Lab 2 - Spark and Spark SQL

Amir H. Payberah
payberah@kth.se

1 Introduction

In this lab you will practice the basic operations of Spark (RDDs) and Spark SQL (DataFrames). We have used the Jupyter Notebooks for this lab assignment. Notebooks are documents that contain both the programming code, as well as human-readable text elements. Below, we first explain how to install Spark and test it, and then we go through the steps to install Jupyter Notebook on a Linux machine. Then, we show how to use this environment to do your assignment.

2 Installing Spark

This section presents the steps you need to do to install Spark.

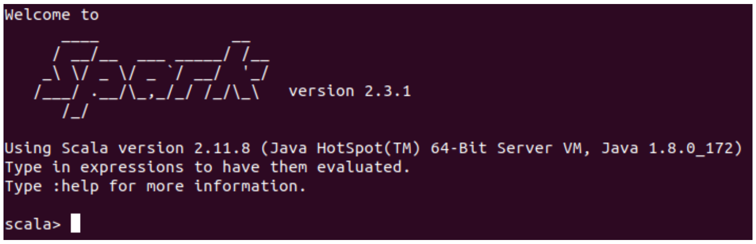
1. Download and install Java SDK 8. You can download it from the following link:
   http://www.oracle.com/technetwork/pt/java/javase/downloads/jdk8-downloads-2133151.html

2. Download Apache Spark 2.3.1 from the following link:
   https://www.apache.org/dyn/closer.lua/spark/spark-2.3.1/spark-2.3.1-bin-hadoop2.7.tgz

3. Set the following environment variables.

```
export JAVA_HOME="/path/to/the/java/folder"
export SPARK_HOME="/path/to/the/spark/folder"
export PATH=$JAVA_HOME/bin:$SPARK_HOME/bin:$PATH
```

4. Run the command `spark-shell` in a terminal. If it works, you should see:



```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.3.1
      /_/

Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_172)
Type in expressions to have them evaluated.
Type :help for more information.

scala> █
```

5. Now, we want to write a self-contained word count application using the Spark API in Scala (with SBT). This code is available in the zip file under the folder `src/hellospark`.

```
import org.apache.spark.sql.SparkSession

object HelloSpark {
  def main(args: Array[String]) {
    val logFile = "/path/to/a/text/file"
    val spark = SparkSession.builder.appName("Hello Spark").master("local[2]").getOrCreate()
    val sc = spark.sparkContext
    val logData = sc.textFile(logFile).cache()
    val wordCounts = logData.flatMap(line => line.split(" "))
                            .map(word => (word, 1))
                            .reduceByKey((a, b) => a + b)
    wordCounts.foreach(println(_))
    spark.stop()
  }
}
```

6. We also need to include a SBT configuration file, `build.sbt`, which explains that Spark is a dependency.

```
name := "Simple Project"

version := "1.0"

scalaVersion := "2.11.8"

libraryDependencies += "org.apache.spark" %% "spark-sql" % "2.3.1"
```

7. To compile and run the code you should run `sbt run` command. If you do not have SBT on your machine, you can install it as shown below.

```
echo "deb https://dl.bintray.com/sbt/debian /" | sudo tee -a /etc/apt/sources.list.d/sbt.list
sudo apt-key adv --keyserver hkp://keyserver.ubuntu.com:80 --recv 2EE0EA64E40A89B84B2DF73499E82A75642AC823
sudo apt-get update
sudo apt-get install sbt
```

3 Installing Jupyter Notebook

Here we present how to install Jupyter Notebook.

1. Download and install Anaconda (Python 2.7). You can download it from the following link: https://repo.anaconda.com/archive/Anaconda2-5.2.0-Linux-x86_64.sh
2. Set the following environment variables.

```
export PYTHONPATH="/path/to/the/python/folder"
export PATH=$PYTHONPATH/bin:$PATH
```

3. Install the Jupyter Notebook.

```
pip install notebook
```

4. Now, we need to install Apache Toree and load it into Jupyter. Apache Toree is a kernel for the Jupyter Notebook platform providing interactively access to Spark.

```
pip install --upgrade toree
jupyter toree install --spark_home=$SPARK_HOME --kernel_name="Spark" --spark_opts="--master=local[*]"
```

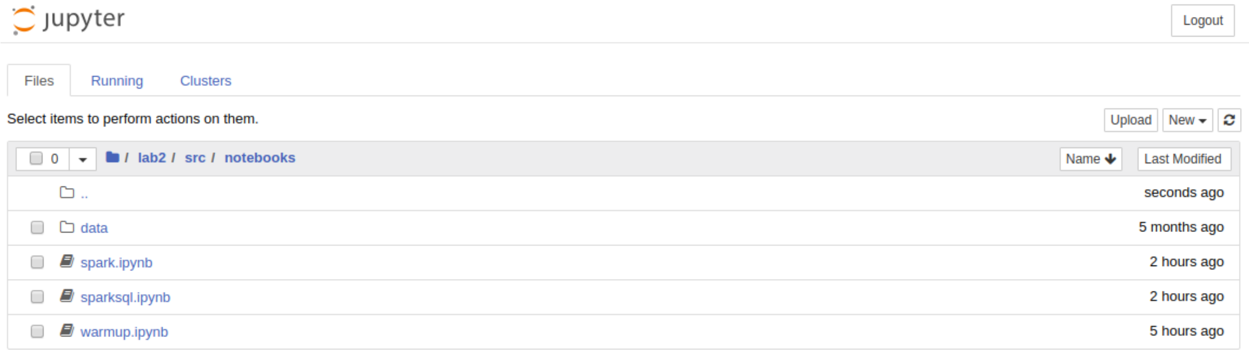
5. We can get the Notebook server running now.

```
jupyter notebook
```

6. Once you run the Jupyter Notebook, you can see it on your browser on the address `localhost:8888`.

4 You Assignment

Unzip the given zip file `lab2.zip`, and copy the Notebooks and the `data` folder from `src/notebook` to the folder you have started the Jupyter Notebook. Then, you should be able to see the files in Jupyter on your browser on the address `localhost:8888` (as shown below). There are three Jupyter Notebooks called `warmup.ipynb`, `spark.ipynb`, and `sparksql.ipynb`, in which the first one (`warmup.ipynb`) is just for practice, and the next two Notebooks are the ones you need to complete. The files are self-explanatory that describe what you need to do.



Select items to perform actions on them.

Name	Last Modified
..	seconds ago
data	5 months ago
spark.ipynb	2 hours ago
sparksql.ipynb	2 hours ago
warmup.ipynb	5 hours ago

What to deliver: you should complete the two Notebooks `spark.ipynb` and `sparksql.ipynb` and zip them in a single file. Please use the filename format `lab2_groupname.zip`.