

# Google

## Exam Questions Professional-Data-Engineer

Google Professional Data Engineer Exam



### NEW QUESTION 1

- (Exam Topic 1)

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

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

**Answer: B**

### NEW QUESTION 2

- (Exam Topic 1)

You need to store and analyze social media postings in Google BigQuery at a rate of 10,000 messages per minute in near real-time. Initially, design the application to use streaming inserts for individual postings. Your application also performs data aggregations right after the streaming inserts. You discover that the queries after streaming inserts do not exhibit strong consistency, and reports from the queries might miss in-flight data. How can you adjust your application design?

- A. Re-write the application to load accumulated data every 2 minutes.
- B. Convert the streaming insert code to batch load for individual messages.
- C. Load the original message to Google Cloud SQL, and export the table every hour to BigQuery via streaming inserts.
- D. Estimate the average latency for data availability after streaming inserts, and always run queries after waiting twice as long.

**Answer: D**

#### Explanation:

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

### NEW QUESTION 3

- (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 4

- (Exam Topic 1)

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

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

**Answer: D**

### NEW QUESTION 5

- (Exam Topic 1)

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

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

**Answer: C**

### NEW QUESTION 6

- (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 typ

- 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 value
- 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 value
- 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 7

- (Exam Topic 1)

Your company is running their first dynamic campaign, serving different offers by analyzing real-time data during the holiday season. The data scientists are collecting terabytes of data that rapidly grows every hour during their 30-day campaign. They are using Google Cloud Dataflow to preprocess the data and collect the feature (signals) data that is needed for the machine learning model in Google Cloud Bigtable. The team is observing suboptimal performance with reads and writes of their initial load of 10 TB of data. They want to improve this performance while minimizing cost. What should they do?

- A. Redefine the schema by evenly distributing reads and writes across the row space of the table.
- B. The performance issue should be resolved over time as the size of the Bigtable cluster is increased.
- C. Redesign the schema to use a single row key to identify values that need to be updated frequently in the cluster.
- D. Redesign the schema to use row keys based on numeric IDs that increase sequentially per user viewing the offers.

**Answer:** A

#### NEW QUESTION 8

- (Exam Topic 1)

Business owners at your company have given you a database of bank transactions. Each row contains the user ID, transaction type, transaction location, and transaction amount. They ask you to investigate what type of machine learning can be applied to the data. Which three machine learning applications can you use? (Choose three.)

- A. Supervised learning to determine which transactions are most likely to be fraudulent.
- B. Unsupervised learning to determine which transactions are most likely to be fraudulent.
- C. Clustering to divide the transactions into N categories based on feature similarity.
- D. Supervised learning to predict the location of a transaction.
- E. Reinforcement learning to predict the location of a transaction.
- F. Unsupervised learning to predict the location of a transaction.

**Answer:** BCD

#### NEW QUESTION 9

- (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 10

- (Exam Topic 1)

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

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

**Answer:** B

#### NEW QUESTION 10

- (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 11**

- (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 12**

- (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 15**

- (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 20

- (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 25

- (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 26

- (Exam Topic 5)

Cloud Bigtable is Google's Big Data database service.

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

**Answer:** C

**Explanation:**

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

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

#### NEW QUESTION 28

- (Exam Topic 5)

The for Cloud Bigtable makes it possible to use Cloud Bigtable in a Cloud Dataflow pipeline.

- A. Cloud Dataflow connector
- B. DataFlow SDK
- C. BigQuery API
- D. BigQuery Data Transfer Service

**Answer:** A

**Explanation:**

The Cloud Dataflow connector for Cloud Bigtable makes it possible to use Cloud Bigtable in a Cloud Dataflow pipeline. You can use the connector for both batch and streaming operations.

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

**NEW QUESTION 30**

- (Exam Topic 5)

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

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

**Answer:** C

**Explanation:**

The YARN ResourceManager and the HDFS NameNode interfaces are available on a Cloud Dataproc cluster master node. The cluster master-host-name is the name of your Cloud Dataproc cluster followed by an -m suffix—for example, if your cluster is named "my-cluster", the master-host-name would be "my-cluster-m".

