Google Cloud

Big Data and Machine Learning Fundamentals with Google Cloud Platform

An explosion of data

“By 2020, some 50 billion smart devices will be connected, along with additional billions of smart sensors, ensuring that the **global supply of data will continue to more than double every two years**”

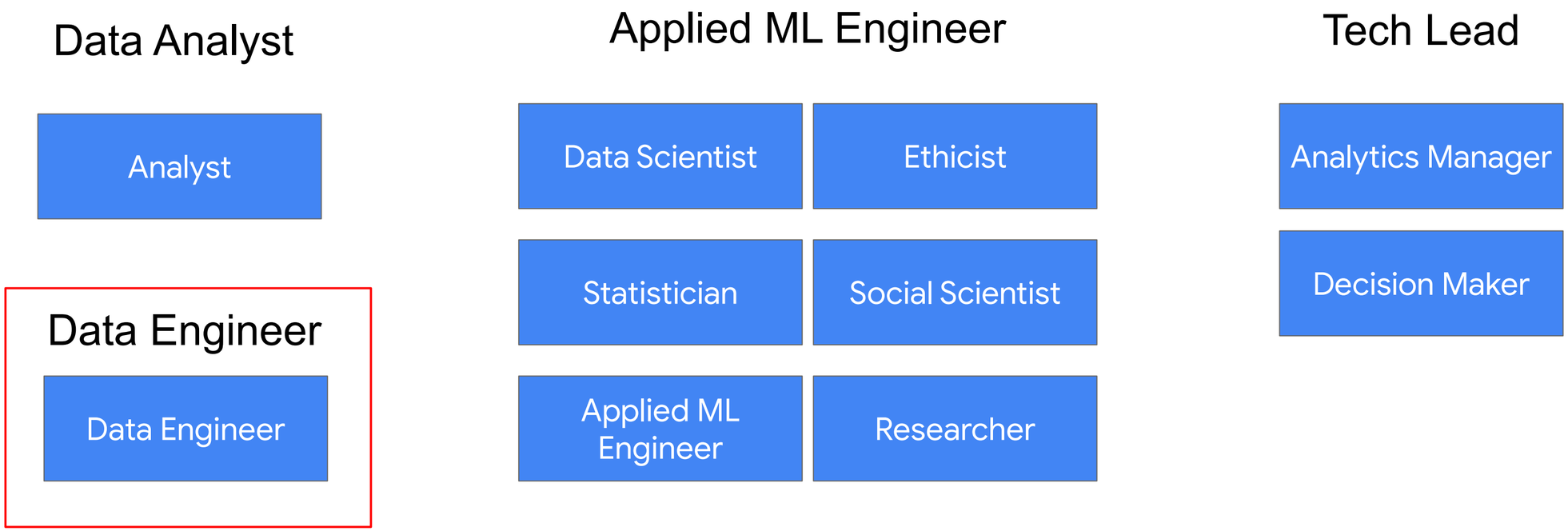
https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/straight-talk-about-big-data

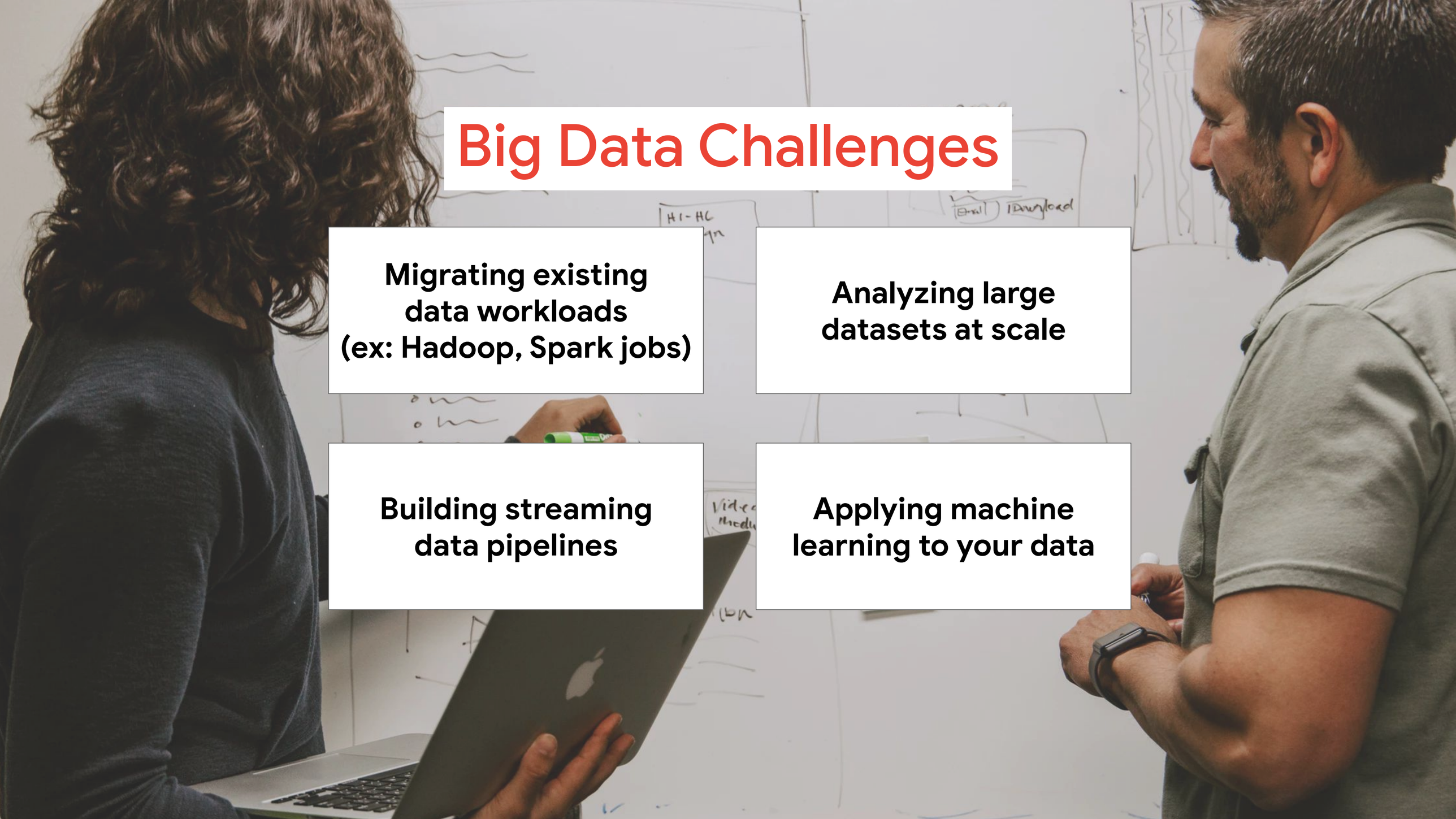
An explosion of data

… and only about 1% of the data generated
today is actually analyzed

https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/straight-talk-about-big-data

There is a great demand for data skills





Big Data Challenges

Migrating existing data workloads (ex: Hadoop, Spark jobs)

Analyzing large datasets at scale

Building streaming data pipelines

Applying machine learning to your data

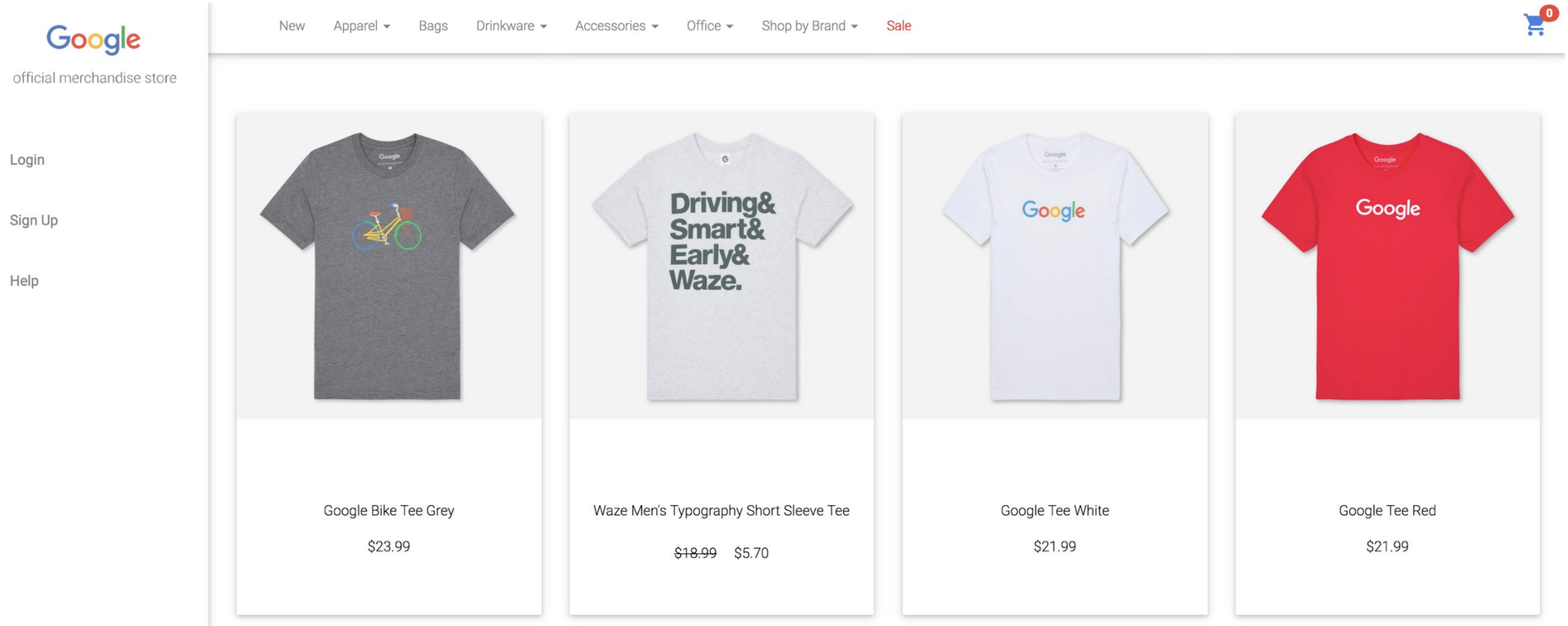
Product recommendations using Cloud SQL and Spark

What housing rentals should I recommend to my customers based on their history?

Cloud SQL

Cloud Dataproc

Classify returning customers with BigQuery ML

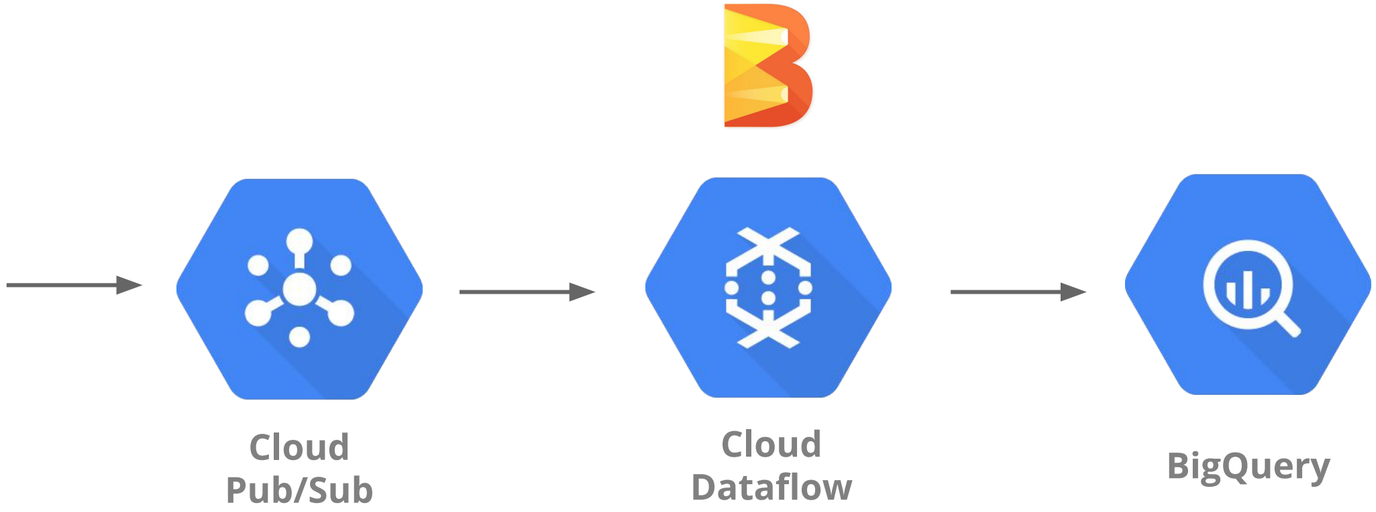


Which ecommerce customers are likely to return and purchase?

BigQuery

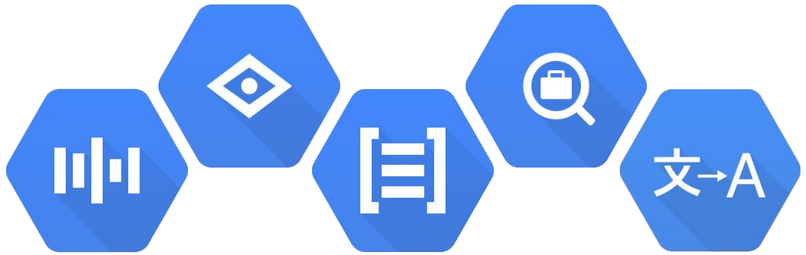
Real-time dashboards with Pub/Sub, Dataflow, and Data Studio

How can I monitor streaming insights from my business?



Classify images with ML two-ways using pre-built models

How can I leverage pre-trained models for image classification? What about my own image datasets?



ML APIs

Cloud AutoML

Agenda

Google Cloud Platform infrastructure

- Compute
- Storage
- Networking
- Security

Big data and ML products

- Google innovation timeline
- Choosing the right approach

What you can do with GCP

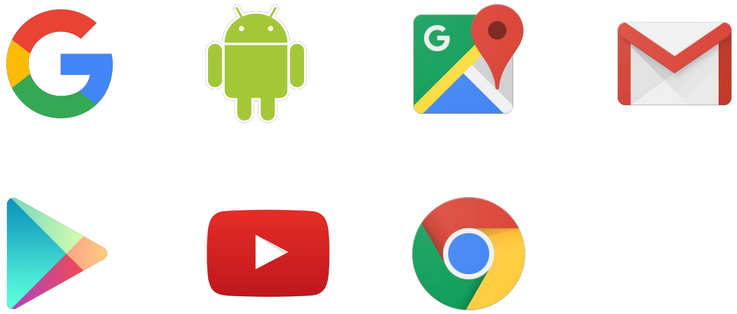
Activity: Explore a customer use case

The different data roles in an organization

Google's mission

Organize the world's information and make it universally accessible and useful.

