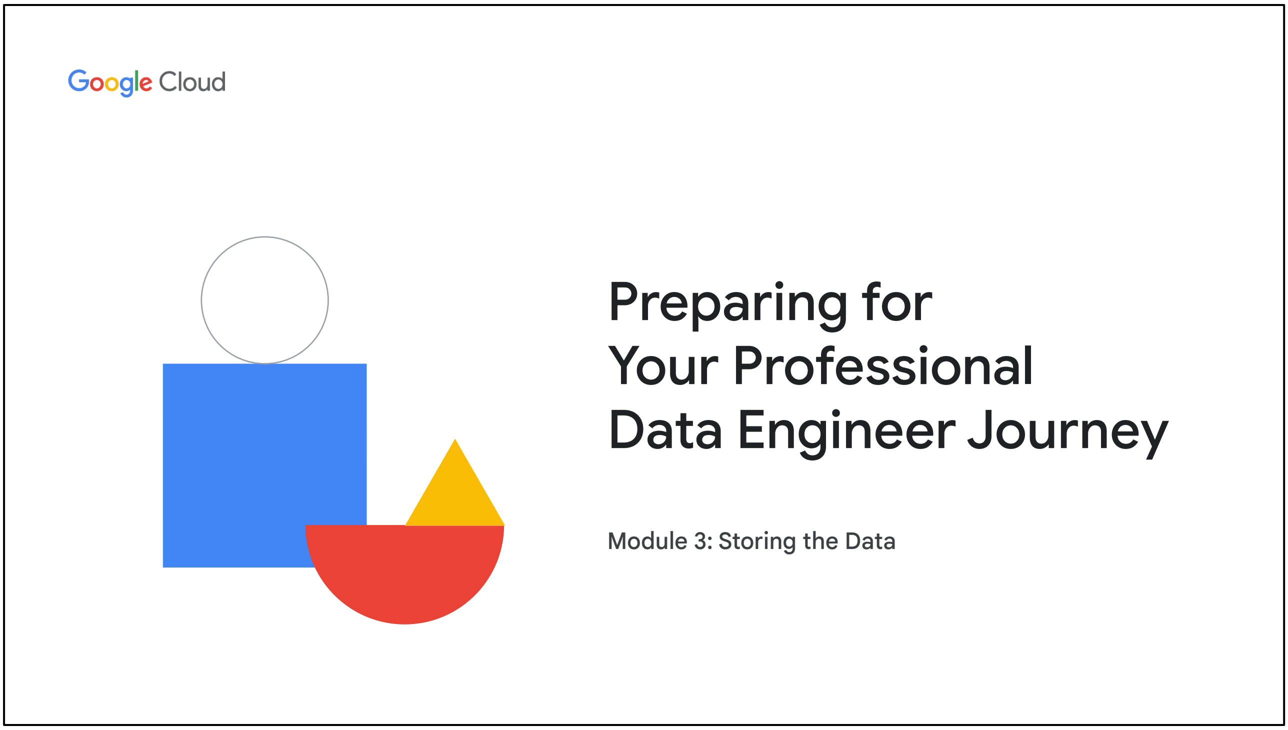
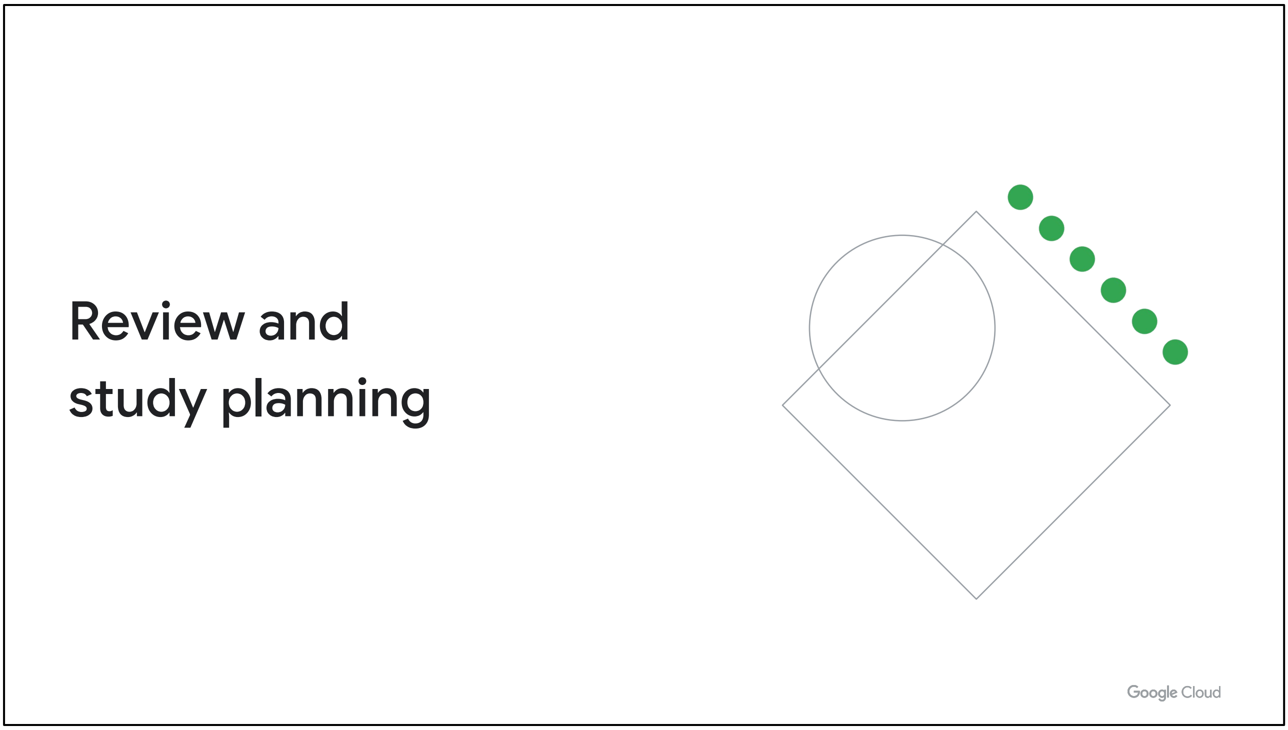


Google Cloud

Preparing for Your Professional Data Engineer Journey

Module 3: Storing the Data

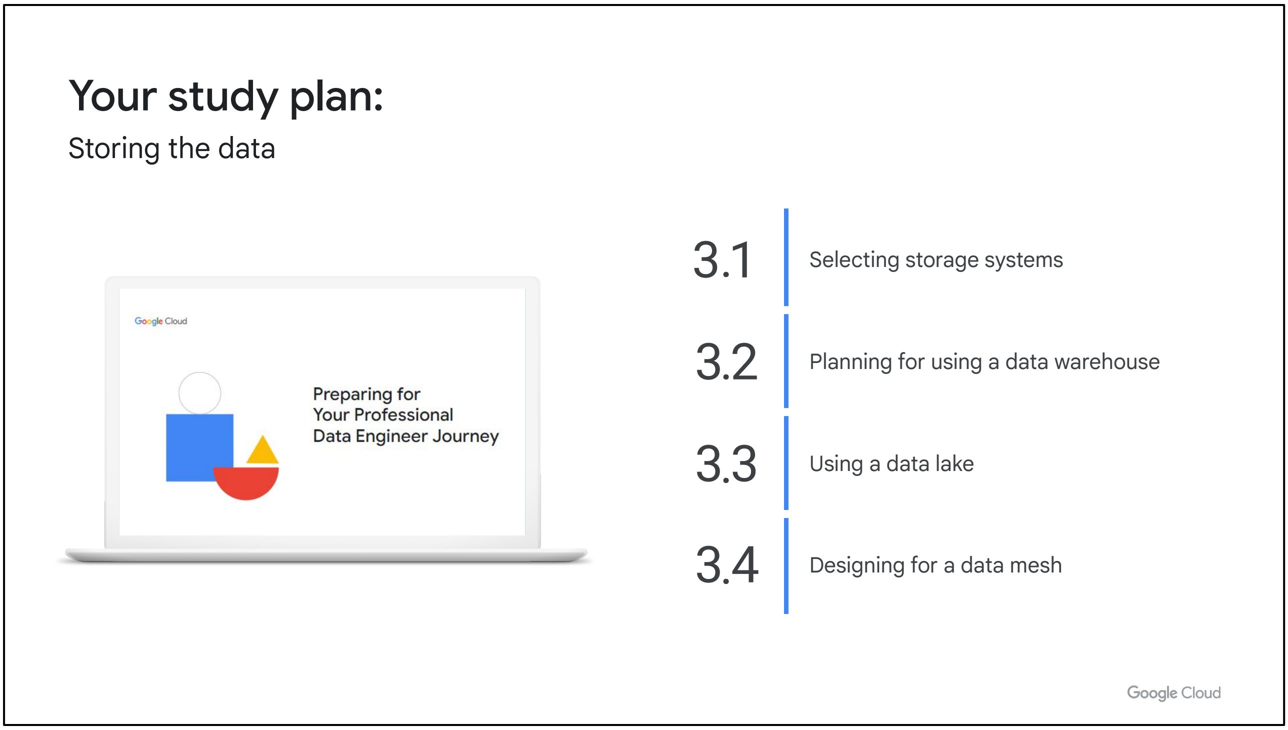
Welcome to Module 3: Storing the Data.



Review and study planning

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.



