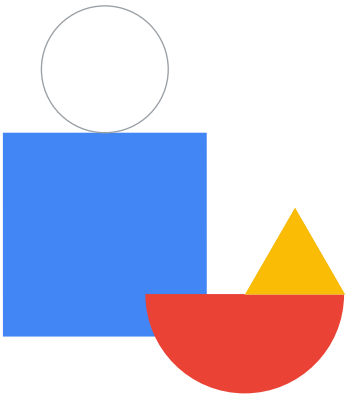
Google Cloud

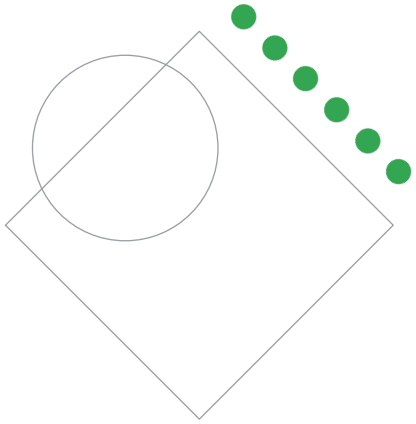


Preparing for Your Professional Data Engineer Journey

Module 1: Designing Data Processing Systems

Welcome to Module 1: Designing Data Processing Systems

Review and study planning



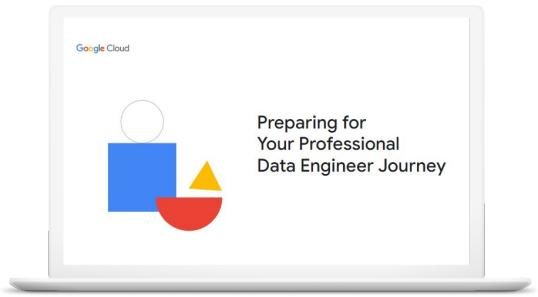
Google Cloud

Now let’s review how to use these diagnostic questions to help you identify what to include in your study plan.

As a reminder - this course isn’t designed to teach you everything you need to know for the exam - and the diagnostic questions don’t cover everything that could be on the exam. Instead, this activity is meant to give you a better sense of the scope of this section and the different skills you’ll want to develop as you prepare for the certification.

Your study plan:

Designing data processing systems



1.1 Designing for security and compliance

1.2 Designing for reliability and fidelity

1.3 Designing for flexibility and portability

1.4 Designing data migrations

Google Cloud

You'll approach this review by looking at the objectives of this exam section and the questions you just answered about each one. Let's introduce an objective, briefly review the answers to the related questions, then explain where you can find out more in the learning resources and/or in Google documentation. As you go through each section objective, use the page in your workbook to mark the specific documentation, courses, and skill badges you'll want to emphasize in your study plan.

1.1 Designing for security and compliance

Considerations include:

- Identity and Access Management (e.g., IAM and organization policies)
- Data security (encryption and key management)
- Privacy (e.g., personally identifiable information, and Cloud Data Loss Prevention API)
- Regional considerations (data sovereignty) for data access and storage
- Legal and regulatory compliance

Google Cloud

Although a Professional Data Engineer's primary responsibility is to ingest, process, store and analyze data, there are other important aspects to consider. You need to provide granular access to data so that users only get what they need to complete the job. When you are storing sensitive data, it should be encrypted so that even if an unauthorized person or entity gains access to the data, they will not be able to read it. As a Professional Data Engineer, you should also be able to redact personally identifiable information (PII) from your data. In addition, you need to understand various legal and regulatory requirements regarding data storage and data access.

- Question 1 tested your familiarity with how IAM roles and permissions are assigned to users and applications in Google Cloud.
- Question 2 asked you to describe the Google Cloud resource hierarchy and inheritance of permissions.
- Question 3 tested your knowledge of how to use the Cloud Data Loss Prevention API to identify, classify, and protect sensitive data.
- Question 4 tested your ability to differentiate between data encryption and key management options in Google Cloud.

1.1 Diagnostic Question 01 Discussion

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?

A. bigquery.resourceViewer and bigquery.dataViewer
B. bigquery.user and bigquery.dataViewer
C. bigquery.dataOwner
D. storage.objectViewer and bigquery.user

Google Cloud

Feedback:

A. Incorrect. The resourceViewer role grants permissions to view capacity and reservations, which are not sufficient to conduct analysis.
B. Correct. The analysts need to view the data and run queries on it, which are granted by these predefined roles.
C. Incorrect. The dataOwner role has more permissions that the business analysts need.
D. Incorrect. The storage.objectViewer role is useful when running federated queries with Cloud Storage, but not for this use case. The business analysts will also require permissions in the bigquery.dataViewer role.