Seven cloud products with
one billion users



Google Cloud

Google Cloud

Big Data and ML Products

Compute Power | Storage | Networking

Security

Google Cloud

Big Data and ML Products

Compute Power

Storage

Networking

Security

Machine Learning Models require significant compute resources

Shown: Automatic **Video Stabilization** for Google Photos

Data sources:

1. Image frames (stills from video)
2. Phone gyroscope
3. Lens motion

A single high-res image represents millions of data points to learn

8 Megapixel resolution

3264 (w) x 2448 (h) x 3 (RGB) =

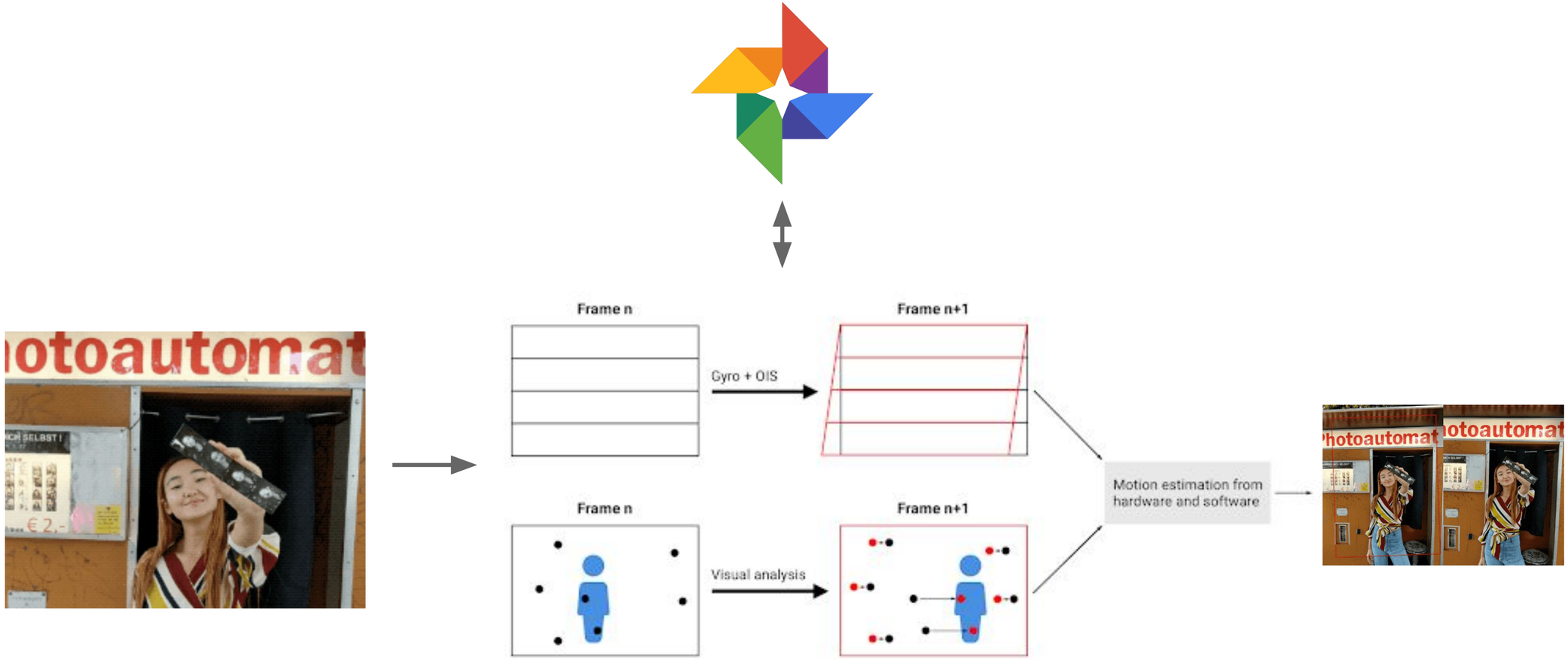
23,970,816
data points per image*

* More data = longer model training times + more storage needed

How many photos are uploaded daily to Google Photos?

How many hours of video are uploaded every minute to YouTube?

Google trains on its infrastructure and deploys ML to phone hardware



Leverage Google's AI research with pre-trained AI building blocks

Sight

- Cloud Vision
- Cloud Video Intelligence
- AutoML Vision

Language

- Cloud Translation
- Cloud Natural Language
- AutoML Translation
- AutoML Natural Language

Conversation

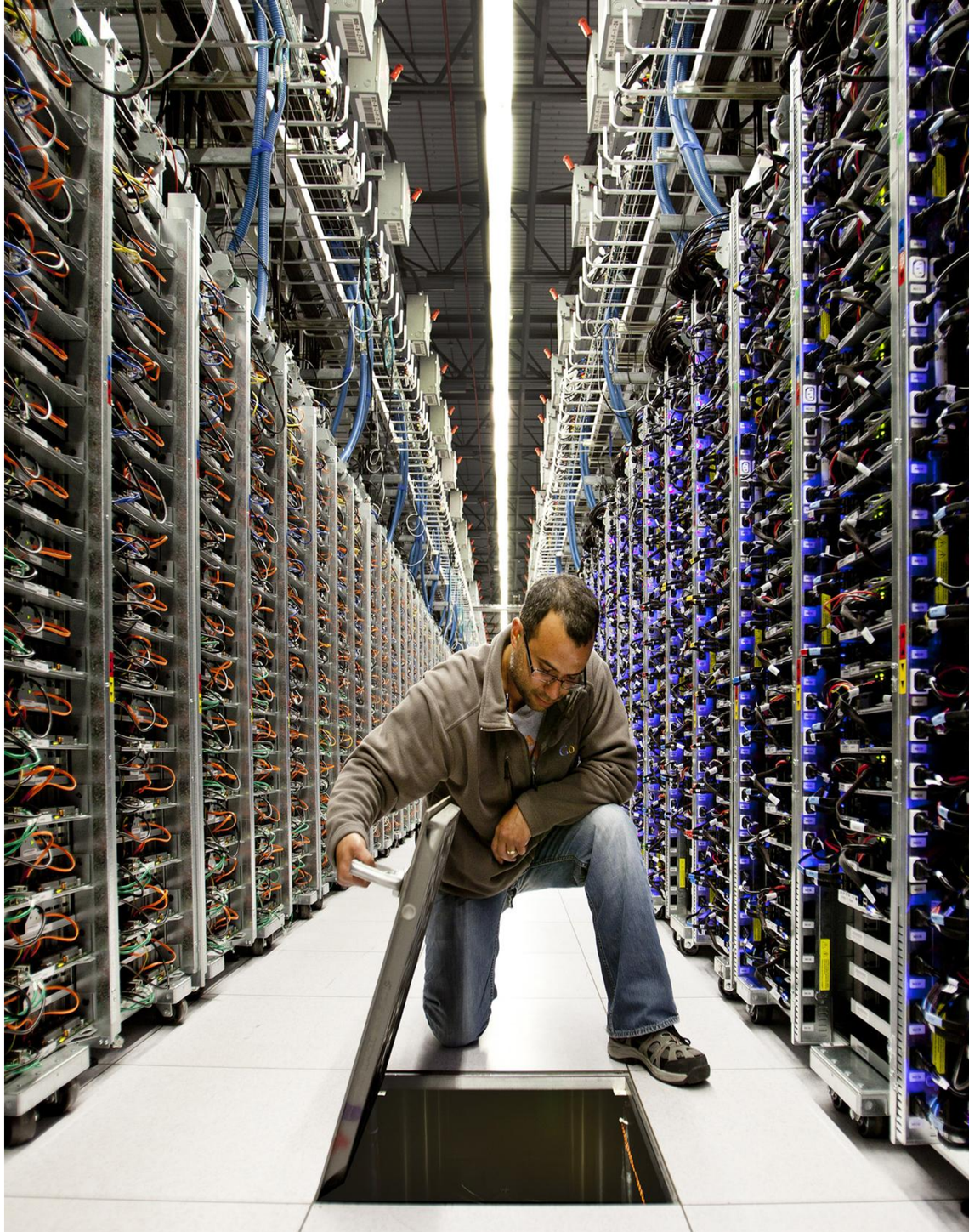
- Dialogflow Enterprise Edition
- Cloud Text-to-Speech
- Cloud Speech-to-Text

https://cloud.google.com/video-intelligence/

Build on Google infrastructure

This is what makes Google Google: its physical network, its thousands of fiber miles, and those many thousands of servers that, in aggregate, add up to the **mother of all clouds**.”

- Wired

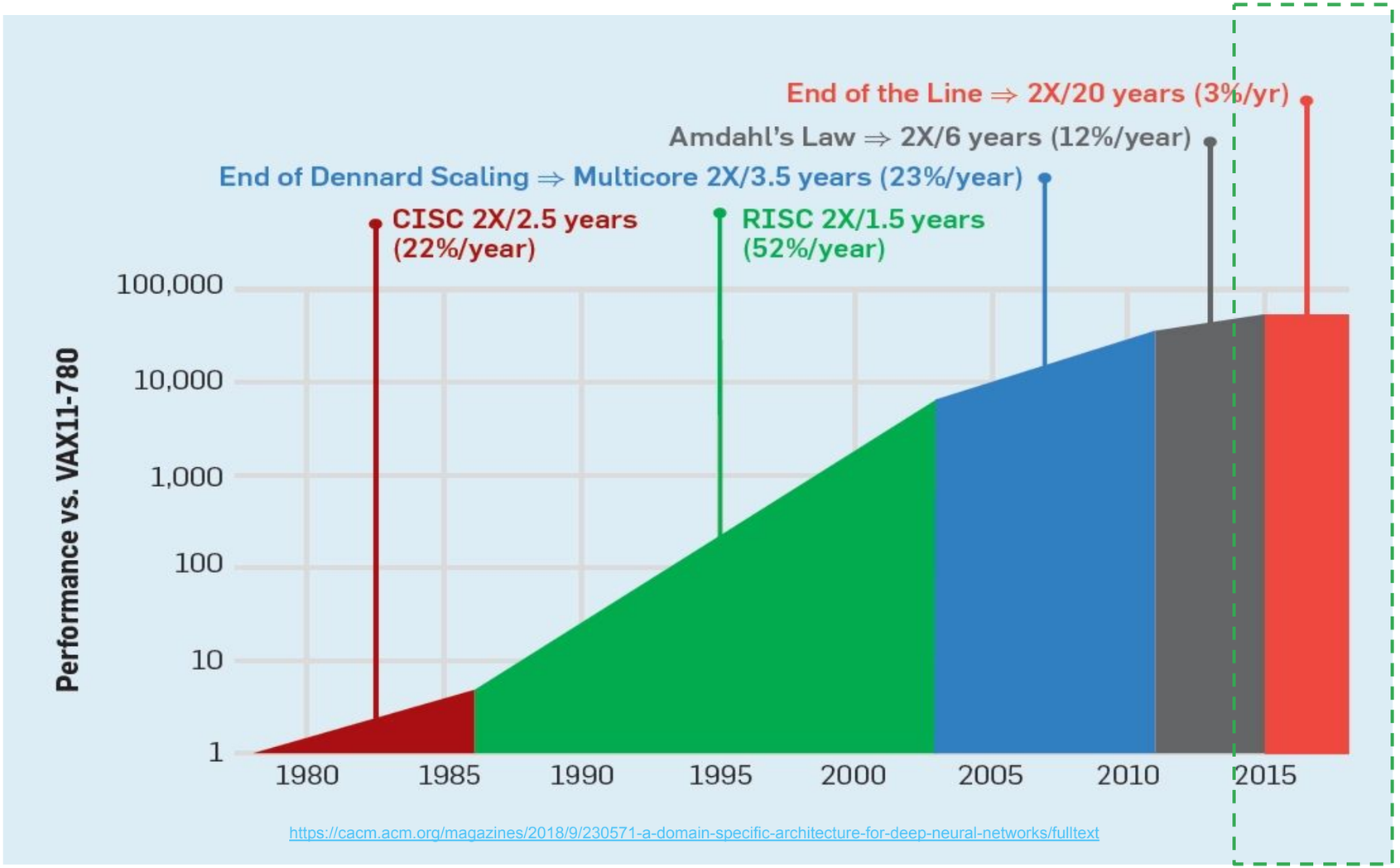


Simply scaling the raw number of servers in Google's data centers isn't enough.

"If everyone spoke to their phone for 3 minutes, we'd exhaust all available computing resources"

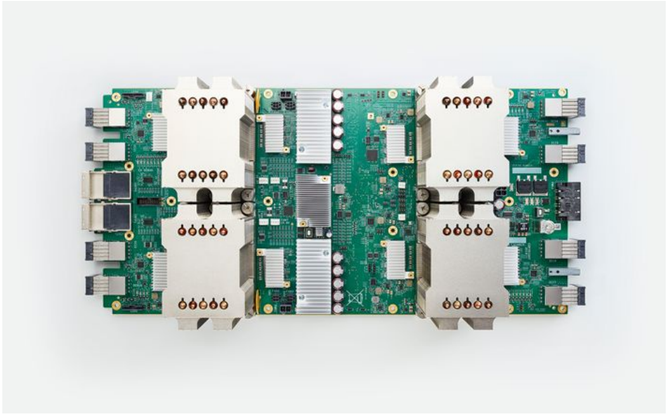
— Jeff Dean, 2014

Will Moore's Law save us?

