Analysis of the Factors that Affect Bus Delay in Toronto

Team Members

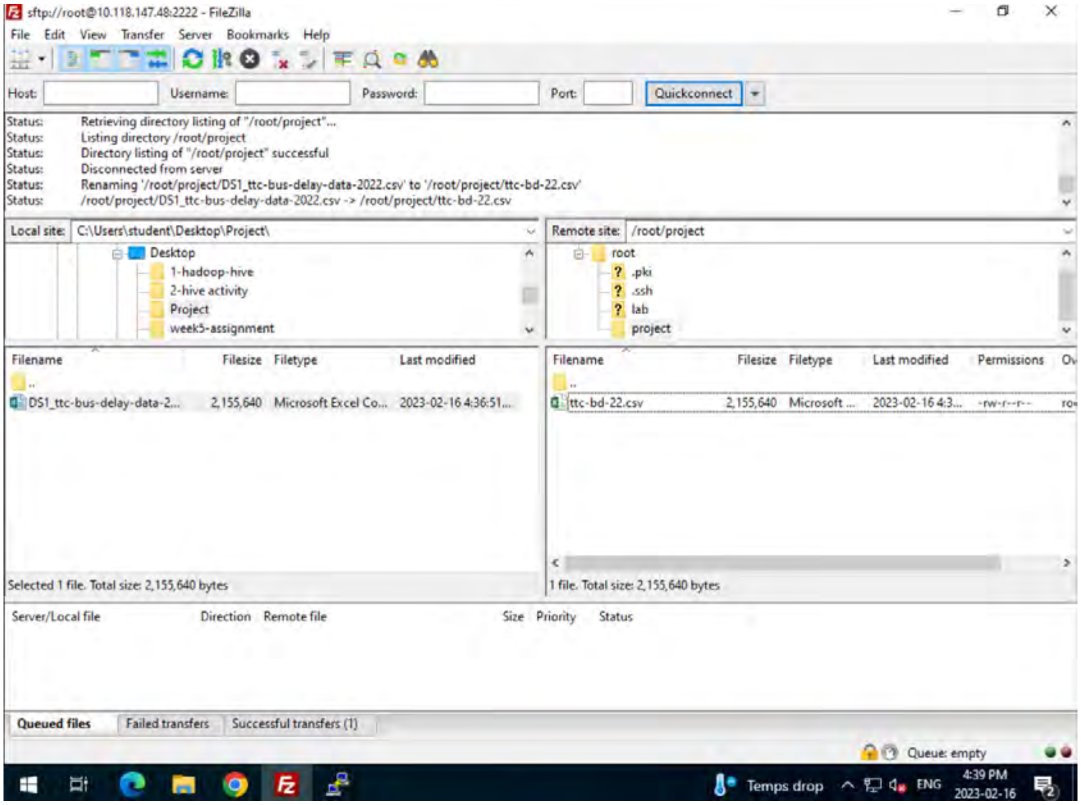
Jaydenn Chang	**N01511476**
Kaiyan Chen	**N01489178**
Mira Philip	**N01495720**
Simul Bista	**N01489966**

Problem Statement:

The project aims to analyze the factors that affect bus delay in Toronto based on the dataset that is sourced from the Toronto Transit Commission (TTC), which provides detailed records of bus delays across the city. It focuses on four key factors, namely bus route, days of the week, time, and incident type. The findings of the project to can be used to inform policy and decision-makers to improve the reliability and efficiency of the public transit system.

Steps

1. **HDFS**
    a. First, we create a directory in the sandbox called project i.e., /root/project and using FileZilla, we upload our dataset into the sandbox.



```
Status:     Retrieving directory listing of "/root/project"...
Status:     Listing directory /root/project
Status:     Directory listing of "/root/project" successful
Status:     Disconnected from server
Status:     Renaming '/root/project/DS1_ttc-bus-delay-data-2022.csv' to '/root/project/ttc-bd-22.csv'
Status:     /root/project/DS1_ttc-bus-delay-data-2022.csv -> /root/project/ttc-bd-22.csv
```

Filename	Filesize	Filetype	Last modified
..			
DS1_ttc-bus-delay-data-2...	2,155,640	Microsoft Excel Co...	2023-02-16 4:36:51...

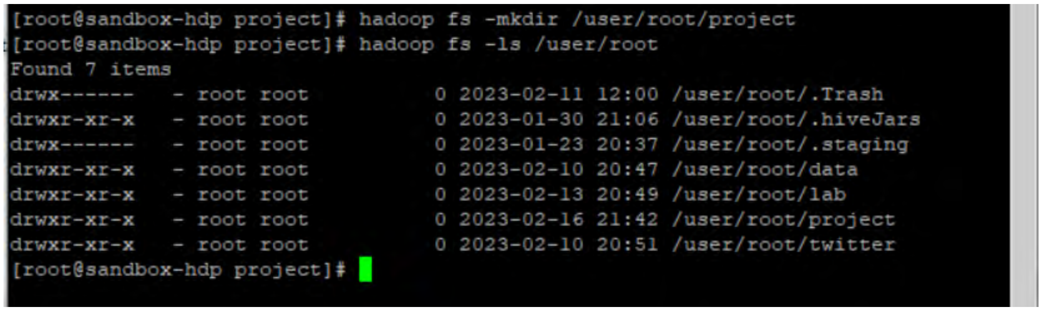
Selected 1 file. Total size: 2,155,640 bytes

Filename	Filesize	Filetype	Last modified	Permissions	Ov
..					
ttc-bd-22.csv	2,155,640	Microsoft ...	2023-02-16 4:3...	-rw-r--r--	ro

