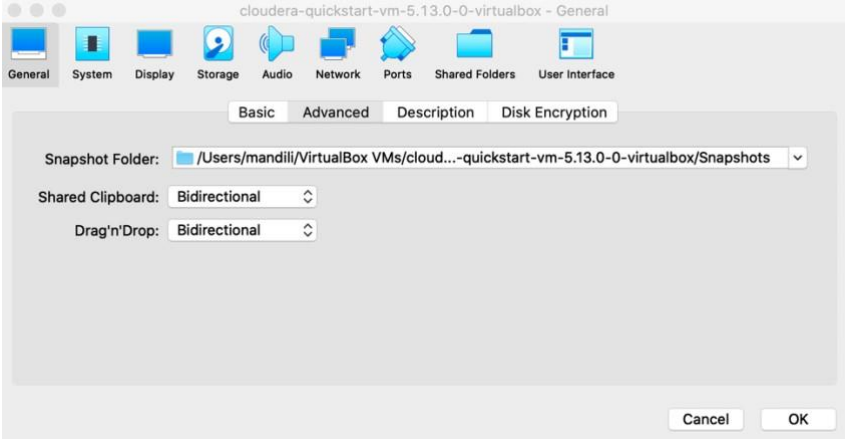
Lab 4 Hive on Cloudera (VirtualBox)

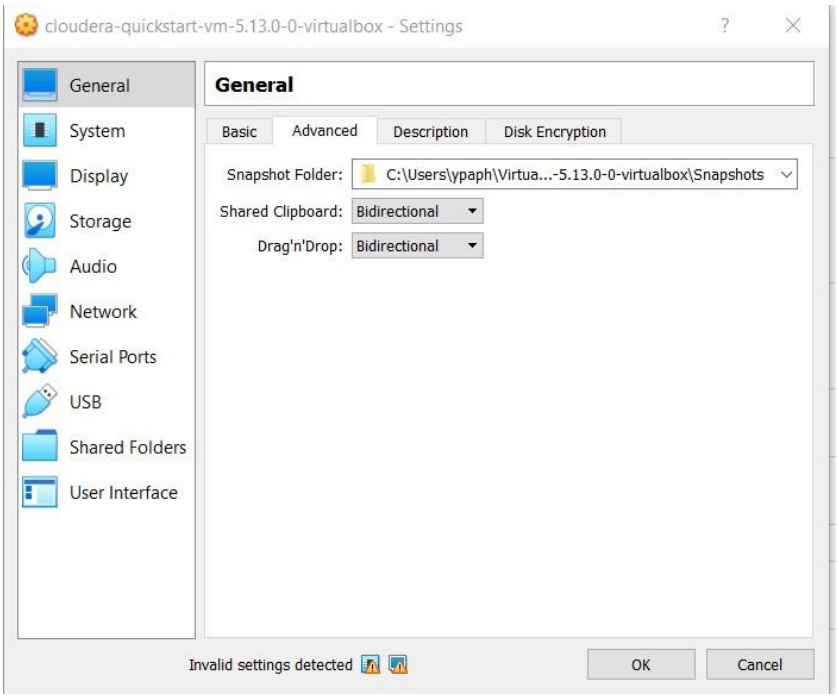
Step 1: Additional Settings

1. Open VirtualBox and change settings for your Cloudera QuickStart VM.
2. Go to Settings →General → Advanced. Change Shared Clipboard and Drag'n'Drop to Bidirectional. Now, you are able to copy text from local to VM.

   Mac view:



   Windows view:

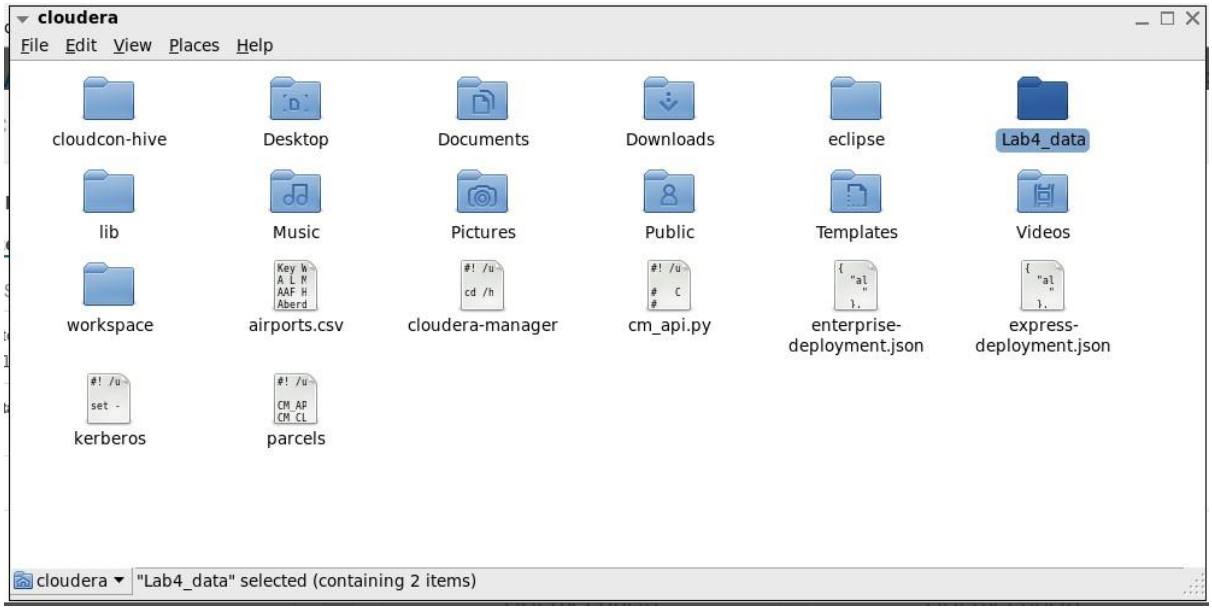


3. Start your VM and go to Device → Insert Guest Additions CD image… and follow the pop up in VM to finish the installation. Now, you are able to drag/copy files from local to VM.

