DataEng: Project Assignment 2

DataEng Project Assignment 2 Submission Document

We could not completely get through the project. We are attaching the work that we have completed so far. If given an opportunity we would like to make another submission with complete work.

Construct a table showing each day for which your pipeline successfully, automatically processed one complete day's worth of sensor readings. The table should look like this:

Date	Day of Week	# Sensor Readings	# updates/insertions into your database

Documentation of Each of the Original Data Fields

For each of the fields of the bread crumb data, provide any documentation or information that you can determine about it. Include bounds or distribution data where appropriate. For example, for something like "Vehicle ID", say something more than "It is the identification number for the vehicle". Instead, add useful information such as "the integers in this field range from <min val>

to <max val>, and there are <n> distinct vehicles identified in the data. Every vehicle is used on weekdays but only 50% of the vehicles are active on weekends.”

EVENT_NO_TRIP
EVENT_NO_STOP
OPD_DATE
VEHICLE_ID
METERS
ACT_TIME
VELOCITY
DIRECTION
RADIO_QUALITY
GPS_LONGITUDE
GPS_LATITUDE:
GPS_SATELLITES
GPS_HDOP
SCHEDULE_DEVIATION

Data Validation Assertions

List 20 or more data validation assertion statements here. These should be English language sentences similar to “The VELOCITY field
exceeds 5000000”. You will only implement a subset of them, so feel free to write assertions that might be difficult to evaluate. Create assertions for all of the fields, even those (like RADIO_QUALITY) that might not be used in your database schema.

- Every record should have a Event-No-trip
- Every record should have an Event-No-stop.
- The Act_time shouldn’t be empty
- Trip id is unique and not null
- Every record should have a latitude value other than 0
- Every record should have a longitude value other than 0
- Every record should have a time value of when the bread crumb was taken
- Every record should have a date value
- Velocity should be greater than 0 and less than 100
- The Act_time should be greater than 0 and less than 86400
- Dates should be greater than 2020
- Longitude values should be between -180 and 180
- Latitude values should be between -90 and 99
- Every record should have a velocity
- Every record should have a longitude
- Every record should have a latitude
- Every record should have a Direction

- Velocity is distance in metres divided by act_time in seconds(difference from the previous entry)
- The total distance travelled by a vehicle with a unique vehicle id is the distance recorded as meters for the last entry for the day.
- Service key is either weekday, saturday or a sunday to indicate the day of travel
- Every vehicle should have a unique id
- Every Trip must only have one vehicle associated with it.

Data Transformations

Describe any transformations that you implemented either to react to validation violations or to shape your data to fit the schema. For each, give a brief description of the transformation along with a reason for the transformation.
Timestamp column: This column creation required transformation of OPD_DATE to date format and ACT_TIME to time format. Then the date and time was combined to create a timestamp column.

Speed: Had to change the velocity column from string and float and fill the empty values with 0. Then multiplied the column with 2.237 to convert it to miles/hour.

Creates breadcrumb data frame for Breadcrumb table.

Example Queries

Provide your responses to the questions listed in Section E above. For each question, provide the SQL you used to answer the questions along with the count of the number of rows returned (where applicable) and a listing of the first 5 rows returned (where applicable).

https://github.com/sayeghmutaz-001/Data-Engineering-Prog2

Your Code

Provide a reference to the repository where you store your python code. If you are keeping it private then share it with Bruce (bruce.irvin@gmail.com), David and Aman (github references TBD).

This code needs to be integrated with consumer.py

https://github.com/Raghu-Srungavarapu/dataengineering/tree/main/Project%20Assignment%202