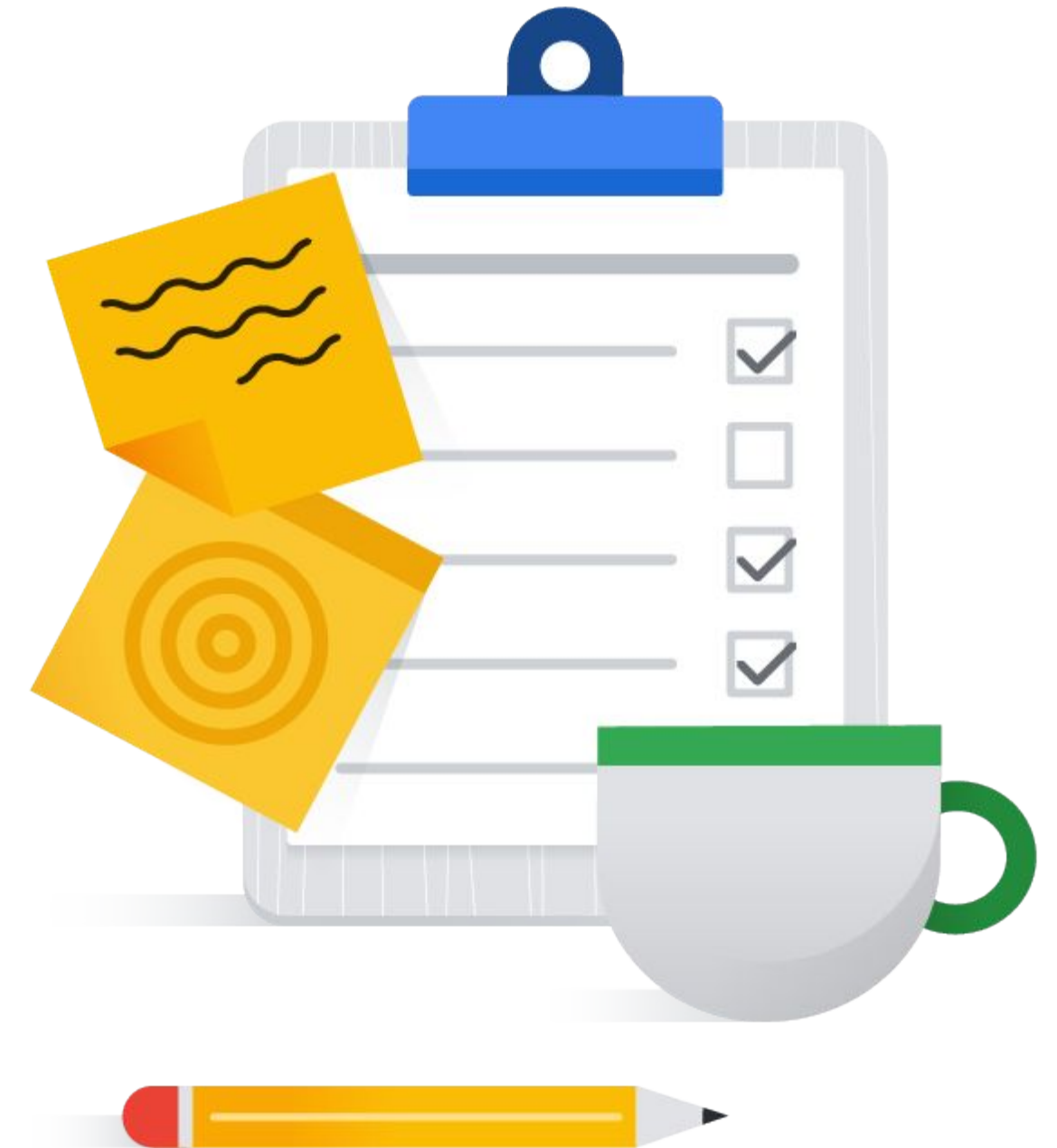




# Building a Data Lake

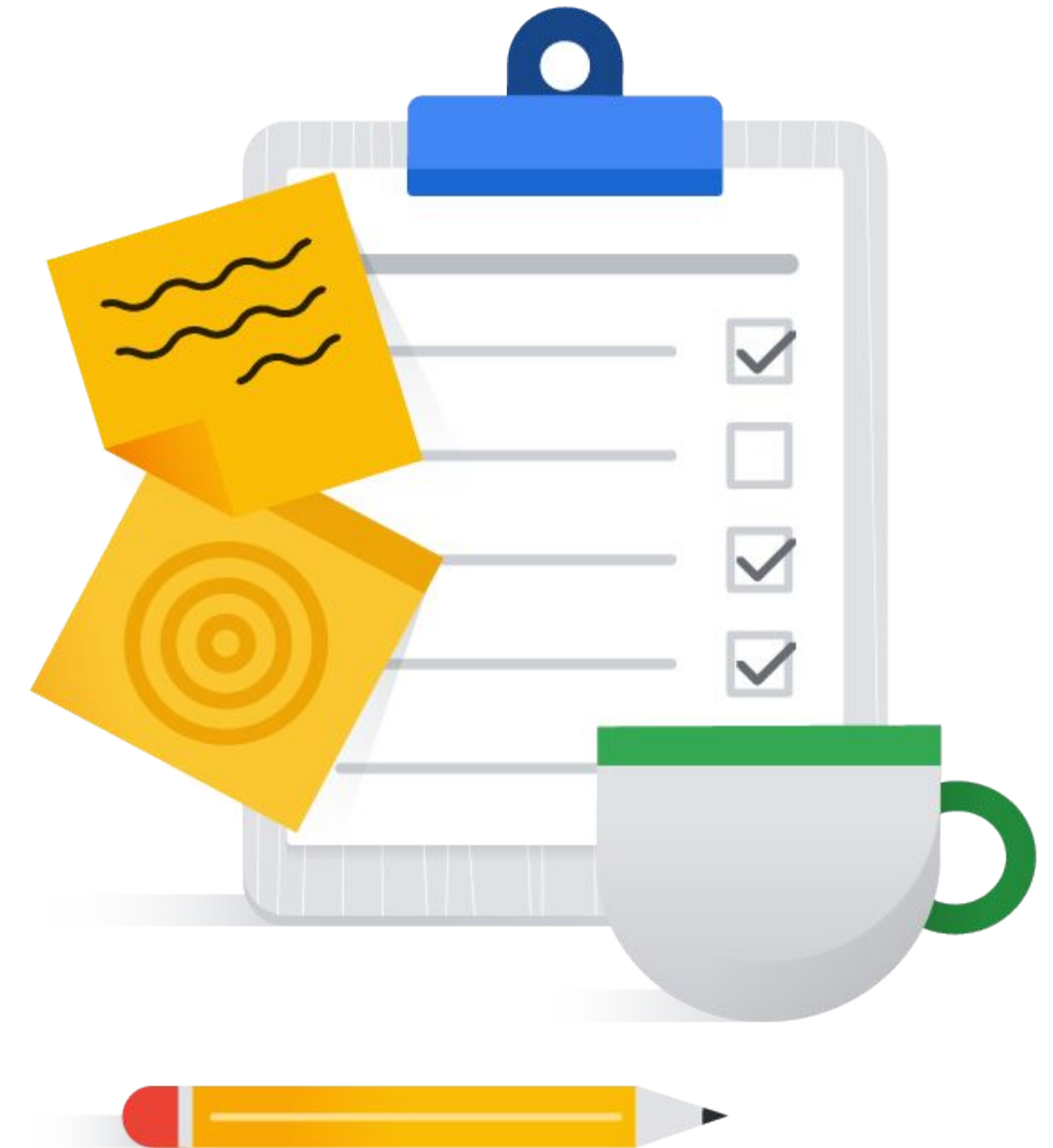
# Building a Data Lake

01	Introduction to data lakes
02	Data storage and ETL options on Google Cloud
03	Build a data lake using Cloud Storage
04	Secure Cloud Storage
05	Store all sorts of data types
06	Cloud SQL as a relational data lake



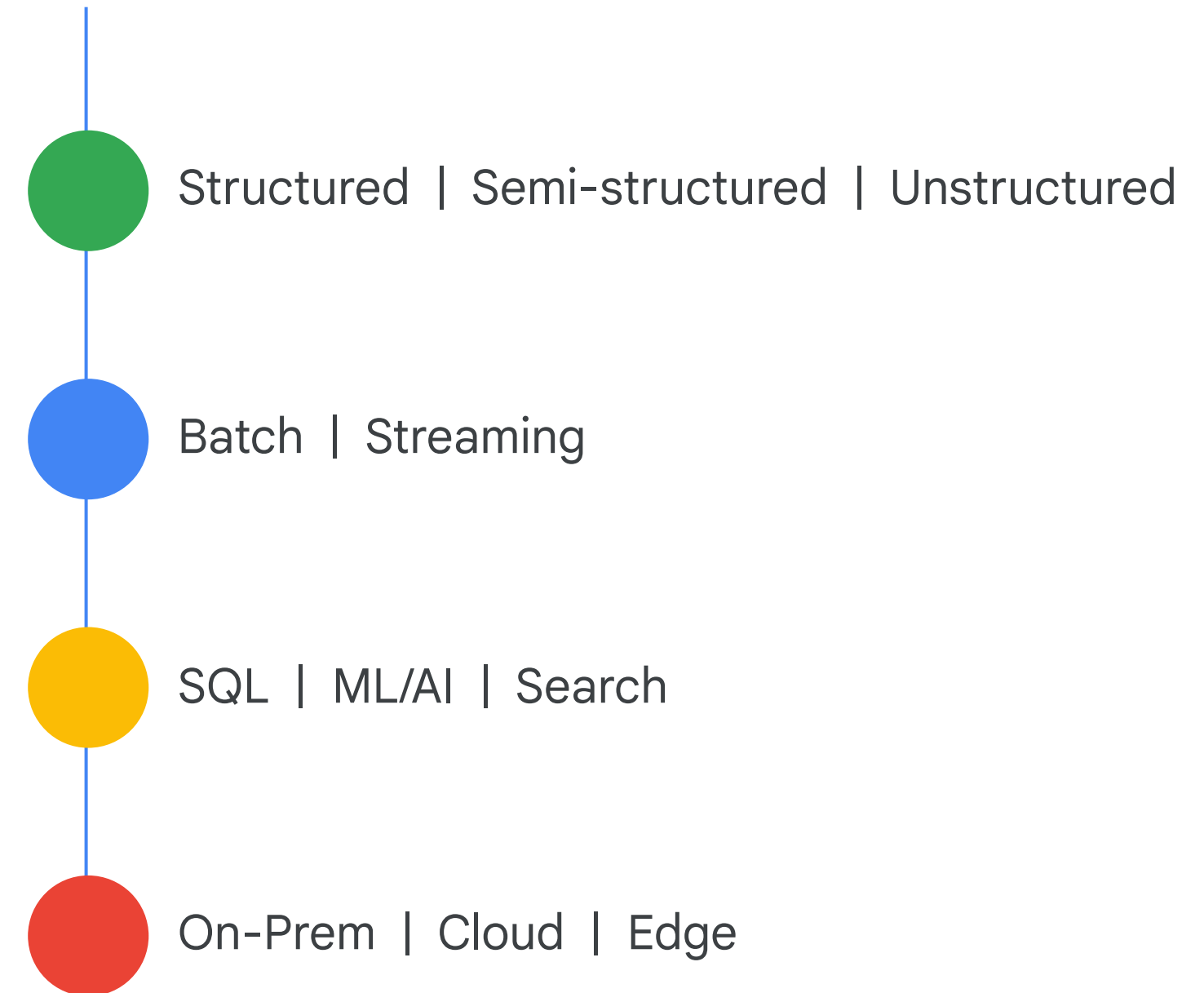
# Building a Data Lake

01	Introduction to data lakes
02	Data storage and ETL options on Google Cloud
03	Build a data lake using Cloud Storage
04	Secure Cloud Storage
05	Store all sorts of data types
06	Cloud SQL as a relational data lake



# 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.



# 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

Our focus in this module

Our focus in the next module

Our focus in other courses



# Data Engineering is like Civil Engineering

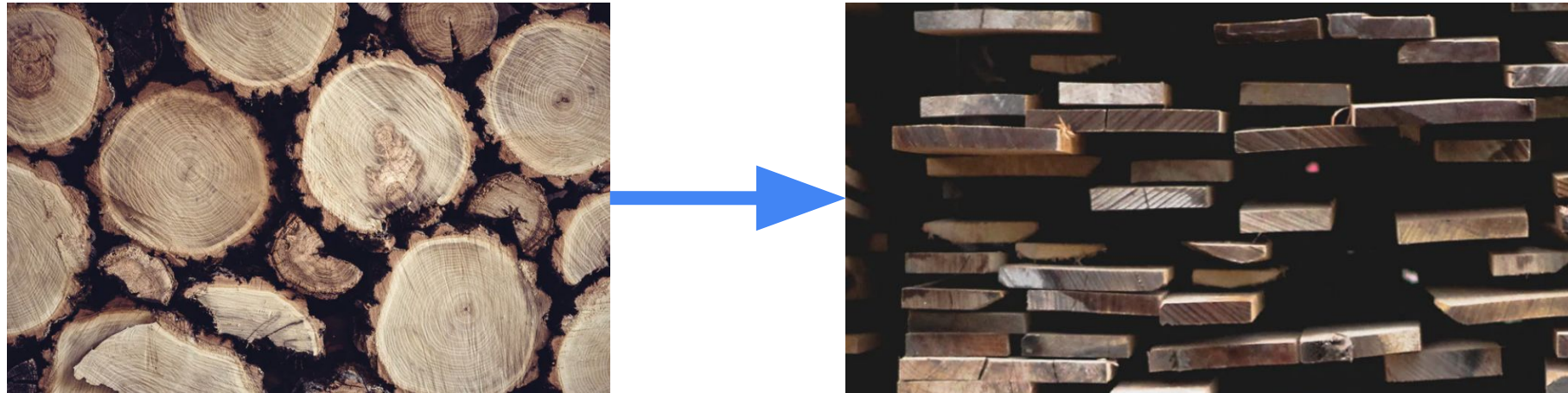


1

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.



