**rCSDAIA24GP003**

**JAYANTH GOLTI (2320397)**

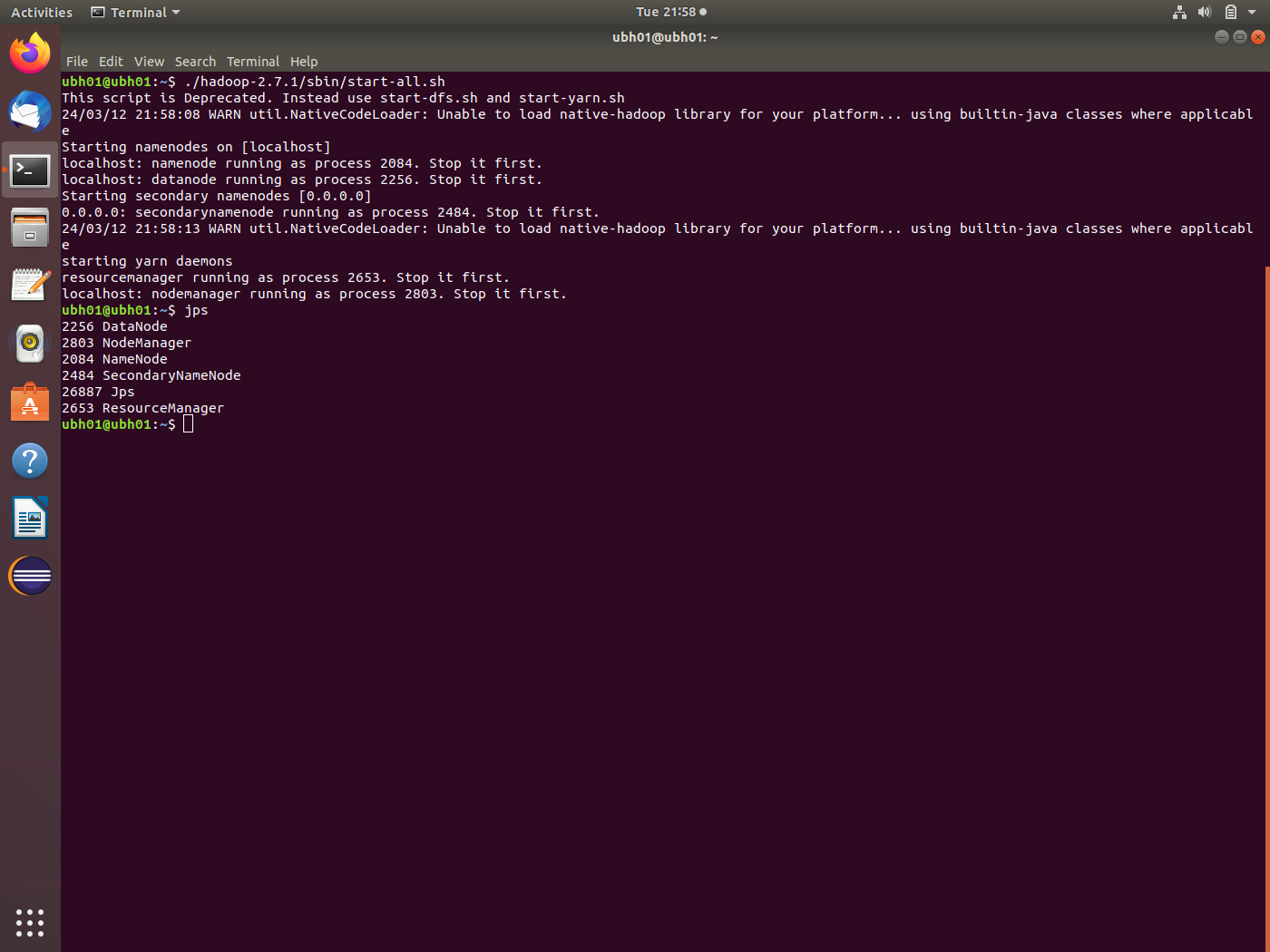
**BIG DATA**

To start the Hadoop we use – **./Hadoop-2.7.1/sbin/start-all.sh**

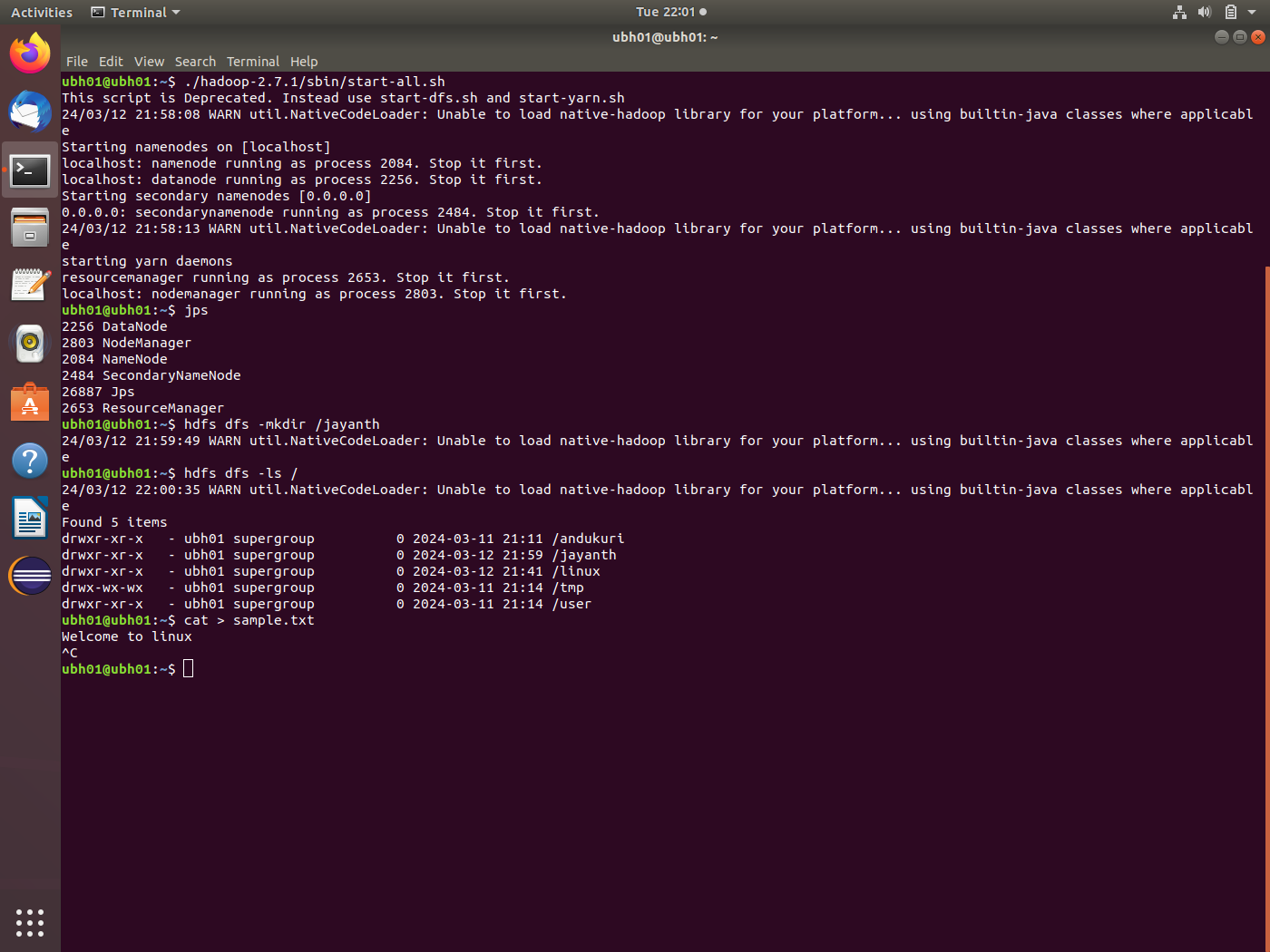
**A screenshot of a computer

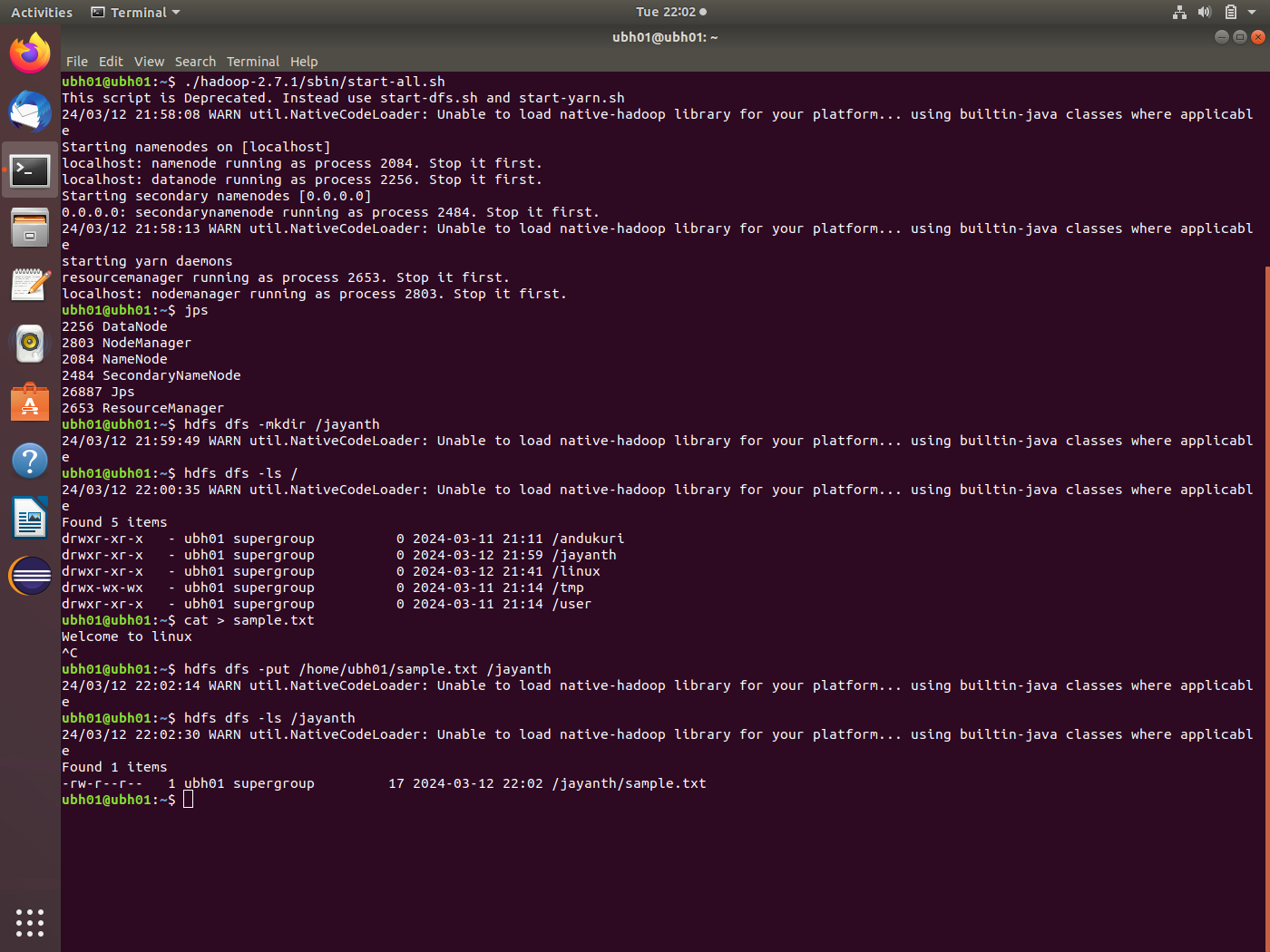
Description automatically generated**

To check whether Hadoop daemons are running or not by --- **jps.**

****

Created a Hadoop directory named Jayanth, and a text file is created named as sample.txt and checked the items in Jayanth directory by **hdfs dfs -ls /Jayanth.**

****

****

After that local file is copied to Hadoop by put command,

Hdfs dfs -put /home/ubh01/sample.txt /Jayanth.

Start MYSQL and create a database and use that database.

A screenshot of a computer

Description automatically generated

Create table with table name student.

