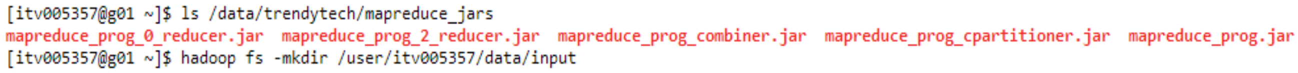
ASSIGNMENT SOLUTION

Q1. Execute all the mapreduce jars as shown in the sessions

1. mapreduce_prog.jar

You can find all the jars in the below mentioned path:

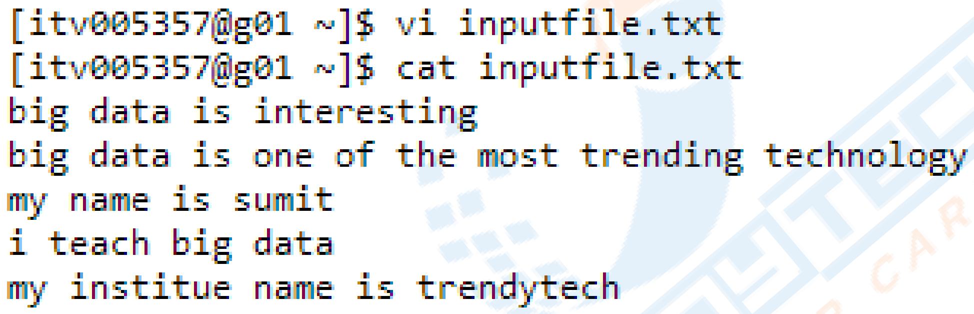
```
ls /data/trendytech/mapreduce_jars

hadoop fs -mkdir /user/itv005357/data/input
```



```
[itv005357@g01 ~]$ ls /data/trendytech/mapreduce_jars
mapreduce_prog_0_reducer.jar  mapreduce_prog_2_reducer.jar  mapreduce_prog_combiner.jar  mapreduce_prog_cpartitioner.jar  mapreduce_prog.jar
[itv005357@g01 ~]$ hadoop fs -mkdir /user/itv005357/data/input
```

Now create an input file using vi editor:

```
vi inputfile.txt
```



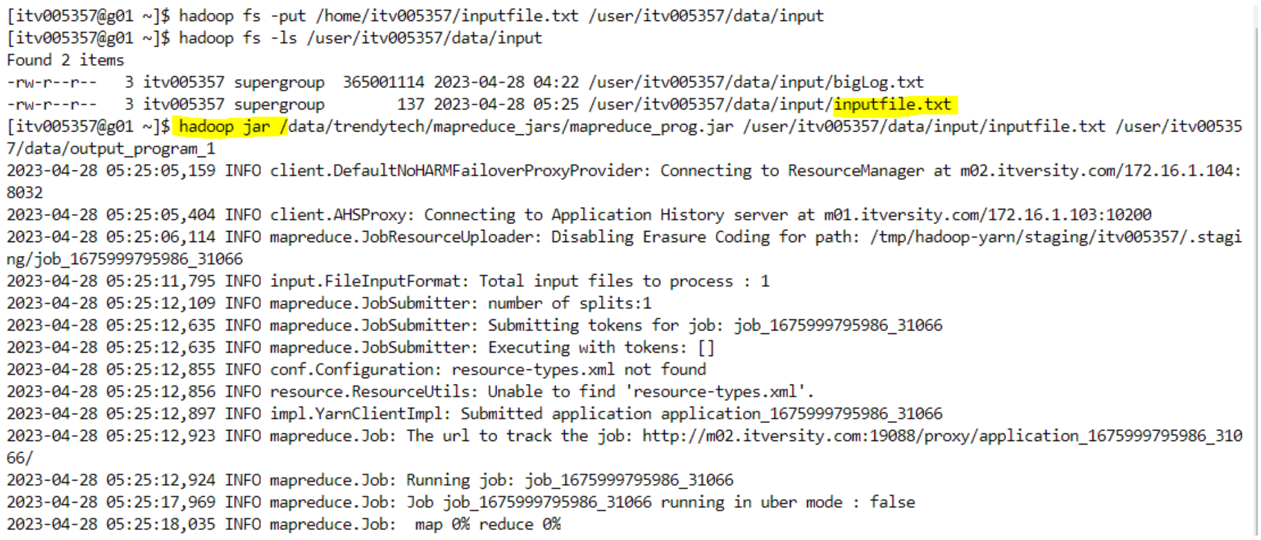
```
[itv005357@g01 ~]$ vi inputfile.txt
[itv005357@g01 ~]$ cat inputfile.txt
big data is interesting
big data is one of the most trending technology
my name is sumit
i teach big data
my institue name is trendytech
```

Now copy the input file from gateway node to HDFS (/data/input)

```
hadoop fs -mkdir /user/itv005357/data/input

hadoop fs -put /home/itv005357/inputfile.txt /user/itv005357/data/input

hadoop fs -ls /user/itv005357/data/input
```

Now Execute the jar file:

```
hadoop jar /data/trendytech/mapreduce_jars/mapreduce_prog.jar
/user/itv005357/data/input/inputfile.txt /user/itv005357/data/output_program_1
```

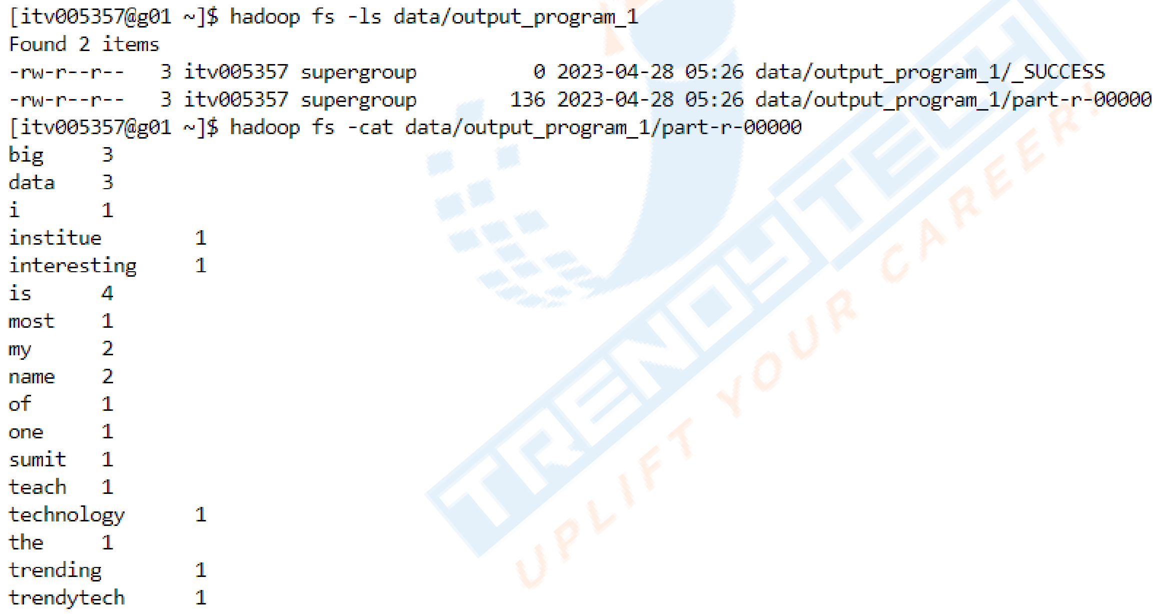


```
[itv005357@g01 ~]$ hadoop fs -put /home/itv005357/inputfile.txt /user/itv005357/data/input
[itv005357@g01 ~]$ hadoop fs -ls /user/itv005357/data/input
Found 2 items
-rw-r--r--   3 itv005357 supergroup  365001114 2023-04-28 04:22 /user/itv005357/data/input/bigLog.txt
-rw-r--r--   3 itv005357 supergroup        137 2023-04-28 05:25 /user/itv005357/data/input/inputfile.txt
[itv005357@g01 ~]$ hadoop jar /data/trendytech/mapreduce_jars/mapreduce_prog.jar /user/itv005357/data/input/inputfile.txt /user/itv00535
7/data/output_program_1
2023-04-28 05:25:05,159 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at m02.itversity.com/172.16.1.104:
8032
2023-04-28 05:25:05,404 INFO client.AHSProxy: Connecting to Application History server at m01.itversity.com/172.16.1.103:10200
2023-04-28 05:25:06,114 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/itv005357/.stagi
ng/job_1675999795986_31066
2023-04-28 05:25:11,795 INFO input.FileInputFormat: Total input files to process : 1
2023-04-28 05:25:12,109 INFO mapreduce.JobSubmitter: number of splits:1
2023-04-28 05:25:12,635 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1675999795986_31066
2023-04-28 05:25:12,635 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-04-28 05:25:12,855 INFO conf.Configuration: resource-types.xml not found
2023-04-28 05:25:12,856 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-04-28 05:25:12,897 INFO impl.YarnClientImpl: Submitted application application_1675999795986_31066
2023-04-28 05:25:12,923 INFO mapreduce.Job: The url to track the job: http://m02.itversity.com:19088/proxy/application_1675999795986_310
66/
2023-04-28 05:25:12,924 INFO mapreduce.Job: Running job: job_1675999795986_31066
2023-04-28 05:25:17,969 INFO mapreduce.Job: Job job_1675999795986_31066 running in uber mode : false
2023-04-28 05:25:18,035 INFO mapreduce.Job:  map 0% reduce 0%
```

hadoop fs -ls data/output_program_1

To see the content of part file:

hadoop fs -cat data/output_program_1/part-r-00000

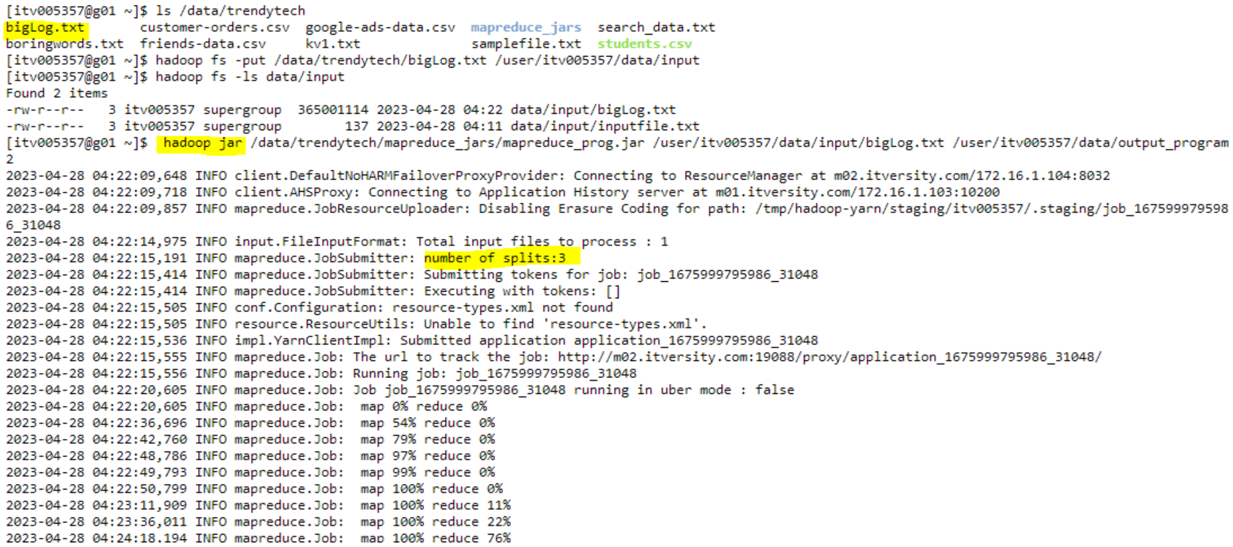


```
[itv005357@g01 ~]$ hadoop fs -ls data/output_program_1
Found 2 items
-rw-r--r--   3 itv005357 supergroup          0 2023-04-28 05:26 data/output_program_1/_SUCCESS
-rw-r--r--   3 itv005357 supergroup        136 2023-04-28 05:26 data/output_program_1/part-r-00000
[itv005357@g01 ~]$ hadoop fs -cat data/output_program_1/part-r-00000
big     3
data    3
i       1
institue        1
interesting     1
is      4
most    1
my      2
name    2
of      1
one     1
sumit   1
teach   1
technology      1
the     1
trending        1
trendytech      1
```

