Ecommerce pipeline project documentation

1) **Mysql :** For this we to create the separate _E_commerce_database in mysql then we have to create schema of 3 tables , after that we import the csv file into 3 tables which is already created.

   Commands are :

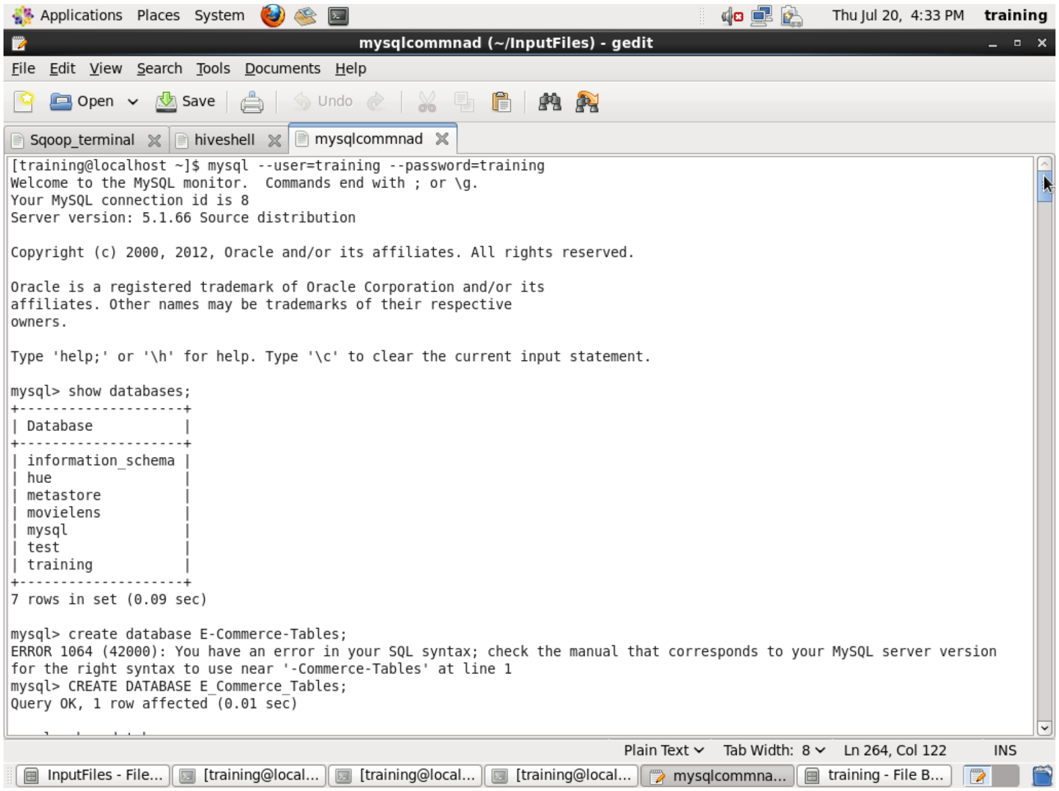
```
mysql> CREATE TABLE clickStream_data (
    -> userID INT,
    -> timestamp TIMESTAMP,
    -> page VARCHAR(50)
    -> );
mysql> CREATE TABLE customer_data(
    -> userID INT,
    -> name VARCHAR(50),
    -> email VARCHAR(100)
    -> );

mysql> CREATE TABLE purchase_data( userID INT,
    -> timestamp TIMESTAMP,
    -> amount DECIMAL(10, 2)
    -> );
```

```
mysql> LOAD DATA LOCAL INFILE '/home/training/InputFiles/customer.csv' INTO
TABLE customer_data FIELDS TERMINATED BY ',' IGNORE 1 LINES;

Query OK, 5 rows affected (0.00 sec)


mysql> LOAD DATA LOCAL INFILE '/home/training/InputFiles/purchase.csv' INTO
TABLE purchase_data FIELDS TERMINATED BY ',' IGNORE 1 LINES;

Query OK, 5 rows affected (0.00 sec)

Records: 5  Deleted: 0  Skipped: 0  Warnings: 0


mysql> SELECT * FROM CUSTOMER_DATA;

ERROR 1146 (42S02): Table 'E_Commerce_Tables.CUSTOMER_DATA' doesn't exist

mysql> SELECT * FROM customer_data;
```

⇨ There is some picture of terminal of mysql commands then I use sqoop for transfer the table from mqsql to hdfs after this I use hive shell for transfer the data from hdfs to hive default storage
/user/hive/warehouse

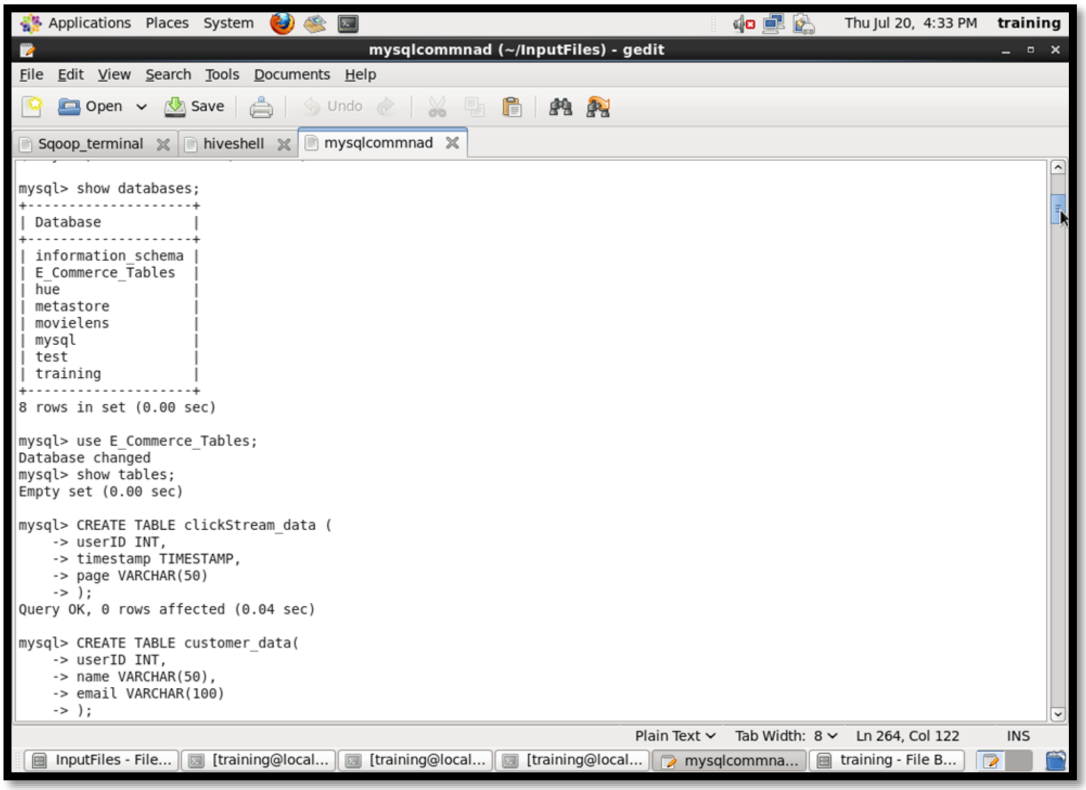
After this run some query for see there is no duplicate in table and there is no vacant row in a table and also find there is some data max and min to purchase table.

