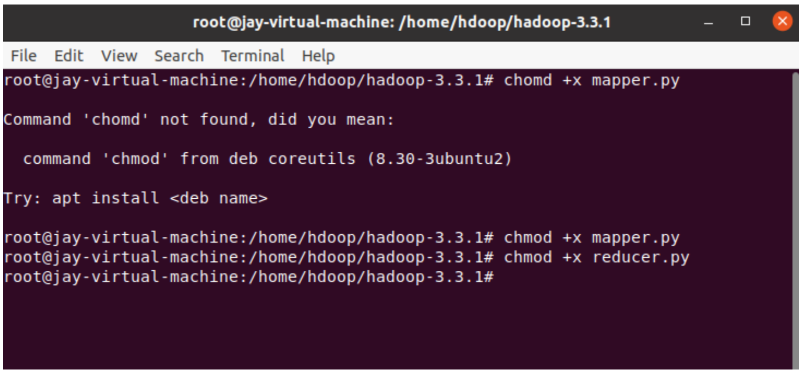
1. Describe the setup architecture of your exercise.
   I divide the whole program into three parts, including Mapper, Shuffle, Reducer respectively. Mapper is aimed to process matrix's data and Shuffle will distribute data with same key into same iterate, and Mapper will calculate the data from Shuffle, finally return and print result.

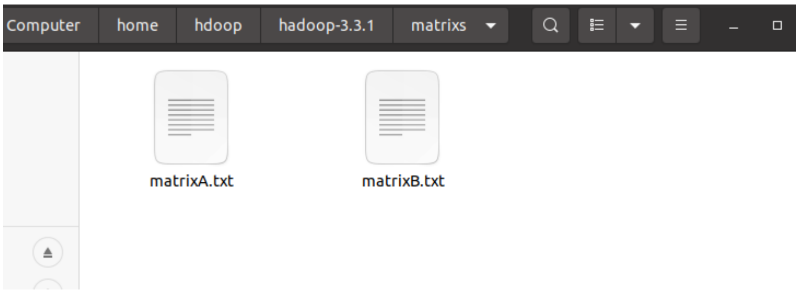
2. Describe Mapper and Reducer function in Python.
   Mapper will mark the data (aij) from matrix A with form of <key, value>, key="i, k", i is number of row and k is number of matrix B's column, value = "a:j, aij". As well as mark the data (bij) from matrix B with form of <key, value>, key="k, j", j is number of column and k is number of matrix A's row, value = "b:i, bij". Based on this, it can easily tell which matrix elements belong and specific position in matrix. And cij can be calculated with data companied with same key value.
   Reducer will do work of calculation, it will multiply data with same number after type of matrix and clearly know the position of data regarding to key value.

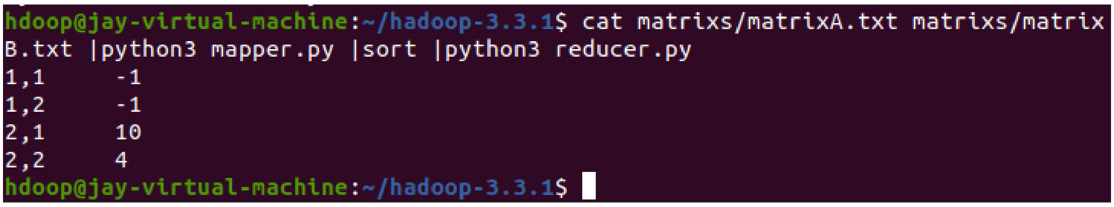
3. Describe your experience step-by-step in your own words and provide screenshots of executed MapReduce programs.


root@jay-virtual-machine: /home/hdoop/hadoop-3.3.1
File Edit View Search Terminal Help

```
root@jay-virtual-machine:/home/hdoop/hadoop-3.3.1# chomd +x mapper.py

Command 'chomd' not found, did you mean:

  command 'chmod' from deb coreutils (8.30-3ubuntu2)

Try: apt install <deb name>

root@jay-virtual-machine:/home/hdoop/hadoop-3.3.1# chmod +x mapper.py
root@jay-virtual-machine:/home/hdoop/hadoop-3.3.1# chmod +x reducer.py
root@jay-virtual-machine:/home/hdoop/hadoop-3.3.1#
```

