

Exam Questions Professional-Data-Engineer

Google Professional Data Engineer Exam

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



NEW QUESTION 1

- (Exam Topic 1)

You want to use a database of information about tissue samples to classify future tissue samples as either normal or mutated. You are evaluating an unsupervised anomaly detection method for classifying the tissue samples. Which two characteristics support this method? (Choose two.)

- A. There are very few occurrences of mutations relative to normal samples.
- B. There are roughly equal occurrences of both normal and mutated samples in the database.
- C. You expect future mutations to have different features from the mutated samples in the database.
- D. You expect future mutations to have similar features to the mutated samples in the database.
- E. You already have labels for which samples are mutated and which are normal in the database.

Answer: AD

Explanation:

Unsupervised anomaly detection techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal by looking for instances that seem to fit least to the remainder of the data set. https://en.wikipedia.org/wiki/Anomaly_detection

NEW QUESTION 2

- (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 3

- (Exam Topic 1)

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

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

Answer: A

NEW QUESTION 4

- (Exam Topic 1)

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 2)

Flowlogistic wants to use Google BigQuery as their primary analysis system, but they still have Apache Hadoop and Spark workloads that they cannot move to BigQuery. Flowlogistic does not know how to store the data that is common to both workloads. What should they do?

- A. Store the common data in BigQuery as partitioned tables.
- B. Store the common data in BigQuery and expose authorized views.
- C. Store the common data encoded as Avro in Google Cloud Storage.
- D. Store the common data in the HDFS storage for a Google Cloud Dataproc cluster.

Answer: B

NEW QUESTION 6

- (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 7

- (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 8

- (Exam Topic 4)

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

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

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

Answer: CE

NEW QUESTION 9

- (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 10

- (Exam Topic 4)

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

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

Answer: A

NEW QUESTION 10

- (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 15

- (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

NEW QUESTION 17

- (Exam Topic 5)

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

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

Answer: BD

Explanation:

Online prediction

Optimized to minimize the latency of serving predictions.

Predictions returned in the response message. Batch prediction

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

Reference:

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

NEW QUESTION 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 24

- (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 27

- (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. BiqQuery 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 29

- (Exam Topic 5)

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

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

Answer: C

Explanation:

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

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

NEW QUESTION 30

- (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 32

- (Exam Topic 5)

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

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

Answer: D

Explanation:

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

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

NEW QUESTION 36

- (Exam Topic 5)

Which software libraries are supported by Cloud Machine Learning Engine?

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

Answer: C

Explanation:

Cloud ML Engine mainly does two things:

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

Hosts those trained models for you in the cloud so that you can use them to get predictions about new data.
Reference: https://cloud.google.com/ml-engine/docs/technical-overview#what_it_does

NEW QUESTION 38

- (Exam Topic 5)

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

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

Answer: A

Explanation:

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

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

NEW QUESTION 43

- (Exam Topic 5)

Which is the preferred method to use to avoid hotspotting in time series data in Bigtable?

- A. Field promotion
- B. Randomization
- C. Salting
- D. Hashing

Answer: A

Explanation:

By default, prefer field promotion. Field promotion avoids hotspotting in almost all cases, and it tends to make it easier to design a row key that facilitates queries.

Reference:

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

NEW QUESTION 46

- (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 49

- (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

NEW QUESTION 54

- (Exam Topic 5)

Which of the following are examples of hyperparameters? (Select 2 answers.)

- A. Number of hidden layers

- B. Number of nodes in each hidden layer
- C. Biases
- D. Weights

Answer: AB

Explanation:

If model parameters are variables that get adjusted by training with existing data, your hyperparameters are the variables about the training process itself. For example, part of setting up a deep neural network is deciding how many "hidden" layers of nodes to use between the input layer and the output layer, as well as how many nodes each layer should use. These variables are not directly related to the training data at all. They are configuration variables. Another difference is that parameters change during a training job, while the hyperparameters are usually constant during a job.

