G.AI.L

Tecnológico de Monterrey

COECYTJAL
Consejo Estatal de Ciencia y Tecnología de Jalisco

Machine Learning Operations Bootcamp

Module 3

MODULE 3: Big Data

Session 2

About us

Grisell Reyes Rios

Josué Ruiz

- I am a chemist who turned into a Data Scientist/ Data Engineer
- I like to travel at every opportunity that I have
- I speak German

Important Notes

Use your name and last name to identify yourself in Meet

Mute your microphone

Raise your hand to participate

Turn off your camera if you have connection issues

Academy Code of Conduct

Be respectful
All questions and ideas are valid and welcome

Be welcoming and patient

Be careful with the words you use

Session Goals

At the end, you will be able to:

- **Understand** the fundamentals of Big Data
- **Better understanding** of Big Data in MLOps
- Examples of Big Data in MLOps
- **Hands-on** on a Spark example in Machine Learning

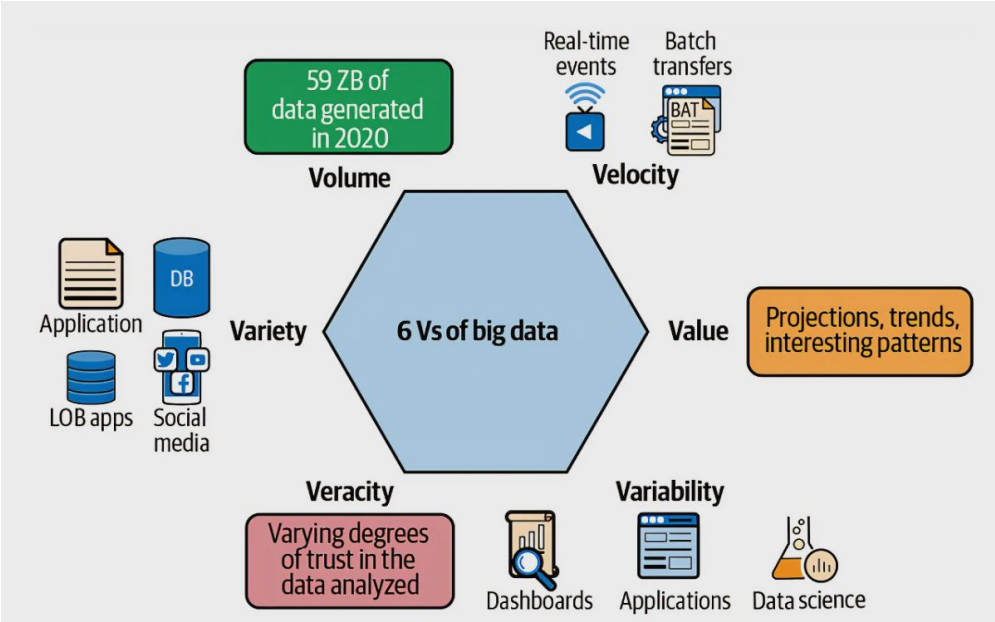
Table of Contents

Introduction to Big Data

Big Data definition and history overview

Big Data Definition

Even though the term “big” is used in big data, it’s not just about the size of the data. It’s also about all the data, big or small, within your company and all the data outside your company that would be helpful to you. The data could be in any type of format and can be collected in any type of regularity. So, the best way to define big data is to think of it as all data, no matter its size (volume), speed (velocity), or type (variety)



Source: The Cloud Data Lake by Rukmani Gopalan (O’Reilly, 2023).

Table of Contents

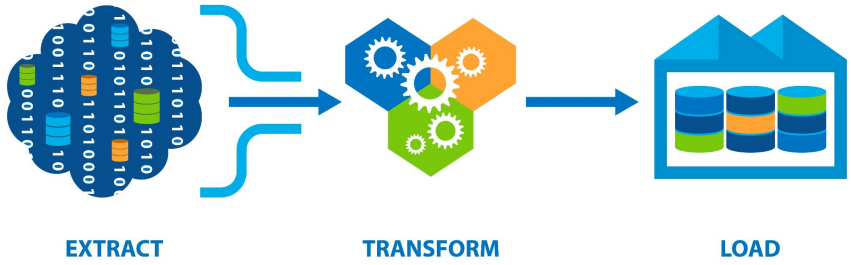
Fundamentals

Get a better understanding of a modern Big Data Platform

Fundamentals

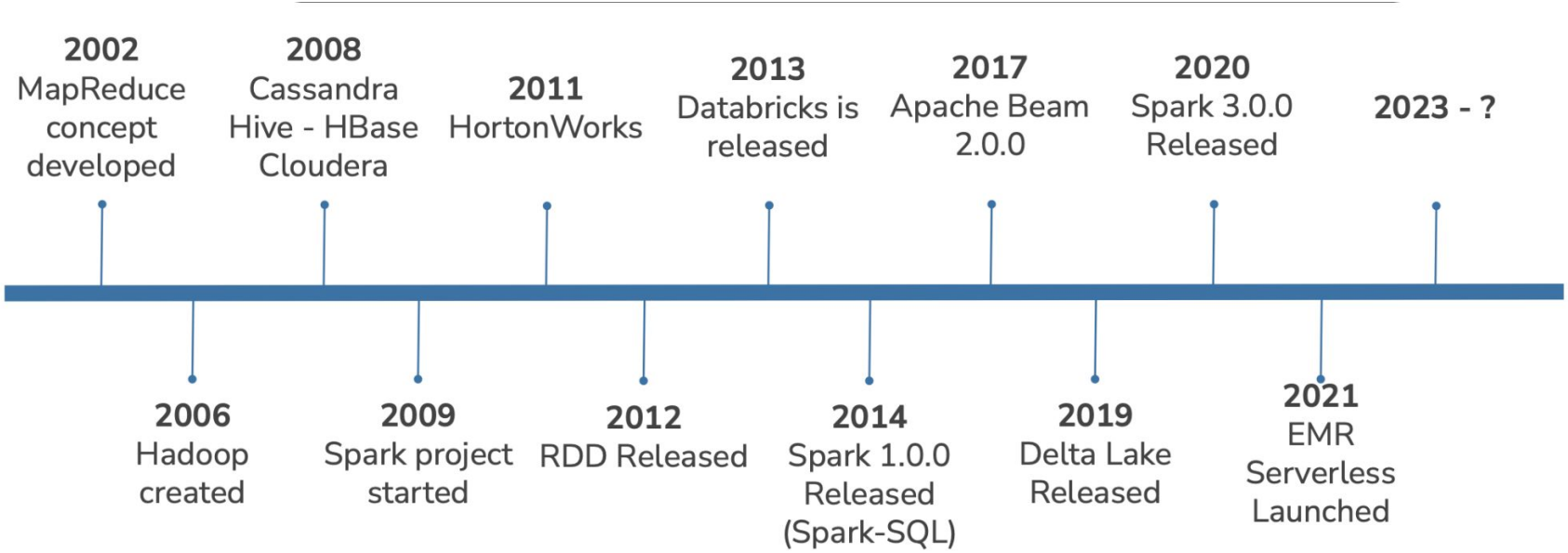
What makes a modern Big Data platform remains unclear. A modern Big Data platform has several requirements, and to meet them correctly, expectations about data should be set. Once a base is established for expectations from data, we can then reason about a modern platform that can serve it.

