Wrocław University of Science and Technology

Large Scale Data Processing

Lecture 4 – Spark - streaming

dr hab. inż. Tomasz Kajdanowicz, Piotr Bielak, Roman Bartusiak, Krzysztof Rajda

November 18, 2021

HR EXCELLENCE IN RESEARCH

Wroclaw University of Science and Technology

Overview

Spark - streaming

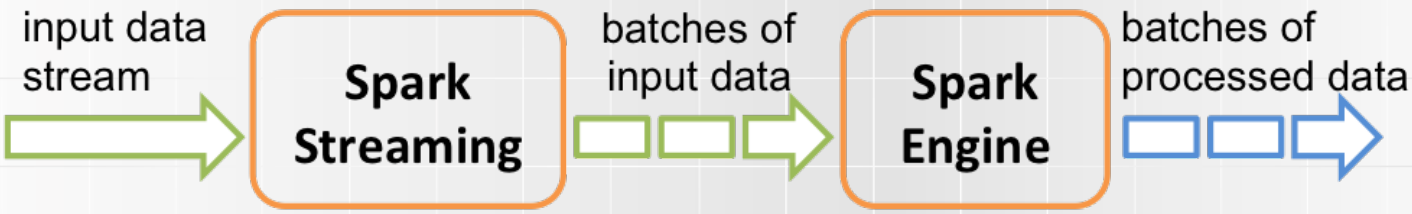
Spark - streaming

Architecture

Spark - streaming

