Big Data Application Development - Summer 2017

Homework 3, Part 2 Answer Sheet

4. Use the REPL to explore Spark RDDs.

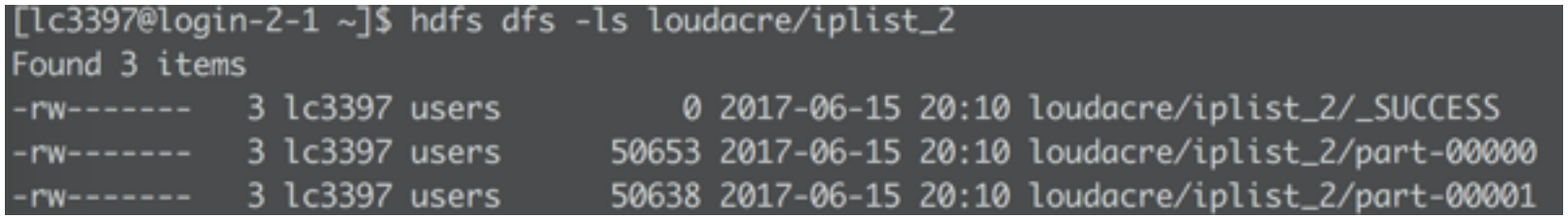
1) Provide the command you used to create your RDD.	val mydata = sc.textFile("/user/lc3397/class3_hw/frostroad.txt")
2) Provide the command you used to count the elements (lines) in your RDD.	mydata.count()
3) Provide the number of elements.	23
4) Provide the collect command you used.	mydata.collect()
5) Provide the command you used to create the HDFS directory.	[lc3397@login-1-1 ~]$ hdfs dfs -mkdir loudacre [lc3397@login-1-1 ~]$ hdfs dfs -mkdir loudacre/weblog
6) Provide the command you used to put the file into HDFS.	[lc3397@login-1-1 ~]$ hdfs dfs -put 2014-03-15.log loudacre/weblog
7) Provide the command you used to view the file.	hdfs dfs -cat loudacre/weblog/2014-03-15.log

5. Transform a small dataset using RDDs.

8) Initialize **`logfile`**.	val logfile_name = "/user/lc3397/loudacre/weblog/2014-03-15.log"
9) Create an RDD from the file.	val logfile = sc.textFile(logfile_name)
10) View the first 10 lines of the data.	logfile.take(10)
11) Create an RDD containing only lines that are requests for **`jpg`** files.	val pattern = ".jpg|.jpeg|.JPG|.JPEG".r val jpg_data_rdd = logfile.filter(line => pattern.findFirstIn(line).isDefined)
12) View the first 10 lines of the data.	jpg_data_rdd.take(10).foreach(println)
13) **Chain** the previous commands into a single command that counts the number of JPG requests.	logfile.map(line=>line.toLowerCase()).filter(line=>line.contains("jpg")).count()
14) Create an RDD using the **`map`** function to return the length of each line of the log file.	val line_length = logfile.map(s => s.length)
15) Create an RDD using the **`map`** and **`split`** functions to map an array of words for each line.	val splitarray = logfile.map(s => s.split(" "))
16) Create an RDD containing only the IP addresses from each line.	val ip_addrs = splitarray.map(s => s(0))
17) Use **`foreach(println)`** to output IP addresses.	ip_addrs.collect().foreach(println)
18) Save the list of IP addresses to an HDFS directory named **`loudacre/iplist`** using **`saveAsTextFile`**.	ip_addrs.saveAsTextFile("loudacre/iplist_2")

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 logfiles_name = "/user/lc3397/loudacre/weblogs/*"
21) Create an RDD from the file.	val logfiles = sc.textFile(logfiles_name)
22) View the first 10 lines of the data.	logfiles.take(10)
23) Create an RDD containing only lines that are requests for **`jpg`** files.	val big_jpg_data = logfiles.map(line=>line.toLowerCase()).filter(line=>line.contains(".jpg"))
24) View the first 10 lines of the data.	big_jpg_data.take(10).foreach(println)
25) Chain the previous commands into a single command that counts the number of JPG requests.	logfiles.map(line=>line.toLowerCase()).filter(line=>line.contains(“.jpg”)).count()
26) Create an RDD using the **`map`** function to return the length of each line of the log file	val line_length = logfiles.map(s => s.length())
27) Create an RDD using the **`map`** and **`split`** functions to map an array of words for each line.	val splitarray_large = logfiles.map(s => s.split(" "))
28) Create an RDD containing only the IP addresses from each line.	val ip_addrs_large = splitarray_large.map(s => s(0))
29) Use **`foreach(println)`** to output IP addresses.	ip_addrs_large.collect().foreach(println)
30) Save the list of IP addresses to a file in an HDFS directory named **`loudacre/bigiplist`** - use **`saveAsTextFile`**.	ip_addrs_large.saveAsTextFile(“loudacre/iplist_large”)

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

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

```
[lc3397@login-2-1 ~]$ hdfs dfs -ls loudacre/iplist_large
Found 494 items
-rw-------   3 lc3397 users          0 2017-06-15 22:12 loudacre/iplist_large/_SUCCESS
-rw-------   3 lc3397 users      49265 2017-06-15 22:11 loudacre/iplist_large/part-00000
-rw-------   3 lc3397 users      45854 2017-06-15 22:11 loudacre/iplist_large/part-00001
-rw-------   3 lc3397 users      50031 2017-06-15 22:11 loudacre/iplist_large/part-00002
-rw-------   3 lc3397 users      45898 2017-06-15 22:11 loudacre/iplist_large/part-00003
-rw-------   3 lc3397 users      48070 2017-06-15 22:11 loudacre/iplist_large/part-00004
-rw-------   3 lc3397 users      46430 2017-06-15 22:11 loudacre/iplist_large/part-00005
-rw-------   3 lc3397 users      46177 2017-06-15 22:11 loudacre/iplist_large/part-00006
-rw-------   3 lc3397 users      50720 2017-06-15 22:11 loudacre/iplist_large/part-00007
-rw-------   3 lc3397 users      47314 2017-06-15 22:11 loudacre/iplist_large/part-00008
-rw-------   3 lc3397 users      46282 2017-06-15 22:11 loudacre/iplist_large/part-00009
-rw-------   3 lc3397 users      45998 2017-06-15 22:11 loudacre/iplist_large/part-00010
-rw-------   3 lc3397 users      49261 2017-06-15 22:11 loudacre/iplist_large/part-00011
```

```
[lc3397@login-2-1 ~]$ hdfs dfs -ls loudacre/
Found 5 items
drwx------   - lc3397 users          0 2017-06-15 20:07 loudacre/iplist
drwx------   - lc3397 users          0 2017-06-15 20:10 loudacre/iplist_2
drwx------   - lc3397 users          0 2017-06-15 22:12 loudacre/iplist_large
drwx------   - lc3397 users          0 2017-06-14 23:36 loudacre/weblog
drwx------   - lc3397 users          0 2017-06-15 20:15 loudacre/weblogs
```