Main menu

Preparing for your Professional Data Engineer Journey

Course · 5 hours · Complete

Course overview

Introduction to the Professional Data Engineer (PDE) Certification

Designing Data Processing Systems

Module Overview

Migrating data from private data centers to Google Cloud for Cymbal Retail

Introduction: Diagnostic questions

Diagnostic Questions Practice Demo

Diagnostic questions

Your study plan

Study plan resources

Knowledge Check

Ingesting and Processing Data

Module Overview

Ingesting and processing data for

Preparing for your Professional Data Engineer Journey > Designing Data Processing Systems

Diagnostic questions

Your score: 80% Passing score: 80% Retake

Congratulations! You passed this assessment.

✓ 1. You are migrating on-premises data to a data warehouse on Google Cloud. This data will be made available to business analysts. Local regulations require that customer information including credit card numbers, phone numbers, and email IDs be captured, but not used in analysis. You need to use a reliable, recommended solution to redact the sensitive data. What should you do?

- Create a regular expression to identify and delete patterns that resemble credit card numbers, phone numbers, and email IDs.
- Use the Cloud Data Loss Prevention API (DLP API) to perform date shifting of any entries with credit card numbers, phone numbers, and email IDs.
- ✓ Use the Cloud Data Loss Prevention API (DLP API) to identify and redact data that matches infoTypes like credit card numbers, phone numbers, and email IDs.
- Delete all columns with a title similar to "credit card," "phone," and "email."

Correct. The Cloud Data Loss Prevention API, part of Sensitive Data Protection, helps you discover, classify, and protect your most sensitive data. There are predefined infoTypes that you can employ to identify and redact specific data types.

✓ 2. Cymbal Retail is migrating its private data centers to Google Cloud. Over many years, hundreds of terabytes of data were accumulated. You currently have a 100 Mbps line and you need to transfer this data reliably before commencing operations on Google Cloud in 45 days. What should you do?

- Store the data in an HTTPS endpoint, and configure Storage Transfer Service to copy the data to Cloud Storage.
- Zip and upload the data to Cloud Storage buckets by using the Google Cloud console.
- Upload the data to Cloud Storage by using gcloud storage.
- ✓ Order a transfer appliance, export the data to it, and ship it to Google.

Correct. For large amounts of data that need to be transferred within a month, a transfer appliance is the right choice.

✓ 3. You have a Dataflow pipeline that runs data processing jobs. You need to identify the parts of the pipeline code that consume the most resources. What should you do?

- ✓ Use Cloud Profiler
- Use Cloud Monitoring
- Use Cloud Audit Logs
- Use Cloud Logging

Correct. Cloud Profiler shows you a flame graph of statistics of the running jobs, which can be used to evaluate resource usage.

✓ 4. Cymbal Retail has a team of business analysts who need to fix and enhance a set of large input data files. For example, duplicates need to be removed, erroneous rows should be deleted, and missing data

should be added. These steps need to be performed on all the present set of files and any files received in the future in a repeatable, automated process. The business analysts are not adept at programming. What should they do?

- ✓ Load the data into Dataprep, explore the data, and edit the transformations as needed.
- ○ Create a Dataflow pipeline with the data fixes you need.
- ○ Create a Dataproc job to perform the data fixes you need.
- ○ Load the data into Google Sheets, explore the data, and fix the data as needed.

Correct. Dataprep lets you load large amounts of data and visually fix it, which would be very convenient for those who are unfamiliar with programming. The data wrangling steps can be captured as a series of transformations that can be reapplied later to future data.

✓ 5. Business analysts in your team need to run analysis on data that was loaded into BigQuery. You need to follow recommended practices and grant permissions. What role should you grant the business analysts?

- ○ storage.objectViewer and bigquery.user
- ✓ bigquery.user and bigquery.dataViewer
- ○ bigquery.dataOwner
- ○ bigquery.resourceViewer and bigquery.dataViewer

Correct. The analysts need to view the data and run queries on it, which are granted by these predefined roles.

