**start hive**

->hive

-----hive started

hive>

hive>

hive>show databases;

hive>use adhoc;

---hive/mysql database location---

-> /var/lib/mysql

-> /usr/hive/warehouse

hive> create table student(id INT,name STRING, email STRING)

> row format delimited

> fields terminated by ','

> lines terminated by '\n'

> ;

hive> show tables;

hive> desc student;

**----->MYSQL --->9SQL)--->database ---> directory ---> table --> file**

hive> select \* from student;

**\*\*creating data set\*\***

->vi sql-dataset.py

----

#!/usr/bin/python3

import sys, time

#no of datasets

nod=sys.argv[1]

#this is string type data

print(type(nod))

#creating empty file to store data

f=open('student\_data.txt','w+')

#converting into int --typecast

num=int(nod)

#apply for loop

for i in range(num) :

print(i)

print('jack',+str(i))

print('jack'+str(i)+'@adhcocnw.org')

time.sleep(1)

f.write(str(i))

f.write(',')

f.write('jack'+str(i))

f.write(',')

f.write('jack'+str(i)+'@adhocnw.org')

f.write('\n')

#closing file

f.close()

**----enter the data----**

->python3 dataset.py 1000

-

--

->cat student\_data.txt

**\*\*store the data in hdfs\*\***

->hdfs dfs -ls/

->hdfs dfs -moveFromLocal student\_data.txt /hello1

->hdfs dfs -ls /hello1

**\*\*loading data from hdfs to HIVE\*\***

->hive

hive>show databases

hive>use adhoc

hive>load data inpath '/hello1/student\_data.txt' into table student;

hive> select count(email) from student;

---------------------------------------------------------------------------------------

**\*\*TASK\*\***

HIVE --> database --dir --- table ---

1-Data --extract from any of website

2-store into a file

3-upload into HDFS

make sure if it is related to hive then upload it

4-launch yarn job and count word

5-plot graph of counted data

------

**1.)**

Website data extract --- http/ https

web scrapping --- wiki, php.net

BeautifulSoup

->pip3 install bs4 #it is for BeautifulSoup

->pip3 install requests

->vim web\_scrape.py

--

#!/usr/bin/python3

Import requests

# to download source code and data of a website

web=requests.get(‘https://php.net’)

#print(dir(web))

textdata=web.txt

print(textdata)

#now here we can apply web-scraping by BeautifulSoup