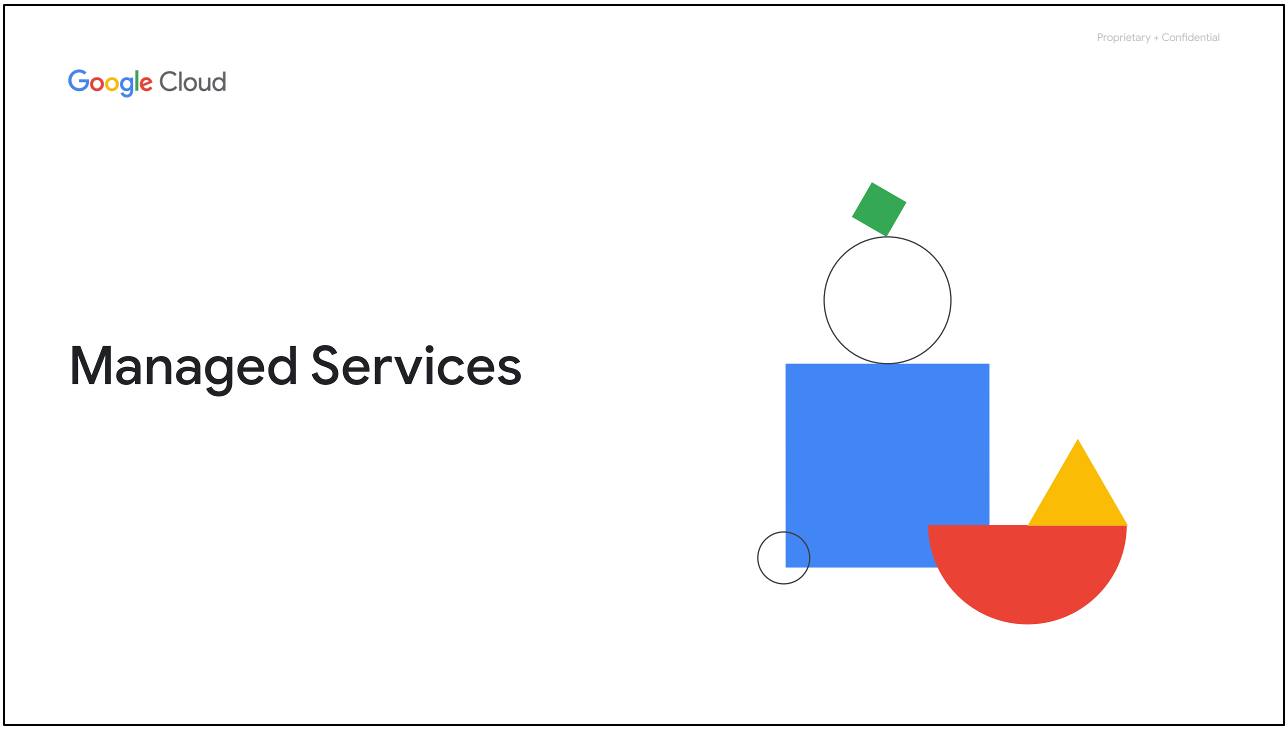


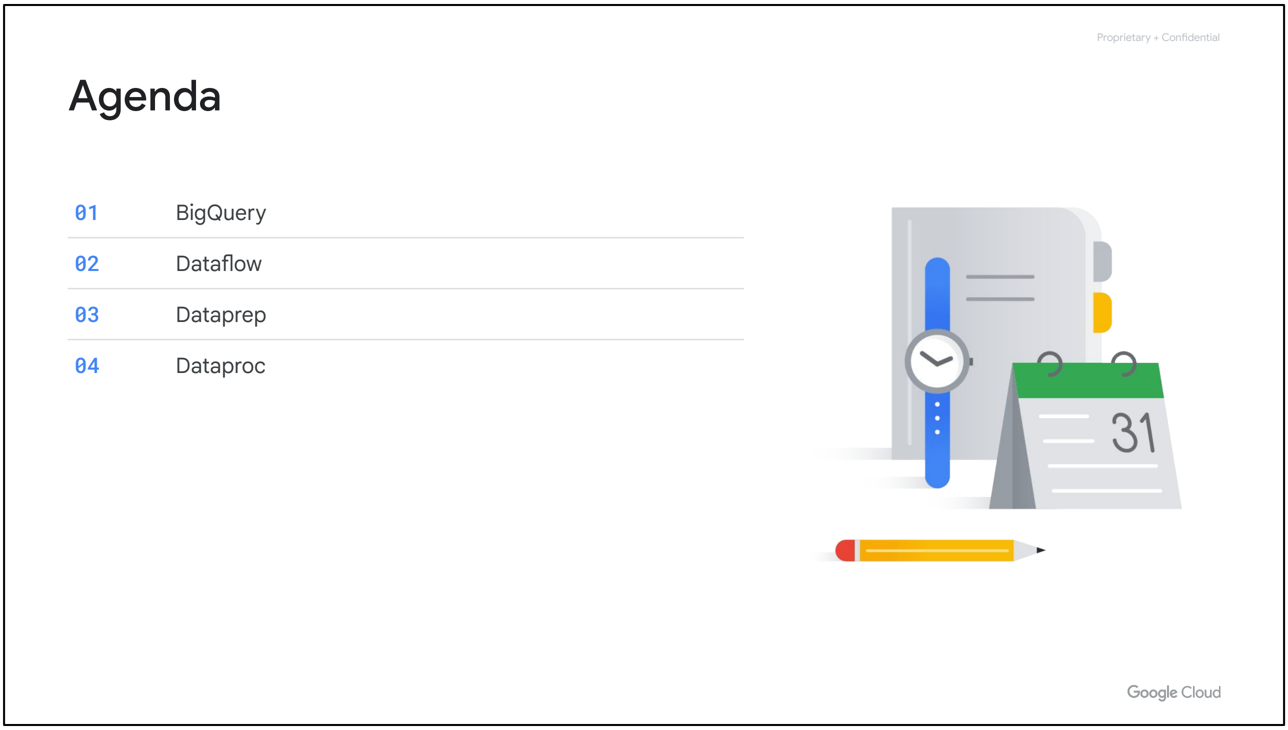
Proprietary + Confidential

Google Cloud

Managed Services

In the last module, we discussed how to automate the creation of infrastructure. As an alternative to infrastructure automation, you can eliminate the need to create infrastructure by leveraging a managed service.

Managed services are partial or complete solutions offered as a service. They exist on a continuum between platform as a service and software as a service, depending on how much of the internal methods and controls are exposed. Using a managed service allows you to outsource a lot of the administrative and maintenance overhead to Google, if your application requirements fit within the service offering.



Agenda

01	BigQuery
02	Dataflow
03	Dataprep
04	Dataproc

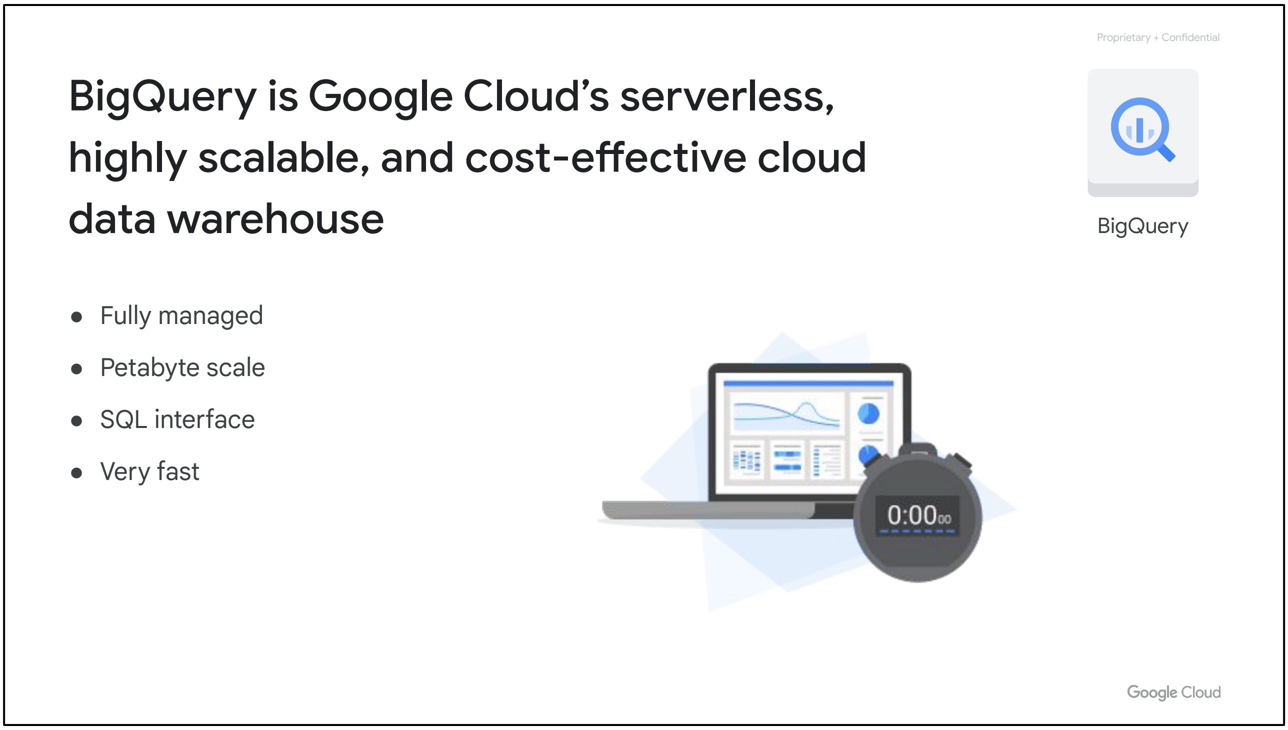
Google Cloud

In this module, we give you an overview of BigQuery, Dataflow, Dataprep by Trifacta, and Dataproc. Now all of these services are for data analytics purposes, and since that’s not the focus of this course, there won’t be any labs in this module. Instead, we’ll have a quick demo to illustrate how easy it is to use managed services.

Let’s start by talking about BigQuery.

