Ex no: 5 **210701143**

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)

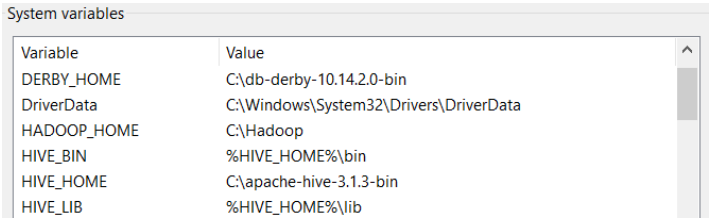
- 10.14.2.0 (May 3, 2018 / SVN 1828579)

-Download and install Apache Hive version 3.1.3:

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

2. Add environment variables:

Environment variables > System variables > Add the below paths



-> (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-dfs.cmd

start-yarn.cmd

```
PS C:\Windows\system32> cd C:\Hadoop\sbin
PS C:\Hadoop\sbin> start-dfs.cmd
PS C:\Hadoop\sbin> start-yarn.cmd
starting yarn daemons
PS C:\Hadoop\sbin> jps
8080 NameNode
11572 NodeManager
11484 Jps
3596 DataNode
7180 ResourceManager
```

6. Derby Network Server:

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

StartNetworkServer -h 0.0.0.0

```
PS C:\Windows\system32> StartNetworkServer -h 0.0.0.0
Sat Aug 31 20:11:02 IST 2024 : Security manager installed using the Basic server security policy.
Sat Aug 31 20:11:07 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.

```
PS C:\Hadoop\sbin> jps
12480 NetworkServerControl
8080 NameNode
11572 NodeManager
12180 Jps
3596 DataNode
7180 ResourceManager
```

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
```

```
PS C:\Hadoop\sbin> cd C:\apache-hive-3.1.3-bin\bin
PS C:\apache-hive-3.1.3-bin\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-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/C:/apache-hive-3.1.3-bin/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.Reload4jLoggerFactory]
2024-08-31 20:12:45,641 INFO conf.HiveConf: Found configuration file null
2024-08-31 20:12:46,492 INFO tools.HiveSchemaHelper: Metastore connection URL:   jdbc:derby:;databaseName=metastore_db;create=true
Metastore connection URL:        jdbc:derby:;databaseName=metastore_db;create=true
2024-08-31 20:12:46,494 INFO tools.HiveSchemaHelper: Metastore Connection Driver :       org.apache.derby.jdbc.EmbeddedDriver
Metastore Connection Driver :    org.apache.derby.jdbc.EmbeddedDriver
2024-08-31 20:12:46,495 INFO tools.HiveSchemaHelper: Metastore connection User:  APP
Metastore connection User:       APP
Starting metastore schema initialization to 3.1.0
Initialization script hive-schema-3.1.0.derby.sql
```

```
Initialization script completed
schemaTool completed
```

8. Open Hive shell by typing:

```
hive
```

```
PS C:\apache-hive-3.1.3-bin\bin> hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/C:/Hadoop/share/hadoop/common/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/C:/apache-hive-3.1.3-bin/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.Reload4jLoggerFactory]
2024-08-31 20:13:15,204 INFO conf.HiveConf: Found configuration file null
2024-08-31 20:13:18,554 WARN common.LogUtils: hive-site.xml not found on CLASSPATH
Hive Session ID = 272282ae-ff6f-4567-bab6-f339170eaaea
2024-08-31 20:13:18,670 INFO SessionState: Hive Session ID = 272282ae-ff6f-4567-bab6-f339170eaaea
```

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 mydata;
```

```
hive> CREATE DATABASE mydata;
2024-08-31 20:14:50,553 INFO conf.HiveConf: Using the default value passed in for log id: 272282ae-ff6f-4567-bab6-f339170eaaea
2024-08-31 20:14:50,793 INFO ql.Driver: Compiling command(queryId=Admin_20240831201450_7bc12345-452d-4cbc-90e6-fd90d0b13853): CREATE DATABASE mydata
2024-08-31 20:14:51,713 INFO ql.Driver: Concurrency mode is disabled, not creating a lock manager
2024-08-31 20:14:51,746 INFO ql.Driver: Semantic Analysis Completed (retrial = false)
2024-08-31 20:14:51,754 INFO ql.Driver: Returning Hive schema: Schema(fieldSchemas:null, properties:null)
```

2. Verify the database is present:

```
SHOW DATABASES;
```

```
hive> SHOW DATABASES;
2024-08-31 20:15:10,184 INFO conf.HiveConf: Using the default value passed in for log
2024-08-31 20:15:10,185 INFO session.SessionState: Updating thread name to 272282ae-ff
2024-08-31 20:15:10,187 INFO ql.Driver: Compiling command(queryId=Admin_20240831201510
2024-08-31 20:15:10,210 INFO ql.Driver: Concurrency mode is disabled, not creating a l
2024-08-31 20:15:10,226 INFO ql.Driver: Semantic Analysis Completed (retrial = false)
2024-08-31 20:15:10,366 INFO ql.Driver: Returning Hive schema: Schema(fieldSchemas:[Fi
2024-08-31 20:15:10,555 INFO exec.ListSinkOperator: Initializing operator LIST_SINK[0]
2024-08-31 20:15:10,574 INFO ql.Driver: Completed compiling command(queryId=Admin_2024
2024-08-31 20:15:10,574 INFO reexec.ReExecDriver: Execution #1 of query
2024-08-31 20:15:10,575 INFO ql.Driver: Concurrency mode is disabled, not creating a l
2024-08-31 20:15:10,575 INFO ql.Driver: Executing command(queryId=Admin_20240831201510
2024-08-31 20:15:10,576 INFO ql.Driver: Starting task [Stage-0:DDL] in serial mode
2024-08-31 20:15:10,578 INFO metastore.HiveMetaStore: 0: get_databases: @hive#
2024-08-31 20:15:10,579 INFO HiveMetaStore.audit: ugi=Admin     ip=unknown-ip-addr

2024-08-31 20:15:10,594 INFO exec.DDLTask: results : 2
2024-08-31 20:15:10,705 INFO ql.Driver: Completed executing command(queryId=Admin_2024
OK
2024-08-31 20:15:10,708 INFO ql.Driver: OK
2024-08-31 20:15:10,710 INFO ql.Driver: Concurrency mode is disabled, not creating a l
2024-08-31 20:15:10,743 INFO Configuration.deprecation: mapred.input.dir is deprecated
2024-08-31 20:15:10,848 INFO mapred.FileInputFormat: Total input files to process : 1
2024-08-31 20:15:10,946 INFO exec.ListSinkOperator: RECORDS_OUT_INTERMEDIATE:0, RECORD
default
mydata
```

3. Switch to the new database:

```
USE mydata;
```

```
hive> USE mydata;
2024-08-31 20:15:47,457 INFO conf.HiveConf: Using the default value passed in for log id: 272282ae-ff6f-4567-bab6-f339170eaaea
2024-08-31 20:15:47,457 INFO session.SessionState: Updating thread name to 272282ae-ff6f-4567-bab6-f339170eaaea main
2024-08-31 20:15:47,460 INFO ql.Driver: Compiling command(queryId=Admin_20240831201547_c04c3733-1ee9-4ebb-a2c9-3b28d8b9ecf7): USE mydata
2024-08-31 20:15:47,482 INFO ql.Driver: Concurrency mode is disabled, not creating a lock manager
2024-08-31 20:15:47,482 INFO metastore.HiveMetaStore: 0: get_database: @hive#mydata
2024-08-31 20:15:47,482 INFO HiveMetaStore.audit: ugi=Admin     ip=unknown-ip-addr      cmd=get_database: @hive#mydata
```

Create a Table in Hive:

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

