IS669 – Group Project

- Asritha Vemireddy
- Lakshmi Akshara Meka
- Darpan Patel
- Sarvesh

Project Overview

Flight Delay Analysis

- Analyze flight delay data for 2001
- Data source: CSV file from Harvard Dataverse
- Dataset: Flights from 1987 to 2008

Objectives:

- Identify airports with highest delays
- Determine airlines with highest delays
- Compare arrival vs. departure delays
- Uncover patterns and insights

Methodology:

- Load data into Hive table on EMR cluster
- Execute SQL-like queries for analysis

Data Loading and Sample

Data Loading:

- Uploaded CSV file to Amazon S3 bucket
- Created external Hive table pointing to S3 location
- Used CREATE EXTERNAL TABLE statement

```
hive> CREATE EXTERNAL TABLE asritharedddy (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, DepDe
lay INT, Origin STRING, Dest STRING, Distance INT, TaxiIn INT, TaxiOut INT, Cancelled INT, Ca
ncellationCode STRING, Diverted INT, CarrierDelay INT, WeatherDelay INT, NASDelay INT, Securi
tyDelay INT, LateAircraftDelay INT)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION 's3://grouppr
oject2001/data/';
OK
Time taken: 1.441 seconds
```

Loaded data into table and Verified data loading with SELECT query

```
hive> LOAD DATA INPATH 's3://groupproject2001/2001.csv' INTO TABLE asrithareddy;
Loading data to table default.asrithareddy
OK
Time taken: 3.683 seconds
hive> SELECT * FROM asrithareddy LIMIT 5;
OK
NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    UniqueCarrier   NULL    TailN
um      NULL    NULL    NULL    NULL    NULL    Origin  Dest    NULL    NULL    NULL    NULLC
ancellationCode NULL    NULL    NULL    NULL    NULL    NULL
2001    1       17      3       1806    1810    1931    1934    US      375     N700□□  85  8
4       60      -3      -4      BWI     CLT     361     5       20      0       NA      0   N
ULL     NULL    NULL    NULL    NULL
2001    1       18      4       1805    1810    1938    1934    US      375     N713□□  93  8
4       64      4       -5      BWI     CLT     361     9       20      0       NA      0   N
ULL     NULL    NULL    NULL    NULL
2001    1       19      5       1821    1810    1957    1934    US      375     N702□□  96  8
4       80      23      11      BWI     CLT     361     6       10      0       NA      0   N
ULL     NULL    NULL    NULL    NULL
2001    1       20      6       1807    1810    1944    1934    US      375     N701□□  97  8
4       66      10      -3      BWI     CLT     361     4       27      0       NA      0   N
ULL     NULL    NULL    NULL    NULL
Time taken: 1.247 seconds, Fetched: 5 row(s)
```

Top 3 Airports with Highest Delay Time (in hours)

Query:

```
hive> SELECT Origin, SUM(ArrDelay + DepDelay) / 60.0 AS TotalDelayHours FROM asritharedddy WHE
RE Year = 2001 GROUP BY Origin ORDER BY TotalDelayHours DESC LIMIT 3;
Query ID = hadoop_20240329000824_4c2489ba-5e99-4bc2-b1cc-0c743da489ab
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1711668553416_0004)
```

Result:

```
--------------------------------------------------------------------------------
-
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 21.39 s
--------------------------------------------------------------------------------
-
OK
ORD     111174.083333
DFW     80119.300000
ATL     67723.116667
Time taken: 24.358 seconds, Fetched: 3 row(s)
hive>
```

Top 3 Carriers with Highest Delay Time

Query:

```
hive> SELECT UniqueCarrier, SUM(ArrDelay + DepDelay) / 60.0 AS TotalDelayHours FROM asrithare
ddy WHERE Year = 2001 GROUP BY UniqueCarrier ORDER BY TotalDelayHours DESC LIMIT 5;
Query ID = hadoop_20240329001311_a6eca646-784e-4657-b20f-17c0374f0cf1
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1711668553416_0004)
```

Result:

```
-
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED

--------------------------------------------------------------------------------
-
Map 1 .......... container     SUCCEEDED     13         13        0        0       0       0
 --------------------------------------------------------------------------------
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0

Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0

--------------------------------------------------------------------------------
-
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 20.72 s
--------------------------------------------------------------------------------
-
OK
UA      218840.750000
WN      212004.033333
AA      168121.516667
DL      166055.233333
MQ      137420.033333
Time taken: 21.085 seconds, Fetched: 5 row(s)
hive>
```

Arrival vs. Departure Delays

Query:

```
hive> SELECT 'ArrivalDelays' AS DelayType, SUM(ArrDelay)/ 60.0 AS TotalDelayHours FROM asrith
areddy WHERE Year = 2001 UNION ALL SELECT 'DepartureDelays',SUM(DepDelay) / 60.0 FROM asritha
reddy WHERE Year = 2001;
Query ID = hadoop_20240329001738_37fc2d2c-c086-4228-867a-f75b922cc83d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1711668553416_0004)
```

Result:

```
----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     13         13        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Map 4 .......... container     SUCCEEDED     13         13        0        0       0       0
Reducer 5 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 20.92 s
----------------------------------------------------------------------------------------------
OK
ArrivalDelays   527364.800000
DepartureDelays 779681.566667
Time taken: 21.515 seconds, Fetched: 2 row(s)
hive>
```

Key Observations and Findings

- The airport with the highest delay time (in hours) for the assigned year was ORD (Chicago O'Hare International Airport), followed by DFW (Dallas/Fort Worth International Airport) and ATL (Hartsfield-Jackson Atlanta International Airport).
- Among the carriers, UA (United Airlines) experienced the highest delay time, closely followed by WN (Southwest Airlines) and AA (American Airlines).
- Departure delays were more prevalent than arrival delays for the assigned year, with

Conclusion

- The analysis of flight delay data for 2001 revealed significant delays experienced by major airports and carriers, with ORD, DFW, and ATL being the most affected airports, and UA, WN, and AA being the carriers with the highest delay times.
- Departure delays were more prominent than arrival delays, suggesting potential areas for improvement in airline operations and scheduling.
- ***Limitations:*** The analysis focused solely on the assigned year and did not consider factors like weather conditions or specific flight routes, which could have contributed to delays.
- ***Future Improvements:*** Incorporating additional data sources, such as weather data and flight schedules, could provide deeper insights into the root causes of