PR OFESSIONAL-DATA-ENGINEER Q&As

Professional Data Engineer on Google Cloud Platform

Pass Google PROFESSIONAL-DATA-ENGINEER Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.certbus.com/professional-data-engineer.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Google Official Exam Center

QUESTION 1

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

Correct Answer: A

QUESTION 2

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

A. Create a Cloud SQL instance in one zone, and create a failover replica in another zone within the same region.

B. Create a Cloud SQL instance in one zone, and create a read replica in another zone within the same region.

C. Create a Cloud SQL instance in one zone, and configure an external read replica in a zone in a different region.

D. Create a Cloud SQL instance in a region, and configure automatic backup to a Cloud Storage bucket in the same region.

Correct Answer: C

QUESTION 3

You need to create a near real-time inventory dashboard that reads the main inventory tables in your BigQuery data warehouse. Historical inventory data is stored as inventory balances by item and location. You have several thousand

updates to inventory every hour. You want to maximize performance of the dashboard and ensure that the data is accurate.

What should you do?

A. Leverage BigQuery UPDATE statements to update the inventory balances as they are changing.

B. Partition the inventory balance table by item to reduce the amount of data scanned with each inventory update.

C. Use the BigQuery streaming the stream changes into a daily inventory movement table.Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.

D. 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. Update the inventory balance table nightly.

Correct Answer: A

QUESTION 4

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

Correct Answer: A

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

QUESTION 5

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

1.

Real-time event stream

2.

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

Which solution should you use?

A. Cloud Dataflow, Cloud SQL, Cloud Spanner

B. Cloud Pub/Sub, Cloud Storage, BigQuery

C. Cloud Dataproc, Cloud Dataflow, BigQuery

D. Cloud Pub/Sub, Cloud Dataproc, Cloud SQL

Correct Answer: A

QUESTION 6

You decided to use Cloud Datastore to ingest vehicle telemetry data in real time. You want to build a storage system that will account for the long-term data growth, while keeping the costs low. You also want to create snapshots of the data periodically, so that you can make a point-in-time (PIT) recovery, or clone a copy of the data for Cloud Datastore in a different environment. You want to archive these snapshots for a long time. Which two methods can accomplish this? Choose 2 answers.

A. Use managed export, and store the data in a Cloud Storage bucket using Nearline or Coldline class.

B. Use managed exportm, and then import to Cloud Datastore in a separate project under a unique namespace reserved for that export.

C. Use managed export, and then import the data into a BigQuery table created just for that export, and delete temporary export files.

D. Write an application that uses Cloud Datastore client libraries to read all the entities. Treat each entity as a BigQuery table row via BigQuery streaming insert. Assign an export timestamp for each export, and attach it as an extra column for each row. Make sure that the BigQuery table is partitioned using the export timestamp column.

E. Write an application that uses Cloud Datastore client libraries to read all the entities. Format the exported data into a JSON file. Apply compression before storing the data in Cloud Source Repositories.

Correct Answer: CE

QUESTION 7

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. Use a standard persistent disk and 50% preemptible workers. Store data in Cloud Storage, and change references in scripts from hdfs:// to gs://

B. Deploy a Cloud Dataproc cluster. Use an SSD persistent disk and 50% preemptible workers. Store data in Cloud Storage, and change references in scripts from hdfs:// to gs://

C. Install Hadoop and Spark on a 10-node Compute Engine instance group with standard instances. Install the Cloud Storage connector, and store the data in Cloud Storage. Change references in scripts from hdfs:// to gs://

D. Install Hadoop and Spark on a 10-node Compute Engine instance group with preemptible instances. Store data in HDFS. Change references in scripts from hdfs:// to gs://

Correct Answer: A

QUESTION 8

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.

Correct Answer: D

QUESTION 9

Does Dataflow process batch data pipelines or streaming data pipelines?

A. Only Batch Data Pipelines

B. Both Batch and Streaming Data Pipelines

C. Only Streaming Data Pipelines

D. None of the above

Correct Answer: B

Dataflow is a unified processing model, and can execute both streaming and batch data pipelines Reference: https://cloud.google.com/dataflow/

QUESTION 10

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

A. Change the processing job to use Google Cloud Dataproc instead.

B. Manually start the Cloud Dataflow job each morning when you get into the office.

C. Create a cron job with Google App Engine Cron Service to run the Cloud Dataflow job.

D. Configure the Cloud Dataflow job as a streaming job so that it processes the log data immediately.

Correct Answer: C

QUESTION 11

Your team is building a data lake platform on Google Cloud. As a part of the data foundation design, you are planning to store all the raw data in Cloud Storage You are expecting to ingest approximately 25 GB of data a day and your billing department is worried about the increasing cost of storing old data. The current business requirements are:

1.

The old data can be deleted anytime

2.

You plan to use the visualization layer for current and historical reporting

3.

The old data should be available instantly when accessed

4.

There should not be any charges for data retrieval.

What should you do to optimize for cost?

A. Create the bucket with the Autoclass storage class feature.

B. Create an Object Lifecycle Management policy to modify the storage class for data older than 30 days to nearline, 90 days to coldline. and 365 days to archive storage class. Delete old data as needed.

C. Create an Object Lifecycle Management policy to modify the storage class for data older than 30 days to coldline, 90 days to nearline. and 365 days to archive storage class Delete old data as needed.

D. Create an Object Lifecycle Management policy to modify the storage class for data older than 30 days to nearlme. 45 days to coldline. and 60 days to archive storage class Delete old data as needed.

Correct Answer: A

Autoclass automatically moves objects between storage classes without impacting performance or availability, nor incurring retrieval costs. It continuously optimizes storage costs based on access patterns without the need to set specific lifecycle management policies.

QUESTION 12

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

Correct Answer: A

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

---

QUESTION 13

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

You have the following requirements:

1.

You will batch-load the posts once per day and run them through the Cloud Natural Language API.

2.

You will extract topics and sentiment from the posts.

3.

You must store the raw posts for archiving and reprocessing.

4.

You will create dashboards to be shared with people both inside and outside your organization.

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

A. Store the social media posts and the data extracted from the API in BigQuery.

B. Store the social media posts and the data extracted from the API in Cloud SQL.

C. Store the raw social media posts in Cloud Storage, and write the data extracted from the API into BigQuery.

D. Feed to social media posts into the API directly from the source, and write the extracted data from the API into BigQuery.

Correct Answer: D

QUESTION 14

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

Correct Answer: A

QUESTION 15

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

Correct Answer: C

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