

# Cloud Computing Normal and Big Data

---

CIS437

Erik Fredericks // [frederer@gvsu.edu](mailto:frederer@gvsu.edu)

*Adapted from Google Cloud Computing Foundations, Overview of Cloud Computing (Wufka & Canonico)*

# FOR LATER

---

[https://www.skills.google/course\\_templates/649/labs/592584](https://www.skills.google/course_templates/649/labs/592584)

First off, what types of data *do we have to deal with?*



# First off, what types of data do we have to deal with?

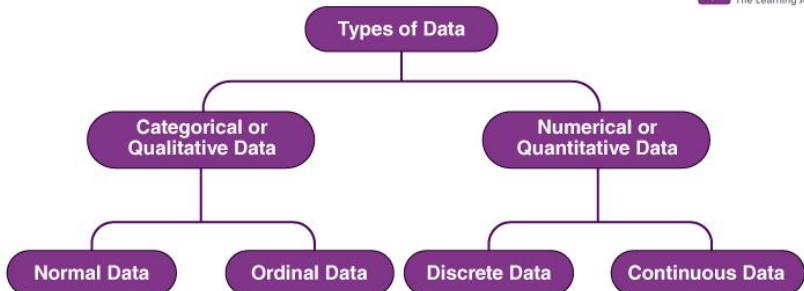
## Binary

- Blobs ... think images, archives, general files, etc.

## Text

- Strings, numbers, JSON, etc.

*Note, if you look to data science you'll have many, many more classifications  
We just care about how to store and retrieve it ... for now*



# Interestingly, that's not as important for us

The question becomes more:

- 1) How *much* data do we have to deal with?
- 2) How *often* does the data change?
- 3) At what *speed* does the data come in/out
- 4) How *reliable* is the data?
- 5) How *valuable* is the data?

Anybody know what this is leading to?

## 5 V's OF DATA

### VOLUME

Amount of Data



### VALUE

Worth of Data



### VARIETY

Diversity of Data



### VELOCITY

Speed of  
Data Generation



### VERACITY

Accuracy of Data



# Essentially, big or normal data?

## Big data:

- Enormous amount of data to manage (think, petabytes/exabytes)
- Exceedingly complex

## Normal data:

- Things you can store in a database without performance considerations
- Files you can store off to some sort of bucket-like system

	<b>Big Data</b>	<b>Small Data</b>
<b>Data Condition</b>	Always unstructured, not ready for analysis, many relational database tables that need merged	Ready for analysis, flat file, no need for merging tables.
<b>Location</b>	Cloud, Offshore, SQL Server, etc.	Database, local PC
<b>Data Size</b>	Over 50K Variables, over 50K individuals, random samples, unstructured	File that is in a spreadsheet, that can be viewed on a few sheets of paper
<b>Data Purpose</b>	No intended purpose	Intended purpose for Data Collection

<https://www.312analytics.com/what-is-the-difference-between-big-and-small-data/>

# Storage options

## Buckets

- File storage

## CloudSQL / Spanner

- Relational database

## BigTable

- NoSQL

## BigQuery:

- Relational (big) data warehouse

*BigLake*: storage/analytics for data lakes



# BIGTABLE VS BIGQUERY

@PVERGADIA

NOSQL  
WIDE-COLUMN  
DATABASE

USE ME FOR  
HEAVY  
READ/WRITE  
EVENTS

DATA WAREHOUSE  
FOR RELATIONAL  
STRUCTURED  
DATA

OH HEY!  
I AM FAST>>

SINGLE DIGIT  
MILLISECOND  
LATENCY PER  
ENTRY/ACCESS

WE ARE BOTH  
CLOUD-NATIVE

I AM YOUR FRIEND  
FOR LARGE SCALE,  
AD-HOC  
SQL-BASED OLAP  
ANALYSIS

BIGTABLE

BIGQUERY



REAL-TIME  
ANALYSIS OF  
LARGE  
DATASETS

ADTECH, FINTECH, IOT, GAMING,  
TIME SERIES ANALYSIS



STORE AND  
ANALYSIS  
LARGE  
DATASETS

PREDICTIVE ANALYTICS, ML

# What is a data lake?

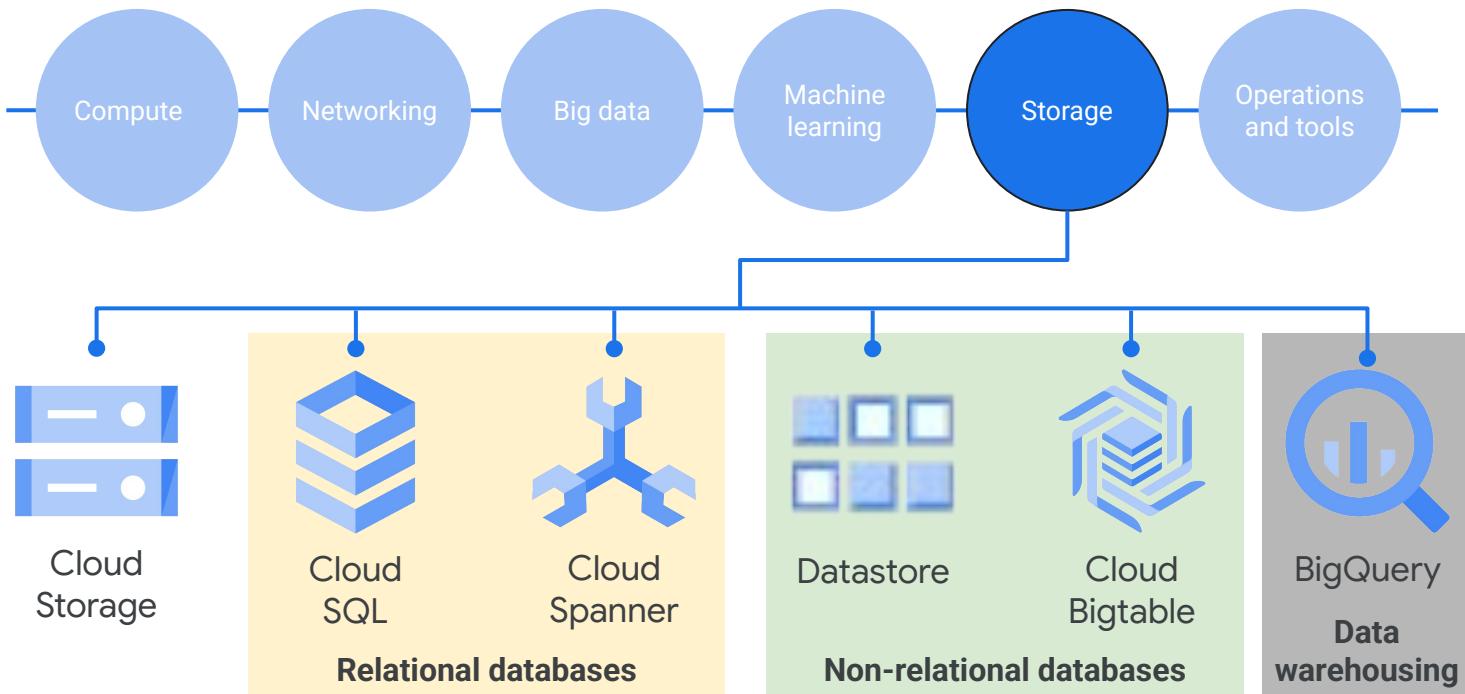
<https://aws.amazon.com/what-is/data-lake/>

"A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale. You can store your data as-is, without having to first structure the data, and run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions."

Think of it as a place to store different types of data for generating reports/analytics to make business decisions

- Or, a data warehouse at scale

# Google Cloud has many storage options



# There are three common use cases for cloud storage

- 1 Content storage and delivery
- 2 Storage for data analytics and general compute
- 3 Backup and archival storage



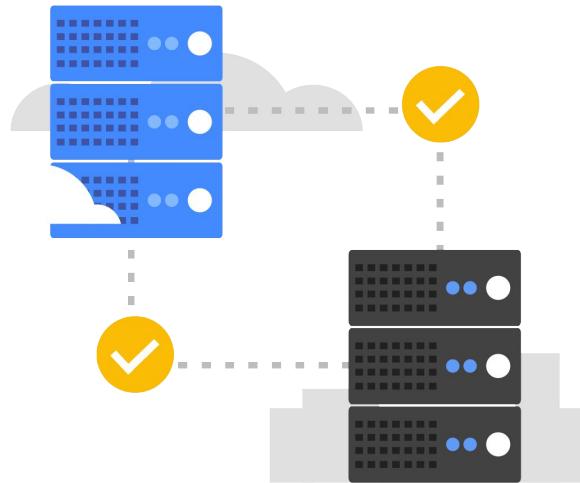
# For users with databases, Google has two priorities



Migrate existing databases to the cloud, and move them to the right service.



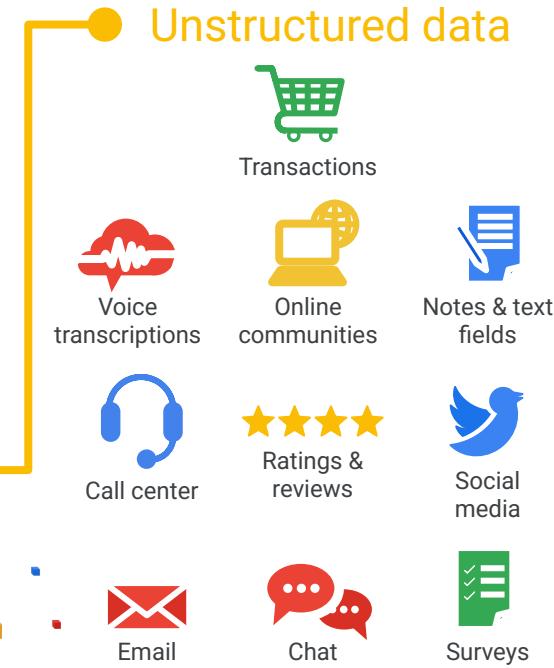
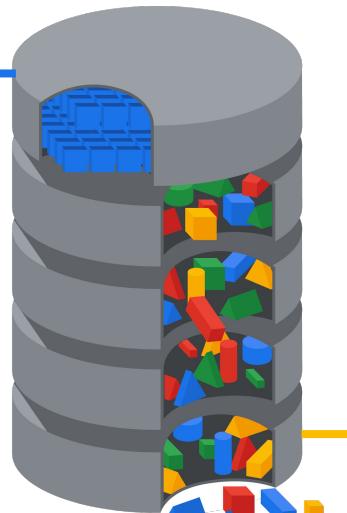
Innovate, build, or rebuild for the cloud, take advantage of mobile, and plan for future growth.



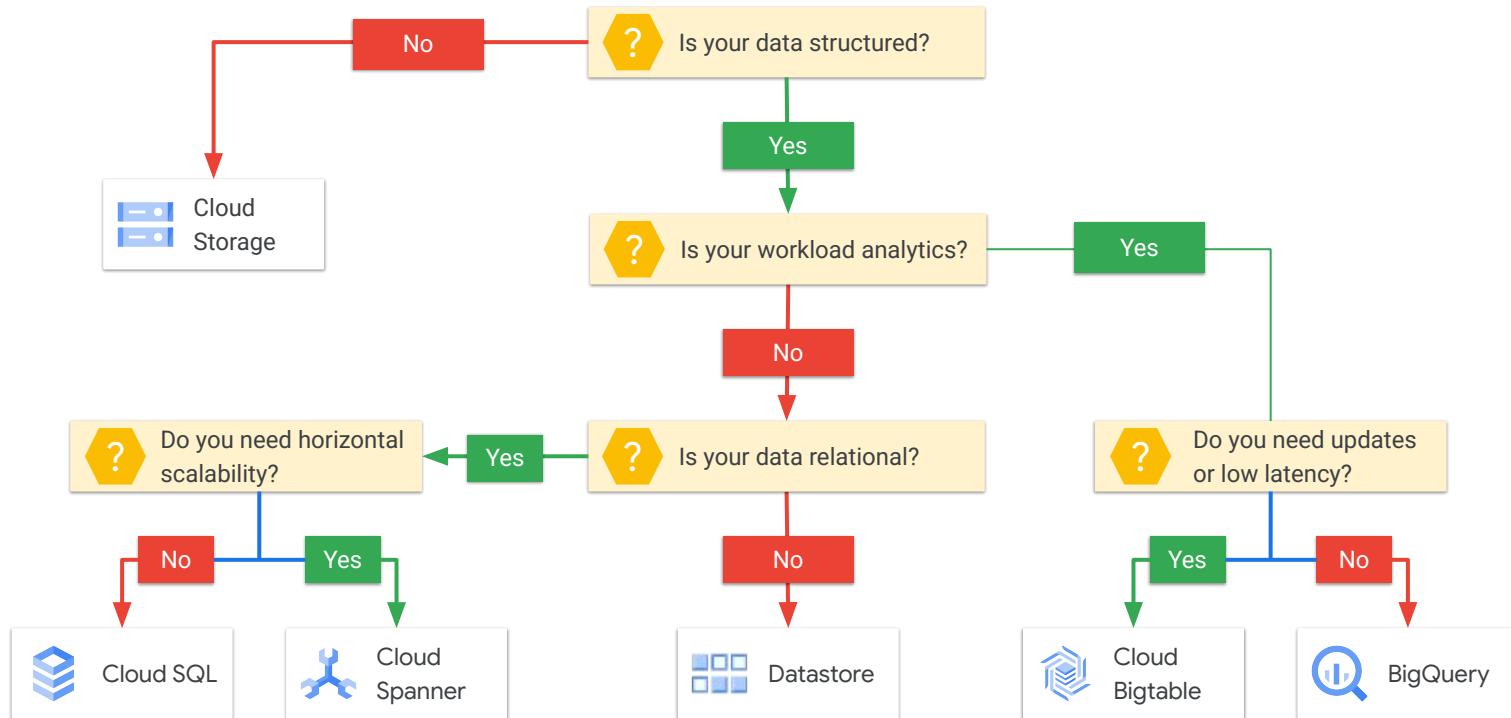
# Structured versus unstructured data

First_Name	Last_Name	Address	City	Age
Sherlock	Holmes	12 Main St	Mesa	60
James	Bond	23 Old St	Napa	43
Scarlett	O'Hara	34 New St	Derby	23
Marge	Simpson	56 West St	Cody	36

