CIS-4560 Term Project Tutorial

Authors: Rodolfo Hoyo, Amy Wong, Jessie Estrada, Alan Kuang, Brayan Gonzalez Luna

Instructor: Jongwook Woo

Date: May 8, 2023

Lab Tutorial

rhoyo@calstatela.edu, awong7@calstatela.edu, jestra126@calstatela.edu, ckuang4@calstatela.edu, bgonza120@calstatela.edu

Objective:

- Download Data
- Unzip and Relocate Data
- Create Tables
- Clean Up Data
- Download Data to the Local Machine
- Analyze Data & Visualization

Platform Specs:

- **CPU:** Intel(R) Xeon(R) Platinum 8167M
- **CPU Speed:** 2.00GHz
- **# CPU Cores:** 8
- **Memory Size:** 4000 Gib

Downloading Data

1. Download the tripdata zip and zones file from Dropbox

```
wget https://www.dropbox.com/s/4yrui8y6gumd6s1/fhvh_tripdata_2021_CSV.zip
wget https://www.dropbox.com/s/q6rvxp3a4t8apfm/taxi_zone_lookup.csv
```

2. Create a directory tripdata to put the file to HDFS

```
hdfs dfs -mkdir tripdata
hdfs dfs -mkdir taxi_zone
hdfs dfs -mkdir tip_data_export
```

Unzip and Relocate Data

3. Unzip the tripdata file

```
 unzip fhvh_tripdata_2021_CSV.zip -d tripdata
```

4. Next, you can run the following shell command to put the files we unzipped into HDFS DFS

```
hdfs dfs -put tripdata/fhvh_tripdata_2021_CSV/* tripdata
hdfs dfs -put /home/<YOUR_USERNAME>/taxi_zone_lookup.csv taxi_zone
```

Create Tables

5. This code will create the table “tripdata” and load the data from the zip files into the table “tripdata.”

```
DROP TABLE IF EXISTS tripdata;
CREATE EXTERNAL TABLE if not exists tripdata (
count BIGINT,
Hvfhs_license_num VARCHAR(10),
Dispatching_base_num VARCHAR(10),
```

```
originating_base_num VARCHAR(10),
request_datetime TIMESTAMP,
on_scene_datetime TIMESTAMP,
Pickup_datetime TIMESTAMP,
DropOff_datetime TIMESTAMP,
PULocationID INT,
DOLocationID INT,
trip_miles FLOAT,
trip_time INT,
base_passenger_fare FLOAT,
tolls FLOAT,
bcf FLOAT,
sales_tax FLOAT,
congestion_surcharge FLOAT,
airport_fee FLOAT,
tips FLOAT,
driver_pay FLOAT,
shared_request_flag CHAR(4),
shared_match_flag CHAR(4),
access_a_ride_flag CHAR(4),
wav_request_flag CHAR(4),
wav_match_flag CHAR(4)
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ","
STORED AS TEXTFILE
LOCATION '/user/<YOUR_USERNAME>/tripdata'
tblproperties ("skip.header.line.count"="1");
```

6. Create a table for taxi zones.

```
DROP TABLE IF EXISTS taxi_zones;
CREATE EXTERNAL TABLE if not exists taxi_zones (
locationId INT,
borough STRING,
zone STRING,
service_zone STRING
 )
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ","
STORED AS TEXTFILE
LOCATION '/user/<YOUR_USERNAME>/taxi_zone'
tblproperties ("skip.header.line.count"="1");
```

7. Run the SQL command below to test the tripdata table

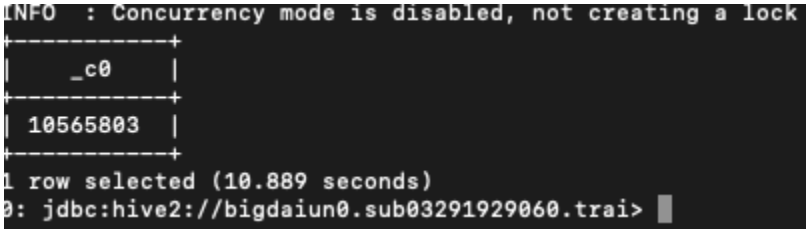
```
SELECT hvfhs_license_num, dispatching_base_num,  request_datetime,
trip_miles, base_passenger_fare FROM tripdata LIMIT 10;
```