Create table student(name varchar(20),roll int primary key, age int);

Inserted the values manually or Bulk load with below command:

Load data local infile ‘/home/ubh01/Desktop/student.csv’ into table student fields terminated by ‘,’ ignore 1 rows;

Syntax:

Insert into student values(“Priyanaka”,1,12);

Insert into student values(“Priya”,4,11);

Insert into student values(“Priyaswi”,6,31);

* To display the table:

Select \* from student;

* exit

**SQOOP:**

It is a command line tool that helps to import and export data from database.

* **Sqoop import:**

Sqoop import is used to import the data from RDBMS to HDFS.

**\*\*\*To start the Hadoop we use ./hadoop-2.7.1/sbin/start-all.sh**

* By sqoop import command import the file into the hdfs by using mapreduce

A computer screen shot of text

Description automatically generated

* Listing the output files of sqoopdata file which includes \_SUCCESS and PART FILES using ls command.
* Display the content inside PART FILE using cat command.

A computer screen shot of text

Description automatically generated

* Sqoop import command with **where** clause.

A screen shot of a computer code

Description automatically generated

* If we want to verify whether values that satisfies where condition imported to hdfs or not, run the below hdfs command **--hdfs dfs -cat /sqoopdata2/part\***

A computer screen shot of white text

Description automatically generated

* Sqoop import with –query command where we will specify mysql query and \$CONDITIONS that is a placeholder in query which divide the data with more where clauses internally and –split-by is used to split the data by particular column that is primary key.

A computer screen shot of text

Description automatically generated

* If we want to verify whether values that satisfies QUERY imported to hdfs or not, run the below hdfs command **--hdfs dfs -cat /sqoopdata3/part\***

A screenshot of a computer program

Description automatically generated

* **Sqoop export**:
* Sqoop export is used to import the data from HDFS to RDBMS.
* Here we have to mention export keyword and –export-directory and before this we have to create a empty schema or table in mysql.

A screenshot of a computer program

Description automatically generated

* If we want to verify whether values are exported from hdfs to mysql or not by running the below command in mysql—select \* from students1;

A screenshot of a computer program

Description automatically generated

**HIVE**

Hive is a data warehouse system which is used for querying and analysing large datasets stored in HDFS.

In hive here we are downloading the dataset(student-dataset.csv) from github.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* To entering into the hive, use **hive** command.
* Create a database and use database.

A computer screen shot of text

Description automatically generated

* Here we are creating a hive table and inserted values into table using load command.
* To verify whether the data is loaded are not, use command**--select \* from student3;**

A computer screen shot of text

Description automatically generated

**Hive Partitioning**:

Create a partition table and follow either static or dynamic partitioning to insert values from main table following conditions.

* Static partitioning:
* It is setup when the table is created and remains static, meaning the partitions don’t change unless explicitly altered by the user.
* Here we are inserting values of a main table into partition table using static partitioning.

A computer screen shot of a computer error

Description automatically generated

* Here is the partition table for partition gender = ’M’ shown below

A computer screen shot of a computer code

Description automatically generated

* **Dynamic partitioning**:
  + It involves automatically creating partitions for data based on the values of specified columns during the insertion process.
  + Here we are inserting values of a main table into partition table using dynamic partitioning.

A computer screen shot of text

Description automatically generated

* To verify whether the data is loaded are not, use command**--select \* from dp;**

A computer screen with white dots

Description automatically generated

**Bucketing**

It is a method of dividing data within partitions into fixed-size buckets based on the hash value of a specified column. It helps distribute data evenly, Improving the query performance by reducing the data.

* Create a bucket table and mention how many buckets you need.
* Now insert main table values into bucket table using--

**insert into table buckettablename select \* from maintablename**

A computer screen shot of text

Description automatically generated

* Check for bucket creation in hive directory of hdfs:

**MYSQL:**

Mysql is initiated and an empty database named gcp created.

A screenshot of a computer

Description automatically generated

Using gcp database a table named **nycbike** created.

