

DATA SCIENCE LAB WORK

Connecting with hive:

```
Microsoft Windows [Version 10.0.19042.867]
(c) 2020 Microsoft Corporation. All rights reserved.

C:\Windows\system32>ssh agarwalshubham238226@f.cloudxlab.com
agarwalshubham238226@f.cloudxlab.com's password:
Permission denied, please try again.
agarwalshubham238226@f.cloudxlab.com's password:
Last failed login: Fri Mar 19 13:05:53 UTC 2021 from 117.225.64.155 on ssh:notty
There was 1 failed login attempt since the last successful login.
Last login: Thu Mar 18 09:13:23 2021 from 103.47.126.228
[agarwalshubham238226@cxln5 ~]$ hive_
```

```
[agarwalshubham238226@cxln5 ~]$ hive
log4j:WARN No such property [maxFileSize] in org.apache.log4j.DailyRollingFileAppender.

Logging initialized using configuration in file:/etc/hive/2.6.2.0-205/0/hive-log4j.properties
hive> _
```

Creating database:

```
hive> create database dist;
OK
Time taken: 0.341 seconds
hive>
```

Change to database dist:

```
hive> use dist;
OK
Time taken: 0.228 seconds
hive> _
```

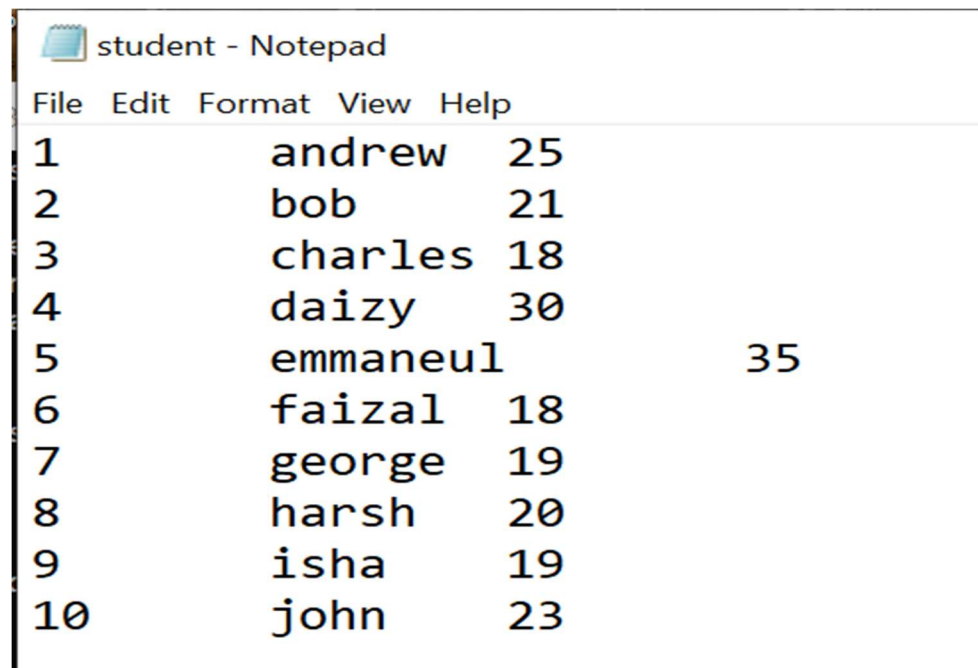
Creating table

```
Time taken: 0.1220 seconds
hive> create table student(rollno int, name string, age int)
> row format delimited
> fields terminated by '\t'
> lines terminated by '\n'
> stored as textfile;
OK
Time taken: 0.297 seconds
hive> _
```

Describing table student:

```
hive> describe student;
OK
rollno          int
name            string
age             int
Time taken: 0.479 seconds, Fetched: 3 row(s)
hive> _
```

Creating the data to load into table student: (fields terminated by \t and lines by \n)



1	andrew	25
2	bob	21
3	charles	18
4	daizy	30
5	emmaneul	35
6	faizal	18
7	george	19
8	harsh	20
9	isha	19
10	john	23

Loading data into table student:

```
hive> LOAD DATA LOCAL INPATH 'student.txt' OVERWRITE INTO TABLE student;
Loading data to table dist.student
Table dist.student stats: [numFiles=1, numRows=0, totalSize=123, rawDataSize=0]
OK
Time taken: 1.105 seconds
hive> _
```

Selecting data from table student:

```
hive> SELECT * FROM student;
OK
1      andrew    25
2      bob       21
3      charles   18
4      daizy     30
5      emmaneul      35
6      faizal    18
7      george    19
8      harsh     20
9      isha      19
10     john      23
Time taken: 0.43 seconds, Fetched: 10 row(s)
hive>
```

Inserting into table student with insert query:

```
hive> INSERT INTO student VALUES(11, 'kamal', 24);
Query ID = agarwalshubham238226_20210319132245_1610a84a-ba1e-4242-ba7f-5793b40a6b2f
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1607429201608_10530, Tracking URL = http://cxln2.c.thelab-240901.internal:8088/proxy/application_1607429201608_10530/
Kill Command = /usr/hdp/2.6.2.0-205/hadoop/bin/hadoop job -kill job_1607429201608_10530
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2021-03-19 13:22:53,930 Stage-1 map = 0%, reduce = 0%
2021-03-19 13:22:59,201 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.2 sec
MapReduce Total cumulative CPU time: 3 seconds 200 msec
Ended Job = job_1607429201608_10530
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://cxln1.c.thelab-240901.internal:8020/apps/hive/warehouse/dist.db/student/.hive-staging_hive_2021-03-19_13-22-45_568_3125588278972548516-1/-ext-10000
Loading data to table dist.student
Table dist.student stats: [numFiles=2, numRows=0, totalSize=135, rawDataSize=0]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 3.2 sec HDFS Read: 4651 HDFS Write: 80 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 200 msec
OK
Time taken: 15.382 seconds
```

