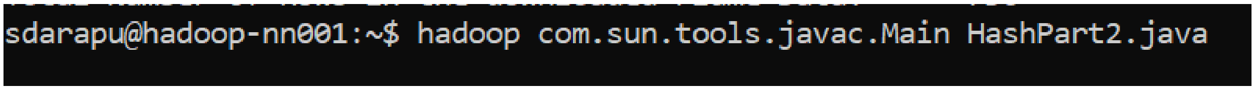
CS 5433: Bigdata Management

Programming Assignment 1

PART 3 – ReadMe for MapReduce Program for Hashtag Count using Partitioner

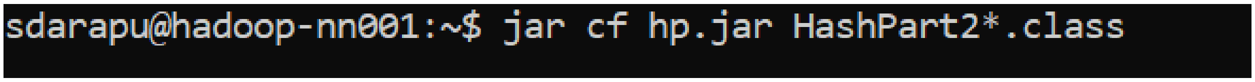
CWID: A20343337

1. Create a java file for Hashtag Count using nano command or in WinSCP.
2. Compile the java program by using below command. (Refer the HashPart2.java code for SatyaRajyaSaiTejaswini_Darapureddy_Program_PA3)



```
sdarapu@hadoop-nn001:~$ hadoop com.sun.tools.javac.Main HashPart2.java
```

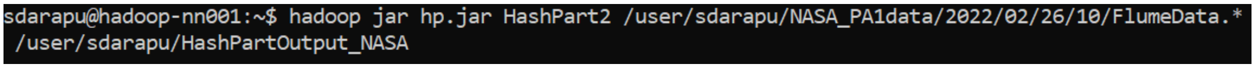
3. Once the program is compiled successfully, create a jar file as shown below.



```
sdarapu@hadoop-nn001:~$ jar cf hp.jar HashPart2*.class
```

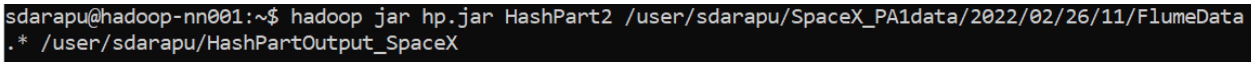
4. Now, run the jar file

For NASA data:



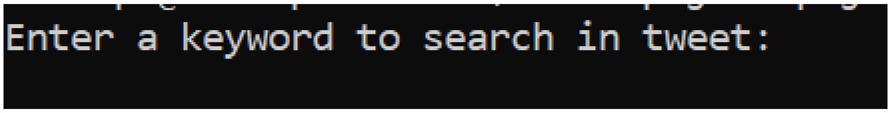
```
sdarapu@hadoop-nn001:~$ hadoop jar hp.jar HashPart2 /user/sdarapu/NASA_PA1data/2022/02/26/10/FlumeData.*
 /user/sdarapu/HashPartOutput_NASA
```

For SpaceX data:



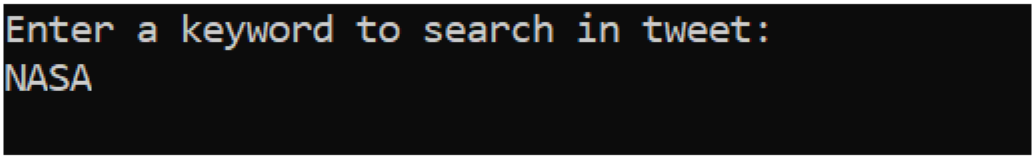
```
sdarapu@hadoop-nn001:~$ hadoop jar hp.jar HashPart2 /user/sdarapu/SpaceX_PA1data/2022/02/26/11/FlumeData
.* /user/sdarapu/HashPartOutput_SpaceX
```

5. Now, System prompts for the input keyword to search in the tweet as shown below. So, enter keyword for searching.



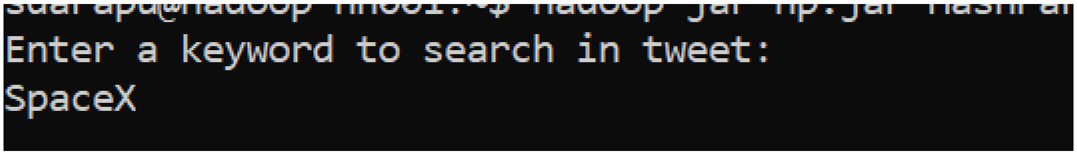
```
Enter a keyword to search in tweet:
```

