

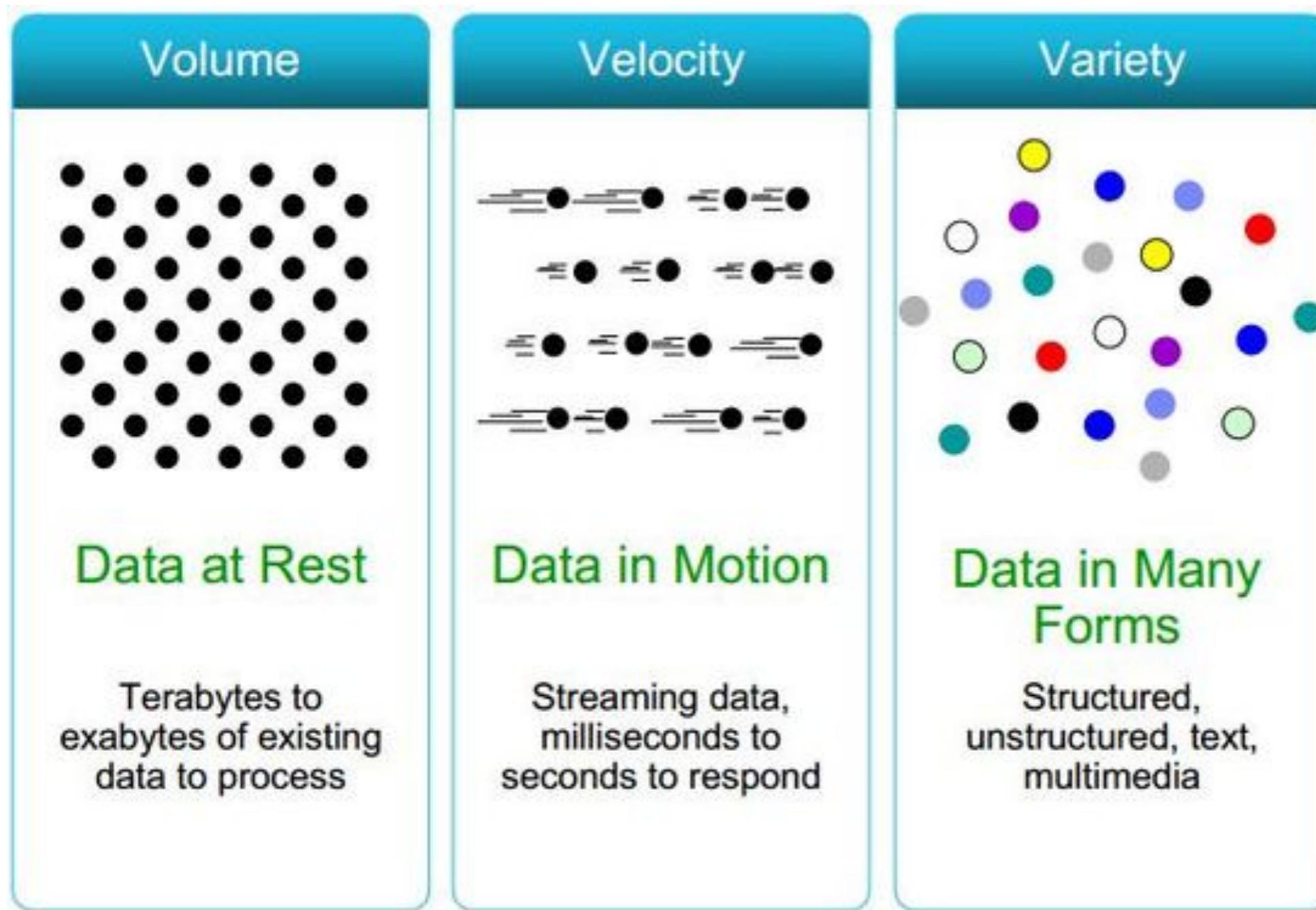


2110531 - Data Science and Data Engineering Tools

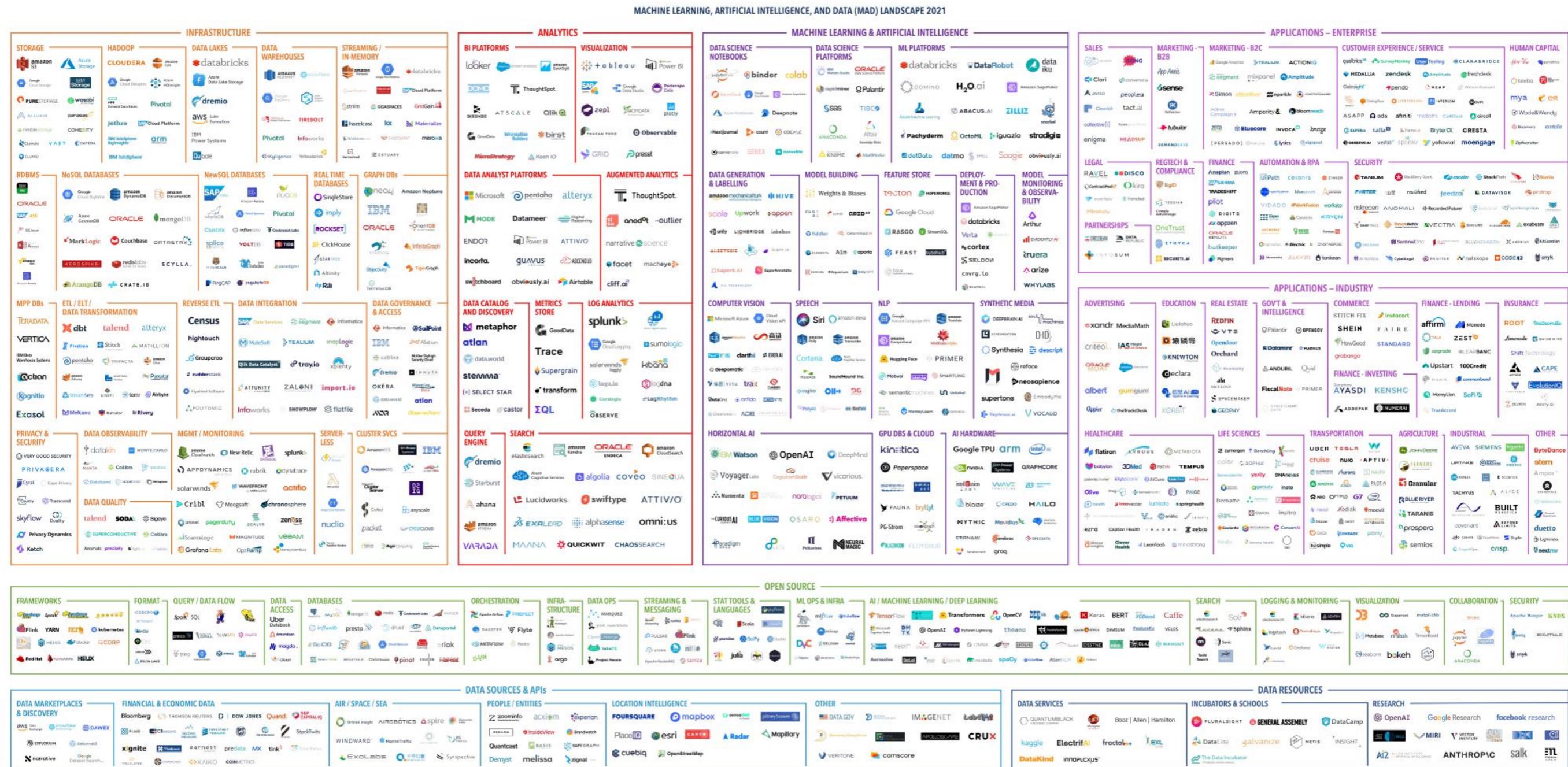
Big Data Architecture

Asst.Prof. Natawut Nupairoj, Ph.D.
Department of Computer Engineering
Chulalongkorn University
natawut.n@chula.ac.th

