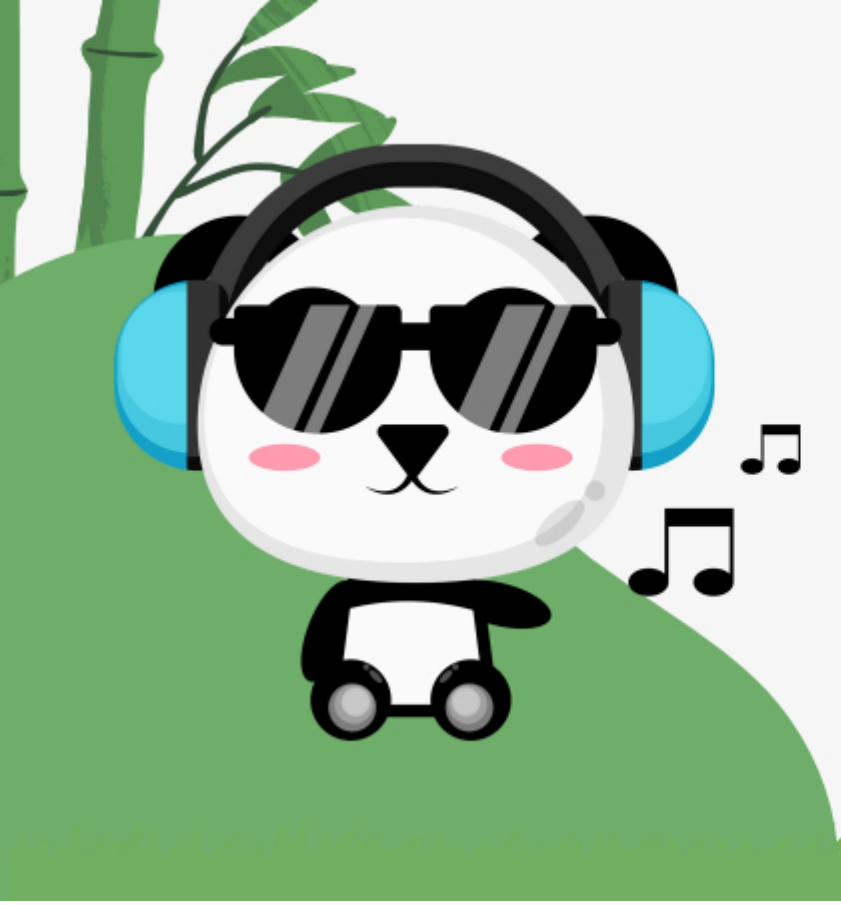


K3

Lesson 16

Data Architecture

One step backward,
think about the data

Learn About

- Data Architectures
- Data Management Systems
- Data Processing Architectures

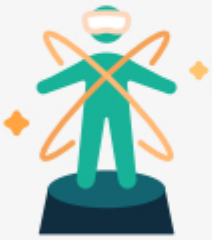
Data Architectures

Concept

Data architectures describe how data is managed

According to IBM

Collection



Transformation

Distribution

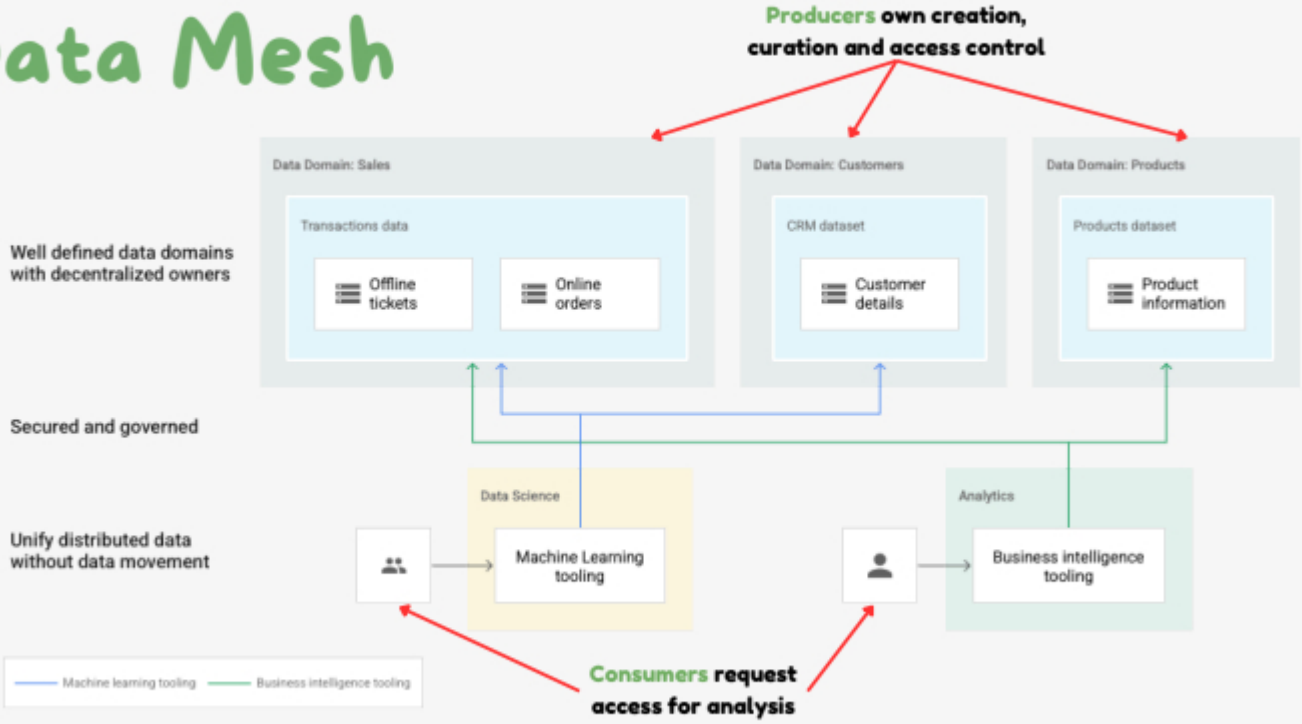
Consumption

Types of Data Architectures

- **Data Fabric:** Unify multiple and disjoint data sources in various environments.
  - Data sources: data warehouses, data lakes, and data marts
  - Environments: on-prem, cloud, and edge

- **Data Mesh:** Distribute data ownership to domain-specific teams.
  - Each team manages, owns, and serves the data as a product

Data Mesh



https://cloud.google.com/dataplex/docs/introduction

Data Fabric

A virtual layer to manage all data
(especially heterogeneous)

https://www.k2view.com/what-is-data-fabric/

Dataplex - A GCP data fabric service

Data Fabric