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

**NEW QUESTION 33**

- (Exam Topic 5)

How would you query specific partitions in a BigQuery table?

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

**Answer:** C

**Explanation:**

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

```
WHERE __PARTITIONTIME BETWEEN TIMESTAMP('2017-01-01') AND TIMESTAMP('2017-01-02')
```

Reference: [https://cloud.google.com/bigquery/docs/partitioned-tables#the\\_partitiontime\\_pseudo\\_column](https://cloud.google.com/bigquery/docs/partitioned-tables#the_partitiontime_pseudo_column)

**NEW QUESTION 37**

- (Exam Topic 5)

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

```
SELECT person FROM `project1.example.table1` WHERE city = "London" How would you correct the error?
```

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

**Answer:** A

**Explanation:**

To access the person.city column, you need to "UNNEST(person)" and JOIN it to table1 using a comma. Reference:

[https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql#nested\\_repeated\\_resu](https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql#nested_repeated_resu)

**NEW QUESTION 39**

- (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 44

- (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 47

- (Exam Topic 5)

Which software libraries are supported by Cloud Machine Learning Engine?

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

**Answer: C**

#### Explanation:

Cloud ML Engine mainly does two things:

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

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

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

#### NEW QUESTION 50

- (Exam Topic 5)

Which of these statements about BigQuery caching is true?

- A. By default, a query's results are not cached.
- B. BigQuery caches query results for 48 hours.
- C. Query results are cached even if you specify a destination table.
- D. There is no charge for a query that retrieves its results from cache.

**Answer: D**

#### Explanation:

When query results are retrieved from a cached results table, you are not charged for the query. BigQuery caches query results for 24 hours, not 48 hours.

Query results are not cached if you specify a destination table.

A query's results are always cached except under certain conditions, such as if you specify a destination table. Reference:

<https://cloud.google.com/bigquery/querying-data#query-caching>

#### NEW QUESTION 54

- (Exam Topic 5)

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

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

**Answer: CD**

#### Explanation:

Selecting and crafting the right set of feature columns is key to learning an effective model.

Bucketization is a process of dividing the entire range of a continuous feature into a set of consecutive bins/buckets, and then converting the original numerical feature into a bucket ID (as a categorical feature) depending on which bucket that value falls into.

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

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

#### NEW QUESTION 57

- (Exam Topic 5)

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

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

**Answer:** A

**Explanation:**

Service accounts used with Cloud Dataproc must have Dataproc/Dataproc Worker role (or have all the permissions granted by Dataproc Worker role).  
Reference: [https://cloud.google.com/dataproc/docs/concepts/service-accounts#important\\_notes](https://cloud.google.com/dataproc/docs/concepts/service-accounts#important_notes)

**NEW QUESTION 62**

- (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 65**

- (Exam Topic 5)

Which of the following is not true about Dataflow pipelines?

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

**Answer:** D

**Explanation:**

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

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

**NEW QUESTION 67**

- (Exam Topic 5)

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

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

**Answer:** C

**Explanation:**

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

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

**NEW QUESTION 70**

- (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 71

- (Exam Topic 5)

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

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

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

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

#### NEW QUESTION 73

- (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 77

- (Exam Topic 5)

Which of the following job types are supported by Cloud Dataproc (select 3 answers)?

- A. Hive
- B. Pig
- C. YARN
- D. Spark

**Answer: ABD**

#### Explanation:

Cloud Dataproc provides out-of-the box and end-to-end support for many of the most popular job types, including Spark, Spark SQL, PySpark, MapReduce, Hive, and Pig jobs.

Reference: [https://cloud.google.com/dataproc/docs/resources/faq#what\\_type\\_of\\_jobs\\_can\\_i\\_run](https://cloud.google.com/dataproc/docs/resources/faq#what_type_of_jobs_can_i_run)

#### NEW QUESTION 82

- (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 86

- (Exam Topic 6)

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

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

- A. Create an application that provides an AP
- B. Write a tool to poll the API and write data to Cloud Storage as gzipped JSON files.

