BIG DATA ASSIGNMENT #4

Name: Akshat Mishra | net-id: am15111 | Spring 2025

Q (A) Using Sales.csv file create a Hive table with name, Net-ID_sales (am15111_sales).

Creating the Table using:

```
CREATE TABLE am15111_sales (
    customer_id INT,
    transaction_id INT,
    product_category STRING,
    product_name STRING,
    quantity INT,
    sales_amount INT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;
```

```
0: jdbc:hive2://localhost:10000> CREATE TABLE am15111_sales (
. . . . . .>     customer_id INT,
. . . . . .>     transaction_id INT,
. . . . . .>     product_category STRING,
. . . . . .>     product_name STRING,
. . . . . .>     quantity INT,
. . . . . .>     sales_amount INT
. . . . . .> )
. . . . . .> ROW FORMAT DELIMITED
. . . . . .> FIELDS TERMINATED BY ','
. . . . . .> STORED AS TEXTFILE;
INFO  : Compiling command(queryId=hive_20250314214924_61007fd6-e925-45b1-b59f-ac66bf63008a): CREATE TABLE am15111_sales (
customer_id INT,
transaction_id INT,
product_category STRING,
product_name STRING,
quantity INT,
sales_amount INT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20250314214924_61007fd6-e925-45b1-b59f-ac66bf63008a); Time taken: 0.035 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250314214924_61007fd6-e925-45b1-b59f-ac66bf63008a): CREATE TABLE am15111_sales (
customer_id INT,
transaction_id INT,
product_category STRING,
product_name STRING,
quantity INT,
sales_amount INT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20250314214924_61007fd6-e925-45b1-b59f-ac66bf63008a); Time taken: 0.065 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.107 seconds)
```

1.List table columns of sales with describe command.

Command: DESC am15111_sales;

```
0: jdbc:hive2://localhost:10000> DESC am15111_sales;
INFO  : Compiling command(queryId=hive_20250314215042_0ad6b5bf-7f83-4cd2-9e56-370fce9b8ed9): DESC am15111_sales
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:col_name, type:string, comment:from deserializer), FieldSch
rom deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250314215042_0ad6b5bf-7f83-4cd2-9e56-370fce9b8ed9); Time taken: 0.045 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250314215042_0ad6b5bf-7f83-4cd2-9e56-370fce9b8ed9): DESC am15111_sales
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20250314215042_0ad6b5bf-7f83-4cd2-9e56-370fce9b8ed9); Time taken: 0.021 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------------------+------------+----------+
|     col_name      | data_type  | comment  |
+-------------------+------------+----------+
| customer_id       | int        |          |
| transaction_id    | int        |          |
| product_category  | string     |          |
| product_name      | string     |          |
| quantity          | int        |          |
| sales_amount      | int        |          |
+-------------------+------------+----------+
```

2.Add a column birth date with appropriate datatype.

Command: ALTER TABLE am15111_sales ADD COLUMNS (birth_date STRING);

```
0: jdbc:hive2://localhost:10000> ALTER TABLE am15111_sales ADD COLUMNS (birth_date DATE);
INFO  : Compiling command(queryId=hive_20250314215233_409e6b9f-b967-4a7c-8389-c6ef1158453b): ALTER TABLE am
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20250314215233_409e6b9f-b967-4a7c-8389-c6ef1158453b); Time
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250314215233_409e6b9f-b967-4a7c-8389-c6ef1158453b): ALTER TABLE am
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20250314215233_409e6b9f-b967-4a7c-8389-c6ef1158453b); Time
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.135 seconds)
0: jdbc:hive2://localhost:10000> DESC am15111_sales;
INFO  : Compiling command(queryId=hive_20250314215241_06e04df6-0209-4d01-b1c2-ef2d6a7c22f3): DESC am15111_s
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:col_name, type:string, comment:from de
rom deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250314215241_06e04df6-0209-4d01-b1c2-ef2d6a7c22f3); Time
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250314215241_06e04df6-0209-4d01-b1c2-ef2d6a7c22f3): DESC am15111_s
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20250314215241_06e04df6-0209-4d01-b1c2-ef2d6a7c22f3); Time
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------------------+------------+----------+
|     col_name      | data_type  | comment  |
+-------------------+------------+----------+
| customer_id       | int        |          |
| transaction_id    | int        |          |
| product_category  | string     |          |
| product_name      | string     |          |
| quantity          | int        |          |
| sales_amount      | int        |          |
| birth_date        | date       |          |
+-------------------+------------+----------+
```

3.Create a test table, testsales by selecting all records from the sales table.

