Kalyan Big Data Projects – Project 8
How To Stream CSV Data Into Hive Using Apache Flume

Pre-Requisites of Flume + Hive Project:

hadoop-2.6.0
flume-1.6.0
hive-1.2.1
java-1.7

NOTE: Make sure that install all the above components

Flume + Hive Project Download Links:

`hadoop-2.6.0.tar.gz` ==> link
(https://archive.apache.org/dist/hadoop/core/hadoop-2.6.0/hadoop-2.6.0.tar.gz)

`apache-flume-1.6.0-bin.tar.gz` ==> link
(https://archive.apache.org/dist/flume/1.6.0/apache-flume-1.6.0-bin.tar.gz)

`apache-hive-1.2.1-bin.tar.gz` ==> link
(http://mirror.fibergrid.in/apache/hive/hive-1.2.1/apache-hive-1.2.1-bin.tar.gz)

`kalyan-csv-hive-agent.conf` ==> link
(https://github.com/kalyanhadooptraining/kalyan-bigdata-realtime-projects/blob/master/flume/project8-hive-csv/kalyan-csv-hive-agent.conf)

`kalyan-bigdata-examples.jar` ==> link
(https://github.com/kalyanhadooptraining/kalyan-bigdata-realtime-projects/blob/master/kalyan/kalyan-bigdata-examples.jar)

Learnings of this Project:

- We will learn Flume Configurations and Commands
- Flume Agent
  1. Source (Exec Source)
  2. Channel (Memory Channel)
  3. Sink (Hive Sink)
- Major project in Real Time `Product Log Analysis`
  1. We are extracting the data from server logs
  2. This data will be useful to do analysis on product views
  3. CSV is the output format
- We can use hive to analyze this data

---

1. create "**kalyan-csv-hive-agent.conf**" file with below content

```
agent.sources = EXEC
agent.sinks = HIVE
agent.channels = MemChannel

agent.sources.EXEC.type = exec
agent.sources.EXEC.command = tail -F /tmp/users.csv
agent.sources.EXEC.channels = MemChannel

agent.sinks.HIVE.type = hive
agent.sinks.HIVE.hive.metastore = thrift://localhost:9083
agent.sinks.HIVE.hive.database = kalyan
agent.sinks.HIVE.hive.table = users1
agent.sinks.HIVE.serializer = DELIMITED
agent.sinks.HIVE.serializer.delimiter = ","
agent.sinks.HIVE.serializer.fieldnames=userid,username,password,email,country,state,city,dt
agent.sinks.HIVE.channel = MemChannel

agent.channels.MemChannel.type = memory
agent.channels.MemChannel.capacity = 1000
agent.channels.MemChannel.transactionCapacity = 100
```

2. Copy "**kalyan-csv-hive-agent.conf**" file into "**$FUME_HOME/conf**" folder

3. Copy "**kalyan-bigdata-examples.jar**" file into "**$FLUME_HOME/lib**" folder

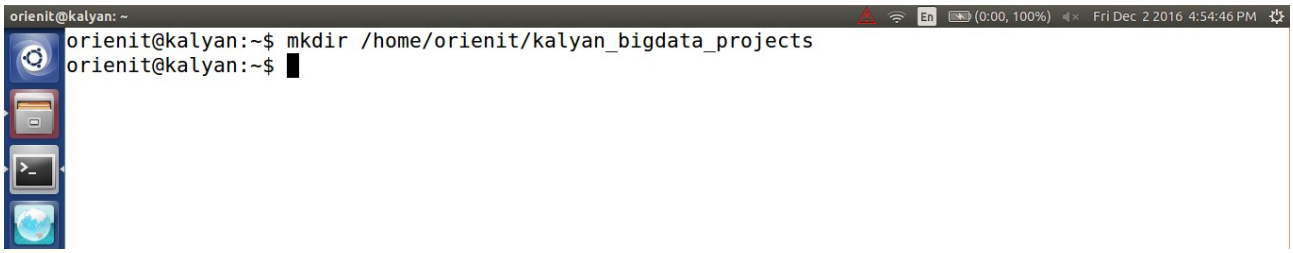
4. Generate Large Amount of Sample CSV data follow this article.

