MONASH University

Stream Processing Technology

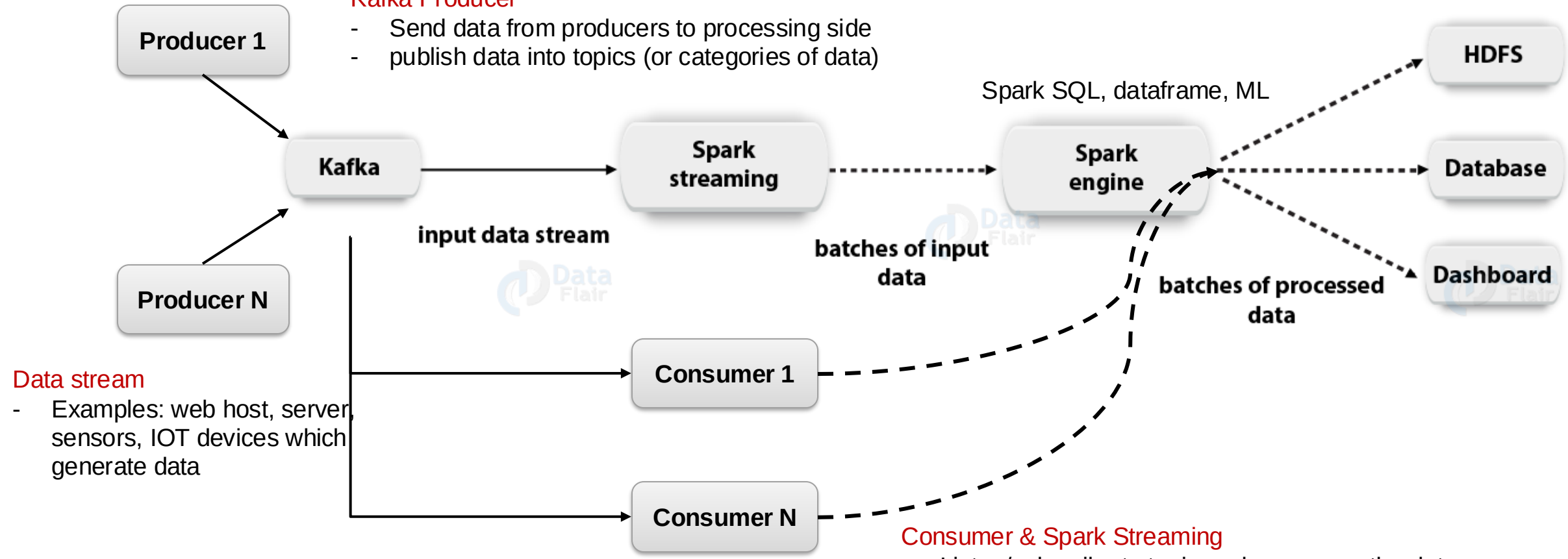
Streaming Data Processing

Prajwol Sangat, Ting Chee Ming
Updated by Jay Zhao(Sep/2024)

Streaming Platforms

Real-Time Streaming Architecture

Kafka-Spark Streaming Integration



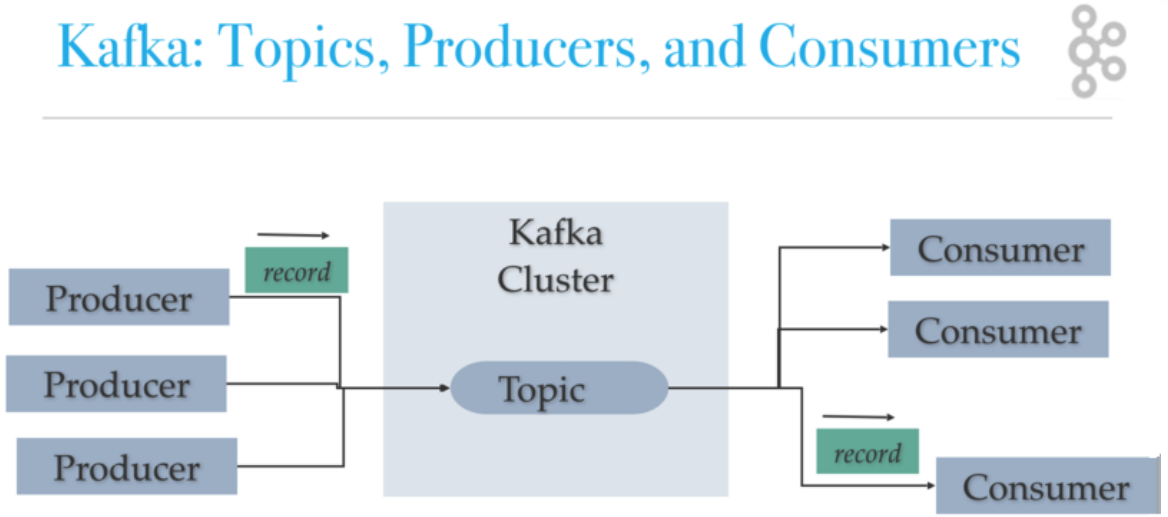
Kafka Producer
- Send data from producers to processing side
- publish data into topics (or categories of data)

Data stream
- Examples: web host, server, sensors, IOT devices which generate data

Consumer & Spark Streaming
- Listen/subscribe to topic and consumes the data

What is Apache Kafka?