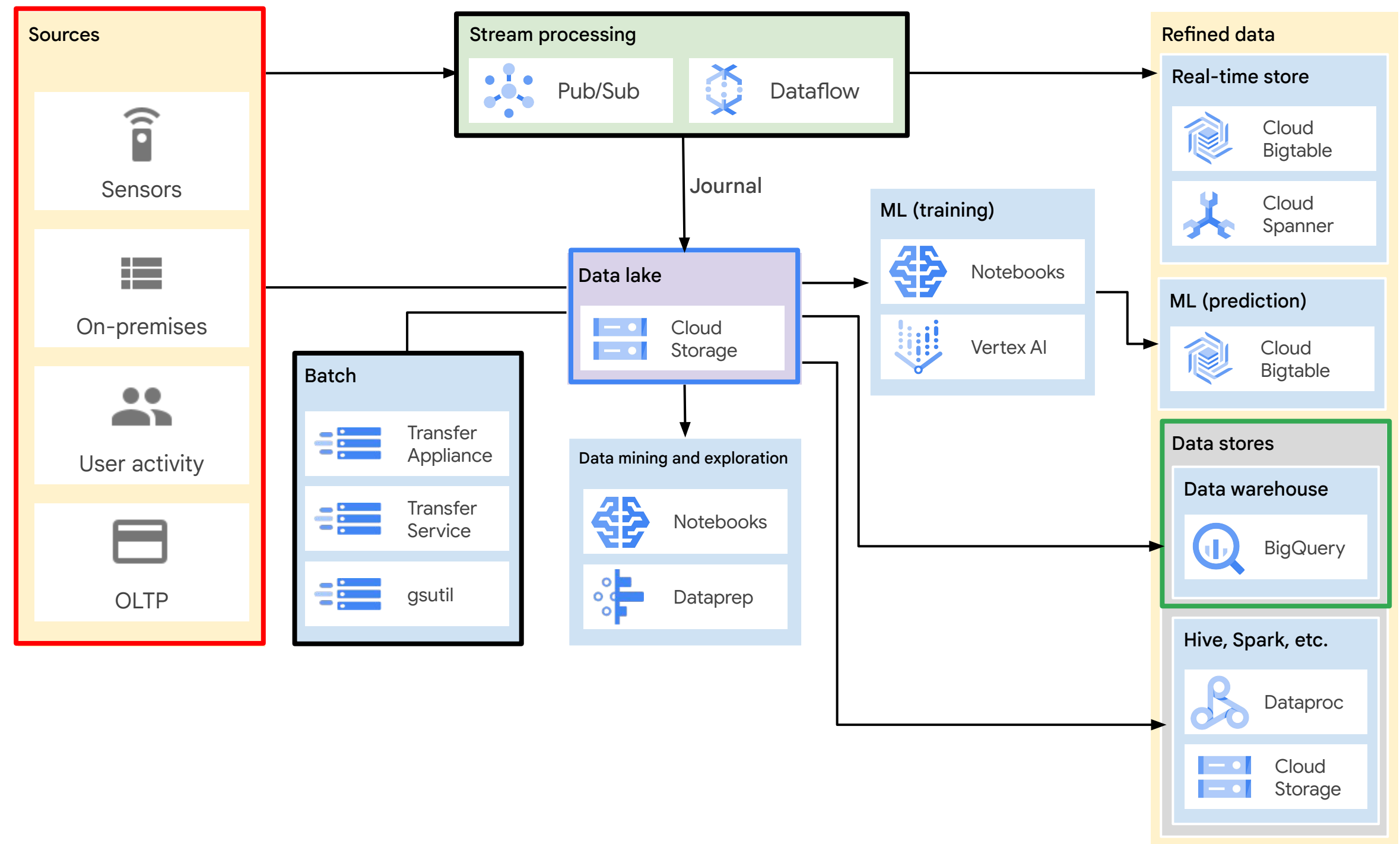
# An orchestrator governs all aspects of the workflow



- 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).

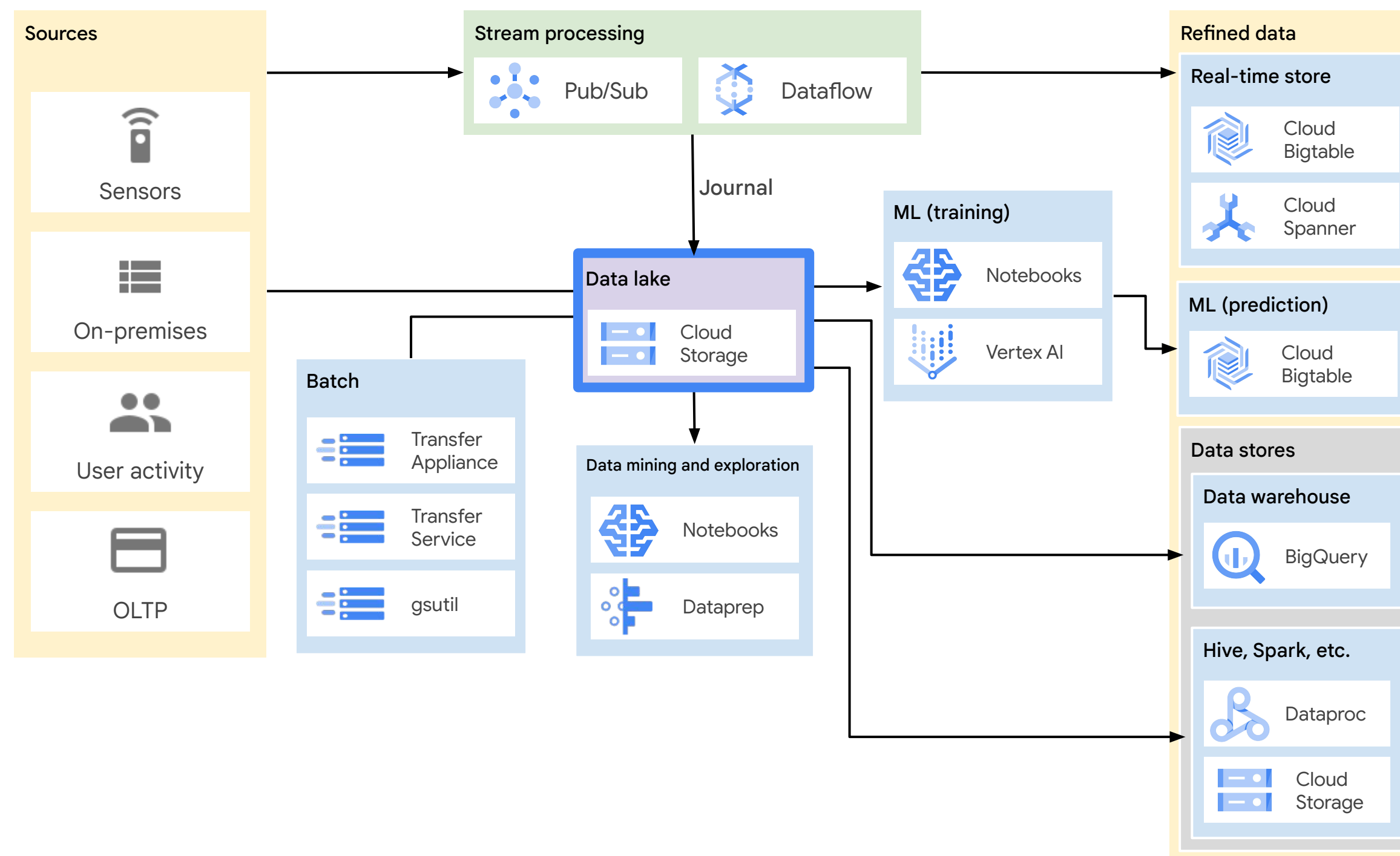
# Example architecture

1. **Data sources**
2. **Data lake**
3. **Data pipelines**
4. **Data warehouse**
5. Used for ML and analytics workloads



# Example architecture

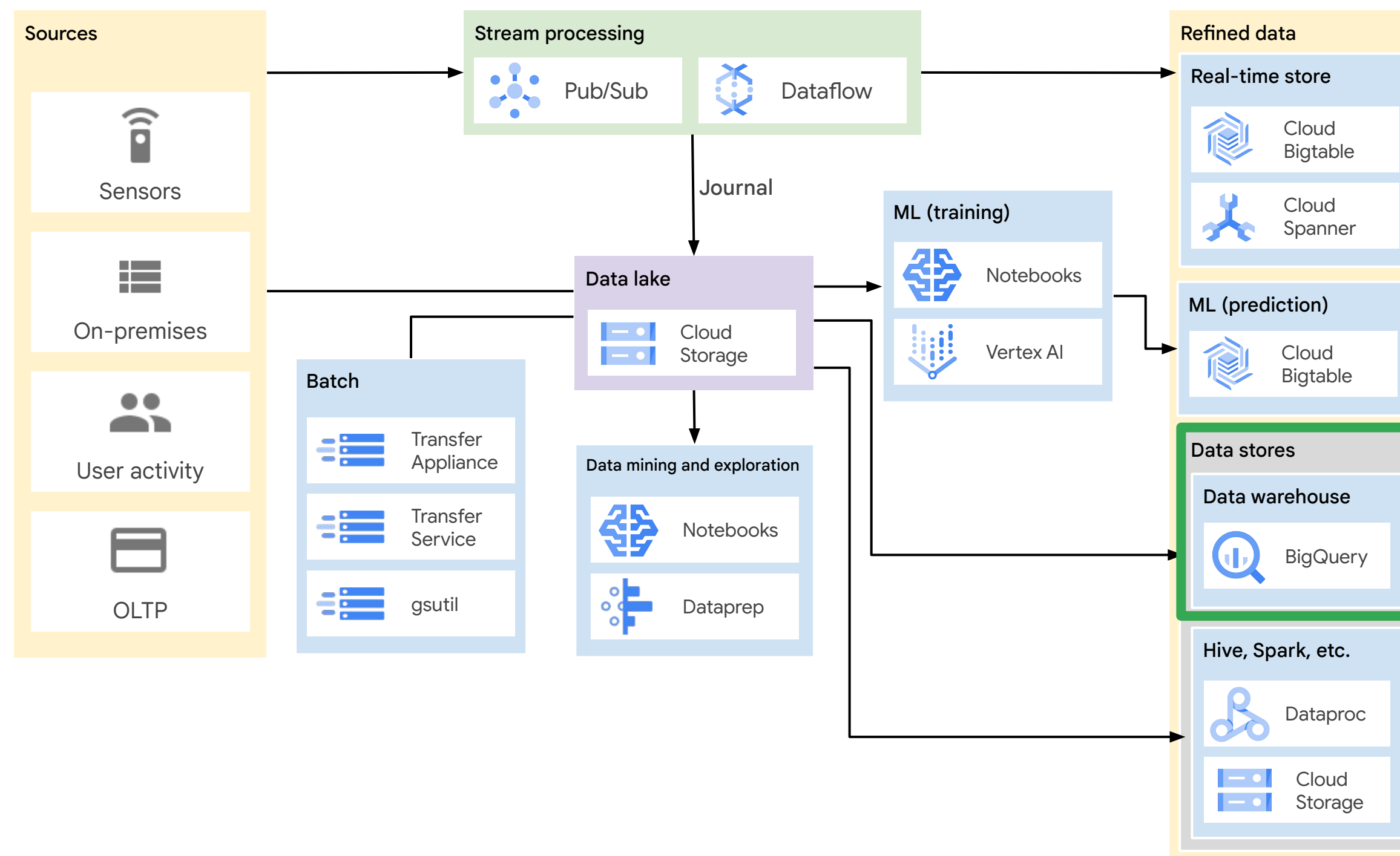
1. **Data sources**
2. **Data lake**
3. **Data pipelines**
4. **Data warehouse**
5. Used for ML and analytics workloads





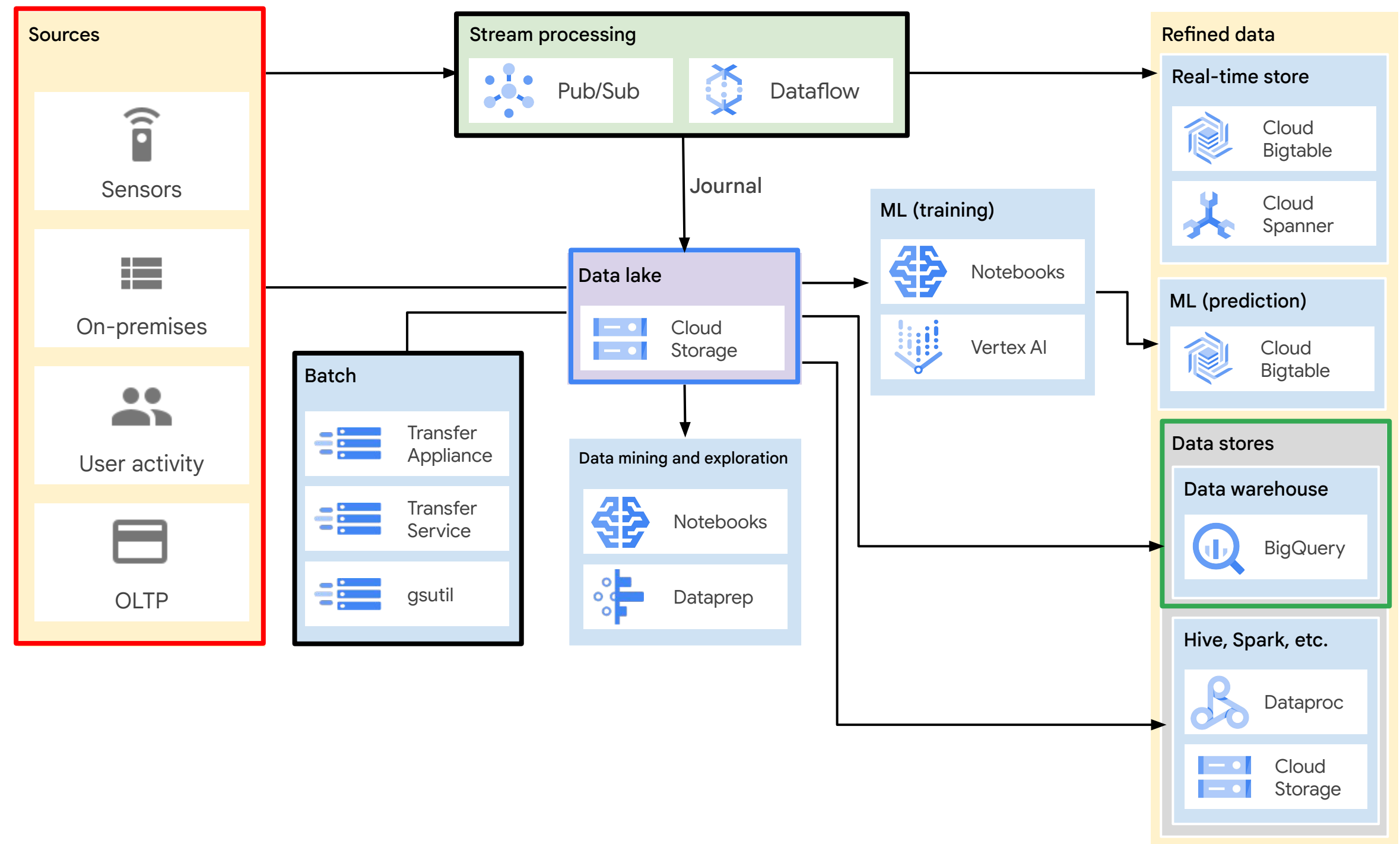
# Example architecture

1. **Data sources**
2. **Data lake**
3. **Data pipelines**
4. **Data warehouse**
5. Used for ML and analytics workloads

