Spark Homework Logs Lizi Chen - lc3397@nyu.edu

alldata.count()

182 Files

```
scala> val alldata = sc.textFile("/user/lc3397/weblogs")
alldata: org.apache.spark.rdd.RDD[String] = /user/lc3397/weblogs MapPartitionsRDD[3] at
 textFile at <console>:27

scala> alldata.count()
res3: Long = 1079891

[Stage 0:=============================================>     (164 + 18) / 182]

scala> alldata.count()
res4: Long = 1079891
```

alldata.take(4).foreach(println)

```
scala> alldata.take(4)
res5: Array[String] = Array(3.94.78.5 - 69827 [15/Sep/2013:23:58:36 +0100] "GET /KBDOC-00033.html HTTP/1.0" 200 14417 "http://www.loudacre.com"  "Loudacre Mobile Browser iFruit
 1", 3.94.78.5 - 69827 [15/Sep/2013:23:58:36 +0100] "GET /theme.css HTTP/1.0" 200 3576 "http://www.loudacre.com"  "Loudacre Mobile Browser iFruit 1", 19.38.140.62 - 21475 [15/S
ep/2013:23:58:34 +0100] "GET /KBDOC-00277.html HTTP/1.0" 200 15517 "http://www.loudacre.com"  "Loudacre Mobile Browser Ronin S1", 19.38.140.62 - 21475 [15/Sep/2013:23:58:34 +01
00] "GET /theme.css HTTP/1.0" 200 13353 "http://www.loudacre.com"  "Loudacre Mobile Browser Ronin S1")

scala> alldata.take(4).foreach(println)
3.94.78.5 - 69827 [15/Sep/2013:23:58:36 +0100] "GET /KBDOC-00033.html HTTP/1.0" 200 14417 "http://www.loudacre.com"  "Loudacre Mobile Browser iFruit 1"
3.94.78.5 - 69827 [15/Sep/2013:23:58:36 +0100] "GET /theme.css HTTP/1.0" 200 3576 "http://www.loudacre.com"  "Loudacre Mobile Browser iFruit 1"
19.38.140.62 - 21475 [15/Sep/2013:23:58:34 +0100] "GET /KBDOC-00277.html HTTP/1.0" 200 15517 "http://www.loudacre.com"  "Loudacre Mobile Browser Ronin S1"
19.38.140.62 - 21475 [15/Sep/2013:23:58:34 +0100] "GET /theme.css HTTP/1.0" 200 13353 "http://www.loudacre.com"  "Loudacre Mobile Browser Ronin S1"
```

val oneLog = sc.textFile("/user/lc3397/weblogs/2014-03-15.log")
oneLog.count()
=> 7097 Lines

```
scala>  val oneLog = sc.textFile("/user/lc3397/weblogs/2014-03-15.log")
oneLog: org.apache.spark.rdd.RDD[String] = /user/lc3397/weblogs/2014-03-15.log MapParti
tionsRDD[5] at textFile at <console>:27

scala> oneLog.count()
res7: Long = 7097
```

oneLog.take(1).foreach(println)

```
scala> oneLog.take(1).foreach(println)
234.206.18.239 - 8495 [15/Mar/2014:23:59:30 +0100] "GET /KBDOC-00082.html HTTP/1.0" 200
 9054 "http://www.loudacre.com"  "Loudacre Mobile Browser Titanic 2200"
```

var jpglines = alldata.filter(line => line.contains(".jpg"))
jpglines.count()
=> 64978

```
scala> var jpglines = alldata.filter(line => line.contains(".jpg"))
jpglines: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[6] at filter at <console>
:29

scala> jpglines.count()
res10: Long = 64978
```

Running on Hive: 17.22 seconds

```
INFO  : Completed executing command(queryId=hive_20170510201414_4bdf3008-60f7-4710-89f0
-2449b851cc01); Time taken: 16.897 seconds
INFO  : OK
+----------+--+
|   _c0    |
+----------+--+
| 1079891  |
+----------+--+
1 row selected (17.22 seconds)
0: jdbc:hive2://babar.es.its.nyu.edu:10000/>
```