Name: Tania Rajabally

Branch: Comps

Batch: C

Roll no: 43

UID:2017130047

Date: 21/10/2020

**Experiment 5**

**Aim:** Map Reduce Using Scala

**Theory:**

Scala is a modern multi-paradigm programming language designed to express common programming patterns in a concise, elegant, and type-safe way. Scala has been created by Martin Odersky and he released the first version in 2003. Scala smoothly integrates the features of object-oriented and functional languages. This tutorial explains the basics of Scala in a simple and reader-friendly way.

Scala Programming is based on Java, so if you are aware of Java syntax, then it's pretty easy to learn Scala. Further if you do not have expertise in Java but if you know any other programming language like C, C++ or Python then it will also help in grasping Scala concepts very quickly.

A MapReduce is a data processing tool which is used to process the data parallely in a distributed form. It was developed in 2004, on the basis of paper titled as "MapReduce: Simplified Data Processing on Large Clusters," published by Google.

The MapReduce is a paradigm which has two phases, the mapper phase, and the reducer phase. In the Mapper, the input is given in the form of a key-value pair. The output of the Mapper is fed to the reducer as input. The reducer runs only after the Mapper is over. The reducer too takes input in key-value format, and the output of reducer is the final output.

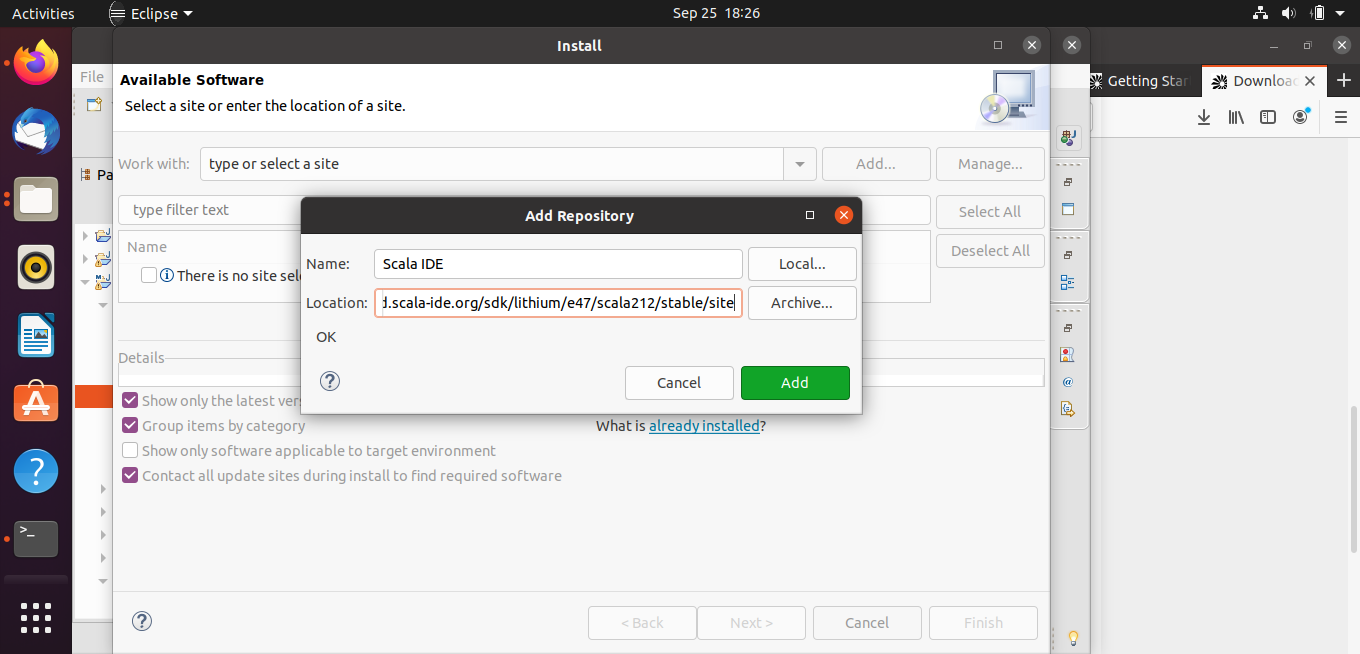
* The map takes data in the form of pairs and returns a list of <key, value> pairs. The keys will not be unique in this case.
* Using the output of Map, sort and shuffle are applied by the Hadoop architecture. This sort and shuffle acts on these list of <key, value> pairs and sends out unique keys and a list of values associated with this unique key <key, list(values)>.
* An output of sort and shuffle sent to the reducer phase. The reducer performs a defined function on a list of values for unique keys, and Final output <key, value> will be stored/displayed.

