Google Cloud

Introduction to Cloud Dataproc

Data Engineering on Google Cloud Platform

Agenda

Why unstructured data?

Why Cloud Dataproc?

Creating a Dataproc cluster + Lab

Custom machine types

Sources of data

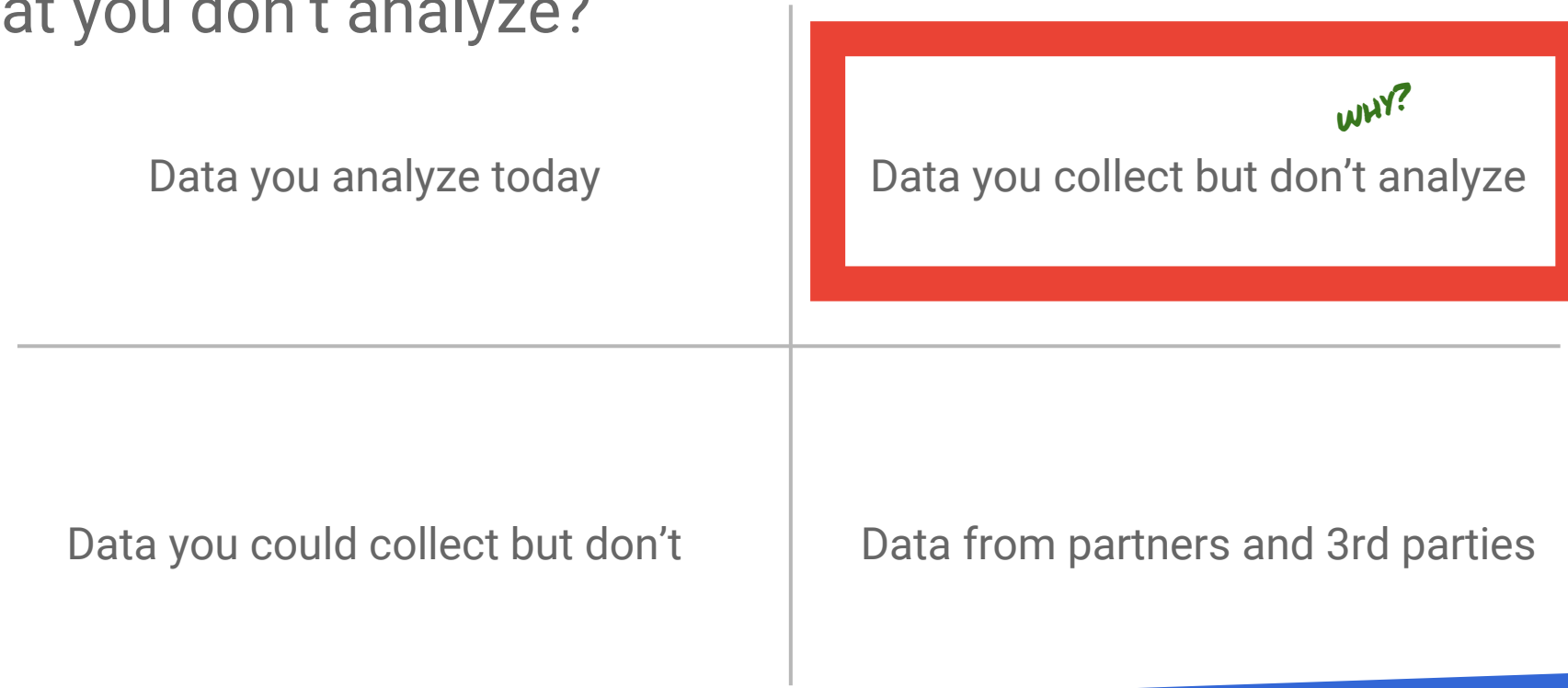
Data you analyze today

Data you collect but don’t analyze

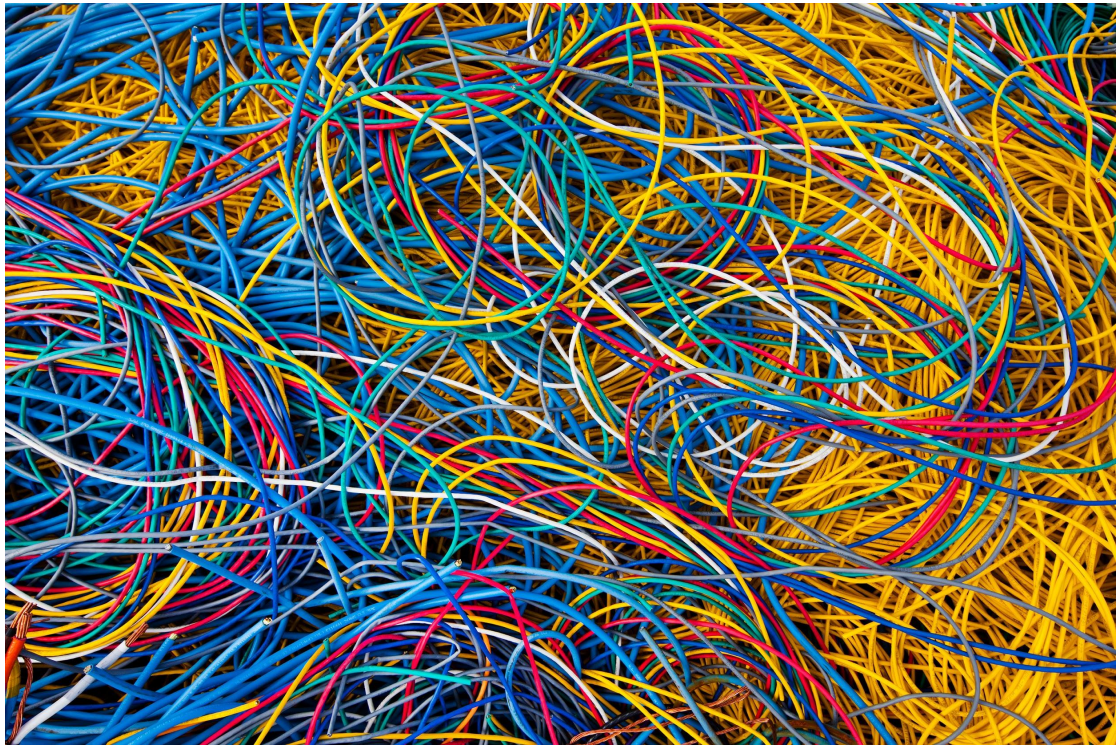
Data you could collect but don’t

Data from partners and 3rd parties

What are some reasons that you have data that you don't analyze?

