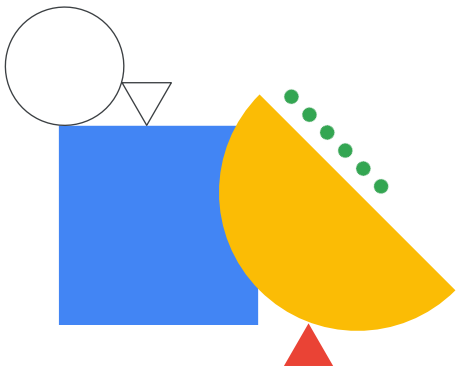
Google Cloud

Analyzing Large Datasets with BigQuery



Agenda

01 Data Analyst Tasks, Challenges, and Google Cloud Data Tools
   Demo: Analyze 10 billion records with BigQuery

02 Fundamental BigQuery Features

03 Google Cloud Tools for Analysts, Data Scientists, and Data Engineers
   Lab: Exploring a BigQuery Public Dataset

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

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.

Data Analyst Tasks, Challenges, and Google Cloud Data Tools

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.

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

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.

Challenges
- Data Volume
- Data Variety
- Data Velocity

Challenges
- Slow Exploration
- Slow Processing
- Unclear Logic

Challenges
- Storage Cost
- Hard to Scale
- Latency Issues

Challenges
- Slow Queries
- Data Volume
- Siloed Data

Challenges
- Dataset Size
- Tool Latency

Google Cloud offers scalable big data tools to overcome data challenge

Ingest

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

Transform

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

Store

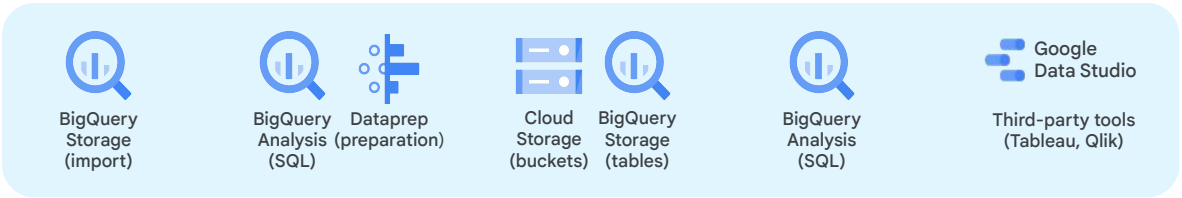
Create, save, and store datasets **inexpensively**.

Analyze

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

Visualize

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

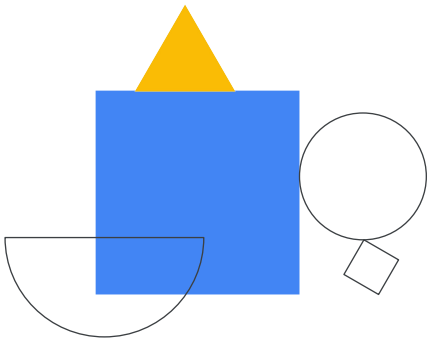


Google Cloud big data tools:
https://cloud.google.com/solutions/big-data/

Demo

Analyze 10 billion records with BigQuery

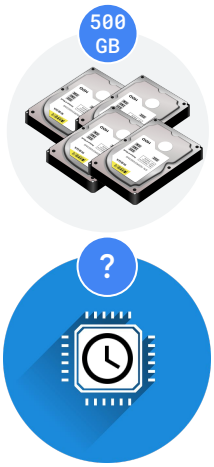
Fully-managed data analysis that scales



Refer to
https://github.com/GoogleCloudPlatform/training-data-analyst/tree/master/courses/data-to-insights/demos/wikipedia-10-billion.sql

