

## Map Reducer in collab using python

Name : Shyam Sundar S

Roll No : 24MX347

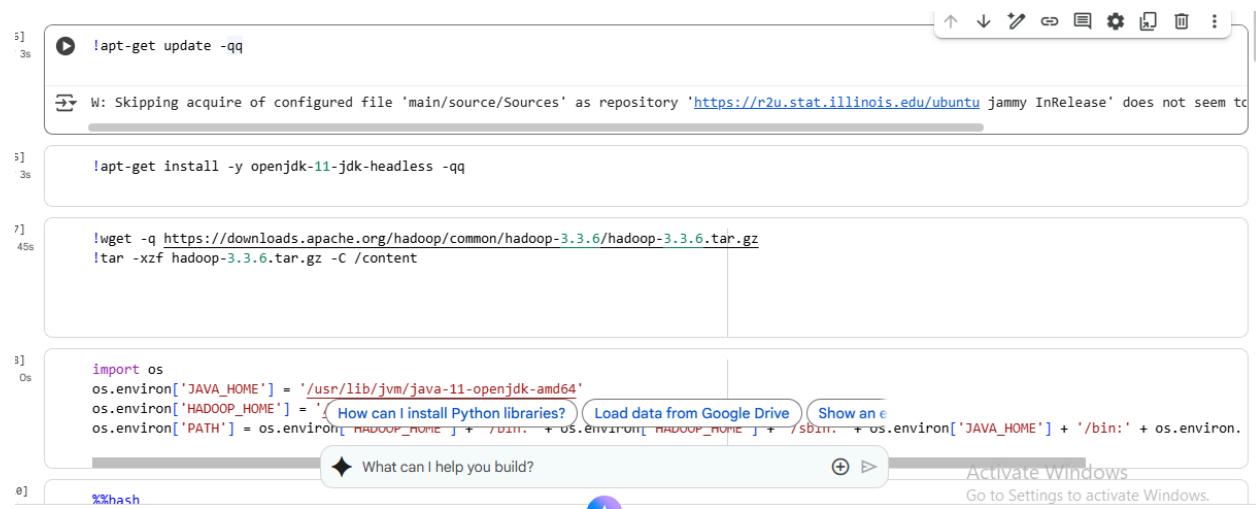
Step 01:

**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:



```
3] !apt-get update -qq  
3s W: Skipping acquire of configured file 'main/source/Sources' as repository 'https://r2u.stat.illinois.edu/ubuntu jammy InRelease' does not seem to  
4] !apt-get install -y openjdk-11-jdk-headless -qq  
4s  
7] 45s !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  
8] 0s 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', '')  
8] %%hash
```

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:

```
[11] 1s
os.environ['JAVA_HOME'] + '/bin:' + \
os.environ['PATH']

# Test hadoop
!hadoop version

→ Hadoop 3.3.6
Source code repository https://github.com/apache/hadoop.git -r 1be78238728da9266a4f88195058f08fd012bf9c
Compiled by ubuntu on 2023-06-18T08:22Z
Compiled on platform linux-x86_64
Compiled with protoc 3.7.1
From source with checksum 5652179ad55f76cb287d9c633bb53bb
This command was run using /content/hadoop-3.3.6/share/hadoop/common/hadoop-common-3.3.6.jar

[12] 0s
!mkdir -p /content/hadoop_input
!echo "Hadoop is powerful. Hadoop is open source. Hadoop runs on big data." > /content/hadoop_input/input.txt
```

### 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

```
[14] 5s
!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

2025-09-26 08:49:11,766 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/content/mapper.py, /content/reducer.py] [] /tmp/streamjob17983904673172037853.jar tmpDir=null
2025-09-26 08:49:12,669 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-09-26 08:49:12,941 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-09-26 08:49:12,941 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-09-26 08:49:12,968 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-09-26 08:49:13,278 INFO mapred.FileInputFormat: Total input files to process : 1
2025-09-26 08:49:13,312 INFO mapreduce.JobSubmitter: number of splits:1
2025-09-26 08:49:13,736 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1414981540_0001
2025-09-26 08:49:13,736 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-09-26 08:49:14,185 INFO mapred.LocalDistributedCacheManager: Localized file:/content/mapper.py as file:/tmp/hadoop-root/mapred/local/job_local1414981540_0001/mapper.py
2025-09-26 08:49:14,211 INFO mapred.LocalDistributedCacheManager: Localized file:/content/reducer.py as file:/tmp/hadoop-root/mapred/local/job_local1414981540_0001/reducer.py
2025-09-26 08:49:14,318 INFO mapreduce.Job: Running job: job_local1414981540_0001
2025-09-26 08:49:14,319 INFO mapreduce.Job: ◆ What can I help you build?
2025-09-26 08:49:14,328 INFO mapreduce.Job: OutputCommitter: File Output Committer Algorithm version is 2
2025-09-26 08:49:14,337 INFO mapreduce.Job: Go to Settings to activate Windows.
```

## 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

The screenshot shows a Jupyter Notebook interface with two code cells.

**Cell 1:** Displays log output from a MapReduce job. The logs show various INFO messages related to configuration deprecation, streaming, and map tasks. Key messages include:

```
2025-09-26 08:49:14,664 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
2025-09-26 08:49:14,665 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
2025-09-26 08:49:14,695 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
2025-09-26 08:49:14,752 INFO streaming.PipeMapRed: Records R/W=1/1
2025-09-26 08:49:14,761 INFO streaming.PipeMapRed: MRerrorThread done
2025-09-26 08:49:14,762 INFO streaming.PipeMapRed: mapRedFinished
2025-09-26 08:49:14,765 INFO mapred.LocalJobRunner:
2025-09-26 08:49:14,765 INFO mapred.MapTask: Starting flush of map output
2025-09-26 08:49:14,765 INFO mapred.MapTask: Spilling map output
2025-09-26 08:49:14,765 INFO mapred.MapTask: bufstart = 0; bufend = 89; bufvoid = 104857600
2025-09-26 08:49:14,765 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214352(104857408); length = 45/6553600
2025-09-26 08:49:14,775 INFO mapred.MapTask: Finished spill 0
2025-09-26 08:49:14,790 INFO mapred.Task: Task:attempt_local1414981540_0001_m_000000_0 is done. And is in the process of committing
2025-09-26 08:49:14,797 INFO mapred.LocalJobRunner: Records R/W=1/1
2025-09-26 08:49:14,797 INFO mapred.Task: Task 'attempt_local1414981540_0001_m_000000_0' done.
2025-09-26 08:49:14,805 INFO mapred.Task: Final Counters for attempt local1414981540_0001_m_000000_0:
Counters: 17
```

**Cell 2:** Displays the output of the command `!cat /content/hadoop_output/part-00000`. The output shows a word count of words in the HDFS output file. The words and their counts are:

Word	Count
big	1
data	1
hadoop	3
is	2
on	1
open	1
powerful	1

Below the code cells, there are several interactive buttons and a message about activating Windows.