Introducing Google Cloud Platform

Google Cloud Platform Fundamentals: Big Data and Machine Learning

Version #1.1

Google Cloud

Notes:

30 minutes + 15 minutes lab

Notes:

1. Introduction

Overview of GCP as a whole, but with emphasis on the data-handling aspects of the platform

- GCP, GCP big data
- Usage scenarios
- Create an account on GCP

2. Foundation of GCP

Compute and Storage with a focus on their value in data ingest, storage, and federated analysis

- Compute Engine
- Cloud Storage
- Start GCE instance
- Upload data to GCS

3. Data analytics on the Cloud

Common use cases that Google manages for you and for which there is an easy migration path to the Cloud

- Cloud SQL
- Dataproc
- Import data into and query Cloud SQL
- Machine Learning with Dataproc

In the morning, we will complete Modules 1 and 2 and get halfway through Module 3.

4a. Scaling data analysis
Change how you compute, not just where you compute with GCP
- Datalab
- Datastore, Big Table
- BigQuery

5. TensorFlow
Change how you compute, not just where you compute with GCP
- TensorFlow
- Datalab instance
- BigQuery
- Demand forecasting with ML

6. Data processing architectures
Scaleable, reliable data processing on GCP
- Pub/Sub
- Dataflow

7. Summary
Course summary
- Resources

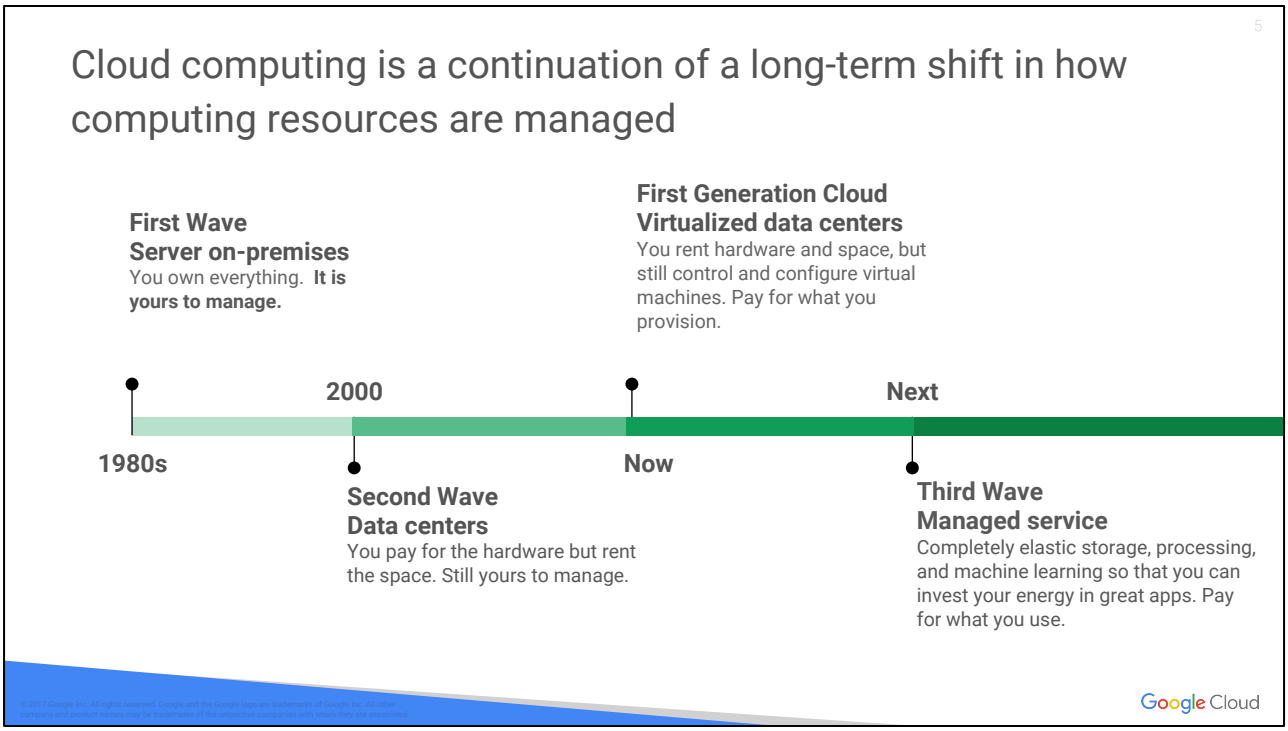
Please feel free to use the appendixes for self-study.
In the morning, we will get halfway through Module 3.

Please feel free to use the appendixes for self-study.

Agenda

What is Google Cloud Platform

Google Cloud Big Data products



Cloud computing is essentially the continuation of a model where you get to rent out computing infrastructure and have it managed by dedicated professionals. Equinix and CenturyLink are two of the largest data center providers in the US. They are not exactly household names. So, why are Amazon, Google, and Microsoft even in this business? In particular, why is Google doing Cloud?

