

# 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:** C

#### NEW QUESTION 2

- (Exam Topic 1)

You are building a model to make clothing recommendations. You know a user's fashion preference is likely to change over time, so you build a data pipeline to stream new data back to the model as it becomes available. How should you use this data to train the model?

- A. Continuously retrain the model on just the new data.
- B. Continuously retrain the model on a combination of existing data and the new data.
- C. Train on the existing data while using the new data as your test set.
- D. Train on the new data while using the existing data as your test set.

**Answer:** C

#### Explanation:

<https://cloud.google.com/automl-tables/docs/prepare>

#### NEW QUESTION 3

- (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 4

- (Exam Topic 1)

Your company handles data processing for a number of different clients. Each client prefers to use their own suite of analytics tools, with some allowing direct query access via Google BigQuery. You need to secure the data so that clients cannot see each other's data. You want to ensure appropriate access to the data. Which three steps should you take? (Choose three.)

- A. Load data into different partitions.
- B. Load data into a different dataset for each client.
- C. Put each client's BigQuery dataset into a different table.
- D. Restrict a client's dataset to approved users.
- E. Only allow a service account to access the datasets.
- F. Use the appropriate identity and access management (IAM) roles for each client's users.

**Answer:** BDF

#### NEW QUESTION 5

- (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 6

- (Exam Topic 1)

Your software uses a simple JSON format for all messages. These messages are published to Google Cloud Pub/Sub, then processed with Google Cloud Dataflow to create a real-time dashboard for the CFO. During testing, you notice that some messages are missing in the dashboard. You check the logs, and all

messages are being published to Cloud Pub/Sub successfully. What should you do next?

- A. Check the dashboard application to see if it is not displaying correctly.
- B. Run a fixed dataset through the Cloud Dataflow pipeline and analyze the output.
- C. Use Google Stackdriver Monitoring on Cloud Pub/Sub to find the missing messages.
- D. Switch Cloud Dataflow to pull messages from Cloud Pub/Sub instead of Cloud Pub/Sub pushing messages to Cloud Dataflow.

**Answer: B**

#### NEW QUESTION 7

- (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 8

- (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 9

- (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: D**

#### NEW QUESTION 10

- (Exam Topic 1)

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

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

**Answer: D**

#### NEW QUESTION 10

- (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:** BCD

#### NEW QUESTION 13

- (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 14

- (Exam Topic 1)

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

- A. Disable caching by editing the report settings.
- B. Disable caching in BigQuery by editing table details.
- C. Refresh your browser tab showing the visualizations.
- D. Clear your browser history for the past hour then reload the tab showing the visualizations.

**Answer:** A

#### Explanation:

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

#### NEW QUESTION 15

- (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 18

- (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 19

- (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 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)

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 31

- (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 35

- (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 40

- (Exam Topic 4)

Your company produces 20,000 files every hour. Each data file is formatted as a comma separated values (CSV) file that is less than 4 KB. All files must be ingested on Google Cloud Platform before they can be processed. Your company site has a 200 ms latency to Google Cloud, and your Internet connection bandwidth is limited as 50 Mbps. You currently deploy a secure FTP (SFTP) server on a virtual machine in Google Compute Engine as the data ingestion point. A local SFTP client runs on a dedicated machine to transmit the CSV files as is. The goal is to make reports with data from the previous day available to the executives by 10:00 a.m. each day. This design is barely able to keep up with the current volume, even though the bandwidth utilization is rather low.

You are told that due to seasonality, your company expects the number of files to double for the next three months. Which two actions should you take? (choose two.)

- A. Introduce data compression for each file to increase the rate file of file transfer.
- B. Contact your internet service provider (ISP) to increase your maximum bandwidth to at least 100 Mbps.
- C. Redesign the data ingestion process to use gsutil tool to send the CSV files to a storage bucket in parallel.
- D. Assemble 1,000 files into a tape archive (TAR) fil
- E. Transmit the TAR files instead, and disassemble the CSV files in the cloud upon receiving them.
- F. Create an S3-compatible storage endpoint in your network, and use Google Cloud Storage Transfer Service to transfer on-premises data to the designated storage bucket.