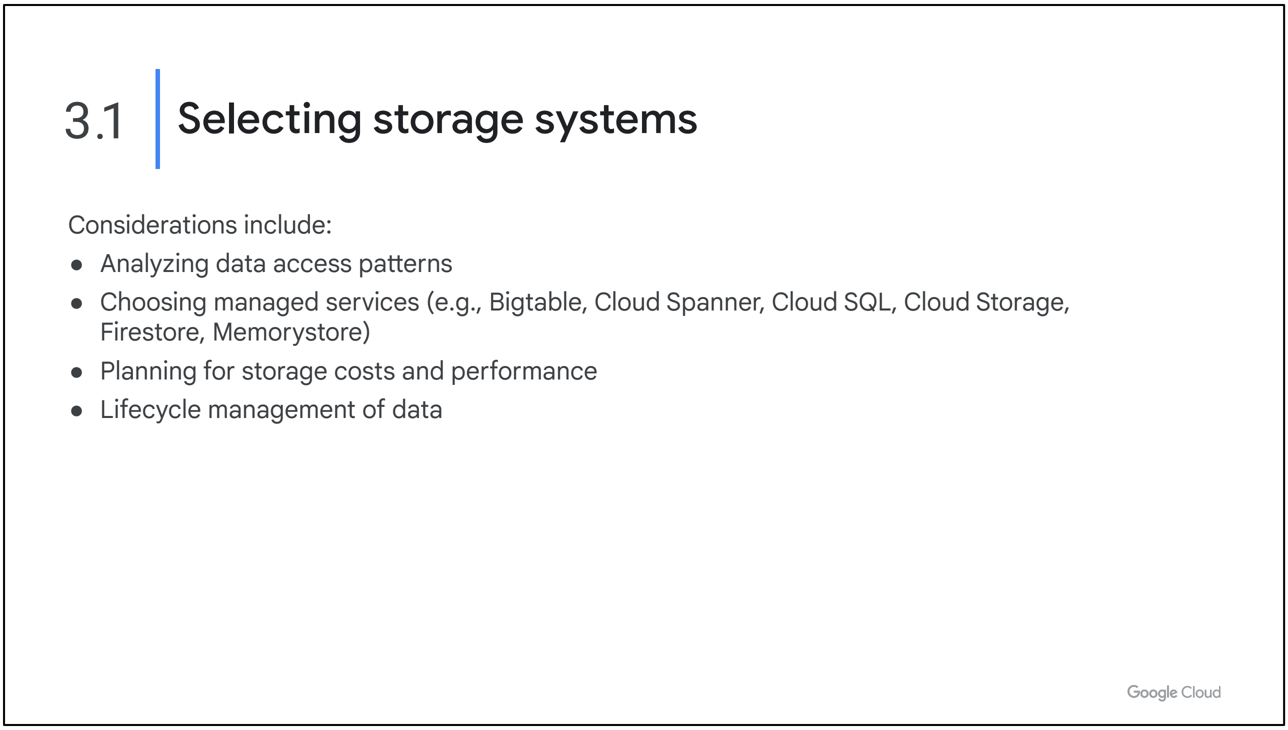
Your study plan:

Storing the data

Preparing for
Your Professional
Data Engineer Journey

3.1	Selecting storage systems
3.2	Planning for using a data warehouse
3.3	Using a data lake
3.4	Designing for a data mesh

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.



3.1 Selecting storage systems

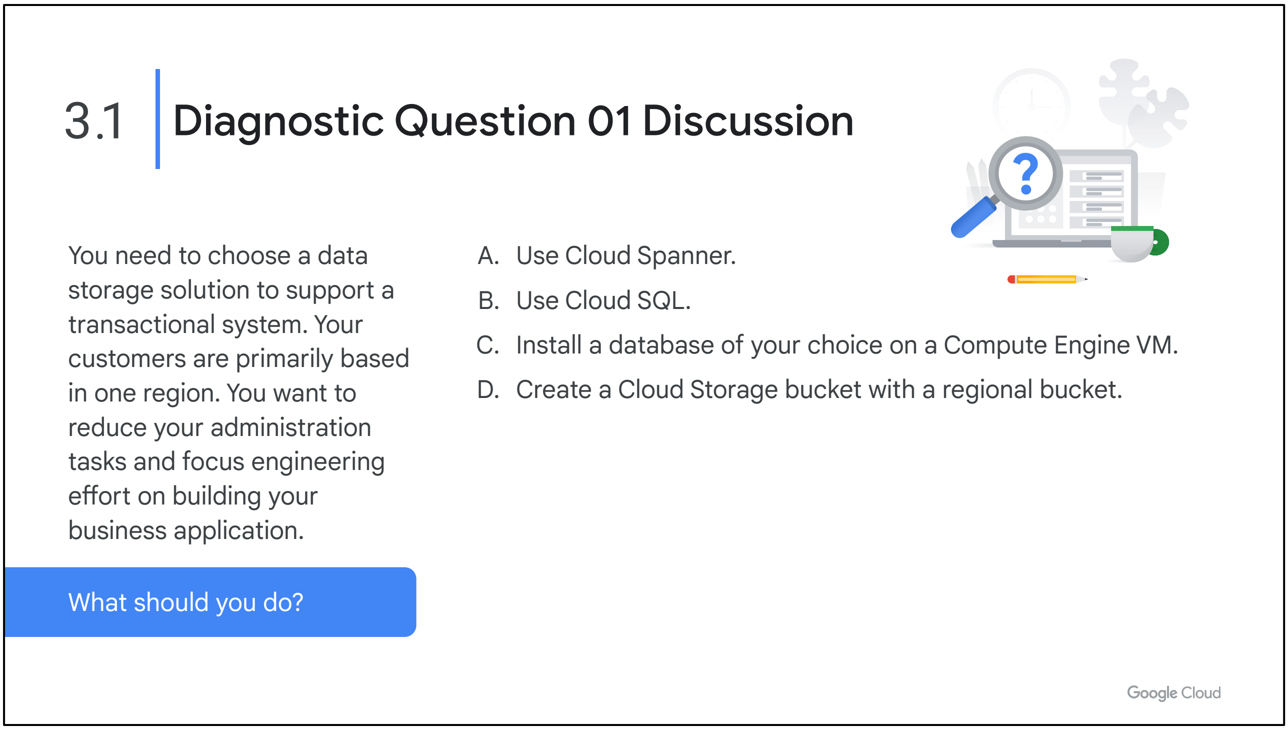
Considerations include:

- Analyzing data access patterns
- Choosing managed services (e.g., Bigtable, Cloud Spanner, Cloud SQL, Cloud Storage, Firestore, Memorystore)
- Planning for storage costs and performance
- Lifecycle management of data

Google Cloud

Google Cloud offers multiple services to store your data. As a Professional Data Engineer, you need to be familiar with all these offerings and choose an appropriate service depending on your use case. For example, do you need a relational SQL database for storing transactional data? Or do you need a centralized repository to store your unstructured data? Along with the type of the data, you also need to consider storage costs and performance implications of each storage service.

Question 1 challenged you to differentiate between managed services for data storage to select the most appropriate option for a given use case. Question 2 tested your ability to determine how to plan for storage costs and performance.



3.1 Diagnostic Question 01 Discussion

You need to choose a data storage solution to support a transactional system. Your customers are primarily based in one region. You want to reduce your administration tasks and focus engineering effort on building your business application.

What should you do?

A. Use Cloud Spanner.

B. Use Cloud SQL.

C. Install a database of your choice on a Compute Engine VM.

D. Create a Cloud Storage bucket with a regional bucket.

Feedback:
A: Incorrect. Cloud Spanner is more suitable for a global, transactional database requirement.

B: Correct. Cloud SQL is a managed service that supports the popular transactional databases MySQL, PostgreSQL, and SQL Server.

C: Incorrect. Installing and maintaining your own database instance is more effort on the engineering team.

D: Incorrect. A Cloud Storage bucket is not a transactional database.

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

More information:
Courses:
Google Cloud Big Data and Machine Learning Fundamentals

- Big Data and Machine Learning on Google Cloud

Modernizing Data Lakes and Data Warehouses on Google Cloud

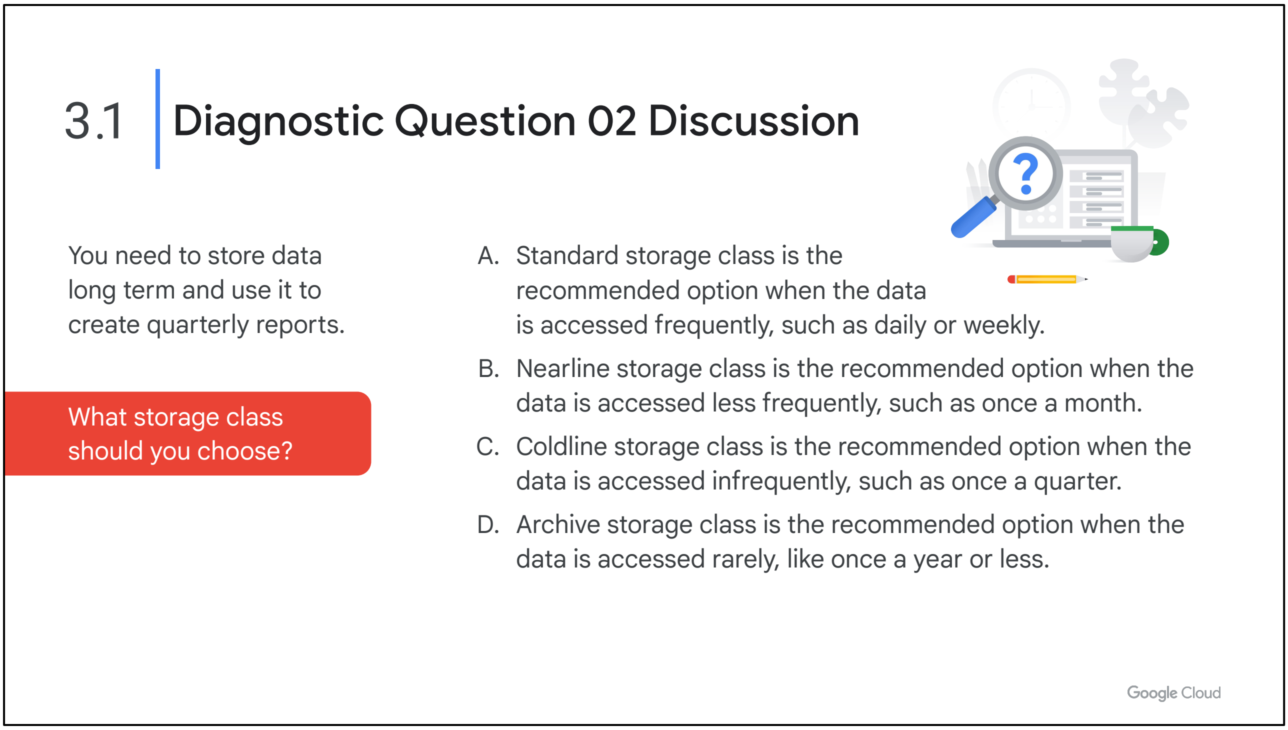
- Introduction to Data Engineering
- Building a Data Lake
- Building a Data Warehouse

Building Resilient Streaming Analytics Systems on Google Cloud

- High-Throughput BigQuery and Bigtable Streaming Features

Summary:

A Professional Data Engineer is required to understand both technical and business requirements. Sometimes, the effort and expense of administering databases is not suitable for the business. You can choose among a variety of managed services on Google Cloud.



