Document 2: Loading Data

Objective: To Load the required Data into HDFS

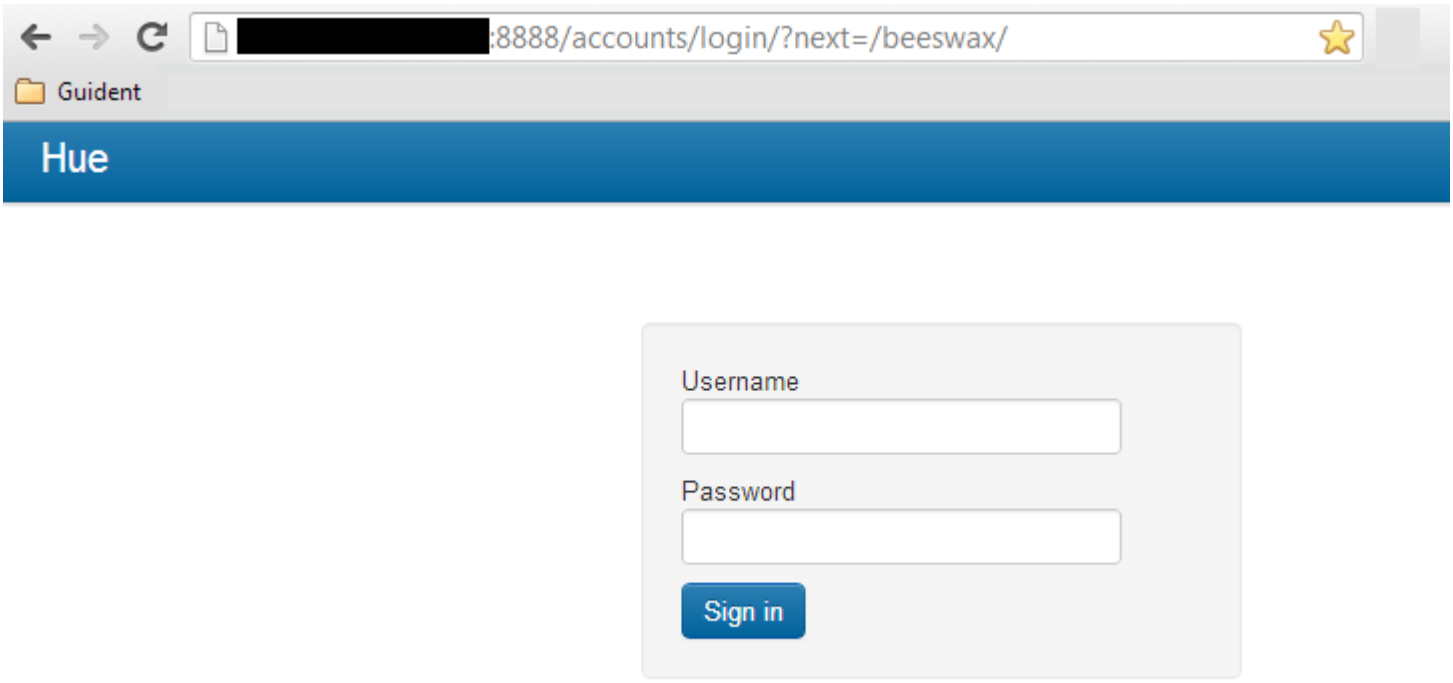
Beeswax:

The Beeswax application enables you to perform queries on Apache Hive, a data warehousing system designed to work with Hadoop. You can create Hive tables, load data, run and manage Hive queries, and download the results in a Microsoft Office Excel worksheet file or a comma-separated values file.

