

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

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

Answer: C

Explanation:

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

NEW QUESTION 3

- (Exam Topic 1)

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

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

Answer: A

NEW QUESTION 4

- (Exam Topic 1)

You are building a new real-time data warehouse for your company and will use Google BigQuery streaming inserts. There is no guarantee that data will only be sent in once but you do have a unique ID for each row of data and an event timestamp. You want to ensure that duplicates are not included while interactively querying data. Which query type should you use?

- A. Include ORDER BY DESK on timestamp column and LIMIT to 1.
- B. Use GROUP BY on the unique ID column and timestamp column and SUM on the values.
- C. Use the LAG window function with PARTITION by unique ID along with WHERE LAG IS NOT NULL.
- D. Use the ROW_NUMBER window function with PARTITION by unique ID along with WHERE row equals 1.

Answer: D

Explanation:

<https://cloud.google.com/bigquery/docs/reference/standard-sql/analytic-function-concepts>

NEW QUESTION 5

- (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 6

- (Exam Topic 1)

Your company has hired a new data scientist who wants to perform complicated analyses across very large datasets stored in Google Cloud Storage and in a Cassandra cluster on Google Compute Engine. The scientist primarily wants to create labelled data sets for machine learning projects, along with some visualization tasks. She reports that her laptop is not powerful enough to perform her tasks and it is slowing her down. You want to help her perform her tasks. What should you do?

- A. Run a local version of Jupiter on the laptop.
- B. Grant the user access to Google Cloud Shell.
- C. Host a visualization tool on a VM on Google Compute Engine.
- D. Deploy Google Cloud Datalab to a virtual machine (VM) on Google Compute Engine.

Answer: B

NEW QUESTION 7

- (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 8

- (Exam Topic 1)

You are building a model to predict whether or not it will rain on a given day. You have thousands of input features and want to see if you can improve training speed by removing some features while having a minimum effect on model accuracy. What can you do?

- A. Eliminate features that are highly correlated to the output labels.
- B. Combine highly co-dependent features into one representative feature.
- C. Instead of feeding in each feature individually, average their values in batches of 3.
- D. Remove the features that have null values for more than 50% of the training records.

Answer: B

NEW QUESTION 9

- (Exam Topic 1)

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

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

Answer: B

NEW QUESTION 10

- (Exam Topic 1)

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

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

Answer: C

NEW QUESTION 10

- (Exam Topic 1)

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

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

Answer: D

NEW QUESTION 15

- (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 clini
- E. Run queries against the smaller table pairs, and use unions for consolidated reports.

Answer: C

NEW QUESTION 20

- (Exam Topic 1)

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

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

Answer: C

Explanation:

Reference

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

NEW QUESTION 22

- (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 27

- (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 28

- (Exam Topic 3)

MJTelco is building a custom interface to share data. They have these requirements:

- They need to do aggregations over their petabyte-scale datasets.
- They need to scan specific time range rows with a very fast response time (milliseconds). Which combination of Google Cloud Platform products should you recommend?

- A. Cloud Datastore and Cloud Bigtable
- B. Cloud Bigtable and Cloud SQL
- C. BigQuery and Cloud Bigtable
- D. BigQuery and Cloud Storage

Answer: C

NEW QUESTION 31

- (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 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 40

- (Exam Topic 4)

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

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

Answer: B

NEW QUESTION 44

- (Exam Topic 4)

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 49

- (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 52

- (Exam Topic 5)

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

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

Answer: C

Explanation:

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

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

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

NEW QUESTION 55

- (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 58

- (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 60

- (Exam Topic 5)

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

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

Answer: D

Explanation:

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

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

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

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

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

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

NEW QUESTION 64

- (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 67

- (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 72

- (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 75

- (Exam Topic 5)

How would you query specific partitions in a BigQuery table?

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

Answer: C

Explanation:

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

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

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

NEW QUESTION 80

- (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 85

- (Exam Topic 5)

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

gives you an error.