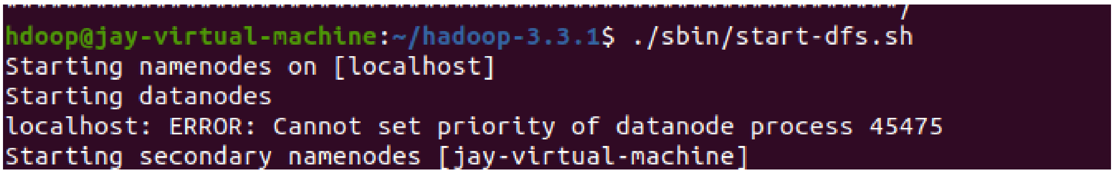
Modify permission of mapper file and reducer file.



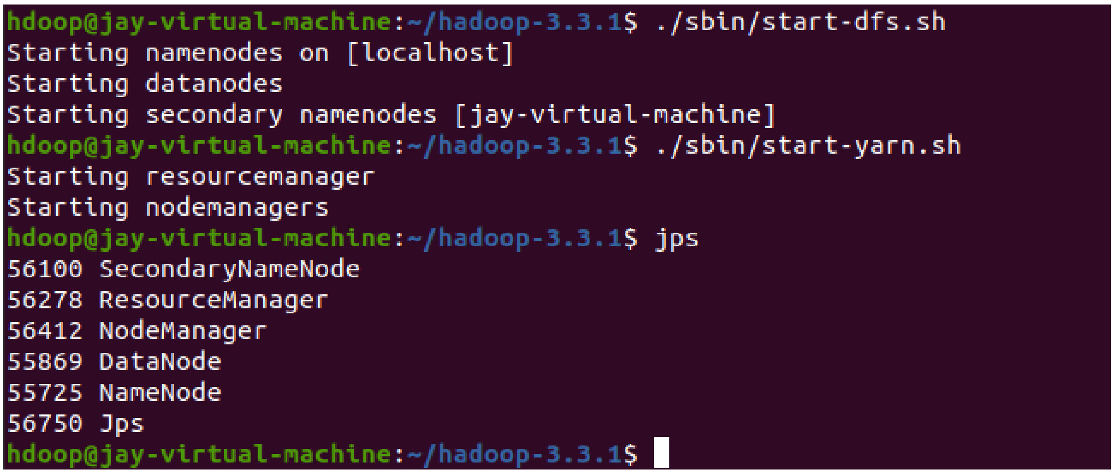


```
hdoop@jay-virtual-machine:~/hadoop-3.3.1$ cat matrixs/matrixA.txt matrixs/matrix
B.txt |python3 mapper.py |sort |python3 reducer.py
1,1     -1
1,2     -1
2,1     10
2,2     4
hdoop@jay-virtual-machine:~/hadoop-3.3.1$
```

Test the result on my own localhost.

