Assignment-3

Student ID	Name	Email
8800149	Basavraj Jaliminche	Bjaliminche0149@econestogac.on.ca
8800060	Zarana Gohil	Zgohil0060@conestogac.on.ca
8807575	Dharti Patel	Dpatel757@conestogac.on.ca

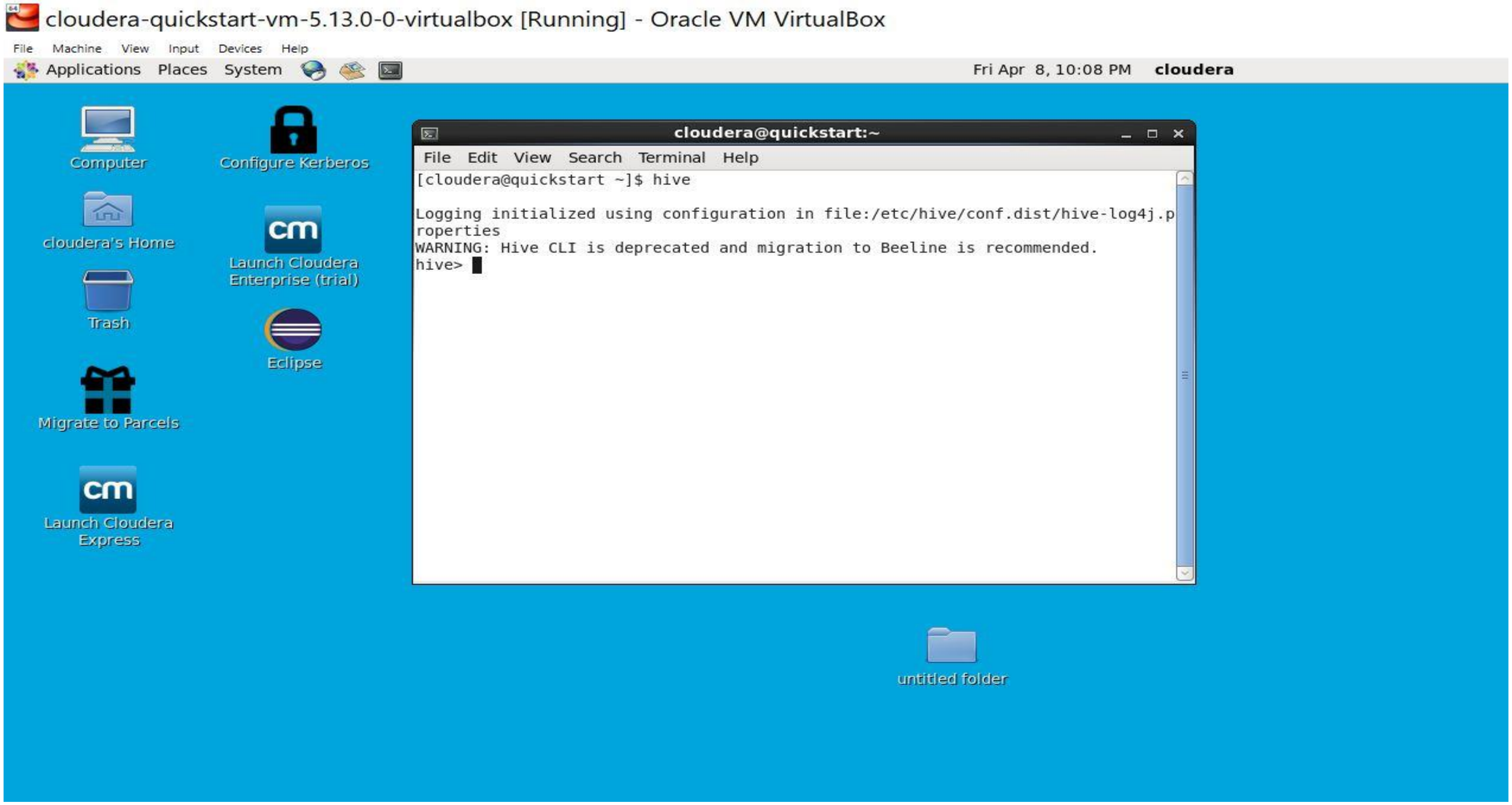
Assignment-3

Laboratory Exercise-05

1.Install Software of cloudera

2. Run Hive from the command line.

---open command line and type as $ hive

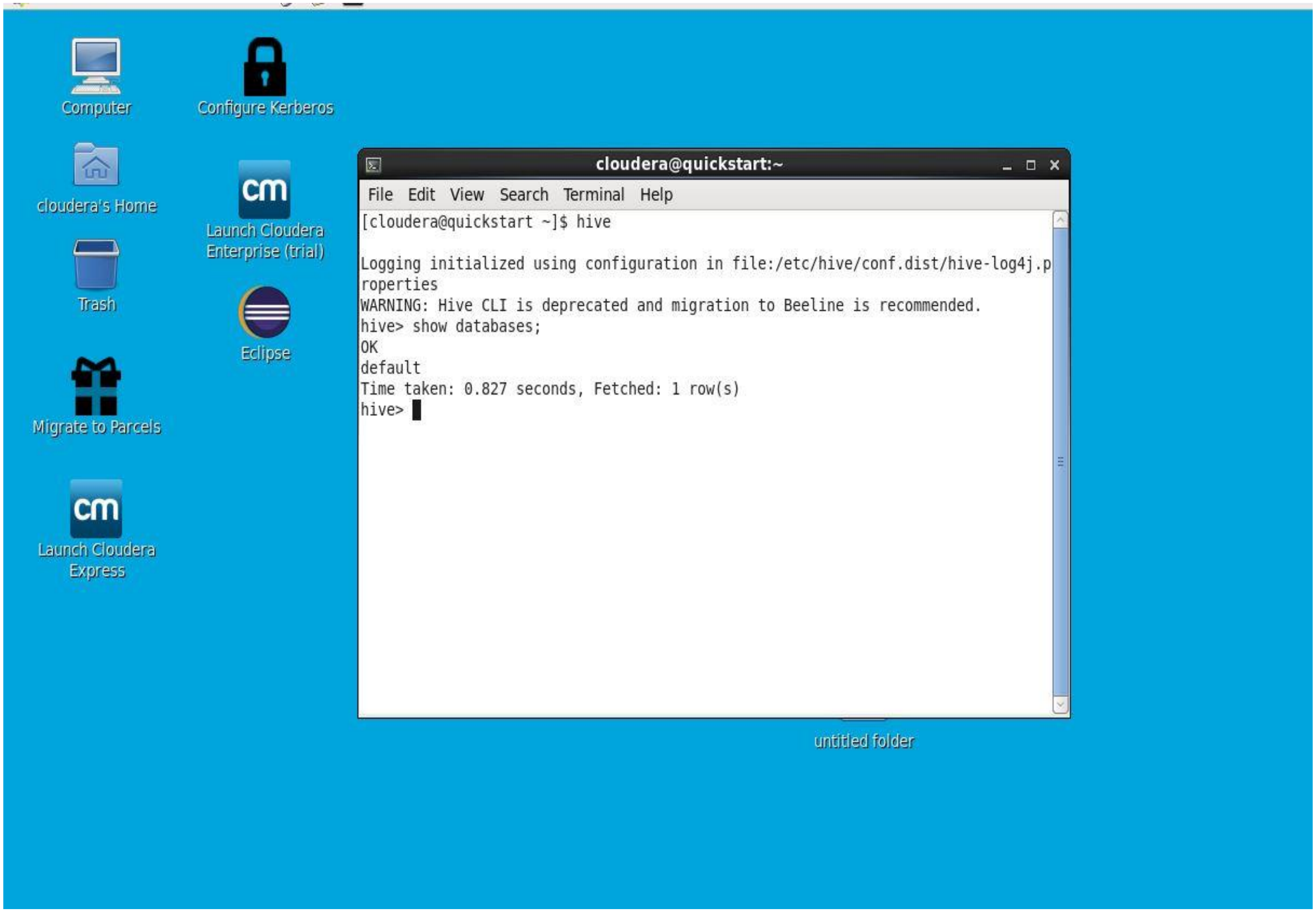


cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Fri Apr 8, 10:08 PM cloudera
Computer
Configure Kerberos
cloudera's Home
Launch Cloudera Enterprise (trial)
Trash
Eclipse
Migrate to Parcels
Launch Cloudera Express
untitled folder
cloudera@quickstart:~
File Edit View Search Terminal Help

```
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.p
roperties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> 
```

3. Display all databases:

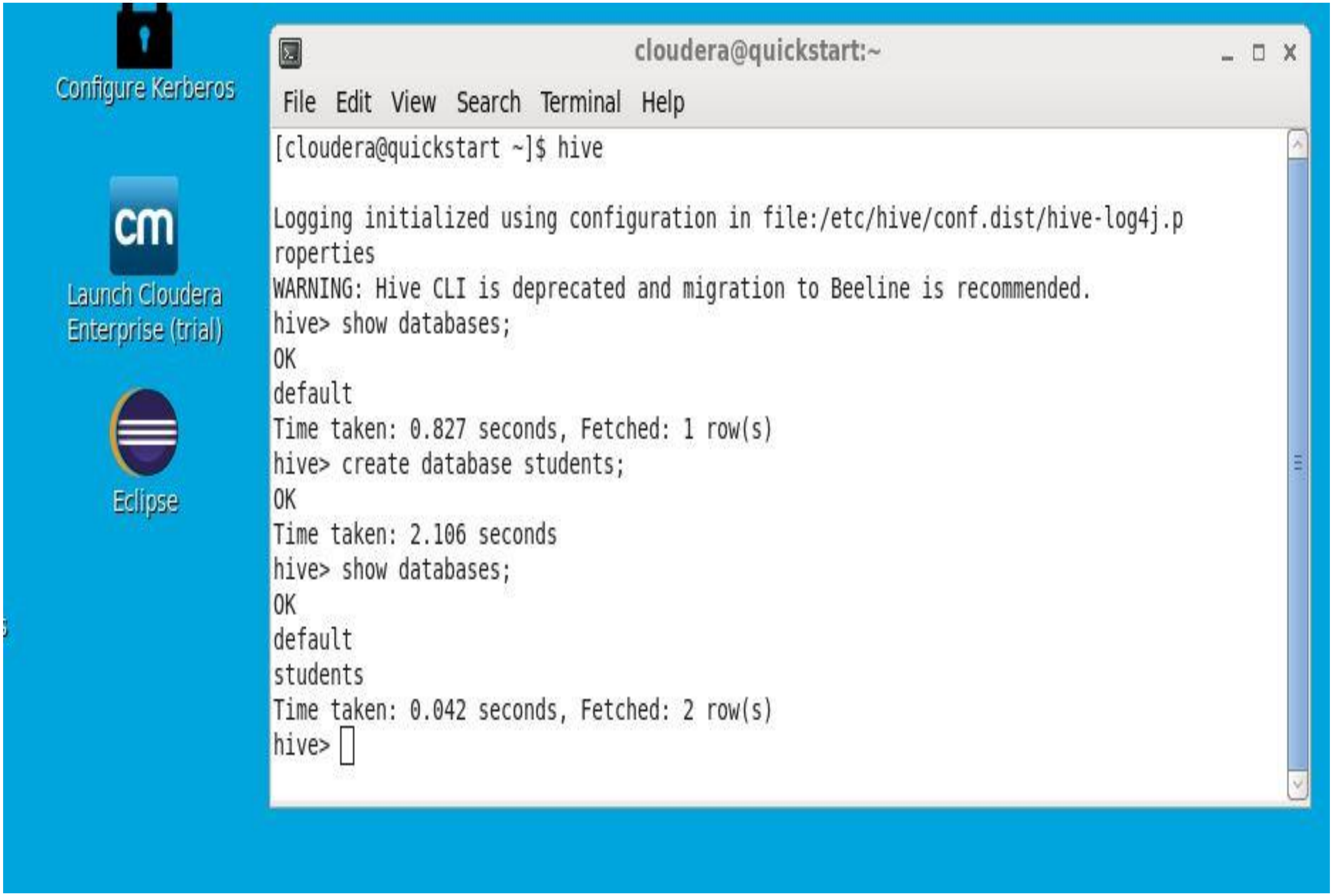
-- On command line type as ‘show databases’;



```
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.p
roperties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> show databases;
OK
default
Time taken: 0.827 seconds, Fetched: 1 row(s)
hive> 
```

