Google Cloud

Make the cloud work for you

Easy, fast, and low-cost streaming with Apache Flink on Google Cloud Platform

James Malone [jamesmalone@google.com]

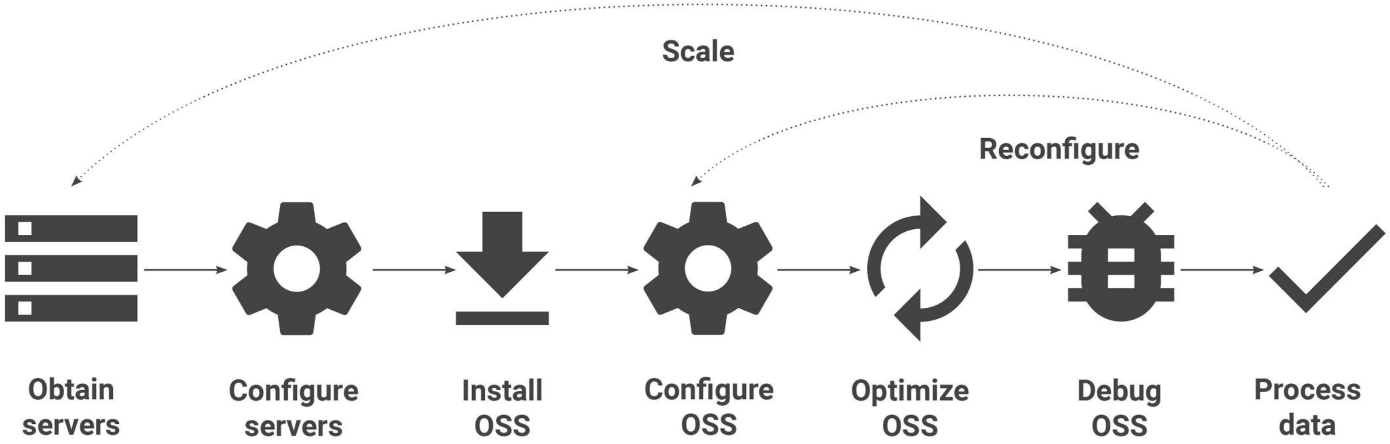
Open source

powerful but **complex**

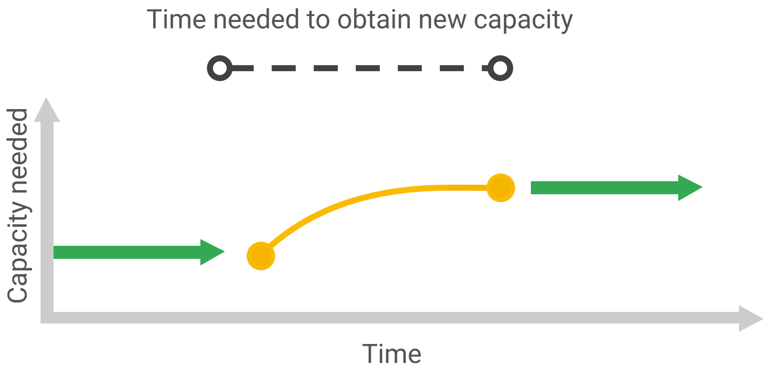
The Apache ecosystem

kafka
Spark
hadoop
APACHE DRILL
HIVE
APACHE HBASE
ZOO
FLUME
OOZIE

Typical OSS deployments

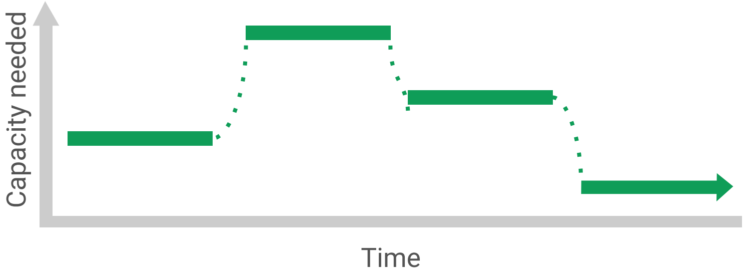


Scaling makes your life difficult



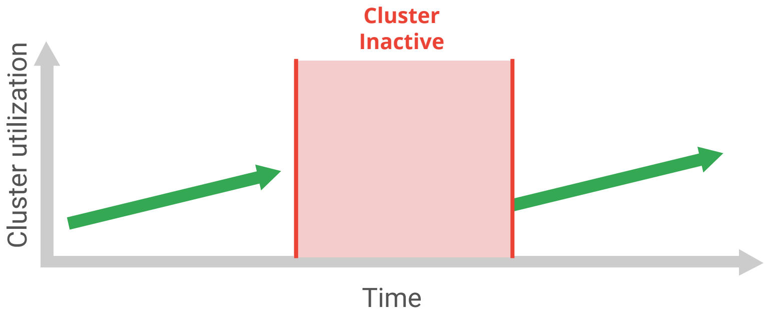
Scaling can take hours, days, or weeks to perform which may delay needed data processing

Scaling should be painless and fast



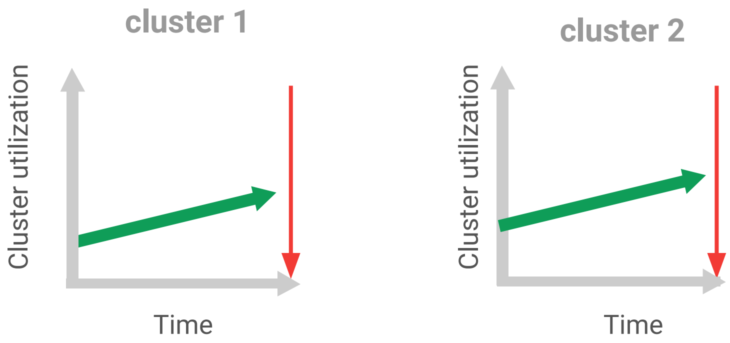
Things take seconds to minutes,
not hours or weeks.

You have to babysit utilization



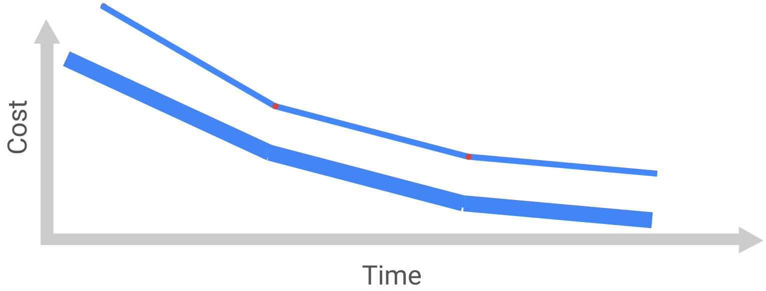
Requires effort to pack clusters so the it does not have periods of inactivity and wasted resources