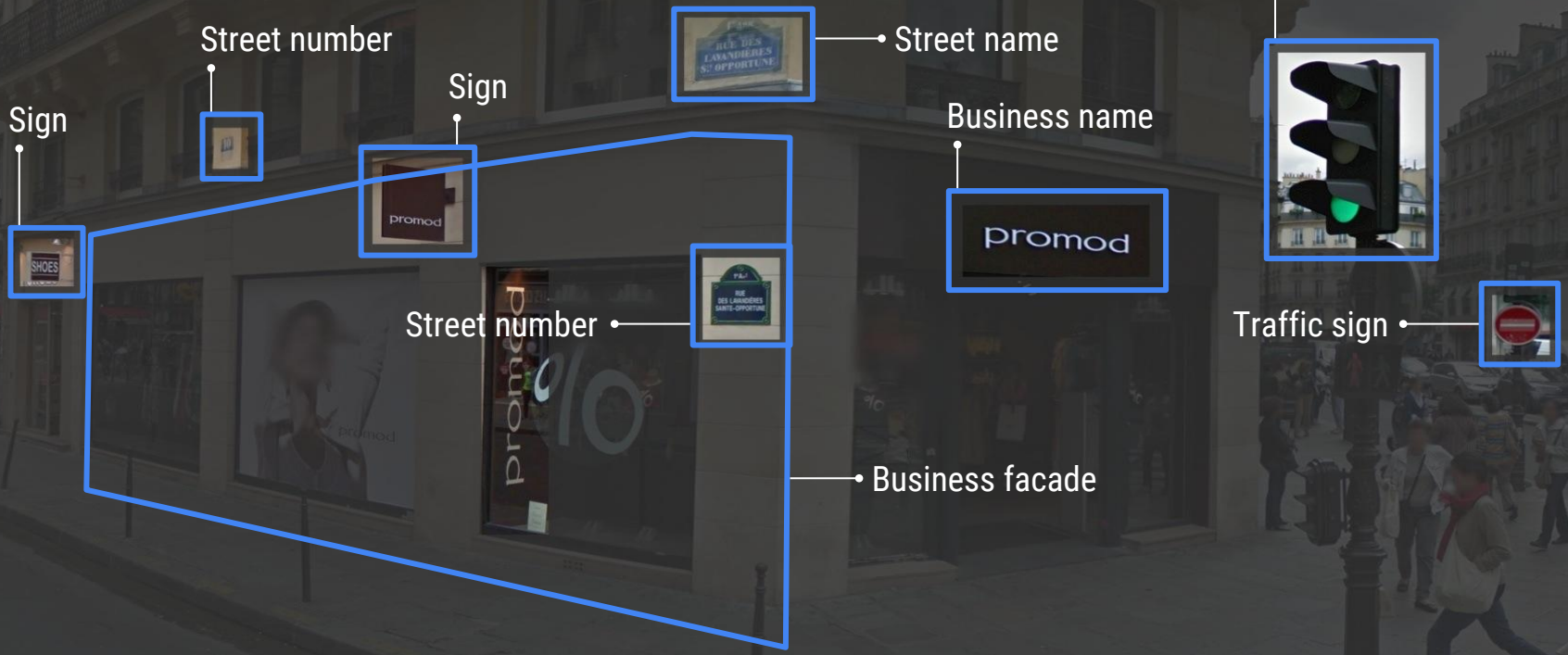


Unstructured data accounts for 90% of enterprise data*




Street View - Apr 2016
promod
promod
promod
CELIOCLUB

Finding new value in data



Some Big Data applications involve a human

Human

Real-time insight into supply chain operations. Which partner is causing issues?

Drive product decisions. How do people really use feature X?

Counting problems are perfect for big data analytical tools

Human

Real-time insight into supply chain operations. Which partner is causing issues?

Drive product decisions. How do people really use feature X?

Easy counting problems

Did error rates decrease after the bug fix was applied?

Which stores are experiencing long delays in payment processing?

These are also counting problems, but they are not as easy...

Human

Real-time insight into supply chain operations. Which partner is causing issues?

Drive product decisions. How do people really use feature X?

Easy counting problems

Did error rates decrease after the bug fix was applied?

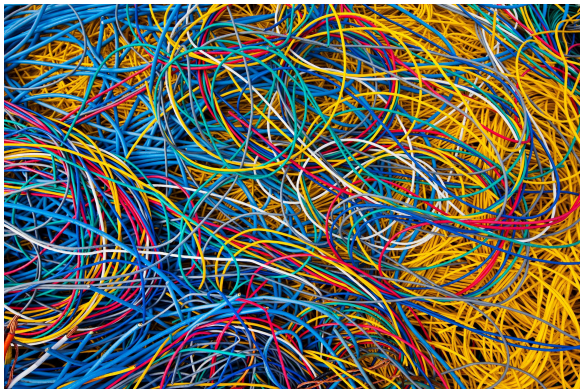
Which stores are experiencing long delays in payment processing?

Harder counting problems

Are programmers checking in low-quality code?

Which stores are experiencing lacking of parking space?

Build on top of Google



Images
Audio
Video
Free-form text

ML API

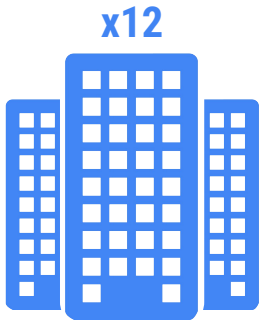
Places
Labels
People
Events
...

Cloud ML Engine

Agenda

Why Cloud Dataproc?

Will you ever have a PetaByte of data?



A stack of floppy disks higher than twelve empire state buildings

27 years to download over 4G

100 Libraries of Congress

Every tweet ever twittered...50 times

How *small* is a PetaByte?

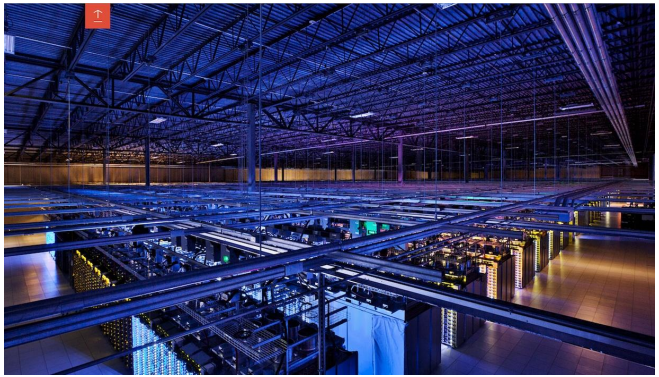
2 micrograms of DNA

1 day's worth of video uploaded to YouTube

200 servers logging at 50 entries per second for 3 years

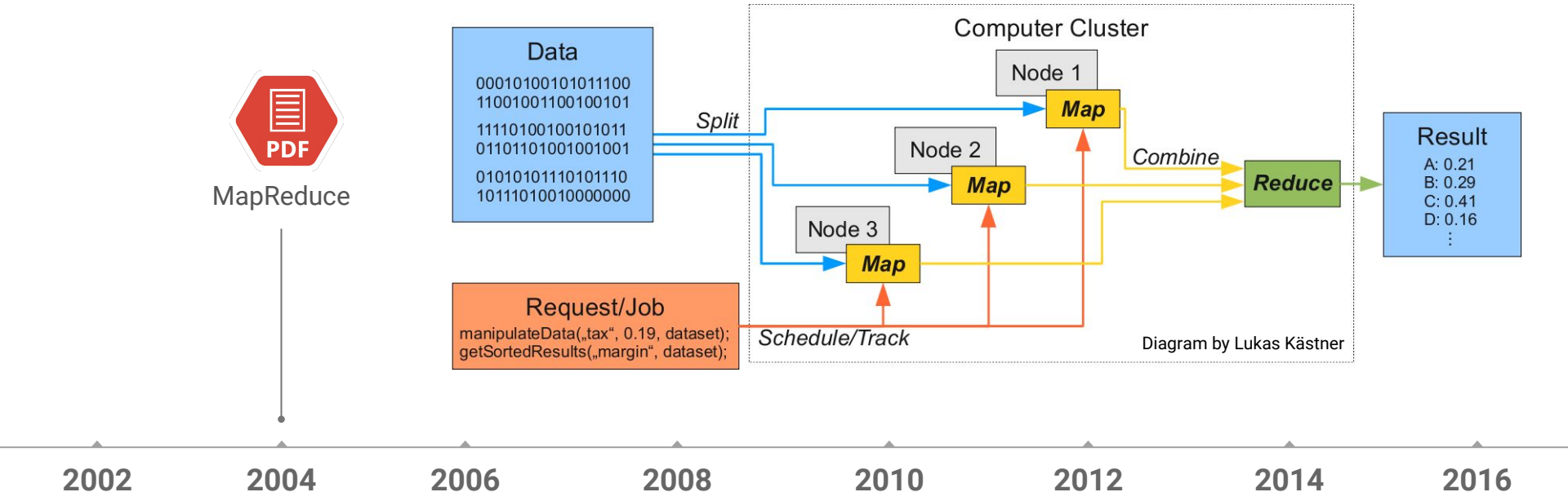
How do you process large amounts of data?

