

Soudabeh Rafiei

1. Creating directory and copying file into created directory

```
[hadoop@ip-172-31-31-104 ~]$ hadoop fs -mkdir SudiHW4
^[[A[hadoop@ip-172-31-31-104 ~]$ hadoop fs -mkdir SudiHW4/HWData
[hadoop@ip-172-31-31-104 ~]$ hadoop fs -mkdir SudiHW4/HWData/Input
[hadoop@ip-172-31-31-104 ~]$ hadoop fs -copyFromLocal NYSE.csv SudiHW4/HWData/Input/data.csv
[hadoop@ip-172-31-31-104 ~]$
```

Hadoop fs -copyFromLocal EntrKeyWords.txt sudi/EnterKeyWords.txt

2. Creating Mapper and Reducer using Nano

```
[hadoop@ip-172-31-31-104 ~]$ nano mapper.py
[hadoop@ip-172-31-31-104 ~]$ nano reducer.py
[hadoop@ip-172-31-31-104 ~]$ chmod +x mapper.py
[hadoop@ip-172-31-31-104 ~]$ chmod +x reducer.py
```

3. Search for Hadoop Jar

```
[hadoop@ip-172-31-31-104 ~]$ find /usr/lib/ -name *hadoop*streaming*.jar
/usr/lib/hadoop/hadoop-streaming-3.3.6-amzn-2.jar
/usr/lib/hadoop/hadoop-streaming.jar
/usr/lib/hadoop-mapreduce/hadoop-streaming-3.3.6-amzn-2.jar
/usr/lib/hadoop-mapreduce/hadoop-streaming.jar
```

4. Running the code

```
[hadoop@ip-172-31-31-104 ~]$ hadoop jar /usr/lib/hadoop/hadoop-streaming.jar -files mapper.py,reducer.py -mapper mapper.py -reducer reducer.py -input SudiHW4/HWData/Input/data.csv -output SudiHW4/HWData/output
```

Hadoop jar /usr/lib/Hadoop/Hadoop-streaming.jar -files mapper.py, reducer.py -mapper mapper.py -reducer reducer.py -input EntrKey.txt -output frequency

5. Search for output Directory

```
[hadoop@ip-172-31-31-104 ~]$ hadoop fs -ls SudiHW4/HWData/output
Found 4 items
-rw-r--r-- 1 hadoop hdfsadmingroup 0 2024-02-29 05:14 SudiHW4/HWData/output/_SUCCESS
-rw-r--r-- 1 hadoop hdfsadmingroup 484 2024-02-29 05:14 SudiHW4/HWData/output/part-00000
-rw-r--r-- 1 hadoop hdfsadmingroup 513 2024-02-29 05:14 SudiHW4/HWData/output/part-00001
-rw-r--r-- 1 hadoop hdfsadmingroup 500 2024-02-29 05:14 SudiHW4/HWData/output/part-00002
[hadoop@ip-172-31-31-104 ~]$
```

6. To see the result

```
[hadoop@ip-172-31-31-104 ~]$ hadoop fs -cat SudiHW4/HWData/output/p*
GAI      14.62
GAR      26.11
GB        44.6
GBB      52.36
GBE      14.05
GCA      18.0
GCS      17.57
GCV      11.2
GDF      15.48
GDI      56.55
GDL      19.99
GDO      20.02
GE       160.0
GET      56.72
GFA      37.17
GFY      19.16
GGC      50.41
GIM      9.98
GIS      83.25
GJF      25.3
GJI      24.65
GJL      24.7
GJO      24.5
```

Searching for Hadoop jar file: `find /usr/lib/ -name '*hadoop*streaming*.jar'`