- Publish-subscribe messaging system
- Scalable, Fault-tolerant
- Enables distributed applications
- Powers web-scale Internet companies such as LinkedIn, NetFlix, AirBnB, and many others.



https://dzone.com/articles/kafka-architecture

- ❑ Producer publish streams of data records into topics
- ❑ Consumers subscribe to the topics and process the stream of records

- ❑ Data in Kafka is stored in topic
  - Topic is category/feed name where records are stored
  - A topic is associated with a log – data structure on disk
  - Each topic is indexed and stored with timestamp

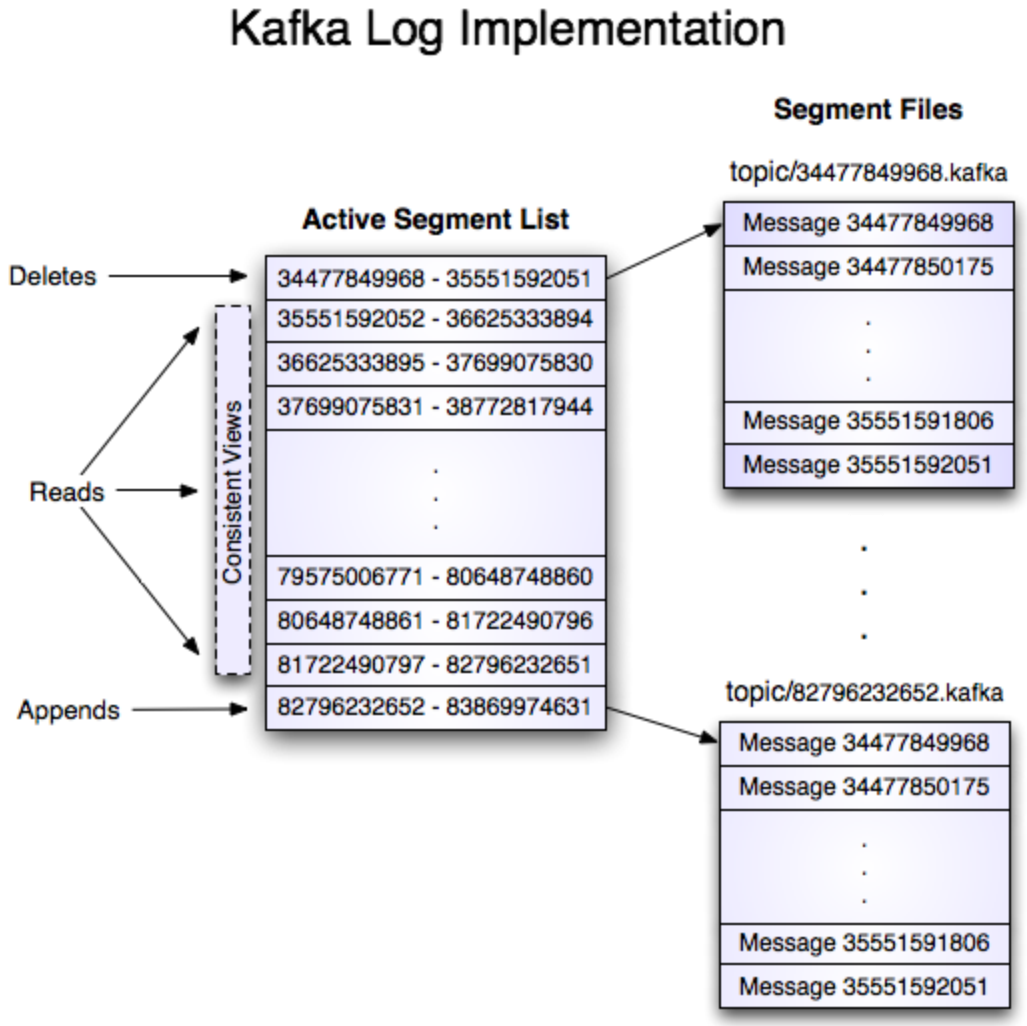
- ❑ Kafka is run as a cluster comprised of one or more servers (called brokers)