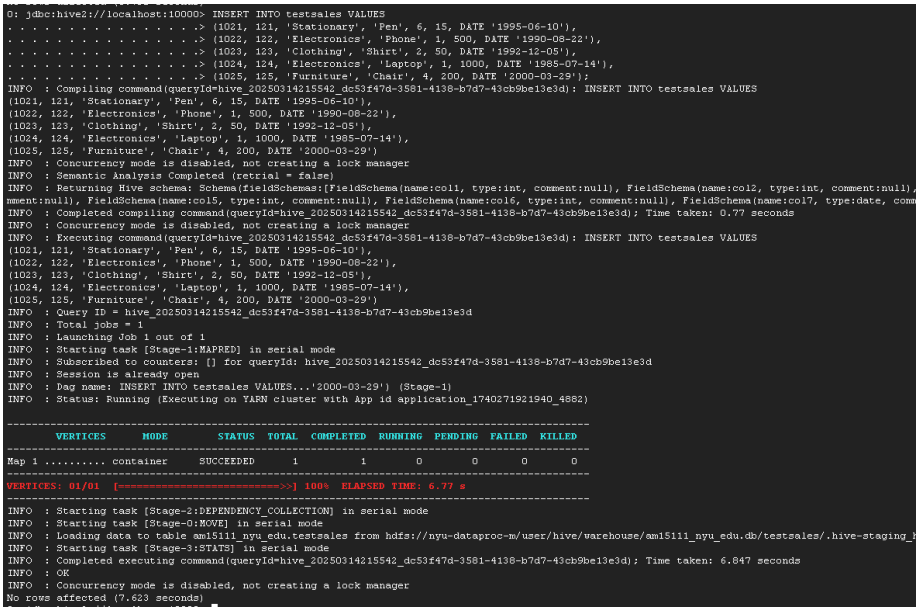
Command: CREATE TABLE testsales AS SELECT * FROM am15111_sales;

```
0: jdbc:hive2://localhost:10000> CREATE TABLE testsales AS SELECT * FROM am15111_sales;
INFO  : Compiling command(queryId=hive_20250314215335_b05fcbf1-6300-479a-bdd4-576991223ec7): CREATE TABLE testsales AS SELECT * FROM am1511
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:am15111_sales.customer_id, type:int, comment:null), FieldSchema(name:a
ory, type:string, comment:null), FieldSchema(name:am15111_sales.product_name, type:string, comment:null), FieldSchema(name:am15111_sales.qu
ll), FieldSchema(name:am15111_sales.birth_date, type:date, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250314215335_b05fcbf1-6300-479a-bdd4-576991223ec7); Time taken: 0.089 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250314215335_b05fcbf1-6300-479a-bdd4-576991223ec7): CREATE TABLE testsales AS SELECT * FROM am1511
INFO  : Query ID = hive_20250314215335_b05fcbf1-6300-479a-bdd4-576991223ec7
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20250314215335_b05fcbf1-6300-479a-bdd4-576991223ec7
INFO  : Tez session hasn't been created yet. Opening session
INFO  : Dag name: CREATE TABLE testsales AS SE...am15111_sales (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1740271921940_4882)

INFO  : Starting task [Stage-2:DEPENDENCY_COLLECTION] in serial mode
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Moving data to directory hdfs://nyu-dataproc-m/user/hive/warehouse/am15111_nyu_edu.db/testsales from hdfs://nyu-dataproc-m/user/hiv
/-ext-10001
INFO  : Starting task [Stage-4:DDL] in serial mode
INFO  : Starting task [Stage-3:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20250314215335_b05fcbf1-6300-479a-bdd4-576991223ec7); Time taken: 8.653 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1            container     SUCCEEDED      0          0        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 00/01  [>>--------------------------] 0%    ELAPSED TIME: 0.79 s
----------------------------------------------------------------------------------------------
No rows affected (8.792 seconds)
```

4.Insert 5 new records in test table.

Command: INSERT INTO testsales VALUES
(1021,121, 'Stationary', 'Pen', 6, 15, '1995-06-10'),
(1022,122, 'Electronics', 'Phone', 1, 500, '1990-08-22'),
(1023,123, 'Clothing', 'Shirt', 2, 50, '1992-12-05'),
(1024,124, 'Electronics', 'Laptop', 1, 1000, '1985-07-14'),
(1025,125, 'Furniture', 'Chair', 4, 200, '2000-03-29');



5.Query all records from the test table!

