**A PROOF OF CONCEPT**

**On**

**Big Data Analytics**

**SIX WEEKS SUMMER TRAINING REPORT**

***Submitted by***

*KANISHKA PAL*

Reg. No.:

***Under the Guidance of***

Allsoft Solution (IBM)

Ms. Jyoti Taylor

Ms. Manorma Sharma

***In partial fulfillment for the award of the degree of***

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE AND ENGINEERING**



**Faculty of Engineering & Technology**

**Manav Rachna International Institute of Research and Studies, Faridabad**

**ACKNOWLEDGEMENT**

The successful realization of project is an outgrowth of a consolidated effort of people from desperate fronts. We are thankful to **MS. JYOYI TAYLOR** and **MS. MANORMA SHARMA** for their variable advice and support extended to us without which we could not be able to complete our project for a success.

We are thankful to **MS. JYOYI TAYLOR** and **MS. MANORMA SHARMA,** Project Coordinator, for her guidance and support.

I would like to express my special thanks of gratitude to my teacher (*Ms. Jyoti*

*Taylor and Ms. Manorma Sharma*) as Well as our Director who gave me the golden chance to do this project on the vibrant field of **Big Data Analytics (Rainfall in India (1901-2017))**, which also helped me in doing a lot of Research about the rainfall of our country and I came to know about so many new things.

I am really thankful to them.

*- KANISHKA PAL*

**Declaration**

I hereby declare that I have completed my six-week summer training at (Allsoft

Solution (IBM), from (22/05/2020) to (30/06/2020) under the guidance of (Ms.

Jyoti Taylor and Ms. Manorma Sharma). I hereby undertake that the project undertaken by me is the genuine work of mine.

(Signature of student)  
 *KANISHKA PAL*

Reg. No.:



**Certificate**

This is to certify that this project report entitled “Big Data Analytics” *by KANISHKA PAL* submitted in partial fulfillment of the requirements for the summer internship (Allsoft Solution (IBM)), during the six week summer training, is a bonafide record of work carried out under my guidance and supervision. I hereby declare that the work has been carried out under my supervision and has not been submitted elsewhere for any other purpose.

**(Signature of Project Guide)                  (Signature of Project Guide)**

Ms. Jyoti Taylor   Ms. Manorma Sharma

IBM Trainer                                                 Carrier Education Trainer

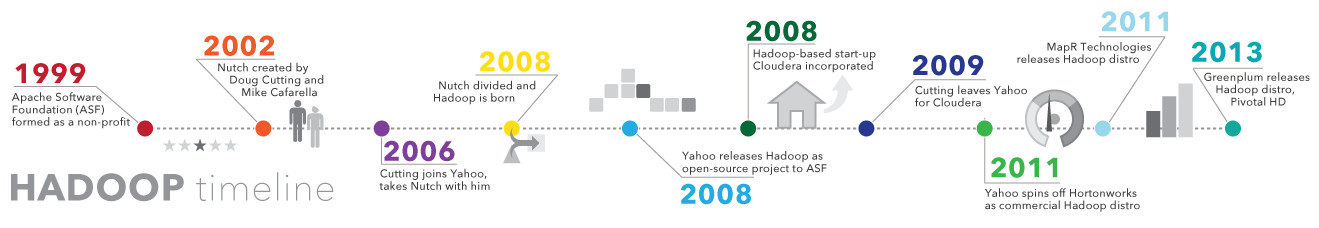
IBM

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| S. NO. | TOPICS | PAGE NO. |
| \* | History and Introduction of Hadoop | vi-vii |
|  |  |  |
| 1. | Introduction | 1 - 4 |
|  | 1.1 Purpose |  |
|  | 1.2 Project Scope |  |
|  | 1.3 References |  |
|  | 1.4 About Data Set |  |
|  | 1.5 Details Of Data Set |  |
|  | 1.6 Attribute Information |  |
|  | 1.7 Screenshots |  |
|  |  |  |
| 2. | Tools and Working Essentials | 5 |
|  | 2.1 Dependencies |  |
|  |  |  |
| 3. | Moving file On Tools | 5 – 12 |
|  | 3.1 HDFS |  |
|  | 3.2 Sqoop |  |
|  | 3.3 HIVE |  |
|  | 3.4 PIG |  |
|  |  |  |
| 4. | Queries | 13 – 26 |
|  |  |  |
| 5. | Partition with Bucketing | 27 - 29 |
|  |  |  |
| 6. | Pig Queries |  |
|  |  |  |
| 7. | Bibliography | 30 |

* **HISTORY OF HADOOP**

As the World Wide Web grew in the late 1900s and early 2000s, search engines and indexes were created to help locate relevant information amid the text-based content. In the early years, search results were returned by humans. But as the web grew from dozens to millions of pages, automation was needed. Web crawlers were created, many as university-led research projects, and search engine start-ups took off (Yahoo, AltaVista, etc.).

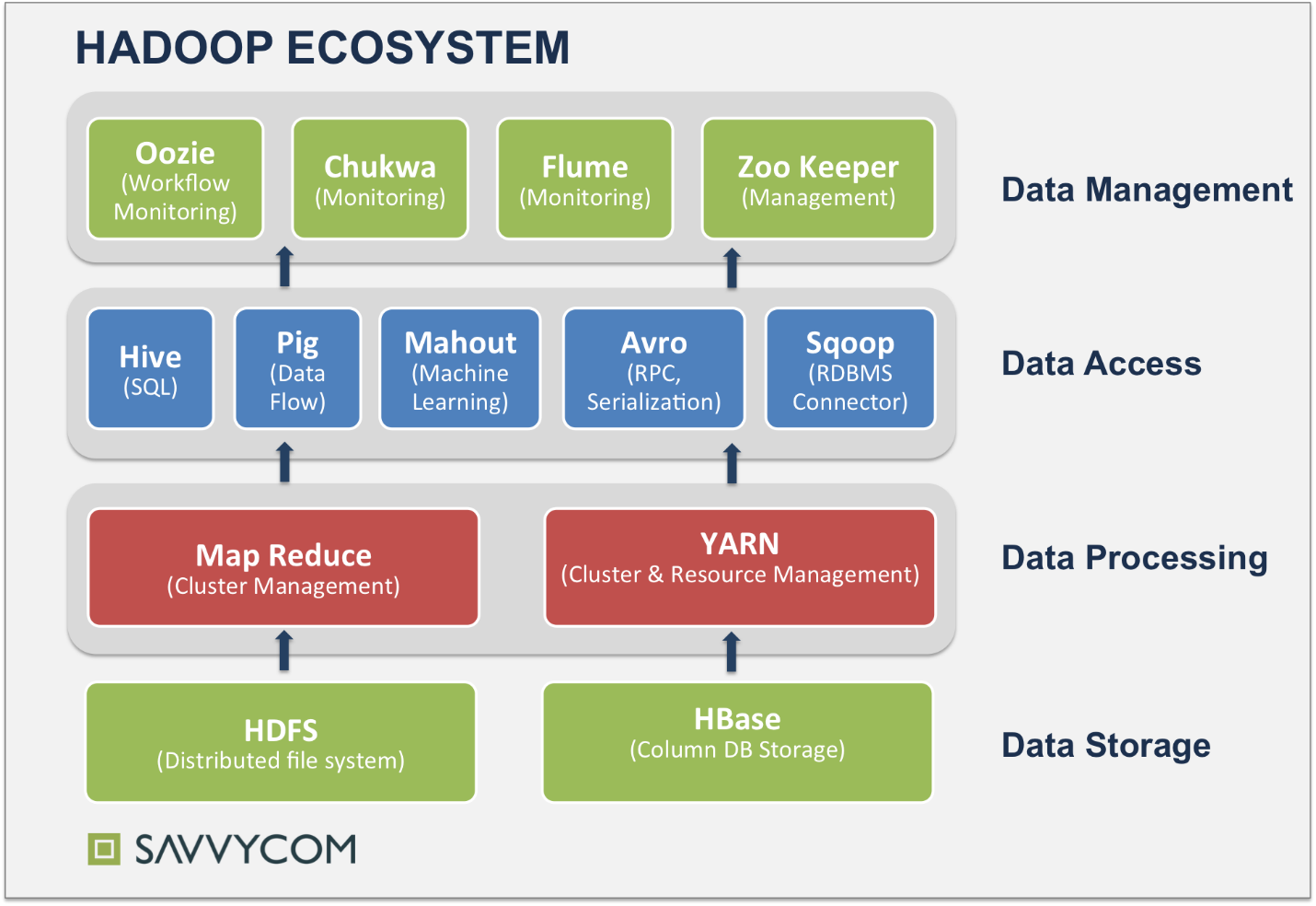


One such project was an open-source web search engine called Nutch – the brainchild of Doug Cutting and Mike Cafarella. They wanted to return web search results faster by distributing data and calculations across different computers so multiple tasks could be accomplished simultaneously. During this time, another search engine project called Google was in progress. It was based on the same concept – storing and processing data in a distributed, automated way so that relevant web search results could be returned faster.

In 2006, Cutting joined Yahoo and took with him the Nutch project as well as ideas based on Google’s early work with automating distributed data storage and processing. The Nutch project was divided – the web crawler portion remained as Nutch and the distributed computing and processing portion became Hadoop (named after Cutting’s son’s toy elephant). In 2008, Yahoo released Hadoop as an open-source project. Today, Hadoop’s framework and ecosystem of technologies are managed and maintained by the non-profit Apache Software Foundation (ASF), a global community of software developers and contributors.

* **WHAT IS HADOOP**

Apache **Hadoop** is an open source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. Instead of using one large computer to store and process the data, **Hadoop** allows clustering multiple computers to analyze massive datasets in parallel more quickly.Hadoop is a framework that allows you to first store Big Data in a distributed environment, so that, you can process it parallel.



1. **Introduction**

**Data Set: - Rainfall in India (1901-2017)**

****

This project is to analyses a data set related to Rainfall in India within the time-period of 1901-2017. It includes the information and raw data about the rain occurred in different regions of India.

It gives a region wise perspective analysis of the deviation between rains occurred in different months.

*The features of this project are given below –*

1. Framing the data region wise for better understanding.
2. The precipitation statistics include annual precipitation for the years 1901-2017.
3. Descriptive Analysis (i.e. it is limited to past data).

**1.1 Purpose**

This project is to analyses a data set related to Rainfall in India within the time-period of 1901-2017. It includes the information and raw data about the rain occurred in different regions of India. It gives a state wise perspective analysis of deviation between various Indian states.

**1.2 Project Scope**

1. It is helpful for checking details of the rain occurred in the past years.

2. It is helpful in maintaining records.

3. We can also compare past rainfall data with present.

4. It is valid for descriptive analysis.

5. It is valid for particular set of compatible attributes.

**GOALS: -**

1. How did the rain fall varies in different states/region?

2. Did seasons played a major role in measurements of rainfall occurred?

3. Was there any state of India which had most rain fall during year?

4. Can it helps in analyzing the drought states and flooded states of India?

**1.3. References**

<https://www.kaggle.com>

<http://www.data.gov.com>

**1.4 About Data Set**

This data set contains information related to Rainfall in India. It contains state/region name, year (1901-2017), month’s name, annual rainfall, and many more attributes. This POC contains the Hive and Pig queries to manipulate and analyze the data. This also contains the word count for the same data set.

