Google Professional-Data-Engineer Mock Exam

Shared by Talley on 17-06-2026

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

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?

Options:

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

Question 2

Question Type: MultipleChoice

You are developing a software application using Google's Dataflow SDK, and want to use conditional, for loops and other complex programming structures to create a branching pipeline. Which component will be used for the data processing operation?

Options:

A- PCollection
B- Transform
C- Pipeline
D- Sink API

Answer:

B

Explanation:

In Google Cloud, the Dataflow SDK provides a transform component. It is responsible for the data processing operation. You can use conditional, for loops, and other complex programming structure to create a branching pipeline.

Question 3

Question Type: MultipleChoice

You have one BigQuery dataset which includes customers' street addresses. You want to retrieve all occurrences of street addresses from the dataset. What should you do?

Options:

A- Create a deep inspection job on each table in your dataset with Cloud Data Loss Prevention and create an inspection template that includes the STREET_ADDRESS infoType.
B- Create a de-identification job in Cloud Data Loss Prevention and use the masking transformation.
C- Write a SQL query in BigQuery by using REGEXP_CONTAINS on all tables in your dataset to find rows where the word 'street' appears.
D- Create a discovery scan configuration on your organization with Cloud Data Loss Prevention and create an inspection template that
includes the STREET_ADDRESS infoType.

Answer:

A

Explanation:

To retrieve all occurrences of street addresses from a BigQuery dataset, the most effective and comprehensive method is to use Cloud Data Loss Prevention (DLP). Here's why option A is the best choice:

Cloud Data Loss Prevention (DLP):

Cloud DLP is designed to discover, classify, and protect sensitive information. It includes pre-defined infoTypes for various kinds of sensitive data, including street addresses.

Using Cloud DLP ensures thorough and accurate detection of street addresses based on advanced pattern recognition and contextual analysis.

Deep Inspection Job:

A deep inspection job allows you to scan entire tables for sensitive information.

By creating an inspection template that includes the STREET_ADDRESS infoType, you can ensure that all instances of street addresses are detected across your dataset.

Scalability and Accuracy:

Cloud DLP is scalable and can handle large datasets efficiently.

It provides a high level of accuracy in identifying sensitive data, reducing the risk of missing any occurrences.

Steps to Implement:

Set Up Cloud DLP:

Enable the Cloud DLP API in your Google Cloud project.

Create an Inspection Template:

Create an inspection template in Cloud DLP that includes the STREET_ADDRESS infoType.

Run Deep Inspection Jobs:

Create and run a deep inspection job for each table in your dataset using the inspection template.

Review the inspection job results to retrieve all occurrences of street addresses.

Cloud DLP Documentation

Creating Inspection Jobs

Question 4

Question Type: MultipleChoice

Your organization stores highly personal data in BigQuery and needs to comply with strict data privacy regulations. You need to ensure that sensitive data values are rendered unreadable whenever an employee leaves the organization. What should you do?

Options:

A- Use dynamic data masking and revoke viewer permissions when employees leave the

organization.
B- Use column-level access controls with policy tags and revoke viewer permissions when employees leave the organization.
C- Use AEAD functions and delete keys when employees leave the organization.
D- Use customer-managed encryption keys (CMEK) and delete keys when employees leave the organization.

Answer:

C

Explanation:

Comprehensive and Detailed

The core requirement is to make specific data values permanently unreadable, a concept often referred to as cryptographic erasure or 'crypto-shredding.'

Option C is the correct answer because it achieves this at a granular, field-level. BigQuery's AEAD (Authenticated Encryption with Associated Data) functions allow you to encrypt individual field values within your tables using a keyset from Cloud KMS. You can encrypt the sensitive data, store the encrypted value in BigQuery, and when an employee leaves, you simply destroy the specific KMS key version used for their data. Once the key is destroyed, the encrypted data is permanently unrecoverable, thus rendering it unreadable.

Option A and B are incorrect because data masking and column-level access controls are about controlling access to the data, not about making the underlying data itself unreadable. If an administrator with sufficient privileges were to query the table, the data would still be there in its original form. Revoking permissions prevents a specific user from seeing the data, but it doesn't erase it.