Now we can try the same use case with a big file - bigLog.txt

hadoop fs -put /data/trendytech/bigLog.txt /user/itv005357/data/input

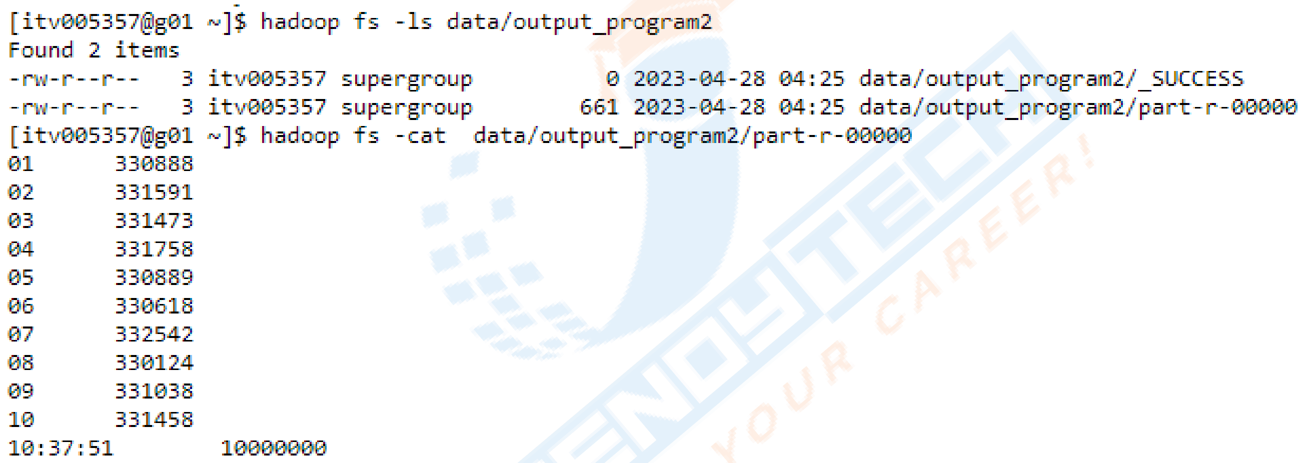
hadoop jar /data/trendytech/mapreduce_jars/mapreduce_prog.jar /user/itv005357/data/input/bigLog.txt /user/itv005357/data/output_program2



```
[itv005357@g01 ~]$ ls /data/trendytech
bigLog.txt        customer-orders.csv  google-ads-data.csv  mapreduce_jars  search_data.txt
boringwords.txt  friends-data.csv     kv1.txt              samplefile.txt  students.csv
[itv005357@g01 ~]$ hadoop fs -put /data/trendytech/bigLog.txt /user/itv005357/data/input
[itv005357@g01 ~]$ hadoop fs -ls data/input
Found 2 items
-rw-r--r--   3 itv005357 supergroup  365001114 2023-04-28 04:22 data/input/bigLog.txt
-rw-r--r--   3 itv005357 supergroup        137 2023-04-28 04:11 data/input/inputfile.txt
[itv005357@g01 ~]$  hadoop jar /data/trendytech/mapreduce_jars/mapreduce_prog.jar /user/itv005357/data/input/bigLog.txt /user/itv005357/data/output_program
2
2023-04-28 04:22:09,648 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at m02.itversity.com/172.16.1.104:8032
2023-04-28 04:22:09,718 INFO client.AHSProxy: Connecting to Application History server at m01.itversity.com/172.16.1.103:10200
2023-04-28 04:22:09,857 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/itv005357/.staging/job_167599979598
6_31048
2023-04-28 04:22:14,975 INFO input.FileInputFormat: Total input files to process : 1
2023-04-28 04:22:15,191 INFO mapreduce.JobSubmitter: number of splits:3
2023-04-28 04:22:15,414 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1675999795986_31048
2023-04-28 04:22:15,414 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-04-28 04:22:15,505 INFO conf.Configuration: resource-types.xml not found
2023-04-28 04:22:15,505 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-04-28 04:22:15,536 INFO impl.YarnClientImpl: Submitted application application_1675999795986_31048
2023-04-28 04:22:15,555 INFO mapreduce.Job: The url to track the job: http://m02.itversity.com:19088/proxy/application_1675999795986_31048/
2023-04-28 04:22:15,556 INFO mapreduce.Job: Running job: job_1675999795986_31048
2023-04-28 04:22:20,605 INFO mapreduce.Job: Job job_1675999795986_31048 running in uber mode : false
2023-04-28 04:22:20,605 INFO mapreduce.Job:  map 0% reduce 0%
2023-04-28 04:22:36,696 INFO mapreduce.Job:  map 54% reduce 0%
2023-04-28 04:22:42,760 INFO mapreduce.Job:  map 79% reduce 0%
2023-04-28 04:22:48,786 INFO mapreduce.Job:  map 97% reduce 0%
2023-04-28 04:22:49,793 INFO mapreduce.Job:  map 99% reduce 0%
2023-04-28 04:22:50,799 INFO mapreduce.Job:  map 100% reduce 0%
2023-04-28 04:23:11,909 INFO mapreduce.Job:  map 100% reduce 11%
2023-04-28 04:23:36,011 INFO mapreduce.Job:  map 100% reduce 22%
2023-04-28 04:24:18.194 INFO mapreduce.Job:  map 100% reduce 76%
```

hadoop fs -ls data/output_program2

hadoop fs -cat data/output_program2/part-r-00000

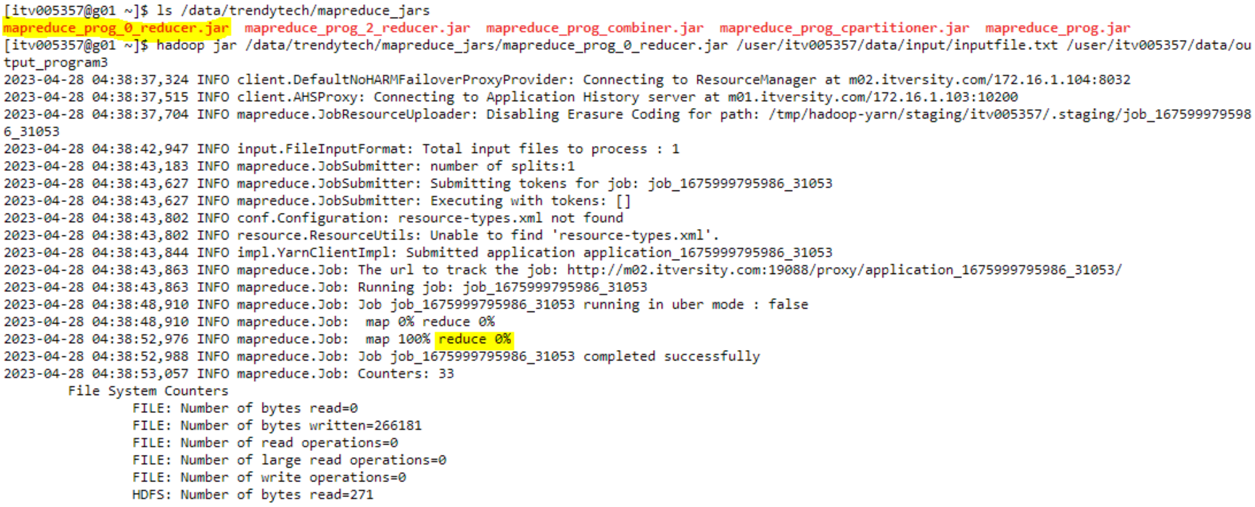


```
[itv005357@g01 ~]$ hadoop fs -ls data/output_program2
Found 2 items
-rw-r--r--   3 itv005357 supergroup          0 2023-04-28 04:25 data/output_program2/_SUCCESS
-rw-r--r--   3 itv005357 supergroup        661 2023-04-28 04:25 data/output_program2/part-r-00000
[itv005357@g01 ~]$ hadoop fs -cat  data/output_program2/part-r-00000
01      330888
02      331591
03      331473
04      331758
05      330889
06      330618
07      332542
08      330124
09      331038
10      331458
10:37:51        10000000
```

2. Now let’s consider the use case with 0 reducer.

mapreduce_prog_0_reducer.jar

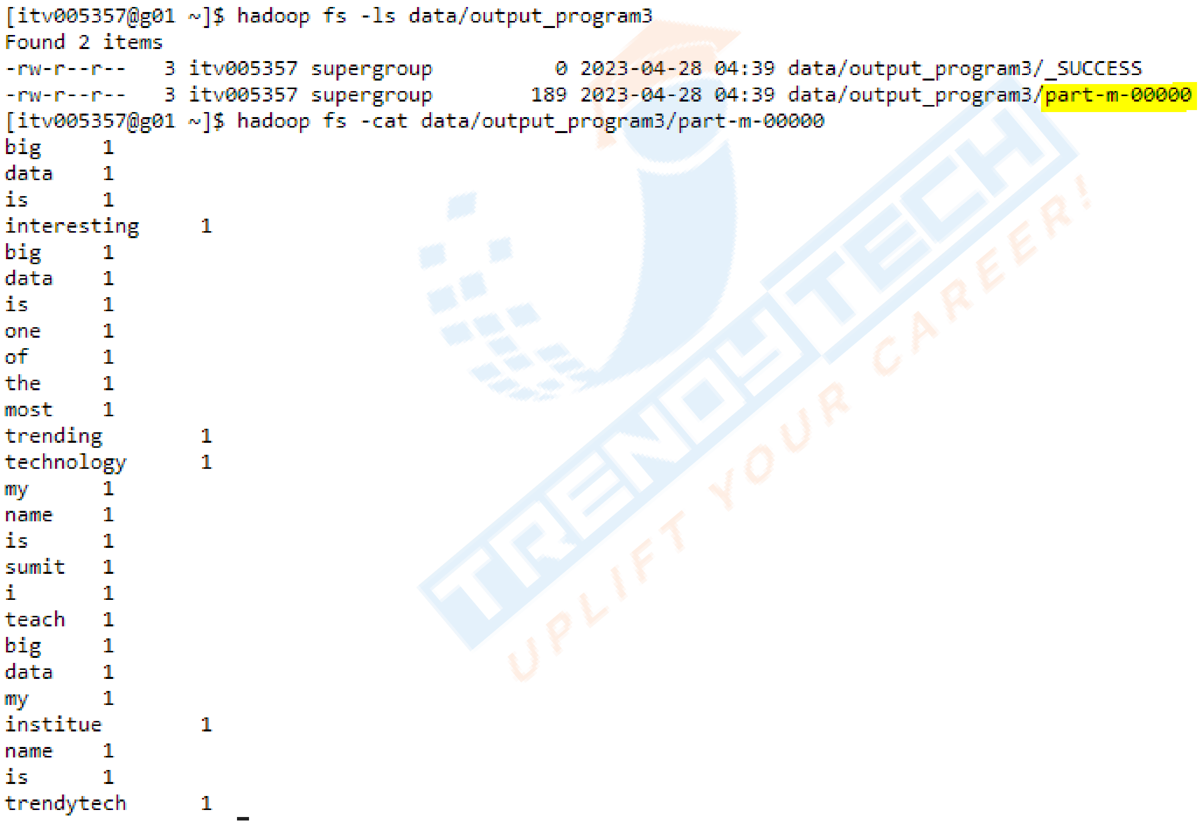
hadoop jar /data/trendytech/mapreduce_jars/mapreduce_prog_0_reducer.jar /user/itv005357/data/input/inputfile.txt /user/itv005357/data/output_program3



```
[itv005357@g01 ~]$ ls /data/trendytech/mapreduce_jars
mapreduce_prog_0_reducer.jar  mapreduce_prog_2_reducer.jar  mapreduce_prog_combiner.jar  mapreduce_prog_cpartitioner.jar  mapreduce_prog.jar
[itv005357@g01 ~]$ hadoop jar /data/trendytech/mapreduce_jars/mapreduce_prog_0_reducer.jar /user/itv005357/data/input/inputfile.txt /user/itv005357/data/ou
tput_program3
2023-04-28 04:38:37,324 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at m02.itversity.com/172.16.1.104:8032
2023-04-28 04:38:37,515 INFO client.AHSProxy: Connecting to Application History server at m01.itversity.com/172.16.1.103:10200
2023-04-28 04:38:37,704 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/itv005357/.staging/job_167599979598
6_31053
2023-04-28 04:38:42,947 INFO input.FileInputFormat: Total input files to process : 1
2023-04-28 04:38:43,183 INFO mapreduce.JobSubmitter: number of splits:1
2023-04-28 04:38:43,627 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1675999795986_31053
2023-04-28 04:38:43,627 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-04-28 04:38:43,802 INFO conf.Configuration: resource-types.xml not found
2023-04-28 04:38:43,802 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-04-28 04:38:43,844 INFO impl.YarnClientImpl: Submitted application application_1675999795986_31053
2023-04-28 04:38:43,863 INFO mapreduce.Job: The url to track the job: http://m02.itversity.com:19088/proxy/application_1675999795986_31053/
2023-04-28 04:38:43,863 INFO mapreduce.Job: Running job: job_1675999795986_31053
2023-04-28 04:38:48,910 INFO mapreduce.Job: Job job_1675999795986_31053 running in uber mode : false
2023-04-28 04:38:48,910 INFO mapreduce.Job:  map 0% reduce 0%
2023-04-28 04:38:52,976 INFO mapreduce.Job:  map 100% reduce 0%
2023-04-28 04:38:52,988 INFO mapreduce.Job: Job job_1675999795986_31053 completed successfully
2023-04-28 04:38:53,057 INFO mapreduce.Job: Counters: 33
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=266181
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=271
```