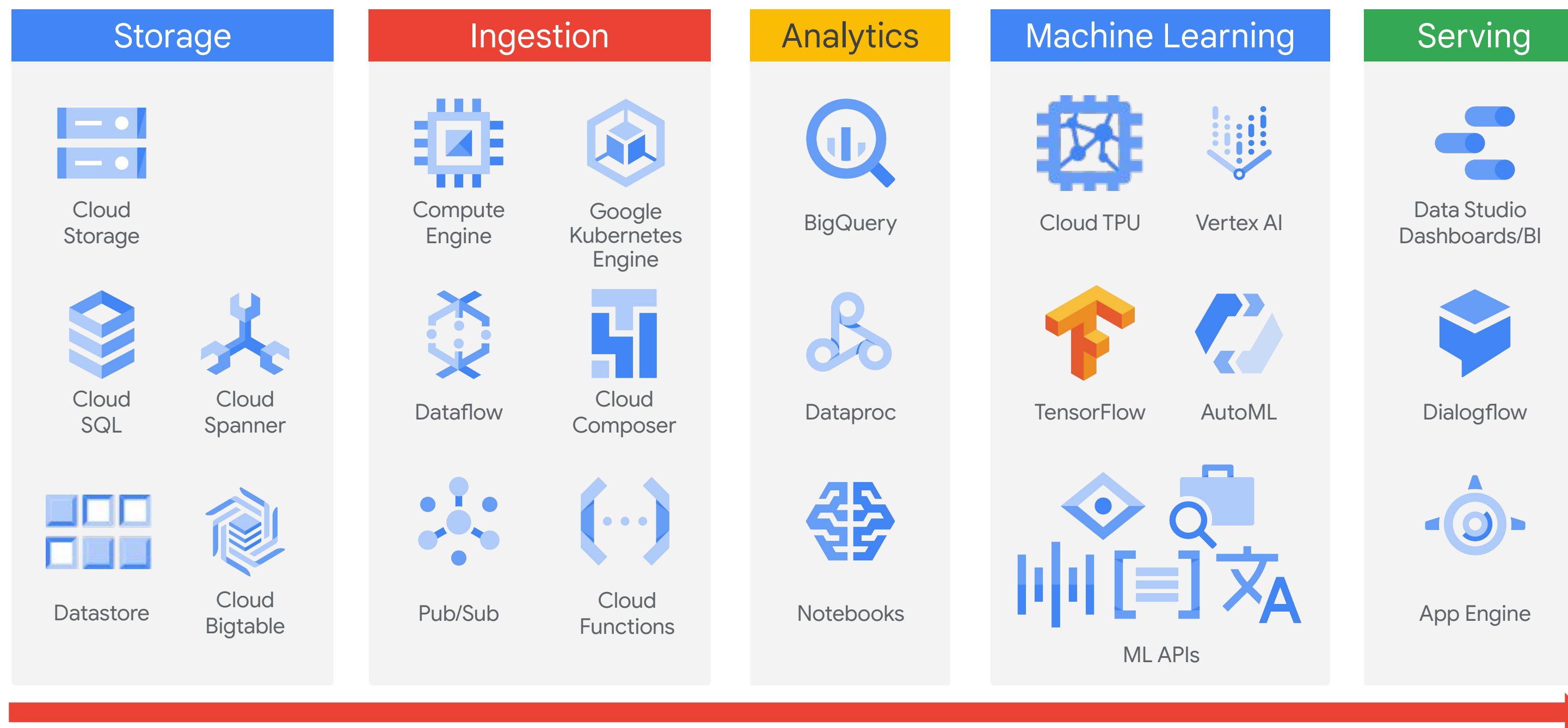


# Example architecture

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 Google Cloud storage solutions



# 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

# 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**



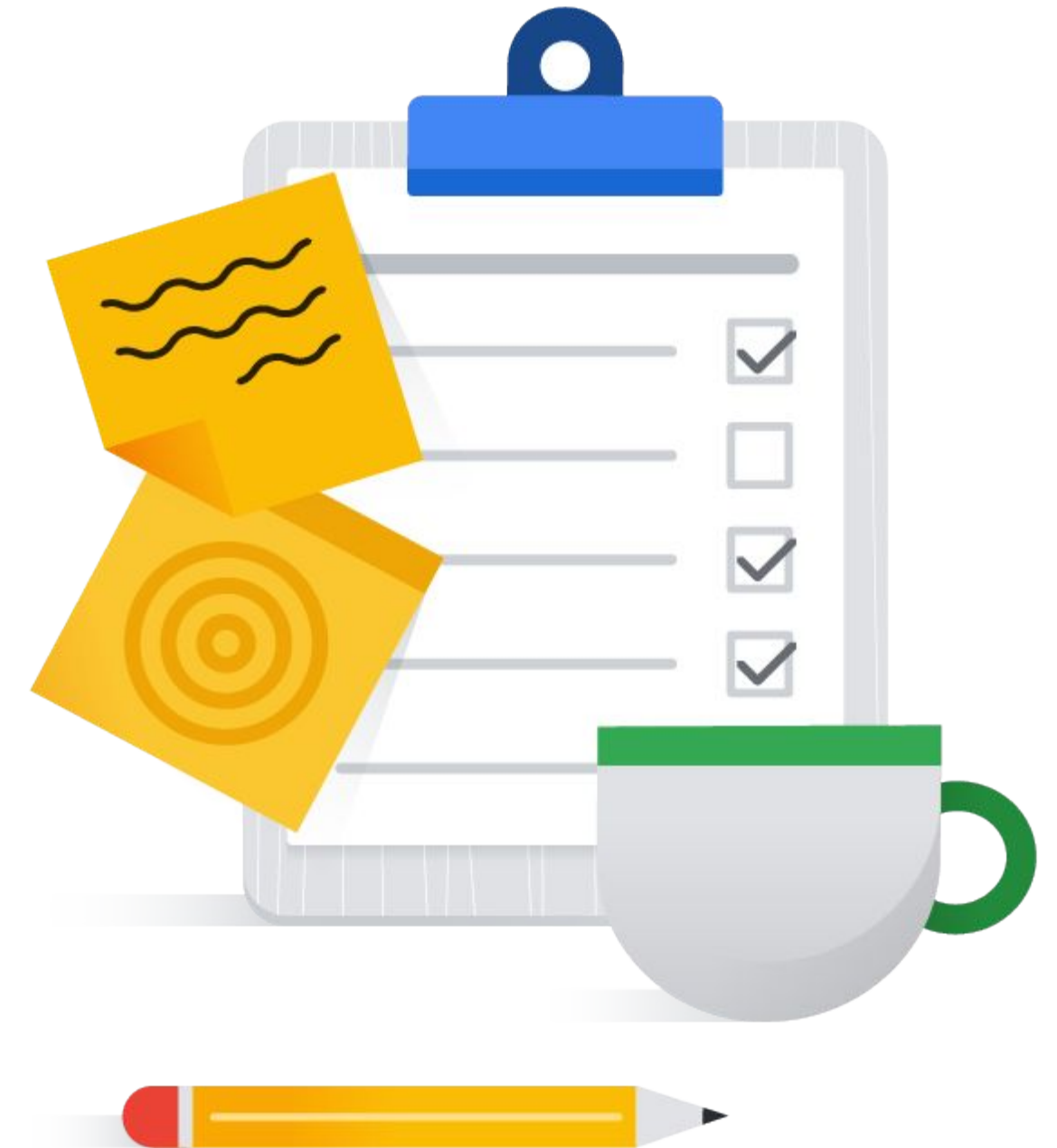
# Data lake versus data warehouse

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

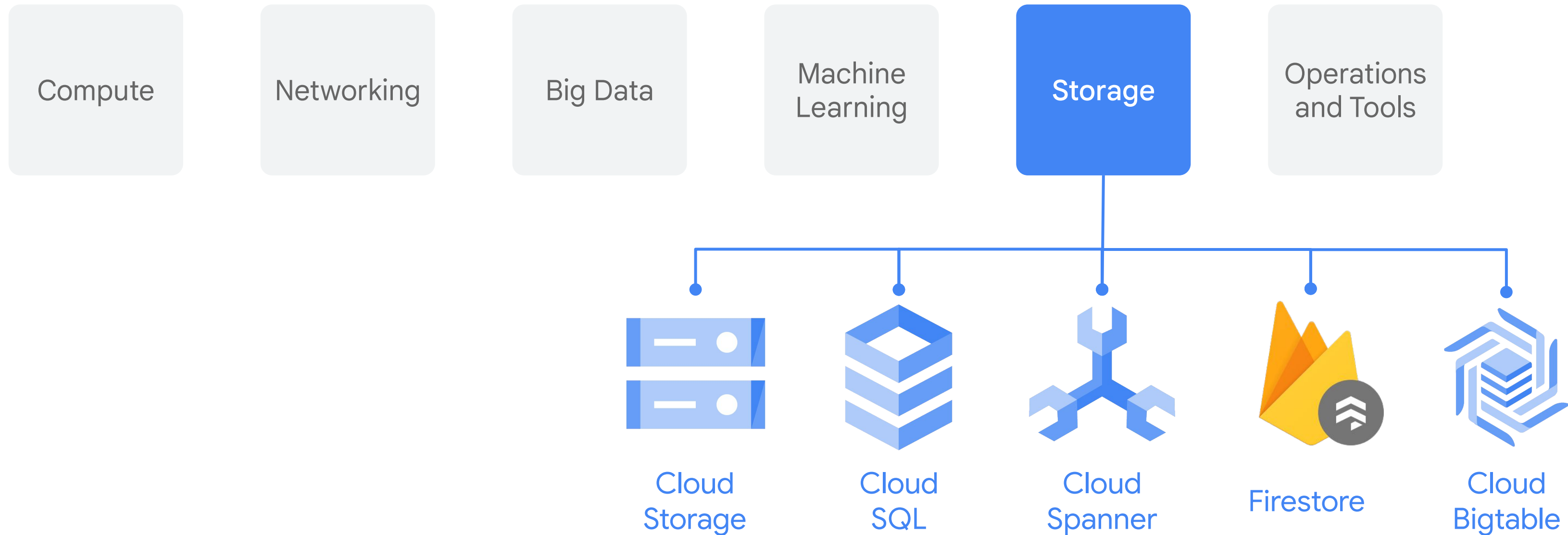
- Typically loaded only after a use case is defined.
- Processed/organized/transformed.
- Provide faster insights.
- Current/historical data for reporting.
- Tends to have consistent schema shared across applications.

# Building a Data Lake

01	Introduction to data lakes
02	Data storage and ETL options on Google Cloud
03	Build a data lake using Cloud Storage
04	Secure Cloud Storage
05	Store all sorts of data types
06	Cloud SQL as a relational data lake



# Storage options for your data on Google Cloud





# 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

## ELT



Extract, Load, and Transform

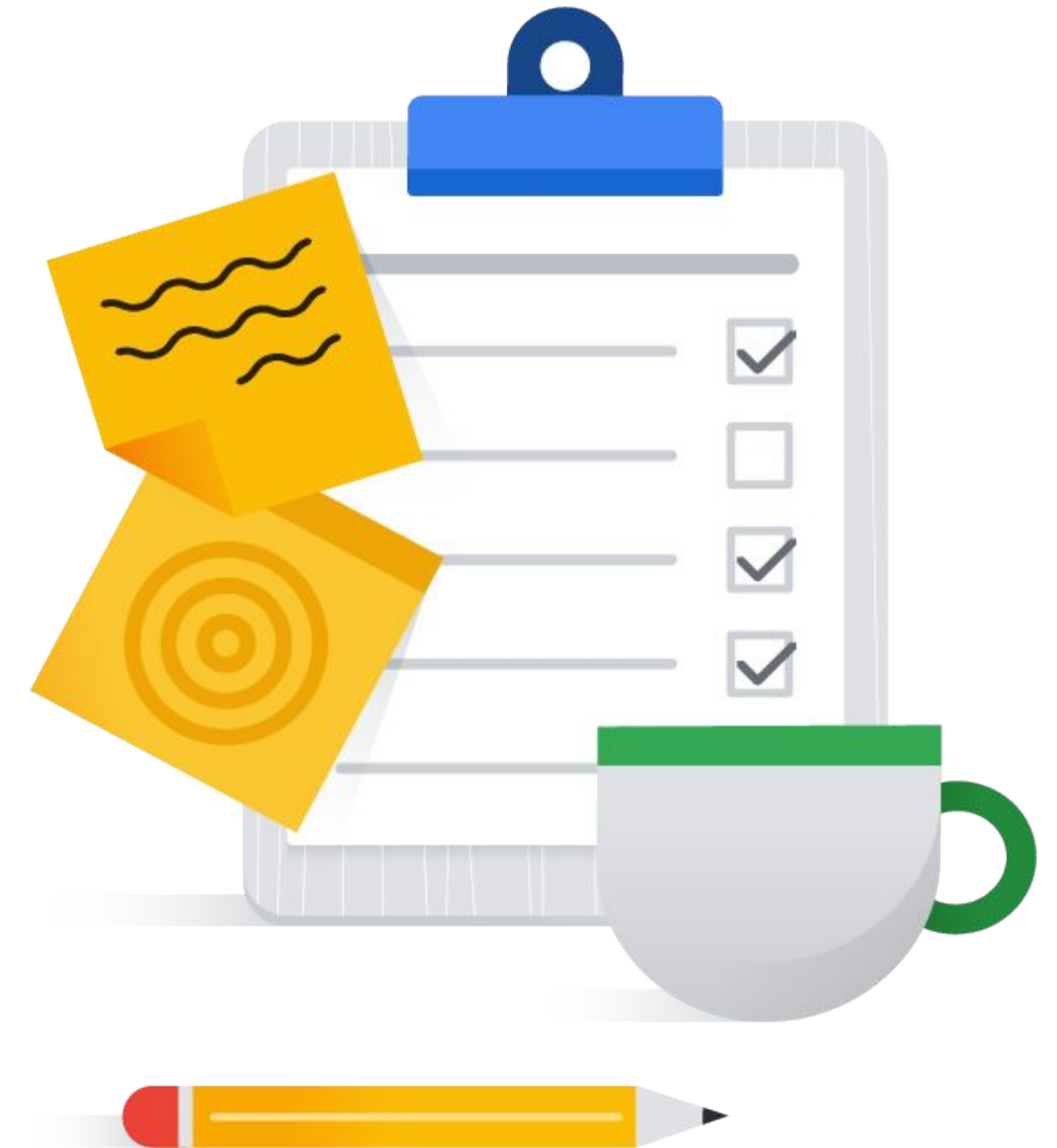
## ETL



Extract, Transform, and Load

# Building a Data Lake

01	Introduction to data lakes
02	Data storage and ETL options on Google Cloud
03	<b>Build a data lake using Cloud Storage</b>
04	Secure Cloud Storage
05	Store all sorts of data types
06	Cloud SQL as a relational data lake