(http://kalyanbigdatatraining.blogspot.com/2016/12/how-to-generate-large-amount-of-sample.html)

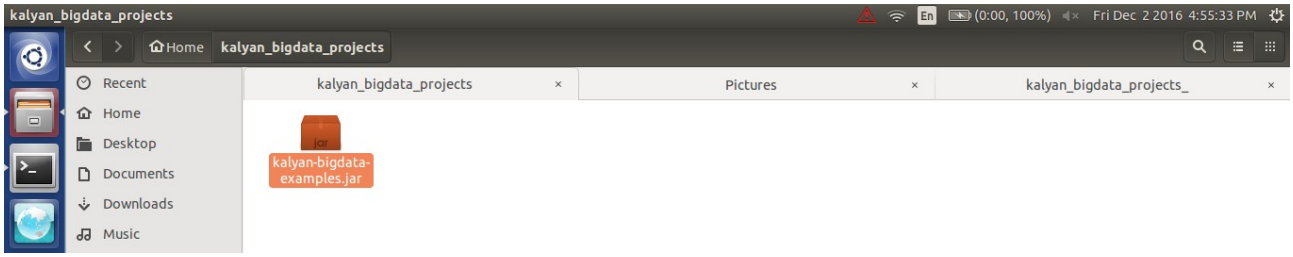
5. Follow below steps...

i) Create '**kalyan_bigdata_projects**' folder in **user home** (i.e **/home/orienit**)

Command: *mkdir /home/orienit/kalyan_bigdata_projects*

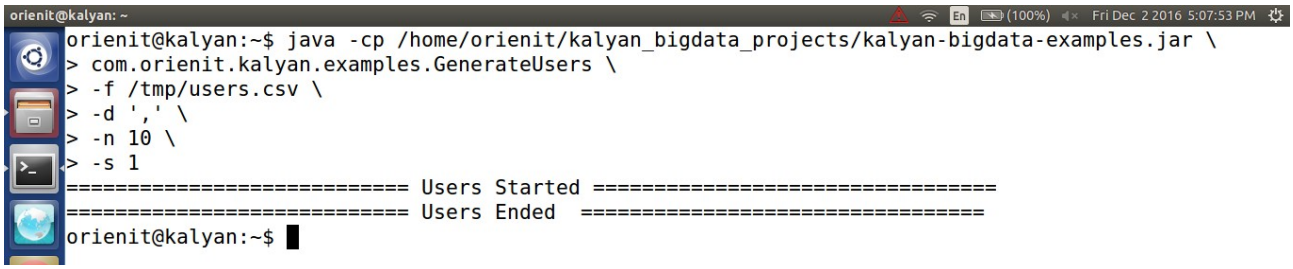


ii) Copy '**kalyan-bigdata-examples.jar**' jar file into '**/home/orienit/kalyan_bigdata_projects**' folder



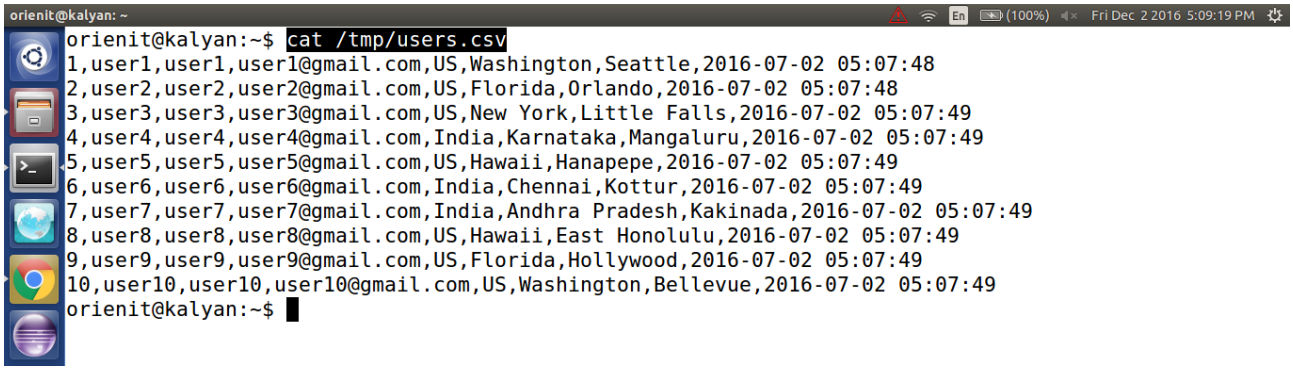
iii) Execute Below Command to Generate Sample CSV data with 100 lines. Increase this number to get more data ...

```
java -cp /home/orienit/kalyan_bigdata_projects/kalyan-bigdata-examples.jar \
com.orienit.kalyan.examples.GenerateUsers \
-f /tmp/users.csv \
-d ',' \
-n 100 \
-s 1
```



