



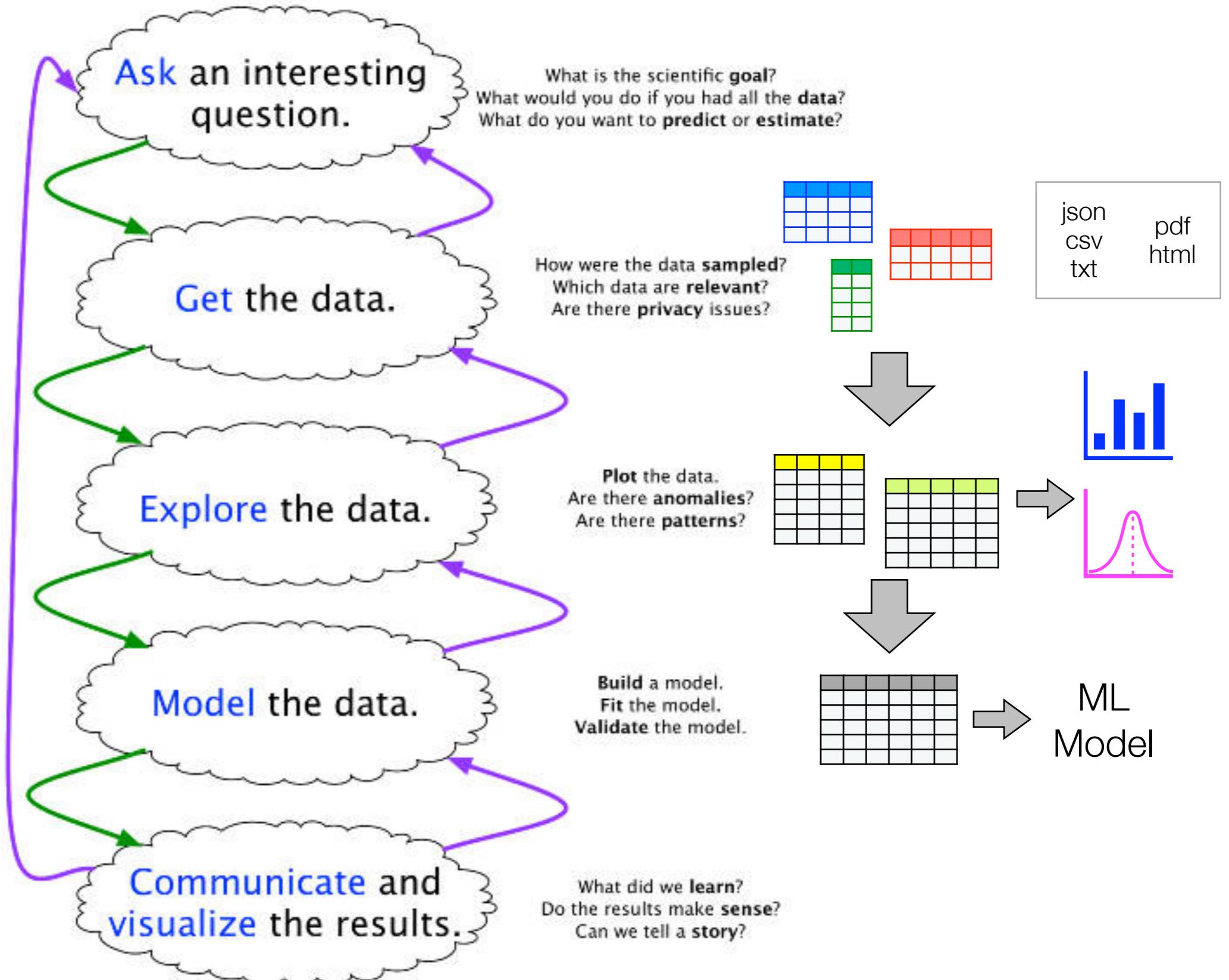
2110446 - Data Science and Data Engineering

## Introduction to Data Engineering

**Asst.Prof. Natawut Nupairoj, Ph.D.**

Department of Computer Engineering  
Chulalongkorn University  
[natawut.n@chula.ac.th](mailto:natawut.n@chula.ac.th)

## The Data Science Process



Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course <http://cs109.org/>.

# THE DATA SCIENCE HIERARCHY OF NEEDS

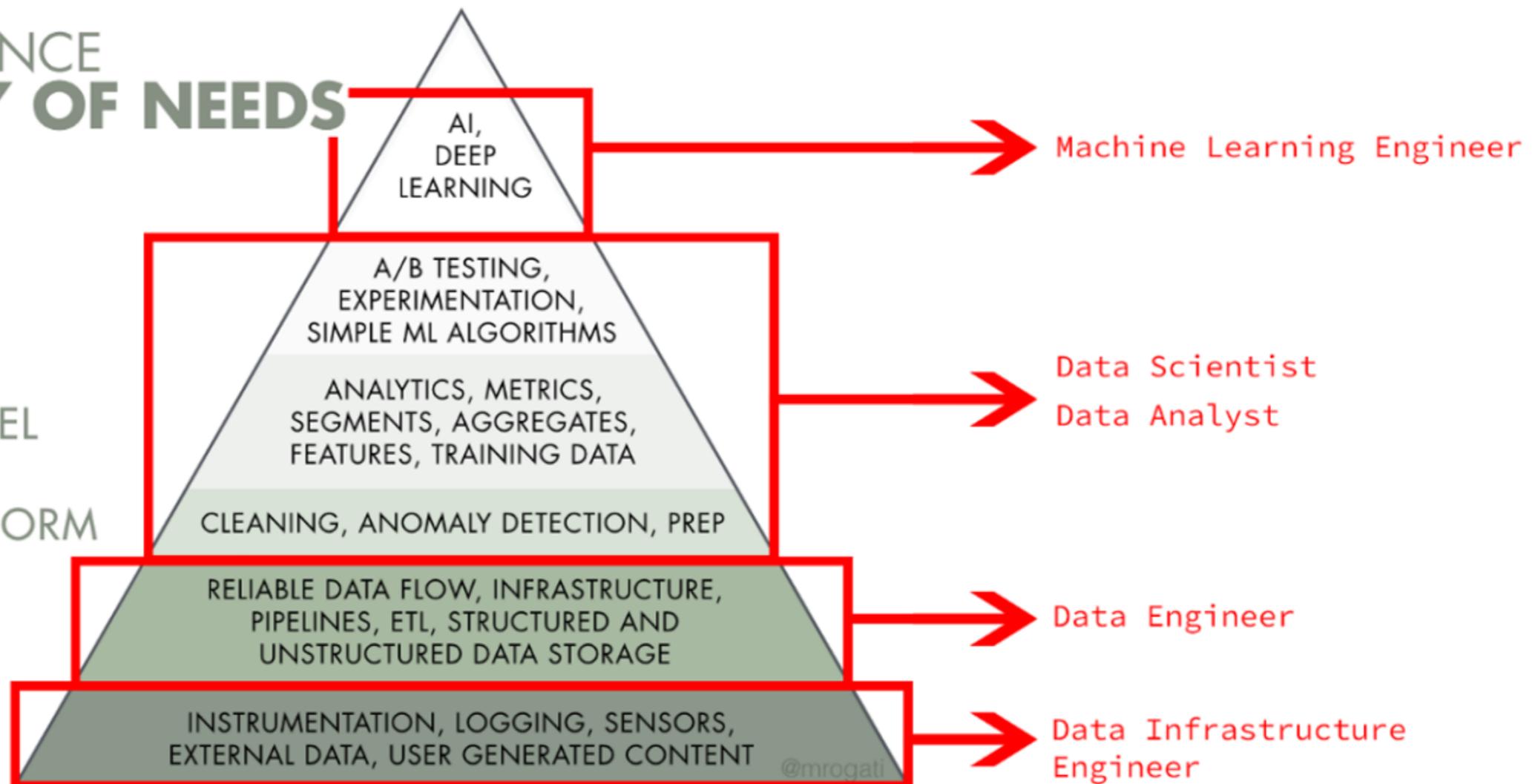
LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT



# What is Data Engineering? (Wikipedia)

---

- Data engineering refers to the building of systems to enable the **collection and usage of data**
- This data is usually used to **enable subsequent analysis** and data science, which often involves **machine learning**
- Making the data usable usually involves substantial compute and storage, as well as data processing

# The Data Engineer Job Market in 2024 [Research on 1,000 Job Postings]

*Join over 2 million students who advanced their careers with 365 Data Science. Learn from instructors who have worked at Meta, Spotify, Google, IKEA, Netflix, and Coca-Cola and master Python, SQL, Excel, machine learning, data analysis, AI fundamentals, and more.*

