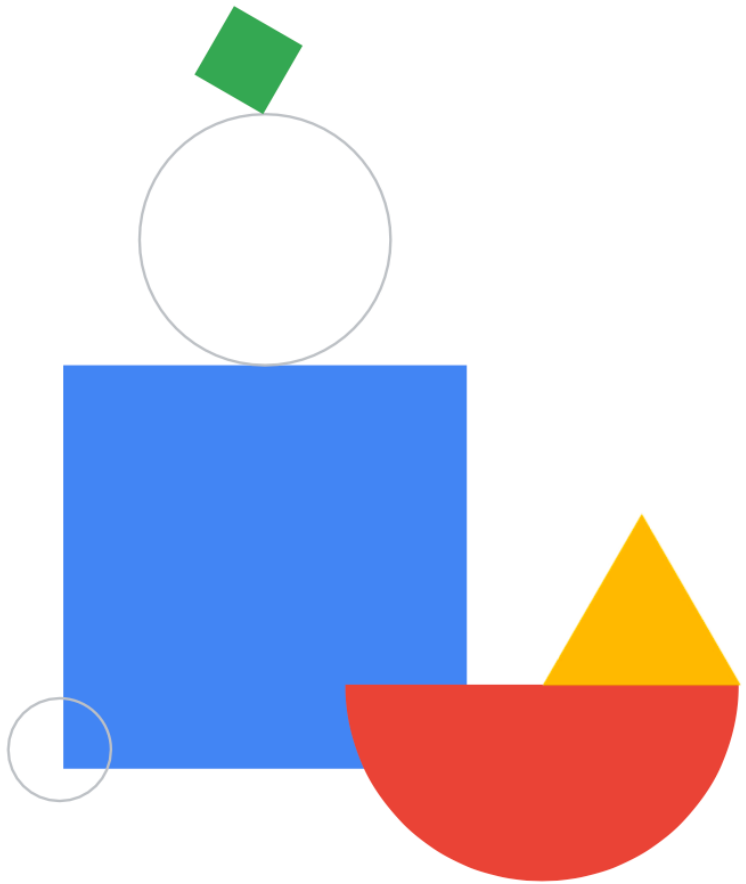
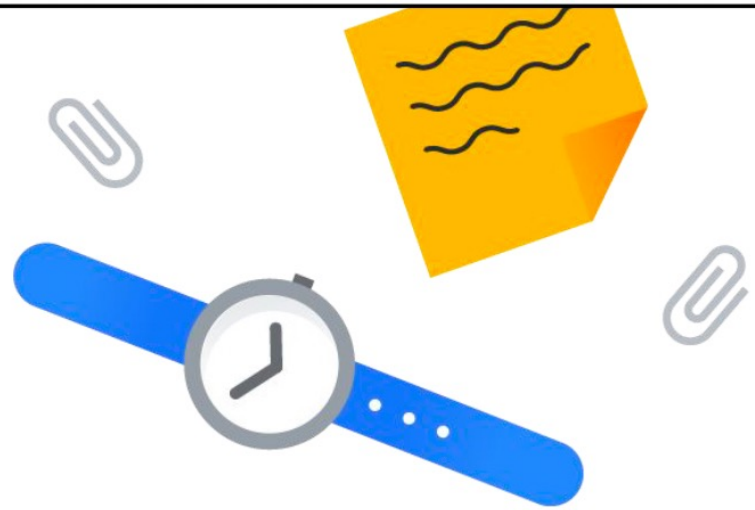
Google Cloud

Introduction to Data Engineering





Module agenda

01 The Role of a Data Engineer

02 Data Engineering Challenges

03 Introduction to BigQuery

04 Data Lakes and Data Warehouses

05 Transactional Databases Versus Data Warehouses

06 Partner Effectively with Other Data Teams

07 Manage Data Access and Governance

08 Build Production-ready Pipelines

09 Google Cloud Customer Case Study

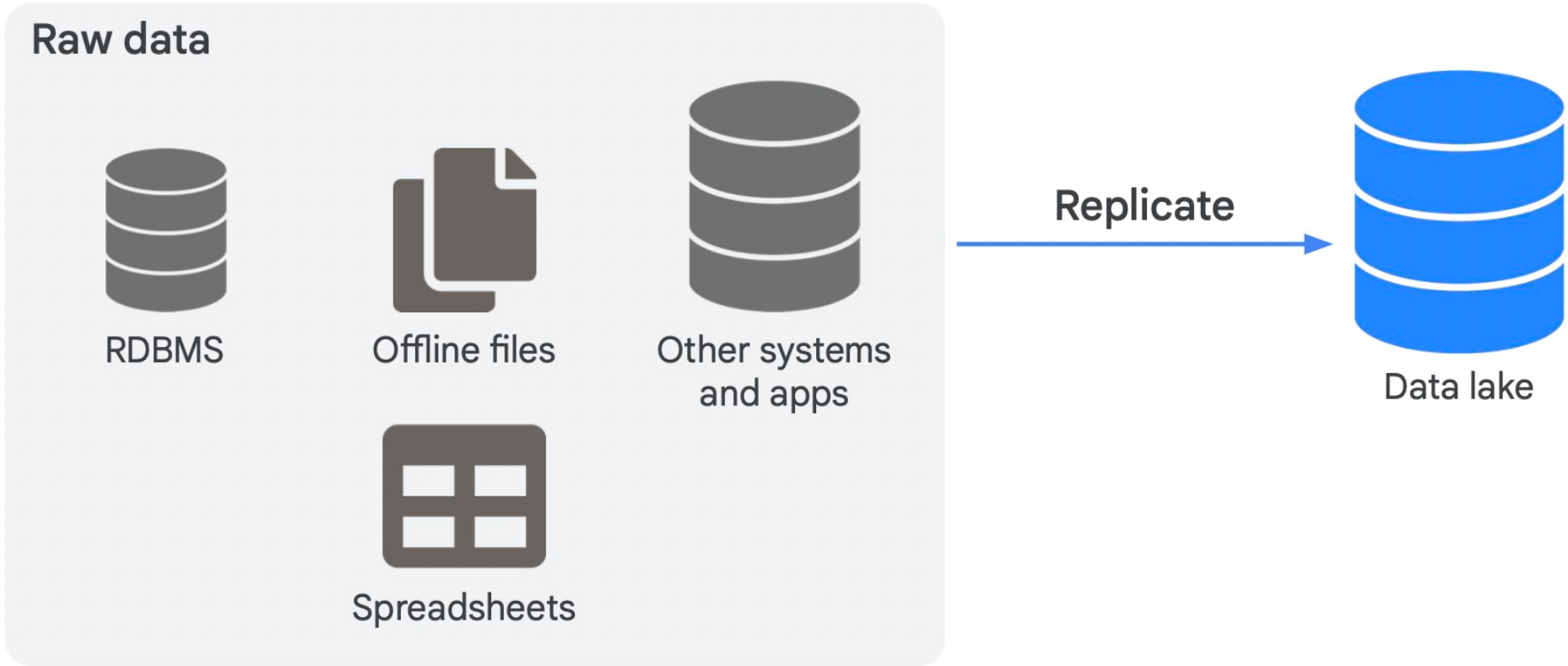
The Role of a Data Engineer

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

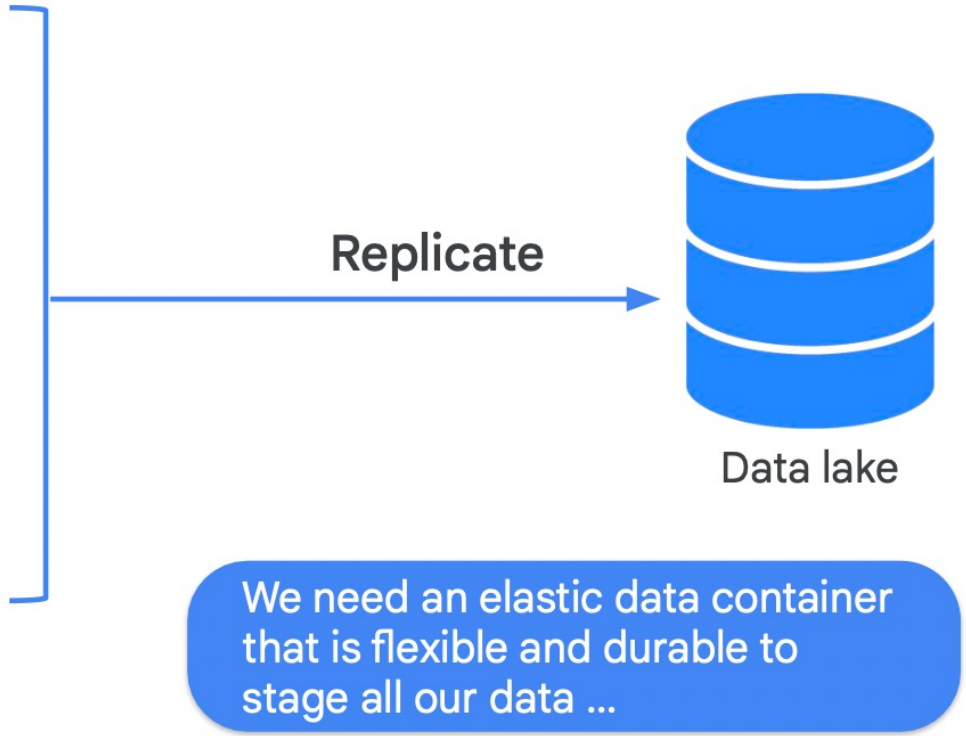
So... how do we get the raw data from multiple systems and where can we store it durably?

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

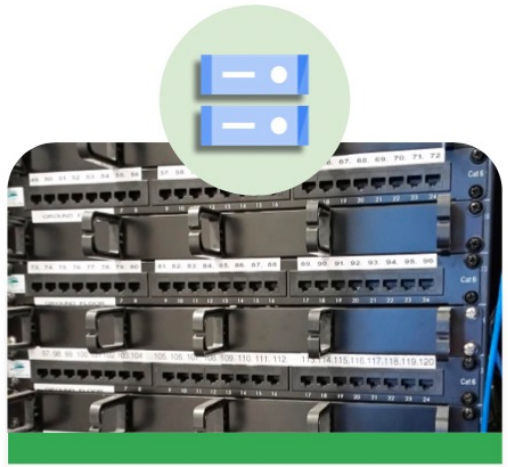


Key considerations when building a data lake

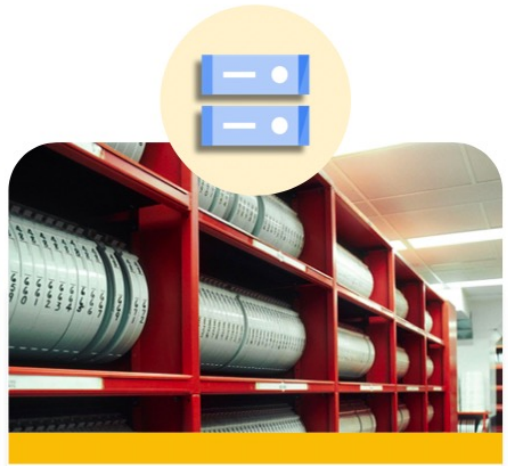
1. Can your data lake handle all the types of data you have?
2. Can it scale to meet the demand?
3. Does it support high-throughput ingestion?
4. Is there fine-grained access control to objects?
5. Can other tools connect easily?



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, **T**ransform, and **L**oad

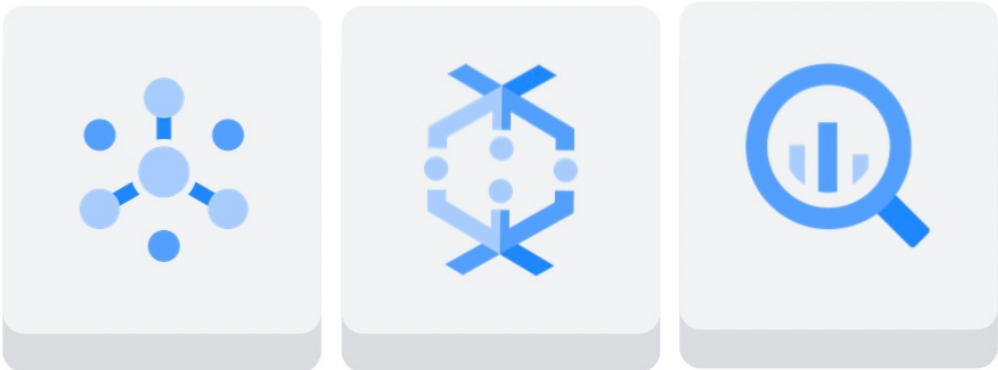
Data processing

Dataproc

Dataflow

What if your data arrives continuously and endlessly?