For NASA data:



```
Enter a keyword to search in tweet:
NASA
```

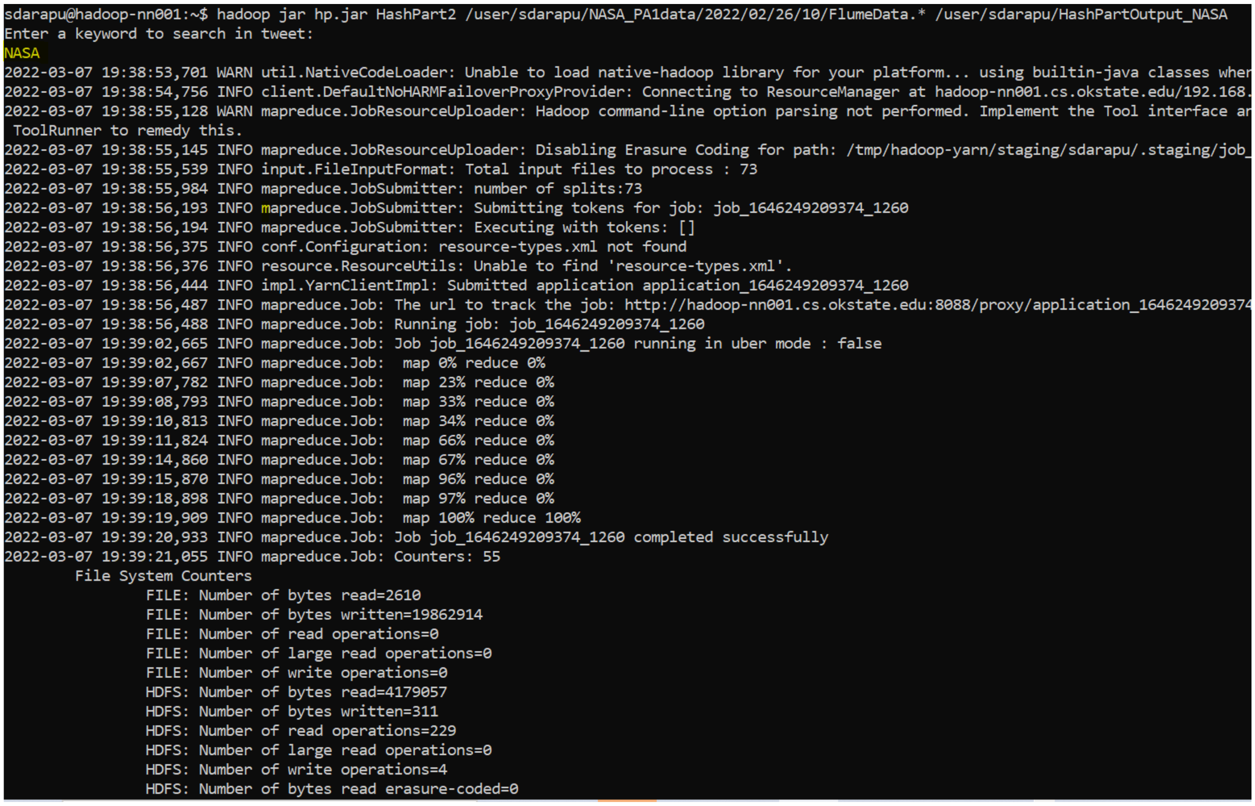
For SpaceX data:



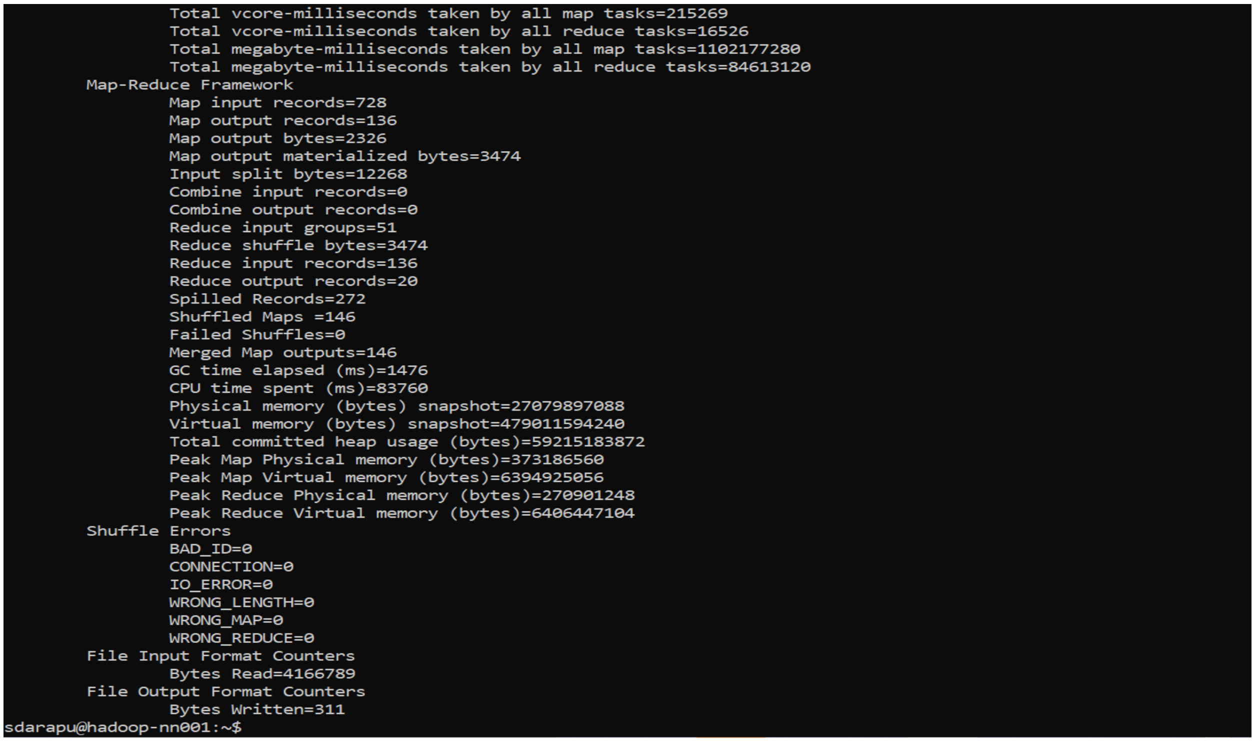
```
sdarapu@hadoop-nn001:~$ hadoop jar hp.jar HashPar
Enter a keyword to search in tweet:
SpaceX
```