Structured data



# What type of storage will meet my needs best?



# Cloud Storage

## Object-based storage

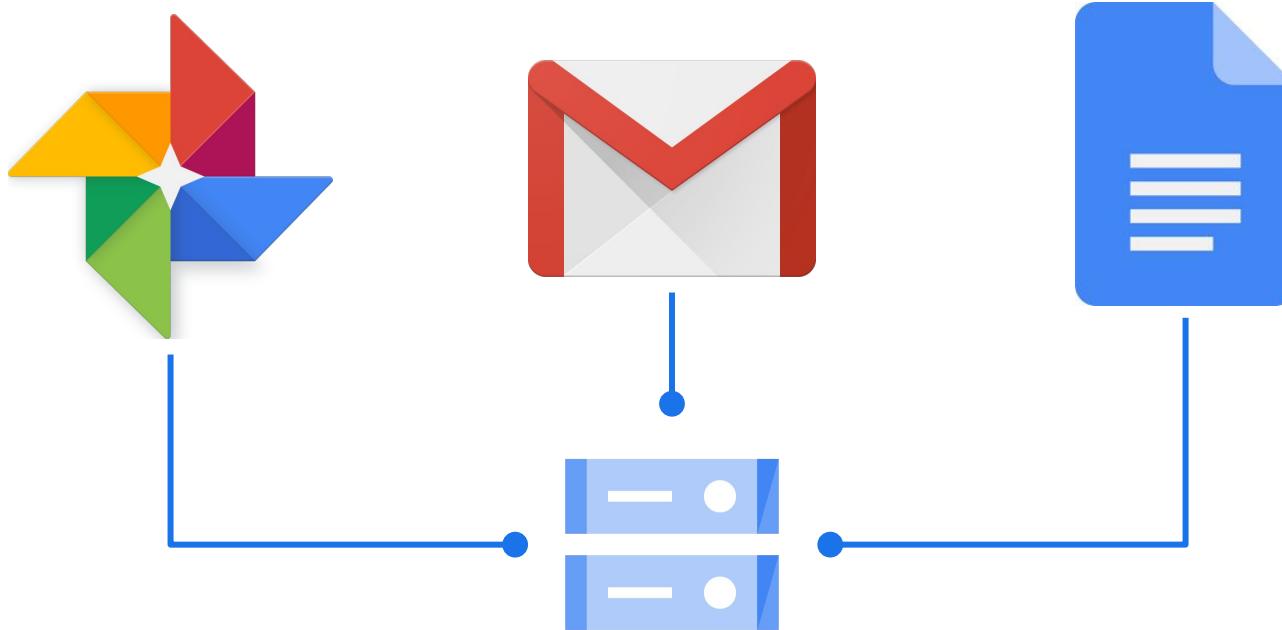
Uses concept of "buckets"

- Logical containers for files
- Usual access rights apply

<https://cloud.google.com/storage>

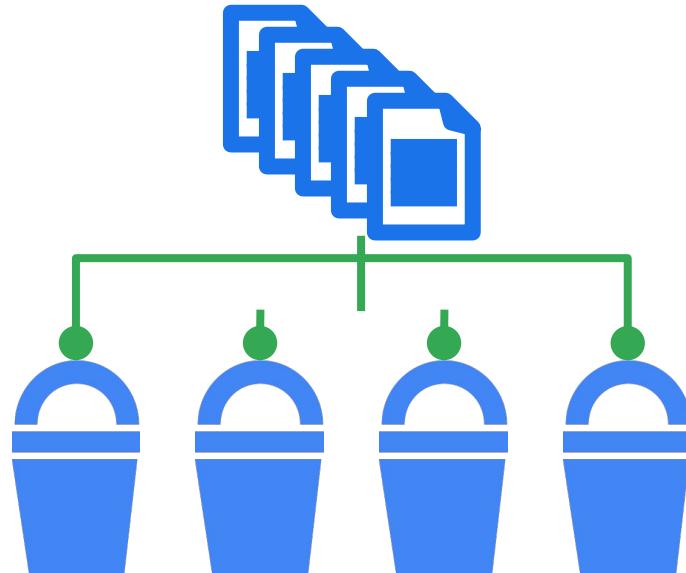
Storage type	Description	Best for
<a href="#">Standard storage</a>	Storage for data that is frequently accessed ("hot" data) and/or stored for only brief periods of time.	"Hot" data, including websites, streaming videos, and mobile apps.
<a href="#">Nearline storage</a>	Low cost, highly durable storage service for storing infrequently accessed data.	Data that can be stored for 30 days.
<a href="#">Coldline storage</a>	A very low cost, highly durable storage service for storing infrequently accessed data.	Data that can be stored for 90 days.
<a href="#">Archive storage</a>	The lowest cost, highly durable storage service for data archiving, online backup, and disaster recovery.	Data that can be stored for 365 days.

# Google uses Cloud Storage too!



# Cloud Storage files are organized into buckets

- Globally unique name
- Location (region, dual-region, or multi-region)
- Storage class
- IAM policies or access-control lists
- Object versioning setting
- Object lifecycle management rules



---

# Lab Intro

## Cloud Storage: Qwik Start - CLI/SDK

Create a storage bucket, upload objects, create folders and subfolders, and make objects publicly accessible using the Google Cloud command line.

You can find the lab [here](#).



# Cloud SQL

Now, time for database storage

Options are relational or non-relational (NoSQL)

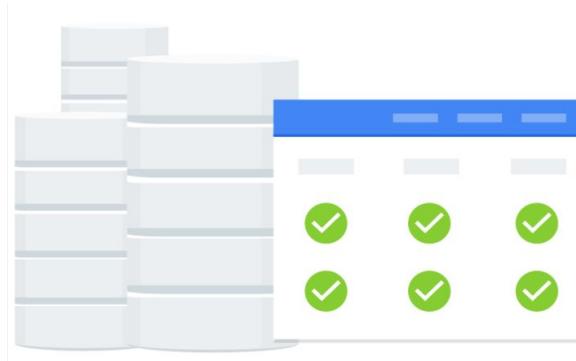
# What is a database and how is it used?

A collection of information organized so that it can easily be accessed and managed.

Computer applications run databases to get a fast answer to questions.



# Relational databases are the most common



Relational database management systems

= RDBMS

= relational databases

= SQL databases

Suitable use cases:

- Have a well-structured data model.
- Need transactions.
- Ability to join data across tables to retrieve complex data combinations.

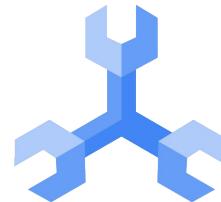
# Options for SQL-based managed services



## Cloud SQL

MySQL, PostgreSQL, and SQL Server databases as a service

- Automatic replication
- Managed backups
- Vertical scaling (read and write)
- Horizontal scaling (read)



## Cloud Spanner

- Automatic replication
- Strong global consistency
- Managed instances with high availability
- SQL (ANSI SQL 2011 with extensions)

# Cloud Spanner

Note: \$\$\$

---

# The difference between Cloud Spanner and other databases

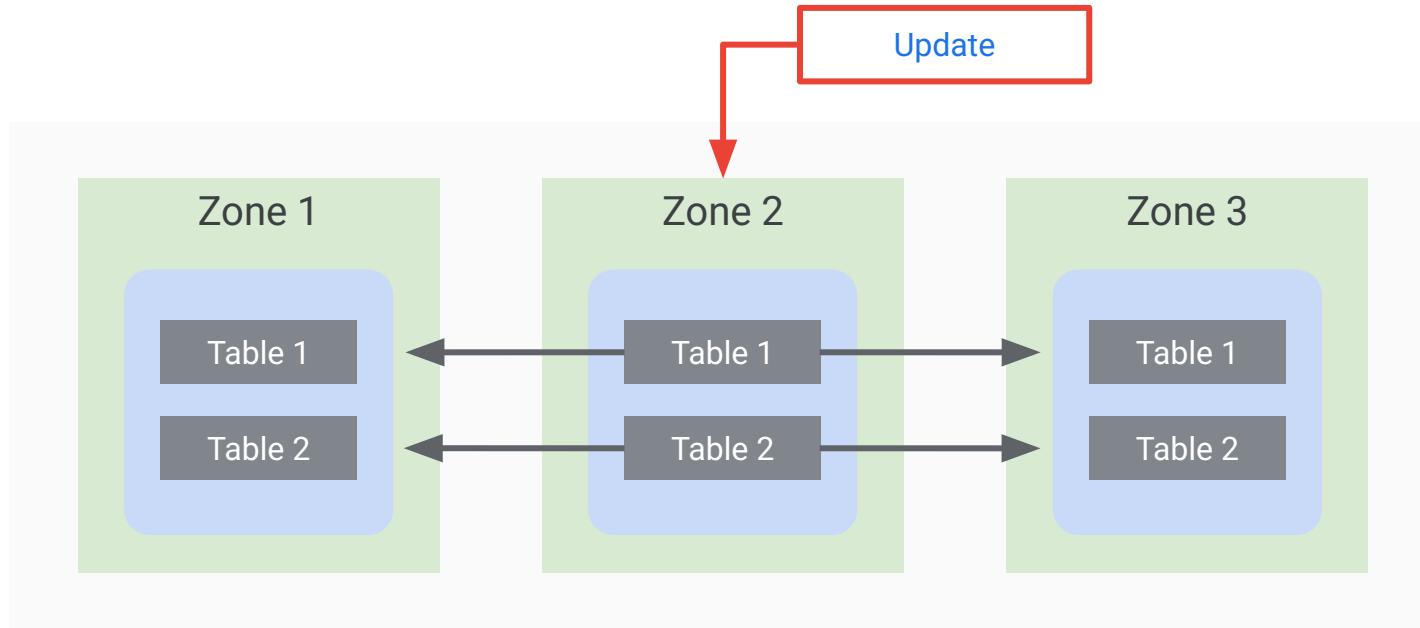
- 1 Familiar relational database structure
- 2 Scales to very large databases
- 3 Strong external consistency
- 4 Reduces operational overheads

# Get the best of relational database structure and non-relational database scale and performance



Scale + SQL	Fully managed	Launch faster	Enterprise grade security
Scales horizontally.  Low latency, transactional consistency, and high availability.  Future-proofs database backends.	Create or scale a globally replicated database in a few clicks.  Synchronous replication and maintenance built in.	Relational semantics.  ACID transactions.  Schemas.	Data-layer encryption.  IAM integration.  Audit logging.

# How Cloud Spanner works

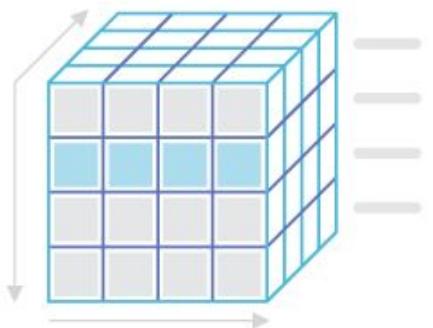
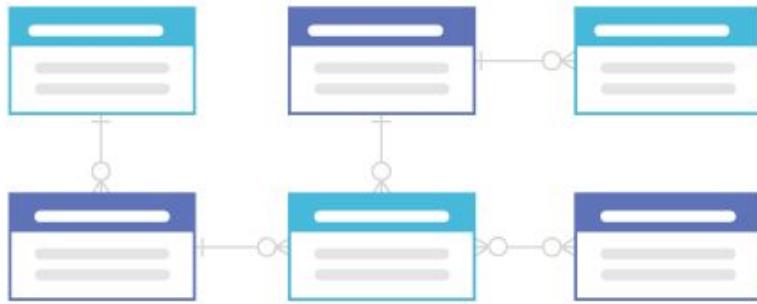


# NoSQL options

Difference to relational databases?

