THE DATA SCIENCE LAB
Introduction to Data Stream Processing

COM 490 – Spring 2025

Week 9

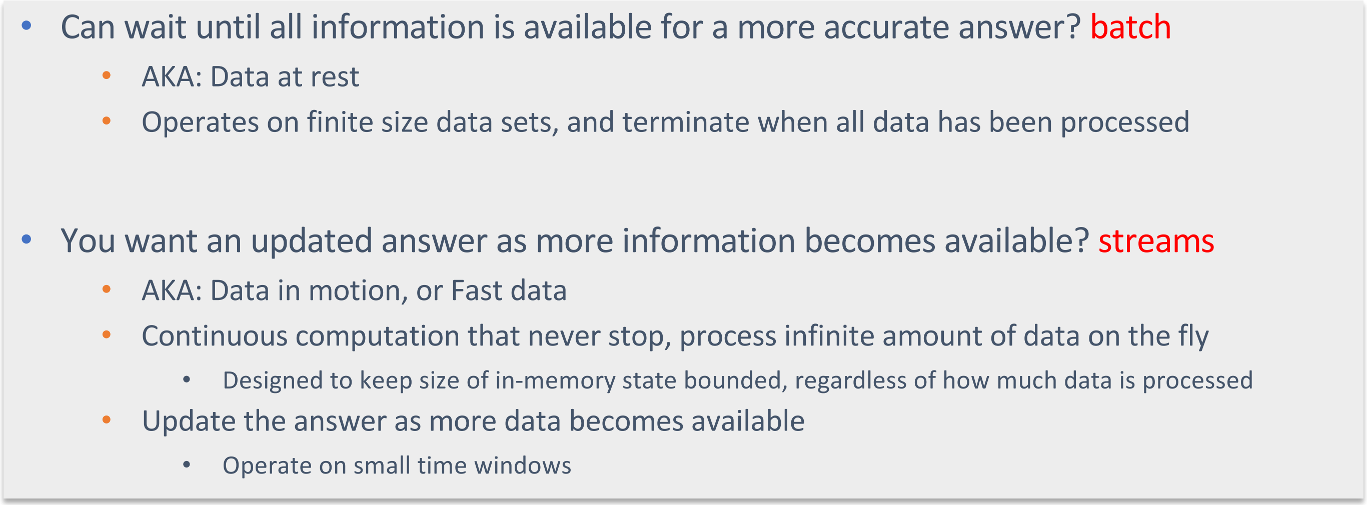
EPFL

Stream Processing Module

- Objectives
  - Review concepts of stream processing
  - Experiment with typical tools for
    - Data ingestion and processing
- Week 9
  - Concepts
  - Experiments
- Week 10
  - Advanced topics
    - Operations on streaming data (joins)
    - Time constraints
- Week 11
  - Analytics on data at rest and data in motion

Why Stream Processing?

- **Reminder from module 2 (Big Data)**
  - Batch vs Stream



- Can wait until all information is available for a more accurate answer? batch
  - AKA: Data at rest
  - Operates on finite size data sets, and terminate when all data has been processed
- You want an updated answer as more information becomes available? streams
  - AKA: Data in motion, or Fast data
  - Continuous computation that never stop, process infinite amount of data on the fly
    - Designed to keep size of in-memory state bounded, regardless of how much data is processed
  - Update the answer as more data becomes available
    - Operate on small time windows

Why Stream Processing?