6. Now, the program file gets executed.

   For NASA data:

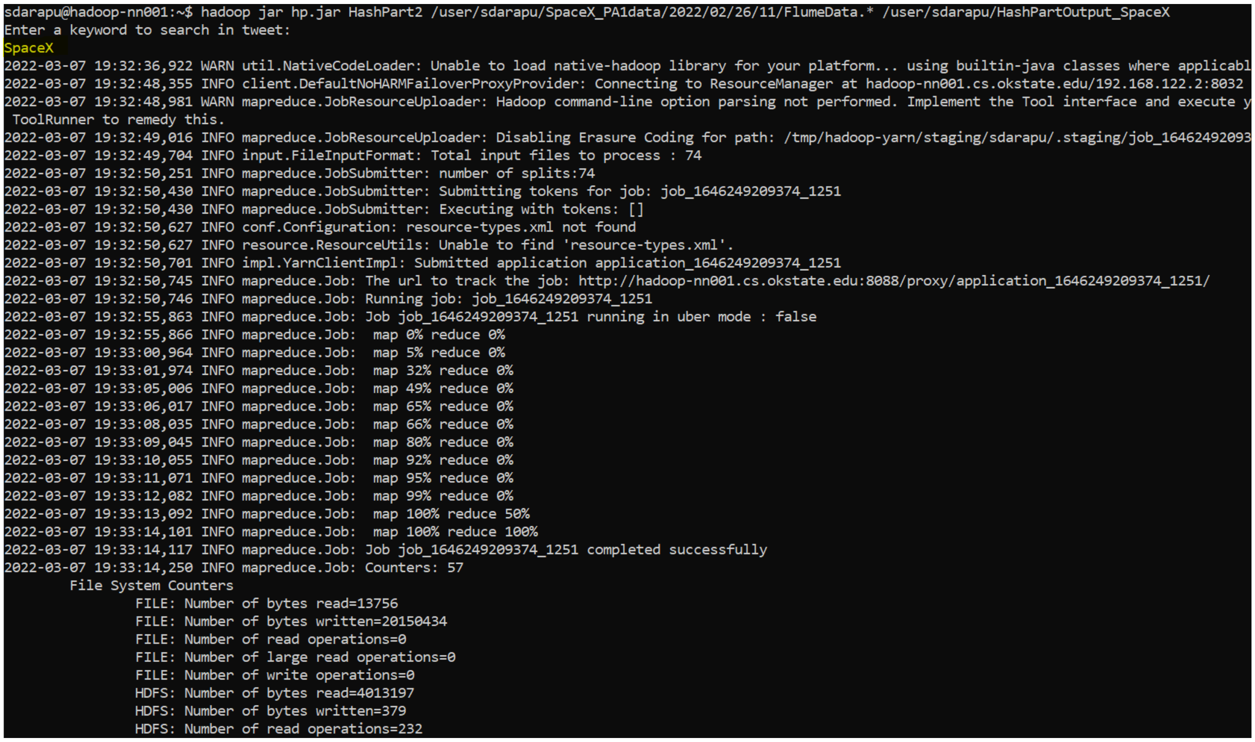


```
sdarapu@hadoop-nn001:~$ hadoop jar hp.jar HashPart2 /user/sdarapu/NASA_PA1data/2022/02/26/10/FlumeData.* /user/sdarapu/HashPartOutput_NASA
Enter a keyword to search in tweet:
NASA
2022-03-07 19:38:53,701 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes wher
2022-03-07 19:38:54,756 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at hadoop-nn001.cs.okstate.edu/192.168.
2022-03-07 19:38:55,128 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface ar
 ToolRunner to remedy this.
2022-03-07 19:38:55,145 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/sdarapu/.staging/job_
2022-03-07 19:38:55,539 INFO input.FileInputFormat: Total input files to process : 73
2022-03-07 19:38:55,984 INFO mapreduce.JobSubmitter: number of splits:73
2022-03-07 19:38:56,193 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1646249209374_1260
2022-03-07 19:38:56,194 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-03-07 19:38:56,375 INFO conf.Configuration: resource-types.xml not found
2022-03-07 19:38:56,376 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-03-07 19:38:56,444 INFO impl.YarnClientImpl: Submitted application application_1646249209374_1260
2022-03-07 19:38:56,487 INFO mapreduce.Job: The url to track the job: http://hadoop-nn001.cs.okstate.edu:8088/proxy/application_1646249209374
2022-03-07 19:38:56,488 INFO mapreduce.Job: Running job: job_1646249209374_1260
2022-03-07 19:39:02,665 INFO mapreduce.Job: Job job_1646249209374_1260 running in uber mode : false
2022-03-07 19:39:02,667 INFO mapreduce.Job:  map 0% reduce 0%
2022-03-07 19:39:07,782 INFO mapreduce.Job:  map 23% reduce 0%
2022-03-07 19:39:08,793 INFO mapreduce.Job:  map 33% reduce 0%
2022-03-07 19:39:10,813 INFO mapreduce.Job:  map 34% reduce 0%
2022-03-07 19:39:11,824 INFO mapreduce.Job:  map 66% reduce 0%
2022-03-07 19:39:14,860 INFO mapreduce.Job:  map 67% reduce 0%
2022-03-07 19:39:15,870 INFO mapreduce.Job:  map 96% reduce 0%
2022-03-07 19:39:18,898 INFO mapreduce.Job:  map 97% reduce 0%
2022-03-07 19:39:19,909 INFO mapreduce.Job:  map 100% reduce 100%
2022-03-07 19:39:20,933 INFO mapreduce.Job: Job job_1646249209374_1260 completed successfully
2022-03-07 19:39:21,055 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=2610
                FILE: Number of bytes written=19862914
                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
                HDFS: Number of bytes written=311
                HDFS: Number of read operations=229
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
                HDFS: Number of bytes read erasure-coded=0
```

