HUMBER INSTITUTE OF TECHNOLOGY
AND ADVANCED LEARNING
(HUMBER COLLEGE)

ASSIGNMENT: Big Data 1 - Group Project

Team Members:

First Name	Last Name	Student Number
Anjali	Patel	N01510772
Meutia	Putri	N01510783
Sairohit	Chowdhary	N01510624

Course: Big Data 1 - ITE-5200-0LA
Submitted to: Prabhpreet Sidhu

Submission Date:
07/28/2022

Introduction

The Covid-19 pandemic lasted more than two years and forced countries around the globe to place specific restrictions. Almost if not all industries are significantly impacted; however, one sector, in particular, was hit the hardest. According to the S&P Global Market Intelligence analysis, the airline industry is the most impacted industry, mainly because of the sudden border closures, mass grounding of air traffic, and lockdown policies all around the world, which is disastrous for the aviation industry (Haydon & Kumar, 2020).

As of now, when life begins to return to normal and most countries have lifted travel restrictions and started to open their borders for travelers, a new problem arises. Since last year, the travel industry, notably airlines, has been understaffed, making it challenging to manage the travel spikes. Domestic and international airlines and airports are still trying to completely staff up, making them more vulnerable to delays (Nguyen, 2022). According to Josephs (2022), the flight delays and cancellations this year were significantly higher than before the pandemic, with more than 1,100 U.S. flights canceled and more than 12,000 flights delayed during the hectic fourth of July weekend.

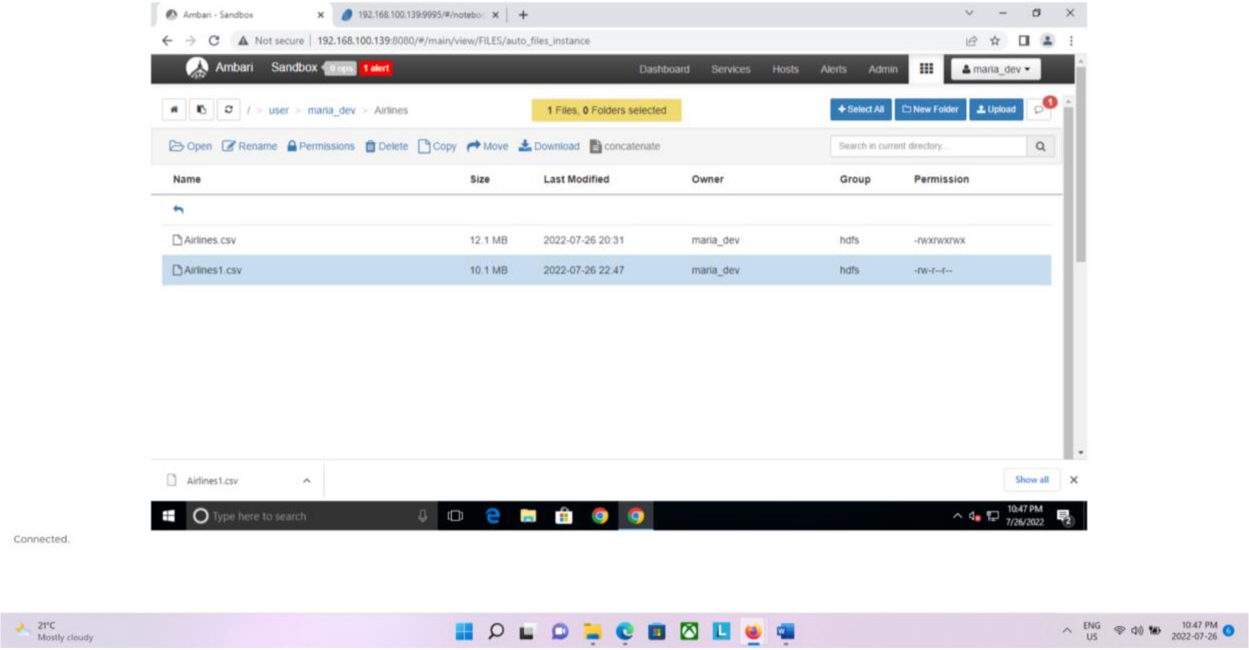
Delays impact not only passengers but airlines as well. They induce stress and anger among passengers through increased travel time and expenses on meals and hotels. Furthermore, they undermine the goal of air travel, which are quick and economical. On the other hand, airlines incur extra crew costs and additional costs associated with accommodating interrupted customers and aircraft re-positioning because airline fleet and crew timetables mainly rely on scheduled timings (Trefis, 2016).

To minimize customers' dissatisfaction, one of the most well-known online travel agencies, Expedia, has hired us as industry analysts to provide insights on flight delays among 16 airlines in North America. Our analysis and insights are provided below.

Coding and Analysis

The first step is to upload the two CSV files (airlines.csv and airlines1.csv) downloaded from Kaggle onto HDFS to make it accessible on Zeppelin, Hive and Hbase so we can easily manipulate the data, query the data and get meaningful visualizations and ultimately derive insights for the business.

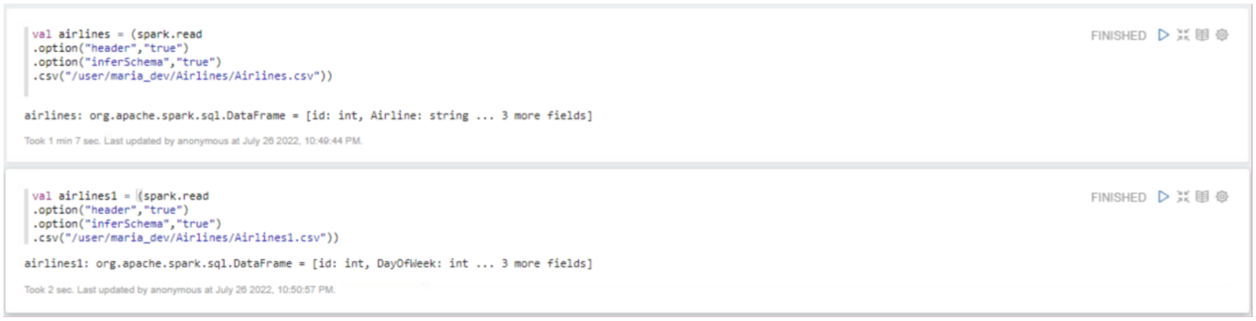
The screenshot of the file is successfully uploaded in the HDFS file system is attached below:



Name	Size	Last Modified	Owner	Group	Permission
Airlines.csv	12.1 MB	2022-07-26 20:31	maria_dev	hdfs	-rwxrwxrwx
Airlines1.csv	10.1 MB	2022-07-26 22:47	maria_dev	hdfs	-rw-r--r--

Now, to perform any type of analytical operations on the file, it is required to create a data frame in zeppelin by reading the CSV file which we saved in HDFS.

Code snip for the two data frames:



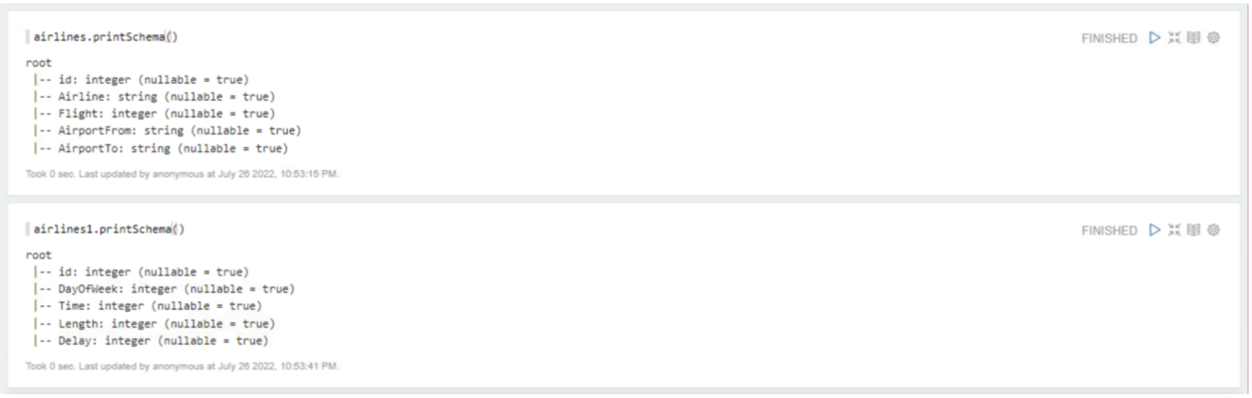
```
val airlines = (spark.read
.option("header","true")
.option("inferSchema","true")
.csv("/user/maria_dev/Airlines/Airlines.csv"))

airlines: org.apache.spark.sql.DataFrame = [id: int, Airline: string ... 3 more fields]
```

Took 1 min 7 sec. Last updated by anonymous at July 26 2022, 10:49:44 PM.

```
val airlines1 = (spark.read
.option("header","true")
.option("inferSchema","true")
.csv("/user/maria_dev/Airlines/Airlines1.csv"))
airlines1: org.apache.spark.sql.DataFrame = [id: int, DayOfWeek: int ... 3 more fields]
```

Took 2 sec. Last updated by anonymous at July 26 2022, 10:50:57 PM.

