Name: Tania Rajabally

Branch: Comps

Batch: C

Roll no: 43

UID:2017130047

Date: 09/10/2020

**Experiment 6**

**Aim:** To install Apache Pig and perform some operations.

**Theory:**

Apache Pig is an abstraction over MapReduce. It is a tool/platform which is used to analyze larger sets of data representing them as data flows. Pig is generally used with Hadoop; we can perform all the data manipulation operations in Hadoop using Apache Pig.

To write data analysis programs, Pig provides a high-level language known as Pig Latin. This language provides various operators using which programmers can develop their own functions for reading, writing, and processing data.

To analyze data using Apache Pig, programmers need to write scripts using Pig Latin language. All these scripts are internally converted to Map and Reduce tasks. Apache Pig has a component known as Pig Engine that accepts the Pig Latin scripts as input and converts those scripts into MapReduce jobs.

Programmers who are not so good at Java normally used to struggle working with Hadoop, especially while performing any MapReduce tasks. Apache Pig is a boon for all such programmers.

* Using Pig Latin, programmers can perform MapReduce tasks easily without having to type complex codes in Java.
* Apache Pig uses multi-query approach, thereby reducing the length of codes. For example, an operation that would require you to type 200 lines of code (LoC) in Java can be easily done by typing as less as just 10 LoC in Apache Pig. Ultimately Apache Pig reduces the development time by almost 16 times.
* Pig Latin is SQL-like language and it is easy to learn Apache Pig when you are familiar with SQL.

Apache Pig comes with the following features −

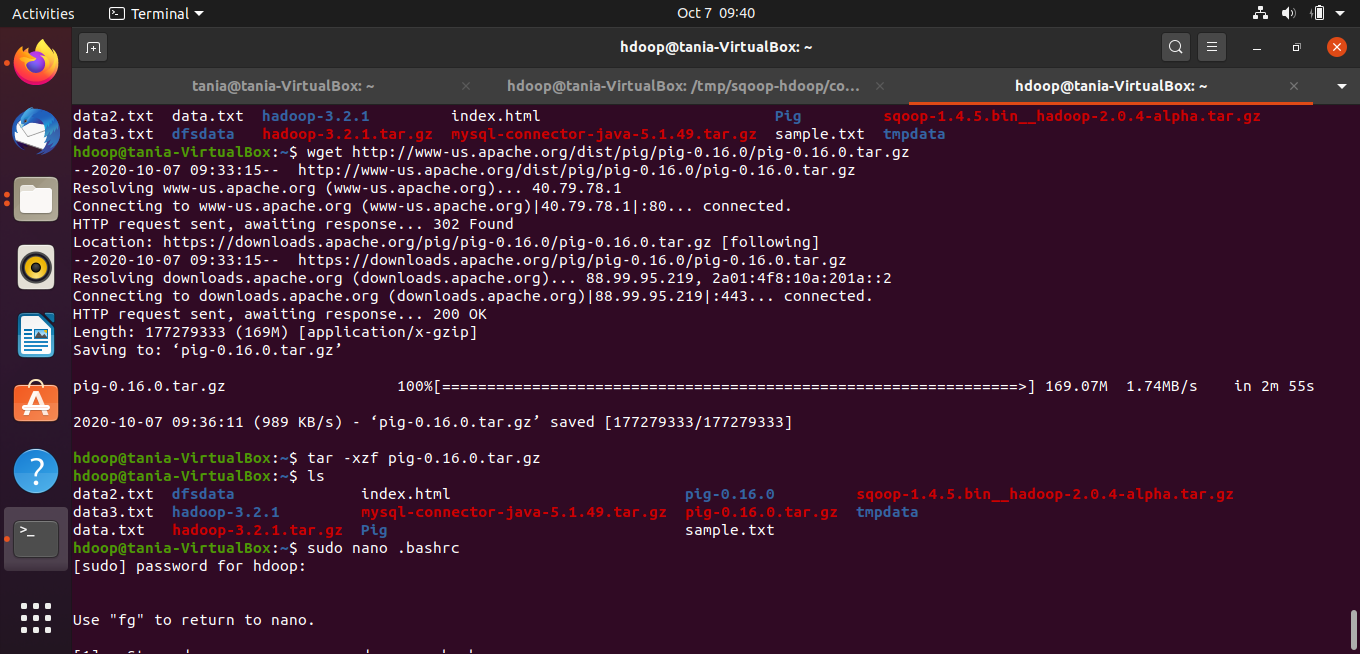
* Rich set of operators − It provides many operators to perform operations like join, sort, filer, etc.
* Ease of programming − Pig Latin is similar to SQL and it is easy to write a Pig script if you are good at SQL.
* Optimization opportunities − The tasks in Apache Pig optimize their execution automatically, so the programmers need to focus only on semantics of the language.
* Extensibility − Using the existing operators, users can develop their own functions to read, process, and write data.
* UDF’s − Pig provides the facility to create User-defined Functions in other programming languages such as Java and invoke or embed them in Pig Scripts.
* Handles all kinds of data − Apache Pig analyzes all kinds of data, both structured as well as unstructured. It stores the results in HDFS.

**Steps:**

Step 1: Install Pig

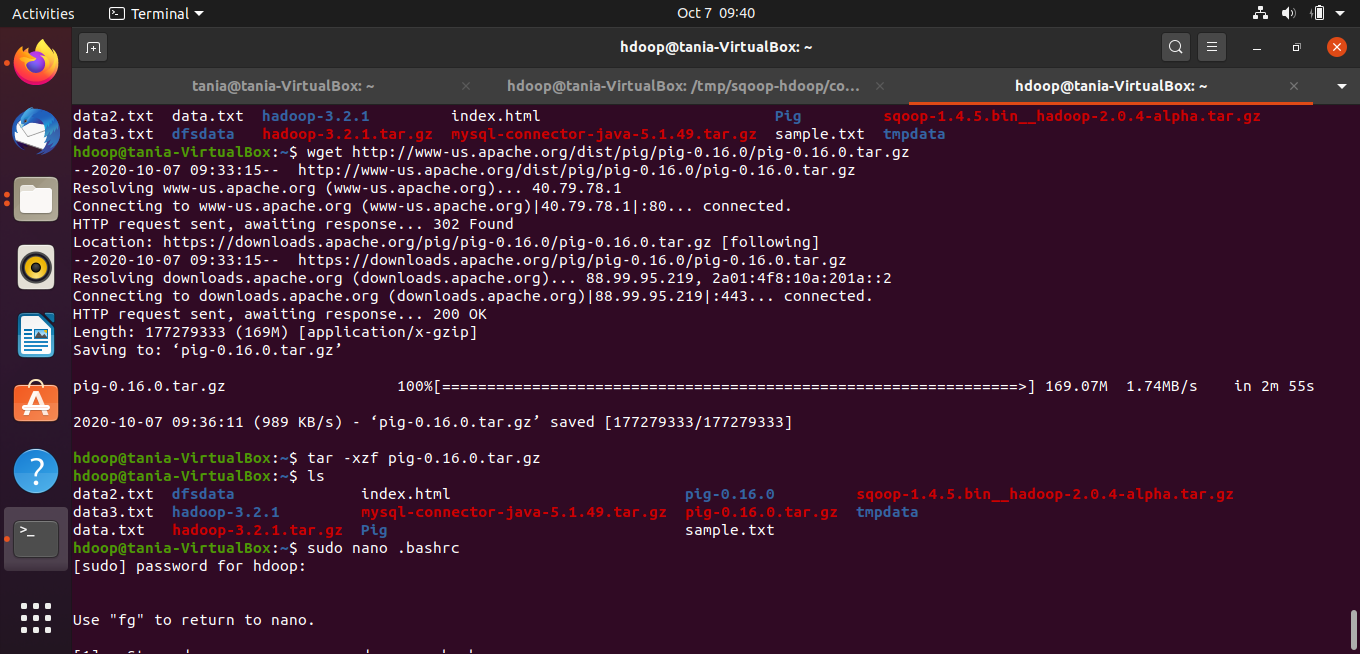
Step 1.1: Download Pig tar file.