6. Verify the Sample CSV data in Console, using below command

cat /tmp/users.csv



7. To work with **Flume + Hive Integration**, Follow the below steps

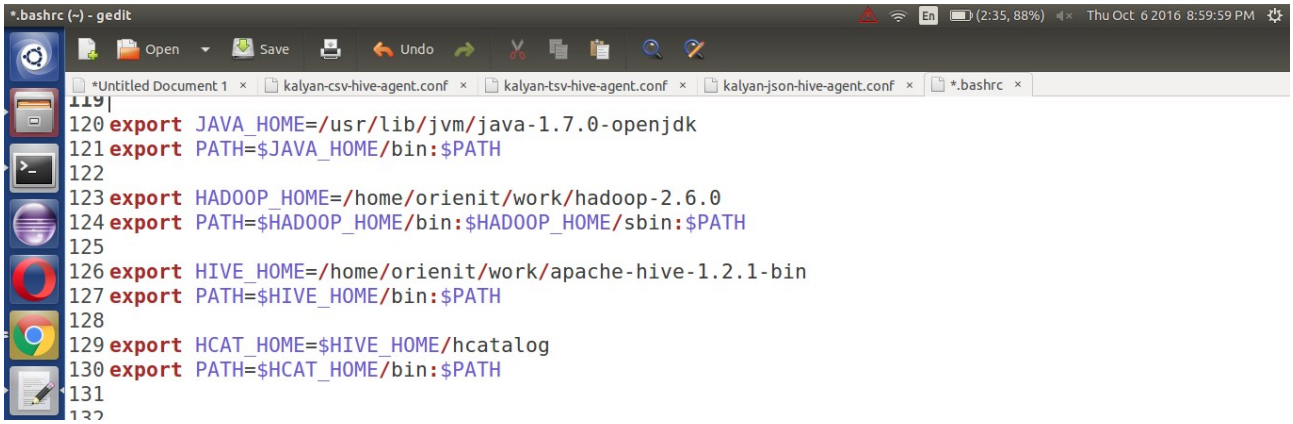
Follow this aritcle to work with below procedure.

Refer: http://kalyanbigdatatraining.blogspot.in/2016/10/how-to-work-with-acid-functionality-in.html