1 file. Total size: 2,155,640 bytes

Server/Local file	Direction	Remote file	Size	Priority	Status

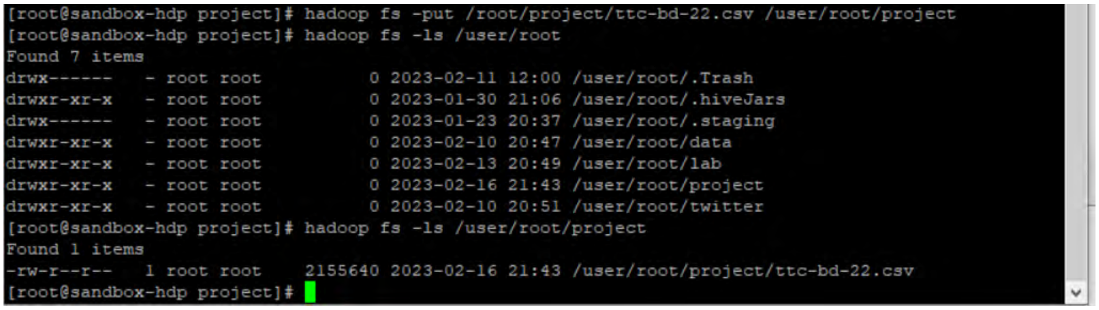
b. We create a project directory in Hadoop using the command:
   hadoop fs -mkdir /user/root/project



```
[root@sandbox-hdp project]# hadoop fs -mkdir /user/root/project
[root@sandbox-hdp project]# hadoop fs -ls /user/root
Found 7 items
drwx------   - root root          0 2023-02-11 12:00 /user/root/.Trash
drwxr-xr-x   - root root          0 2023-01-30 21:06 /user/root/.hiveJars
drwx------   - root root          0 2023-01-23 20:37 /user/root/.staging
drwxr-xr-x   - root root          0 2023-02-10 20:47 /user/root/data
drwxr-xr-x   - root root          0 2023-02-13 20:49 /user/root/lab
drwxr-xr-x   - root root          0 2023-02-16 21:42 /user/root/project
drwxr-xr-x   - root root          0 2023-02-10 20:51 /user/root/twitter
[root@sandbox-hdp project]#
```

c. We generally give read and write access permission using chmod to the folder, however, we already being root, we don't do that as of now.

d. We then load the data from the sandbox to HDFS using the command:
   hadoop fs -put /root/project/ttc-bd-22.csv /user/root/project



```
[root@sandbox-hdp project]# hadoop fs -put /root/project/ttc-bd-22.csv /user/root/project
[root@sandbox-hdp project]# hadoop fs -ls /user/root
Found 7 items
drwx------   - root root          0 2023-02-11 12:00 /user/root/.Trash
drwxr-xr-x   - root root          0 2023-01-30 21:06 /user/root/.hiveJars
drwx------   - root root          0 2023-01-23 20:37 /user/root/.staging
drwxr-xr-x   - root root          0 2023-02-10 20:47 /user/root/data
drwxr-xr-x   - root root          0 2023-02-13 20:49 /user/root/lab
drwxr-xr-x   - root root          0 2023-02-16 21:43 /user/root/project
drwxr-xr-x   - root root          0 2023-02-10 20:51 /user/root/twitter
[root@sandbox-hdp project]# hadoop fs -ls /user/root/project
Found 1 items
-rw-r--r--   1 root root    2155640 2023-02-16 21:43 /user/root/project/ttc-bd-22.csv
[root@sandbox-hdp project]#
```

2. HIVE

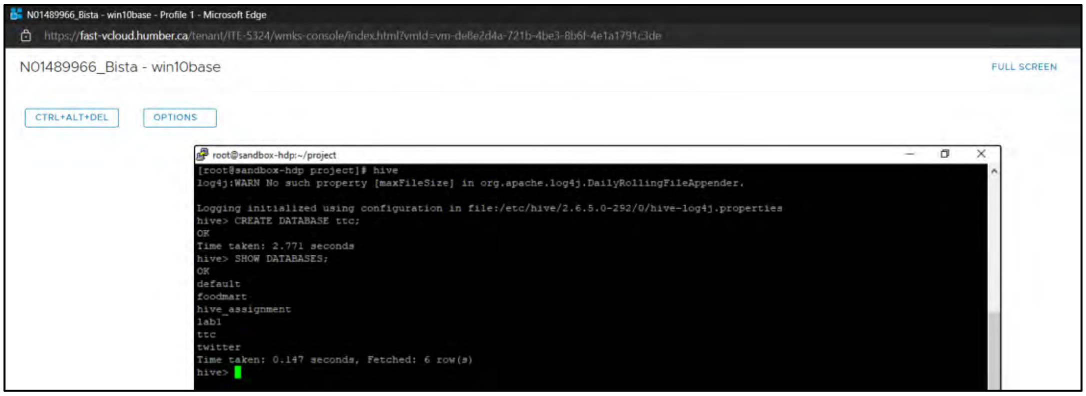
a. Access hive using the command:
   Hive

b. Now, we create a database called ttc using the following command:

   CREATE DATABASE ttc;

c. We can verify that the database has been created using:

   SHOW DATABASES;



N01489966_Bista - win10base - Profile 1 - Microsoft Edge
https://fast-vcloud.humber.ca/tenant/ITE-5324/wmks-console/index.html?vmId=vm-de8e2d4a-721b-4be3-8b6f-4e1a1791c3de
N01489966_Bista - win10base
FULL SCREEN
CTRL+ALT+DEL
OPTIONS
root@sandbox-hdp:~/project

```
[root@sandbox-hdp project]# hive
log4j:WARN No such property [maxFileSize] in org.apache.log4j.DailyRollingFileAppender.

Logging initialized using configuration in file:/etc/hive/2.6.5.0-292/0/hive-log4j.properties
hive> CREATE DATABASE ttc;
OK
Time taken: 2.771 seconds
hive> SHOW DATABASES;
OK
default
foodmart
hive_assignment
lab1
ttc
twitter
Time taken: 0.147 seconds, Fetched: 6 row(s)
hive> 
```

d. Before creating the table, we must make sure that we are inside the correct database since by default a default database is selected. We do this using:

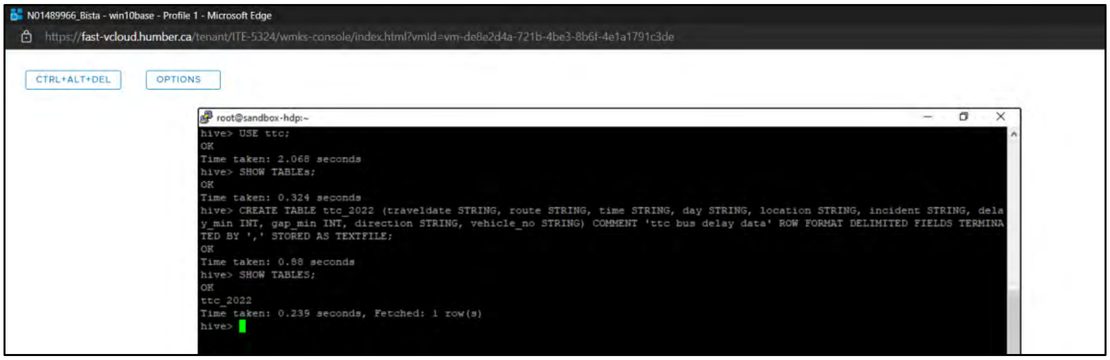
USE ttc;

e. We then create the table in which we want to load the data from our dataset which is in hdfs. The command is:

CREATE EXTERNAL TABLE ttc_bus_delay_2022 (traveldate STRING,route STRING, time TIMESTAMP, day STRING, location STRING, incident STRING, delay_min INT, gap_min INT, vehicle_no STRING);

f. We can check if the table has been created using the command:

SHOW TABLES IN ttc;



N01489966_Bista - win10base - Profile 1 - Microsoft Edge
https://fast-vcloud.humber.ca/tenant/ITE-5324/wmks-console/index.html?vmId=vm-de8e2d4a-721b-4be3-8b6f-4e1a1791c3de
CTRL+ALT+DEL
OPTIONS
root@sandbox-hdp:~

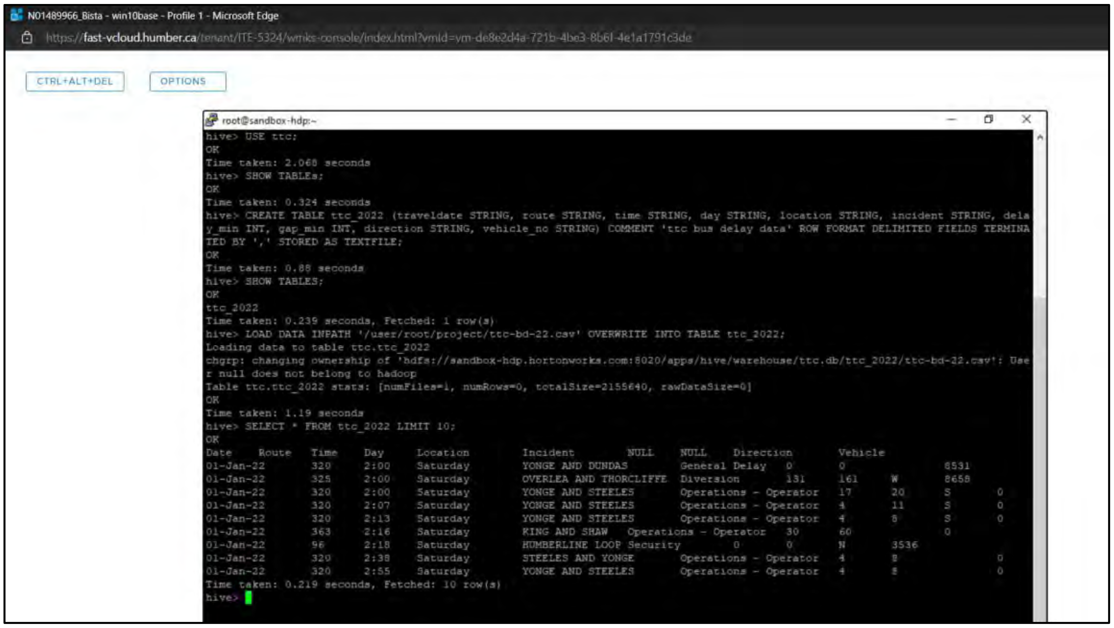
```
hive> USE ttc;
OK
Time taken: 2.068 seconds
hive> SHOW TABLEs;
OK
Time taken: 0.324 seconds
hive> CREATE TABLE ttc_2022 (traveldate STRING, route STRING, time STRING, day STRING, location STRING, incident STRING, dela
y_min INT, gap_min INT, direction STRING, vehicle_no STRING) COMMENT 'ttc bus delay data' ROW FORMAT DELIMITED FIELDS TERMINA
TED BY ',' STORED AS TEXTFILE;
OK
Time taken: 0.88 seconds
hive> SHOW TABLES;
OK
ttc_2022
Time taken: 0.239 seconds, Fetched: 1 row(s)
hive> 
```

g. Now its time to load the data into the table, we do that using:

LOAD DATA INPATH ‘/user/root/project/ttc-bd-22.csv’ OVERWRITE INTO TABLE ttc_bus_delay_2022;

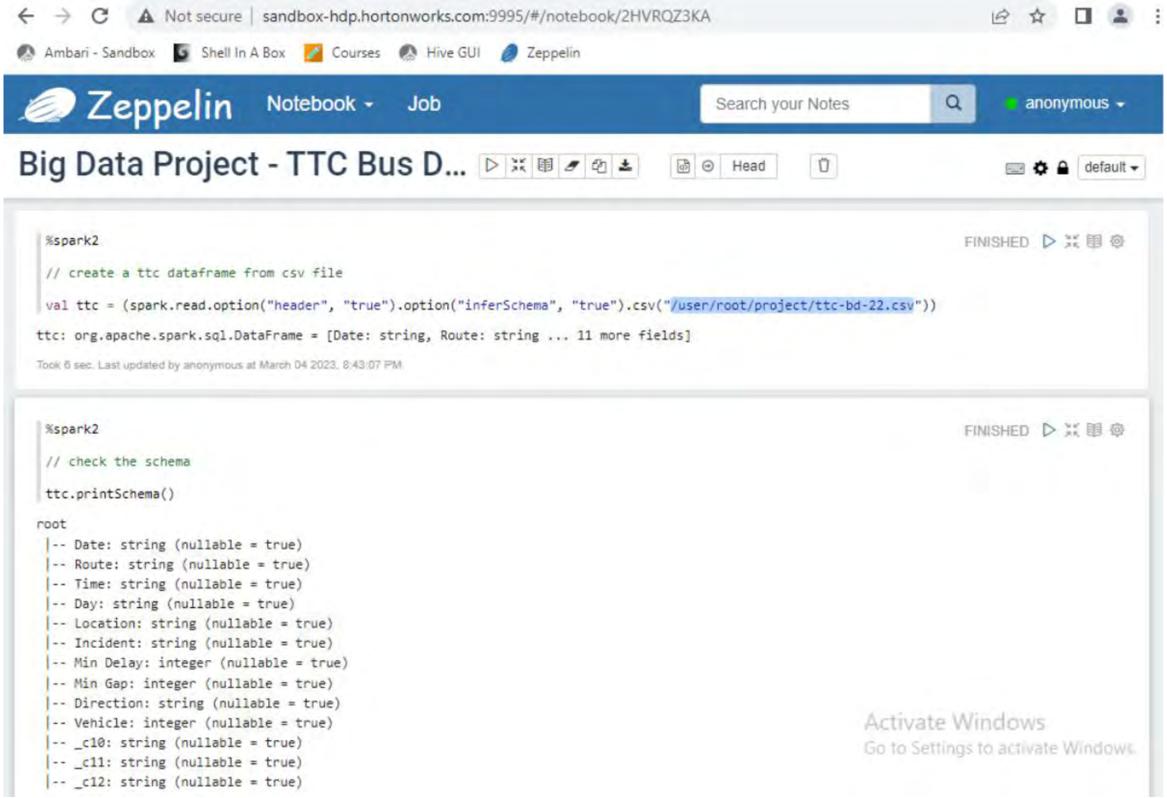
h. Use a select query to view the first couple of records that has been loaded into the table:

SELECT * FROM ttc_bus_delay_2022 LIMIT 10;


N01489966_Bista - win10base - Profile 1 - Microsoft Edge
https://fast-vcloud.humber.ca/tenant/ITE-5324/wmks-console/index.html?vmId=vm-de8e2d4a-721b-4be3-8b6f-4e1a1791c3de
CTRL+ALT+DEL
OPTIONS
root@sandbox-hdp:~

```
hive> USE ttc;
OK
Time taken: 2.068 seconds
hive> SHOW TABLEs;
OK
Time taken: 0.324 seconds
hive> CREATE TABLE ttc_2022 (traveldate STRING, route STRING, time STRING, day STRING, location STRING, incident STRING, dela
y_min INT, gap_min INT, direction STRING, vehicle_no STRING) COMMENT 'ttc bus delay data' ROW FORMAT DELIMITED FIELDS TERMINA
TED BY ',' STORED AS TEXTFILE;
OK
Time taken: 0.88 seconds
hive> SHOW TABLES;
OK
ttc_2022
Time taken: 0.239 seconds, Fetched: 1 row(s)
hive> LOAD DATA INPATH '/user/root/project/ttc-bd-22.csv' OVERWRITE INTO TABLE ttc_2022;
Loading data to table ttc.ttc_2022
chgrp: changing ownership of 'hdfs://sandbox-hdp.hortonworks.com:8020/apps/hive/warehouse/ttc.db/ttc_2022/ttc-bd-22.csv': Use
r null does not belong to hadoop
Table ttc.ttc_2022 stats: [numFiles=1, numRows=0, totalSize=2155640, rawDataSize=0]
OK
Time taken: 1.19 seconds
hive> SELECT * FROM ttc_2022 LIMIT 10;
OK
Date    Route   Time    Day     Location        Incident        NULL    NULL    Direction       Vehicle
01-Jan-22       320     2:00    Saturday        YONGE AND DUNDAS        General Delay   0       0               8531
01-Jan-22       325     2:00    Saturday        OVERLEA AND THORCLIFFE  Diversion       131     161     W       8658
01-Jan-22       320     2:00    Saturday        YONGE AND STEELES       Operations - Operator   17      20      S       0
01-Jan-22       320     2:07    Saturday        YONGE AND STEELES       Operations - Operator   4       11      S       0
01-Jan-22       320     2:13    Saturday        YONGE AND STEELES       Operations - Operator   4       8       S       0
01-Jan-22       363     2:16    Saturday        KING AND SHAW   Operations - Operator   30      60              0
01-Jan-22       96      2:18    Saturday        HUMBERLINE LOOP Security        0       0       N       3536
01-Jan-22       320     2:38    Saturday        STEELES AND YONGE       Operations - Operator   4       8               0
01-Jan-22       320     2:55    Saturday        YONGE AND STEELES       Operations - Operator   4       8               0
Time taken: 0.219 seconds, Fetched: 10 row(s)
hive> 
```

3. **Zeppelin (Spark)**
    a. We create the data frame called ttc from the given dataset (csv file)



```
%spark2
// create a ttc dataframe from csv file
val ttc = (spark.read.option("header", "true").option("inferSchema", "true").csv("/user/root/project/ttc-bd-22.csv"))
```

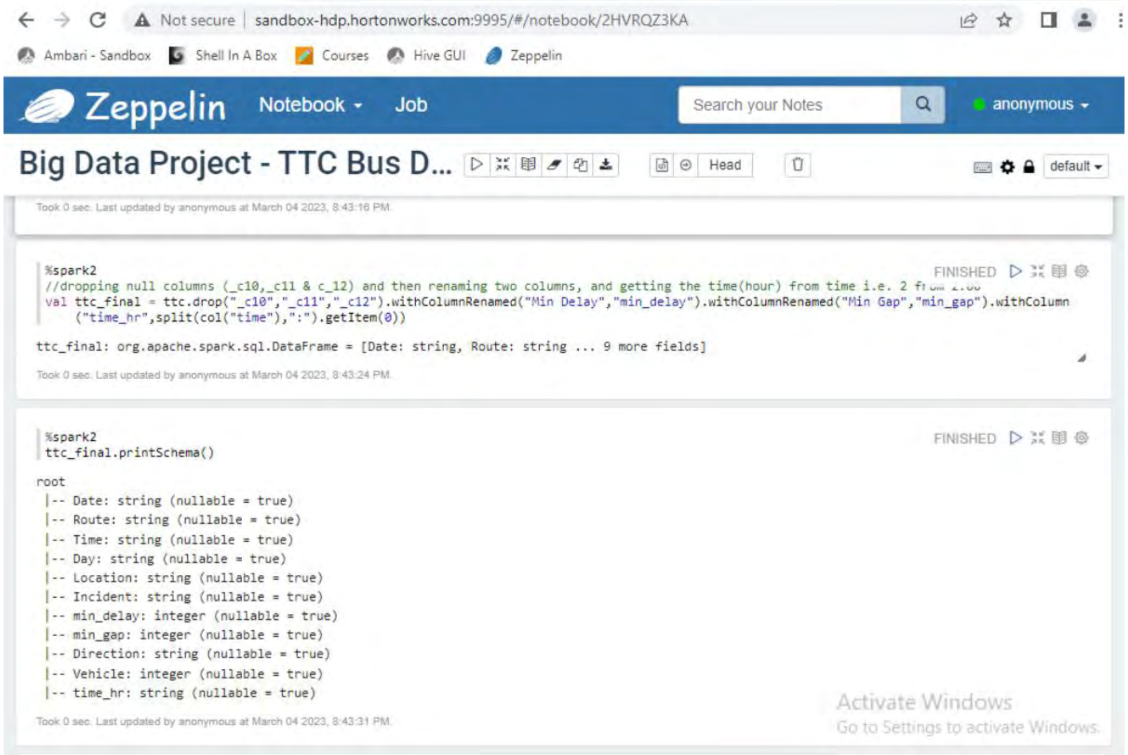
```
ttc: org.apache.spark.sql.DataFrame = [Date: string, Route: string ... 11 more fields]
```