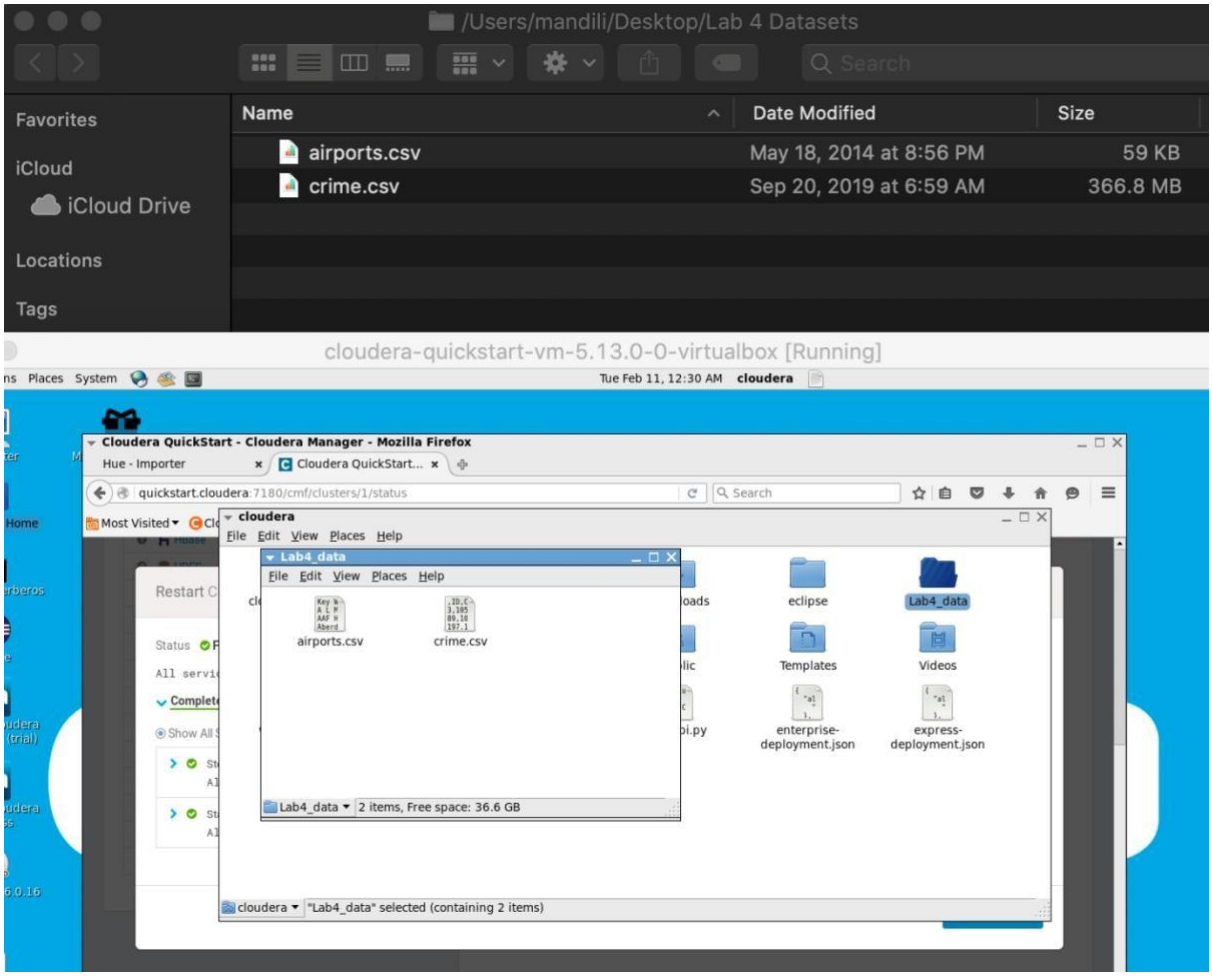
Attention Windows Users: Some of you may notice that you do not have a toolbar on the top of the VirtualBox. Do not worry, please check out this video: How to get menus back in VirtualBox:
https://www.youtube.com/watch?v=KOwCoZSsD7U

Step 2: Move files from our local laptop to VM (Option 1: Drag)

1. Start VM if you haven't done so and restart all the services if necessary.
2. Download airports.csv and crime.csv from BB.
3. We can open the local folder and the virtual machine folder, just drag the csv files from your local folder to the virtual machine folder to finish the copy.
4. I created a folder called Lab4_data under cloudera folder on VM to store the copied files from my local laptop.