Streaming data processing



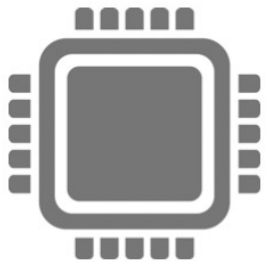
Pub/Sub Dataflow BigQuery

Data Engineering Challenges

Common challenges encountered by data engineers

Access to data

Data accuracy and quality



Availability of computational resources

Query performance

Challenge: Consolidating disparate datasets, data formats, and manage access at scale

Access to data

Data accuracy and quality

Availability of computational resources

Query performance

Getting insights across multiple datasets is difficult without a data lake

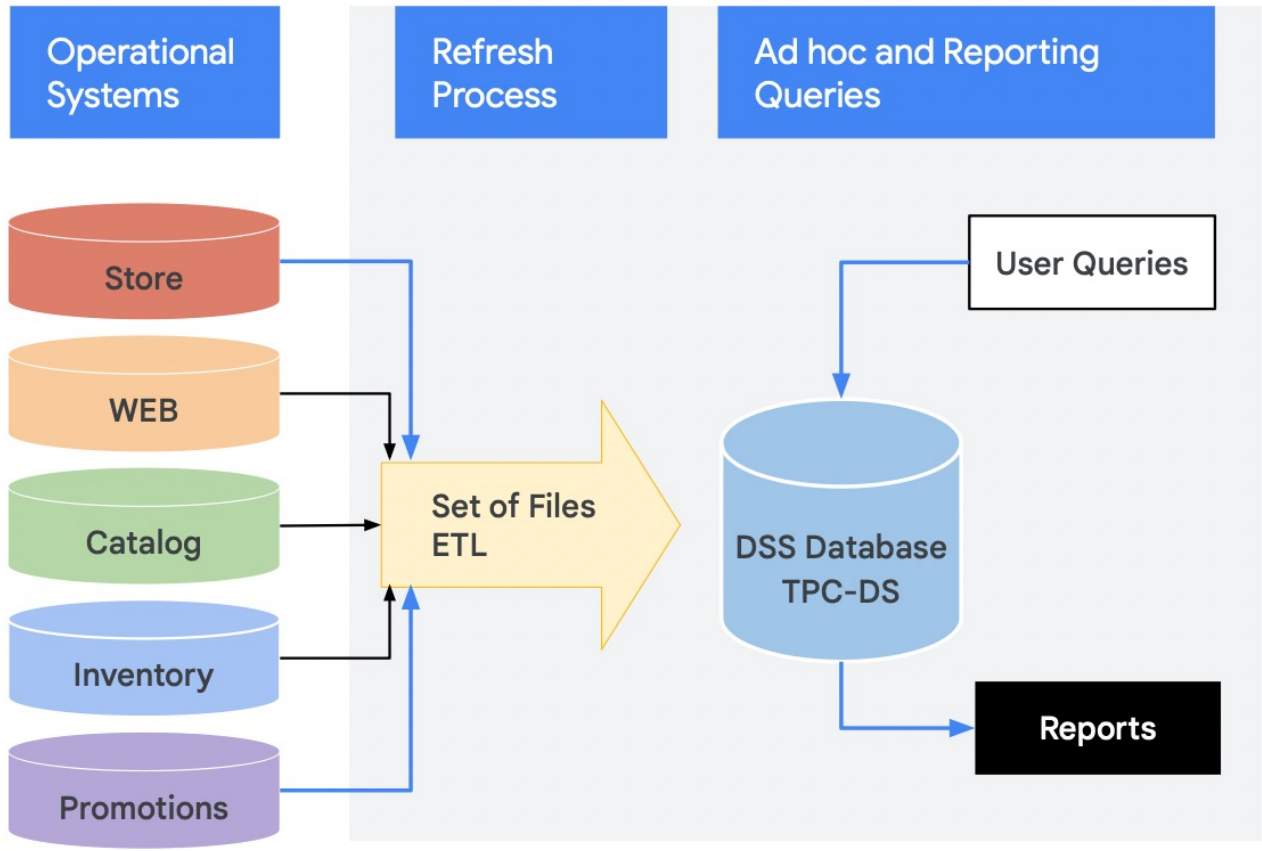
Data is scattered across Google Analytics 360, CRM, and Campaign Manager products, among other sources.

Customer and sales data is stored in a CRM system.

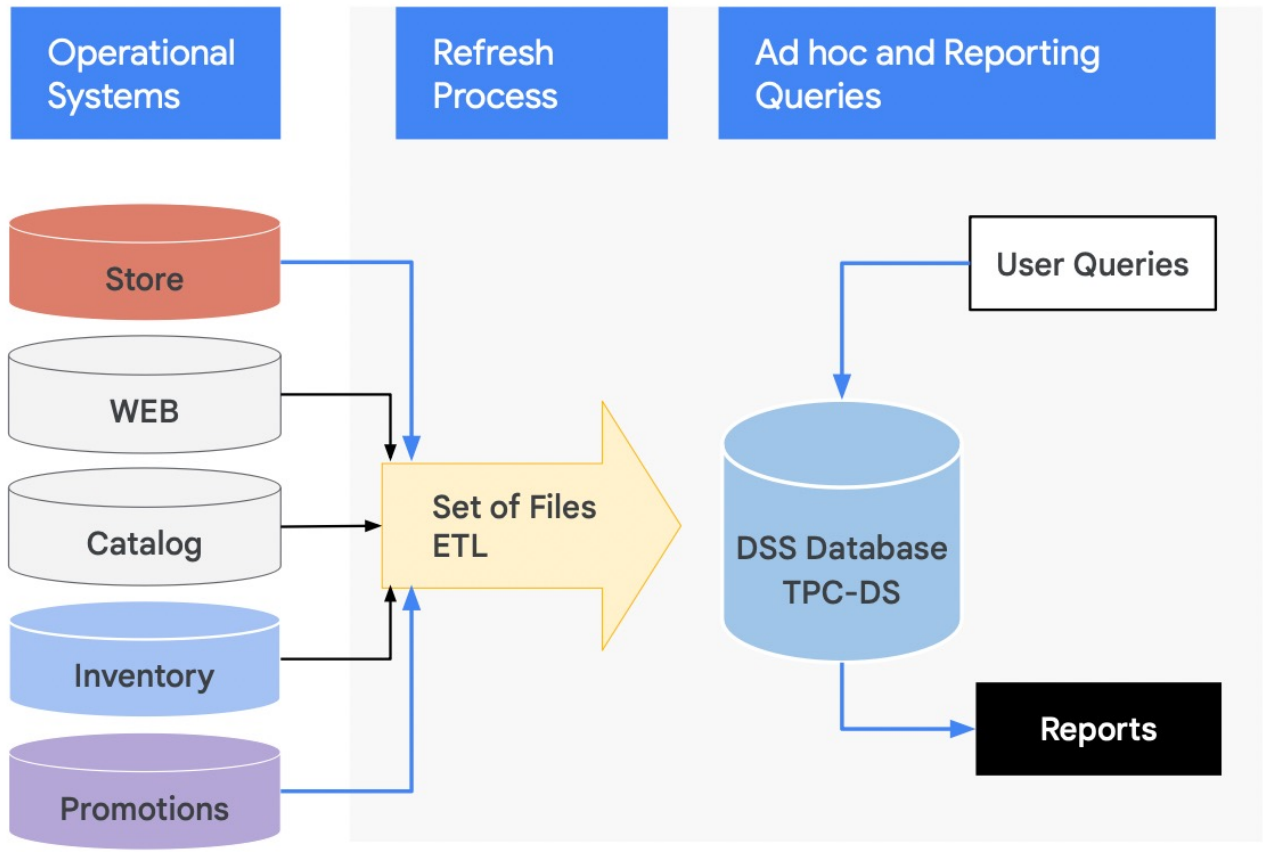
No common tool exists to analyze data and share results with the rest of the organization.

Some data is not in a queryable format.

Data is often siloed in many upstream source systems

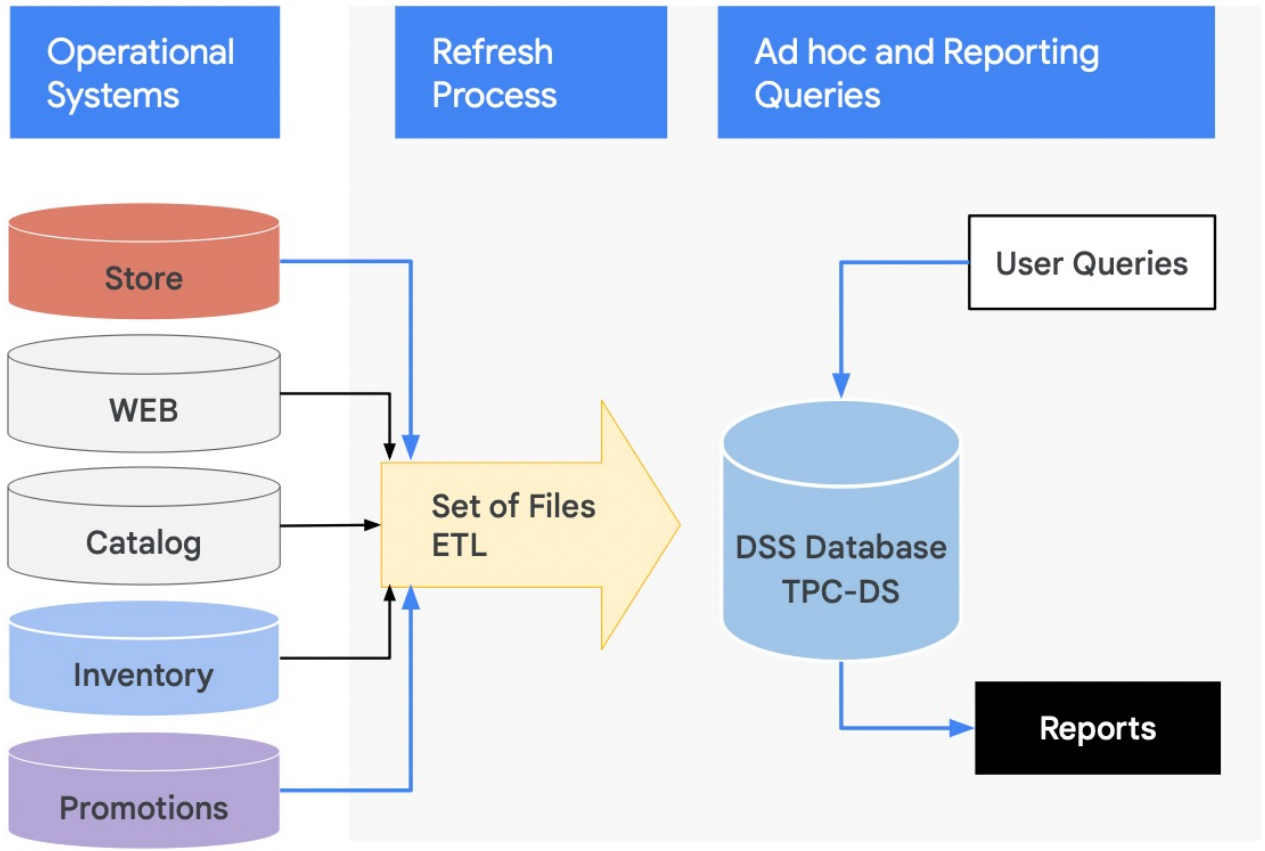


Data is often siloed in many upstream source systems



Data is often siloed in many upstream source systems

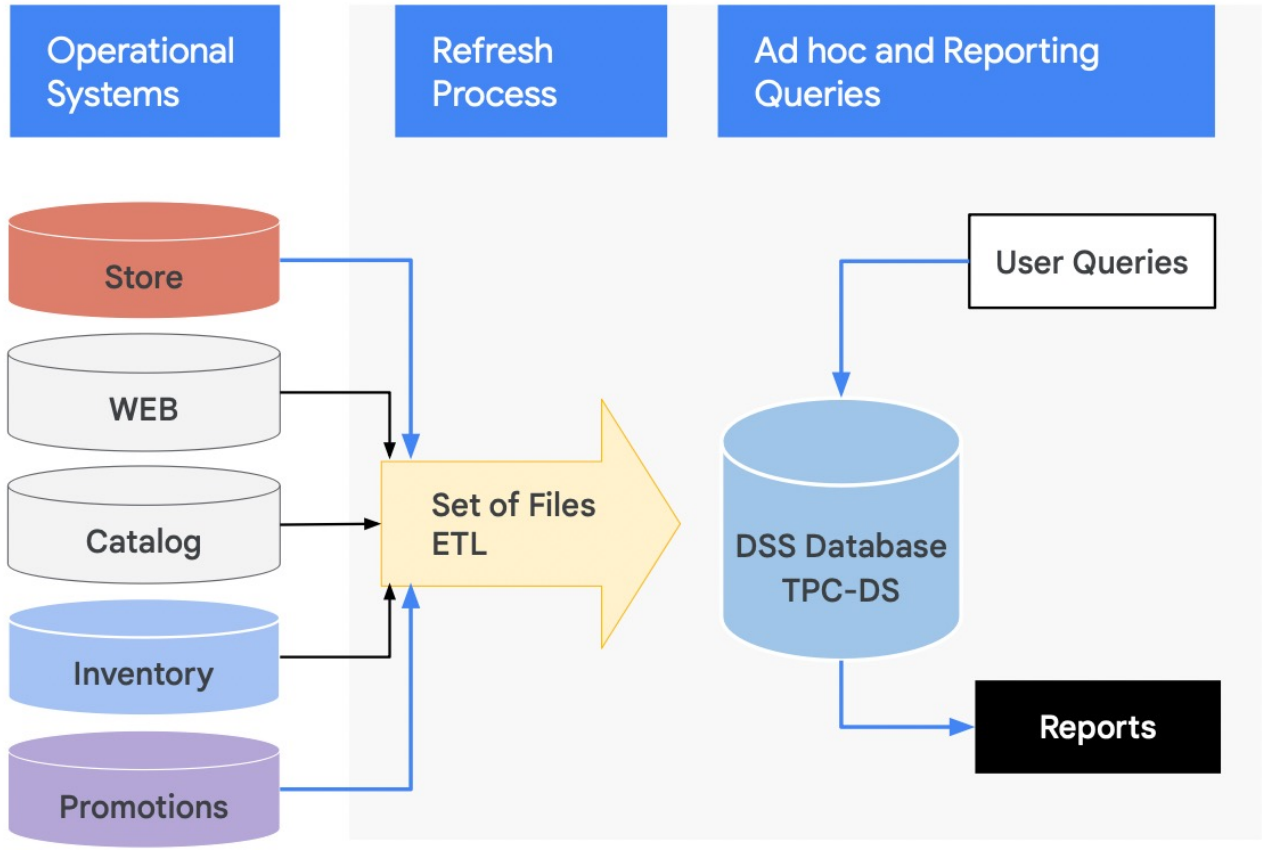
Example query:
Give me all the in-store promotions for recent orders and their **inventory levels.**



Data is often siloed in many upstream source systems

Example query:
Give me all the in-store promotions for recent orders and their **inventory levels.**

Stored in a separate system and restricted access



Challenge: Cleaning, formatting, and getting the data ready for useful business insights in a data warehouse