- Expectations of data
  - Ease of access
  - Security
  - Quality
  - Extensibility
- Expectations of platform
  - Storage Layer
  - Resource Management
  - ETL
  - Discovery
  - Monitoring
  - Testing
  - Lifecycle Management



Source: https://www.datachannel.co/blogs/what-is-etl-and-how-the-etl-process-works

A Look in History...



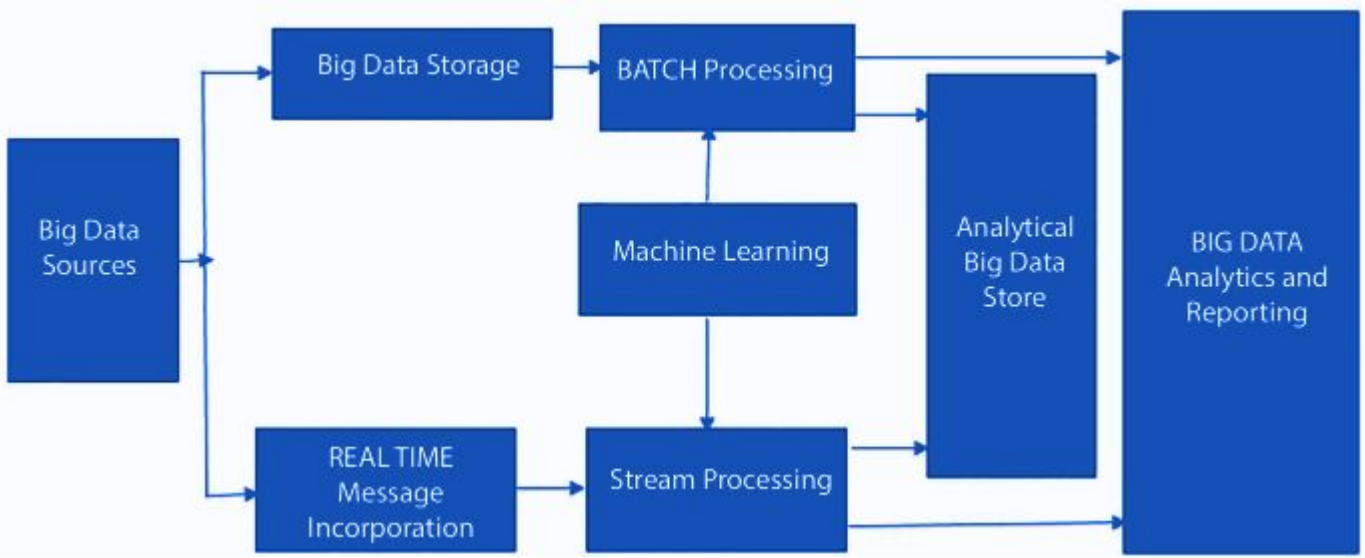
Types of Data Sources and Memory Symbolize and Size Handle by Big Data

Format	Binary/Text	Human-readable	Example use cases
JSON	Text	Yes	Everywhere
CSV	*Text*	*Yes*	*Everywhere*
Parquet	*Binary*	*No*	*Hadoop, Amazon Redshift*
Avro	Binary primary	No	Hadoop
Protobuf	Binary primary	No	Google, TensorFlow (TFRecord)
Pickle	Binary	No	Python, PyTorch serialization

S. No.	Memory	Symbolize	Size
1	1 Kilobyte	KB	10^3
2	1 Megabyte	MB	10^6
3	1 Gigabyte	GB	10^9
4	1 Terabyte	TB	10^{12}
5	1 Petabyte	PB	10^{15}
6	1 Exabyte	EB	10^{18}
7	1 Zettabyte	ZB	10^{21}
8	1 Yottabyte	YB	10^{24}

Source: Dulhare, U.N., Ahmad, K, Amali Bin Ahmad, K. (2020). Machine Learning and Big Data. Wiley-Scrivener.

Architecture of Big Data



Source: Adapted from Dulhare, U.N., Ahmad, K, Amali Bin Ahmad, K. (2020). Machine Learning and Big Data. Wiley-Scrivener.

Big Data Analytical Tools

Apache Hadoop

CDH (Cloudera Distribution for Hadoop)

Cassandra

MongoDB

Rapidminer

Source: Adapted from Dulhare, U.N., Ahmad, K, Amali Bin Ahmad, K. (2020). Machine Learning and Big Data. Wiley-Scrivener.

