Basic Spark-Shell

|  |  |
| --- | --- |
| **Objective** | Perform basic spark-shell operations |
| **Location of Files** | ~/materials/data |
| **Expected Outcome** | Compare its performance with a traditional Hadoop MapReduce program. |
| **Before You Begin** | Open New Terminal Window |

In this exercise, we will upload a relatively big data set (NASDAQ Exchange Daily 1970-2010) to Hadoop file system (HDFS). We will compute **“Number of occurrence for the stock-ticker ‘AMG’ in the dataset”** using Spark.

**Step 1:** Copy NASDAQ data to HDFS.

$ hdfs dfs –mkdir nasdaq

$ hdfs dfs –put NYSE\* nasdaq

$ hdfs dfs –ls nasdaq

**Step 2:** Count number of occurrence for the stock-ticker ‘AMG’ using **SPARK**.

1**.** Start Spark shell.

$ spark-shell

NOTE**:** Spark context is available as ‘sc’.

2**.** To check what different spark-commands are available, type following and **press ’TAB’ button**:

scala> sc

3. Execute following on **scala>** prompt:

scala> val file = sc.textFile("hdfs://sandbox:8020/user/ root/nasdaq")

scala> val counts = file.flatMap(line => line.split(",")).map(word => (word, 1)).reduceByKey(\_ + \_)

scala> val amg = file.filter(line => line.contains("AMG"))

scala> amg.count()

It takes few seconds to compute and finally it will give count as **3065**.

4**.** Exit from **scala>** prompt:

scala> exit

**END**