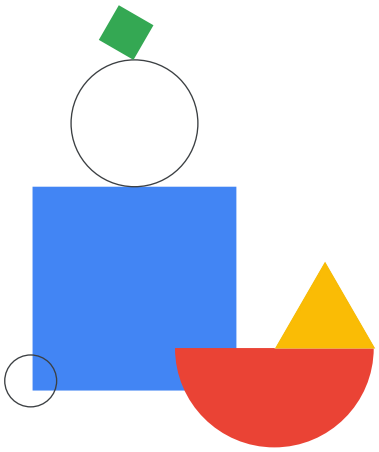
Google Cloud

# Introduction to Data on Google Cloud



# Agenda

01 Analytics Challenges Faced by Data Analysts

02 Big Data On-premise Versus on the Cloud

03 Real-world Use Cases of Companies Transformed Through Analytics on the Cloud

04 Google Cloud Project Basics




In this module, we will highlight analytics challenges faced by data analysts and compare big data on-premise versus on the Cloud.

We'll learn from real-world use cases of companies transformed through analytics on the Cloud, and then navigate Google Cloud Project basics.


# Analytics Challenges Faced by Data Analysts



Let’s start by looking at common analytics challenges faced by you, the data analyst.

# Data analysts face query, infrastructure, and storage challenges

My queries are taking way **too long** to run and is stalling my analysis.

We're a data department, not an **infrastructure** department. Maintaining and upgrading our own servers is unsustainable.

We can only **afford to store a subset** of the data our business generates.

I have no easy way to **combine and query** all the data I've collected.

My on premise clusters **aren't scaling** with my analysis.

We don't have a **central data** analytics warehouse or set of tools.



Two of the most common barriers a data analyst will run into is either too much data or data that is not connected together.

Green = Querying

Orange = Infrastructure

Blue = Storage


# Big Data On-Premise Versus on the Cloud

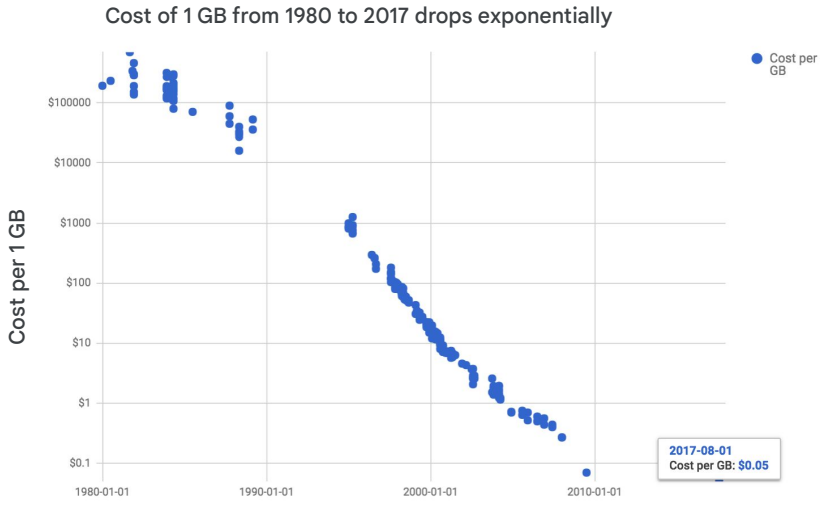


Next up, we will compare on-premise servers and infrastructure with the Google Cloud.

