

Exam Questions Professional-Data-Engineer

Google Professional Data Engineer Exam

<https://www.2passeasy.com/dumps/Professional-Data-Engineer/>



NEW QUESTION 1

- (Exam Topic 1)

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

- ☐ No interaction by the user on the site for 1 hour
- ☐ Has added more than \$30 worth of products to the basket
- ☐ Has not completed a transaction

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

- A. Use a fixed-time window with a duration of 60 minutes.
- B. Use a sliding time window with a duration of 60 minutes.
- C. Use a session window with a gap time duration of 60 minutes.
- D. Use a global window with a time based trigger with a delay of 60 minutes.

Answer: D

NEW QUESTION 2

- (Exam Topic 1)

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

- A. Create a Google Cloud Dataflow job to process the data.
- B. Create a Google Cloud Dataproc cluster that uses persistent disks for HDFS.
- C. Create a Hadoop cluster on Google Compute Engine that uses persistent disks.
- D. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.
- E. Create a Hadoop cluster on Google Compute Engine that uses Local SSD disks.

Answer: A

NEW QUESTION 3

- (Exam Topic 1)

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
- B. Reload the data.
- C. 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
- D. Reference the column TS instead of the column DT from now on.
- E. Create a view CLICK_STREAM_V, where strings from the column DT are cast into TIMESTAMP values
- F. Reference the view CLICK_STREAM_V instead of the table CLICK_STREAM from now on.
- G. Add two columns to the table CLICK_STREAM: TS of the TIMESTAMP type and IS_NEW of the BOOLEAN type
- H. Reload all data in append mode
- I. For each appended row, set the value of IS_NEW to true
- J. 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.
- K. 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
- L. Run the query into a destination table NEW_CLICK_STREAM, in which the column TS is the TIMESTAMP type
- M. Reference the table NEW_CLICK_STREAM instead of the table CLICK_STREAM from now on
- N. In the future, new data is loaded into the table NEW_CLICK_STREAM.

Answer: D

NEW QUESTION 4

- (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 5

- (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 6

- (Exam Topic 1)

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

- A. Make a call to the Stackdriver API to list all logs, and apply an advanced filter.
- B. In the Stackdriver logging admin interface, and enable a log sink export to BigQuery.
- C. In the Stackdriver logging admin interface, enable a log sink export to Google Cloud Pub/Sub, and subscribe to the topic from your monitoring tool.
- D. Using the Stackdriver API, create a project sink with advanced log filter to export to Pub/Sub, and subscribe to the topic from your monitoring tool.

Answer: B

NEW QUESTION 7

- (Exam Topic 1)

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

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

Answer: A

NEW QUESTION 8

- (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 9

- (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 10

- (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 10

- (Exam Topic 1)

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

- A. Add a node to the MySQL cluster and build an OLAP cube there.
- B. Use an ETL tool to load the data from MySQL into Google BigQuery.
- C. Connect an on-premises Apache Hadoop cluster to MySQL and perform ETL.
- D. Mount the backups to Google Cloud SQL, and then process the data using Google Cloud Dataproc.

Answer: C

NEW QUESTION 12

- (Exam Topic 1)

Your company built a TensorFlow neural-network model with a large number of neurons and layers. The model fits well for the training data. However, when tested against new data, it performs poorly. What method can you employ to address this?

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

Answer: C

Explanation:

Reference

<https://medium.com/mlreview/a-simple-deep-learning-model-for-stock-price-prediction-using-tensorflow-30505>

NEW QUESTION 14

- (Exam Topic 1)

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.

Answer: D

NEW QUESTION 15

- (Exam Topic 1)

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

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

Answer: B

NEW QUESTION 17

- (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 19

- (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 20

- (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 21

- (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 26

- (Exam Topic 3)

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

Answer: C

NEW QUESTION 27

- (Exam Topic 3)

Given the record streams MJTelco is interested in ingesting per day, they are concerned about the cost of Google BigQuery increasing. MJTelco asks you to provide a design solution. They require a single large data table called tracking_table. Additionally, they want to minimize the cost of daily queries while performing fine-grained analysis of each day's events. They also want to use streaming ingestion. What should you do?

- A. Create a table called tracking_table and include a DATE column.
- B. Create a partitioned table called tracking_table and include a TIMESTAMP column.
- C. Create sharded tables for each day following the pattern tracking_table_YYYYMMDD.
- D. Create a table called tracking_table with a TIMESTAMP column to represent the day.