**Start for Free**

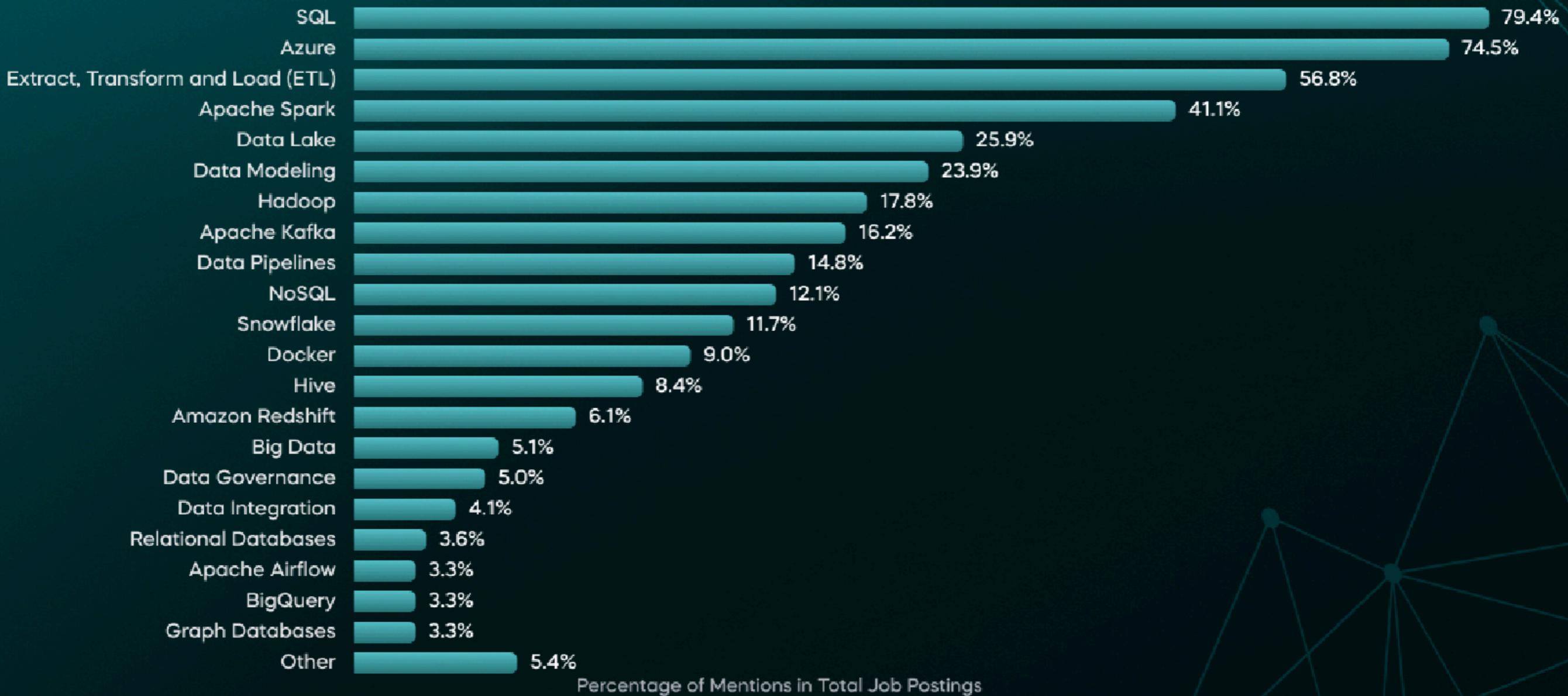


Sophie Magnet • 11 Sep 2024 • 18 min read

With the emergence of ChatGPT, recent years have seen concerns about how artificial intelligence (AI) could lead to job losses in data-related fields like data science and data engineering. There was speculation that AI could replace these roles because of its capacity to automate tasks traditionally performed by humans.

Source: <https://365datascience.com/career-advice/data-engineer-job-market/>

# Data Engineering Skills



# Data Engineering Skill Groups

---

- Database Management
- Big Data Technologies
- Data Warehouse / ETL
- Cloud
- Data Pipeline and Workflow Management
- Data Governance
- Containerization (e.g. docker, k8s)