- **Relevance (vs batch)**
  - Insight more valuable shortly after events happen
    - (Near) real-time: from milliseconds to seconds, or minutes
  - It allows faster reaction
    - Detecting patterns, setting alerts
  - Some data is naturally unbounded (e.g. sensor data)
  - Resource constraints (storage and compute)
    - process large large volumes of data arriving at high velocities
    - Retain only what is useful
  - Continuous processing

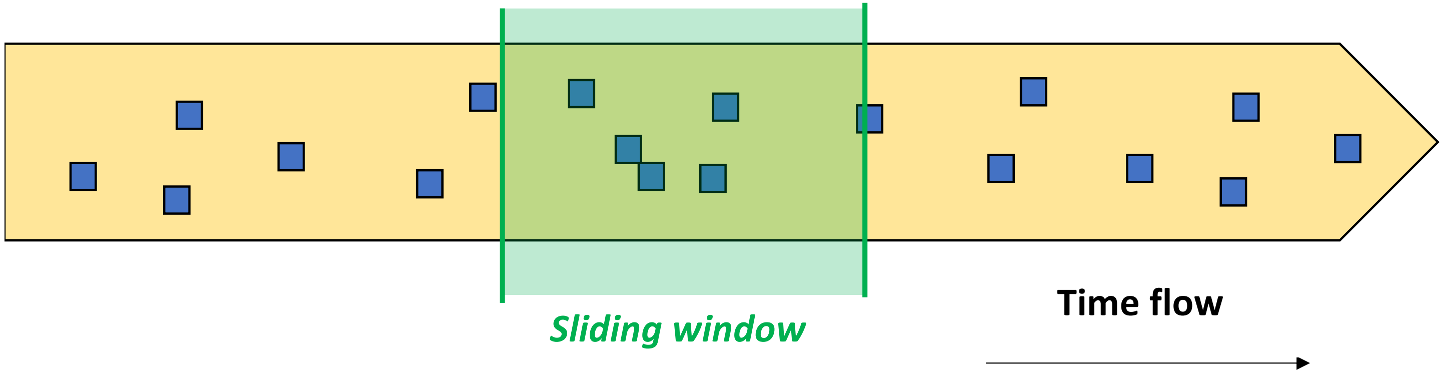
Applications of Stream Processing

- Computing
  - Log analysis,
  - Detection of DoS attacks,
  - Scaling service capacities
- Real-time monitoring
  - Fraud detection (credit cards),
  - Intrusion detection (surveillance)
- Sensor data processing
  - Weather,
  - Transportation
  - Traffic
  - Patient health
- Social media
  - Trend analysis
- Industry
  - Process optimization
  - Predictive maintenance
  - Logistics
- Advertising and promotions
  - Contextualized to user behavior or geolocation
- Financial trading
  - Algorithmic trading
  - Risk analysis
- …

Constraints and challenges

- **Inputs**
  - **Time constraints**
  - **Data elements**
    - **Unbound**
    - **Unordered**
    - **Uncomplete**
- **Outputs**
  - **Approximate answers**

