



Document 2: Loading Data

Objective: To Load the required Data into HDFS

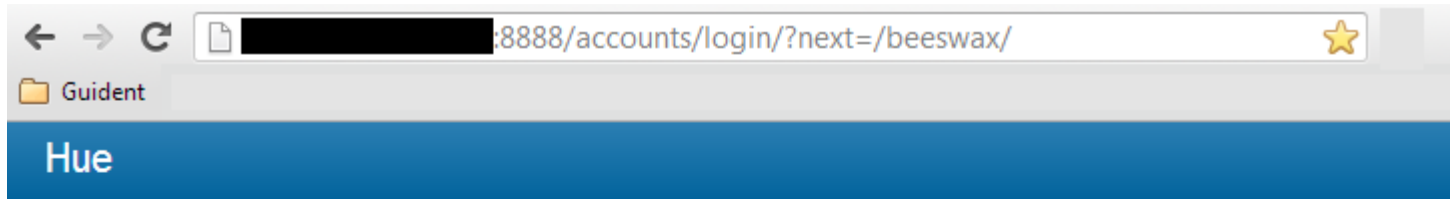
Beeswax:

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.

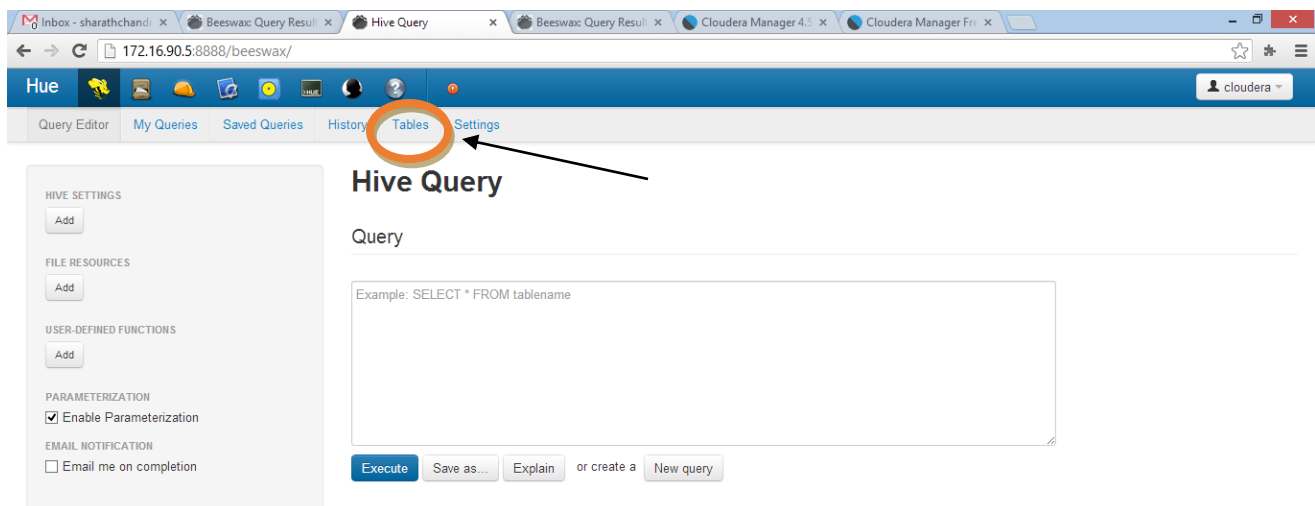
Beeswax, the Hive user interface in Hue, uses your system's Hive installation and is compatible with Hive 0.7. Your Hive data is stored in the Hadoop Distributed File System (HDFS), typically in the `/user/hive/warehouse` directory (or the directory you specify as `hive.metastore.warehouse.dir` in the `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:

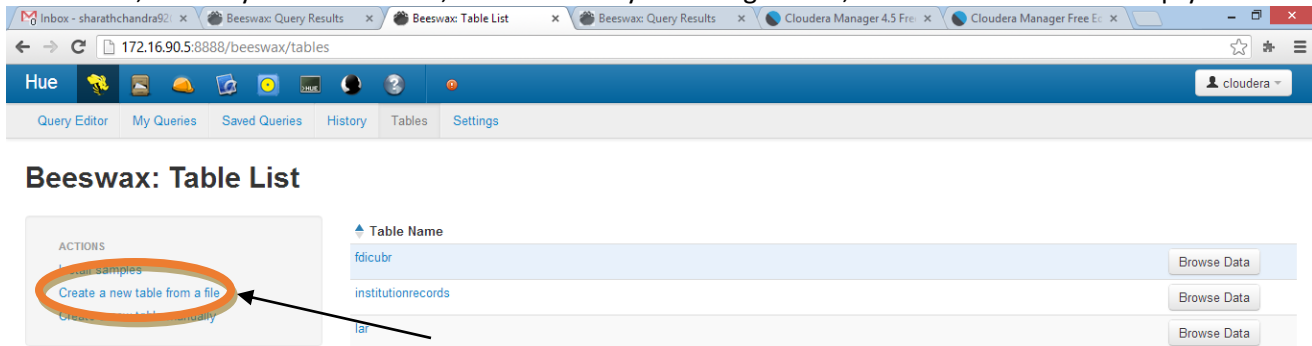
A close-up of the login form from the previous image. It features a light gray background. At the top is the label 'Username' above a white text input field. Below that is the label 'Password' above another white text input field. At the bottom is a blue button with the text 'Sign in' in white.