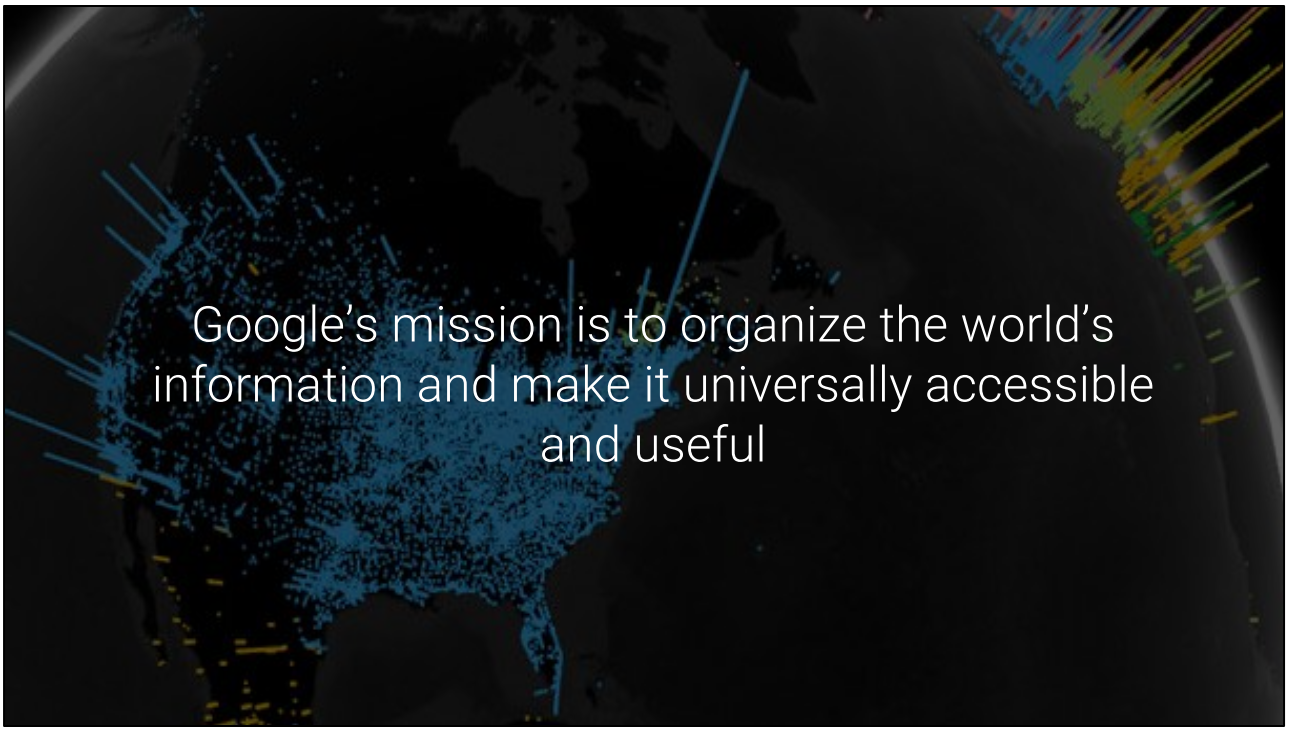
The concept of cloud computing began with colocation. Instead of operating your own data center, you rented space in a colocation facility. This was the first wave of outsourcing IT. With colocation, the transfer of ownership was minimal: you still owned the machines and you maintained them. Traditionally, colocation is not thought of as cloud computing, but it did begin the process of transferring IT infrastructure out of your organization.

Today, cloud computing involves virtualized data centers—virtual machines and APIs. Virtualization provides elasticity. You automate infrastructure procurement instead of purchasing hardware. With virtualization, you still maintain the infrastructure. It is still a user-controlled/user-configured environment. This is the same as an on-premises data center, but now, the hardware is in a different location. Virtualization does provide a number of benefits: your development teams can move faster, and you can turn capital

expenses into operating expenses.