Only use clusters when you need them



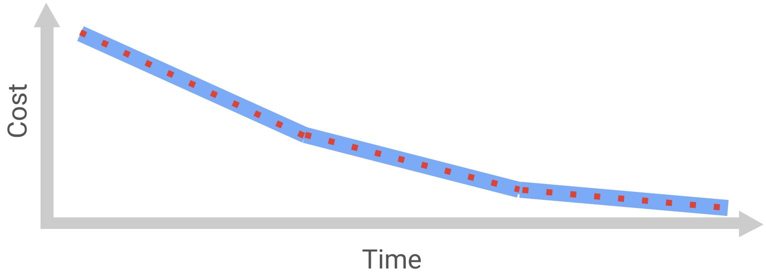
Be an expert with your data, not
your infrastructure

You are not paying for what you use



You are paying for more (spare) capacity than you actually need to process your data

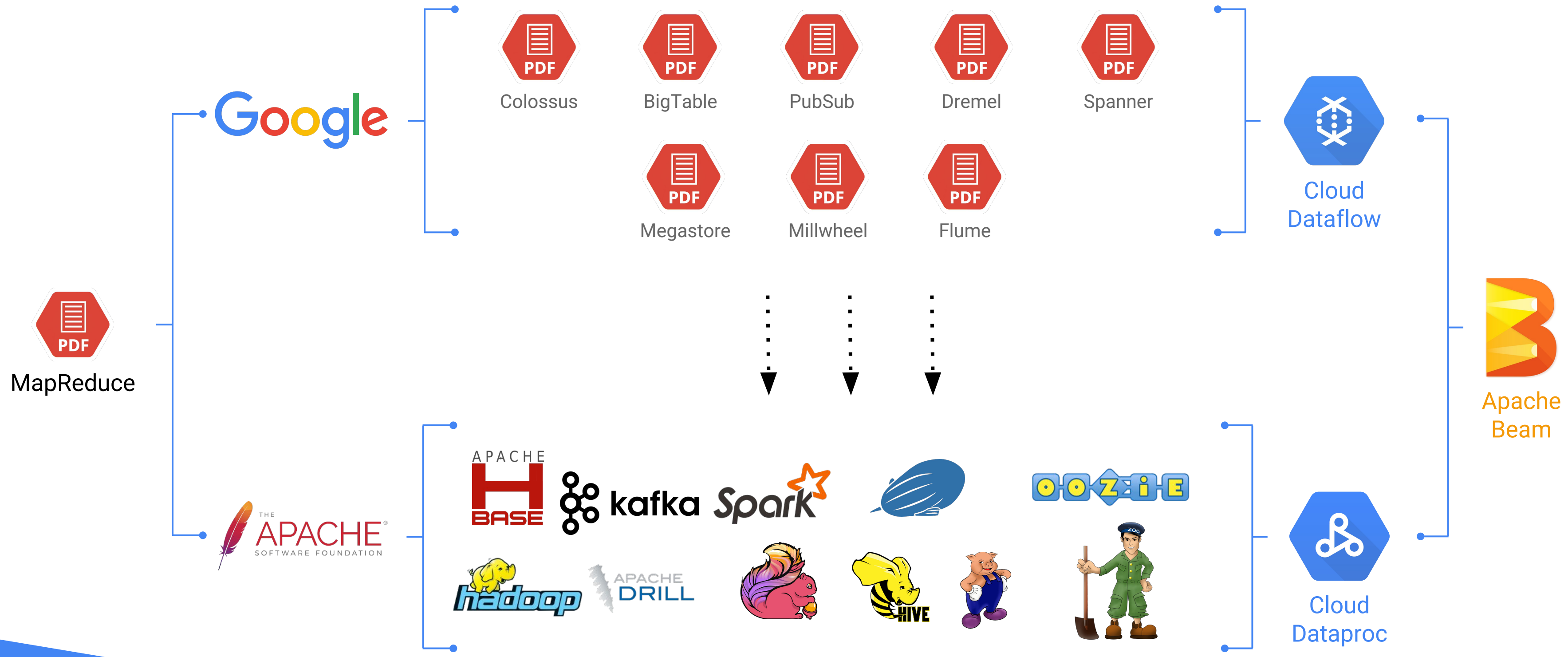
Pay for exactly what you use



You only pay for resources when you need them

Open source on **Google Cloud**

Google is passionate about open source



MapReduce

Colossus

BigTable

PubSub

Dremel

Spanner

Megastore

Millwheel

Flume

Cloud Dataflow

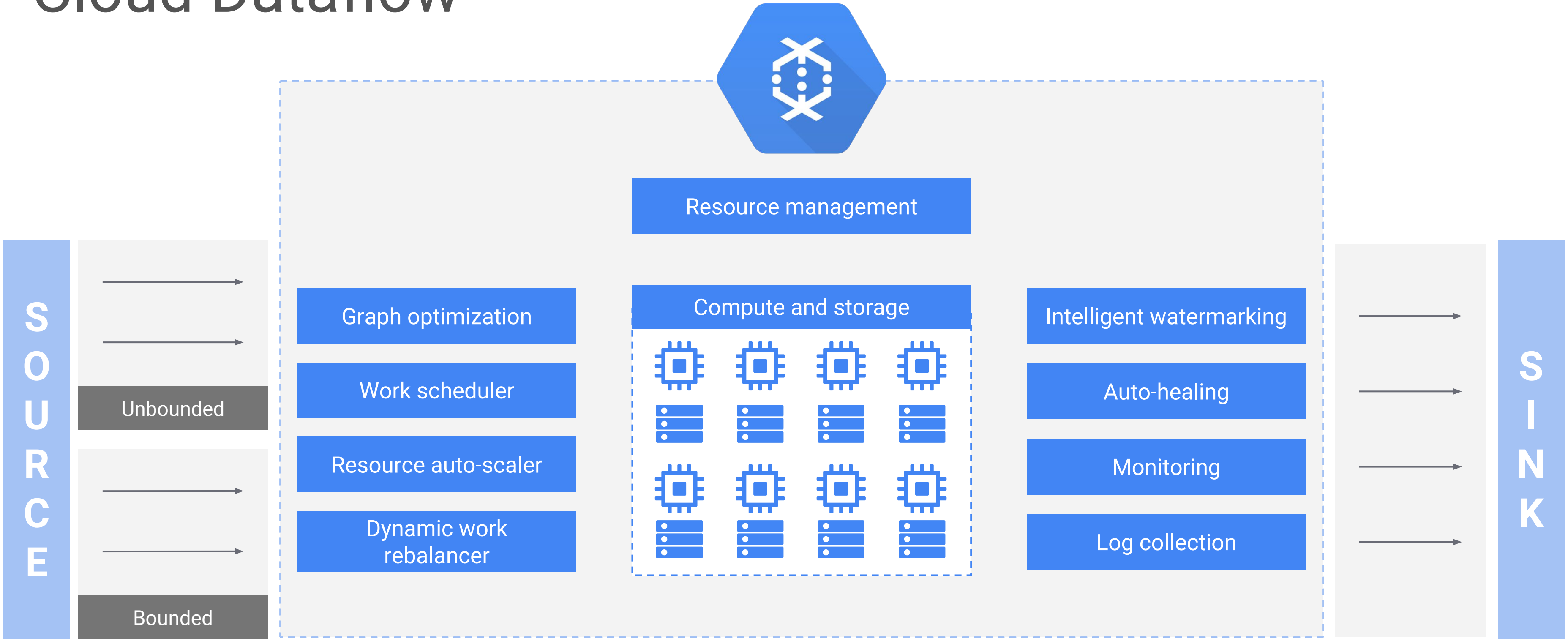
Apache Beam

Cloud Dataproc

What is Cloud Dataflow?