hadoop fs -ls data/output_program3

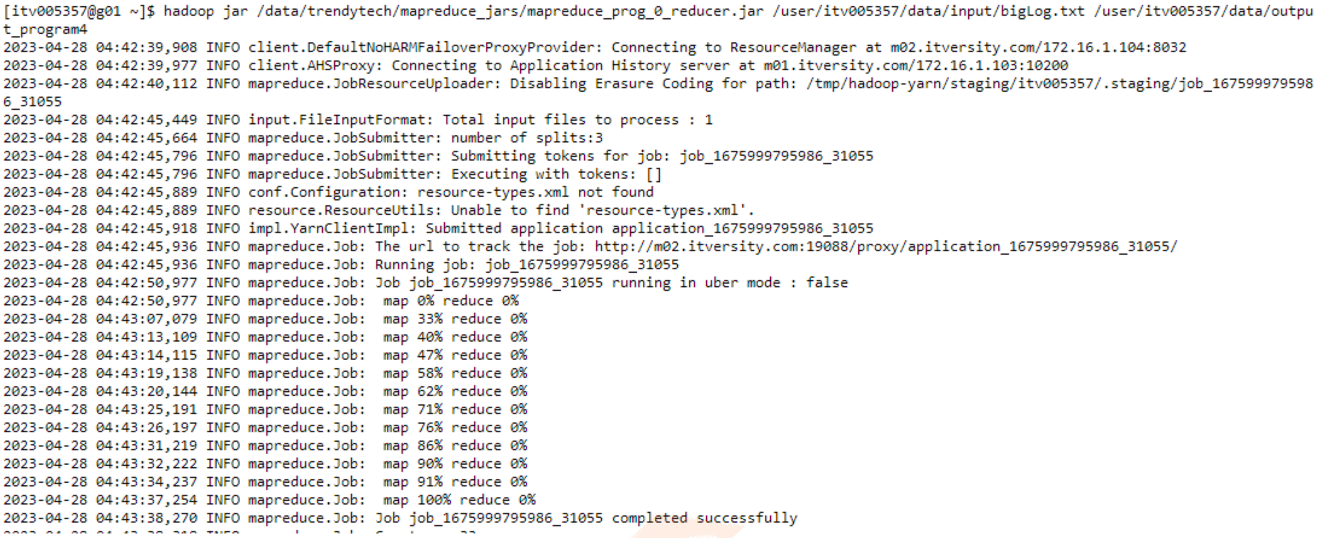
hadoop fs -cat data/output_program3/part-m-00000



```
[itv005357@g01 ~]$ hadoop fs -ls data/output_program3
Found 2 items
-rw-r--r--   3 itv005357 supergroup          0 2023-04-28 04:39 data/output_program3/_SUCCESS
-rw-r--r--   3 itv005357 supergroup        189 2023-04-28 04:39 data/output_program3/part-m-00000
[itv005357@g01 ~]$ hadoop fs -cat data/output_program3/part-m-00000
big     1
data    1
is      1
interesting     1
big     1
data    1
is      1
one     1
of      1
the     1
most    1
trending        1
technology      1
my      1
name    1
is      1
sumit   1
i       1
teach   1
big     1
data    1
my      1
institue        1
name    1
is      1
trendytech      1
```

Now lets consider the same use case with a big file – bigLog.txt

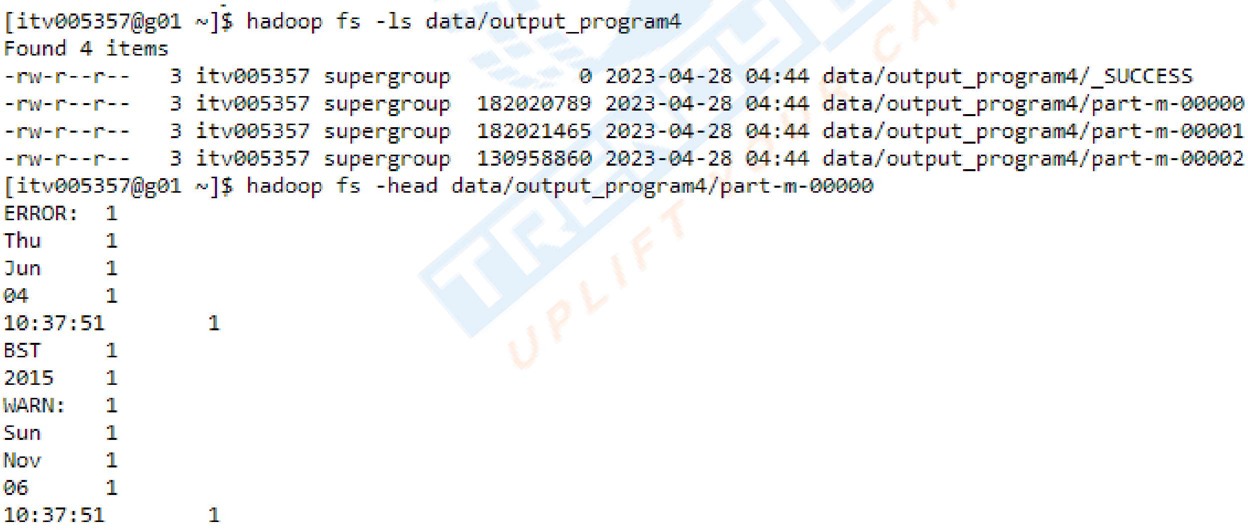
hadoop jar /data/trendytech/mapreduce_jars/mapreduce_prog_0_reducer.jar /user/itv005357/data/input/bigLog.txt /user/itv005357/data/output_program4



```
[itv005357@g01 ~]$ hadoop jar /data/trendytech/mapreduce_jars/mapreduce_prog_0_reducer.jar /user/itv005357/data/input/bigLog.txt /user/itv005357/data/outpu
t_program4
2023-04-28 04:42:39,908 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at m02.itversity.com/172.16.1.104:8032
2023-04-28 04:42:39,977 INFO client.AHSProxy: Connecting to Application History server at m01.itversity.com/172.16.1.103:10200
2023-04-28 04:42:40,112 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/itv005357/.staging/job_167599979598
6_31055
2023-04-28 04:42:45,449 INFO input.FileInputFormat: Total input files to process : 1
2023-04-28 04:42:45,664 INFO mapreduce.JobSubmitter: number of splits:3
2023-04-28 04:42:45,796 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1675999795986_31055
2023-04-28 04:42:45,796 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-04-28 04:42:45,889 INFO conf.Configuration: resource-types.xml not found
2023-04-28 04:42:45,889 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-04-28 04:42:45,918 INFO impl.YarnClientImpl: Submitted application application_1675999795986_31055
2023-04-28 04:42:45,936 INFO mapreduce.Job: The url to track the job: http://m02.itversity.com:19088/proxy/application_1675999795986_31055/
2023-04-28 04:42:45,936 INFO mapreduce.Job: Running job: job_1675999795986_31055
2023-04-28 04:42:50,977 INFO mapreduce.Job: Job job_1675999795986_31055 running in uber mode : false
2023-04-28 04:42:50,977 INFO mapreduce.Job:  map 0% reduce 0%
2023-04-28 04:43:07,079 INFO mapreduce.Job:  map 33% reduce 0%
2023-04-28 04:43:13,109 INFO mapreduce.Job:  map 40% reduce 0%
2023-04-28 04:43:14,115 INFO mapreduce.Job:  map 47% reduce 0%
2023-04-28 04:43:19,138 INFO mapreduce.Job:  map 58% reduce 0%
2023-04-28 04:43:20,144 INFO mapreduce.Job:  map 62% reduce 0%
2023-04-28 04:43:25,191 INFO mapreduce.Job:  map 71% reduce 0%
2023-04-28 04:43:26,197 INFO mapreduce.Job:  map 76% reduce 0%
2023-04-28 04:43:31,219 INFO mapreduce.Job:  map 86% reduce 0%
2023-04-28 04:43:32,222 INFO mapreduce.Job:  map 90% reduce 0%
2023-04-28 04:43:34,237 INFO mapreduce.Job:  map 91% reduce 0%
2023-04-28 04:43:37,254 INFO mapreduce.Job:  map 100% reduce 0%
2023-04-28 04:43:38,270 INFO mapreduce.Job: Job job_1675999795986_31055 completed successfully
```

hadoop fs -ls data/output_program4

hadoop fs -head data/output_program4/part-m-00000

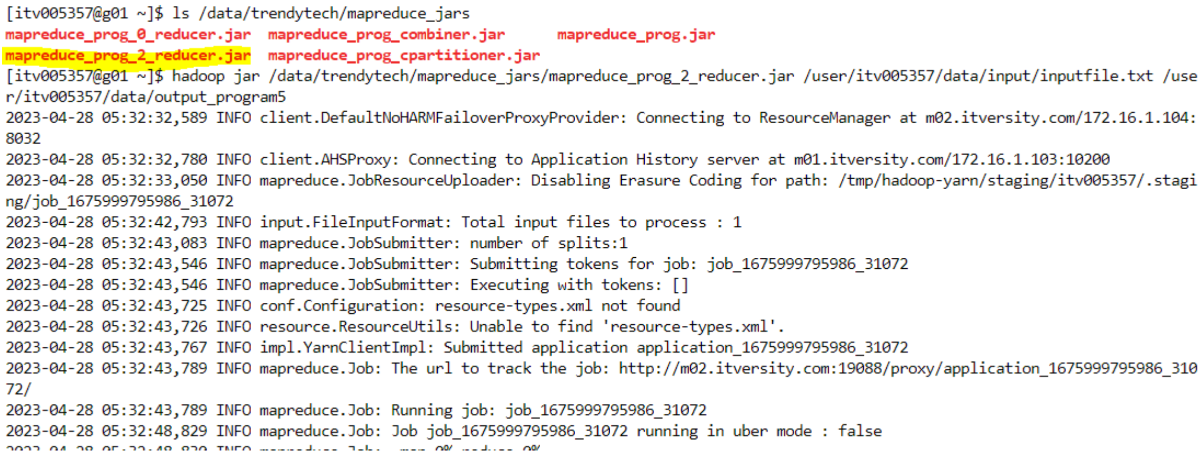


```
[itv005357@g01 ~]$ hadoop fs -ls data/output_program4
Found 4 items
-rw-r--r--   3 itv005357 supergroup          0 2023-04-28 04:44 data/output_program4/_SUCCESS
-rw-r--r--   3 itv005357 supergroup  182020789 2023-04-28 04:44 data/output_program4/part-m-00000
-rw-r--r--   3 itv005357 supergroup  182021465 2023-04-28 04:44 data/output_program4/part-m-00001
-rw-r--r--   3 itv005357 supergroup  130958860 2023-04-28 04:44 data/output_program4/part-m-00002
[itv005357@g01 ~]$ hadoop fs -head data/output_program4/part-m-00000
ERROR:  1
Thu     1
Jun     1
04      1
10:37:51        1
BST     1
2015    1
WARN:   1
Sun     1
Nov     1
06      1
10:37:51        1
```