Table of Contents

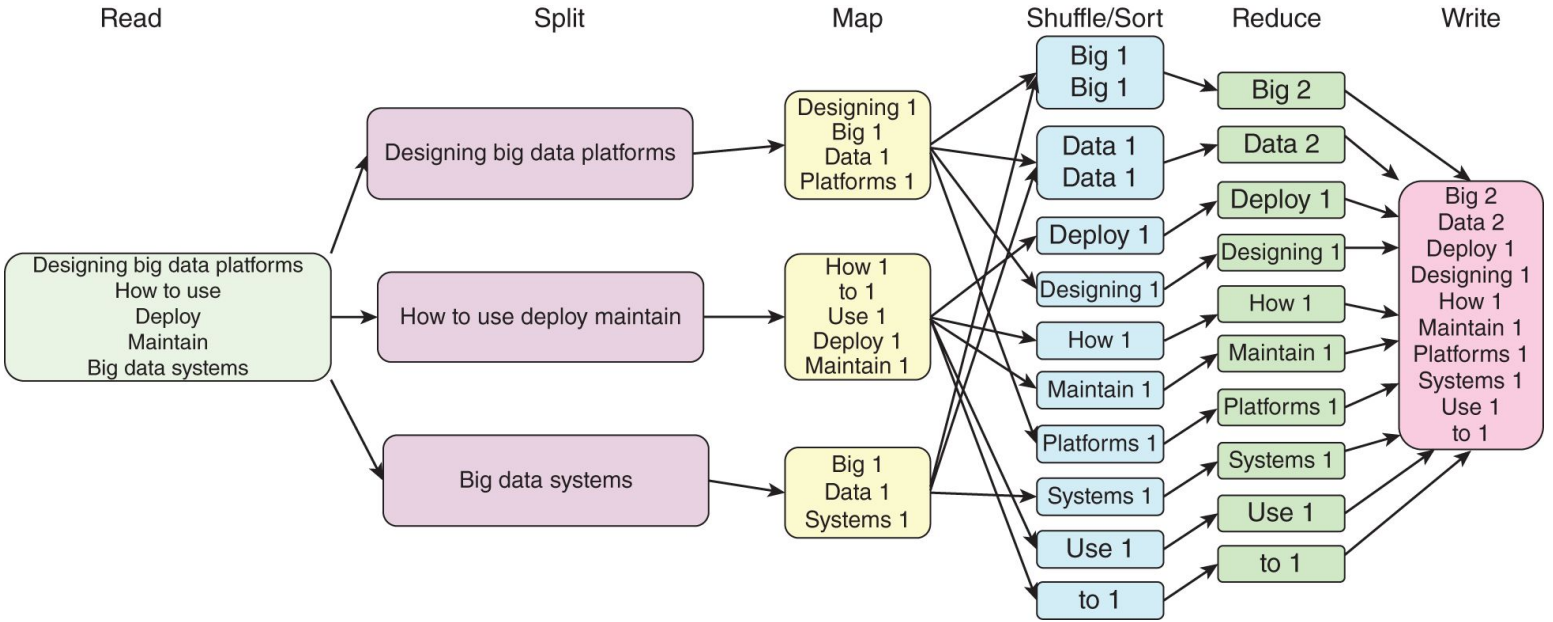
Apache Spark: Big Data Processing

Definition, clusters components, and differences between RDD, Dataframe and Dataset

Apache Hadoop

1. **MapReduce:** Of all the approaches to data processing in Hadoop, MapReduce requires the most code by far.

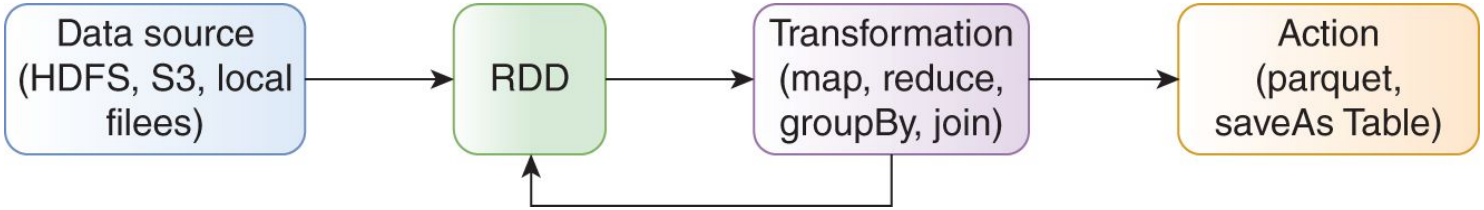
Let us see this powerful programming model in action with the algorithm for the *Word Count Problem*.



Source: Aytas, Y.(2021). Designing Big Data Platforms. Wiley.

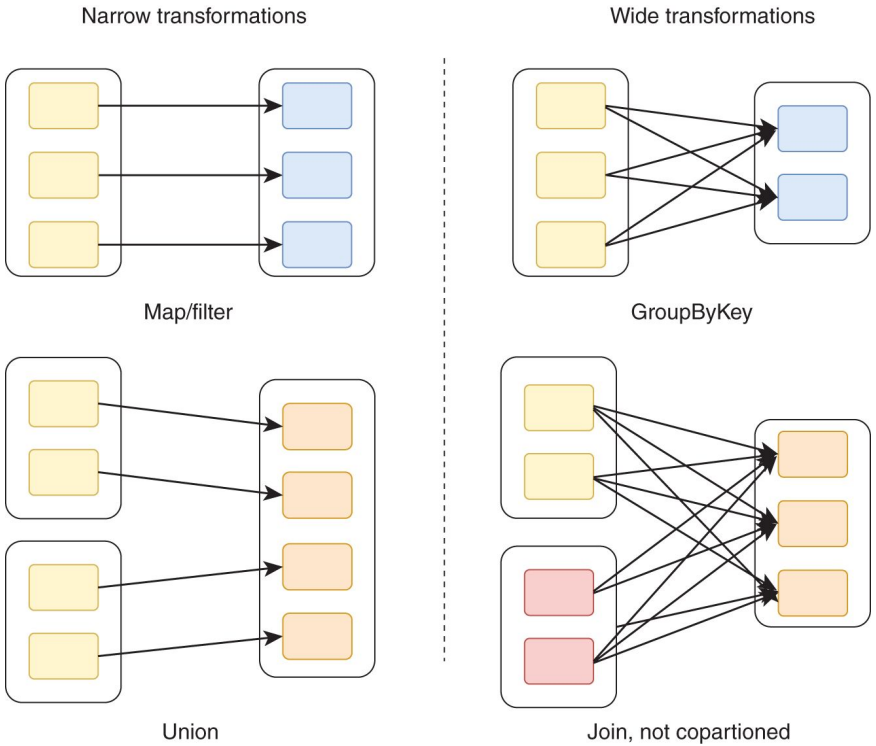
2. Spark: Apache Spark is a unified analytics engine for Big Data processing. Spark provides a collection of libraries, APIs in multiple languages, and a diverse set of data science tools, graph processing, data analytics, and stream processing. Apache Spark can run on different platforms ranging from a laptop to clusters. It can scale to thousands of machines.