Weights and biases are variables that get adjusted during the training process, so they are not hyperparameters. Reference: <https://cloud.google.com/ml-engine/docs/hyperparameter-tuning-overview>

NEW QUESTION 58

- (Exam Topic 5)

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

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

Answer: CD

Explanation:

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

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

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

Reference: https://www.tensorflow.org/tutorials/wide#selecting_and_engineering_features_for_the_model

NEW QUESTION 62

- (Exam Topic 5)

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

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

Answer: A

Explanation:

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

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

This is especially useful in the case of testing distributed models, as it allows you to validate that you are properly interacting with the Cloud ML Engine cluster configuration.

Reference: <https://cloud.google.com/sdk/gcloud/reference/ml-engine/local/train>

NEW QUESTION 63

- (Exam Topic 5)

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

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

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

Answer: B

Explanation:

Using the Drain option to stop your job tells the Dataflow service to finish your job in its current state. Your job will immediately stop ingesting new data from input sources, but the Dataflow service will preserve any existing resources (such as worker instances) to finish processing and writing any buffered data in your pipeline.

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

NEW QUESTION 66

- (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 71

- (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 76

- (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 80

- (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 83

- (Exam Topic 5)

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

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

Answer: B

Explanation:

For Cloud Bigtable, you can configure access control at the project level. For example, you can grant the ability to:
Read from, but not write to, any table within the project.

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

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

NEW QUESTION 86

- (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 87

- (Exam Topic 5)

Which Google Cloud Platform service is an alternative to Hadoop with Hive?

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

Answer: C

Explanation:

Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data summarization, query, and analysis.

Google BigQuery is an enterprise data warehouse. Reference: https://en.wikipedia.org/wiki/Apache_Hive

NEW QUESTION 89

- (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 93

- (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 97

- (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 101

- (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 105

- (Exam Topic 6)

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

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

Answer: B

NEW QUESTION 106

- (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 dat
- 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 107

- (Exam Topic 6)

You need to set access to BigQuery for different departments within your company. Your solution should comply with the following requirements:

- Each department should have access only to their data.
- Each department will have one or more leads who need to be able to create and update tables and provide them to their team.
- Each department has data analysts who need to be able to query but not modify data. How should you set access to the data in BigQuery?

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

Answer: D

NEW QUESTION 109

- (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 111

- (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 112

- (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 114

- (Exam Topic 6)

You need ads data to serve AI models and historical data for analytics. Longtail and outlier data points need to be identified. You want to cleanse the data in near-real time before running it through AI models. What should you do?

- A. Use BigQuery to ingest, prepare, and then analyze the data and then run queries to create views
- B. Use Cloud Storage as a data warehouse, shell scripts for processing, and BigQuery to create views for desired datasets
- C. Use Dataflow to identify longtail and outlier data points programmatically with BigQuery as a sink
- D. Use Cloud Composer to identify longtail and outlier data points, and then output a usable dataset to BigQuery

Answer: A

NEW QUESTION 119

- (Exam Topic 6)

The marketing team at your organization provides regular updates of a segment of your customer dataset. The marketing team has given you a CSV with 1 million records that must be updated in BigQuery. When you use the UPDATE statement in BigQuery, you receive a quotaExceeded error. What should you do?

- A. Reduce the number of records updated each day to stay within the BigQuery UPDATE DML statement limit.
- B. Increase the BigQuery UPDATE DML statement limit in the Quota management section of the Google Cloud Platform Console.
- C. Split the source CSV file into smaller CSV files in Cloud Storage to reduce the number of BigQuery UPDATE DML statements per BigQuery job.
- D. Import the new records from the CSV file into a new BigQuery table.
- E. Create a BigQuery job that merges the new records with the existing records and writes the results to a new BigQuery table.

Answer: D

NEW QUESTION 122

- (Exam Topic 6)

You use BigQuery as your centralized analytics platform. New data is loaded every day, and an ETL pipeline modifies the original data and prepares it for the final users. This ETL pipeline is regularly modified and can generate errors, but sometimes the errors are detected only after 2 weeks. You need to provide a method to recover from these errors, and your backups should be optimized for storage costs. How should you organize your data in BigQuery and store your backups?

