Week-4

- Create 2 .py files → Mapper.py, Reducer.py

- To Find the location of python → which python3
  /usr/bin/python3

→ mapper3.py

```
#!/usr/bin/python3
"""mapper3.py"""
Import sys

for line in sys.stdin:
    line = line.strip()
    words = line.split()
    for word in words:
        print('%s\t%s' % (word, 1))
```

→ reducer3.py

```
#!/usr/bin/python3
"""reducer3.py"""
Import sys

current_word = None
current_count = 0
word = None
for line in sys.stdin:
    line = line.strip()
    word, count = line.split('\t', 1)
    try:
        count = int(count)
```

```
except ValueError:
    continue
if current_word == word:
    current_count += count
else:
    if current_word:
        print('%s \t %s' % (current_word, current_count))
    current_count = count
    current_word = word
if current_word == word:
    print('%s \t %s' % (current_word, current_count))
```

- Create an input file. (Vi p3.txt)

```
Hello
hello hi
good afternoon
This is BDA lab week 4
```

- Giving the file as input to mapper file

```
cat p3.txt | python3 mapper3.py
```

```
Hello   1
good    1
morning 1
hi      1
BDA     1
week    1
4       1
Lab     1
```

- Sending this output to reducer file

```
cat p3.txt | python3 mapper3.py | sort | python3 reducer3.py
# sorting is present as default. No shuffling
```

DATE / /

BD A 1
hello 2
thes 3
week 1

To run python file on Hadoop Cluster, we need support of Streaming

- /home/hdoop/hadoop-3.3.4/bin/hadoop jar '../hadoop-streaming-3.3.4.jar' \
  > -file /home/hdoop/mapper3.py -mapper mapper3.py \
  (Defining the mapper file on local system to read as mapper function)
  > -file /home/hdoop/reducer3.py -reducer reducer3.py \
  > -input /batch-3/200968048/p3.txt \ (Input File)
  > -output /batch-3/200968048/py_out (Output Directory)