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

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

Answer: A

Explanation:

To access the person.city column, you need to "UNNEST(person)" and JOIN it to table1 using a comma. Reference: https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql#nested_repeated_resu

NEW QUESTION 86

- (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 89

- (Exam Topic 5)

Which of these rules apply when you add preemptible workers to a Dataproc cluster (select 2 answers)?

- A. Preemptible workers cannot use persistent disk.
- B. Preemptible workers cannot store data.
- C. If a preemptible worker is reclaimed, then a replacement worker must be added manually.
- D. A Dataproc cluster cannot have only preemptible workers.

Answer: BD

Explanation:

The following rules will apply when you use preemptible workers with a Cloud Dataproc cluster: Processing only—Since preemptibles can be reclaimed at any time, preemptible workers do not store data.

Preemptibles added to a Cloud Dataproc cluster only function as processing nodes.

No preemptible-only clusters—To ensure clusters do not lose all workers, Cloud Dataproc cannot create preemptible-only clusters.

Persistent disk size—As a default, all preemptible workers are created with the smaller of 100GB or the primary worker boot disk size. This disk space is used for local caching of data and is not available through HDFS.

The managed group automatically re-adds workers lost due to reclamation as capacity permits. Reference:

<https://cloud.google.com/dataproc/docs/concepts/preemptible-vms>

NEW QUESTION 93

- (Exam Topic 5)

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

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

Answer: A

Explanation:

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

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

NEW QUESTION 96

- (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 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 101

- (Exam Topic 5)

Which of these statements about BigQuery caching is true?

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

Answer: D

Explanation:

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

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

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

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

NEW QUESTION 102

- (Exam Topic 5)

What is the general recommendation when designing your row keys for a Cloud Bigtable schema?

- A. Include multiple time series values within the row key
- B. Keep the row key as an 8 bit integer
- C. Keep your row key reasonably short
- D. Keep your row key as long as the field permits

Answer: C

Explanation:

A general guide is to, keep your row keys reasonably short. Long row keys take up additional memory and storage and increase the time it takes to get responses from the Cloud Bigtable server.

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

NEW QUESTION 104

- (Exam Topic 5)

Suppose you have a dataset of images that are each labeled as to whether or not they contain a human face. To create a neural network that recognizes human faces in images using this labeled dataset, what approach would likely be the most effective?

- A. Use K-means Clustering to detect faces in the pixels.
- B. Use feature engineering to add features for eyes, noses, and mouths to the input data.
- C. Use deep learning by creating a neural network with multiple hidden layers to automatically detect features of faces.
- D. Build a neural network with an input layer of pixels, a hidden layer, and an output layer with two categories.

Answer: C

Explanation:

Traditional machine learning relies on shallow nets, composed of one input and one output layer, and at most one hidden layer in between. More than three layers (including input and output) qualifies as "deep" learning. So deep is a strictly defined, technical term that means more than one hidden layer.

In deep-learning networks, each layer of nodes trains on a distinct set of features based on the previous layer's output. The further you advance into the neural net, the more complex the features your nodes can recognize, since they aggregate and recombine features from the previous layer.

A neural network with only one hidden layer would be unable to automatically recognize high-level features of faces, such as eyes, because it wouldn't be able to "build" these features using previous hidden layers that detect low-level features, such as lines.

Feature engineering is difficult to perform on raw image data.

K-means Clustering is an unsupervised learning method used to categorize unlabeled data. Reference: <https://deeplearning4j.org/neuralnet-overview>

NEW QUESTION 108

- (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 109

- (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 112

- (Exam Topic 5)

Dataprox 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 117

- (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

NEW QUESTION 119

- (Exam Topic 5)

For the best possible performance, what is the recommended zone for your Compute Engine instance and Cloud Bigtable instance?

- A. Have the Compute Engine instance in the furthest zone from the Cloud Bigtable instance.
- B. Have both the Compute Engine instance and the Cloud Bigtable instance to be in different zones.
- C. Have both the Compute Engine instance and the Cloud Bigtable instance to be in the same zone.
- D. Have the Cloud Bigtable instance to be in the same zone as all of the consumers of your data.

