Google Professional-Data-Engineer Practice Questions

Shared by Robinson on 24-08-2026

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

You want to migrate an Apache Spark 3 batch job from on-premises to Google Cloud. You need to minimally change the job so that the job reads from Cloud Storage and writes the result to BigQuery. Your job is optimized for Spark, where each executor has 8 vCPU and 16 GB memory, and you want to be able to choose similar settings. You want to minimize installation and management effort to run your job. What should you do?

Options:

A- Execute the job in a new Dataproc cluster.
B- Execute as a Dataproc Serverless job.
C- Execute the job as part of a deployment in a new Google Kubernetes Engine cluster.
D- Execute the job from a new Compute Engine VM.

Answer:

B

Explanation:

The key requirements are:

Migrate Spark 3 batch job.

Minimally change the job (reads from GCS, writes to BQ -- standard for Spark on GCP).

Job optimized for Spark (specific executor vCPU/memory).

Ability to choose similar executor settings.

Minimize installation and management effort.

Dataproc Serverless (Option A)is designed for these use cases.

Spark 3 Support:Dataproc Serverless supports various Spark runtimes, including Spark 3.

Minimal Changes:Spark jobs reading from GCS and writing to BigQuery (using the Spark-BigQuery connector) are standard. Minimal code changes are generally needed.

Customizable Resources:Dataproc Serverless allows you to specify resources for the driver and executors, including vCPU and memory. You can configure these to match your optimized on-premises settings (e.g., 8 vCPU, 16 GB memory per executor, though specific available configurations should be checked).

Minimal Installation and Management:This is the core benefit of 'serverless.' You don't need to provision, manage, or scale clusters. You submit your batch job, and Google Cloud handles the underlying infrastructure. This significantly reduces operational overhead.

Let's analyze why other options are less suitable:

B (Compute Engine VM):You would need to manually install Spark, configure it, manage dependencies, and manage the VM itself. This is high management effort.

C (Google Kubernetes Engine cluster):While you can run Spark on GKE (e.g., using Spark on Kubernetes operator), it involves managing the GKE cluster, Spark deployment configurations, Docker images, etc. This is also significant management effort, more than Dataproc Serverless.

D (Dataproc cluster):A standard Dataproc cluster provides more control than serverless but also requires more management (creating, scaling, and managing the cluster lifecycle). Dataproc Serverless is specifically designed to minimize this management for batch jobs. Given the 'minimize installation and management effort' requirement, serverless is preferred over a managed cluster if it meets the job's needs.

Google Cloud Documentation: Dataproc Serverless > Overview. 'Dataproc Serverless for Spark lets you run Spark batch workloads without requiring you to provision and manage your own cluster... Submit your Spark workload to the Dataproc Serverless service. The service will run the workload on a managed compute infrastructure, autoscaling resources as needed.'

Google Cloud Documentation: Dataproc Serverless > Submitting Spark batch workloads > Spark batch workload properties. This documentation shows how you can specify properties for driver and executor cores (spark.driver.cores, spark.executor.cores) and memory (spark.driver.memory, spark.executor.memory), allowing you to choose settings similar to your existing optimized job.

Question 2

Question Type: MultipleChoice

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 dat

a. They should only see certain tables based on their team membership. How should you set user permissions?

Options:

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. Assign the authorized views data viewer access to the dataset in which the data resides. Assign the users/groups data viewer access to the datasets in which the authorized views reside

Answer:

A

Question 3

Question Type: MultipleChoice

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?

Options:

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

Question 4

Question Type: MultipleChoice

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?

Options:

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

Question 5

Question Type: MultipleChoice

You have a streaming pipeline that ingests data from Pub/Sub in production. You need to update this streaming pipeline with improved business logic. You need to ensure that the updated pipeline reprocesses the previous two days of delivered Pub/Sub messages. What should you do?