Took 6 sec. Last updated by anonymous at March 04 2023, 8:43:07 PM

```
%spark2
// check the schema
ttc.printSchema()
```

```
root
 |-- Date: string (nullable = true)
 |-- Route: string (nullable = true)
 |-- Time: string (nullable = true)
 |-- Day: string (nullable = true)
 |-- Location: string (nullable = true)
 |-- Incident: string (nullable = true)
 |-- Min Delay: integer (nullable = true)
 |-- Min Gap: integer (nullable = true)
 |-- Direction: string (nullable = true)
 |-- Vehicle: integer (nullable = true)
 |-- _c10: string (nullable = true)
 |-- _c11: string (nullable = true)
 |-- _c12: string (nullable = true)
```

b. We then clean the data - remove null columns, perform some column renaming and then format the data in the time column to show the hours only (truncating the mins).



Took 0 sec. Last updated by anonymous at March 04 2023, 8:43:16 PM.

```
%spark2
//dropping null columns (_c10,_c11 & c_12) and then renaming two columns, and getting the time(hour) from time i.e. 2 from 2:00
val ttc_final = ttc.drop("_c10","_c11","_c12").withColumnRenamed("Min Delay","min_delay").withColumnRenamed("Min Gap","min_gap").withColumn
    ("time_hr",split(col("time"),":").getItem(0))

ttc_final: org.apache.spark.sql.DataFrame = [Date: string, Route: string ... 9 more fields]
```

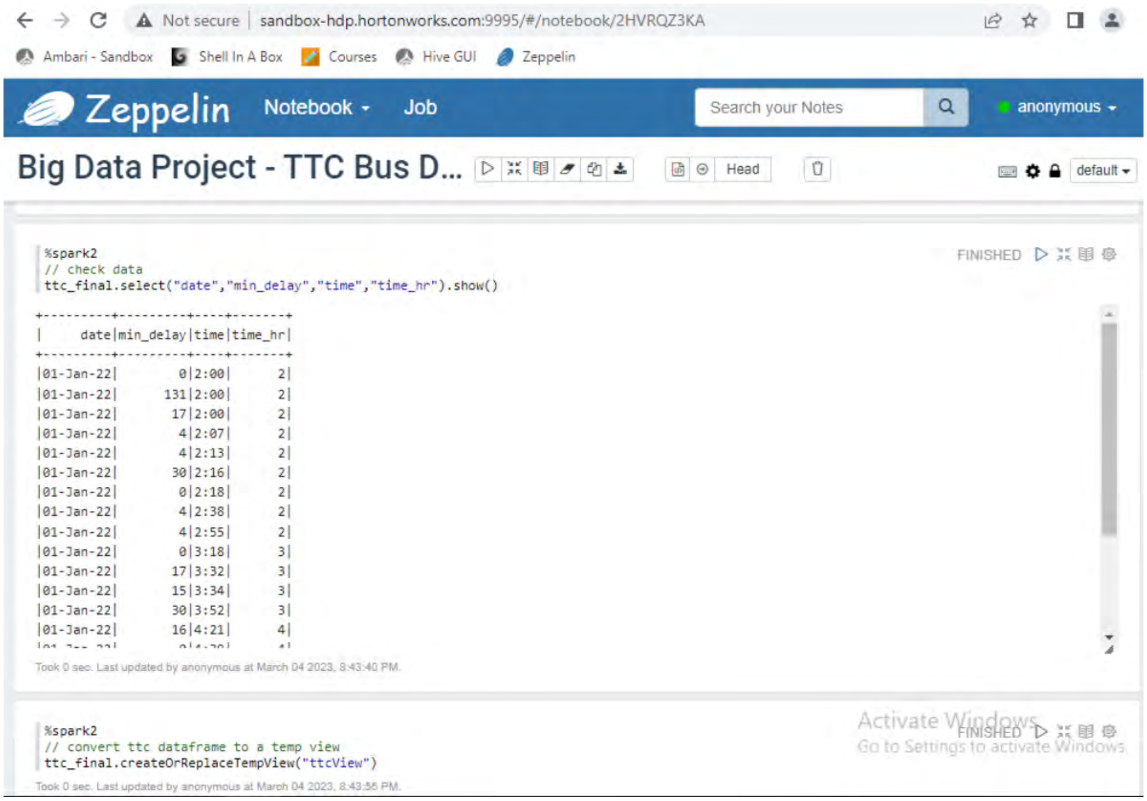
Took 0 sec. Last updated by anonymous at March 04 2023, 8:43:24 PM.

```
%spark2
ttc_final.printSchema()

root
 |-- Date: string (nullable = true)
 |-- Route: string (nullable = true)
 |-- Time: string (nullable = true)
 |-- Day: string (nullable = true)
 |-- Location: string (nullable = true)
 |-- Incident: string (nullable = true)
 |-- min_delay: integer (nullable = true)
 |-- min_gap: integer (nullable = true)
 |-- Direction: string (nullable = true)
 |-- Vehicle: integer (nullable = true)
 |-- time_hr: string (nullable = true)
```

Took 0 sec. Last updated by anonymous at March 04 2023, 8:43:31 PM.