**Answer: CE**

#### NEW QUESTION 42

- (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 45

- (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 tabl
- 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 fil
- 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 FullNam
- G. Run a BigQuery load job to load the new CSV file into BigQuery.

**Answer: C**

#### NEW QUESTION 50

- (Exam Topic 4)

You are designing the database schema for a machine learning-based food ordering service that will predict what users want to eat. Here is some of the information you need to store:

- > The user profile: What the user likes and doesn't like to eat
- > The user account information: Name, address, preferred meal times
- > The order information: When orders are made, from where, to whom

The database will be used to store all the transactional data of the product. You want to optimize the data schema. Which Google Cloud Platform product should you use?

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

D. Cloud Datastore

**Answer:** A

#### NEW QUESTION 51

- (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 54

- (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 59

- (Exam Topic 5)

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

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

**Answer:** C

#### Explanation:

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

You can easily use the `--properties` option of the `dataproc` command in the Google Cloud SDK to modify many common configuration files when creating a cluster. When creating a Cloud Dataproc cluster, you can specify initialization actions in executables and/or scripts that Cloud Dataproc will run on all nodes in your Cloud Dataproc cluster immediately after the cluster is set up. [<https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/init-actions>]

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

#### NEW QUESTION 64

- (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 67

- (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 71

- (Exam Topic 5)

If a dataset contains rows with individual people and columns for year of birth, country, and income, how many of the columns are continuous and how many are categorical?

- A. 1 continuous and 2 categorical
- B. 3 categorical
- C. 3 continuous
- D. 2 continuous and 1 categorical

**Answer:** D

#### Explanation:

The columns can be grouped into two types—categorical and continuous columns:

A column is called categorical if its value can only be one of the categories in a finite set. For example, the native country of a person (U.S., India, Japan, etc.) or the education level (high school, college, etc.) are categorical columns.

A column is called continuous if its value can be any numerical value in a continuous range. For example, the capital gain of a person (e.g. \$14,084) is a continuous column.

Year of birth and income are continuous columns. Country is a categorical column.

You could use bucketization to turn year of birth and/or income into categorical features, but the raw columns are continuous.

Reference: [https://www.tensorflow.org/tutorials/wide#reading\\_the\\_census\\_data](https://www.tensorflow.org/tutorials/wide#reading_the_census_data)

#### NEW QUESTION 75

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

#### NEW QUESTION 78

- (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 79

- (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 81**

- (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)

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 88**

- (Exam Topic 5)

To give a user read permission for only the first three columns of a table, which access control method would you use?

- A. Primitive role
- B. Predefined role
- C. Authorized view
- D. It's not possible to give access to only the first three columns of a table.

**Answer:** C

**Explanation:**

An authorized view allows you to share query results with particular users and groups without giving them read access to the underlying tables. Authorized views can only be created in a dataset that does not contain the tables queried by the view.

When you create an authorized view, you use the view's SQL query to restrict access to only the rows and columns you want the users to see.

Reference: <https://cloud.google.com/bigquery/docs/views#authorized-views>

**NEW QUESTION 91**

- (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](https://cloud.google.com/dataproc/docs/concepts/service-accounts#important_notes)

#### NEW QUESTION 96

- (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 99

- (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 101

- (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 105

- (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 106

- (Exam Topic 5)

Why do you need to split a machine learning dataset into training data and test data?

- A. So you can try two different sets of features
- B. To make sure your model is generalized for more than just the training data
- C. To allow you to create unit tests in your code
- D. So you can use one dataset for a wide model and one for a deep model

**Answer: B**

**Explanation:**

The flaw with evaluating a predictive model on training data is that it does not inform you on how well the model has generalized to new unseen data. A model that is selected for its accuracy on the training dataset rather than its accuracy on an unseen test dataset is very likely to have lower accuracy on an unseen test dataset. The reason is that the model is not as generalized. It has specialized to the structure in the training dataset. This is called overfitting.

