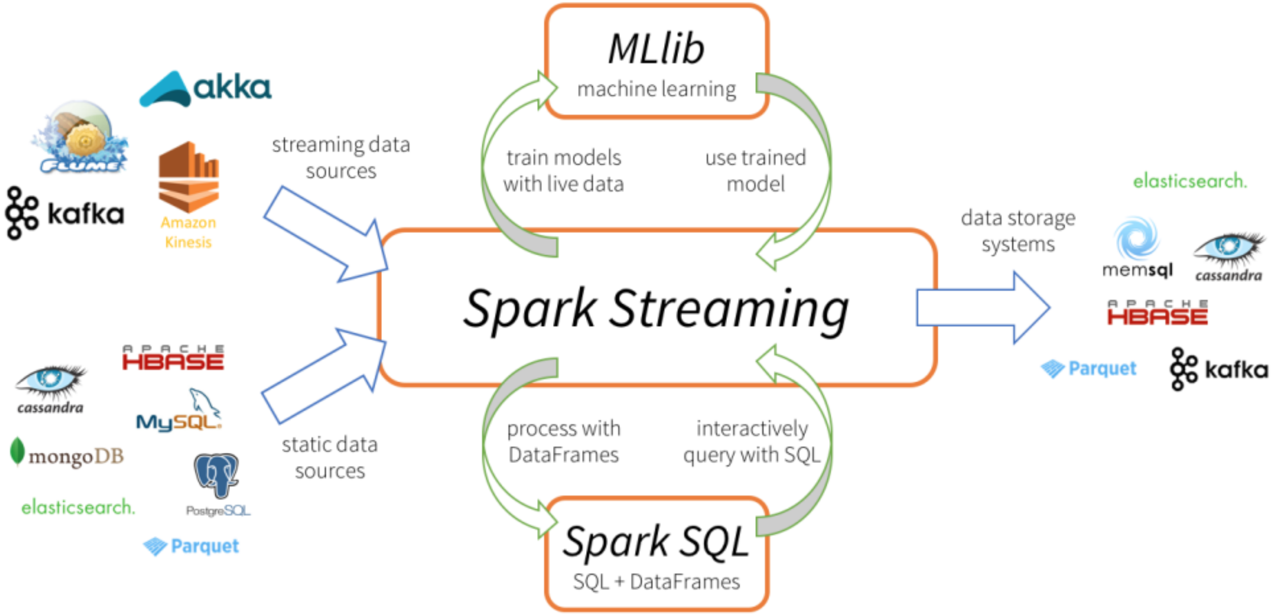
PROCESSING OF BIG DATA

SPARK

SESSION-4



MLlib

machine learning

train models with live data

use trained model

streaming data sources

Spark Streaming

data storage systems

static data sources

process with DataFrames

interactively query with SQL

Spark SQL

SQL + DataFrames

Agenda

- Unstructured Streaming
- Structured Streaming
- Deploying Scala code in production

Unstructured Streaming

- Source
  - Socket: exercise 1
  - File: exercise 2
  - Kafka: exercise 3

- Sink
  - File: exercise 4
  - Database: exercise 5

Unstructured Streaming: Cont..

- ETL on streaming data
    - Transform operation: exercise 6

- Checkpoint for restartability
    - Metadata checkpoint: exercise 7
    - Data checkpoint: self study

- Social media feed processing
    - Twitter feed analysis: self study

Structured Streaming

- Source
  - Socket: exercise 8
  - File: exercise 9
  - Kafka: exercise 11

- Sink
  - File: exercise 10
  - Kafka: exercise 12, exercise 13 (self study)

Structured Streaming: Cont..

- ETL
  - Apply all possible transformation as you do on normal dataframe: exercise 14

- Window aggregation
  - Analyze aggregated data over a period of time:

    exercise 15

Reference

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

- https://spark.apache.org/docs/2.4.0/structured-streaming-kafka-integration.html

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

Thank you!