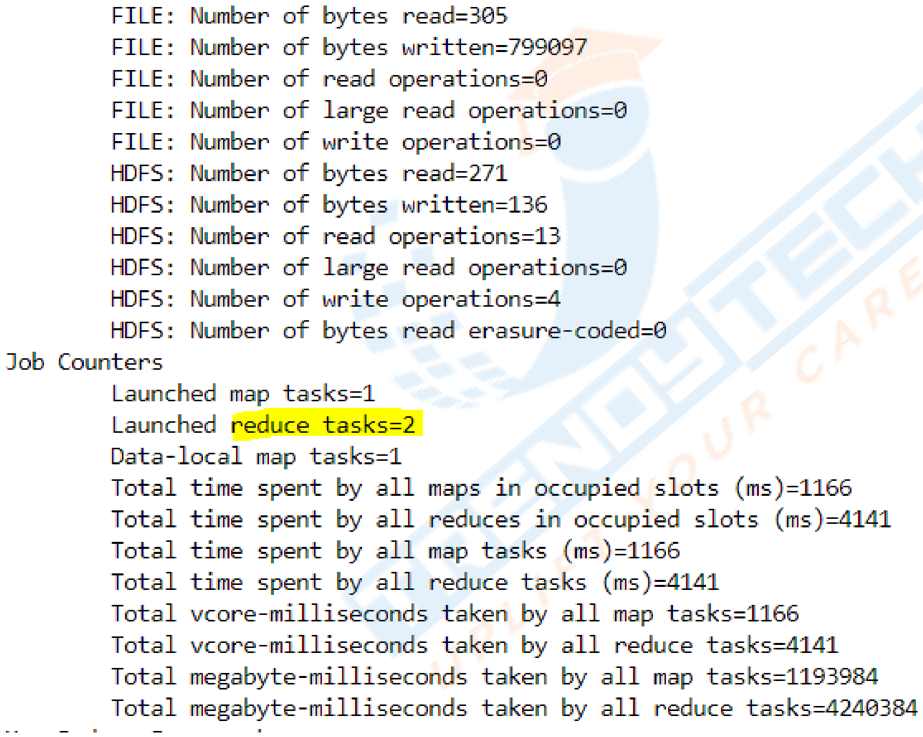
3. Now let’s consider a use case with 2 reducers.

mapreduce_prog_2_reducer.jar

hadoop jar /data/trendytech/mapreduce_jars/mapreduce_prog_2_reducer.jar /user/itv005357/data/input/inputfile.txt /user/itv005357/data/output_program5



```
[itv005357@g01 ~]$ ls /data/trendytech/mapreduce_jars
mapreduce_prog_0_reducer.jar  mapreduce_prog_combiner.jar      mapreduce_prog.jar
mapreduce_prog_2_reducer.jar  mapreduce_prog_cpartitioner.jar
[itv005357@g01 ~]$ hadoop jar /data/trendytech/mapreduce_jars/mapreduce_prog_2_reducer.jar /user/itv005357/data/input/inputfile.txt /use
r/itv005357/data/output_program5
2023-04-28 05:32:32,589 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at m02.itversity.com/172.16.1.104:
8032
2023-04-28 05:32:32,780 INFO client.AHSProxy: Connecting to Application History server at m01.itversity.com/172.16.1.103:10200
2023-04-28 05:32:33,050 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/itv005357/.stagi
ng/job_1675999795986_31072
2023-04-28 05:32:42,793 INFO input.FileInputFormat: Total input files to process : 1
2023-04-28 05:32:43,083 INFO mapreduce.JobSubmitter: number of splits:1
2023-04-28 05:32:43,546 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1675999795986_31072
2023-04-28 05:32:43,546 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-04-28 05:32:43,725 INFO conf.Configuration: resource-types.xml not found
2023-04-28 05:32:43,726 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-04-28 05:32:43,767 INFO impl.YarnClientImpl: Submitted application application_1675999795986_31072
2023-04-28 05:32:43,789 INFO mapreduce.Job: The url to track the job: http://m02.itversity.com:19088/proxy/application_1675999795986_310
72/
2023-04-28 05:32:43,789 INFO mapreduce.Job: Running job: job_1675999795986_31072
2023-04-28 05:32:48,829 INFO mapreduce.Job: Job job_1675999795986_31072 running in uber mode : false
```

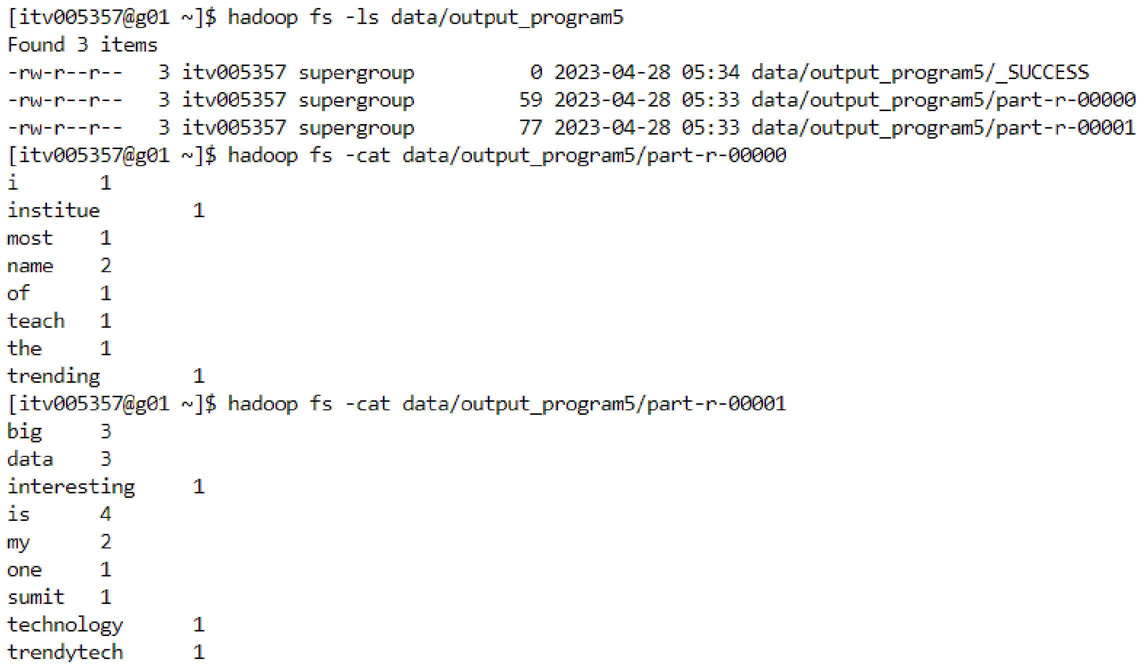


```
                FILE: Number of bytes read=305
                FILE: Number of bytes written=799097
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=271
                HDFS: Number of bytes written=136
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=2
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=1166
                Total time spent by all reduces in occupied slots (ms)=4141
                Total time spent by all map tasks (ms)=1166
                Total time spent by all reduce tasks (ms)=4141
                Total vcore-milliseconds taken by all map tasks=1166
                Total vcore-milliseconds taken by all reduce tasks=4141
                Total megabyte-milliseconds taken by all map tasks=1193984
                Total megabyte-milliseconds taken by all reduce tasks=4240384
```

hadoop fs -ls data/output_program5

hadoop fs -cat data/output_program5/part-r-00000

hadoop fs -cat data/output_program5/part-r-00001



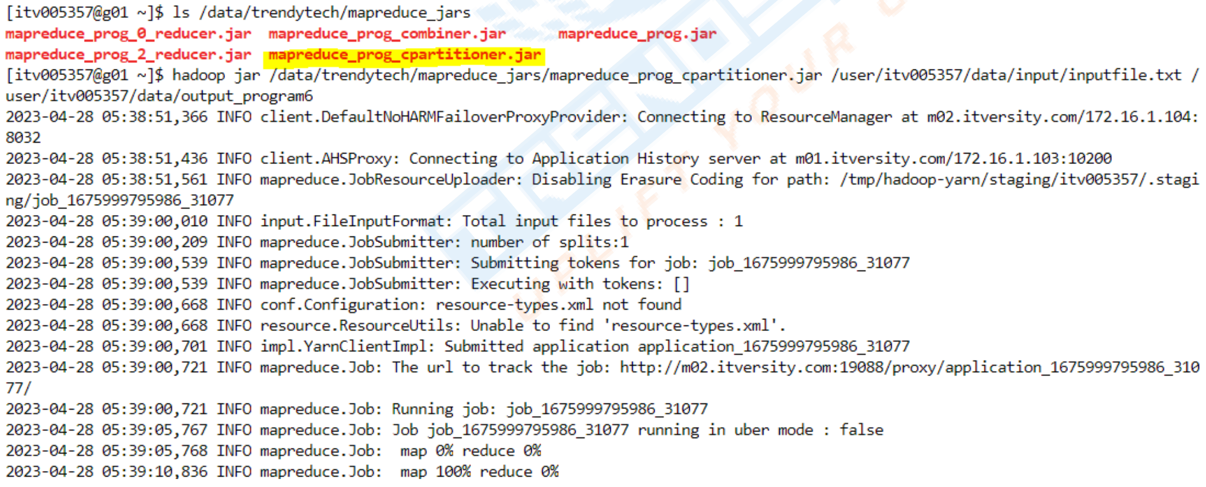
```
[itv005357@g01 ~]$ hadoop fs -ls data/output_program5
Found 3 items
-rw-r--r--   3 itv005357 supergroup          0 2023-04-28 05:34 data/output_program5/_SUCCESS
-rw-r--r--   3 itv005357 supergroup         59 2023-04-28 05:33 data/output_program5/part-r-00000
-rw-r--r--   3 itv005357 supergroup         77 2023-04-28 05:33 data/output_program5/part-r-00001
[itv005357@g01 ~]$ hadoop fs -cat data/output_program5/part-r-00000
i       1
institue        1
most    1
name    2
of      1
teach   1
the     1
trending        1
[itv005357@g01 ~]$ hadoop fs -cat data/output_program5/part-r-00001
big     3
data    3
interesting     1
is      4
my      2
one     1
sumit   1
technology      1
trendytech      1
```

4. Now let’s consider a use case with a custom partitioner.

Partitioner logic : If the word length is less than or equal to 3 then it should go to reducer0 and all the other words should go to the second reducer- reducer1

mapreduce_prog_cpartitioner.jar

hadoop jar /data/trendytech/mapreduce_jars/mapreduce_prog_cpartitioner.jar /user/itv005357/data/input/inputfile.txt /user/itv005357/data/output_program6