5. Then, simply drag two files from my local machine to the VM and put them inside Lab4_data folder to finish the copy.

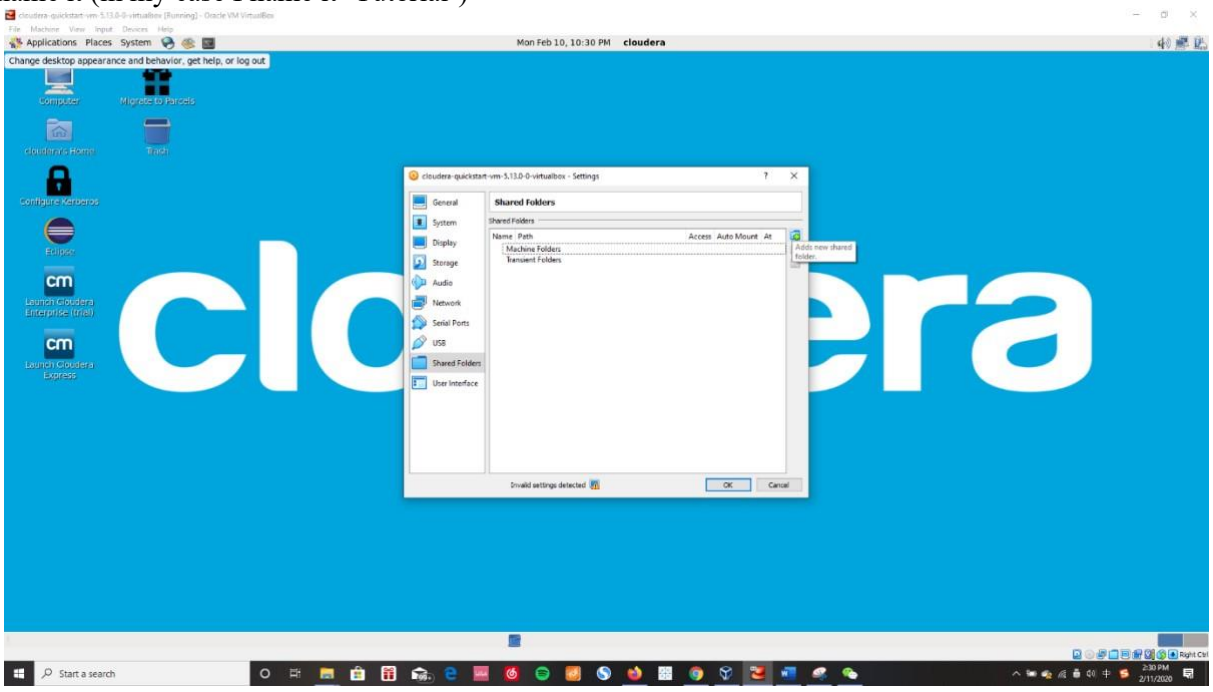


If you are unable to drag and drop files from your local machine to the VM, unmount (right click the disk like icon and choose unmount or eject) the disk like icon on the desktop of your Virtual Machine and then restart the VM, then try Option 2 below.

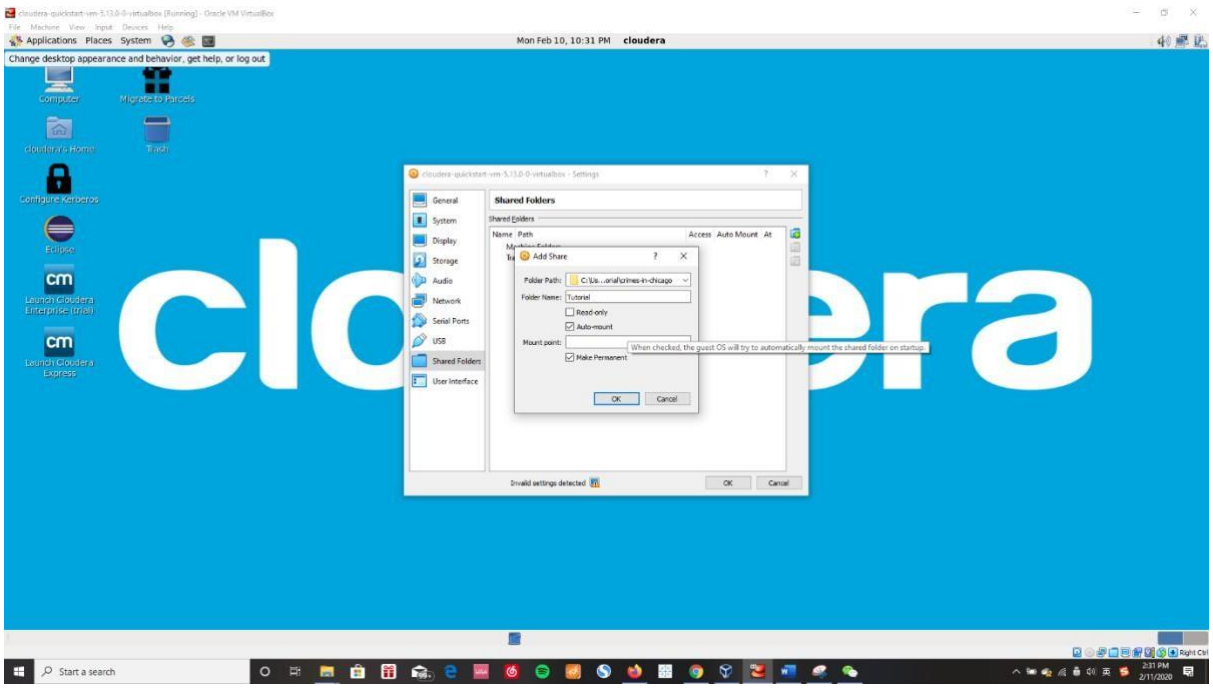
Step 2: Unable to drag'n'drop (Option 2)

If you are unable to drag or copy a folder to virtual machine, like what we did in previous step, please strictly follow the steps below. (If you were able to copy the file, just skip this part) First,
go to Devices -> Shared Folders -> Shared Folder Settings.

