This this module, we will highlight the five common tasks of any data analyst and map those to their respective tools in the Google Cloud Platform

After that, we'll head into a demo showing BigQuery operating on billions of records. Following the demo, we will explore the BigQuery featureset and end with a discussion and comparison of data analysts, data scientists, and data engineers.

A data analyst is responsible for analyzing and gleaning insights from data

Ingest

Get data in.

Transform

Prepare, clean, and transform data.

Store

Create, save, and store datasets.

Analyze

Derive insights from data.

Visualize

Explore and present data insights.

Google Cloud

Challenges in each task prevent data analysts from getting to scalable insights

Ingest

Get data in.

Challenges

Data Volume
Data Variety
Data Velocity

Transform

Prepare, clean, and transform data.

Challenges

Slow Exploration
Slow Processing
Unclear Logic

Store

Create, save, and store datasets.

Challenges

Storage Cost
Hard to Scale
Latency Issues

Analyze

Derive insights from data.

Challenges

Slow Queries
Data Volume
Siloed Data

Visualize

Explore and present data insights.

Challenges

Dataset Size
Tool Latency

Google Cloud

Choosing the Right Tools

How many of you have used business intelligence and data analysis tools? There is no one-size fits all solution for all of those common data analysis tasks but rather a toolkit.

Let’s explore the big data toolkit available on the Google Cloud Platform.

https://unsplash.com/search/photos/tools?photo=iCtJF-A5hvs
Photo by Jeff Hopper on Unsplash

Google Cloud Platform offers scalable big data tools to overcome data challenges

Ingest

Get **petabytes** of data in from a **variety of formats**.

BigQuery Storage (import)

Transform

Prepare, clean, and transform data **quickly and easily**.

BigQuery Analysis (SQL)

Cloud Dataprep (preparation)

Store

Create, save, and store datasets **inexpensively**.

Cloud Storage (buckets)

BigQuery Storage (tables)

Analyze

Derive insights from data **at scale and without managing servers**.

BigQuery Analysis (SQL)

Visualize

Explore and present **interactive and impactful** data insights.

Google Data Studio

Third-party Tools (Tableau, Looker, Qlik)

Google Cloud

Google Cloud Platform big data tools:
https://cloud.google.com/solutions/big-data/

Module 2

Big Data Tools Overview

In this module we will:

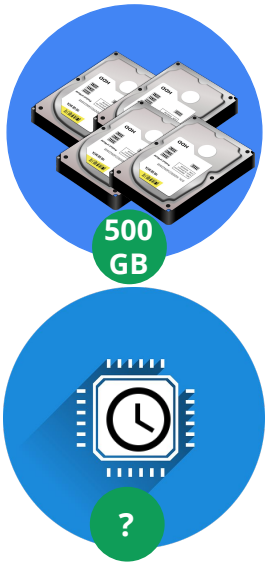
- Walkthrough Data Analyst Tasks, Challenges, and Introduce Google Cloud Platform Data Tools
- **Demo: Analyze 10 Billion Records with Google BigQuery**
- Explore 9 Fundamental BigQuery Features
- Compare GCP Tools for Analysts, Data Scientists, and Data Engineers

Google Cloud

BigQuery Demo using 10 Billion+ rows

```
#standardSQL

# Demo processing 10 Billion Wikipedia records
SELECT
  language,
  title,
  SUM(views) AS views
FROM
  `bigquery-samples.wikipedia_benchmark.Wiki10B`
WHERE
  title LIKE '%Google%'
GROUP BY
  language,
  title
ORDER BY
  views DESC;
```



Google Cloud

Open the following saved query:
https://bigquery.cloud.google.com/savedquery/133415875420:55739a3b2c5941a68f44ef8da0ff2c37

Google BigQuery has numerous Public Datasets that anyone can query. One of these is all public wikipedia page metadata.
Let's run a SQL query to see how fast we can scan and process 10 Billion rows looking for the word "Google" in the Wikipedia Page Title.

On BigQuery it's a best practice to use #standardSQL because it is standards compliant (ANSI 2011) and has significant performance advantages that we will cover later. You can enable standardSQL in the options and/or write #standardSQL as a comment in the first line of your query.

Point out the amount of data the query will process by ***clicking on the validator*** *(around 500 GB)*

Run the Query

Point out the processing time *(should be around 12 seconds)*

Point out the results of the Query *showing the main "Google" wikipedia page has been* ***viewed over 9 Billion times***

Lastly, click on the query Explanation button to show how many input and output rows.

Poll the class: *What do the 10 Billion input rows signify? What about the resulting 200K+ output? The 10 Billion rows correspond to the count of Wikipedia pages and the 200K+ final result is the count of pages that contained the word "Google" somewhere in the title.*

Poll the class: *Do you think the query will run faster, slower, or depends on the resources if we re-ran it right now?*

Re-Run the Query

The same query executed much faster as it is now pulling from ***query cache****. We'll discuss this more in the Performance section of the course.*

Last Poll: *Do you think it matters that we spelled Google with a capital "G" when matching against title? Is SQL case sensitive? In the LIKE operator, yes!*

