Google Cloud

Building a Data Lake

강사 : 고병화

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

Introduction to Data Lakes

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.

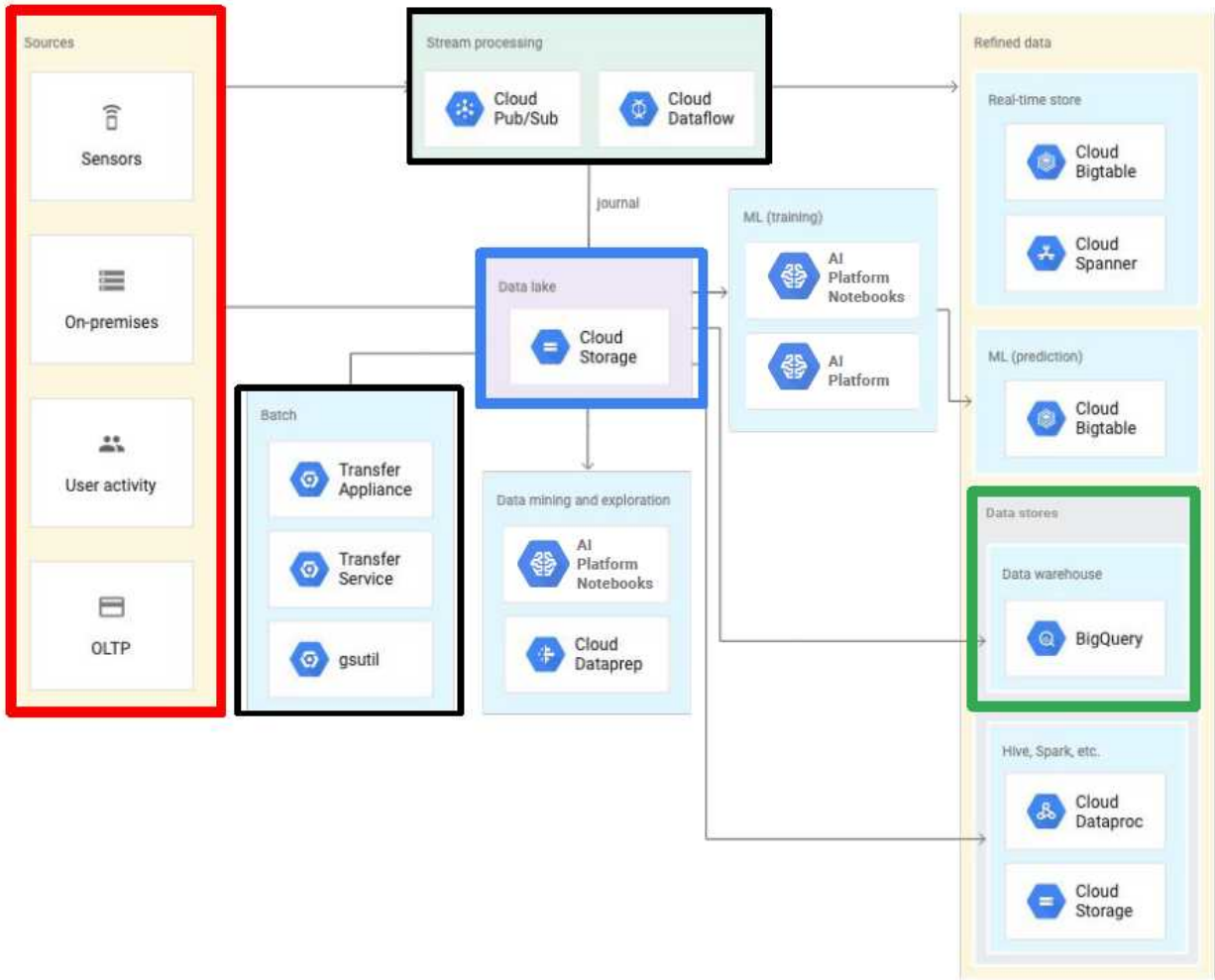
- Structured | Semi-structured | Unstructured
- Batch | Streaming
- SQL | ML/AI | Search
- On-Prem | Cloud | Edge

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

Example Architecture

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

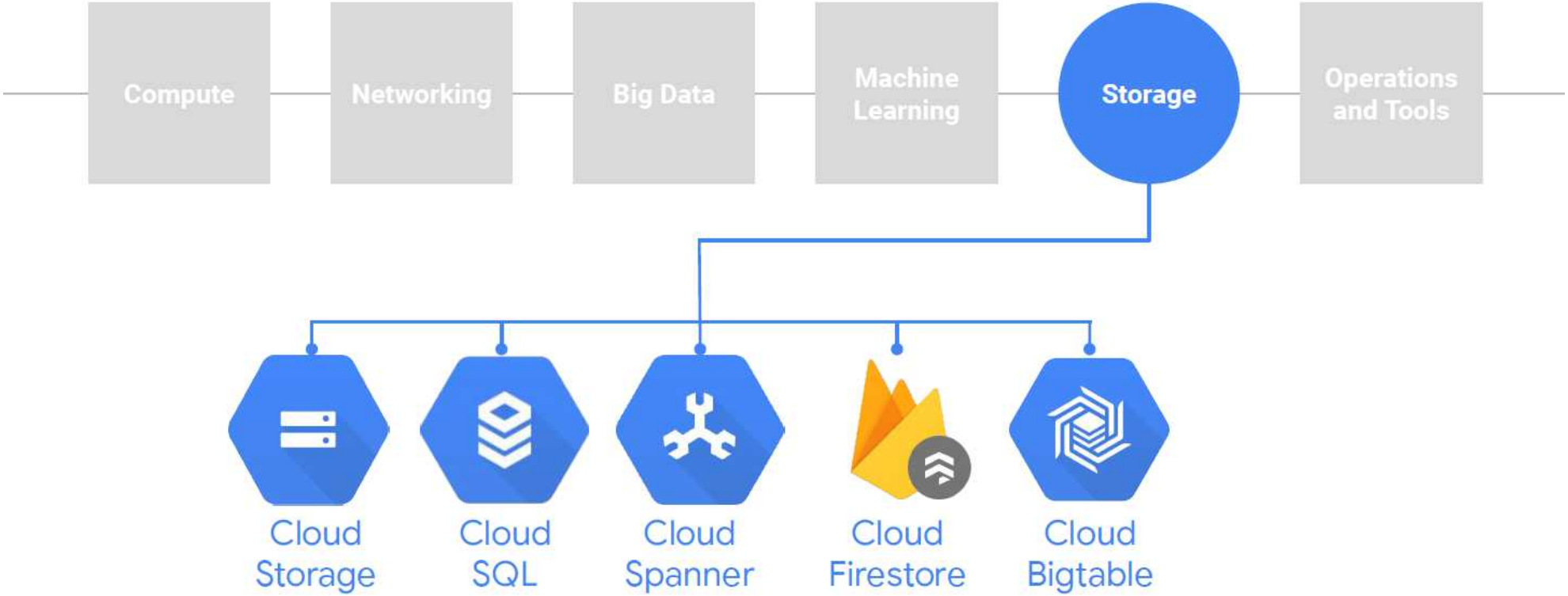


The suite of big data products on Google Cloud

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

Data Storage and ETL options on GCP

Storage options for your data on GCP



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.

