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

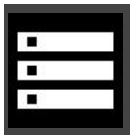
A data analyst is responsible for analyzing and gleaning insights from data

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.

Challenges

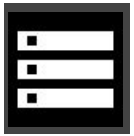
Data Volume
Data Variety
Data Velocity

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.

Slow Queries
Data Volume
Siloed Data

Visualize

Explore and present data insights.

Dataset Size
Tool Latency

Choosing the Right Tools

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

Ingest

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

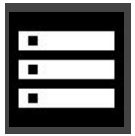
BigQuery Storage (import)

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

Module 2

Big Data Tools Overview

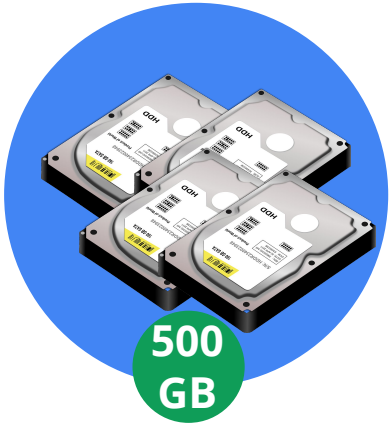
In this module we will:

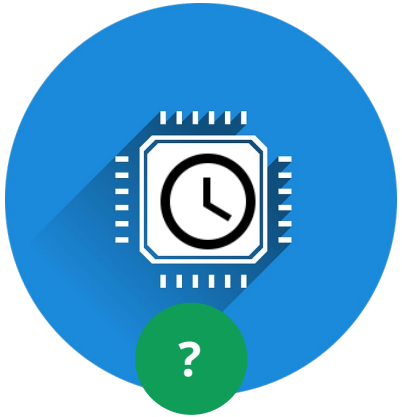
- Walkthrough Data Analyst Tasks, Challenges, and Introduce Google Cloud Platform Data Tools
- **Demo: Analyze 10 Billion Records with Google BigQuery**
- Explore 9 Fundamental BigQuery Features
- Compare GCP Tools for Analysts, Data Scientists, and Data Engineers