8. Run the SQL command below to test the data in the taxi_zone table

```
SELECT * FROM taxi_zones LIMIT 15;
```

9. Run the SQL command below to count the total amount of records.
    a. 10,565,803 Records

```
SELECT COUNT(*) FROM tripdata;
```



Clean Up Data

10. Run the SQL below on Hive to create a join table between “tripdata” and “taxi_zones.” This will help us when visualization of the data. At the same time, we are replacing any HV-code with their assigned Rideshare Company.

```
DROP TABLE IF EXISTS trips_join_table;
CREATE TABLE IF NOT EXISTS trips_join_table row format delimited fields
terminated BY "," stored AS textfile location
"/user/<YOUR_USERNAME>/tmp/taxi_zones" AS
SELECT
       CASE
              WHEN hvfhs_license_num = "HV0002" THEN
regexp_replace(hvfhs_license_num, 'HV0002', 'Juno')
              WHEN hvfhs_license_num = "HV0003" THEN
regexp_replace(hvfhs_license_num, 'HV0003', 'Uber')
              WHEN hvfhs_license_num = "HV0004" THEN
regexp_replace(hvfhs_license_num, 'HV0004', 'Via')
```

```
            WHEN hvfhs_license_num = "HV0005" THEN
regexp_replace(hvfhs_license_num, 'HV0005', 'Lyft')
       END AS ridesharing_company,
       dispatching_base_num,
       request_datetime,
       pickup_datetime,
       dropoff_datetime,
       pulocationid,
       zone AS pu_zone,
       dolocationid,
       trip_miles,
       trip_time,
       base_passenger_fare,
       tolls,
       bcf,
       sales_tax,
       congestion_surcharge,
       airport_fee,
       tips,
       driver_pay
FROM   taxi_zones t
JOIN   tripdata s
ON     s.pulocationid = t.locationid;
```

11. Run this command to verify both tables are joined properly.

```
SELECT ridesharing_company, request_datetime, pickup_datetime,
dropoff_datetime, pu_zone FROM trips_join_table limit 10;
```

```
+---------------------+---------------------+---------------------+---------------------+--------------------+
| ridesharing_company |   request_datetime  |   pickup_datetime   |   dropoff_datetime  |       pu_zone      |
+---------------------+---------------------+---------------------+---------------------+--------------------+
| Uber                | 2021-09-02 14:47:08.0 | 2021-09-02 15:01:26.0 | 2021-09-02 15:52:19.0 | "Roosevelt Island" |
| Uber                | 2021-09-02 15:36:53.0 | 2021-09-02 15:41:06.0 | 2021-09-02 16:33:32.0 | "Roosevelt Island" |
| Lyft                | 2021-09-02 15:27:27.0 | 2021-09-02 15:36:04.0 | 2021-09-02 16:25:03.0 | "Roosevelt Island" |
| Lyft                | 2021-09-02 15:14:33.0 | 2021-09-02 15:22:19.0 | 2021-09-02 15:59:46.0 | "Roosevelt Island" |
| Lyft                | 2021-09-02 15:02:46.0 | 2021-09-02 15:05:45.0 | 2021-09-02 16:15:03.0 | "Roosevelt Island" |
| Uber                | 2021-09-02 15:40:00.0 | 2021-09-02 15:33:25.0 | 2021-09-02 16:25:43.0 | "Roosevelt Island" |
| Lyft                | 2021-09-02 14:02:10.0 | 2021-09-02 14:18:55.0 | 2021-09-02 14:42:43.0 | "Roosevelt Island" |
| Uber                | 2021-09-02 14:10:00.0 | 2021-09-02 14:04:01.0 | 2021-09-02 14:36:28.0 | "Roosevelt Island" |
| Lyft                | 2021-09-02 14:08:19.0 | 2021-09-02 14:28:38.0 | 2021-09-02 14:49:39.0 | "Roosevelt Island" |
| Uber                | 2021-09-02 14:11:15.0 | 2021-09-02 14:28:47.0 | 2021-09-02 14:38:31.0 | "Roosevelt Island" |
+---------------------+---------------------+---------------------+---------------------+--------------------+
```