  - A broker receive messages from producers and stores them on disk keyed by unique offset.
  - A broker allows consumers to fetch messages by topic, partition and offset

https://kafka.apache.org/081/documentation.html

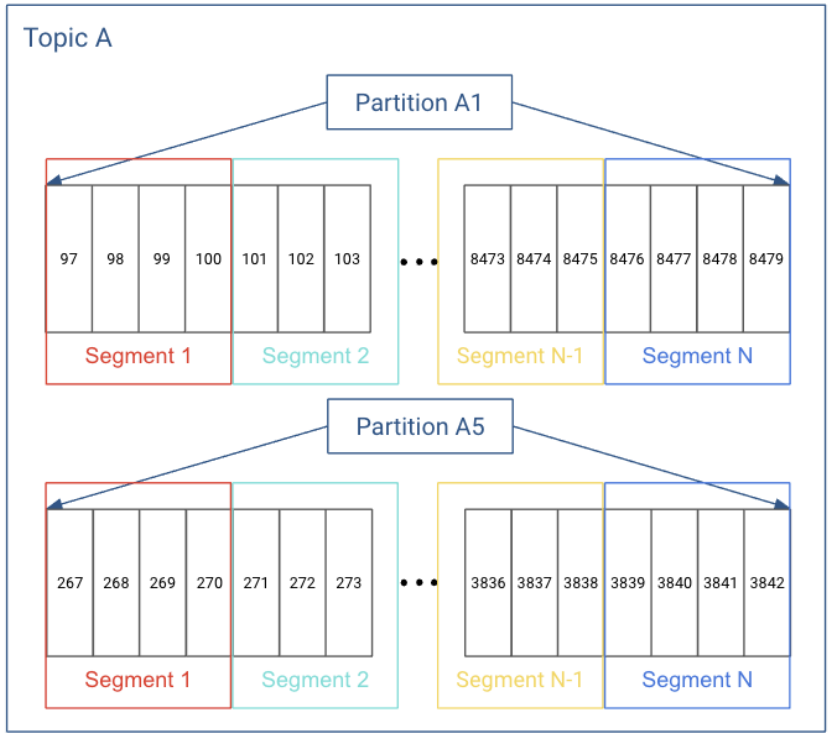
How does Kafka Work?

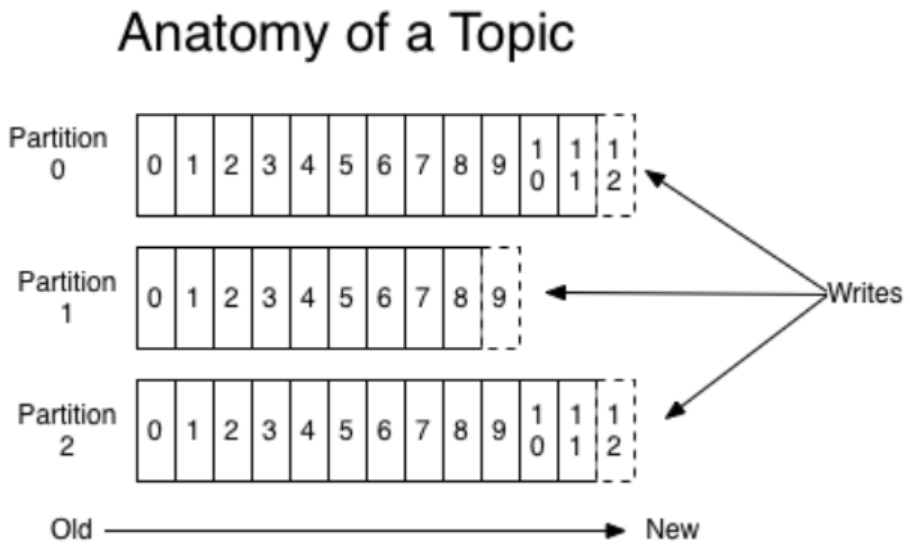
- Topics represent commit log data structures stored on disk
  - ➢ Topics are divided into partitions, each partition is further divided into segments
  - ➢ Each segment has a log file to store the actual message
  - ➢ Log – time-ordered, sequence of messages that is continually appended (each log entry can be array of bytes)
  - ➢ Partitions are replicated and distributed over servers in Kafka cluster (brokers) for fault tolerance (when a server in the cluster fails so messages remain available)
  - ➢ Messages stay around for a configurable period of time (i.e., 7 days, 30 days, etc.).
    - Can recover lost messages during time out or lost connection



https://medium.com/@durgaswaroop/a-practical-introduction-to-kafka-storage-internals-d5b544f6925f

Kafka Storage Structure





Messages are written to it in an append-only fashion

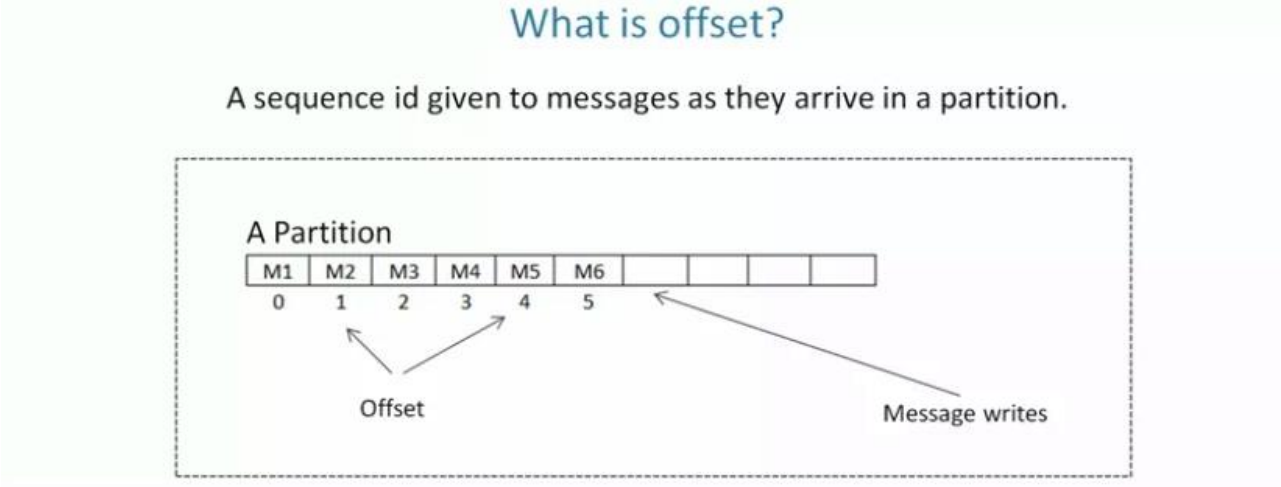
- Topic is logical grouping
- Partition is actual unit of data storage

