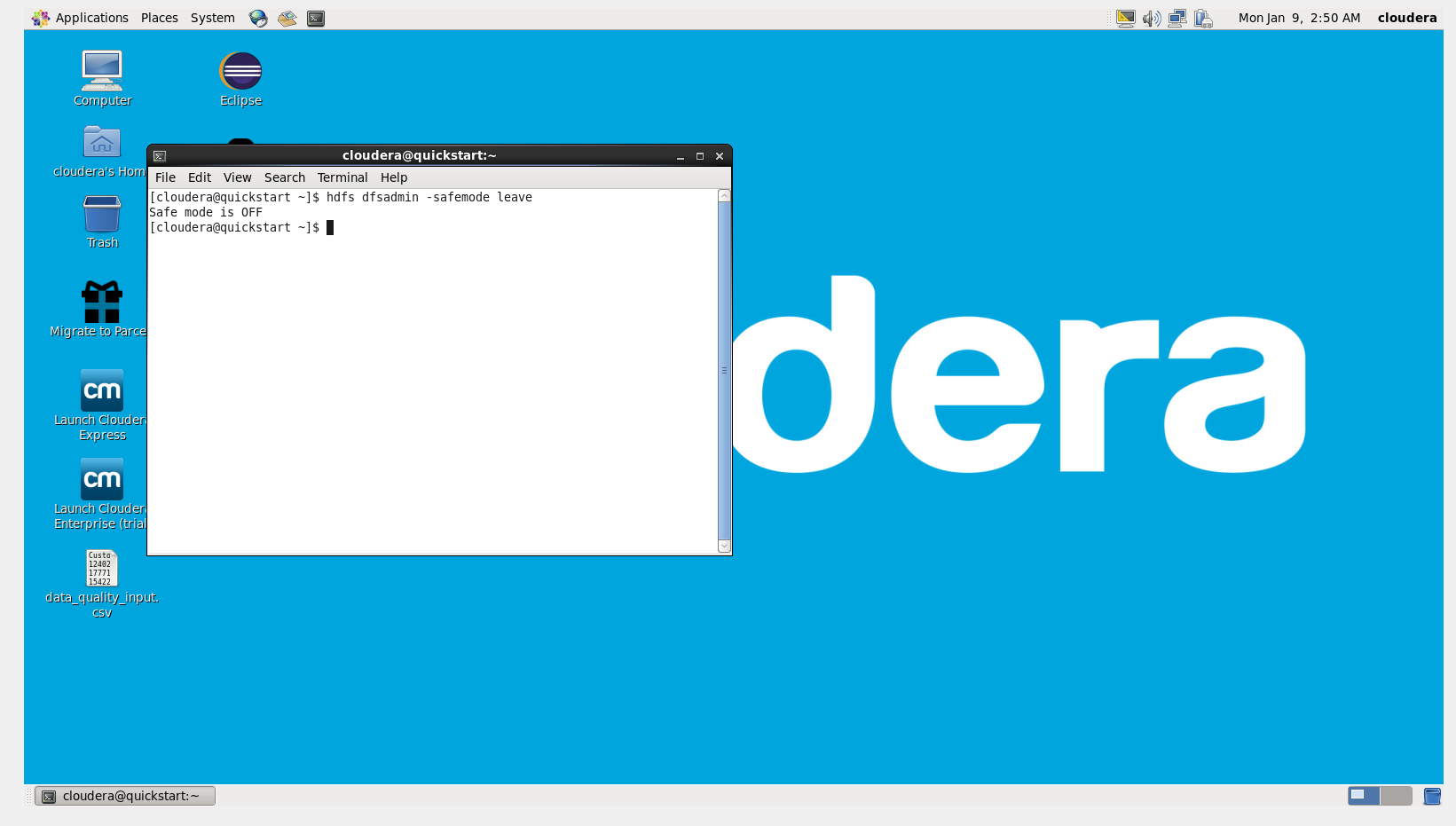
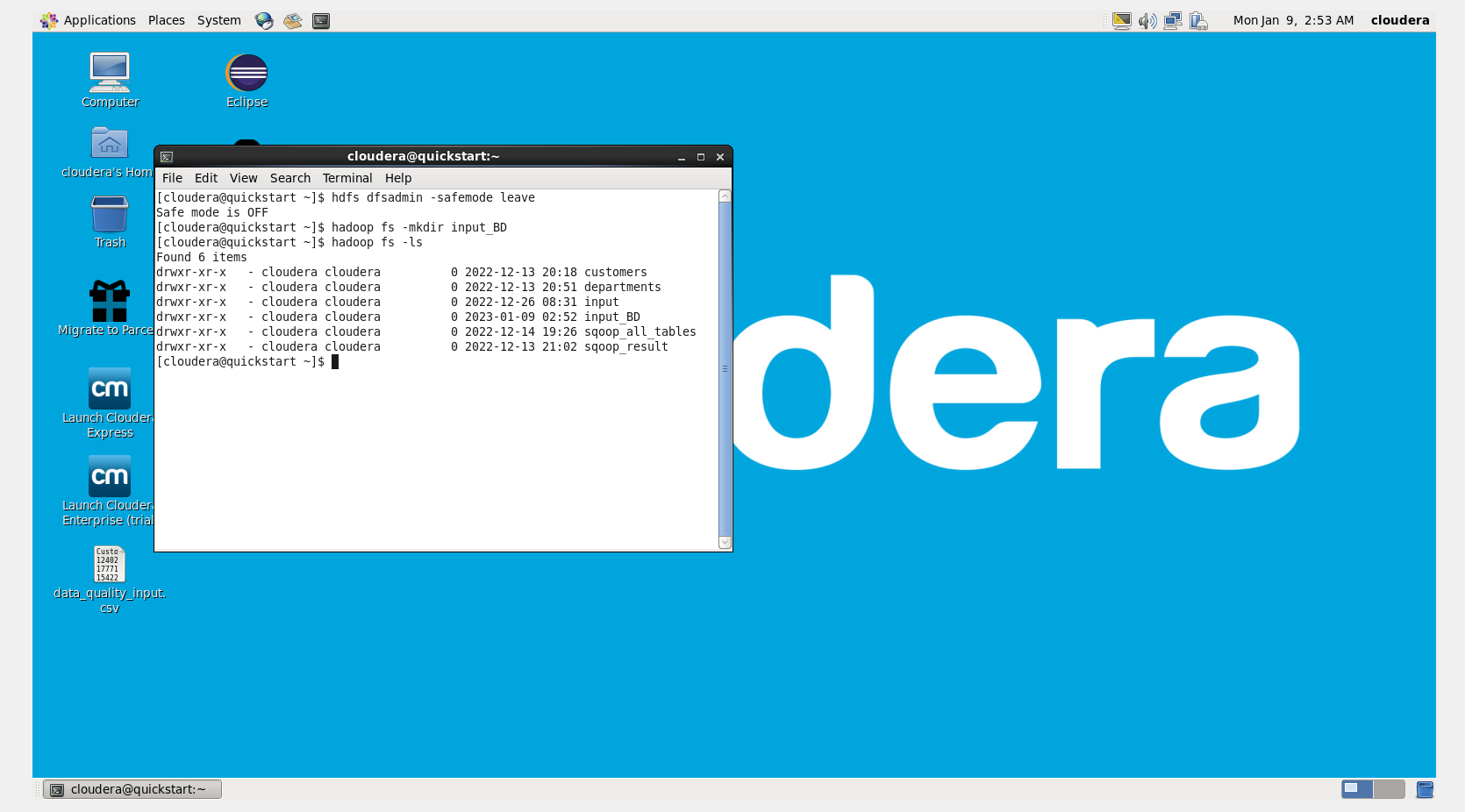
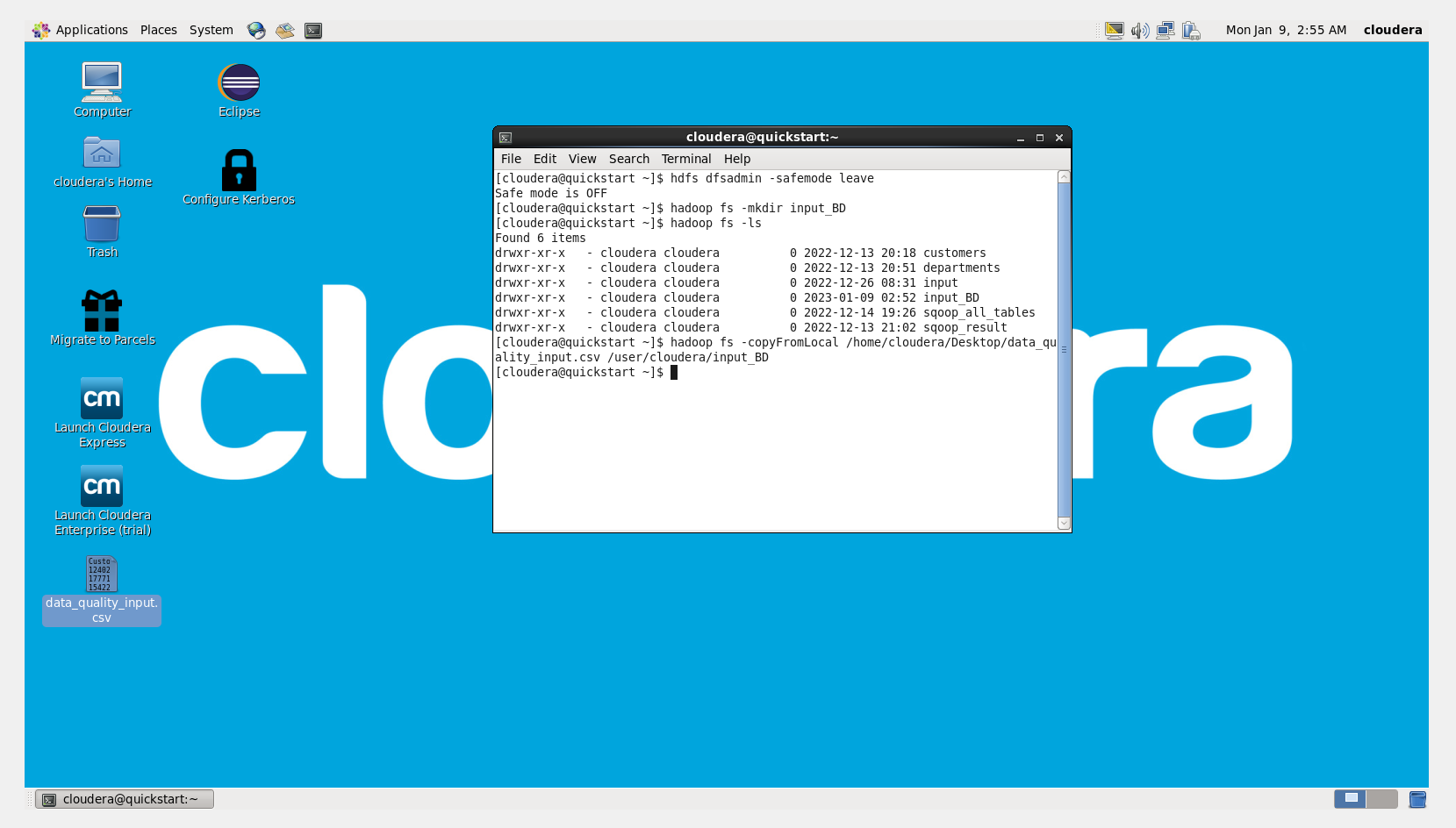
1. **Create a directory in HDFS called input**