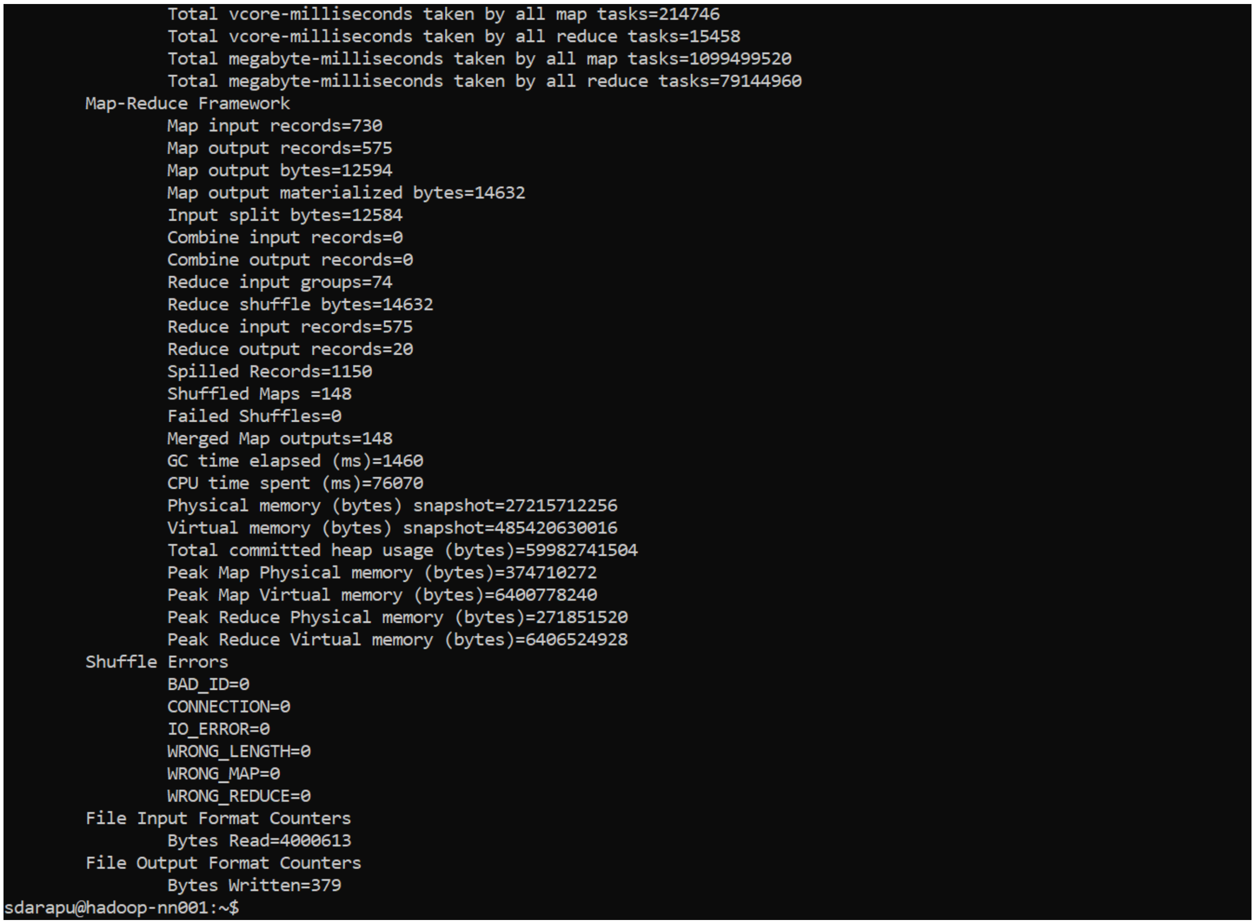


```
                Total vcore-milliseconds taken by all map tasks=215269
                Total vcore-milliseconds taken by all reduce tasks=16526
                Total megabyte-milliseconds taken by all map tasks=1102177280
                Total megabyte-milliseconds taken by all reduce tasks=84613120
        Map-Reduce Framework
                Map input records=728
                Map output records=136
                Map output bytes=2326
                Map output materialized bytes=3474
                Input split bytes=12268
                Combine input records=0
                Combine output records=0
                Reduce input groups=51
                Reduce shuffle bytes=3474
                Reduce input records=136
                Reduce output records=20
                Spilled Records=272
                Shuffled Maps =146
                Failed Shuffles=0
                Merged Map outputs=146
                GC time elapsed (ms)=1476
                CPU time spent (ms)=83760
                Physical memory (bytes) snapshot=27079897088
                Virtual memory (bytes) snapshot=479011594240
                Total committed heap usage (bytes)=59215183872
                Peak Map Physical memory (bytes)=373186560
                Peak Map Virtual memory (bytes)=6394925056
                Peak Reduce Physical memory (bytes)=270901248
                Peak Reduce Virtual memory (bytes)=6406447104
        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=311
sdarapu@hadoop-nn001:~$
```

