Google Cloud

Building a Data Lake

Agenda

Introduction to Data Lakes

Data Storage and ETL options on GCP

Building a Data Lake using Cloud Storage

Securing Cloud Storage

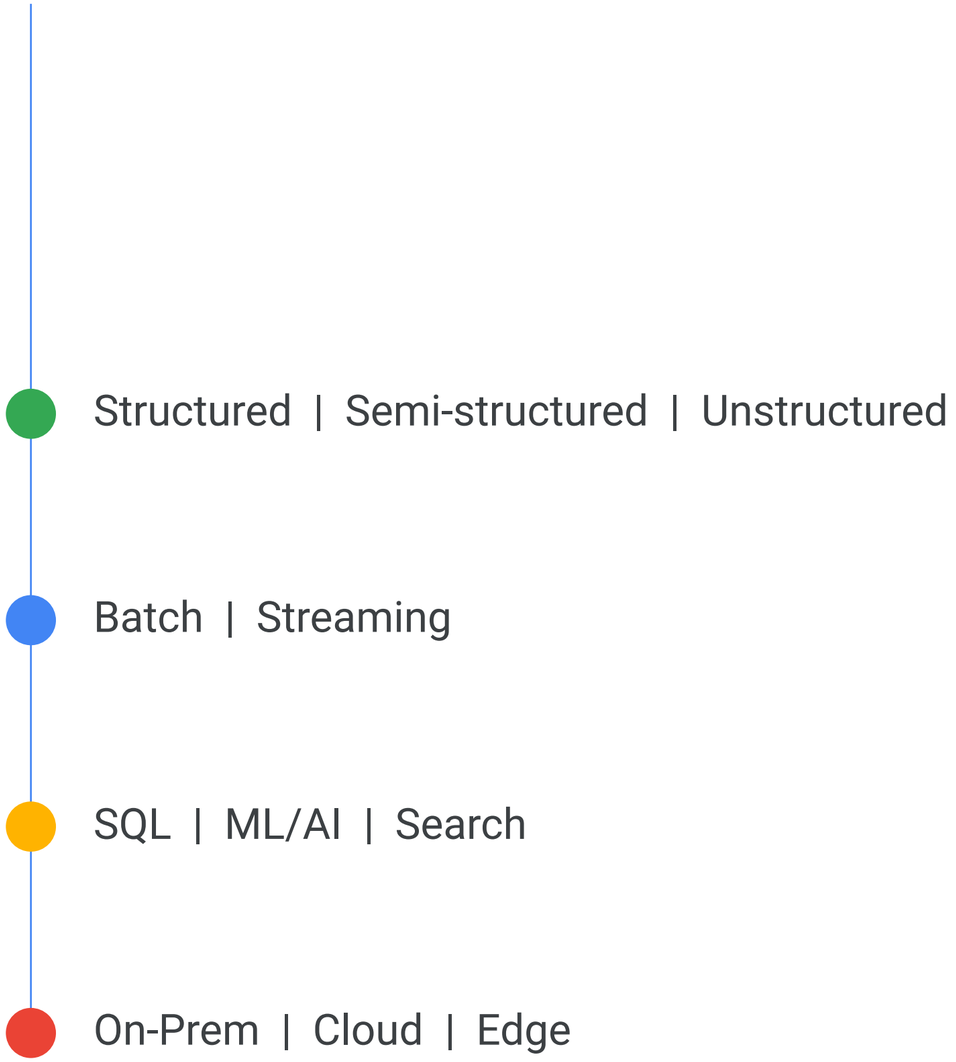
Storing All Sorts of Data Types

Cloud SQL as a relational Data Lake

Lab: Loading Taxi Data into Cloud SQL

What is a data lake?

A scalable and secure data platform that allows enterprises to ingest, store, process, and analyze any type or volume of information.



Google Cloud

Components of a Data Engineering ecosystem

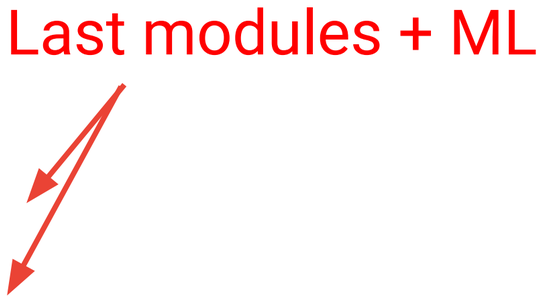
- Data sources
- Data sinks
  - **Central Data Lake repository** ← Our focus in this module
  - Data Warehouse
- Data pipelines (batch and streaming)
- High-level orchestration workflows

Components of a Data Engineering ecosystem

- Data sources
- Data sinks
  - Central Data Lake repository
  - **Data Warehouse**
- Data pipelines (batch and streaming)
- High-level orchestration workflows

Components of a Data Engineering ecosystem

- Data sources
- Data sinks
  - Central Data Lake repository
  - Data Warehouse
- Data pipelines (batch and streaming)
- High-level orchestration workflows



Data Engineering is like Civil Engineering

1. Raw Materials need to be brought to the job site (into the Data Lake)

Transform raw materials into a useful form

1. Raw Materials need to be brought to the job site (into the Data Lake)

2. Materials need to be cut and transformed for purpose and stored (pipelines to data sinks)

The new building is the new insight, ML model, etc.

1. Raw Materials need to be brought to the job site (into the Data Lake)

2. Materials need to be cut and transformed for purpose and stored (pipelines to data sinks)

3. The actual building is the new insight or ML model etc.

An orchestrator governs all aspects of the workflow

1. Raw Materials need to be brought to the job site (into the Data Lake)

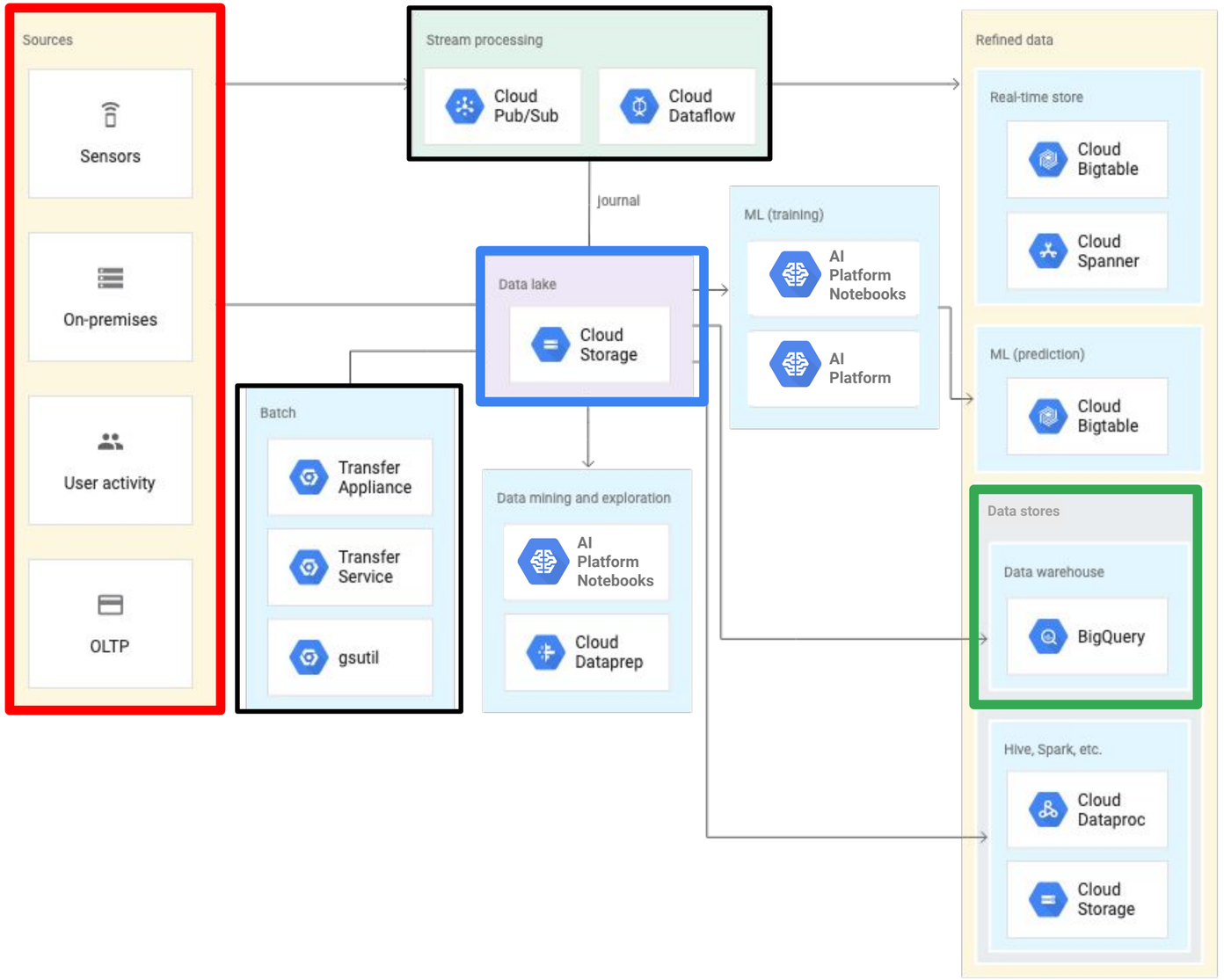
2. Materials need to be cut and transformed for purpose and stored (pipelines to data sinks)