Back to Basic - Data Characteristics - 3V



Source: IBM



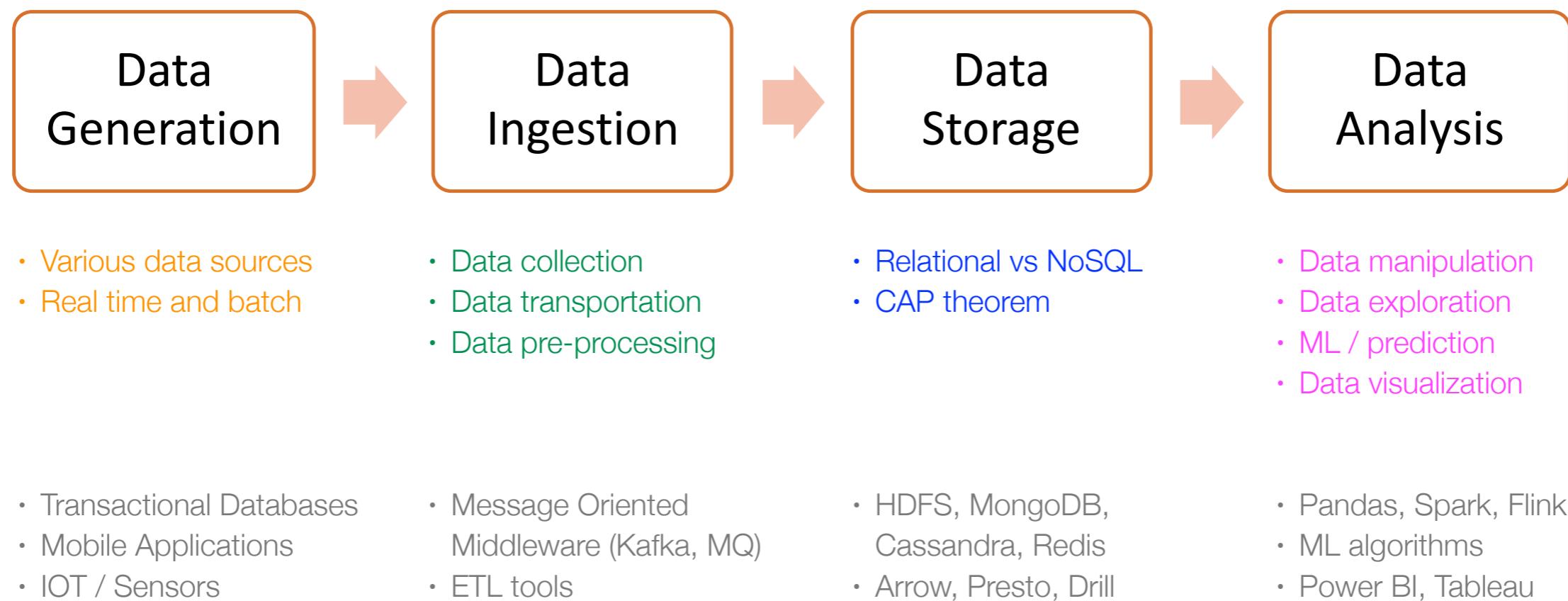




Data Pipeline Analogy

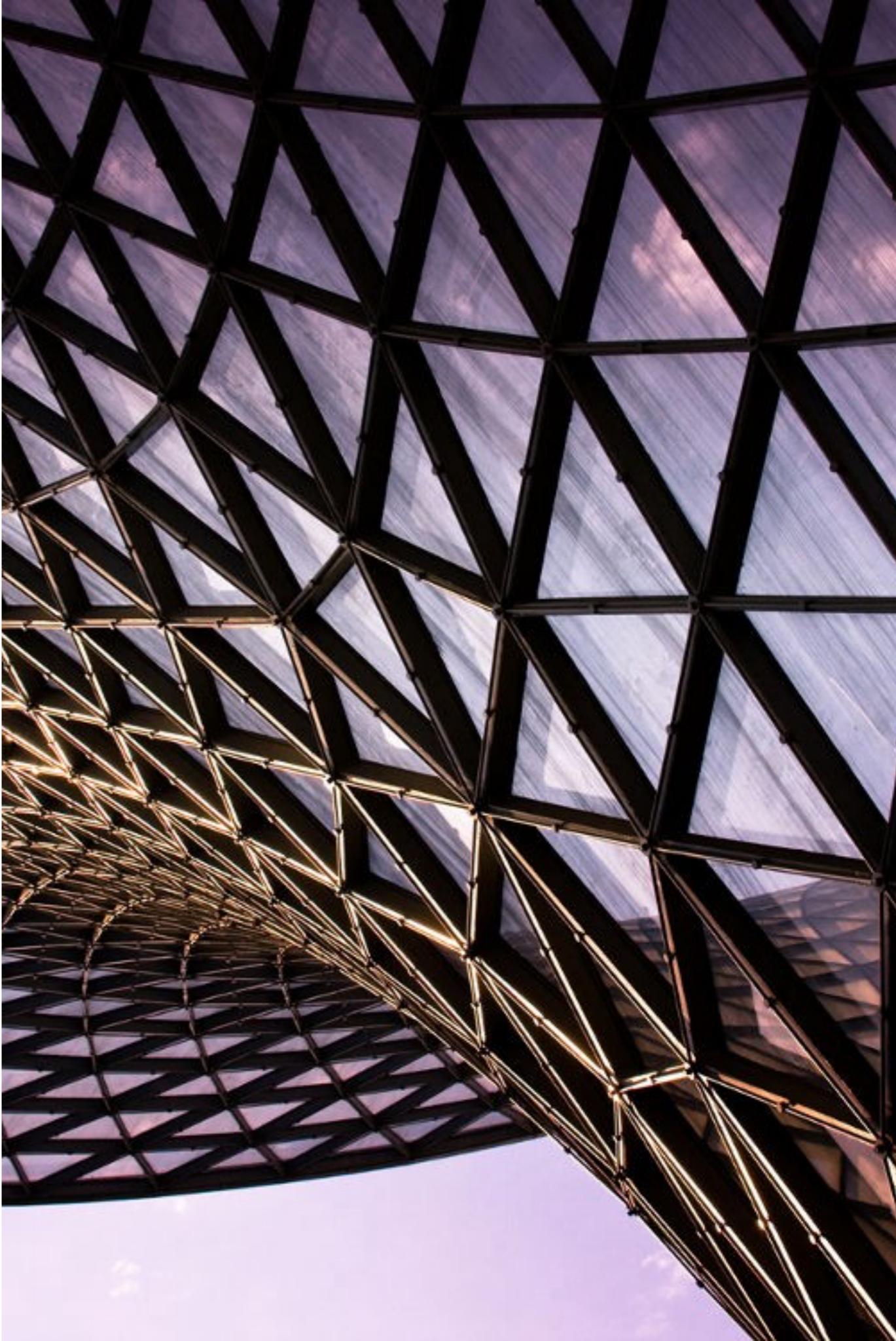


Data Science Lifecycle



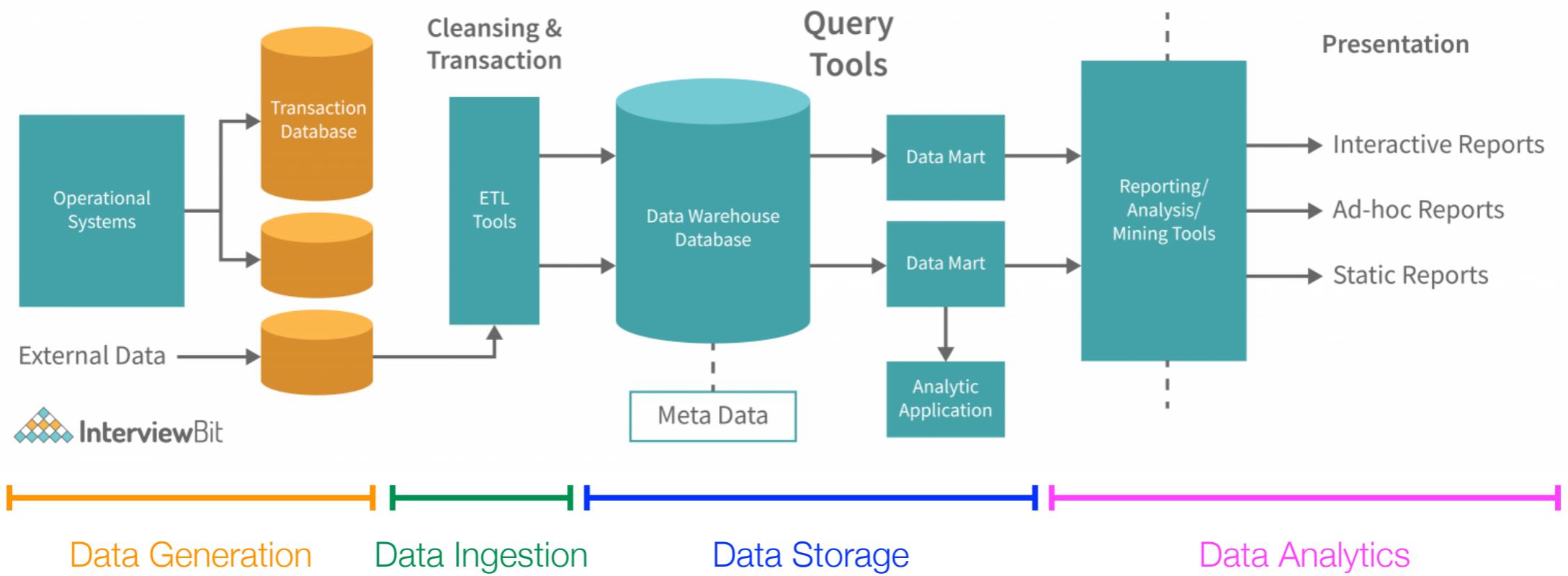