**1.5 The details of the data set are given below:**

**Format:**

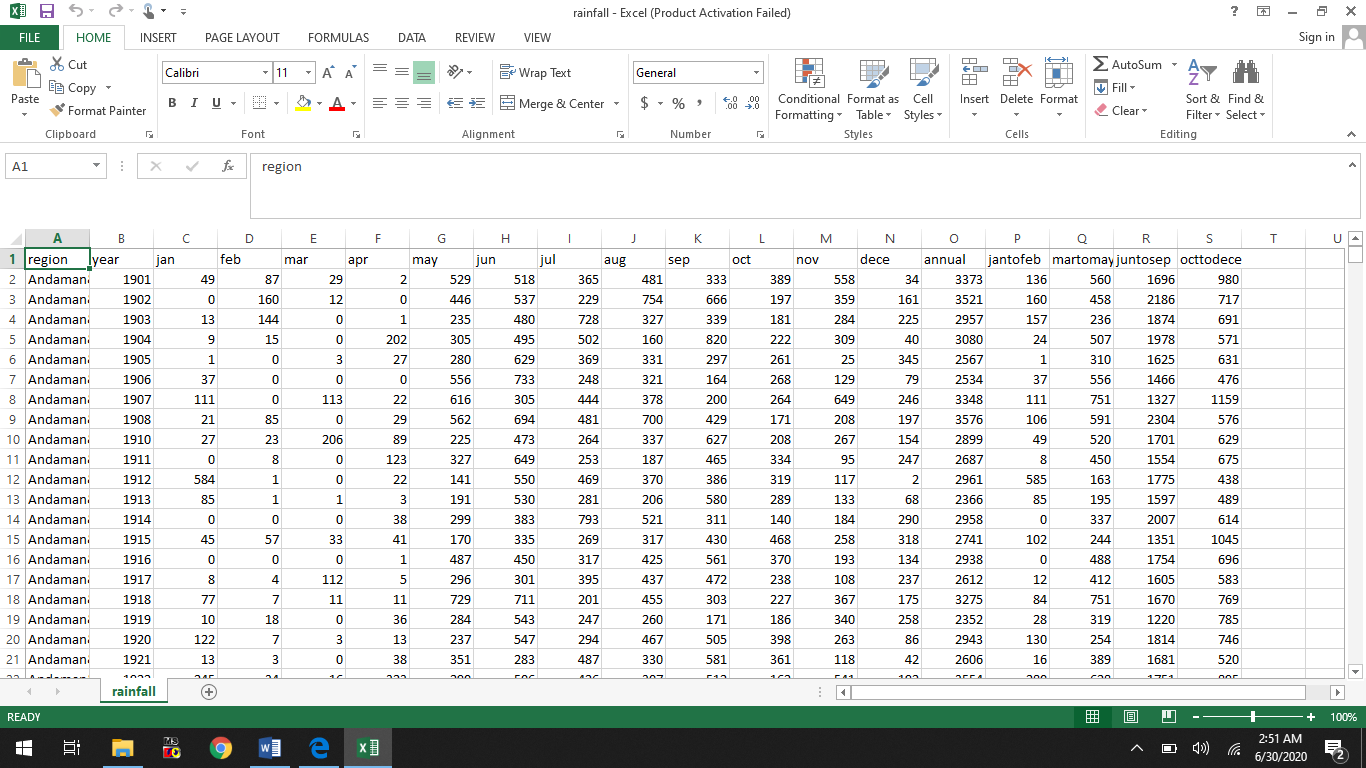
The data set contains the following information about Rainfall in India, they are divided by ‘,’ in the data file:

Region name, year, January, February, March, April, May, June, July, August, September, October, November, December, Annual sum of months, JanuarytoFebruary, MarchtoMay, JunetoSeptember, OctobertoDecember.

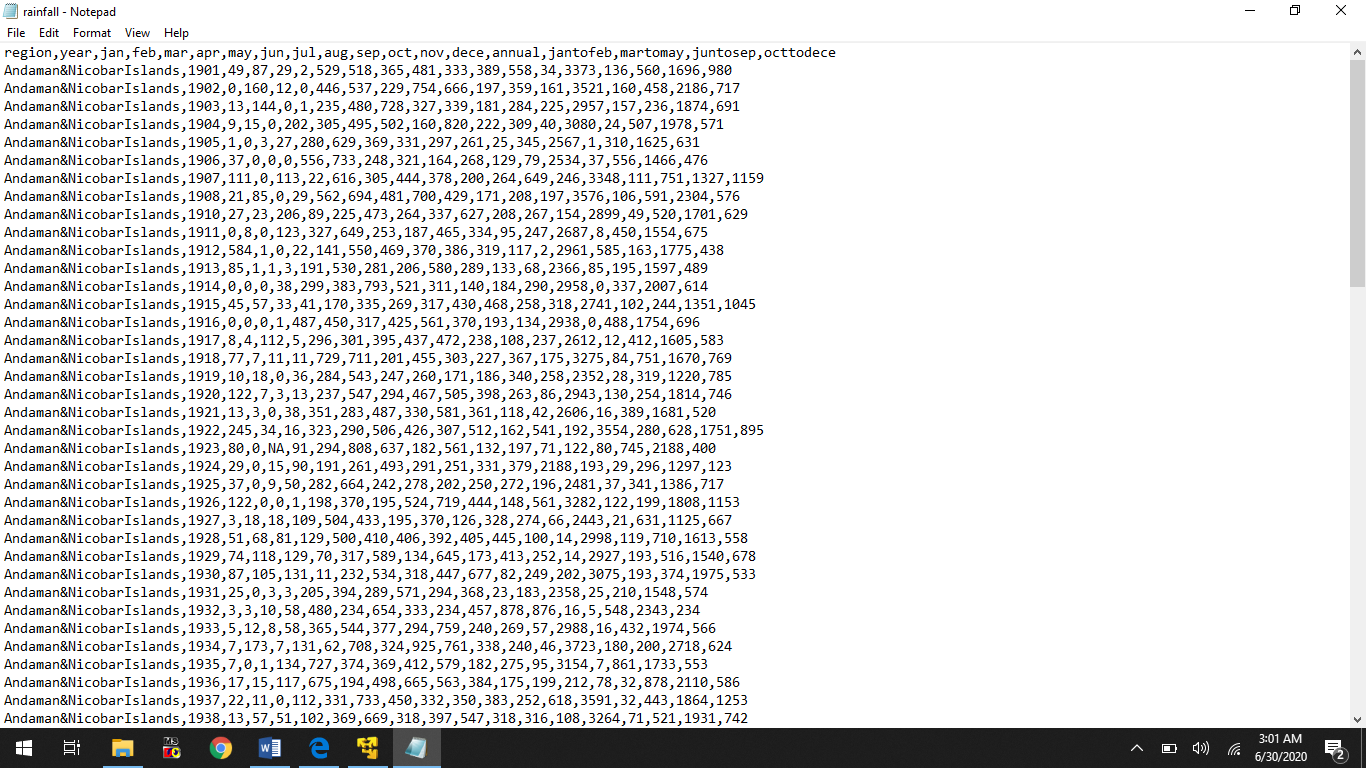
**1.6 Attribute Information**

|  |  |
| --- | --- |
| **Attribute Name** | **Description** |
| Region | A string of region/state name. |
| Year | An integer value which have data within the time-period of 1901-2017. |
| Jan | An integer value having data of January month. |
| Feb | An integer value having data of February month. |
| Mar | An integer value having data of March month. |
| Apr | An integer value having data of April month. |
| May | An integer value having data of May month. |
| Jun | An integer value having data of June month. |
| Jul | An integer value having data of July month. |
| Aug | An integer value having data of August month. |
| Sep | An integer value having data of September month. |
| Oct | An integer value having data of October month. |
| Nov | An integer value having data of November month. |
| Dece | An integer value having data of December month. |
| Annual | An integer value which provides us sum of rain in mm occurred in whole year. |
| Jantofeb | An integer value to store sum of data from January to February. |
| Martomay | An integer value to store sum of data from March to May. |
| Juntosep | An integer value to store sum of data from June to September. |
| Octtodece | An integer value to store sum of data from October to December. |

**1.7 Screenshots of Dataset**



*In Comma Separated View stored in notepad++ file.*



**2. TOOLS AND WORKING ENVIRONMENT:**

Project will be made using Cloudera working in VMware. It includes data manipulation and visualization libraries such as:-

1. *Linux*
2. *HDFS*
3. *Sqoop*
4. *Hive*
5. *Pig*

**2.1 Dependencies**

The dependencies that could be involved throughout the project is VMware Workstation 14 Player.

**3. Moving file on tools**

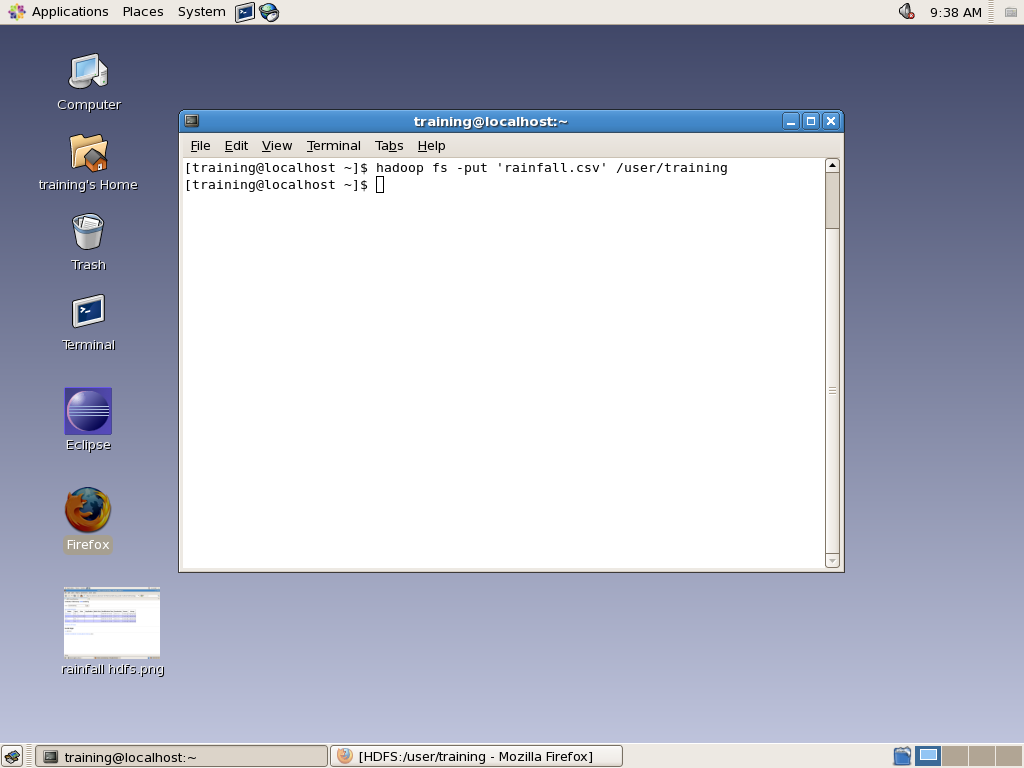
1. Hadoop Distributed File System
2. Sqoop
3. Hive
4. Pig

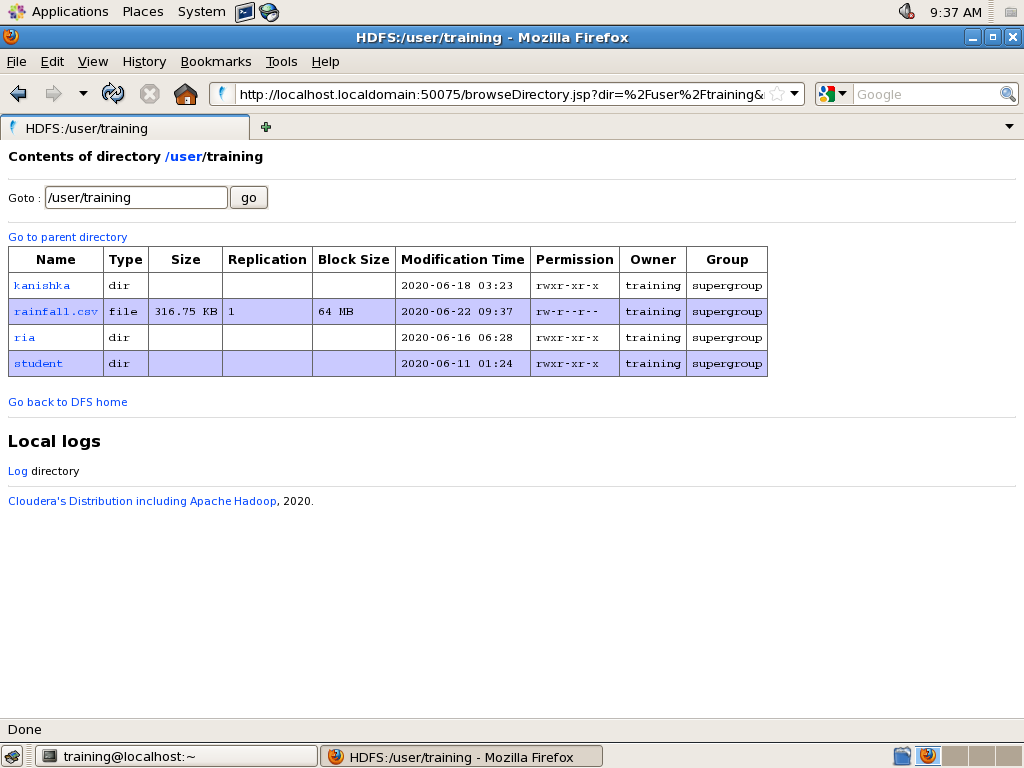
**3.1. HADOOP DISTRIBUTED FILE SYSTEM**

*Step 1: Moving data into Hadoop Distributed File System (HDFS).*

For any queries in Hive and Pig, we need to load the data in HDFS. For queries in hive we will load the data in HDFS and for queries in Pig we will use the data from local file system.