3. The actual building is the new insight or ML model etc.

4. The supervisor directs all aspects and teams on the project (workflow orchestration)

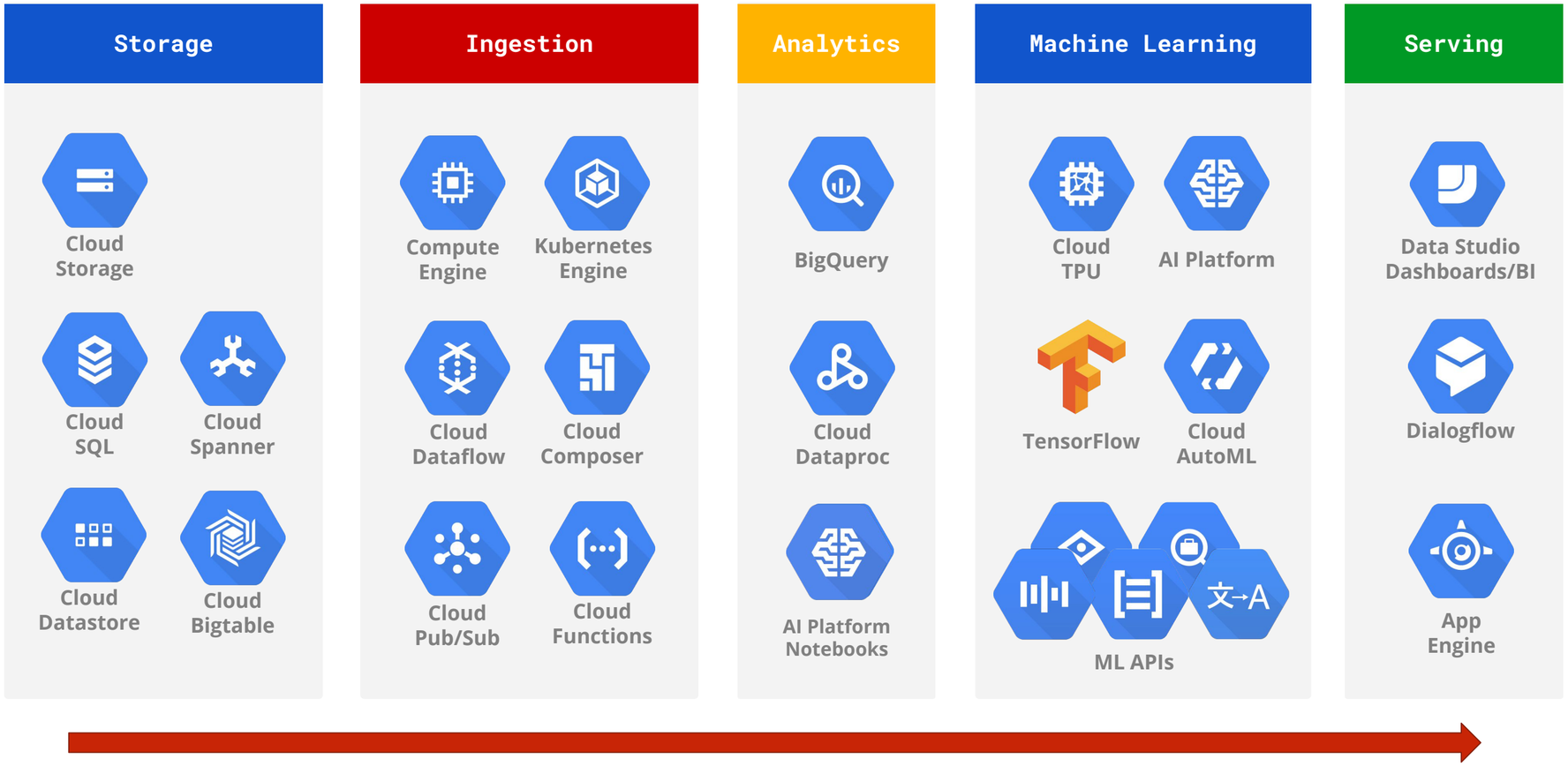
Example Architecture

1. **Data sources**
2. **Data Lake**
3. **Data Pipelines**
4. **Data Warehouse**
5. Used for ML and analytics workloads



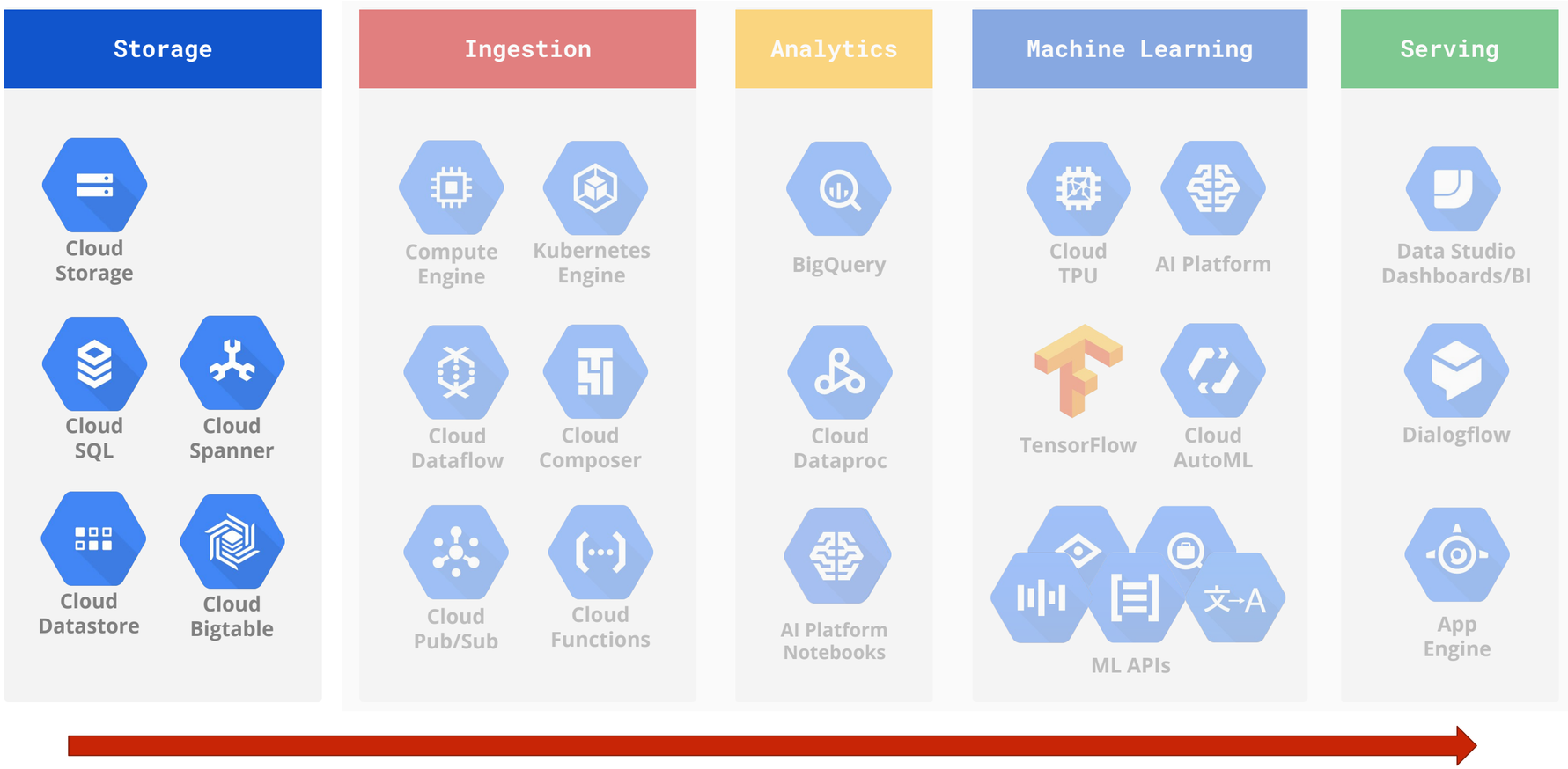
Google Cloud

The suite of big data products on Google Cloud Platform



Google Cloud

You will build scalable, durable, Data Lakes with GCP storage solutions



Google Cloud

Data lake versus data warehouse

A data lake is a capture of every aspect of your business operation. The data is stored in its natural/raw format, usually as object blobs or files.

- Retain all data in its native format
- Support all data types and all users
- Adapt to changes easily
- Tends to be application-specific

Data lake versus data warehouse

In contrast, a data warehouse typically has the following characteristics:

- Typically loaded only after a use case is defined
- Processed/organized/transformed
- Provide faster insights
- Current/historical data for reporting
- Tends to have consistent schema shared across applications

Agenda

Introduction to Data Lakes

Data Storage and ETL options on GCP

Building a Data Lake using Cloud Storage

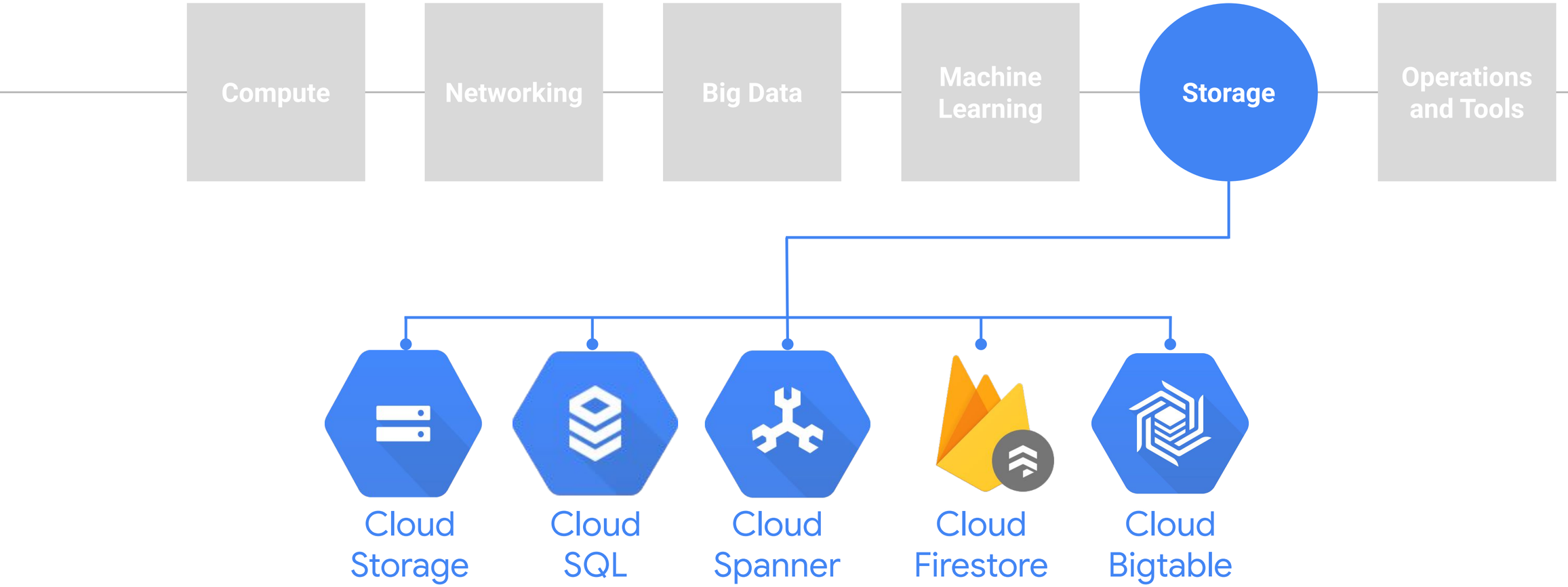
Securing Cloud Storage

Storing All Sorts of Data Types

Cloud SQL as a relational Data Lake

Lab: Loading Taxi Data into Cloud SQL

Storage options for your data on GCP



The path your data takes to get to the cloud depends on

- Where your data is now.