4. Create a DB named students

- Type on command line as ‘create database students’
- then you will get message as ‘ok’.
- For showing databases use command as ‘show database’

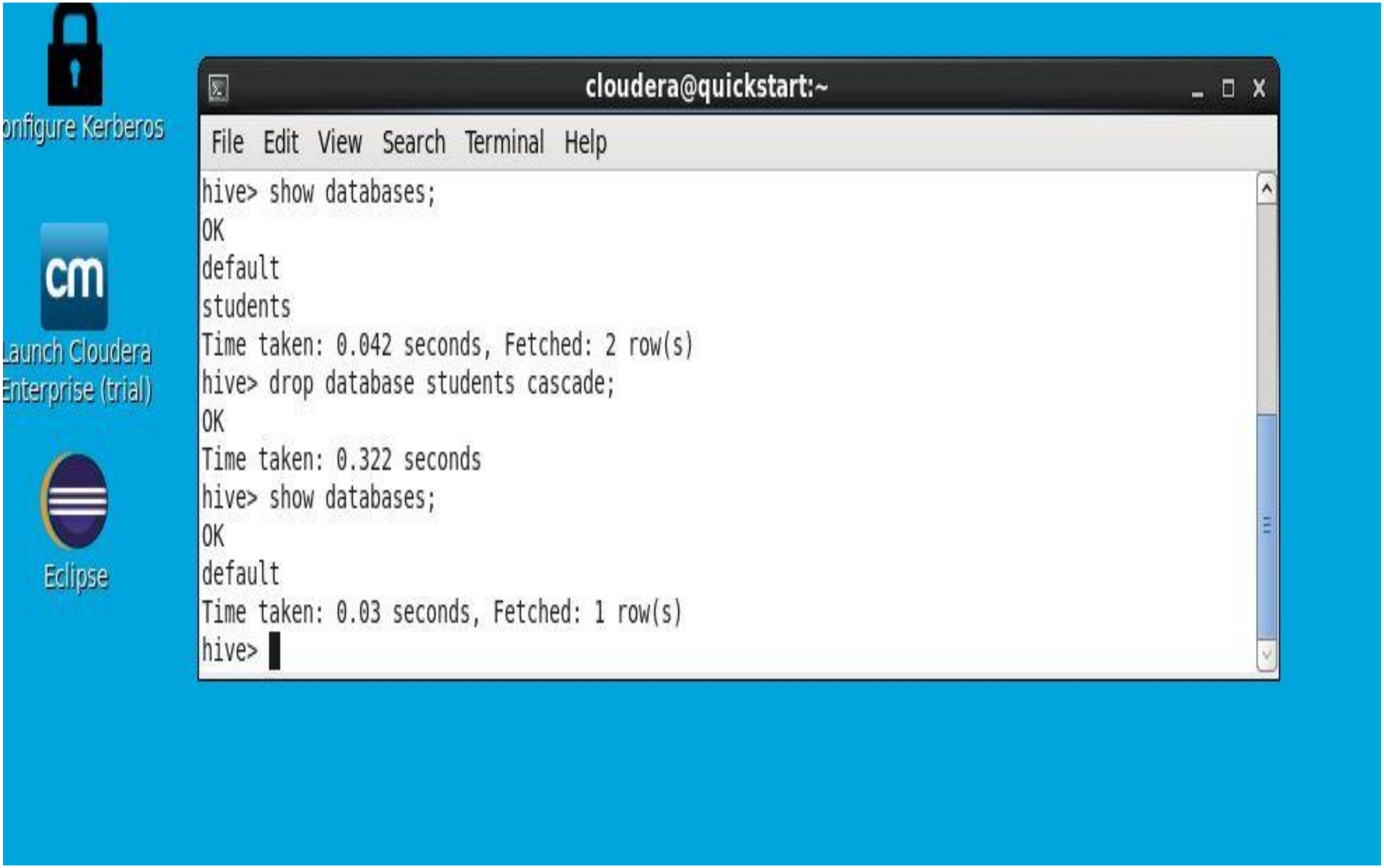


Configure Kerberos
Launch Cloudera Enterprise (trial)
Eclipse
cloudera@quickstart:~
File Edit View Search Terminal Help

```
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.p
roperties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> show databases;
OK
default
Time taken: 0.827 seconds, Fetched: 1 row(s)
hive> create database students;
OK
Time taken: 2.106 seconds
hive> show databases;
OK
default
students
Time taken: 0.042 seconds, Fetched: 2 row(s)
hive> 
```

5. Dropping the DB students

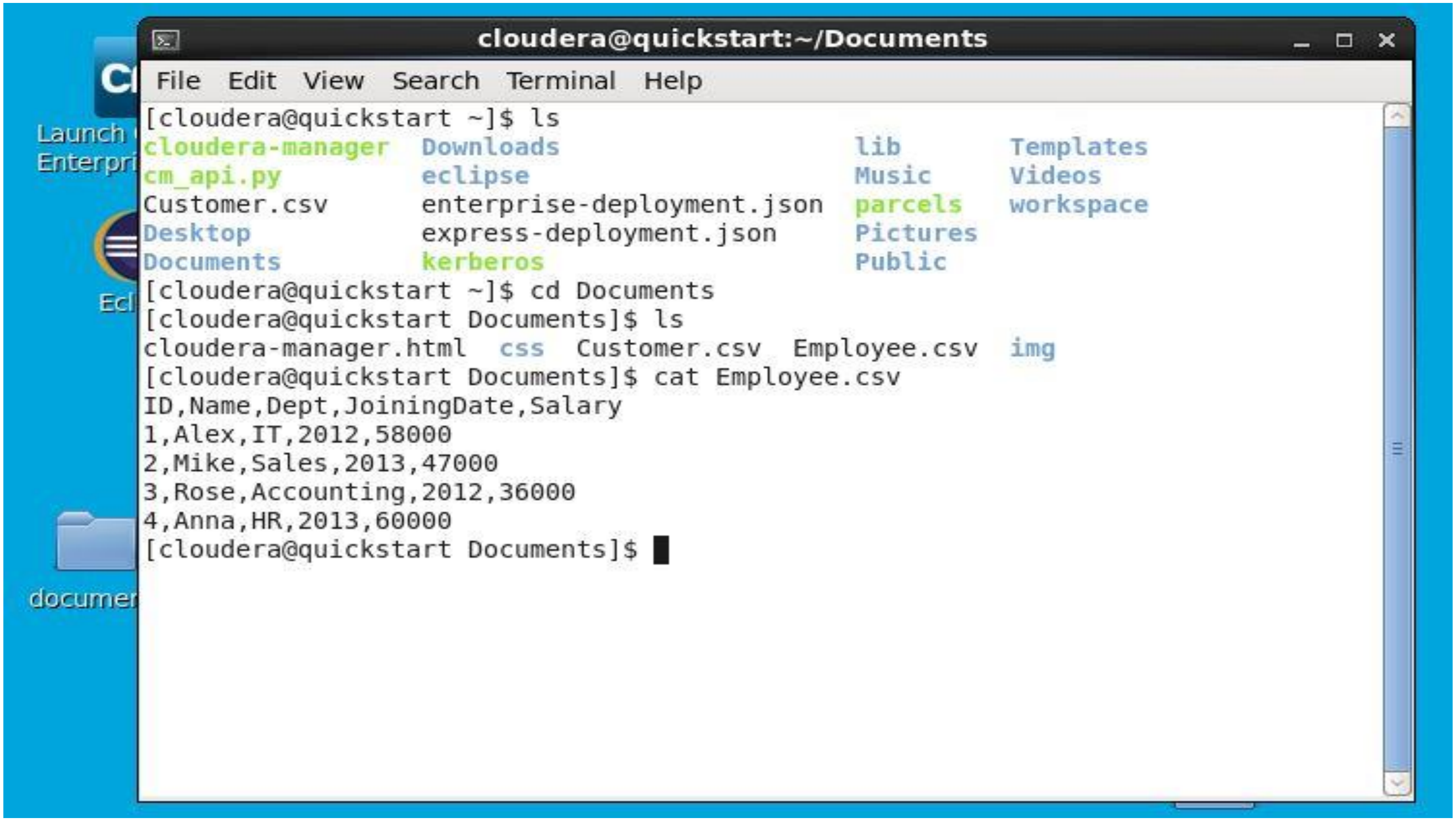
- Type command as ‘drop database students’. Then student’s database is dropped.
- To obtain the result we will type command as ‘show databases’.


onfigure Kerberos
Launch Cloudera Enterprise (trial)
Eclipse
cloudera@quickstart:~
File Edit View Search Terminal Help

```
hive> show databases;
OK
default
students
Time taken: 0.042 seconds, Fetched: 2 row(s)
hive> drop database students cascade;
OK
Time taken: 0.322 seconds
hive> show databases;
OK
default
Time taken: 0.03 seconds, Fetched: 1 row(s)
hive> 
```

6. Display the Employee.csv

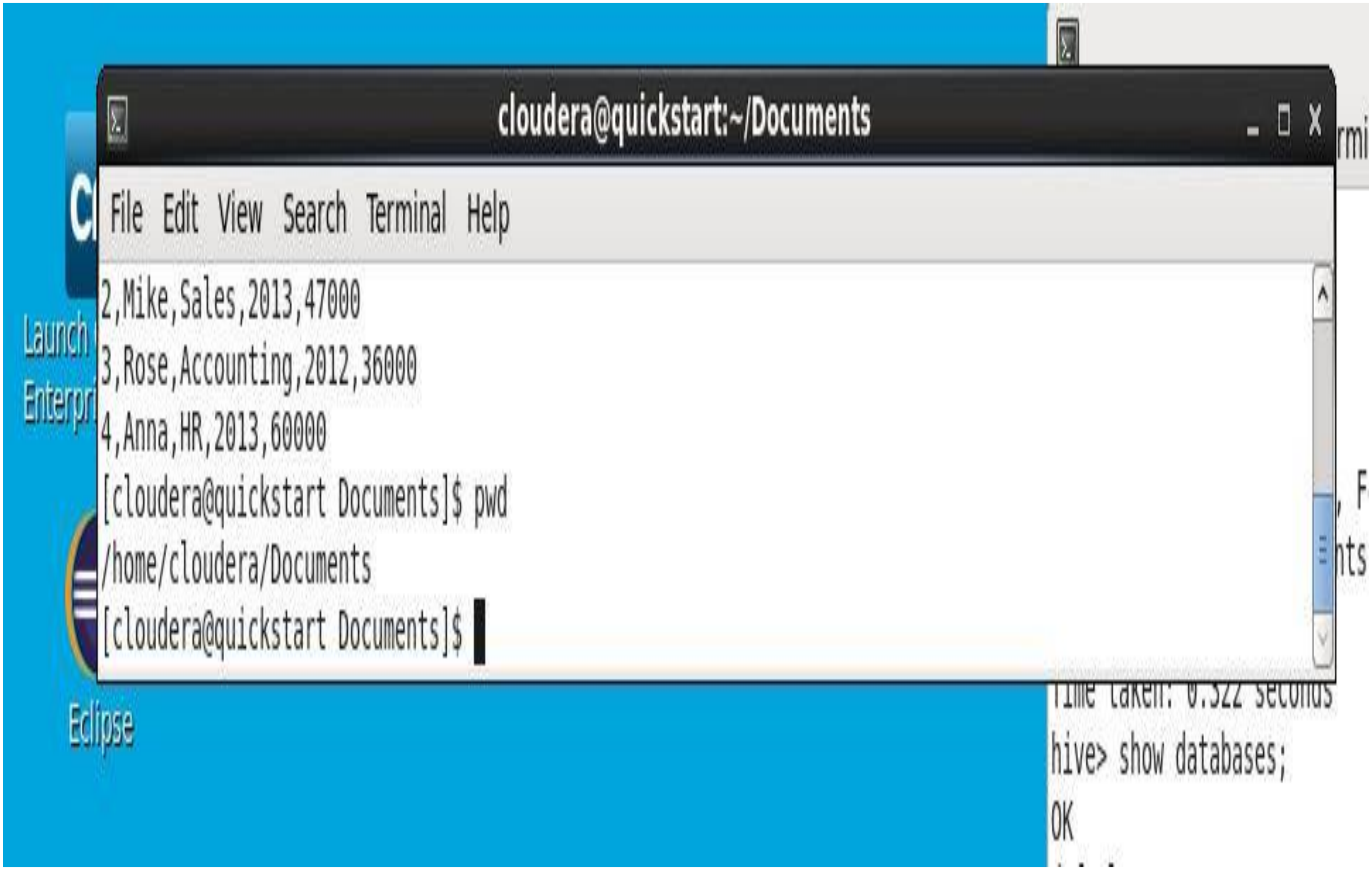
- We have a header line
- The data is comma-separated
- In Hadoop, you write the data once, and you read it many times. You must clean the data before it gets into the DB
- Do the Extract Transform Load (ETL)
- Schema: Integer, String, string, Integer, Integer


