Welcome to Day 1!

About MatrixC

matrixc

Partner
Google Cloud

Google Cloud
Partner Since 2008

Work from anywhere, on any device

G Suite

- Unified Communication
- Store and Share
- Productivity Apps
- Collaboration & Social
- Manage & Archiving
- Search & Discovery

Google Cloud Platform

- Auto-scale Applications
- Virtual Machines
- Big Data Analytics
- Cloud Storage
- Managed Database
- Machine Learning
- Google APIs

3rd Party Solution Partners

acer Chrome for Work
Symple
Barracuda
tableau
CloudEndure™

and many more ...

Infrastructure

Google Data Centers

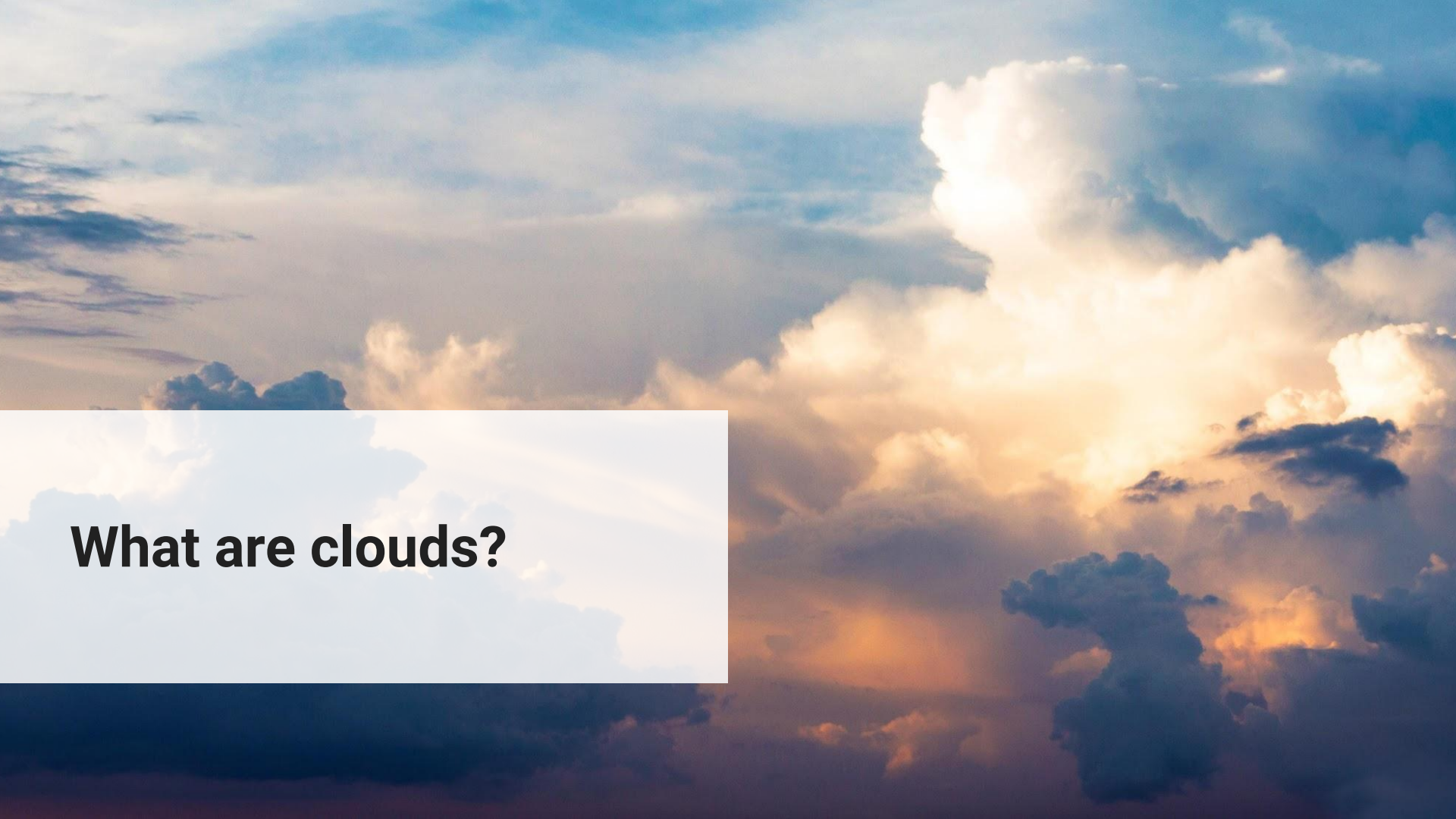
- Globally Aware Data Centers
- Dedicated worldwide IP network
- 99.9% Uptime SLA
- Audits & Certifications
- Scale & Security
- Energy Efficient
- Disaster Recovery
- Redundancy

Wilson Chuah
Cloud Architect and Data Engineer for Matrix Connexion

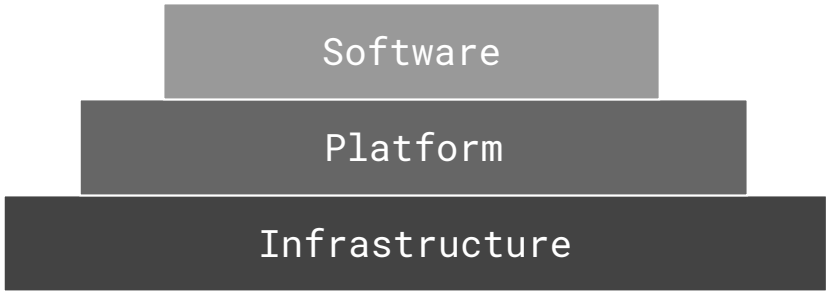
wilson@matrixc.com

Topics That We'll Be Covering

- Introduction to Cloud Concepts
- Introduction to Google Cloud
- Google Cloud SDK & Cloud Shell
- Google Cloud Identity and Access Management
- Google Compute Engine & Networking
- Google Cloud Functions [Beta]
- Google Kubernetes Engine
- Google App Engine
- Google Cloud Storage
- Google Pub/Sub
- Google Dataflow
- Google BigQuery
- Google DataStudio



What are clouds?

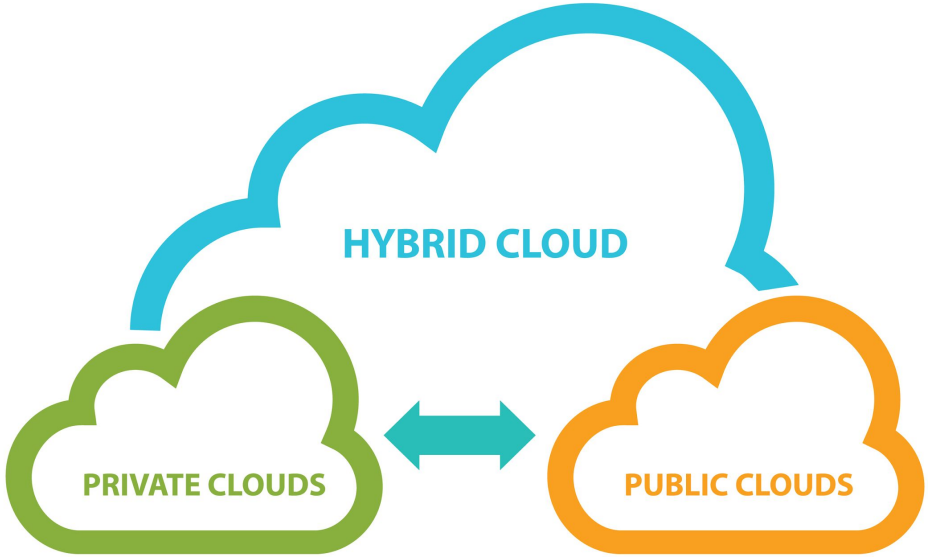


Cloud Lingo

- IaaS - Infrastructure as a Service
- PaaS - Platform as a Service
- SaaS - Software as a Service

matrixc

3 types of clouds

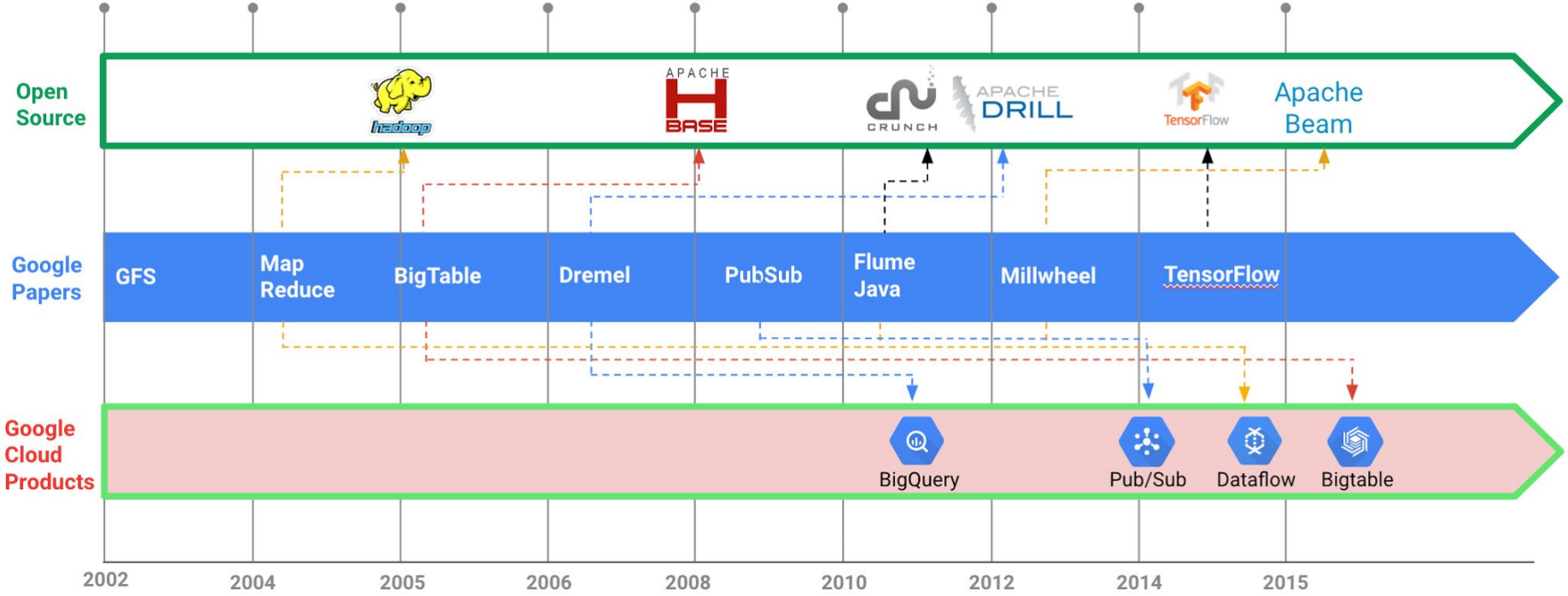


Google Cloud

Google

7 Cloud products with 1 billion users

Google Cloud Platform is not an afterthought products



Purpose-built hardware infrastructure

Provenance from the bottom of the stack to the top

Purpose-built chips

Purpose-built servers

Purpose-built storage

Purpose-built network

Purpose-built data centers

Reduced “vendor in the middle” risk

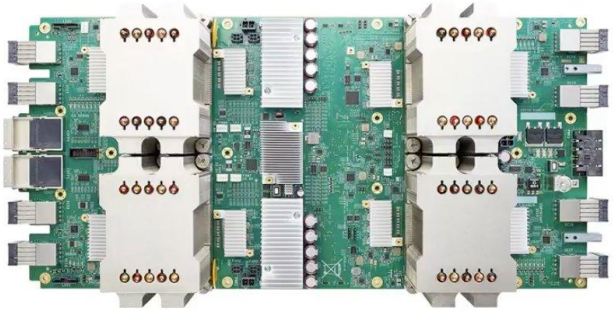
Google Cloud

Google.ai - Machine Learning

- Specialized Neural Network powered by Unique Tensor Processing Unit (TPU)
- Custom ASIC built and optimized for **TensorFlow** (open-source software library for Machine Intelligence)
- Used in production at Google for over 17+ months

TensorFlow

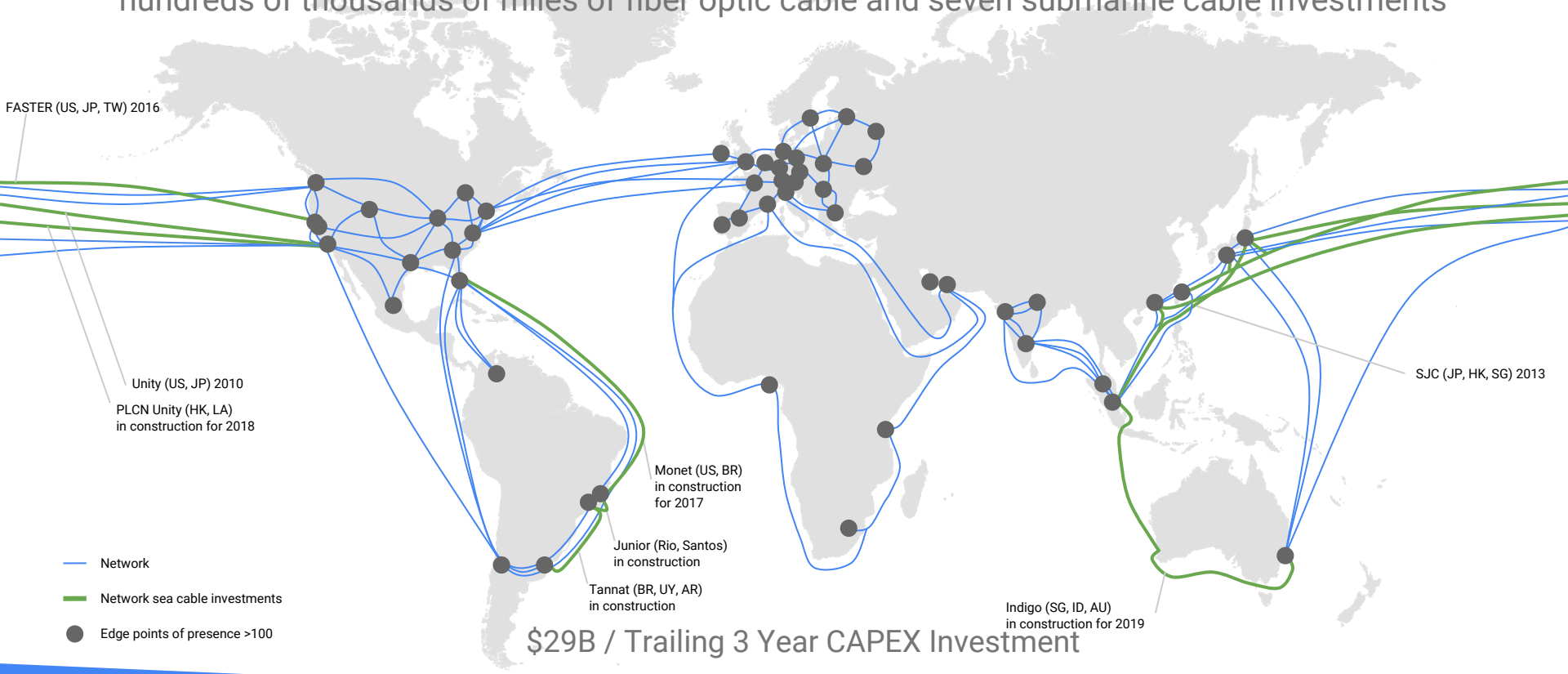
Cloud TPU



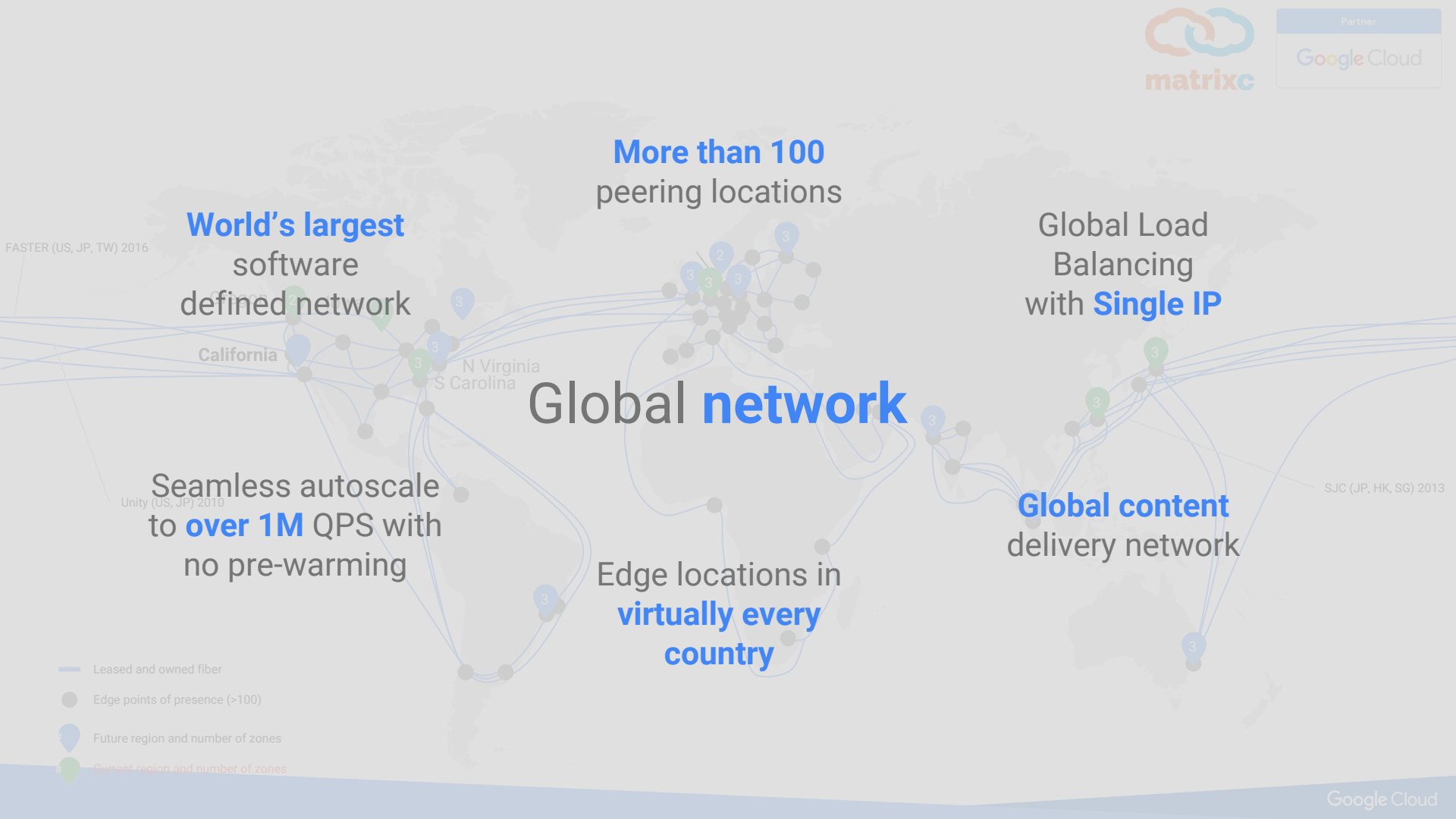

Per-second billing =
Truly Elastic Costs
Live Migration =
Less Downtime
Custom Machine Types =
No over-provisioning
Resize disks on the fly =
Easier operations
Move and Improve
Sustained Use Discounts =
No upfront payments
Automatic app-specific sizing
recommendations
High throughput storage at
no extra cost
Super-fast startup: VMs boot
in seconds, not minutes

