

**STATE WISE DATA ANALYSIS**

Submitted to the

**Department of Master of Computer Applications**

in partial fulfilment of the requirements of

**Hadoop Ecosystems (MAE03)**

by

**VINAYKUMAR**

**1MY18MCA28**

**RAMAIAH INSTITUTE OF TECHNOLOGY**

(Autonomous Institute, Affiliated to VTU)

Accredited by National Board of Accreditation & NAAC with ‘A’ Grade

MSR Nagar, MSRIT Post, Bangalore-560054

www.msrit.edu

2019-2020



**CERTIFICATE**

This is to certify that the Mini Project entitled **STATE WISE DATA ANALYSIS (USA)** is carried out by VINAYKUMAR, 1MY18MCA28, a bonafide student of Dept. of Master of Computer Applications, Ramaiah Institute of Technology, Bangalore, in partial fulfilment for the **Hadoop Ecosystems (MCAE03)** during the year **2019-2020.** It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the mini-project report deposited in the department.

**Faculty Signature**

**1. Abhishek K L**

**2. Mrs. Sailaja Kumar K**

**CONTENTS**

**Title**  Page no

1) Abstract----------------------------------------------------------------4

2) Implementation----------------------------------------------------5-18

a) Load the data into hdfs using flume

b) Copy the data into table using pig

c) Queries

3) CONCLUSION -------------------------------------------------------------19

**ABSTRACT**

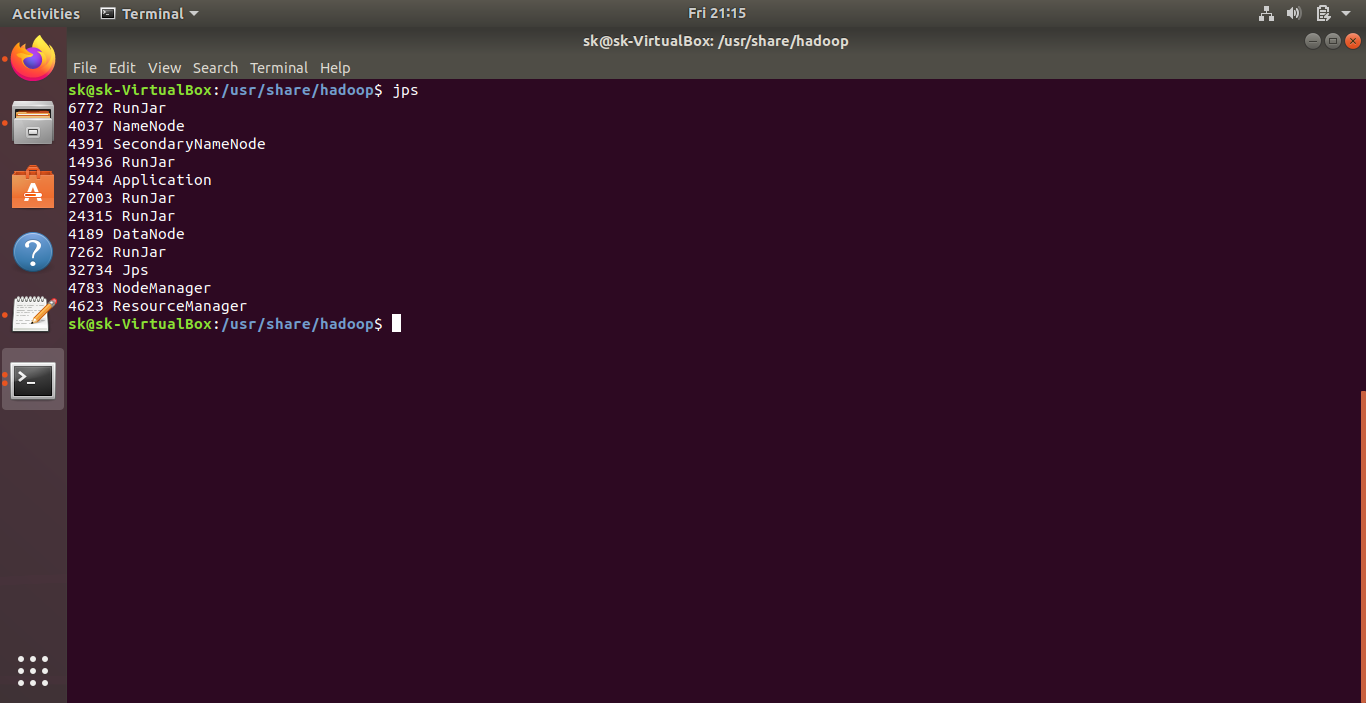
The study of STATE WISE data is interesting for several reasons. They will try to find in which state highest population is, identify gdp growth and internet households or explore the more about state and other related information. Second, the dataset has rich information on each state and their data allowing for a wide variety of data analyses. Hadoop is used for storing and processing big data. It is a distributed file system allowing concurrent processing and fault tolerance. In this project the tools used are FLUME AND PIG. FLUME is used to load large amounts of data from the local system to the Hadoop file system (HDFS).Here we have used flume to load the state wise data. PIG is used to analyse the data present in the HDFS, by using operations such as MAX, MIN, SUM, and AVG. also, it includes grouping of data for better understating. The main objective of this project is to demonstrate data by using Hadoop ecosystems, how data generated from state wise, and how it is utilized to make target, real time and informed decisions.