The path your data takes to get to the cloud depends on

- Where your data is now.
- How big your data is.

The path your data takes to get to the cloud depends on

- Where your data is now.
- How big your data is.
- Where it has to go.

The path your data takes to get to the cloud depends on

- Where your data is now.
- How big your data is.
- Where it has to go.
- How much transformation is needed.

The method you use to load data depends on how much transformation is needed

EL

Extract and Load

The method you use to load data depends on how much transformation is needed

EL	ELT
	
Extract and Load	Extract, Load, and Transform

The method you use to load data depends on how much transformation is needed

EL	ELT	ETL
		
Extract and Load	Extract, Load, and Transform	Extract, Transform, and Load

Agenda

Introduction to Data Lakes

Data Storage and ETL options on GCP

Building a Data Lake using Cloud Storage

Securing Cloud Storage

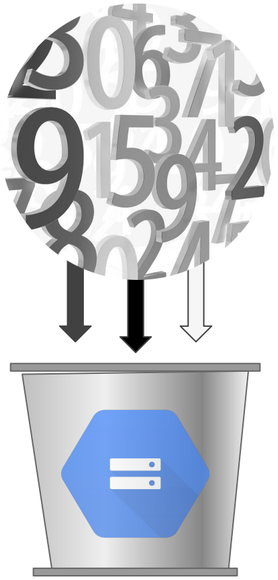
Storing All Sorts of Data Types

Cloud SQL as a relational Data Lake

Lab: Loading Taxi Data into Cloud SQL

Cloud Storage

Cloud Storage



Qualities that Cloud Storage contributes to data engineering solutions:

Persistence Durability Strong consistency Availability High throughput

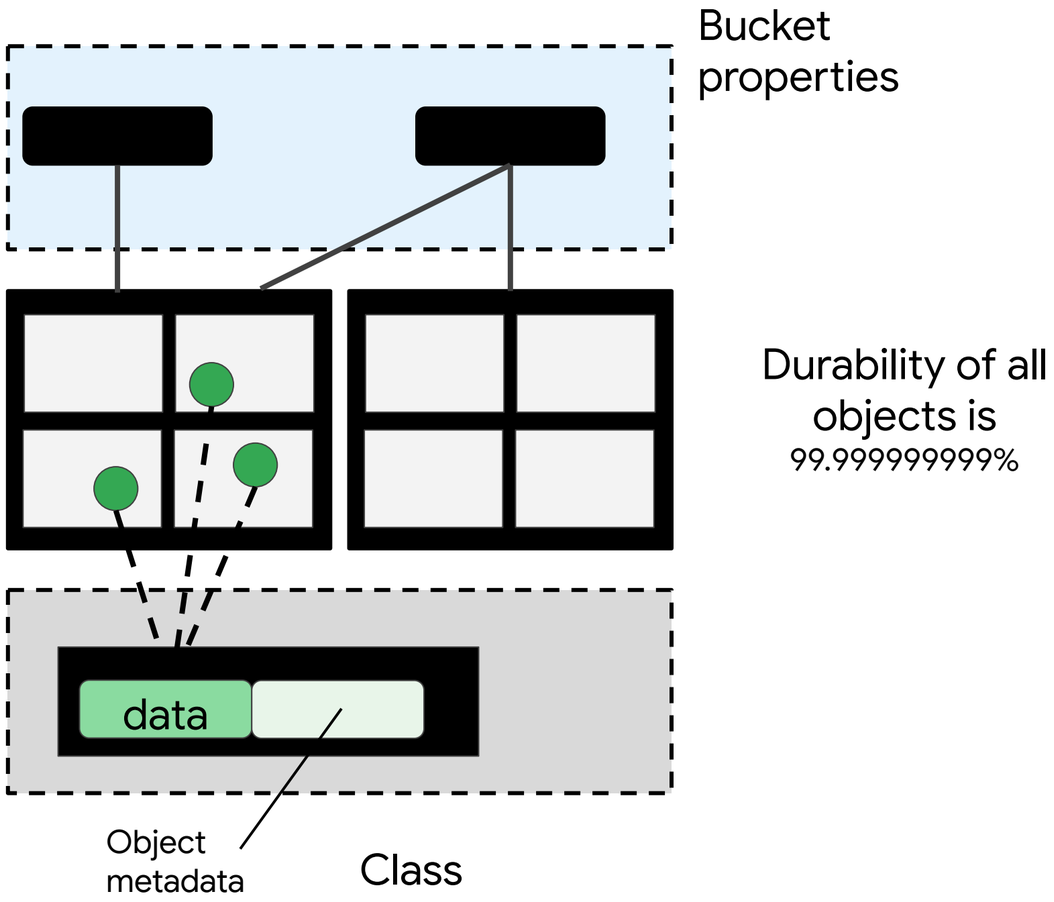
How does Cloud Storage work?

Single global namespace simplifies locating buckets and objects

Location to control latency

Durability and availability

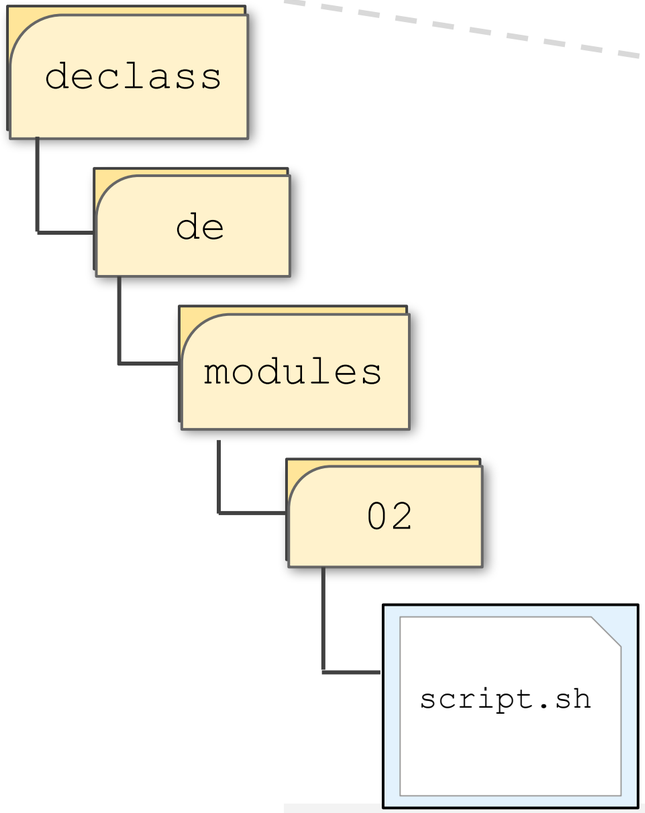
Long object names simulate structure

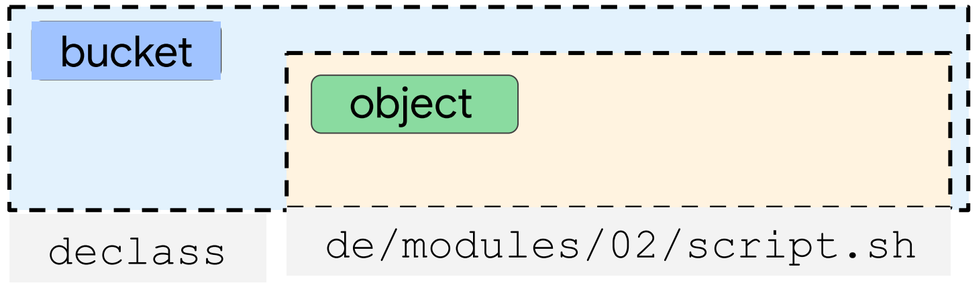


Overview of storage classes

	Standard	Nearline	Coldline	Archive
Use case	“Hot” data and/or stored for only brief periods of time like data-intensive computations	Infrequently accessed data like data backup, long-tail multimedia content, and data archiving	Infrequently accessed data that you read or modify at most once a quarter	Data archiving, online backup, and disaster recovery
Minimum storage duration	None	30 days	90 days	365 days
Retrieval cost	None	$0.01 per GB	$0.02 per GB	$0.05 per GB
Availability SLA	99.95% (multi/dual) 99.90% (region)	99.90% (multi/dual) 99.00% (region)		None
Durability	99.999999999%			

Cloud Storage simulates a file system





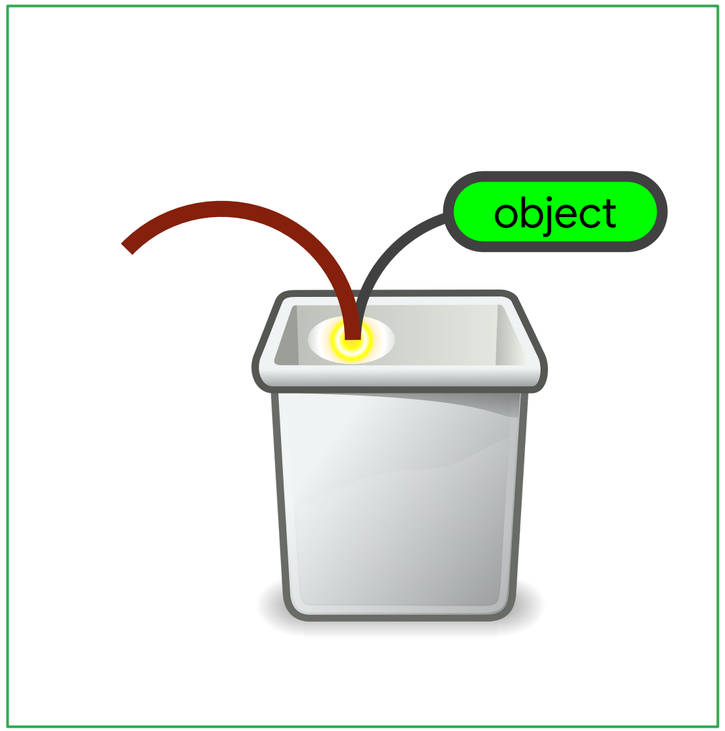
Long object name
simulates file structure

