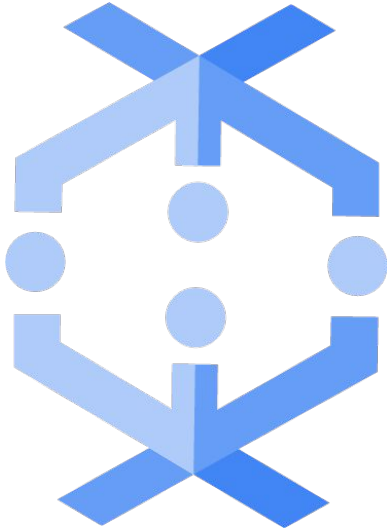
Introduction

- Dataflow
- Dataproc
- Dataflow VS. Dataproc



Dataflow

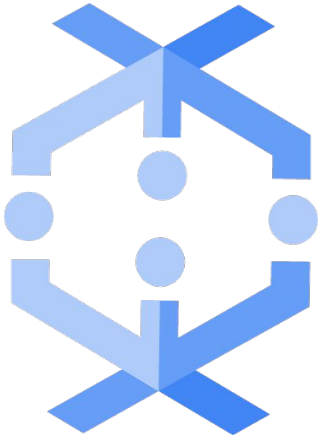
Dataflow

Dataflow is a Google Cloud service that provides ***unified stream*** and ***batch*** data processing at scale

Cloud
DataFlow

beam

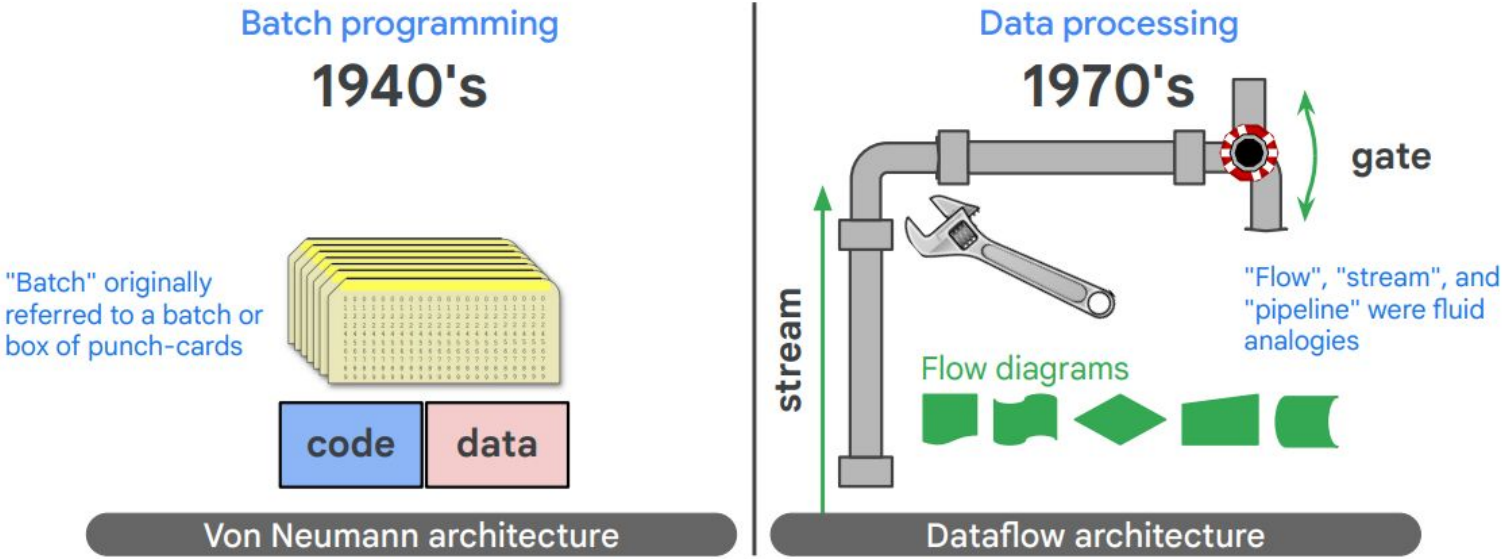
Dataflow



Qualities that Dataflow contributes to data engineering solutions:

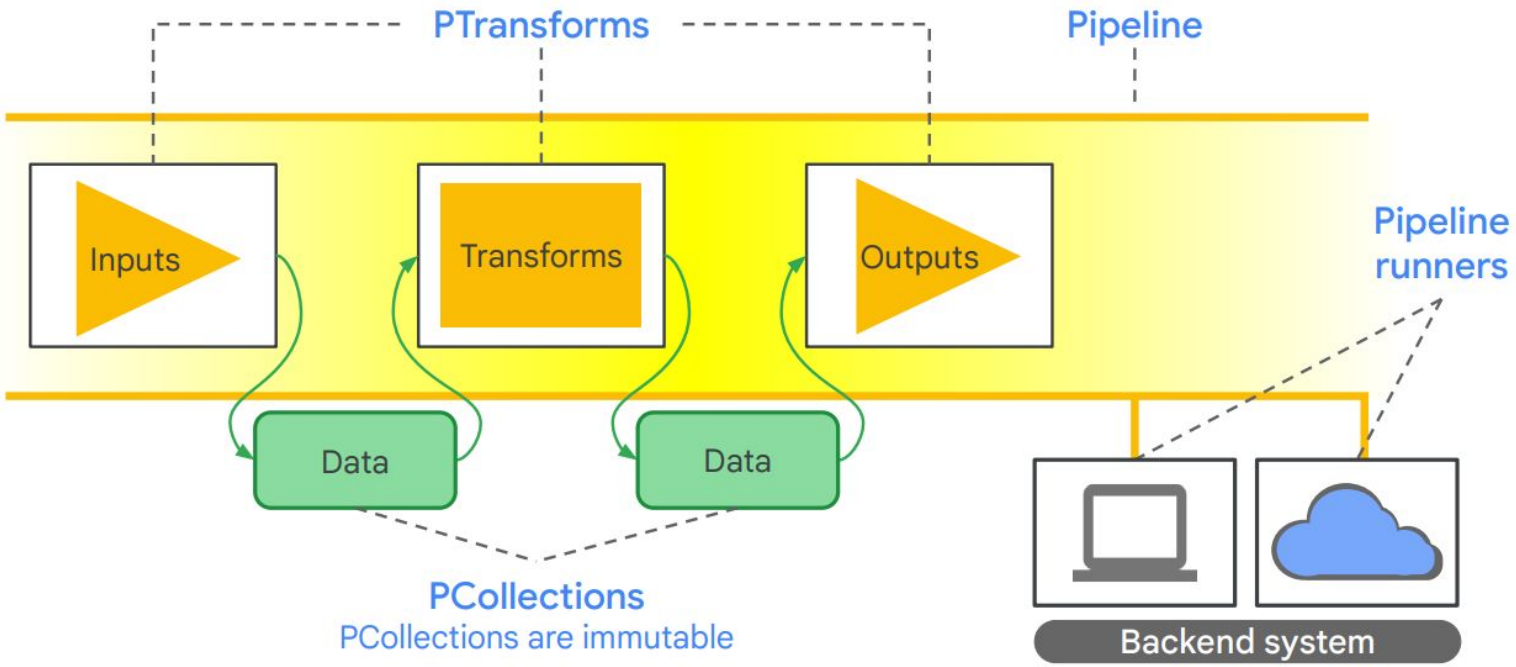
- Scalability
- Low latency

Batch programming and data processing used to be two very separate and different things