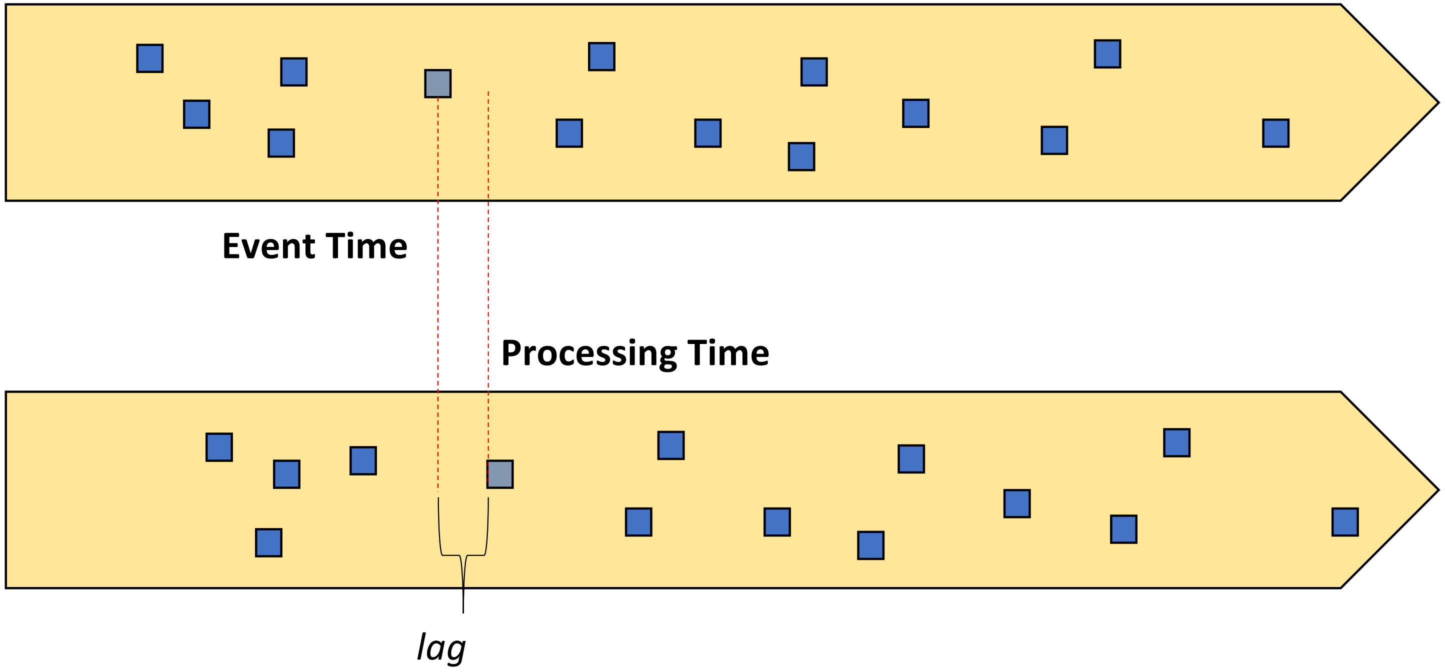
Sliding Window



Related Concepts

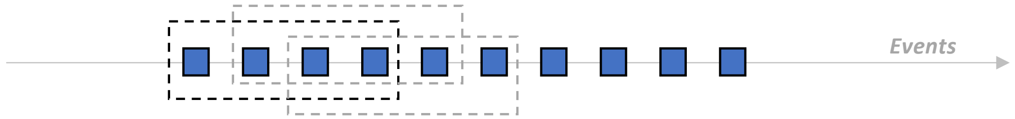
- **Event Time vs Processing Time**
- **Types of Windows**
  - **Sliding**
  - **Tambling**
  - **Time-based vs count-based**
- **Window Operations (transformations)**
- **Stateful / Stateless Operations (transormations)**

