Adrien Protzel
2/19/2025

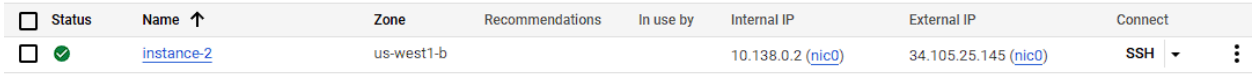
Group Homework: BigQuery - Bigish Data

Steps Followed:

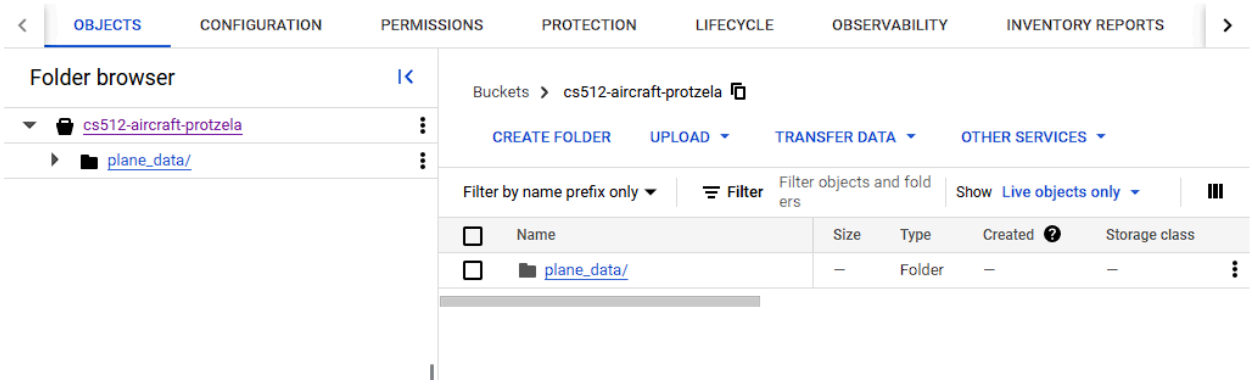
1. Set up compute engine - for cloud vm instead of local
    a. Create instance
    b. Region - west Oregon
    c. Set memory to 100GB
2. Connect to VM
    a. Click SSH
    b. Make directory: mkdir plane_data
    c. cd plane_data
    d. sudo apt-get install wget
    e. sudo apt-get install unzip
    f. wget https://web.engr.oregonstate.edu/~wolfordj/plane_data.zip
    g. unzip <tab>
3. Upload files to cloud bucket
    a. Cloud storage > create bucket
        i. cs512_aircraft
        ii. <Change nothing>
    b. <in SSH window>
    c. gcloud init
    d. Create new account: 2
        i. Copy link
        ii. Copy key code
        iii. Create project
        iv. Move zip up one directory: mv plane_data.zip ../
        v. cd ..
        vi. gsutil -m cp -r plane_data/ gs://cs512-aircraft-protzela
4. Load data on dataprep
    a. Open dataprep
    b. Import data
        i. Google cloud
        ii. Select plane_data folder
        iii. Add description
            1. If import button does not show, click continue
            2. Remove structure of imported data folder
            3. Use in new flow
            4. Edit recipe to break on '}, '
            5. Add step to add suffix } to column 1
        iv. import
    c. Add recipe steps, 'filter contains' out data

5. <make BigQuery Database>
    a. +ADD
    b. Google Cloud Storage
    c. URI: wolford-cs512-aircraft-data/BQ_Table.csv
    d. Project: cs512-aircraft-protzela
    e. Dataset: aircraft_data
    f. Table: plane_data
    g. Auto detect schema
    h. <Create table>
    i. Run fixing query: ALTER TABLE aircraft_data.plane_data RENAME COLUMN Long1 TO Long;
6. Run query to find answers on data set:
    a. SELECT count(distinct Icao) FROM
       `cs512-aircraft-protzela.aircraft_data.plane_data`
       WHERE (Lat between (44.497222 - 0.2) AND (44.497222 + 0.2))
       AND (Long between (-123.289444 - 0.2) AND (-123.289444 + 0.2))