- Unified batch and streaming processing
- Fully managed, no-ops data processing
- Open source programming model
- Intelligently scales to millions of QPS

Cloud Dataflow



Cloud Dataproc offers a spectrum

Cloud Dataflow

Cloud Dataproc

Cloud Dataflow

Cloud Dataflow is a real-time data processing service for batch and stream data processing.

- Fully managed
- Unified programming model
- Integrated and open source
- Resource management
- Autoscaling
- Monitoring

What is Cloud Dataproc?

Google Cloud Dataproc is a fast, easy to use, low cost and fully-managed service, powered by Google Cloud Platform, that helps you take advantage of the Spark, Flink, and Hadoop ecosystem.

Google Cloud Dataproc

Fast

Things take seconds to minutes, not hours or weeks

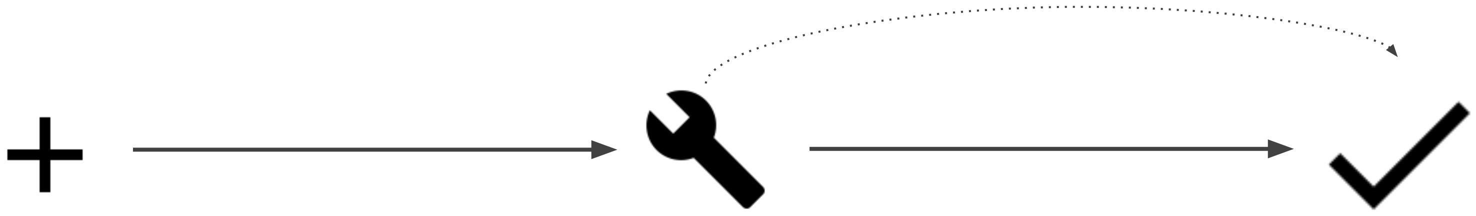
Easy

Be an expert with your data, not your data infrastructure

Cost-effective

Pay for exactly what you use to process your data, not more

Cloud Dataproc clusters



Scale anytime

Create cluster

(0 seconds)

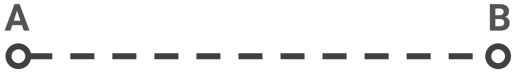
Configure cluster

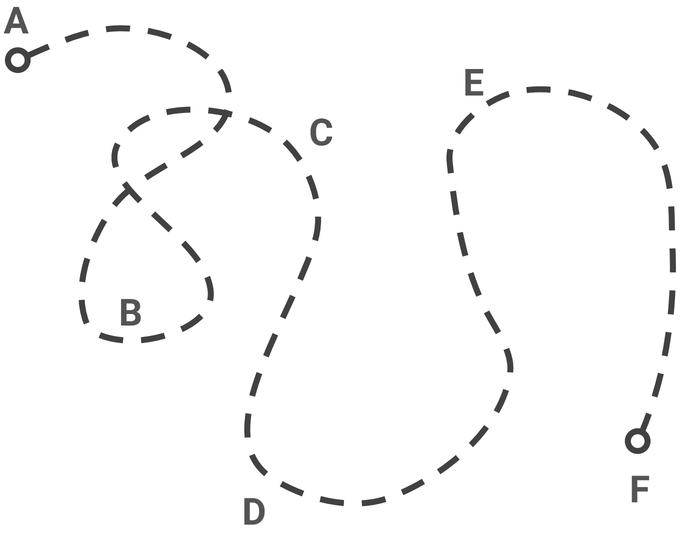
(20 seconds)

Use cluster

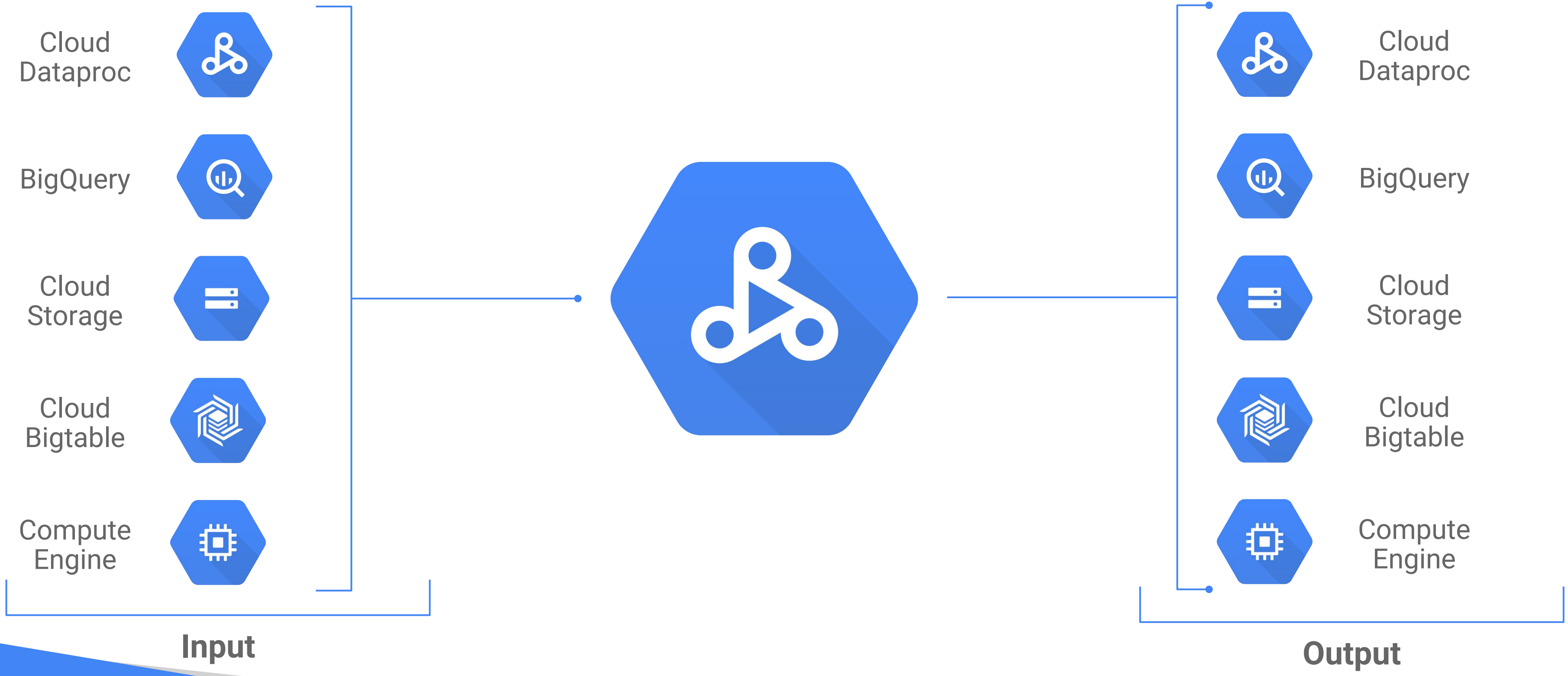
(90 seconds)

