Ex no: 5 210701018

Create tables in Hive and write queries to access the data in the table

Aim:

To create tables in Hive and write queries to access the data in the table.

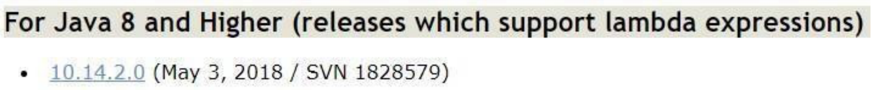
Procedure:

Hive Download and installation:

1. Hive Installation setup:

- Download and install Apache Derby version 10.14.2.0:

https://db.apache.org/derby/derby_downloads.html#For+Java+8+and+Higher

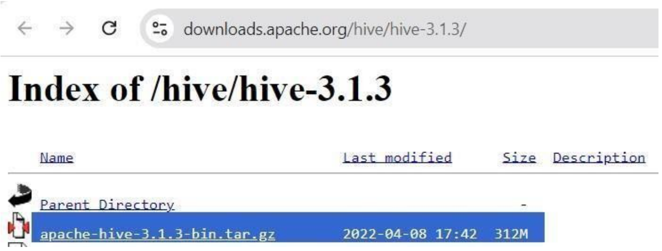


For Java 8 and Higher (releases which support lambda expressions)

- 10.14.2.0 (May 3, 2018 / SVN 1828579)

-Download and install Apache Hive version 3.1.3:

https://downloads.apache.org/hive/hive-3.1.3/



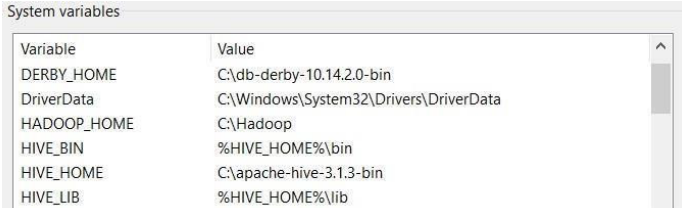
downloads.apache.org/hive/hive-3.1.3/

Index of /hive/hive-3.1.3

Name	Last modified	Size	Description
Parent Directory		-	
apache-hive-3.1.3-bin.tar.gz	2022-04-08 17:42	312M	

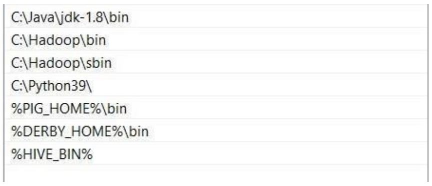
2. Add environment variables:

Environment variables > System variables > Add the below paths



System variables

Variable	Value
DERBY_HOME	C:\db-derby-10.14.2.0-bin
DriverData	C:\Windows\System32\Drivers\DriverData
HADOOP_HOME	C:\Hadoop
HIVE_BIN	%HIVE_HOME%\bin
HIVE_HOME	C:\apache-hive-3.1.3-bin
HIVE_LIB	%HIVE_HOME%\lib



```
C:\Java\jdk-1.8\bin
C:\Hadoop\bin
C:\Hadoop\sbin
C:\Python39\
%PIG_HOME%\bin
%DERBY_HOME%\bin
%HIVE_BIN%
```

-> (Inside Path)

3. Copy Derby libraries:

Go to the Derby libraries directory (db-derby-10.14.2.0\lib) and copy all *.jar files. Then, paste them within the Hive libraries directory.

