Implementing ETL on Yelp Restaurant Data in Mount Pleasant, MI using Google Cloud Platform

Yasaswi Avula
dept. Computer Science
Central Michigan University
Mt Pleasant, MI
avula1y@cmich.edu

Gowthami Nathani
dept. Computer Science
Central Michigan University
Mount Pleasant, MI
natha1g@cmich.edu

Manikanta Prayaga
dept. Computer Science
Central Michigan University
Mount Pleasant, MI
praya1mn@cmich.edu

***Abstract*—The goal of this project is to demonstrate the us of Google Cloud Platform for deploying the Extract, Transform and Load (ETL) on Yelp Website. This ETL (Extract-Transform Load) processes in Cloud or on-premises are responsible fo integrating data into a place called Data warehouse. Specifically the aim is to perform Extraction of the Yelp restaurant data i Mt Pleasant, MI and then transforming it into structured dat using a Virtual Machine and then loading the .csv file containin the yelp restaurant data into Cloud Storage**

I. Introduction

Yelp.com is a business review website which contain ratings and reviews of different businesses located in specific locations. For business owners, these yelp reviews can improve the businesses by attracting more potential customers, and owner can address customer needs in real time by launching new products or services tailored by their reviews This is a very important step in the decision- making process and called ETL (Extract - Transform - Load). Here we chose to perform ETL on Restaurants located in Mount Pleasant, MI. Then we wanted to implement the project in Google Cloud Platform because of its cost efficiency, Quick Collaboration and ease of Cloud Management.

II. Challenges

- The foremost challenge is the Code deployment process in the Local System. It was a bit trouble as the data of each page is not extracted I.e., when there are about 100+ restaurants in Mount Pleasant, it just extracted the first 10 restaurants displayed on the first page. So, we implemented Pagination and wrote the code using recursive functions which enabled us to extract all page's data.
- The very next challenge is Redundant data. While implementing the transformation process, we've removed all the redundant data, cleaned and reported a structured data frame, available for converting it into .csv file.
- When we implemented the functioning code in Server less cloud functions in GCP, it returned errors like, cannot handle the request. So, we Changed code as advised by the professor but still faced same errors.

Identify applicable funding agency here. If none, delete this.

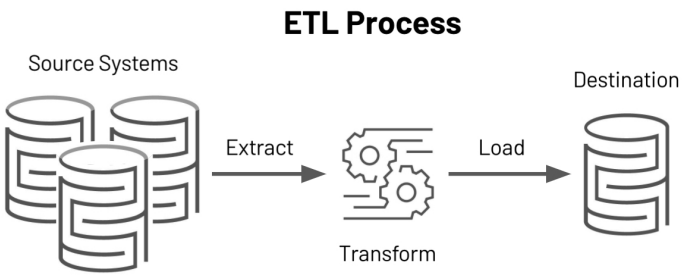

Fig. 1. ETL Process.

- We've created a Virtual Machine Instance and performed the code there in Secure shell (SSH) which gave us the output we expected but faced issues while connecting with GS (Google Storage) and installing python. Later we connected to GS and stored the contents in the Storage Bucket.

III. ETL Use case

In Real time, How the ETL data is used for any Business or Organization? Any business will only thrive when it meets customer expectations and needs. Consider this, If the owner doesn't have the knowledge about how a customer thinks or feels about their business, it will become the dooms day for the business. So, the business owners need to have an eye on these reviews, data from ETL pipelines running in cloud and perform BI analysis to understand about their business. For Example, when a user had a bad experience with certain restaurant in Location A of a big Chain Restaurant like Chick-Fil-A, He will post the review about how he feels in Yelp. These reviews may catch sight for other customers as well, which may impact the revenue of that restaurant in location A. The owner will never know the cause until he finds it.The owners need to compensate the user and try not to repeat the mistakes again. This benefits both users and the management. This is how the ETL is implemented by data in real time.

IV. ETL IMPLEMENTATION

A. Extract

Yelp website is extracted from the local system to the Debian virtual machine by creating virtual instances. This instance name is used to run workloads on Google infrastructure, and we can execute the program by connecting to Debian Linux using secure shell. Performing extraction on the Python platform and required libraries by giving the required commands in shell to run the file.