# Reasons why Google Cloud is used for data analysis

- Storage is cheap
- Focus on queries, not Infrastructure
- Massive scalability

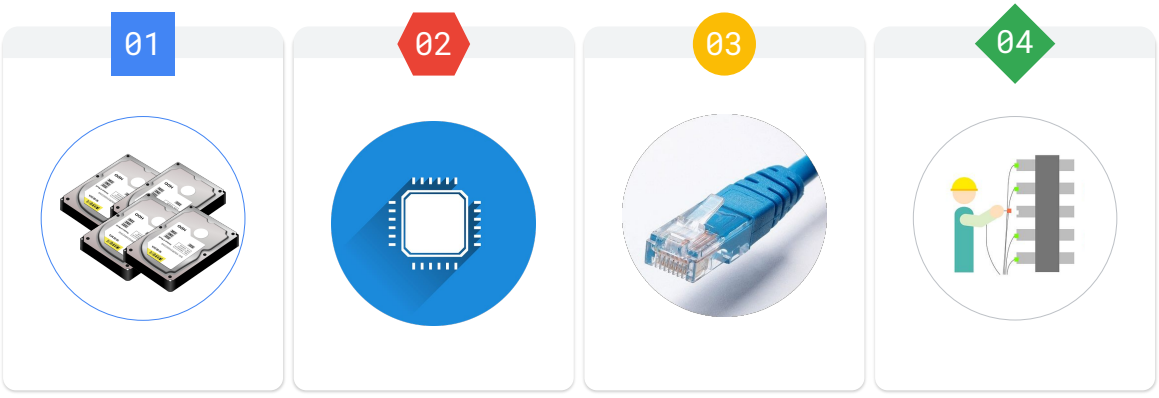
# The cost of 1GB of storage has dropped dramatically






In December 1981, the cost of a 10 MB Hard Drive was $300,000. In August of 2017, a 1 TB Hard Drive is roughly $50. Storage is cheap.

# Traditional big data platforms require an investment in infrastructure





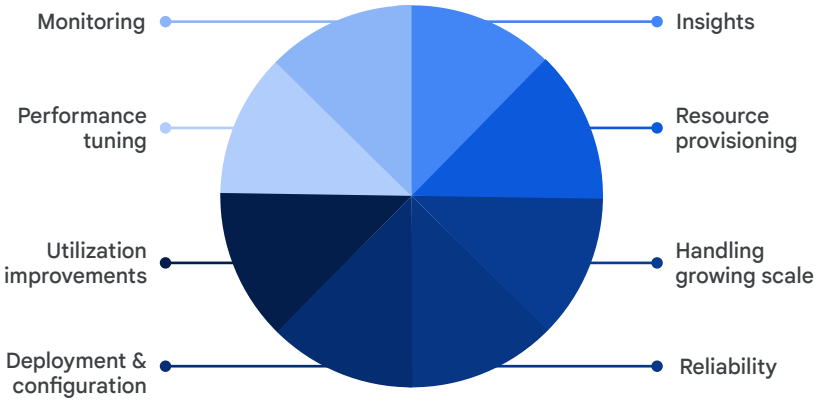

Although hard drives are cheap, they're not the only thing you need to have to query big data.

You will also need:
1 - Storage
2 - Computing Power
3 - Networking
4 - Admin and hardware teams to maintain and upgrade your infrastructure

(Not to mention software and software license costs)

# Typical big data processing

Time to understanding





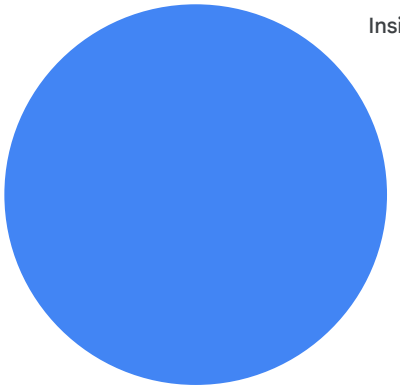

Then what? How do you find the needle in the haystack of data? Traditional DB not up to the challenge of data on this scale.

Most BigData today is Map Reduce - a research paper originally published by Google back in 2004.

But if you look at most big data projects and really look and where time and money is being spent, you see that most of the time isn't spent getting insights from the data; it's spent on the care and feeding of the machinery - managing infrastructure, manipulating data, monitoring and performance.