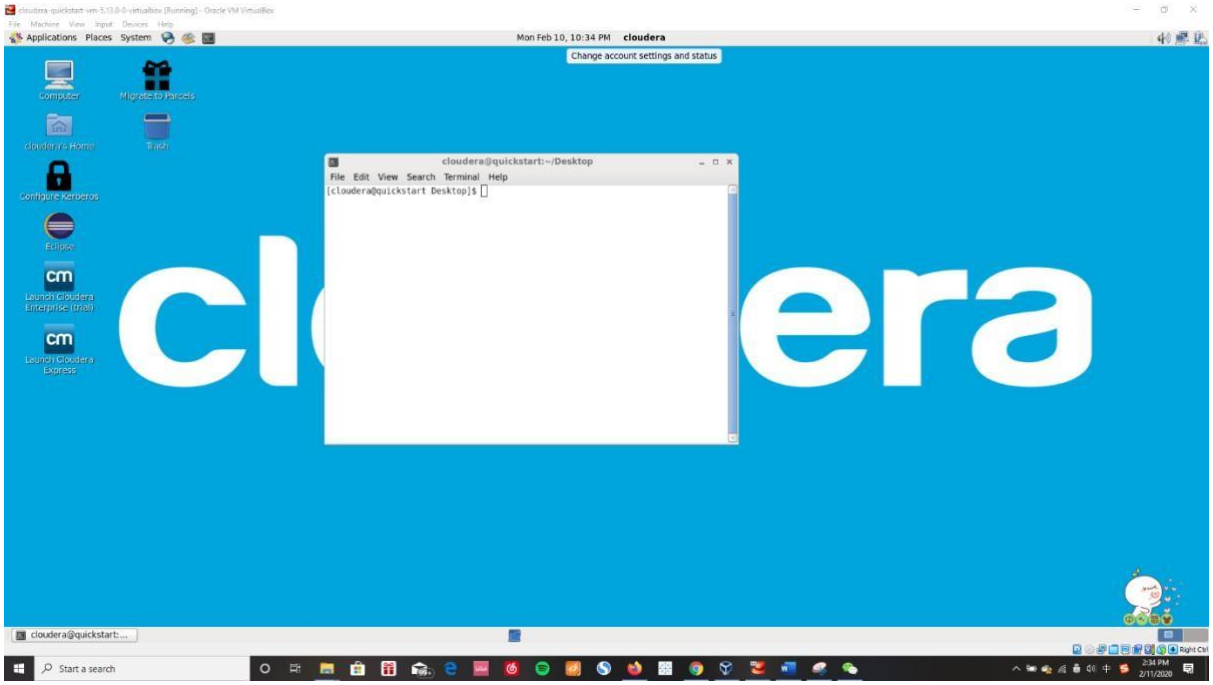
Then, select ‘Machine Folders’ and add a new shared folder and input your desired local path and name it (in my case I name it ‘Tutorial’)



Tick on button Anto-mount and Make Permanent, then hit OK.

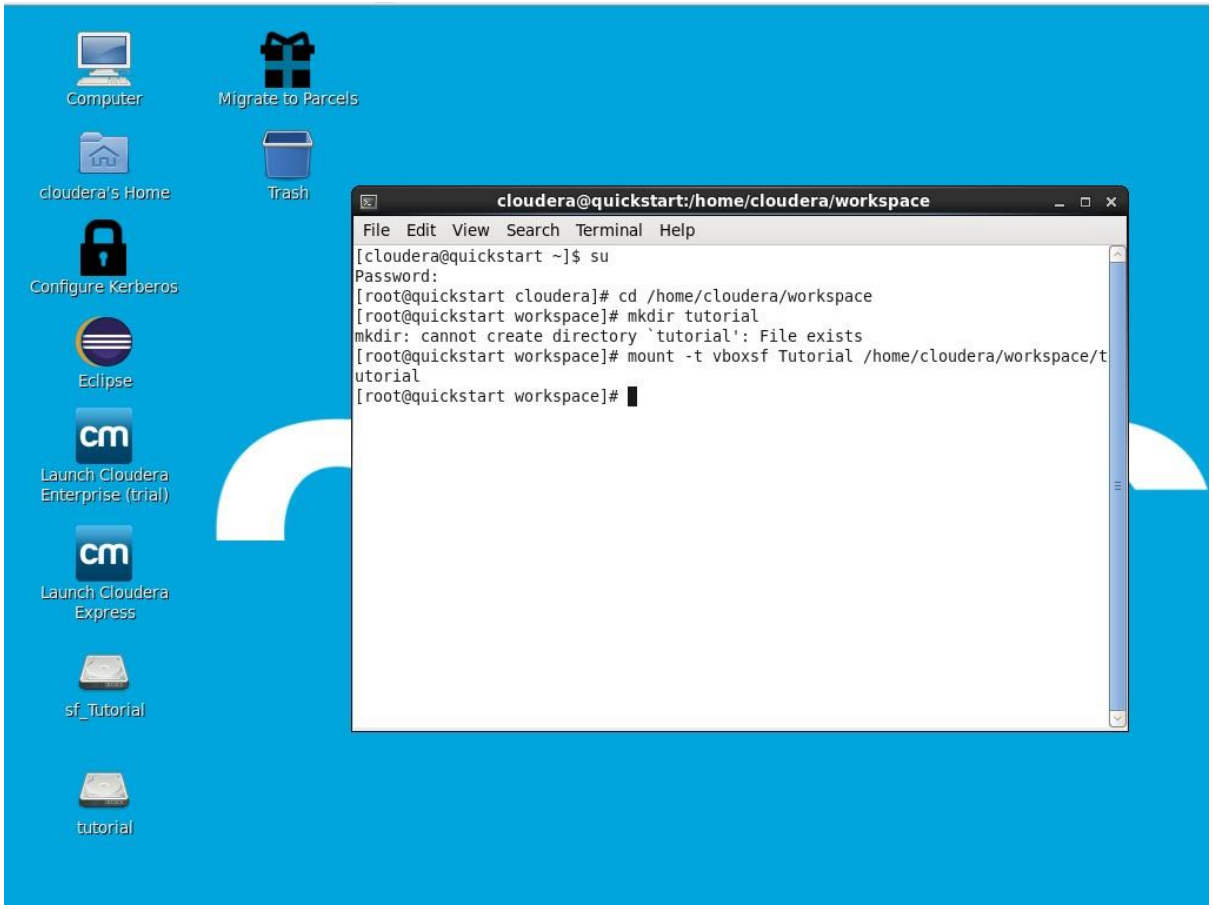


Go to Linux Terminal.