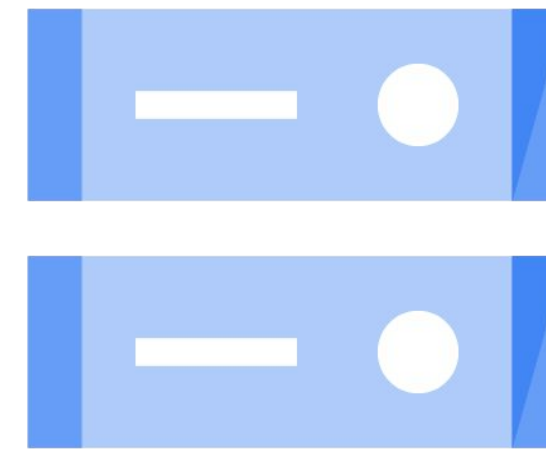




# Cloud Storage

Qualities that Cloud Storage contributes to data engineering solutions:

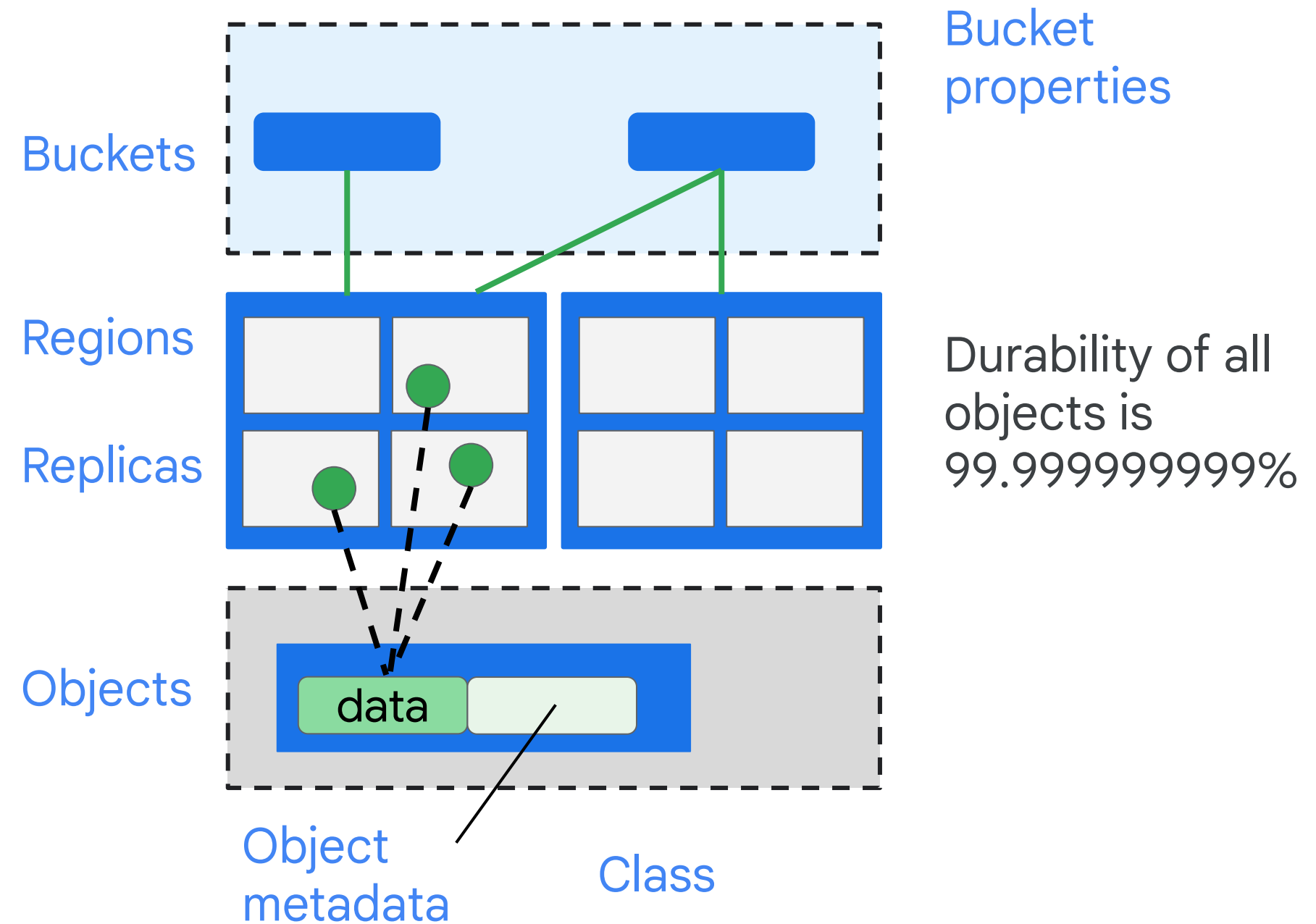
- ✓ Persistence
- ✓ Durability
- ✓ Strong consistency
- ✓ Availability
- ✓ High throughput



Cloud Storage

# How does Cloud Storage work?

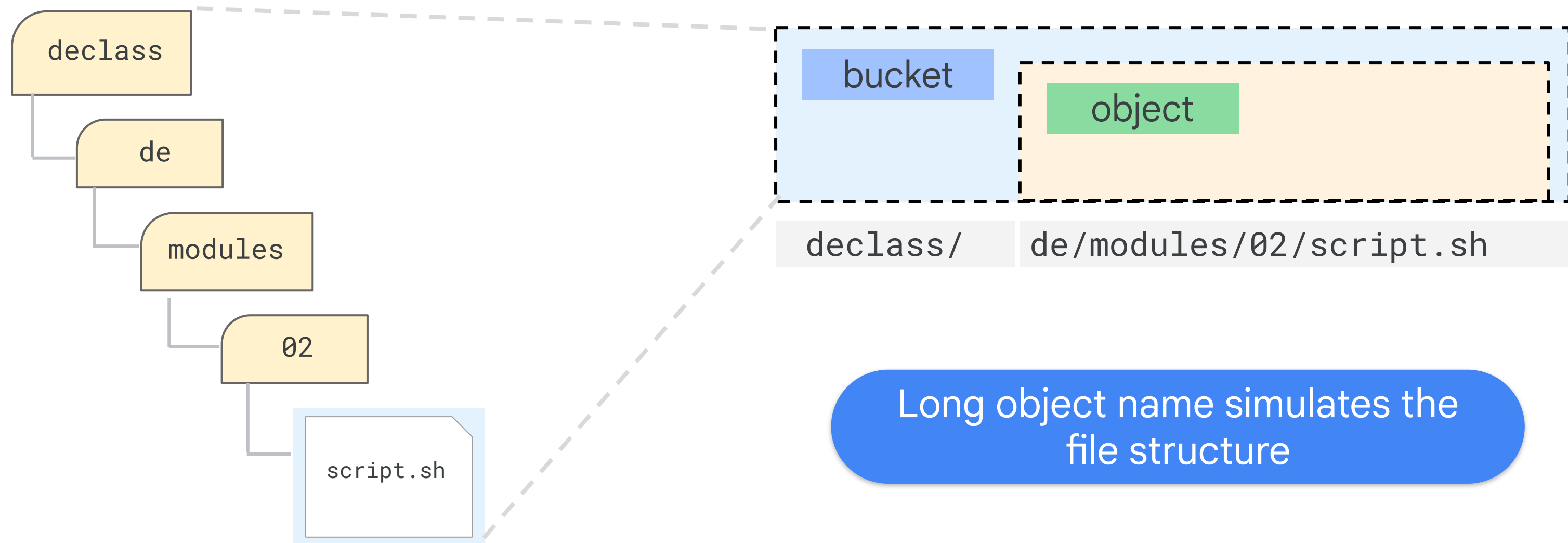
- Single global namespace simplifies locating buckets and objects
- Location to control latency
- Durability and availability
- Long object names simulate structure



# Overview of storage classes

Storage Class	Minimum duration	Availability SLA	Typical monthly availability	Use cases	Name for APIs and gsutil
Standard Storage	None	Multi-region 99.95% Dual-region 99.95% Region 99.9%	>99.99% availability in multi-regions and dual-regions; 99.99% in regions	Access data frequently ("hot" data) and/or store for brief periods <ul style="list-style-type: none"> <li>• Serve website content</li> <li>• Stream videos</li> <li>• Interactive workloads</li> <li>• Mobile and gaming apps</li> </ul>	STANDARD
Nearline Storage	30 days	Multi-region 99.9% Dual-region 99.9% Region 99.0%	99.95% availability in multi-regions and dual-regions; 99.9% in regions	Read/modify data ≤ once per month <ul style="list-style-type: none"> <li>• Data backup</li> <li>• Serve long-tail multimedia content</li> </ul>	NEARLINE
Coldline Storage	90 days	Multi-region 99.9% Dual-region 99.9% Region 99.0%	99.95% availability in multi-regions and dual-regions; 99.9% in regions	Read/modify data no more than once a quarter	COLDLINE
Archive Storage	365 days	None	99.95% availability in multi-regions and dual-regions; 99.9% in regions	Read/modify data < once a year <ul style="list-style-type: none"> <li>• Cold data storage</li> <li>• Disaster recovery</li> </ul>	ARCHIVE