Reference: <https://machinelearningmastery.com/a-simple-intuition-for-overfitting/>

**NEW QUESTION 108**

- (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 113**

- (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 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 119**

- (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 123

- (Exam Topic 5)

You want to use a BigQuery table as a data sink. In which writing mode(s) can you use BigQuery as a sink?

- A. Both batch and streaming
- B. BigQuery cannot be used as a sink
- C. Only batch
- D. Only streaming

**Answer:** A

#### Explanation:

When you apply a BigQueryIO.Write transform in batch mode to write to a single table, Dataflow invokes a BigQuery load job. When you apply a BigQueryIO.Write transform in streaming mode or in batch mode using a function to specify the destination table, Dataflow uses BigQuery's streaming inserts

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

### NEW QUESTION 125

- (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 130

- (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 134

- (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://elitedatascience.com/machine-learning-algorithms>

### NEW QUESTION 135

- (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] 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](https://www.tensorflow.org/tutorials/linear#feature_columns_and_transformations)

**NEW QUESTION 140**

- (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 145**

- (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 149**

- (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 151**

- (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 155**

- (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 160**

- (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 162**

- (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 164**

- (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 168**

- (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 170**

- (Exam Topic 5)

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

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

**Answer:** B

**Explanation:**

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

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

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

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

**NEW QUESTION 173**

- (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 175**

- (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 179**

- (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 183**

- (Exam Topic 6)

You decided to use Cloud Datastore to ingest vehicle telemetry data in real time. You want to build a storage system that will account for the long-term data growth, while keeping the costs low. You also want to create snapshots of the data periodically, so that you can make a point-in-time (PIT) recovery, or clone a copy of the data for Cloud Datastore in a different environment. You want to archive these snapshots for a long time. Which two methods can accomplish this? Choose 2 answers.

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

**Answer:** CE

**NEW QUESTION 188**

- (Exam Topic 6)

You use a dataset in BigQuery for analysis. You want to provide third-party companies with access to the same dataset. You need to keep the costs of data sharing low and ensure that the data is current. Which solution should you choose?

- A. Create an authorized view on the BigQuery table to control data access, and provide third-party companies with access to that view.
- B. Use Cloud Scheduler to export the data on a regular basis to Cloud Storage, and provide third-party companies with access to the bucket.
- C. Create a separate dataset in BigQuery that contains the relevant data to share, and provide third-party companies with access to the new dataset.
- D. Create a Cloud Dataflow job that reads the data in frequent time intervals, and writes it to the relevant BigQuery dataset or Cloud Storage bucket for third-party companies to use.

**Answer:** B

**NEW QUESTION 189**

- (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

**Explanation:**

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

**NEW QUESTION 193**

- (Exam Topic 6)

You have Cloud Functions written in Node.js that pull messages from Cloud Pub/Sub and send the data to BigQuery. You observe that the message processing rate on the Pub/Sub topic is orders of magnitude higher than anticipated, but there is no error logged in Stackdriver Log Viewer. What are the two most likely causes of this problem? Choose 2 answers.

- A. Publisher throughput quota is too small.
- B. Total outstanding messages exceed the 10-MB maximum.
- C. Error handling in the subscriber code is not handling run-time errors properly.
- D. The subscriber code cannot keep up with the messages.
- E. The subscriber code does not acknowledge the messages that it pulls.

**Answer:** CD

**NEW QUESTION 194**

- (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 198

- (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 201

- (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 204

- (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 205

- (Exam Topic 6)

You are migrating a table to BigQuery and are deciding on the data model. Your table stores information related to purchases made across several store locations and includes information like the time of the transaction, items purchased, the store ID and the city and state in which the store is located. You frequently query this table to see how many of each item were sold over the past 30 days and to look at purchasing trends by state, city, and individual store. You want to model this table to minimize query time and cost. What should you do?

