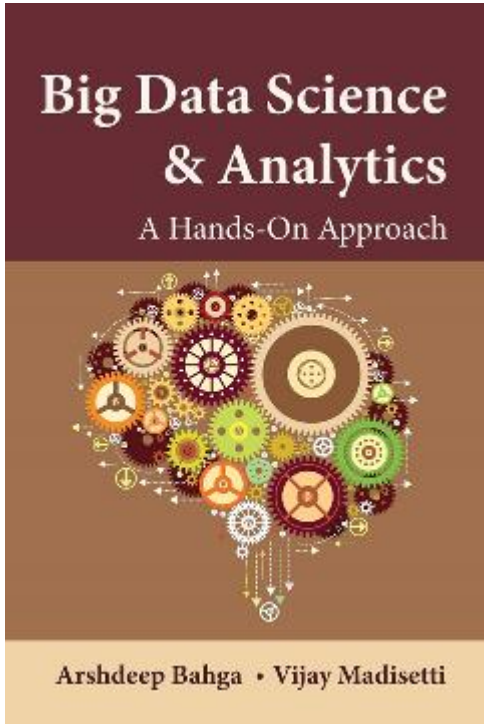
Real-time Analysis

Prof. Gheith Abandah

Reference



- Chapter 8: **Real-time Analysis**

- Arshdeep Bahga and Vijay Madisetti, **Big Data Science and Analytics: A Hands-On Approach**, 2019.
  - Web site: http://www.hands-on-books-series.com/

Outline

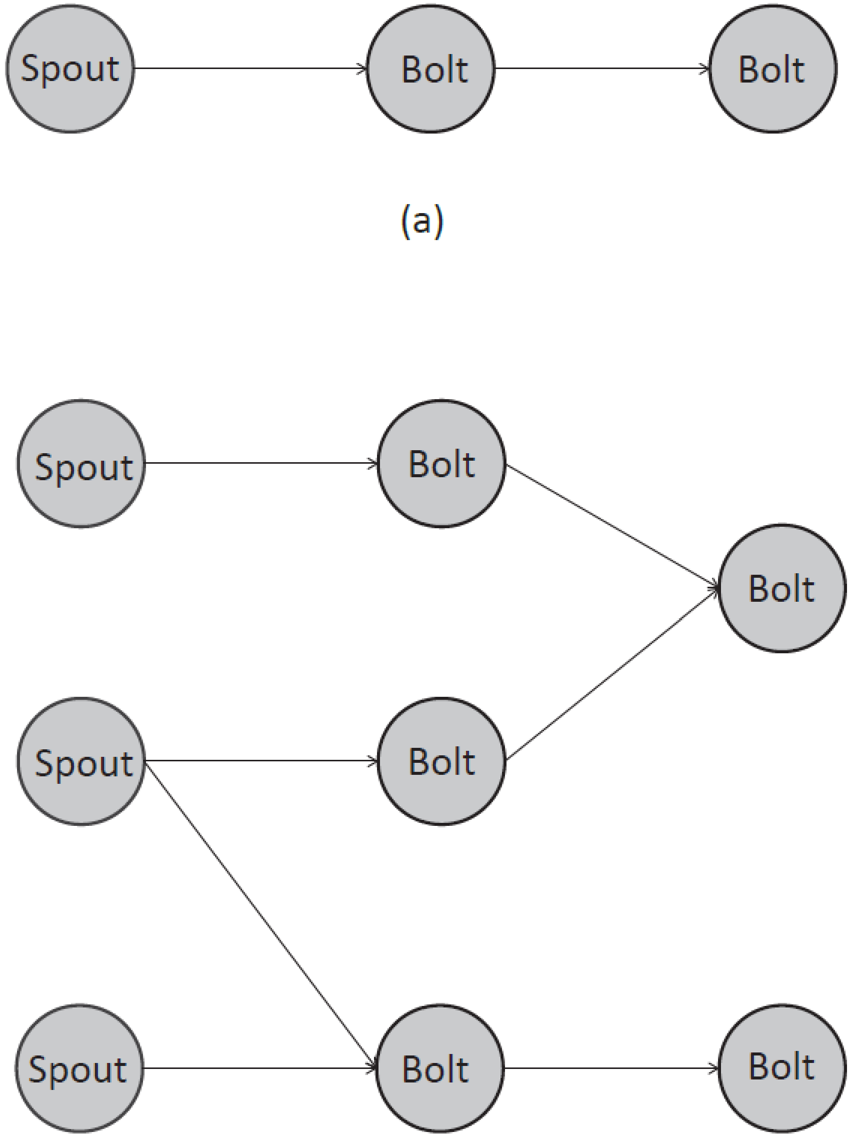
- Stream Processing with Apache Storm
- In-Memory Processing with Apache Spark