File access `gs://declass/de/modules/02/script.sh`

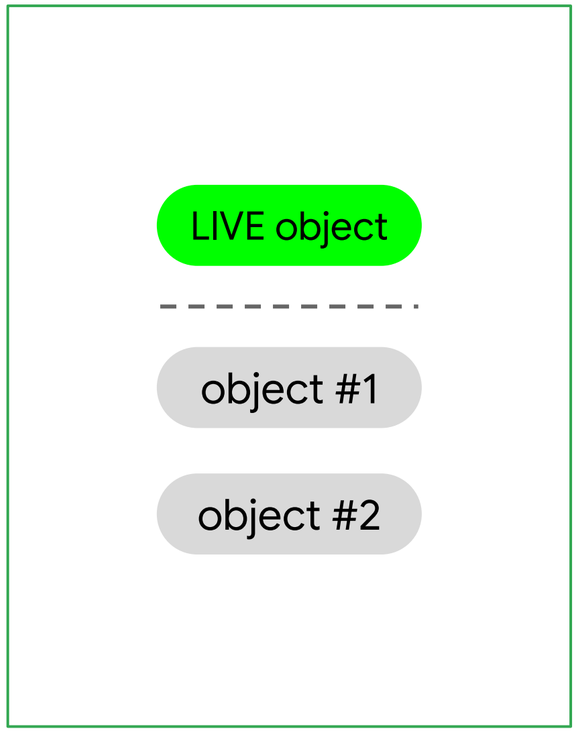
Web access `https://storage.cloud.google.com/declass/de/modules/02/script.sh`

Google Cloud

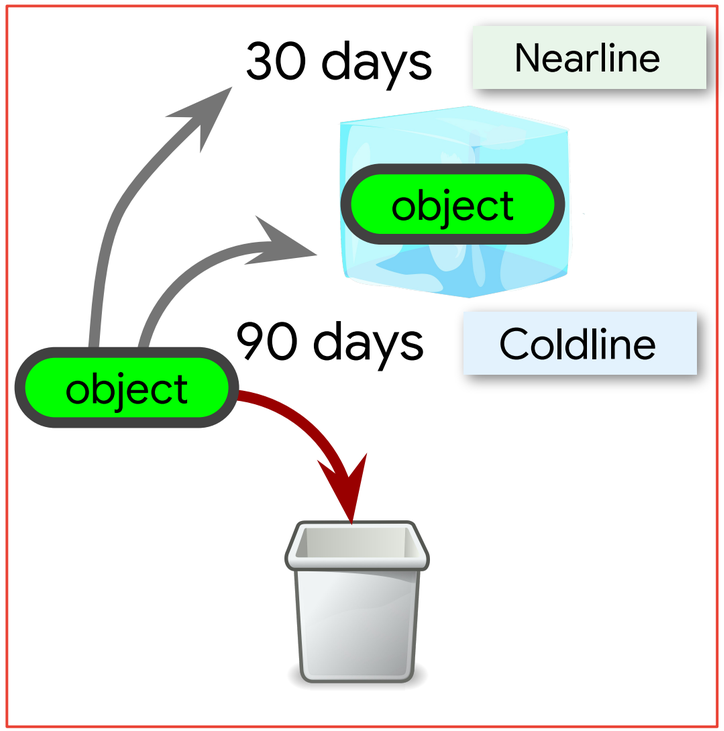
Cloud Storage has many object management features



Retention Policy



Versioning



Lifecycle Management

Agenda

Introduction to Data Lakes

Data Storage and ETL options on GCP

Building a Data Lake using Cloud Storage

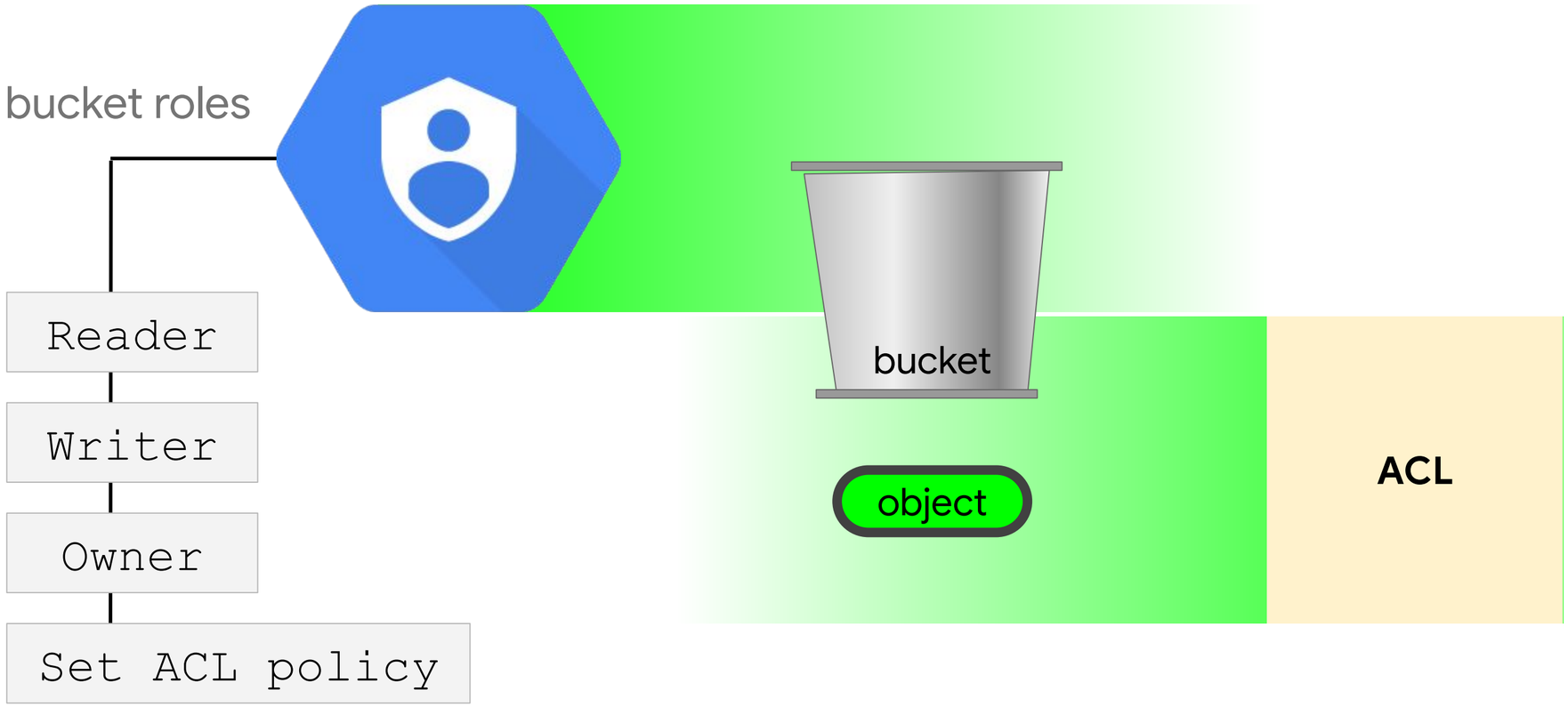
Securing Cloud Storage

Storing All Sorts of Data Types

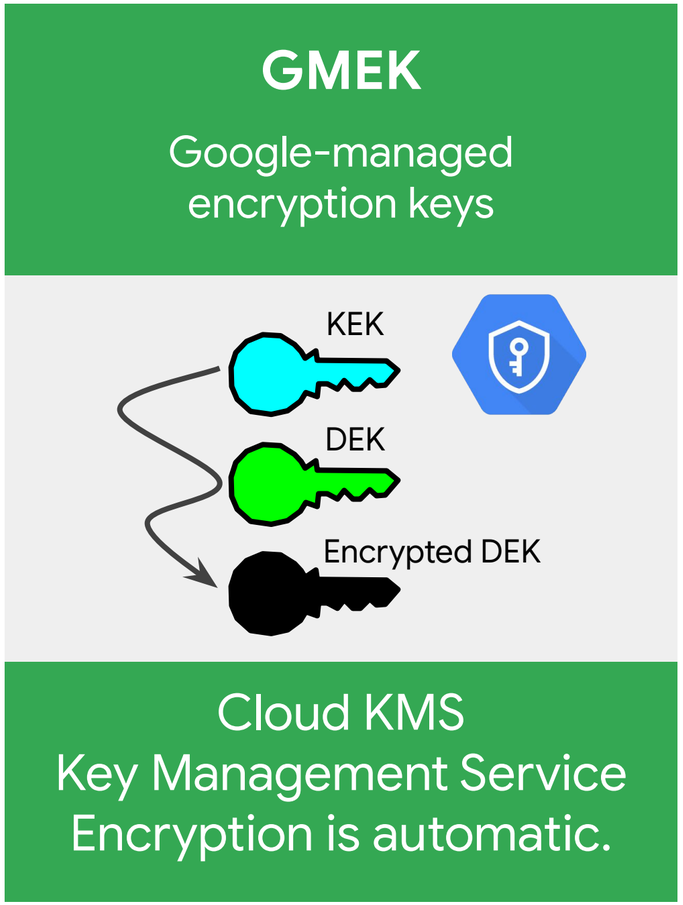
Cloud SQL as a relational Data Lake

Lab: Loading Taxi Data into Cloud SQL

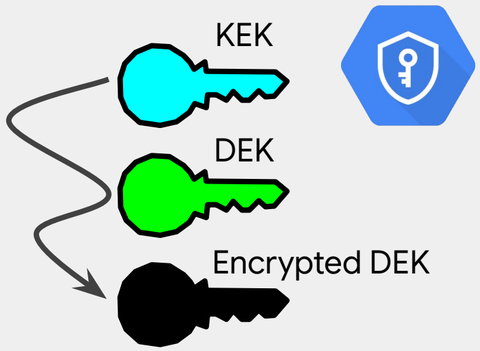
Controlling access with Cloud IAM and access lists