01

BigQuery



Proprietary + Confidential

BigQuery is Google Cloud’s serverless, highly scalable, and cost-effective cloud data warehouse

BigQuery

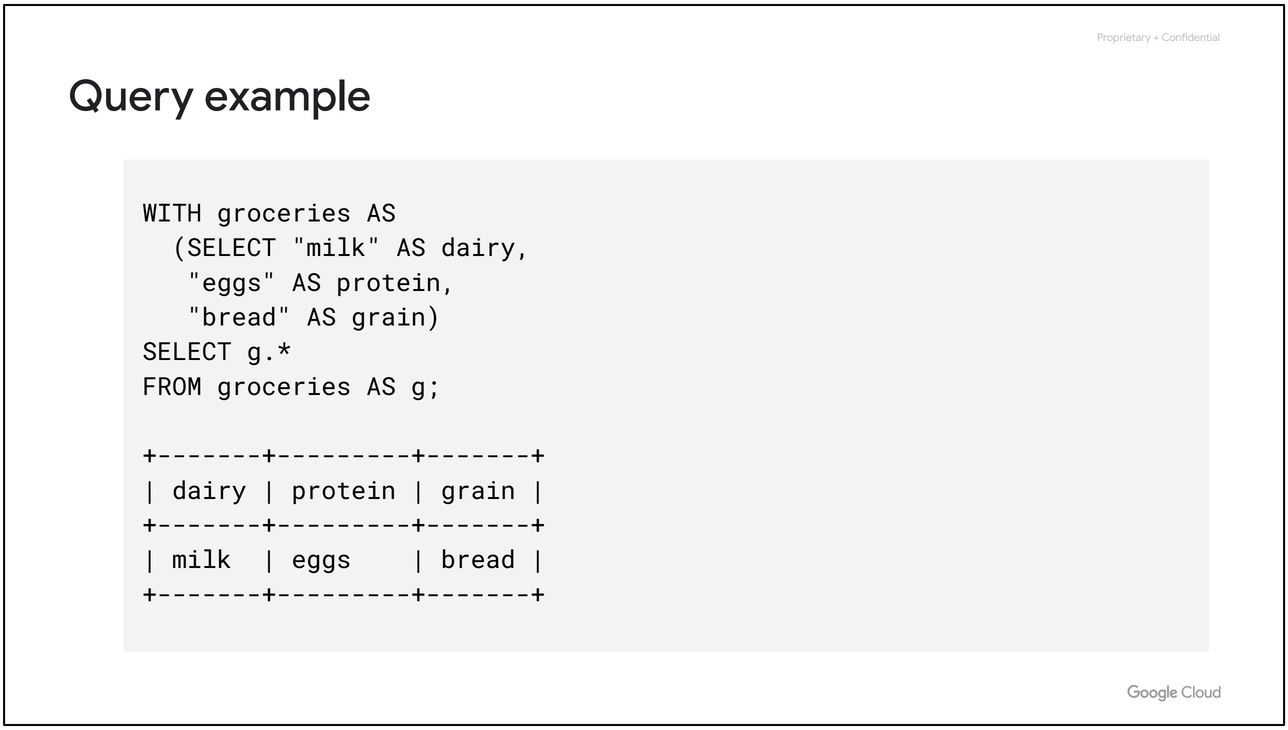
- Fully managed
- Petabyte scale
- SQL interface
- Very fast

Google Cloud

BigQuery is Google Cloud’s serverless, highly scalable, and cost-effective cloud data warehouse.

It is a petabyte-scale data warehouse that allows for super-fast queries using the processing power of Google's infrastructure. Because there is no infrastructure for you to manage, you can focus on uncovering meaningful insights using familiar SQL without the need for a database administrator.

BigQuery is used by all types of organizations.



Proprietary + Confidential

Query example

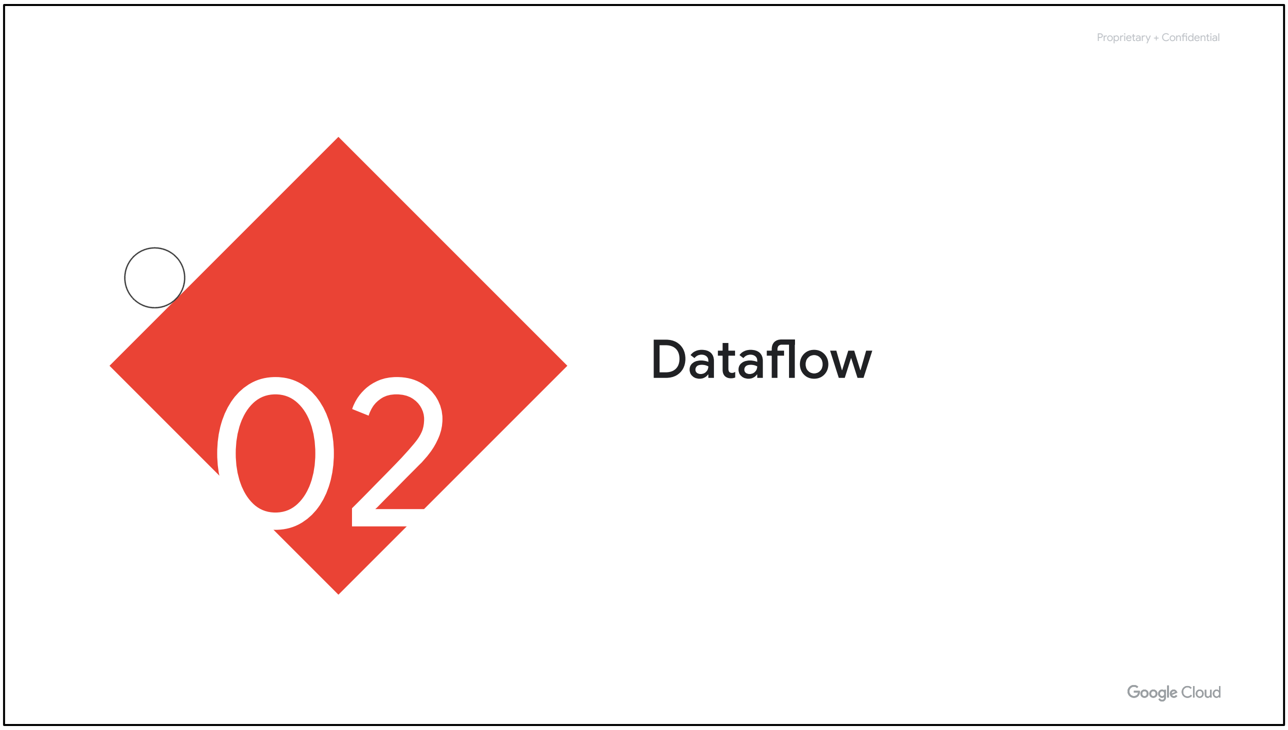
```
WITH groceries AS
  (SELECT "milk" AS dairy,
   "eggs" AS protein,
   "bread" AS grain)
SELECT g.*
FROM groceries AS g;

+-------+---------+-------+
| dairy | protein | grain |
+-------+---------+-------+
| milk  | eggs    | bread |
+-------+---------+-------+
```

Google Cloud

You can access BigQuery by using the Google Cloud console, by using a command-line tool, or by making calls to the BigQuery REST API using a variety of client libraries such as Java, .NET, or Python. There are also several third-party tools that you can use to interact with BigQuery, such as visualizing the data or loading the data.

Here is an example of a Standard SQL query on a table called groceries. This query produces one output column for each column in the table groceries, aliased as g.

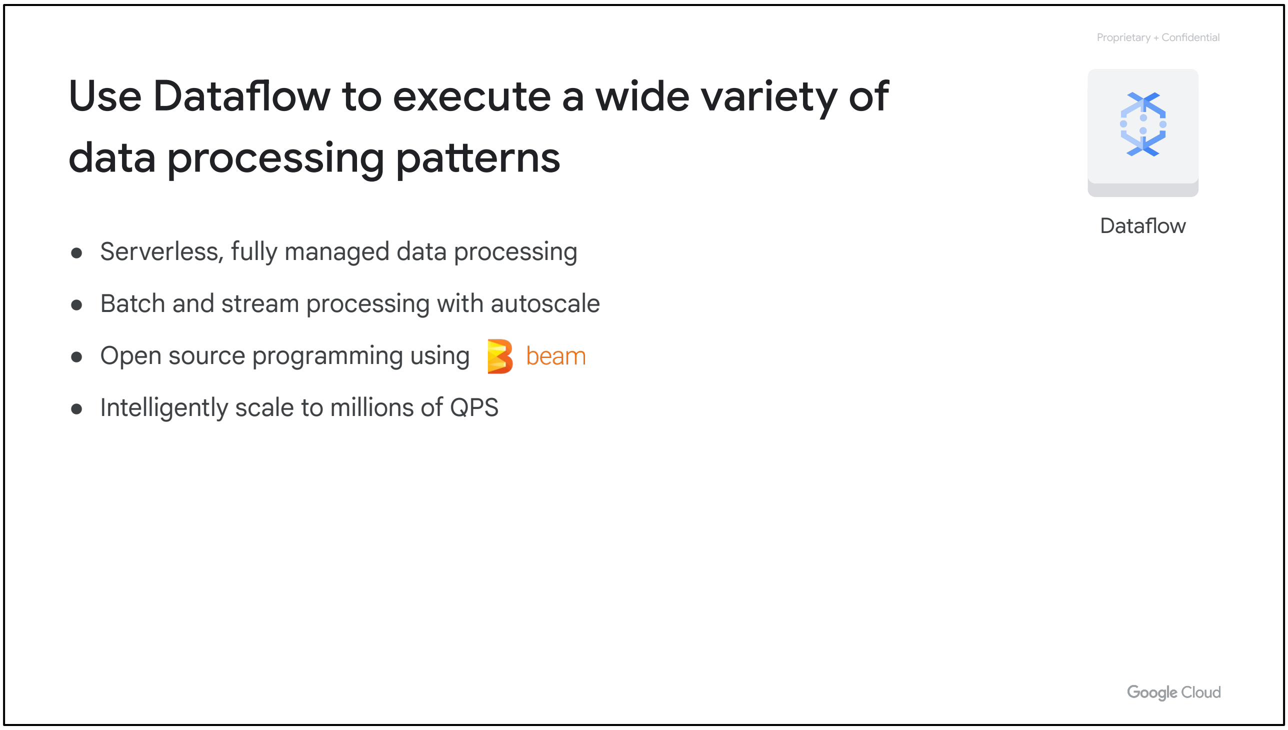


02

Dataflow

Google Cloud

Let’s learn a little bit about Dataflow.



