Google Cloud

Introduction to Data Engineering

Agenda

Explore the role of a data engineer

Analyze data engineering challenges

Intro to BigQuery

Data Lakes and Data Warehouses

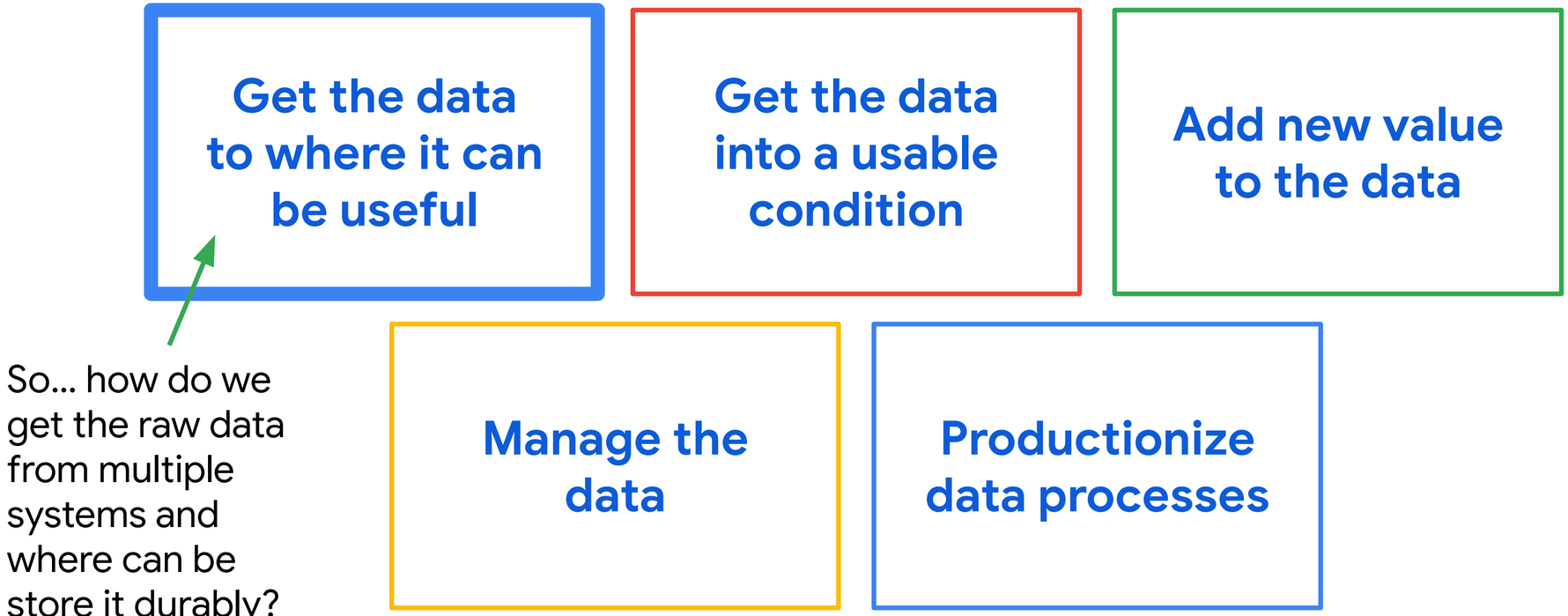
Transactional Databases vs Data Warehouses

Partner effectively with other data teams
- Manage data access and governance
- Build production-ready pipelines

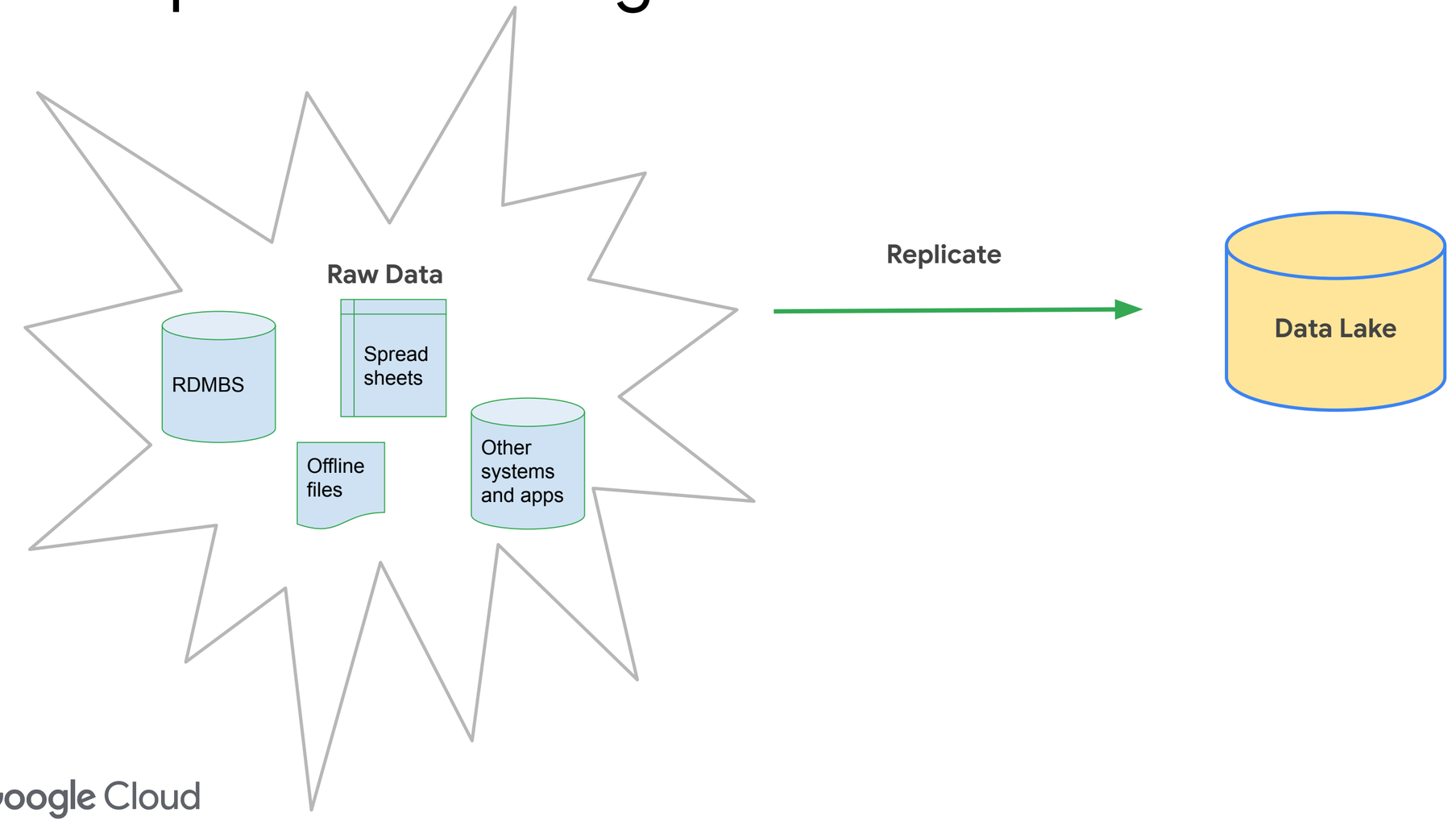
Review GCP customer case study

Lab: Analyzing Data with BigQuery

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



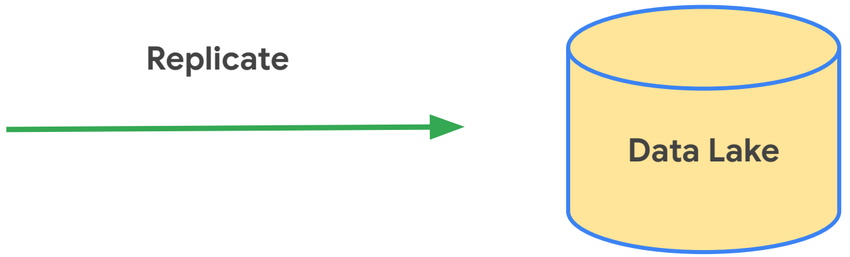
A data lake brings together data from across the enterprise into a single location



Google Cloud

Key considerations when building a Data Lake

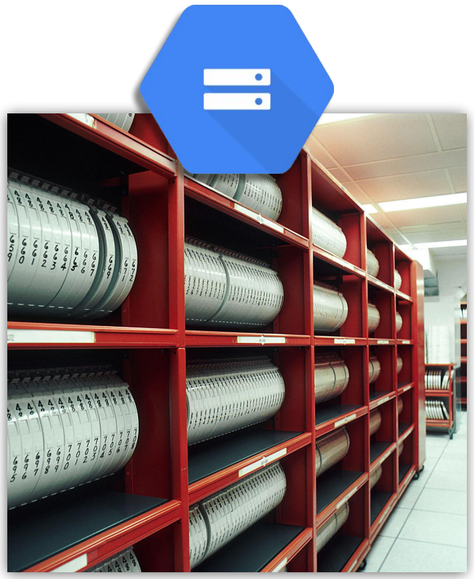
1. Can your data lake handle all the types of data you have?
2. Can it scale to meet the demand?
3. Does it support high-throughput ingestion?
4. Is there fine-grained access control to objects?
5. Can other tools connect easily?



We need an elastic data container that is flexible and durable to stage all our data ...

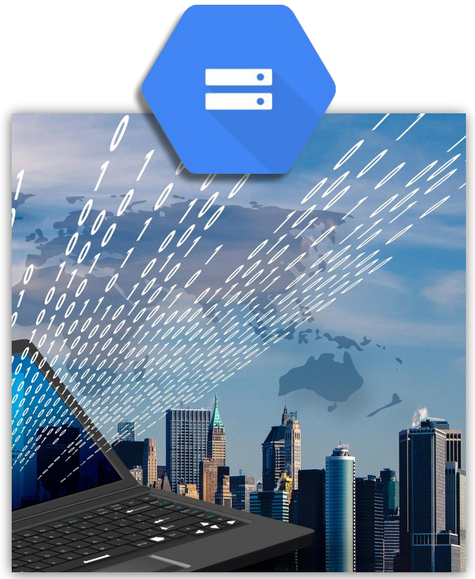
Cloud Storage is designed for 99.999999999% annual durability

Backup



Replace/decommission infrastructure

Analytics and ML



Content storage and delivery

Quickly create buckets with cloud shell

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

Google Cloud

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?

Streaming Data Processing

Cloud Pub/Sub

Cloud Dataflow

BigQuery

Agenda

Explore the role of a data engineer

Analyze data engineering challenges

Intro to BigQuery

Data Lakes and Data Warehouses

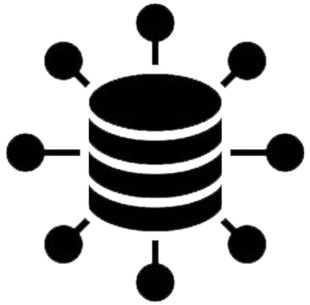
Transactional Databases vs Data Warehouses

Partner effectively with other data teams
- Manage data access and governance
- Build production-ready pipelines

Review GCP customer case study

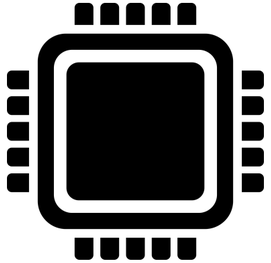
Lab: Analyzing Data with BigQuery

Common challenges encountered by data engineers



Access to data

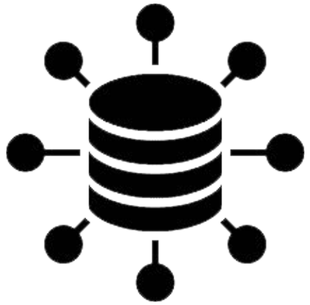
Data accuracy and quality



Availability of computational resources

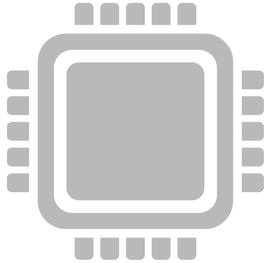
Query performance

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

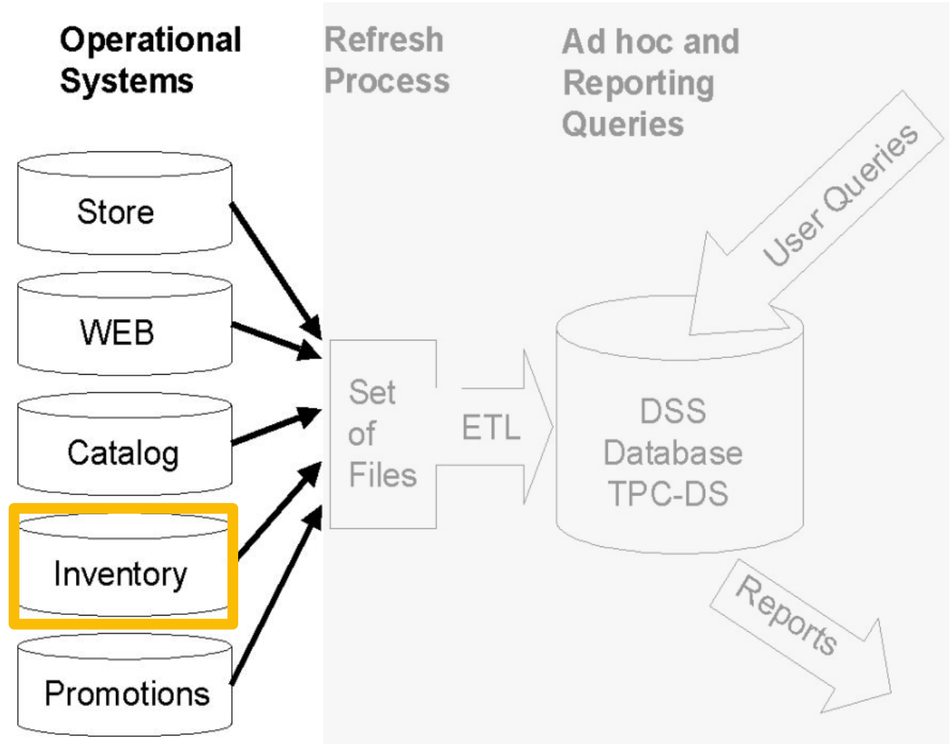
Data is scattered across Google Analytics 360, CRM, and Campaign Manager products, among other sources.