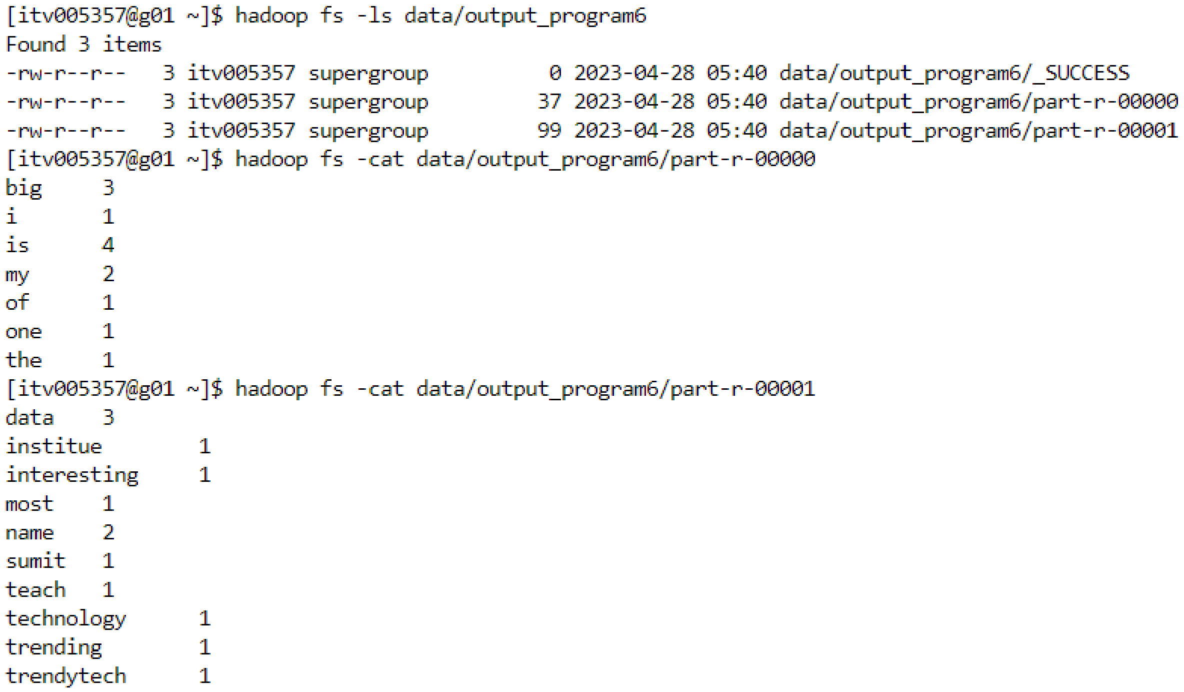


```
[itv005357@g01 ~]$ ls /data/trendytech/mapreduce_jars
mapreduce_prog_0_reducer.jar  mapreduce_prog_combiner.jar      mapreduce_prog.jar
mapreduce_prog_2_reducer.jar  mapreduce_prog_cpartitioner.jar
[itv005357@g01 ~]$ hadoop jar /data/trendytech/mapreduce_jars/mapreduce_prog_cpartitioner.jar /user/itv005357/data/input/inputfile.txt /
user/itv005357/data/output_program6
2023-04-28 05:38:51,366 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at m02.itversity.com/172.16.1.104:
8032
2023-04-28 05:38:51,436 INFO client.AHSProxy: Connecting to Application History server at m01.itversity.com/172.16.1.103:10200
2023-04-28 05:38:51,561 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/itv005357/.stagi
ng/job_1675999795986_31077
2023-04-28 05:39:00,010 INFO input.FileInputFormat: Total input files to process : 1
2023-04-28 05:39:00,209 INFO mapreduce.JobSubmitter: number of splits:1
2023-04-28 05:39:00,539 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1675999795986_31077
2023-04-28 05:39:00,539 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-04-28 05:39:00,668 INFO conf.Configuration: resource-types.xml not found
2023-04-28 05:39:00,668 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-04-28 05:39:00,701 INFO impl.YarnClientImpl: Submitted application application_1675999795986_31077
2023-04-28 05:39:00,721 INFO mapreduce.Job: The url to track the job: http://m02.itversity.com:19088/proxy/application_1675999795986_310
77/
2023-04-28 05:39:00,721 INFO mapreduce.Job: Running job: job_1675999795986_31077
2023-04-28 05:39:05,767 INFO mapreduce.Job: Job job_1675999795986_31077 running in uber mode : false
2023-04-28 05:39:05,768 INFO mapreduce.Job:  map 0% reduce 0%
2023-04-28 05:39:10,836 INFO mapreduce.Job:  map 100% reduce 0%
```

hadoop fs -ls data/output_program6

hadoop fs -cat data/output_program6/part-r-00000

hadoop fs -cat data/output_program6/part-r-00001

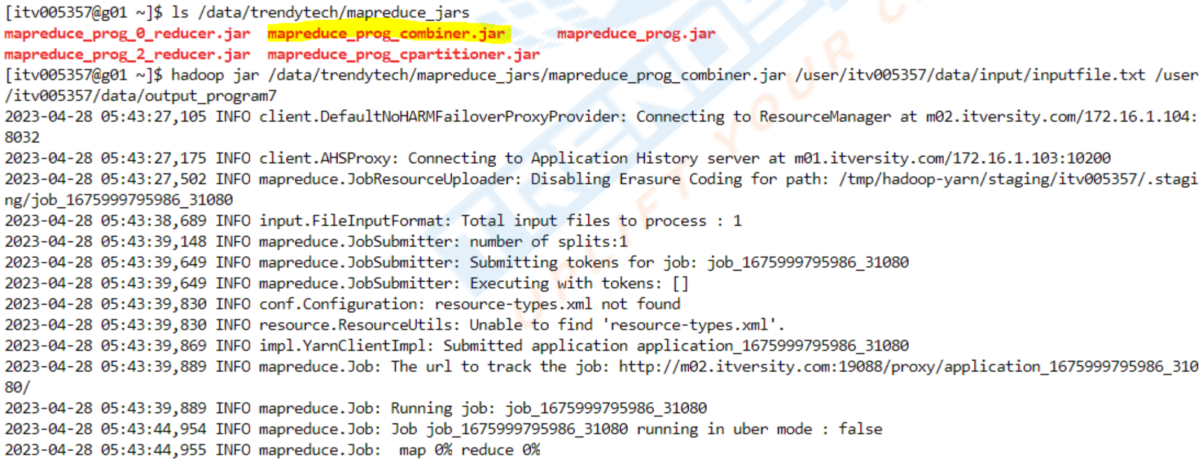


```
[itv005357@g01 ~]$ hadoop fs -ls data/output_program6
Found 3 items
-rw-r--r--   3 itv005357 supergroup          0 2023-04-28 05:40 data/output_program6/_SUCCESS
-rw-r--r--   3 itv005357 supergroup         37 2023-04-28 05:40 data/output_program6/part-r-00000
-rw-r--r--   3 itv005357 supergroup         99 2023-04-28 05:40 data/output_program6/part-r-00001
[itv005357@g01 ~]$ hadoop fs -cat data/output_program6/part-r-00000
big     3
i       1
is      4
my      2
of      1
one     1
the     1
[itv005357@g01 ~]$ hadoop fs -cat data/output_program6/part-r-00001
data    3
institue        1
interesting     1
most    1
name    2
sumit   1
teach   1
technology      1
trending        1
trendytech      1
```

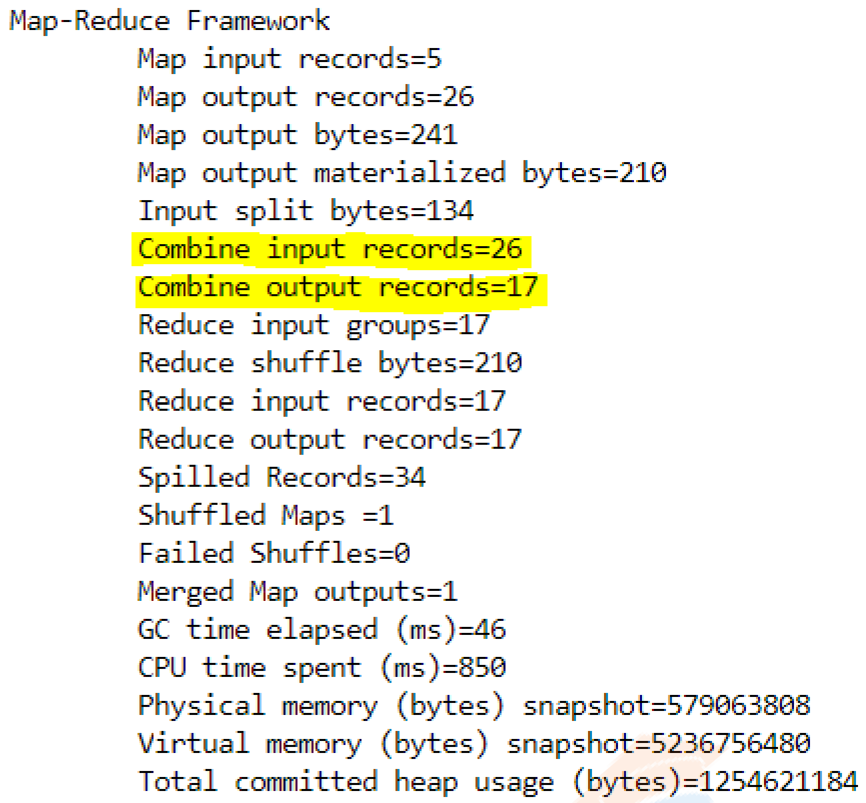
5. Now let’s consider a use case with combiner.

mapreduce_prog_combiner.jar

hadoop jar /data/trendytech/mapreduce_jars/mapreduce_prog_combiner.jar /user/itv005357/data/input/inputfile.txt /user/itv005357/data/output_program7