The printing schema in a tree format:



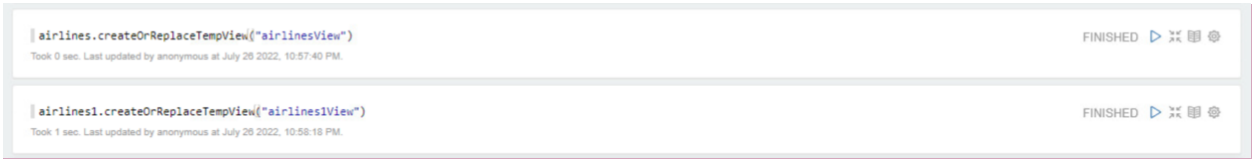
```
airlines.printSchema()

root
 |-- id: integer (nullable = true)
 |-- Airline: string (nullable = true)
 |-- Flight: integer (nullable = true)
 |-- AirportFrom: string (nullable = true)
 |-- AirportTo: string (nullable = true)
```

Took 0 sec. Last updated by anonymous at July 26 2022, 10:53:15 PM.

```
airlines1.printSchema()

root
 |-- id: integer (nullable = true)
 |-- DayOfWeek: integer (nullable = true)
 |-- Time: integer (nullable = true)
 |-- Length: integer (nullable = true)
 |-- Delay: integer (nullable = true)
```

Took 0 sec. Last updated by anonymous at July 26 2022, 10:53:41 PM.

Converting data frames into temp view:



```
airlines.createOrReplaceTempView("airlinesView")
```

Took 0 sec. Last updated by anonymous at July 26 2022, 10:57:40 PM.

```
airlines1.createOrReplaceTempView("airlines1View")
```

Took 1 sec. Last updated by anonymous at July 26 2022, 10:58:18 PM.

Data frame 1 (airlineView) has 5 attributes:

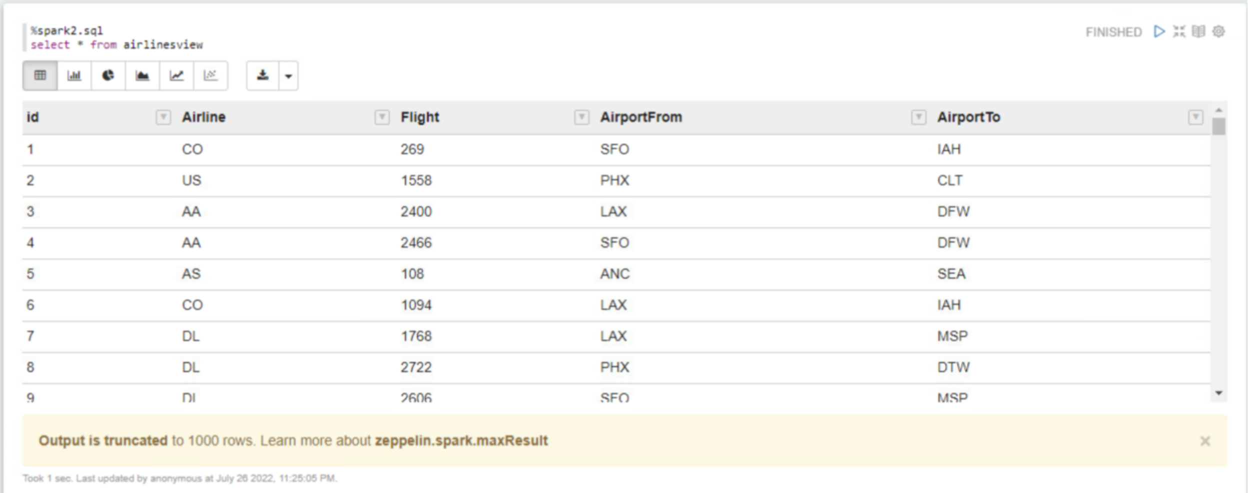
Attribute	**Data Type**	**Explanation**
ID	INT	ID for each row/entry
Airline	String	Abbreviation of Airline’s name
Flight	INT	Flight number
AirportFrom	String	Airport code of departing flight
AirportTo	String	Airport code of arriving flight

Data frame 2 (airline1View) has 5 attributes:

Attribute	**DataType**	**Explanation**
Id	INT	id for each row/entry
DayOfWeek	INT	1 = Monday 2 = Tuesday 3= Wednesday 4= Thursday 5= Friday 6=Saturday 7= Sunday
Time	INT	Scheduled time of departure
Length	INT	Duration in Length
Delay	Binary	0= Not Delayed 1 = Delayed

Following the above steps, the data frames are ready to use and printed for an easier understanding when querying the data in Zeppelin.

Here are the screenshots of the temporary views we created in Zeppelin with the data:

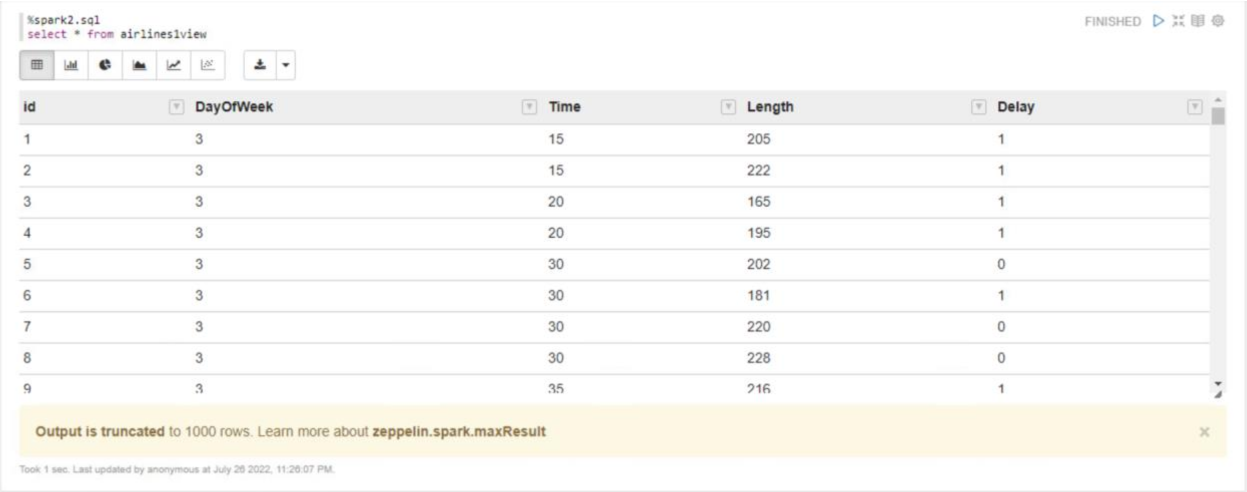


```
%spark2.sql
select * from airlinesview
```

id	Airline	Flight	AirportFrom	AirportTo
1	CO	269	SFO	IAH
2	US	1558	PHX	CLT
3	AA	2400	LAX	DFW
4	AA	2466	SFO	DFW
5	AS	108	ANC	SEA
6	CO	1094	LAX	IAH
7	DL	1768	LAX	MSP
8	DL	2722	PHX	DTW
9	DL	2606	SFO	MSP

Output is truncated to 1000 rows. Learn more about **zeppelin.spark.maxResult**

Took 1 sec. Last updated by anonymous at July 26 2022, 11:25:05 PM.



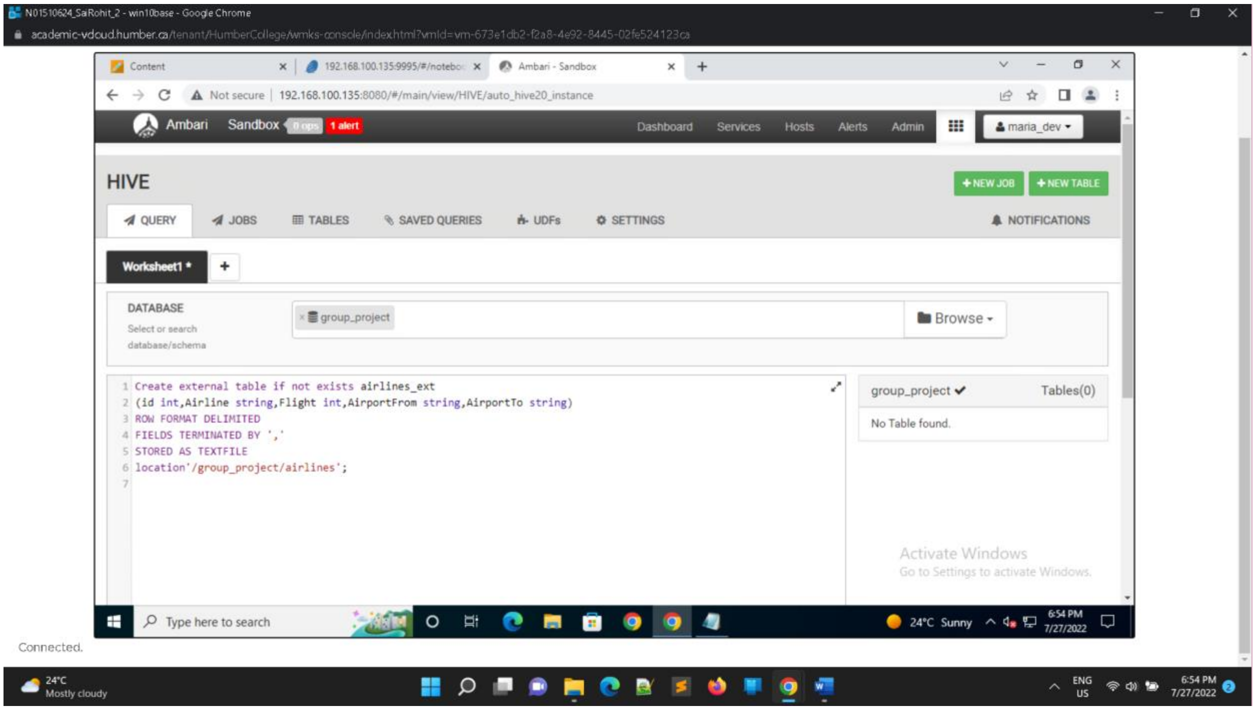
```
%spark2.sql
select * from airlines1view
```

id	DayOfWeek	Time	Length	Delay
1	3	15	205	1
2	3	15	222	1
3	3	20	165	1
4	3	20	195	1
5	3	30	202	0
6	3	30	181	1
7	3	30	220	0
8	3	30	228	0
9	3	35	216	1