## Big data with Google: Focus on insights, not infrastructure

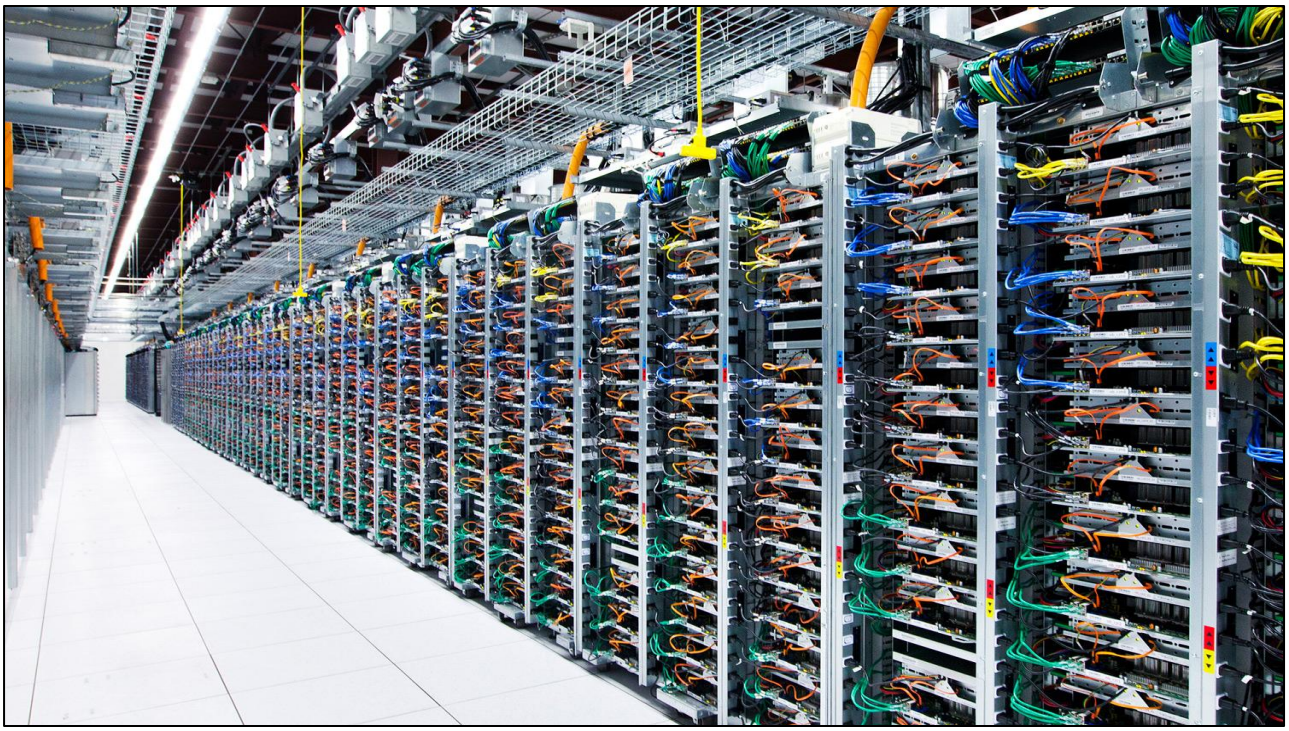
Time to understanding






We realized this ourselves at Google several years ago, so we started developing systems that let us:

1) **scale with your data growth** even as it explodes
2) are **managed** so that you aren't wasting time on dealing with all of the underlying complexities
3) are just generally **magically awesome** so you can get back to **data insights, and not data mgmt**



