1. What is Hive?

Apache Hive

The Apache Hive ™ data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage using SQL (Standard Query Language).

Hive is an initiative started by Facebook to provide a traditional Data Warehouse interface for MapReduce programming. For writing queries for MapReduce in SQL fashion, the Hive compiler converts them in the background to be executed in the Hadoop cluster. It helps the programmers to use their SQL knowledge rather than focusing on developing a new language

Features of Hive:

* Provide SQL type language which is called HQL.
* Helps in querying large data sets stored in HDFS(Hadoop Distributed File System).
* It is an open-source tool.
* It supports flexible project views and makes data visualisation easy.

1. Hive Vs Map Reduce

| **MapReduce** | **Hive** |
| --- | --- |
| 1. It is a Data Processing Language | It is a SQL-like Query Language. |
| 1. It converts the job into map-reduce functions. | It converts the SQL queries to HQL(Hive-QL) |
| 1. It provides low level of abstraction | It provides a high level of abstraction |
| 1. It is difficult for the user to perform join operations | It makes it easy for the user to perform SQL-like operations on HDFS. |
| 1. The user has to write 10 times more lines of code to perform a similar task than Pig. | The user has to write a few lines of code than MapReduce. |
| 1. It has several jobs therefore execution time is more | The code execution time is more but development effort is less |
| 1. It is supported by versions of the Hadoop. | It is also supported with recent versions of Hadoop. |

3. Hive Vs Relational Databases

| **RDBMS** | **Hive** |
| --- | --- |
| 1. It is used to maintain databases. | It is used to maintain data warehouse |
| 1. It uses SQL (Structured Query Language). | It uses HQL (Hive Query Language). |
| 1. Schema is fixed in RDBMS | Schema varies in it. |
| 1. Normalised data is stored. | Normalised and denormalized both type of data is stored. |
| 1. Tables in rdms are sparse | Tables in hive are dense. |
| 1. It doesn’t support partitioning | It supports automation partition. |
| 1. No partition method is used. | Sharding method is used for partition. |

4. Installation and setup of hive

**1. Prerequisites**

1. Hardware Requirement

RAM — Min. 8GB, if you have an SSD in your system then 4GB RAM would also work.

CPU — Min. Quad core, with at least 1.80GHz

1. JRE 1.8 — Offline installer for JRE

<https://www.java.com/en/download/windows_offline.jsp>

1. Java Development Kit — 1.8

<https://www.oracle.com/java/technologies/downloads/#java8>

1. A Software for Unzipping like 7Zip or Win Rar

<https://www.7-zip.org/download.html>

<https://www.win-rar.com/download.html?L=0>

1. Hadoop

I am using Hadoop-2.9.2, you can also use any other STABLE version for Hadoop.

1. MySQL Query Browser

<https://dev.mysql.com/downloads/mysql/>

1. Download Hive zip

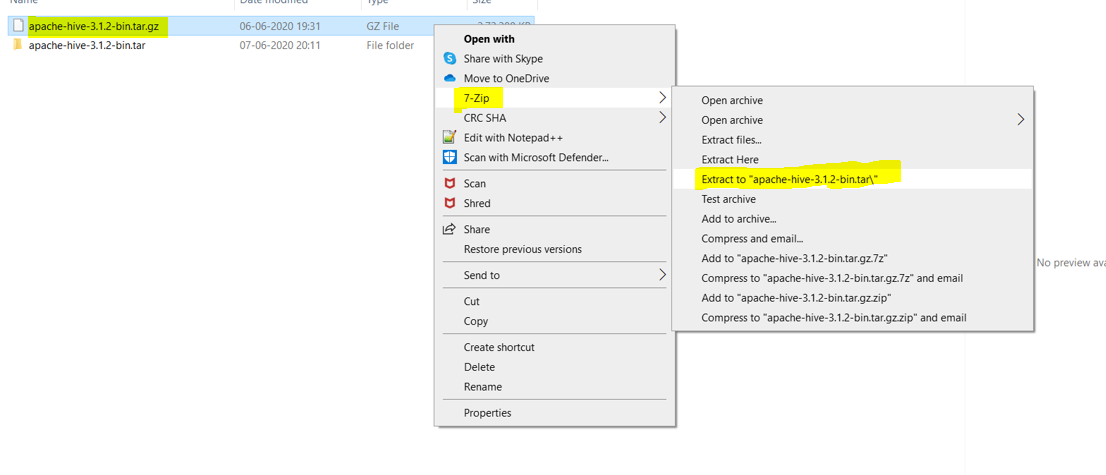
<https://downloads.apache.org/hive/hive-3.1.2/>

I am using Hive-3.1.2, you can also use any other STABLE version for Hive.