Usage of Map Reduce:

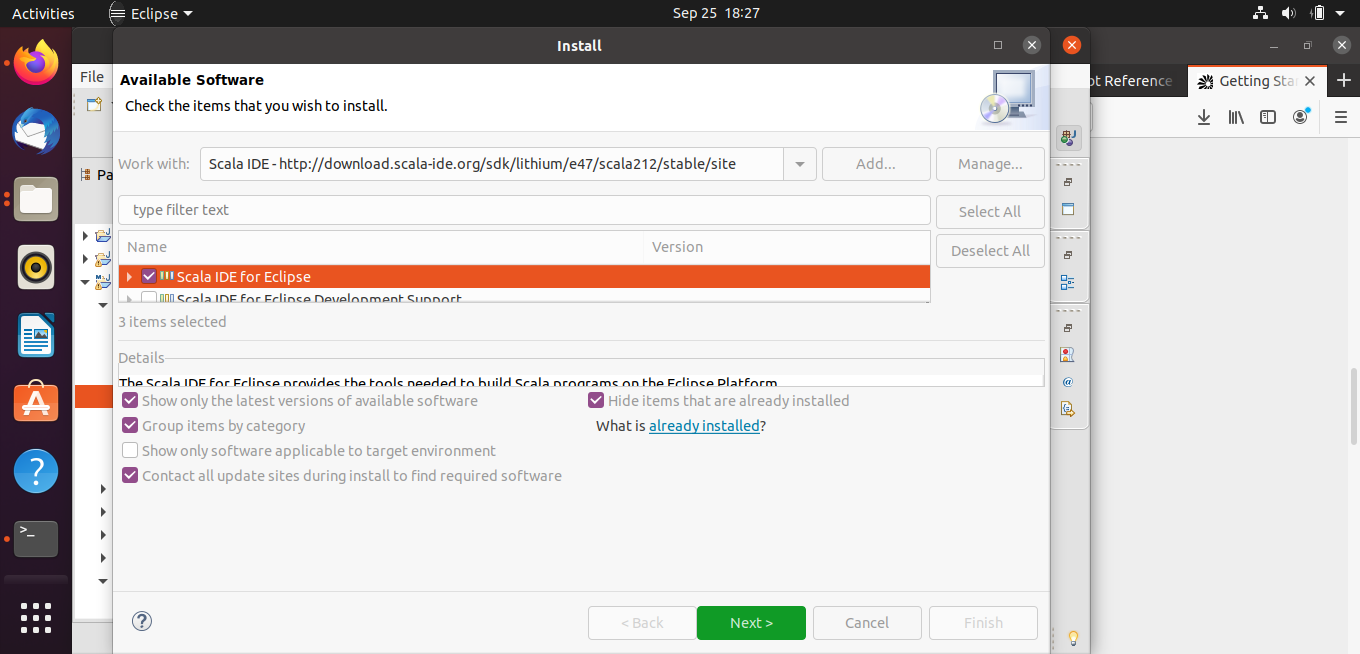
* It can be used in various applications like document clustering, distributed sorting, and web link-graph reversal.
* It can be used for distributed pattern-based searching.
* We can also use MapReduce in machine learning.
* It was used by Google to regenerate Google's index of the World Wide Web.
* It can be used in multiple computing environments such as multi-cluster, multi-core, and mobile environment.