Google Cloud

BigQuery Demo using 10 Billion+ rows

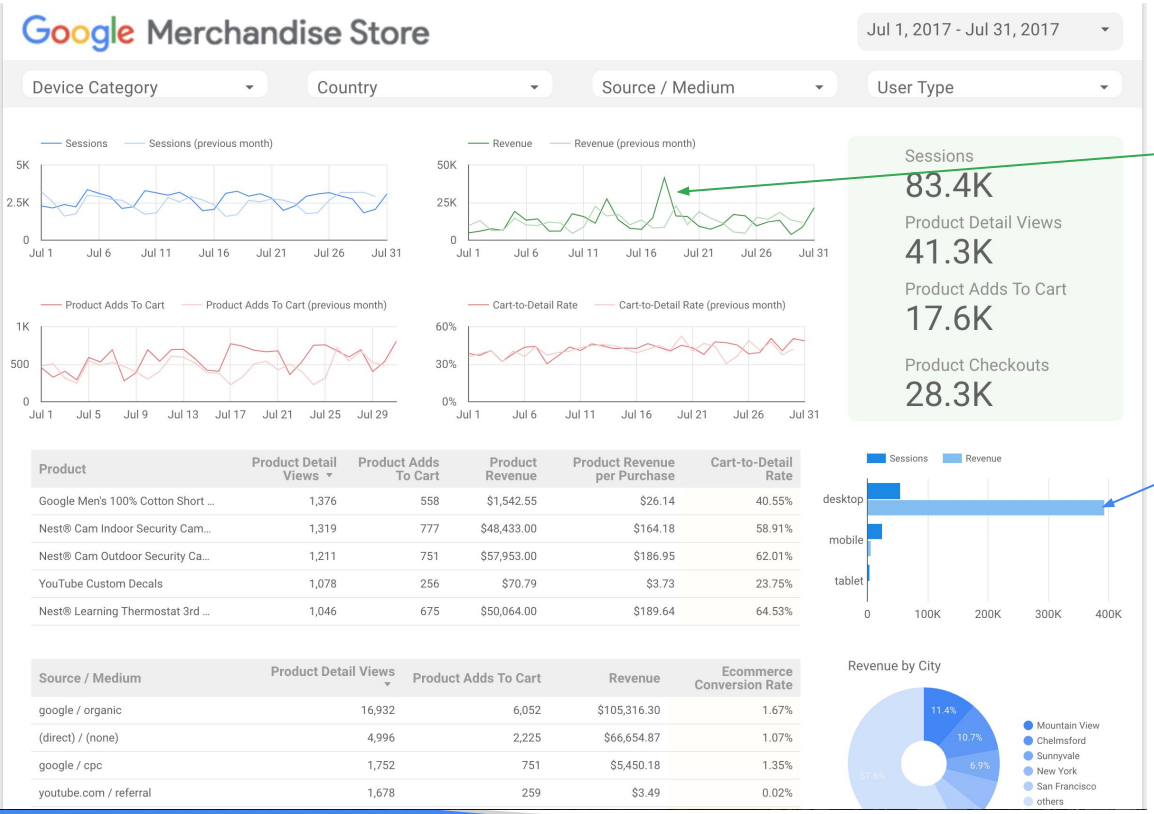
```
#standardSQL

# 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

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

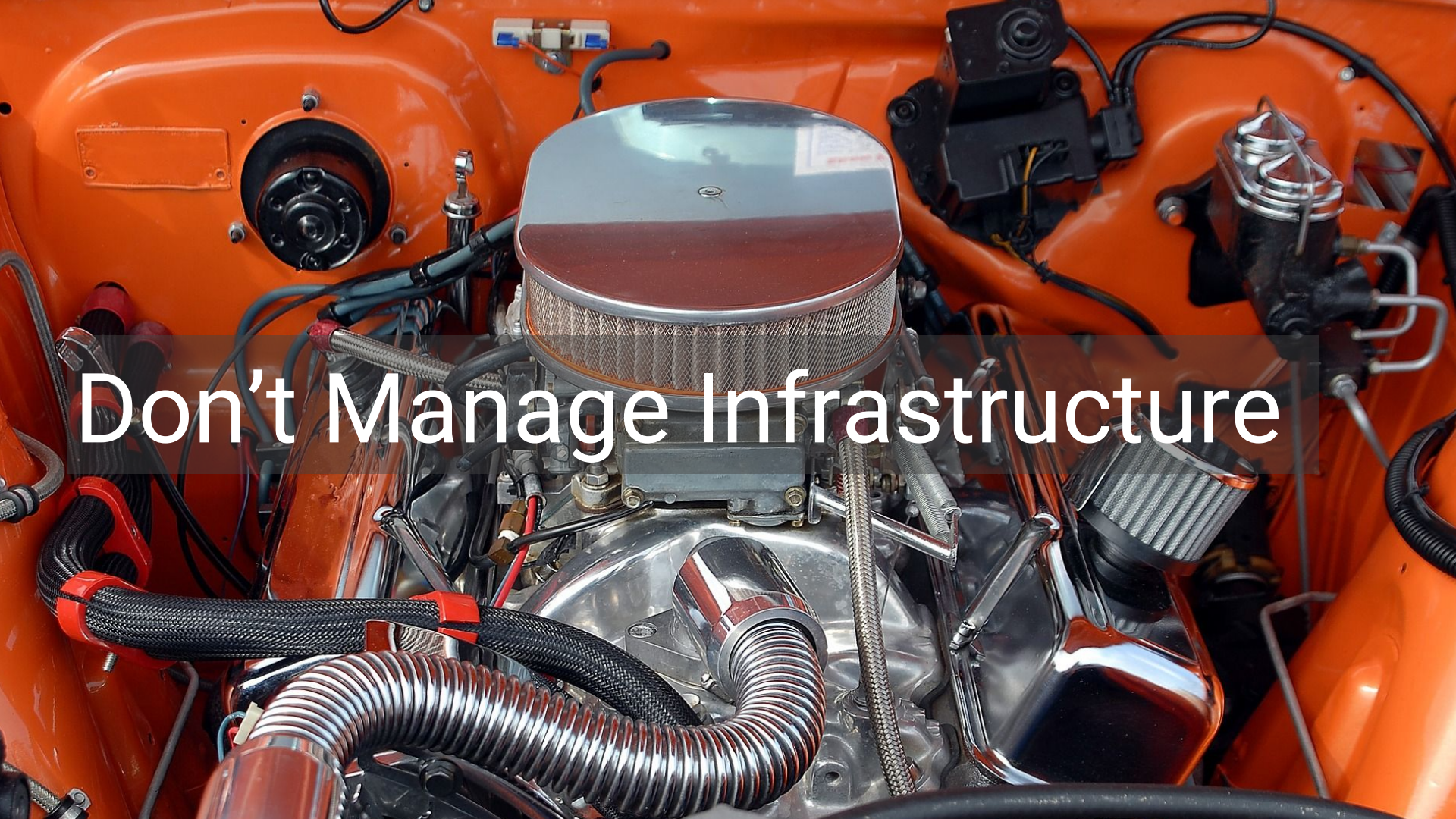
Module 2

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


Don’t Manage Infrastructure


Focus on Finding Insights

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

Google
Big Query

1. **Fully-Managed Data Warehouse:** No-Ops, Petabyte-Scale
