Implementation of Dataflow pipeline

NYC AIRBNB

Akhilesh Borgaonkar

Contact: 203-570-1279
Email: borgaonkar.akhilesh@gmail.com

Problem statement:

Create a pipeline in Dataflow that reads data from a csv file, applies transformations, and inserts resulting data into BigQuery table.

Implementation:

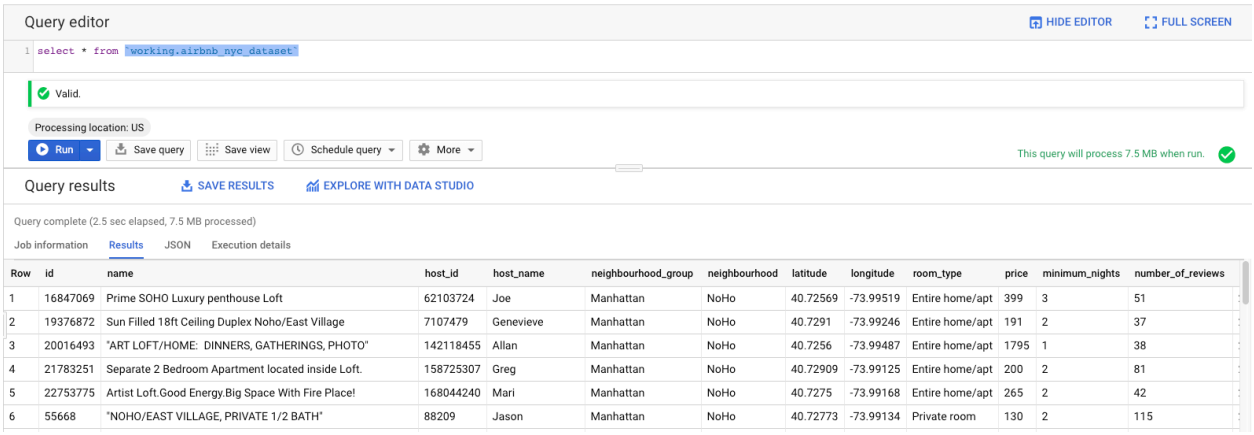
I have implemented the Dataflow pipeline with the help of "google-cloud-dataflow-java-sdk" and "Google Cloud Tools". The steps in the implementation of Java application are as follows:

1. Initialized an object of "org.apache.beam.sdk.Pipeline" class to leverage the Dataflow execution phases in the form of Directed Acyclic Graph.
2. Customized configurations for the pipeline object according to the job details.
3. Created nodes for the DAG execution of Pipeline. Find the stages of DAG execution as below:
    i. **Read** the input NYC-airbnb CSV file stored in GCS bucket.
    ii. **Extract** the fields and rows from the input file & load into Collections.
    iii. **Transform** the input rows to BigQuery compatible row format. Perform Group By aggregation on Neighborhood field.
    iv. **Load** the data to the BigQuery table.

Output:

I have compiled set of screenshots to leverage the successful ETL processing of NYC-airbnb dataset to BigQuery. Please find them below:

Figure 1: Job execution DAG in Dataflow UI



Query editor

HIDE EDITOR FULL SCREEN

```
select * from `working.airbnb_nyc_dataset`
```

Valid.

Processing location: US

Run | Save query | Save view | Schedule query | More

This query will process 7.5 MB when run.