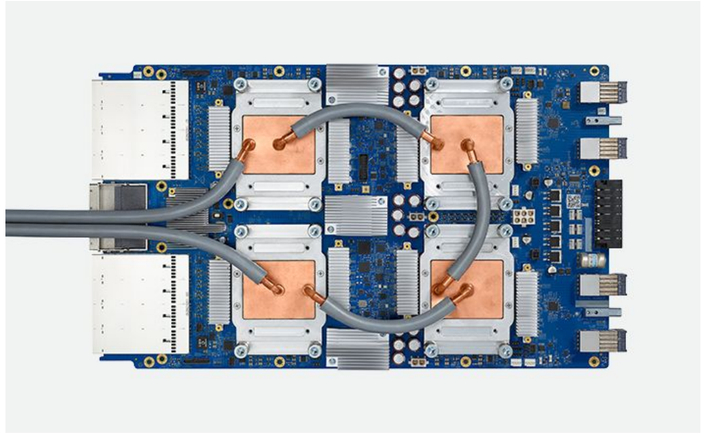


https://cacm.acm.org/magazines/2018/9/230571-a-domain-specific-architecture-for-deep-neural-networks/fulltext

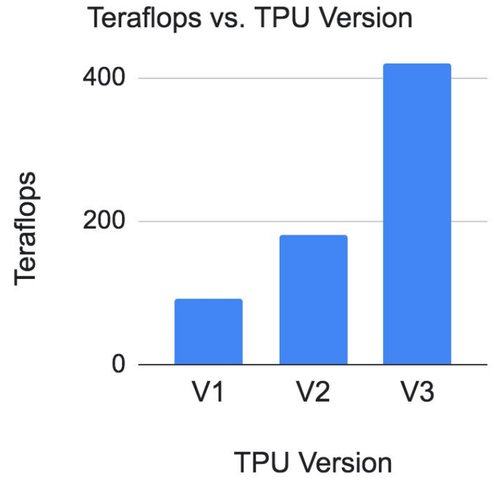
Tensor Processing Units (TPUs) are specialized ML hardware



Cloud TPU v2
180 teraflops
64-GB High Bandwidth Memory (HBM



Cloud TPU v3
420 teraflops
128-GB HBM

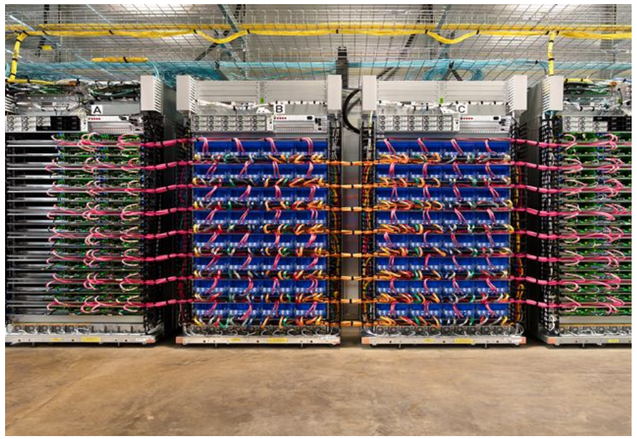


Google Cloud

TPUs enable faster models and more iterations

ebay™

+



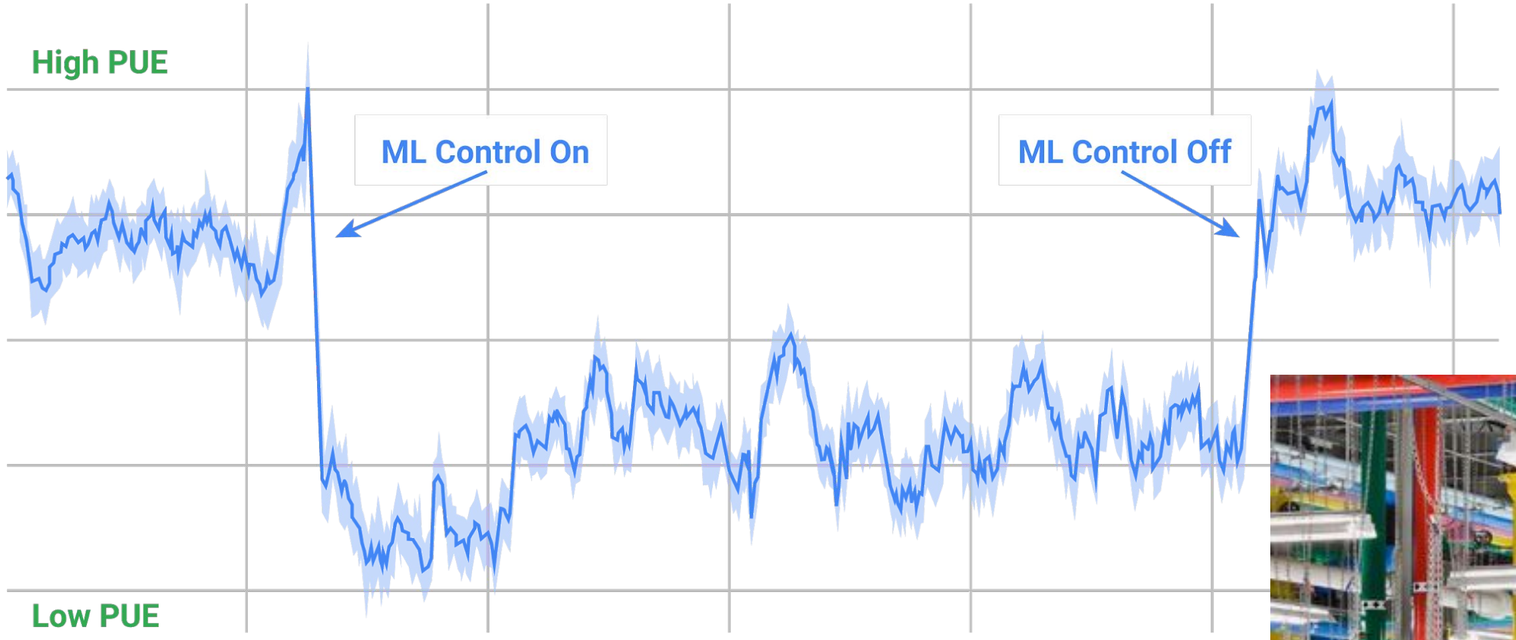
"**Cloud TPU Pods** have transformed our approach to visual shopping by delivering a **10X speedup** over our previous infrastructure.

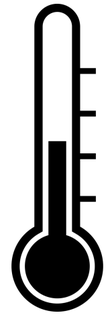
We used to spend months training a single image recognition model, whereas now we can train much more accurate models in a few days on Cloud TPU Pods.

— **Larry Colagiovanni**
VP of New Product Development

Google Cloud

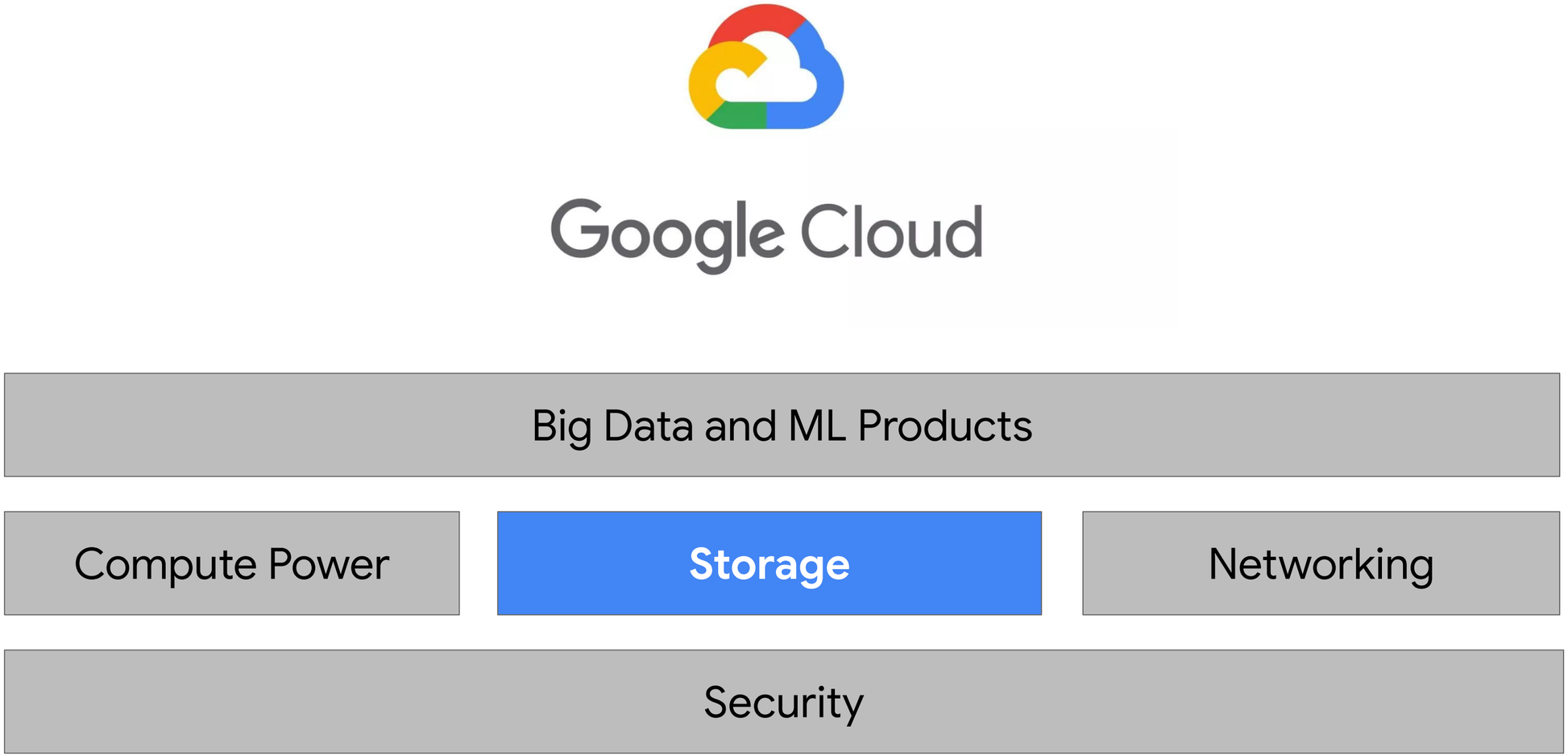
Google saved it's own data center cooling energy by 40%
improved Power Usage Effectiveness (PUE) by 15%





https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/

Google Cloud


Google Cloud
Big Data and ML Products
Compute Power
Storage
Networking
Security

1.2 billion photos and videos are uploaded to Google Photos every day.

Total size of over **13 PB** of photo data.

1PB or 400 hours of video uploaded **every minute**

Leverage Google's 99.999999999% durability storage

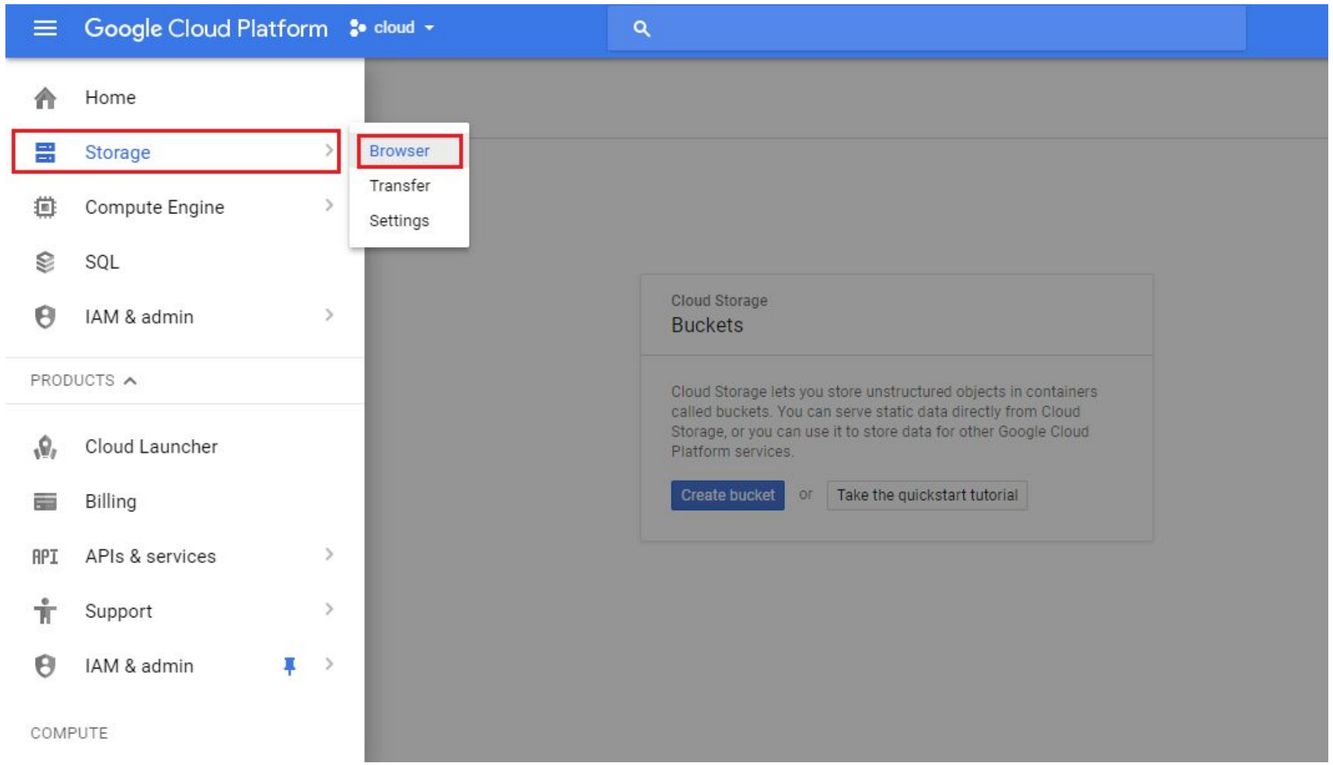
Cloud Storage

Spotify · vimeo · Coca-Cola · WiX.com · BROAD INSTITUTE · CLOUDIAN · EVERNOTE · motorola · PHILIPS · zulily