Answer: C

Explanation:

It is recommended to create your Compute Engine instance in the same zone as your Cloud Bigtable instance for the best possible performance, If it's not possible to create a instance in the same zone, you should create your instance in another zone within the same region. For example, if your Cloud Bigtable instance is located in us-central1-b, you could create your instance in us-central1-f. This change may result in several milliseconds of additional latency for each Cloud Bigtable request.

It is recommended to avoid creating your Compute Engine instance in a different region from your Cloud Bigtable instance, which can add hundreds of milliseconds of latency to each Cloud Bigtable request.

Reference: <https://cloud.google.com/bigtable/docs/creating-compute-instance>

NEW QUESTION 121

- (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 125

- (Exam Topic 5)

When running a pipeline that has a BigQuery source, on your local machine, you continue to get permission denied errors. What could be the reason for that?

- A. Your gcloud does not have access to the BigQuery resources
- B. BigQuery cannot be accessed from local machines
- C. You are missing gcloud on your machine
- D. Pipelines cannot be run locally

Answer: A

Explanation:

When reading from a Dataflow source or writing to a Dataflow sink using DirectPipelineRunner, the Cloud Platform account that you configured with the gcloud executable will need access to the corresponding source/sink

Reference:

<https://cloud.google.com/dataflow/java-sdk/JavaDoc/com/google/cloud/dataflow/sdk/runners/DirectPipelineRun>

NEW QUESTION 127

- (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 132

- (Exam Topic 5)

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

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

Answer: ABD

Explanation:

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

Reference: https://cloud.google.com/dataproc/docs/resources/faq#what_type_of_jobs_can_i_run

NEW QUESTION 135

- (Exam Topic 5)

Which methods can be used to reduce the number of rows processed by BigQuery?

- A. Splitting tables into multiple tables; putting data in partitions
- B. Splitting tables into multiple tables; putting data in partitions; using the LIMIT clause
- C. Putting data in partitions; using the LIMIT clause
- D. Splitting tables into multiple tables; using the LIMIT clause

Answer: A

Explanation:

If you split a table into multiple tables (such as one table for each day), then you can limit your query to the data in specific tables (such as for particular days). A better method is to use a partitioned table, as long as your data can be separated by the day.

If you use the LIMIT clause, BigQuery will still process the entire table. Reference: <https://cloud.google.com/bigquery/docs/partitioned-tables>

NEW QUESTION 140

- (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 144

- (Exam Topic 6)

You have Cloud Functions written in Node.js that pull messages from Cloud Pub/Sub and send the data to

BigQuery. You observe that the message processing rate on the Pub/Sub topic is orders of magnitude higher than anticipated, but there is no error logged in Stackdriver Log Viewer. What are the two most likely causes of this problem? Choose 2 answers.

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

Answer: CD

NEW QUESTION 148

- (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 150

- (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 155

- (Exam Topic 6)

You need to copy millions of sensitive patient records from a relational database to BigQuery. The total size of the database is 10 TB. You need to design a solution that is secure and time-efficient. What should you do?

- A. Export the records from the database as an Avro file
- B. Upload the file to GCS using gsutil, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.
- C. Export the records from the database as an Avro file
- D. Copy the file onto a Transfer Appliance and send it to Google, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.

- E. Export the records from the database into a CSV file
- F. Create a public URL for the CSV file, and then use Storage Transfer Service to move the file to Cloud Storage
- G. Load the CSV file into BigQuery using the BigQuery web UI in the GCP Console.
- H. Export the records from the database as an Avro file
- I. Create a public URL for the Avro file, and then use Storage Transfer Service to move the file to Cloud Storage
- J. Load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.

Answer: A

NEW QUESTION 158

- (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 162

- (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 166

- (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 169

- (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 171

- (Exam Topic 6)

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

What should you do?

