**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. |  |
| 2) Provide the command you used to count the elements (lines) in your RDD. |  |
| 3) Provide the number of elements. |  |
| 4) Provide the collect command you used. |  |
| 5) Provide the command you used to create the HDFS directory. |  |
| 6) Provide the command you used to put the file into HDFS. |  |
| 7) Provide the command you used to view the file. |  |

5. Transform a small dataset using RDDs.

|  |  |
| --- | --- |
| 8) Initialize **logfile**. |  |
| 9) Create an RDD from the file. |  |
| 10) View the first 10 lines of the data. |  |
| 11) Create an RDD containing only lines that are requests for **jpg** files. |  |
| 12) View the first 10 lines of the data. |  |
| 13) Chain the previous commands into a single command that counts the number of JPG requests. |  |
| 14) Create an RDD using the **map** function to return the length of each line of the log file. |  |
| 15) Create an RDD using the **map** and **split** functions to map an array of words for each line. |  |
| 16) Create an RDD containing only the IP addresses from each line. |  |
| 17) Use **foreach(println)** to output IP addresses. |  |
| 18) Save the list of IP addresses to an HDFS directory named **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.)

6. Transform a large dataset using RDDs.

|  |  |
| --- | --- |
| 20) Initialize **logfile**. |  |
| 21) Create an RDD from the file. |  |
| 22) View the first 10 lines of the data. |  |
| 23) Create an RDD containing only lines that are requests for **jpg** files. |  |
| 24) View the first 10 lines of the data. |  |
| 25) Chain the previous commands into a single command that counts the number of JPG requests. |  |
| 26) Create an RDD using the **map** function to return the length of each line of the log file |  |
| 27) Create an RDD using the **map** and **split** functions to map an array of words for each line. |  |
| 28) Create an RDD containing only the IP addresses from each line. |  |
| 29) Use **foreach(println)** to output IP addresses. |  |
| 30) Save the list of IP addresses to a file in an HDFS directory named **loudacre/bigiplist** - use **saveAsTextFile**. |  |

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

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