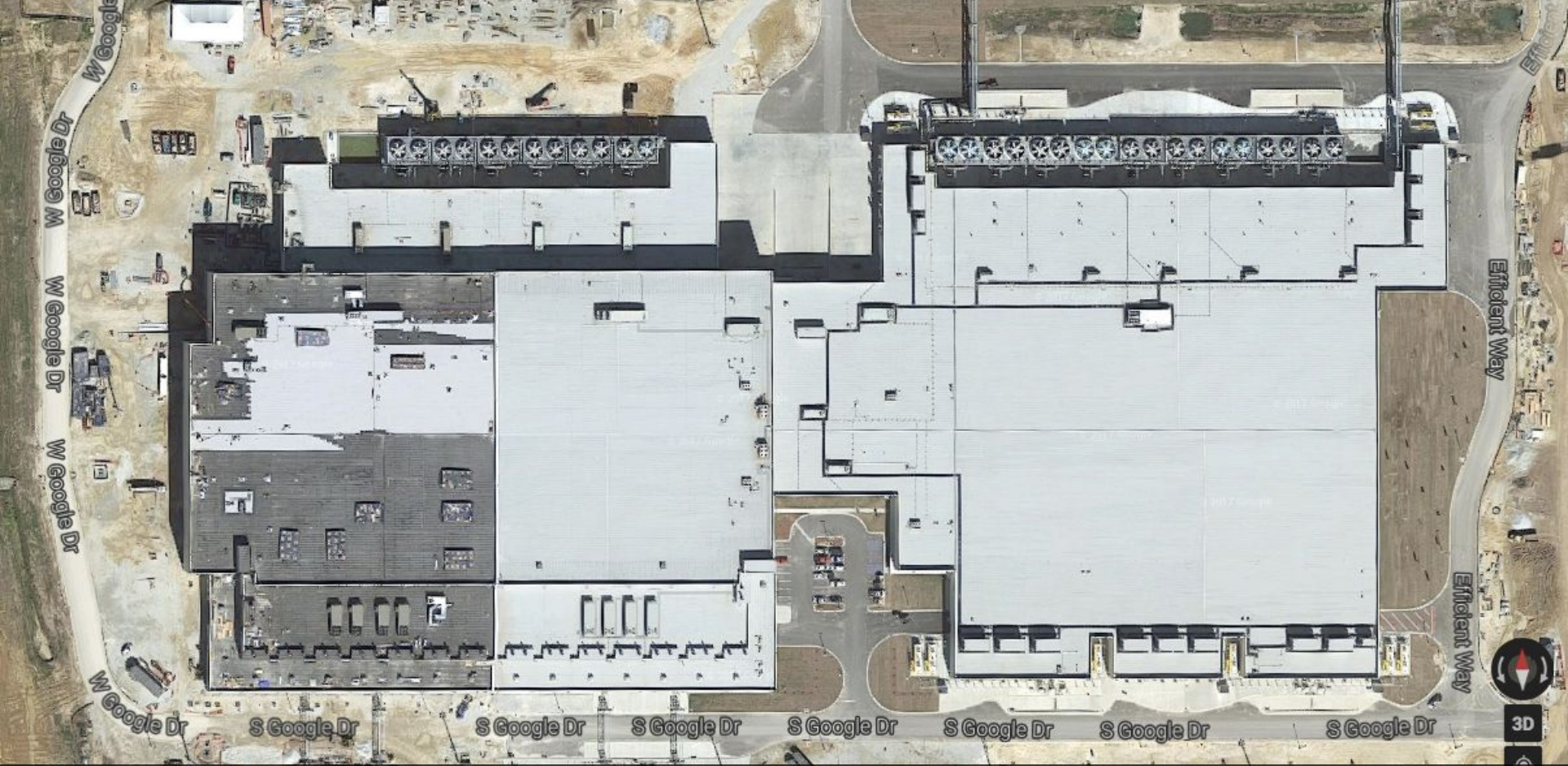
Google Cloud Network

Google Cloud's well-provisioned global network is comprised of hundreds of thousands of miles of fiber optic cable and seven submarine cable investments

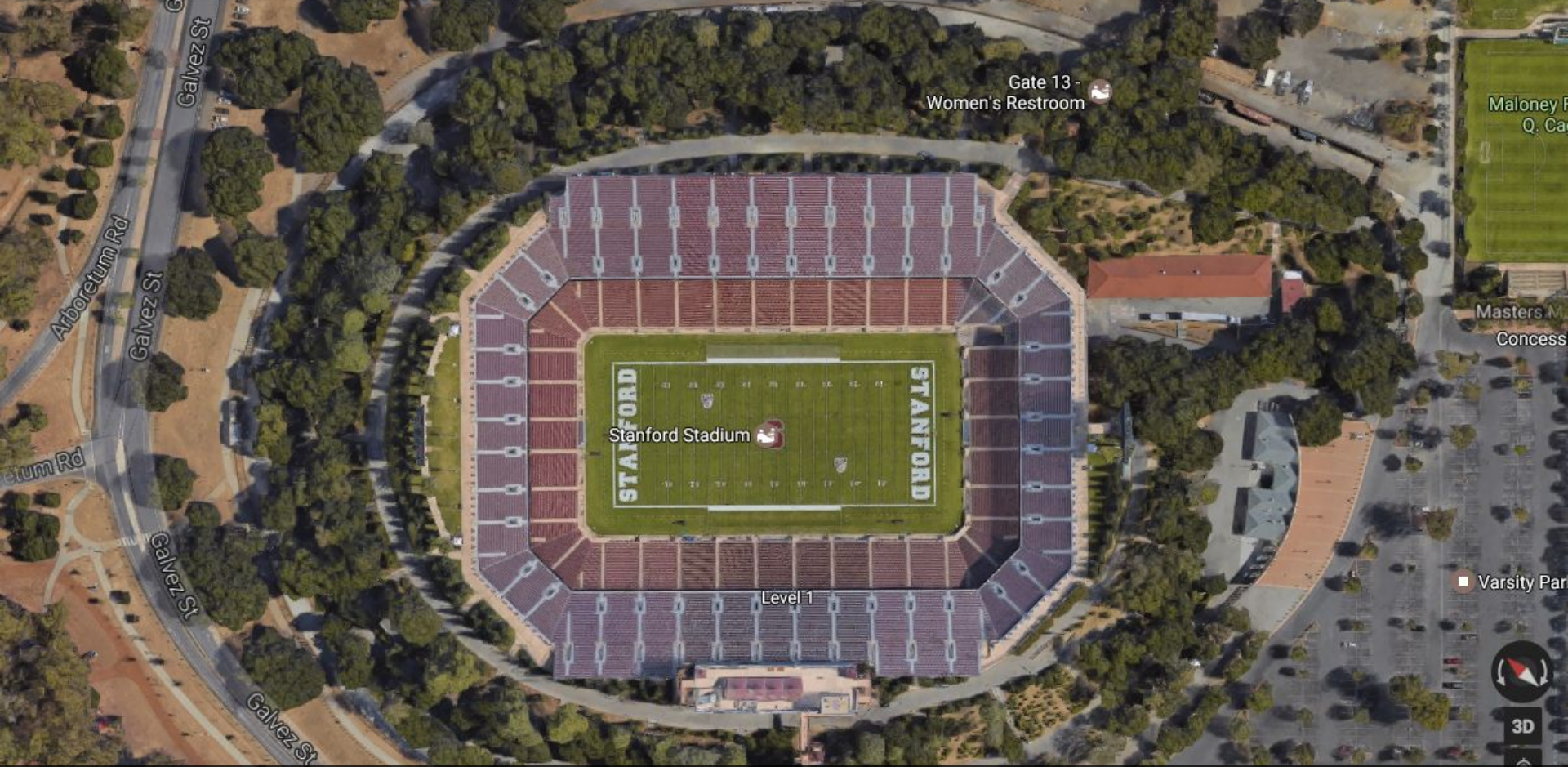


$29B / Trailing 3 Year CAPEX Investment


matrixc
Partner
Google Cloud
More than 100
peering locations
World's largest
software
defined network
Global Load
Balancing
with Single IP
FASTER (US, JP, TW) 2016
California
N Virginia
S Carolina
Global network
Seamless autoscale
to over 1M QPS with
no pre-warming
Unity (US, JP) 2010
Global content
delivery network
SJC (JP, HK, SG) 2013
Edge locations in
virtually every
country
Leased and owned fiber
Edge points of presence (>100)
Future region and number of zones
Current region and number of zones
Google Cloud



One building in our South Carolina campus...



...is larger than Stanford University Stadium

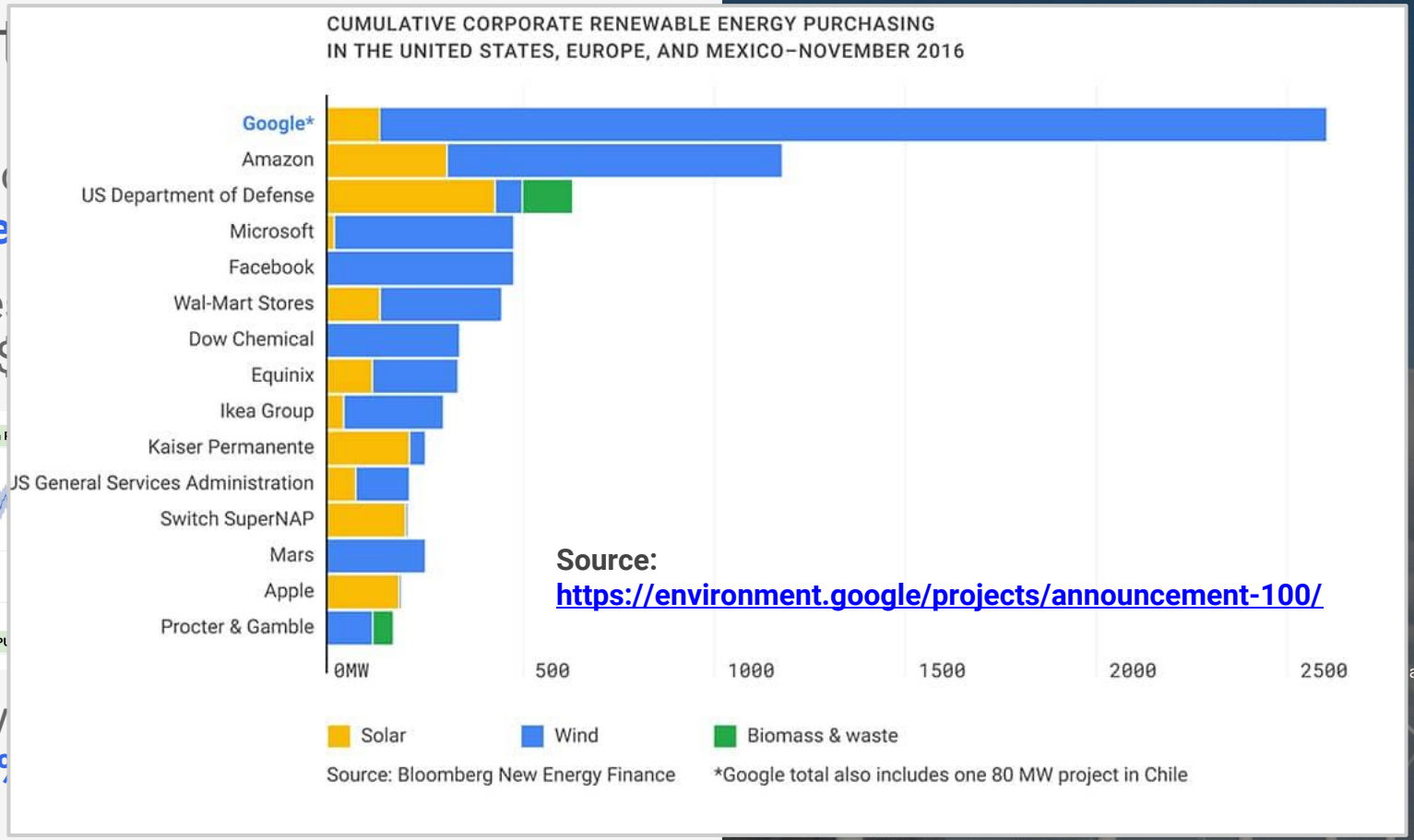
Netherlands complex has 10,000 miles of cables

Sust

Goo
overhe

Large
s

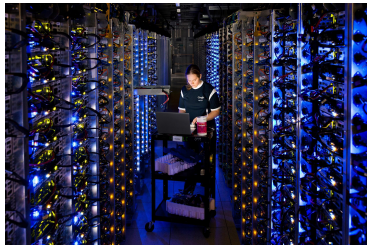
Apply
40%



Source:
https://environment.google/projects/announcement-100/

all the processing

Google Security approach

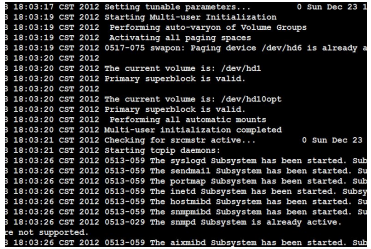


Defense in depth by default at scale

Hardshell perimeter model is insufficient

Trust through transparency

Trust not just with technology, but through transparency



Abstraction and automation

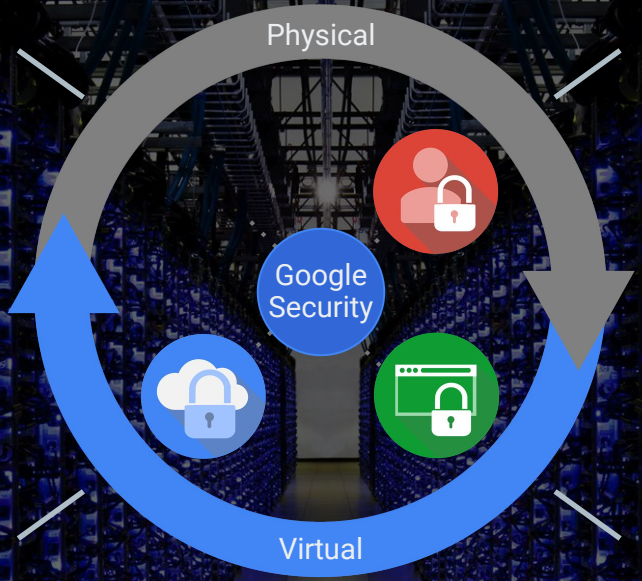
Automate best practices and prevent common mistakes at scale

Innovation

Focus on your business and we take care of the rest

Security and data protection is among our primary design criteria

Multi-layered, advanced technology, laser beam intrusion detection, biometrics, video analytics



First major cloud provider to enable HTTPS/TLS by default

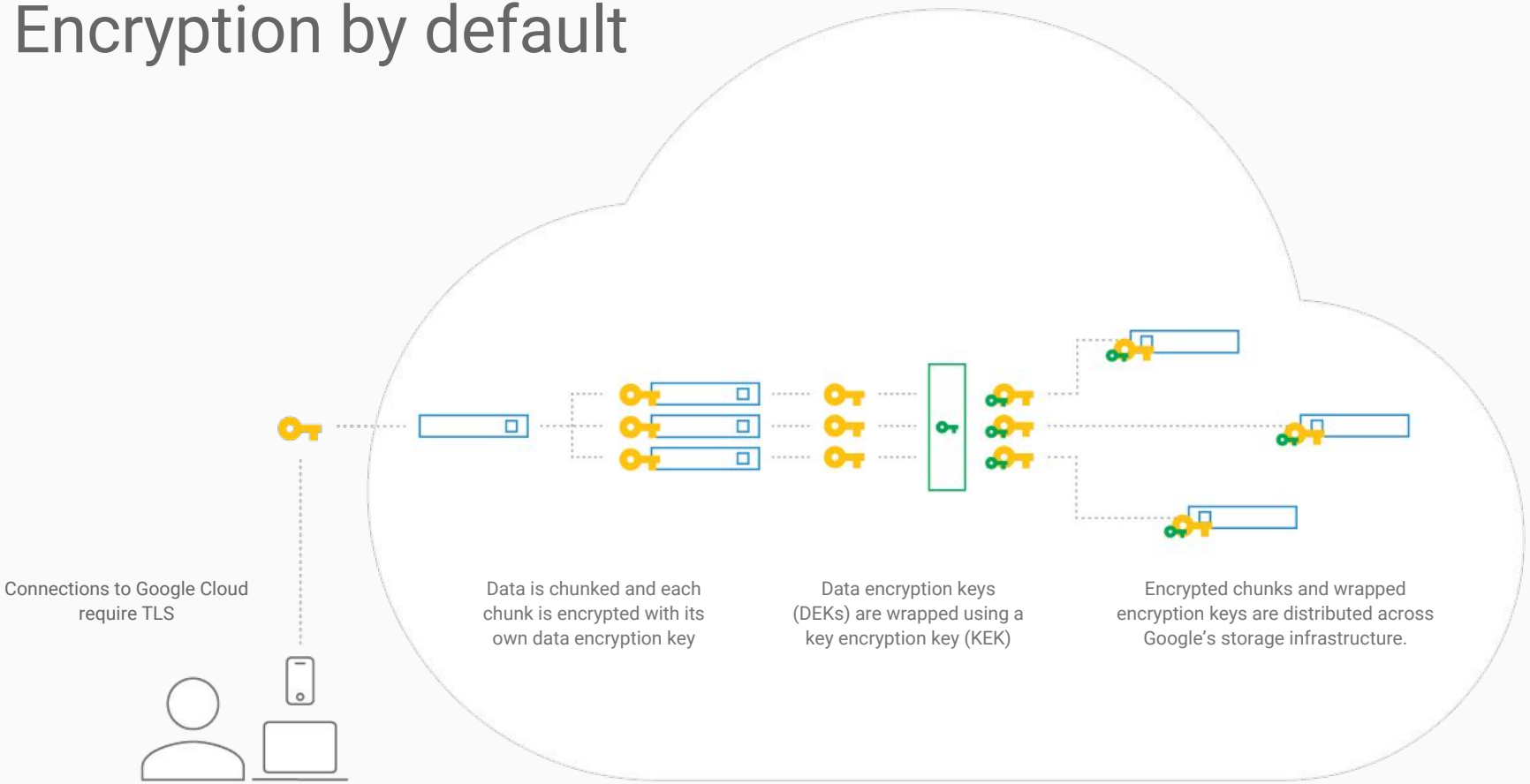
Encryption between customer device and Google

500+ Security Experts

First to start reward program

Backed by industry certifications: ISO, PCI, ...