Answer: B

NEW QUESTION 32

- (Exam Topic 3)

You need to compose visualizations for operations teams with the following requirements: Which approach meets the requirements?

- A. Load the data into Google Sheets, use formulas to calculate a metric, and use filters/sorting to show only suboptimal links in a table.
- B. Load the data into Google BigQuery tables, write Google Apps Script that queries the data, calculates the metric, and shows only suboptimal rows in a table in Google Sheets.
- C. Load the data into Google Cloud Datastore tables, write a Google App Engine Application that queries all rows, applies a function to derive the metric, and then renders results in a table using the Google charts and visualization API.
- D. Load the data into Google BigQuery tables, write a Google Data Studio 360 report that connects to your data, calculates a metric, and then uses a filter expression to show only suboptimal rows in a table.

Answer: C

NEW QUESTION 34

- (Exam Topic 4)

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

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

- A. Redis

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

Answer: BDF

NEW QUESTION 39

- (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 Datastor
- 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 bucke
- F. Use Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Google Cloud Storage.

Answer: A

NEW QUESTION 42

- (Exam Topic 4)

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

- A. The CSV data loaded in BigQuery is not flagged as CSV.
- B. The CSV data has invalid rows that were skipped on import.
- C. The CSV data loaded in BigQuery is not using BigQuery's default encoding.
- D. The CSV data has not gone through an ETL phase before loading into BigQuery.

Answer: B

NEW QUESTION 44

- (Exam Topic 4)

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

- A. Rewrite the job in Pig.
- B. Rewrite the job in Apache Spark.
- C. Increase the size of the Hadoop cluster.
- D. Decrease the size of the Hadoop cluster but also rewrite the job in Hive.

Answer: A

NEW QUESTION 49

- (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 50

- (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 51

- (Exam Topic 5)

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

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

Answer: C

Explanation:

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

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

NEW QUESTION 52

- (Exam Topic 5)

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

- A. [0, 5, 0, 0, 0, 0]
- B. [0, 0, 0, 1, 0, 0, 1]
- C. [0, 1]
- D. [1, 0, 0, 0, 0, 0, 0]

Answer: CD

Explanation:

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

[0, 0, 0, 1, 0, 0, 1] is not a sparse vector because it has two 1s in it. A sparse vector contains only a single 1. [0, 5, 0, 0, 0, 0, 0] is not a sparse vector because it has a 5 in it. Sparse vectors only contain 0s and 1s. Reference: https://www.tensorflow.org/tutorials/linear#feature_columns_and_transformations

NEW QUESTION 54

- (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 55

- (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-vms>

NEW QUESTION 56

- (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.
- D. table names that include their associated project name).
- E. 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 59

- (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 61

- (Exam Topic 5)

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

- A. To export more than 1 GB of data, you need to put a wildcard in the destination filename.
- B. The only supported export destination is Google Cloud Storage.
- C. Data can only be exported in JSON or Avro format.
- D. The only compression option available is GZIP.

Answer: C

Explanation:

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

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

NEW QUESTION 62

- (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 64

- (Exam Topic 5)

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

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

Answer: D

Explanation:

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

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

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

You can think of each value in the embedding vector as a feature of the category. So, if two categories are very similar to each other, then their embedding vectors should be very similar too.

Reference:

<https://cloudacademy.com/google/introduction-to-google-cloud-machine-learning-engine-course/a-wide-and-dee>

NEW QUESTION 67

- (Exam Topic 5)

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

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

Answer: C

Explanation:

Cloud Bigtable is a sparsely populated table that can scale to billions of rows and thousands of columns, allowing you to store terabytes or even petabytes of data.

A single value in each row is indexed; this value is known as the row key.

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

NEW QUESTION 71

- (Exam Topic 5)

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

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

Answer: AB

Explanation:

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

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

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

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

Reference:

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

NEW QUESTION 74

- (Exam Topic 5)

How would you query specific partitions in a BigQuery table?

- A. Use the DAY column in the WHERE clause
- B. Use the EXTRACT(DAY) clause
- C. Use the PARTITIONTIME pseudo-column in the WHERE clause
- D. Use DATE BETWEEN in the WHERE clause

Answer: C

Explanation:

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

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

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

NEW QUESTION 79

- (Exam Topic 5)

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

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

Answer: C

Explanation:

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

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

NEW QUESTION 84

- (Exam Topic 5)

Scaling a Cloud Dataproc cluster typically involves .

