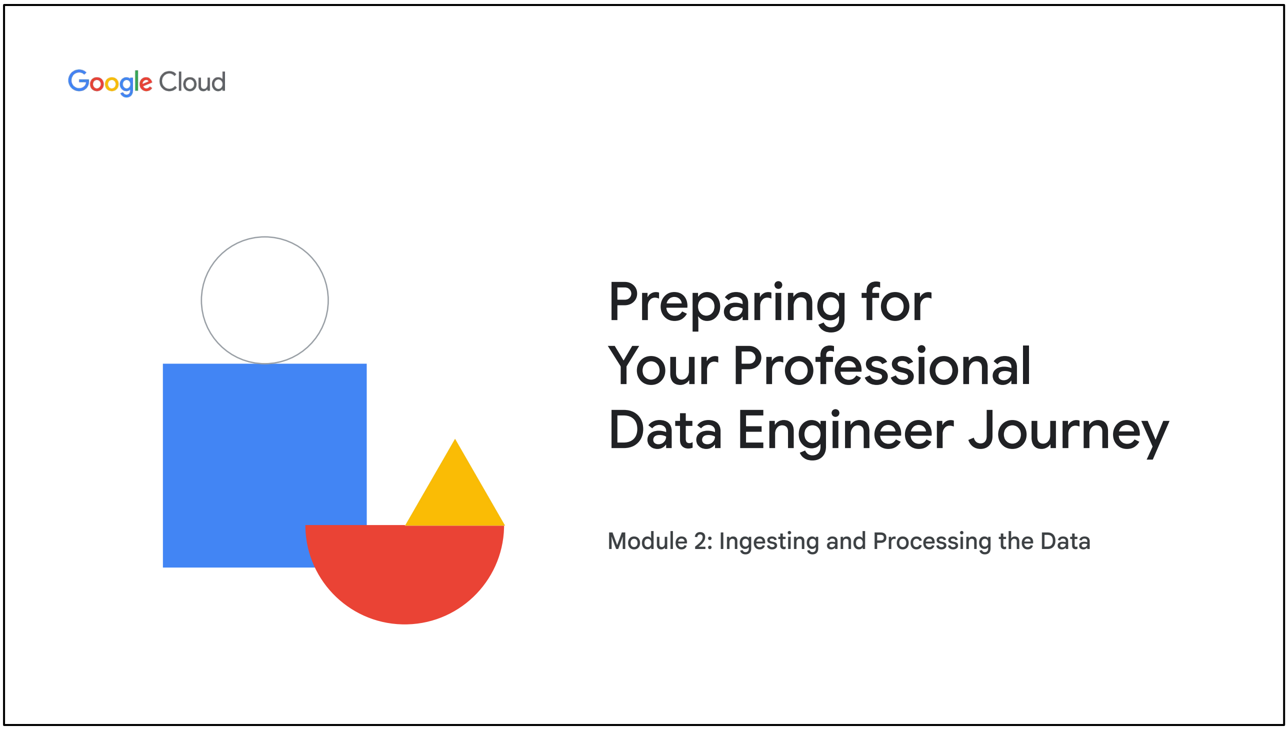
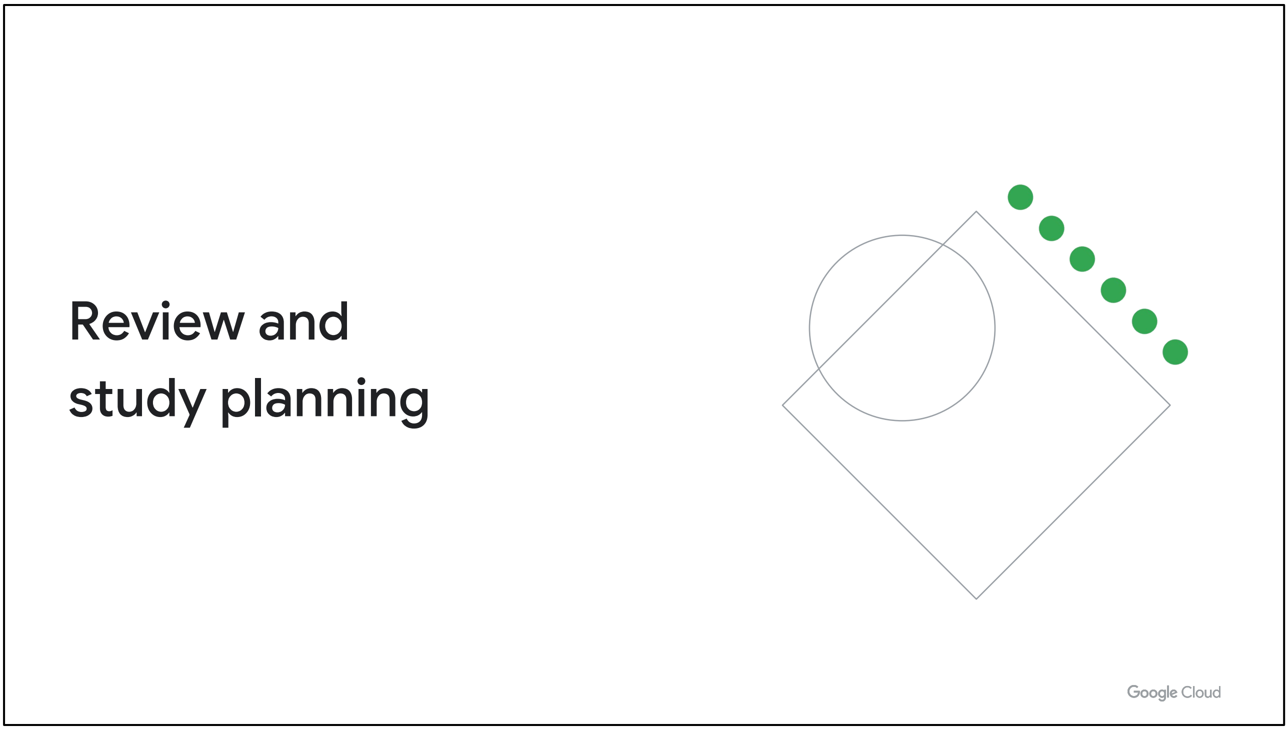


Google Cloud

Preparing for Your Professional Data Engineer Journey

Module 2: Ingesting and Processing the Data

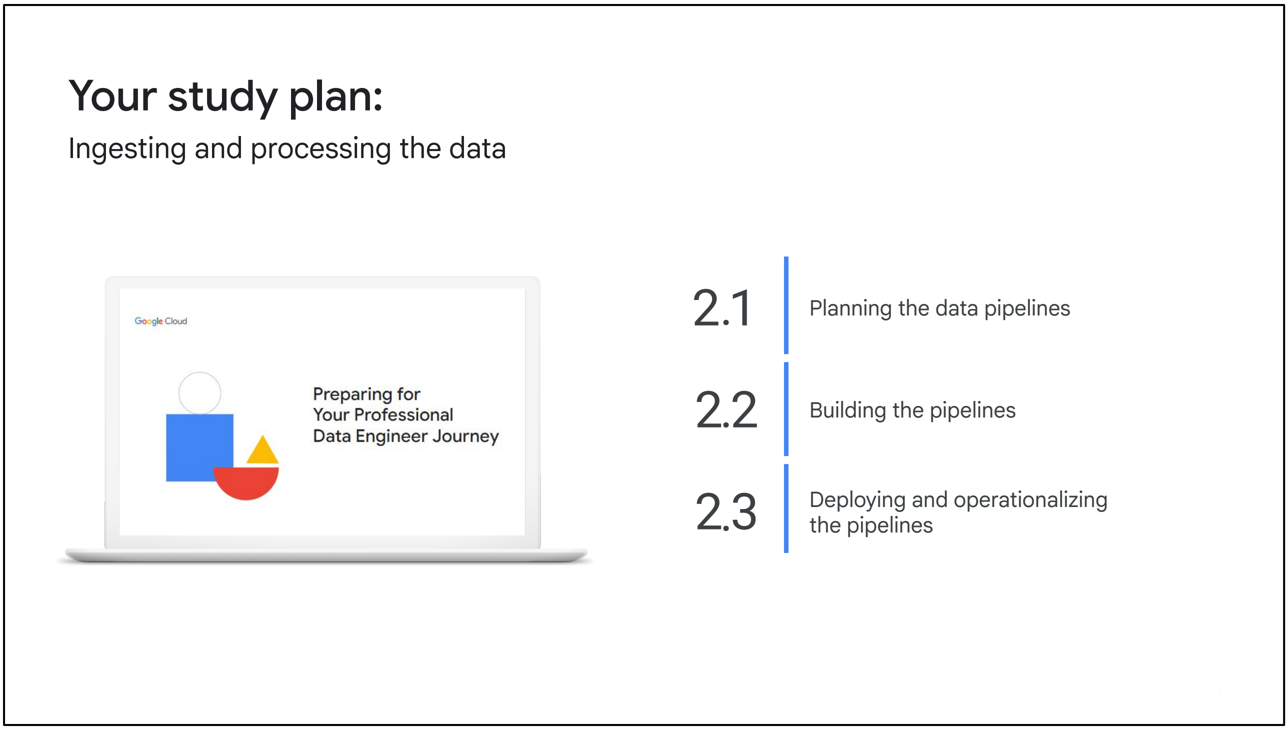
Welcome to Module 2: Ingesting and Processing Data.