- C. Create an application that writes to a Cloud SQL database to store the data.
- D. Set up periodic exports of the database to write to Cloud Storage and load into BigQuery.
- E. Create an application that publishes events to Cloud Pub/Sub, and create Spark jobs on Cloud Dataproc to convert the JSON data to Avro format, stored on HDFS on Persistent Disk.
- F. Create an application that publishes events to Cloud Pub/Sub, and create a Cloud Dataflow pipeline that transforms the JSON event payloads to Avro, writing the data to Cloud Storage and BigQuery.

**Answer:** A

#### NEW QUESTION 87

- (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 89

- (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 90

- (Exam Topic 6)

You store historic data in Cloud Storage. You need to perform analytics on the historic data. You want to use a solution to detect invalid data entries and perform data transformations that will not require programming or knowledge of SQL. What should you do?

- A. Use Cloud Dataflow with Beam to detect errors and perform transformations.
- B. Use Cloud Dataprep with recipes to detect errors and perform transformations.
- C. Use Cloud Dataproc with a Hadoop job to detect errors and perform transformations.
- D. Use federated tables in BigQuery with queries to detect errors and perform transformations.

**Answer:** B

#### NEW QUESTION 91

- (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 increasing the size of vocabularies or n-grams used.

**Answer:** D

#### NEW QUESTION 92

- (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 94**

- (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?

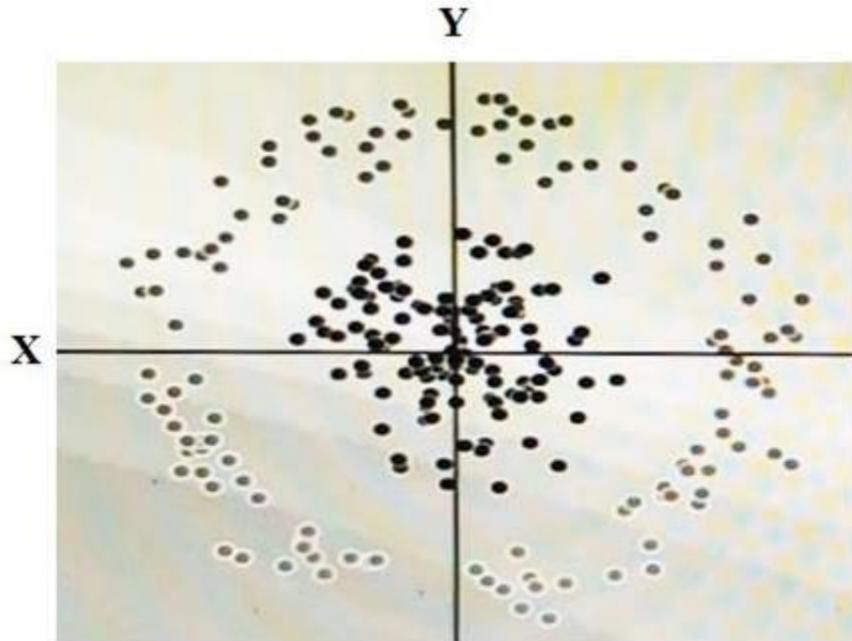
- A. Cloud SQL
- B. Cloud Bigtable
- C. Cloud Spanner
- D. Cloud Datastore

**Answer: C**

**NEW QUESTION 95**

- (Exam Topic 6)

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



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

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

**Answer: D**

**NEW QUESTION 99**

- (Exam Topic 6)

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

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

**Answer: C**

**NEW QUESTION 101**

- (Exam Topic 6)

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

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

- L. Use Cloud Scheduler to run a script every five minutes that counts the number of rows created in BigQuery in the last hour.
- M. If that number falls below 4000, send an alert.

