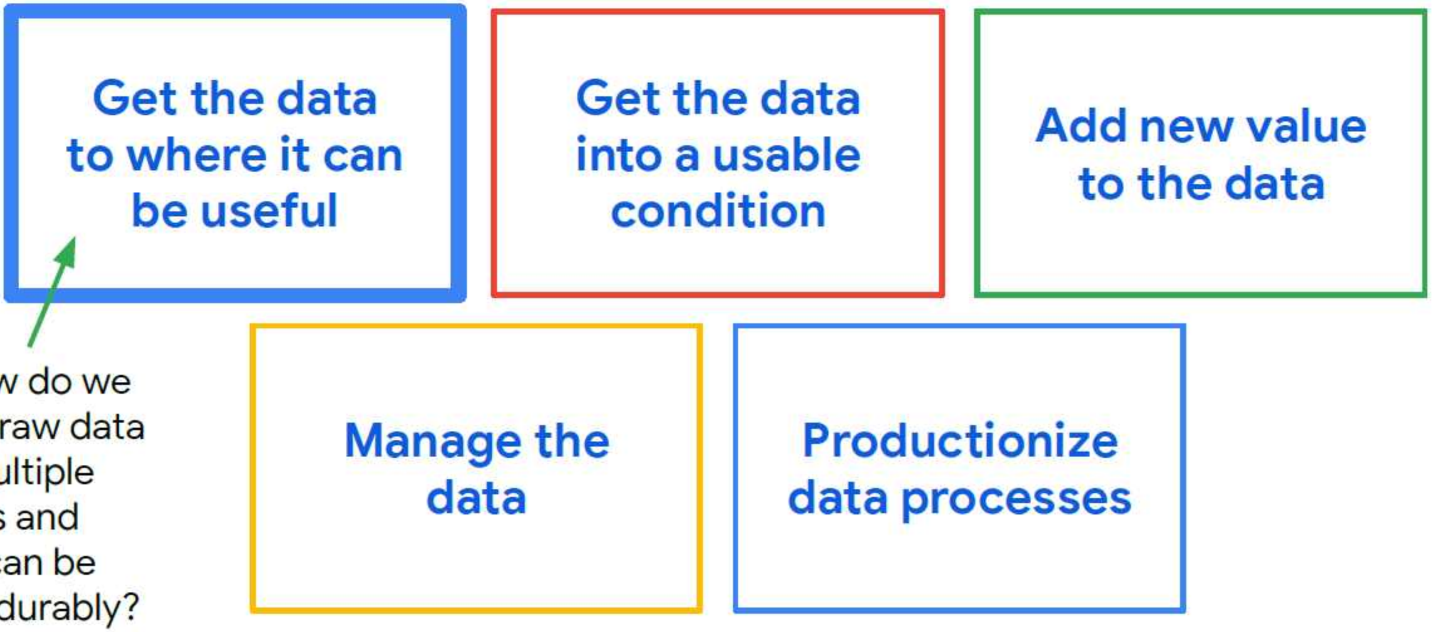
Google Cloud

Introduction to Data Engineering

강사 : 고병화

Explore the role of a data engineer

A data engineer builds data pipelines to enable data-driven decisions



Get the data to where it can be useful

Get the data into a usable condition

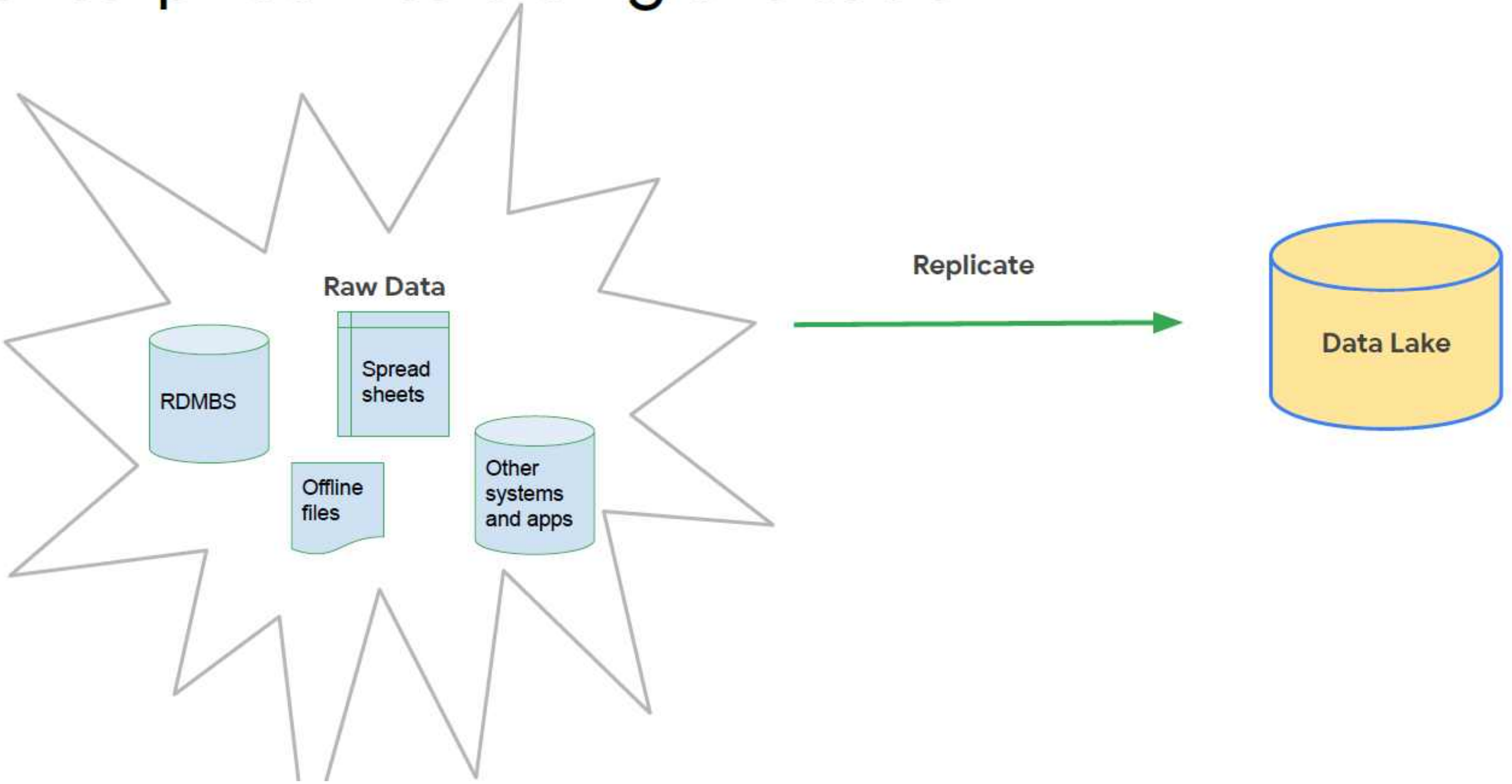
Add new value to the data

Manage the data

Productionize data processes

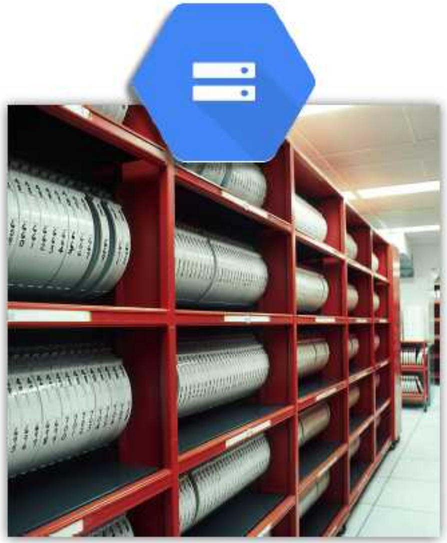
So... how do we get the raw data from multiple systems and where can be store it durably?

A data lake brings together data from across the enterprise into a single location



Cloud Storage is designed for 99.999999999% annual durability

Backup



Replace/decommission infrastructure

Analytics and ML

Content storage and delivery

Quickly create buckets with cloud shell

```
gsutil mb gs://your-project-name
```

What if your data is not usable in its original form?

Extract, Transform, and Load

Data Processing

Cloud Dataproc

Cloud Dataflow

What if your data arrives continuously and endlessly?