Input the same command in the screenshot

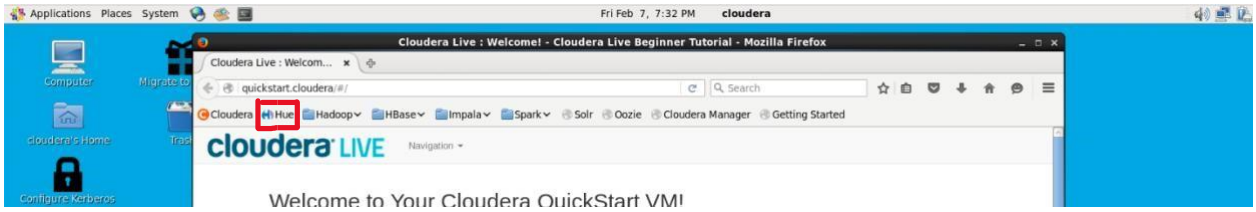
1. Input *su*, and then type *cloudera* as password
2. Input *cd /home/cloudera/workspace*
3. Input *mkdir tutorial*
4. Input *mount -t vboxsf Tutorial /home/cloudera/workspace/tutorial*
   Notice that in step 4 Tutorial should be changed to the name of your local folder



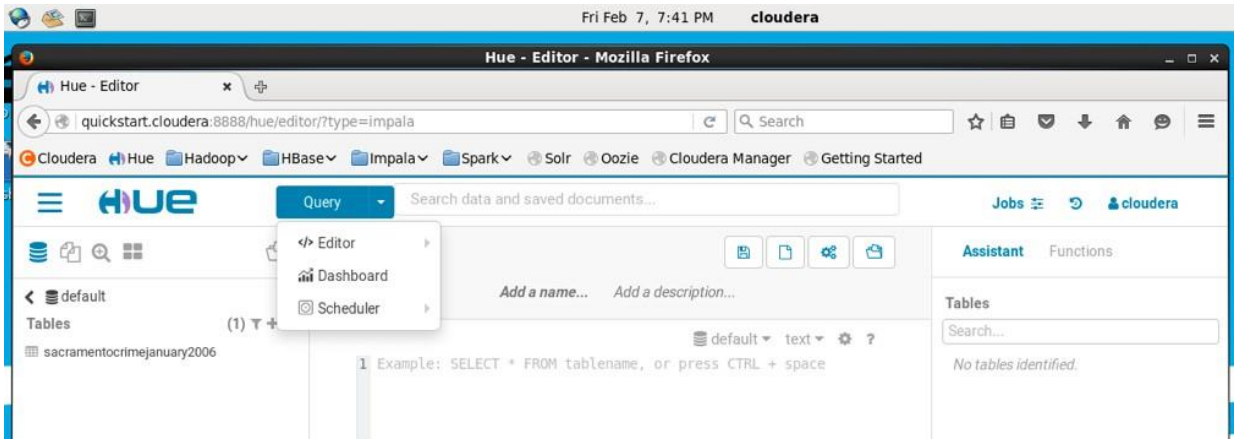
After that, you can see a file called ‘tutorial’ on your virtual machine desktop.

Step 3: Find and start learning about Hive

1. Open the browser and click on Hue on the top bar.



2. If you are asked to enter username and password, type in "cloudera" for both.
3. Locate the drill down menu and click on the downward arrow. Go to Editor → Hive.



Now, you will be able to see the Hive editor where you can create tables and run Hive queries to analyze your data.

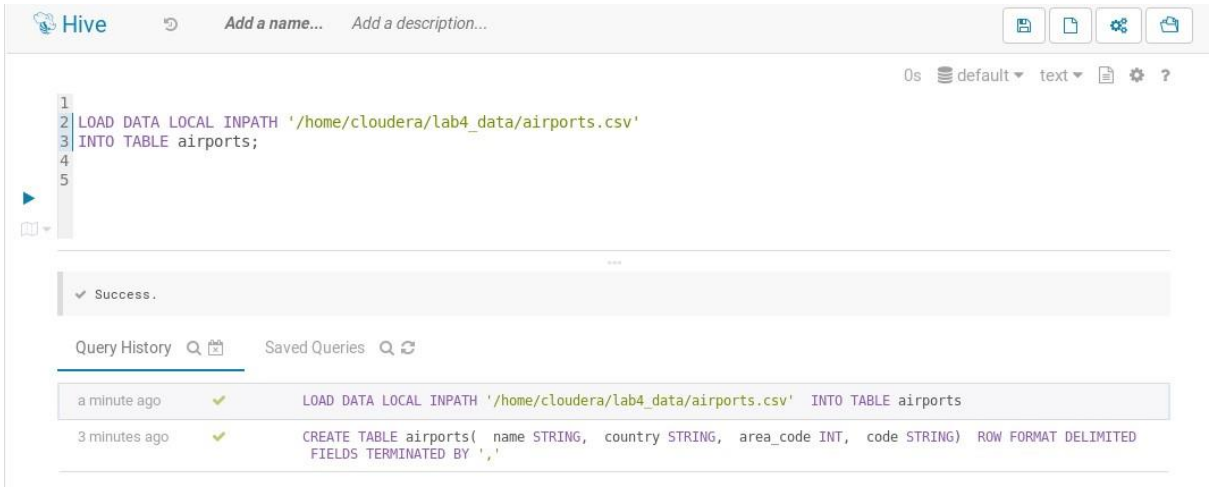
Step 4: Import csv to Hive with HQL (Optional) 1. Open Hue and run the following code to create a table.

```
CREATE TABLE airports(
name STRING,
country STRING,
area_code INT,    code
STRING) ROW FORMAT
DELIMITED
FIELDS TERMINATED BY ',';
```

2. Once the table has been created you can browse the table in left bar (Click default). We will also load data into it using the command below.