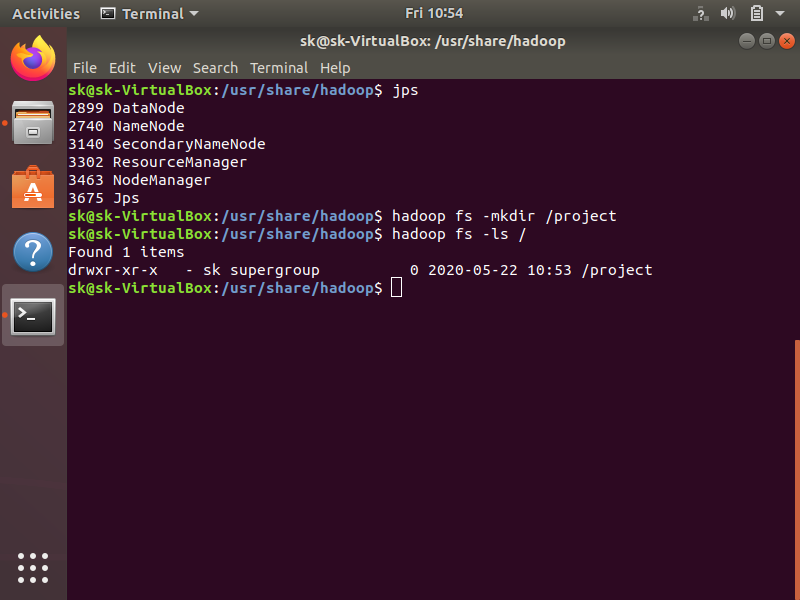
**SCREENSHOTS**

1. **JPS Screenshot**

**(Java Virtual Machine Process Status Tool)**

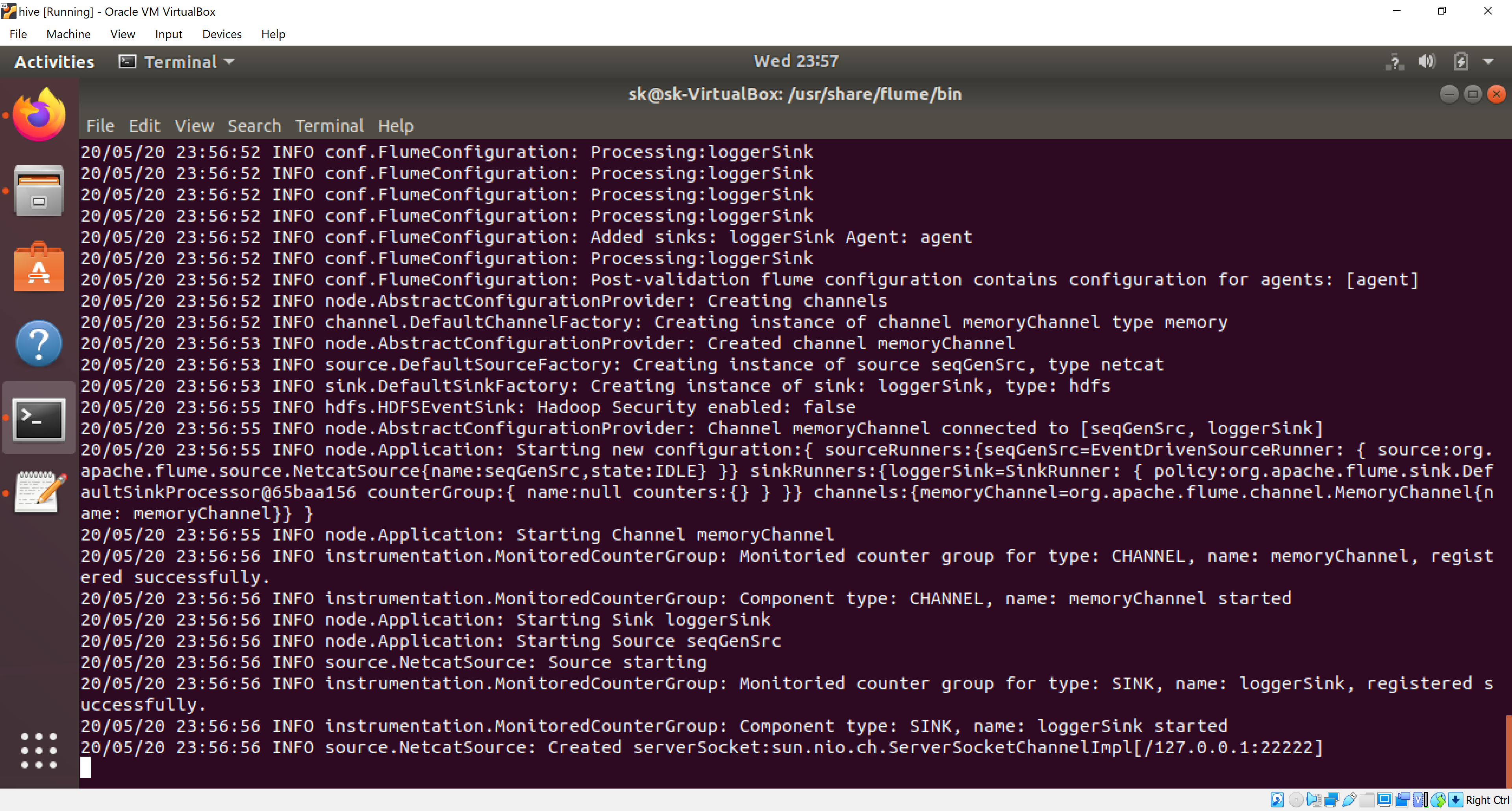
Description: It is a command used to check all the Hadoop daemons like Name Node, Data Node, Resource Manager, Node Manager etc. which are running on the machine.