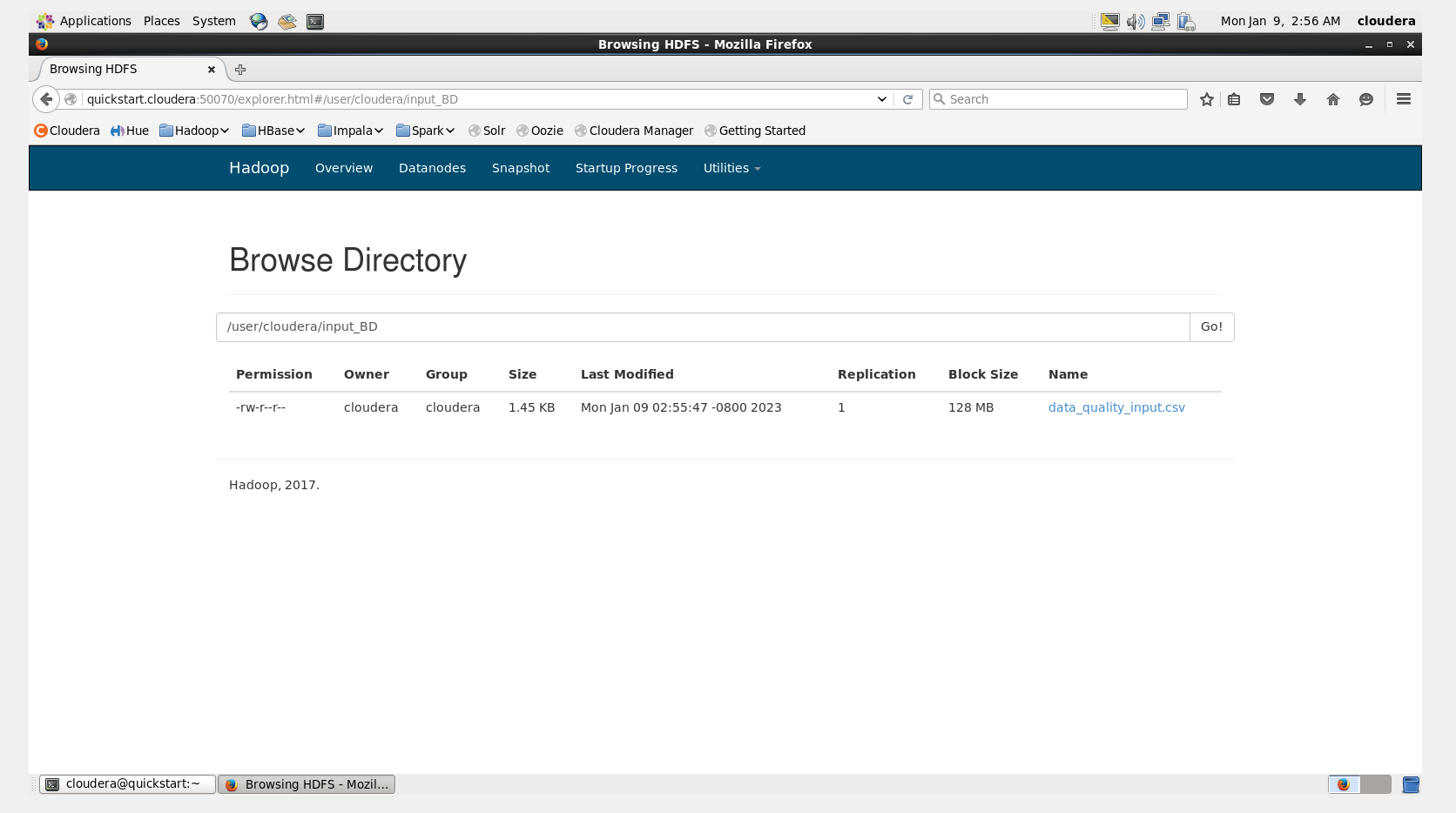




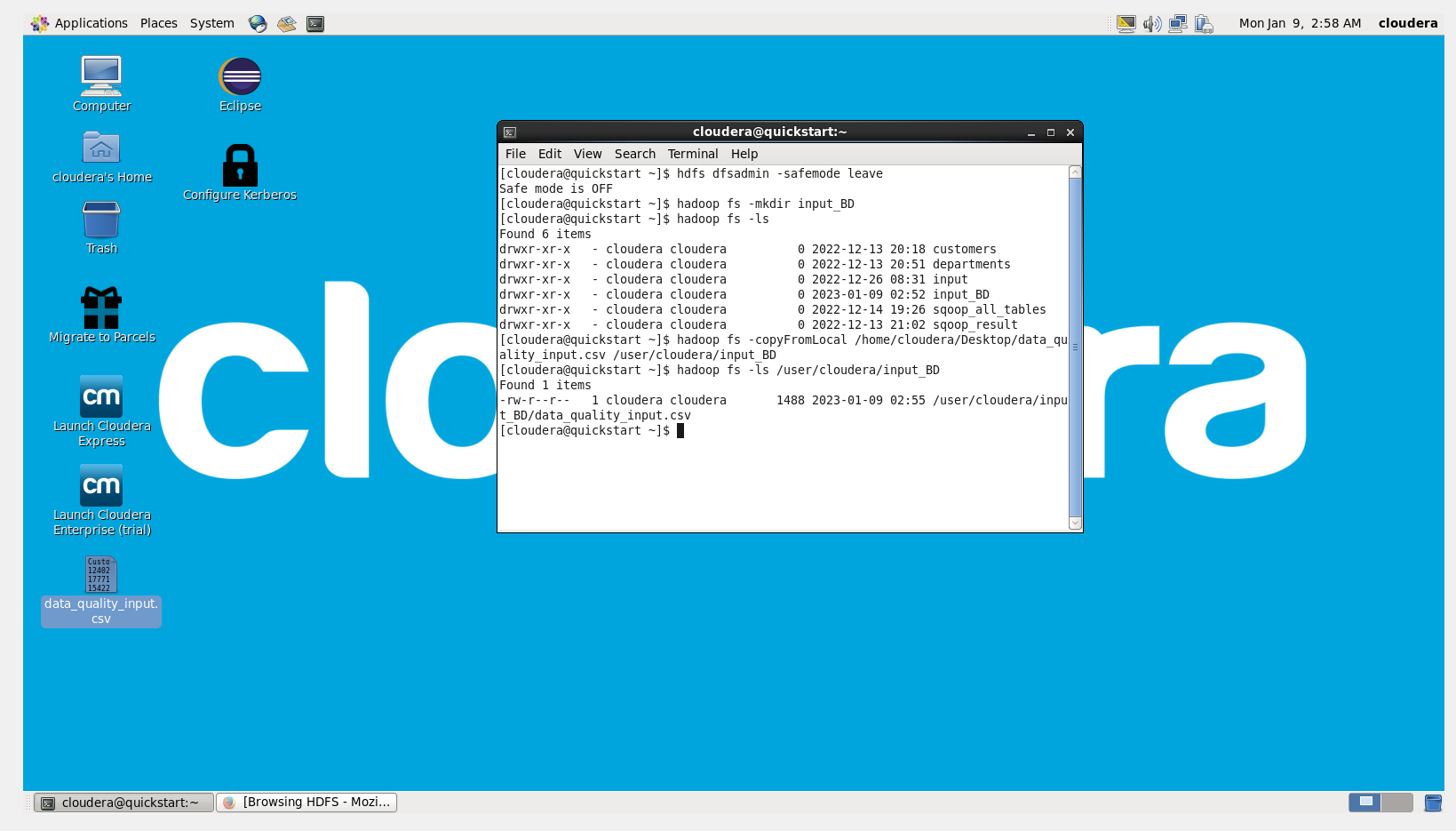
1. **Load the file data\_quality\_input.csv into this directory**