2. **Reliability:** Backed by Google Datacenters
3. **Economical:** Pay only for the processing and storage you use

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

Google
Big Query

4 **Security:** Role ACLs, Data Encrypted in Transport and at Rest

5 **Auditable:** Every Transaction Logged and Queryable

6 **Scalable:** Highly Parallel Processing Model means Fast Queries

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

Google
Big Query

7 **Flexible:** Mashup Data across Multiple Datasets

8 **Easy-to-use:** Familiar SQL, No Indexes, Open Standards

9 **Public Datasets:** Explore and Practice with Real Datasets (NOAA, IRS, GitHub, NYC Taxi etc.)

Three Ways to **Interface with BigQuery**

Web UI

Build, validate, and run queries quickly through the Web UI.

This will be our primary focus for this course.

Command-Line Interface (CLI)

Use Cloud Shell or the Google Cloud SDK (gcloud) to interact through a terminal

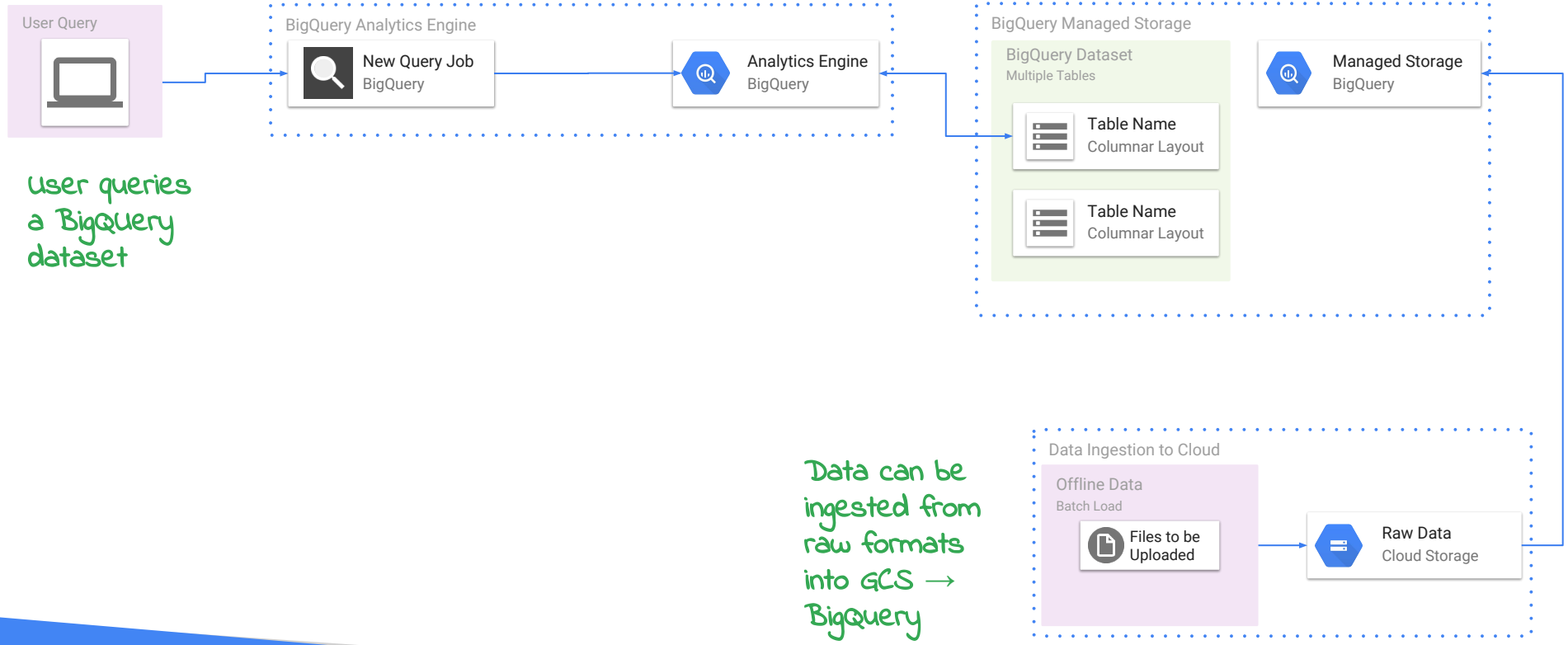
```
bq mk [DATASET_ID]
```

REST API

Programmatically run queries using languages like Java and Python over HTTP

```
GET
https://www.googleapis.com/bigquery/v2/projects/projectId/queries/jobId
```

Creating and Querying Datasets: BigQuery Terminology

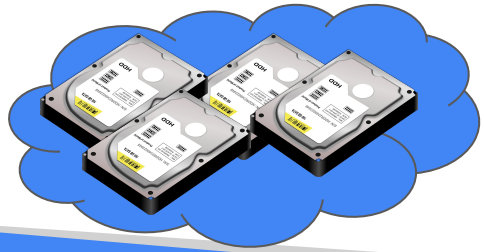


Google BigQuery is actually two services in one

Google
Big Query

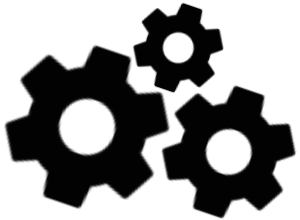
BigQuery
Managed Storage

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

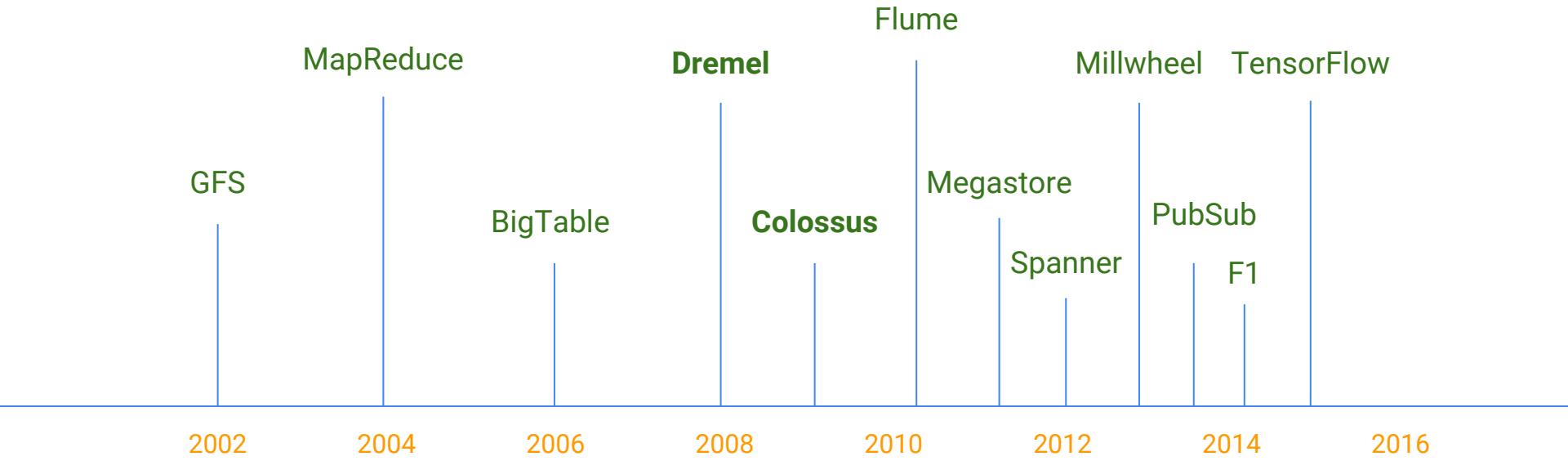


BigQuery
Analysis

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

