CM2606 Data Engineering

Tutorial 04

The aim of this tutorial is to:

- Give students hands-on experience with working Hive Data warehouse.

1. If you are using the sandbox for Hadoop Log into the sandbox VM as the root using the link http://localhost:1080 Or else, you can use your local machine with Hadoop installed.

2. This tutorial assumes that you have csv files copied to your hdfs folder (/user/hadoop/geolocation) as instructed in Tutorial 3.

3. Start a **Hive shell** by typing `hive` at the command prompt and enter the following commands.

4. Use <show tables;> command to see all the tables in hive.

5. Use the below command to create your first table in hive

```
< CREATE TABLE `default`.`geolocation`(
  `truckid` STRING,
  `driverid` STRING,
  `event` STRING,
  `latitude` DOUBLE,
  `longitude` DOUBLE,
  `city` STRING,
  `state` STRING,
  `velocity` INT,
```

```
  `event_ind` INT,
  `idling_ind` INT
)
COMMENT 'Locations'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/user/hadoop/geolocation/';>
```

6. You might hit the below error when you try to execute the above command.

   Error: Error while compiling statement: FAILED: HiveAccessControlException Permission denied: user [hive] does not have [ALL] privilege on [hdfs://sandbox-hdp.hortonworks.com:8020/user/hadoop/geolocation] (state=42000,code=40000) 0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> Closing: 0: jdbc:hive2://sandbox-hdp.hortonworks.com:2181/default;password=hive;serviceDiscov eryMode=zooKeeper;user=hive;zooKeeperNamespace=hiveserver2

7. If you hit the error, make sure you give the required permission for the csv file to be used. Exit the hive shell, login as hdfs user (refer tutorial 3) and use the below command

```
[root@sandbox-hdp ~]# su hdfs
[hdfs@sandbox-hdp root]$ hdfs dfs -chmod 777 /user/hadoop/geolocation/geolocation.csv
[hdfs@sandbox-hdp root]$ exit
```

8. Once permission granted you should be able to create the table successfully

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> CREATE TABLE `default`.`geolocation` (
. . . . . .>
. . . . . .>   `truckid` STRING,
. . . . . .>
. . . . . .>   `driverid` STRING,
. . . . . .>
. . . . . .>   `event` STRING,
. . . . . .>
. . . . . .>   `latitude` DOUBLE,
. . . . . .>
. . . . . .>   `longitude` DOUBLE,
. . . . . .>
. . . . . .>   `city` STRING,
. . . . . .>
. . . . . .>   `state` STRING,
. . . . . .>
. . . . . .>   `velocity` INT,
. . . . . .>
. . . . . .>   `event_ind` INT,
. . . . . .>
. . . . . .>   `idling_ind` INT
. . . . . .>
. . . . . .> ) COMMENT 'Locations' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/user/hadoop/geolocation/'
INFO  : Compiling command(queryId=hive_20220202212524_1042d1ab-7b2c-465a-b001-d2caa1aec1d1): CREATE TABLE `default`.`geolocation` (
`truckid` STRING,
`driverid` STRING,
`event` STRING,
`latitude` DOUBLE,
`longitude` DOUBLE,
`city` STRING,
`state` STRING,
`velocity` INT,
`event_ind` INT,
`idling_ind` INT
) COMMENT 'Locations' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/user/hadoop/geolocation/'
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20220202212524_1042d1ab-7b2c-465a-b001-d2caa1aec1d1); Time taken: 0.03 seconds
INFO  : Executing command(queryId=hive_20220202212524_1042d1ab-7b2c-465a-b001-d2caa1aec1d1): CREATE TABLE `default`.`geolocation` (
`truckid` STRING,
`driverid` STRING,
`event` STRING,
`latitude` DOUBLE,
`longitude` DOUBLE,
`city` STRING,
`state` STRING,
`velocity` INT,
`event_ind` INT,
`idling_ind` INT
) COMMENT 'Locations' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/user/hadoop/geolocation/'
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20220202212524_1042d1ab-7b2c-465a-b001-d2caa1aec1d1); Time taken: 0.098 seconds
INFO  : OK
No rows affected (0.16 seconds)
```

9. Now Use <show tables;> command again to see the created table.

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> show tables;
INFO  : Compiling command(queryId=hive_20220202205148_843428be-2f69-453d-a6b7-1cfb8e7e4346): show tables
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20220202205148_843428be-2f69-453d-a6b7-1cfb8e7e4346); Time taken: 0.022 seconds
INFO  : Executing command(queryId=hive_20220202205148_843428be-2f69-453d-a6b7-1cfb8e7e4346): show tables
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20220202205148_843428be-2f69-453d-a6b7-1cfb8e7e4346); Time taken: 0.011 seconds
INFO  : OK
+--------------+
|   tab_name   |
+--------------+
| geolocation  |
+--------------+
1 row selected (0.051 seconds)
```