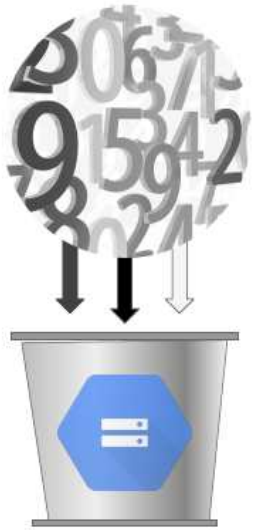
The method you use to load data depends on how much transformation is needed

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

Building a Data Lake using Cloud Storage

Cloud Storage

Cloud Storage

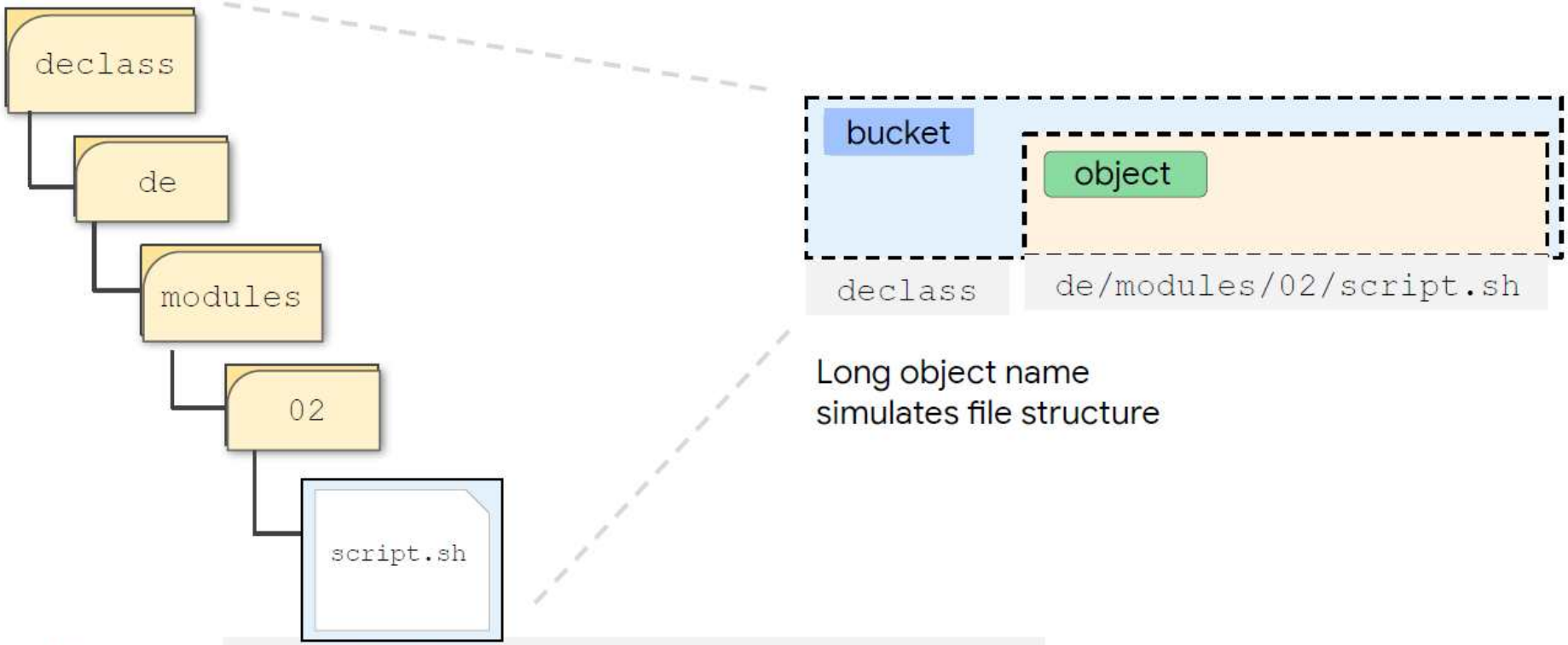


Qualities that Cloud Storage contributes to data engineering solutions:	Persistence Durability Strong consistency Availability High throughput