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

Answer: B

NEW QUESTION 174

- (Exam Topic 6)

You work on a regression problem in a natural language processing domain, and you have 100M labeled examples in your dataset. You have randomly shuffled

your data and split your dataset into train and test samples (in a 90/10 ratio). After you trained the neural network and evaluated your model on a test set, you discover that the root-mean-squared error (RMSE) of your model is twice as high on the train set as on the test set. How should you improve the performance of your model?

- A. Increase the share of the test sample in the train-test split.
- B. Try to collect more data and increase the size of your dataset.
- C. Try out regularization techniques (e.g., dropout of batch normalization) to avoid overfitting.
- D. Increase the complexity of your model by, e.g., introducing an additional layer or increase sizing the size of vocabularies or n-grams used.

Answer: D

NEW QUESTION 179

- (Exam Topic 6)

Your team is working on a binary classification problem. You have trained a support vector machine (SVM) classifier with default parameters, and received an area under the Curve (AUC) of 0.87 on the validation set. You want to increase the AUC of the model. What should you do?

- A. Perform hyperparameter tuning
- B. Train a classifier with deep neural networks, because neural networks would always beat SVMs
- C. Deploy the model and measure the real-world AUC; it's always higher because of generalization
- D. Scale predictions you get out of the model (tune a scaling factor as a hyperparameter) in order to get the highest AUC

Answer: A

Explanation:

<https://towardsdatascience.com/understanding-hyperparameters-and-its-optimisation-techniques-f0debba07568>

NEW QUESTION 184

- (Exam Topic 6)

You architect a system to analyze seismic data. Your extract, transform, and load (ETL) process runs as a series of MapReduce jobs on an Apache Hadoop cluster. The ETL process takes days to process a data set because some steps are computationally expensive. Then you discover that a sensor calibration step has been omitted. How should you change your ETL process to carry out sensor calibration systematically in the future?

- A. Modify the transformMapReduce jobs to apply sensor calibration before they do anything else.
- B. Introduce a new MapReduce job to apply sensor calibration to raw data, and ensure all other MapReduce jobs are chained after this.
- C. Add sensor calibration data to the output of the ETL process, and document that all users need to apply sensor calibration themselves.
- D. Develop an algorithm through simulation to predict variance of data output from the last MapReduce job based on calibration factors, and apply the correction to all data.

Answer: A

NEW QUESTION 189

- (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 194

- (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 198

- (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: D

NEW QUESTION 199

- (Exam Topic 6)

You need to choose a database for a new project that has the following requirements:

- Fully managed
- Able to automatically scale up
- Transactionally consistent
- Able to scale up to 6 TB
- Able to be queried using SQL Which database do you choose?

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

Answer: C

NEW QUESTION 203

- (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 DD
- B. Partition the tables by a column containing a TIMESTAMP or DATETIME.
- 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 207

- (Exam Topic 6)

You have data pipelines running on BigQuery, Cloud Dataflow, and Cloud Dataproc. You need to perform health checks and monitor their behavior, and then notify the team managing the pipelines if they fail. You also need to be able to work across multiple projects. Your preference is to use managed products or features of the platform. What should you do?