Encryption by default



Compliance audits

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
For G Suite and Google App Engine

Privacy Shield Framework

Google Security for G Cloud

- Google - Our Secure Data Centers
  - https://www.youtube.com/watch?v=7pkNt3szF1A
- Google Security Whitepaper
  - https://cloud.google.com/security/whitepaper
- Google Infrastructure Security Design Overview
  - Google Cloud Whitepaper
  - https://cloud.google.com/security/security-design/resources/google_infrastructure_whitepaper_fa.pdf

Google Security for G Suite

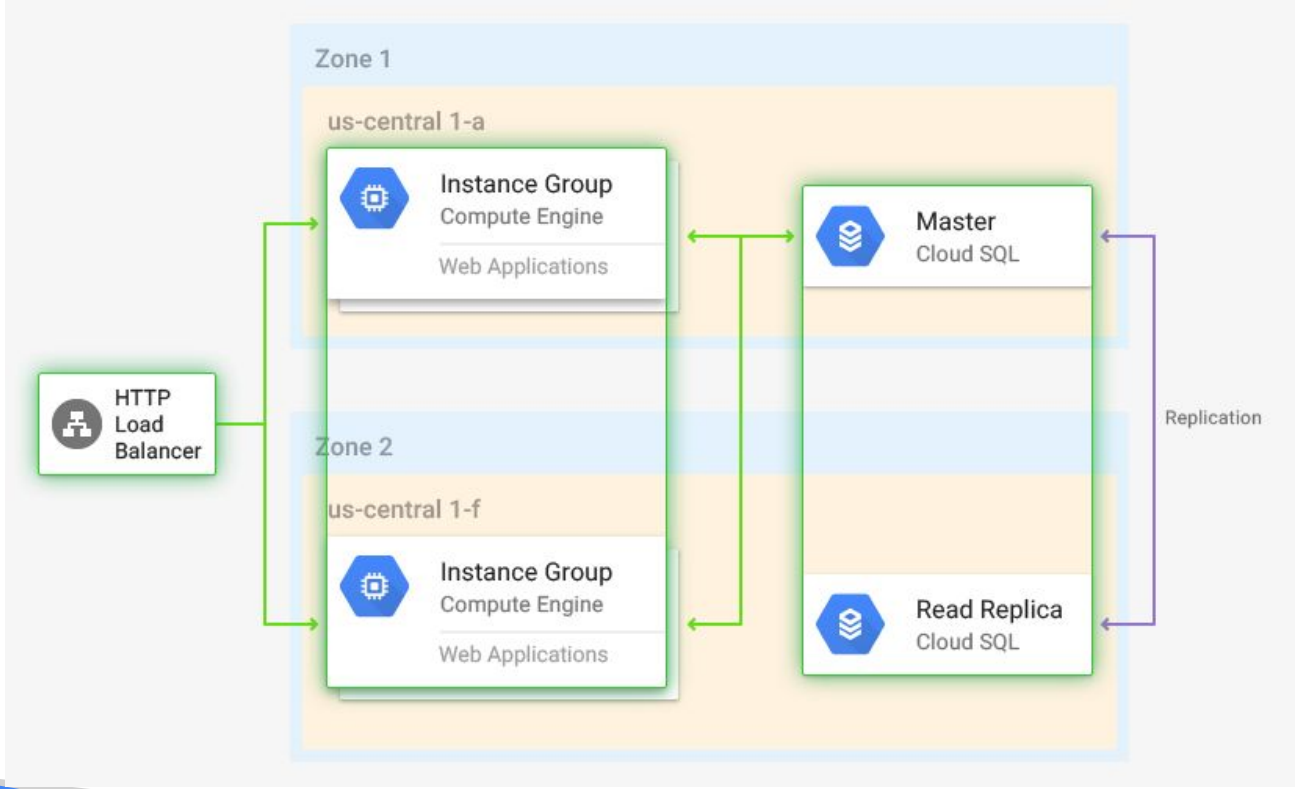
- Inside a Google Data Center:
  - https://www.youtube.com/watch?v=XZmGGAbHqa0

- G Suite Security & Privacy
  - Google is committed to the security and privacy of your organization's data
  - https://gsuite.google.com/learn-more/security-google-apps.html

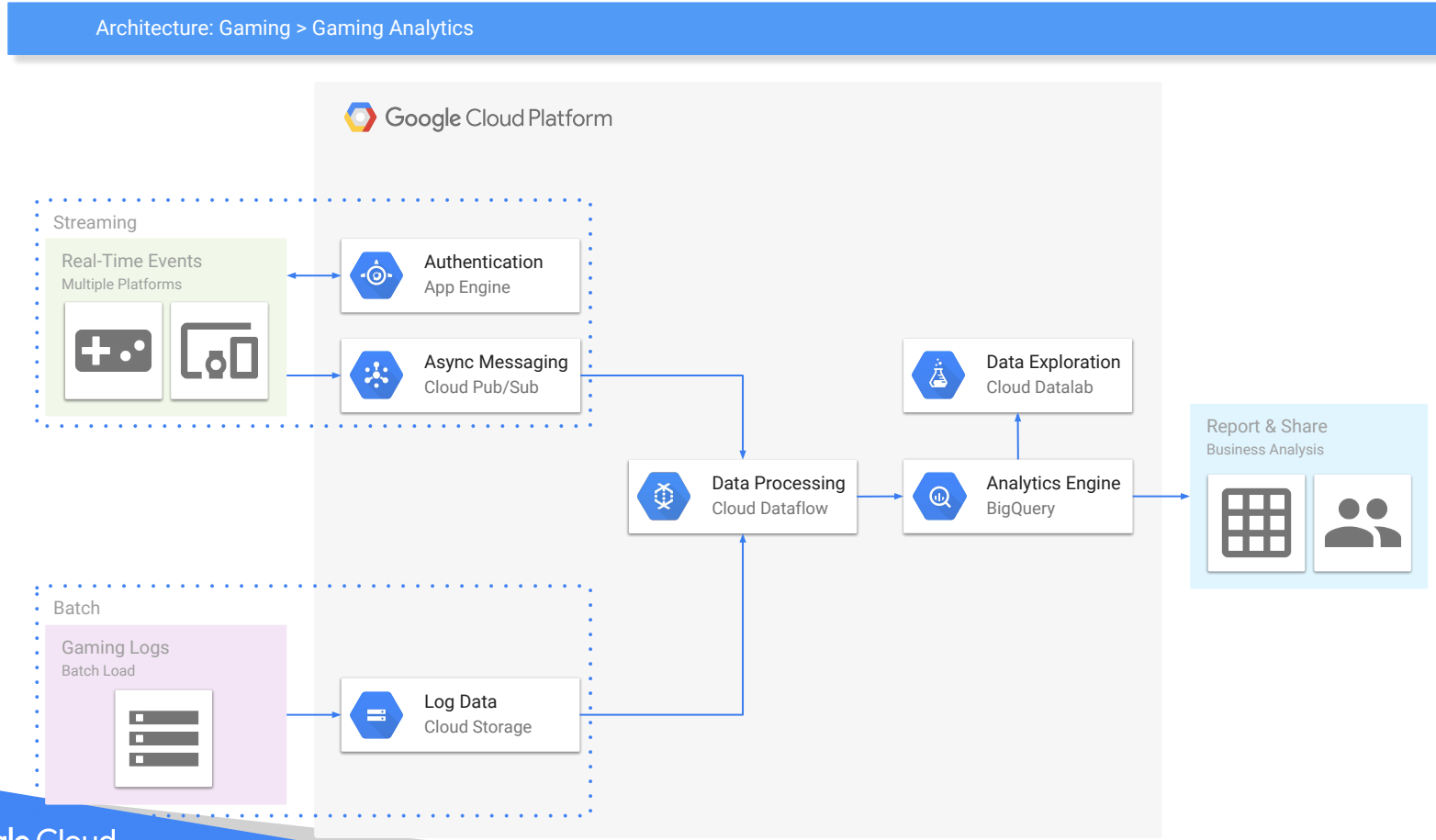
- Google for Work Security and Compliance Whitepaper
  - How Google protects your data.
  - https://static.googleusercontent.com/media/gsuite.google.com/en//files/google-apps-security-and-compliance-whitepaper.pdf

Sample GCP Reference Architectures

Web Apps: Dynamic Hosting

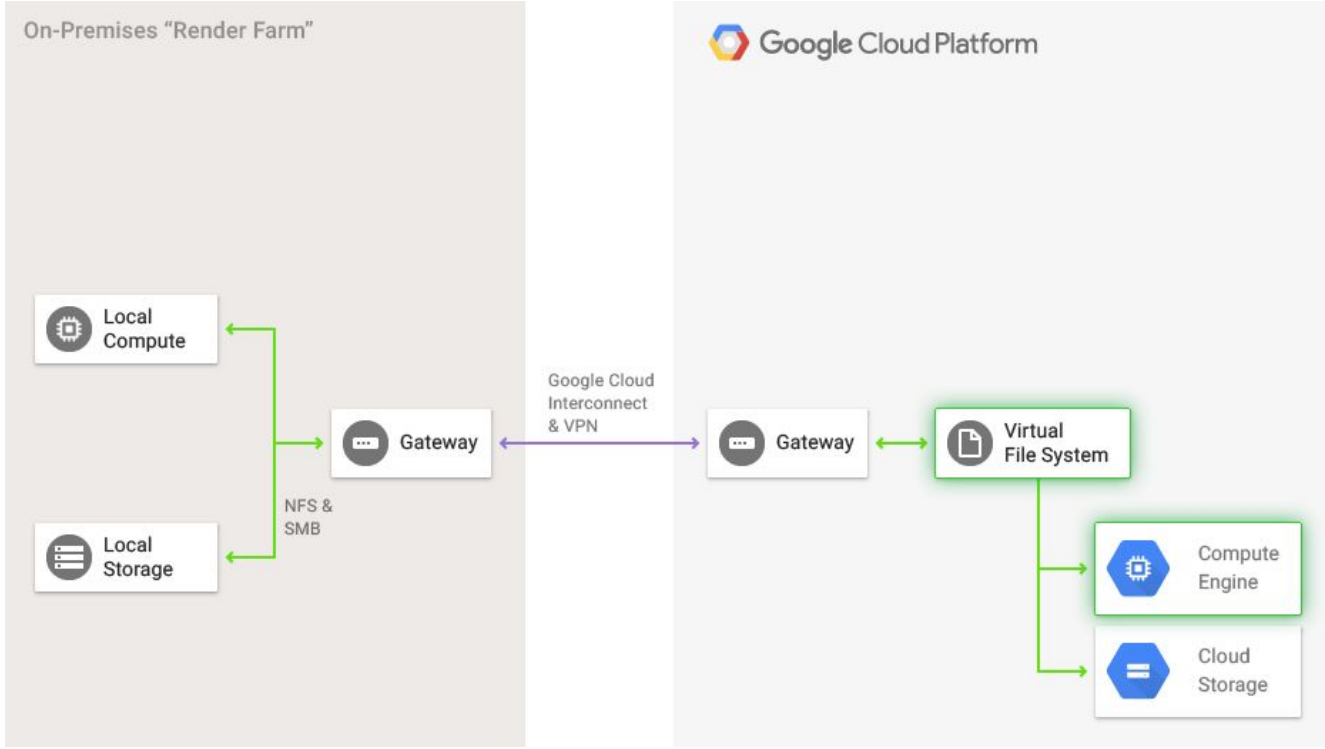


Gaming > Gaming Analytics



Google Cloud

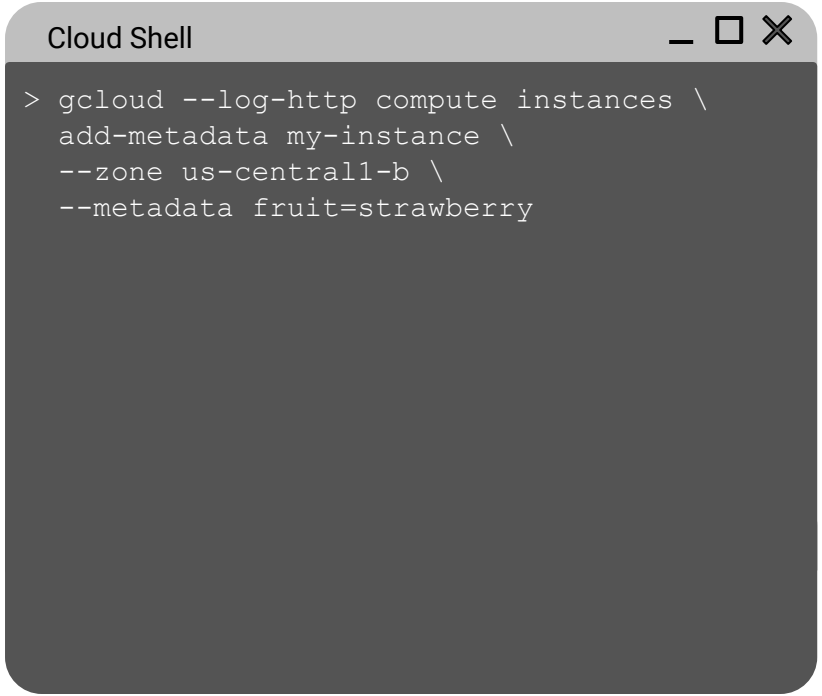
Media: Rendering



Cloud SDK & Cloud Shell

Google Cloud

Cloud SDK



- https://cloud.google.com/sdk/
- Makes Google Cloud scriptable
- Available for all major Operating Systems
- Auto-Installed in all Google Cloud Images

Simple Cloud SDK Exercise

- Login using Google Account

```
gcloud auth login
```

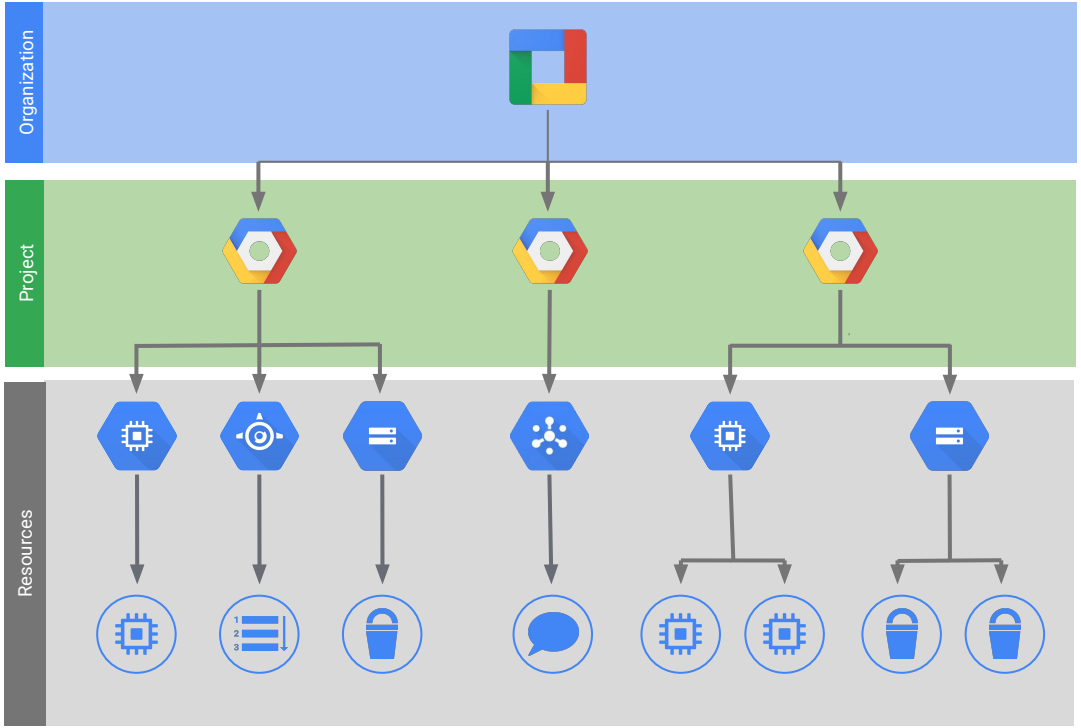
- Query projects

```
gcloud projects list
```

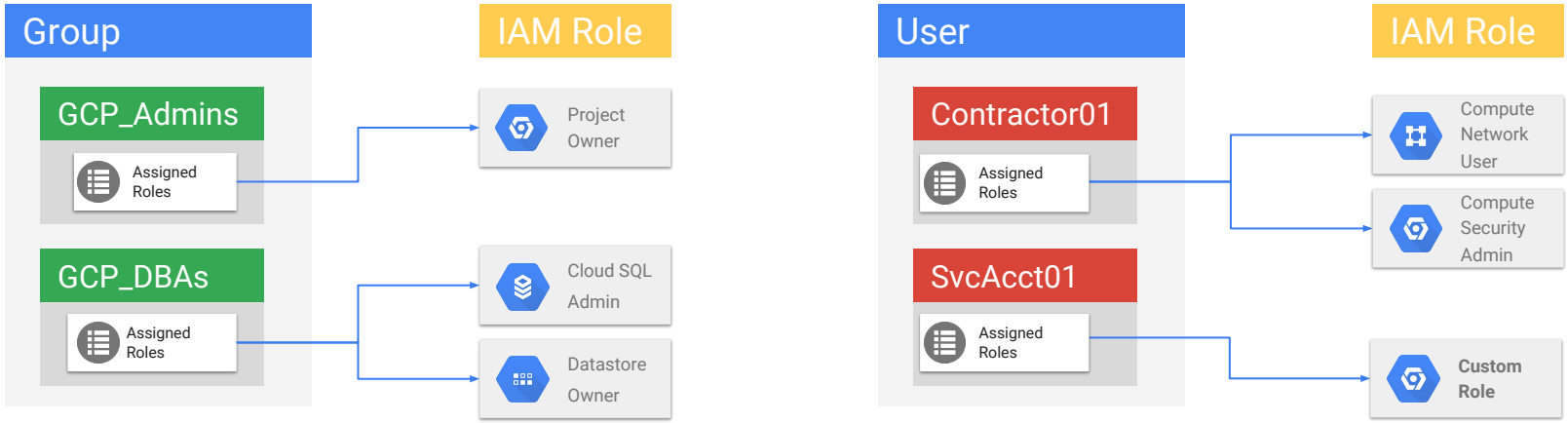
Cloud IAM

IAM - Cloud Resource Manager

Top-down inheritance: additive only



IAM - Assigning Roles



Google Compute Engine & Google Cloud Networking Functions

Google Compute Engine

IaaS		PaaS		SaaS
CPUs, Memory Disks, Interfaces	Servers VM Instances	Clusters Cluster Management	Serverless, Autoscaling	
IT Ops	**Sys Ops**	**Dev Ops**	**Low Ops**	**No Ops**

Managing components:
component health, component function and performance, component backup and replacement

Managing systems:
system availability, system performance, system costs

Managing resources:
users and identities, access to resources, and resource consumption

GCE Standard Machines

Machine Name	Virtual CPUs	Memory (GB)	Operating Systems Available on GCP	Standard Machine Use Cases
n1-standard-1	1	3.75	CentOS, CoreOS, Debian, Red Hat Enterprise Linux (RHEL), SUSE, Ubuntu, Windows Server, SQL Server on Windows Server*	Web Servers, CRM Applications
n1-standard-2	2	7.50		
n1-standard-4	4	15		
n1-standard-8	8	30		
n1-standard-16	16	60		
n1-standard-32	32	120		

GCE High-CPU Machines

Machine Name	Virtual CPUs	Memory (GB)	Operating Systems Available on GCP	High-CPU Machine Use Cases
n1-highcpu-2	2	1.80	CentOS, CoreOS, Debian, Red Hat Enterprise Linux (RHEL), SUSE, Ubuntu, Windows Server, SQL Server on Windows Server*	Genomics, Analytics, IoT
n1-highcpu-4	4	3.60		
n1-highcpu-8	8	7.20		
n1-highcpu-16	16	14.4		
n1-highcpu-32	32	28.8		