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

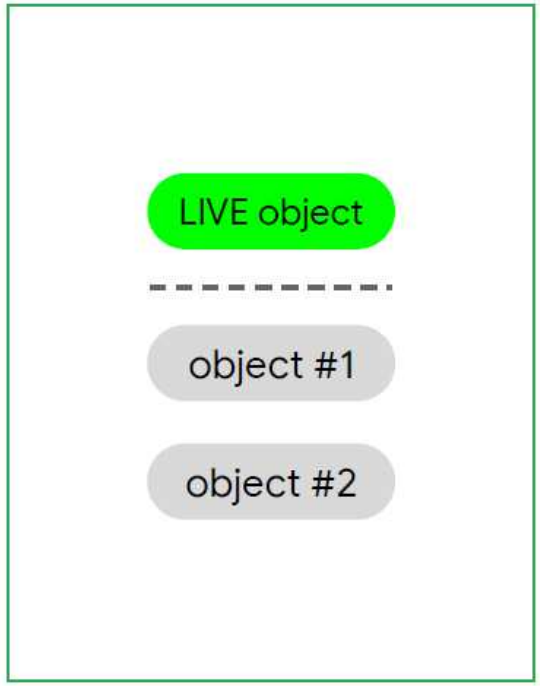


File access `gs://declass/de/modules/02/script.sh`

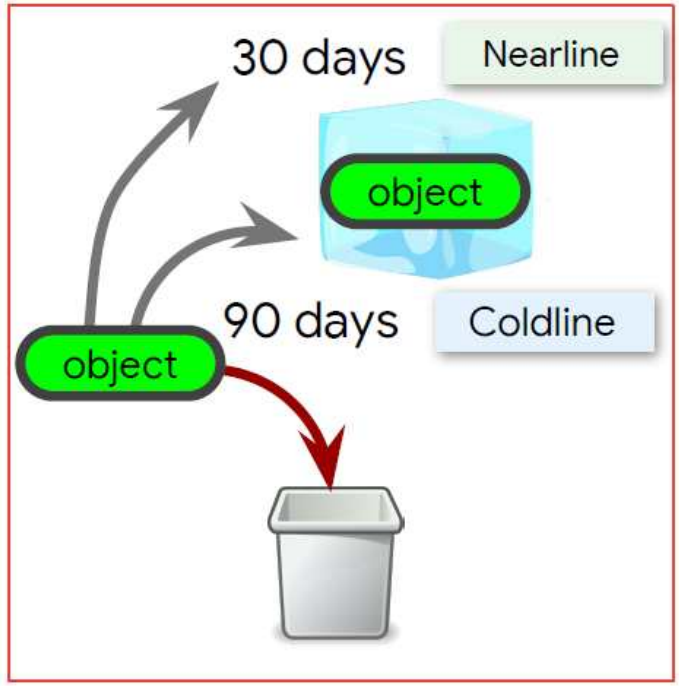
Web access `https://storage.cloud.google.com/declass/de/modules/02/script.sh`

Cloud Storage has many object management features

Retention Policy



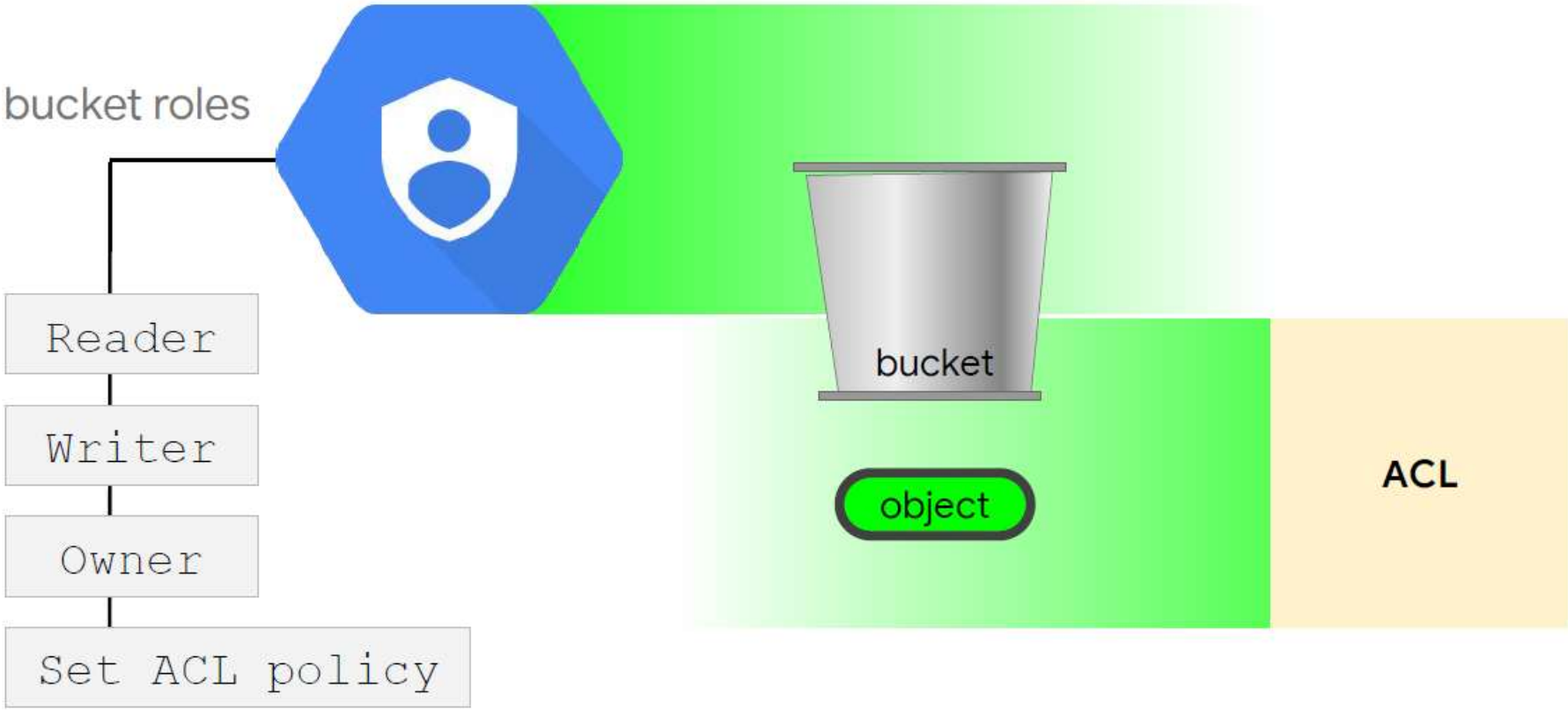
Versioning



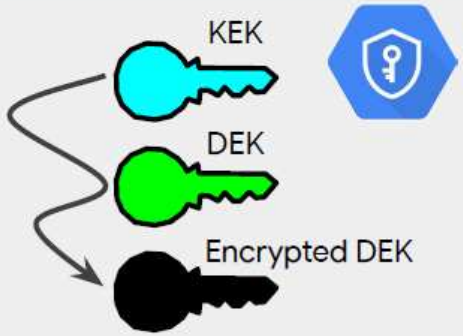
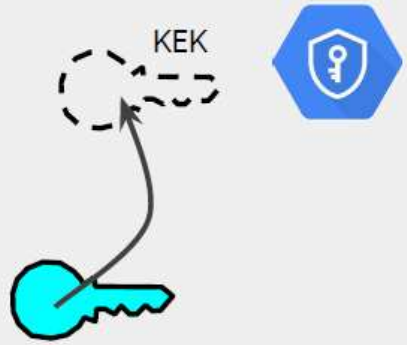
Lifecycle Management

