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

CWID: A20343337

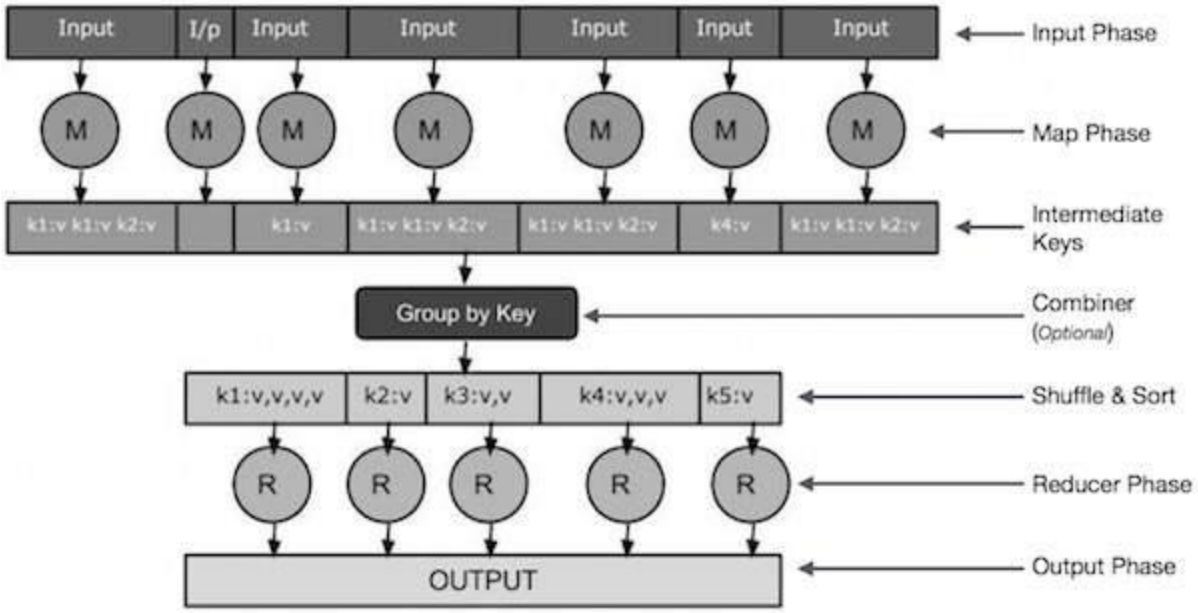
MapReduce: MapReduce is a programming model which makes to write the applications easily and process the enormous amounts of data in parallel on large clusters of commodity hardware in a fault-tolerant and reliable way.

The MapReduce algorithm mainly contains two tasks, called Map and Reduce.

- ➔ Map task takes inputs as set of data and converts it into new set of data in which individual elements are split down into tuples called key-value pairs.
- ➔ Reduce task takes the Map's output as an input and merges those key-value pairs into smaller set of tuples.

There are different phases in MapReduce algorithm, below are the details of those phases,

1. Input Phase: Here, the given data set is taken by the Record Reader which then parses each record in an input file and passes the processed data to the mapper as key-value pairs.
2. Map Phase: Map is a user-defined procedure. It takes series of key-value pairs from input phase and processes each of them into produce zero or more key-value pairs.
3. Intermediate Keys: The key-value pairs produced by the Mapper are called as intermediate keys.
4. Combiner Phase: It is an optional phase in MapReduce algorithm. It is a form of local reducer which groups the similar or related data from the map phase into discoverable sets. It accepts the intermediate keys from the mapper as input and aggregates the values in a narrow scope of one mapper using a user-defined function.
5. Shuffle and Sort Phase: The Reducer task begins with the shuffle and sort phase. It takes the intermediate key-value pairs as input and sorts each key-value pairs based on the key into a larger data list. The equivalent keys in the data list are grouped together so that their values can be iterated flexibly and easily in the reducer phase.
6. Reduce Phase: The Reducer takes the input as grouped key-value paired data and applies a reducer function to each of them individually. Here, the data can be filtered, aggregated, or combined in a variety of ways and it necessitates a variety of processing. When the execution is complete, the final step is given zero or more key-value pairs.
7. Output Phase: In this phase, the key-value pairs from the reducer function are taken by the output formatter which translates them and writes them onto a file using record writer.