BigQuery demo using 10 billion+ rows

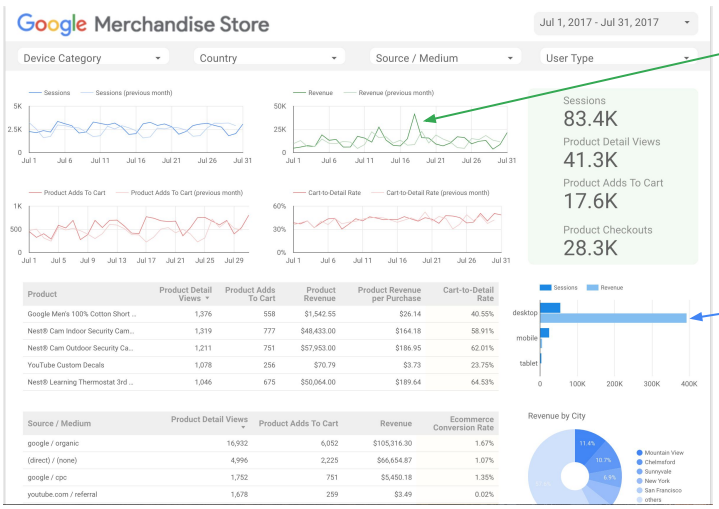
```
#standardSQL

# 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;
```



Refer to https://github.com/GoogleCloudPlatform/training-data-analyst/tree/master/courses/data-to-insights/demos/ folder

Demo: wikipedia-10-billion.sql

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?

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

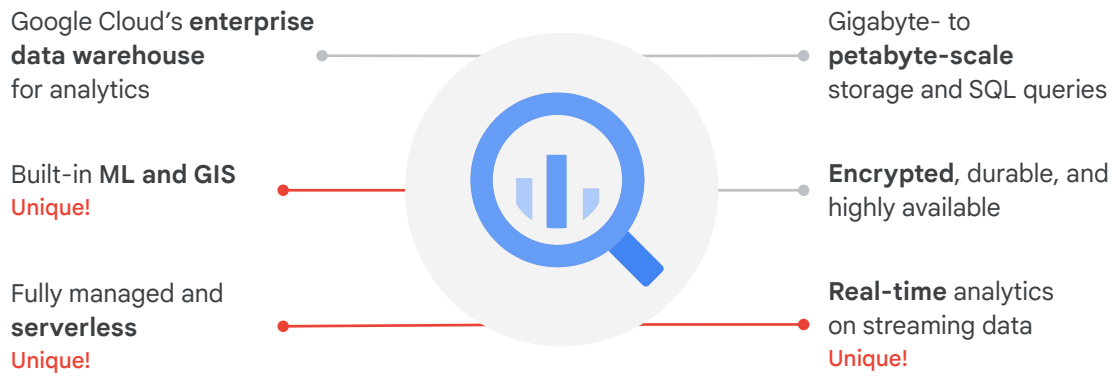
Here take a look at a merchandise dashboard and the highlighted insights and recommended actions.

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

Fundamental BigQuery Features

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

What is BigQuery ?



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.

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

All your focus should be on finding interesting places to see.

BigQuery is a petabyte-scale data analytics warehouse

1. Fully-Managed Data Warehouse	2. Reliable	3. Economical
No-ops, petabyte-scale	Backed by Google data centers	Pay only for the processing and storage you use

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

Reliable

Data replicated across multiple data centers

Economical

Only pay for storage and processing used

BigQuery is a petabyte-scale data analytics warehouse

4. Secure	5. Auditable	6. Scalable
Role ACLs, data encrypted in transport and at rest	Every transaction logged and queryable	Highly parallel processing model means fast queries