Google Cloud

Creating a Cloud Storage bucket for your data is easy

UI



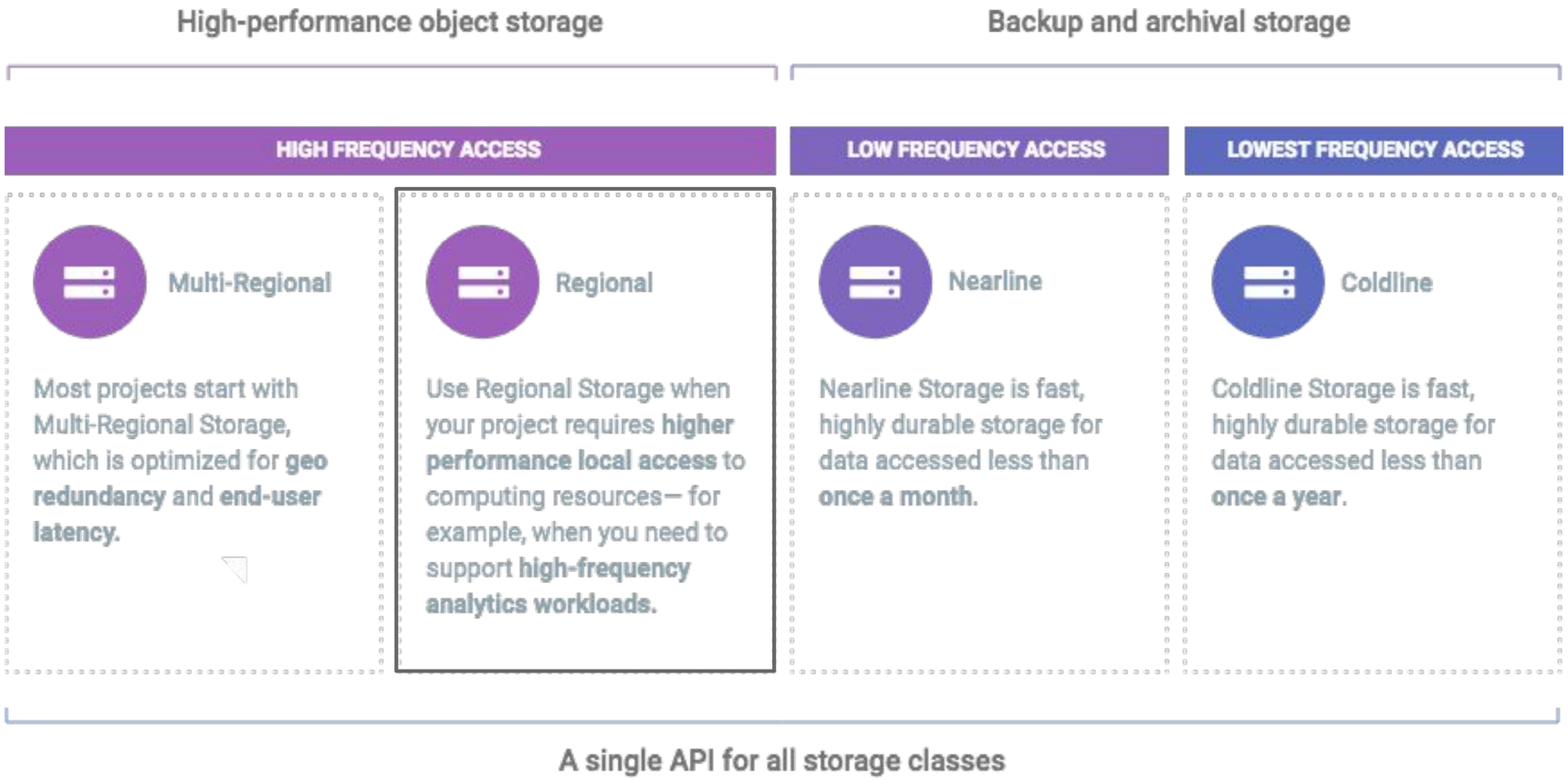
Cloud Storage

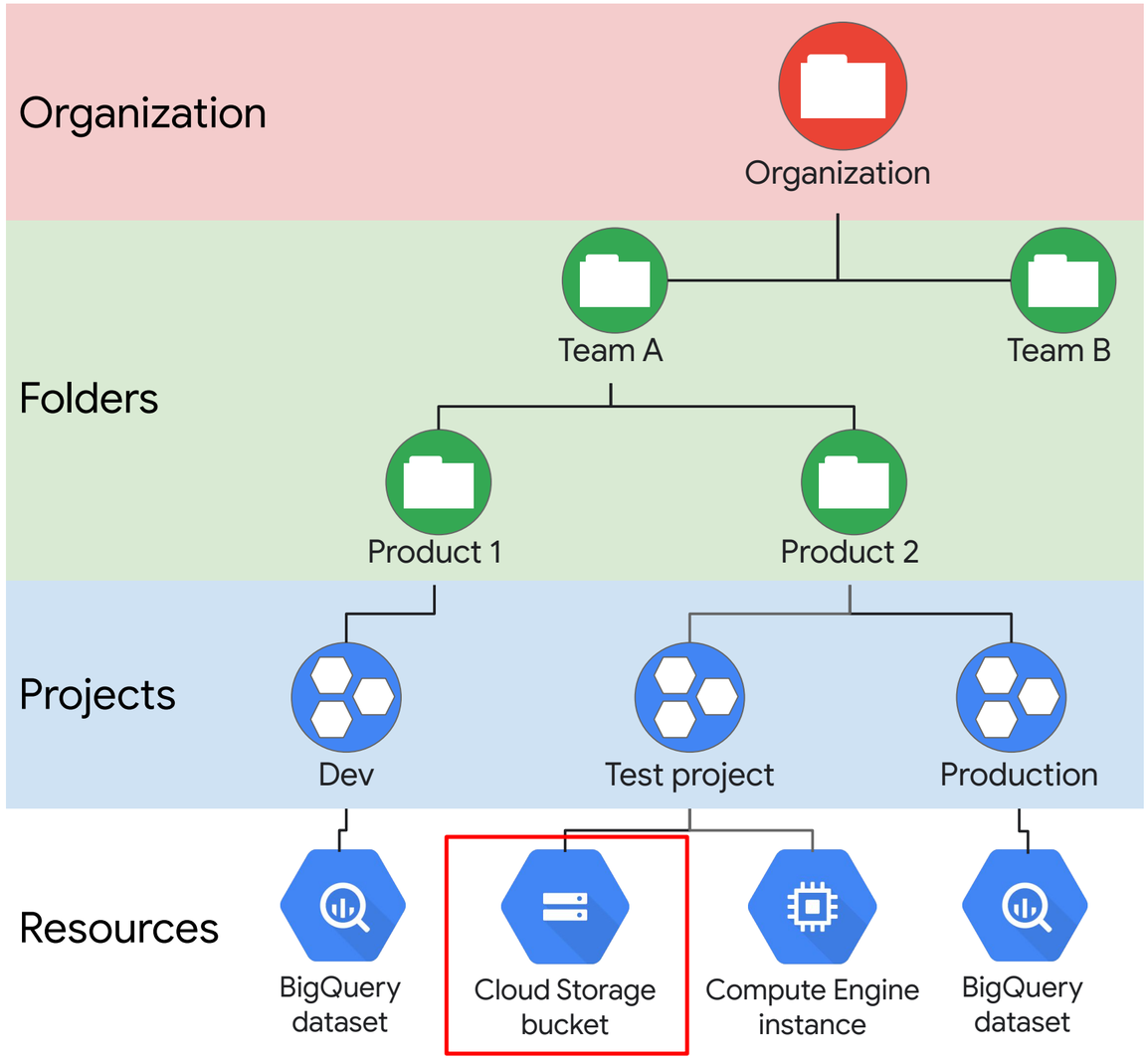
CLI

```
gsutil mb -p [PROJECT_NAME] -c [STORAGE_CLASS]
-l [BUCKET_LOCATION] gs://[BUCKET_NAME]/
```

Google Cloud

Typical big data analytics workloads run in Regional Storage

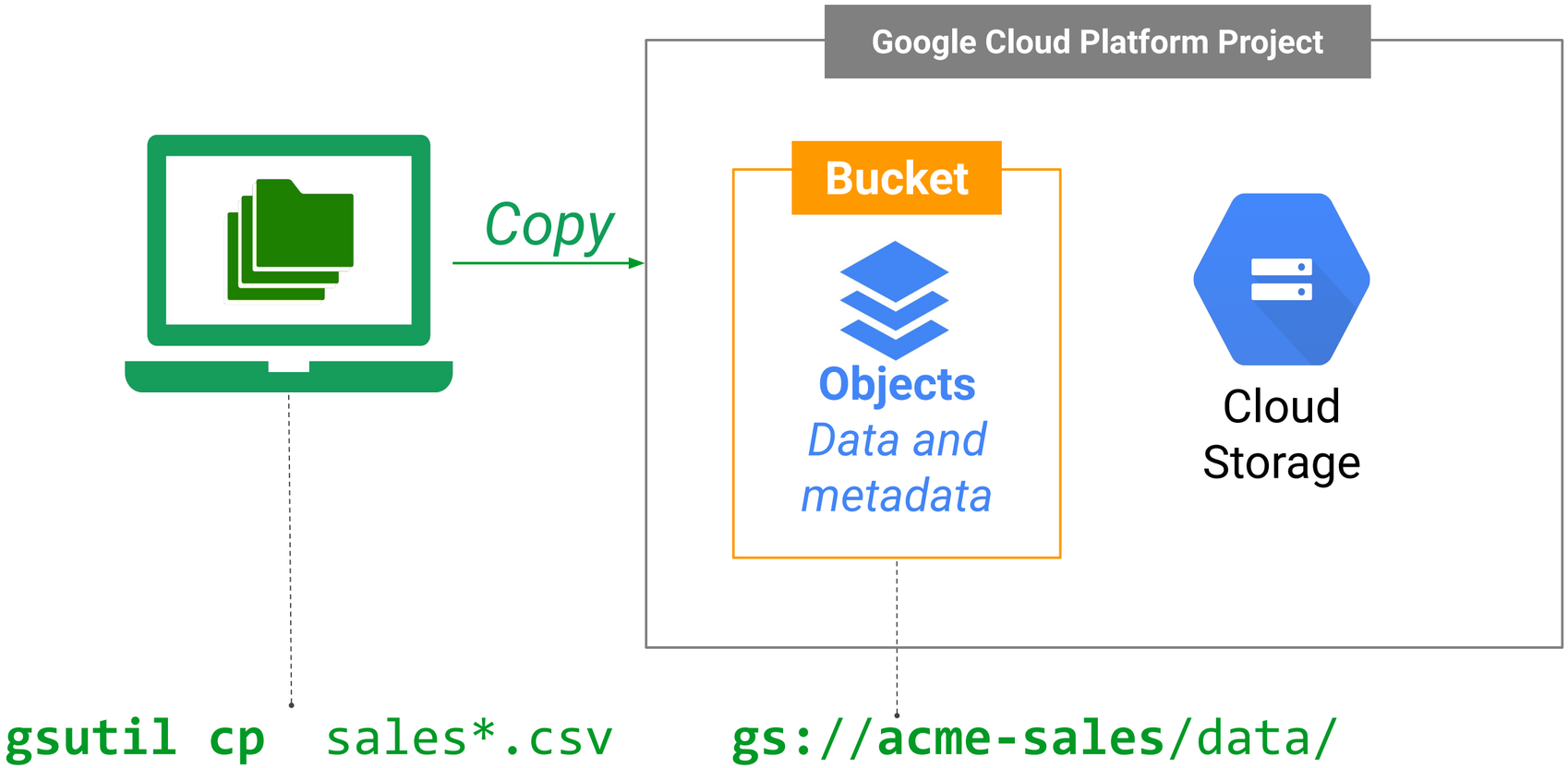




Cloud Storage buckets are one of many resources of the Google Cloud Platform

You can collaborate with many other teams in your organization across many projects

Got data? Quickly migrate your data to the cloud using **gsutil** tool



gsutil = google storage utility, cp = copy

Google Cloud

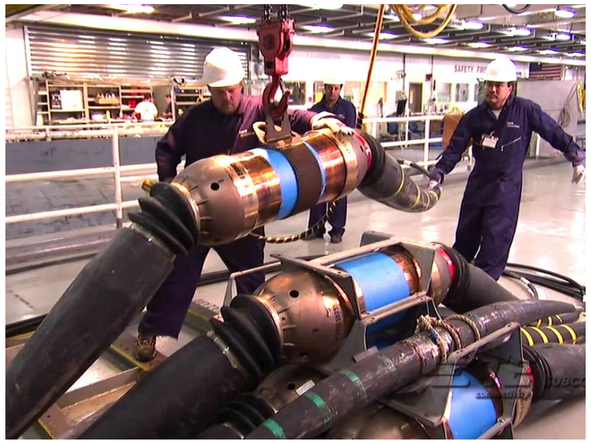
Google Cloud

Big Data and ML Products

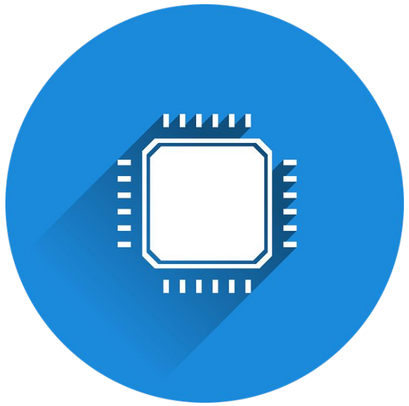
Compute Power | Storage | **Networking**

Security

Google’s private network carries as much as 40% of the world’s internet traffic every day