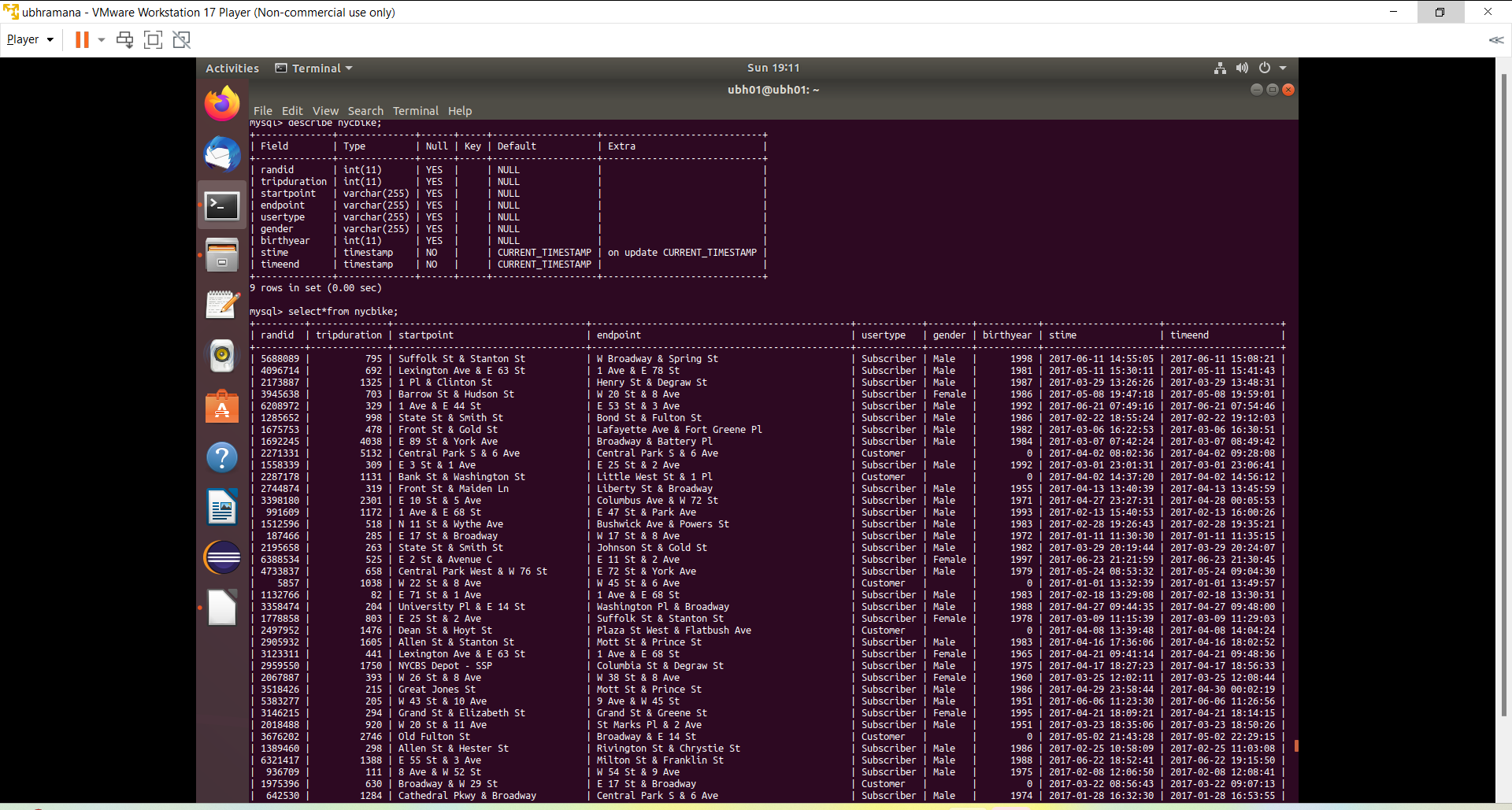
A screenshot of a computer

Description automatically generated

Using the below displayed command the csv data file is imported from local to mysql.A computer screen shot of a code

Description automatically generated

Ensured the data is stored(837 records are loaded).



**HADOOP:**

Initialization of Hadoop setup is completed and ensured.

A screenshot of a computer

Description automatically generated

**SQOOP:**

Using the below command the table nycbike from database gcp is imported to hdfs through sqoop.