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:

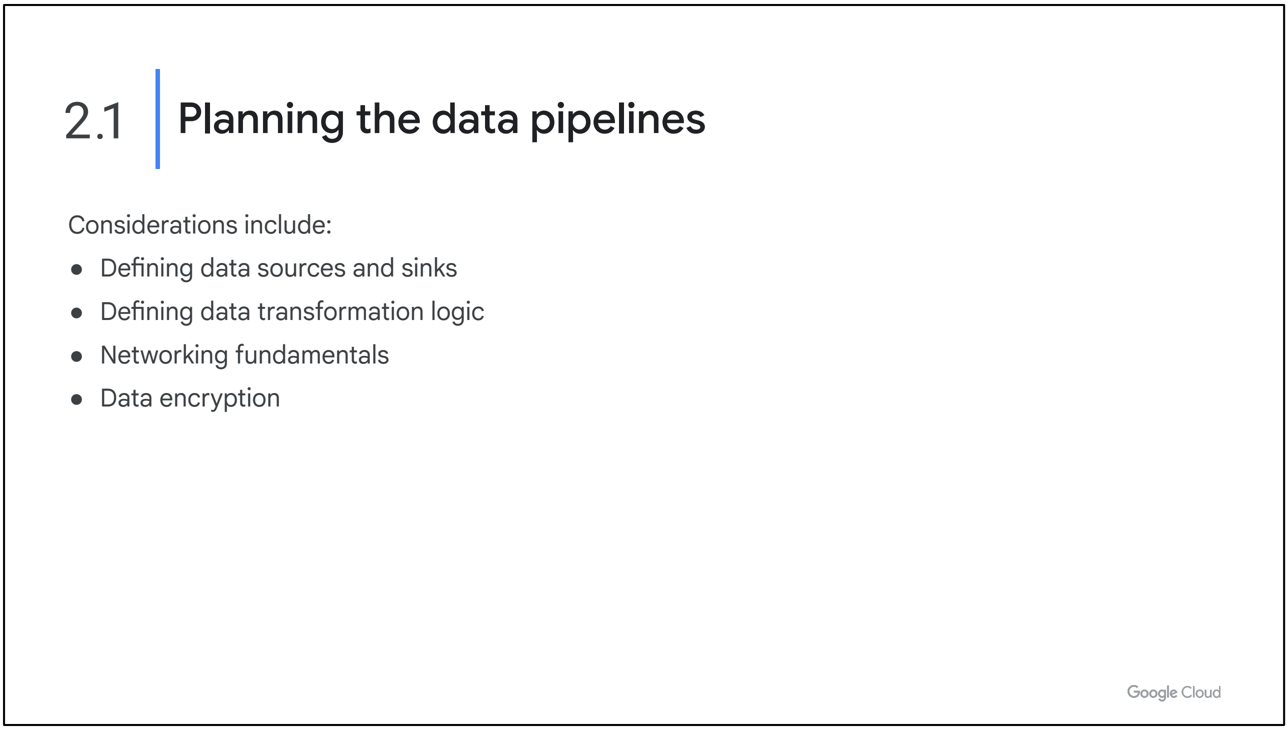
Ingesting and processing the data

2.1 Planning the data pipelines

2.2 Building the pipelines

2.3 Deploying and operationalizing the pipelines

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.



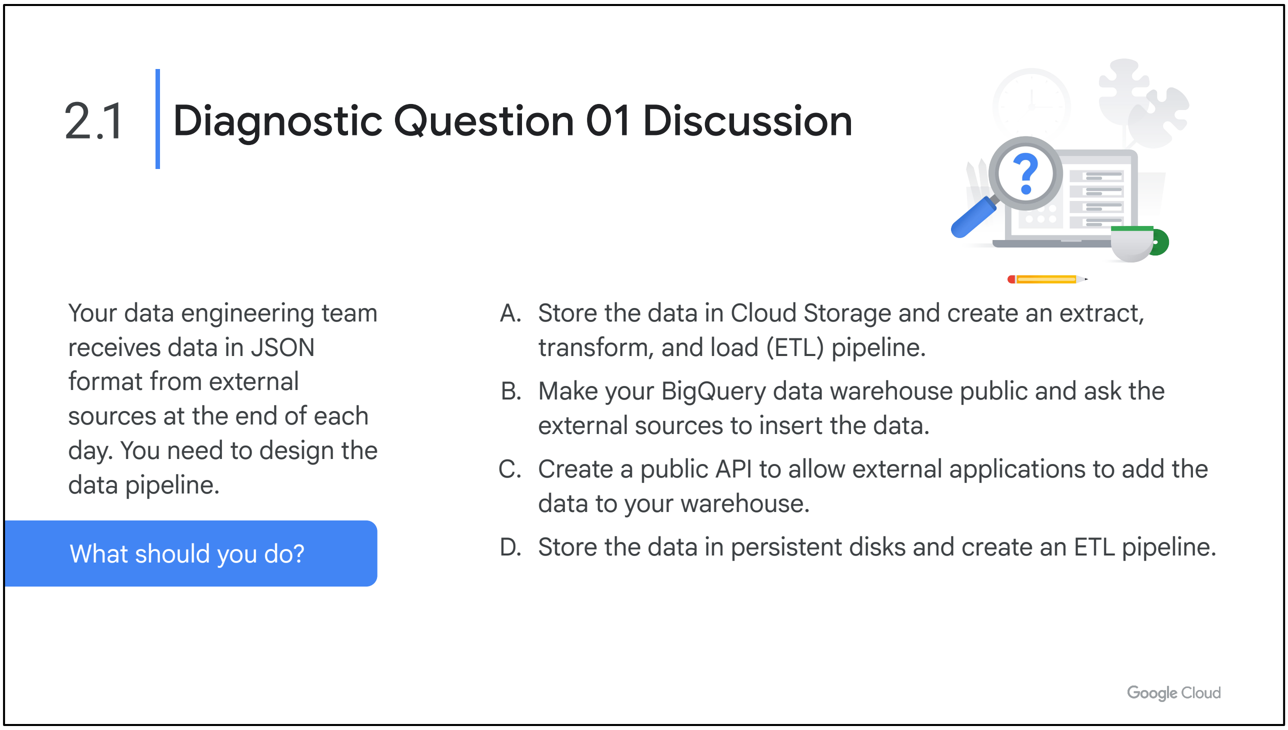
2.1 Planning the data pipelines

Considerations include:

- Defining data sources and sinks
- Defining data transformation logic
- Networking fundamentals
- Data encryption

A fundamental part of the role of a Professional Data Engineer is to design, build and maintain data pipelines. The first step in designing a data pipeline is to identify where the data originates and where the data should finally reside. A data sink is typically a data warehouse like BigQuery and enables the subsequent analysis. In addition, you also need to look at the data security aspects as the data moves from the source to the sink.

Question 1 asked you to differentiate between data source options to select the most appropriate option for a data pipeline use case. Question 2 challenged you to differentiate between data sink options to select the most appropriate option for a data pipeline use case. In Question 3, you defined appropriate transformation logic for your data pipeline goals.



2.1 Diagnostic Question 01 Discussion

Your data engineering team receives data in JSON format from external sources at the end of each day. You need to design the data pipeline.

What should you do?

A. Store the data in Cloud Storage and create an extract, transform, and load (ETL) pipeline.
B. Make your BigQuery data warehouse public and ask the external sources to insert the data.
C. Create a public API to allow external applications to add the data to your warehouse.
D. Store the data in persistent disks and create an ETL pipeline.

Google Cloud

Feedback:

A. Correct. The recommended approach for batch data pipelines is to store data in Cloud Storage. Then, create an ETL (or ELT, depending on the use case) pipeline to move the data into a data warehouse.
B. Incorrect. Making your data warehouse public is not recommended. The data also might not have the transformations you require.
C. Incorrect. Data added by external sources might not have the transformations you require.
D. Incorrect. Storing the data in persistent disks is not recommended for data analytics pipelines. A centralized, serverless storage, such as Cloud Storage, is preferable.

Links:
https://cloud.google.com/blog/topics/developers-practitioners/what-data-pipeline-architecture-should-i-use

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

- Building a Data Lake
- Building a Data Warehouse

Build Batch Data Pipelines on Google Cloud

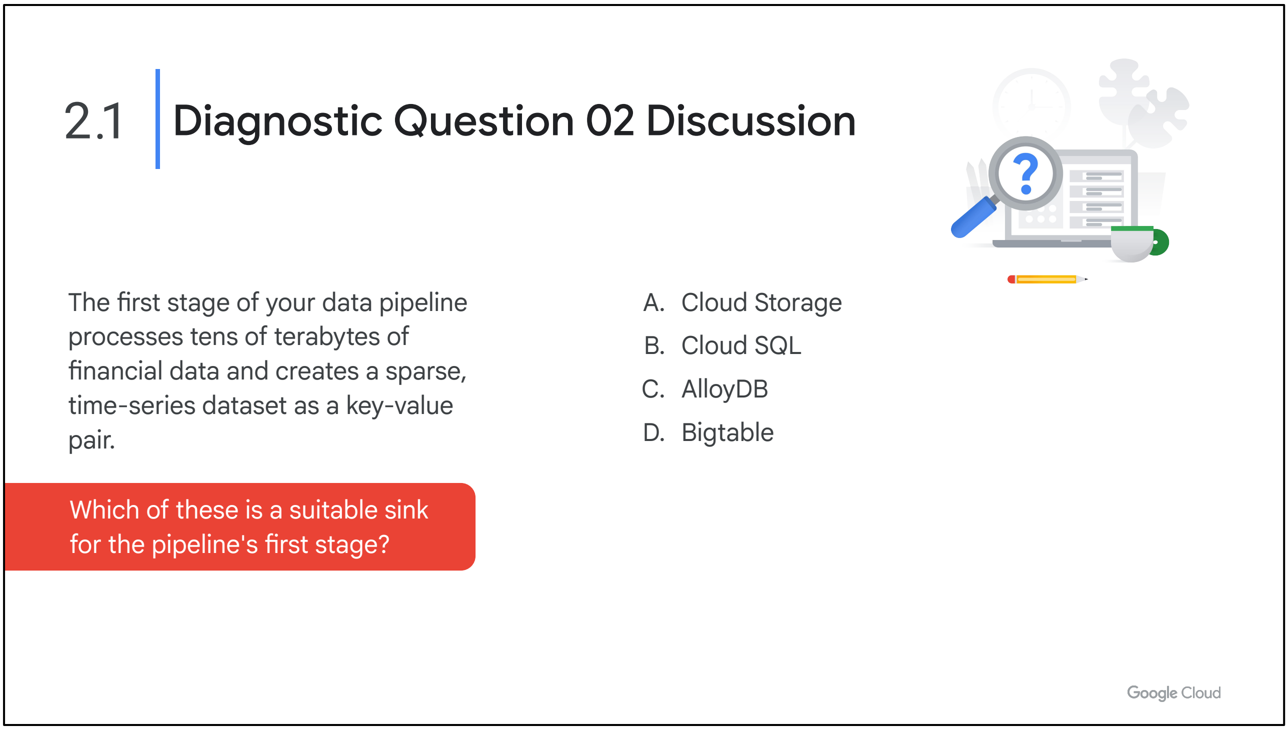
- Executing Spark on Dataproc
- Manage Data Pipelines with Cloud Data Fusion and Cloud Composer

Serverless Data Processing with Dataflow: Develop Pipelines

- Sources and Sinks

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

Summary:
Cloud Storage is recommended for storage for data pipelines that insert batch data, including those that arrive from external sources. Cloud Storage also serves as a long term backup.