cloudera@quickstart:~/Documents
File Edit View Search Terminal Help
Launch
Enterpri
Ecl
documen

```
[cloudera@quickstart ~]$ ls
cloudera-manager  Downloads                   lib       Templates
cm_api.py         eclipse                     Music     Videos
Customer.csv      enterprise-deployment.json  parcels   workspace
Desktop           express-deployment.json     Pictures
Documents         kerberos                    Public
[cloudera@quickstart ~]$ cd Documents
[cloudera@quickstart Documents]$ ls
cloudera-manager.html  css  Customer.csv  Employee.csv  img
[cloudera@quickstart Documents]$ cat Employee.csv
ID,Name,Dept,JoiningDate,Salary
1,Alex,IT,2012,58000
2,Mike,Sales,2013,47000
3,Rose,Accounting,2012,36000
4,Anna,HR,2013,60000
[cloudera@quickstart Documents]$
```

7. Display the full path of the source of data

- Type command as '$pwd'
- Here u get the full path of the source of data



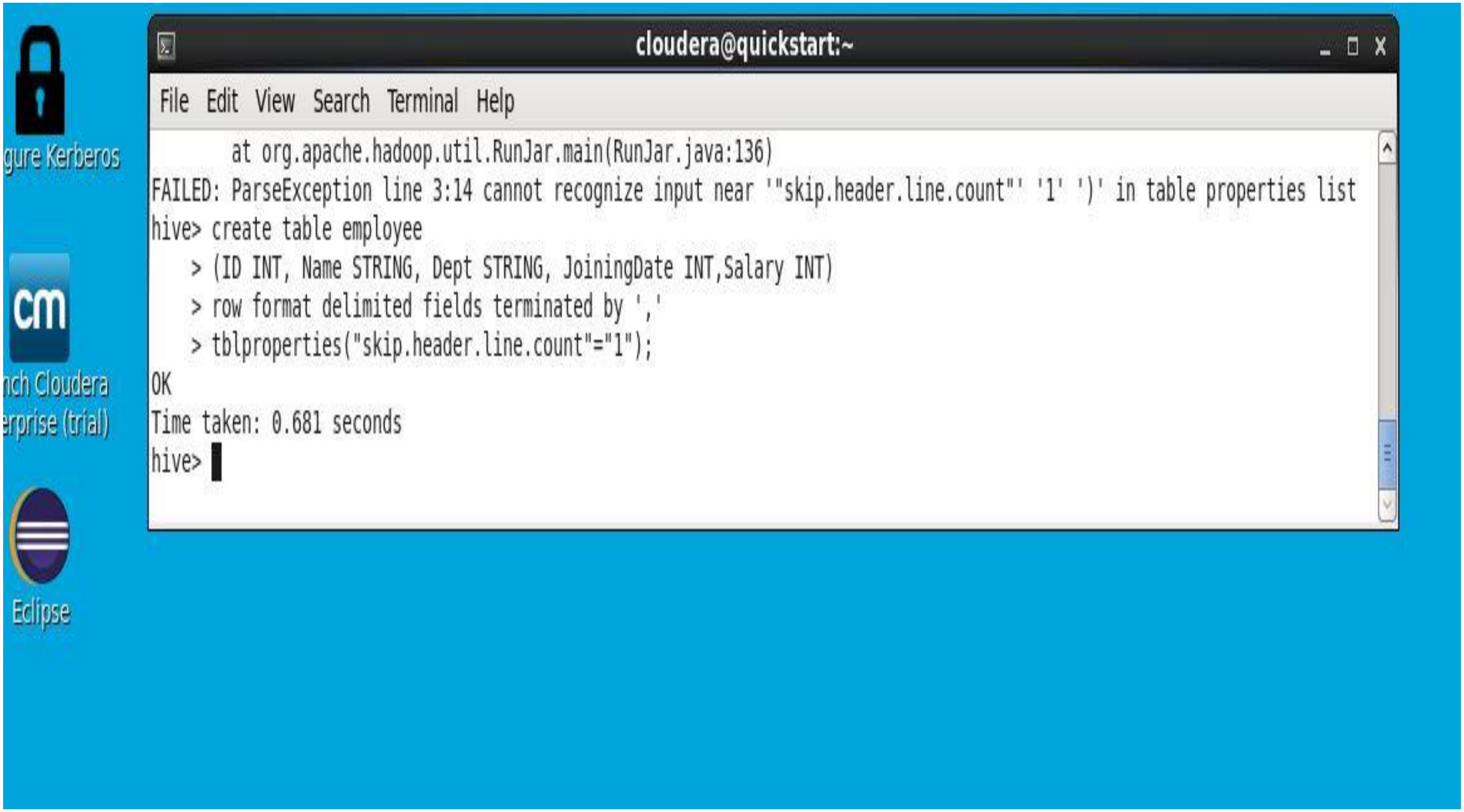
cloudera@quickstart:~/Documents
File Edit View Search Terminal Help
Launch
Enterpri
Eclipse

```
2,Mike,Sales,2013,47000
3,Rose,Accounting,2012,36000
4,Anna,HR,2013,60000
[cloudera@quickstart Documents]$ pwd
/home/cloudera/Documents
[cloudera@quickstart Documents]$
```

```
Time taken: 0.322 seconds
hive> show databases;
OK
```

8. Create a table named employee

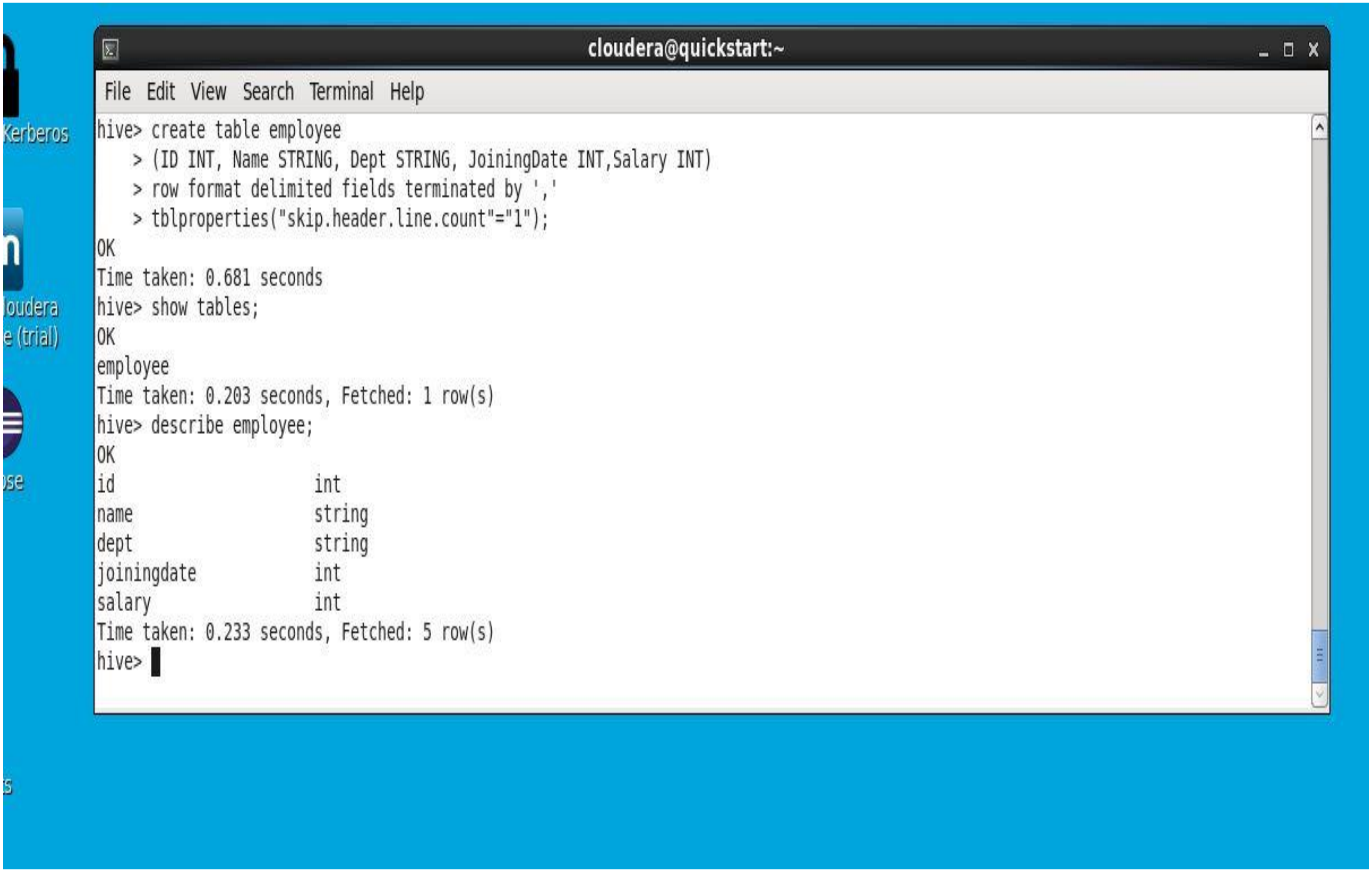
- Write the commands as
  1. Create table employee
  2. (ID INT, Name STRING, JoiningDate INT, Salary INT)
  3. row format delimited fields terminated by ‘,’
  4. tblproperties(“skip.header.line.count”=”1”);
- Then the table is created



```
        at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
FAILED: ParseException line 3:14 cannot recognize input near '"skip.header.line.count"' '1' ')' in table properties list
hive> create table employee
    > (ID INT, Name STRING, Dept STRING, JoiningDate INT,Salary INT)
    > row format delimited fields terminated by ','
    > tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.681 seconds
hive> 
```

