

# Google

## Exam Questions Professional-Data-Engineer

Google Professional Data Engineer Exam



**NEW QUESTION 1**

- (Exam Topic 1)

Your startup has never implemented a formal security policy. Currently, everyone in the company has access to the datasets stored in Google BigQuery. Teams have freedom to use the service as they see fit, and they have not documented their use cases. You have been asked to secure the data warehouse. You need to discover what everyone is doing. What should you do first?

- A. Use Google Stackdriver Audit Logs to review data access.
- B. Get the identity and access management (IAM) policy of each table.
- C. Use Stackdriver Monitoring to see the usage of BigQuery query slots.
- D. Use the Google Cloud Billing API to see what account the warehouse is being billed to.

**Answer: C**

**NEW QUESTION 2**

- (Exam Topic 1)

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

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

**Answer: A**

**NEW QUESTION 3**

- (Exam Topic 1)

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)

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 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 6**

- (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 7

- (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 8

- (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:** D

#### NEW QUESTION 9

- (Exam Topic 1)

You designed a database for patient records as a pilot project to cover a few hundred patients in three clinics. Your design used a single database table to represent all patients and their visits, and you used self-joins to generate reports. The server resource utilization was at 50%. Since then, the scope of the project has expanded. The database must now store 100 times more patient records. You can no longer run the reports, because they either take too long or they encounter errors with insufficient compute resources. How should you adjust the database design?

- A. Add capacity (memory and disk space) to the database server by the order of 200.
- B. Shard the tables into smaller ones based on date ranges, and only generate reports with prespecified date ranges.
- C. Normalize the master patient-record table into the patient table and the visits table, and create other necessary tables to avoid self-join.
- D. Partition the table into smaller tables, with one for each clinic.
- E. Run queries against the smaller table pairs, and use unions for consolidated reports.

**Answer:** B

#### NEW QUESTION 10

- (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 10

- (Exam Topic 1)

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

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

**Answer:** A

#### NEW QUESTION 12

- (Exam Topic 2)

Flowlogic is rolling out their real-time inventory tracking system. The tracking devices will all send package-tracking messages, which will now go to a single Google Cloud Pub/Sub topic instead of the Apache Kafka cluster. A subscriber application will then process the messages for real-time reporting and store them in Google BigQuery for historical analysis. You want to ensure the package data can be analyzed over time. Which approach should you take?

- A. Attach the timestamp on each message in the Cloud Pub/Sub subscriber application as they are received.
- B. Attach the timestamp and Package ID on the outbound message from each publisher device as they are sent to Cloud Pub/Sub.
- C. Use the NOW () function in BigQuery to record the event's time.
- D. Use the automatically generated timestamp from Cloud Pub/Sub to order the data.

**Answer:** B

#### NEW QUESTION 15

- (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 17

- (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 20

- (Exam Topic 3)

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

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

**Answer:** B

#### NEW QUESTION 25

- (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 29

- (Exam Topic 3)

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

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

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

**Answer:** BD

#### NEW QUESTION 33

- (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 35

- (Exam Topic 4)

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

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

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

**Answer: BDF**

#### NEW QUESTION 38

- (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 41

- (Exam Topic 4)

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

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

**Answer: A**

#### NEW QUESTION 44

- (Exam Topic 4)

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

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

**Answer: C**



#### NEW QUESTION 49

- (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 53

- (Exam Topic 5)

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

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

**Answer: C**

#### Explanation:

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

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

#### NEW QUESTION 55

- (Exam Topic 5)

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

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

**Answer: A**

#### Explanation:

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

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

#### NEW QUESTION 59

- (Exam Topic 5)

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 62

- (Exam Topic 5)

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

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

**Answer: A**

#### Explanation:

In a Cloud Bigtable architecture all client requests go through a front-end server before they are sent to a Cloud Bigtable node. The nodes are organized into a Cloud Bigtable cluster, which belongs to a Cloud Bigtable instance, which is a container for the cluster. Each node in the cluster handles a subset of the requests to the cluster. When additional nodes are added to a cluster, you can increase the number of simultaneous requests that the cluster can handle, as well as the maximum throughput for the entire cluster.  
Reference: <https://cloud.google.com/bigtable/docs/overview>

**NEW QUESTION 66**

- (Exam Topic 5)

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

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

**Answer:** B

**Explanation:**

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

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

**NEW QUESTION 71**

- (Exam Topic 5)

Which of these operations can you perform from the BigQuery Web UI?