- Every piece of data stored in segment file is a message
- Each message in partition is uniquely identified by an ID called offset

Image source: https://docs.cloudera.com/runtime/7.2.8/kafka-overview/topics/kafka-overview-logs-and-log-segments.html
https://kafka.apache.org/081/documentation.html

What Kafka doesn't do?

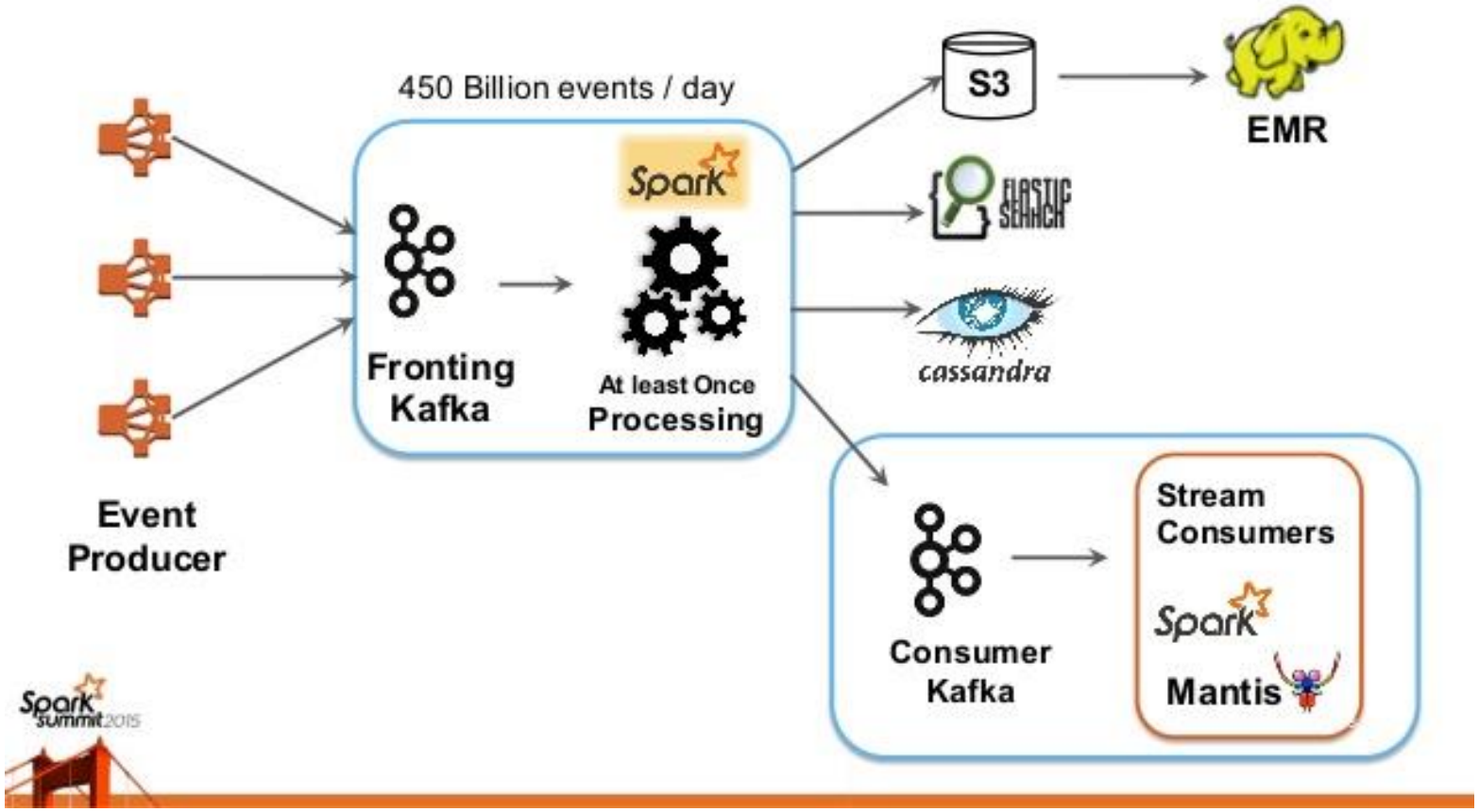
- Kafka does not have individual message IDs. Messages are simply addressed by their offset in the log.
- Kafka also does not track the consumers that a topic has or who has consumed what messages.
- There are no deletes. Kafka keeps all parts of the log for the specified time.