Cloud Dataproc offers a spectrum

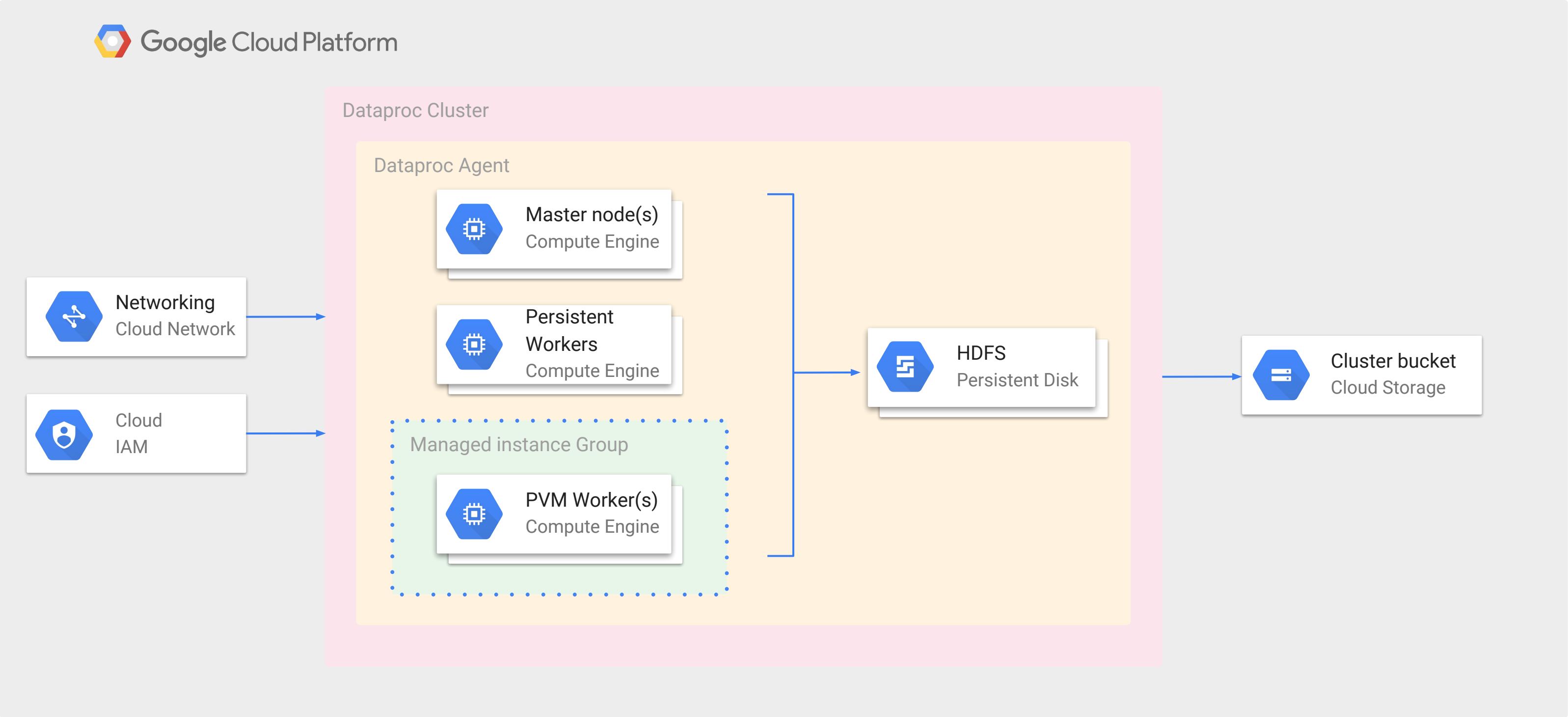




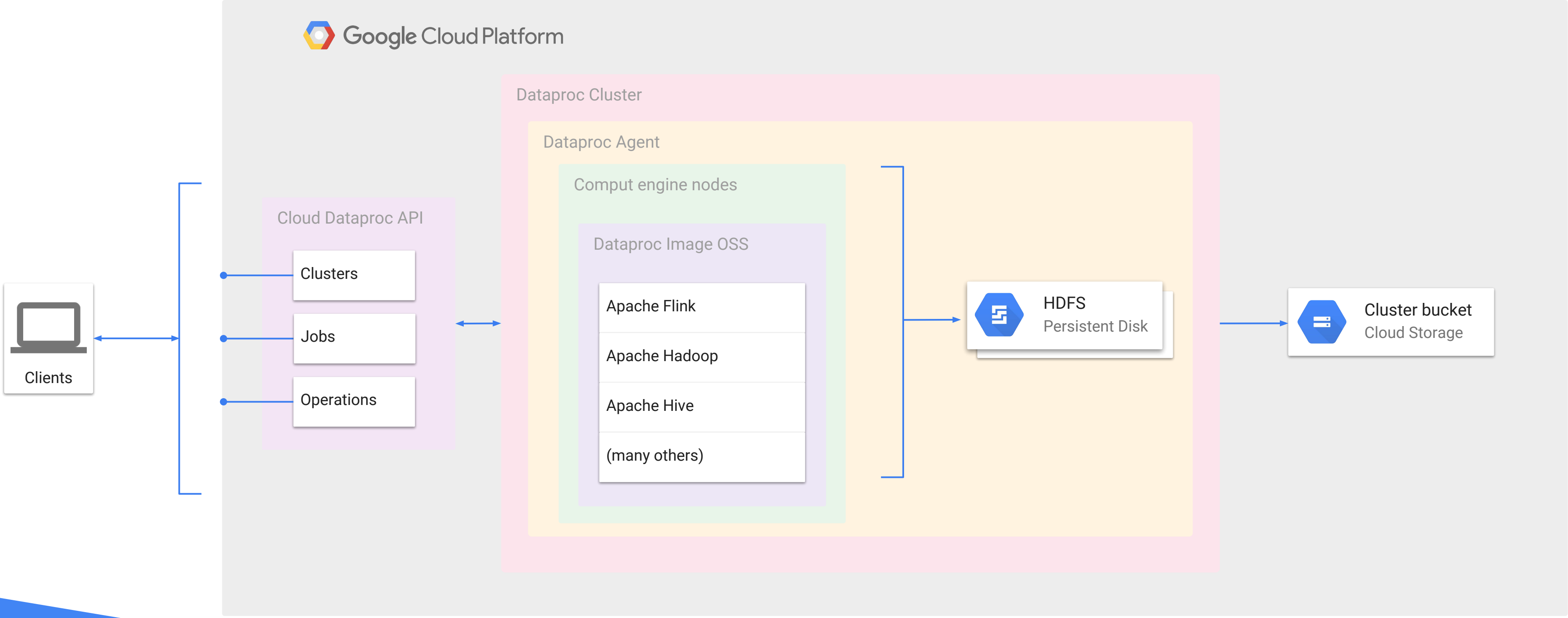
Connecting OSS to Cloud Platform



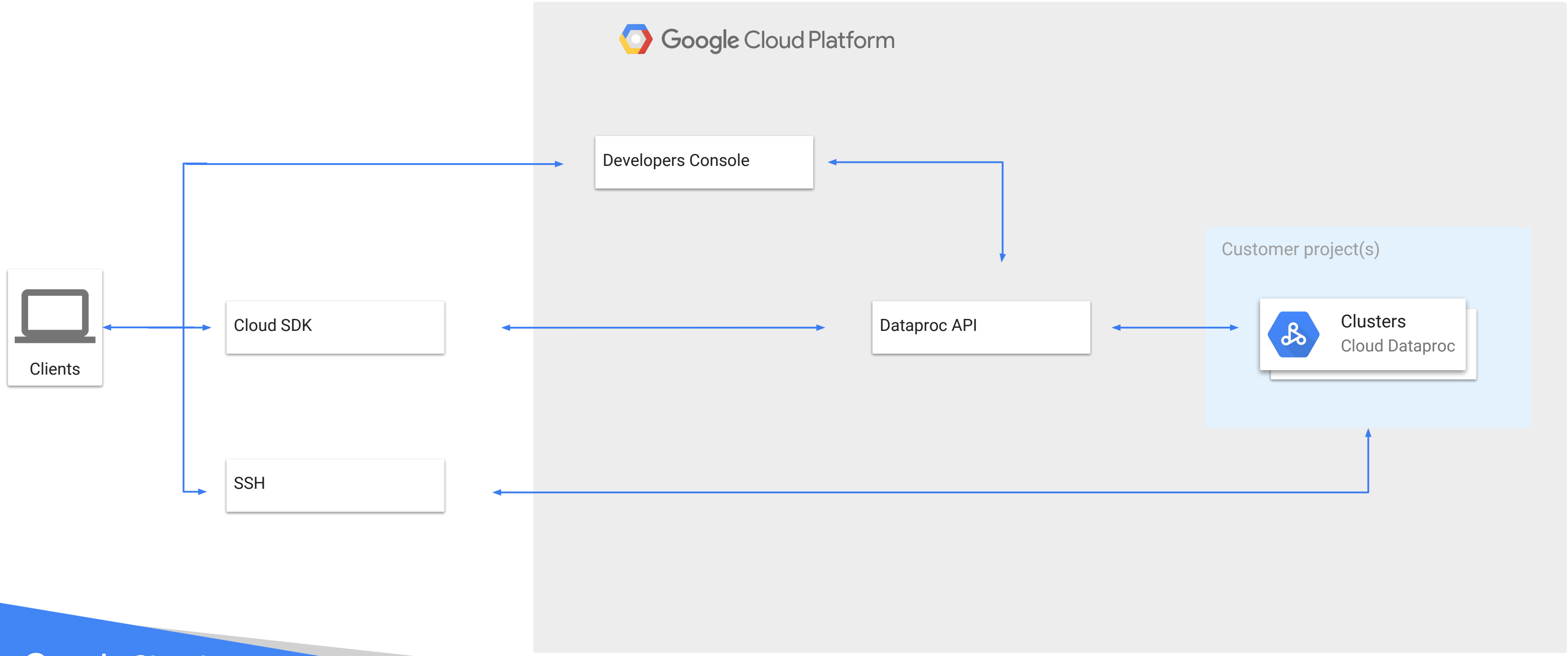
Cloud Dataproc - under the hood



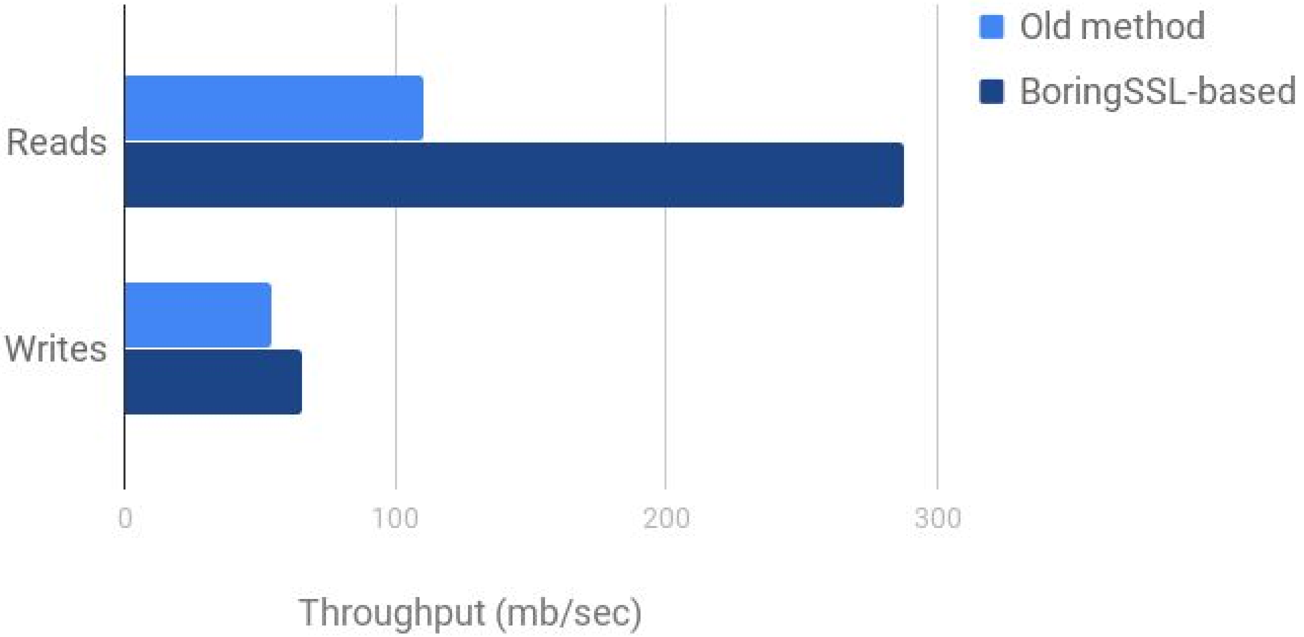
Cloud Dataproc - under the hood



Cloud Dataproc - under the hood



Cloud Storage performance (and improvement)



Cloud Dataproc **demo**

FEDERAL RESERVE NOTE
THE UNITED STATES OF AMERICA
THIS NOTE IS LEGAL TENDER FOR ALL DEBTS, PUBLIC AND PRIVATE
L 11180916 G
FEDERAL RESERVE BANK OF SAN FRANCISCO CALIFORNIA
L
12
H3
WASHINGTON, D.C.
ONE
THE DEPARTMENT OF THE TREASURY 1789
FW H 57
SERIES 2003 A
Anna Escobedo Cabral
Treasurer of the United States.
John W Snow
Secretary of the Treasury.
WASHINGTON
ONE DOLLAR

$12.85

Example new features and their impact

Restartable jobs (beta)

- Any job submitted through the Cloud Dataproc Jobs API can now be set to automatically restart on failure

- Very useful for both batch **and** streaming jobs. Jobs which checkpoint can also be automatically restarted