Join the table for better clarification of customer and purchase information .



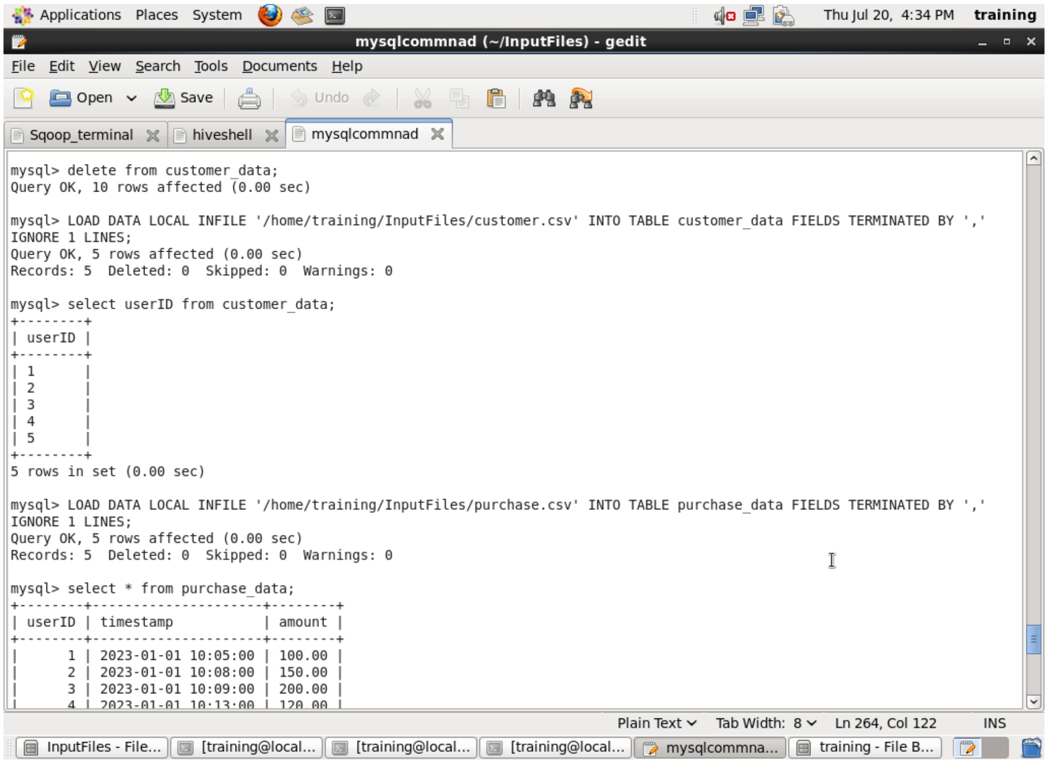
Applications Places System
Thu Jul 20, 4:33 PM training
mysqlcommnad (~/InputFiles) - gedit
File Edit View Search Tools Documents Help
Open Save Undo
Sqoop_terminal hiveshell mysqlcommnad
Plain Text Tab Width: 8 Ln 264, Col 122 INS
InputFiles - File... [training@local... [training@local... [training@local... mysqlcommna... training - File B...

```
[training@localhost ~]$ mysql --user=training --password=training
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 8
Server version: 5.1.66 Source distribution

Copyright (c) 2000, 2012, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| hue                |
| metastore          |
| movielens          |
| mysql              |
| test               |
| training           |
+--------------------+
7 rows in set (0.09 sec)

mysql> create database E-Commerce-Tables;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version
for the right syntax to use near '-Commerce-Tables' at line 1
mysql> CREATE DATABASE E_Commerce_Tables;
Query OK, 1 row affected (0.01 sec)
```