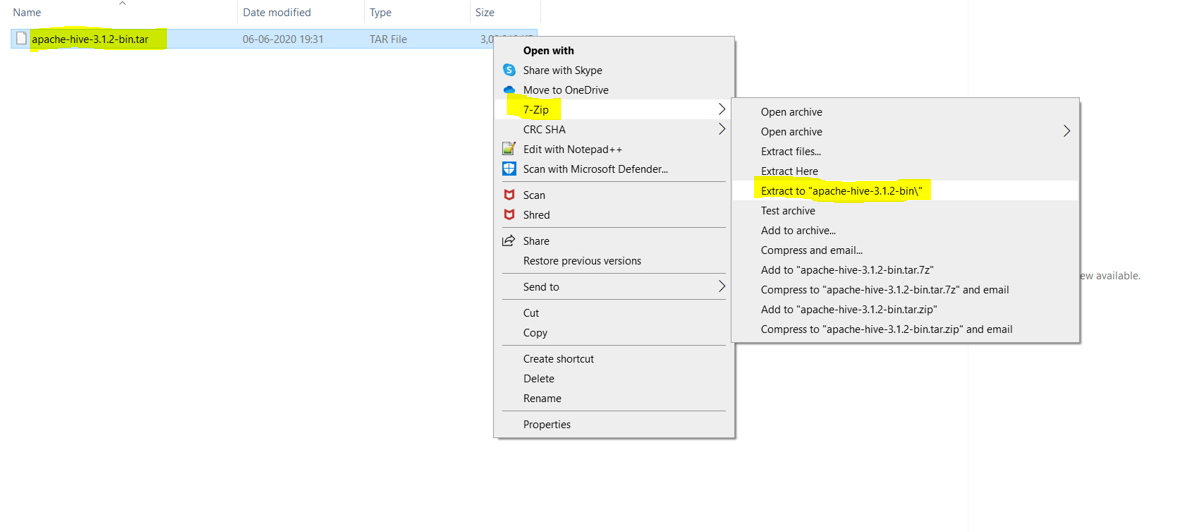
2. Unzip and Install Hive

After Downloading the Hive, we need to Unzip the apache-hive-3.1.2-bin.tar.gz file.

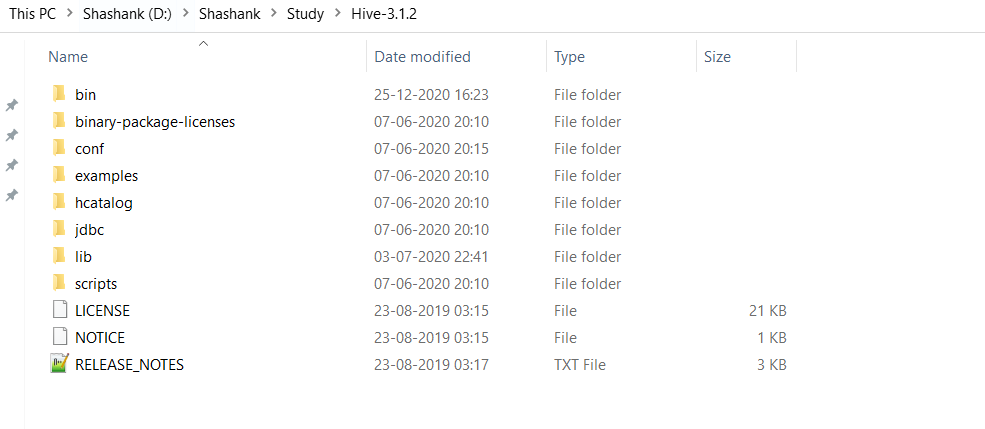


Once extracted, we would get a new file apache-hive-3.1.2-bin.tar

Now, once again we need to extract this tar file.



Now we can organize our Hive installation, we can create a folder and move the final extracted file in it. For Eg. :-



Please note while creating folders, DO NOT ADD SPACES IN BETWEEN THE FOLDER NAME.(it can cause issues later)