# Cloud Storage simulates a file system

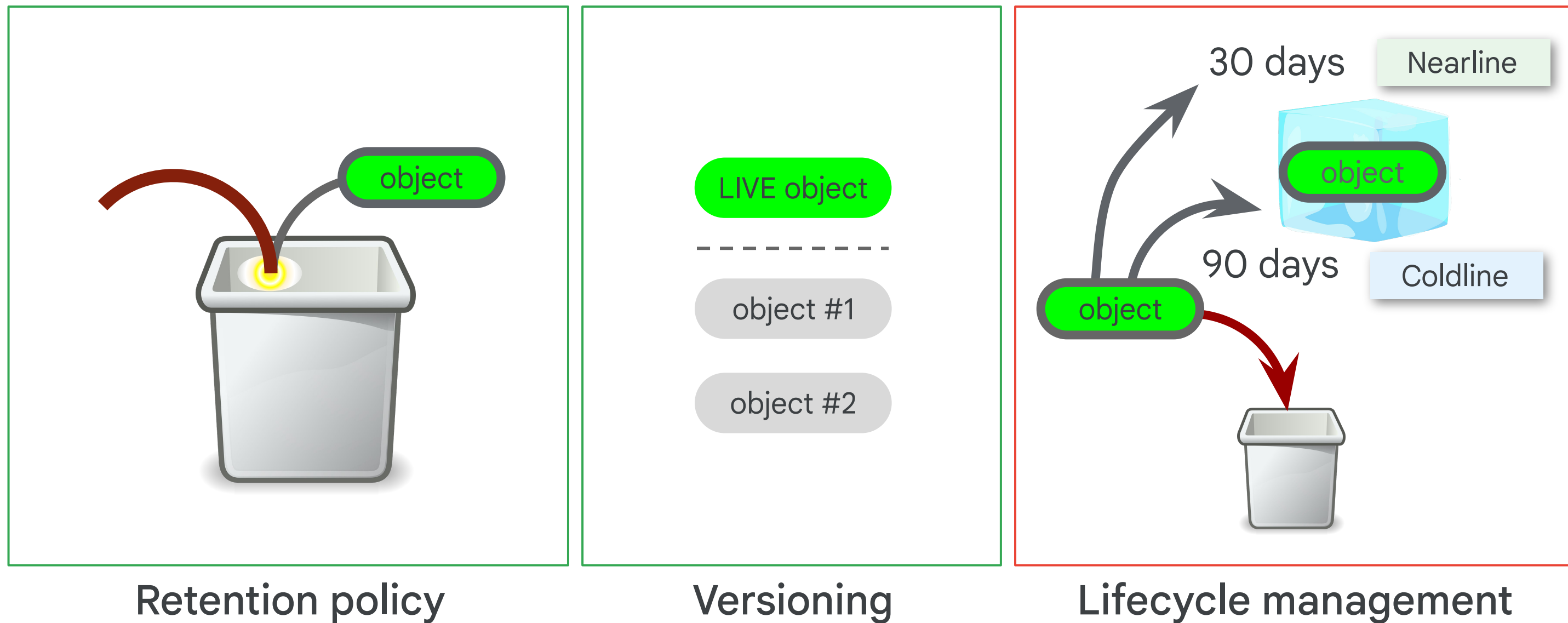


File access `gs://declass/de/modules/02/script.sh`

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

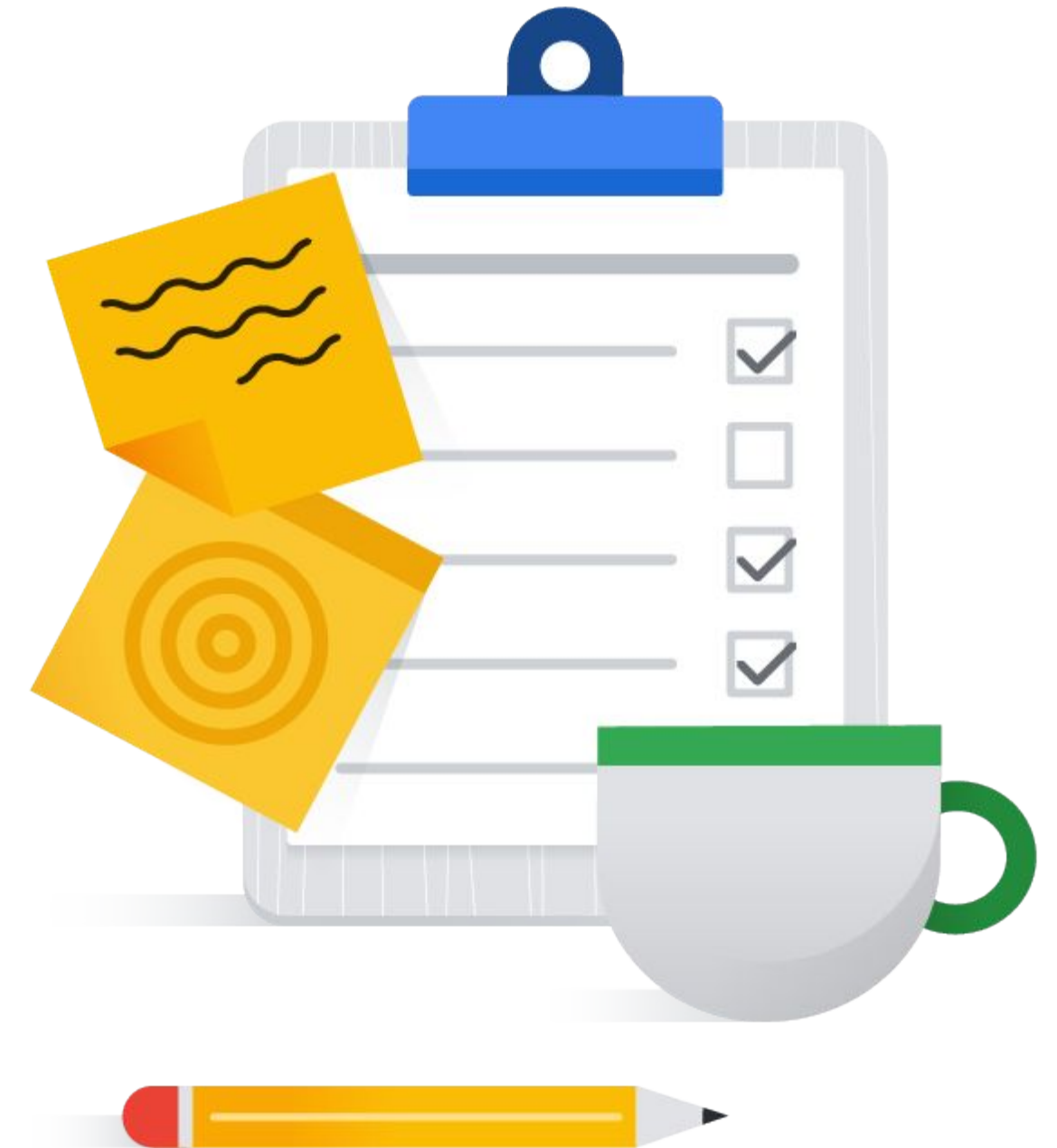


# Cloud Storage has many object management features

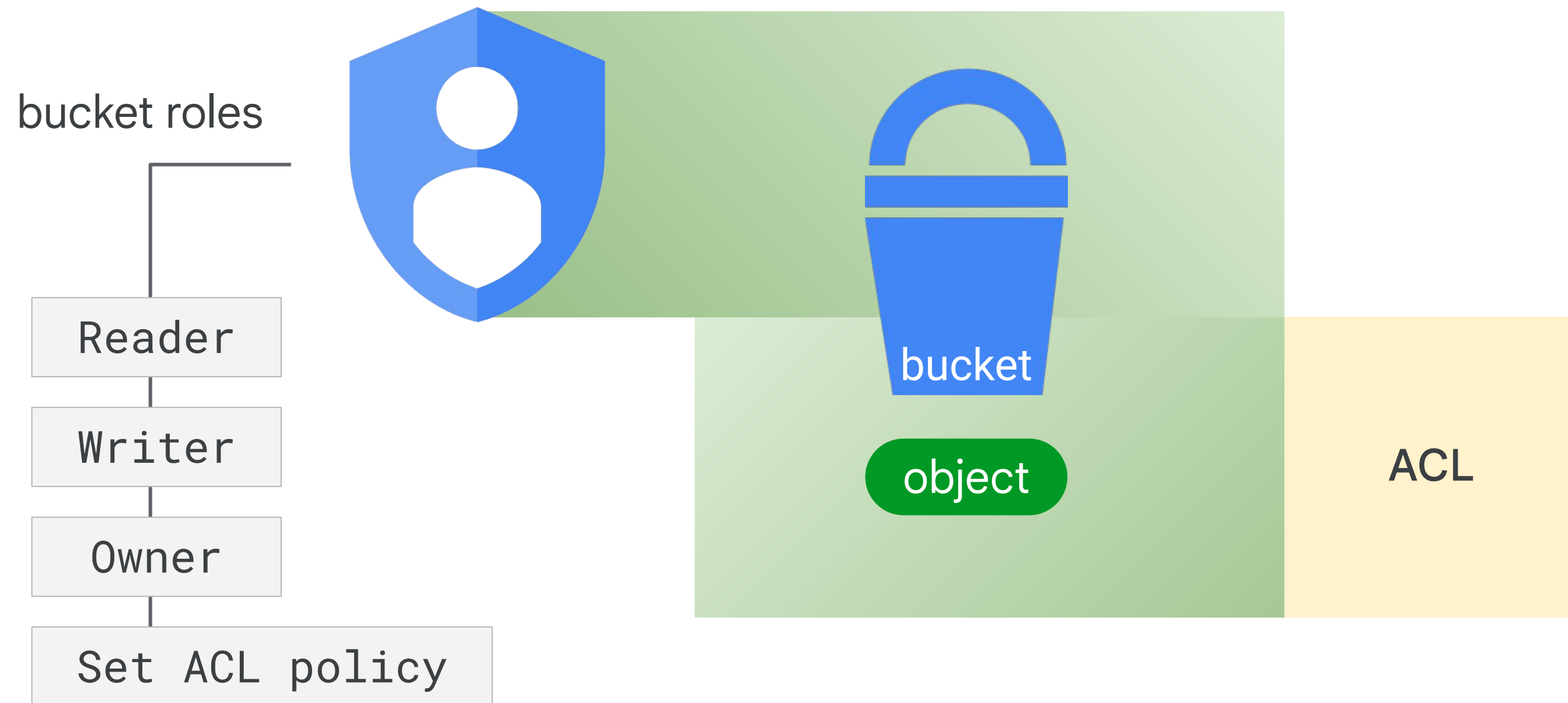


# Building a Data Lake

01	Introduction to data lakes
02	Data storage and ETL options on Google Cloud
03	Build a data lake using Cloud Storage
04	<a href="#">Secure Cloud Storage</a>
05	Store all sorts of data types
06	Cloud SQL as a relational data lake



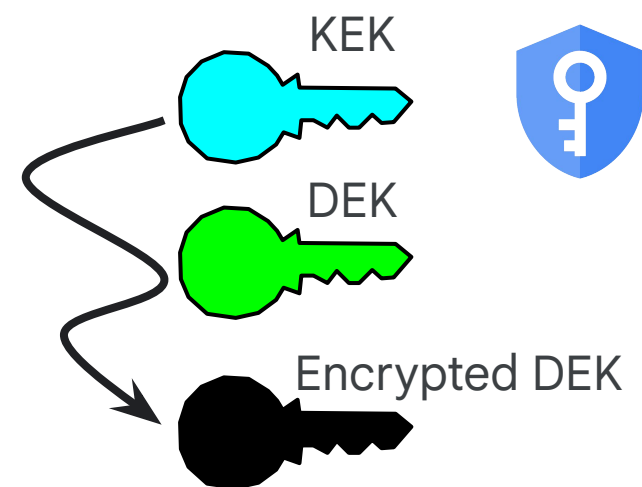
# Controlling access with IAM and access lists