Architectural Patterns

Common design patterns being used in data science

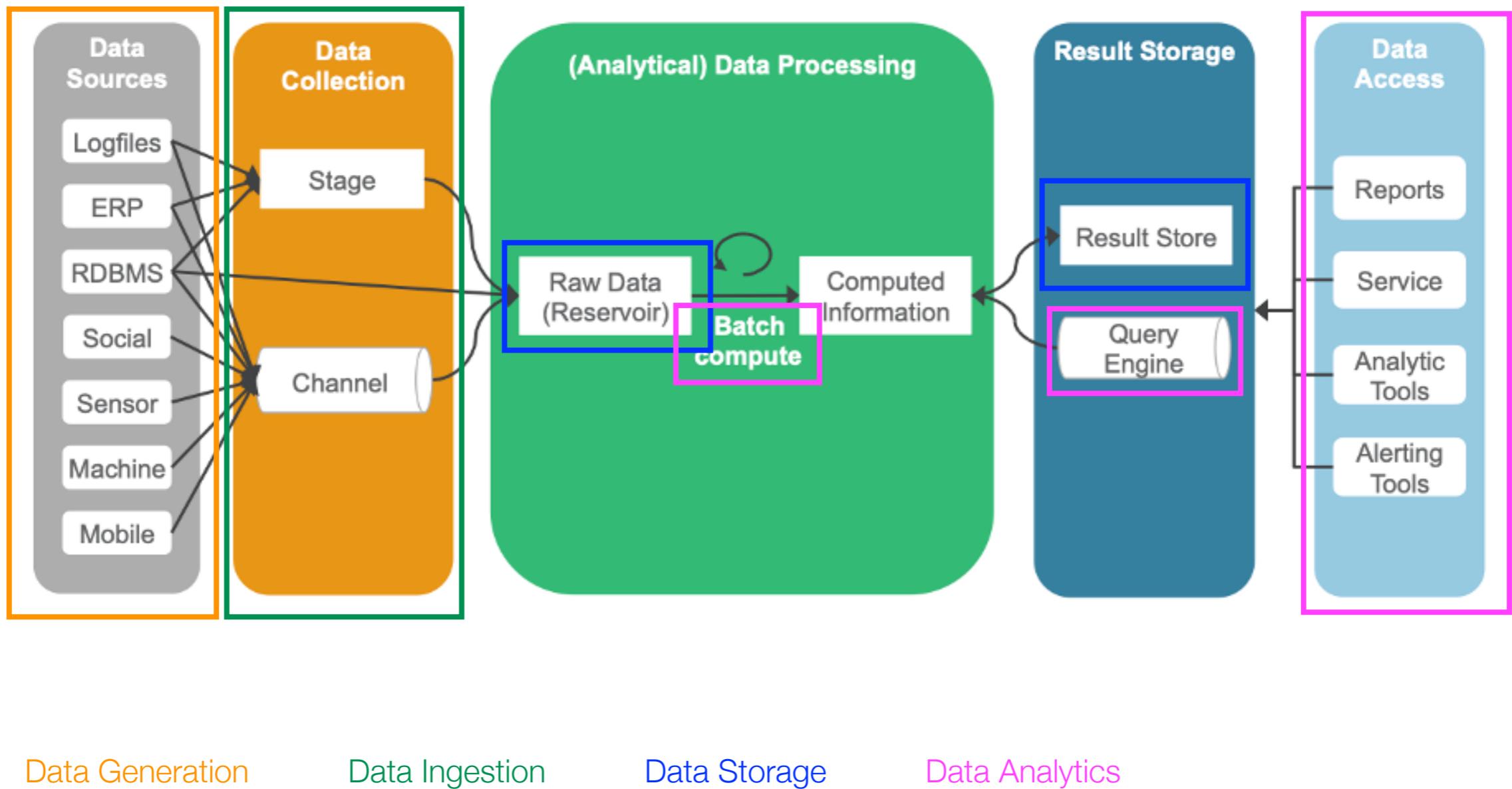


Data Warehouse

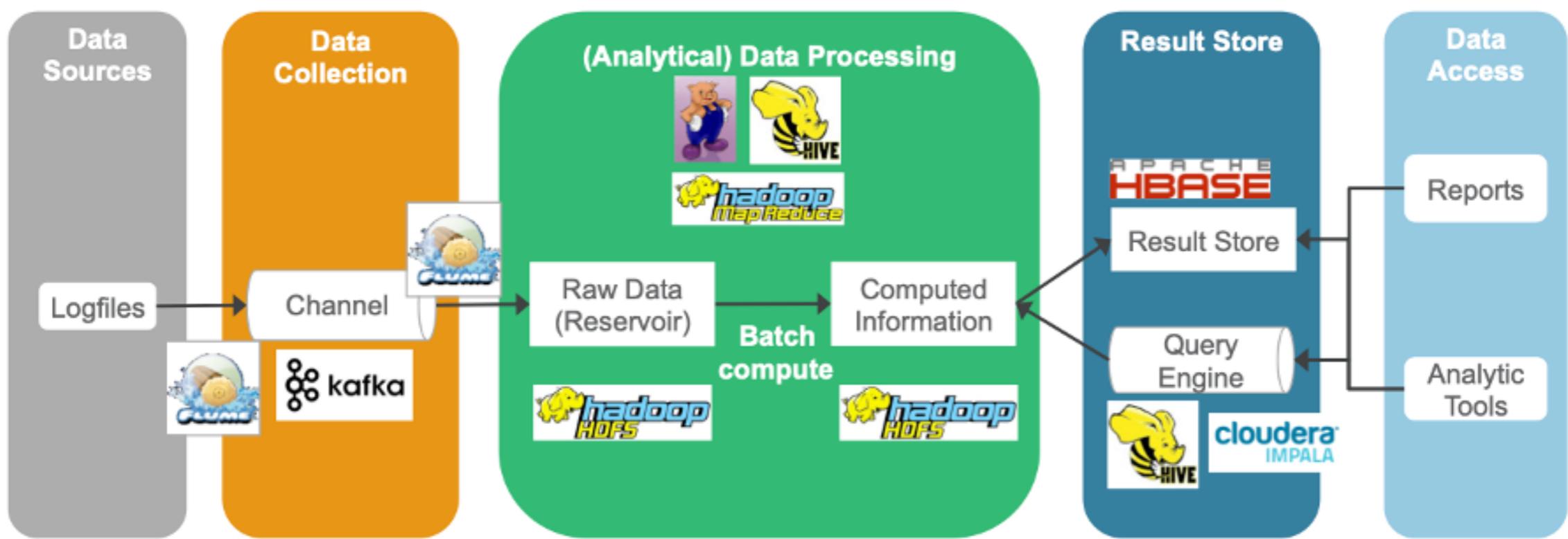
Pre-Big-Data Era Data Analytics



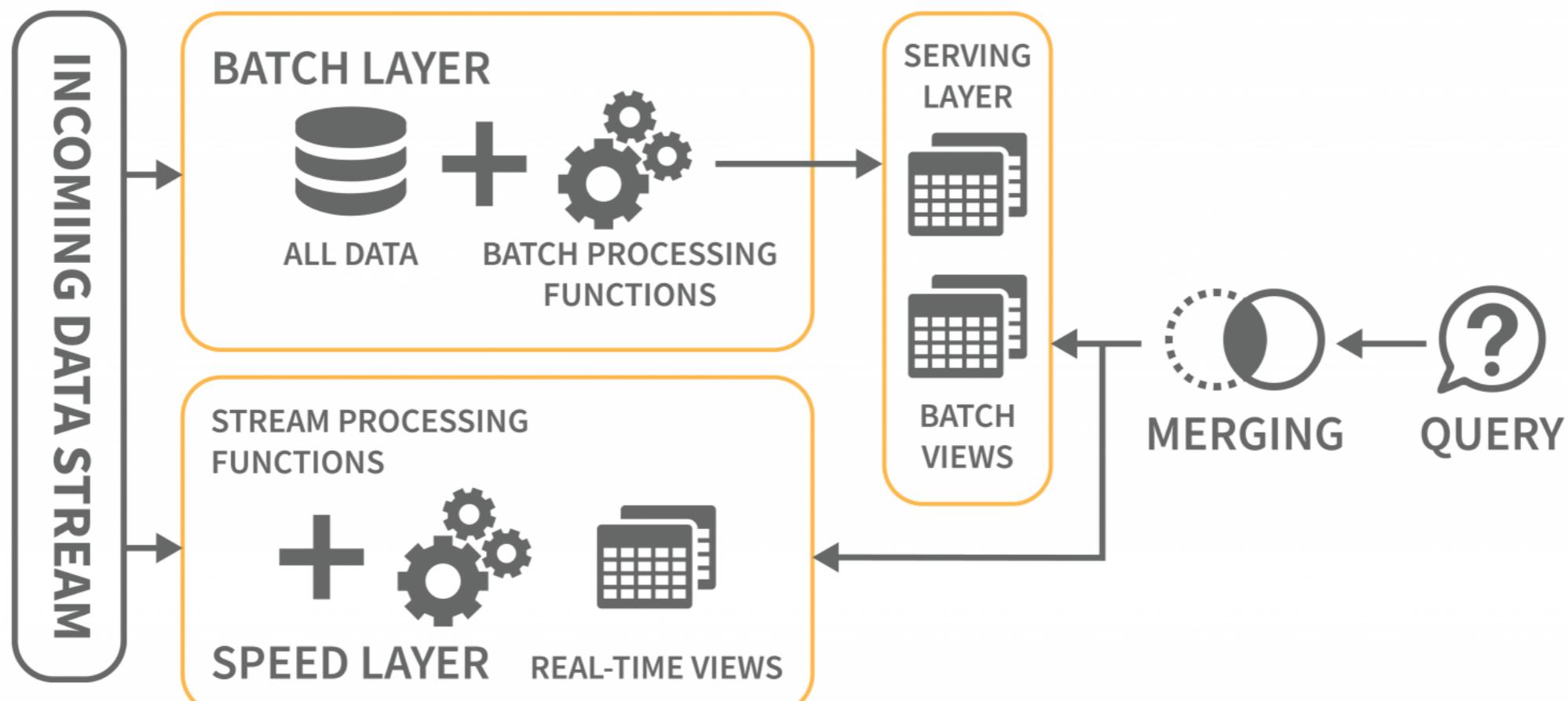
Simple Big Data Analytic Architecture



Example: Facebook Data Pipeline (early days)

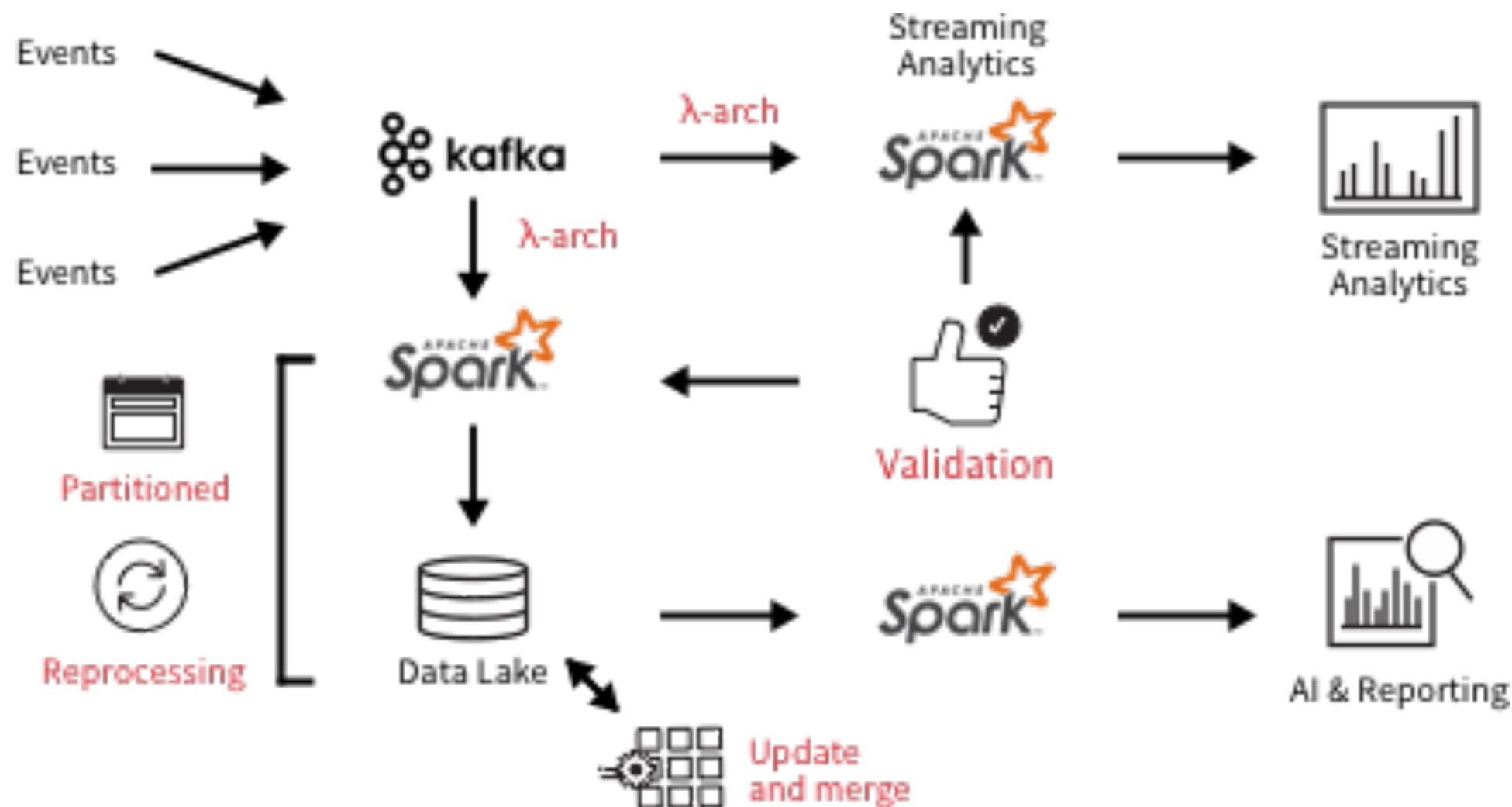


Lambda Architecture (Nathan Marz)



InterviewBit

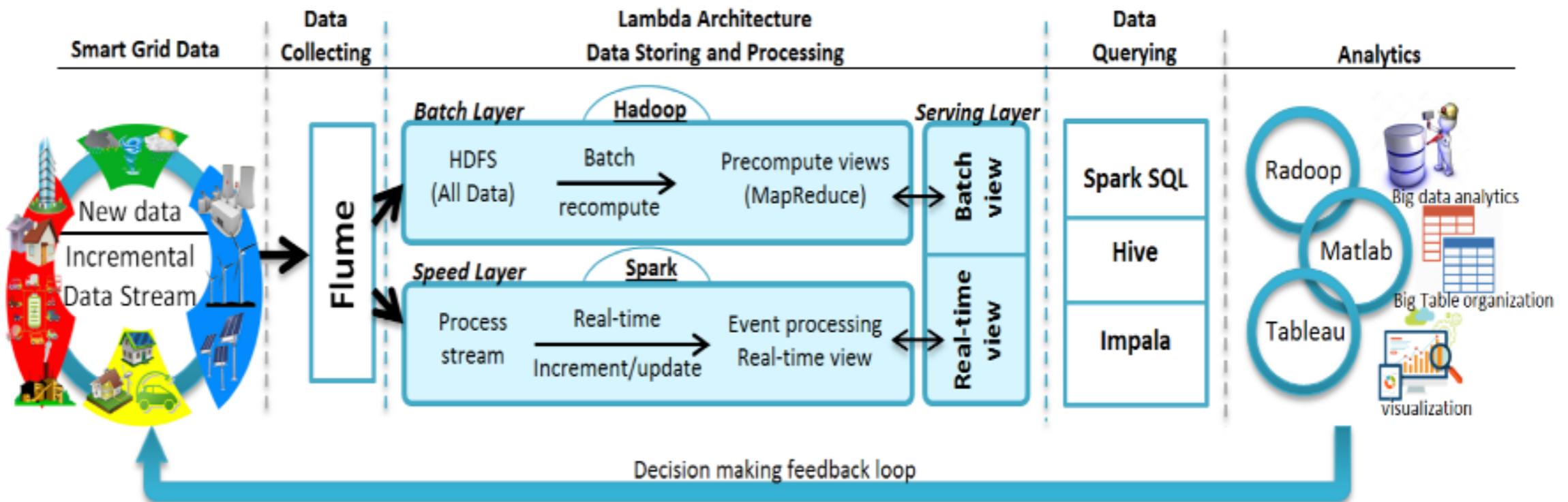
Lambda Architecture in Practice



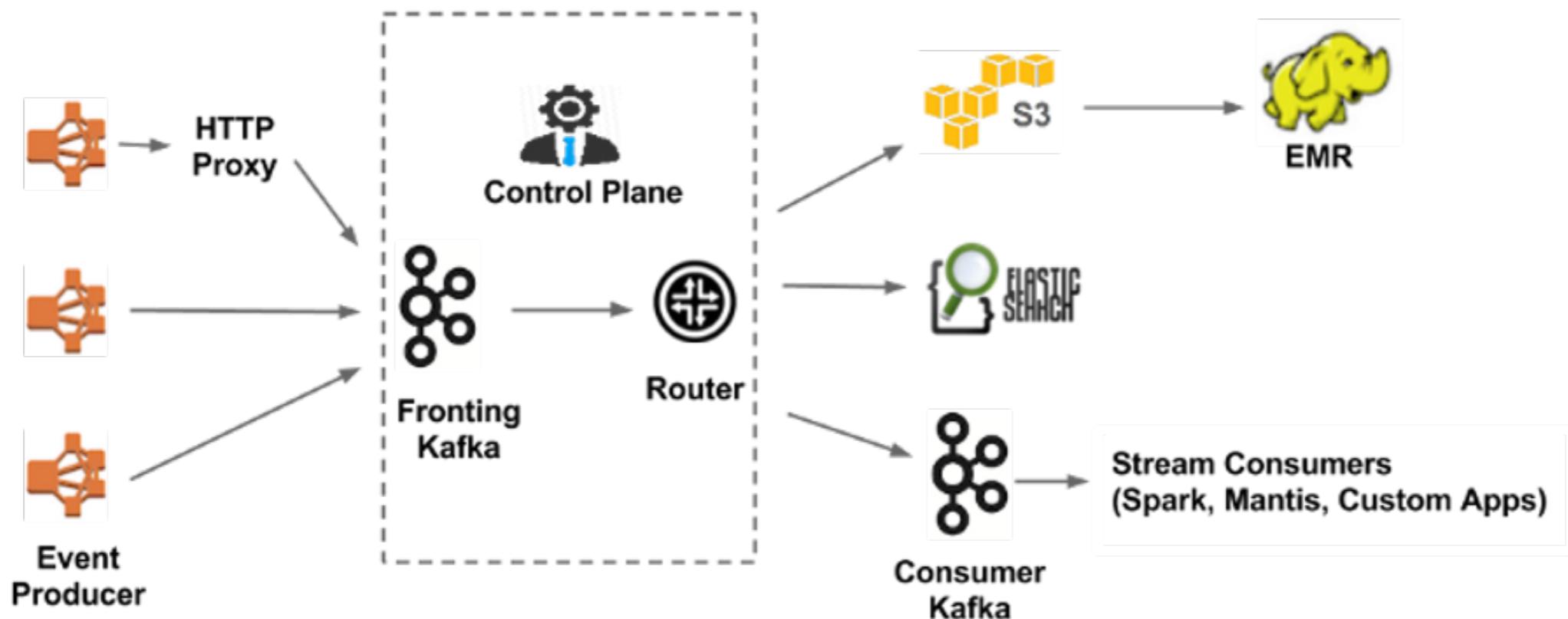
Principles of Lambda Architecture

- 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

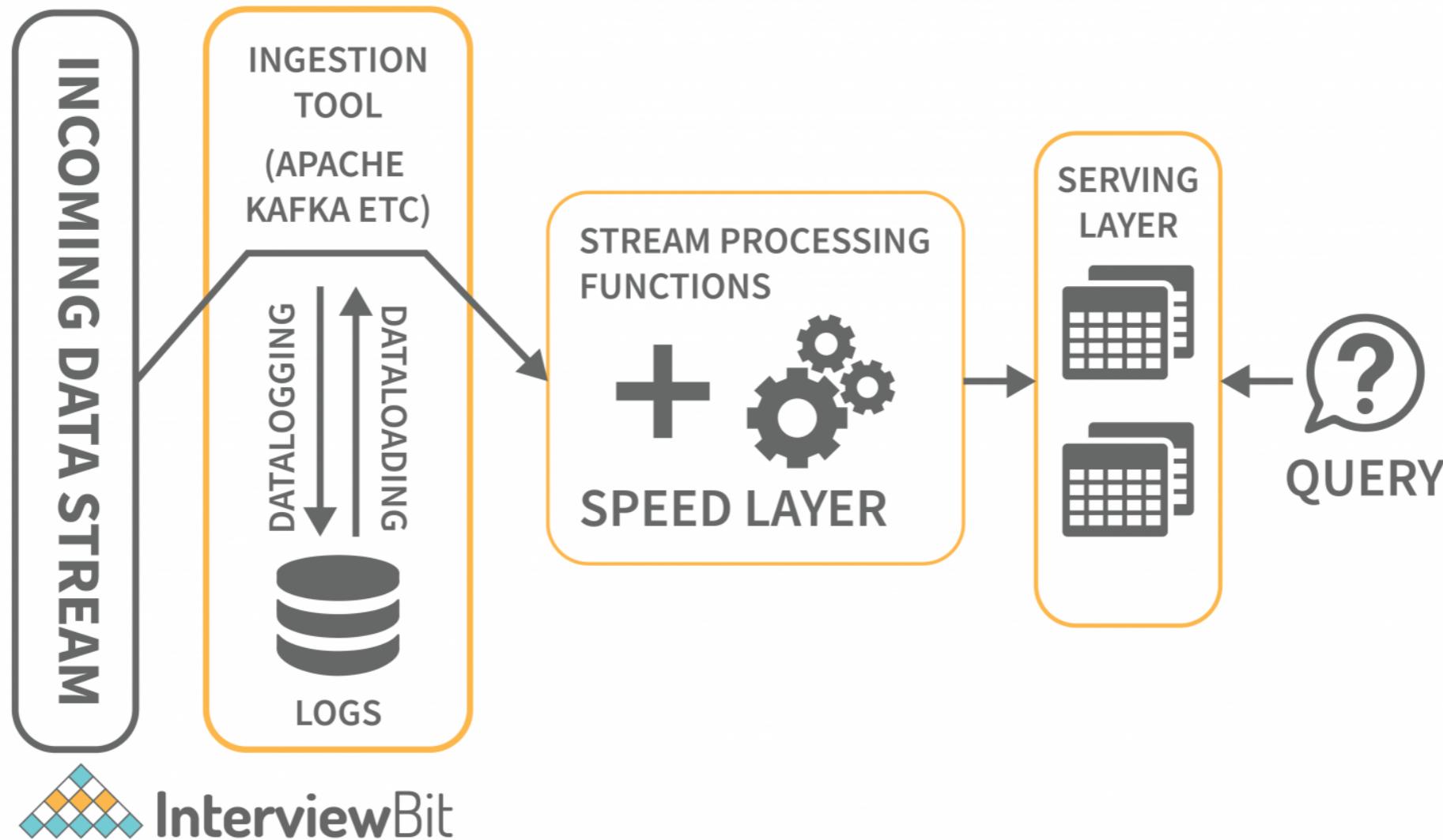
Example: Smart Grid Architecture



Example: Netflix's Keystone Architecture



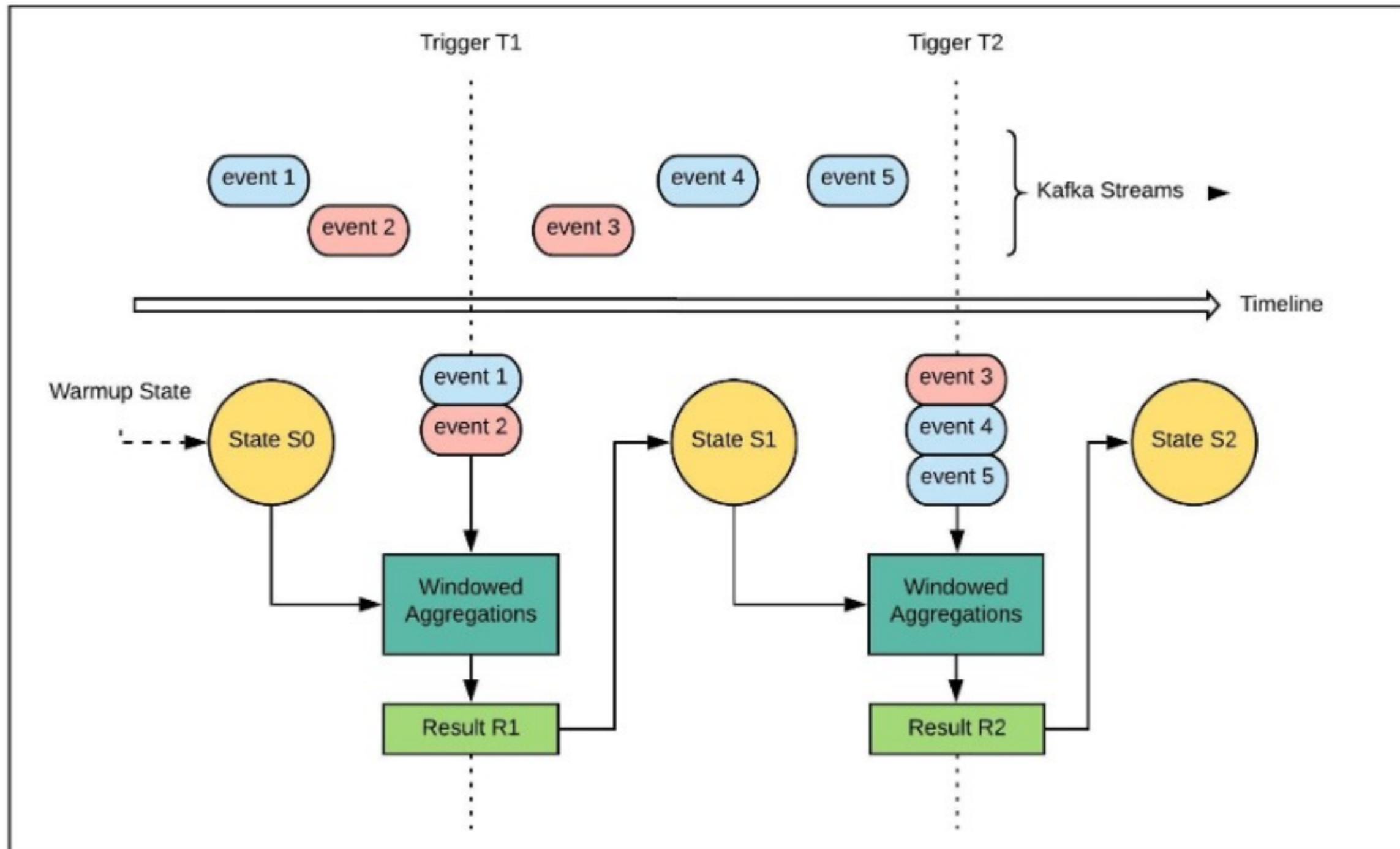
Kappa Architecture (Jay Krepsen)