- A. Upload a file in SQL format.
- B. Load data with nested and repeated fields.
- C. Upload a 20 MB file.
- D. Upload multiple files using a wildcard.

**Answer:** B

**Explanation:**

You can load data with nested and repeated fields using the Web UI. You cannot use the Web UI to:

- Upload a file greater than 10 MB in size
- Upload multiple files at the same time
- Upload a file in SQL format

All three of the above operations can be performed using the "bq" command. Reference: <https://cloud.google.com/bigquery/loading-data>

**NEW QUESTION 76**

- (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 77**

- (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 79**

- (Exam Topic 5)

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

- A. Standard SQL is the preferred query language for BigQuery.

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

**Answer:** D

**Explanation:**

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

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

In legacy SQL, to query a table with a project-qualified name, you use a colon, :, as a separator. In standard SQL, you use a period, ., instead.

Due to the differences in syntax between the two query languages (such as with project-qualified table names), if you write a query in Legacy SQL, it might generate an error if you try to run it with Standard SQL.

Reference:

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

**NEW QUESTION 83**

- (Exam Topic 5)

What are two methods that can be used to denormalize tables in BigQuery?

- A. 1) Split table into multiple tables; 2) Use a partitioned table  
B. 1) Join tables into one table; 2) Use nested repeated fields  
C. 1) Use a partitioned table; 2) Join tables into one table  
D. 1) Use nested repeated fields; 2) Use a partitioned table

**Answer:** B

**Explanation:**

The conventional method of denormalizing data involves simply writing a fact, along with all its dimensions, into a flat table structure. For example, if you are dealing with sales transactions, you would write each individual fact to a record, along with the accompanying dimensions such as order and customer information. The other method for denormalizing data takes advantage of BigQuery's native support for nested and repeated structures in JSON or Avro input data. Expressing records using nested and repeated structures can provide a more natural representation of the underlying data. In the case of the sales order, the outer part of a JSON structure would contain the order and customer information, and the inner part of the structure would contain the individual line items of the order, which would be represented as nested, repeated elements.

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

**NEW QUESTION 88**

- (Exam Topic 5)

The CUSTOM tier for Cloud Machine Learning Engine allows you to specify the number of which types of cluster nodes?

- A. Workers  
B. Masters, workers, and parameter servers  
C. Workers and parameter servers  
D. Parameter servers

**Answer:** C

**Explanation:**

The CUSTOM tier is not a set tier, but rather enables you to use your own cluster specification. When you use this tier, set values to configure your processing cluster according to these guidelines:

You must set TrainingInput.masterType to specify the type of machine to use for your master node. You may set TrainingInput.workerCount to specify the number of workers to use.

You may set TrainingInput.parameterServerCount to specify the number of parameter servers to use.

You can specify the type of machine for the master node, but you can't specify more than one master node. Reference: [https://cloud.google.com/ml-engine/docs/training-overview#job\\_configuration\\_parameters](https://cloud.google.com/ml-engine/docs/training-overview#job_configuration_parameters)

**NEW QUESTION 91**

- (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 96**

- (Exam Topic 5)

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

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



**Answer:** B

**Explanation:**

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

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

Reference:

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

**NEW QUESTION 97**

- (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 99**

- (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 103**

- (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 105**

- (Exam Topic 5)

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

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

**Answer:** C

**Explanation:**

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

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

#### NEW QUESTION 106

- (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 111

- (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 116

- (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 118

- (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 120

- (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 122**

- (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 125**

- (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 127**

- (Exam Topic 5)

Which action can a Cloud Dataproc Viewer perform?

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

**Answer:** D

**Explanation:**

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

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

**NEW QUESTION 129**

- (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 130

- (Exam Topic 5)

Which of the following is NOT true about Dataflow pipelines?

- A. Dataflow pipelines are tied to Dataflow, and cannot be run on any other runner
- B. Dataflow pipelines can consume data from other Google Cloud services
- C. Dataflow pipelines can be programmed in Java
- D. Dataflow pipelines use a unified programming model, so can work both with streaming and batch data sources

**Answer:** A

#### Explanation:

Dataflow pipelines can also run on alternate runtimes like Spark and Flink, as they are built using the Apache Beam SDKs

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

#### NEW QUESTION 135

- (Exam Topic 5)

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

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

**Answer:** BD

#### Explanation:

Online prediction

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

Batch prediction

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

Reference:

[https://cloud.google.com/ml-engine/docs/prediction-overview#online\\_prediction\\_versus\\_batch\\_prediction](https://cloud.google.com/ml-engine/docs/prediction-overview#online_prediction_versus_batch_prediction)

#### NEW QUESTION 138

- (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 139

- (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 141

- (Exam Topic 6)

You are designing storage for two relational tables that are part of a 10-TB database on Google Cloud. You want to support transactions that scale horizontally. You also want to optimize data for range queries on nonkey columns. What should you do?