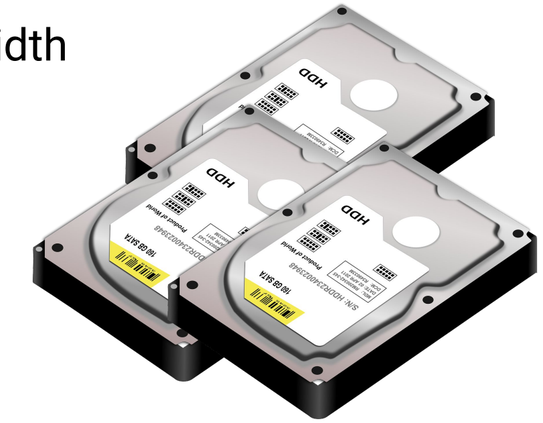


Google's data center network speed enables the separation of compute and storage



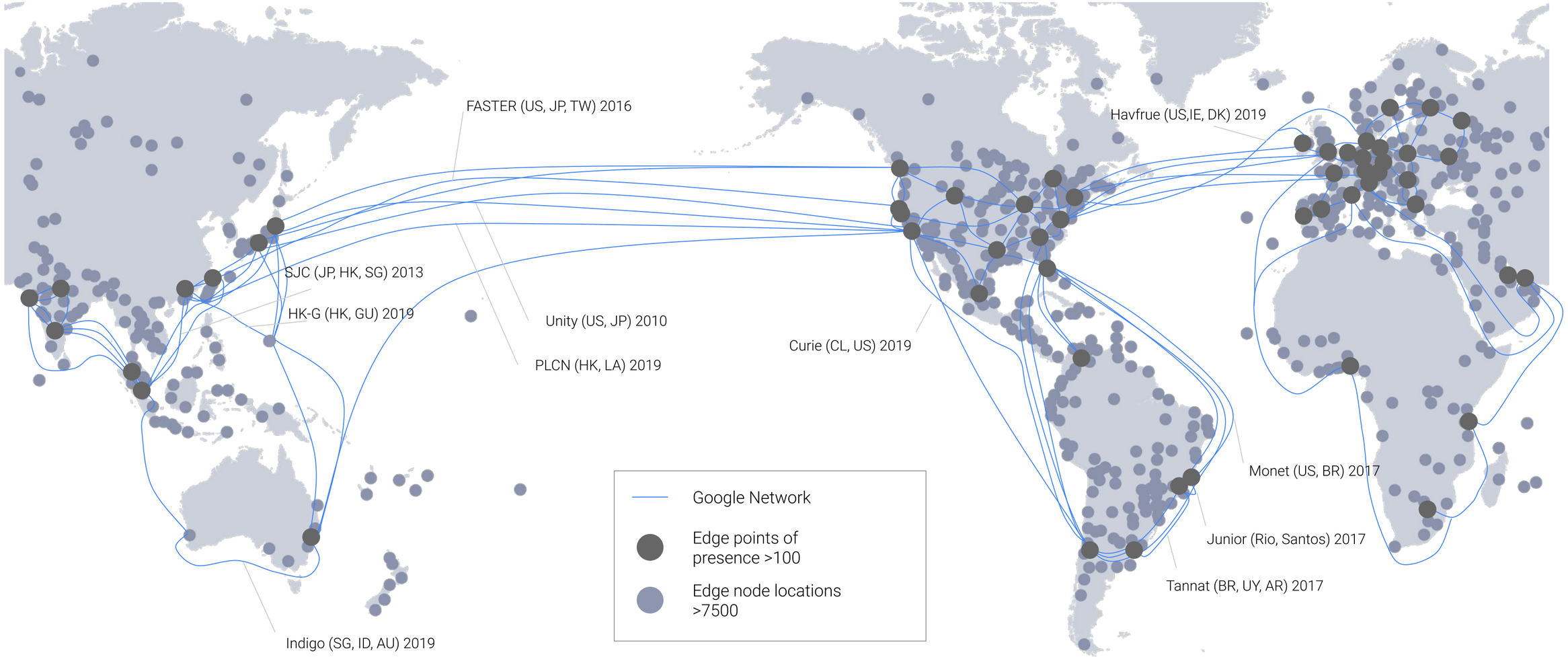
Servers doing compute tasks don't need to have the data on their disks

1 Petabit/sec of total bisection bandwidth



Data can be "shuffled" between compute workers at over 10GBs

Google’s cable network spans the globe



Big Data and ML Products

Compute Power | Storage | Networking

Security

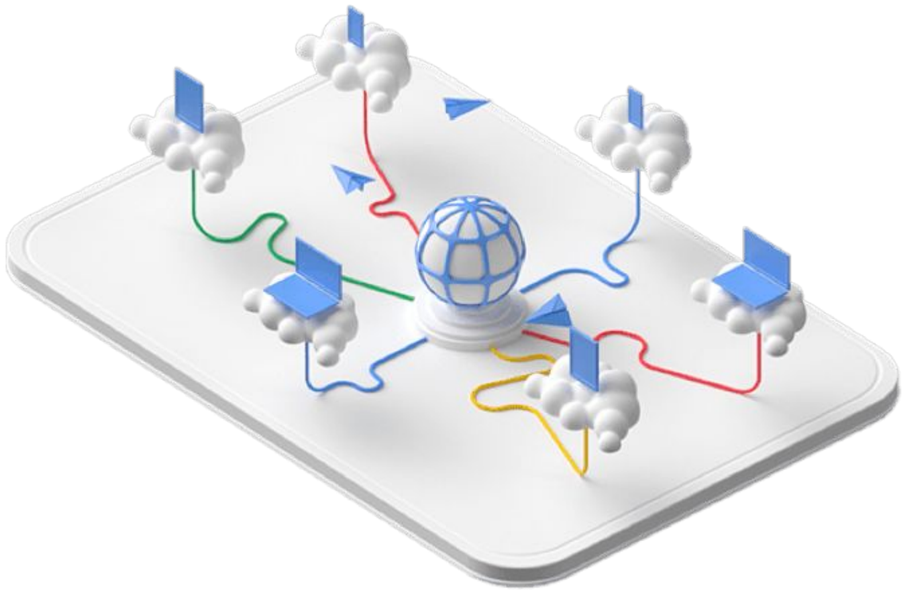
On-premise → you manage all security layers

Responsibility	On-premises
Content	
Access policies	
Usage	
Deployment	
Web app security	
Identity	
Operations	
Access and authentication	
Network security	
OS, data, and content	
Audit logging	
Network	
Storage and encryption	
Hardware	

Google Cloud Platform offers fully-managed services

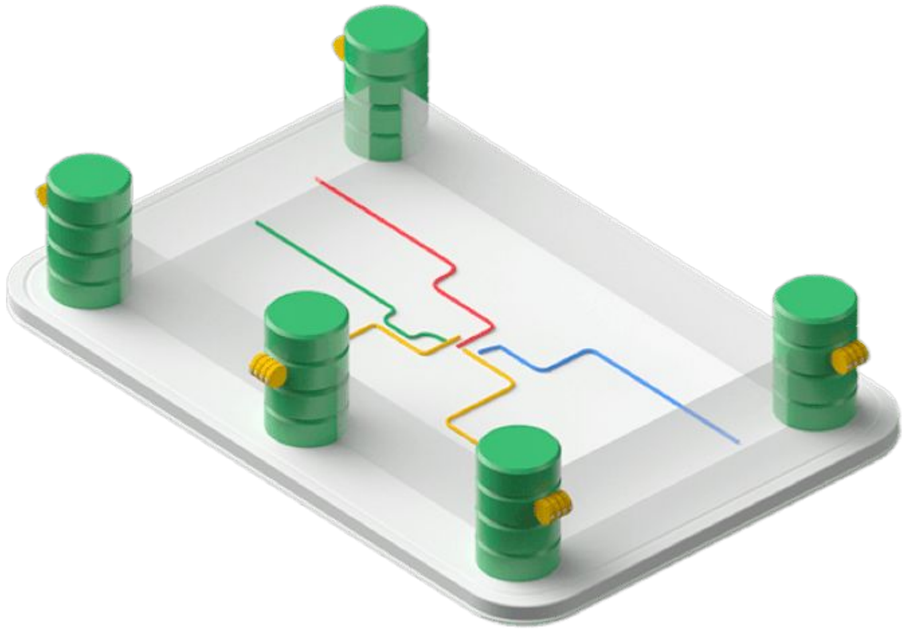
Google Cloud

Communications to Google Cloud are encrypted in transit



- In-transit encryption
- Multiple layers of security
- Backed by Google security teams 24/7

Stored data is encrypted at rest and distributed



- Data automatically encrypted at rest
- Distributed for availability and reliability

Spotlight: BigQuery granular control over data access

- BigQuery table data encrypted with keys (and those keys are also encrypted)
- Monitor and flag queries for anomalous behavior
- Limit data access with authorized views

Agenda

Google Cloud Platform infrastructure

- Compute
- Storage
- Networking
- Security

Big data and ML products

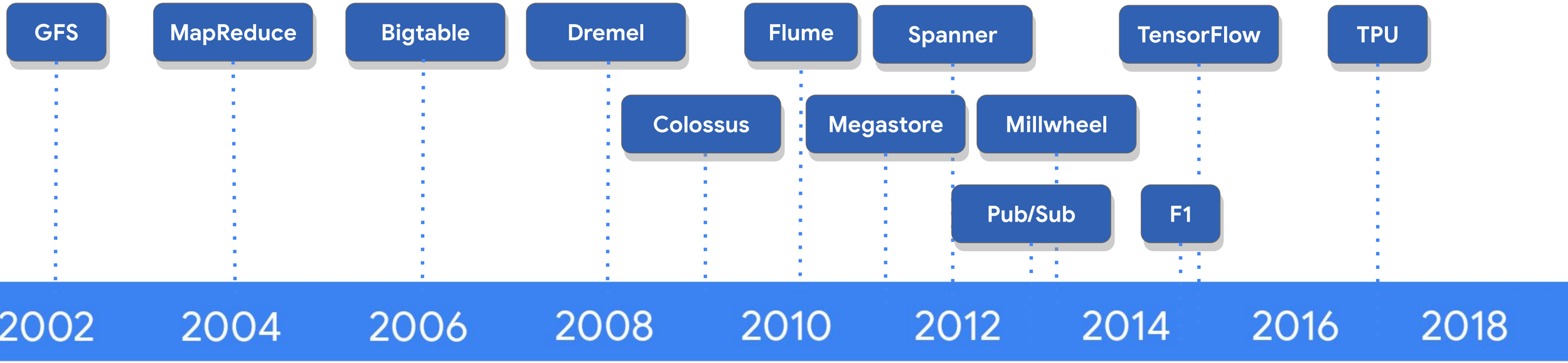
- Google innovation timeline
- Choosing the right approach

What you can do with GCP

Activity: Explore a customer use case

The different data roles in an organization

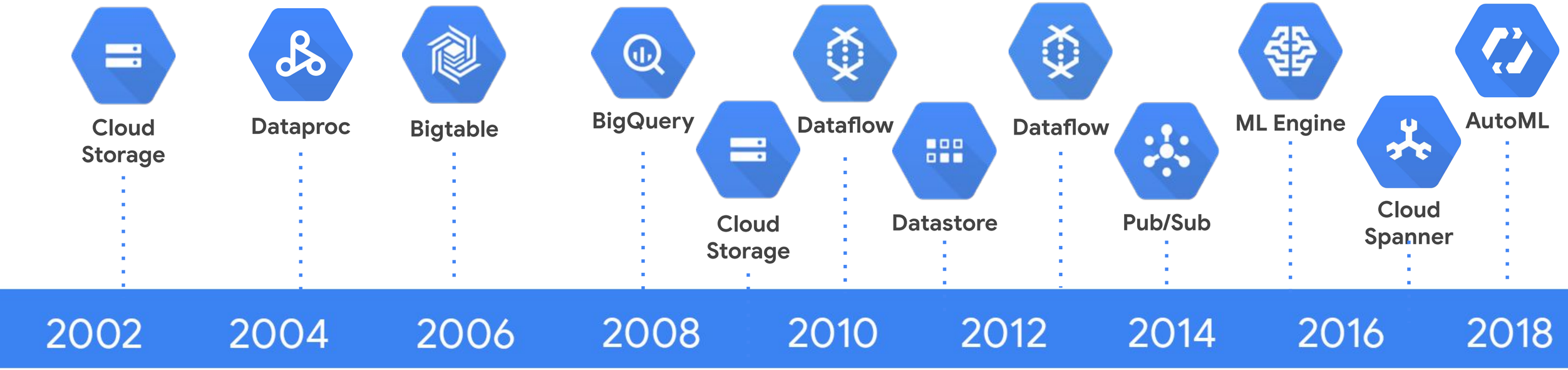
Google invented new data processing methods as it grew



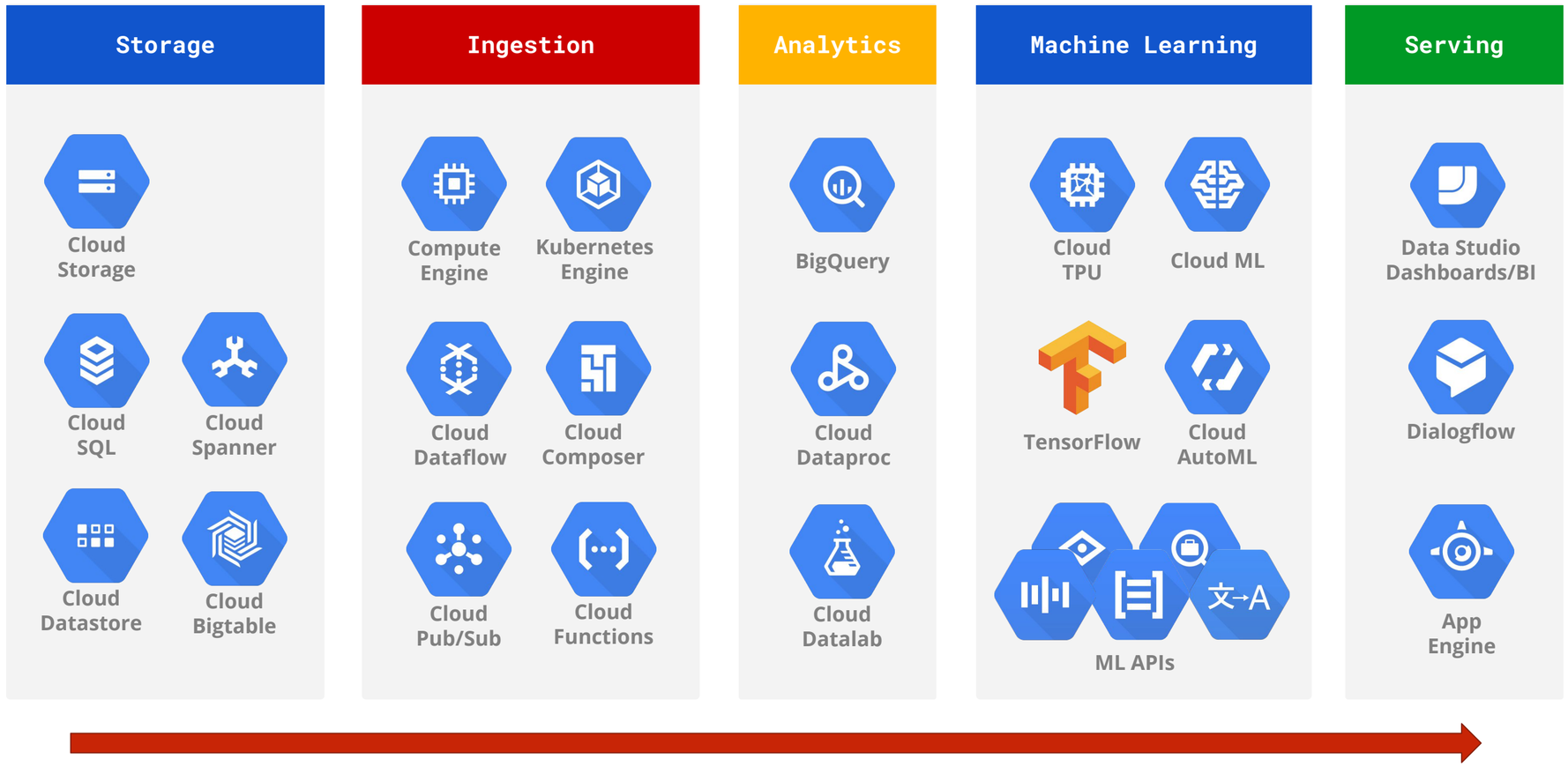
http://research.google.com/pubs/papers.html

Google Cloud

Google Cloud opens up that innovation and infrastructure to you



The suite of big data products on Google Cloud Platform



Google Cloud

Demo

Query 2 billion lines of code in less than 30 seconds

Github on BigQuery

Agenda

Google Cloud Platform infrastructure

- Compute
- Storage
- Networking
- Security

Big data and ML products

- Google innovation timeline
- Choosing the right approach

What you can do with GCP

Activity: Explore a customer use case

The different data roles in an organization

Keller Williams uses AutoML Vision to automatically recognize common elements of house furnishings and architecture

Cloud AutoML Vision

“Modern” style

Granite countertops

Google Cloud

Ocado routes emails based on NLP

Improves natural language processing of customer service claims

“Hi Ocado, I love your website. I have children so it’s easier for me to do the shopping online. Many thanks for saving my time! Regards”

Feedback

Customer is happy

“Thanks to the Google Cloud Platform, Ocado was able to use the power of cloud computing and train our models in parallel.”