9. Verify the table employee

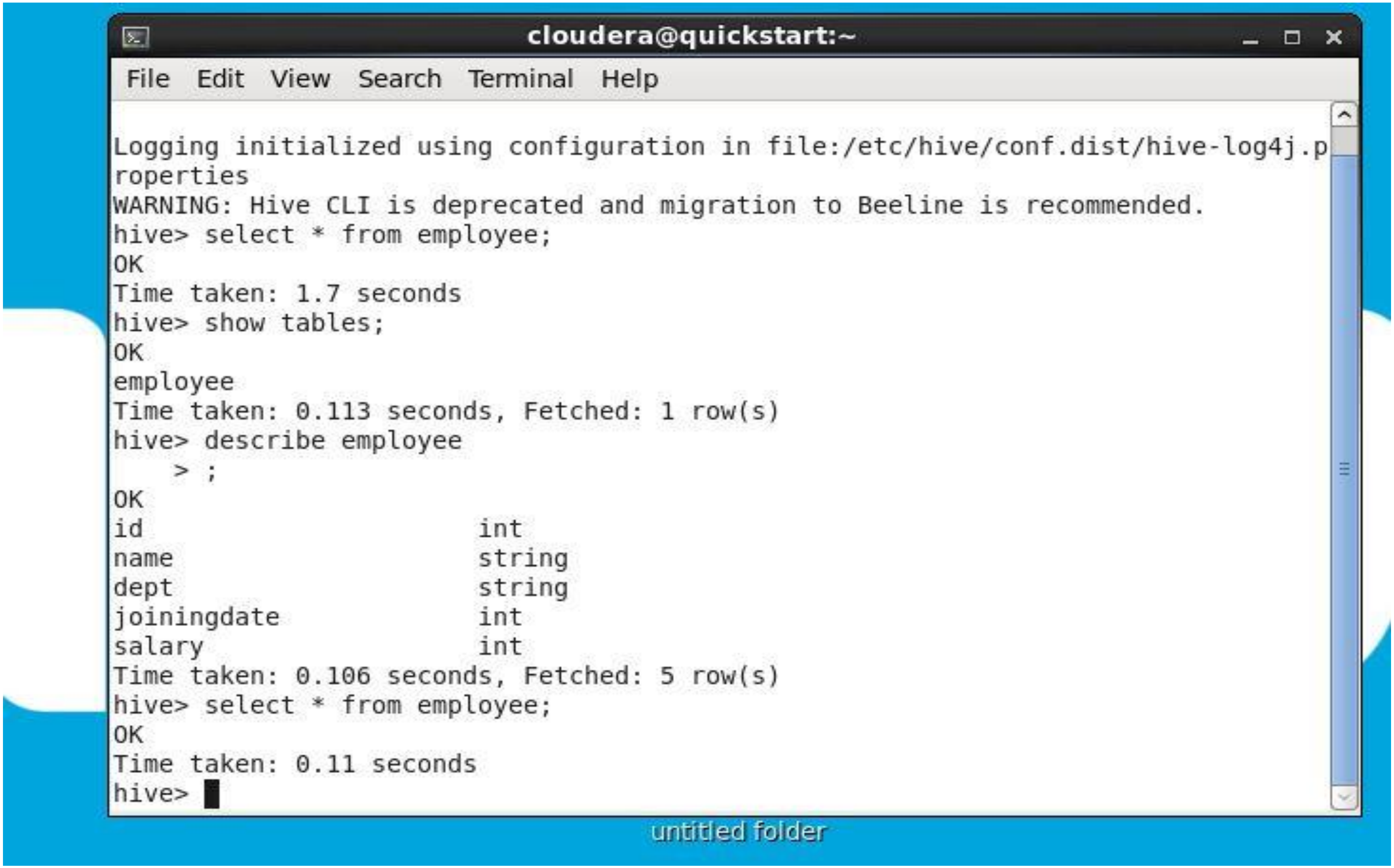
- For verifying the tables type command as ‘show tables ‘.
- For showing the data from table employee – ‘describe employee’.



```
hive> create table employee
    > (ID INT, Name STRING, Dept STRING, JoiningDate INT,Salary INT)
    > row format delimited fields terminated by ','
    > tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.681 seconds
hive> show tables;
OK
employee
Time taken: 0.203 seconds, Fetched: 1 row(s)
hive> describe employee;
OK
id                      int
name                    string
dept                    string
joiningdate             int
salary                  int
Time taken: 0.233 seconds, Fetched: 5 row(s)
hive>
```

10. Display the data in the table

- For displaying all data from table type command as ‘select * from employee;’.

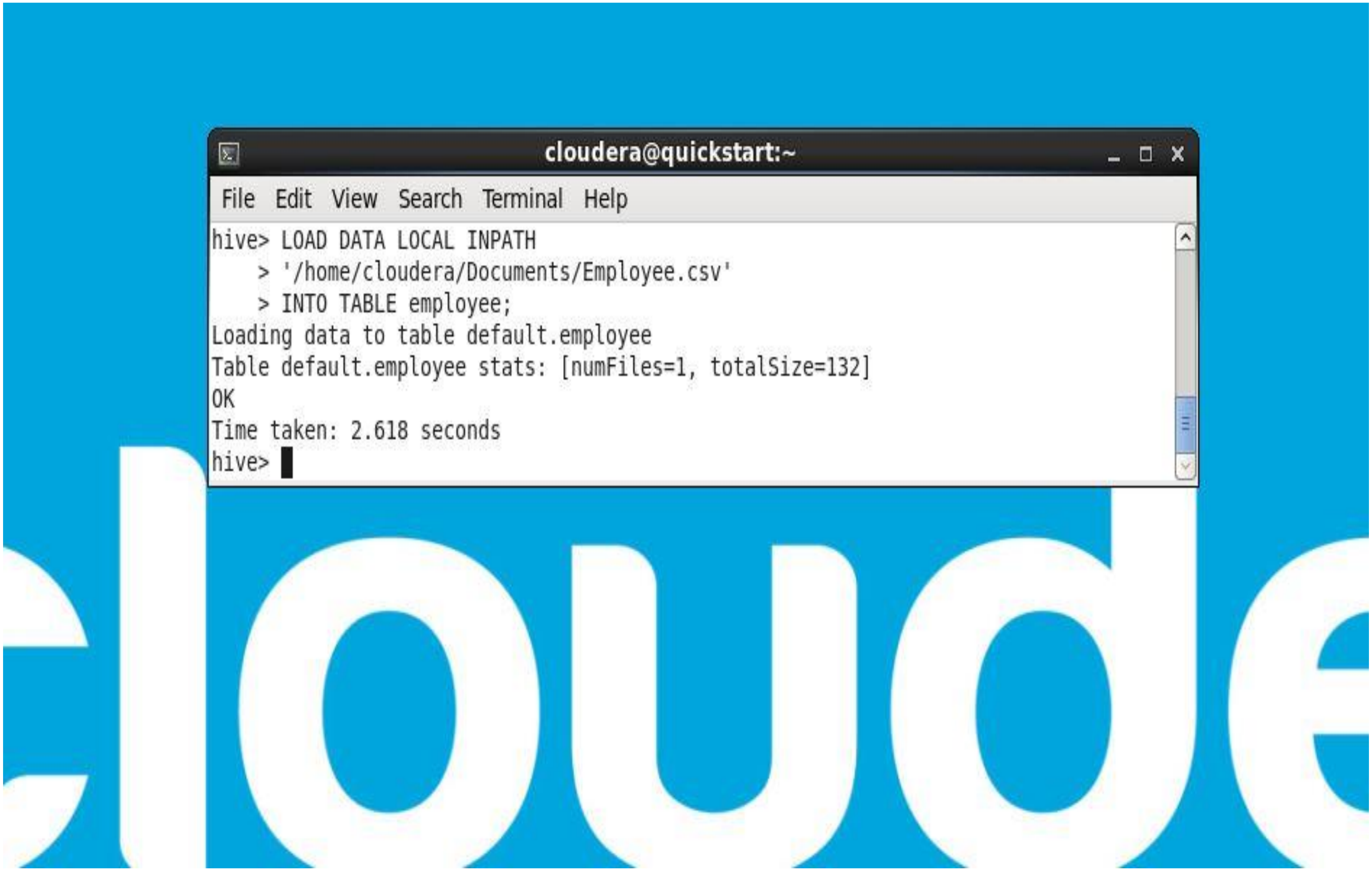


```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.p
roperties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> select * from employee;
OK
Time taken: 1.7 seconds
hive> show tables;
OK
employee
Time taken: 0.113 seconds, Fetched: 1 row(s)
hive> describe employee
    > ;
OK
id                      int
name                    string
dept                    string
joiningdate             int
salary                  int
Time taken: 0.106 seconds, Fetched: 5 row(s)
hive> select * from employee;
OK
Time taken: 0.11 seconds
hive> 
```

11. Loading the data to Hive

- Type command as ‘LOAD DATA LOACAL INPATH
  ‘/HOME/CLODERA/Documents/Employee.csv’

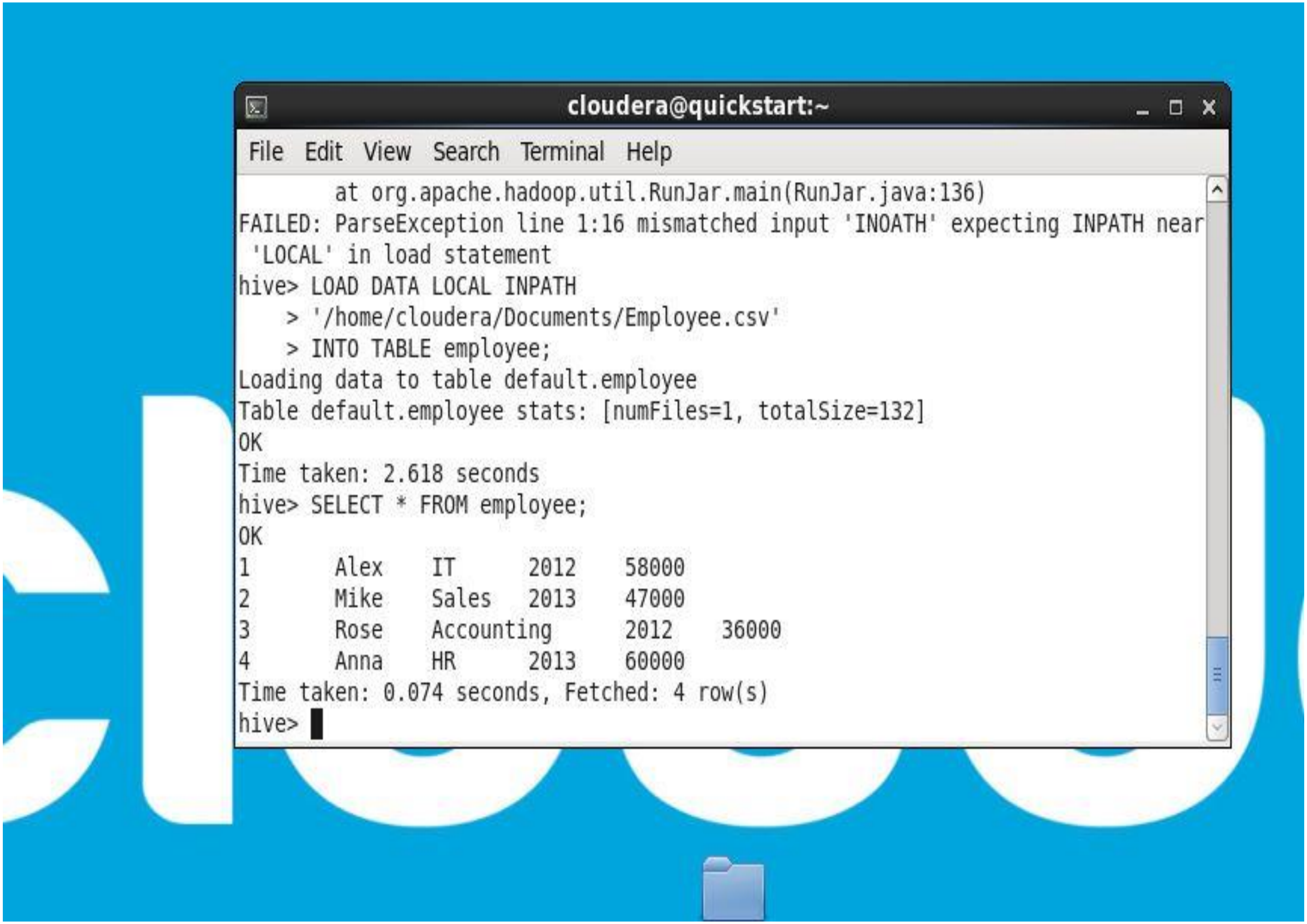
  INTO TABLE employee;’



```
hive> LOAD DATA LOCAL INPATH
    > '/home/cloudera/Documents/Employee.csv'
    > INTO TABLE employee;
Loading data to table default.employee
Table default.employee stats: [numFiles=1, totalSize=132]
OK
Time taken: 2.618 seconds
hive> 
```