c. Next, we check the data (a couple of columns that we just cleaned) to verify that everything is as planned. And then we convert the data frame to a temp view to start visualizing and gain some meaningful information from the representation.



```
%spark2
// check data
ttc_final.select("date","min_delay","time","time_hr").show()
```

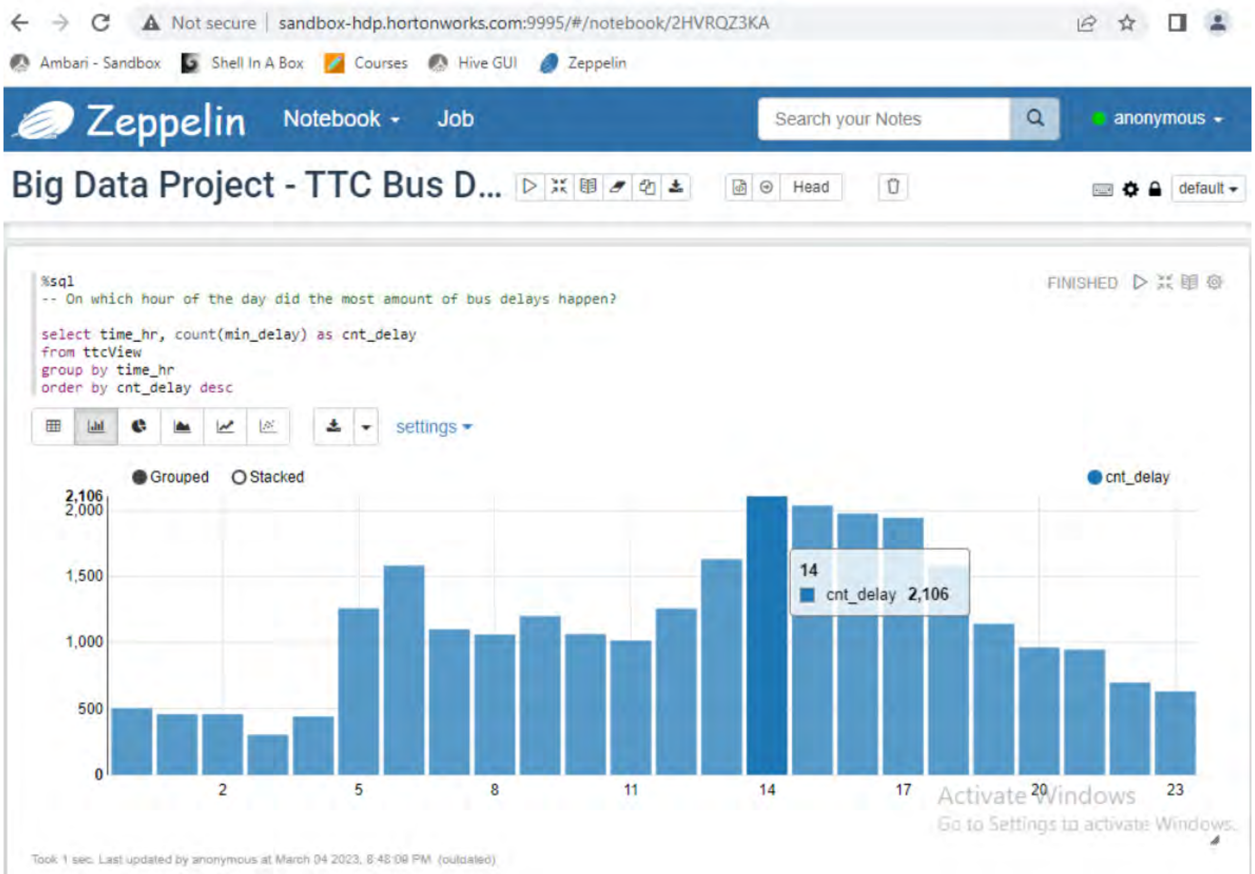
```
+---------+---------+----+-------+
|     date|min_delay|time|time_hr|
+---------+---------+----+-------+
|01-Jan-22|        0|2:00|      2|
|01-Jan-22|      131|2:00|      2|
|01-Jan-22|       17|2:00|      2|
|01-Jan-22|        4|2:07|      2|
|01-Jan-22|        4|2:13|      2|
|01-Jan-22|       30|2:16|      2|
|01-Jan-22|        0|2:18|      2|
|01-Jan-22|        4|2:38|      2|
|01-Jan-22|        4|2:55|      2|
|01-Jan-22|        0|3:18|      3|
|01-Jan-22|       17|3:32|      3|
|01-Jan-22|       15|3:34|      3|
|01-Jan-22|       30|3:52|      3|
|01-Jan-22|       16|4:21|      4|
```

Took 0 sec. Last updated by anonymous at March 04 2023, 8:43:40 PM.

```
%spark2
// convert ttc dataframe to a temp view
ttc_final.createOrReplaceTempView("ttcView")
```

Took 0 sec. Last updated by anonymous at March 04 2023, 8:43:56 PM.

d. **Visualization 1:** We figured out the time of the day in which most bus delays happened. The result shows that around 2-5pm was most likely for the bus delay to happen.



