Map Reducer in collab using python

Name : Shyam Sundar S
Roll No : 24MX347

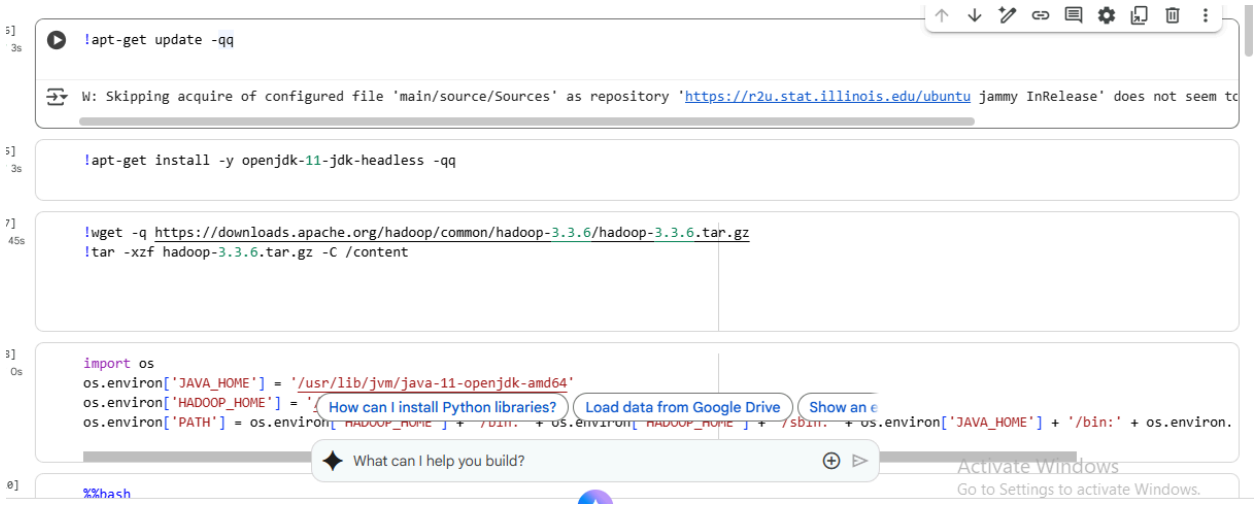
Step 01:

Map reducer Using python Google collab notebook Link(24MX347(shyam Sundar S) :

https://colab.research.google.com/drive/1DqXpRbO4BRGAF0c1Sber8Z3q0R3FK7DA?usp=sharing

```
!apt-get update -qq
!apt-get install -y openjdk-11-jdk-headless -qq
!wget -q https://downloads.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz
!tar -xzf hadoop-3.3.6.tar.gz -C /content
```

Output:



Step 02:

```
import os
os.environ['JAVA_HOME'] = '/usr/lib/jvm/java-11-openjdk-amd64'
os.environ['HADOOP_HOME'] = '/content/hadoop-3.3.6'
os.environ['PATH'] = os.environ['HADOOP_HOME'] + '/bin:' + os.environ['HADOOP_HOME'] +
'/sbin:' + os.environ['JAVA_HOME'] + '/bin:' + os.environ.get('PATH','')
```

```
%%bash
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
export HADOOP_HOME=/content/hadoop-3.3.6
```

Map Reducer in collab using python

Name : Shyam Sundar S
Roll No : 24MX347

```
export PATH=$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$JAVA_HOME/bin:$PATH


# test
echo "JAVA_HOME is $JAVA_HOME"

import os
os.environ['JAVA_HOME'] = '/usr/lib/jvm/java-11-openjdk-amd64'
os.environ['HADOOP_HOME'] = '/content/hadoop-3.3.6'
os.environ['PATH'] = os.environ['HADOOP_HOME'] + '/bin:' + \
                     os.environ['HADOOP_HOME'] + '/sbin:' + \
                     os.environ['JAVA_HOME'] + '/bin:' + \
                     os.environ['PATH']


# Test hadoop
!hadoop version
```

output:

Step 03:

```
!mkdir -p /content/hadoop_input
!echo "Hadoop is powerful. Hadoop is open source. Hadoop runs on big data." >
/content/hadoop_input/input.txt


%%bash
```

Map Reducer in collab using python

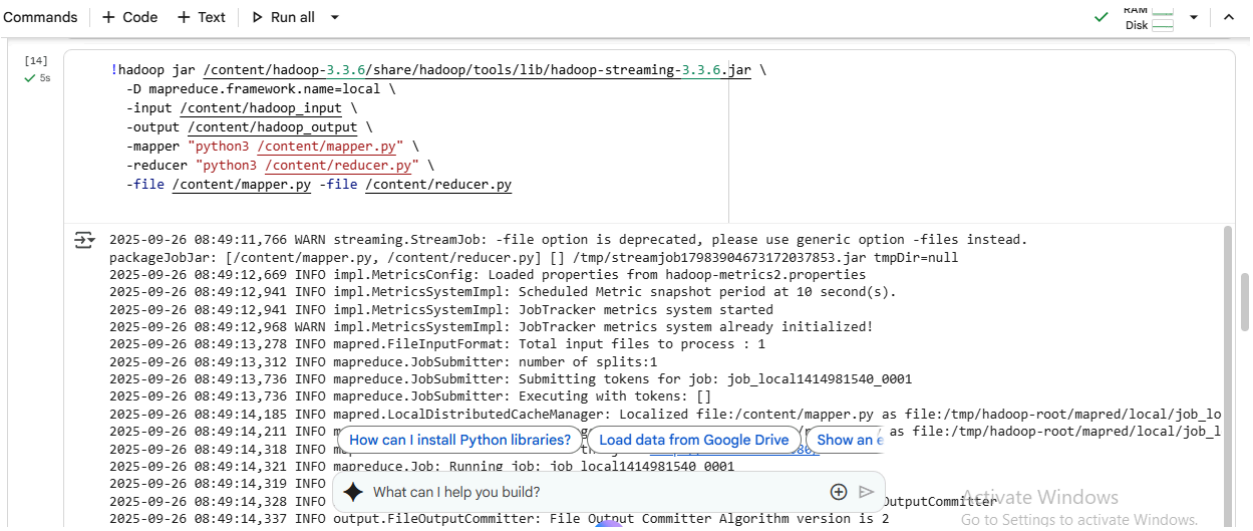
Name : Shyam Sundar S
Roll No : 24MX347

```
cat > /content/mapper.py <<'PY'
#!/usr/bin/env python3
import sys
for line in sys.stdin:
    for word in line.strip().split():
        w = word.strip().lower().strip('.,!?:;"\'')
        if w:
            print(f"{w}\t1")
PY


cat > /content/reducer.py <<'PY'
#!/usr/bin/env python3
import sys
from collections import defaultdict
counts = defaultdict(int)
for line in sys.stdin:
    parts = line.strip().split('\t',1)
    if len(parts)==2:
        word, cnt = parts
        try:
            counts[word] += int(cnt)
        except:
            pass
for w in sorted(counts):
    print(f"{w}\t{counts[w]}")
PY


chmod +x /content/mapper.py /content/reducer.py
```

Output :

Map Reducer in collab using python

Name : Shyam Sundar S
Roll No : 24MX347



Step 04:

```
!hadoop jar /content/hadoop-3.3.6/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar \
  -D mapreduce.framework.name=local \
  -input /content/hadoop_input \
  -output /content/hadoop_output \
  -mapper "python3 /content/mapper.py" \
  -reducer "python3 /content/reducer.py" \
  -file /content/mapper.py -file /content/reducer.py
```

```
!cat /content/hadoop_output/part-00000
```

Output :

Map Reducer in collab using python

Name : Shyam Sundar S
Roll No : 24MX347