**TERMWORK5**

**1.** hduser@ubuntu:~$ cd Documents/

2. hduser@ubuntu:~/Documents$ nano Term5.csv

E001, Sunita, Accounts, 15000

E002, Harsh, IT, 50000

E003, Ragini, IT, 75000

E004, Mithun, Accounts, 20000

E005, Pruthavi, Marketing, 45000

E006, Anjali, IT, 70000

E007, Kunal, Marketing, 60000

E008, Mitali, Accounts, 55000

E009, Roopa, IT, 70000

E010, Deepti, Accounts, 30000

E011, Janavi, Marketing, 25000

E012, Lata, Accounts, 30000

E013, Brijmohan, Marketing, 45000

E014, Nina, Accounts, 50000

E015, Pallavi, Marketing, 25000

3. hduser@ubuntu:~/Documents$ nano mapper.py

#!/usr/bin/env python

import sys

for line in sys.stdin:

line = line.strip()

line = line.split(",")

if len(line) >=2:

dept = line[2]

sal = line[3]

print '%s\t%s' % (dept, sal)

4. hduser@ubuntu:~/Documents$ cat Term5.csv| python mapper.py

5.hd.user@ubuntu:~/Documents$ nano reducer.py

#!/usr/bin/env python

import sys

deptdic={}

for line in sys.stdin:

line = line.strip()

dept,sal = line.split('\t')

if dept in deptdic:

deptdic[dept].append(int(sal))

else:

deptdic[dept] = []

deptdic[dept].append(int(sal))

for dept in deptdic.keys():

sum\_sal = sum(deptdic[dept])

print '%s\t%s'% (dept,sum\_sal)

6. hduser@ubuntu:~/Documents$ cat Term5.csv| python mapper.py | python reducer.py

7. hduser@ubuntu:~/Documents$start-all.sh

8. hduser@ubuntu:~/Documents$ jps

9. hduser@ubuntu:~/Documents$ hdfs dfs -mkdir /T5

10 hduser@ubuntu:~/Documents$ hdfs dfs -copyFromLocal /home/hduser/Documents/Term5.csv /T5

11. hduser@ubuntu:~/Documents$ hdfs dfs -ls /

12. hduser@ubuntu:~/Documents$ chmod 777 mapper.py reducer.py

13. hduser@ubuntu:~/Documents$ ls -l

14. hduser@ubuntu:~/Documents$ hadoop jar /home/hduser/Documents/hadoop-streaming-2.7.3.jar \

> -input /T5/ Term5.csv \

> -output /T5/output \

> -mapper /home/hduser/Documents/mapper.py \

> -reducer /home/hduser/Documents/reducer.py

15. hduser@ubuntu:~/Documents$ hdfs dfs -cat /T5 /output/part-00000