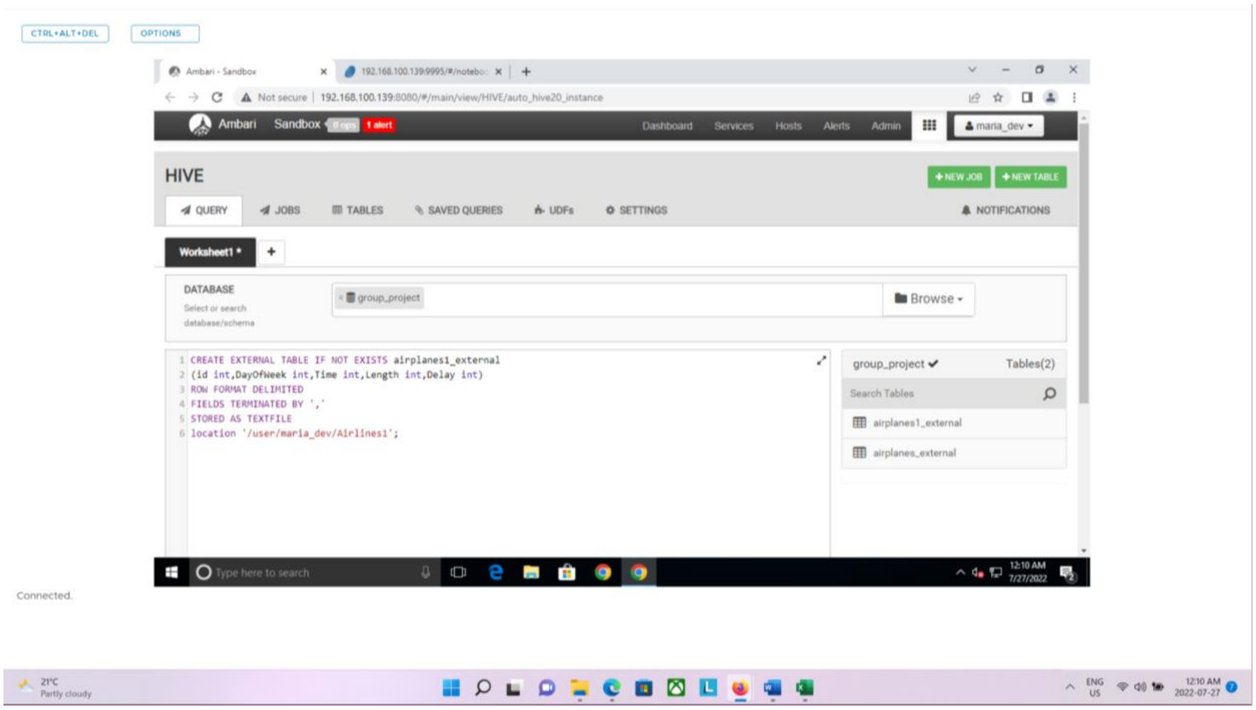
Output is truncated to 1000 rows. Learn more about **zeppelin.spark.maxResult**

Took 1 sec. Last updated by anonymous at July 26 2022, 11:26:07 PM.

Apart from zeppelin, we created external and internal tables for both files in hive to read the data in CSV format which is uploaded to HDFS

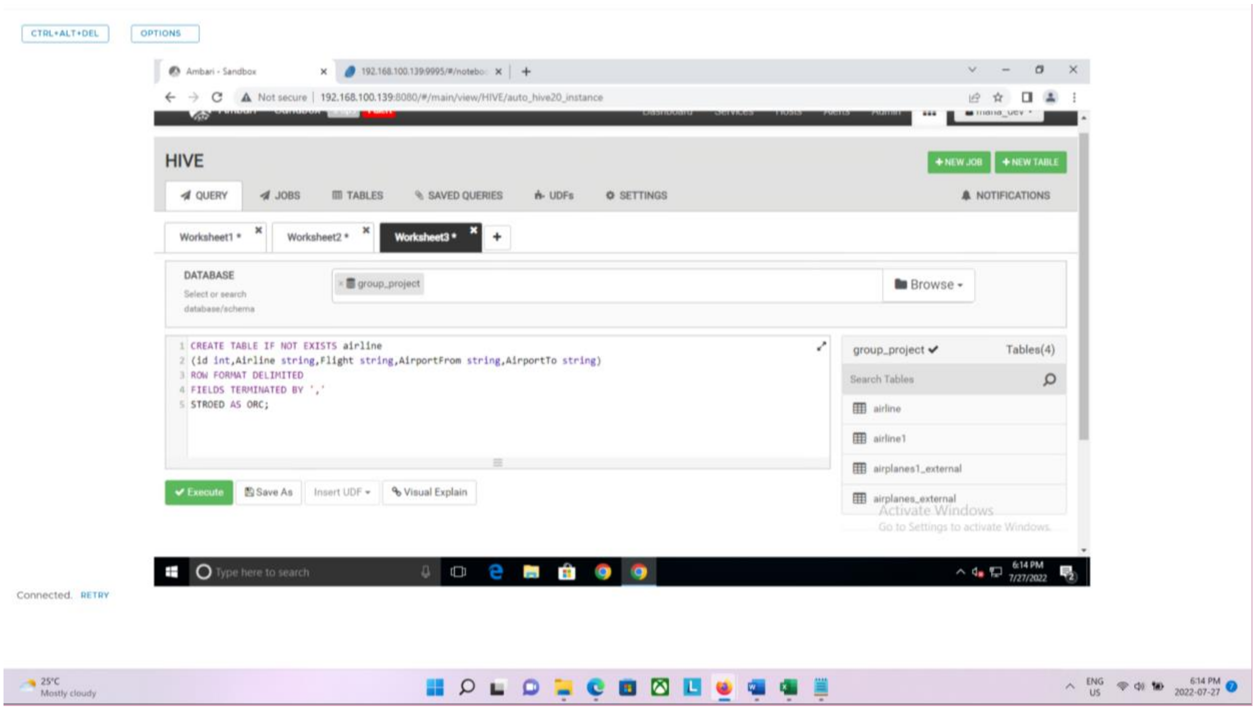


```
Create external table if not exists airlines_ext
(id int,Airline string,Flight int,AirportFrom string,AirportTo string)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
location'/group_project/airlines';
```

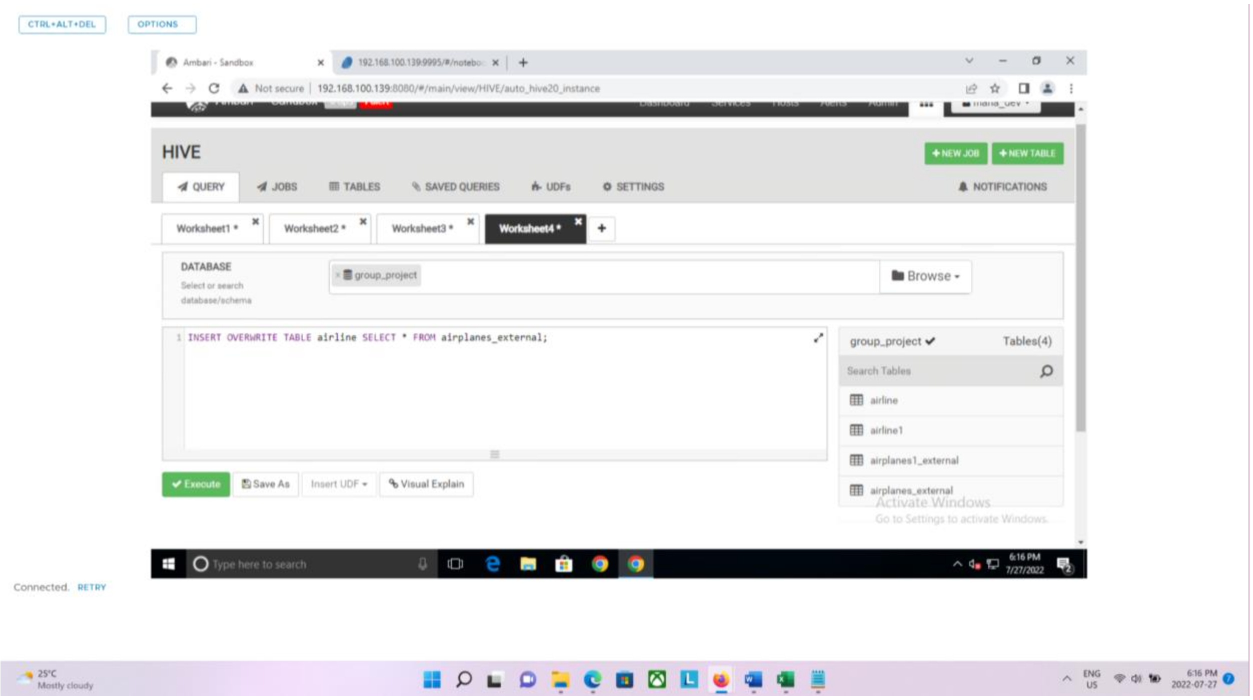


```
CREATE EXTERNAL TABLE IF NOT EXISTS airplanes1_external
(id int,DayOfWeek int,Time int,Length int,Delay int)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
location '/user/maria_dev/Airlines1';
```