For SpaceX data:



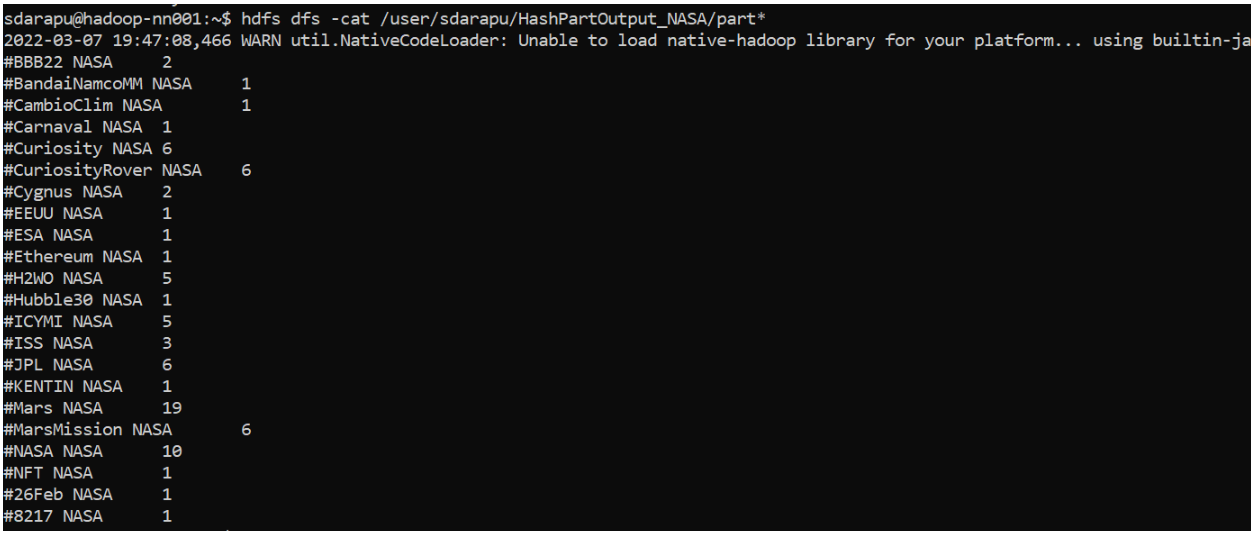
```
sdarapu@hadoop-nn001:~$ hadoop jar hp.jar HashPart2 /user/sdarapu/SpaceX_PA1data/2022/02/26/11/FlumeData.* /user/sdarapu/HashPartOutput_SpaceX
Enter a keyword to search in tweet:
SpaceX
2022-03-07 19:32:36,922 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicabl
2022-03-07 19:32:48,355 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at hadoop-nn001.cs.okstate.edu/192.168.122.2:8032
2022-03-07 19:32:48,981 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute y
 ToolRunner to remedy this.
2022-03-07 19:32:49,016 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/sdarapu/.staging/job_16462492093
2022-03-07 19:32:49,704 INFO input.FileInputFormat: Total input files to process : 74
2022-03-07 19:32:50,251 INFO mapreduce.JobSubmitter: number of splits:74
2022-03-07 19:32:50,430 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1646249209374_1251
2022-03-07 19:32:50,430 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-03-07 19:32:50,627 INFO conf.Configuration: resource-types.xml not found
2022-03-07 19:32:50,627 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-03-07 19:32:50,701 INFO impl.YarnClientImpl: Submitted application application_1646249209374_1251
2022-03-07 19:32:50,745 INFO mapreduce.Job: The url to track the job: http://hadoop-nn001.cs.okstate.edu:8088/proxy/application_1646249209374_1251/
2022-03-07 19:32:50,746 INFO mapreduce.Job: Running job: job_1646249209374_1251
2022-03-07 19:32:55,863 INFO mapreduce.Job: Job job_1646249209374_1251 running in uber mode : false
2022-03-07 19:32:55,866 INFO mapreduce.Job:  map 0% reduce 0%
2022-03-07 19:33:00,964 INFO mapreduce.Job:  map 5% reduce 0%
2022-03-07 19:33:01,974 INFO mapreduce.Job:  map 32% reduce 0%
2022-03-07 19:33:05,006 INFO mapreduce.Job:  map 49% reduce 0%
2022-03-07 19:33:06,017 INFO mapreduce.Job:  map 65% reduce 0%
2022-03-07 19:33:08,035 INFO mapreduce.Job:  map 66% reduce 0%
2022-03-07 19:33:09,045 INFO mapreduce.Job:  map 80% reduce 0%
2022-03-07 19:33:10,055 INFO mapreduce.Job:  map 92% reduce 0%
2022-03-07 19:33:11,071 INFO mapreduce.Job:  map 95% reduce 0%