Google Cloud

Kewpie uses ML to sort out the bad potatoes in baby food

kewpie

Original process required humans to identify low-quality ingredients, which was expensive and stressful.

Machine learning was used to replicate the quality control process.

Google Cloud

Agenda

Google Cloud Platform infrastructure

- Compute
- Storage
- Networking
- Security

Big data and ML products

- Google innovation timeline
- Choosing the right approach

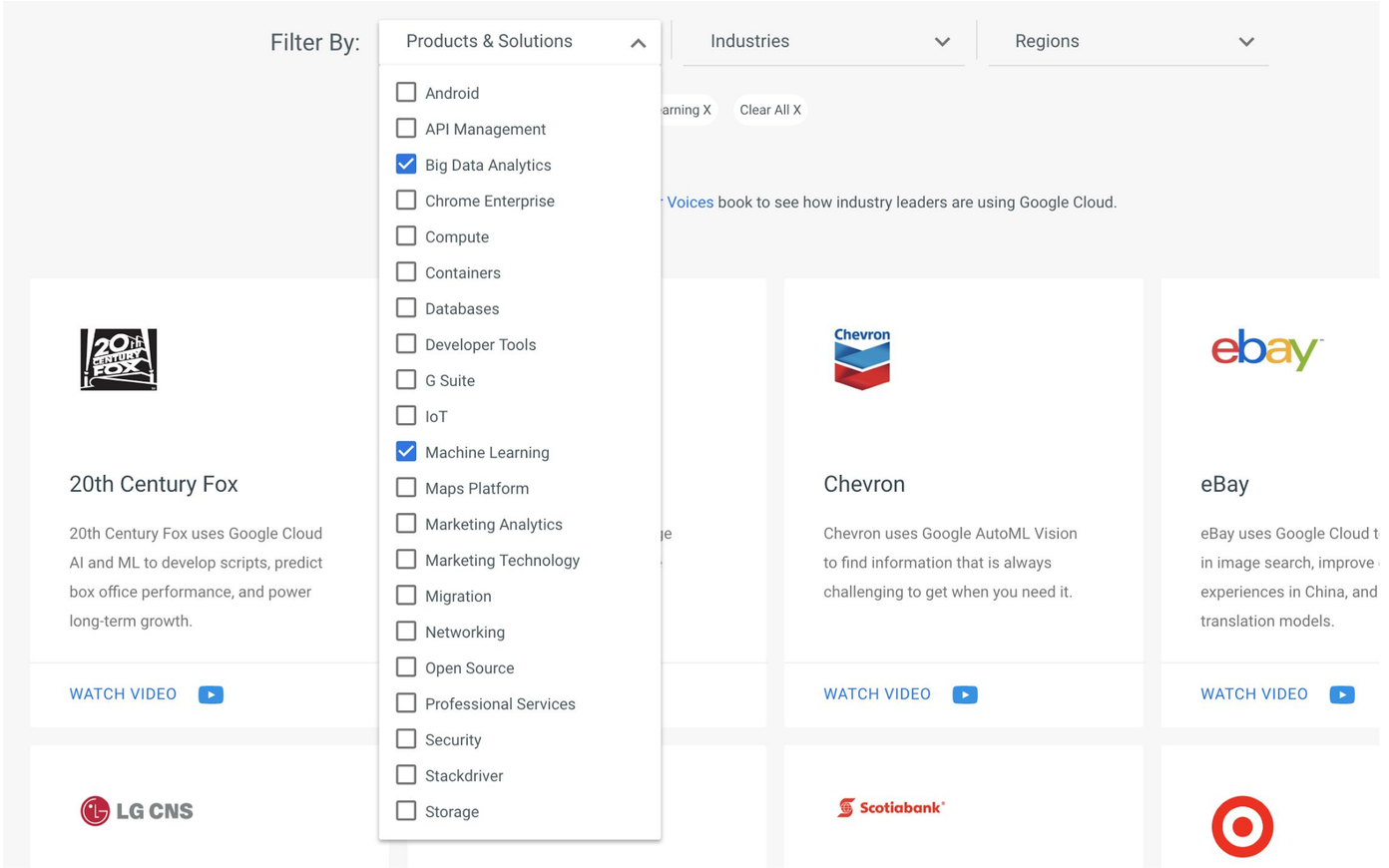
What you can do with GCP

Activity: Explore a customer use case

The different data roles in an organization

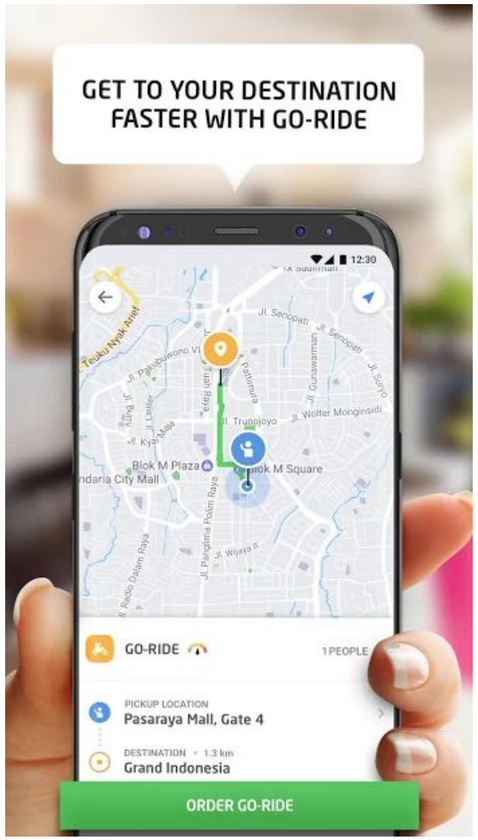
Your turn: Analyze real customer big data use cases

1. Navigate to cloud.google.com/customers/.
2. Filter Products & Solutions for Big Data Analytics.
3. Find an interesting customer use case.
4. Identify the key challenges, how they were solved with the cloud, and impact.



Google Cloud

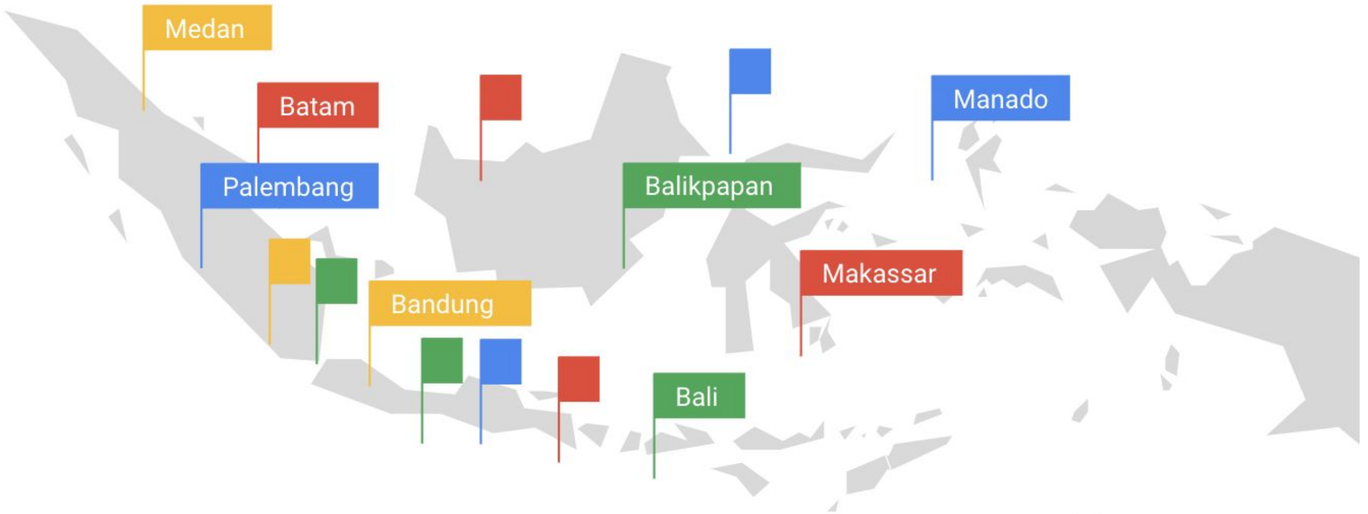
GO-JEK brings goods and services to over 2 million families in 50 cities in Indonesia