Related Concepts

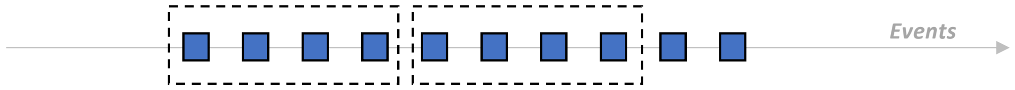


Related Concepts

Sliding Windows

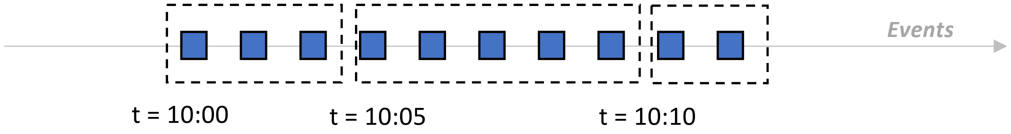


Tumbling windows

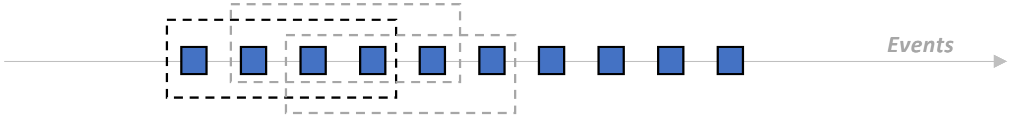


Related Concepts

Time-based windows



Count-based windows



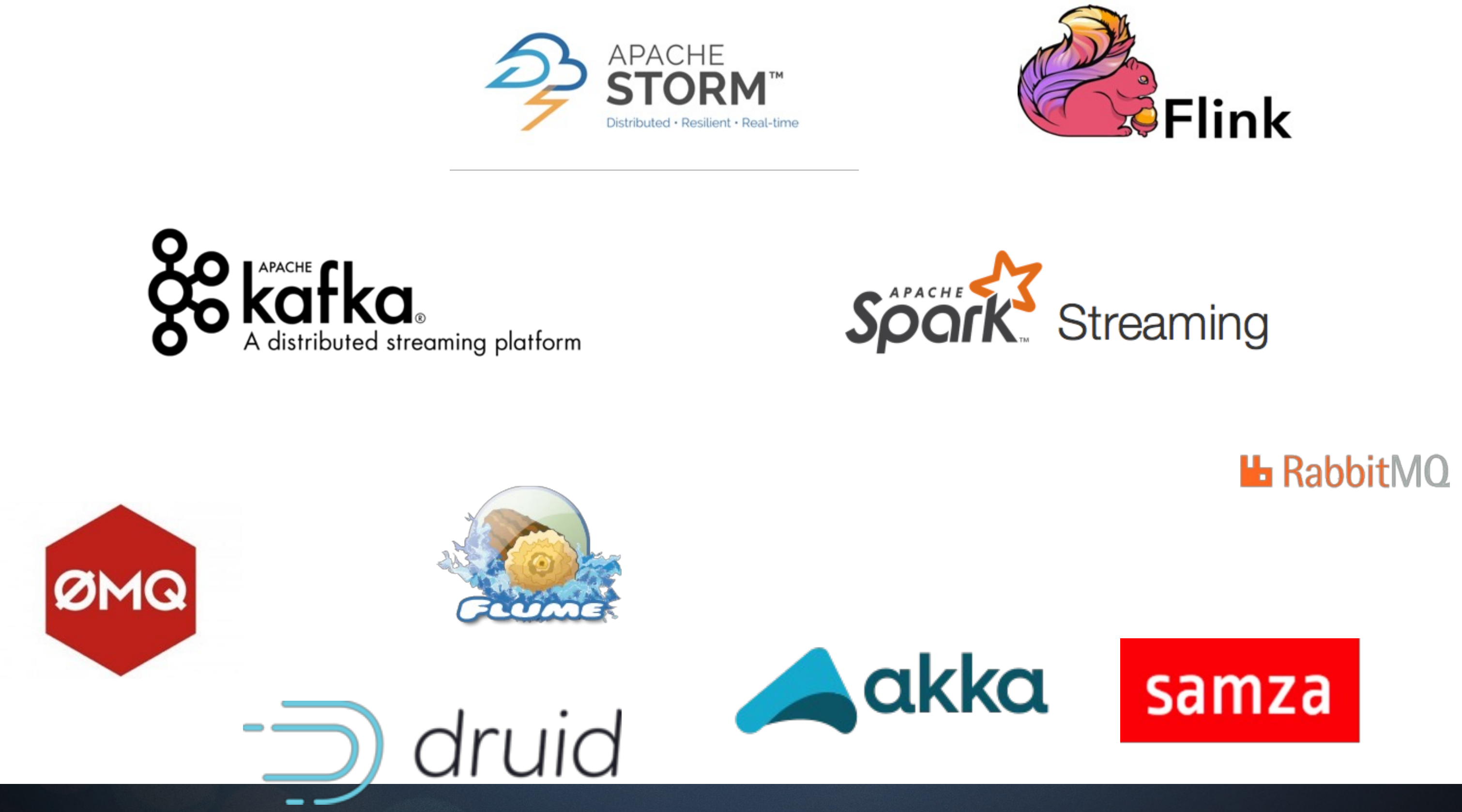
Related Concepts

- **Window Operations (transformations)**
  - **Aggregations**
    - **Sums, averages, counts, maximum, …**
  - **Filtering**
    - **By type, IDs, …**
  - **…**