Securing Cloud Storage

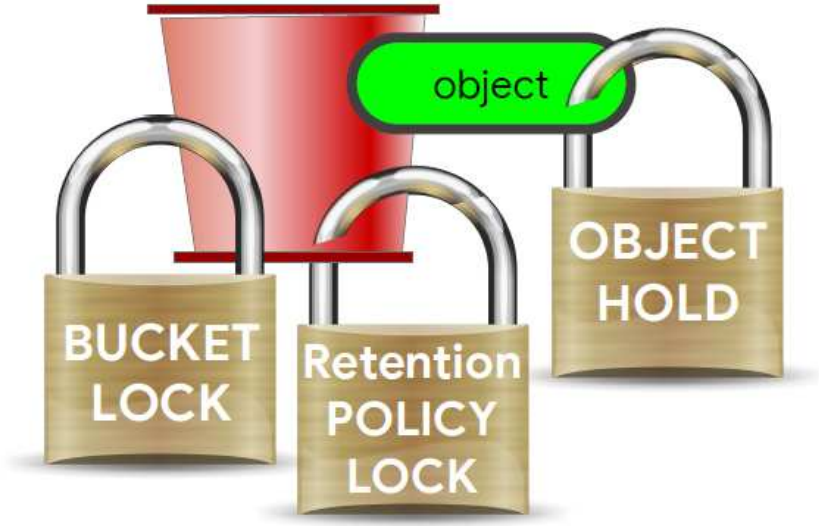
Controlling access with Cloud IAM and access lists



Data encryption options for many requirements

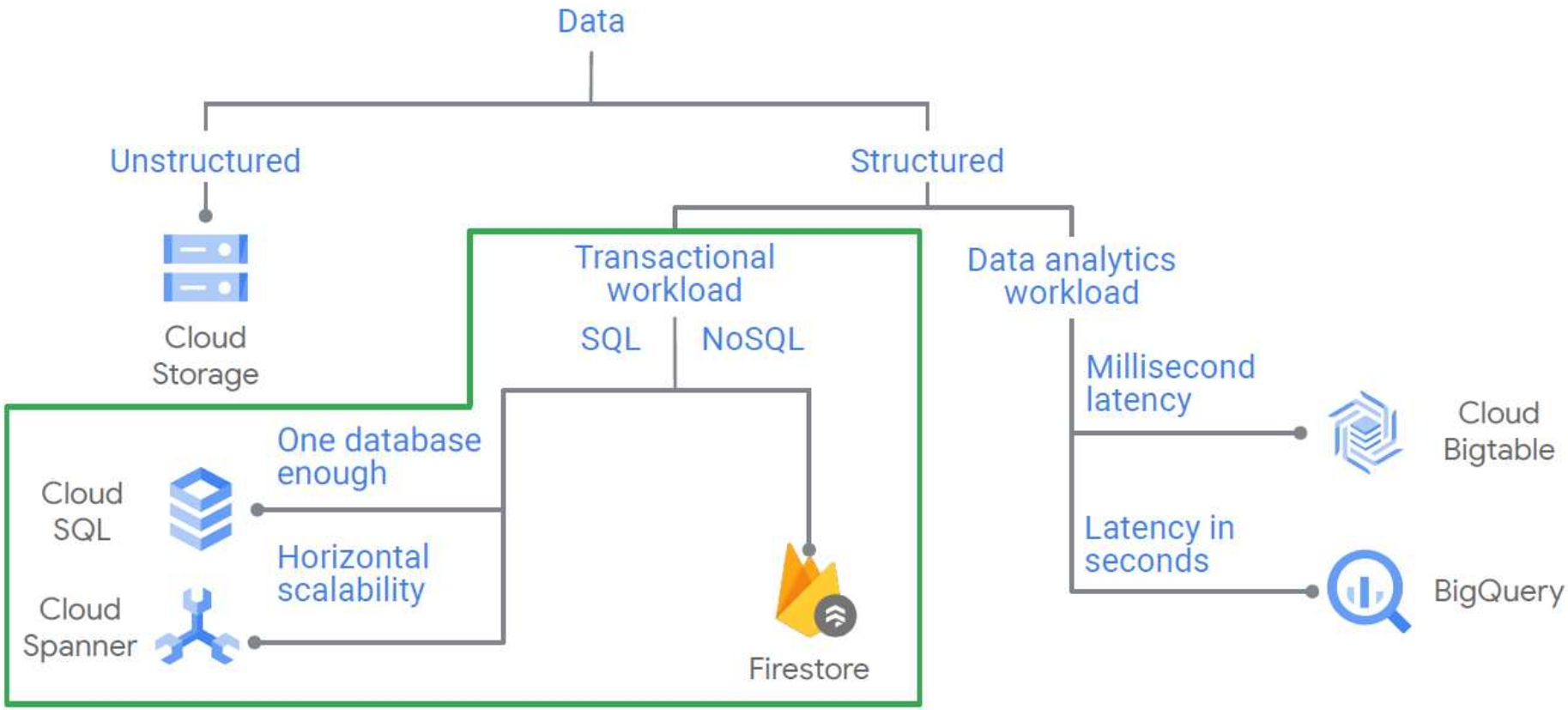
GMEK	CMEK	CSEK
Google-managed encryption keys	Customer-managed encryption keys	Customer-supplied encryption keys
		
Cloud KMS Key Management Service Encryption is automatic.	You control the creation and existence of the KEK key in Cloud KMS.	You provide the KEK key.

Cloud Storage supports many special use cases

Client-side encryption	Data locking for audits	✚
 Client-side encryption		Decompressive coding Requester pays Signed URLs for anonymous sharing Period expirations Composite objects ...

