

# Big Data Application Development - Spring 2019

## Homework 3, Part 2 Answer Sheet

4. Use the REPL to explore Spark RDDs.

<p>1) Provide the command you used to create your RDD.</p> <p>2) Provide the command you used to count the elements (lines) in your RDD.</p> <p>3) Provide the number of elements.</p> <p>4) Provide the collect command you used.</p> <p>5) Provide the command you used to create the HDFS directory.</p>	<pre>val mydata = sc.textFile("file:///home/drr342/bdad/hw3/frostroad.txt")  mydata.count  23  mydata.collect  hdfs dfs -mkdir bdad/hw3 hdfs dfs -mkdir bdad/hw3/loudacre hdfs dfs -mkdir bdad/hw3/loudacre/weblog</pre>
<p>6) Provide the command you used to put the file into HDFS.</p>	<pre>hdfs dfs -put 2014-03-15.log bdad/hw3/loudacre/weblog/</pre>
<p>7) Provide the command you used to view the file.</p>	<pre>hdfs dfs -cat bdad/hw3/</pre>

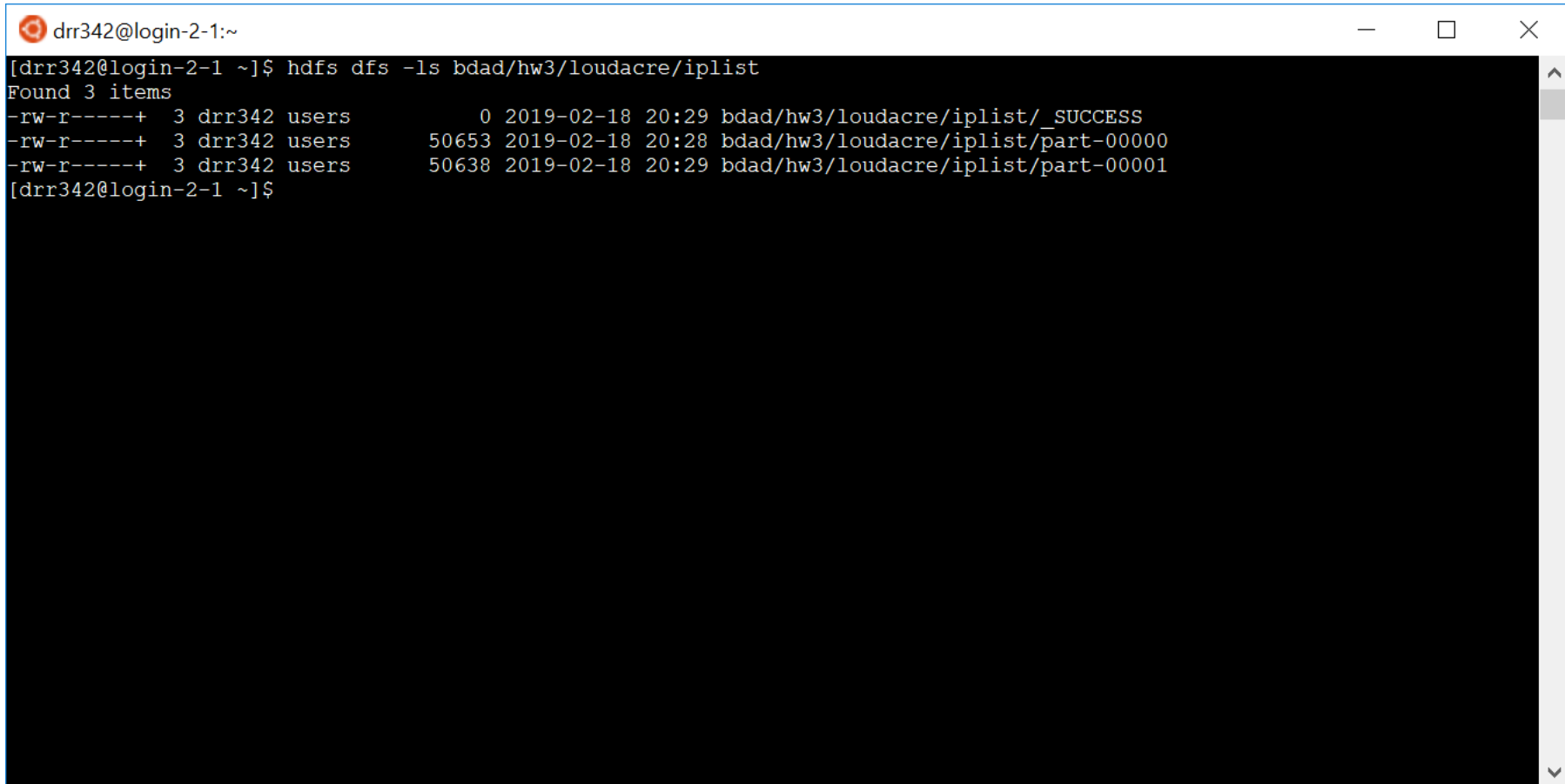
## 5. Transform a small dataset using RDDs.