12. Display the data

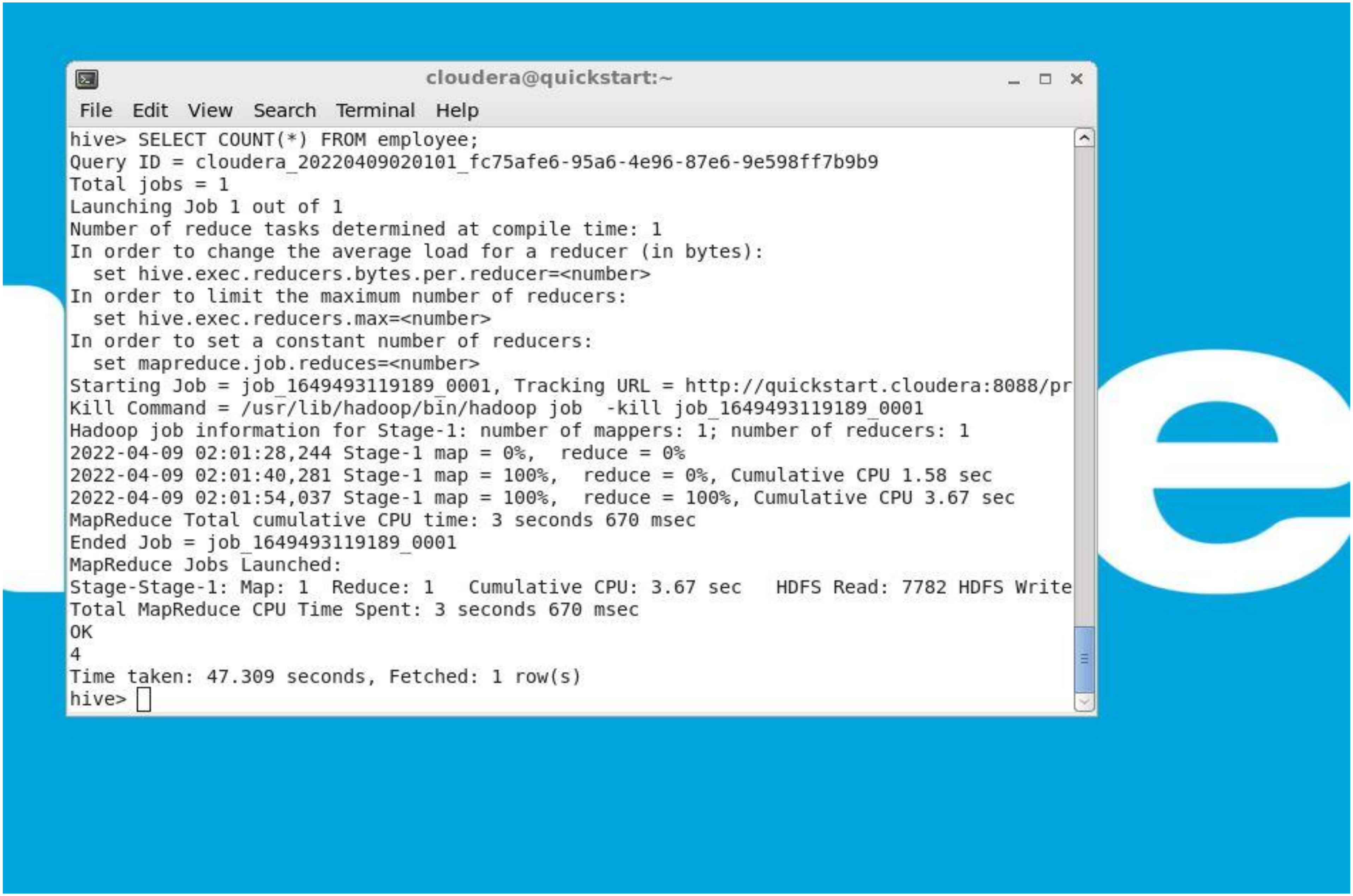
- Select * from employee;



```
        at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
FAILED: ParseException line 1:16 mismatched input 'INOATH' expecting INPATH near
 'LOCAL' in load statement
hive> LOAD DATA LOCAL INPATH
    > '/home/cloudera/Documents/Employee.csv'
    > INTO TABLE employee;
Loading data to table default.employee
Table default.employee stats: [numFiles=1, totalSize=132]
OK
Time taken: 2.618 seconds
hive> SELECT * FROM employee;
OK
1       Alex    IT      2012    58000
2       Mike    Sales   2013    47000
3       Rose    Accounting      2012    36000
4       Anna    HR      2013    60000
Time taken: 0.074 seconds, Fetched: 4 row(s)
hive> 
```

13. Try the full Map Reduce Phase by the Count(*)

Select count(*) from employee;



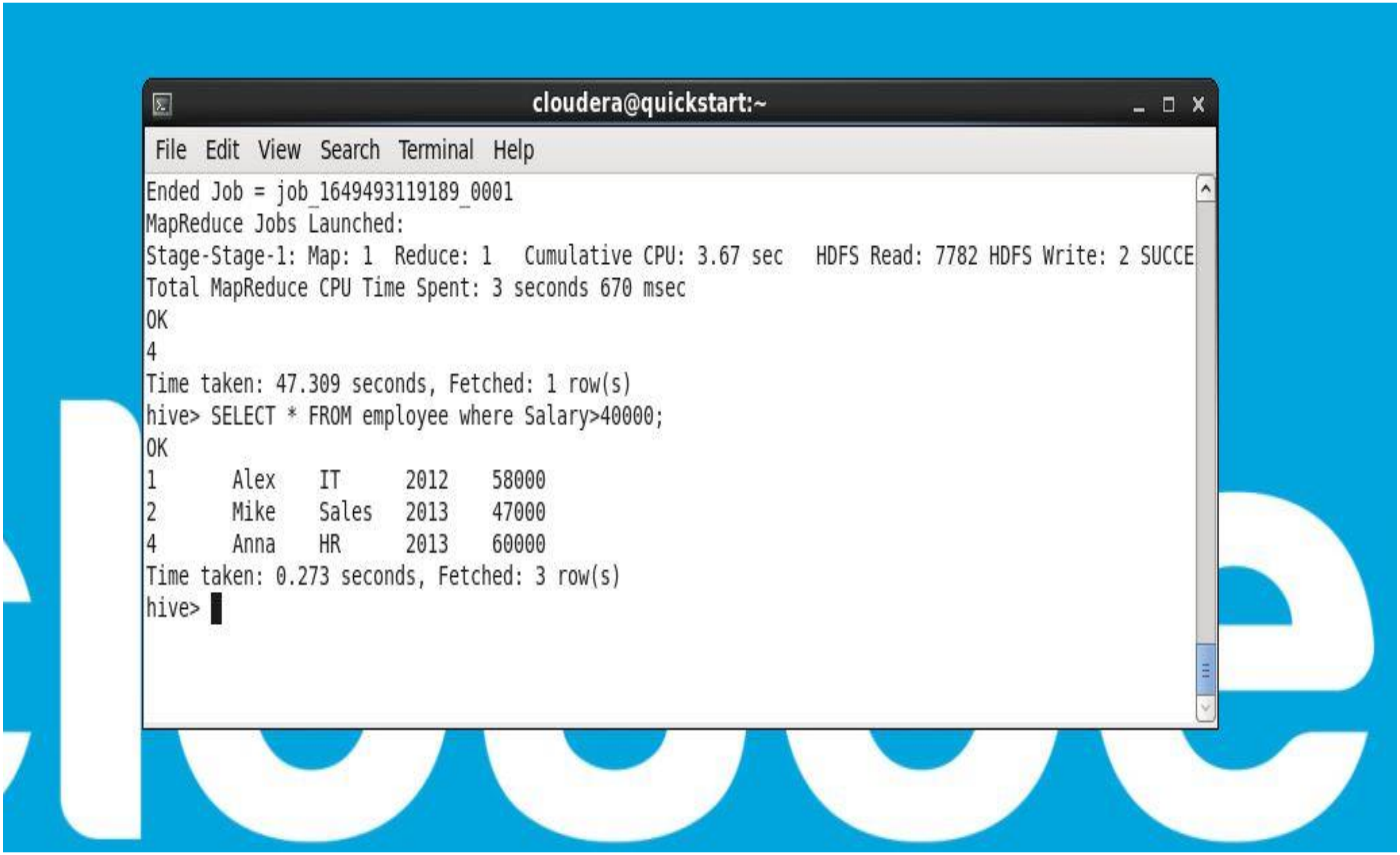
```
hive> SELECT COUNT(*) FROM employee;
Query ID = cloudera_20220409020101_fc75afe6-95a6-4e96-87e6-9e598ff7b9b9
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1649493119189_0001, Tracking URL = http://quickstart.cloudera:8088/pr
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1649493119189_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-04-09 02:01:28,244 Stage-1 map = 0%,  reduce = 0%
2022-04-09 02:01:40,281 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.58 sec
2022-04-09 02:01:54,037 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.67 sec
MapReduce Total cumulative CPU time: 3 seconds 670 msec
Ended Job = job_1649493119189_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.67 sec   HDFS Read: 7782 HDFS Write
Total MapReduce CPU Time Spent: 3 seconds 670 msec
OK
4
Time taken: 47.309 seconds, Fetched: 1 row(s)
hive> 
```

14. More Enhanced query

select *

from employee

where salary>40000;

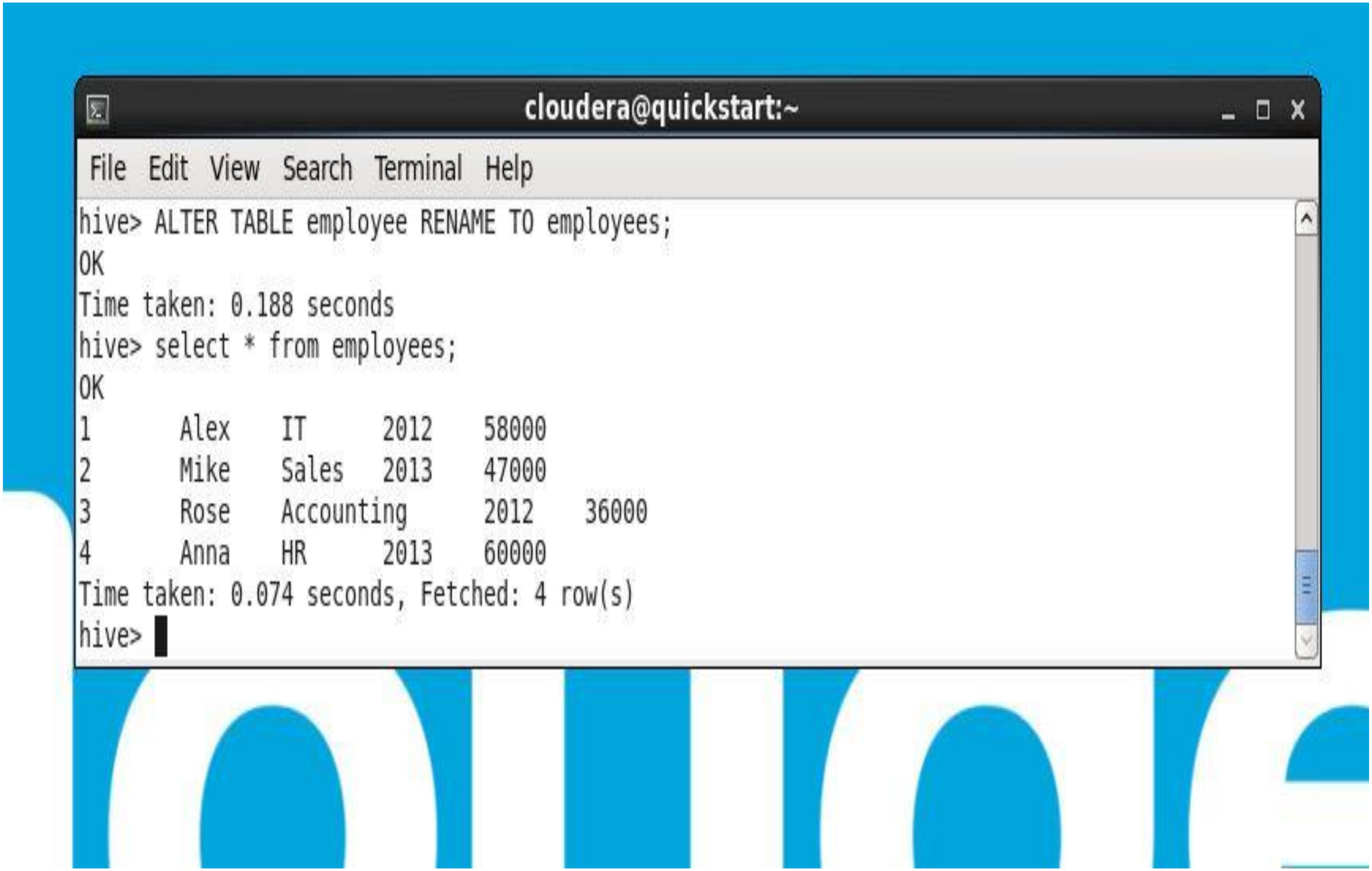


```
Ended Job = job_1649493119189_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.67 sec   HDFS Read: 7782 HDFS Write: 2 SUCCE
Total MapReduce CPU Time Spent: 3 seconds 670 msec
OK
4
Time taken: 47.309 seconds, Fetched: 1 row(s)
hive> SELECT * FROM employee where Salary>40000;
OK
1       Alex    IT      2012    58000
2       Mike    Sales   2013    47000
4       Anna    HR      2013    60000
Time taken: 0.273 seconds, Fetched: 3 row(s)
hive> 
```