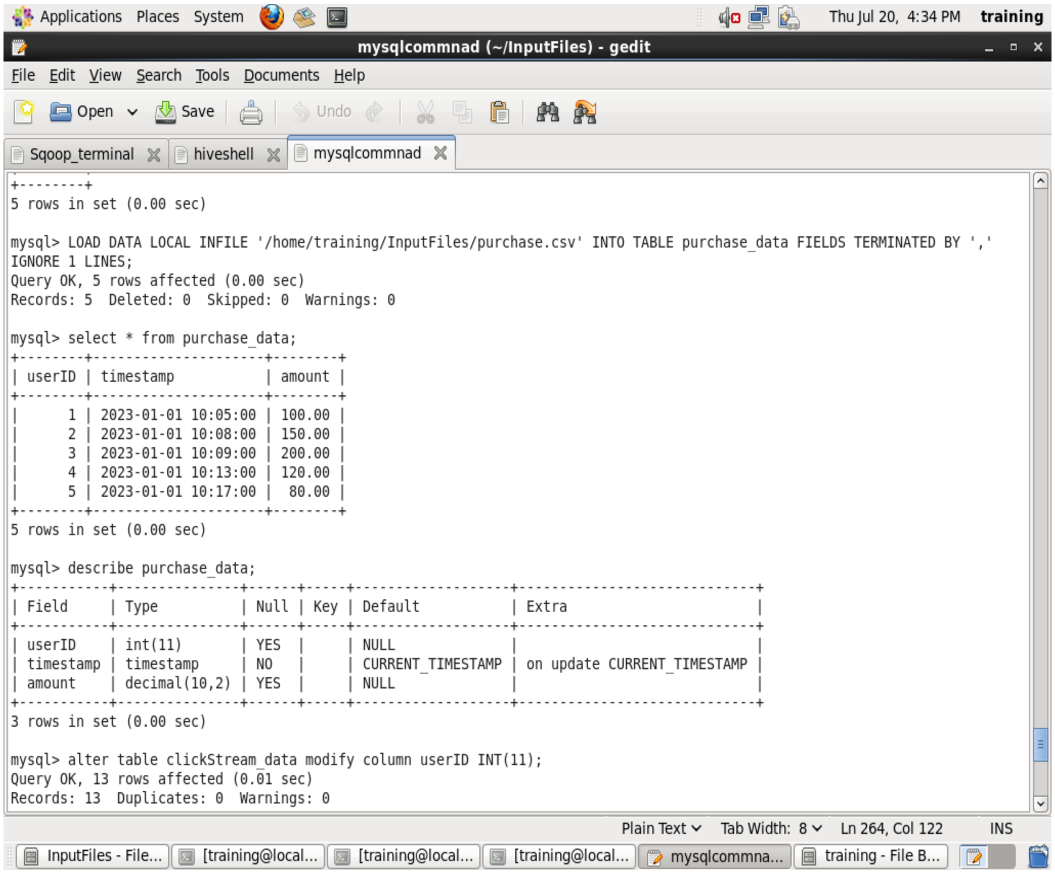


Applications Places System
Thu Jul 20, 4:33 PM training
mysqlcommnad (~/InputFiles) - gedit
File Edit View Search Tools Documents Help
Open Save Undo
Sqoop_terminal hiveshell mysqlcommnad
Plain Text Tab Width: 8 Ln 264, Col 122 INS
InputFiles - File... [training@local... [training@local... [training@local... mysqlcommna... training - File B...

```
mysql> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| E_Commerce_Tables  |
| hue                |
| metastore          |
| movielens          |
| mysql              |
| test               |
| training           |
+--------------------+
8 rows in set (0.00 sec)

mysql> use E_Commerce_Tables;
Database changed
mysql> show tables;
Empty set (0.00 sec)

mysql> CREATE TABLE clickStream_data (
    -> userID INT,
    -> timestamp TIMESTAMP,
    -> page VARCHAR(50)
    -> );
Query OK, 0 rows affected (0.04 sec)

mysql> CREATE TABLE customer_data(
    -> userID INT,
    -> name VARCHAR(50),
    -> email VARCHAR(100)
    -> );
```

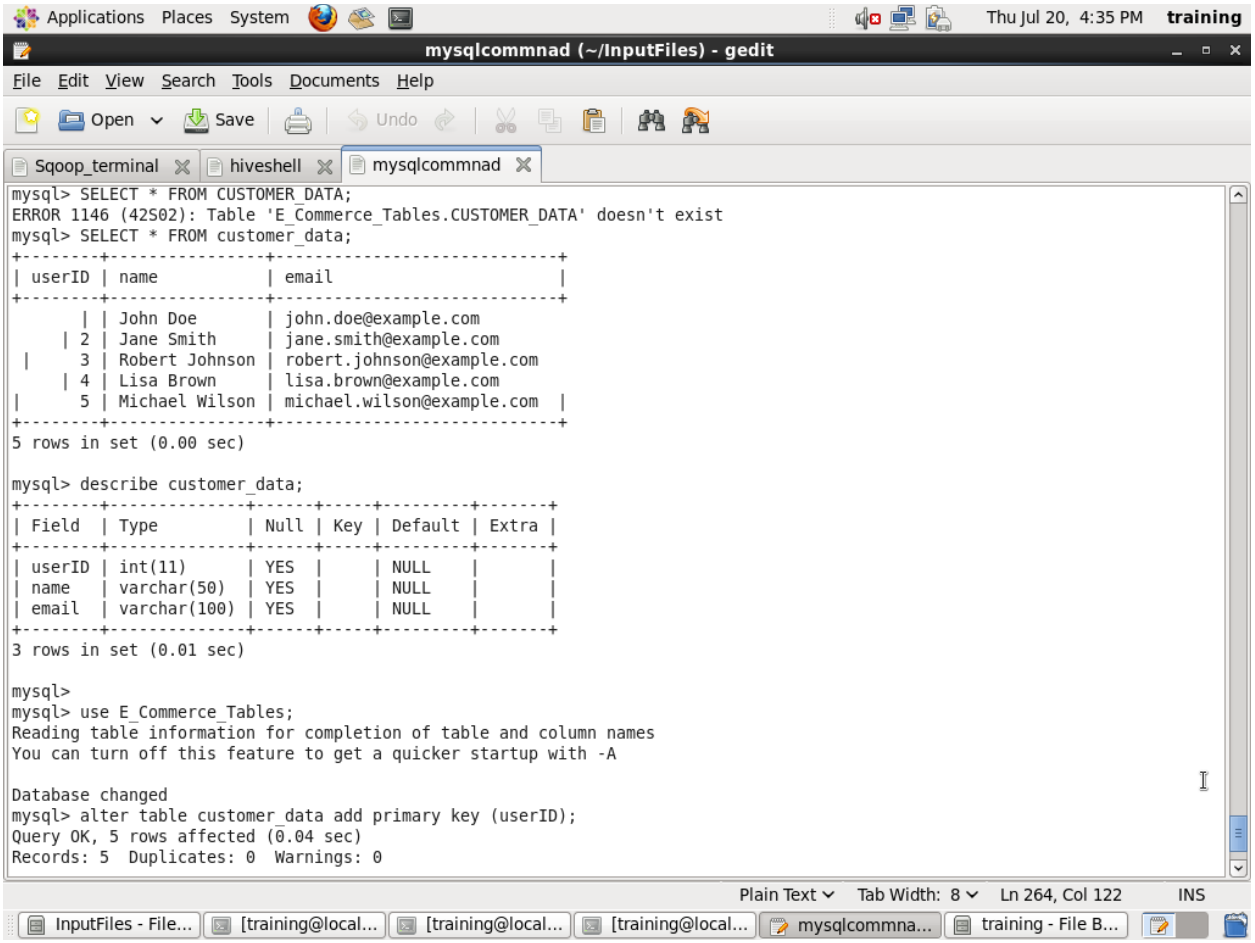

Applications Places System
Thu Jul 20, 4:34 PM training
mysqlcommnad (~/InputFiles) - gedit
File Edit View Search Tools Documents Help
Open Save Undo
Sqoop_terminal hiveshell mysqlcommnad
Plain Text Tab Width: 8 Ln 264, Col 122 INS
InputFiles - File... [training@local... [training@local... [training@local... mysqlcommna... training - File B...

```
mysql> delete from customer_data;
Query OK, 10 rows affected (0.00 sec)

mysql> LOAD DATA LOCAL INFILE '/home/training/InputFiles/customer.csv' INTO TABLE customer_data FIELDS TERMINATED BY ','
IGNORE 1 LINES;
Query OK, 5 rows affected (0.00 sec)
Records: 5  Deleted: 0  Skipped: 0  Warnings: 0

mysql> select userID from customer_data;
+--------+
| userID |
+--------+
| 1      |
| 2      |
| 3      |
| 4      |
| 5      |
+--------+
5 rows in set (0.00 sec)

mysql> LOAD DATA LOCAL INFILE '/home/training/InputFiles/purchase.csv' INTO TABLE purchase_data FIELDS TERMINATED BY ','
IGNORE 1 LINES;
Query OK, 5 rows affected (0.00 sec)
Records: 5  Deleted: 0  Skipped: 0  Warnings: 0

mysql> select * from purchase_data;
+--------+---------------------+--------+
| userID | timestamp           | amount |
+--------+---------------------+--------+
|      1 | 2023-01-01 10:05:00 | 100.00 |
|      2 | 2023-01-01 10:08:00 | 150.00 |
|      3 | 2023-01-01 10:09:00 | 200.00 |
|      4 | 2023-01-01 10:13:00 | 120.00 |
```