2.1 Diagnostic Question 02 Discussion

The first stage of your data pipeline processes tens of terabytes of financial data and creates a sparse, time-series dataset as a key-value pair.

Which of these is a suitable sink for the pipeline's first stage?

A. Cloud Storage
B. Cloud SQL
C. AlloyDB
D. Bigtable

Feedback:

A. Incorrect. Cloud Storage is used for object storage and is not suitable as a database.
B. Incorrect. Cloud SQL supports relational databases, which are not as suitable as Bigtable for sparse, time-series data.
C. Incorrect. AlloyDB is a relational database, which is not as suitable as Bigtable for sparse, time-series data.
D. Correct. Bigtable is ideal for applications that need high throughput and scalability for key/value data, where each value is typically no larger than 10 MB. Bigtable is suitable for applications that work on time-series data, such as financial applications.

Links:
https://cloud.google.com/bigtable/docs/overview

More information:
Courses:
Introduction to Data Engineering on Google Cloud

Build Data Lakes and Data Warehouses on Google Cloud

- Building a Data Warehouse

Build Streaming Data Pipelines on Google Cloud

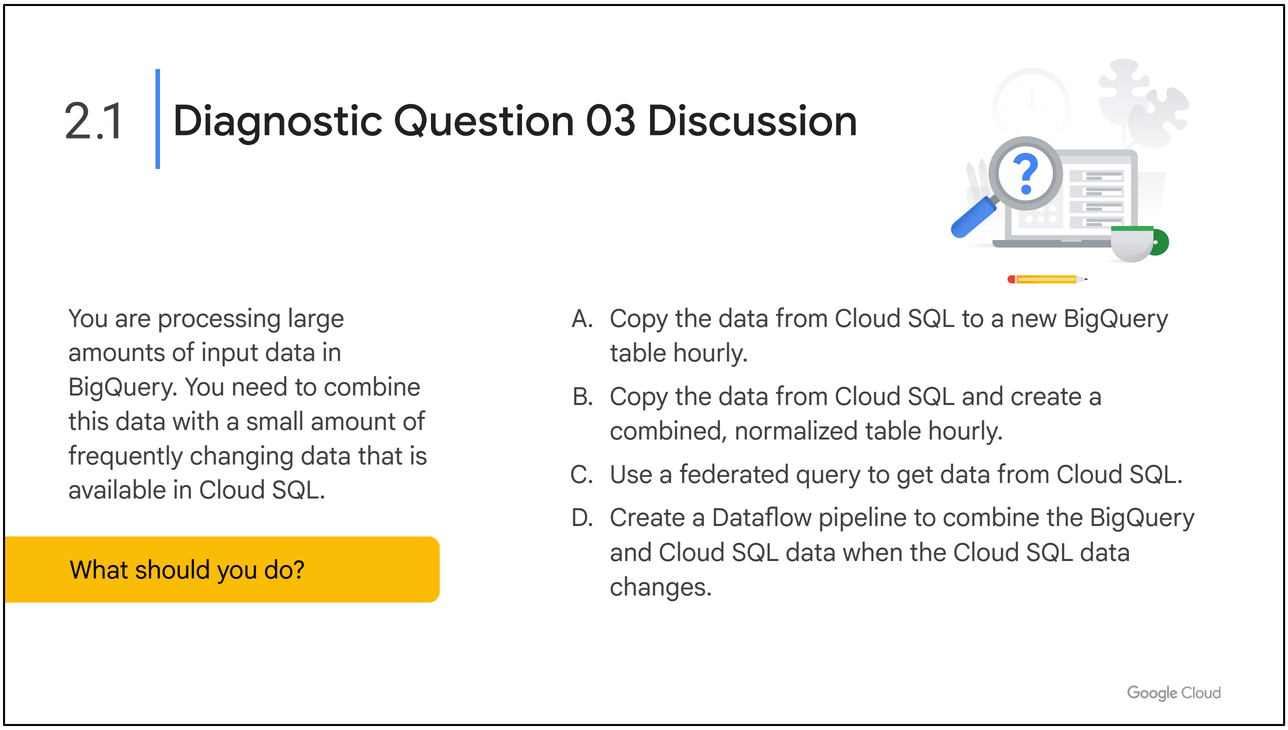
- High-Throughput BigQuery and Bigtable Streaming Features

Serverless Data Processing with Dataflow: Develop Pipelines

- Sources and Sinks

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

Summary:
Google Cloud supports a variety of managed databases. When choosing one for your application, identify the kind of data you want to store, the type of usage you have for it: analytical or transactional, and scaling.



2.1 Diagnostic Question 03 Discussion

You are processing large amounts of input data in BigQuery. You need to combine this data with a small amount of frequently changing data that is available in Cloud SQL.

What should you do?

A. Copy the data from Cloud SQL to a new BigQuery table hourly.

B. Copy the data from Cloud SQL and create a combined, normalized table hourly.

C. Use a federated query to get data from Cloud SQL.

D. Create a Dataflow pipeline to combine the BigQuery and Cloud SQL data when the Cloud SQL data changes.

Google Cloud

Feedback:

A. Incorrect. Because the data is frequently changing, importing the table hourly could result in inaccurate results.

B. Incorrect. Because the data is frequently changing, importing the table hourly could result in inaccurate results.

C. Correct. Because the data is frequently changing, you can query the data in-place by using federated queries from BigQuery.

D. Incorrect. Because the data is frequently changing, computation and costs might be excessive.

Links:

https://cloud.google.com/bigquery/docs/cloud-sql-federated-queries

https://cloud.google.com/blog/products/data-analytics/exploring-new-features-in-bigquery-federated-queries

More information:

Courses:

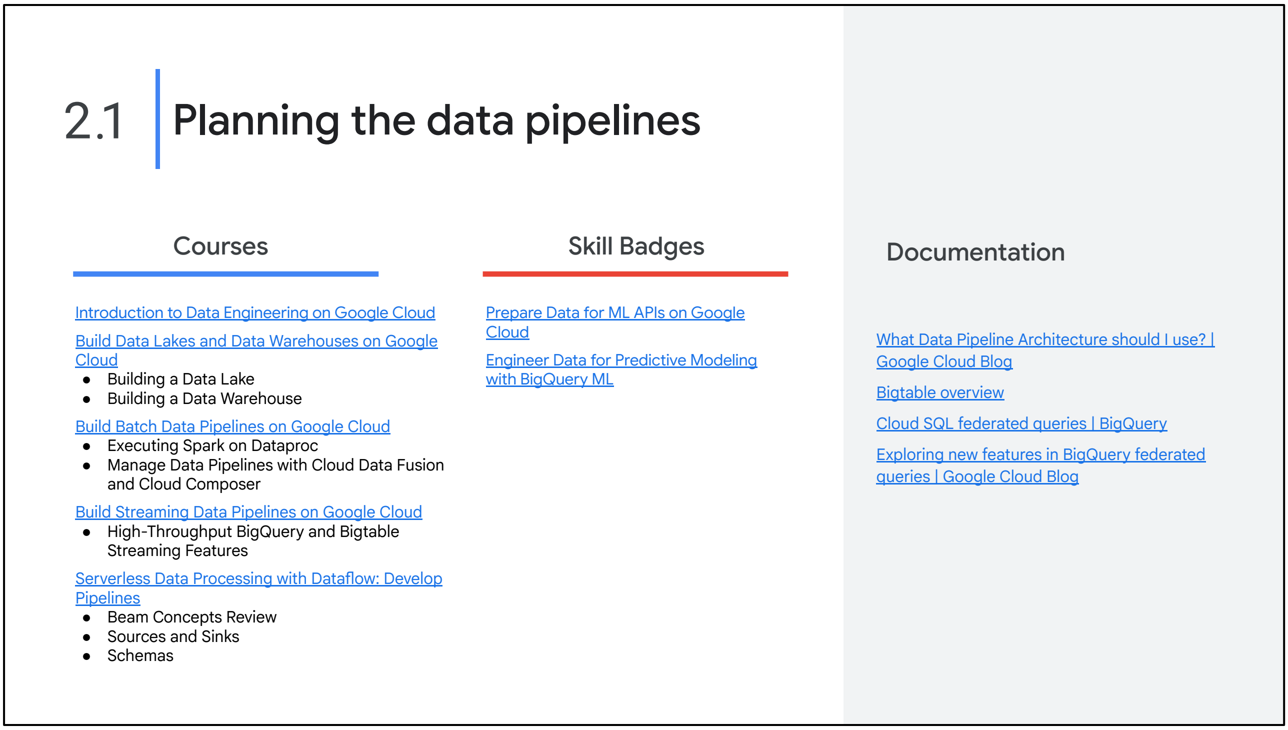
Build Batch Data Pipelines on Google Cloud

Serverless Data Processing with Dataflow: Develop Pipelines

- Beam Concepts Review
- Schemas

Summary:

Some data is static, whereas others are dynamic. A PDE has to build systems that integrate data from different sources and transform them as the business requirements demand. Federated queries allow access of Cloud SQL data directly from BigQuery. This is convenient when the data is changing frequently.



