Name: Gouri Ramachandran

Roll no:22MIA1074

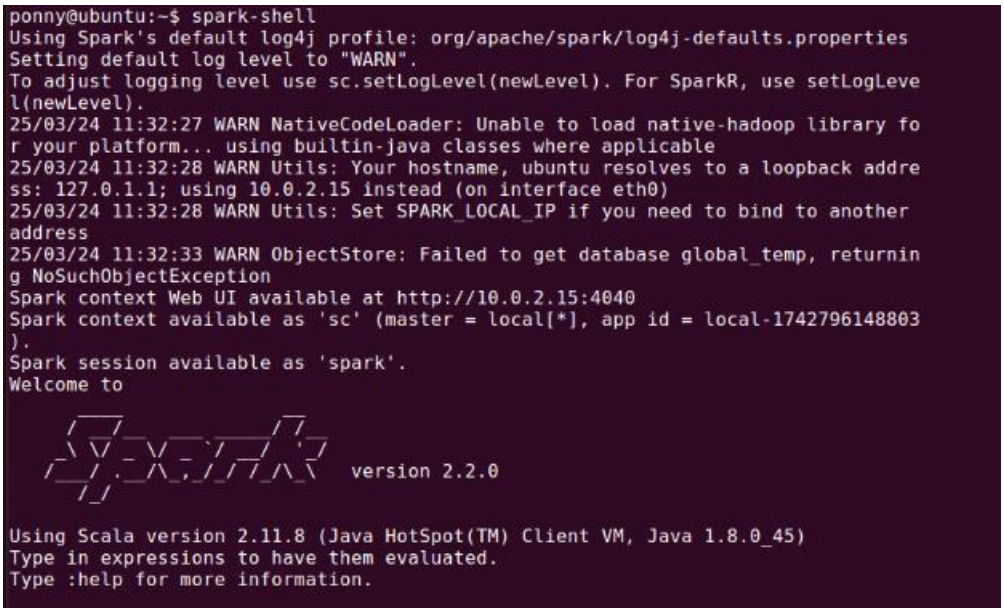
Experiment name: SPARK Framework Word Count Program Experiment

Aim

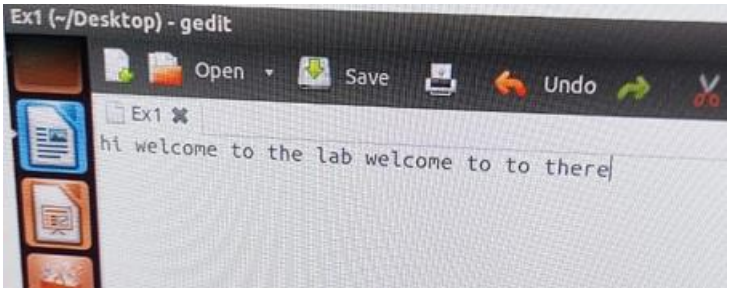
To implement a simple Word Count program using the Apache Spark framework in Scala, and execute it to count the frequency of words in a text file.

Algorithm/Procedure

1. **Environment Setup**
    - Open **Oracle VirtualBox**.
    - Double-click the **Hadoop** virtual machine.
    - Open the terminal and type:
    - spark-shell
    - Confirm Spark version and ensure Spark shell is running.



2. **Creating Input File**
    - On the Ubuntu desktop, create a sample text file named Ex1.txt.
    - Add sample content
    - Save this file on the desktop.



3. **Running Word Count Program in Spark**
    - In the Spark shell (Scala prompt), run the following commands:
    - Load text file from desktop (use appropriate path)
    - Split lines into words
    - Map words to (word, 1) pairs
    - Reduce by key to count occurrences
    - Collect and print the result

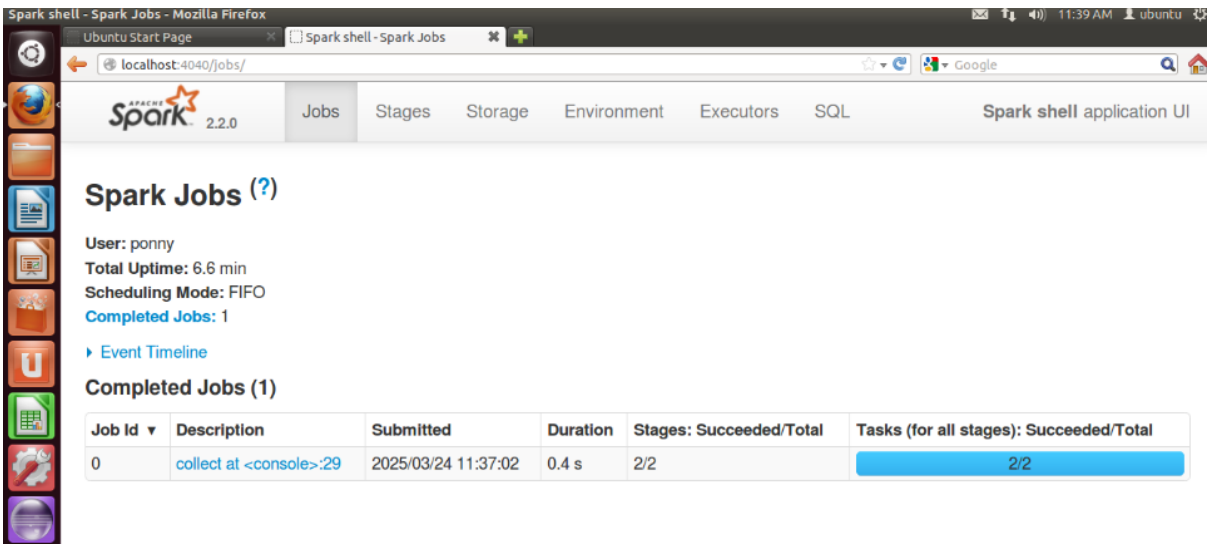
```
scala> var a = sc.textFile("/home/ponny/Desktop/Ex1").flatMap(line => line.split(" ")).map(word => (word,1))
25/03/24 11:36:08 WARN SizeEstimator: Failed to check whether UseCompressedOops is set; assuming yes
a: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[3] at map at <console>:24

scala> var b = a.reduceByKey(_+_);
b: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[4] at reduceByKey at <console>:26

scala> b.collect
res0: Array[(String, Int)] = Array((lab,1), (hi,1), (to,2), (welcome,2), (the,2), (world,1))

scala>
```