Links:

https://cloud.google.com/architecture/help-secure-the-pipeline-from-your-data-lake-to-your-data-warehouse#business_analyst
https://cloud.google.com/iam/docs/understanding-roles#bigquery-roles

More information:

Courses:
Modernizing Data Lakes and Data Warehouses on Google Cloud

- Building a Data Lake
- Building a Data Warehouse

Smart Analytics, Machine Learning, and AI on Google Cloud

- Prebuilt ML Model APIs for Unstructured Data

Serverless Data Processing with Dataflow: Foundations
- IAM, Quotas, and Permissions
- Security

Skill Badges:
Implement Load Balancing on Compute Engine
Prepare Data for ML APIs on Google Cloud

Summary:
The principle of least privilege says that no more permissions ought to be granted than is required to do the job. So it's important to understand the scope of the job role and how it maps to roles within Identity and Access Management (IAM). Ideally, choose predefined roles, but if they do not provide the required set of permissions, create a custom role.

1.1 Diagnostic Question 02 Discussion

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?

A. Create a new organization for all projects in Europe and assign policies in each organization that comply with regional laws.

B. Implement a flat hierarchy, and assign policies to each project according to its region.

C. Create top level folders for each region, and assign policies at the folder level.

D. Implement policies at the resource level that comply with regional laws.

Google Cloud

Feedback:

A. Incorrect. Creating a new organization to apply separate policies or to work with a different region is not a recommended practice. This requirement can be managed within the same organization by using folders.

B. Incorrect. A flat hierarchy is not recommended, because policy settings need to be individually applied for each project.

C. Correct. Folders are used to group related projects within an organization. They can also be used to apply consistent policies for projects within the folder.

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

Links:

https://cloud.google.com/resource-manager/docs/creating-managing-folders
https://cloud.google.com/resource-manager/docs/cloud-platform-resource-hierarchy

More information:

Courses:

Modernizing Data Lakes and Data Warehouses on Google Cloud

- Building a Data Lake
- Building a Data Warehouse

BigQuery Fundamentals for Redshift Professionals

- BigQuery and Google Cloud IAM

Summary:
Google Cloud's resource hierarchy, access control, and organizational policies let you configure Identity and Access Management (IAM) settings at a higher level in the hierarchy that will be inherited by child resources. Grouping resources in folders also lets you have consistent policies for all projects and resources within it.

1.1 Diagnostic Question 03 Discussion

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?

A. 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.

B. Delete all columns with a title similar to "credit card," "phone," and "email."

C. Create a regular expression to identify and delete patterns that resemble credit card numbers, phone numbers, and email IDs.

D. 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.

Google Cloud

Feedback:

A. 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.

B. Incorrect. Deleting columns based on title is not a reliable way to protect sensitive data, because columns might contain sensitive data, even if it is not indicated by the title. For example, there could be a comments column with sensitive private information.

C. Incorrect. Determining the regular expressions that match different data types can be more tedious and less accurate than using Sensitive Data Protection.

D. Date shifting the entries does not redact the personal identifiable information (PII) like credit card numbers, phone numbers, and email IDs.

Links:

https://cloud.google.com/security/products/sensitive-data-protection
https://cloud.google.com/sensitive-data-protection/docs/infotypes-reference

More information:

Courses:
Modernizing Data Lakes and Data Warehouses on Google Cloud

- Introduction to Data Engineering

BigQuery Fundamentals for Redshift Professionals

- BigQuery and Google Cloud IAM

Summary:
Regulations often require you to capture different kinds of data, but there are other regulations that have strict requirements on how you handle and use the data. Frequently, the sharing and usage of personally identifiable information (PII) is not allowed. Sensitive Data Protection has over 150 infoTypes that can detect various kinds of data. Based on your business needs, you can then take different kinds of action, such as redacting or masking sensitive data, to comply with regulations.

1.1 Diagnostic Question 04 Discussion

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?

A. Enable confidential computing for all your virtual machines.

B. Store keys in Cloud Key Management Service (Cloud KMS), and reduce the number of days for automatic key rotation.

C. Store your keys in Cloud Hardware Security Module (Cloud HSM), and retrieve keys from it when required.

D. Store your keys on a supported external key management partner, and use Cloud External Key Manager (Cloud EKM) to get keys when required.