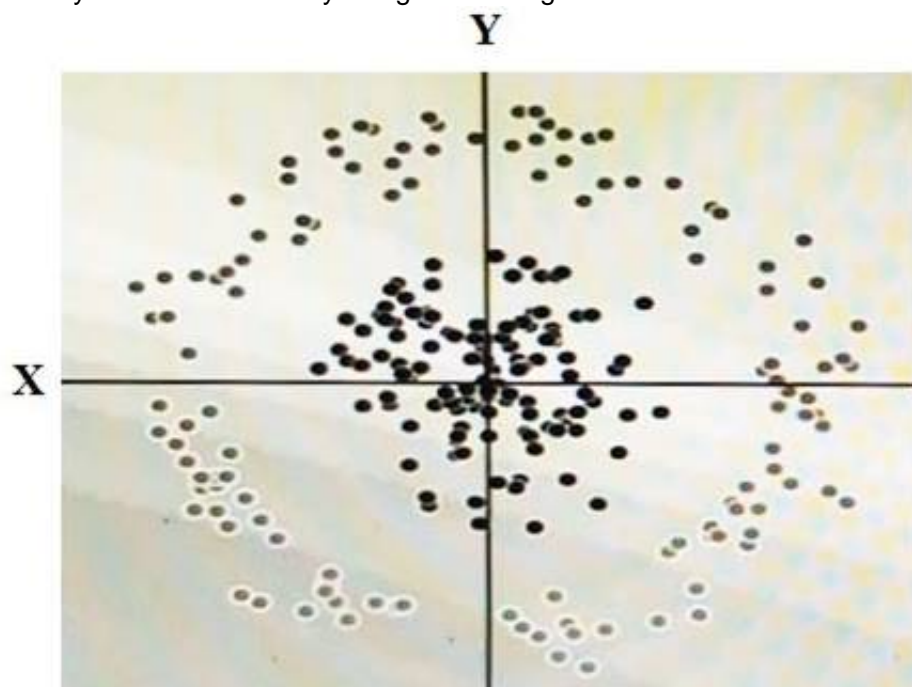
- A. Export the information to Cloud Stackdriver, and set up an Alerting policy
- B. Run a Virtual Machine in Compute Engine with Airflow, and export the information to Stackdriver
- C. Export the logs to BigQuery, and set up App Engine to read that information and send emails if you find a failure in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

Answer: B

NEW QUESTION 209

- (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?

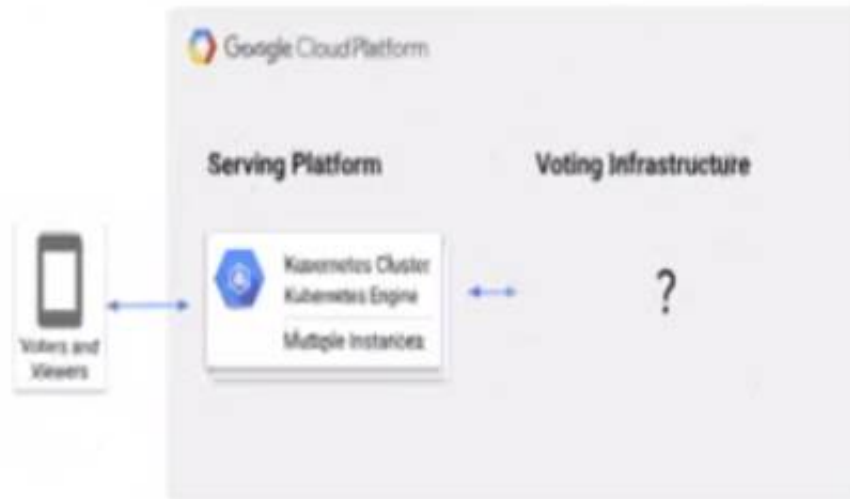
- A. X^2+Y^2
- B. X^2
- C. Y^2
- D. $\cos(X)$

Answer: D

NEW QUESTION 210

- (Exam Topic 6)

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



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

Answer: C

NEW QUESTION 213

- (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 217

- (Exam Topic 6)

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

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

Answer: C

NEW QUESTION 219

- (Exam Topic 6)

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

- A. Perform a DML INSERT UPDATE, or DELETE to replicate each CDC record in the reporting table m real time.

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

Answer: BD

NEW QUESTION 222

- (Exam Topic 6)

You are selecting services to write and transform JSON messages from Cloud Pub/Sub to BigQuery for a data pipeline on Google Cloud. You want to minimize service costs. You also want to monitor and accommodate input data volume that will vary in size with minimal manual intervention. What should you do?

- A. Use Cloud Dataproc to run your transformation
- B. Monitor CPU utilization for the cluster
- C. Resize the number of worker nodes in your cluster via the command line.
- D. Use Cloud Dataproc to run your transformation
- E. Use the diagnose command to generate an operational output archive
- F. Locate the bottleneck and adjust cluster resources.
- G. Use Cloud Dataflow to run your transformation
- H. Monitor the job system lag with Stackdriver
- I. Use the default autoscaling setting for worker instances.
- J. Use Cloud Dataflow to run your transformation
- K. Monitor the total execution time for a sampling of jobs
- L. Configure the job to use non-default Compute Engine machine types when needed.