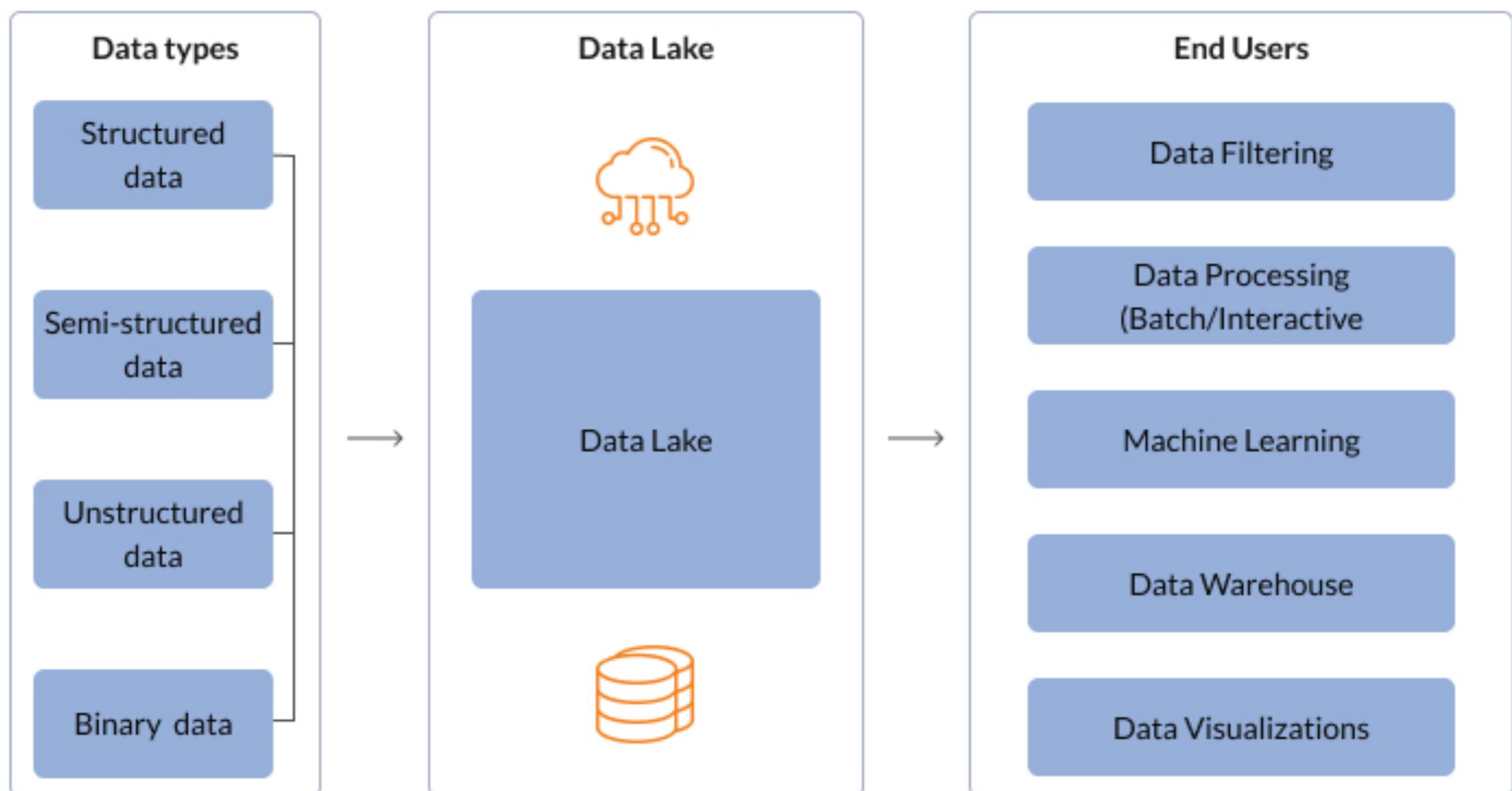
Principles of Kappa Architecture

- Real-time data or “stream” is everything, batch is just a subset of stream
- Raw data is never changed
- Multiple views can be defined on one raw data stream
- Single and unified data processing framework (for both real-time and batch)
- Handle late data arrival and failure recovery can be done by reprocessing or “replaying” historical data (because raw data is never changed)