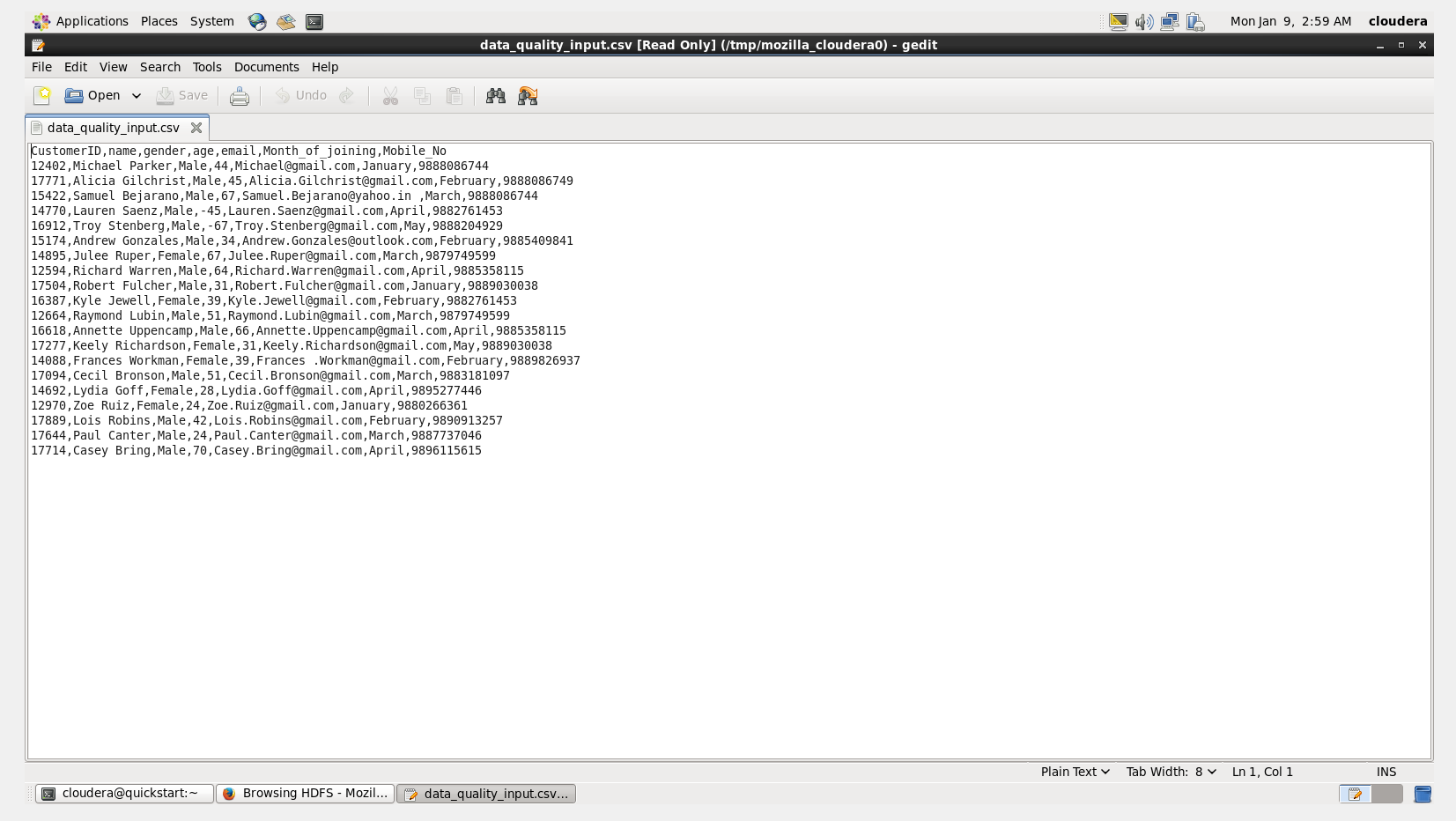
hadoop fs -copyFromLocal /home/cloudera/Desktop/data\_quality\_input.csv /user/cloudera/input\_BD