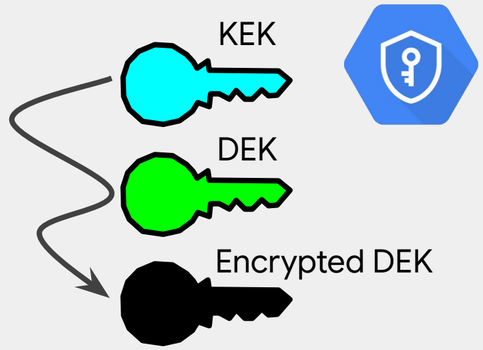
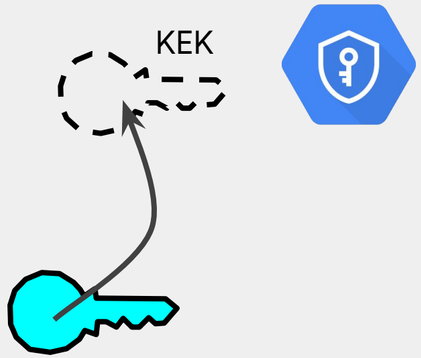
Data encryption options for many requirements



Data encryption options for many requirements

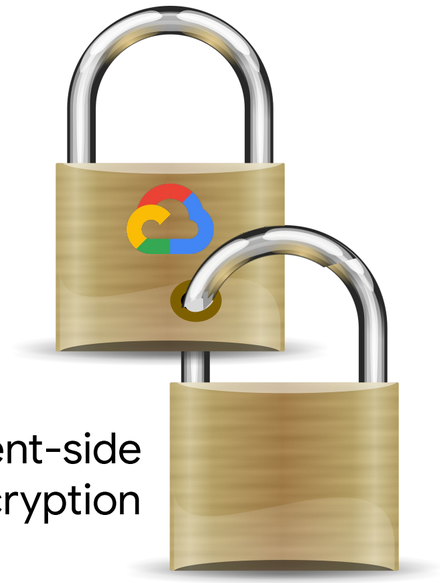
GMEK	CMEK
Google-managed encryption keys	Customer-managed encryption keys
	
Cloud KMS Key Management Service Encryption is automatic.	You control the creation and existence of the KEK key in Cloud KMS.

Data encryption options for many requirements

GMEK	CMEK	CSEK
Google-managed encryption keys	Customer-managed encryption keys	Customer-supplied encryption keys
		
Cloud KMS Key Management Service Encryption is automatic.	You control the creation and existence of the KEK key in Cloud KMS.	You provide the KEK key.

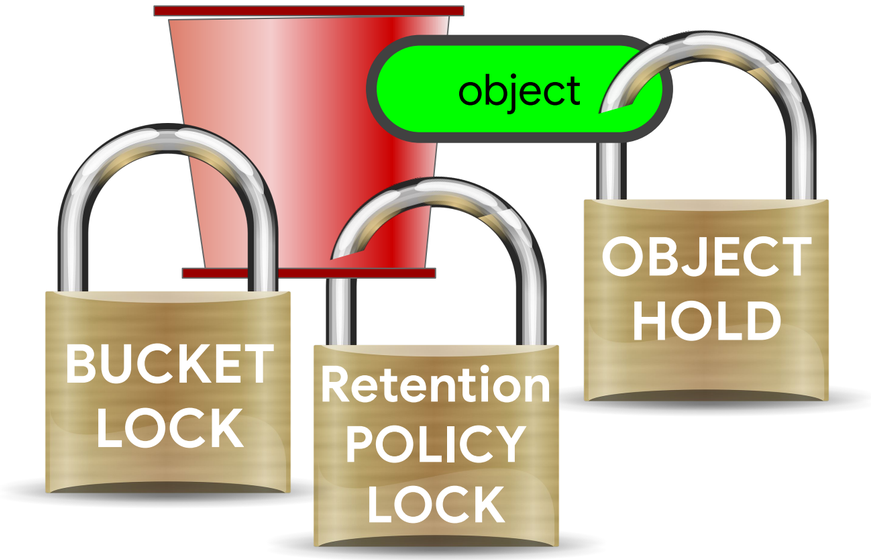
Cloud Storage supports many special use cases

Client-side encryption



Client-side encryption

Data locking for audits



+

Decompressive coding
Requester pays
Signed URLs for anonymous sharing
Period expirations
Composite objects
...