I have placed my Hive in D: drive you can use C: or any other drive also

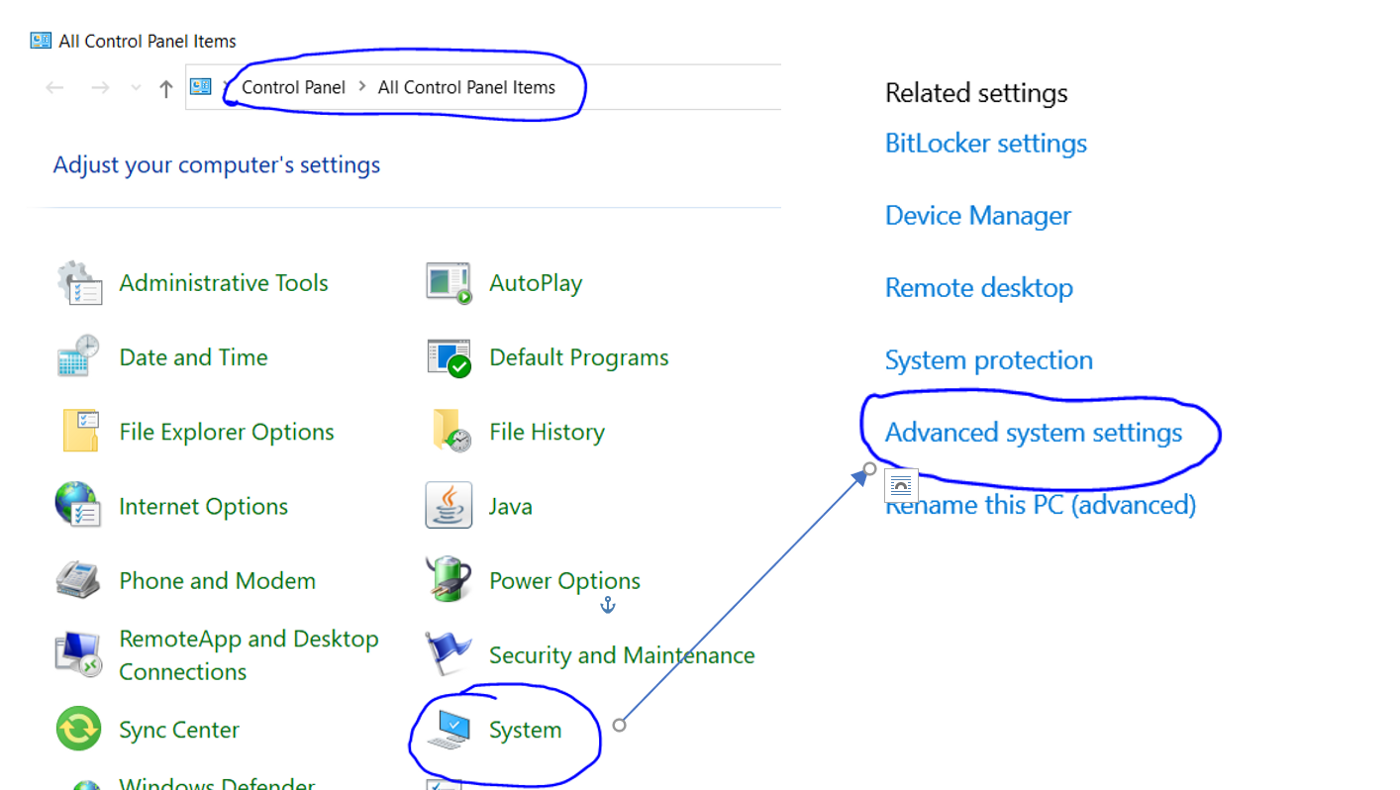
3. Setting Up Environment Variables

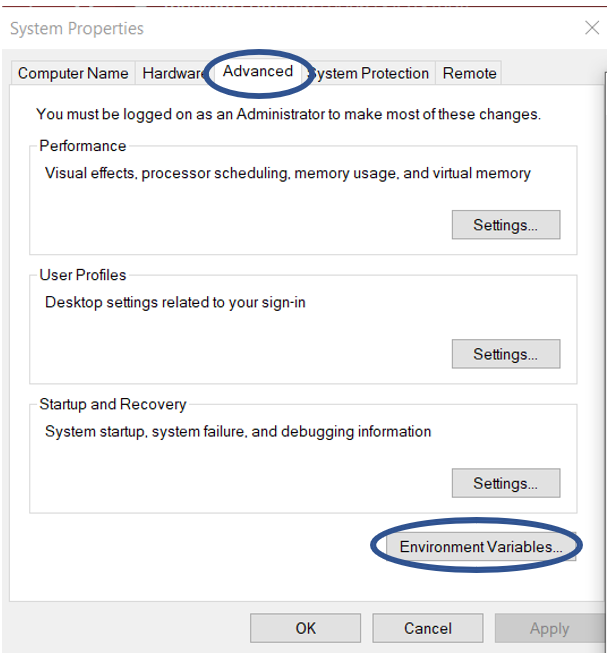
Another important step in setting up a work environment is to set your Systems environment variable.

To edit environment variables, go to Control Panel > System > click on the “Advanced system settings” link

Alternatively, We can Right click on This PC icon and click on Properties and click on the “Advanced system settings” link

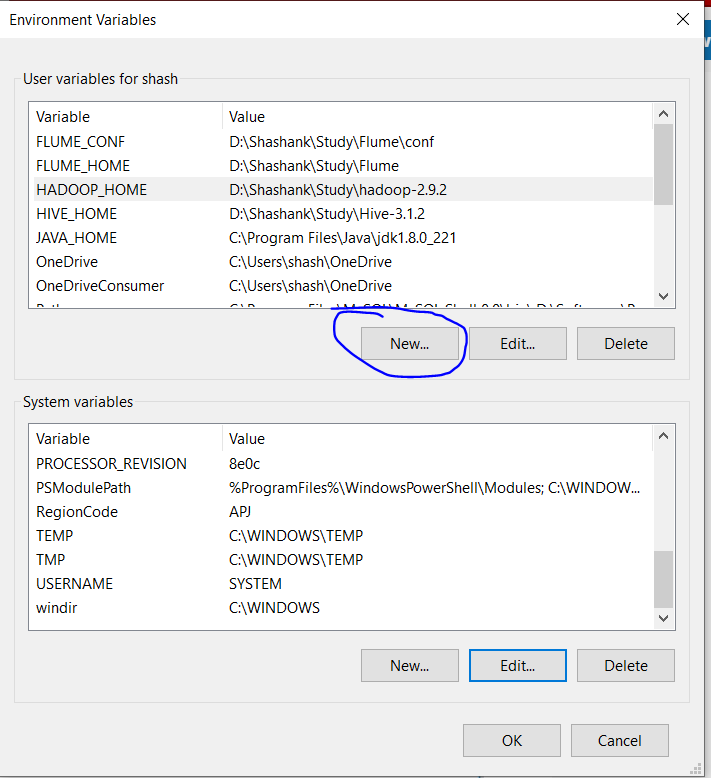
Or, easiest way is to search for Environment Variable in search bar and there you GO



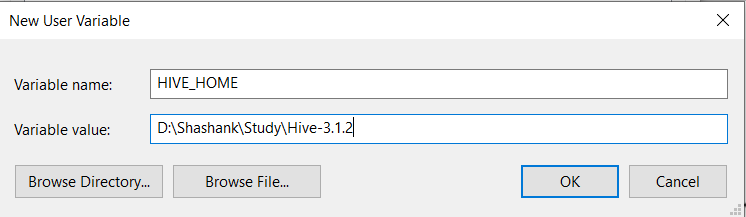


3.1 Setting HIVE\_HOME

Open environment Variable and click on “New” in “User Variable”



On clicking “New”, we get below screen.

