Vendor: Google

Exam Code:PROFESSIONAL-DATA-ENGINEER

Exam Name: Professional Data Engineer on Google Cloud Platform

Version: Demo

QUESTION 1

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

Correct Answer: B

QUESTION 2

When a Cloud Bigtable node fails, ____ is lost.

A. all data

B. no data

C. the last transaction

D. the time dimension

Correct Answer: B

A Cloud Bigtable table is sharded into blocks of contiguous rows, called tablets, to help balance the workload of queries. Tablets are stored on Colossus, Google\\'s file system, in SSTable format. Each tablet is associated with a specific Cloud Bigtable node.

Data is never stored in Cloud Bigtable nodes themselves; each node has pointers to a set of tablets that are stored on Colossus. As a result:

Rebalancing tablets from one node to another is very fast, because the actual data is not copied. Cloud Bigtable simply updates the pointers for each node. Recovery from the failure of a Cloud Bigtable node is very fast, because only metadata needs to be migrated to the replacement node.

When a Cloud Bigtable node fails, no data is lost Reference: https://cloud.google.com/bigtable/docs/overview

---

QUESTION 3

Cloud Dataproc is a managed Apache Hadoop and Apache _____ service.

A. Blaze

B. Spark

C. Fire

D. Ignite

Correct Answer: B

Cloud Dataproc is a managed Apache Spark and Apache Hadoop service that lets you use open source data tools for batch processing, querying, streaming, and machine learning. Reference: https://cloud.google.com/dataproc/docs/

QUESTION 4

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.

Correct Answer: C

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

QUESTION 5

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 IO. Set a sliding time window of 1 hour every 5 minutes. Compute the average when the window closes, and send an alert if the average is less than 4000 messages.

B. Consume the stream of data in Cloud Dataflow using Kafka IO. 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.

C. Use Kafka Connect to link your Kafka message queue to Cloud Pub/Sub. Use a Cloud Dataflow template to write your messages from Cloud Pub/Sub to Cloud Bigtable. Use Cloud Scheduler to run a script every hour that counts the number of rows created in Cloud Bigtable in the last hour. If that number falls below 4000, send an alert.

D. Use Kafka Connect to link your Kafka message queue to Cloud Pub/Sub. Use a Cloud Dataflow template to write your messages from Cloud Pub/Sub to BigQuery. Use Cloud Scheduler to run a script every five minutes that counts the number of rows created in BigQuery in the last hour. If that number falls below 4000, send an alert.

Correct Answer: A

QUESTION 6

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.

Correct Answer: B

QUESTION 7

What is the HBase Shell for Cloud Bigtable?

A. The HBase shell is a GUI based interface that performs administrative tasks, such as creating and deleting tables.

B. The HBase shell is a command-line tool that performs administrative tasks, such as creating and deleting tables.

C. The HBase shell is a hypervisor based shell that performs administrative tasks, such as creating and deleting new virtualized instances.

D. The HBase shell is a command-line tool that performs only user account management functions to grant access to Cloud Bigtable instances.

Correct Answer: B

The HBase shell is a command-line tool that performs administrative tasks, such as creating and deleting tables. The

Cloud Bigtable HBase client for Java makes it possible to use the HBase shell to connect to Cloud Bigtable. Reference: https://cloud.google.com/bigtable/docs/installing-hbase-shell

---

QUESTION 8

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

A. cron

B. Cloud Composer

C. Cloud Scheduler

D. Workflow Templates on Cloud Dataproc

Correct Answer: B

QUESTION 9

Which of the following is NOT true about Dataflow pipelines?

A. Dataflow pipelines are tied to Dataflow, and cannot be run on any other runner

B. Dataflow pipelines can consume data from other Google Cloud services

C. Dataflow pipelines can be programmed in Java

D. Dataflow pipelines use a unified programming model, so can work both with streaming and batch data sources

Correct Answer: A

Dataflow pipelines can also run on alternate runtimes like Spark and Flink, as they are built using the Apache Beam SDKs Reference: https://cloud.google.com/dataflow/

QUESTION 10

You want to rebuild your batch pipeline for structured data on Google Cloud You are using PySpark to conduct data transformations at scale, but your pipelines are taking over twelve hours to run To expedite development and pipeline run time, you want to use a serverless tool and SQL syntax You have already moved your raw data into Cloud Storage

How should you build the pipeline on Google Cloud while meeting speed and processing requirements?

A. Convert your PySpark commands into SparkSQL queries to transform the data; and then run your pipeline on Dataproc to write the data into BigQuery

B. Ingest your data into Cloud SQL, convert your PySpark commands into SparkSQL queries to transform the data, and then use federated queries from BigQuery for machine learning.

C. Ingest your data into BigQuery from Cloud Storage, convert your PySpark commands into BigQuery SQL queries to transform the data, and then write the transformations to a new table

D. Use Apache Beam Python SDK to build the transformation pipelines, and write the data into BigQuery

Correct Answer: C

QUESTION 11

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.

Correct Answer: B

QUESTION 12

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

A. Primitive role

B. Predefined role

C. Authorized view

D. It\\'s not possible to give access to only the first three columns of a table.

Correct Answer: C

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

view.

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

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