Agenda

Introduction to Data Lakes

Data Storage and ETL options on GCP

Building a Data Lake using Cloud Storage

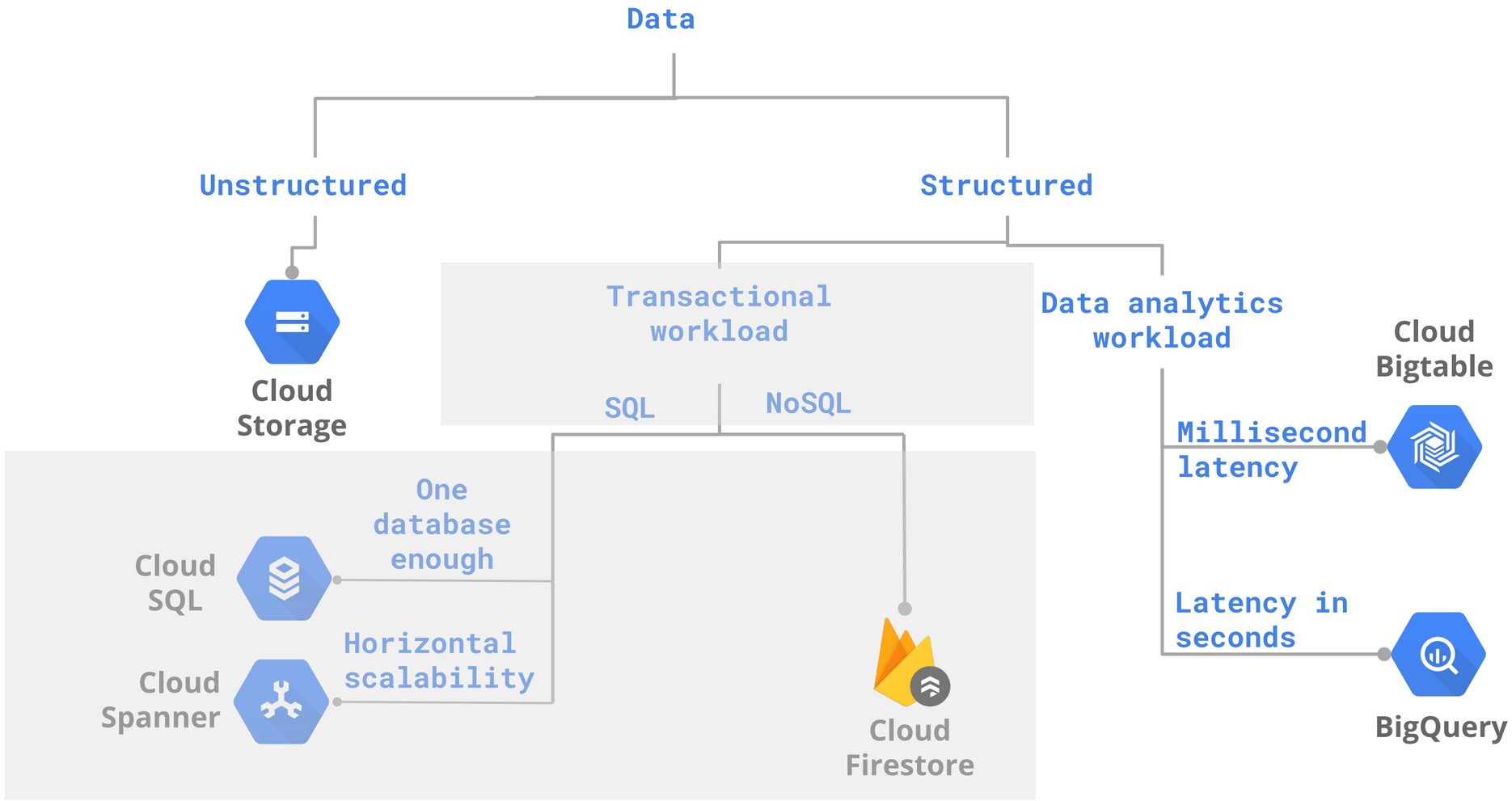
Securing Cloud Storage

Storing All Sorts of Data Types

Cloud SQL as a relational Data Lake

Lab: Loading Taxi Data into Cloud SQL

Different considerations for transactional workloads



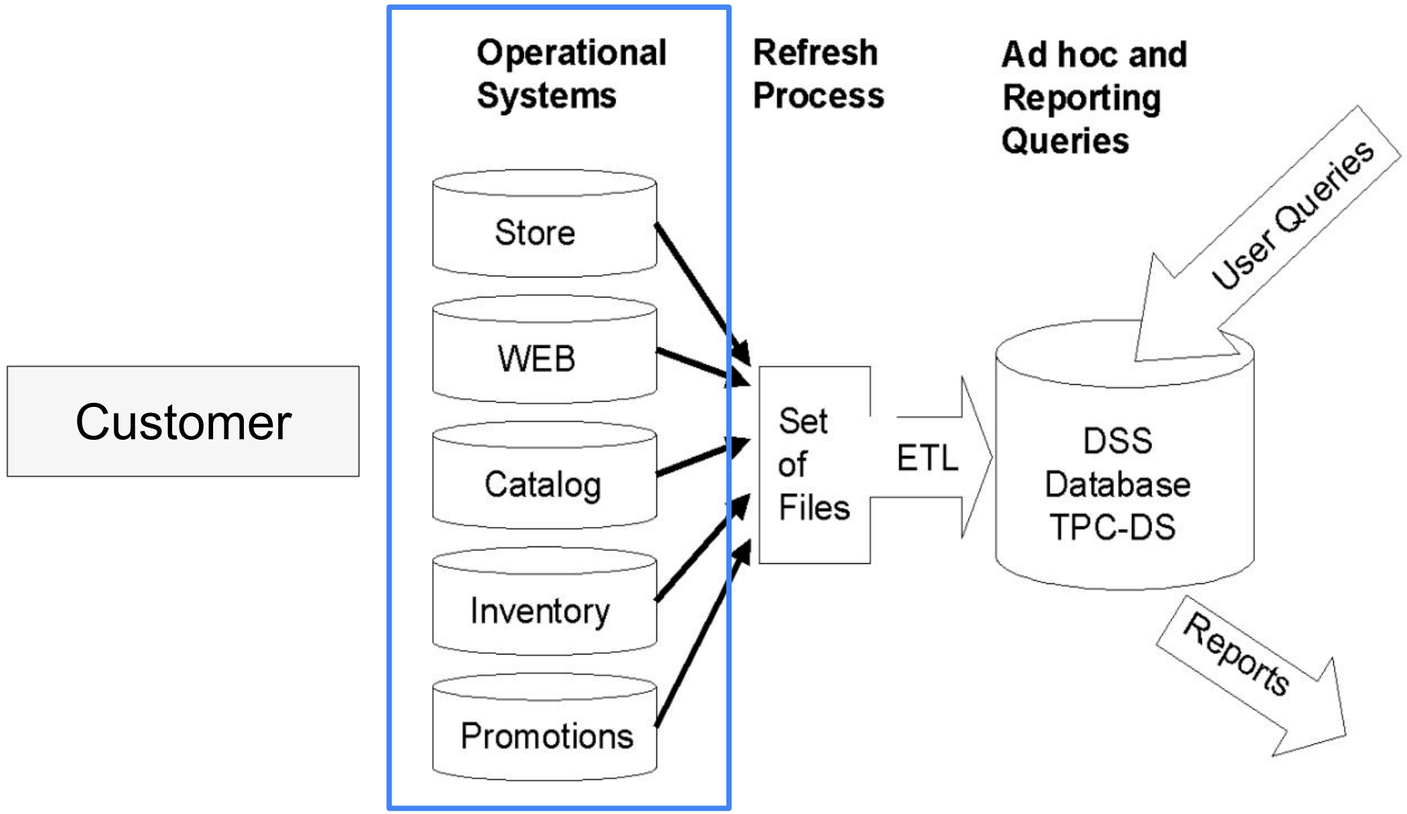
Transactional versus analytical

	Transactional	Analytical
Source of data	Operational data; OLTPs are the original source of the data	Consolidation data; OLAP data comes from the various OLTP databases
Purpose of data	Control and run fundamental business tasks	Help with planning, problem solving, and decision support
What the data shows	Reveals snapshot of ongoing business processes	Multi-dimensional views of various kinds of business activities
Inserts and updates	Short and fast inserts and updates initiated by end users	Periodic long-running batch jobs refresh the data
Queries	Relatively standardized and simple queries returning relatively few records	Often complex queries involving aggregations
Processing speed	Typically very fast	Depends on amount of data involved; improve query speed with indexes
Space requirements	Can be relatively small if historical data is archived	Larger, more indexes than OLTP

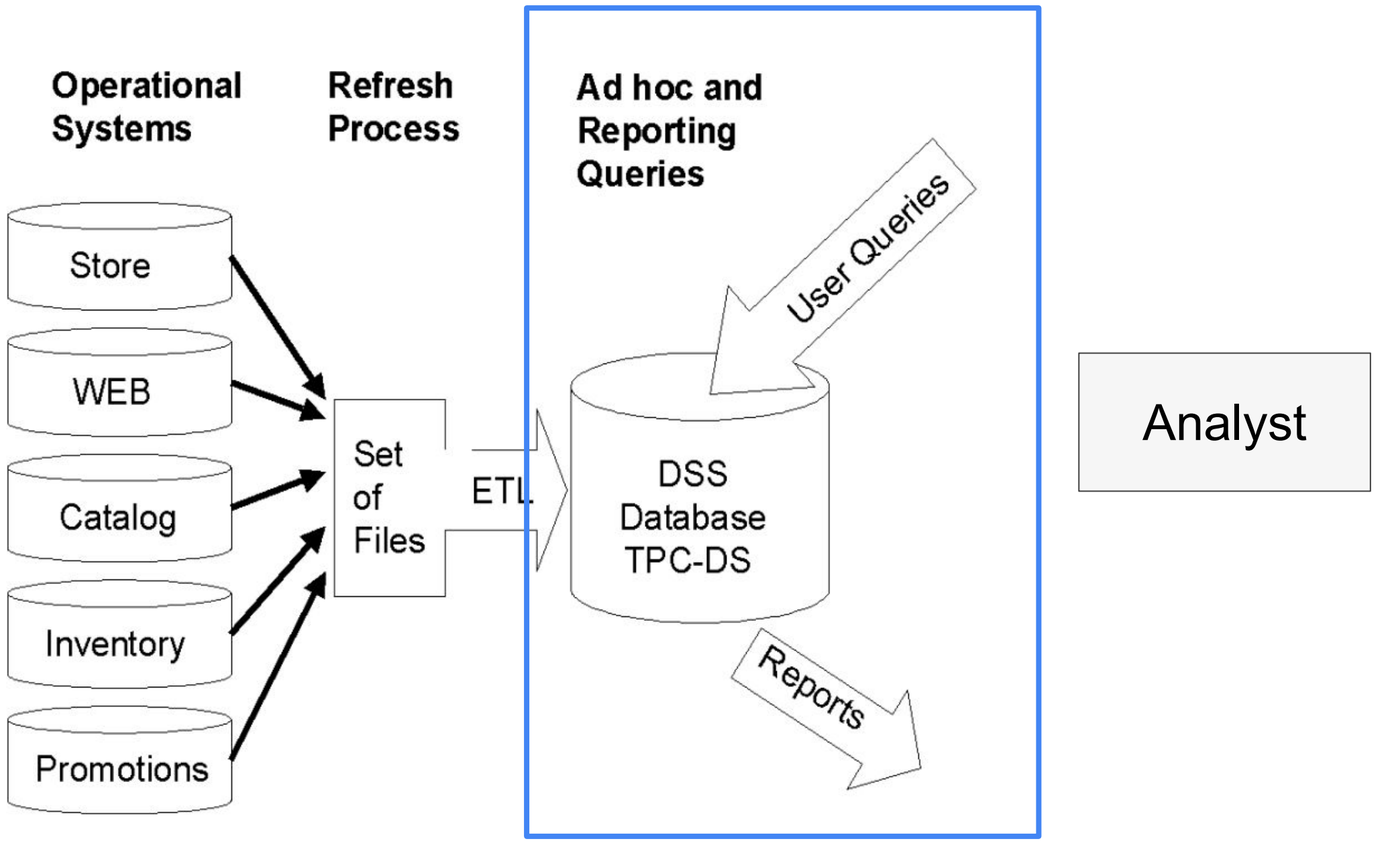
Transactional versus analytical

	Transactional	Analytical
Source of data	Operational data; OLTPs are the original source of the data	Consolidation data; OLAP data comes from the various OLTP databases
Purpose of data	Control and run fundamental business tasks	Help with planning, problem solving, and decision support
What the data shows	Reveals snapshot of ongoing business processes	Multi-dimensional views of various kinds of business activities
Inserts and updates	Short and fast inserts and updates initiated by end users	Periodic long-running batch jobs refresh the data
Queries	Relatively standardized and simple queries returning relatively few records	Often complex queries involving aggregations
Processing speed	Typically very fast	Depends on amount of data involved; improve query speed with indexes
Space requirements	Can be relatively small if historical data is archived	Larger, more indexes than OLTP

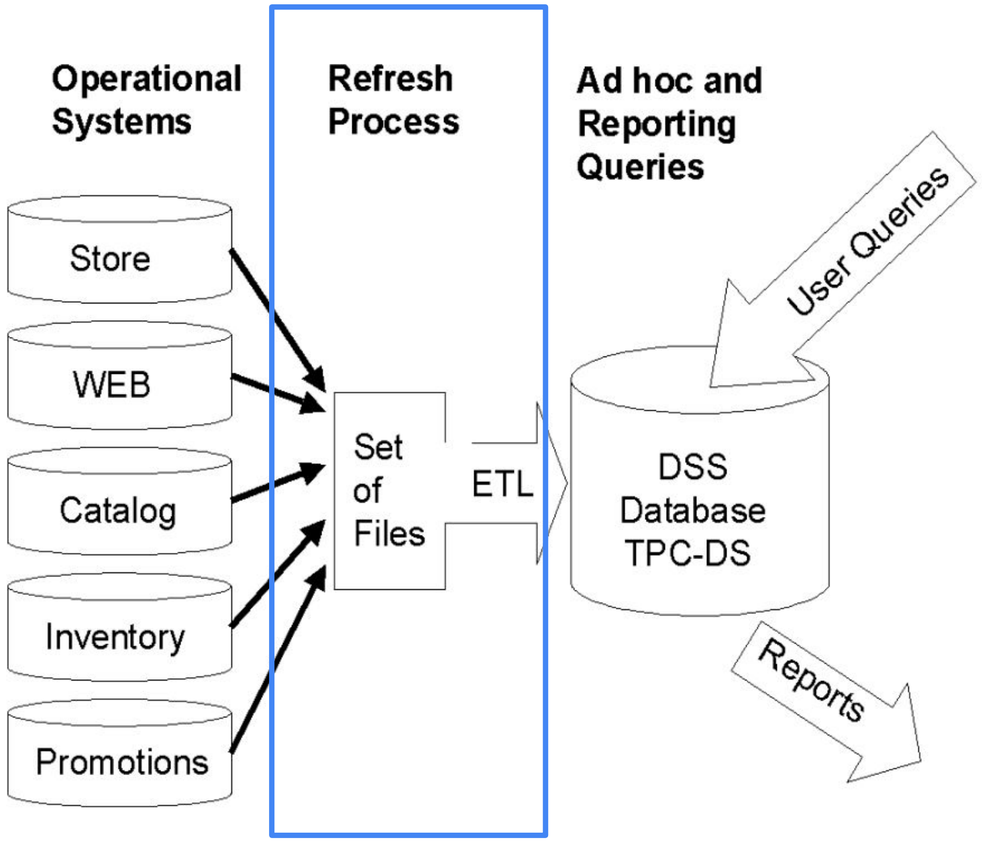
Transactional systems are 80% writes and 20% reads*