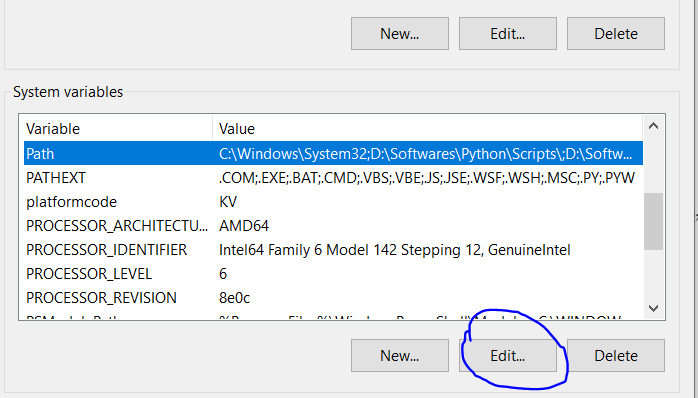


Now as shown, add HIVE\_HOME in variable name and path of Hive in Variable Value.

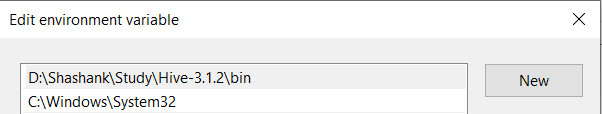
Click OK and we are half done with setting HIVE\_HOME.

3.2 Setting Path Variable

Last step in setting the Environment variable is setting Path in System Variable.



Select Path variable in the system variables and click on “Edit”.



Now we need to add these paths to Path Variable :-

\* %HIVE\_HOME%\bin

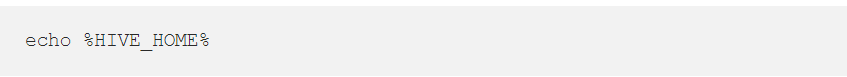
Click OK and OK. & we are done with Setting Environment Variables.

3.4 Verify the Paths

Now we need to verify that what we have done is correct and reflecting.

Open a NEW Command Window

Run following commands



4. Editing Hive

Once we have configured the environment variables the next step is to configure Hive. It has 7 parts:-

4.1 Replacing bins

First step in configuring the hive is to download and replace the bin folder.

\* Go to this GitHub Repo and [download](https://github.com/HadiFadl/Hive-cmd) the bin folder as a zip.

\* Extract the zip and replace all the files present under bin folder to %HIVE\_HOME%\bin

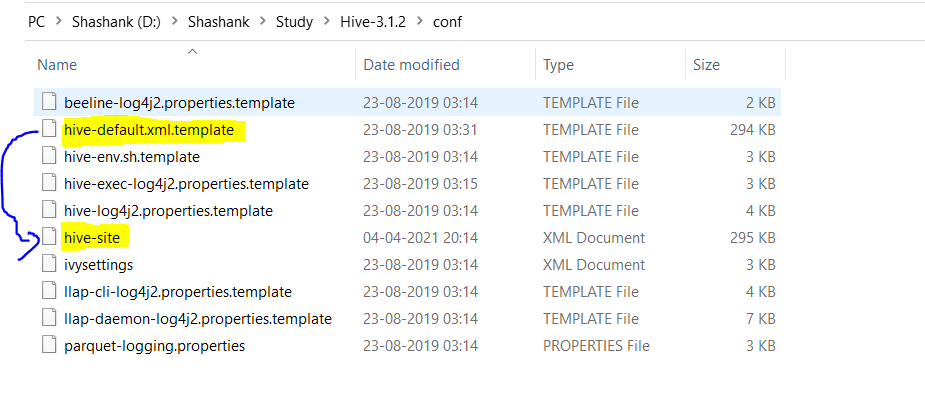
Note:- If you are using different version of HIVE then please search for its respective bin folder and download it.

4.2 Creating File Hive-site.xml

Now we need to create the Hive-site.xml file in hive for configuring it :-

(We can find these files in Hive -> conf -> hive-default.xml.template)

We need to copy the hive-default.xml.template file and paste it in the same location and rename it to hive-site.xml. This will act as our main Config file for Hive.



4.3 Editing Configuration Files

4.3.1 Editing the Properties

Now Open the newly created Hive-site.xml and we need to edit the following properties



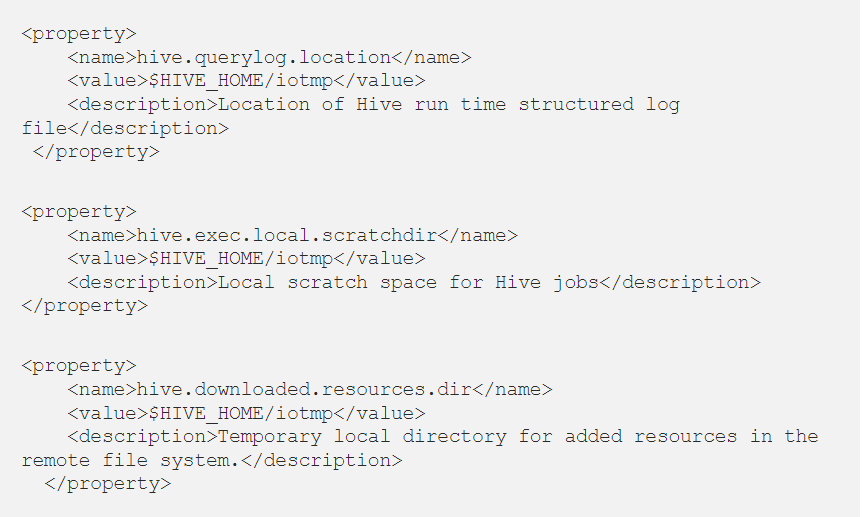
Replace the value for <Your IP Address> with the IP Address of your System and replace <Your drive Folder> with the Hive folder Path.