4. **View the activity**
    - Open your browser and go to Spark UI:
        - http://localhost:4040
    - You can monitor job progress, stages, tasks, and storage here.



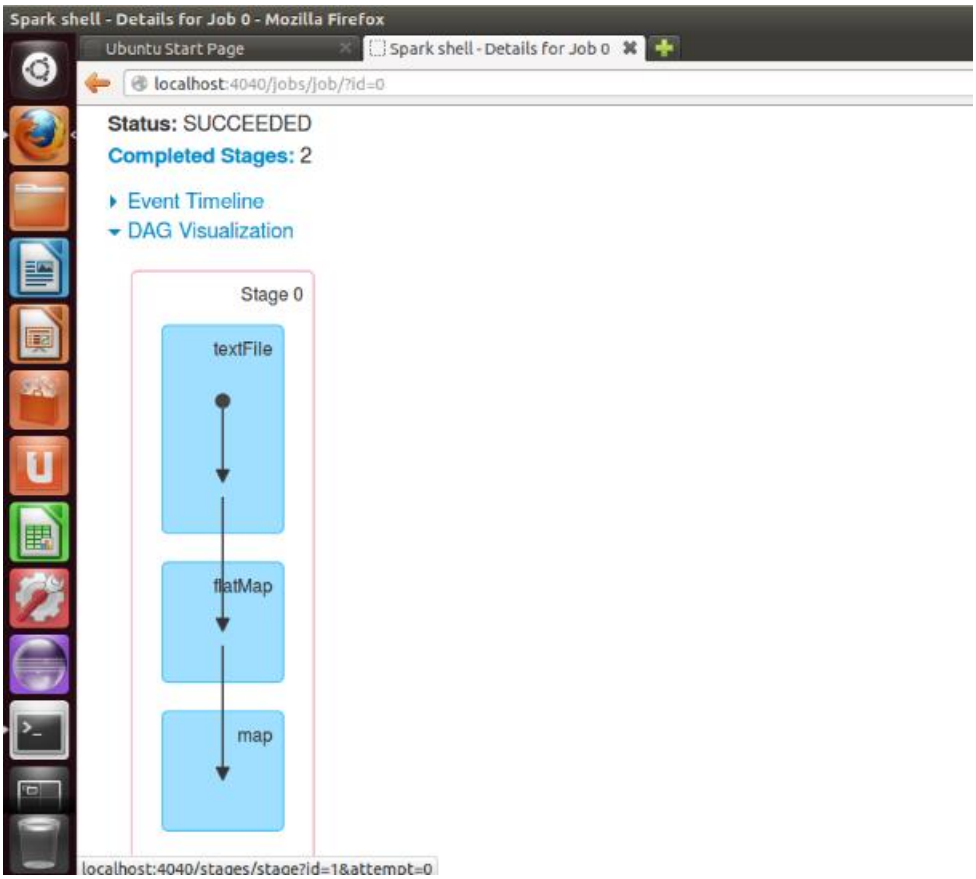
Details for Job 0

Status: SUCCEEDED
Completed Stages: 2

- Event Timeline
- DAG Visualization

Completed Stages (2)

Stage Id ▼	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
1	collect at <console>:29 +details	2025/03/24 11:37:03	57 ms	1/1			86.0 B	
0	map at <console>:24 +details	2025/03/24 11:37:02	0.2 s	1/1	43.0 B			86.0 B



Program (Word Count in Spark using Scala)

```
var a= sc.textFile("/home/ponny/Desktop/Ex1").flatMap(line => line.split("")).map(word => (word, 1))

var ba.reduceByKey(_+_);

b.collect
```

Output

For the sample text:

hi welcome to the lab welcome to the world

The output in Spark shell will be:

Array[(String, Int)] = Array((lab,1), (hi,1), (to,2), (welcome, 2), (the, 2), (world, 1))

```
scala> b.collect
res0: Array[(String, Int)] = Array((lab,1), (hi,1), (to,2), (welcome,2), (the,2), (world,1))
```

Result

The Word Count program was successfully executed using the Spark framework. The frequency of each word in the input file was counted and displayed as output.