```
hive> CREATE TABLE students_table(name STRING, roll INT, dept STRING);
2024-08-31 22:34:29,463 INFO conf.HiveConf: Using the default value passed in for log id: 02b967f6-c082-40d3-a79
bf2de5
2024-08-31 22:34:29,464 INFO session.SessionState: Updating thread name to 02b967f6-c082-40d3-a797-6db099bf2de5
2024-08-31 22:34:29,466 INFO ql.Driver: Compiling command(queryId=Admin_20240831223429_082a0028-2ab3-4ce6-8d36-0
676): CREATE TABLE students_table(name STRING, roll INT, dept STRING)
2024-08-31 22:34:29,492 INFO ql.Driver: Concurrency mode is disabled, not creating a lock manager
2024-08-31 22:34:29,493 INFO parse.CalcitePlanner: Starting Semantic Analysis
2024-08-31 22:34:29,496 INFO parse.CalcitePlanner: Creating table mydata.students_table position=13
2024-08-31 22:34:29,501 INFO metastore.HiveMetaStore: 0: get_database: @hive#mydata
2024-08-31 22:34:29,501 INFO HiveMetaStore.audit: ugi=Admin     ip=unknown-ip-addr      cmd=get_database: @hive#m

2024-08-31 22:34:29,507 INFO ql.Driver: Semantic Analysis Completed (retrial = false)
2024-08-31 22:34:29,508 INFO ql.Driver: Returning Hive schema: Schema(fieldSchemas:null, properties:null)
2024-08-31 22:34:29,508 INFO ql.Driver: Completed compiling command(queryId=Admin_20240831223429_082a0028-2ab3-4
-0f67e321d676); Time taken: 0.042 seconds
```

Add it to hadoop using –put command:

```
PS C:\Hadoop\sbin> hdfs dfs -put C:/Users/Admin/Documents/Hive/student_data.csv /user/hive
PS C:\Hadoop\sbin> hdfs dfs -ls /user/hive
Found 2 items
-rw-r--r--   1 Admin supergroup         87 2024-08-31 22:23 /user/hive/student_data.csv
```

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;
2024-08-31 22:40:35,595 INFO conf.HiveConf: Using the default value passed in for log id: 94d05ec8-
7bf6a2
2024-08-31 22:40:35,595 INFO session.SessionState: Updating thread name to 94d05ec8-002c-4d94-9297-
2024-08-31 22:40:35,598 INFO ql.Driver: Compiling command(queryId=Admin_20240831224035_83fccc17-867
318): LOAD DATA INPATH '/user/hive/student_data.csv' INTO TABLE students_table
2024-08-31 22:40:35,618 INFO ql.Driver: Concurrency mode is disabled, not creating a lock manager
2024-08-31 22:40:35,619 INFO metastore.HiveMetaStore: 0: get_table : tbl=hive.mydata.students_table
2024-08-31 22:40:35,619 INFO HiveMetaStore.audit: ugi=Admin     ip=unknown-ip-addr      cmd=get_tab
students_table
2024-08-31 22:40:36,187 INFO ql.Driver: Semantic Analysis Completed (retrial = false)
2024-08-31 22:40:36,188 INFO ql.Driver: Returning Hive schema: Schema(fieldSchemas:null, properties
2024-08-31 22:40:36,188 INFO ql.Driver: Completed compiling command(queryId=Admin_20240831224035_83
-a5fd93a24318); Time taken: 0.59 seconds
2024-08-31 22:40:36,188 INFO reexec.ReExecDriver: Execution #1 of query
2024-08-31 22:40:36,188 INFO ql.Driver: Concurrency mode is disabled, not creating a lock manager
2024-08-31 22:40:36,189 INFO ql.Driver: Executing command(queryId=Admin_20240831224035_83fccc17-867
318): LOAD DATA INPATH '/user/hive/student_data.csv' INTO TABLE students_table
2024-08-31 22:40:36,189 INFO ql.Driver: Starting task [Stage-0:MOVE] in serial mode
2024-08-31 22:40:36,190 INFO metastore.HiveMetaStore: 0: Cleaning up thread local RawStore...
```

List Hive Tables and Data:

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

```
SHOW TABLES;
```

```
hive> SHOW TABLES;
2024-08-31 22:37:47,100 INFO conf.HiveConf: Using the default value passed in for log
bf2de5
2024-08-31 22:37:47,100 INFO session.SessionState: Updating thread name to 02b967f6-c
2024-08-31 22:37:47,115 INFO ql.Driver: Compiling command(queryId=Admin_2024083122374
dfd): SHOW TABLES
2024-08-31 22:37:47,145 INFO ql.Driver: Concurrency mode is disabled, not creating a
2024-08-31 22:37:47,148 INFO metastore.HiveMetaStore: 0: get_database: @hive#mydata
2024-08-31 22:37:47,149 INFO HiveMetaStore.audit: ugi=Admin     ip=unknown-ip-addr

2024-08-31 22:37:47,151 INFO ql.Driver: Semantic Analysis Completed (retrial = false)
2024-08-31 22:37:47,152 INFO ql.Driver: Returning Hive schema: Schema(fieldSchemas:[F
ng, comment:from deserializer)], properties:null)
2024-08-31 22:37:47,155 INFO exec.ListSinkOperator: Initializing operator LIST_SINK[0
2024-08-31 22:37:47,156 INFO ql.Driver: Completed compiling command(queryId=Admin_202
-cbce9f8d2dfd); Time taken: 0.041 seconds
2024-08-31 22:37:47,156 INFO reexec.ReExecDriver: Execution #1 of query
2024-08-31 22:37:47,156 INFO ql.Driver: Concurrency mode is disabled, not creating a
2024-08-31 22:37:47,156 INFO ql.Driver: Executing command(queryId=Admin_2024083122374
dfd): SHOW TABLES
2024-08-31 22:37:47,157 INFO ql.Driver: Starting task [Stage-0:DDL] in serial mode
2024-08-31 22:37:47,157 INFO metastore.HiveMetaStore: 0: get_database: @hive#mydata
2024-08-31 22:37:47,157 INFO HiveMetaStore.audit: ugi=Admin     ip=unknown-ip-addr

2024-08-31 22:37:47,165 INFO metastore.HiveMetaStore: 0: get_tables: db=@hive#mydata
2024-08-31 22:37:47,166 INFO HiveMetaStore.audit: ugi=Admin     ip=unknown-ip-addr
pat=.*
2024-08-31 22:37:47,173 INFO ql.Driver: Completed executing command(queryId=Admin_202
-cbce9f8d2dfd); Time taken: 0.017 seconds
OK
2024-08-31 22:37:47,173 INFO ql.Driver: OK
2024-08-31 22:37:47,174 INFO ql.Driver: Concurrency mode is disabled, not creating a
2024-08-31 22:37:47,180 INFO mapred.FileInputFormat: Total input files to process : 1
2024-08-31 22:37:47,199 INFO exec.ListSinkOperator: RECORDS_OUT_INTERMEDIATE:0, RECOR
students
students_data
students_table
Time taken: 0.061 seconds, Fetched: 3 row(s)
2024-08-31 22:37:47,208 INFO CliDriver: Time taken: 0.061 seconds, Fetched: 3 row(s)
```