2.1 Planning the data pipelines

Courses

Introduction to Data Engineering on Google Cloud

Build Data Lakes and Data Warehouses on Google Cloud

- Building a Data Lake
- Building a Data Warehouse

Build Batch Data Pipelines on Google Cloud

- Executing Spark on Dataproc
- Manage Data Pipelines with Cloud Data Fusion and Cloud Composer

Build Streaming Data Pipelines on Google Cloud

- High-Throughput BigQuery and Bigtable Streaming Features

Serverless Data Processing with Dataflow: Develop Pipelines

- Beam Concepts Review
- Sources and Sinks
- Schemas

Skill Badges

Prepare Data for ML APIs on Google Cloud

Engineer Data for Predictive Modeling with BigQuery ML

Documentation

What Data Pipeline Architecture should I use? | Google Cloud Blog

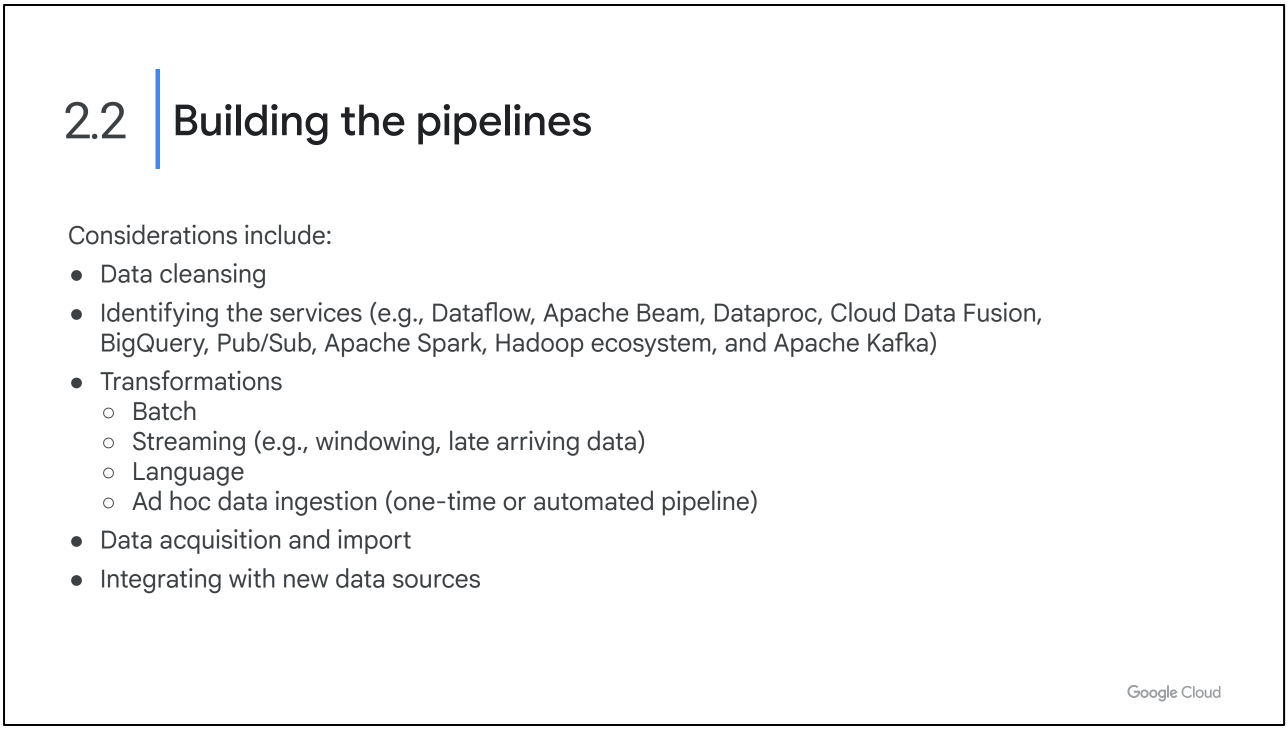
Bigtable overview

Cloud SQL federated queries | BigQuery

Exploring new features in BigQuery federated queries | Google Cloud Blog

You just reviewed several diagnostic questions that addressed different aspects of planning data pipelines. 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/what-data-pipeline-architecture-should-i-use
https://cloud.google.com/bigtable/docs/overview
https://cloud.google.com/bigquery/docs/cloud-sql-federated-queries
https://cloud.google.com/blog/products/data-analytics/exploring-new-features-in-bigquery-federated-queries



2.2 Building the pipelines

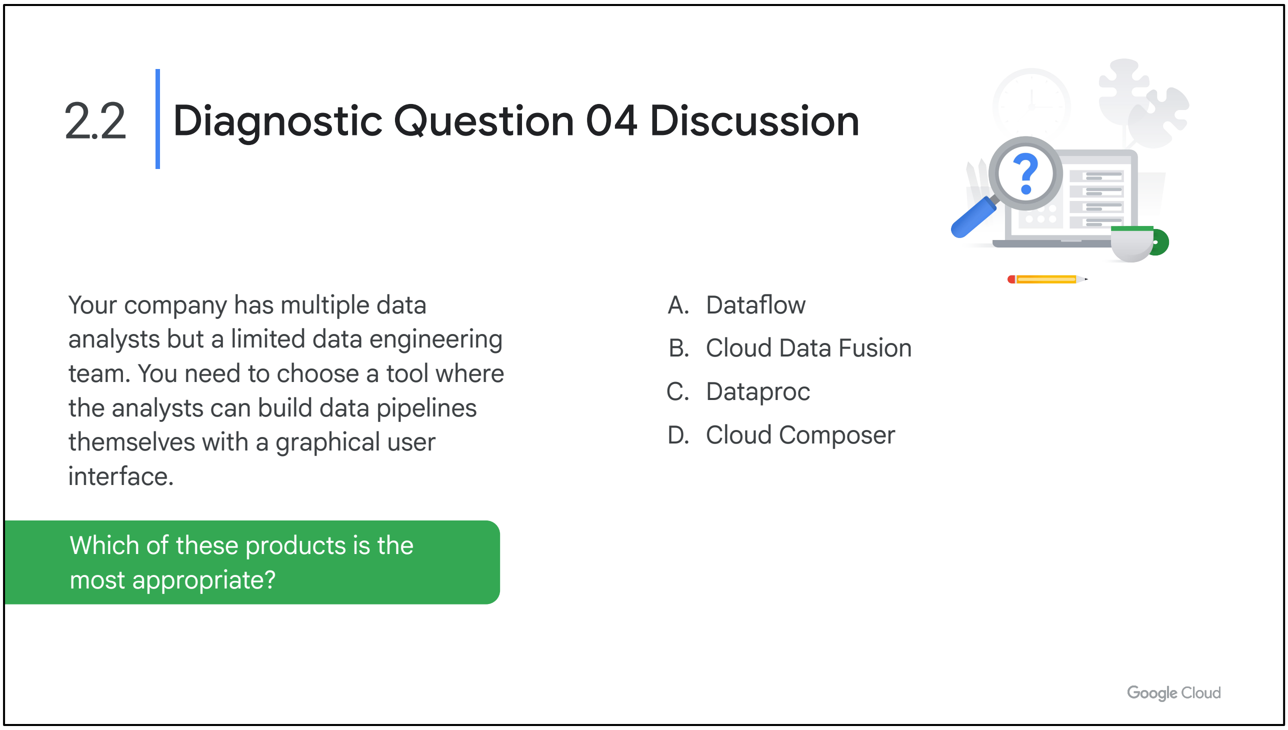
Considerations include:

- Data cleansing
- Identifying the services (e.g., Dataflow, Apache Beam, Dataproc, Cloud Data Fusion, BigQuery, Pub/Sub, Apache Spark, Hadoop ecosystem, and Apache Kafka)
- Transformations
  - Batch
  - Streaming (e.g., windowing, late arriving data)
  - Language
  - Ad hoc data ingestion (one-time or automated pipeline)
- Data acquisition and import
- Integrating with new data sources

Google Cloud

As a Professional Data Engineer, you will spend a significant amount of your time on the data cleansing activities. If the incoming data is mostly clean, you can use tools like Dataprep to do the remaining cleanup. But most of the time, the incoming data will be in a raw format and you will need to do complex data processing to transform that data into a suitable form. As a Professional Data Engineer, you should be familiar with multiple tools including Dataproc, Dataflow, Data Fusion, and Dataprep, among others, and use an appropriate tool based on your use case. You will be handling both batch data as well as real time streaming data. Streaming data pipelines are significantly more complex than the batch pipelines and you should be familiar with the concepts like windowing, late inputs, and early evaluation.

Question 4 asked you to differentiate between options for processing data. Question 5 tested your knowledge of options and considerations for batch data workflows, and Question 6 asked about tools for batch data workflows in Google Cloud. Question 7 tested your knowledge of options and considerations for streaming data workflows, and Question 8 asked about tools for batch data workflows in Google Cloud.



2.2 Diagnostic Question 04 Discussion

Your company has multiple data analysts but a limited data engineering team. You need to choose a tool where the analysts can build data pipelines themselves with a graphical user interface.

Which of these products is the most appropriate?

A. Dataflow
B. Cloud Data Fusion
C. Dataproc
D. Cloud Composer

Google Cloud

Feedback:

A. Incorrect. Dataflow requires programming knowledge.
B. Correct. The Cloud Data Fusion web UI lets you build scalable data integration solutions to clean, prepare, blend, transfer, and transform data, without having to manage the infrastructure.
C. Incorrect. Dataproc requires considerable technical knowledge.
D. Incorrect. With Cloud Composer, you need to create a pipeline with code.