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

Answer: A

Explanation:

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

- 1) increase the number of workers to make a job run faster
 - 2) decrease the number of workers to save money
 - 3) increase the number of nodes to expand available Hadoop Distributed Filesystem (HDFS) storage
- Reference: <https://cloud.google.com/dataproc/docs/concepts/scaling-clusters>

NEW QUESTION 88

- (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 93

- (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

NEW QUESTION 98

- (Exam Topic 5)

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

- A. Unsupervised learning
- B. Regressor
- C. Classifier
- D. Clustering estimator

Answer: B

Explanation:

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

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

Clustering is an unsupervised learning task for finding natural groupings of observations (i.e. clusters) based on the inherent structure within your dataset.

Examples include customer segmentation, grouping similar items in e-commerce, and social network analysis.

Reference: <https://elitedatasience.com/machine-learning-algorithms>

NEW QUESTION 100

- (Exam Topic 5)

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

- A. Apache Spark
- B. Apache Hadoop
- C. Apache Kafka
- D. Apache Beam

Answer: D

Explanation:

Dataflow SDKs are being transitioned to Apache Beam, as per the latest Google directive Reference: <https://cloud.google.com/dataflow/docs/>

NEW QUESTION 101

- (Exam Topic 5)

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

- A. Execute to Google Cloud Storage; write to Google Cloud Logging
- B. Write to Google Cloud Storage; read to Google Cloud Logging
- C. Execute to Google Cloud Storage; execute to Google Cloud Logging
- D. Read and write to Google Cloud Storage; write to Google Cloud Logging

Answer: D

Explanation:

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

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

NEW QUESTION 106

- (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

NEW QUESTION 107

- (Exam Topic 5)

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

- A. dataflow.worker
- B. dataflow.compute
- C. dataflow.developer
- D. dataflow.viewer

Answer: A

Explanation:

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

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

NEW QUESTION 111

- (Exam Topic 5)

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

- A. Windows at every 100 MB of data
- B. Single, Global Window
- C. Windows at every 1 minute
- D. Windows at every 10 minutes

Answer: B

Explanation:

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

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

NEW QUESTION 116

- (Exam Topic 5)

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

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

Answer: C

Explanation:

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

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

NEW QUESTION 121

- (Exam Topic 5)

When a Cloud Bigtable node fails, is lost.

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

Answer: B

Explanation:

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

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

Rebalancing tablets from one node to another is very fast, because the actual data is not copied. Cloud Bigtable simply updates the pointers for each node.

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

When a Cloud Bigtable node fails, no data is lost Reference: <https://cloud.google.com/bigtable/docs/overview>

NEW QUESTION 124

- (Exam Topic 5)

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

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

Answer: B

Explanation:

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

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

NEW QUESTION 128

- (Exam Topic 5)

Cloud Dataproc is a managed Apache Hadoop and Apache service.

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

Answer: B

Explanation:

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

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

NEW QUESTION 132

- (Exam Topic 5)

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

- A. It is optimized to handle a high volume of data instances in a job and to run more complex models.
- B. Predictions are returned in the response message.
- C. Predictions are written to output files in a Cloud Storage location that you specify.
- D. It is optimized to minimize the latency of serving predictions.

Answer: BD

Explanation:

Online prediction

Optimized to minimize the latency of serving predictions. Predictions returned in the response message.

Batch prediction

Optimized to handle a high volume of instances in a job and to run more complex models. Predictions written to output files in a Cloud Storage location that you specify.

Reference:

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

NEW QUESTION 136

- (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 139

- (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

NEW QUESTION 142

- (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

NEW QUESTION 146

- (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 151

- (Exam Topic 6)

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
- E. 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
- F. Create a temporary table in Cloud Spanner that will act as a buffer for new data
- G. Create a new step in your pipeline to write to this table first, and then create a new pipeline to write from Cloud Spanner to BigQuery

Answer: BE

NEW QUESTION 153

- (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 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 156

- (Exam Topic 6)

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

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

Answer: A

NEW QUESTION 160

- (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 164

- (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 169

- (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 174

- (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 179

- (Exam Topic 6)

Your company maintains a hybrid deployment with GCP, where analytics are performed on your anonymized customer data. The data are imported to Cloud Storage from your data center through parallel uploads to a data transfer server running on GCP. Management informs you that the daily transfers take too long and have asked you to fix the problem. You want to maximize transfer speeds. Which action should you take?

