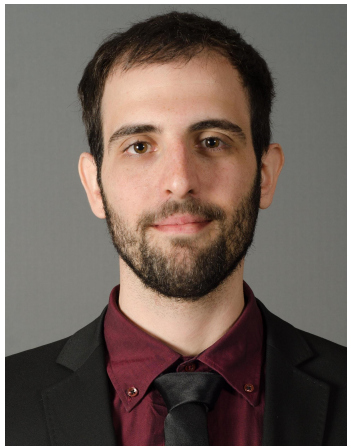
Conformal Prediction in Spark

Tutorial session - COPA 2017

Marco Capuccini
PharmB.io
Uppsala University, Sweden

Who am I?



PhD student – Uppsala University
Department of Information Technology
Department of Pharmaceutical Biosciences

Rome

Uppsala

Background
Computer Science
Bioinformatics

Today's plan

1. Introduction to Apache Spark
2. Demo: CP in Spark using Scala-CP
    a. **GitHub**: https://github.com/mcapuccini/scala-cp
    b. *M. Capuccini, L. Carlsson, U. Norinder and O. Spjuth, "Conformal Prediction in Spark: Large-Scale Machine Learning with Confidence," 2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC), Limassol, 2015, pp. 61-67.*
3. Hands-on/Hackaton
    a. Install TheSparkBox: https://github.com/mcapuccini/TheSparkBox
    b. Reproduce demo: link
    c. Tune the Zeppelin notebook, try some of your use cases

Takeaways: build large-scale CP, large-scale interactive analysis and visualization

Today's plan

1. **Introduction to Apache Spark**
2. Demo: CP in Spark using Scala-CP
   a. **GitHub**: https://github.com/mcapuccini/scala-cp
   b. *M. Capuccini, L. Carlsson, U. Norinder and O. Spjuth, "Conformal Prediction in Spark: Large-Scale Machine Learning with Confidence," 2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC), Limassol, 2015, pp. 61-67.*
3. Hands-on/Hackaton
   a. Install TheSparkBox: https://github.com/mcapuccini/TheSparkBox
   b. Reproduce demo: link
   c. Tune the Zeppelin notebook, try some of your use cases

Takeaways: build large-scale CP, large-scale interactive analysis and visualization

Why Apache Spark?