Google Cloud

Feedback:

A. Incorrect. Confidential computing does not affect the storage of data.

B. Incorrect. Cloud KMS stores keys on Google Cloud; therefore, this won't meet the requirement for the data and the keys that will be stored in different provider environments.

C. Incorrect. Cloud HSM is a Google Cloud service that stores keys on hardware security modules on Google Cloud. Hence, this won't meet the requirement for the data and the keys that will be stored in different provider environments.

D. 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.

Links:

Cloud KMS - https://cloud.google.com/kms/docs/key-management-service
Cloud HSM - https://cloud.google.com/kms/docs/hsm
Cloud EKM - https://cloud.google.com/kms/docs/ekm

More information:

Lab:
Getting Started with Cloud KMS

Summary:

Google Cloud provides a variety of options for storing secure keys for controlled

access to assets. These range from storing keys directly in the cloud using Cloud KMS, storing keys in hardware security modules using Cloud HSM, and storing keys outside of Google Cloud as externally managed keys (Cloud EKM).

1.1 Designing for security and compliance

Courses

[Modernizing Data Lakes and Data Warehouses with Google Cloud](#)
- Introduction to Data Engineering
- Building a Data Lake
- Building a Data Warehouse

[Smart Analytics, Machine Learning, and AI on Google Cloud](#)
- Prebuilt ML Model APIs for Unstructured Data

[Serverless Data Processing with Dataflow: Foundations](#)
- IAM, Quotas, and Permissions
- Security

[BigQuery Fundamentals for Redshift Professionals](#)
- BigQuery and Google Cloud IAM

Skill Badges

[Implement Load Balancing on Compute Engine](#)

[Prepare Data for ML APIs on Google Cloud](#)

Documentation

[Import data from Google Cloud into a secured BigQuery data warehouse](#)
[IAM basic and predefined roles reference](#)
[Creating and managing Folders](#)
[Resource hierarchy](#)
[Sensitive Data Protection](#)
[InfoType detector reference](#)
[Cloud External Key Manager](#)
[Hold your own key with Google Cloud External Key Manager](#)
[Evolving Cloud External Key Manager – What's new with Cloud EKM | Google Cloud Blog](#)

You just reviewed several diagnostic questions that addressed different aspects of designing data processing systems for security and compliance. These are some courses, skill badges, and links to learn more about the concepts in these questions. They provide a starting point to explore Google-recommended practices.

Links:
https://cloud.google.com/architecture/confidential-data-warehouse-blueprint#business_analyst
https://cloud.google.com/iam/docs/understanding-roles#bigquery-roles
https://cloud.google.com/resource-manager/docs/creating-managing-folders
https://cloud.google.com/resource-manager/docs/cloud-platform-resource-hierarchy
https://cloud.google.com/security/products/sensitive-data-protection
https://cloud.google.com/sensitive-data-protection/docs/infotypes-reference
https://cloud.google.com/kms/docs/ekm
https://cloud.google.com/blog/products/identity-security/hold-your-own-key-with-google-cloud-external-key-manager
https://cloud.google.com/blog/products/identity-security/whats-new-with-cloud-ekm

1.2 Designing for reliability and fidelity

Considerations include:

- Preparing and cleaning data (e.g., Dataprep, Dataflow, and Cloud Data Fusion)
- Monitoring and orchestration of data pipelines
- Disaster recovery and fault tolerance
- Making decisions related to ACID (atomicity, consistency, isolation, and durability) compliance and availability
- Data validation

Google Cloud

Data cleaning and data transformation activities are a primary concern for a Professional Data Engineer. Google Cloud offers multiple tools like Dataflow, Cloud Data Fusion, Dataprep etc. to perform data cleaning and data transformation. As a Professional Data Engineer, you should be able to select the most appropriate tool depending on your use case. A Professional Data Engineer often builds complex data pipelines so that the entire data processing activity can be automated. Once data cleaning and data transformation steps are complete, the output data needs to be stored in an appropriate database. A Professional Data Engineers needs to make the selection based on the type of data, i.e. is data structured or unstructured, is it for analytical use case or transactional use case etc.

Question 5 asked you about strategies and options for data preparation and quality control. Question 6 asked about strategies and options for monitoring and ensuring reliability of data pipelines.

1.2 Diagnostic Question 05 Discussion

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?