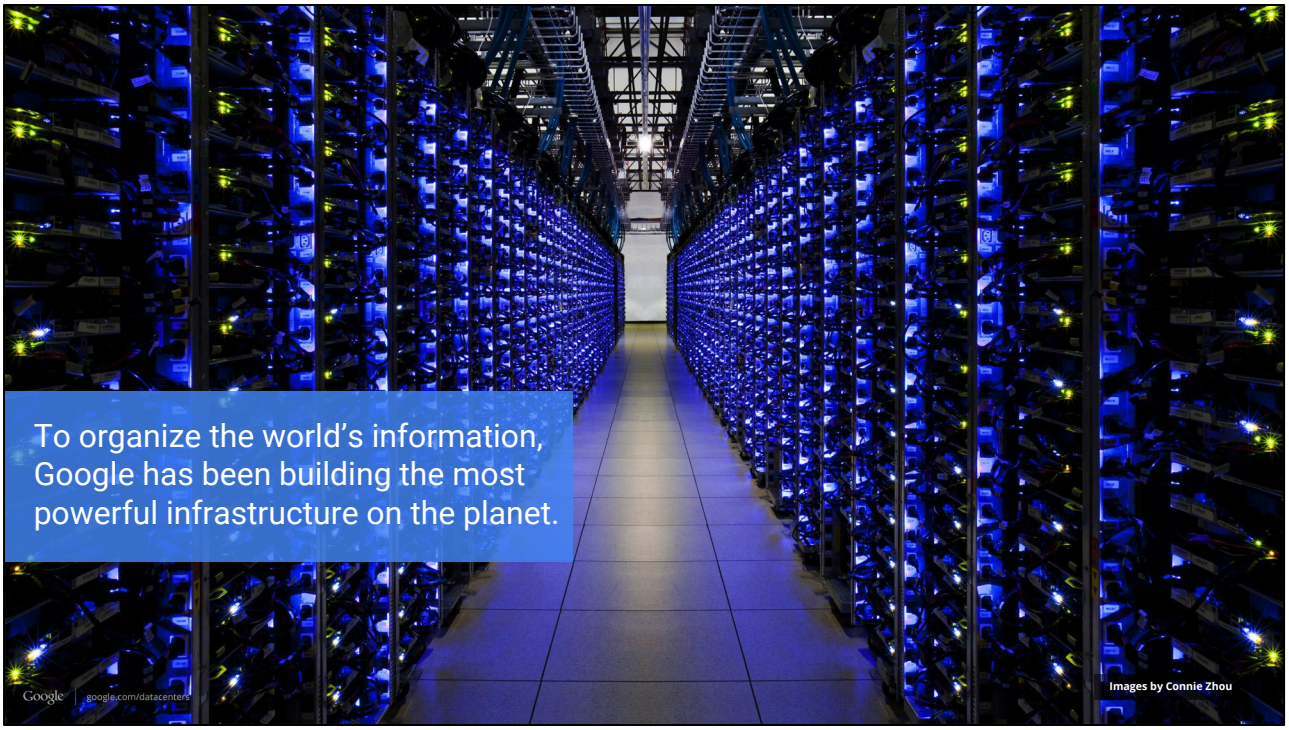
The next wave of cloud computing is a fully automated, elastic cloud. This involves a move from user-maintained infrastructure to automated services. In a fully automated environment, developers do not think about individual machines. The service automatically provisions and configures the infrastructure used to run your applications. Google is uniquely positioned to propel organizations into the next wave of cloud computing.

But what does Google have to do with Cloud?



Google Cloud is so powerful because Google has had to deal with lots more data than anyone else, much earlier than anyone else. Using Google Cloud, you can take advantage of Google's powerful data infrastructure.

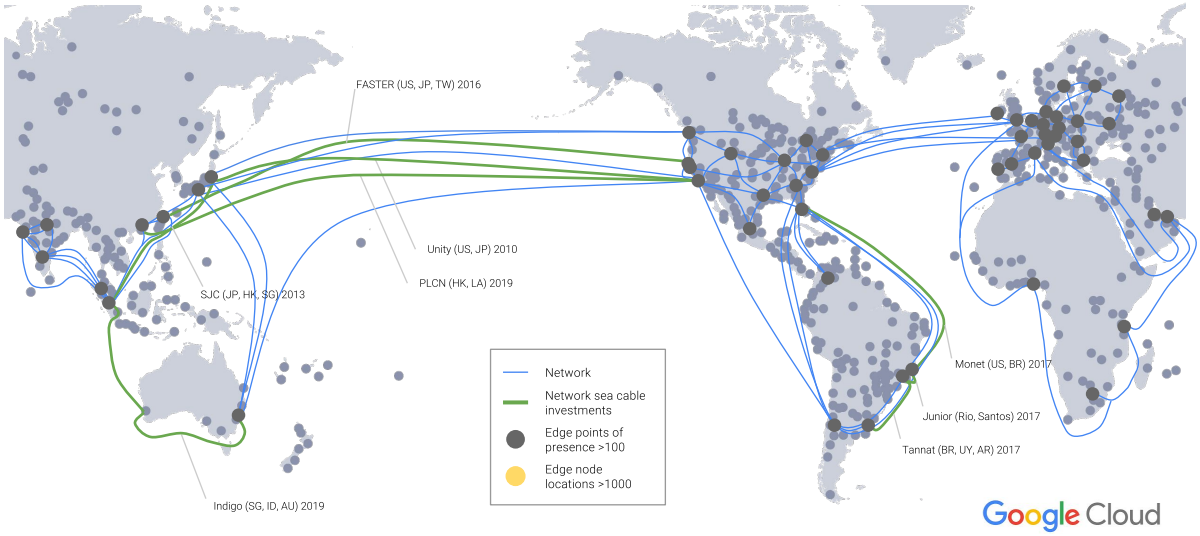
- Our mission is to organize the world's information and make it universally accessible and useful.
- This enables people all around the world—from farmers in California to small business owners in Kenya—to find and connect to the information they are looking for.
- If we *understand* the information, we can help them even more. This is where machine intelligence and learning comes in



