

SPARK framework introduction

Go to the Bigdata User login

Pwd - hadoop

Now search for Oracle virtual box

Double Click Hadoop

Go to login,

Login - Ubuntu

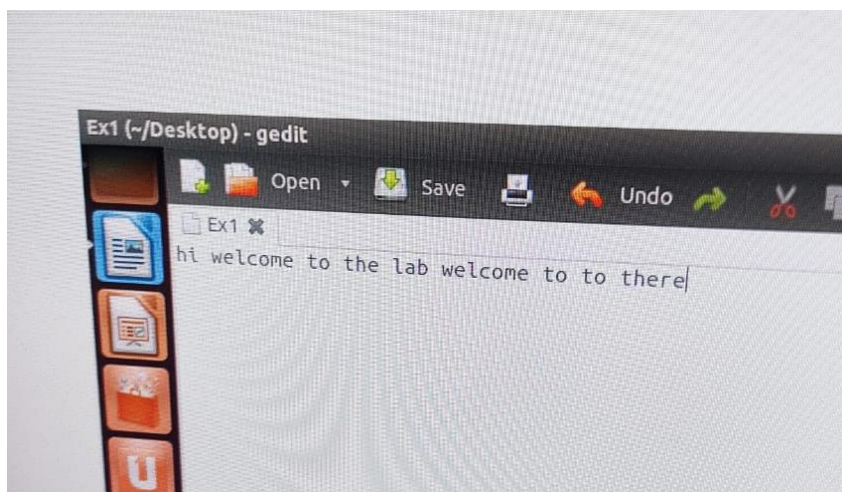
Passwd - vitcc123

Open terminal

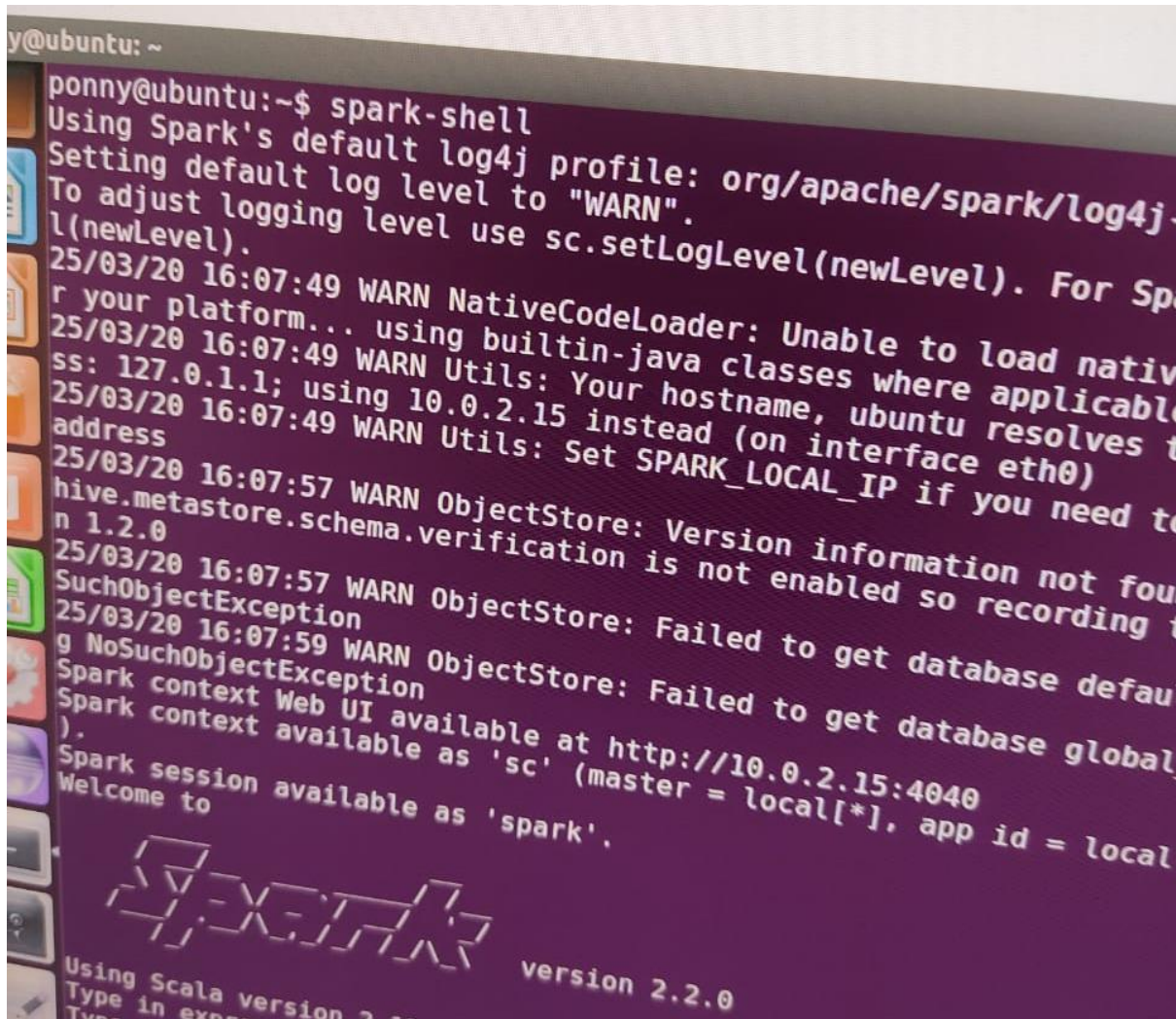
Type the command: spark-shell

Experiment 1: Word count program:

1. Create sample text file name “Ex1” and save it in desktop



2. Give the command: spark-shell, to check the spark framework version



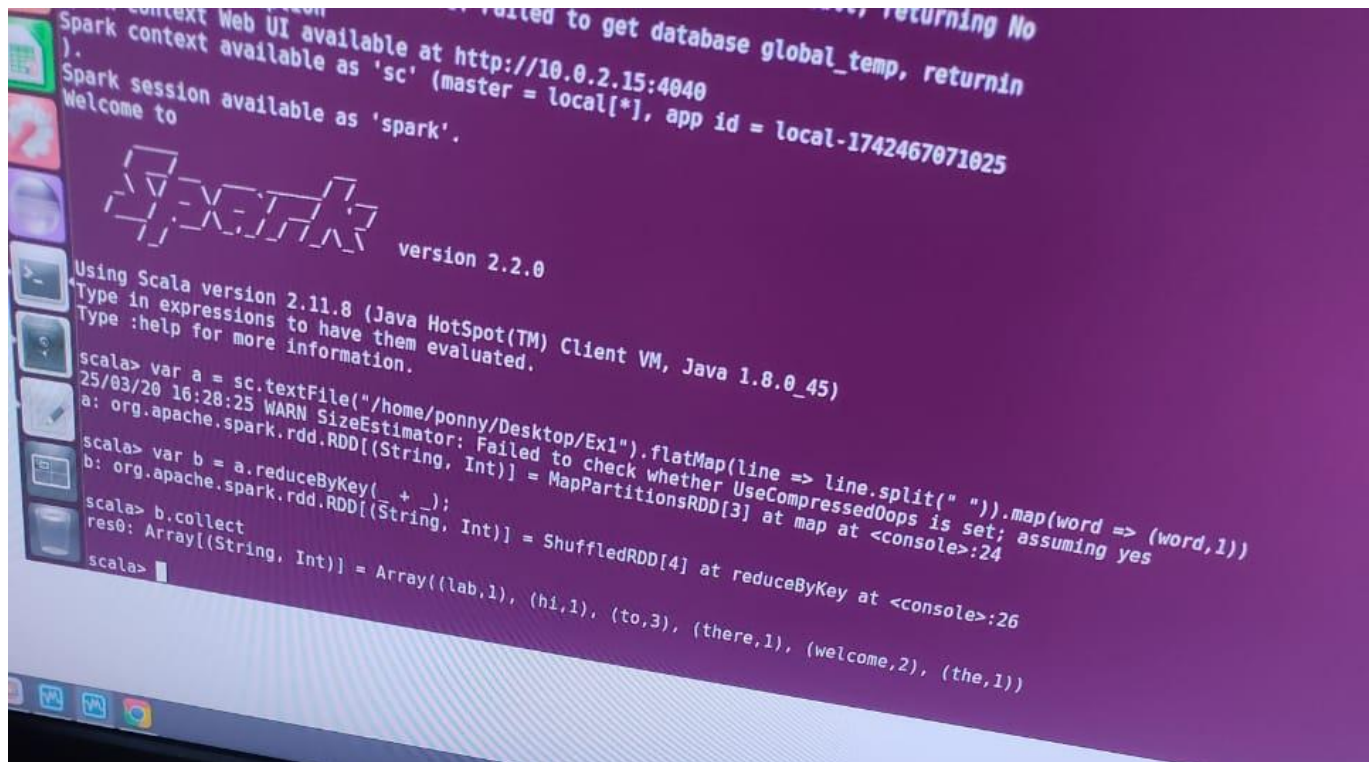
```
y@ubuntu: ~
ponny@ubuntu:~$ spark-shell
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.conf
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For Spark2, use
l(newLevel).
25/03/20 16:07:49 WARN NativeCodeLoader: Unable to load native code loader
r your platform... using builtin-java classes where applicable
25/03/20 16:07:49 WARN Utils: Your hostname, ubuntu resolves to 127.0.1.1; using 10.0.2.15 instead (on interface eth0)
25/03/20 16:07:49 WARN Utils: Set SPARK_LOCAL_IP if you need to specify the local address
25/03/20 16:07:57 WARN ObjectStore: Version information not found in 1.2.0
hive.metastore.schema verification is not enabled so recording schema change
25/03/20 16:07:57 WARN ObjectStore: Failed to get database default
SuchObjectException
25/03/20 16:07:59 WARN ObjectStore: Failed to get database global
g NoSuchObjectException
Spark context Web UI available at http://10.0.2.15:4040
Spark context available as 'sc' (master = local[*], app id = local-1503201607591234)
Spark session available as 'spark'.
Welcome to

Spark

Using Scala version 2.10.4
Type in expressions to interact with the interpreter
Type :help for more information
```