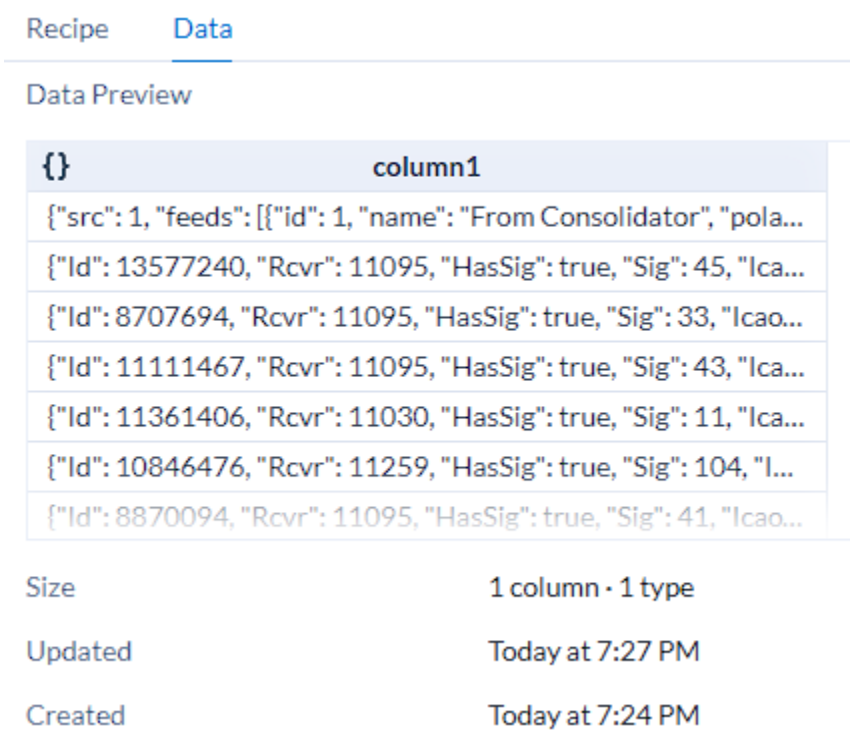
Snippets of each rubric step is listed below in order:



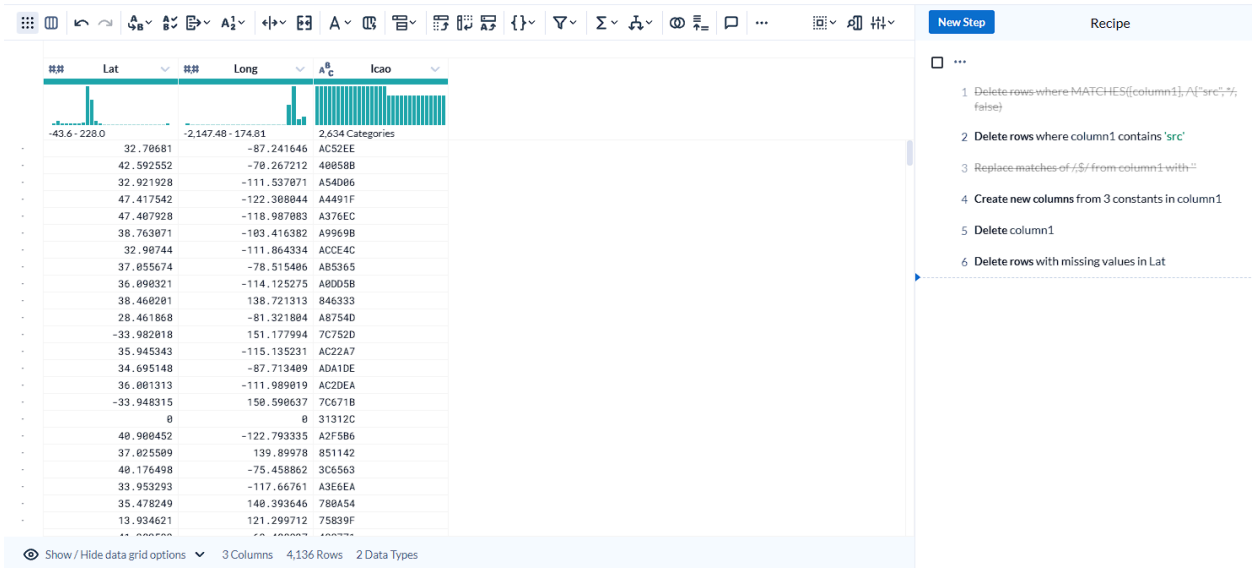
Get Zip file onto Compute Engine Instance