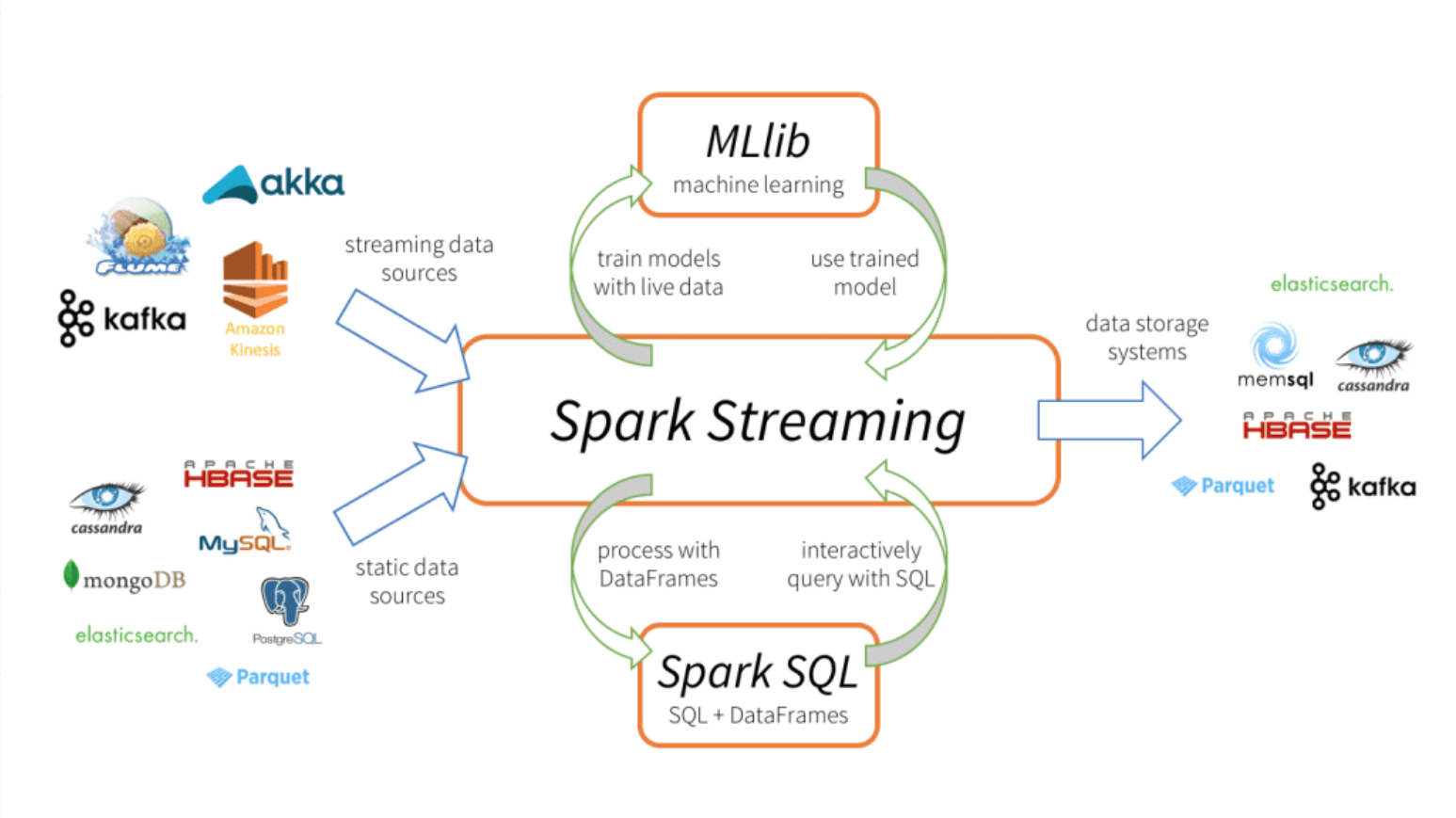
Architecture

Spark - streaming



Architecture

Spark - streaming



Concepts - DStream

Spark - streaming

- Utilizes low-level Spark API
- discretized-stream
- micro-batches
- "old"

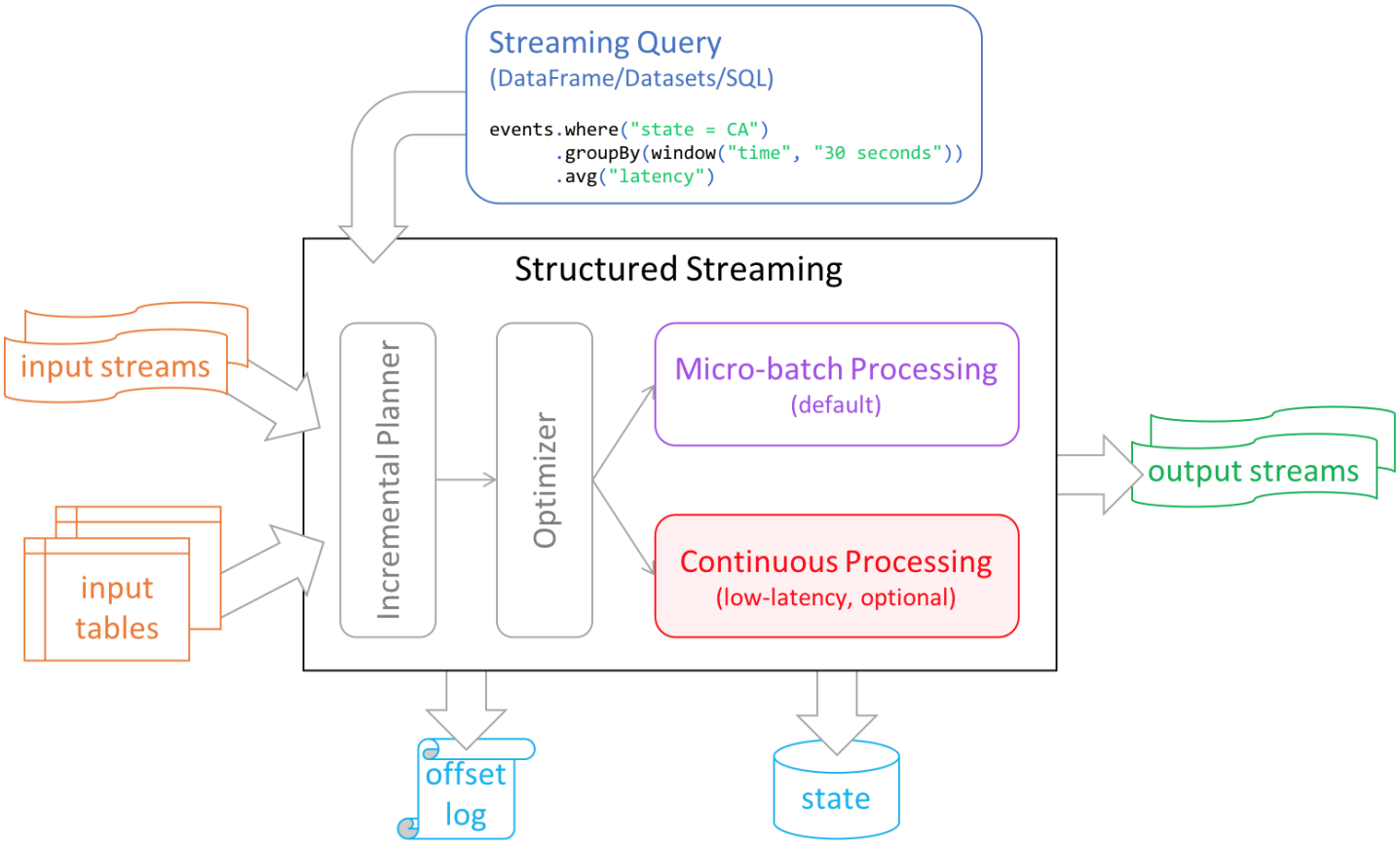
Concepts - Structured Streaming

Spark - streaming

- On top of Spark SQL
- batches and continous
- streaming DataSets
- streaming DataFrames
- "new"