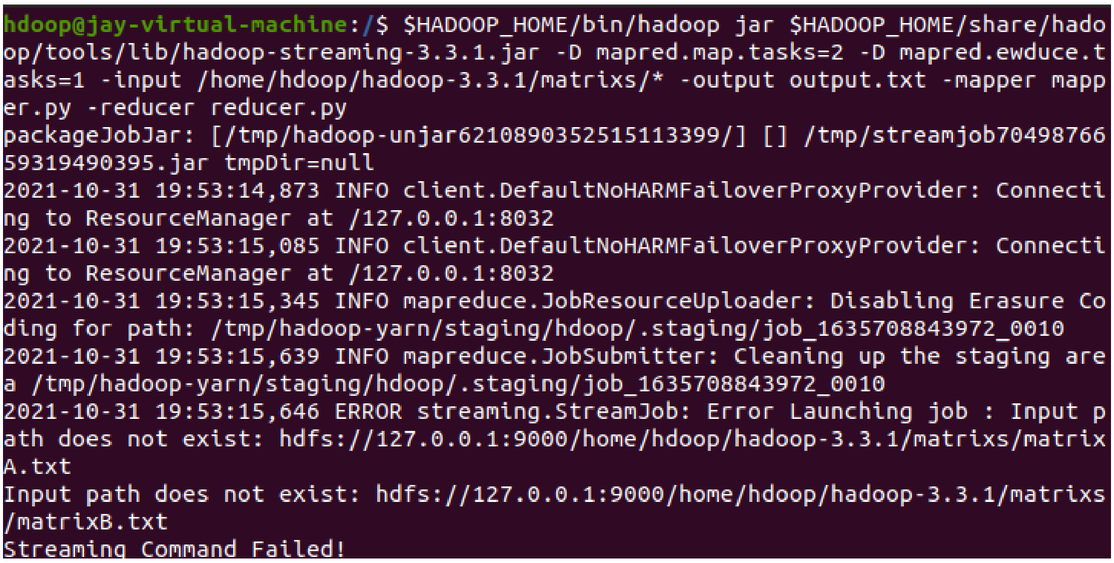


```
hdoop@jay-virtual-machine:~/hadoop-3.3.1$ ./sbin/start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
localhost: ERROR: Cannot set priority of datanode process 45475
Starting secondary namenodes [jay-virtual-machine]
```

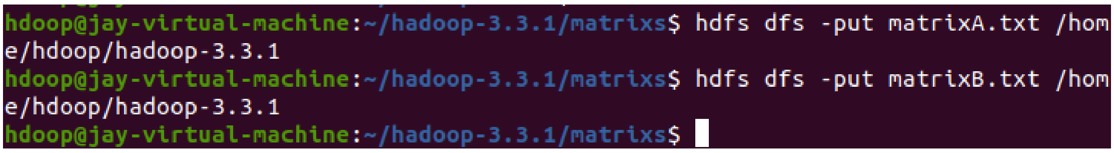


```
hdoop@jay-virtual-machine:~/hadoop-3.3.1$ ./sbin/start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [jay-virtual-machine]
hdoop@jay-virtual-machine:~/hadoop-3.3.1$ ./sbin/start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hdoop@jay-virtual-machine:~/hadoop-3.3.1$ jps
56100 SecondaryNameNode
56278 ResourceManager
56412 NodeManager
55869 DataNode
55725 NameNode
56750 Jps
hdoop@jay-virtual-machine:~/hadoop-3.3.1$
```

Run hdfs file system and yarn resource manager. Type ips command to see what services are open now.

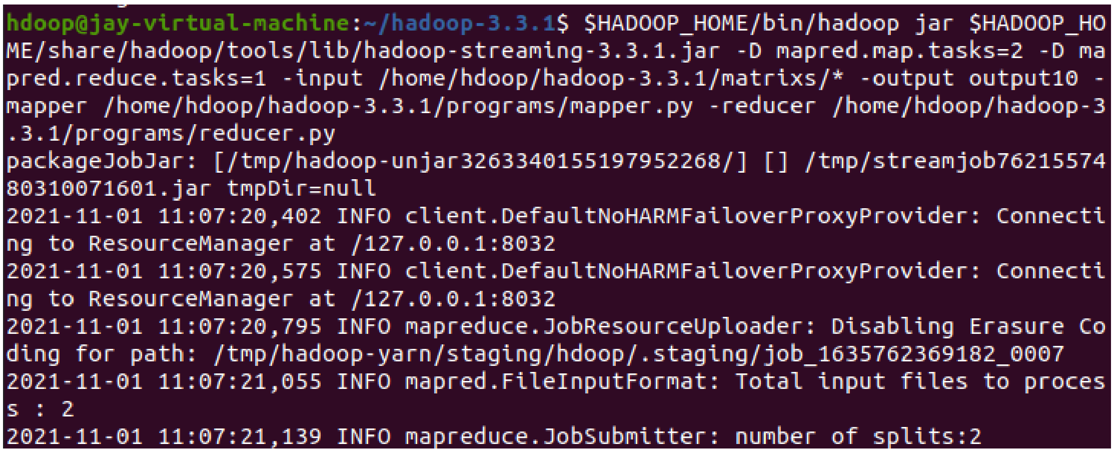


```
hdoop@jay-virtual-machine:/$ $HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/share/hado
op/tools/lib/hadoop-streaming-3.3.1.jar -D mapred.map.tasks=2 -D mapred.ewduce.t
asks=1 -input /home/hdoop/hadoop-3.3.1/matrixs/* -output output.txt -mapper mapp
er.py -reducer reducer.py
packageJobJar: [/tmp/hadoop-unjar6210890352515113399/] [] /tmp/streamjob70498766
59319490395.jar tmpDir=null
2021-10-31 19:53:14,873 INFO client.DefaultNoHARMFailoverProxyProvider: Connecti
ng to ResourceManager at /127.0.0.1:8032
2021-10-31 19:53:15,085 INFO client.DefaultNoHARMFailoverProxyProvider: Connecti
ng to ResourceManager at /127.0.0.1:8032
2021-10-31 19:53:15,345 INFO mapreduce.JobResourceUploader: Disabling Erasure Co
ding for path: /tmp/hadoop-yarn/staging/hdoop/.staging/job_1635708843972_0010
2021-10-31 19:53:15,639 INFO mapreduce.JobSubmitter: Cleaning up the staging are
a /tmp/hadoop-yarn/staging/hdoop/.staging/job_1635708843972_0010
2021-10-31 19:53:15,646 ERROR streaming.StreamJob: Error Launching job : Input p
ath does not exist: hdfs://127.0.0.1:9000/home/hdoop/hadoop-3.3.1/matrixs/matrix
A.txt
Input path does not exist: hdfs://127.0.0.1:9000/home/hdoop/hadoop-3.3.1/matrixs
/matrixB.txt
Streaming Command Failed!
```