Command: SELECT * FROM testsales;

```
0: jdbc:hive2://localhost:10000> SELECT * FROM testsales;
INFO  : Compiling command(queryId=hive_20250314210935_a2240604-932a-44f2-a7f1-58528a4a9c1b): SELECT * FROM testsales
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:testsales.customer_id, type:int, comment:null), FieldSchema(name:testsales.transaction_id, type:int, comment:null), FieldSchema(name:te
ring, comment:null), FieldSchema(name:testsales.product_name, type:string, comment:null), FieldSchema(name:testsales.quantity, type:int, comment:null), FieldSchema(name:testsales.sales_amount, type:double
:testsales.birth_date, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250314210935_a2240604-932a-44f2-a7f1-58528a4a9c1b); Time taken: 0.285 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250314210935_a2240604-932a-44f2-a7f1-58528a4a9c1b): SELECT * FROM testsales
INFO  : Completed executing command(queryId=hive_20250314210935_a2240604-932a-44f2-a7f1-58528a4a9c1b); Time taken: 0.0 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----------------------+--------------------------+-----------------------------+-------------------------+---------------------+-------------------------+-----------------------+
| testsales.customer_id | testsales.transaction_id | testsales.product_category  | testsales.product_name  | testsales.quantity  | testsales.sales_amount  | testsales.birth_date  |
+-----------------------+--------------------------+-----------------------------+-------------------------+---------------------+-------------------------+-----------------------+
| NULL                  | NULL                     | product_category            | product_name            | NULL                | NULL                    | NULL                  |
| 1001                  | 101                      | Electronics                 | Laptop                  | 2                   | 1200.0                  | NULL                  |
| 1002                  | 102                      | Clothing                    | T-Shirt                 | 5                   | 75.0                    | NULL                  |
| 1003                  | 103                      | Home & Garden               | Lawn Mower              | 1                   | 299.99                  | NULL                  |
| 1004                  | 104                      | Electronics                 | Smartphone              | 3                   | 900.0                   | NULL                  |
| 1005                  | 105                      | Books                       | Novel - The Catcher     | 2                   | 29.99                   | NULL                  |
| 1006                  | 106                      | Electronics                 | Headphones              | 1                   | 49.99                   | NULL                  |
| 1007                  | 107                      | Clothing                    | Jeans                   | 2                   | 49.95                   | NULL                  |
| 1008                  | 108                      | Home & Garden               | Vacuum Cleaner          | 1                   | 199.99                  | NULL                  |
| 1009                  | 109                      | Electronics                 | Smart TV                | 1                   | 799.0                   | NULL                  |
| 1010                  | 110                      | Books                       | Cookbook                | 3                   | 24.99                   | NULL                  |
| 1011                  | 111                      | Electronics                 | Tablet                  | 2                   | 350.0                   | NULL                  |
| 1012                  | 112                      | Clothing                    | Dress                   | 1                   | 99.99                   | NULL                  |
| 1013                  | 113                      | Home & Garden               | Garden Hose             | 3                   | 19.99                   | NULL                  |
| 1014                  | 114                      | Electronics                 | Wireless Mouse          | 4                   | 14.95                   | NULL                  |
| 1015                  | 115                      | Books                       | Sci-Fi Novel            | 1                   | 9.99                    | NULL                  |
| 1016                  | 116                      | Electronics                 | Bluetooth Speaker       | 2                   | 59.99                   | NULL                  |
| 1017                  | 117                      | Clothing                    | Sneakers                | 1                   | 89.95                   | NULL                  |
| 1018                  | 118                      | Home & Garden               | BBQ Grill               | 1                   | 349.99                  | NULL                  |
| 1019                  | 119                      | Electronics                 | Camera                  | 1                   | 499.0                   | NULL                  |
| 1020                  | 120                      | Books                       | Mystery Novel           | 2                   | 12.99                   | NULL                  |
| 1021                  | 121                      | Stationary                  | Pen                     | 6                   | 15.0                    | 1995-06-10            |
| 1022                  | 122                      | Electronics                 | Phone                   | 1                   | 500.0                   | 1990-08-22            |
| 1023                  | 123                      | Clothing                    | Shirt                   | 2                   | 50.0                    | 1992-12-05            |
| 1024                  | 124                      | Electronics                 | Laptop                  | 1                   | 1000.0                  | 1985-07-14            |
| 1025                  | 125                      | Furniture                   | Chair                   | 4                   | 200.0                   | 2000-03-29            |
+-----------------------+--------------------------+-----------------------------+-------------------------+---------------------+-------------------------+-----------------------+
26 rows selected (0.313 seconds)
```

6.Write three queries with filters (where clause) and show result of queries.