Download Data to the Local Machine

12. Create a CSV file from the join table.

```
INSERT OVERWRITE DIRECTORY '/user/<your_username/trip_data_export/'
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
select * from trips_join_table;
```

13. Export CSV file from Hive to Local Machine

```
hdfs dfs -get /user/<your_username>/trip_data_export/0000*
/home/<your_username>/tripdata
```

14. Compile all CSV files into one CSV file

```
cat /home/<your_username>/tripdata/0000* > tripdata.txt
```

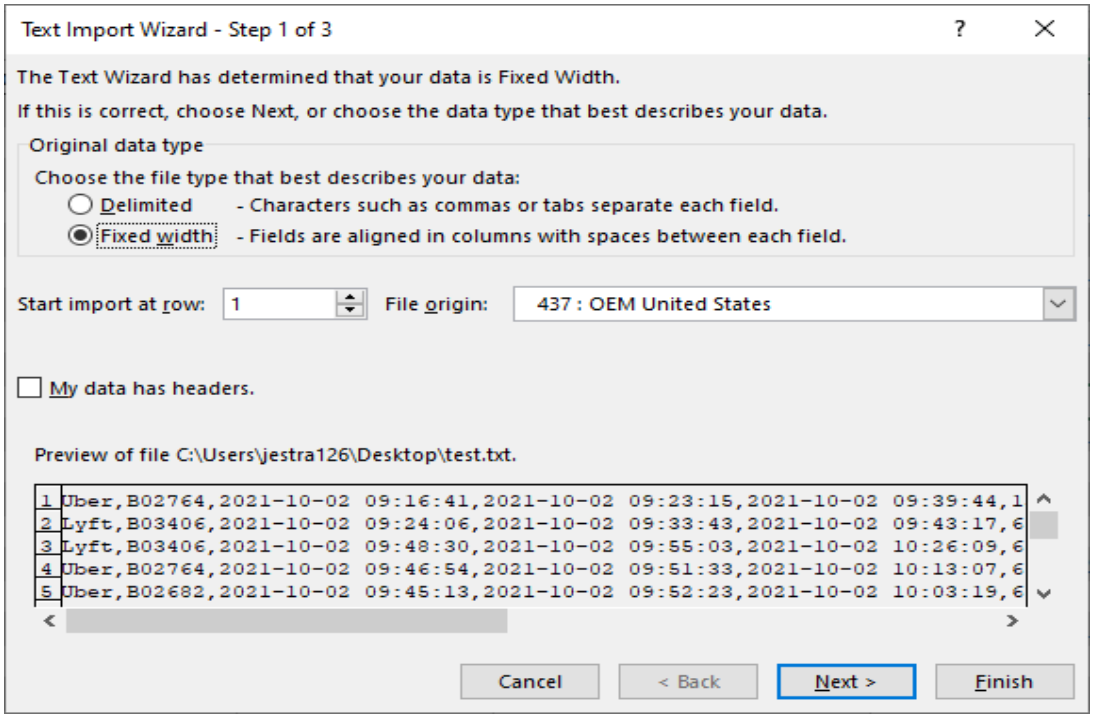
15. Download the file from the server to the Desktop of your home computer

```
scp <your_username>@129.153.114.72:/home/<your_username>/tripdata.txt
~/Desktop/tripdata.txt
```