To show table column names and data types, run:

```
DESC students_table;
```

```
hive> DESC students_table;
2024-08-31 22:36:47,409 INFO conf.HiveConf: Using the default value passed in for log
bf2de5
2024-08-31 22:36:47,409 INFO session.SessionState: Updating thread name to 02b967f6-c0
2024-08-31 22:36:47,411 INFO ql.Driver: Compiling command(queryId=Admin_20240831223647
3f3): DESC students_table
2024-08-31 22:36:47,439 INFO ql.Driver: Concurrency mode is disabled, not creating a lo
2024-08-31 22:36:47,443 INFO metastore.HiveMetaStore: 0: get_table : tbl=hive.mydata.st
2024-08-31 22:36:47,444 INFO HiveMetaStore.audit: ugi=Admin     ip=unknown-ip-addr
students_table
2024-08-31 22:36:47,463 INFO parse.DDLSemanticAnalyzer: analyzeDescribeTable done
2024-08-31 22:36:47,464 INFO ql.Driver: Semantic Analysis Completed (retrial = false)
2024-08-31 22:36:47,464 INFO ql.Driver: Returning Hive schema: Schema(fieldSchemas:[Fie
ng, comment:from deserializer), FieldSchema(name:data_type, type:string, comment:from d
mment, type:string, comment:from deserializer)], properties:null)
2024-08-31 22:36:47,466 INFO exec.ListSinkOperator: Initializing operator LIST_SINK[0]
2024-08-31 22:36:47,467 INFO ql.Driver: Completed compiling command(queryId=Admin_2024
-b70c0809a3f3); Time taken: 0.056 seconds
2024-08-31 22:36:47,467 INFO reexec.ReExecDriver: Execution #1 of query
2024-08-31 22:36:47,467 INFO ql.Driver: Concurrency mode is disabled, not creating a lo
2024-08-31 22:36:47,467 INFO ql.Driver: Executing command(queryId=Admin_20240831223647
3f3): DESC students_table
2024-08-31 22:36:47,468 INFO ql.Driver: Starting task [Stage-0:DDL] in serial mode
2024-08-31 22:36:47,469 INFO metastore.HiveMetaStore: 0: get_table : tbl=hive.mydata.st
2024-08-31 22:36:47,469 INFO HiveMetaStore.audit: ugi=Admin     ip=unknown-ip-addr
students_table
2024-08-31 22:36:47,510 INFO ql.Driver: Completed executing command(queryId=Admin_2024
-b70c0809a3f3); Time taken: 0.043 seconds
OK
2024-08-31 22:36:47,512 INFO ql.Driver: OK
2024-08-31 22:36:47,513 INFO ql.Driver: Concurrency mode is disabled, not creating a lo
2024-08-31 22:36:47,521 INFO mapred.FileInputFormat: Total input files to process : 1
2024-08-31 22:36:47,536 INFO exec.ListSinkOperator: RECORDS_OUT_INTERMEDIATE:0, RECORDS
name                    string
roll                    int
dept                    string
Time taken: 0.106 seconds, Fetched: 3 row(s)
2024-08-31 22:36:47,545 INFO CliDriver: Time taken: 0.106 seconds, Fetched: 3 row(s)
```

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

```
SELECT * FROM students;
```

```
21687106974588382 99-1/-ext-10000
Loading data to table default.students
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 1.313 sec   HDFS Read: 17847 HDFS Write: 411 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 313 msec
OK
Time taken: 33.608 seconds
2024-09-07T20:01:34,011 INFO [2e42a8b6-ae15-4125-9e35-9f5467922de6 main] org.apache.hadoop.hive.conf.HiveConf - Using t
he default value passed in for log id: 2e42a8b6-ae15-4125-9e35-9f5467922de6
2024-09-07T20:01:34,011 INFO [2e42a8b6-ae15-4125-9e35-9f5467922de6 main] org.apache.hadoop.hive.ql.session.SessionState
 - Resetting thread name to  main
hive> SELECT * FROM students;
2024-09-07T20:01:59,989 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id
: 2e42a8b6-ae15-4125-9e35-9f5467922de6
2024-09-07T20:01:59,989 INFO [main] org.apache.hadoop.hive.ql.session.SessionState - Updating thread name to 2e42a8b6-a
e15-4125-9e35-9f5467922de6 main
2024-09-07T20:02:00,149 INFO [2e42a8b6-ae15-4125-9e35-9f5467922de6 main] org.apache.hadoop.hive.common.FileUtils - Crea
ting directory if it doesn't exist: hdfs://localhost:9000/tmp/hive/ASUS/2e42a8b6-ae15-4125-9e35-9f5467922de6/hive_2024-
09-07_20-02-00_005_7144496833936708017-1/-mr-10001/.hive-staging_hive_2024-09-07_20-02-00_005_7144496833936708017-1
OK
1       Pragadeesh      20      CSE C
2       Jazil   21      CSE C
3       Murshid 22      CSE C
Time taken: 0.192 seconds, Fetched: 3 row(s)
2024-09-07T20:02:00,203 INFO [2e42a8b6-ae15-4125-9e35-9f5467922de6 main] org.apache.hadoop.hive.conf.HiveConf - Using t
he default value passed in for log id: 2e42a8b6-ae15-4125-9e35-9f5467922de6
2024-09-07T20:02:00,203 INFO [2e42a8b6-ae15-4125-9e35-9f5467922de6 main] org.apache.hadoop.hive.ql.session.SessionState
 - Resetting thread name to  main
```

Result:

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