**Steps:**

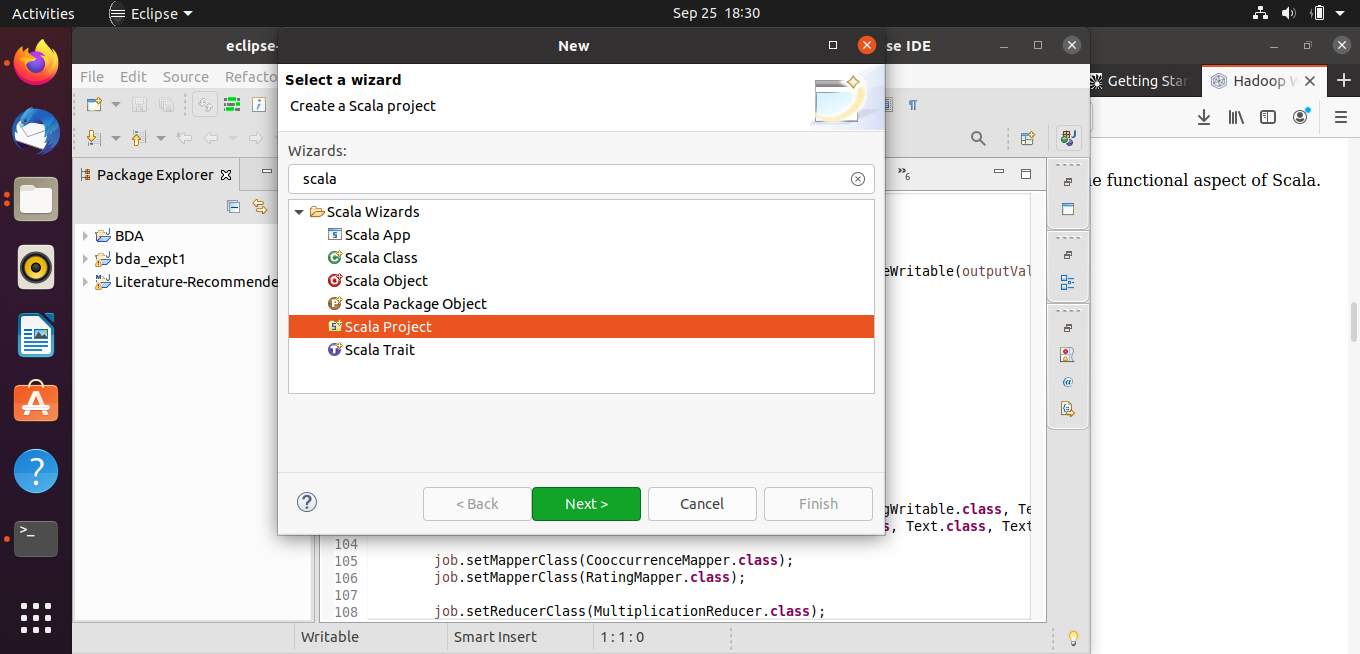
Add scala to eclipse

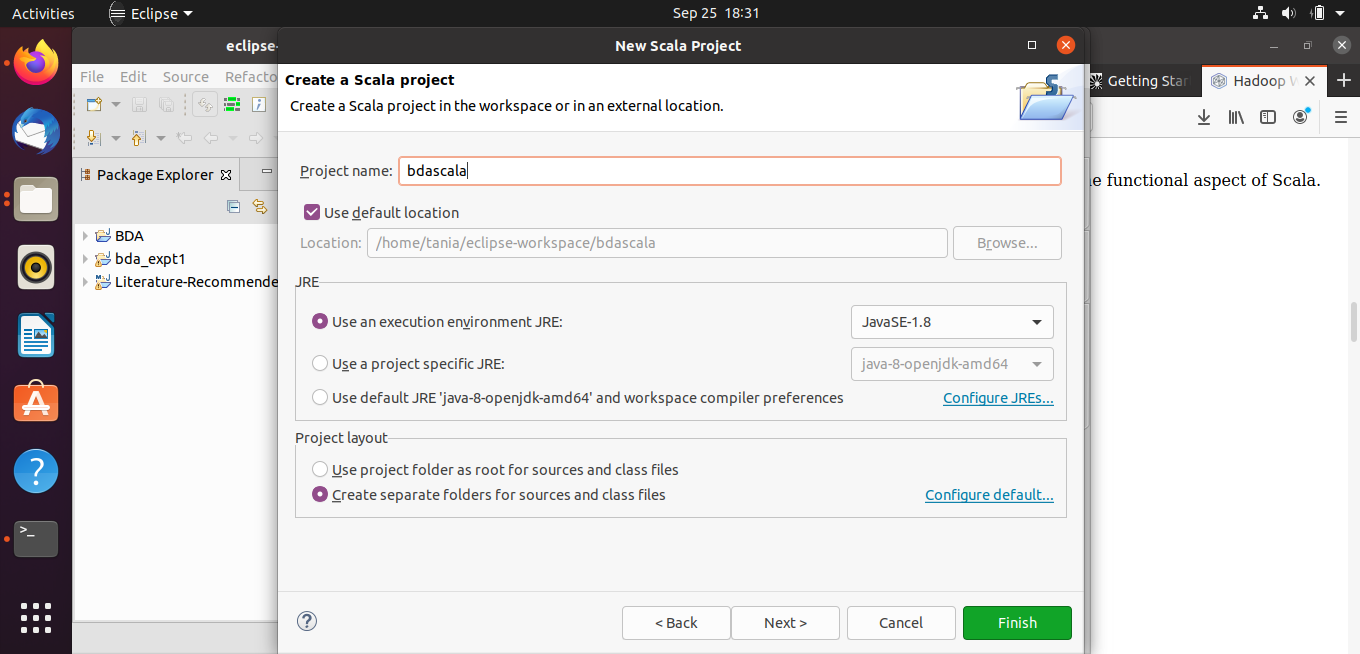


Choose Scala IDE for Eclipse

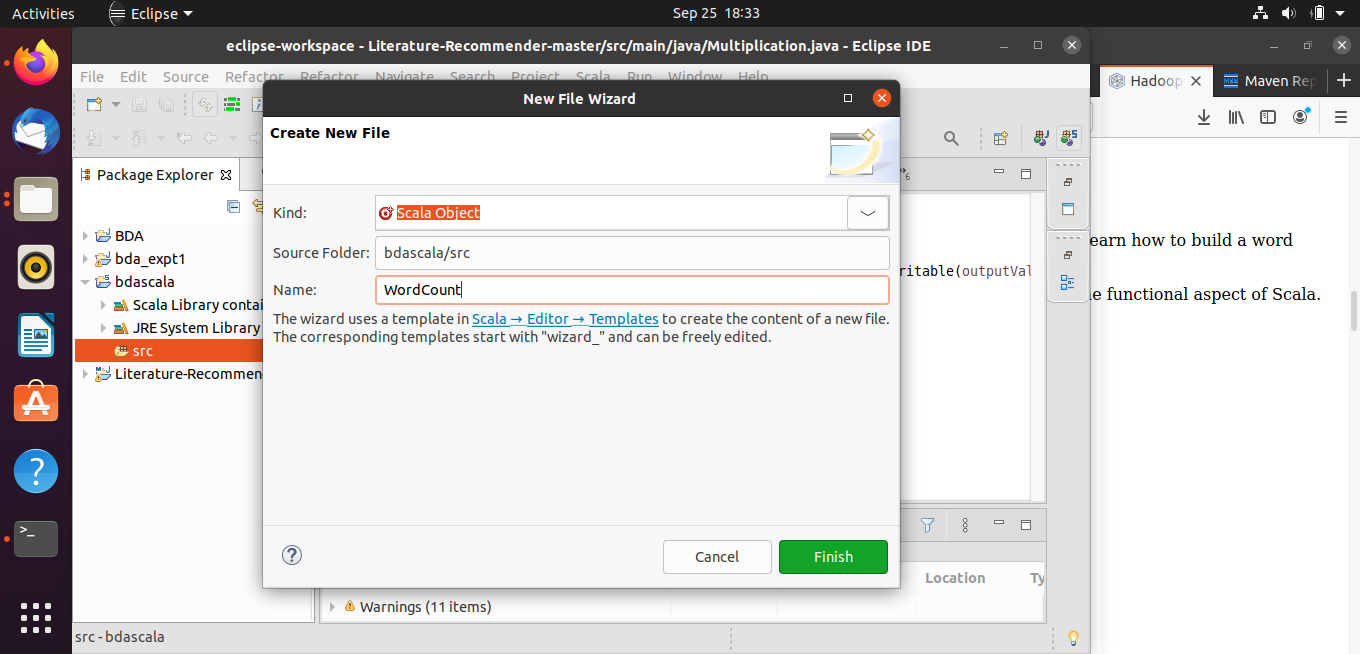