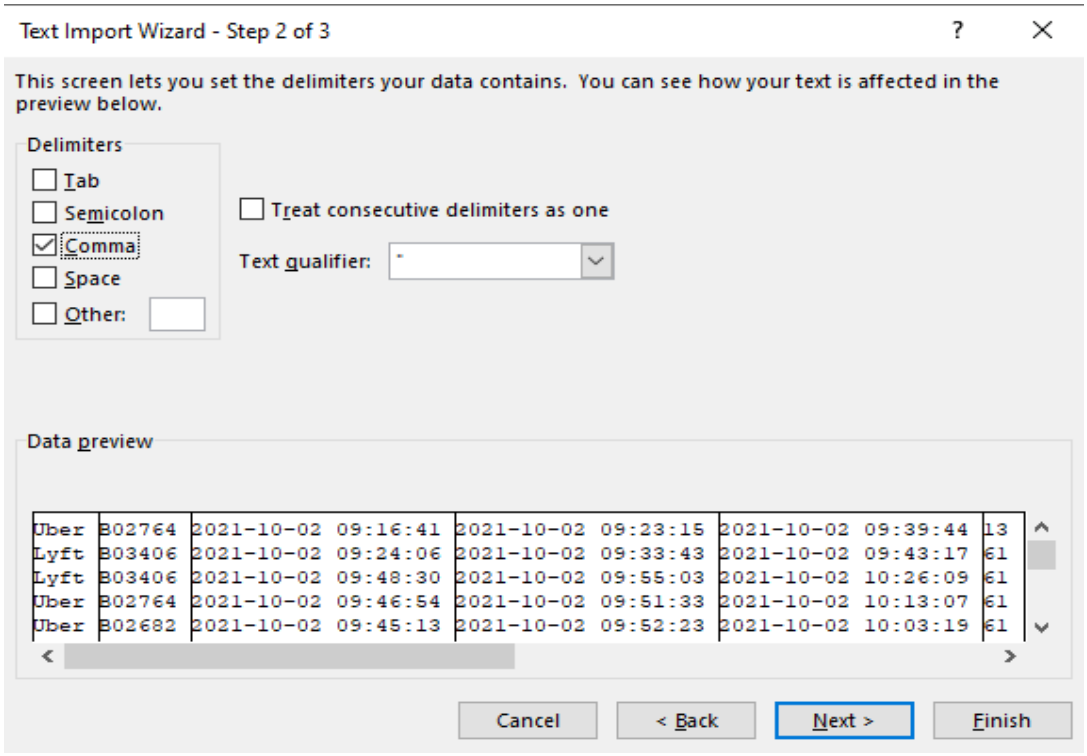
Analyze Data & Visualization

In this step, we will upload our .txt file onto Excel to create our desired 3D Map. To correctly upload data to Excel, we will:

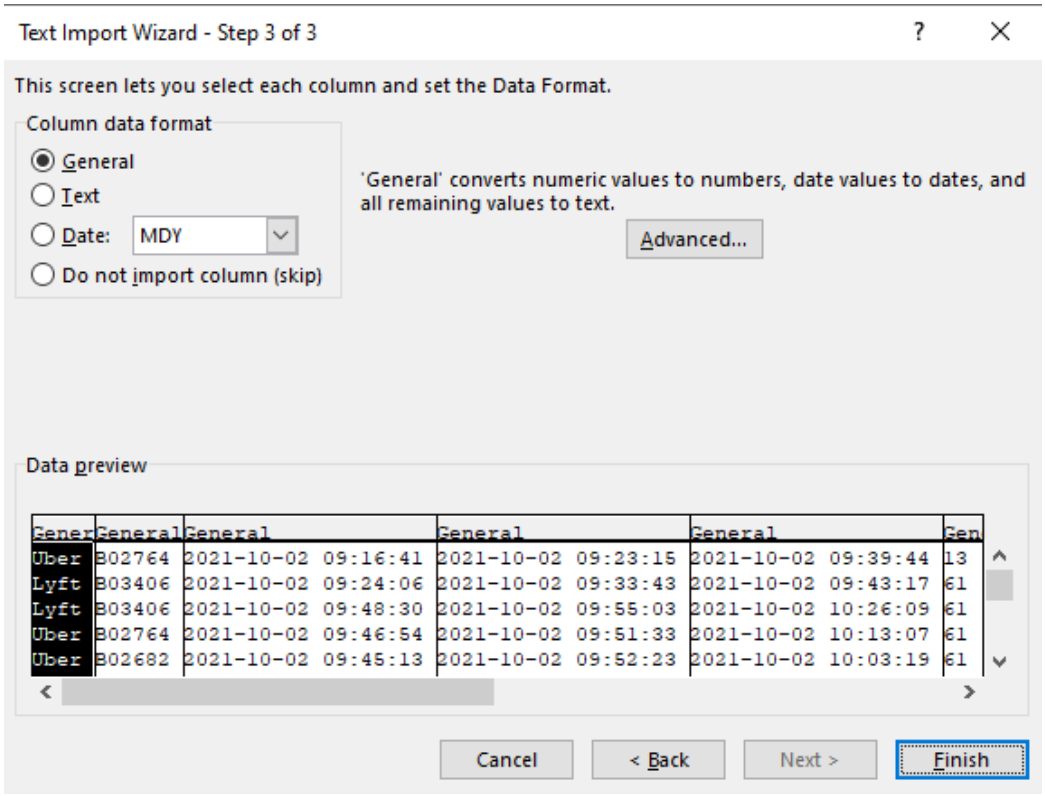
1. First, open the Excel application and then the file utilizing the Text Import Wizard. We will be choosing the Delimited description for our data.



