Building and Developing Science Gateways with Google Cloud Platform

Alicia Salmeron (salmeron@google.com) Higher-Ed & Research
Karan Bhatia, Phd (karanbhatia@google.com) Scientific Computing

Seven products with over
one billion users each,
all powered by the cloud.

Google Search

Android

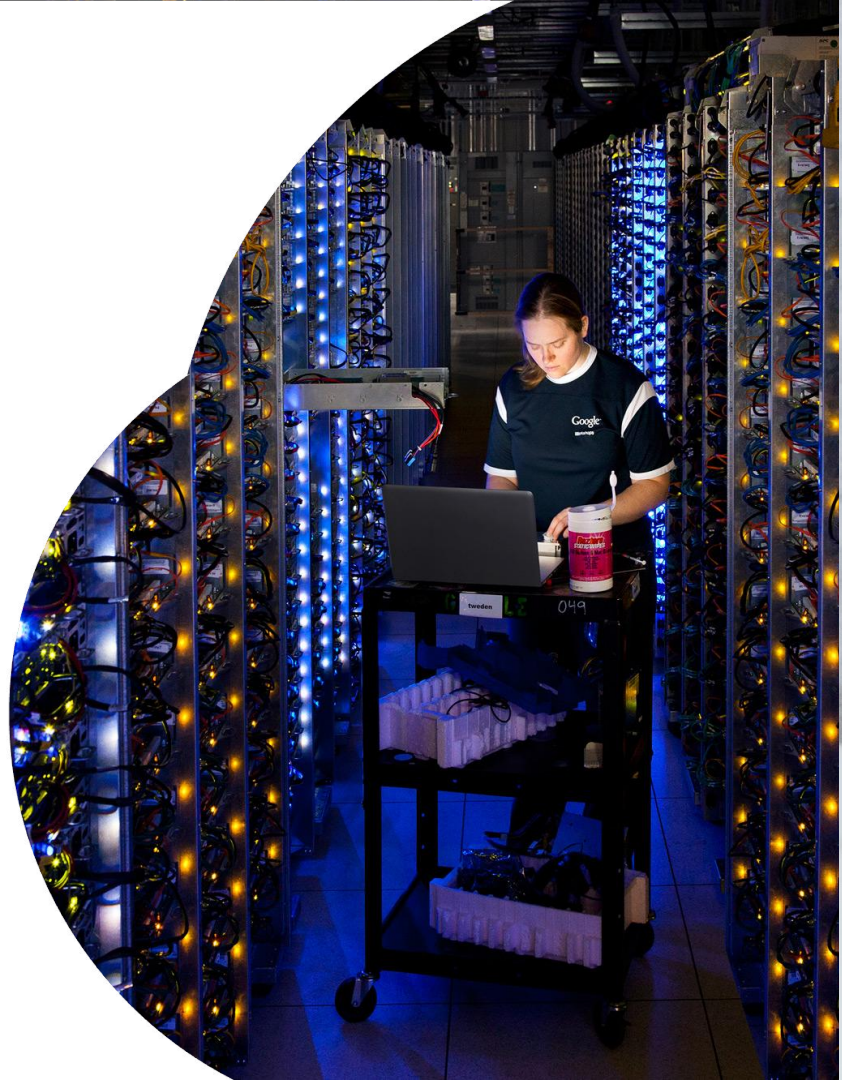
Chrome

Google Maps

Google Play

YouTube

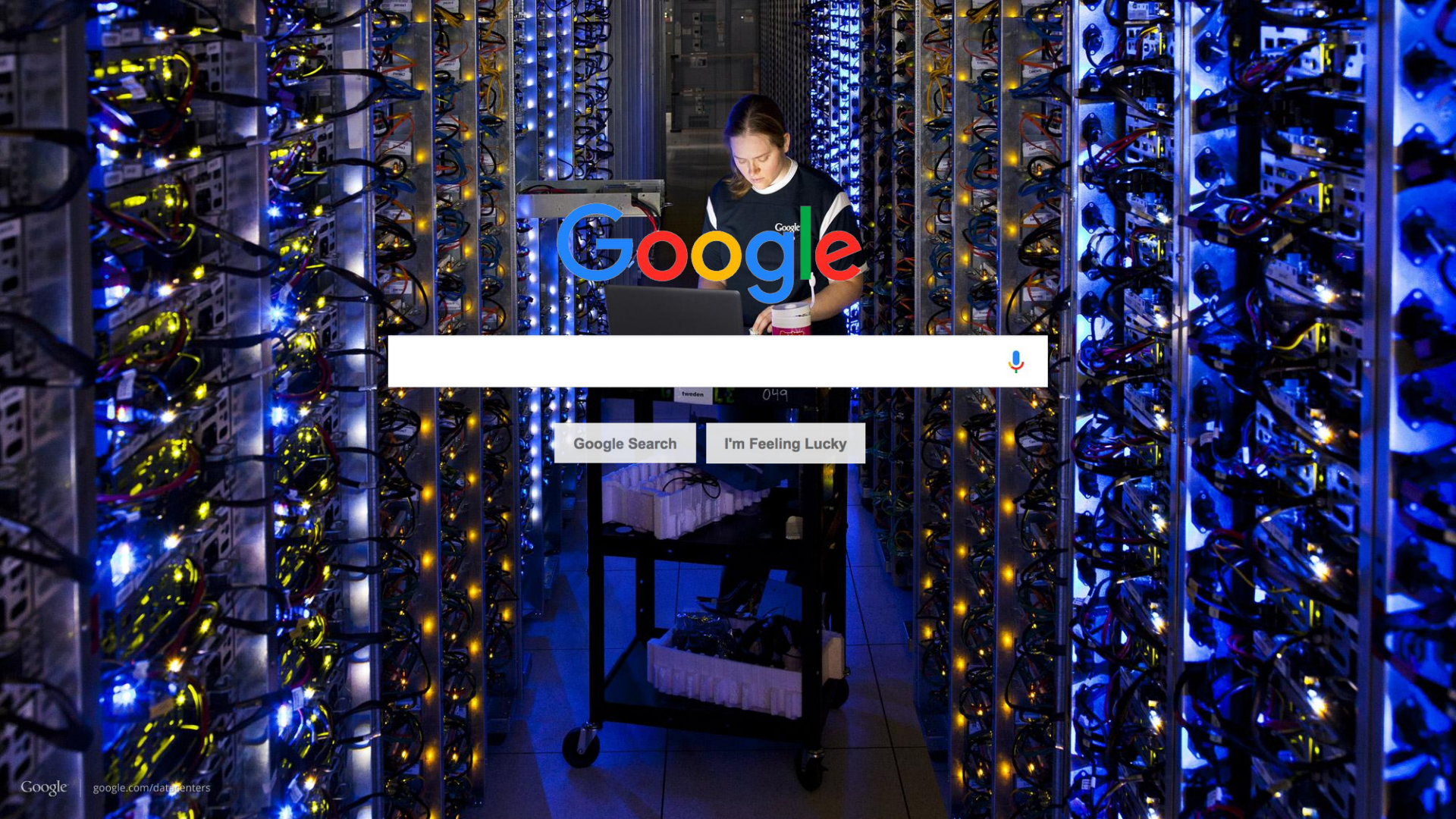
Gmail



Google Cloud

Google Search I'm Feeling Lucky

Google
Google Search
I'm Feeling Lucky


Google
Google Search
I'm Feeling Lucky

Our customers serve over
1 Billion End Users each day

How is Google Cloud Platform Different?