****

Create a Scala Project

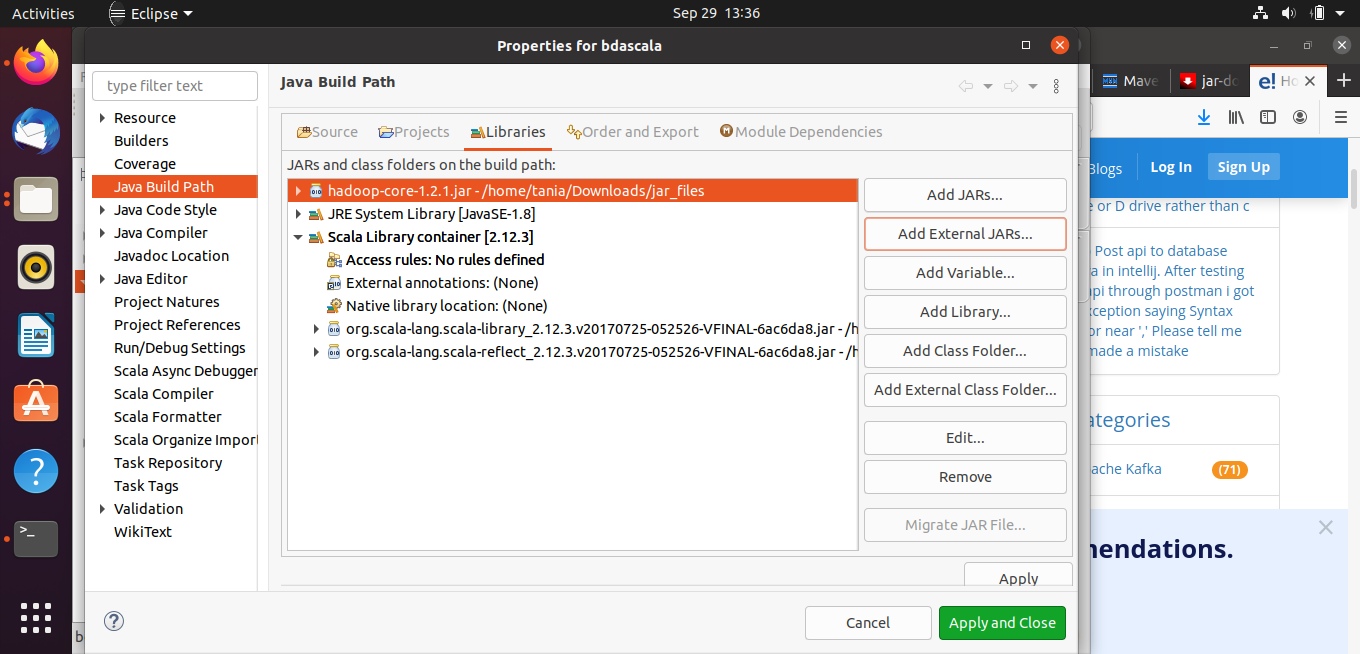
****

****

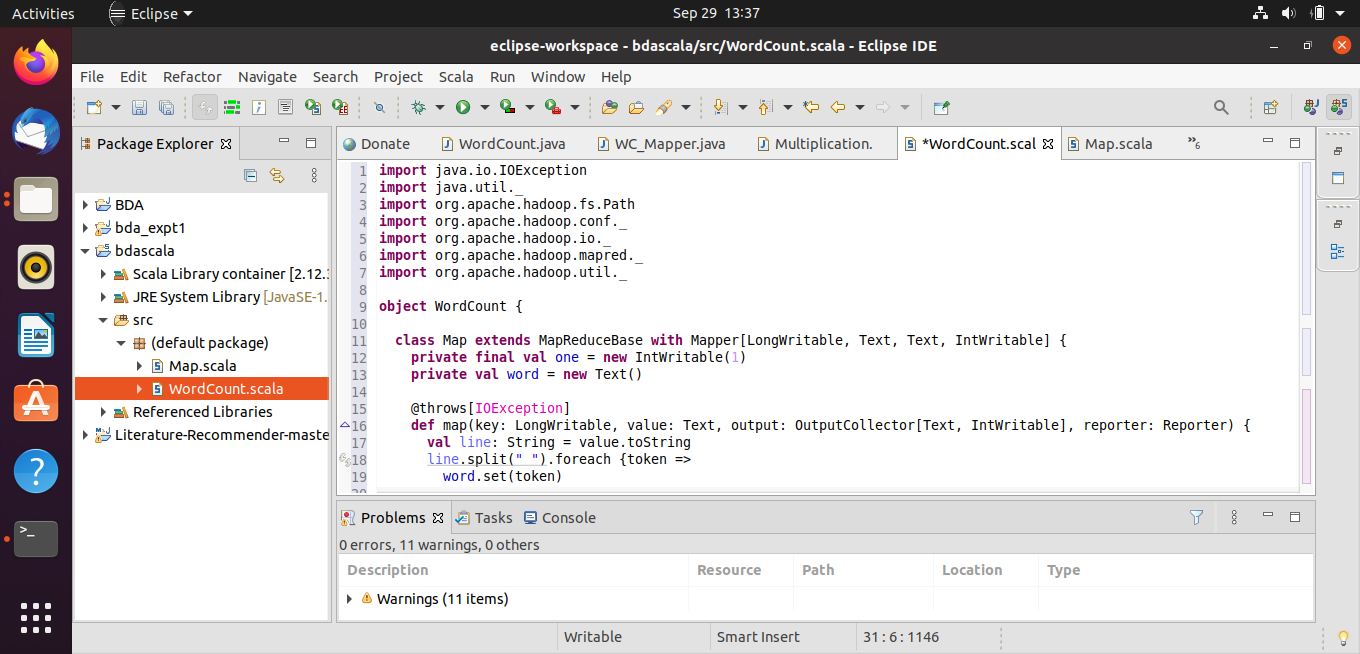
Create a Scala Object

