**Task:**

Build a data solution on AWS, **using free tier resources only,** to ingest data from NYC “For Hire Vehicles (FHV) – Active” dataset

* Dataset: <https://data.cityofnewyork.us/Transportation/For-Hire-Vehicles-FHV-Active/8wbx-tsch>
* API: <https://dev.socrata.com/foundry/data.cityofnewyork.us/8wbx-tsch>

**Notes or observations while learning about the dataset:**

* ‘Last Update Date’ column gets updated for all rows every day when data gets refreshed. Delta can’t be captured as all rows have same ‘Last Update Date’.
* Data gets updated everyday evening (4 to 7pm EST).
* Last update date and time are 2 different columns.
* ‘Vehicle License Number’ seems unique and can be served as primary key.
* Most of the columns with data type ‘date time’ have just date or time.
* Name field has commas in it.
* Seems like API doesn’t need authentication.
* <https://data.cityofnewyork.us/api/views/8wbx-tsch/rows.csv?accessType=DOWNLOAD> – mentioned in documentation is working well when I used in browser.
* Full load file is of around 20 MB of data.

**Thought process for the solution:**

* Limited time and resources – 1 to 2 hours for the task and AWS free tier
* IDE (PyCharm community edition) – development can be done quick here.
* Download the file to working directory using the URL ‘<https://data.cityofnewyork.us/api/views/8wbx-tsch/rows.csv?accessType=DOWNLOAD>’ . Python in-built library ‘urllib’ should serve the purpose.
* Create a s3 bucket for the task.
* I think I must go with ‘boto3’ package to upload file to s3 bucket.
* Redshift / Athena may be a good option as we are going to load the files in s3 bucket.
* Take the columns and data types from API documentation and create the table in DB.
* Redshift copy command should be perfect to load the file from s3. ‘psycopg2’ package may be needed in python script to load the file into Redshift table.
* Add a column for row hash (excluding the last updated date and time). This can be useful to capture incremental data.

**Solution:**

* Screen shot of solution using python in IDE (I will share socrata\_task\_for\_bezos.py file)A screenshot of a computer program

  Description automatically generated

A screenshot of a computer program

Description automatically generated

* Screenshot of files loaded to s3A screenshot of a computer

  Description automatically generated
* Screenshot of table in Redshift using query editorA screenshot of a computer

  Description automatically generated

**Queries:**

* Median age of vehicles

**A screen shot of a computer

Description automatically generated**

* Stats by region A screenshot of a computer

  Description automatically generated

**Improvements if time permits:**

* Incremental logic can be implemented by joining the current file with yesterday's file on vehicle\_license\_number and compare the hash generated on entire row except last\_updated\_date and last\_updated\_time.
* Error handling can be added.
* Audit log table can be added to capture the row count for each run (inserts, updates, and deletes).