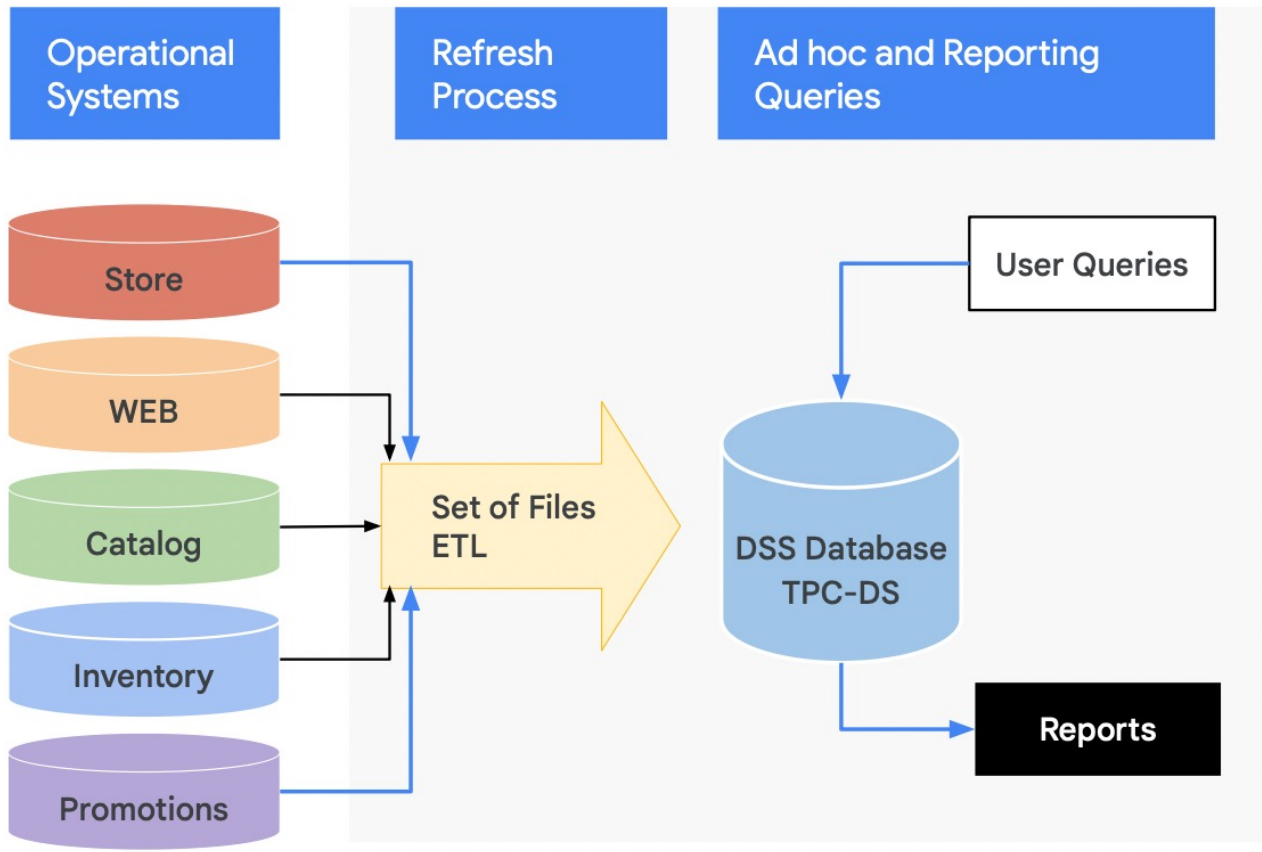
Access to data

Data accuracy and quality

Availability of computational resources

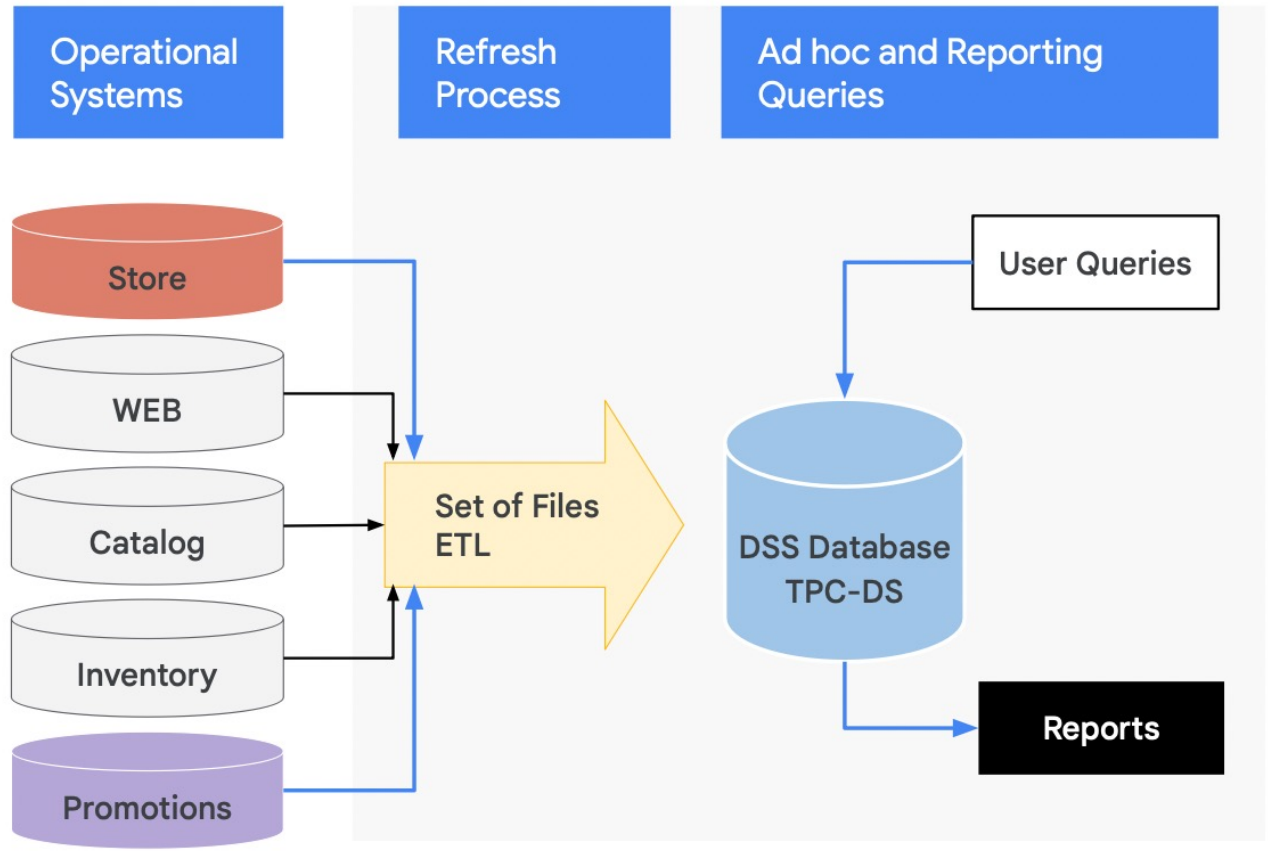
Query performance

Assume that any raw data from source systems needs to be cleaned, transformed, and stored in a data warehouse



Assume that any raw data from source systems needs to be cleaned, transformed, and stored in a data warehouse

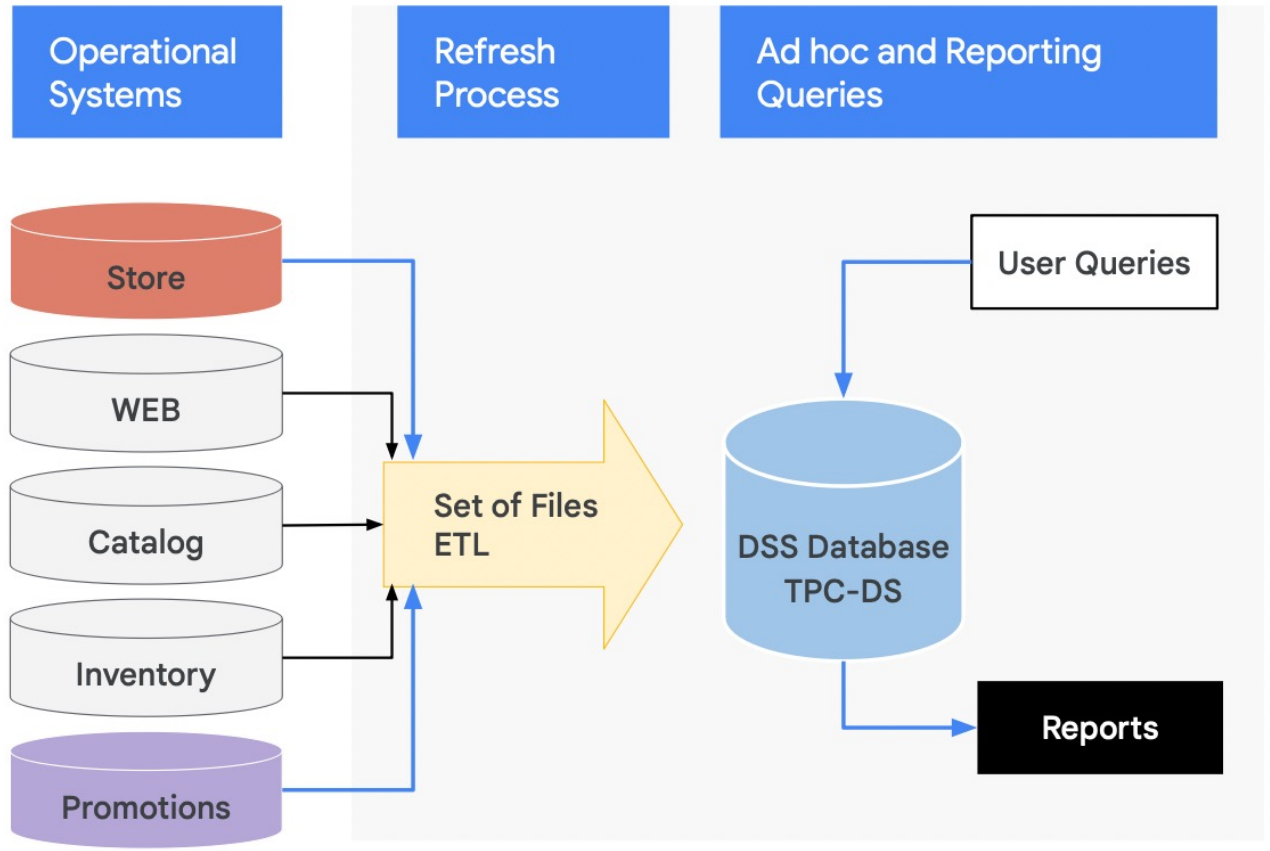
Query:
Give me the best performing in-store promotions in France.



Assume that any raw data from source systems needs to be cleaned, transformed, and stored in a data warehouse

Missing data and all timestamps stored as UTC

Query:
Give me the best performing in-store promotions in France.

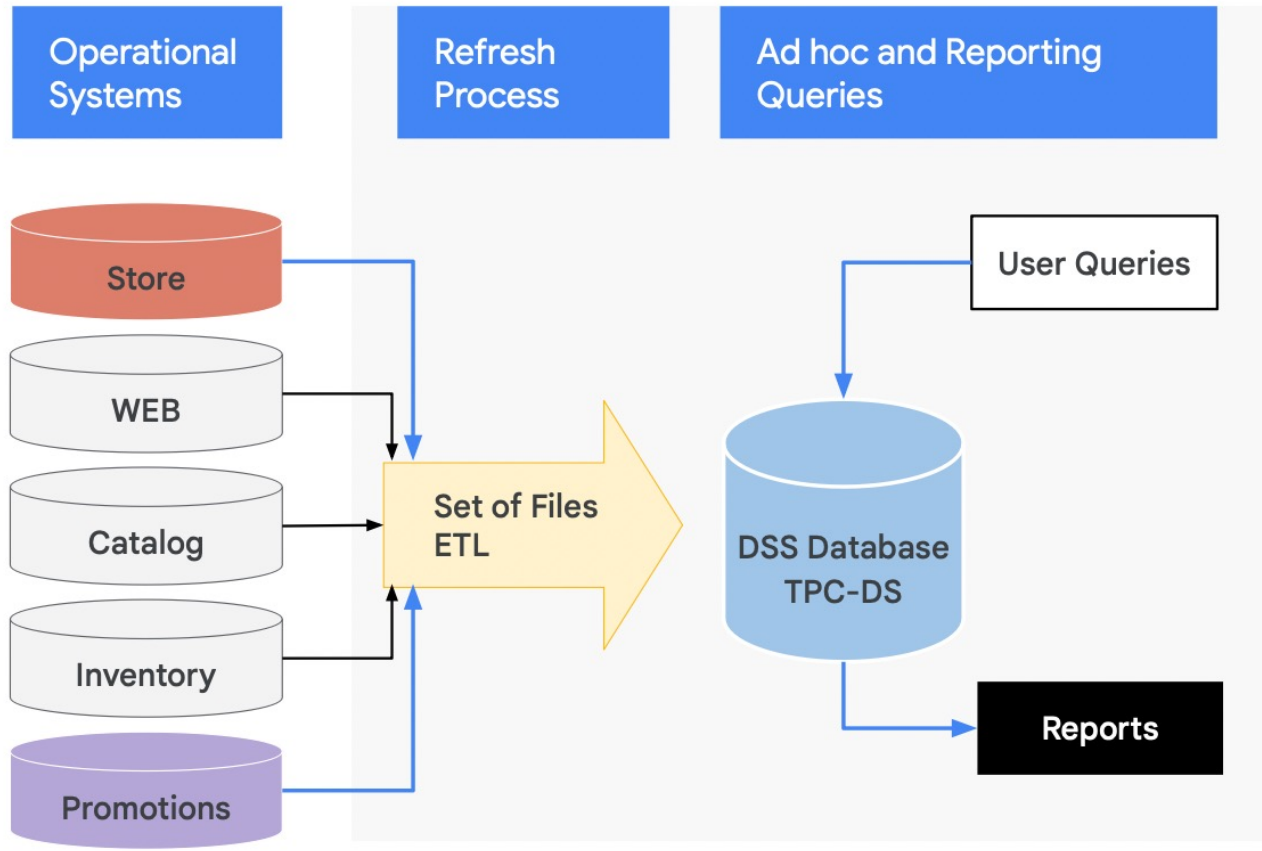


Assume that any raw data from source systems needs to be cleaned, transformed, and stored in a data warehouse

Missing data and all timestamps stored as UTC

Query:
Give me the best performing in-store promotions in France.