4. Configuring hive-site.xml and Hive's Bin folder:

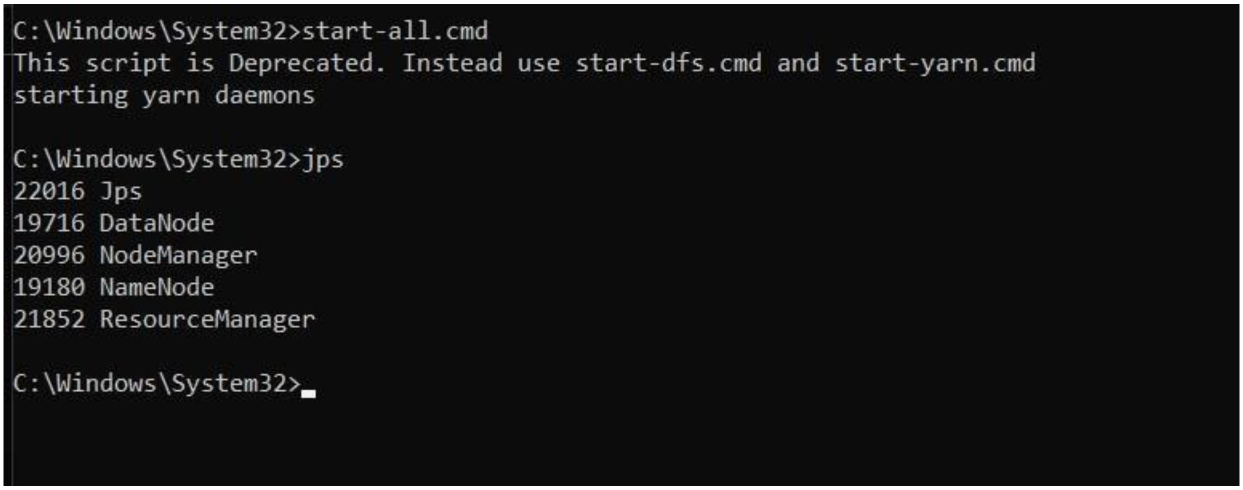
Refer following link to download the file. Also download the guava file. Put hive-site.xml file to hive's conf location and replace hive's current guava file with this one in lib location. Also download the bin folder from link and replace the existing hive's bin folder.

https://1drv.ms/f/s!ArSg3Xpur4Grmw0SDqW0g44T7HYU?e=wDsoBn

5. Starting Hadoop Services

Open PowerShell as administrator and go to Hadoop sbin directory and start hadoop services using the following commands:

```
Start-all.cmd
```

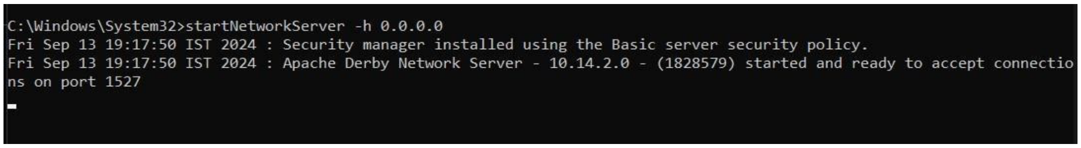


```
C:\Windows\System32>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

C:\Windows\System32>jps
22016 Jps
19716 DataNode
20996 NodeManager
19180 NameNode
21852 ResourceManager

C:\Windows\System32>
```

6. Derby Network Server:

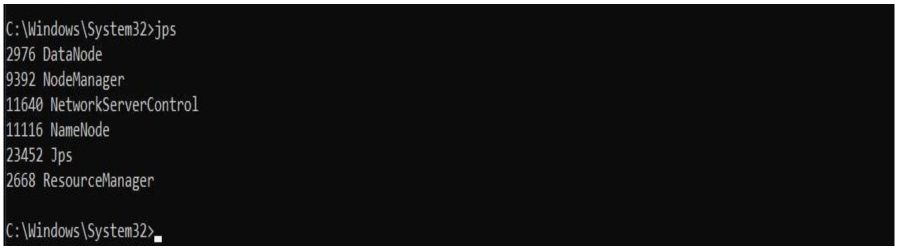
Open another PowerShell window and run the following command to open Derby:

StartNetworkServer -h 0.0.0.0



```
C:\Windows\System32>startNetworkServer -h 0.0.0.0
Fri Sep 13 19:17:50 IST 2024 : Security manager installed using the Basic server security policy.
Fri Sep 13 19:17:50 IST 2024 : Apache Derby Network Server - 10.14.2.0 - (1828579) started and ready to accept connectio
ns on port 1527
```

Go to first PowerShell window and check whether NetworkServerControl is running.