Stream Processing with Apache Storm

- Framework for **distributed** and **fault-tolerant real-time** computation that can be used for **real-time processing** of **streams** of data.
- **Ingests** data from a **variety of sources**.
- Storm is a **scalable**, **distributed** framework and offers **reliable** processing of messages.
- Designed to **run indefinitely** and process streams of data in real-time.
- Its processing latencies are in the order of **milliseconds**.

Storm Concepts

- **Topology**: A computation job that is a graph of computation.
- **Tuples**: nodes consume and emit data in the form of tuples.
- **Stream** is an unbounded sequence of tuples.
- **Spout** is a source node that receives data from external sources.
- **Bolt** is a node that processes tuples.
- **Workers** are processes in spouts and bolts with multiple threads for parallel processing.



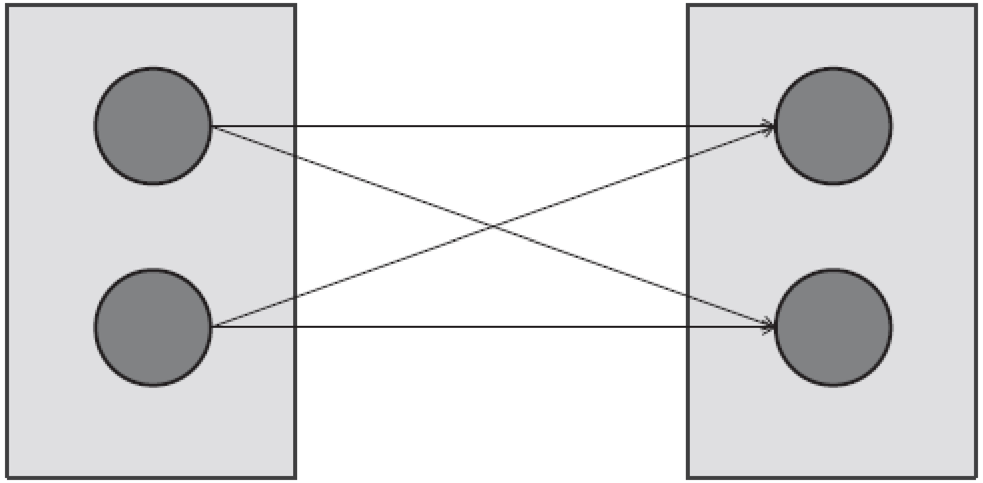
(a)

(b)