4.3.2 Removing Special Characters

This is a short step and we need to remove all the &#8 character present in the hive-site.xml file.

4.3.3 Adding few More Properties

Now we need to add the following properties as it is in the hive-site.xml File.



Great..!!! We are almost done with the Hive part, for configuring MySQL database as Metastore for Hive, we need to follow below steps:-

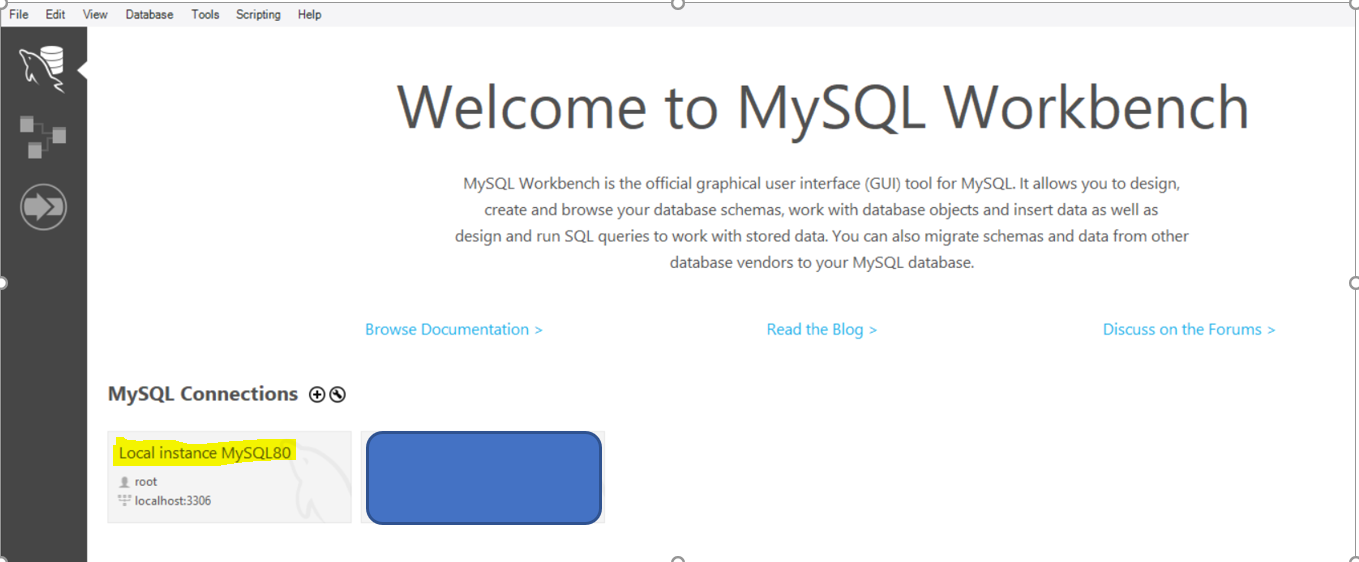
4.4 Creating Hive User in MySQL

The next important step in configuring Hive is to create users for MySQL.

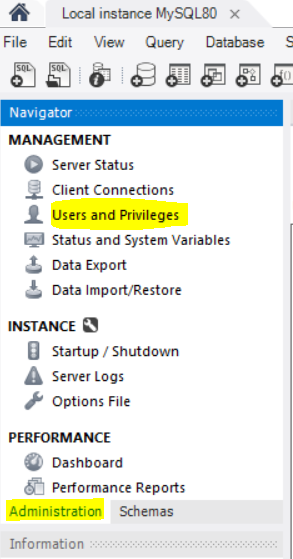
These Users are used for connecting Hive to MySQL Database for reading and writing data from it.

Note:- You can skip this step if you have created the hive user while SQOOP installation