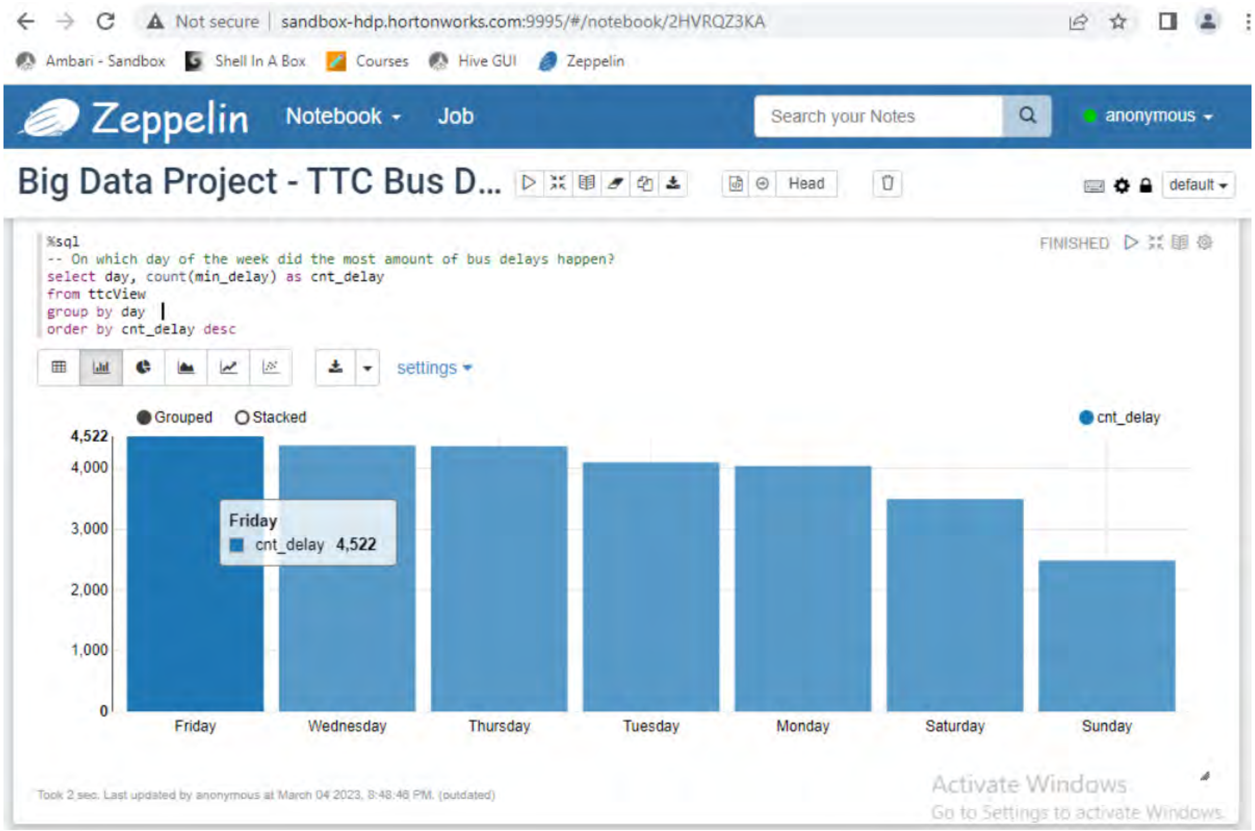
Big Data Project - TTC Bus D...

```
%sql
-- On which hour of the day did the most amount of bus delays happen?

select time_hr, count(min_delay) as cnt_delay
from ttcView
group by time_hr
order by cnt_delay desc
```

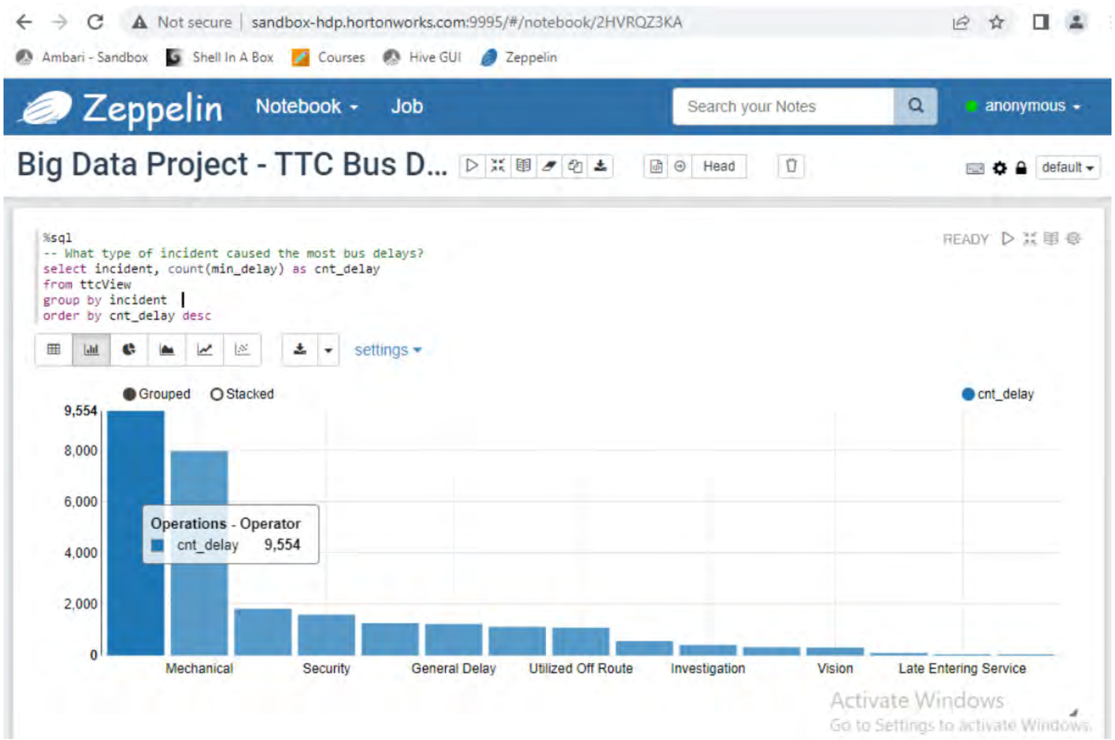
Took 1 sec. Last updated by anonymous at March 04 2023, 8:48:09 PM. (outdated)

e. **Visualization 2:** We figure out the day of the week in which most bus delays occurred. The result shows the day of the week doesn’t show significant difference.