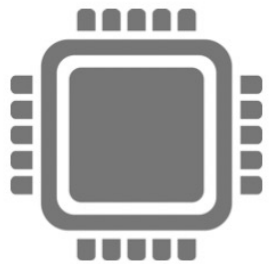
Promotion list stored as .csv files or manual spreadsheets



Challenge: Ensuring you have the compute capacity to meet peak-demand for your team

Access to data

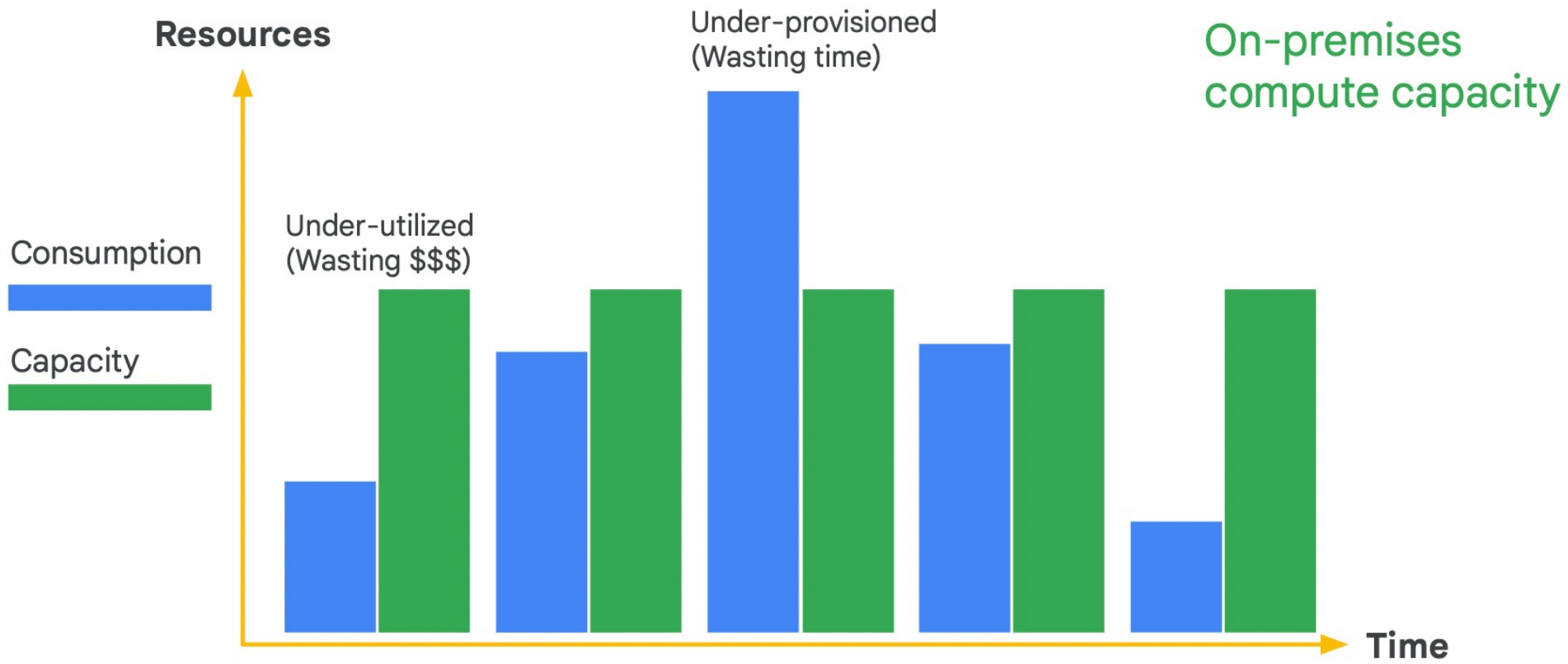
Data accuracy and quality



Availability of computational resources

Query performance

Challenge: Data Engineers need to manage server and cluster capacity if using on-premise



Challenge: Queries need to be optimized for performance (caching, parallel execution)

Access to data

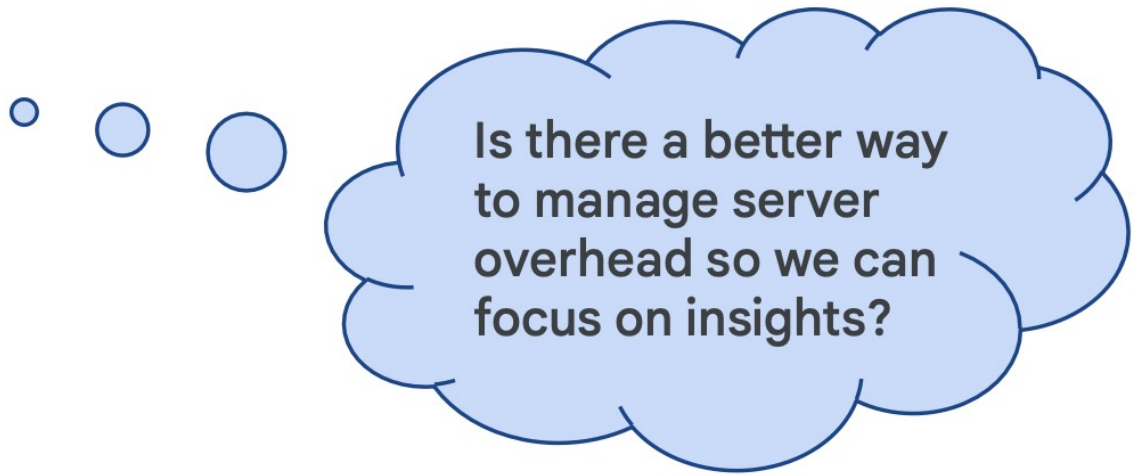
Data accuracy and quality

Availability of computational resources

Query performance

Challenge: Managing query performance on-premise comes with added overhead

- Choosing a query engine.
- Continually patching and updating query engine software.
- Managing clusters and when to re-cluster.
- Optimize for concurrent queries and quota / demand between teams.



Introduction to BigQuery

BigQuery is Google's data warehouse solution

Data warehouse

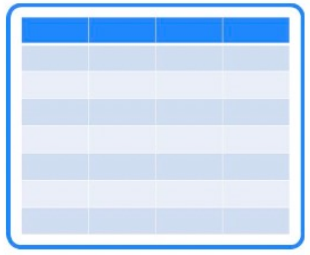
BigQuery replaces a typical data warehouse hardware setup.

Data mart

BigQuery organizes data tables into units called datasets.

Data lake

BigQuery defines schemas and issues queries directly on external data sources.



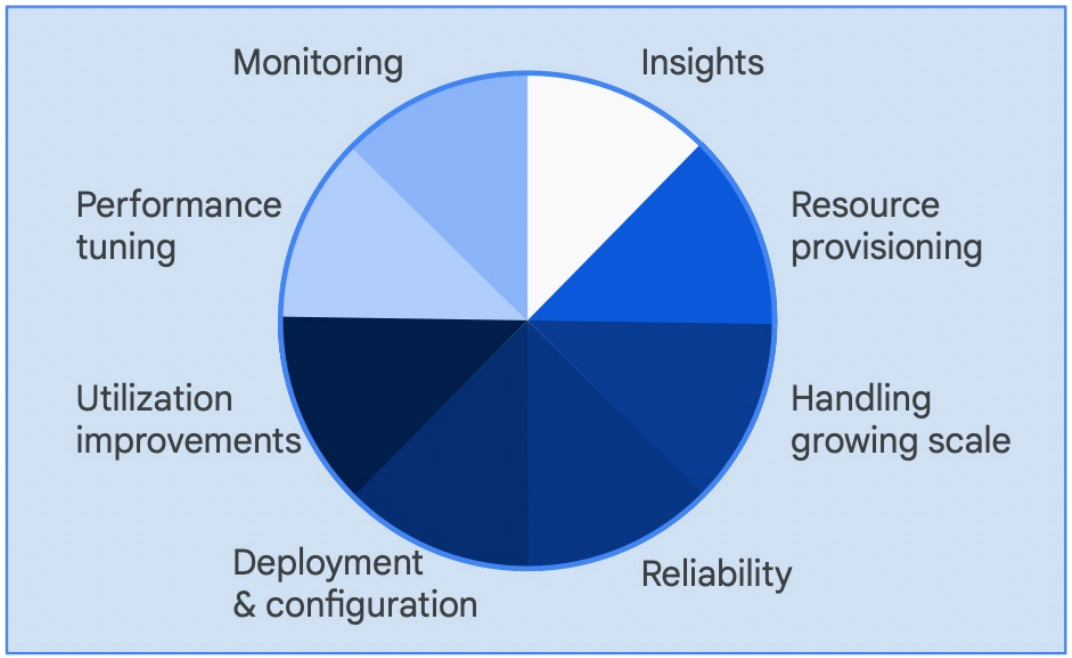
Tables and views

Function the same way as in a traditional data warehouse.

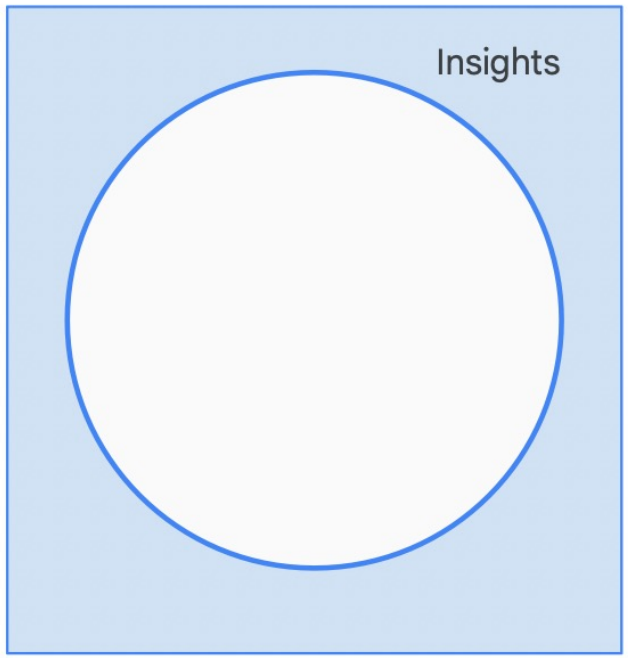
Grants

IAM grants permission to perform specific actions.

Cloud allows data engineers to spend less time managing hardware and enabling scale. Let Google do that for you