3.1

Diagnostic Question 02 Discussion

You need to store data long term and use it to create quarterly reports.

What storage class should you choose?

A. Standard storage class is the recommended option when the data is accessed frequently, such as daily or weekly.

B. Nearline storage class is the recommended option when the data is accessed less frequently, such as once a month.

C. Coldline storage class is the recommended option when the data is accessed infrequently, such as once a quarter.

D. Archive storage class is the recommended option when the data is accessed rarely, like once a year or less.

Google Cloud

Feedback:
A: Incorrect. Cloud Spanner is more suitable for a global, transactional database requirement.

B: Correct. Cloud SQL is a managed service that supports the popular transactional databases MySQL, PostgreSQL, and SQL Server.

C: Incorrect. Installing and maintaining your own database instance is more effort on the engineering team.

D: Incorrect. A Cloud Storage bucket is not a transactional database.

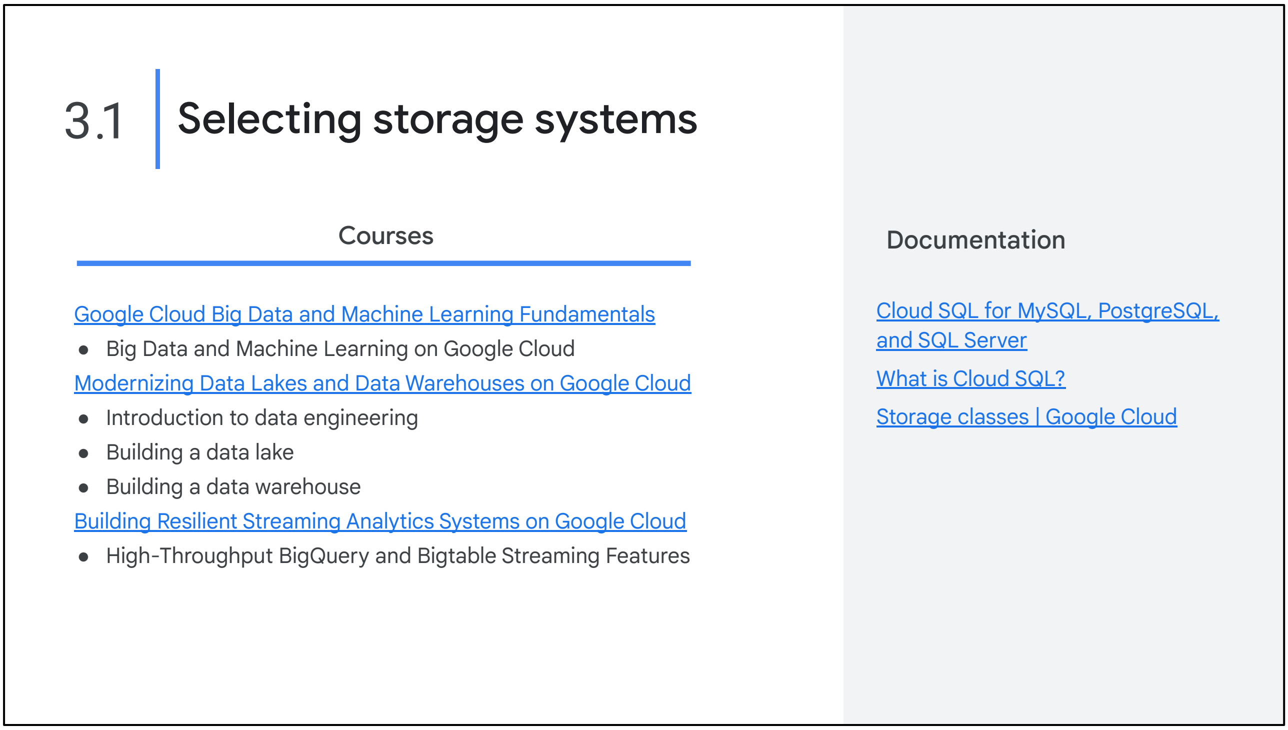
Links:
https://cloud.google.com/storage/docs/storage-classes

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

- Building a Data Lake

Summary:
Google Cloud has different classes of storage that are priced on both the amount of data stored and how frequently it is accessed. A Professional Data Engineer needs to understand the expected data access frequency and plan an appropriate storage

class.



3.1 Selecting storage systems

Courses

Google Cloud Big Data and Machine Learning Fundamentals

- Big Data and Machine Learning on Google Cloud

Modernizing Data Lakes and Data Warehouses on Google Cloud

- Introduction to data engineering
- Building a data lake
- Building a data warehouse

Building Resilient Streaming Analytics Systems on Google Cloud

- High-Throughput BigQuery and Bigtable Streaming Features

Documentation

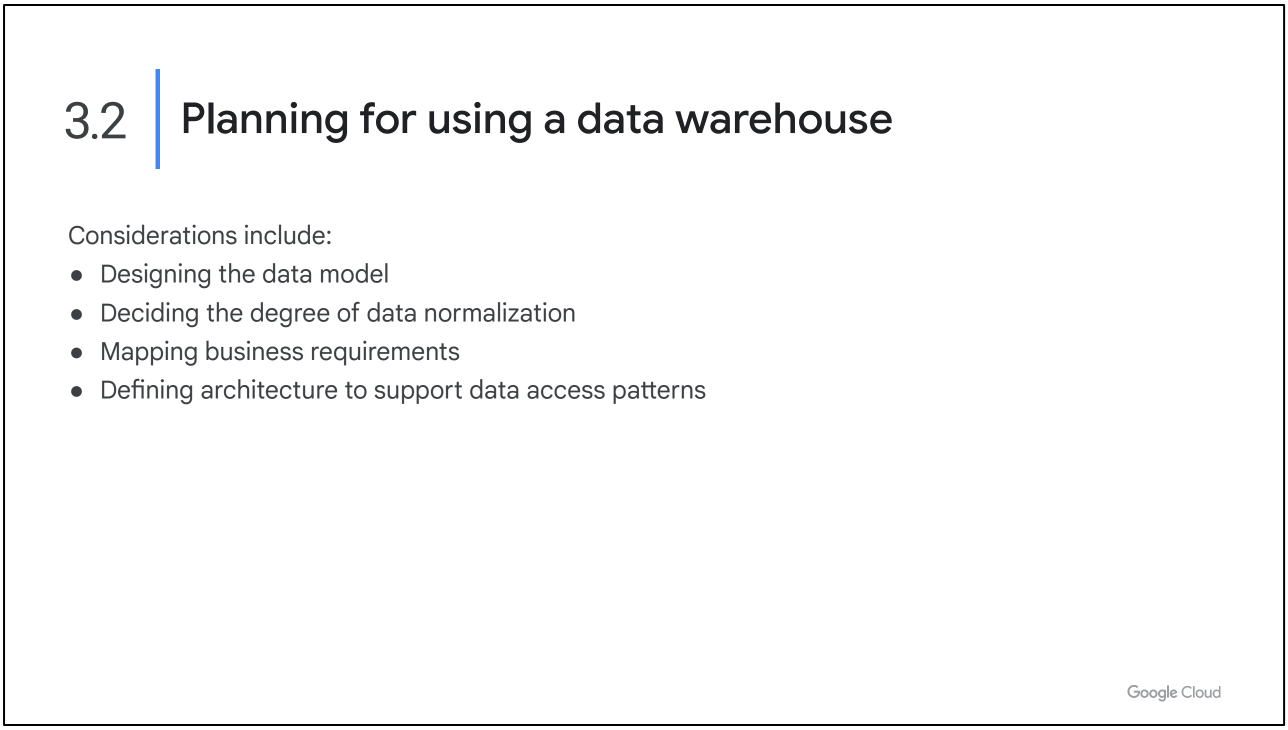
Cloud SQL for MySQL, PostgreSQL, and SQL Server

What is Cloud SQL?

Storage classes | Google Cloud

You just reviewed several diagnostic questions that addressed different aspects of selecting storage systems for your data. 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/sql
https://cloud.google.com/sql/docs/introduction
https://cloud.google.com/storage/docs/storage-classes



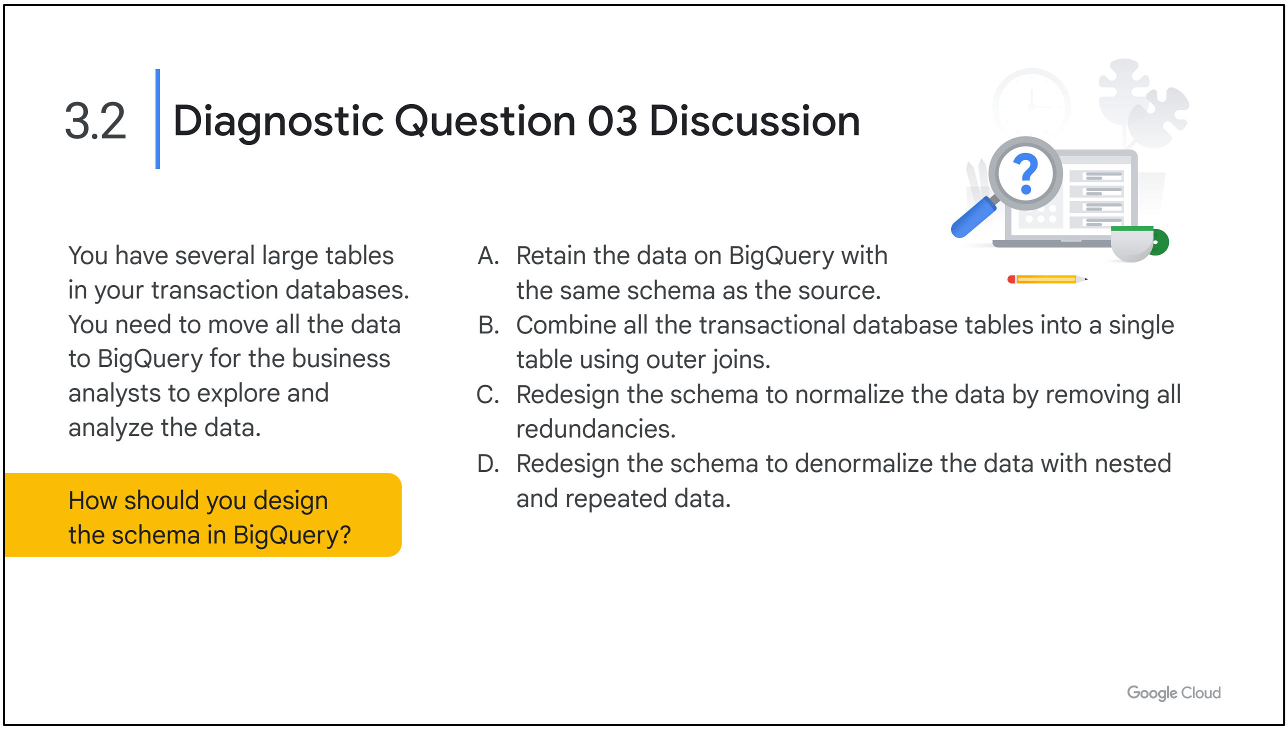
3.2 Planning for using a data warehouse

Considerations include:

- Designing the data model
- Deciding the degree of data normalization
- Mapping business requirements
- Defining architecture to support data access patterns

As a Professional Data Engineer, you will implement a data warehouse to store and analyze your data. Google Cloud offers BigQuery, a serverless and cost-effective enterprise data warehouse that works across clouds and scales with your data. As a Professional Data Engineer, you will build data models, decide the degree of data normalization and finalize the schema. A key skill is to understand the business requirements, map them to the most frequently used data access patterns, and optimize your architecture for the way you use data.