Proprietary + Confidential

Use Dataflow to execute a wide variety of data processing patterns

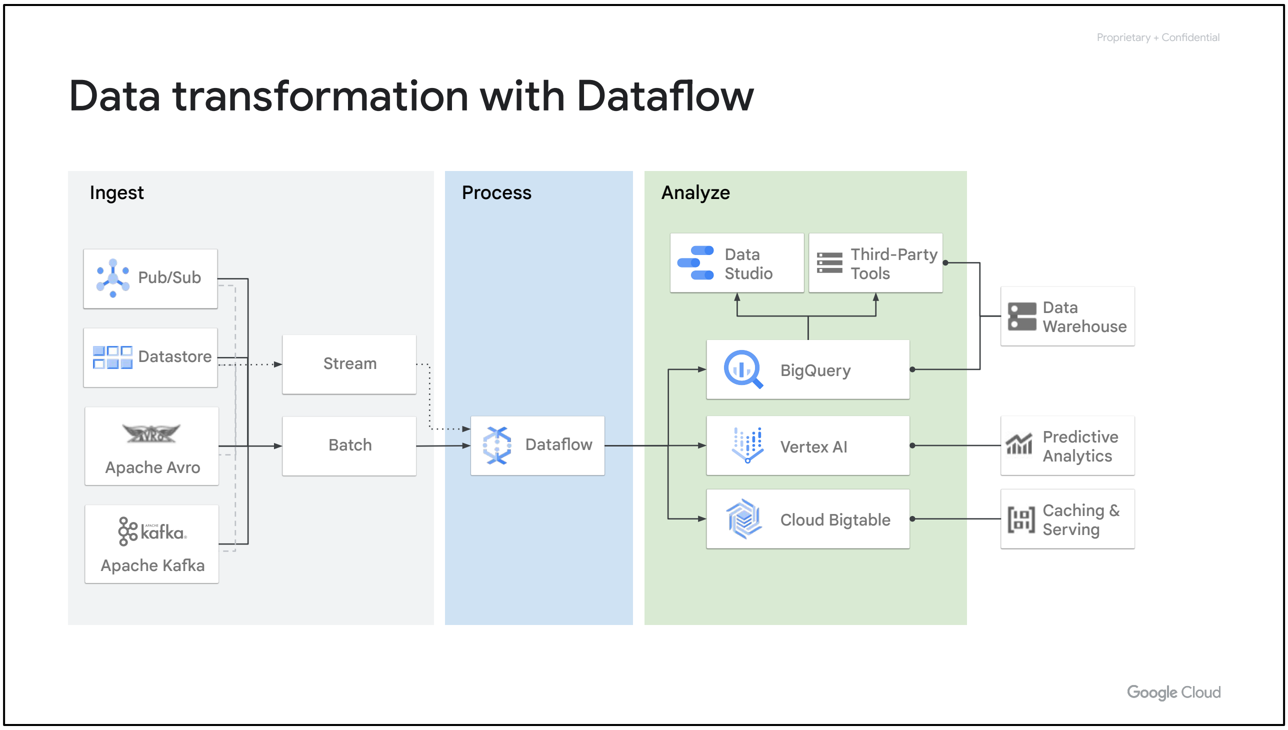
Dataflow

- Serverless, fully managed data processing
- Batch and stream processing with autoscale
- Open source programming using beam
- Intelligently scale to millions of QPS

Google Cloud

Dataflow is a managed service for executing a wide variety of data processing patterns. It’s essentially a fully managed service for transforming and enriching data in stream and batch modes with equal reliability and expressiveness. With Dataflow, a lot of the complexity of infrastructure setup and maintenance is handled for you. It’s built on Google Cloud infrastructure and autoscales to meet the demands of your data pipelines, allowing it to intelligently scale to millions of queries per second.

Dataflow supports fast, simplified pipeline development via expressive SQL, Java, and Python APIs in the Apache Beam SDK, which provides a rich set of windowing and session analysis primitives as well as an ecosystem of source and sink connectors. Dataflow is also tightly coupled with other Google Cloud services like Google Cloud’s operations suite, so you can set up priority alerts and notifications to monitor your pipeline and the quality of data coming in and out.

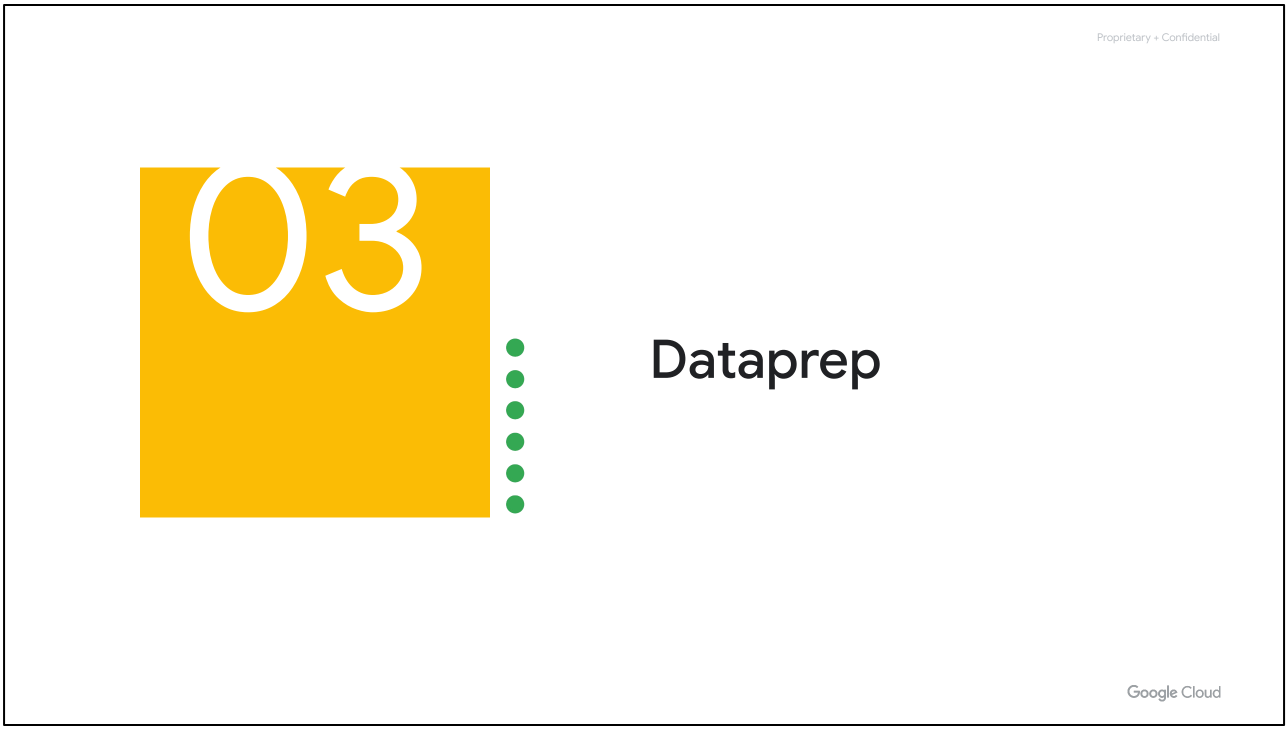


Proprietary + Confidential

Data transformation with Dataflow

This diagram shows some example uses cases of Dataflow. As we just mentioned, Dataflow processes stream and batch data. This data could come from other Google Cloud services like Datastore or Pub/Sub, which is Google's messaging and publishing service. The data could also be ingested from third-party services like Apache Avro and Apache Kafka.

After you transform the data with Dataflow, you can analyze it in BigQuery, Vertex AI, or even Cloud Bigtable. Using Data Studio, you can even build real-time dashboards for IoT devices.

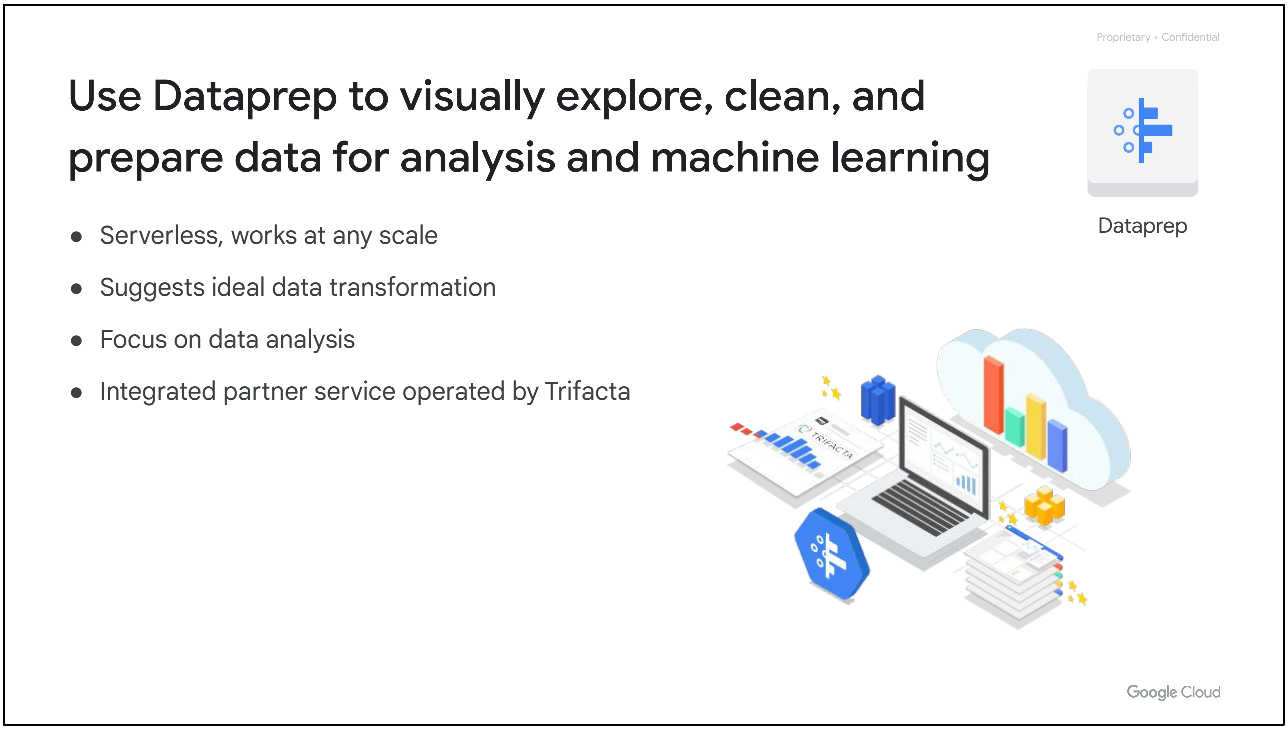


03

Dataprep

Google Cloud

Let’s learn a little bit about Dataprep.



