databricks Quick Start Using Scala

(http://databricks.com)

Import Notebook

Quick Start Using Scala

- Using a Databricks notebook to showcase RDD operations using Scala
- Reference http://spark.apache.org/docs/latest/quick-start.html (http://spark.apache.org/docs/latest/quick-start.html)

```
> // Take a look at the file system
  display(dbutils.fs.ls("/mnt/tardis6/docs"))
```

path	name
dbfs:/mnt/tardis6/docs/README.md	READM

```
> // Setup the textFile RDD to read the README.md file
  //   Note this is lazy
  val textFile = sc.textFile("/mnt/tardis6/docs/README.md")
```

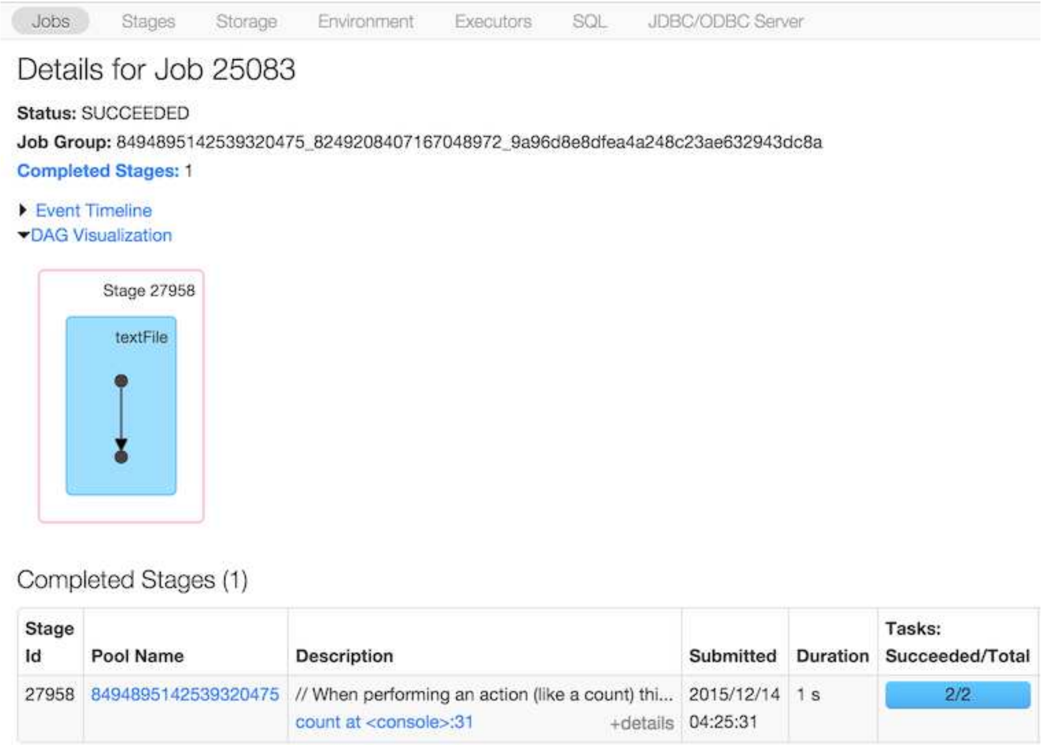
```
textFile: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[103321] at textFile at <console>:27
```

RDDs have ***actions***, which return values, and ***transformations***, which return pointers to new RDDs.

```
> // When performing an action (like a count) this is when the textFile is read and aggregate calculated
  //    Click on [View] to see the stages and executors
  textFile.count()
```

```
res3: Long = 82
```

Scala Count (Jobs)



Details for Job 25083

Status: SUCCEEDED

Job Group: 8494895142539320475_8249208407167048972_9a96d8e8dfea4a248c23ae632943dc8a

Completed Stages: 1

Completed Stages (1)

Stage Id	Pool Name	Description	Submitted	Duration	Tasks: Succeeded/Total
27958	8494895142539320475	// When performing an action (like a count) thi... count at <console>:31 +details	2015/12/14 04:25:31	1 s	2/2

Scala Count (Stages)

- Notice how the file is read during the *.count()* action
- Many Spark operations are lazy and executed upon some action