i) update '**~/.bashrc**' file with below changes

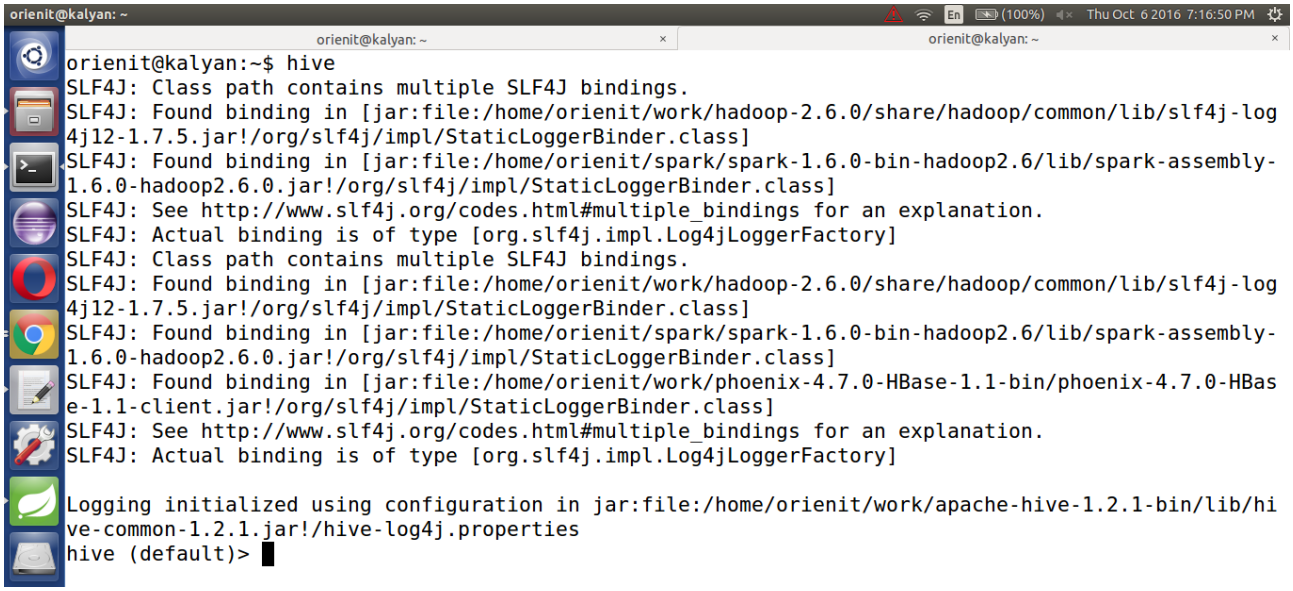
export HIVE_HOME=/home/orienit/work/apache-hive-1.2.1-bin
export PATH=$HIVE_HOME/bin:$PATH

export HCAT_HOME=$HIVE_HOME/hcatalog
export PATH=$HCAT_HOME/bin:$PATH

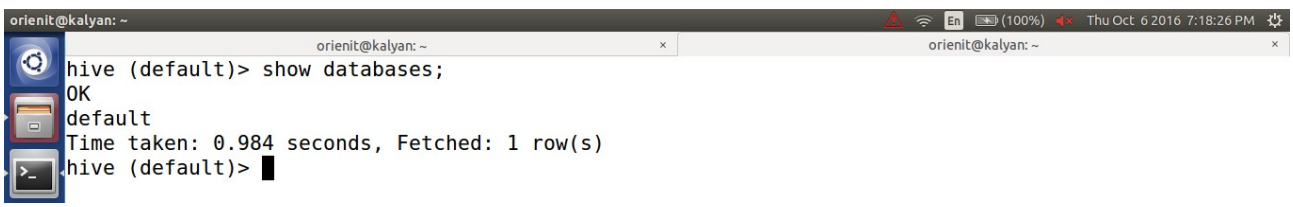


ii. reopen the Terminal

iii. start the hive using '**hive**' command.