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

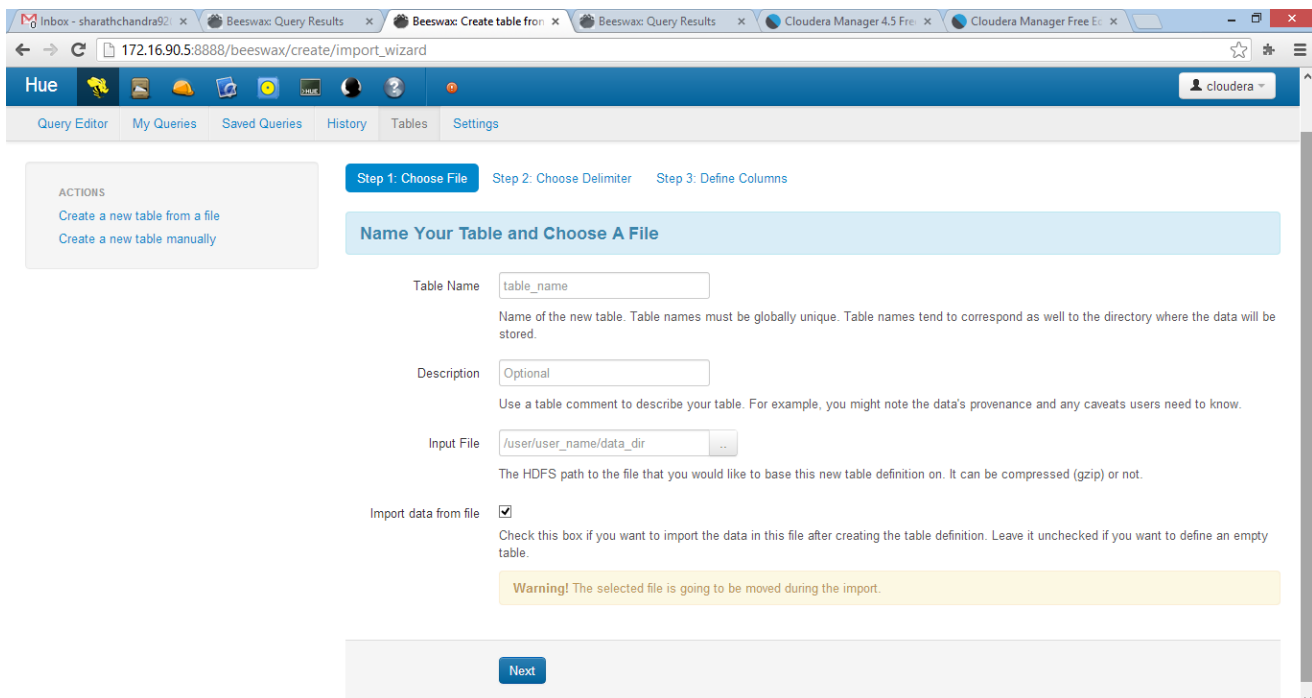


Here to the right on the toolbar pane, you can see 'Tables', click that

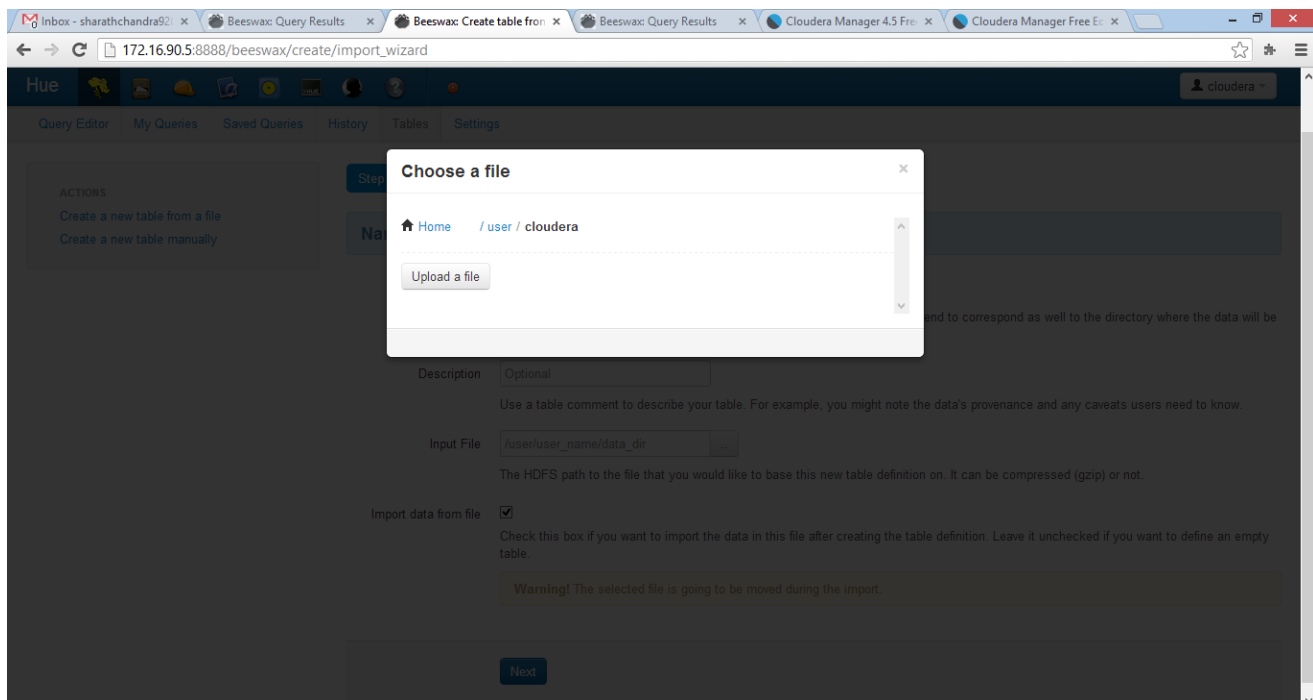
And you are shown a screen where existing tables/option to create tables is presented with:
In our case, I already loaded 3 tables, and hence they are being shown, otherwise it would be empty.



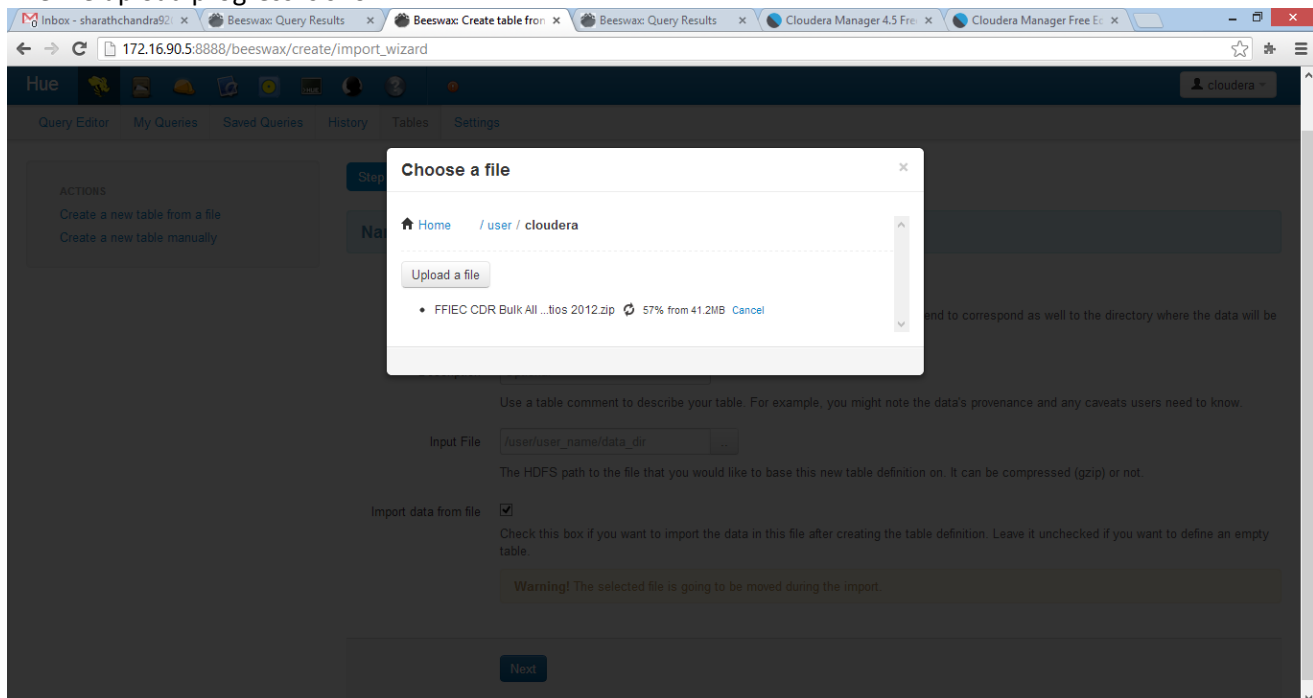
Now, Click on Create a new table from a file:
For this document sake, I am showing how to load one file and the same can be repeated for other files.



