

# Spark Examples

---

## 1. Basic WordCount:

```
val textFile = sc.textFile("hdfs://<Path>")
val counts = textFile.flatMap(line => line.split(" ")).map(word => (word, 1)).reduceByKey(_ + _)
counts.saveAsTextFile("sparkout")
```

## 1b. Find line having largest count of words:

```
scala> import java.lang.Math
import java.lang.Math

scala> textFile.map(line => line.split(" ").size).reduce((a, b) => Math.max(a, b))
res5: Int = 15
```

## 2. Running from command file:

Create a file called wc.scala containing the commands below:

```
val f = sc.textFile("file:///usr/local/spark-1.6.1-bin-hadoop2.4/README.md")
val words = f.flatMap(l => l.split(" ")).map(word => (word, 1))
words.reduceByKey(_ + _).collect.foreach(println)
System.exit(0)
```

Run it as:

```
spark-shell -i wc.scala
```

Other examples available at:

<http://spark.apache.org/examples.html>

## 3. Other examples are available:

```
cd $SPARK_HOME
```

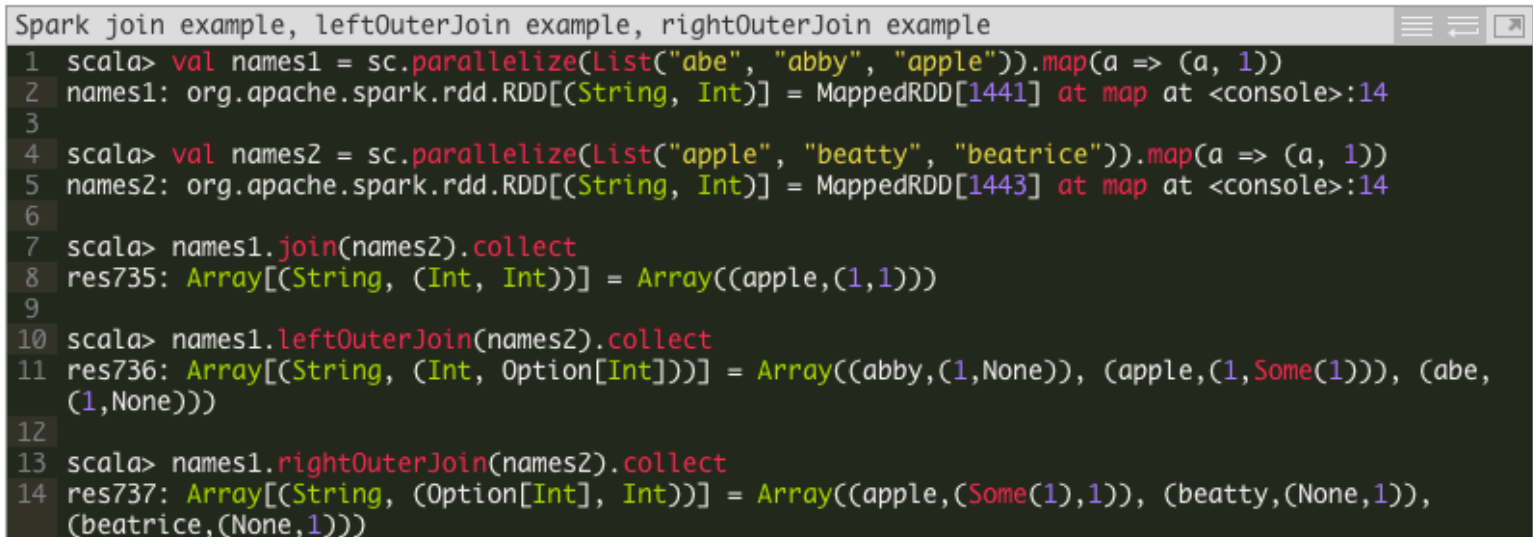
Look at examples here `examples/src/main/scala/org/apache/spark/examples/`

How to run:

To run SparkPi code, run the following in the \$SPARK\_HOME directory:

```
./bin/spark-submit --class org.apache.spark.examples.SparkPi --master yarn lib/spark-examples-1.4.1-hadoop2.4.0.jar 100
```

#### 4. Spark Join examples

A screenshot of a terminal window with a dark background and light-colored text. The window title is "Spark join example, leftOuterJoin example, rightOuterJoin example". It shows a series of Scala commands and their outputs. Line 1: "scala> val names1 = sc.parallelize(List("abe", "abby", "apple")).map(a => (a, 1))". Line 2: "names1: org.apache.spark.rdd.RDD[(String, Int)] = MappedRDD[1441] at map at <console>:14". Line 3: A blank line. Line 4: "scala> val names2 = sc.parallelize(List("apple", "beatty", "beatrice")).map(a => (a, 1))". Line 5: "names2: org.apache.spark.rdd.RDD[(String, Int)] = MappedRDD[1443] at map at <console>:14". Line 6: A blank line. Line 7: "scala> names1.join(names2).collect". Line 8: "res735: Array[(String, (Int, Int))] = Array((apple,(1,1)))". Line 9: A blank line. Line 10: "scala> names1.leftOuterJoin(names2).collect". Line 11: "res736: Array[(String, (Int, Option[Int]))] = Array((abby,(1,None)), (apple,(1,Some(1))), (abe,(1,None)))". Line 12: A blank line. Line 13: "scala> names1.rightOuterJoin(names2).collect". Line 14: "res737: Array[(String, (Option[Int], Int))] = Array((apple,(Some(1),1)), (beatty,(None,1)), (beatrice,(None,1)))".

```
Spark join example, leftOuterJoin example, rightOuterJoin example
1  scala> val names1 = sc.parallelize(List("abe", "abby", "apple")).map(a => (a, 1))
2  names1: org.apache.spark.rdd.RDD[(String, Int)] = MappedRDD[1441] at map at <console>:14
3
4  scala> val names2 = sc.parallelize(List("apple", "beatty", "beatrice")).map(a => (a, 1))
5  names2: org.apache.spark.rdd.RDD[(String, Int)] = MappedRDD[1443] at map at <console>:14
6
7  scala> names1.join(names2).collect
8  res735: Array[(String, (Int, Int))] = Array((apple,(1,1)))
9
10 scala> names1.leftOuterJoin(names2).collect
11 res736: Array[(String, (Int, Option[Int]))] = Array((abby,(1,None)), (apple,(1,Some(1))), (abe,
    (1,None)))
12
13 scala> names1.rightOuterJoin(names2).collect
14 res737: Array[(String, (Option[Int], Int))] = Array((apple,(Some(1),1)), (beatty,(None,1)),
    (beatrice,(None,1)))
```

#### 5. To run compiled jar file

```
spark-submit --class com.examples.MainExample \
  --master yarn-cluster \
  spark-scala-maven-project-0.0.1-SNAPSHOT-with-dependencies.jar \
  inputhdfspath \
  outputhdfspath
```