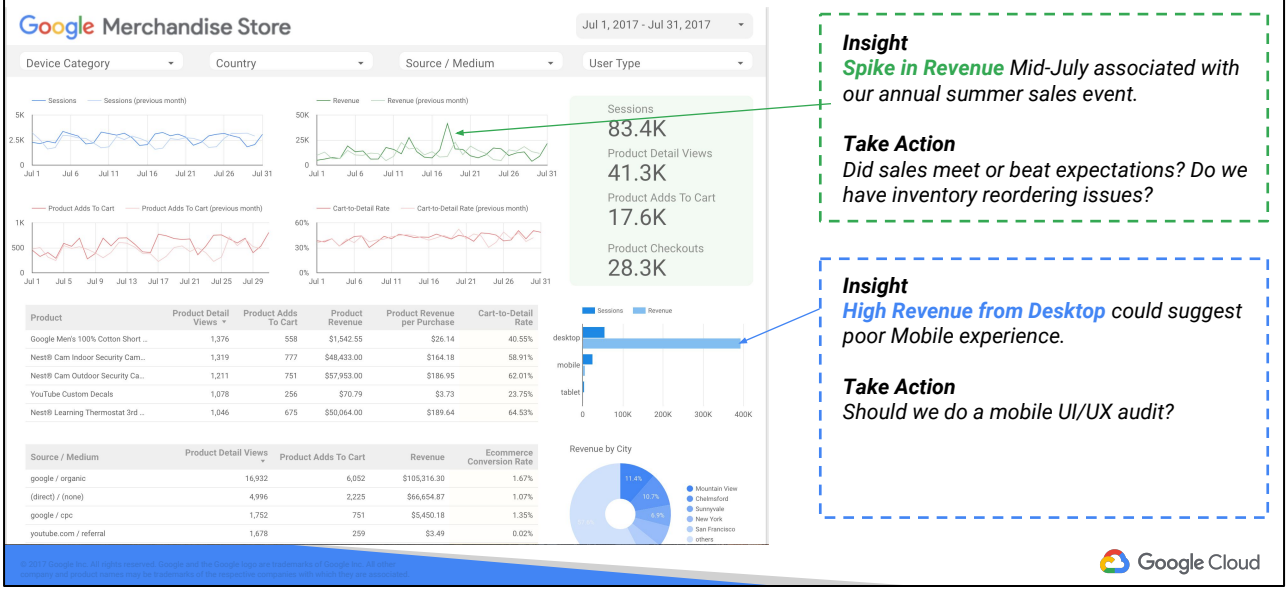
Re-Run the Query changing '%Google%' to '%google%' in the LIKE operator

The results are wildly different now. We will review pitfalls like these in our SQL exercises as part of this course. It is your responsibility as a data analyst to understand what your query is actually doing.

```
# Another Query you could try if you dont like Wikipedia
SELECT
hour,
AVG(fare) AS avg_fare
FROM (
SELECT
EXTRACT(HOUR
FROM
trip_start_timestamp) AS hour,
fare
FROM
`bigquery-public-data.chicago_taxi_trips.taxi_trips`
WHERE pickup_community_area = 32
)
GROUP BY
hour
ORDER BY
avg_fare DESC
```

Finds taxifare by hour for picks that happen in downtown Chicago. Using area = 76 will give you O'Hare

Explore and visualize large datasets with Data Studio



Insight
Spike in Revenue *Mid-July associated with our annual summer sales event.*

Take Action
Did sales meet or beat expectations? Do we have inventory reordering issues?

Insight
High Revenue from Desktop *could suggest poor Mobile experience.*

Take Action
Should we do a mobile UI/UX audit?

Google Cloud

Another tool that we will be covering in this course is Data Studio which can connect to BigQuery to visualize your insights.

Here take a look at a merchandise dashboard and the highlighted insights and recommended actions.

Link to Data Studio example merchandise store dashboard: https://datastudio.google.com/c/u/0/org/UTgoe29uR0C3F1FBAYBSww/reporting/0B2-rNcnRS4x5UG50LTBMT0E4aXM/page/nQN

Module 2

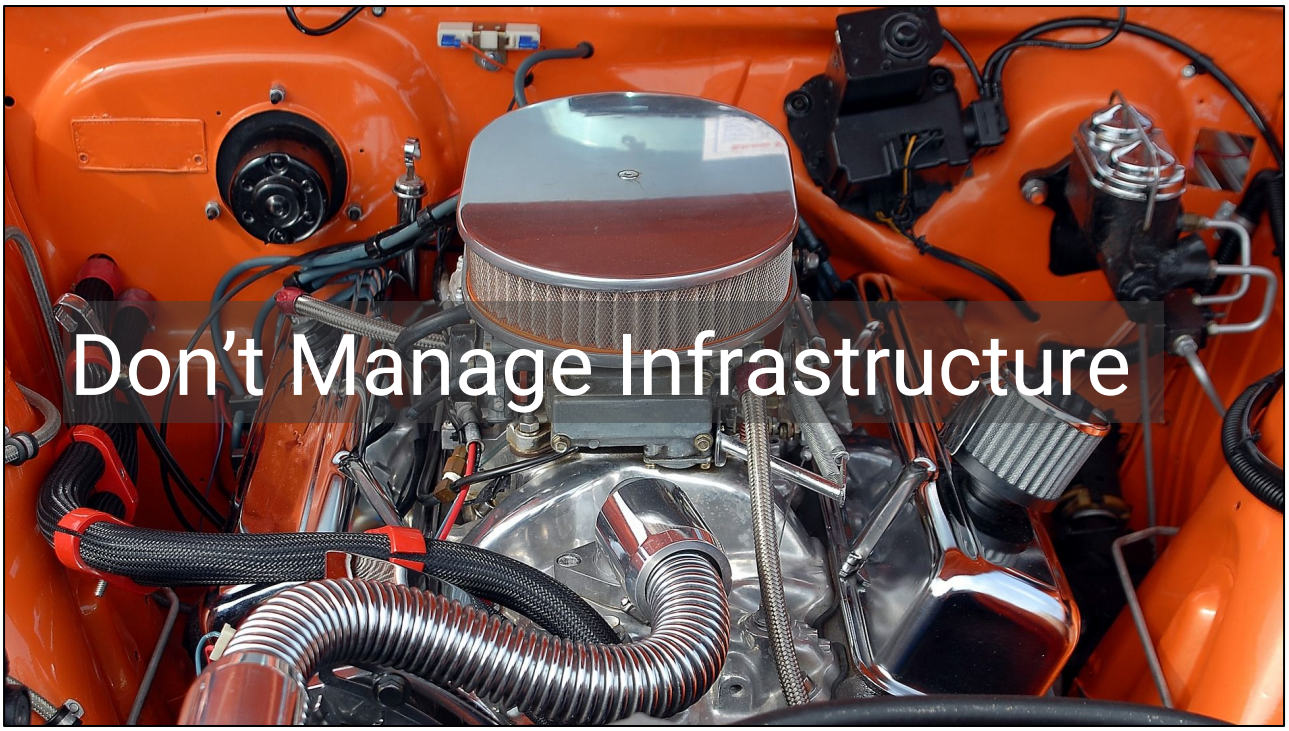
Big Data Tools Overview

In this module we will:

- Walkthrough Data Analyst Tasks, Challenges, and Introduce Google Cloud Platform Data Tools
- Demo: Analyze 10 Billion Records with Google BigQuery
- **Explore 9 Fundamental BigQuery Features**
- Compare GCP Tools for Analysts, Data Scientists, and Data Engineers

Google Cloud

In this section we will explore the core featureset of BigQuery that enables you to query petabyte-scale datasets within tens of seconds.



With BigQuery you get the benefit of Google datacenter backed infrastructure that is fully managed. That means no-operations, no car mechanics, and no debating over whether your engine is too small or too big for the job.

The best part is that you don’t need to spend your time optimizing the specific hardware, and networking. You can focus on just using the engine and writing queries for insights.

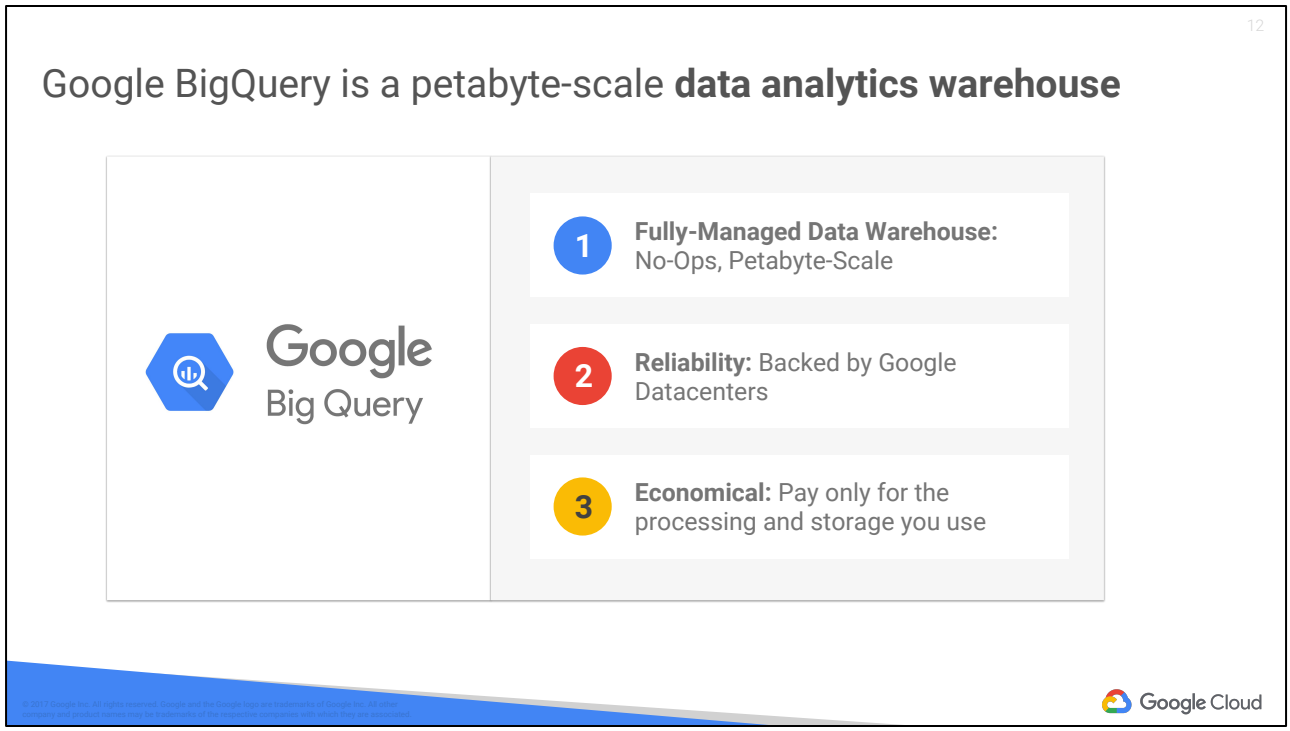
Now let’s expand on specific features of BigQuery (next slide)

Image (engine) cc0: https://pixabay.com/en/car-engine-engine-motor-car-1738309/

Your job as a data analyst is to focus on asking great questions of your dataset and hunt down interesting insights.

This much like a traveller or an explorer with a map -- all your focus should be on finding interesting places to see.

https://unsplash.com/search/photos/map?photo=kZO9xqmO_TA
Photo by Annie Spratt on Unsplash