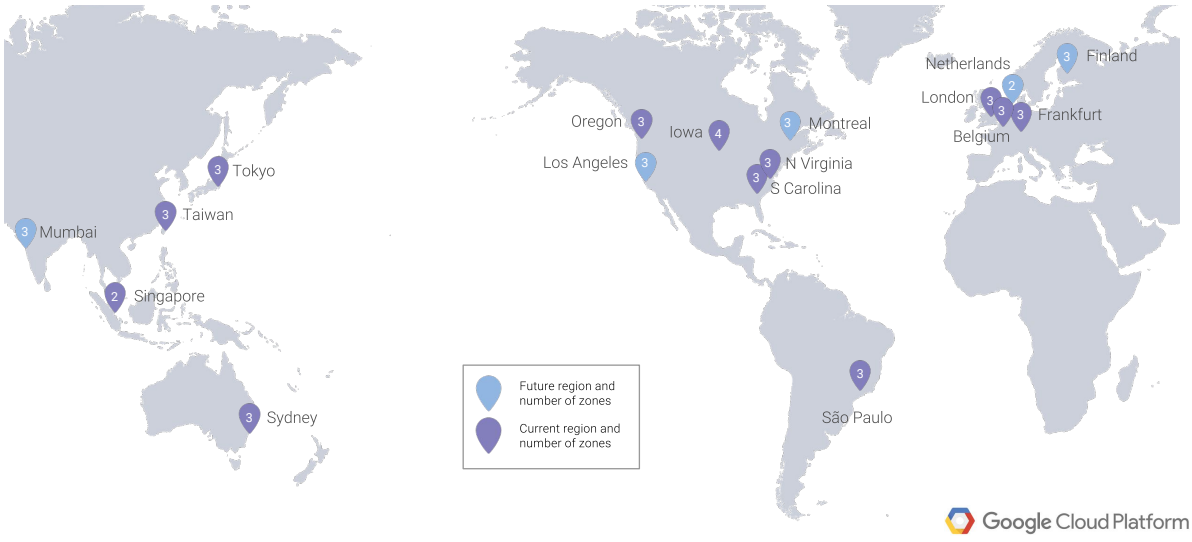
Notes:

In terms of hardware, that infrastructure consists of a global network of data centers and private fiber connections
Google Cloud has 12 regions, 36 zones, over 100 points of presence, and a well-provisioned global network with 100,000s of miles of fiber optic cable.
Some (not all) of Google's data centers. The ones shown are the Cloud ones, and ten new are expected in 2017. https://cloud.google.com/about/locations/ has an interactive graphic that will stay up-to-date.
Google lays its own transoceanic fibers
Google doesn't release the locations of their edge centers, but an interesting project tries to map them anyway using network speeds. The point is that these are much more widespread than the data centers, and because of software-defined networking (see below), this really matters ...
Software-defined networking means that any machine within Google Cloud gets load-balancing, auto-migration, etc. without you (the customer) having to do anything. Also, every machine within a region appears to be just one hop away.

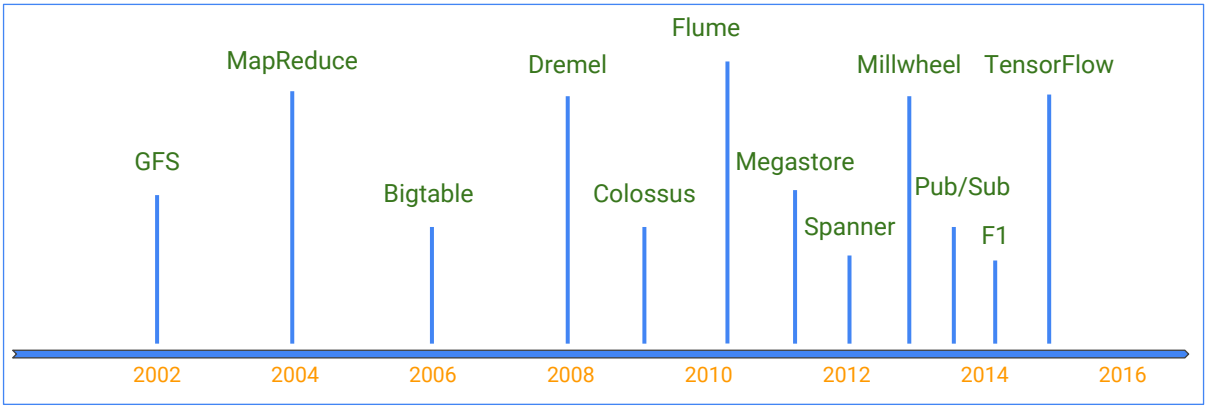
In terms of hardware, Google Cloud has the largest cloud network, with more than 100 points of presence