Creating internal tables in hive and input data from the external tables



```
CREATE TABLE IF NOT EXISTS airline
(id int,Airline string,Flight string,AirportFrom string,AirportTo string)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STROED AS ORC;
```

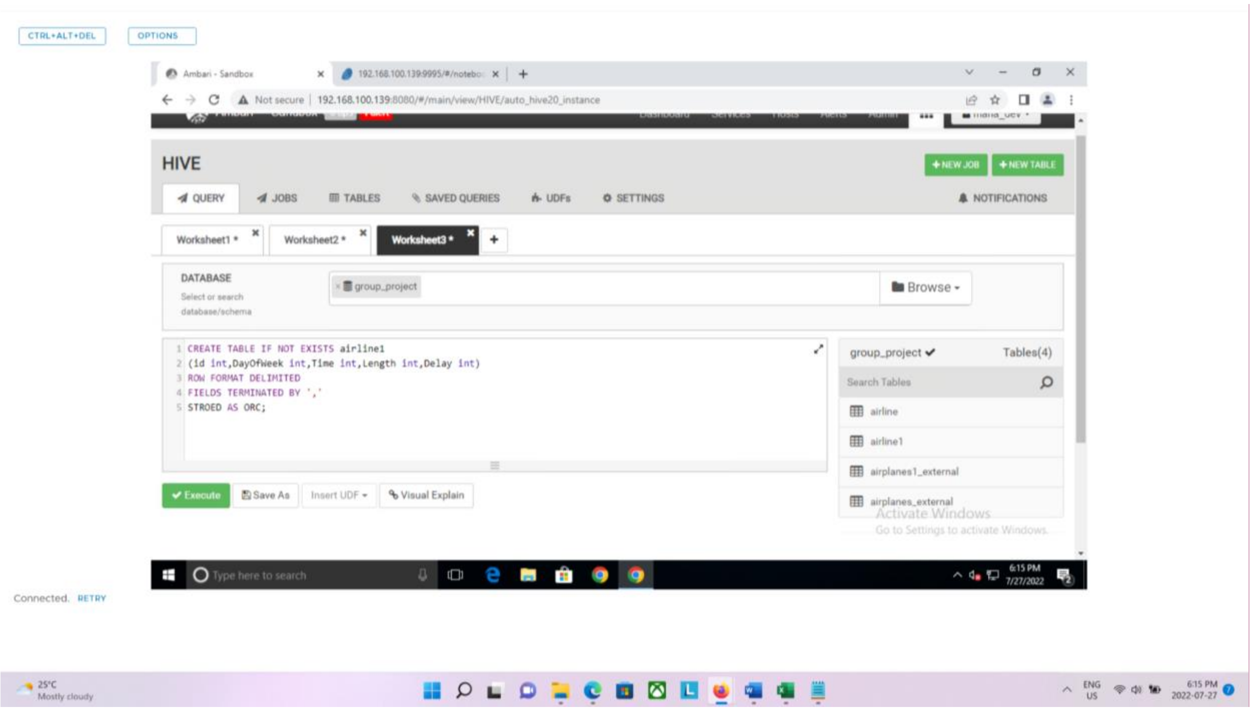


DATABASE: group_project

```
INSERT OVERWRITE TABLE airline SELECT * FROM airplanes_external;
```

group_project ✓ Tables(4)

- airline
- airline1
- airplanes1_external
- airplanes_external

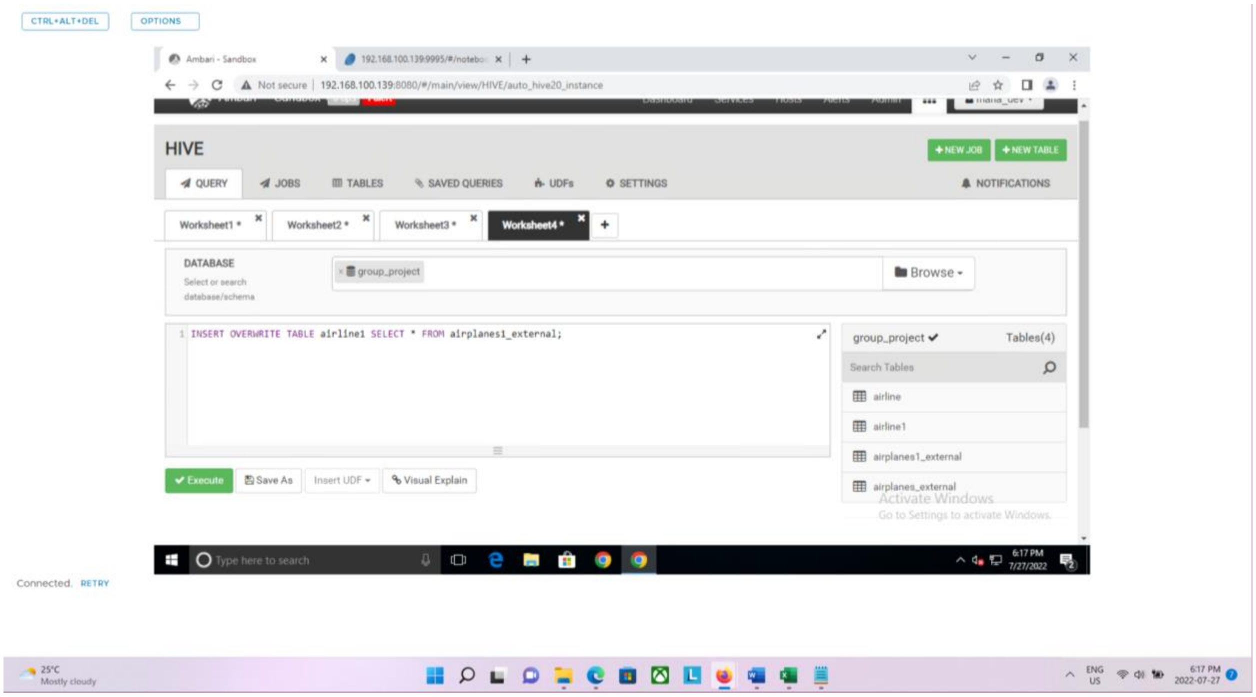


DATABASE: group_project

```
CREATE TABLE IF NOT EXISTS airline1
(id int,DayOfWeek int,Time int,Length int,Delay int)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STROED AS ORC;
```

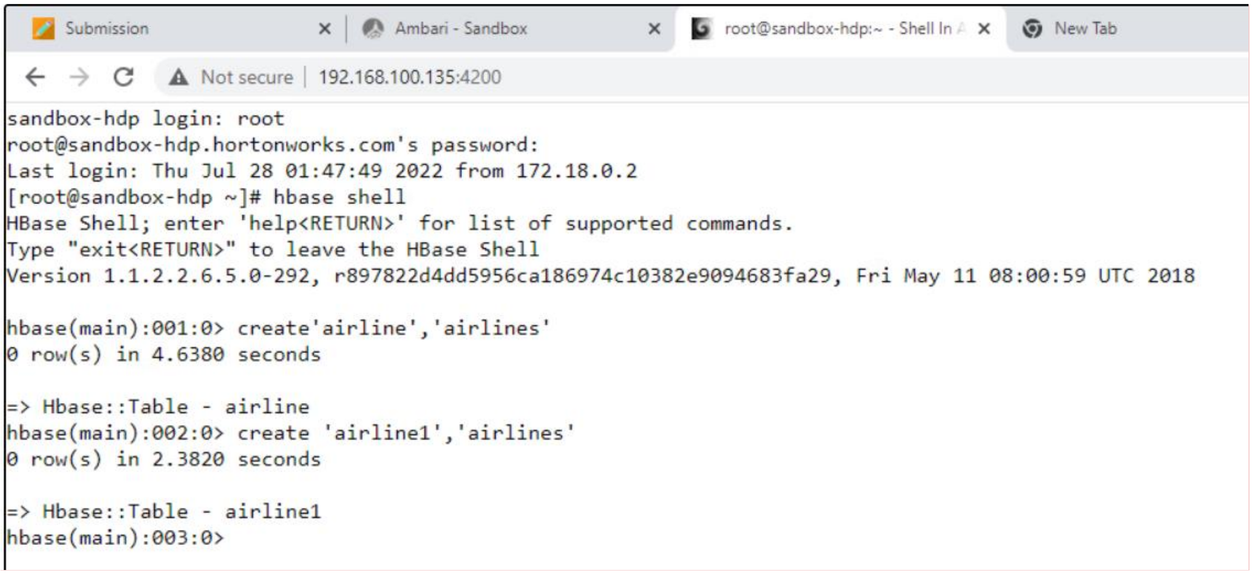
group_project ✓ Tables(4)

