BDA Practical Examination

Name: Simran Biswas Roll No: 8591

Word count program using MapReduce/PySpark

Colab Link:

https://colab.research.google.com/drive/1CkqcXTWQrYBbTxEo3Yo9GHrPC3dJDWmS?usp=sharing

Code:

```
#Mounting uploaded .txt file to drive
from google.colab import drive
drive.mount('/content/gdrive')


!pip install pyspark
import sys
from operator import add
from pyspark.sql import SparkSession


#Uploading file.txt
from google.colab import files
uploaded = files.upload()
FILE_LOCATION = './file.txt'


#Create a Spark Session

#read each line of file.txt
spark = SparkSession.builder.appName("WordCount").getOrCreate()
lines = spark.read.text(FILE_LOCATION).rdd.map(lambda r: r[0])


#Mapper of the application
mapper_output = lines.flatMap(lambda line: line.split(' ')).map(lambda
word:(word, 1))


#Shuffling & Sorting
mapper_output = mapper_output.sortByKey()
```

Name: Simran Biswas Roll No: 8591

```
#Reducer
reducer_output = mapper_output.reduceByKey(add).collect()

#Print count of each word
for (word, count) in reducer_output:
  print("{}: {}".format(word, count))

#Stop the Spark session
spark.stop()
```

Output:

```
[31] #Print count of each word
     for (word, count) in reducer_output:
       print("{}: {}".format(word, count))

     Bear: 2
     Car: 3
     Dear: 1
     Deer: 1
     River: 2
```

Conclusion:

With the help of MapReduce, we have implemented word count program. We have understood the basics of MapReduce and how it is useful for big data processing.