Concepts - Structured Streaming

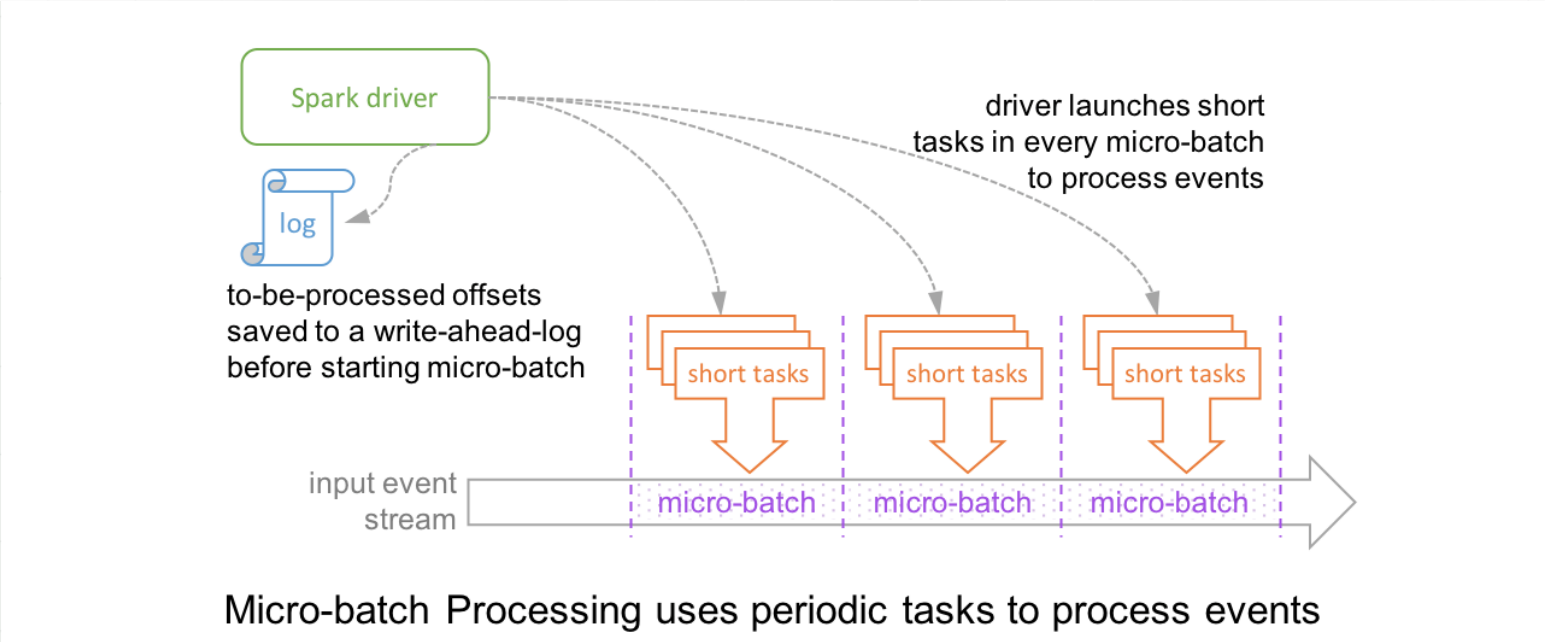
Spark - streaming



Structured Streaming processing modes: Micro-batch and Continuous Processing

Concepts - Structured Streaming

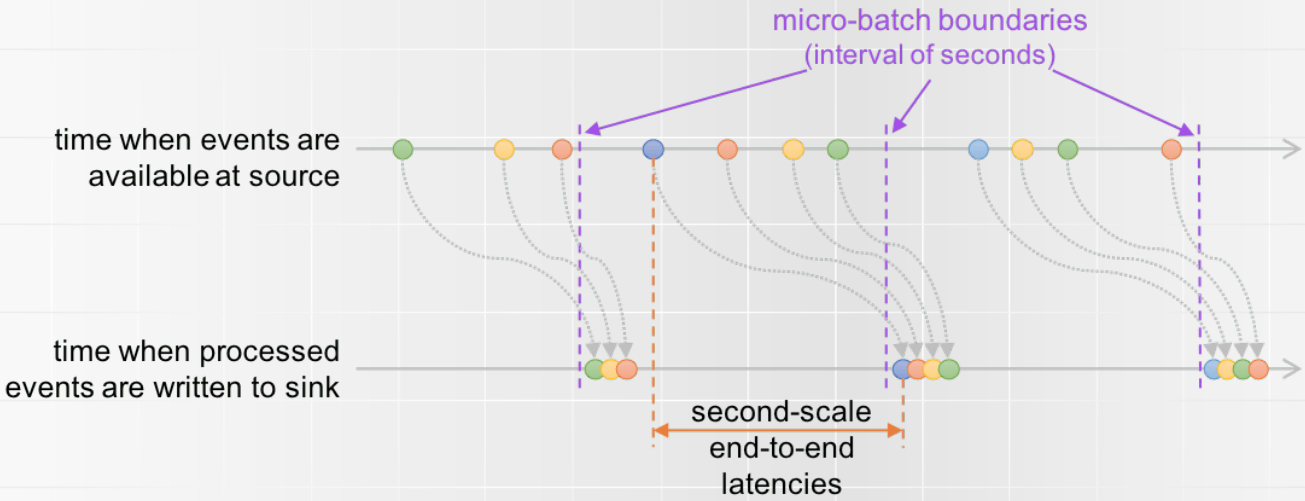
Spark - streaming



Micro-batch Processing uses periodic tasks to process events

Concepts - Structured Streaming

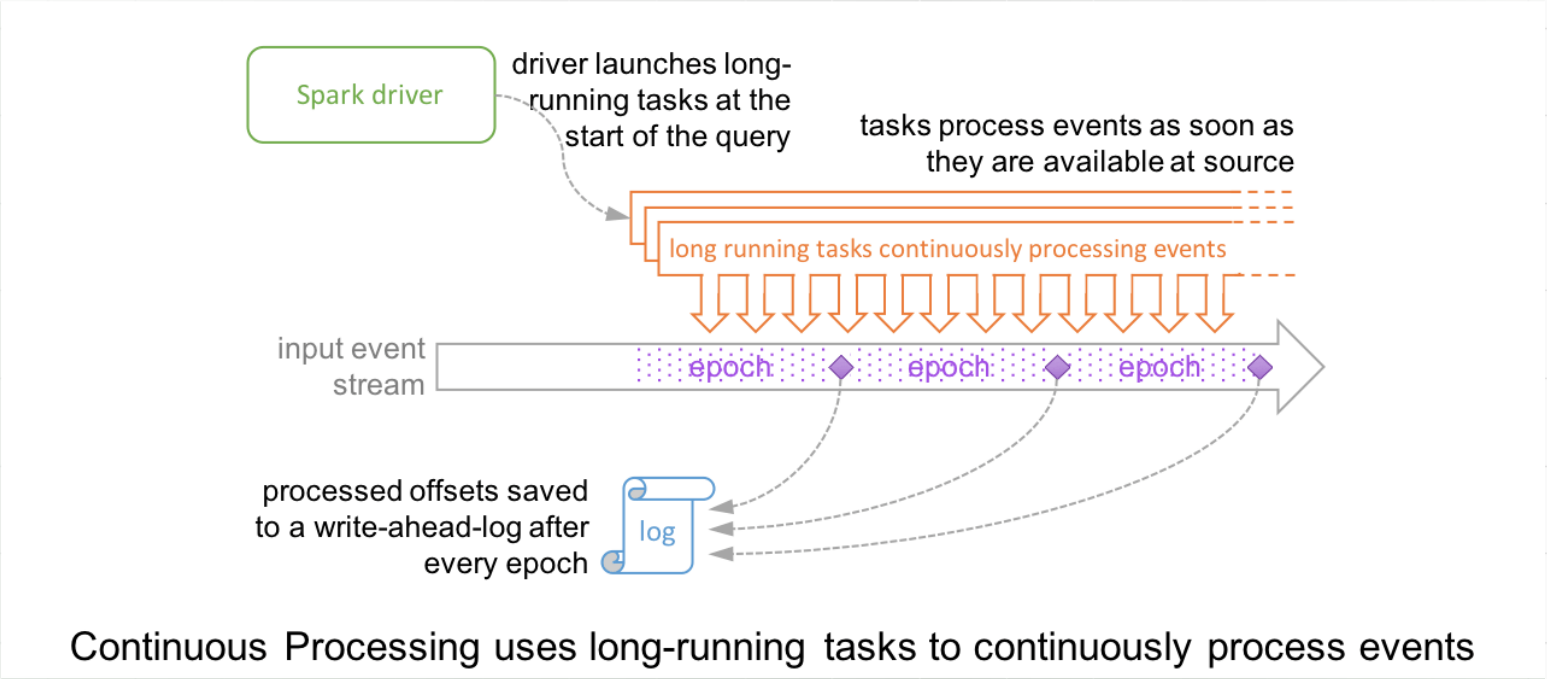
Spark - streaming



Second-scale end-to-end latencies with Micro-batch Processing

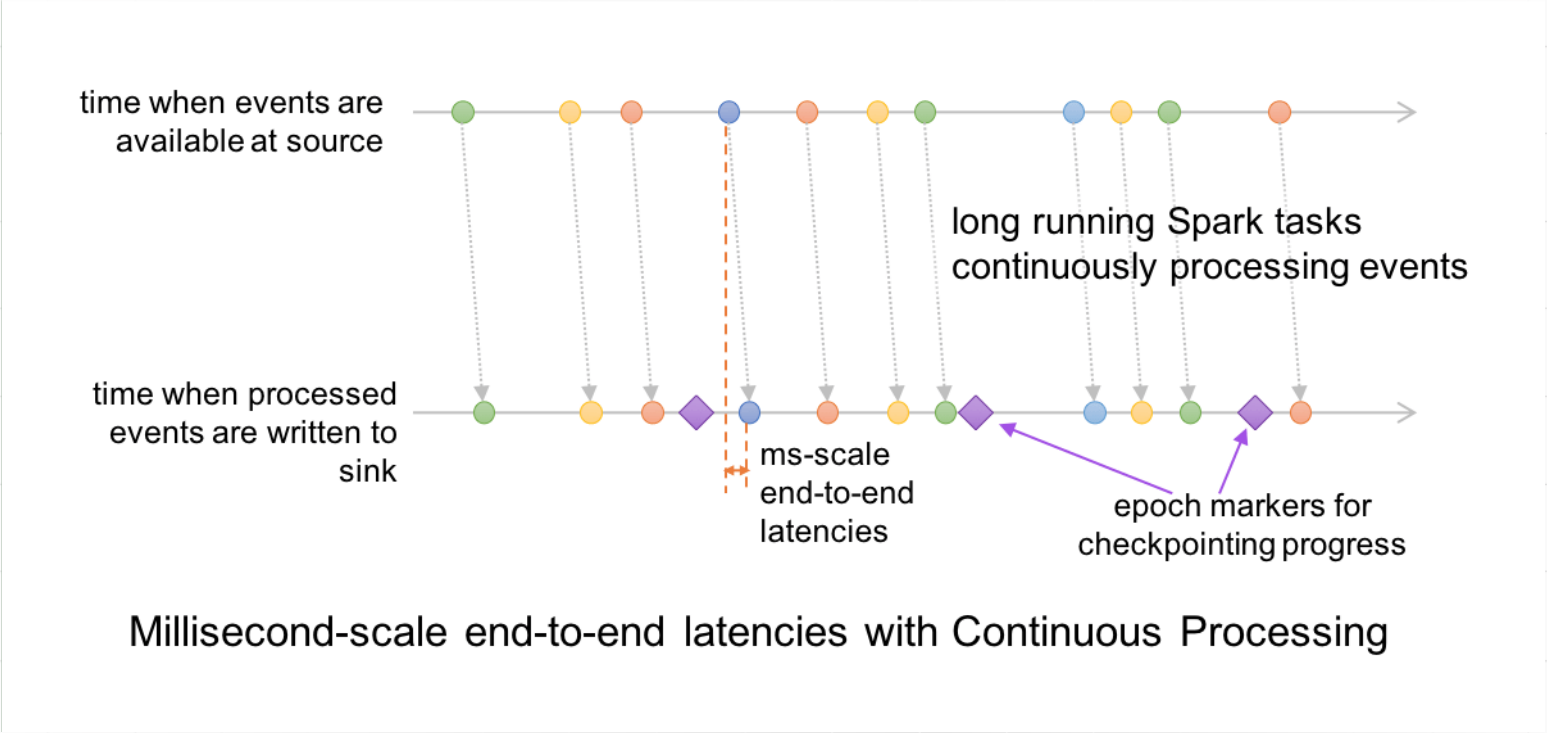
Concepts - Structured Streaming

Spark - streaming



Concepts - Structured Streaming

