## Hive Query

**To start the hive** => hive

**To create table** =>

```
CREATE TABLE IF NOT EXISTS delay_flights (
 id int,
 Year int,
 Month int,
 DayofMonth int,
 DayOfWeek int,
 DepTime string,
 CRSDepTime string,
 ArrTime string,
 CRSArrTime string,
 UniqueCarrier string,
 FlightNum int,
 TailNum string,
 ActualElapsedTime string,
 CRSElapsedTime string,
 AirTime string,
 ArrDelay string,
 DepDelay string,
 Origin string,
 Dest string,
 Distance int,
 TaxiIn int,
 TaxiOut int,
 Cancelled int,
 CancellationCode string,
 Diverted int,
 CarrierDelay string,
 WeatherDelay string,
 NASDelay string,
 SecurityDelay string,
 LateAircraftDelay string)
 COMMENT 'Delay flights Table'
 ROW FORMAT DELIMITED
 FIELDS TERMINATED BY ',';
```

**To load the data from s3** =>

```
LOAD DATA INPATH 's3://bigdataassignm/DelayedFlights-updated.csv' INTO TABLE delay_flights;
```

**To print the names of the columns as headers in result sets** =>

```
set hive.cli.print.header=true;
```

```
SELECT Year, avg((CarrierDelay /ArrDelay)*100) from delay_flights WHERE Year>=2003 AND Year<=2010
GROUP BY Year ORDER BY Year;

SELECT Year, avg((NASDelay /ArrDelay)*100) from delay_flights WHERE Year>=2003 AND Year<=2010
GROUP BY Year ORDER BY Year;

SELECT Year, avg((WeatherDelay /ArrDelay)*100) from delay_flights WHERE Year>=2003 AND Year<=2010
GROUP BY Year ORDER BY Year;

SELECT Year, avg((SecurityDelay /ArrDelay)*100) from delay_flights WHERE Year>=2003 AND Year<=2010
GROUP BY Year ORDER BY Year;

SELECT Year, avg((LateAircraftDelay /ArrDelay)*100) from delay_flights WHERE Year>=2003 AND
Year<=2010 GROUP BY Year ORDER BY Year;
```