Apache Spark installation

PySpark

Step 1: Install PySpark

- pip install pyspark==2.3.2
- pip install findspark

Step 2: Set Environment variables

- set SPARK_HOME={path to pyspark directory}
    - if using Anaconda : “C:\Users\NAMRA\Anaconda3\Lib\site-packages\pyspark”
- set HADOOP_HOME={path to hadoop directory}
    - from previous hadoop installation guide: “C:\hadoop”
- set PATH=%SPARK_HOME%\bin;%PATH%
- set PATH=%HADOOP_HOME%\bin;%PATH%
- set PYTHONPATH= {Path to python installation directory}
    - if using Anaconda : “C:\Users\NAMRA\Anaconda3”

Step 3: Run example WordCount.py

- `spark-submit WordCount.py`
    - See WordCount.py file opens “word_count.text” from local directory.
    - You can also open file directly from Hadoop Cluster (HDFS).
- When program is running, we can see details of various jobs at `localhost:4040`

Install Spark to run with Java

Step 1: Download necessary files

- Eclipse : https://www.eclipse.org/downloads/
- Apache Spark : http://spark.apache.org/downloads.html (2.3.3)
- Apache Maven : https://maven.apache.org/download.cgi

Step 2: Put spark and maven in C:\spark and C:\Maven

Step 3: Set Environment variables

- set SPARK_HOME={path to pyspark directory}
  - if using Anaconda : "C:\Users\NAMRA\Anaconda3\Lib\site-packages\pyspark"
- set HADOOP_HOME={path to hadoop directory}
  - from previous hadoop installation guide: "C:\hadoop"
- set PATH=%SPARK_HOME%\bin;%PATH%
- set PATH=%HADOOP_HOME%\bin;%PATH%
- set PYTHONPATH= {Path to python installation directory}
  - if using Anaconda : "C:\Users\NAMRA\Anaconda3"
- mkdir C:\tmp\hive
- cd c:\hadoop\bin
- winutils.exe chmod -R 777 C:\tmp\hive

Step 4: Open Eclipse and Create new Maven Project

Step 5: Modify the file pom.xml and add inside the « <dependencies>…</dependencies> »,

add the following:

```
<dependency>
    <groupId>org.apache.spark</groupId>
    <artifactId>spark-core_2.11</artifactId>
    <version>2.1.0</version>
</dependency>
<dependency>
<groupId>org.scala-lang</groupId>
    <artifactId>scala-library</artifactId>
    <version>2.11.8</version>
</dependency>
<dependency>
    <groupId>org.scala-lang</groupId>
    <artifactId>scala-xml_2.11.0-M4</artifactId>
    <version>1.0-RC1</version>
</dependency>
```

Step 6: Create/Modify your Java File

```
package com.mycompany.app;

import org.apache.spark.SparkConf;

import org.apache.spark.api.java.JavaRDD;

import org.apache.spark.api.java.JavaSparkContext;

public class App

{

    public static void main( String[] args )

    {

        System.out.println( "Hello World!" );

        SparkConf conf = new
SparkConf().setAppName("firstSparkProject").setMaster("local[*]");

        JavaSparkContext sc = new JavaSparkContext(conf);

        String path = "linescount.txt";

        System.out.println("Trying to open: " + path);

        JavaRDD<String> lines = sc.textFile(path.toString());

        System.out.println("Lines count: " + lines.count());

        sc.stop();

    }

}
```

Step 7: Compile Project into .JAR

Step 8: spark-submit --class {YOUR_CLASS_NAME} {PATH_TO JAR} {INPUT_ARGUMENTS}

- If You've Written code for HDFS it will be as follows:

  spark-submit –class {class_name} {input_files} {output_directory}

Step 9: Try running TFIDF.jar file from Previous LAB.