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

Customer and sales data is stored in a CRM system.

Some data is not in a queryable format.

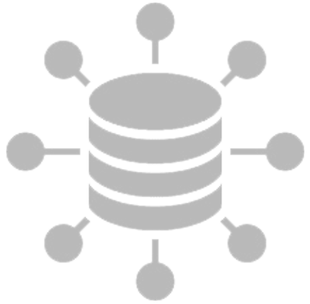
Data is often siloed in many upstream source systems



Stored in a separate system and restricted access

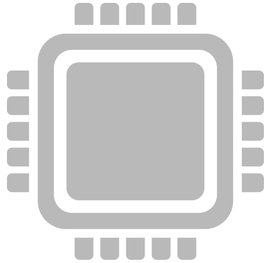
Example Query:
Give me all the in-store promotions for recent orders and their **inventory levels**

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



Access to data

Data accuracy and quality

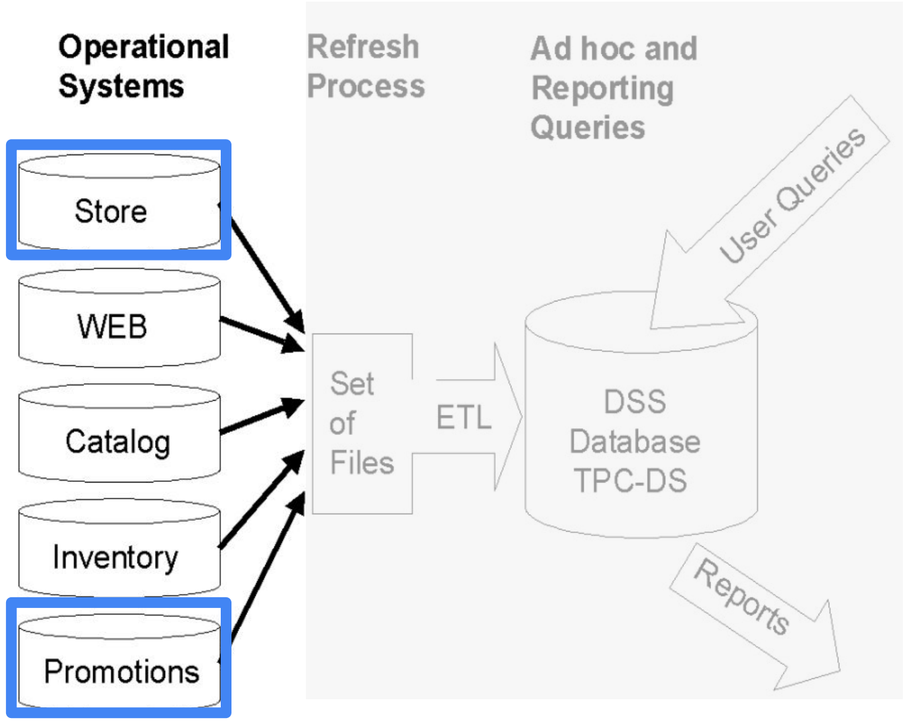


Availability of computational resources

Query performance

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

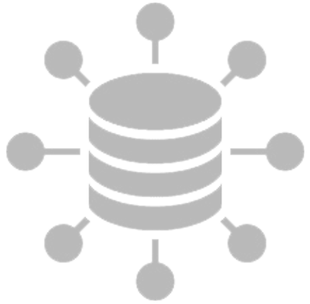
Missing data and all timestamps stored as UTC



Promotion list stored as .csv files or manual spreadsheets

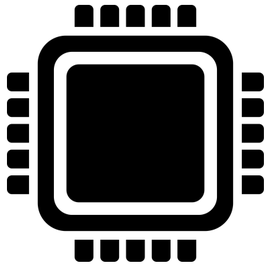
Query: Give me the best performing in-store promotions in France

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



Access to data

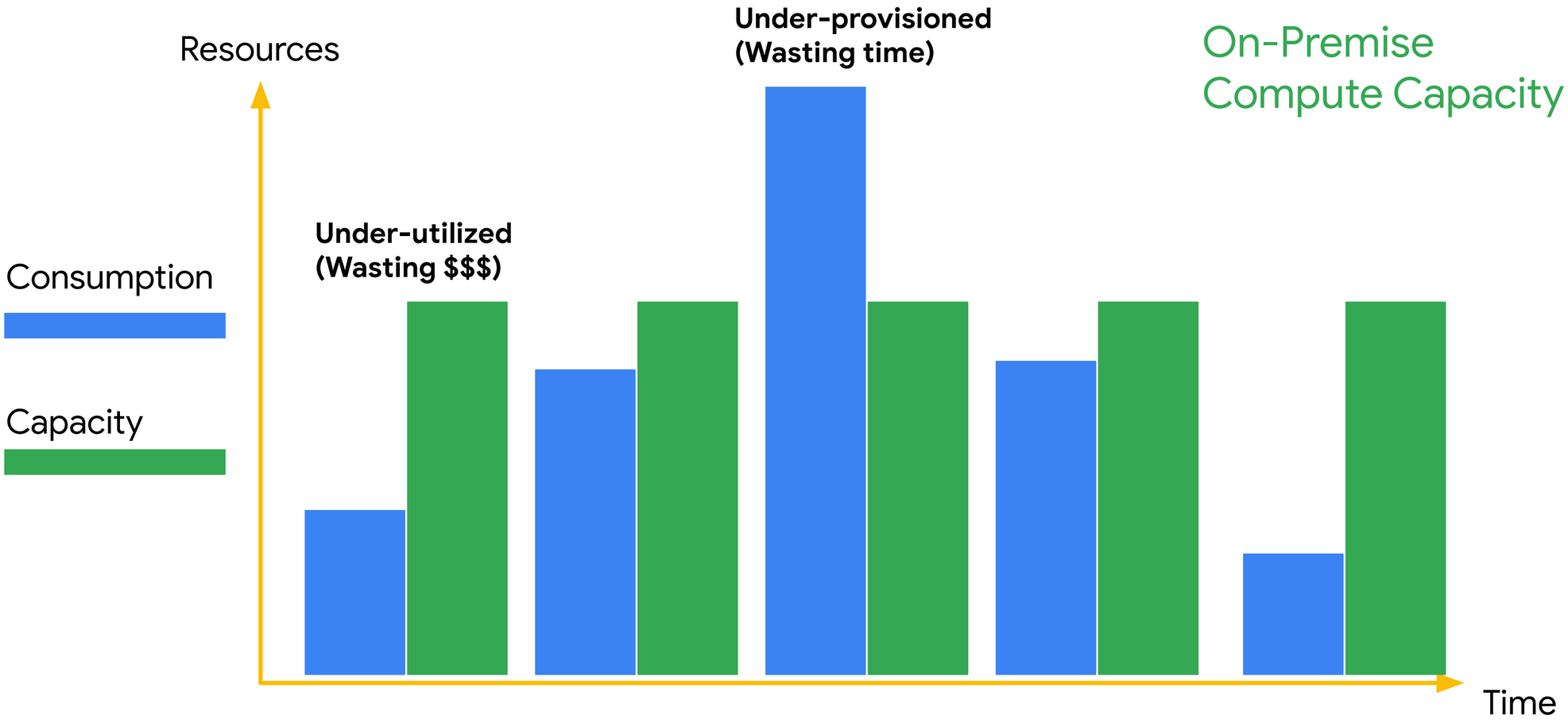
Data accuracy and quality



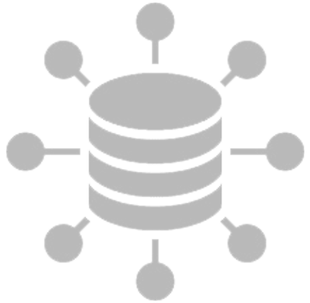
Availability of computational resources

Query performance

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

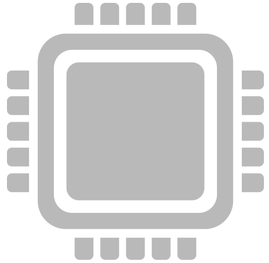


Challenge: Queries need to be optimized for performance (caching, parallel execution)



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

Is there a better way to manage server overhead so we can focus on insights?

Agenda

Explore the role of a data engineer

Analyze data engineering challenges

Intro to BigQuery

Data Lakes and Data Warehouses

Transactional Databases vs Data Warehouses

Partner effectively with other data teams
- Manage data access and governance
- Build production-ready pipelines

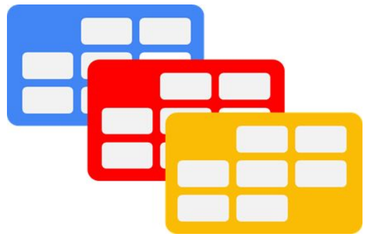
Review GCP customer case study

Lab: Analyzing Data with BigQuery

BigQuery is Google's data warehouse solution

Data warehouse

BigQuery replaces a typical data warehouse hardware setup

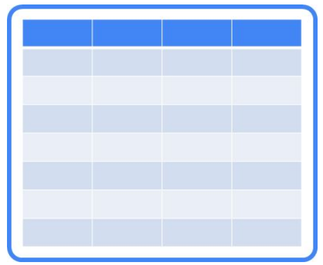


Data mart

BigQuery organizes data tables into units called datasets

Data lake

BigQuery defines schemas and issues queries directly on external data sources



Tables and views

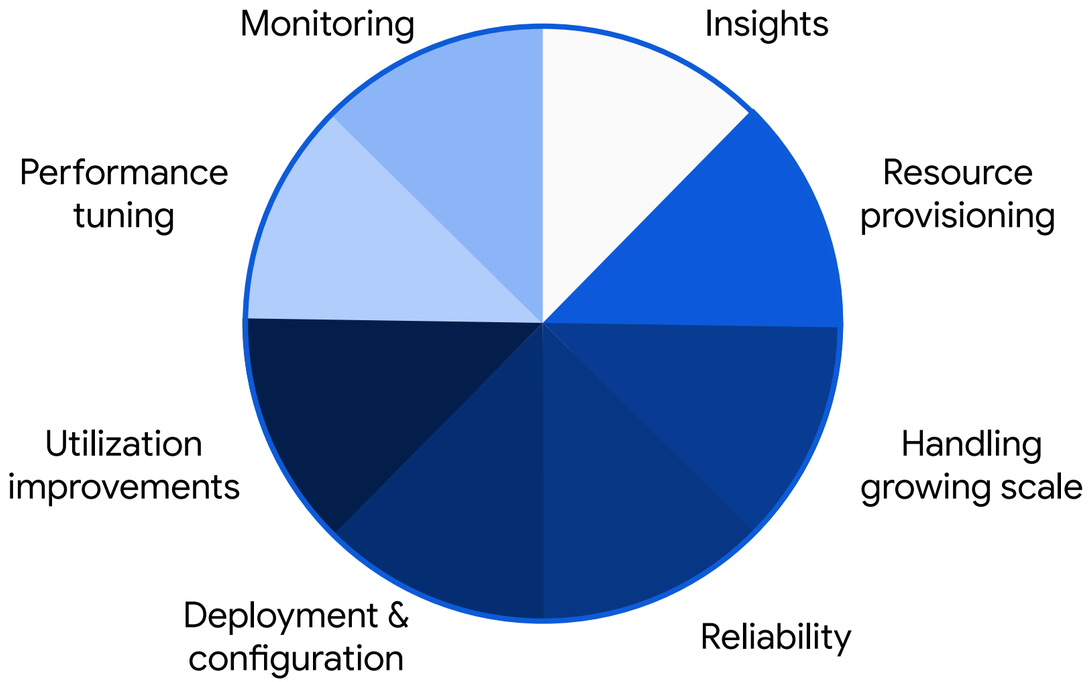
Function the same way as in a traditional data warehouse

Grants

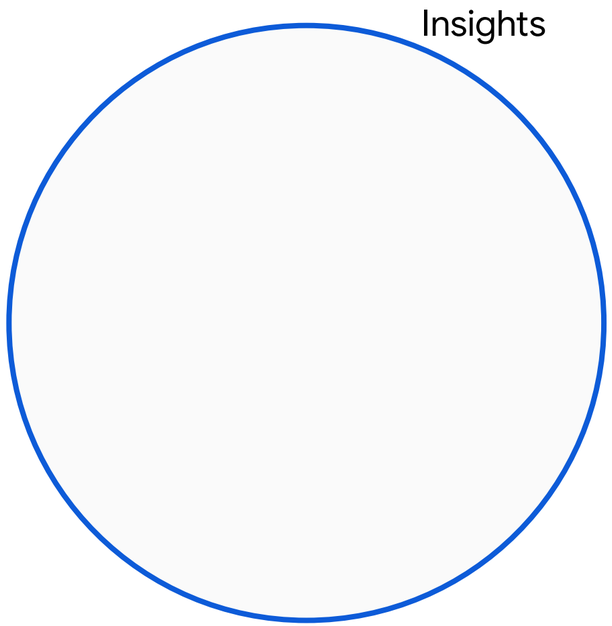
Cloud IAM grants permission to perform specific actions