**Answer: C**

#### NEW QUESTION 104

- (Exam Topic 6)

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

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

**Answer: C**

#### NEW QUESTION 105

- (Exam Topic 6)

Your company is migrating its on-premises data warehousing solution to BigQuery. The existing data warehouse uses trigger-based change data capture (CDC) to apply daily updates from transactional database sources. Your company wants to use BigQuery to improve its handling of CDC and to optimize the performance of the data warehouse. Source system changes must be available for query in near-real time using trigger-based CDC streams. You need to ensure that changes in the BigQuery reporting table are available with minimal latency and reduced overhead. What should you do? Choose 2 answers.

- A. Perform a DML INSERT, UPDATE, or DELETE to replicate each CDC record in the reporting table in real time.
- B. Periodically DELETE outdated records from the reporting table. Periodically use a DML MERGE to simultaneously perform DML INSERT, UPDATE, and DELETE operations in the reporting table.
- C. Update, and DELETE operations in the reporting table.
- D. Insert each new CDC record and corresponding operation type into a staging table in real time.
- E. Insert each new CDC record and corresponding operation type into the reporting table in real time and use a materialized view to expose only the current version of each unique record.

**Answer: BD**

#### NEW QUESTION 110

- (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 112

- (Exam Topic 6)

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

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

**Answer: B**

#### NEW QUESTION 115

- (Exam Topic 6)

A TensorFlow machine learning model on Compute Engine virtual machines (n2-standard-32) takes two days to complete training. The model has custom TensorFlow operations that must run partially on a CPU. You want to reduce the training time in a cost-effective manner. What should you do?

- A. Change the VM type to n2-highmem-32.
- B. Change the VM type to e2-standard-32.
- C. Train the model using a VM with a GPU hardware accelerator.
- D. Train the model using a VM with a TPU hardware accelerator.

**Answer: C**

#### NEW QUESTION 118

- (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 122

- (Exam Topic 6)

You are designing an Apache Beam pipeline to enrich data from Cloud Pub/Sub with static reference data from BigQuery. The reference data is small enough to fit in memory on a single worker. The pipeline should write enriched results to BigQuery for analysis. Which job type and transforms should this pipeline use?

- A. Batch job, PubSubIO, side-inputs
- B. Streaming job, PubSubIO, JdbcIO, side-outputs
- C. Streaming job, PubSubIO, BigQueryIO, side-inputs
- D. Streaming job, PubSubIO, BigQueryIO, side-outputs

**Answer: C**

#### NEW QUESTION 124

- (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 125

- (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 129

- (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 134

- (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 136

- (Exam Topic 6)

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

What should you do?

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

**Answer:** B

#### NEW QUESTION 137

- (Exam Topic 6)

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

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

**Answer:** A

#### Explanation:

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

#### NEW QUESTION 141

- (Exam Topic 6)

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

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

**Answer:** C

#### NEW QUESTION 142

- (Exam Topic 6)

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

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

**Answer:** D

#### Explanation:

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

#### NEW QUESTION 143

- (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 148

- (Exam Topic 6)

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

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

**Answer:** BC

#### NEW QUESTION 151

- (Exam Topic 6)

You want to optimize your queries for cost and performance. How should you structure your data?

- A. Partition table data by create\_date, location\_id and device\_version
- B. Partition table data by create\_date cluster table data by location\_id and device\_version
- C. Cluster table data by create\_date location\_id and device\_version
- D. Cluster table data by create\_date partition by location and device\_version

**Answer:** B

#### NEW QUESTION 155

- (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 158

- (Exam Topic 6)

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

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

**Answer:** A

#### NEW QUESTION 160

- (Exam Topic 6)

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

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

Answer: D

#### NEW QUESTION 162

- (Exam Topic 6)

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

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

Answer: C

#### NEW QUESTION 164

- (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 166

- (Exam Topic 6)

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

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

Answer: A

#### NEW QUESTION 170

- (Exam Topic 6)

Your organization has been collecting and analyzing data in Google BigQuery for 6 months. The majority of the data analyzed is placed in a time-partitioned table named events\_partitioned. To reduce the cost of queries, your organization created a view called events, which queries only the last 14 days of data. The view is described in legacy SQL. Next month, existing applications will be connecting to BigQuery to read the events data via an ODBC connection. You need to ensure the applications can connect. Which two actions should you take? (Choose two.)

