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 will use the Jupyter Notebooks for this 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.

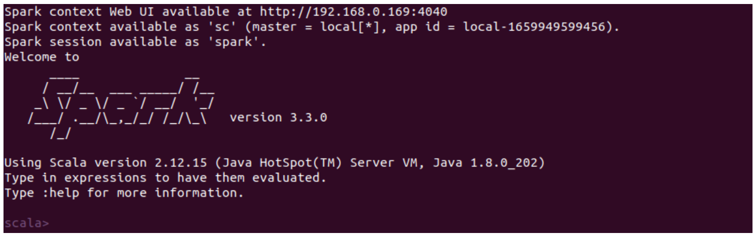
1. Download and install Java SDK 8. You can download it from the following link:
   http://www.oracle.com/technetwork/pt/java/javase/downloads/jdk8-downloads-2133151.html

2. Download Apache Spark 3.3.0 from the following link:
   https://www.apache.org/dyn/closer.lua/spark/spark-3.3.0/spark-3.3.0-bin-hadoop3.tgz

3. Set the following environment variables.

```
export JAVA_HOME=<path to the Java folder>
export SPARK_HOME=<path to the Spark folder>
```

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



```
Spark context Web UI available at http://192.168.0.169:4040
Spark context available as 'sc' (master = local[*], app id = local-1659949599456).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.3.0
      /_/

Using Scala version 2.12.15 (Java HotSpot(TM) Server VM, Java 1.8.0_202)
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/hellosession`.

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

object HelloSpark {
  def main(args: Array[String]) {
    val logFile = "data/story/hamlet.txt"
    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 := "2.12.15"

libraryDependencies += "org.apache.spark" %% "spark-sql" % "3.3.0"
```

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://repo.scala-sbt.org/scalasbt/debian all main" | sudo tee /etc/apt/sources.list.d/sbt.list
echo "deb https://repo.scala-sbt.org/scalasbt/debian /" | sudo tee /etc/apt/sources.list.d/sbt_old.list
curl -sL "https://keyserver.ubuntu.com/pks/lookup?op=get&search=0x2EE0EA64E40A89B84B2DF73499E82A75642AC823"
     | sudo apt-key add
sudo apt-get update
sudo apt-get install sbt
```

3 Installing Jupyter Notebook and Apache Toree

Here we present how to install Jupyter Notebook.

1. Download and install Anaconda. You can download it from the following link: https://www.anaconda.com/products/distribution
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
```

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

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`. There are four notebooks, which are self-explanatory that describe what you need to do.