Applications Places System
Thu Jul 20, 4:34 PM training
mysqlcommnad (~/InputFiles) - gedit
File Edit View Search Tools Documents Help
Open Save Undo
Sqoop_terminal hiveshell mysqlcommnad
Plain Text Tab Width: 8 Ln 264, Col 122 INS
InputFiles - File... [training@local... [training@local... [training@local... mysqlcommna... training - File B...

```
+--------+
5 rows in set (0.00 sec)

mysql> LOAD DATA LOCAL INFILE '/home/training/InputFiles/purchase.csv' INTO TABLE purchase_data FIELDS TERMINATED BY ','
IGNORE 1 LINES;
Query OK, 5 rows affected (0.00 sec)
Records: 5  Deleted: 0  Skipped: 0  Warnings: 0

mysql> select * from purchase_data;
+--------+---------------------+--------+
| userID | timestamp           | amount |
+--------+---------------------+--------+
|      1 | 2023-01-01 10:05:00 | 100.00 |
|      2 | 2023-01-01 10:08:00 | 150.00 |
|      3 | 2023-01-01 10:09:00 | 200.00 |
|      4 | 2023-01-01 10:13:00 | 120.00 |
```

```
|      5 | 2023-01-01 10:17:00 |  80.00 |
+--------+---------------------+--------+
5 rows in set (0.00 sec)

mysql> describe purchase_data;
+-----------+---------------+------+-----+-------------------+-----------------------------+
| Field     | Type          | Null | Key | Default           | Extra                       |
+-----------+---------------+------+-----+-------------------+-----------------------------+
| userID    | int(11)       | YES  |     | NULL              |                             |
| timestamp | timestamp     | NO   |     | CURRENT_TIMESTAMP | on update CURRENT_TIMESTAMP |
| amount    | decimal(10,2) | YES  |     | NULL              |                             |
+-----------+---------------+------+-----+-------------------+-----------------------------+
3 rows in set (0.00 sec)

mysql> alter table clickStream_data modify column userID INT(11);
Query OK, 13 rows affected (0.01 sec)
Records: 13  Duplicates: 0  Warnings: 0
```



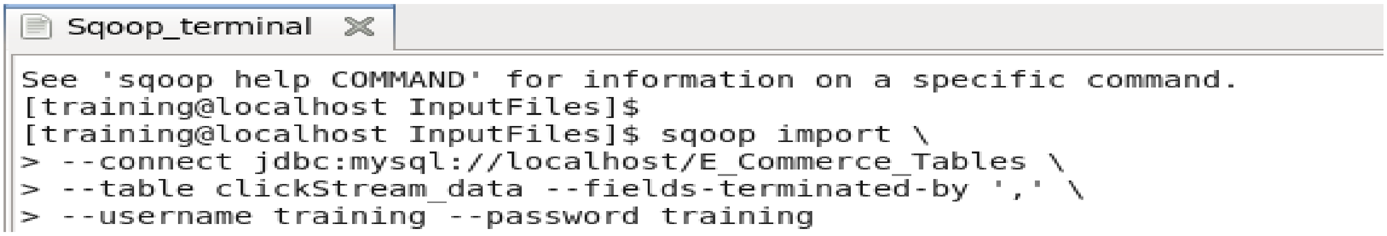
```
mysql> SELECT * FROM CUSTOMER_DATA;
ERROR 1146 (42S02): Table 'E_Commerce_Tables.CUSTOMER_DATA' doesn't exist
mysql> SELECT * FROM customer_data;
+--------+----------------+-----------------------------+
| userID | name           | email                       |
+--------+----------------+-----------------------------+
       | | John Doe       | john.doe@example.com
     | 2 | Jane Smith     | jane.smith@example.com
 |     3 | Robert Johnson | robert.johnson@example.com
     | 4 | Lisa Brown     | lisa.brown@example.com
|      5 | Michael Wilson | michael.wilson@example.com  |
+--------+----------------+-----------------------------+
5 rows in set (0.00 sec)

mysql> describe customer_data;
+--------+--------------+------+-----+---------+-------+
| Field  | Type         | Null | Key | Default | Extra |
+--------+--------------+------+-----+---------+-------+
| userID | int(11)      | YES  |     | NULL    |       |
| name   | varchar(50)  | YES  |     | NULL    |       |
| email  | varchar(100) | YES  |     | NULL    |       |
+--------+--------------+------+-----+---------+-------+
3 rows in set (0.01 sec)

mysql>
mysql> use E_Commerce_Tables;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> alter table customer_data add primary key (userID);
Query OK, 5 rows affected (0.04 sec)
Records: 5  Duplicates: 0  Warnings: 0
```

SQOOP COMMANDS FOR IMPORT THE MYSQL TABLES INTO HDFS

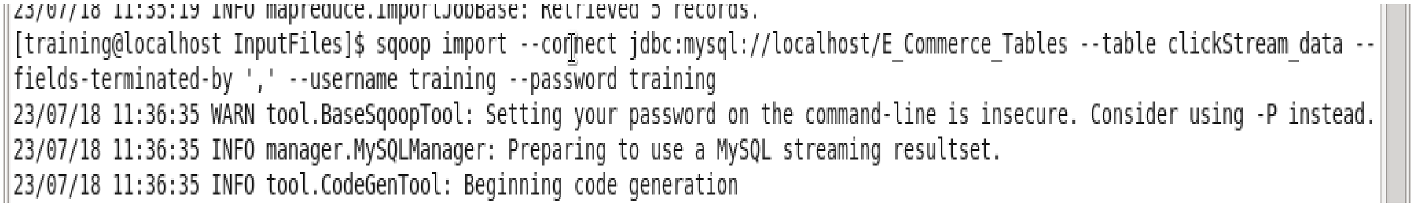
```
[training@localhost InputFiles]$ sqoop import \
> --connect jdbc:mysql://localhost/E_Commerce_Tables \
> --table clickStream_data --fields-terminated-by ',' \
> --username training --password training
```

```
[training@localhost InputFiles]$ sqoop import --connect jdbc:mysql://localhost/E_Commerce_Tables --table purchase_data --fields-terminated-by ',' --username training --password training
```

[training@localhost InputFiles]$ sqoop import --connect jdbc:mysql://localhost/E_Commerce_Tables --table clickStream_data --fields-terminated-by ',' --username training --password training