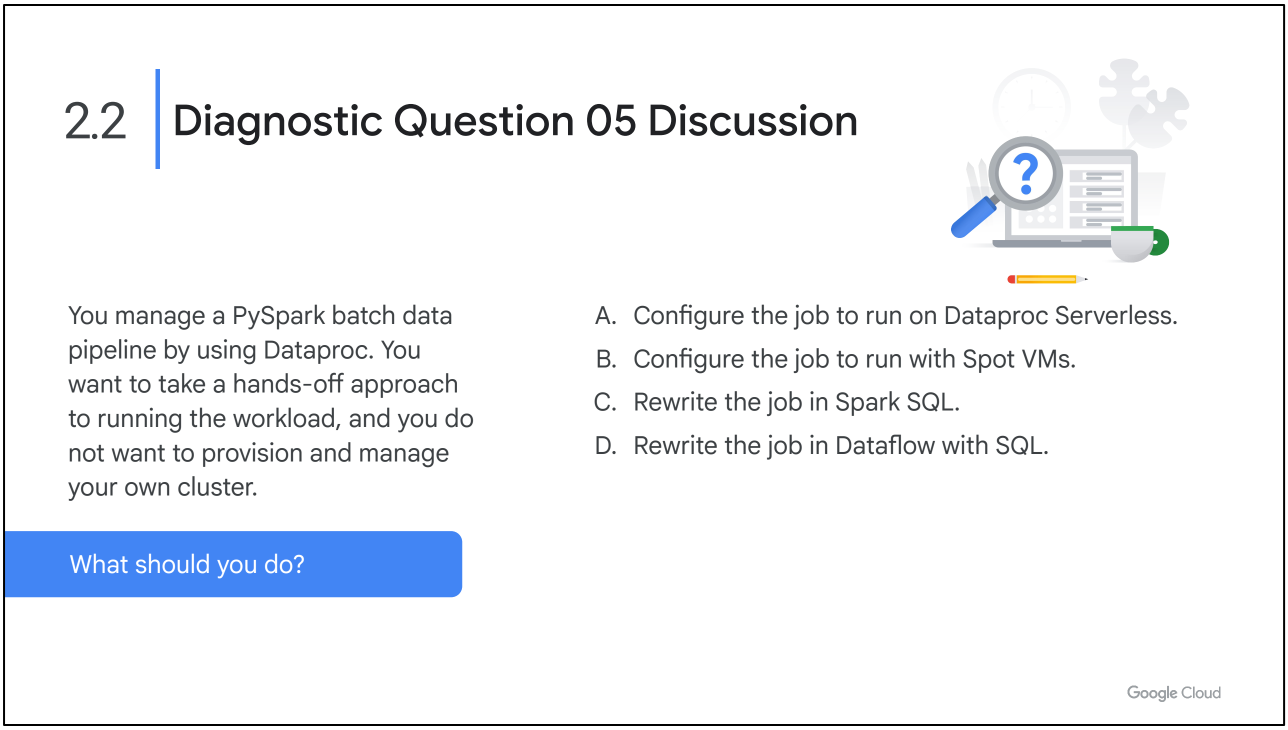
Links:
https://cloud.google.com/data-fusion/docs/concepts/overview

More information:
Courses:
Build Batch Data Pipelines on Google Cloud

- Introduction to Building Batch Data Pipelines
- Executing Spark on Dataproc
- Serverless Data Processing with Dataflow
- Manage Data Pipelines with Cloud Data Fusion and Cloud Composer

Summary:
Google Cloud offers various tools to create data pipelines. Identify the use case for the data pipeline, and choose the tool that is most suitable. For example, Dataflow and Cloud Composer require programming knowledge. Others such as Dataprep and

Cloud Fusion require less technical knowledge. Google Cloud's data processing tools often also have open source equivalents. If such data processing jobs are being migrated to Google Cloud, choose the corresponding managed solution.



2.2 Diagnostic Question 05 Discussion

You manage a PySpark batch data pipeline by using Dataproc. You want to take a hands-off approach to running the workload, and you do not want to provision and manage your own cluster.

What should you do?

A. Configure the job to run on Dataproc Serverless.
B. Configure the job to run with Spot VMs.
C. Rewrite the job in Spark SQL.
D. Rewrite the job in Dataflow with SQL.

Google Cloud

Feedback:

A. Correct. Dataproc Serverless will automatically provision the resources to run your Dataproc jobs.
B. Incorrect. Configuring Spot VMs is a cost reduction strategy, and not an effort reduction.
C. Incorrect. Running a Spark SQL job does not automatically take care of cluster management.
D. Incorrect. Rewriting the jobs in Dataflow requires more effort.

Links:
https://cloud.google.com/dataproc-serverless/docs/overview

More information:
Courses:
Serverless Data Processing with Dataflow: Foundations

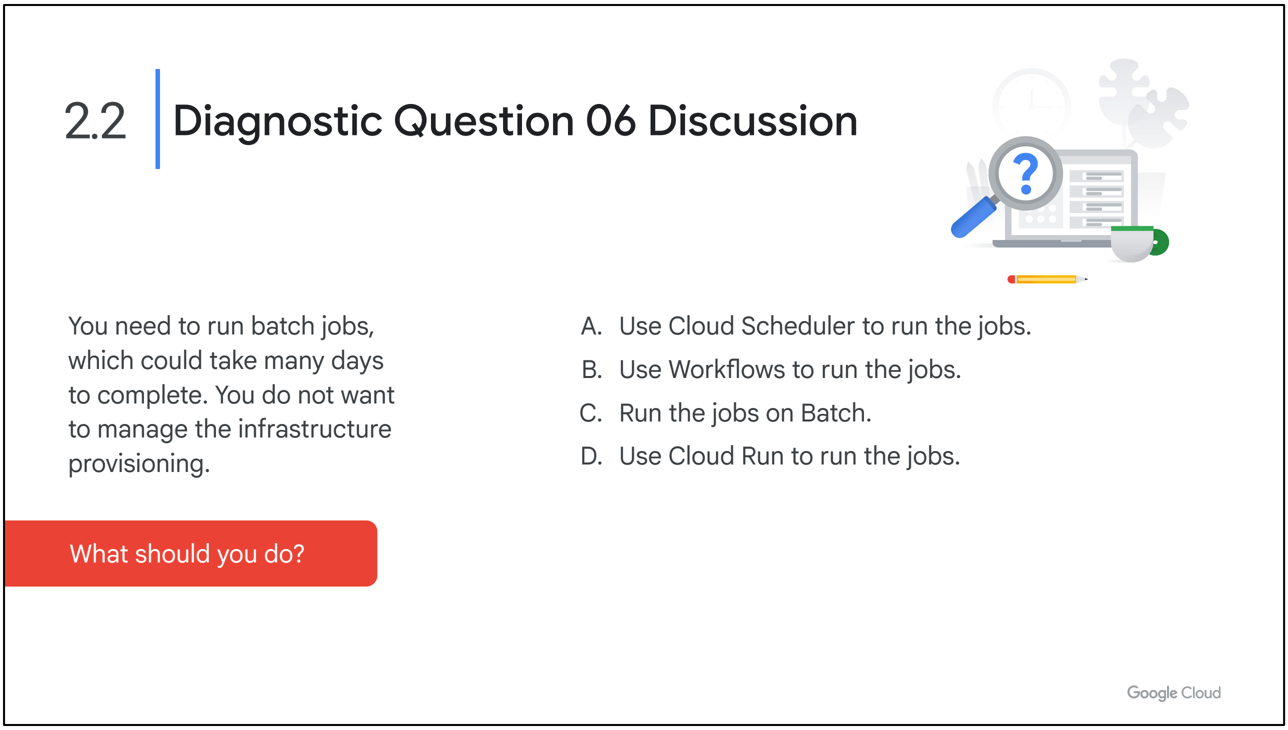
- Separating Compute and Storage with Dataflow

Serverless Data Processing with Dataflow: Operations

- Performance

Summary:
In your role as a Professional Data Engineer, you want to reduce time on repetitive tasks. Provisioning and managing resources for long running, batch data jobs is one

of them. Batch jobs can run for hours and days, and you don't want to be manually monitoring capacity. Google Cloud has multiple serverless options, including with some Dataproc workloads. Using them can save effort on cluster management.



2.2 Diagnostic Question 06 Discussion

You need to run batch jobs, which could take many days to complete. You do not want to manage the infrastructure provisioning.

What should you do?

A. Use Cloud Scheduler to run the jobs.
B. Use Workflows to run the jobs.
C. Run the jobs on Batch.
D. Use Cloud Run to run the jobs.

Feedback:

A. Incorrect. Cloud Scheduler lets you schedule an event in the future, but it does not automatically allocate any resources.
B. Incorrect. Workflows is recommended when you want to connect a series of shorter tasks.
C. Correct. Batch is a fully managed service that schedules, queues, and executes batch processing workloads on Google Cloud. Resources and capacity are provisioned and managed for you based on your requirements.
D. Incorrect. Cloud Run has limits on the maximum time of a running process and is not a viable option for long running data processing jobs.