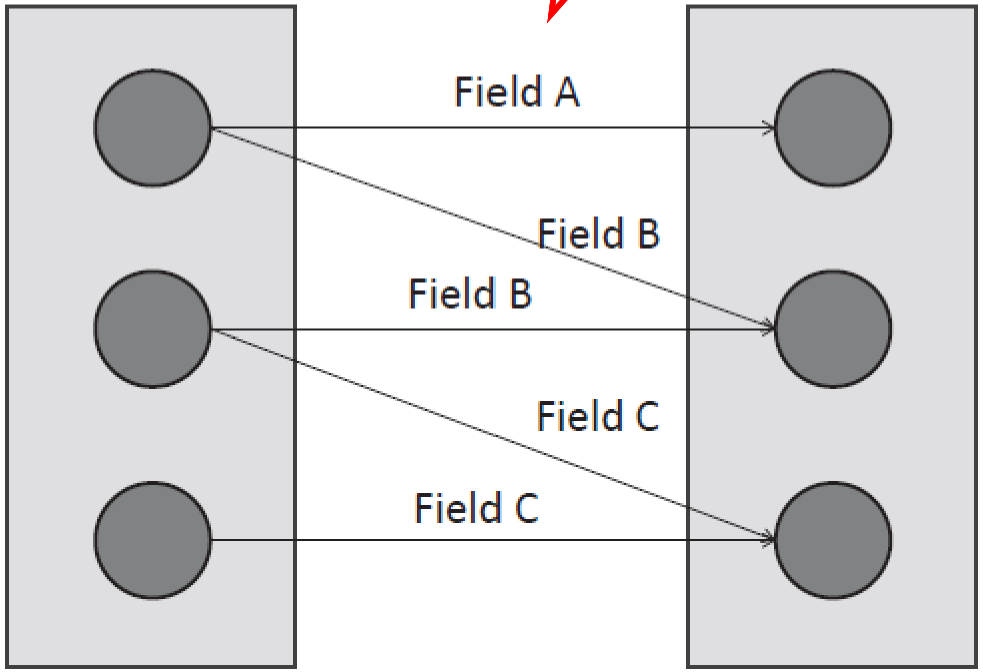
Stream Groupings: how streams are partitioned among the threads

Tuples with the same value of a specified grouping field are always sent to the same task.