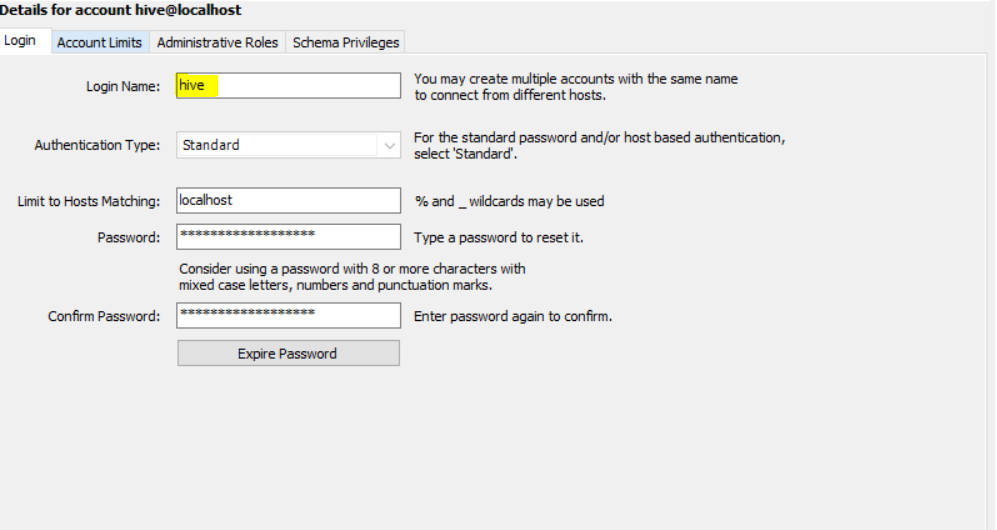
Firstly, we need to open the MySQL Workbench and open the workspace(default or any specific, if you want). We will be using the default workspace only for now.



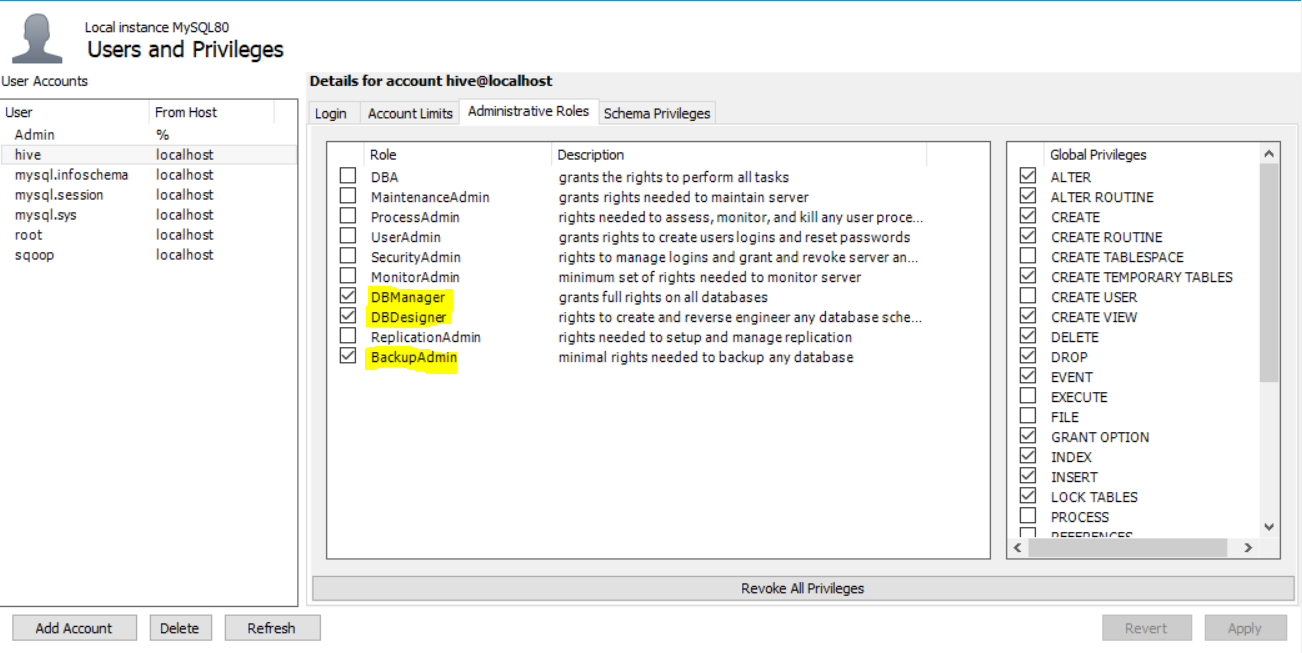
Now Open the Administration option in the Workspace and select Users and privileges option under Management



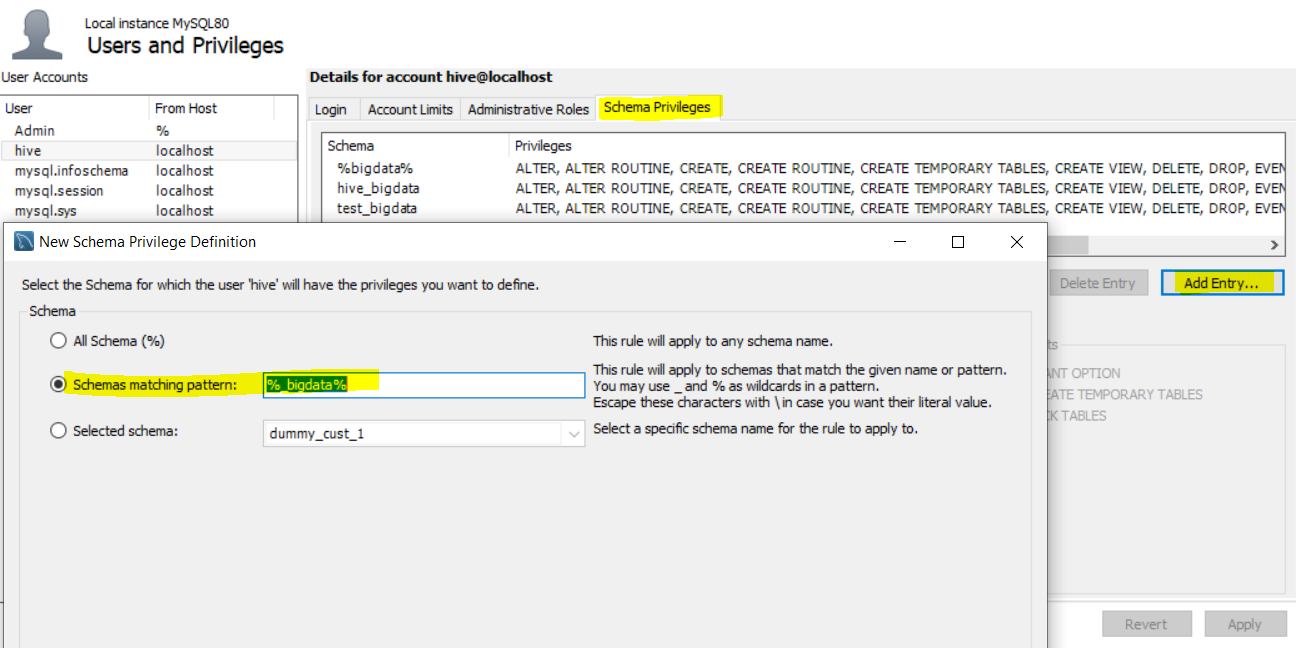
Now select Add Account option and Create an new user with Login Name as hive and Limit to Host Mapping as the localhost and Password of your choice.