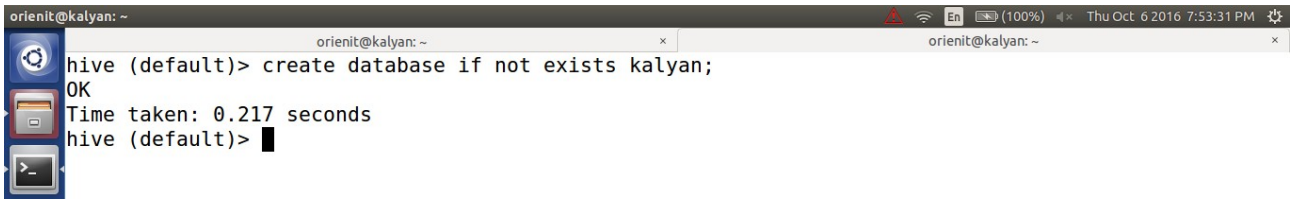


iv. list out all the databases in hive using '**show databases;**' command

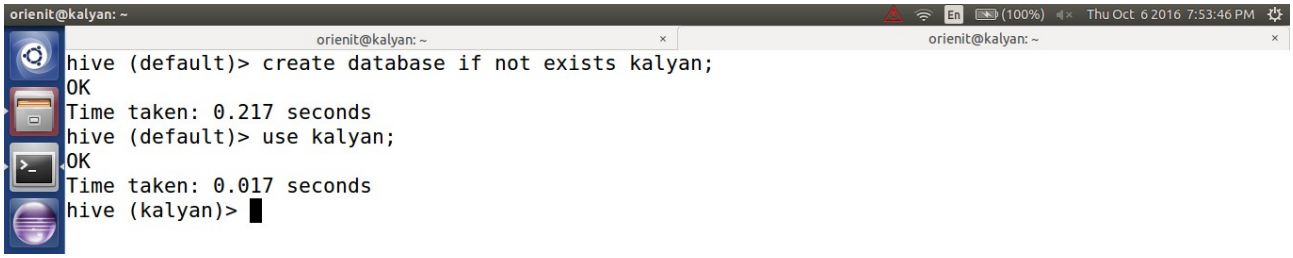


v. create a new database (**kalyan**) in hive using below command.

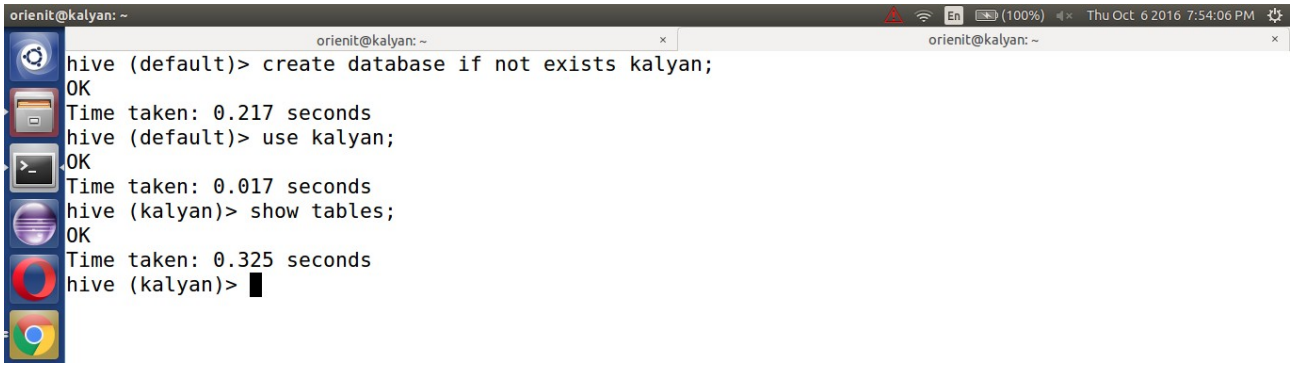
create database if not exists kalyan;



vi. use kalyan database using '**use kalyan;**' command

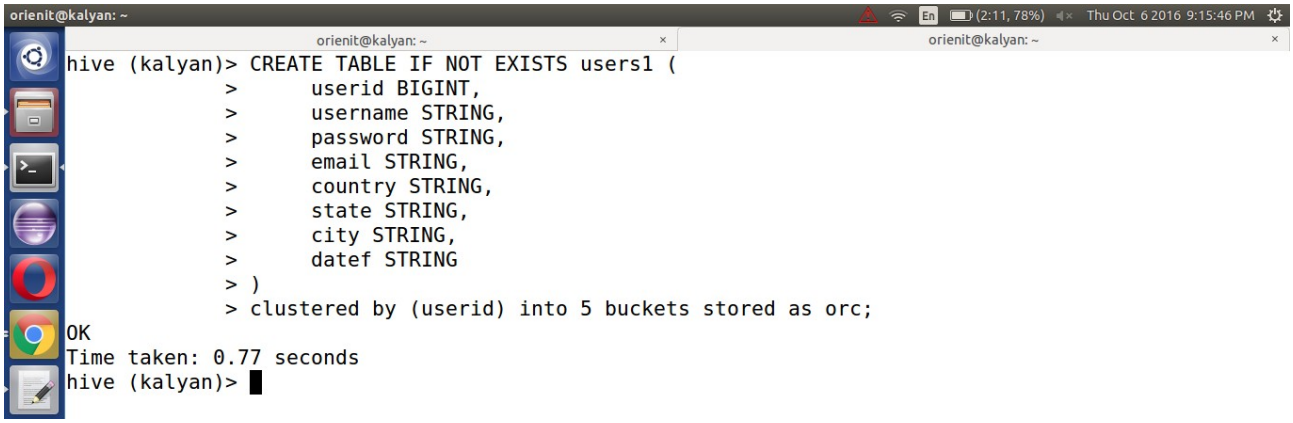


vii. list out all the tables in kalyan database using '**show tables;**' command.