Scaling up

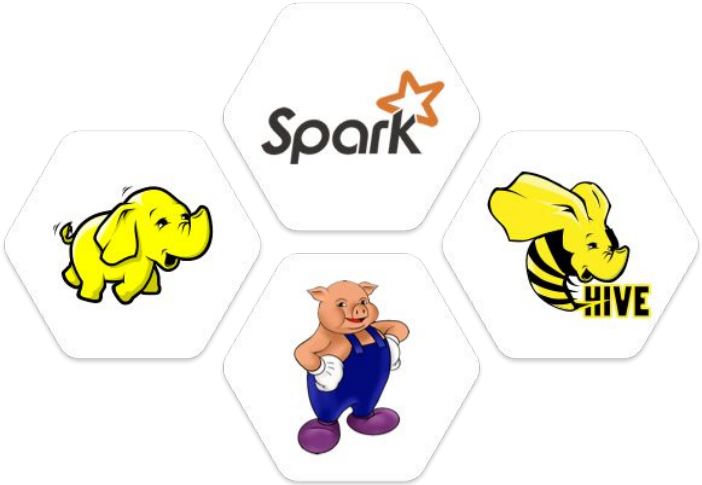


Scaling out

MapReduce approach splits Big Data so that each compute node processes data local to it

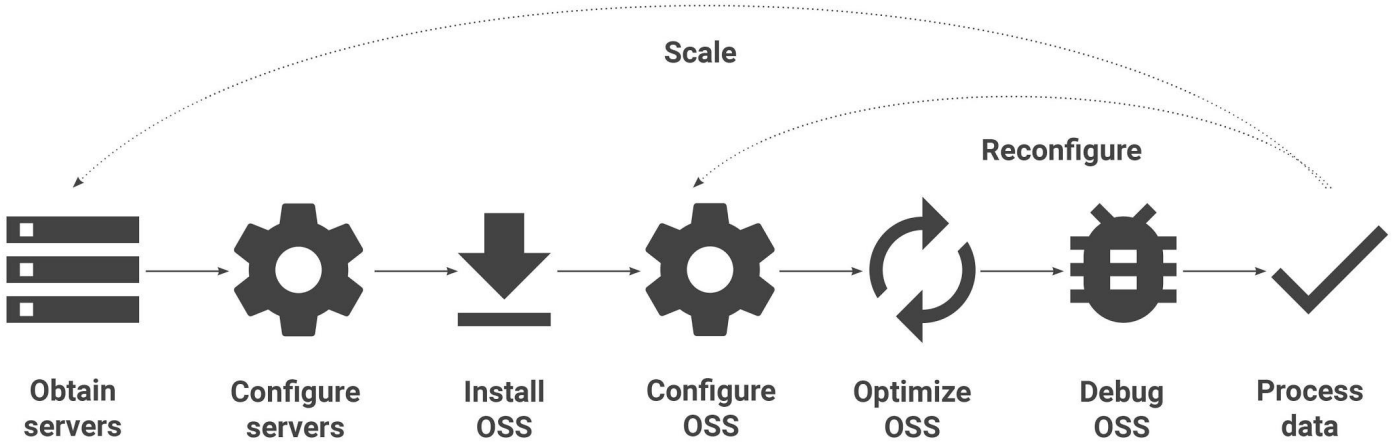


Many on-premise applications for Big Data are built using the open-source stack

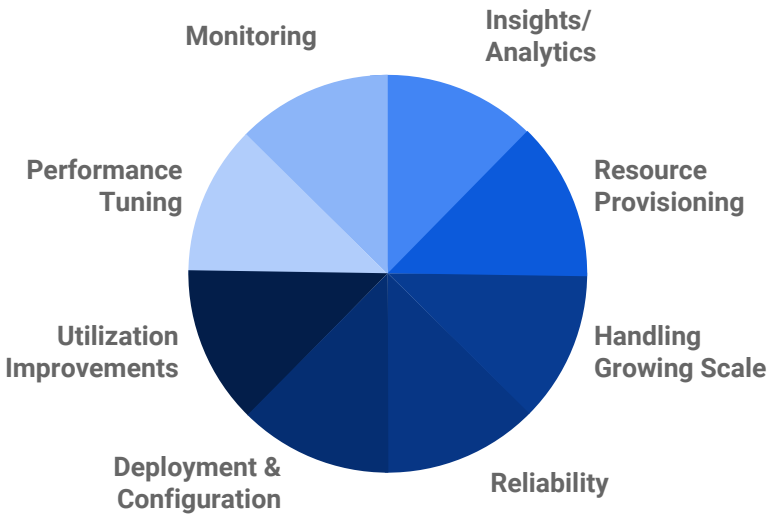


Apache Spark is a popular, flexible, powerful way to process large datasets

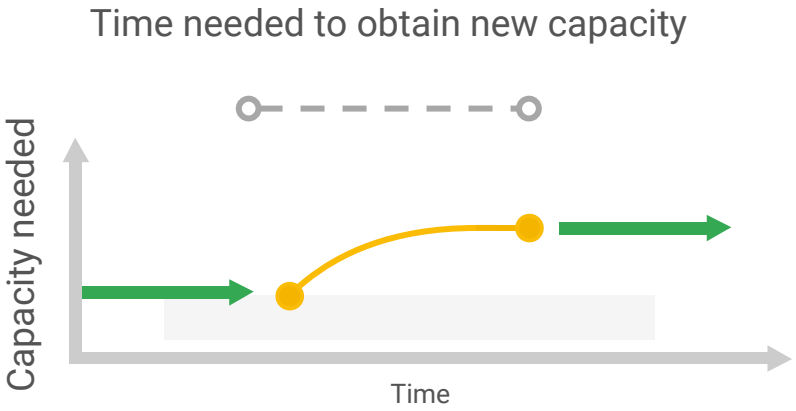
Typical Spark and Hadoop deployments involve...