B. Transform

For effective decision making, data must go through a transformation process that involves six basic steps: 1) data collection, 2) data organization, 3) data processing, 4) data integration, 5) data reporting, and finally 6) data utilization. We collected the data in an unstructured format and organized that data by using response and request URLs. Following this step the processing takes place, like converting the data to a structured format to perform operations easily . For example a specific restaurant link was integrated to the default website link to direct it to the working restaurant link. Finally,the data is converted and retrieved by merging all the above steps.

C. Load

The data from the virtual instance can be loaded into a storage bucket, and the CSV file can be downloaded from there for future reference.

V. CLOUD SERVICES

A. Identity and Access Management

Identity and Access Management (IAM) lets you create and manage permissions for Google Cloud resources. IAM unifies access control for Google Cloud services into a single system and presents a consistent set of operations. So, here we assigned roles among the teammates such as viewing, modifying and manage access to the resources. In IAM, access is granted through allow policies, also known as IAM policies. An allow policy is attached to a Google Cloud resource. This makes easy to gain access and make changes from anywhere.

B. SSH connections

Compute Engine uses key based SSH authentication to establish connections to all Linux virtual machine instances. You can enable SSH for Windows VMs. By default, passwords aren't configured for local users on Linux VMs. several configurations must be performed before you connect to a VM. If you use the Google Cloud console or the Google Cloud CLI to connect to your VMs, Compute Engine performs these configurations on your behalf. Compute Engine performs different configurations depending on which tool you use to connect and whether you manage access to VMs through metadata.

C. Cloud Storage

Buckets are the basic containers that hold your data. Everything that you store in Cloud Storage must be contained in a bucket. You can use buckets to organize your data and control access to your data, but unlike directories and folders. We performed data loading from the VM instance created and retrieved the data.

D. VM instances

A VM is a virtualized instance of a computer that can perform almost all the same functions as a computer, including running applications and operating systems. Virtual machines run on a physical machine and access computing resources from software called a hypervisor. Here, we created a VM instance, performed operations like loading data , data transformation and data retrieval.

E. Serverless

Serverless is faster and cost effective compared to other compute engines. Cloud functions can connect to google cloud via triggers and can be integrated via multiple environments. We created a serverless function to deploy our ETL model but unable to display the output without any errors. Serverless will be a great service for someone who is building on-top of already existing functionality.

VI. CONCLUSION AND FUTURE SCOPE

Finally we have achieved the goal of our project by demonstrating the use of google cloud platform by performing ETL on yelp restaurant data in Mount Pleasant data.Cloud fusion and big query can be used for data analysis and data warehousing, where massive amounts of data can be stored and pipelines run continuously to modify and update the data. In our scenario, Yelp website data can be stored in a big query and the owners can easily maintain the businesses effectively without manual changes.

REFERENCES

[1] H. S. and R. Ramathmika, "Sentiment Analysis of Yelp Reviews by Machine Learning," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 700-704, doi: 10.1109/ICCS45141.2019.9065812.

[2] Hajas, Peter and Gutierrez, Louis Krishnamoorthy, Mukkai. (2014). Analysis of Yelp Reviews.

[3] Luca, Michael, (2016), Reviews, Reputation, and Revenue: The Case of Yelp.com, No 12-016, Harvard Business School Working Papers, Harvard Business School, https://EconPapers.repec.org/RePEc:hbs:wpaper:12-016.

[4] P. S. Diouf, A. Boly and S. Ndiaye, "Variety of data in the ETL processes in the cloud: State of the art," 2018 IEEE International Conference on Innovative Research and Development (ICIRD), 2018, pp. 1-5, doi: 10.1109/ICIRD.2018.8376308.

[5] Frank Escobedo-Bailon1, Antonio Arqque-Pantigozo2, Carlos Alzamora-Aragon3, Blanca Pasco-Barriga4, Soledad Olivares-Zegarra5, Katherin Rodriguez-Zevallos6, "Cloud Technology As A Support For The ETL Process And Its Influence On Decision Making." International Journal of Aquatic Science ISSN: 2008-8019 Vol 12, Issue 02, 2021.

[6] E. M. Haryono, Fahmi, A. S. Tri W, I. Gunawan, A. Nizar Hidayanto and U. Rahardja, "Comparison of the E-LT vs ETL Method in Data Warehouse Implementation: A Qualitative Study," 2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), 2020, pp. 115-120, doi: 10.1109/ICIMCIS51567.2020.9354284