Proprietary + Confidential

Use Dataprep to visually explore, clean, and prepare data for analysis and machine learning

- Serverless, works at any scale
- Suggests ideal data transformation
- Focus on data analysis
- Integrated partner service operated by Trifacta

Dataprep

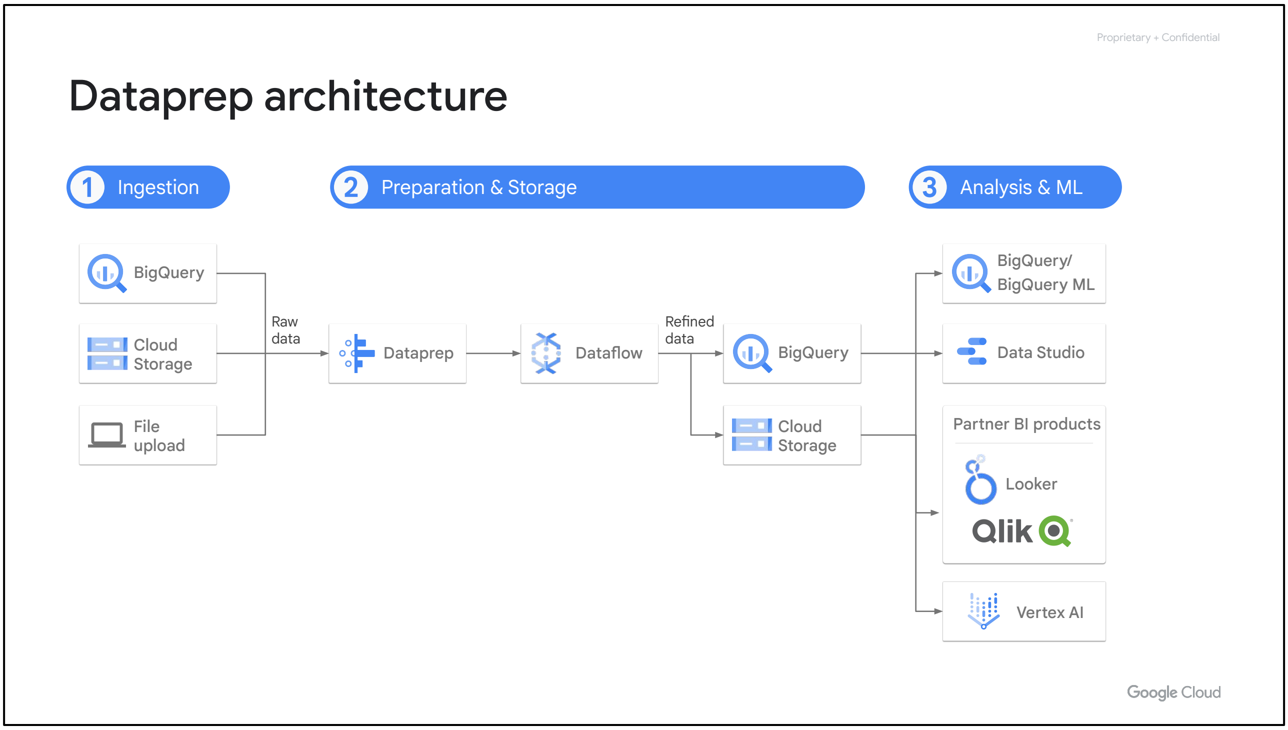
Google Cloud

Dataprep is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis, reporting, and machine learning.

Because Dataprep is serverless and works at any scale, there is no infrastructure to deploy or manage. Your next ideal data transformation is suggested and predicted with each UI input, so you don't have to write code.

With automatic schema, datatype, possible joins, and anomaly detection, you can skip time-consuming data profiling and focus on data analysis.

Dataprep is an integrated partner service operated by Trifacta and based on their industry-leading data preparation solution, Trifacta Wrangler. Google works closely with Trifacta to provide a seamless user experience that removes the need for up-front software installation, separate licensing costs, or ongoing operational overhead. Dataprep is fully managed and scales on demand to meet your growing data preparation needs, so you can stay focused on analysis.



Proprietary + Confidential

Dataprep architecture

1 Ingestion

2 Preparation & Storage

3 Analysis & ML

BigQuery

Cloud Storage

File upload

Raw data

Dataprep

Dataflow

Refined data

BigQuery

Cloud Storage

BigQuery/ BigQuery ML

Data Studio

Partner BI products

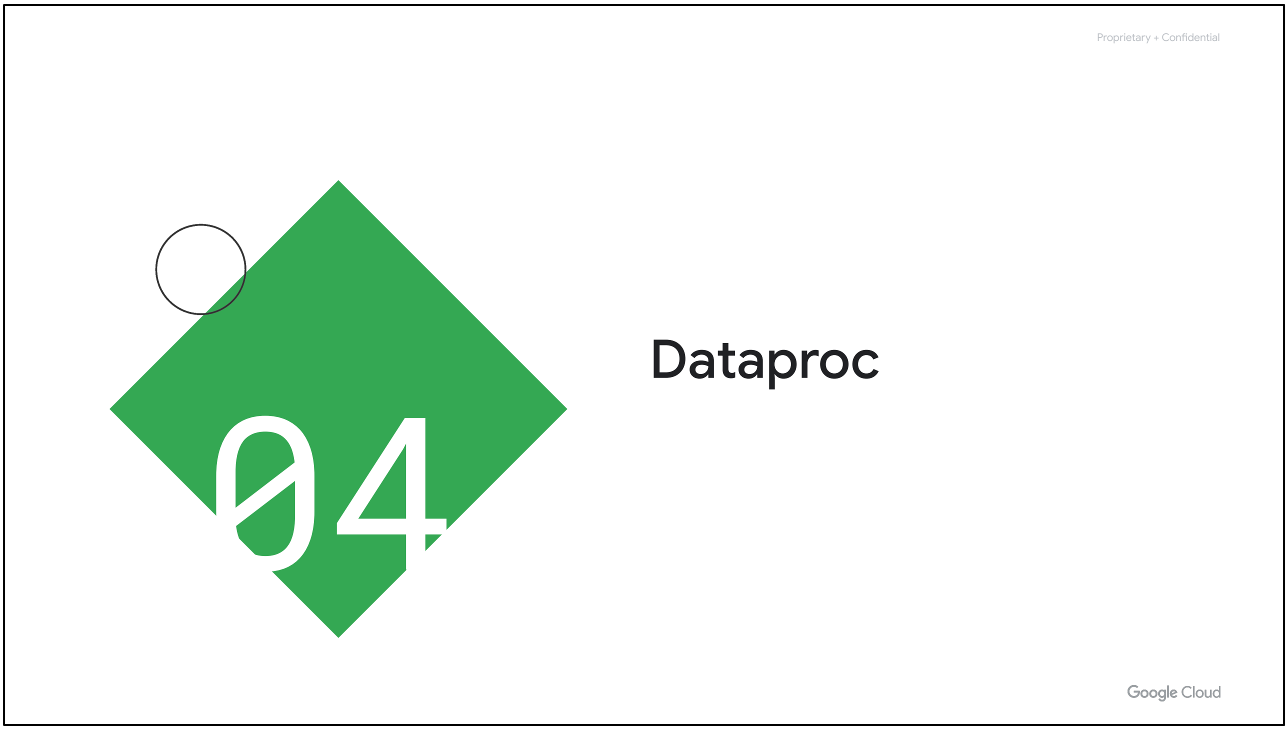
Looker

Qlik

Vertex AI

Google Cloud

Here's an example of a Dataprep architecture. As you can see, Dataprep can be leveraged to prepare raw data from BigQuery, Cloud Storage, or a file upload before ingesting it onto a transformation pipeline like Dataflow. The refined data can then be exported to BigQuery or Cloud Storage for analysis and machine learning.



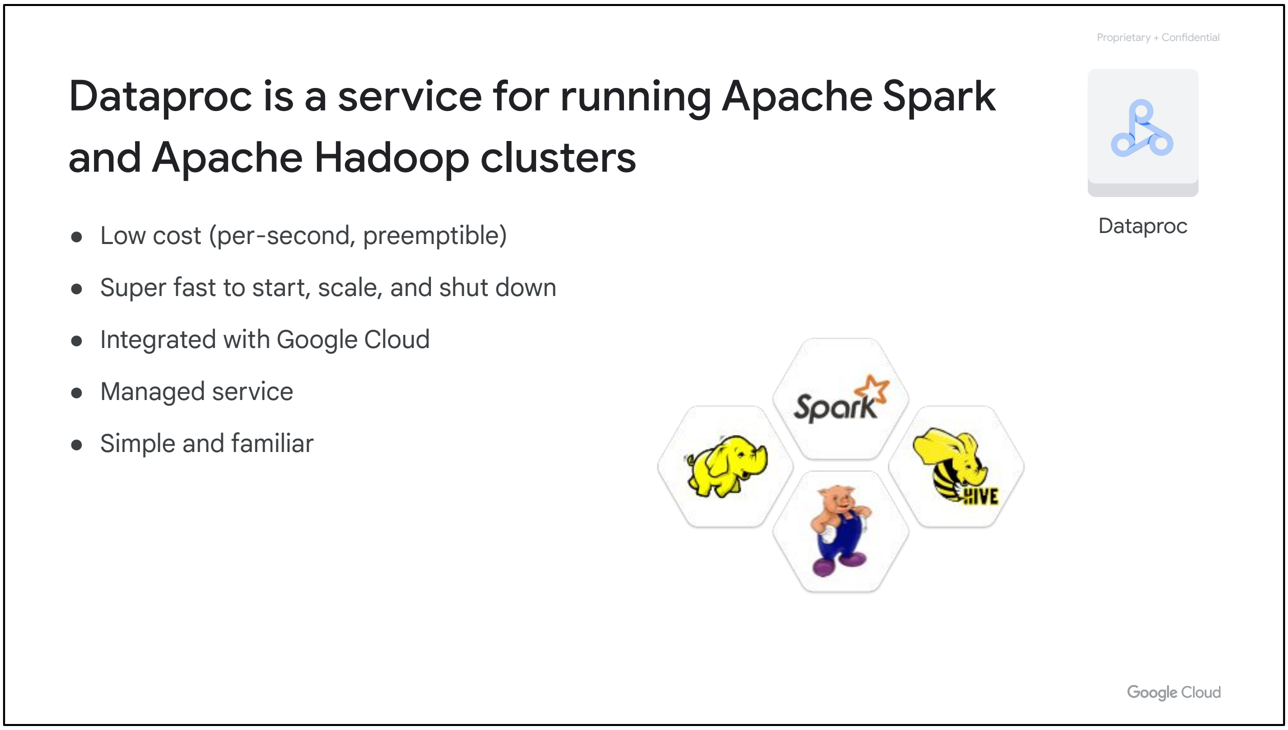
Proprietary + Confidential

04

Dataproc

Google Cloud

Let’s learn a little bit about Dataproc.