*import*

Sqoop import –connectjdbc:mysql://localhost:3306/gcp –username root -p –query –table nycbike –target-dir /sqoopdata -m 1

A screenshot of a computer

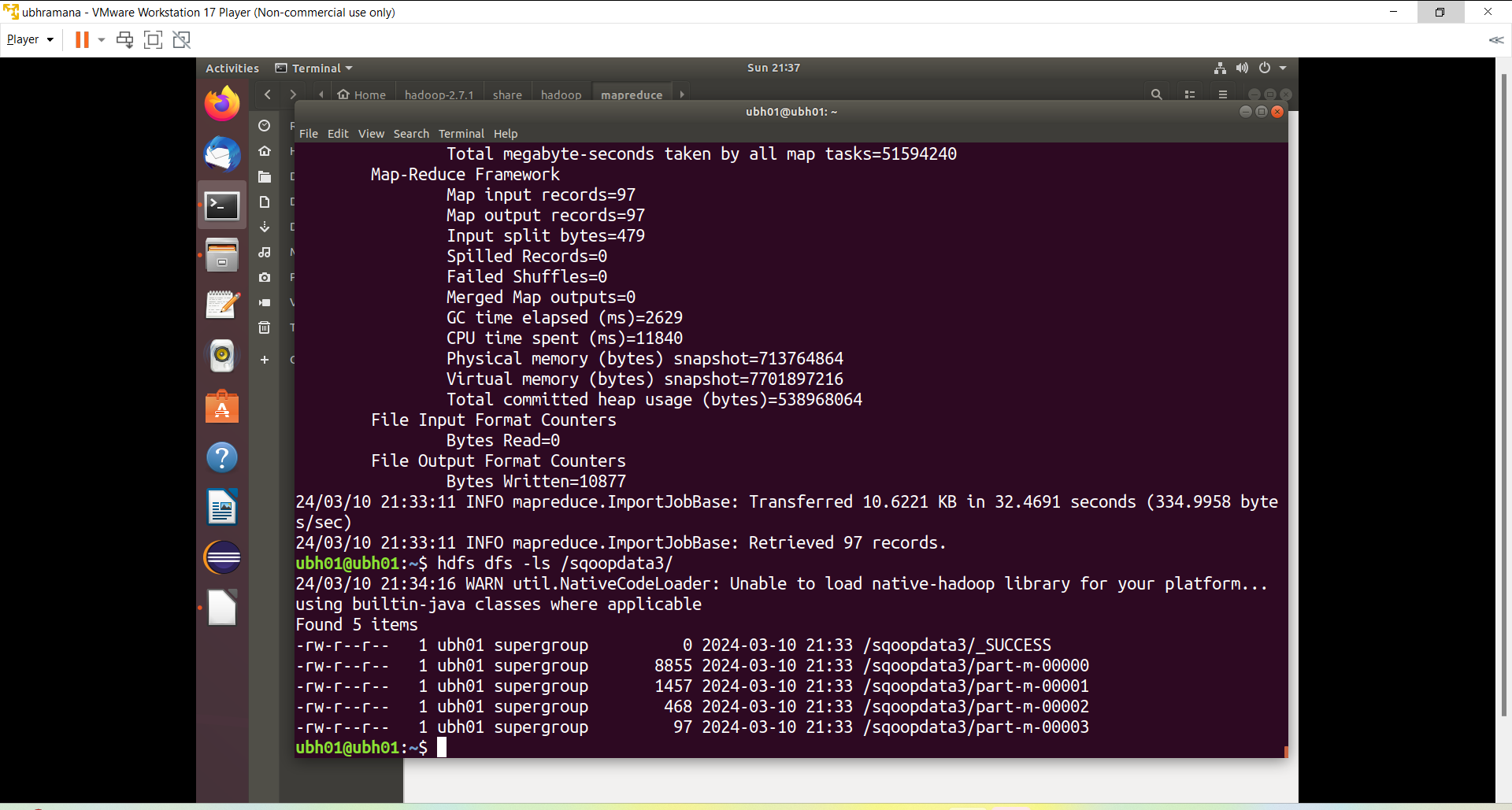
Description automatically generated

Sqoop import –connectjdbc:mysql://localhost:3306/gcp –username root -p –query –table nycbike –where “usertype =’Subscriber’” –target-dir /sqoopdata2 -m 1