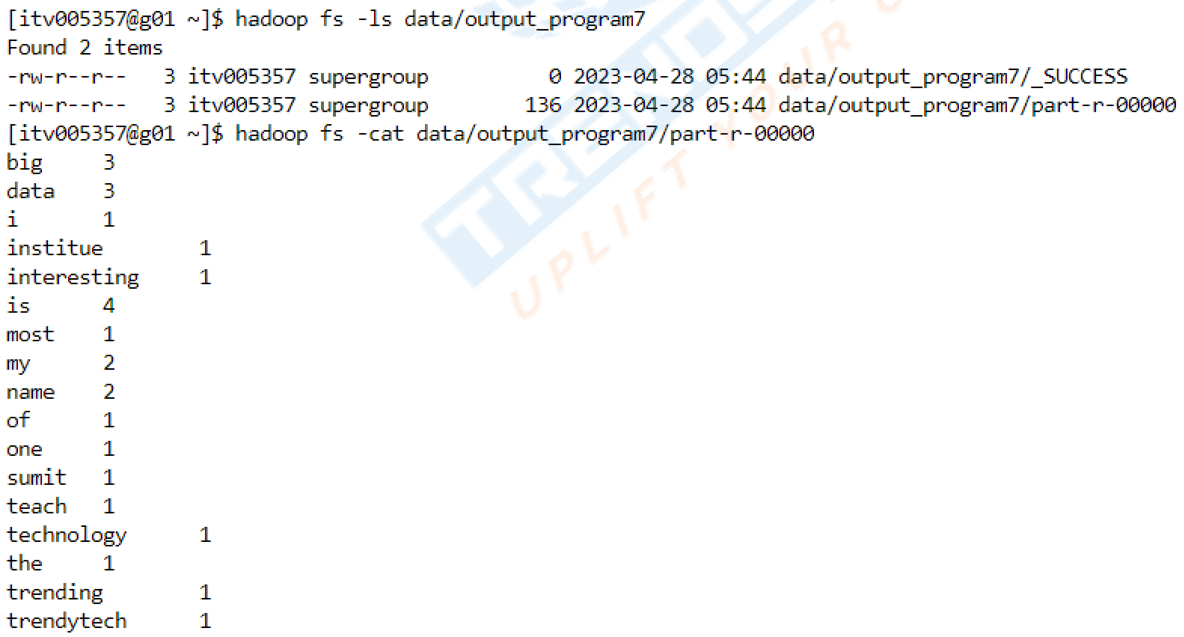
```
[itv005357@g01 ~]$ ls /data/trendytech/mapreduce_jars
mapreduce_prog_0_reducer.jar  mapreduce_prog_combiner.jar      mapreduce_prog.jar
mapreduce_prog_2_reducer.jar  mapreduce_prog_cpartitioner.jar
[itv005357@g01 ~]$ hadoop jar /data/trendytech/mapreduce_jars/mapreduce_prog_combiner.jar /user/itv005357/data/input/inputfile.txt /user
/itv005357/data/output_program7
2023-04-28 05:43:27,105 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at m02.itversity.com/172.16.1.104:
8032
2023-04-28 05:43:27,175 INFO client.AHSProxy: Connecting to Application History server at m01.itversity.com/172.16.1.103:10200
2023-04-28 05:43:27,502 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/itv005357/.stagi
ng/job_1675999795986_31080
2023-04-28 05:43:38,689 INFO input.FileInputFormat: Total input files to process : 1
2023-04-28 05:43:39,148 INFO mapreduce.JobSubmitter: number of splits:1
2023-04-28 05:43:39,649 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1675999795986_31080
2023-04-28 05:43:39,649 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-04-28 05:43:39,830 INFO conf.Configuration: resource-types.xml not found
2023-04-28 05:43:39,830 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-04-28 05:43:39,869 INFO impl.YarnClientImpl: Submitted application application_1675999795986_31080
2023-04-28 05:43:39,889 INFO mapreduce.Job: The url to track the job: http://m02.itversity.com:19088/proxy/application_1675999795986_310
80/
2023-04-28 05:43:39,889 INFO mapreduce.Job: Running job: job_1675999795986_31080
2023-04-28 05:43:44,954 INFO mapreduce.Job: Job job_1675999795986_31080 running in uber mode : false
2023-04-28 05:43:44,955 INFO mapreduce.Job:  map 0% reduce 0%
```



```
Map-Reduce Framework
        Map input records=5
        Map output records=26
        Map output bytes=241
        Map output materialized bytes=210
        Input split bytes=134
        Combine input records=26
        Combine output records=17
        Reduce input groups=17
        Reduce shuffle bytes=210
        Reduce input records=17
        Reduce output records=17
        Spilled Records=34
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=46
        CPU time spent (ms)=850
        Physical memory (bytes) snapshot=579063808
        Virtual memory (bytes) snapshot=5236756480
        Total committed heap usage (bytes)=1254621184
```

hadoop fs -ls data/output_program7

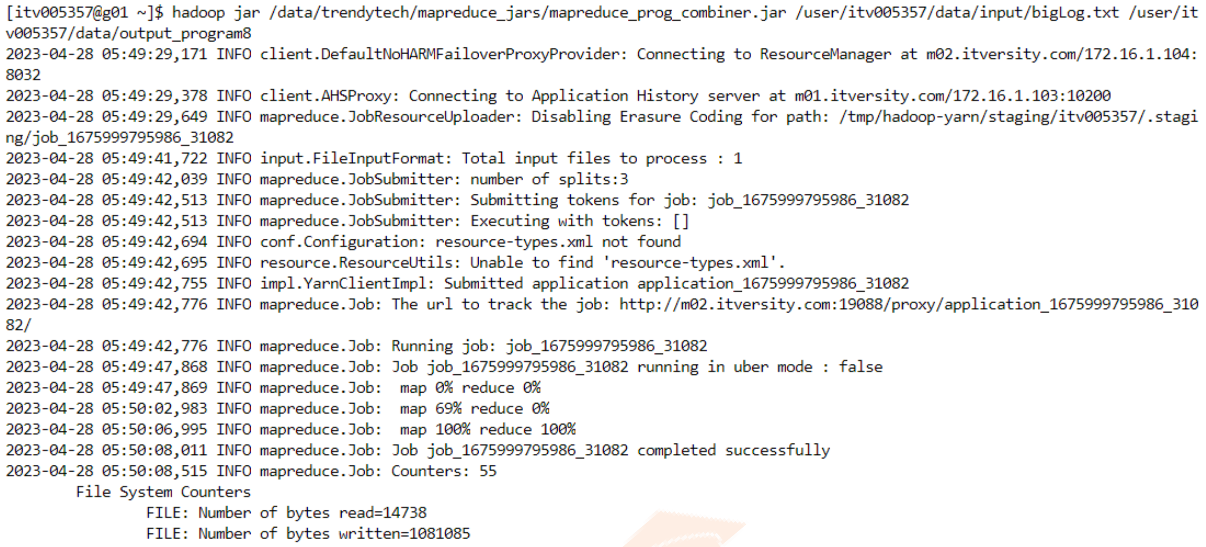
hadoop fs -cat data/output_program7/part-r-00000



```
[itv005357@g01 ~]$ hadoop fs -ls data/output_program7
Found 2 items
-rw-r--r--   3 itv005357 supergroup          0 2023-04-28 05:44 data/output_program7/_SUCCESS
-rw-r--r--   3 itv005357 supergroup        136 2023-04-28 05:44 data/output_program7/part-r-00000
[itv005357@g01 ~]$ hadoop fs -cat data/output_program7/part-r-00000
big     3
data    3
i       1
institue        1
interesting     1
is      4
most    1
my      2
name    2
of      1
one     1
sumit   1
teach   1
technology      1
the     1
trending        1
trendytech      1
```

Now we can use the same jar file with a big file – bigLog.txt

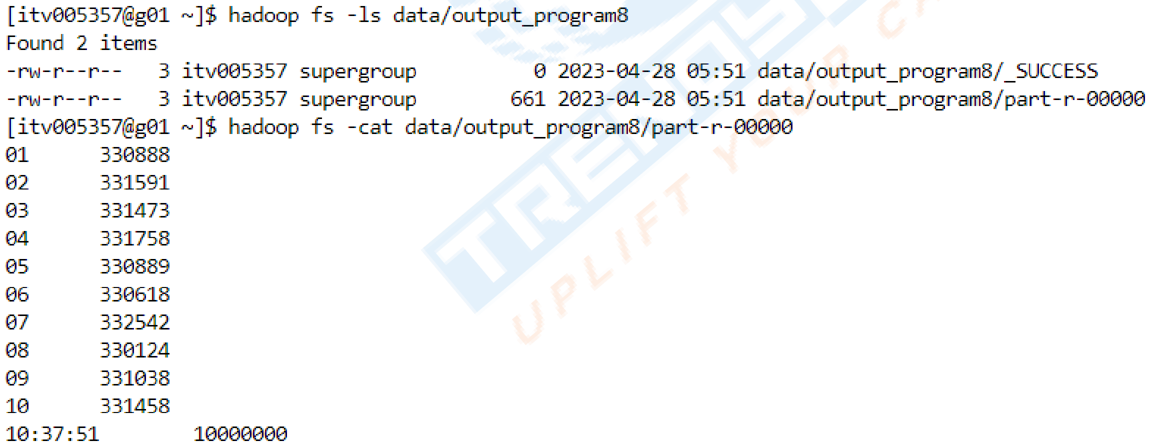
hadoop jar /data/trendytech/mapreduce_jars/mapreduce_prog_combiner.jar /user/itv005357/data/input/bigLog.txt /user/itv005357/data/output_program8



```
[itv005357@g01 ~]$ hadoop jar /data/trendytech/mapreduce_jars/mapreduce_prog_combiner.jar /user/itv005357/data/input/bigLog.txt /user/it
v005357/data/output_program8
2023-04-28 05:49:29,171 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at m02.itversity.com/172.16.1.104:
8032
2023-04-28 05:49:29,378 INFO client.AHSProxy: Connecting to Application History server at m01.itversity.com/172.16.1.103:10200
2023-04-28 05:49:29,649 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/itv005357/.stagi
ng/job_1675999795986_31082
2023-04-28 05:49:41,722 INFO input.FileInputFormat: Total input files to process : 1
2023-04-28 05:49:42,039 INFO mapreduce.JobSubmitter: number of splits:3
2023-04-28 05:49:42,513 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1675999795986_31082
2023-04-28 05:49:42,513 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-04-28 05:49:42,694 INFO conf.Configuration: resource-types.xml not found
2023-04-28 05:49:42,695 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-04-28 05:49:42,755 INFO impl.YarnClientImpl: Submitted application application_1675999795986_31082
2023-04-28 05:49:42,776 INFO mapreduce.Job: The url to track the job: http://m02.itversity.com:19088/proxy/application_1675999795986_310
82/
2023-04-28 05:49:42,776 INFO mapreduce.Job: Running job: job_1675999795986_31082
2023-04-28 05:49:47,868 INFO mapreduce.Job: Job job_1675999795986_31082 running in uber mode : false
2023-04-28 05:49:47,869 INFO mapreduce.Job:  map 0% reduce 0%
2023-04-28 05:50:02,983 INFO mapreduce.Job:  map 69% reduce 0%
2023-04-28 05:50:06,995 INFO mapreduce.Job:  map 100% reduce 100%
2023-04-28 05:50:08,011 INFO mapreduce.Job: Job job_1675999795986_31082 completed successfully
2023-04-28 05:50:08,515 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=14738
                FILE: Number of bytes written=1081085
```

hadoop fs -ls data/output_program8

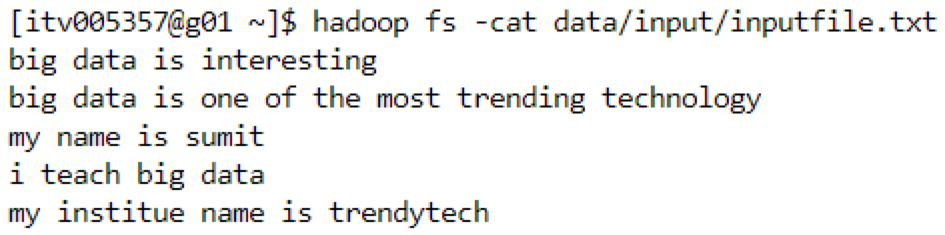
hadoop fs -cat data/output_program8/part-r-00000



```
[itv005357@g01 ~]$ hadoop fs -ls data/output_program8
Found 2 items
-rw-r--r--   3 itv005357 supergroup          0 2023-04-28 05:51 data/output_program8/_SUCCESS
-rw-r--r--   3 itv005357 supergroup        661 2023-04-28 05:51 data/output_program8/part-r-00000
[itv005357@g01 ~]$ hadoop fs -cat data/output_program8/part-r-00000
01      330888
02      331591
03      331473
04      331758
05      330889
06      330618
07      332542
08      330124
09      331038
10      331458
10:37:51        10000000
```

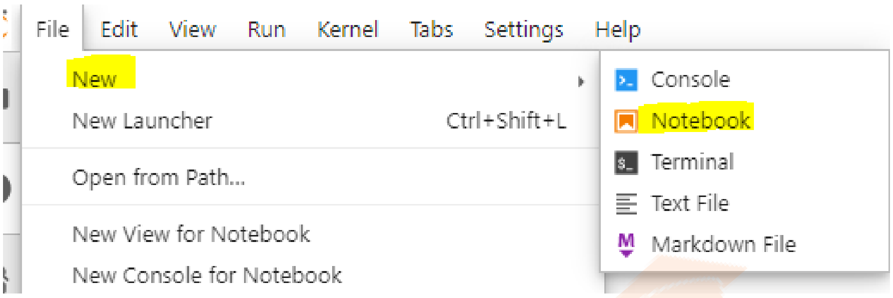
Q2. Execute the spark program to find the frequency of each word, same like shown in the sessions