Secure

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

BigQuery is a petabyte-scale data analytics warehouse

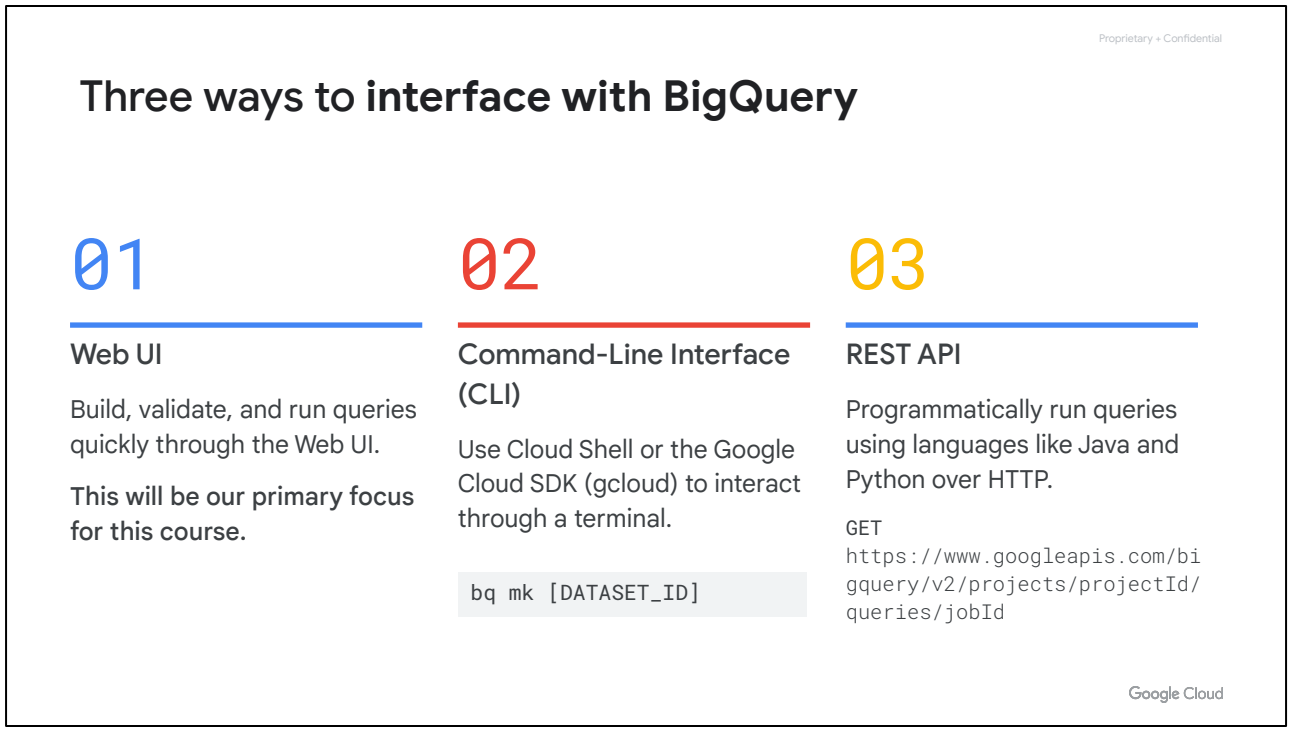
7. Flexible	8. Easy-to-use	9. Public Datasets
Mashup data across multiple datasets	Familiar SQL, no indexes, open standards	Explore and practice with real datasets (NOAA, IRS, GitHub, NYC Taxi etc.)