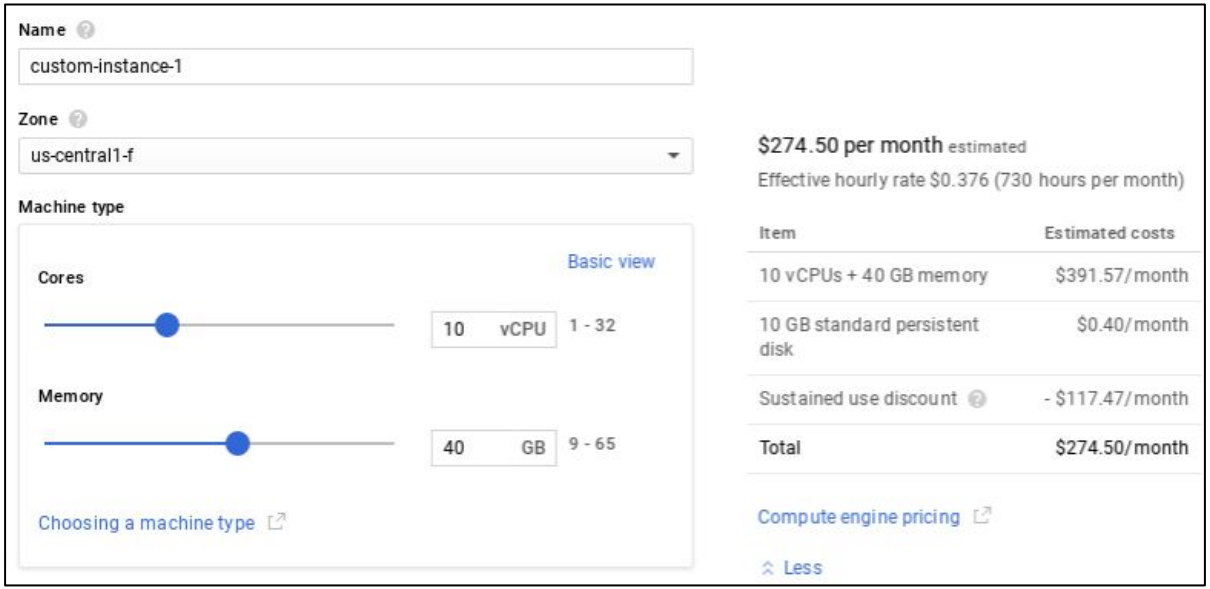
GCE High-Memory Machines

Machine Name	Virtual CPUs	Memory (GB)	Operating Systems Available on GCP	High-Memory Machine Use Cases
n1-highmem-2	2	13	CentOS, CoreOS, Debian, Red Hat Enterprise Linux (RHEL), SUSE, Ubuntu, Windows Server, SQL Server on Windows Server*	Dataproc, Hadoop, Databases
n1-highmem-4	4	26		
n1-highmem-8	8	52		
n1-highmem-16	16	104		
n1-highmem-32	32	208		

GCE Shared-Core Machines

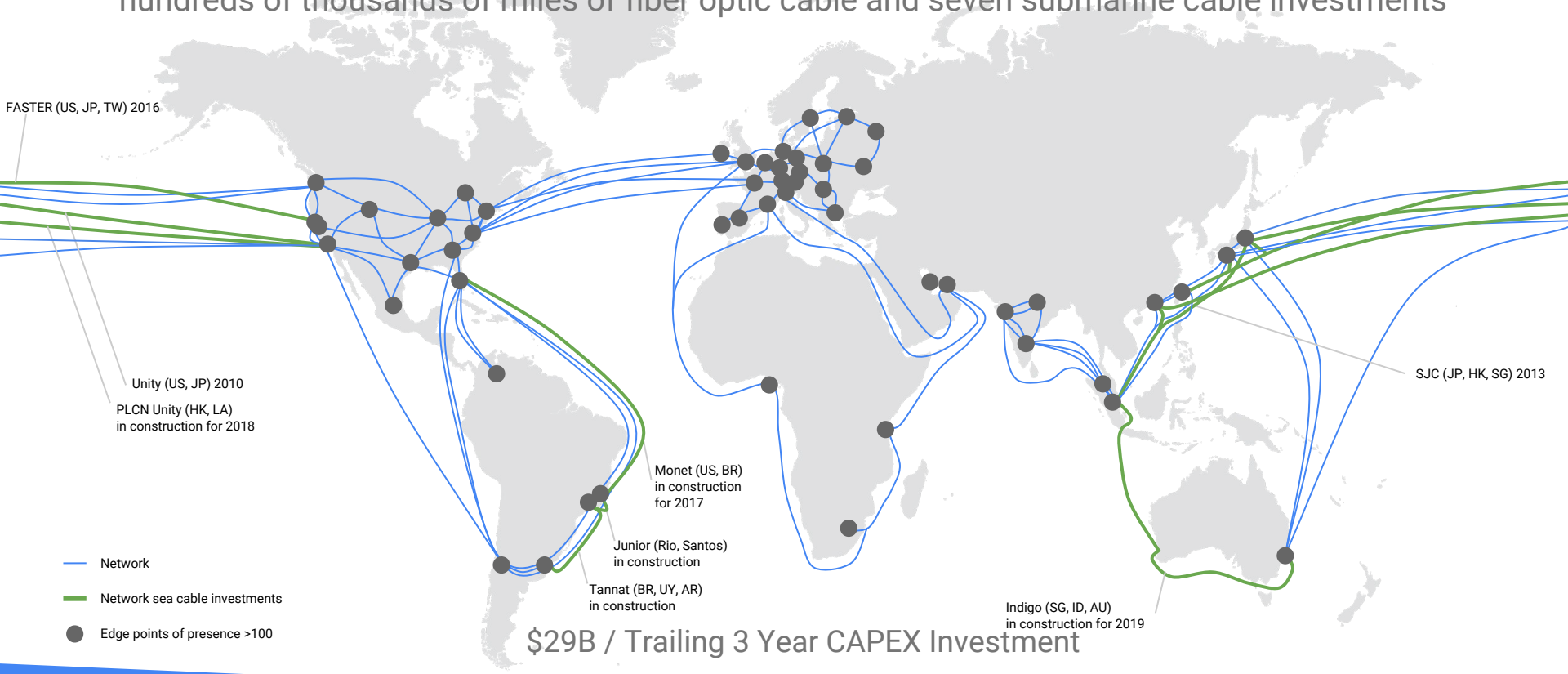
Machine Name	Virtual CPUs	Memory (GB)	Operating Systems Available on GCP	Shared-Core Machine Use Cases
f1-micro	0.2	0.60	CentOS, CoreOS, Debian, Red Hat Enterprise Linux (RHEL), SUSE, Ubuntu	Batch Processing
g1-small	0.5	1.70		

GCE Custom Machines



Google Cloud Network

Google Cloud's well-provisioned global network is comprised of hundreds of thousands of miles of fiber optic cable and seven submarine cable investments



$29B / Trailing 3 Year CAPEX Investment

matrixc

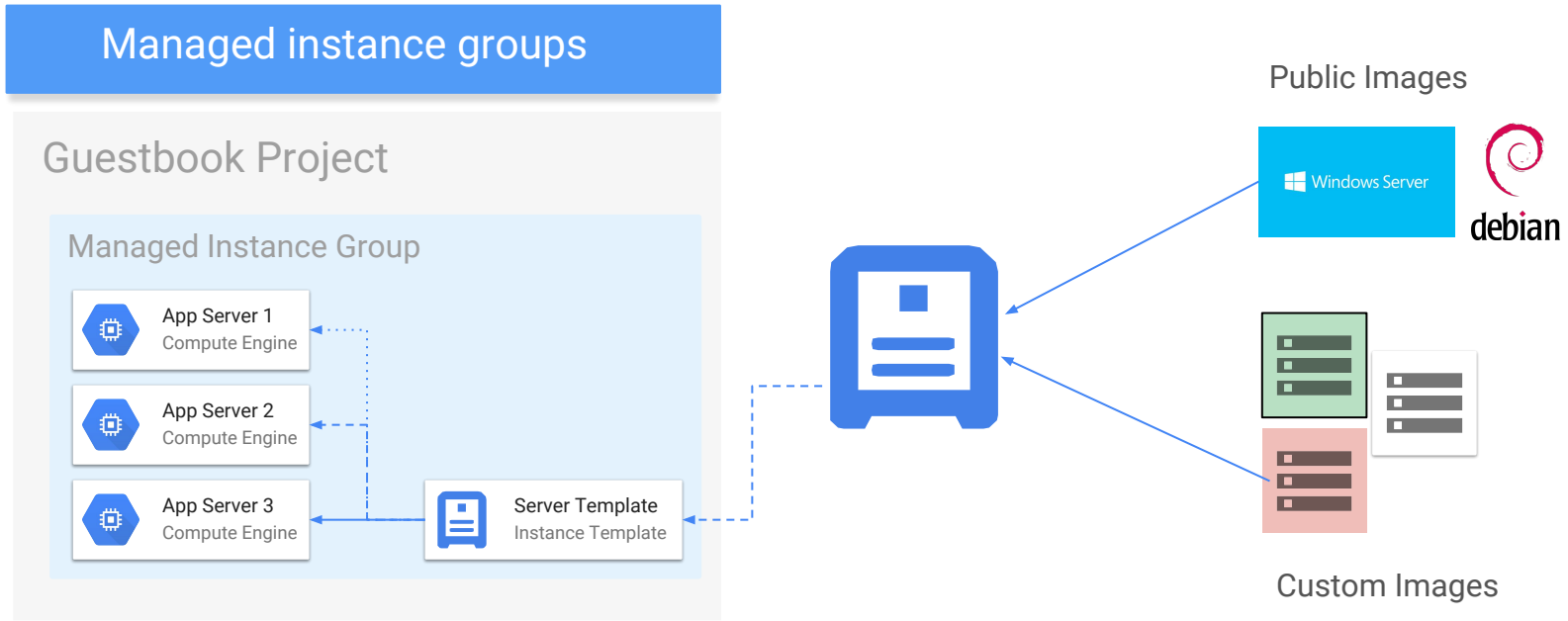
Partner

Google Cloud

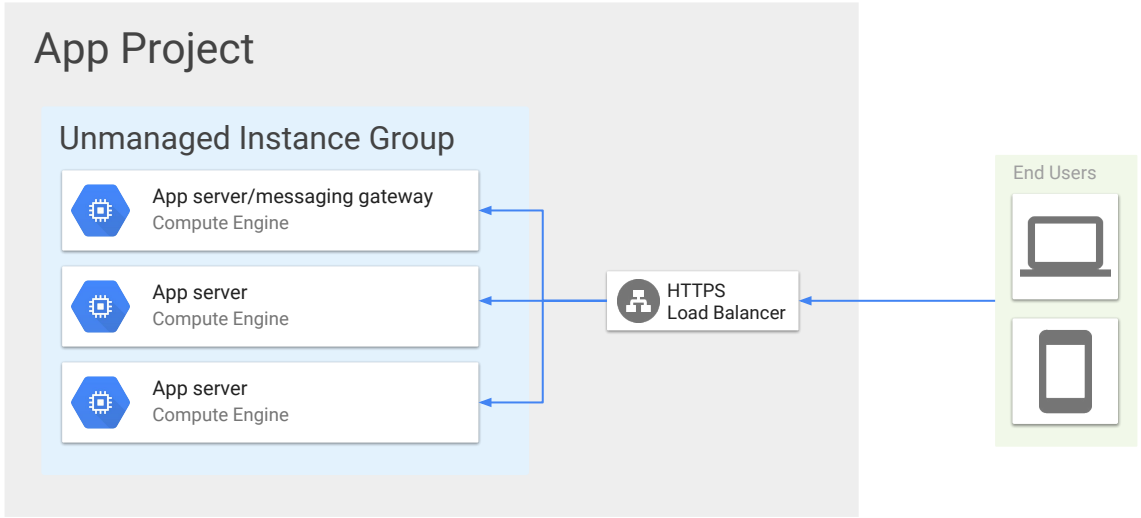
How do you decide the best region to host your system?

gcping.com

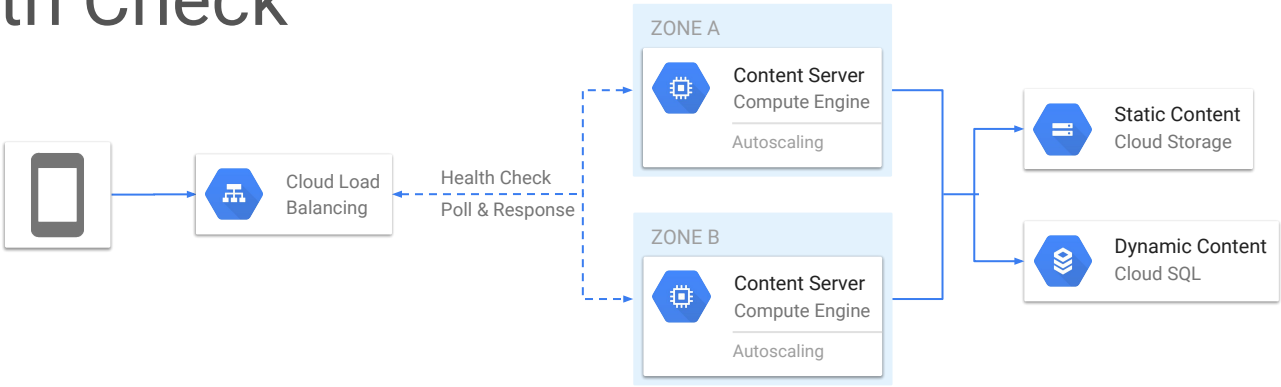
Managed Instance Groups



Unmanaged Instance Groups



Health Check



Health Check Type	HTTP/HTTPS	SSL/TLS
How it works	Probes instance on selected port specified number of times over a configured interval to determine instance health	Simple handshake or request/response
Healthy or unhealthy	Healthy instances return code 200 Failed check marks instance as unhealthy, load balancers stop serving traffic but existing connections are unaffected	Handshake is successful or response is provided Failed check marks instance as unhealthy, load balancers stop serving traffic but existing connections are unaffected
Good for	Use in conjunction with load-balanced managed instance groups to maintain application capacity	

Persistent Disks and Local SSDs

Persistent Disks

Good For	Block storage for GCE and GKE
Use Cases	Snapshots for backups, boot disks
Storage Type	Block storage
Overall Capacity	Up to 64 TB


Snapshots

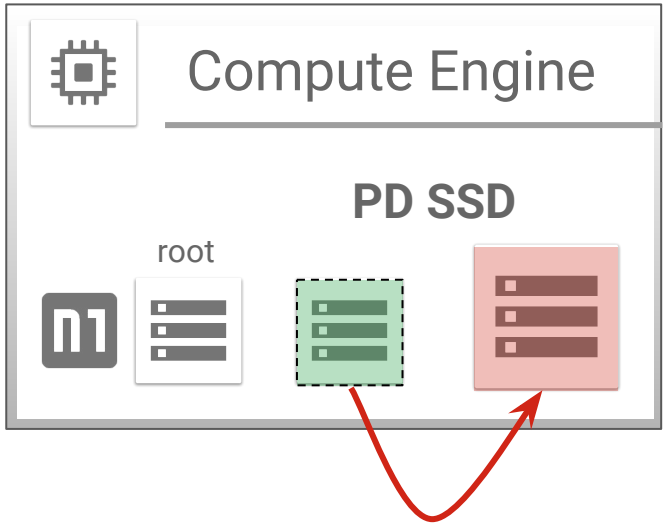

Boot Disks

Local SSDs

Good For	High IOPS, low latency
Use Cases	Temp cache or data processing space
Storage Type	Locally attached SSD
Overall Capacity	Up to 3TB


Low value data

Resizing Persistent Disks



Improve I/O performance by increasing storage capacity of a persistent disk

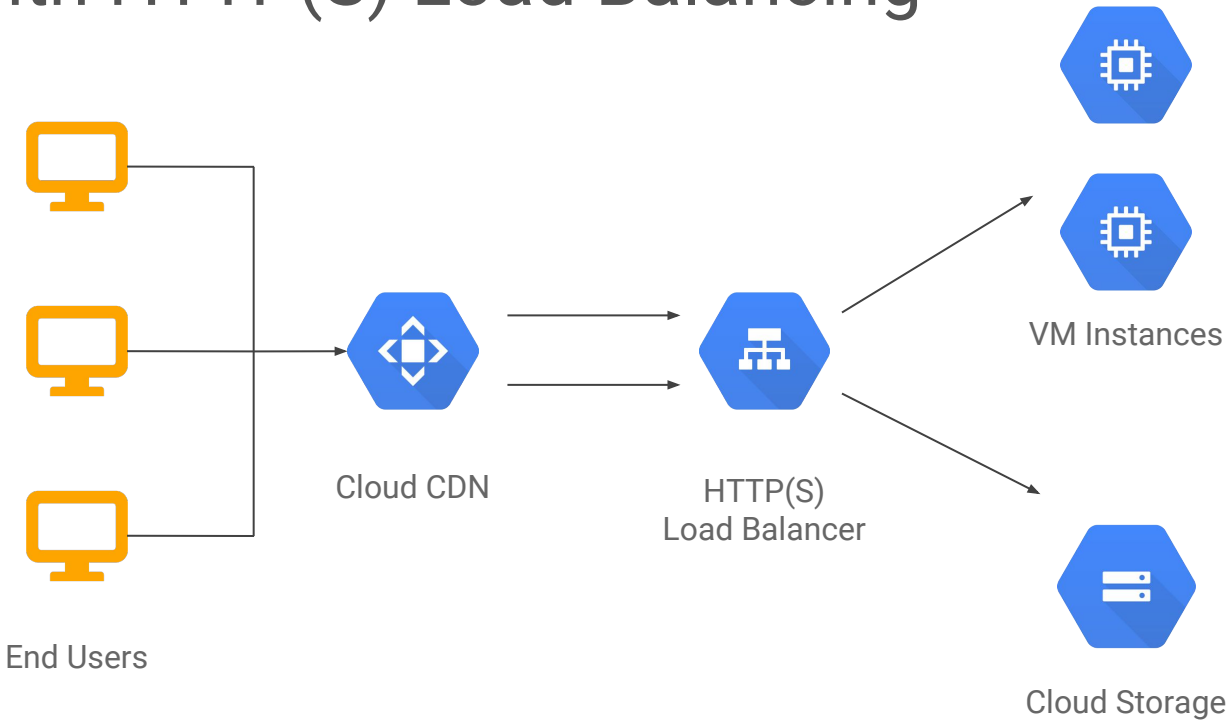
Edit the persistent disk to increase its size and IOPS capacity

	Sustained Random IOPS Limit (Read or Write)	Sustained throughput limit (mb/s)
10 GB SSD	300	4.8
100 GB SSD	3000	48
500 GB SSD	15000	240

Load Balancing

	HTTP(S) Load Balancing (lvl 7)	TCP/UDP Load Balancing (lvl 4)
Good For	Provides Global load balancing for HTTP(S) requests. Use for cross-regional load balancing	Can load-balance additional TCP/UDP protocols. Intended for Non HTTP traffic, but can be set up to balance HTTPS traffic.
Considerations	Must be internet-facing	Cannot automatically forward traffic across regions like HTTP(S) LB.
Supported ports	Only supports ports 80 and 443	Supports ports: 25, 43, 110, 143, 195, 443, 465, 587, 700, 993, 995
Configure	HTTP Load Balancer HTTPS Load Balancer	TCP/UDP Load Balancer SSL Proxy
Options	Internet-facing (only) Single or Multi-Region	Internet-facing or Internal Single or Multi-Region

CDN with HTTP(S) Load Balancing



Short Exercise

- Spin up a Linux Compute Instance
- Upload a file to Cloud Storage Bucket via UI
- Download a file from Cloud Storage via CLI

Cloud Functions [Beta]

What are Google Cloud Functions?

- Google Cloud Functions is a serverless execution environment for building and connecting cloud services.
- Write simple, single-purpose functions that are attached to events emitted
- Google Cloud Function is triggered when an event being watched is fired.
- The code executes in a fully managed environment. There is no need to provision any infrastructure or worry about managing any servers.
- Cloud Functions are written in Javascript and execute in a Node.js v6.11.1 environment on Google Cloud Platform.