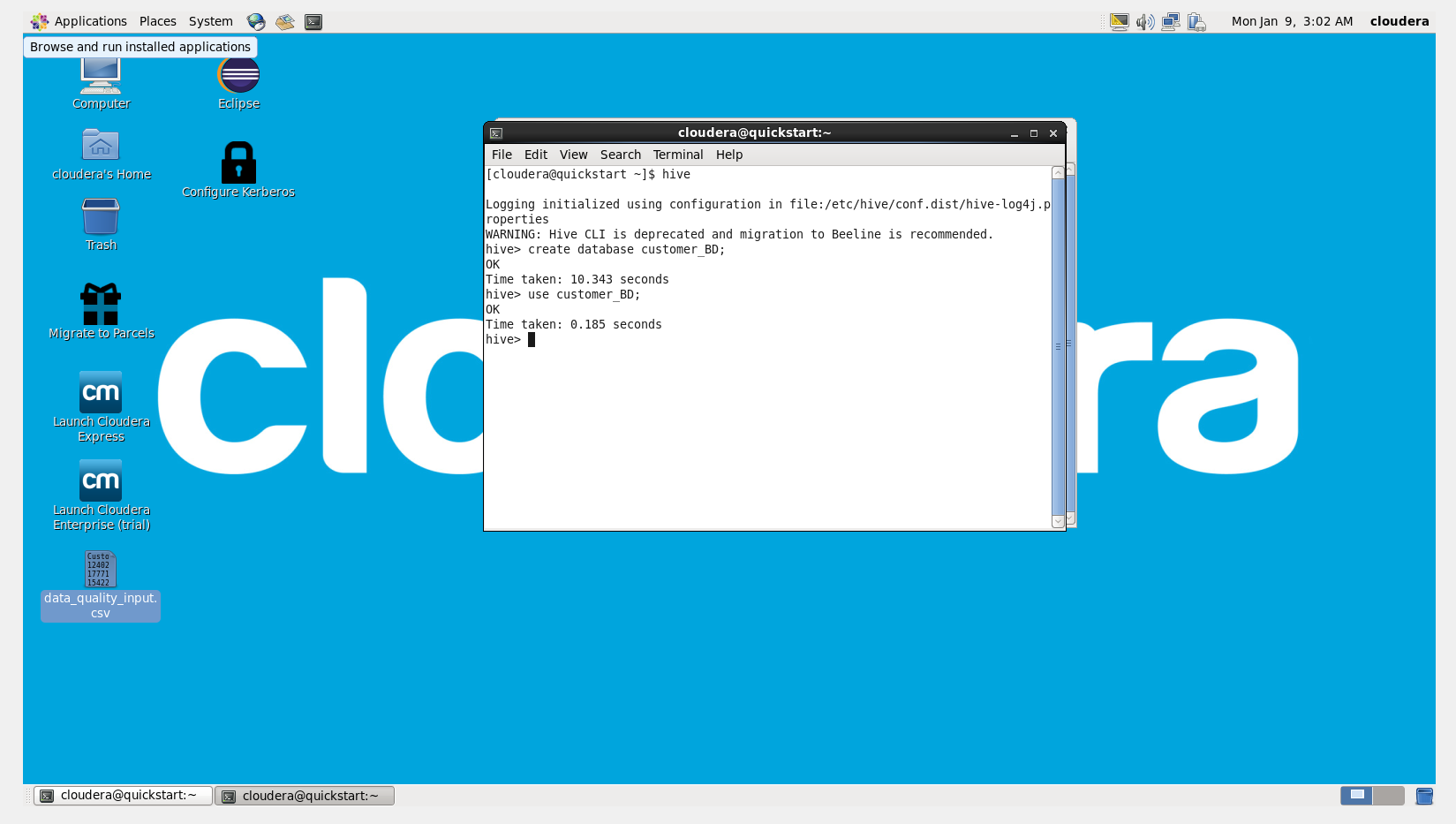


hadoop fs -ls /user/cloudera/input\_BD



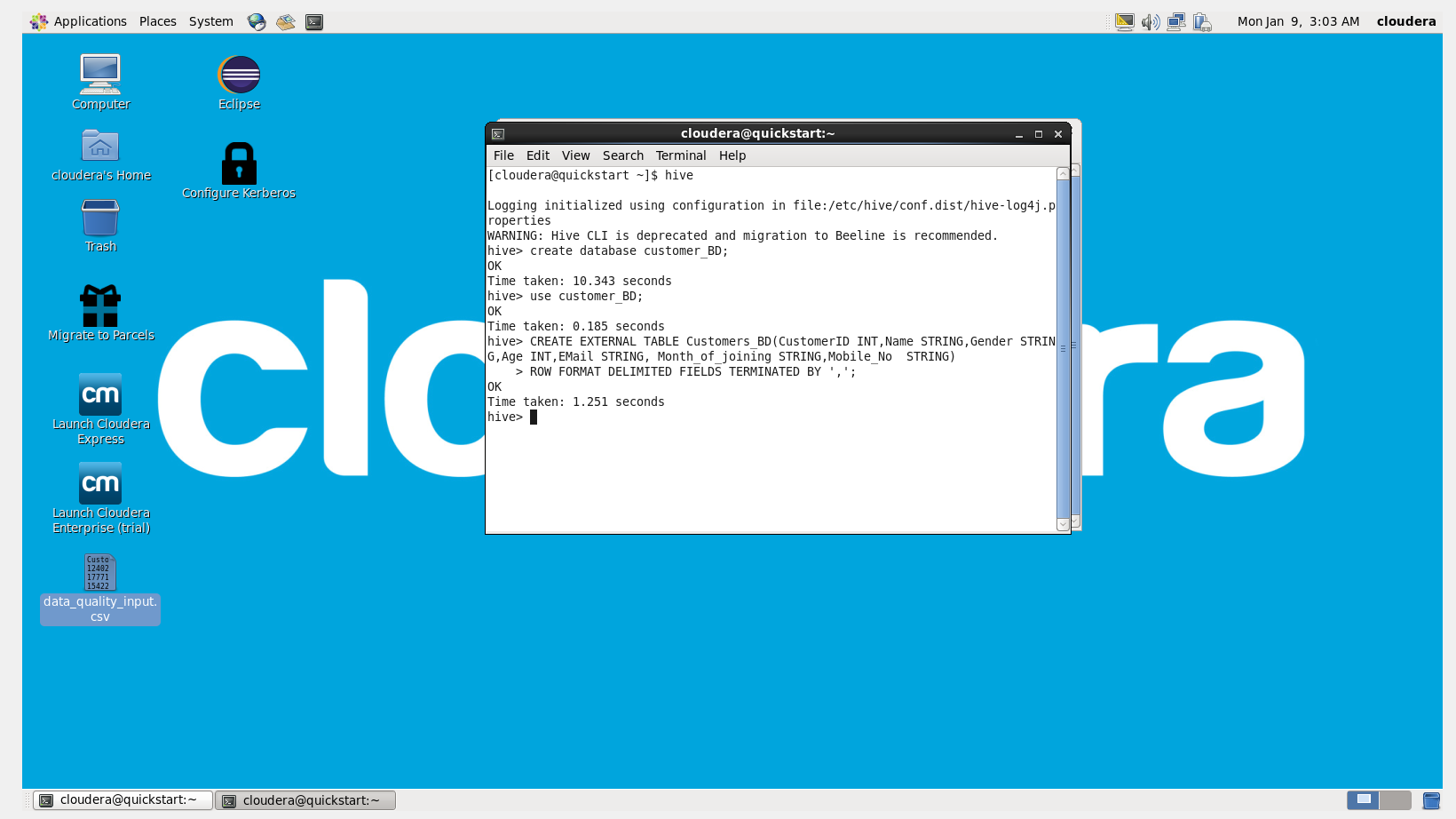


1. **Understand the given input data and create hive external table for this**

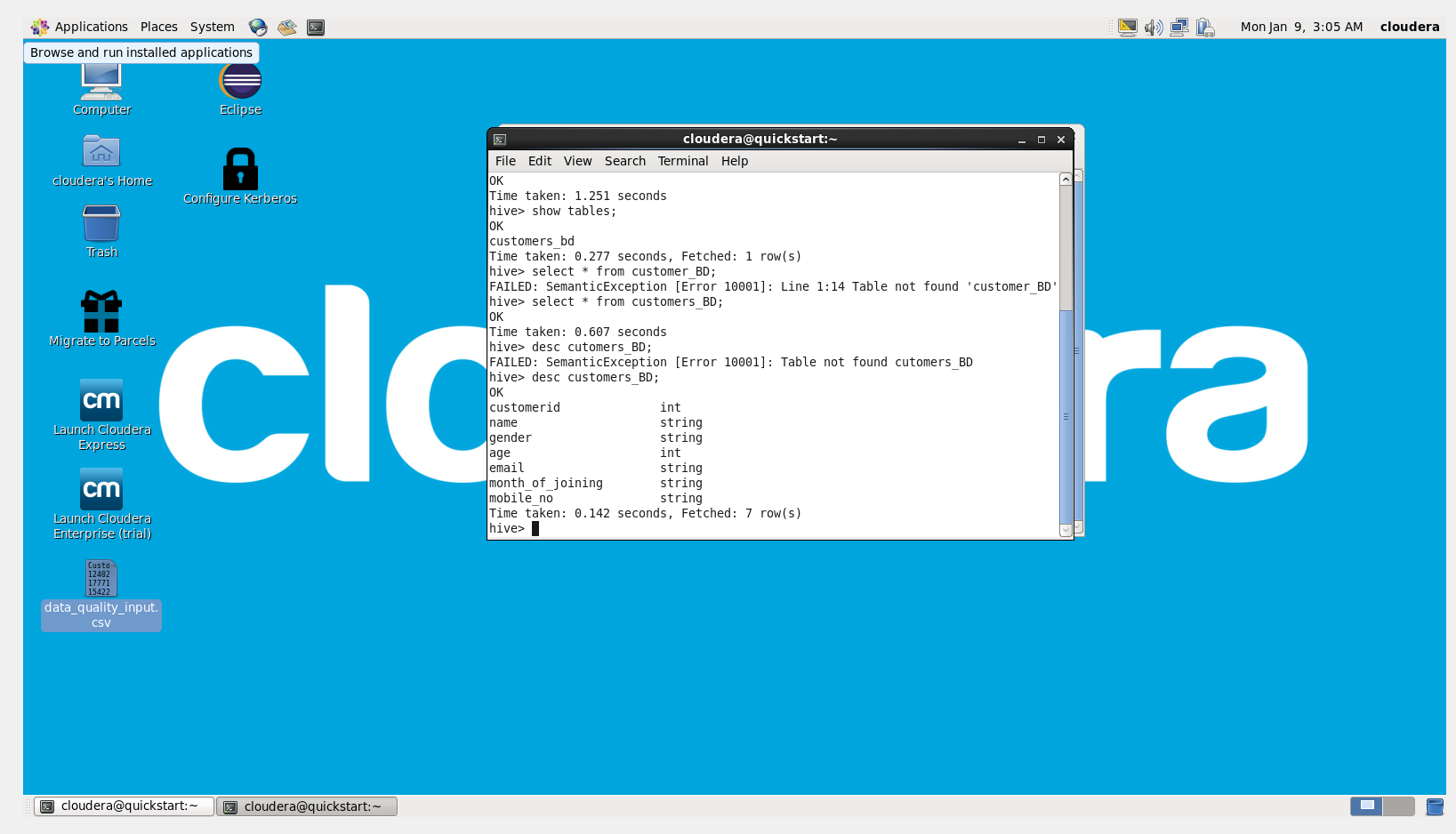


CREATE EXTERNAL TABLE Customers\_BD(CustomerID INT,Name STRING,Gender STRING,Age INT,EMail STRING, Month\_of\_joining STRING,Mobile\_No STRING)