- airline
- airline1
- airplanes1_external
- airplanes_external



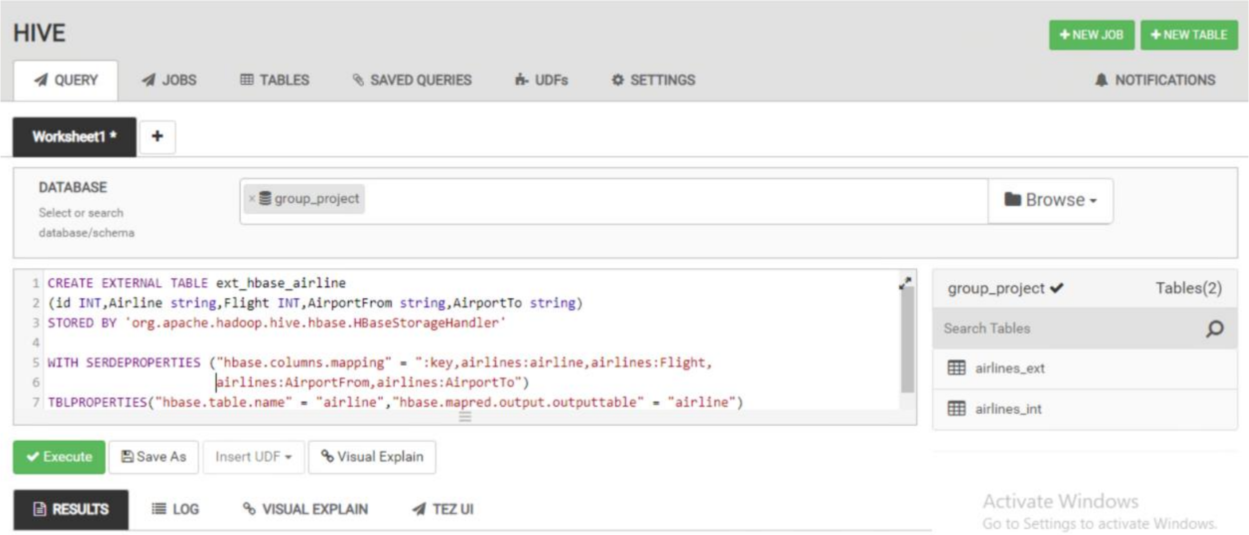
```
INSERT OVERWRITE TABLE airline1 SELECT * FROM airplanes1_external;
```

Now we created two HBase tables:

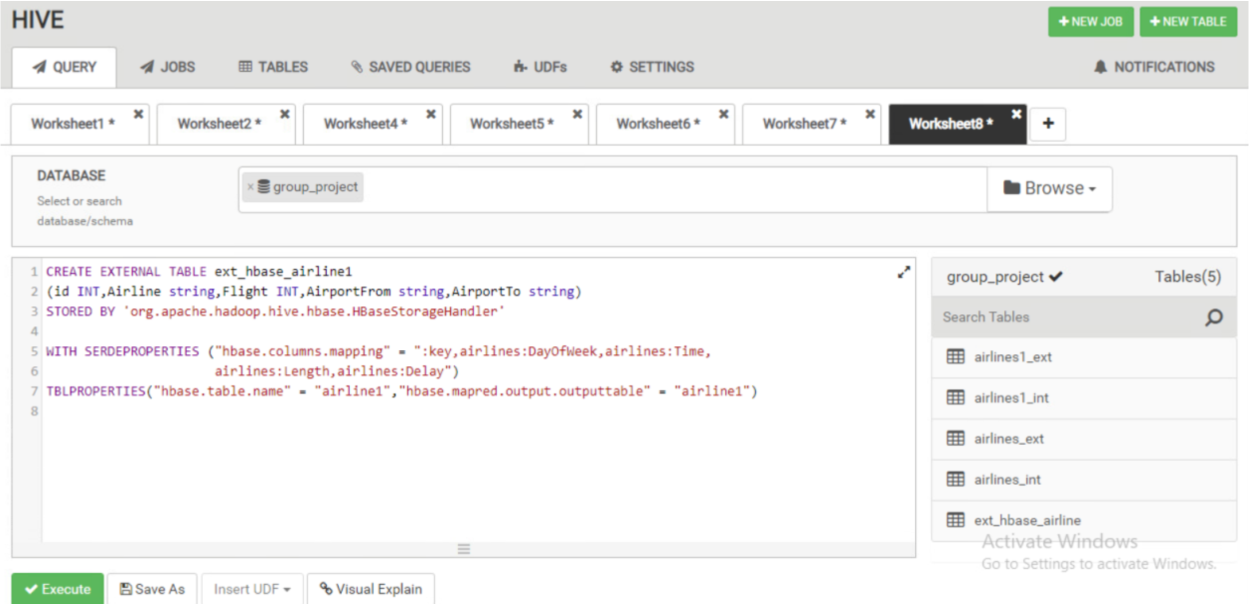

Submission
Ambari - Sandbox
root@sandbox-hdp:~ - Shell In A
New Tab
Not secure | 192.168.100.135:4200

```
sandbox-hdp login: root
root@sandbox-hdp.hortonworks.com's password:
Last login: Thu Jul 28 01:47:49 2022 from 172.18.0.2
[root@sandbox-hdp ~]# hbase shell
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.1.2.2.6.5.0-292, r897822d4dd5956ca186974c10382e9094683fa29, Fri May 11 08:00:59 UTC 2018

hbase(main):001:0> create'airline','airlines'
0 row(s) in 4.6380 seconds

=> Hbase::Table - airline
hbase(main):002:0> create 'airline1','airlines'
0 row(s) in 2.3820 seconds

=> Hbase::Table - airline1
hbase(main):003:0>
```

Creating and connecting data from Hive external tables to Hbase tables:

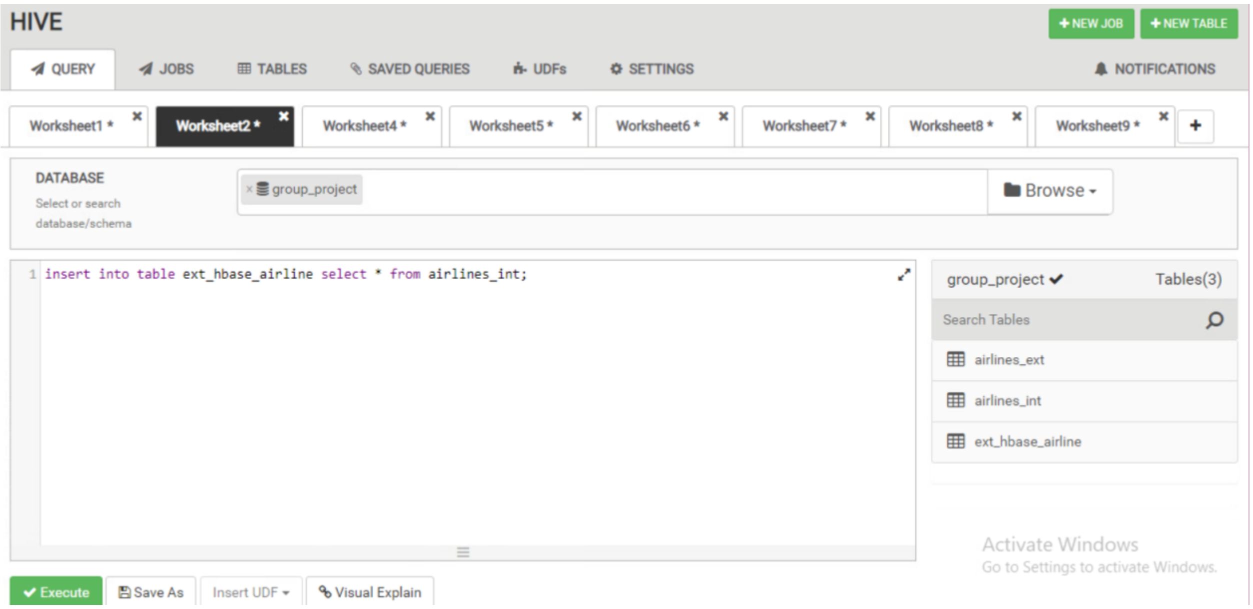


```
CREATE EXTERNAL TABLE ext_hbase_airline
(id INT,Airline string,Flight INT,AirportFrom string,AirportTo string)
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'

WITH SERDEPROPERTIES ("hbase.columns.mapping" = ":key,airlines:airline,airlines:Flight,
                      airlines:AirportFrom,airlines:AirportTo")
TBLPROPERTIES("hbase.table.name" = "airline","hbase.mapred.output.outputtable" = "airline")
```