Now we have to define the roles for this user under Administrative Roles and select DBManager ,DBDesigner and BackupAdmin Roles

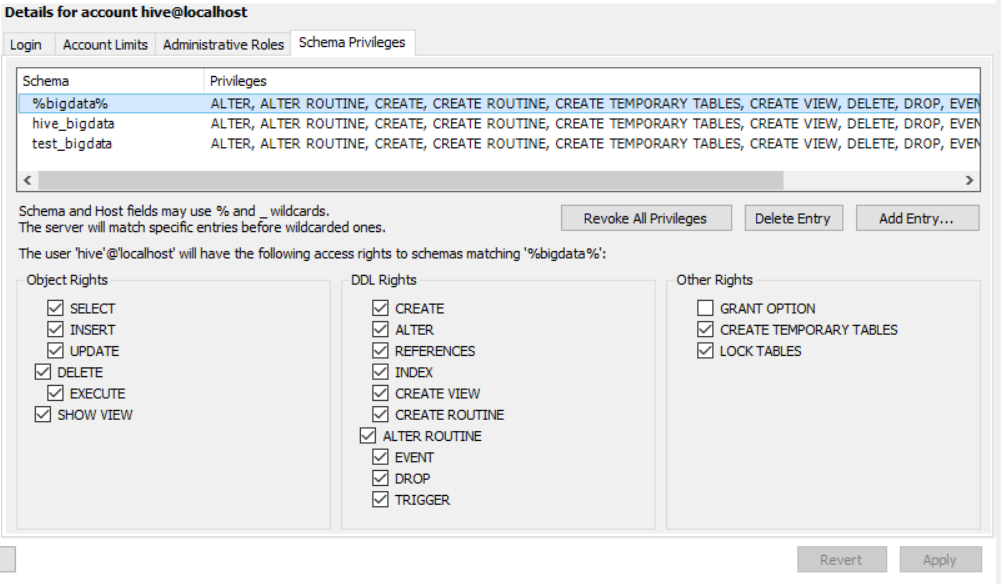


Now we need to grant schema privileges for the user by using Add Entry option and selecting the schemas we need access to.



I am using schema matching pattern as %\_bigdata% for all my bigdata related schemas. You can use other 2 options also.

After clicking OK we need to select All the privileges for this schema.

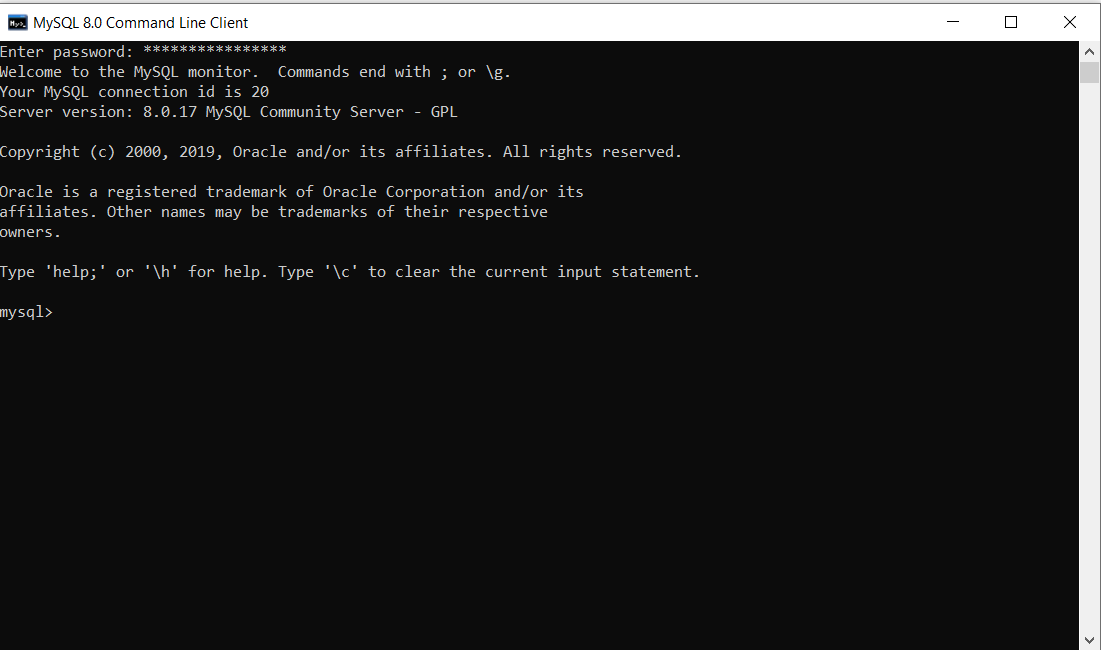


Click Apply and we are done with the creating Hive user.