Google Cloud

GO-JEK's footprint nationwide

Operating in 50 cities throughout Indonesia



+77m
app downloads

+150k
merchants

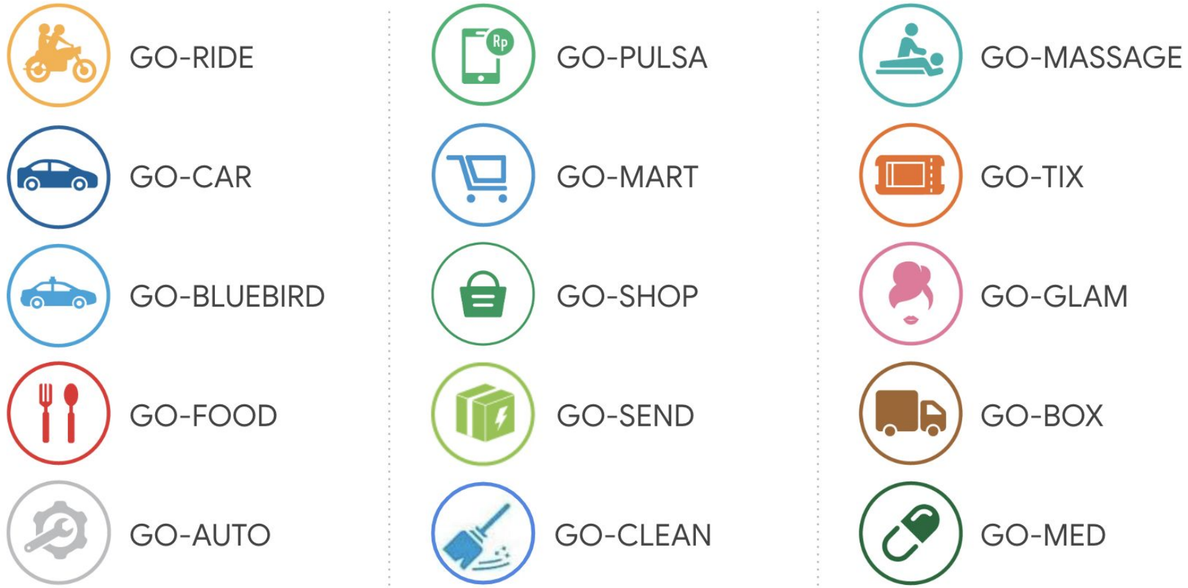
50
cities

+1m
drivers

2m
families

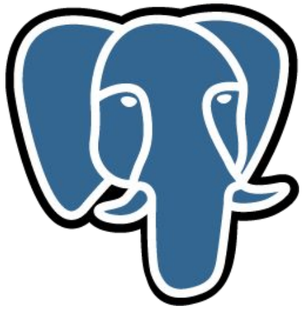
Google Cloud

GO-JEK manages 5 TB+ per day for analysis



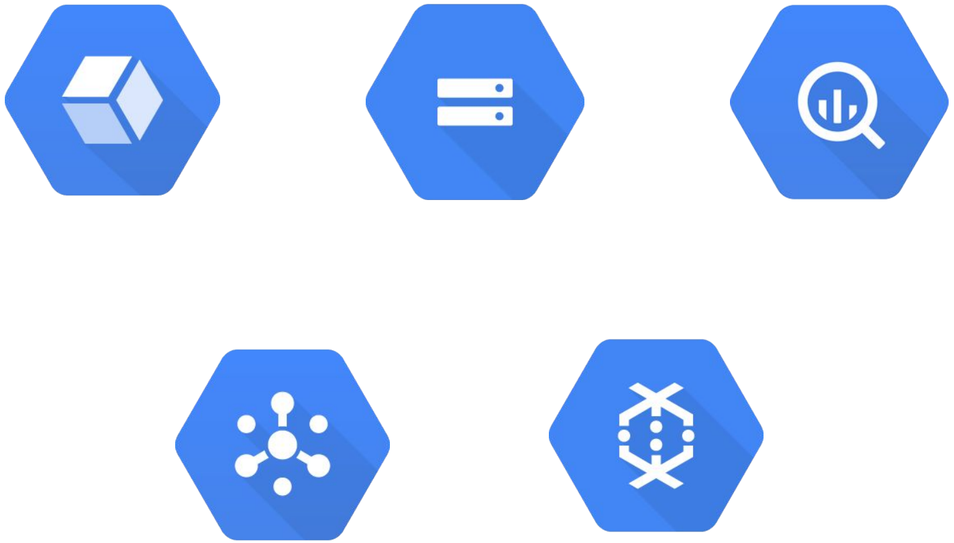
Google Cloud

GO-JEK soon faced data scale and latency challenges



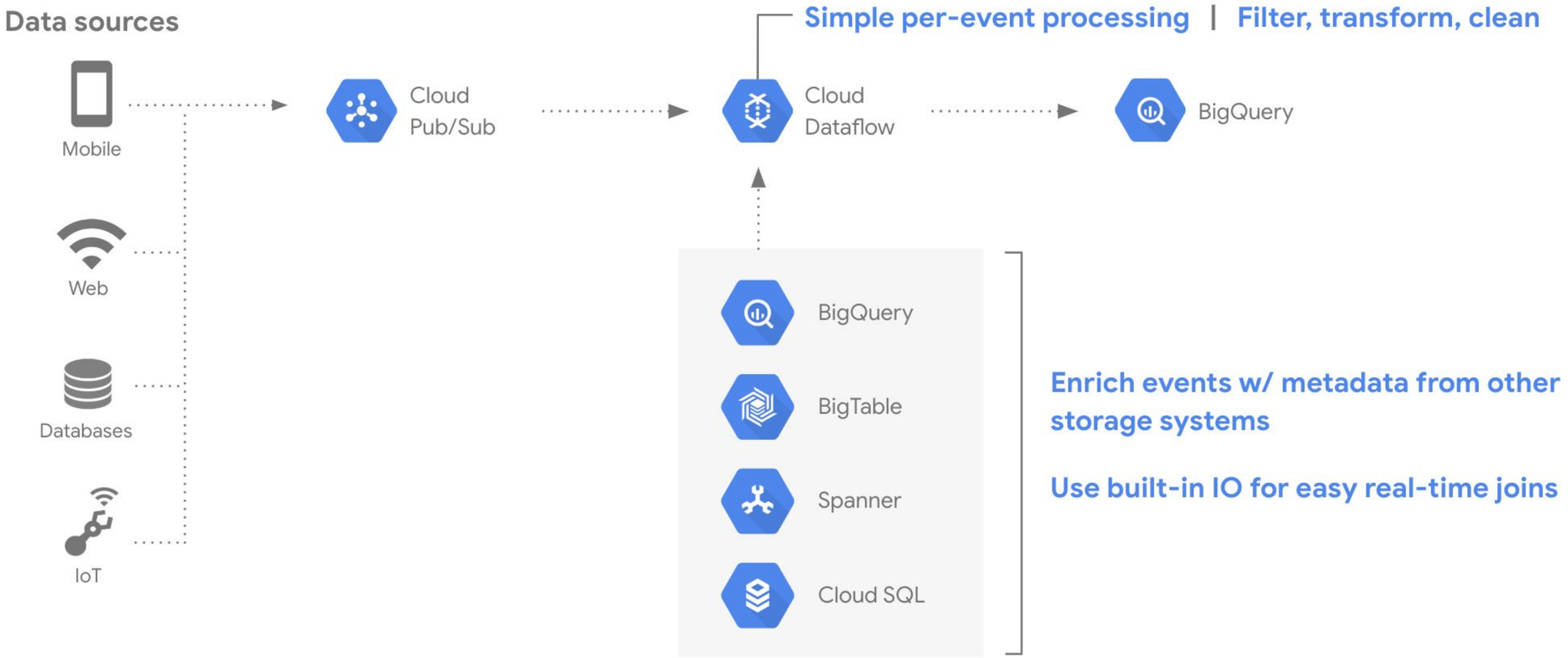
“Most of the reports are Day +1, so we couldn’t identify the problems as soon as possible.”

GO-JEK migrated their data pipelines to GCP



- High performance scalability with minimal operational maintenance
- More granular data with high velocity and less latency (stream processing)
- The ability to solve business problems with real time data insights

GO-JEK architecture review



GO-JEK ride-share supply/demand use case

Business question

Which locations have mismatched supply and demand in real time.

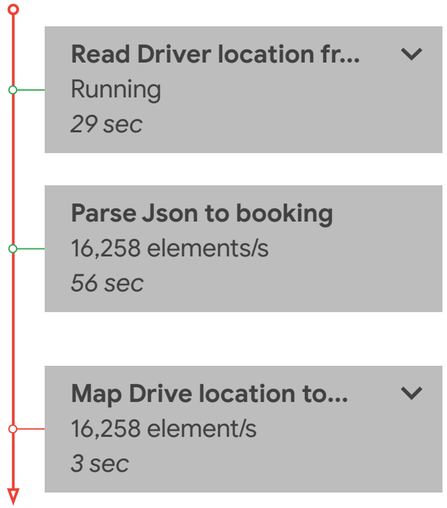
Challenge

We **ping every one of our drivers every 10 seconds**, which means 6 million pings per minute and 8 billion pings per day. How do we stream and report on such a volume of data?

Autoscale streaming pipelines with Cloud Dataflow

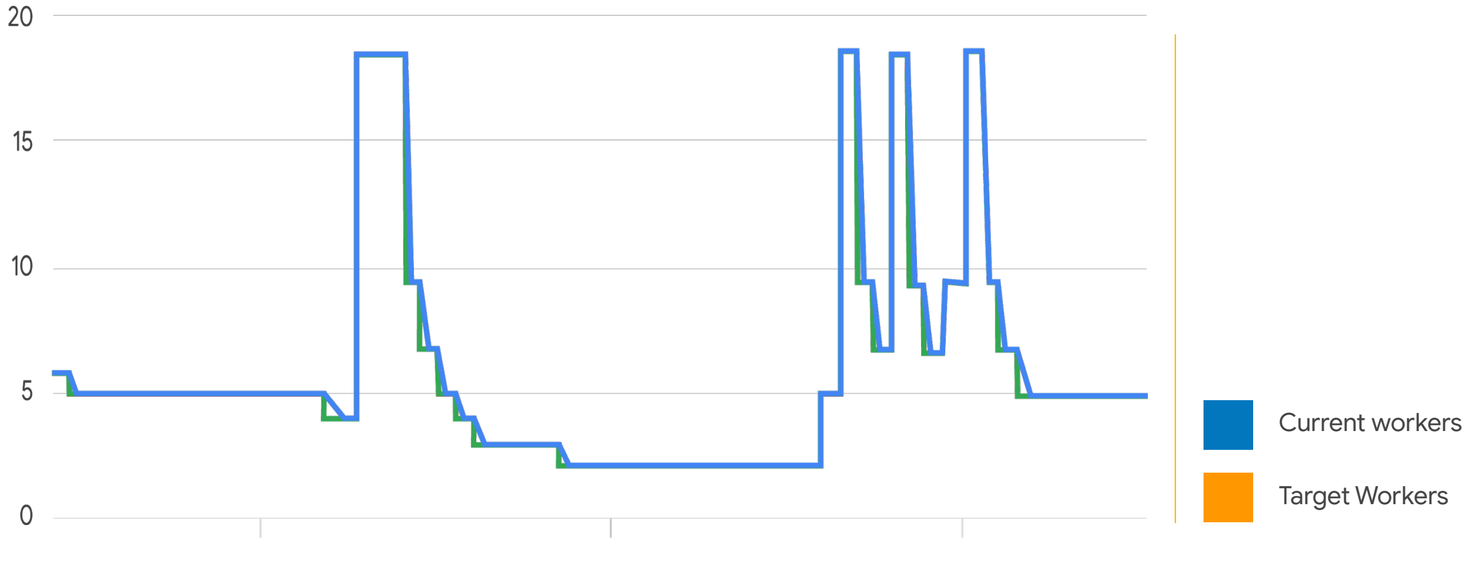
Driver location ping

Dataflow autoscale based on throughput data



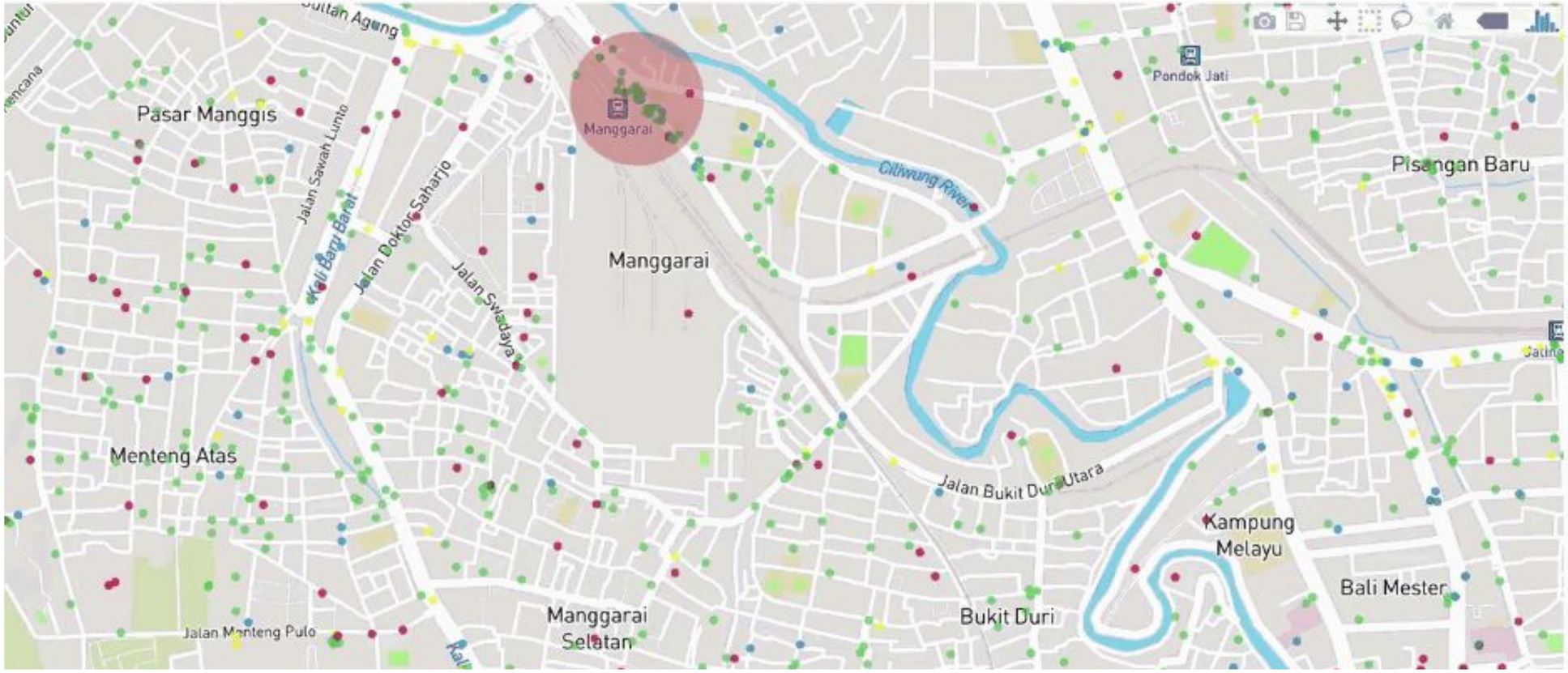
Autoscaling

Automatically adjust the number of workers based on demand



Google Cloud

Visualize demand/supply mismatches with GIS data



Google Cloud

Agenda

Google Cloud Platform infrastructure

- Compute
- Storage
- Networking
- Security

Big data and ML products

- Google innovation timeline
- Choosing the right approach

What you can do with GCP

Activity: Explore a customer use case

The different data roles in an organization

Review: Who are the people solving these challenges?

Data Analyst

- Analyst

Data Engineer

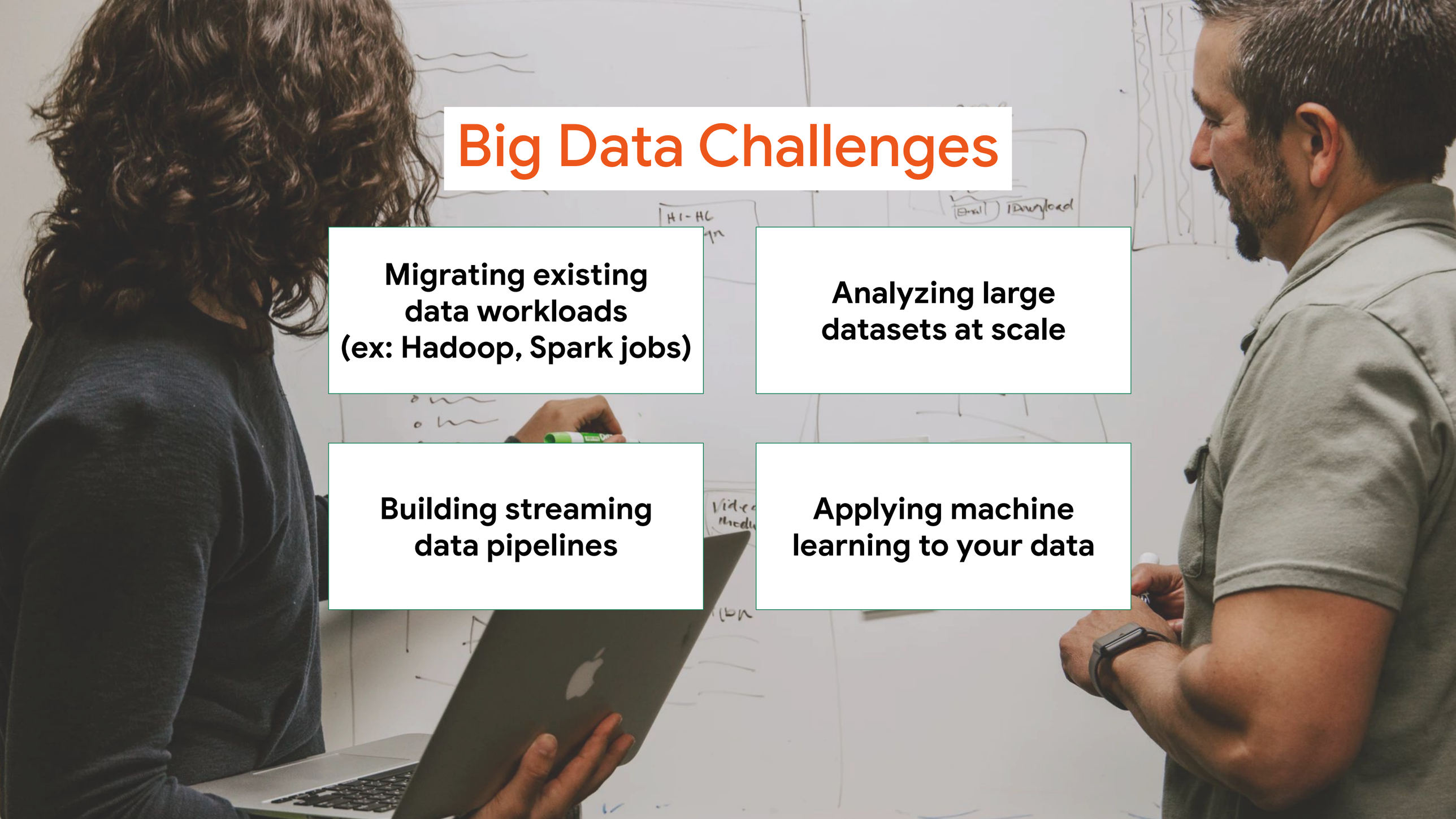
- Data Engineer

Applied ML Engineer

- Data Scientist
- Ethicist
- Statistician
- Social Scientist
- Applied ML Engineer
- Researcher

Tech Lead

- Analytics Manager
- Decision Maker



Big Data Challenges

Migrating existing data workloads (ex: Hadoop, Spark jobs)

Analyzing large datasets at scale

Building streaming data pipelines

Applying machine learning to your data

Personas: Who are the people solving these challenges?