# Data encryption options for many requirements

## GMEK

Google-managed  
encryption keys



Cloud KMS Key Management  
Service Encryption is automatic

## CMEK

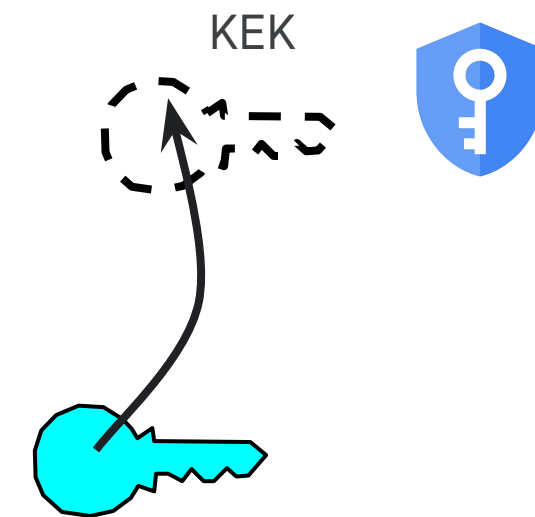
Customer-managed  
encryption keys



You control the creation and  
existence of the KEK key in  
Cloud KMS

## CSEK

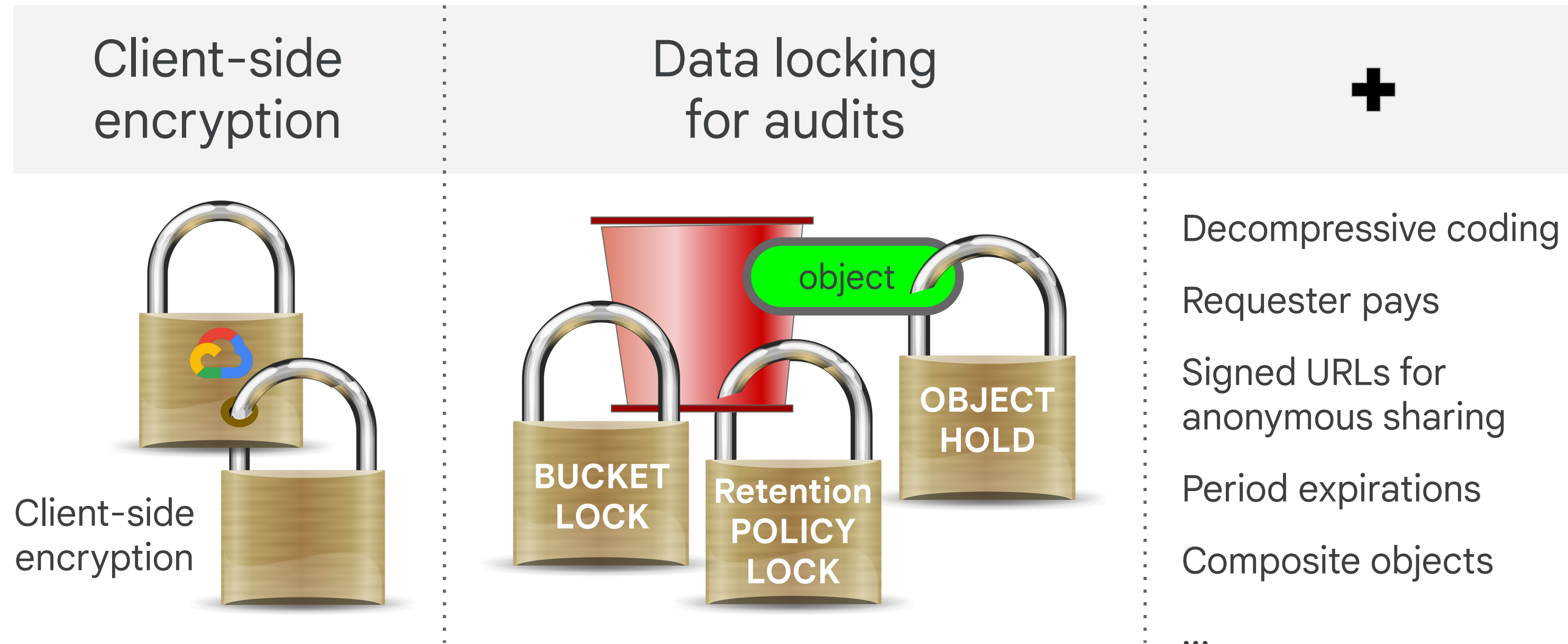
Customer-supplied  
encryption keys



You provide the KEK key

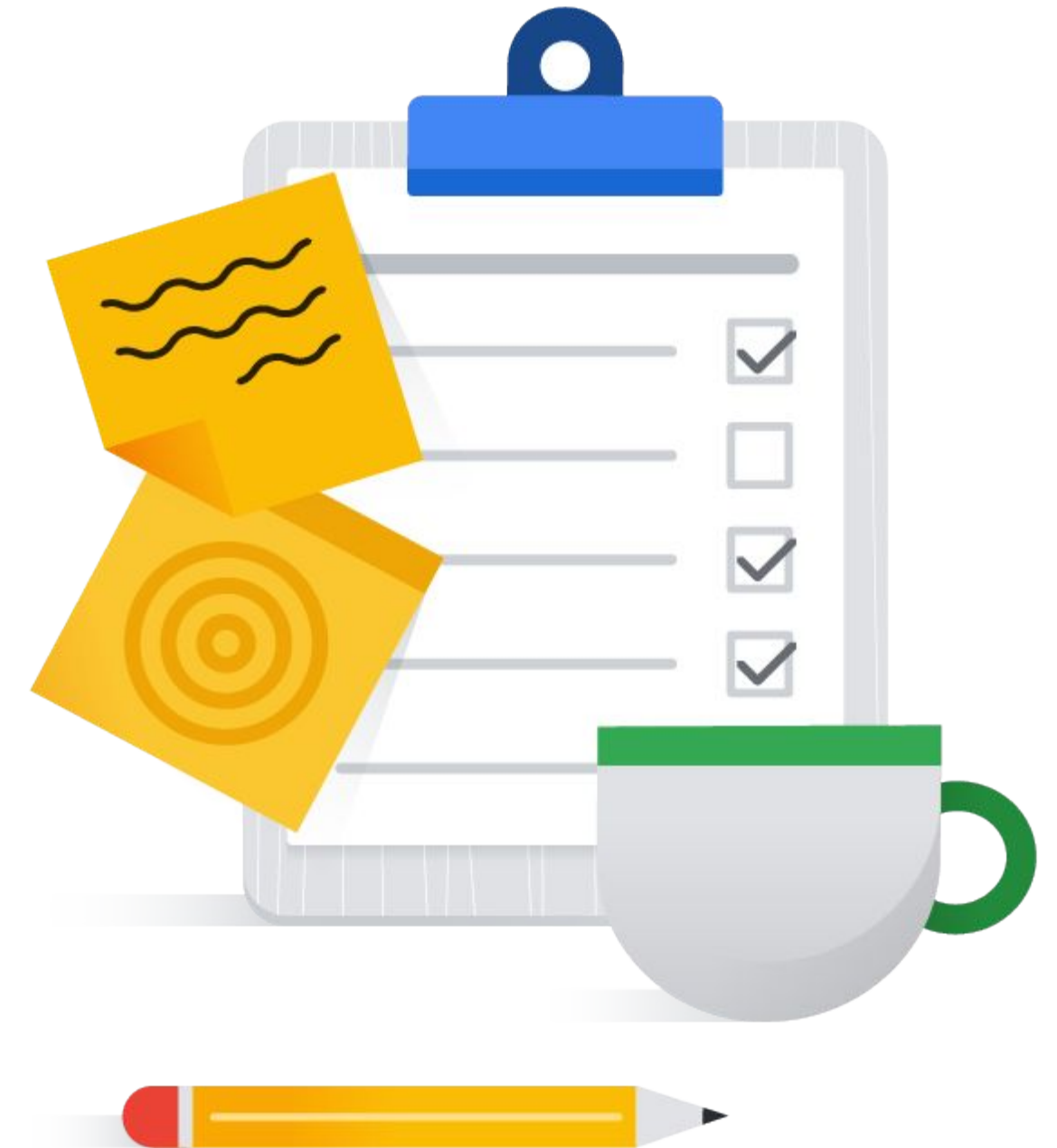


# Cloud Storage supports many special use cases

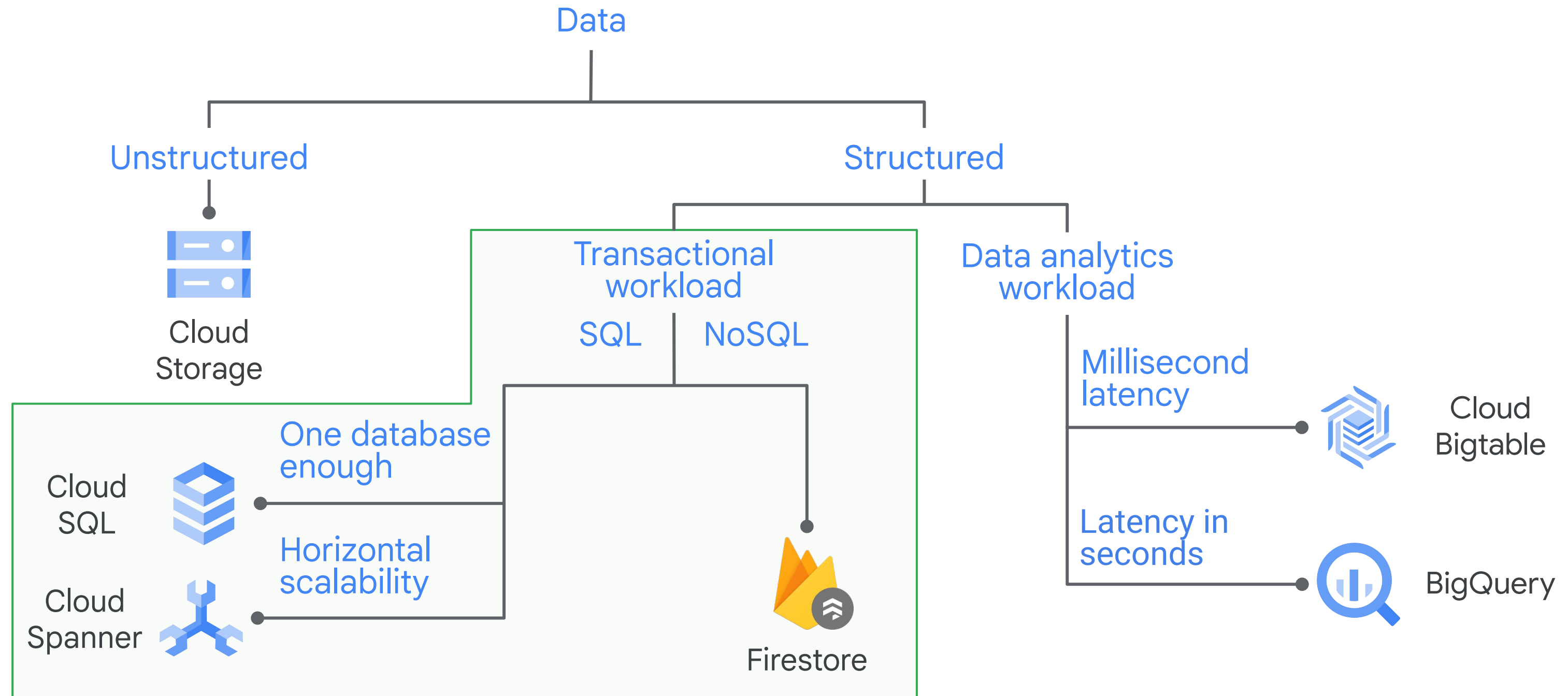


# Building a Data Lake

01	Introduction to data lakes
02	Data storage and ETL options on Google Cloud
03	Build a data lake using Cloud Storage
04	Secure Cloud Storage
05	Store all sorts of data types
06	Cloud SQL as a relational data lake



# Different considerations for transactional workloads

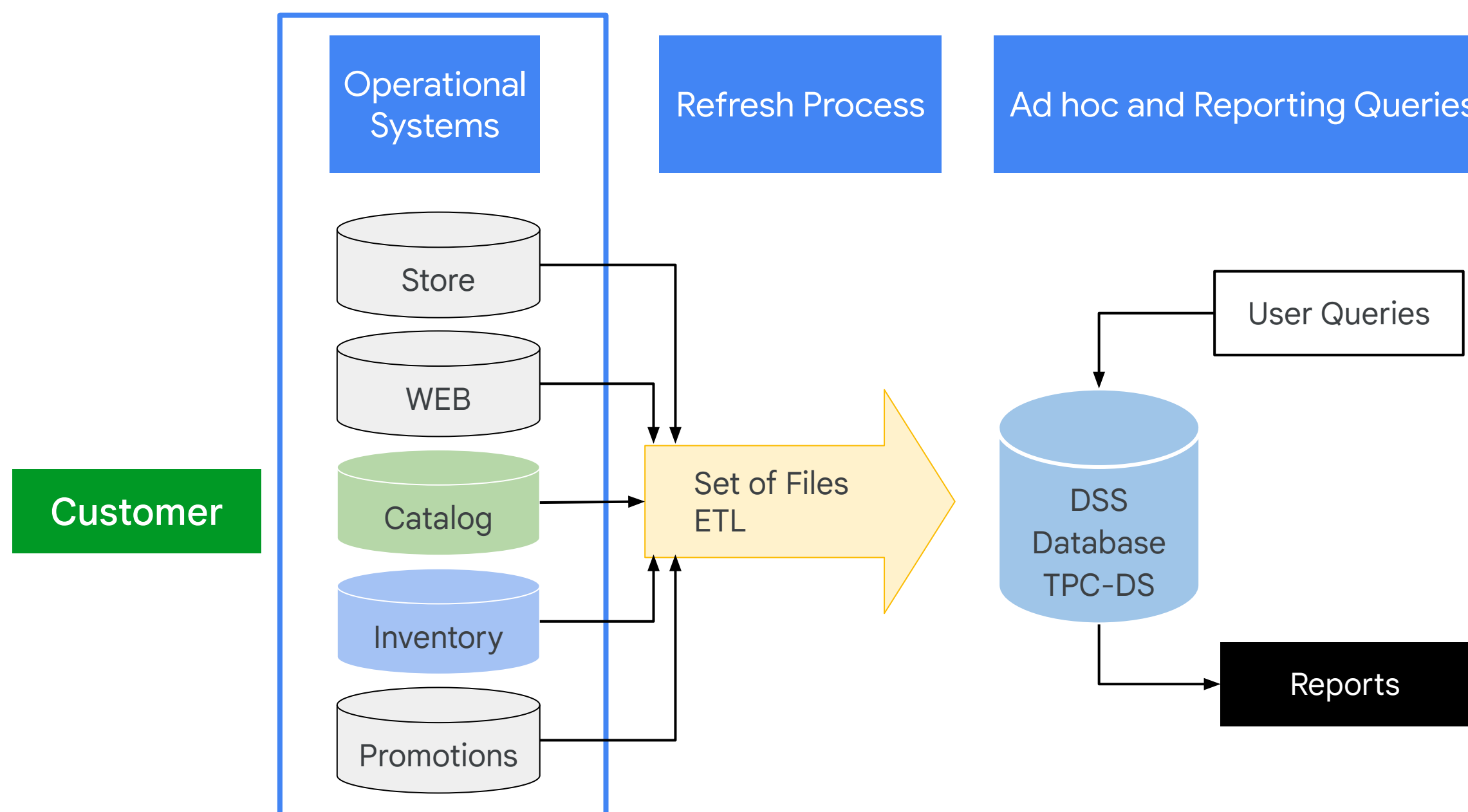


# Transactional versus analytical

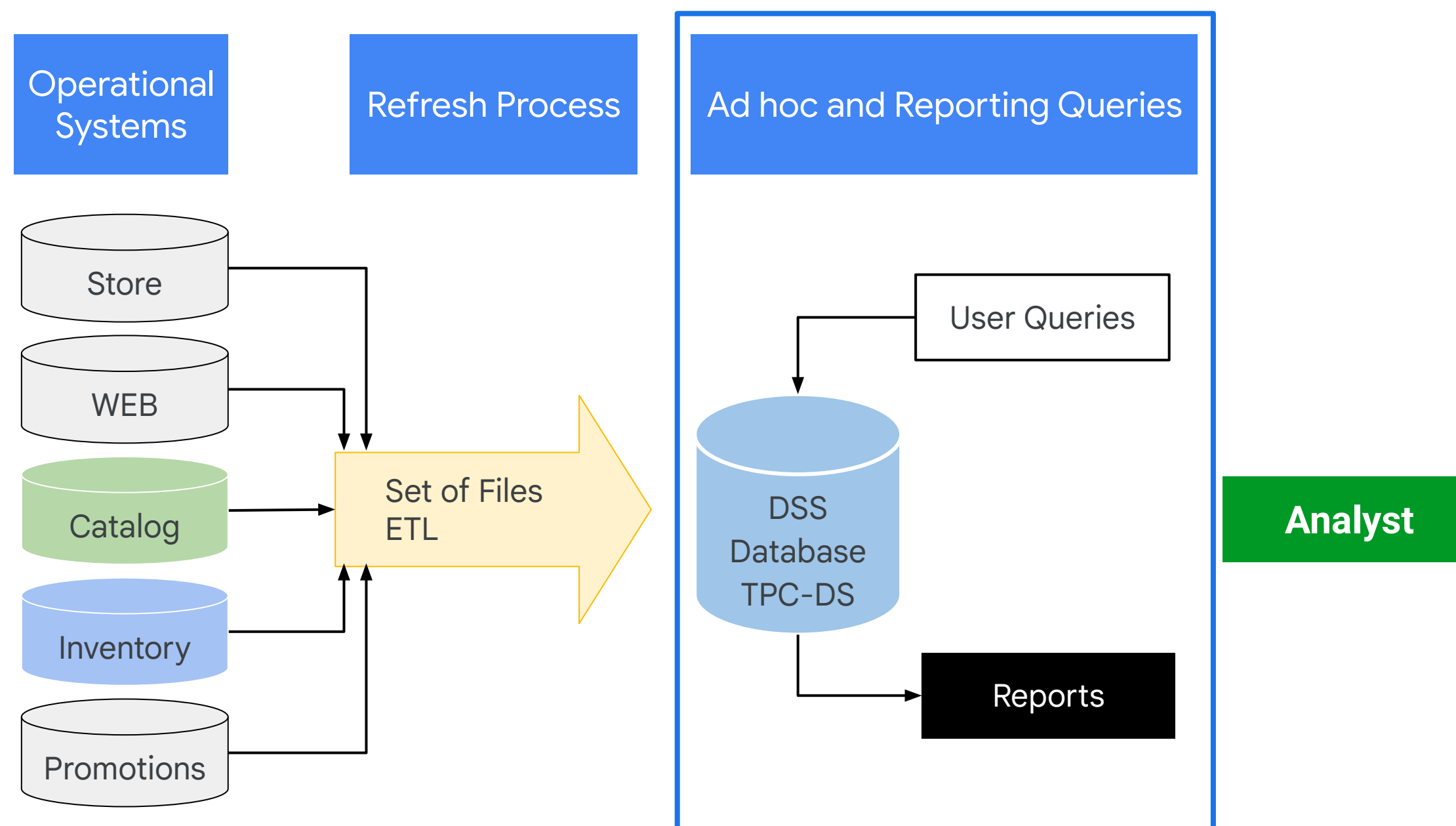
	Transactional	Analytical
Source of data	Operational data; OLTPs are the original source of the data	Consolidation data; OLAP data comes from the various OLTP databases
Purpose of data	Control and run fundamental business tasks	Help with planning, problem solving, and decision support
What the data shows	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



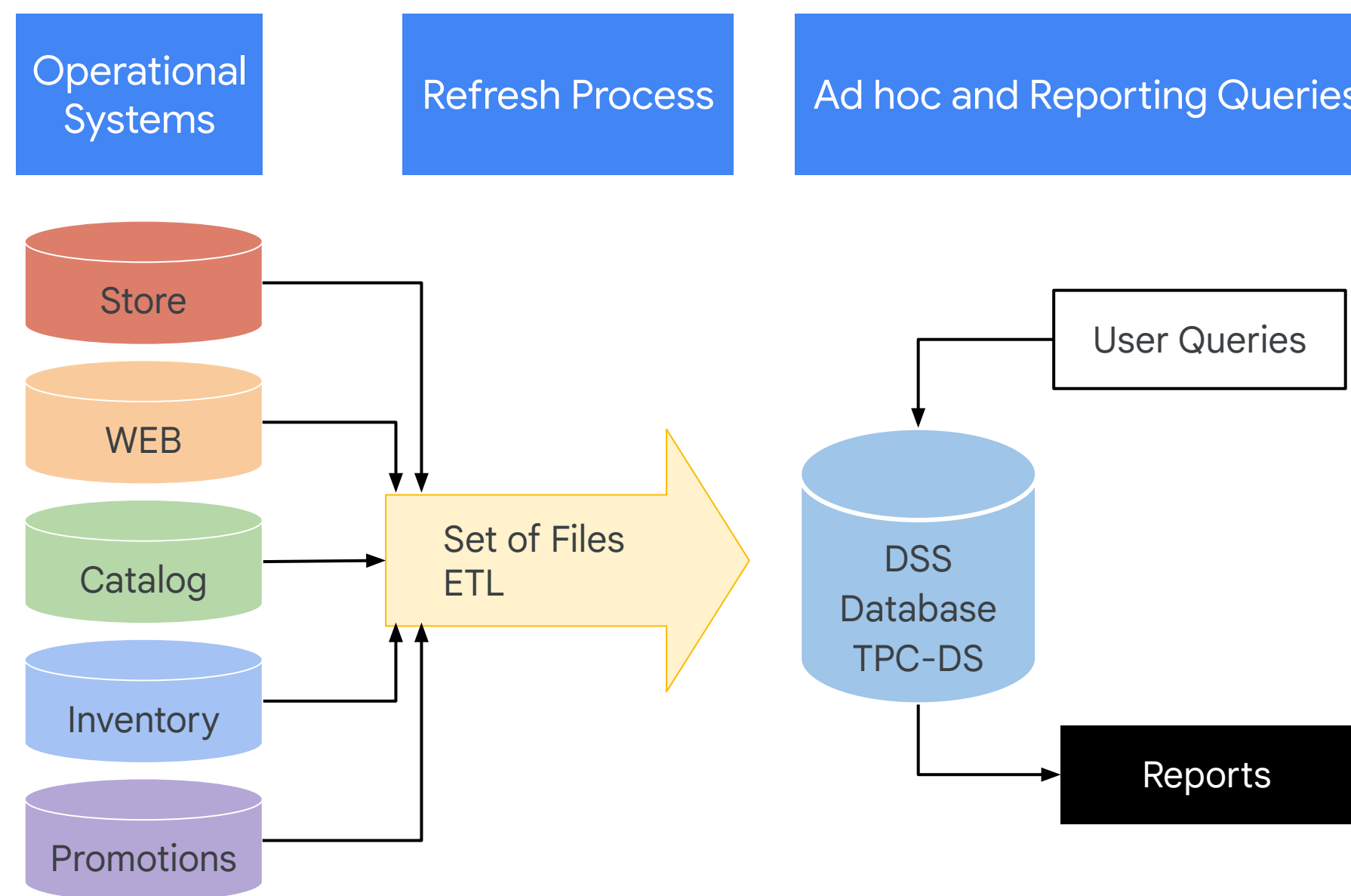
# Transactional systems are 80% writes and 20% reads