## SQL

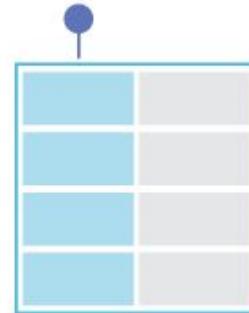
*Relational Database Management Systems (RDBMS)*



*Online Analytical Processing (OLAP) Cube*

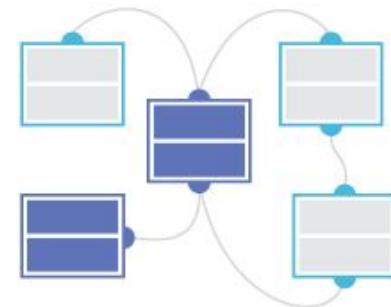
## NoSQL

*Key-Value*



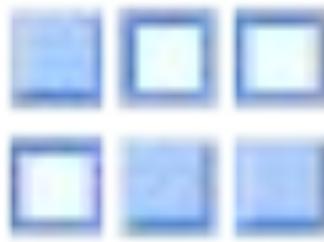
*Document*

*Graph*



*Column store*

# Options for NoSQL based managed services



Datastore



Cloud Bigtable

# Agenda

Exploring Cloud SQL

Lab: Cloud SQL for MySQL: Qwik Start

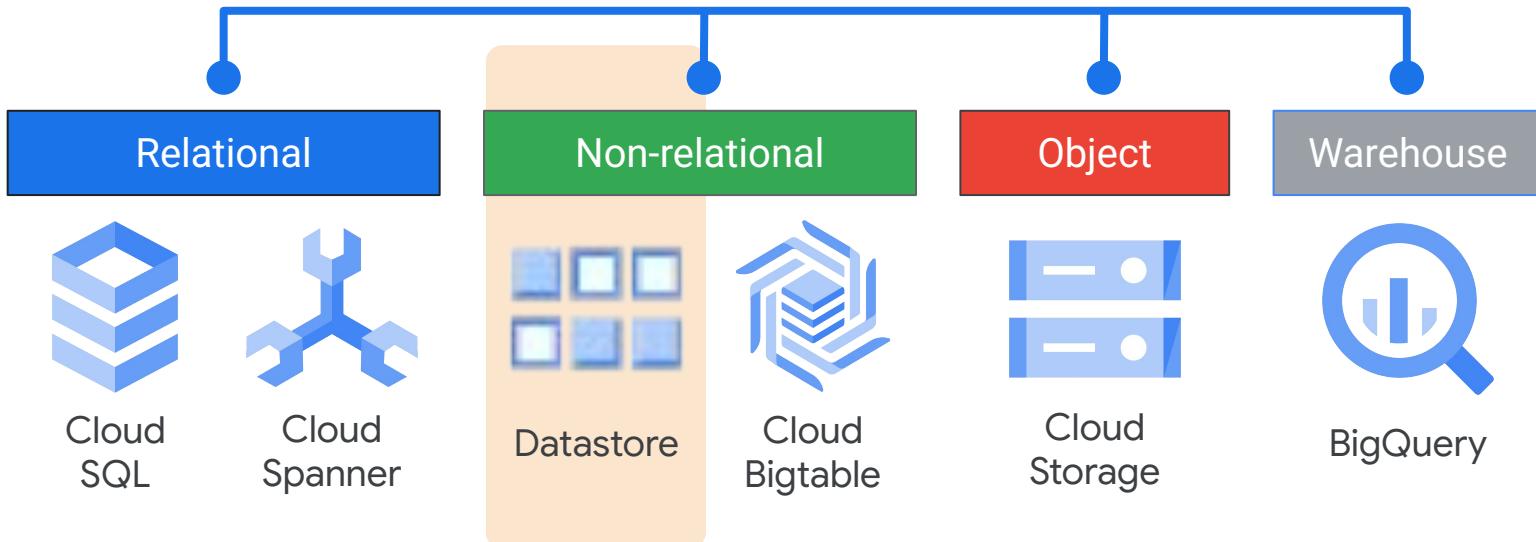
Cloud Spanner as a Managed Service

NoSQL Managed Services Options

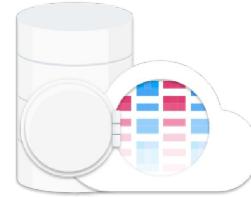
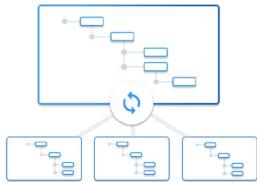
Datastore, a NoSQL Document Store



# Datastore in Google Cloud



# Datastore is a highly available and durable NoSQL database for low-latency serving of data



## Schema-less

Change your data structure as your app evolves.

## Fast and highly scalable

High-speed queries no matter the size of the database.

Seamless scaling.

## Fully managed

Instantly provision a scalable and available NoSQL database.

Automatic sharding and replication.

## Integrated and secure

RESTful interface makes data accessible by any deployment target.

Serves as an integration point.

# Examples of Datastore use cases

- 1 User profiles
- 2 Product catalogues
- 3 Recording transactions
- 4 Mobile games

# Agenda

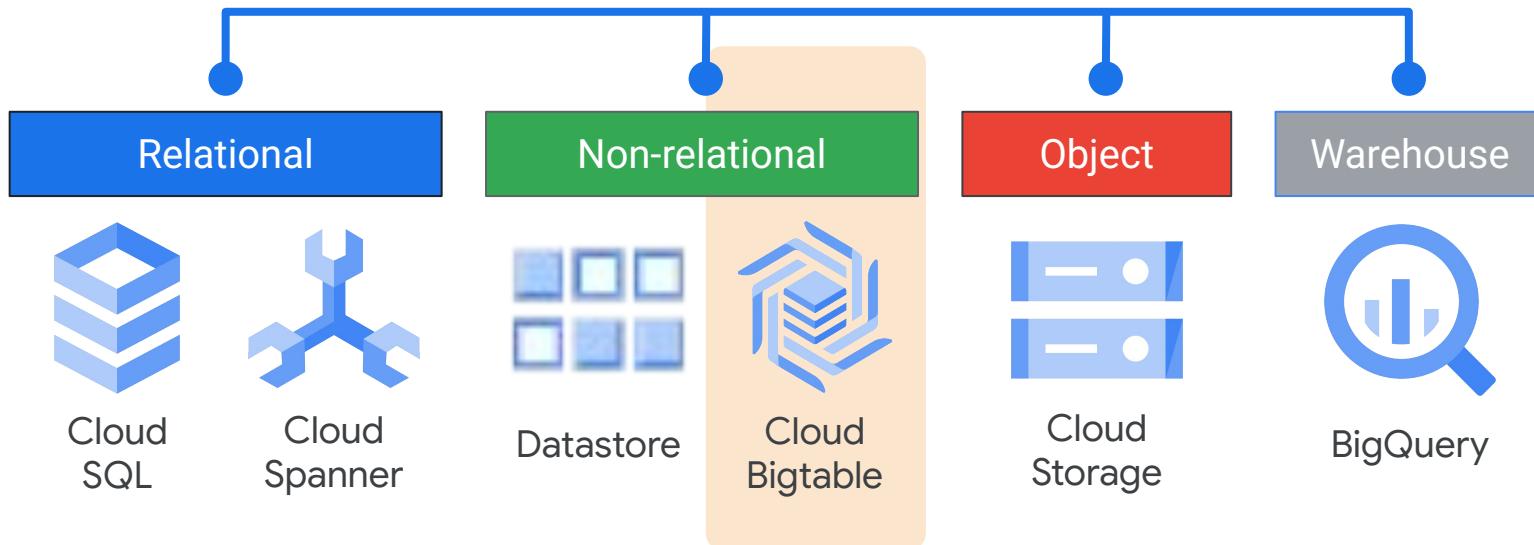
Cloud Bigtable as a NoSQL Option

Quiz

Summary



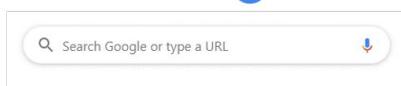
# Cloud Bigtable in Google Cloud



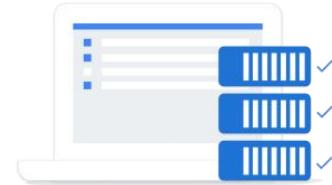
# Google uses Cloud Bigtable too!



Google Analytics



# Cloud Bigtable is a fully managed NoSQL database for large analytical and operational workloads



## Fast and performant

High performance under high loads.

Faster, more reliable, and more efficient.

Low latency

## Seamless scaling and replication

Billions of rows and thousands of columns.

No downtime during reconfiguration.

Replication adds high availability.

## Fully managed

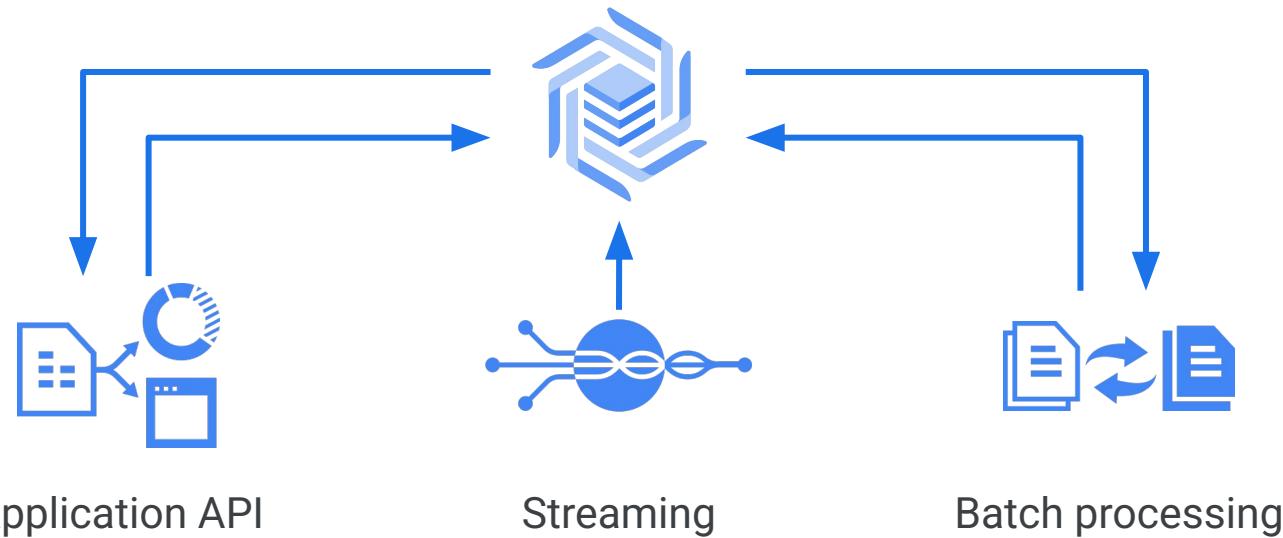
Database configuration and tuning handled by Google.

Data backups created for disaster recovery.

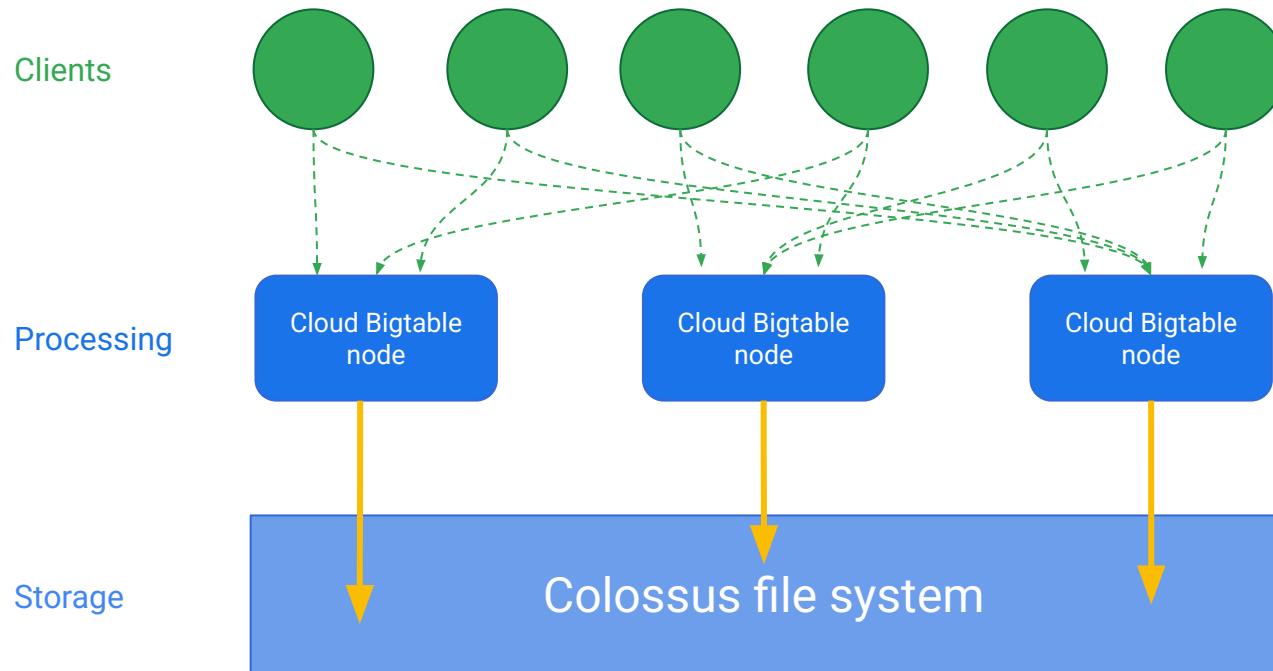
## Integrated and secure

Integrated with open-source big data tools for powerful data analysis.