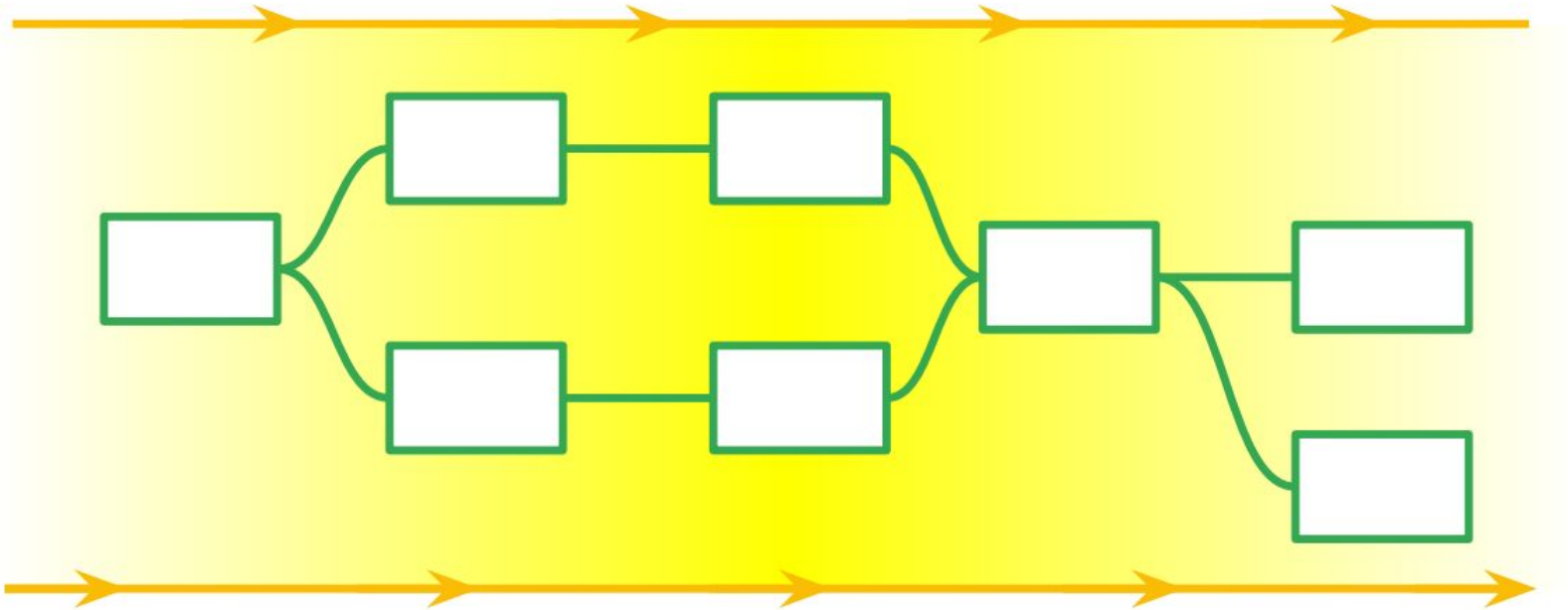


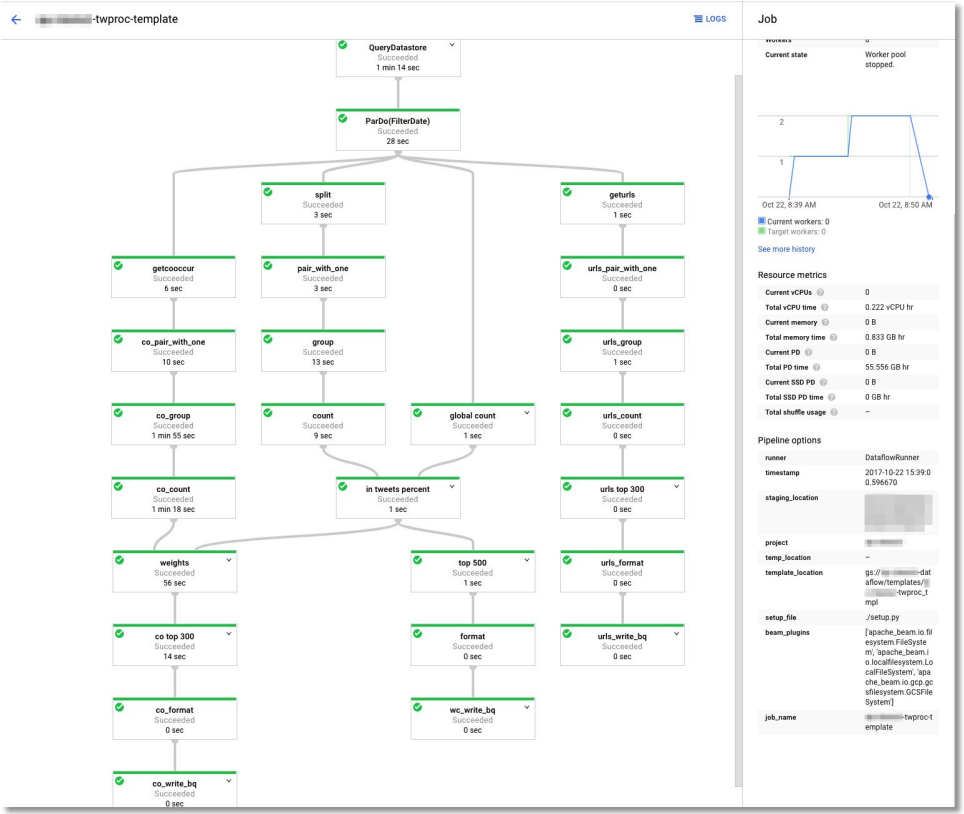
Different tools, different platforms, different concepts, different methods.