8) Initialize <code>logfile</code> .	<code>val logfile: String = "/user/drr342/bdad/hw3/loudacre/weblog/2014-03-15.log"</code>
9) Create an RDD from the file.	<code>val logrdd = sc.textFile(logfile)</code>
10) View the first 10 lines of the data.	<code>logrdd.take(10)</code>
11) Create an RDD containing only lines that are requests for <code>jpg</code> files.	<code>val jpgrdd = logrdd.filter(_.toLowerCase().contains("jpg"))</code>
12) View the first 10 lines of the data.	<code>jpgrdd.take(10)</code>
13) Chain the previous commands into a single command that counts the number of JPG requests.	<code>val jpgcount = logrdd.filter(_.toLowerCase().contains("jpg")).count</code>
14) Create an RDD using the <code>map</code> function to return the length of each line of the log file.	<code>val logLengthsrdd = logrdd.map(_.length)</code>
15) Create an RDD using the <code>map</code> and <code>split</code> functions to map an array of words for each line.	<code>val logWordsrdd = logrdd.map(_.split(" "))</code>
16) Create an RDD containing only the IP addresses from each line.	<code>val ip = raw"(\d{1,3}\.){3}\d{1,3}".r val logIPrdd = logrdd.map(ip.findFirstIn(_).getOrElse("No IP address found!"))</code>
17) Use <code>foreach(println)</code> to output IP addresses.	<code>logIPrdd.collect.foreach(println)</code>
18) Save the list of IP addresses to an HDFS directory named	<code>val ipfile: String = "/user/drr342/bdad/hw3/loudacre/iplist" logIPrdd.saveAsTextFile(ipfile)</code>

loudacre/iplist using  
saveAsTextFile.

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

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

A terminal window with a dark background and light text. The window title is 'drr342@login-2-1:~'. The command 'hdfs dfs -ls bdad/hw3/loudacre/iplist' has been executed. The output shows three items: a directory named '\_SUCCESS' and two files named 'part-00000' and 'part-00001'. The file sizes are 50653 and 50638 bytes respectively. The terminal window has standard window controls (minimize, maximize, close) in the top right corner.

```
drr342@login-2-1:~$ hdfs dfs -ls bdad/hw3/loudacre/iplist
Found 3 items
-rw-r-----+ 3 drr342 users      0 2019-02-18 20:29 bdad/hw3/loudacre/iplist/_SUCCESS
-rw-r-----+ 3 drr342 users  50653 2019-02-18 20:28 bdad/hw3/loudacre/iplist/part-00000
-rw-r-----+ 3 drr342 users  50638 2019-02-18 20:29 bdad/hw3/loudacre/iplist/part-00001
drr342@login-2-1:~$
```