Query 1: Select sales above $100

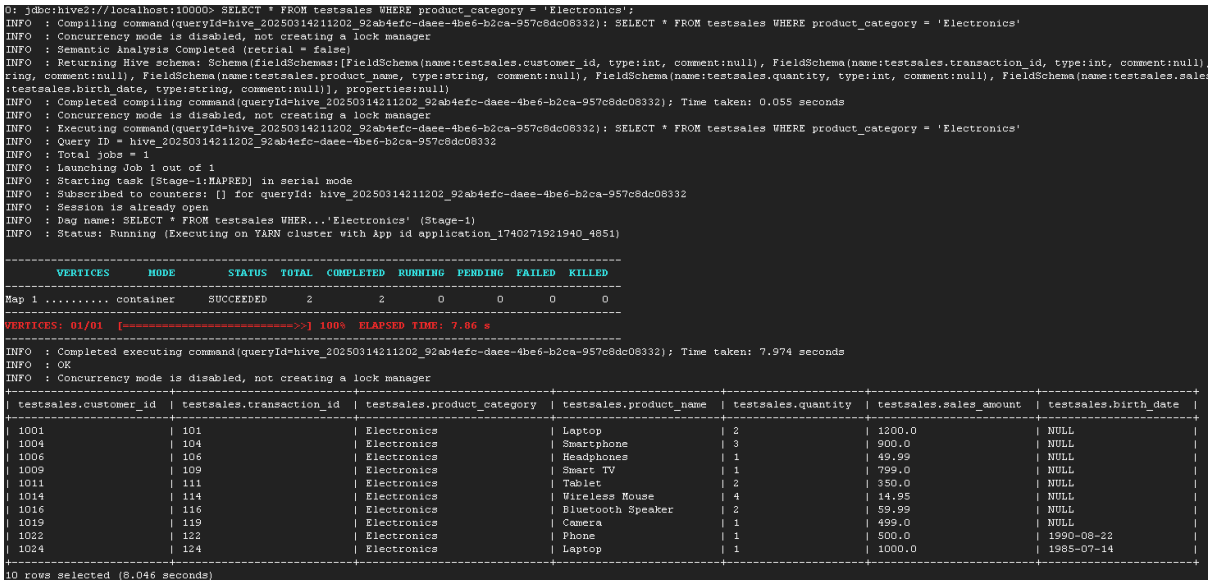
Command: SELECT * FROM testsales WHERE sales_amount > 100;

```
0: jdbc:hive2://localhost:10000> SELECT * FROM testsales WHERE sales_amount > 100;
INFO  : Compiling command(queryId=hive_20250314211104_50b6207f-105c-4ac2-920e-6845a32ee89e): SELECT * FROM testsales WHERE sales_amount > 100
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:testsales.customer_id, type:int, comment:null), FieldSchema(name:testsales.transaction_id, type:int, comment:null), FieldSche
ring, comment:null), FieldSchema(name:testsales.product_name, type:string, comment:null), FieldSchema(name:testsales.quantity, type:int, comment:null), FieldSchema(name:testsales.sales_amount, t
:testsales.birth_date, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250314211104_50b6207f-105c-4ac2-920e-6845a32ee89e); Time taken: 0.053 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250314211104_50b6207f-105c-4ac2-920e-6845a32ee89e): SELECT * FROM testsales WHERE sales_amount > 100
INFO  : Query ID = hive_20250314211104_50b6207f-105c-4ac2-920e-6845a32ee89e
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20250314211104_50b6207f-105c-4ac2-920e-6845a32ee89e
INFO  : Session is already open
INFO  : Dag name: SELECT * FROM testsales WHERE sales_am...100 (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1740271921940_4851)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 7.26 s
--------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20250314211104_50b6207f-105c-4ac2-920e-6845a32ee89e); Time taken: 7.39 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----------------------+--------------------------+-----------------------------+-------------------------+---------------------+-------------------------+-----------------------+
| testsales.customer_id | testsales.transaction_id | testsales.product_category  | testsales.product_name  | testsales.quantity  | testsales.sales_amount  | testsales.birth_date  |
+-----------------------+--------------------------+-----------------------------+-------------------------+---------------------+-------------------------+-----------------------+
| 1001                  | 101                      | Electronics                 | Laptop                  | 2                   | 1200.0                  | NULL                  |
| 1003                  | 103                      | Home & Garden               | Lawn Mower              | 1                   | 299.99                  | NULL                  |
| 1004                  | 104                      | Electronics                 | Smartphone              | 3                   | 900.0                   | NULL                  |
| 1008                  | 108                      | Home & Garden               | Vacuum Cleaner          | 1                   | 199.99                  | NULL                  |
| 1009                  | 109                      | Electronics                 | Smart TV                | 1                   | 799.0                   | NULL                  |
| 1011                  | 111                      | Electronics                 | Tablet                  | 2                   | 350.0                   | NULL                  |
| 1018                  | 118                      | Home & Garden               | BBQ Grill               | 1                   | 349.99                  | NULL                  |
| 1019                  | 119                      | Electronics                 | Camera                  | 1                   | 499.0                   | NULL                  |
| 1022                  | 122                      | Electronics                 | Phone                   | 1                   | 500.0                   | 1990-08-22            |
| 1024                  | 124                      | Electronics                 | Laptop                  | 1                   | 1000.0                  | 1985-07-14            |
| 1025                  | 125                      | Furniture                   | Chair                   | 4                   | 200.0                   | 2000-03-29            |
+-----------------------+--------------------------+-----------------------------+-------------------------+---------------------+-------------------------+-----------------------+
```

Query 2: Select products in the Electronics category

Command: SELECT * FROM testsales WHERE product_category = 'Electronics';