*The snapshot contains hadoop fs command to put dataset from local system to hdfs file system.*





**3.2. SQOOP**

Sqoop is used for import and export of data from\to RDBMS (MySQL).

3.2.1. *Sqoop import command.*

It imports data from MySQL and put it on hadoop cluster.

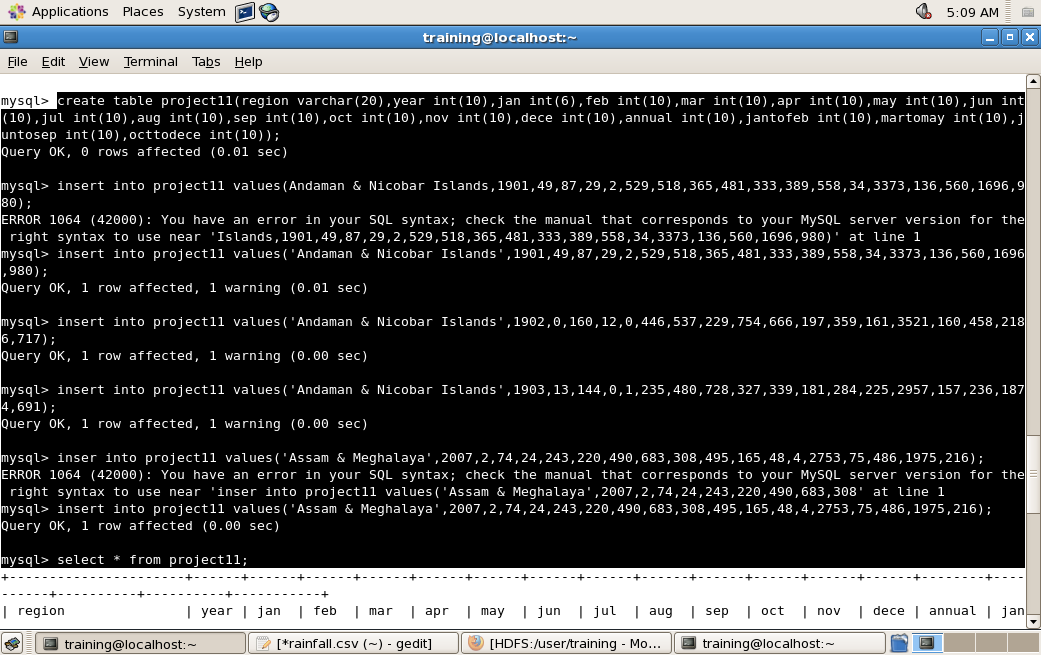
1. Open the terminal.

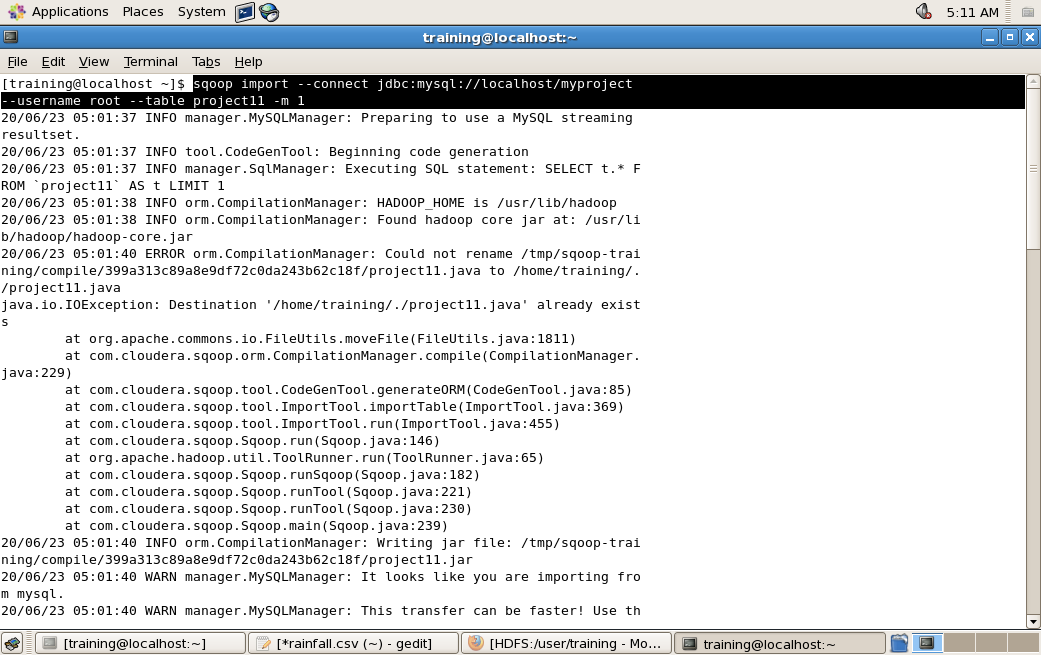
2. Now you need to create the table in sql and insert data into created table.

3. Write the Sqoop import command on the terminal.

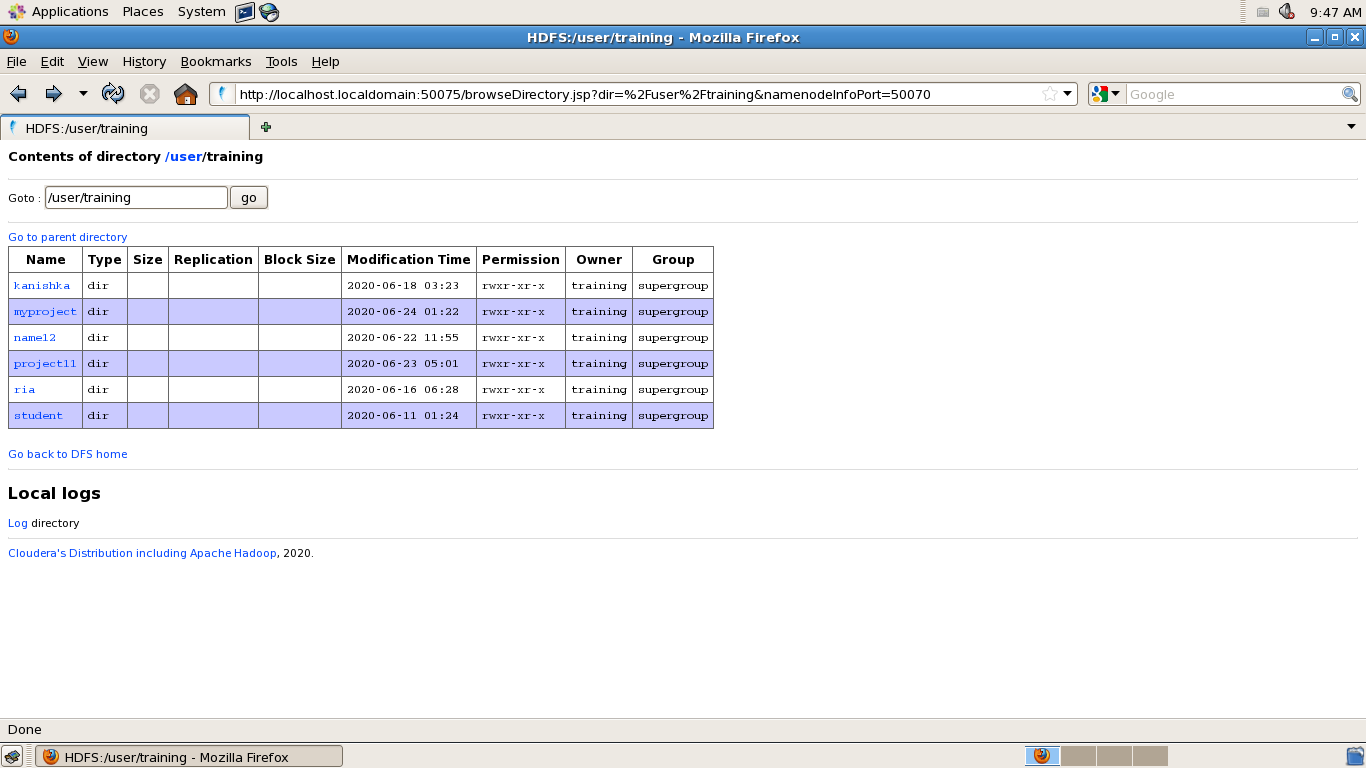
3. Check the file which we have imported on hdfs.

*The snapshot contains Sqoop import command to put dataset from MySQL to hdfs.*





Output:-



3.2.2. *Sqoop export command.*

It exports data from hdfs and put it on MySQL table.

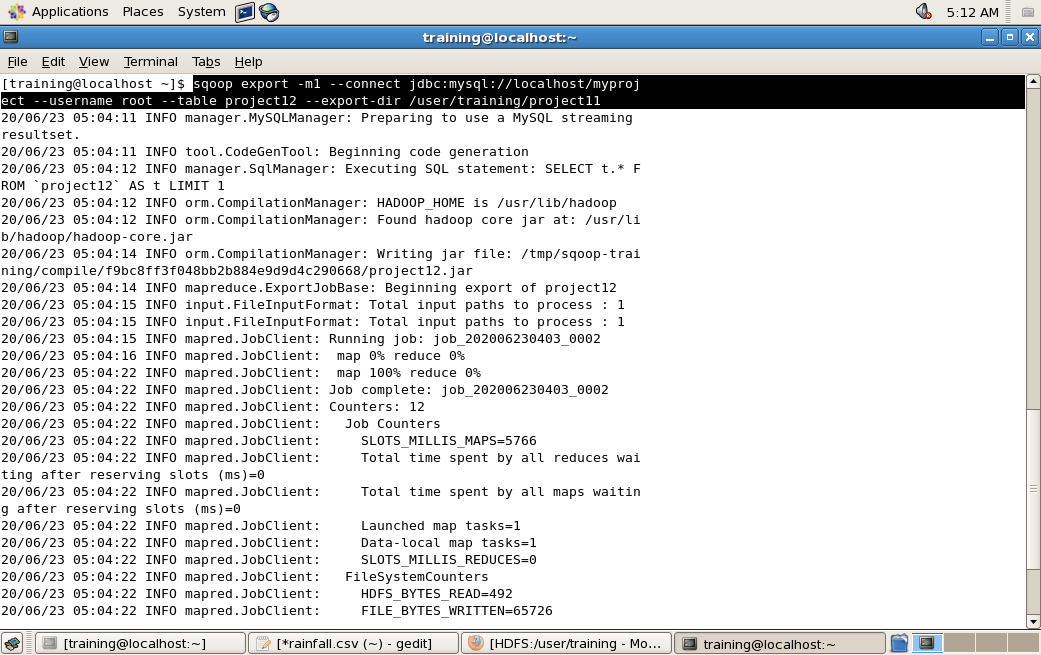
1. Open the terminal.