Dataproc is a service for running Apache Spark and Apache Hadoop clusters

Dataproc

- Low cost (per-second, preemptible)
- Super fast to start, scale, and shut down
- Integrated with Google Cloud
- Managed service
- Simple and familiar

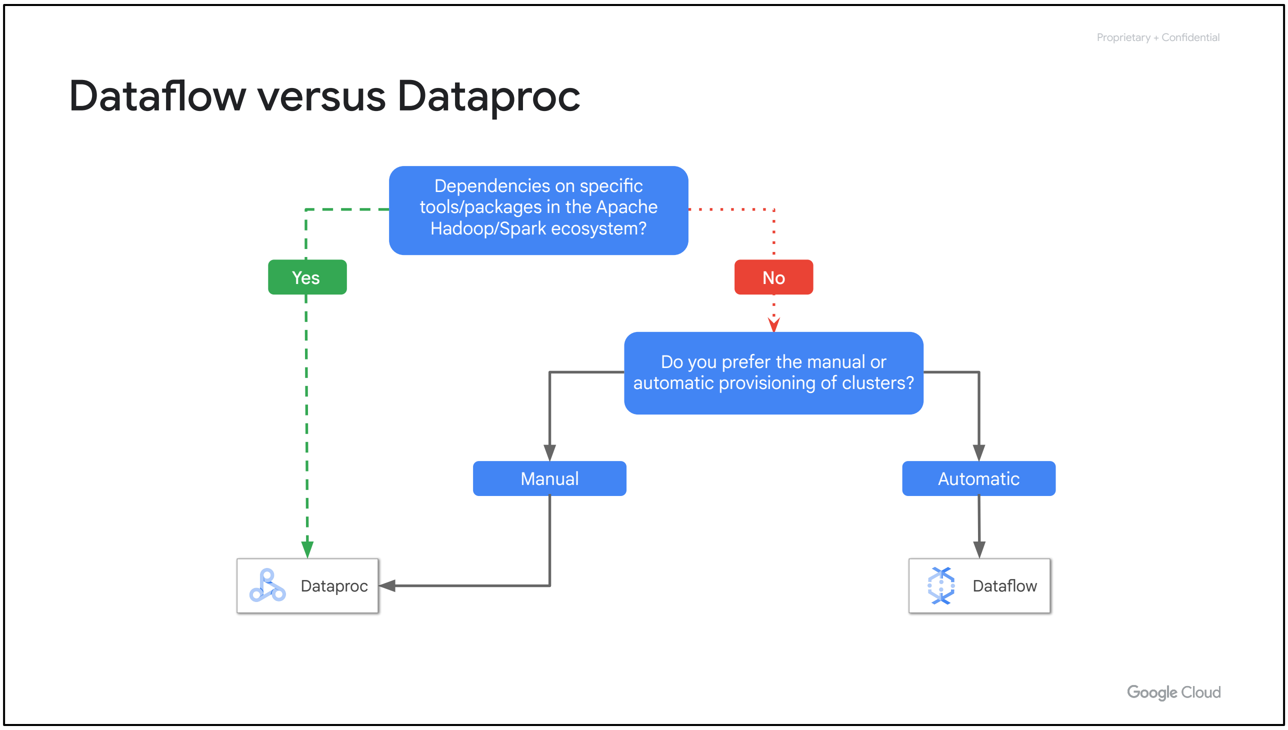
Dataproc is a fast, easy-to-use, fully managed cloud service for running Apache Spark and Apache Hadoop clusters in a simpler way. You only pay for the resources you use with per-second billing. If you leverage preemptible instances in your cluster, you can reduce your costs even further.

Without using Dataproc, it can take from five to 30 minutes to create Spark and Hadoop clusters on-premises or through other Infrastructure-as-a-Service providers. Dataproc clusters are quick to start, scale, and shut down, with each of these operations taking 90 seconds or less, on average. This means you can spend less time waiting for clusters and more hands-on time working with your data.

Dataproc has built-in integration with other Google Cloud services, such as BigQuery, Cloud Storage, Cloud Bigtable, Cloud Logging, and Cloud Monitoring. This provides you with a complete data platform rather than just a Spark or Hadoop cluster.

As a managed service, you can create clusters quickly, manage them easily, and save money by turning clusters off when you don't need them. With less time and money spent on administration, you can focus on your jobs and your data.

If you’re already using Spark, Hadoop, Pig, or Hive, you don’t even need to learn new tools or APIs to use Dataproc. This makes it easy to move existing projects into Dataproc without redevelopment.



Dataflow versus Dataproc

Dependencies on specific tools/packages in the Apache Hadoop/Spark ecosystem?

Yes

No

Do you prefer the manual or automatic provisioning of clusters?

Manual

Automatic

Dataproc

Dataflow

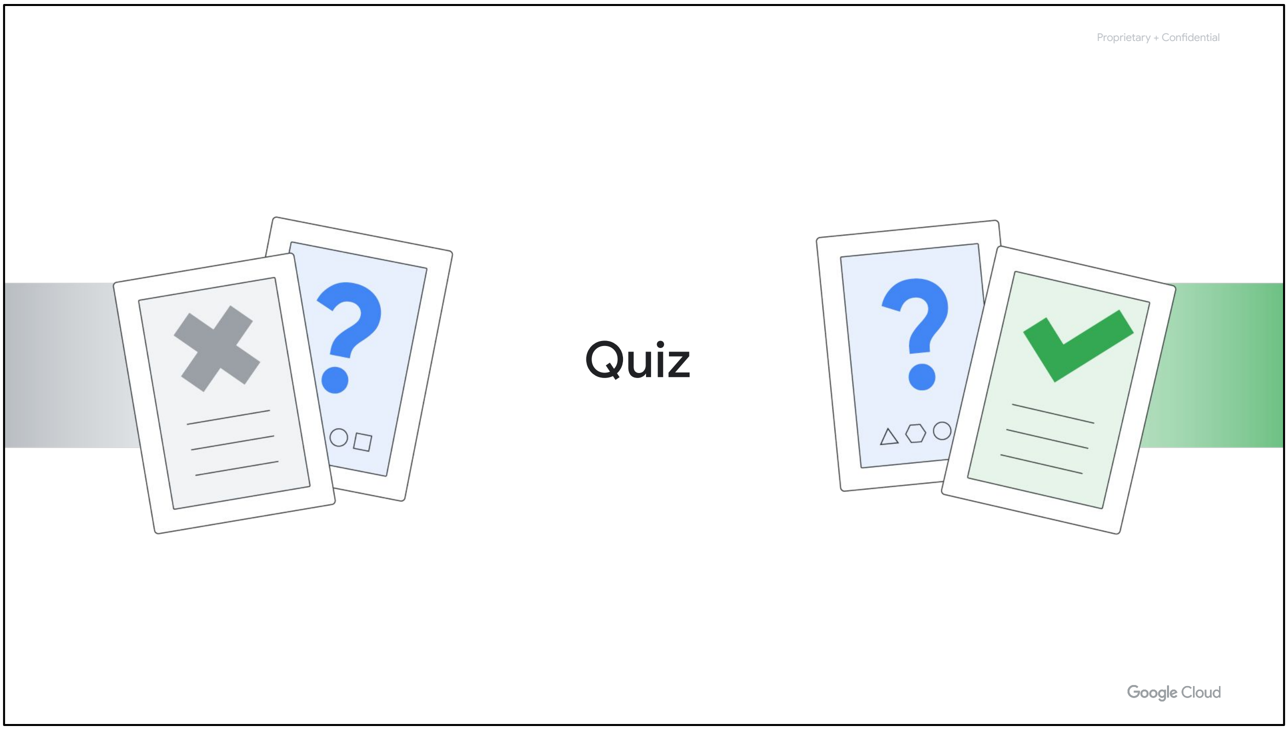
Google Cloud

Now, Dataproc and Dataflow can both be used for data processing, and there’s overlap in their batch and streaming capabilities. So, how do you decide which product is a better fit for your environment?

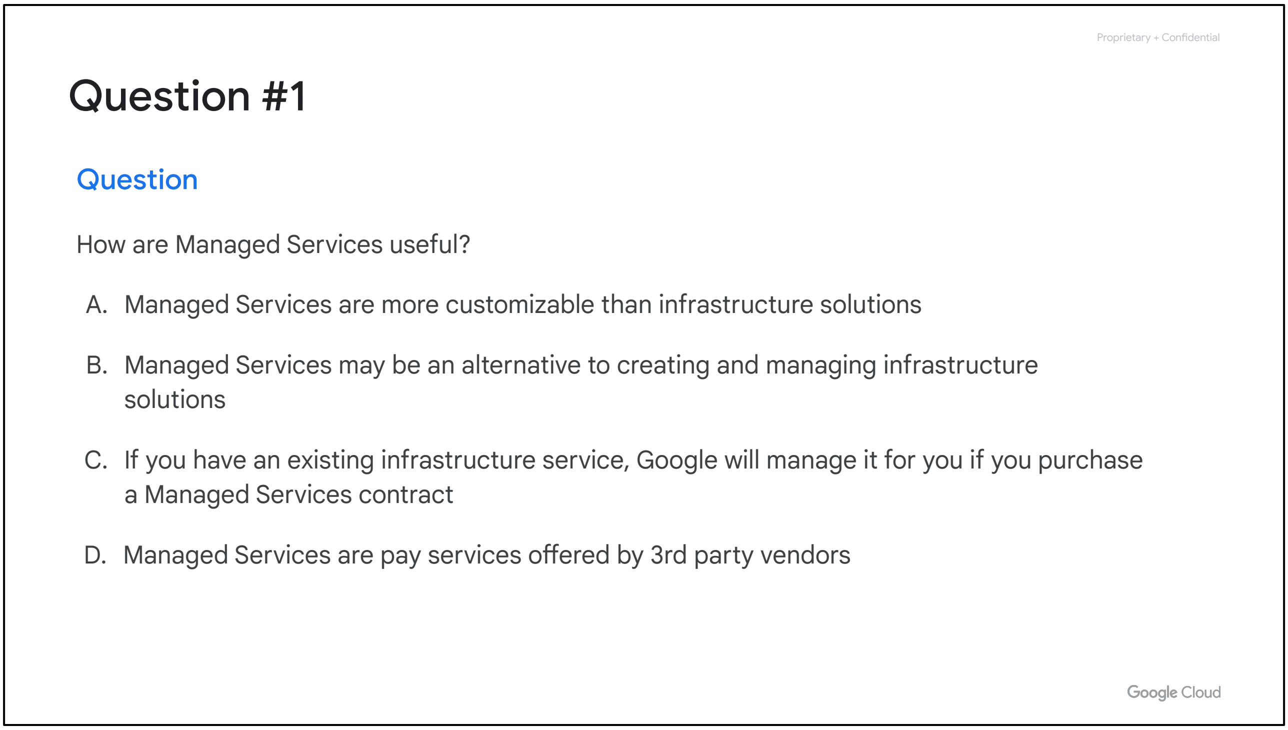
Well, first, ask yourself whether you have dependencies on specific tools or packages in the Apache Hadoop or Spark ecosystem. If that’s the case, you’ll obviously want to use Dataproc.