Command: wget <http://www-us.apache.org/dist/pig/pig-0.16.0/pig-0.16.0.tar.gz>

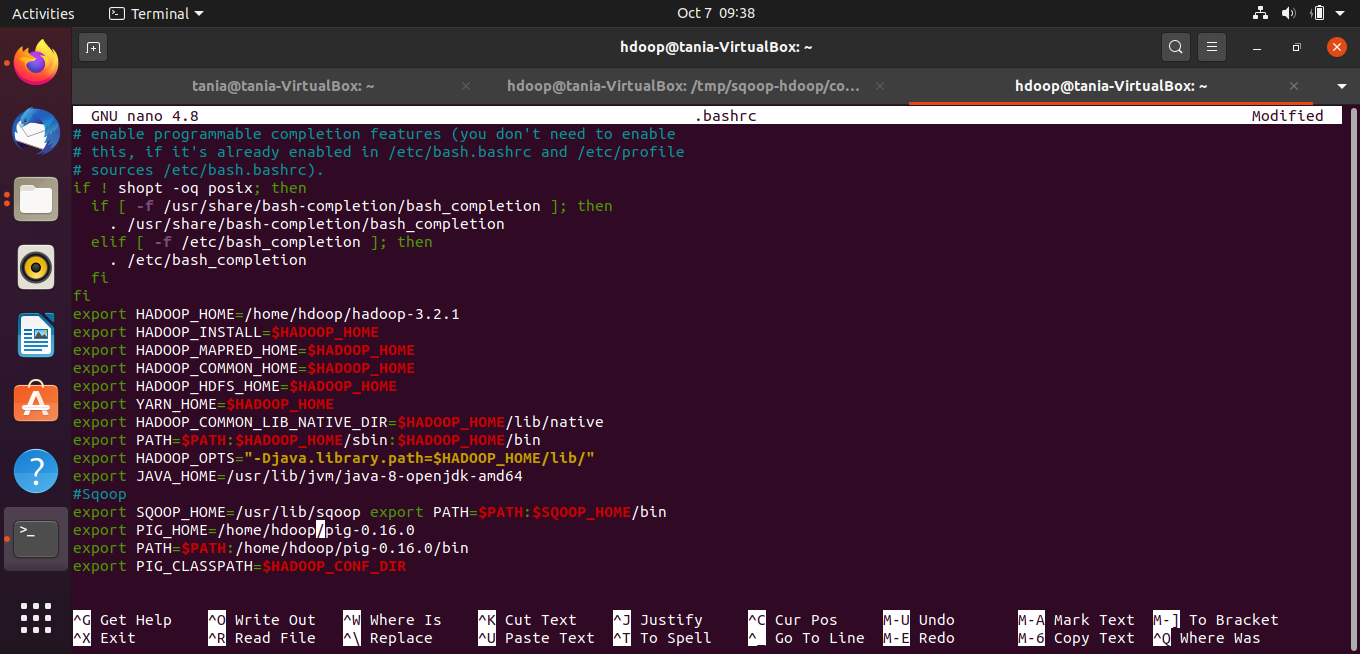


Step 1.2: Extract the tar file using tar command. In below tar command, x means extract an archive file, z means filter an archive through gzip, f means filename of an archive file.

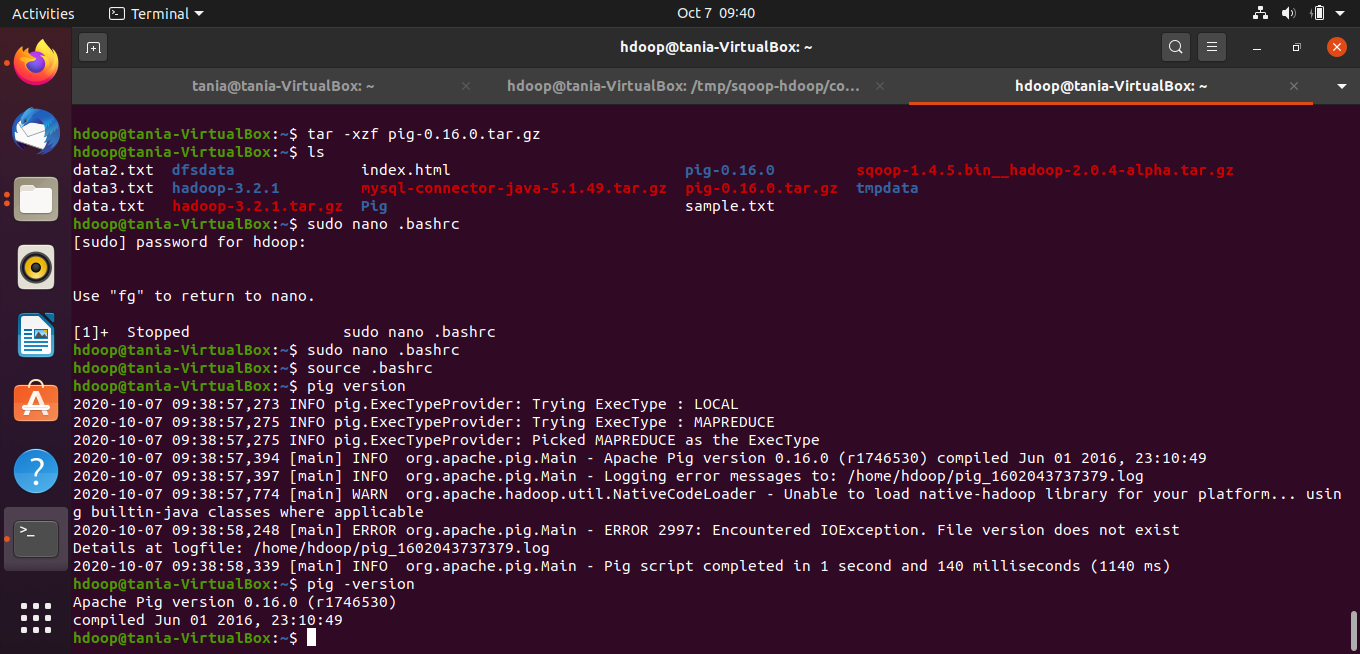
Command: tar -xzf pig-0.16.0.tar.gz



Step 1.3 : Edit the “.bashrc” file to update the environment variables of Apache Pig. We are setting it so that we can access pig from any directory, we need not go to pig directory to execute pig commands. Also, if any other application is looking for Pig, it will get to know the path of Apache Pig from this file.



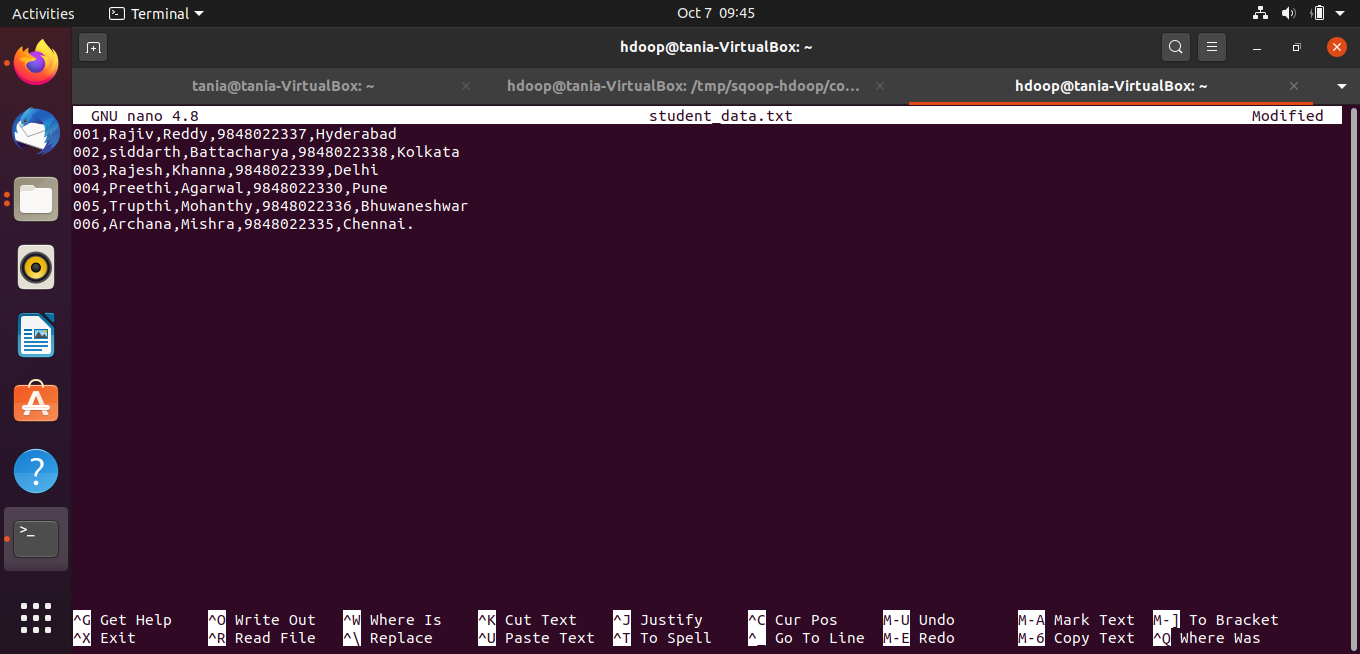
Step 1.4: Check pig version.