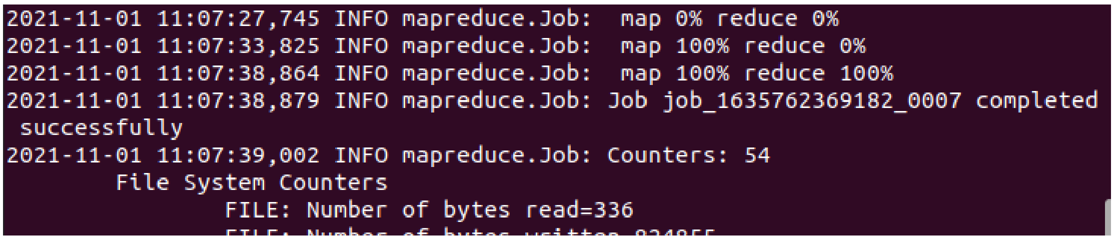
```
hdoop@jay-virtual-machine:~/hadoop-3.3.1/matrixs$ hdfs dfs -put matrixA.txt /hom
e/hdoop/hadoop-3.3.1
hdoop@jay-virtual-machine:~/hadoop-3.3.1/matrixs$ hdfs dfs -put matrixB.txt /hom
e/hdoop/hadoop-3.3.1
hdoop@jay-virtual-machine:~/hadoop-3.3.1/matrixs$
```

Have to put input matrixs on hdfs in advance.