Question 3 tested your understanding of how to design the data model and table schemas for a data warehouse. Question 4 asked you to describe how to map business requirements to a data warehouse design and implementation. Question 5 tested your ability to define a data warehouse architecture to support data access.



3.2 Diagnostic Question 03 Discussion

You have several large tables in your transaction databases. You need to move all the data to BigQuery for the business analysts to explore and analyze the data.

How should you design the schema in BigQuery?

A. Retain the data on BigQuery with the same schema as the source.
B. Combine all the transactional database tables into a single table using outer joins.
C. Redesign the schema to normalize the data by removing all redundancies.
D. Redesign the schema to denormalize the data with nested and repeated data.

Feedback:
A: Incorrect. The normalized form in transactional databases are efficient for writes, but not efficient for running queries against.

B: Incorrect. Combining multiple tables into a single table using only outer joins could include significantly more redundancy than is required. Fewer tables are better on BigQuery, but the schema needs to be thought through.

C: Incorrect. Normalizing the data is not the recommended approach for BigQuery.

D: Correct. Denormalizing the data is the recommended approach. Joining large amounts of data repeatedly during data analysis increases costs.

Links:
https://cloud.google.com/bigquery/docs/best-practices-performance-overview

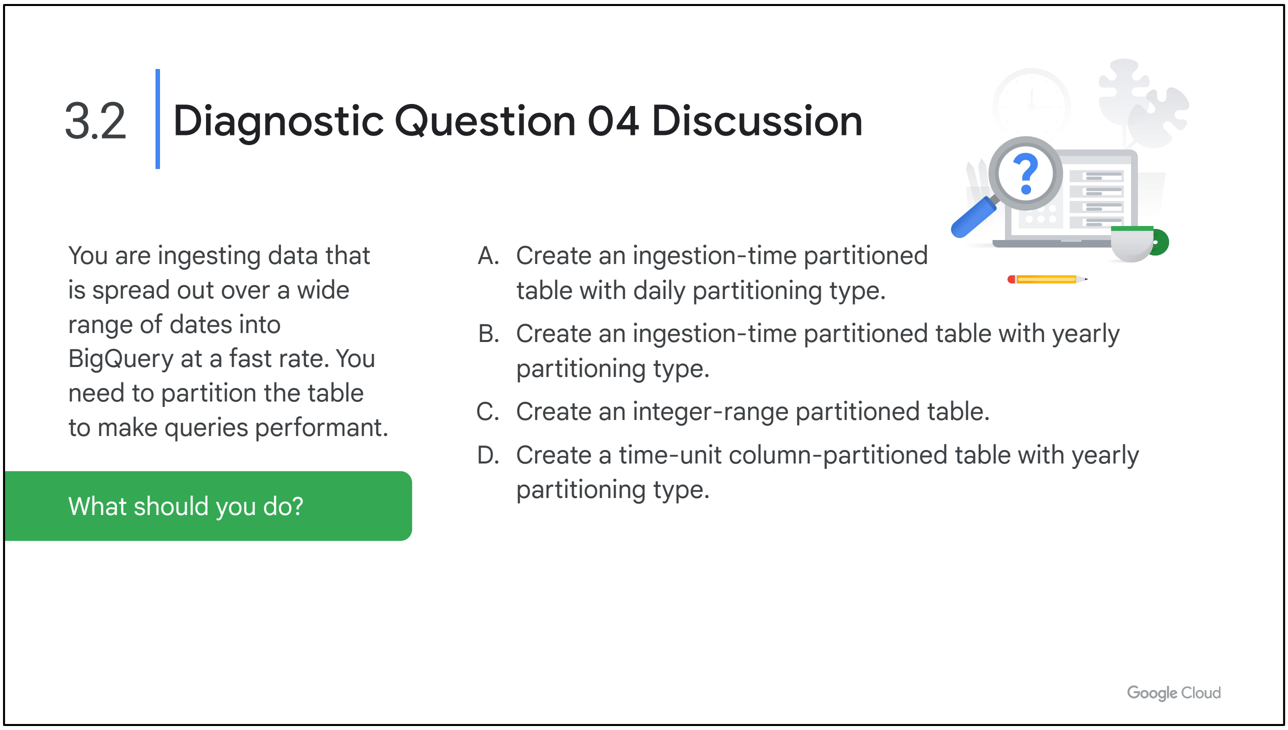
More information:
Courses:
Modernizing Data Lakes and Data Warehouses on Google Cloud

- Building a Data Warehouse

Skill Badges:
Build and Optimize Data Warehouses with BigQuery

Summary:

Analytics oriented databases such as BigQuery have significantly more read requests than write requests. In this scenario, the analytics engine optimizes for query performance at the cost of data redundancy. BigQuery can also use other optimizations like partitioning and clustering to further improve the performance of queries.



3.2 | Diagnostic Question 04 Discussion

You are ingesting data that is spread out over a wide range of dates into BigQuery at a fast rate. You need to partition the table to make queries performant.

What should you do?

A. Create an ingestion-time partitioned table with daily partitioning type.

B. Create an ingestion-time partitioned table with yearly partitioning type.

C. Create an integer-range partitioned table.

D. Create a time-unit column-partitioned table with yearly partitioning type.

Google Cloud

Feedback:
A: Correct. A daily partition type is the most suitable given the volume of the data and the range of dates.

B: Incorrect. A yearly partition type has too much data per partition, which makes queries inefficient.

C: Incorrect. An integer-range partition type is not appropriate given that the data is defined by dates.

D: Incorrect. A yearly partition type has too much data per partition, which makes queries inefficient.

Links:
https://cloud.google.com/bigquery/docs/partitioned-tables#date_timestamp_partitioned_tables
https://cloud.google.com/bigquery/docs/creating-partitioned-tables#create_a_time-unit_column-partitioned_table

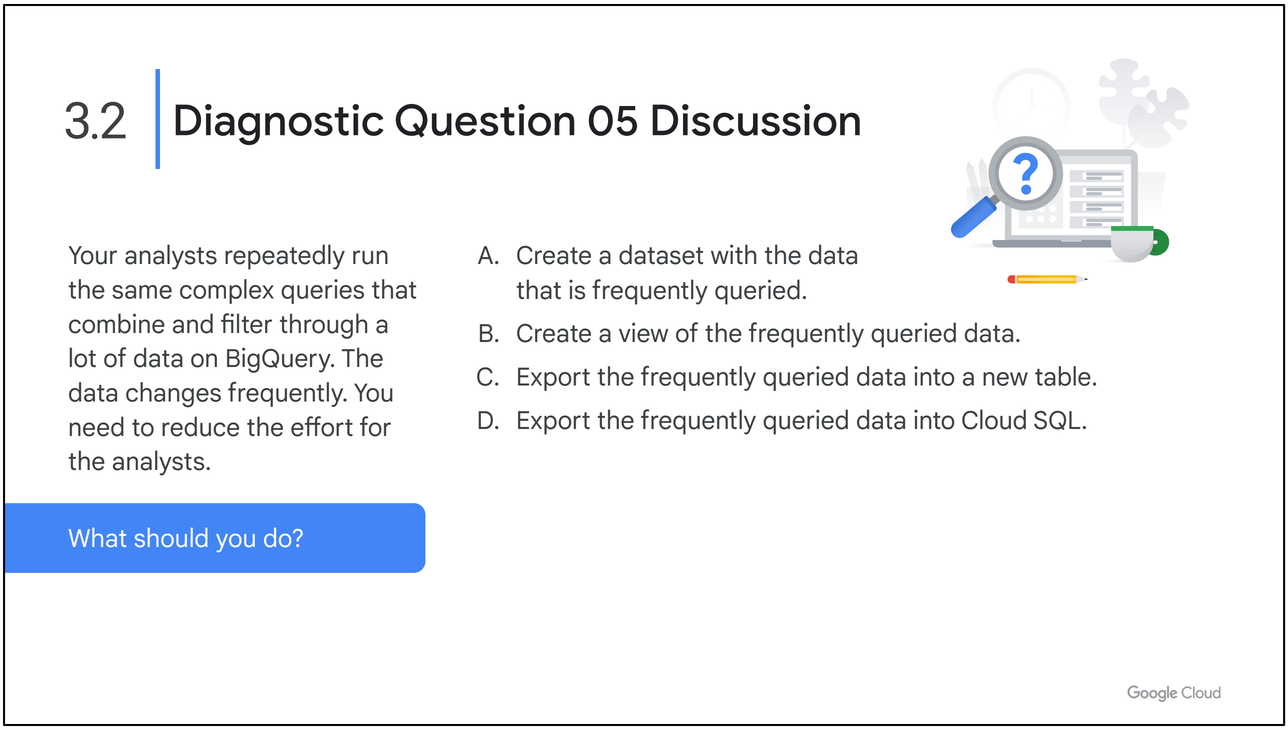
More information:
Courses:
Modernizing Data Lakes and Data Warehouses on Google Cloud
- Building a Data Warehouse

Building Resilient Streaming Analytics Systems on Google Cloud

- Advanced BigQuery Functionality and Performance

Skill Badges:
Build and Optimize Data Warehouses with BigQuery

Summary:
When performance and cost-effectiveness is required on BigQuery, the data engineer can reduce the amount of data queried by partitioning the table. Depending on the type of data and the volume of data, you can choose from many types of partitioning.