Example: Uber Real-Time Dynamic Pricing

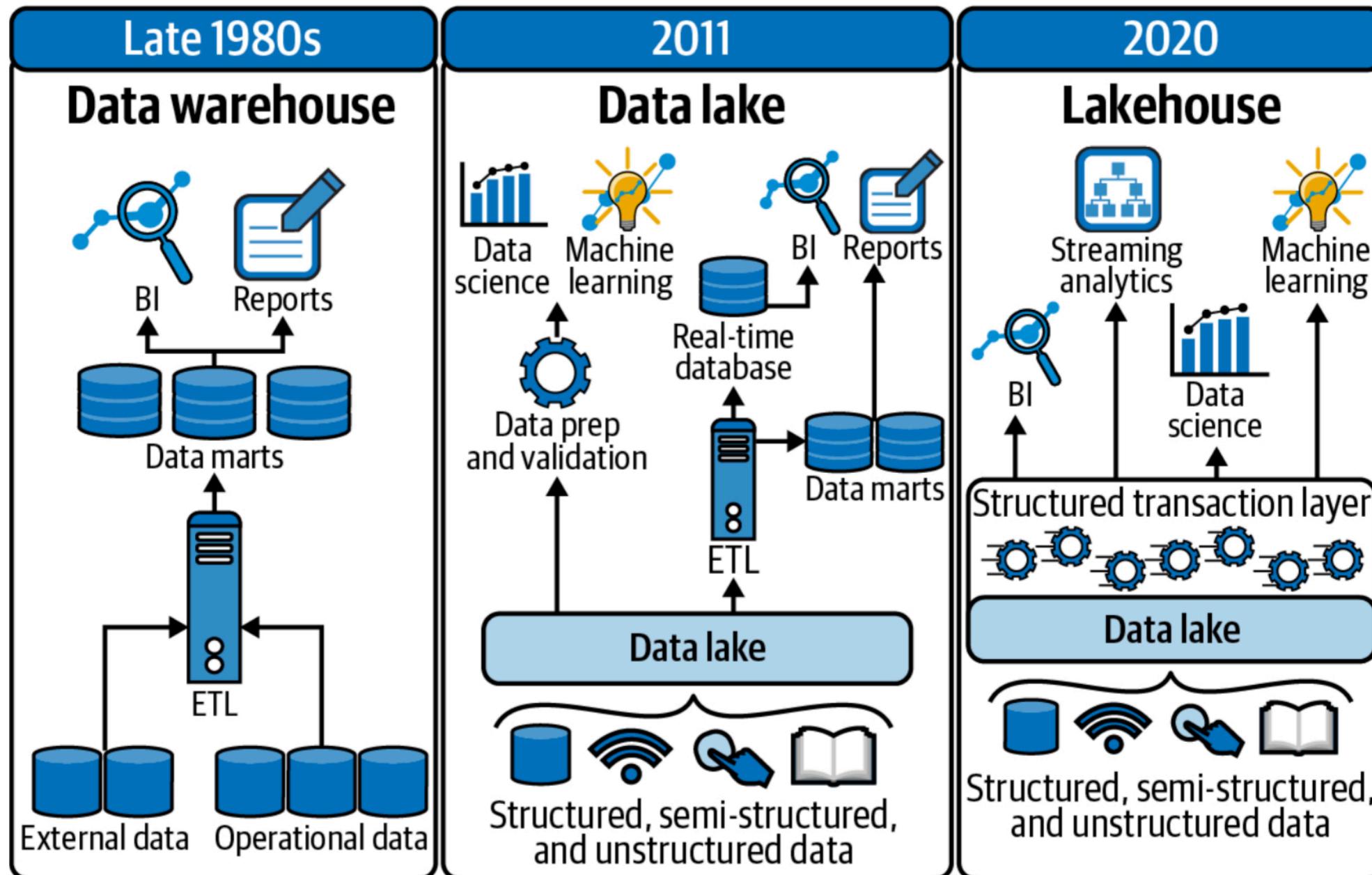


Data Lake Architecture



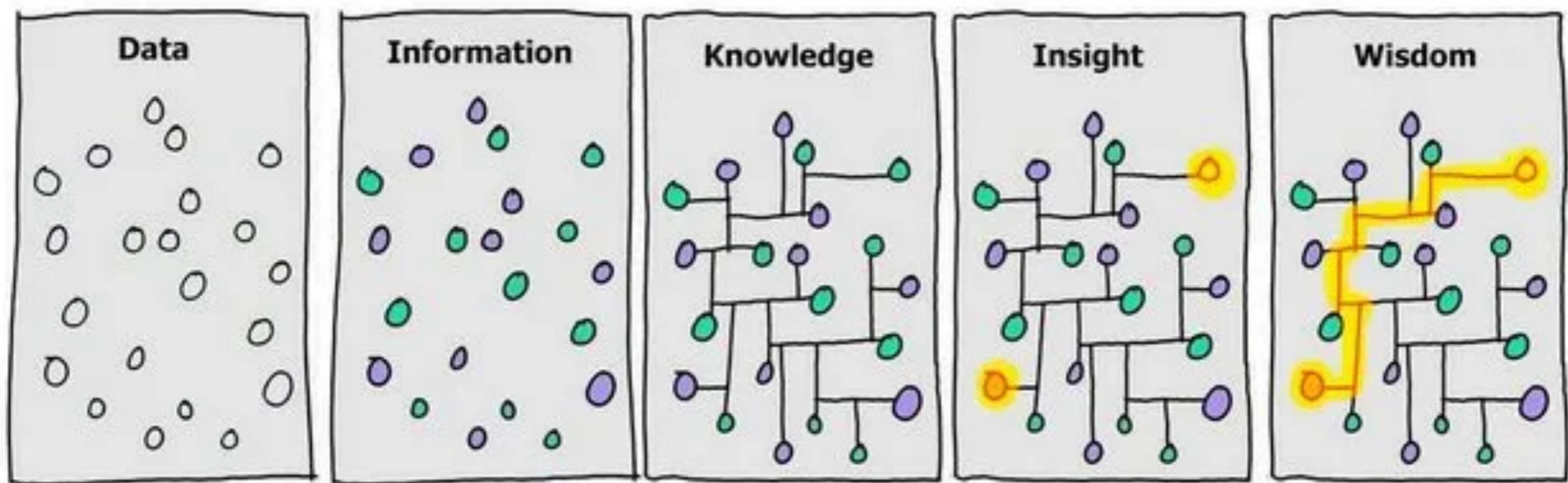
Source: <https://www.n-ix.com/data-lake-vs-data-warehouse/>

Data Lakehouse Architecture



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

Data Evolution



Bronze

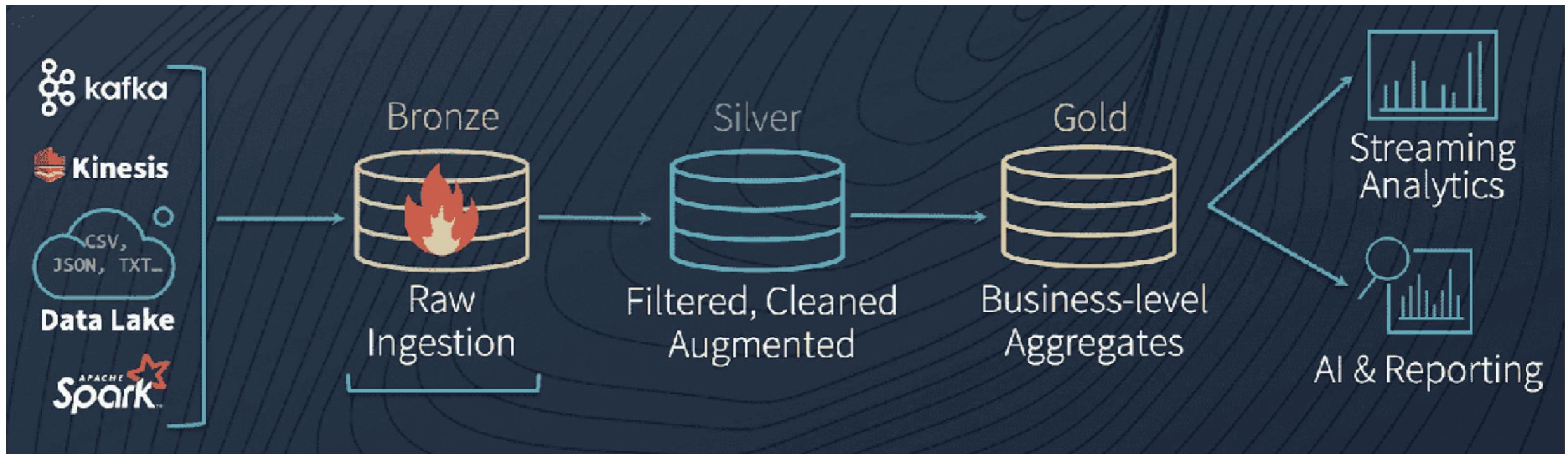
Silver

Gold

Visualization /
BI Tools

Personalization
AI/ML

Delta Lake Architecture



Source: <https://www.databricks.com/blog/2020/11/20/delta-vs-lambda-why-simplicity-trumps-complexity-for-data-pipelines.html>

- 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 Delta Lake Architecture

- Lambda architecture is complex and requires data validation between real-time and batch processing paths
- Kappa architecture does not perform well when large amount of historical data playback is required
- Delta Lake architecture focuses on simplicity by processing data in the data store from bronze to silver to gold
- Single source of truth as there is only one path