Apache Spark is the most active open source large-scale data processing engine

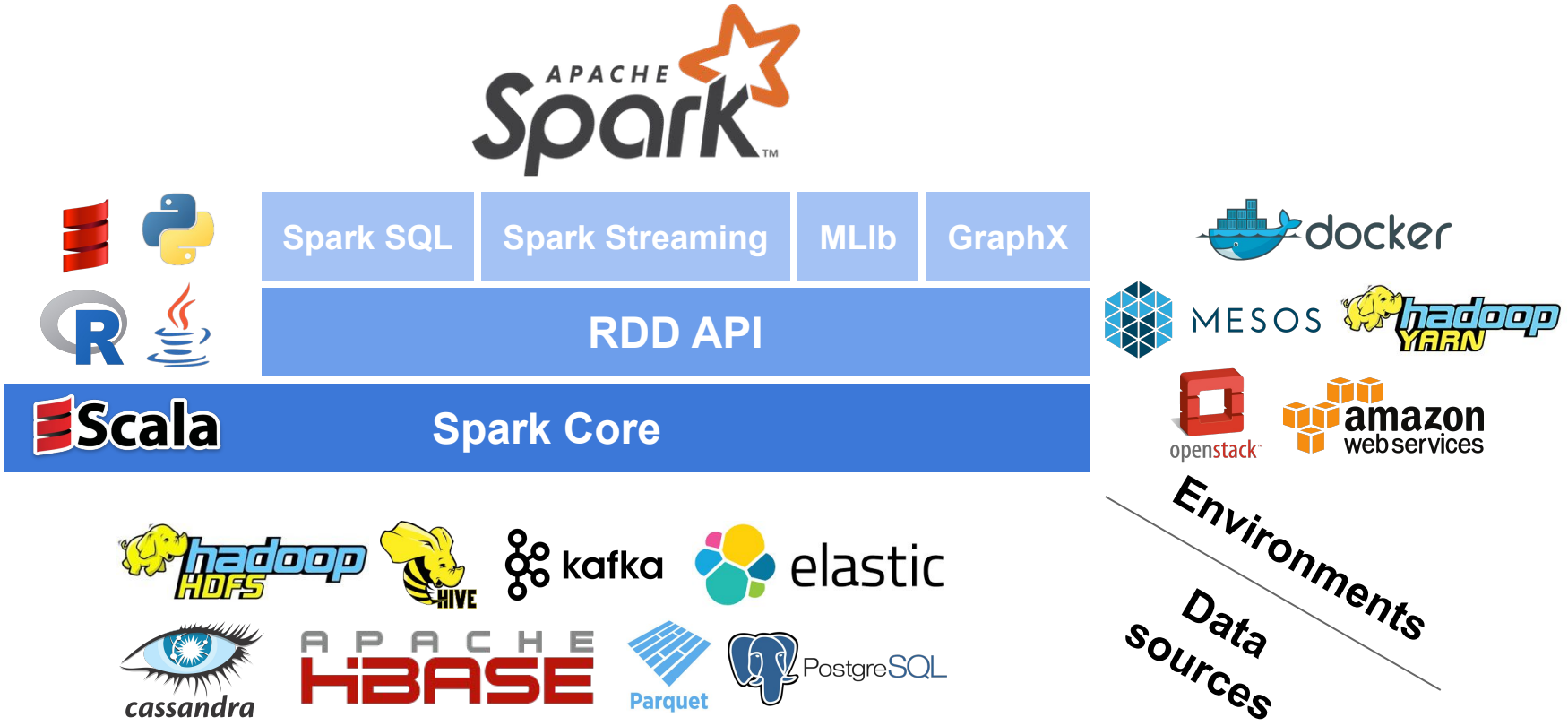
1000+ contributors from over 250 organizations

Originally born to overcome MapReduce **lack of dataset caching**

Spark: Cluster Computing with Working Sets, Zaharia *et al. (2010)*

It allows for **interactive analysis**

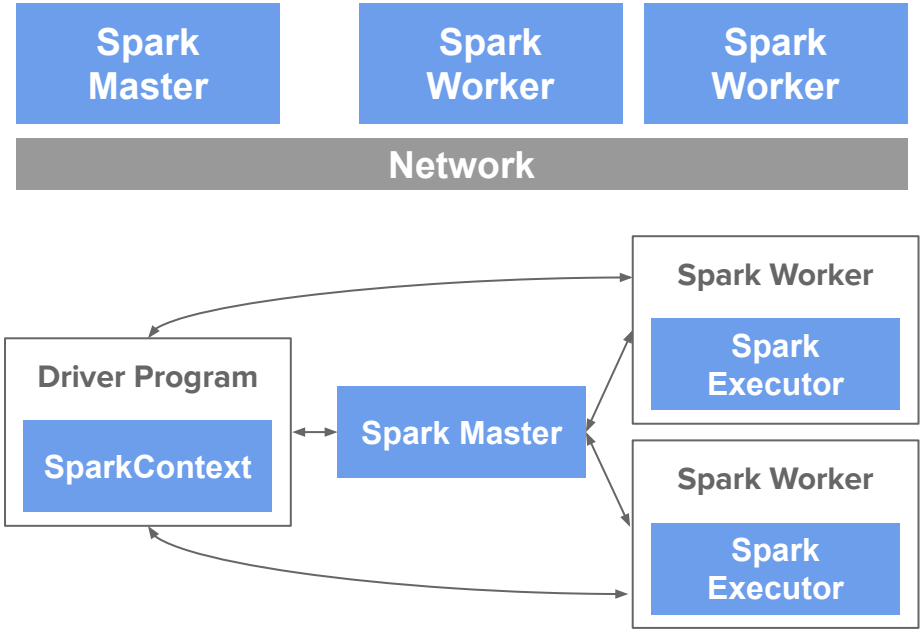
A unified computing engine



Apache Spark architecture (1)