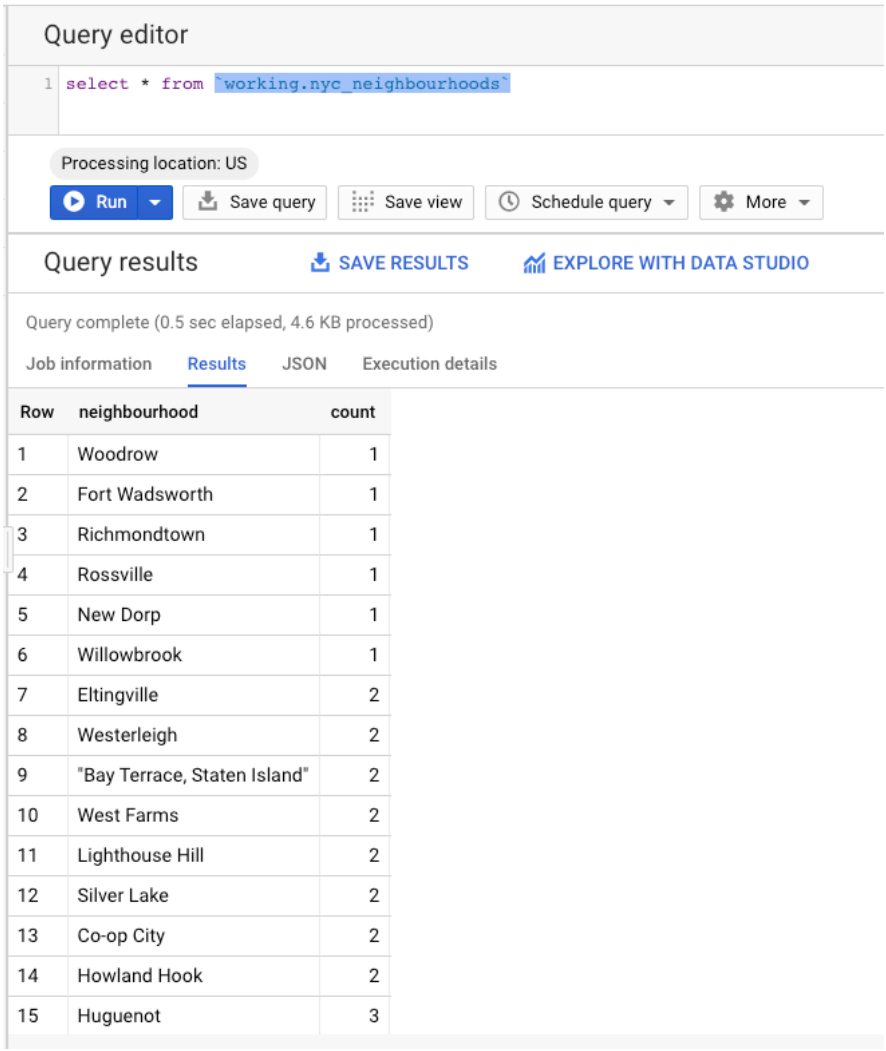
Query results SAVE RESULTS EXPLORE WITH DATA STUDIO

Query complete (2.5 sec elapsed, 7.5 MB processed)

Job information | Results | JSON | Execution details

Row	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews
1	16847069	Prime SOHO Luxury penthouse Loft	62103724	Joe	Manhattan	NoHo	40.72569	-73.99519	Entire home/apt	399	3	51
2	19376872	Sun Filled 18ft Ceiling Duplex Noho/East Village	7107479	Genevieve	Manhattan	NoHo	40.7291	-73.99246	Entire home/apt	191	2	37
3	20016493	"ART LOFT/HOME: DINNERS, GATHERINGS, PHOTO"	142118455	Allan	Manhattan	NoHo	40.7256	-73.99487	Entire home/apt	1795	1	38
4	21783251	Separate 2 Bedroom Apartment located inside Loft.	158725307	Greg	Manhattan	NoHo	40.72909	-73.99125	Entire home/apt	200	2	81
5	22753775	Artist Loft.Good Energy.Big Space With Fire Place!	168044240	Mari	Manhattan	NoHo	40.7275	-73.99168	Entire home/apt	265	2	42
6	55668	"NOHO/EAST VILLAGE, PRIVATE 1/2 BATH"	88209	Jason	Manhattan	NoHo	40.72773	-73.99134	Private room	130	2	115

Figure 2: BigQuery table view - Parsed dataset (Module 1)



Query editor

```
select * from `working.nyc_neighbourhoods`
```

Processing location: US

Run | Save query | Save view | Schedule query | More

Query results SAVE RESULTS EXPLORE WITH DATA STUDIO

Query complete (0.5 sec elapsed, 4.6 KB processed)

Job information | Results | JSON | Execution details

Row	neighbourhood	count
1	Woodrow	1
2	Fort Wadsworth	1
3	Richmondtown	1
4	Rossville	1
5	New Dorp	1
6	Willowbrook	1
7	Eltingville	2
8	Westerleigh	2
9	"Bay Terrace, Staten Island"	2
10	West Farms	2
11	Lighthouse Hill	2
12	Silver Lake	2
13	Co-op City	2
14	Howland Hook	2
15	Huguenot	3

Figure 3: Output of Group By transform on Neighbourhood field (Module 2)

Note: For detailed output, please find the attached files.