3.2 Diagnostic Question 05 Discussion

Your analysts repeatedly run the same complex queries that combine and filter through a lot of data on BigQuery. The data changes frequently. You need to reduce the effort for the analysts.

What should you do?

A. Create a dataset with the data that is frequently queried.

B. Create a view of the frequently queried data.

C. Export the frequently queried data into a new table.

D. Export the frequently queried data into Cloud SQL.

Google Cloud

Feedback:
A: Incorrect. This is not a recommended approach, because there will be redundant data.

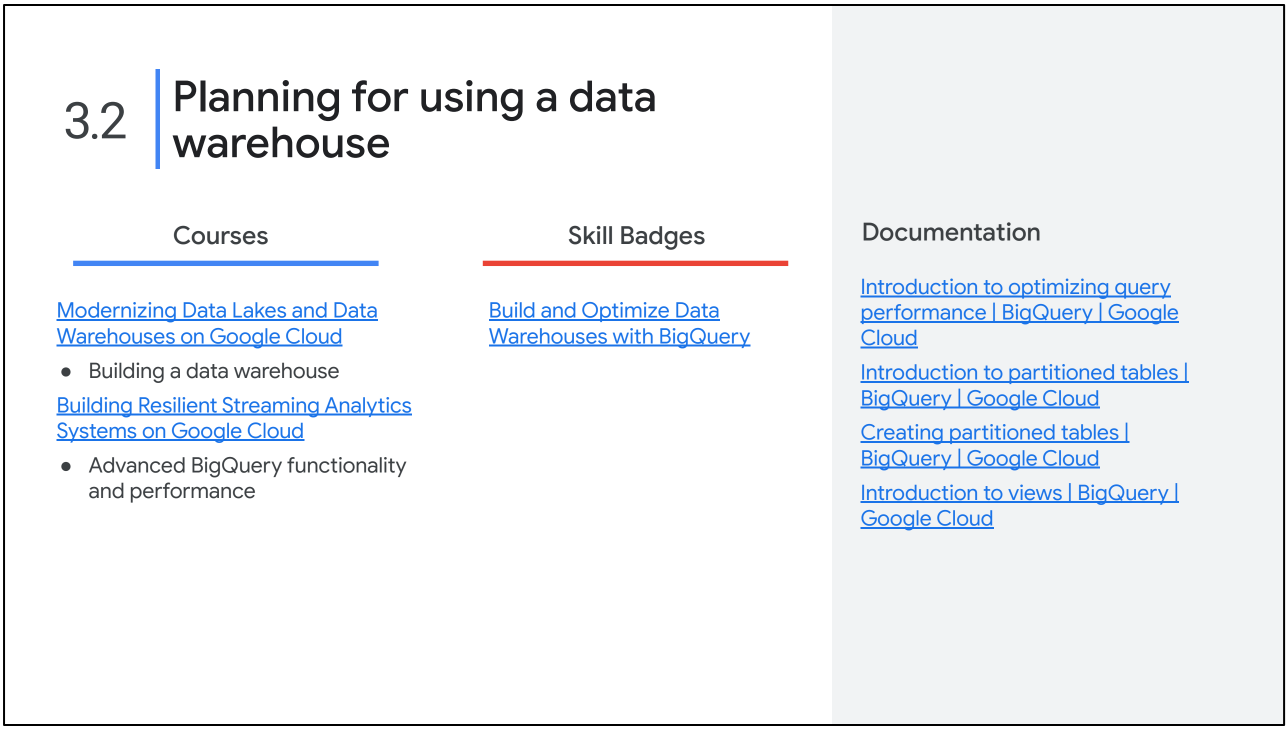
B: Correct. Creating a view will rerun the complex query automatically. Meanwhile, the analysts need only reference the name of the view, which makes it easier for them.

C: Incorrect. This is not a recommended approach, because there will be redundant data.

D: Incorrect. It requires more effort and cost to duplicate the data in another database.

Links:
https://cloud.google.com/bigquery/docs/views-intro

Summary:
A Professional Data Engineer should cater to internal customers and ease the process of working with data for them. BigQuery supports a variety of data definitions, including datasets, tables, views, and materialized views. Defining a view with a query will make it easy for others to reuse the same query without having to know the complex details.



3.2 Planning for using a data warehouse

Courses

Modernizing Data Lakes and Data Warehouses on Google Cloud

- Building a data warehouse

Building Resilient Streaming Analytics Systems on Google Cloud

- Advanced BigQuery functionality and performance

Skill Badges

Build and Optimize Data Warehouses with BigQuery

Documentation

Introduction to optimizing query performance | BigQuery | Google Cloud

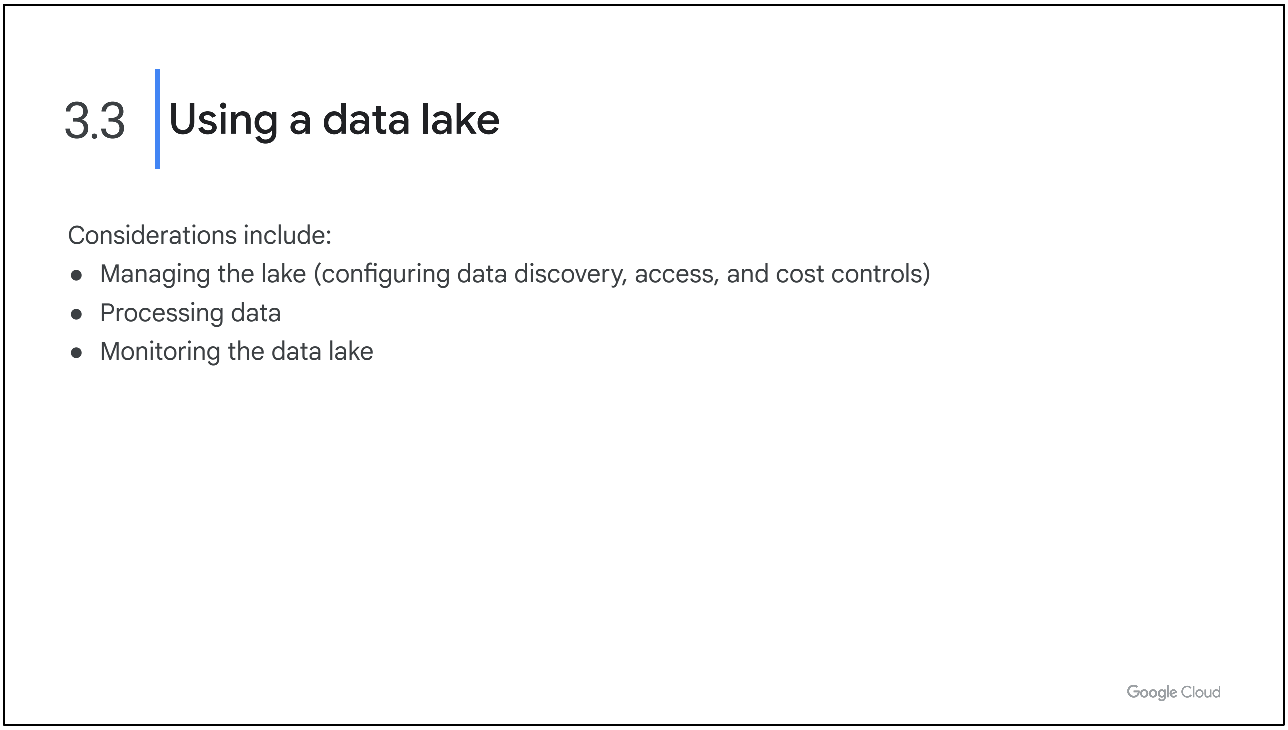
Introduction to partitioned tables | BigQuery | Google Cloud

Creating partitioned tables | BigQuery | Google Cloud

Introduction to views | BigQuery | Google Cloud

The diagnostic questions you just reviewed explored some aspects of planning for using a data warehouse. 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://cloud.google.com/bigquery/docs/best-practices-performance-overview
https://cloud.google.com/bigquery/docs/partitioned-tables#date_timestamp_partitioned_tables
https://cloud.google.com/bigquery/docs/creating-partitioned-tables#create_a_time-unit_column-partitioned_table
https://cloud.google.com/bigquery/docs/views-intro



3.3 Using a data lake

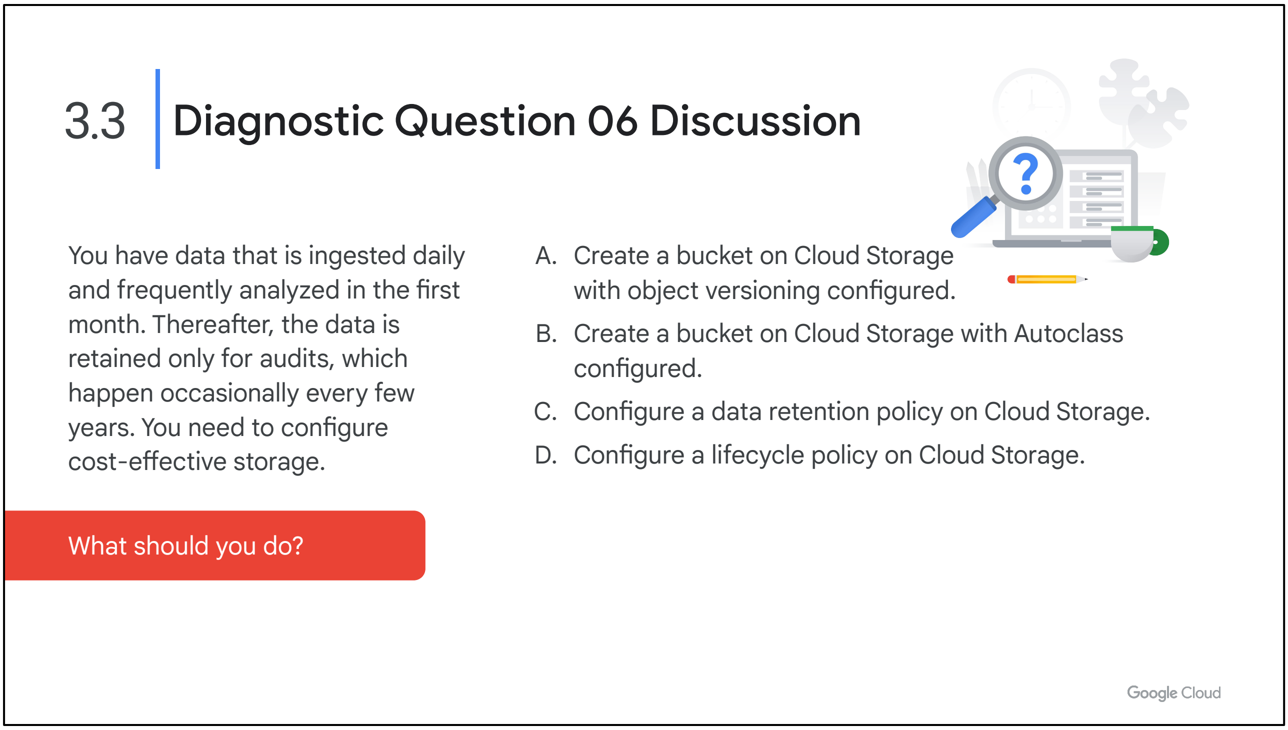
Considerations include:

- Managing the lake (configuring data discovery, access, and cost controls)
- Processing data
- Monitoring the data lake

Google Cloud

As a Professional Data Engineer, you will implement a data lake for your organization. You will be responsible for guaranteeing access performance, reducing storage costs, ensuring data availability, and securing the data.