Global Network & Security

Compute & Data Analytics

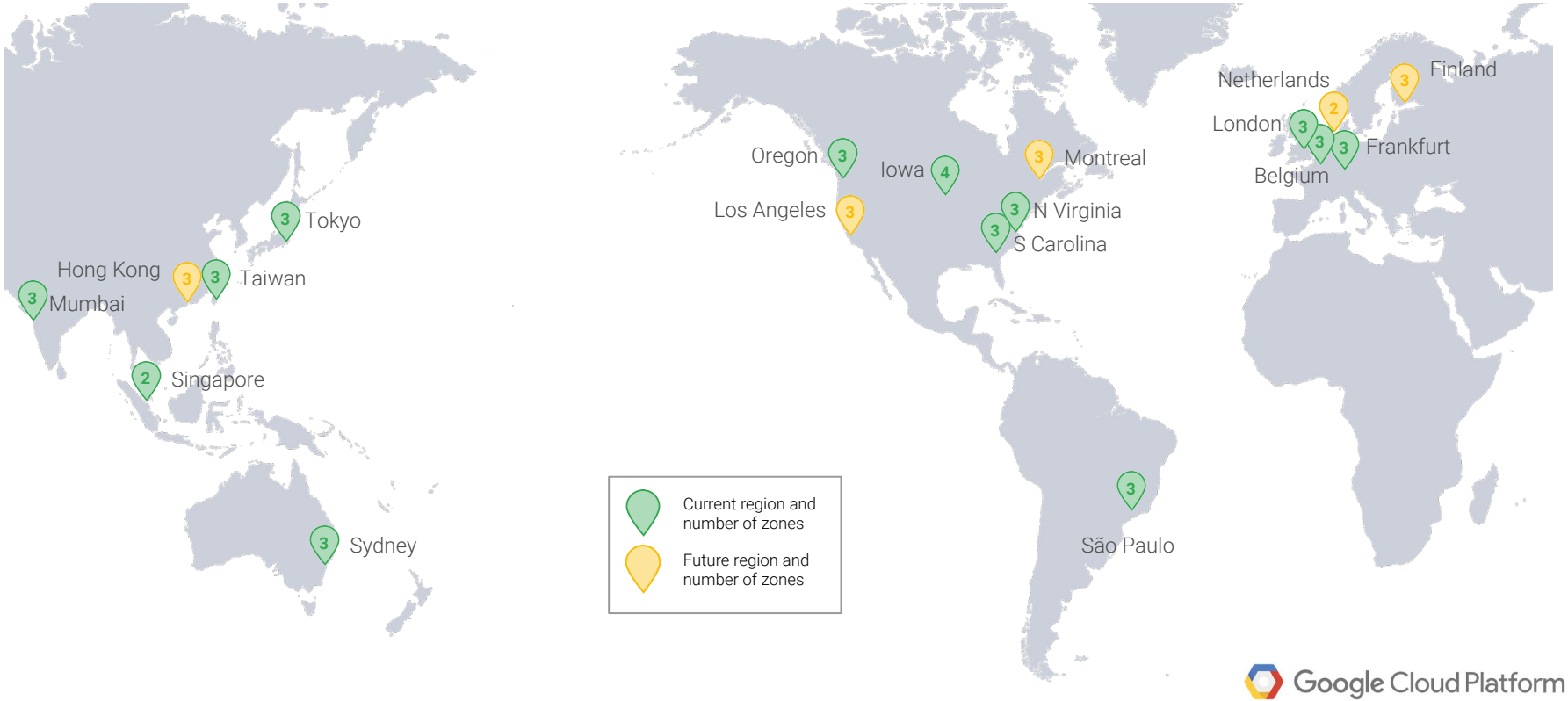
Dev Tools and Open Source Ecosystem

Use Cases & EDU Programs

Google Cloud

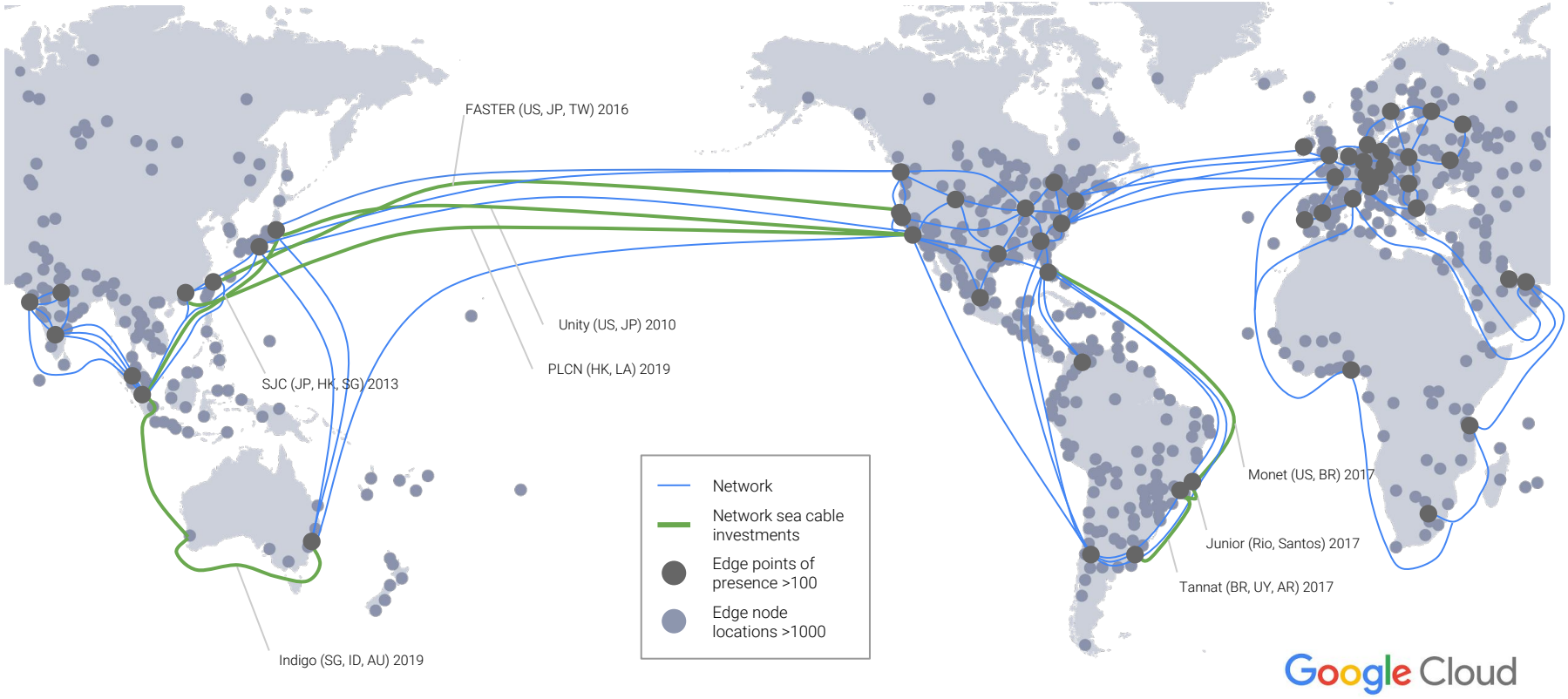
Google Cloud Platform Regions

Select from 13 regions. 5 new regions coming in 2018.



Google Cloud Network

The largest cloud network, comprised of >100 points of presence



Google Cloud Network

The largest cloud network, comprised of >100 points of presence

More than 100
Peering Locations

World's Largest
Software Defined
Network

Global Load
Balancing with
Single IP

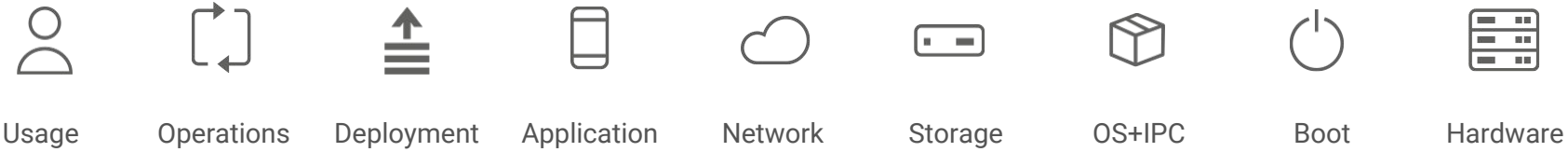
Seamless Autoscale
to **Over 1M** Queries
Per Second with
no pre-warming

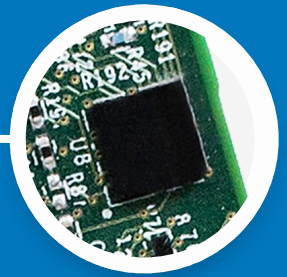
Global Content
Delivery Network

Edge Locations in
Virtually Every Country

Google Cloud

Layered **Defense in Depth** Security

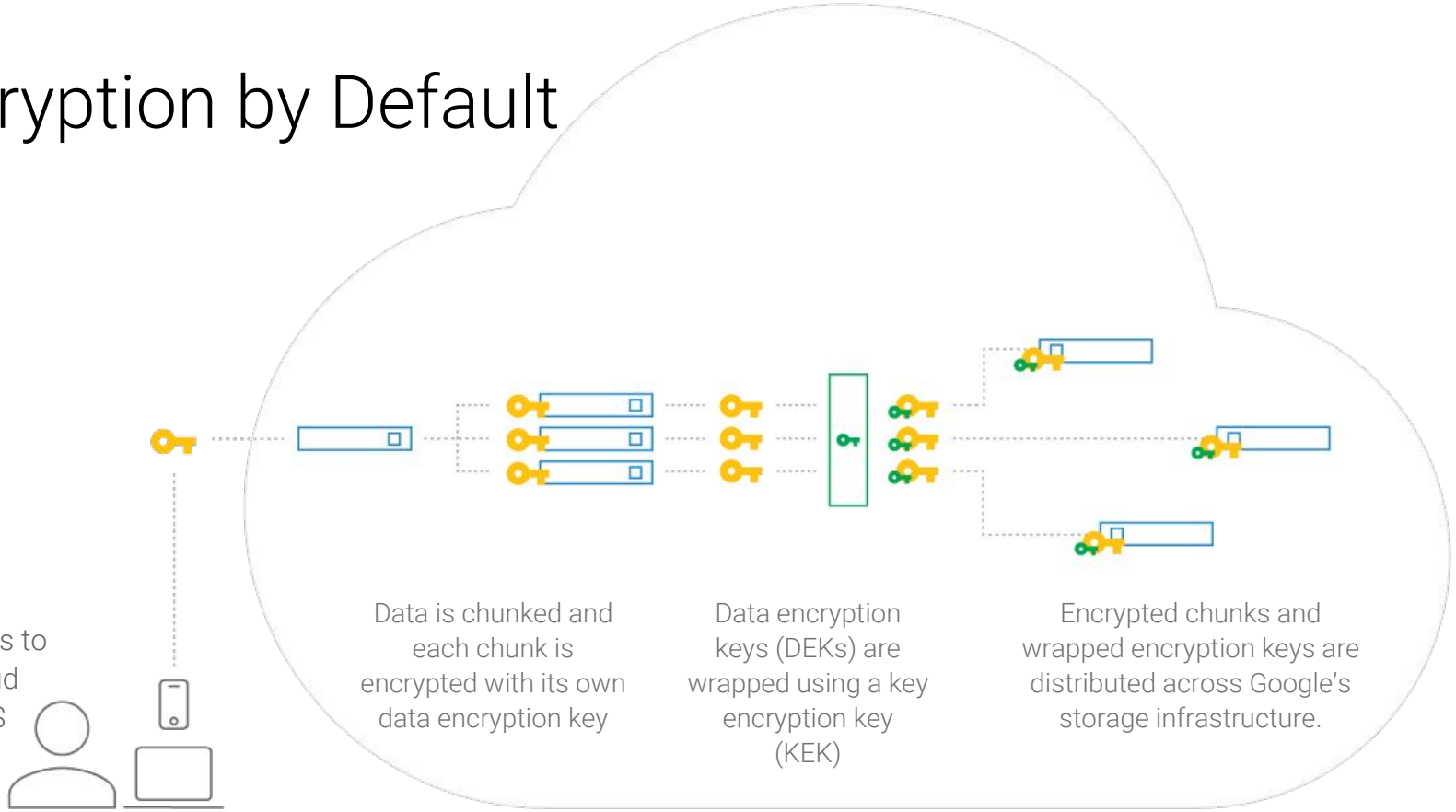




Titan

Google's purpose-built chip to establish **hardware root of trust** for both machines and peripherals on cloud infrastructure

Encryption by Default



All connections to Google Cloud require TLS

Data is chunked and each chunk is encrypted with its own data encryption key

Data encryption keys (DEKs) are wrapped using a key encryption key (KEK)

Encrypted chunks and wrapped encryption keys are distributed across Google's storage infrastructure.

Third-party audits and certification

ISO 27001

ISO 27017

ISO 27018

HIPAA

ISAE 3402 Type II

AICPA SOC

AICPA SOC

SSAE 15 Type II

PCI DSS v3.1

FedRAMP ATO

Privacy Shield
Framework

CSA STAR

What makes Google Cloud Platform Different?

Global Network & Security

Compute & Data Analytics

Dev Tools and Open Source Ecosystem

Use Cases & EDU Programs

Google Cloud

Infrastructure

Lightning fast & scalable: Fast VM startup time, millisecond access for all storage classes, high IOPS for VCPUs, high bandwidth global networking

Reliable: Built-in redundancy and scale, live Migration, Google Site Reliability Engineering for your workload.