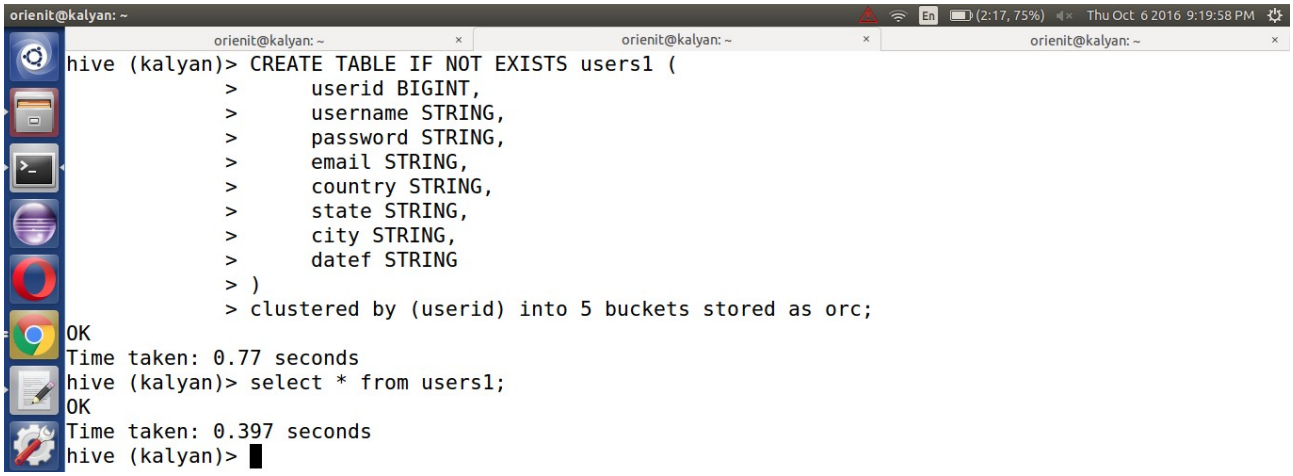
viii. create '**users1**' table in kalyan database using below command.

```
CREATE TABLE IF NOT EXISTS kalyan.users1 (
  userid BIGINT,
  username STRING,
  password STRING,
  email STRING,
  country STRING,
  state STRING,
  city STRING,
  dt STRING
)
clustered by (userid) into 5 buckets stored as orc;
```