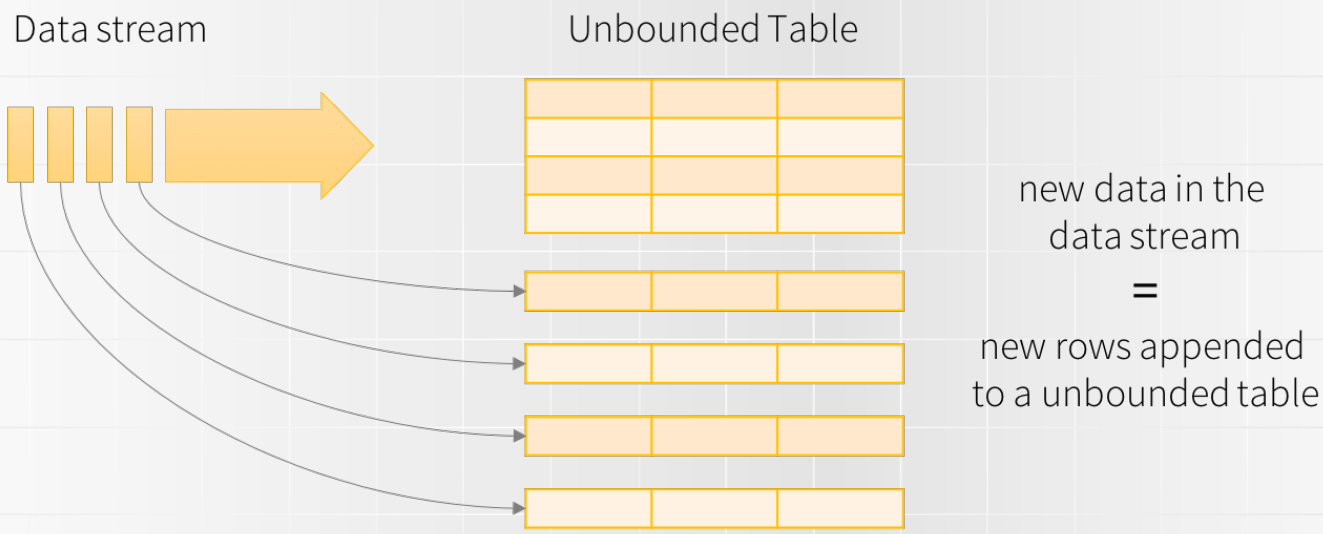
Spark - streaming



Millisecond-scale end-to-end latencies with Continuous Processing

Concepts - Data frames

Spark - streaming



Data stream as an unbounded table

Concepts - Input and Receiver

Spark - streaming

- Each data input is consuming resources
- Each input is connected with receiver
- Receiver is responsible for fetching data and pushing to Spark
- Number of cores must be tuned to number of receivers
- Too low number of cores - no resources to process data as all resources will be occupied by receivers

Concepts - Kafka

Spark - streaming

- Distributed event streaming platform
- Events - records (key+value+timestamp)
- Events are organized in topics
- Producers and consumers
- Each topic is partitioned by event key - scalability
- Events in partitions are guaranteed to be returned to consumers in write order
- Offset - tracking already consumed events by consumer in given partition
- Data can be replicated

Concepts - Kafka

Spark - streaming

- Kafka Clients need to connect to many nodes
- Addresses of nodes are fetched during connection to the cluster - subset is specified pragmatically, full list resolved automatically
- Remember about that aspect during connecting to clusters
- Telepresence - might be usefull

Concepts - Kafka

Spark - streaming

- Take care of memory
- Java Kafka Client consumes direct memory - before deserialization
- Direct memory is not part of heap
- By default that part of memory is same as heap size

Concepts - Custom source

Spark - streaming

- No available in Python
- Implement simple interface
- onStart, onStop
- call store(element) for each element

Concepts - Reliability

Spark - streaming

- at-least-once, at-most-once, exactly-once
- Spark is observably exactly-once
- receivers can be reliable and unreliable - acknowledgements to the source