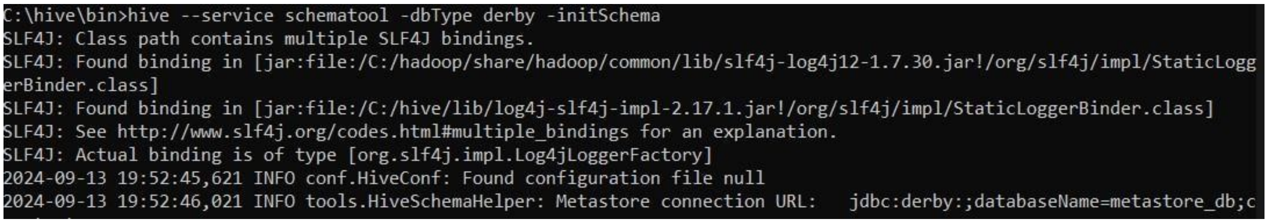


```
C:\Windows\System32>jps
2976 DataNode
9392 NodeManager
11640 NetworkServerControl
11116 NameNode
23452 Jps
2668 ResourceManager

C:\Windows\System32>
```

7. Starting Apache Hive:

Go to Apache Hive’s bin location with cd command and run the following command:

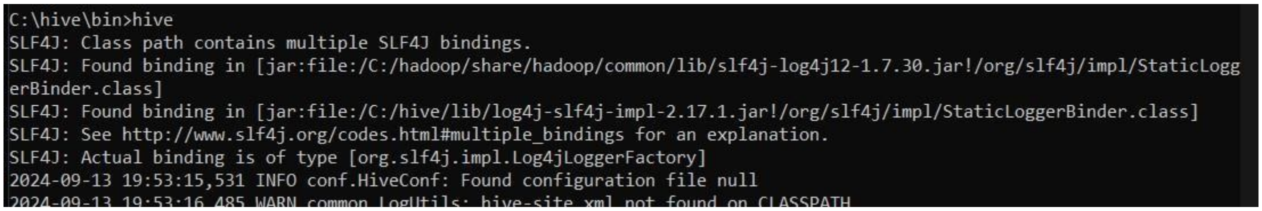
hive --service schematool -dbType derby –initSchema



```
C:\hive\bin>hive --service schematool -dbType derby -initSchema
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/C:/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.30.jar!/org/slf4j/impl/StaticLogg
erBinder.class]
SLF4J: Found binding in [jar:file:/C:/hive/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2024-09-13 19:52:45,621 INFO conf.HiveConf: Found configuration file null
2024-09-13 19:52:46,021 INFO tools.HiveSchemaHelper: Metastore connection URL:   jdbc:derby:;databaseName=metastore_db;c
```

8. Open Hive shell by typing:

hive



```
C:\hive\bin>hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/C:/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.30.jar!/org/slf4j/impl/StaticLogg
erBinder.class]
SLF4J: Found binding in [jar:file:/C:/hive/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2024-09-13 19:53:15,531 INFO conf.HiveConf: Found configuration file null
2024-09-13 19:53:16,485 WARN common.LogUtils: hive-site.xml not found on CLASSPATH
```

Create a Database:

Start by creating a database. Open the Hive CLI and follow the steps below:

1. Use the **CREATE DATABASE** statement to create a new database:

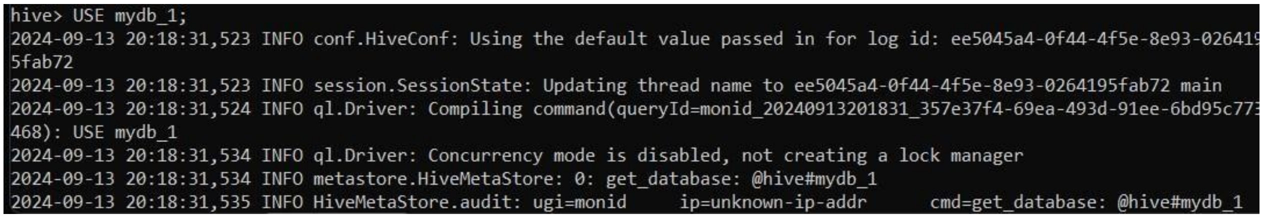
```
CREATE DATABASE mydb_1;
```

2. Verify the database is present:

```
SHOW DATABASES;
```

3. Switch to the new database:

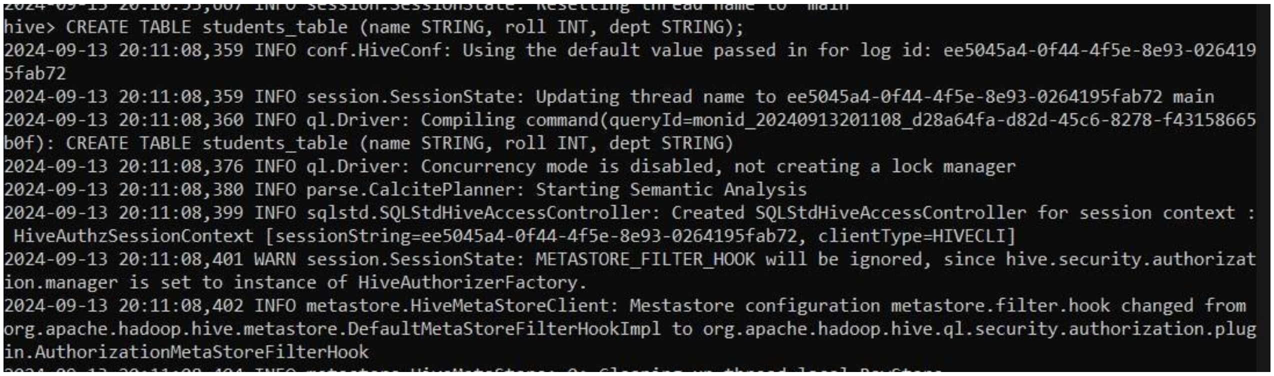
```
USE mydata;
```



```
hive> USE mydb_1;
2024-09-13 20:18:31,523 INFO conf.HiveConf: Using the default value passed in for log id: ee5045a4-0f44-4f5e-8e93-026419
5fab72
2024-09-13 20:18:31,523 INFO session.SessionState: Updating thread name to ee5045a4-0f44-4f5e-8e93-0264195fab72 main
2024-09-13 20:18:31,524 INFO ql.Driver: Compiling command(queryId=monid_20240913201831_357e37f4-69ea-493d-91ee-6bd95c773
468): USE mydb_1
2024-09-13 20:18:31,534 INFO ql.Driver: Concurrency mode is disabled, not creating a lock manager
2024-09-13 20:18:31,534 INFO metastore.HiveMetaStore: 0: get_database: @hive#mydb_1
2024-09-13 20:18:31,535 INFO HiveMetaStore.audit: ugi=monid     ip=unknown-ip-addr      cmd=get_database: @hive#mydb_1
```

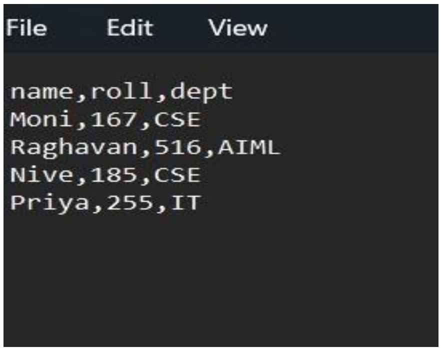
Create a Table in Hive:

```
CREATE TABLE students_table (name STRING, roll INT, dept STRING);
```