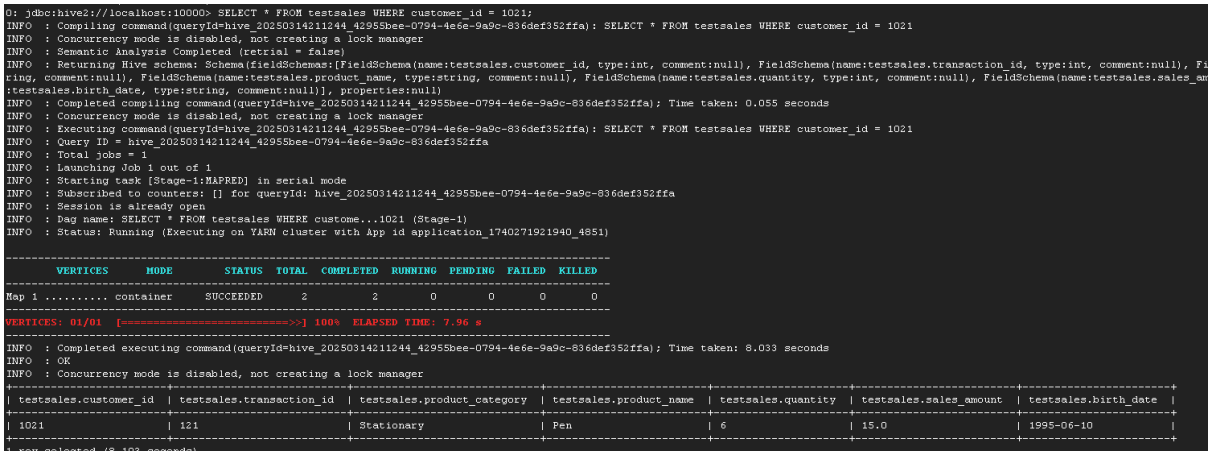


```
0: jdbc:hive2://localhost:10000> SELECT * FROM testsales WHERE product_category = 'Electronics';
INFO  : Compiling command(queryId=hive_20250314211202_92ab4efc-daee-4be6-b2ca-957c8dc08332): SELECT * FROM testsales WHERE product_category = 'Electronics'
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:testsales.customer_id, type:int, comment:null), FieldSchema(name:testsales.transaction_id, type:int, comment:null),
ring, comment:null), FieldSchema(name:testsales.product_name, type:string, comment:null), FieldSchema(name:testsales.quantity, type:int, comment:null), FieldSchema(name:testsales.sale
:testsales.birth_date, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250314211202_92ab4efc-daee-4be6-b2ca-957c8dc08332); Time taken: 0.055 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250314211202_92ab4efc-daee-4be6-b2ca-957c8dc08332): SELECT * FROM testsales WHERE product_category = 'Electronics'
INFO  : Query ID = hive_20250314211202_92ab4efc-daee-4be6-b2ca-957c8dc08332
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20250314211202_92ab4efc-daee-4be6-b2ca-957c8dc08332
INFO  : Session is already open
INFO  : Dag name: SELECT * FROM testsales WHER...'Electronics' (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1740271921940_4851)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 7.86 s
--------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20250314211202_92ab4efc-daee-4be6-b2ca-957c8dc08332); Time taken: 7.974 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----------------------+--------------------------+-----------------------------+-------------------------+---------------------+-------------------------+-----------------------+
| testsales.customer_id | testsales.transaction_id | testsales.product_category  | testsales.product_name  | testsales.quantity  | testsales.sales_amount  | testsales.birth_date  |
+-----------------------+--------------------------+-----------------------------+-------------------------+---------------------+-------------------------+-----------------------+
| 1001                  | 101                      | Electronics                 | Laptop                  | 2                   | 1200.0                  | NULL                  |
| 1004                  | 104                      | Electronics                 | Smartphone              | 3                   | 900.0                   | NULL                  |
| 1006                  | 106                      | Electronics                 | Headphones              | 1                   | 49.99                   | NULL                  |
| 1009                  | 109                      | Electronics                 | Smart TV                | 1                   | 799.0                   | NULL                  |
| 1011                  | 111                      | Electronics                 | Tablet                  | 2                   | 350.0                   | NULL                  |
| 1014                  | 114                      | Electronics                 | Wireless Mouse          | 4                   | 14.95                   | NULL                  |
| 1016                  | 116                      | Electronics                 | Bluetooth Speaker       | 2                   | 59.99                   | NULL                  |
| 1019                  | 119                      | Electronics                 | Camera                  | 1                   | 499.0                   | NULL                  |
| 1022                  | 122                      | Electronics                 | Phone                   | 1                   | 500.0                   | 1990-08-22            |
| 1024                  | 124                      | Electronics                 | Laptop                  | 1                   | 1000.0                  | 1985-07-14            |
+-----------------------+--------------------------+-----------------------------+-------------------------+---------------------+-------------------------+-----------------------+
10 rows selected (8.046 seconds)
```

Query 3: Select transactions for customer_id 1021

Command: SELECT * FROM testsales WHERE customer_id = 1021;



```
0: jdbc:hive2://localhost:10000> SELECT * FROM testsales WHERE customer_id = 1021;
INFO  : Compiling command(queryId=hive_20250314211244_42955bee-0794-4e6e-9a9c-836def352ffa): SELECT * FROM testsales WHERE customer_id = 1021
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:testsales.customer_id, type:int, comment:null), FieldSchema(name:testsales.transaction_id, type:int, comment:null), Fi
ring, comment:null), FieldSchema(name:testsales.product_name, type:string, comment:null), FieldSchema(name:testsales.quantity, type:int, comment:null), FieldSchema(name:testsales.sales_am
:testsales.birth_date, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250314211244_42955bee-0794-4e6e-9a9c-836def352ffa); Time taken: 0.055 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250314211244_42955bee-0794-4e6e-9a9c-836def352ffa): SELECT * FROM testsales WHERE customer_id = 1021
INFO  : Query ID = hive_20250314211244_42955bee-0794-4e6e-9a9c-836def352ffa
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20250314211244_42955bee-0794-4e6e-9a9c-836def352ffa
INFO  : Session is already open
INFO  : Dag name: SELECT * FROM testsales WHERE custome...1021 (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1740271921940_4851)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 7.96 s
--------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20250314211244_42955bee-0794-4e6e-9a9c-836def352ffa); Time taken: 8.033 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----------------------+--------------------------+-----------------------------+-------------------------+---------------------+-------------------------+-----------------------+
| testsales.customer_id | testsales.transaction_id | testsales.product_category  | testsales.product_name  | testsales.quantity  | testsales.sales_amount  | testsales.birth_date  |
+-----------------------+--------------------------+-----------------------------+-------------------------+---------------------+-------------------------+-----------------------+
| 1021                  | 121                      | Stationary                  | Pen                     | 6                   | 15.0                    | 1995-06-10            |
+-----------------------+--------------------------+-----------------------------+-------------------------+---------------------+-------------------------+-----------------------+
1 row selected (8.103 seconds)
```

7.Show the list of tables.

Command: show tables;

```
0: jdbc:hive2://localhost:10000> show tables;
INFO  : Compiling command(queryId=hive_20250314220102_00d6
INFO  : Concurrency mode is disabled, not creating a lock
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldS
INFO  : Completed compiling command(queryId=hive_202503142
INFO  : Concurrency mode is disabled, not creating a lock
INFO  : Executing command(queryId=hive_20250314220102_00d6
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_202503142
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock
+----------------+
|    tab_name    |
+----------------+
| am15111_sales  |
| testsales      |
+----------------+
```

8.Drop the test table.