Select 2 answers

Options:

A- Use Pub/Sub Seek with a timestamp.
B- Use the Pub/Sub subscription clear-retry-policy flag.
C- Create a new Pub/Sub subscription two days before the deployment.
D- Use the Pub/Sub subscription retain-asked-messages flag.
E- Use Pub/Sub Snapshot capture two days before the deployment.

Answer:

A, E

Explanation:

To update a streaming pipeline with improved business logic and reprocess the previous two days of delivered Pub/Sub messages, you should use Pub/Sub Seek with a timestamp and Pub/Sub Snapshot capture two days before the deployment. Pub/Sub Seek allows you to replay or purge messages in a subscription based on a time or a snapshot. Pub/Sub Snapshot allows you to capture the state of a subscription at a given point in time and replay messages from that point. By using these features, you can ensure that the updated pipeline can process the messages that were delivered in the past two days without losing any data.Reference:

Pub/Sub Seek

Pub/Sub Snapshot

Question 6

Question Type: MultipleChoice

You are running your BigQuery project in the on-demand billing model and are executing a change data capture (CDC) process that ingests data. The CDC process loads 1 GB of data every 10 minutes into a temporary table, and then performs a merge into a 10 TB target table. This process is very scan intensive and you want to explore options to enable a predictable cost model. You need to create a BigQuery reservation based on utilization information gathered from BigQuery Monitoring and apply the reservation to the CDC process. What should you do?

Options:

A- Create a BigQuery reservation for the job.
B- Create a BigQuery reservation for the service account running the job.
C- Create a BigQuery reservation for the dataset.
D- Create a BigQuery reservation for the project.

Answer:

D

Explanation:

https://cloud.google.com/blog/products/data-analytics/manage-bigquery-costs-with-custom-quotas.

Here's why creating a BigQuery reservation for the project is the most suitable solution:

Project-Level Reservation:BigQuery reservations are applied at the project level. This means that the reserved slots (processing capacity) are shared across all jobs and queries running within that project. Since your CDC process is a significant contributor to your BigQuery usage, reserving slots for the entire project ensures that your CDC process always has access to the necessary resources, regardless of other activities in the project.

Predictable Cost Model:Reservations provide a fixed, predictable cost model. Instead of paying the on-demand price for each query, you pay a fixed monthly fee for the reserved slots. This eliminates the variability of costs associated with on-demand billing, making it easier to budget and forecast your BigQuery expenses.

BigQuery Monitoring:You can use BigQuery Monitoring to analyze the historical usage patterns of your CDC process and other queries within your project. This informationhelps you determine the appropriate amount of slots to reserve, ensuring that you have enough capacity to handle your workload while optimizing costs.

Why other options are not suitable:

A . Create a BigQuery reservation for the job:BigQuery does not support reservations at the individual job level. Reservations are applied at the project or assignment level.

B . Create a BigQuery reservation for the service account running the job:While you can create reservations for assignments (groups of users or service accounts), it's less efficient than a project-level reservation in this scenario. A project-level reservation covers all jobs within the project, regardless of the service account used.

C . Create a BigQuery reservation for the dataset:BigQuery does not support reservations at the dataset level.

By creating a BigQuery reservation for your project based on your utilization analysis, you can achieve a predictable cost model while ensuring that your CDC process and other queries have the necessary resources to run smoothly.

Question 7

Question Type: MultipleChoice

You are creating a new pipeline in Google Cloud to stream IoT data from Cloud Pub/Sub through Cloud Dataflow to BigQuery. While previewing the data, you notice that roughly 2% of the data appears to be corrupt. You need to modify the Cloud Dataflow pipeline to filter out this corrupt data. What should you do?

Options:

A- Add a SideInput that returns a Boolean if the element is corrupt.
B- Add a ParDo transform in Cloud Dataflow to discard corrupt elements.
C- Add a Partition transform in Cloud Dataflow to separate valid data from corrupt data.
D- Add a GroupByKey transform in Cloud Dataflow to group all of the valid data together and discard the rest.

Answer:

B