Big Data Application Development - Fall 2017

Homework 3, Part 2 Answer Sheet

4. Use the REPL to explore Spark RDDs.

1) Provide the command you used to create your RDD.	val mydata = sc.textFile("file:/home/cloudera/hw3/frostroad.txt")
2) Provide the command you used to count the elements (lines) in your RDD.	mydata.count()
3) Provide the number of elements.	23
4) Provide the collect command you used.	mydata.collect()
5) Provide the command you used to create the HDFS directory.	hdfs dfs -mkdir bdad17/loudacre/ hdfs dfs -mkdir bdad17/loudacre/weblog
6) Provide the command you used to put the file into HDFS.	hdfs dfs -put 2014-03-15.log bdad17/loudacre/weblog/
7) Provide the command you used to view the file.	hdfs dfs -text bdad17/loudacre/weblog/2014-03-15.log

5. Transform a small dataset using RDDs.

8) Initialize **`logfile`**.	val logfile = "bdad17/loudacre/weblog/2014-03-15.log"
9) Create an RDD from the file.	var mydata = sc.textFile(logfile)
10) View the first 10 lines of the data.	mydata.take(10)
11) Create an RDD containing only lines that are requests for **`jpg`** files.	val mydata_filter = mydata.filter(line => line.contains(".jpg"))
12) View the first 10 lines of the data.	mydata_filter.take(10)
13) Chain the previous commands into a single command that counts the number of JPG requests.	mydata.filter(line => line.contains(".jpg")).count() res14: Long = 423
14) Create an RDD using the **`map`** function to return the length of each line of the log file.	val mydata_map1= mydata.map(line => line.length())
15) Create an RDD using the **`map`** and **`split`** functions to map an array of words for each line.	val mydata_map2= mydata.map(line => line.split(" "))
16) Create an RDD containing only the IP addresses from each line.	val mydata_map3= mydata.map(line => line.split(" ")(0))
17) Use **`foreach(println)`** to output IP addresses.	mydata_map3.foreach(println)
18) Save the list of IP addresses to an HDFS directory named **`loudacre/iplist`** using **`saveAsTextFile`**.	mydata_map3.saveAsTextFile("bdad17/loudacre/iplist")

5. Transform a small dataset using RDDs. (continued)

19) Provide a screenshot of the contents of the **`loudacre/iplist`** folder. (Paste it below.)

```
[cloudera@quickstart ~]$ hdfs dfs -ls bdad17/loudacre
Found 2 items
drwxr-xr-x   - cloudera cloudera          0 2017-10-01 18:28 bdad17/loudacre/iplist
drwxr-xr-x   - cloudera cloudera          0 2017-10-01 17:38 bdad17/loudacre/weblog
[cloudera@quickstart ~]$ hdfs dfs -ls bdad17/loudacre/iplist
Found 3 items
-rw-r--r--   1 cloudera cloudera          0 2017-10-01 18:28 bdad17/loudacre/iplist/_SUCCESS
-rw-r--r--   1 cloudera cloudera      50653 2017-10-01 18:28 bdad17/loudacre/iplist/part-00000
-rw-r--r--   1 cloudera cloudera      50638 2017-10-01 18:28 bdad17/loudacre/iplist/part-00001
```

6. Transform a large dataset using RDDs.

20) Initialize **`logfile`**.	val logfile = "bdad17/loudacre/weblogs"

21) Create an RDD from the file.	val mydata = sc.textFile(logfile)
22) View the first 10 lines of the data.	mydata.take(10)
23) Create an RDD containing only lines that are requests for **`jpg`** files.	val mydata_filter = mydata.filter(line => line.contains(".jpg"))
24) View the first 10 lines of the data.	mydata_filter.take(10)
25) Chain the previous commands into a single command that counts the number of JPG requests.	mydata.filter(line => line.contains(".jpg")).count() res21: Long = 64978
26) Create an RDD using the **`map`** function to return the length of each line of the log file	val mydata_map1= mydata.map(line => line.length())
27) Create an RDD using the **`map`** and **`split`** functions to map an array of words for each line.	val mydata_map2= mydata.map(line => line.split(" "))
28) Create an RDD containing only the IP addresses from each line.	val mydata_map3= mydata.map(line => line.split(" ")(0))
29) Use **`foreach(println)`** to output IP addresses.	mydata_map3.foreach(println)
30) Save the list of IP addresses to a file in an HDFS directory named **`loudacre/bigiplist`** - use **`saveAsTextFile`**.	mydata_map3.saveAsTextFile("bdad17/loudacre/bigiplist")

6. Transform a large dataset using RDDs. (continued)

31) Provide a screenshot of the contents of the **`loudacre/bigiplist`** folder. (Paste it below.)

```
[cloudera@quickstart ~]$ hdfs dfs -ls bdad17/loudacre/bigiplist
Found 312 items
-rw-r--r--   1 cloudera cloudera          0 2017-10-01 18:55 bdad17/loudacre/bigiplist/_SUCCESS
-rw-r--r--   1 cloudera cloudera      49904 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00000
-rw-r--r--   1 cloudera cloudera      49904 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00001
-rw-r--r--   1 cloudera cloudera      49811 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00002
-rw-r--r--   1 cloudera cloudera      50010 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00003
-rw-r--r--   1 cloudera cloudera      49840 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00004
-rw-r--r--   1 cloudera cloudera      49663 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00005
-rw-r--r--   1 cloudera cloudera      50035 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00006
-rw-r--r--   1 cloudera cloudera      49867 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00007
-rw-r--r--   1 cloudera cloudera      49915 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00008
-rw-r--r--   1 cloudera cloudera      49964 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00009
-rw-r--r--   1 cloudera cloudera      49862 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00010
-rw-r--r--   1 cloudera cloudera      50016 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00011
-rw-r--r--   1 cloudera cloudera      49900 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00012
-rw-r--r--   1 cloudera cloudera      49840 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00013
-rw-r--r--   1 cloudera cloudera      49854 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00014
-rw-r--r--   1 cloudera cloudera      49846 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00015
-rw-r--r--   1 cloudera cloudera      50041 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00016
-rw-r--r--   1 cloudera cloudera      49611 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00017
-rw-r--r--   1 cloudera cloudera      49828 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00018
-rw-r--r--   1 cloudera cloudera      49879 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00019
-rw-r--r--   1 cloudera cloudera      49966 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00020
-rw-r--r--   1 cloudera cloudera      49973 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00021
-rw-r--r--   1 cloudera cloudera      49898 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00022
-rw-r--r--   1 cloudera cloudera      49846 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00023
-rw-r--r--   1 cloudera cloudera      49914 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00024
-rw-r--r--   1 cloudera cloudera      49682 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00025
-rw-r--r--   1 cloudera cloudera      50058 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00026
-rw-r--r--   1 cloudera cloudera      49853 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00027
-rw-r--r--   1 cloudera cloudera      50019 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00028
-rw-r--r--   1 cloudera cloudera      49911 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00029
-rw-r--r--   1 cloudera cloudera      49762 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00030
-rw-r--r--   1 cloudera cloudera      49896 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00031
-rw-r--r--   1 cloudera cloudera      49774 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00032
-rw-r--r--   1 cloudera cloudera      50011 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00033
-rw-r--r--   1 cloudera cloudera      49903 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00034
-rw-r--r--   1 cloudera cloudera      49760 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00035
-rw-r--r--   1 cloudera cloudera      49795 2017-10-01 18:55 bdad17/loudacre/bigiplist/part-00036
```