- A. Partition by transaction time; cluster by state first, then city then store ID
- B. Partition by transaction time; cluster by store ID first, then city, then state
- C. Top-level cluster by state first, then city then store
- D. Top-level cluster by store ID first, then city then state.

**Answer:** C

#### NEW QUESTION 210

- (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 211

- (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 214

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

**Answer:** D

#### NEW QUESTION 216

- (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 220

- (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 222

- (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 227**

- (Exam Topic 6)

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

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

**Answer:** A

**NEW QUESTION 232**

- (Exam Topic 6)

You are updating the code for a subscriber to a Pub/Sub feed. You are concerned that upon deployment the subscriber may erroneously acknowledge messages, leading to message loss. Your subscriber is not set up to retain acknowledged messages. What should you do to ensure that you can recover from errors after deployment?

- A. Use Cloud Build for your deployment. If an error occurs after deployment, use a Seek operation to locate a timestamp logged by Cloud Build at the start of the deployment.
- B. Create a Pub/Sub snapshot before deploying new subscriber code.
- C. Use a Seek operation to re-deliver messages that became available after the snapshot was created.
- D. Set up the Pub/Sub emulator on your local machine. Validate the behavior of your new subscriber code before deploying it to production.
- E. Enable dead-lettering on the Pub/Sub topic to capture messages that aren't successfully acknowledged. If an error occurs after deployment, re-deliver any messages captured by the dead-letter queue.

**Answer:** B

**NEW QUESTION 233**

- (Exam Topic 6)

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

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

**Answer:** C

**Explanation:**

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

**NEW QUESTION 235**

- (Exam Topic 6)

Your company currently runs a large on-premises cluster using Spark, Hive, and Hadoop Distributed File System (HDFS) in a colocation facility. The cluster is designed to support peak usage on the system, however, many jobs are batch in nature, and usage of the cluster fluctuates quite dramatically. Your company is eager to move to the cloud to reduce the overhead associated with on-premises infrastructure and maintenance and to benefit from the cost savings. They are also hoping to modernize their existing infrastructure to use more server offerings in order to take advantage of the cloud. Because of the timing of their contract renewal with the colocation facility, they have only 2 months for their initial migration. How should you recommend they approach their upcoming migration strategy so they can maximize their cost savings in the cloud while still executing the migration in time?

- A. Migrate the workloads to Dataproc plus HOPS, modernize later.
- B. Migrate the workloads to Dataproc plus Cloud Storage, modernize later.
- C. Migrate the Spark workload to Dataproc plus HDFS, and modernize the Hive workload for BigQuery.
- D. Modernize the Spark workload for Dataflow and the Hive workload for BigQuery.

**Answer:** D

**NEW QUESTION 237**

- (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 242**

- (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 243

- (Exam Topic 6)

You work for a shipping company that uses handheld scanners to read shipping labels. Your company has strict data privacy standards that require scanners to only transmit recipients' personally identifiable information (PII) to analytics systems, which violates user privacy rules. You want to quickly build a scalable solution using cloud-native managed services to prevent exposure of PII to the analytics systems. What should you do?

- A. Create an authorized view in BigQuery to restrict access to tables with sensitive data.
- B. Install a third-party data validation tool on Compute Engine virtual machines to check the incoming data for sensitive information.
- C. Use Stackdriver logging to analyze the data passed through the total pipeline to identify transactions that may contain sensitive information.
- D. Build a Cloud Function that reads the topics and makes a call to the Cloud Data Loss Prevention API
- E. Use the tagging and confidence levels to either pass or quarantine the data in a bucket for review.

**Answer:** D

#### NEW QUESTION 248

- (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 252

- (Exam Topic 6)

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

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

**Answer:** D

#### NEW QUESTION 255

- (Exam Topic 6)

An aerospace company uses a proprietary data format to store its night data. You need to connect this new data source to BigQuery and stream the data into BigQuery. You want to efficiently import the data into BigQuery where consuming as few resources as possible. What should you do?