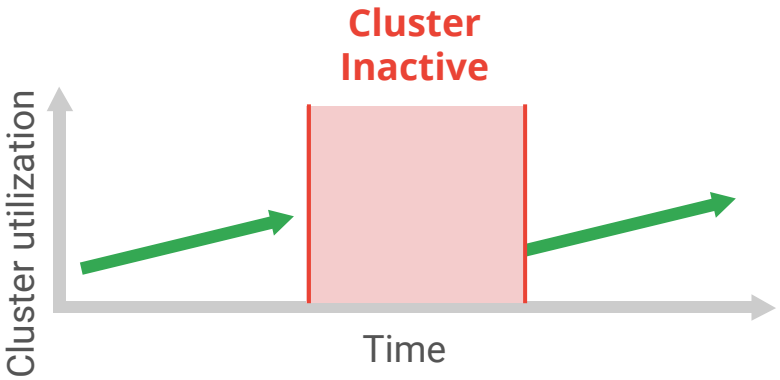
Lots of time is spent on administration and operational issues



Scaling can take hours, days, or weeks

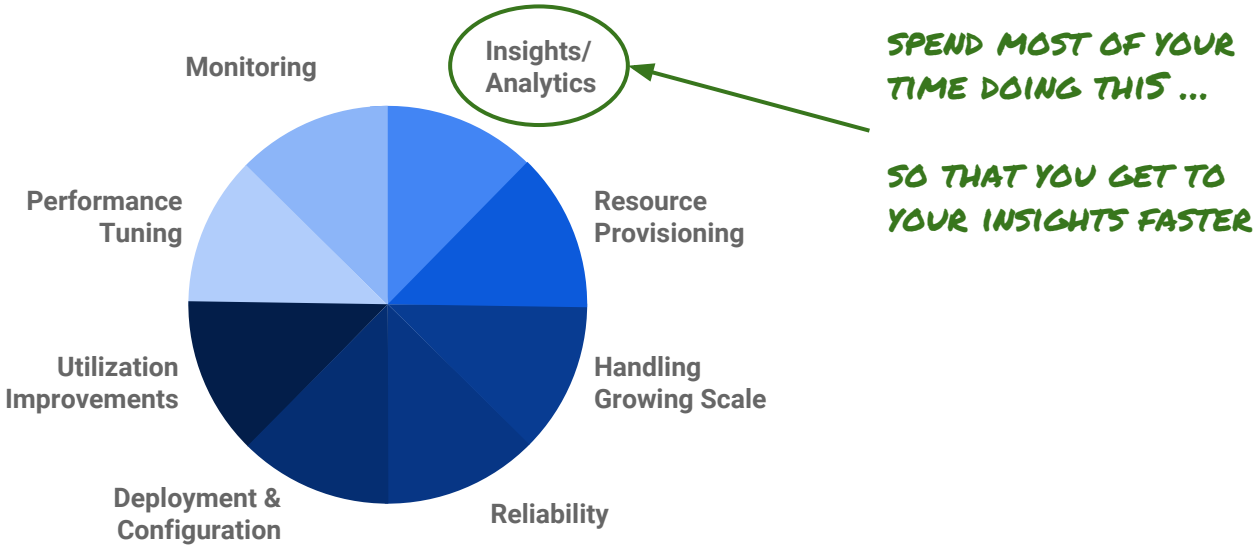


You have to babysit utilization

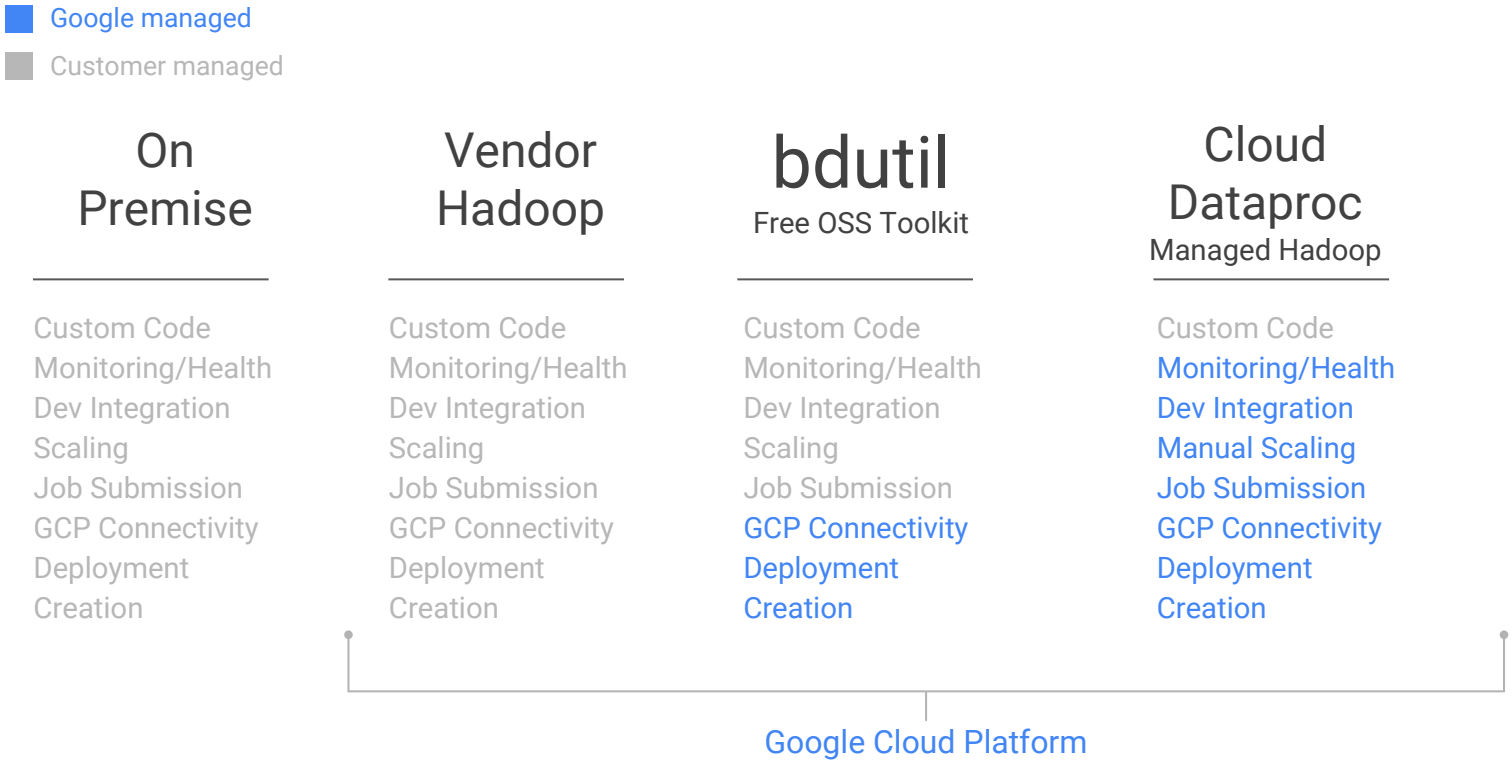


ENSURING CLUSTER IS ALWAYS BEING UTILIZED IS HARD

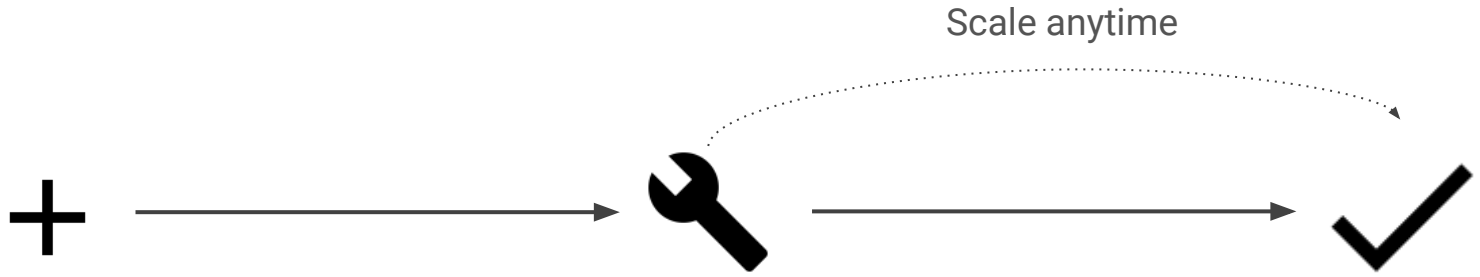
But you want to focus on insights and analytics



Dataproc eases Hadoop management



Typical Dataproc deployments involve...