- A. Increase the CPU size on your server.
- B. Increase the size of the Google Persistent Disk on your server.
- C. Increase your network bandwidth from your datacenter to GCP.
- D. Increase your network bandwidth from Compute Engine to Cloud Storage.

Answer: C

NEW QUESTION 183

- (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 185

- (Exam Topic 6)

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 Datapro
- B. Call the model from your application.
- C. Build and train a classification model with Spark MLlib to generate label
- D. Build and train a second classification model with Spark MLlib to filter results to match customer preference
- E. Deploy the Models using Cloud Datapro
- F. Call the models from your application.
- G. Build an application that calls the Cloud Video Intelligence API to generate label
- H. Store data in Cloud Bigtable, and filter the predicted labels to match the user's viewing history to generate preferences.
- I. Build an application that calls the Cloud Video Intelligence API to generate label
- J. Store data in Cloud SQL, and join and filter the predicted labels to match the user's viewing history to generate preferences.

Answer: C

NEW QUESTION 189

- (Exam Topic 6)

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 applicatio
- B. Process the generated Entity Analysis as labels.
- C. Call the Cloud Natural Language API from your applicatio

- D. Process the generated Sentiment Analysis as labels.
- E. Build and train a text classification model using TensorFlow
- F. Deploy the model using Cloud Machine Learning Engine
- G. Call the model from your application and process the results as labels.
- H. Build and train a text classification model using TensorFlow
- I. Deploy the model using a KubernetesEngine cluster
- J. Call the model from your application and process the results as labels.

Answer: B

NEW QUESTION 192

- (Exam Topic 6)

You need to choose a database to store time series CPU and memory usage for millions of computers. You need to store this data in one-second interval samples. Analysts will be performing real-time, ad hoc analytics against the database. You want to avoid being charged for every query executed and ensure that the schema design will allow for future growth of the dataset. Which database and data model should you choose?

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

Answer: D

NEW QUESTION 195

- (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?

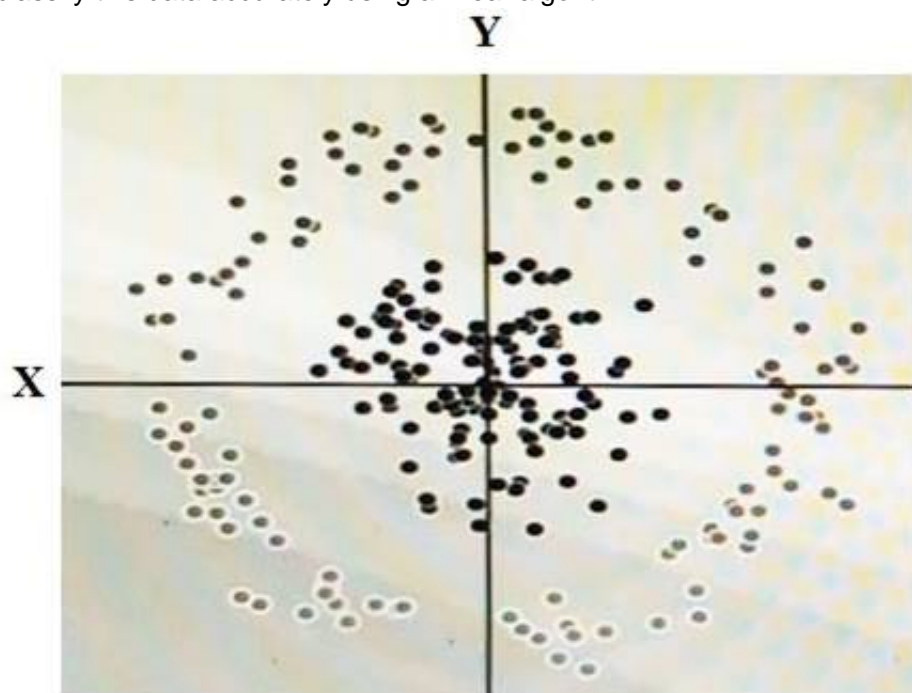
- A. Cloud SQL
- B. Cloud Bigtable
- C. Cloud Spanner
- D. Cloud Datastore

Answer: C

NEW QUESTION 196

- (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 200

- (Exam Topic 6)

Your company has a hybrid cloud initiative. You have a complex data pipeline that moves data between cloud provider services and leverages services from each of the cloud providers. Which cloud-native service should you use to orchestrate the entire pipeline?