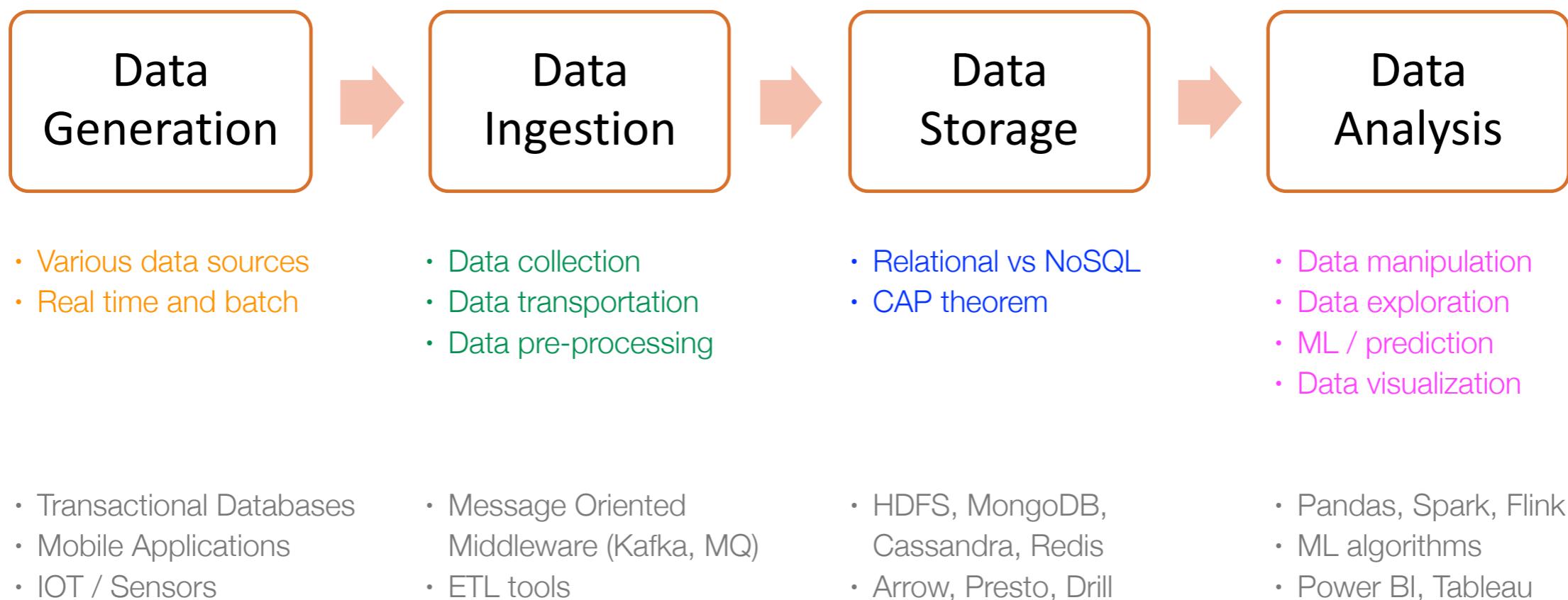


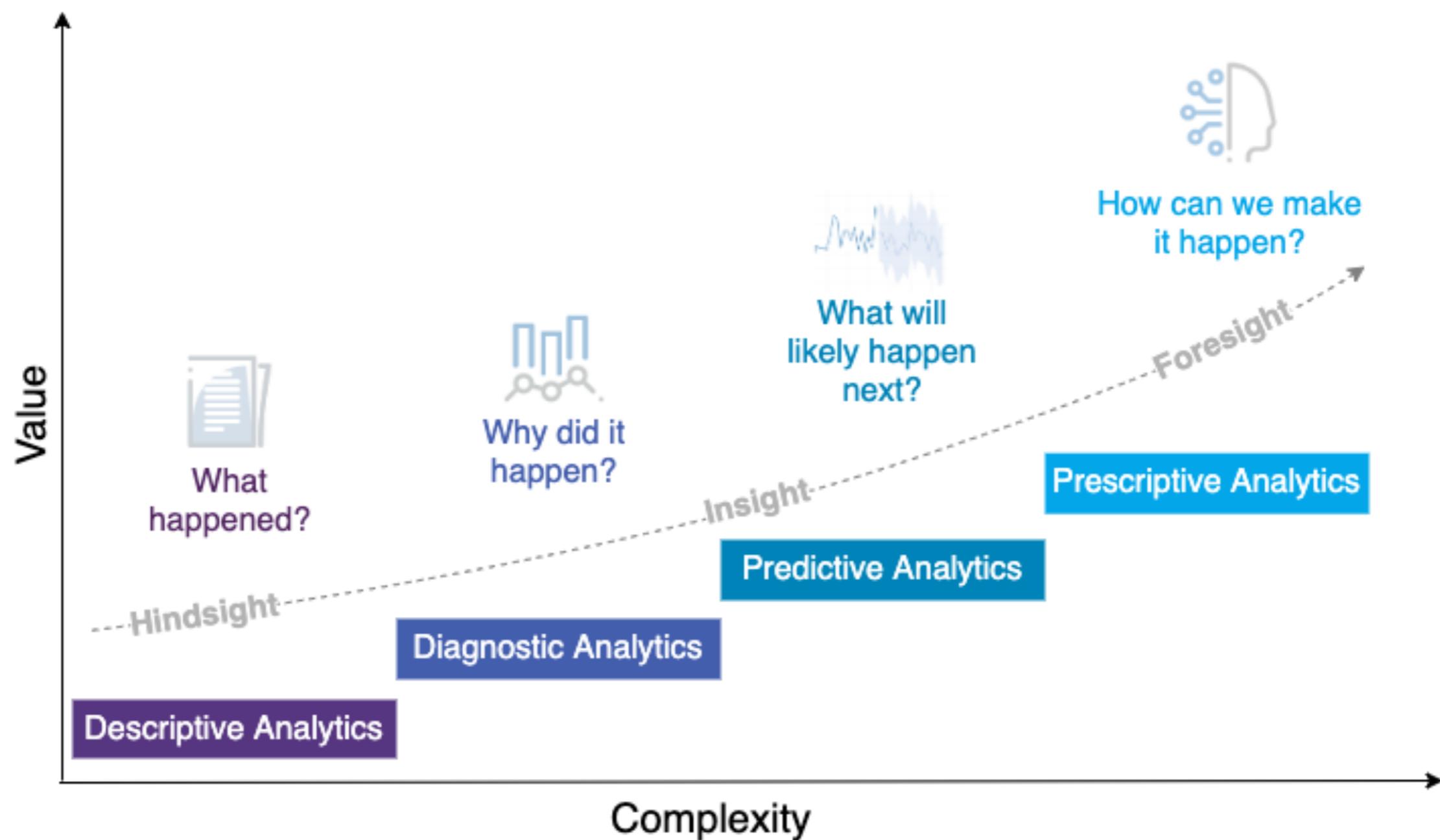
# Data Pipeline Analogy



# Data Lifecycle

---





The 4 types of Data Analytics (Adapted from Davenport & Harris 2007 / Gartner 2012)

# Data Analytics Simplified

---

Descriptive

“A.Natawut drinks about 1 cup of coffee a day”

Diagnostic

“Number of cups that A.Natawut drinks depends on number of meetings he has each day”

Predictive

“Tomorrow, A.Natawut has 2 meetings. It is very likely that A.Natawut will drink 2 cups.”

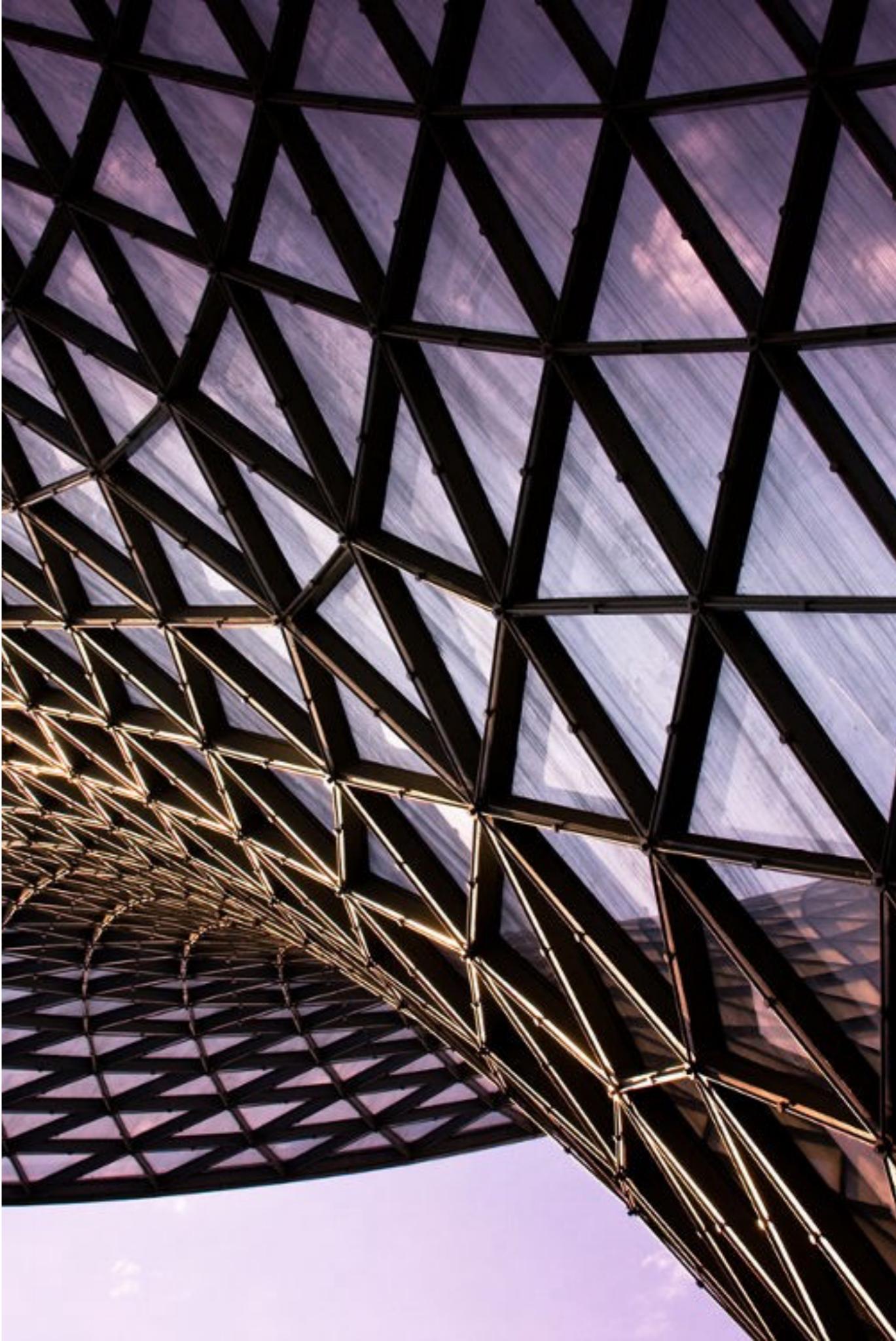
Prescriptive

“Inform secretary to prepare one cup in the morning and one in the afternoon for A.Natawut”

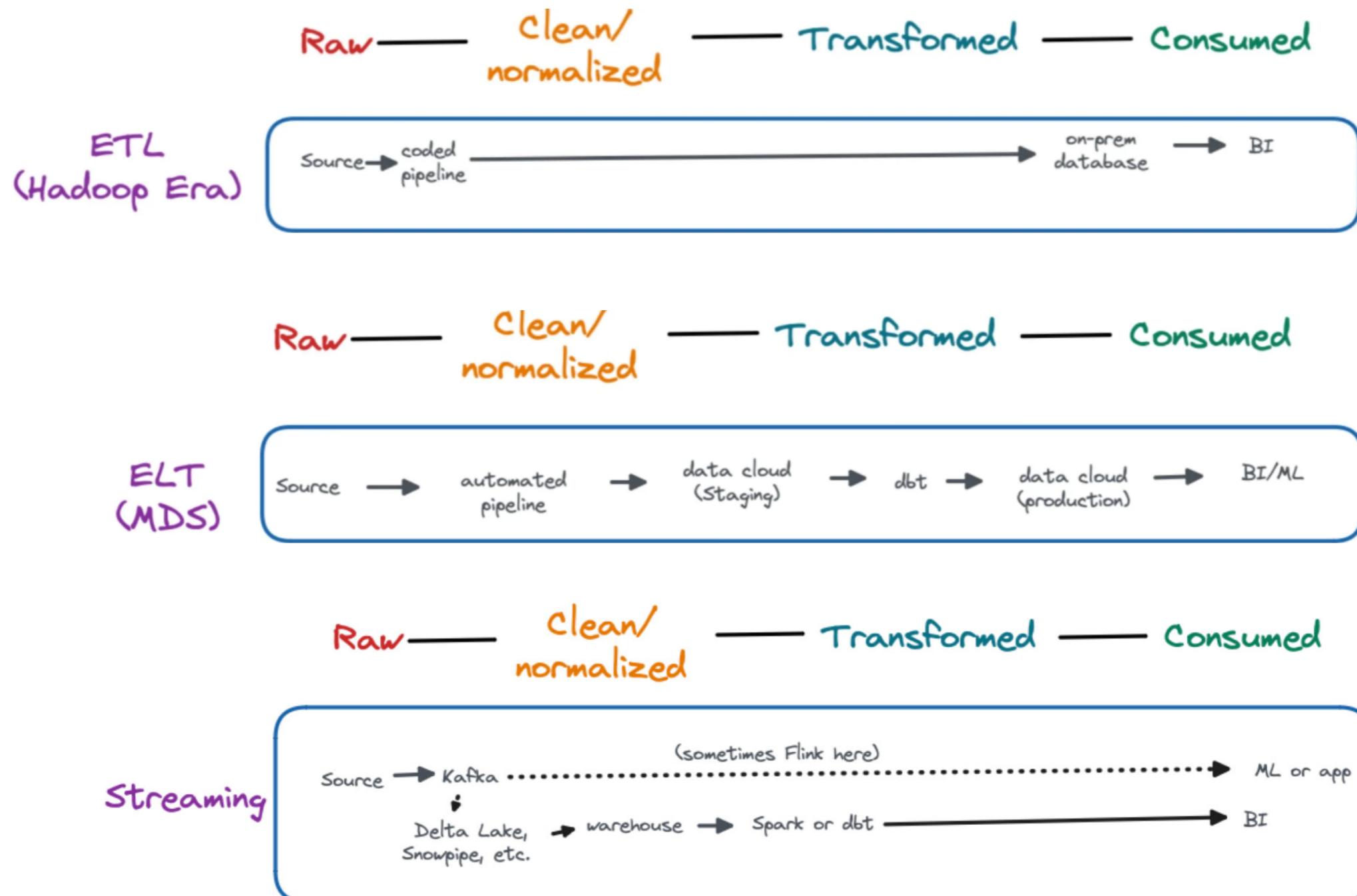
# Data Pipeline and Data Platform Architectural Patterns

---

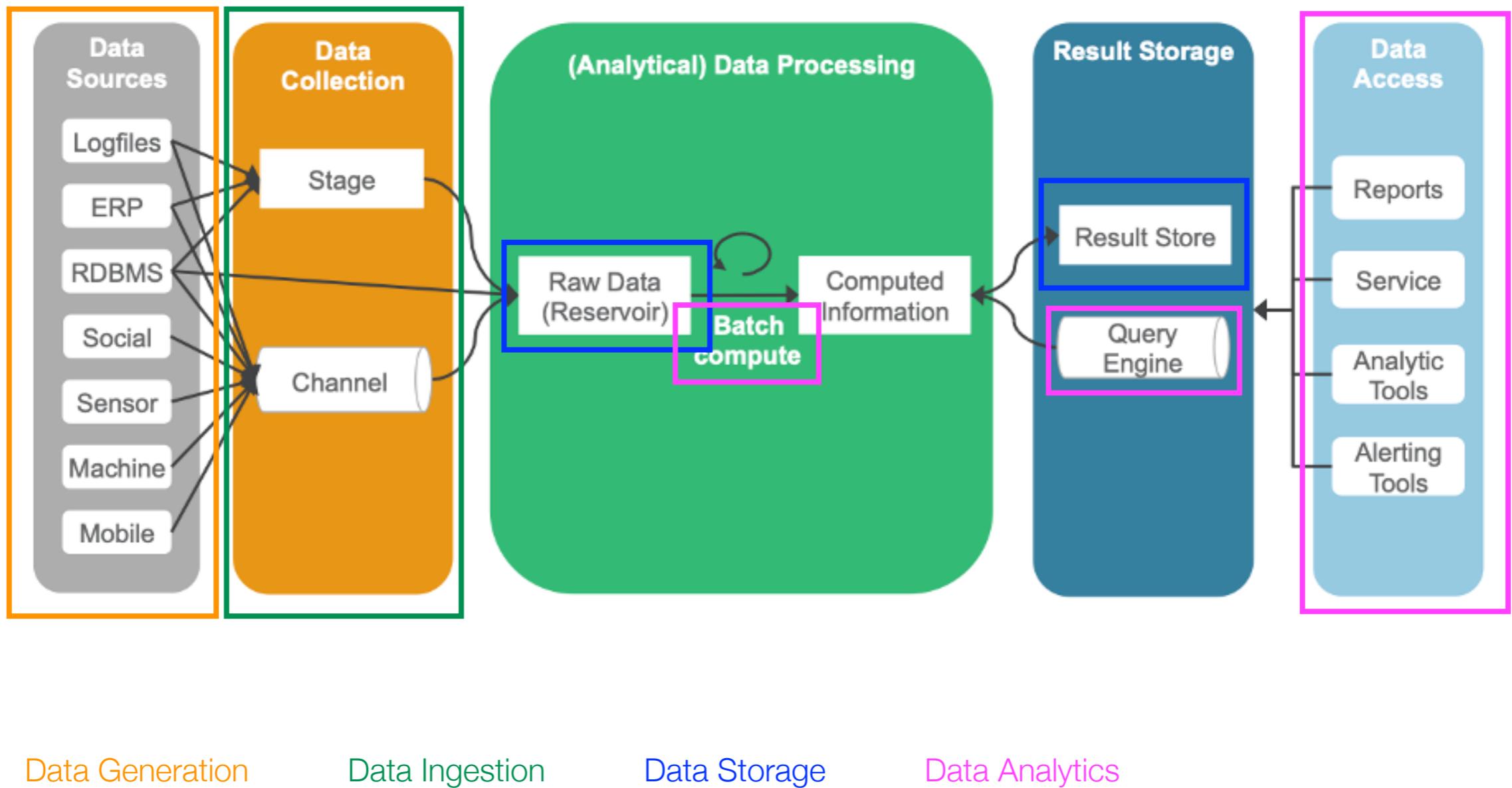
Common design patterns being  
used in data science



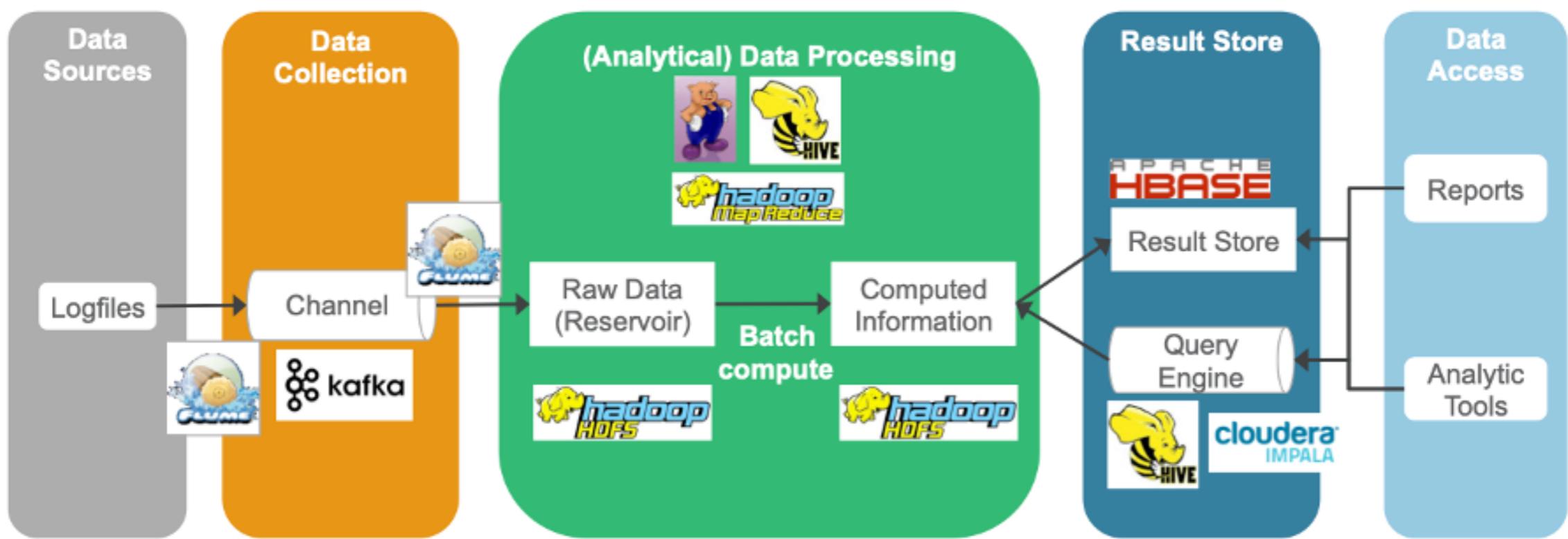
# Basic Design Patterns



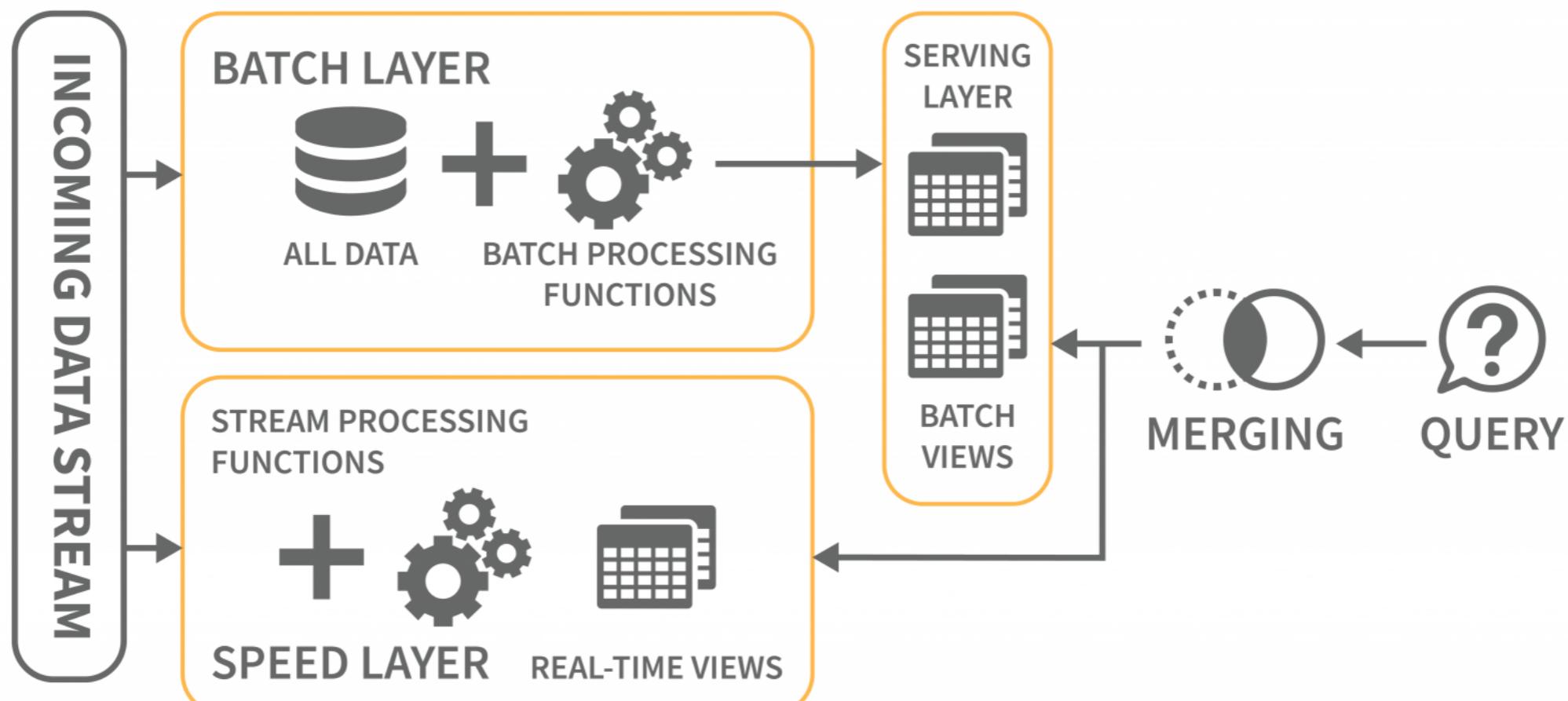
# Simple Big Data Analytic Architecture



# Example: Facebook Data Pipeline (early days)



# Lambda Architecture (Nathan Marz)



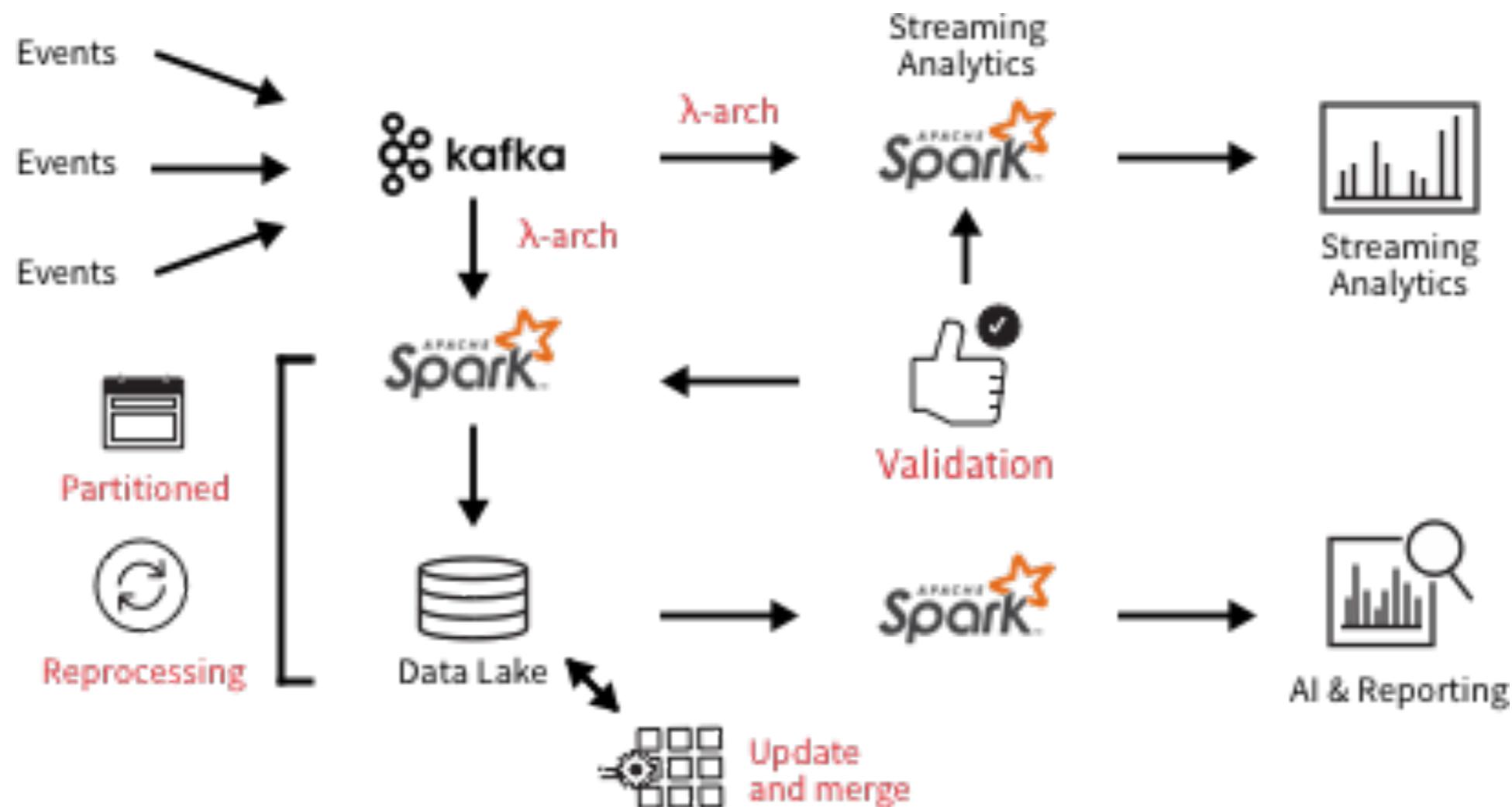
InterviewBit

# Principles of Lambda Architecture

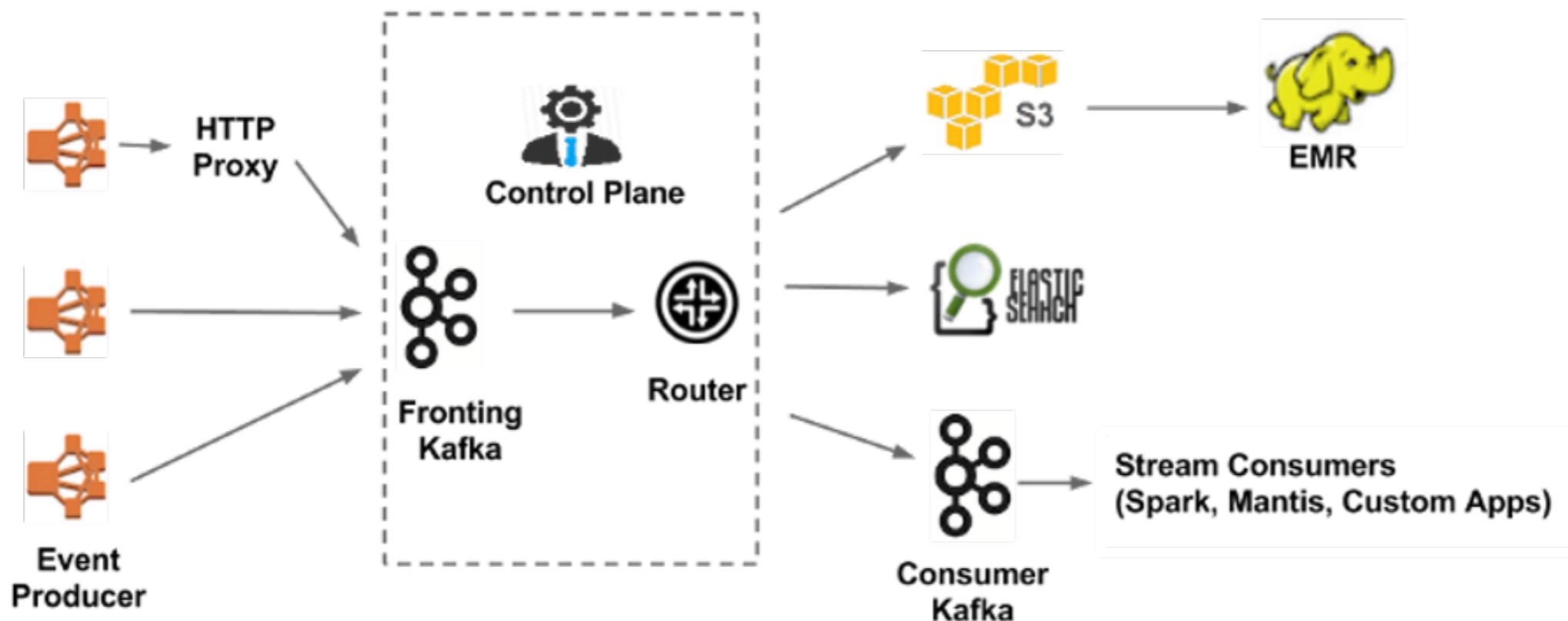
---

- Assume on a data model with an append-only, immutable data source that serves as a system of record
- Intended for ingesting and processing timestamped events that are appended to existing events rather than overwriting them
- Balance latency, throughput, and fault-tolerance
  - Batch processing to provide comprehensive and accurate views of batch data
  - Real-time stream processing to provide views of online data

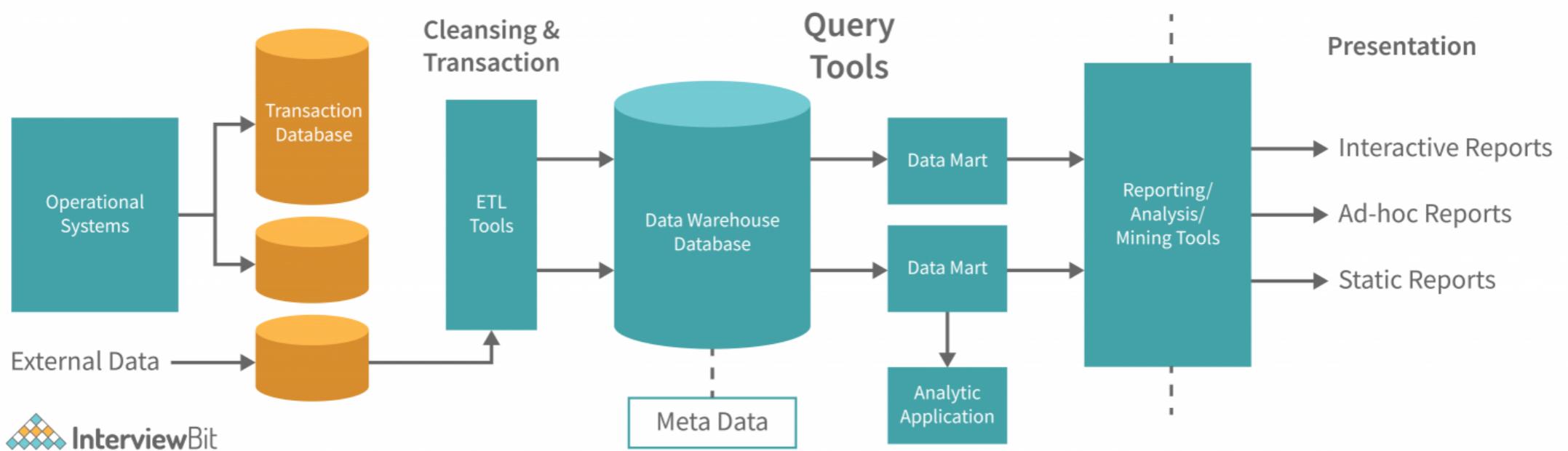
# Lambda Architecture in Practice



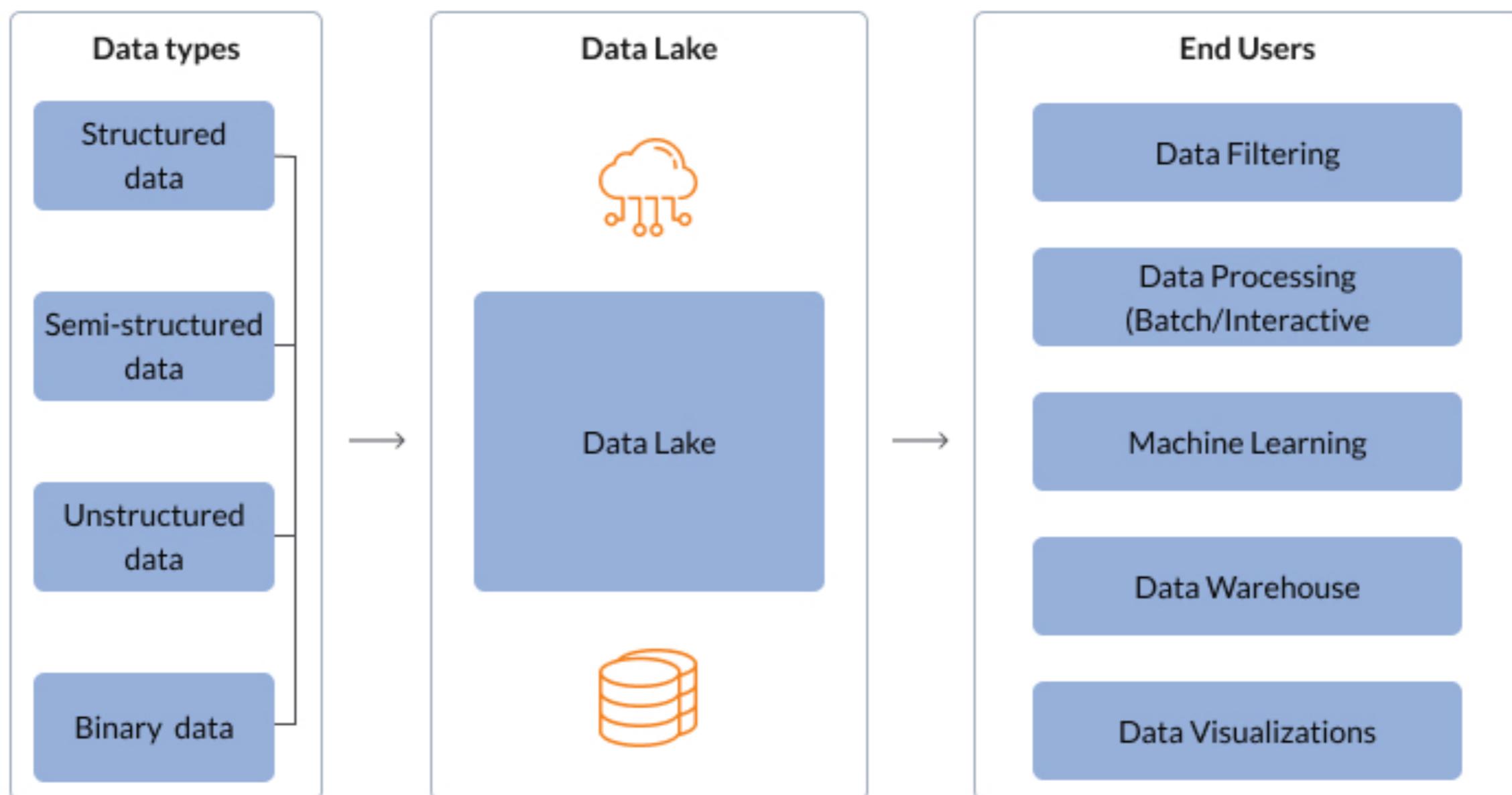
# Example: Netflix's Keystone Architecture