Option D is incorrect because Customer-Managed Encryption Keys (CMEK) operate at the entire table level. Deleting a CMEK key would make the entire table unreadable, which is far too broad and disruptive. The goal is to target specific sensitive values, not destroy whole tables.

Reference (Google Cloud Documentation Concepts):

Question 5

Question Type: MultipleChoice

After migrating ETL jobs to run on BigQuery, you need to verify that the output of the migrated jobs is the same as the output of the original. You've loaded a table containing the output of the original job and want to compare the contents with output from the migrated job to show that

they are identical. The tables do not contain a primary key column that would enable you to join them together for comparison.

What should you do?

Options:

A- Select random samples from the tables using the RAND() function and compare the samples.
B- Select random samples from the tables using the HASH() function and compare the samples.
C- Use a Dataproc cluster and the BigQuery Hadoop connector to read the data from each table and calculate a hash from non-timestamp columns of the table after sorting. Compare the hashes of each table.
D- Create stratified random samples using the OVER() function and compare equivalent samples from each table.

Answer:

B

Question 6

Question Type: MultipleChoice

You use BigQuery as your centralized analytics platform. New data is loaded every day, and an ETL pipeline modifies the original data and prepares it for the final users. This ETL pipeline is regularly modified and can generate errors, but sometimes the errors are detected only after 2 weeks. You need to provide a method to recover from these errors, and your backups should be optimized for storage costs. How should you organize your data in BigQuery and store your backups?

Options:

A- Organize your data in a single table, export, and compress and store the BigQuery data in Cloud Storage.
B- Organize your data in separate tables for each month, and export, compress, and store the data in Cloud Storage.
C- Organize your data in separate tables for each month, and duplicate your data on a separate dataset in BigQuery.
D- Organize your data in separate tables for each month, and use snapshot decorators to restore the table to a time prior to the corruption.

Answer:

D

Question 7

Question Type: MultipleChoice

You need to deploy additional dependencies to all of a Cloud Dataproc cluster at startup using an existing initialization action. Company security policies require that Cloud Dataproc nodes do not have access to the Internet so public initialization actions cannot fetch resources. What should you do?

Options:

A- Deploy the Cloud SQL Proxy on the Cloud Dataproc master
B- Use an SSH tunnel to give the Cloud Dataproc cluster access to the Internet
C- Copy all dependencies to a Cloud Storage bucket within your VPC security perimeter
D- Use Resource Manager to add the service account used by the Cloud Dataproc cluster to the Network User role

Answer:

C

Question 8

Question Type: MultipleChoice

You are operating a streaming Cloud Dataflow pipeline. Your engineers have a new version of the pipeline with a different windowing algorithm and triggering strategy. You want to update the running pipeline with the new version. You want to ensure that no data is lost during the update. What should you do?

Options:

A- Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to the existing job name
B- Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to a new unique job name
C- Stop the Cloud Dataflow pipeline with the Cancel option. Create a new Cloud Dataflow job with

the updated code
D- Stop the Cloud Dataflow pipeline with the Drain option. Create a new Cloud Dataflow job with the updated code

Answer:

A

Question 9

Question Type: MultipleChoice

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?

Options:

A- Use Cloud SQL for storage. Add secondary indexes to support query patterns.
B- Use Cloud SQL for storage. Use Cloud Dataflow to transform data to support query patterns.
C- Use Cloud Spanner for storage. Add secondary indexes to support query patterns.
D- Use Cloud Spanner for storage. Use Cloud Dataflow to transform data to support query patterns.

Answer:

D

Question 10

Question Type: MultipleChoice

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

Options:

A- Include multiple time series values within the row key
B- Keep the row keep 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.

P2P exams

P2P exams

To Get Premium Files for Professional-Data-Engineer Visit
https://www.p2pexams.com/products/professional-data-engineer

For More Free Questions Visit
https://www.p2pexams.com/google/pdf/professional-data-engineer