Google Cloud Functions

matrixc

Example: Logging Changes in Google Cloud Storage

```
exports.processFile = function(event, callback) {
  console.log('Processing file: ' + event.data.name);

  const file = event.data;

  if (file.resourceState === 'not_exists') {
                console.log(`File ${file.name} deleted.`);
                callback();
  } else if (file.metageneration === '1') {
    // metageneration attribute is updated on metadata changes.
    // on create value is 1
    console.log(`File ${file.name} uploaded.`);
        }
  callback();
};
```

Example: Pulling a JSON from Pub/Sub

```
exports.pubSubJSONStream = function (event, callback) {
    const pubsubMessage = event.data;
    const jsonStr = Buffer.from(pubsubMessage.data,
'base64').toString();
    const messages = JSON.parse(jsonStr);
    const rows = messages.map(function(x) {
       return JSON.parse(x);
    });
    console.log(rows); //don't log too many though
 };
```

Google Kubernetes Engine

Why use Google Kubernetes Engine?

You have a container-centric view of the world.

- **Deploying or maintaining a fleet of VMs has been a challenge** and you've determined that **containers** are the solution.
- You've **containerized your workload** and need a system on which to run and manage it.
- You **never want to touch a server or infrastructure.**

Kubernetes Engine

Cluster manager and orchestration engine built on Google's container experience

What are containers?

- Just like shipping containers, a software container make it easier for you to package, manage, and ship your code.
- You write software applications that run in a **container.** The container provides the operating system (OS) you need to run your application. The container will run on any container platform.
    - This can save a lot of time and cost compared to running servers or virtual machines.
    - Two short articles for more detail on Linux Containers and Container OS.

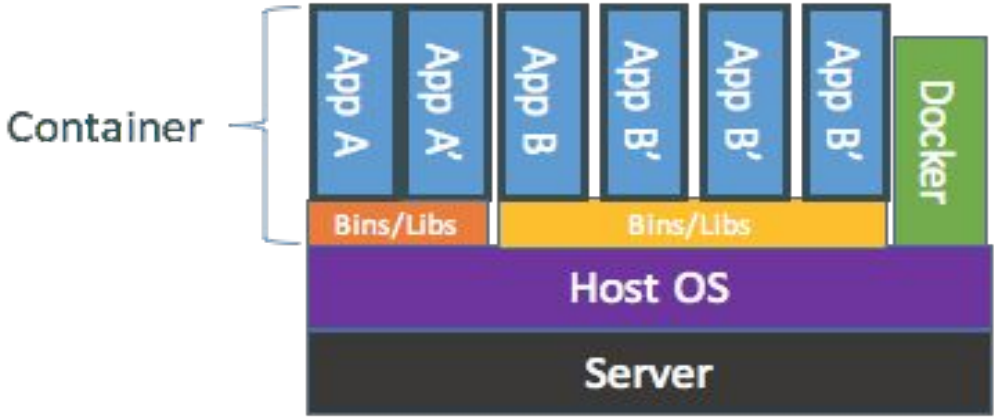
matrixc

What are containers? (continued)

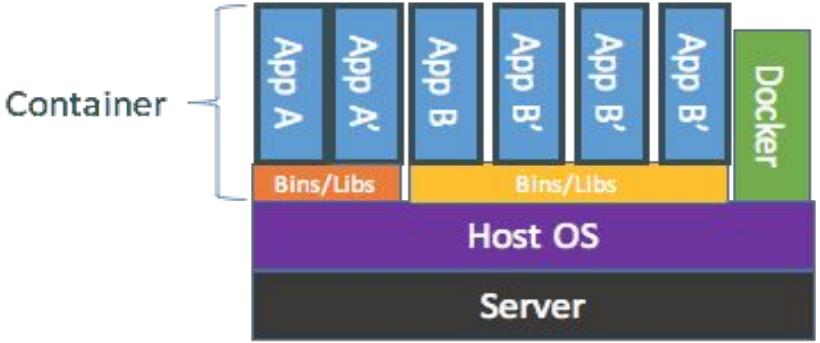
- Docker is the tool that puts your application and everything it needs in the container.
- Once your application is in a container, you can move it anywhere that will run Docker containers—any laptop, server, or cloud provider.
    - This portability makes code easier to produce, manage, troubleshoot, and update, which creates opportunities for Google partners to sell services.
- As a service provider, containers makes it easy for you to develop code that can be **ported** to your customer and back.

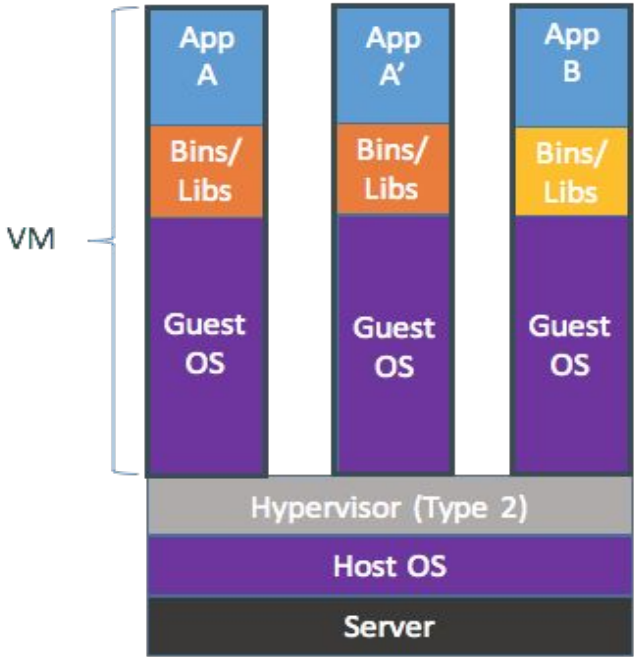
What do containers look like?

Containers are “isolated” from one another: each container is it’s own application and it can be turned on and off without affecting other containers. However it may share an OS and other needed things like bins and libraries, which can be used by multiple applications. From the container’s point of view, it thinks that it’s running each resource on it’s own machine.



Why are containers better than virtual machines?



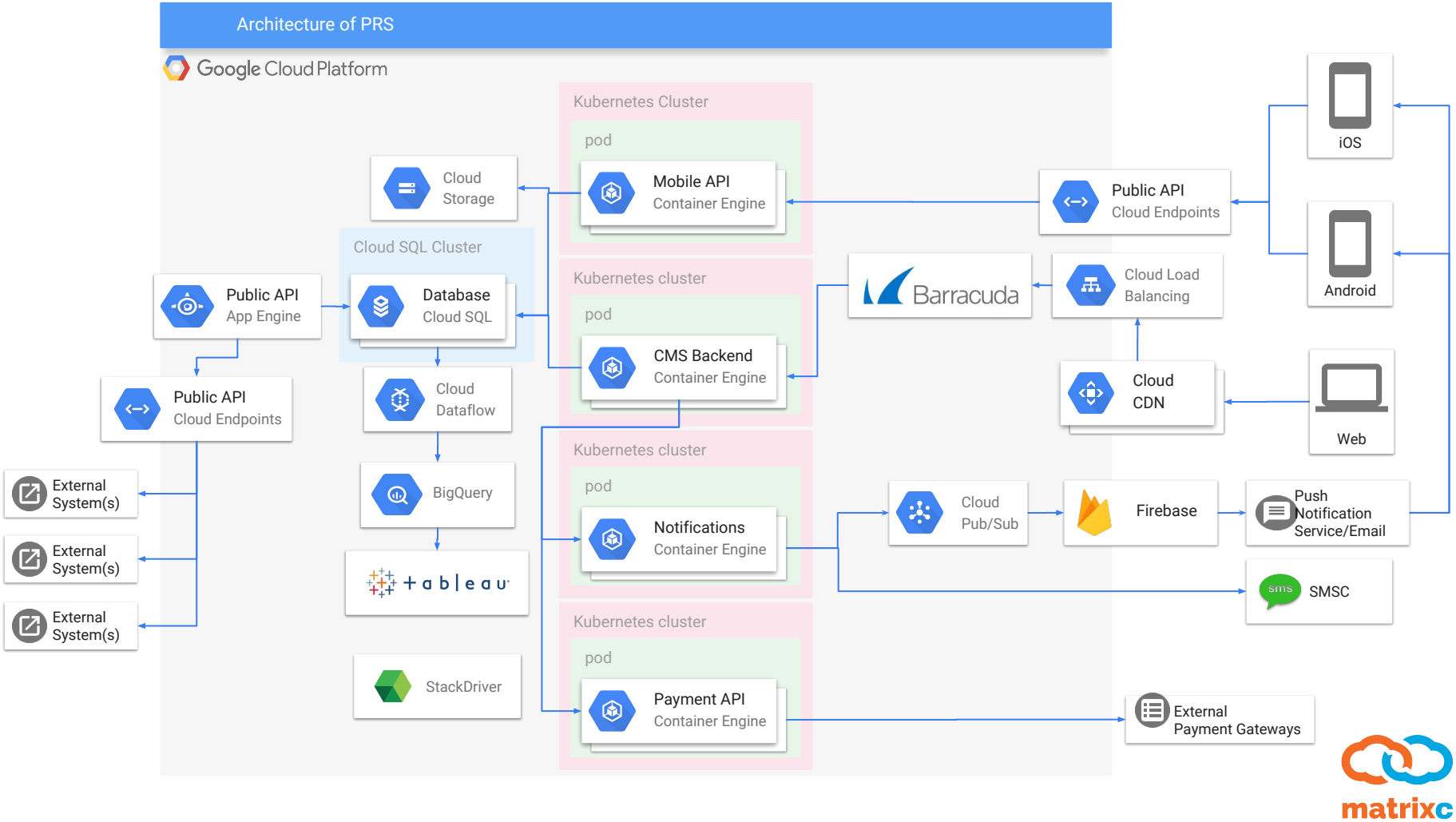


In these two graphics comparing containers to VMs, you can see that containers can run more applications with less resources. Containers are much simpler. The result is significantly faster deployment, much less overhead, easier migration, and faster restart.

What to avoid

1. Although existing monolithic applications can be moved to Container Engine, these are non-ideal applications for now, as Container Engine is geared for stateless workloads, where any container can be killed at any time.
2. Storing persistent data, such as in NoSQL databases, is currently *possible* but complex.

† Kubernetes is on a 3-month release cycle. New versions of Kubernetes are traditionally available on Container Engine for new clusters on the release day, and within a week for upgrading existing clusters.


Architecture of PRS
Google Cloud Platform
Kubernetes Cluster
pod
Mobile API
Container Engine
Cloud
Storage
Cloud SQL Cluster
Public API
App Engine
Database
Cloud SQL
Kubernetes cluster
pod
CMS Backend
Container Engine
Barracuda
Cloud Load
Balancing
Public API
Cloud Endpoints
iOS
Android
Cloud
CDN
Web
Public API
Cloud Endpoints
Cloud
Dataflow
BigQuery
+ableau
External
System(s)
External
System(s)
External
System(s)
Kubernetes cluster
pod
Notifications
Container Engine
Cloud
Pub/Sub
Firebase
Push
Notification
Service/Email
SMSC
sms
Kubernetes cluster
pod
Payment API
Container Engine
StackDriver
External
Payment Gateways
matrixc

Codelabs: Kubernetes Hello World

Before you begin, sign into your Google Cloud Project & enable Kubernetes Engine

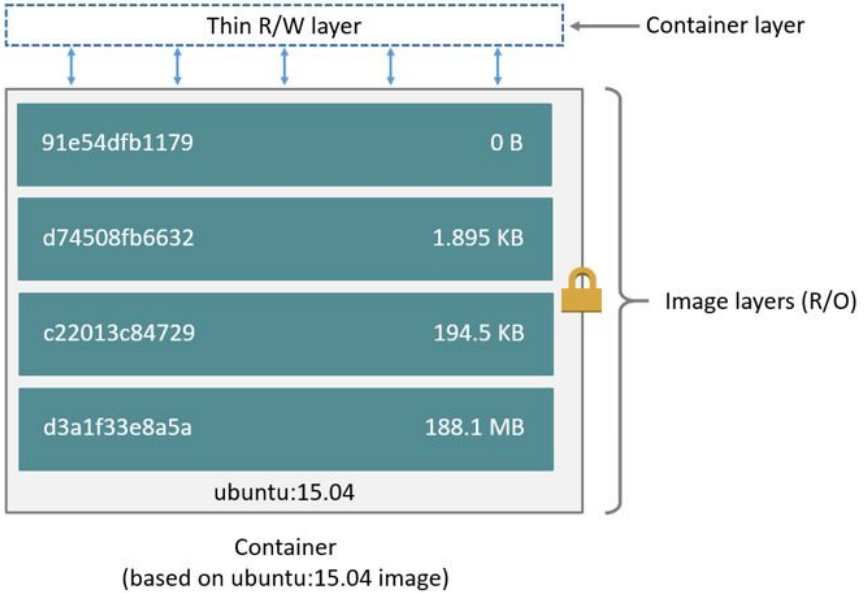
https://goo.gl/qWbheq

matrixc

Understanding Docker Layers

Sample Code

```
FROM ubuntu:15.04
COPY . /app
RUN make /app
CMD python /app/app.py
```



matrixc

What’s gcr.io?

- Google’s Container Registry Address!
- This is where we keep our container images (defaults to US)
- When we build clusters, this is where we can pull them from to make our life easier

Spinning up your first cluster

- Default Container Size: 2
- You specify the specs & the location
- OS depends on the OS you built the cluster registry from!

What you have learned

- Difference between compute and containers
- How to spin up containers
- How to update containers

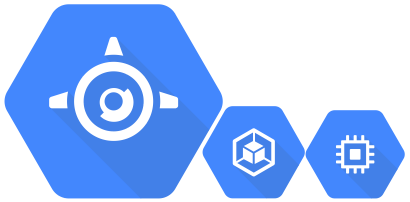
Welcome to Day 2!

Google App Engine

Why use Google App Engine?

You have an app-centric view of the world.

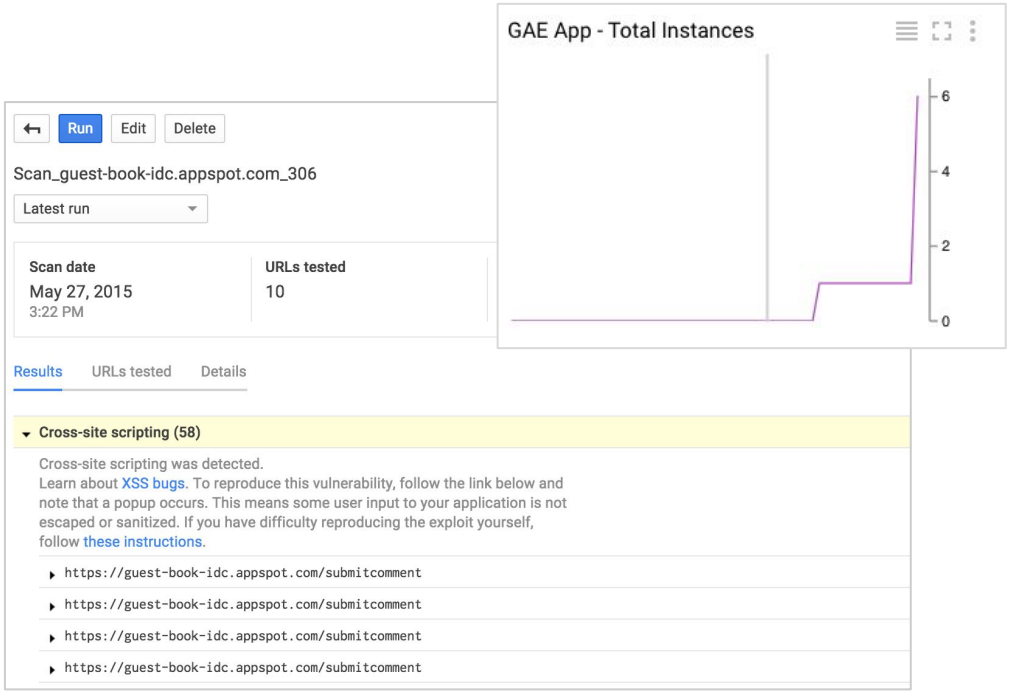
- You want to **focus on writing code** and never touch a server, cluster, or infrastructure.
- Building **quickly** and **time to market** are highly valued.
- You want to sleep at night and **not worry about a pager going off** or **5xx errors**.
- You **expect your app to have high availability** without a complex architecture.



App Engine

A flexible, **zero ops** platform for building highly available apps

Zero ops



Focus on code, not managing infrastructure. Fully managed by Google site reliability engineers.

What workloads are ideal?

App Engine's benefits make it ideally suited for building

Mobile backends, especially social and casual games

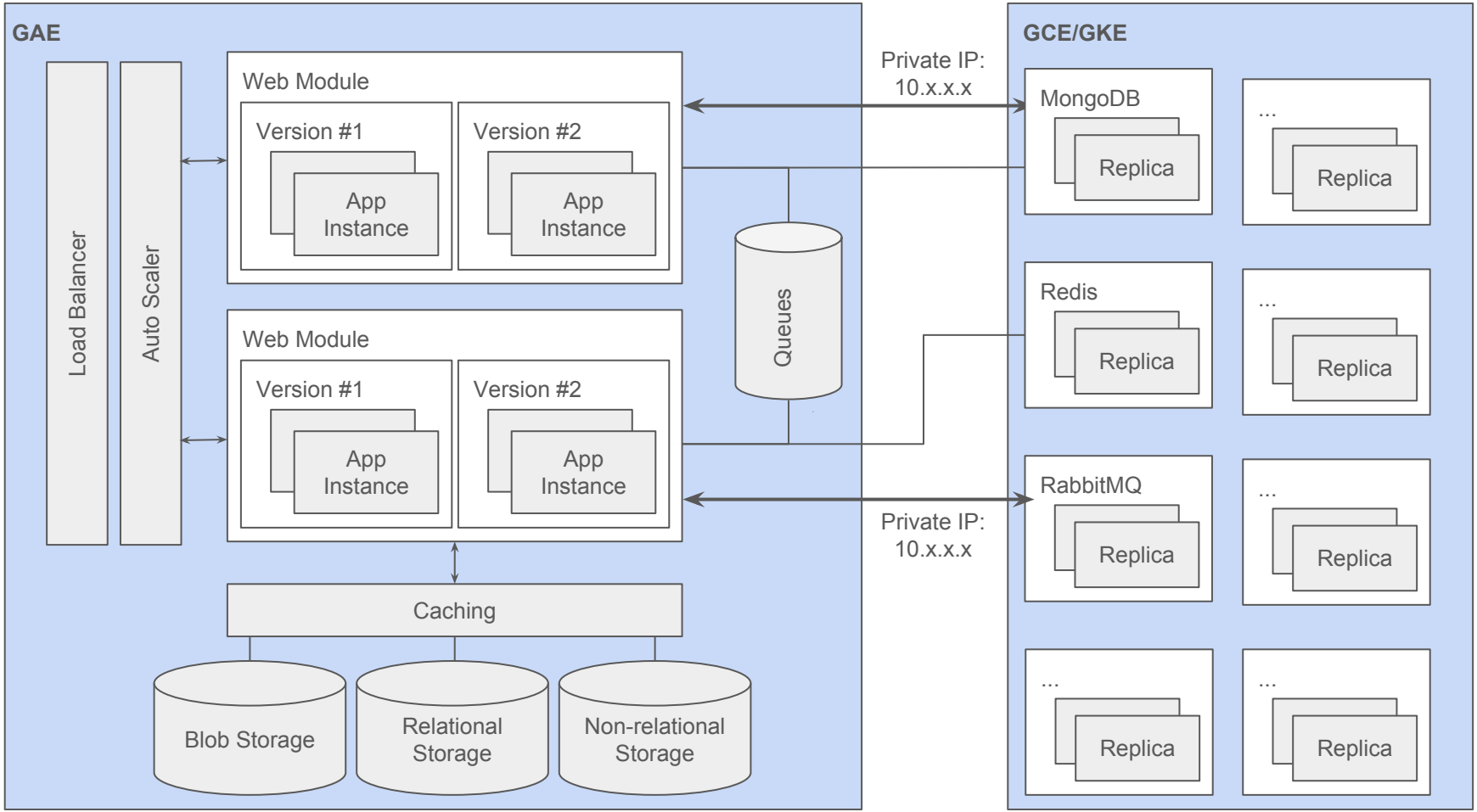
Software as a Service (SaaS) applications that can disrupt stagnant industries

Internal IT apps that improve productivity and revenue (think Googelplex)

Internet of Things (IoT) front end and backend workloads.