- A. Use a standard Dataflow pipeline to store the raw data in BigQuery and then transform the format later when the data is used.
- B. Write a shell script that triggers a Cloud Function that performs periodic ETL batch jobs on the new data source
- C. Use Apache Hive to write a Dataproc job that streams the data into BigQuery in CSV format
- D. Use an Apache Beam custom connector to write a Dataflow pipeline that streams the data into BigQuery in Avro format

**Answer:** D

#### NEW QUESTION 259

- (Exam Topic 6)

You want to automate execution of a multi-step data pipeline running on Google Cloud. The pipeline includes Cloud Dataproc and Cloud Dataflow jobs that have multiple dependencies on each other. You want to use managed services where possible, and the pipeline will run every day. Which tool should you use?

- A. cron
- B. Cloud Composer
- C. Cloud Scheduler
- D. Workflow Templates on Cloud Dataproc

**Answer:** B

**NEW QUESTION 264**

- (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.

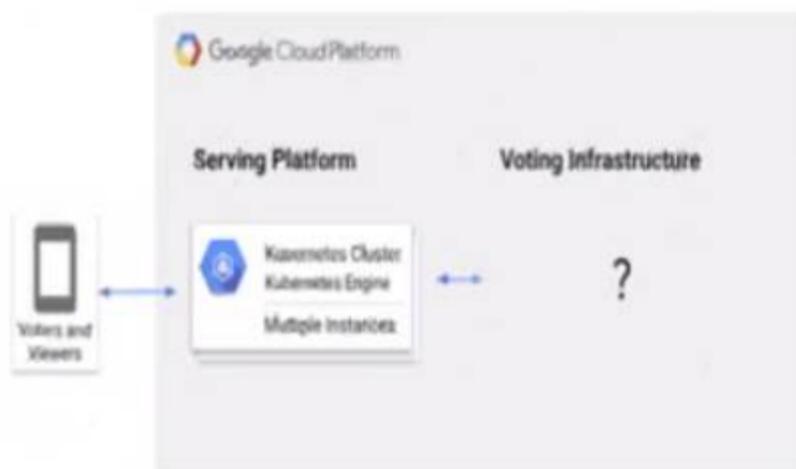
- A. Review Key Visualizer metric
- B. Increase the size of the Cloud Bigtable cluster when the Read pressure index is above 100.
- C. Review Key Visualizer metric
- D. Increase the size of the Cloud Bigtable cluster when the Write pressure index is above 100.
- E. Monitor the latency of write operation
- 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 266**

- (Exam Topic 6)

A live TV show asks viewers to cast votes using their mobile phones. The event generates a large volume of data during a 3 minute period. You are in charge of the Voting restructure\* and must ensure that the platform can handle the load and that all votes are processed. You must display partial results while voting is open. After voting ends you need to count the votes exactly once while optimizing cost. What should you do?



- A. Create a Memorystore instance with a high availability (HA) configuration
- B. Write votes to a Pub Sub topic and have Cloud Functions subscribe to it and write votes to BigQuery
- C. Write votes to a Pub/Sub topic and load into both Bigtable and BigQuery via a Dataflow pipeline. Query Bigtable for real-time results and BigQuery for later analysis. Shutdown the Bigtable instance when voting concludes.
- D. Create a Cloud SQL for PostgreSQL database with high availability (HA) configuration and multiple read replicas

**Answer: C**

**NEW QUESTION 270**

- (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
- B. Set a sliding time window of 1 hour every 5 minutes
- 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
- 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/Sub
- G. Use a Cloud Dataflow template to write your messages from Cloud Pub/Sub to Cloud Bigtable
- H. Use Cloud Scheduler to run a script every hour that counts the number of rows created in Cloud Bigtable in the last hour
- I. If that number falls below 4000, send an alert.
- J. Use Kafka Connect to link your Kafka message queue to Cloud Pub/Sub
- K. Use a Cloud Dataflow template to write your messages from Cloud Pub/Sub to BigQuery
- L. Use Cloud Scheduler to run a script every five minutes that counts the number of rows created in BigQuery in the last hour
- M. If that number falls below 4000, send an alert.