If not, ask yourself whether you prefer-manual provisioning of clusters, in this case you would choose Dataproc. If you prefer Serverless, automatic provisioning of clusters, than choose Dataflow.

For quick walkthrough on how to create a Dataproc cluster, modify the number of workers in the cluster, and submit a simple Apache Spark job, refer to this video.



Quiz



Question #1

Question

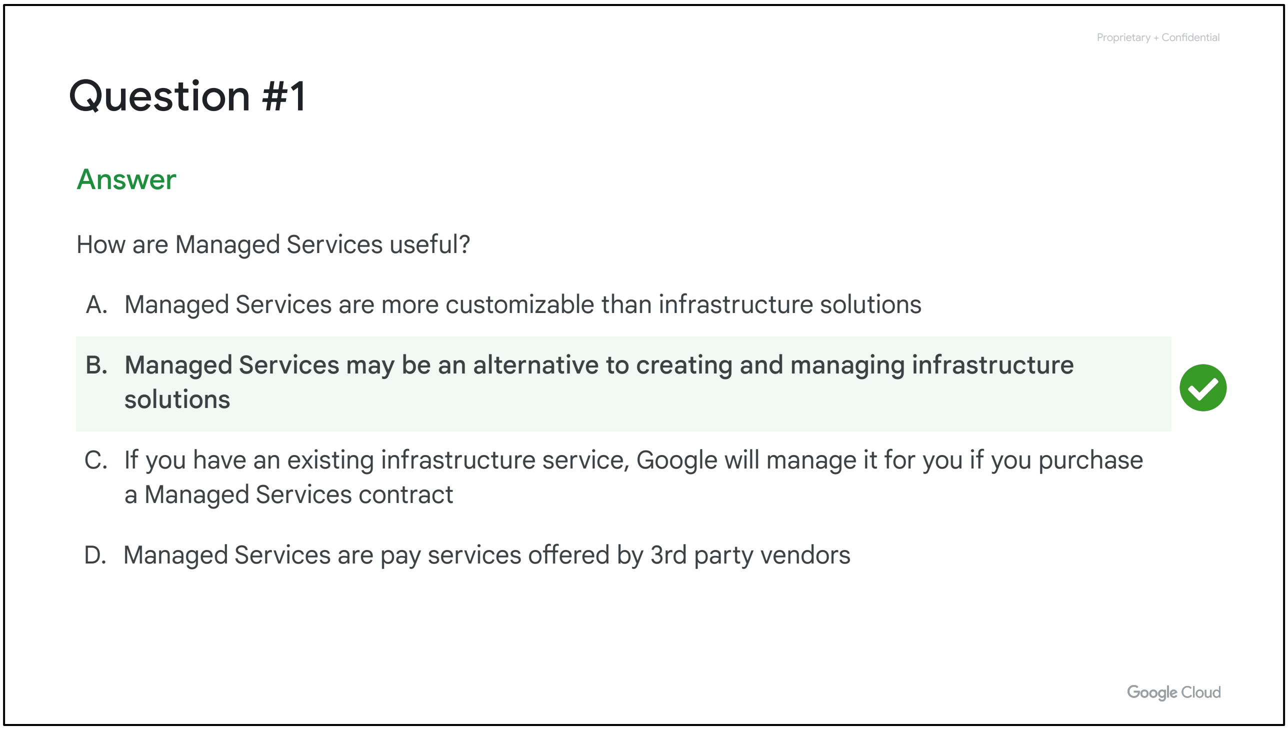
How are Managed Services useful?

A. Managed Services are more customizable than infrastructure solutions

B. Managed Services may be an alternative to creating and managing infrastructure solutions

C. If you have an existing infrastructure service, Google will manage it for you if you purchase a Managed Services contract

D. Managed Services are pay services offered by 3rd party vendors



Question #1

Answer

How are Managed Services useful?

A. Managed Services are more customizable than infrastructure solutions

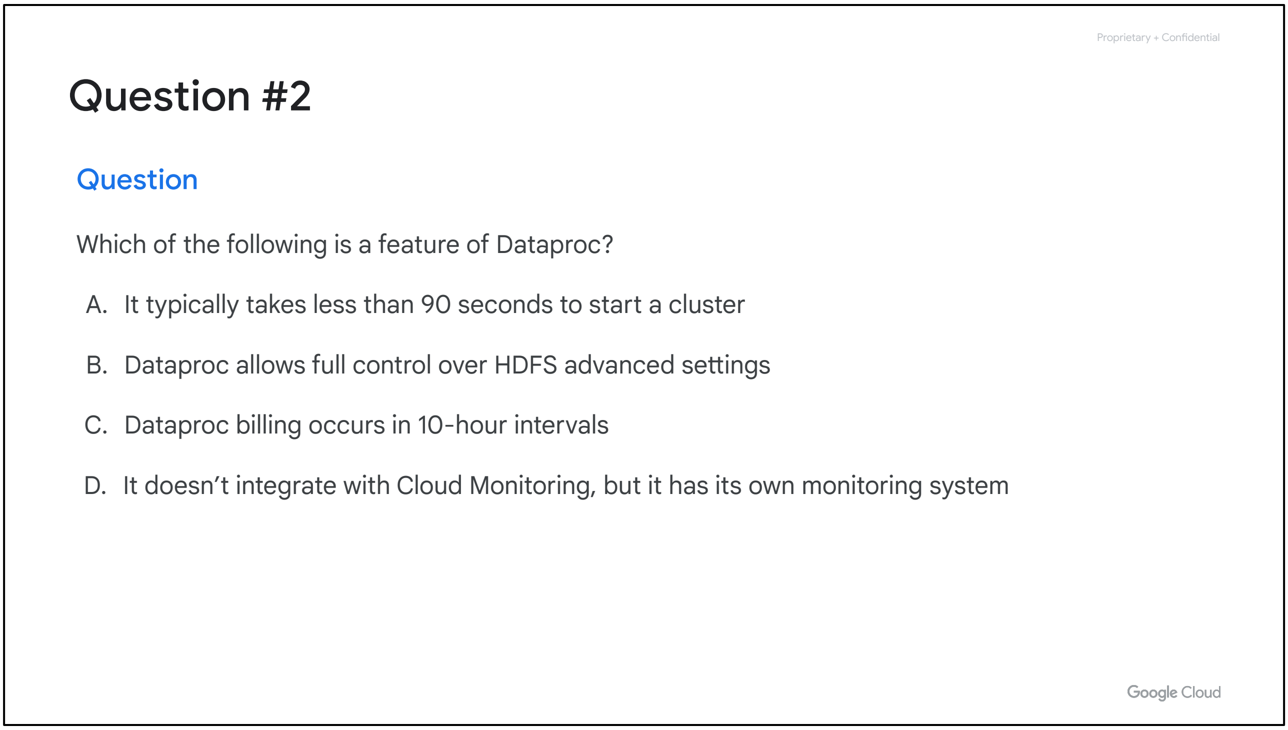
B. Managed Services may be an alternative to creating and managing infrastructure solutions ✓

C. If you have an existing infrastructure service, Google will manage it for you if you purchase a Managed Services contract

D. Managed Services are pay services offered by 3rd party vendors

Google Cloud

Explanation:
Managed Services in this class are presented as a possible alternative to building your own infrastructure data processing solution.



Question #2

Question

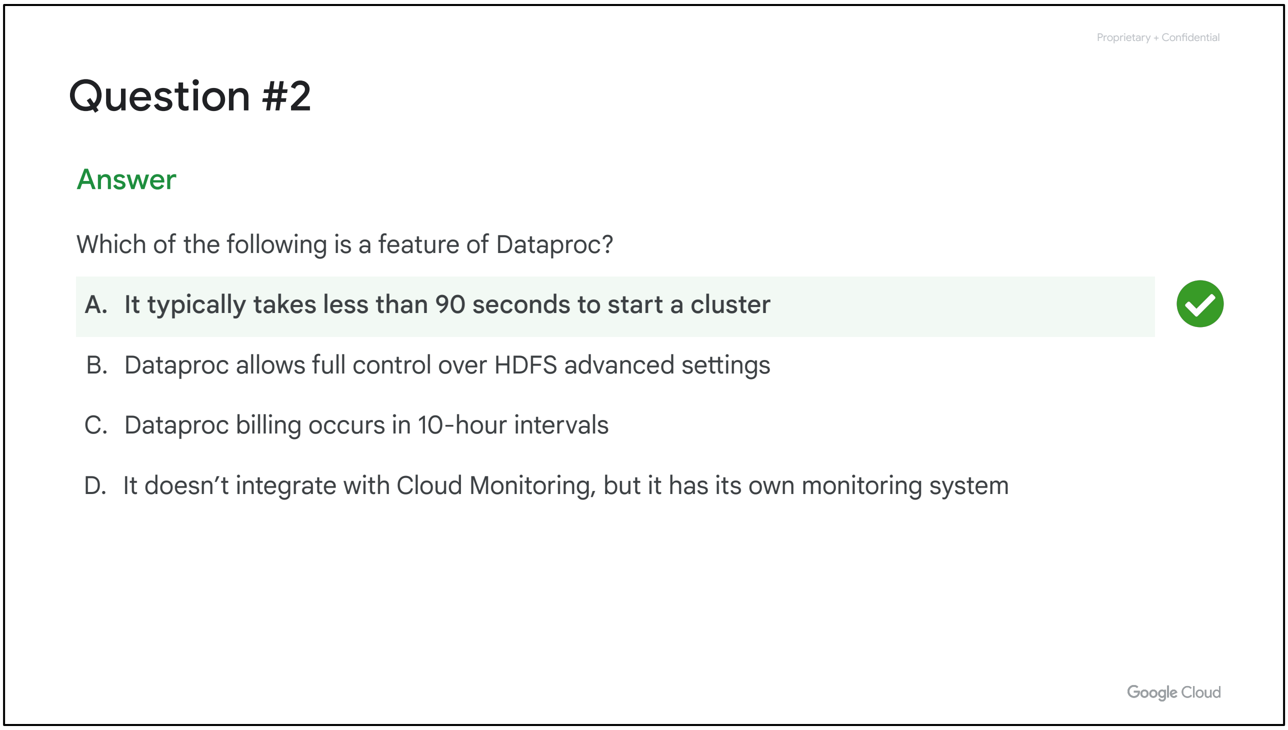
Which of the following is a feature of Dataproc?