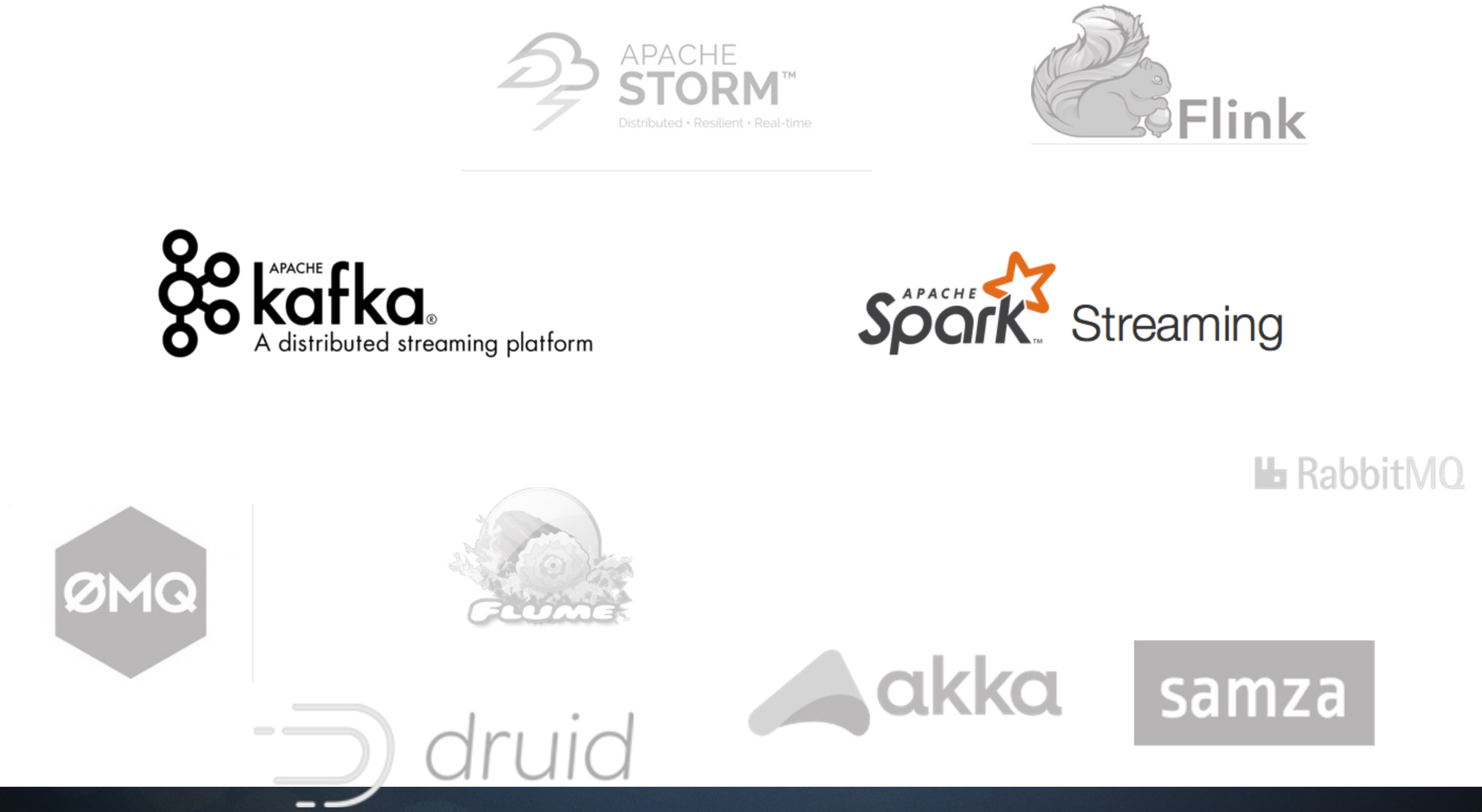
Related Concepts

- **Stateful vs Stateless Operations (Transformations)**
  - **Stateful: need to memorize records or partial results**
    - **e.g. Min, Max and average temperature of a sensor**
  - **Stateless: rely only on information within the window**
    - **e.g. Average temperature of sensor over last 5 minutes**