Storing All Sorts of Data Types

Different considerations for transactional workloads



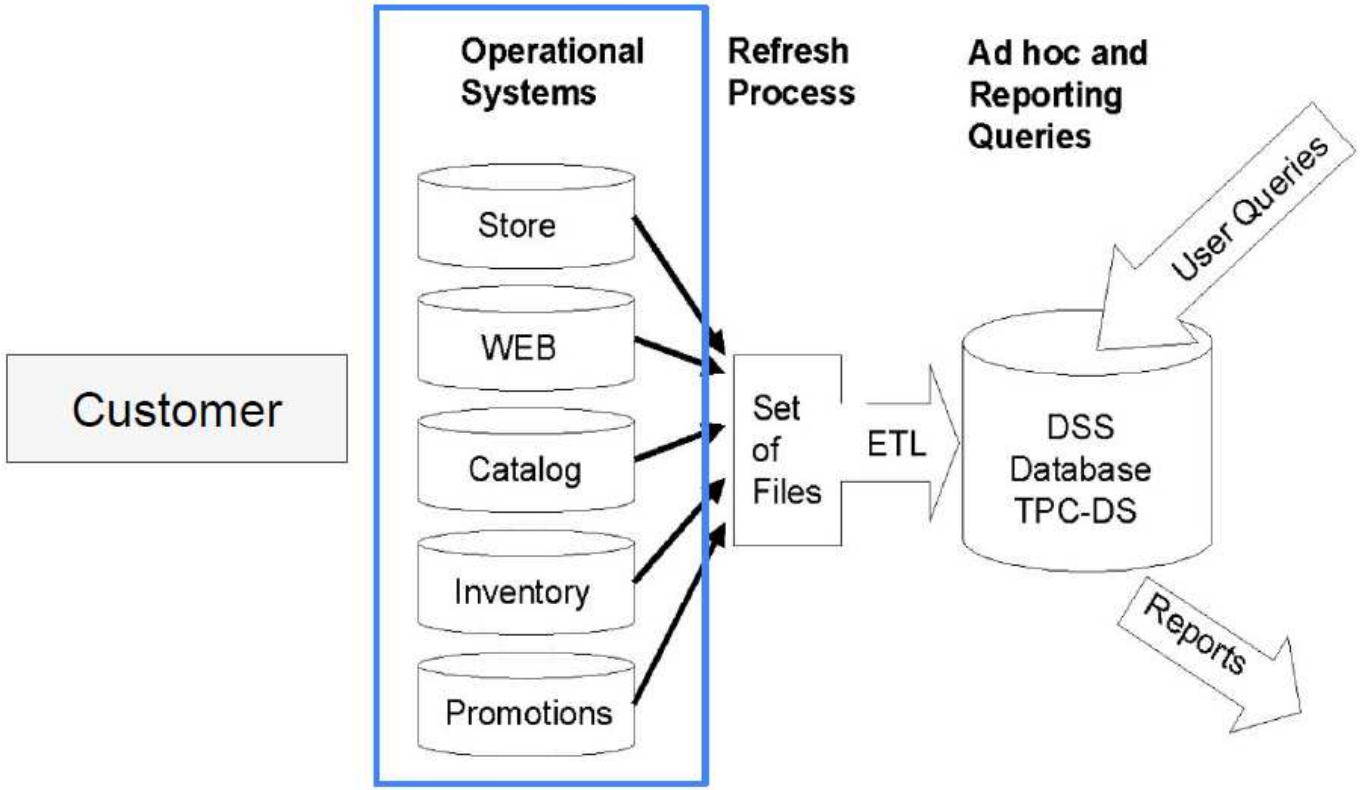
Transactional versus analytical

	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

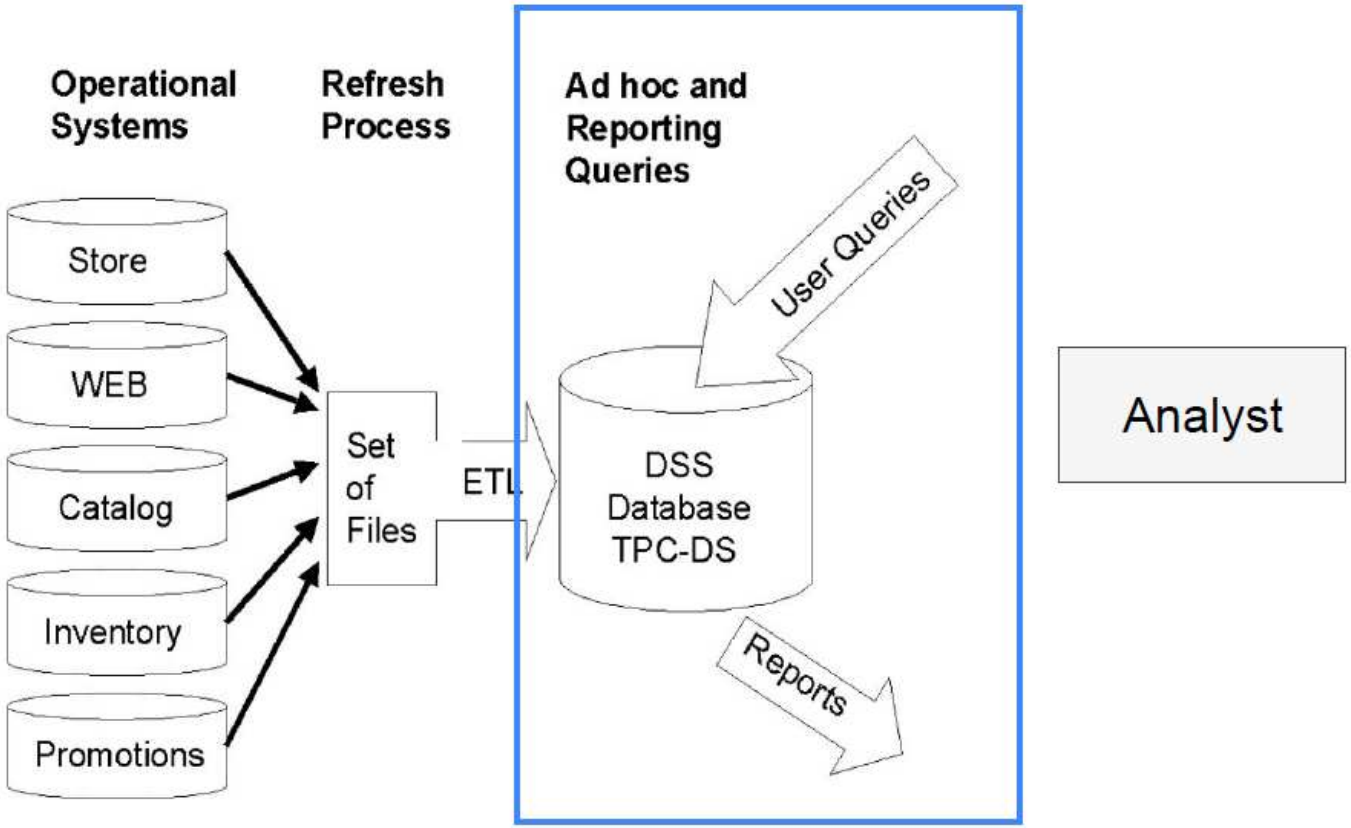
