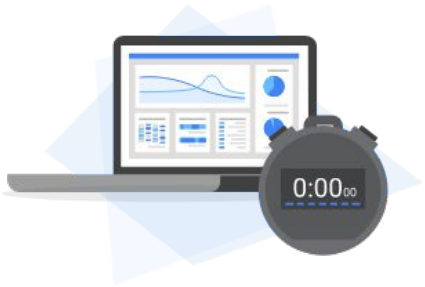
In the last module, we discussed how to automate the creation of infrastructure. As an alternative to infrastructure automation, you can eliminate the need to create infrastructure by leveraging a managed service.

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.

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

BigQuery

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



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.

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 |
+-------+---------+-------+
```


You can access BigQuery by using the 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.

Use Cloud Dataflow to execute a wide variety of data processing patterns

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

Cloud Dataflow

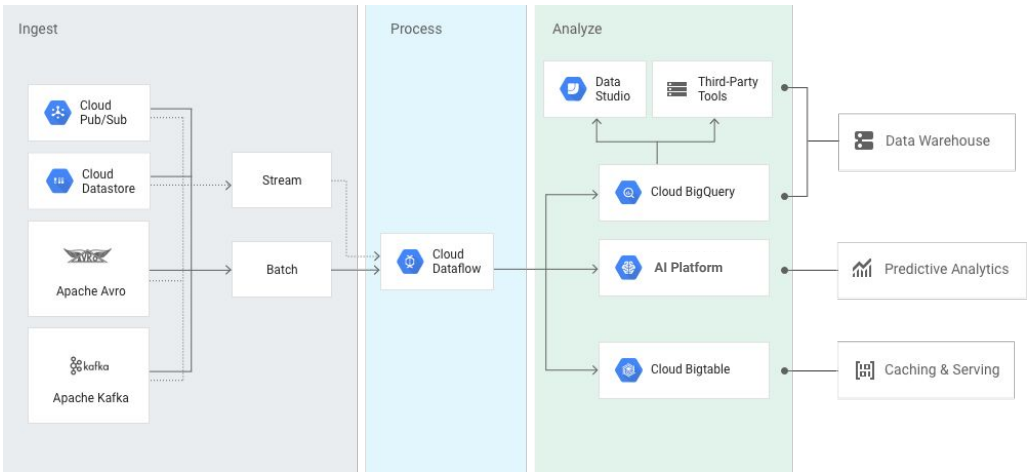
Stackdriver is now Google Cloud's operations suite

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.

Data transformation with Cloud Dataflow



This diagram shows some example uses cases of Dataflow. As I 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, AI Platform, or even Cloud Bigtable. Using Data Studio, you can even build real-time dashboards for IoT devices.

Use Cloud 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

Cloud Dataprep

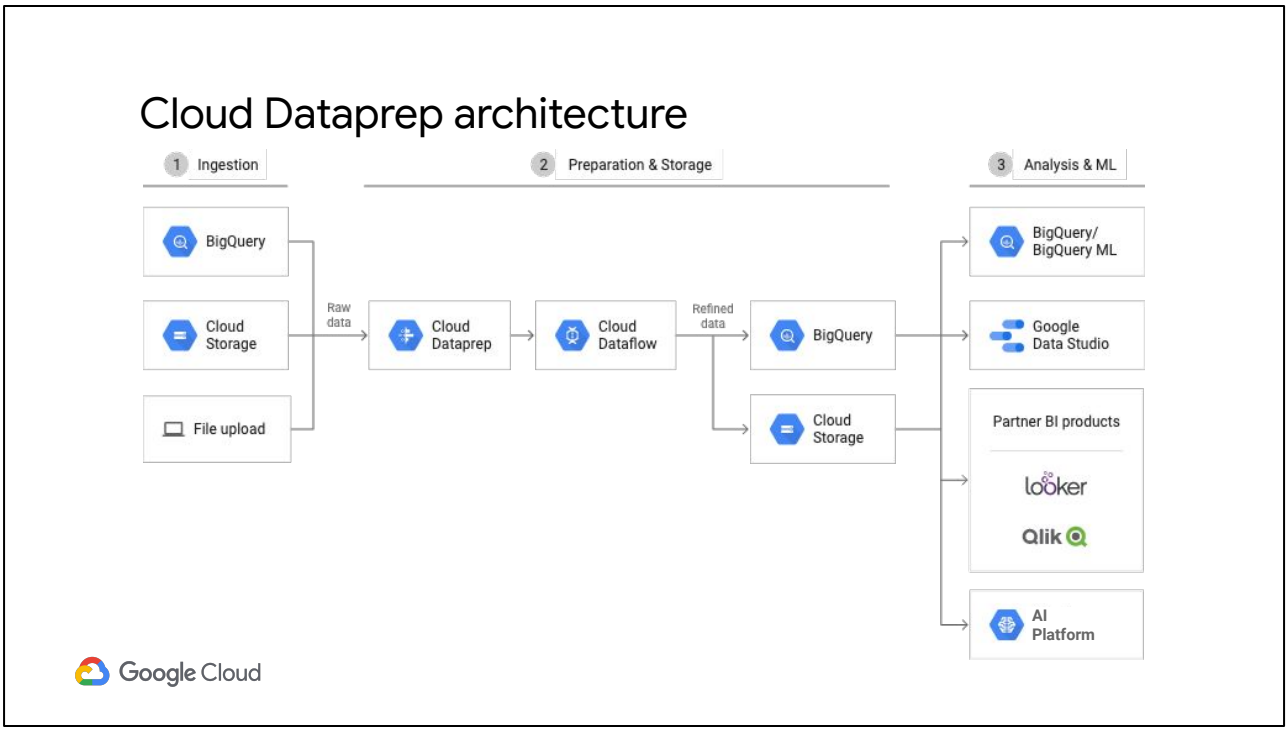
Google Cloud

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

Because Cloud 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.

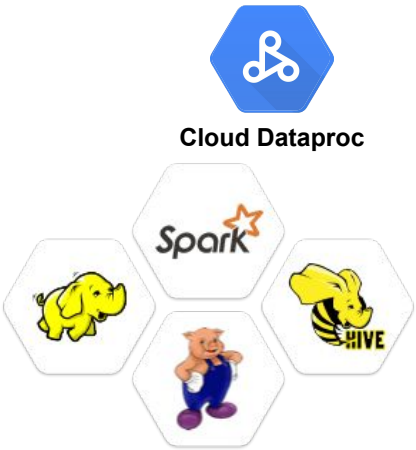