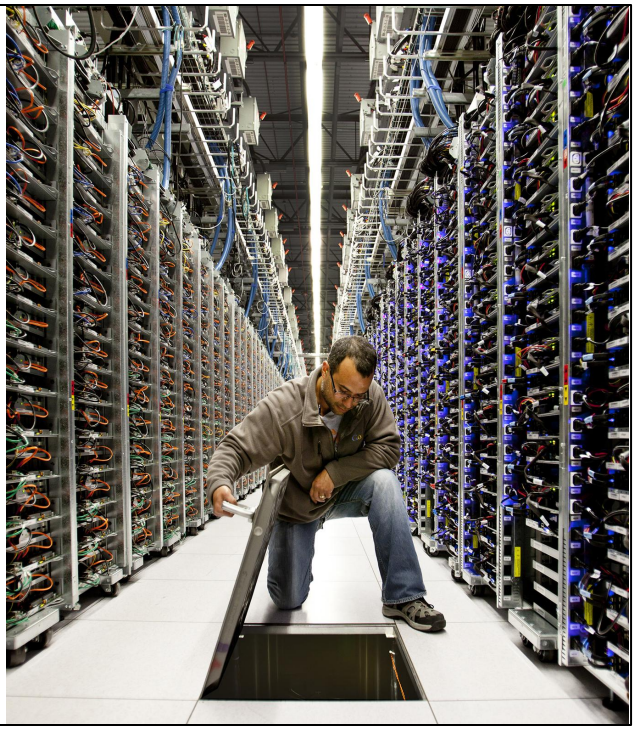
Don’t build it yourself, leverage Google’s massive infrastructure.

“

“[Google's] ability to build, organize, and operate a huge network of servers and fiber-optic cables with an efficiency and speed that rocks physics on its heels.

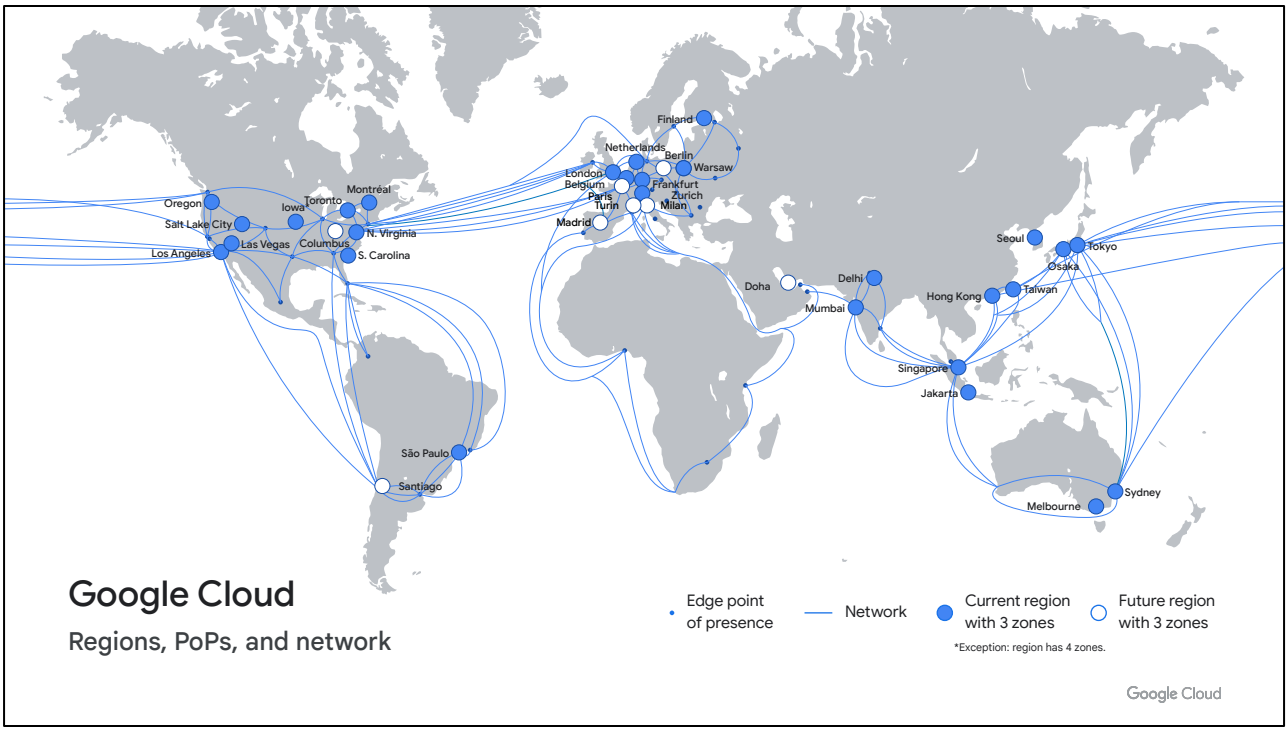
This is what makes Google Google: its physical network, its thousands of fiber miles, and those many thousands of servers that, in aggregate, add up to the mother of all clouds.”