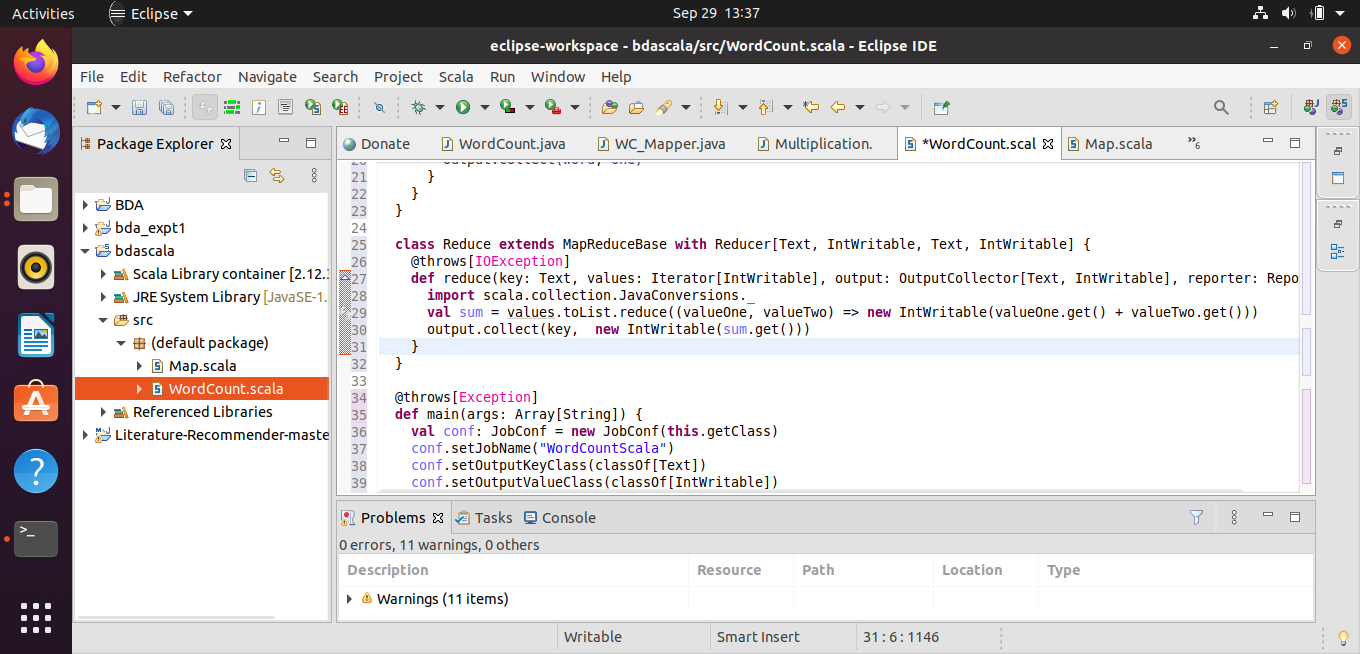
****

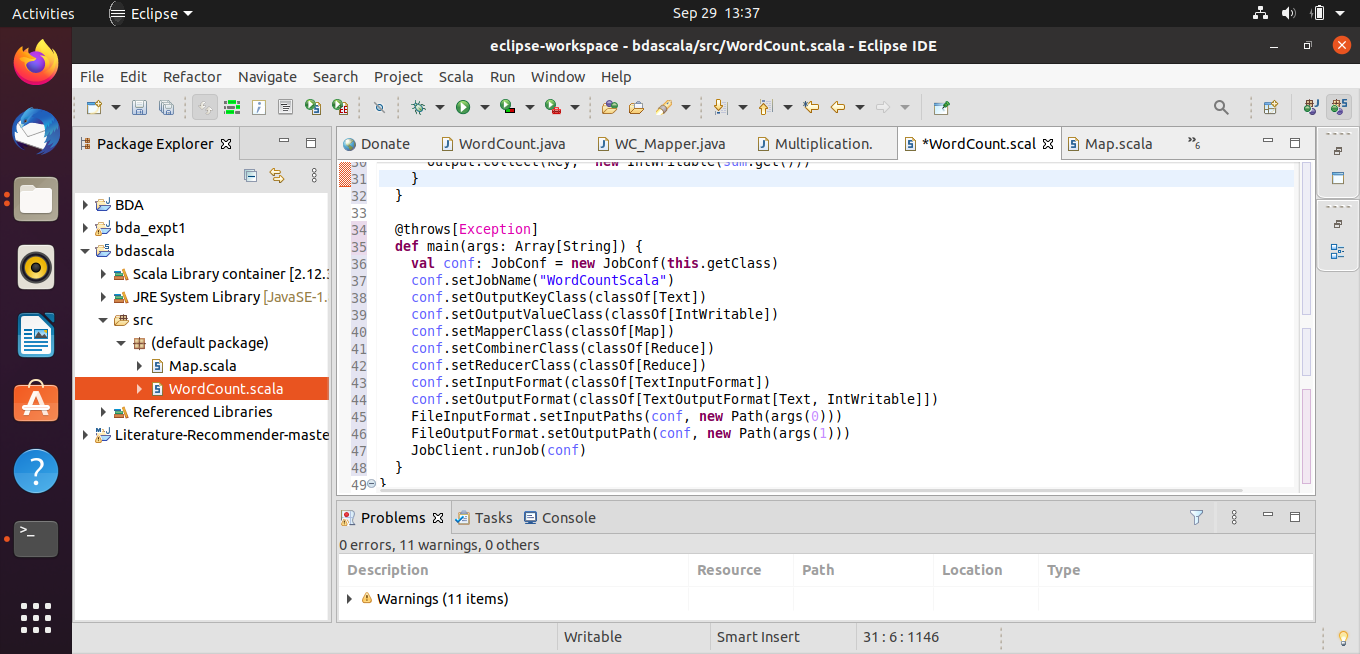
Import the external jar files

****

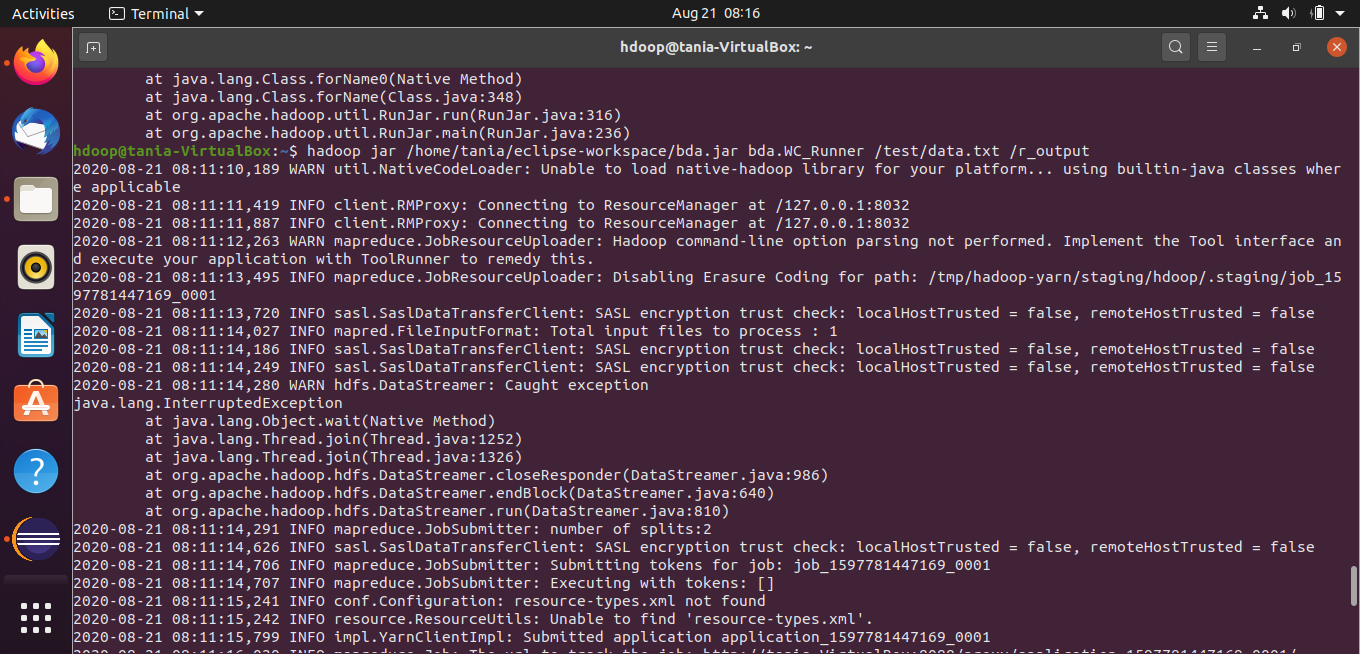
Write the WordCount Program