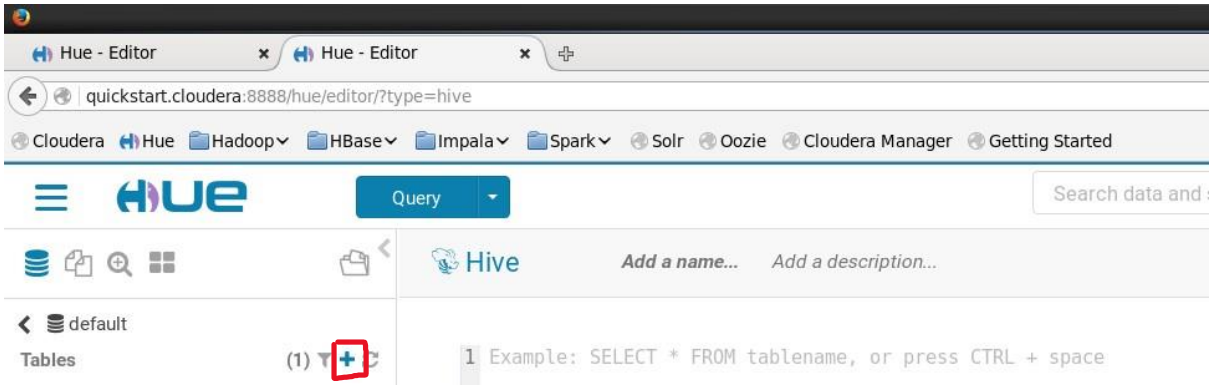
```
LOAD DATA LOCAL INPATH '/home/cloudera/lab4_data/airports.csv' INTO TABLE
airports;
```

Note: Your path maybe different than mine. Make sure you use the correct path where the file is located.

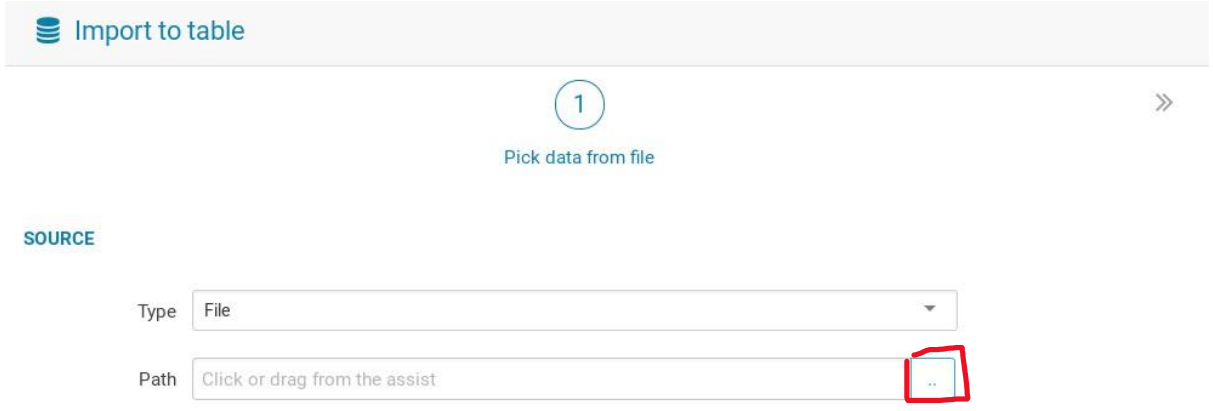


Step 5: Import csv file into Hive

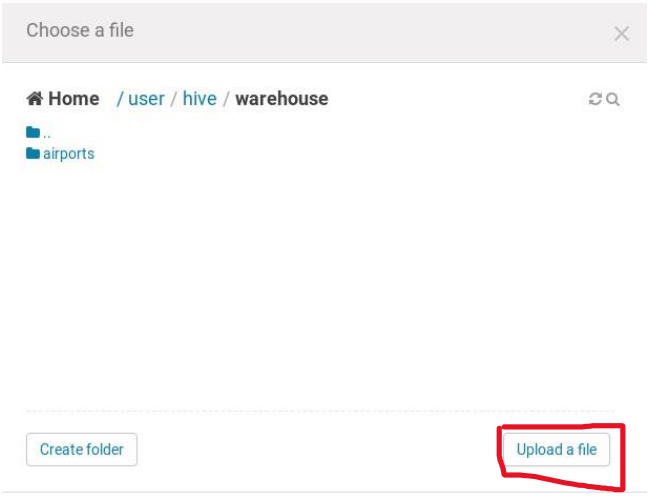
1. Open your HUE in Firefox. Under table spreadsheet, Click the '+' to create a new table with this interface.