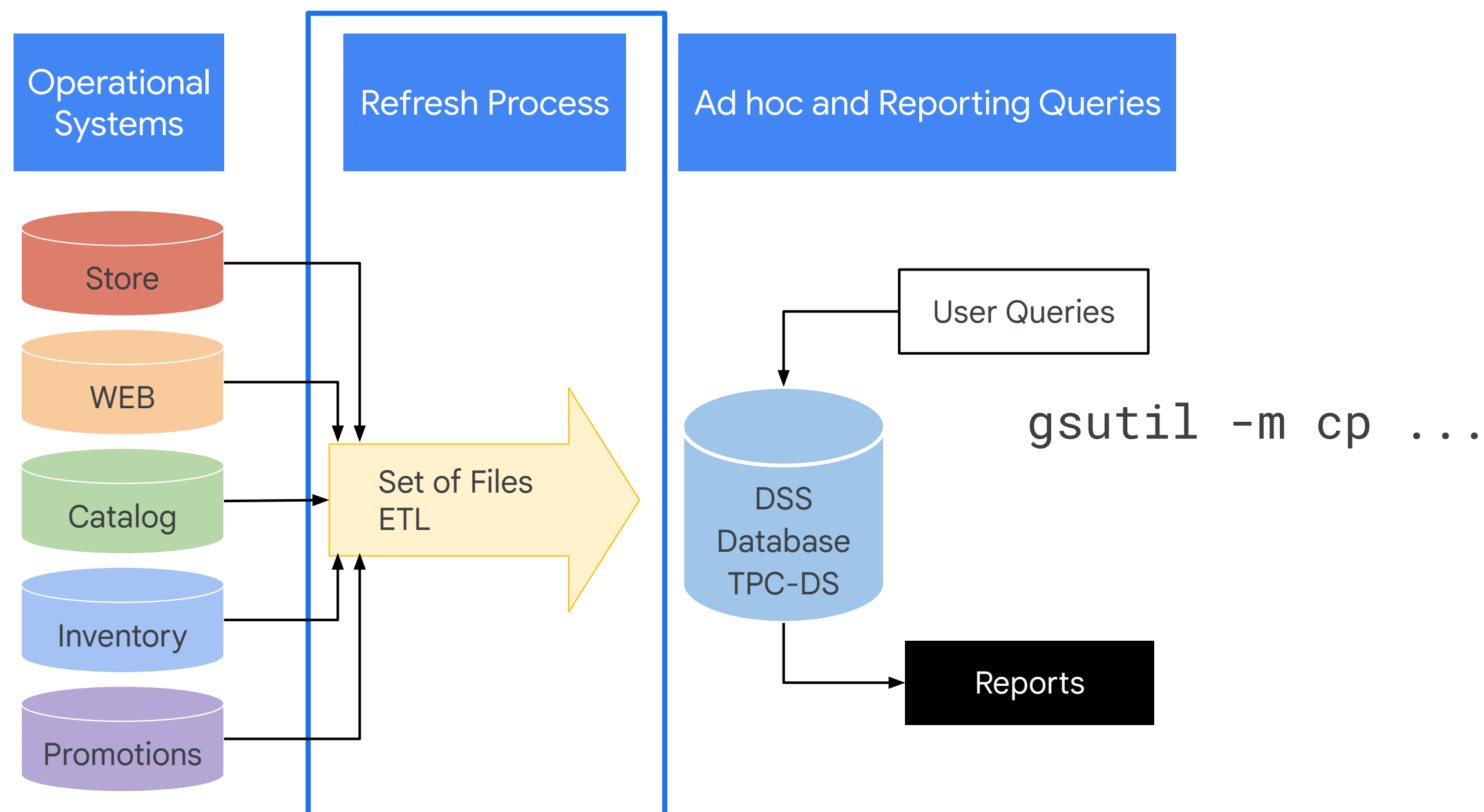
# Analytical systems are 20% writes and 80% reads



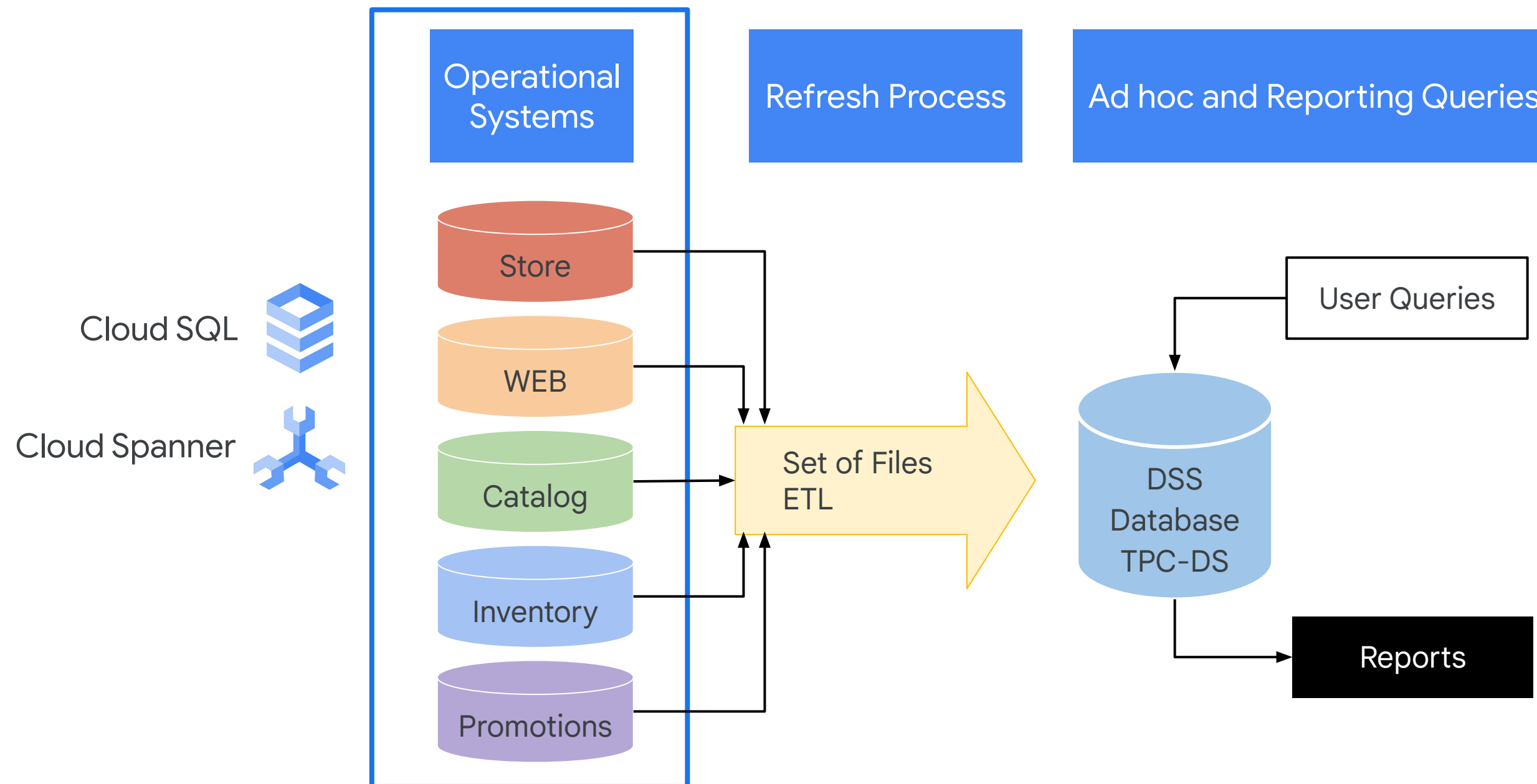
# Data engineers build the pipelines between the systems



# Use Cloud Storage for scalable staging of raw data

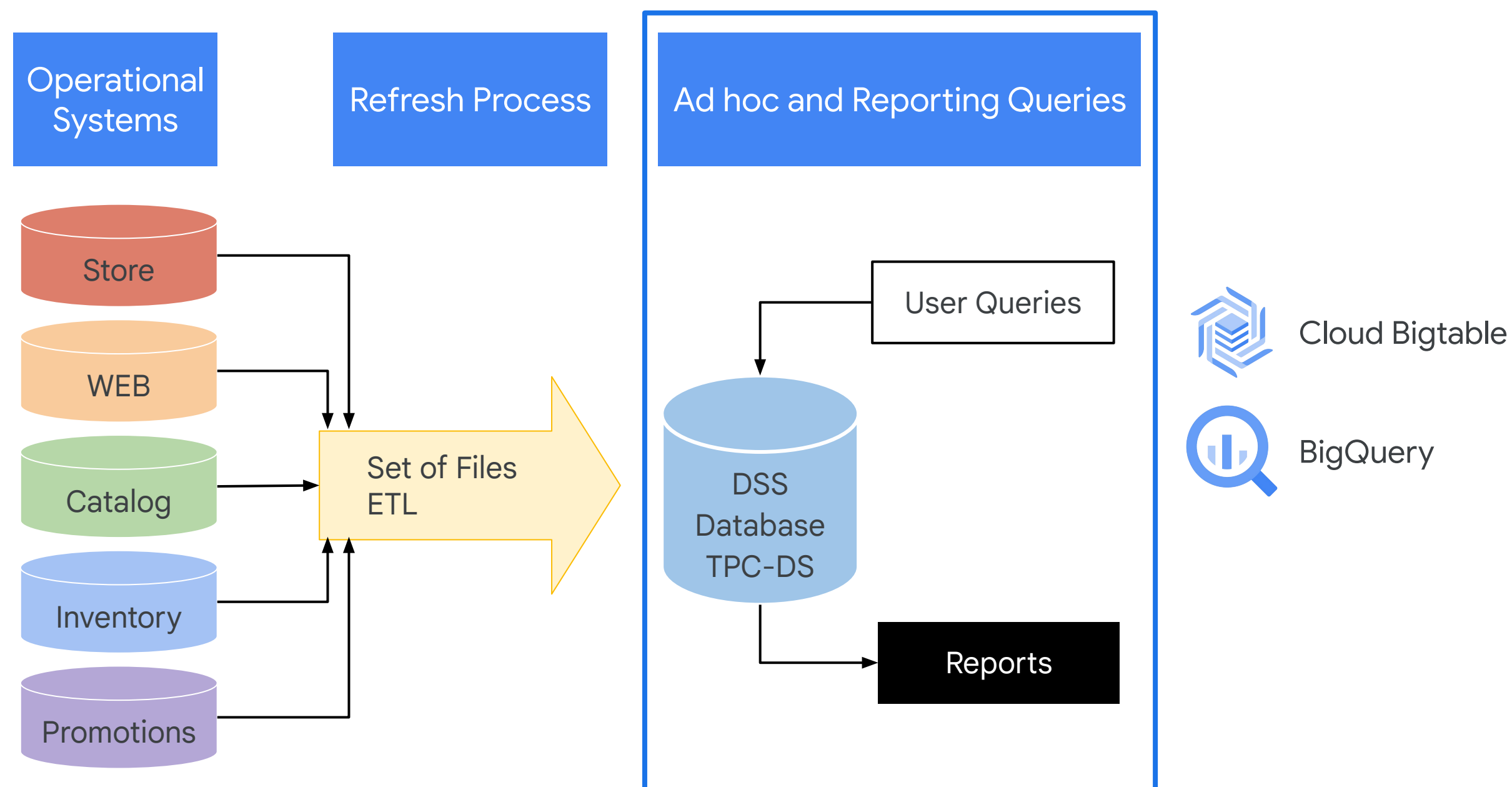


# Choose from cloud relational databases for transactional workloads



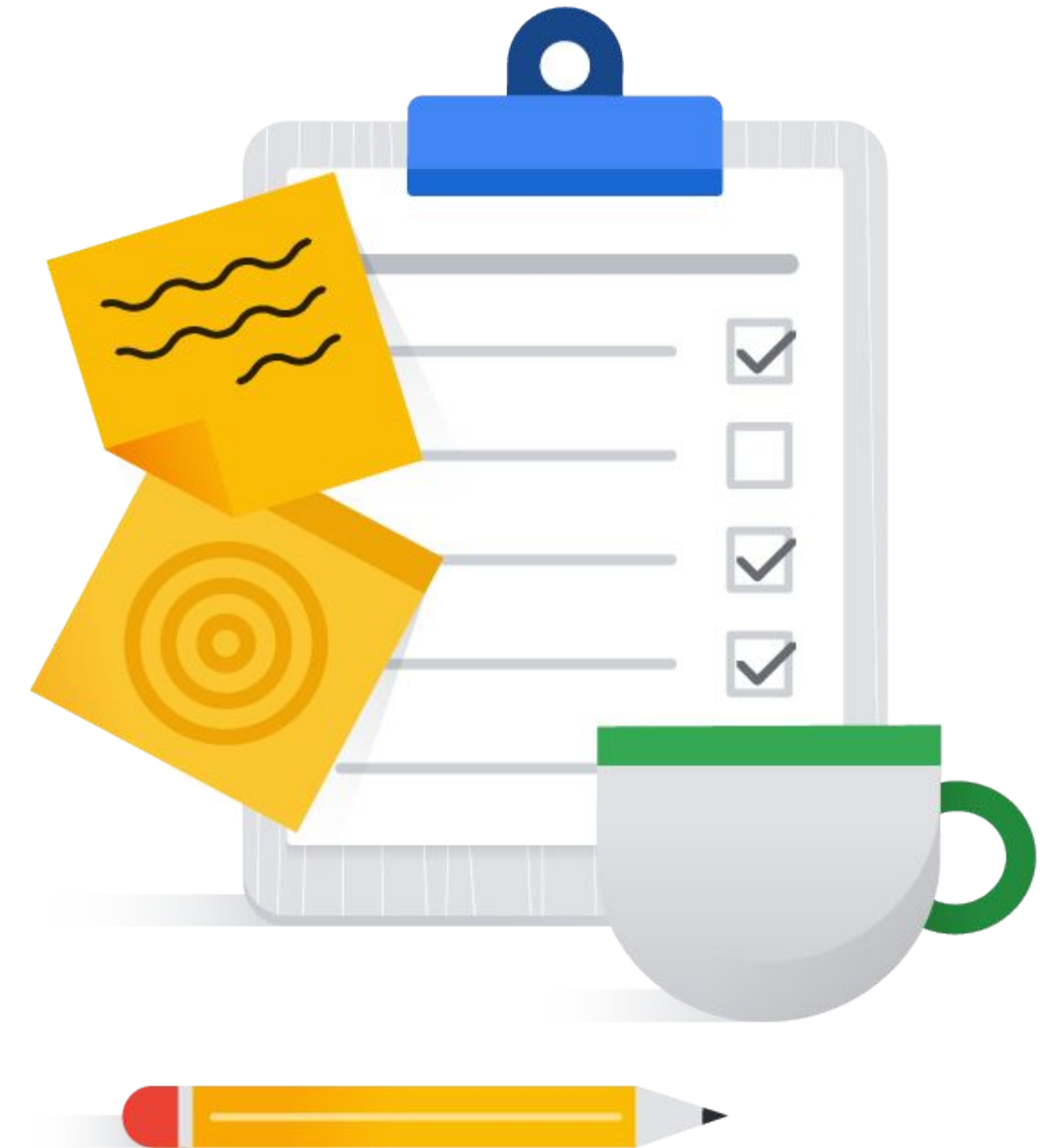


# Choose from cloud data warehouses for analytic workloads



# Building a Data Lake

01	Introduction to data lakes
02	Data storage and ETL options on Google Cloud
03	Build a data lake using Cloud Storage
04	Secure Cloud Storage
05	Store all sorts of data types
06	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 databases in the cloud

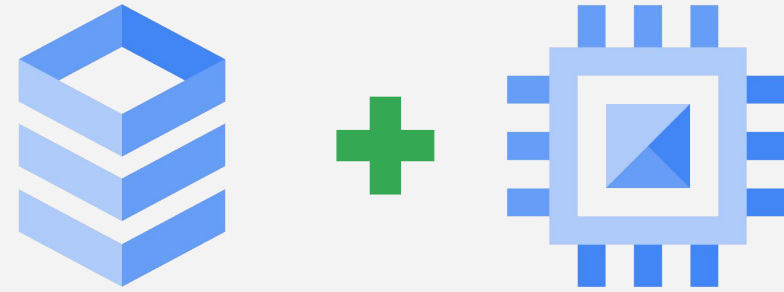


# Cloud SQL can be used with other Google Cloud services



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.