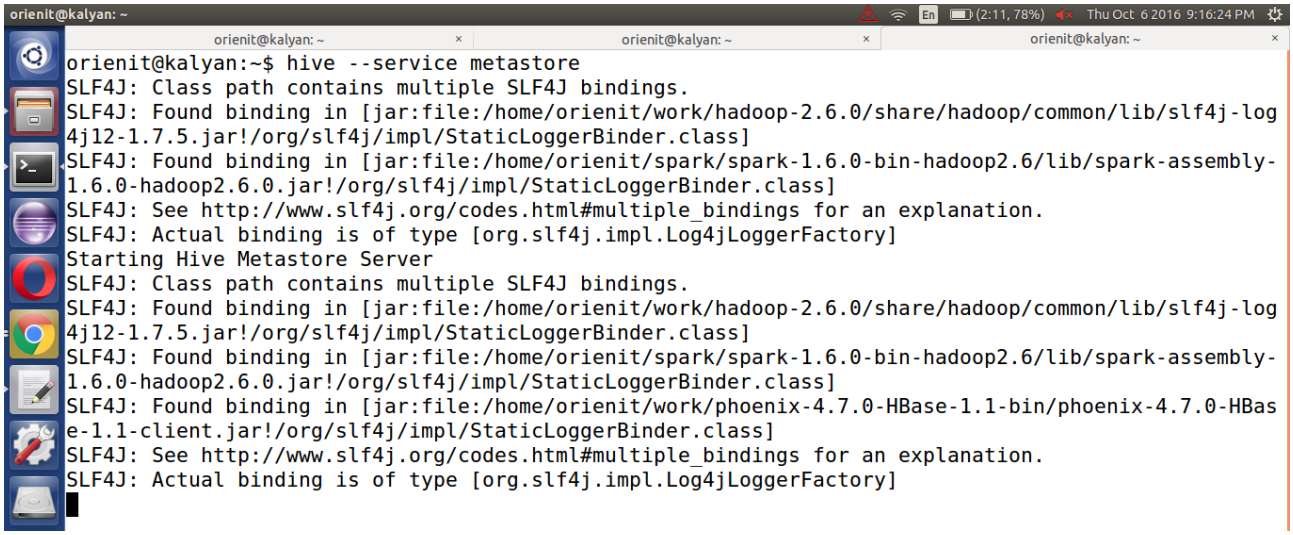
ix. Display the data from '**users1**' table using below command

select * from users1;



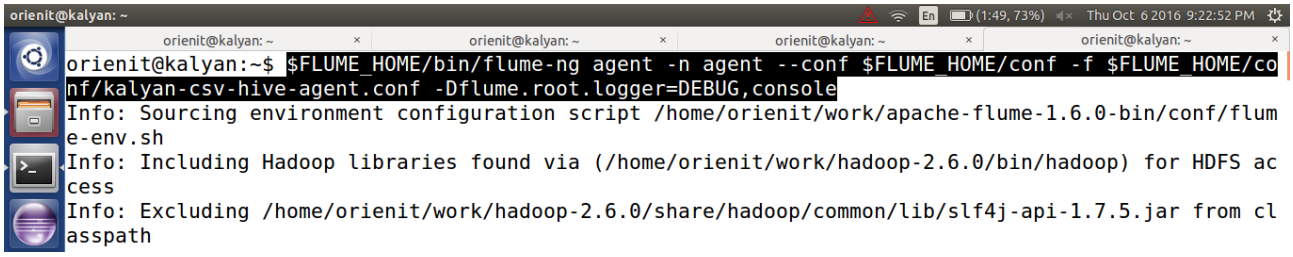
x. start the hive in external metastore db mode using below command

hive --service metastore

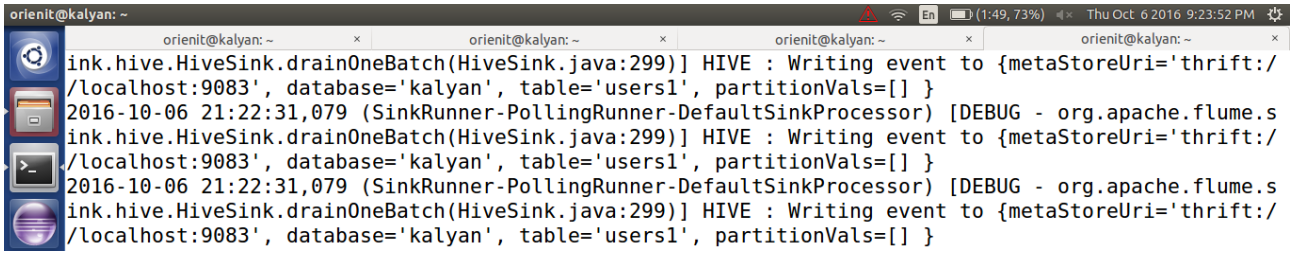


8. Execute the below command to `**Extract data from CSV data into Hive using Flume**`

$FLUME_HOME/bin/flume-ng agent -n agent --conf $FLUME_HOME/conf -f $FLUME_HOME/conf/kalyan-csv-hive-agent.conf -Dflume.root.logger=DEBUG,console



9. Verify the data in console



10. Verify the data in Hive

Execute below command to get the data from hive table '**users1**'

select * from users1;