Cloud Bigtable can interact with other Google Cloud services and third-party clients

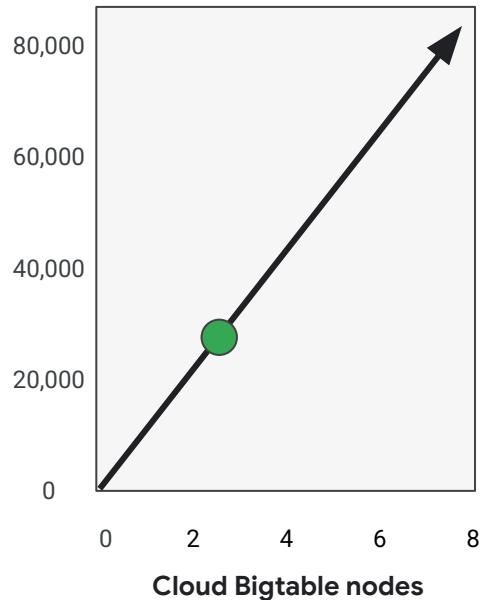


# Cloud Bigtable structure

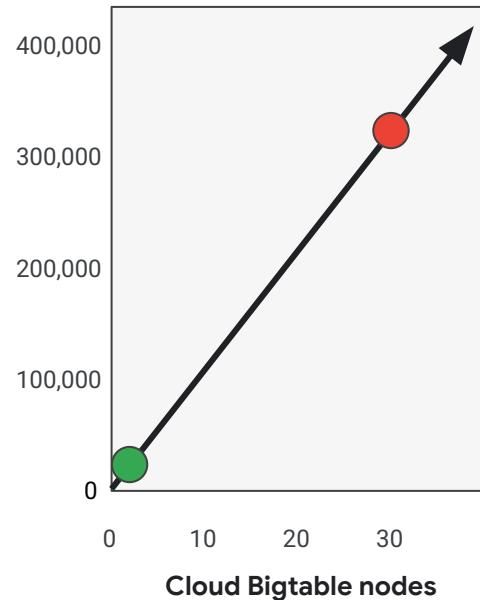


# Scaling Cloud Bigtable

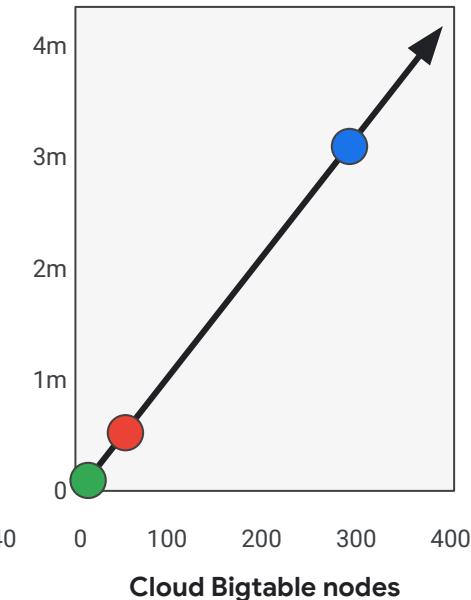
Queries per second



Queries per second



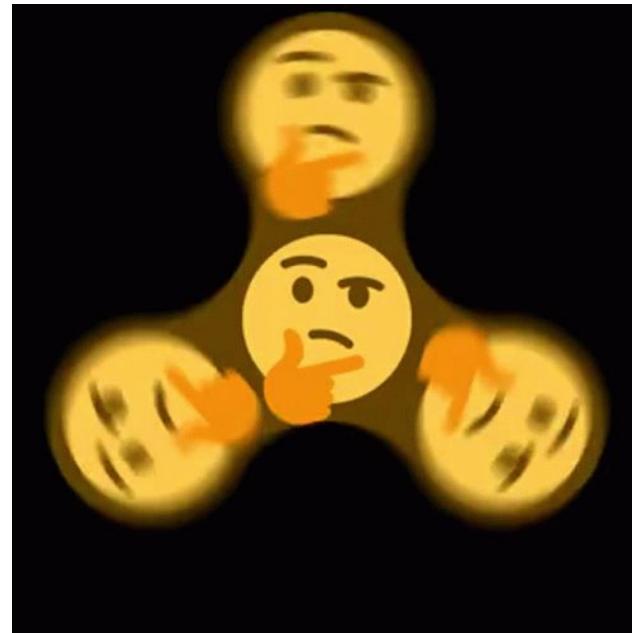
Queries per second



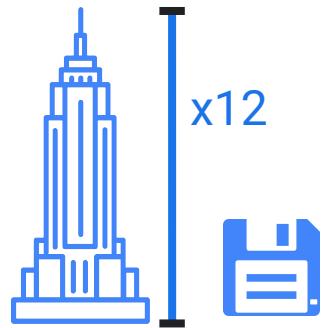
# BigQuery / BigLake Codelab

[https://codelabs.cs.pdx.edu/labs/C09.3g\\_bq\\_bl/index.html?index=..%2F..cs430#0](https://codelabs.cs.pdx.edu/labs/C09.3g_bq_bl/index.html?index=..%2F..cs430#0)

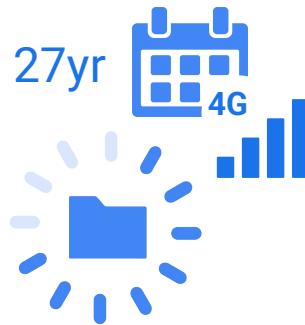
And now, what to do with all this data



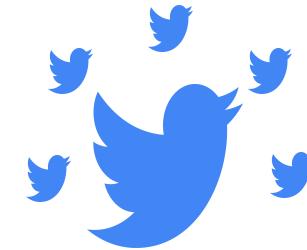
# How big is a petabyte of data?



A stack of floppy  
disks higher than 12  
Empire State  
Buildings

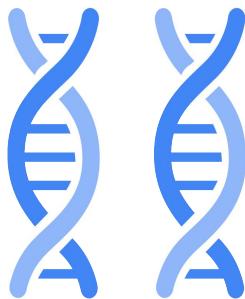


27 years to  
download over  
4G



Every tweet ever  
tweeted ...  
x 50

# How small is a petabyte of data?



---

2 micrograms of  
DNA



---

1 day's worth of  
video uploaded  
to YouTube

# Overview of big data managed services



Dataproc

Process big data  
with Hadoop/Spark



Dataflow

Analyze streaming  
data in real time



BigQuery

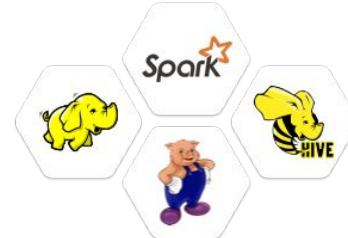
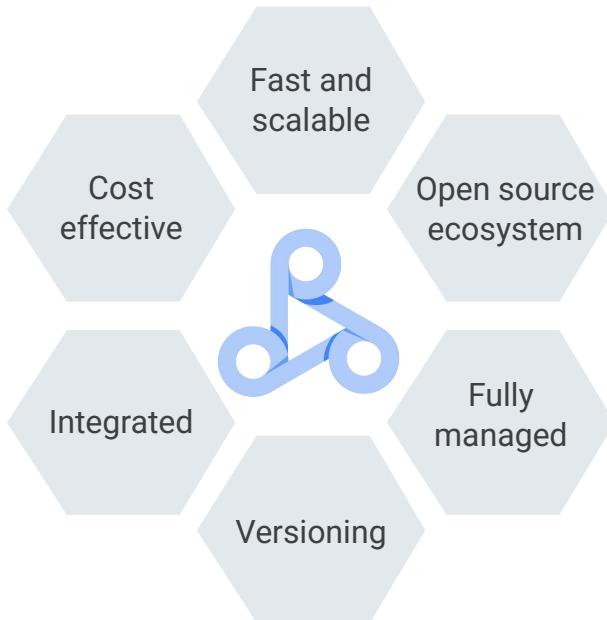
Modernize a data  
warehouse  
foundation

---

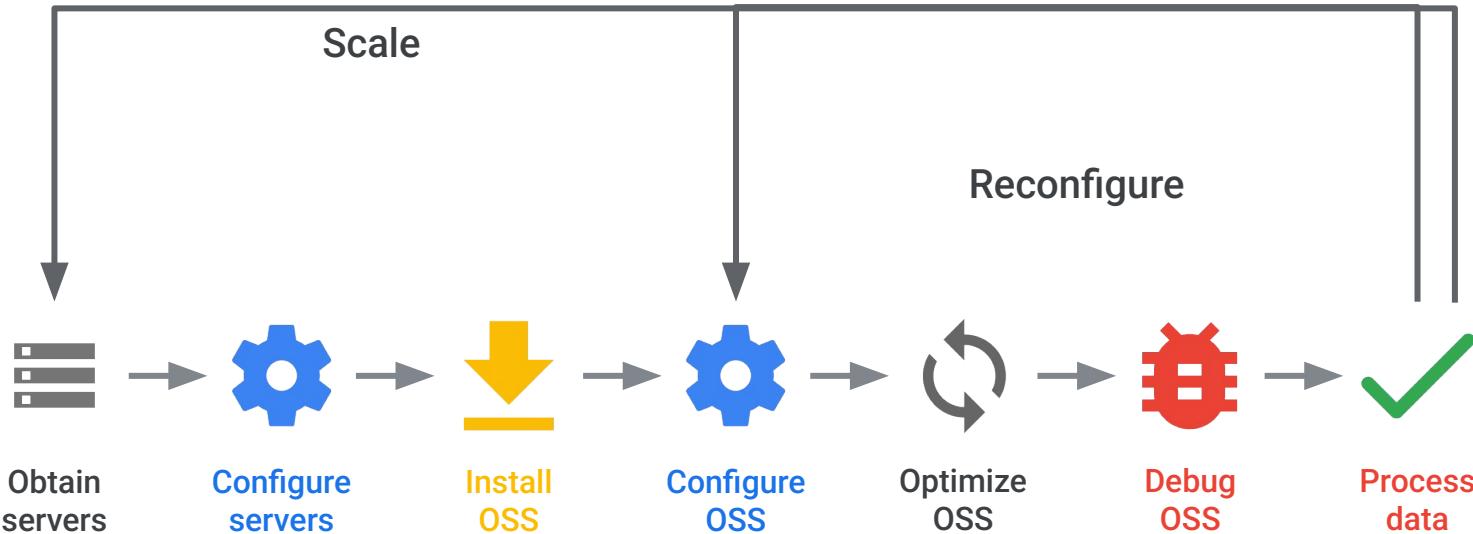
Hadoop and Spark are open source technologies that often form the backbone of big data processing



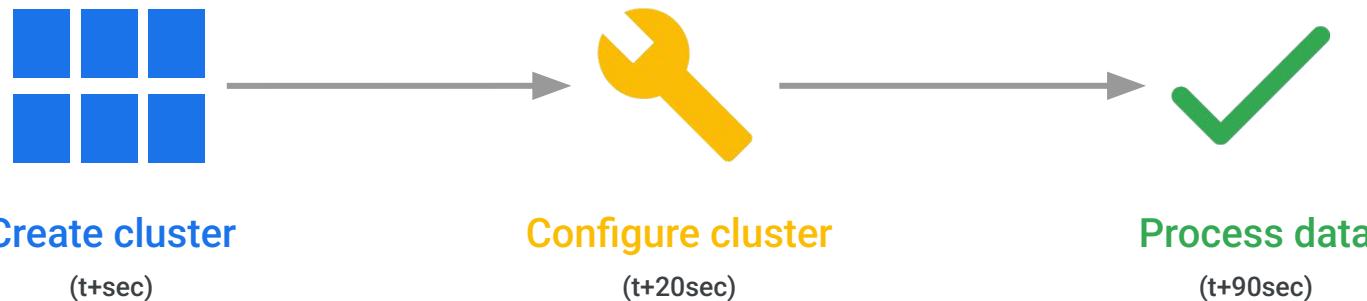
# Dataproc is a managed service for batch processing, querying, streaming, and ML



# Typical Spark/Hadoop clusters

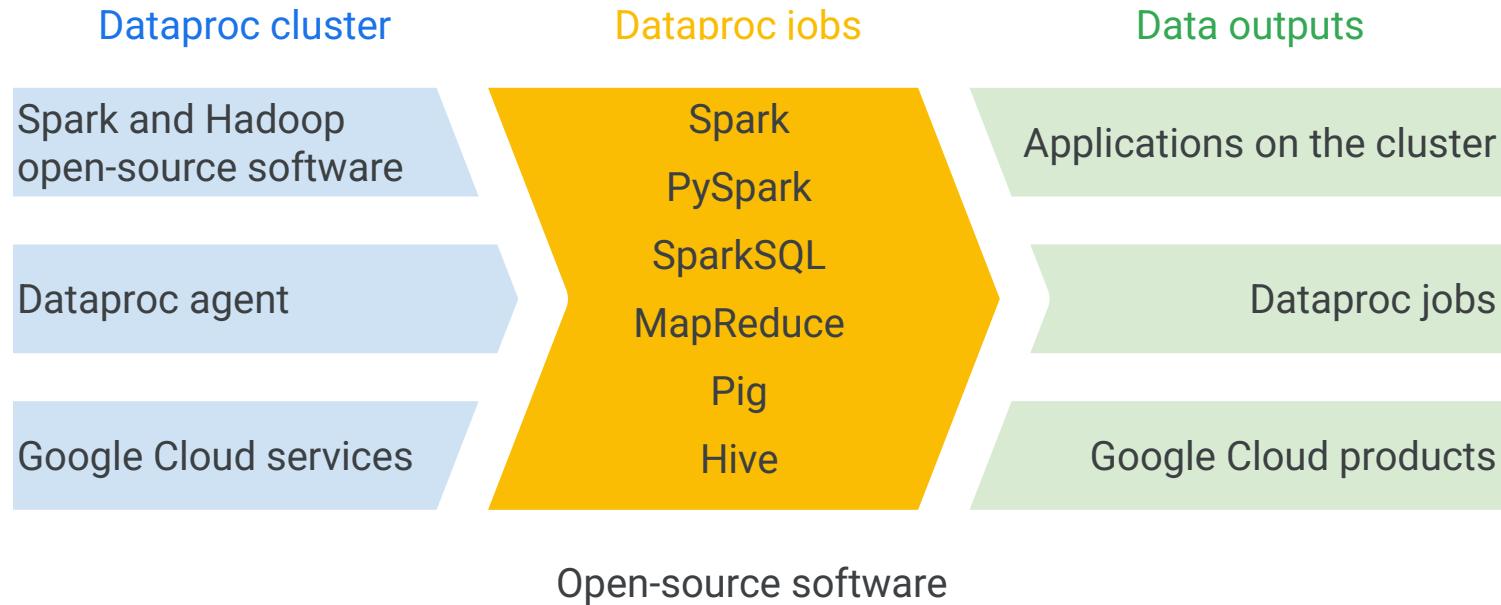


# Dataproc separates storage and compute

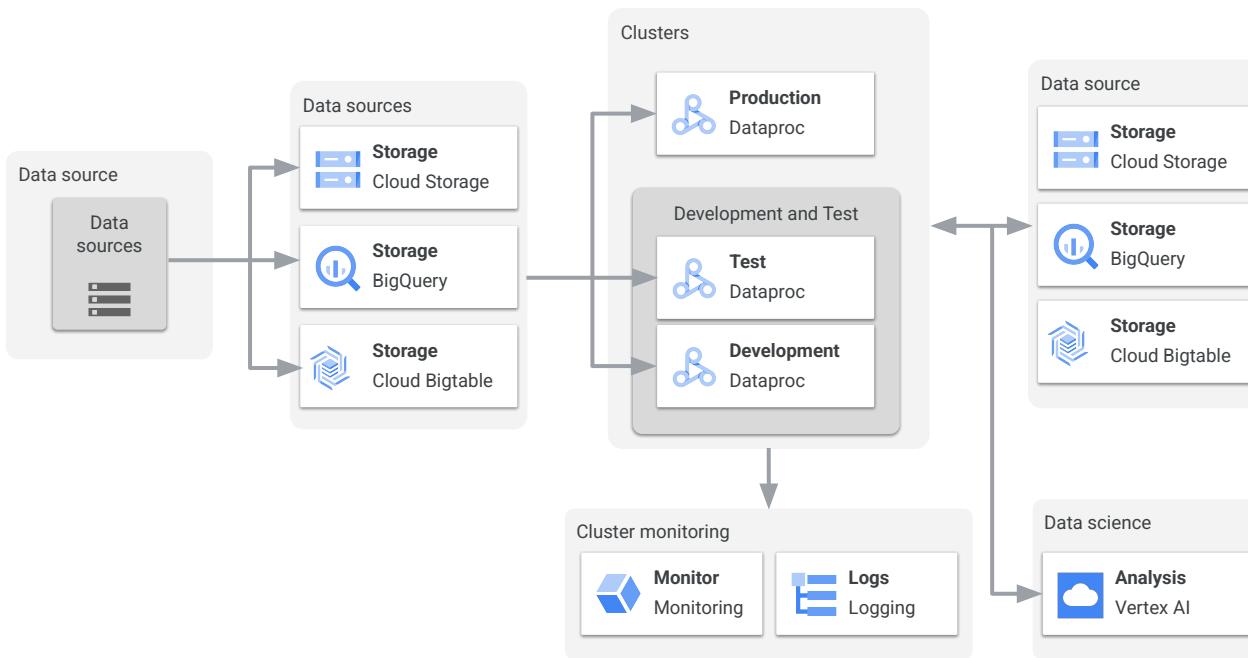


hdfs:// ➔ gs://

# Hadoop and Spark jobs and workflows



# No management or maintenance required!



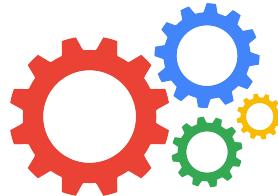
# Dataproc can help with log processing



The  
need

Large volumes of data from several sources are aggregated and loaded into databases so metrics can be gathered for daily reporting, management dashboards, and analysis.

A dedicated on-premises cluster is currently used to store and process the logs with MapReduce.



The  
solution

Cloud Storage provides a low-cost storage option.  
An ephemeral Dataproc cluster can be created in less than 2 mins.

Data is processed with existing MapReduce.



The  
value

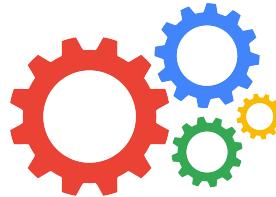
Saves money and reduces complexity.

# Dataproc can help with ad-hoc data analysis



The  
need

Analysts are using Spark Shell but are concerned about increase in usage.  
Unsure on how to scale their cluster, which is running in standalone mode.



The  
solution

Creates clusters that scale for speed and mitigate failure.  
Can use web interface, Cloud SDK, or native Spark Shell via SSH.



The  
value

Unlocks the cloud without technical complexity.  
Complex computations take seconds, not hours.

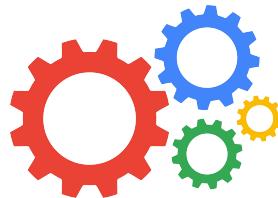
# Dataproc can help with machine learning



The  
need

Spark Machine Learning Libraries (MLlib) are used to run classification algorithms on large datasets.

There is a reliance on cloud based machines to install and customize Spark.



The  
solution

Spark and MLlib can be installed on any Dataproc cluster.  
Customizations can be applied to clusters via initialization actions.  
Use Cloud Monitoring to monitor workflows.



The  
value

Resources can focus on data, not cluster creation and management.  
Integration with Google Cloud unlocks new Spark features.

# Agenda

Introduction to Big Data Managed Services in the Cloud

Leverage Big Data Operations with Dataproc

[Lab: Dataproc: Qwik Start: Console](#)

Lab: Dataproc: Qwik Start:  
Command Line

Build Extract, Transform, and Load Pipelines using Dataflow



---

# Lab Intro

## Dataproc: Qwik Start - Console

Create a Dataproc cluster, run a simple Apache Spark job in the cluster, and modify the number of workers in the cluster using the Cloud Console.

The lab can be found [here](#).

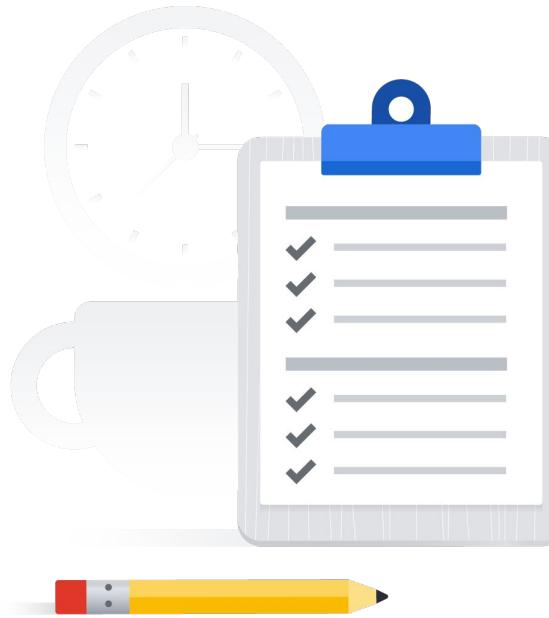


# Lab objectives

Create a cluster.

Submit a job.

View the job output.



---

# Lab Intro

Introduction to Dataproc: Hadoop  
and Spark on Google Cloud  
(Alternative)

Create a Dataproc cluster, submit a Spark job,  
and shut down your cluster.

The lab can be found [here](#).

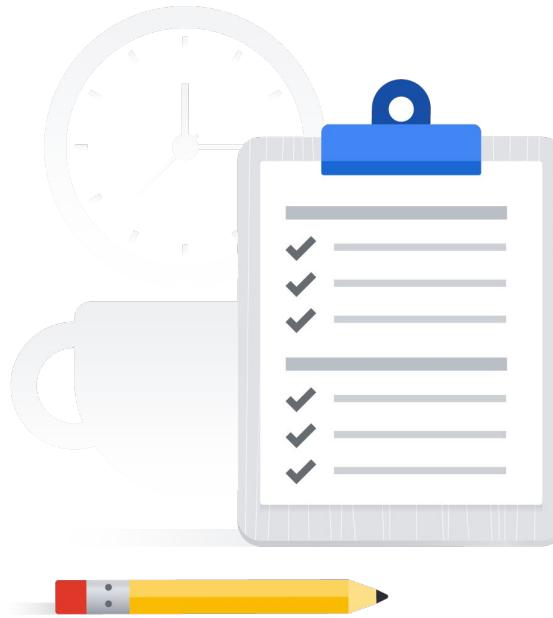


# Lab objectives

Create a Dataproc cluster.

Submit a Spark job to the cluster.

Shut down the cluster.



# Agenda

Introduction to Big Data Managed Services in the Cloud

Leverage Big Data Operations with Dataproc

Lab: Dataproc: Qwik Start: Console

Lab: Dataproc: Qwik Start: Command Line

Build Extract, Transform, and Load Pipelines using Dataflow



---

# Lab Intro

## Dataproc: Qwik Start - Command Line

Create a Dataproc cluster, run a simple Apache Spark job in the cluster, and modify the number of workers in the cluster using gcloud on Google Cloud.

The lab can be found [here](#).

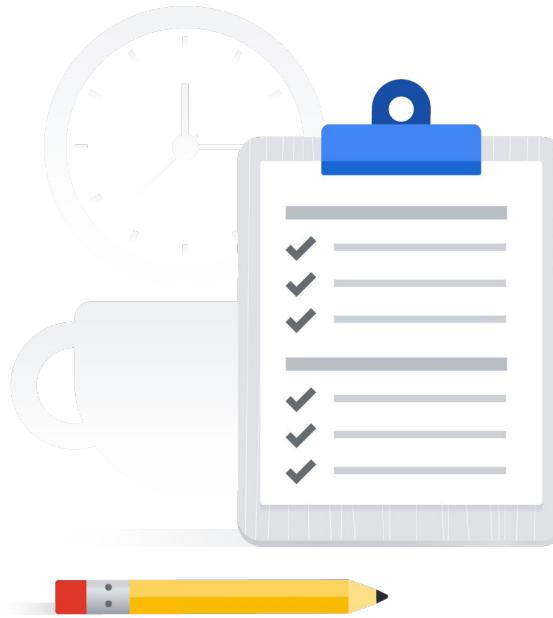


# Lab objectives

Create a cluster.

Submit a job.

Update a cluster.



---

## Lab Intro

Introduction to Dataproc: Hadoop  
and Spark on Google Cloud  
(Alternative)

Create a Dataproc cluster, submit a Spark job,  
and shut down your cluster.

The lab can be found [here](#).

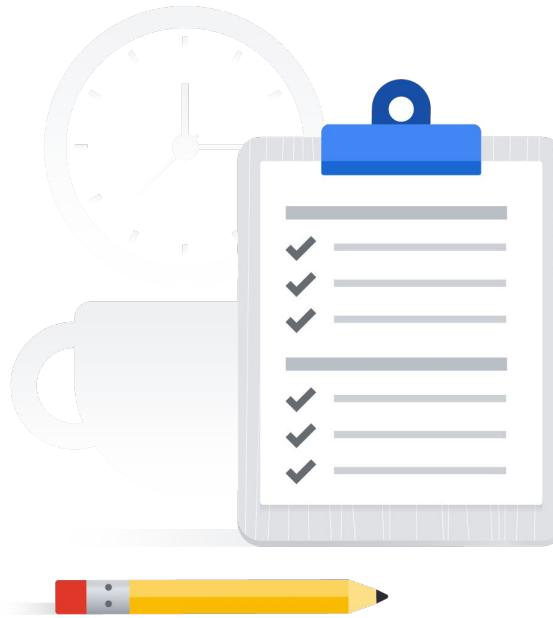


# Lab objectives

Create a Dataproc cluster.

Submit a Spark job to the cluster.

Shut down the cluster.



# Agenda

Introduction to Big Data Managed Services in the Cloud

Leverage Big Data Operations with Dataproc

Lab: Dataproc: Qwik Start: Console

Lab: Dataproc: Qwik Start: Command Line

Build Extract, Transform, and Load Pipelines using Dataflow



# Dataflow offers simplified stream and batch data processing



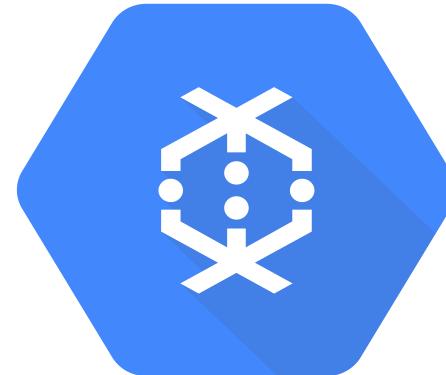
Unified programming model



Fully-managed service



Integrated



# Dataflow templates enable the rapid deployment of standard job types

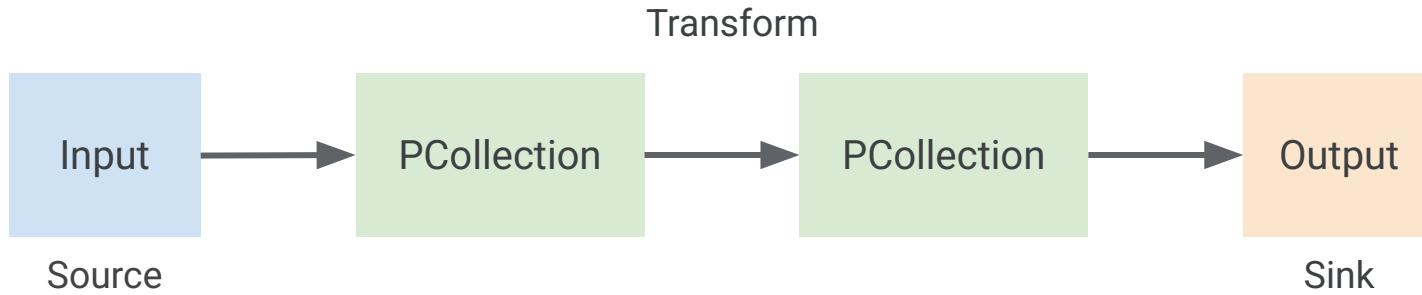
The screenshot shows the 'Create job from template' interface in the Google Cloud Platform Dataflow section. On the left, there's a 'Job name' input field containing 'my-job-name'. Below it is a 'Cloud Dataflow template' dropdown menu with the placeholder 'Select a template'. At the bottom are 'Run job' and 'Cancel' buttons. A large blue arrow points from the right side of the 'Select a template' dropdown towards a list of available templates on the right. This list is organized into sections: 'Get Started' (containing 'Word Count'), 'Process Data Continuously (stream)' (containing 'Cloud Pub/Sub Subscription to BigQuery', 'Cloud Pub/Sub Topic to BigQuery', 'Cloud Pub/Sub to Text Files on Cloud Storage', 'Cloud Pub/Sub to Avro Files on Cloud Storage', 'Cloud Pub/Sub to Cloud Pub/Sub', 'Text Files on Cloud Storage to Cloud Pub/Sub', 'Text Files on Cloud Storage to BigQuery', and 'Data Masking/Tokenization using Cloud DLP from GCS to BigQuery'), 'Process Data in Bulk (batch)' (containing 'Text Files on Cloud Storage to Cloud Pub/Sub', 'Text Files on Cloud Storage to BigQuery', 'Cloud Datastore to Text Files on Cloud Storage', 'Text Files on Cloud Storage to Cloud Datastore', 'Cloud Spanner to Text Files on Cloud Storage', 'Cloud Spanner to Avro Files on Cloud Storage', 'Avro Files on Cloud Storage to Cloud Spanner', 'Cloud BigTable to SequenceFile Files on Cloud Storage', 'SequenceFile Files on Cloud Storage to Cloud BigTable', 'Cloud Bigtable to Avro Files on Cloud Storage', 'Avro Files on Cloud Storage to Cloud Bigtable', and 'Jdbc to BigQuery').

Get Started
Word Count

Process Data Continuously (stream)
Cloud Pub/Sub Subscription to BigQuery
Cloud Pub/Sub Topic to BigQuery
Cloud Pub/Sub to Text Files on Cloud Storage
Cloud Pub/Sub to Avro Files on Cloud Storage
Cloud Pub/Sub to Cloud Pub/Sub
Text Files on Cloud Storage to Cloud Pub/Sub
Text Files on Cloud Storage to BigQuery
Data Masking/Tokenization using Cloud DLP from GCS to BigQuery

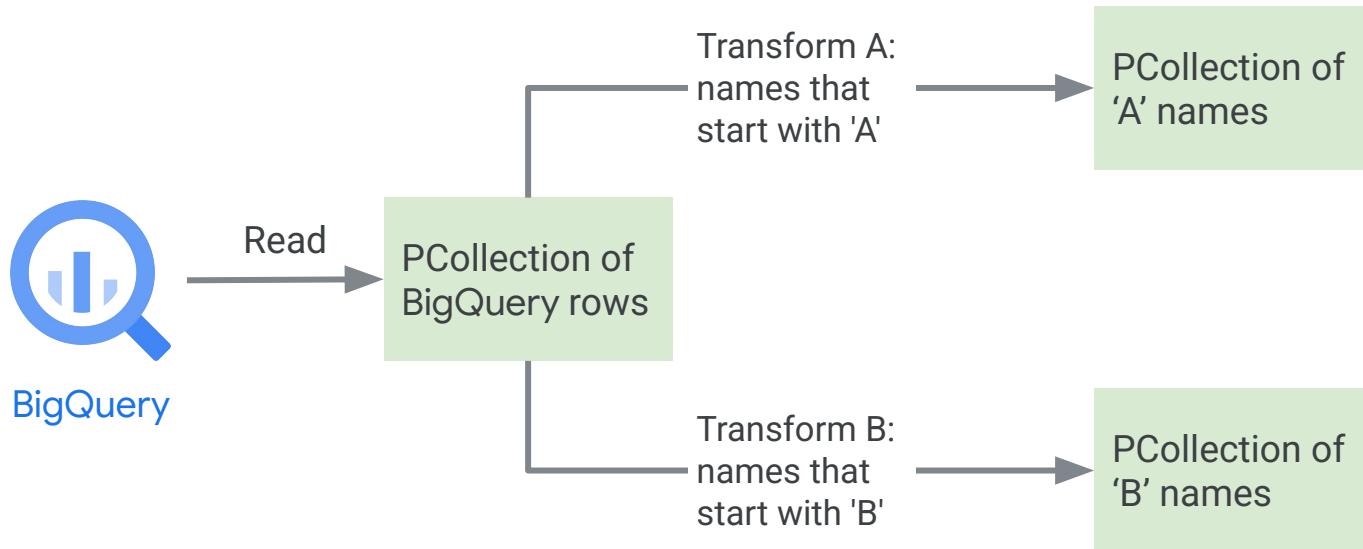
Process Data in Bulk (batch)
Text Files on Cloud Storage to Cloud Pub/Sub
Text Files on Cloud Storage to BigQuery
Cloud Datastore to Text Files on Cloud Storage
Text Files on Cloud Storage to Cloud Datastore
Cloud Spanner to Text Files on Cloud Storage
Cloud Spanner to Avro Files on Cloud Storage
Avro Files on Cloud Storage to Cloud Spanner
Cloud BigTable to SequenceFile Files on Cloud Storage
SequenceFile Files on Cloud Storage to Cloud BigTable
Cloud Bigtable to Avro Files on Cloud Storage
Avro Files on Cloud Storage to Cloud Bigtable
Jdbc to BigQuery