**Answer: C**

**NEW QUESTION 272**

- (Exam Topic 6)

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

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

Answer: D

#### NEW QUESTION 274

- (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: C

#### Explanation:

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

#### NEW QUESTION 276

- (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: A

#### NEW QUESTION 277

- (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 278

- (Exam Topic 6)

You need to give new website users a globally unique identifier (GUID) using a service that takes in data points and returns a GUID. This data is sourced from both internal and external systems via HTTP calls that you will make via microservices within your pipeline. There will be tens of thousands of messages per second and that can be multithreaded, and you worry about the backpressure on the system. How should you design your pipeline to minimize that backpressure?

- A. Call out to the service via HTTP
- B. Create the pipeline statically in the class definition
- C. Create a new object in the startBundle method of DoFn
- D. Batch the job into ten-second increments

Answer: A

#### NEW QUESTION 282

- (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 selects 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: C

#### NEW QUESTION 283

- (Exam Topic 6)

You are migrating an application that tracks library books and information about each book, such as author or year published, from an on-premises data warehouse to BigQuery. In your current relational database, the author information is kept in a separate table and joined to the book information on a common key. Based on Google's recommended practice for schema design, how would you structure the data to ensure optimal speed of queries about the author of each book that has been borrowed?

- A. Keep the schema the same, maintain the different tables for the book and each of the attributes, and query as you are doing today
- B. Create a table that is wide and includes a column for each attribute, including the author's first name, last name, date of birth, etc
- C. Create a table that includes information about the books and authors, but nest the author fields inside the author column
- D. Keep the schema the same, create a view that joins all of the tables, and always query the view

**Answer:** C

#### NEW QUESTION 287

- (Exam Topic 6)

As your organization expands its usage of GCP, many teams have started to create their own projects. Projects are further multiplied to accommodate different stages of deployments and target audiences. Each project requires unique access control configurations. The central IT team needs to have access to all projects. Furthermore, data from Cloud Storage buckets and BigQuery datasets must be shared for use in other projects in an ad hoc way. You want to simplify access control management by minimizing the number of policies. Which two steps should you take? Choose 2 answers.

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

**Answer:** AC

#### NEW QUESTION 289

- (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 291

- (Exam Topic 6)

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

- Executing the transformations on a schedule
- Enabling non-developer analysts to modify transformations
- Providing a graphical tool for designing transformations

What should you do?

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

**Answer:** A

#### Explanation:

you can use dataprep for continuously changing target schema

In general, a target consists of the set of information required to define the expected data in a dataset. Often referred to as a "schema," this target schema information can include:

Names of columns

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

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

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

#### NEW QUESTION 296

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

**Answer:** C

#### NEW QUESTION 297

- (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: C**

#### NEW QUESTION 299

- (Exam Topic 6)

Your globally distributed auction application allows users to bid on items. Occasionally, users place identical bids at nearly identical times, and different application servers process those bids. Each bid event contains the item, amount, user, and timestamp. You want to collate those bid events into a single location in real time to determine which user bid first. What should you do?

- A. Create a file on a shared file and have the application servers write all bid events to that file
- B. Process the file with Apache Hadoop to identify which user bid first.
- C. Have each application server write the bid events to Cloud Pub/Sub as they occur
- D. Push the events from Cloud Pub/Sub to a custom endpoint that writes the bid event information into Cloud SQL.
- E. Set up a MySQL database for each application server to write bid events into
- F. Periodically query each of those distributed MySQL databases and update a master MySQL database with bid event information.
- G. Have each application server write the bid events to Google Cloud Pub/Sub as they occur
- H. Use a pull subscription to pull the bid events using Google Cloud Dataflow
- I. Give the bid for each item to the user in the bid event that is processed first.