2022-03-07 19:33:12,082 INFO mapreduce.Job:  map 99% reduce 0%
2022-03-07 19:33:13,092 INFO mapreduce.Job:  map 100% reduce 50%
2022-03-07 19:33:14,101 INFO mapreduce.Job:  map 100% reduce 100%
2022-03-07 19:33:14,117 INFO mapreduce.Job: Job job_1646249209374_1251 completed successfully
2022-03-07 19:33:14,250 INFO mapreduce.Job: Counters: 57
        File System Counters
                FILE: Number of bytes read=13756
                FILE: Number of bytes written=20150434
                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=379
                HDFS: Number of read operations=232
```



```
                Total vcore-milliseconds taken by all map tasks=214746
                Total vcore-milliseconds taken by all reduce tasks=15458
                Total megabyte-milliseconds taken by all map tasks=1099499520
                Total megabyte-milliseconds taken by all reduce tasks=79144960
        Map-Reduce Framework
                Map input records=730
                Map output records=575
                Map output bytes=12594
                Map output materialized bytes=14632
                Input split bytes=12584
                Combine input records=0
                Combine output records=0
                Reduce input groups=74
                Reduce shuffle bytes=14632
                Reduce input records=575
                Reduce output records=20
                Spilled Records=1150
                Shuffled Maps =148
                Failed Shuffles=0
                Merged Map outputs=148
                GC time elapsed (ms)=1460
                CPU time spent (ms)=76070
                Physical memory (bytes) snapshot=27215712256
                Virtual memory (bytes) snapshot=485420630016
                Total committed heap usage (bytes)=59982741504
                Peak Map Physical memory (bytes)=374710272
                Peak Map Virtual memory (bytes)=6400778240
                Peak Reduce Physical memory (bytes)=271851520
                Peak Reduce Virtual memory (bytes)=6406524928
        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=379
sdarapu@hadoop-nn001:~$
```