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

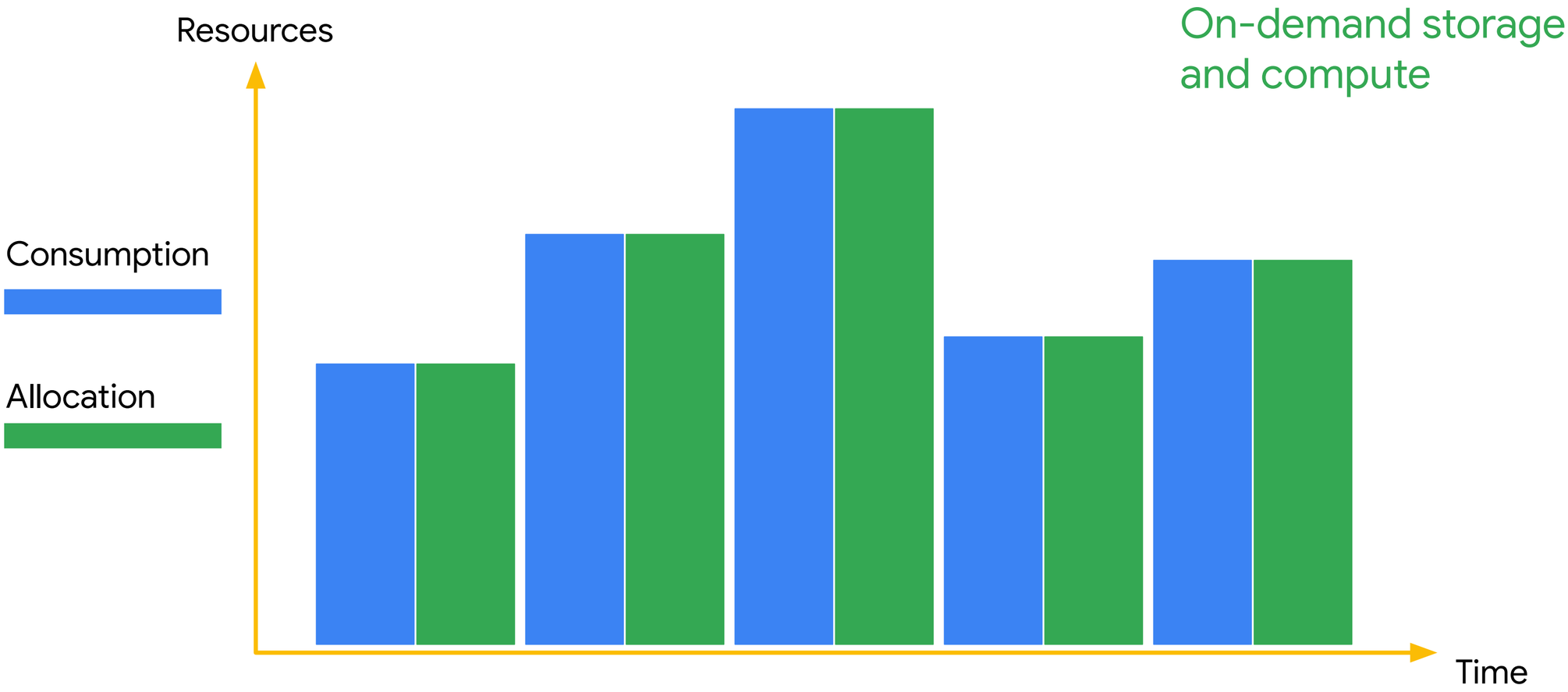
Typical Big Data Processing



With Google



You don't need to provision resources before using BigQuery



Google Cloud

Agenda

Explore the role of a data engineer

Analyze data engineering challenges

Intro to BigQuery

Data Lakes and Data Warehouses

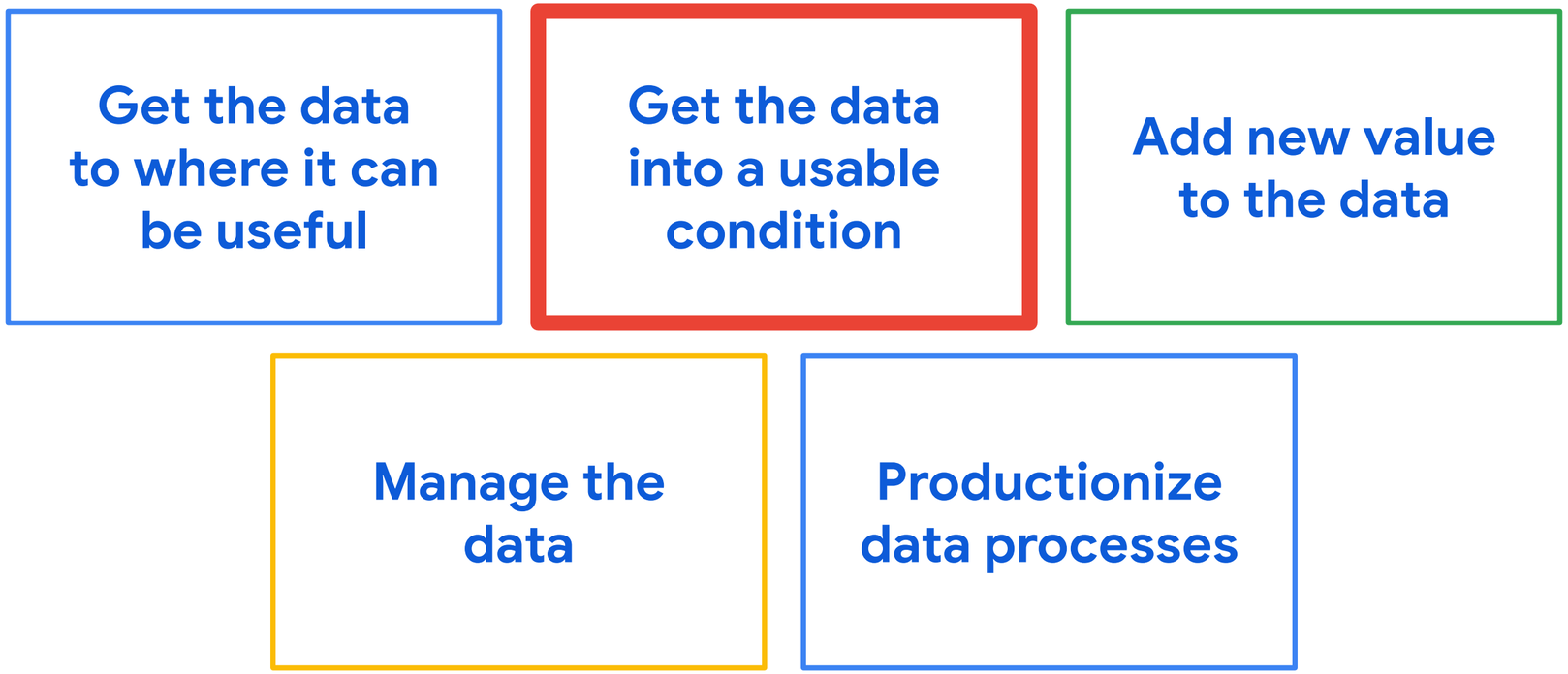
Transactional Databases vs Data Warehouses

Partner effectively with other data teams
- Manage data access and governance
- Build production-ready pipelines

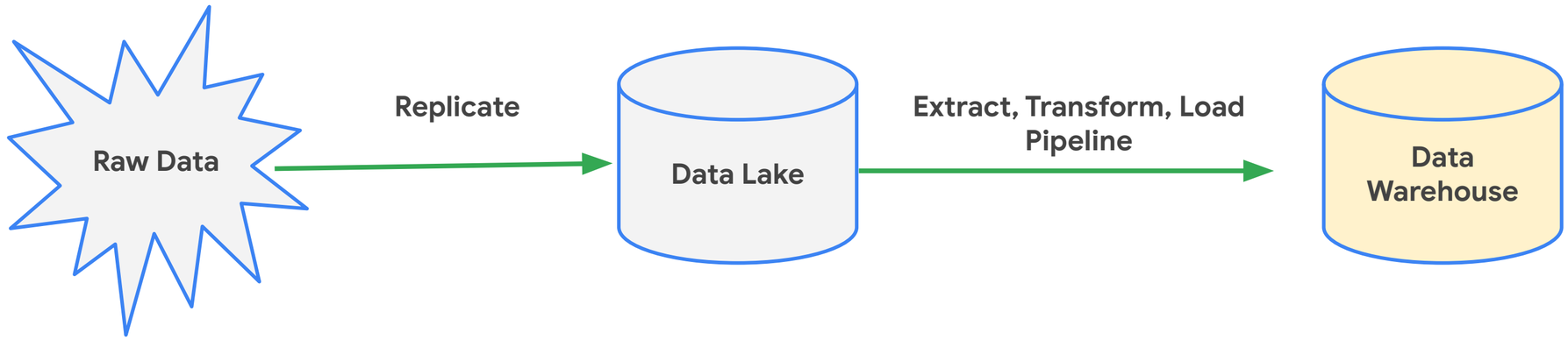
Review GCP customer case study

Lab: Analyzing Data with BigQuery

A data engineer gets data into a useable condition



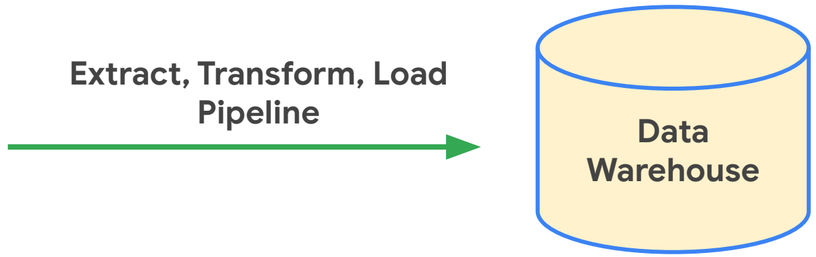
A data warehouse stores transformed data in a usable condition for business insights



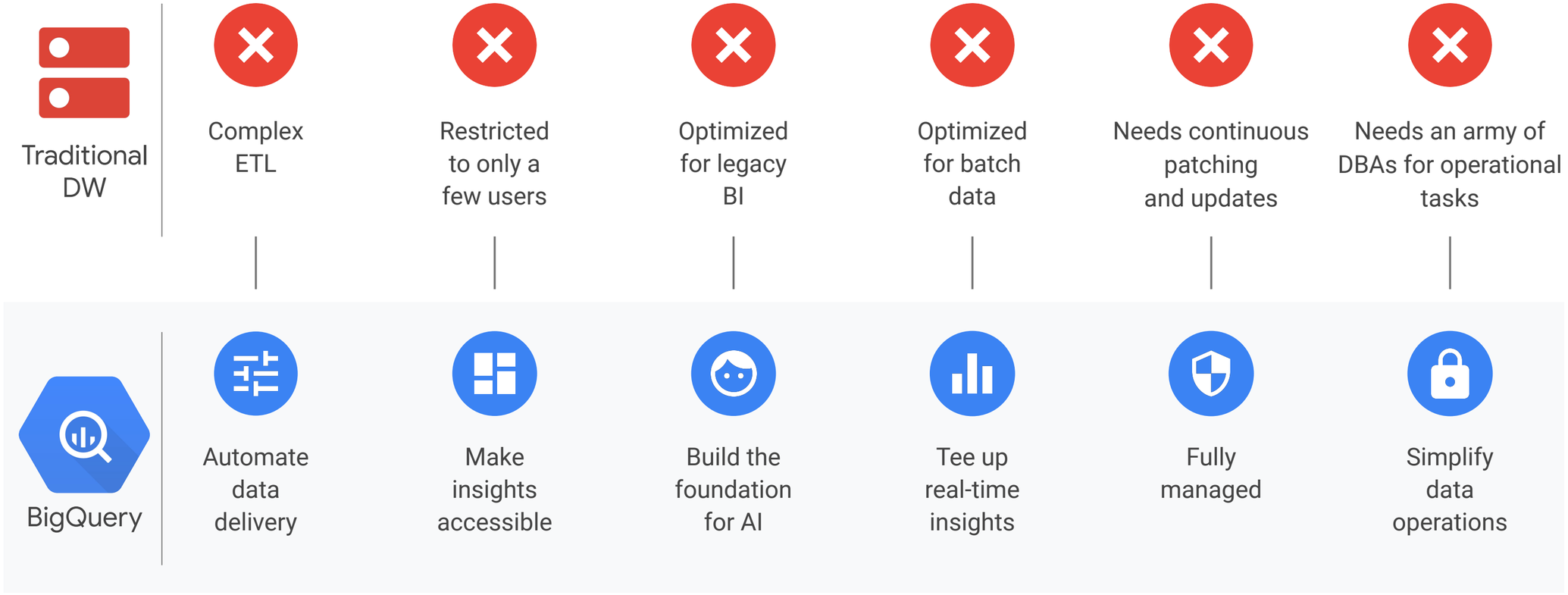
What are the key considerations when deciding between data warehouse options?