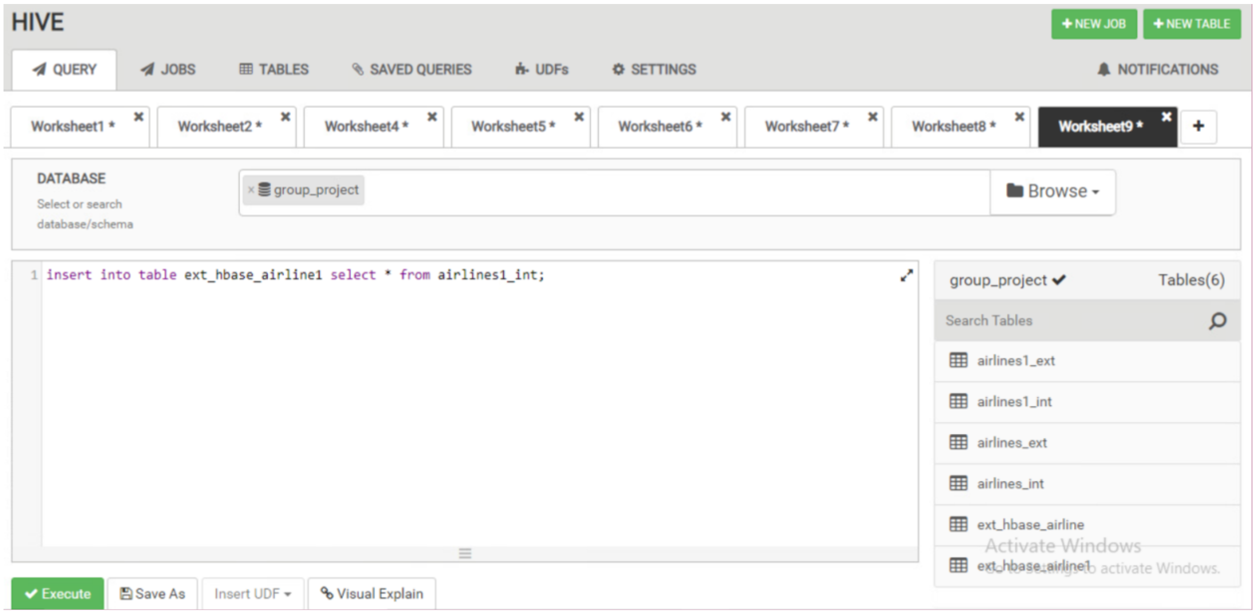
```
CREATE EXTERNAL TABLE ext_hbase_airline1
(id INT,Airline string,Flight INT,AirportFrom string,AirportTo string)
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'

WITH SERDEPROPERTIES ("hbase.columns.mapping" = ":key,airlines:DayOfWeek,airlines:Time,
                      airlines:Length,airlines:Delay")
TBLPROPERTIES("hbase.table.name" = "airline1","hbase.mapred.output.outputtable" = "airline1")

```

Inserting data into HBase table



DATABASE
Select or search database/schema: group_project

```
insert into table ext_hbase_airline select * from airlines_int;
```



DATABASE
Select or search database/schema: group_project

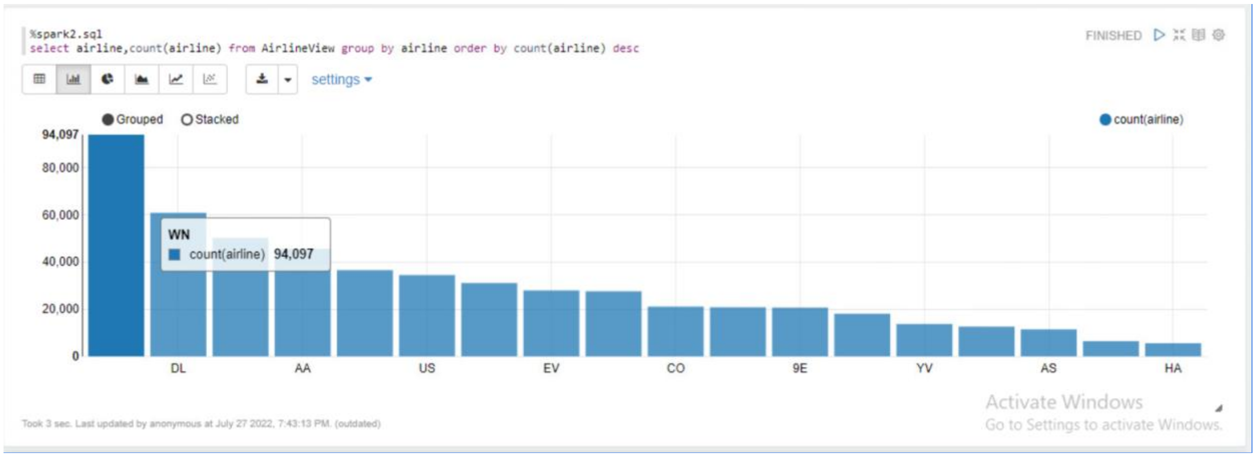
```
insert into table ext_hbase_airline1 select * from airlines1_int;
```

The data is now stored in Hbase so that it can be used for future batch processing as it is more scalable and faster than traditional RDBMS.

From the data above, we can answer the following 5 analytical questions which can help render visualizations and get insights to improve the business.

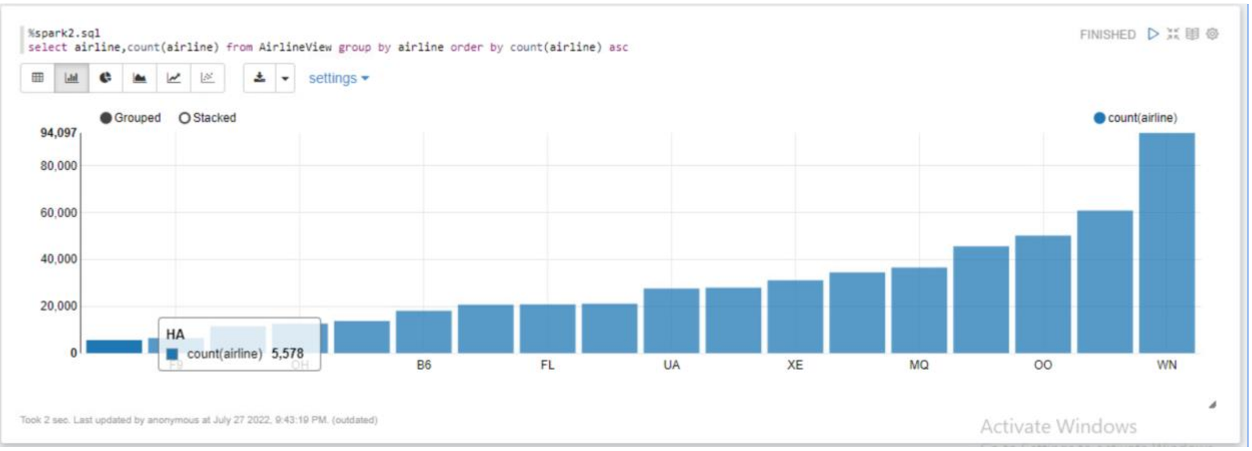
Analytics Questions

1. Which Airlines have the most and least Flights?



```
%spark2.sql
select airline,count(airline) from AirlineView group by airline order by count(airline) desc
```

Took 3 sec. Last updated by anonymous at July 27 2022, 7:43:13 PM. (outdated)



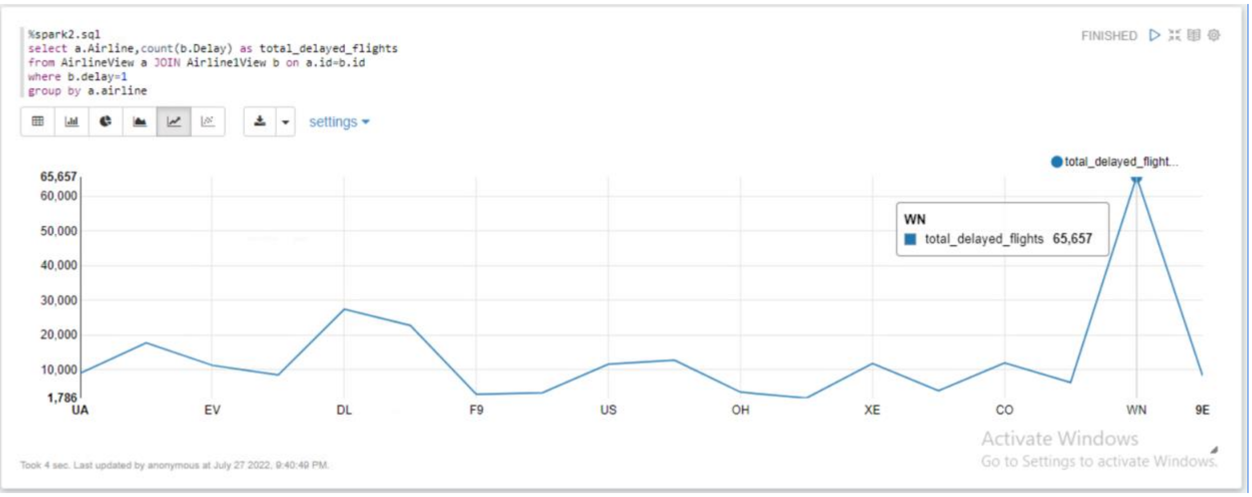
```
%spark2.sql
select airline,count(airline) from AirlineView group by airline order by count(airline) asc
```

Took 2 sec. Last updated by anonymous at July 27 2022, 9:43:19 PM. (outdated)

From the Bar chart above, it seems that Northwestern Airlines has the highest frequency of flights in North America, with flights over 94000.