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



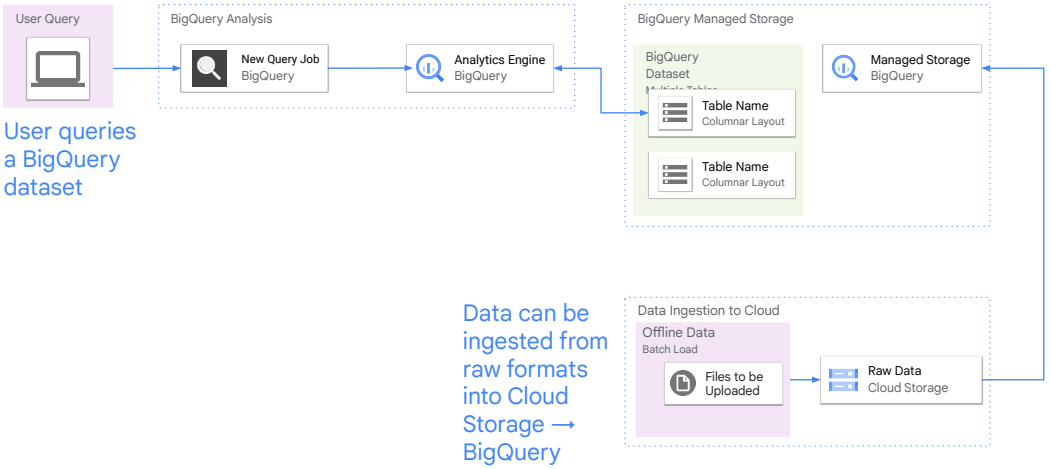
There are three ways to interact with BigQuery – the web UI, the command-line interface (CLI), and the REST API.

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



BigQuery is actually two services in one

BigQuery

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.)

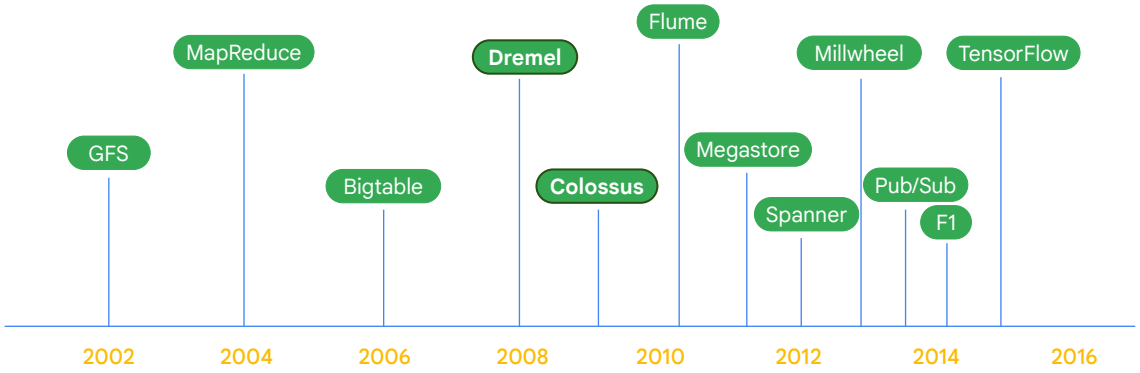
BigQuery Analysis

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

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

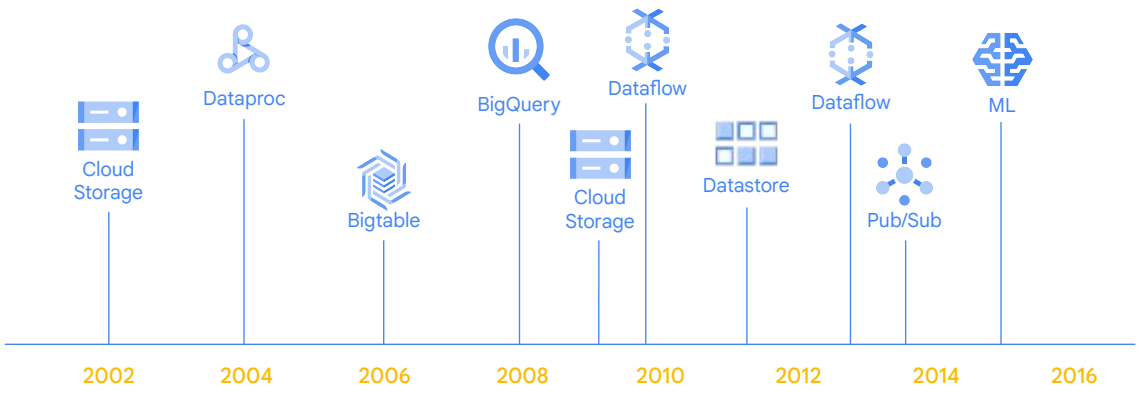
Google innovates data technologies



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

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 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

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 just so they don't have to worry about determining how much hardware they need.