Any web frontend (are you running Tomcat or nginx? stop.)

App Engine Flexible



Exercise: Hello World on Python w/ App Engine

https://goo.gl/NAVnxb

matrixc

What you have learned

- Difference between App Engine and Container Engine
- App Engine is actually Container Engine as a managed service
- If Container Engine was easy, App Engine is even easier (but with limitations)

Google Cloud Storage

Data Sources

amazon web services

Microsoft Azure

EMC²

NetApp™

DoubleClick by Google

YouTube

Google Analytics 360 Suite

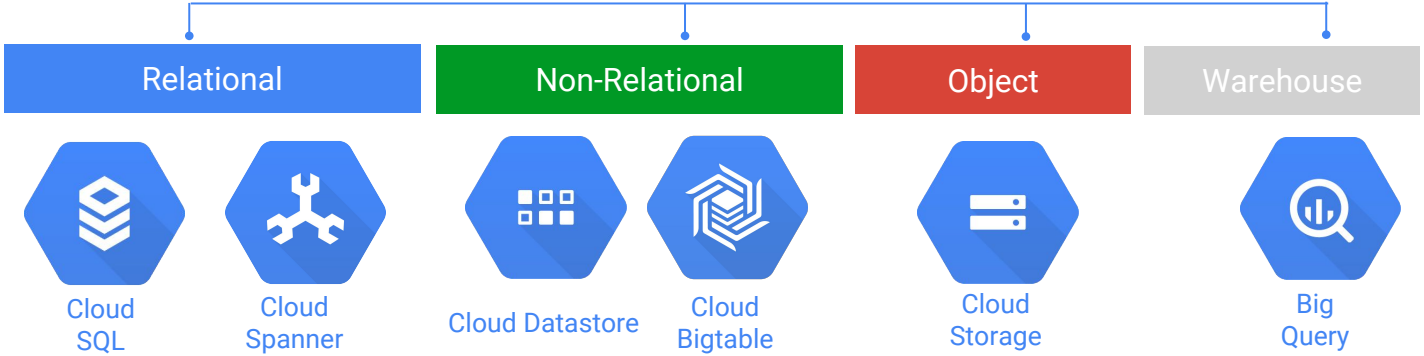
Google AdWords

salesforce

krux

Cloud

GCP Storage Portfolio

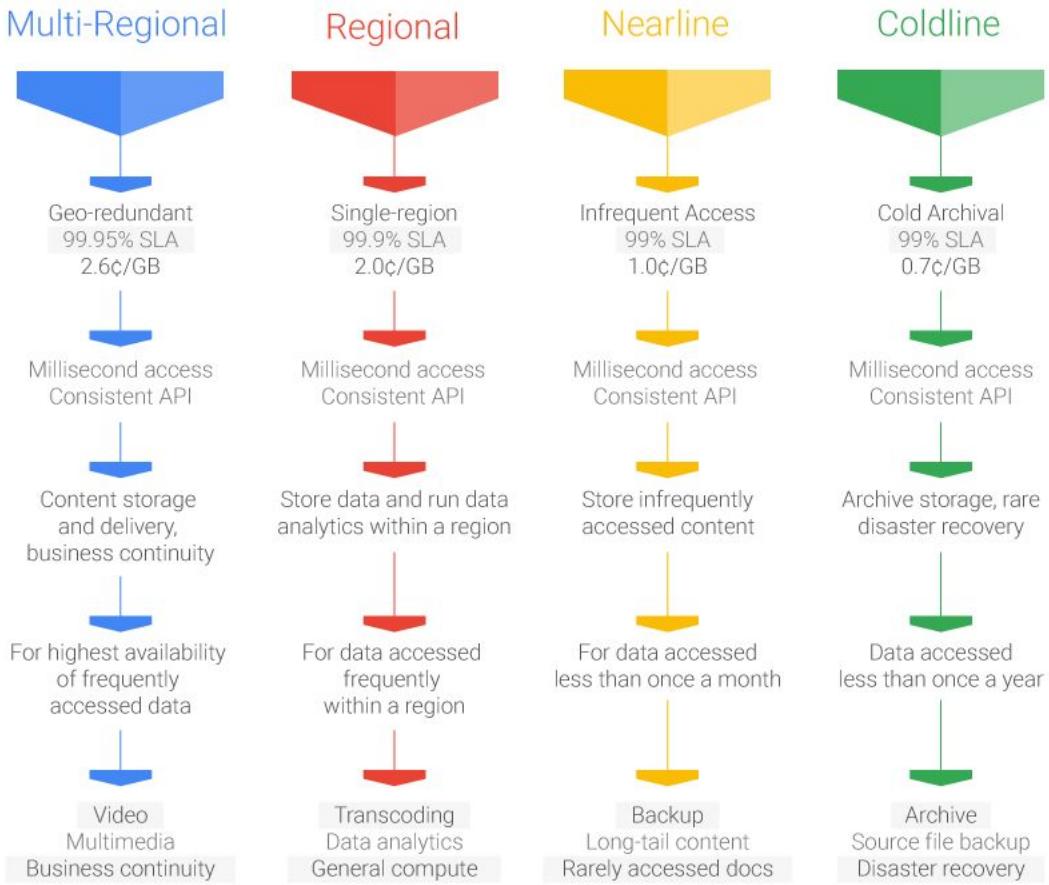


Google Cloud Storage

Foundational component to any cloud platform.

How to get data into GCS?

- APIs
- GCP Console
- gsutil
- GCS FUSE
- Cloud Storage Transfer Service
- Offline media import/export
- Project Fireball



Laws of physics and speed of light

	Bandwidth (assuming 100% utilization)					
Data Size	1 Mbps	10 Mbps	100 Mbps	1 Gbps	10 Gbps	100 Gbps
1 GB	3 hrs	18 mins	2 mins	11 secs	1 sec	0.1 secs
10 GB	30 hrs	3 hrs	18 mins	2 mins	11 secs	1 sec
100 GB	12 days	30 hrs	3 hrs	18 mins	2 mins	11 secs
1 TB	124 days	12 days	30 hrs	3 hrs	18 mins	2 mins
10 TB	3 years	124 days	12 days	30 hrs	3 hrs	18 mins
100 TB	34 years	3 years	124 days	12 days	30 hrs	3 hrs
1 PB	340 years	34 years	3 years	124 days	12 days	30 hrs
10 PB	3404 years	340 years	34 years	3 years	124 days	12 days
100 PB	34048 years	3404 years	340 years	34 years	3 years	124 days

OLAP vs OLTP: Which Fits My Use Case?

	OLTP	OLAP
	OnLine Transaction Processing	OnLine Analytical Processing
Data Source	Operational	Historical
Focus	Updating/Retrieve	Reporting
Queries	Simple	Complex
Query Latency	Low	High
Google Cloud Platform Products	 Cloud SQL, Cloud Datastore, Cloud Spanner, BigTable	 BigQuery

Google Cloud

Database Characteristics

	Product	Interface	Query Latency	Typical Size	Storage Structure
	**CloudSQL**	SQL	Low (ms)	< 10TB	Relational
	**Datastore**	Proprietary / NoSQL	Medium (10s of ms)	< 200TB	Document
	**Bigtable**	HBase API	Low (ms)	Terabytes - Petabytes	Key/Value
	**BigQuery**	REST / SQL / WebUI	High (s)	Terabytes - Petabytes	Columnar
	**Spanner**	SQL	Low (ms)	Terabytes	Relational

Google Cloud Storage

	Access frequency	At rest pricing	Retrieval pricing	SLA
Multi-regional	Frequent, cross-regional	$0.026 per GB/month	FREE	99.95%
Regional	Frequent, single-region	$0.02 per GB/month	FREE	99.9%
Nearline	Less than once per month	$0.01 per GB/month	$0.01 per GB	99.0%
Coldline	Less than once per year	$0.007 per GB/month	$0.05 per GB	99.0%

Short Exercise: Installing and Using GCSFuse

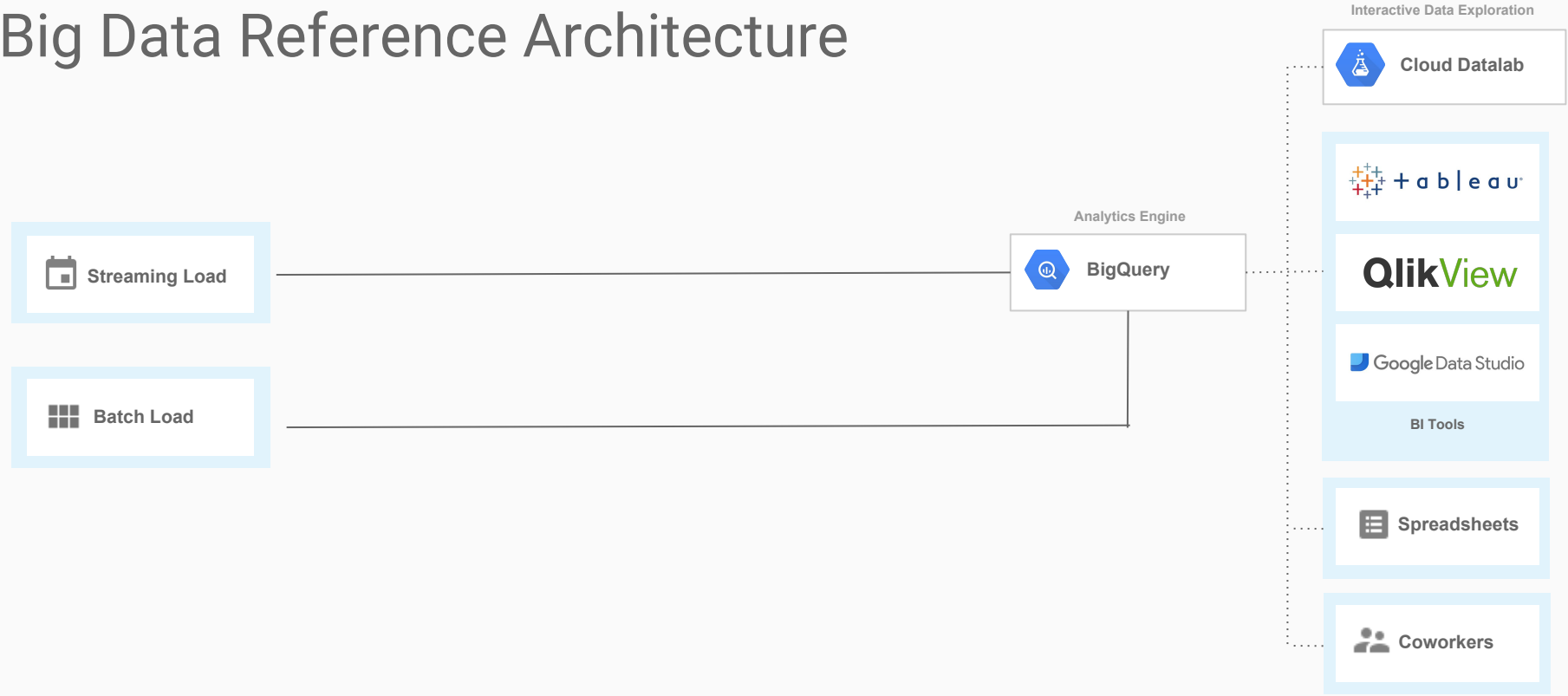
- First of, be aware of the latency between regions. The further the region, the more latency.
- Installation Steps: https://goo.gl/xYsM4q
- GCSFuse Example:

```
mkdir /path/to/mount
gcsfuse example-bucket /path/to/mount
ls /path/to/mount
```