A. Load the data into Dataprep, explore the data, and edit the transformations as needed.

B. Create a Dataproc job to perform the data fixes you need.

C. Create a Dataflow pipeline with the data fixes you need.

D. Load the data into Google Sheets, explore the data, and fix the data as needed.

Google Cloud

Feedback:

A. 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.

B. Incorrect. Dataproc requires programming knowledge and does not provide a visual data wrangling interface, which is required for this team.

C. Incorrect. Dataflow requires you to write new code, which is not possible for the data analysts.

D. Incorrect. Performing data wrangling in Google Sheets is not an automatable process in the data pipeline; therefore, it's ineffective for this requirement.

Links:

https://help.alteryx.com/dataprep/en/trifacta-application/basics.html
https://help.alteryx.com/dataprep/en/trifacta-application/wrangle-language.html

More information:

Courses:

Building Batch Data Pipelines on Google Cloud

- Introduction to Building Batch Data Pipelines

Serverless Data Processing with Dataflow: Develop Pipelines

- Best Practices

Serverless Data Processing with Dataflow: Operations

- Troubleshooting and Debug

Skill Badges:
Prepare Data for ML APIs on Google Cloud
Engineer Data for Predictive Modeling with BigQuery ML

Summary:
Dataprep provides a visual interface to wrangle data. The data can be visualized as a table with rows and columns. The user can then make edits on the data, which are then captured and rerun on other datasets.

1.2 Diagnostic Question 06 Discussion

You are running a user-supplied DoFn method signature pipeline in Dataflow. The function has been defined by you. The code is running slow and you want to further examine the pipeline code to get better visibility of why.

What should you do?

A. Use Cloud Monitoring
B. Use Cloud Logging
C. Use Cloud Profiler
D. Use Cloud Audit Logs

Google Cloud

Feedback:

A. Incorrect. Cloud Monitoring provides useful diagnostics about Dataflow jobs, but for this requirement, you need the features in Cloud Profiler.
B. Incorrect. Results captured by Cloud Logging are not inherently enough to identify resource usage.
C. Correct. Cloud Profiler shows you a flame graph of statistics of the running jobs, which can be used to evaluate resource usage.
D. Incorrect. Cloud Audit Logs capture administrative events. They do not provide sufficient data to identify Dataflow resource usage.

Links:
https://cloud.google.com/dataflow/docs/guides/profiling-a-pipeline

More information:
Courses:
Modernizing Data Lakes and Data Warehouses on Google Cloud
- Building a Data Warehouse

Building Batch Data Pipelines on Google Cloud
- Manage Data Pipelines with Cloud Data Fusion and Cloud Composer

Building Resilient Streaming Analytics Systems on Google Cloud
- Serverless Messaging with Pub/Sub

Serverless Data Processing with Dataflow: Operations

- Monitoring
- Logging and Error Reporting
- Troubleshooting and Debug
- Testing and CI/CD
- Reliability

Summary:
Google Cloud has multiple tools to monitor and evaluate resources and running workloads. They are strongly integrated into many products. Enable the tools and review the data to identify bottlenecks or errors in your data pipelines.

1.2 Designing for reliability and fidelity

Courses

Modernizing Data Lakes and Data Warehouses on Google Cloud
- Building a Data Warehouse

Building Batch Data Pipelines on Google Cloud
- Introduction to Building Batch Data Pipelines
- Manage Data Pipelines with Cloud Data Fusion and Cloud Composer

Building Resilient Streaming Analytics Systems on Google Cloud
- Serverless Messaging with Pub/Sub

Serverless Data Processing with Dataflow: Develop Pipelines
- Best Practices

Serverless Data Processing with Dataflow: Operations
- Monitoring
- Logging and Error Reporting
- Troubleshooting and Debug
- Testing and CI/CD
- Reliability

Skill Badges

Prepare Data for ML APIs on Google Cloud

Engineer Data for Predictive Modeling with BigQuery ML

Documentation

Dataprep Basics
Dataprep Wrangle Language
Monitoring pipeline performance using Cloud Profiler | Cloud Dataflow

The diagnostic questions we just reviewed explored some aspects of designing for reliability and fidelity. These are some courses and links to learn more about the concepts in these questions. They provide a starting point to explore Google-recommended practices.