The network connects 12 regions, with 5 more coming in 2017 through 2018



In terms of software, organizing the world's information has meant that Google needed to invent data processing methods



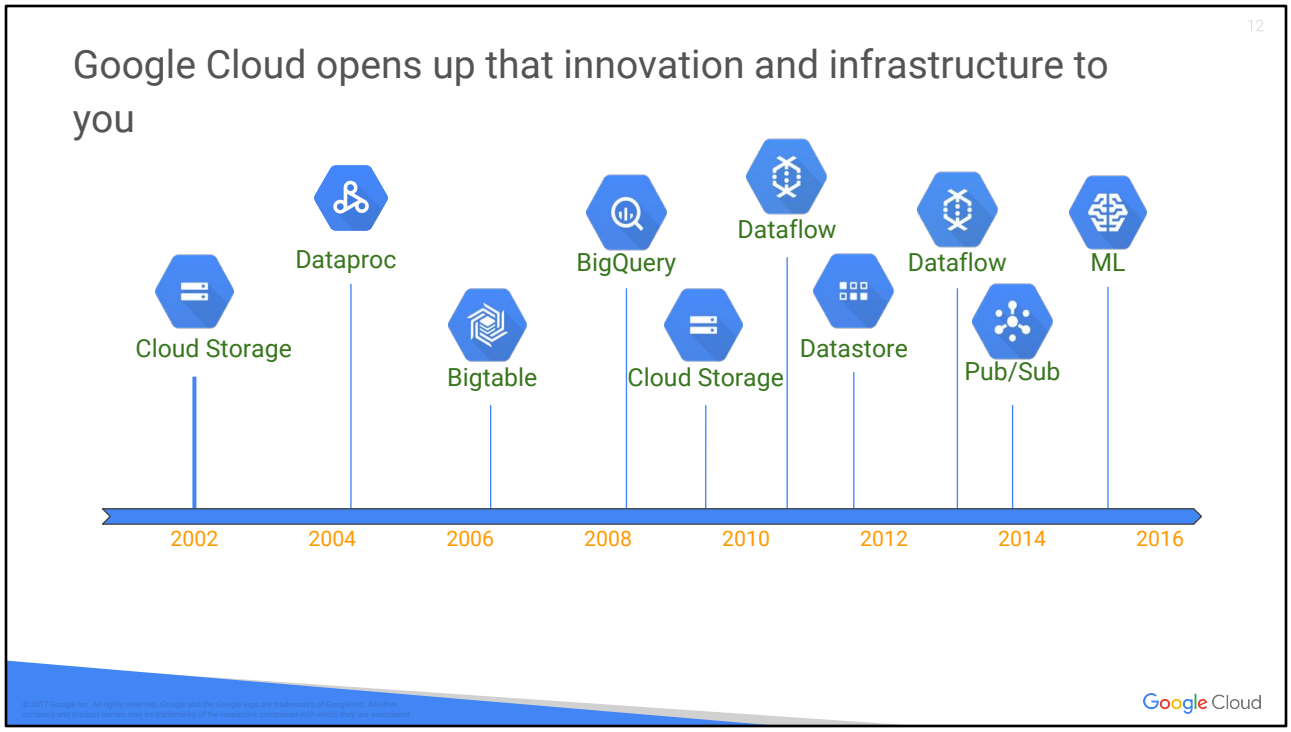
http://research.google.com/pubs/papers.html

Google Cloud

Notes:

Google Research Publications referenced are available here:
http://research.google.com/pubs/papers.html
The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines, 2009
http://research.google.com/pubs/pub35290.html

Organizing the world's information at never-before-heard-of scales means that Google had to invent new ways of doing data processing. Your standard database technology wouldn't do it. So, Google innovated technologies and wrote white papers on them, and these became the basis of the Hadoop ecosystem. The problem? Even though Google's implementations are much better, and Google has moved on from those early technologies, other organizations haven't been able to use our newer technologies.



Notes:

Google Research Publications referenced are available here:
http://research.google.com/pubs/papers.html
The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines, 2009
http://research.google.com/pubs/pub35290.html

So, the mode is now to provide the exact implementations that Google uses, and give you a way to use them directly. The APIs are open-sourced, but not Google's implementations (the Apache Beam/DataFlow model). Starting with Bigtable, there are no exact equivalents any more. (Bigtable != HBase/MongoDB and BigQuery != Amazon RedShift).

http://db-engines.com/en/system/Google+Cloud+Bigtable%3BHBase%3BMongoDB: The main difference is that Bigtable is no-ops (hosted). It is also more performant for very, very large databases.