Customer friendly pricing: simple and efficient: Pay-per-second, custom VMs, automatic discounts, flexible buy-in-bulk discounts

Geographic coverage: 11 new regions in 2017-18 for a total of 17, HA in each region

- Significant “per core” performance improvements
- Intel® Advanced Vector Extension 51 (Intel® AVX-512)
  - 2x flops/second
- Accelerated IO with Intel® Omni-Path Architecture (Fabric)
- Integrated Intel® QuickAssist Technology (crypto & compression offload)
- Intel® Resource Director Technology (Intel® RDT) for Efficiency & TCO

Google Cloud Platform Blog

Product updates, customer stories, and tips and tricks on Google Cloud Platform

Google Cloud Platform is the first cloud provider to offer Intel Skylake

Friday, February 24, 2017

By Urs Hölzle, Senior Vice President, Google Cloud Infrastructure

I’m excited to announce that Google Cloud Platform (GCP) is the first cloud provider to offer the next generation Intel Xeon processor, codenamed Skylake.

Hardware Accelerated

- Available Today: **NVIDIA K80 GPU, P100s**
- Coming Soon: **Tensor Processing Unit (TPU)**
- Custom ASIC built and optimized for TensorFlow
- Used in production at Google for over 16 months
- 7 years ahead of GPU performance per watt

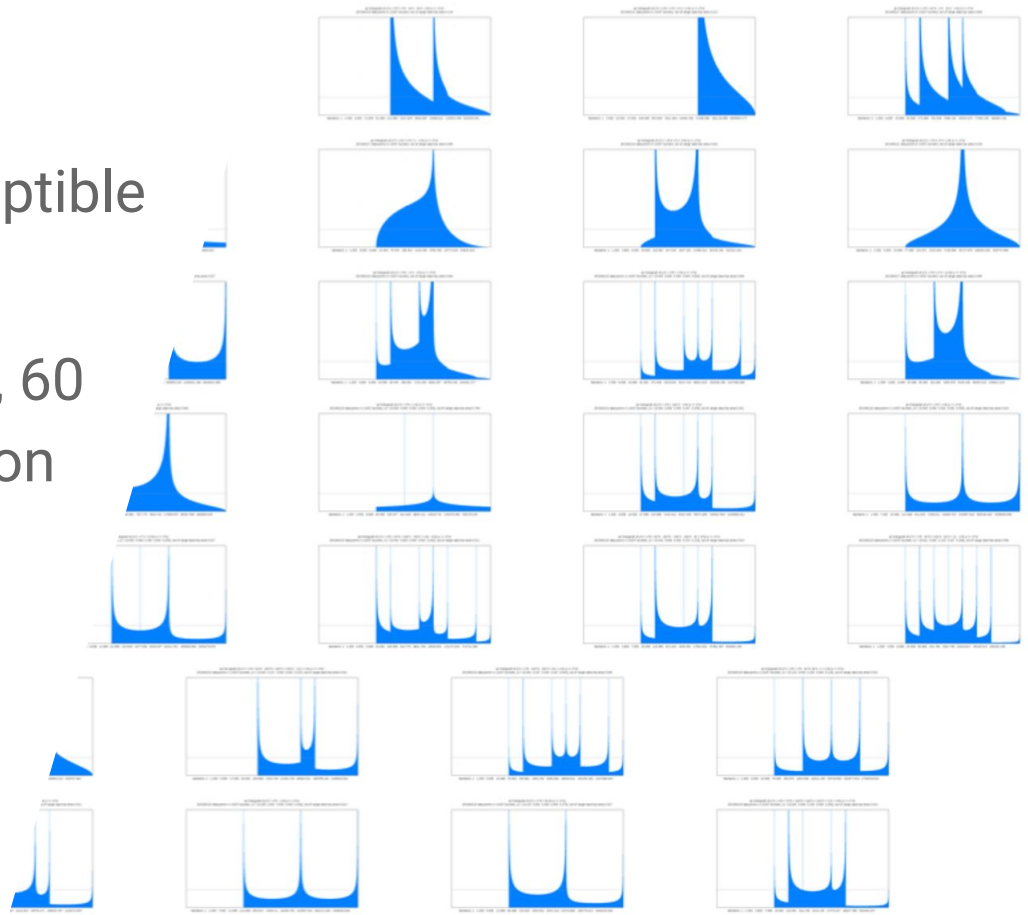
MIT Research w/ VMs

580,000 cores

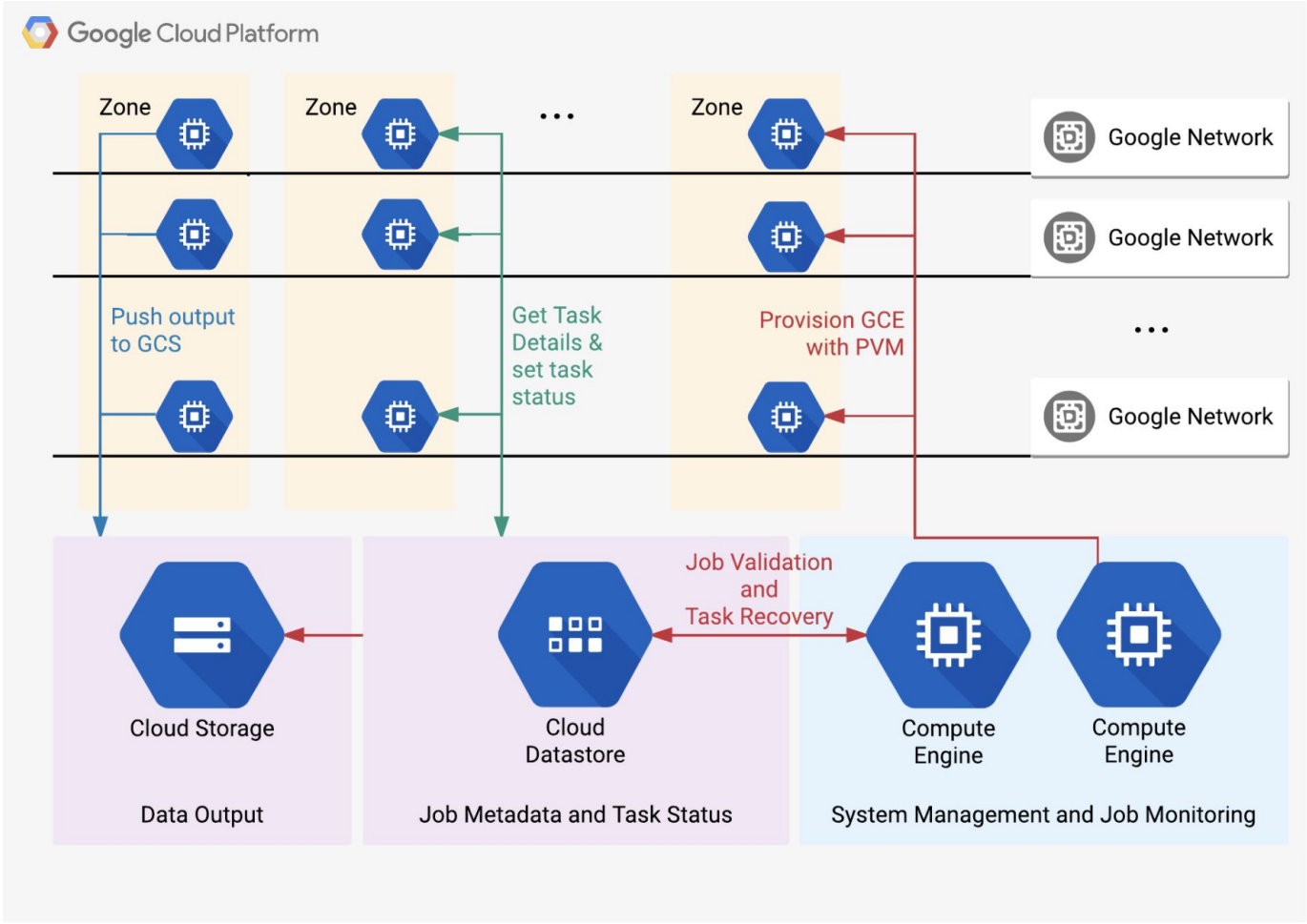
~~220,000~~ cores on preemptible VMs

2,250 32-core instances, 60 CPU-years of computation in a single afternoon

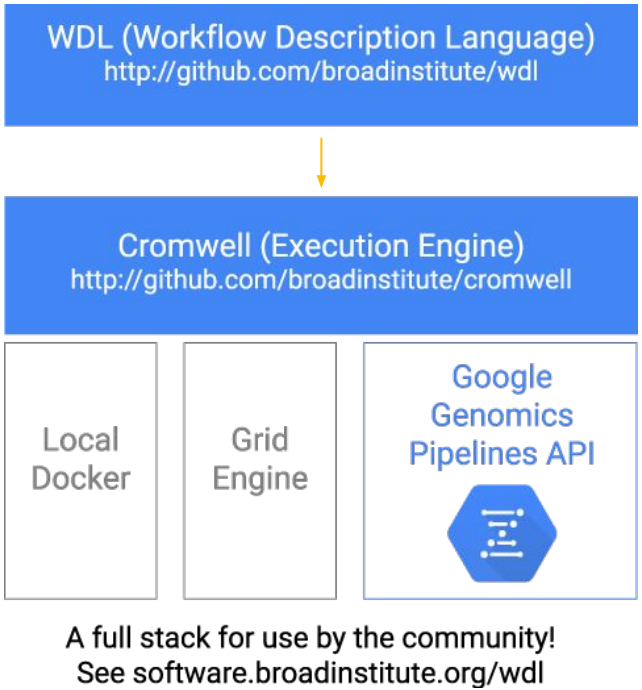
Answers in hours v. months



Products used: Google Compute Engine, Cloud Storage, DataStore


Google Cloud Platform
Zone
Zone
Zone
Google Network
Google Network
Google Network
Push output to GCS
Get Task Details & set task status
Provision GCE with PVM
Job Validation and Task Recovery
Cloud Storage
Cloud Datastore
Compute Engine
Compute Engine
Data Output
Job Metadata and Task Status
System Management and Job Monitoring

Broad Firecloud:
WDL, Cromwell and Google Genomics



A full stack for use by the community!
See software.broadinstitute.org/wdl

WDL: an external DSL used by computational biologists to express the analytical pipelines

Cromwell: a scalable, robust engine for executing WDL against pluggable backends including local, Docker, Grid Engine or …

Google Genomics Pipelines API: co-developed by Broad and Google Genomics, a scalable Docker-as-a-Service with data scheduling