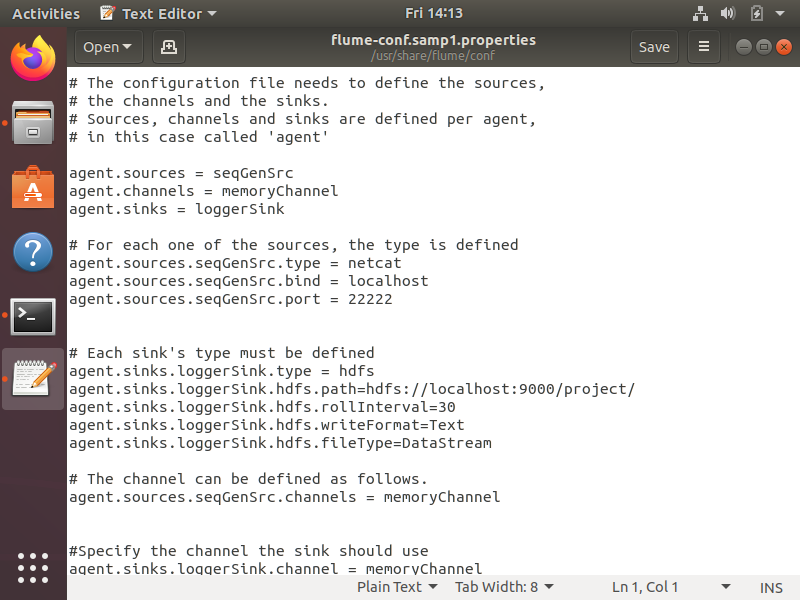


**2) Creation of directory in hadoop**

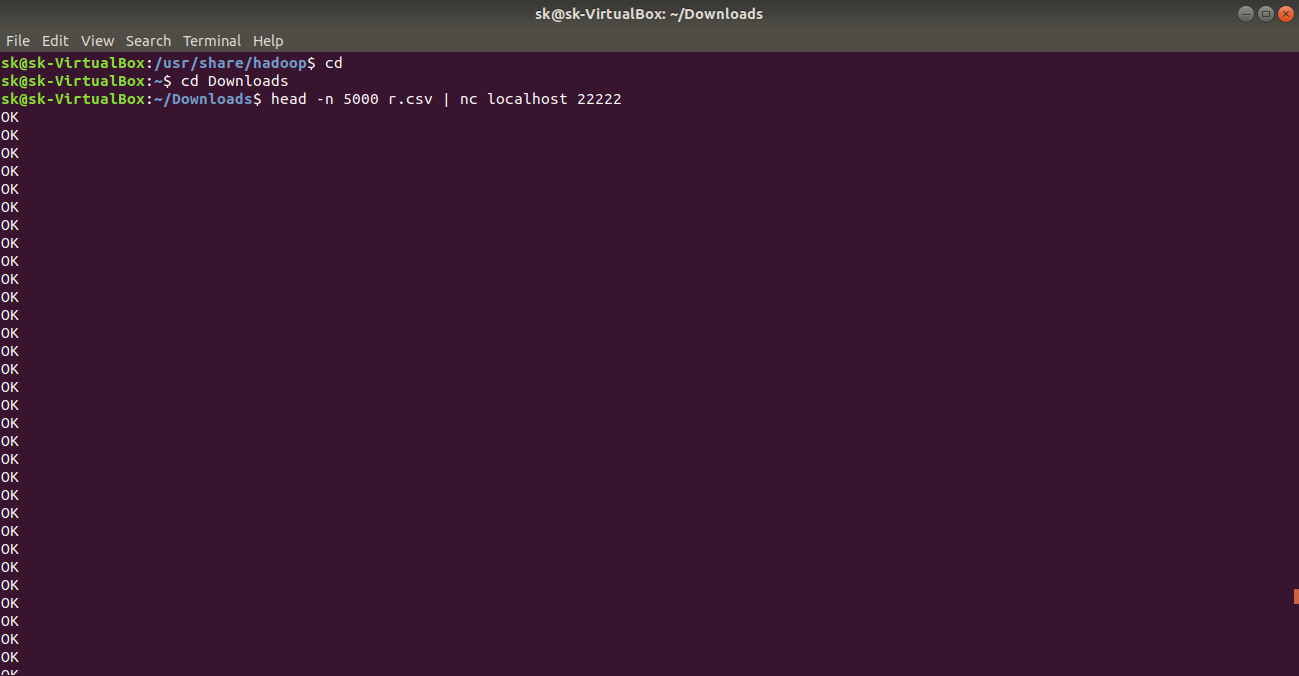
**3) Flume command execution screen**

Query: /flume-ng agent -n agent -c conf -f ../conf/flume-conf.samp1.properties. Description: Starting the Flume.

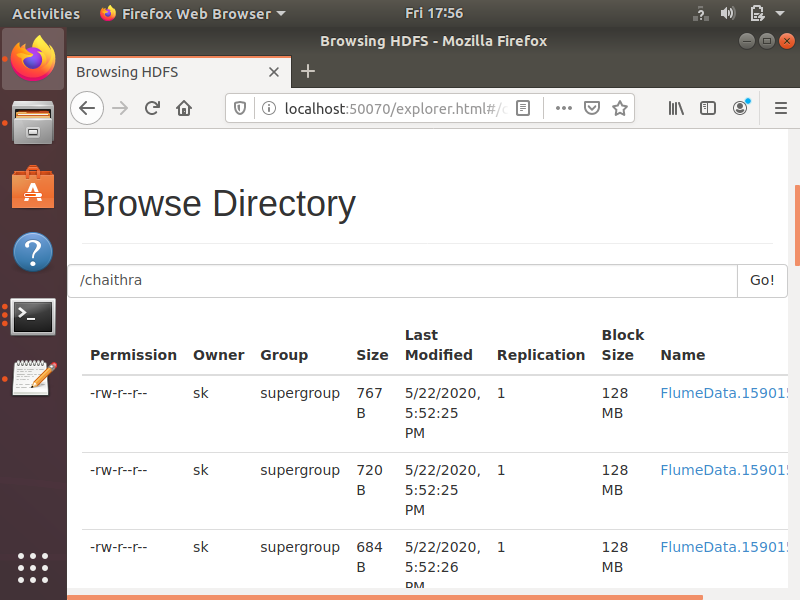


**4) Configuration file screen**

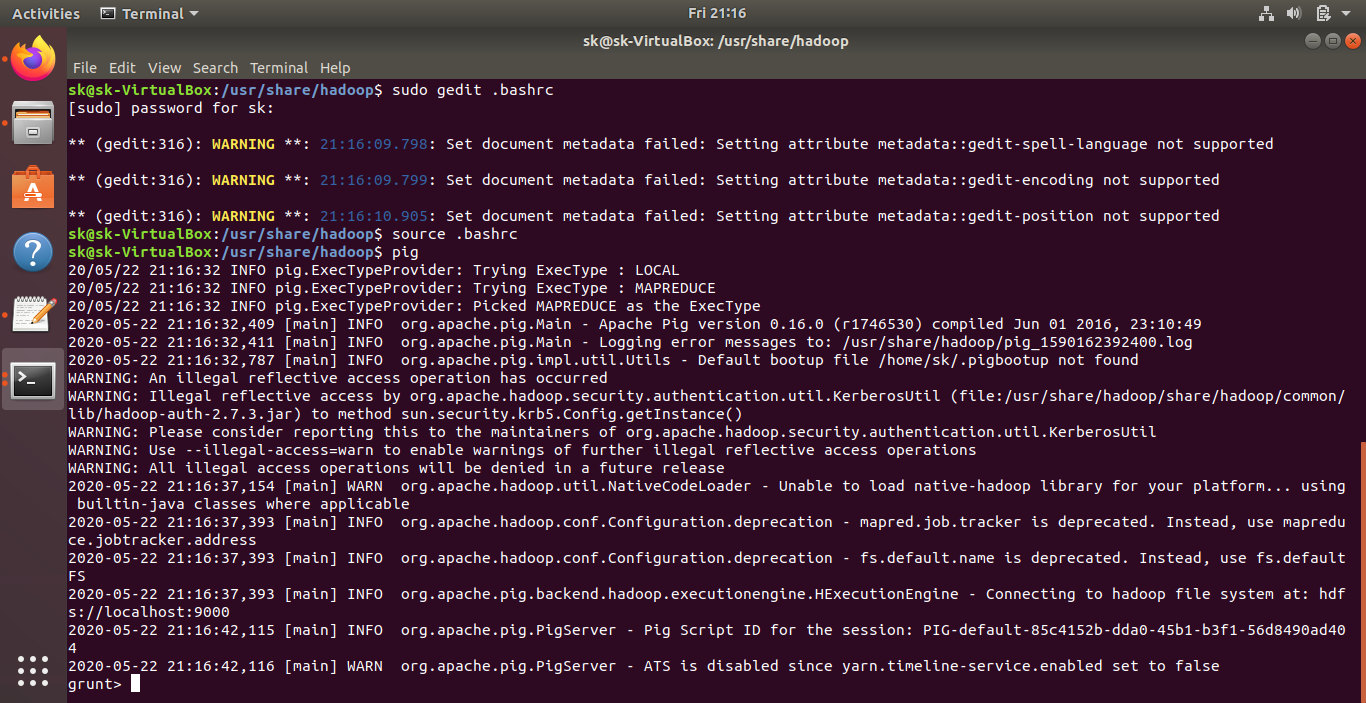
**5) Loading the data into Flume from Desktop using Head command.**