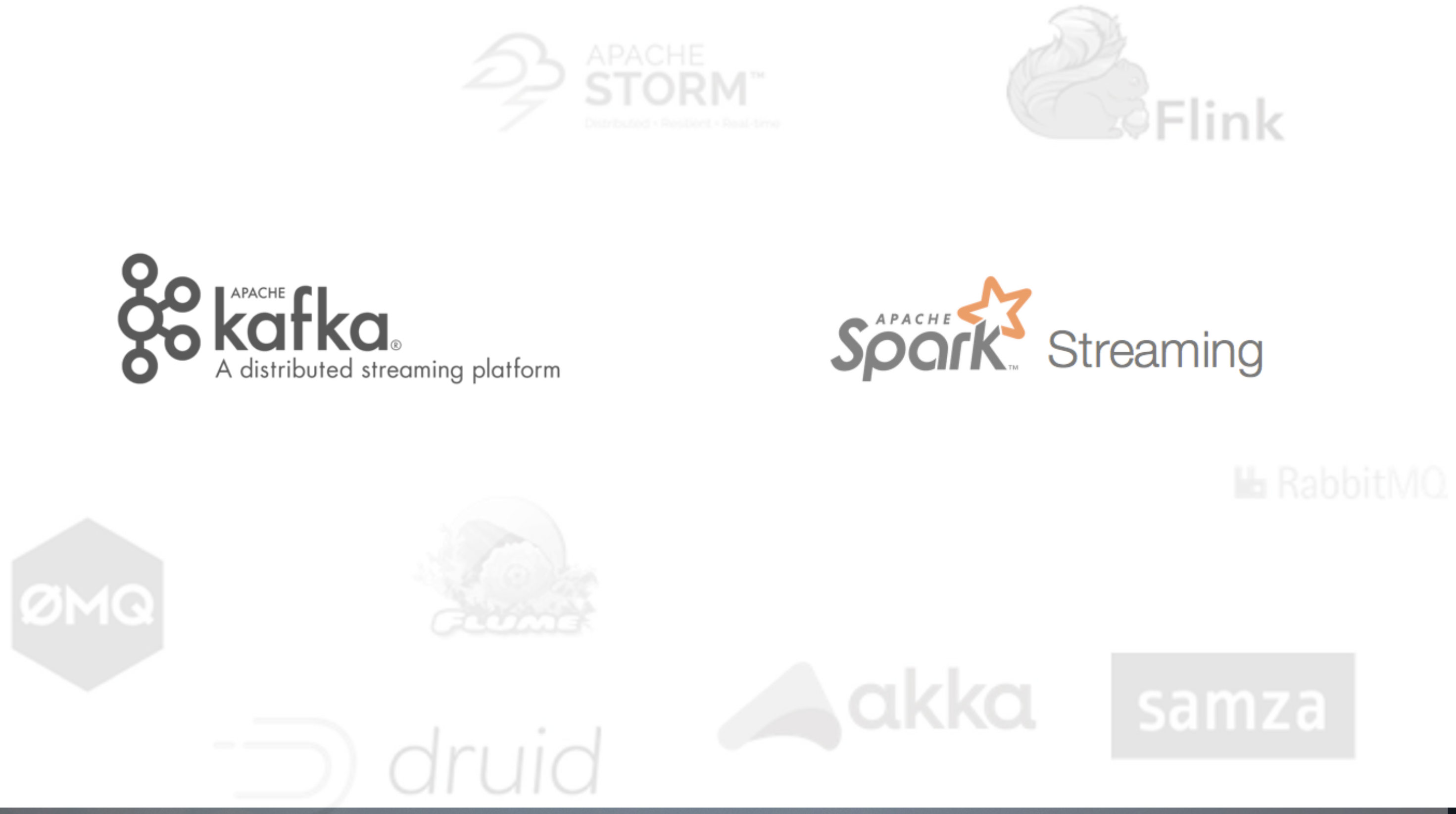
Stream Processing - Tools



Stream Processing - Tools

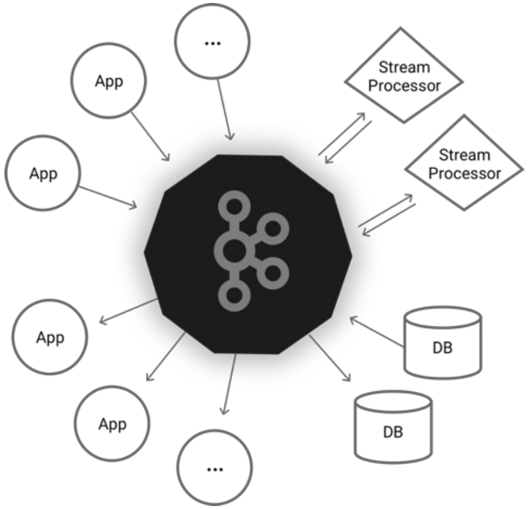


Stream Processing - Tools