- Wired



Google faced big data problems and as a result

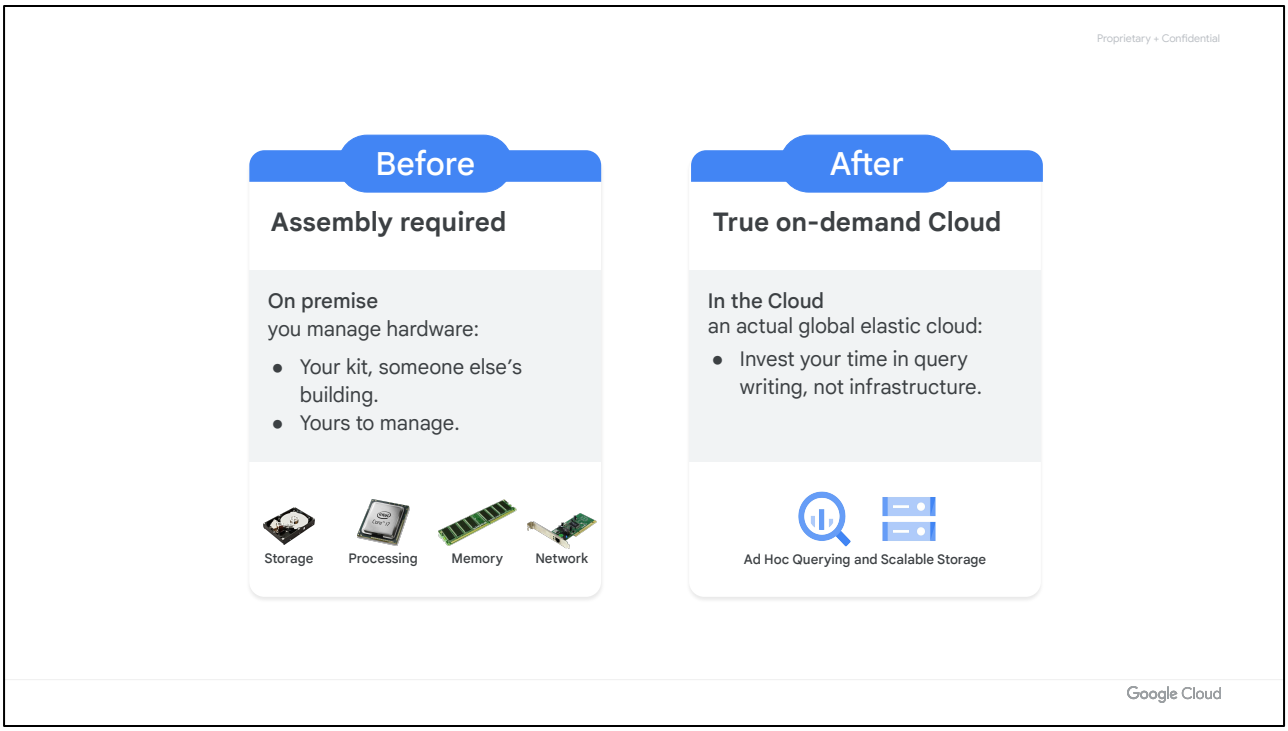
WIRED Article:
https://www.wired.com/2012/10/ff-inside-google-data-center/




The number of regions and zones is continually increasing. An up-to-date view is here: https://cloud.google.com/about/locations/.

The network and points of presence are here: https://cloud.google.com/about/locations/?tab=network.

The edge points of presence are the locations where Google networks are connected with internet service providers to allow users to connect. The Google Cloud network strongly distinguishes Google from other cloud service providers. The points of presence allow Google to provide very low latency network performance.

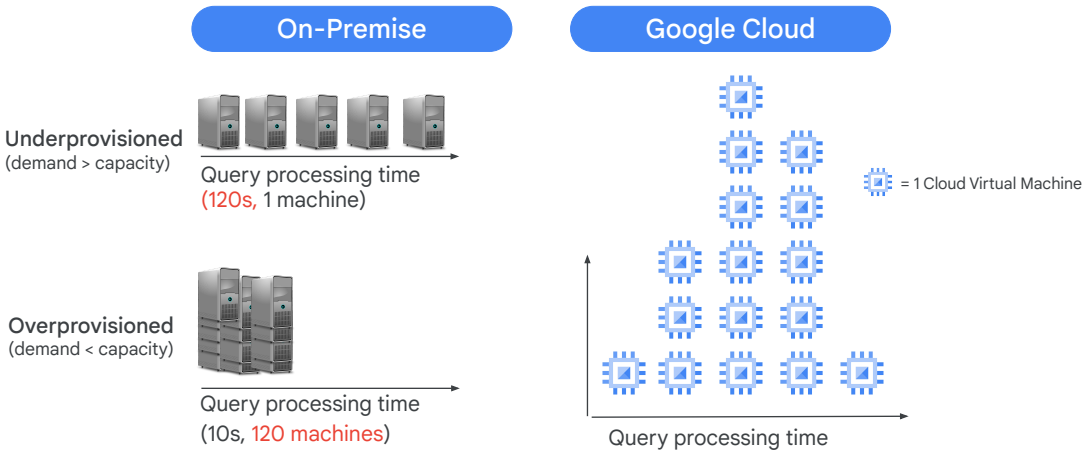



The concept of cloud computing began with colocation. Instead of operating your own data center, you rented space in a colocation facility. This was the first wave of outsourcing IT. With colocation, the transfer of ownership was minimal - you still owned the machines and you maintained them. Traditionally, 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 datacenters - 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 datacenter, 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.

# Google Cloud enables on-demand scalability





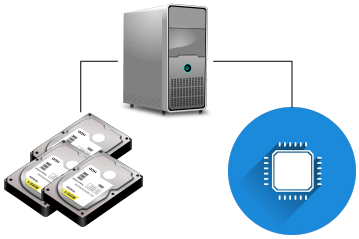

Cloud Economies of Scale

You still pay for power for unused CPUs on premise (especially if you have storage HDDs and CPUs on the same servers)

You can use the same processing power as having 150 machines to process 1 TB in 1 second just as if you had one machine process the query in 150 seconds. And the scaling up and down is automatic so there is no guesswork for demand -- if you need more, the cloud scales up elastically and then goes away when not in use.

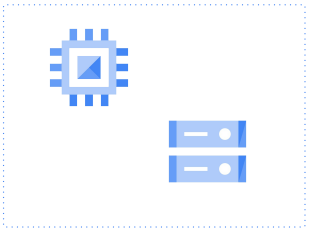
# Separation of storage and computing power enables efficient resource allocation

**On-Premise**



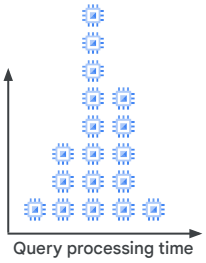
Pay for ability to use processing power even when no queries running

**Google Cloud**



Pay for only the resources you are using and no more

# BigQuery scales automatically and you only pay for what you use




Fully-managed infrastructure scales to process faster...