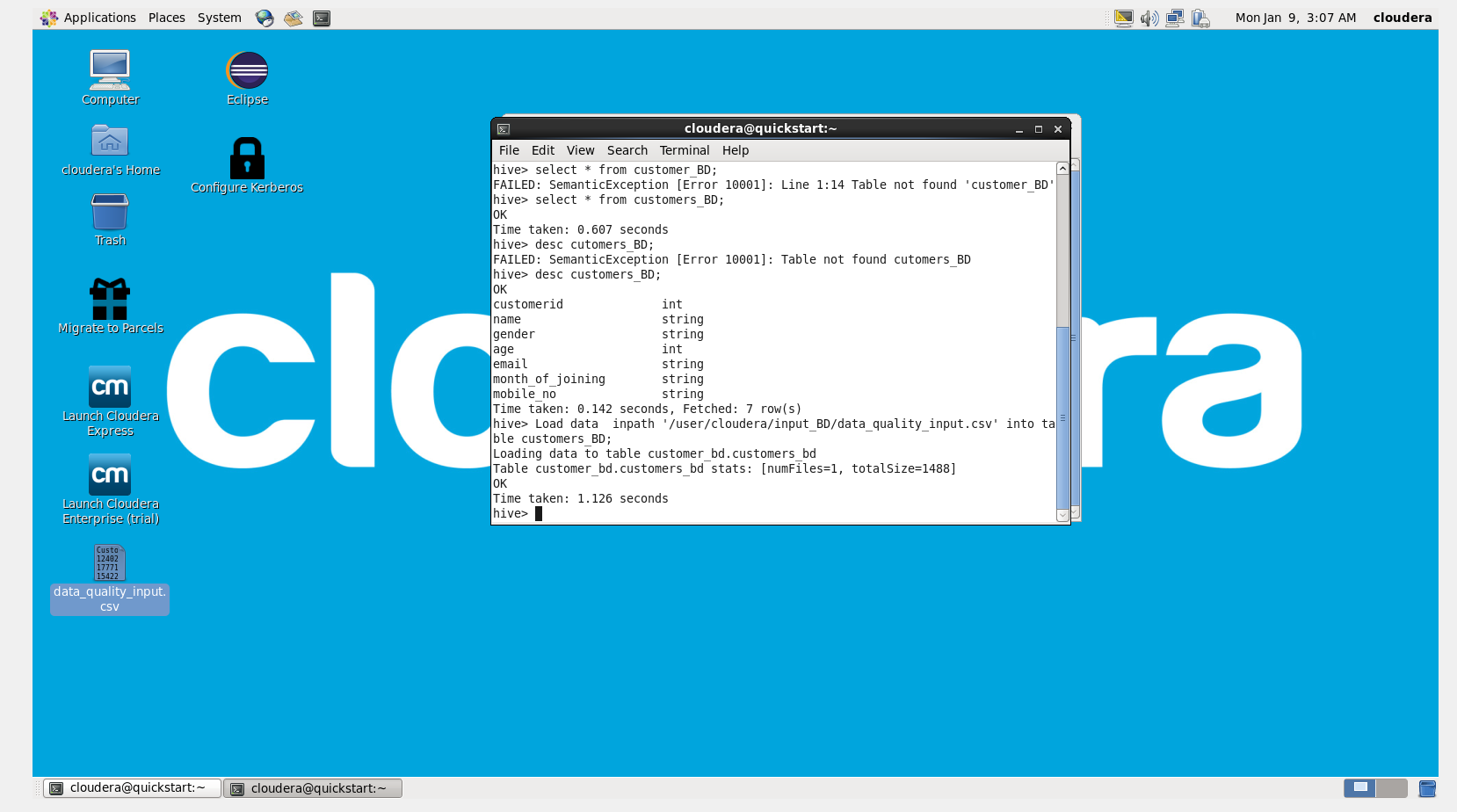
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';



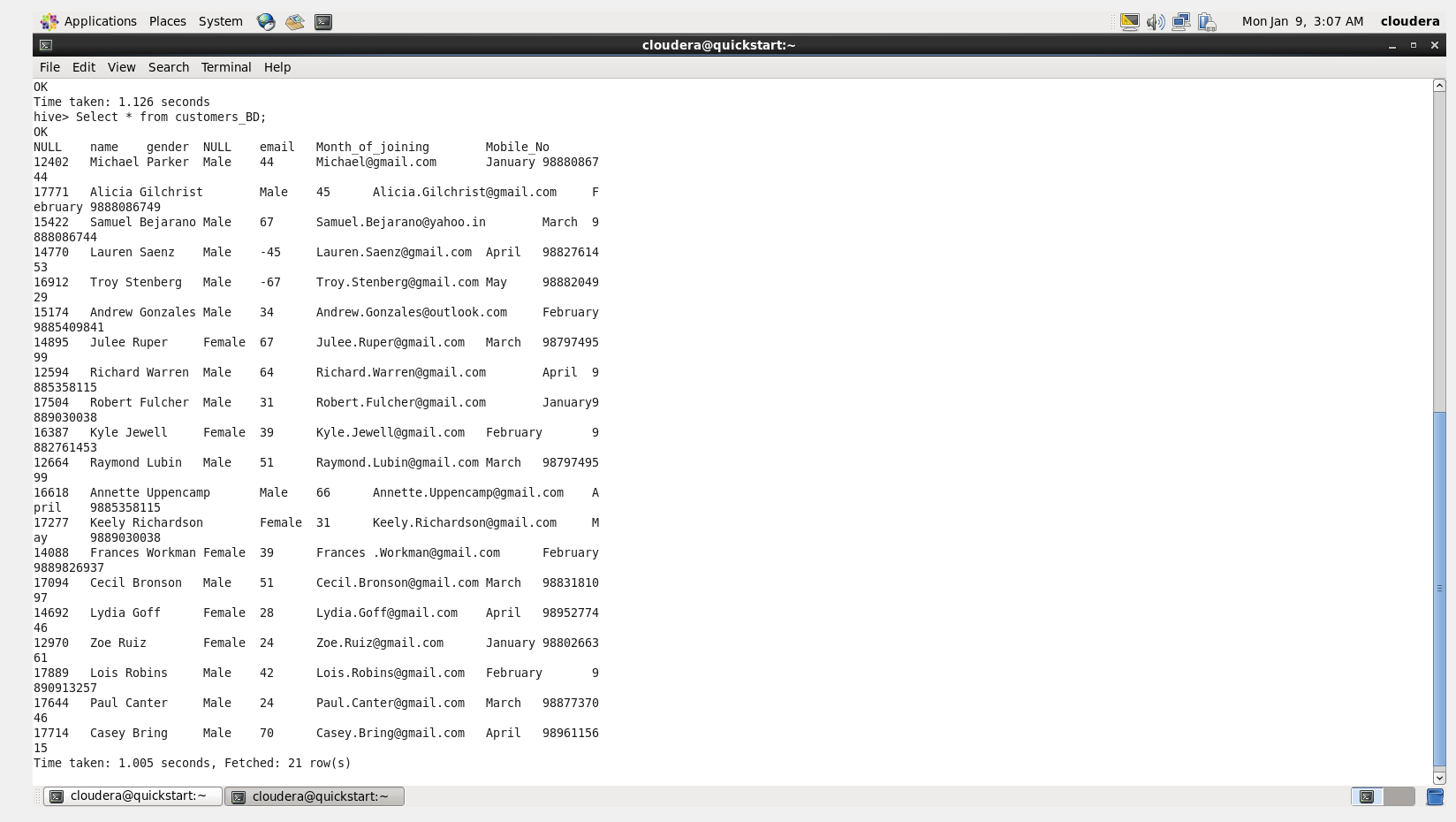
show tables;



Load data inpath '/user/cloudera/input\_BD/data\_quality\_input.csv' into table customers\_BD;