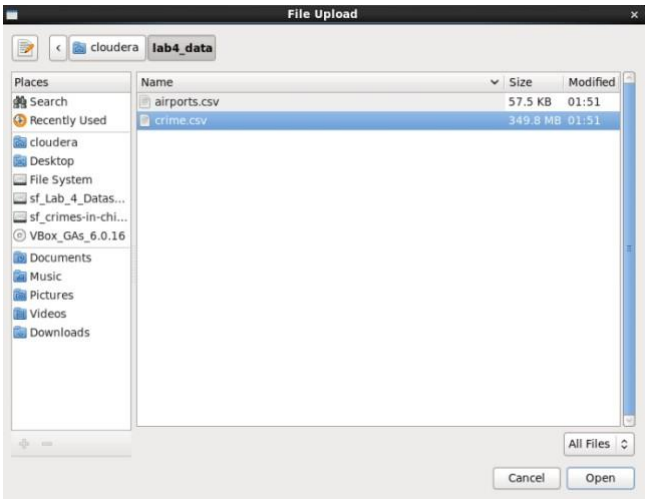


2. Click '..' to choose the file you want to load.

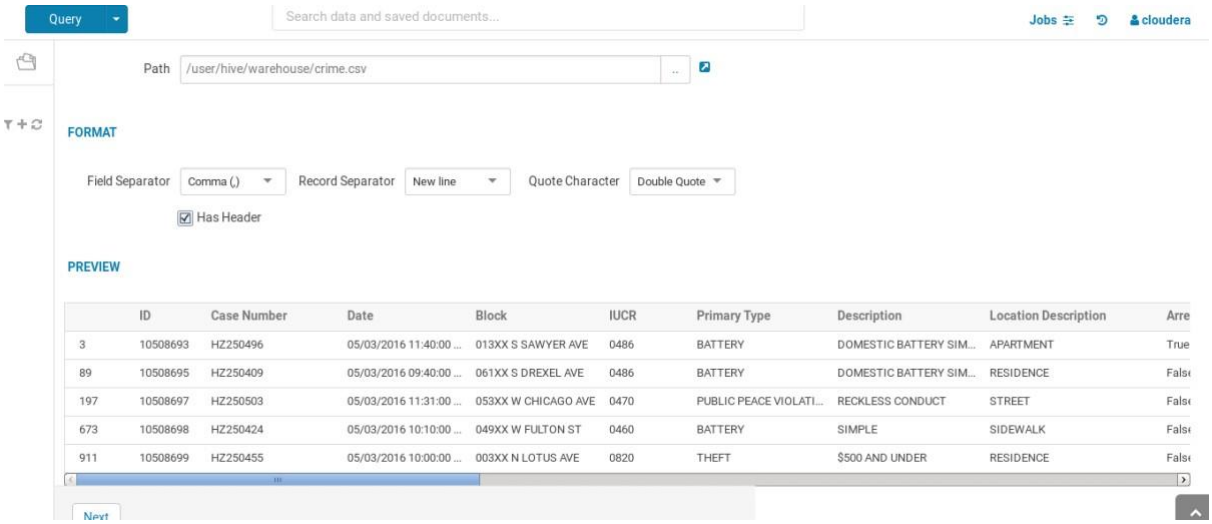


3. However, we do not have any files in our HDFS system. Click 'upload a file' to upload our crime.csv file to HDFS system (Make sure upload your csv file under '**/user/hive/warehouse**').



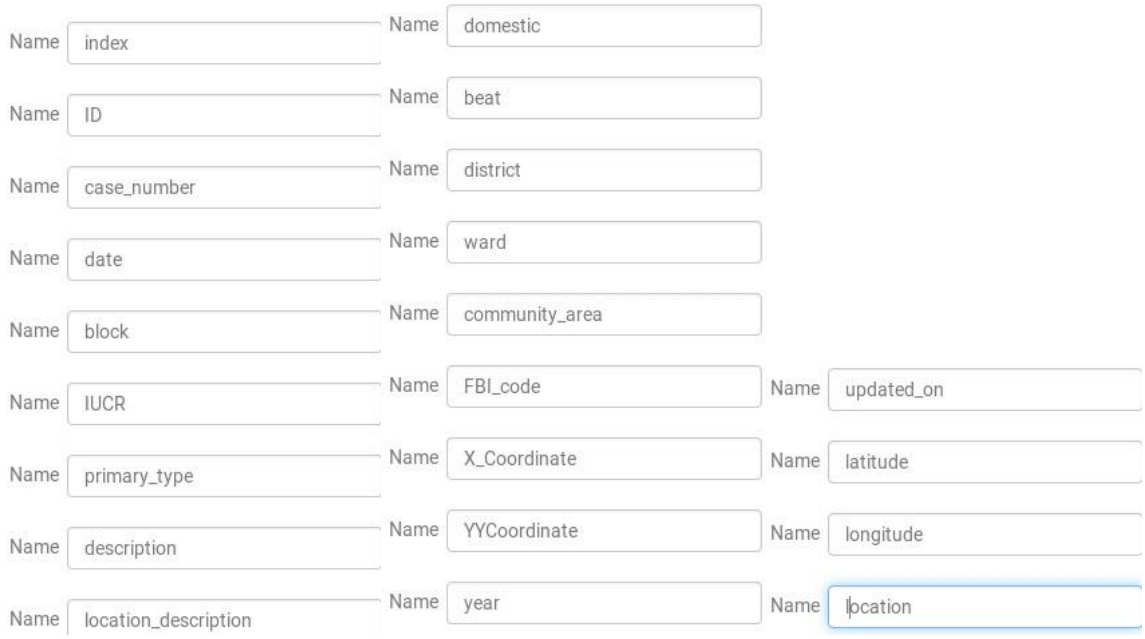


4. After uploading your csv file, click the file you want load and the interface will show as follow.

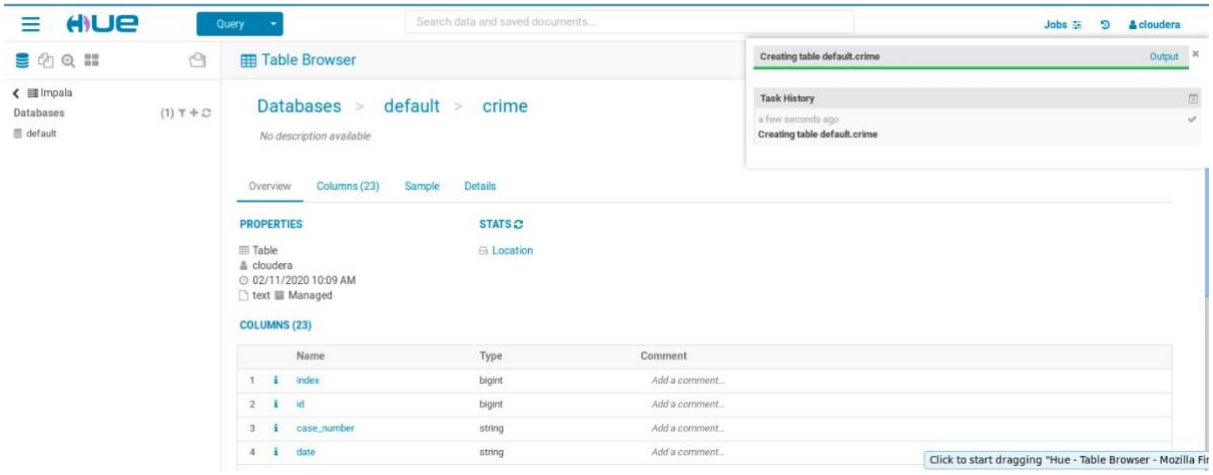


5. Click ‘Next’ and get ready to import our data.
    a. Give your table a name.

b. Fill the missing column names and make sure all the names all acceptable.
c. Check the datatype of each column.