- A. Cloud Dataflow
- B. Cloud Composer
- C. Cloud Dataprep
- D. Cloud Dataproc

Answer: D

NEW QUESTION 204

- (Exam Topic 6)

You have developed three data processing jobs. One executes a Cloud Dataflow pipeline that transforms data uploaded to Cloud Storage and writes results to BigQuery. The second ingests data from on-premises servers and uploads it to Cloud Storage. The third is a Cloud Dataflow pipeline that gets information from third-party data providers and uploads the information to Cloud Storage. You need to be able to schedule and monitor the execution of these three workflows and manually execute them when needed. What should you do?

- A. Create a Direct Acyclic Graph in Cloud Composer to schedule and monitor the jobs.
- B. Use Stackdriver Monitoring and set up an alert with a Webhook notification to trigger the jobs.
- C. Develop an App Engine application to schedule and request the status of the jobs using GCP API calls.
- D. Set up cron jobs in a Compute Engine instance to schedule and monitor the pipelines using GCP API calls.

Answer: D

NEW QUESTION 209

- (Exam Topic 6)

After migrating ETL jobs to run on BigQuery, you need to verify that the output of the migrated jobs is the same as the output of the original. You've loaded a table containing the output of the original job and want to compare the contents with output from the migrated job to show that they are identical. The tables do not contain a primary key column that would enable you to join them together for comparison. What should you do?

- A. Select random samples from the tables using the RAND() function and compare the samples.
- B. Select random samples from the tables using the HASH() function and compare the samples.
- C. Use a Dataproc cluster and the BigQuery Hadoop connector to read the data from each table and calculate a hash from non-timestamp columns of the table after sortin
- D. Compare the hashes of each table.
- E. Create stratified random samples using the OVER() function and compare equivalent samples from each table.

Answer: B

NEW QUESTION 211

- (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 accoun
- D. Use the service account's private key to access the dataset
- E. Create a dummy user and grant dataset access to that use
- 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 215

- (Exam Topic 6)

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

Answer: A

NEW QUESTION 216

- (Exam Topic 6)

You are managing a Cloud Dataproc cluster. You need to make a job run faster while minimizing costs, without losing work in progress on your clusters. What should you do?

- A. Increase the cluster size with more non-preemptible workers.
- B. Increase the cluster size with preemptible worker nodes, and configure them to forcefully decommission.
- C. Increase the cluster size with preemptible worker nodes, and use Cloud Stackdriver to trigger a script to preserve work.

D. Increase the cluster size with preemptible worker nodes, and configure them to use graceful decommissioning.

Answer: D

Explanation:

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

NEW QUESTION 220

- (Exam Topic 6)

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
- D. Create a new Cloud Dataflow job with the updated code
- E. Stop the Cloud Dataflow pipeline with the Drain option
- F. Create a new Cloud Dataflow job with the updated code

Answer: A

NEW QUESTION 221

- (Exam Topic 6)

You have a data pipeline that writes data to Cloud Bigtable using well-designed row keys. You want to monitor your pipeline to determine when to increase the size of your Cloud Bigtable cluster. Which two actions can you take to accomplish this? Choose 2 answers.

- A. Review Key Visualizer metric
- B. Increase the size of the Cloud Bigtable cluster when the Read pressure index is above 100.
- C. Review Key Visualizer metric
- D. Increase the size of the Cloud Bigtable cluster when the Write pressure index is above 100.
- E. Monitor the latency of write operation
- F. Increase the size of the Cloud Bigtable cluster when there is a sustained increase in write latency.
- G. Monitor storage utilization
- H. Increase the size of the Cloud Bigtable cluster when utilization increases above 70% of max capacity.
- I. Monitor latency of read operation
- J. Increase the size of the Cloud Bigtable cluster if read operations take longer than 100 ms.

Answer: AC

NEW QUESTION 226

- (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 231

- (Exam Topic 6)

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
- F. Use BigQuery's support for external data sources to query.

Answer: DE

NEW QUESTION 233

- (Exam Topic 6)

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

How should you securely run this workload?