A typical spark application would load data from a data source, and then transform the RDD into another one. A new RDD points back to its parent, which creates a lineage graph or directed acyclic graph (DAG). DAG instructs Spark about how to execute these transformations.



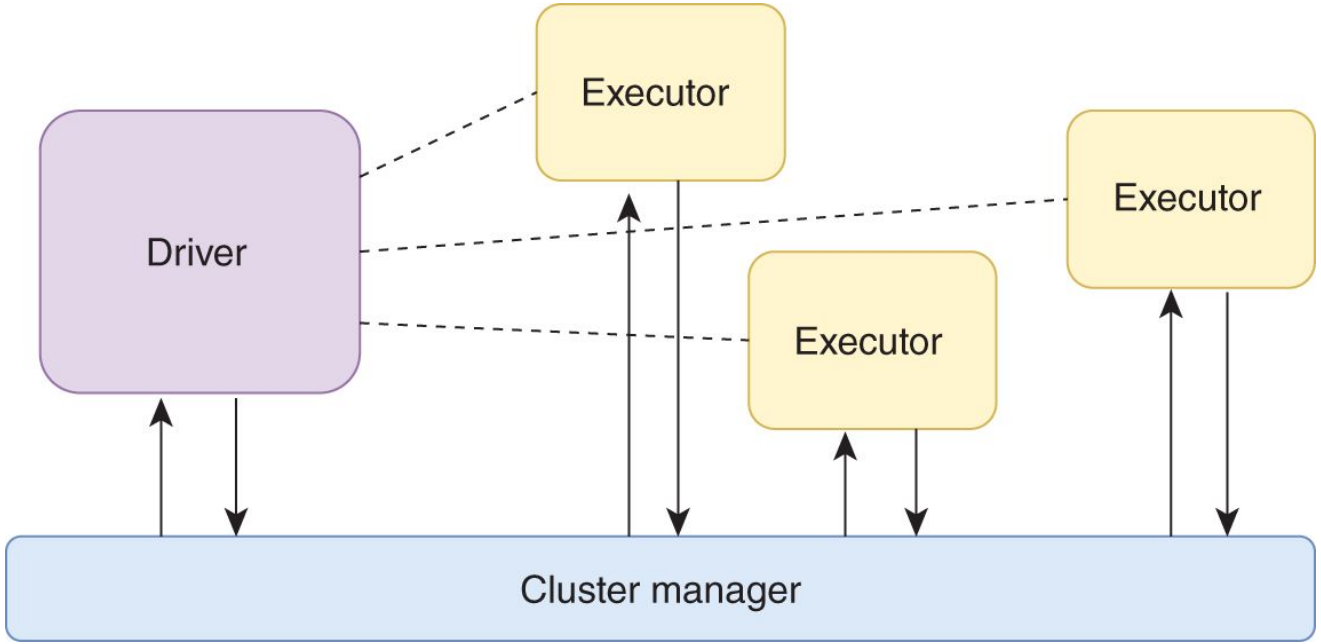
Source: Aytas, Y.(2021). Designing Big Data Platforms. Wiley.

Apache Hadoop



Source: Aytas, Y.(2021). Designing Big Data Platforms. Wiley.

Spark High Level Architecture



Source: Aytas, Y.(2021). Designing Big Data Platforms. Wiley.

Comparation between RDD, Dataset and Dataframes

	RDD	Dataset	Dataframes
Similarities & Differences	Is a distributed collection of data that allow for parallel processing.	Is a distributed collection of data that allow for parallel processing.	Is a distributed collection of data that allow for parallel processing.
	Are unstructured, meaning they don't have a schema.	Have a schema, meaning that their data is organized into named columns.	Have a schema, meaning that their data is organized into named columns.
	Don't take advantage of Spark's Catalyst Optimizer.	Take advantage of Spark's Catalyst Optimizer.	Take advantage of Spark's Catalyst Optimizer.
	Strongly-typed (variables are checked at compile-time and type errors are detected before the program is executed)	Strongly-typed (variables are checked at compile-time and type errors are detected before the program is executed)	Weakly-typed (not checked until runtime and type errors may not be detected until the program is executed)
Best use cases	Complex data processing, machine learning algorithms that require low-level transformation and control.	High-performance data processing and analysis, machine learning algorithms that require type safety and low-level transformation.	Structured data processing, ETL (Extract, Transform, Load) operations, exploratory data analysis.
	Unstructured data processing	Is a hybrid of RDDs and Dataframes and offer benefits of both.	Data exploration and analysis and Running ad-hoc queries on large datasets with Spark-SQL.

Table of Contents

Apache Spark: Big Data Processing

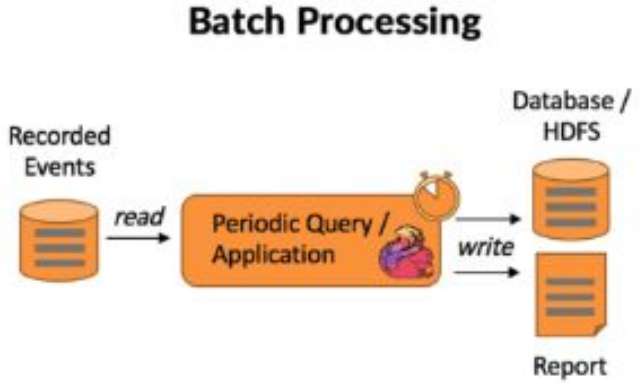
Definition, clusters components and differences between RDD, Dataframe and Dataset

Batch and Stream Processing

Batch processing is the method computers use to periodically complete high-volume, repetitive data jobs. Data systems process such tasks in batches, often in off-peak times when computing resources are more commonly available, such as at the end of the day or overnight. For example, consider an e-commerce system that receives orders throughout the day. Instead of processing every order as it occurs, the system might collect all orders at the end of each day and share them in one batch with the order fulfillment team.