Transactional versus analytical

	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

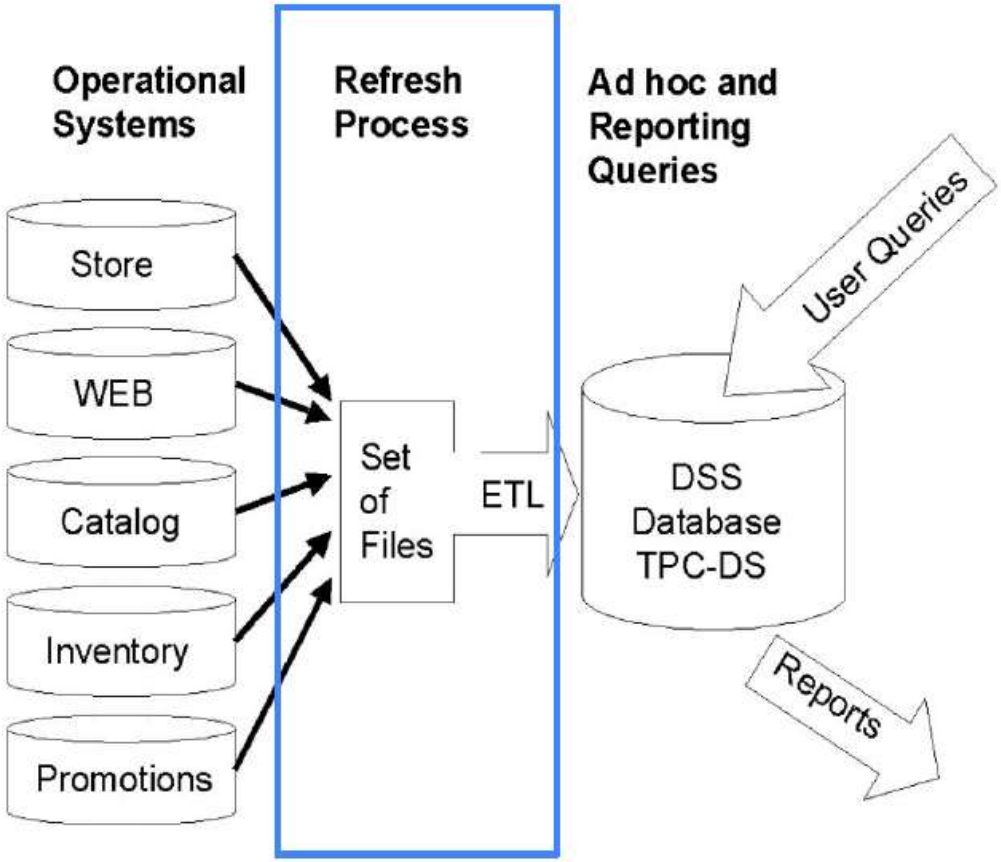
Transactional systems are 80% writes and 20% reads*



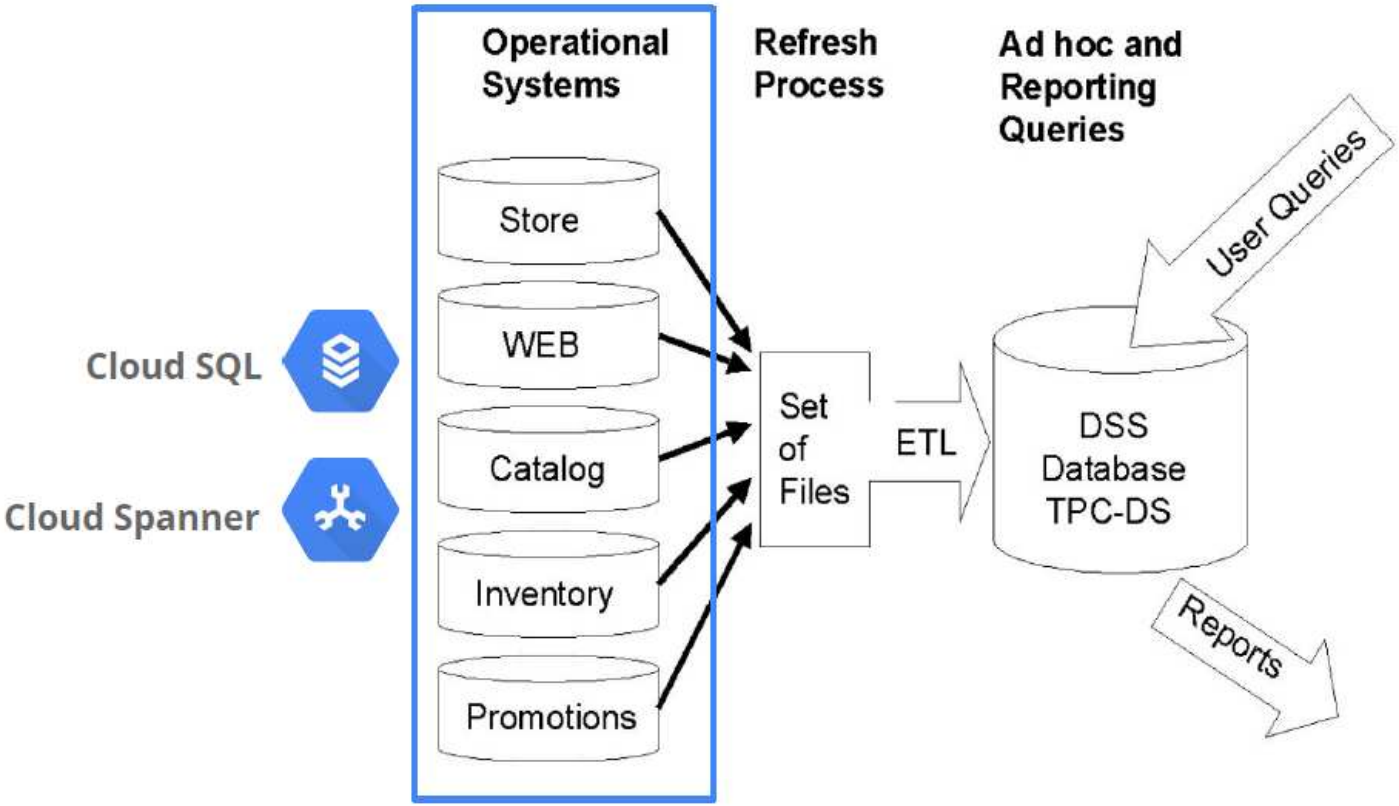
Analytical systems are 20% writes and 80% reads*