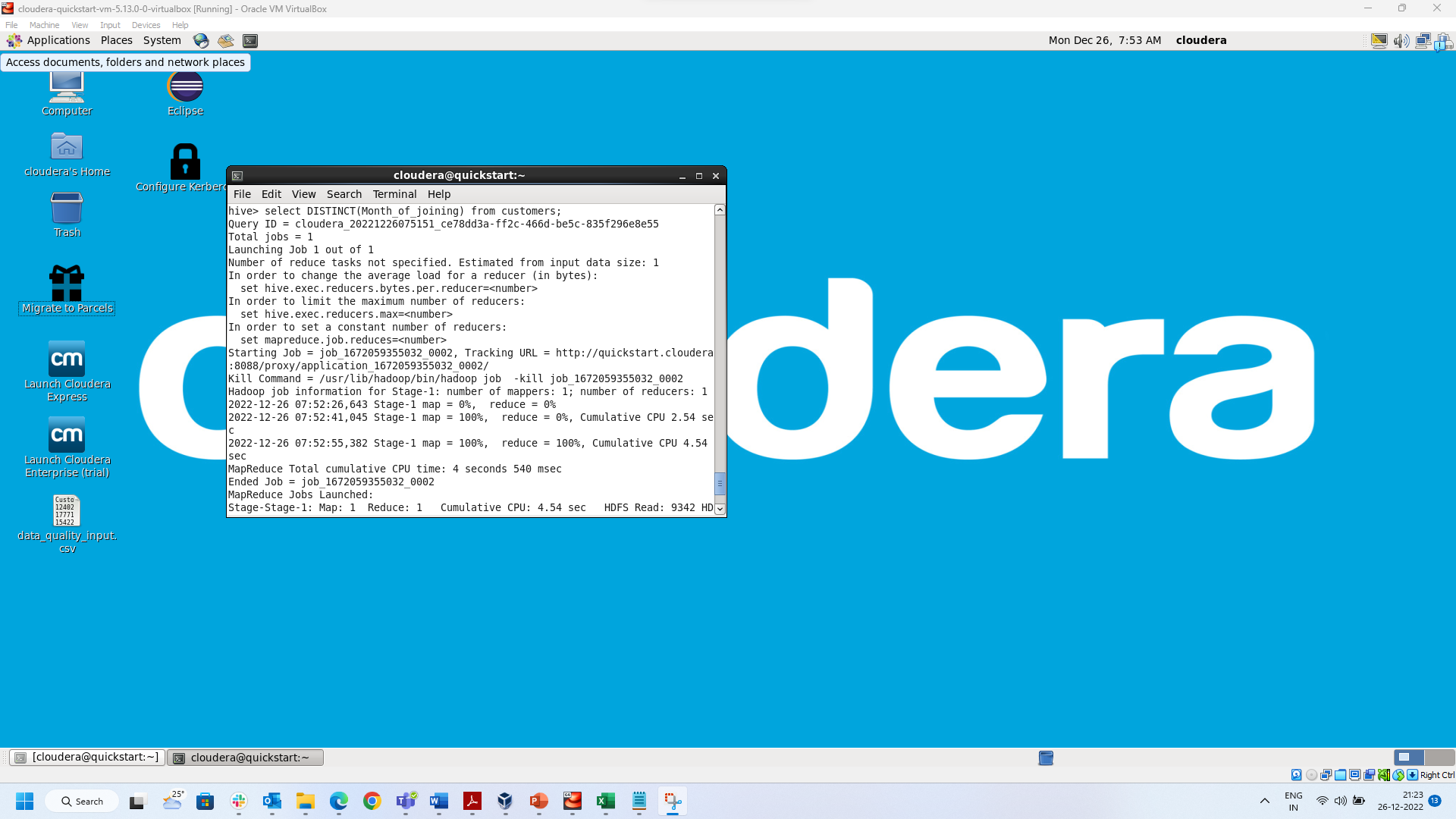


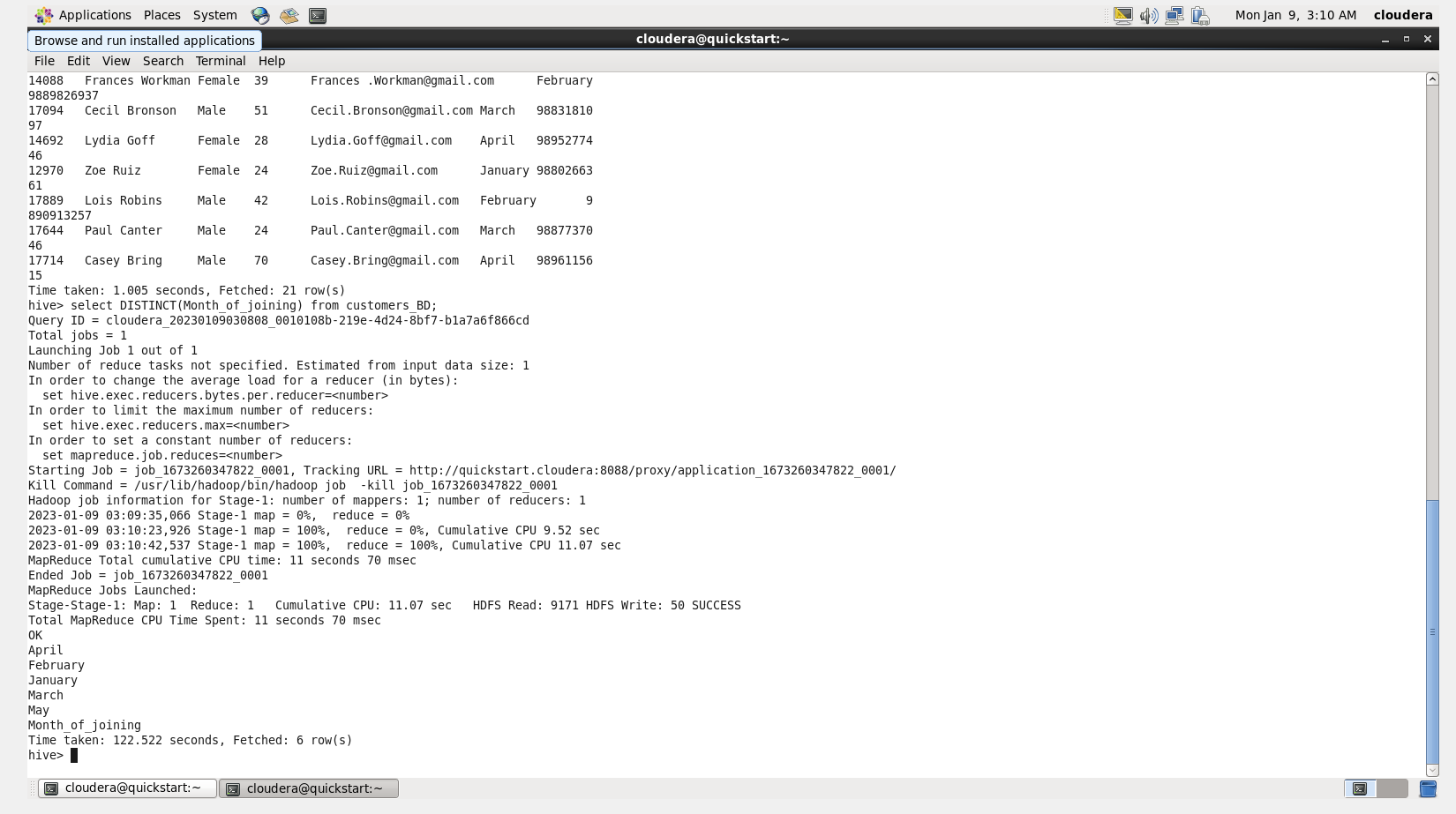
Select \* from customers\_BD;



1. **List down the distinct month names in the given input**

select DISTINCT(Month\_of\_joining) from customers\_BD;

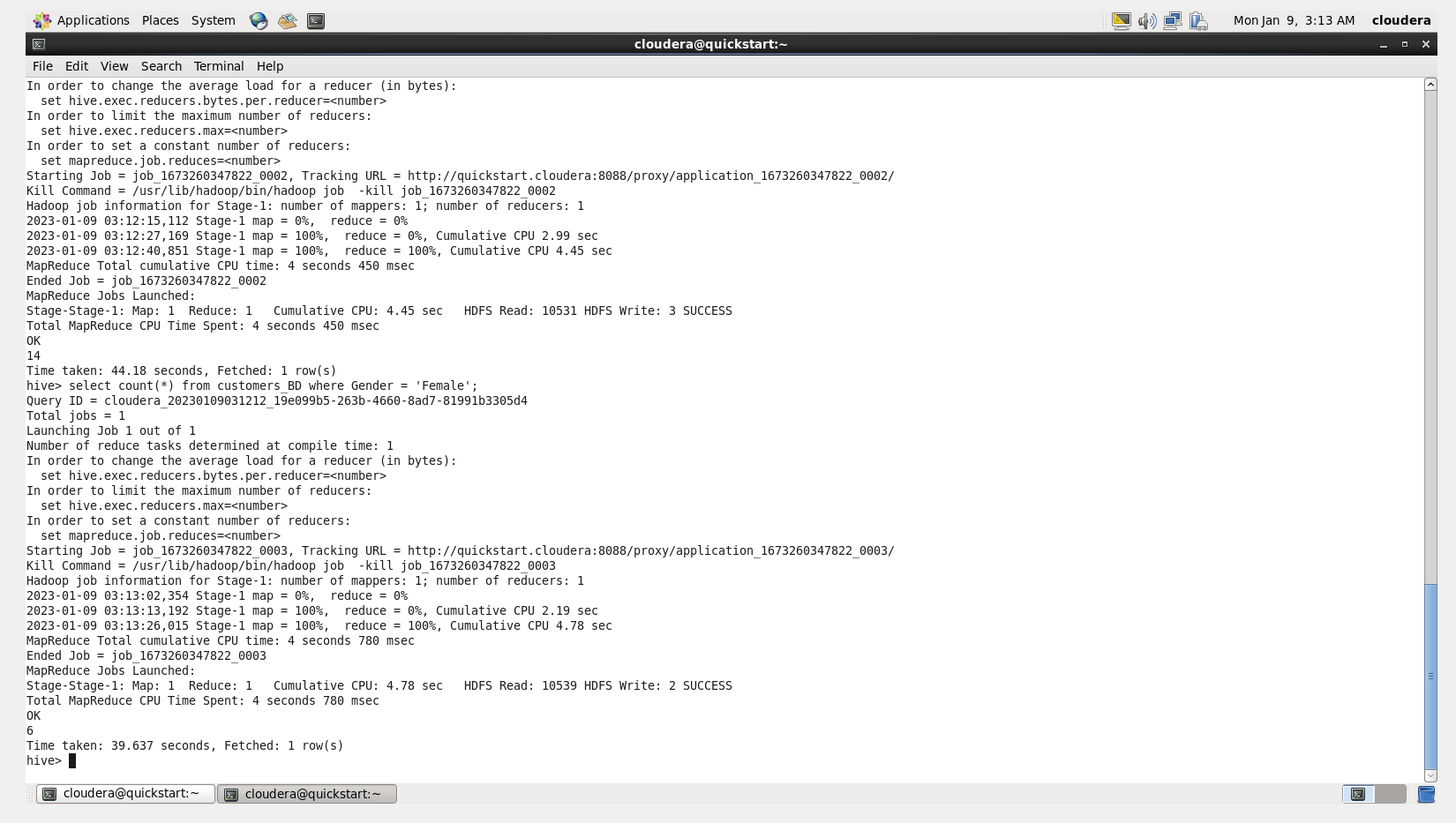




1. **Calculate the total number of Male and Female in this data**

select count(\*) from customers\_BD where Gender = 'Male';

select count(\*) from customers\_BD where Gender = 'Female';



To be started from here

1. **Write down the role of these services in these above processes**
2. **HADOOP – NameNode:**

NameNode maintains the file system tree and the metadata for all the files and directories in the tree.

1. **HADOOP – DataNodes:**

Datanodes are the work-horses of the filesystem and store the actual data.

Responsible for serving read and write requests from the file system’s clients.

Also perform block creation, deletion, and replication upon instruction from the NameNode.

Periodically send a Heartbeat and a Blockreport to NameNode.

Heartbeat indicates that the DataNode is functioning properly. A Blockreport contains a list of all blocks on a DataNode.