Batch process systems are used to process various types of data and requests. Some of the most common types of batch-processing jobs include:

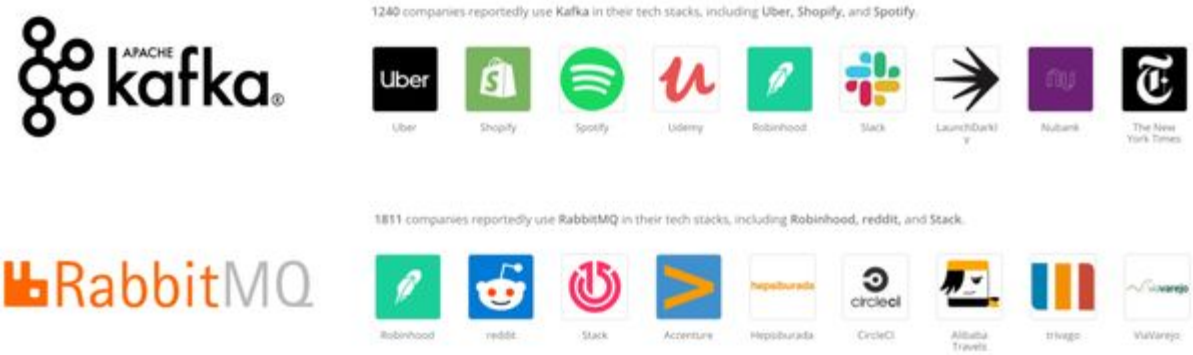
- Weekly/monthly billing
- Payroll
- Inventory processing
- Report generation
- Data conversion
- Subscription cycles
- Supply chain fulfillment



Batch and Streaming Processing

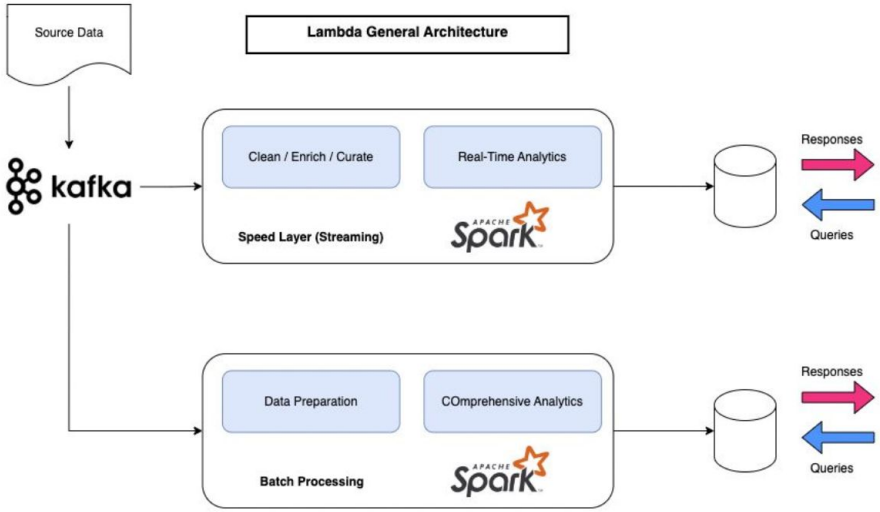
When you have data in real-time transports like Apache Kafka and Amazon Kinesis, we say that you have streaming data. *Stream processing* refers to doing computation on streaming data. Computation on streaming data can also be kicked off periodically, but the periods are usually much shorter than the periods for batch jobs (e.g., every five minutes instead of every day).

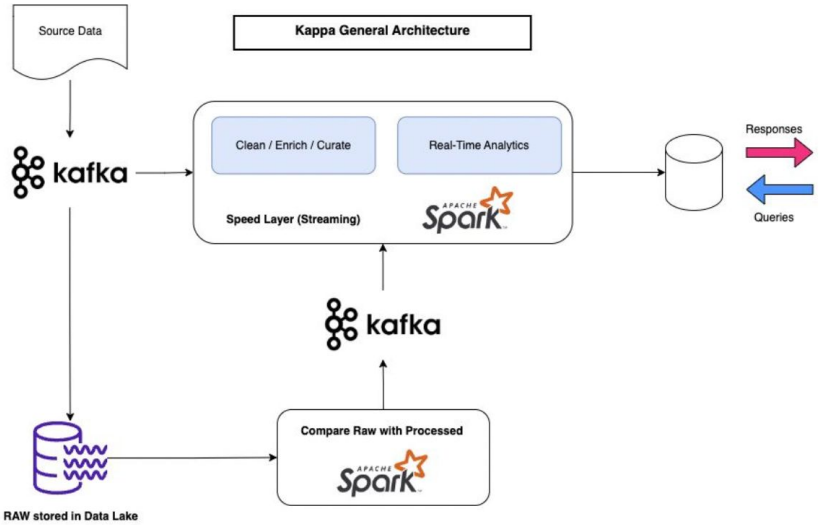
A piece of data broadcast to a real-time transport is called an event. This architecture is, therefore, also called event-drive, real-time transport is sometimes called an event bus. The two most common types of real-time transports are pub-sub, which is short for publish-subscribe, and message queue. In the pub-sub model, any service can publish to different topics in a real-time transport, and any service that subscribes to a topic can read all the events on that topic.