Command: drop table testsales;

```
0: jdbc:hive2://localhost:10000> drop table testsales;
INFO  : Compiling command(queryId=hive_20250314220136_09b68faf-6432-4eb1-b99d-3540cf4b253f): drop table testsales
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20250314220136_09b68faf-6432-4eb1-b99d-3540cf4b253f); Time taken: 0.042 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250314220136_09b68faf-6432-4eb1-b99d-3540cf4b253f): drop table testsales
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20250314220136_09b68faf-6432-4eb1-b99d-3540cf4b253f); Time taken: 0.126 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.175 seconds)
```

9.Show the list of tables after dropping test table

Command: show tables;

```
0: jdbc:hive2://localhost:10000> show tables;
INFO  : Compiling command(queryId=hive_20250314220159_b201f908-8b87-424c-af5f-529715f1fccb): show tables
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, type:string, comment:from deserializer)], properties:nu
INFO  : Completed compiling command(queryId=hive_20250314220159_b201f908-8b87-424c-af5f-529715f1fccb); Time taken: 0.034 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250314220159_b201f908-8b87-424c-af5f-529715f1fccb): show tables
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20250314220159_b201f908-8b87-424c-af5f-529715f1fccb); Time taken: 0.021 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+----------------+
|    tab_name    |
+----------------+
| am15111_sales  |
+----------------+
```

Q (B) Use following code to create a Hive table, customers with name, Net-Id_customers (e.g. am15111_customers).

Created Table using:

```
CREATE TABLE am15111_customers (
    customer_id INT,
    customer_name STRING,
    customer_email STRING,
    customer_address STRING
);
```

Insert Data into Customers Table

```
INSERT INTO TABLE am15111_customers VALUES
(1001, 'John Doe', 'john@example.com', '123 Main St'),
(1002, 'Alice Smith', 'alice@example.com', '456 Elm St'),
(1003, 'Bob Johnson', 'bob@example.com', '789 Oak St'),
(7001, 'John Doe', 'john@test.com', '123 Main St'),
(7002, 'Alice Smith', 'alice@test.com', '456 Elm St'),
(7003, 'Bob Johnson', 'bob@test.com', '789 Oak St');
```

Using Sales and Customers tables, write quires with INNER JOIN, LEFT OUTER JOIN, RIGHT OUTER JOIN, and FULL OUTER JOIN. Submit SQL queries and screenshot of their results.

```
INNER JOIN: SELECT s.transaction_id, s.product_name, s.sales_amount,
c.customer_name, c.customer_email
FROM am15111_sales s
JOIN am15111_customers c
ON s.customer_id = c.customer_id;
```

s.transaction_id	s.product_name	s.sales_amount	c.customer_name	c.customer_email
103	Lawn Mower	299.99	Bob Johnson	bob@example.com
101	Laptop	1200.0	John Doe	john@example.com
102	T-Shirt	75.0	Alice Smith	alice@example.com

LEFT OUTER JOIN: SELECT c.customer_id, c.customer_name, s.transaction_id, s.product_name
FROM am15111_customers c
LEFT JOIN am15111_sales s
ON c.customer_id = s.customer_id;

s.transaction_id	s.product_name	s.sales_amount	c.customer_name	c.customer_email
NULL	product_name	NULL	NULL	NULL
101	Laptop	1200.0	John Doe	john@example.com
102	T-Shirt	75.0	Alice Smith	alice@example.com
103	Lawn Mower	299.99	Bob Johnson	bob@example.com
104	Smartphone	900.0	NULL	NULL
105	Novel - The Catcher	29.99	NULL	NULL
106	Headphones	49.99	NULL	NULL
107	Jeans	49.95	NULL	NULL
108	Vacuum Cleaner	199.99	NULL	NULL
109	Smart TV	799.0	NULL	NULL
110	Cookbook	24.99	NULL	NULL
111	Tablet	350.0	NULL	NULL
112	Dress	99.99	NULL	NULL
113	Garden Hose	19.99	NULL	NULL
114	Wireless Mouse	14.95	NULL	NULL
115	Sci-Fi Novel	9.99	NULL	NULL
116	Bluetooth Speaker	59.99	NULL	NULL
117	Sneakers	89.95	NULL	NULL
118	BBQ Grill	349.99	NULL	NULL
119	Camera	499.0	NULL	NULL
120	Mystery Novel	12.99	NULL	NULL

RIGHT OUTER JOIN: SELECT s.transaction_id, s.product_name, s.sales_amount, c.customer_name, c.customer_email
FROM am15111_sales s
RIGHT JOIN am15111_customers c
ON s.customer_id = c.customer_id;

s.transaction_id	s.product_name	s.sales_amount	c.customer_name	c.customer_email
NULL	NULL	NULL	John Doe	john@test.com
NULL	NULL	NULL	Alice Smith	alice@test.com
103	Lawn Mower	299	Bob Johnson	bob@example.com
101	Laptop	1200	John Doe	john@example.com
NULL	NULL	NULL	Bob Johnson	bob@test.com
102	T-Shirt	75	Alice Smith	alice@example.com

