GOOGLE CLOUD SERVICES PROJECT

INTRODUCTION

Google cloud platform stands out from the rest of the cloud service providers because of its Community support, cost effective comparatively, flexible, Scalable, reliable, fully managed, and fast services. Through google cloud services one can concentrate on discovering insights instead of dealing with the infrastructure and can join cloud services with open source tools as required in both batch and streaming mode(Liu *et al.* 2012). It offers various services and one among them is Big data services. They are as follows

1. **Google BigQuery** is completely managed, minimal cost datawarehouse. Its serverless, there is no framework to manage and there is no compelling reason to figure the limit or overprovision. There is no need of data organization, one can simply concentrate on analysing the information and full insights.
2. **Google Dataproc** is managed spark and Hadoop administration which is utilized to effectively process enormous datasets utilizing open tools of apache Hadoop framework. One can control the cost utilizing clusters of any size in about a moment and can kill them when done.
3. **Google cloud Datalab** is an intelligent scratch pad which depends on jupyter and is utilized to perform various functions like exploring, visualizing information. It is coordinated with BigQuery and google cloud AI to give the client a simple access to the key information handling administrations.
4. **Google Pub/Sub** is a serverless, enormous scale, dependable framework administration that enables one to send and get messages between free application. One can use its adaptability to decouple framework and parts on the cloud stage.
5. **Google dataflow** is a processing service which enables users to build a pipeline for analzing the dataset using templates for both batch and streaming data.

BACKGROUND

As of late, the market for cloud computing has developed surprisingly be it a large organisation or small start-ups. They all use the cloud services in one form or the another. According to the Gartner’s prediction the Worldwide public class service market will increase from one hundred fourty billion dollars in 2019 to one hundred seventy billion dollars company in 2020 and will continue to grow with alarming rate of twenty two percent annually(Challita *et al.* 2018). Google provides this cloud services. Firstly, Google cloud platform is a set of computing, networking, storage, big data and machine learning and management services provided by google that stands on the same cloud infrastructure that google uses internally for user products.

USER REQUIREMENT

In this project starbucks.csv file is downloaded from Kaggle which was publicly available. User requirement for this project is to know which countries in the world has the highest number of Starbucks coffeehouse.

DATA PIPELINE DESIGN

dataset

Dataprep

Dataflow

BigQuery

The data pipeline design for this project was as per the design shown above. Its working is as follows:

1. Initially the dataset is uploaded into the storage service on Google cloud platform
2. In order to clean the dataset it is exported to Dataprep service where appropriate cleaning tools are available for different types of cleaning purposes.
3. Once the dataset is cleaned in Dataprep it automatically gets stored in the storage console under the storage bucket created. Now in order to transform the cleaned dataset into Bigquery a pipeline is built using Dataflow service
4. Further to check whether the datapipeline is successfully built, a simple sql is executed in the Bigquery console. Here, the preview of the dataset will be displayed only if the dataset is successfully transformed from the cloud storage into the table created to hold the dataset in Bigquery.
5. In order to uncover the insights of the dataset and to meet the user requirement, Google data studio is used where the reports are generated by using different charts to visualize and to carry out this process, the dataset present in Bigquery is exported to google data studio by clicking on export to Google data studio option on the panel in Bigquery console.

MAJOR DEPLOYMENT STEPS

**Section A: Uploading the dataset and cleaning it using Dataprep**

**Step 1**: Firstly, the storage bucket is created by clicking on the **Storage** option on **Navigation menu** in the left hand side of the console in order to hold the files

**Step 2:** Secondly, Dataprep service is used to check whether the dataset i,e **starbucks.csv** file is clean or not. If not, its cleaned by using different cleaning tools available. Here, starbucks.csv file is imported and columns showing red colour on the title will be edited as it indicates the column is unclean.

**Step 3:** Once after the dataset is clean, **Run job** option is clicked to store the cleaned dataset under the storage bucket created earlier. Besides, by clicking on the project id reflected on the left side of the console in BigQuery service, a destination table is created to hold the dataset.

**Section B: Building a data pipeline using Dataflow service**

**Step 1:** In order to build a data pipeline to transform the cleaned batch data stored under the storage bucket to BigQuery, Dataflow service is used. Here to execute the job, **Text files on Cloud storage to BigQuery** template is selected.

**Step 2:** The template selected requires **json** and **.js** file where the schema of the dataset and logic to transform the lines of text should be defined respectively. Based on these files a data pipeline is built to transform the batch data from the storage to Bigquery. Therefore, json file along with .js file is created and uploaded into google cloud storage console.

**Step 3:** Further, the job is executed, and the data pipeline is built

**Section C: Execution of sql command in BigQuery to check whether the data pipeline is created successfully**

**Step 1:** In the BigQuery console a sql command is exceuted to display the preview of the file to check whether the starbucks.csv file is successfully loaded into the destination table named ‘realtime’, from the cloud storage.

**Step 2:** Once after confirming the presence of a file in the destination table, its exported to Google data studio by clicking on **Explore in data studio** option present on the top section of the console

**Section D: Exploring Data studio to visualize the dataset**

**Step1:** After exporting the file from bigquery to data studio it can be visualized by using various charts like tables, bar charts, time series, pie charts, scatter and bubble charts and so on.

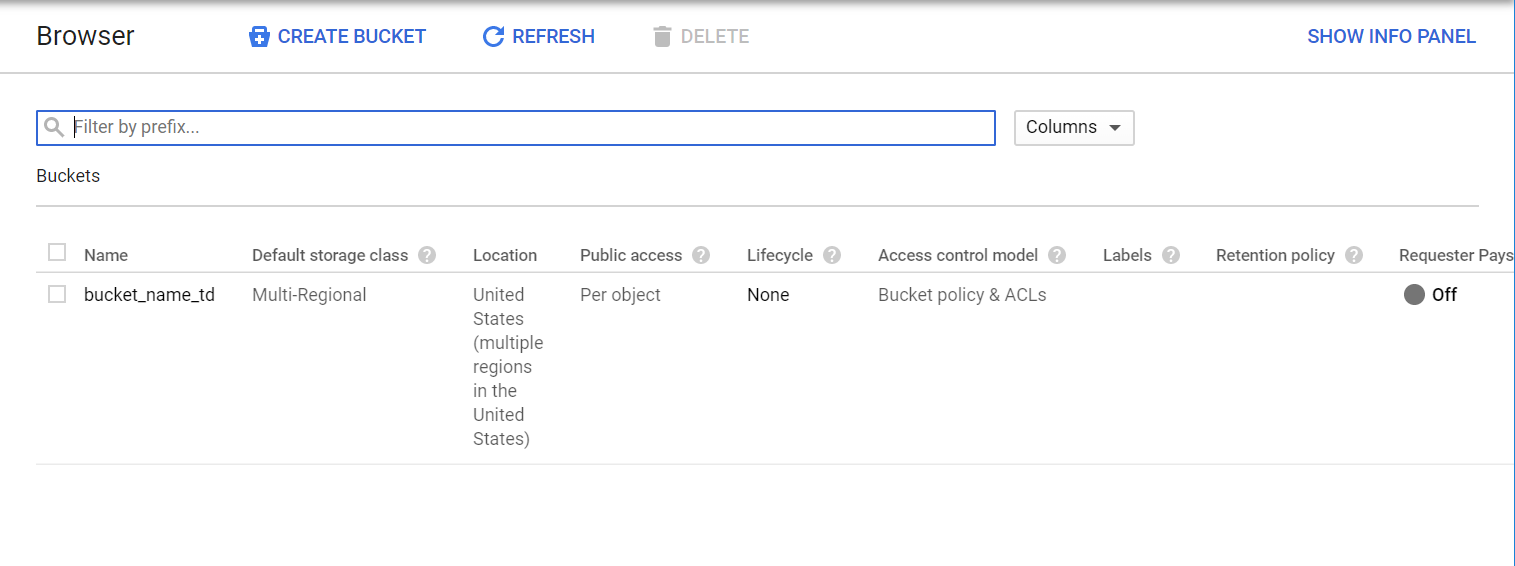
**Step 2:** Here, the starbucks dataset is visualized using **geo map** chart by selecting **country** in dimension section and **store\_number** in metric section.

**Step 3:** Further to make the visualization report more attractive, legend can be enabled where different colours can be selected that distinguishes countries having different data. Similarly background and border colour can also be set.

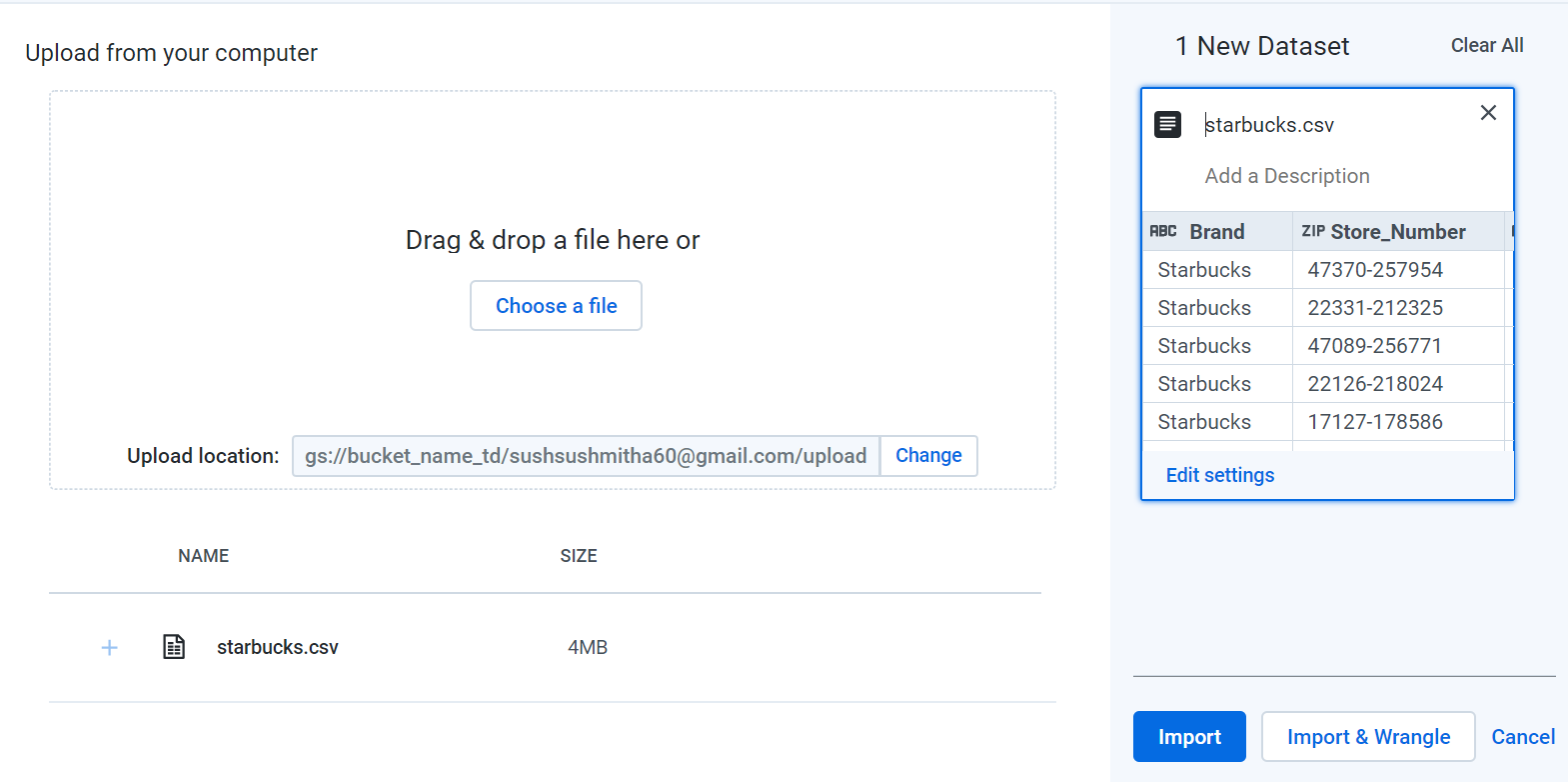
RESULTS

**Results of Section A:**

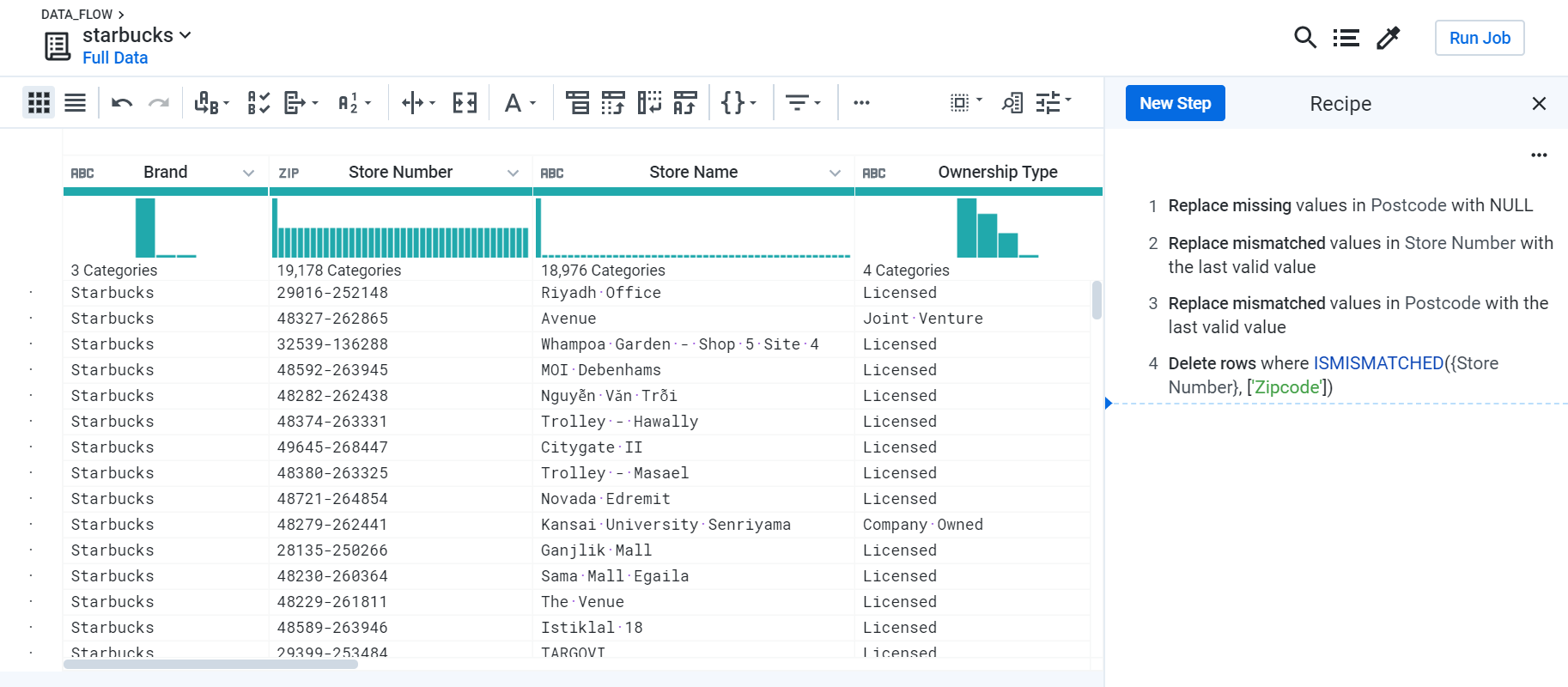
**Task 1:** A bucket is created named ‘bucket\_name\_td’ as shown below by clicking on the **create bucket** option present at the top of the panel



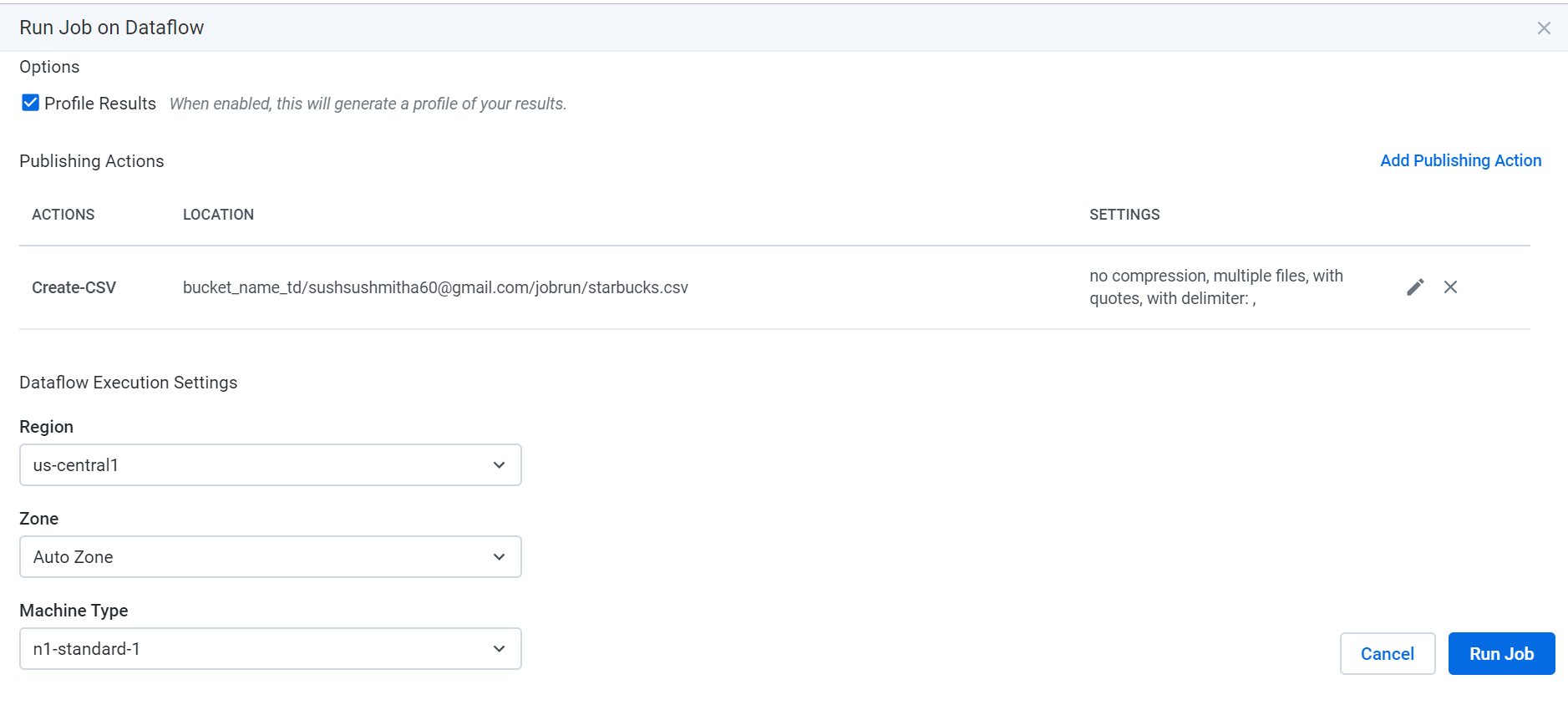
**Task 2:** Here starbucks.csv file is imported from local storage in Dataprep as shown below to clean the file



**Task 3:** In starbucks.csv file, missing values found in Postcode column were cleaned by replacing it with NULL values. In Store Number and Postcode columns there were mismatched values and to clean them they were replaced with the last valid value as shown below. Here all these steps are displayed in the right side of the window under Recipe section as shown below

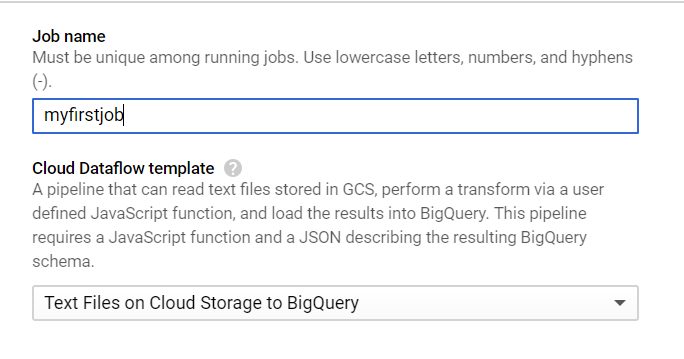


**Taks 4:** Once after the file is free from all errors, **Run job** option is clicked and a new window pops up showing the location where the cleaned file will be stored as shown below

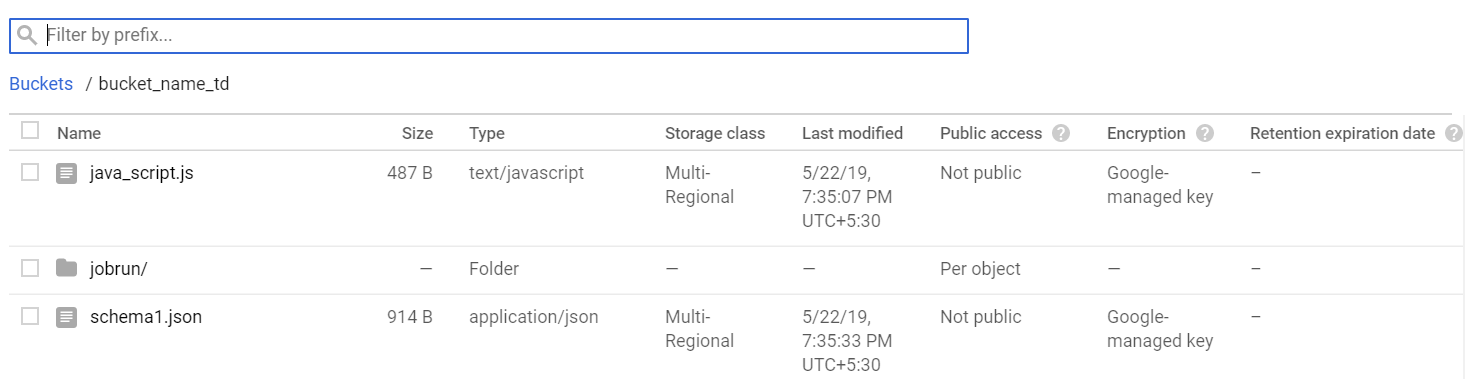


**Results of Section B:**

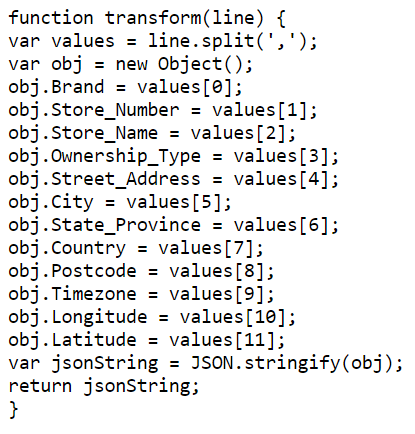
**Task 1:** To execute a job to transform the dataset from the cloud storage to Bigquery ‘**Text Files on Cloud Storage BigQuery’** template is selected as shown



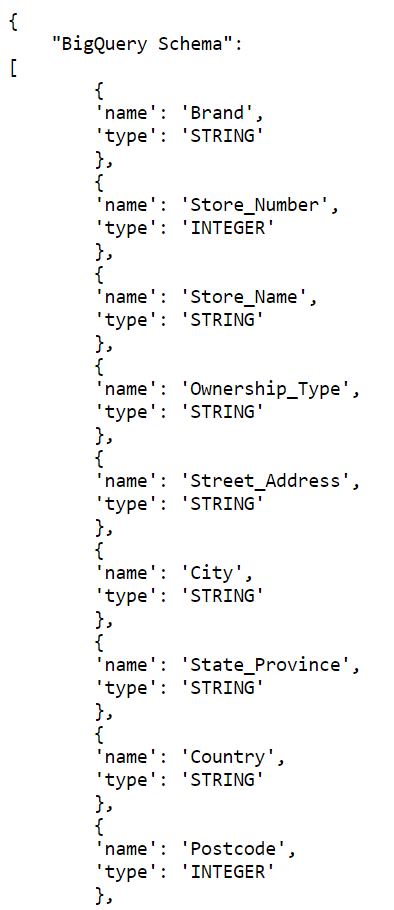
**Task 2:** The template selected requires **json** and **.js** file where the schema of the dataset and logic to transform the lines of text is defined. Therefore, **json** file along with **.js** file is created and uploaded into google cloud storage console and based on these files a datapipeline is built to transform the batch data from the storage to Bigquery



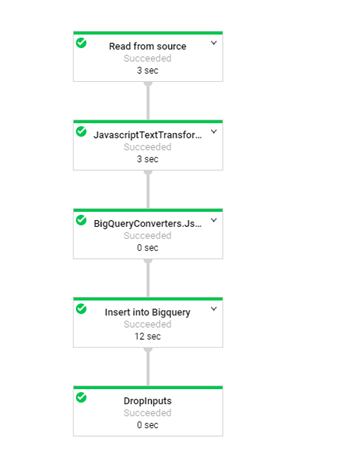
The **.js** file contains the logic to transform the lines of text as shown below



The **.json** file is created where the schema of the starbucks.csv file is defined as shown below

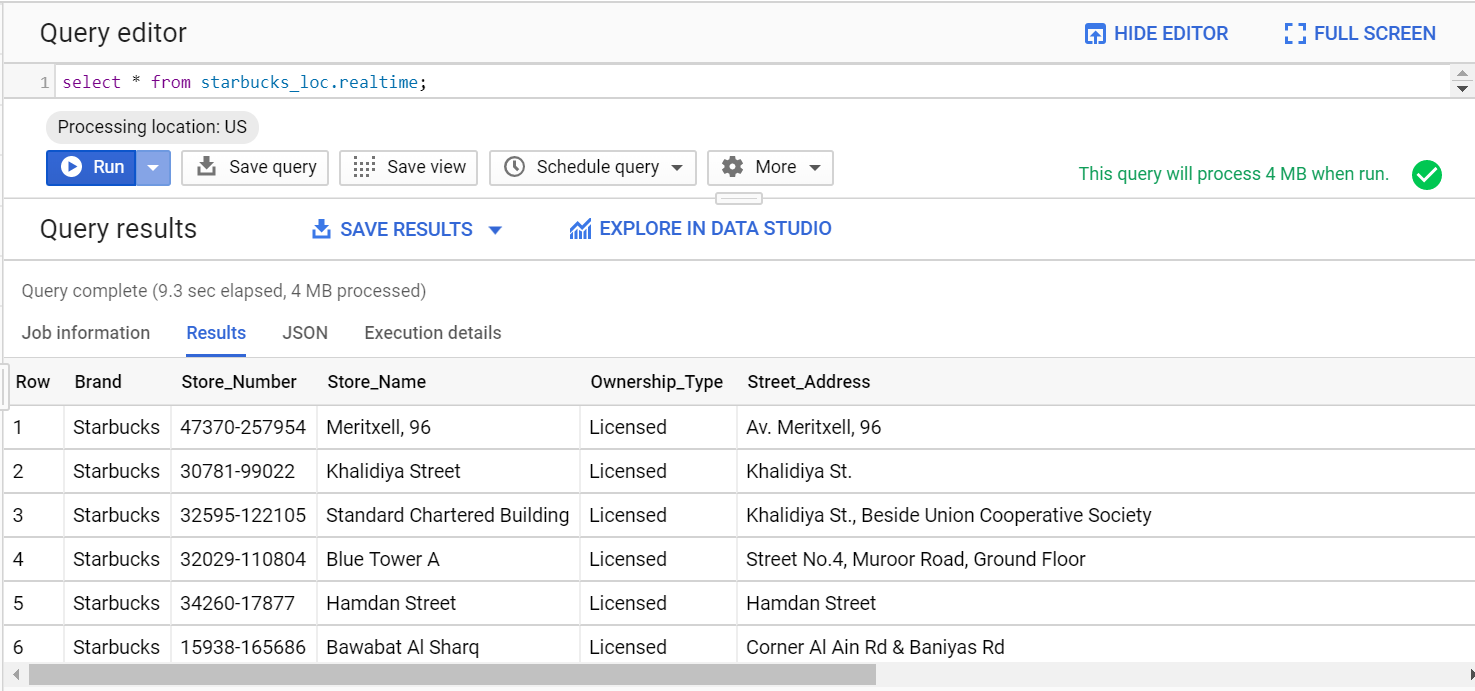


**Task 3:** Once after the required configurations is set up, the job is executed and the data pipeline is created as shown



**Results of Section C:**

**Task 1:** A sql command is exceuted to display the preview of the file to check whether the starbucks.csv file is successfully loaded into the destination table named ‘realtime’, from the cloud storage in BigQuery service as shown



**Task 2:** Further, the file is exported to Google data studio by clicking on **Explore in data studio** option present on the top section of the console

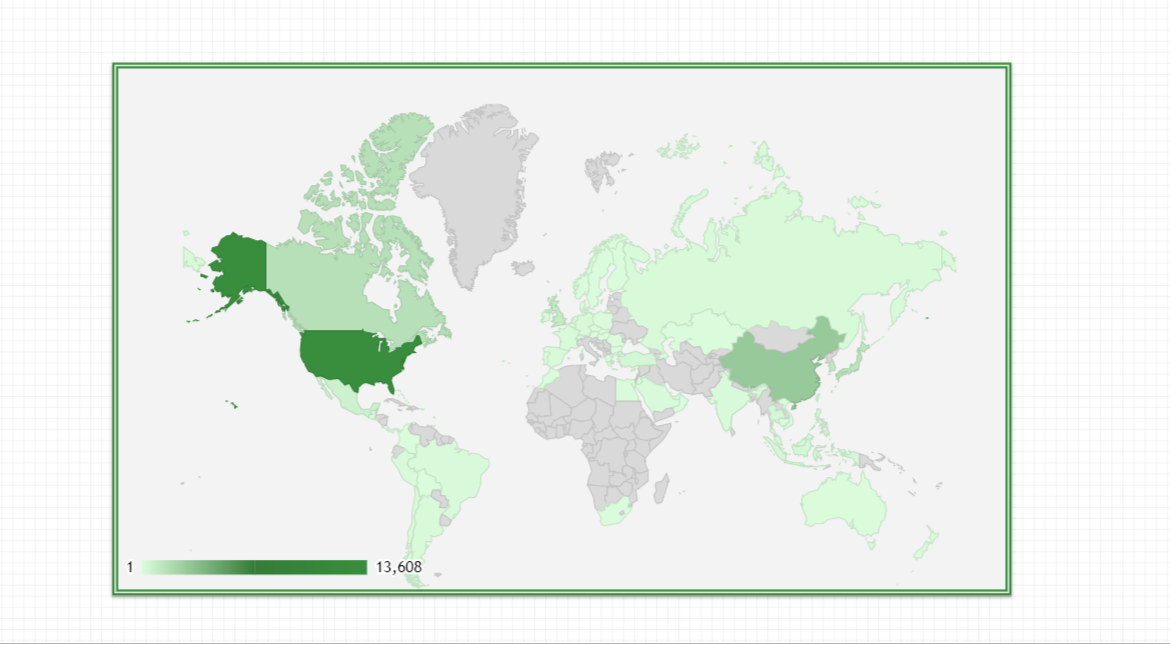
**Results of Section D:**

**Task 1:** In the google data studio in order to seek the answer for the user requirement i,e which country has the highest number of Starbucks coffeehouse?, **Geomap** chart is used for the visualization. Here, since the user needs to know which country and hence the **country** column is dragged and placed in the data dimension section and similarly **Store\_Number** column is placed in Metric section as shown below since Store\_Number is unique for each Store\_Name and this enables the chart to effectively to show the number of Starbucks located in each country.

A screenshot of a cell phone

Description automatically generated

**Task 2:** After placing all the required columns in each section, the **Geomap** is displayed as shown where it indicates **United States** and **Alaska** shaded in dark green has got more number of Starbucks coffeehouse when compared to **China** and **Canada** shaded in neither too dark nor pale green. On the other hand, countries shaded with pale green and grey has fewer and no Starbucks coffeehouse at all respectively.



CONCLUSION

Building a data pipeline by using Google cloud services available on Google cloud platform was easy after having hands on experience with qwik labs where each step was explained in detail. Perhaps, it enabled to uncover the insights to know the actual working of each cloud services namely Dataprep, dataflow, Bigquery and Google data studio in building up datapipeline and also to visualize using different charts to meet the user requirement.

In this project since the user requirement was to know the country having maximum number of Starbucks coffeehouse, a datapipeline was built and then it was visualized using geomaps to enable the user to actually visualize the country having more number of starbucks coffeehouse by shades of colours marked on different countries. As a result, it was observed countries namely **United States** and **Alaska** has got a greater number of Starbucks coffeehouse when compared to **China** and **Canada**. On the other hand, it was known that countries shaded with pale green and grey has fewer and no Starbucks coffeehouse at all respectively.

Overall, Google cloud platform was a great tool that allowed to build solutions very quickly and deploy it where that can scale to whatever the user needs.

REFERENCES

1. Challita, S., Zalila, F., Gourdin, C., Merle, P. (2018) ‘A Precise Model for Google Cloud Platform’, in *2018 IEEE International Conference on Cloud Engineering (IC2E)*, Presented at the 2018 IEEE International Conference on Cloud Engineering (IC2E), IEEE: Orlando, FL, 177–183, available: https://ieeexplore.ieee.org/document/8360326/ [accessed 26 May 2019].
2. Liu, J., Yang, G., Wu, H., Zheng, L. (2012) ‘Logistics information management system based on Google cloud computing platform’, in *2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet)*, Presented at the 2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet), IEEE: Yichang, China, 2685–2687, available: http://ieeexplore.ieee.org/document/6201876/ [accessed 26 May 2019].