8. Create External table and load a csv file from local storage. The data should be stored in ORC format in specified folder in HDFS. Perform query operations and verify the data in HDFS. Delete the table and verify the data in HDFS.

To create an external table in Hive, load a CSV file from local storage, store the data in ORC format in a specified folder in HDFS, perform query operations, and then delete the table while ensuring the data remains in HDFS, follow these steps:

**Step 1: Prepare HDFS Directory**

1. **Create an HDFS Directory**:

hdfs dfs -mkdir -p /user/hive/external/employee\_orc

### Step 2: Create and Load the External Table

1. **Start Hive**:

hive

Create the External Table

CREATE EXTERNAL TABLE employee\_external (

id INT,

name STRING,

salary DOUBLE

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE

LOCATION '/user/hive/external/employee\_orc';

**Load Data from Local Storage**: Assuming your CSV file data.csv is in the local filesystem, load it into the Hive table:

LOAD DATA LOCAL INPATH '/path/to/data.csv' INTO TABLE employee\_external;

### Step 3: Convert Data to ORC Format

1. **Create a Table in ORC Format**:

CREATE TABLE employee\_orc (

id INT,

name STRING,

salary DOUBLE

)

STORED AS ORC;

**Insert Data into ORC Table**:

INSERT INTO TABLE employee\_orc

SELECT \* FROM employee\_external;

**Move Data to Specified HDFS Folder**:

INSERT OVERWRITE DIRECTORY '/user/hive/external/employee\_orc'

STORED AS ORC

SELECT \* FROM employee\_orc;

**Step 4: Verify the Data in HDFS**

1. **Check the HDFS Path**

hdfs dfs -ls /user/hive/external/employee\_orc

**Step 5: Perform Query Operations**

1. **Run Some Queries on the ORC Table**

-- Select all records

SELECT \* FROM employee\_orc;

-- Count the number of records

SELECT COUNT(\*) FROM employee\_orc;

-- Calculate the total salary

SELECT SUM(salary) FROM employee\_orc;

-- Find the highest salary

SELECT MAX(salary) FROM employee\_orc;

### Step 6: Delete the Table and Verify the Data in HDFS

1. **Drop the External Table**:

DROP TABLE employee\_external;

**Verify the Data in HDFS**: The data should still be in HDFS because it was loaded as an external table.

hdfs dfs -ls /user/hive/external/employee\_orc

### Complete Commands

Here's the complete sequence of commands to run in the Hive terminal and the shell:

**Hive Terminal:**

-- Create the external table

CREATE EXTERNAL TABLE employee\_external (

id INT,

name STRING,

salary DOUBLE

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE

LOCATION '/user/hive/external/employee\_orc';

-- Load data from local storage

LOAD DATA LOCAL INPATH '/path/to/data.csv' INTO TABLE employee\_external;

-- Create the ORC table

CREATE TABLE employee\_orc (

id INT,

name STRING,

salary DOUBLE

)

STORED AS ORC;

-- Insert data into the ORC table

INSERT INTO TABLE employee\_orc

SELECT \* FROM employee\_external;

-- Move data to the specified HDFS folder in ORC format

INSERT OVERWRITE DIRECTORY '/user/hive/external/employee\_orc'

STORED AS ORC

SELECT \* FROM employee\_orc;

-- Run queries

SELECT \* FROM employee\_orc;

SELECT COUNT(\*) FROM employee\_orc;

SELECT SUM(salary) FROM employee\_orc;

SELECT MAX(salary) FROM employee\_orc;

-- Drop the external table

DROP TABLE employee\_external;

Shell:

# Create the HDFS directory

hdfs dfs -mkdir -p /user/hive/external/employee\_orc

# Verify the HDFS path before dropping the table

hdfs dfs -ls /user/hive/external/employee\_orc

# Verify the HDFS path after dropping the table

hdfs dfs -ls /user/hive/external/employee\_orc

Replace /path/to/data.csv with the actual path to your CSV file. This sequence will create an external table, load data, convert it to ORC format, run queries, drop the table, and verify the data's presence in HDFS at each step.