Links:
https://help.alteryx.com/dataprep/en/trifacta-application/basics.html
https://help.alteryx.com/dataprep/en/trifacta-application/wrangle-language.html
https://cloud.google.com/dataflow/docs/guides/profiling-a-pipeline

1.3 Designing for flexibility and portability

Considerations include:

- Mapping current and future business requirements to the architecture
- Designing for data and application portability (e.g., multi-cloud and data residency requirements)
- Data staging, cataloging, and discovery (data governance)

Google Cloud

While designing a data processing architecture, a Professional Data Engineer needs to look at the current as well as the future business requirements. The designed architecture needs to be highly scalable as the data is growing at an exponential rate. Many companies use a multi-cloud approach, so a Professional Data Engineer needs to be able to design a data architecture that is highly portable and can seamlessly work on different types of infrastructure.

Question 7 tested your knowledge of how to design data pipelines for portability of data and applications. Question 8 asked you to explain how Google Cloud tools support data cataloging and discovery.

13 Diagnostic Question 07 Discussion

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?

A. Cloud SQL
B. Zonal persistent disks
C. Local SSD
D. Cloud Storage

Google Cloud

Feedback:

A. Incorrect. Cloud SQL is appropriate for transactional data, but it is not the recommended option for Dataproc data processing.
B. Incorrect. Zonal Persistent Disks are not the recommended option for Dataproc storage.
C. Incorrect. Local SSDs are not persistent, so they are not a good storage option when you need to retain intermediate data for longer than the life of the worker node.
D. 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).

Links:
https://cloud.google.com/blog/topics/developers-practitioners/dataproc-best-practices-guide
https://cloud.google.com/blog/products/storage-data-transfer/hdfs-vs-cloud-storage-pros-cons-and-migration-tips

More information:
Courses:
Modernizing Data Lakes and Data Warehouses on Google Cloud

- Building a Data Lake

Serverless Data Processing with Dataflow: Foundations
- Beam Portability

Summary:
When designing data pipelines, a Professional Data Engineer (PDE) should be able to design for portability of data and applications. One aspect of this involves choosing an appropriate storage location for your data.

Cloud Storage is a Hadoop Compatible File System (HCFS) that supports Hadoop and Spark jobs with minimal changes. It is tightly integrated into Google Cloud and supports multiple features such as Identity and Access Management (IAM), redundancy, high availability, and durability, which makes this Google product easy to work with.

13 Diagnostic Question 08 Discussion

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?

A. Implement a data mesh with Dataplex and have producers tag data when created.

B. Implement a data lake with Cloud Storage, and create buckets for each team such as retail, sales, marketing.

C. Implement a data warehouse by using BigQuery, and create datasets for each team such as retail, sales, marketing.

D. Implement Looker dashboards that provide views of the data that meet each teams' requirements.

Google Cloud

Feedback:

A. 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.

B. Incorrect. Cloud Storage is an object store solution. Though we can separate the data in buckets, it does not have rich tagging capabilities.

C. Incorrect. BigQuery can separate the data in datasets, but it is not a centralized data discovery solution that can hold different types of data.

D. Incorrect. Looker provides visualizations of the data based on the data itself. It does not have inherent data tagging facilities.

Links:
https://cloud.google.com/dataplex/docs/introduction

More information:
Courses:
Modernizing Data Lakes and Data Warehouses on Google Cloud

- Introduction to Data Engineering

Building Batch Data Pipelines on Google Cloud

- Introduction to Building Batch Data Pipelines

Skills badge:
Get Started with Dataplex

Summary:
As data volumes increase, it becomes difficult to monitor the data. Also, its lineage, access controls, security, and other metadata can overwhelm data producers and consumers. A data mesh like Dataplex can maintain metadata in a Data Catalog; some of the data can be auto-tagged and some other can be manually tagged, which provides rich information that makes the data readily consumable across the organization.

1.3 Designing for flexibility and portability

Courses

Modernizing Data Lakes and Data Warehouses on Google Cloud
- Introduction to Data Engineering
- Building a Data Lake

Building Batch Data Pipelines on Google Cloud
- Introduction to Building Batch Data Pipelines

Serverless Data Processing with Dataflow: Foundations
- Beam Portability

Skill Badges

Get Started with Dataplex

Documentation

Dataproc best practices | Google Cloud Blog
HDFS vs. Cloud Storage: Pros, cons and migration tips | Google Cloud Blog
Dataplex overview

You just reviewed diagnostic questions that addressed aspects of designing your data processing systems for flexibility and portability. These are some courses, skill badges, and links to learn more about the concepts in these questions. They provide a starting point to explore Google-recommended practices.