A virtual layer to manage all data
(especially heterogeneous)

https://www.k2view.com/what-is-data-fabric/

Dataplex - A GCP data fabric service

- **Data Fabric and Data Mesh are NOT mutually exclusive.**
- **Data Fabric focuses on technologies, while Data Mesh concentrates on cultures.**

Data Management Systems

Concept

Software systems to **organize**, **secure**, and make data accessible for authorized users

Organize

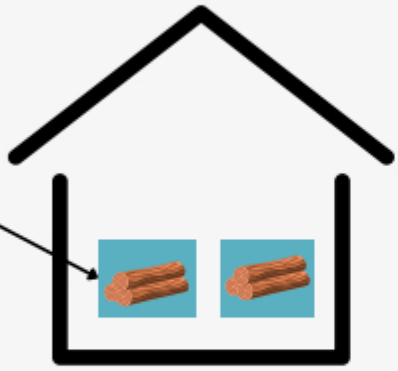
Secure

Accessible

Types of DMS

2. Data Mart

A subset of a warehouse to serve a specific domain.
E.g., sales, accounting



1. Data Warehouse

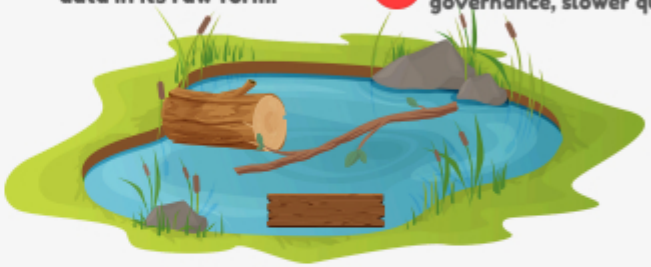
A central data hub containing highly formatted and structured data for analytics
E.g., GCP BigQuery, AWS Redshift, Clickhouse

- ✓ Optimize for structured data, easier for data governance and security
- ✗ Only for structured data, high cost for maintenance, expensive scaling

3. Data Lake

A central location for both structured and unstructured data in its raw form.

