A) CODE: Word Count

1. Open Terminal and switch to Hadoop user

```
pvg@pvg-HP-ProDesk-400-G4-SFF:~$ su hduser
Password:
```

2. Create a text file to count words

```
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ nano word_count.txt
```

3. Start HDFS

```
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ start-dfs.sh
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ start-yarn.sh
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ jps
```

4. Create an input directory and upload your file to HDFS:

```
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ hdfs dfs -ls /
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ hdfs dfs -rm -r /input
#Similarly, delete any previous output files if present using: hdfs dfs
rm -r /output
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ hdfs dfs -mkdir -p /input
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ hdfs dfs -ls /

hduser@pvg-HP-ProDesk-400-G4-SFF:~$ hdfs dfs -put word_count.txt /input/
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ hdfs dfs -ls /input/
```

5. Run the word count program:

```
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ whereis hadoop
```

```
hadoop: /usr/local/hadoop /usr/local/hadoop/bin/hadoop.cmd
/usr/local/hadoop/bin/Hadoop
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ hadoop jar
/usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.4.jar
wordcount /input /output
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ hdfs dfs -ls /
```

6. View Output:

```
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ hdfs dfs -ls /output/
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ hdfs dfs -cat /output/part-r-00000

hduser@pvg-HP-ProDesk-400-G4-SFF:~$ stop-dfs.sh
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ stop-yarn.sh
```

B) CODE: Character Count

mapper.py

```
#!/usr/bin/env python3
import sys
for line in sys.stdin:
   for char in line.strip():
      print(f"{char}\t1")
```

reducer.py

```
#!/usr/bin/env python3
import sys
```

```
from collections import defaultdict

counts = defaultdict(int)

for line in sys.stdin:
    line = line.strip()
    if not line:
        continue  # skip empty lines
    parts = line.split("\t")
    if len(parts) != 2:
        continue  # skip malformed lines
    key, val = parts
    try:
        counts[key] += int(val)
    except ValueError:
        continue  # skip lines with non-integer values

for key in sorted(counts):
    print(f"{key}\t{counts[key]}")
```

1. Open Terminal and switch to Hadoop user

```
pvg@pvg-HP-ProDesk-400-G4-SFF:~$ su hduser
Password:
```

2. Start HDFS

```
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ start-dfs.sh
```

```
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ start-yarn.sh
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ jps
```

3. Create an input directory

```
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ hdfs dfs -ls /
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ hdfs dfs -rm -r /input
#Similarly, delete any previous output files if present using: hdfs dfs
rm -r /output
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ hdfs dfs -mkdir -p /input
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ hdfs dfs -ls /
```

4. Create a text file and upload it to HDFS

```
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ nano character_count.txt
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ hdfs dfs -put character_count.txt /input/
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ hdfs dfs -ls /input/
```

5. Similarly, create a mapper.py and reducer.py file

```
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ nano mapper.py
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ nano reducer.py
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ chmod +x mapper.py
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ chmod +x reducer.py
```

6. Run Hadoop streaming jar using the mapper and reducer scripts

```
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ whereis hadoop
hadoop: /usr/local/hadoop /usr/local/hadoop/bin/hadoop.cmd
/usr/local/hadoop/bin/Hadoop
```

```
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ hadoop jar
/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.4.jar \
> -input /input/character_count.txt \
> -output /output/character_output \
> -mapper mapper.py \
> -reducer reducer.py \
> -file mapper.py \
> -file reducer.py
```

7. View Output

```
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ hdfs dfs -ls /output/character_output/
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ hdfs dfs -cat /output/character_output/part-00000
```

8. Stop HDFS

```
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ stop-dfs.sh
hduser@pvg-HP-ProDesk-400-G4-SFF:~$ stop-yarn.sh
```