Data Engineering for Java Developers

Ricardo Martinelli de Oliveira
Senior Software Engineer - Open Data Hub/RHODS

Agenda

- Data Engineering is not for few people
- Java tools for Data Engineering
- Demos
- AMA (Ask Me Anything)

Disclaimer

- This talk is based on the speakers experience with the related topic
- There might be other solutions that fit better, but remember on the previous bullet
- The main purpose of this talk is to show how a Java Developer can study Data Analysis and Data Engineering without the need of learning a new language
- We don’t want to prove that one language is better than another, but to show some simple concepts for people who already know Java, as there are tons of resources for other languages
- On the other hand, there is no way to truly learn Data Analysis and Data Engineering without proper techniques and tools

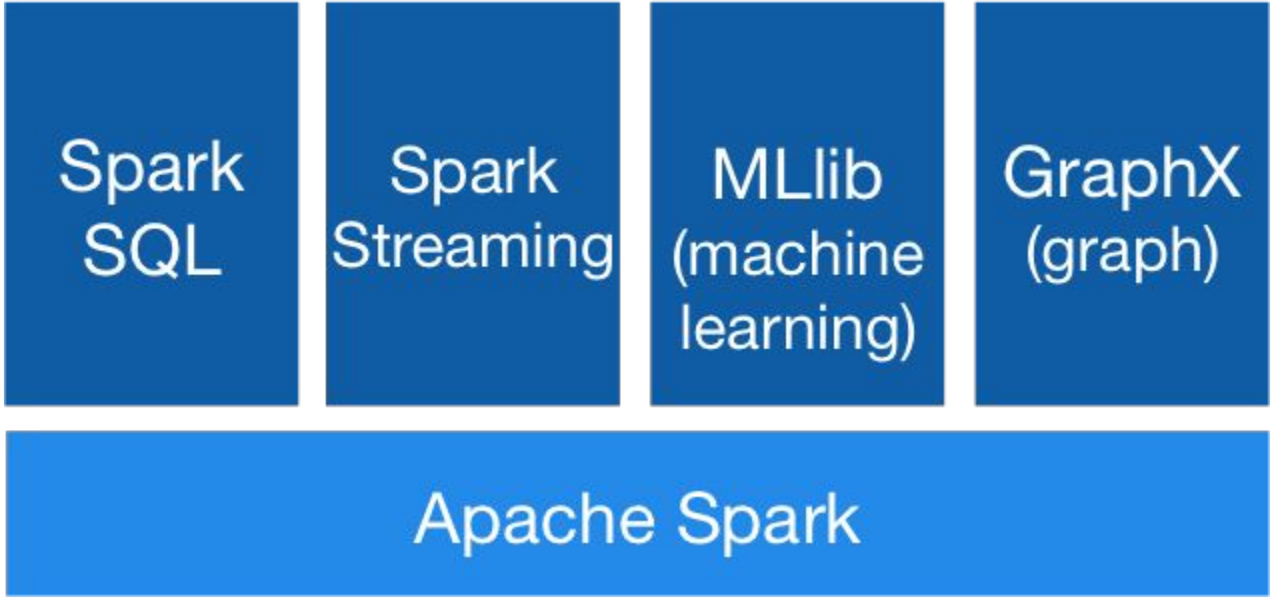
Python <=> Java

- Pandas? Stream API
- Distributed Pandas? Spark
- MongoDB as NoSQL? Cassandra
- Airflow? Nifi or Argo Workflows(Cloud-Native)
- Jupyter? Apache Zeppelin

Stream API

- Processing Data with Java SE 8 Streams, Part 1
- Part 2: Processing Data with Java SE 8 Streams

Spark



Spark SQL

Spark Streaming

MLlib (machine learning)

GraphX (graph)

Apache Spark

Cassandra

- NoSQL
- Replication
- SQL as query language
- And others...

Nifi/Argo

- Nifi
  - Visual Data Pipeline builder
  - Many connectors
- Argo Workflows
  - Build Data pipelines declaratively
  - K8s native

Apache Zeppelin

- No need to code spark connection details
- Lots of “kernels”
- Good for dashboards

Thank you!

Twitter: @rimolive
LinkedIn: https://www.linkedin.com/in/rmartinelli/
GitHub: https://github.com/rimolive/data-engineering-for-java-developers