**6) Flume files in browser**

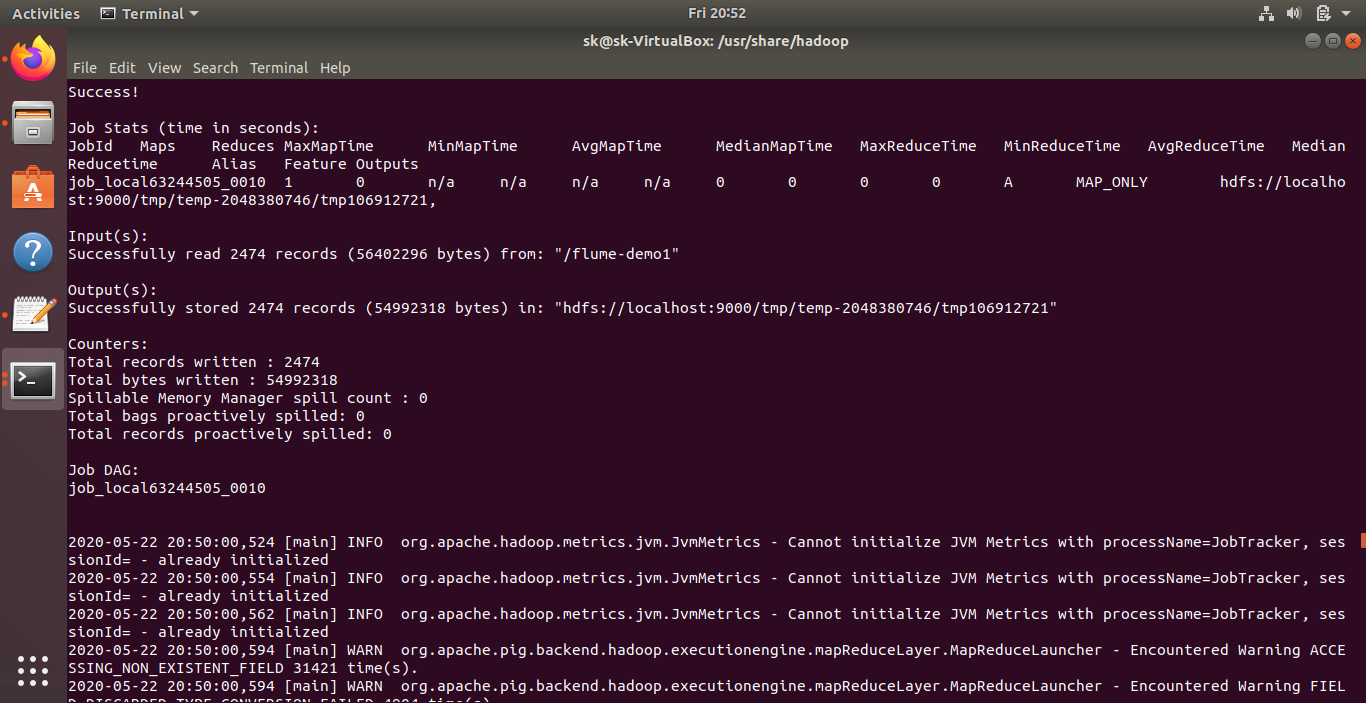


6) Data copied into table using pig and Displaying the Loaded Data:



**QUERIES USING PIG**

**A = LOAD '/flume-demo1' USING PigStorage(',')AS(state:Chararray,year:int,internet\_usage\_household:float,bachelors\_degree\_or\_up:float,gdppcgrowth:float,unemployment:float,population:int,farmemploy:int,mining\_employ:int,utilitiesemploy:int,constructionemploy:int,manufactureemploy:int,wholesaletradeemploy:int,retailtradeemploy:int,transportemploy:int,financeinsuranceemploy:int,realtyemploy:int,managementemploy:int,educationemploy:int,healthcareemploy:int,artsemploy:int,governmentemploy:int,male:float,female:float);**