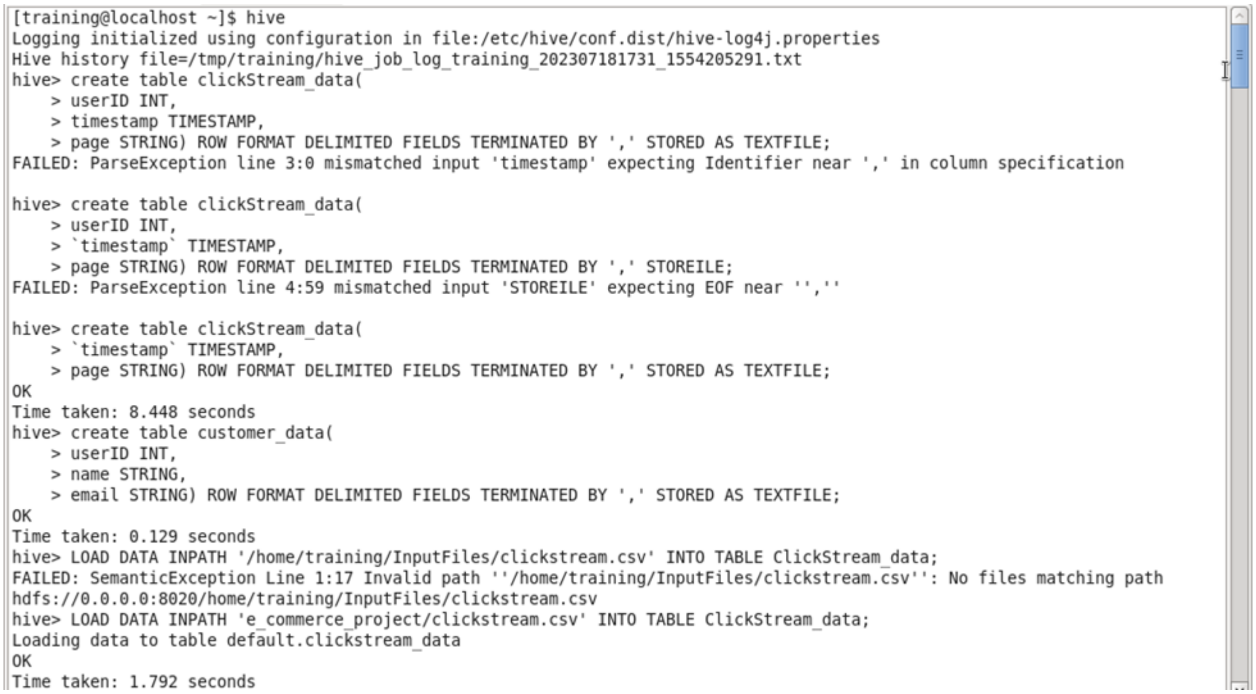


```
See 'sqoop help COMMAND' for information on a specific command.
[training@localhost InputFiles]$
[training@localhost InputFiles]$ sqoop import \
> --connect jdbc:mysql://localhost/E_Commerce_Tables \
> --table clickStream_data --fields-terminated-by ',' \
> --username training --password training
```

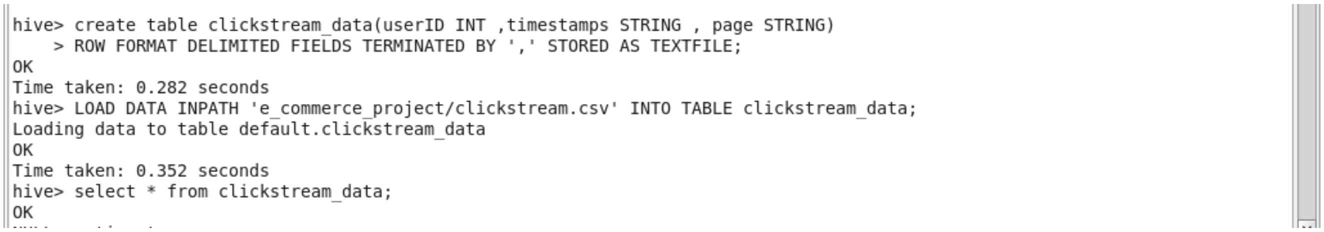


```
23/07/18 11:35:19 INFO mapreduce.ImportJobBase: Retrieved 5 records.
[training@localhost InputFiles]$ sqoop import --connect jdbc:mysql://localhost/E_Commerce_Tables --table clickStream_data --
fields-terminated-by ',' --username training --password training
23/07/18 11:36:35 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
23/07/18 11:36:35 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/07/18 11:36:35 INFO tool.CodeGenTool: Beginning code generation
```

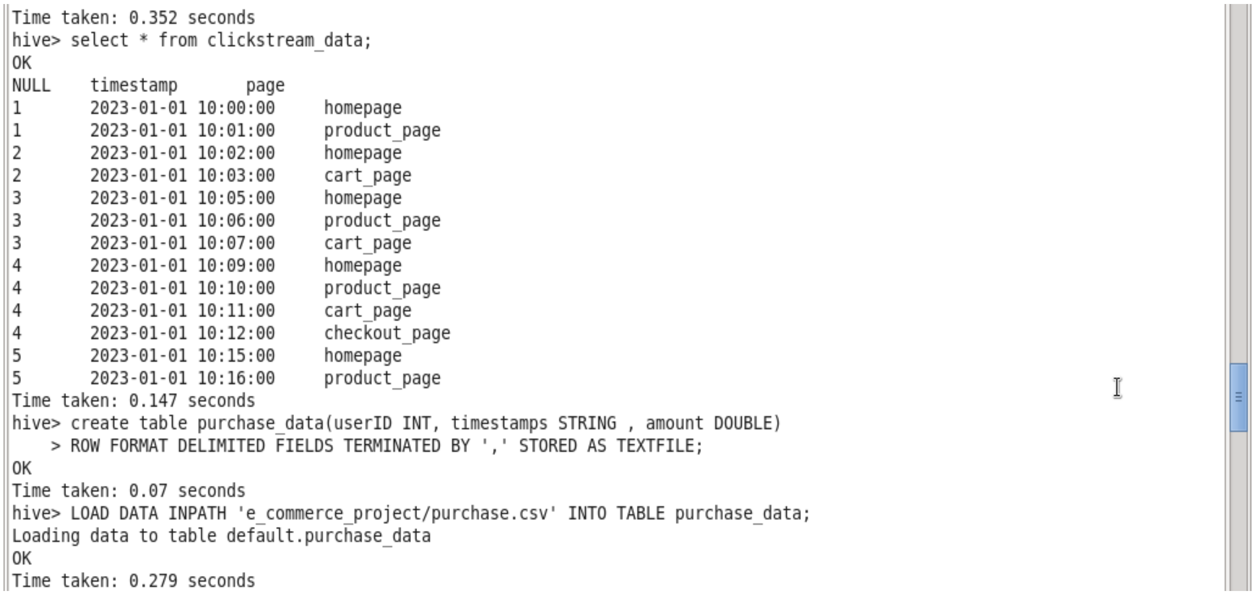
After this I move this table in hive with the help of hive shell



```
[training@localhost ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
Hive history file=/tmp/training/hive_job_log_training_202307181731_1554205291.txt
hive> create table clickStream_data(
    > userID INT,
    > timestamp TIMESTAMP,
    > page STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
FAILED: ParseException line 3:0 mismatched input 'timestamp' expecting Identifier near ',' in column specification

hive> create table clickStream_data(
    > userID INT,
    > `timestamp` TIMESTAMP,
    > page STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STOREILE;
FAILED: ParseException line 4:59 mismatched input 'STOREILE' expecting EOF near '','''

hive> create table clickStream_data(
    > `timestamp` TIMESTAMP,
    > page STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
OK
Time taken: 8.448 seconds
hive> create table customer_data(
    > userID INT,
    > name STRING,
    > email STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
OK
Time taken: 0.129 seconds
hive> LOAD DATA INPATH '/home/training/InputFiles/clickstream.csv' INTO TABLE ClickStream_data;
FAILED: SemanticException Line 1:17 Invalid path ''/home/training/InputFiles/clickstream.csv'': No files matching path
hdfs://0.0.0.0:8020/home/training/InputFiles/clickstream.csv
hive> LOAD DATA INPATH 'e_commerce_project/clickstream.csv' INTO TABLE ClickStream_data;
Loading data to table default.clickstream_data
OK
Time taken: 1.792 seconds
```

