Free Questions for Professional-Data-Engineer by dumpssheet

Shared by Schneider on 12-12-2023

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

Your company currently runs a large on-premises cluster using Spark Hive and Hadoop Distributed File System (HDFS) in a colocation facility. The duster is designed to support peak usage on the system, however, many jobs are batch n 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 servers offerings m order to take advantage of the cloud Because of the tuning of their contract renewal with the colocation facility they have only 2 months for their initial migration How should you recommend they approach thee upcoming migration strategy so they can maximize their cost savings in the cloud will still executing the migration in time?

Options:

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

Question 2

Question Type: MultipleChoice

You've migrated a Hadoop job from an on-premises cluster to Dataproc and Good Storage. Your Spark job is a complex analytical workload fiat consists of many shuffling operations, and initial data are parquet toes (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 Idee to continue using Dataproc on preemptibles (with 2 non-preemptibles workers only) for this workload. What should you do?

Options:

A- Switch from HODs to SSDs override the preemptible VMs configuration to increase the boot disk size

B- Increase the see of your parquet files to ensure them to be 1 GB minimum

C- Switch to TFRecords format (appr 200 MB per We) instead of parquet files

D- Switch from HDDs to SSDs. 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

Question 3

Question Type: **MultipleChoice**

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

Options:

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 table
Periodically use a DML MERGE to simultaneously perform DML INSERT. UPDATE, and DELETE operations in the reporting table

C- Insert each new CDC record and corresponding operation type into a staging table in real time

D- 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:

B, D

Question 4

Question Type: MultipleChoice

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?

Options:

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

Question 5

Question Type: MultipleChoice

An online brokerage company requires a high volume trade processing architecture. You need to create a secure queuing system that triggers jobs. The jobs will run in Google Cloud and cat the company's Python API to execute trades. You need to efficiently implement a solution. What should you do?

Options:

A- Use Cloud Composer to subscribe to a Pub/Sub tope and can the Python API.

B- Use a Pub/Sub push subscription to trigger a Cloud Function to pass the data to tie Python API.

C- Write an application that makes a queue in a NoSQL database

D- Write an application hosted on a Compute Engine instance that makes a push subscription to the Pub/Sub topic

Answer:

C

Question 6

Question Type: **MultipleChoice**

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?

Options:

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

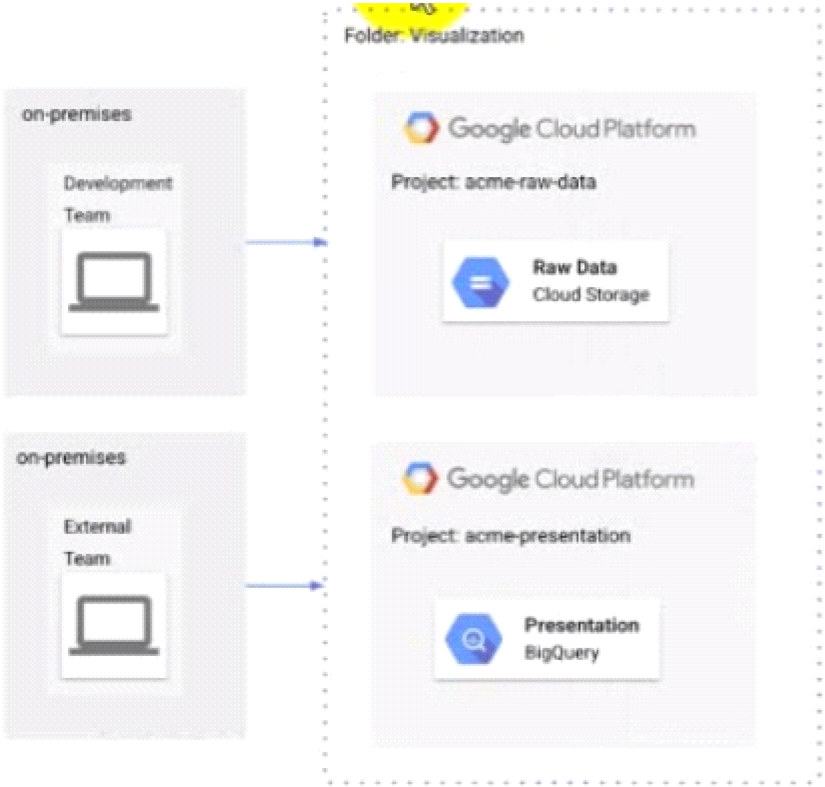
Answer:

C

Question 7

Question Type: MultipleChoice

The Development and External teams nave the project viewer Identity and Access Management (1AM) role m a folder named Visualization. You want the Development Team to be able to read data from both Cloud Storage and BigQuery, but the External Team should only be able to read data from BigQuery. What should you do?



Options:

A- Remove Cloud Storage IAM permissions to the External Team on the acme-raw-data project

B- Create Virtual Private Cloud (VPC) firewall rules on the acme-raw-data protect that deny all Ingress traffic from the External Team CIDR range

C- Create a VPC Service Controls perimeter containing both protects and BigQuery as a restricted API Add the External Team users to the perimeter s Access Level

D- Create a VPC Service Controls perimeter containing both protects and Cloud Storage as a restricted API. Add the Development Team users to the perimeter's Access Level

Answer:

C