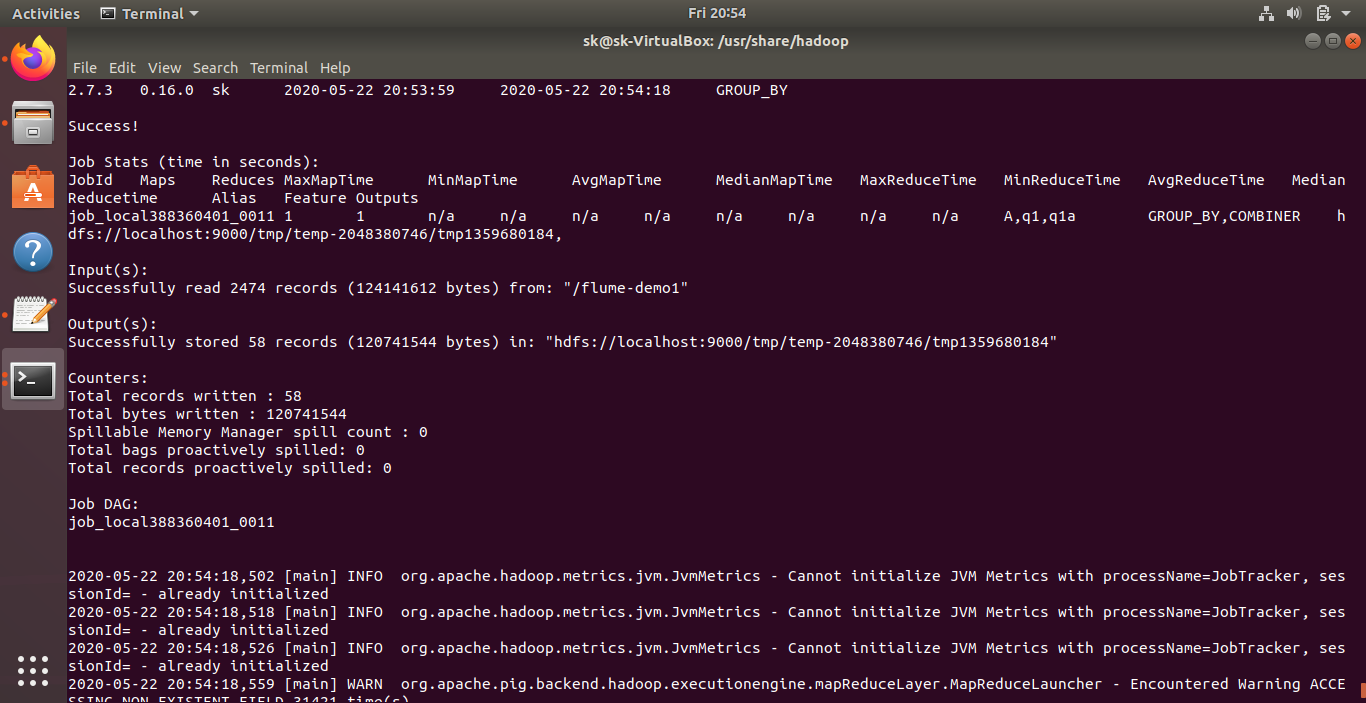
****

**Query 1)** **Display name of the state who had a heighest internet\_usage\_household??**

**Answer:** q1 = GROUP A by state;

q1a = FOREACH q1 GENERATE MAX (A.internet\_usage\_household);

dump q1a;

**Output: **

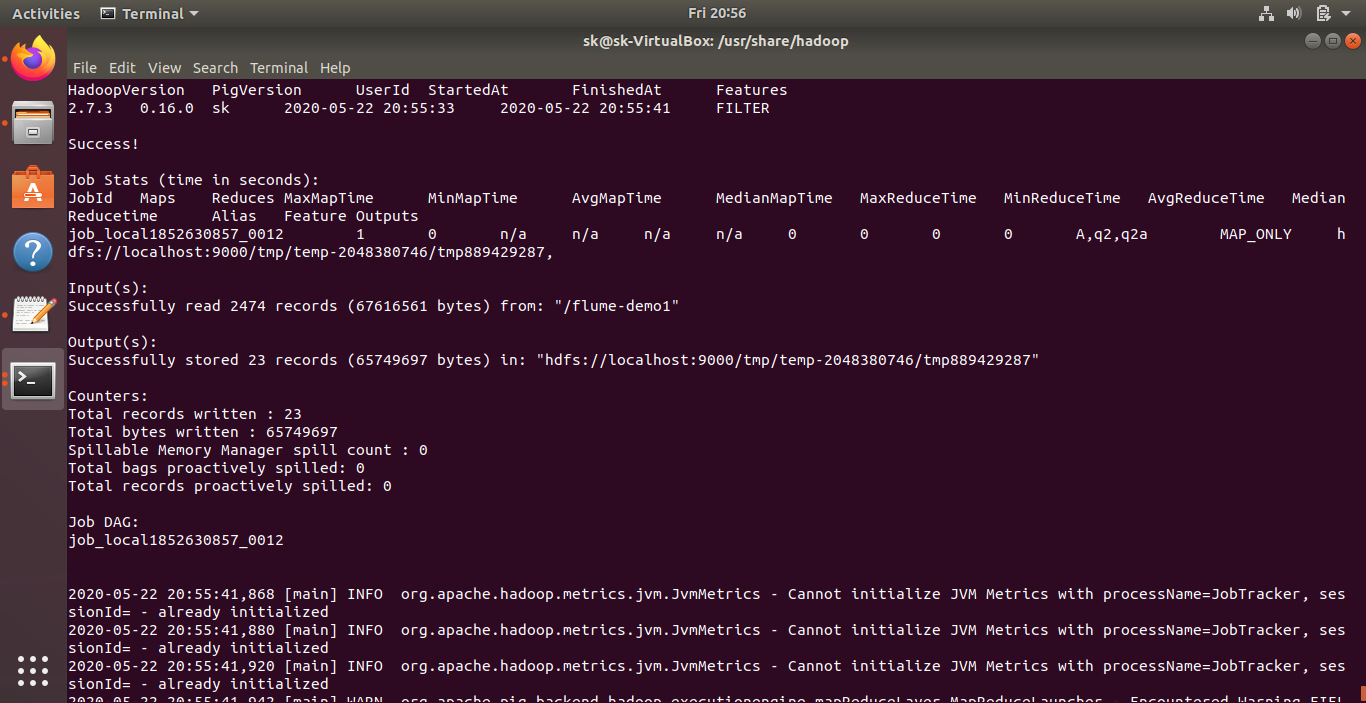
**Query 2) Display the name of the state and bachelors\_degree\_or\_up who had a population above 5000000?**

**Answer:** q2 = FILTER A by population > 5000000;

q2a = foreach q2 generate state,bachelors\_degree\_or\_up;

dump q2a;