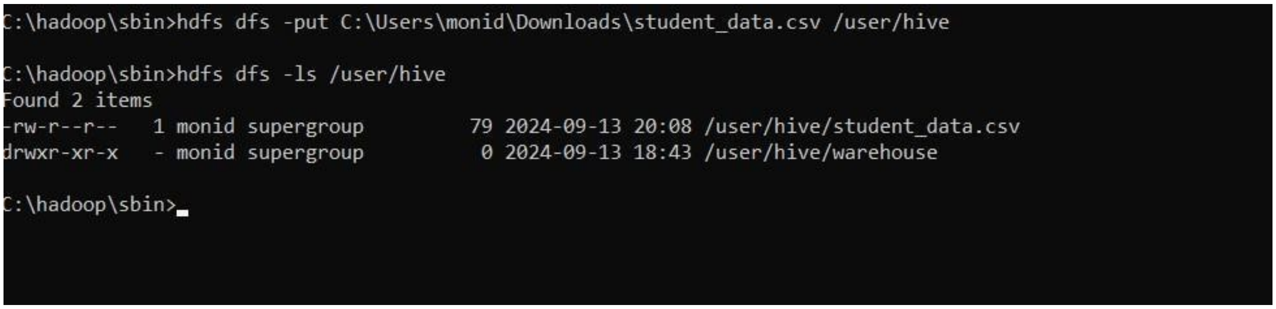
```
2024-09-13 20:10:53,607 INFO session.SessionState: Resetting thread name to  main
hive> CREATE TABLE students_table (name STRING, roll INT, dept STRING);
2024-09-13 20:11:08,359 INFO conf.HiveConf: Using the default value passed in for log id: ee5045a4-0f44-4f5e-8e93-026419
5fab72
2024-09-13 20:11:08,359 INFO session.SessionState: Updating thread name to ee5045a4-0f44-4f5e-8e93-0264195fab72 main
2024-09-13 20:11:08,360 INFO ql.Driver: Compiling command(queryId=monid_20240913201108_d28a64fa-d82d-45c6-8278-f43158665
b0f): CREATE TABLE students_table (name STRING, roll INT, dept STRING)
2024-09-13 20:11:08,376 INFO ql.Driver: Concurrency mode is disabled, not creating a lock manager
2024-09-13 20:11:08,380 INFO parse.CalcitePlanner: Starting Semantic Analysis
2024-09-13 20:11:08,399 INFO sqlstd.SQLStdHiveAccessController: Created SQLStdHiveAccessController for session context :
 HiveAuthzSessionContext [sessionString=ee5045a4-0f44-4f5e-8e93-0264195fab72, clientType=HIVECLI]
2024-09-13 20:11:08,401 WARN session.SessionState: METASTORE_FILTER_HOOK will be ignored, since hive.security.authorizat
ion.manager is set to instance of HiveAuthorizerFactory.
2024-09-13 20:11:08,402 INFO metastore.HiveMetaStoreClient: Mestastore configuration metastore.filter.hook changed from
org.apache.hadoop.hive.metastore.DefaultMetaStoreFilterHookImpl to org.apache.hadoop.hive.ql.security.authorization.plug
in.AuthorizationMetaStoreFilterHook
```

Create a .csv file:



```
name,roll,dept
Moni,167,CSE
Raghavan,516,AIML
Nive,185,CSE
Priya,255,IT
```

Add it to hadoop using –put command:

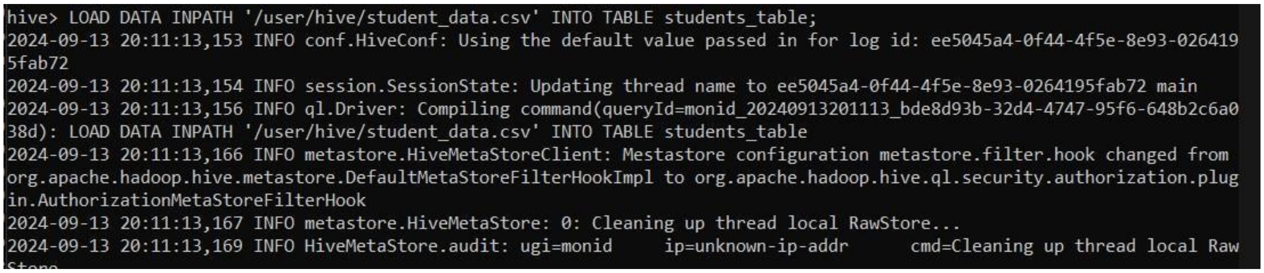


```
C:\hadoop\sbin>hdfs dfs -put C:\Users\monid\Downloads\student_data.csv /user/hive

C:\hadoop\sbin>hdfs dfs -ls /user/hive
Found 2 items
-rw-r--r--   1 monid supergroup         79 2024-09-13 20:08 /user/hive/student_data.csv
drwxr-xr-x   - monid supergroup          0 2024-09-13 18:43 /user/hive/warehouse

C:\hadoop\sbin>
```

Add Data to the TABLE:

Run the **LOAD DATA LOCAL INPATH** command:

LOAD DATA INPATH '/user/hive/student_data.csv' INTO TABLE students_table;

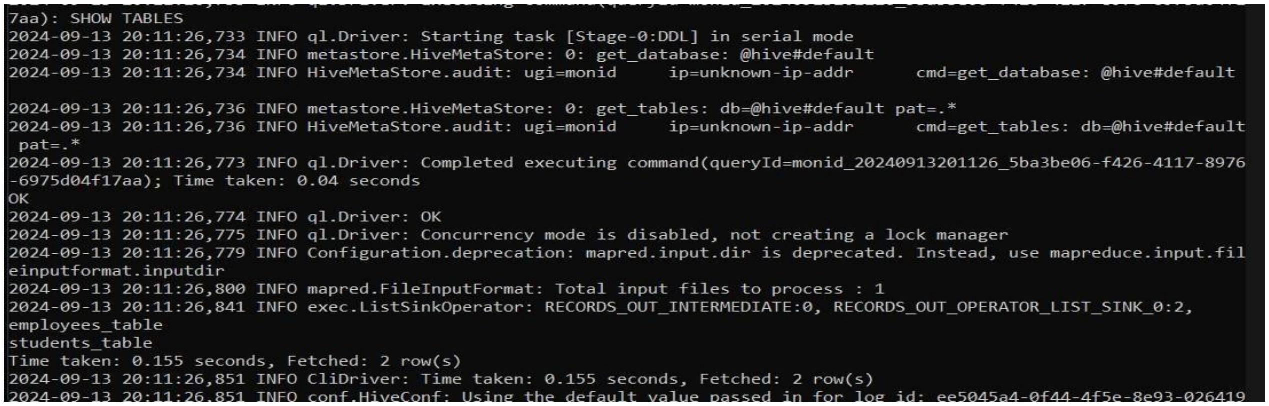


```
hive> LOAD DATA INPATH '/user/hive/student_data.csv' INTO TABLE students_table;
2024-09-13 20:11:13,153 INFO conf.HiveConf: Using the default value passed in for log id: ee5045a4-0f44-4f5e-8e93-026419
5fab72
2024-09-13 20:11:13,154 INFO session.SessionState: Updating thread name to ee5045a4-0f44-4f5e-8e93-0264195fab72 main
2024-09-13 20:11:13,156 INFO ql.Driver: Compiling command(queryId=monid_20240913201113_bde8d93b-32d4-4747-95f6-648b2c6a0
38d): LOAD DATA INPATH '/user/hive/student_data.csv' INTO TABLE students_table
2024-09-13 20:11:13,166 INFO metastore.HiveMetaStoreClient: Mestastore configuration metastore.filter.hook changed from
org.apache.hadoop.hive.metastore.DefaultMetaStoreFilterHookImpl to org.apache.hadoop.hive.ql.security.authorization.plug
in.AuthorizationMetaStoreFilterHook
2024-09-13 20:11:13,167 INFO metastore.HiveMetaStore: 0: Cleaning up thread local RawStore...
2024-09-13 20:11:13,169 INFO HiveMetaStore.audit: ugi=monid     ip=unknown-ip-addr      cmd=Cleaning up thread local Raw
Store
```