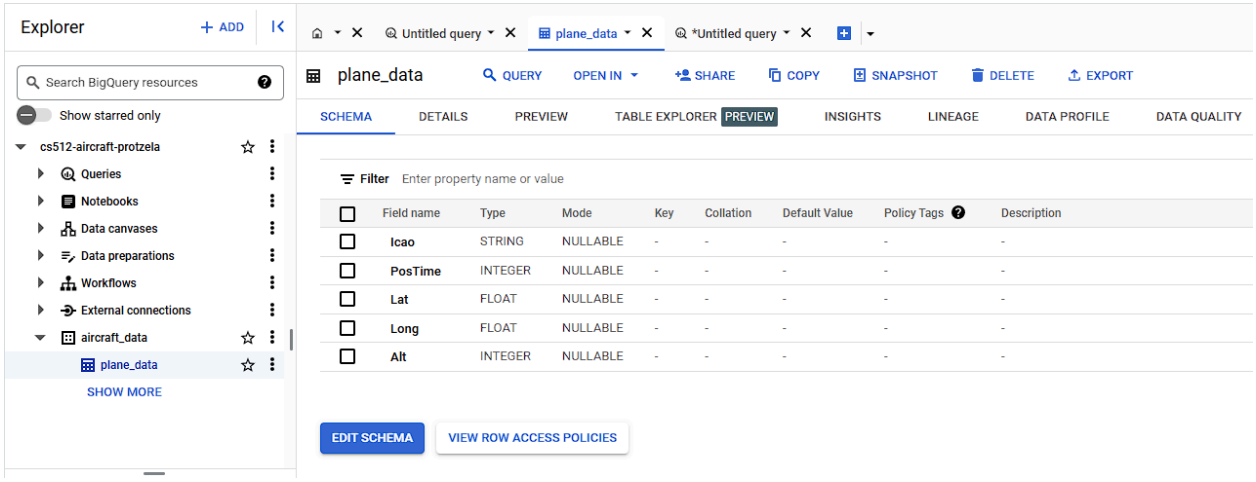
Load JSON files into Google Cloud Storage



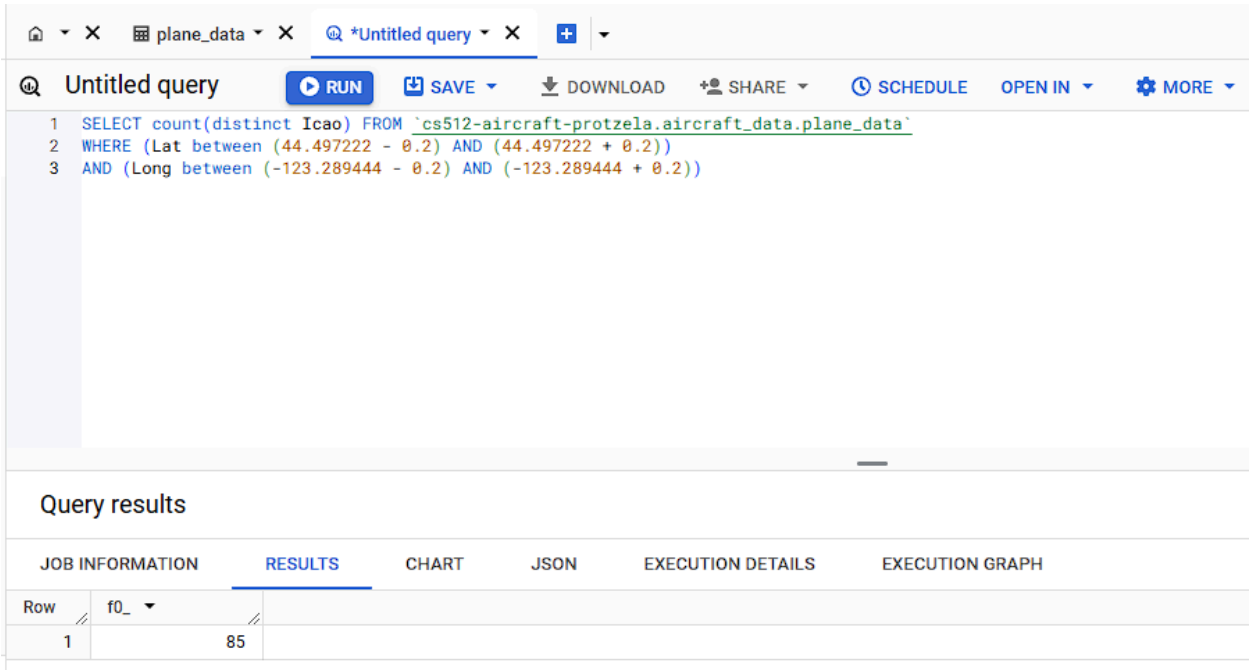
Load JSON files as a data set into Google Cloud Dataprep



Parse JSON into appropriate columns in Dataprep



Export Dataprep job into BigQuery



BigQuery SQL to compute the answer