BigQuery background
https://cloud.google.com/bigquery/

Fully-managed, enterprise data warehouse

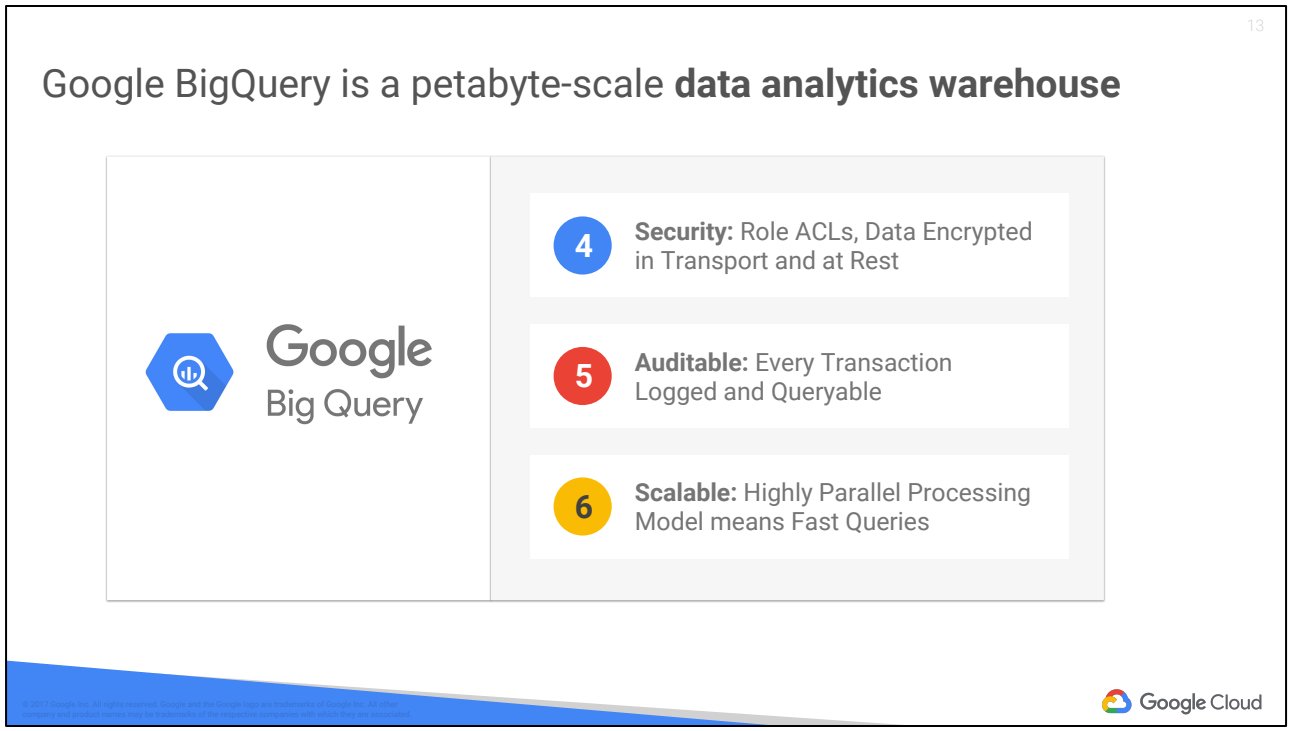
- Provides **near real-time interactive analysis** of massive datasets
- Runs on Google's fully managed, secure, high-performance infrastructure
- "NoOps" - No administration for performance and scale

Reliability

- Data replicated across multiple data centers

Economical

- Only pay for storage and processing used



Security

Secured through Access Control Lists (ACLs) and Identity and Access Management (IAM)

Data is encrypted in transport and at rest

Auditable

Google Cloud Audit Logs track Admin Activity and Data Access

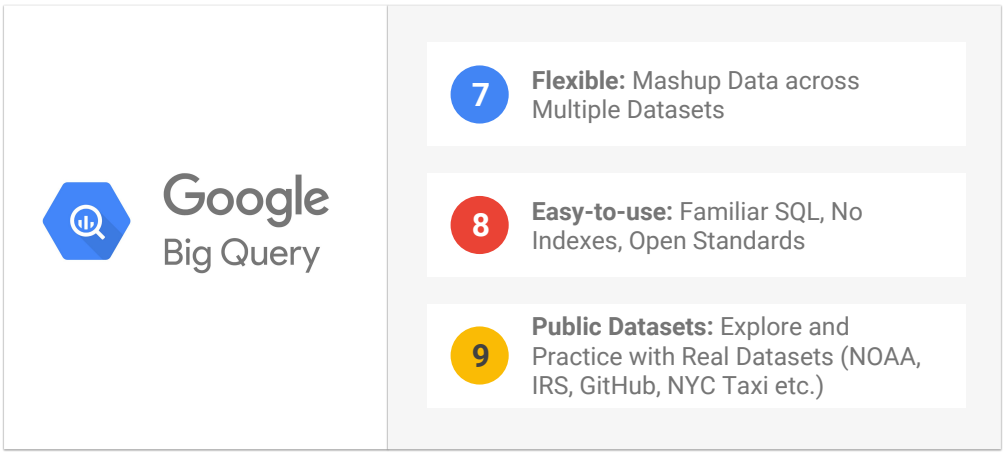
Immutable logs - “who did what, where, and when?” in BigQuery