Standalone cluster mode

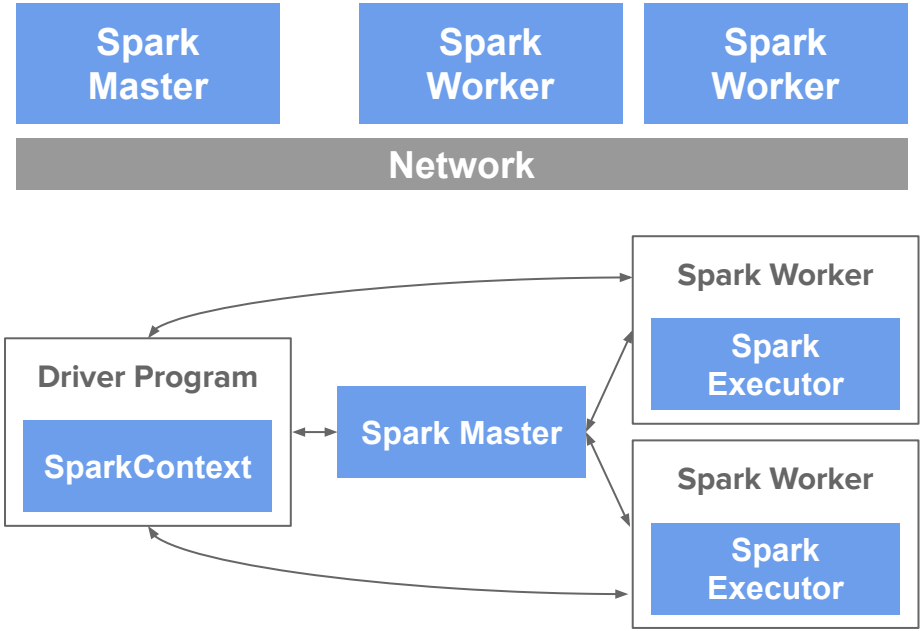
- **Spark Master:** it acts as a *cluster manager*, it maintains the *workers quorum* and it *manages the resources*
- **Spark Worker:** it receive instructions from the *Spark Master*, it launches *SparkExecutors*



Apache Spark architecture (2)

Execution model

- **Driver Program:** it is the program written by the Spark developer. It allocates a **SparkConext**, which is a conduit to access all of the Spark's functionalities
- **Spark Executor:** a container with an allocated amount of *cores* and *memory*. It executes *Tasks* and it stores *Data Partitions*



Today's plan

1. Introduction to Apache Spark
2. **Demo: CP in Spark using Scala-CP**
    a. **GitHub**: https://github.com/mcapuccini/scala-cp
    b. *M. Capuccini, L. Carlsson, U. Norinder and O. Spjuth, "Conformal Prediction in Spark: Large-Scale Machine Learning with Confidence," 2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC), Limassol, 2015, pp. 61-67.*
3. Hands-on/Hackaton
    a. Install TheSparkBox: https://github.com/mcapuccini/TheSparkBox
    b. Reproduce demo: link
    c. Tune the Zeppelin notebook, try some of your use cases

Takeaways: build large-scale CP, large-scale interactive analysis and visualization

Today's plan

1. Introduction to Apache Spark
2. Demo: CP in Spark using Scala-CP
    a. **GitHub**: https://github.com/mcapuccini/scala-cp
    b. *M. Capuccini, L. Carlsson, U. Norinder and O. Spjuth, "Conformal Prediction in Spark: Large-Scale Machine Learning with Confidence," 2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC), Limassol, 2015, pp. 61-67.*
3. **Hands-on/Hackaton**
    a. Install TheSparkBox: https://github.com/mcapuccini/TheSparkBox
    b. Reproduce demo: link
    c. Tune the Zeppelin notebook, try some of your use cases

Takeaways: build large-scale CP, large-scale interactive analysis and visualization

Questions?