Considerations when choosing a data warehouse include:

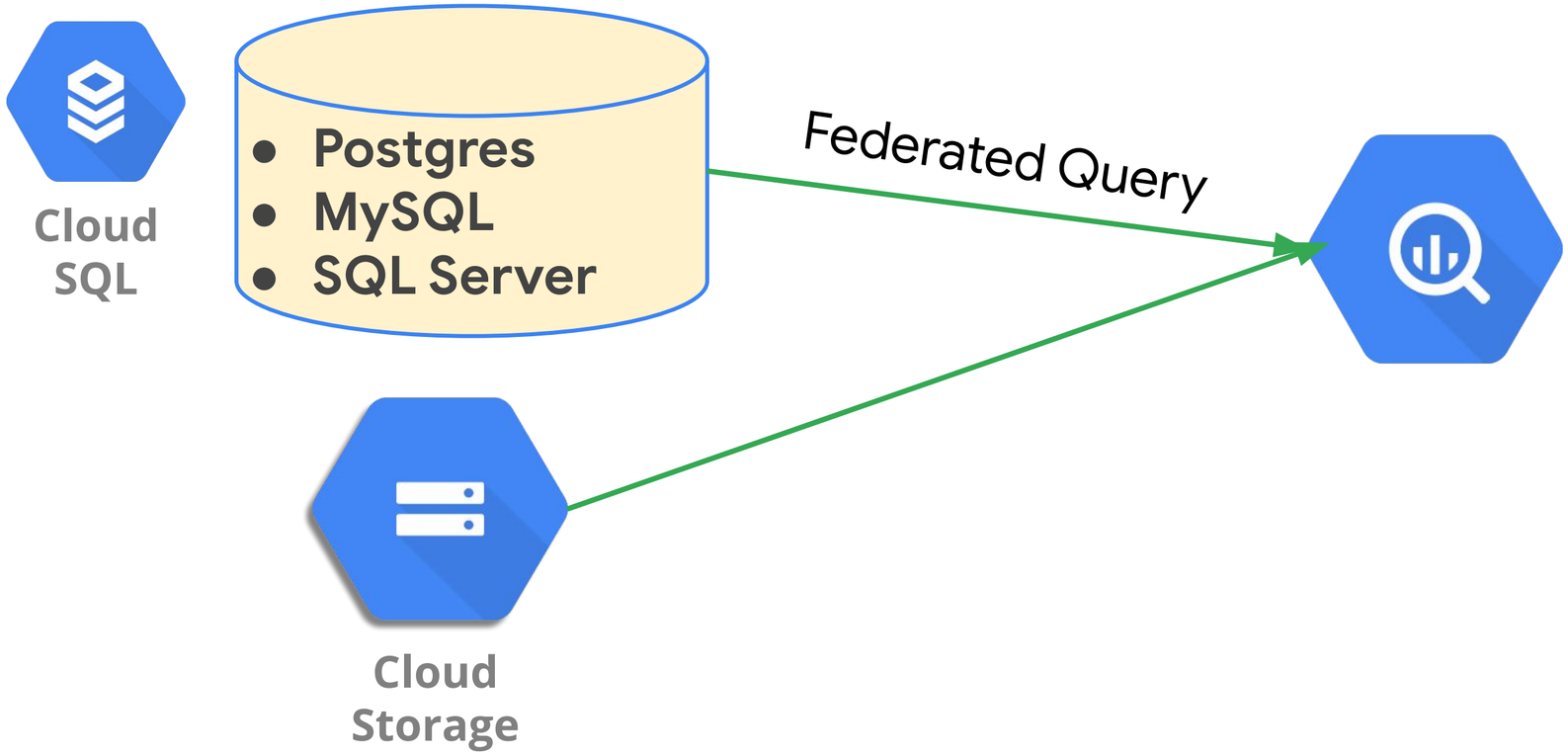
- Can it serve as a sink for both batch and streaming data pipelines?
- Can the data warehouse scale to meet my needs?
- How is the data organized, cataloged, and access controlled?
- Is the warehouse designed for performance?
- What level of maintenance is required by our engineering team?



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



Demo

Federated Queries with BigQuery

Agenda

Explore the role of a data engineer

Analyze data engineering challenges

Intro to BigQuery

Data Lakes and Data Warehouses

Transactional Databases vs Data Warehouses

Partner effectively with other data teams
- Manage data access and governance
- Build production-ready pipelines

Review GCP customer case study

Lab: Analyzing Data with BigQuery

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

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

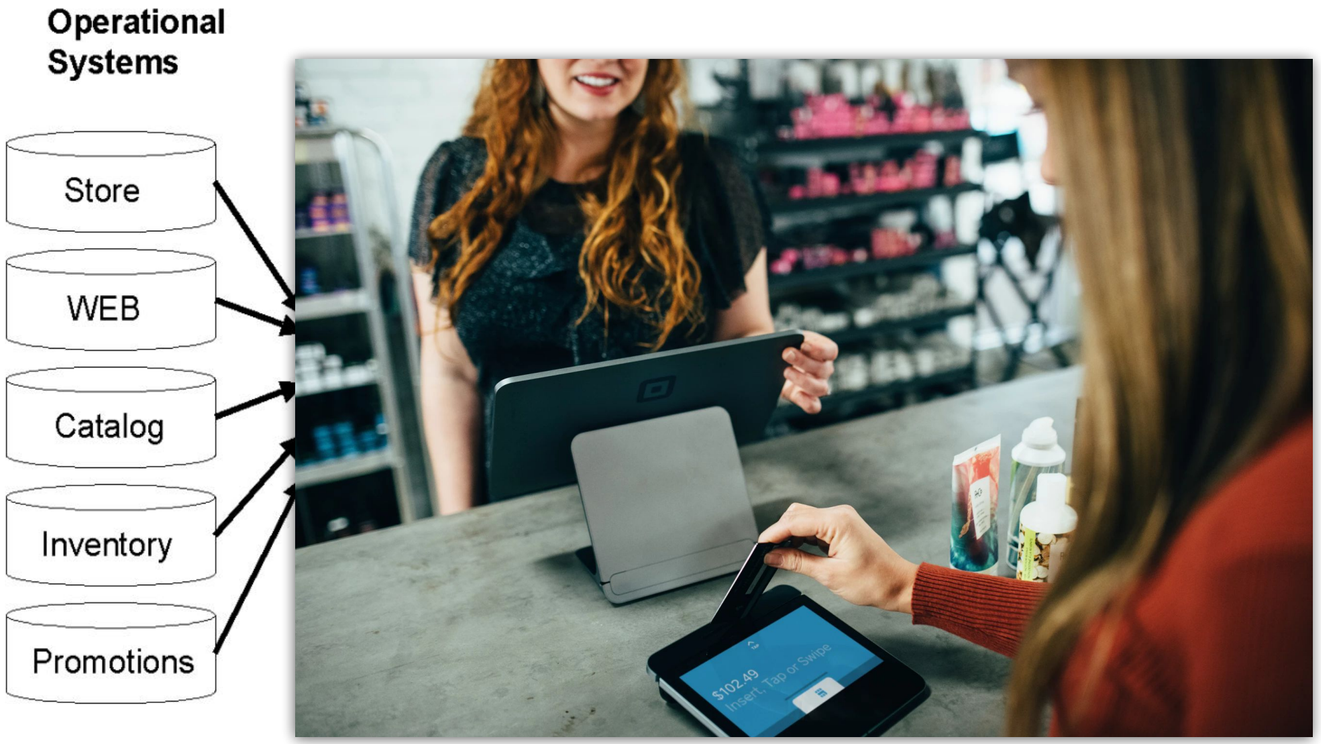
BigQuery

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

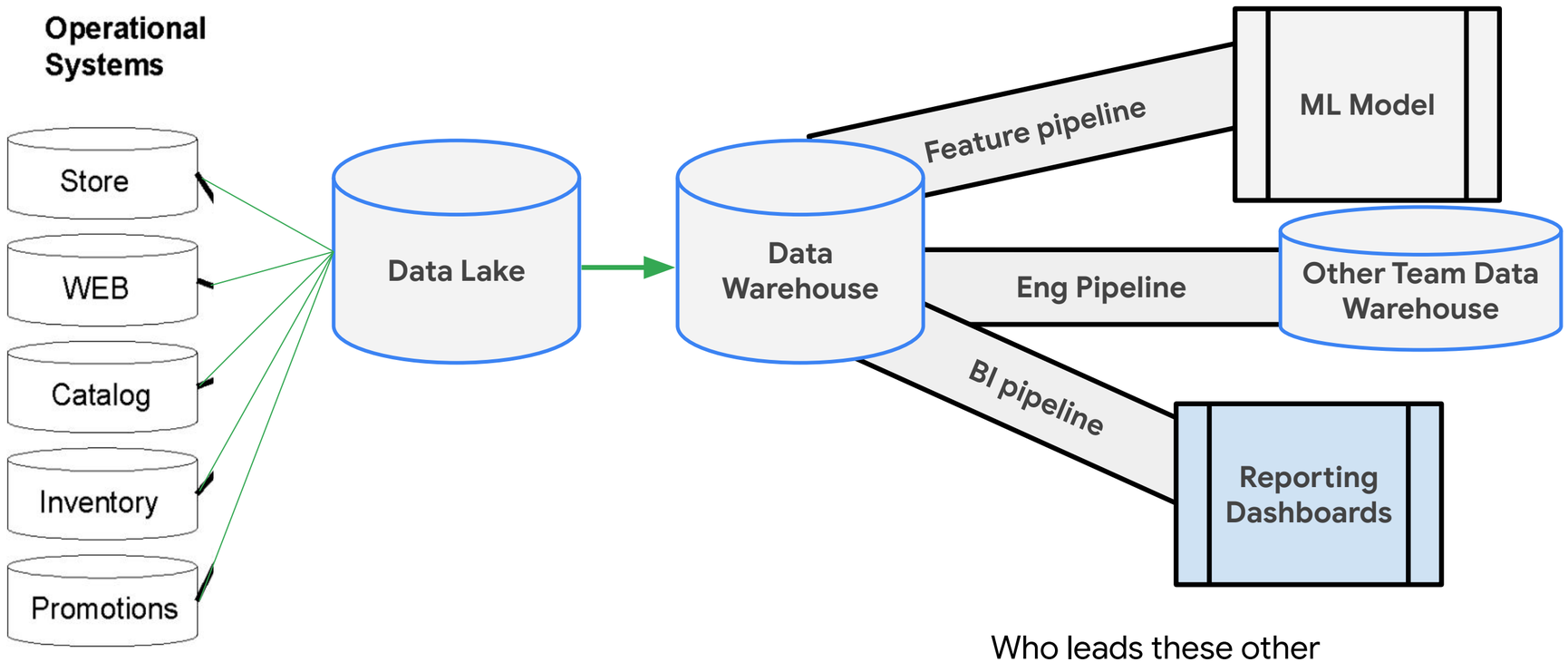
- Scales to PB
- Easily connect to external data sources for ingestion
- Column based storage

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



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

Agenda

Explore the role of a data engineer

Analyze data engineering challenges

Intro to BigQuery

Data Lakes and Data Warehouses

Transactional Databases vs Data Warehouses

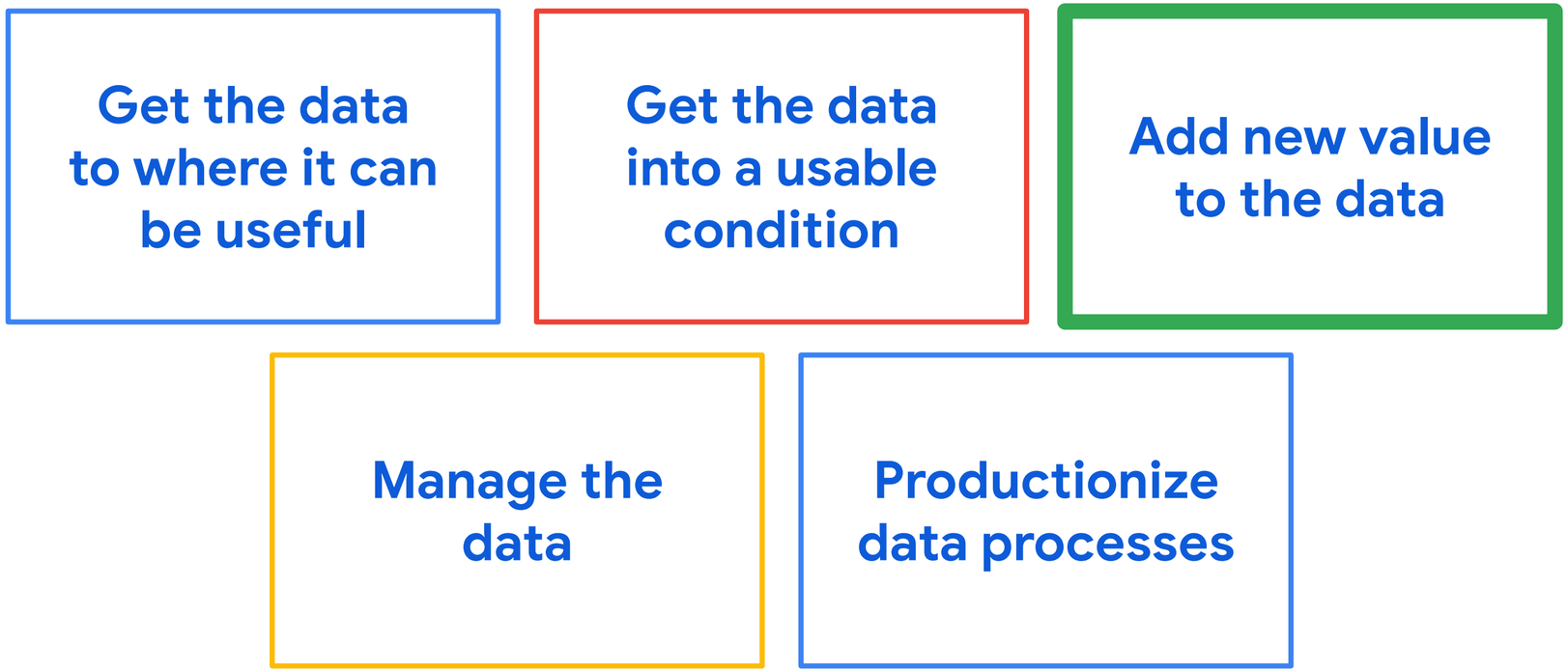
Partner effectively with other data teams
- Manage data access and governance
- Build production-ready pipelines

Review GCP customer case study

Lab: Analyzing Data with BigQuery

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

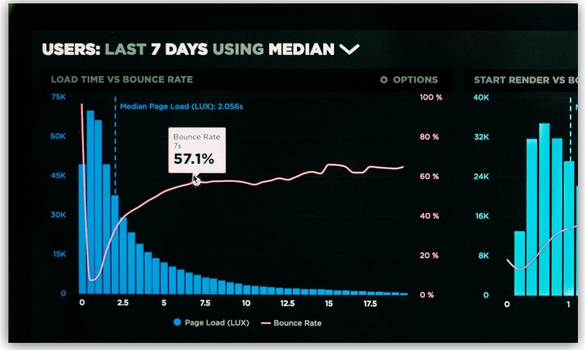
What teams rely on these pipelines?