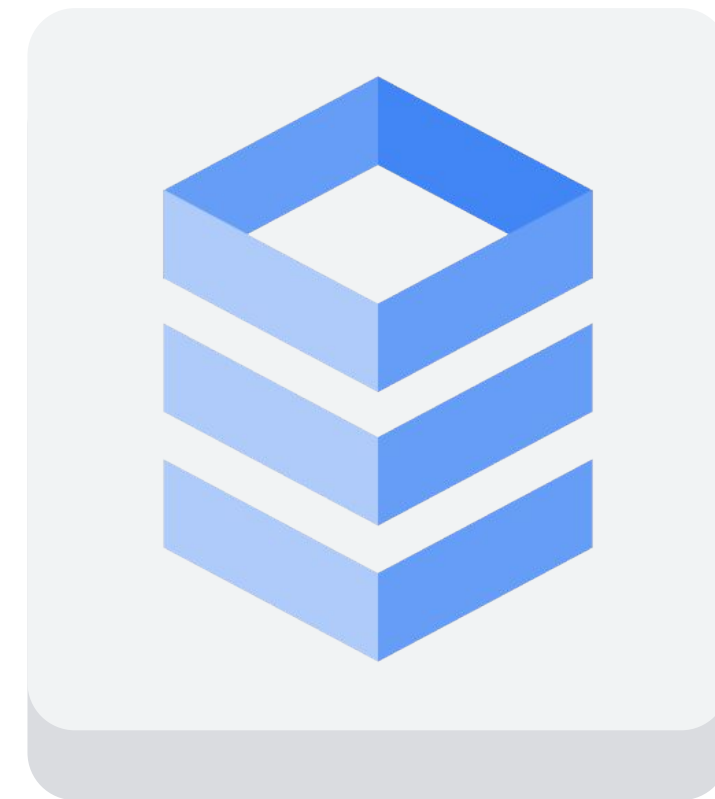
Cloud SQL can be used with external applications and clients.

Standard tools can be used to administer databases.

External read replicas can be configured.

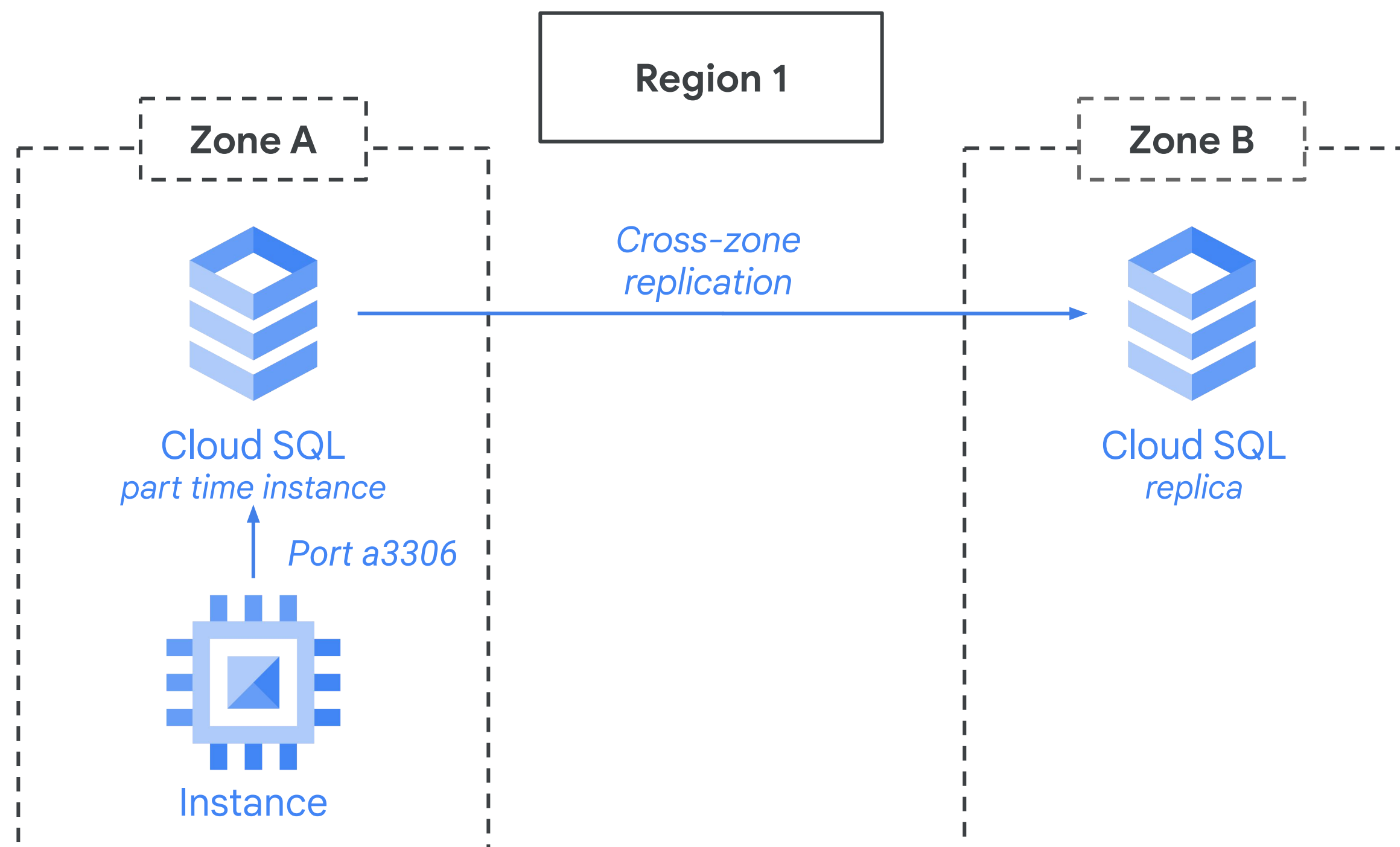
# Backup, recovery, scaling, and security is managed for you

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











# Cloud SQL replication



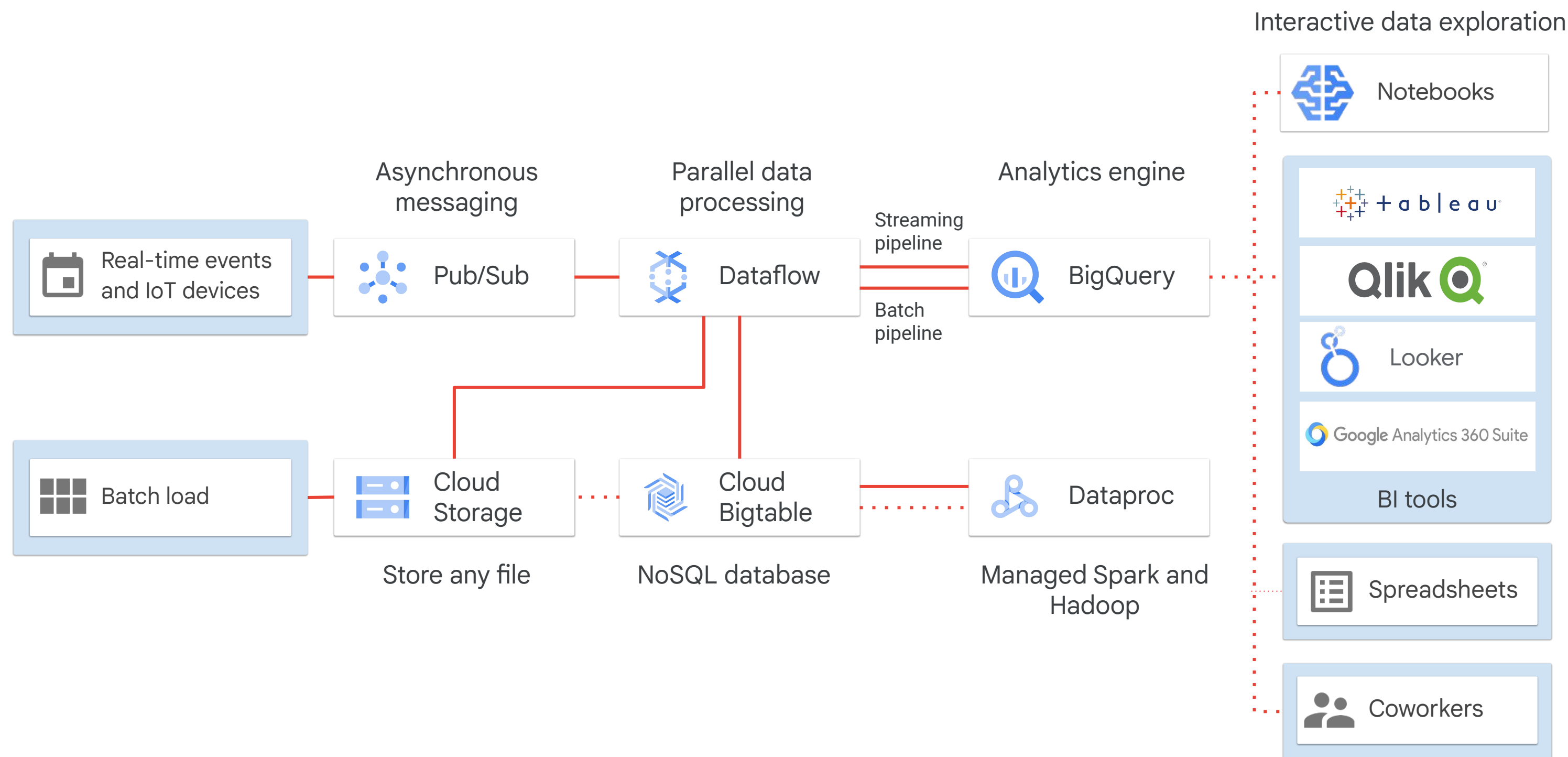
# Cloud SQL replication caveats

- ✓ The failover replica is charged as a separate instance
- ✓ Existing connections to the instance are closed
- ✓ The replica becomes the primary

# Fully managed versus serverless

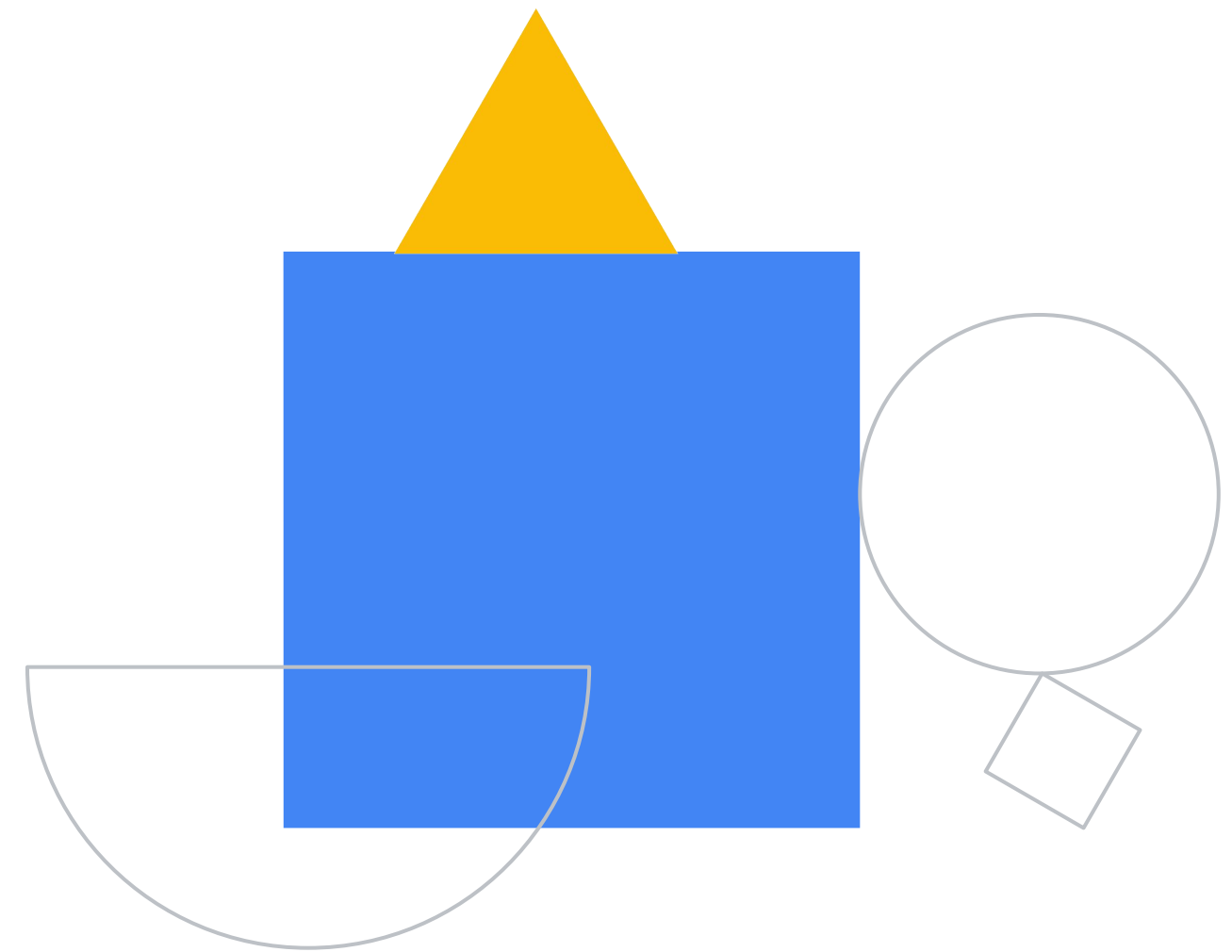
Fully managed	 No setup	 Automated backups, updates, etc.	 Replicated, highly available
Serverless	 No server management	 Fully managed security	 Pay only for usage

# Modern serverless data management architecture



# Lab Intro

Loading Taxi Data into Cloud SQL



# Lab objectives

01

Create a Cloud SQL instance which can hold multiple databases.

02

Create a new Cloud SQL database.

03

Import text data into Cloud SQL.

04

Check the data for integrity.