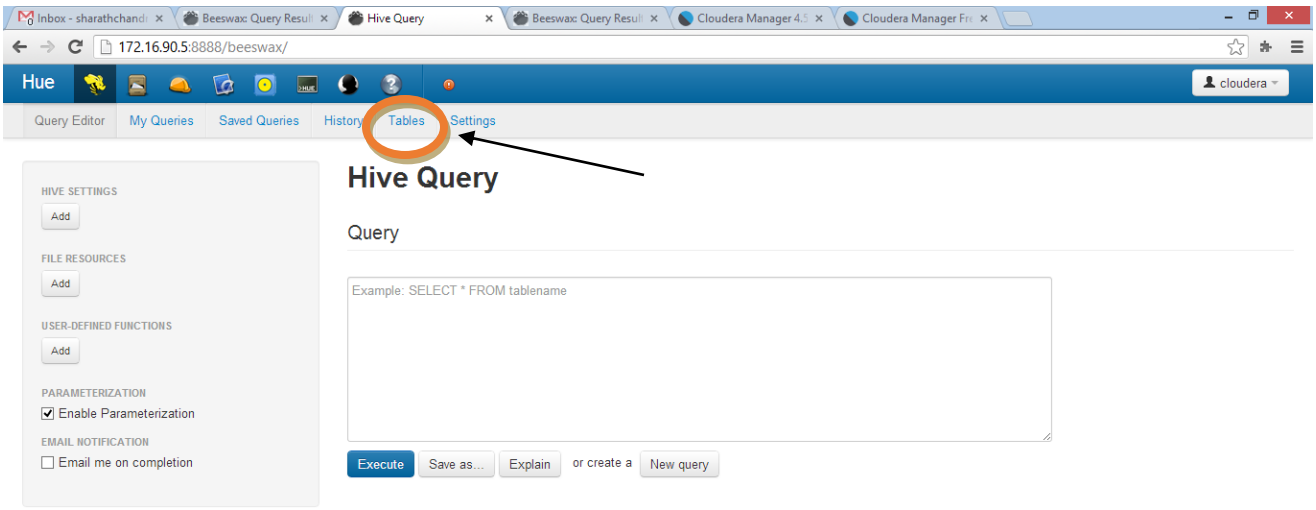
Beeswax, the Hive user interface in Hue, uses your system's Hive installation and is compatible with Hive 0.7. Your Hive data is stored in the Hadoop Distributed File System (HDFS), typically in the `/user/hive/warehouse`directory (or the directory you specify as `hive.metastore.warehouse.dir` in the `hive-site.xml` file). Make sure this directory exists and is writable by the users whom you expect to be creating tables. The directory `/tmp` (on the local file system) must also be world-writable because Hive uses it extensively.

Logging into the Web UI:

This screen appears as soon as you key in 172.16.90.5:8888 on your web browser:

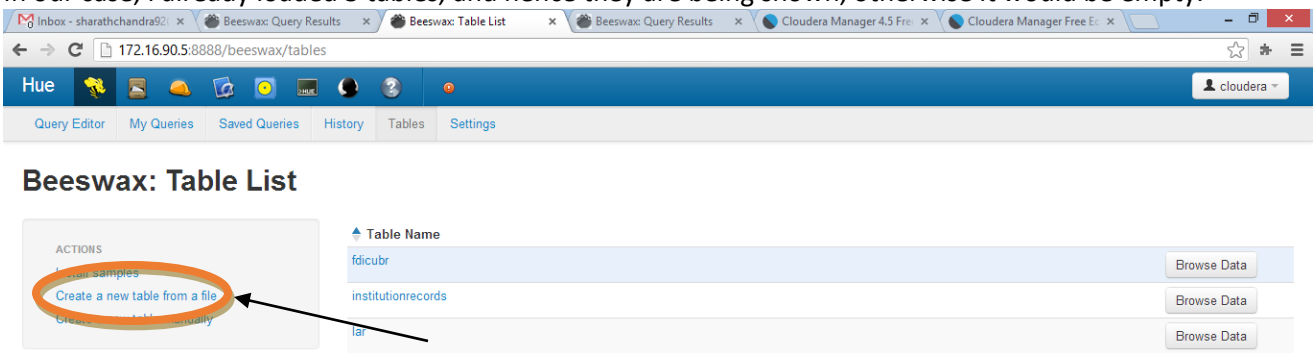


After entering the username and password, which are cloudera and wh!tesw@n (apparently), you are shown the services that exist on the VM:

