Big Data Application Development - Summer 2017

Homework 3, Part 2 Answer Sheet

4. Use the REPL to explore Spark RDDs.

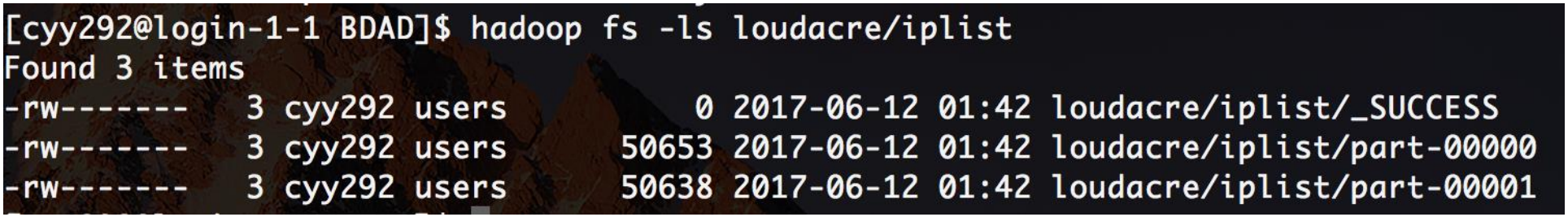
1) Provide the command you used to create your RDD.	`val mydata = sc.textFile("File:///home/cyy292/BDAD/frostroad.txt")`
2) Provide the command you used to count the elements (lines) in your RDD.	`mydata.count`
3) Provide the number of elements.	`res1: Long = 23`
4) Provide the collect command you used.	`mydata.collect`
5) Provide the command you used to create the HDFS directory.	`hadoop fs -mkdir loudacre` `hadoop fs -mkdir loudacre/weblog`
6) Provide the command you used to put the file into HDFS.	`hadoop fs -put 2014-03-15.log loudacre/weblog`
7) Provide the command you used to view the file.	`hadoop fs -cat loudacre/weblog/2014-03-15.log`

5. Transform a small dataset using RDDs.

8) Initialize **logfile**.	`val logfile = "/user/cyy292/loudacre/weblog/2014-03-15.log"`
9) Create an RDD from the file.	`val logRDD = sc.textFile(logfile)`
10) View the first 10 lines of the data.	`logRDD.take(10).foreach(println)`
11) Create an RDD containing only lines that are requests for **jpg** files.	`val logRDD_jpg = logRDD.filter(_.contains("jpg"))`
12) View the first 10 lines of the data.	`logRDD_jpg.take(10).foreach(println)`
13) Chain the previous commands into a single command that counts the number of JPG requests.	`scala> logRDD.filter(_.contains("jpg")).count` `res24: Long = 423`
14) Create an RDD using the **map** function to return the length of each line of the log file.	`val line_len = logRDD.map(_.length)`
15) Create an RDD using the **map** and **split** functions to map an array of words for each line.	`val line_split = logRDD.map(_.split(' '))`
16) Create an RDD containing only the IP addresses from each line.	`val line_IP = line_split.map(_(0))`
17) Use **foreach(println)** to output IP addresses.	`line_IP.foreach(println)`
18) Save the list of IP addresses to an HDFS directory named **loudacre/iplist** using **saveAsTextFile**.	`line_IP.saveAsTextFile("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.)



6. Transform a large dataset using RDDs.

20) Initialize **logfile**.	`val logfile = "loudacre/weblogs/FlumeData.1424275921"`