A. It typically takes less than 90 seconds to start a cluster

B. Dataproc allows full control over HDFS advanced settings

C. Dataproc billing occurs in 10-hour intervals

D. It doesn’t integrate with Cloud Monitoring, but it has its own monitoring system



Question #2

Answer

Which of the following is a feature of Dataproc?

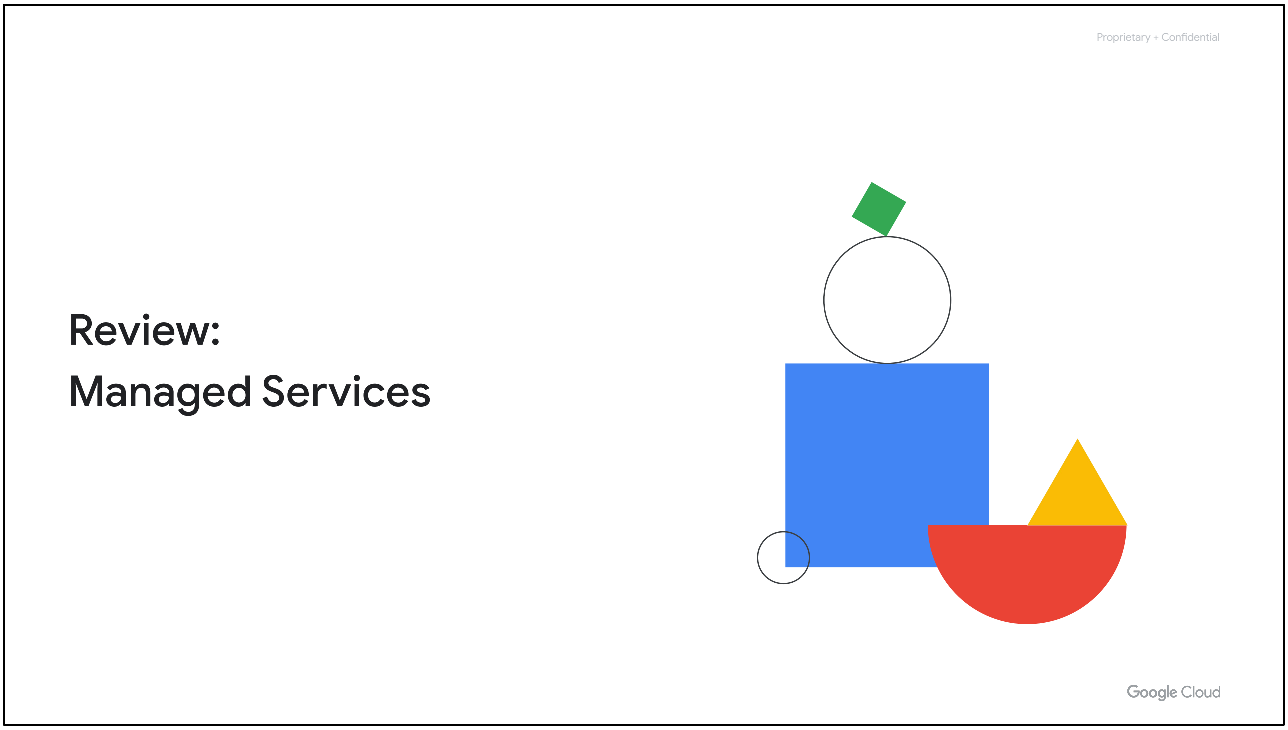
A. **It typically takes less than 90 seconds to start a cluster** ✓

B. Dataproc allows full control over HDFS advanced settings

C. Dataproc billing occurs in 10-hour intervals

D. It doesn’t integrate with Cloud Monitoring, but it has its own monitoring system

Explanation:
Fast to start a cluster.

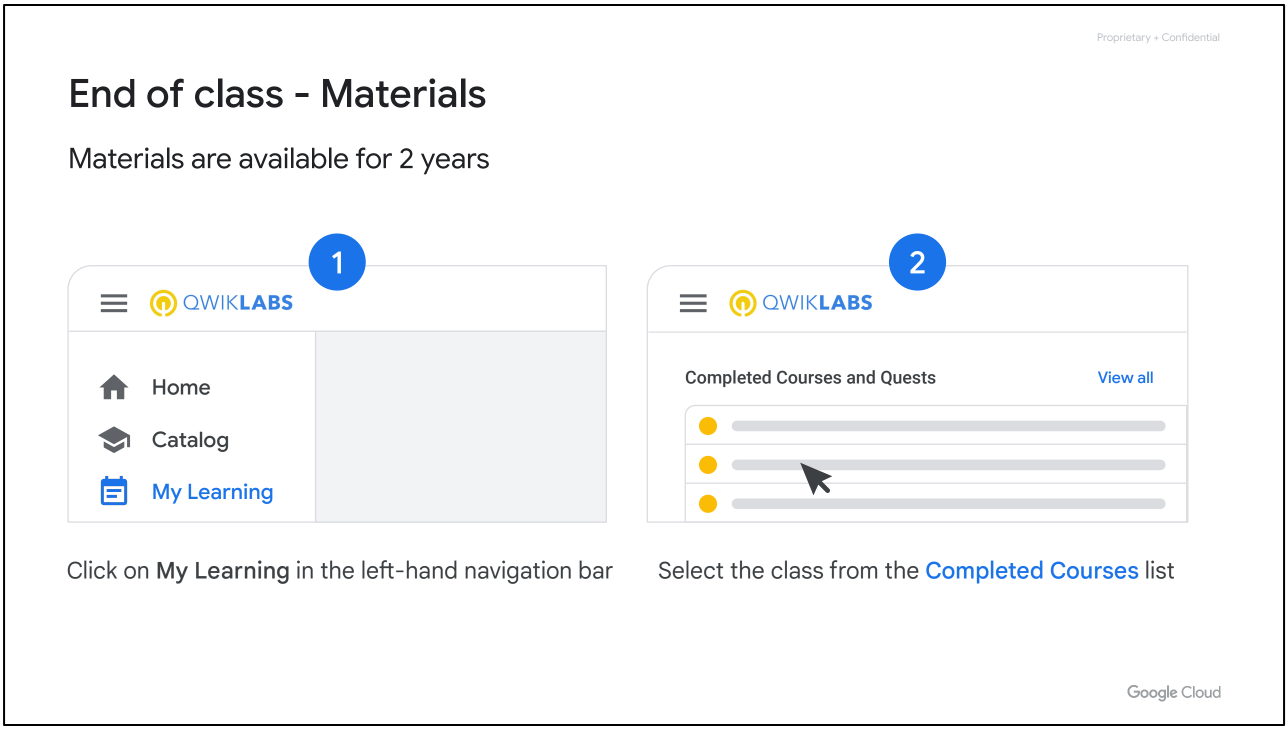


Review:
Managed Services

In this module, we provided you with an overview of managed services for data processing in Google Cloud, namely BigQuery, Dataflow, Dataprep, and Dataproc.

Managed services allow you to outsource a lot of the administrative and maintenance overhead to Google, so you can focus on your workloads, instead of the infrastructure. Speaking of infrastructure, most of the services that we covered are serverless. Now, this doesn’t mean that there aren’t any actual servers processing your data. Serverless means that servers or Compute Engine instances are obfuscated so that you don’t have to worry about the infrastructure.

Dataproc isn’t a serverless service, because you are able to view and manages the underlying master and worker instances.



Proprietary + Confidential

End of class - Materials

Materials are available for 2 years

Click on **My Learning** in the left-hand navigation bar

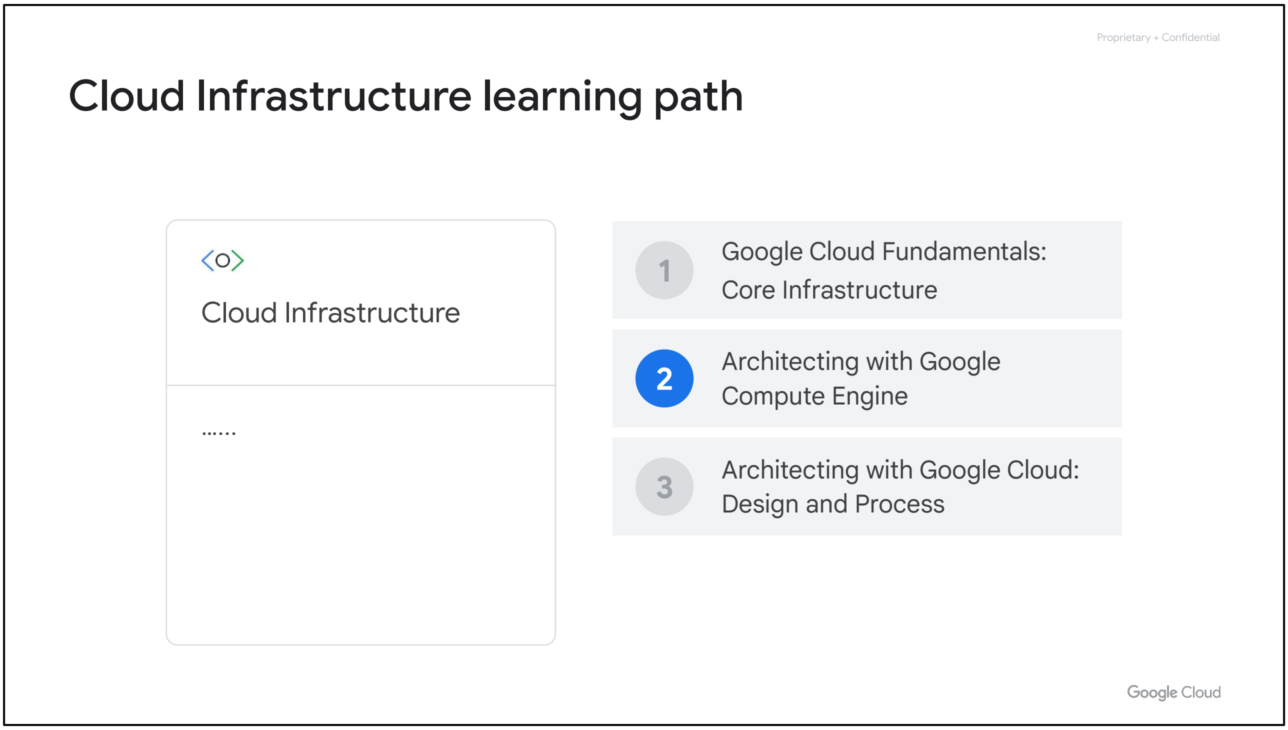
Select the class from the Completed Courses list

Google Cloud

You can view the course materials within Qwiklabs as follows:

1. Click on ***My Learning*** in the left-hand navigation bar.
2. Select the class from the ***Completed Courses*** list.