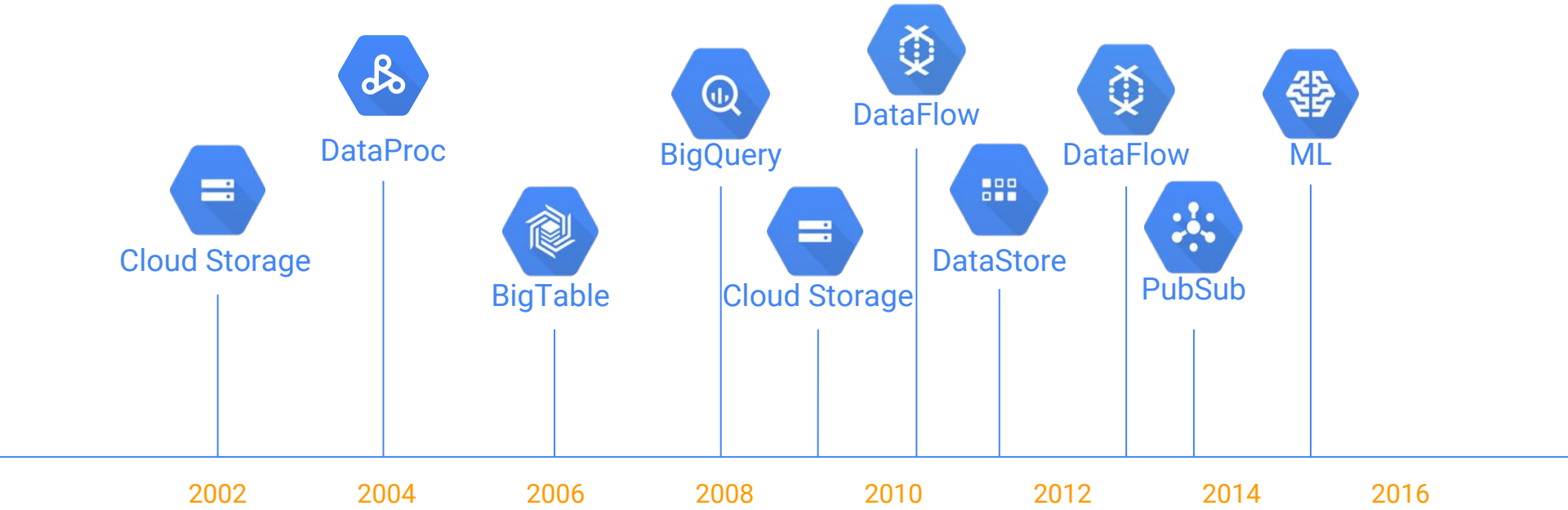


Google innovates data technologies



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 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

Module 2

Big Data Tools Overview

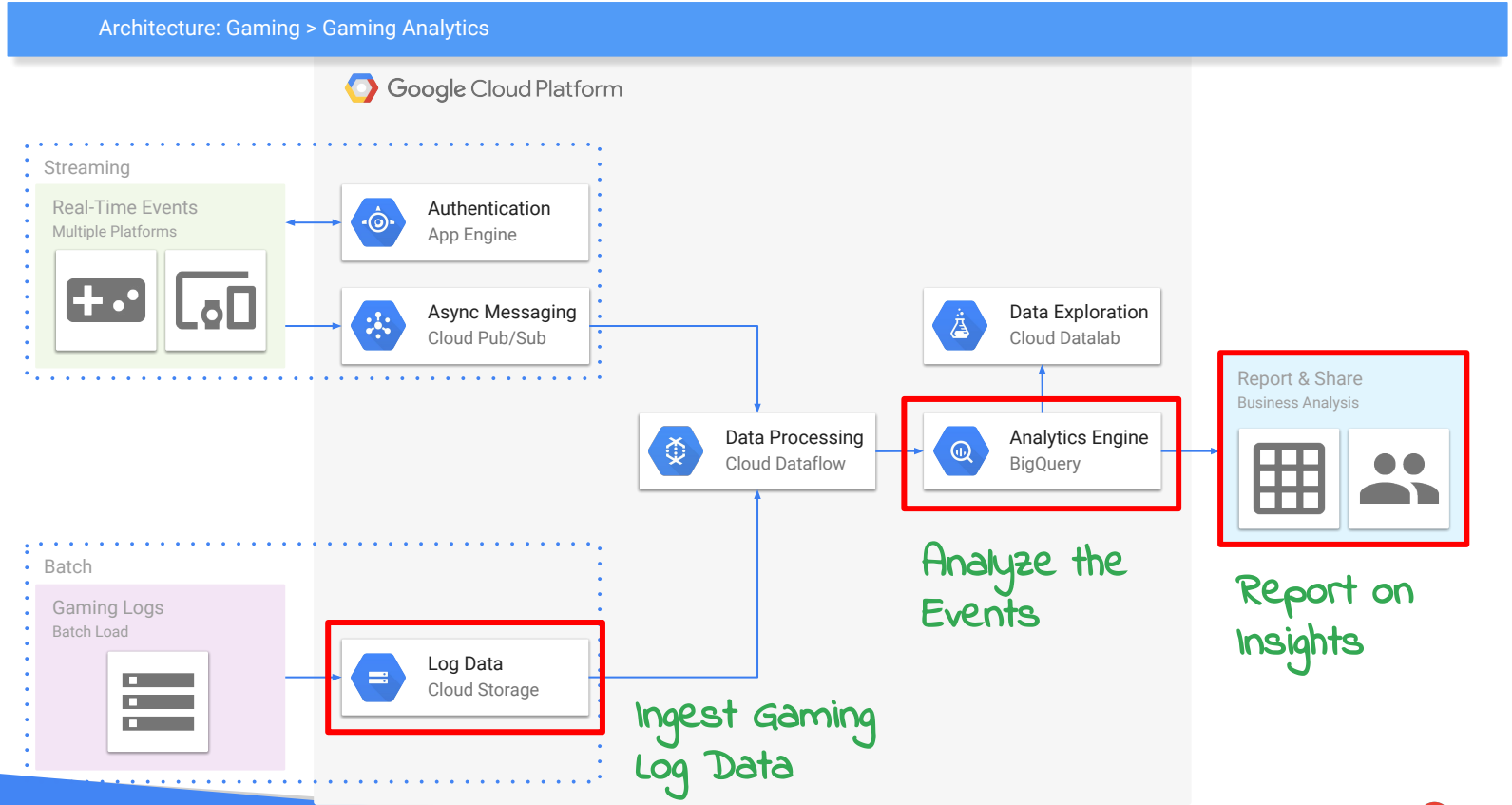
In this module we will:

- 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**

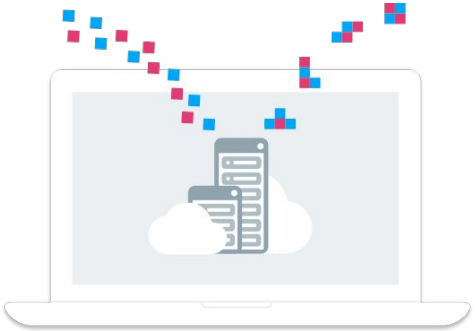
Each data-related role uses a different suite of tools

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:			