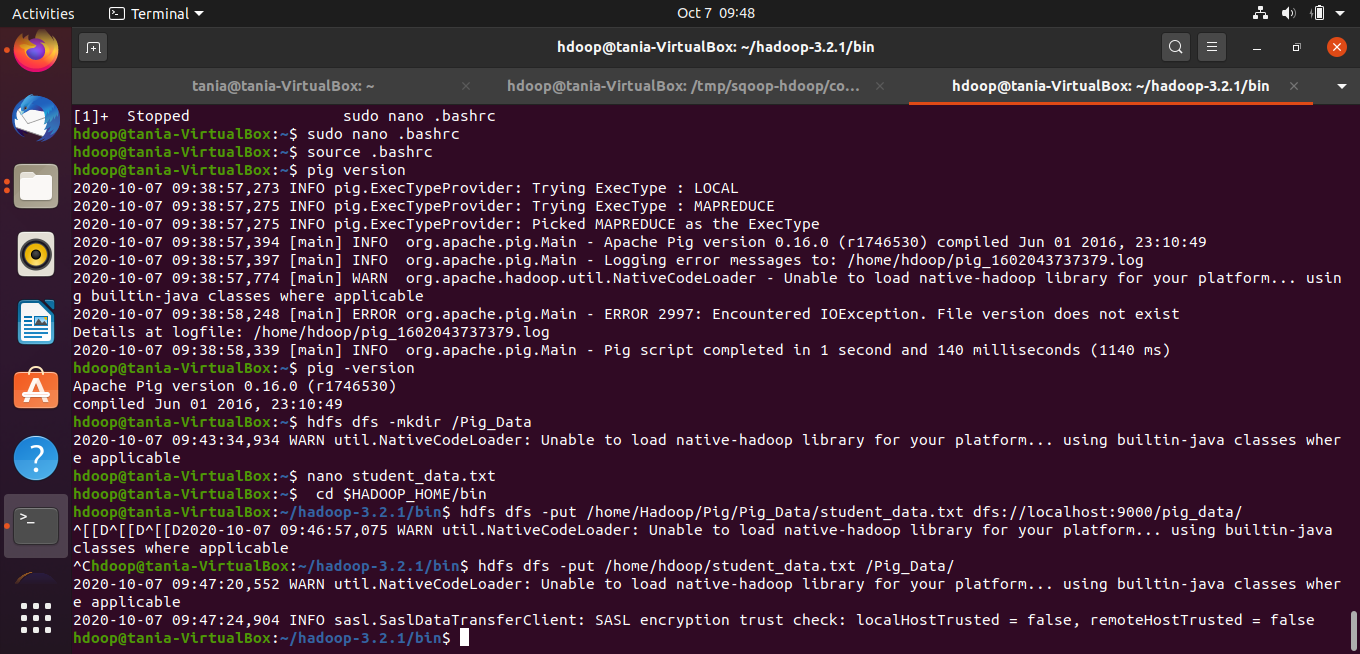
Step 2: Operations in Pig

Step 2.1 : Reading Data

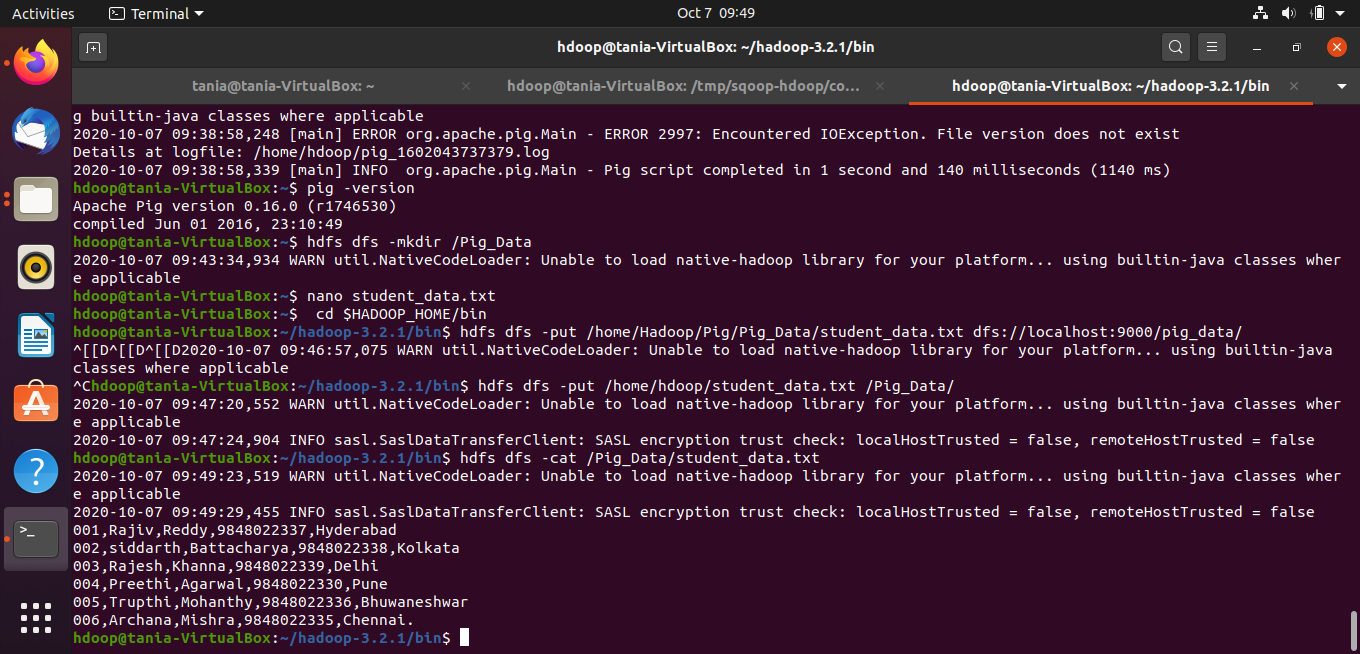
Create a txt file



Create a directory in HDFS and place the data in HDFS



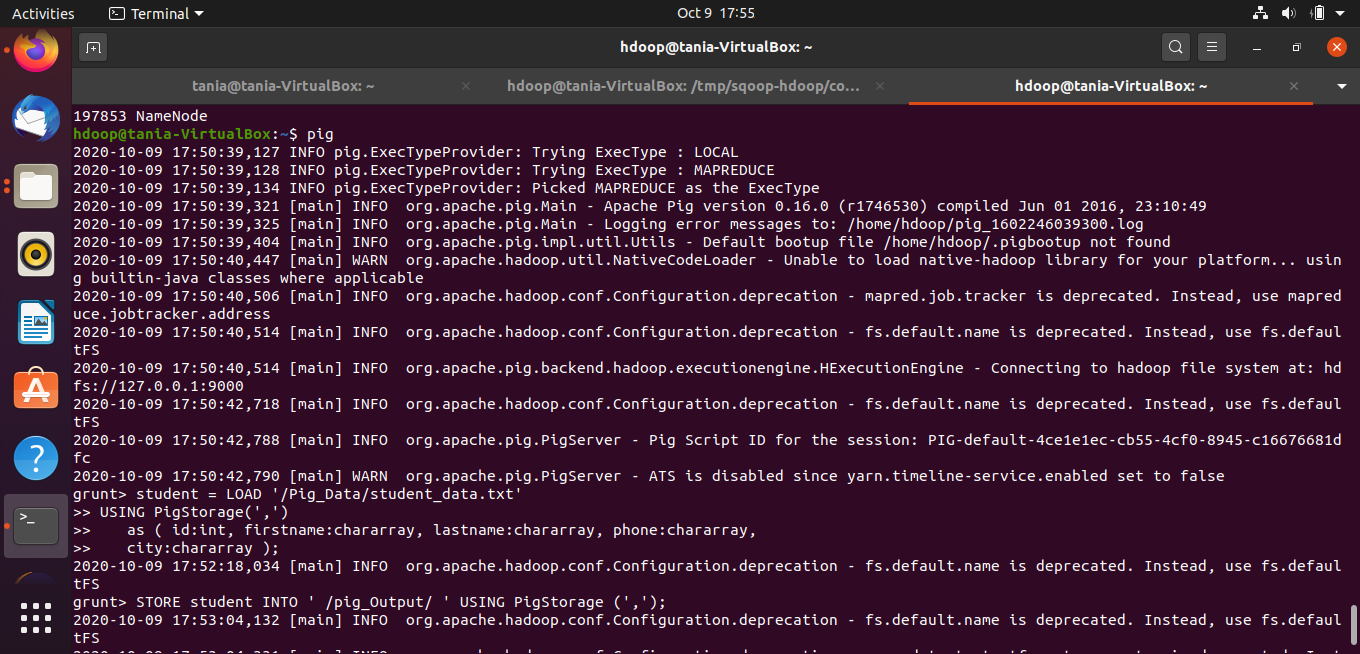