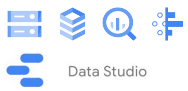
Google Cloud Tools for Analysts, Data Scientists, and Data Engineers

In this last lesson, 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

Data Analyst

- What they do:
  Derive data insights from queries and visualization
- Background:
  Data analysis using SQL
- Google Cloud tools used:



Data Scientist

- What they do:
  Analyze data and model systems using statistics and machine learning
- Background:
  Statistical analysis using SQL, R, Python
- Google Cloud tools used:

Data Engineer

- What they do:
  Design, build, and maintain data processing systems
- Background:
  Computer Engineering
- Google Cloud tools used:

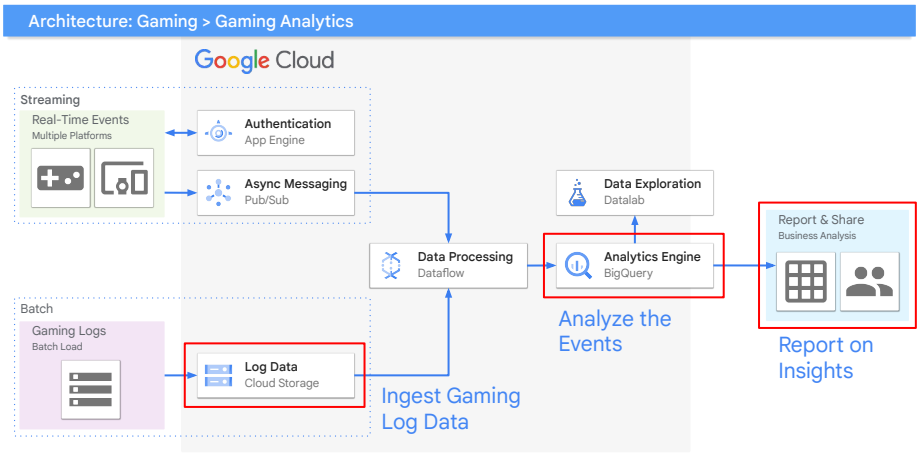
Spotlight on Certifications and Additional Courses
https://cloud.google.com/certification/data-engineer

Data Analyst
Cloud Storage
BigQuery
Dataprep
Google Data Studio

Data Scientist
Datalab
BigQuery
AI Platform

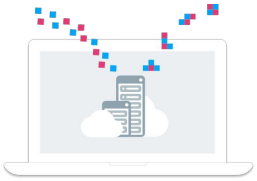
Data Engineer
Compute Engine
Cloud Storage
Dataproc DataStore
Dataflow Cloud SQL
Bigtable Spanner

End-to-end gaming analytics example highlighting Google Cloud tools



Additional background on the life of a BigQuery Query:
https://cloud.google.com/blog/big-data/2016/01/anatomy-of-a-bigquery-query

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, Dataprep, and Google Data Studio.

Explored the 9 features that make BigQuery a petabyte-scale data analytics warehouse.

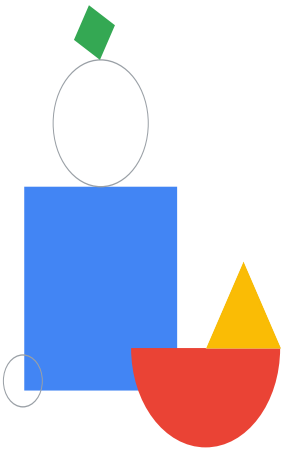
Compared data analysts, data scientists, and data engineers.

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. 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.

Lab Intro

Exploring a BigQuery Public Dataset



Lab objectives

01 Query a public dataset (USA Names)

02 Create a custom table from a CSV

03 Load data into a table

04 Query a table