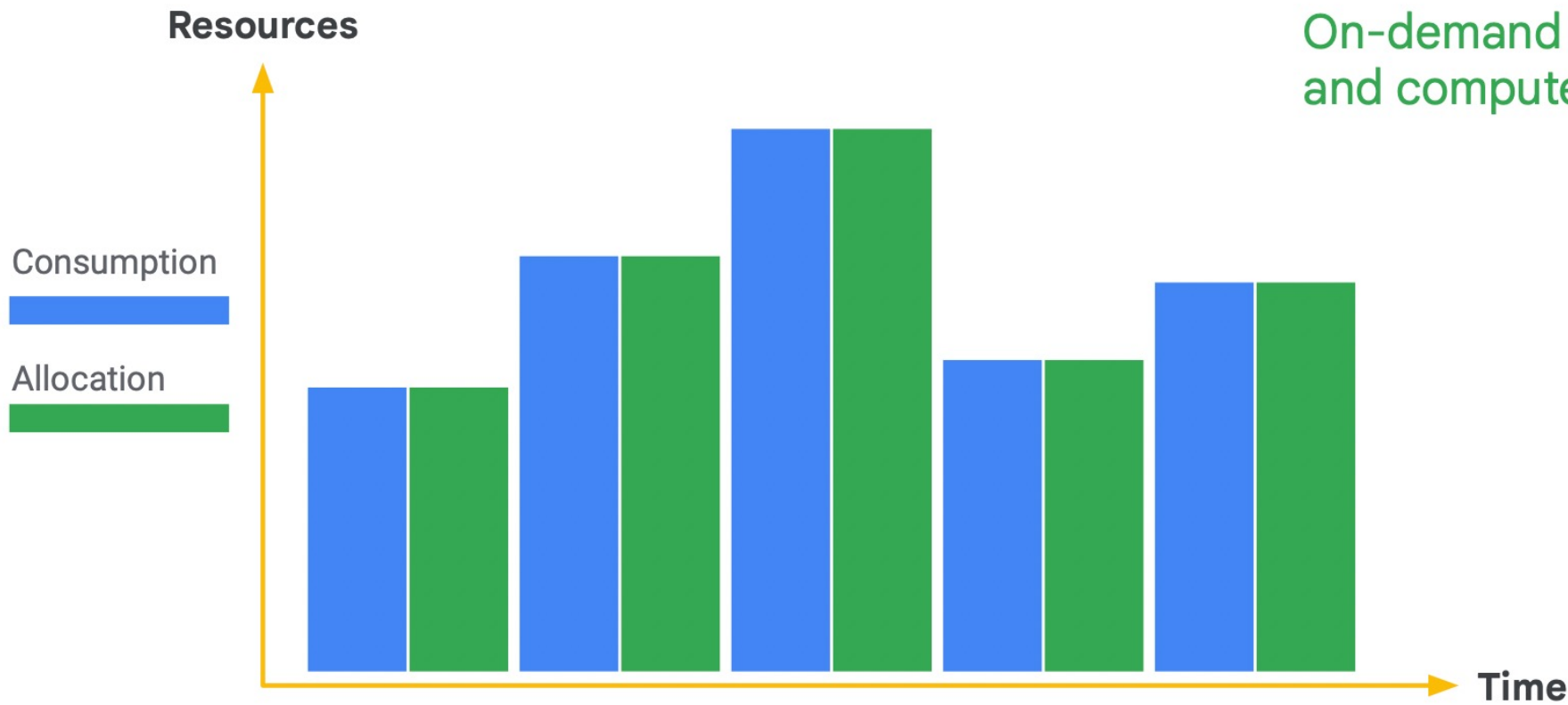


Typical Big Data Processing



With Google

You don't need to provision resources before using BigQuery



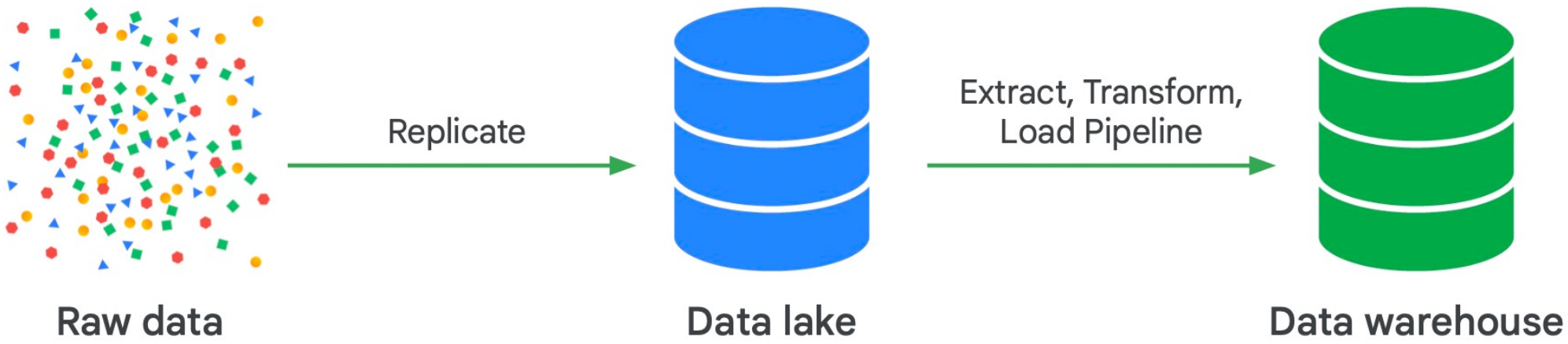
On-demand storage and compute

Data Lakes and Data Warehouses

A data engineer gets data into a useable condition

- 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

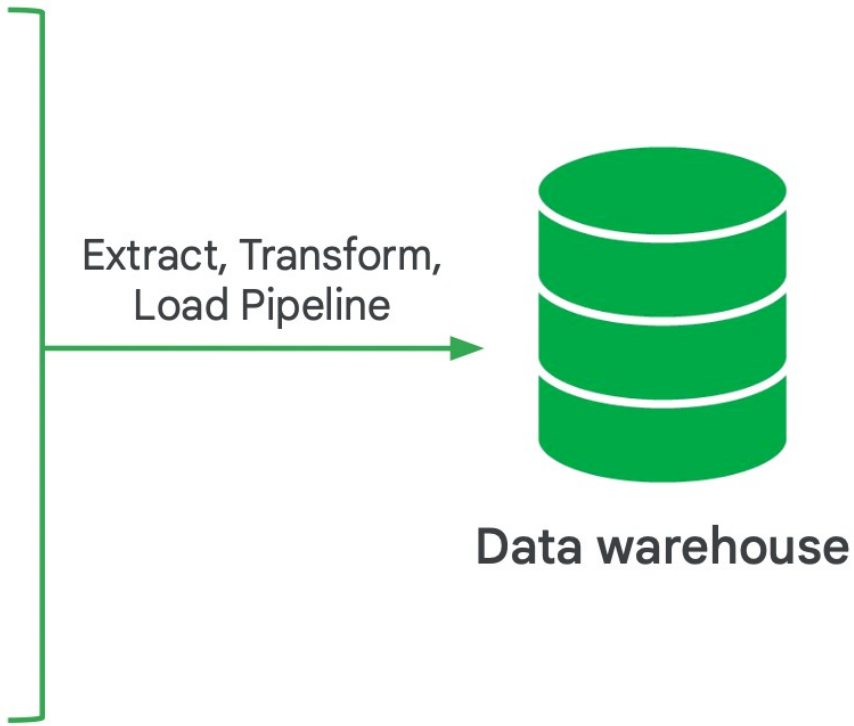
A data warehouse stores transformed data in a usable condition for business insights



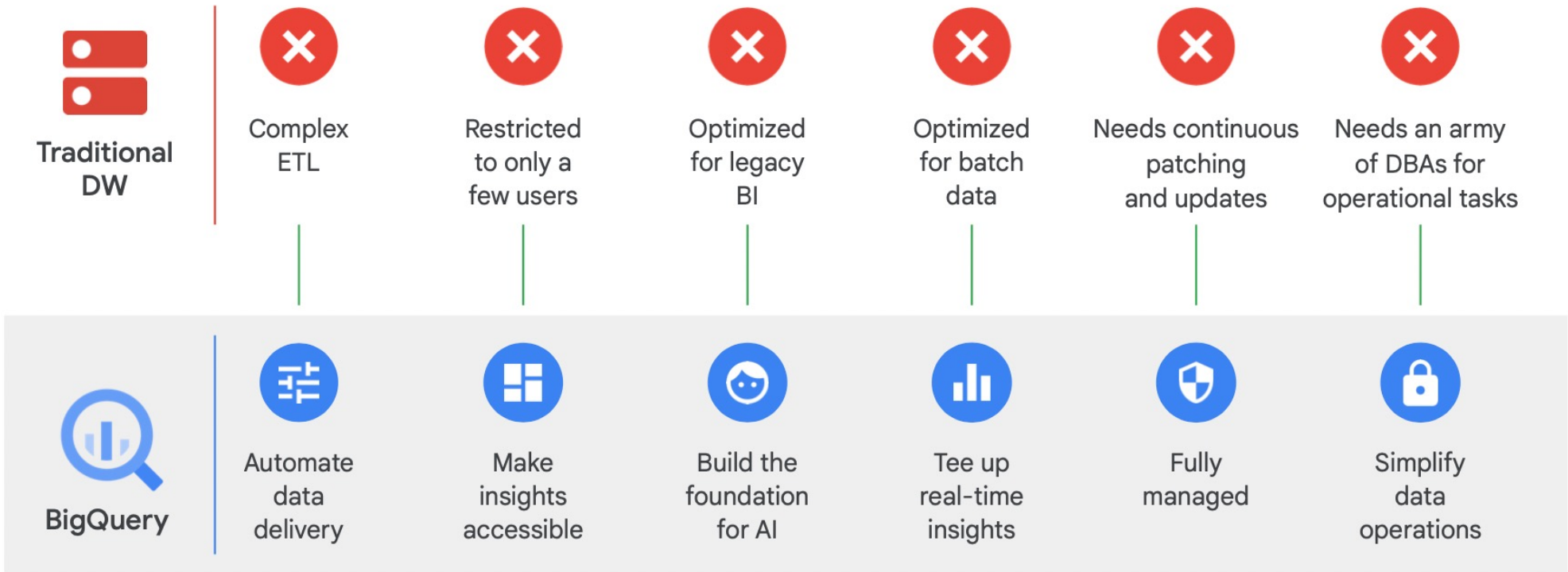
What are the key considerations when deciding between data warehouse options?

Considerations when choosing a data warehouse

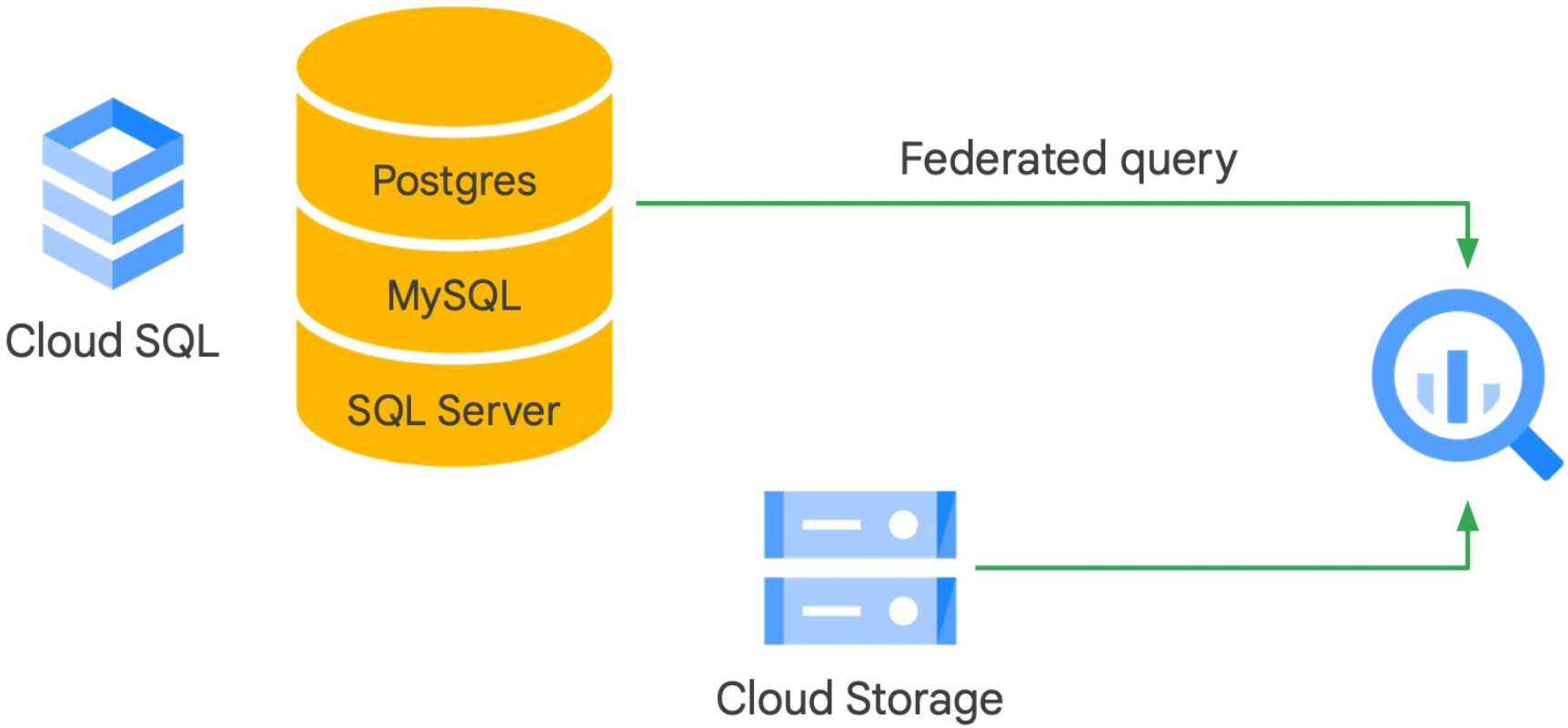
- Can it serve as a sink for both batch and streaming data pipelines?
- Can the data warehouse scale to meet my needs?
- How is the data organized, cataloged, and access controlled?
- Is the warehouse designed for performance?
- What level of maintenance is required by our engineering team?



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



You can simplify Data Warehouse ETL pipelines with external connections to Cloud Storage and Cloud SQL



Transactional Databases Versus Data Warehouses

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

Why not simply use Cloud SQL for reporting workflows?

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

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

Cloud SQL *

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

BigQuery

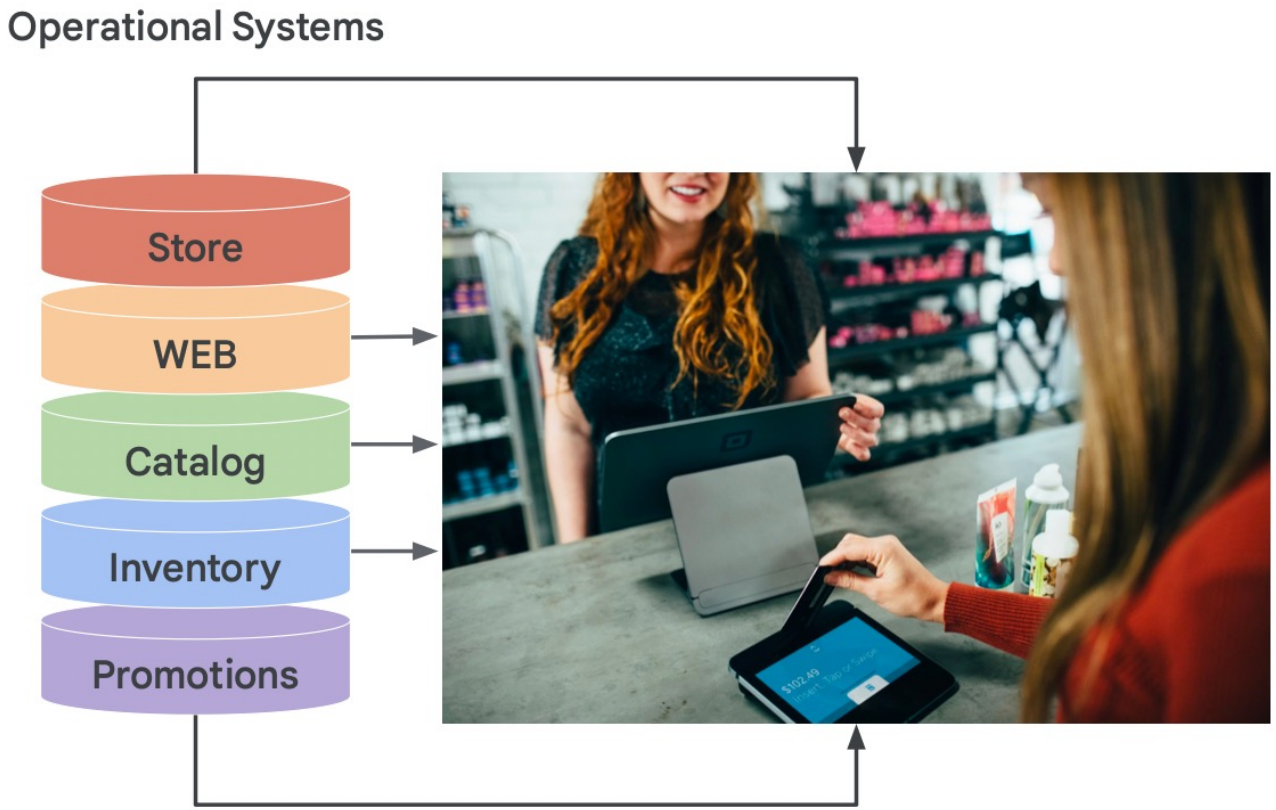
- Scales to PB.
- Easily connect to external data sources for ingestion.
- Column-based storage.