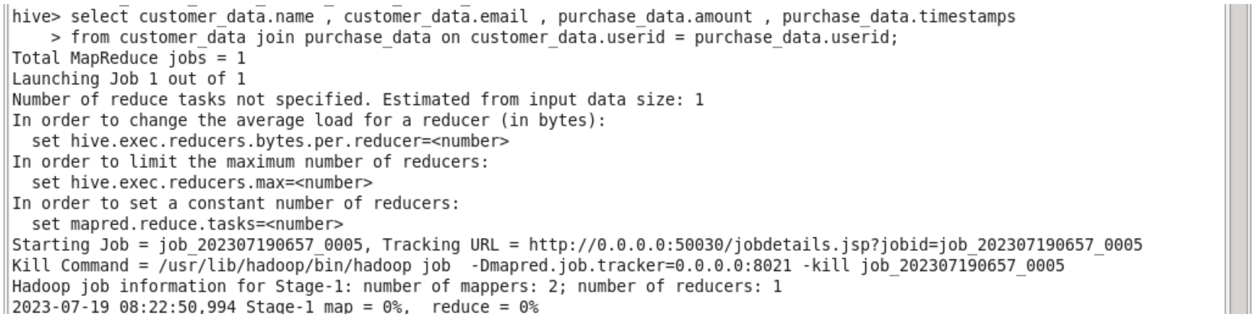


```
hive> create table clickstream_data(userID INT ,timestamps STRING , page STRING)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
OK
Time taken: 0.282 seconds
hive> LOAD DATA INPATH 'e_commerce_project/clickstream.csv' INTO TABLE clickstream_data;
Loading data to table default.clickstream_data
OK
Time taken: 0.352 seconds
hive> select * from clickstream_data;
OK
```

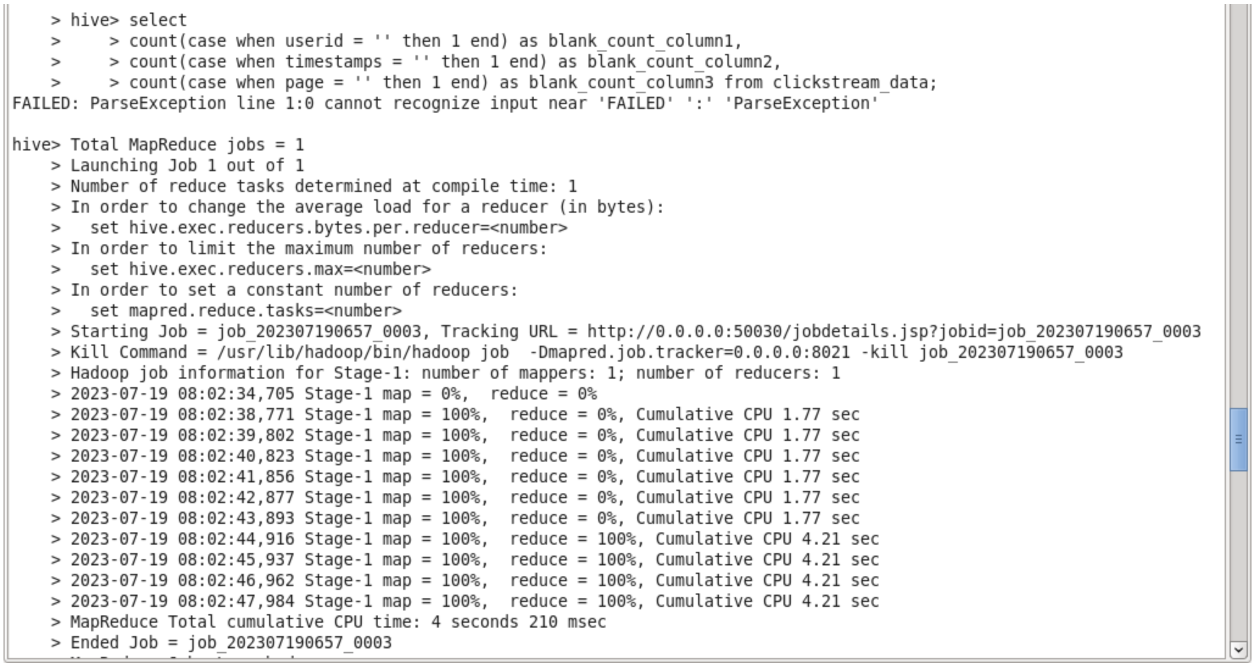


```
Time taken: 0.352 seconds
hive> select * from clickstream_data;
OK
NULL    timestamp       page
1       2023-01-01 10:00:00     homepage
1       2023-01-01 10:01:00     product_page
2       2023-01-01 10:02:00     homepage
2       2023-01-01 10:03:00     cart_page
3       2023-01-01 10:05:00     homepage
3       2023-01-01 10:06:00     product_page
3       2023-01-01 10:07:00     cart_page
4       2023-01-01 10:09:00     homepage
4       2023-01-01 10:10:00     product_page
4       2023-01-01 10:11:00     cart_page
4       2023-01-01 10:12:00     checkout_page
5       2023-01-01 10:15:00     homepage
5       2023-01-01 10:16:00     product_page
Time taken: 0.147 seconds
hive> create table purchase_data(userID INT, timestamps STRING , amount DOUBLE)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
OK
Time taken: 0.07 seconds
hive> LOAD DATA INPATH 'e_commerce_project/purchase.csv' INTO TABLE purchase_data;
Loading data to table default.purchase_data
OK
Time taken: 0.279 seconds
```