**Output:**

****

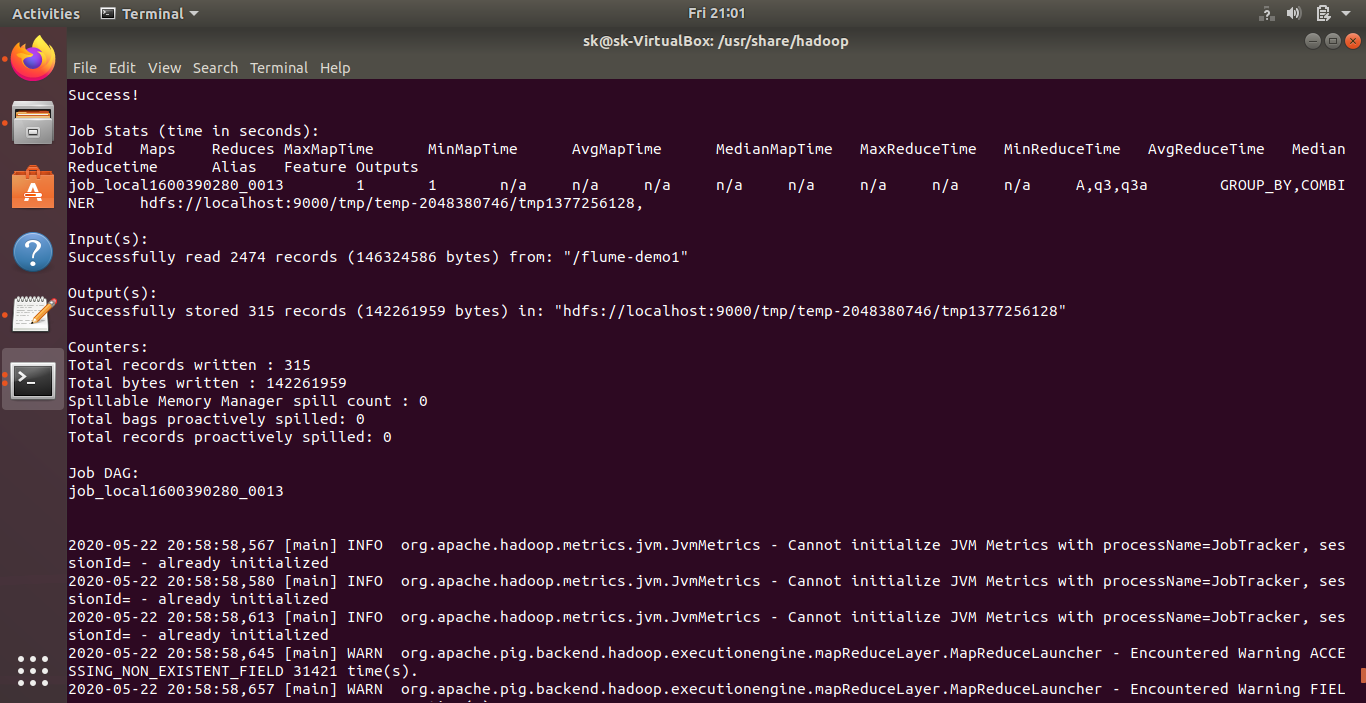
**Query 3)** **Display the unemployment state whose population is more?**

q3 = GROUP A BY (state,unemployment);

q3a = FOREACH q3 GENERATE $0,MAX(A.population);

dump q3a;

**Output:**



**Query 4:** **List all state desc ?**

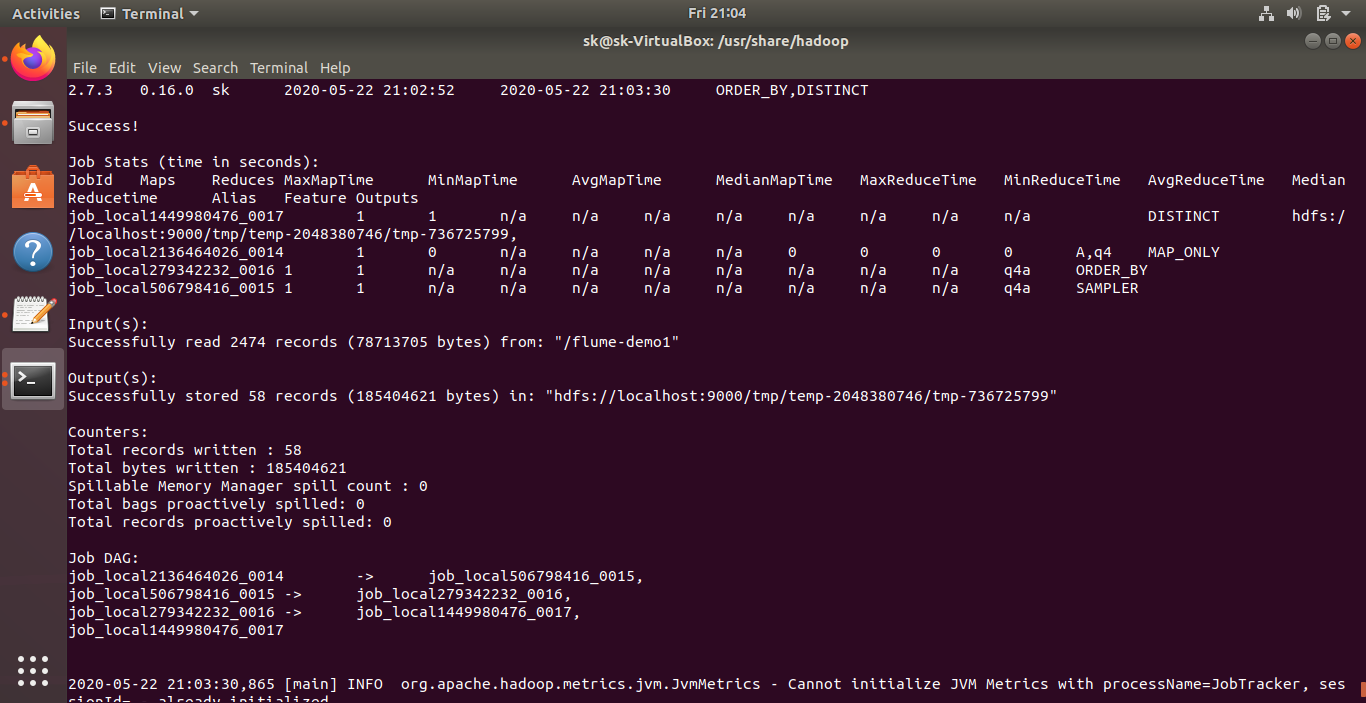
q4 = foreach A generate $0;

q4a = order q4 by state;

q4aa = distinct q4a;

dump q4aa;