Learner context and backgrounds

Goals/needs

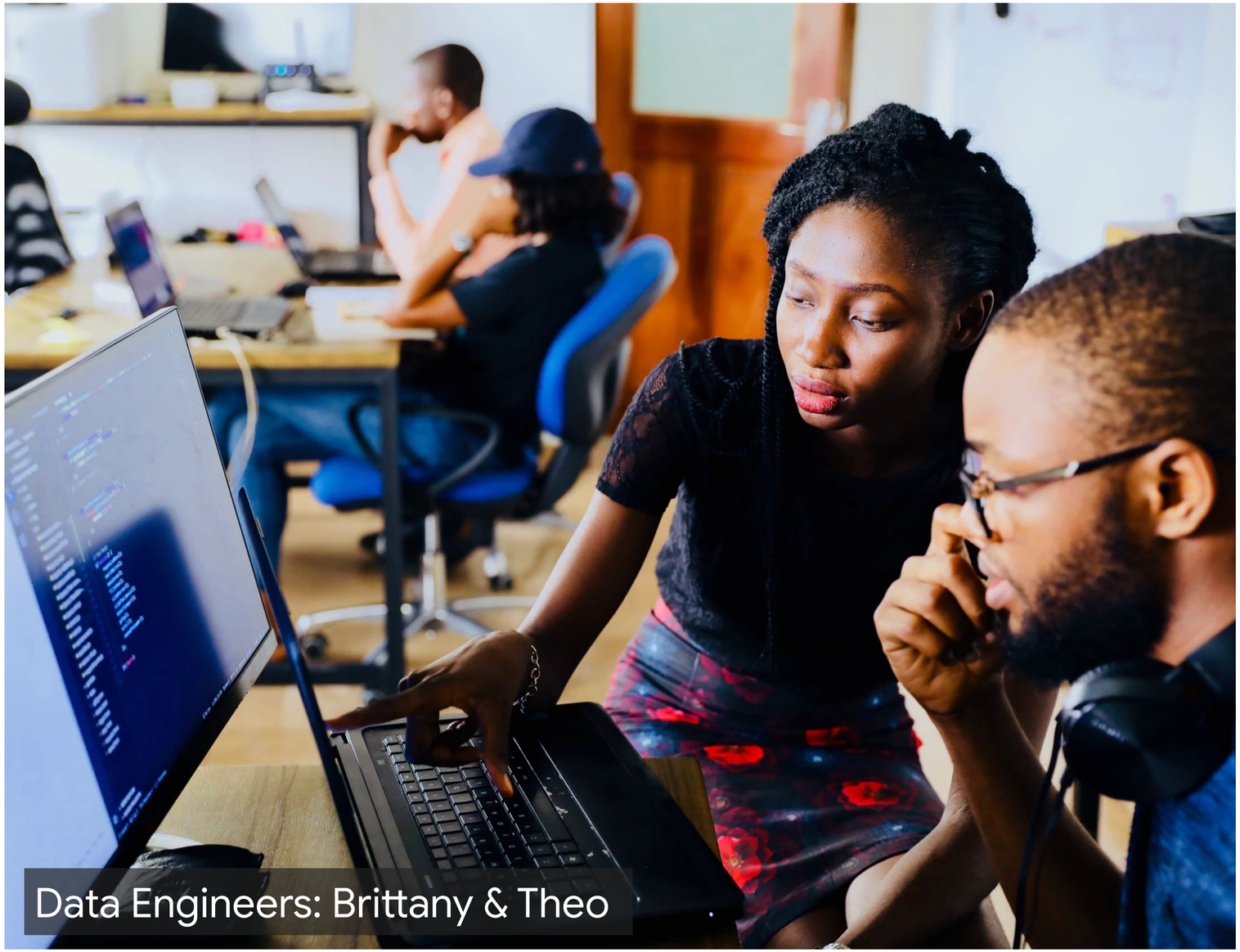
Competencies

Challenges

We want to migrate
on-premise
Hadoop workloads

“Our CTO has challenged our data engineering team to find ways we can spend less on managing our on-prem cluster.

Right now, we just want to show her options that don’t require any code changes to our 100+ Hadoop jobs.”

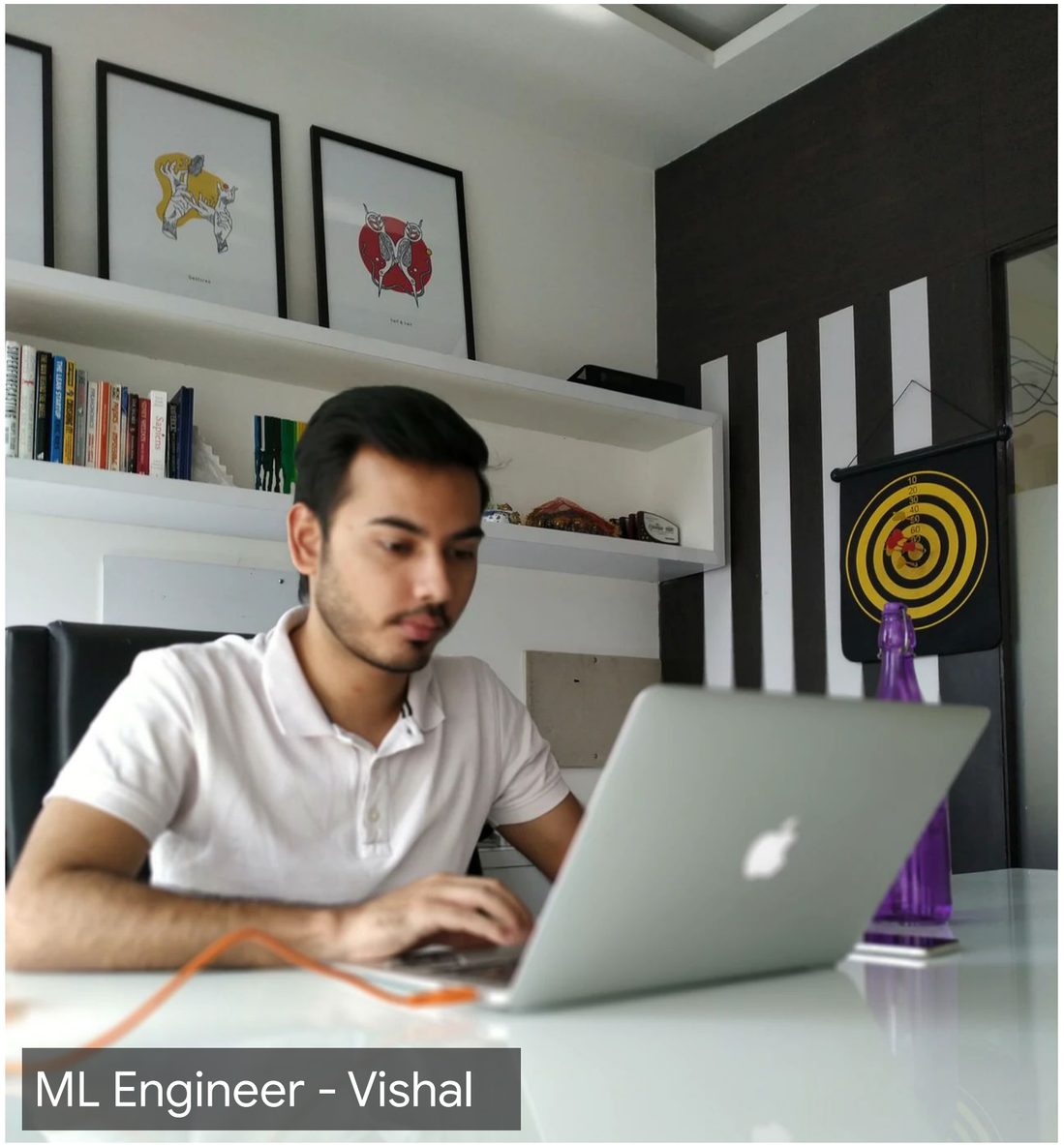


Google Cloud

I want to create real-time dashboards from streaming data sources

“I really want to design our data pipelines for the future. For us that means lots and lots of streaming data from our IoT devices with low latency.”

Google Cloud



I want to open up AI and ML
for everyone to use

“I pitched my team on the value ML can add, and I’ve got buy-in for a prototype.

What are some of the easiest ways I can see whether ML is feasible for my data?”

I want to load and analyze all my data at scale

“I’ve been asked to find a way to ingest and query my 5 TB of company data for fast insights... and I don’t have time to manage hardware.”

Google Cloud

BigQuery has over 130 Public Datasets to explore

- Advertising (7)
- Analytics (6)
- Big data (27)
- Climate (20)
- Databases (1)
- Developer tools (20)
- Economics (27)
- Encyclopedic (29)
- Finance (3)
- Genomics (3)
- Health (8)
- Machine learning (1)
- Maps (1)
- Public safety (13)
- Science & research (47)
- Social (3)
- Transportation (1)
- Other (11)

GitHub

Lab

Exploring Public Datasets using the BigQuery Web UI

- Open BigQuery
- Query a Public Dataset
- Create a custom table
- Load data into a new table
- Querying basics

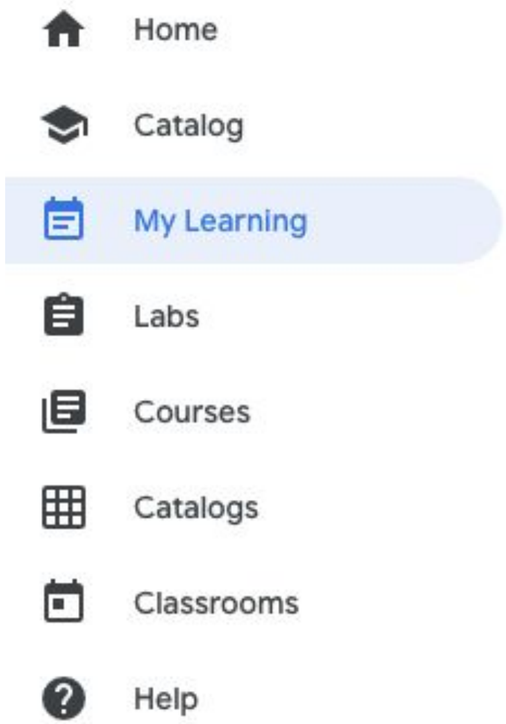
Open Qwiklabs

1. **Open an incognito window**
(or private/anonymous window)

2. **Go** to events.qwiklabs.com

3. **Sign In** with existing account or **Join** with new account (with email you used to register for the bootcamp)

4. Launch the course from **My Learning**

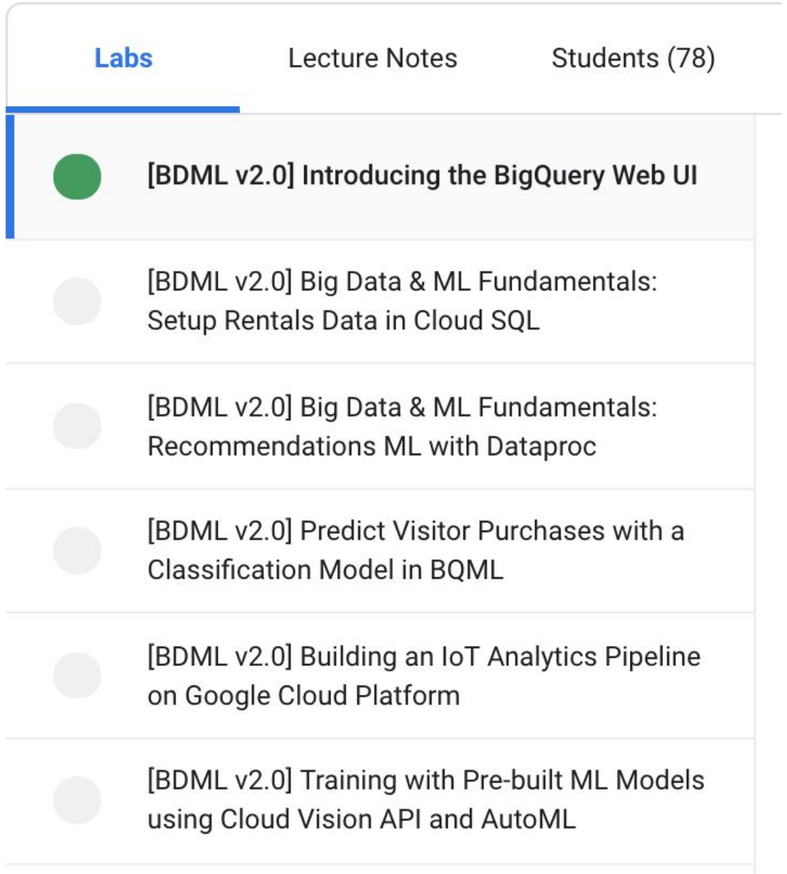


Don’t remember what email you used to register with or don’t have access to it?

https://goo.gl/xrVBpM

If so, use the link or QR code to add your email address for access to the course material.

View your lab



Note: You can access the course PDFs under Lecture Notes

Launch the lab and start on Lab 1.

Start Lab 00:40:00

Score —/15

Labs will last for 40 minutes or so (and materials can be accessed for 2 years)

Tip: Track your progress with Score X/15

[BDML v2.0] Introducing the BigQuery Web UI

40 minutes Free ☆☆☆☆☆ Rate Lab

Overview

Storing and querying massive datasets can be time consuming and expensive without the right hardware and infrastructure. Google BigQuery is an enterprise data warehouse that solves this problem by enabling super-fast SQL queries using the processing power of Google's infrastructure. Simply move your data into BigQuery and let us handle the hard work. You can control access to both the project and your data based on your business needs, such as giving others the ability to view or query

Pro tip: Use the table of contents on the right to quickly navigate

Do not click
End Lab until you are done with that lab

(note: each lab is independent)

Google Cloud

Do Ask Questions

(we have a talented team of experts to help!)