FULL OUTER JOIN: SELECT s.transaction_id, s.product_name, s.sales_amount, c.customer_name, c.customer_email
FROM am15111_customers c
FULL OUTER JOIN am15111_sales s
ON c.customer_id = s.customer_id;

s.transaction_id	s.product_name	s.sales_amount	c.customer_name	c.customer_email
NULL	product_name	NULL	NULL	NULL
101	Laptop	1200.0	John Doe	john@example.com
102	T-Shirt	75.0	Alice Smith	alice@example.com
103	Lawn Mower	299.99	Bob Johnson	bob@example.com
104	Smartphone	900.0	NULL	NULL
105	Novel - The Catcher	29.99	NULL	NULL
106	Headphones	49.99	NULL	NULL
107	Jeans	49.95	NULL	NULL
108	Vacuum Cleaner	199.99	NULL	NULL
109	Smart TV	799.0	NULL	NULL
110	Cookbook	24.99	NULL	NULL
111	Tablet	350.0	NULL	NULL
112	Dress	99.99	NULL	NULL
113	Garden Hose	19.99	NULL	NULL
114	Wireless Mouse	14.95	NULL	NULL
115	Sci-Fi Novel	9.99	NULL	NULL
116	Bluetooth Speaker	59.99	NULL	NULL
117	Sneakers	89.95	NULL	NULL
118	BBQ Grill	349.99	NULL	NULL
119	Camera	499.0	NULL	NULL
120	Mystery Novel	12.99	NULL	NULL
NULL	NULL	NULL	John Doe	john@test.com
NULL	NULL	NULL	Alice Smith	alice@test.com
NULL	NULL	NULL	Bob Johnson	bob@test.com

Q (C) Using Zipcodes.csv file, create Hive table Net-ID_zipcodes (e.g. am15111_zipcodes). This table should have partitions by state and with 3 buckets by zipcode.

Provide screenshot of

1.hdfs directory and subdirectories of partitions, also show files under partition state='AL'

Command: hdfs dfs -ls /user/hive/warehouse/am15111_nyu_edu.db/am15111_zipcodes

```
am15111_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -ls /user/hive/warehouse/am15111_nyu_edu.db/am15111_zipcodes
Found 6 items
drwxr-xr-x   - am15111_nyu_edu am15111_nyu_edu          0 2025-03-14 21:38 /user/hive/warehouse/am15111_nyu_edu.db/am15111_zipcodes/state=AL
drwxr-xr-x   - am15111_nyu_edu am15111_nyu_edu          0 2025-03-14 21:38 /user/hive/warehouse/am15111_nyu_edu.db/am15111_zipcodes/state=AZ
drwxr-xr-x   - am15111_nyu_edu am15111_nyu_edu          0 2025-03-14 21:38 /user/hive/warehouse/am15111_nyu_edu.db/am15111_zipcodes/state=FL
drwxr-xr-x   - am15111_nyu_edu am15111_nyu_edu          0 2025-03-14 21:38 /user/hive/warehouse/am15111_nyu_edu.db/am15111_zipcodes/state=NC
drwxr-xr-x   - am15111_nyu_edu am15111_nyu_edu          0 2025-03-14 21:38 /user/hive/warehouse/am15111_nyu_edu.db/am15111_zipcodes/state=PR
drwxr-xr-x   - am15111_nyu_edu am15111_nyu_edu          0 2025-03-14 21:38 /user/hive/warehouse/am15111_nyu_edu.db/am15111_zipcodes/state=TX
```

Command: hdfs dfs -ls /user/hive/warehouse/am15111_nyu_edu.db/am15111_zipcodes/state='AL'

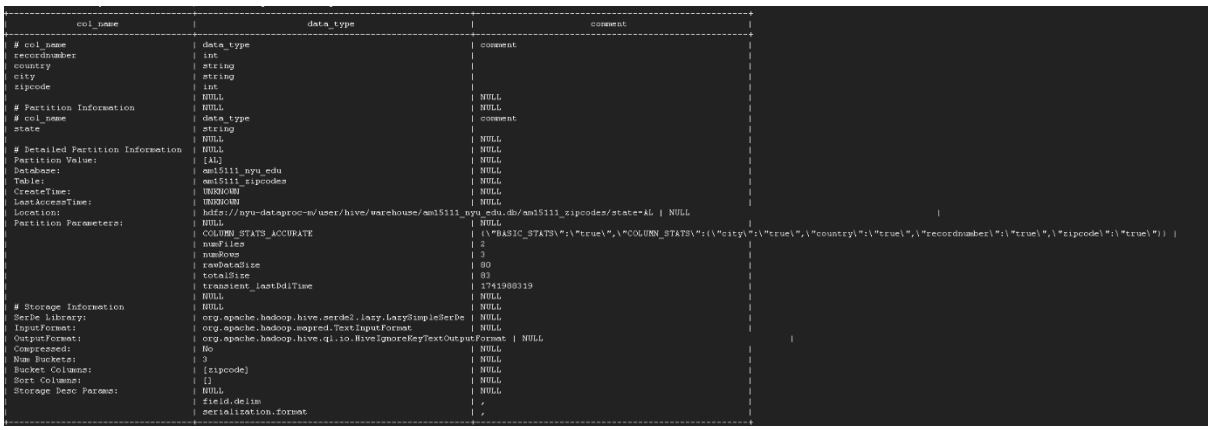
```
am15111_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -ls /user/hive/warehouse/am15111_nyu_edu.db/am15111_zipcodes/state='AL'
Found 2 items
-rw-r--r--   1 am15111_nyu_edu am15111_nyu_edu         27 2025-03-14 21:38 /user/hive/warehouse/am15111_nyu_edu.db/am15111_zipcodes/state=AL/000001_0
-rw-r--r--   1 am15111_nyu_edu am15111_nyu_edu         56 2025-03-14 21:38 /user/hive/warehouse/am15111_nyu_edu.db/am15111_zipcodes/state=AL/000002_0
am15111_nyu_edu@nyu-dataproc-m:~$
```