15. Renaming the table

ALTER TABLE employee

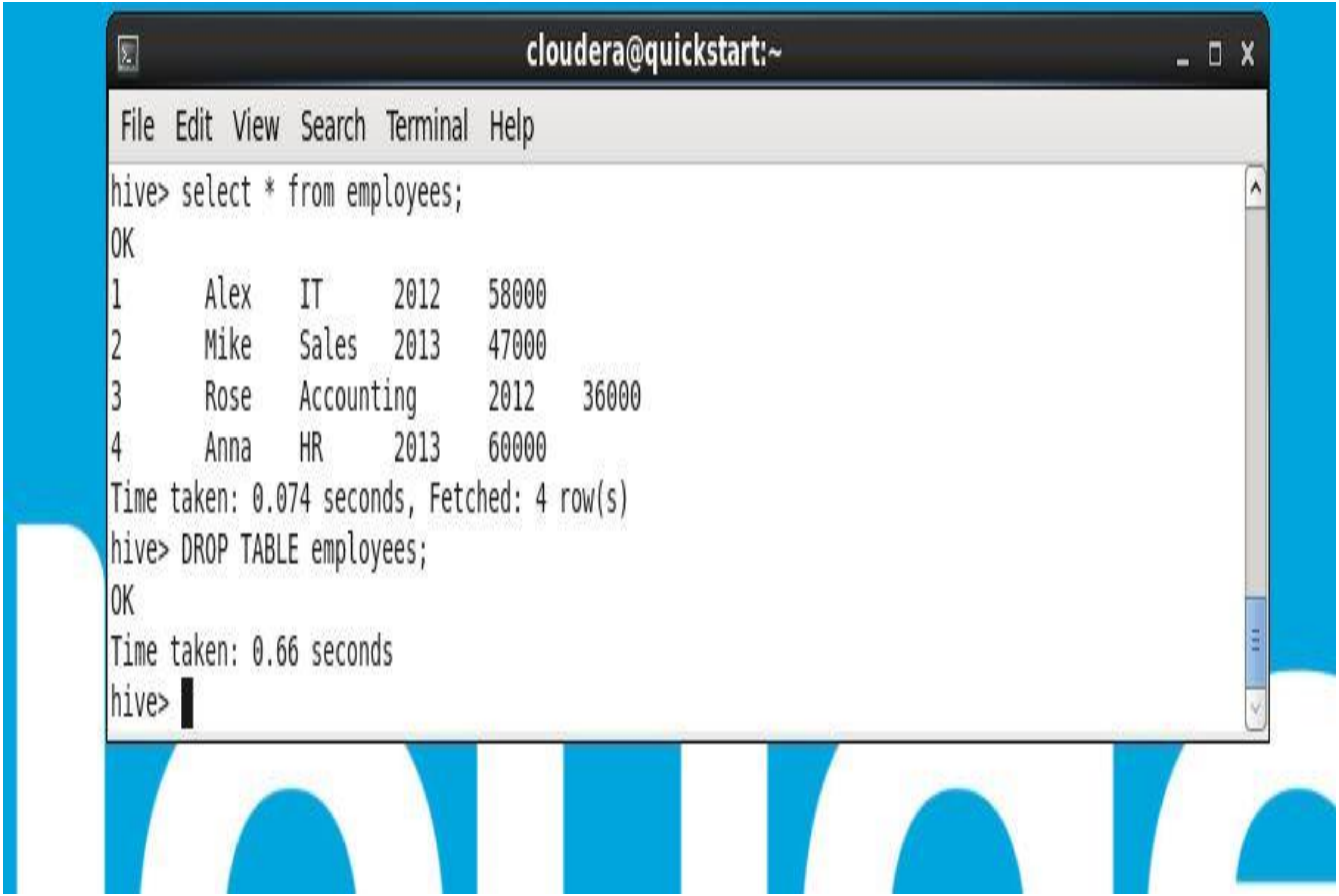
RENAME TO employees;



```
hive> ALTER TABLE employee RENAME TO employees;
OK
Time taken: 0.188 seconds
hive> select * from employees;
OK
1       Alex    IT      2012    58000
2       Mike    Sales   2013    47000
3       Rose    Accounting      2012    36000
4       Anna    HR      2013    60000
Time taken: 0.074 seconds, Fetched: 4 row(s)
hive> 
```

16. Drop the table employees

Drop table employees;



```
hive> select * from employees;
OK
1       Alex    IT      2012    58000
2       Mike    Sales   2013    47000
3       Rose    Accounting      2012    36000
4       Anna    HR      2013    60000
Time taken: 0.074 seconds, Fetched: 4 row(s)
hive> DROP TABLE employees;
OK
Time taken: 0.66 seconds
hive> 
```

17. Create the table order

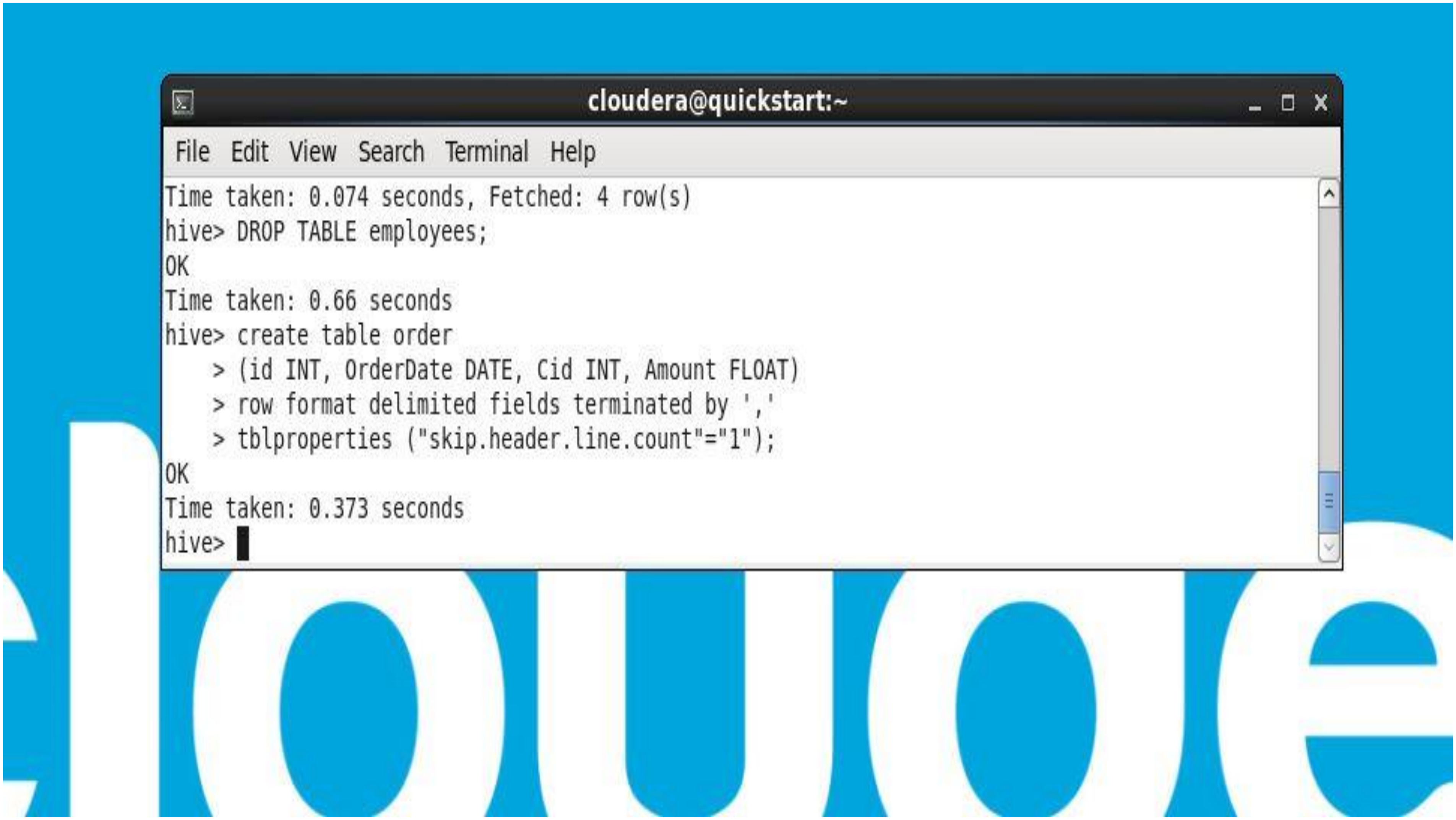
- We have to create the table called order.

Create table order

(id INT, orderdateDate DATE, Cid INT,Amount Float)

Row format delimited fields terminated by ‘,’

Tbl properties(“skip.header.line.count”=”1”);



```
Time taken: 0.074 seconds, Fetched: 4 row(s)
hive> DROP TABLE employees;
OK
Time taken: 0.66 seconds
hive> create table order
    > (id INT, OrderDate DATE, Cid INT, Amount FLOAT)
    > row format delimited fields terminated by ','
    > tblproperties ("skip.header.line.count"="1");
OK
Time taken: 0.373 seconds
hive> 
```

To load data in the table we have to use command

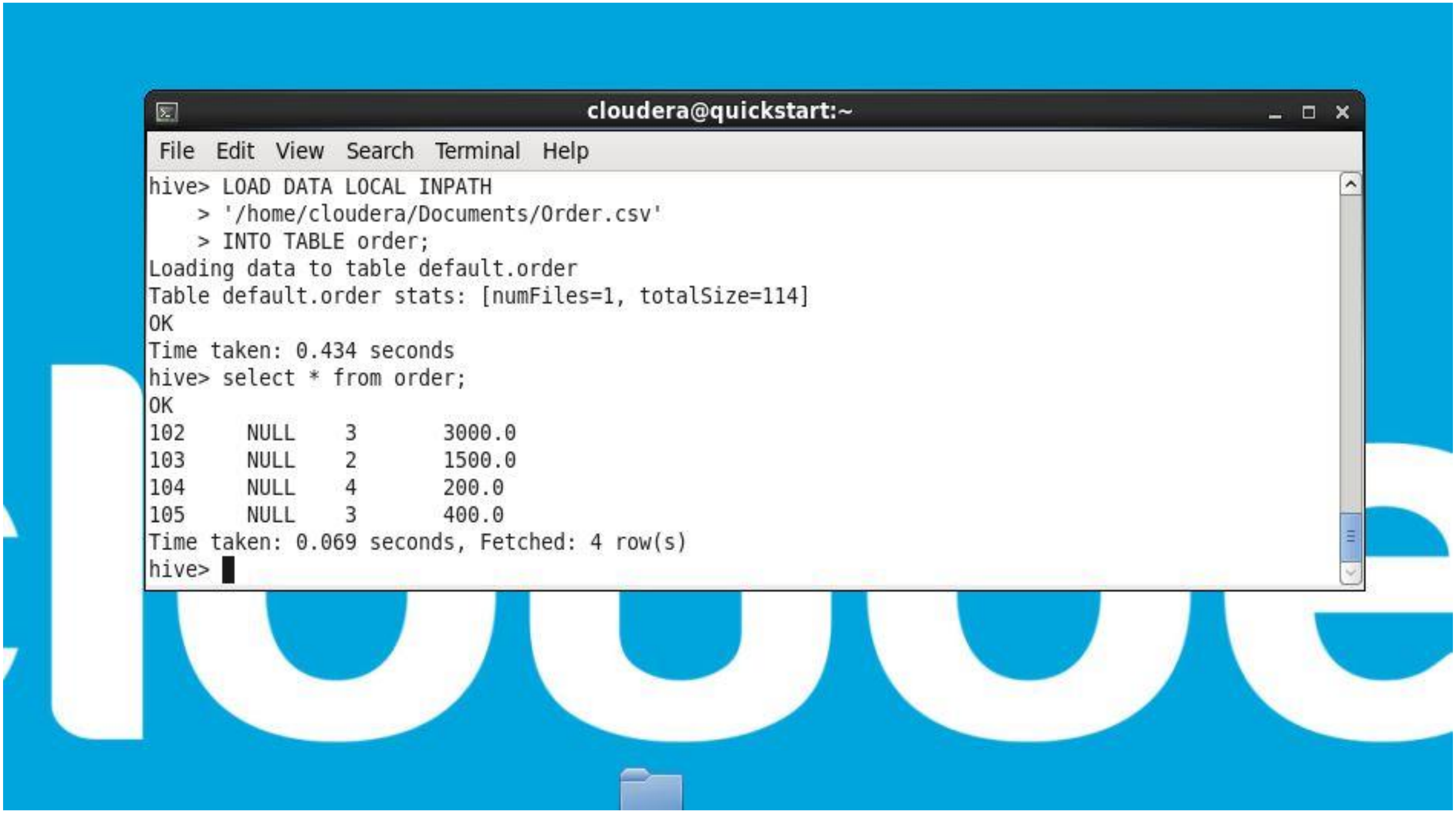
LOAD DATA LOCAL INPATH

‘/home/cloudera/Documents/Order.csv’