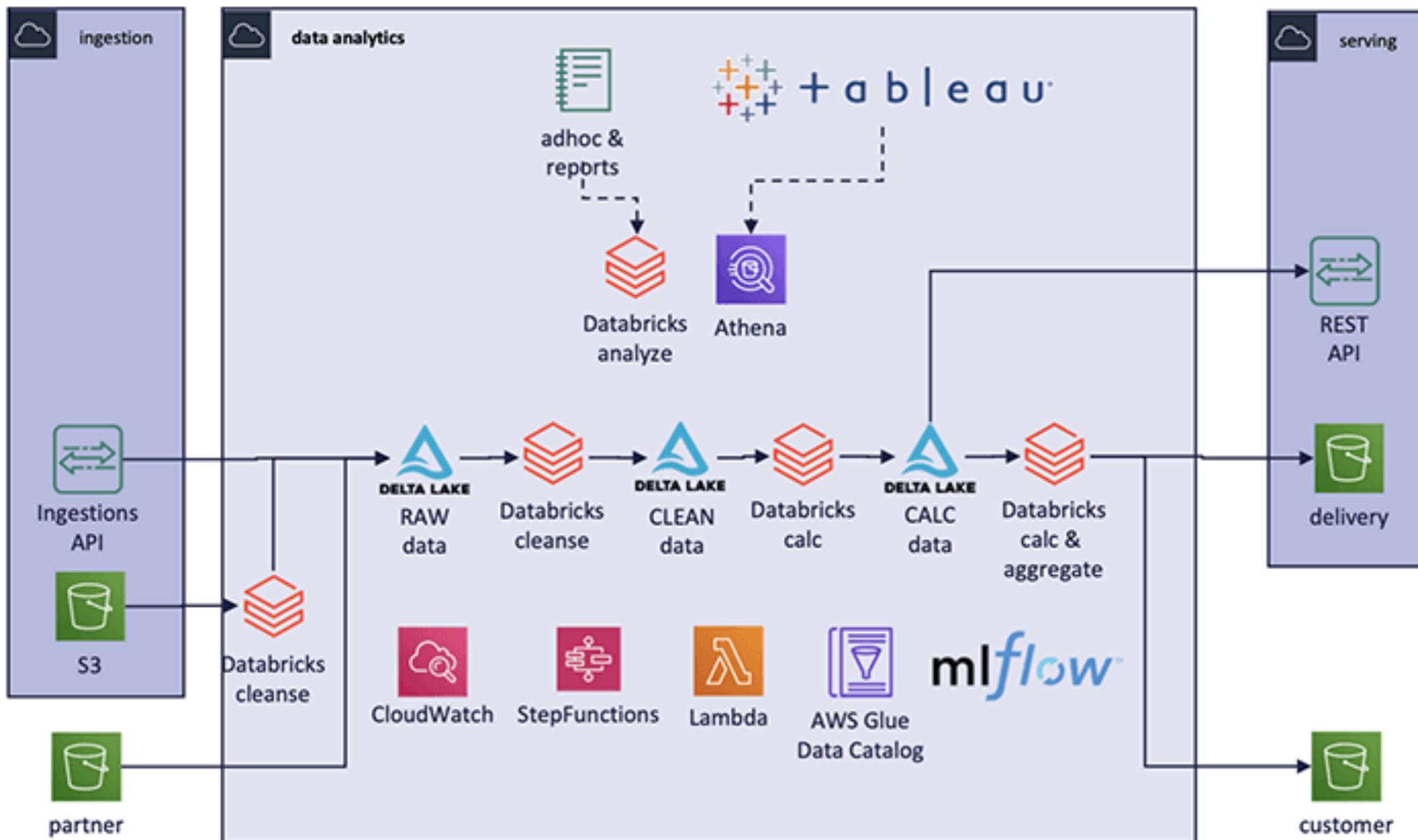
Key Benefits of Delta Lake Architecture

- 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)
- Built-in data governance including ACID transaction control, schema enforcement, etc.
- Unifying batch and stream processing simplifies data pipeline development and maintenance

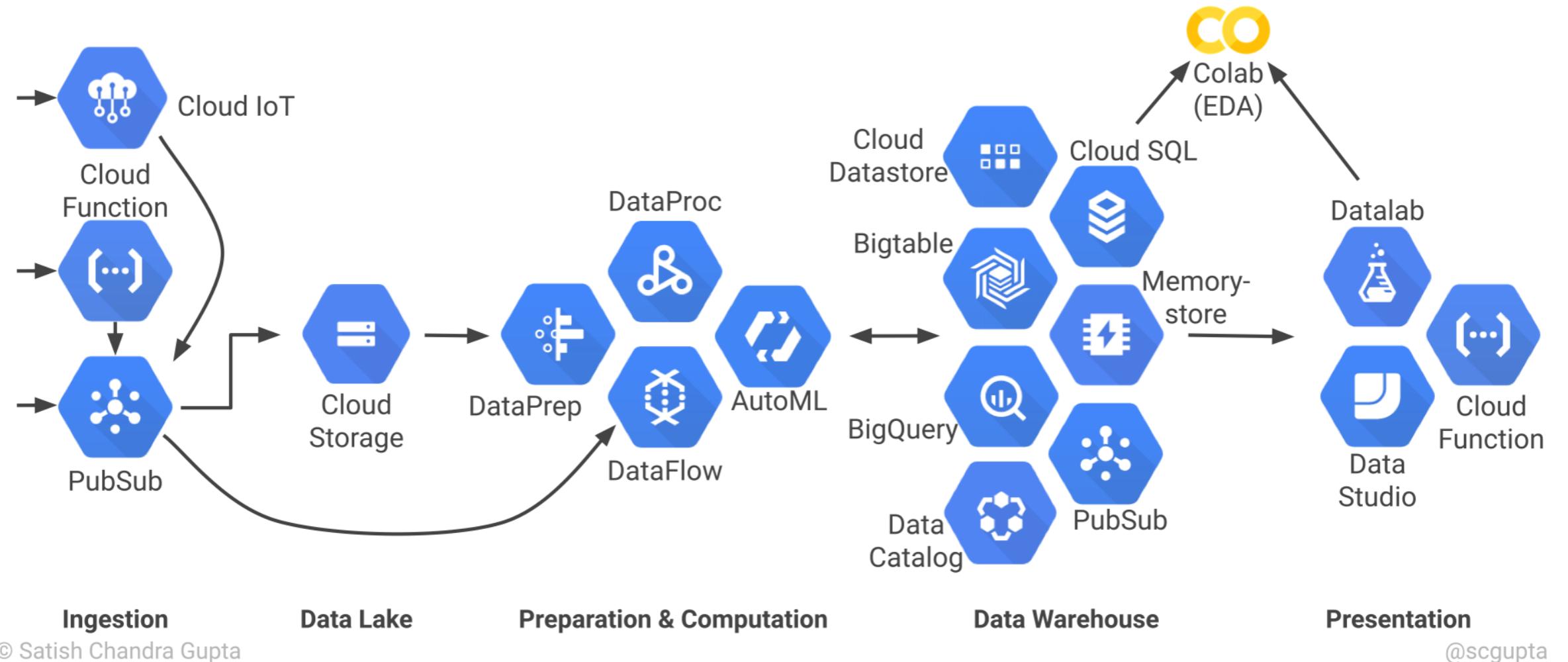
Example: wetter.com

- 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 Delta Lake architecture

Example: wetter.com



Cloud Serverless Architecture



Key Takeaways

- Big Data is just a foundation for data processing; data analytics is also important as it creates business values
- Architectural design patterns can provide templates or guidelines to big data pipeline solutions
- Lambda Architecture is one of the most popular architecture as it can support both batch and real-time data processing
- Data Lakehouse is an emerging architecture, combining best features from data lake (flexibility, low-cost) and data warehouse (structured, governance)

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.
- Real-Time Data Infrastructure Team, “Evolution of the Netflix Data Pipeline,” Netflix Technology Blog. <https://netflixtechblog.com/evolution-of-the-netflix-data-pipeline-da246ca36905> (accessed Aug. 05, 2021).
- Jay Kreps, “Questioning the Lambda Architecture,” O’Reilly Data Newsletter. <https://www.oreilly.com/radar/questioning-the-lambda-architecture/> (accessed Aug. 06, 2021).
- “Data Warehouse Architecture – Detailed Explanation”, <https://www.interviewbit.com/blog/data-warehouse-architecture/>

References

- N. Seyvet and I. M. Viela, “Applying the Kappa architecture in the telco industry.” <https://www.oreilly.com/content/applying-the-kappa-architecture-in-the-telco-industry/> (accessed Aug. 06, 2021).
- Amey Chaugule, “Designing a Production-Ready Kappa Architecture for Timely Data Stream Processing,” Uber Engineering. <https://eng.uber.com/kappa-architecture-data-stream-processing/> (accessed Aug. 06, 2021).
- N. Serheichuk, O. Semkiv, and I. Tymchuk, “Data lake vs data warehouse: Which one to choose for your business?” <https://www.n-ix.com/data-lake-vs-data-warehouse/> (accessed Aug. 06, 2021).
- A. A. Munshi and Y. A.-R. I. Mohamed, “Data Lake Lambda Architecture for Smart Grids Big Data Analytics,” IEEE Access, vol. 6, 2018, doi: 10.1109/ACCESS.2018.2858256.
- <https://www.databricks.com/blog/2020/11/20/delta-vs-lambda-why-simplicity-trumps-complexity-for-data-pipelines.html>
- <https://towardsdatascience.com/scalable-efficient-big-data-analytics-machine-learning-pipeline-architecture-on-cloud-4d59efc092b5>
- 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>