

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:

<input type="checkbox"/> Status	Name ↑	Zone	Recommendations	In use by	Internal IP	External IP	Connect
<input type="checkbox"/>	instance-2	us-west1-b			10.138.0.2 (nic0)	34.105.25.145 (nic0)	SSH ▾

Get Zip file onto Compute Engine Instance

OBJECTS

CONFIGURATION

PERMISSIONS

PROTECTION

LIFECYCLE

OBSERVABILITY

INVENTORY REPORTS

Folder browser

cs512-aircraft-protzela

plane_data/

Buckets > cs512-aircraft-protzela

CREATE FOLDER

UPLOAD ▾

TRANSFER DATA ▾

OTHER SERVICES ▾

Filter by name prefix only ▾

Filter

Filter objects and folders

Show [Live objects only](#) ▾

<input type="checkbox"/>	Name	Size	Type	Created	Storage class
<input type="checkbox"/>	plane_data/	—	Folder	—	—

Load JSON files into Google Cloud Storage

Recipe

Data

Data Preview

{}	column1
{ "src": 1, "feeds": [{ "id": 1, "name": "From Consolidator", "pola...	
{ "Id": 13577240, "Rcvr": 11095, "HasSig": true, "Sig": 45, "lca...	
{ "Id": 8707694, "Rcvr": 11095, "HasSig": true, "Sig": 33, "lcao...	
{ "Id": 11111467, "Rcvr": 11095, "HasSig": true, "Sig": 43, "lca...	
{ "Id": 11361406, "Rcvr": 11030, "HasSig": true, "Sig": 11, "lca...	
{ "Id": 10846476, "Rcvr": 11259, "HasSig": true, "Sig": 104, "l...	
{ "Id": 8870094, "Rcvr": 11095, "HasSig": true, "Sig": 41, "lcao...	

Size

1 column · 1 type

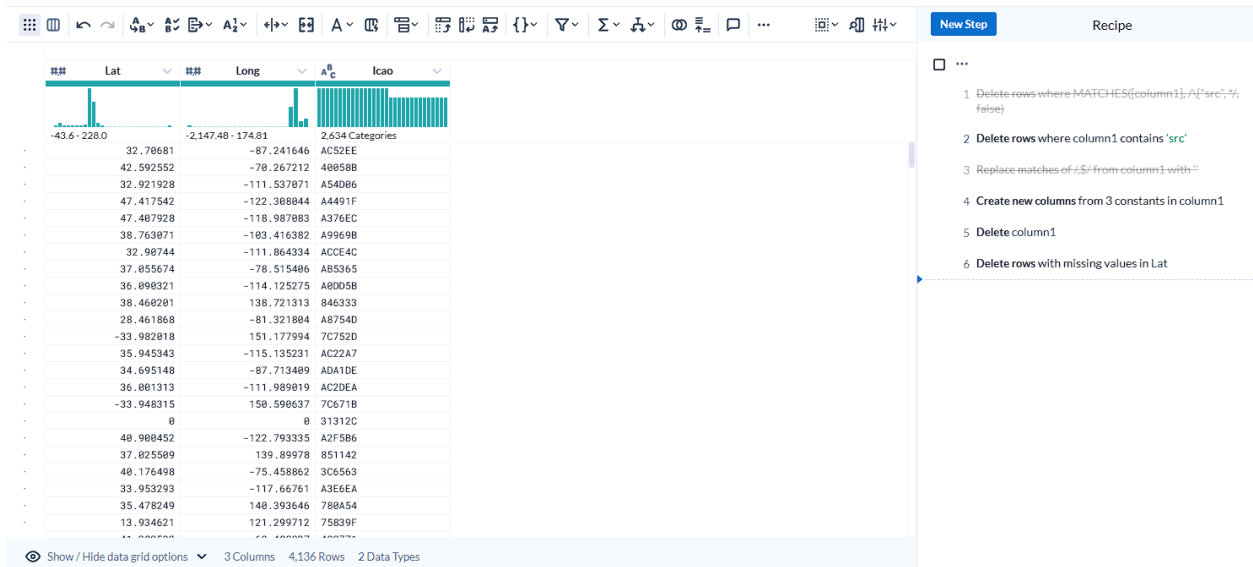
Updated

Today at 7:27 PM

Created

Today at 7:24 PM

Load JSON files as a data set into Google Cloud Dataprep



Parse JSON into appropriate columns in Dataprep

The screenshot shows the BigQuery Explorer interface. The left sidebar shows the 'aircraft_data' dataset with the 'plane_data' table selected. The main panel shows the schema for the 'plane_data' table.

Field name	Type	Mode	Key	Collation	Default Value	Policy Tags	Description
Icao	STRING	NULLABLE	-	-	-	-	-
PosTime	INTEGER	NULLABLE	-	-	-	-	-
Lat	FLOAT	NULLABLE	-	-	-	-	-
Long	FLOAT	NULLABLE	-	-	-	-	-
Alt	INTEGER	NULLABLE	-	-	-	-	-

Export Dataprep job into BigQuery

plane_data

*Untitled query

Untitled query

RUN

SAVE

DOWNLOAD

SHARE

SCHEDULE

OPEN IN

MORE

1

SELECT count(distinct Icao) FROM `cs512-aircraft-protzela.aircraft_data.plane_data`

2

WHERE (Lat between (44.497222 - 0.2) AND (44.497222 + 0.2))

3

AND (Long between (-123.289444 - 0.2) AND (-123.289444 + 0.2))

Query results

JOB INFORMATION

RESULTS

CHART

JSON

EXECUTION DETAILS

EXECUTION GRAPH

Row	f0_
1	85

BigQuery SQL to compute the answer