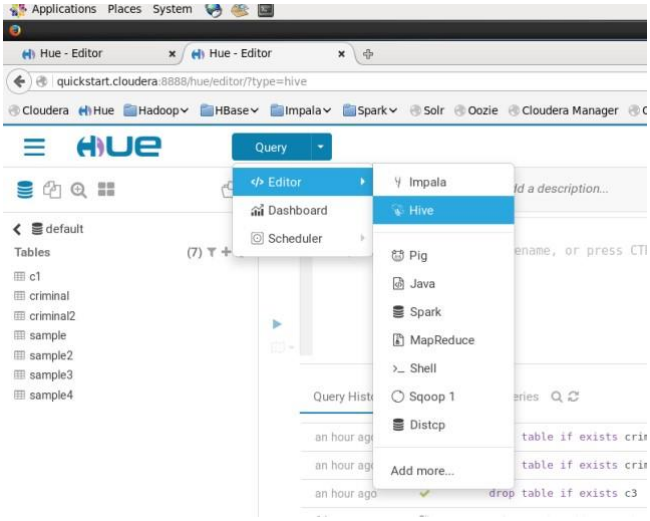
6. Click Submit to finish loading.



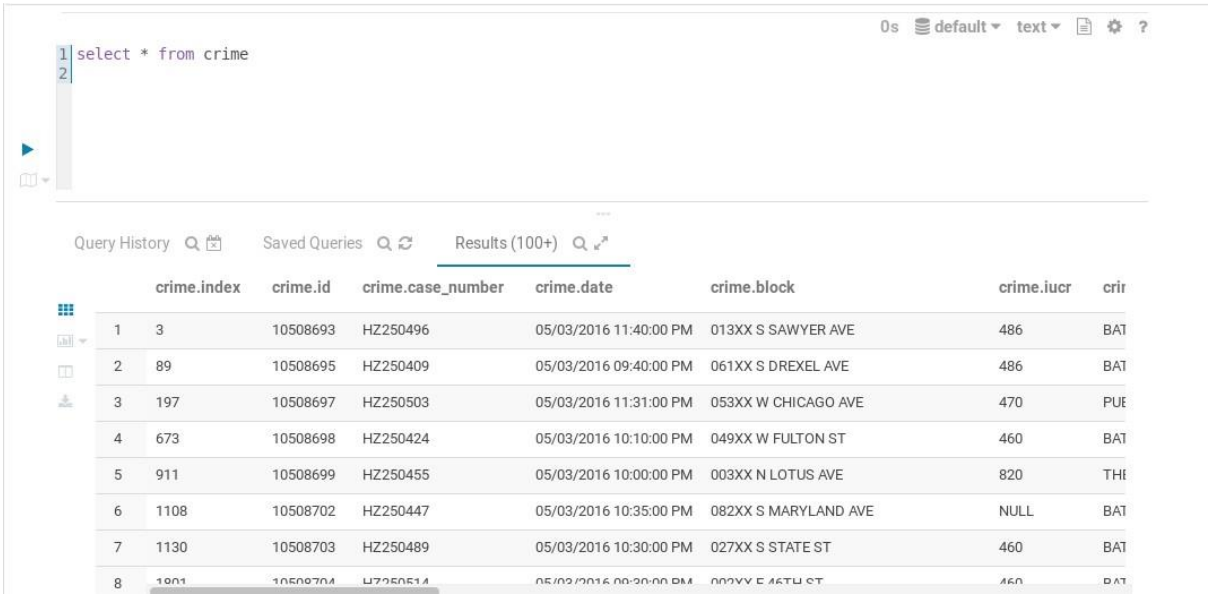
Now we can try some query in Hive. Click Query and choose Hive.

Step 6: Queries in Hive (HQL)

1. Go to Hive editor.



2. Check the data in our table.



	crime.index	crime.id	crime.case_number	crime.date	crime.block	crime.iucr	crir
1	3	10508693	HZ250496	05/03/2016 11:40:00 PM	013XX S SAWYER AVE	486	BAT
2	89	10508695	HZ250409	05/03/2016 09:40:00 PM	061XX S DREXEL AVE	486	BAT
3	197	10508697	HZ250503	05/03/2016 11:31:00 PM	053XX W CHICAGO AVE	470	PUE
4	673	10508698	HZ250424	05/03/2016 10:10:00 PM	049XX W FULTON ST	460	BAT
5	911	10508699	HZ250455	05/03/2016 10:00:00 PM	003XX N LOTUS AVE	820	THE
6	1108	10508702	HZ250447	05/03/2016 10:35:00 PM	082XX S MARYLAND AVE	NULL	BAT
7	1130	10508703	HZ250489	05/03/2016 10:30:00 PM	027XX S STATE ST	460	BAT
8	1801	10508704	HZ250514	05/03/2016 09:30:00 PM	002XX E 46TH ST	460	BAT

Now try to answer these questions with your query.

1. Please find Top 5 primary criminal type in all crimes of Chicago
2. Please retrieve # of theft in each year
3. Which month has the most # of theft?
4. For each month during the year of 2010 to 2018, count all types of crime and order them by desc.
5. Find out the latitude and longitude where NARCOTICS happened in the year of 2016.
6. Which location was most vulnerable during midnight to 11AM?
7. List the ward with most thefts in year 2018
8. Find assault and theft for each ward between 2016 and 2017.