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

Machine Learning Engineer

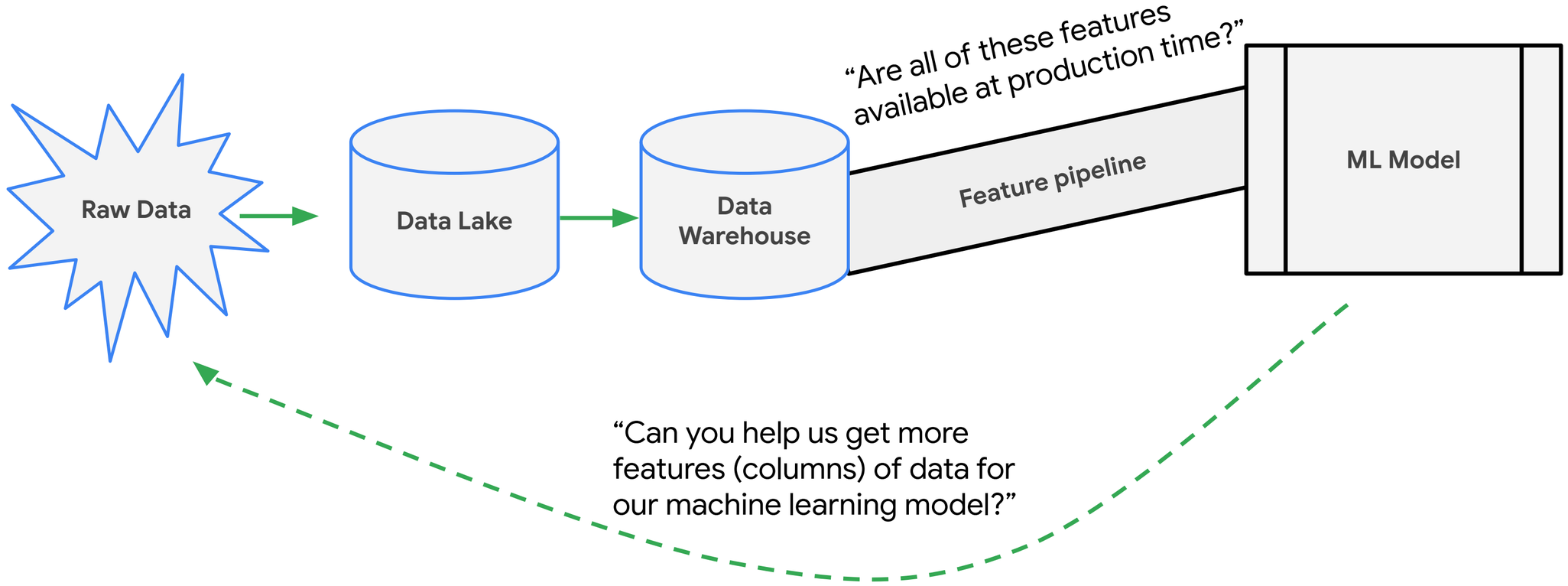
Data Analyst



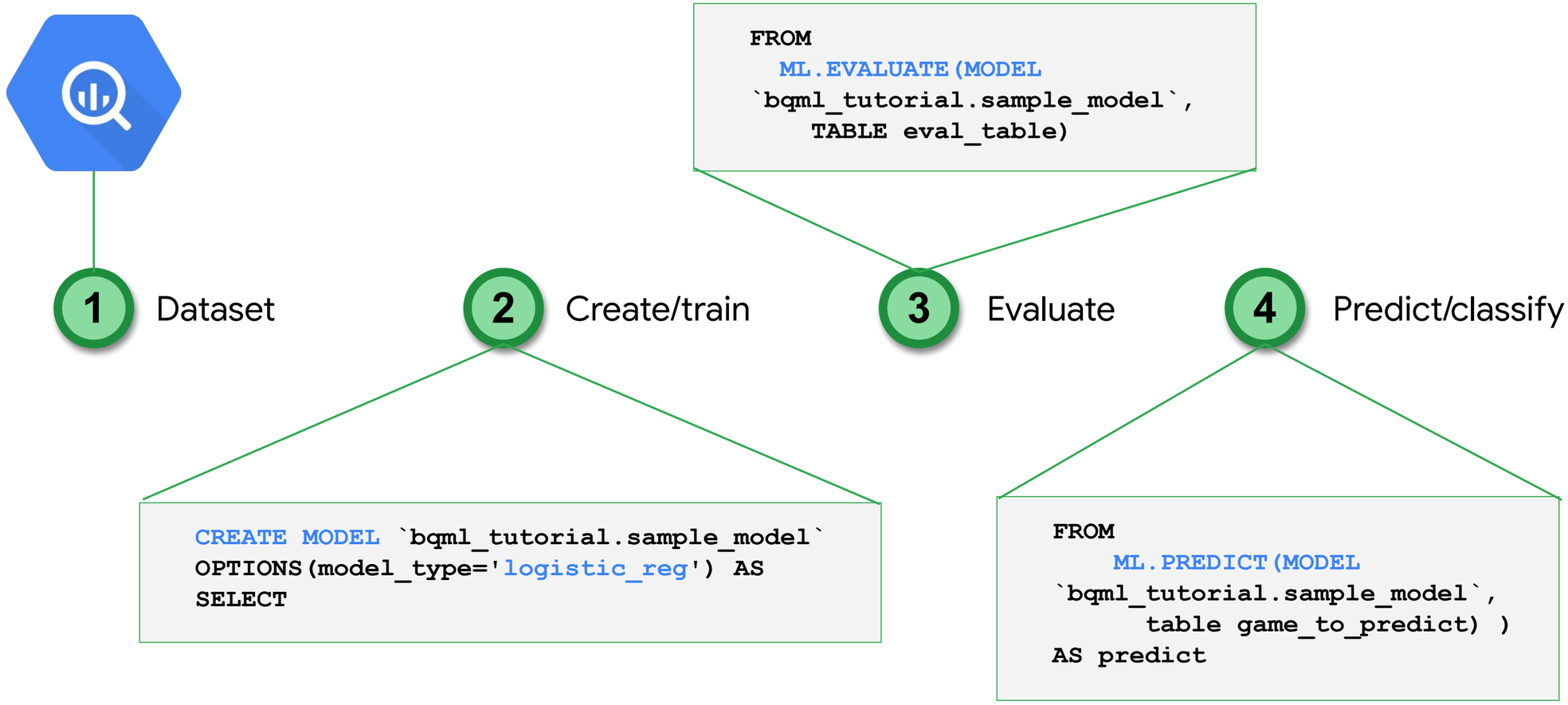
Data Engineer

How might each of these teams rely on data engineering?

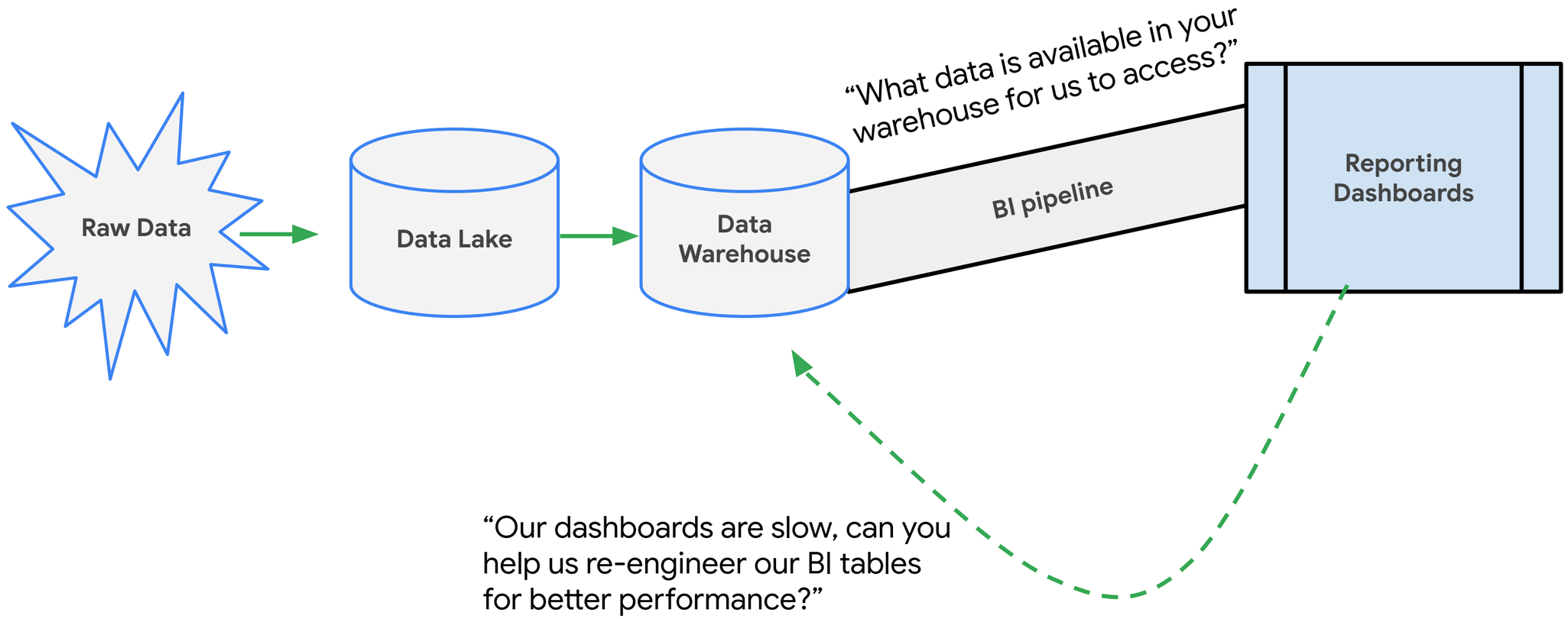
Machine learning teams need data engineers to help them capture new features in a stable pipeline



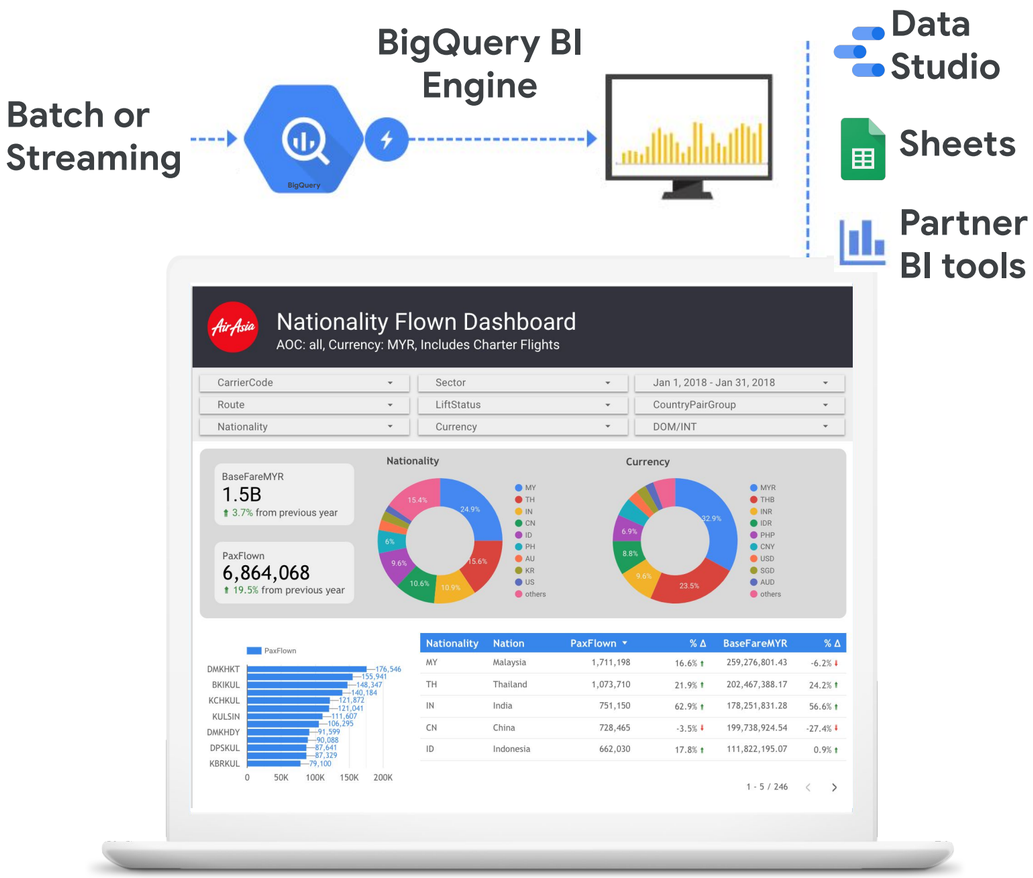
Add value: Machine learning directly in BigQuery