# Understanding pipelines

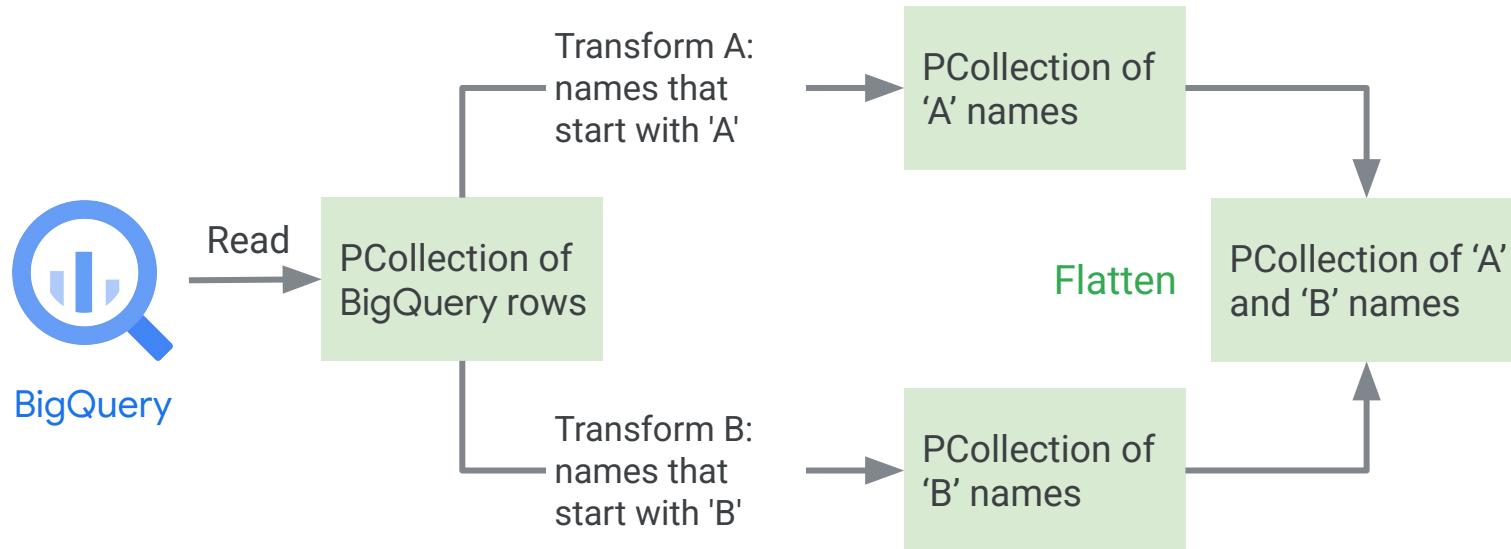


Simple pipeline example

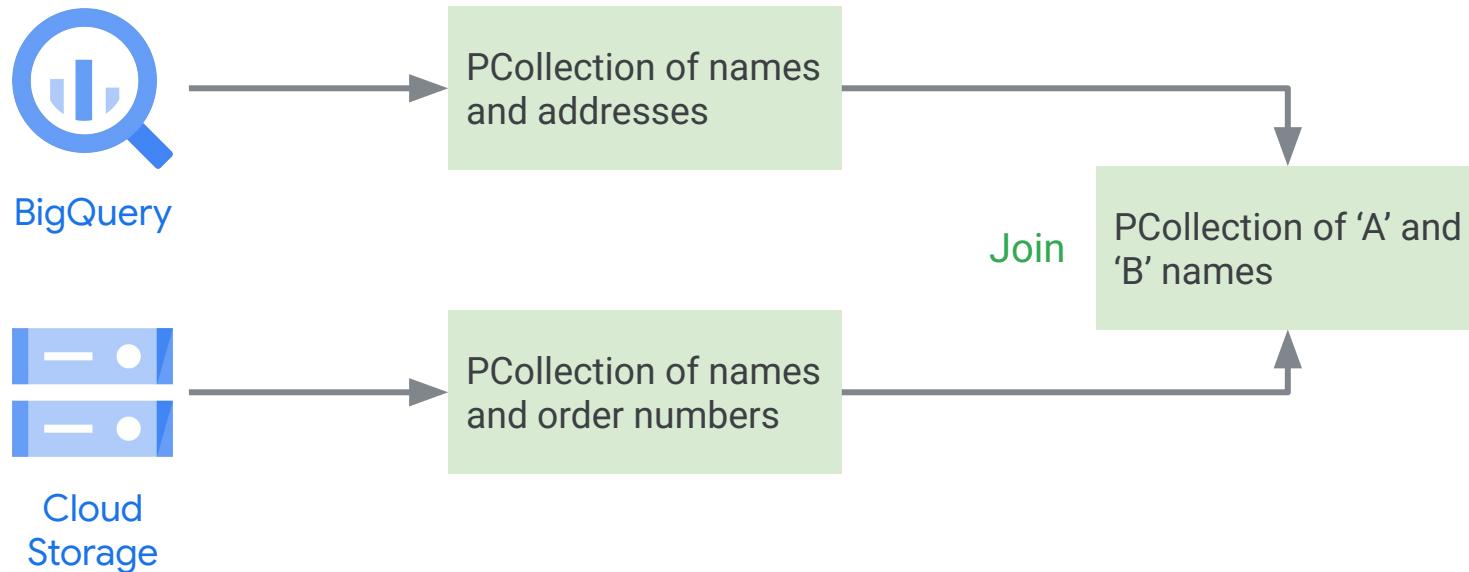
# A multiple transform pipeline extract



# A merge pipeline extract



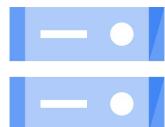
# A multiple input pipeline



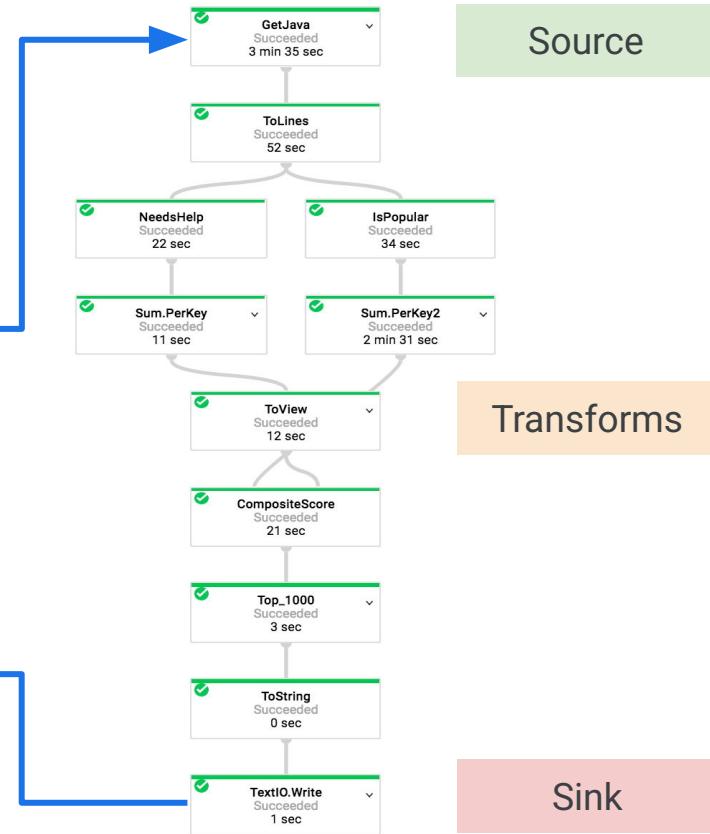
# A simple but real Dataflow pipeline example



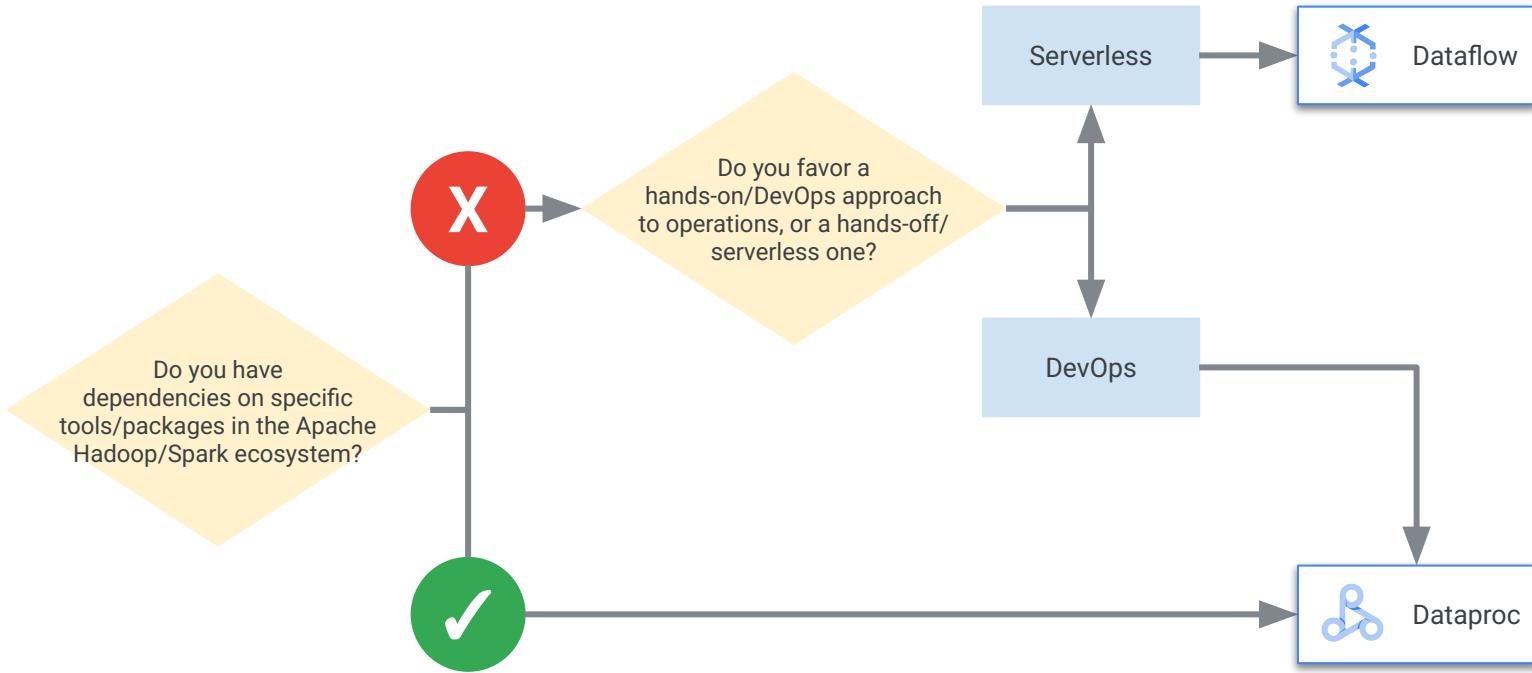
BigQuery



Cloud Storage



# Dataproc versus Dataflow



# Agenda

Lab: Dataflow: Qwik Start -  
Templates

Lab: Dataflow: Qwik Start - Python

BigQuery, Google's Enterprise Data  
Warehouse

Lab: Dataprep: Qwik Start

Quiz

Summary



---

# Lab Intro

## Dataflow: Qwik Start - Templates

Create a streaming pipeline using a  
Google-provided Dataflow template.

You can find the lab [here](#).

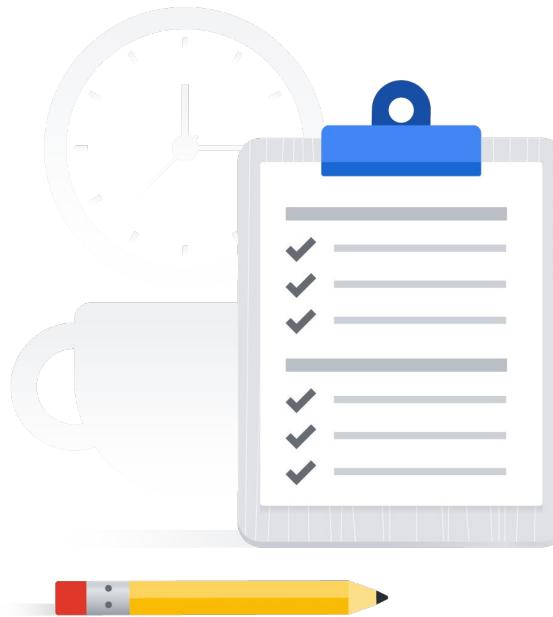


# Lab objectives

Create a BigQuery dataset and table using Cloud Shell and/or the Cloud Console.

Run the pipeline.

Submit a query.



# Agenda

Lab: Dataflow: Qwik Start -  
Templates

[Lab: Dataflow: Qwik Start - Python](#)

BigQuery, Google's Enterprise Data  
Warehouse

Lab: Dataprep: Qwik Start

Quiz

Summary



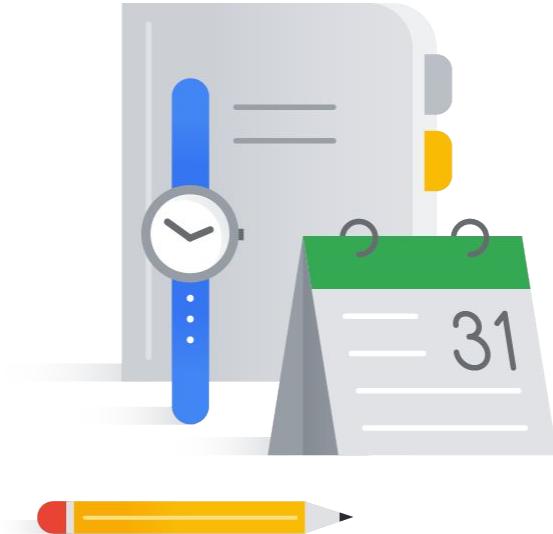
---

Lab!

## Serverless Data

### Processing with Dataflow - Writing an ETL Pipeline using Apache Beam and Dataflow (Python)

[https://www.cloudskillsboost.google/focuses/64780?catalog\\_rank=%7B%22rank%22%3A3%2C%22num\\_filters%22%3A0%2C%22has\\_search%22%3Atrue%7D&parent=catalog&search\\_id=38607186](https://www.cloudskillsboost.google/focuses/64780?catalog_rank=%7B%22rank%22%3A3%2C%22num_filters%22%3A0%2C%22has_search%22%3Atrue%7D&parent=catalog&search_id=38607186)



---

# Lab Intro

## Dataflow: Qwik Start - Python

Set up a Python development environment, get the Dataflow SDK for Python, and run an example pipeline using the Cloud Console.

You can find the lab [here](#).

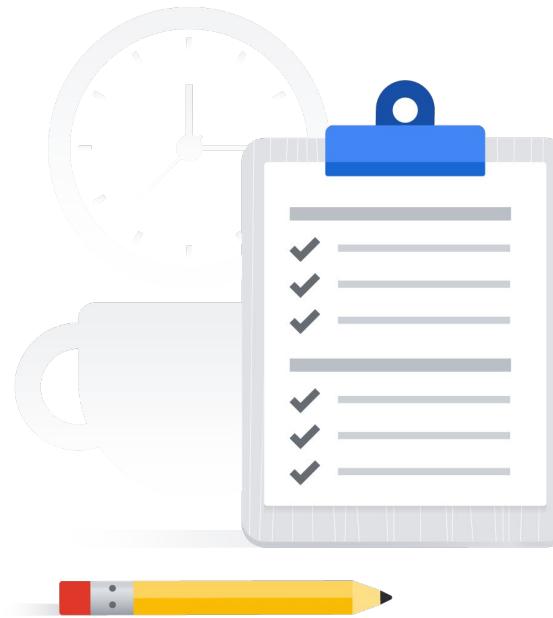


# Lab objectives

Set up a Python development environment.

Get the Dataflow SDK for Python.

Run an example pipeline using the Cloud Console.



---

# Lab Intro

## Processing Data with Dataflow (Alternative)

Process a set of text files from a real-time real world historical dataset using Python and Dataflow, and use BigQuery to analyze some features.

You can find the lab [here](#).



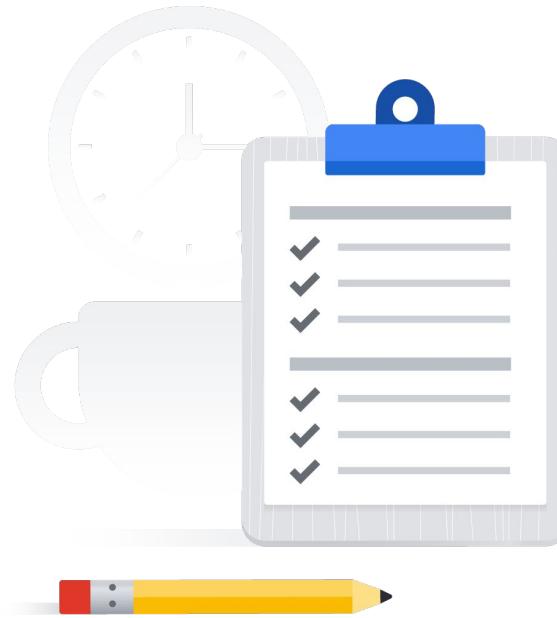
# Lab objectives

Configure a Python application to create a simulated real-time data stream from historical data.

Use Apache Beam locally to test Dataflow locally.

Use Apache Beam to process data using Dataflow to create a simulated real-time data set.

Query the simulated real-time data stream using BigQuery.