- A. Use Cloud SQL for storag
- B. Add secondary indexes to support query patterns.
- C. Use Cloud SQL for storag
- D. Use Cloud Dataflow to transform data to support query patterns.
- E. Use Cloud Spanner for storag
- F. Add secondary indexes to support query patterns.
- G. Use Cloud Spanner for storag
- H. Use Cloud Dataflow to transform data to support query patterns.

**Answer:** D

**Explanation:**

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

**NEW QUESTION 142**

- (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 145**

- (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 146**

- (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 151**

- (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 entities
- E. Treat each entity as a BigQuery table row via BigQuery streaming insert
- F. Assign an export timestamp for each export, and attach it as an extra column for each row
- 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 entities
- I. Format the exported data into a JSON file
- J. Apply compression before storing the data in Cloud Source Repositories.

**Answer:** CE

**NEW QUESTION 152**

- (Exam Topic 6)

You are designing a data processing pipeline. The pipeline must be able to scale automatically as load increases. Messages must be processed at least once, and must be ordered within windows of 1 hour. How should you design the solution?

- A. Use Apache Kafka for message ingestion and use Cloud Dataproc for streaming analysis.
- B. Use Apache Kafka for message ingestion and use Cloud Dataflow for streaming analysis.
- C. Use Cloud Pub/Sub for message ingestion and Cloud Dataproc for streaming analysis.
- D. Use Cloud Pub/Sub for message ingestion and Cloud Dataflow for streaming analysis.

**Answer:** C



#### NEW QUESTION 154

- (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 159

- (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 file system, and use those credentials to access the BigQuery dataset

**Answer: C**

#### NEW QUESTION 160

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

**Answer: A**

#### NEW QUESTION 163

- (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 167

- (Exam Topic 6)

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

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

**Answer: D**

#### NEW QUESTION 169

- (Exam Topic 6)

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

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

**Answer:** A

#### NEW QUESTION 172

- (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:** A

#### NEW QUESTION 177

- (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 179

- (Exam Topic 6)

You have historical data covering the last three years in BigQuery and a data pipeline that delivers new data to BigQuery daily. You have noticed that when the Data Science team runs a query filtered on a date column and limited to 30–90 days of data, the query scans the entire table. You also noticed that your bill is increasing more quickly than you expected. You want to resolve the issue as cost-effectively as possible while maintaining the ability to conduct SQL queries. What should you do?

- A. Re-create the tables using DDL
- B. Partition the tables by a column containing a TIMESTAMP or DATE Type.
- C. Recommend that the Data Science team export the table to a CSV file on Cloud Storage and use Cloud Datalab to explore the data by reading the files directly.
- D. Modify your pipeline to maintain the last 30–90 days of data in one table and the longer history in a different table to minimize full table scans over the entire history.
- E. Write an Apache Beam pipeline that creates a BigQuery table per day
- F. Recommend that the Data Science team use wildcards on the table name suffixes to select the data they need.

**Answer:** C

#### NEW QUESTION 183

- (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 187

- (Exam Topic 6)

You are responsible for writing your company's ETL pipelines to run on an Apache Hadoop cluster. The pipeline will require some checkpointing and splitting pipelines. Which method should you use to write the pipelines?

- A. PigLatin using Pig
- B. HiveQL using Hive
- C. Java using MapReduce
- D. Python using MapReduce

**Answer:** D

**NEW QUESTION 188**

- (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 AP
- E. Use the tagging and confidence levels to either pass or quarantine the data in a bucket for review.

**Answer:** A

**NEW QUESTION 191**

- (Exam Topic 6)

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

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

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



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

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

**Answer:** A

**NEW QUESTION 195**

- (Exam Topic 6)

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

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

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

**Answer:** C

**NEW QUESTION 197**

- (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 202**

- (Exam Topic 6)

You've migrated a Hadoop job from an on-prem cluster to dataproc and GCS. Your Spark job is a complicated 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. You need to keep in mind that 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. Increase the size of your parquet files to ensure them to be 1 GB minimum.
- B. Switch to TFRecords formats (app
- C. 200MB per file) instead of parquet files.
- D. Switch from HDDs to SSDs, copy initial data from GCS to HDFS, run the Spark job and copy results back to GCS.
- E. Switch from HDDs to SSDs, override the preemptible VMs configuration to increase the boot disk size.