https://shubhamagtech.home.blog/2019/07/31/kafka-offset-management/

Kafka and big data at web-scale companies

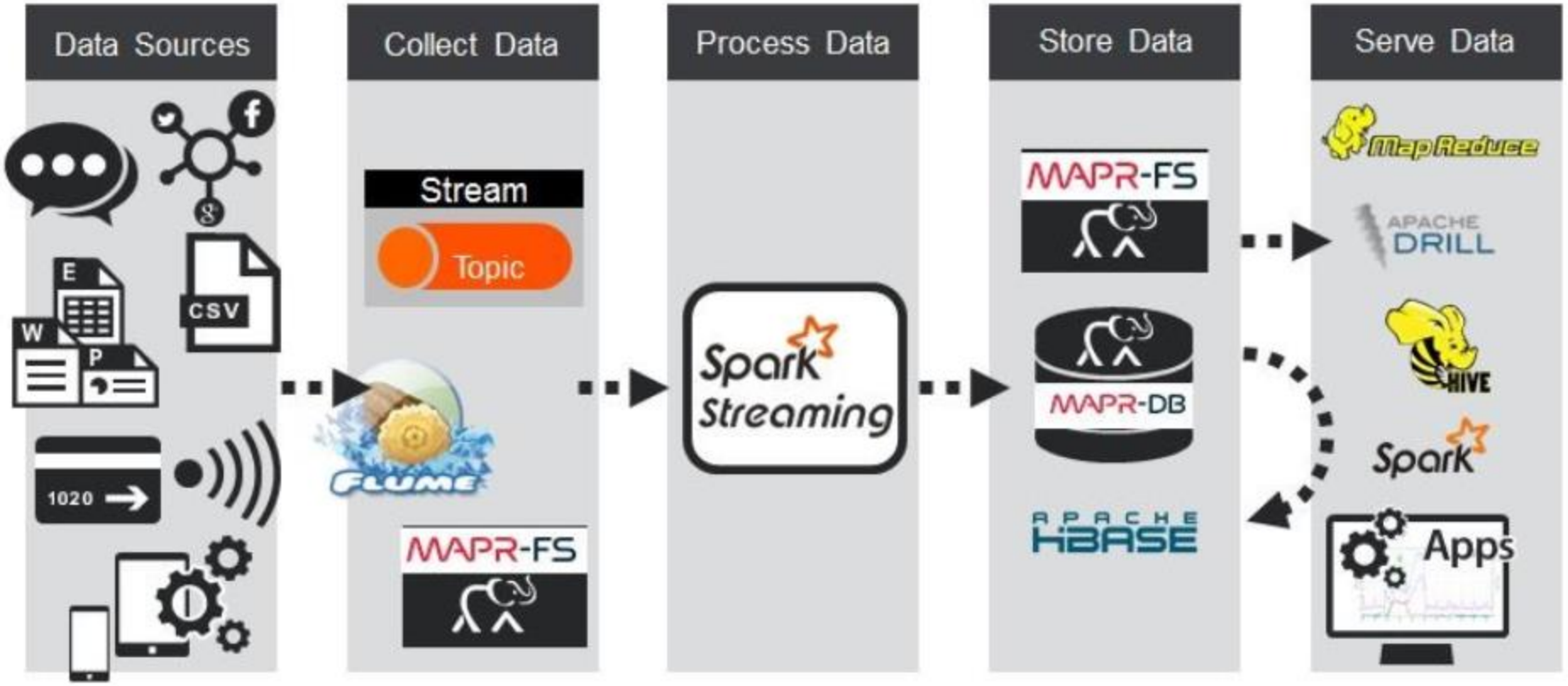
- **Big Data ingestion at Netflix**



https://www.slideshare.net/SparkSummit/spark-and-spark-streaming-at-netfix-sedakar-daxini