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

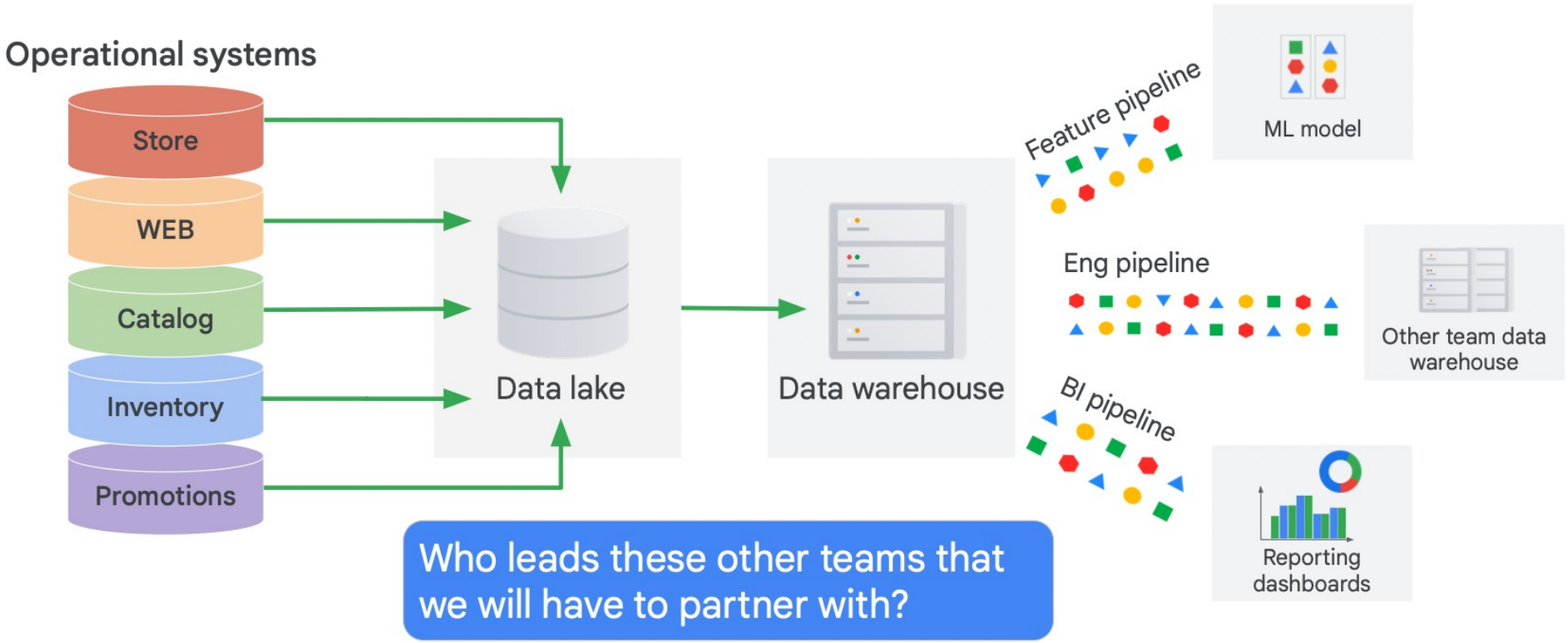
* Cloud SQL is one of several RDBMS options on Google Cloud

Relational database management systems (RDBMS) are critical for managing new transactions

RDBMS are optimized for high throughput **WRITES** to **RECORDS**.



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



Partner Effectively with Other Data Teams

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

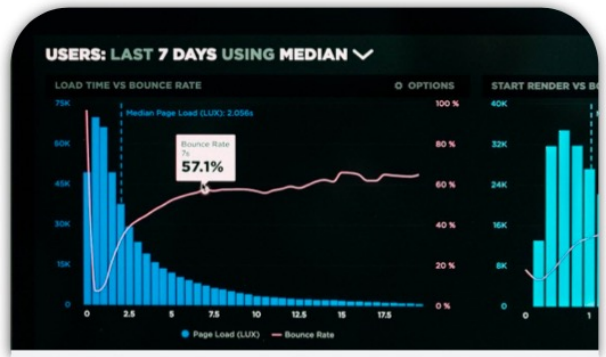
What teams rely on these pipelines?

Many teams rely on partnerships with data engineering to get value out of their data

How might each of these teams rely on data engineering?

ML Engineer

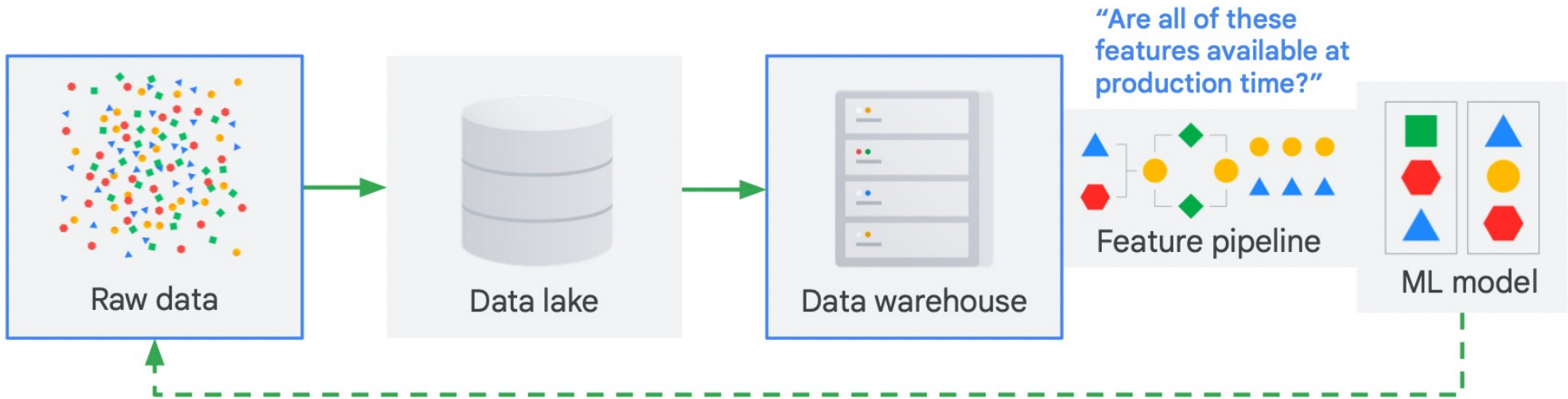
Data Analyst



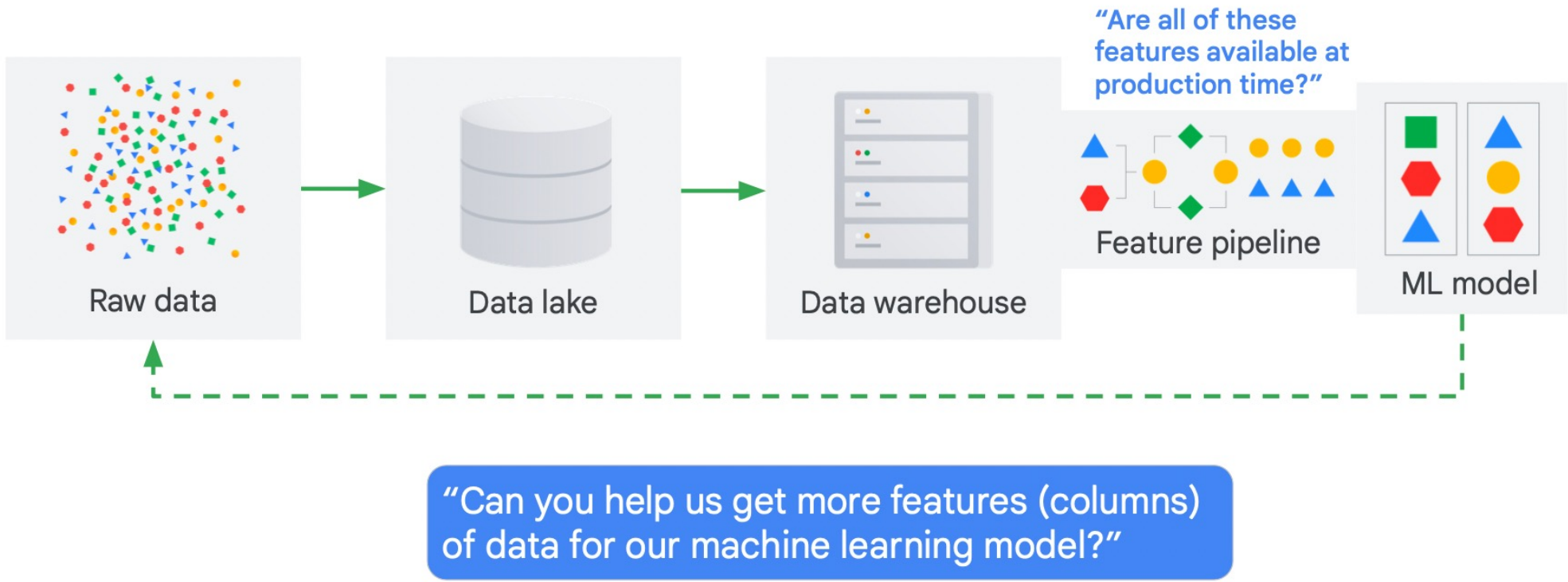
Data Engineer

Machine learning teams need data engineers to help them capture new features in a stable pipeline

Machine learning teams need data engineers to help them capture new features in a stable pipeline



Machine learning teams need data engineers to help them capture new features in a stable pipeline



Add value: Machine learning directly in BigQuery

1 Dataset

2 Create/train

```
CREATE MODEL
`bqml_tutorial.sample_model`
OPTIONS(model_type='logistic_reg') AS
SELECT
```

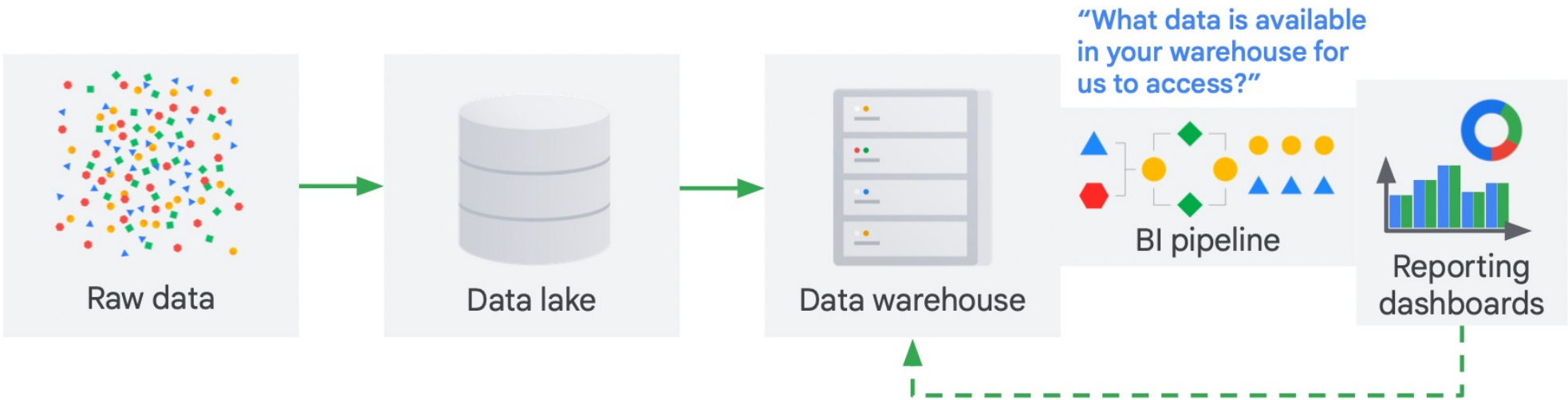
```
FROM
  ML.EVALUATE(MODEL
`bqml_tutorial.sample_model`,
    TABLE eval_table)
```

3 Evaluate

4 Predict/classify

```
FROM
    ML.PREDICT(MODEL
`bqml_tutorial.sample_model`,
      table game_to_predict) )
AS predict
```