Apache BEAM = Batch + strEAM



A Dataflow pipeline is a directed graph of steps



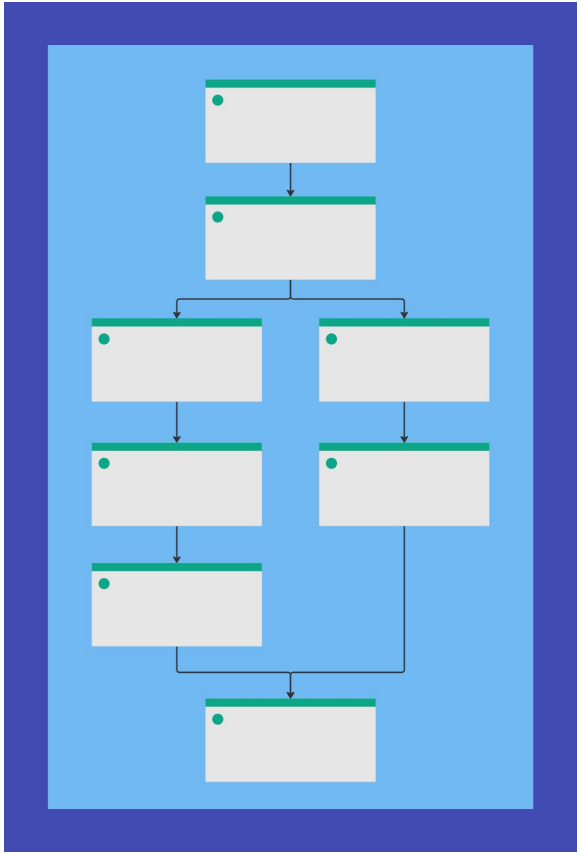

-twproc-template
LOGS
Job
QueryDatastore
Succeeded
1 min 14 sec
ParDo(FilterDate)
Succeeded
28 sec
split
Succeeded
3 sec
geturls
Succeeded
1 sec
getcooccur
Succeeded
6 sec
pair_with_one
Succeeded
3 sec
urls_pair_with_one
Succeeded
0 sec
co_pair_with_one
Succeeded
10 sec
group
Succeeded
13 sec
urls_group
Succeeded
1 sec
co_group
Succeeded
1 min 55 sec
count
Succeeded
9 sec
global count
Succeeded
1 sec
urls_count
Succeeded
0 sec
co_count
Succeeded
1 min 18 sec
in tweets percent
Succeeded
1 sec
urls top 300
Succeeded
0 sec
weights
Succeeded
56 sec
top 500
Succeeded
1 sec
urls_format
Succeeded
0 sec
co top 300
Succeeded
14 sec
format
Succeeded
0 sec
urls_write_bq
Succeeded
0 sec
co_format
Succeeded
0 sec
wc_write_bq
Succeeded
0 sec
co_write_bq
Succeeded
0 sec
Current state
Worker pool stopped.
Oct 22, 8:39 AM
Oct 22, 8:50 AM
Current workers: 0
Target workers: 0
See more history
Resource metrics
Current vCPUs 0
Total vCPU time 0.222 vCPU hr
Current memory 0 B
Total memory time 0.833 GB hr
Current PD 0 B
Total PD time 55.556 GB hr
Current SSD PD 0 B
Total SSD PD time 0 GB hr
Total shuffle usage –
Pipeline options
runner DataflowRunner
timestamp 2017-10-22 15:39:00.596670
staging_location
project
temp_location –
template_location gs://-dataflow/templates/-twproc_tmpl
setup_file ./setup.py
beam_plugins ['apache_beam.io.filesystem.FileSystem', 'apache_beam.io.localfilesystem.LocalFileSystem', 'apache_beam.io.gcp.gcsfilesystem.GCSFileSystem']
job_name -twproc-template


ReadLines
Succeeded
1 sec
1 of 1 stage succeeded
WordCount.CountWords
Succeeded
1 sec
2 of 2 stages succeeded
MapElements
Succeeded
0 sec
1 of 1 stage succeeded
WriteCounts
Succeeded
3 sec
4 of 4 stages succeeded

Dataflow templates

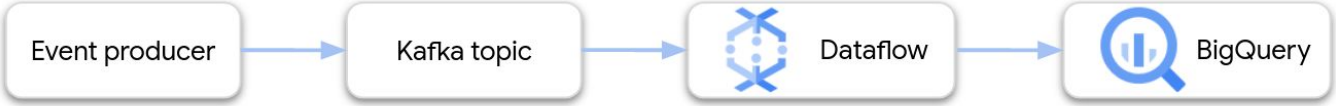
Dataflow templates allow you to package a Dataflow pipeline for deployment.