Links:
https://www.youtube.com/watch?v=RS7UJhD4R48
https://cloud.google.com/batch/docs/get-started

More information:
Courses:
Build Batch Data Pipelines on Google Cloud

- Executing Spark on Dataproc
- Manage Data Pipelines with Cloud Data Fusion and Cloud Composer

Serverless Data Processing with Dataflow: Foundations

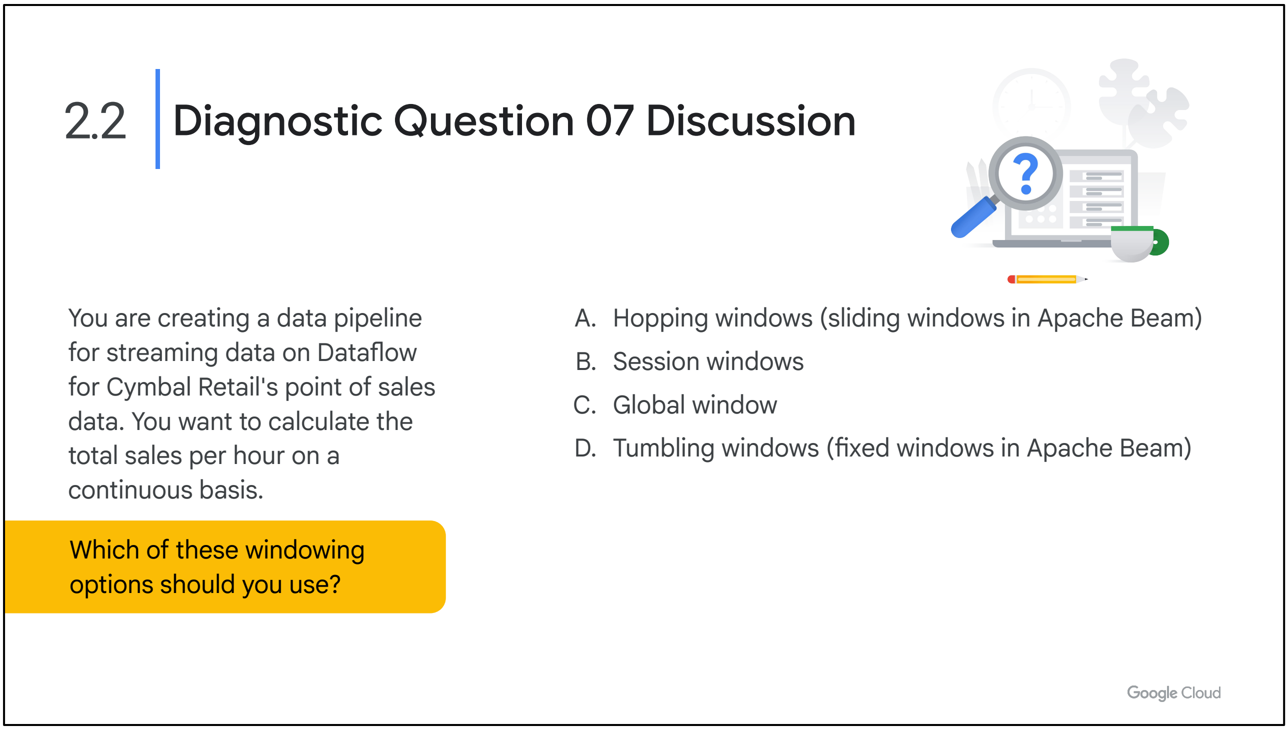
- Separating Compute and Storage with Dataflow

Serverless Data Processing with Dataflow: Develop Pipelines

- Windows, Watermarks, and Triggers

Skill badge:
Prepare Data for ML APIs on Google Cloud

Summary:
Batch is built for running long running batch jobs. It's a managed service that automatically provisions resources to run the batch processing that you configured.



2.2 Diagnostic Question 07 Discussion

You are creating a data pipeline for streaming data on Dataflow for Cymbal Retail's point of sales data. You want to calculate the total sales per hour on a continuous basis.

Which of these windowing options should you use?

A. Hopping windows (sliding windows in Apache Beam)
B. Session windows
C. Global window
D. Tumbling windows (fixed windows in Apache Beam)

Google Cloud

Feedback:

A. Incorrect. Hopping windows (or sliding windows in Apache Beam) can overlap and are not the right option for this requirement.
B. Incorrect. Session windows are appropriate for use cases with gaps in the activity, which is not the case here.
C. Incorrect. A global window applies to the whole dataset and is not the appropriate choice for intermediate results.
D. Correct. A tumbling window (or fixed window in Apache Beam) is fixed duration and non-overlapping, which is the right option for this requirement.

Links:
https://cloud.google.com/dataflow/docs/concepts/streaming-pipelines
https://beam.apache.org/documentation/basics/#window

More information:
Courses:
Serverless Data Processing with Dataflow: Develop Pipelines

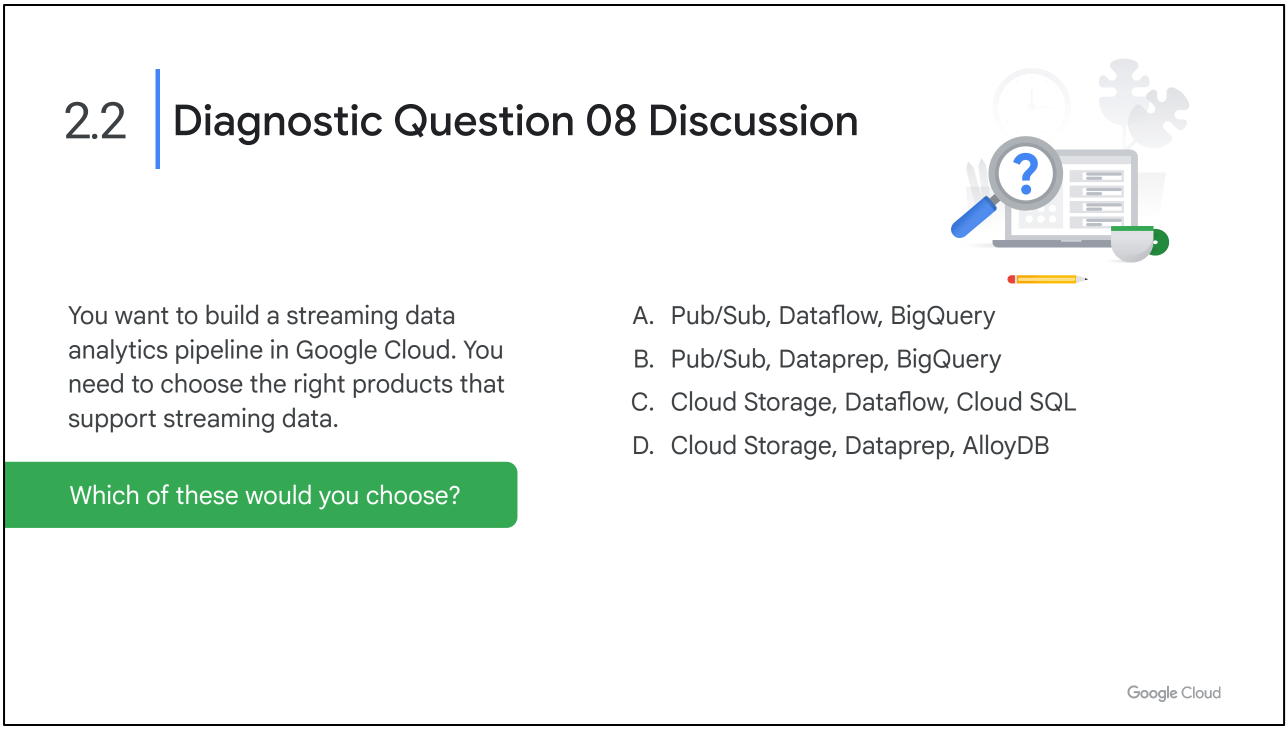
- Windows, Watermarks, and Triggers
- State and Timers
- Dataflow SQL and DataFrames

Serverless Data Processing with Dataflow: Operations

- Performance

- Testing and CI/CD
- Flex Templates

Summary:
When data is non-finite but you need intermediate results, windowing helps separate the entire time period into intermediate time periods of processing. Combined with watermarks and triggers, windowing gives the developer the flexibility to control when data processing occurs.



2.2 Diagnostic Question 08 Discussion

You want to build a streaming data analytics pipeline in Google Cloud. You need to choose the right products that support streaming data.

Which of these would you choose?

A. Pub/Sub, Dataflow, BigQuery
B. Pub/Sub, Dataprep, BigQuery
C. Cloud Storage, Dataflow, Cloud SQL
D. Cloud Storage, Dataprep, AlloyDB

Feedback:

A. Correct. Pub/Sub, Dataflow, and BigQuery support streaming data and form the recommended pipeline for continuous data processing.
B. Incorrect. Dataprep does not support streaming data processing.
C. Incorrect. Cloud Storage is appropriate for a data lake but not for streaming data. Cloud SQL is a transactional storage solution and does not have streaming capabilities.
D. Incorrect. Cloud Storage is appropriate for a data lake but not for streaming data. AlloyDB is a transactional database.

Links:
https://cloud.google.com/solutions/stream-analytics

More information:
Courses:
Build Streaming Data Pipelines on Google Cloud

- Serverless Messaging with Pub/Sub
- Dataflow Streaming Features

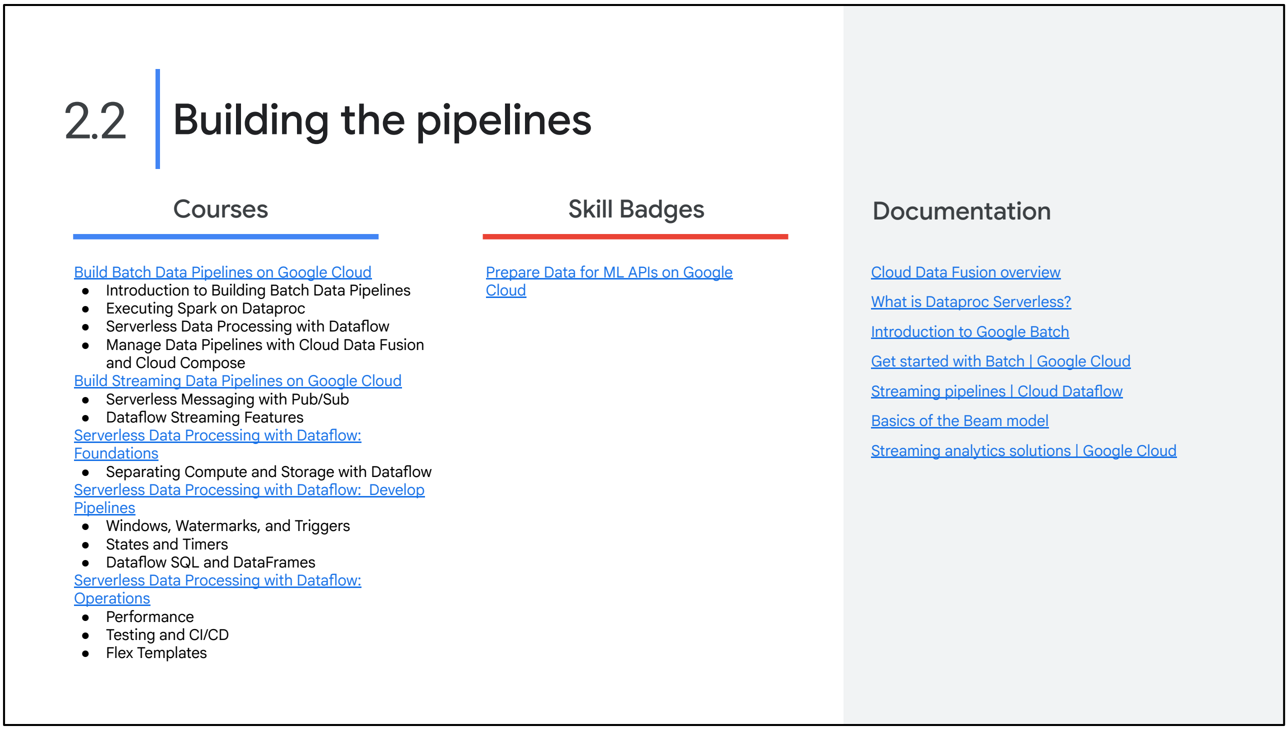
Serverless Data Processing with Dataflow: Foundations