4.5 Granting permission to Users

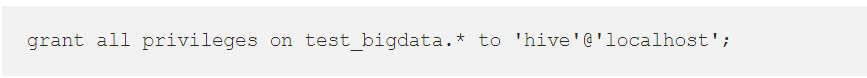
Once we have created the user hive the next step is to Grant All privileges to this user for all the Tables in the previously selected Schema.

Open the MySQL cmd Window. We can open it by using the Window’s Search bar.



Upon opening it will ask for your root user password(created while setting up MySQL).

Now we need to run the below command in the cmd window



where test\_bigdata will be you schema name and hive@localhost will be the user name @ Host name.

4.6 Creating Metastore

Now we need to create our own metastore for Hive in MySQL..

Firstly, we need to create a database for metastore in MySQL OR we can use the one which used in previous step test\_bigdata in my case.

Now Navigate to the below path

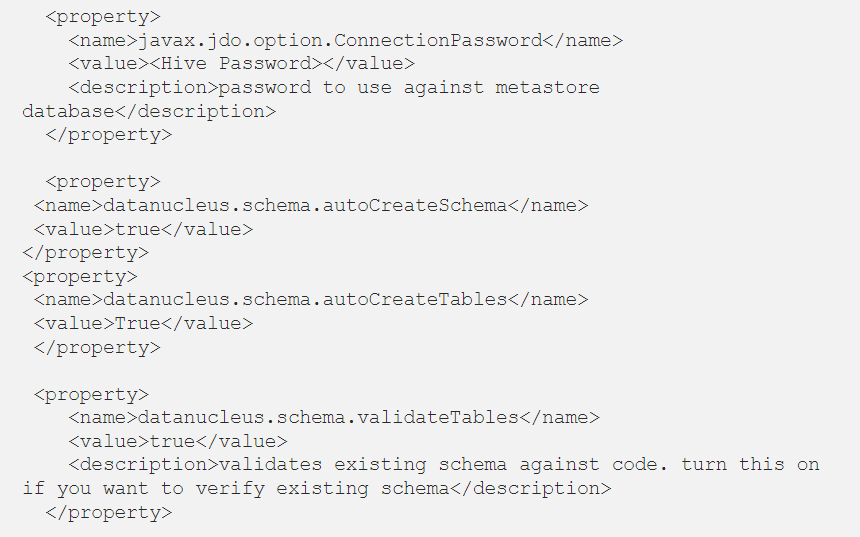
hive -> scripts -> metastore -> upgrade -> mysql and execute the file hive-schema-3.1.0.mysql in MySQL in your database.

Note:- If you are using different Database, select the folder for same in upgrade folder and execute the hive-schema file.

4.7 Adding Few More Properties(Metastore related Properties)

Finally, we need to open our hive-site.xml file once again and make some changes their, these are related to Hive metastore that’s why did not add them in starting so as to distinguish between the different set of properties.



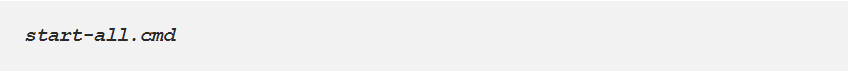


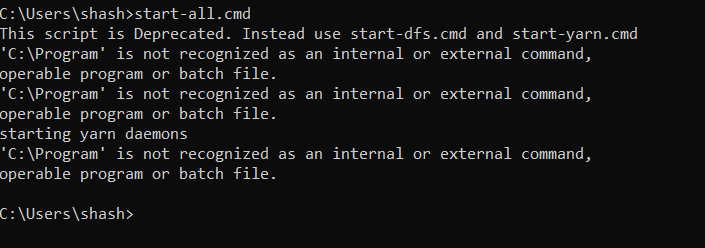
Replace the value for <Hive Password> with the hive user password that we created in MySQL user creation. And <Your Database> with the database that we used for metastore in MySQL.

5. Starting Hive

5.1 Starting Hadoop

Now we need to start a new Command Prompt remember to run it as administrator to avoid permission issues and execute below commands

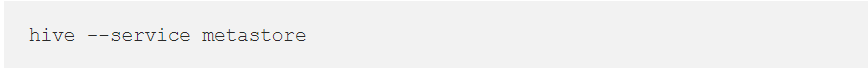


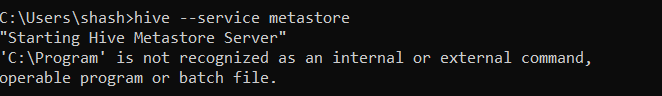


All the 4 daemons should be UP and running.

5.2 Starting Hive Metastore

Open a cmd window, run below command to start the Hive metastore.





5.3 Starting Hive

Now open a new cmd window and run the below command to start Hive