- Specified with the switch `--max_retries_per_hour` when using the Cloud SDK (`gcloud`) the `max_failures_per_hour` in the Jobs API

Clusters with GPUs (beta)

- Cloud Dataproc clusters support Compute Engine nodes with Nvidia Tesla K80 GPUs attached to them

- We expect GPU support will continue to grow in the open source data processing ecosystem throughout 2017

- Easily add GPUs to a Cloud Dataproc cluster with the switch `--master/worker_accelerator` with the Cloud SDK (`gcloud`)

Single-node clusters (beta)

- Create a sandbox Cloud Dataproc “cluster” with only one node instead of the typical three node design (1 master, 2 workers)

- Great for lightweight data science, small-scale testing, proof of concept building, and education

- Use the `--single-node` argument in the Cloud SDK or select “Single node” when creating a cluster in the Google Cloud Console

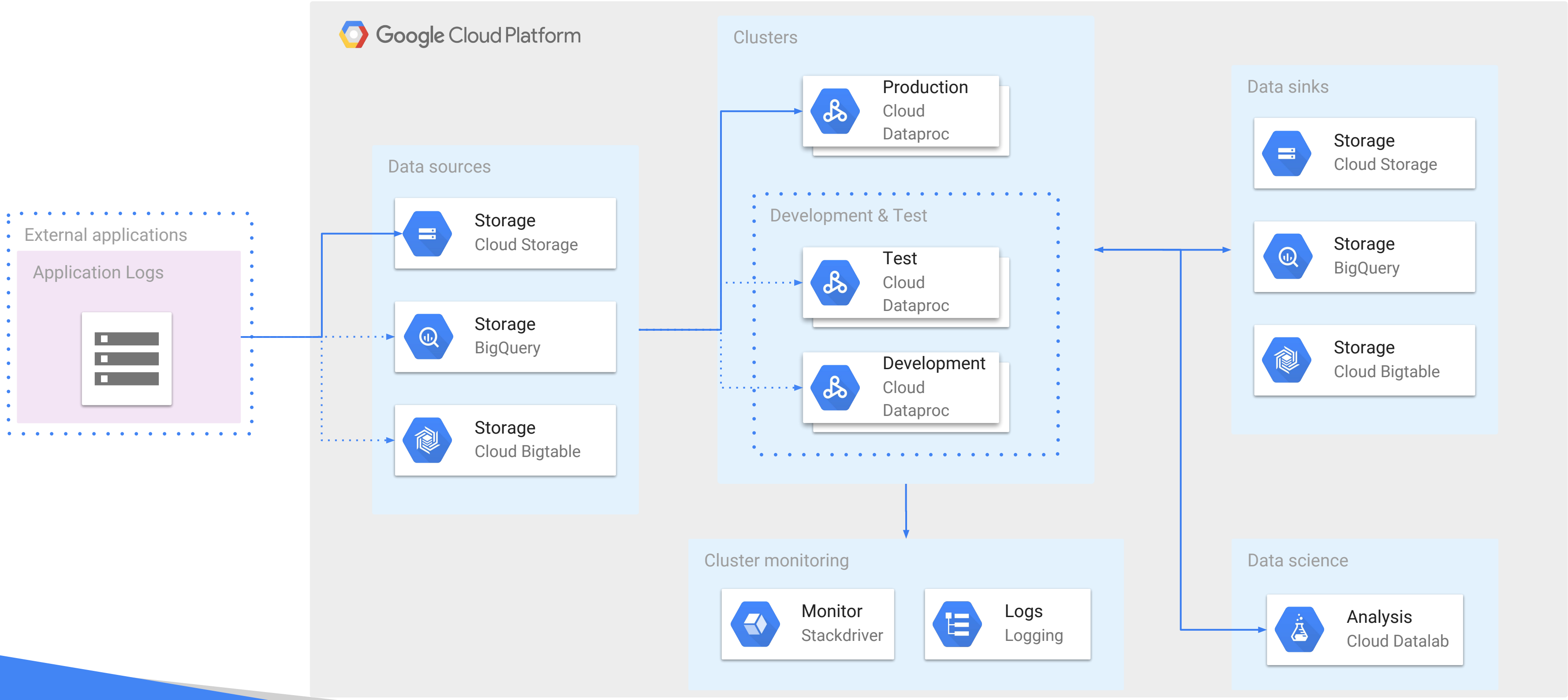
Regional endpoints & private IP clusters (beta)

- Cloud Dataproc supports a “global” endpoint and “regional endpoints” in each Compute Engine region. This allows you to isolate Cloud Dataproc interactions to one specific region

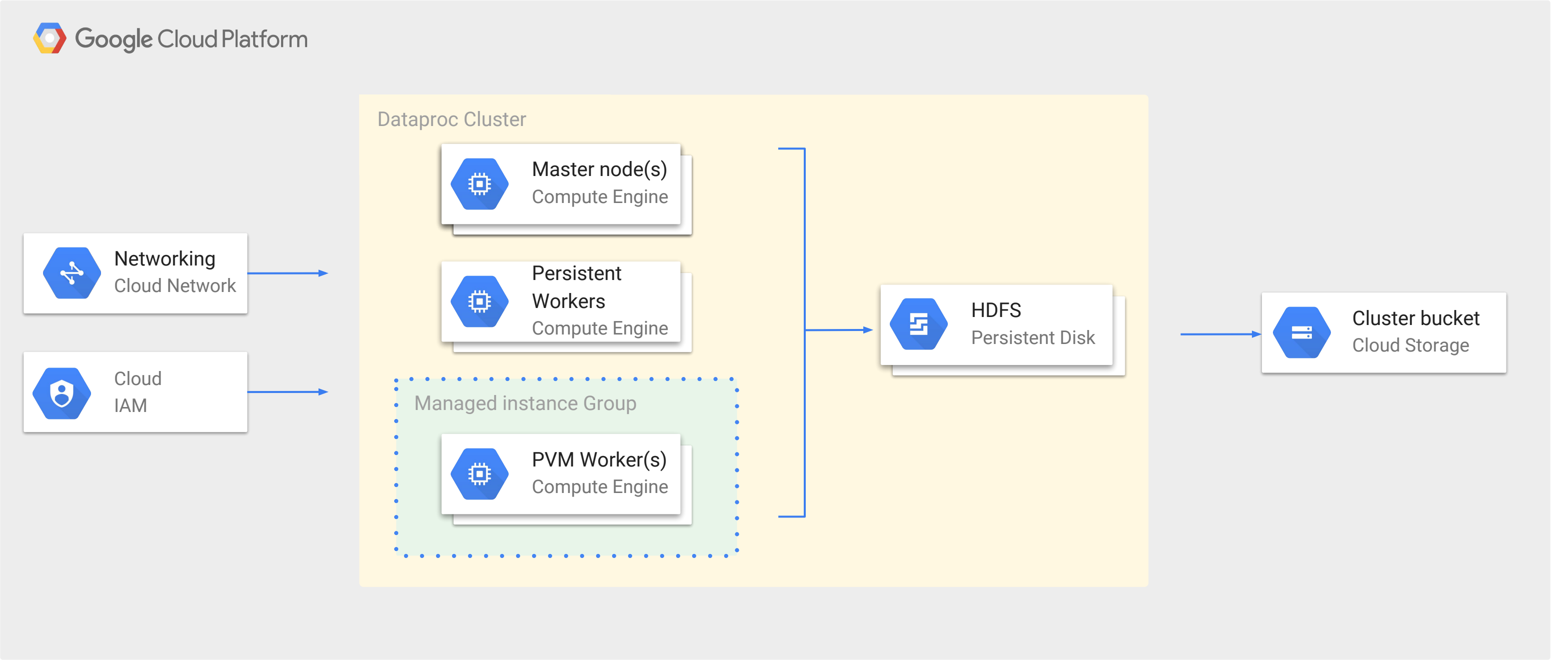
- Traditionally clusters have needed a public IP attached to them. Cloud Dataproc now supports (easy to setup) “private IP only” clusters which do not require a public IP address on Compute Engine nodes