Pipeline definition

```
{
  "name": "samtools index",
  "description": "Run samtools index to generate a BAM index file",
  "inputParameters": [
    {"name": "inputFile",
      "localCopy": {
        "disk": "data",
        "path": "input.bam"
      }
    },
    {"name": "outputFile",
      "localCopy": {
        "disk": "data",
        "path": "output.bam.bai"
      }
    },
  ],
  "resources": {
    "minimumCpuCores": 1,
    "minimumRamGb": 1,
    "disks": [{
      "name": "data",
      "type": "PERSISTENT_HDD"
      "sizeGb": 200,
      "mountPoint": "/mnt/data",
    }]
  },
  "docker": {
    "imageName": "quay.io/cancercollaboratory/dockstore-tool-samtools-index",
    "cmd": "samtools index /mnt/data/input.bam /mnt/data/output.bam.bai"
  }
}
```

Create, run, monitor, and kill pipelines

Create

```
$ gcloud alpha genomics pipelines create --pipeline-json-file PIPELINE-FILE.json --pipeline-json-file samtools_index.json
Created samtools index, id: PIPELINE-ID
```

Run

```
$ gcloud alpha genomics pipelines run --pipeline_id PIPELINE-ID \
--logging gs://YOUR-BUCKET/YOUR-DIRECTORY/logs \
--inputs inputFile=gs://genomics-public-data/gatk-examples/example1/NA12878_chr22.bam \
--outputs outputFile=gs://YOUR-BUCKET/YOUR-DIRECTORY/output/NA12878_chr22.bam.bai
Running: operations/OPERATION-ID
```

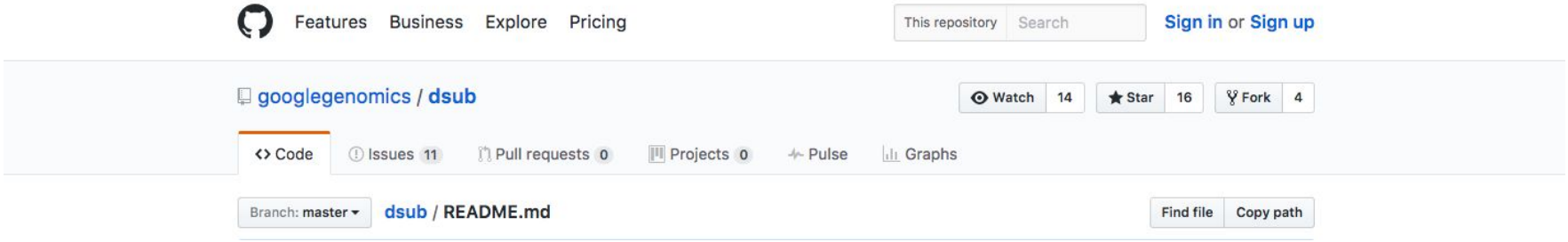
Status

```
$ gcloud alpha genomics operations describe OPERATION-ID
```

Kill

```
$ gcloud alpha genomics operations cancel OPERATION-ID
```

DSUB (google genomics pipelines)



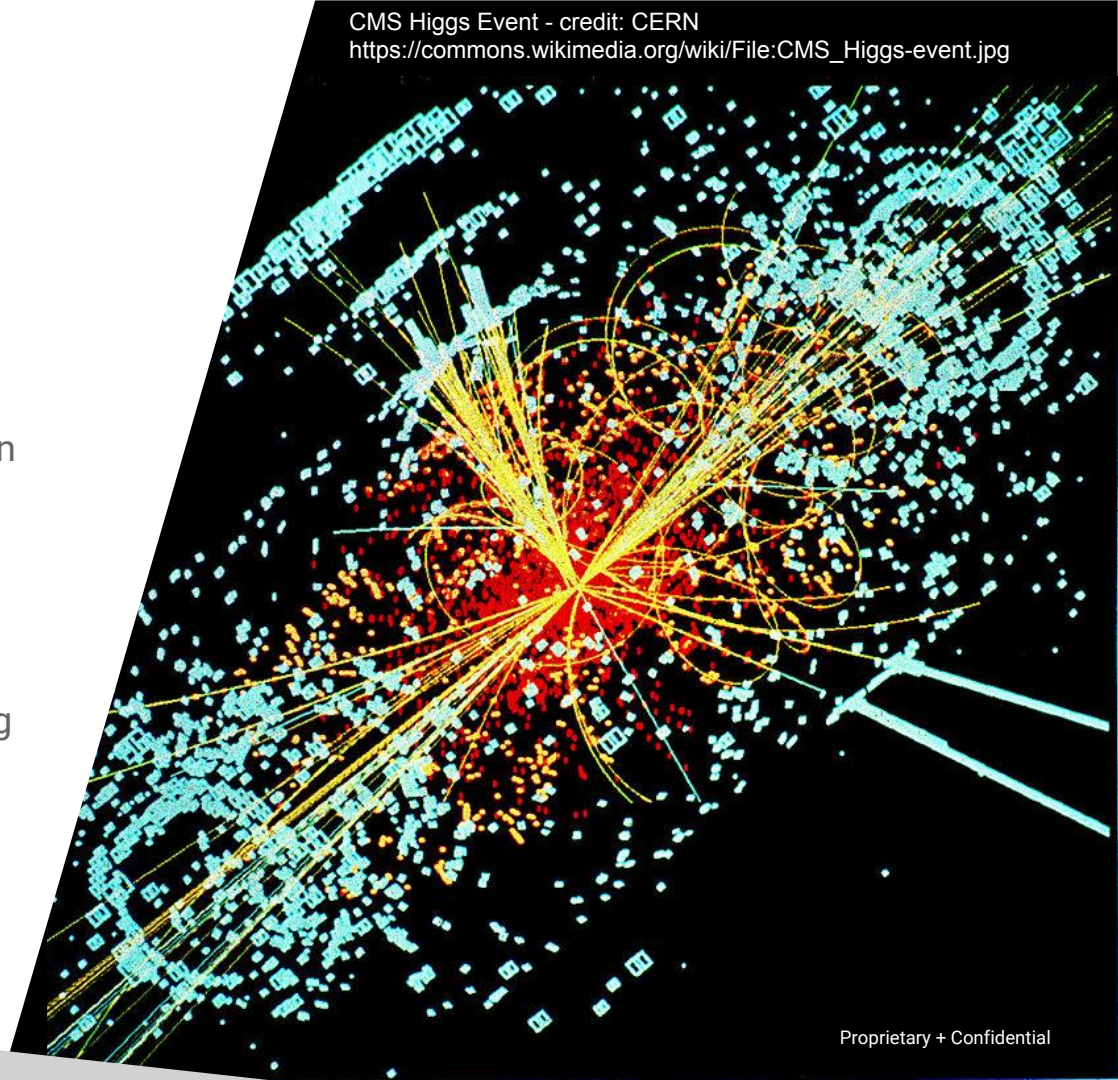

```
./dsub \
    --project my-cloud-project \
    --zones "us-*" \
    --logging gs://my-bucket/logs \
    --input BAM=gs://genomics-public-data/1000-genomes/bam/HG00114.mapped.ILLUMINA.bwa.GBR.low_covera
    --output BAI=gs://my-bucket/HG00114.mapped.ILLUMINA.bwa.GBR.low_coverage.20120522.bam.bai \
    --image quay.io/cancercollaboratory/dockstore-tool-samtools-index \
    --command 'samtools index ${BAM} ${BAI}' \
    --wait
```

SC16 CMS Demonstrator

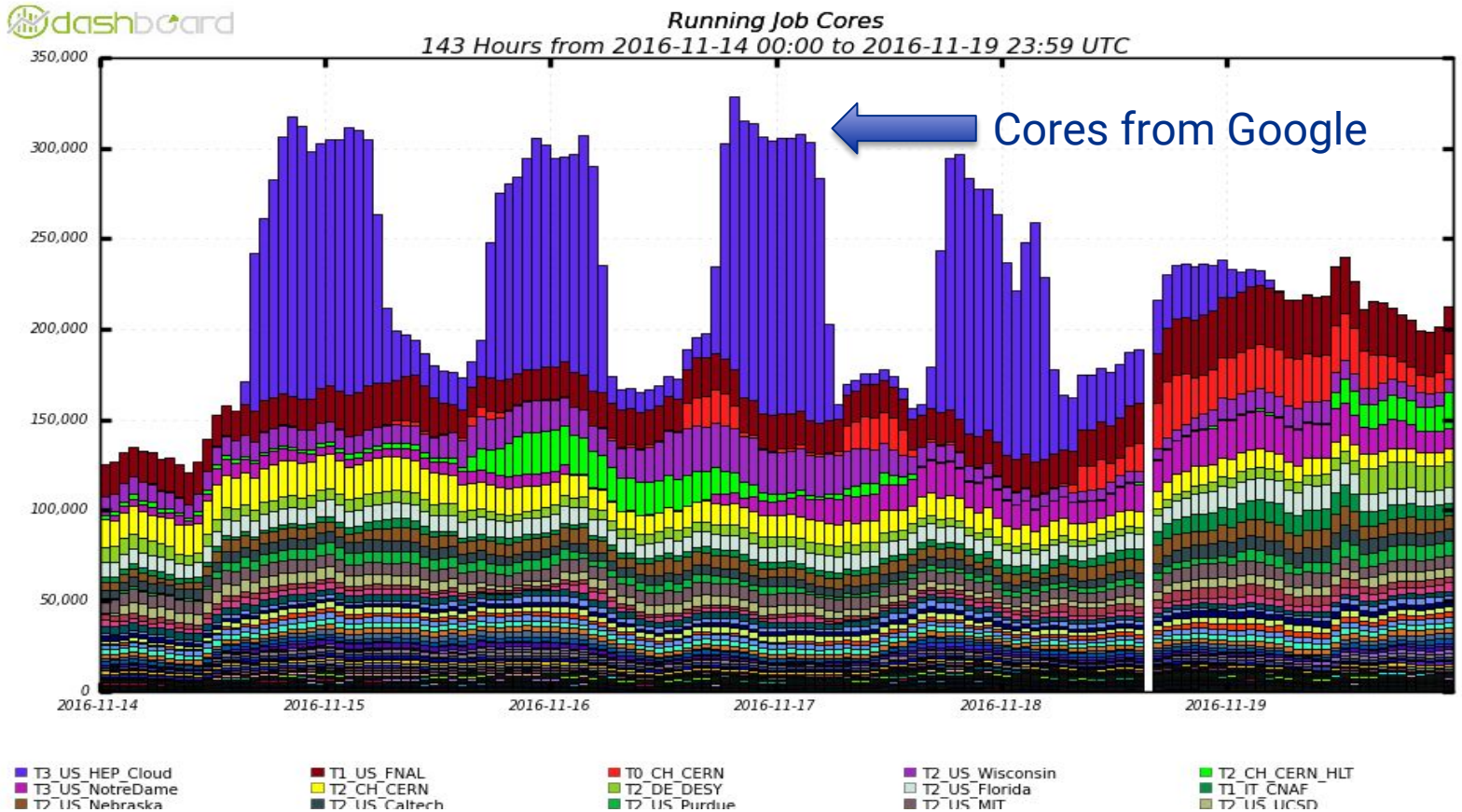
Target: generate 1 Billion events in 48 hours during Supercomputing 2016 on Google Cloud via HEPCloud