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



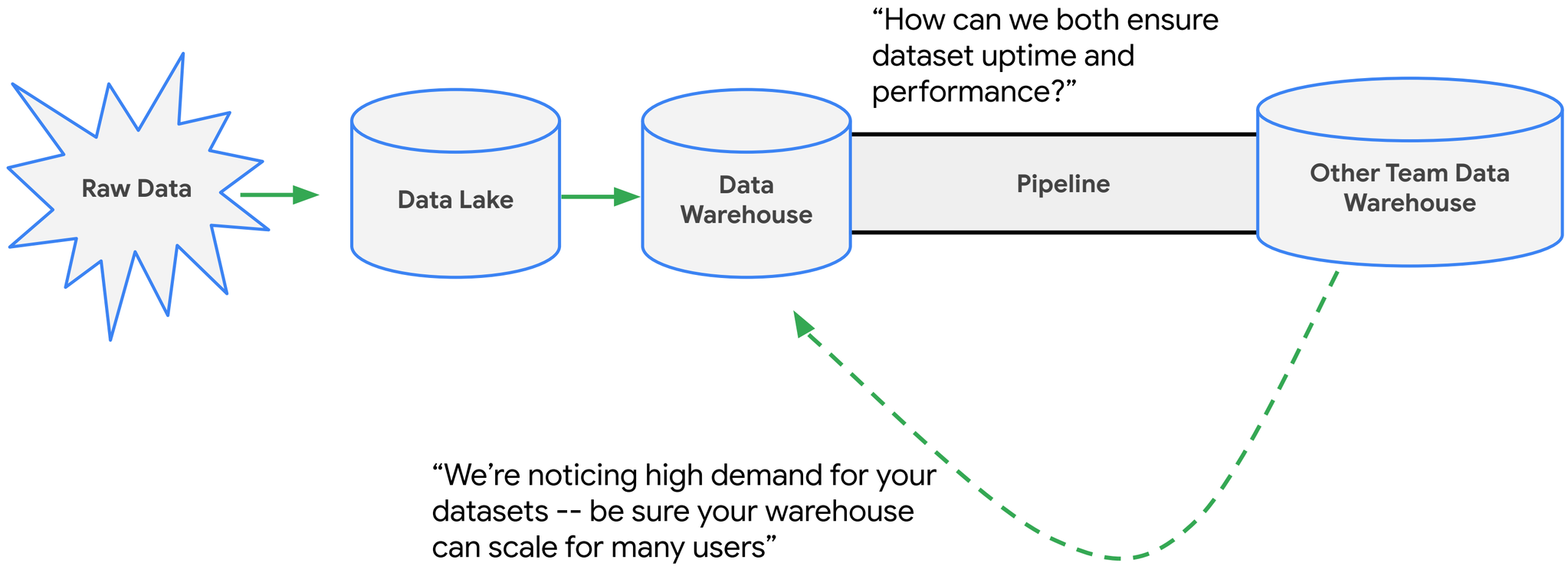
Add value: BI Engine for dashboard performance



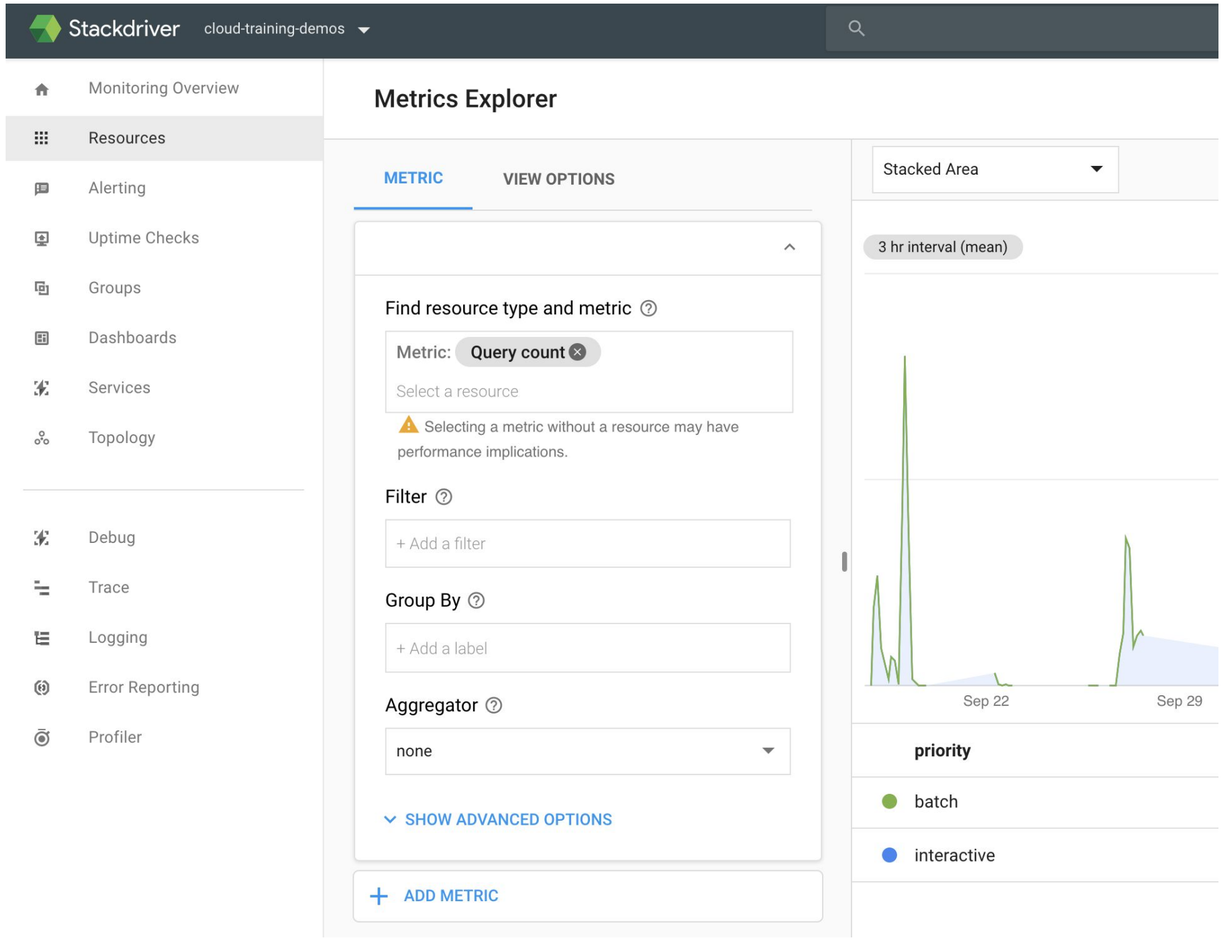
- No need to manage OLAP cubes or separate BI servers for dashboard performance
- Natively integrates with BigQuery streaming for real-time data refresh
- Column oriented in-memory BI execution engine

Google Cloud

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



Add value: Stackdriver monitoring for performance



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

Agenda

Explore the role of a data engineer

Analyze data engineering challenges

Intro to BigQuery

Data Lakes and Data Warehouses

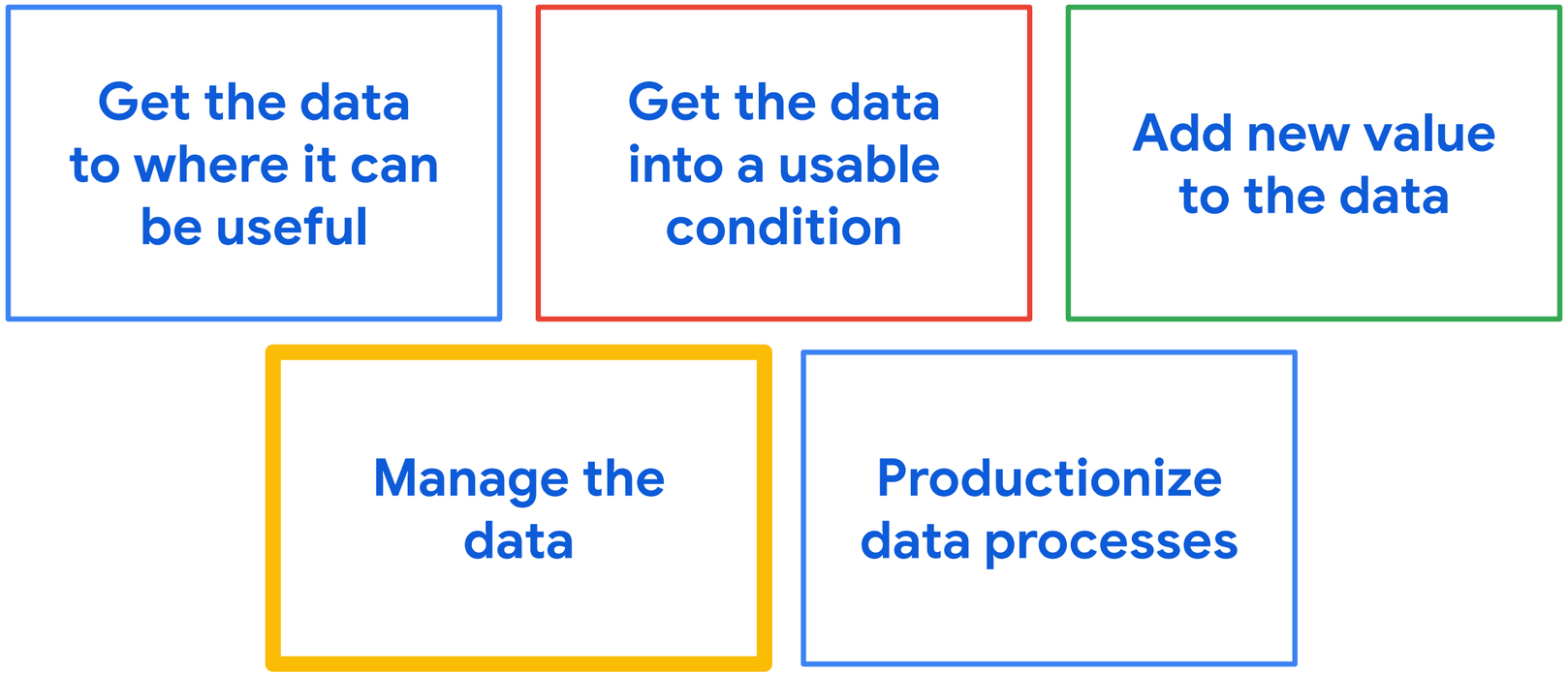
Transactional Databases vs Data Warehouses

Partner effectively with other data teams
- **Manage data access and governance**
- Build production-ready pipelines