- Separating Compute and Storage with Dataflow

Serverless Data Processing with Dataflow: Develop Pipelines

- Windows, Watermarks, and Triggers
- State and Timers
- Dataflow SQL and DataFrames

Summary:
Streaming analytics requires tools that are tuned to continuous processing. On Google Cloud, Pub/Sub, BigQuery, Dataflow, and Datastream are a few of the tools that are recommended for streaming analytics.



2.2 Building the pipelines

Courses

[Build Batch Data Pipelines on Google Cloud]
- Introduction to Building Batch Data Pipelines
- Executing Spark on Dataproc
- Serverless Data Processing with Dataflow
- Manage Data Pipelines with Cloud Data Fusion and Cloud Compose

Build Streaming Data Pipelines on Google Cloud
- Serverless Messaging with Pub/Sub
- Dataflow Streaming Features

Serverless Data Processing with Dataflow: Foundations
- Separating Compute and Storage with Dataflow

Serverless Data Processing with Dataflow: Develop Pipelines
- Windows, Watermarks, and Triggers
- States and Timers
- Dataflow SQL and DataFrames

Serverless Data Processing with Dataflow: Operations
- Performance
- Testing and CI/CD
- Flex Templates

Skill Badges

Prepare Data for ML APIs on Google Cloud

Documentation

Cloud Data Fusion overview

What is Dataproc Serverless?

Introduction to Google Batch

Get started with Batch | Google Cloud

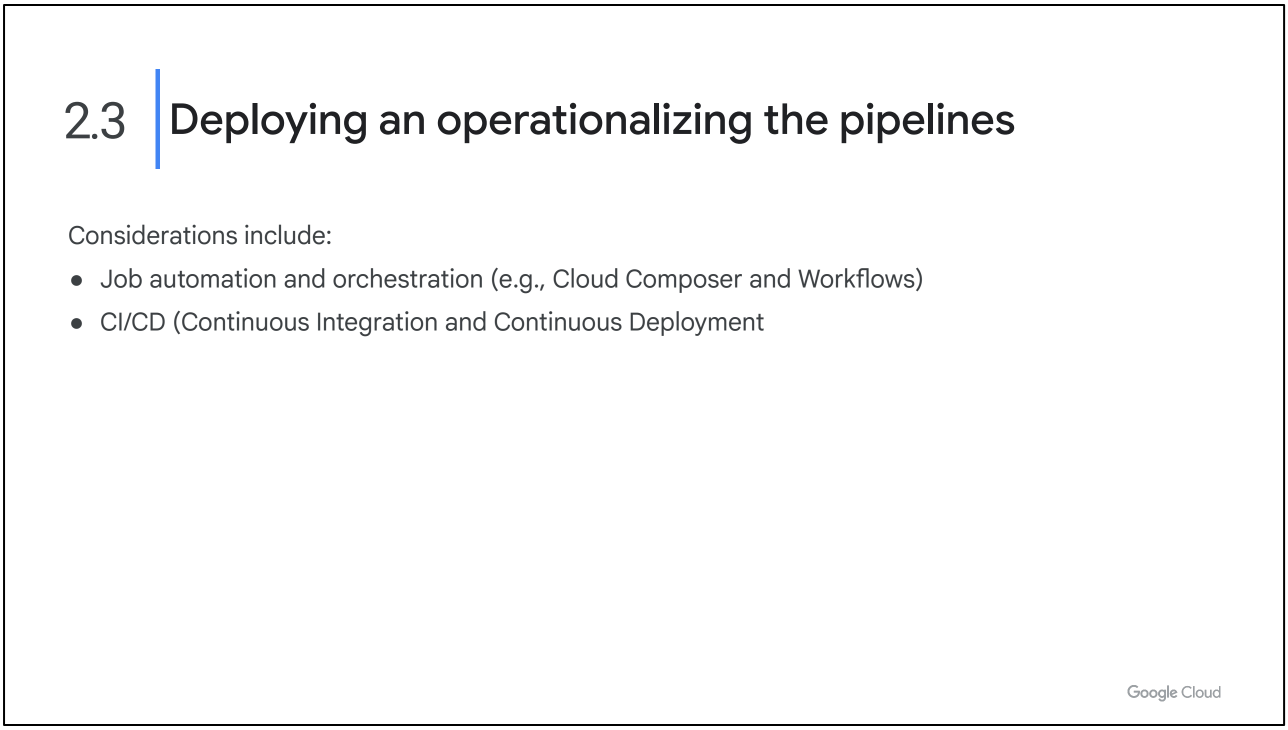
Streaming pipelines | Cloud Dataflow

Basics of the Beam model

Streaming analytics solutions | Google Cloud

The diagnostic questions we just reviewed explored some aspects of building data pipelines. 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/data-fusion/docs/concepts/overview
https://cloud.google.com/dataproc-serverless/docs/overview
https://www.youtube.com/watch?v=RS7UJhD4R48
https://cloud.google.com/batch/docs/get-started
https://cloud.google.com/dataflow/docs/concepts/streaming-pipelines
https://beam.apache.org/documentation/basics/#window
https://cloud.google.com/solutions/stream-analytics



2.3 Deploying an operationalizing the pipelines

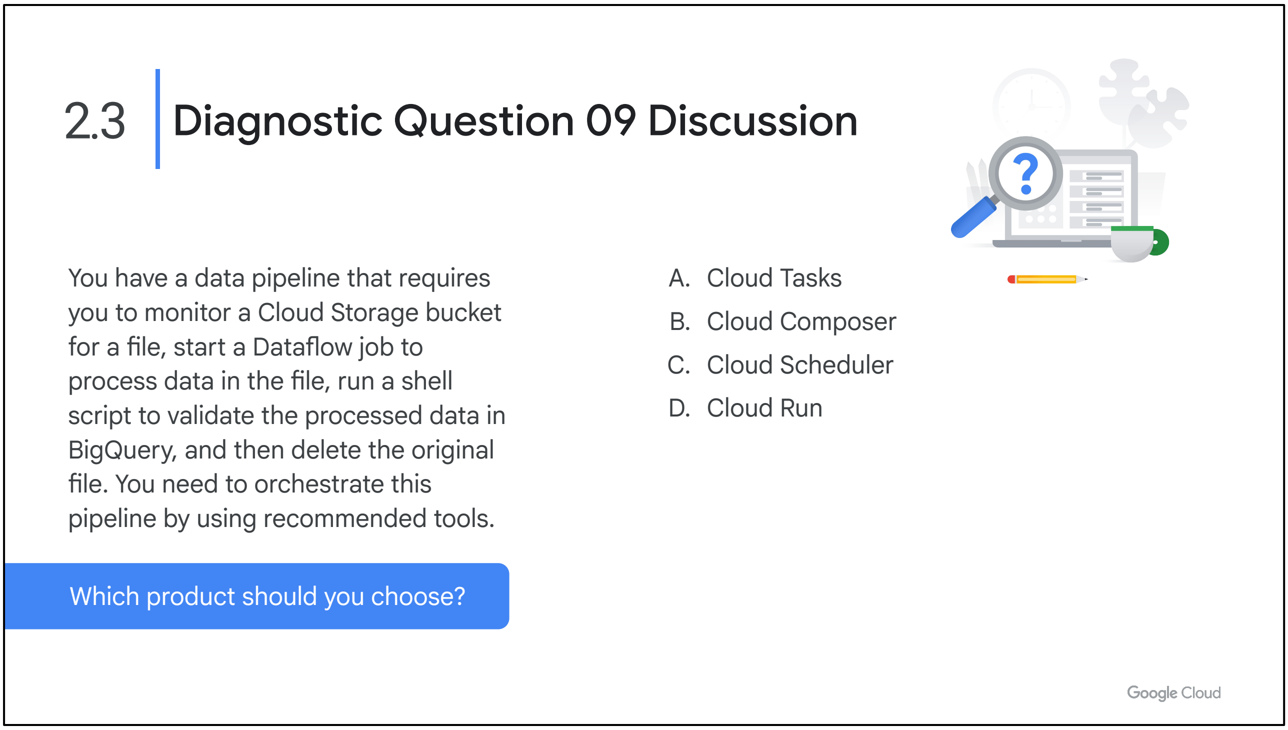
Considerations include:

- Job automation and orchestration (e.g., Cloud Composer and Workflows)
- CI/CD (Continuous Integration and Continuous Deployment

Google Cloud

Professional Data Engineers need to programmatically author, schedule, and monitor workflows. Cloud Composer, which is built on the Apache Airflow project, provides a single orchestration tool—whether your pipeline lives on-premises, in multiple clouds, or fully within Google Cloud. As a Professional Data Engineer, you need to architect and implement end-to-end continuous integration and continuous delivery pipelines.