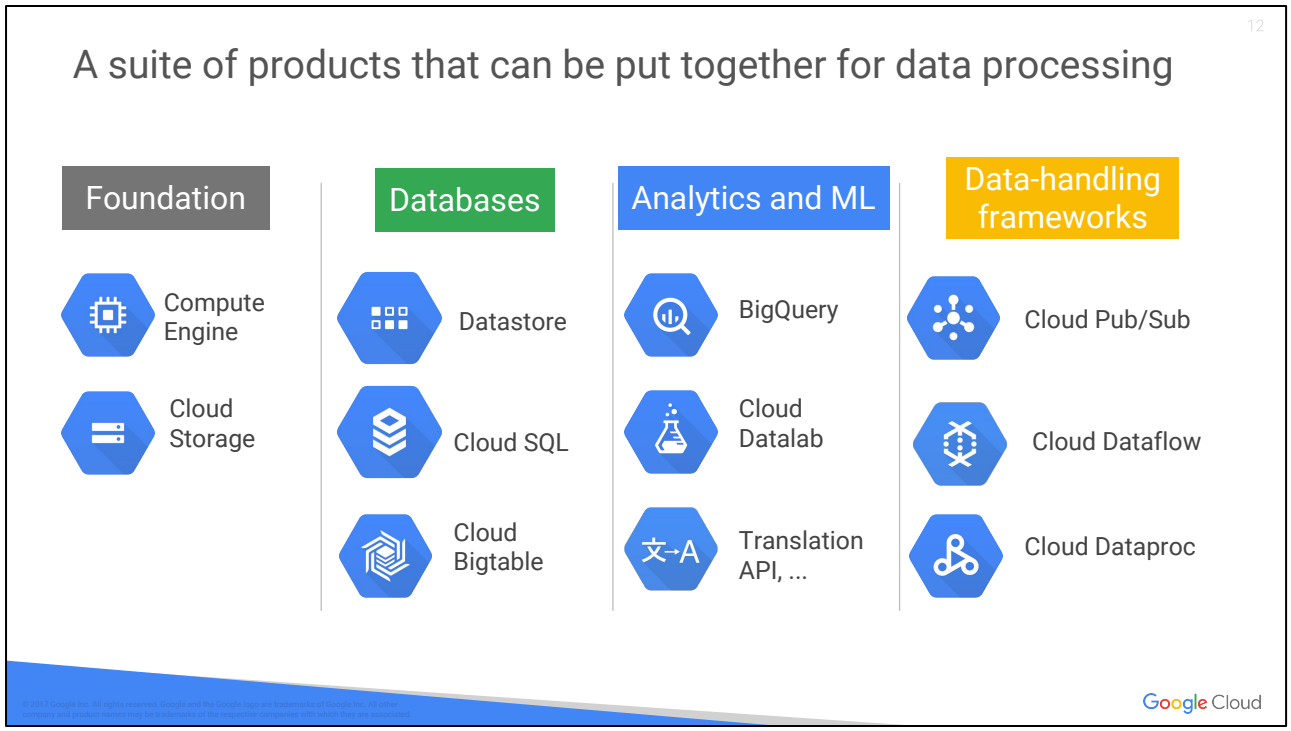
https://www.quora.com/How-good-is-Googles-BigQuery-as-compared-to-Amazons-Redshift: The differences here are similar. BigQuery is no-ops, but Amazon Redshift requires provisioning. The quora answer by Peter Mueller says what the bloodless word "provisioning" means in practice: They move data from Amazon S3 to Google Cloud Platform just so they don't have to

worry about determining how much hardware they need.

Agenda

What is Google Cloud Platform

Google Cloud big data products



Notes:

A suite of blue hexagons ...
But why do you have different databases?

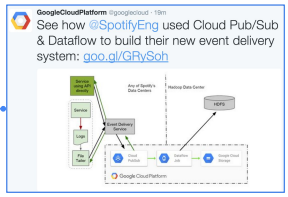
Spotify illustrates the typical journey of companies that come to Google Cloud: From lower costs to increased reliability to business transformation

Spend less
No-ops, Pay for use, Secure

1

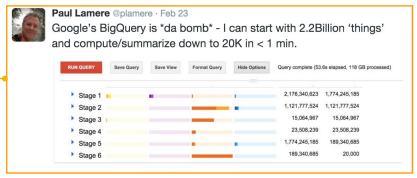
Flexible
Complete

2



Innovative
Powerful

3



https://cloud.google.com/blog/big-data/2016/03/big-data-sessions-by-customers-at-gcp-next-2016