Links:
https://cloud.google.com/blog/topics/developers-practitioners/dataproc-best-practices-guide
https://cloud.google.com/blog/products/storage-data-transfer/hdfs-vs-cloud-storage-pros-cons-and-migration-tips
https://cloud.google.com/dataplex/docs/introduction

1.4 Designing data migrations

Considerations include:

- Analyzing current stakeholder needs, users, processes, and technologies and creating a plan to get to desired state
- Planning migration to Google Cloud (e.g., BigQuery Data Transfer Service, Database Migration Service, Transfer Appliance, Google Cloud networking, Datastream)
- Designing the migration validation strategy
- Designing the project, dataset, and table architecture to ensure proper data governance

Google Cloud

A Professional Data Engineer needs to pay attention to various considerations while migrating data from private data centers to Google Cloud. You should be familiar with the multiple tools available to facilitate your data migration to Google Cloud and be able to choose the most appropriate tool depending on the use case. An integral part of data migration involves setting internal standards and policies for gathering, storing, processing, and disposing of the data.

Question 9 asked you to describe how to design Google Cloud projects and data storage to support data governance. Question 10 tested your knowledge of resources and tools that can help you migrate data to Google Cloud.

14 | Diagnostic Question 09 Discussion

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?

A. Store the data in a Cloud Storage bucket, and enable object versioning and delete any version older than 365 days.

B. Store the data in a Cloud Storage bucket, and specify a retention period.

C. Store the data in a Cloud Storage bucket, and set a lifecycle policy to delete the file after 365 days.

D. Store the data in a Cloud Storage bucket, enable object versioning, and delete any version greater than 365.

Google Cloud

Feedback:
A: Incorrect. Object versioning does not restrict the files from being modified or deleted.
B: Correct. Object retention is a built-in option that lets you configure how long the files should remain without allowing any changes.
C: Incorrect. This option automates the deletion of the file after 365 days, but it does not restrict the deletion of the file before that.
D: Incorrect. Object versioning does not restrict the files from being modified or deleted. Having multiple versions of the objects is also not cost-effective.

Links:
https://cloud.google.com/storage/docs/bucket-lock

More information:
Courses:
Modernizing Data Lakes and Data Warehouses on Google Cloud
- Building a Data Lake
- Building a Data Warehouse

BigQuery Fundamentals for Redshift Professionals
- BigQuery and Google Cloud IAM

Summary:

Cloud Storage's built-in capabilities for lifecycle management are simple to configure and ease such important tasks without the need to build, operate, or maintain any custom solutions.

14 | Diagnostic Question 10 Discussion

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?

A. Store the data in an HTTPS endpoint, and configure Storage Transfer Service to copy the data to Cloud Storage.

B. Upload the data to Cloud Storage by using `gcloud storage`.

C. Zip and upload the data to Cloud Storage buckets by using the Google Cloud console.

D. Order a transfer appliance, export the data to it, and ship it to Google.

Google Cloud

Feedback:
A: Incorrect. The bandwidth available is not sufficient to finish the data transfer in the required time.
B: Incorrect. `gcloud storage` is useful for a few files of small to medium sizes. For larger data, other data transfer options should be considered.
C: Incorrect. Uploading large amounts of data through the Google Cloud console is not convenient or performant.
D: Correct. For large amounts of data that need to be transferred within a month, a transfer appliance is the right choice.

Links:
https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#transfer-options

More information:
Courses:
Modernizing Data Lakes and Data Warehouses on Google Cloud

- Building a Data Warehouse

BigQuery Fundamentals for Redshift Professionals

- SQL in BigQuery

Summary:
For data transfers, you need to consider network bandwidth, reliability of the network, and the timeline you need to meet. When the available network bandwidth is low, transfer time will be extended, which might not be viable for the business. In such circumstances, bringing the data into Google Cloud with a transfer appliance is more convenient.

1.4 Designing data migrations

Courses

Modernizing Data Lakes and Data Warehouses on Google Cloud
- Building a Data Lake
- Building a Data Warehouse

BigQuery Fundamentals for Redshift Professionals
- BigQuery and Google Cloud IAM

Documentation

Retention policies and retention policy locks | Cloud Storage

Migration to Google Cloud: Transferring your large datasets

You just reviewed diagnostic questions that addressed different aspects of designing data migrations. These are some courses, skill badges, and links to learn more about the concepts in these questions. They provide a starting point to explore Google-recommended practices.

Links:
https://cloud.google.com/storage/docs/bucket-lock
https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#transfer-options