**Answer:** C

**NEW QUESTION 206**

- (Exam Topic 6)

You have a query that filters a BigQuery table using a WHERE clause on timestamp and ID columns. By using bq query --dry\_run you learn that the query

triggers a full scan of the table, even though the filter on timestamp and ID select a tiny fraction of the overall data. You want to reduce the amount of data scanned by BigQuery with minimal changes to existing SQL queries. What should you do?

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

**Answer: B**

#### NEW QUESTION 210

- (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 215

- (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 217

- (Exam Topic 6)

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

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

**Answer: D**

#### NEW QUESTION 220

- (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 225

- (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 230

- (Exam Topic 6)

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

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

**Answer:** A

#### NEW QUESTION 234

- (Exam Topic 6)

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

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

**Answer:** D

#### Explanation:

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

#### NEW QUESTION 239

- (Exam Topic 6)

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

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

**Answer:** B

#### NEW QUESTION 243

- (Exam Topic 6)

You need to create a near real-time inventory dashboard that reads the main inventory tables in your BigQuery data warehouse. Historical inventory data is stored as inventory balances by item and location. You have several thousand updates to inventory every hour. You want to maximize performance of the dashboard and ensure that the data is accurate. What should you do?

- A. Leverage BigQuery UPDATE statements to update the inventory balances as they are changing.
- B. Partition the inventory balance table by item to reduce the amount of data scanned with each inventory update.
- C. Use the BigQuery streaming table to stream changes into a daily inventory movement table
- D. Calculate balances in a view that joins it to the historical inventory balance table
- E. Update the inventory balance table nightly.
- F. Use the BigQuery bulk loader to batch load inventory changes into a daily inventory movement table. Calculate balances in a view that joins it to the historical inventory balance table
- G. Update the inventory balance table nightly.

**Answer:** A

#### NEW QUESTION 247

- (Exam Topic 6)

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

- ▶ The data being analyzed is in CSV, Avro, and PDF formats and will be accessed by multiple analysis tools including Cloud Dataproc, BigQuery, and Compute Engine.
- ▶ A streaming data pipeline stores new data daily.



- ▶ Performance is not a factor in the solution.
  - ▶ The solution design should maximize availability.
- How should you design data storage for this solution?

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

**Answer: C**

#### NEW QUESTION 251

- (Exam Topic 6)

You're using Bigtable for a real-time application, and you have a heavy load that is a mix of read and writes. You've recently identified an additional use case and need to perform hourly an analytical job to calculate certain statistics across the whole database. You need to ensure both the reliability of your production application as well as the analytical workload. What should you do?

- A. Export Bigtable dump to GCS and run your analytical job on top of the exported files.
- B. Add a second cluster to an existing instance with a multi-cluster routing, use live-traffic app profile for your regular workload and batch-analytics profile for the analytics workload.
- C. Add a second cluster to an existing instance with a single-cluster routing, use live-traffic app profile for your regular workload and batch-analytics profile for the analytics workload.
- D. Increase the size of your existing cluster twice and execute your analytics workload on your new resized cluster.

**Answer: B**

#### NEW QUESTION 256

- (Exam Topic 6)

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

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

**Answer: B**

#### NEW QUESTION 260

- (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 HDFS
- 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 cluster
- 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 cluster
- F. Then run the Hadoop utility to copy them to HDFS
- 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 BigQuery
- 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 261

- (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 customs ops.
- D. Stay on CPUs, and increase the size of the cluster you're training your model on.

**Answer:** B

#### NEW QUESTION 265

- (Exam Topic 6)

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

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

**Answer:** D

#### NEW QUESTION 266

- (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:** A

#### NEW QUESTION 267

- (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:** A

#### NEW QUESTION 268

- (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 271

- (Exam Topic 6)

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

- A. Use BigQuery machine learning to be able to train the model at scale, so you can analyze the packages in batches.
- B. Train an AutoML model on your corpus of images, and build an API around that model to integrate with the package tracking applications.
- C. Use the Cloud Vision API to detect for damage, and raise an alert through Cloud Function
- D. Integrate the package tracking applications with this function.
- E. Use TensorFlow to create a model that is trained on your corpus of image
- F. Create a Python notebook in Cloud Datalab that uses this model so you can analyze for damaged packages.

**Answer:** A

#### NEW QUESTION 275

- (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 276

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

**Answer:** C

#### NEW QUESTION 277

.....

## 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](#)**