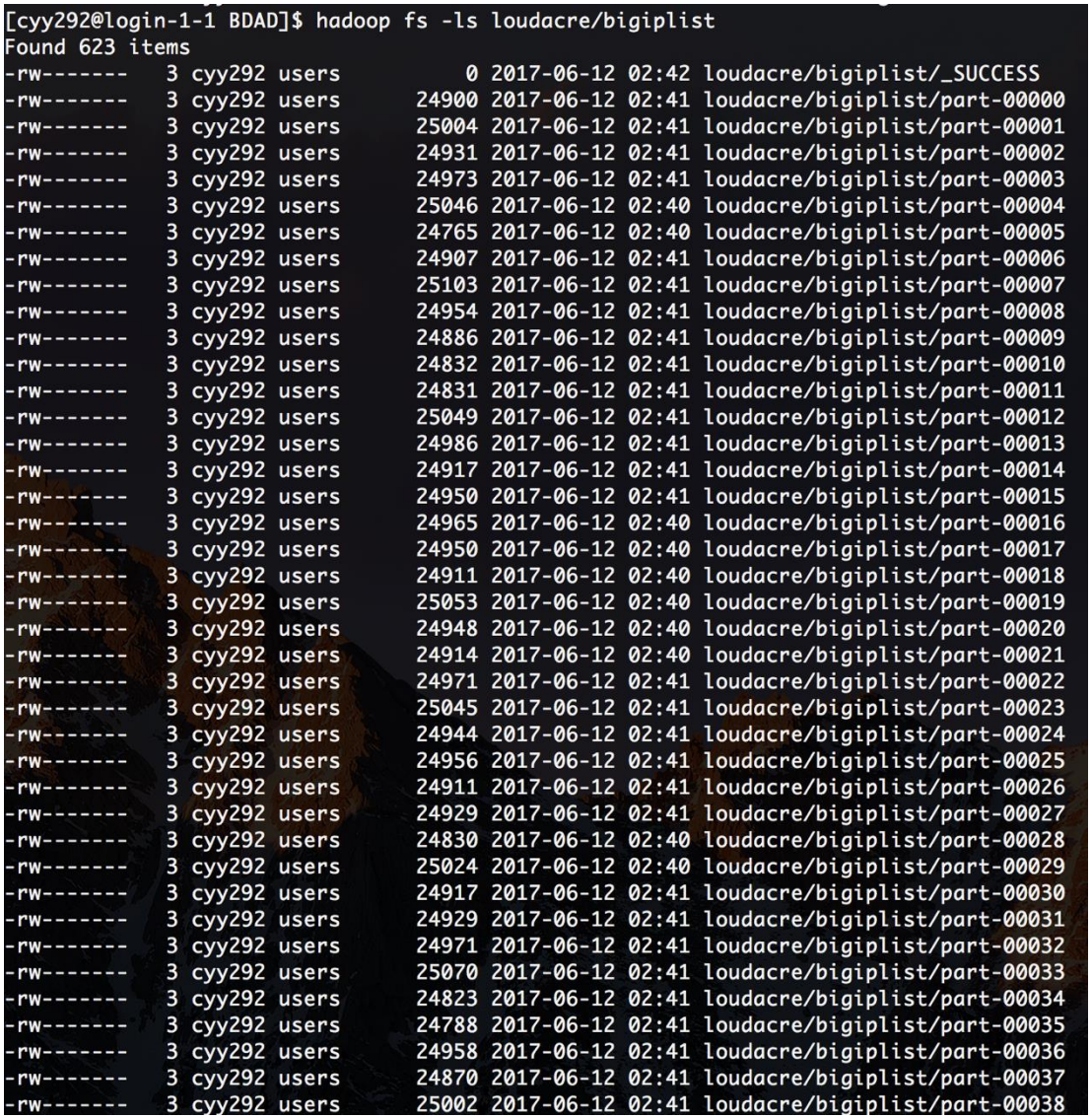
21) Create an RDD from the file.	`var logsRDD: org.apache.spark.rdd.RDD[String] = sc.emptyRDD`
	`scala> for (a <- 226 to 536)`
	`
		logsRDD = logsRDD.union(sc.textFile(logfile + a.toString))`

22) View the first 10 lines of the data.	`logsRDD.take(10).foreach(println)`
23) Create an RDD containing only lines that are requests for **jpg** files.	`val logsRDD_jpg = logsRDD.filter(_.contains("jpg"))`
24) View the first 10 lines of the data.	`logsRDD_jpg.take(10).foreach(println)`
25) Chain the previous commands into a single command that counts the number of JPG requests.	`scala> logsRDD.filter(_.contains("jpg")).count` `res53: Long = 64978`
26) Create an RDD using the **map** function to return the length of each line of the log file	`val logsRDD_length = logsRDD.map(_.length)`
27) Create an RDD using the **map** and **split** functions to map an array of words for each line.	`val logsRDD_split = logsRDD.map(_.split(' '))`
28) Create an RDD containing only the IP addresses from each line.	`val logsRDD_IP = logsRDD_split.map(_(0))`
29) Use **foreach(println)** to output IP addresses.	`logsRDD_IP.foreach(println)`