Cloud Dataprep is an integrated partner service operated by Trifacta and based on their industry-leading data preparation solution, Trifacta Wrangler. Google works closely with Trifacta to provide a seamless user experience that removes the need for up-front software installation, separate licensing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



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

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

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



Google Cloud

Cloud 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.

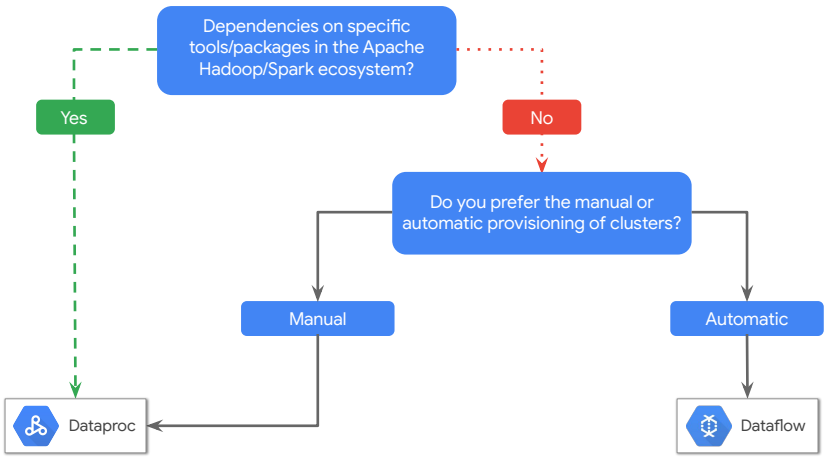
Without using Cloud Dataproc, it can take from five to 30 minutes to create Spark and Hadoop clusters on-premises or through other Infrastructure-as-a-Service providers. Cloud Dataproc clusters are quick to start, scale, and shut down, with each of these operations taking 90 seconds or less, on average.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

Cloud Dataproc has built-in integration with other GCP services, such as BigQuery, Cloud Storage, Cloud Bigtable, Stackdriver Logging, and Stackdriver 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 Cloud Dataproc. This makes it easy to move existing projects into Cloud Dataproc without redevelopment

Dataflow versus Dataproc



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 a hands-on or DevOps approach to operations, or a hands-off or serverless approach. If you opt for the DevOps approach, you want to use Dataproc; otherwise, use Dataflow.

Review

Architecting with Google Compute Engine

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

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

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

See you next time!