Sample Reference Architecture for Big Data

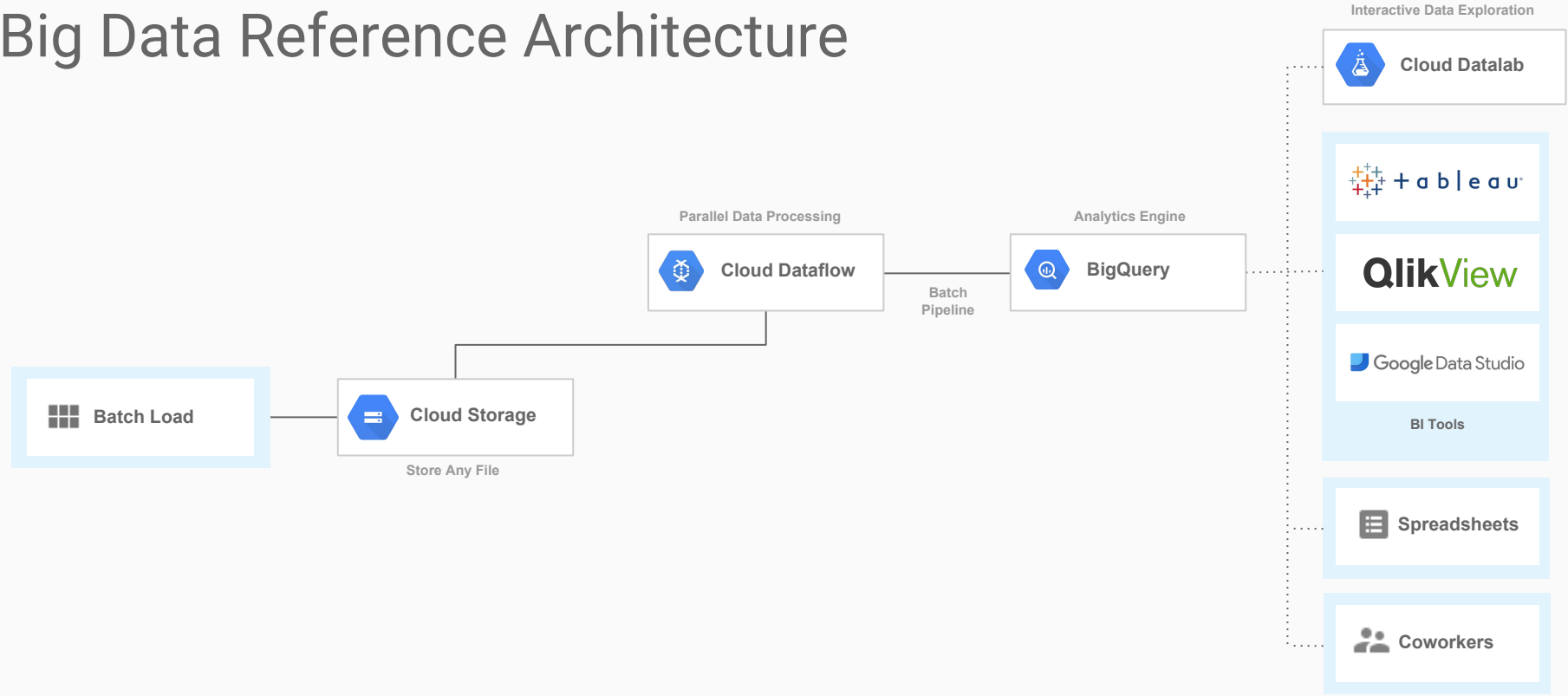
Direct to BigQuery

Big Data Reference Architecture



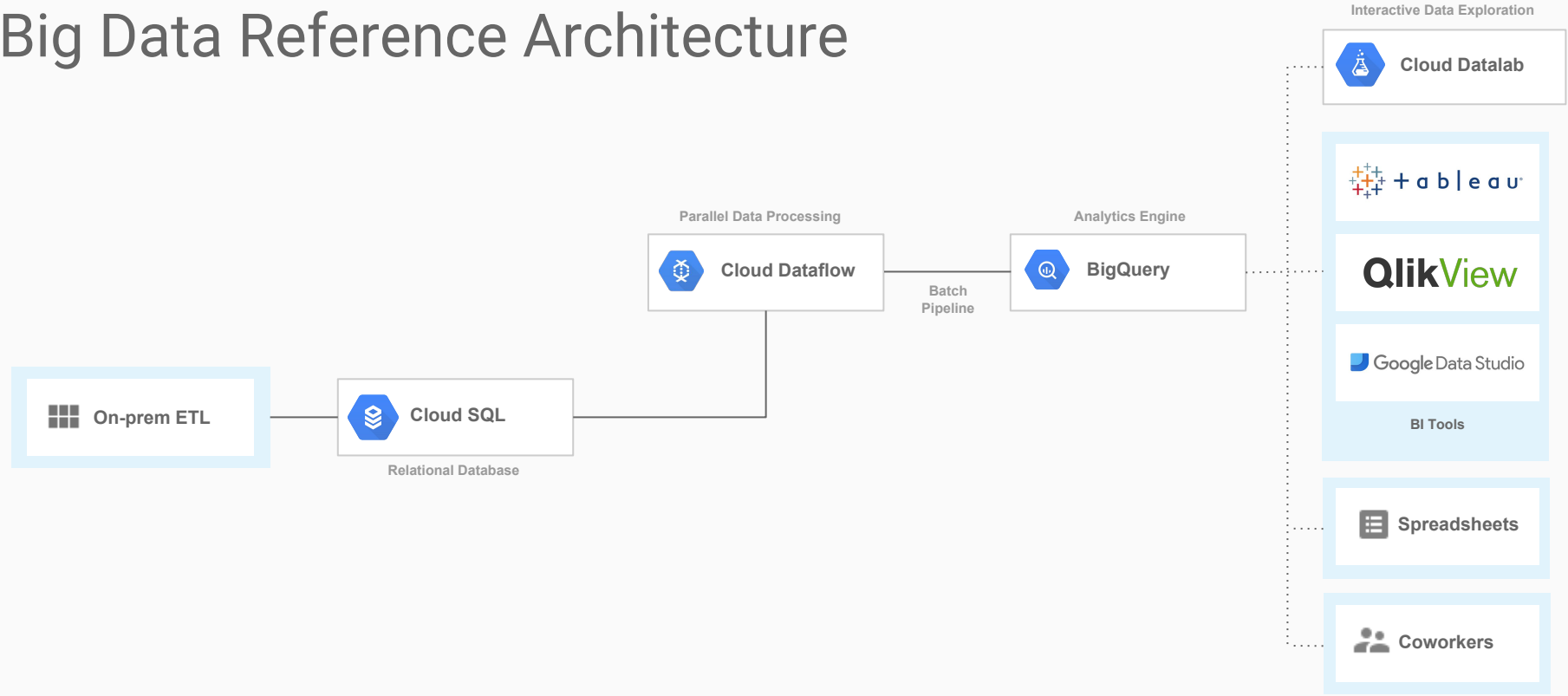
Batch Ingest, Transform, Load

Big Data Reference Architecture

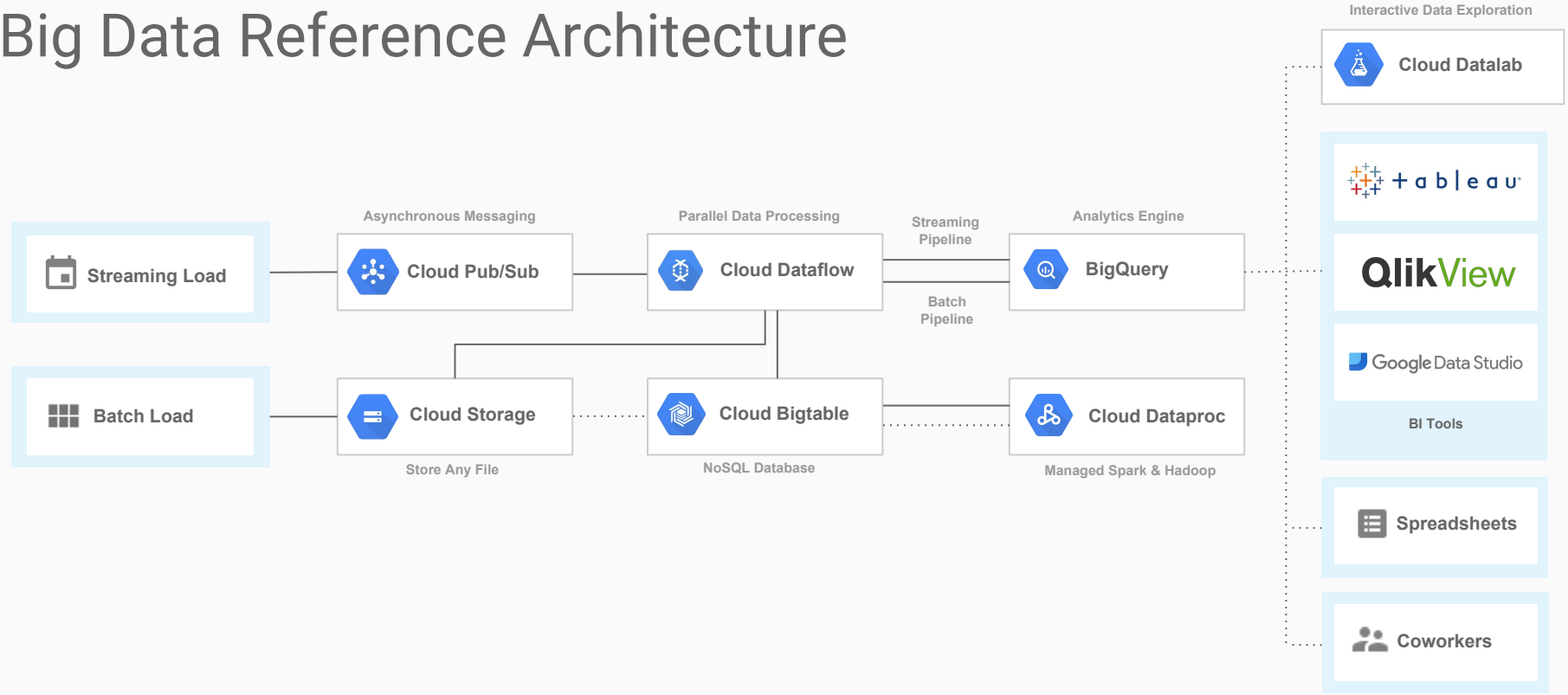


Stage, Transform, Load

Big Data Reference Architecture



Big Data Reference Architecture



Google Cloud

Data Lifecycle Steps

Ingest

The first stage is to pull in the raw data, such as streaming data from devices, on-premises batch data, application logs, or mobile-app user events and analytics.

Store

After the data has been retrieved, it needs to be stored in a format that is durable and can be easily accessed.

Process & Analyze

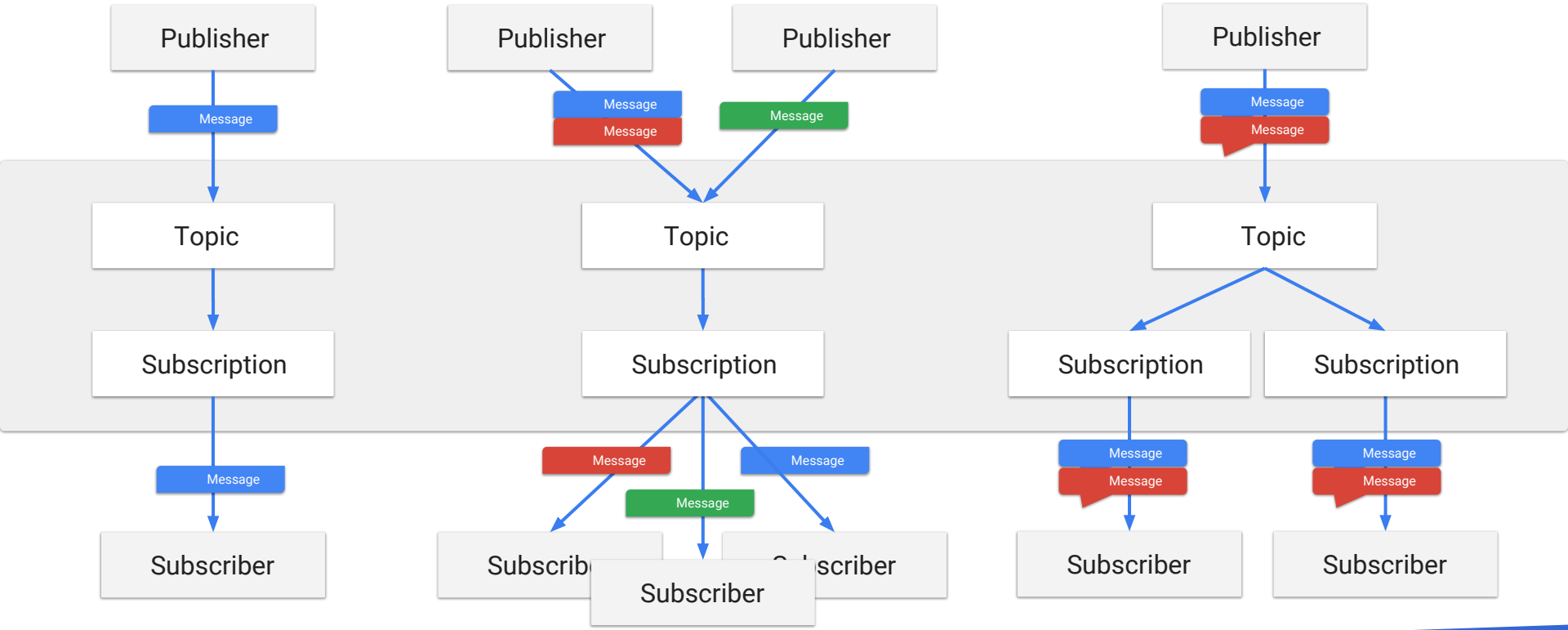
In this stage, the data is transformed from raw form into actionable information.

Explore & Visualize

The final stage is to convert the results of the analysis into a format that is easy to draw insights from and to share with colleagues and peers.

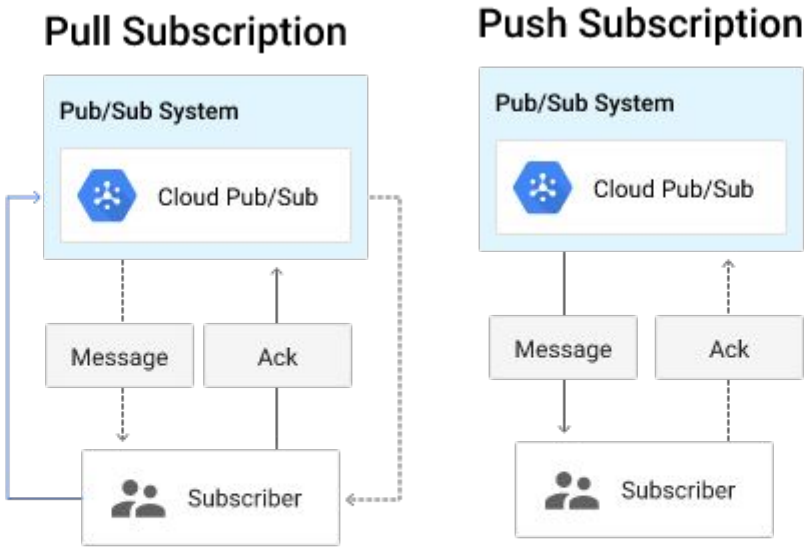
Google Pub/Sub

Publish/Subscribe patterns



At-Least-Once Delivery Guarantee

- A subscriber ACKs each message for every subscription
- A message is resent if subscriber takes more than “ackDeadline” to respond
- A subscriber can extend the deadline per message



Order & Duplication: No Guarantees

- Messages can be delivered in any order, especially with large backlog
- Duplication will happen
- Dataflow solves all of this @ cost of latency
- New features coming up to reduce ordering problems.

How Do We Subscribe to Pub/Sub?

The Current Way: Pull

The Other Way: Push

Pushing from Pub/Sub

- Set a URL for the push subscription
- The “Webhook” receiving the push must respond with a 200, 201, 202, 204 or 102
- Example message:

```
{
 "message": {
    "attributes": {
      "key": "value"
    },
    "data": "SGVsbG8gQ2xvdWQgUHViL1N1YiEgSGVyZSBpcyBteSBtZXNzYWdlIQ==",
    "message_id": "136969346945"
  },
  "subscription": "projects/myproject/subscriptions/mysubscription"
}
```

Service Features

- Single, global service: move data from anywhere to anywhere
- Scalable from 1 KB/s to 100 GB/s, with consistent performance.
- Durable: sync data replication before ACK, for 7 days
- Secure: encryption in transit and rest, private network
- Managed: round-the-clock ops team

Exercise: Pub/Sub publishing and pulling using Cloud SDK

1. Install Google SDK Beta functions:

```
gcloud components install beta
```

2. Create a new Pub/Sub Topic

```
gcloud beta pubsub topics create test-topic
```

3. Create a new Pub/Sub Subscription

```
gcloud beta pubsub subscriptions create test-sub --topic=test-topic
```

4. Publish a message to Topic

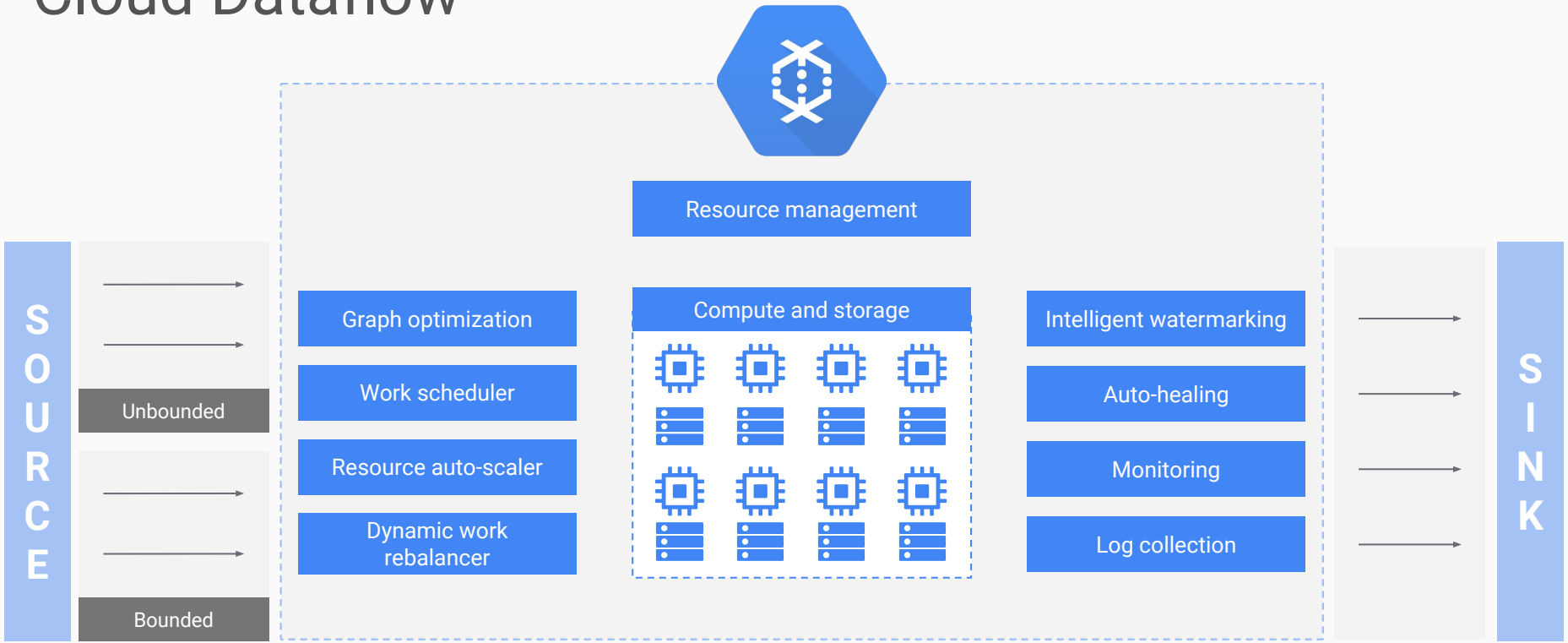
```
gcloud beta pubsub topics publish test-topic --message="Hello" --attribute
KEY1="Hi",KEY2="Guys"
```

5. Pull a message from topic

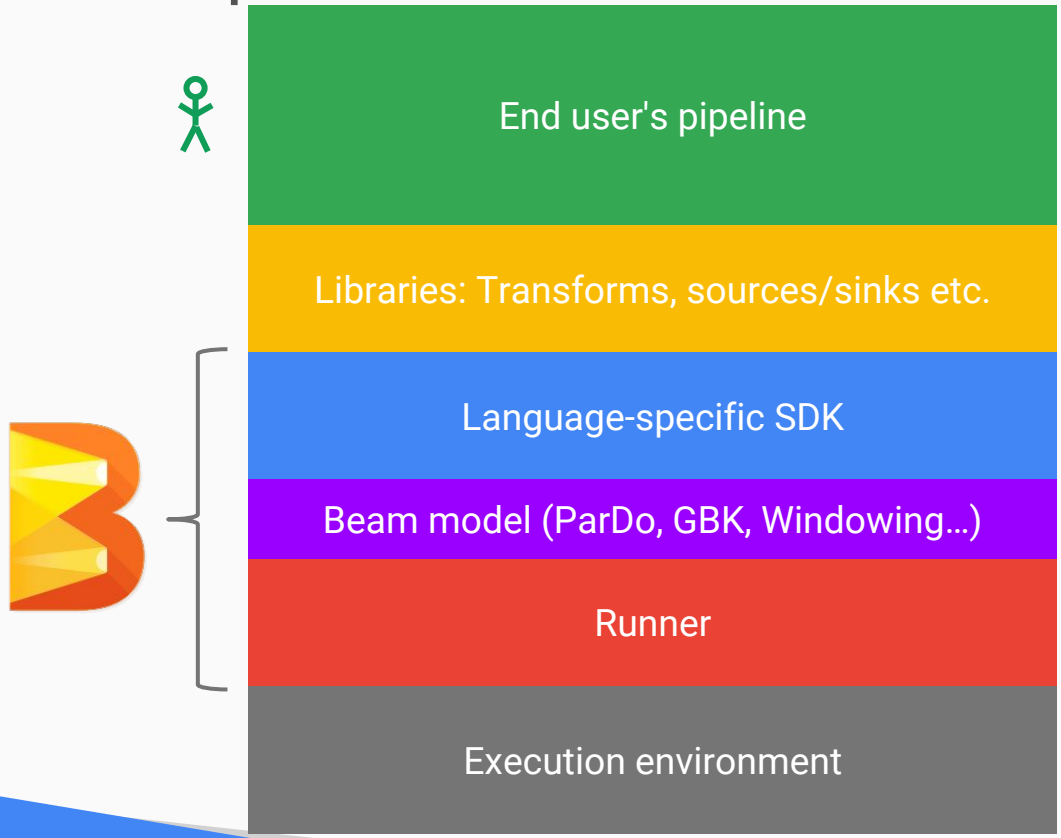
```
gcloud beta pubsub subscriptions pull
projects/<project-id>/subscriptions/test-sub --limit=5 --auto-ack
```

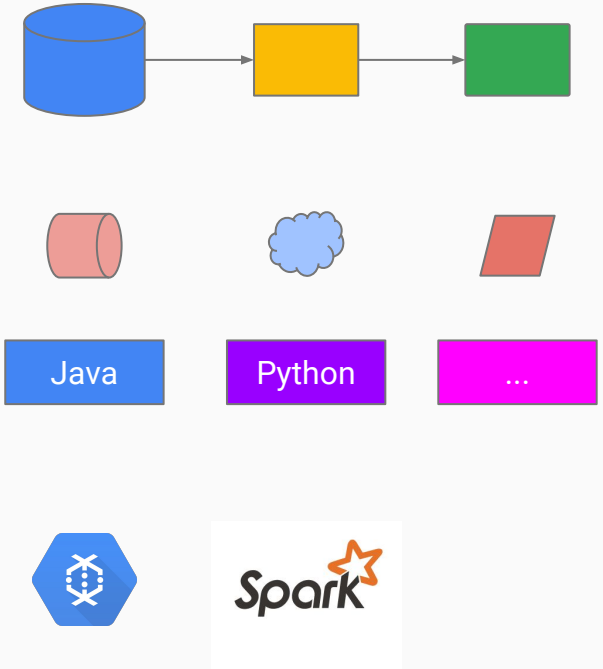
Dataflow Best Practices

Cloud Dataflow



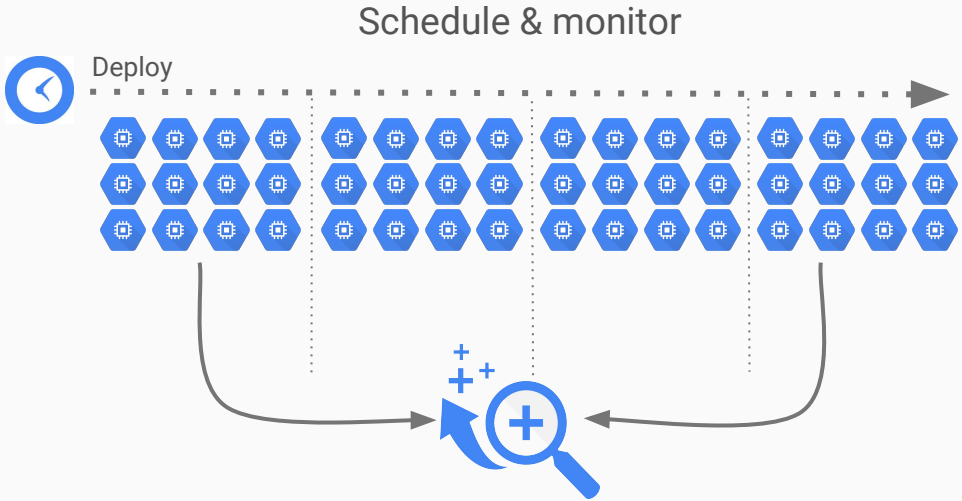
Developer's View





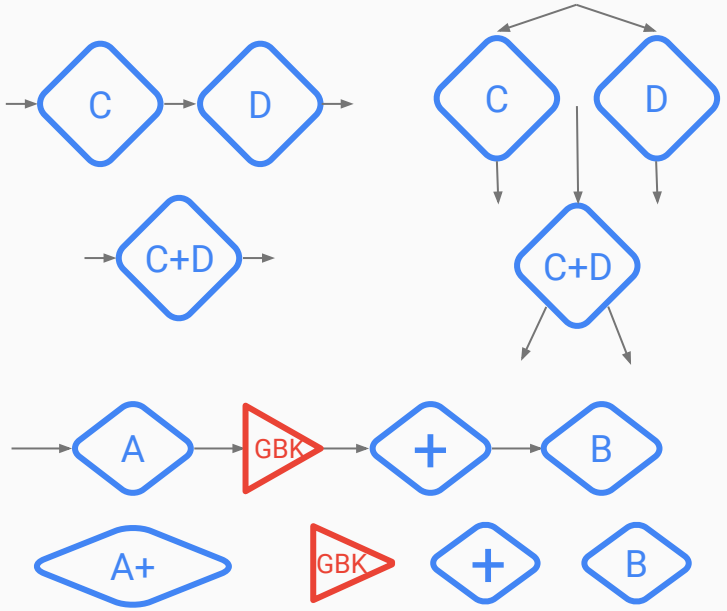
Why Customers Value Dataflow (Technical Perspective)