- Apache Spark is an open source, in-memory distributed computing engine created to address the problem of processing large datasets for data analytics and machine learning. Spark is written in Scala and it's native integration with Spark APIs

3. Now follow the commands using the Scala version, and check for the word count in the sample text file.



```
Spark context Web UI available at http://10.0.2.15:4040
Spark context available as 'sc' (master = local[*], app id = local-1742467071025)
Spark session available as 'spark'.
Welcome to

    ____
   / ____ \
  / /  _  \
 / /  | | | \
/ /___| | | |
\_____| |_| |
      \_____/

version 2.2.0

Using Scala version 2.11.8 (Java HotSpot(TM) Client VM, Java 1.8.0_45)
Type in expressions to have them evaluated.
Type :help for more information.

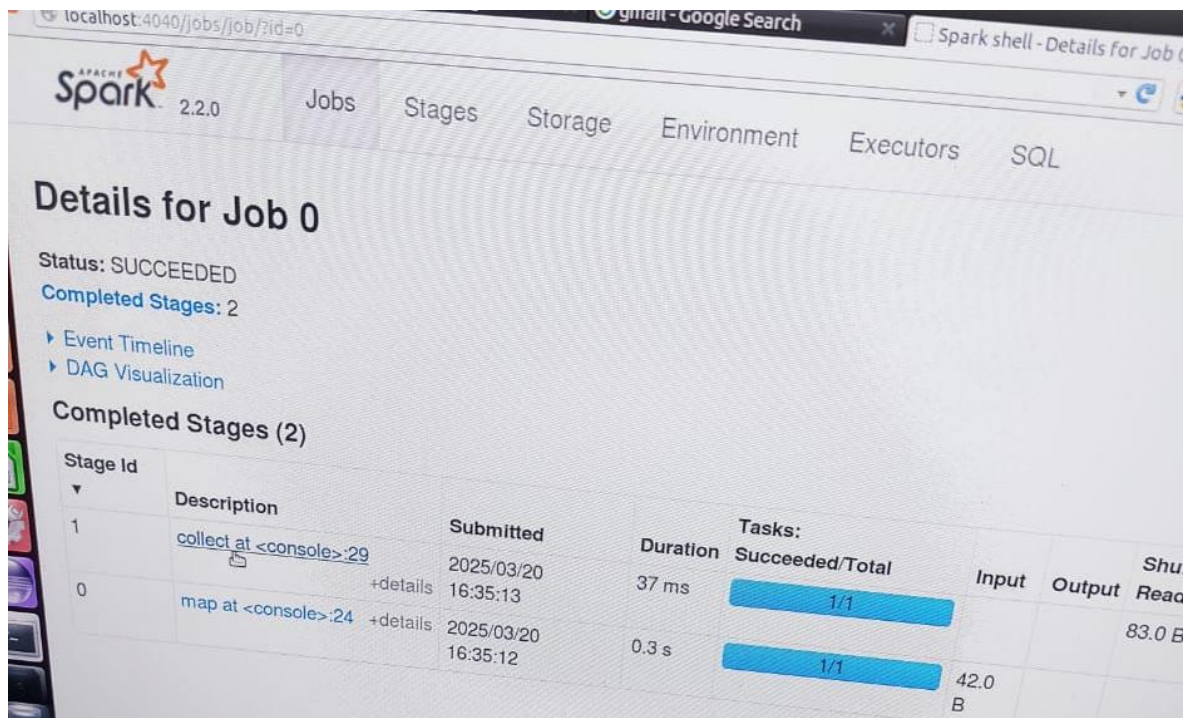
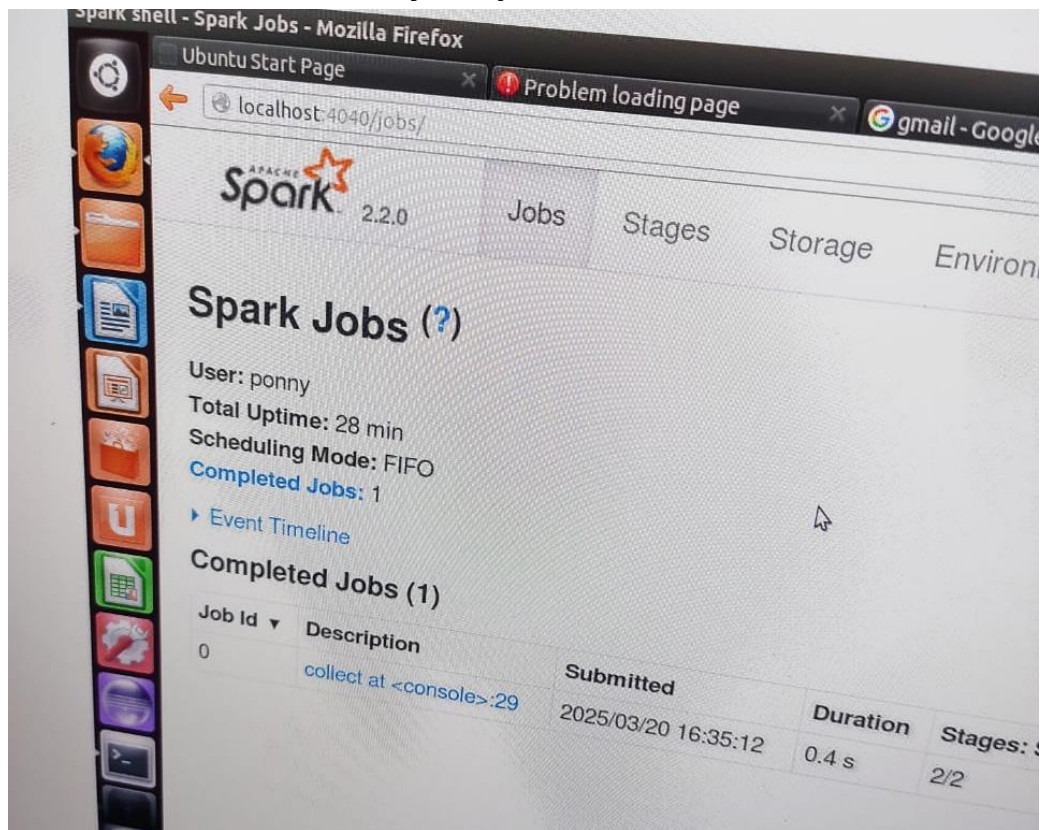
scala> var a = sc.textFile("/home/ponny/Desktop/Ex1").flatMap(line => line.split(" ")).map(word => (word,1))
25/03/20 16:28:25 WARN SizeEstimator: Failed to check whether UseCompressedOops is set; assuming yes
a: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[3] at map at <console>:24

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,3), (there,1), (welcome,2), (the,1))

scala>
```


4. U can check the activity in spark framework ---



Spark shell - Details for Job 0 - Mozilla Firefox

Ubuntu Start Page

localhost:4040/jobs/job/?id=0

Problem loading page

gmail - Google Search

Completed Stages: 2

- ▶ Event Timeline
- ▼ DAG Visualization

Stage 0

textFile

flatMap

map

Shows a graph of stages executed for this job, each of which can contain multiple RDD operations (e.g. map() and filter()), and of RDDs inside each operation (shown as dots).