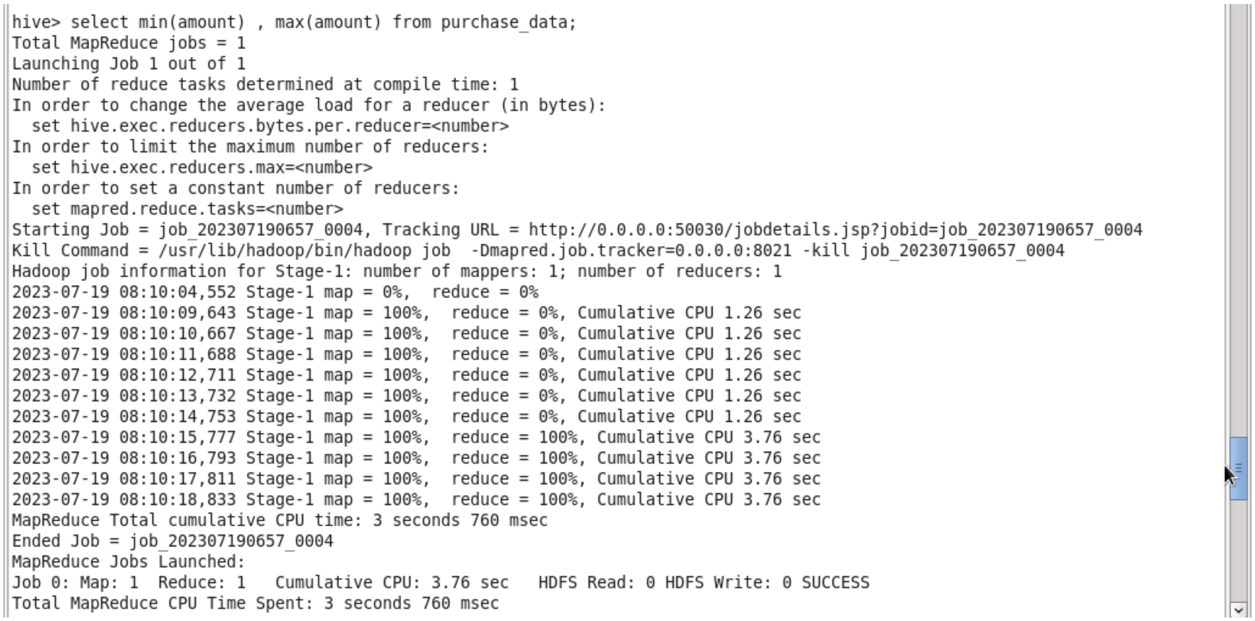
Some hive Query for data enrichment and cleansing



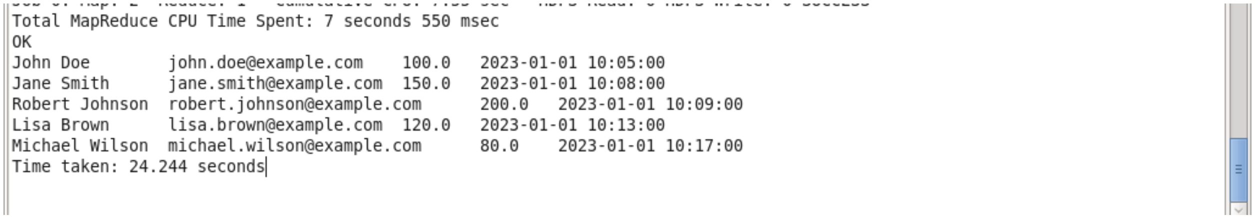
```
hive> select customer_data.name , customer_data.email , purchase_data.amount , purchase_data.timestamps
    > from customer_data join purchase_data on customer_data.userid = purchase_data.userid;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 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 mapred.reduce.tasks=<number>
Starting Job = job_202307190657_0005, Tracking URL = http://0.0.0.0:50030/jobdetails.jsp?jobid=job_202307190657_0005
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=0.0.0.0:8021 -kill job_202307190657_0005
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1
2023-07-19 08:22:50,994 Stage-1 map = 0%,  reduce = 0%
```



```
    > hive> select
    >     > count(case when userid = '' then 1 end) as blank_count_column1,
    >     > count(case when timestamps = '' then 1 end) as blank_count_column2,
    >     > count(case when page = '' then 1 end) as blank_count_column3 from clickstream_data;
FAILED: ParseException line 1:0 cannot recognize input near 'FAILED' ':' 'ParseException'

hive> Total MapReduce 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 mapred.reduce.tasks=<number>
    > Starting Job = job_202307190657_0003, Tracking URL = http://0.0.0.0:50030/jobdetails.jsp?jobid=job_202307190657_0003
    > Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=0.0.0.0:8021 -kill job_202307190657_0003
    > Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
    > 2023-07-19 08:02:34,705 Stage-1 map = 0%,  reduce = 0%
    > 2023-07-19 08:02:38,771 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.77 sec
    > 2023-07-19 08:02:39,802 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.77 sec
    > 2023-07-19 08:02:40,823 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.77 sec
    > 2023-07-19 08:02:41,856 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.77 sec
    > 2023-07-19 08:02:42,877 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.77 sec
    > 2023-07-19 08:02:43,893 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.77 sec
    > 2023-07-19 08:02:44,916 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.21 sec
    > 2023-07-19 08:02:45,937 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.21 sec
    > 2023-07-19 08:02:46,962 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.21 sec
    > 2023-07-19 08:02:47,984 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.21 sec
    > MapReduce Total cumulative CPU time: 4 seconds 210 msec
    > Ended Job = job_202307190657_0003
```