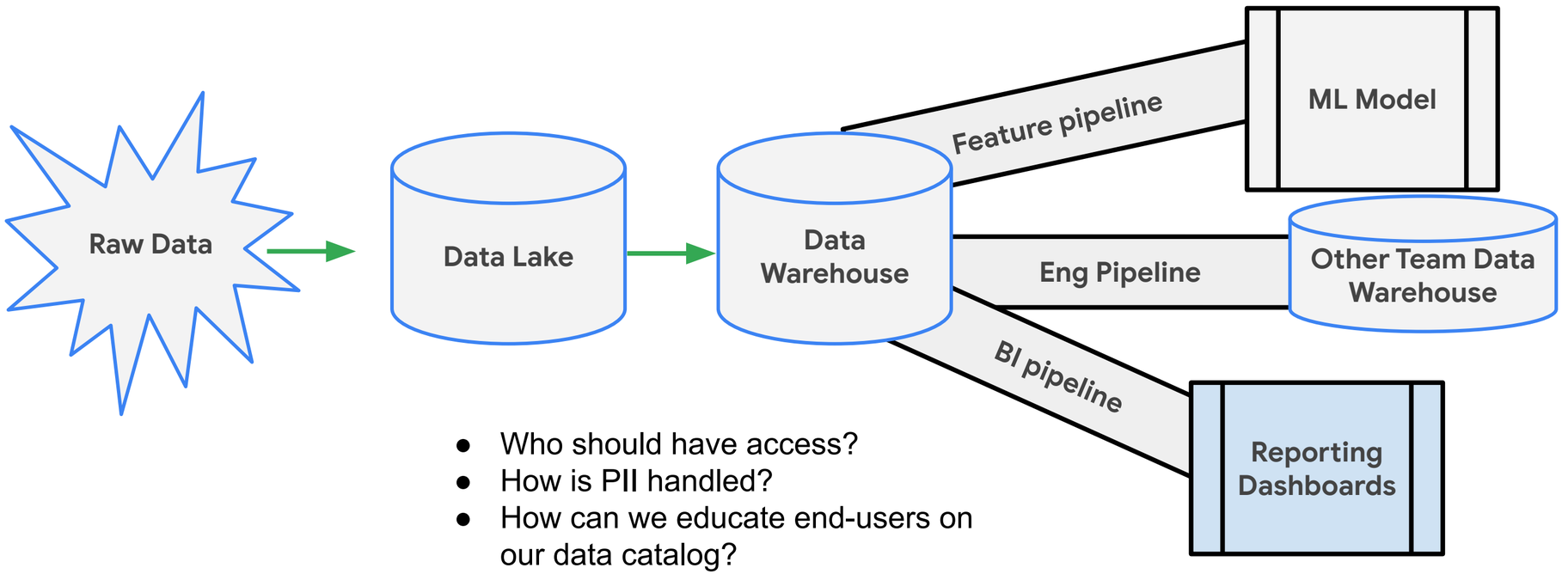
Review GCP customer case study

Lab: Analyzing Data with BigQuery

A data engineer manages data access and governance

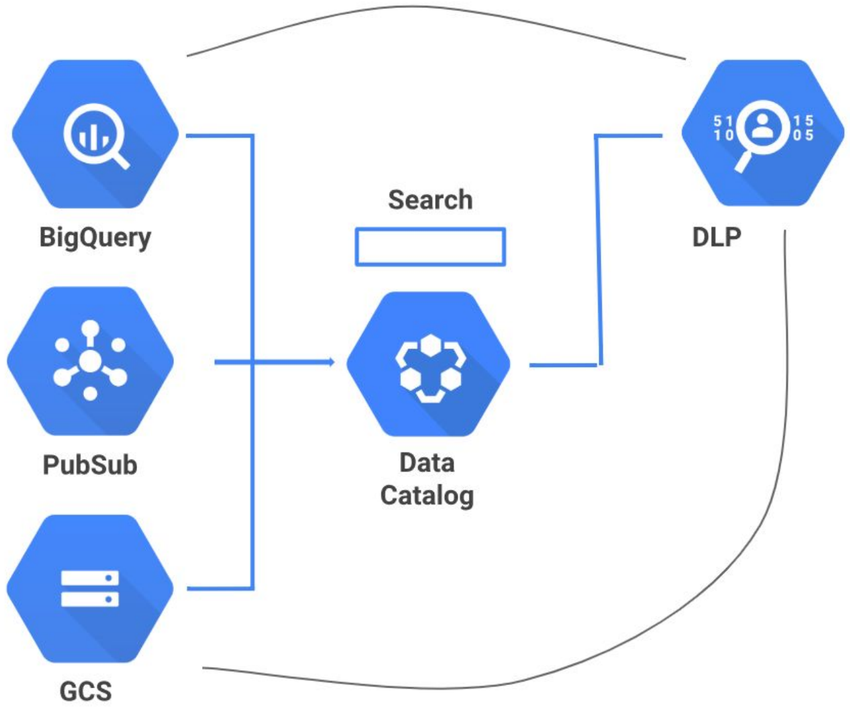


Data engineering must set and communicate a responsible data governance model



- 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

Demo

Finding PII in your dataset with DLP API

Agenda

Explore the role of a data engineer

Analyze data engineering challenges

Intro to BigQuery

Data Lakes and Data Warehouses

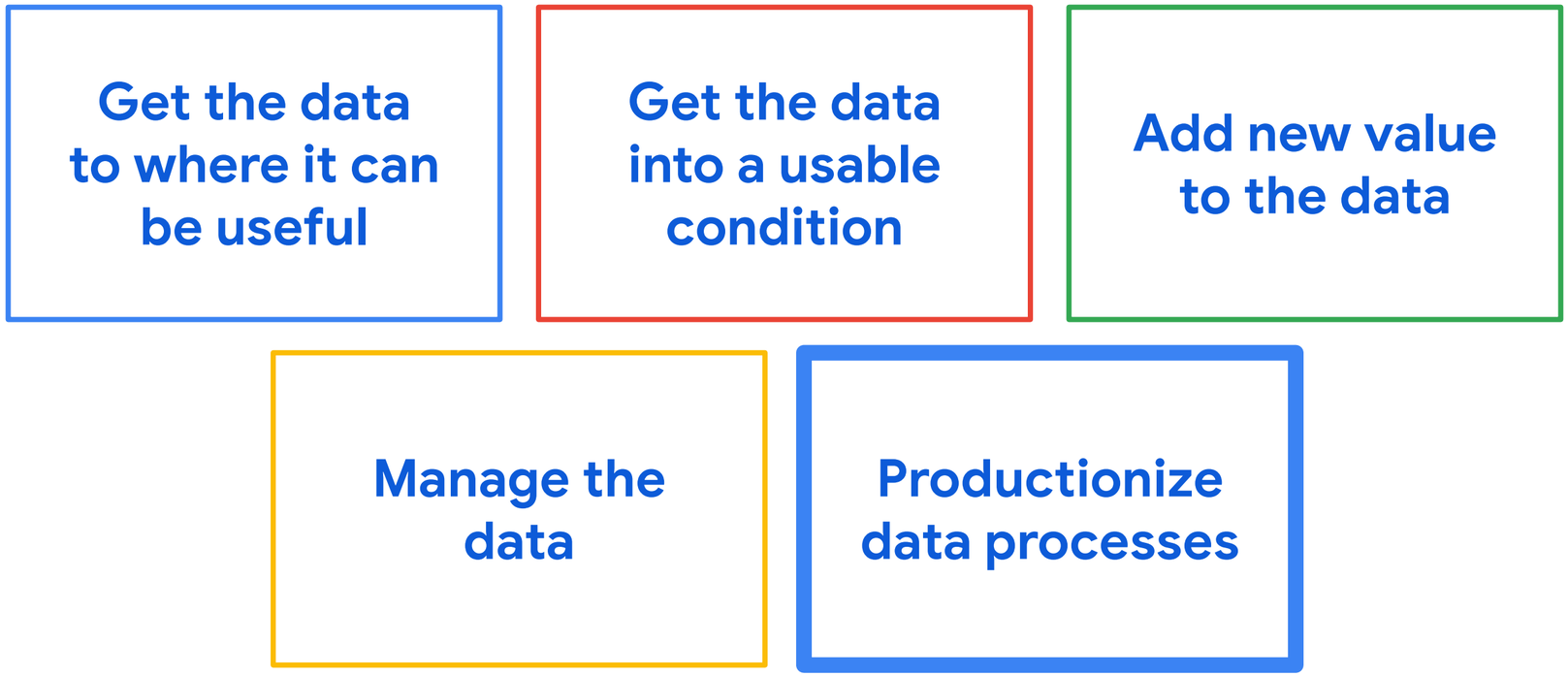
Transactional Databases vs Data Warehouses

Partner effectively with other data teams
- Manage data access and governance
- **Build production-ready pipelines**

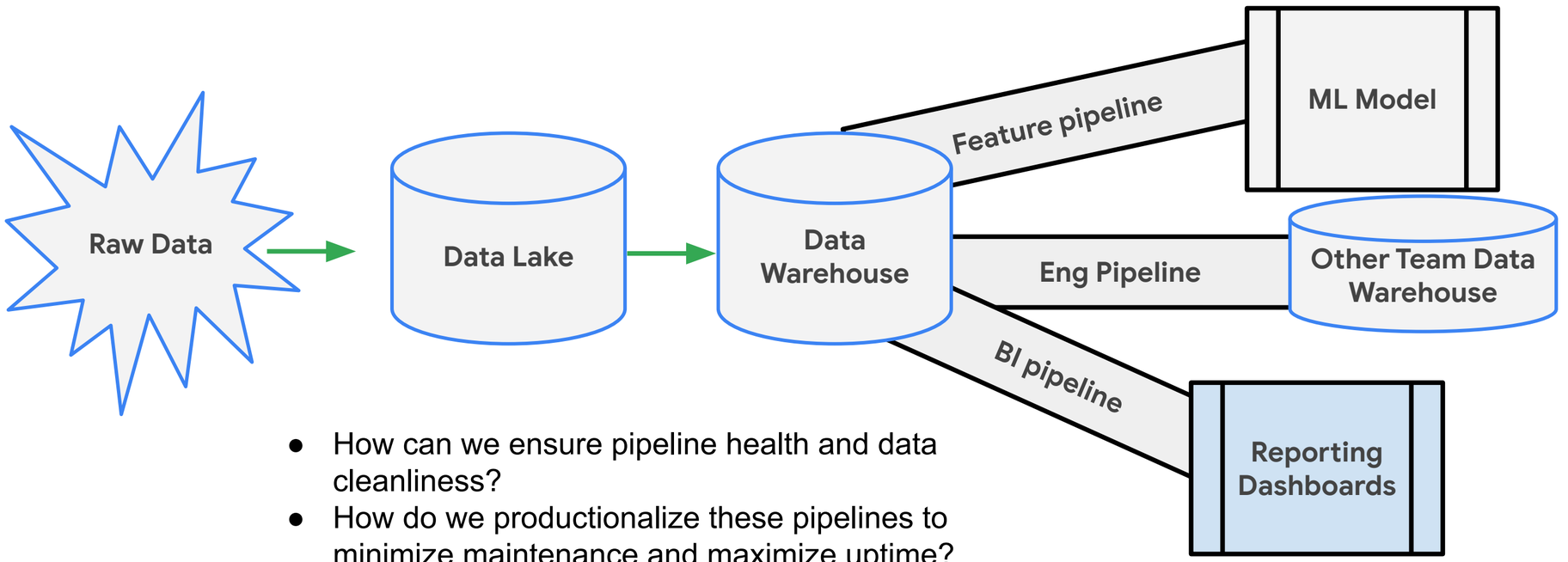
Review GCP customer case study

Lab: Analyzing Data with BigQuery

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

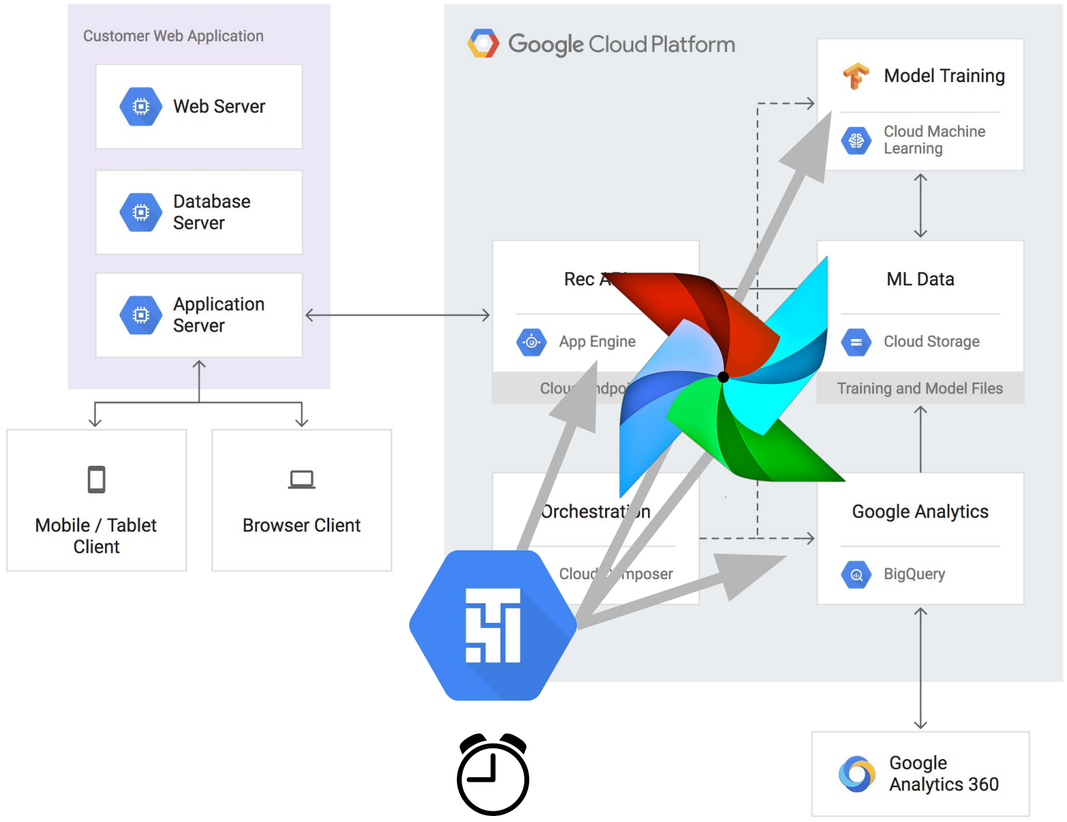


Data engineering owns the health and future of their production data pipelines



- 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



Agenda

Explore the role of a data engineer

Analyze data engineering challenges

Intro to BigQuery

Data Lakes and Data Warehouses

Transactional Databases vs Data Warehouses

Partner effectively with other data teams
- Manage data access and governance
- Build production-ready pipelines

Review GCP customer case study

Lab: Analyzing Data with BigQuery

Ocado's customer service department is bombarded with messages

Can we use ML to prioritize these messages?

"My order is missing 12 steaks. I need them for a party I'm hosting in 6 hours."