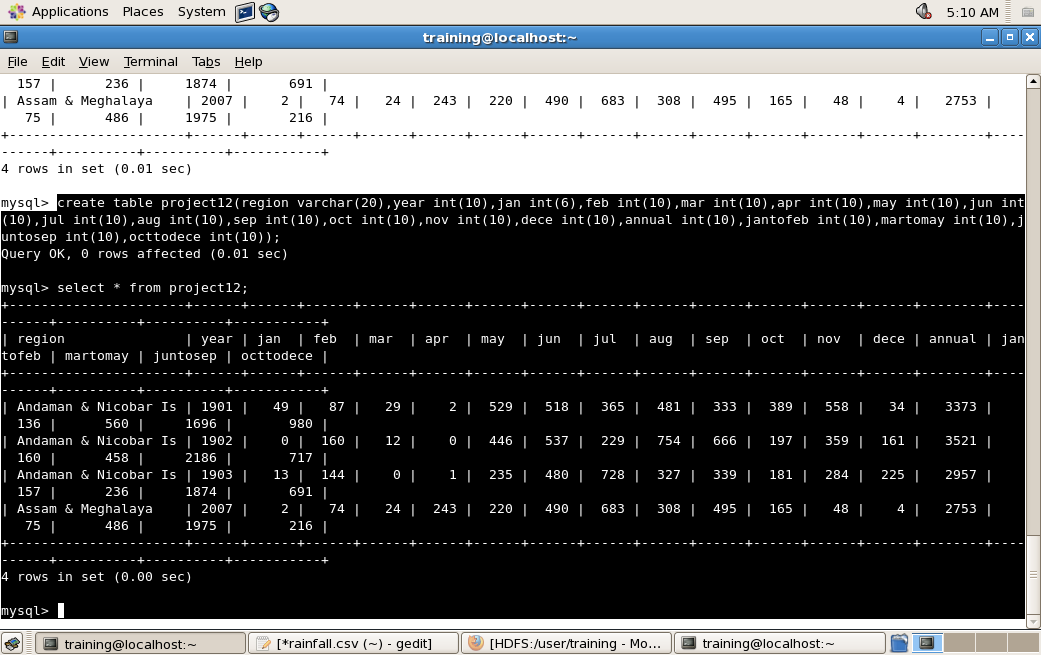
2. Now you need to create the table in sql and remain it empty.

3. Write the Sqoop export command on the terminal.

3. Check the file which we have exported on MySQL table.

*The snapshot contains Sqoop export command to put dataset from hdfs to MySQL.*





**3.3.** **Hive**

Load Data into hive environments.

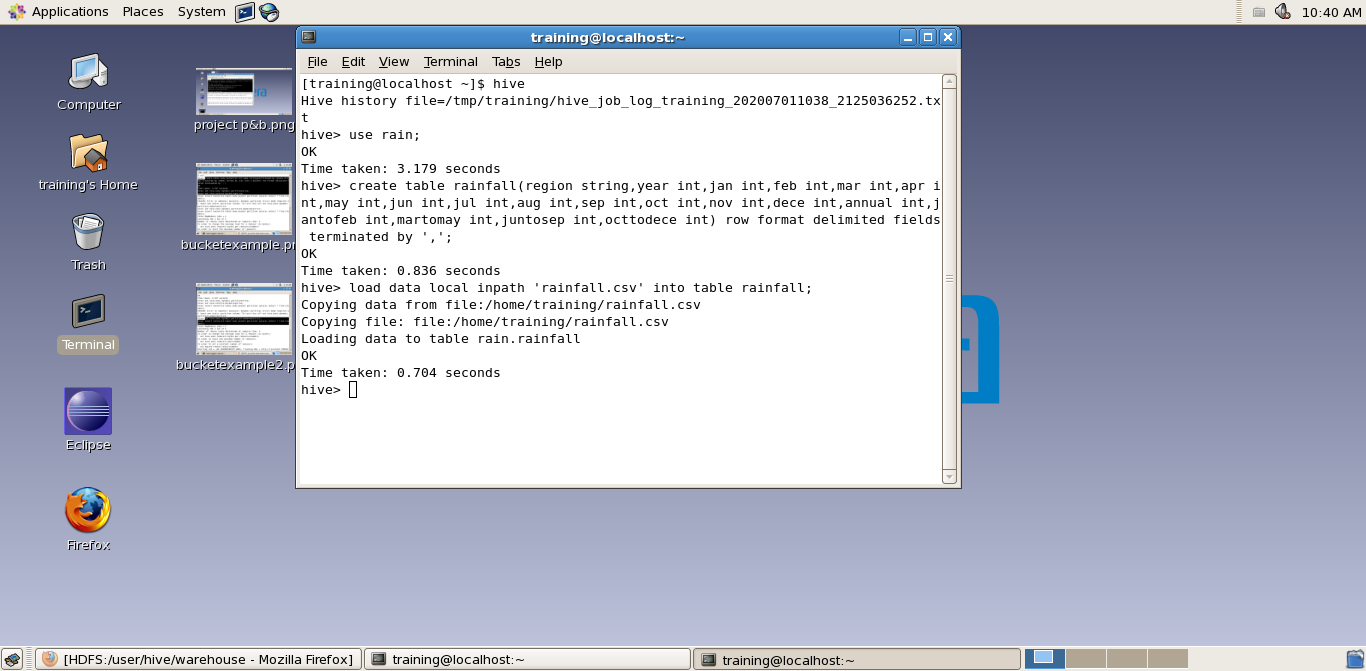
1. Open the terminal.

2. Write the Hive command on the terminal.

3. This will open the hive shell.

4. Now you need to create the table in hive and load data into it from HDFS.

*The snapshot contains hive command to put dataset from local to hive.*



**3.4 PIG**

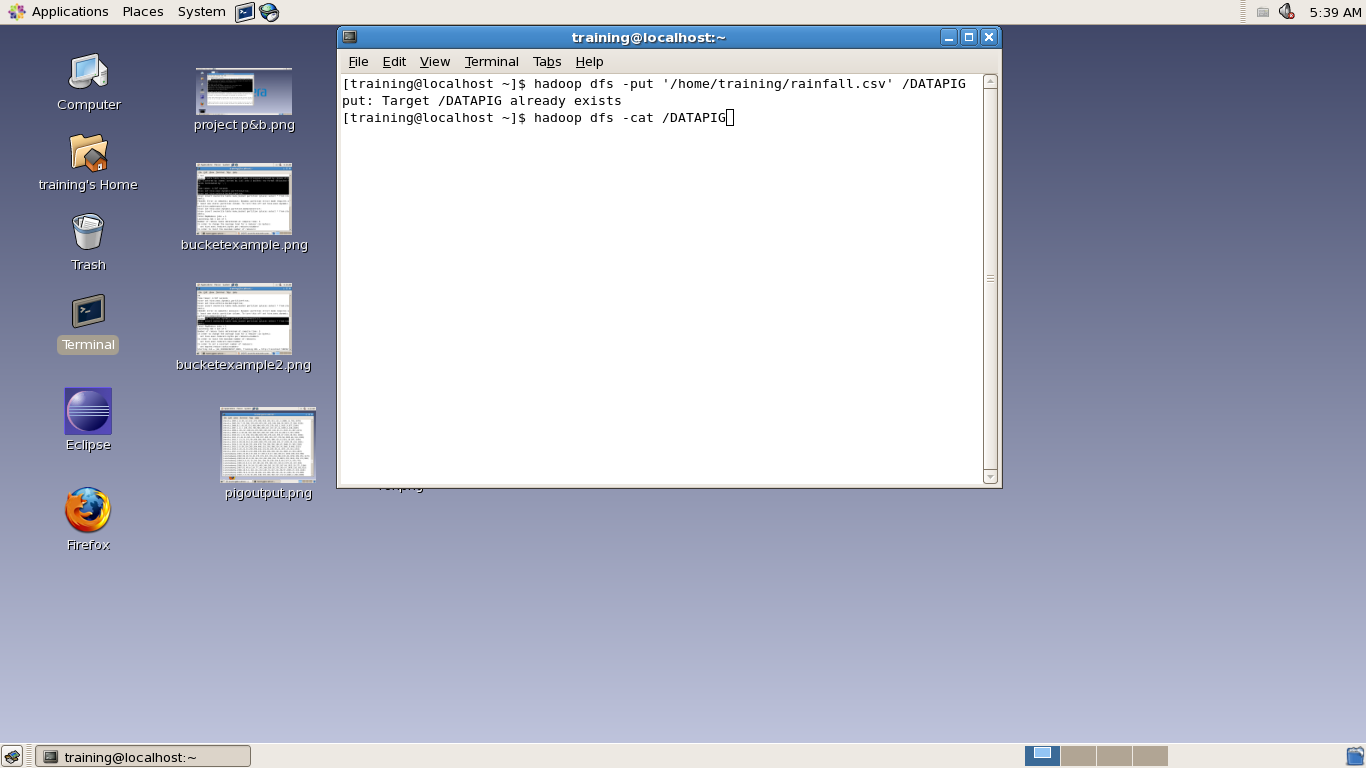
Load Data into Pig environments.

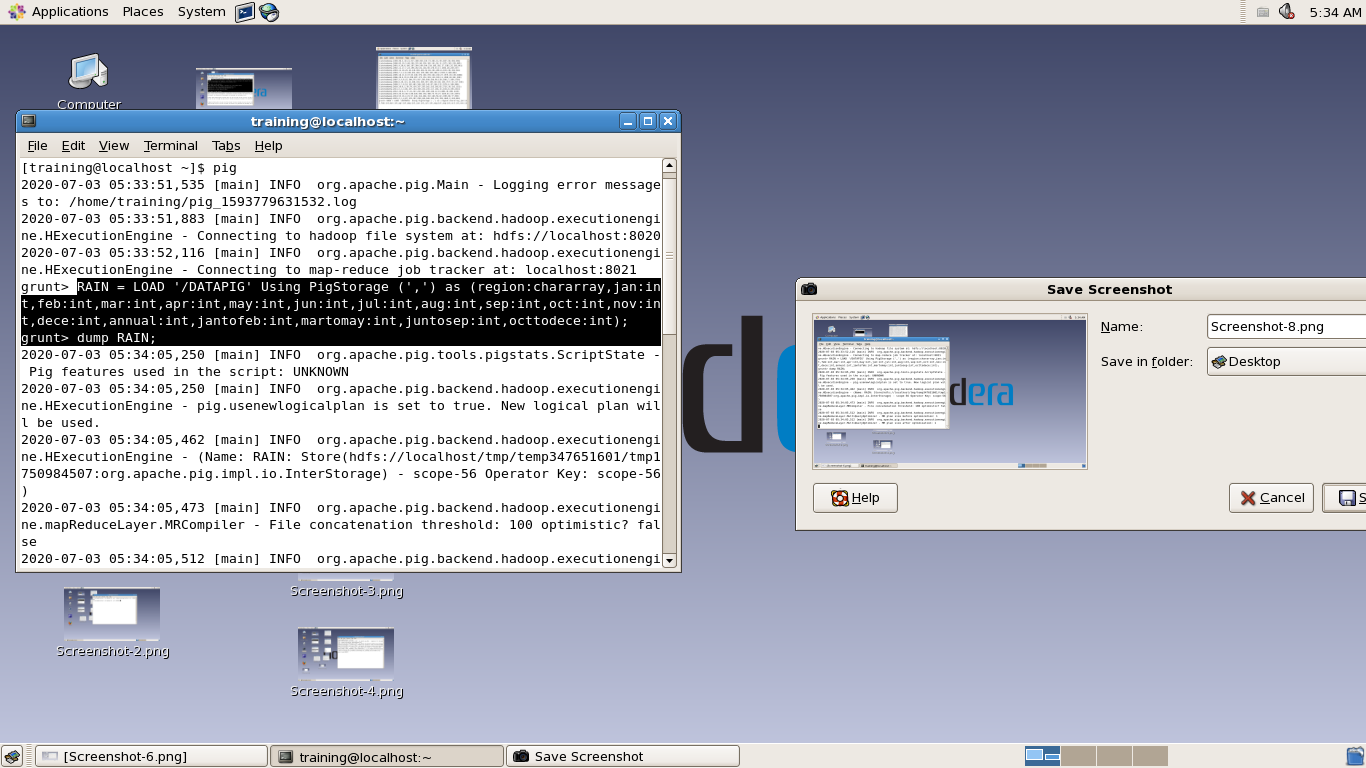
1. Open the terminal.