INTO TABLE order;

Then data is loaded in file and the for showing the data we have to use

Select * from order;



```
hive> LOAD DATA LOCAL INPATH
    > '/home/cloudera/Documents/Order.csv'
    > INTO TABLE order;
Loading data to table default.order
Table default.order stats: [numFiles=1, totalSize=114]
OK
Time taken: 0.434 seconds
hive> select * from order;
OK
102     NULL    3       3000.0
103     NULL    2       1500.0
104     NULL    4       200.0
105     NULL    3       400.0
Time taken: 0.069 seconds, Fetched: 4 row(s)
hive> 
```

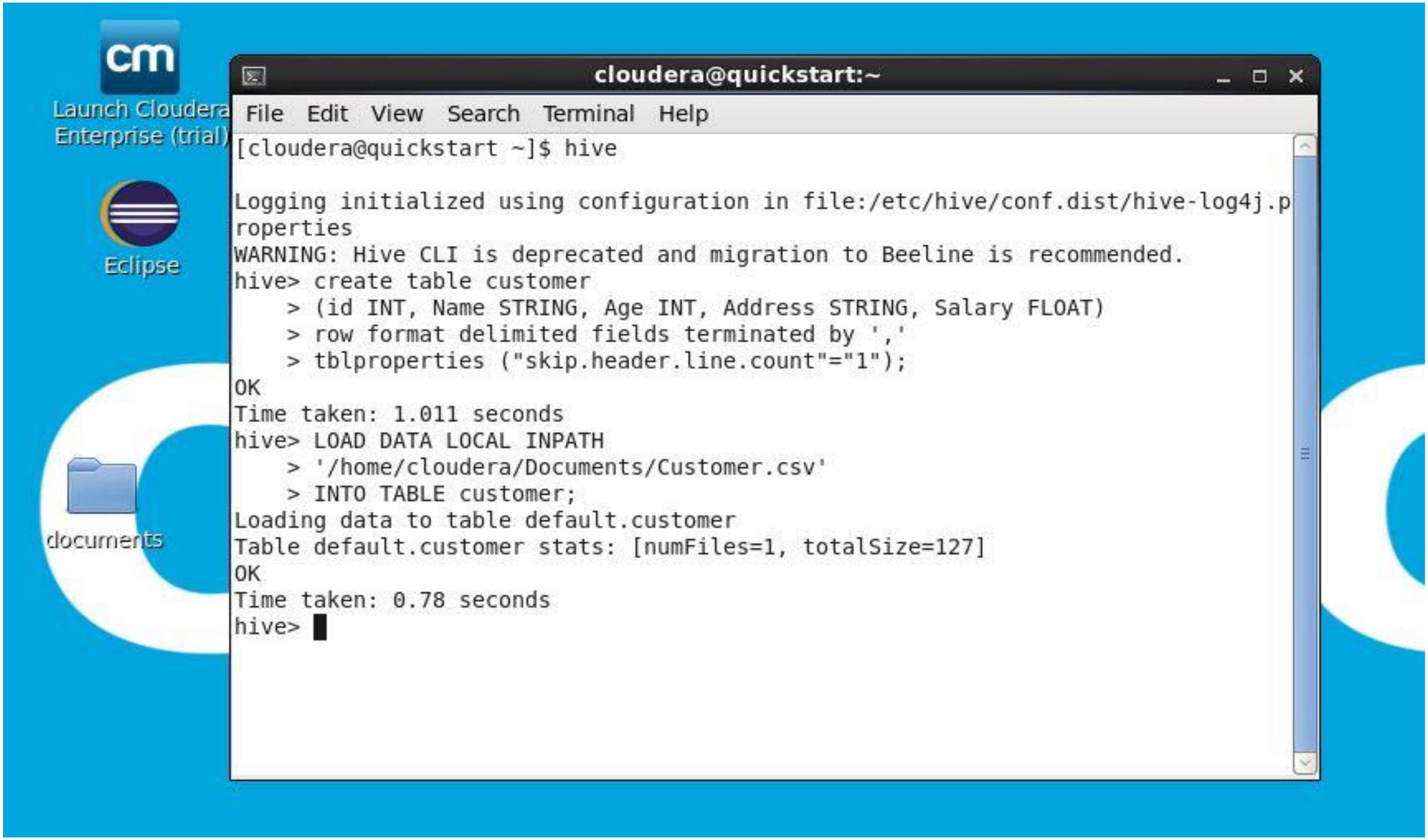
18. Same for table Customer

Create table order

(id INT, orderdateDate DATE, Cid INT,Amount Float)

Row format delimited fields terminated by ‘,’

Tbl properties(“skip.header.line.count”=”1”);



Launch Cloudera Enterprise (trial)
Eclipse
documents
cloudera@quickstart:~
File Edit View Search Terminal Help

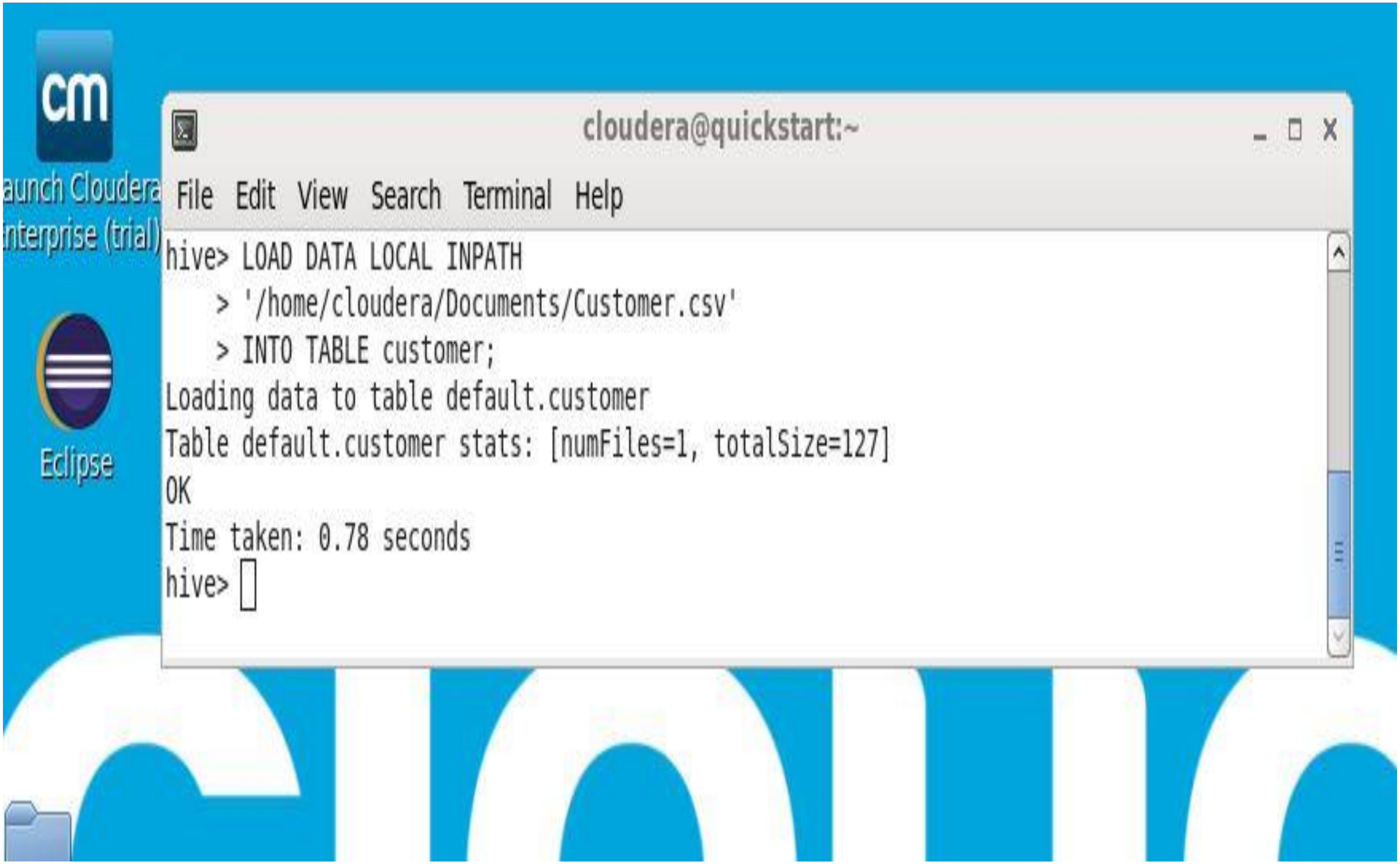
```
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.p
roperties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> create table customer
    > (id INT, Name STRING, Age INT, Address STRING, Salary FLOAT)
    > row format delimited fields terminated by ','
    > tblproperties ("skip.header.line.count"="1");
OK
Time taken: 1.011 seconds
hive> LOAD DATA LOCAL INPATH
    > '/home/cloudera/Documents/Customer.csv'
    > INTO TABLE customer;
Loading data to table default.customer
Table default.customer stats: [numFiles=1, totalSize=127]
OK
Time taken: 0.78 seconds
hive> 
```

For loading data into the table from file path:

LOAD DATA LOCAL INPATH

‘/home/clodera/Documents/Customer.csv’