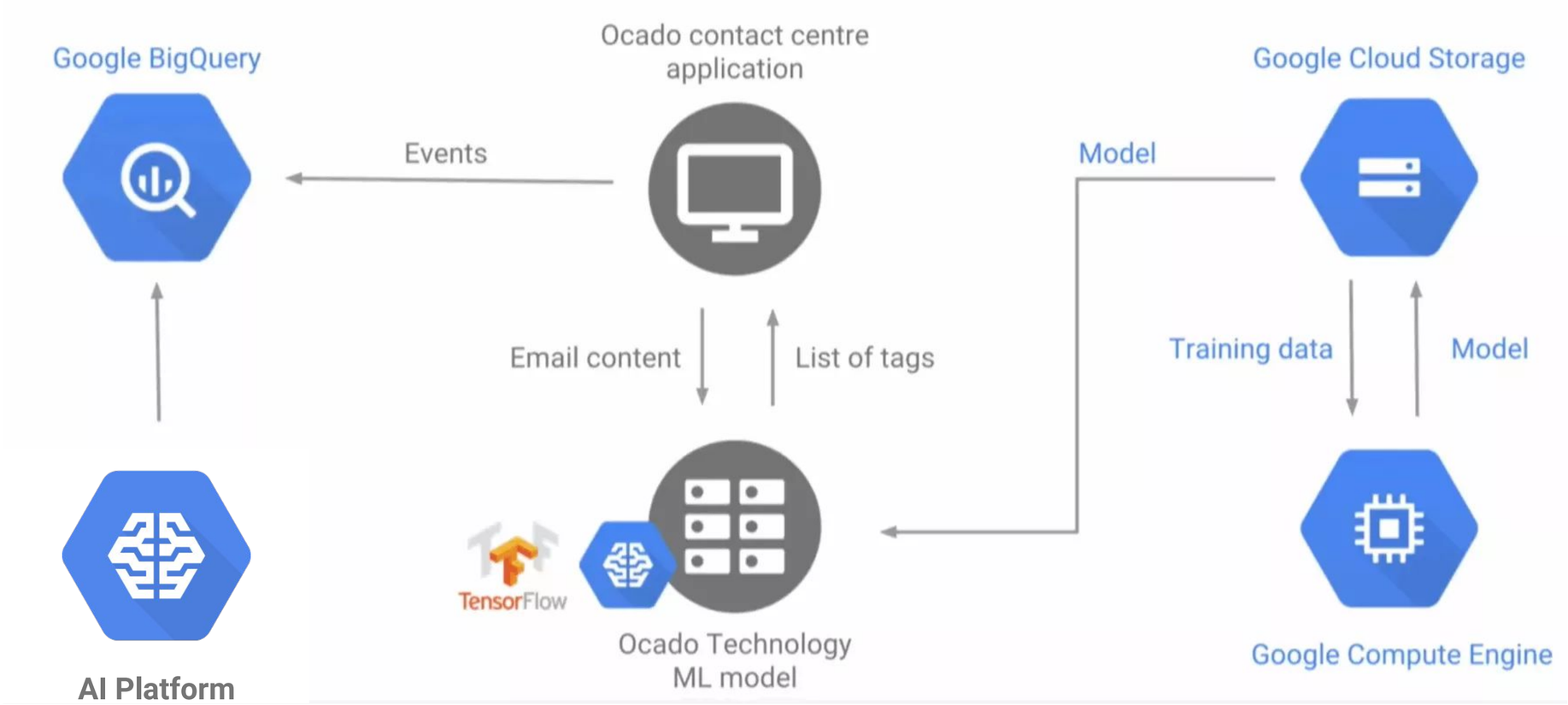
"My delivery is five hours late. Can I get a full refund?"

"I love Ocado! You're the best."

"I scheduled my delivery for 3 PM today. Something came up and I won't be home at that time. Can I reschedule?"

Ocado's GCP solution helps them respond to urgent customer emails 4x faster with ML

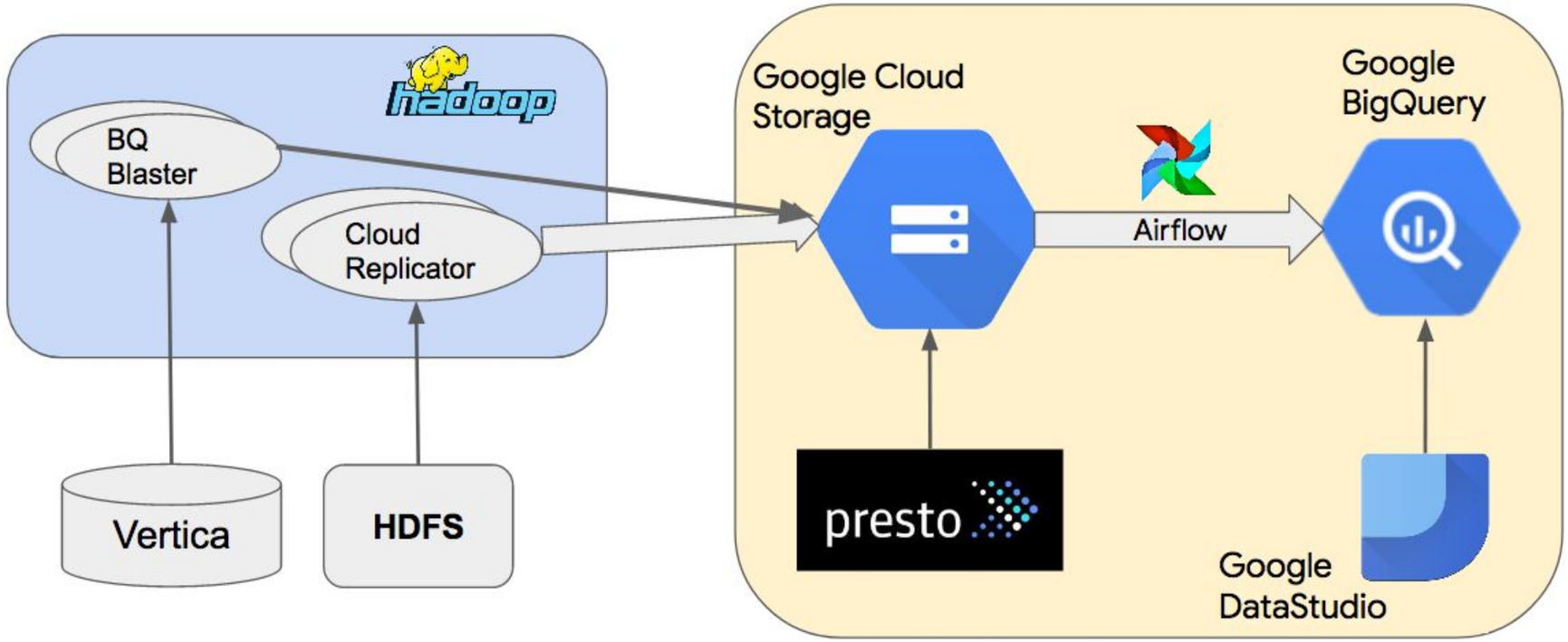
Increased contact center efficiency enables representatives to spend extra time on high-priority tasks



http://www.multichannel-blog.co.uk/2017/05/03/google-the-future-of-cloud-conference-in-london-3-4th-may/

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



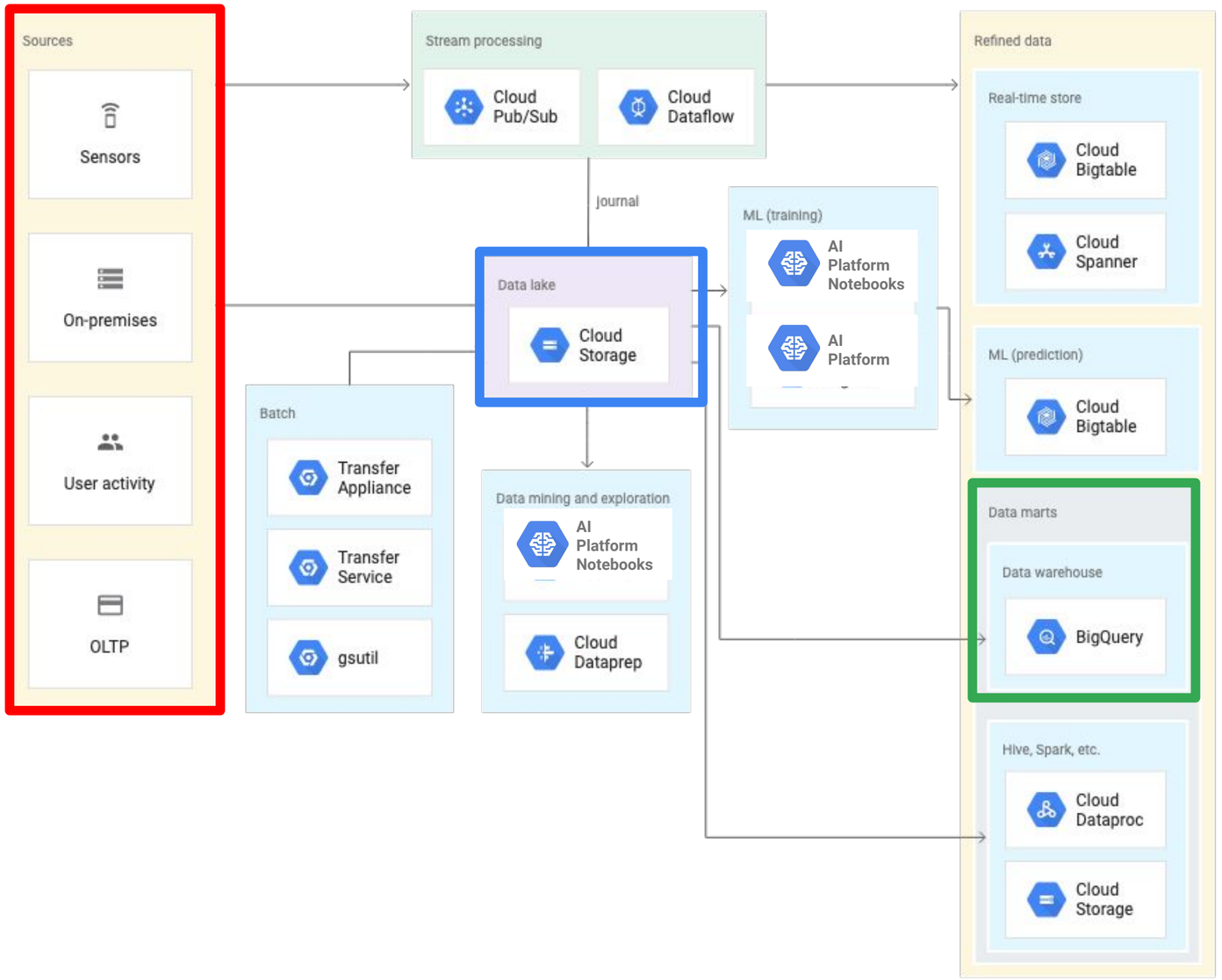
https://blog.twitter.com/engineering/en_us/topics/infrastructure/2019/democratizing-data-analysis-with-google-bigquery.html

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

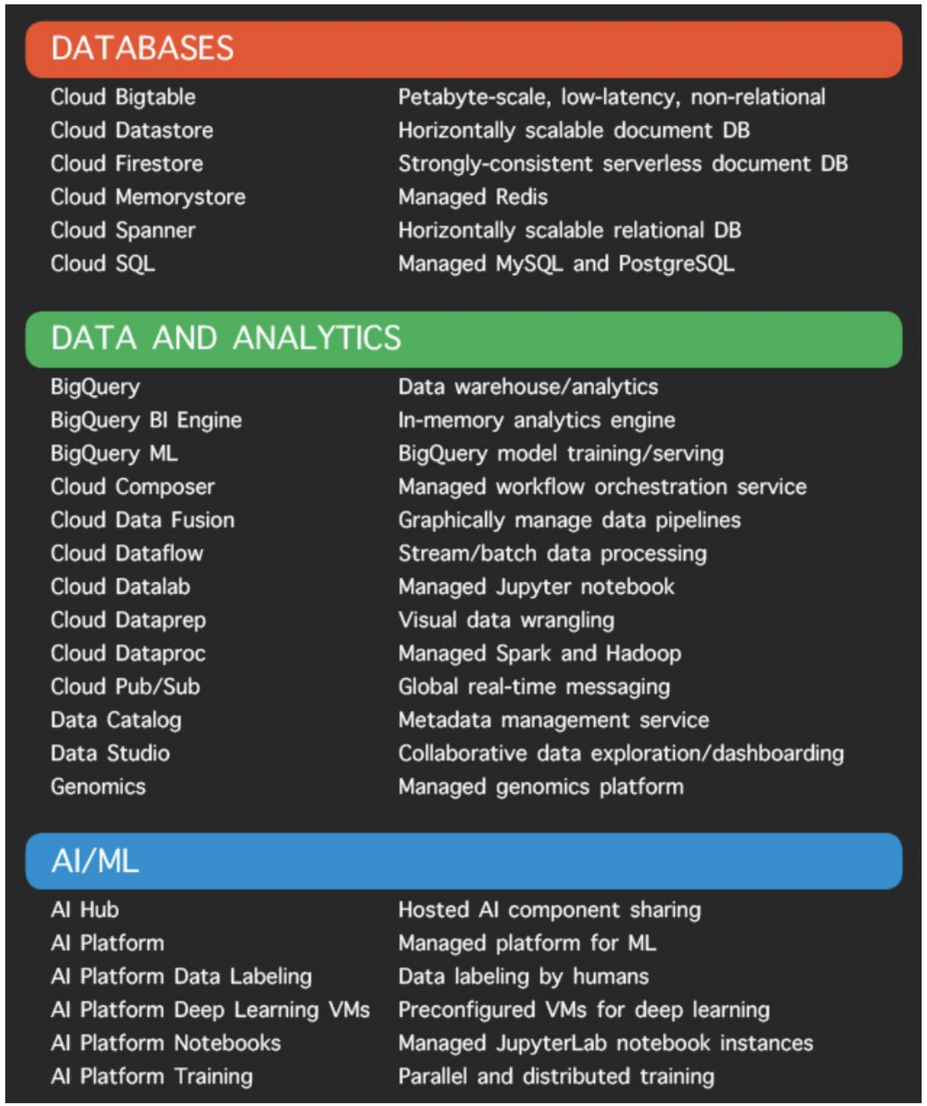


Google Cloud

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

Agenda

Explore the role of a data engineer

Analyze data engineering challenges

Intro to BigQuery

Data Lakes and Data Warehouses

Transactional Databases vs Data Warehouses

Partner effectively with other data teams
- Manage data access and governance
- Build production-ready pipelines

Review GCP customer case study

Lab: Analyzing Data with BigQuery

Lab

Using BigQuery to do Analysis

Objectives

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