****

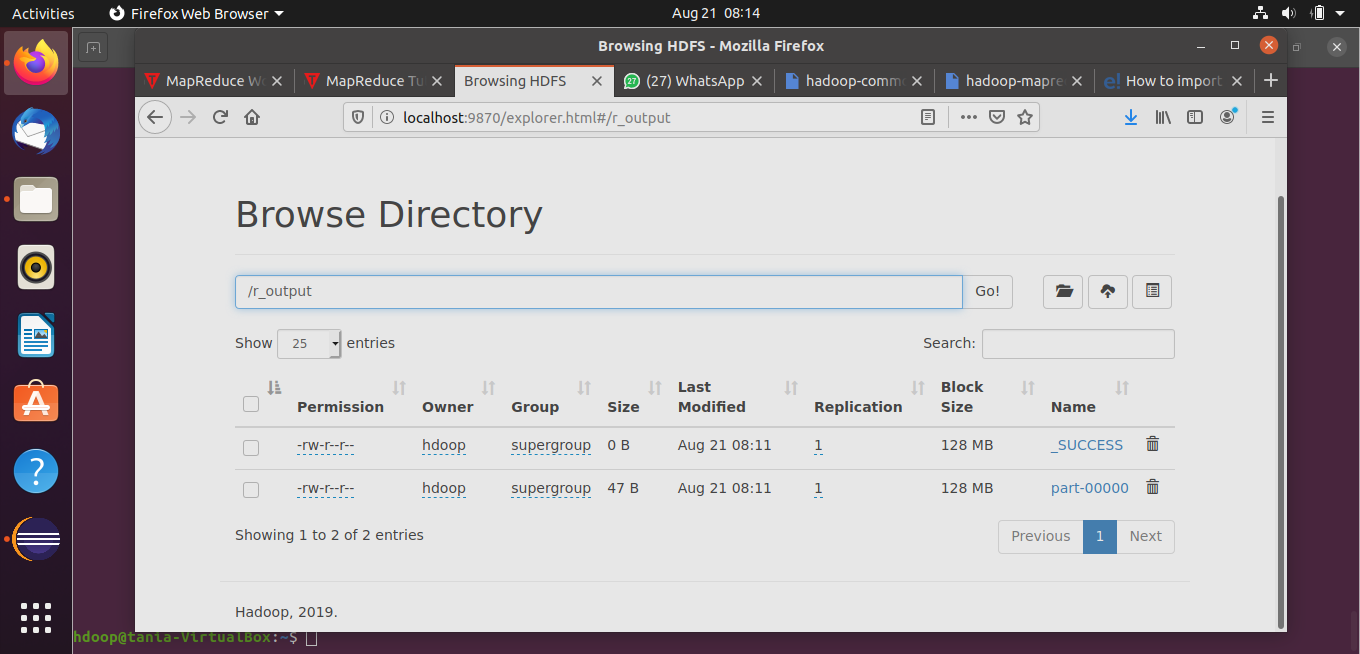
****

****

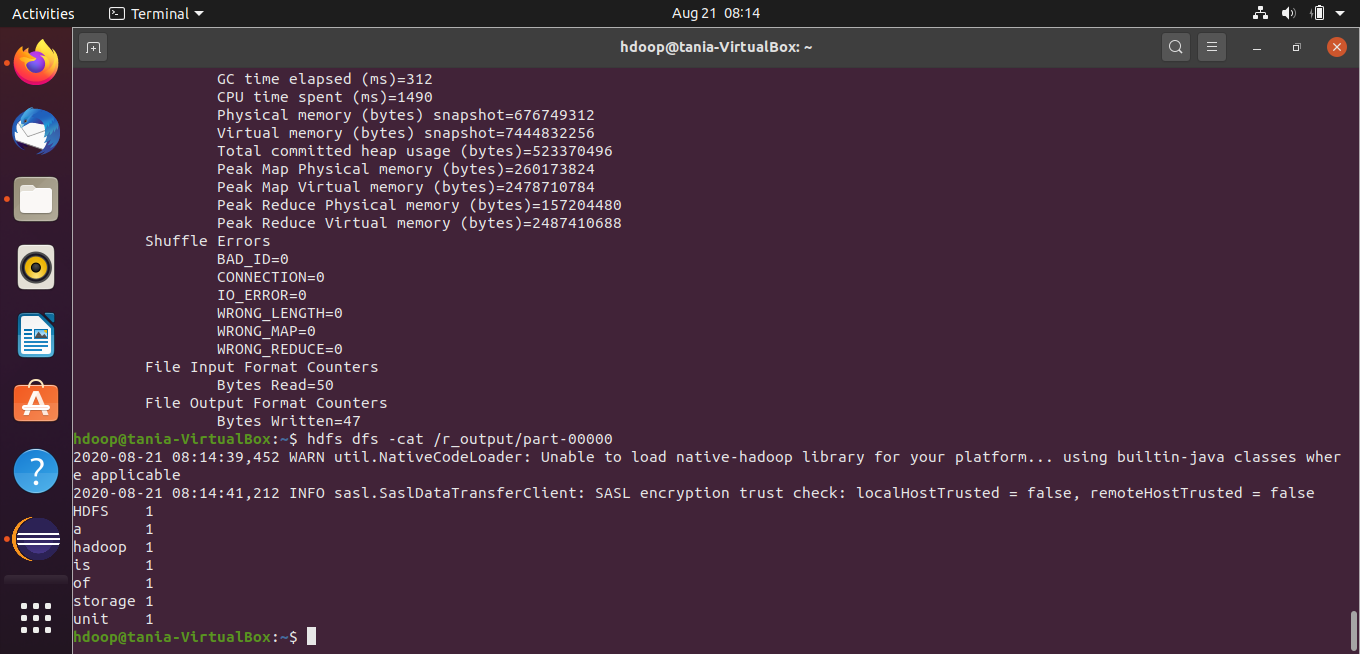
After creating the jar file, run it

****

Check in the hdfs webview if the file is created

****

See the output that is printed

****

**Conclusion:**

The MapReduce word count problem was implemented on the Hadoop system with Scala using Eclipse and jar files.

We had to import external jar files. We created a scala object and wrote the WordCount problem. We then created a jar file and ran it to get the output.