Spark

Installation:

Recommendation: Use Intellij
https://www.jetbrains.com/idea/download

Follow the steps in this tutorial:
https://medium.com/@Sushil_Kumar/setting-up-spark-with-scala-development-environment-using-intellij-idea-b22644f73ef1

He uses Scala 2.11.11, change it to the latest version of Scala 2.12

Also use the latest dependencies

compile group: 'org.apache.spark', name: 'spark-core_2.12', version: '3.0.0'

compile group: 'org.apache.spark', name: 'spark-sql_2.12', version: '3.0.0'

And for streaming

providedCompile group: 'org.apache.spark', name: 'spark-streaming_2.12', version: '3.0.0'

Deployment

You can either create your custom environment with Spark 3.0 and Hadoop, or use the provided docker-compose.yml file to setup the system.

If you use the Docker compose file:

With Hue you can access the HDFS.

Interface:

- Namenode: http://localhost:50070
- Datanode: http://localhost:50075
- Spark-master: http://localhost:8080
- Spark-notebook: http://localhost:9001
- Hue (HDFS Filebrowser): http://localhost:8088/home