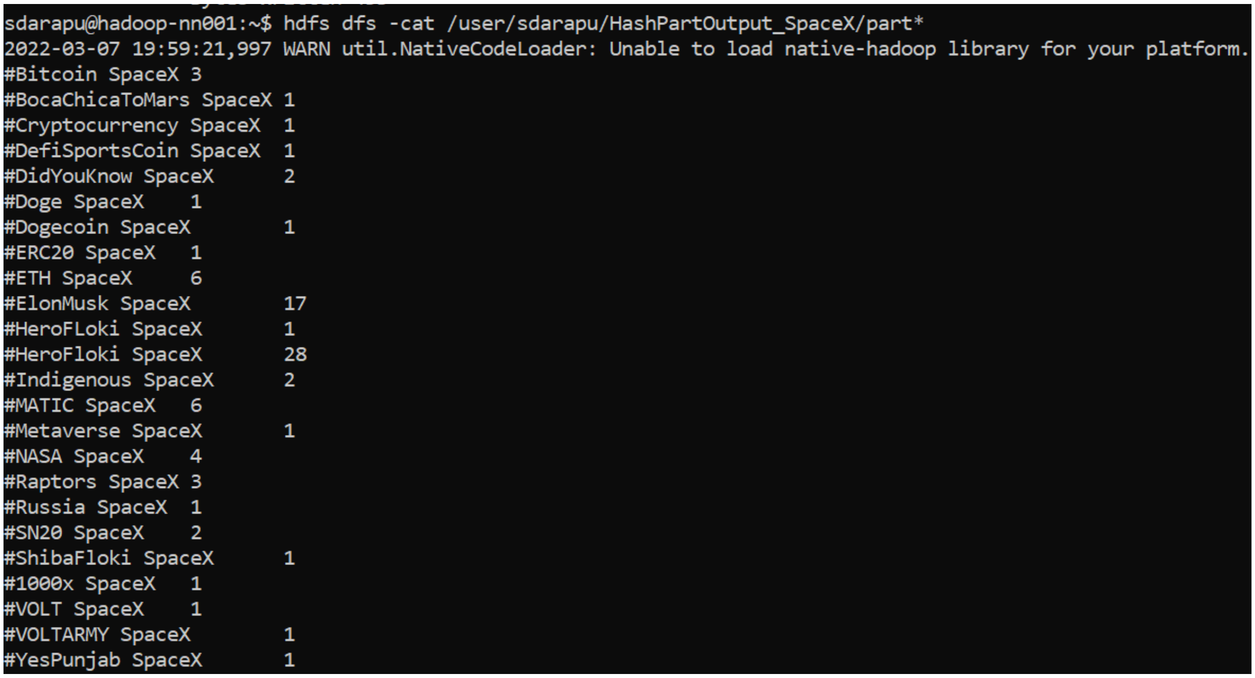
7. Now, to view the output in all the reducers, execute the below command.

   For NASA data:



```
sdarapu@hadoop-nn001:~$ hdfs dfs -cat /user/sdarapu/HashPartOutput_NASA/part*
2022-03-07 19:47:08,466 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
#BBB22 NASA     2
#BandaiNamcoMM NASA     1
#CambioClim NASA        1
#Carnaval NASA  1
#Curiosity NASA 6
#CuriosityRover NASA    6
#Cygnus NASA    2
#EEUU NASA      1
#ESA NASA       1
#Ethereum NASA  1
#H2WO NASA      5
#Hubble30 NASA  1
#ICYMI NASA     5
#ISS NASA       3
#JPL NASA       6
#KENTIN NASA    1
#Mars NASA      19
#MarsMission NASA       6
#NASA NASA      10
#NFT NASA       1
#26Feb NASA     1
#8217 NASA      1
```

   For SpaceX data:

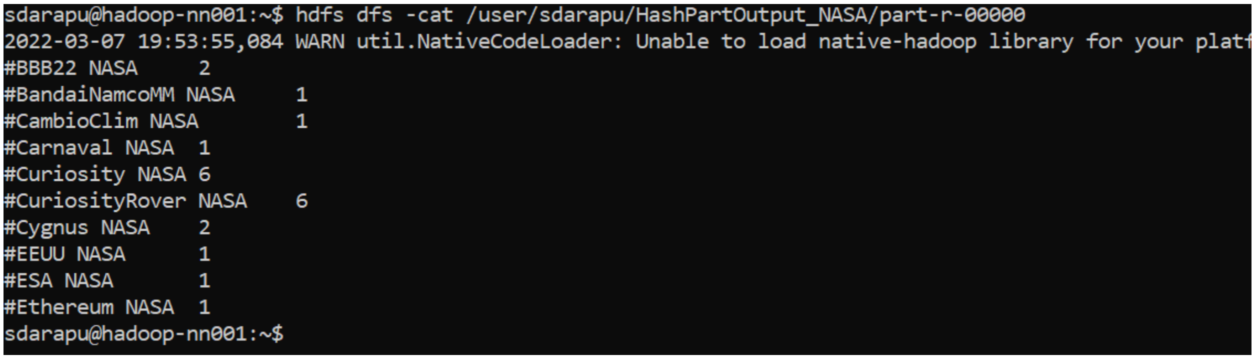


```
sdarapu@hadoop-nn001:~$ hdfs dfs -cat /user/sdarapu/HashPartOutput_SpaceX/part*
2022-03-07 19:59:21,997 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform.
#Bitcoin SpaceX 3
#BocaChicaToMars SpaceX 1
#Cryptocurrency SpaceX  1
#DefiSportsCoin SpaceX  1
#DidYouKnow SpaceX      2
#Doge SpaceX    1
#Dogecoin SpaceX        1
#ERC20 SpaceX   1
#ETH SpaceX     6
#ElonMusk SpaceX        17
#HeroFLoki SpaceX       1
#HeroFloki SpaceX       28
#Indigenous SpaceX      2
#MATIC SpaceX   6
#Metaverse SpaceX       1
#NASA SpaceX    4
#Raptors SpaceX 3
#Russia SpaceX  1
#SN20 SpaceX    2
#ShibaFloki SpaceX      1
#1000x SpaceX   1
#VOLT SpaceX    1
#VOLTARMY SpaceX        1
#YesPunjab SpaceX       1
```