## 6. Transform a large dataset using RDDs.

20) Initialize <code>logfile</code> .	<code>val weblogfile: String = "/user/drr342/bdad/hw3/loudacre/weblogs/"</code>
21) Create an RDD from the file.	<code>val weblogsrdd = sc.textFile(weblogfile)</code>
22) View the first 10 lines of the data.	<code>weblogsrdd.take(10)</code>
23) Create an RDD containing only lines that are requests for <code>jpg</code> files.	<code>val webjpgrrdd = weblogsrdd.filter(_.toLowerCase().contains("jpg"))</code>
24) View the first 10 lines of the data.	<code>webjpgrrdd.take(10)</code>
25) Chain the previous commands into a single command that counts the number of JPG requests.	<code>val webjpgcount = weblogsrdd.filter(_.toLowerCase().contains("jpg")).count</code>
26) Create an RDD using the <code>map</code> function to return the length of each line of the log file	<code>val weblogLengthsrdd = weblogsrdd.map(_.length)</code>
27) Create an RDD using the <code>map</code> and <code>split</code> functions to map an array of words for each line.	<code>val weblogWordsrdd = weblogsrdd.map(_.split(" "))</code>
28) Create an RDD containing only the IP addresses from each line.	<code>val weblogIPrdd = weblogsrdd.map(ip.findFirstIn(_).getOrElse("No IP address found!"))</code>
29) Use <code>foreach</code> ( <code>println</code> ) to output IP addresses.	<code>weblogIPrdd.collect.foreach(println)</code>

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

```
val bigipfile: String = "/user/drr342/bdad/hw3/loudacre/bigiplist"
weblogsIPrdd.saveAsTextFile(bigipfile)
```

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

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

```
drr342@login-2-1:~/bdad/hw3
[drr342@login-2-1 hw3]$ hdfs dfs -ls bdad/hw3/loudacre/bigiplist
Found 312 items
-rw-r----- 3 drr342 users          0 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/_SUCCESS
-rw-r----- 3 drr342 users    49904 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00000
-rw-r----- 3 drr342 users    49904 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00001
-rw-r----- 3 drr342 users    49811 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00002
-rw-r----- 3 drr342 users    50010 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00003
-rw-r----- 3 drr342 users    49840 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00004
-rw-r----- 3 drr342 users    49663 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00005
-rw-r----- 3 drr342 users    50035 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00006
-rw-r----- 3 drr342 users    49867 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00007
-rw-r----- 3 drr342 users    49915 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00008
-rw-r----- 3 drr342 users    49964 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00009
-rw-r----- 3 drr342 users    49862 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00010
-rw-r----- 3 drr342 users    50016 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00011
-rw-r----- 3 drr342 users    49900 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00012
-rw-r----- 3 drr342 users    49840 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00013
-rw-r----- 3 drr342 users    49854 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00014
-rw-r----- 3 drr342 users    49846 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00015
-rw-r----- 3 drr342 users    50041 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00016
-rw-r----- 3 drr342 users    49611 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00017
-rw-r----- 3 drr342 users    49828 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00018
-rw-r----- 3 drr342 users    49879 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00019
-rw-r----- 3 drr342 users    49966 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00020
-rw-r----- 3 drr342 users    49973 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00021
-rw-r----- 3 drr342 users    49898 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00022
-rw-r----- 3 drr342 users    49846 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00023
-rw-r----- 3 drr342 users    49914 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00024
-rw-r----- 3 drr342 users    49682 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00025
-rw-r----- 3 drr342 users    50058 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00026
-rw-r----- 3 drr342 users    49853 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00027
-rw-r----- 3 drr342 users    50019 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00028
-rw-r----- 3 drr342 users    49911 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00029
-rw-r----- 3 drr342 users    49762 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00030
-rw-r----- 3 drr342 users    49896 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00031
-rw-r----- 3 drr342 users    49774 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00032
-rw-r----- 3 drr342 users    50011 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00033
-rw-r----- 3 drr342 users    49903 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00034
-rw-r----- 3 drr342 users    49760 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00035
-rw-r----- 3 drr342 users    49795 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00036
-rw-r----- 3 drr342 users    49911 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00037
-rw-r----- 3 drr342 users    49784 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00038
-rw-r----- 3 drr342 users    49893 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00039
-rw-r----- 3 drr342 users    49774 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00040
-rw-r----- 3 drr342 users    49825 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00041
-rw-r----- 3 drr342 users    49947 2019-02-18 20:51 bdad/hw3/loudacre/bigiplist/part-00042
```