Question 6 asked you to describe how to design and manage a data lake to control cost. Question 7 tested your knowledge of how to manage a data lake for data discovery and security. Question 8 tested your knowledge of options for monitoring a data lake.



3.3 Diagnostic Question 06 Discussion

You have data that is ingested daily and frequently analyzed in the first month. Thereafter, the data is retained only for audits, which happen occasionally every few years. You need to configure cost-effective storage.

What should you do?

A. Create a bucket on Cloud Storage with object versioning configured.

B. Create a bucket on Cloud Storage with Autoclass configured.

C. Configure a data retention policy on Cloud Storage.

D. Configure a lifecycle policy on Cloud Storage.

Google Cloud

Feedback:
A: Incorrect. Object versioning retains versions of the object, but it will add to costs.

B: Incorrect. An Autoclass configuration will automatically move data to different classes, but you will not have granular control over when it happens and to which class.

C: Incorrect. Retention policies restrict changing the object, but do not reduce cost.

D: Correct. A lifecycle policy can be configured to automatically move the objects between different storage classes on schedules that you determine.

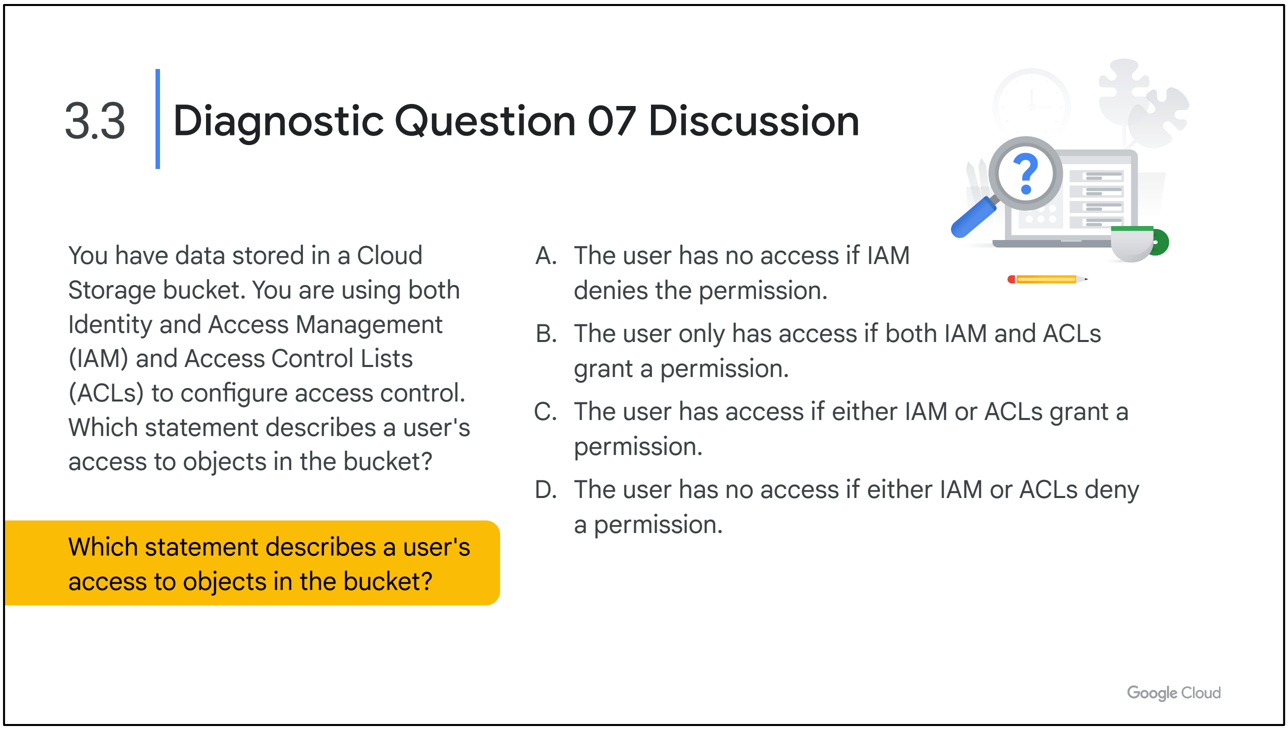
Links:
https://cloud.google.com/storage
https://cloud.google.com/storage/docs/lifecycle

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

- Building a Data Lake

Summary:
Organizations ingest large amounts of data. Storage incurs a cost, and yet, for various requirements like audits and regulations, we have to store it. A Professional

Data Engineer should look for cost-effective ways to store data depending on the expected usage frequency.



3.3 Diagnostic Question 07 Discussion

You have data stored in a Cloud Storage bucket. You are using both Identity and Access Management (IAM) and Access Control Lists (ACLs) to configure access control. Which statement describes a user's access to objects in the bucket?

Which statement describes a user's access to objects in the bucket?

A. The user has no access if IAM denies the permission.

B. The user only has access if both IAM and ACLs grant a permission.

C. The user has access if either IAM or ACLs grant a permission.

D. The user has no access if either IAM or ACLs deny a permission.

Feedback:
A: Incorrect. ACLs can override IAM permissions.

B: Incorrect. Permissions need not be congruently set in both; either can give access.

C: Correct. IAM and ACLs work in parallel. If either of the systems grants the user access permission, the user will have access.

D: Incorrect. Either system can override the permission set by the other.

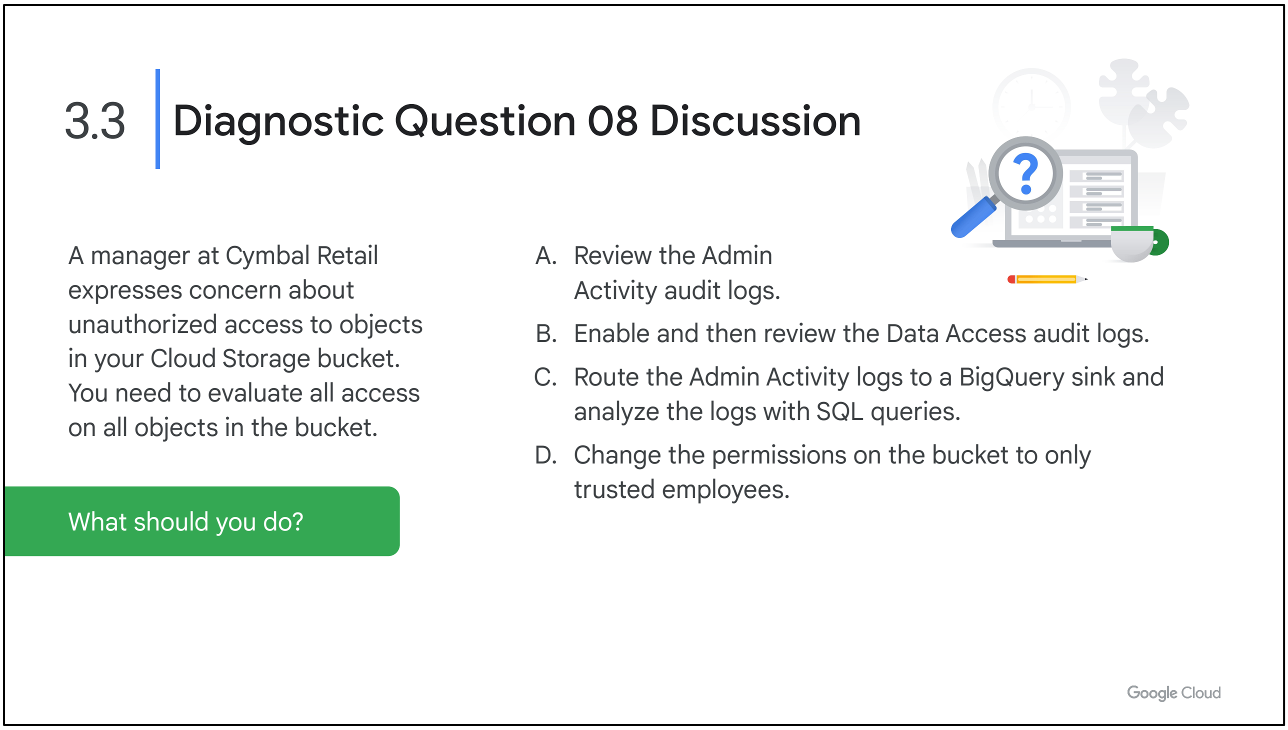
Links:
https://cloud.google.com/storage/docs/access-control

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

- Building a Data Lake

Summary:
As a Professional Data Engineer responsible for the data and its security, Google Cloud gives you the power to configure granular control over access permissions. You have options such as IAM, ACL, signed URLs, and signed policy documents. IAM gives you bucket-level control, and ACLs give you object-level control. You can use

either to configure access.



3.3 Diagnostic Question 08 Discussion

A manager at Cymbal Retail expresses concern about unauthorized access to objects in your Cloud Storage bucket. You need to evaluate all access on all objects in the bucket.

What should you do?