Google Cloud

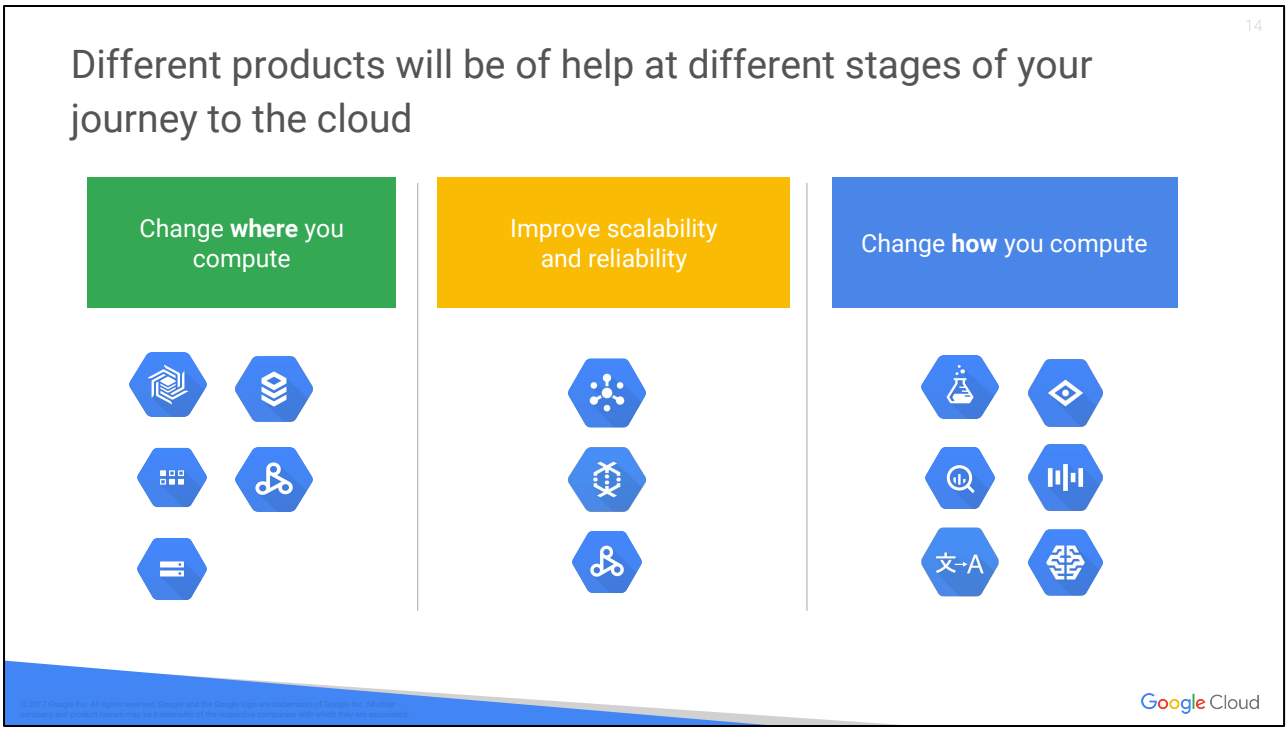
Notes:

The key reason that companies move? They save money on operations and administration, and they only pay for what they use (not what they reserve!). But also because they realize that their data is more secure on Google Cloud Platform (How many security engineers would Spotify have had? How about Google?).

The second quote is about completeness and flexibility -- Google Cloud Platform has a messaging platform *and* a data pipeline API, and these are open enough that Spotify could build their custom tool on top of them.

The “da bomb” quote is about innovativeness (BigQuery) and the power of the tools. It surprises even people who live in the data analysis world and can help you build citizen data scientists.

There is no quote for “easy to use” -- that’s the purpose of the course today. You can be the judge today in the afternoon. But realize that as easy as everything looks, Google Cloud Platform is also cost-effective, secure, innovative, powerful, flexible and complete.



Notes:

In the green tier, you see solutions for storage and fast retrieval. You probably use variations of these in your enterprise now. The products here are meant to ease your transition to the cloud by providing completely managed versions of these services so that you spend less on ops and admin.

In the orange tier, you have products that help you build more reliable, scalable data ingest and processing frameworks.

In the blue-tier, you see products that will change the way you do computing.

We will focus in this course on the green and blue boxes. The orange boxes are meant more for systems programmers. We'll cover the green tier in Module 3, the orange tier in Module 6, and the blue tier in Modules 4 and 5.

Atomic Fiction lowered their costs with per-minute (now per-second) billing

Change **where** you compute

https://www.youtube.com/watch?v=mBY-RjE15WA

Google Cloud

Notes:

Change where you compute. It can save you tons of money.

There is a 2 minute video. It's pretty awesome. I will take a break and let Atomic Fiction tell you how they spent their customers' money more wisely.
Compute Engine's per-minute billing resulted in 4 more minutes of film

FIS was able to improve reliability and scalability on a massive data-processing challenge

The Consolidated Audit Trail (CAT) is a data repository of all equities and options orders, quotes, and events; FIS processed the CAT to organize 100 billion market events into an “order lifecycle” in a 4-hour window using Cloud Bigtable.