Create cluster

(0 seconds)

Configure cluster

(20 seconds)

Use cluster

(90 seconds)

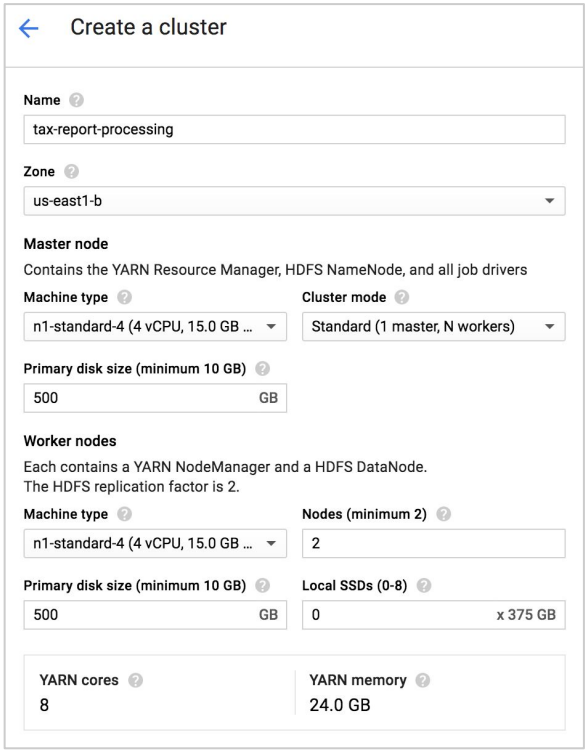
Cloud Dataproc provides compelling reasons to run open-source tools on GCP

- Stateless clusters in <90 seconds
- Supports Hadoop, Spark, Pig, Hive, etc.
- High-level APIs for job submission
- Connectors to Bigtable, BigQuery, Cloud Storage