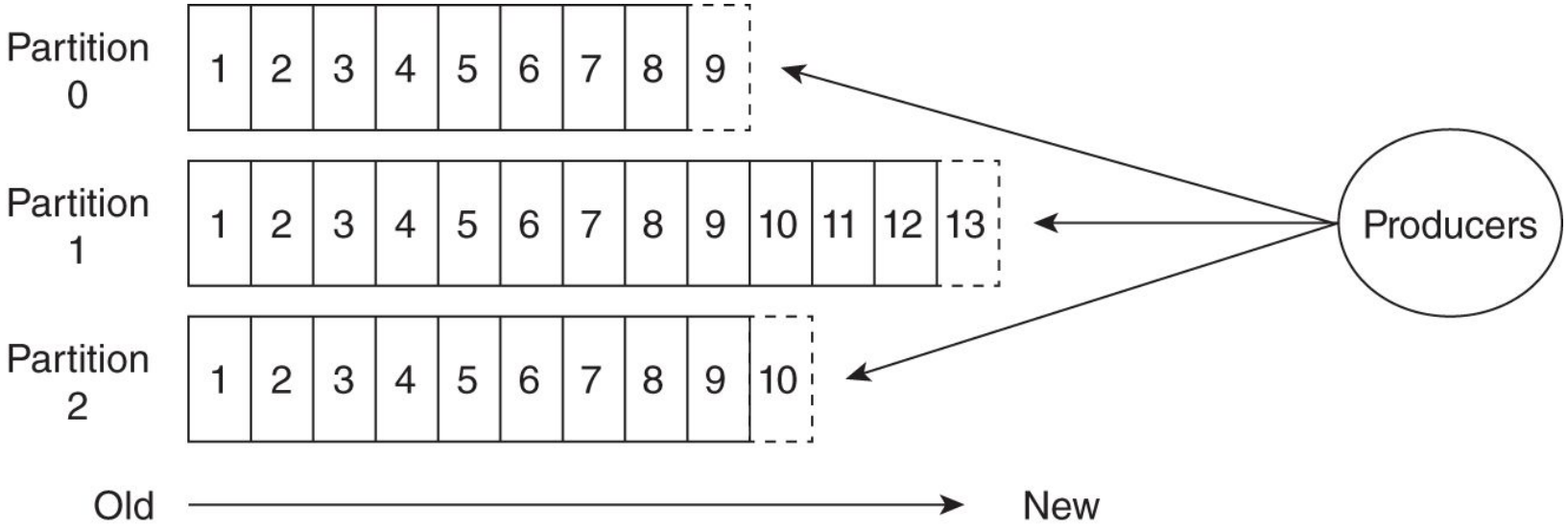
Source: Huyen, C.(2021). Designing Machine Learning Systems. O´Reily Media, Inc.