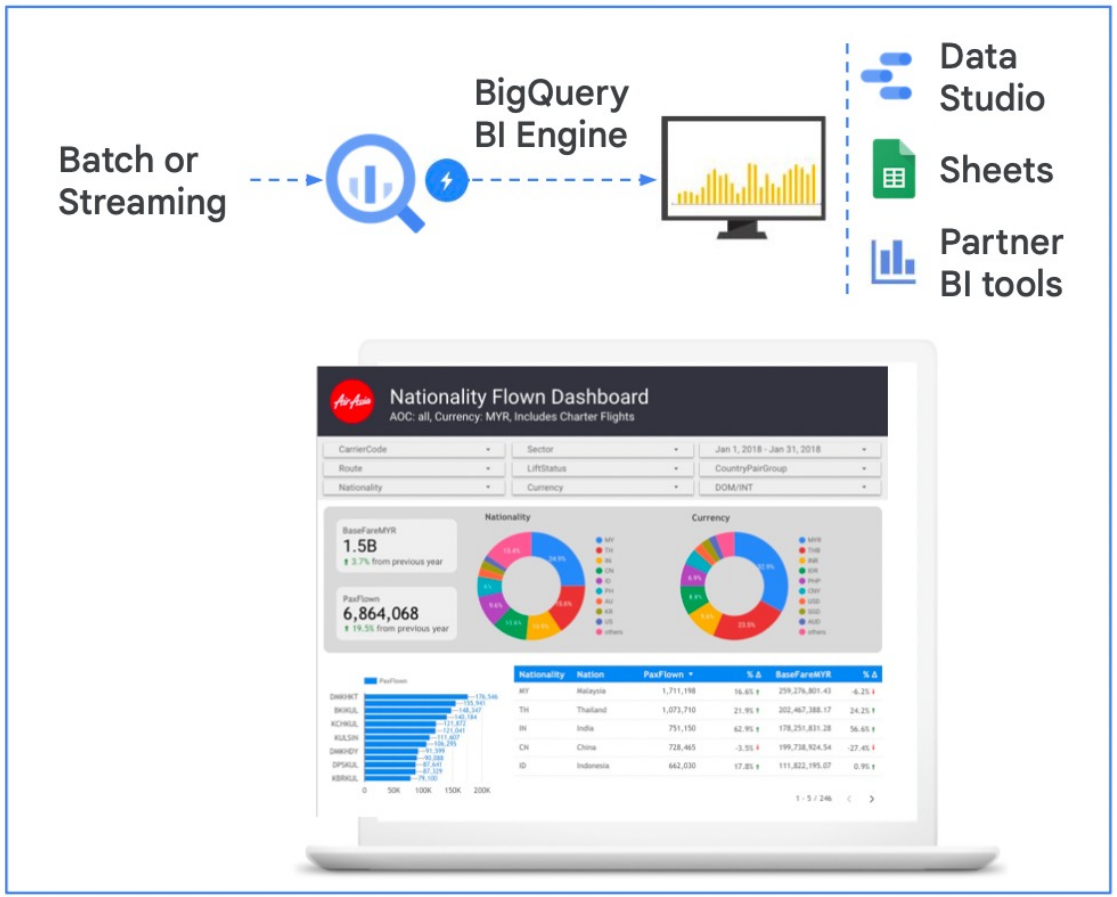
Data analysis and business intelligence teams rely on data engineering to showcase the latest insights



"Our dashboards are slow, can you help us re-engineer our BI tables for better performance?"

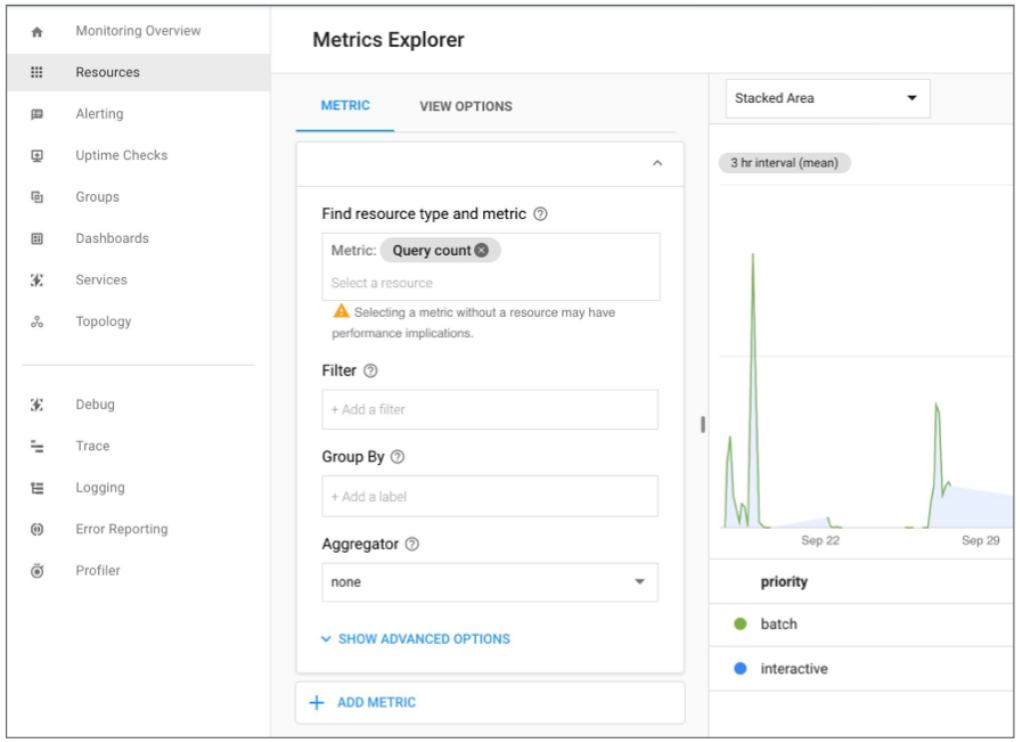
Add value: BI Engine for dashboard performance

- No need to manage OLAP cubes or separate BI servers for dashboard performance.
- Natively integrates with BigQuery streaming for real-time data refresh.
- Column oriented in-memory BI execution engine.



Other data engineering teams may rely on your pipelines being timely and error free

Add value: Cloud Monitoring for performance



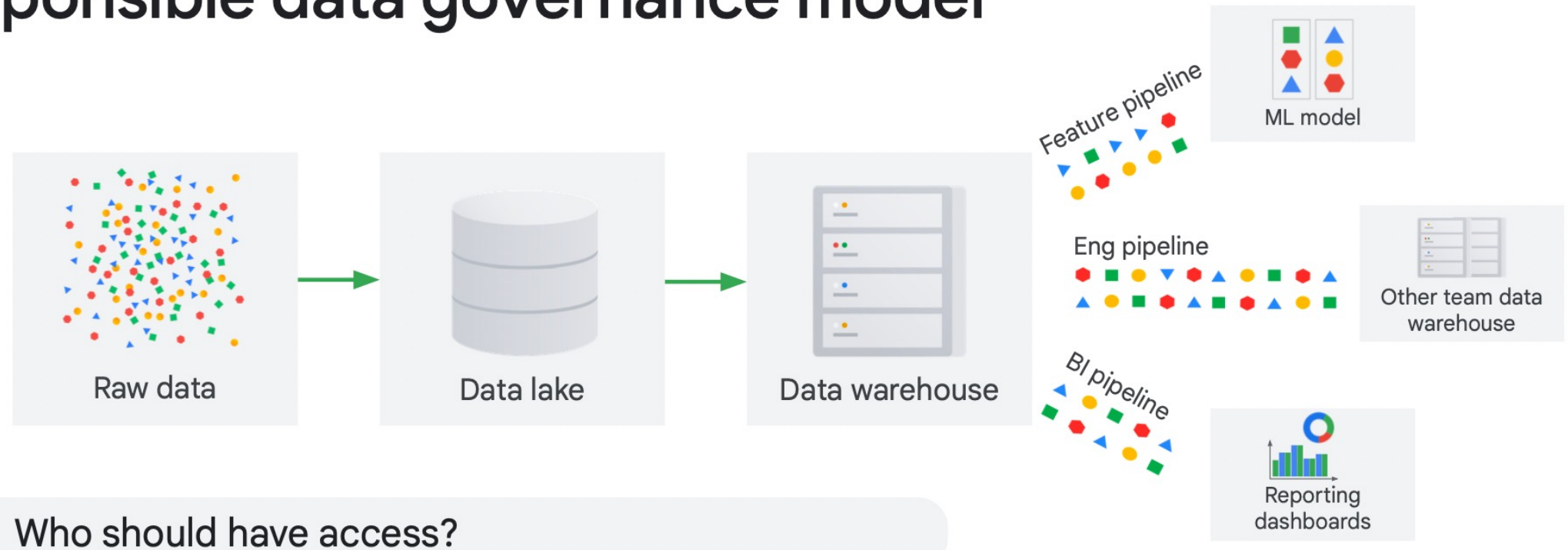
- View in-flight and completed queries.
- Create alerts and send notifications.
- Track spending on BigQuery resources.
- Use Cloud Audit Logs to view actual job information (who executed, what query was ran).

Manage Data Access and Governance

A data engineer manages data access and governance

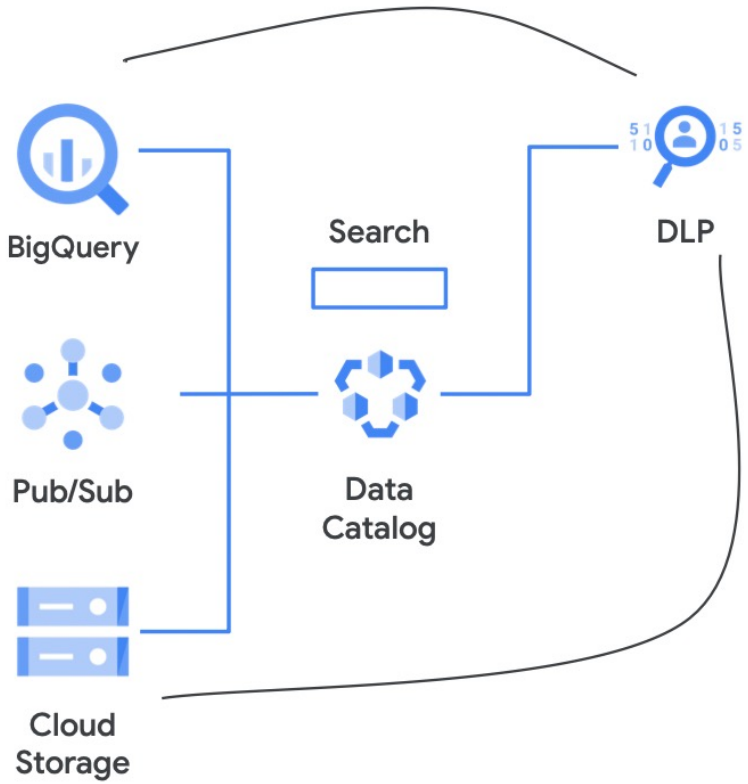
- Get the data to where it can be useful
- Get the data into a usable condition
- Add new value to the data
- **Manage the data**
- Productionize data processes

Data engineering must set and communicate a responsible data governance model



- Who should have access?
- How is PII handled?
- How can we educate end-users on our data catalog?

Cloud Data Catalog is a managed data discovery + Data Loss Prevention API for guarding PII



Data Catalog

Simplify data discovery at any scale:
Fully managed metadata management service with no infrastructure to set up or manage.

Unified view of all datasets:
Central and secure data catalog across Google Cloud with metadata capture and tagging.

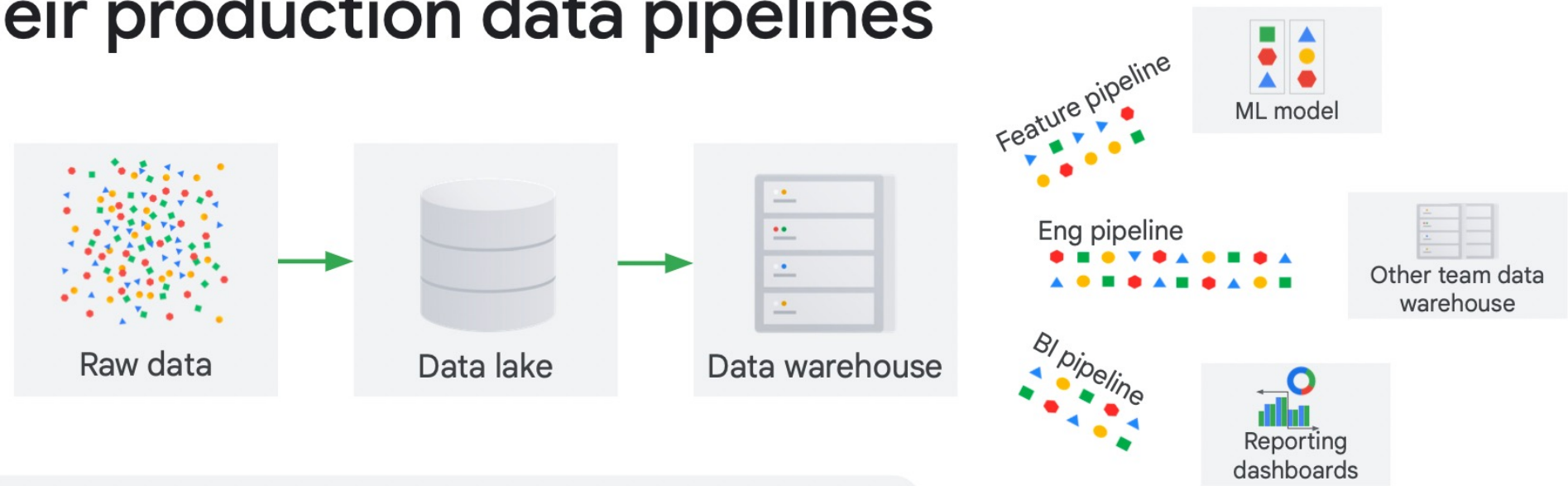
Data governance foundation:
Security compliance with access level controls along with Cloud Data Loss Prevention integration for handling sensitive data.