- A. Create a new view over events using standard SQL
- B. Create a new partitioned table using a standard SQL query
- C. Create a new view over events\_partitioned using standard SQL
- D. Create a service account for the ODBC connection to use for authentication
- E. Create a Google Cloud Identity and Access Management (Cloud IAM) role for the ODBC connection and shared "events"

Answer: AE

#### NEW QUESTION 175

- (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 179

- (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 180

- (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 connector
- 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 connector
- 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 connector
- I. Use a Dataflow job to read from PubSub and write to GCS.

**Answer: A**

#### NEW QUESTION 185

- (Exam Topic 6)

You want to analyze hundreds of thousands of social media posts daily at the lowest cost and with the fewest steps. You have the following requirements:

- > You will batch-load the posts once per day and run them through the Cloud Natural Language API.
- > You will extract topics and sentiment from the posts.
- > You must store the raw posts for archiving and reprocessing.
- > You will create dashboards to be shared with people both inside and outside your organization.

You need to store both the data extracted from the API to perform analysis as well as the raw social media posts for historical archiving. What should you do?

- A. Store the social media posts and the data extracted from the API in BigQuery.
- B. Store the social media posts and the data extracted from the API in Cloud SQL.
- C. Store the raw social media posts in Cloud Storage, and write the data extracted from the API into BigQuery.
- D. Feed to social media posts into the API directly from the source, and write the extracted data from the API into BigQuery.

**Answer: D**

#### NEW QUESTION 188

- (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 that consists of many shuffling operations, and initial data are parquet files (on average 200-400 MB size each) You see some degradation in performance after the migration to Dataproc so you'd like to optimize for it. Your organization is very cost-sensitive so you'd like to continue using Dataproc on preemptibles (with 2 non-preemptible workers only) for this workload. What should you do?

- A. Switch from HDDs to SSDs override the preemptible VMs configuration to increase the boot disk size
- B. Increase the size of your parquet files to ensure them to be 1 GB minimum
- C. Switch to TFRecords format (approx 200 MB per file) 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 190

- (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 195

- (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 197

- (Exam Topic 6)

You are migrating your data warehouse to Google Cloud and decommissioning your on-premises data center. Because this is a priority for your company, you know that bandwidth will be made available for the initial data load to the cloud. The files being transferred are not large in number, but each file is 90 GB. Additionally, you want your transactional systems to continually update the warehouse on Google Cloud in real time. What tools should you use to migrate the data and ensure that it continues to write to your warehouse?

- A. Storage Transfer Service for the migration, Pub/Sub and Cloud Data Fusion for the real-time updates
- B. BigQuery Data Transfer Service for the migration, Pub/Sub and Dataproc for the real-time updates
- C. gsutil for the migration; Pub/Sub and Dataflow for the real-time updates
- D. gsutil for both the migration and the real-time updates

**Answer: A**

#### NEW QUESTION 202

- (Exam Topic 6)

You are building a data pipeline on Google Cloud. You need to prepare data using a casual method for a machine-learning process. You want to support a logistic regression model. You also need to monitor and adjust for null values, which must remain real-valued and cannot be removed. What should you do?

- A. Use Cloud Dataprep to find null values in sample source data
- B. Convert all nulls to 'none' using a Cloud Dataproc job.
- C. Use Cloud Dataprep to find null values in sample source data
- D. Convert all nulls to 0 using a Cloud Dataprep job.
- E. Use Cloud Dataflow to find null values in sample source data
- F. Convert all nulls to 'none' using a Cloud Dataprep job.
- G. Use Cloud Dataflow to find null values in sample source data
- H. Convert all nulls to using a custom script.

**Answer: C**

#### NEW QUESTION 205

.....

## **Thank You for Trying Our Product**

### **We offer two products:**

1st - We have Practice Tests Software with Actual Exam Questions

2nd - Questions and Answers in PDF Format

### **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

**100% Actual & Verified — Instant Download, Please Click**  
**[Order The Professional-Data-Engineer Practice Test Here](#)**