2. Write the Pig Command (pig –x local) on the terminal.

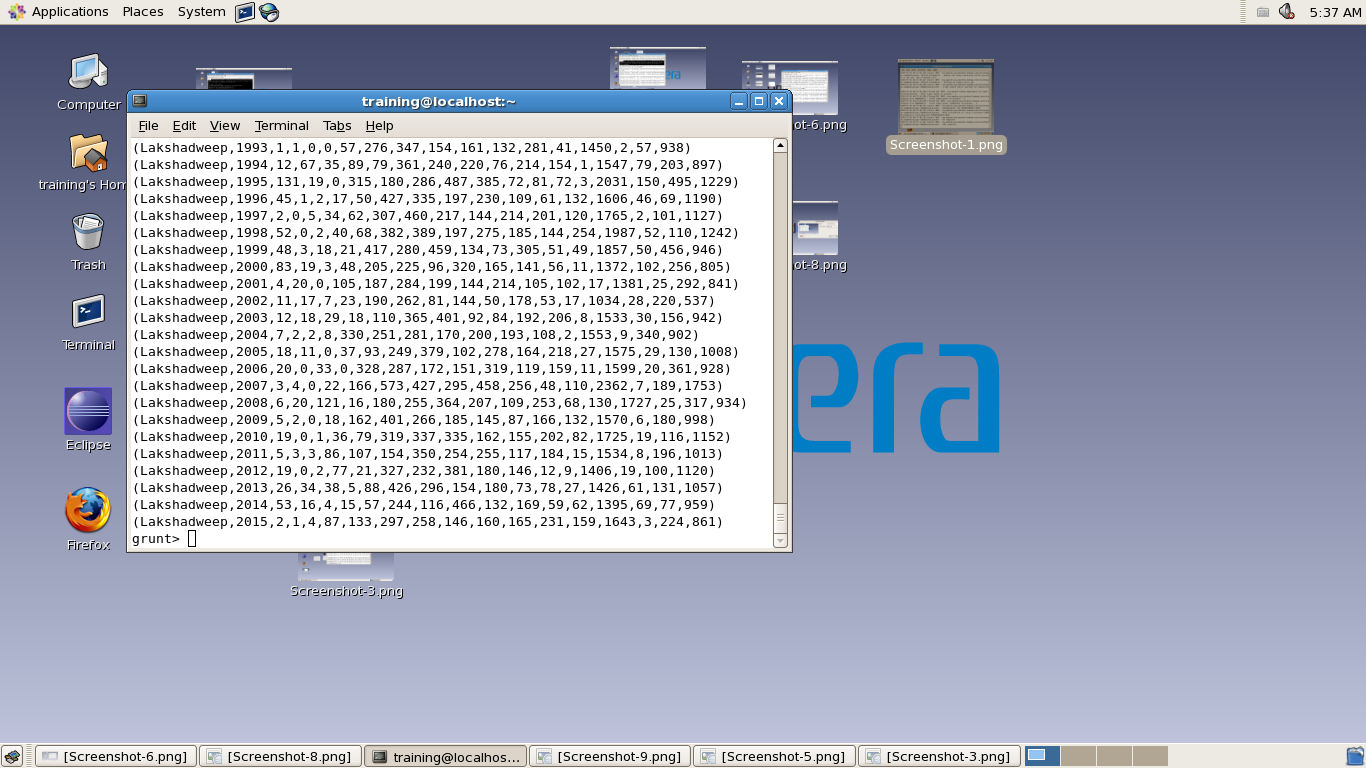
3. This will open the grunt shell in the local mode.

4. Now you need to create a table and load data into it from local file system.





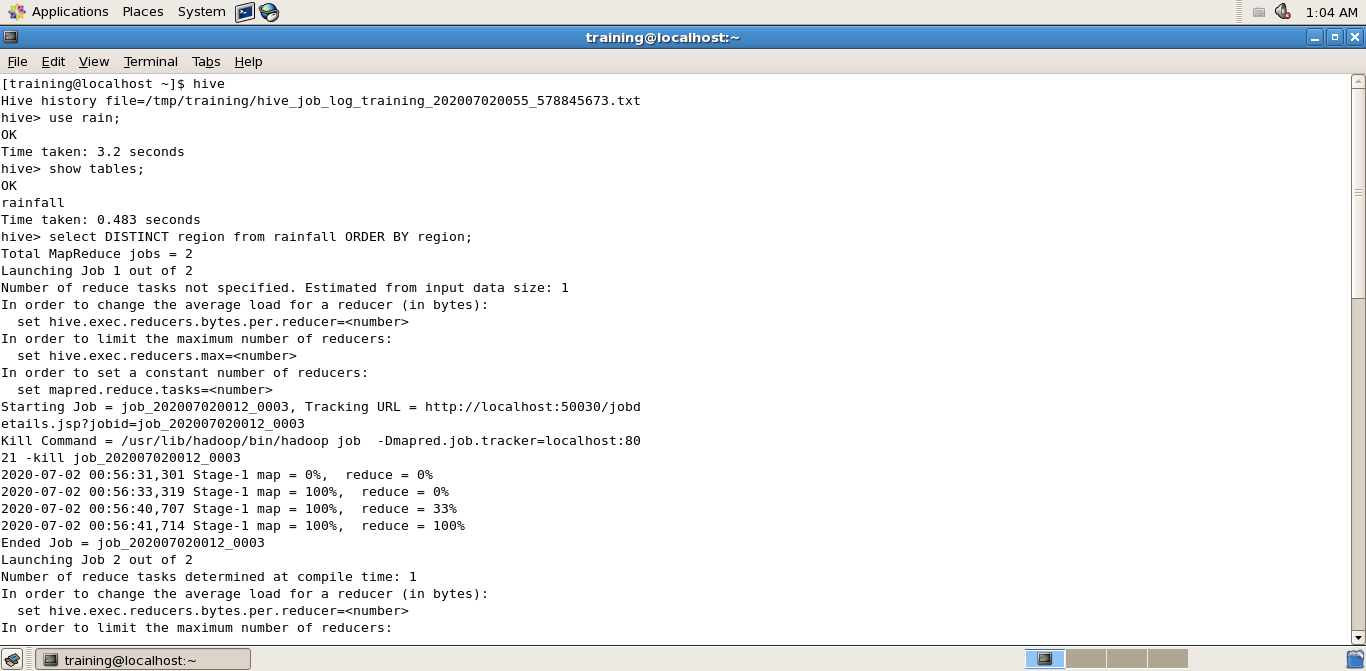


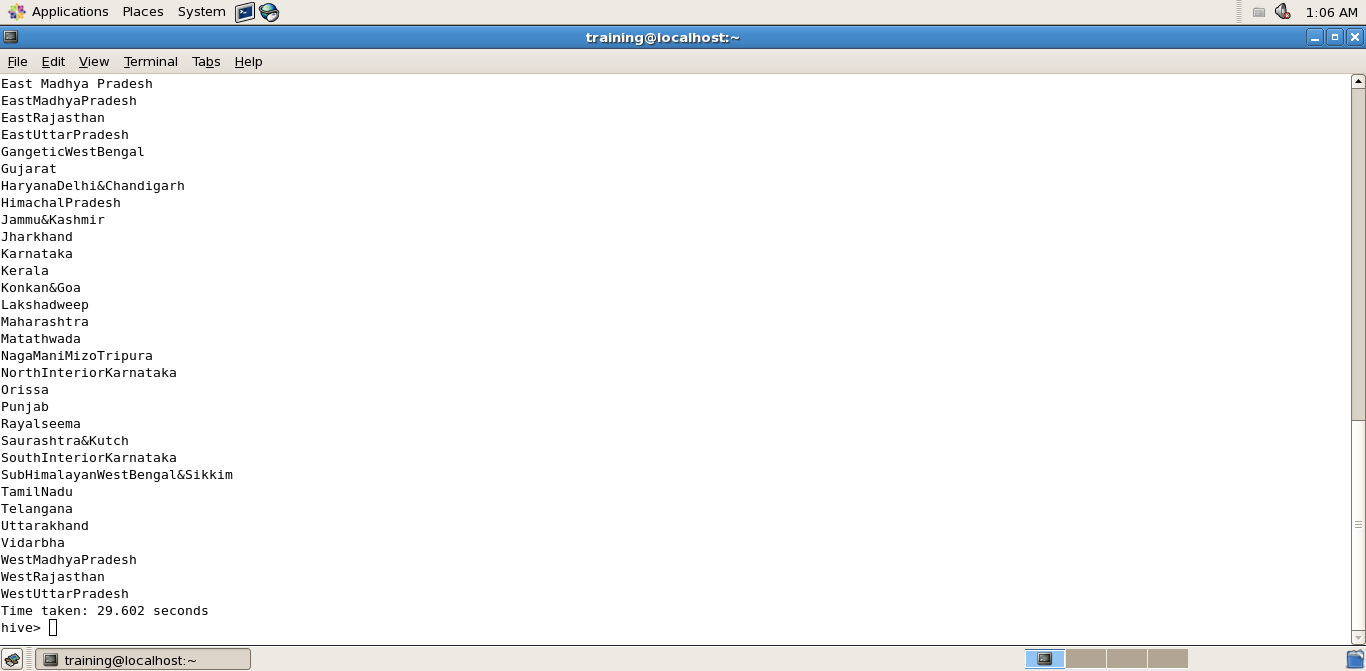


**4. QUERIES**

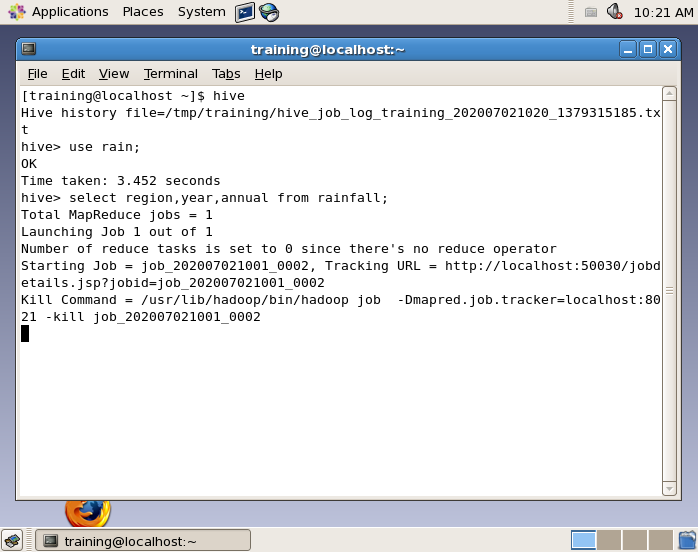
1. Name all states available in this dataset.
2. Annual rainfall in all states from 1901 to 2017.
3. Rainfall in particular state between jan-feb in time-period of 2000-2017.
4. Rainfall in particular state between July and December in particular year.
5. Annual rainfall in all states within last 50 years.
6. Show maximum rainfall of all states in a time period of 1901-2017.
7. Average rainfall on particular state within last 15 years.
8. Table that shows rainfall does not occur in respective months along with respective year.
9. What’s the value of maximum annual rainfall?
10. Which year and state have annual rainfall equal to 8999?
11. In which year and states have annual rainfall greater than 6000?
12. Average of annual rainfall occurs in all states from 1901 to 2017.
13. Average of rainfall in each month in each region.

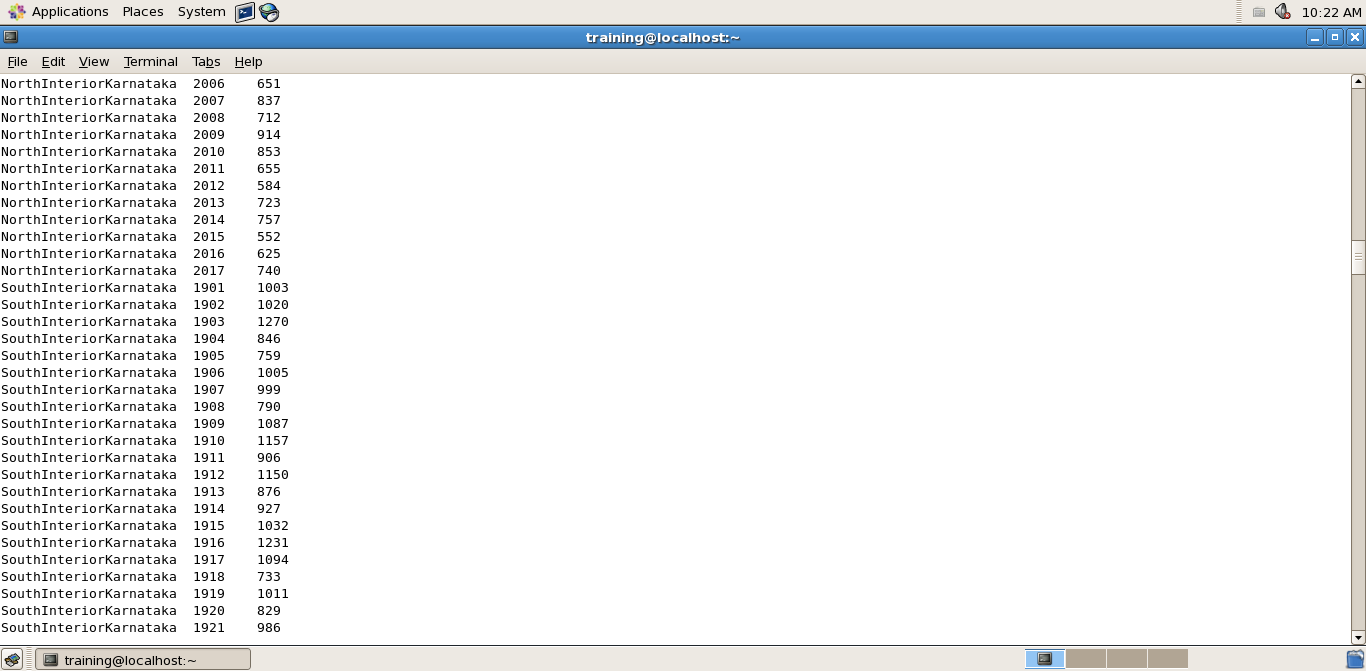
Query 1: Name all states available in this dataset.