Answer: B

NEW QUESTION 226

- (Exam Topic 6)

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

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

Answer: C

Explanation:

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

NEW QUESTION 228

- (Exam Topic 6)

You are migrating your data warehouse to BigQuery. You have migrated all of your data into tables in a dataset. Multiple users from your organization will be using the data. They should only see certain tables based on their team membership. How should you set user permissions?

- A. Assign the users/groups data viewer access at the table level for each table
- B. Create SQL views for each team in the same dataset in which the data resides, and assign the users/groups data viewer access to the SQL views
- C. Create authorized views for each team in the same dataset in which the data resides, and assign the users/groups data viewer access to the authorized views
- D. Create authorized views for each team in datasets created for each team
- E. Assign the authorized views data viewer access to the dataset in which the data reside
- F. Assign the users/groups data viewer access to the datasets in which the authorized views reside

Answer: A

NEW QUESTION 231

- (Exam Topic 6)

You have a requirement to insert minute-resolution data from 50,000 sensors into a BigQuery table. You expect significant growth in data volume and need the data to be available within 1 minute of ingestion for real-time analysis of aggregated trends. What should you do?

- A. Use bq load to load a batch of sensor data every 60 seconds.
- B. Use a Cloud Dataflow pipeline to stream data into the BigQuery table.
- C. Use the INSERT statement to insert a batch of data every 60 seconds.
- D. Use the MERGE statement to apply updates in batch every 60 seconds.

Answer: C

NEW QUESTION 232

- (Exam Topic 6)

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

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

- C. Call the Cloud Natural Language API from your application
- D. Process the generated Sentiment Analysis as labels.
- E. Build and train a text classification model using TensorFlo
- F. Deploy the model using Cloud MachineLearning Engin
- G. Call the model from your application and process the results as labels.
- H. Build and train a text classification model using TensorFlo
- I. Deploy the model using a Kubernetes Engine cluste
- J. Call the model from your application and process the results as labels.

Answer: B

NEW QUESTION 236

- (Exam Topic 6)

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

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

Answer: C

NEW QUESTION 237

- (Exam Topic 6)

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

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

Answer: A

NEW QUESTION 241

- (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 243

- (Exam Topic 6)

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

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

What should you do?

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

Answer: A

Explanation:

you can use dataprep for continuously changing target schema

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

Names of columns

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

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

https://cloud.google.com/dataprep/docs/html/Overview-of-RapidTarget_136155049

NEW QUESTION 244

- (Exam Topic 6)

You are designing a pipeline that publishes application events to a Pub/Sub topic. You need to aggregate events across hourly intervals before loading the results to BigQuery for analysis. Your solution must be scalable so it can process and load large volumes of events to BigQuery. What should you do?

- A. Create a streaming Dataflow job to continually read from the Pub/Sub topic and perform the necessary aggregations using tumbling windows
- B. Schedule a batch Dataflow job to run hourly, pulling all available messages from the Pub-Sub topic and performing the necessary aggregations
- C. Schedule a Cloud Function to run hourly, pulling all avertable messages from the Pub/Sub topic and performing the necessary aggregations
- D. Create a Cloud Function to perform the necessary data processing that executes using the Pub/Sub trigger every time a new message is published to the topic.

Answer: A

NEW QUESTION 249

- (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 252

- (Exam Topic 6)

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

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

Answer: C

NEW QUESTION 256

- (Exam Topic 6)

You have a data pipeline with a Cloud Dataflow job that aggregates and writes time series metrics to Cloud Bigtable. This data feeds a dashboard used by thousands of users across the organization. You need to support additional concurrent users and reduce the amount of time required to write the data. Which two actions should you take? (Choose two.)

