Ex no: 5   **210701171**

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)
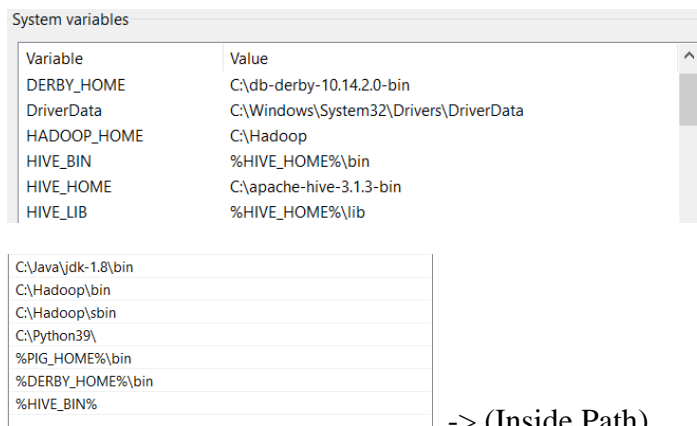
- 10.14.2.0 (May 3, 2018 / SVN 1828579)

-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

```
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
13136 NodeManager
1696 NameNode
15624 ResourceManager
10252 Jps
3532 DataNode

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
Tue Sep 17 18:54:04 IST 2024 : Security manager installed using the Basic server security policy.
Tue Sep 17 18:54:05 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
13136 NodeManager
1696 NameNode
2208 Jps
29172 NetworkServerControl
15624 ResourceManager
3532 DataNode

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

```
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#

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
```

SELECT * FROM students;

```
1        murshid 21
Time taken: 3.798 seconds, Fetched: 1 row(s)
2024-09-17 12:17:28,933 INFO CliDriver: Time taken: 3.798 seconds, Fetched: 1 row(s)
2024-09-17 12:17:28,933 INFO conf.HiveConf: Using the default value passed in for log id: 423831dd-7ad6-4562-b701-f8bf7d
fb651e
2024-09-17 12:17:28,933 INFO session.SessionState: Resetting thread name to  main
hive>
```

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...
```

File information - 000000_0   ×

Download   Head the file (first 32K)   Tail the file (last 32K)

Block information –   Block 0   ✓

Block ID: 1073741882

Block Pool ID: BP-986848115-192.168.52.206-1724747253865

Generation Stamp: 1058

Size: 13

Availability:

  • LAPTOP-17OLBCES

File contents

  1☐murshid☐21

Result:

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