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.

Step 1: Choose File Step 2: Choose Delimiter Step 3: Define Columns

Choose a Delimiter
Beeswax has determined that this file is delimited by tabs.

Delimiter: Tab (\t) Preview

Enter the column delimiter. Must be a single character. Use syntax like "\001" or "\t" for special characters.

Table preview

col_1	col_2
10180	ABILENE, TX
10380	AGUADILLA-ISABELA-SAN SEBASTIAN, PR
10420	AKRON, OH
10500	ALBANY, GA
10580	ALBANY-SCHENECTADY-TROY, NY
10740	ALBUQUERQUE, NM
10780	ALEXANDRIA, LA
10900	ALLENTOWN-BETHLEHEM-EASTON, PA-NJ
11020	ALTOONA, PA
11100	AMARILLO, TX

Previous Next

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

Name the columns accordingly. Here, care has to be exercised to trim the first row from the file, in case that represents the column 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.

Step 1: Choose File Step 2: Choose Delimiter Step 3: Define Columns

Define your columns

Column Name	Column Name
col_0	col_1

Column Type	Column Type
string	string

Row #1 10180 ABILENE, TX

Row #2 10380 AGUADILLA-ISABELA-SAN SEBASTIAN, PR

Previous Create Table

Create a new table from a file

Press Create Table!

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

The screenshot shows the Beeswax Query Results page in Hue. The URL is `172.16.90.5:8888/ beeswax/results/84/0?table=Sample1&on_success_url=%2Fbeeswax%2Fcreate%2Fauto_load&path=%2Fuser%2Fhive%2F2011HMDAMSADescription.txt`. The page title is "Beeswax: Query Results". On the left, a message says "MR JOBS No Hadoop jobs were launched in running this query." The main content area shows an "Error!" message with a stack trace. The error message is: "Driver returned: 1. Errors: Hive history file=/tmp/hue/hive_job_log_hue_201302210457_777611269.txt FAILED: Error in metadata: MetaException(message:Got exception: org.apache.hadoop.ipc.RemoteException Cannot create directory /user/hive/warehouse/sample1. Name node is in safe mode. Resources are low on NN. Safe mode must be turned off manually. at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.mkdirsInternal(FSNamesystem.java:2866) at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.mkdirsInt(FSNamesystem.java:2844) at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.mkdirs(FSNamesystem.java:2823) at org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.mkdirs(NameNodeRpcServer.java:639) at org.apache.hadoop.hdfs.protocolPB.ClientNameNodeProtocolServerSideTranslatorPB.mkdirs(ClientNameNodeProtocolServerSideTranslatorPB.java:417) at org.apache.hadoop.hdfs.protocol.proto.ClientNameNodeProtocolProtos\$ClientNameNodeProtocol\$2.callBlockingMethod(ClientNameNodeProtocolProtos.java:44096) at org.apache.hadoop.ipc.ProtobufRpcEngine\$Server\$ProtoBufRpcInvoker.call(ProtobufRpcEngine.java:453) at org.apache.hadoop.ipc.RPC\$Server.call(RPC.java:898) at org.apache.hadoop.ipc.Server\$Handler\$1.run(Server.java:1693) at org.apache.hadoop.ipc.Server\$Handler\$1.run(Server.java:1689) at java.security.AccessController.doPrivileged(Native Method) at javax.security.auth.Subject.doAs(Subject.java:396) at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1332) at org.apache.hadoop.ipc.Server\$Handler.run(Server.java:1687)) FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.q1.exec.DDLTask".

1. The file name contains 'Spaces/other invalid characters' in it. Remove all the spaces and redo the process.
2. The columns were given invalid names: again having special characters and so on. Rename the columns.
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:

The screenshot shows the Beeswax Table List page in Hue. The URL is `172.16.90.5:8888/ beeswax/tables`. The page title is "Beeswax: Table List". On the left, there are links: "Install samples", "Create a new table from a file", and "Create a new table manually". The main content area shows a table with the following data:

Table Name
fdicubr
institutionrecords
lar

For each table, there is a "Browse Data" button. The "Browse Data" button for the "fdicubr" table is highlighted with an orange circle and an arrow.

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.