Kafka and big data at web-scale companies

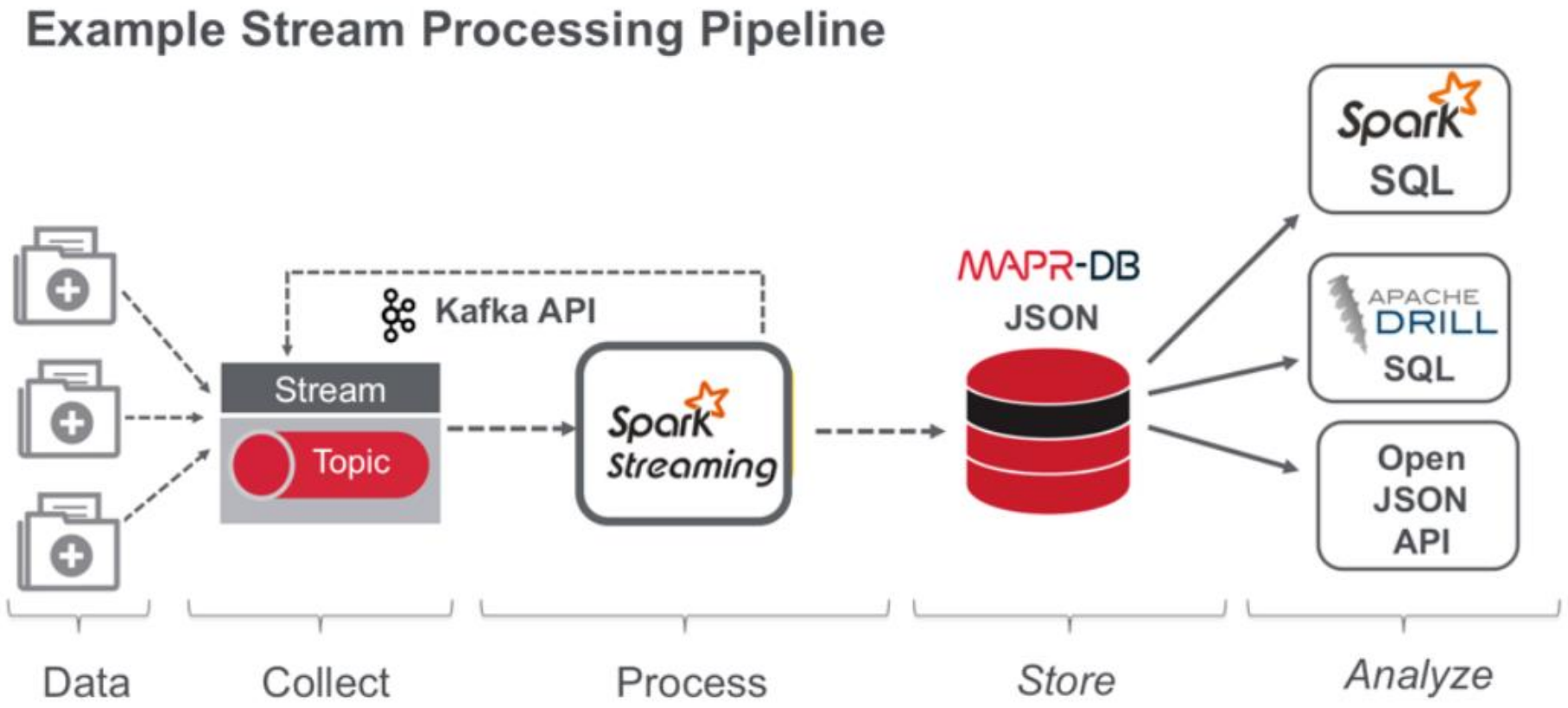
- **BP OIL USE CASE :**



https://www.linkedin.com/pulse/real-time-streaming-data-pipelines-apache-kafka-spark-steven-murhula/

Kafka and big data at web-scale companies

- **Transform, Store and Explore Healthcare Dataset**



https://mapr.com/blog/streaming-data-pipeline-transform-store-explore-healthcare-dataset-mapr-db/

Should you use Apache Kafka?

- Kafka fits a class of problems that many web-scale companies and enterprises have, but just as the traditional message broker is not a one-size-fits-all solution, neither is Kafka.
- If you're looking to build a set of resilient data services and applications, Kafka can serve as the source of truth by collecting and keeping all the "facts" or "events" for a system.

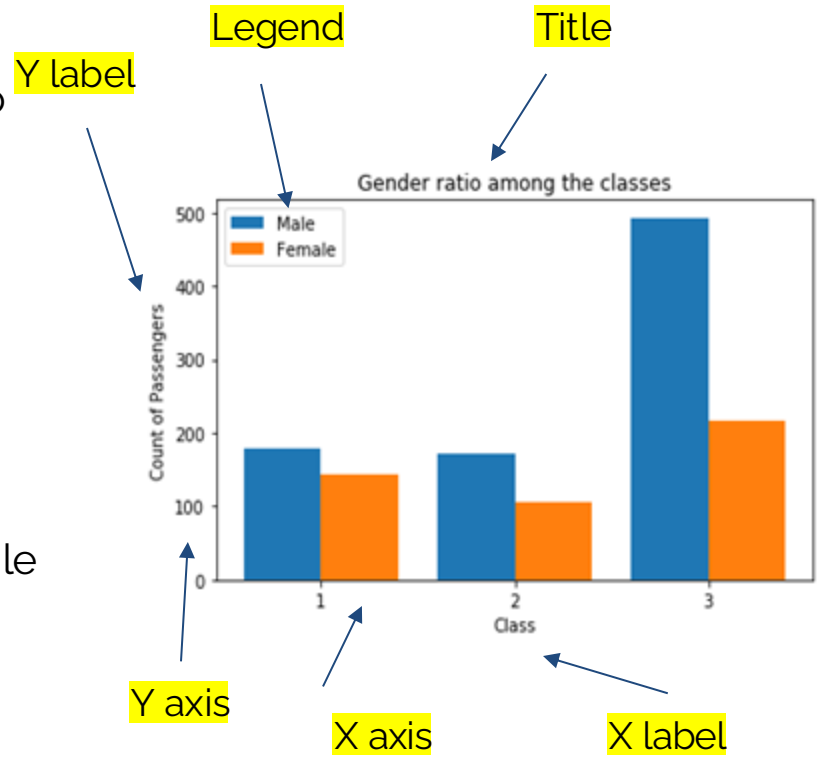
https://techbeacon.com/app-dev-testing/what-apache-kafka-why-it-so-popular-should-you-use-it