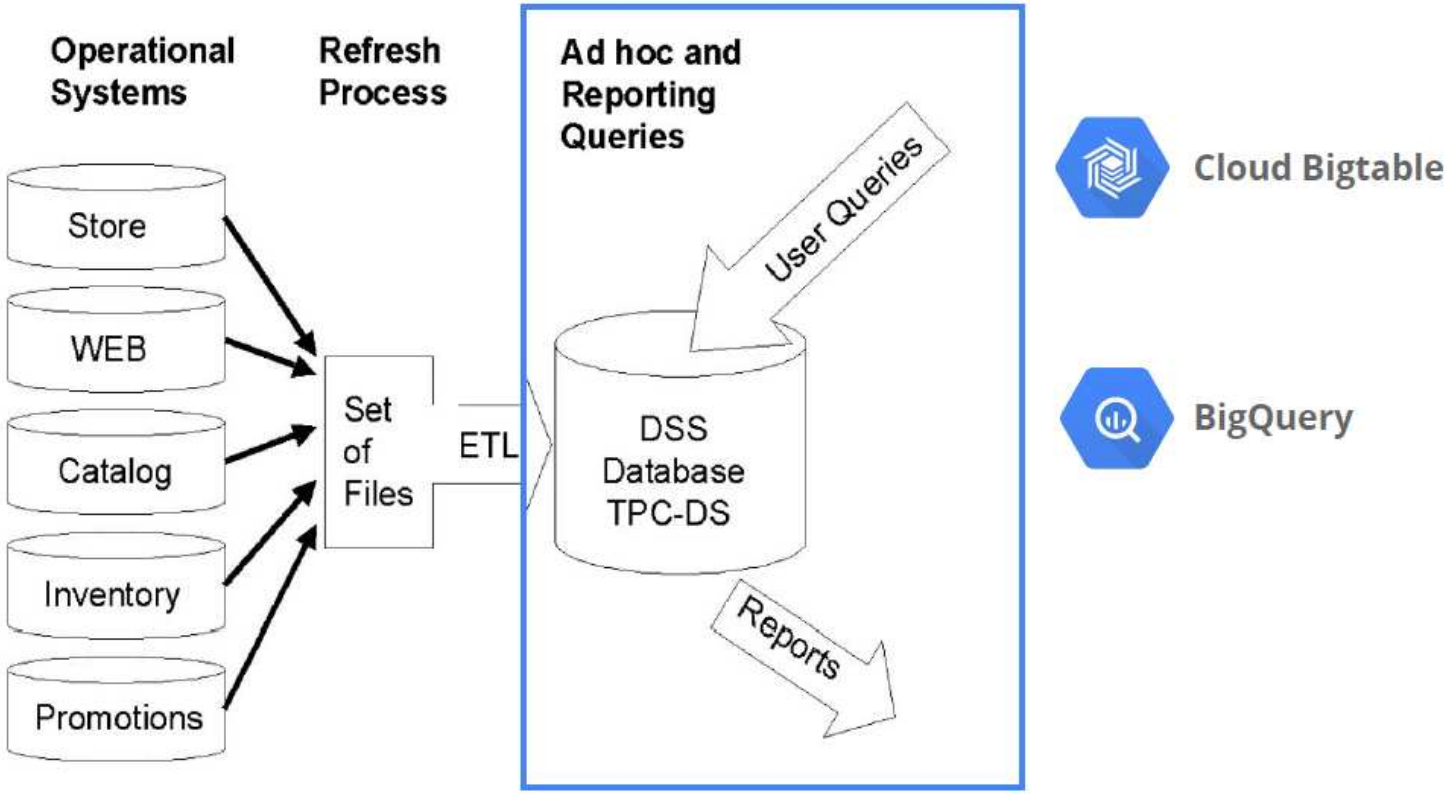
Data engineers build the pipelines between the systems



Choose from cloud relational databases for transactional workloads



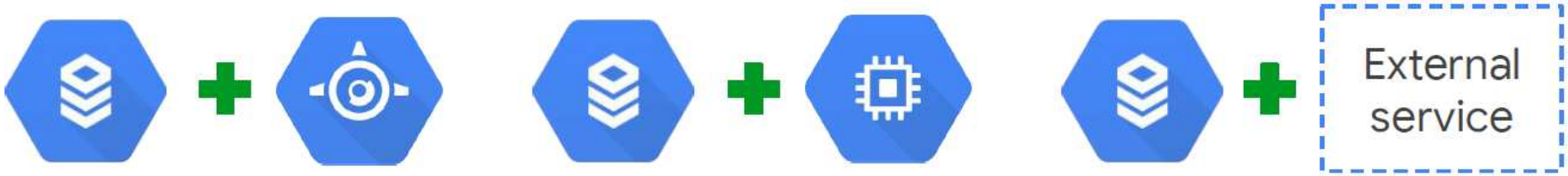
Choose from cloud data warehouses for analytic workloads



Cloud SQL as a relational Data Lake

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.