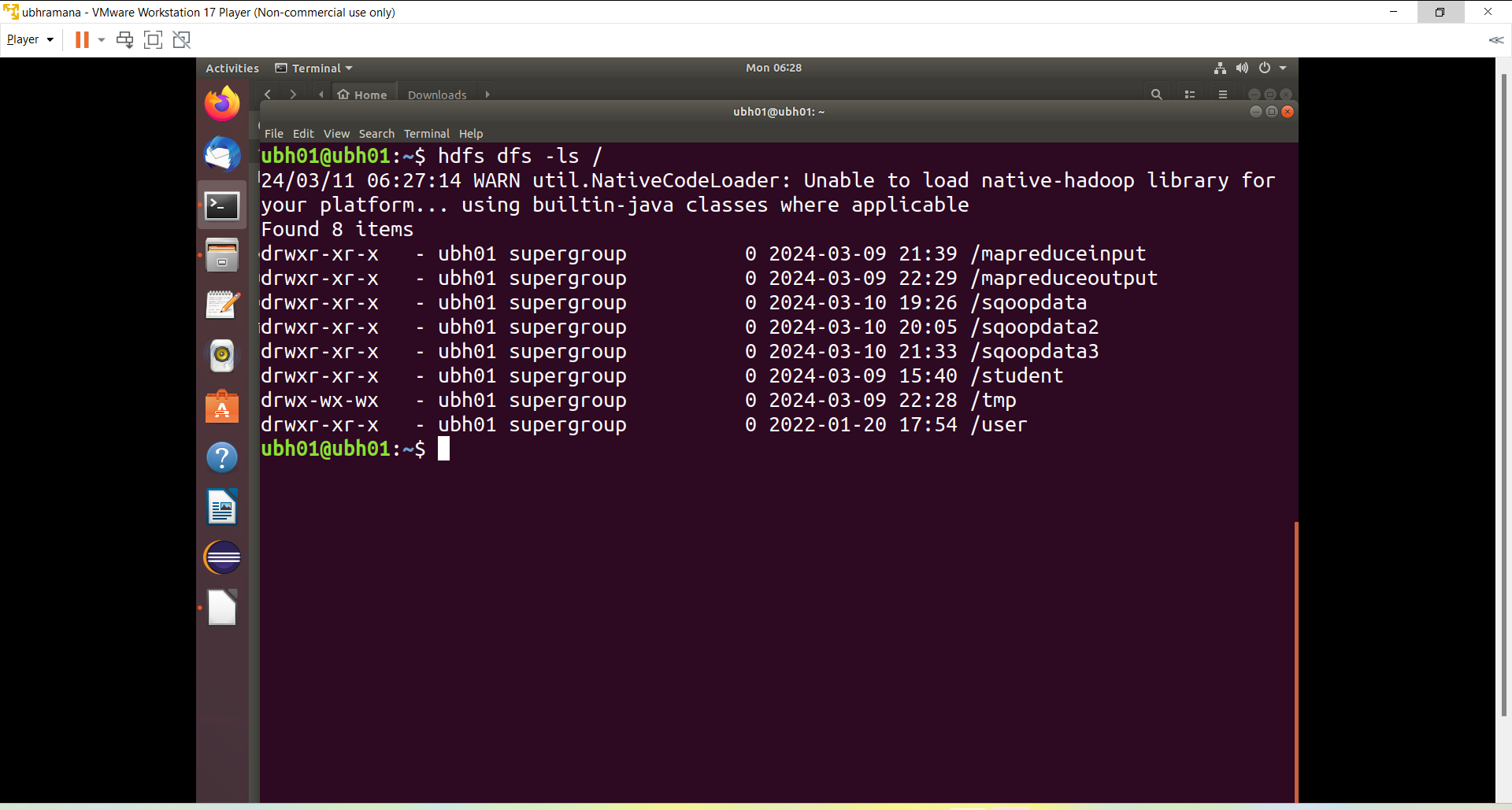
A screenshot of a computer screen

Description automatically generated

Sqoop import –connect jdbc:mysql://localhost:3306/gcp –username root -p –query “select\*from nycbike where usertype=’customer’ and \$CONDITIONS” –split-by tripduration –target-dir /sqoopdata3



Finally, the above all Sqoop command processes are stored in Hadoop and it is ensured (through hdfs dfs -ls / command)



**HIVE:**

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated