

CS 5433: Bigdata Management
Programming Assignment 1
PART 2 – ReadMe for MapReduce Program for Row Count

CWID: A20343337

1. Create a java file for Row Count using nano command or in WinSCP.
2. Compile the java program by using below command. (Refer the RowCount.java code for SatyaRajyaSaiTejaswini_Darapureddy_Program_PA2)

```
sdarapu@hadoop-nn001:~$ hadoop com.sun.tools.javac.Main RowCount.java
```

3. Once the program is compiled successfully, create a jar file as shown below.

```
sdarapu@hadoop-nn001:~$ jar cf rc.jar RowCount*.class
```

4. Now, run the jar file

[For NASA data:](#)

```
sdarapu@hadoop-nn001:~$ hadoop jar rc.jar RowCount /user/sdarapu/NASA_PA1data/2022/02/26/10/FlumeData.* /user/sdarapu/RowCountOutput_NASA
```

[For SpaceX data:](#)

```
sdarapu@hadoop-nn001:~$ hadoop jar rc.jar RowCount /user/sdarapu/SpaceX_PA1data/2022/02/26/11/FlumeData.* /user/sdarapu/RowCountOutput_SpaceX
```

5. The program file gets executed.

[For NASA data:](#)

```
sdarapu@hadoop-nn001:~$ hadoop jar rc.jar RowCount /user/sdarapu/NASA_PA1data/2022/02/26/10/FlumeData.* /user/sdarapu/RowCountOutput_NASAtOutput_NASA
2022-03-06 11:41:49,909 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2022-03-06 11:41:50,701 INFO client.DefaultNoHARMFoverProxyProvider: Connecting to ResourceManager at hadoop-nn001.cs.okstate.edu/192.168.122.2:8032
2022-03-06 11:41:51,182 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute
ToolRunner to remedy this.
2022-03-06 11:41:51,203 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/sdarapu/.staging/job_1646249209374_0437
2022-03-06 11:41:51,557 INFO input.FileInputFormat: Total input files to process : 73
2022-03-06 11:41:51,937 INFO mapreduce.JobSubmitter: number of splits:73
2022-03-06 11:41:52,106 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1646249209374_0437
2022-03-06 11:41:52,106 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-03-06 11:41:52,314 INFO conf.Configuration: resource-types.xml not found
2022-03-06 11:41:52,315 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-03-06 11:41:52,414 INFO impl.YarnClientImpl: Submitted application application_1646249209374_0437
2022-03-06 11:41:52,465 INFO mapreduce.Job: The url to track the job: http://hadoop-nn001.cs.okstate.edu:8088/proxy/application_1646249209374_0437/
2022-03-06 11:41:52,466 INFO mapreduce.Job: Running job: job_1646249209374_0437
2022-03-06 11:41:58,582 INFO mapreduce.Job: Job job_1646249209374_0437 running in uber mode : false
2022-03-06 11:41:58,584 INFO mapreduce.Job: map 0% reduce 0%
2022-03-06 11:42:03,696 INFO mapreduce.Job: map 33% reduce 0%
2022-03-06 11:42:06,726 INFO mapreduce.Job: map 40% reduce 0%
2022-03-06 11:42:07,740 INFO mapreduce.Job: map 66% reduce 0%
2022-03-06 11:42:09,762 INFO mapreduce.Job: map 70% reduce 0%
2022-03-06 11:42:10,772 INFO mapreduce.Job: map 78% reduce 0%
2022-03-06 11:42:11,784 INFO mapreduce.Job: map 97% reduce 0%
2022-03-06 11:42:13,803 INFO mapreduce.Job: map 100% reduce 0%
2022-03-06 11:42:14,814 INFO mapreduce.Job: map 100% reduce 100%
2022-03-06 11:42:15,837 INFO mapreduce.Job: Job job_1646249209374_0437 completed successfully
2022-03-06 11:42:15,976 INFO mapreduce.Job: Counters: 55
File System Counters
FILE: Number of bytes read=4167
FILE: Number of bytes written=19588064
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=4170857
```

OpenSSH SSH client

```
Total vcore-milliseconds taken by all reduce tasks=7582
Total megabyte-milliseconds taken by all map tasks=1067494400
Total megabyte-milliseconds taken by all reduce tasks=38819840
Map-Reduce Framework
  Map input records=728
  Map output records=728
  Map output bytes=40040
  Map output materialized bytes=4599
  Input split bytes=12268
  Combine input records=728
  Combine output records=73
  Reduce input groups=1
  Reduce shuffle bytes=4599
  Reduce input records=73
  Reduce output records=1
  Spilled Records=146
  Shuffled Maps =73
  Failed Shuffles=0
  Merged Map outputs=73
  GC time elapsed (ms)=1484
  CPU time spent (ms)=43040
  Physical memory (bytes) snapshot=26263769088
  Virtual memory (bytes) snapshot=472132898816
  Total committed heap usage (bytes)=58321797120
  Peak Map Physical memory (bytes)=368611328
  Peak Map Virtual memory (bytes)=6385541120
  Peak Reduce Physical memory (bytes)=271253504
  Peak Reduce Virtual memory (bytes)=6395412480
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=4166789
File Output Format Counters
  Bytes Written=55
sdarapu@hadoop-nn001:~$
```

For SpaceX data:

```
sdarapu@hadoop-nn001:~$ hadoop jar rc.jar RowCount /user/sdarapu/SpaceX_PA1data/2022/02/26/11/FlumeData.* /user/sdarapu/RowCountOutput_SpaceX
2022-03-06 11:45:26,039 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2022-03-06 11:45:26,941 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at hadoop-nn001.cs.okstate.edu/192.168.122.2:8032
2022-03-06 11:45:27,411 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute
ToolRunner to remedy this.
2022-03-06 11:45:27,429 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/sdarapu/.staging/job_1646249209374_0439
2022-03-06 11:45:27,834 INFO input.FileInputFormat: Total input files to process : 74
2022-03-06 11:45:28,254 INFO mapreduce.JobSubmitter: number of splits:74
2022-03-06 11:45:28,442 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1646249209374_0439
2022-03-06 11:45:28,442 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-03-06 11:45:28,666 INFO conf.Configuration: resource-types.xml not found
2022-03-06 11:45:28,667 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-03-06 11:45:28,805 INFO impl.YarnClientImpl: Submitted application application_1646249209374_0439
2022-03-06 11:45:28,857 INFO mapreduce.Job: The url to track the job: http://hadoop-nn001.cs.okstate.edu:8088/proxy/application_1646249209374_0439/
2022-03-06 11:45:28,858 INFO mapreduce.Job: Running job: job_1646249209374_0439
2022-03-06 11:45:34,978 INFO mapreduce.Job: Job job_1646249209374_0439 running in uber mode : false
2022-03-06 11:45:34,980 INFO mapreduce.Job: map 0% reduce 0%
2022-03-06 11:45:40,088 INFO mapreduce.Job: map 32% reduce 0%
2022-03-06 11:45:43,127 INFO mapreduce.Job: map 38% reduce 0%
2022-03-06 11:45:44,139 INFO mapreduce.Job: map 65% reduce 0%
2022-03-06 11:45:46,158 INFO mapreduce.Job: map 68% reduce 0%
2022-03-06 11:45:47,167 INFO mapreduce.Job: map 70% reduce 0%
2022-03-06 11:45:48,178 INFO mapreduce.Job: map 96% reduce 0%
2022-03-06 11:45:49,188 INFO mapreduce.Job: map 99% reduce 0%
2022-03-06 11:45:50,198 INFO mapreduce.Job: map 100% reduce 0%
2022-03-06 11:45:51,208 INFO mapreduce.Job: map 100% reduce 100%
2022-03-06 11:45:51,225 INFO mapreduce.Job: Job job_1646249209374_0439 completed successfully
2022-03-06 11:45:51,397 INFO mapreduce.Job: Counters: 56
File System Counters
  FILE: Number of bytes read=4167
  FILE: Number of bytes written=19852956
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=4013197
  HDFS: Number of bytes written=55
  HDFS: Number of read operations=227
```

OpenSSH SSH client

```
Total vcore-milliseconds taken by all map tasks=208291
Total vcore-milliseconds taken by all reduce tasks=7283
Total megabyte-milliseconds taken by all map tasks=1066449920
Total megabyte-milliseconds taken by all reduce tasks=37288960
Map-Reduce Framework
  Map input records=730
  Map output records=730
  Map output bytes=40150
  Map output materialized bytes=4605
  Input split bytes=12584
  Combine input records=730
  Combine output records=73
  Reduce input groups=1
  Reduce shuffle bytes=4605
  Reduce input records=73
  Reduce output records=1
  Spilled Records=146
  Shuffled Maps =74
  Failed Shuffles=0
  Merged Map outputs=74
  GC time elapsed (ms)=1302
  CPU time spent (ms)=44720
  Physical memory (bytes) snapshot=26582155264
  Virtual memory (bytes) snapshot=478494691328
  Total committed heap usage (bytes)=59089354752
  Peak Map Physical memory (bytes)=366718976
  Peak Map Virtual memory (bytes)=6387228672
  Peak Reduce Physical memory (bytes)=276185088
  Peak Reduce Virtual memory (bytes)=6409326592
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=4000613
File Output Format Counters
  Bytes Written=55
sdarapu@hadoop-nn001:~$
```

6. Now, copy the output file to Hadoop local by using the below command

For NASA data:

```
sdarapu@hadoop-nn001:~$ hadoop fs -get /user/sdarapu/RowCountOutput_NASA /home/sdarapu
```

For SpaceX data:

```
sdarapu@hadoop-nn001:~$ hadoop fs -get /user/sdarapu/RowCountOutput_SpaceX /home/sdarapu
```

7. To display the output, use the below command

For NASA data:

```
sdarapu@hadoop-nn001:~$ hadoop fs -cat /user/sdarapu/RowCountOutput_NASA/part*
2022-03-06 13:10:05,323 WARN util.NativeCodeLoader: Unable to load native-hadoop lib
tin-java classes where applicable
Total Number of Rows in the downloaded Flume Data:      728
```

For SpaceX data:

```
sdarapu@hadoop-nn001:~$ hadoop fs -cat /user/sdarapu/RowCountOutput_SpaceX/part*
2022-03-06 13:21:33,979 WARN util.NativeCodeLoader: Unable to load native-hadoop lib
tin-java classes where applicable
Total Number of Rows in the downloaded Flume Data:      730
sdarapu@hadoop-nn001:~$
```