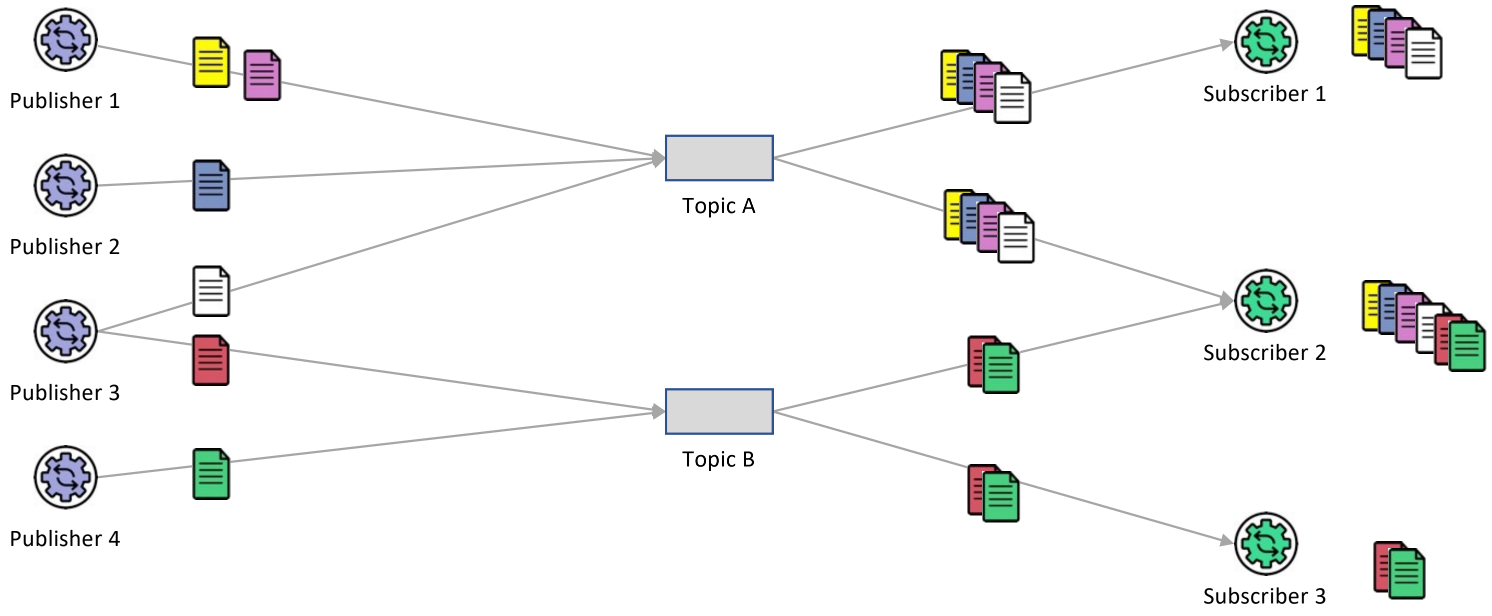
Kafka

- Messaging system
  - Publish & Subscribe
- Distributed
- Fault tolerant
- Scalable (large data volumes)
- Real-time
- Low latency