- A. Organize your data in a single table, export, and compress and store the BigQuery data in Cloud Storage.
- B. Organize your data in separate tables for each month, and export, compress, and store the data in Cloud Storage.
- C. Organize your data in separate tables for each month, and duplicate your data on a separate dataset in BigQuery.
- D. Organize your data in separate tables for each month, and use snapshot decorators to restore the table to a time prior to the corruption.

Answer: D

NEW QUESTION 127

- (Exam Topic 6)

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

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

Answer: A

NEW QUESTION 130

- (Exam Topic 6)

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

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

Answer: D

NEW QUESTION 131

- (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 132

- (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 136

- (Exam Topic 6)

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

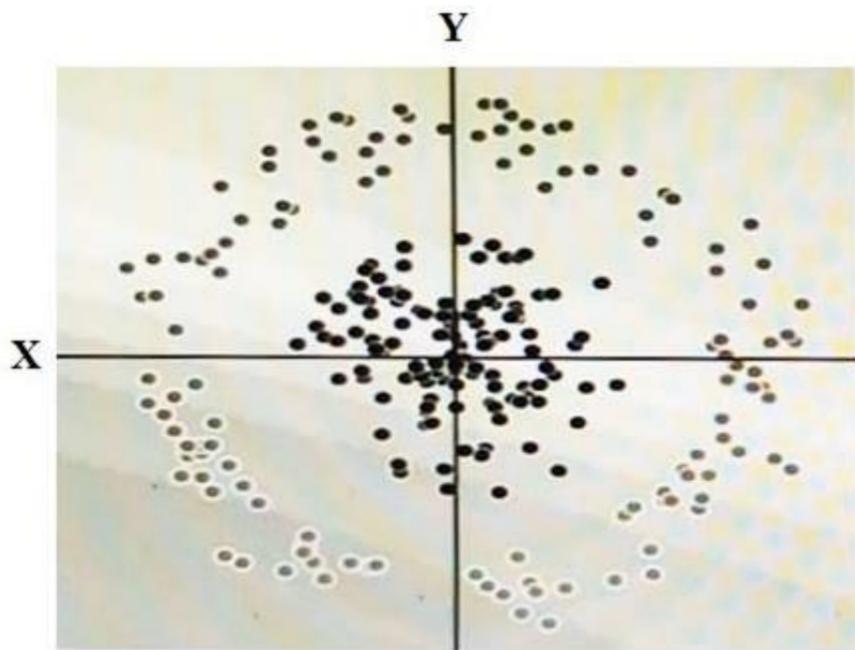
- A. Create a Stackdriver Monitoring dashboard based on the BigQuery metric query/scanned_bytes
- B. Create a Stackdriver Monitoring dashboard based on the BigQuery metric slots/allocated_for_project
- C. Create a log export for each project, capture the BigQuery job execution logs, create a custom metric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

Answer: D

NEW QUESTION 141

- (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 144

- (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 149

- (Exam Topic 6)

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

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

Answer: C

NEW QUESTION 154

- (Exam Topic 6)

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

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

Answer: C

Explanation:

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

NEW QUESTION 157

- (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 161

- (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 166

- (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 169

- (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 172

- (Exam Topic 6)

You are using Cloud Bigtable to persist and serve stock market data for each of the major indices. To serve the trading application, you need to access only the most recent stock prices that are streaming in. How should you design your row key and tables to ensure that you can access the data with the most simple query?

- A. Create one unique table for all of the indices, and then use the index and timestamp as the row key design
- B. Create one unique table for all of the indices, and then use a reverse timestamp as the row key design.
- C. For each index, have a separate table and use a timestamp as the row key design
- D. For each index, have a separate table and use a reverse timestamp as the row key design

Answer: A

NEW QUESTION 174

- (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 178

- (Exam Topic 6)

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

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

Which solution should you use?

