1) Connect to the EMR cluster using SSH.

2)Open the Hive command-line interface by typing the following command:

**hive**

3) Create a database for storing the table by typing the following command:

**CREATE DATABASE IF NOT EXISTS airline\_db;**

This will create a database named 'airline\_db' if it does not already exist.

4) Create an external table using the 'DelayedFlights-updated.csv' file stored in the S3 bucket by typing the following command:

**CREATE EXTERNAL TABLE IF NOT EXISTS airline\_db.airline\_delay (**

**Year INT,**

**Month INT,**

**DayofMonth INT,**

**DayOfWeek INT,**

**DepTime INT,**

**CRSDepTime INT,**

**ArrTime INT,**

**CRSArrTime INT,**

**UniqueCarrier STRING,**

**FlightNum INT,**

**TailNum STRING,**

**ActualElapsedTime INT,**

**CRSElapsedTime INT,**

**AirTime INT,**

**ArrDelay INT,**

**DepDelay INT,**

**Origin STRING,**

**Dest STRING,**

**Distance INT,**

**TaxiIn INT,**

**TaxiOut INT,**

**Cancelled INT,**

**CancellationCode STRING,**

**Diverted INT,**

**CarrierDelay INT,**

**WeatherDelay INT,**

**NASDelay INT,**

**SecurityDelay INT,**

**LateAircraftDelay INT**

**)**

**ROW FORMAT DELIMITED**

**FIELDS TERMINATED BY ','**

**LINES TERMINATED BY '\n'**

**LOCATION 's3://your-bucket-name/DelayedFlights-updated.csv'**

**TBLPROPERTIES ('skip.header.line.count'='1');**

This will create an external table named 'airline\_delay' in the 'airline\_db' database. The table schema is defined based on the structure of the 'DelayedFlights-updated.csv' file. The table is stored in the S3 bucket location specified in the 'LOCATION' clause. The 'skip.header.line.count' property is used to skip the header row of the CSV file.

5) Verify that the table is created successfully by typing the following command:

**SHOW TABLES;**

SHOW TABLES;This will show the list of tables in the 'airline\_db' database. You should see the 'airline\_delay' table in the list.

Now, we can run HiveQL queries on the 'airline\_delay' table to analyze the Airline Delay data.

1. Year-wise carrier delay from 2003-2010

**SELECT Year, SUM(CarrierDelay) AS TotalCarrierDelay**

**FROM airline\_db.airline\_delay**

**WHERE Year BETWEEN 2003 AND 2010**

**GROUP BY Year**

**ORDER BY Year ASC;**

1. Year-wise NAS delay from 2003-2010

**SELECT Year, SUM(NASDelay) AS TotalNASDelay**

**FROM airline\_db.airline\_delay**

**WHERE Year BETWEEN 2003 AND 2010**

**GROUP BY Year**

**ORDER BY Year ASC;**

3)Year wise Weather delay from 2003-2010

**SELECT Year, SUM(WeatherDelay) AS TotalWeatherDelay**

**FROM airline\_db.airline\_delay**

**WHERE Year BETWEEN 2003 AND 2010**

**GROUP BY Year**

**ORDER BY Year ASC**;

4) Year-wise late aircraft delay from 2003-2010

**SELECT Year, SUM(LateAircraftDelay) AS TotalLateAircraftDelay**

**FROM airline\_db.airline\_delay**

**WHERE Year BETWEEN 2003 AND 2010**

**GROUP BY Year**

**ORDER BY Year ASC;**

5) Year-wise security delay from 2003-2010

**SELECT Year, SUM(SecurityDelay) AS TotalSecurityDelay**

**FROM airline\_db.airline\_delay**

**WHERE Year BETWEEN 2003 AND 2010**

**GROUP BY Year**

**ORDER BY Year ASC;**