Exploratory Data Analysis

- **Exploratory Data Analysis (EDA)** is an approach to process data to summaries their main characteristics, often with visual and non-visual methods.”
- **Visual Methods:** For example; Scatter Plots, Box Plots, Histograms etc.
- **Non-visual Methods:** For example; Central Locations, Variability, Correlation etc.

Exploratory Data Analysis

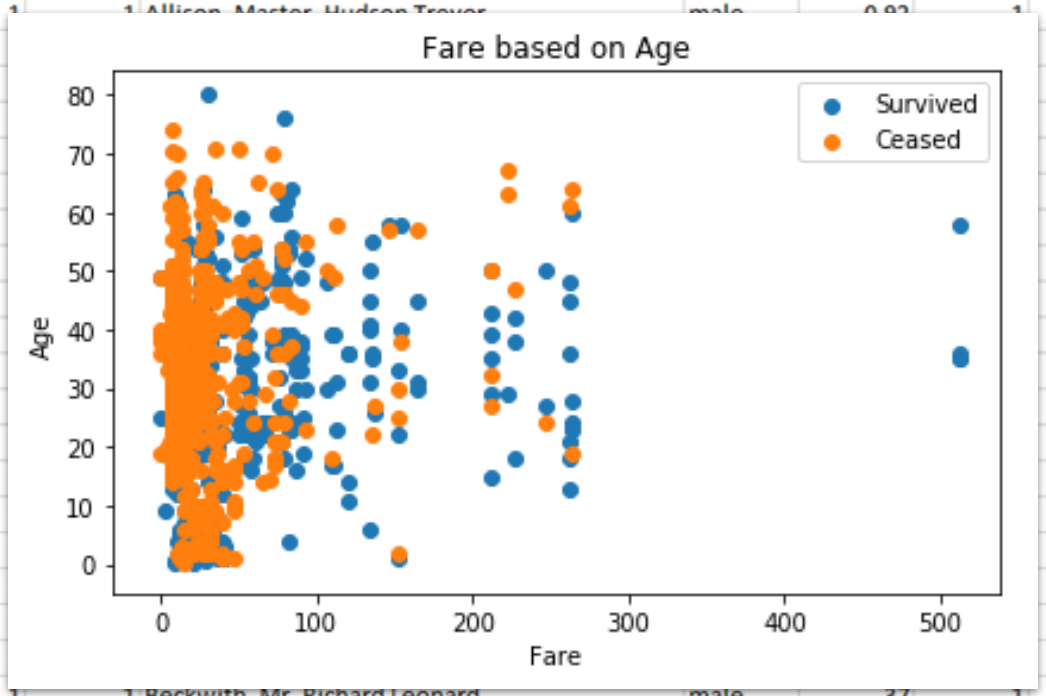
- **Factors to be aware of in visualising the data**
  - *Visual effects*
    - Includes usages of appropriate shapes, colors and size to represent the analysed data.
  - *Coordinate system*
    - It helps to organise the data points within the provided coordinates.
  - *Data type and scale*
    - Choose the type of data such as numeric or categorical.
  - *Informative interpretation*
    - Helps create visuals in an effective and easily interpretable manner using labels, title, legends and pointers.



Exploratory Data Analysis

- **Why use Visual Methods?**

pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home_dest
1	1	Allen, Miss. Elisabeth Walton	female	29	0	0	24160	211.3375	B5	S	2		St Louis, MO
1	1	Beckwith, Mr. Richard Leonard	male	37	1	1	11751	52.5542	D35	S	5		New York, NY
1	1	Beckwith, Mrs. Richard Leonard (Sallie Monypeny	female	47	1	1	11751	52.5542	D35	S	5		New York, NY
1	1	Behr, Mr. Karl Howell	male	26	0	0	111369	30	C148	C	5		New York, NY



Exploratory Data Analysis

- **Plotting Graphs**
  - Depends on data size
  - Plotting libraries run on a single machine
    - Expects data that fits into the main memory
    - Uses vectors or matrices as data source
  - Too much data leads to…
    - Main memory problems
    - Performance problems