10. You can use the <describe formatted> command to check the schema and other meta data of the table created.

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> describe formatted geolocation;
INFO  : Compiling command(queryId=hive_20220202212659_197556a3-e3cb-4bab-bb51-217c8ef550c9): describe formatted geolocation
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:col_name, type:string, comment:from deserializer), FieldSchema(name:data_type, type:string, comment:from deserializer), FieldSchema(name:comment, type:strin
g, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20220202212659_197556a3-e3cb-4bab-bb51-217c8ef550c9); Time taken: 0.035 seconds
INFO  : Executing command(queryId=hive_20220202212659_197556a3-e3cb-4bab-bb51-217c8ef550c9): describe formatted geolocation
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20220202212659_197556a3-e3cb-4bab-bb51-217c8ef550c9); Time taken: 0.031 seconds
INFO  : OK
+-------------------------------+----------------------------------------------------+-----------------------+
|           col_name            |                     data_type                      |        comment        |
+-------------------------------+----------------------------------------------------+-----------------------+
| # col_name                    | data_type                                          | comment               |
| truckid                       | string                                             |                       |
| driverid                      | string                                             |                       |
| event                         | string                                             |                       |
| latitude                      | double                                             |                       |
| longitude                     | double                                             |                       |
| city                          | string                                             |                       |
| state                         | string                                             |                       |
| velocity                      | int                                                |                       |
| event_ind                     | int                                                |                       |
| idling_ind                    | int                                                |                       |
|                               | NULL                                               | NULL                  |
| # Detailed Table Information  | NULL                                               | NULL                  |
| Database:                     | default                                            | NULL                  |
| OwnerType:                    | USER                                               | NULL                  |
| Owner:                        | hive                                               | NULL                  |
| CreateTime:                   | Wed Feb 02 21:25:24 UTC 2022                       | NULL                  |
| LastAccessTime:               | UNKNOWN                                            | NULL                  |
| Retention:                    | 0                                                  | NULL                  |
| Location:                     | hdfs://sandbox-hdp.hortonworks.com:8020/user/hadoop/geolocation | NULL                  |
| Table Type:                   | MANAGED_TABLE                                      | NULL                  |
| Table Parameters:             | NULL                                               | NULL                  |
|                               | bucketing_version                                  | 2                     |
|                               | comment                                            | Locations             |
|                               | numFiles                                           | 1                     |
|                               | totalSize                                          | 526677                |
|                               | transactional                                      | true                  |
|                               | transactional_properties                           | insert_only           |
|                               | transient_lastDdlTime                              | 1643837124            |
|                               | NULL                                               | NULL                  |
| # Storage Information         | NULL                                               | NULL                  |
| SerDe Library:                | org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe | NULL                  |
| InputFormat:                  | org.apache.hadoop.mapred.TextInputFormat           | NULL                  |
| OutputFormat:                 | org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat | NULL                  |
| Compressed:                   | No                                                 | NULL                  |
| Num Buckets:                  | -1                                                 | NULL                  |
| Bucket Columns:               | []                                                 | NULL                  |
| Sort Columns:                 | []                                                 | NULL                  |
| Storage Desc Params:          | NULL                                               | NULL                  |
|                               | field.delim                                        | ,                     |
|                               | serialization.format                               | ,                     |
+-------------------------------+----------------------------------------------------+-----------------------+
41 rows selected (0.099 seconds)
```

11. Use general sql queries to query the table

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> select * from geolocation limit 5;
INFO  : Compiling command(queryId=hive_20220202212824_d61bc5c0-a65b-4833-aae7-a9e412856331): select * from geolocation limit 5
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:geolocation.truckid, type:string, comment:null), FieldSchema(name:geolocation.driverid, type:string, comment:null), FieldSchema(name:geolocation.event, type
:string, comment:null), FieldSchema(name:geolocation.latitude, type:double, comment:null), FieldSchema(name:geolocation.longitude, type:double, comment:null), FieldSchema(name:geolocation.city, type:string, comment:null), Fie
ldSchema(name:geolocation.state, type:string, comment:null), FieldSchema(name:geolocation.velocity, type:int, comment:null), FieldSchema(name:geolocation.event_ind, type:int, comment:null), FieldSchema(name:geolocation.idling
_ind, type:int, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20220202212824_d61bc5c0-a65b-4833-aae7-a9e412856331); Time taken: 0.19 seconds
INFO  : Executing command(queryId=hive_20220202212824_d61bc5c0-a65b-4833-aae7-a9e412856331): select * from geolocation limit 5
INFO  : Completed executing command(queryId=hive_20220202212824_d61bc5c0-a65b-4833-aae7-a9e412856331); Time taken: 0.032 seconds
INFO  : OK
+---------------------+---------------------+--------------------+---------------------+----------------------+--------------------+---------------------+-----------------------+---------------------+-----------------
--------+
| geolocation.truckid  | geolocation.driverid  | geolocation.event  | geolocation.latitude  | geolocation.longitude  | geolocation.city  | geolocation.state  | geolocation.velocity  | geolocation.event_ind  | geolocation.idli
ng_ind  |
+---------------------+---------------------+--------------------+---------------------+----------------------+--------------------+---------------------+-----------------------+---------------------+-----------------
--------+
| truckid             | driverid            | event              | NULL                | NULL                 | city               | state               | NULL                  | NULL                | NULL
        |
| A54                 | A54                 | normal             | 38.440467           | -122.714431          | Santa Rosa         | California          | 17                    | 0                   | 0
        |
| A20                 | A20                 | normal             | 36.977173           | -121.899402          | Aptos              | California          | 27                    | 0                   | 0
        |
| A40                 | A40                 | overspeed          | 37.957702           | -121.29078           | Stockton           | California          | 77                    | 1                   | 0
        |
| A31                 | A31                 | normal             | 39.409608           | -123.355566          | Willits            | California          | 22                    | 0                   | 0
        |
+---------------------+---------------------+--------------------+---------------------+----------------------+--------------------+---------------------+-----------------------+---------------------+-----------------
--------+
5 rows selected (0.29 seconds)
```

12. If you are using the sandbox explore how you can use Data Analytics Studio to create the same table via UI by simply uploading a csv file. Refer the tutorial here.