You can create your own ***custom Dataflow templates***, and Google provides ***pre-built templates*** for common scenarios.

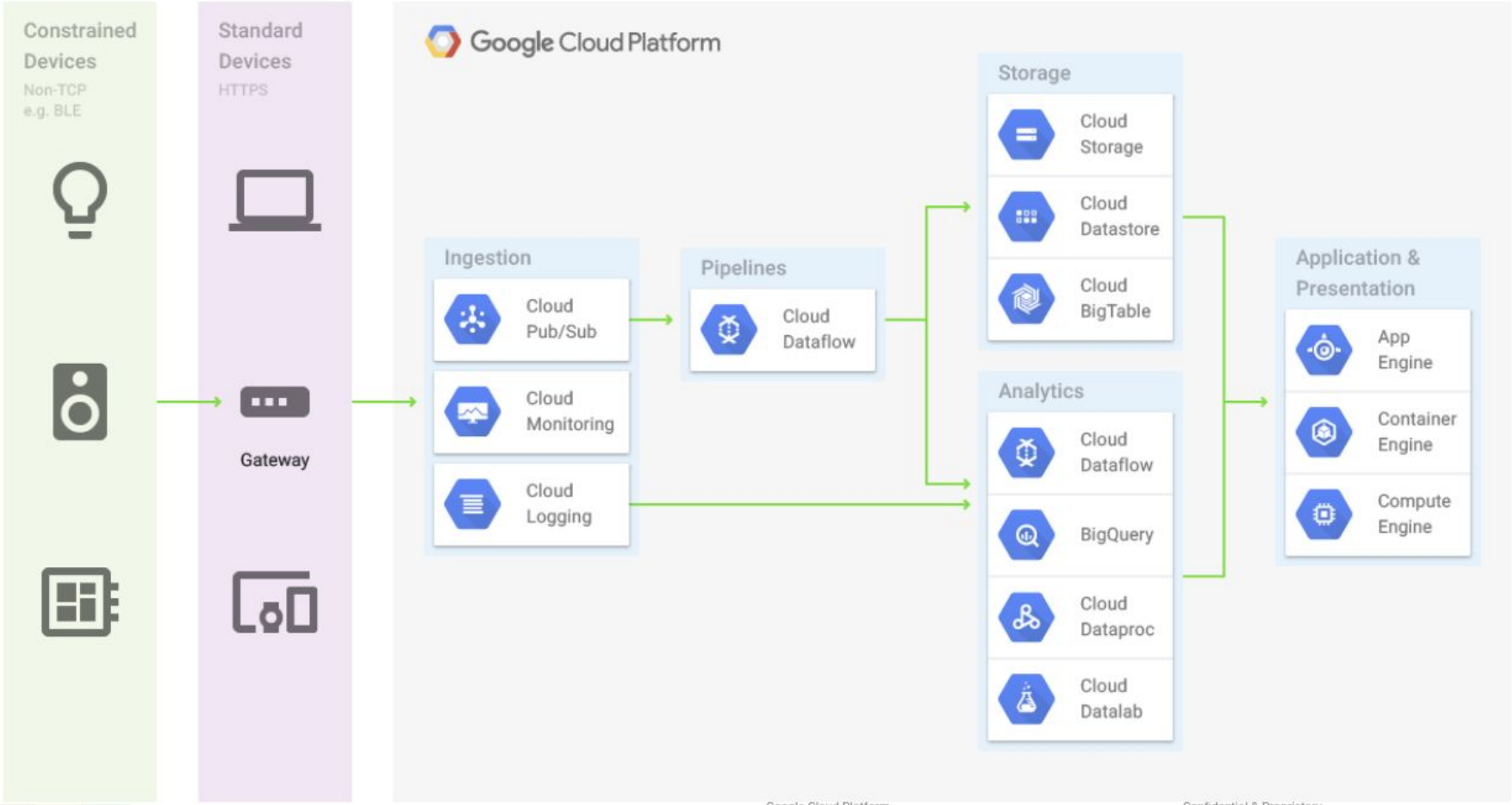


Dataflow templates Pre-built templates

- Apache Kafka to BigQuery
- Change Data Capture from MySQL to BigQuery (Stream)
- MongoDB to BigQuery (CDC)
- Pub/Sub to Pub/Sub