A. Review the Admin Activity audit logs.

B. Enable and then review the Data Access audit logs.

C. Route the Admin Activity logs to a BigQuery sink and analyze the logs with SQL queries.

D. Change the permissions on the bucket to only trusted employees.

Feedback:
A: Incorrect. Admin Activity audit logs do not show who is reading and writing the objects.

B: Correct. Data Access audit logs have to be specifically enabled first, because they could generate a lot of logs for all reads and writes.

C: Incorrect. Routing Admin Activity audit logs to BigQuery for analysis will not help because they will not show who is reading and writing the objects.

D: Incorrect. Changing permissions will restrict usage; however, you won't be able to discover who is currently accessing the objects.

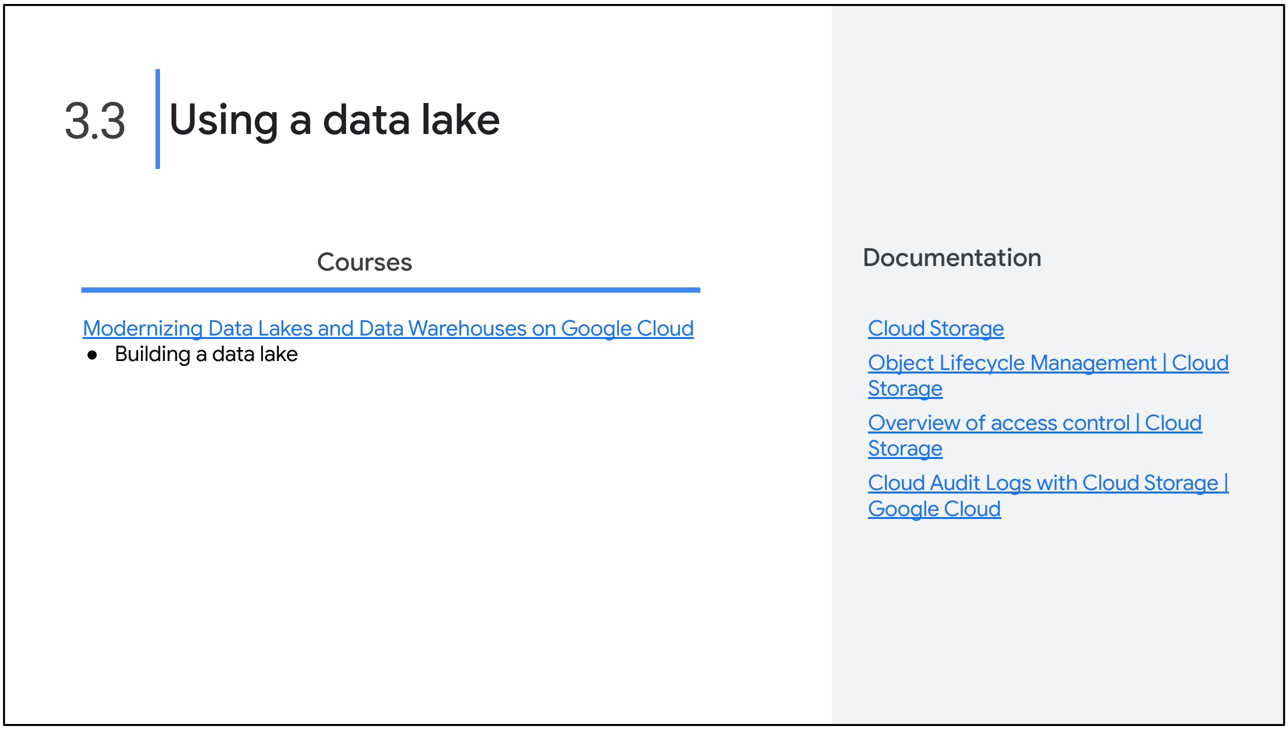
Links:
https://cloud.google.com/storage/docs/audit-logging

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

- Building a Data Lake

Summary:
Logs help a Professional Data Engineer analyze what happened: who accessed which documents and when. Admin Activity audit logs are always on, whereas Data

Access audit logs have to be enabled.



3.3 Using a data lake

Courses

[Modernizing Data Lakes and Data Warehouses on Google Cloud]

- Building a data lake

Documentation

Cloud Storage

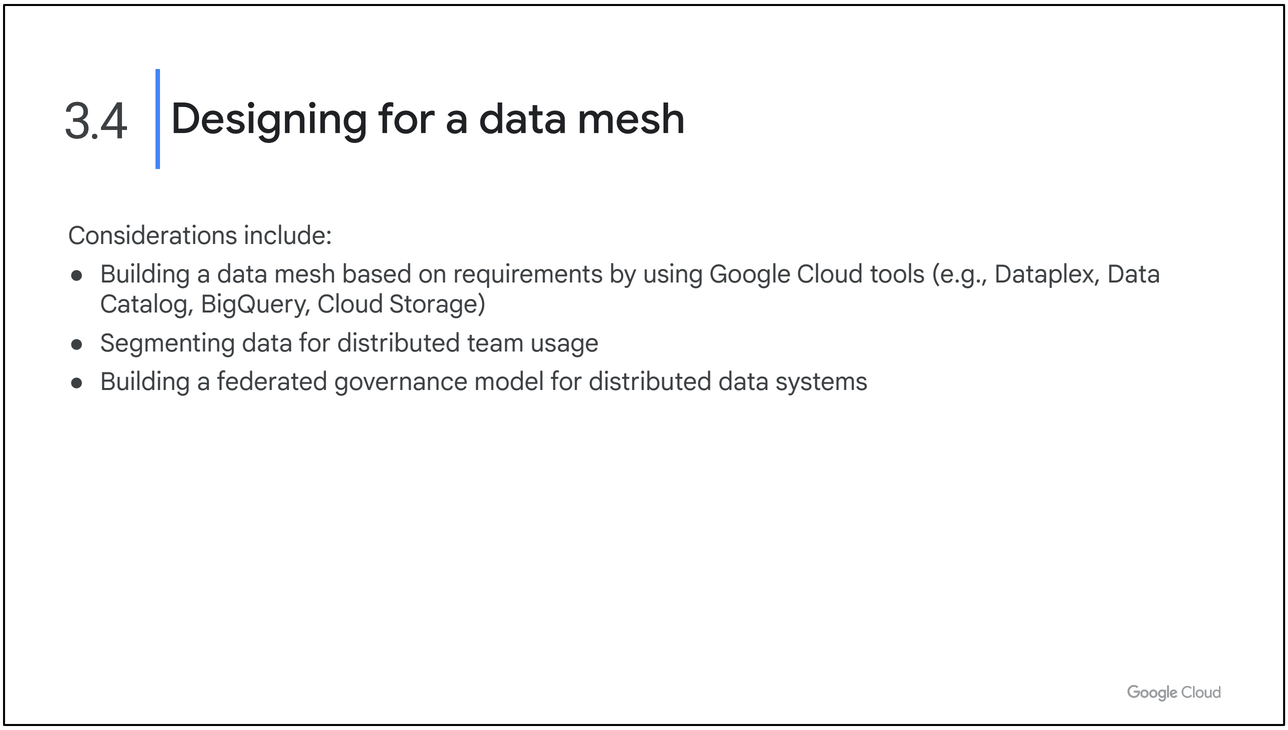
Object Lifecycle Management | Cloud Storage

Overview of access control | Cloud Storage

Cloud Audit Logs with Cloud Storage | Google Cloud

You just reviewed diagnostic questions that addressed considerations related to using a data lake. 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
https://cloud.google.com/storage/docs/lifecycle
https://cloud.google.com/storage/docs/access-control
https://cloud.google.com/storage/docs/audit-logging



3.4 Designing for a data mesh

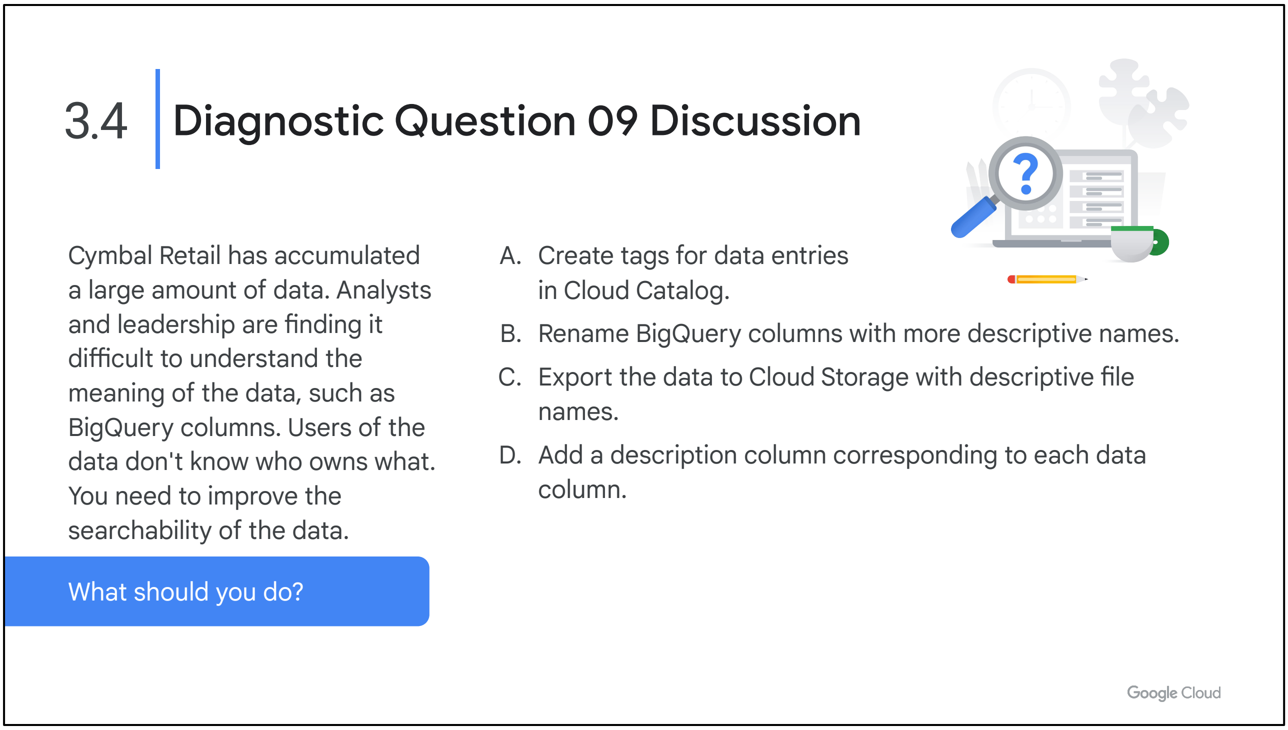
Considerations include:

- Building a data mesh based on requirements by using Google Cloud tools (e.g., Dataplex, Data Catalog, BigQuery, Cloud Storage)
- Segmenting data for distributed team usage
- Building a federated governance model for distributed data systems

Google Cloud

The concept of a data mesh is becoming very popular today as companies need the ability to distribute ownership of data across teams that own the business context. As a Professional Data Engineer, you should be able to build a data mesh, segment data for distributed team usage, and ensure that the overall data lifecycle management and governance is consistently applied across the distributed data landscape.

