket directly using Cloud Dataproc, BigQuery, and Compute Engine.
- G. Store the data in a multi-regional Cloud Storage bucket.
- H. Access the data directly using Cloud Dataproc, BigQuery, and Compute Engine.

**Answer: C**

**NEW QUESTION 304**

- (Exam Topic 6)

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 transform MapReduce jobs to apply sensor calibration before they do anything else.
- B. Introduce a new MapReduce job to apply sensor calibration to raw data, and ensure all other MapReduce jobs are chained after this.
- C. Add sensor calibration data to the output of the ETL process, and document that all users need to apply sensor calibration themselves.
- D. Develop an algorithm through simulation to predict variance of data output from the last MapReduce job based on calibration factors, and apply the correction to all data.

**Answer: A**

**NEW QUESTION 307**

- (Exam Topic 6)

You have a data stored in BigQuery. The data in the BigQuery dataset must be highly available. You need to define a storage, backup, and recovery strategy of this data that minimizes cost. How should you configure the BigQuery table?

- A. Set the BigQuery dataset to be regional.
- B. In the event of an emergency, use a point-in-time snapshot to recover the data.
- C. Set the BigQuery dataset to be regional.
- D. Create a scheduled query to make copies of the data to tables suffixed with the time of the backup.
- E. In the event of an emergency, use the backup copy of the table.
- F. Set the BigQuery dataset to be multi-regional.
- G. In the event of an emergency, use a point-in-time snapshot to recover the data.

- H. Set the BigQuery dataset to be multi-regiona
- I. Create a scheduled query to make copies of the data to tables suffixed with the time of the backu
- J. In the event of an emergency, use the backup copy of the table.

**Answer: B**

#### NEW QUESTION 310

- (Exam Topic 6)

You want to migrate an on-premises Hadoop system to Cloud Dataproc. Hive is the primary tool in use, and the data format is Optimized Row Columnar (ORC). All ORC files have been successfully copied to a Cloud Storage bucket. You need to replicate some data to the cluster's local Hadoop Distributed File System (HDFS) to maximize performance. What are two ways to start using Hive in Cloud Dataproc? (Choose two.)

- A. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to HDF
- B. Mount the Hive tables locally.
- C. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to any node of the Dataproc cluste
- D. Mount the Hive tables locally.
- E. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to the master node of the Dataproc cluste
- F. Then run the Hadoop utility to copy them do HDF
- G. Mount the Hive tables from HDFS.
- H. Leverage Cloud Storage connector for Hadoop to mount the ORC files as external Hive table
- I. Replicate external Hive tables to the native ones.
- J. Load the ORC files into BigQuer
- K. Leverage BigQuery connector for Hadoop to mount the BigQuery tables as external Hive table
- L. Replicate external Hive tables to the native ones.

**Answer: BC**

#### NEW QUESTION 313

- (Exam Topic 6)

You are deploying MariaDB SQL databases on GCE VM Instances and need to configure monitoring and alerting. You want to collect metrics including network connections, disk IO and replication status from MariaDB with minimal development effort and use StackDriver for dashboards and alerts. What should you do?

- A. Install the OpenCensus Agent and create a custom metric collection application with a StackDriver exporter.
- B. Place the MariaDB instances in an Instance Group with a Health Check.
- C. Install the StackDriver Logging Agent and configure fluentd in\_tail plugin to read MariaDB logs.
- D. Install the StackDriver Agent and configure the MySQL plugin.

**Answer: C**

#### NEW QUESTION 315

- (Exam Topic 6)

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 BigQuer
- E. Keep this ratio as 80% warm and 20% active.

**Answer: D**

#### NEW QUESTION 319

- (Exam Topic 6)

You are creating a new pipeline in Google Cloud to stream IoT data from Cloud Pub/Sub through Cloud Dataflow to BigQuery. While previewing the data, you notice that roughly 2% of the data appears to be corrupt. You need to modify the Cloud Dataflow pipeline to filter out this corrupt data. What should you do?

- A. Add a SideInput that returns a Boolean if the element is corrupt.
- B. Add a ParDo transform in Cloud Dataflow to discard corrupt elements.
- C. Add a Partition transform in Cloud Dataflow to separate valid data from corrupt data.
- D. Add a GroupByKey transform in Cloud Dataflow to group all of the valid data together and discard the rest.

**Answer: B**

#### NEW QUESTION 323

- (Exam Topic 6)

You have an Apache Kafka Cluster on-prem with topics containing web application logs. You need to replicate the data to Google Cloud for analysis in BigQuery and Cloud Storage. The preferred replication method is mirroring to avoid deployment of Kafka Connect plugins. What should you do?

- A. Deploy a Kafka cluster on GCE VM Instance
- B. Configure your on-prem cluster to mirror your topics tothe cluster running in GC
- C. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.
- D. Deploy a Kafka cluster on GCE VM Instances with the PubSub Kafka connector configured as a Sink connecto
- E. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.
- F. Deploy the PubSub Kafka connector to your on-prem Kafka cluster and configure PubSub as a Source connecto
- G. Use a Dataflow job to read fron PubSub and write to GCS.

- H. Deploy the PubSub Kafka connector to your on-prem Kafka cluster and configure PubSub as a Sink connector
- I. Use a Dataflow job to read from PubSub and write to GCS.

**Answer:** A

#### NEW QUESTION 328

- (Exam Topic 6)

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
- E. Assign the authorized views data viewer access to the dataset in which the data reside
- F. Assign the users/groups data viewer access to the datasets in which the authorized views reside

**Answer:** C

#### NEW QUESTION 331

- (Exam Topic 6)

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
- B. 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.
- C. Use `gcloud kms keys create` to create a symmetric key
- D. Then use `gcloud kms encrypt` to encrypt each archival file with the key
- E. Use `gsutil cp` to upload each encrypted file to the Cloud Storage bucket
- F. Manually destroy the key previously used for encryption, and rotate the key once and rotate the key once.
- G. Specify customer-supplied encryption key (CSEK) in the `.boto` configuration file
- H. Use `gsutil cp` to upload each archival file to the Cloud Storage bucket
- I. Save the CSEK in Cloud Memorystore as permanent storage of the secret.
- J. Specify customer-supplied encryption key (CSEK) in the `.boto` configuration file
- K. Use `gsutil cp` to upload each archival file to the Cloud Storage bucket
- L. Save the CSEK in a different project that only the security team can access.

**Answer:** B

#### NEW QUESTION 335

.....

## Thank You for Trying Our Product

\* 100% Pass or Money Back

All our products come with a 90-day Money Back Guarantee.

\* One year free update

You can enjoy free update one year. 24x7 online support.

\* Trusted by Millions

We currently serve more than 30,000,000 customers.

\* Shop Securely

All transactions are protected by VeriSign!

**100% Pass Your Professional-Data-Engineer Exam with Our Prep Materials Via below:**

<https://www.certleader.com/Professional-Data-Engineer-dumps.html>