**Output:**



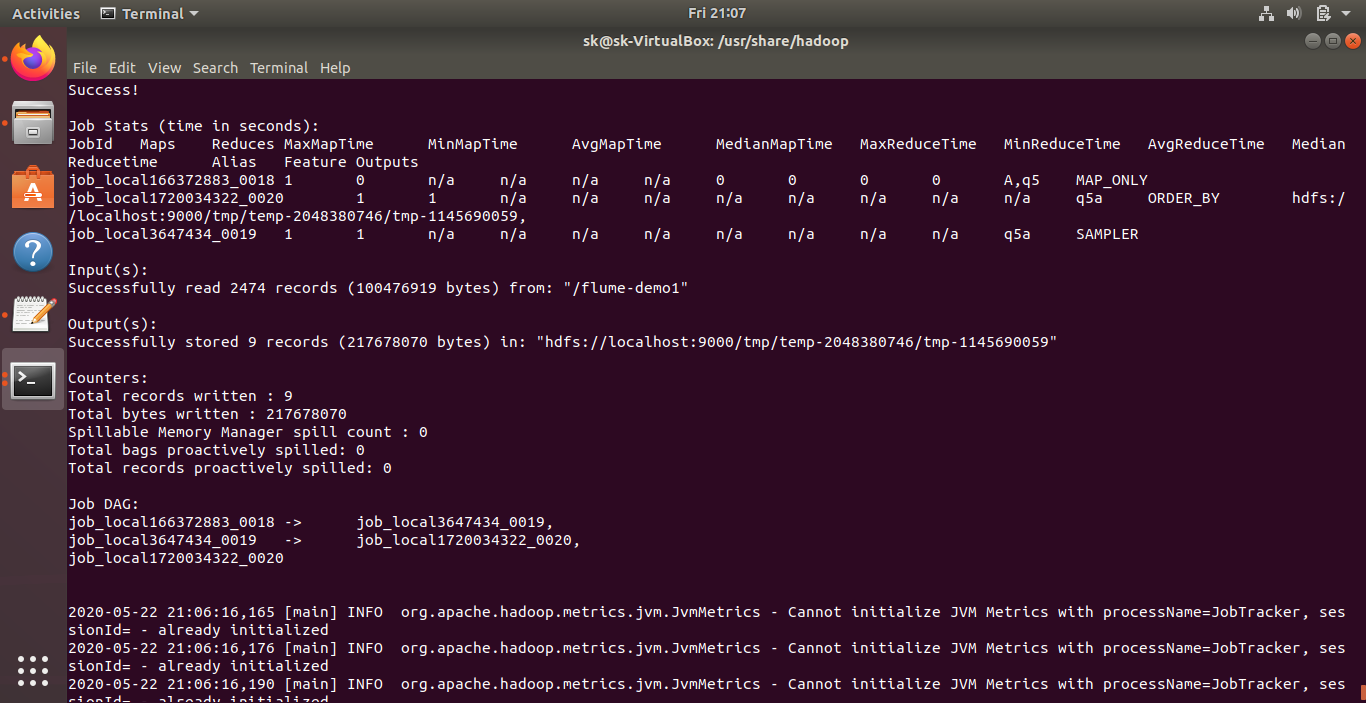
**Query 5: Display the name of the state who had internet\_usage\_household greater than 80?**

q5 = filter A BY $2 > 80;

q5a = order q5 by $0,$2;

dump q5a;

**Output:**

****

**Query 6) Show total case in healthcare emp group waise**

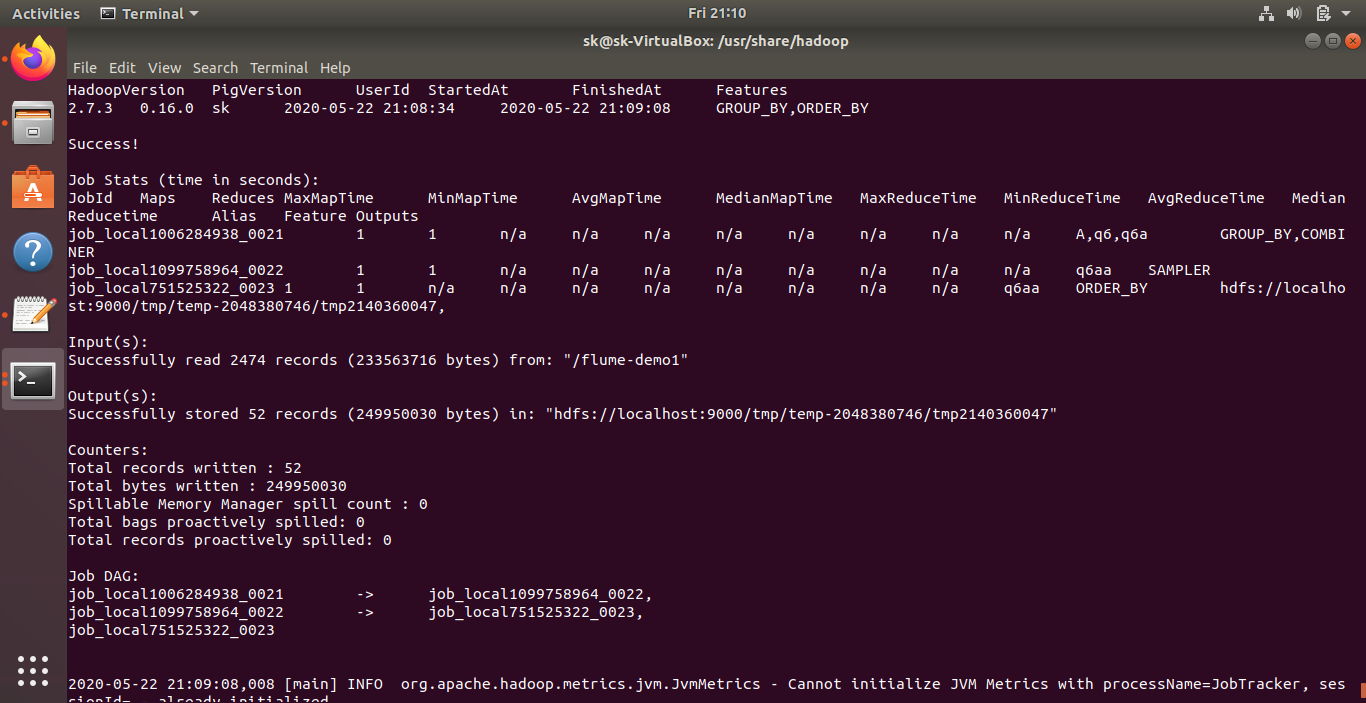
q6 = group A by (healthcareemploy);

q6a = foreach q6 generate $0, COUNT(A.healthcareemploy);

q6aa = Order q6a By $1 desc;

dump q6aa;