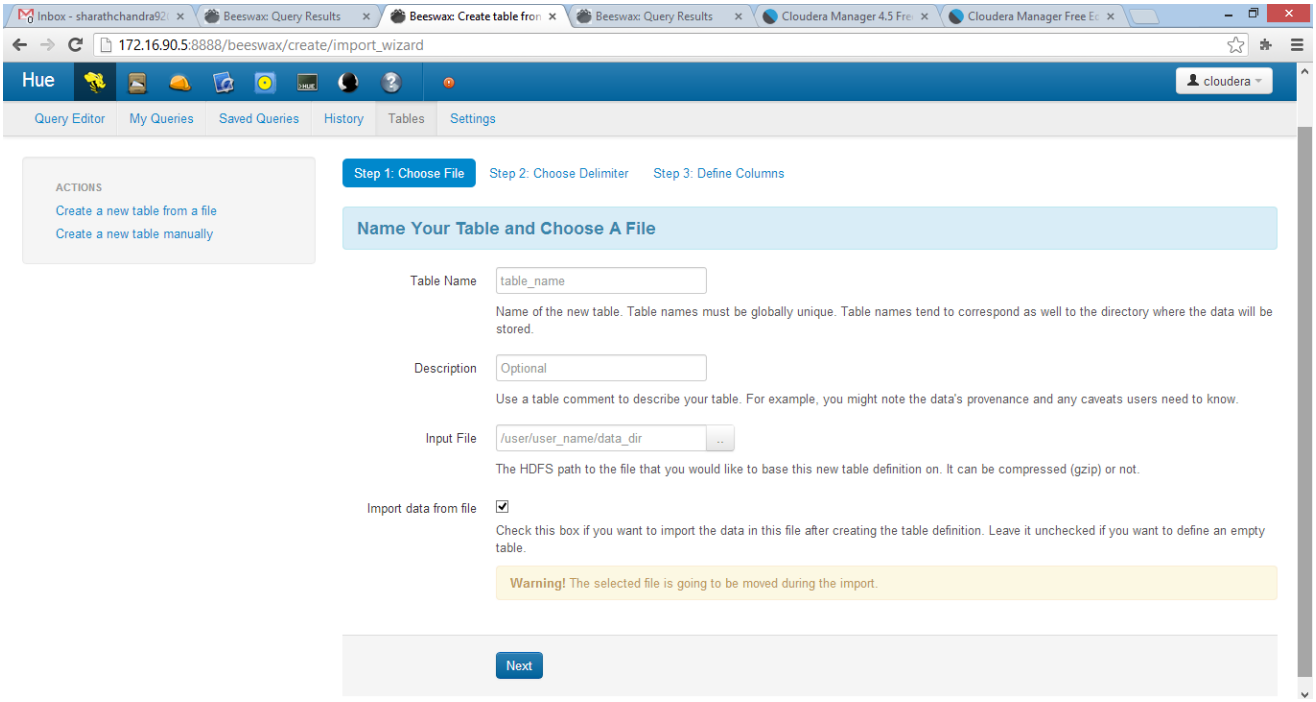


Here to the right on the toolbar pane, you can see ‘Tables’, click that

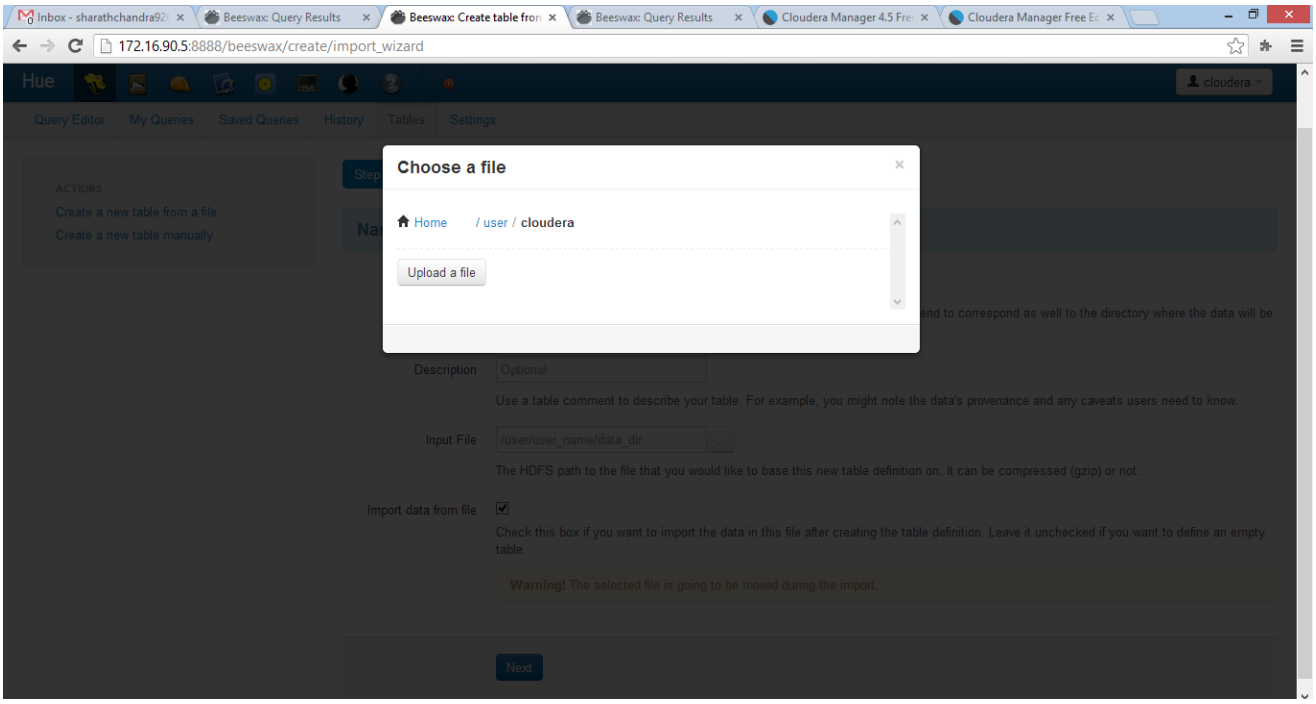
And you are shown a screen where existing tables/option to create tables is presented with:
In our case, I already loaded 3 tables, and hence they are being shown, otherwise it would be empty.



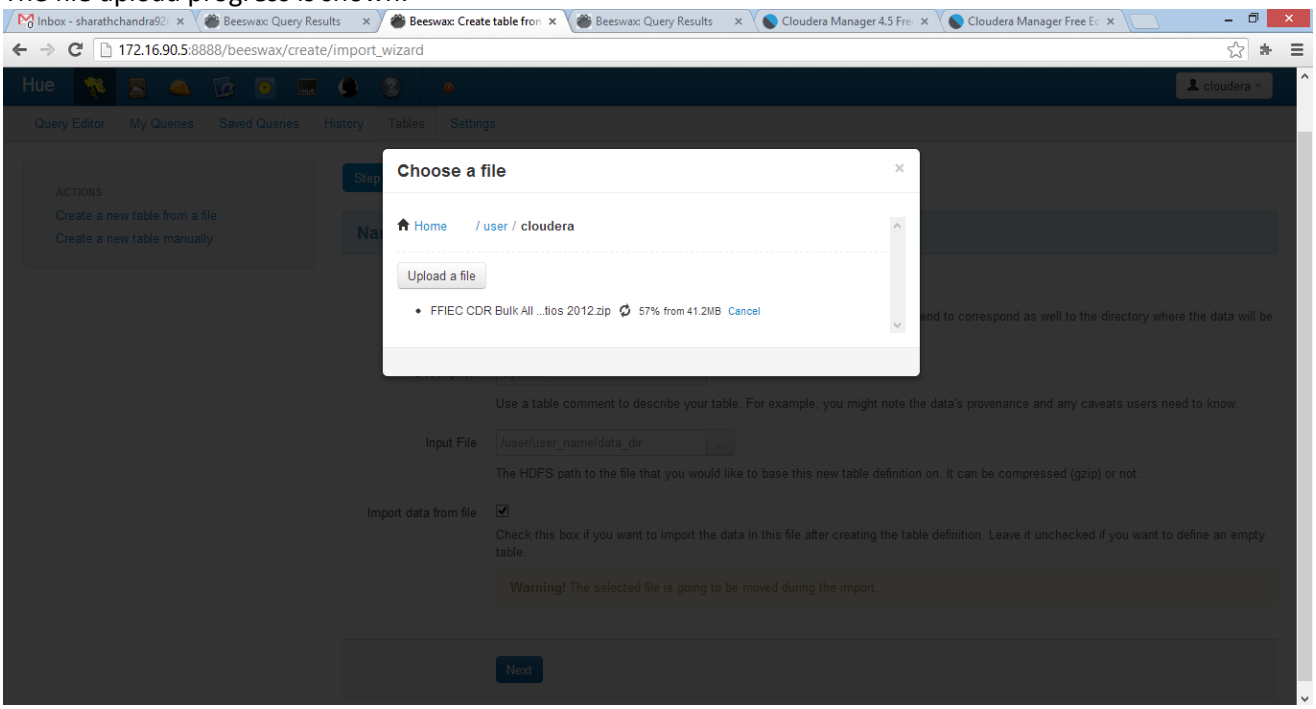
Now, Click on Create a new table from a file:
For this document sake, I am showing how to load one file and the same can be repeated for other files.