# Data Platform: Data Warehouse

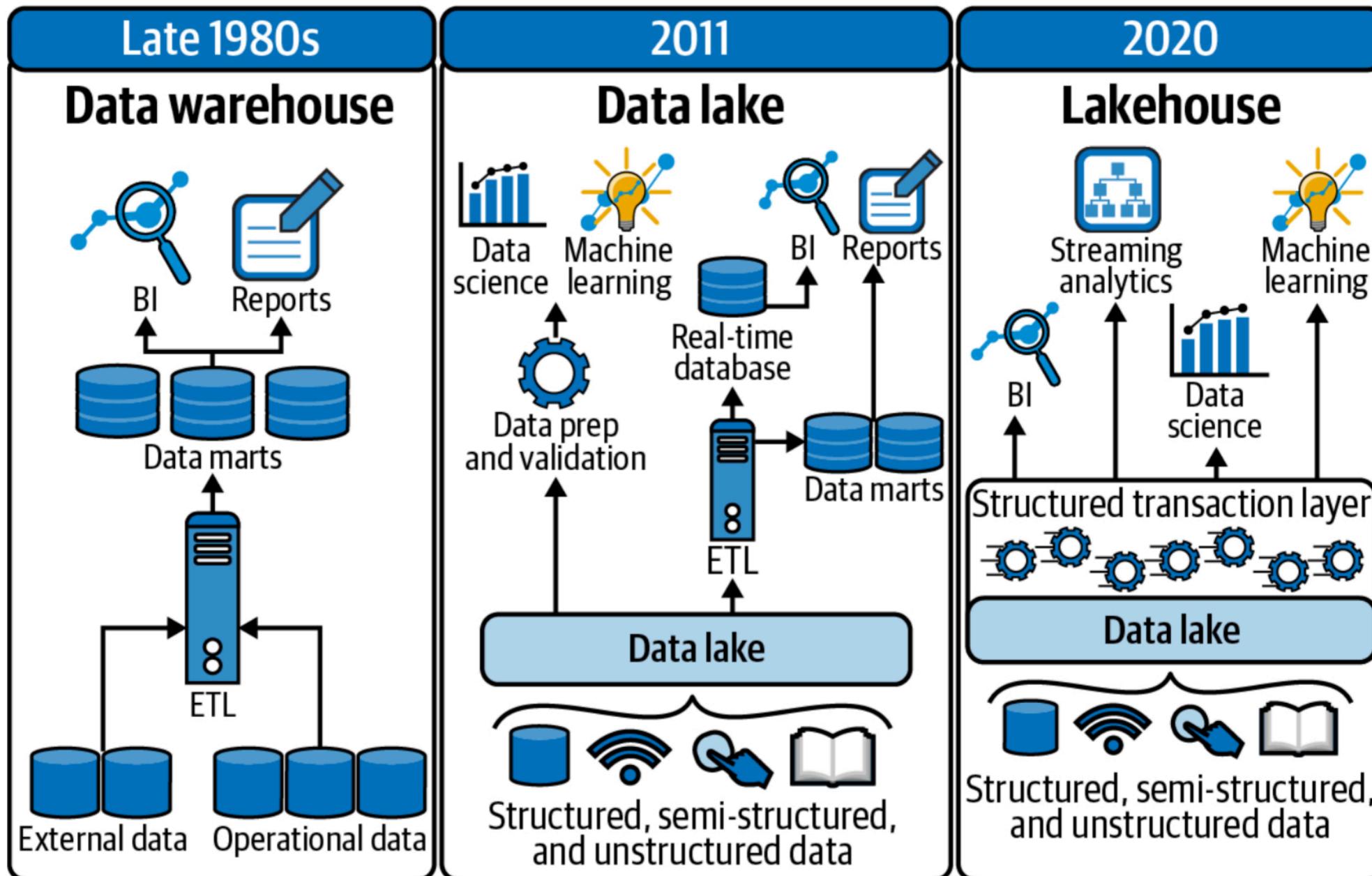


# Data Platform: Data Lake



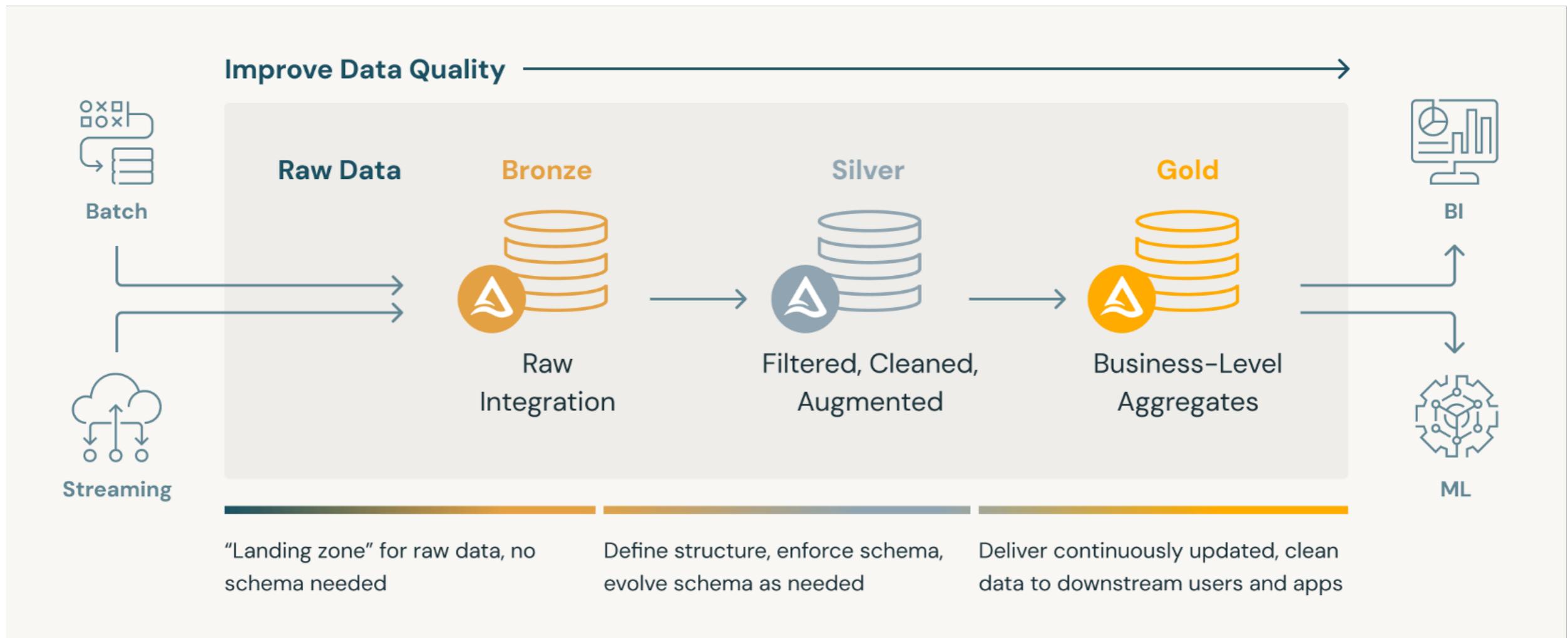
Source: <https://www.n-ix.com/data-lake-vs-data-warehouse/>

# Data Platform: Data Lakehouse



Source: Alice LaPlante, “The Modern Cloud Data Platform”

# Principles of Data Lakehouse Architecture



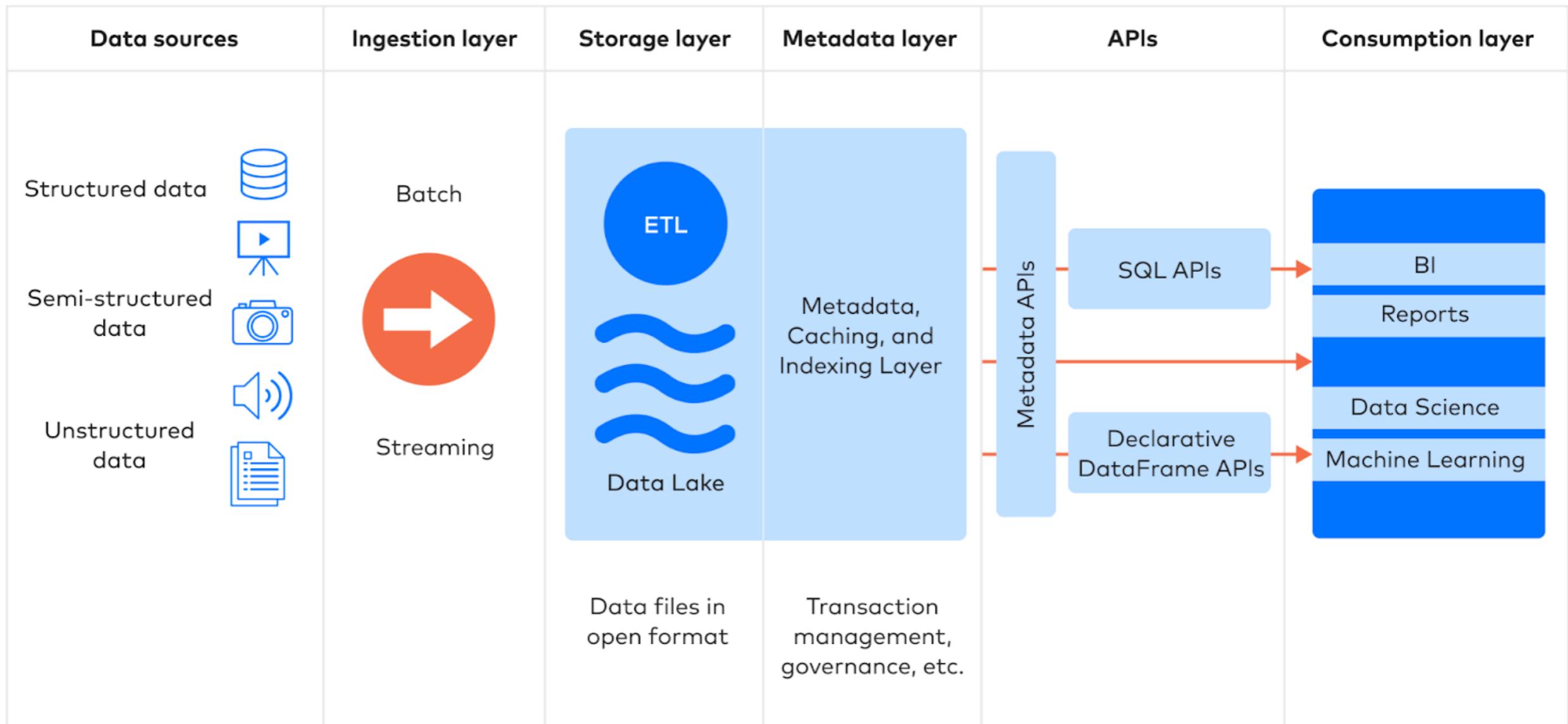
- All the processing and enrichment of data from Bronze (raw data) to Silver (cleansed, filtered) to Gold (structured and ready to be used by analytics, reporting, and data science) happens within Delta Lake

# Principles of Data Lakehouse Architecture

---

- Lambda architecture is complex and requires data validation between real-time and batch processing paths
- Data Lakehouse architecture focuses on simplicity by processing data in the data store from bronze to silver to gold
- Built-in data governance including ACID transaction control, schema enforcement, etc.
  - Recording all changes and commits in transaction log ensures data integrity and reliability
  - Each write creates a new data version, old version can be view and revert (rollback)
- Single source of truth as there is only one path