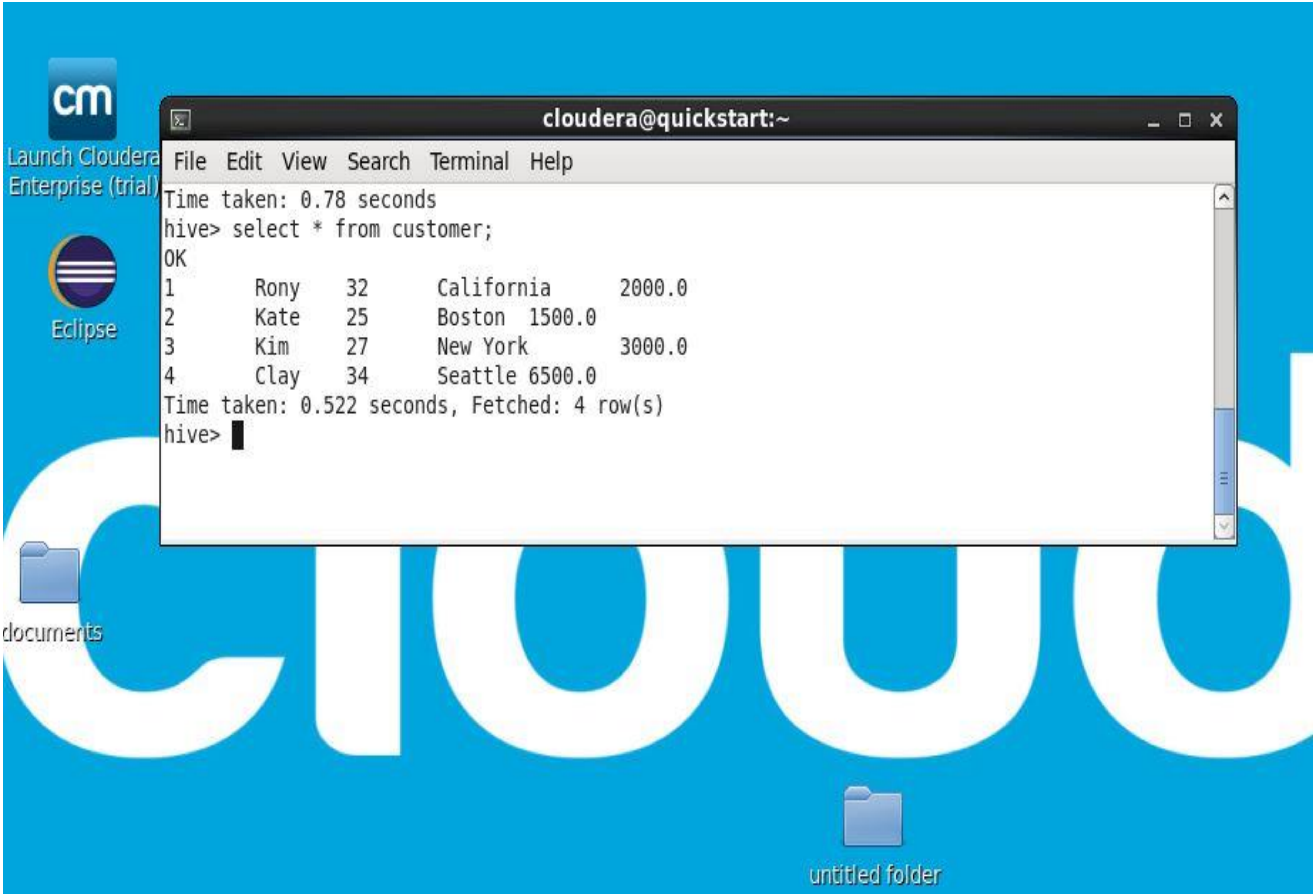
INTO TABLE customer;



```
hive> LOAD DATA LOCAL INPATH
    > '/home/cloudera/Documents/Customer.csv'
    > INTO TABLE customer;
Loading data to table default.customer
Table default.customer stats: [numFiles=1, totalSize=127]
OK
Time taken: 0.78 seconds
hive> 
```

For displaying data, we have to use:

Select * from customer;



Launch Cloudera Enterprise (trial)
Eclipse
documents
untitled folder
cloudera@quickstart:~
File Edit View Search Terminal Help

```
Time taken: 0.78 seconds
hive> select * from customer;
OK
1       Rony    32      California      2000.0
2       Kate    25      Boston  1500.0
3       Kim     27      New York        3000.0
4       Clay    34      Seattle 6500.0
Time taken: 0.522 seconds, Fetched: 4 row(s)
hive> 
```

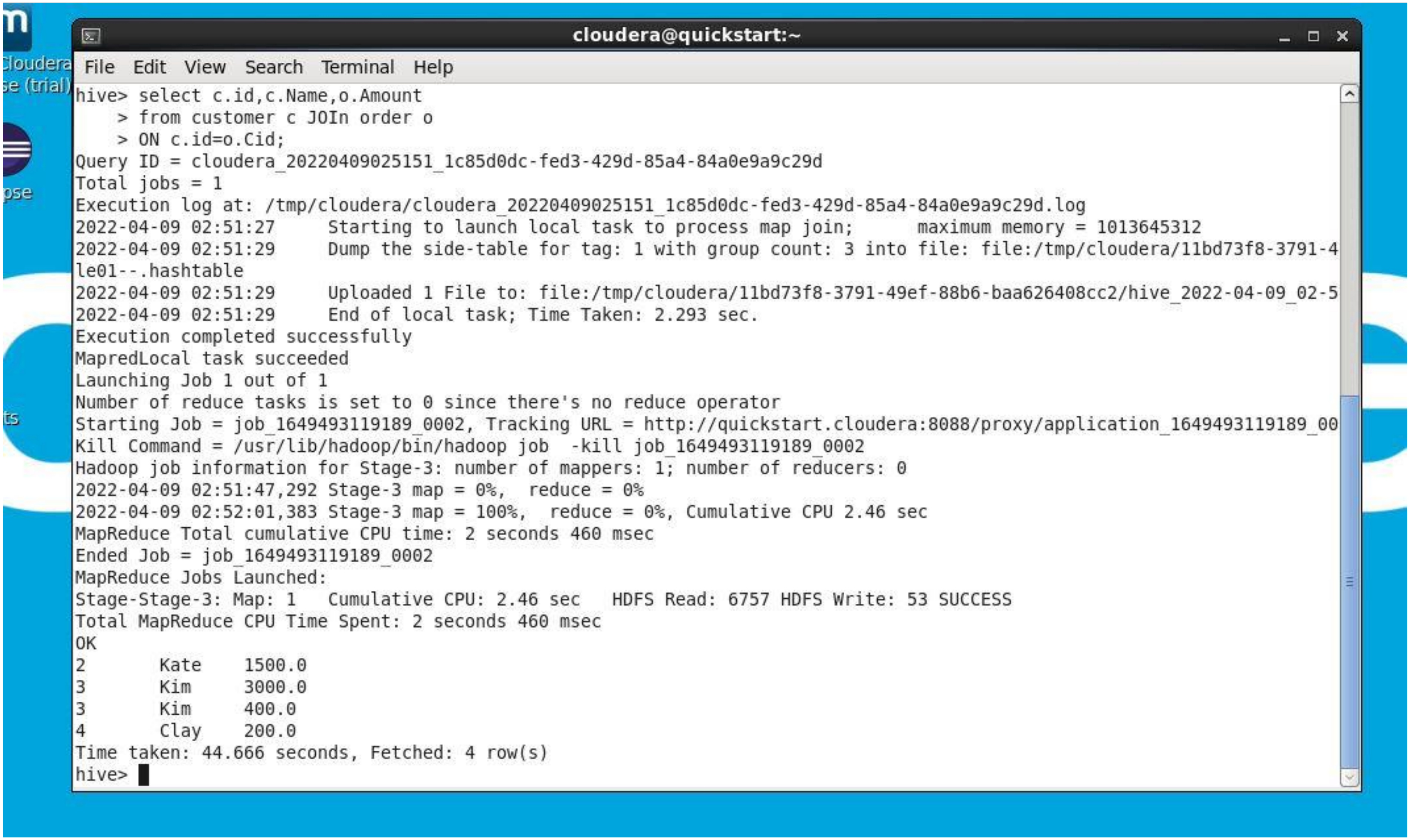
19. A Complex Query

Joining table query:

Select c.id, c.Name, o.Amount

From customer c JOIN order o

ON c.id = o.Cid;

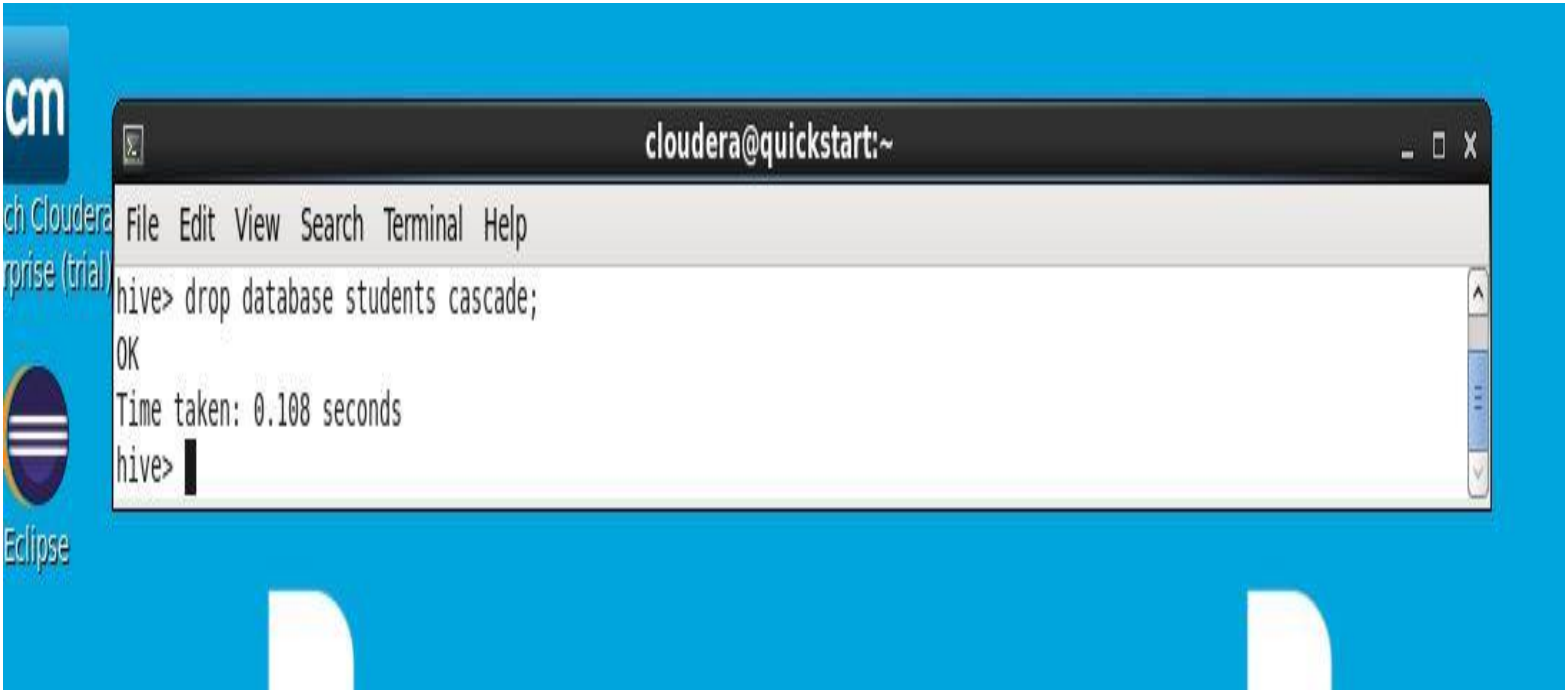


```
hive> select c.id,c.Name,o.Amount
    > from customer c JOIn order o
    > ON c.id=o.Cid;
Query ID = cloudera_20220409025151_1c85d0dc-fed3-429d-85a4-84a0e9a9c29d
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20220409025151_1c85d0dc-fed3-429d-85a4-84a0e9a9c29d.log
2022-04-09 02:51:27     Starting to launch local task to process map join;      maximum memory = 1013645312
2022-04-09 02:51:29     Dump the side-table for tag: 1 with group count: 3 into file: file:/tmp/cloudera/11bd73f8-3791-4
le01--.hashtable
2022-04-09 02:51:29     Uploaded 1 File to: file:/tmp/cloudera/11bd73f8-3791-49ef-88b6-baa626408cc2/hive_2022-04-09_02-5
2022-04-09 02:51:29     End of local task; Time Taken: 2.293 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1649493119189_0002, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1649493119189_00
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1649493119189_0002
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2022-04-09 02:51:47,292 Stage-3 map = 0%,  reduce = 0%
2022-04-09 02:52:01,383 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 2.46 sec
MapReduce Total cumulative CPU time: 2 seconds 460 msec
Ended Job = job_1649493119189_0002
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 2.46 sec   HDFS Read: 6757 HDFS Write: 53 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 460 msec
OK
2       Kate    1500.0
3       Kim     3000.0
3       Kim     400.0
4       Clay    200.0
Time taken: 44.666 seconds, Fetched: 4 row(s)
hive> 
```

20. Drop DB

For dropping table we have to use command as:

Drop database students cascade;



```
hive> drop database students cascade;
OK
Time taken: 0.108 seconds
hive> 
```