Scalable

Virtually unlimited data storage and processing power

Highly parallel/distributed process model

Google BigQuery is a petabyte-scale **data analytics warehouse**



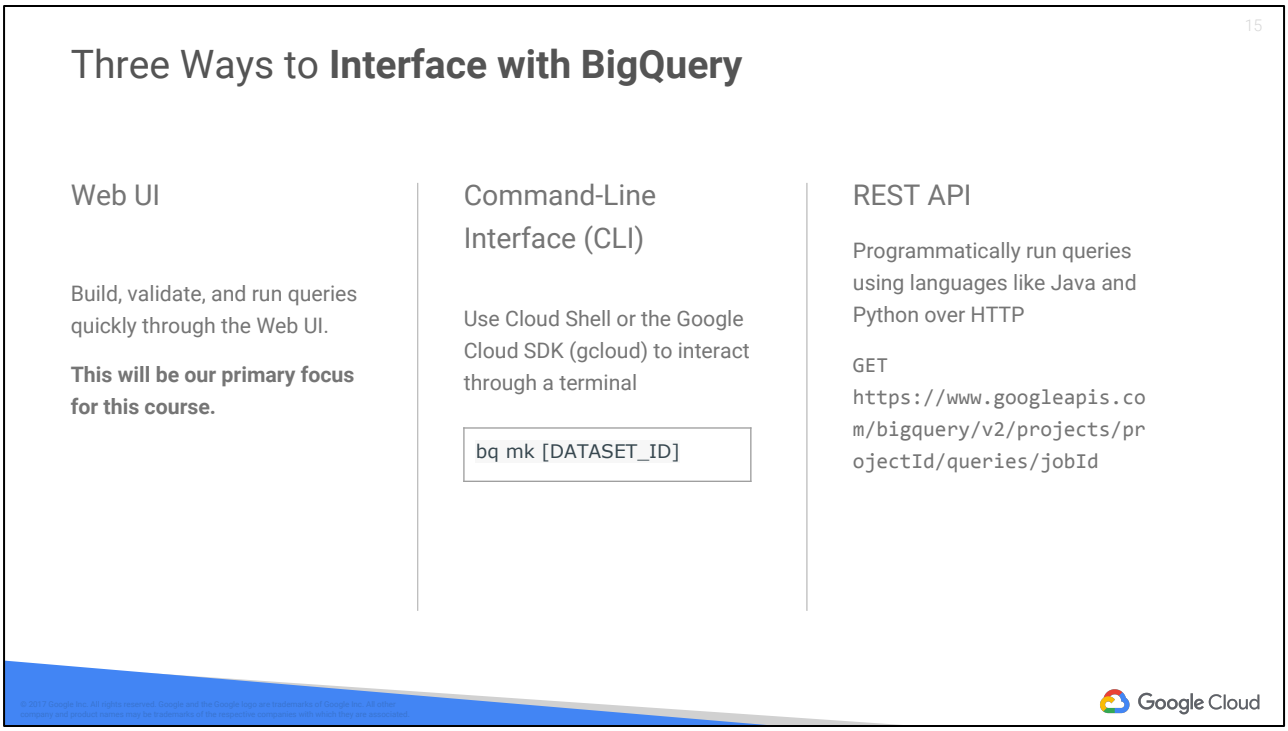
Google Cloud

Flexible

- Streaming ingestion:100K rows/sec per table for real-time data
- Data mashup: JOIN across diverse datasets/projects

Easy to use

- Data stored in denormalized **tables** (simple schemas)
- Columnar storage for high performance
- Requires no indexes, keys, or partitions
- Familiar SQL interface and intuitive UI
- Nested and repeated field support for schema flexibility
- Supports open standards - Analysts can use preferred tools



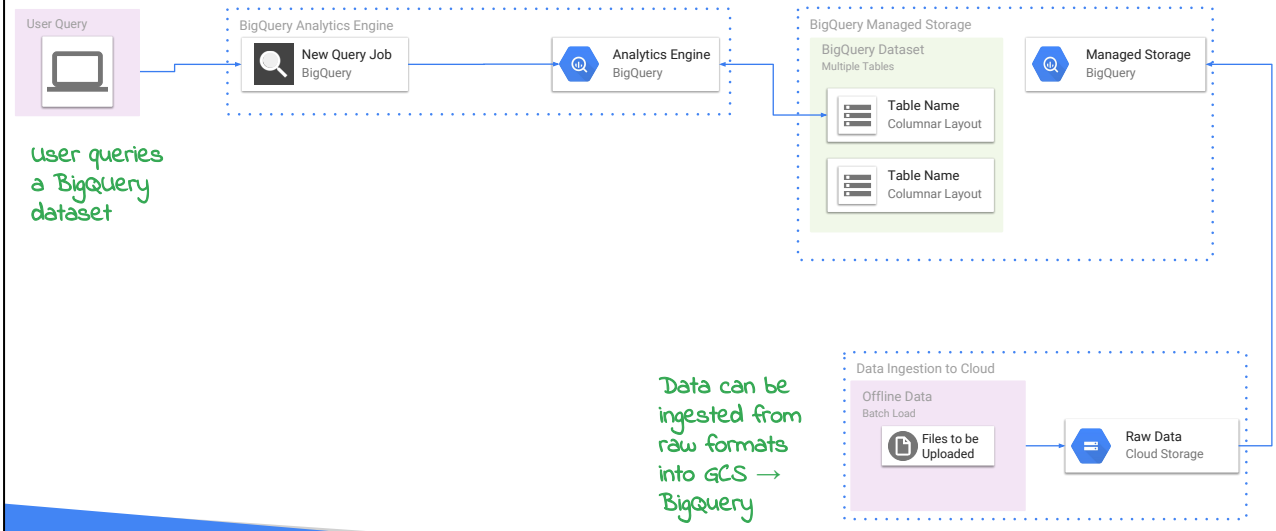
There are three ways to interact with BigQuery – the web UI, the command-line interface (CLI), and the REST API.

Since this course focuses on using BigQuery for data analysis, you spend most of the course using the web UI. In this lab you learn how to examine tables, quickly build queries a few simple mouse clicks, and validate/determine how much the query will process, along with query caching and query priorities.

You also use the CLI to execute queries and explore BigQuery features. The CLI contains a robust set of commands that provide you the flexibility to run commands and queries interactively.

Finally, the REST API is the programmatic interface that programming languages like Java and Python use to communicate with BigQuery. The service receives HTTP requests and returns JSON responses. Both the web UI and the CLI use this API to communicate with BigQuery. Note that the REST API is beyond the scope of this course.

Creating and Querying Datasets: BigQuery Terminology



Google Cloud

Google BigQuery is actually two services in one

Google Big Query

BigQuery Managed Storage

Fully-managed and ***scalable data storage*** that is based on the same technology that stores Google's product data (ads, gmail etc.)