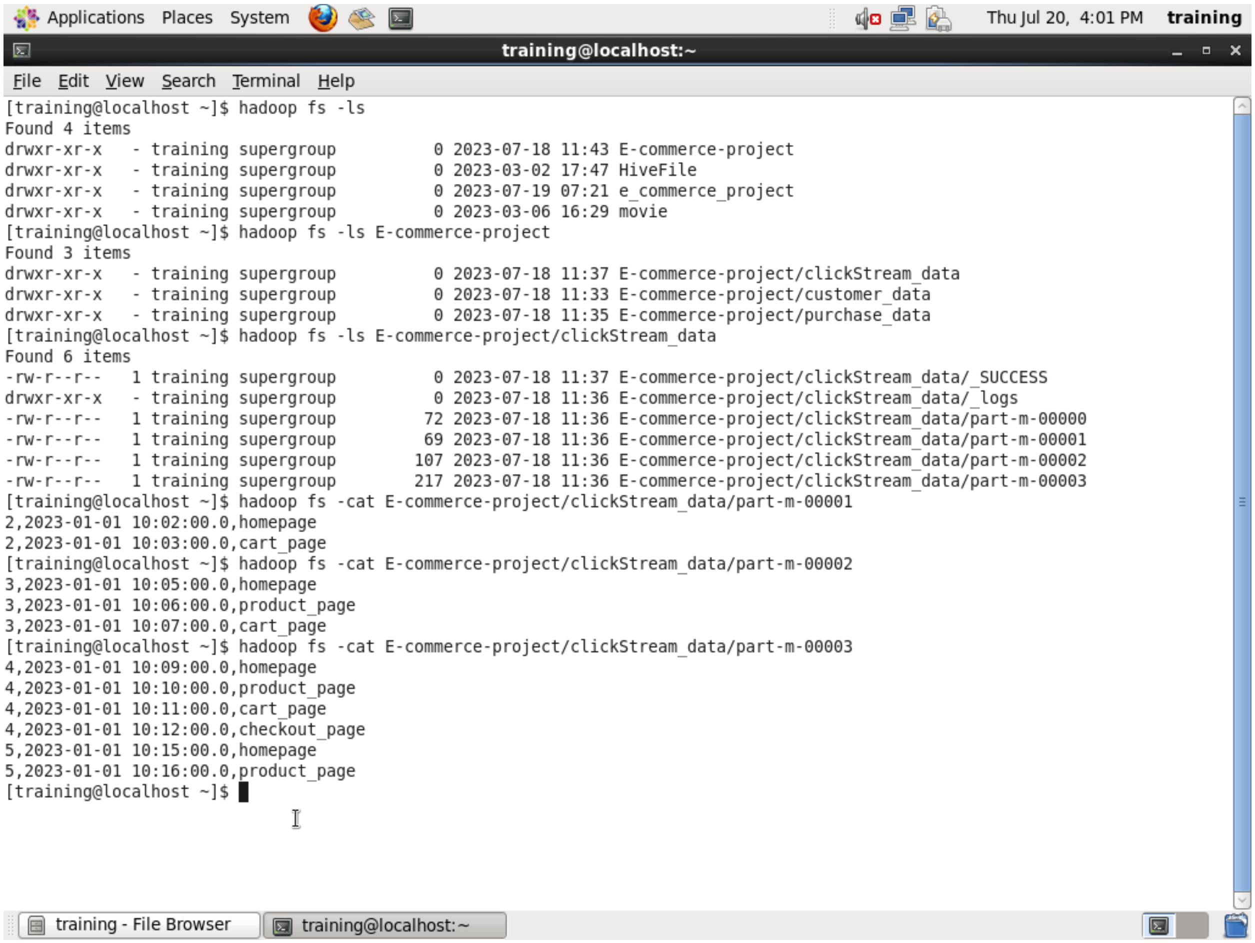


```
hive> select min(amount) , max(amount) from purchase_data;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_202307190657_0004, Tracking URL = http://0.0.0.0:50030/jobdetails.jsp?jobid=job_202307190657_0004
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=0.0.0.0:8021 -kill job_202307190657_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-07-19 08:10:04,552 Stage-1 map = 0%,  reduce = 0%
2023-07-19 08:10:09,643 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.26 sec
2023-07-19 08:10:10,667 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.26 sec
2023-07-19 08:10:11,688 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.26 sec
2023-07-19 08:10:12,711 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.26 sec
2023-07-19 08:10:13,732 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.26 sec
2023-07-19 08:10:14,753 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.26 sec
2023-07-19 08:10:15,777 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.76 sec
2023-07-19 08:10:16,793 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.76 sec
2023-07-19 08:10:17,811 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.76 sec
2023-07-19 08:10:18,833 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.76 sec
MapReduce Total cumulative CPU time: 3 seconds 760 msec
Ended Job = job_202307190657_0004
MapReduce Jobs Launched:
Job 0: Map: 1  Reduce: 1   Cumulative CPU: 3.76 sec   HDFS Read: 0 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 760 msec
```



```
Total MapReduce CPU Time Spent: 7 seconds 550 msec
OK
John Doe        john.doe@example.com    100.0   2023-01-01 10:05:00
Jane Smith      jane.smith@example.com  150.0   2023-01-01 10:08:00
Robert Johnson  robert.johnson@example.com      200.0   2023-01-01 10:09:00
Lisa Brown      lisa.brown@example.com  120.0   2023-01-01 10:13:00
Michael Wilson  michael.wilson@example.com      80.0    2023-01-01 10:17:00
Time taken: 24.244 seconds
```

Hadoop terminal for import data from mysql into hdfs


Applications Places System
Thu Jul 20, 4:01 PM training
training@localhost:~
File Edit View Search Terminal Help
training - File Browser
training@localhost:~

```
[training@localhost ~]$ hadoop fs -ls
Found 4 items
drwxr-xr-x   - training supergroup          0 2023-07-18 11:43 E-commerce-project
drwxr-xr-x   - training supergroup          0 2023-03-02 17:47 HiveFile
drwxr-xr-x   - training supergroup          0 2023-07-19 07:21 e_commerce_project
drwxr-xr-x   - training supergroup          0 2023-03-06 16:29 movie
[training@localhost ~]$ hadoop fs -ls E-commerce-project
Found 3 items
drwxr-xr-x   - training supergroup          0 2023-07-18 11:37 E-commerce-project/clickStream_data
drwxr-xr-x   - training supergroup          0 2023-07-18 11:33 E-commerce-project/customer_data
drwxr-xr-x   - training supergroup          0 2023-07-18 11:35 E-commerce-project/purchase_data
[training@localhost ~]$ hadoop fs -ls E-commerce-project/clickStream_data
Found 6 items
-rw-r--r--   1 training supergroup          0 2023-07-18 11:37 E-commerce-project/clickStream_data/_SUCCESS
drwxr-xr-x   - training supergroup          0 2023-07-18 11:36 E-commerce-project/clickStream_data/_logs
-rw-r--r--   1 training supergroup         72 2023-07-18 11:36 E-commerce-project/clickStream_data/part-m-00000
-rw-r--r--   1 training supergroup         69 2023-07-18 11:36 E-commerce-project/clickStream_data/part-m-00001
-rw-r--r--   1 training supergroup        107 2023-07-18 11:36 E-commerce-project/clickStream_data/part-m-00002
-rw-r--r--   1 training supergroup        217 2023-07-18 11:36 E-commerce-project/clickStream_data/part-m-00003
[training@localhost ~]$ hadoop fs -cat E-commerce-project/clickStream_data/part-m-00001
2,2023-01-01 10:02:00.0,homepage
2,2023-01-01 10:03:00.0,cart_page
[training@localhost ~]$ hadoop fs -cat E-commerce-project/clickStream_data/part-m-00002
3,2023-01-01 10:05:00.0,homepage
3,2023-01-01 10:06:00.0,product_page
3,2023-01-01 10:07:00.0,cart_page
[training@localhost ~]$ hadoop fs -cat E-commerce-project/clickStream_data/part-m-00003
4,2023-01-01 10:09:00.0,homepage
4,2023-01-01 10:10:00.0,product_page
4,2023-01-01 10:11:00.0,cart_page
4,2023-01-01 10:12:00.0,checkout_page
5,2023-01-01 10:15:00.0,homepage
5,2023-01-01 10:16:00.0,product_page
[training@localhost ~]$ 
```