Build Production-ready Pipelines

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

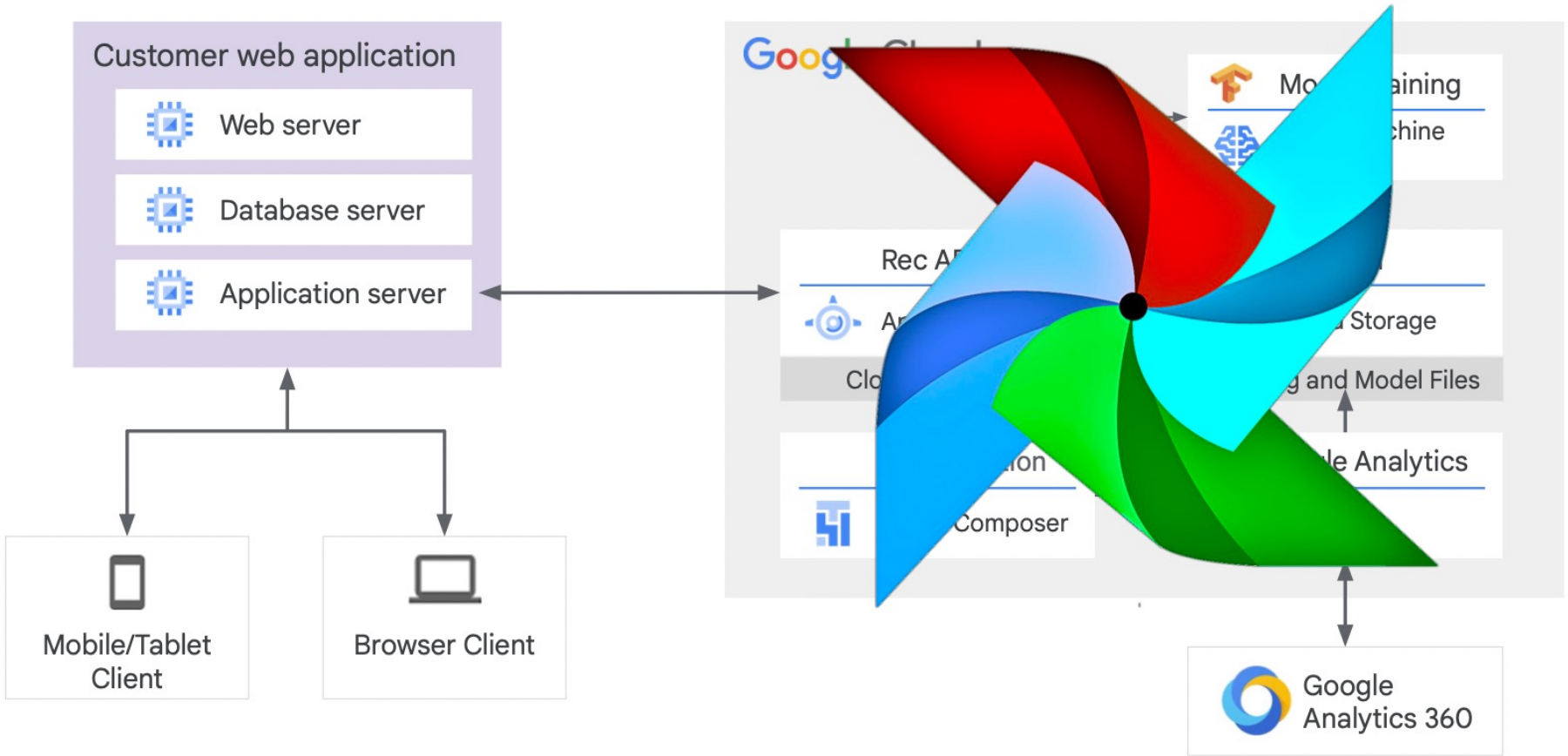
- Get the data to where it can be useful
- Get the data into a usable condition
- Add new value to the data
- Manage the data
- **Productionize data processes**

Data engineering owns the health and future of their production data pipelines

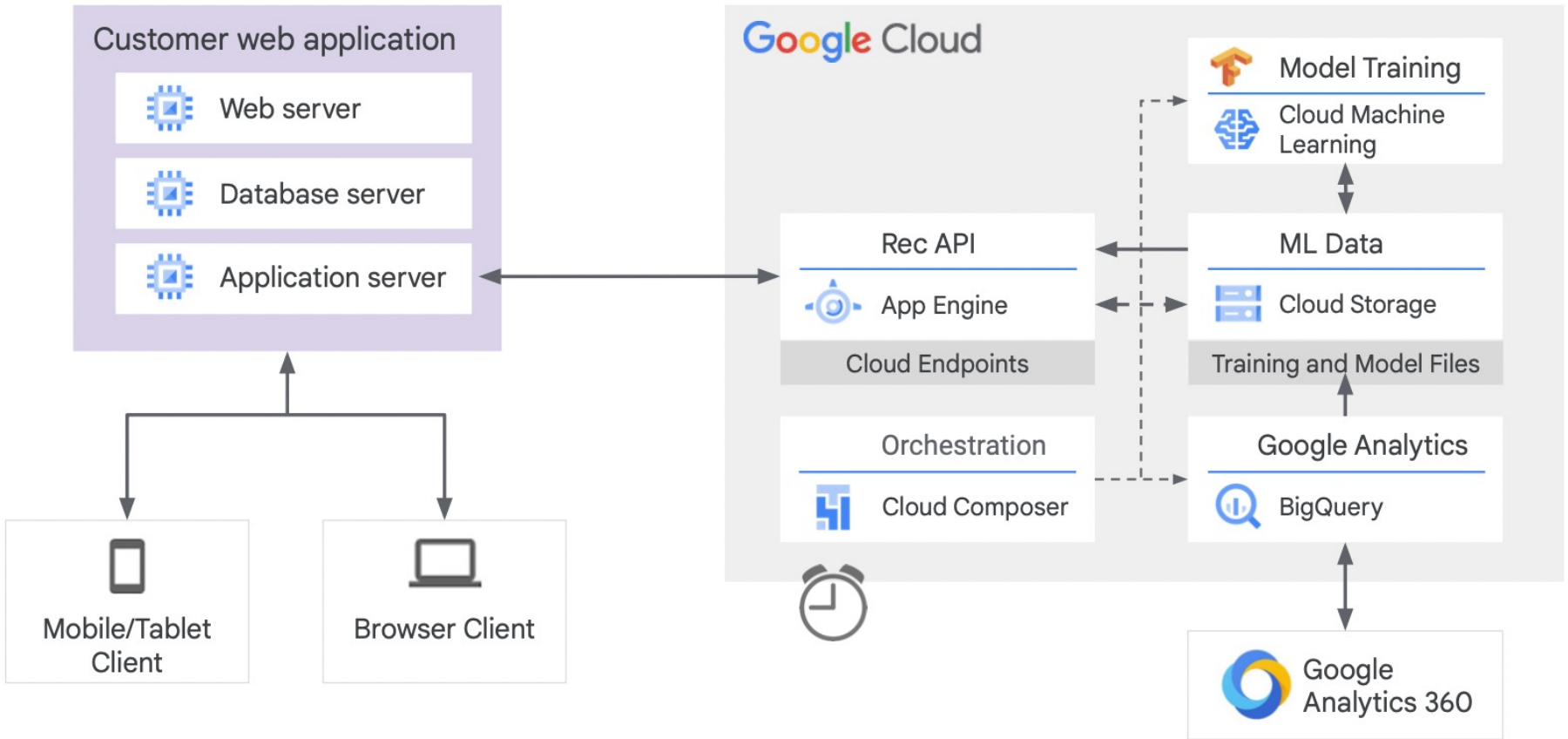


- How can we ensure pipeline health and data cleanliness?
- How do we productionalize these pipelines to minimize maintenance and maximize uptime?
- How do we respond and adapt to changing schemas and business needs?
- Are we using the latest data engineering tools and best practices?

Cloud Composer (managed Apache Airflow) is used to orchestrate production workflows



Cloud Composer (managed Apache Airflow) is used to orchestrate production workflows

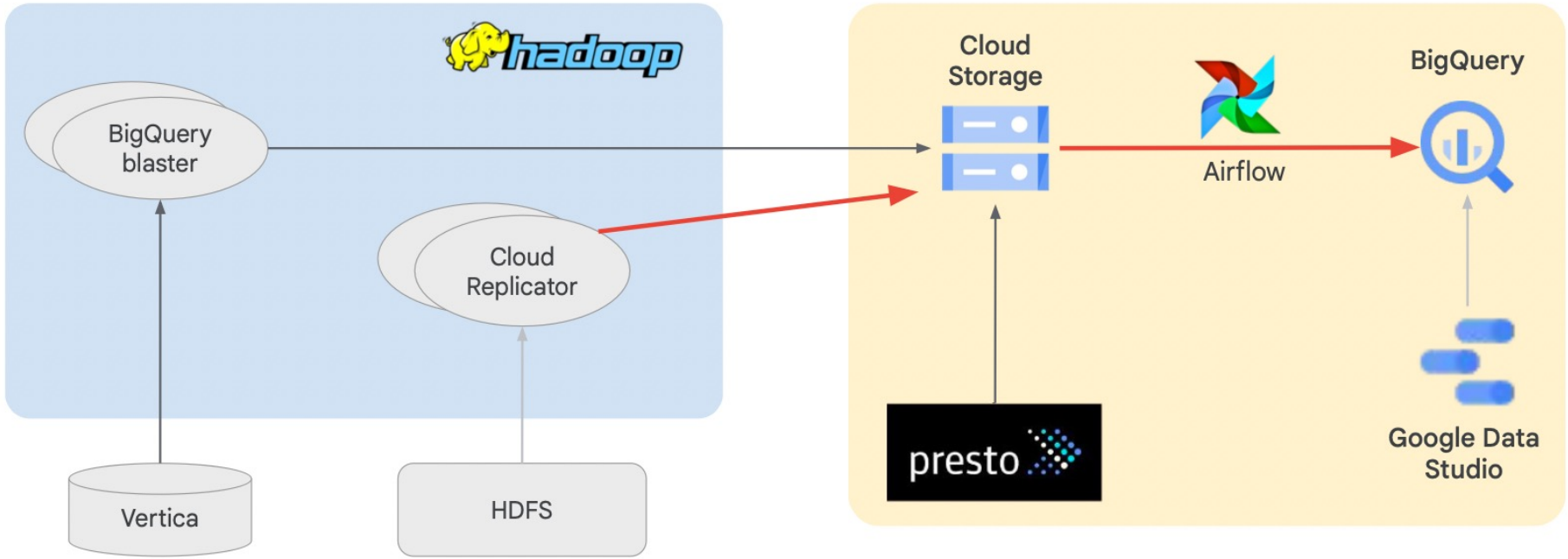


Google Cloud Customer Case Study

Twitter democratized data analysis using BigQuery

"We believe that users with a wide range of technical skills should be able to discover data and have access to SQL-based analysis and visualization tools that perform well"

-- Twitter

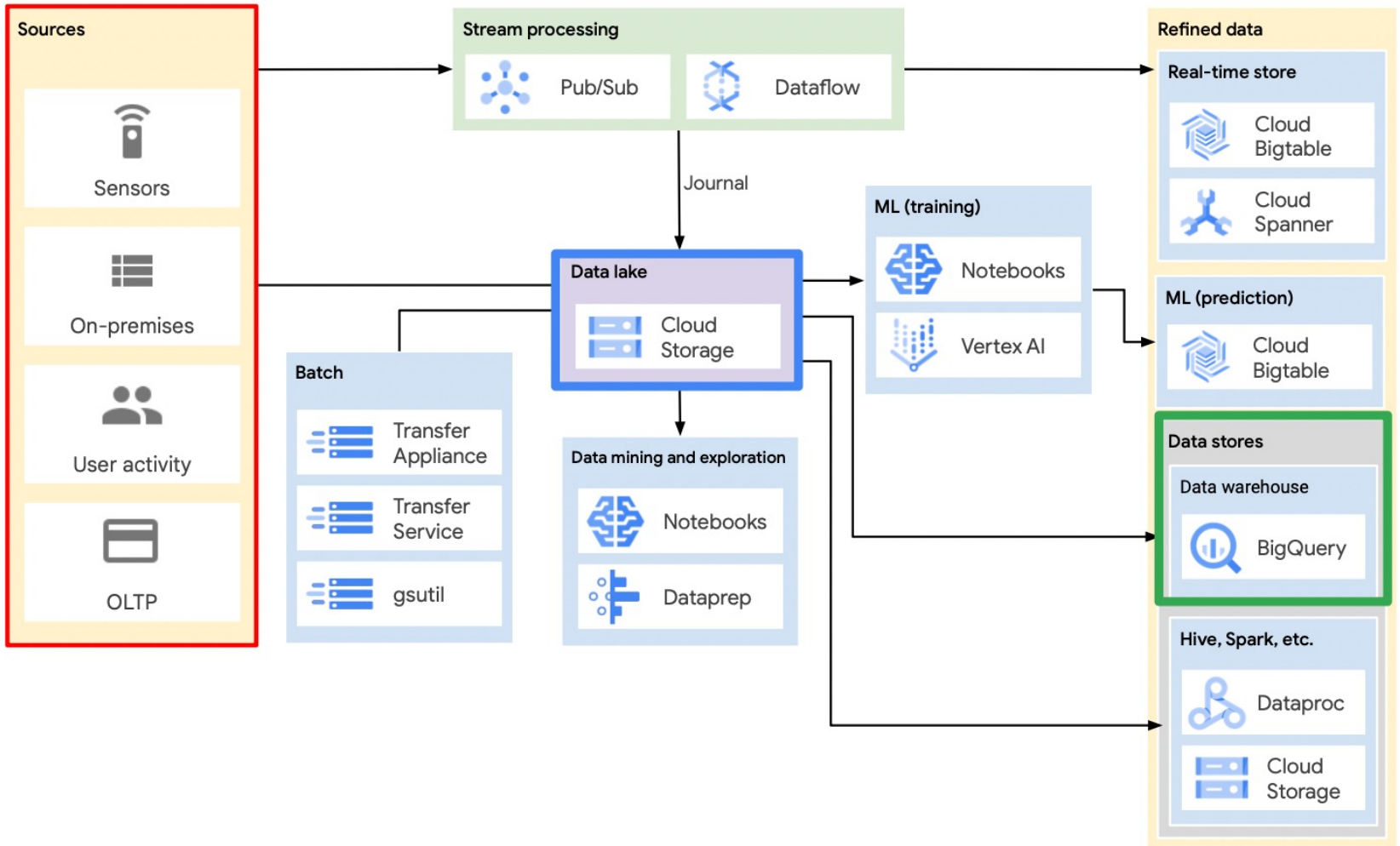


Recap

- Data sources
- Data lakes
- Data warehouses
- Google Cloud solutions for Data Engineering

Concept review

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



Lab Intro

Using BigQuery to do Analysis