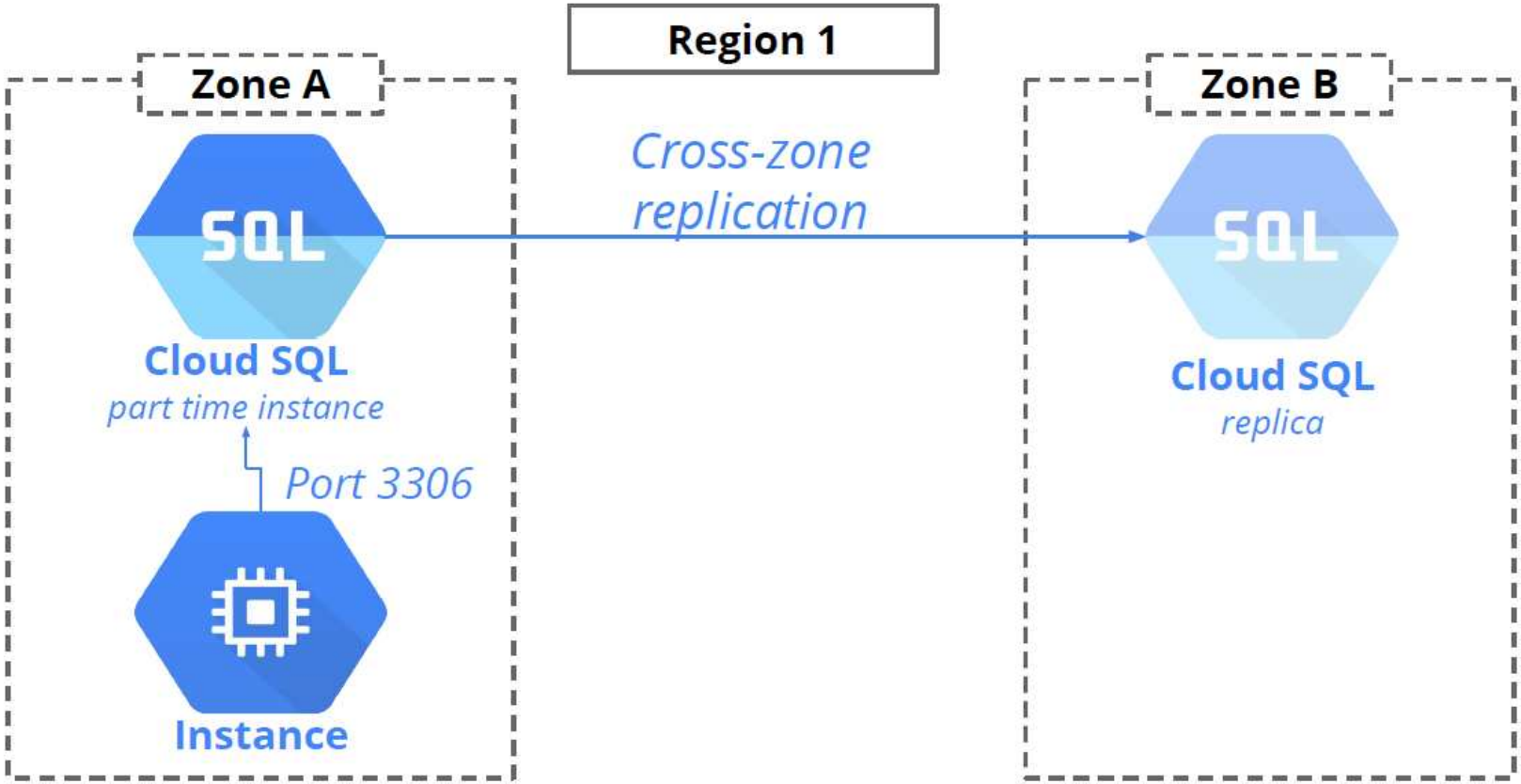
Standard tools can be used to administer databases.

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

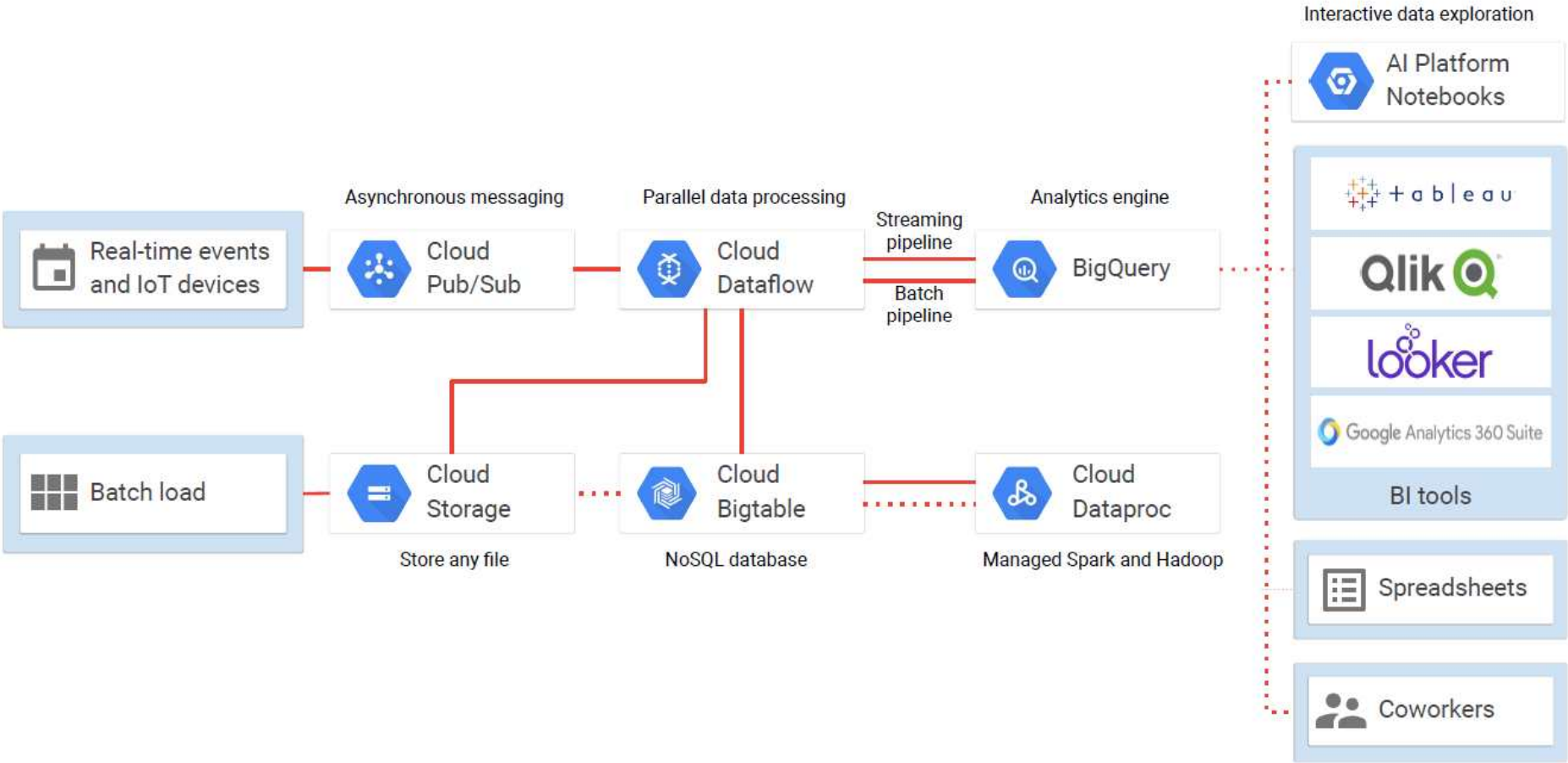
Cloud SQL replication



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



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