1. **Fully managed and auto-configured**
2. Auto-graph-optimized for best execution path
3. Auto-scaling mid-job
4. Dynamic work rebalancing mid-job



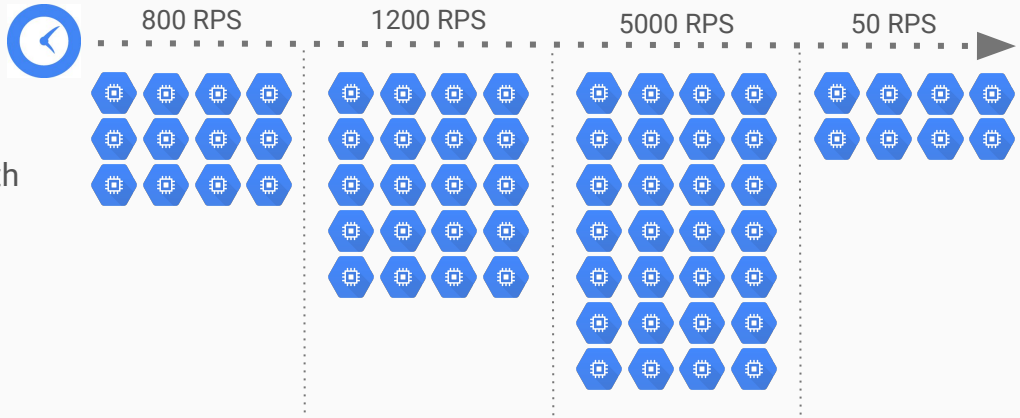
Why Customers Value Dataflow (Technical Perspective)

1. Fully-managed and auto-configured
2. **Auto-graph-optimized for best execution path**
3. Autoscaling mid-job
4. Dynamic work rebalancing mid-job



Why Customers Value Dataflow (Technical Perspective)

1. Fully-managed and auto-configured
2. Auto graph-optimized for best execution path
3. **Auto-scaling mid-job**
4. Dynamic work rebalancing mid-job



Why Customers Value Dataflow (Technical Perspective)

1. Fully-managed and auto-configured
2. Auto graph-optimized for best execution path
3. Autoscaling mid-job
4. **Dynamic work rebalancing mid-job**

Exercise: Dataflow Example

Before you begin, you'll need to download Maven (or, if you know your way around Java IDEs, feel free to try Eclipse)

https://goo.gl/NR58fT

Bonus: Dataflow Templates

- Navigate to Dataflow in the cloud console
- Create a Job from a Template
- Test the PubSub to GCS Text Template

Dataflow Best Practices

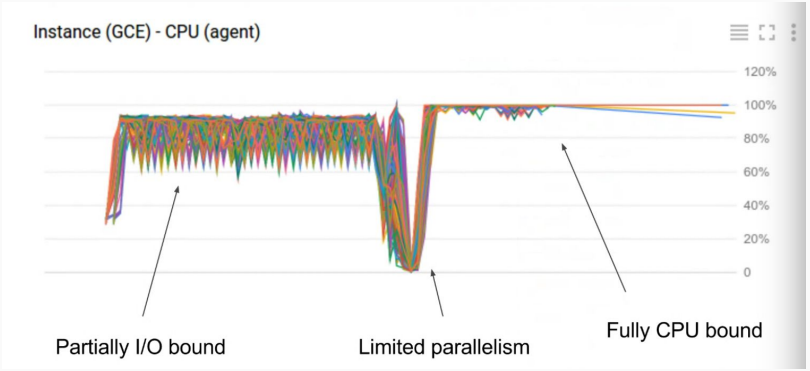
Think Performance: Step Timing

- Java is currently 2x faster than Python and is recommended for performance-critical environments.
- Use the step-timing counters in the UI to identify expensive ParDos. In many cases slowdowns can be due to unnecessary initialization or excessive logging in the ParDo apply(...) function.

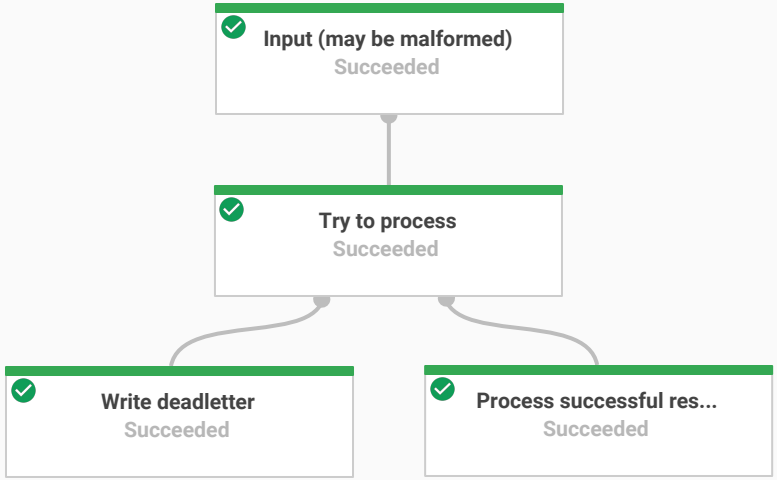
Dataflow Best Practices

Think Performance: Profiling

- Utilize the resource usage graphs in Stackdriver to profile a pipeline:
  - Low CPU utilization suggests an I/O bottleneck or limited parallelism.
  - I/O bottlenecks can be helped by switching to SSD and increasing the size of the disk.



Dataflow Best Practices



Redirect failed records during a load process by catching exceptions and exporting to a side output. This aids in diagnosing issues in large data loads without failing the entire load.

Bonus: Spotify/SCIO REPL

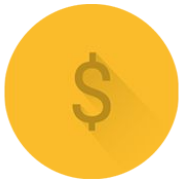
1. Found at https://github.com/spotify/scio
2. Based off the SCALA API
3. Has an REPL for interactive evaluation (or small compilations)
4. Convenient and easy to use

BigQuery

What is BigQuery?

- BigQuery is Google's cloud-based, enterprise data warehouse.
- BigQuery answers queries of very large databases, quickly.
- BigQuery scales to petabytes, but is cost-effective for any organization.
- BigQuery is fully managed. Customers benefit from almost NoOps.
- BigQuery supports the industry standards, such as SQL, ODBC, JDBC.

Recent BigQuery features added for large enterprises



Flat-rate pricing

Standard SQL

ODBC + JDBC connectors

Data manipulation language (DML) – beta

Stackdriver

Identity access and management (IAM)

For more read the “BigQuery for enterprise” blog post.

Why use BigQuery?

- Speed, scale, agility
  - Ask questions over petabytes of data—make decisions at “transformationally fast speed.”
  - Go very big—scale up to tens of petabytes or down to gigabytes and back—no problem.
  - BigQuery’s speed and scale make it suitable for huge data lakes or small data marts.
- Enterprise ready to be a mission-critical part of our customers’ businesses.
- Managed services and cutting edge technology.
  - Google does all the maintenance and monitoring.
  - Google is your partner who will keep you on the cutting edge of technology.
- Input up to 100,000 rows per second into BigQuery...the new data is available for querying instantly.
- Ad hoc queries:
  - No database maintenance tasks. Just ask questions as they come to you.

Why use BigQuery? (continued)

Things that BigQuery doesn’t have to do to run a query over a petabyte:

- Call anyone to warn them.
- Buy any extra resources.
- Create indexing or design keys beforehand.

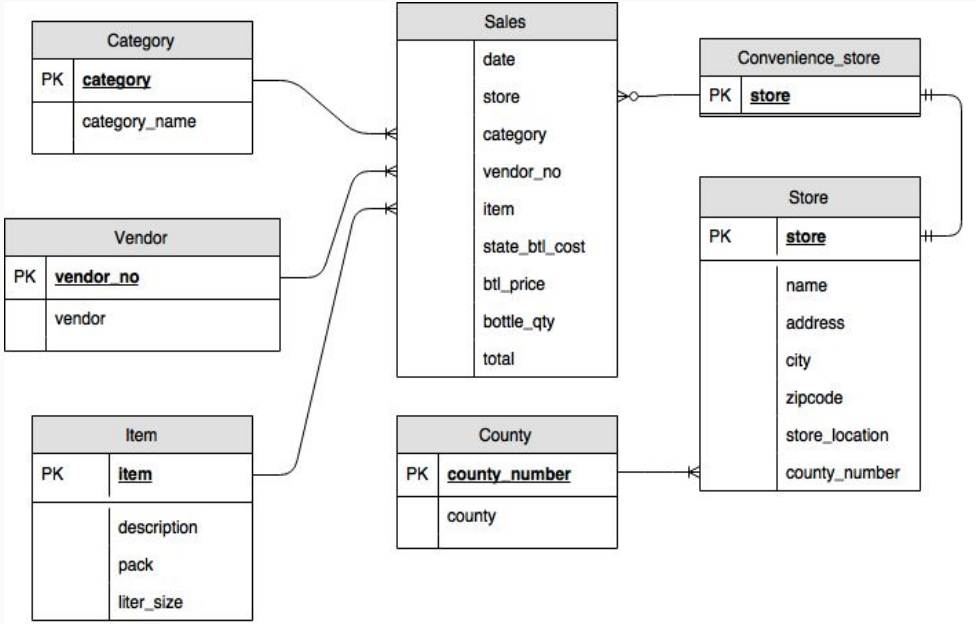
Design Recommendations

- Design the schema based on the need to query it
- Write facts and dimension data into a flat table structure wherever possible
- Limit DML requirements in table design

Design Watchpoints

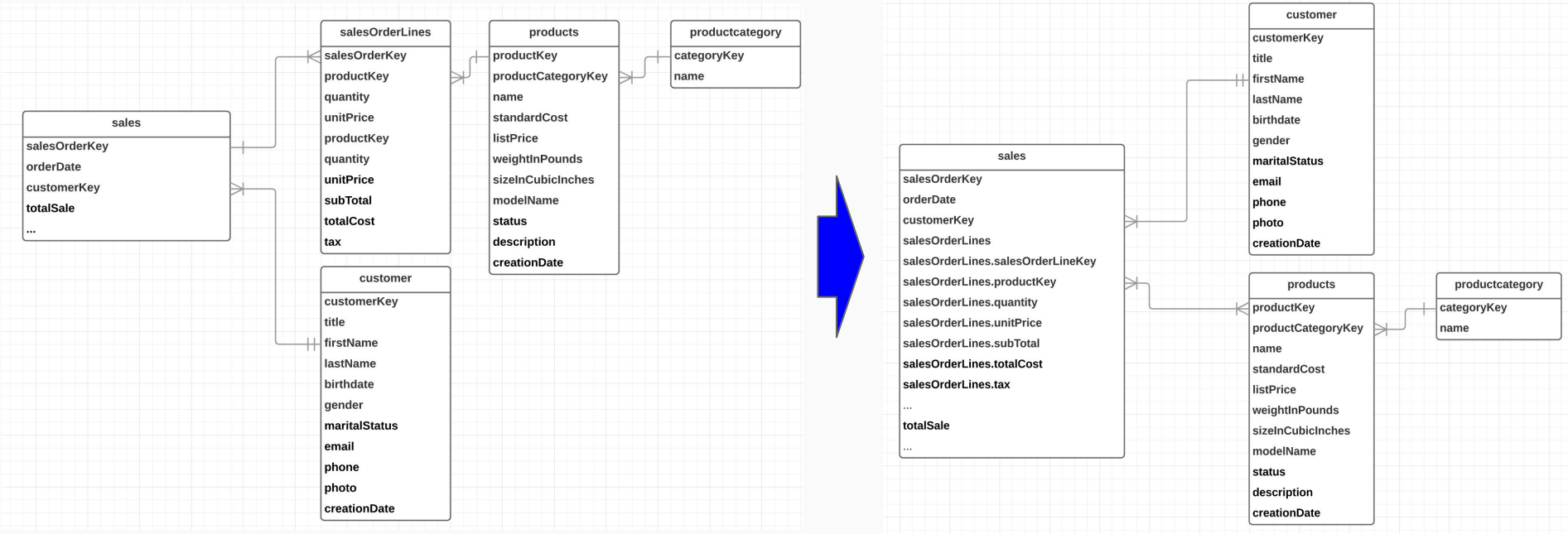
- Pick the right tool for the job (BigQuery vs. DataFlow/DataProc)
- Performing joins on larger dimension tables may impact performance
- Repeated fields may be difficult to work with beyond 1st or 2nd level of nesting

Denormalization Example



iowa_sales_denorm
date
cstore
store
name
address
city
zipcode
store_location
county_number
county
category
category_name
vendor_no
vendor
item
description
pack
liter_size
state_btl_cost
btw_price
bottle_qty
total

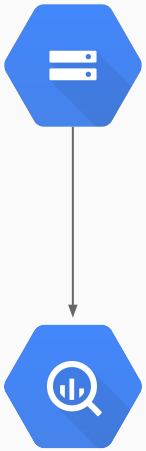
Denormalization Example



Google Cloud

BigQuery Data Loading

Data Loading Tips



Decompress files before loading into BigQuery, if performance is important. Highly parallel load operations allow uncompressed files to load significantly faster than compressed files.

Load data into BigQuery from AVRO formatted files. This will result in a 10x performance gain over JSON or CSV, and since the schema is already embedded, there are no serialization headaches.

Exercise: Load Data Into BigQuery

Let's try loading some data!

https://goo.gl/a5FVSK

Loading using Cloud SDK

```
bq query --destination_table=mydataset.happyhalloween "SELECT name,count FROM
mydataset.babynames WHERE gender = 'M' ORDER BY count DESC LIMIT 6"
```

Short Exercise: Visualizing the Data

Instead of querying the data, let's do something more:

1. Go to https://datastudio.google.com
2. Create a new report
3. Add the table from the exercise as a data source
4. Place a pie chart and a bar graph

Example of a Delete Query in BigQuery

```
bq query --use_legacy_sql=false "delete from
`<project>.<dataset>.<table>` where 1=1"
```

Catch: Doesn’t work on Partitioned Tables

BigQuery Tips & Tricks

Query Performance Tips

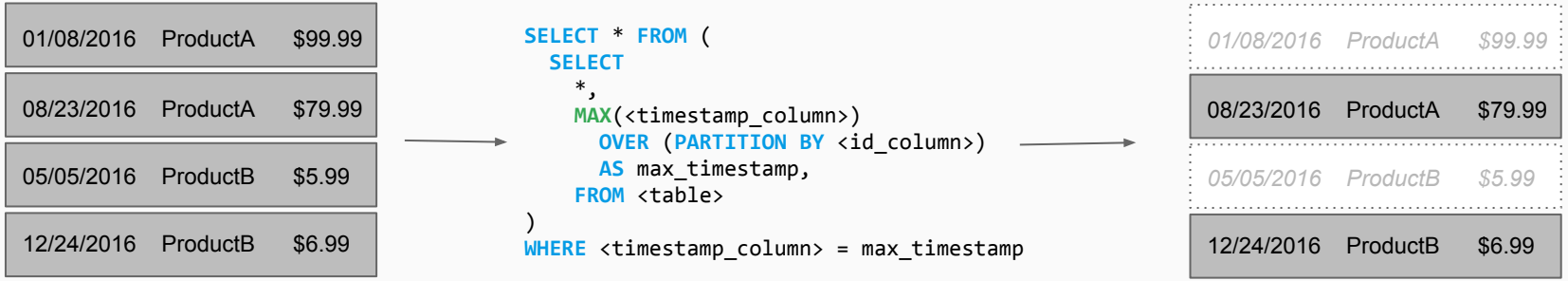
- Flatten normalized tables for super-fast querying.
  - No performance penalty for sparse or duplicate data.
  - Trade JOINS for column scans because storage is more performant and cheaper than compute resources.
  - Use nested repeated schema to simulate normalization benefits (e.g., order containing line items).

Tips to Saving $$$

- Partition your data by date and use partition elimination when querying.
- Query only what you need. Do you REALLY need 1 year's worth of results?
- For commonly used data, create daily aggregates. (e.g., gaming use case: aggregate event tables for player-day combinations)
- Avoid SELECT * LIMIT 1 queries. Instead use the preview tab in the UI, bq head, or tabledata.list() API to get the schema for a table.

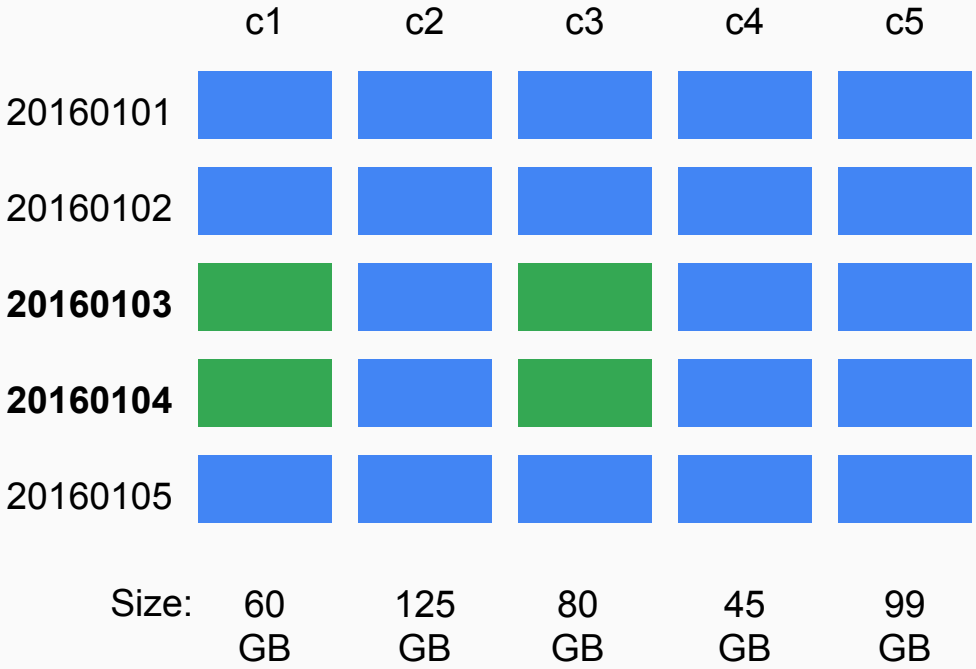
Flattening Duplicate Records

When using an append-only update strategy, utilize WINDOW functions to flatten records and retrieve only the latest versions for the table based on timestamp.



Partition Tables

Table partitioning allows you to efficiently query over the parts of the table you want and cut costs on the data read.



Template Tables

Use template tables to let BigQuery create sharded tables for you on ingestion from the BigQuery API or bq command-line tool.

`<targeted_table_name> + <template_suffix>`

Short Exercise: Table Suffixes Example

- Take a look at the public data set (**bigquery-public-data)** called **noaa_gsod**
- Imagine the incredibly long query used to join all the tables that fall in the 19** range
- Try something like this instead:

```
#standardSQL
SELECT max, ROUND((max-32)*5/9,1) celsius, mo, da, year
FROM
  `bigquery-public-data.noaa_gsod.gsod19*`
WHERE
  max != 9999.9 # code for missing data
AND _TABLE_SUFFIX BETWEEN '29' AND '40'
ORDER BY max DESC
```

Temporary Tables

Utilize a separate dataset with a default table expiration for creating temp tables in BigQuery. Any new table created in the dataset will be automatically deleted the specified number of days after creation.