**Big Data Application Development - Spring 2019**

**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/drr342/bdad/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 bdad/hw3  hdfs dfs -mkdir bdad/hw3/loudacre  hdfs dfs -mkdir bdad/hw3/loudacre/weblog |
| 6) Provide the command you used to put the file into HDFS. | hdfs dfs -put 2014-03-15.log bdad/hw3/loudacre/weblog/ |
| 7) Provide the command you used to view the file. | hdfs dfs -cat bdad/hw3/ |

5. Transform a small dataset using RDDs.

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

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

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

A screenshot of a computer

Description automatically generated

6. Transform a large dataset using RDDs.

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

A screenshot of a computer

Description automatically generated