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

Answer: A

NEW QUESTION 179

- (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 the stream changes into a daily inventory movement tabl
- D. Calculate balances in a view that joins it to the historical inventory balance tabl
- 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 tabl
- G. Update the inventory balance table nightly.

Answer: A

NEW QUESTION 181

- (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 183

- (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 availabilit
- B. Store the data in HDFS, and perform analysis as needed.
- C. Store the data in BigQuer
- D. Access the data using the BigQuery Connector or Cloud Dataproc and Compute Engine.
- E. Store the data in a regional Cloud Storage bucke

- F. Access the bucket directly using Cloud Dataproc, BigQuery, and Compute Engine.
- G. Store the data in a multi-regional Cloud Storage bucket
- H. Access the data directly using Cloud Dataproc, BigQuery, and Compute Engine.

Answer: D

NEW QUESTION 188

- (Exam Topic 6)

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

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

Answer: D

NEW QUESTION 190

- (Exam Topic 6)

A data scientist has created a BigQuery ML model and asks you to create an ML pipeline to serve predictions. You have a REST API application with the requirement to serve predictions for an individual user ID with latency under 100 milliseconds. You use the following query to generate predictions: `SELECT predicted_label, user_id FROM ML.PREDICT (MODEL 'dataset.model', table user_features)`. How should you create the ML pipeline?

- A. Add a WHERE clause to the query, and grant the BigQuery Data Viewer role to the application service account.
- B. Create an Authorized View with the provided query
- C. Share the dataset that contains the view with the application service account.
- D. Create a Cloud Dataflow pipeline using BigQueryIO to read results from the query
- E. Grant the Dataflow Worker role to the application service account.
- F. Create a Cloud Dataflow pipeline using BigQueryIO to read predictions for all users from the query. Write the results to Cloud Bigtable using BigtableIO
- G. Grant the Bigtable Reader role to the application service account so that the application can read predictions for individual users from Cloud Bigtable.

Answer: D

NEW QUESTION 194

- (Exam Topic 6)

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

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

Answer: D

NEW QUESTION 199

- (Exam Topic 6)

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

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

Answer: C

NEW QUESTION 204

- (Exam Topic 6)

Your company is implementing a data warehouse using BigQuery, and you have been tasked with designing the data model. You move your on-premises sales data warehouse with a star data schema to BigQuery but notice performance issues when querying the data of the past 30 days. Based on Google's recommended practices, what should you do to speed up the query without increasing storage costs?

- A. Denormalize the data
- B. Shard the data by customer ID
- C. Materialize the dimensional data in views
- D. Partition the data by transaction date

Answer: C

NEW QUESTION 207

- (Exam Topic 6)

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

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

Answer: B

NEW QUESTION 212

- (Exam Topic 6)

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

What should you do?

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

Answer: B

NEW QUESTION 213

- (Exam Topic 6)

Your infrastructure includes a set of YouTube channels. You have been tasked with creating a process for sending the YouTube channel data to Google Cloud for analysis. You want to design a solution that allows your world-wide marketing teams to perform ANSI SQL and other types of analysis on up-to-date YouTube channels log data. How should you set up the log data transfer into Google Cloud?

- A. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Multi-Regional storage bucket as a final destination.
- B. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Regional bucket as a final destination.
- C. Use BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage Multi-Regional storage bucket as a final destination.
- D. Use BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage Regional storage bucket as a final destination.

Answer: B

NEW QUESTION 218

- (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 223

- (Exam Topic 6)

You want to archive data in Cloud Storage. Because some data is very sensitive, you want to use the "Trust No One" (TNO) approach to encrypt your data to prevent the cloud provider staff from decrypting your data. What should you do?

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

Answer: B

NEW QUESTION 224

- (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 229

.....

THANKS FOR TRYING THE DEMO OF OUR PRODUCT

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

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

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

Money Back Guarantee

Professional-Data-Engineer Practice Exam Features:

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