# Agenda

Lab: Dataflow: Qwik Start -  
Templates

Lab: Dataflow: Qwik Start - Python

[BigQuery, Google's Enterprise Data  
Warehouse](#)

Lab: Dataprep: Qwik Start

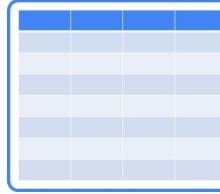
Quiz

Summary

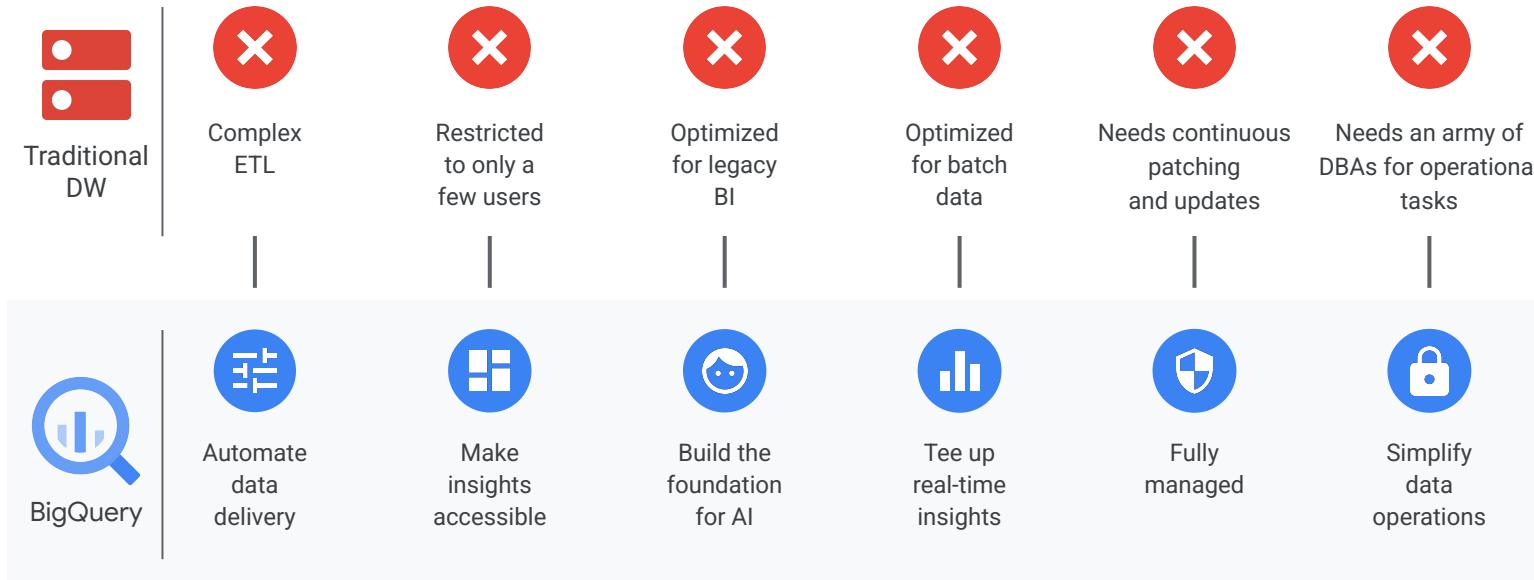




# BigQuery is Google's data warehouse solution

				
Data warehouse	Data mart	Data lake	Tables and views	Grants
BigQuery replaces a typical data warehouse hardware setup	BigQuery organizes data tables into units called datasets	BigQuery defines schemas and issues queries directly on external data sources	Function the same way as in a traditional data warehouse	Cloud IAM grants permission to perform specific actions

# BigQuery is a modern data warehouse that changes the conventional mode of data warehousing



# BigQuery ML enables users to create and execute ML models in BigQuery using standard SQL queries

- 1 Execute ML initiatives without moving data from BigQuery.
- 2 Iterate on models in SQL in BigQuery to increase development speed.
- 3 Automate common ML tasks and hyperparameter tuning.



# BigQuery is a fully-managed service

 Data aging

 Query engine optimization

 Storage management

 Hardware

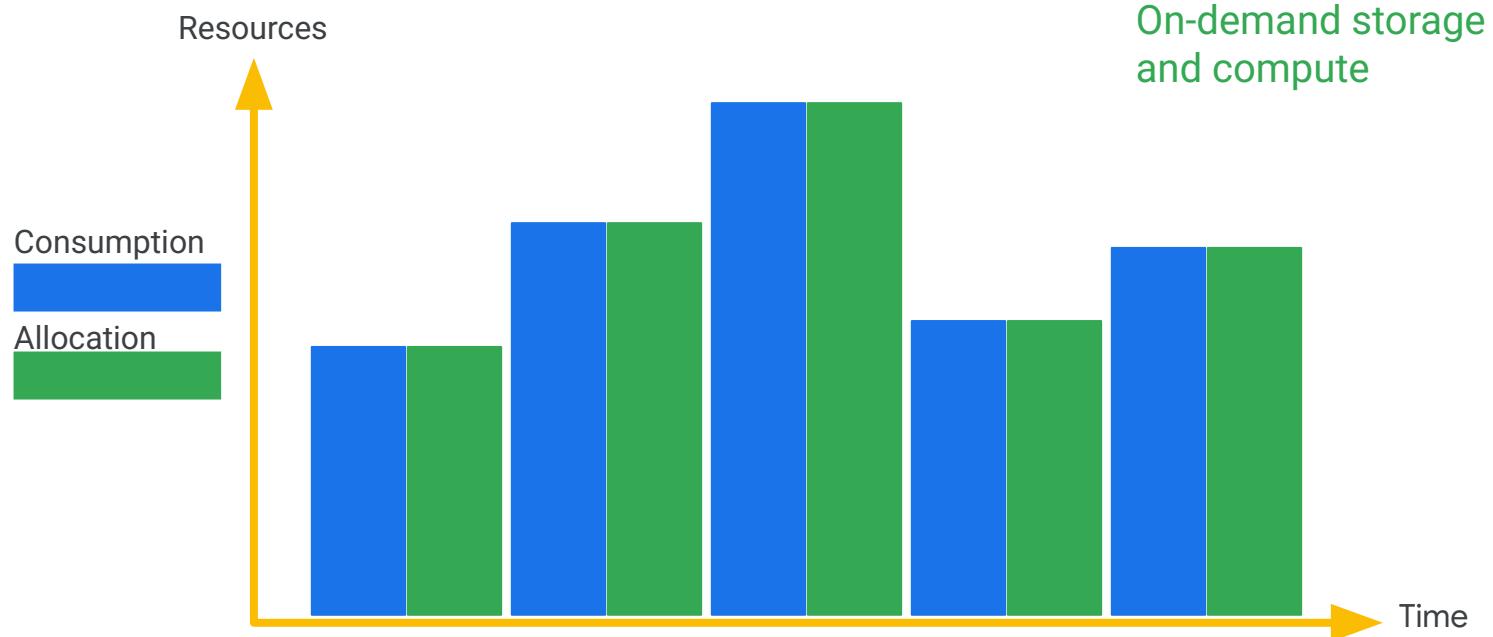
 Fault recovery

 Updates

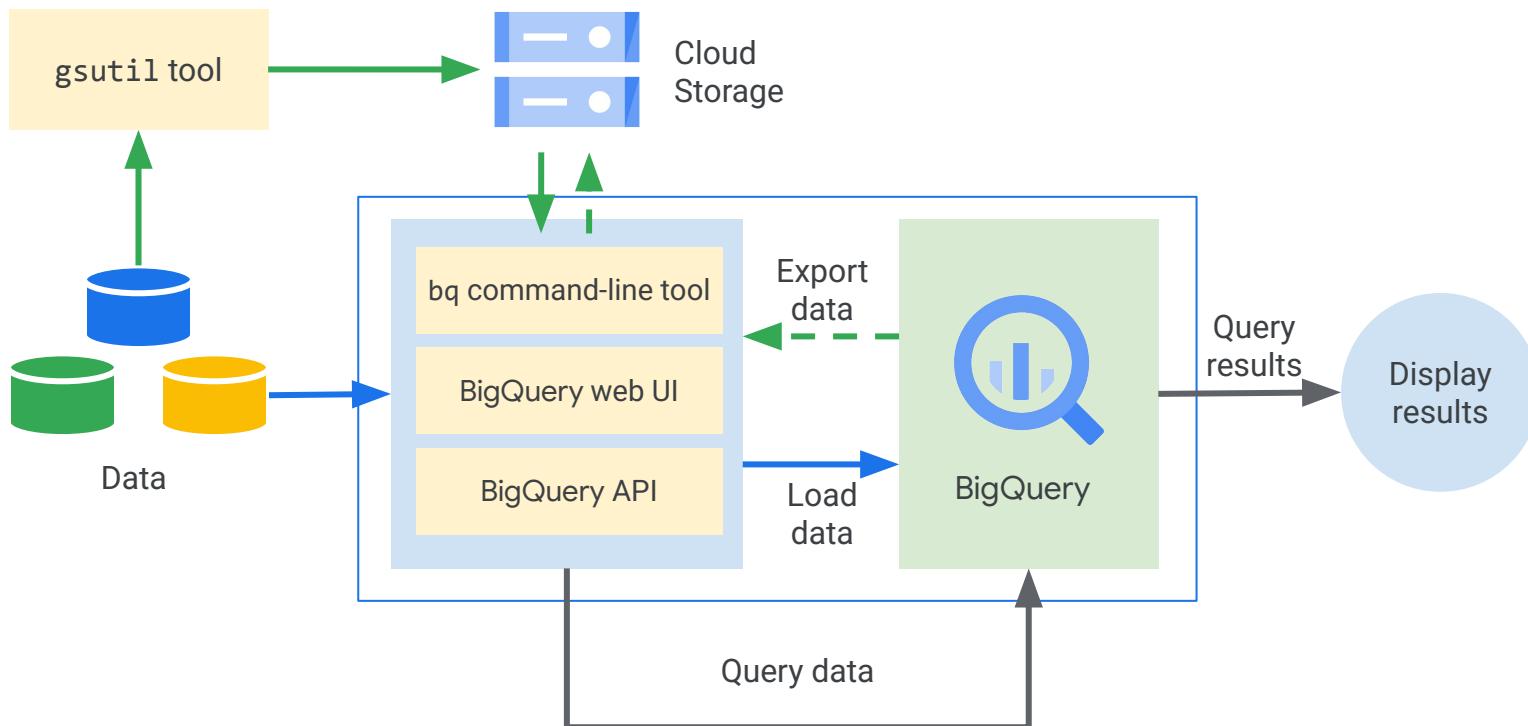
Free up real people-hours by not  
having to worry about common tasks.



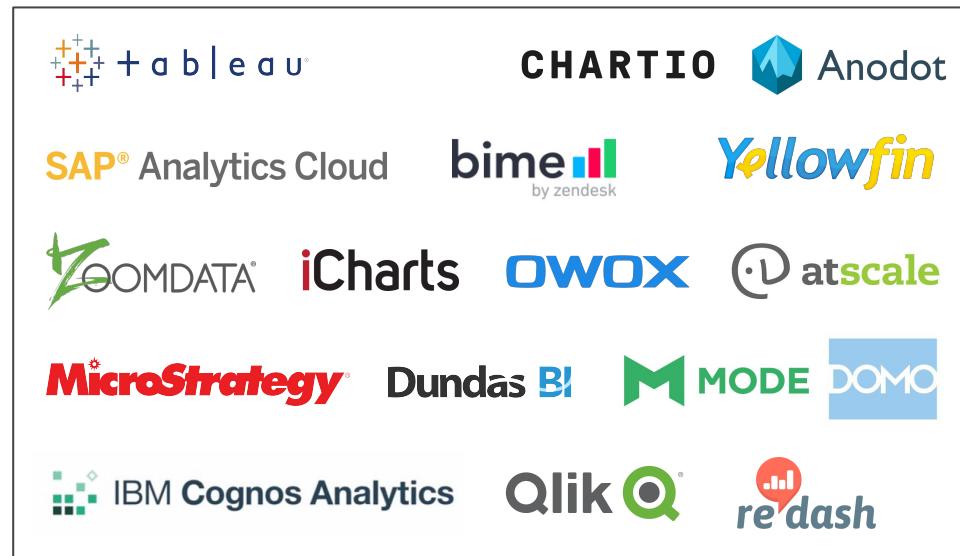
# You don't need to provision resources before using BigQuery



# Loading data into BigQuery



There are various ways that you can connect to BigQuery and analyze the data



Data Analysis and Visualization Partners

# BQ examples

## BigQuery - Google Cloud Big Data → GitHub queries

```
#standardSQL  
SELECT SUM(copies) FROM `bigquery-public-data.github_repos.sample_contents`  
WHERE NOT binary AND content LIKE '%This should never happen'
```

```
#standardSQL  
SELECT SUM(copies) FROM `bigquery-public-data.github_repos.sample_contents`  
WHERE NOT binary AND (content LIKE '%This should never happen%' OR content  
LIKE '%FIXME%' OR content LIKE '%TODO%')
```

---

# Lab Intro

## Dataprep: Qwik Start

Use Dataprep to manipulate a dataset. You import datasets, correct mismatched data, transform data, and join data.

You can find the lab [here](#).