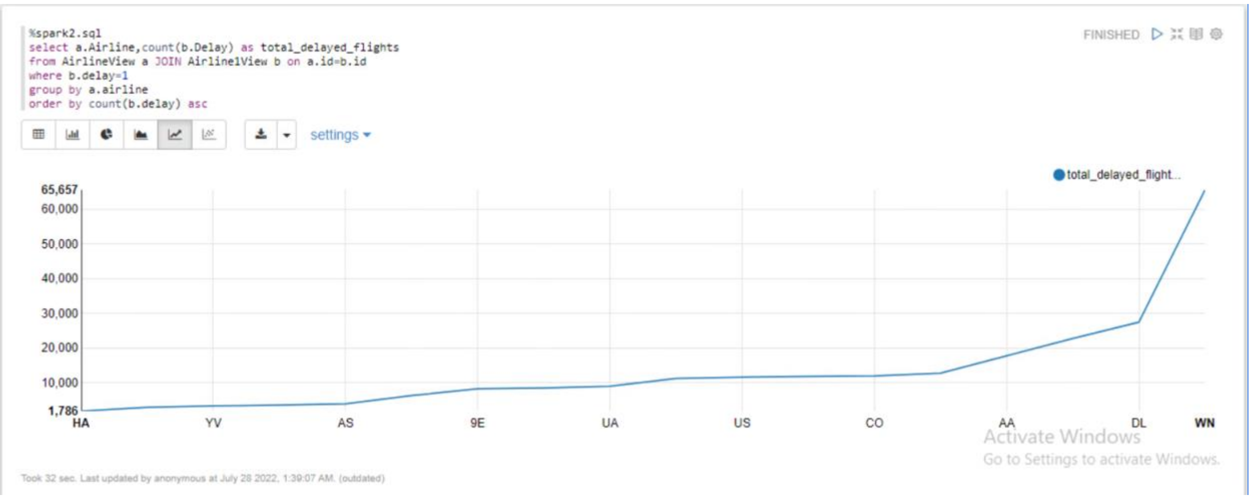
Meanwhile, Hawaiian Airlines had the least number of flights with only 5578 compared to the same period of time.

2. Which Airlines have the most and least Delays?



```
%spark2.sql
select a.Airline,count(b.Delay) as total_delayed_flights
from AirlineView a JOIN Airline1View b on a.id=b.id
where b.delay=1
group by a.airline
```

Took 4 sec. Last updated by anonymous at July 27 2022, 9:40:49 PM.

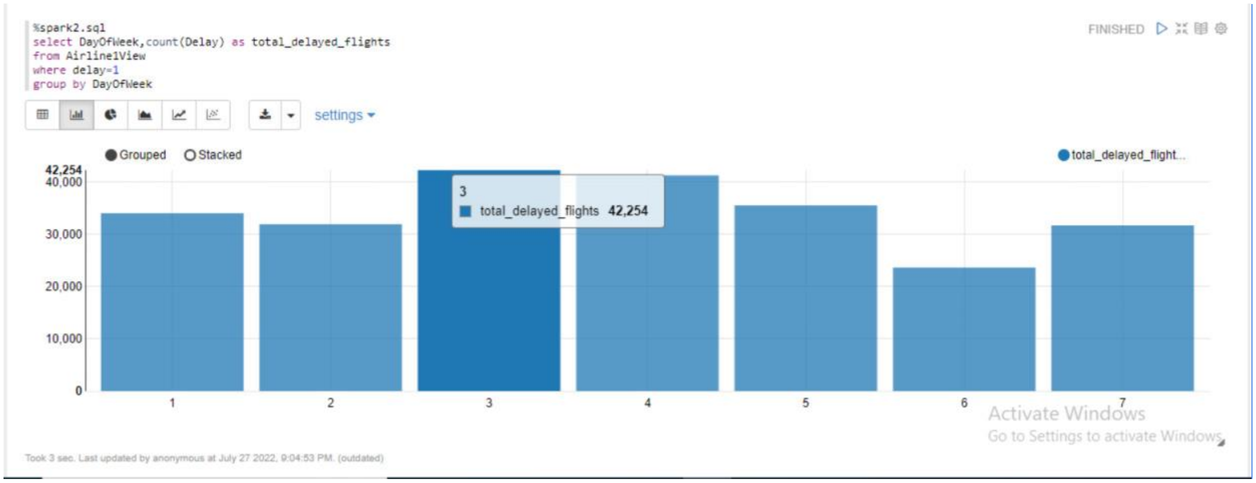


```
%spark2.sql
select a.Airline,count(b.Delay) as total_delayed_flights
from AirlineView a JOIN Airline1View b on a.id=b.id
where b.delay=1
group by a.airline
order by count(b.delay) asc
```

Took 32 sec. Last updated by anonymous at July 28 2022, 1:39:07 AM. (outdated)

We can see from the data visualisation that the majority of Northwestern Airlines flights with over 65k are delayed, while Hawaiian Airlines have the shortest wait times with just 1786 being delayed flights.

3. Does the delay depend on the flight’s departure day of the week?

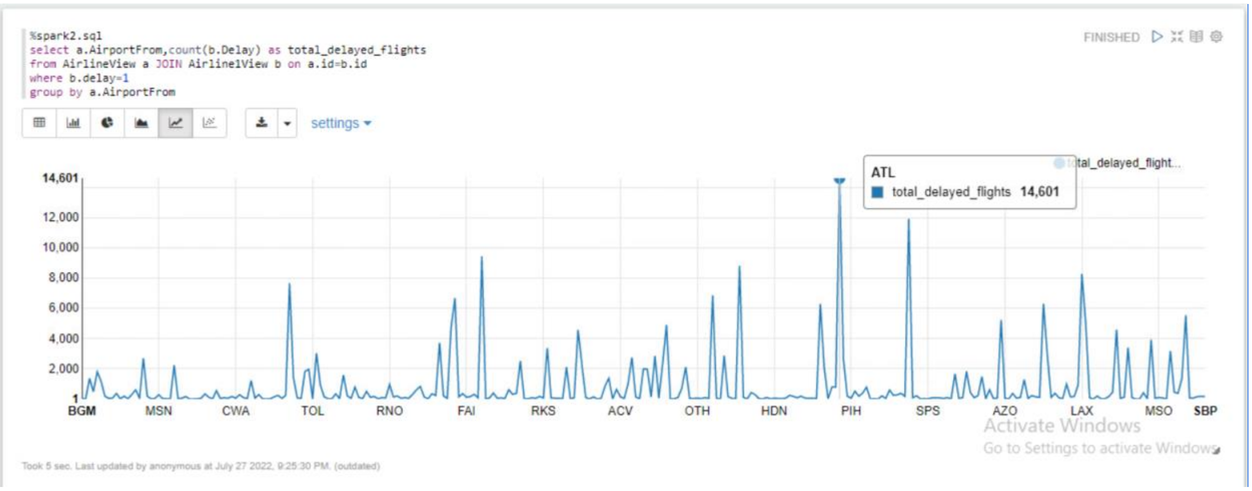


```
%spark2.sql
select DayOfWeek,count(Delay) as total_delayed_flights
from Airline1View
where delay=1
group by DayOfWeek
```

Took 3 sec. Last updated by anonymous at July 27 2022, 9:04:53 PM. (outdated)

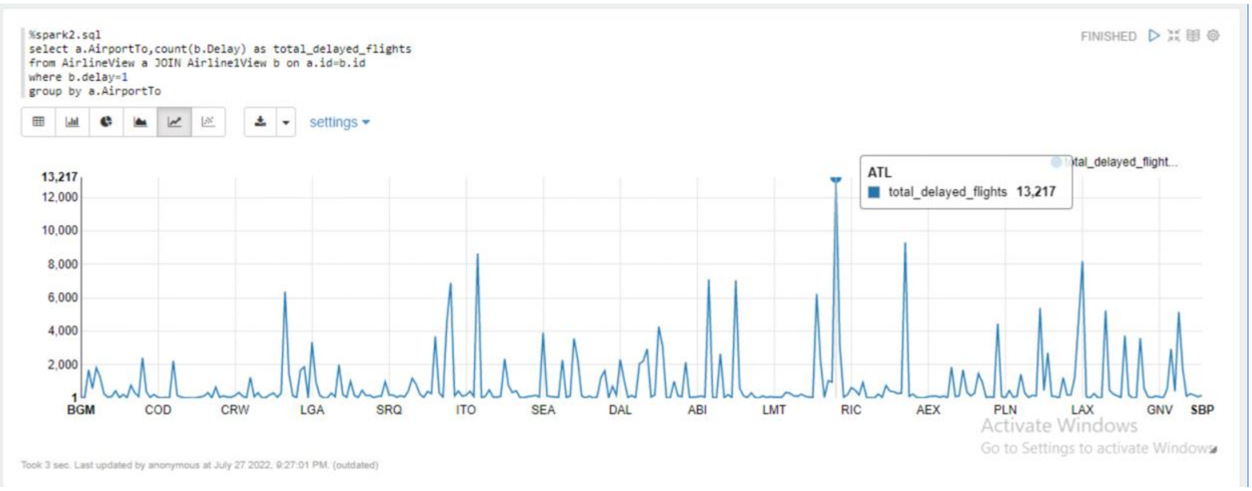
From our findings, Wednesdays and Thursdays are the most delayed days of the week compared to saturdays.