✗ 6. Cymbal Retail has acquired another company in Europe. Data access permissions and policies in this new region differ from those in Cymbal Retail's headquarters, which is in North America. You need to define a consistent set of policies for projects in each region that follow recommended practices. What should you do?

- ○ Create a new organization for all projects in Europe and assign policies in each organization that comply with regional laws.
- ○ Create top level folders for each region, and assign policies at the folder level.
- ✗ ~~Implement policies at the resource level that comply with regional laws.~~
- ○ Implement a flat hierarchy, and assign policies to each project according to its region.

Incorrect. Applying policies at the resource level is time-consuming and might be inconsistent.

✓ 7. Your data and applications reside in multiple geographies on Google Cloud. Some regional laws require you to hold your own keys outside of the cloud provider environment, whereas other laws are less restrictive and allow storing keys with the same provider who stores the data. The management of these keys has increased in complexity, and you need a solution that can centrally manage all your keys. What should you do?

- ○ Store keys in Cloud Key Management Service (Cloud KMS), and reduce the number of days for automatic key rotation.
- ✓ Store your keys on a supported external key management partner, and use Cloud External Key Manager (Cloud EKM) to get keys when required.
- ○ Store your keys in Cloud Hardware Security Module (Cloud HSM), and retrieve keys from it when required.
- ○ Enable confidential computing for all your virtual machines.

Correct. With Cloud EKM, you manage access to your externally managed keys that reside outside of Google Cloud. Because you need a single solution that also has to store keys externally, this would be the appropriate option.

✗ 8. Laws in the region where you operate require that files related to all orders made each day are stored immutably for 365 days. The solution that you recommend has to be cost-effective. What should you do?

- ○ Store the data in a Cloud Storage bucket, and set a lifecycle policy to delete the file after 365 days.
- ○ Store the data in a Cloud Storage bucket, and enable object versioning and delete any version older than 365 days.
- ✗ ~~Store the data in a Cloud Storage bucket, enable object versioning, and delete any version greater than 365.~~
- ○ Store the data in a Cloud Storage bucket, and specify a retention period.

Incorrect. Object versioning does not restrict the files from being modified or deleted. Having multiple versions of the objects is also not cost-effective.

✓ 9. You are managing the data for Cymbal Retail, which consists of multiple teams including retail, sales, marketing, and legal. These teams are consuming data from multiple producers including point of sales systems, industry data, orders, and more. Currently, teams that consume data have to repeatedly ask the teams that produce it to verify the most up-to-date data and to clarify other questions about the data, such as source and ownership. This process is unreliable and time-consuming and often leads to repeated escalations. You need to implement a centralized solution that gains a unified view of the organization's data and improves searchability. What should you do?

- ✓ Implement a data mesh with Dataplex and have producers tag data when created.
- ○ Implement a data warehouse by using BigQuery, and create datasets for each team such as retail, sales, marketing.
- ○ Implement a data lake with Cloud Storage, and create buckets for each team such as retail, sales, marketing.
- ○ Implement Looker dashboards that provide views of the data that meet each teams' requirements.

Correct. Dataplex is a data mesh that also includes data cataloging capability with Data Catalog. Consumers of data can search and discover information readily without having to wait for data producers to respond, which reduces the bottlenecks on data analysis.

✓ 10. You are using Dataproc to process a large number of CSV files. The storage option you choose needs to be flexible to serve many worker nodes in multiple clusters. These worker nodes will read the data and also write to it for intermediate storage between processing jobs. What is the recommended storage option on Google Cloud?

- ○ Zonal persistent disks
- ✓ Cloud Storage
- ○ Local SSD
- ○ Cloud SQL

Correct. Cloud Storage is the recommended, centralized storage option for Dataproc. It offers many benefits such as high data availability, no storage management, quick startup, and consistent Identity and Access Management (IAM).