1. **HADOOP – Secondary NameNode:**

*Secondary namenode,* is run whose main role is to periodically merge the namespace image with the edit log to prevent the edit log from becoming too large.

1. **HADOOP – Resource Manager:**

ResourceManager (RM) is the master that arbitrates all the available cluster resources and thus helps manage the distributed applications running on the Hadoop system.

Resource Manager is responsible for managing global resources among all nodes. It is the ultimate authority to allocate resources among all applications.

1. **HADOOP – Node Manager:**

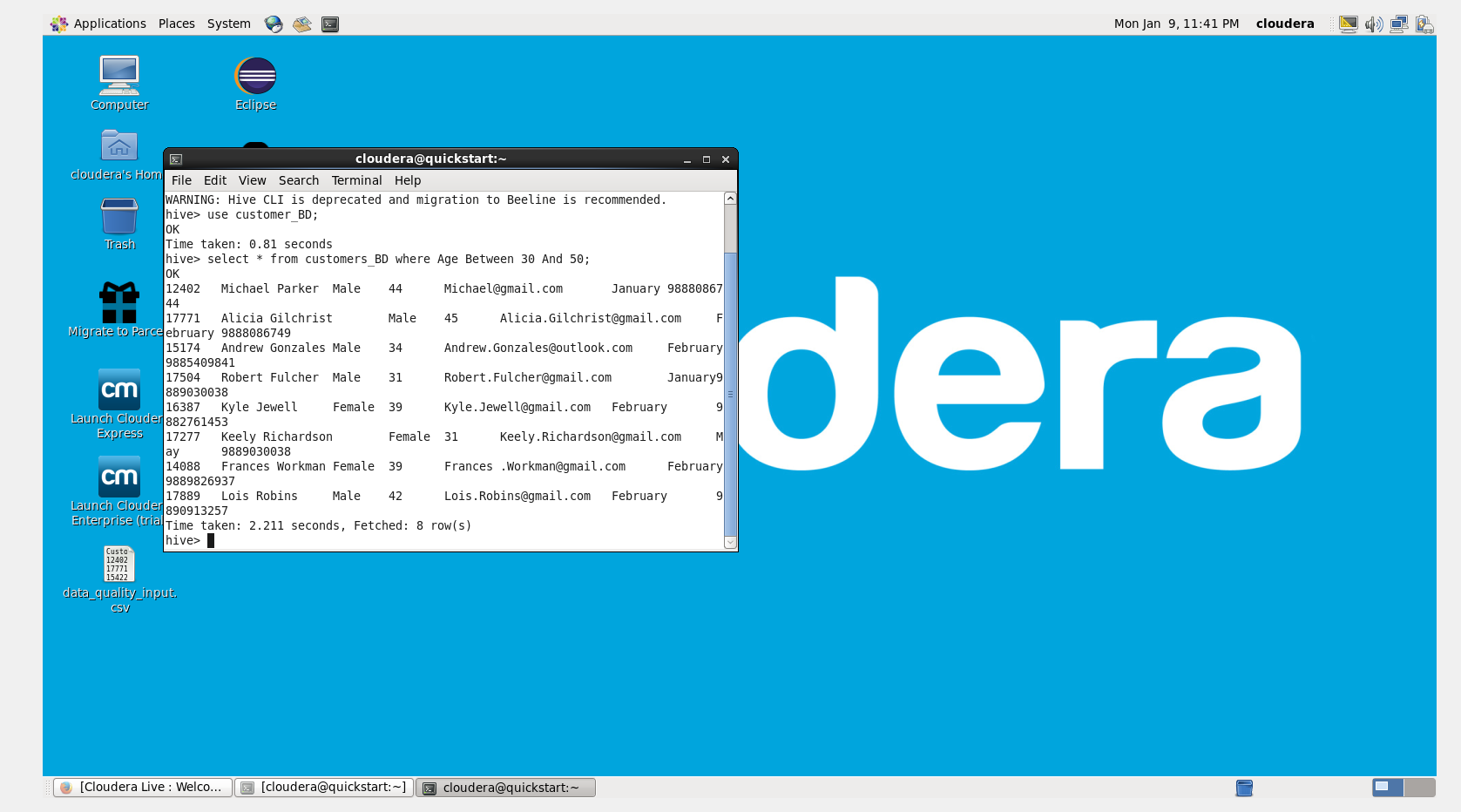
The Hadoop Yarn Node Manager is the per-machine/per-node framework agent who is responsible for containers, monitoring their resource usage and reporting the same to the ResourceManager.

The NodeManager runs services to determine the health of the node it is executing on. The services perform checks on the disk as well as any user specified tests.

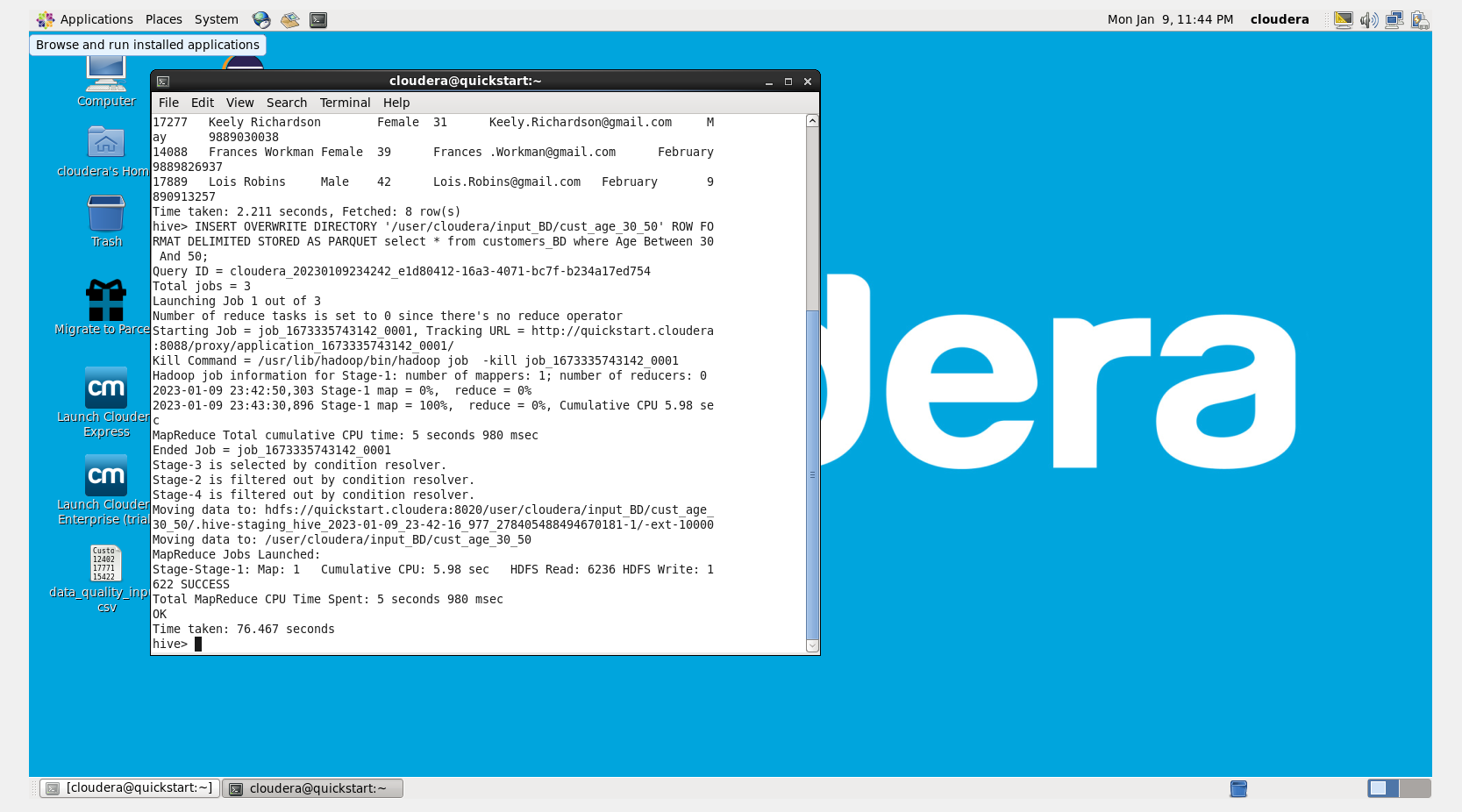
If any health check fails, the NodeManager marks the node as unhealthy and communicates this to the ResourceManager, which then stops assigning containers to the node

1. **Take the records only the age between 30 and 50 and store it in HDFS**

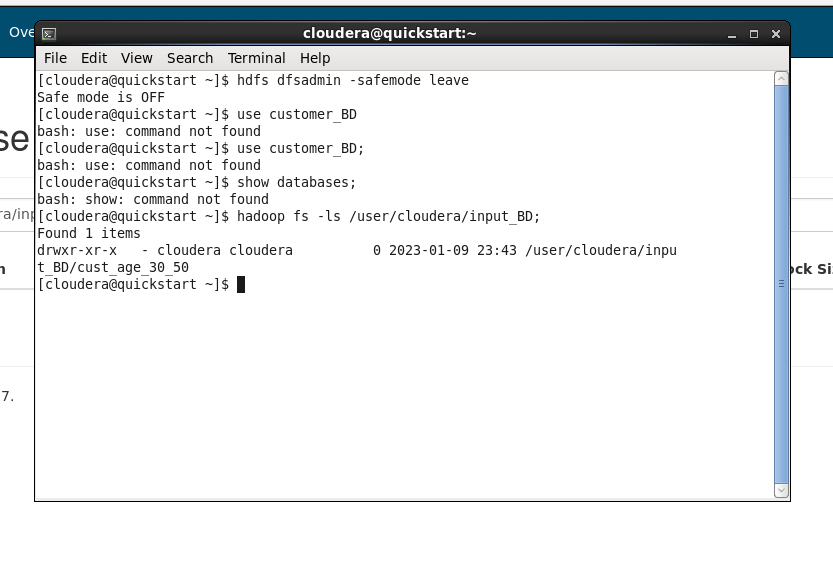
select \* from customers\_BD where Age Between 30 And 50;



