Implementation of GCP Dataflow

NYC AIRBNB

Shriyanshi Shikha

Contact: 765-720-6222

Email: shriyanshikha@gmail.com

Problem Statement:

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

Implementation:

I have implemented the Dataflow pipeline with the help of “Apache Beam SDK for Python” and “Google Cloud Tools”. The steps in the implementation of python application are as below:

I. Created a GCP account and enabled necessary APIs
II. Created a new project
III. Created a Service account to access certain GCP services and also performed the setup of Google Cloud Shell
IV. Created a service account key and downloaded the JSON file that contains the service account key
V. Created a Cloud storage bucket to use the input data i.e. AB_NYC_2019
VI. Created a BigQuery dataset manually under the same GCP project

After setting up the account I used the command prompt to set the virtual environment and once that’s done I defined a pipeline with an Apache Beam program and chose the runner as Dataflow to run the pipeline.

I initialized an object class to leverage the Dataflow execution phases in the form of Directed Acyclic Graph. Find the stages of DAG execution as below:

I. **Read** the input NYC-airbnb CSV file stored in GCS bucket
II. **Extract** the field and rows from the input file and load into collections
III. **Transform** the input rows to BigQuery compatible row format. Perform Group By aggregation on Neighbourhood field
IV. **Load** the data into bigQuery table

```
with beam.Pipeline(options=pipeline_options) as p:
    lines=(p|"ReadFromFile" >> beam.Create([known_args.input])
            |"ParseCSV" >> beam.FlatMap(get_csv_reader)
            |"write to bigquery :" >>
                beam.io.WriteToBigQuery(table="springmltest",dataset="springmltest",project="springmltest",schema=table_schema,create_
                disposition=beam.io.BigQueryDisposition.CREATE_IF_NEEDED))
```

Figure 1: Dataflow pipeline to read the csv file and write to a bigquery table

```
with beam.Pipeline(options=pipeline_options) as p:
    nb_count=(p|'QueryTableStdSQL' >> beam.io.ReadFromBigQuery(

          query='SELECT neighbourhood, sum(calculated_host_listings_count) as count FROM '\
          '`springmltest.springmltest.springmltest` group by neighbourhood',
          use_standard_sql=True)
          |"write count to bigquery :" >>
              beam.io.WriteToBigQuery(table="neighbourhoodcount",dataset="neighbourhoodcount",project="springmltest",schema=table_cou
              nt_schema,create_disposition=beam.io.BigQueryDisposition.CREATE_IF_NEEDED))
```

Figure 2: Dataflow pipeline to read the original data from a bigquery table and run query to calculate listings count by neighbourhood

Output:

Below is a set of screenshots to leverage a successful ETL processing of NYC Airbnb dataset to BigQuery. Please find them below:

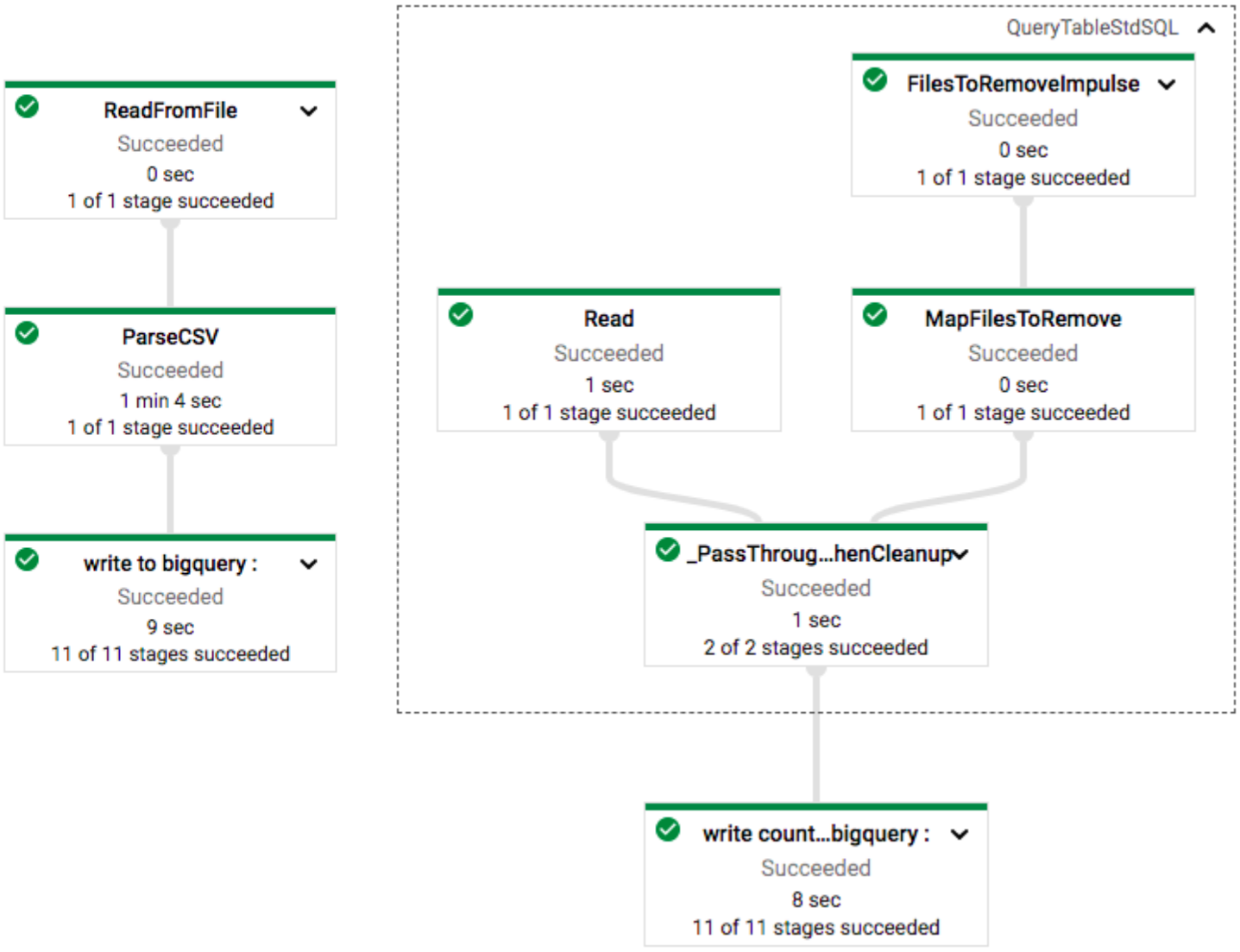
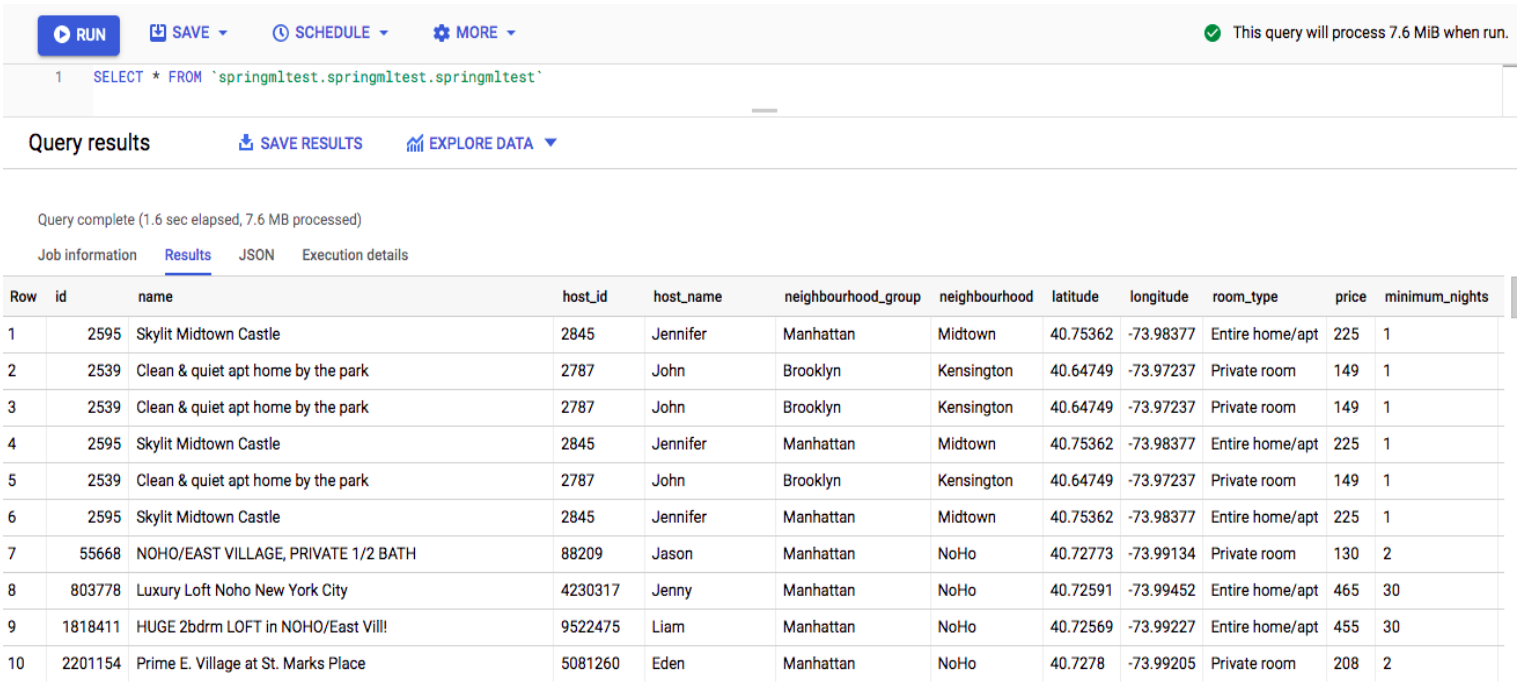
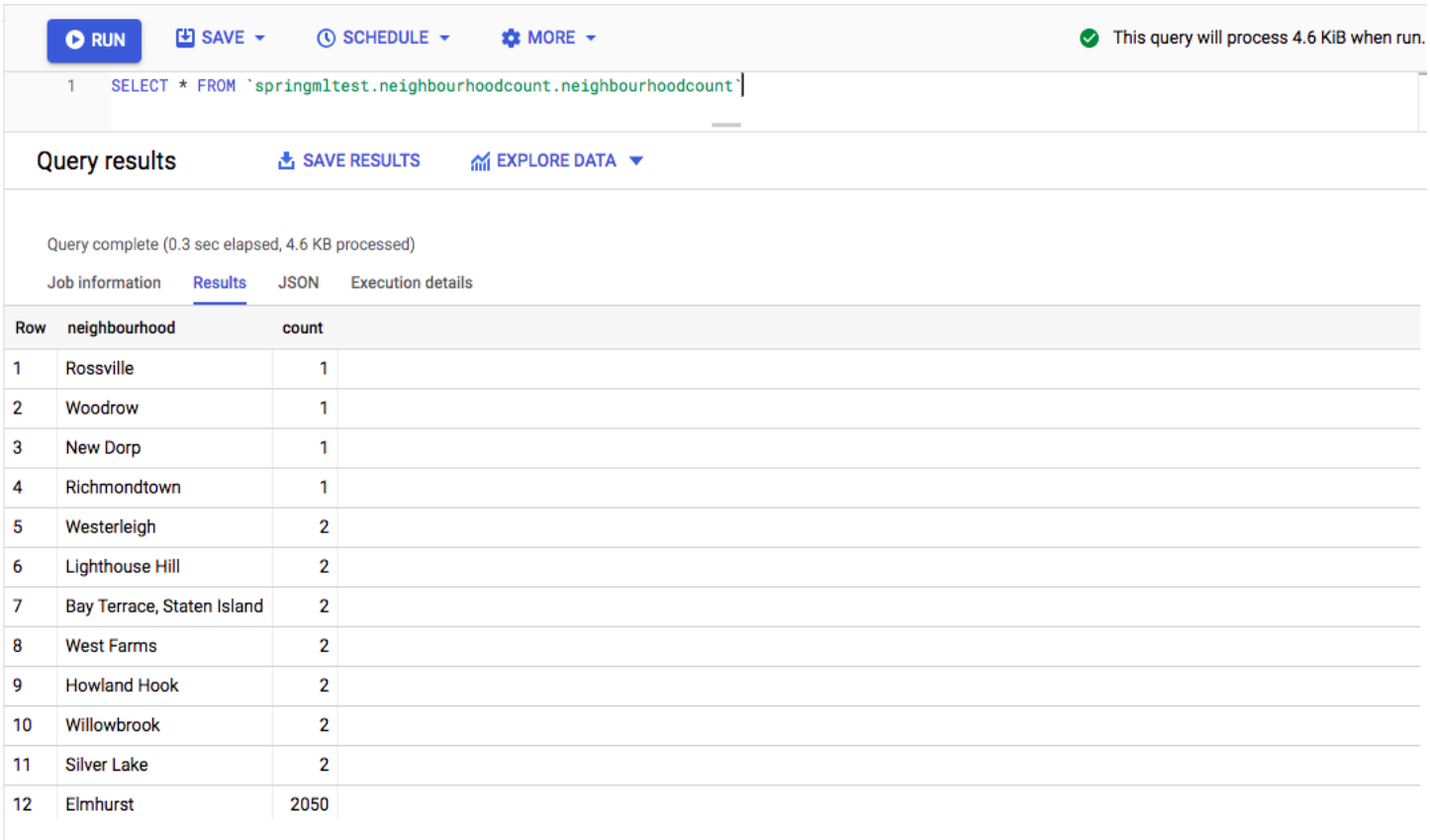


Figure 2: Job execution DAG in Dataflow UI



Row	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1
2	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1
3	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1
4	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1
5	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1
6	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1
7	55668	NOHO/EAST VILLAGE, PRIVATE 1/2 BATH	88209	Jason	Manhattan	NoHo	40.72773	-73.99134	Private room	130	2
8	803778	Luxury Loft Noho New York City	4230317	Jenny	Manhattan	NoHo	40.72591	-73.99452	Entire home/apt	465	30
9	1818411	HUGE 2bdrm LOFT in NOHO/East Vill!	9522475	Liam	Manhattan	NoHo	40.72569	-73.99227	Entire home/apt	455	30
10	2201154	Prime E. Village at St. Marks Place	5081260	Eden	Manhattan	NoHo	40.7278	-73.99205	Private room	208	2

Figure 3: BigQuery table:springmltest (contains data from original csv)



Row	neighbourhood	count
1	Rossville	1
2	Woodrow	1
3	New Dorp	1
4	Richmondtown	1
5	Westerleigh	2
6	Lighthouse Hill	2
7	Bay Terrace, Staten Island	2
8	West Farms	2
9	Howland Hook	2
10	Willowbrook	2
11	Silver Lake	2
12	Elmhurst	2050

Figure 4: BigQuery table:neighbouthoodcount (contains count of listings grouped by neighbourhood)