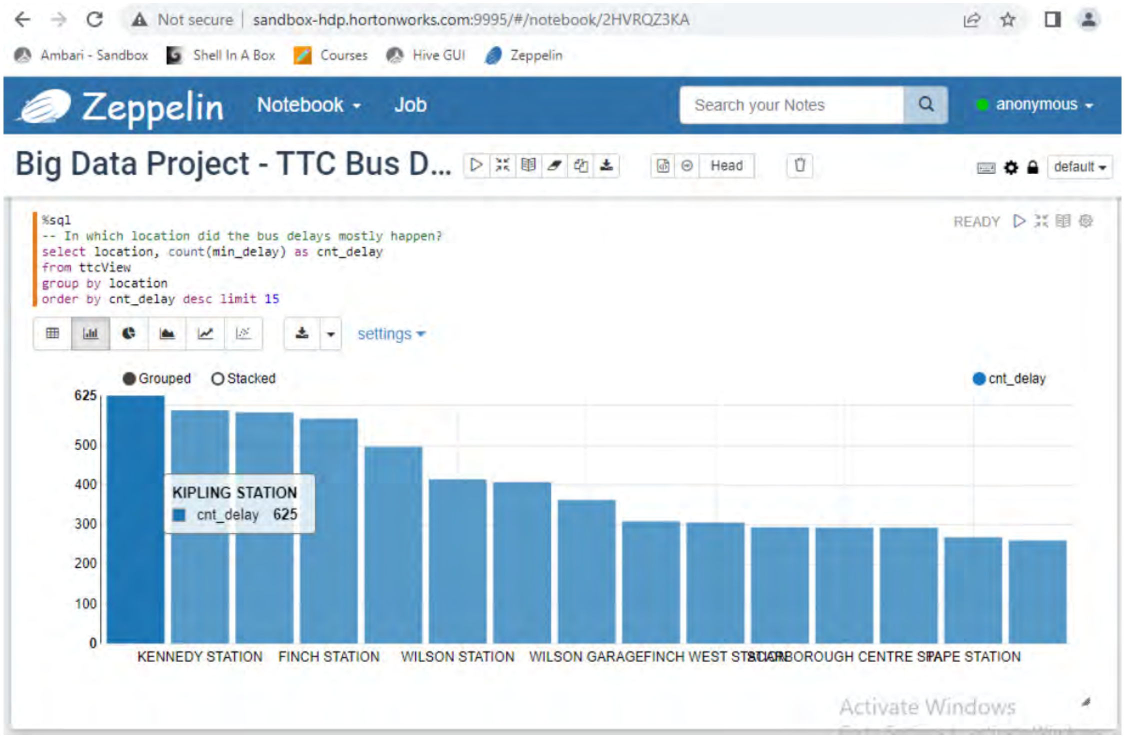
```
%sql
-- On which day of the week did the most amount of bus delays happen?
select day, count(min_delay) as cnt_delay
from ttcView
group by day
order by cnt_delay desc
```

f. **Visualization 3:** Next, we figure out the incident type which caused the most bus delays. The result shows that “Operations/Operators” and the “Mechanical” are the top reason behind bus delays.



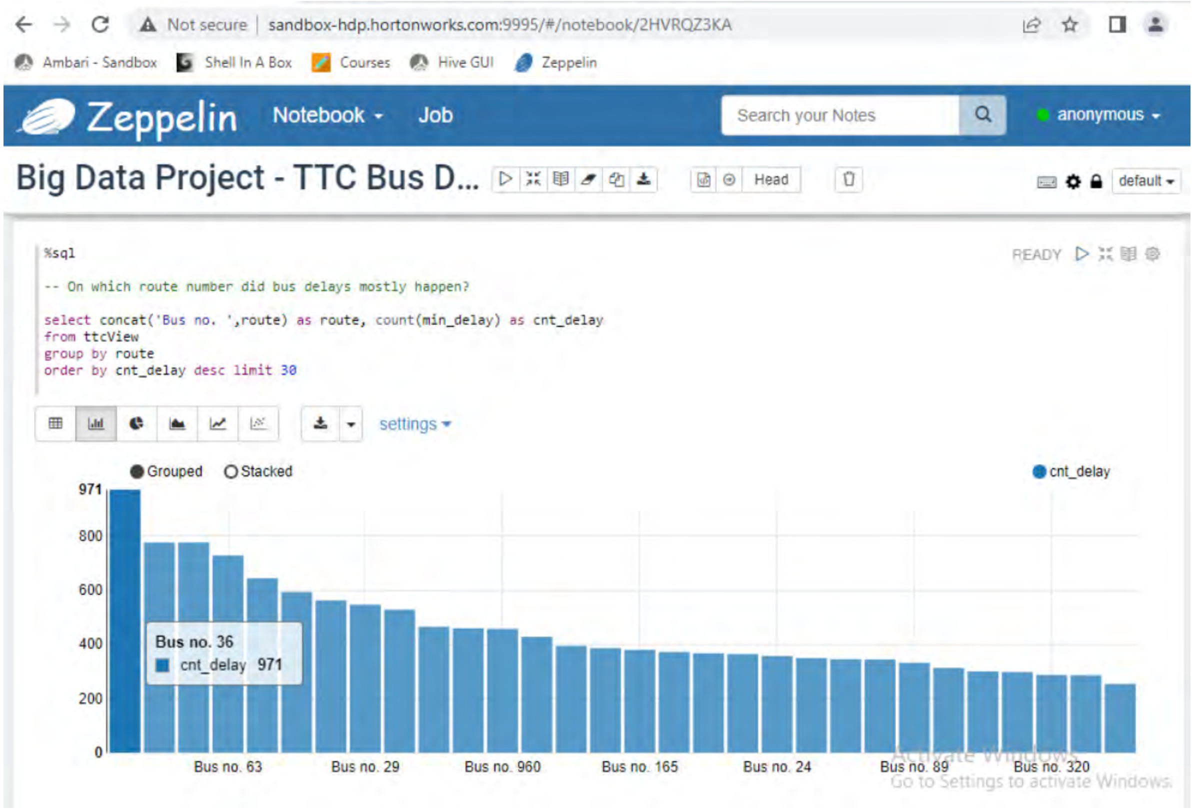
```
%sql
-- What type of incident caused the most bus delays?
select incident, count(min_delay) as cnt_delay
from ttcView
group by incident
order by cnt_delay desc
```

g. **Visualization 4:** We figured out the location(station) where the most bus delays happened. The result shows that at Kipling station and Kennedy station are two major stations where the bus delay happened. They happened both to be the end station of the subway lines.



```
%sql
-- In which location did the bus delays mostly happen?
select location, count(min_delay) as cnt_delay
from ttcView
group by location
order by cnt_delay desc limit 15
```

h. **Visualization 5:** We noticed on which routes the most bus delay happened. The result shows that its Bus Route 36.



```
%sql

-- On which route number did bus delays mostly happen?

select concat('Bus no. ',route) as route, count(min_delay) as cnt_delay
from ttcView
group by route
order by cnt_delay desc limit 30
```

Conclusion:

Bus delays mostly happened in the afternoon around 2-5pm at the end station of subway line 2 with the majority cause categories of “Operations/Operators” and “Mechanical.” We can focus on these causes and dive deeper into the each reason to put together a further conclusion and possibly solutions to the bus delays.