**Answer: C**

#### NEW QUESTION 302

- (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: BC**

#### NEW QUESTION 305

- (Exam Topic 6)

You work for a bank. You have a labelled dataset that contains information on already granted loan applications 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 307

- (Exam Topic 6)

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

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

**Answer: B**

#### NEW QUESTION 311

- (Exam Topic 6)

You are using BigQuery and Data Studio to design a customer-facing dashboard that displays large quantities of aggregated data. You expect a high volume of concurrent users. You need to optimize the dashboard to provide quick visualizations with minimal latency. What should you do?

- A. Use BigQuery BI Engine with materialized views
- B. Use BigQuery BI Engine with streaming data.
- C. Use BigQuery BI Engine with authorized views
- D. Use BigQuery BI Engine with logical reviews

Answer: B

#### NEW QUESTION 315

- (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 318

- (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 320

- (Exam Topic 6)

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

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

Answer: A

#### NEW QUESTION 324

- (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 329

- (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:** D

#### NEW QUESTION 333

- (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 338

- (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 custom ops.
- C. Use Cloud GPUs after implementing GPU kernel support for your custom ops.
- D. Stay on CPUs, and increase the size of the cluster you're training your model on.

**Answer:** B

#### NEW QUESTION 341

- (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:** D

#### NEW QUESTION 346

- (Exam Topic 6)

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

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

**Answer:** C

#### NEW QUESTION 350

- (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 351

- (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 356

- (Exam Topic 6)

Your analytics team wants to build a simple statistical model to determine which customers are most likely to work with your company again, based on a few different metrics. They want to run the model on Apache Spark, using data housed in Google Cloud Storage, and you have recommended using Google Cloud Dataproc to execute this job. Testing has shown that this workload can run in approximately 30 minutes on a 15-node cluster, outputting the results into Google BigQuery. The plan is to run this workload weekly. How should you optimize the cluster for cost?

- A. Migrate the workload to Google Cloud Dataflow
- B. Use pre-emptible virtual machines (VMs) for the cluster
- C. Use a higher-memory node so that the job runs faster
- D. Use SSDs on the worker nodes so that the job can run faster

**Answer: A**

#### NEW QUESTION 361

- (Exam Topic 6)

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

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

**Answer: A**

#### NEW QUESTION 365

- (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 367

- (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 369

- (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:** C

**Explanation:**

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

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

**NEW QUESTION 374**

- (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 regiona
- B. In the event of an emergency, use a point-in-time snapshot to recover the data.
- C. Set the BigQuery dataset to be regiona
- D. Create a scheduled query to make copies of the data to tables suffixed with the time of the backu
- E. In the event of an emergency, use the backup copy of the table.
- F. Set the BigQuery dataset to be multi-regiona
- 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 375**

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

**Answer:** A

**NEW QUESTION 377**

- (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 379**

- (Exam Topic 6)

You've migrated a Hadoop job from an on-premises cluster to Dataproc and Good Storage. Your Spark job is a complex analytical workload fiat consists of many shuffling operations, and initial data are parquet toes (on average 200-400 MB size each) You see some degradation in performance after the migration to Dataproc so you'd like to optimize for it. Your organization is very cost-sensitive so you'd Idee to continue using Dataproc on preemptibles (with 2 non-preemptibles workers only) for this workload. What should you do?

- A. Switch from HODs to SSDs override the preemptible VMs configuration to increase the boot disk size
- B. Increase the see of your parquet files to ensure them to be 1 GB minimum
- C. Switch to TFRecords format (appr 200 MB per We) instead of parquet files
- D. Switch from HDDs to SSD
- E. copy initial data from Cloud Storage to Hadoop Distributed File System (HDFS) run the Spark job and copy results back to Cloud Storage

**Answer:** A

**NEW QUESTION 384**

- (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 386

- (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 390

.....

## 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