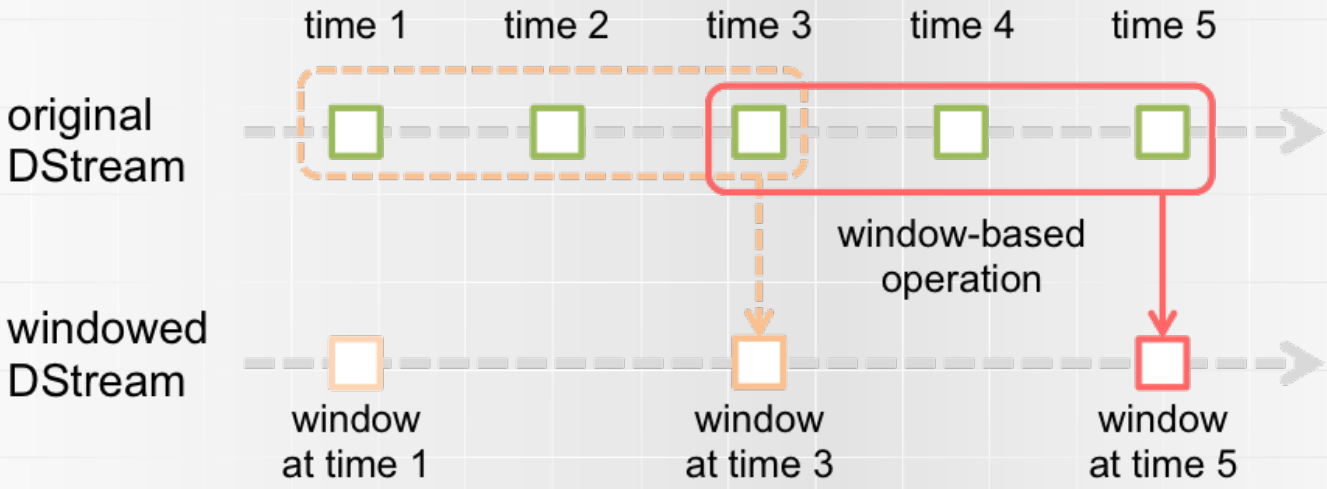
Concepts - Transformation

Spark - streaming

- Access underlying RDD
- Use-full when wanting to use static data with streams

Concepts - Window operations

Spark - streaming



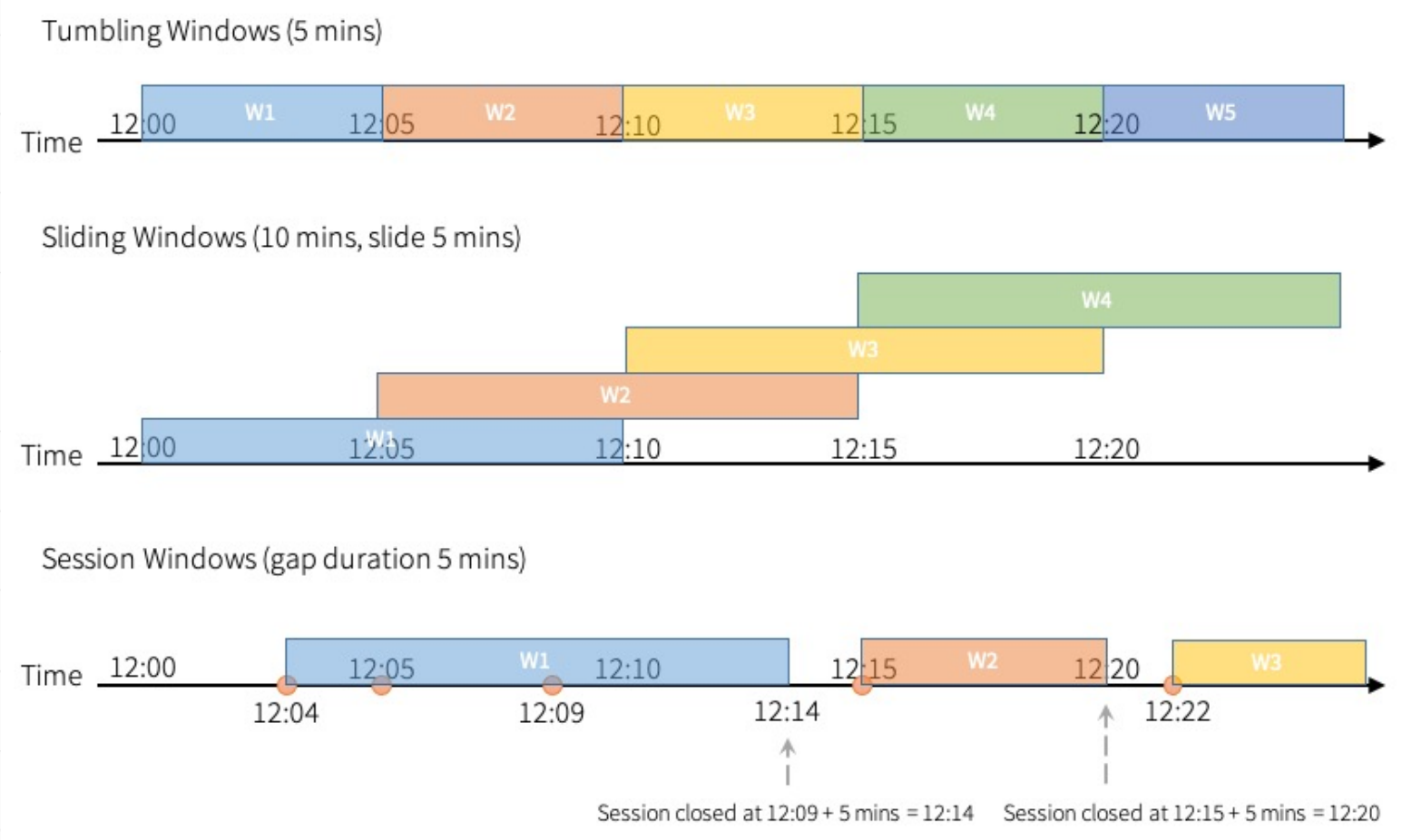
Concepts - Window operations

Spark - streaming

- tumbling
- sliding
- session

Concepts - Window operations

Spark - streaming



Concepts - Window operations

Spark - streaming

Tumbling:

- Windows size

Concepts - Window operations

Spark - streaming

Sliding:

- Windows size
- Interval length

Concepts - Window operations

Spark - streaming

Session:

- Gap size

Concepts - Joins

Spark - streaming

- stream - stream - based on microbatches