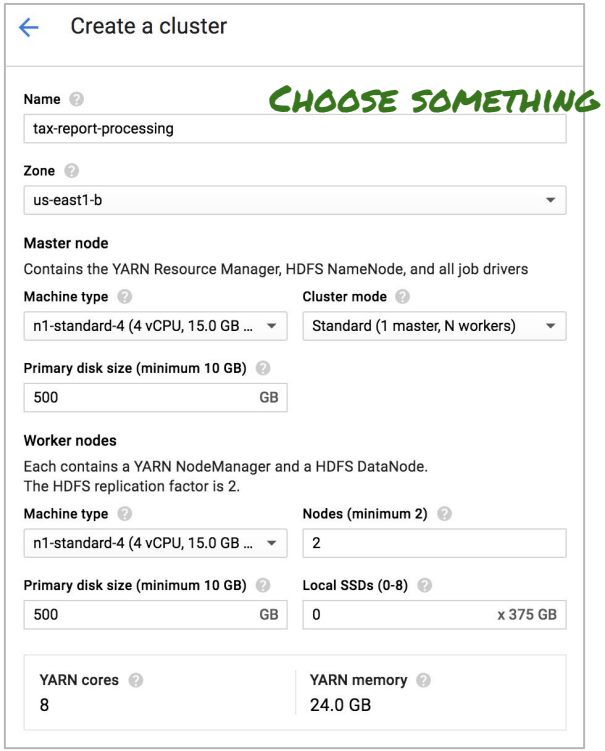
Agenda

Creating a Dataproc cluster + Lab

Create a cluster from the web console

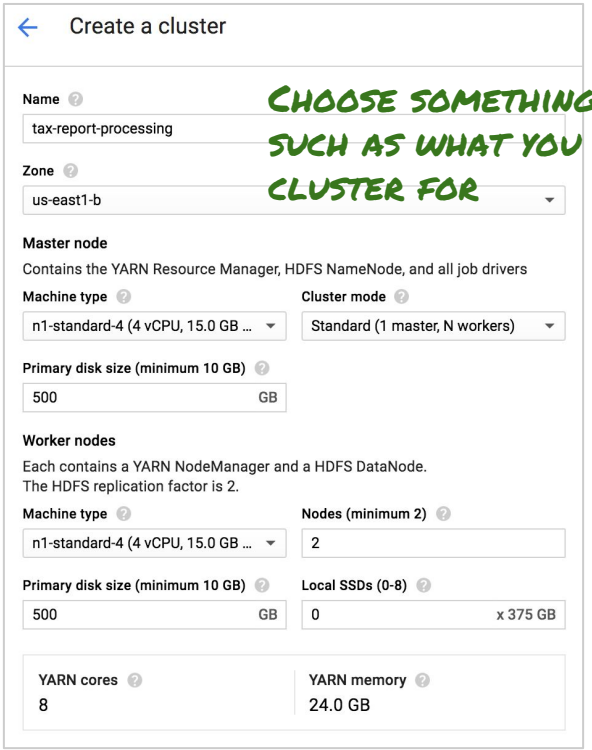


The name needs to be unique within your project



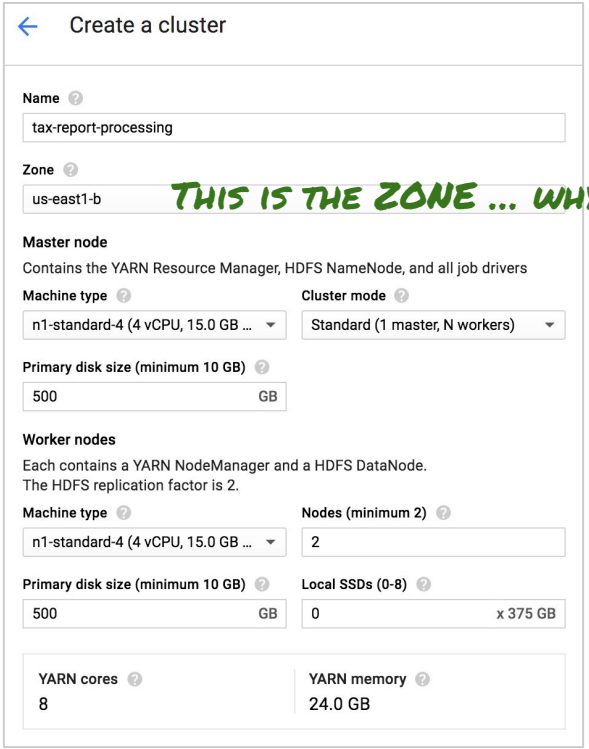
CHOOSE SOMETHING YOU WILL REMEMBER

One cluster per job

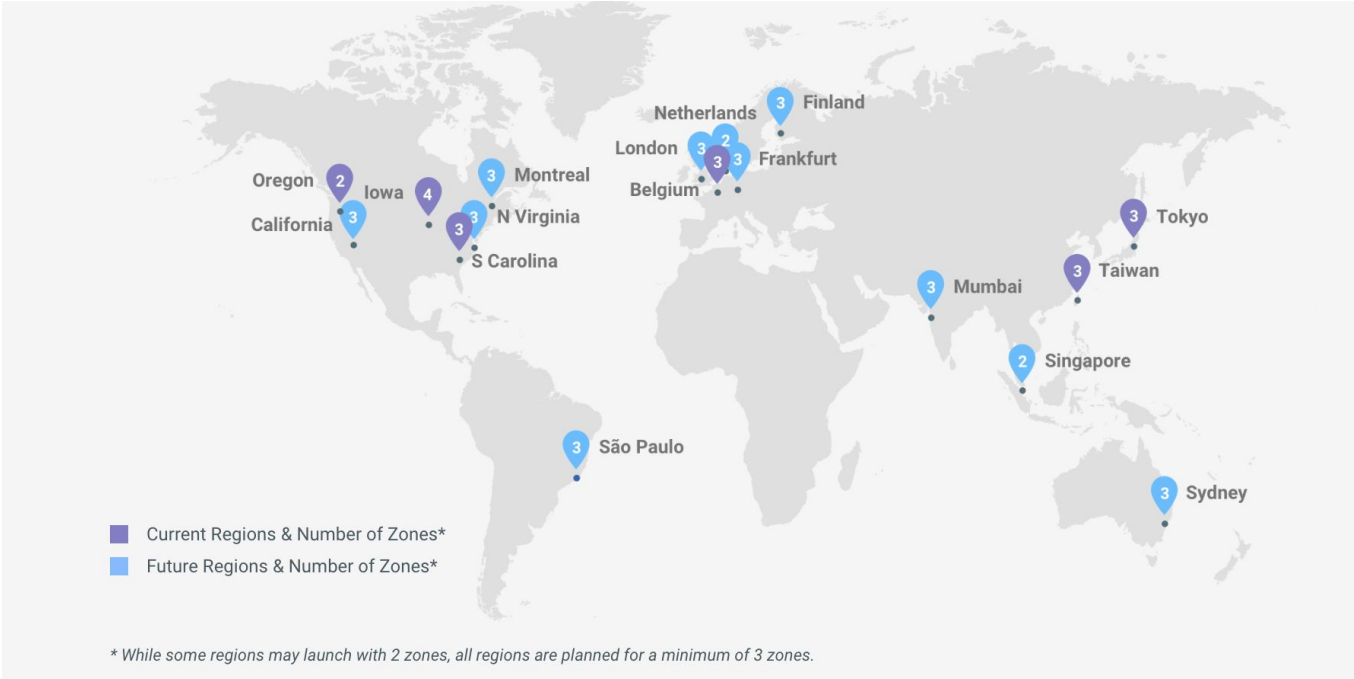


Choose something you will remember, such as what you are going to use the cluster for

The zone is very, very important

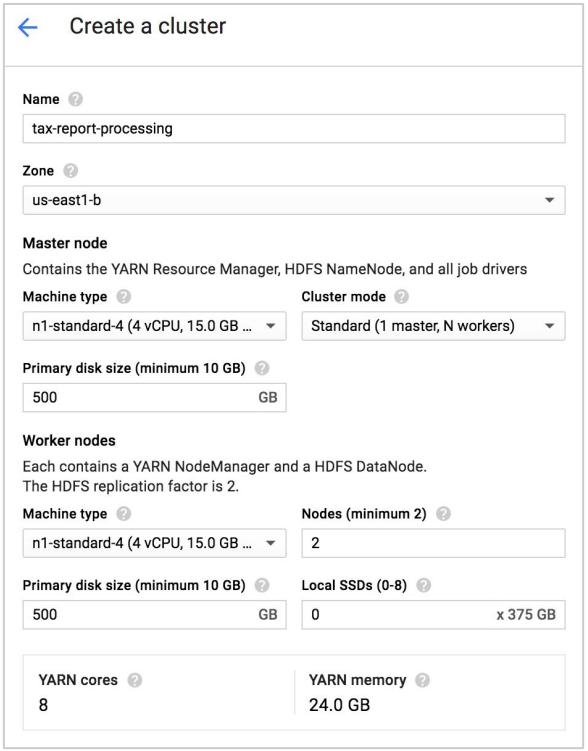


The zone is where the compute nodes will live



Match your data location with your compute location (same region)

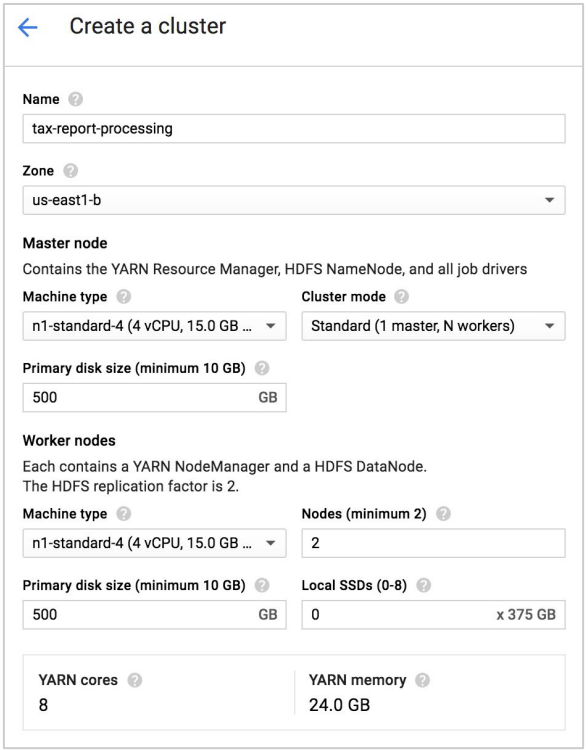
Three cluster configurations possible



The master node manages the cluster
Choose between:

1. Single node (for experimentation)
2. Standard (1 master only)
3. High availability (3 masters)

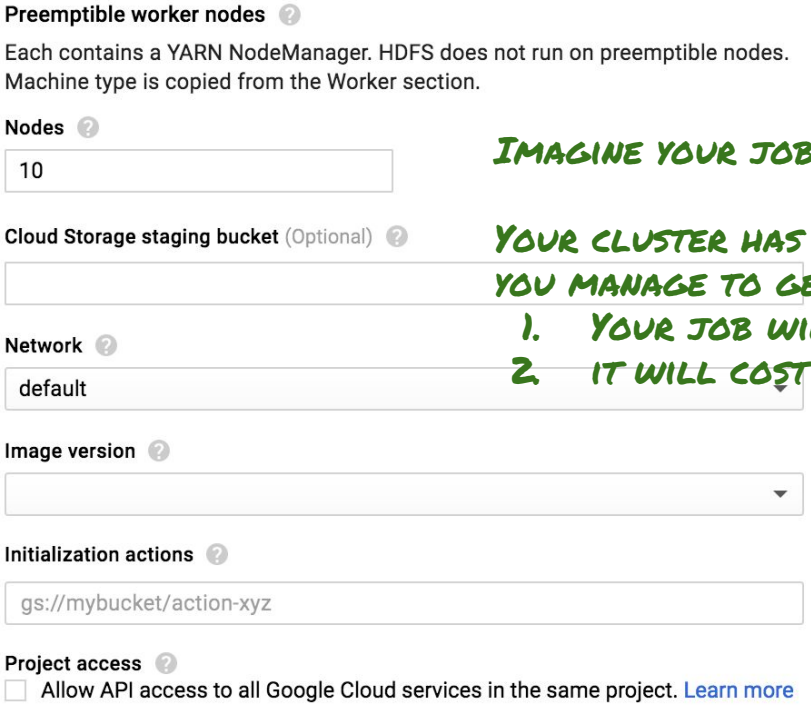
HDFS file system available, but don't use it