Now, selecting all data of table student:

```
hive> SELECT * FROM student;
OK
11      kamal      24
1       andrew     25
2       bob        21
3       charles    18
4       daizy      30
5       emmaneul   35
6       faizal     18
7       george     19
8       harsh      20
9       isha       19
10      john       23
Time taken: 0.049 seconds, Fetched: 11 row(s)
hive>
```

Creating the table business for data analysis:

```
hive> create table business(Series_reference String,
> Period String,
> Data_value String,
> Suppressed String,
> STATUS String,
> UNITS String,
> Magnitude String,
> grp String,
> Series_title_1 String,
> Series_title_2 String,
> Series_title_3 String,
> Series_title_4 String,
> Series_title_5 String)
> row format delimited
> fields terminated by ','
> lines terminated by '\n'
> stored as textfile;
OK
Time taken: 0.183 seconds
hive>
```

Loading the csv file and overwriting it into business table:

```
hive> LOAD DATA LOCAL
> INPATH 'business.csv'
> OVERWRITE INTO TABLE business;
Loading data to table dist.business
Table dist.business stats: [numFiles=1, numRows=0, totalSize=1066318, rawDataSize=0]
OK
Time taken: 0.811 seconds
hive>
```

Counting how many tuples are there in table business:

```
hive> SELECT COUNT(1) FROM BUSINESS;
Query ID = agarwalshubham238226_20210319132807_21a313a4-36d8-42fa-8bd3-d03ddbb787eb
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1607429201608_10531, Tracking URL = http://cxln2.c.thelab-240901.internal:8088/proxy/application_1607429201608_10531/
Kill Command = /usr/hdp/2.6.2.0-205/hadoop/bin/hadoop job -kill job_1607429201608_10531
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-03-19 13:28:15,893 Stage-1 map = 0%, reduce = 0%
2021-03-19 13:28:22,157 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.97 sec
2021-03-19 13:28:28,415 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 5.93 sec
MapReduce Total cumulative CPU time: 5 seconds 930 msec
Ended Job = job_1607429201608_10531
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 5.93 sec HDFS Read: 1076326 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 930 msec
OK
12316
Time taken: 21.799 seconds, Fetched: 1 row(s)
hive> _
```

Describing the table business:

```
hive> DESCRIBE BUSINESS;
OK
series_reference      string
period                string
data_value            string
suppressed            string
status                string
units                 string
magnitude             string
grp                   string
series_title_1        string
series_title_2        string
series_title_3        string
series_title_4        string
series_title_5        string
Time taken: 0.378 seconds, Fetched: 13 row(s)
hive> _
```


Checking for max value in data_value in table business:

```
hive> SELECT MAX(DATA_VALUE)
> FROM BUSINESS;
Query ID = agarwalshubham238226_20210319133007_6ddd0977-3879-4fed-befe-2b9b98ba63c6
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1607429201608_10532, Tracking URL = http://cxln2.c.thelab-240901.internal:8088/proxy/application_1607429201608_10532/
Kill Command = /usr/hdp/2.6.2.0-205/hadoop/bin/hadoop job -kill job_1607429201608_10532
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-03-19 13:30:15,714 Stage-1 map = 0%, reduce = 0%
2021-03-19 13:30:25,150 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.48 sec
2021-03-19 13:30:31,394 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.34 sec
MapReduce Total cumulative CPU time: 6 seconds 340 msec
Ended Job = job_1607429201608_10532
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.34 sec HDFS Read: 1076342 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 340 msec
OK
99941
Time taken: 24.98 seconds, Fetched: 1 row(s)
hive>
```

Counting number of tuples having status as 'F':

```
hive> SELECT COUNT(1) FROM BUSINESS WHERE STATUS='F';
Query ID = agarwalshubham238226_20210319133332_f5cd5f12-e8fa-4101-8cb8-147ba925e0cb
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1607429201608_10533, Tracking URL = http://cxln2.c.thelab-240901.internal:8088/proxy/application_1607429201608_10533/
Kill Command = /usr/hdp/2.6.2.0-205/hadoop/bin/hadoop job -kill job_1607429201608_10533
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-03-19 13:33:38,958 Stage-1 map = 0%, reduce = 0%
2021-03-19 13:33:46,246 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.42 sec
2021-03-19 13:33:51,452 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 7.78 sec
MapReduce Total cumulative CPU time: 7 seconds 780 msec
Ended Job = job_1607429201608_10533
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 7.78 sec HDFS Read: 1077292 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 780 msec
OK
11030
Time taken: 20.628 seconds, Fetched: 1 row(s)
```

Python Program to connect hive:

```
from pyhive import hive

host_name = "f.cloudxlab.com"
port = 10000
user = "agarwalshubham238226"
password = ""
database = "dist"

def hiveconnection(host_name, port, user,password, database):
    conn = hive.Connection(host=host_name, port=port,
username=user, password=password,
                                database=database, auth='CUSTOM')

    cur = conn.cursor()
    cur.execute('select * from business limit 5')
    result = cur.fetchall()

    return result

# Call above function
output = hiveconnection(host_name, port, user,password,
database)
print(output)
```