8. To view the data in each reducer individually, execute the command as shown below

   For NASA data:

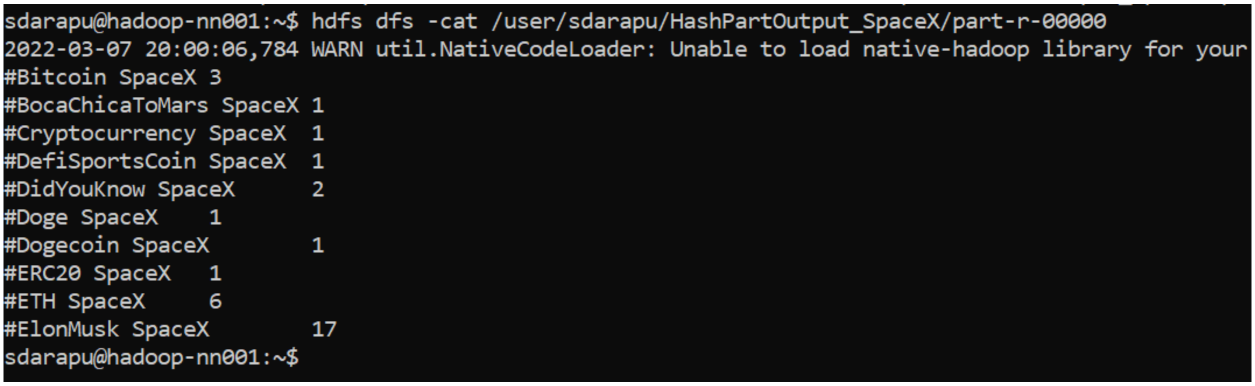
   Example: For reducer 0 (part-r-00000)



```
sdarapu@hadoop-nn001:~$ hdfs dfs -cat /user/sdarapu/HashPartOutput_NASA/part-r-00000
2022-03-07 19:53:55,084 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platf
#BBB22 NASA     2
#BandaiNamcoMM NASA     1
#CambioClim NASA        1
#Carnaval NASA  1
#Curiosity NASA 6
#CuriosityRover NASA    6
#Cygnus NASA    2
#EEUU NASA      1
#ESA NASA       1
#Ethereum NASA  1
sdarapu@hadoop-nn001:~$
```

For SpaceX data:

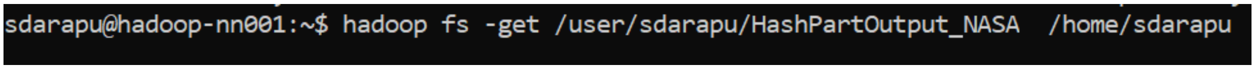
Example: For reducer 0 (part-r-00000)



```
sdarapu@hadoop-nn001:~$ hdfs dfs -cat /user/sdarapu/HashPartOutput_SpaceX/part-r-00000
2022-03-07 20:00:06,784 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
#Bitcoin SpaceX 3
#BocaChicaToMars SpaceX 1
#Cryptocurrency SpaceX  1
#DefiSportsCoin SpaceX  1
#DidYouKnow SpaceX      2
#Doge SpaceX    1
#Dogecoin SpaceX        1
#ERC20 SpaceX   1
#ETH SpaceX     6
#ElonMusk SpaceX        17
sdarapu@hadoop-nn001:~$
```

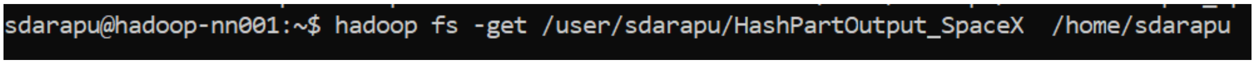
9. To copy the output files to Hadoop local then use the below command

For NASA data:



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

For SpaceX data:



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