- A. Configure your Cloud Dataflow pipeline to use local execution
- B. Increase the maximum number of Cloud Dataflow workers by setting `maxNumWorkers` in `PipelineOptions`
- C. Increase the number of nodes in the Cloud Bigtable cluster
- D. Modify your Cloud Dataflow pipeline to use the Flatten transform before writing to Cloud Bigtable
- E. Modify your Cloud Dataflow pipeline to use the `CoGroupByKey` transform before writing to Cloud Bigtable

Answer: BC

NEW QUESTION 257

- (Exam Topic 6)

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

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

Answer: B

NEW QUESTION 261

- (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 266

- (Exam Topic 6)

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

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

Answer: D

Explanation:

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

NEW QUESTION 270

- (Exam Topic 6)

Your company needs to upload their historic data to Cloud Storage. The security rules don't allow access from external IPs to their on-premises resources. After an initial upload, they will add new data from existing on-premises applications every day. What should they do?

- A. Execute gsutil rsync from the on-premises servers.
- B. Use Cloud Dataflow and write the data to Cloud Storage.
- C. Write a job template in Cloud Dataproc to perform the data transfer.
- D. Install an FTP server on a Compute Engine VM to receive the files and move them to Cloud Storage.

Answer: B

NEW QUESTION 275

- (Exam Topic 6)

You need to create a data pipeline that copies time-series transaction data so that it can be queried from within BigQuery by your data science team for analysis. Every hour, thousands of transactions are updated with a new status. The size of the initial dataset is 1.5 PB, and it will grow by 3 TB per day. The data is heavily structured, and your data science team will build machine learning models based on this data. You want to maximize performance and usability for your data science team. Which two strategies should you adopt? Choose 2 answers.

- A. Denormalize the data as much as possible.
- B. Preserve the structure of the data as much as possible.
- C. Use BigQuery UPDATE to further reduce the size of the dataset.
- D. Develop a data pipeline where status updates are appended to BigQuery instead of updated.
- E. Copy a daily snapshot of transaction data to Cloud Storage and store it as an Avro file.
- F. Use BigQuery's support for external data sources to query.

Answer: AE

NEW QUESTION 279

- (Exam Topic 6)

You set up a streaming data insert into a Redis cluster via a Kafka cluster. Both clusters are running on Compute Engine instances. You need to encrypt data at rest with encryption keys that you can create, rotate, and destroy as needed. What should you do?

- A. Create a dedicated service account, and use encryption at rest to reference your data stored in your Compute Engine cluster instances as part of your API service calls.
- B. Create encryption keys in Cloud Key Management Service.
- C. Use those keys to encrypt your data in all of the Compute Engine cluster instances.
- D. Create encryption keys locally.
- E. Upload your encryption keys to Cloud Key Management Service.
- F. Use those keys to encrypt your data in all of the Compute Engine cluster instances.
- G. Create encryption keys in Cloud Key Management Service.
- H. Reference those keys in your API service calls when accessing the data in your Compute Engine cluster instances.

Answer: C

NEW QUESTION 280

- (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 capture 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 282

- (Exam Topic 6)

You are working on a linear regression model on BigQuery ML to predict a customer's likelihood of purchasing your company's products. Your model uses a city name variable as a key predictive component in order to train and serve the model your data must be organized in columns. You want to prepare your data using the least amount of coding while maintaining the predictable variables. What should you do?

- A. Use SQL in BigQuery to transform the stale column using a one-hot encoding method, and make each city a column with binary values.
- B. Create a new view with BigQuery that does not include a column which city information.
- C. Cloud Data Fusion to assign each city to a region that is labeled as 1, 2 3, 4, or 5, and then use that number to represent the city in the model.
- D. Use TensorFlow to create a categorical variable with a vocabulary lis
- E. Create the vocabulary file and upload that as part of your model to BigQuery ML.

Answer: C

NEW QUESTION 286

- (Exam Topic 6)

You are designing storage for 20 TB of text files as part of deploying a data pipeline on Google Cloud. Your input data is in CSV format. You want to minimize the cost of querying aggregate values for multiple users who will query the data in Cloud Storage with multiple engines. Which storage service and schema design should you use?