$

...and you only pay for bytes processed + storage



Key message: BigQuery scales automatically


# Real-World Use Cases of Companies Transformed Through Analytics on the Cloud



In this lesson, we will examine real-world use cases of companies where moving compute and storage to the cloud has enabled them to scale their data analysis.

## Store petabytes of data

> *[Our mission is] to **make our data so intelligent** it has the answer before the question is even asked. It was a stretch goal but essentially one that means we have to **capture all the data** we produce - both now and in the future.*

**Dan Nelson**
Head of Data - Ocado

ocado



“Ocado is the world's largest online-only grocery retailer, shipping over 150,000 orders a week or 1.1M items a day, and with a delivery area which covers over 70% of British households.”
https://www.youtube.com/watch?v=dUjze7LyDwI

2015
190,000 orders a week, or one every 2.5 seconds
Over 20 million items delivered per week

Data growth
10M routing decisions each day per Customer Fulfillment Centers

# Focus on your business, not hardware

> *The less time that we can spend solving problems that are already solved, like scaling,... **the more time and energy we can spend on turning our data into value.***
>
> Nicholas Harteau
> VP Infrastructure - Spotify

Spotify



Quote from Nicholas Harteau:
https://youtu.be/pm6KZ2xicFA?t=57s

Additional context:
https://www.forbes.com/sites/alexkonrad/2016/02/29/why-spotify-really-chose-google-cloud/#7c8658223ee4

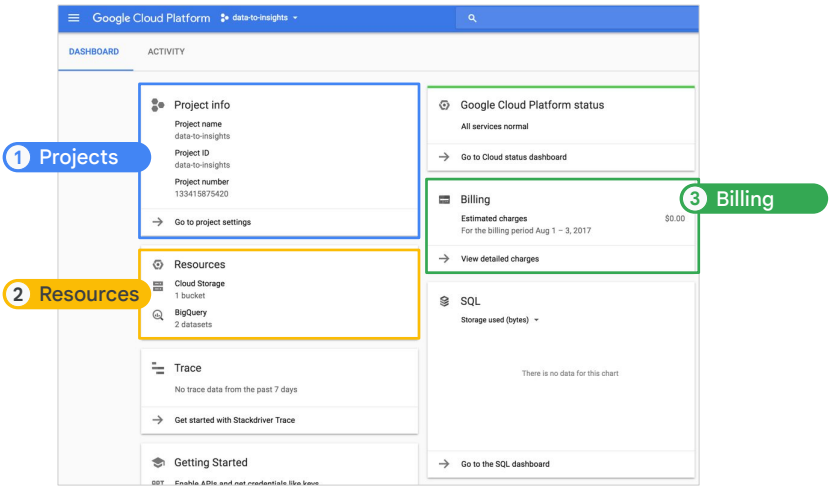

# Google Cloud Project Basics



Now let's dive into the specifics of Google Cloud and a few of the core pieces that you will be building on for the rest of this course.

# Navigate the Google Cloud using the dashboard

1. Projects
2. Resources
3. Billing

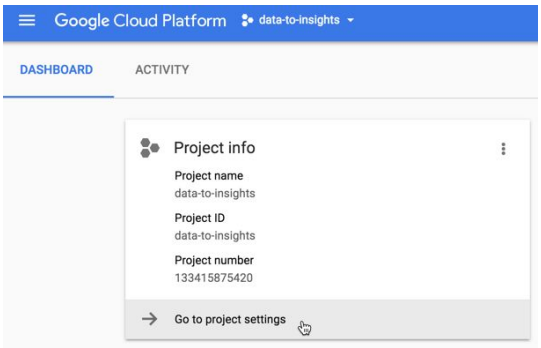


# Projects

Projects organize and govern your activities in the cloud

1 Projects

- **Navigate and launch cloud tools** for your project by exploring the Products and Services menu.
- **Work collaboratively** by adding project users through IAM (Identity and Access Management).
- **Authorize Tools and Apps** through the API manager.