BigQuery Analysis

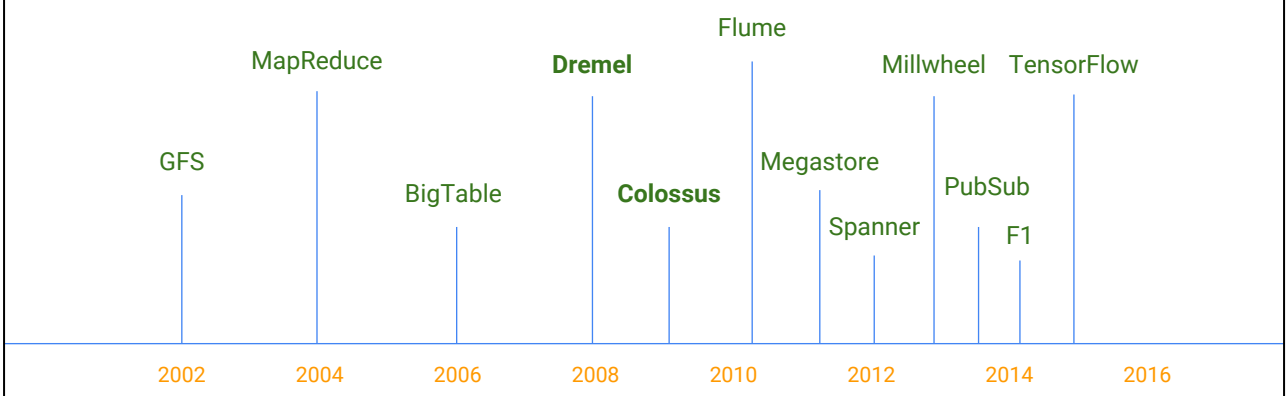
Fast massively parallel ***SQL Engine*** based on Google's own internal Dremel query engine technology

Google Cloud

You don't see the managed storage piece - it just works behind-the-scenes

- Replicating your data
- Mapping which datacenters (and servers) have which pieces of your data

Google innovates data technologies

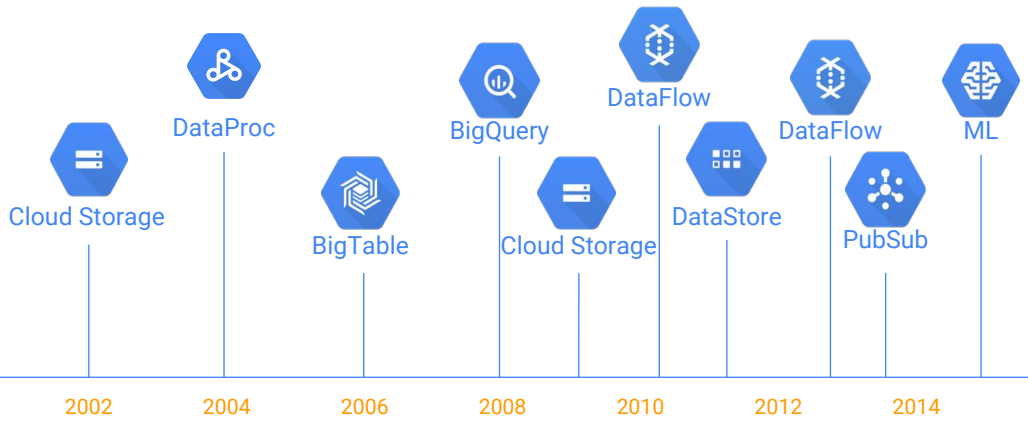


Google Research Publications referenced are available here: http://research.google.com/pubs/papers.html
The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines, 2009
http://research.google.com/pubs/pub35290.html

Google Cloud

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.

Google Cloud Platform opens up that innovation to you



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

Google Cloud

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 where 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.

Module 2

Big Data Tools Overview

In this module we will:

- Walkthrough Data Analyst Tasks, Challenges, and Introduce Google Cloud Platform Data Tools
- Demo: Analyze 10 Billion Records with Google BigQuery
- Explore 9 Fundamental BigQuery Features
- **Compare GCP Tools for Analysts, Data Scientists, and Data Engineers**

Google Cloud

In this last section, we will compare the roles and tools used by data analysts, data scientists, and data engineers.

Each data-related role uses a different suite of tools

Roles:	**Data Analyst**	**Data Scientist**	**Data Engineer**
What they do:	*Derive data insights from queries and visualization.*	*Analyze data and model systems using statistics and machine learning.*	*Designs, builds, and maintains data processing systems.*
Background:	*Data analysis using SQL*	*Statistical analysis using SQL, R, Python*	*Computer Engineering*
GCP Tools Used:	 Google Data Studio		

Google Cloud

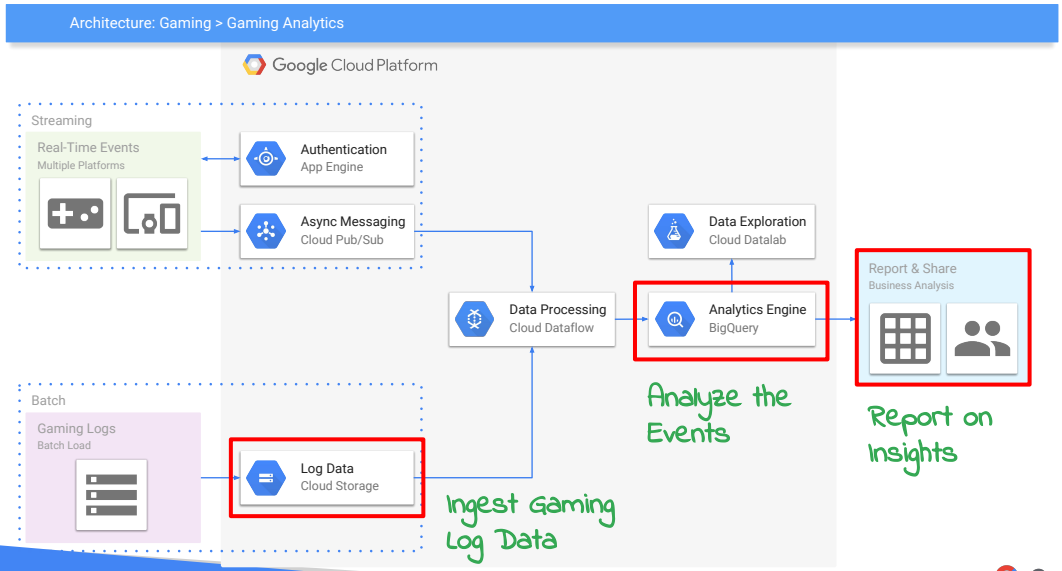
Spotlight on Certifications and Additional Courses
https://cloud.google.com/certification/data-engineer

Data Analyst
Cloud Storage
Google BigQuery
Cloud DataPrep
Google Data Studio

Data Scientist
Cloud DataLab
Google BigQuery
Cloud ML Engine

Data Engineer
Compute Engine
Cloud Storage
DataProc DataStore
DataFlow Cloud SQL
BigTable Spanner

End-to-end gaming analytics example highlighting GCP tools



Additional background on the life of a BigQuery Query:
https://cloud.google.com/blog/big-data/2016/01/anatomy-of-a-bigquery-query
https://docs.google.com/presentation/d/1vjm5YdmOH5LrubFhHf1vlqW2O9Z2UqdWA8biN3e8K5U/edit#slide=id.p99

Summary: Review data analyst tasks and tools

Reviewed Data Analyst tasks: Ingest, Transform, Store, Analyze, and Visualize Data

Data Analysts will use Cloud Storage, BigQuery, Cloud Data Prep, and Google Data Studio

Explored the 9 Features that make BigQuery is a Petabyte-Scale Data Analytics Warehouse

Compared Data Analysts, Data Scientists, and Data Engineers

Google Cloud

In this module, we covered the lifecycle of data analyst tasks and mapped each task to the right tools to use on the Google Cloud Platform. Then we demo'd BigQuery, the petabyte-scale data analytics warehouse, and covered it's core featureset. Lastly, we compared data roles and toolsets used by data analysts, data scientists, and data engineers. And while this course is targeted to data analysts, it will provide a clear ramp into more advanced tools and topics that are covered in other Google Cloud courses like Data Engineering.

Next up, let's continue our foray into BigQuery by practicing dataset exploration.

Image cc0 (data computer): https://cloud.google.com/solutions/big-data/
Image cc0 (tools): https://pixabay.com/en/tool-pliers-screwdriver-145375/
Image cc0 (hats): https://pixabay.com/en/hats-fedora-hat-manufacture-stack-829509/

Lab 1
Exploring your Public Dataset with Google BigQuery

Google Cloud

Lab 1 in Qwiklabs

BigQuery hosts 50+ public datasets for SQL practice

Public datasets include flights, taxi cab logs, weather recordings, and many more

Example SQL code is provided for practice

Google Cloud

U.S. Nonprofit Organizations may gain tax exempt status by filing their financial information each fiscal year through Form 990

All nonprofit Form 990s are open for **public inspection**

There are over 1 Million filings for us to analyze!

Nonprofits include: Teaching Organizations, Hospitals, Pet Relief, Environmental causes and more….

Logo
https://en.wikipedia.org/wiki/Internal_Revenue_Service#/media/File:Logo_of_the_Internal_Revenue_Service.svg

Types of nonprofits
https://www.charitynavigator.org/index.cfm?bay=content.view&cpid=1559

Image (aircraft) cc0: https://pixabay.com/en/aircraft-jet-landing-cloud-537963/
Image (taxi cab) cc0: https://pixabay.com/en/taxi-cab-traffic-cab-new-york-381233/
Image (weather) cc0: https://pixabay.com/en/lightning-storm-weather-sky-399853/
Image (doctor) cc0: https://pixabay.com/en/doctor-medical-medicine-health-563428/

Your course dataset is millions of U.S. charity tax filings




Nonprofit charities like hospitals, schools, animal shelters and more run programs

U.S. Internal Revenue Service (IRS) collects taxes from all individuals and businesses

Form 990

To subsidize charitable efforts, the US Internal Revenue Service (tax) allows these organizations to avoid paying tax by filing a special form annually

Google Cloud

U.S. Nonprofit Organizations may gain tax exempt status by filing their financial information each fiscal year through Form 990

All nonprofit Form 990s are open for **public inspection**

There are over 1 Million filings for us to analyze!

Nonprofits include: Teaching Organizations, Hospitals, Pet Relief, Environmental causes and more….

Logo
https://en.wikipedia.org/wiki/Internal_Revenue_Service#/media/File:Logo_of_the_Internal_Revenue_Service.svg

Types of nonprofits
https://www.charitynavigator.org/index.cfm?bay=content.view&cpid=1559

Images
https://pixabay.com/en/help-child-charity-voluntary-1265227/
https://unsplash.com/photos/FQ1L770x6l8
https://unsplash.com/photos/xuIlYVIbYIc
https://unsplash.com/photos/_h_weGa3eGo

Form 990 Preview

Form 990 is the special form that nonprofit organizations must file annually to receive their special tax exemption

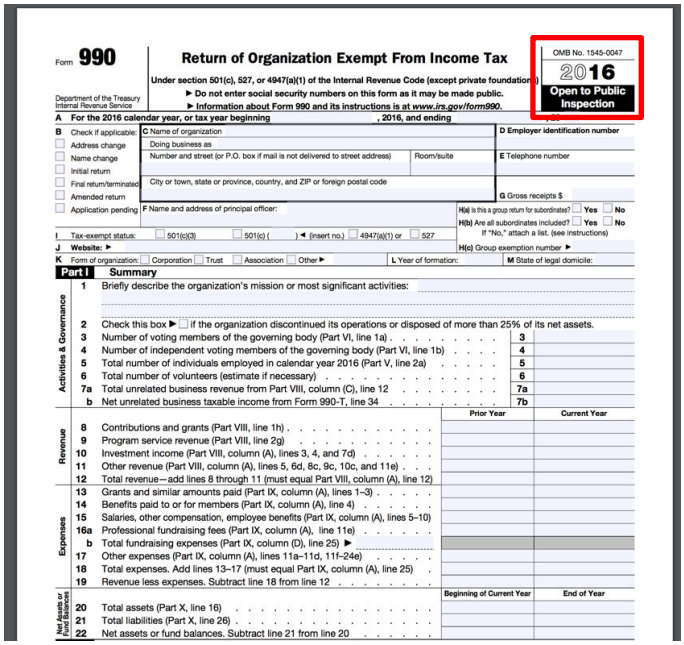
It contains key financial information and is then published publicly by the IRS for the public to also inspect

Employees

Revenue

Expenses

Assets



Form **990** — **Return of Organization Exempt From Income Tax**

Under section 501(c), 527, or 4947(a)(1) of the Internal Revenue Code (except private foundation)

Department of the Treasury
Internal Revenue Service

► Do not enter social security numbers on this form as it may be made public.
► Information about Form 990 and its instructions is at *www.irs.gov/form990*.

OMB No. 1545-0047
2016
Open to Public Inspection

A For the 2016 calendar year, or tax year beginning , 2016, and ending

B Check if applicable:
Address change
Name change
Initial return
Final return/terminated
Amended return
Application pending

C Name of organization
Doing business as
Number and street (or P.O. box if mail is not delivered to street address) Room/suite
City or town, state or province, country, and ZIP or foreign postal code

D Employer identification number
E Telephone number
G Gross receipts $

F Name and address of principal officer:

H(a) Is this a group return for subordinates? Yes No
H(b) Are all subordinates included? Yes No
If "No," attach a list. (see instructions)

I Tax-exempt status: 501(c)(3) 501(c) ( ) ◄ (insert no.) 4947(a)(1) or 527

J Website: ►

H(c) Group exemption number ►

K Form of organization: Corporation Trust Association Other ►

L Year of formation:

M State of legal domicile:

Part I Summary

Activities & Governance

1 Briefly describe the organization's mission or most significant activities:

2 Check this box ► if the organization discontinued its operations or disposed of more than 25% of its net assets.
3 Number of voting members of the governing body (Part VI, line 1a) 3
4 Number of independent voting members of the governing body (Part VI, line 1b) 4
5 Total number of individuals employed in calendar year 2016 (Part V, line 2a) 5
6 Total number of volunteers (estimate if necessary) 6
7a Total unrelated business revenue from Part VIII, column (C), line 12 7a
b Net unrelated business taxable income from Form 990-T, line 34 7b

		Prior Year	Current Year
Revenue	8 Contributions and grants (Part VIII, line 1h)		
	9 Program service revenue (Part VIII, line 2g)		
	10 Investment income (Part VIII, column (A), lines 3, 4, and 7d)		
	11 Other revenue (Part VIII, column (A), lines 5, 6d, 8c, 9c, 10c, and 11e)		
	12 Total revenue—add lines 8 through 11 (must equal Part VIII, column (A), line 12)		
Expenses	13 Grants and similar amounts paid (Part IX, column (A), lines 1–3)		
	14 Benefits paid to or for members (Part IX, column (A), line 4)		
	15 Salaries, other compensation, employee benefits (Part IX, column (A), lines 5–10)		
	16a Professional fundraising fees (Part IX, column (A), line 11e)		
	b Total fundraising expenses (Part IX, column (D), line 25) ►		
	17 Other expenses (Part IX, column (A), lines 11a–11d, 11f–24e)		
	18 Total expenses. Add lines 13–17 (must equal Part IX, column (A), line 25)		
	19 Revenue less expenses. Subtract line 18 from line 12		

		Beginning of Current Year	End of Year
Net Assets or Fund Balances	20 Total assets (Part X, line 16)		
	21 Total liabilities (Part X, line 26)		
	22 Net assets or fund balances. Subtract line 21 from line 20		

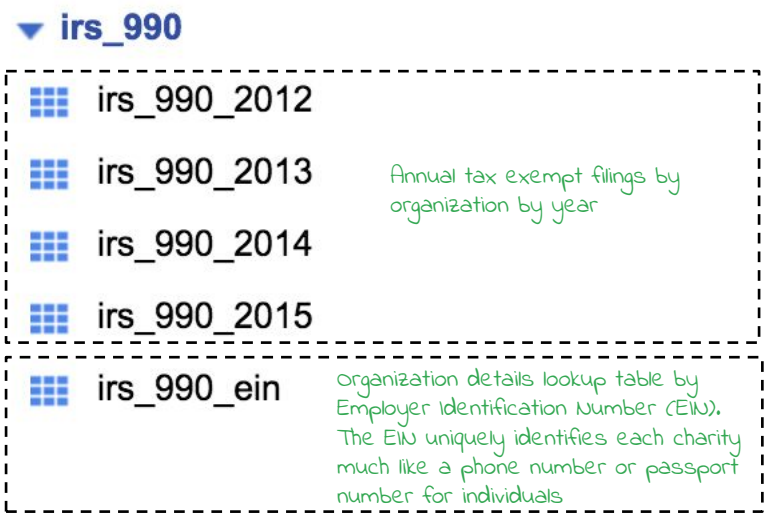
Google Cloud

https://www.irs.gov/pub/irs-pdf/f990.pdf

Logo
https://en.wikipedia.org/wiki/Internal_Revenue_Service#/media/File:Logo_of_the_Internal_Revenue_Service.svg

Example Form:
https://goo.gl/d92L8c

IRS BigQuery public dataset has two primary table types



Google Cloud

Access the course dataset:
https://bigquery.cloud.google.com/dataset/bigquery-public-data:irs_990

Exploring your Dataset with Google BigQuery

- Locate and Query the IRS_990 BigQuery Public Dataset
- Explore dataset and table metadata using the Google BigQuery UI
- Enable the Standard SQL dialect for your queries
- Perform basic stats and counts on data tables using Standard SQL in the Google BigQuery UI
- Find duplicate records in a data table using SQL

U.S. Internal Revenue Service (IRS) collects taxes from all individuals and businesses

Image (IRS Logo):
https://en.wikipedia.org/wiki/Internal_Revenue_Service#/media/File:Logo_of_the_Internal_Revenue_Service.svg