List Hive Tables and Data:

To show all tables in a selected database, use the following statement:

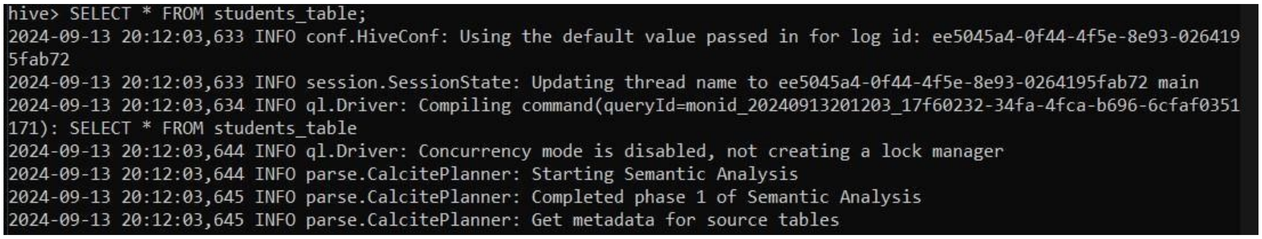
SHOW TABLES;



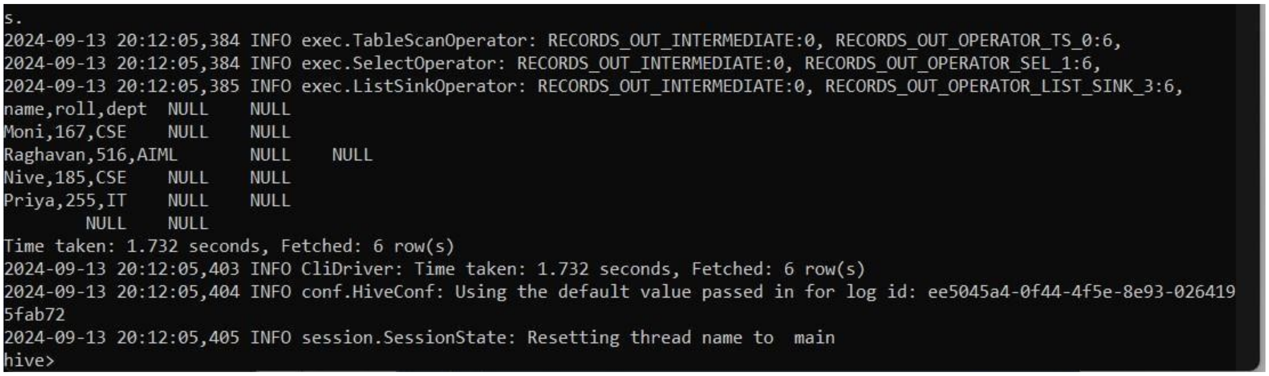
```
7aa): SHOW TABLES
2024-09-13 20:11:26,733 INFO ql.Driver: Starting task [Stage-0:DDL] in serial mode
2024-09-13 20:11:26,734 INFO metastore.HiveMetaStore: 0: get_database: @hive#default
2024-09-13 20:11:26,734 INFO HiveMetaStore.audit: ugi=monid     ip=unknown-ip-addr      cmd=get_database: @hive#default

2024-09-13 20:11:26,736 INFO metastore.HiveMetaStore: 0: get_tables: db=@hive#default pat=.*
2024-09-13 20:11:26,736 INFO HiveMetaStore.audit: ugi=monid     ip=unknown-ip-addr      cmd=get_tables: db=@hive#default
 pat=.*
2024-09-13 20:11:26,773 INFO ql.Driver: Completed executing command(queryId=monid_20240913201126_5ba3be06-f426-4117-8976
-6975d04f17aa); Time taken: 0.04 seconds
OK
2024-09-13 20:11:26,774 INFO ql.Driver: OK
2024-09-13 20:11:26,775 INFO ql.Driver: Concurrency mode is disabled, not creating a lock manager
2024-09-13 20:11:26,779 INFO Configuration.deprecation: mapred.input.dir is deprecated. Instead, use mapreduce.input.fil
einputformat.inputdir
2024-09-13 20:11:26,800 INFO mapred.FileInputFormat: Total input files to process : 1
2024-09-13 20:11:26,841 INFO exec.ListSinkOperator: RECORDS_OUT_INTERMEDIATE:0, RECORDS_OUT_OPERATOR_LIST_SINK_0:2,
employees_table
students_table
Time taken: 0.155 seconds, Fetched: 2 row(s)
2024-09-13 20:11:26,851 INFO CliDriver: Time taken: 0.155 seconds, Fetched: 2 row(s)
2024-09-13 20:11:26,851 INFO conf.HiveConf: Using the default value passed in for log id: ee5045a4-0f44-4f5e-8e93-026419
```