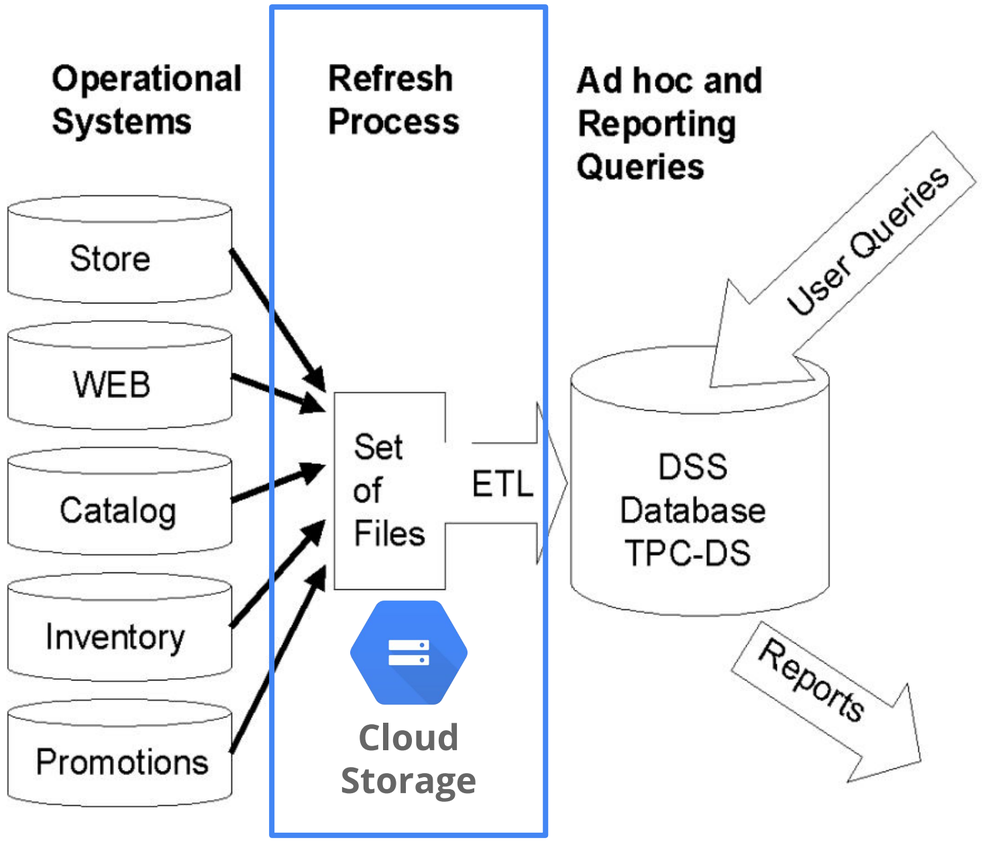
Analytical systems are 20% writes and 80% reads*



Data engineers build the pipelines between the systems



Use Cloud Storage for scalable staging of raw data




```
gsutil -m cp ...
```

Query data directly from GCS in BigQuery

easy

Create BigQuery tables powered by Parquet/ORC

fast

Columnar file format & logical partitions

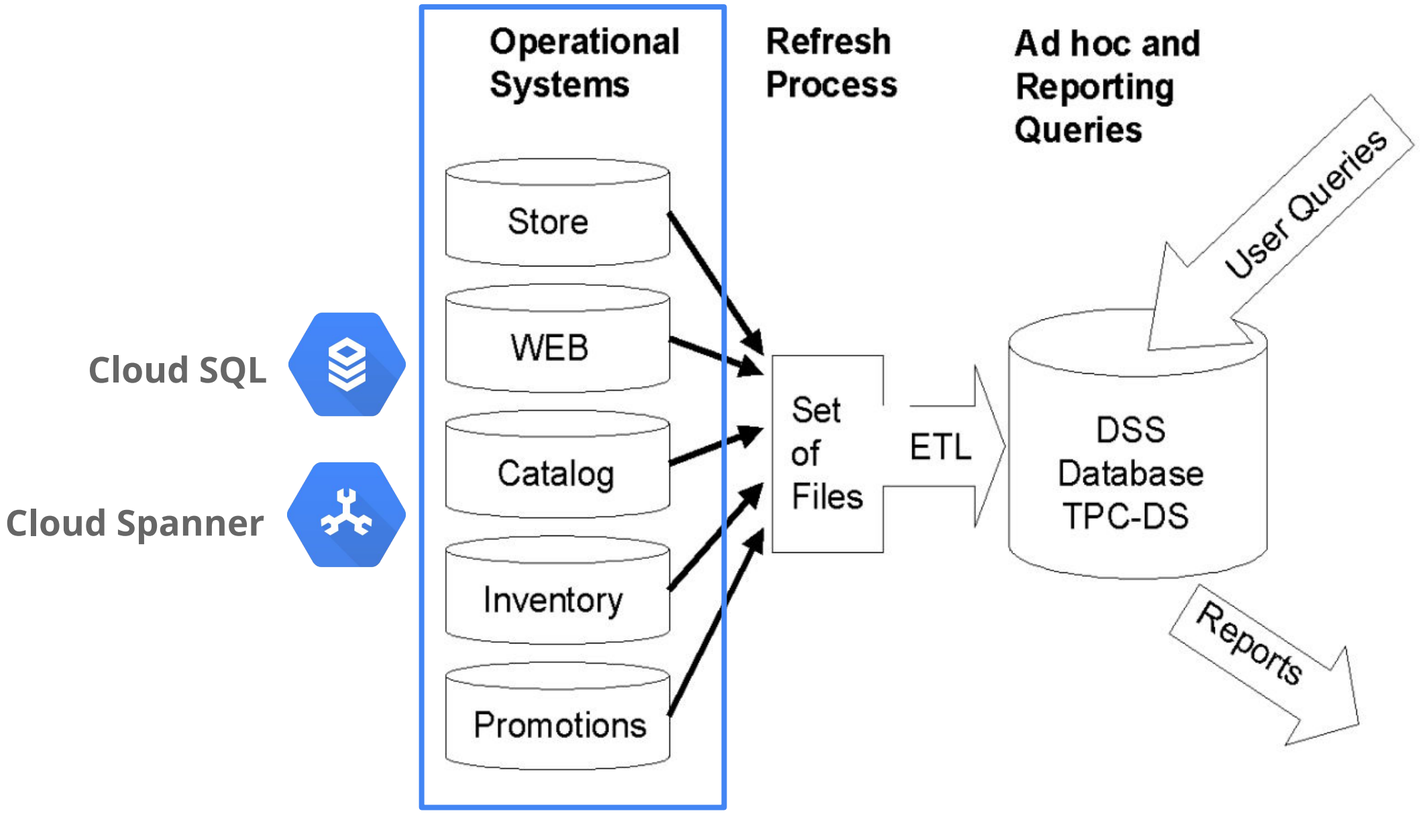
convenient

Hive partitions support for query & load

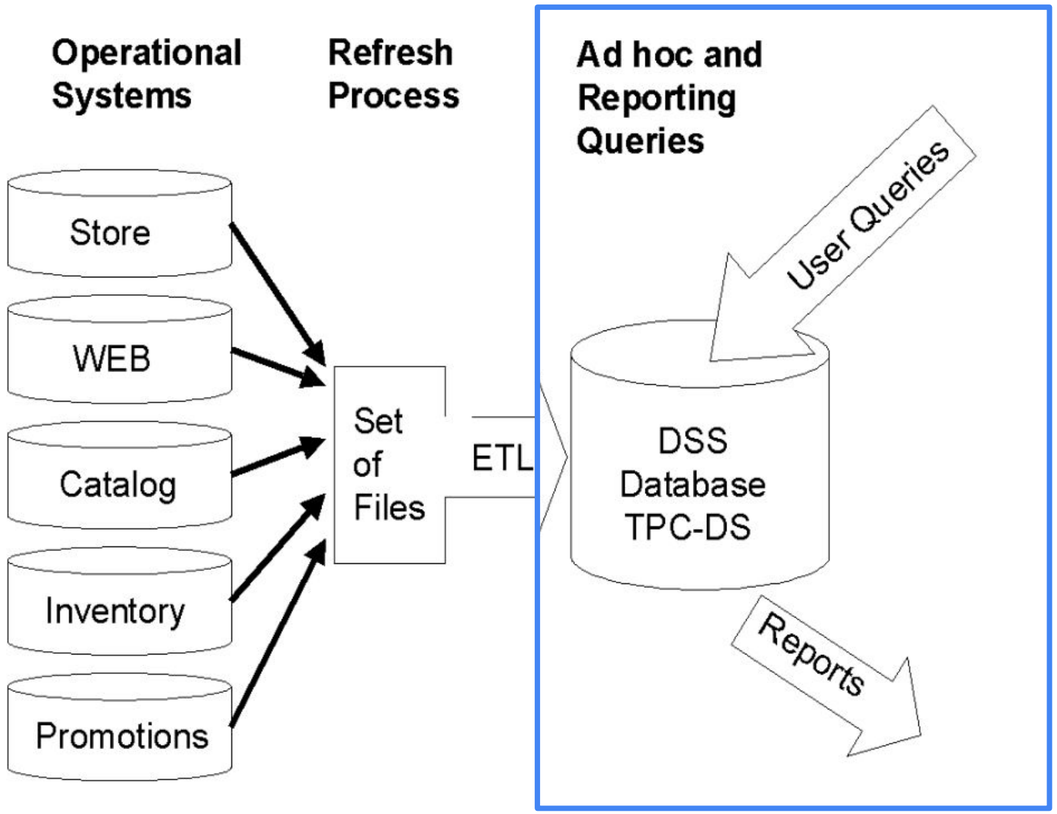
Demo

Running federated queries on Parquet and ORC files in BigQuery

Choose from cloud relational databases for transactional workloads



Choose from cloud data warehouses for analytic workloads



Cloud Bigtable

BigQuery

Google Cloud

Agenda

Introduction to Data Lakes

Data Storage and ETL options on GCP

Building a Data Lake using Cloud Storage

Securing Cloud Storage

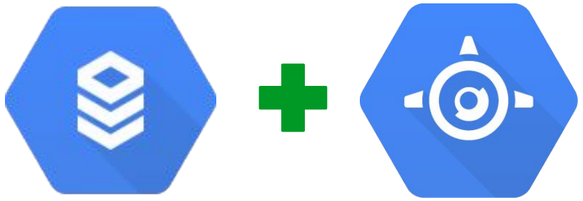
Storing All Sorts of Data Types

Cloud SQL as a relational Data Lake

Lab: Loading Taxi Data into Cloud SQL

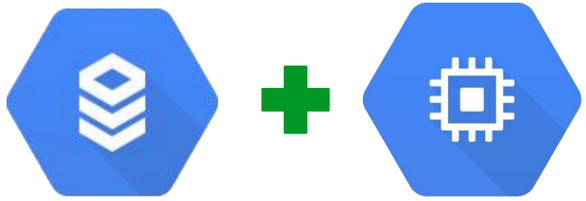
Cloud SQL is a fully managed database service that makes it easy to set up and administer your relational MySQL and PostgreSQL databases in the cloud

Cloud SQL can be used with other GCP services



Cloud SQL can be used with App Engine using standard drivers.