End-to-end gaming analytics example highlighting GCP tools



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

Lab 1
Exploring your Public Dataset with Google BigQuery

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

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

Form 990 Preview

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

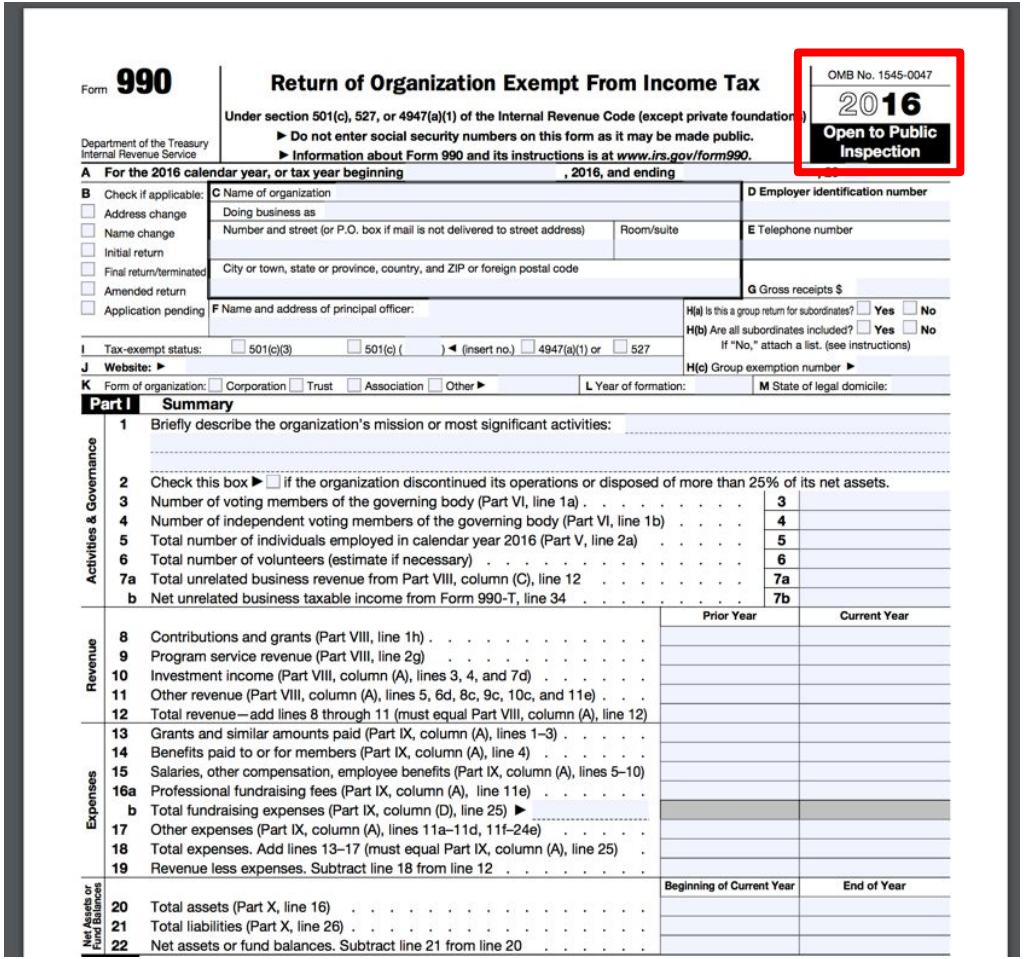
It contains key financial information and is then published publicly by the IRS for the public to also inspect