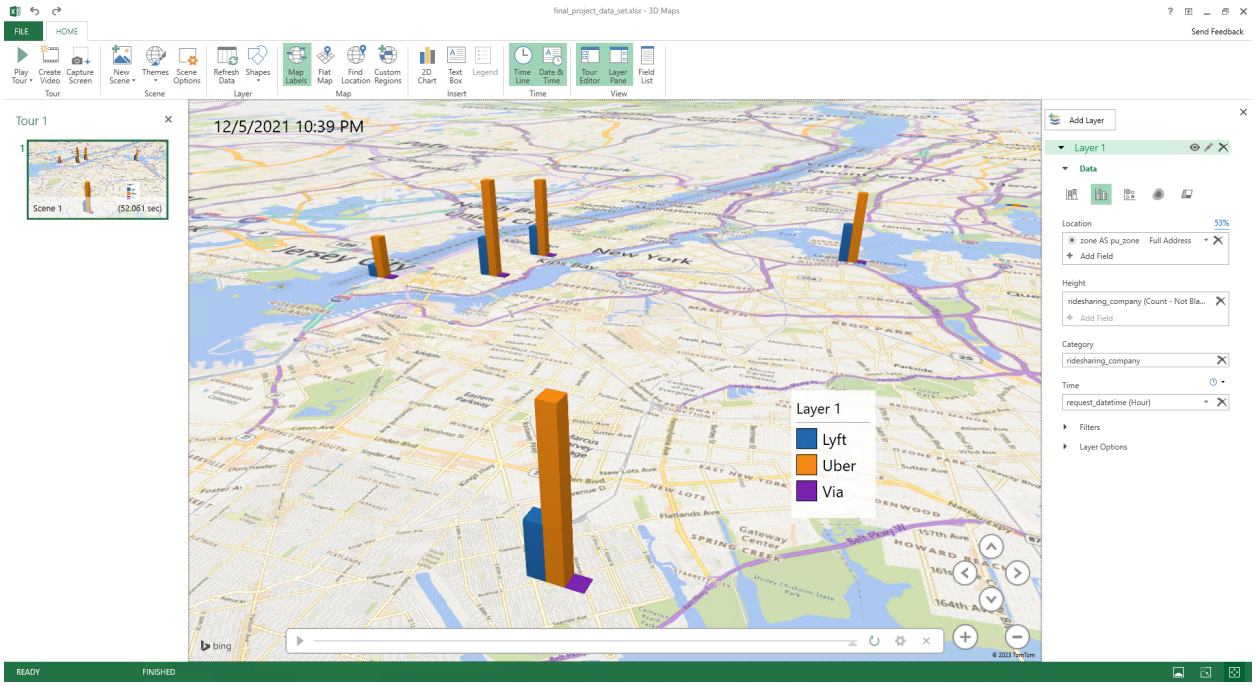
2. Select **Comma** as the Delimiter.



3. Column Data Format will be **General**.



4. To visualize location go to **Insert** tab and then click on the 3D Map button.



References

1. Data Source URL:
    a. https://www.kaggle.com/datasets/shuhengmo/uber-nyc-forhire-vehicles-trip-data-2021
2. RawData in CSV format
    a. https://www.dropbox.com/s/4yrui8y6gumd6s1/fhvh_tripdata_2021_CSV.zip
3. GitHub URL:
    a. https://github.com/JessieEstrada/NYC_Uber_Trip_Data_2021_Project