Question 9 asked about tools for job and workflow orchestration on Google Cloud. Question 10 tested your knowledge of how CI/CD helps you to deploy and operationalize data pipelines.



2.3 Diagnostic Question 09 Discussion

You have a data pipeline that requires you to monitor a Cloud Storage bucket for a file, start a Dataflow job to process data in the file, run a shell script to validate the processed data in BigQuery, and then delete the original file. You need to orchestrate this pipeline by using recommended tools.

Which product should you choose?

A. Cloud Tasks
B. Cloud Composer
C. Cloud Scheduler
D. Cloud Run

Google Cloud

Feedback:

A. Incorrect. Cloud Tasks is for asynchronous task execution and is not the preferred tool for data pipeline orchestration.
B. Correct. Cloud Composer, a managed version of Apache Airflow, can orchestrate a series of data pipeline tasks.
C. Incorrect. Cloud Scheduler is a cron job service that can run applications at a scheduled time.
D. Incorrect. Cloud Run runs container applications in a serverless approach. It is not the appropriate tool to perform data pipeline orchestration.

Links:
https://youtu.be/3UfYwR3Uwgw?t=169
https://cloud.google.com/composer/docs/concepts/overview

More information:
Courses:
Build Batch Data Pipelines on Google Cloud

- Manage Data Pipelines with Cloud Data Fusion and Cloud Composer

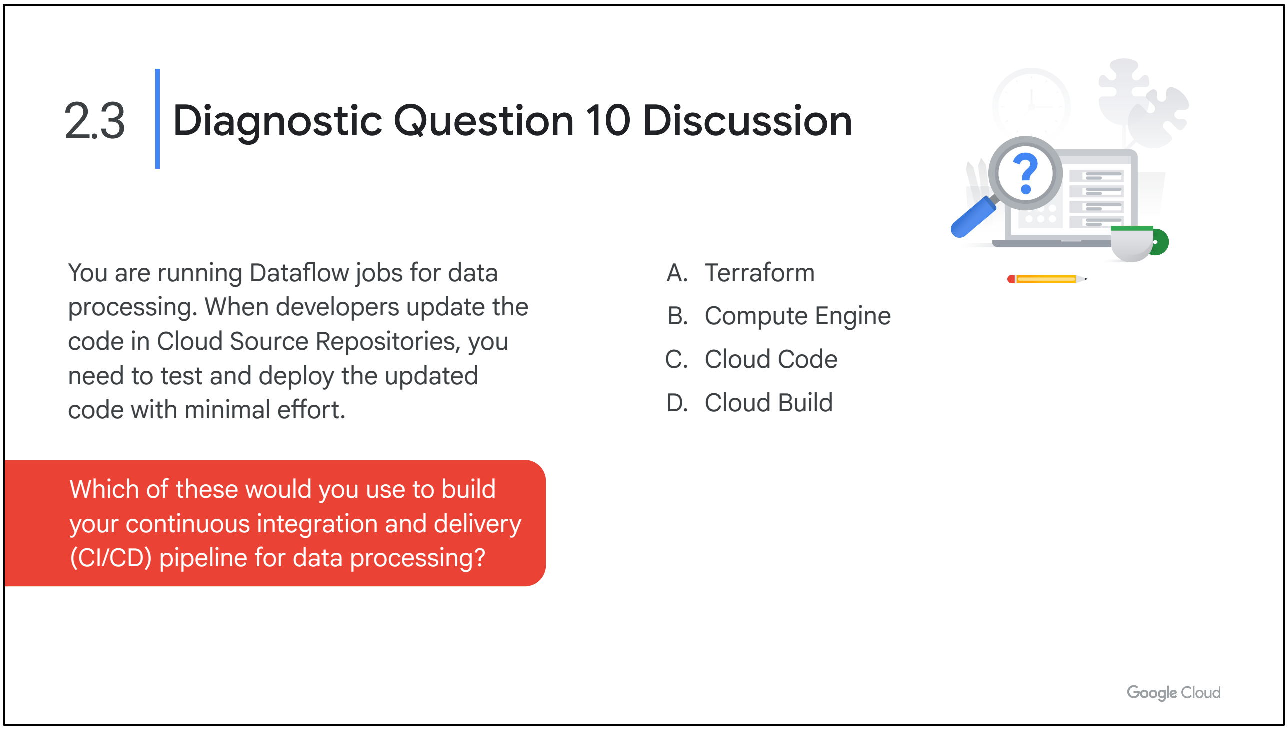
Serverless Data Processing with Dataflow: Operations

- Testing and CI/CD

Skill badge:

Engineer Data for Predictive Modeling with BigQuery ML

Summary:
Google Cloud offers multiple tools such as Dataproc, Dataflow, Data Fusion, and Dataprep. As a PDE, you need to know each product's capabilities and make appropriate choices. Cloud Composer is based on Apache Airflow. You can even create your pipelines with the open source Apache Airflow and execute them on Google Cloud by using Cloud Composer.



2.3 Diagnostic Question 10 Discussion

You are running Dataflow jobs for data processing. When developers update the code in Cloud Source Repositories, you need to test and deploy the updated code with minimal effort.

Which of these would you use to build your continuous integration and delivery (CI/CD) pipeline for data processing?

A. Terraform
B. Compute Engine
C. Cloud Code
D. Cloud Build

Google Cloud

Feedback:

A. Incorrect. Terraform is suitable for automated infrastructure management, but it is not a CI/CD solution.

B. Incorrect. Using Compute Engine requires significantly more effort to install required software and provision build resources.

C. Incorrect. Cloud Code integrates with popular integrated development environments (IDEs) to make it easier to work with Google Cloud. It is not a CI/CD or build system by itself.

D. Correct. Cloud Build can be configured to watch for updates in the source repository and trigger a series of steps, as required, to implement a CI/CD pipeline.

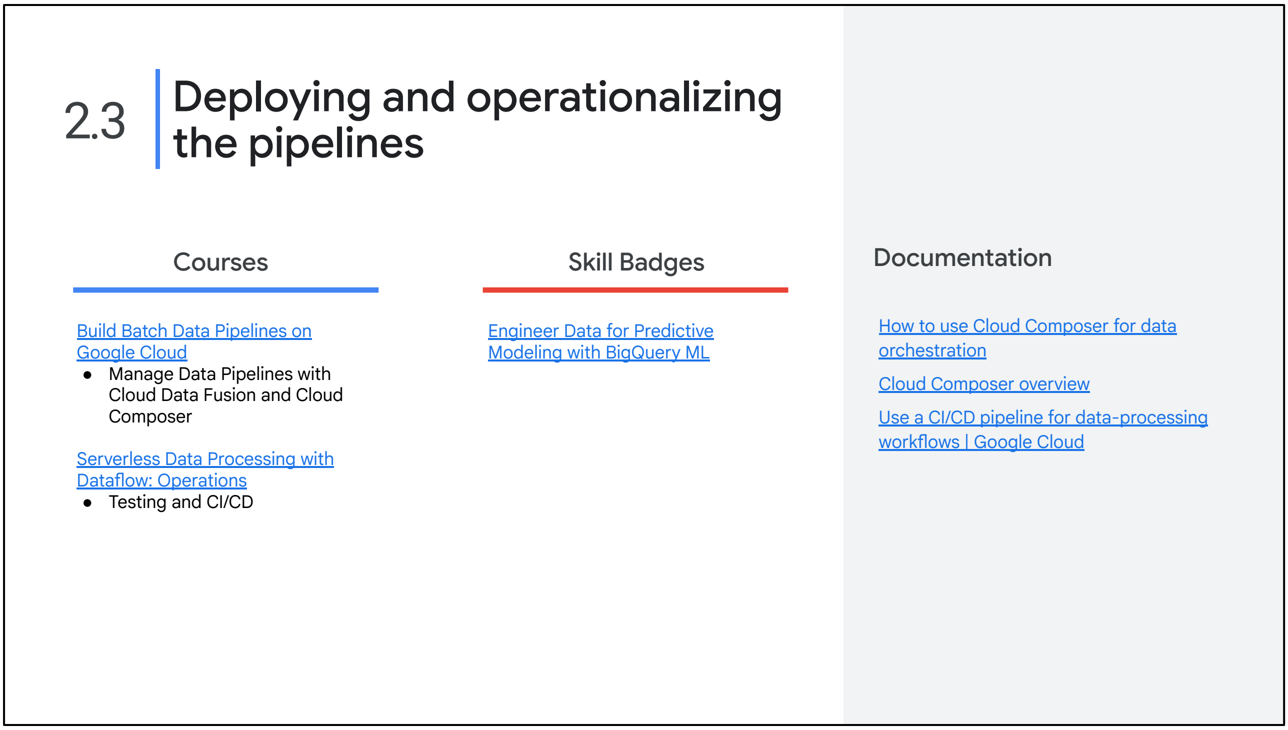
Links:
https://cloud.google.com/architecture/cicd-pipeline-for-data-processing

More information:
Courses:
Serverless Data Processing with Dataflow: Operations

- Testing and CI/CD

Summary:
A PDE must look for ways to automate systems and make them reliable. A manual process where people watch for changes and then take action is not scalable. Cloud

Build is integrated with code repositories and automatically allocates required resources to execute the steps of the CI/CD pipeline.



2.3 Deploying and operationalizing the pipelines

Courses

[Build Batch Data Pipelines on Google Cloud]

- Manage Data Pipelines with Cloud Data Fusion and Cloud Composer

[Serverless Data Processing with Dataflow: Operations]

- Testing and CI/CD

Skill Badges

[Engineer Data for Predictive Modeling with BigQuery ML]

Documentation

[How to use Cloud Composer for data orchestration]

[Cloud Composer overview]

[Use a CI/CD pipeline for data-processing workflows | Google Cloud]

You just reviewed diagnostic questions that addressed aspects of deploying and operationalizing your data pipelines. 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://youtu.be/3UfYwR3Uwgw?t=169
https://cloud.google.com/composer/docs/concepts/overview
https://cloud.google.com/architecture/cicd-pipeline-for-data-processing