Cloud platform **architecture concepts**

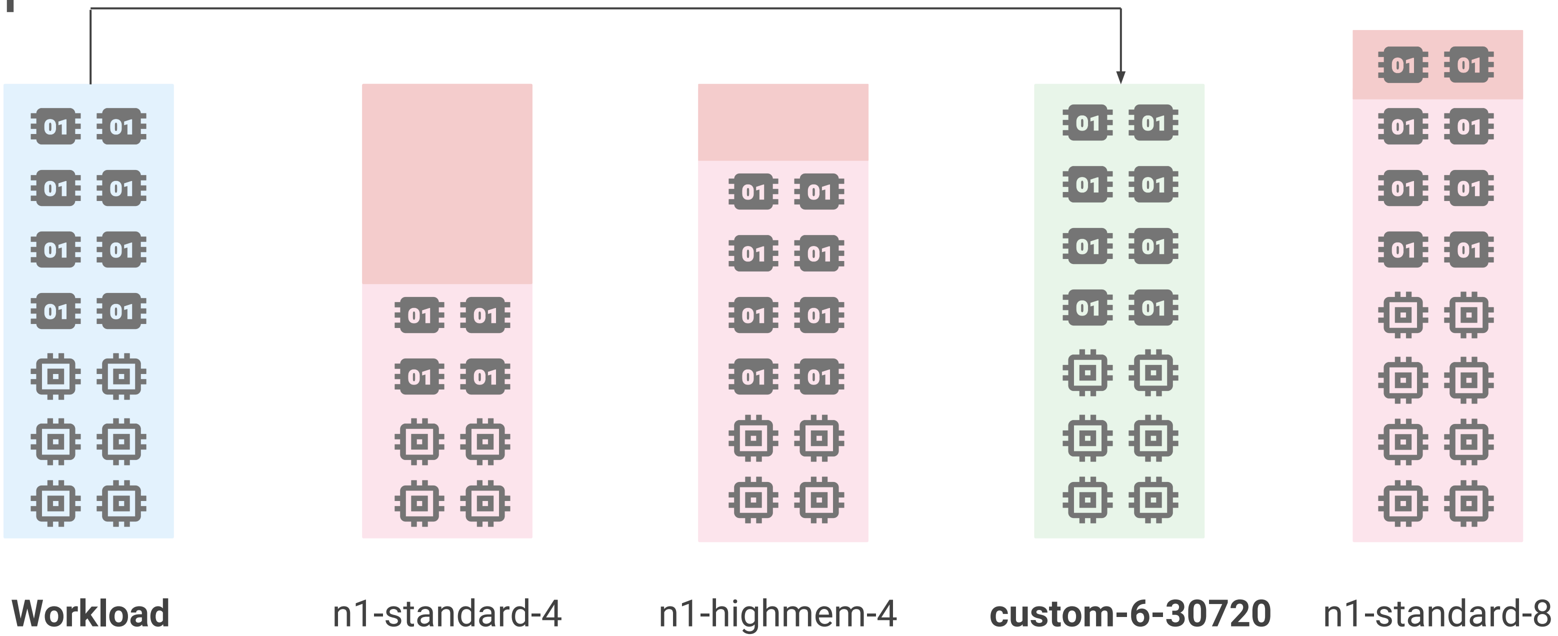
Disaggregation of storage and compute



Cost savings through preemptible VMs

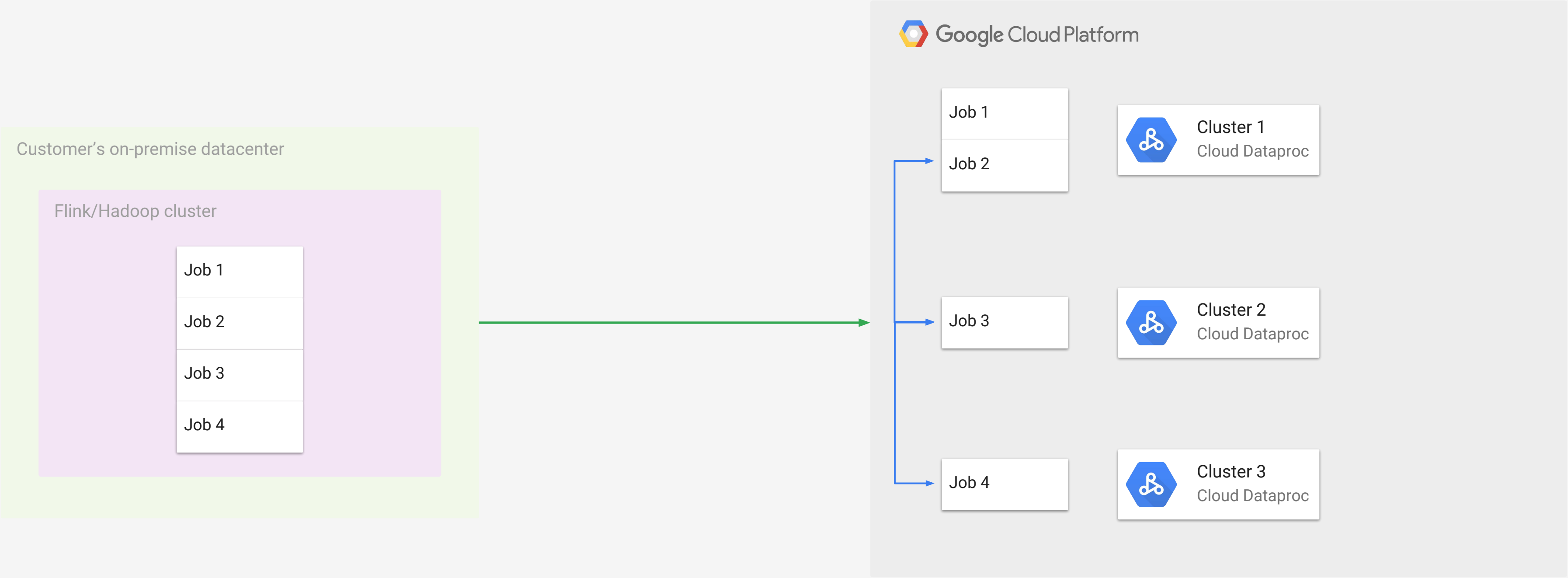


Right-sizing your hardware with custom machine types

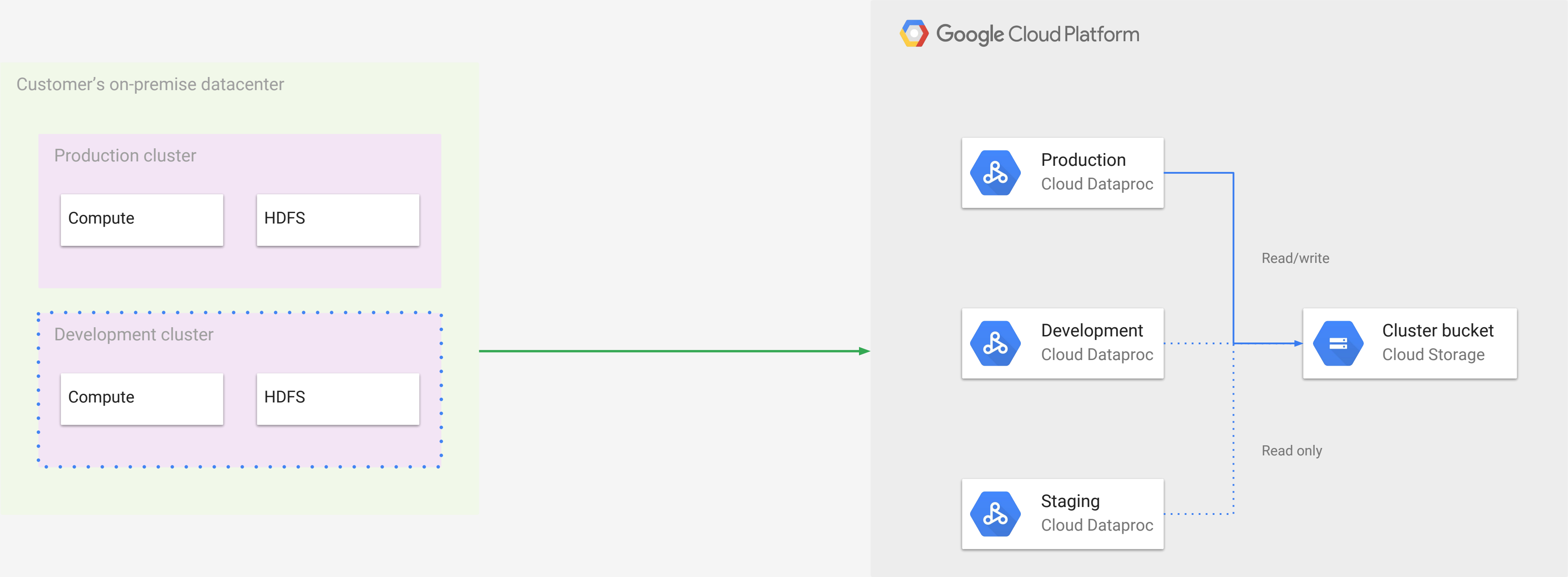


Google Cloud

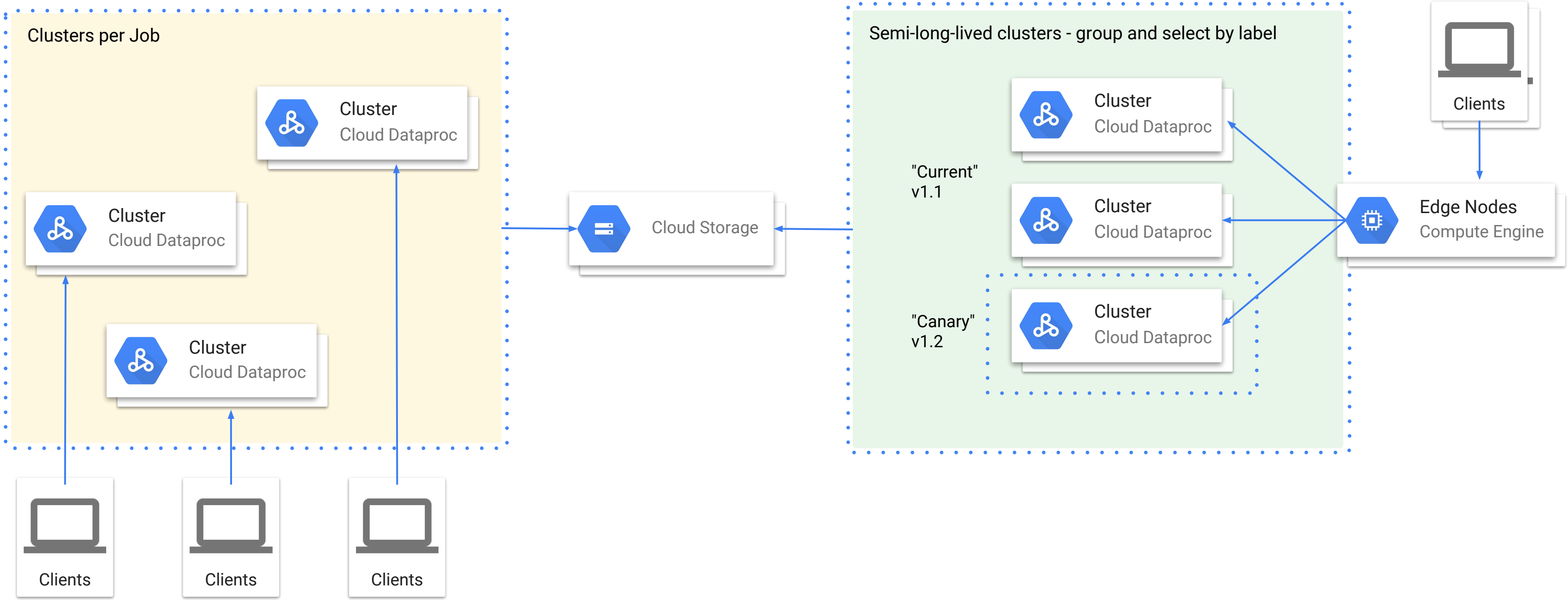
Split clusters and jobs



Separate development and production



Ephemeral and semi-long-lived clusters



Questions and **next steps**

Getting started

Codelabs - codelabs.developers.google.com/codelabs/cloud-dataproc-starter/

Cloud Dataproc quickstarts - cloud.google.com/dataproc/docs/quickstarts

Cloud Dataproc tutorials - cloud.google.com/dataproc/docs/tutorials

Cloud Dataproc initialization actions - github.com/GoogleCloudPlatform/dataproc-initialization-actions

Getting help

Cloud Dataproc documentation - cloud.google.com/dataproc/docs

Cloud Dataproc release notes - cloud.google.com/dataproc/docs

Stack Overflow - google-cloud-dataproc

Cloud Dataproc email discussion - cloud-dataproc-discuss@googlegroups.com

Google Cloud Support - cloud.google.com/support

Thank you

https://goo.gl/Qyf5U7