INSERT OVERWRITE DIRECTORY '/user/cloudera/input\_BD/cust\_age\_30\_50' ROW FORMAT DELIMITED STORED AS PARQUET select \* from customers\_BD where Age Between 30 And 50;



hadoop fs -ls /user/cloudera/input\_BD

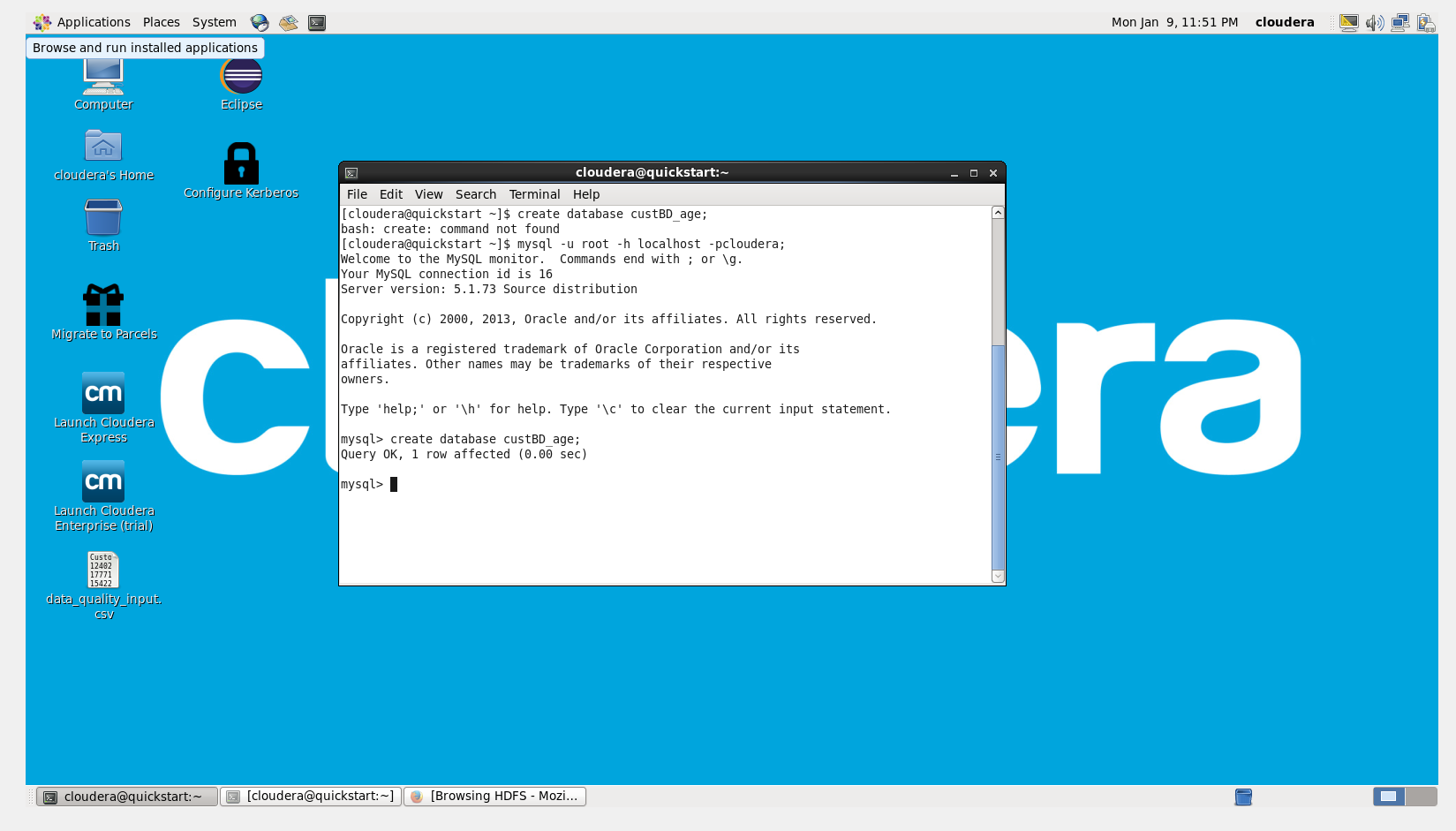


Graphical user interface, text, application, email

Description automatically generated

1. **Create the Database and table for the above results**

create database custBD\_age;



use custBD\_age;

Graphical user interface, text

Description automatically generated

CREATE TABLE customersBD\_age\_30\_50(customer\_id Int,Name varchar(50),Gender varchar(20),Age Int,EMail varchar(50),Month\_of\_joining varchar(10),Mobile\_No varchar(10));

Graphical user interface, text

Description automatically generated

Show tables;

Graphical user interface, text, application

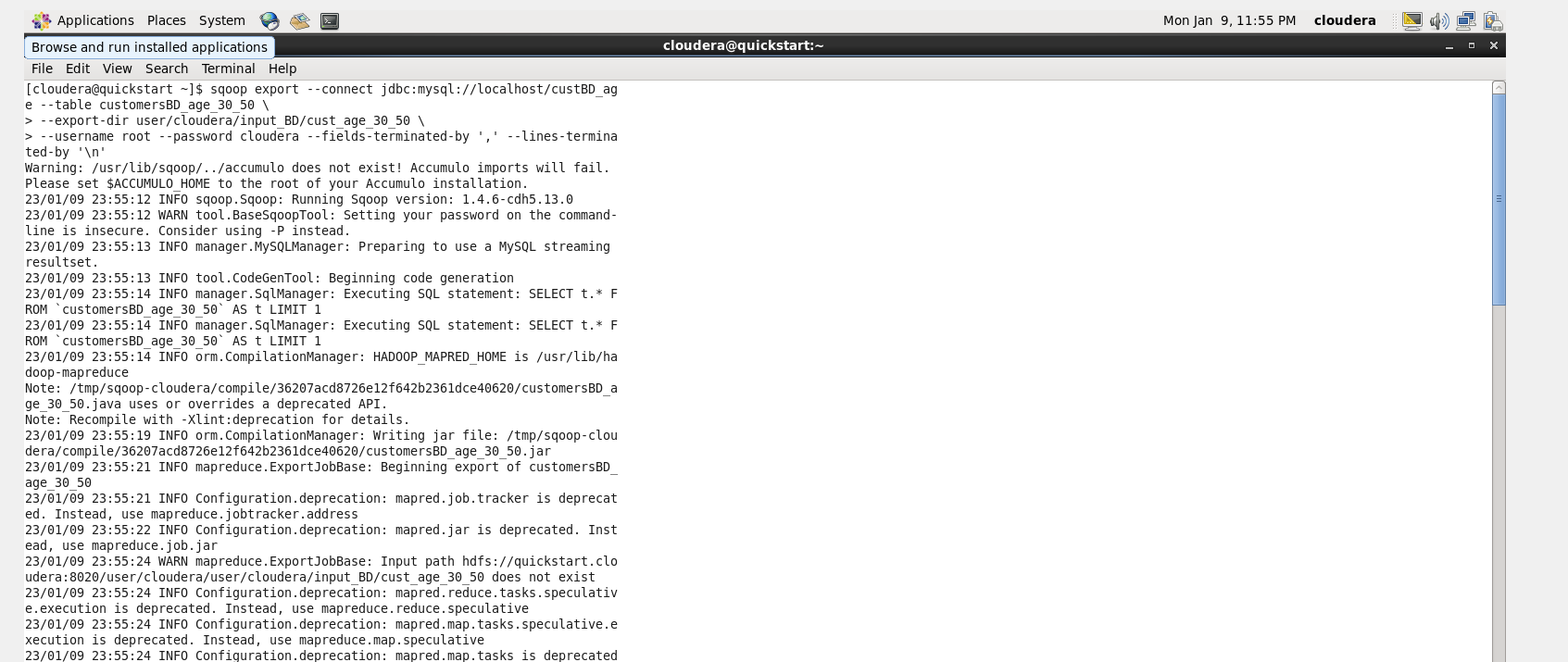
Description automatically generated

1. **Do the data export to transfer the data from HDFS to MySQL**

sqoop export --connect jdbc:mysql://localhost/custBD\_age --table customersBD\_age\_30\_50 \

--export-dir user/cloudera/input\_BD/cust\_age\_30\_50 \

--username root --password cloudera --fields-terminated-by ',' --lines-terminated-by '\n'



**10. Verify all the results are stored in MySQL table**

select \* from customersBD\_age\_30\_50;