35% filter efficiency = stage out 380 million events → 150 TB output

Double the size of global CMS computing resources


CMS Higgs Event - credit: CERN
https://commons.wikimedia.org/wiki/File:CMS_Higgs-event.jpg

Proprietary + Confidential


dashboard
Running Job Cores
143 Hours from 2016-11-14 00:00 to 2016-11-19 23:59 UTC
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
2016-11-14
2016-11-15
2016-11-16
2016-11-17
2016-11-18
2016-11-19
Cores from Google
T3_US_HEP_Cloud
T3_US_NotreDame
T2_US_Nebraska
T1_US_FNAL
T2_CH_CERN
T2_US_Caltech
T0_CH_CERN
T2_DE_DESY
T2_US_Purdue
T2_US_Wisconsin
T2_US_Florida
T2_US_MIT
T2_CH_CERN_HLT
T1_IT_CNAF
T2_US_UCSD

On-prem vs. Cloud

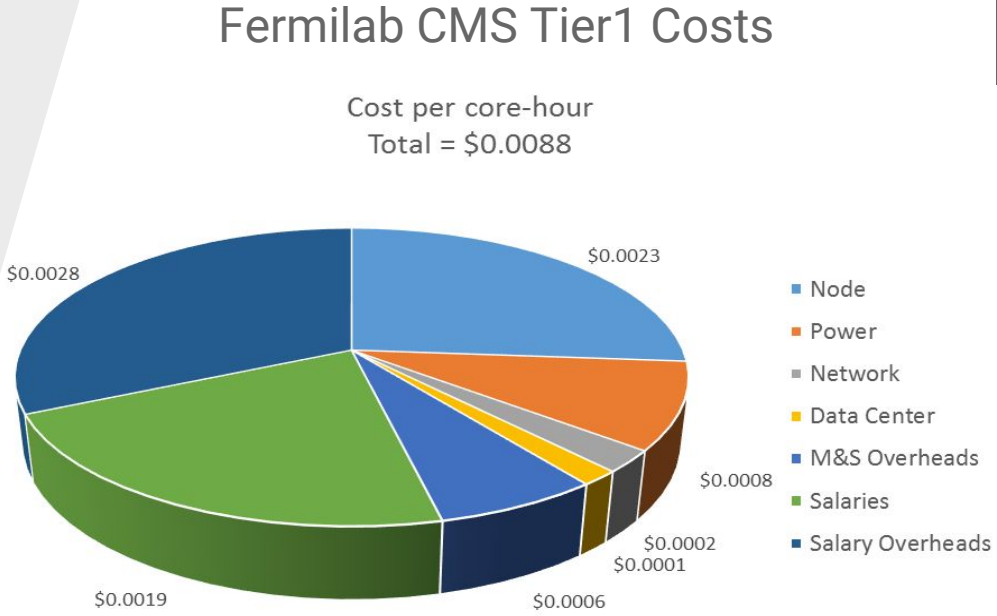
Average cost per core-hour (~25% error)

- On-premises Fermilab:
  0.9 cents per core-hour
  (assumes 100% utilization)
- Google Cloud:
  1.6 cents per core-hour
  (comparable to other vendors)

Fermilab has years of experience in optimizing its facility

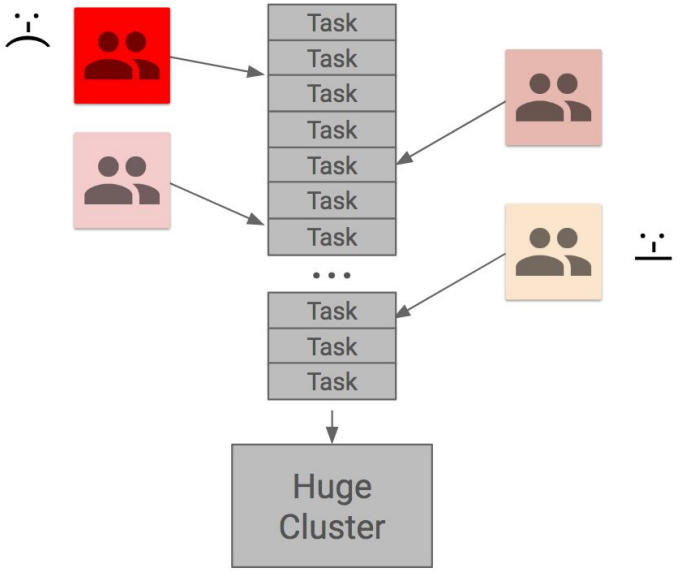
Cloud costs larger, but approaching equivalence

Considered well worth the cost of adding 160,000 core in a few hours

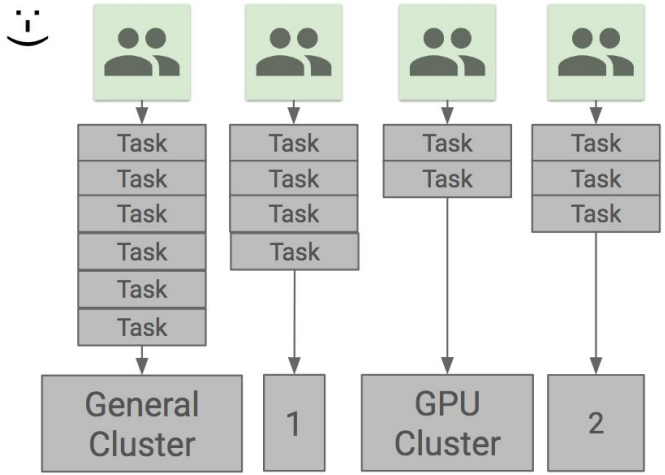


Task Tailored Resources

On-prem: One cluster, one queue, one size fits all hardware, angry users.

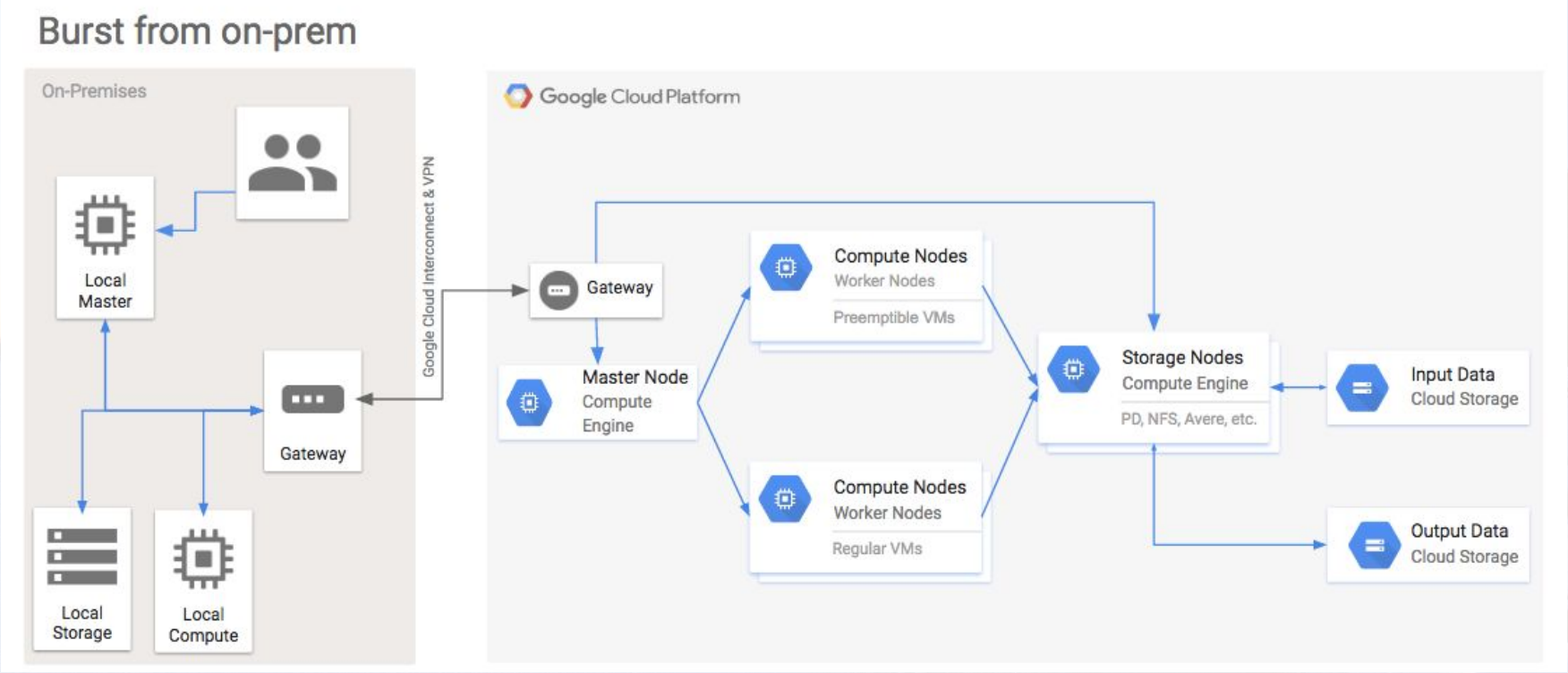


In the cloud: Tailored clusters, less queue sharing, happy users.



1 - n1-highcpu-16, Preemptible VM Cluster
2 - n1-highmem-32 Cluster

Basic Architecture: GCP as overflow capacity



Preemptible VM Instances

- **What Preemptible VMs are**
  - Up to 80% cheaper than regular VMs. (~$0.01 per core hour)
  - Very easy to use -- just flip one switch in the UI, API or command line
  - Many of our biggest customers run huge clusters (10k+ cores) with great success and savings.

- **Things to keep in mind**
  - Same great disk, OS images and network
  - Google Compute Engine can *preempt* (i.e. shutdown/take-away) the VM with 30 seconds of notice
  - Maximum 24 hours of uptime
  - No SLAs or guarantees of any kind but we historically see preemption rates of 5-15%

Fully Managed Storage & Database Services

Object

Cloud Storage

Binary or object data

Images, Media serving, backups

Key-value

App Engine Memcache

Web/mobile applications, gaming

Game state, user sessions

Non-relational

Cloud Datastore

Hierarchical, mobile, web

User profiles, Game State

Cloud Bigtable

Heavy read + write, events

AdTech, Financial, IoT

Relational

Cloud SQL

Web frameworks

CMS, eCommerce

Cloud Spanner

RDBMS+scale, HA, HTAP

Transactions, Ad/Fin/MarTech

Warehouse

BigQuery

Enterprise Data Warehouse

Analytics, Dashboards

Block storage

Reliable, high-performance block storage for any GCE VM instance

	Local SSD Fastest, Attached, Ephemeral	**Persistent Disk: SSD** Fast, Persistent, Durable, Remote	**Persistent Disk: HDD** Cheapest, Persistent, Durable, Remote
Target scenarios	- High-performance scratch space. Frequently accessed data. - Excellent for scientific workloads, especially when combined with fast compute VMs like GPU instances	- Latency sensitive applications and files. - High performance database and enterprise applications - Databases	- Large data processing workloads - Latency incentive tasks with lots of data: Genomics processing, video transcoding in GCE
Features	- Ephemeral storage - Highest-performance ($0.218 GB) - **IOPS:** 680k read / 360k write	- Persistent storage - Performance sensitive ($0.17GB) - **IOPS:** up to 40k read / 30k write	- Persistent storage - Cost sensitive ($.04 GB) - **IOPS:** 3k read / 15k write
	Encryption 3TB - 375 GB per partition, up to 8 partitions	Encryption, Snapshots 64 TB, Disk Size sets performance (Attach larger VMs for max SSD performance)	

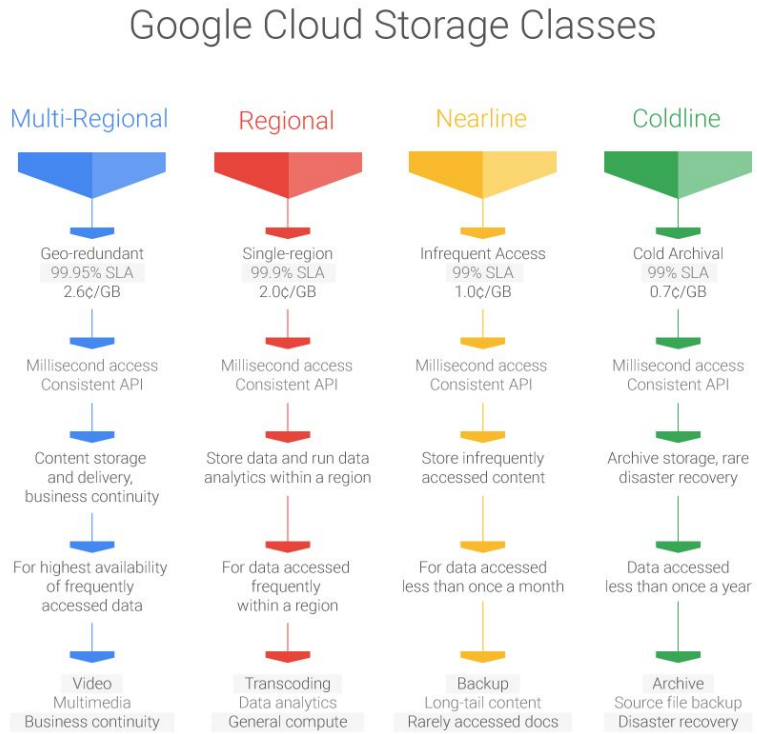
GCS: Object/Blob store

Google Cloud Storage is a scalable object storage service suitable for all kinds of unstructured data

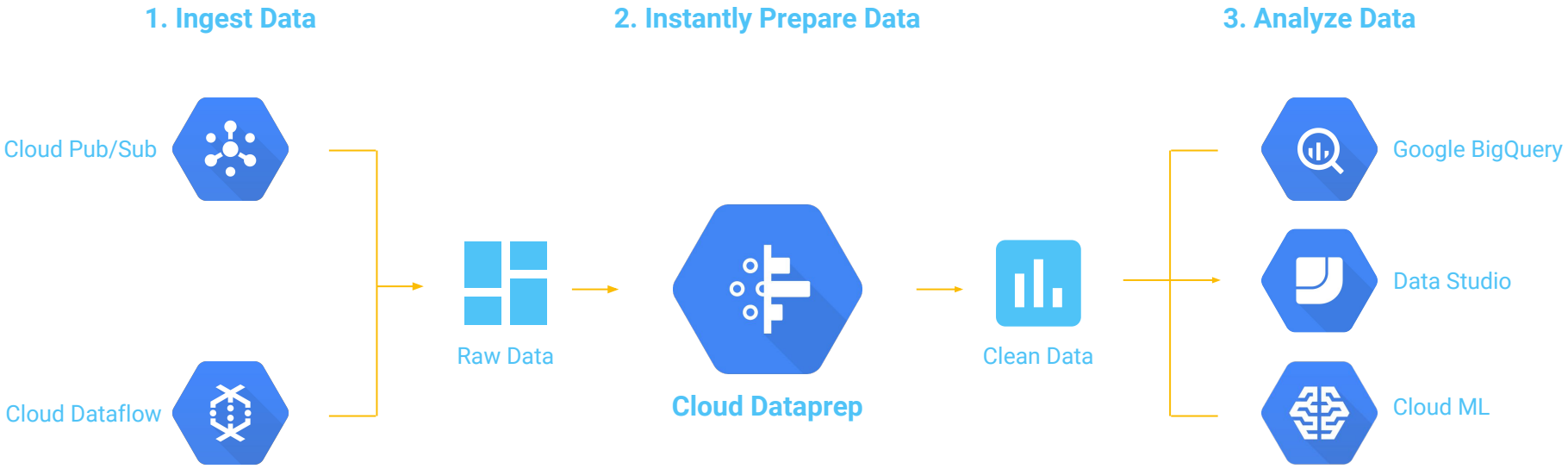
Cloud Storage vs Perst. Disk:

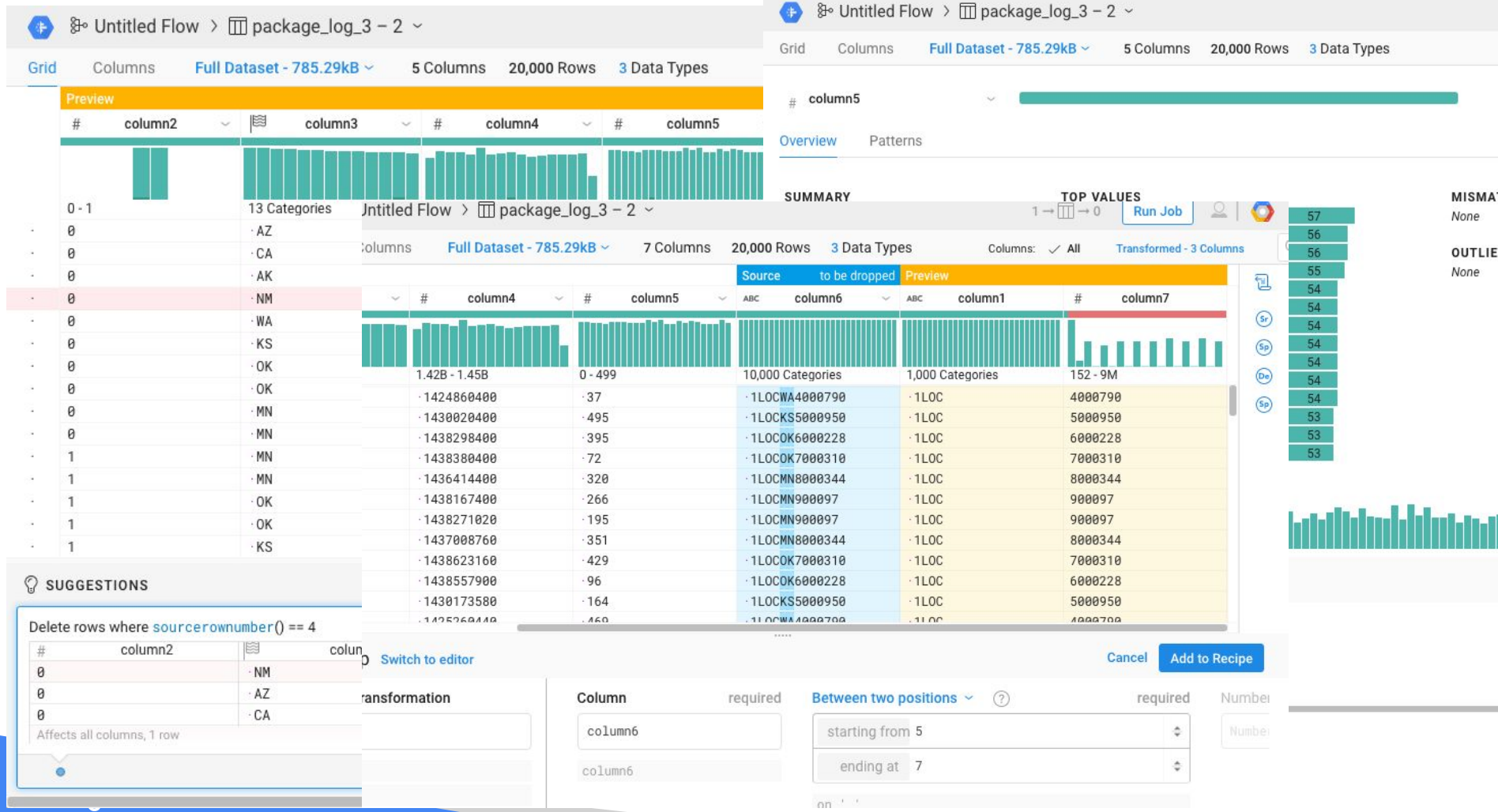
- Scales to exabytes
- Accessible from anywhere; REST interface
- Higher latency than PD
- Write semantics include insert and overwrite file only
- Offers versioning
- Cheaper - put your data here until you need it