30) Save the list of IP addresses to a file in an HDFS directory named **`loudacre/bigiplist`** - use **`saveAsTextFile`**.	`logsRDD_IP.saveAsTextFile("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.)



```
[cyy292@login-1-1 BDAD]$ hadoop fs -ls loudacre/bigiplist
Found 623 items
-rw-------   3 cyy292 users          0 2017-06-12 02:42 loudacre/bigiplist/_SUCCESS
-rw-------   3 cyy292 users      24900 2017-06-12 02:41 loudacre/bigiplist/part-00000
-rw-------   3 cyy292 users      25004 2017-06-12 02:41 loudacre/bigiplist/part-00001
-rw-------   3 cyy292 users      24931 2017-06-12 02:41 loudacre/bigiplist/part-00002
-rw-------   3 cyy292 users      24973 2017-06-12 02:41 loudacre/bigiplist/part-00003
-rw-------   3 cyy292 users      25046 2017-06-12 02:40 loudacre/bigiplist/part-00004
-rw-------   3 cyy292 users      24765 2017-06-12 02:40 loudacre/bigiplist/part-00005
-rw-------   3 cyy292 users      24907 2017-06-12 02:41 loudacre/bigiplist/part-00006
-rw-------   3 cyy292 users      25103 2017-06-12 02:41 loudacre/bigiplist/part-00007
-rw-------   3 cyy292 users      24954 2017-06-12 02:41 loudacre/bigiplist/part-00008
-rw-------   3 cyy292 users      24886 2017-06-12 02:41 loudacre/bigiplist/part-00009
-rw-------   3 cyy292 users      24832 2017-06-12 02:41 loudacre/bigiplist/part-00010
-rw-------   3 cyy292 users      24831 2017-06-12 02:41 loudacre/bigiplist/part-00011
-rw-------   3 cyy292 users      25049 2017-06-12 02:41 loudacre/bigiplist/part-00012
-rw-------   3 cyy292 users      24986 2017-06-12 02:41 loudacre/bigiplist/part-00013
-rw-------   3 cyy292 users      24917 2017-06-12 02:41 loudacre/bigiplist/part-00014
-rw-------   3 cyy292 users      24950 2017-06-12 02:41 loudacre/bigiplist/part-00015
-rw-------   3 cyy292 users      24965 2017-06-12 02:40 loudacre/bigiplist/part-00016
-rw-------   3 cyy292 users      24950 2017-06-12 02:40 loudacre/bigiplist/part-00017
-rw-------   3 cyy292 users      24911 2017-06-12 02:40 loudacre/bigiplist/part-00018
-rw-------   3 cyy292 users      25053 2017-06-12 02:40 loudacre/bigiplist/part-00019
-rw-------   3 cyy292 users      24948 2017-06-12 02:40 loudacre/bigiplist/part-00020
-rw-------   3 cyy292 users      24914 2017-06-12 02:40 loudacre/bigiplist/part-00021
-rw-------   3 cyy292 users      24971 2017-06-12 02:41 loudacre/bigiplist/part-00022
-rw-------   3 cyy292 users      25045 2017-06-12 02:41 loudacre/bigiplist/part-00023
-rw-------   3 cyy292 users      24944 2017-06-12 02:41 loudacre/bigiplist/part-00024
-rw-------   3 cyy292 users      24956 2017-06-12 02:41 loudacre/bigiplist/part-00025
-rw-------   3 cyy292 users      24911 2017-06-12 02:41 loudacre/bigiplist/part-00026
-rw-------   3 cyy292 users      24929 2017-06-12 02:41 loudacre/bigiplist/part-00027
-rw-------   3 cyy292 users      24830 2017-06-12 02:40 loudacre/bigiplist/part-00028
-rw-------   3 cyy292 users      25024 2017-06-12 02:40 loudacre/bigiplist/part-00029
-rw-------   3 cyy292 users      24917 2017-06-12 02:41 loudacre/bigiplist/part-00030
-rw-------   3 cyy292 users      24929 2017-06-12 02:41 loudacre/bigiplist/part-00031
-rw-------   3 cyy292 users      24971 2017-06-12 02:41 loudacre/bigiplist/part-00032
-rw-------   3 cyy292 users      25070 2017-06-12 02:41 loudacre/bigiplist/part-00033
-rw-------   3 cyy292 users      24823 2017-06-12 02:41 loudacre/bigiplist/part-00034
-rw-------   3 cyy292 users      24788 2017-06-12 02:41 loudacre/bigiplist/part-00035
-rw-------   3 cyy292 users      24958 2017-06-12 02:41 loudacre/bigiplist/part-00036
-rw-------   3 cyy292 users      24870 2017-06-12 02:41 loudacre/bigiplist/part-00037
-rw-------   3 cyy292 users      25002 2017-06-12 02:41 loudacre/bigiplist/part-00038
```