You can configure a Cloud SQL instance to follow an App Engine application.



Compute Engine instances can be authorized to access Cloud SQL instances using an external IP address.

Cloud SQL instances can be configured with a preferred zone.

Cloud SQL can be used with external applications and clients.

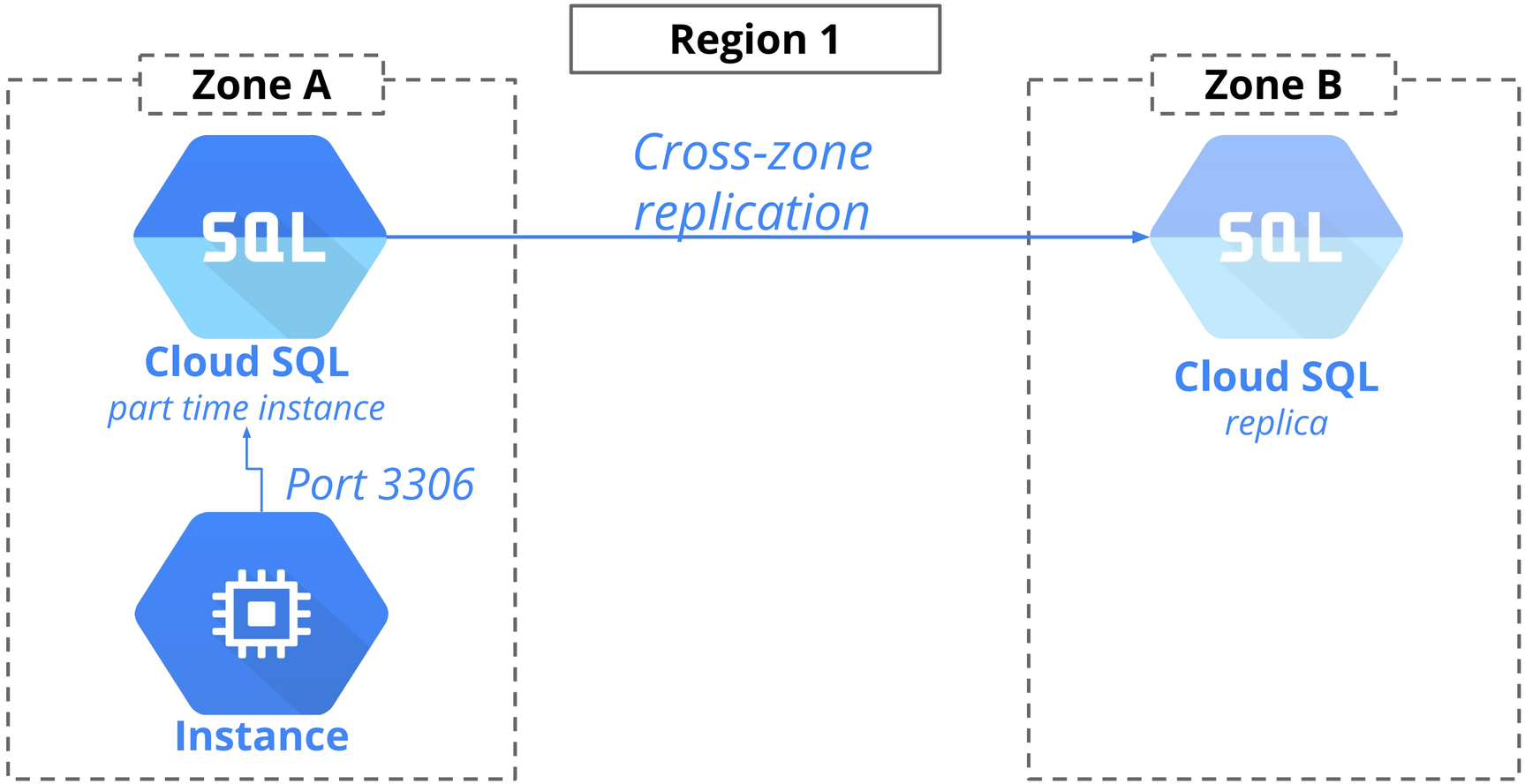
Standard tools can be used to administer databases.

External read replicas can be configured.

Backup, recovery, scaling, and security is managed for you

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

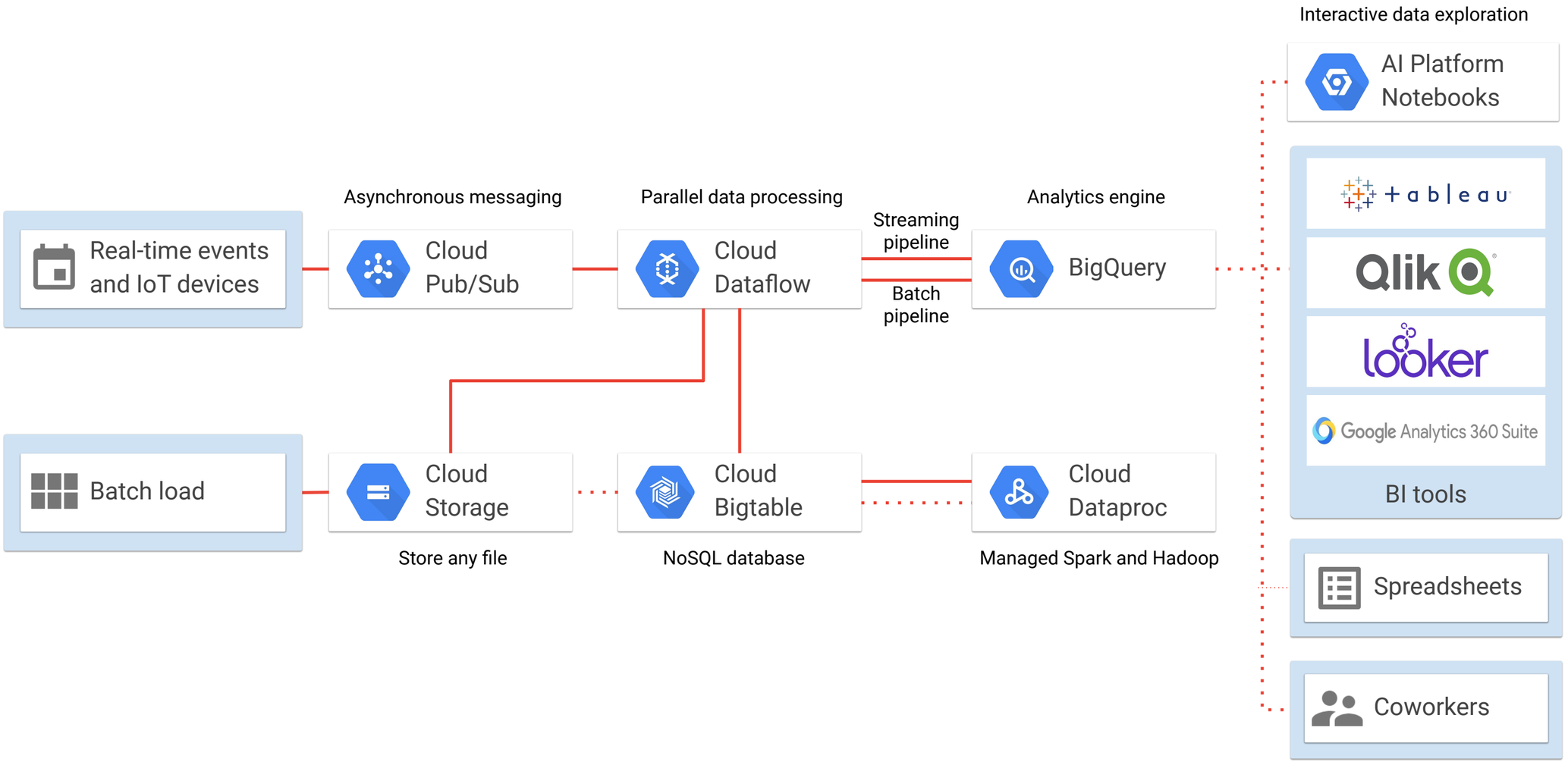
Cloud SQL replication



Fully managed versus serverless

Fully managed	 No setup	 Automated backups, updates, etc.	 Replicated, highly available
Serverless	 No server management	 Fully managed security	 Pay only for usage

Modern serverless data management architecture



Google Cloud

Agenda

Introduction to Data Lakes

Data Storage and ETL options on GCP

Building a Data Lake using Cloud Storage

Securing Cloud Storage

Storing All Sorts of Data Types

Cloud SQL as a relational Data Lake

Lab: Loading Taxi Data into Cloud SQL

Lab

Loading Taxi Data into Cloud SQL

Objectives

- Create Cloud SQL instance
- Create a Cloud SQL database
- Import text data into Cloud SQL
- Check the data for integrity