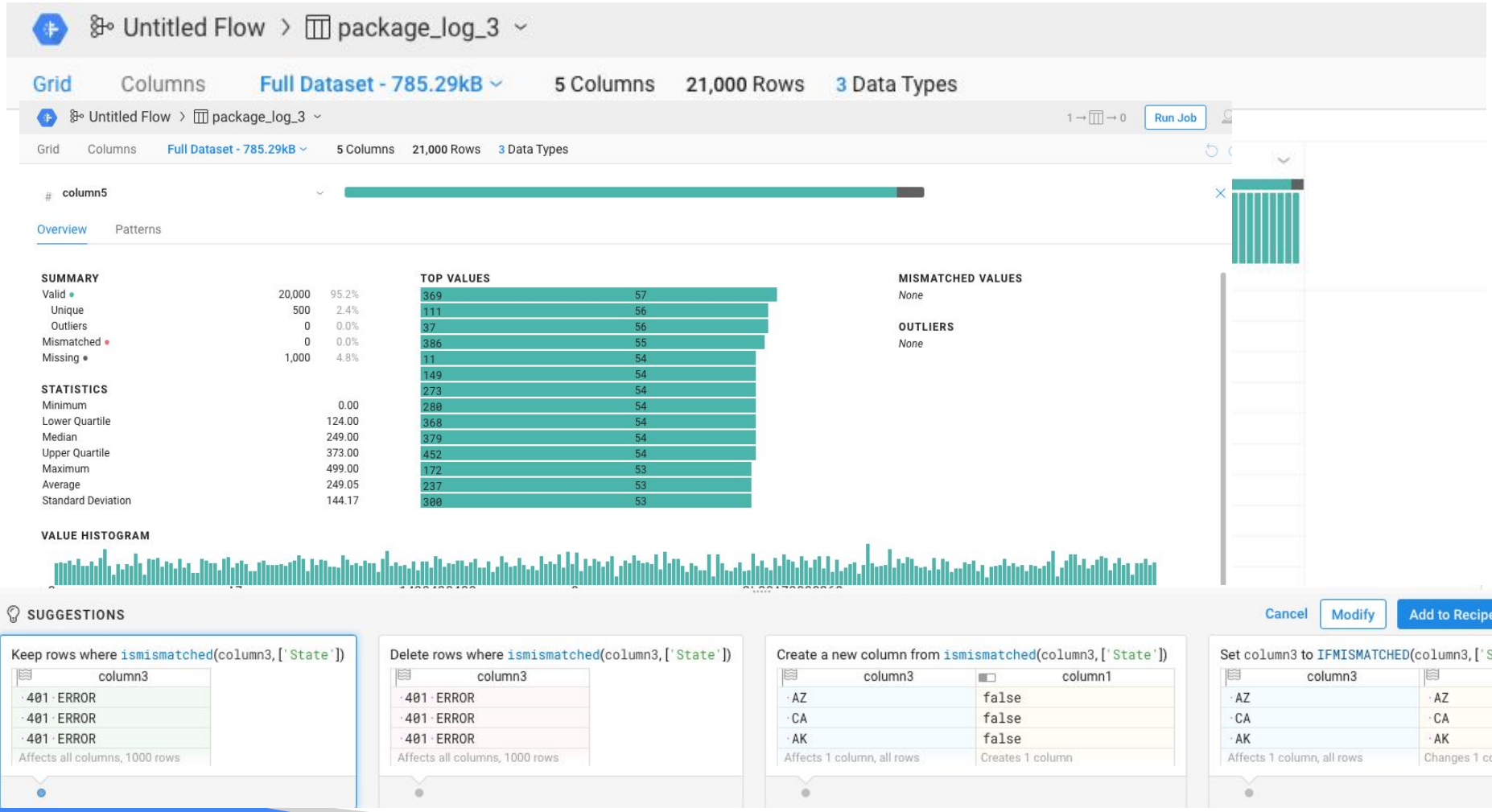
Lots of guidelines on picking storage on our site



Data Prep

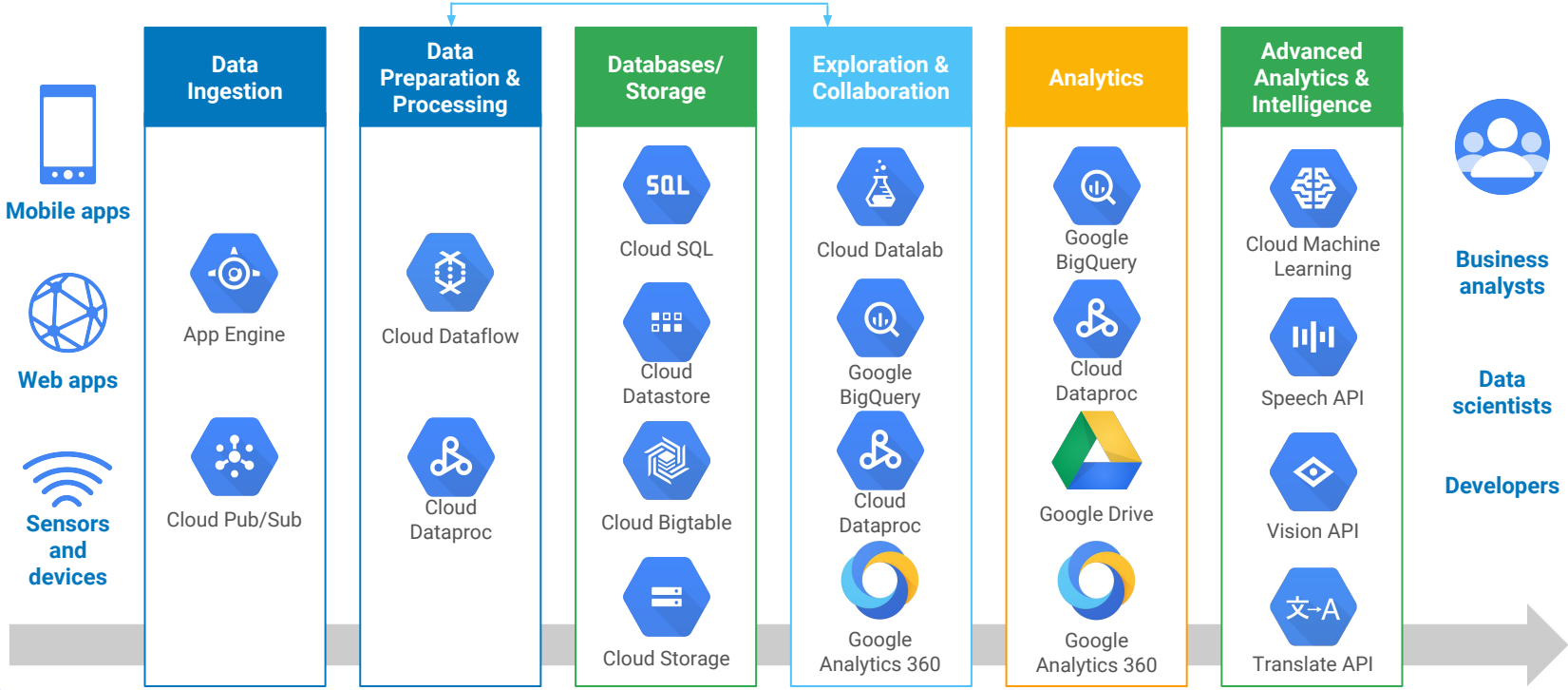

1. Ingest Data
2. Instantly Prepare Data
3. Analyze Data
Cloud Pub/Sub
Cloud Dataflow
Raw Data
Cloud Dataprep
Clean Data
Google BigQuery
Data Studio
Cloud ML


Untitled Flow › package_log_3 – 2
Grid Columns Full Dataset - 785.29kB 5 Columns 20,000 Rows 3 Data Types
Preview
column2 column3 column4 column5
0 - 1 13 Categories
0 AZ
0 CA
0 AK
0 NM
0 WA
0 KS
0 OK
0 OK
0 MN
0 MN
1 MN
1 MN
1 OK
1 OK
1 KS
SUGGESTIONS
Delete rows where sourcerownumber() == 4
column2
0 NM
0 AZ
0 CA
Affects all columns, 1 row
Untitled Flow › package_log_3 – 2
Grid Columns Full Dataset - 785.29kB 5 Columns 20,000 Rows 3 Data Types
column5
Overview Patterns
SUMMARY
TOP VALUES
Run Job
Columns Full Dataset - 785.29kB 7 Columns 20,000 Rows 3 Data Types
Columns: All Transformed - 3 Columns
Source to be dropped Preview
column4 column5 column6 column1 column7
1.42B - 1.45B 0 - 499 10,000 Categories 1,000 Categories 152 - 9M
1424860400 37 1LOCWA4000790 1LOC 4000790
1430020400 495 1LOCKS5000950 1LOC 5000950
1438298400 395 1LOCOK6000228 1LOC 6000228
1438380400 72 1LOCOK7000310 1LOC 7000310
1436414400 320 1LOCMN8000344 1LOC 8000344
1438167400 266 1LOCMN900097 1LOC 900097
1438271020 195 1LOCMN900097 1LOC 900097
1437008760 351 1LOCMN8000344 1LOC 8000344
1438623160 429 1LOCOK7000310 1LOC 7000310
1438557900 96 1LOCOK6000228 1LOC 6000228
1430173580 164 1LOCKS5000950 1LOC 5000950
57 56 56 55 54 54 54 54 54 54 54 53 53 53
MISMA
None
OUTLIE
None
Switch to editor
Cancel
Add to Recipe
ransformation
Column required
column6
Between two positions
required
starting from 5
ending at 7



Proprietary + Confidential

Transform data into actions



What makes Google Cloud Platform Different?