****



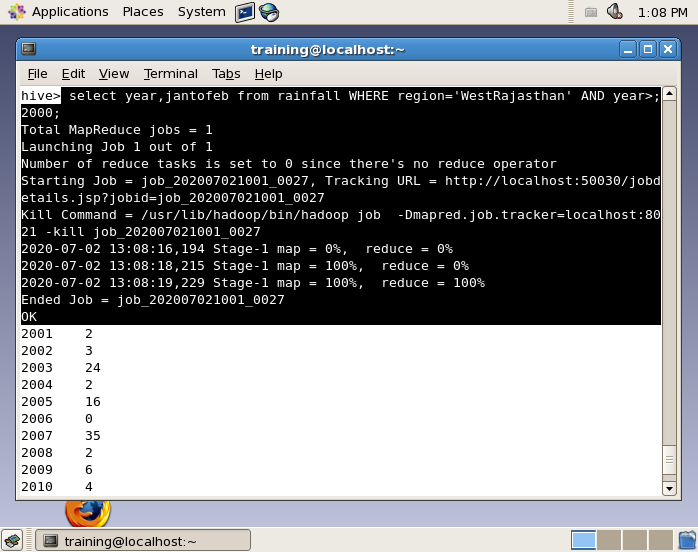
Query 2: Annual rainfall in all states from 1901 to 2017.

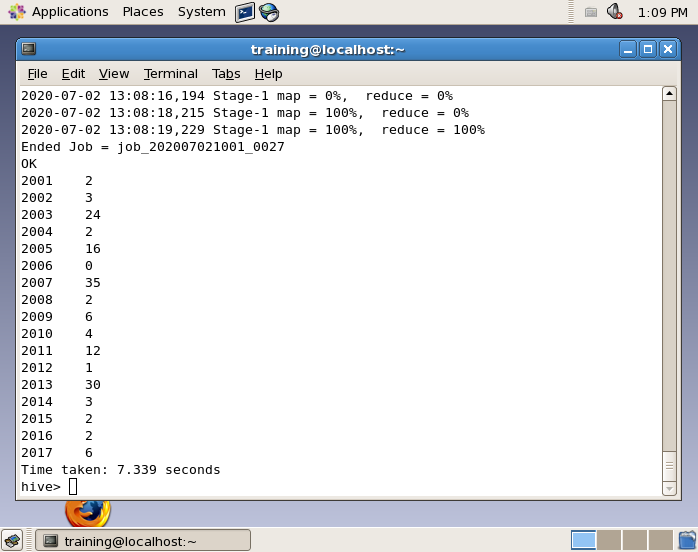
****

****

Query 3: Rainfall in particular state between jan-feb in time-period of 2000-2017.

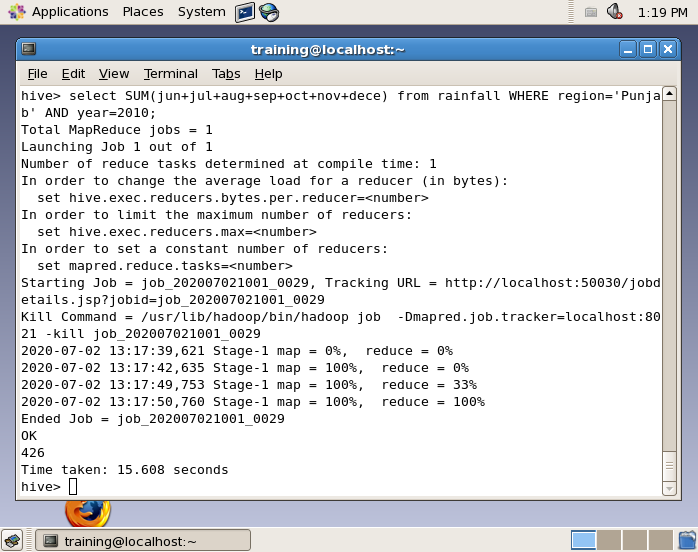
*In the below snapshot, we execute the above query on region “WestRajasthan”.*



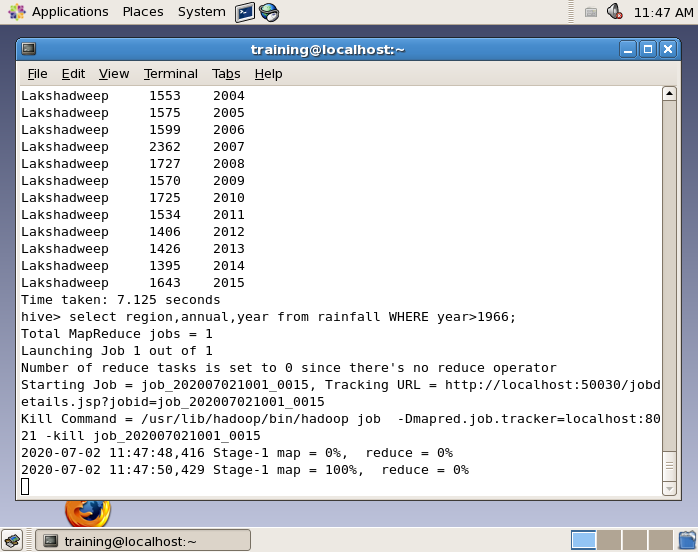
****

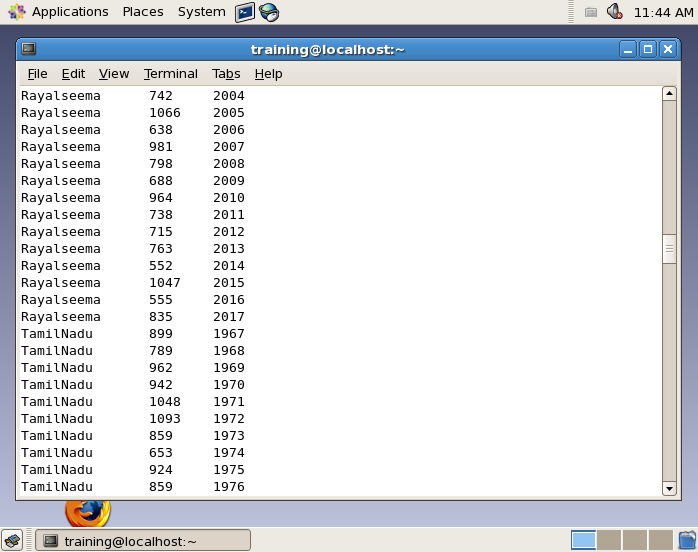
Query 4: Rainfall in particular state between July and December in particular year.

*In the below snapshot, we execute the above query on region “Punjab”.*

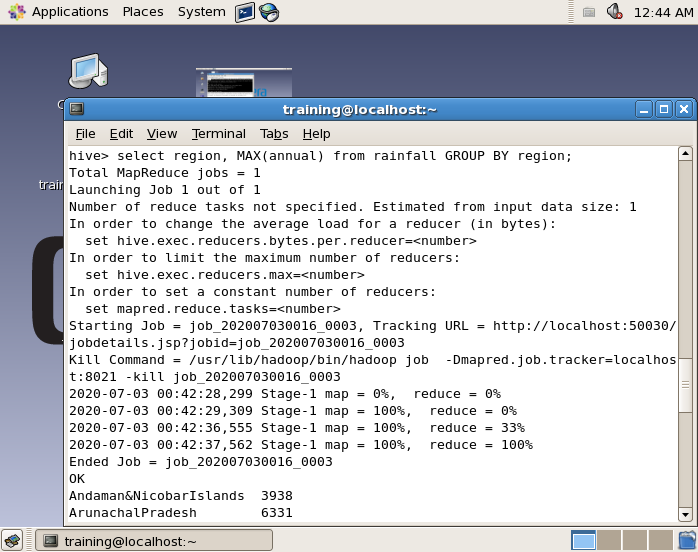
****

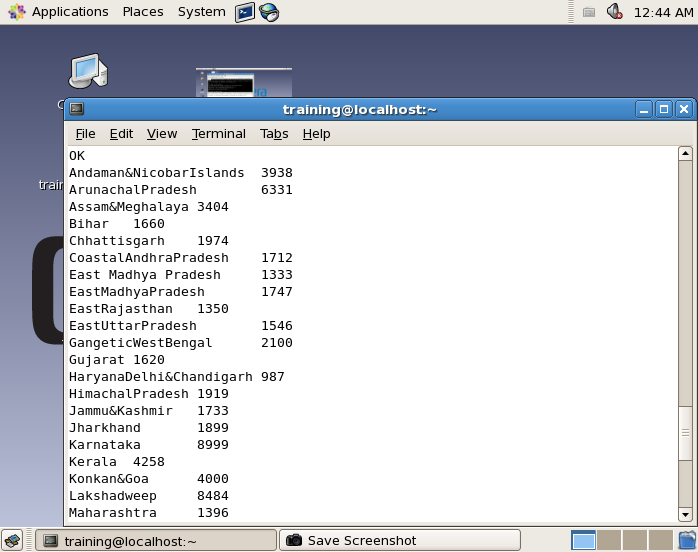
Query 5: Annual rainfall in all states within last 50 years.

****

****

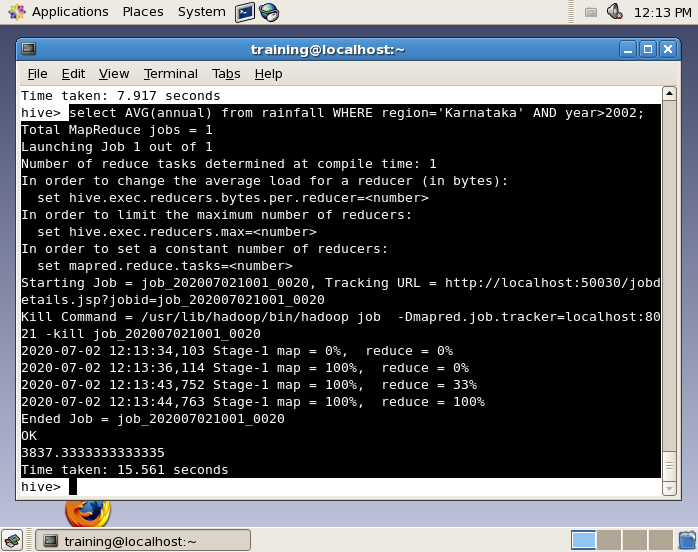
Query 6: Show maximum rainfall of all states in a time period of 1901-2017.