Materials are available for 2 years following the completion of a course.



Cloud Infrastructure learning path

Cloud Infrastructure

……

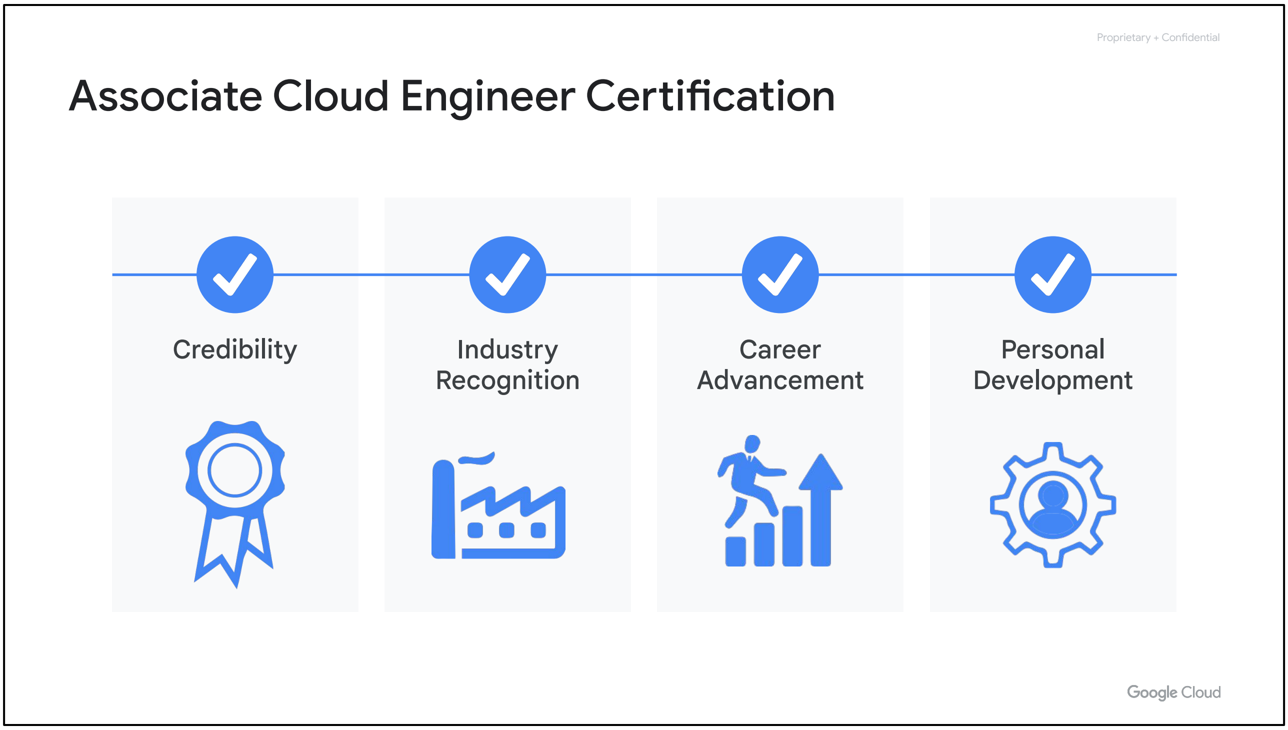
1 Google Cloud Fundamentals: Core Infrastructure

2 Architecting with Google Compute Engine

3 Architecting with Google Cloud: Design and Process

Google Cloud

The “Architecting with Google Compute Engine” course is part of the Cloud Infrastructure learning path. This path is designed for IT professionals who are responsible for implementing, deploying, migrating, and maintaining applications in the cloud. Next, we recommend taking the Architecting with Google Cloud: Design and Process course, which is part of the learning path for the Professional Cloud Architect Certification.



Proprietary + Confidential

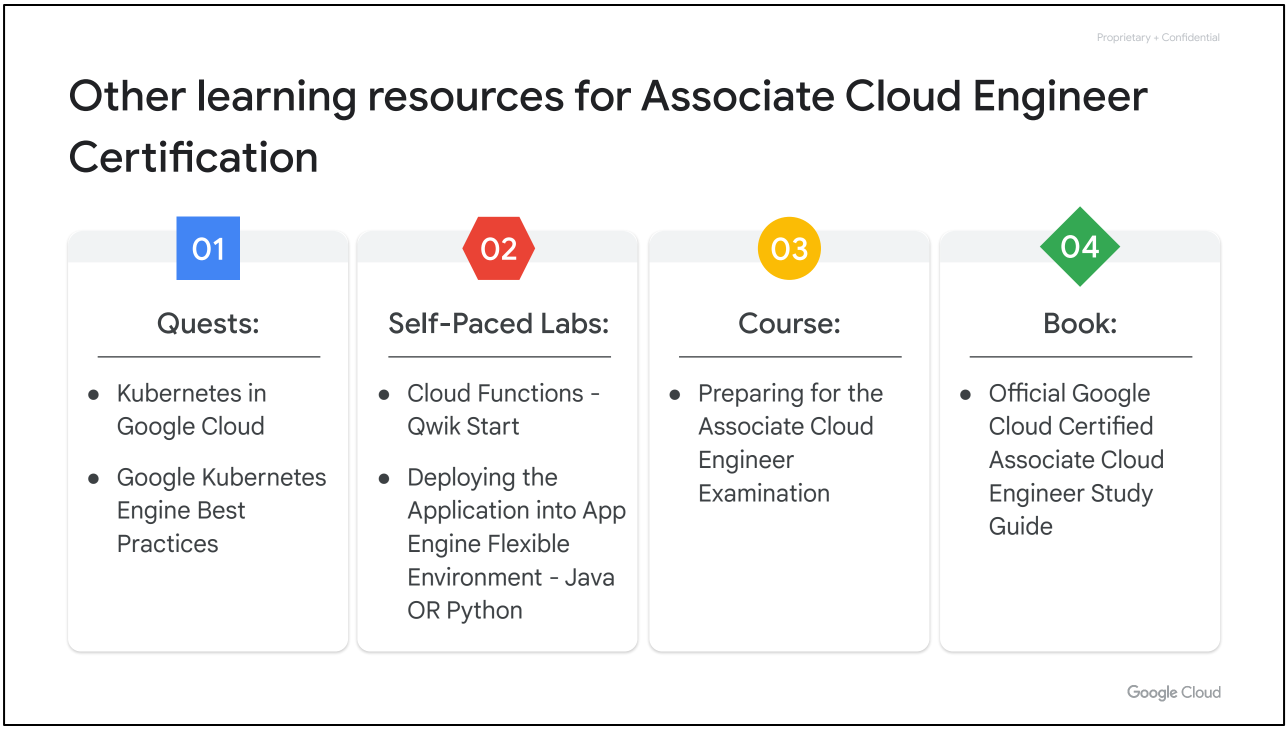
Associate Cloud Engineer Certification

Credibility
Industry Recognition
Career Advancement
Personal Development

Google Cloud

Speaking of certification, we recommend to validate your hands-on Google Cloud skills and advance your career with the **Associate Cloud Engineer** certification. Certification can help you gain credibility and give you an advantage in today's highly competitive market. The Associate Cloud Engineer certification is for individuals who want to demonstrate their ability to deploy applications, monitor operations, and maintain cloud projects on Google Cloud. It is recommended that you have at least 6 months of hands-on experience working with Google Cloud. The exam will assess your ability to:

- Set up a cloud solution environment
- Plan and configure a cloud solution
- Deploy and implement a cloud solution
- Ensure successful operation of a cloud solution
- Configure access and security



Proprietary + Confidential

Other learning resources for Associate Cloud Engineer Certification

01 Quests:

- Kubernetes in Google Cloud
- Google Kubernetes Engine Best Practices

02 Self-Paced Labs:

- Cloud Functions - Qwik Start
- Deploying the Application into App Engine Flexible Environment - Java OR Python

03 Course:

- Preparing for the Associate Cloud Engineer Examination

04 Book:

- Official Google Cloud Certified Associate Cloud Engineer Study Guide

Google Cloud

We recommend broadening your knowledge by completing the following Qwiklabs Self-Paced Labs, which are single-topic hands-on activities; and Qwiklabs Quests, which are groups of self-paced labs on a focused theme:

- Quests:
  a. Kubernetes in Google Cloud
  b. Google Kubernetes Engine Best Practices
- Self-Paced Labs:
  a. Cloud Functions - Qwik Start
  b. Deploying the Application into App Engine Flexible Environment - Java OR Deploying the Application into App Engine Flexible Environment - Python

To help you structure your preparation for the Associate Cloud Engineer exam, we recommend the Preparing for the Associate Cloud Engineer Examination course.

You can also prepare using the Official Google Cloud Certified Associate Cloud Engineer Study Guide, published by Wiley. Visit the Google Cloud Certification website for more information and to register.

Good luck!

Thank you for taking the “Architecting with Google Compute Engine” course!

We hope you have a better understanding of the comprehensive and flexible infrastructure and platform services provided by Google Cloud. We also hope that the demos and labs made you feel more comfortable with using the different Google Cloud services that we covered.

Now it’s your turn. Go ahead and apply what you have learned by architecting your own infrastructure in Google Cloud.

See you next time!