Verifying the output



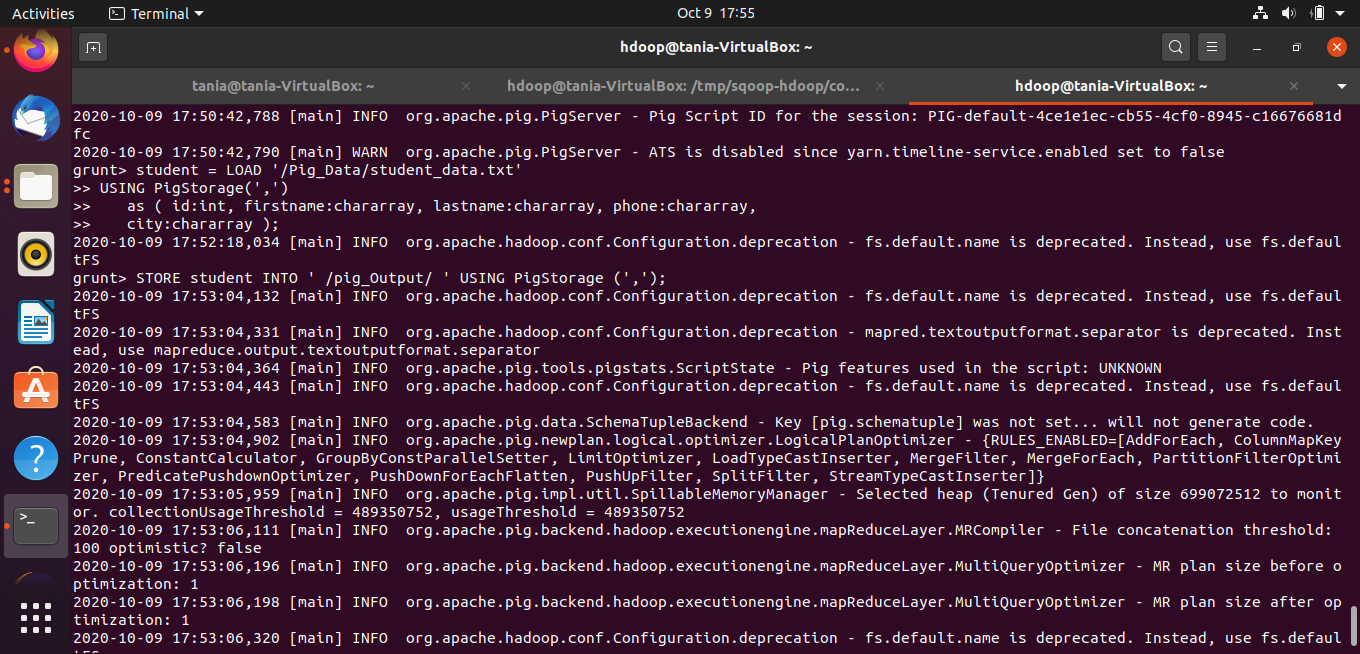


Step 2.2 : Loading and storing the data

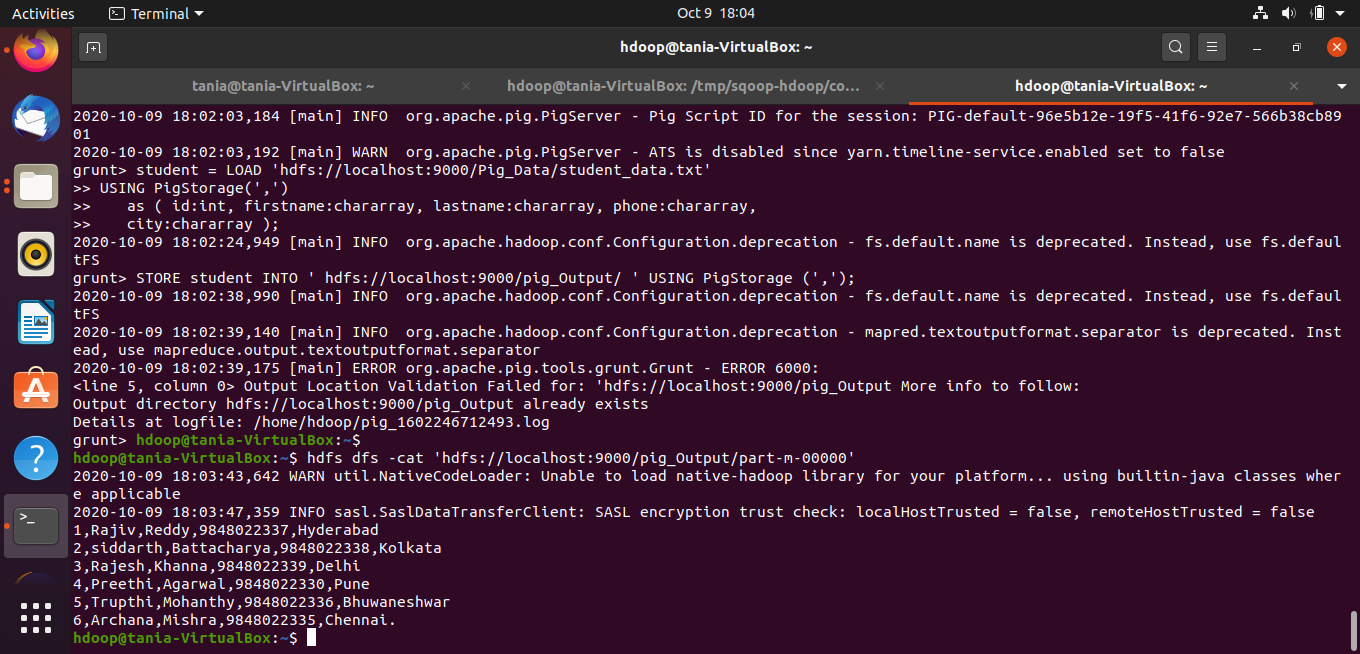
Read the txt file into a relation student using the LOAD operator