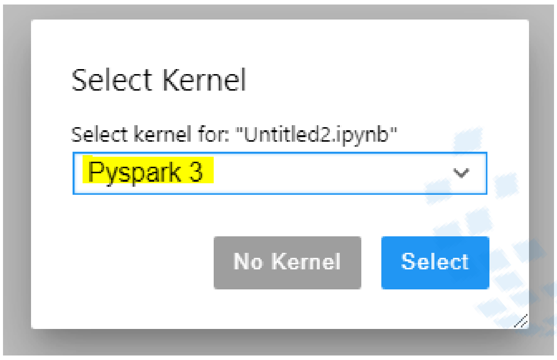
Consider the input file – inputfile.txt



```
[itv005357@g01 ~]$ hadoop fs -cat data/input/inputfile.txt
big data is interesting
big data is one of the most trending technology
my name is sumit
i teach big data
my institue name is trendytech
```

Open the notebook to run the code:





Select Kernel

Select kernel for: "Untitled2.ipynb"

Pyspark 3

```
from pyspark.sql import SparkSession
import getpass
username = getpass.getuser()
spark = SparkSession. \
builder. \
config('spark.ui.port','0'). \
config("spark.sql.warehouse.dir", f"/user/itv000173/warehouse"). \
enableHiveSupport(). \
master('yarn'). \
getOrCreate()
```

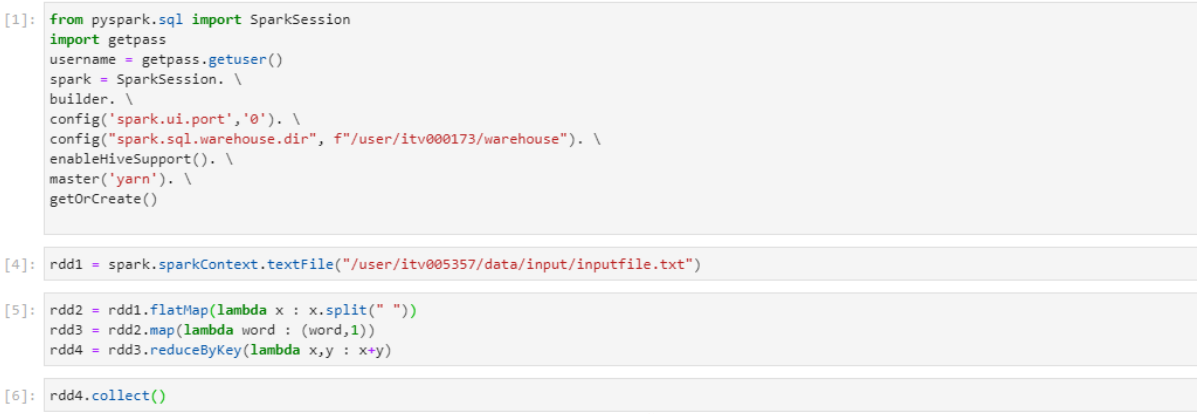
rdd1 = spark.sparkContext.textFile("/user/itv005357/data/input/inputfile.txt")

rdd2 = rdd1.flatMap(lambda x : x.split(" "))

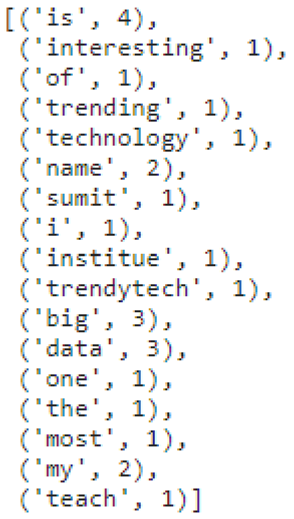
rdd3 = rdd2.map(lambda word : (word,1))

rdd4 = rdd3.reduceByKey(lambda x,y : x+y)

rdd4.collect()



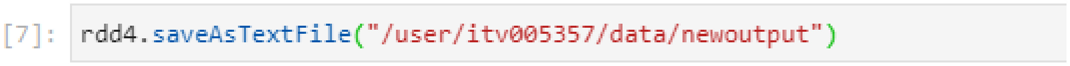
```
[1]: from pyspark.sql import SparkSession
     import getpass
     username = getpass.getuser()
     spark = SparkSession. \
     builder. \
     config('spark.ui.port','0'). \
     config("spark.sql.warehouse.dir", f"/user/itv000173/warehouse"). \
     enableHiveSupport(). \
     master('yarn'). \
     getOrCreate()

[4]: rdd1 = spark.sparkContext.textFile("/user/itv005357/data/input/inputfile.txt")

[5]: rdd2 = rdd1.flatMap(lambda x : x.split(" "))
     rdd3 = rdd2.map(lambda word : (word,1))
     rdd4 = rdd3.reduceByKey(lambda x,y : x+y)

[6]: rdd4.collect()
```



```
[('is', 4),
 ('interesting', 1),
 ('of', 1),
 ('trending', 1),
 ('technology', 1),
 ('name', 2),
 ('sumit', 1),
 ('i', 1),
 ('institue', 1),
 ('trendytech', 1),
 ('big', 3),
 ('data', 3),
 ('one', 1),
 ('the', 1),
 ('most', 1),
 ('my', 2),
 ('teach', 1)]
```

To save the file to HDFS:

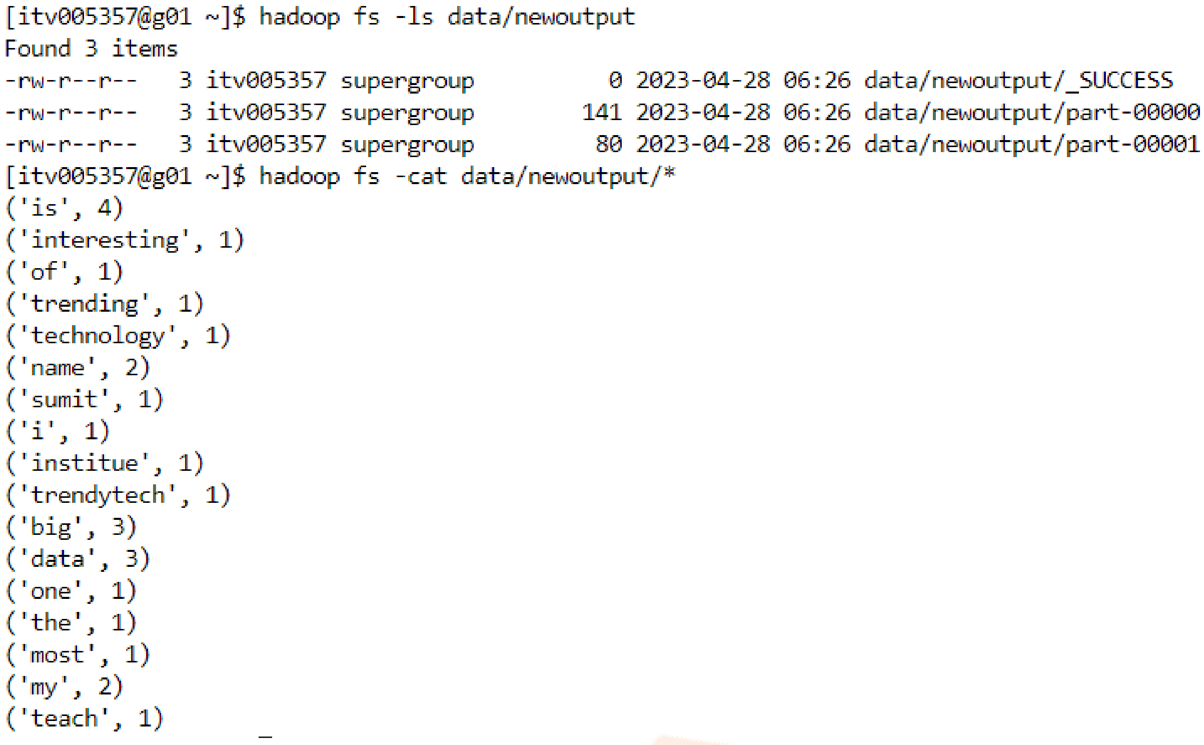
rdd4.saveAsTextFile("/user/itv005357/data/newoutput")



```
[7]: rdd4.saveAsTextFile("/user/itv005357/data/newoutput")
```

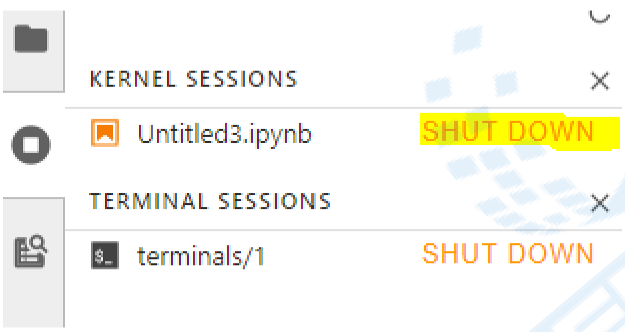
hadoop fs -ls data/newoutput

hadoop fs -cat data/newoutput/*

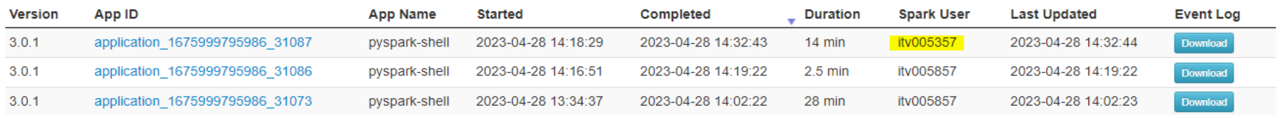


```
[itv005357@g01 ~]$ hadoop fs -ls data/newoutput
Found 3 items
-rw-r--r--   3 itv005357 supergroup          0 2023-04-28 06:26 data/newoutput/_SUCCESS
-rw-r--r--   3 itv005357 supergroup        141 2023-04-28 06:26 data/newoutput/part-00000
-rw-r--r--   3 itv005357 supergroup         80 2023-04-28 06:26 data/newoutput/part-00001
[itv005357@g01 ~]$ hadoop fs -cat data/newoutput/*
('is', 4)
('interesting', 1)
('of', 1)
('trending', 1)
('technology', 1)
('name', 2)
('sumit', 1)
('i', 1)
('institue', 1)
('trendytech', 1)
('big', 3)
('data', 3)
('one', 1)
('the', 1)
('most', 1)
('my', 2)
('teach', 1)
```

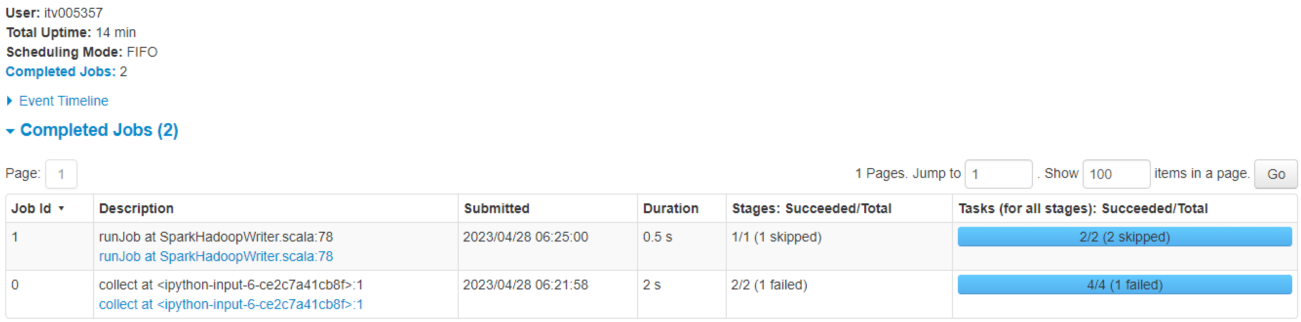
To see the history server, shut down the notebook.