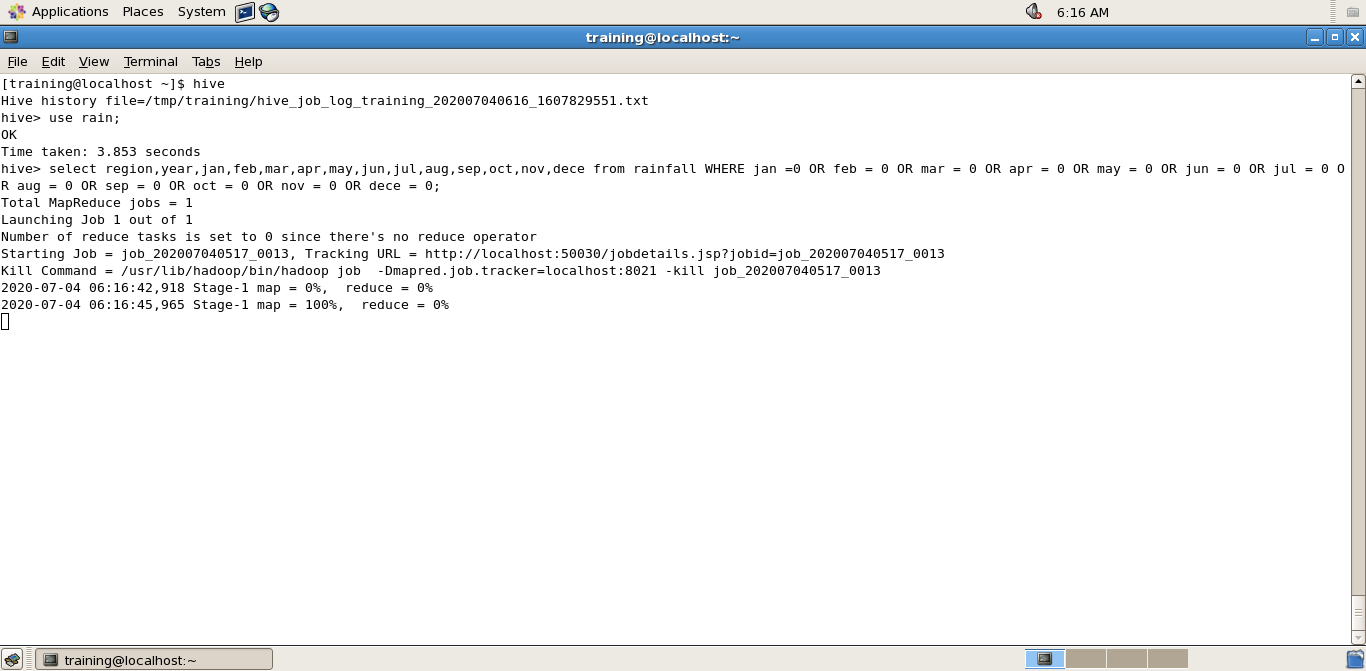


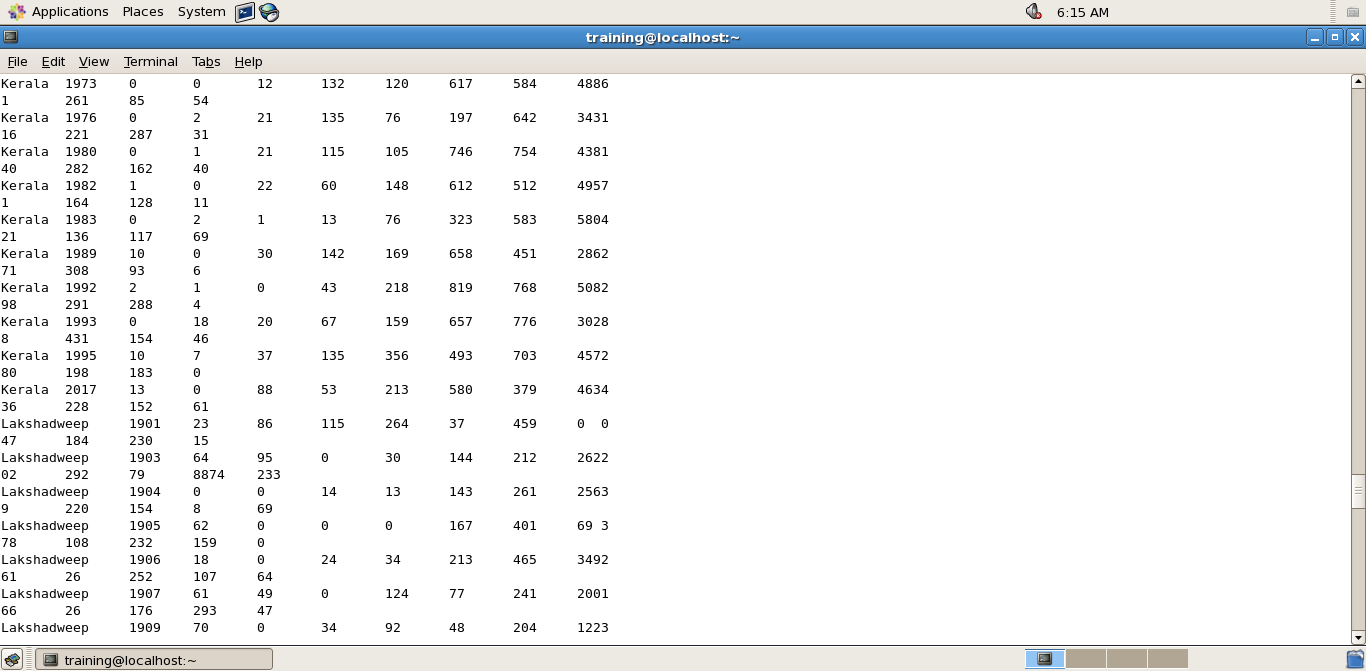
Query 7: Average rainfall on particular state within last 15 years.

*In the below snapshot, we execute the above query on region “Karnataka”.*

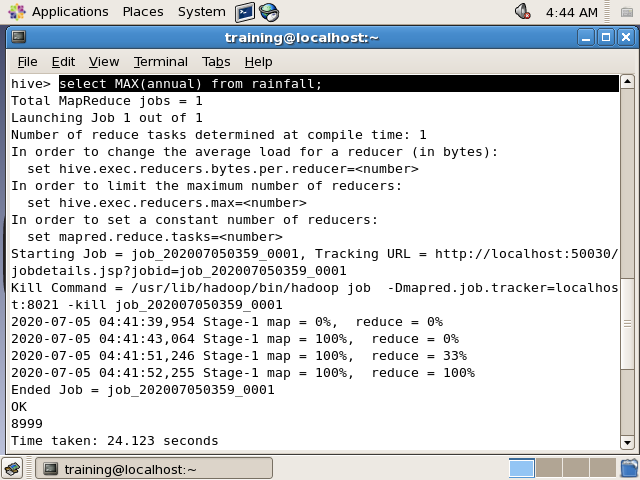


Query 8: Table that shows rainfall does not occur in at least one month along with their respective regions.

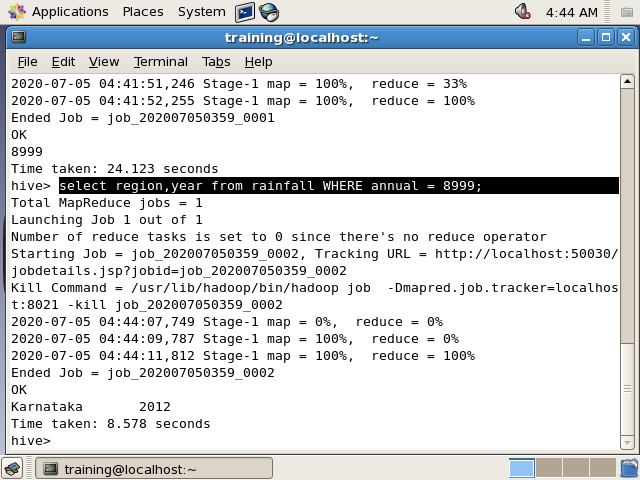


****

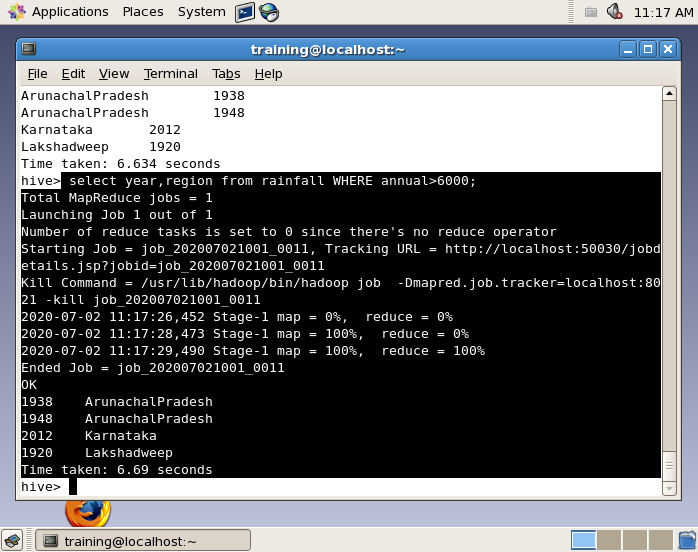
Query 9: What’s the value of maximum annual rainfall.

****

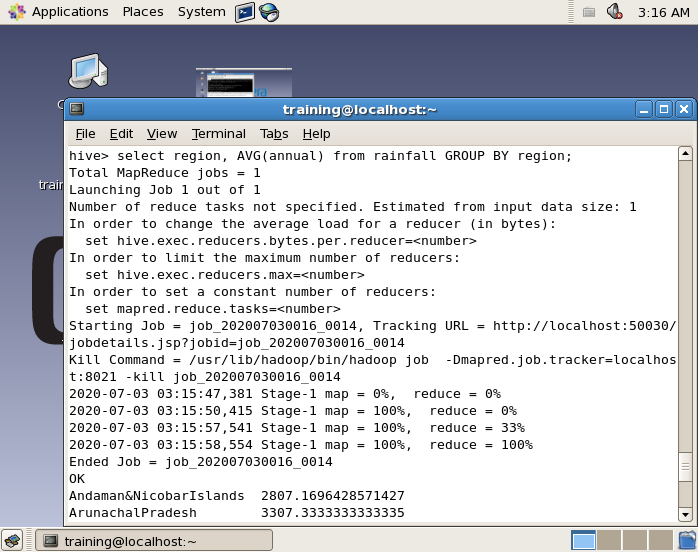
Query 10: which year and state have annual rainfall equal to 8999.

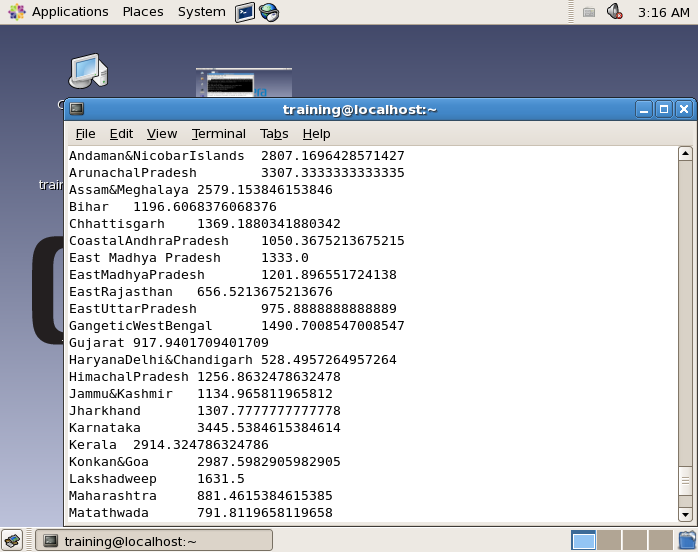
****

Query 11: In which year and states have annual rainfall greater than 6000?

****

Query 12: Average of annual rainfall occurs in all states from 1901 to 2017.