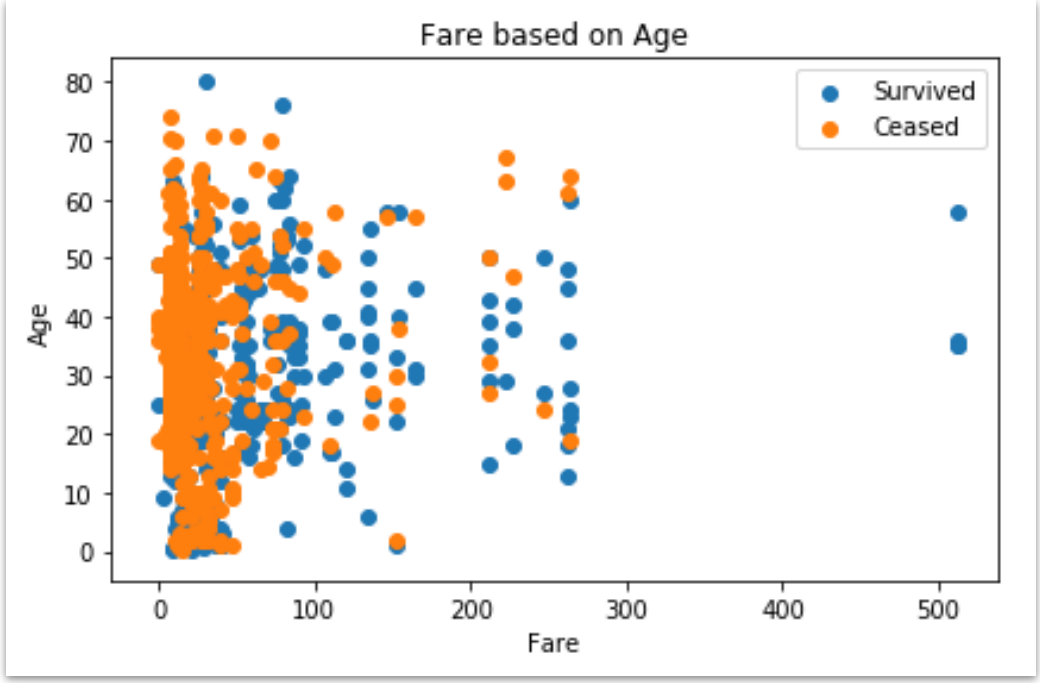
  - **What to do if the data does not fit in the main memory?**

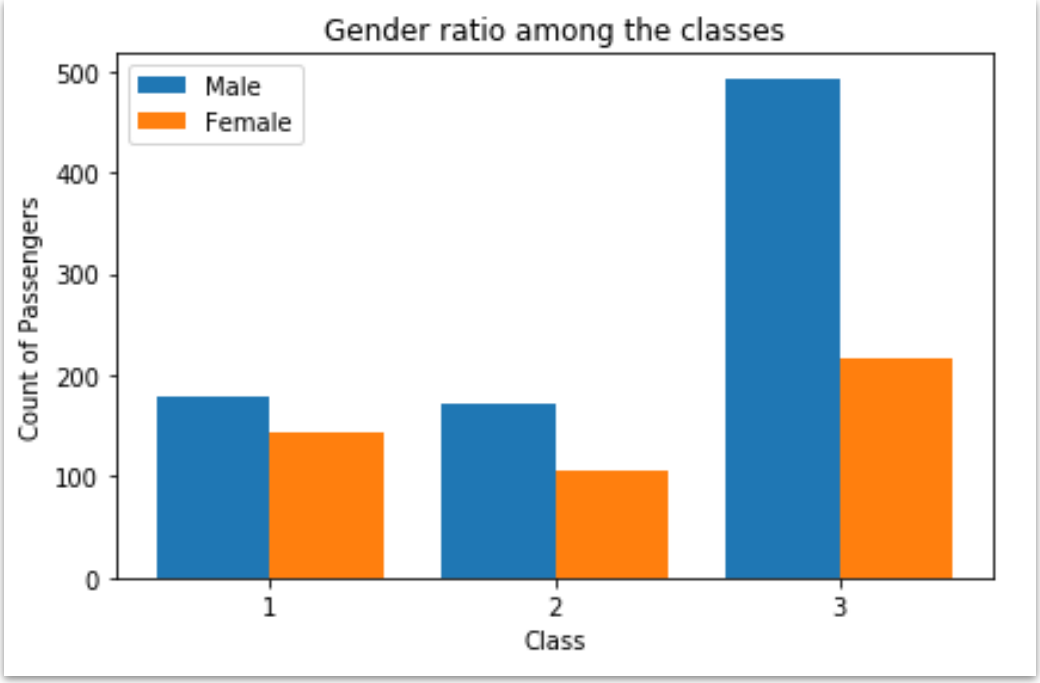
Exploratory Data Analysis: Sampling

- **Plotting Graphs**
  - **What to do if the data does not fit in the main memory?**
    - **Sampling**
      - **What?**
        - The process of selecting a subset of rows
        - Due to inherent randomness, preserves most properties of original data
      - **Why?**
        - Reducing the volume of data (enhance space and speed)
        - Reduces computational cost
      - **How?**
        - Simple random sampling with/without replacement
        - Stratified Sampling (subgroup/strata sampling)
        - Balanced Sampling, bias reduction in ML
        - And many more.

Exploratory Data Analysis

- **Is sampling always appropriate?**





Thank You

Questions?