To display table data, use a **SELECT** statement. For example, to select everything in a table, run:

SELECT * FROM students;



```
hive> SELECT * FROM students_table;
2024-09-13 20:12:03,633 INFO conf.HiveConf: Using the default value passed in for log id: ee5045a4-0f44-4f5e-8e93-026419
5fab72
2024-09-13 20:12:03,633 INFO session.SessionState: Updating thread name to ee5045a4-0f44-4f5e-8e93-0264195fab72 main
2024-09-13 20:12:03,634 INFO ql.Driver: Compiling command(queryId=monid_20240913201203_17f60232-34fa-4fca-b696-6cfaf0351
171): SELECT * FROM students_table
2024-09-13 20:12:03,644 INFO ql.Driver: Concurrency mode is disabled, not creating a lock manager
2024-09-13 20:12:03,644 INFO parse.CalcitePlanner: Starting Semantic Analysis
2024-09-13 20:12:03,645 INFO parse.CalcitePlanner: Completed phase 1 of Semantic Analysis
2024-09-13 20:12:03,645 INFO parse.CalcitePlanner: Get metadata for source tables
```



```
s.
2024-09-13 20:12:05,384 INFO exec.TableScanOperator: RECORDS_OUT_INTERMEDIATE:0, RECORDS_OUT_OPERATOR_TS_0:6,
2024-09-13 20:12:05,384 INFO exec.SelectOperator: RECORDS_OUT_INTERMEDIATE:0, RECORDS_OUT_OPERATOR_SEL_1:6,
2024-09-13 20:12:05,385 INFO exec.ListSinkOperator: RECORDS_OUT_INTERMEDIATE:0, RECORDS_OUT_OPERATOR_LIST_SINK_3:6,
name,roll,dept  NULL    NULL
Moni,167,CSE    NULL    NULL
Raghavan,516,AIML       NULL    NULL
Nive,185,CSE    NULL    NULL
Priya,255,IT    NULL    NULL
        NULL    NULL
Time taken: 1.732 seconds, Fetched: 6 row(s)
2024-09-13 20:12:05,403 INFO CliDriver: Time taken: 1.732 seconds, Fetched: 6 row(s)
2024-09-13 20:12:05,404 INFO conf.HiveConf: Using the default value passed in for log id: ee5045a4-0f44-4f5e-8e93-026419
5fab72
2024-09-13 20:12:05,405 INFO session.SessionState: Resetting thread name to  main
hive>
```

Result:

Thus, to create tables in Hive and write queries to access the data in the table was completed successfully.