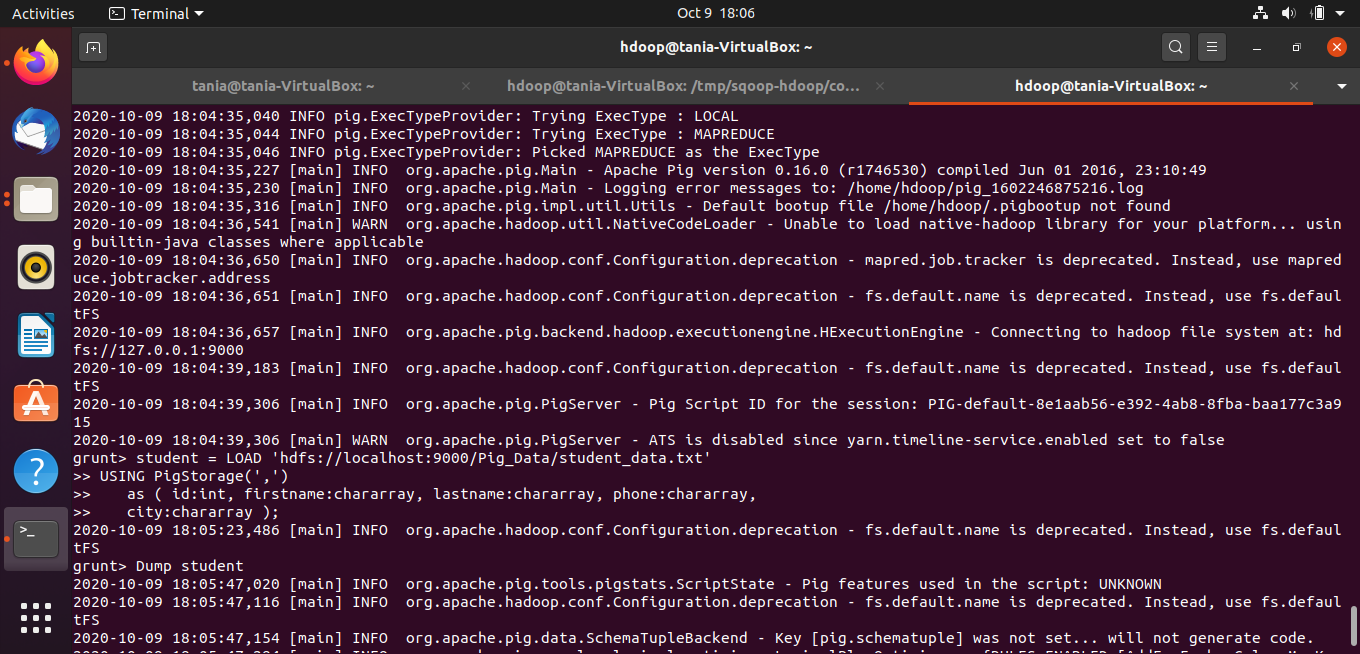
Store the relation in the HDFS directory “/pig\_Output/”