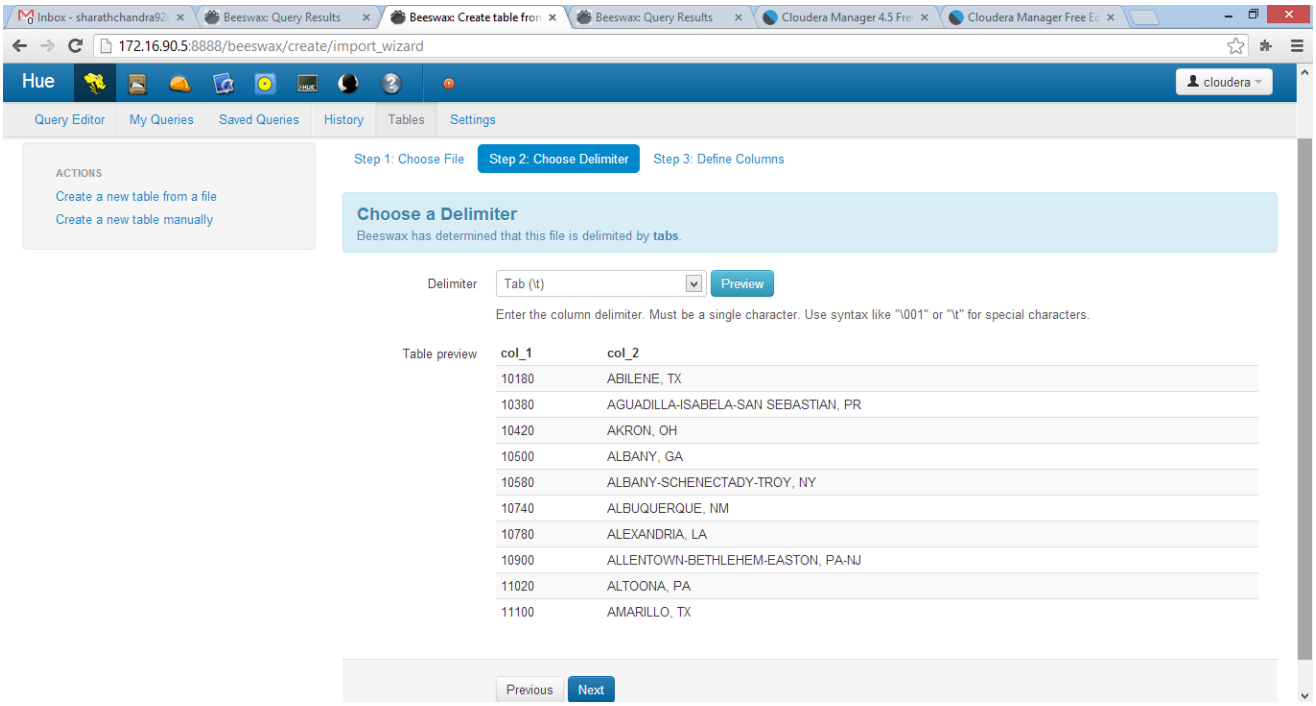
Fill in the details and select the Input file from your hard disk



The file upload progress is shown:

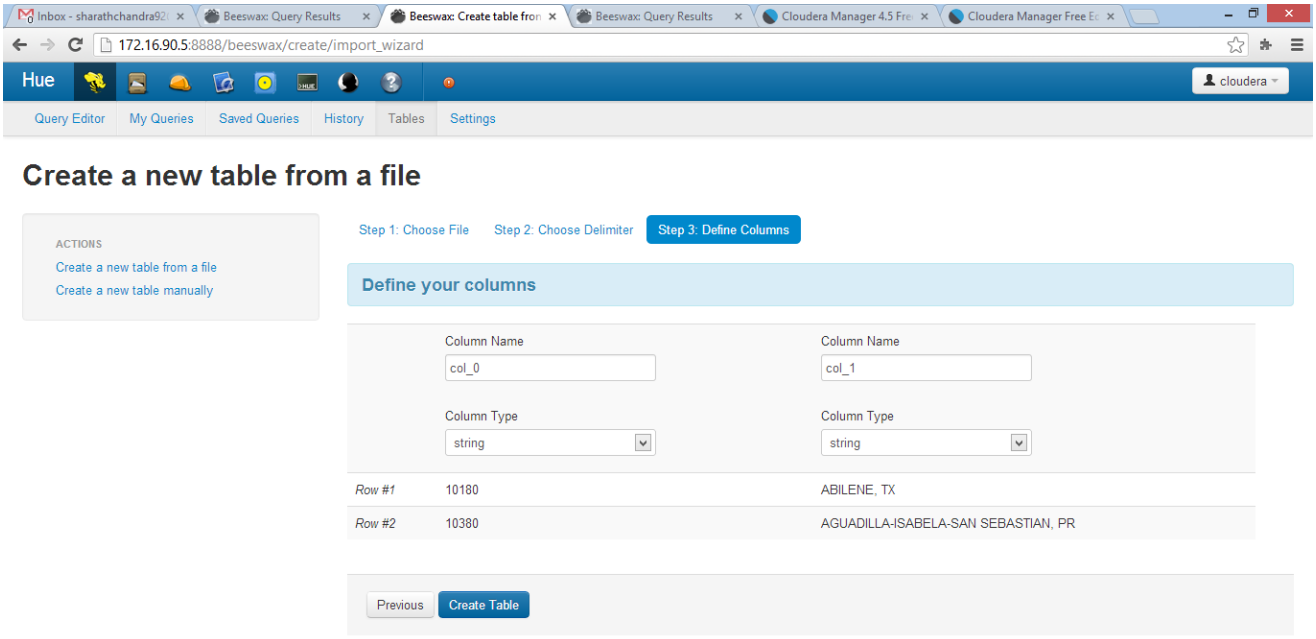


The file is uploaded into the hive directory. Select it and proceed.



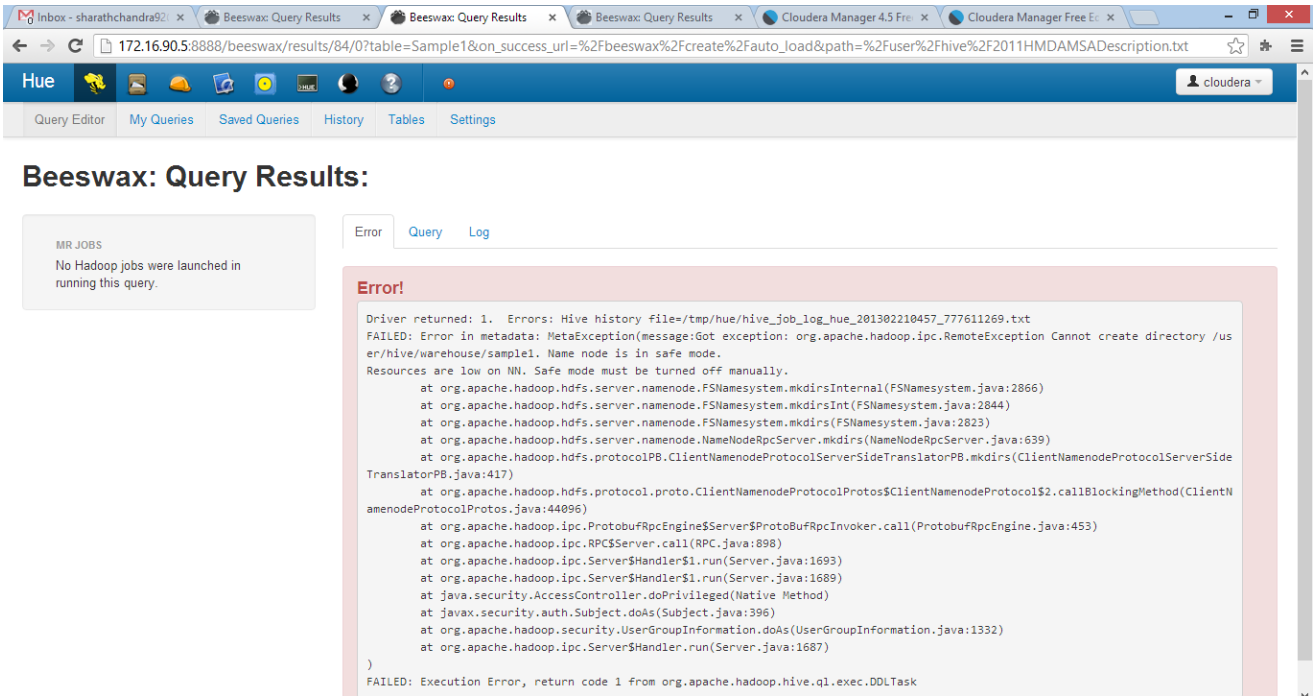
Then choose the appropriate delimiter (most of our files use ‘,’ or ‘\t’)

