Building Spark Application Using Scala/SBT
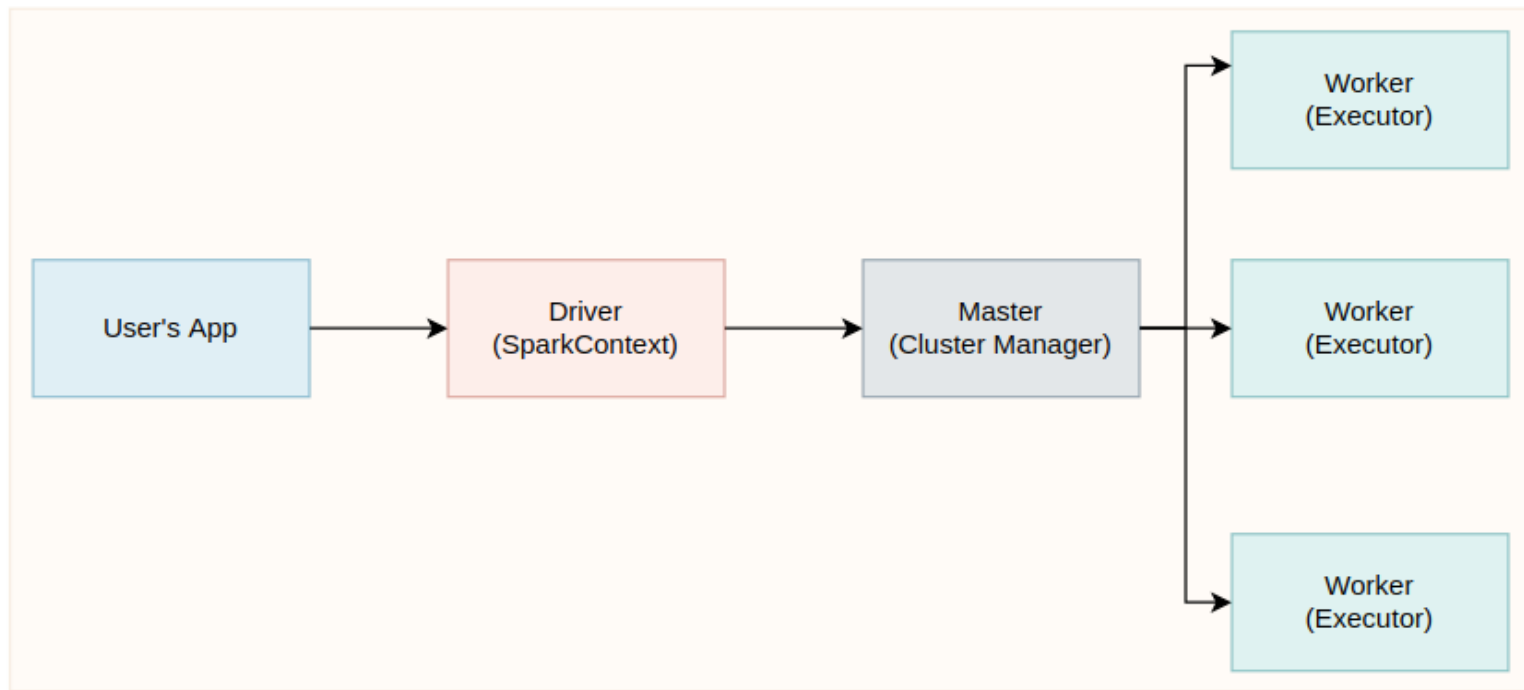
Kodjo Klouvi | kodjo.klouvi@gmail.com
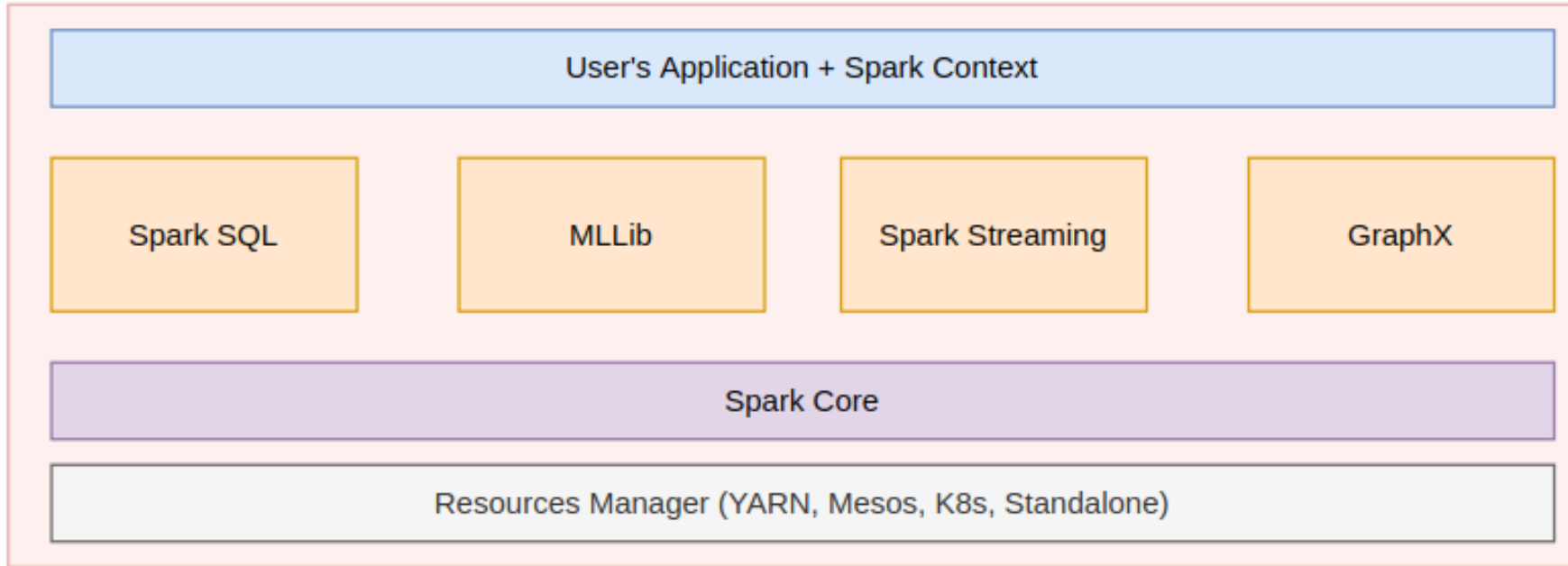
Apache Spark