Lambda and Kappa Architecture



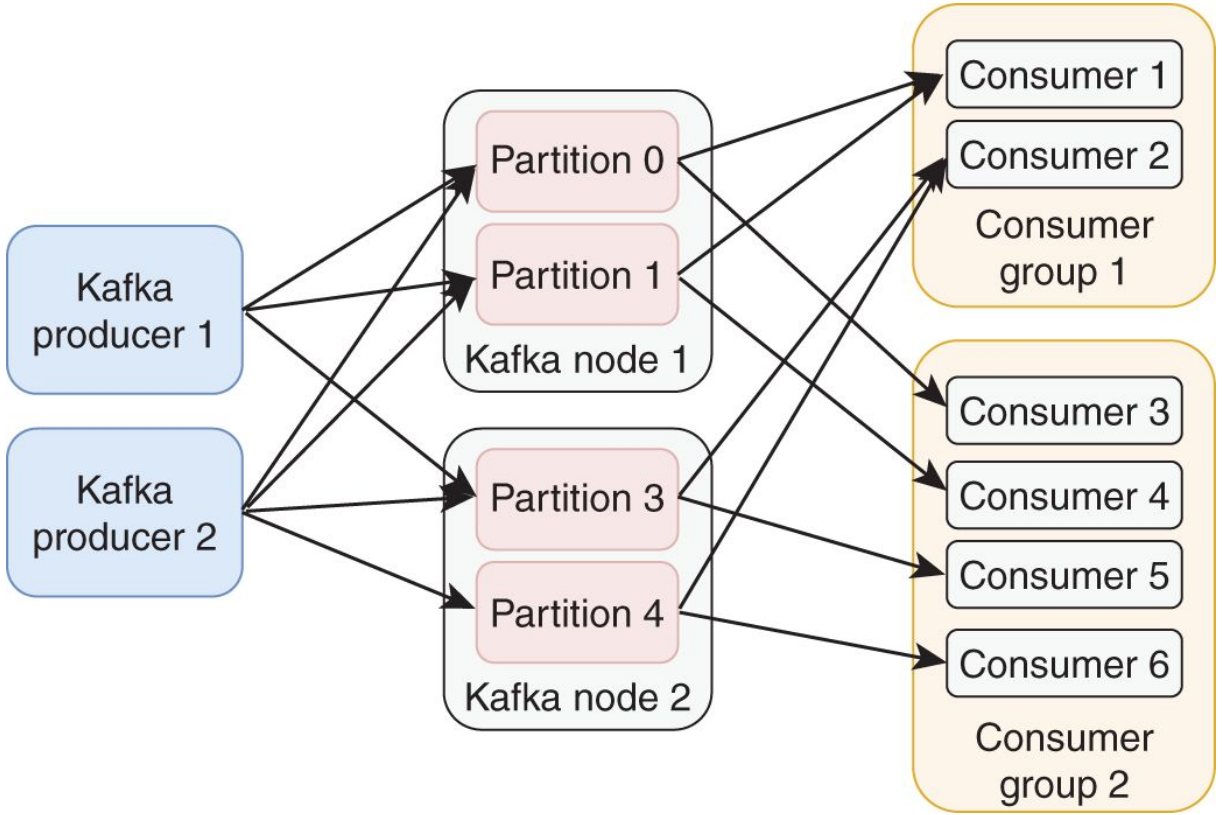


Apache Kafka



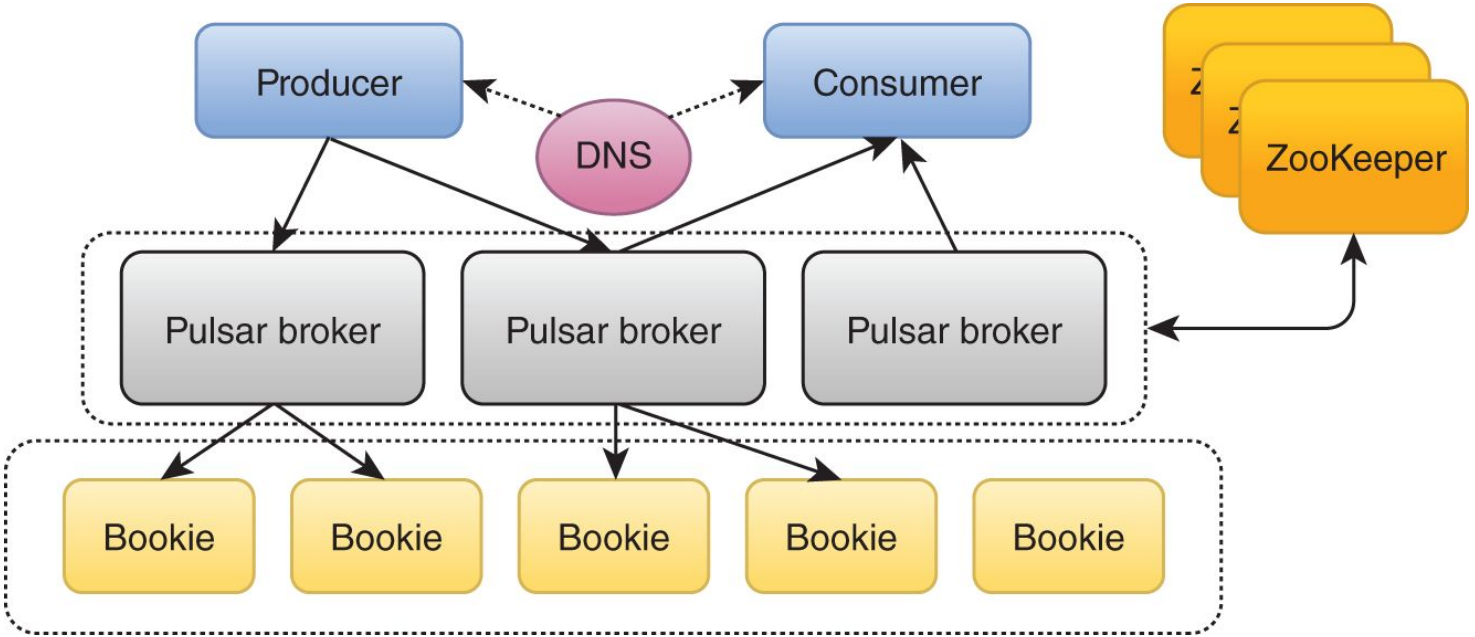
Source: Aytas, Y.(2021). Designing Big Data Platforms. Wiley.

Apache Samza



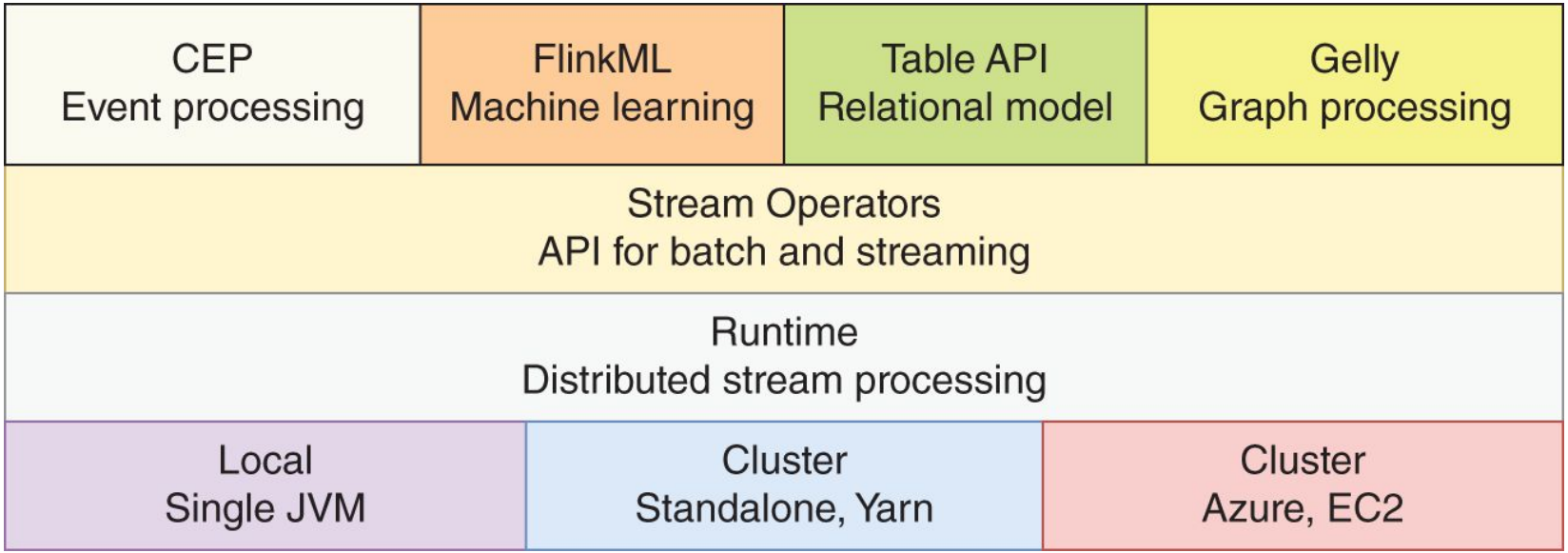
Source: Aytas, Y.(2021). Designing Big Data Platforms. Wiley.