- A. Restrict the Google Cloud Storage bucket so only you can see the files
- B. Grant the Project Owner role to a service account, and run the job with it
- C. Use a service account with the ability to read the batch files and to write to BigQuery
- D. Use a user account with the Project Viewer role on the Cloud Dataproc cluster to read the batch files and write to BigQuery

Answer: B

NEW QUESTION 238

- (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 241

- (Exam Topic 6)

Government regulations in your industry mandate that you have to maintain an auditable record of access to certain types of data. Assuming that all expiring logs will be archived correctly, where should you store data that is subject to that mandate?

- A. Encrypted on Cloud Storage with user-supplied encryption key
- B. 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 use
- E. The Cloud SQL Admin activity logs will be used to provide the auditability.
- F. In a bucket on Cloud Storage that is accessible only by an AppEngine service that collects user information and logs the access before providing a link to the bucket.

Answer: B

NEW QUESTION 244

- (Exam Topic 6)

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

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

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

Answer: C

NEW QUESTION 247

- (Exam Topic 6)

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
- B. Assign the department leads the role of OWNER, and assign the data analysts the role of WRITER on their dataset.
- C. Create a dataset for each department
- D. Assign the department leads the role of WRITER, and assign the data analysts the role of READER on their dataset.
- E. Create a table for each department
- F. Assign the department leads the role of Owner, and assign the data analysts the role of Editor on the project the table is in.
- G. Create a table for each department
- H. Assign the department leads the role of Editor, and assign the data analysts the role of Viewer on the project the table is in.

Answer: D

NEW QUESTION 248

- (Exam Topic 6)

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.

Answer: B

NEW QUESTION 250

- (Exam Topic 6)

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

- A. Convert all daily log tables into date-partitioned tables
- B. Convert the sharded tables into a single partitioned table
- C. Enable query caching so you can cache data from previous months
- D. Create separate views to cover each month, and query from these views

Answer: A

NEW QUESTION 254

- (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 255

- (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 257

- (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 258

- (Exam Topic 6)

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

Answer: ADF

NEW QUESTION 260

- (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 of features of the platform. What should you do?

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

Answer: B

NEW QUESTION 262

- (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 the all streaming insert. What is the most likely cause of this problem?

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

Answer: C

NEW QUESTION 263

- (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 268

- (Exam Topic 6)

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

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

Answer: D

Explanation:

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

NEW QUESTION 270

- (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 274

- (Exam Topic 6)

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

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

Answer: B

NEW QUESTION 279

- (Exam Topic 6)

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 tabl
- E. Create a BigQuery job that merges the new records with the existing records and writes the results to a new BigQuery table.

Answer: A

NEW QUESTION 284

- (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 theusers/groups data viewer access to the authorized views
- D. Create authorized views for each team in datasets created for each tea
- 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 285

- (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 ke
- 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 ke
- D. Then use gcloud kms encrypt to encrypt each archival file with the ke
- E. Use gsutil cp to upload each encrypted file to the Cloud Storage bucke
- 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 fil
- H. Use gsutil cp to upload each archival file to the Cloud Storage bucke
- I. Save the CSEK in Cloud Memorystore as permanent storage of the secret.
- J. Specify customer-supplied encryption key (CSEK) in the .boto configuration fil
- K. Use gsutil cp to upload each archival file to the Cloud Storage bucke
- L. Save the CSEK in a different project that only the security team can access.

Answer: B

NEW QUESTION 288

- (Exam Topic 6)

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

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

Answer: A

NEW QUESTION 291

.....

THANKS FOR TRYING THE DEMO OF OUR PRODUCT

Visit Our Site to Purchase the Full Set of Actual Professional-Data-Engineer Exam Questions With Answers.

We Also Provide Practice Exam Software That Simulates Real Exam Environment And Has Many Self-Assessment Features. Order the Professional-Data-Engineer Product From:

<https://www.2passeasy.com/dumps/Professional-Data-Engineer/>

Money Back Guarantee

Professional-Data-Engineer Practice Exam Features:

- * Professional-Data-Engineer Questions and Answers Updated Frequently
- * Professional-Data-Engineer Practice Questions Verified by Expert Senior Certified Staff
- * Professional-Data-Engineer Most Realistic Questions that Guarantee you a Pass on Your FirstTry
- * Professional-Data-Engineer Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year