2.Results of following commands:

Command: SHOW PARTITIONS am15111_zipcodes;

```
0: jdbc:hive2://localhost:10000> SHOW PARTITIONS am15111_zipcodes;
INFO  : Compiling command(queryId=hive_20250314213915_8bb7638c-c4fb-4112-bede-7028dd436e29): SHOW PARTITIONS am15111_zipcodes
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:partition, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250314213915_8bb7638c-c4fb-4112-bede-7028dd436e29); Time taken: 0.367 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250314213915_8bb7638c-c4fb-4112-bede-7028dd436e29): SHOW PARTITIONS am15111_zipcodes
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20250314213915_8bb7638c-c4fb-4112-bede-7028dd436e29); Time taken: 0.048 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+------------+
| partition  |
+------------+
| state=AL   |
| state=AZ   |
| state=FL   |
| state=NC   |
| state=PR   |
| state=TX   |
+------------+
6 rows selected (0.458 seconds)
```

Command: DESCRIBE FORMATTED am15111_zipcodes PARTITION(state='AL');



col_name	data_type	comment
# col_name	data_type	comment
recordnumber	int	
country	string	
city	string	
zipcode	int	
	NULL	NULL
# Partition Information	NULL	NULL
# col_name	data_type	comment
state	string	
	NULL	NULL
# Detailed Partition Information	NULL	NULL
Partition Value:	[AL]	NULL
Database:	am15111_nyu_edu	NULL
Table:	am15111_zipcodes	NULL
CreateTime:	UNKNOWN	NULL
LastAccessTime:	UNKNOWN	NULL
Location:	hdfs://nyu-dataproc-m/user/hive/warehouse/am15111_nyu_edu.db/am15111_zipcodes/state=AL	NULL
Partition Parameters:	NULL	NULL
	COLUMN_STATS_ACCURATE	{\"BASIC_STATS\":\"true\",\"COLUMN_STATS\":{\"city\":\"true\",\"country\":\"true\",\"recordnumber\":\"true\",\"zipcode\":\"true\"}}
	numFiles	2
	numRows	3
	rawDataSize	80
	totalSize	83
	transient_lastDdlTime	1741988319
	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:	3	NULL
Bucket Columns:	[zipcode]	NULL
Sort Columns:	[]	NULL
Storage Desc Params:	NULL	NULL
	field.delim	,
	serialization.format	,

Command: SHOW TABLE EXTENDED LIKE am15111_zipcodes PARTITION(state='AL');

```
0: jdbc:hive2://localhost:10000> SHOW TABLE EXTENDED LIKE am15111_zipcodes PARTITION(state='AL');
INFO  : Compiling command(queryId=hive_20250314214042_e998659b-ec17-4fc7-bbf7-f1f23aeb506b): SHOW TABLE EXTENDED LIKE am15111_zipcodes PARTITION(state='AL')
INFO  : Concurrency mode is disabled, not creating a lock manager
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_20250314214042_e998659b-ec17-4fc7-bbf7-f1f23aeb506b); Time taken: 0.105 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20250314214042_e998659b-ec17-4fc7-bbf7-f1f23aeb506b): SHOW TABLE EXTENDED LIKE am15111_zipcodes PARTITION(state='AL')
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20250314214042_e998659b-ec17-4fc7-bbf7-f1f23aeb506b); Time taken: 0.068 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+----------------------------------------------------+
|                      tab_name                      |
+----------------------------------------------------+
| tableName:am15111_zipcodes                         |
| owner:am15111_nyu_edu                              |
| location:hdfs://nyu-dataproc-m/user/hive/warehouse/am15111_nyu_edu.db/am15111_zipcodes/state=AL |
| inputformat:org.apache.hadoop.mapred.TextInputFormat |
| outputformat:org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat |
| columns:struct columns { i32 recordnumber, string country, string city, i32 zipcode} |
| partitioned:true                                   |
| partitionColumns:struct partition_columns { string state} |
| totalNumberFiles:2                                 |
| totalFileSize:83                                   |
| maxFileSize:56                                     |
| minFileSize:27                                     |
| lastAccessTime:1741988317265                       |
| lastUpdateTime:1741988320097                       |
|                                                    |
+----------------------------------------------------+
```

Key learnings from this experiment:

1. **Hive Table Creation & Schema Management** – Understanding how to create and modify Hive tables, including adding columns and managing schemas dynamically.
2. **Data Manipulation & Querying** – Learning to insert, update, and query data using HiveQL, along with applying filters (WHERE clause) to extract meaningful insights.
3. **Joins & Data Relationships** – Gaining hands-on experience with different types of joins (INNER, LEFT, RIGHT, FULL OUTER) to combine datasets and analyze customer-sales relationships.
4. **Partitioning & Bucketing** – Learning to optimize Hive tables using partitioning (by state) and bucketing (by zipcode) to enhance query performance.
5. **HDFS File System Interaction** – Understanding how Hive stores data in HDFS, and using commands to navigate, list partitions, and inspect storage structures efficiently.