Apache Pulsar



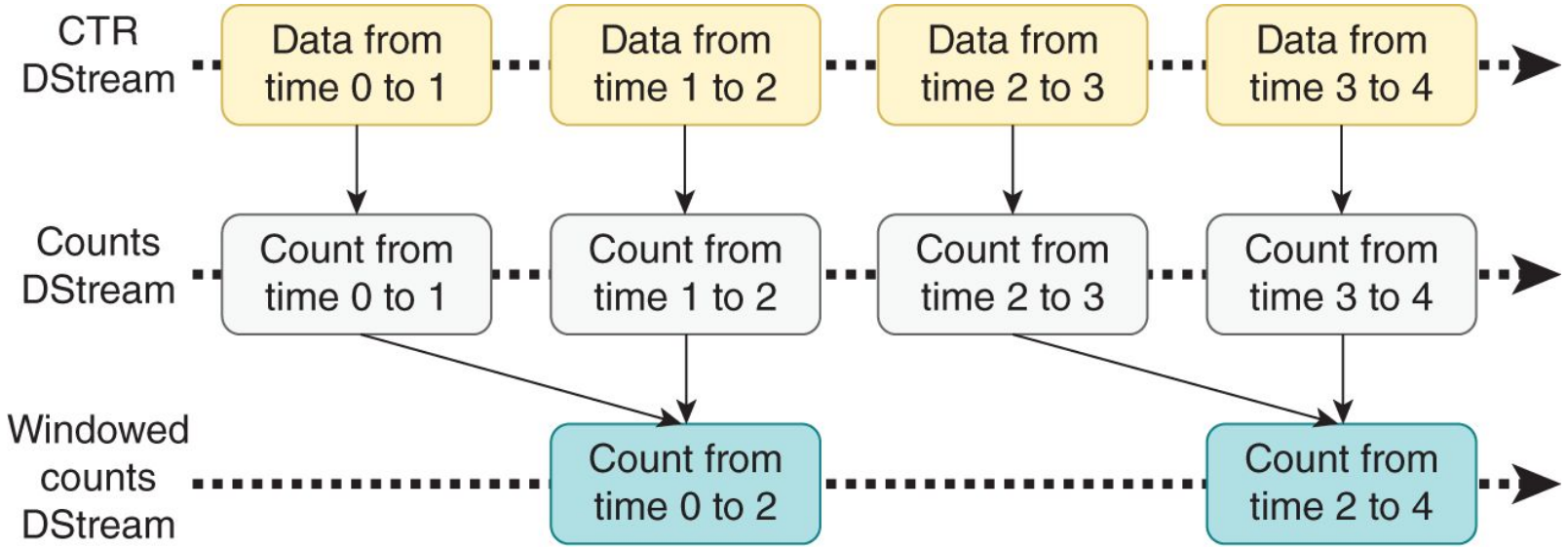
Source: Aytas, Y.(2021). Designing Big Data Platforms. Wiley.

Apache Flink



Source: Aytas, Y.(2021). Designing Big Data Platforms. Wiley.

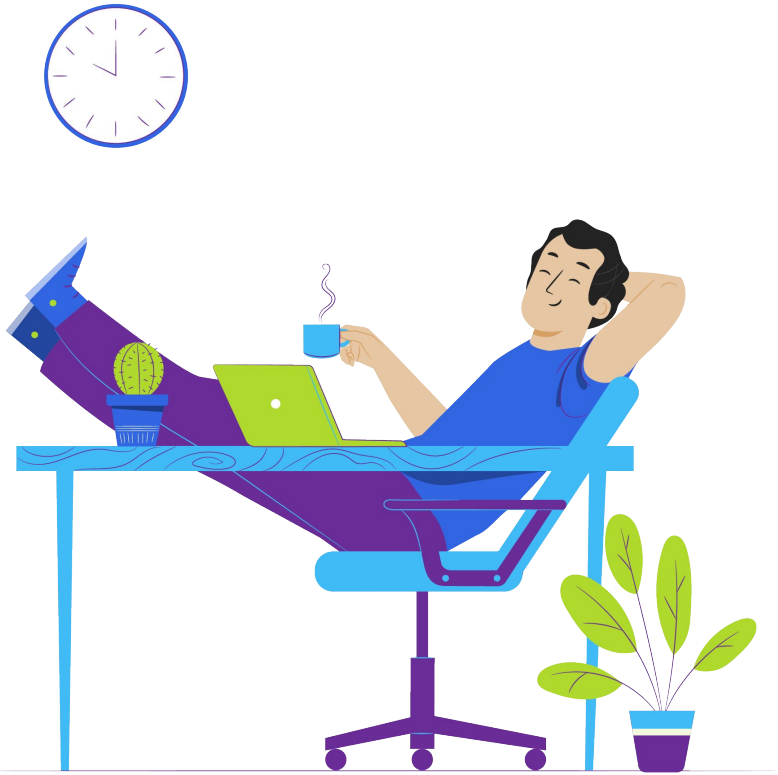
Spark Streaming



Source: Aytas, Y.(2021). Designing Big Data Platforms. Wiley.

Table of Contents

Break Time



Hands-On

Recommendation Engine with Apache Spark

Demo and Hands-On - Small Recommender System with Apache Spark

A link to the code repository will be provided during the session: [Repository]

Recap

The impact of Big Data in MLOps

1. Training better ML models:

- It provides the raw material that ML models need to learn. The larger and more diverse the data set, the better the model will be able to generalize to new data.
- It allows for the training of more complex ML models that can capture more subtle relationships in the data.

2. Improving model performance:

- Big Data can be used to identify and correct biases in ML models.
- Big Data can be used to monitor the performance of ML models in production and identify any issues that need to be addressed.

3. Accelerating the ML lifecycle:

- It can be used to automate the process of building, deploying, and managing ML models.
- It can be used to identify and share best practices for MLOps.

The impact of Big Data in MLOps

4. Democratizing ML:

- It is possible for more businesses to adopt ML, regardless of their size or resources.

- It allows for the development of cloud-based ML platforms that make it easier for businesses to get started with ML.

5. Supporting new ML applications:

- Big Data is essential for developing new ML applications, such as natural language processing and computer vision.

- Big Data is enabling the development of new ML architectures, such as deep learning and reinforcement learning.

Please, give us feedback!

Q&A

Thank you!