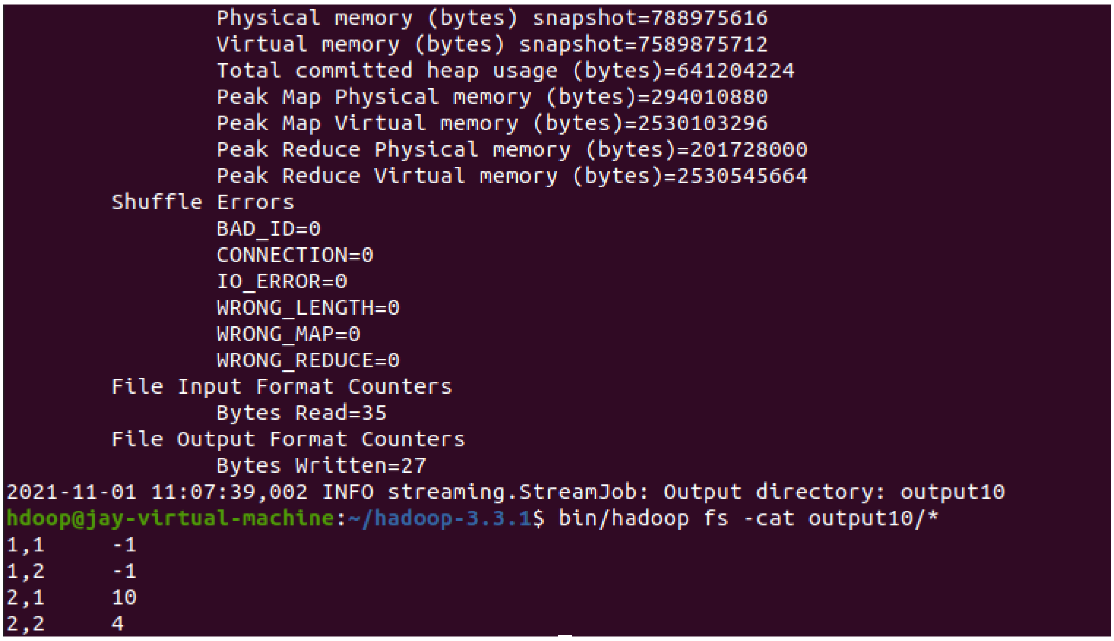
```
hdoop@jay-virtual-machine:~/hadoop-3.3.1$ $HADOOP_HOME/bin/hadoop jar $HADOOP_HO
ME/share/hadoop/tools/lib/hadoop-streaming-3.3.1.jar -D mapred.map.tasks=2 -D ma
pred.reduce.tasks=1 -input /home/hdoop/hadoop-3.3.1/matrixs/* -output output10 -
mapper /home/hdoop/hadoop-3.3.1/programs/mapper.py -reducer /home/hdoop/hadoop-3
.3.1/programs/reducer.py
packageJobJar: [/tmp/hadoop-unjar3263340155197952268/] [] /tmp/streamjob76215574
80310071601.jar tmpDir=null
2021-11-01 11:07:20,402 INFO client.DefaultNoHARMFailoverProxyProvider: Connecti
ng to ResourceManager at /127.0.0.1:8032
2021-11-01 11:07:20,575 INFO client.DefaultNoHARMFailoverProxyProvider: Connecti
ng to ResourceManager at /127.0.0.1:8032
2021-11-01 11:07:20,795 INFO mapreduce.JobResourceUploader: Disabling Erasure Co
ding for path: /tmp/hadoop-yarn/staging/hdoop/.staging/job_1635762369182_0007
2021-11-01 11:07:21,055 INFO mapred.FileInputFormat: Total input files to proces
s : 2
2021-11-01 11:07:21,139 INFO mapreduce.JobSubmitter: number of splits:2
```

Upload input, output, mapper, reducer into Hadoop's HDFS by Streaming, and run program.



```
2021-11-01 11:07:27,745 INFO mapreduce.Job:  map 0% reduce 0%
2021-11-01 11:07:33,825 INFO mapreduce.Job:  map 100% reduce 0%
2021-11-01 11:07:38,864 INFO mapreduce.Job:  map 100% reduce 100%
2021-11-01 11:07:38,879 INFO mapreduce.Job: Job job_1635762369182_0007 completed
 successfully
2021-11-01 11:07:39,002 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=336
                FILE: Number of bytes written=834855
```

Program is completed.



```
                Physical memory (bytes) snapshot=788975616
                Virtual memory (bytes) snapshot=7589875712
                Total committed heap usage (bytes)=641204224
                Peak Map Physical memory (bytes)=294010880
                Peak Map Virtual memory (bytes)=2530103296
                Peak Reduce Physical memory (bytes)=201728000
                Peak Reduce Virtual memory (bytes)=2530545664
        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=35
        File Output Format Counters
                Bytes Written=27
2021-11-01 11:07:39,002 INFO streaming.StreamJob: Output directory: output10
hdoop@jay-virtual-machine:~/hadoop-3.3.1$ bin/hadoop fs -cat output10/*
1,1     -1
1,2     -1
2,1     10
2,2     4
```

Print the result from HDFS and its result is same with one completed on localhost.