THIS DATA DOES NOT WAIT

Streaming Data Processing

Cloud Pub/Sub

Cloud Dataflow

BigQuery

Analyze data engineering challenges

Common challenges encountered by data engineers

Access to data

Data accuracy and quality

Availability of computational resources

Query performance

BigQuery is Google's data warehouse solution

Data warehouse

BigQuery replaces a typical data warehouse hardware setup

Data mart

BigQuery organizes data tables into units called datasets

Data lake

BigQuery defines schemas and issues queries directly on external data sources

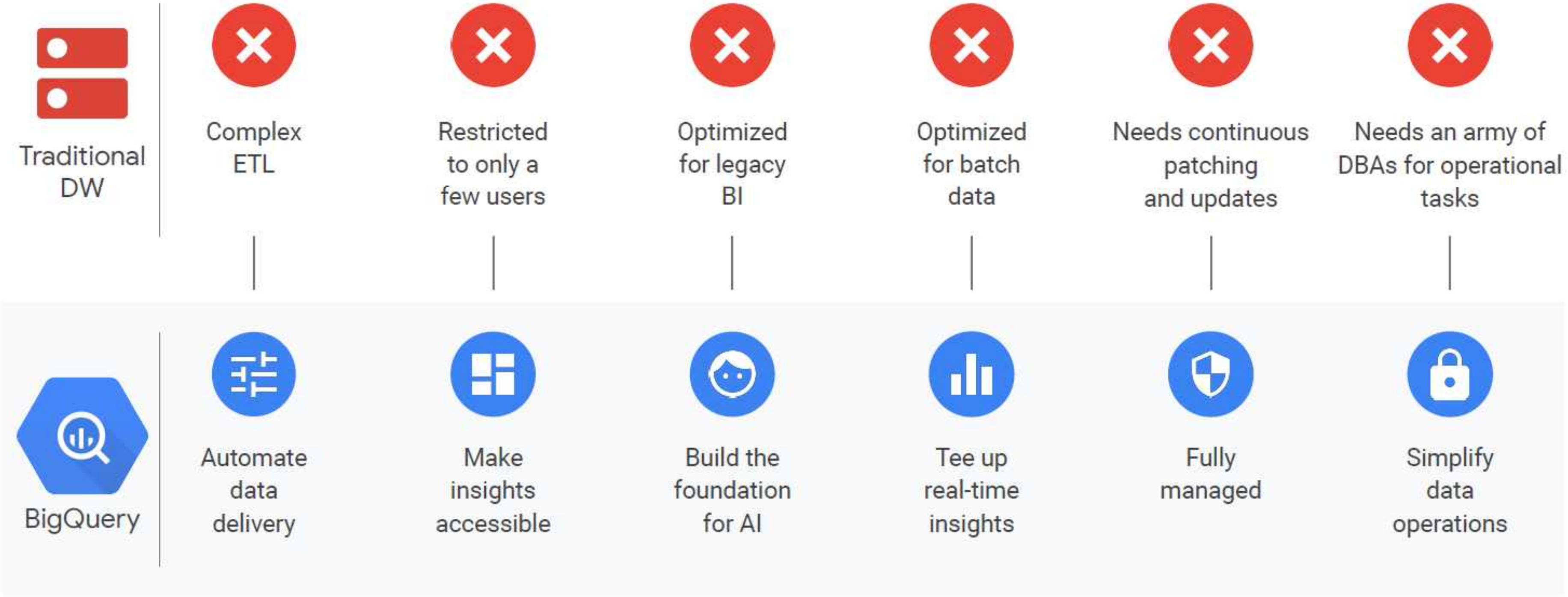
Tables and views

Function the same way as in a traditional data warehouse

Grants

Cloud IAM grants permission to perform specific actions

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



Traditional DW	✗ Complex ETL	✗ Restricted to only a few users	✗ Optimized for legacy BI	✗ Optimized for batch data	✗ Needs continuous patching and updates	✗ Needs an army of DBAs for operational tasks
BigQuery	Automate data delivery	Make insights accessible	Build the foundation for AI	Tee up real-time insights	Fully managed	Simplify data operations

Cloud SQL is fully managed SQL Server, Postgres, or MySQL for your Relational Database (transactional RDBMS)

- Automatic encryption
- 30TB storage capacity
- 60,000 IOPS (read/write per second)
- Auto-scale and auto backup

Why not simply use Cloud SQL for reporting workflows?