- ✓ More flexible and cost effective
- ✗ Hard for security and governance, slower query



4. Lake House

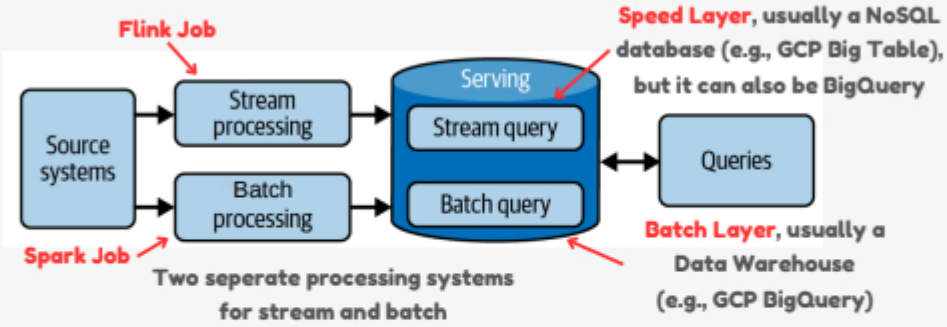
Add layers for data management, governance and query performance on top of Data Lake

- ✓ Address some of Data Lake's issues
- ✗ More complex than the others

Data Processing Architectures

Types of architectures

Lambda Architecture

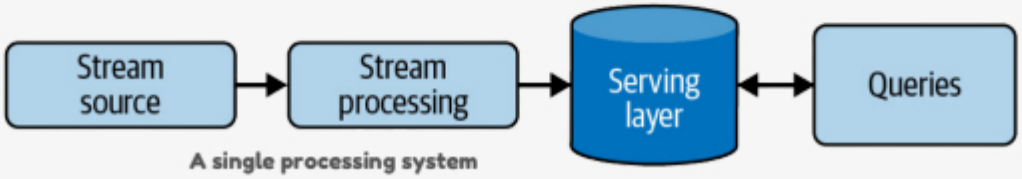


Note:
- Two-separate serving layer
- Often used for historical analysis

✅ Batch processing is easier to code, and it supports complex transformation

❌ Double infra, mismatched between 2 types of processing, 2 code base to maintain

Kappa Architecture



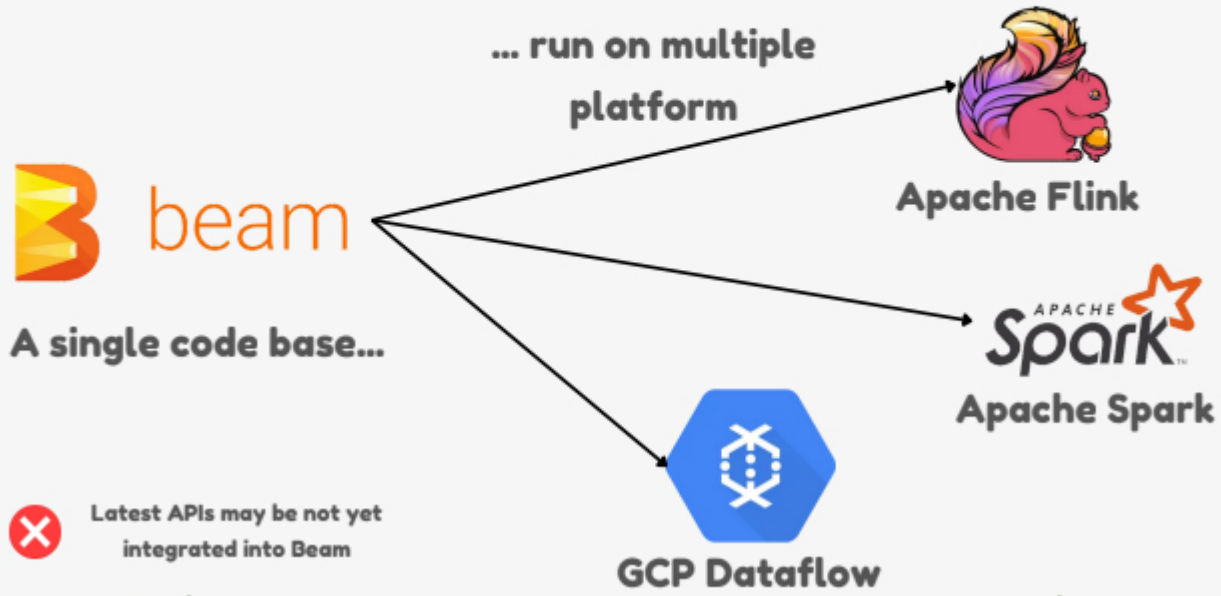
Note:
- Unified serving layer
- Often used for real-time analysis

✅ Single infra and code base

❌ Not easy to implement, complex join (e.g., 30 tables), and expensive for huge volume data

DataFlow Model

Inspired from **"batch is a special case of streaming"** philosophy



References

- https://www.datanami.com/2021/10/25/data-mesh-vs-data-fabric-understanding-the-differences/
- https://medium.com/codex/what-is-a-data-swamp-38b1aed54dc6
- https://www.kai-waehner.de/blog/2021/09/23/real-time-kappa-architecture-mainstream-replacing-batch-lambda/
- Fundamentals of Data Engineering

Thank You!

Created by
Quan Dang