****

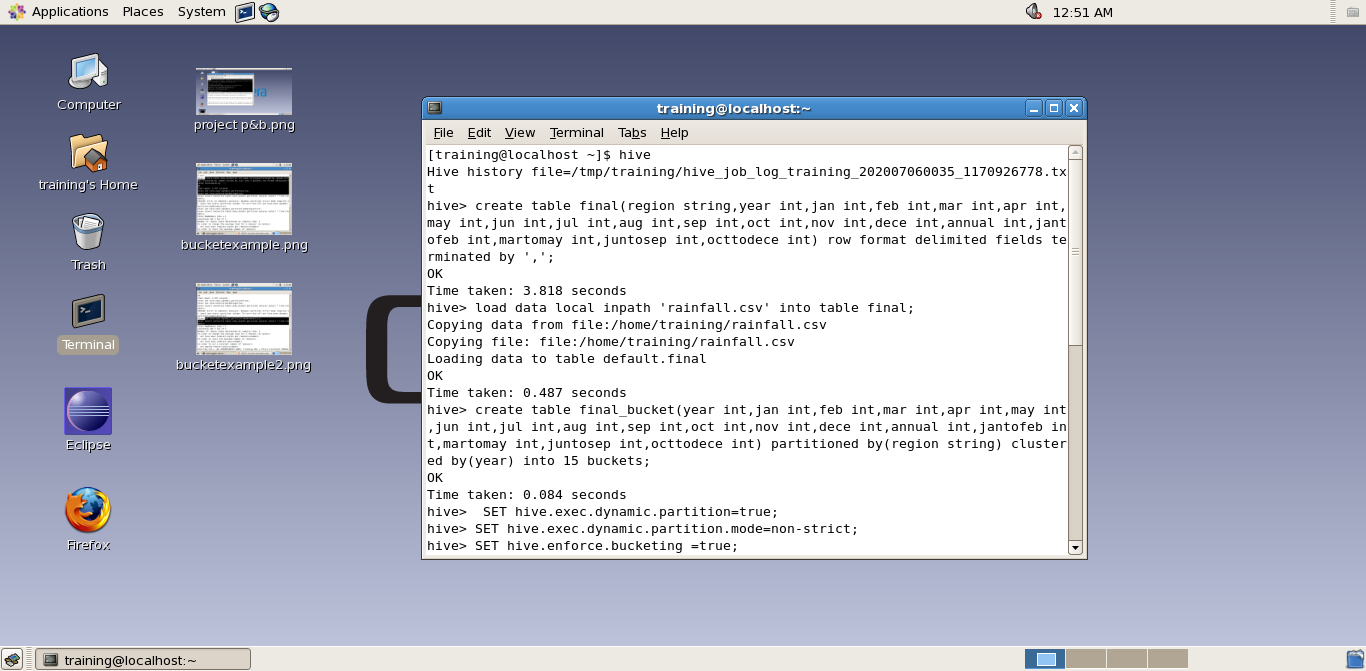
****

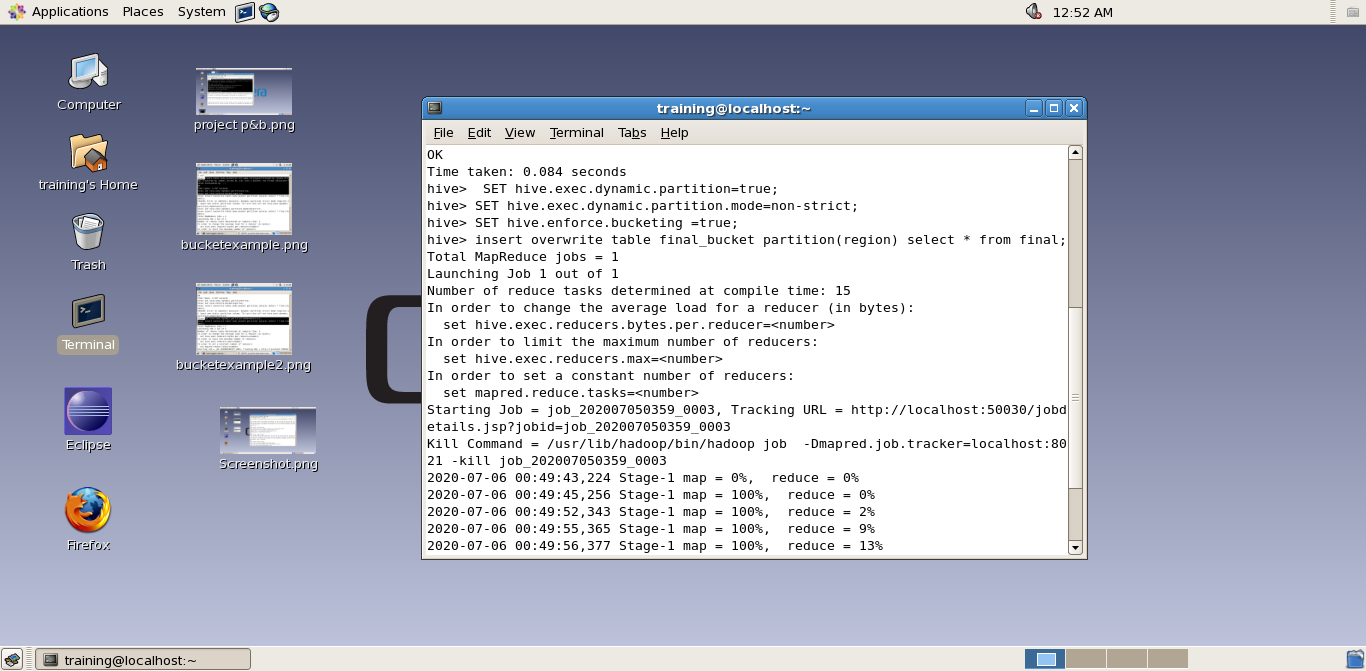
Query 13: Average of rainfall in each month in each region.

****

**5. PARTION WITH BUCKETING**

Hive Buckets is nothing but another technique of decomposing data or decreasing the data into more manageable parts or equal parts. Hive organizes tables into Partitions. It is a way of dividing a table into related parts based on the values of partitioned columns. Using partition, it is easy to query a portion of the data.

****

****

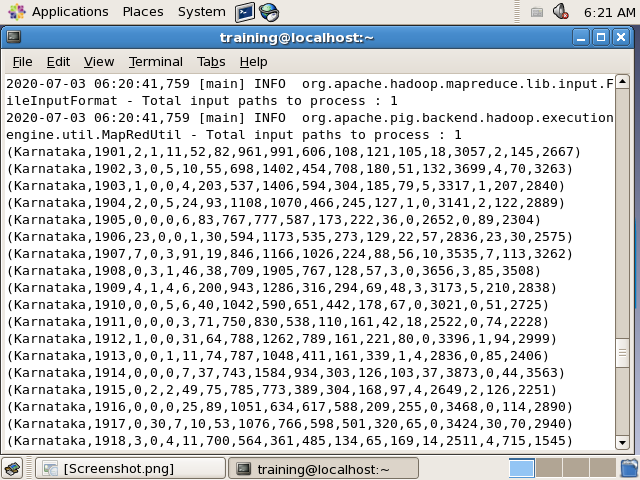
**6. PIG QUERIES**

1. List the data of Karnataka.
2. List the region group by annual.
3. List the first 100 details from the data.
4. Show the data using pig according to months from January to December.

**QUERY 1: List the data of Karnataka.**

grunt> query1 = filter RAIN by region == ‘Karnataka’;

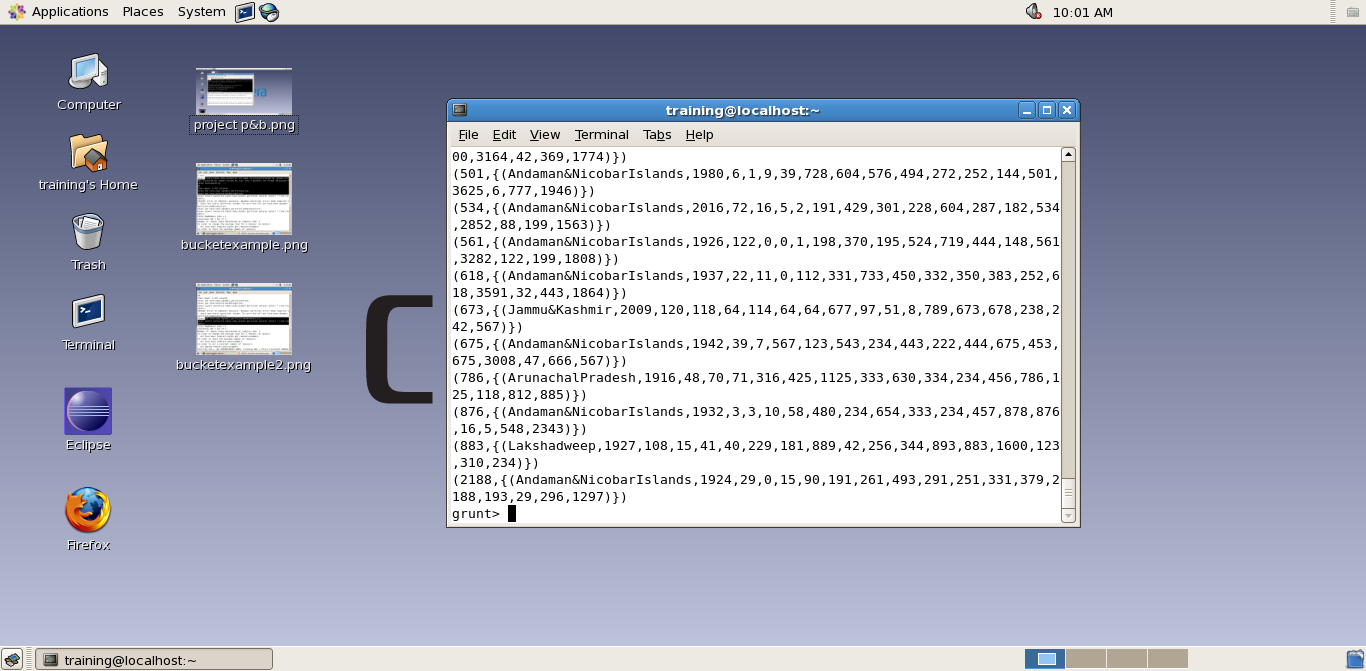
grunt> dump query1;

****

**QUERY 2: List the region group by annual.**

grunt> query2 = group RAIN by annual;

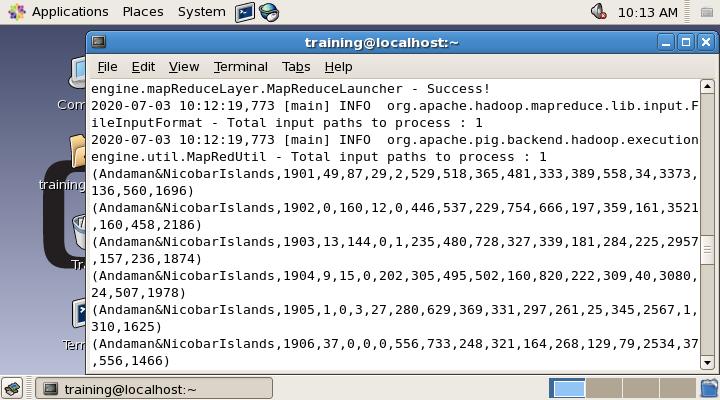
grunt> dump query2;

****

**QUERY 3: List the first 100 details from the data.**

grunt> query3 = limit RAIN 100;

grunt> dump query3;

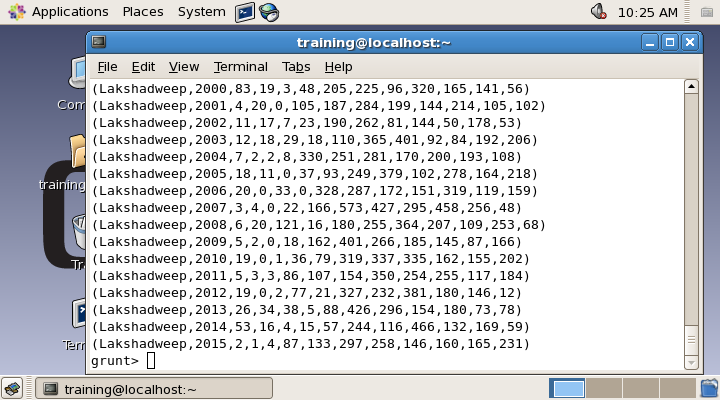
****

**QUERY 4: Show the data using pig according to months from January to December.**

grunt> query4 = foreach RAIN generate region,jan,feb,mar,apr,may,jun,jul,aug,sep

,oct,nov,dece;

grunt> dump query4;



**5. BIBLIOGRAPHY**

I have taken the data of this project from various sites that helps me to improve my dataset and project a lot.

LINKS:

<https://www.kaggle.com/tjjohn/rainfall>

<https://www.guru99.com/introduction-to-pig-and-hive.html>

<https://www.whizlabs.com/blog/big-data-analytics-importance/>

<https://en.wikipedia.org/wiki/Big_data>