Question 9 tested your ability to build a data mesh using Google Cloud tools. Question 10 asked you determine how to segment and govern data for distributed team usage.



3.4 Diagnostic Question 09 Discussion

Cymbal Retail has accumulated a large amount of data. Analysts and leadership are finding it difficult to understand the meaning of the data, such as BigQuery columns. Users of the data don't know who owns what. You need to improve the searchability of the data.

What should you do?

A. Create tags for data entries in Cloud Catalog.

B. Rename BigQuery columns with more descriptive names.

C. Export the data to Cloud Storage with descriptive file names.

D. Add a description column corresponding to each data column.

Google Cloud

Feedback:
A: Correct. Tags enable attaching metadata to data assets and entities. This can improve searchability of the data.

B: Incorrect. Renaming data columns is not a viable option, because it would affect programmes elsewhere. Data might also be ingested from external sources where you do not have the ability to change the column names.

C: Incorrect. Analysts require the data in analytic engines for performance and convenience. Moving the data to files in Cloud Storage is not a good option.

D: Incorrect. Adding description columns for each column is not a convenient approach. It increases data in the tables and also requires changes to existing schemas, whether they were internally or externally defined.

Links:
https://cloud.google.com/data-catalog/docs/tags-and-tag-templates
https://cloud.google.com/data-catalog/docs/tag-bigquery-dataset

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

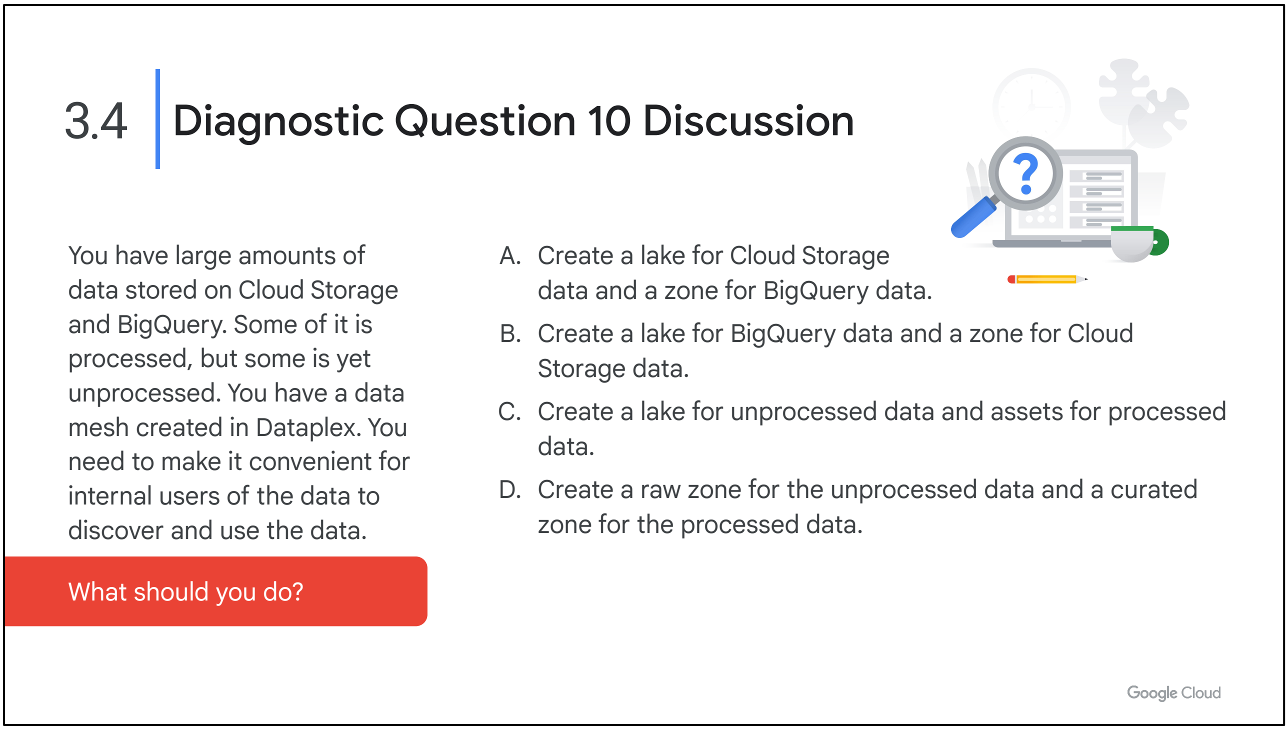
- Introduction to Data Engineering

Building Batch Data Pipelines on Google Cloud

- Introduction to Building Batch Data Pipelines

Skill badges:
Data Catalog Fundamentals

Summary:
As data increases, the challenge of making the data usable also increases. The data engineer can configure ways to attach metadata to the data, automatically and manually, to increase data searchability and understanding.



3.4 Diagnostic Question 10 Discussion

You have large amounts of data stored on Cloud Storage and BigQuery. Some of it is processed, but some is yet unprocessed. You have a data mesh created in Dataplex. You need to make it convenient for internal users of the data to discover and use the data.

What should you do?

A. Create a lake for Cloud Storage data and a zone for BigQuery data.

B. Create a lake for BigQuery data and a zone for Cloud Storage data.

C. Create a lake for unprocessed data and assets for processed data.

D. Create a raw zone for the unprocessed data and a curated zone for the processed data.

Google Cloud

Feedback:
A: Incorrect. A lake represents a data domain or business unit. A lake could encompass both Cloud Storage and BigQuery data.

B: Incorrect. A lake represents a data domain or business unit. A lake could encompass both Cloud Storage and BigQuery data.

C: Incorrect. A zone is where we make a distinction between processed and unprocessed data.

D: Correct. The recommended architecture is to use raw zones for unprocessed data and curated zones for processed data.

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

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

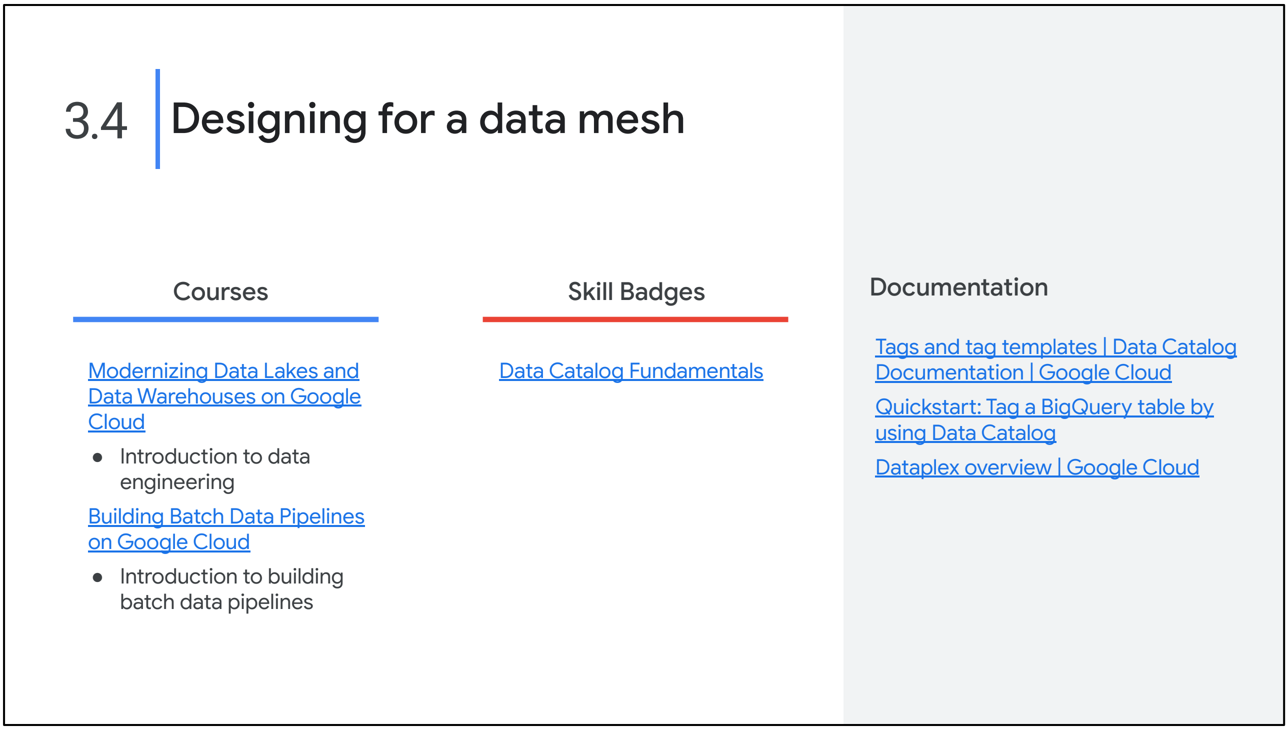
- Introduction to Data Engineering

Building Batch Data Pipelines on Google Cloud

- Introduction to Building Batch Data Pipelines

Skill badges:
Data Catalog Fundamentals

Summary:
Dataplex has logical concepts that abstract the underlying storage technology. There is a hierarchy of lakes, zones, assets, and entities. As a Professional Data Engineer, using these terms consistently and designing with them will make communicating the ideas also consistent.



3.4 Designing for a data mesh

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

Skill Badges

[Data Catalog Fundamentals]

Documentation

[Tags and tag templates | Data Catalog Documentation | Google Cloud]

[Quickstart: Tag a BigQuery table by using Data Catalog]

[Dataplex overview | Google Cloud]

You just reviewed diagnostic questions that addressed different aspects of designing for a data mesh. 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/data-catalog/docs/tags-and-tag-templates
https://cloud.google.com/data-catalog/docs/tag-bigquery-dataset
https://cloud.google.com/dataplex/docs/introduction