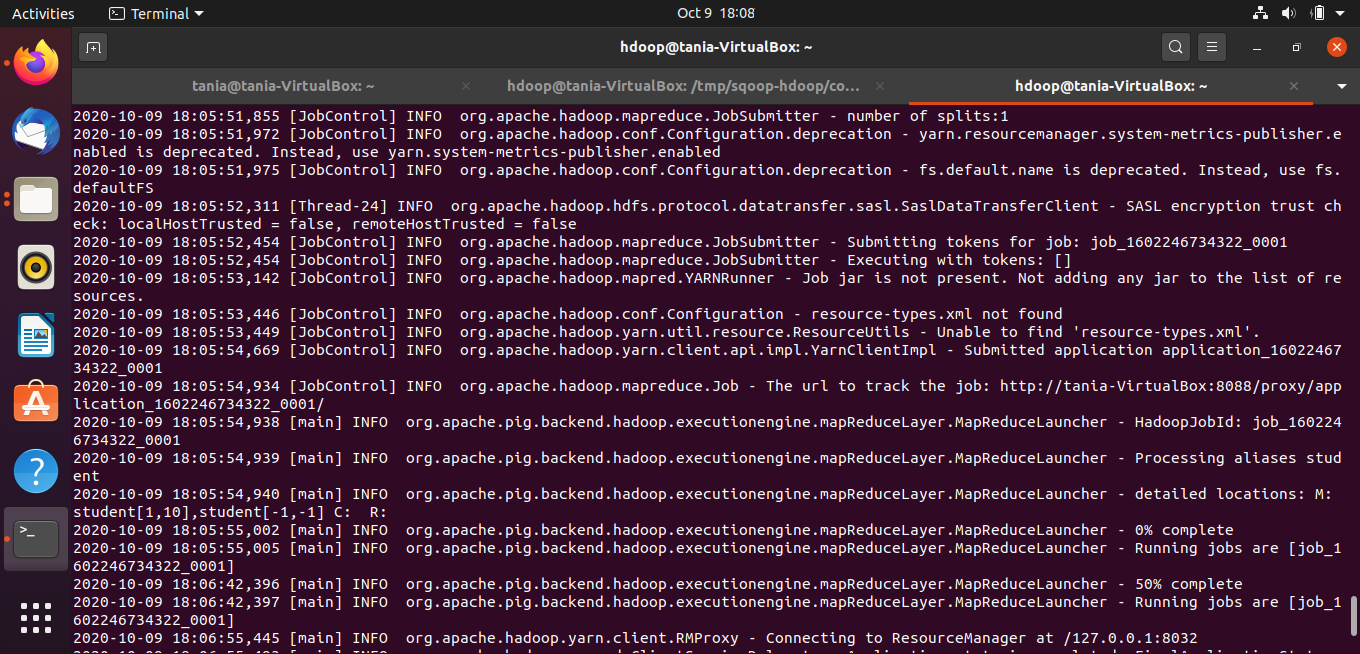


Display the output

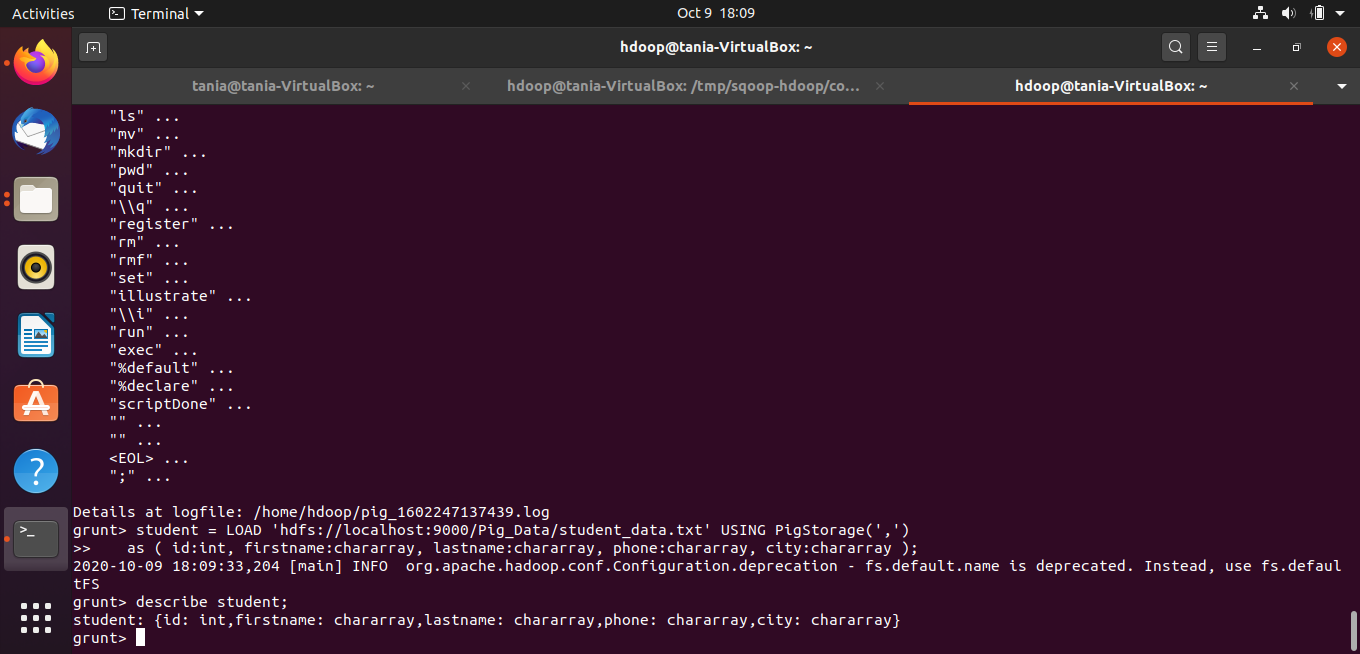


Step 2.3 : Dump Operator

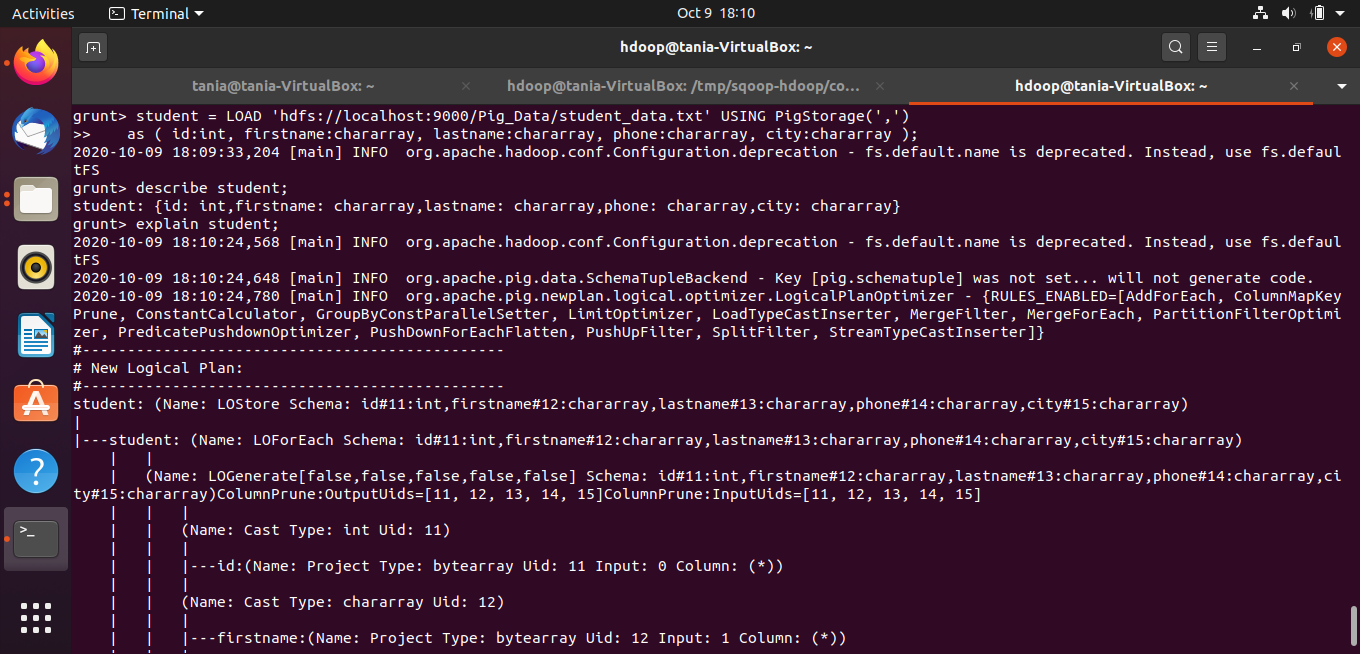


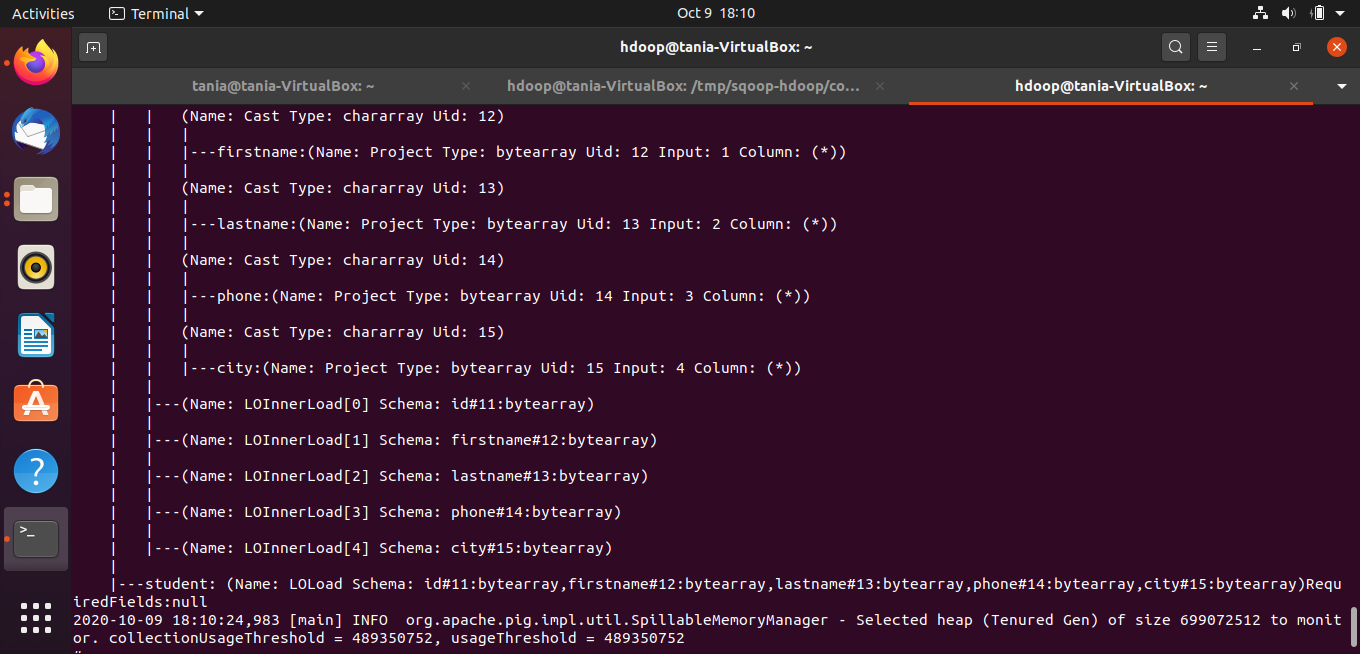


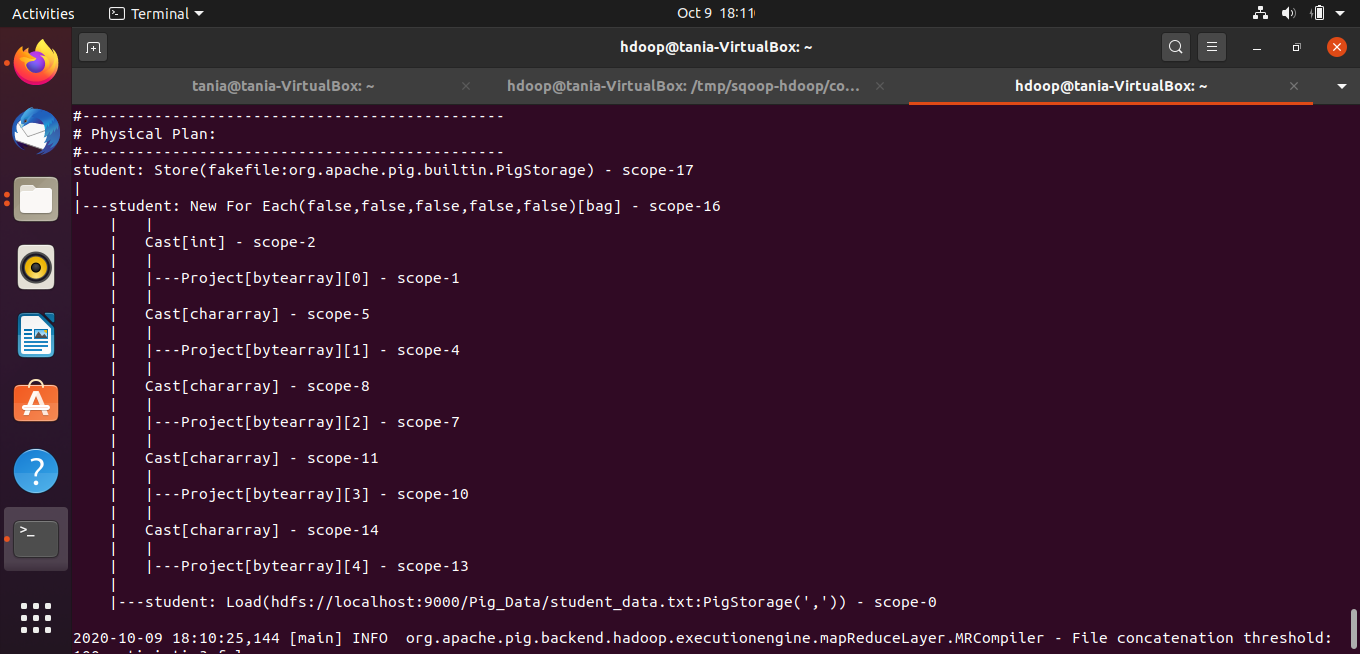