4. Which airports have the most delays based on the flight’s origin and destination Airport?



```
%spark2.sql
select a.AirportFrom,count(b.Delay) as total_delayed_flights
from AirlineView a JOIN Airline1View b on a.id=b.id
where b.delay=1
group by a.AirportFrom
```

Took 5 sec. Last updated by anonymous at July 27 2022, 9:25:30 PM. (outdated)

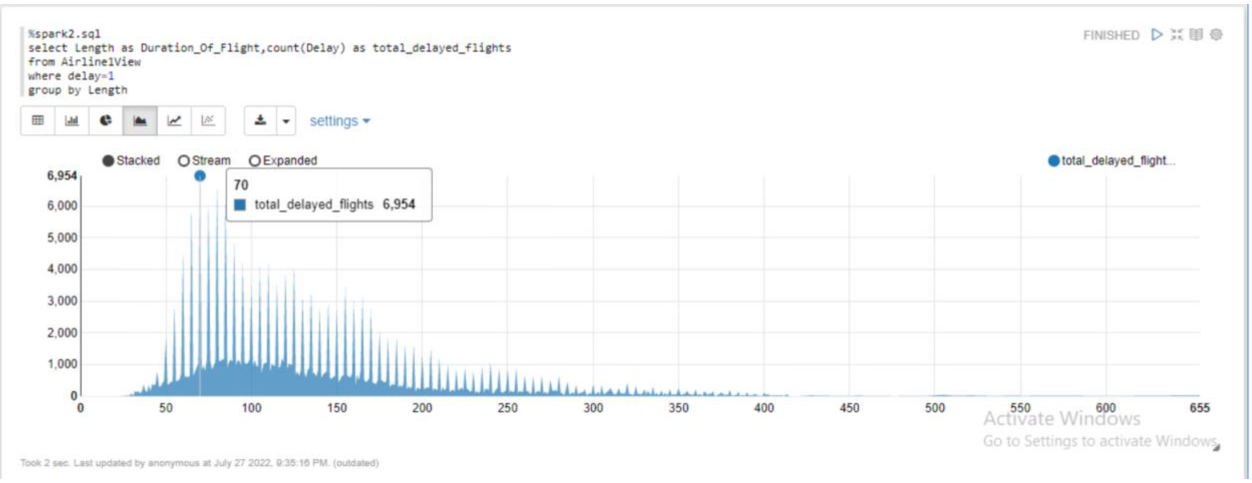


```
%spark2.sql
select a.AirportTo,count(b.Delay) as total_delayed_flights
from AirlineView a JOIN Airline1View b on a.id=b.id
where b.delay=1
group by a.AirportTo
```

Took 3 sec. Last updated by anonymous at July 27 2022, 9:27:01 PM. (outdated)

There are most delays at Atlanta International Airport for both departing and arriving planes.

5. Does the duration of the flight play a role in the flight being delayed?



```
%spark2.sql
select Length as Duration_Of_Flight,count(Delay) as total_delayed_flights
from Airline1View
where delay=1
group by Length
```

Took 2 sec. Last updated by anonymous at July 27 2022, 9:35:16 PM. (outdated)

Yes, from the visuals we can deduce that shorter flight duration ranging from 60 mins to 90 mins tend to experience delays the most.

Insights and Conclusion

Northwestern Airlines has the highest frequency of flights; however, it has the most amount of delays. Hawaiian Airlines have the least delay, but it may be because it has the least amount of flights in North America. Hence, we can say that the amount of flights delayed is directly proportional to the amount of the flights being operated by that particular airline at that period of time.

From our findings, we suggest that booking flights departing on Saturday or Tuesday should be promoted to minimize the chances of their flight being delayed.

While booking the flights, customers can avoid ATL or Atlanta Airport as a Layover Airport as they experience long wait times as an origin and destination airport in North America. Expedia should add a disclaimer for the customers who are intending to book flights from, to and through Atlanta stating that they are most likely to experience delays. Hence, Customers can make well advanced and informed decisions prior to flying.

Additionally, the length of the flights is important when discussing delays. Our analysis revealed that delays are frequently experienced by flights with duration between 60 and 90 minutes.

In conclusion, we as analysts suggest Expedia to optimize their website by using our insights to reduce customer complaints regarding flight delays. Hence, improving their customer satisfaction.

Appendix for the Data

List of Airlines

Alaska Airlines AS / ASA

American Airlines AA/AAL

Air Canada AC/ACA

Aeromexico AM / AMX

Continental Airlines CO / COA

Delta Airlines DL / DAL

FedEx FX / FDX

Hawaiian Airlines HA / HAL

Northwest Airlines NW / NWA/WN

Polar Air Cargo PO / PAC

Southwest Airlines SW / SWA

United Airlines UA / UAL

United Parcel (UPS) 5X / UPS

Virgin Atlantic VS / VIR

VivaAerobús VB / VIV

WestJet WS / WJ

List of Airports

ATL - Hartsfield-Jackson Atlanta International Airport - Georgia

AUS - Austin-Bergstrom International Airport - Texas

BNA - Nashville International Airport - Tennessee

BOS - Boston Logan International Airport - Massachusetts

BWI - Baltimore-Washington International Thurgood Marshall Airport - Washington

CLT - Charlotte Douglas International Airport - North Carolina

DAL - Dallas Love Field - Texas

DCA - Ronald Reagan Washington National Airport - Arlington, Virginia

DEN - Denver International Airport - Colorado

DFW - Dallas/Fort Worth International Airport - Texas

DTW - Detroit Metropolitan Airport - Michigan

EWR - Newark Liberty International Airport - New Jersey

FLL - Fort Lauderdale–Hollywood International Airport - Florida

HNL - Daniel K. Inouye International Airport - Honolulu, Hawaii

HOU - William P. Hobby Airport - Houston, Texas

IAD - Dulles International Airport - Virginia

IAH - George Bush Intercontinental Airport - Houston, Texas

JFK - John F. Kennedy International Airport - Queens, New York

LAS - McCarran International Airport - Las Vegas, Nevada

LAX - Los Angeles International Airport - California

LGA - LaGuardia Airport - Queens, New York

MCO - Orlando International Airport - Florida

MDW - Chicago Midway International Airport - Illinois

MIA - Miami International Airport - Florida

MSP - Minneapolis–Saint Paul International Airport - Minnesota

MSY - Louis Armstrong New Orleans International Airport - Louisiana

OAK - Oakland International Airport - California

ORD - O'Hare International Airport - Chicago, Illinois

PDX - Portland International Airport - Oregon

PHL - Philadelphia International Airport - Pennsylvania

PHX - Phoenix Sky Harbor International Airport - Arizona

RDU - Raleigh-Durham International Airport - North Carolina

SAN - San Diego International Airport - California

SEA - Seattle–Tacoma International Airport - Washington

SFO - San Francisco International Airport - California

SJC - Norman Y. Mineta San Jose International Airport - California

SLC - Salt Lake City International Airport - Utah

SMF - Sacramento International Airport - California

STL - St. Louis Lambert International Airport - Missouri

TPA - Tampa International Airport - Florida

References

Haydon, D. H., & Kumar, N. K. (2020, April 7). *Industries most and least impacted by COVID-19 from a probability of default perspective – March 2020 update*. S&P Global Market Intelligence. Retrieved from https://www.spglobal.com/marketintelligence/en/news-insights/blog/industries-most-and-least-impacted-by-covid-19-from-a-probability-of-default-perspective-march-2020-update

Josephs, L. (2022, July 1). *More than 12,000 flights delayed, hundreds canceled during busy July Fourth Weekend*. CNBC. Retrieved July 28, 2022, from https://www.cnbc.com/2022/07/01/airline-travel-of-july-weekend-puts-airlines-and-travelers-to-the-test-after-difficult-spring.html

Nguyen, T. (2022, June 30). *Flying will be the worst part of your summer vacation*. Vox. Retrieved from https://www.vox.com/the-goods/2022/6/30/23189458/summer-travel-2022-pilot-shortage

Trefis. (2016, August 31). *What is the impact of flight delays?* Trefis. Retrieved July 28, 2022, from https://www.trefis.com/stock/dal/articles/375013/what-is-the-impact-of-flight-delays/2016-08-31#:~:text=Delays%20and%20cancellations%20affect%20both,make%20the%20passengers%20distrust%20airlines

Dataset is available on-

https://www.kaggle.com/datasets/jimschacko/airlines-dataset-to-predict-a-delay