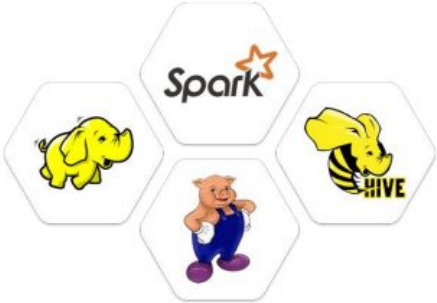


Reference Architecture (Real Time Streaming IoT)



Dataproc

The Hadoop ecosystem developed because of a need to analyze large datasets



Database

Bring the data to
the processor

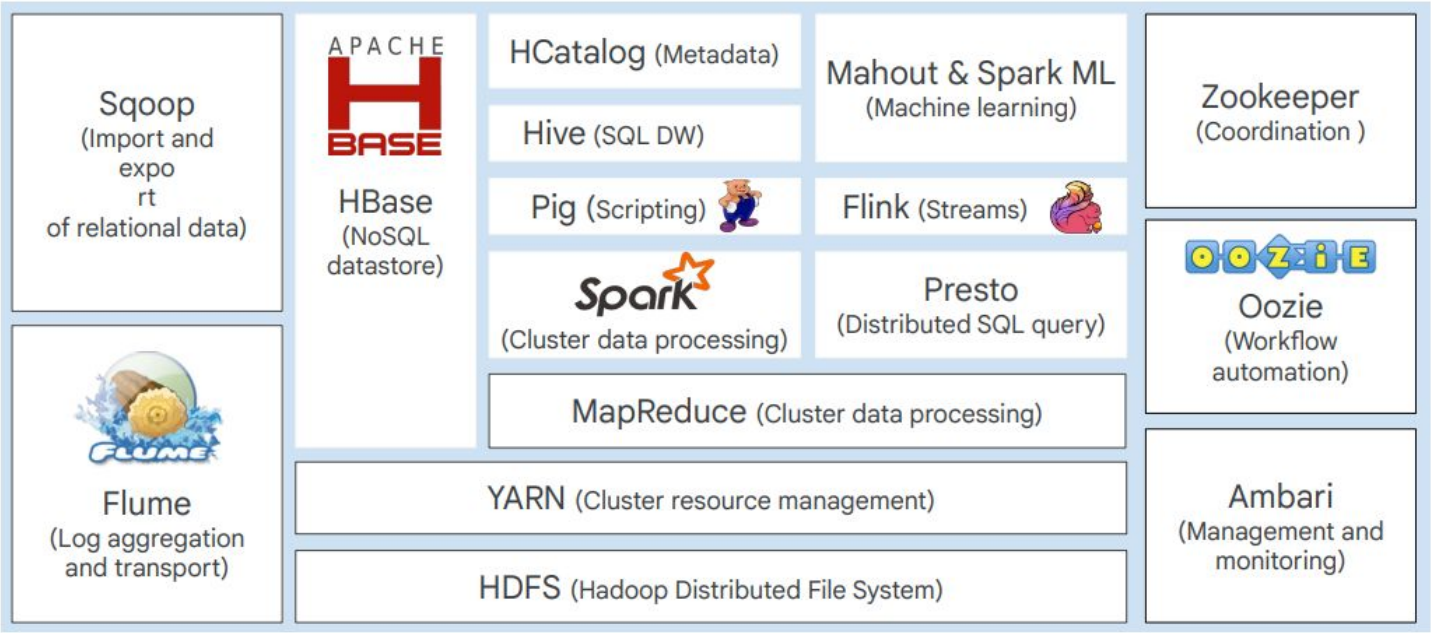
Hadoop

Distribute the processing

Store the data with
The processors

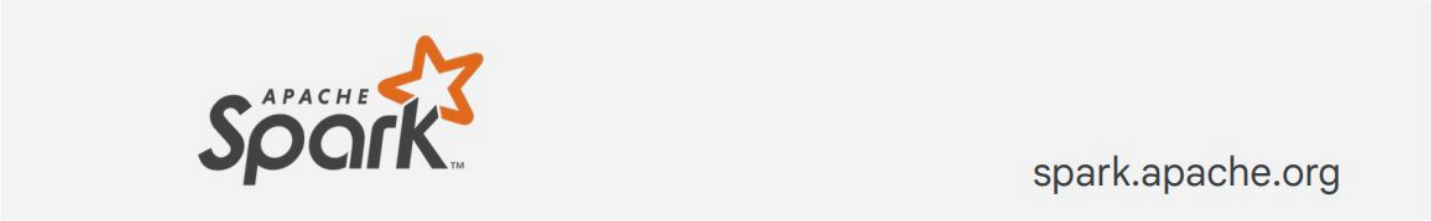
2006

The Hadoop ecosystem is very popular for Big Data workloads

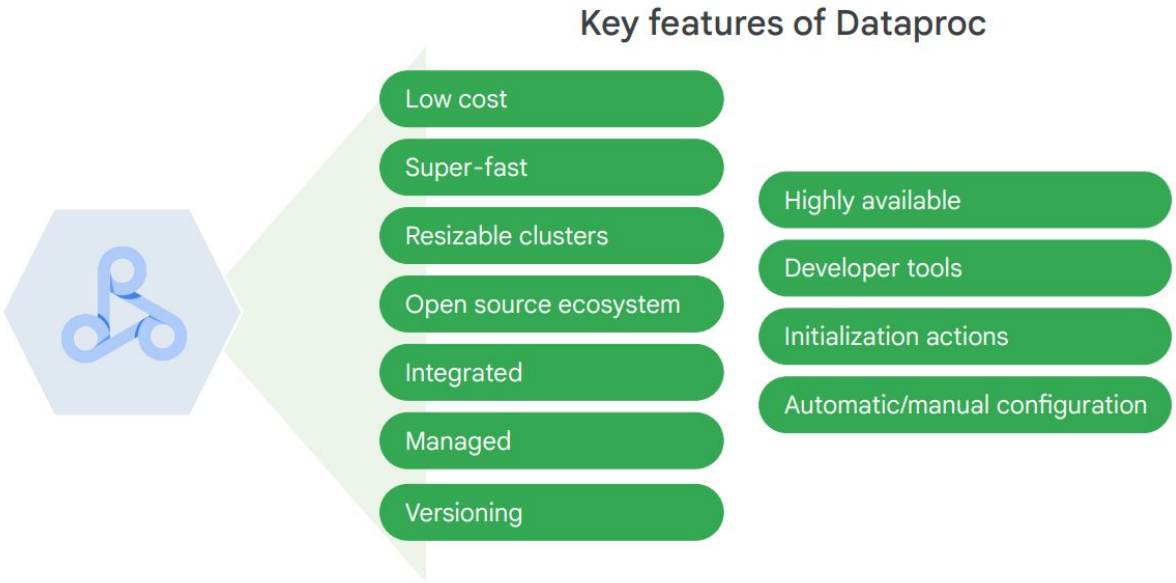


Apache Spark is a popular, flexible, powerful way to process large datasets

etc.



Dataproc is a managed service for running Hadoop and Spark data processing workloads



There are other OSS options available in Dataproc

• Spark (default)	• Hive (default)	• HDFS (default)
• Pig (default)	• Zeppelin	• Zookeeper
• Kafka	• Hue	• Tez
• Presto	• Anaconda	• Cloud SQL Proxy
• Jupyter	• Apache Flink	• Datalab
• IPython	• Oozie	• Sqoop
• Much more...		

Comparison between Dataproc & Dataflow

Dataflow VS. Dataproc

	Dataflow	Dataproc
Recommended for:	New data processing pipelines, unified batch and streaming	Existing Hadoop/Spark applications, machine learning/data science ecosystem, large-batch jobs, preemptible VMs
Serverless:	Yes	No
Auto-scaling:	Yes, transform-by-transform (adaptive)	Yes, based on cluster utilization (reactive)
Expertise:	Apache Beam	Hadoop, Hive, Pig, Apache Big Data ecosystem, Spark, Flink, Presto, Druid

Choosing between Dataflow and Dataproc