**Output:**

****

**Query 7: Dispaly the name of the state who had GDP Growth greater than 3?**

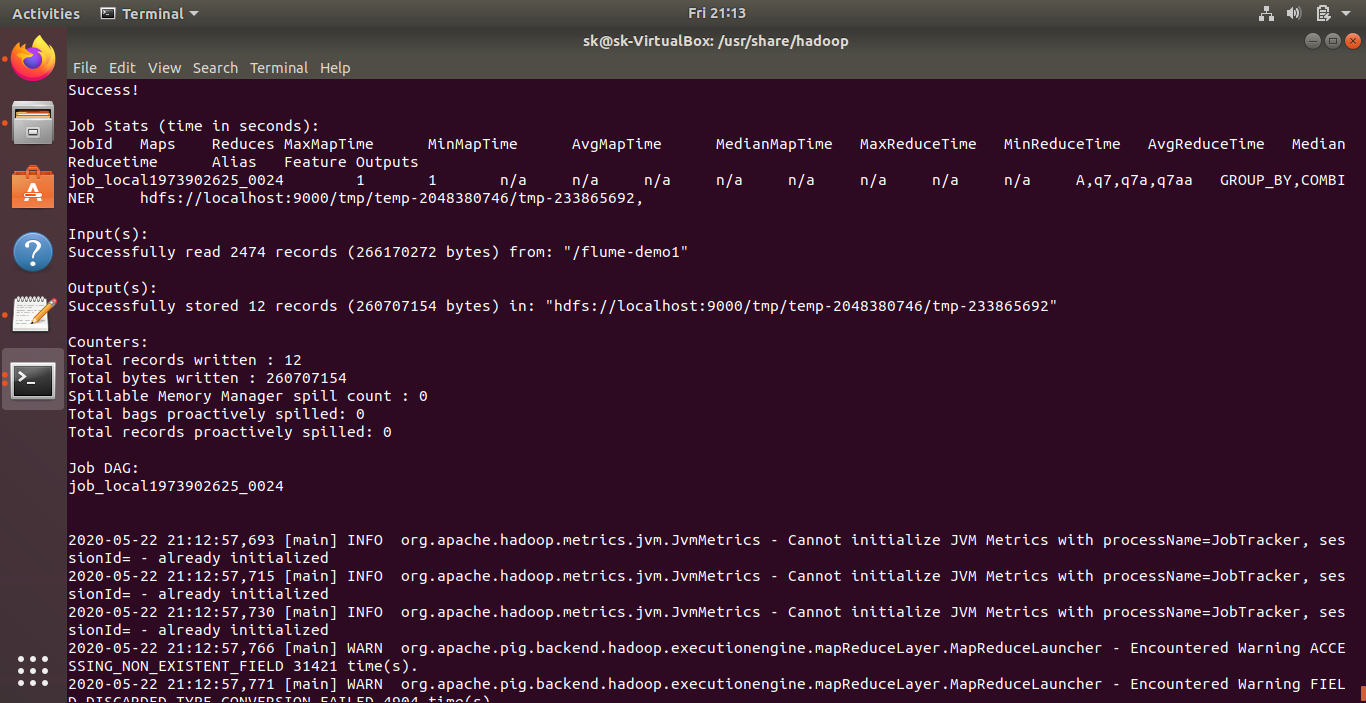
q7 = filter A BY $4 > 3;

q7a = group q7 by $0;

q7aa = foreach q7a generate $0, COUNT(q7.$4);

dump q7aa;

**Output:**



**Query 8)** **.Display the average of unemployment in particular state?**

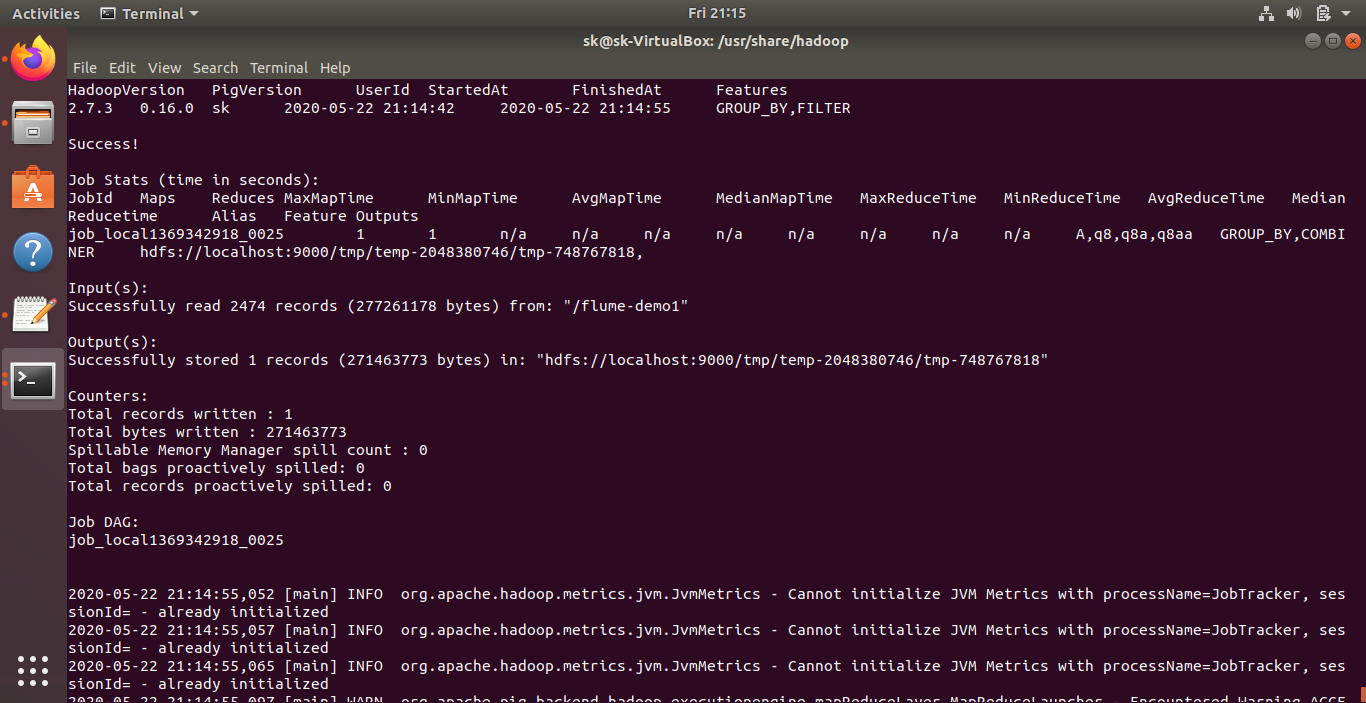
q8 = FILTER A BY state == 'Alaska';

q8a = group q8 all;

q8aa = foreach q8a generate AVG(q8.unemployment);

dump q8aa;

**Output:**

****

**CONCLUSION**

The task of big data analysis is not only important but also a necessity. In fact, many organizations that have implemented Big Data are realizing significant competitive advantage compared to other organizations with no Big Data efforts. The project is intended to analyze the course data and come up with significant insights which cannot be determined otherwise.