1. Shuffle Grouping

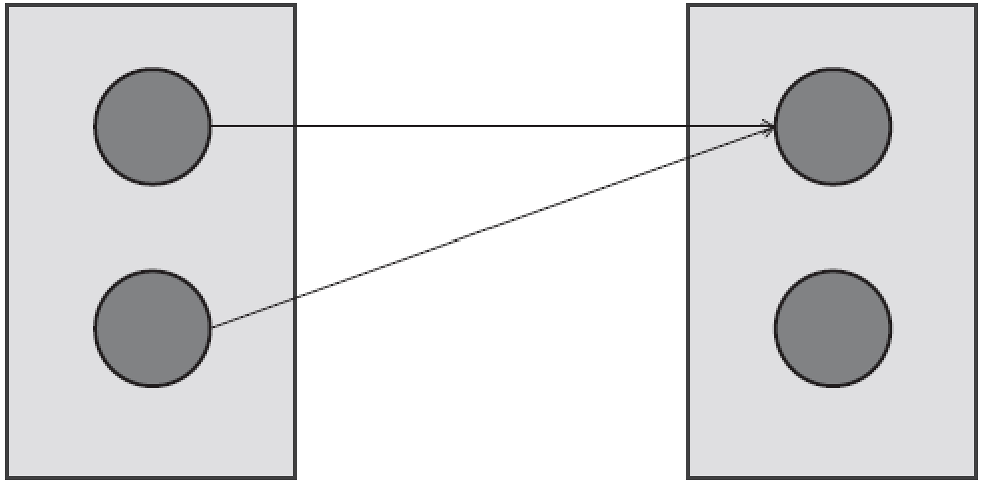


2. Field Grouping

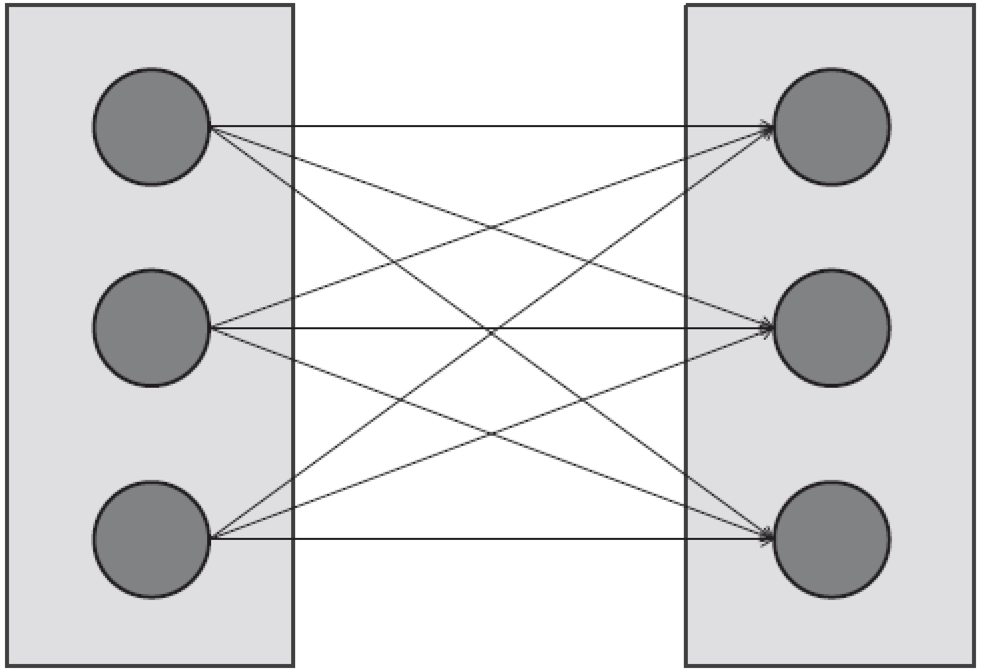


Stream Groupings: how streams are partitioned among the threads

3. Global Grouping

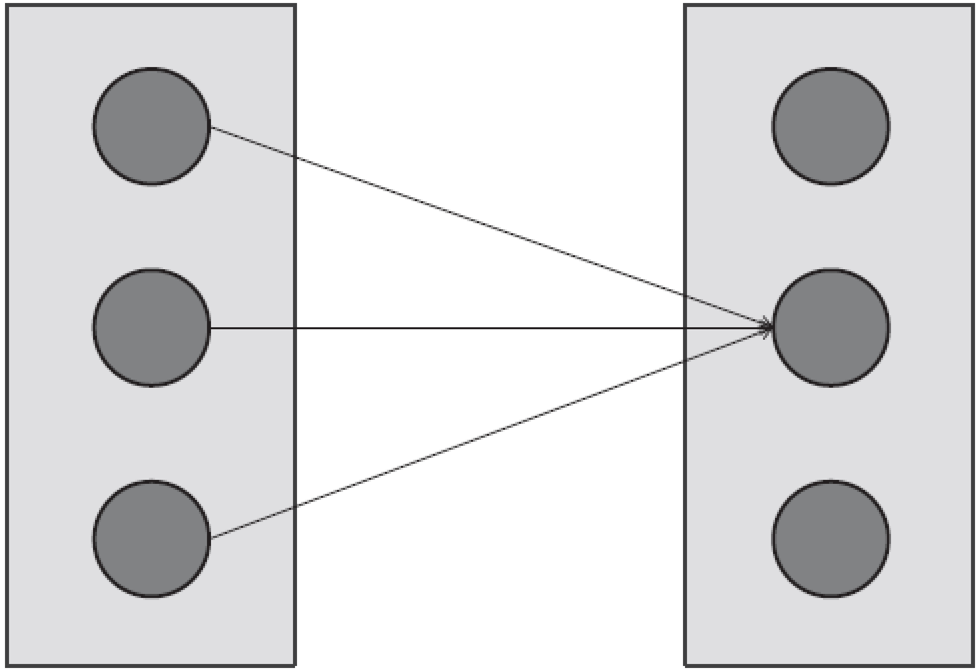


4. All Grouping



Stream Groupings: how streams are partitioned among the threads

5. **Direct Grouping**: the sender node decides which task in the destination bolt should receive the stream.