Qwiklabs will auto create and provision project ids for you. Do not create one.
https://cloud.google.com/resource-manager/docs/cloud-platform-resource-hierarchy#projects

**Background on: The Project resource**

Recall that the project resource is the base level organizing entity. Unlike an Organization, a project is required to use Google Cloud, and forms the basis for creating, enabling and using all Cloud services, managing APIs, enabling billing, adding and removing collaborators, and managing permissions.

All projects consist of the following:

- Two identifiers:
  1. Project ID, which is a unique identifier for the project.
  2. Project number, which is automatically assigned when you create the project. It is read-only.
- One mutable display name.
- The lifecycle state of the project; for example, ACTIVE or DELETE_REQUESTED.
- A collection of labels that can be used for filtering projects.
- The time when the project was created.

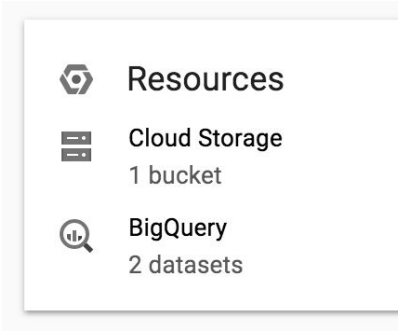
# Resources

## Resources are what you are using in the cloud

2 Resources

Commonly used by data analysts:

- **Storage** in Google Cloud Storage
  - Example: You use a Bucket for uploading large CSV files to ingest later for analysis.
- **Datasets** in BigQuery
  - Example: You perform analysis on raw data and create a brand new dataset.

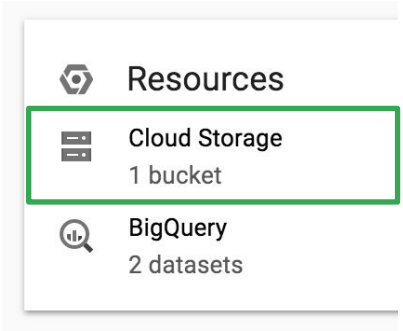


# Resources

The Cloud Storage bucket is your go-to for scalable storage

2 Resources

- Buckets are scalable containers that hold your data.
- You can create and upload files to your buckets within the Google Cloud console.






Cloud Storage buckets are ideal to use as a staging area for raw data.
https://cloud.google.com/storage/docs/key-terms#buckets

# Billing

## You are billed for the resources you use

3 Billing

Commonly used by data analysts:

- **Storage** in Google Cloud Storage
  - Billed for Bucket Storage
- **Datasets** in BigQuery
  - Billed for Query processing
  - Billed for Table Storage


*After this course, try exporting BigQuery logs using this tutorial to recreate the above Data Studio billing dashboard*

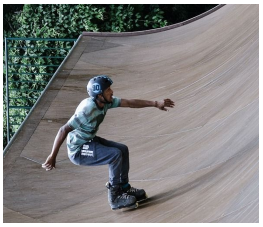


We will be covering pricing and how you can calculate query costs later in this course.

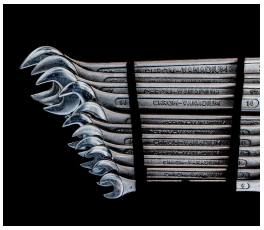
Walkthrough guide on enabling billing to look at BigQuery usage: https://medium.com/google-cloud/visualize-gcp-billing-using-bigquery-and-data-studio-d3e695f90c08

# Module summary: Scale with Google Cloud

Overcome query speed, infrastructure, and cost challenges.



Efficiently scale your compute and storage needs.



Manage and monitor your project resources in one place.

Evangelize data analysis in your organization.



Wrapping up this module, let's review some of the key points about the Google Cloud. We've covered the common challenges data analysts face and how the cloud offers scalable fully-managed tools for any analyst to use. In the following modules, we will introduce these specific tools like BigQuery, Data Studio, and Dataprep and how they build on the compute and storage scalability of the Google Cloud.