Kafka

- Concept of Publish/Subscribe messaging



Icons made by Freepik from www.flaticon.com

Spark Streaming

- Extension to Spark
  - Integrated with Spark API
- Scalable, fault tolerant
- Can read from multiple sources
- Apply ML algorithms to data streams

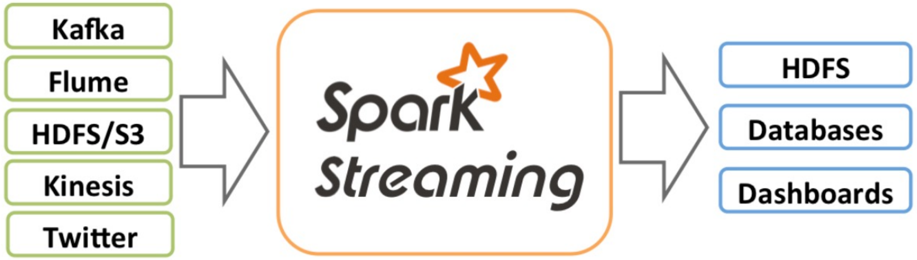
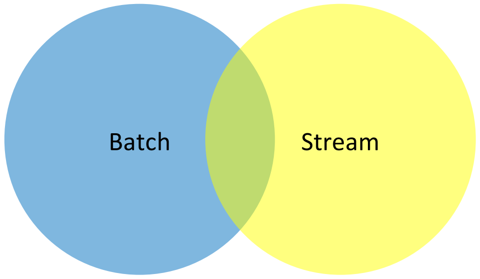


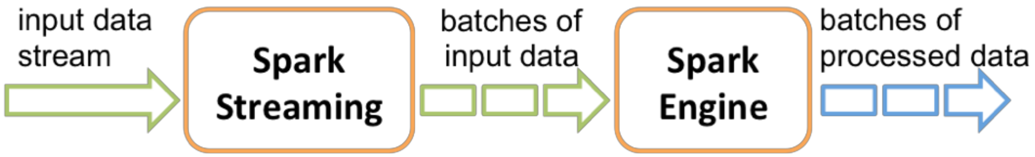
Image credits: https://spark.apache.org/docs/latest/streaming-programming-guide.html

Spark Streaming

- How it works
  - Micro-batches processing



Micro-batch



- DStream: continuous stream of data
  - Created from inputs (e.g. Kafka) or derived from other Dstreams
  - Continuous series of RDDs
  - Supports (many) transformations similar to RDDs
    - (map, count, join, etc)

Image credits: https://spark.apache.org/docs/latest/streaming-programming-guide.html

Exercises

- Documentation and Resources
  - Spark Streaming Programming Guide [1]
  - Kafka Documentation [2]
- Practical Exercise (with solutions)

https://dslabgit.datascience.ch/course/2025/module-4a

`kafka-exercise-solution.py`

[1] https://spark.apache.org/docs/latest/streaming-programming-guide.html

[2] https://kafka.apache.org/documentation/

Exercises

1. Message queue

- Introduction to Apache Kafka
- Topics
  - Creation
  - Publish
  - Subscribe
- Synthetic example
  - Producing and consuming data through Kafka

Exercises (next week)

2. Stream Processing with Spark Streaming and Kafka

- How to properly setup Spark Streaming
- Resume synthetic exercise
- Connect to Kafka and consume stream
- Window operations
- Use real public stream

Useful references

[1] https://spark.apache.org/docs/latest/streaming-programming-guide.html

[2] https://kafka.apache.org/documentation/