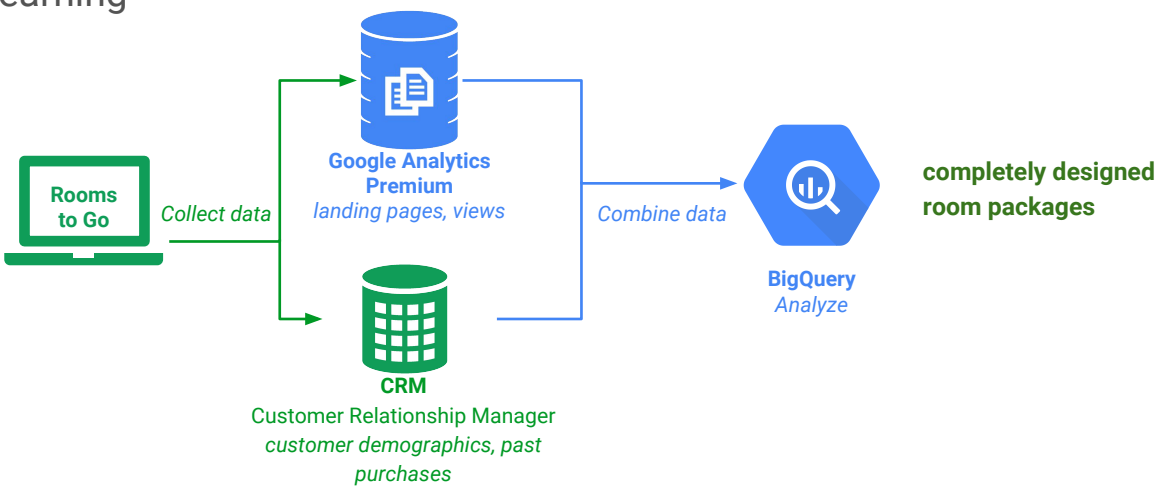
Google Cloud

Notes:
The Consolidated Auditing Trail for the SEC is aimed at providing more transparency into financial markets, partly in response to the computer-driven 2010 "flash crash" which briefly created U.S. stock prices. See: http://www.cbronline.com/news/cloud/aas/sungard-google-take-on-30-petabytes-of-cloud-data-4600745.

Overview: Data to process: Data in the Consolidated Audit Trail (CAT).
A data repository of all equities and options orders, quotes,
and events
Challenges: How to process the CAT and organize 100 billion market events into an “order lifecycle” in a 4 hour window

Store 6 years (~30PB) of data

Rooms to Go transformed its business with data and machine learning



https://www.thinkwithgoogle.com/case-studies/rooms-to-go-improves-the-shopper-experience.html

Notes:

Change how you compute. You can make your company more efficient and effective in your business.

Home furnishing retailer Rooms To Go simplifies the consumer shopping experience by offering completely designed room packages. Given that the company offers variations and add-ons, it wanted to better understand its customers" purchase choices to streamline online customization options. Working with LunaMetrics to integrate Google Analytics Premium with BigQuery, Rooms To Go was able to identify the products that are often purchased together. Armed with these insights, the company has made customization on the site smarter and easier for users, which ultimately boosted sales and improved the overall shopping experience.

In summary, Google Cloud offers you ways to...

Spend less on ops and administration

Incorporate real-time data into apps and architectures

Apply machine learning broadly and easily

Become a truly data-driven company

Google Cloud

Spend less on ops and administration: We've "automated out" the complexity of building and maintaining data and analytics systems.

Incorporate real-time data into apps and architectures: To get the most out of data and secure competitive advantage.

Apply machine learning broadly and easily: We make it simple and practical to incorporate machine learning models within custom applications.

Create citizen data scientists: Transform your organization into a truly data driven company. Putting tools into hands of domain experts.

Module Review

Google Cloud

Module review

Google Cloud Platform is:

*(select **all** of the correct options)*

- ❏ Operated by Google on the same infrastructure it uses
- ❏ A set of modular services from which you can compose cloud-based applications
- ❏ Most cost-effective if you pre-purchase instances on a yearly basis
- ❏ A platform on which to host scalable and fast distributed applications

Module review answer

Google Cloud Platform is:

*(select **all** of the correct options)*

- ✓ Operated by Google on the same infrastructure it uses
- ✓ A set of modular services from which you can compose cloud-based applications
- ❑ Most cost-effective if you pre-purchase instances on a yearly basis
- ✓ A platform on which to host scalable and fast distributed applications

Resources

Google Cloud Platform	https://cloud.google.com/
Datacenters	https://www.google.com/about/datacenters/
Google IT security	https://cloud.google.com/files/Google-CommonSecurity-WhitePaper-v1.4.pdf
Why Google Cloud Platform?	https://cloud.google.com/why-google/
Pricing Philosophy	https://cloud.google.com/pricing/philosophy/

Google Cloud

Notes:

Google Cloud Platform is a different type of cloud. Google Cloud Platform is about more than just renting infrastructure.


cloud.google.com
Image by Connie Zhou