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

Answer: A

NEW QUESTION 287

- (Exam Topic 6)

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

- A. Partition table data by create_date, location_id and device_version
- B. Partition table data by create_date cluster table data by location_Id and device_version
- C. Cluster table data by create_date location_id and device_version
- D. Cluster table data by create_date partition by located and device_version

Answer: B

NEW QUESTION 289

- (Exam Topic 6)

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

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

Answer: D

NEW QUESTION 293

- (Exam Topic 6)

You are integrating one of your internal IT applications and Google BigQuery, so users can query BigQuery from the application's interface. You do not want individual users to authenticate to BigQuery and you do not want to give them access to the dataset. You need to securely access BigQuery from your IT application. What should you do?

- A. Create groups for your users and give those groups access to the dataset
- B. Integrate with a single sign-on (SSO) platform, and pass each user's credentials along with the query request
- C. Create a service account and grant dataset access to that accoun
- D. Use the service account's private key to access the dataset
- E. Create a dummy user and grant dataset access to that use
- F. Store the username and password for that user in a file on the files system, and use those credentials to access the BigQuery dataset

Answer: C

NEW QUESTION 294

- (Exam Topic 6)

You are collecting IoT sensor data from millions of devices across the world and storing the data in BigQuery. Your access pattern is based on recent data tittered

by location_id and device_version with the following query:

```
SELECT
    MAX(temperature)
FROM
    acme_iot_data.sensors
WHERE
    create_date > DATE_SUB(CURRENT_DATE(), INTERVAL 7 day)
    AND location_id = "SW1W9TQ"
    AND device_version = "202007r3"
```

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

- A. Partition table data by create_date, location_id and device_version
- B. Partition table data by create_date cluster table data by location_id and device_version
- C. Cluster table data by create_date location_id and device_version
- D. Cluster table data by create_date, partition by location and device_version

Answer: C

NEW QUESTION 297

- (Exam Topic 6)

You have uploaded 5 years of log data to Cloud Storage. A user reported that some data points in the log data are outside of their expected ranges, which indicates errors. You need to address this issue and be able to run the process again in the future while keeping the original data for compliance reasons. What should you do?

- A. Import the data from Cloud Storage into BigQuery. Create a new BigQuery table, and skip the rows with errors.
- B. Create a Compute Engine instance and create a new copy of the data in Cloud Storage. Skip the rows with errors.
- C. Create a Cloud Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to a new dataset in Cloud Storage.
- D. Create a Cloud Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to the same dataset in Cloud Storage.

Answer: D

NEW QUESTION 301

- (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 302

- (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 303

- (Exam Topic 6)

You are building a report-only data warehouse where the data is streamed into BigQuery via the streaming API. Following Google's best practices, you have both a staging and a production table for the data. How should you design your data loading to ensure that there is only one master dataset without affecting performance on either the ingestion or reporting pieces?

- A. Have a staging table that is an append-only model, and then update the production table every three hours with the changes written to staging.
- B. Have a staging table that is an append-only model, and then update the production table every ninety minutes with the changes written to staging.
- C. Have a staging table that moves the staged data over to the production table and deletes the contents of the staging table every three hours.
- D. Have a staging table that moves the staged data over to the production table and deletes the contents of the staging table every thirty minutes.

Answer: D

NEW QUESTION 308

- (Exam Topic 6)

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

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

Answer: A**NEW QUESTION 309**

- (Exam Topic 6)

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

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

Answer: A**NEW QUESTION 310**

- (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 314**

- (Exam Topic 6)

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

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

Answer: C**Explanation:**

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

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

NEW QUESTION 316

- (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 319

- (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 321

- (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 storage
- B. Add secondary indexes to support query patterns.
- C. Use Cloud SQL for storage
- D. Use Cloud Dataflow to transform data to support query patterns.
- E. Use Cloud Spanner for storage
- F. Add secondary indexes to support query patterns.
- G. Use Cloud Spanner for storage
- 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 324

- (Exam Topic 6)

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

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

Answer: A

NEW QUESTION 327

- (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 328

- (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 329

.....

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