# Data Lakehouse Architecture



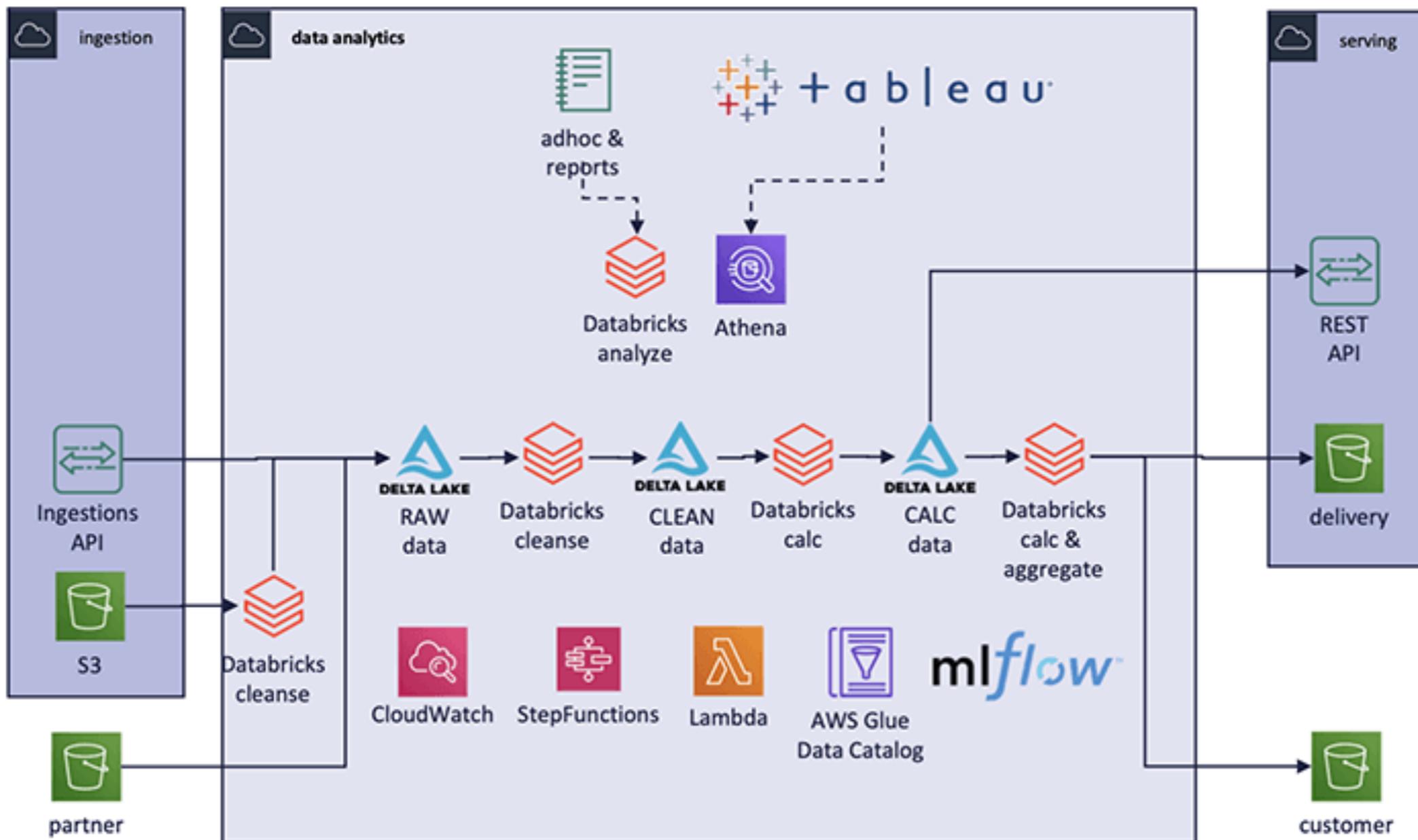
Source: <https://www.fivetran.com/blog/what-is-a-data-lakehouse>

## Example: wetter.com

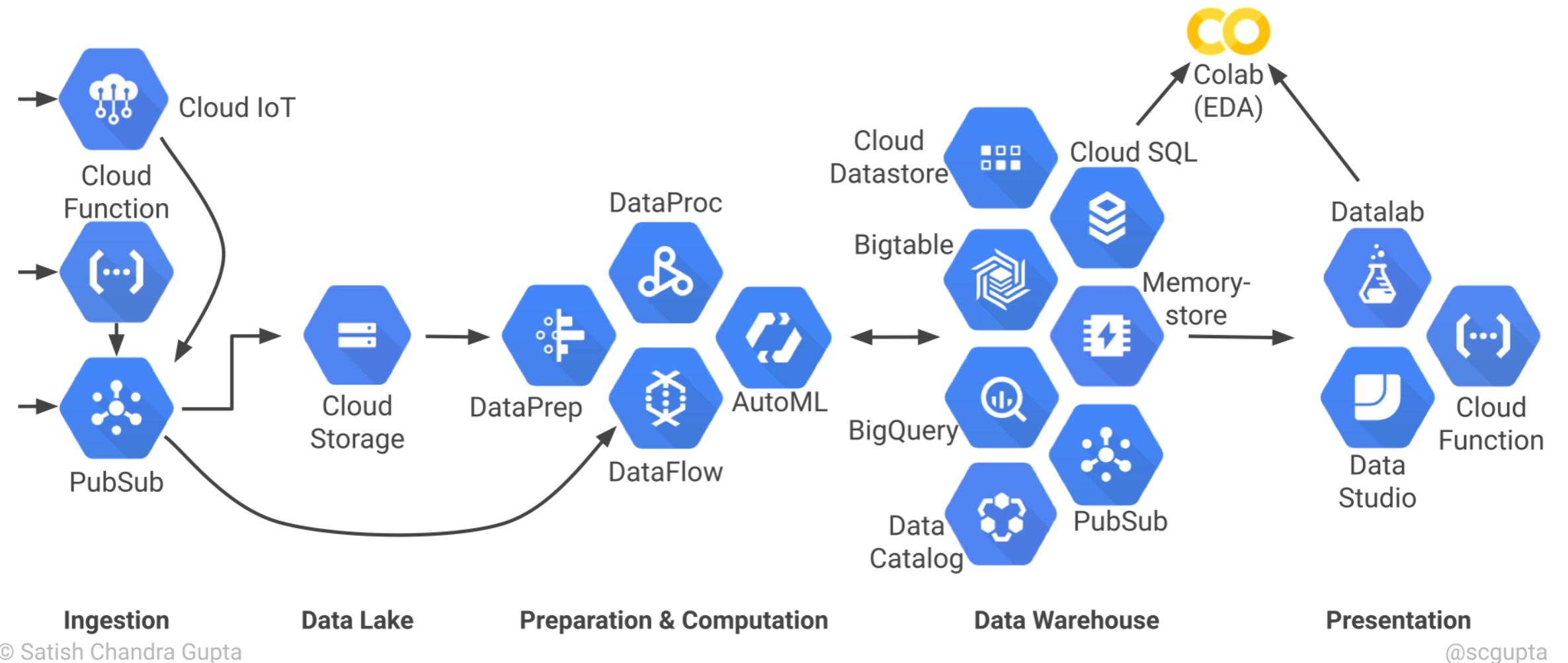
---

- Germany weather portal with up to 20 million monthly unique users
- Monetize data by selling data-products to business customers
- Decode the interrelation between weather, consumer behavior and many other factors used by clients in retail, FMCG, e-commerce, tourism, food and advertising
- Switch from multiple steps Spark/S3 data pipeline to Data Lakehouse architecture

# Example: wetter.com



# Cloud Serverless Architecture



# Conclusion

---

- Data engineering is a process that makes data usable e.g. ready for analytics or ML modeling
- Typical data lifecycle = Generation / Ingestion / Storage / Analysis
- Data pipeline is a workflow consisting of one or more tasks that ingest, move, and transform raw data from one or more sources to a destination
- Existing Data Platform: data warehouse, data lake, and data lakehouse

# References

---

- D. Reinsel, J. Gantz, and J. Rydning, “Data Age 2025: The Digitization of the World From Edge to Core,” International Data Corporation, 2018.
- A. Menon, “Big data@ facebook,” in Proceedings of the 2012 workshop on Management of big data systems, 2012, pp. 31–32.
- J. Warren and N. Marz, Big Data: Principles and best practices of scalable realtime data systems. Simon and Schuster, 2015.
- “Data Warehouse Architecture – Detailed Explanation”, <https://www.interviewbit.com/blog/data-warehouse-architecture/>
- H. Leano, “How to Evaluate Data Pipelines for Cost to Performance”, <https://www.databricks.com/blog/2020/11/13/how-to-evaluate-data-pipelines-for-cost-to-performance.html>
- Data Engineering Wiki, “Data Pipeline”, <https://dataengineering.wiki/Concepts/Data+Pipeline>