Name the coloumns accordingly. Here, care has to exercised to trim the first row from the file, in case that represents the coloumn name. This can be done by importing the file into MS Excel and deleting the first row. Else, there can be issues while processing the data at later stages.



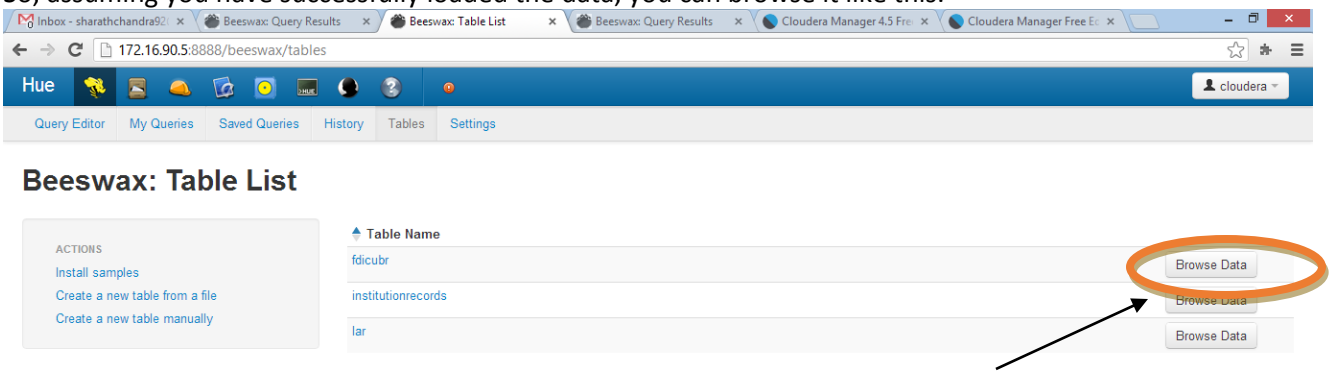
Press Create Table!

A successful upload redirects you to the table. In case there is a failure, the following can be the reasons:



1. The file name contains ‘Spaces/other invalid characters’ in it. Remove all the spaces and redo the process.
2. The coloumns were given invalid names: again having special characters and so on. Rename the coloumns.
3. An incorrect delimiter was selected. Although there wont be an error in this case, but browsing the contents of the file makes it clear that this is not what we want!

So, assuming you have successfully loaded the data, you can browse it like this:



And you can go through the data.
Now, we are all set to play with the data we have on the HDFS.
Btw, this being a 1-node VM, things are quite simpler and easy to achieve and at the same time, there is a heavy constraint on not being able to process heavy chunks of data due to memory shortage, which we experienced while running Hive queries to link datasets. Therefore, we have to select the best possible dataset and then experiment with that.