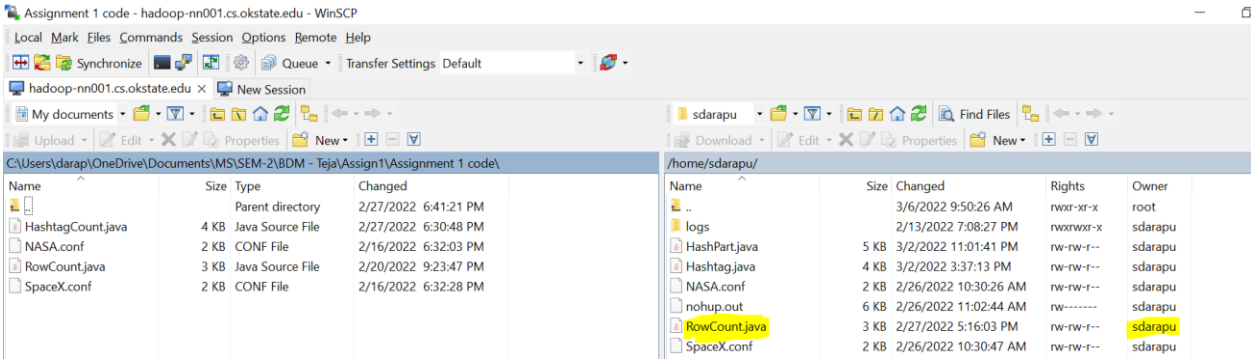
Applying Row Count MapReduce Program on downloaded Flume data:

Here, on downloaded flume data for the keywords NASA and SpaceX, Row count is applied by using MapReduce program where each record in a file is considered as a row and displays the count of those rows as output.

Approach:

Below are the steps I have followed to complete the PART 2 of the assignment,

1) I have created a java file for Row Count using WinSCP instead of nano command as shown below.

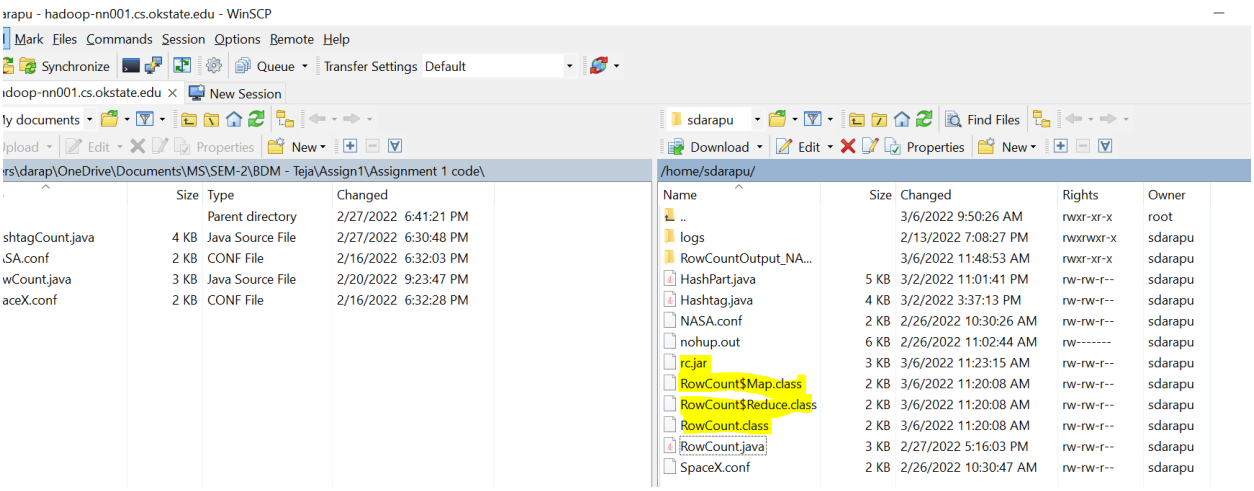


2) Once RowCount.java program is written, I have connected to the hadoop cluster and compiled the java program by using below command. (Refer the RowCount.java code for SatyaRajyaSaiTejaswini_Darapureddy_Program_PA2)

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

3) Once the program is compiled successfully, created a jar file for row count called “rc.jar” by using the below command.

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



4) Now, running 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) Now, 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 applica
2022-03-06 11:41:50,701 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at hadoop-nn001.cs.okstate.edu/192.168.122.2:803
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_164624926
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=4179057
```

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 applicab
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_1646249209
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, copied the output file to Hadoop local by using the below command and also to check the output.

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, below is the command I have executed. We can also use “hadoop fs -cat /path” as we copied the data to Hadoop in the above step.

For NASA data:

```
sdarapu@hadoop-nn001:~$ hdfs dfs -cat /user/sdarapu/RowCountOutput_NASA/part*
2022-03-06 11:50:43,684 WARN util.NativeCodeLoader: Unable to load native-hadoop library
Total Number of Rows in the downloaded Flume Data:      728
sdarapu@hadoop-nn001:~$
```

For total 73 files of downloaded NASA data, I have got 728 rows.

For SpaceX data:

```
sdarapu@hadoop-nn001:~$ hdfs dfs -cat /user/sdarapu/RowCountOutput_SpaceX/part*
2022-03-06 11:52:01,666 WARN util.NativeCodeLoader: Unable to load native-hadoop library
Total Number of Rows in the downloaded Flume Data:      730
sdarapu@hadoop-nn001:~$
```

For total 74 files of downloaded SpaceX data, I have got 730 rows.