RDBMS are optimized for data from a single source and high-throughput writes vs high-read data warehouses

Cloud SQL

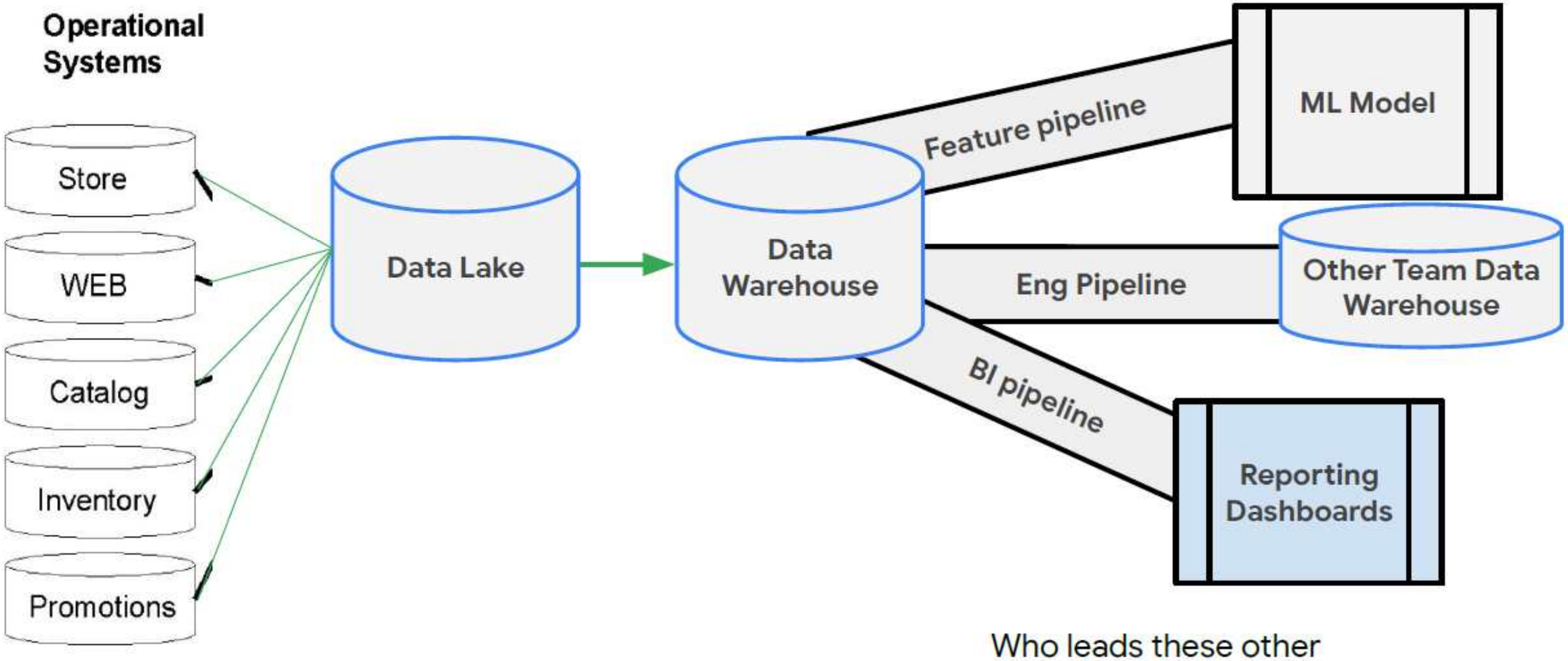
- Scales to GB and TB
- Ideal for back-end database applications
- Record based storage

You will likely need and encounter both a database and data warehouse in your final architecture

BigQuery

- Scales to PB
- Easily connect to external data sources for ingestion
- Column based storage