Now go to the URL: http://m02.itversity.com:18080/



Version	App ID	App Name	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.0.1	application_1675999795986_31087	pyspark-shell	2023-04-28 14:18:29	2023-04-28 14:32:43	14 min	itv005357	2023-04-28 14:32:44	Download
3.0.1	application_1675999795986_31086	pyspark-shell	2023-04-28 14:16:51	2023-04-28 14:19:22	2.5 min	itv005857	2023-04-28 14:19:22	Download
3.0.1	application_1675999795986_31073	pyspark-shell	2023-04-28 13:34:37	2023-04-28 14:02:22	28 min	itv005857	2023-04-28 14:02:23	Download



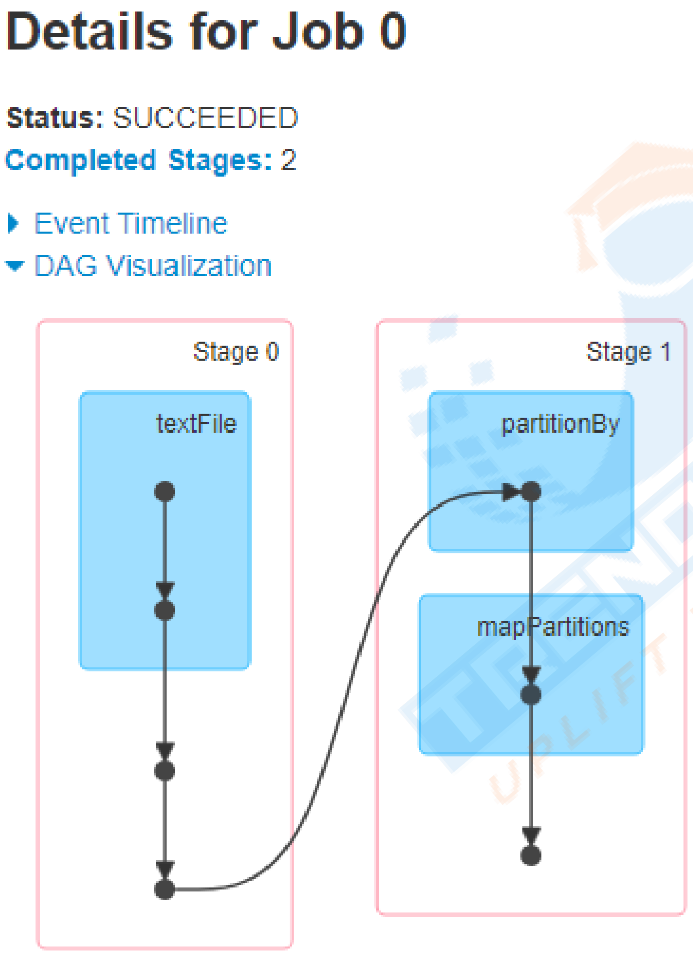
User: itv005357
Total Uptime: 14 min
Scheduling Mode: FIFO
Completed Jobs: 2

▸ Event Timeline

▾ Completed Jobs (2)

Page: 1   1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
1	runJob at SparkHadoopWriter.scala:78 runJob at SparkHadoopWriter.scala:78	2023/04/28 06:25:00	0.5 s	1/1 (1 skipped)	2/2 (2 skipped)
0	collect at <ipython-input-6-ce2c7a41cb8f>:1 collect at <ipython-input-6-ce2c7a41cb8f>:1	2023/04/28 06:21:58	2 s	2/2 (1 failed)	4/4 (1 failed)



Details for Job 0

Status: SUCCEEDED

Completed Stages: 2

▸ Event Timeline

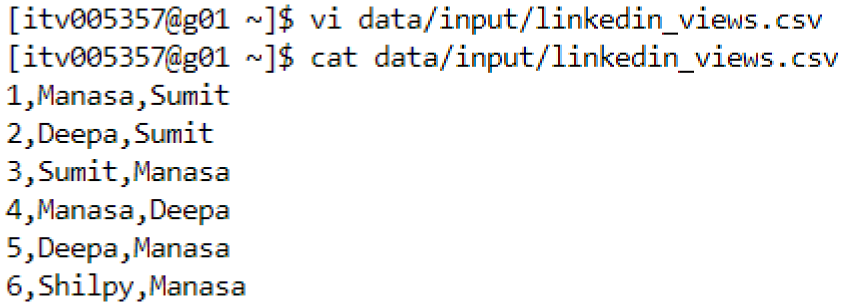
▾ DAG Visualization

Q3. Create logic for linkedIn profile views in Apache Spark.

First create an input file with the mentioned name:

vi data/input/linkedin_views.csv

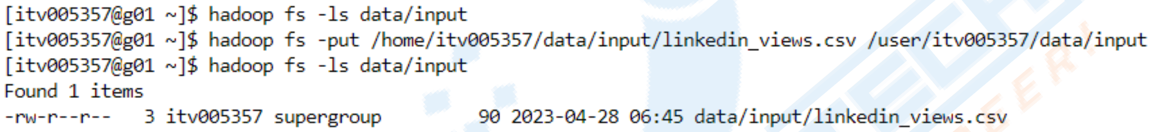
cat data/input/linkedin_views.csv



```
[itv005357@g01 ~]$ vi data/input/linkedin_views.csv
[itv005357@g01 ~]$ cat data/input/linkedin_views.csv
1,Manasa,Sumit
2,Deepa,Sumit
3,Sumit,Manasa
4,Manasa,Deepa
5,Deepa,Manasa
6,Shilpy,Manasa
```

hadoop fs -put /home/itv005357/data/input/linkedin_views.csv /user/itv005357/data/input

hadoop fs -ls data/input



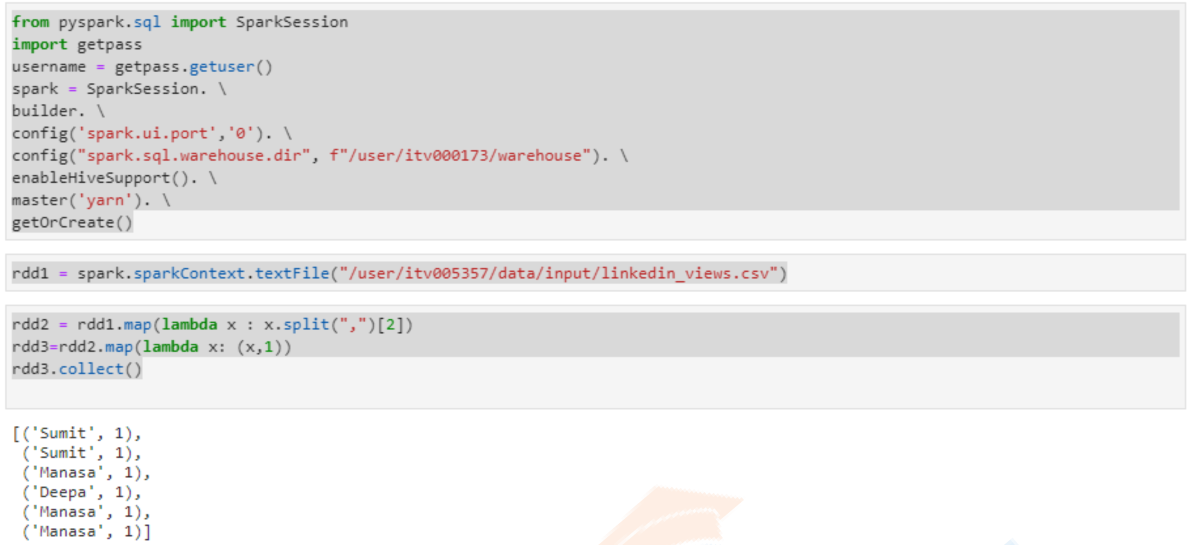
```
[itv005357@g01 ~]$ hadoop fs -ls data/input
[itv005357@g01 ~]$ hadoop fs -put /home/itv005357/data/input/linkedin_views.csv /user/itv005357/data/input
[itv005357@g01 ~]$ hadoop fs -ls data/input
Found 1 items
-rw-r--r--   3 itv005357 supergroup         90 2023-04-28 06:45 data/input/linkedin_views.csv
```

```
from pyspark.sql import SparkSession
import getpass
username = getpass.getuser()
spark = SparkSession. \
builder. \
config('spark.ui.port','0'). \
config("spark.sql.warehouse.dir", f"/user/itv000173/warehouse"). \
enableHiveSupport(). \
master('yarn'). \
getOrCreate()
rdd1 =
spark.sparkContext.textFile("/user/itv005357/data/input/linkedin_views.csv")
rdd2 = rdd1.map(lambda x : x.split(",")[2])
```

rdd3=rdd2.map(lambda x: (x,1))

rdd4 = rdd3.reduceByKey(lambda x,y : x+y)

rdd4.saveAsTextFile("/user/itv005357/data/output")

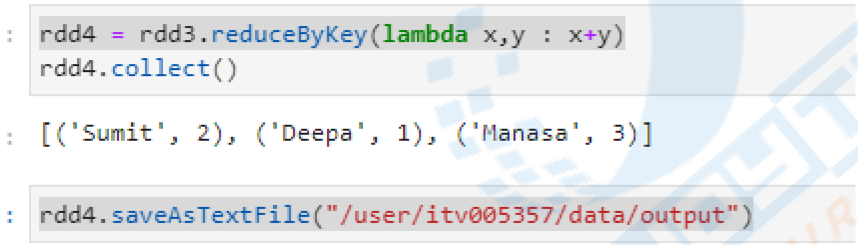


```
from pyspark.sql import SparkSession
import getpass
username = getpass.getuser()
spark = SparkSession. \
builder. \
config('spark.ui.port','0'). \
config("spark.sql.warehouse.dir", f"/user/itv000173/warehouse"). \
enableHiveSupport(). \
master('yarn'). \
getOrCreate()
```

```
rdd1 = spark.sparkContext.textFile("/user/itv005357/data/input/linkedin_views.csv")
```

```
rdd2 = rdd1.map(lambda x : x.split(",")[2])
rdd3=rdd2.map(lambda x: (x,1))
rdd3.collect()
```

```
[('Sumit', 1),
 ('Sumit', 1),
 ('Manasa', 1),
 ('Deepa', 1),
 ('Manasa', 1),
 ('Manasa', 1)]
```



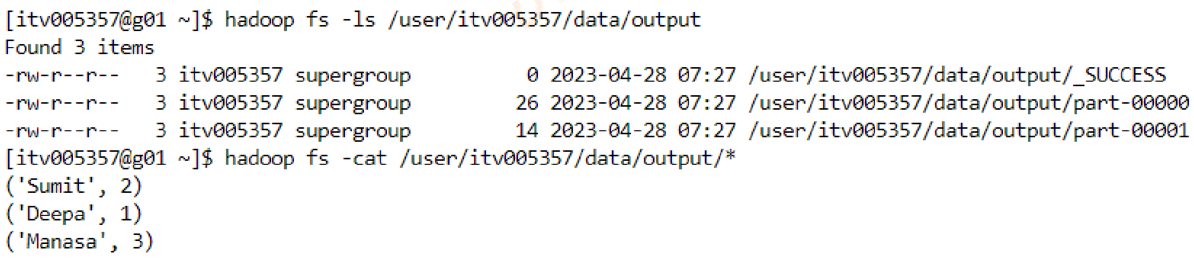
```
rdd4 = rdd3.reduceByKey(lambda x,y : x+y)
rdd4.collect()
```

```
[('Sumit', 2), ('Deepa', 1), ('Manasa', 3)]
```

```
rdd4.saveAsTextFile("/user/itv005357/data/output")
```

hadoop fs -ls /user/itv005357/data/output

hadoop fs -cat /user/itv005357/data/output/*



```
[itv005357@g01 ~]$ hadoop fs -ls /user/itv005357/data/output
Found 3 items
-rw-r--r--   3 itv005357 supergroup          0 2023-04-28 07:27 /user/itv005357/data/output/_SUCCESS
-rw-r--r--   3 itv005357 supergroup         26 2023-04-28 07:27 /user/itv005357/data/output/part-00000
-rw-r--r--   3 itv005357 supergroup         14 2023-04-28 07:27 /user/itv005357/data/output/part-00001
[itv005357@g01 ~]$ hadoop fs -cat /user/itv005357/data/output/*
('Sumit', 2)
('Deepa', 1)
('Manasa', 3)
```