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 foundations)

► 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*.

Department of the Treasury
Internal Revenue Service

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:

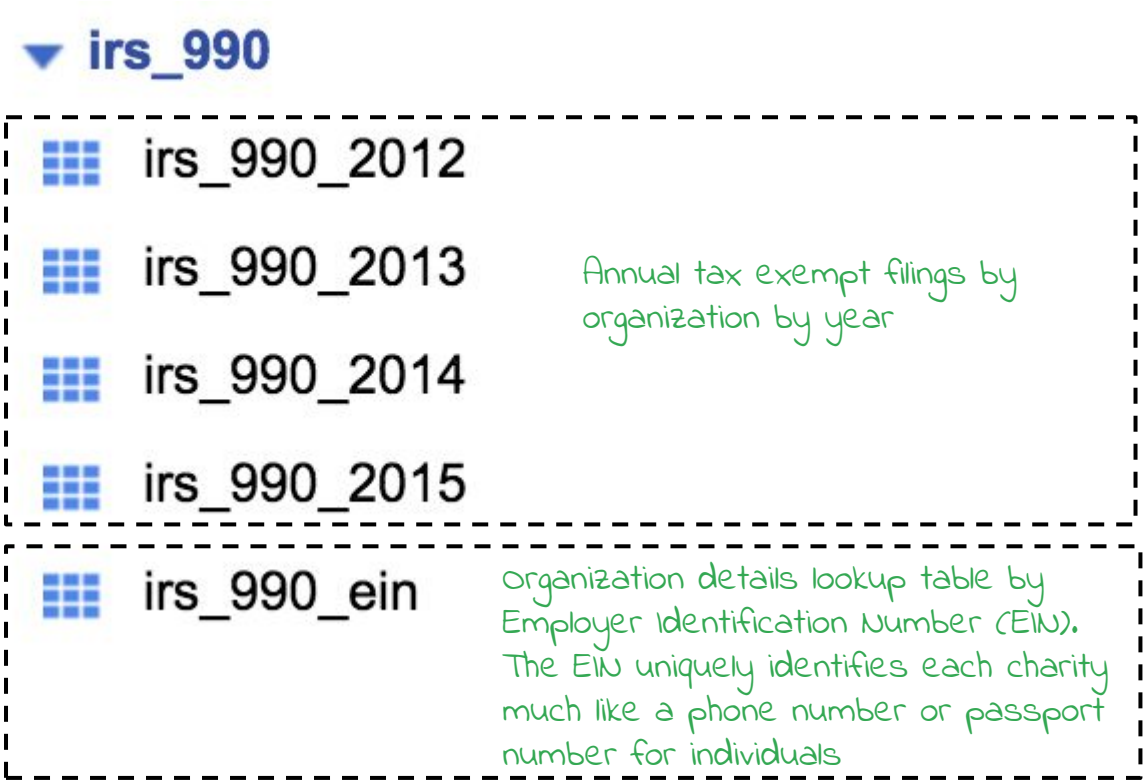
M State of legal domicile:

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

IRS BigQuery public dataset has two primary table types



Google Cloud

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