- Big Data processing Engine (batch and streaming)

- In-memory data pipeline (ETL)

- Parallel execution

- Machine Learning

Spark Architecture

Spark Framework

User's Application + Spark Context			
Spark SQL	MLLib	Spark Streaming	GraphX
Spark Core			
Resources Manager (YARN, Mesos, K8s, Standalone)			

Scala

- Object-oriented programming language

- Coupled with Java (Java bytecode/Gateway)

- Statically typed / Multithreading

- Spark is built in Scala

Hands-On

- https://github.com/osekoo/hands-on-spark-scala
- Preparing dev environment
- Writing code
- Building/testing
- Packaging and submitting tasks

IntelliJ

- Integrated Development Environment

- Multiple modules available for Scala, SBT, Big Data, etc.

- Easy to debug

- Profiler available

Docker

- Platform as a service (PaaS)

- OS abstraction

- Self-contained applications running in containers

- Simplify applications delivery/distribution

- Enable (auto)scaling

Lab Session

- Sync Spark, Hadoop, Scala, Java and SBT versions

- We use here Spark 3.0.2, Hadoop 2.7, Scala 2.12.x, Java 8 and SBT 1.x

- Other details on the Github page

Troubleshooting

- Raise any question during the session or via email
  kodjo.klouvi@gmail.com