Global Network & Security

Compute & Data Analytics

Dev Tools and Open Source Ecosystem

Use Cases & EDU Programs

Google Cloud

Kubernetes

Greek for *“Helmsman”*; also the root of the words *“governor”* and *“cybernetic”*

- Manages container clusters
- Inspired and informed by Google’s experiences and internal systems
- Supports multiple cloud and bare-metal environments
- Supports multiple container runtimes
- **100% Open source**, written in Go

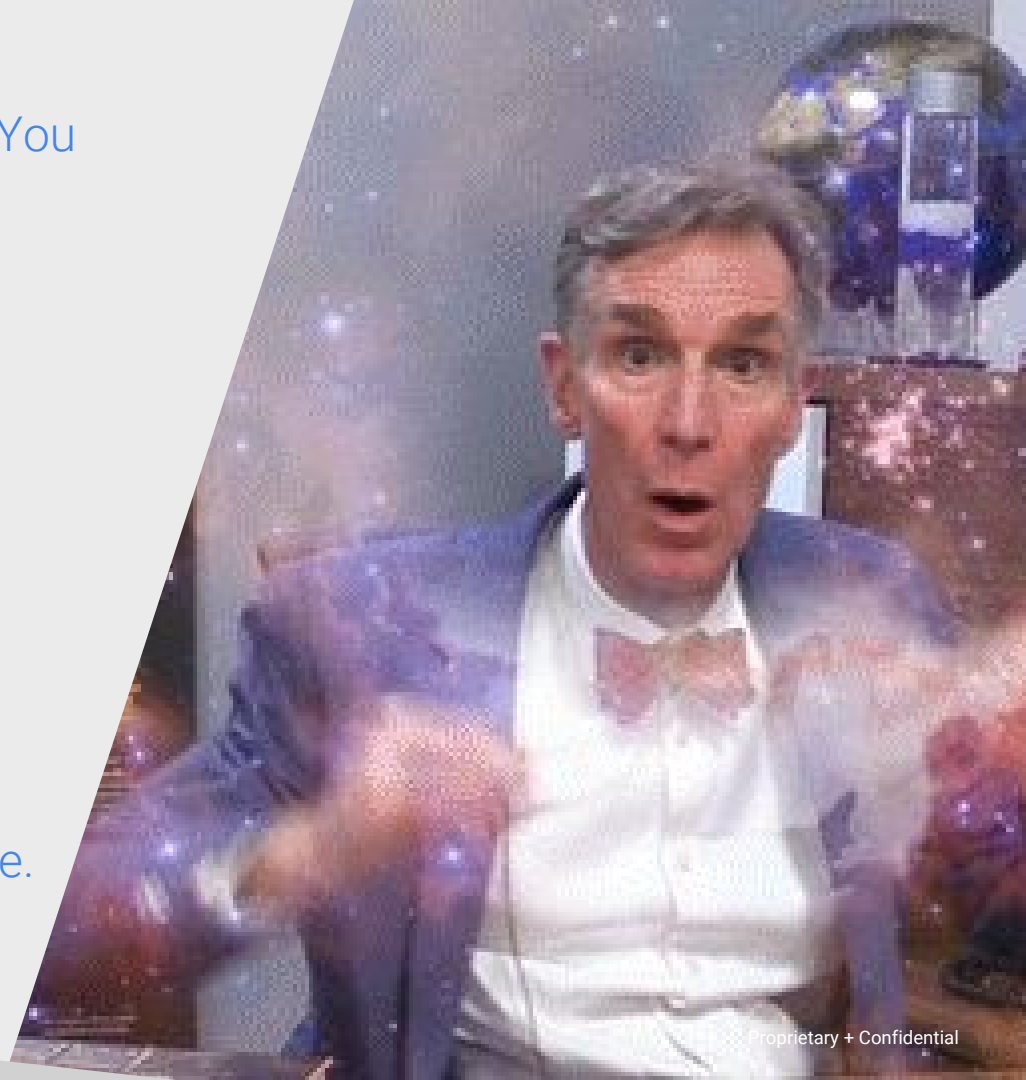
Manage applications, not machines

Question: Why does the Ship’s wheel on the Kubernetes Logo have 7 Handles?

Proprietary + Confidential

If you remember Star Trek Voyager. You may recall a Borg Character named “Seven of Nine”.

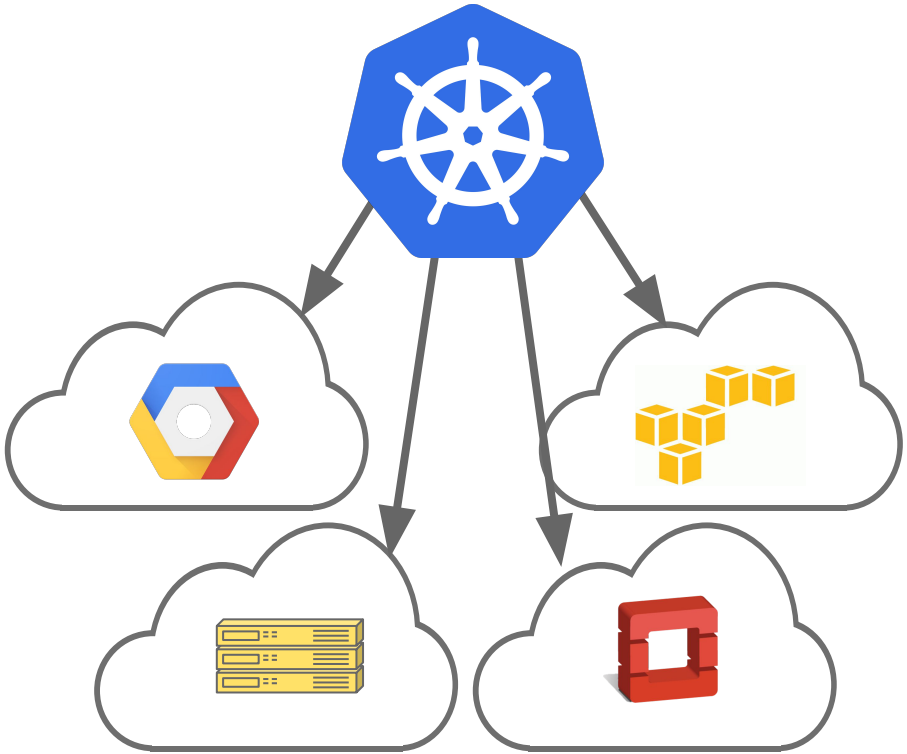
The seven handles are a direct reference to Kubernetes’ Borg lineage. Seven was a more friendly Borg.



Proprietary + Confidential

Kubernetes

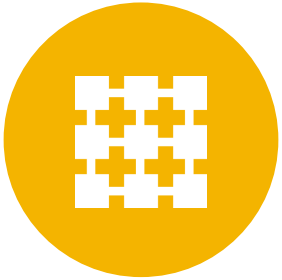
Runs in many environments, including “bare metal” and “your laptop”
The API and the implementation are 100% open
The whole system is modular and replaceable
Build your apps on-prem, lift-and-shift into cloud when you are ready



Container Engine: Managed Kubernetes

Deployment-Ready Cluster

Spin up a managed container cluster in minutes, run Docker containers.



Declarative Management

Declare your container requirements and Container Engine handles the rest

Cloud Flexibility

Built for a multi-cloud world, public, private, or hybrid

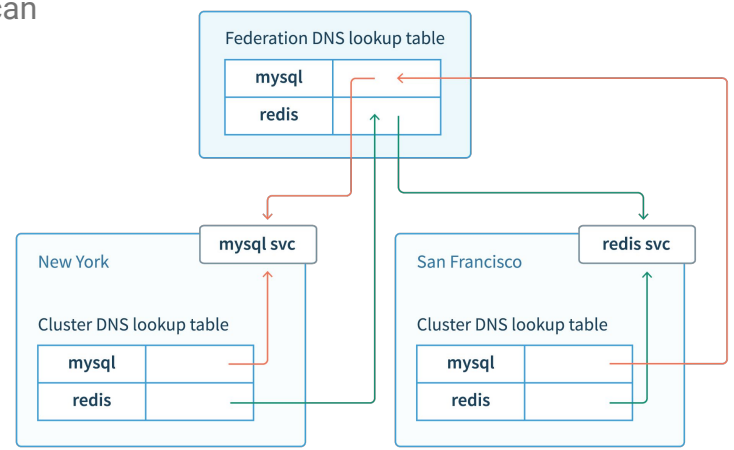
Cluster Federation

Federation now lets you create **services**, **namespaces**, **Replica Sets** and **secrets** across multiple clusters (multi-region, multi-cloud or hybrid), via a single Federated API.

- Deploy a federated control plane to your cluster, add clusters (running on GKE, GCP or elsewhere) to your federation, and you can create services that run across them with one command.
- GKE clusters can use a federated Ingress to expose their service behind a global load balancer with a single IP address.
- Further objects are on the roadmap

More info:

- GKE documentation (1.3)
- Kubernetes documentation
- Blog post (1.3)
- Sneak peek video (1.3)



* Also available in Kubernetes

What makes Google Cloud Platform Different?

Global Network & Security

Compute & Data Analytics

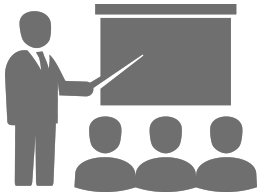
Dev Tools and Open Source Ecosystem

Use Cases & EDU Programs

Google Cloud

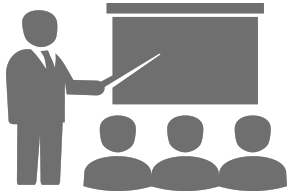
Administration

Research



Teaching & Learning

Google Cloud Platform



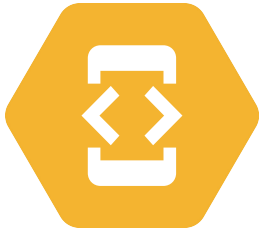
Teaching & Learning

Faculty

+

Teaching university courses

+



In computer science or related fields

Apply for our Education Grants Program: cloud.google.com/edu

Google Cloud Platform Fundamentals Training

Interested? salmeron@google.com

Mountain View, CA	1/12/2018	Google Cloud Platform Fundamentals: Core Infrastructure
Los Angeles, CA	1/22/2018	Google Cloud Platform Fundamentals: Core Infrastructure
Chicago, IL	1/26/2018	Google Cloud Platform Fundamentals: Core Infrastructure
Washington, DC	1/29/2018	Google Cloud Platform Fundamentals: Core Infrastructure
Seattle, WA	1/31/2018	Google Cloud Platform Fundamentals: Core Infrastructure
Atlanta, GA	2/5/2018	Google Cloud Platform Fundamentals: Core Infrastructure
Austin, TX	2/5/2018	Google Cloud Platform Fundamentals: Core Infrastructure
San Francisco, CA	2/12/2018	Google Cloud Platform Fundamentals: Core Infrastructure
Cambridge, MA	2/12/2018	Google Cloud Platform Fundamentals: Core Infrastructure
San Francisco, CA	2/16/2018	Google Cloud Platform Fundamentals: Big Data & Machine Learning
Chicago, IL	2/22/2018	Google Cloud Platform Fundamentals: Big Data & Machine Learning
Chicago, IL	2/23/2018	Google Cloud Platform Fundamentals: Core Infrastructure
Washington, DC	2/26/2018	Google Cloud Platform Fundamentals: Core Infrastructure
Mountain View, CA	3/12/2018	Google Cloud Platform Fundamentals: Core Infrastructure
New York, NY	3/19/2018	Google Cloud Platform Fundamentals: Core Infrastructure
Chicago, IL	3/23/2018	Google Cloud Platform Fundamentals: Core Infrastructure
Washington, DC	3/26/2018	Google Cloud Platform Fundamentals: Core Infrastructure

Questions?

Contact: salmeron@google.com

Google Cloud