Step 2.4 : Describe Operator



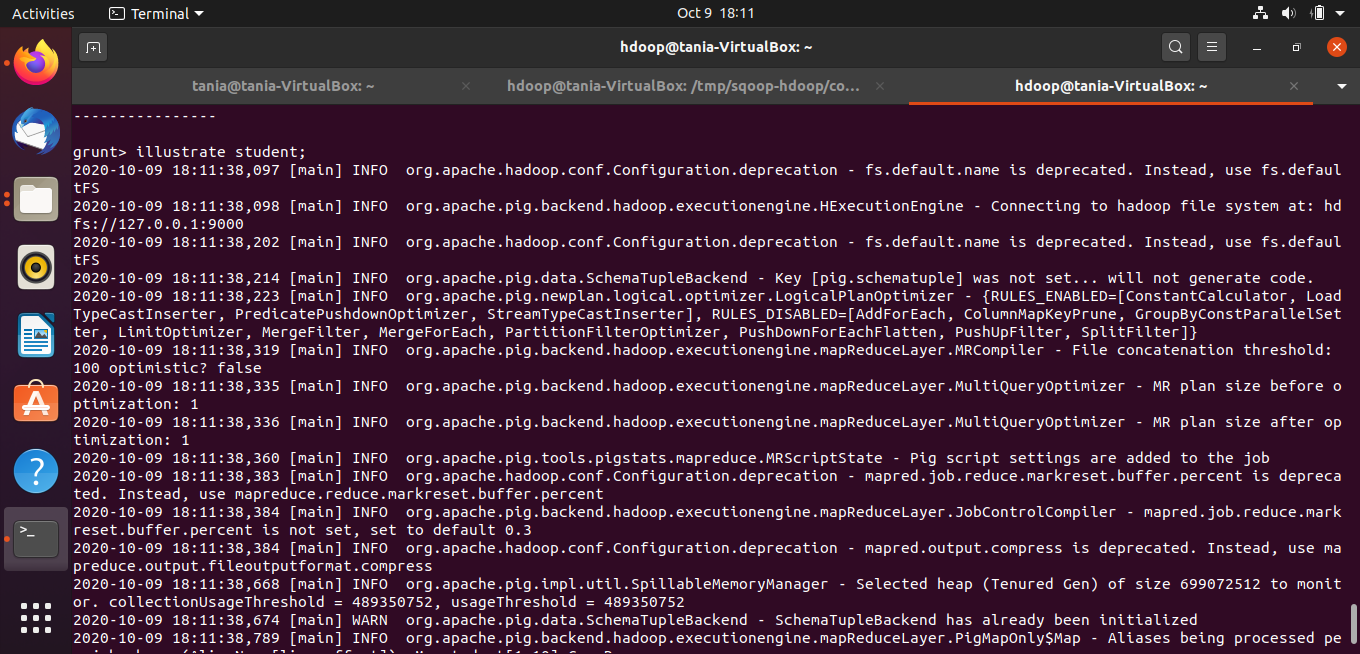
Step 2.5 : Explain Operator

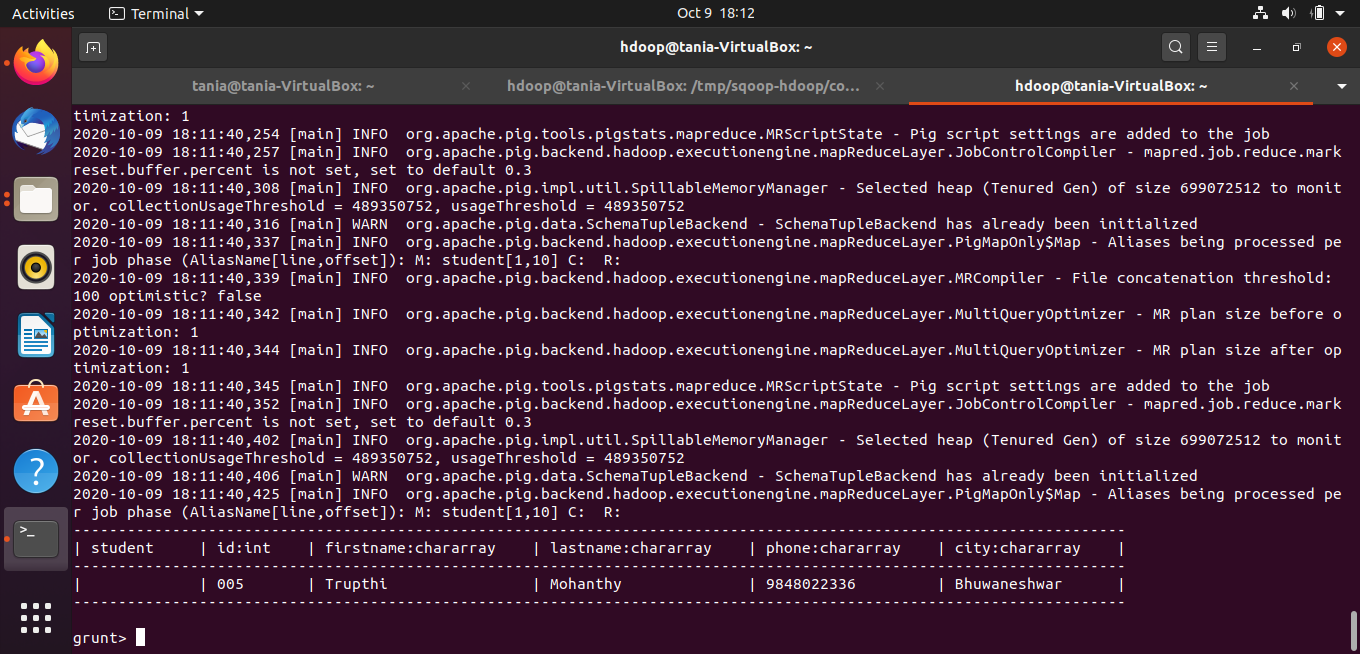




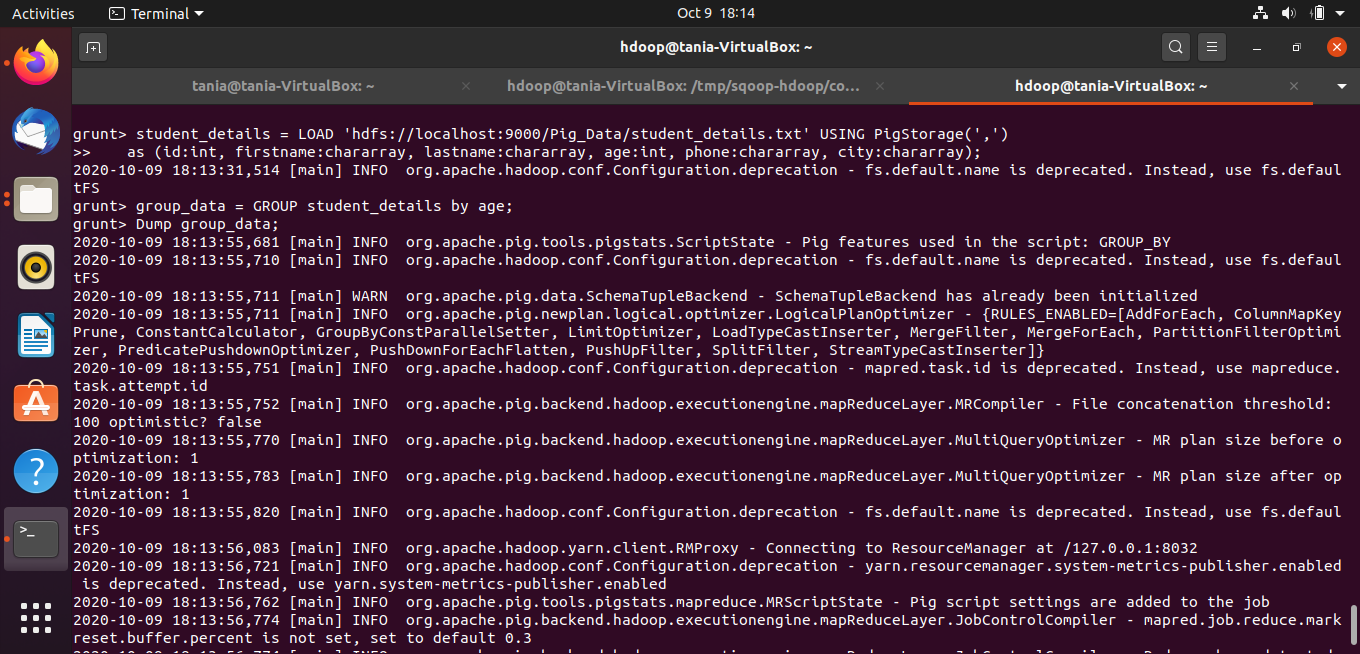


Step 2.6 : Illustrate Operator

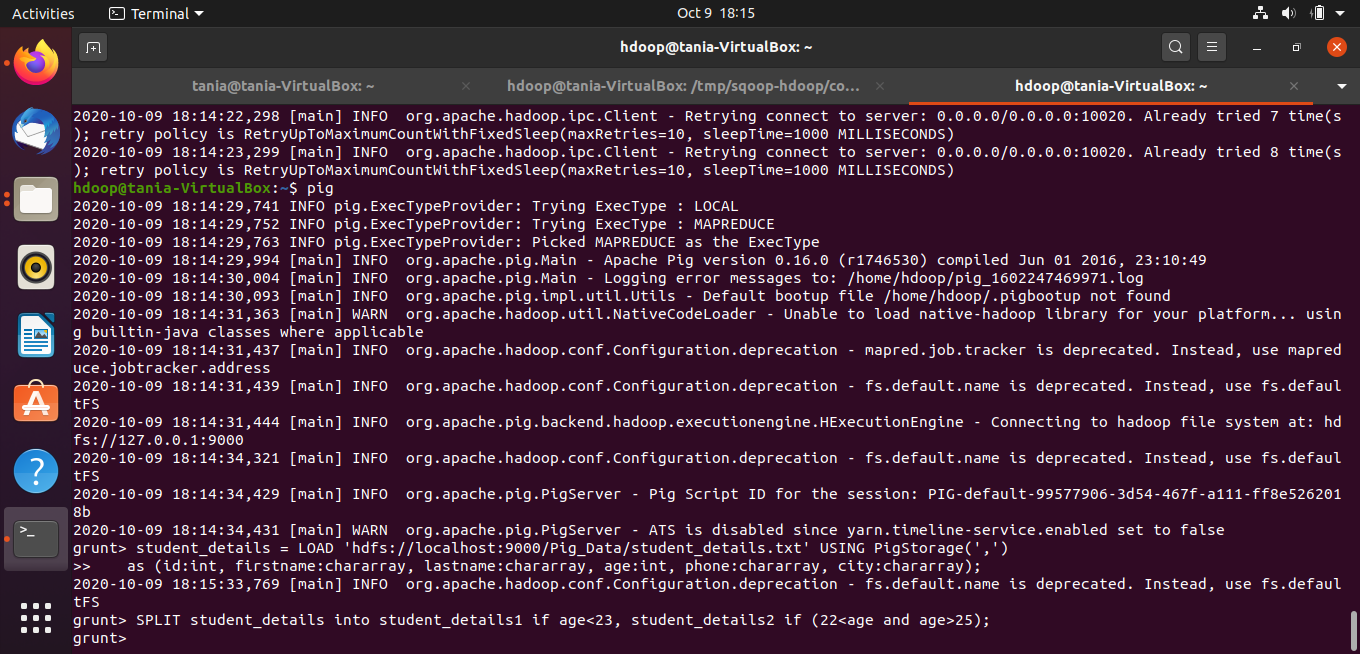




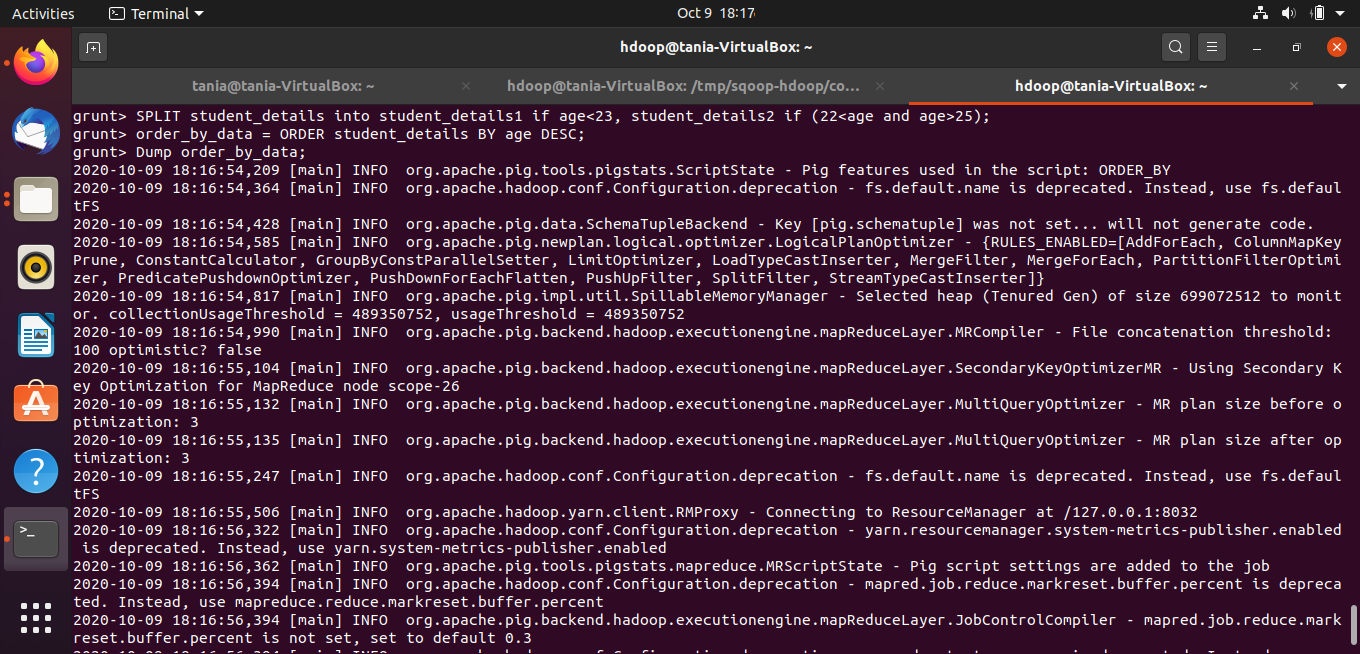
Step 2.7 : Group Operator



Step 2.8 : Split Operator



Step 2.9 : Order By Operator



**Conclusion:**

In this experiment, we installed Apache pig and performed various operations in Pig like loading and storing data, group by, describe, illustrate, etc.