MACHINE TYPE, NUMBER OF WORKERS

DISK PERFORMANCE SCALES WITH SIZE!!!
DON'T USE HDFS TO STORE INPUT/OUTPUT DATA

Preemptible workers can be a good deal

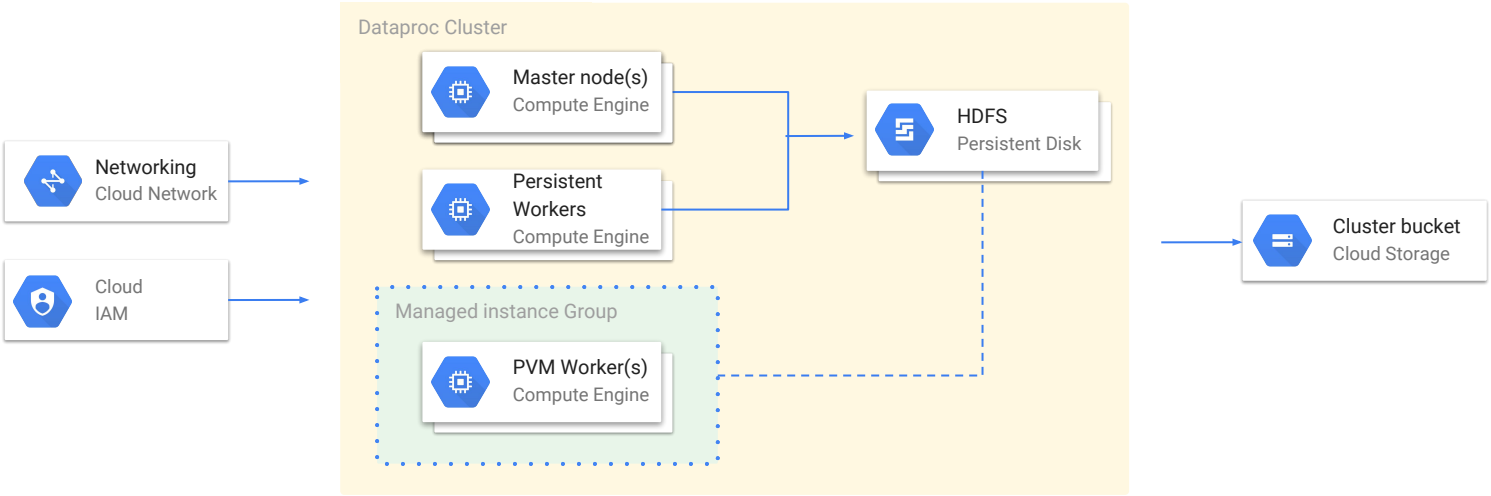


Imagine your job needs 10 machines for 130 minutes

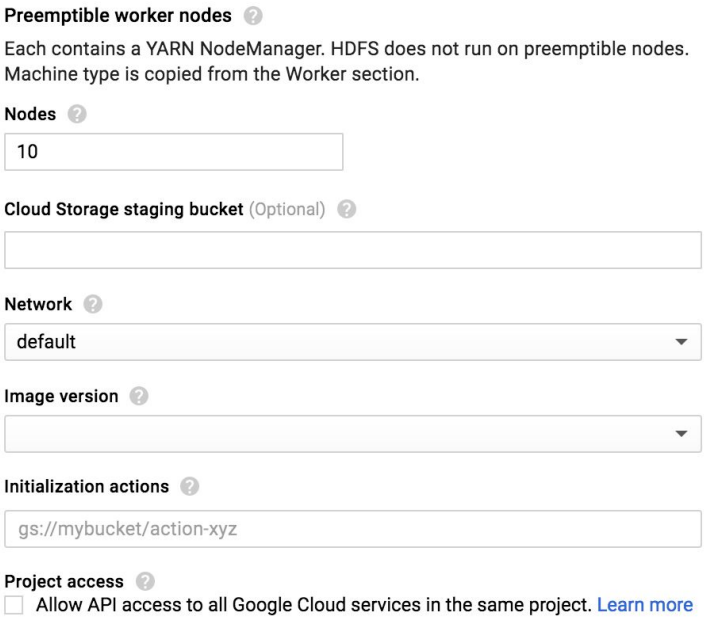
Your cluster has 10 standard workers and ...
you manage to get 10 preemptible machines

1. Your job will now finish in 65 minutes!
2. It will cost 40% less overall!

Dataproc manages joins/leaves of preemptible instances



Can customize the Dataproc cluster



Can set up firewall rules etc.

Can also install custom software on the Dataproc workers and Master

Most things you can do from the web console...

WEB CONSOLE

GCLOUD SDK COMMAND LINE

```
gcloud dataproc clusters
        --master-machine
        --num-workers 2
        --worker-boot-di
```

REST API call

Google Cloud Platform

CUSTOM SOFTWARE ...

Creating a cluster using gcloud SDK

```
gcloud dataproc clusters create my-second-cluster --zone us-central1-a \
        --master-machine-type n1-standard-1 --master-boot-disk-size 50 \
        --num-workers 2 --worker-machine-type n1-standard-1 \
        --worker-boot-disk-size 50
```

CONTEXT-SPECIFIC HELP

```
gcloud dataproc --help
gcloud dataproc clusters --help
gcloud dataproc clusters create --help
```

Lab - Leveraging Unstructured Data : Part 1

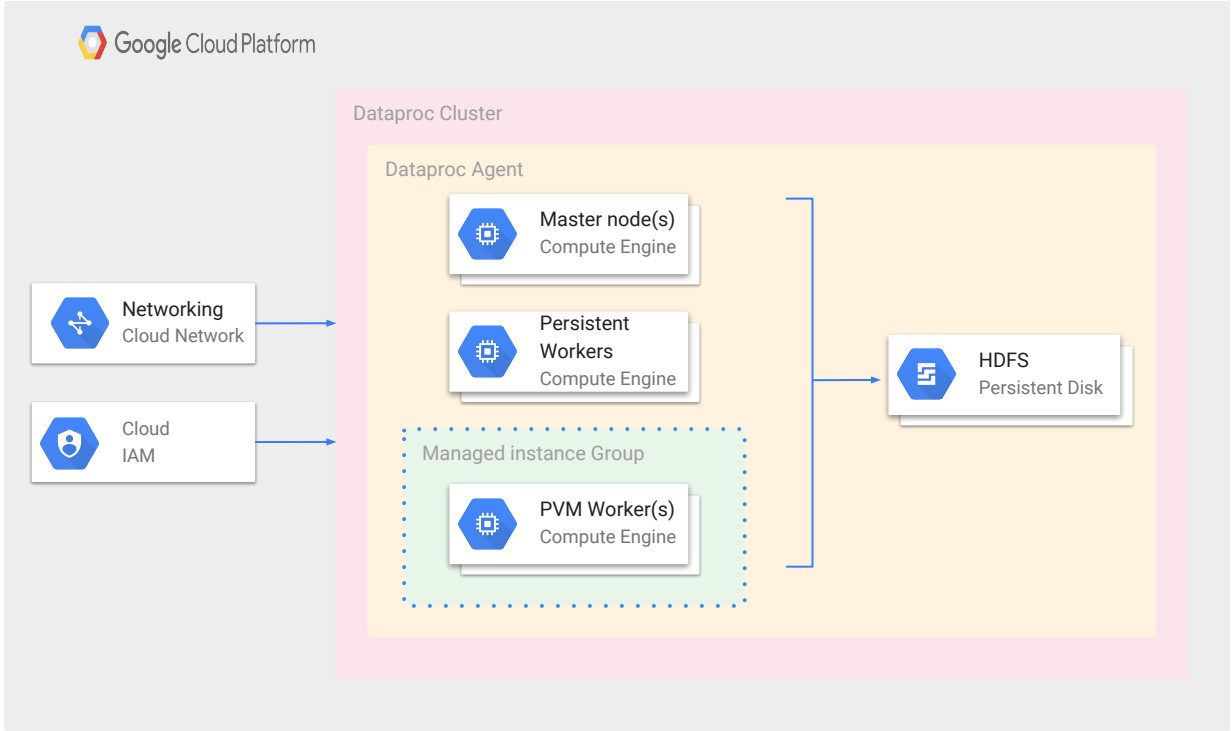
You will learn how to:

- Create a Dataproc cluster from the Web console
- SSH into the cluster
- Add a firewall rule that allows access to your cluster from the browser
- Create, manage and delete Dataproc clusters from the CLI

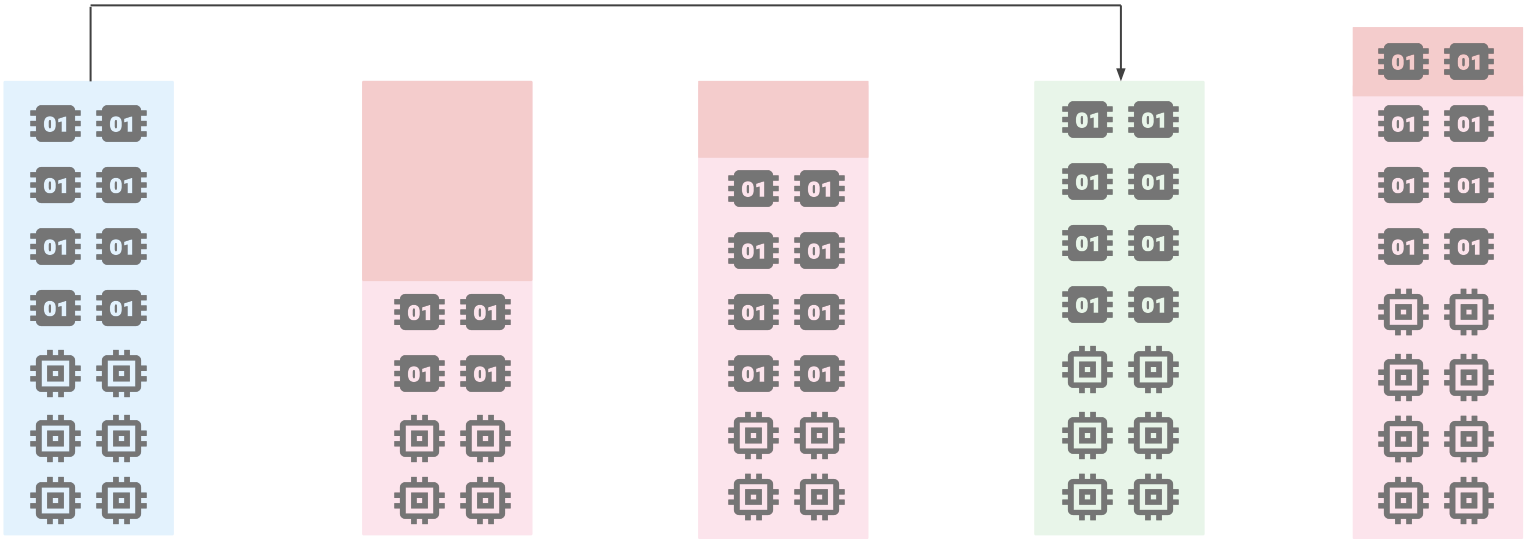
Agenda

Custom machine types

Cloud Dataproc hardware architecture



Right-size your workload using machine types

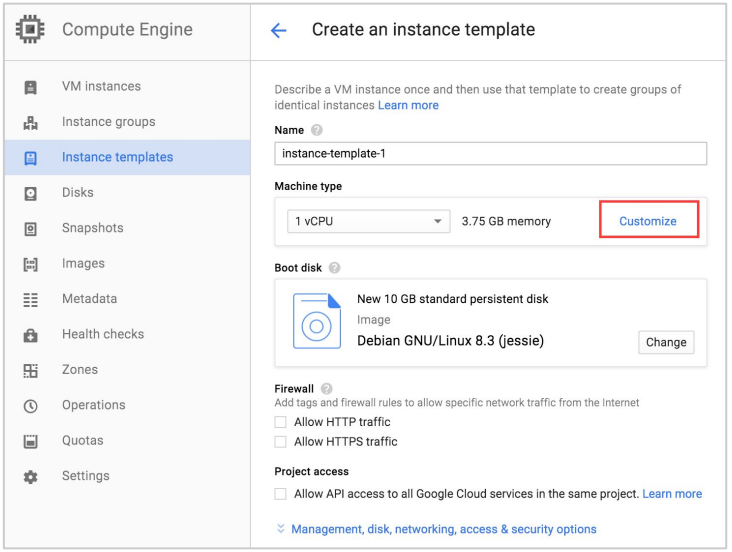


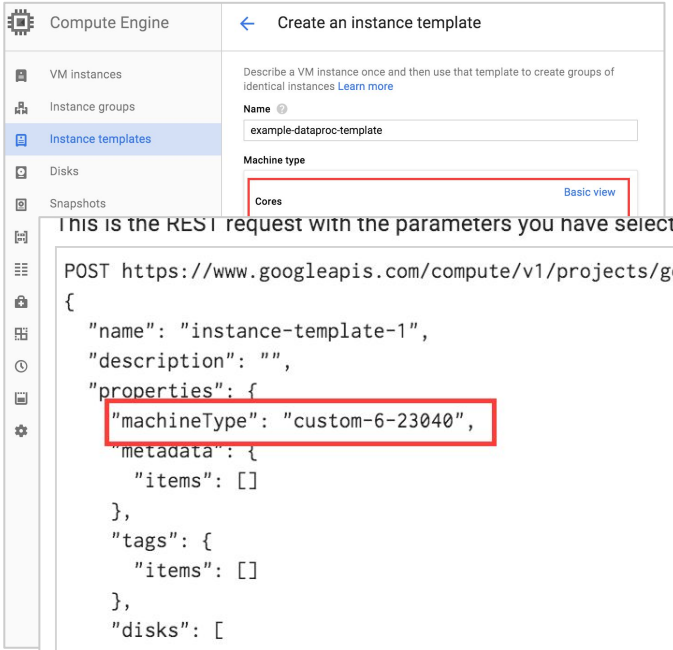
Creating the custom machine type...

```
gcloud dataproc clusters create test-cluster /
    --worker-machine-type custom-6-30720 /
    --master-machine-type custom-6-23040
```

6 CPUs
30 GB * 1024 = 30720

You can find the name from the web console





This is the REST request with the parameters you have selected.

```
POST https://www.googleapis.com/compute/v1/projects/google.com:hadoop-demo/global/
{
  "name": "instance-template-1",
  "description": "",
  "properties": {
    "machineType": "custom-6-23040",
    "metadata": {
      "items": []
    },
    "tags": {
      "items": []
    },
    "disks": [
      {
        "type": "PERSISTENT",
        "boot": true,
        "mode": "READ_WRITE",
```


cloud.google.com
Images by Connie Zhou