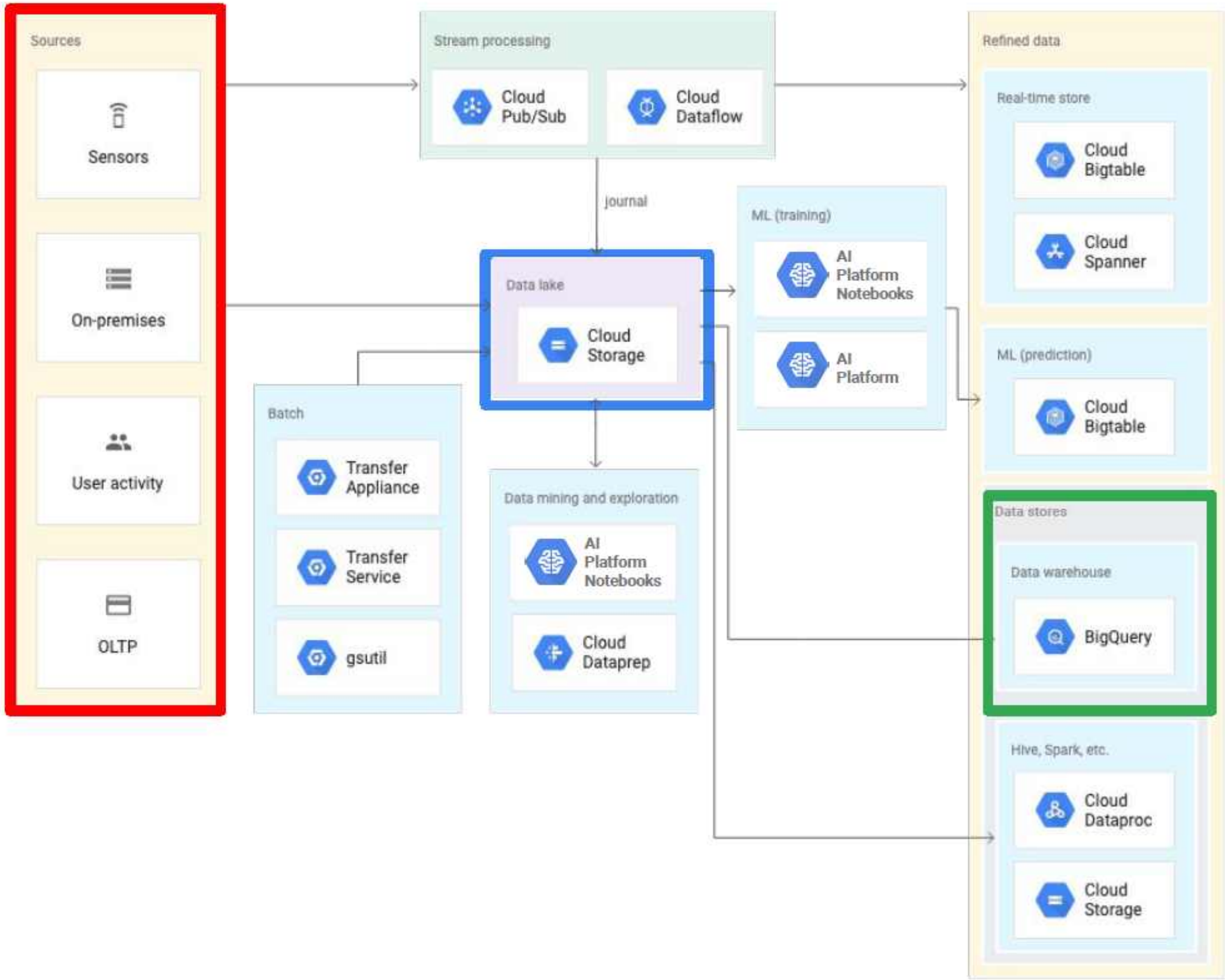
The complete picture: Source data comes into the data lake, is processed into the data warehouse and made available for insights



Who leads these other teams that we will have to partner with?

Concept Review:

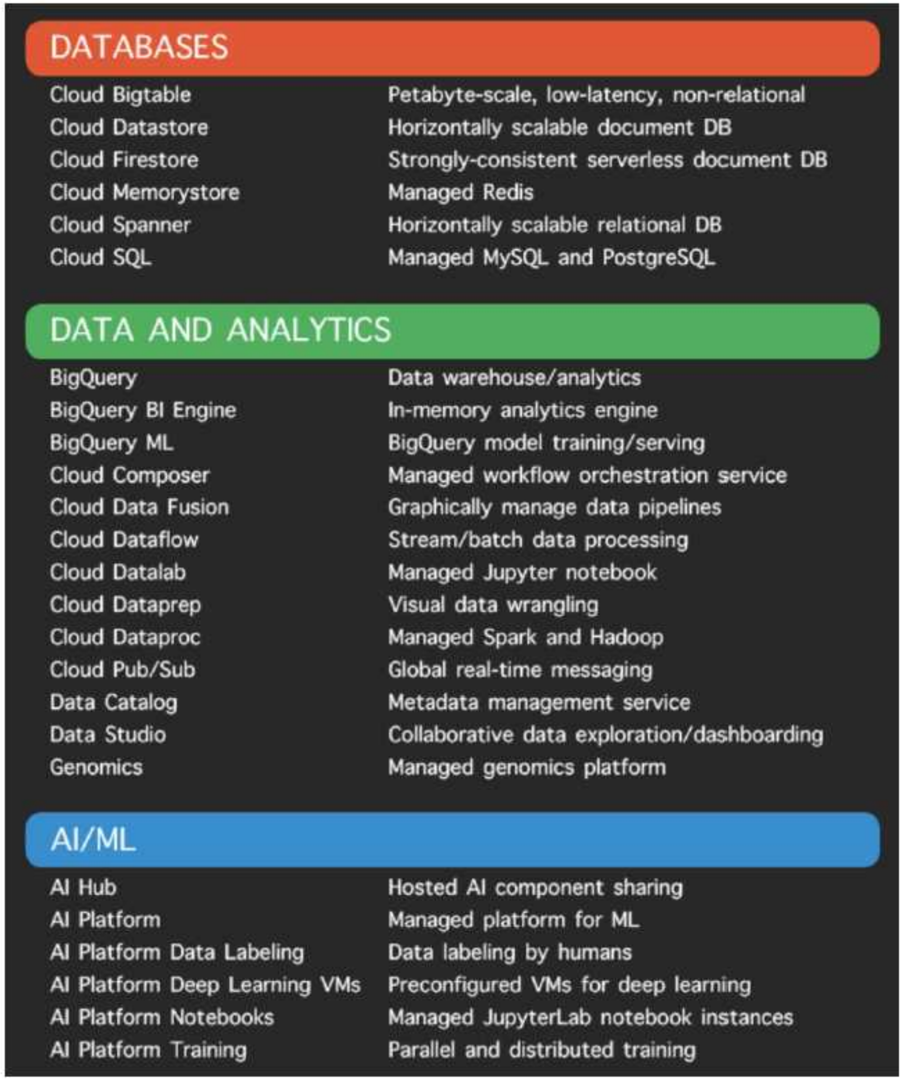
Data sources feed into a Data Lake and are processed into your Data Warehouse for analysis



Here’s a useful guide for “GCP products in 4 words or less”

https://github.com/gregsramblings/google-cloud-4-words

Updated continually By Greg Wilson - Google DevRel



DATABASES

Cloud Bigtable	Petabyte-scale, low-latency, non-relational
Cloud Datastore	Horizontally scalable document DB
Cloud Firestore	Strongly-consistent serverless document DB
Cloud Memorystore	Managed Redis
Cloud Spanner	Horizontally scalable relational DB
Cloud SQL	Managed MySQL and PostgreSQL

DATA AND ANALYTICS

BigQuery	Data warehouse/analytics
BigQuery BI Engine	In-memory analytics engine
BigQuery ML	BigQuery model training/serving
Cloud Composer	Managed workflow orchestration service
Cloud Data Fusion	Graphically manage data pipelines
Cloud Dataflow	Stream/batch data processing
Cloud Datalab	Managed Jupyter notebook
Cloud Dataprep	Visual data wrangling
Cloud Dataproc	Managed Spark and Hadoop
Cloud Pub/Sub	Global real-time messaging
Data Catalog	Metadata management service
Data Studio	Collaborative data exploration/dashboarding
Genomics	Managed genomics platform

AI/ML

AI Hub	Hosted AI component sharing
AI Platform	Managed platform for ML
AI Platform Data Labeling	Data labeling by humans
AI Platform Deep Learning VMs	Preconfigured VMs for deep learning
AI Platform Notebooks	Managed JupyterLab notebook instances
AI Platform Training	Parallel and distributed training

Lab

Using BigQuery to do Analysis

Objectives

- Execute interactive queries in the BigQuery console
- Combine and run analytics on multiple datasets