- stream - data - using transformations

Concepts - Outputs

Spark - streaming

- files
- HDFS
- access each micro-batch RDD
- external libraries

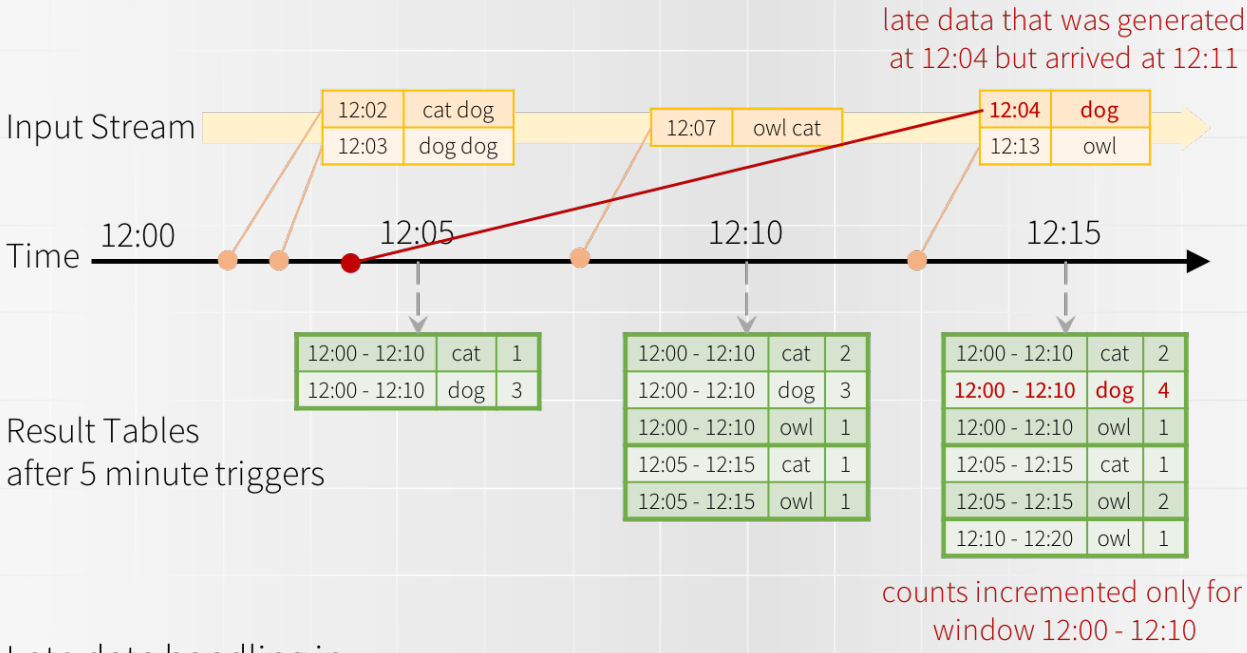
Concepts - watermarking

Spark - streaming

- how to handle late data
- spark is keeping track of current event time
- then it is used to manage internal states

Concepts - watermarking

Spark - streaming



Late data handling in
Windowed Grouped Aggregation

Wroclaw University of Science and Technology

Next week

Spark - streaming

- Flink - streaming

Wrocław University of Science and Technology

Large Scale Data Processing

Lecture 4 – Spark - streaming

dr hab. inż. Tomasz Kajdanowicz, Piotr Bielak, Roman Bartusiak, Krzysztof Rajda

November 18, 2021