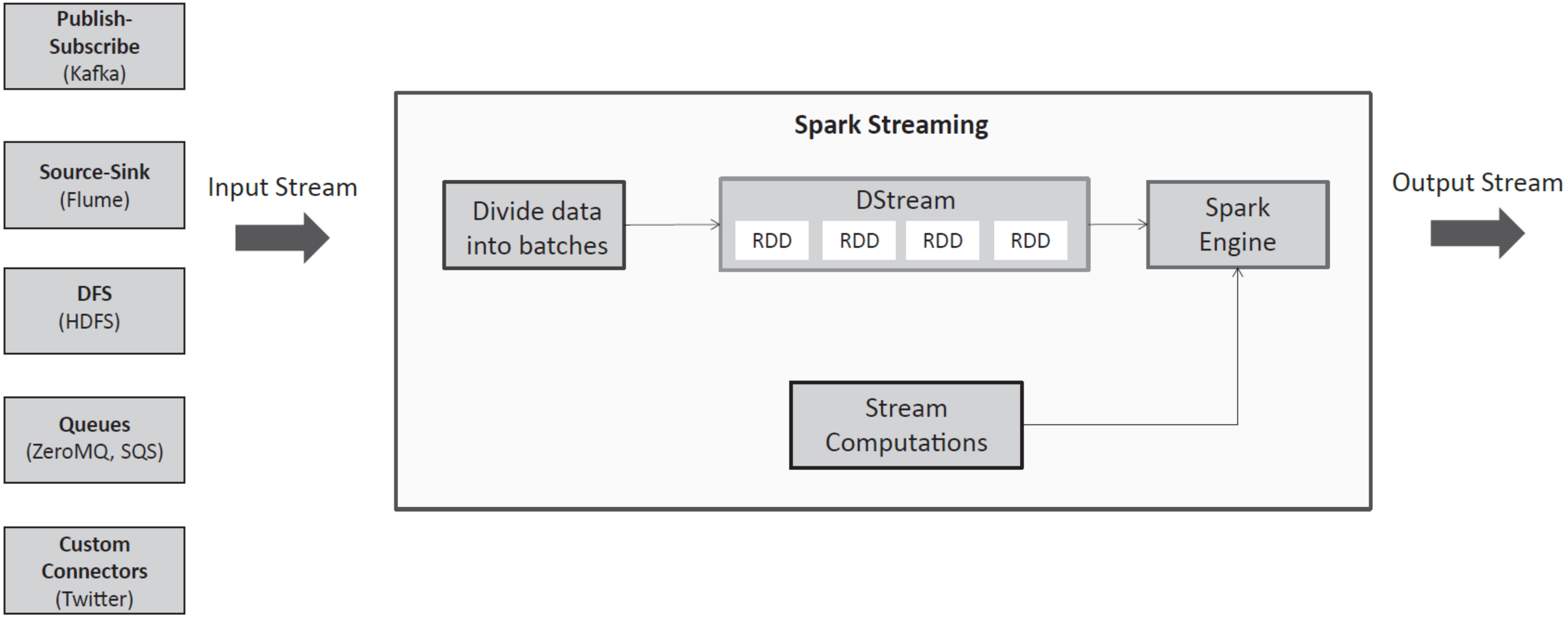
Outline

- Stream Processing with Apache Storm
- In-Memory Processing with Apache Spark

In-Memory Processing with Apache Spark

- Spark streaming enables **scalable**, **high throughput** and **fault-tolerant** stream processing.
- The streaming data is ingested and analyzed in **micro-batches**.
- Spark streaming provides a high-level abstraction called **DStream** (**discretized stream**), which is a sequence of **RDDs**.
- Spark can ingest data from **various types of data sources** into **DStreams**.

Spark Streaming



Publish-Subscribe (Kafka)

Source-Sink (Flume)

DFS (HDFS)

Queues (ZeroMQ, SQS)

Custom Connectors (Twitter)

Input Stream

Spark Streaming

Divide data into batches

DStream

RDD RDD RDD RDD

Spark Engine

Stream Computations

Output Stream

Outline

- Stream Processing with Apache Storm
- In-Memory Processing with Apache Spark