International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

Digital Intelligence on Google Cloud Platform

Munshi Mohammad Tawqeer[1], Dr. Chethana R. Murthy[2]

[1]Department of Information Science and Engineering, RV College of Engineering, Bengaluru, Karnataka, India
[2]Assistant Professor, Department of Information Science and Engineering, RV College of Engineering, Bengaluru, Karnataka, India

---

Abstract - *"Digital Intelligence on GCP" involves development of applications that aim to automate the analytic process using gcp and to build a powerful analytic tool for the company. a lot of resources are already present in the google cloud documentation where implementations are provided. The project aims to overcome the problem of manual involvement and extra effort on auxiliary processes. The methodology used involves firstly, building a gradle based java application using intellij that automates the auxiliary steps in the process of analysis. secondly, using gcp as a tool to leverage various services to automate and schedule transfers and to trigger events.*

The end result of the project shows that the whole analytic process significantly reduces the time and effort and the costs associated. the application does not require any human involvement or supervision, the work done is completely automated. this shall make the whole process over 2x efficient.

This goes to say that while we're aware of how automation is the new future we can also draw a conclusion that gcp is a wonderful tool not just for development but as well for analytics. tools like GCS, GBQ, AutoML make it the powerful tools it is. minimal latency, high efficiency, and cost efficient. Whereas the tool does prove to be very helpful there are a lot of additions that can be carried out in the course of future. Machine learning is one of the most important techniques in the field of analytics, we shall leverage the artificial intelligence tools of gcp like the automl for predictions on data, aligning results and much more. We look forward to adding multiple functionalities to the project.

Key Words: AWS S3, GCS, GBQ, Trigger, AutoML

1. INTRODUCTION

Google, a long-standing pioneer in data analytics, continues to regain its place by continually enhancing its offerings in data analytics. Now you can collect, process, store, and analyze your data in one location with Google Cloud Platform (GCP), enabling you to move your attention from infrastructure to analytics that inform business decisions. Nevertheless, GCP Big Data software can also be used in conjunction with other cloud-native and open-source applications to suit your needs. Below is a summary of GCP Big Data Tools and how you might use them to improve analytics.

BigQuery is probably the most powerful of GCP's Big Data resources that can be used by businesses of any scale. Using SQL, it is a managed, serverless, corporate data warehouse where you can analyze data in real time, but it also helps you to carry data from spreadsheets and object storage. BigQuery and other GCP solutions have been selected by companies like Spotify because Google's offerings are more sophisticated than the tools of other cloud providers, according to Nicholas Harteau, vice president of engineering at Spotify.

BigQuery helps you to free your workers from database management and concentrate instead on delivering information to boost the bottom line using onboard, reliable reporting and data extraction software. For example, Motorola increased its data collection capabilities after switching to BigQuery and App Engine that offered more knowledge to assist engineers with product troubleshooting to assist their customers.

BigQuery is also inexpensive when it comes to the bottom line. You never pay for the services you don't need by allowing you to scale up and down while your needs fluctuate. Google also provides you with 1 TB of processed data and 10 GB of free storage per month to further cut costs. As BigQuery is encrypted at rest and in transit, and further protected by granular access controls, you always keep your data secure. BigQuery also functions to track and sign on with StackDriver. Google also makes it simple to use their BigQuery Data Transfer Service to pass your data into BigQuery. This controlled service assists in transferring data from sources such as AdWords,YouTube, DoubleClick and other SaaS applications.

2. TERMINOLOGY AND ABBREVIATION USED

GCP : Google Cloud Platform
GCS : Google Cloud Storage
GBQ : Google Big Query
Project : In GCP, we can create projects within which all the work is done.
Bucket : All the data in GCS is stored in Buckets. This data may be files, images, videos etc.
Dataset: Tables are created and stored in Datasets in GBQ.
Tables : The data is stored in a structured manner in GBQ in the form of Tables.
AWS S3 : The data source of the Amazon web services.
Objects/Blobs : Objects may be referred to the files or the data stored within a bucket.

3. WHY MIGRATE FROM AWS TO GCP?

Analytics companies have much to do with Machine Learning. GCP offers some breakthrough machine learning capabilities which can be leveraged by us as a company to serve our clients better. The market is changing at a great pace, new innovations are happening particularly in this field. So, we also need to stay in pace with the world and explore other capabilities like GCP's ML tools and be a leader in our field.

Most of the companies have their own machine learning infrastructures that they use according to their own needs. These systems vary widely based on the company and individual needs. For example, an organisation that might be working on Artificial Intelligence has high end machine learning systems and tools to work on, Analytics companies on the other hand want to leverage the machine learning capabilities of various tools to improve their analytic services. Some of these organisations have their own Deep Learning systems called Active Learning Systems. The Google Cloud Platform can easily integrate with the Active Learning System and serve as a data source for training and testing of various models thus continuously improving the efficiency and accuracy of the model. This is complemented by the machine learning tool of Google Cloud Platform called AutoML.

4. WHAT IS AUTOML?

Although the field of AutoML has been around for years (including open-source AutoML libraries, workshops, research, and competitions), in May 2017 Google co-opted the term *AutoML* for its neural architecture search. In blog posts accompanying announcements made at the conference Google I/O, Google CEO Sundar Pichai wrote, *"That's why we've created an approach called AutoML, showing that it's possible for neural nets to design neural nets"* and Google AI researchers Barret Zoph and Quoc Le wrote *"In our approach (which we call "AutoML"), a controller neural net can propose a "child" model architecture…"*

In it's latest updates, the new AutoML vision and video intelligence is a breakthrough in the marketing intelligence industry. This is instrumental for organisations as this is the digital age and most of the marketing initiatives are taken on the web or the media. In such scenarios, the marketing teams of these companies can do a huge amount of analytics at minimal cost. Their marketing campaigns can be analysed by leveraging these tools.

5. AUTOMATING GOOGLE CLOUD PLATFORM

The Google Cloud Platform has high flexibility of development. Cloud Developers, Data Engineers or mere Software Developers can use GCP in any way they want. Google has Open-Sourced almost all of their functionalities in their documentation section. The functionalities are available (in addition to their UI/Console) as CLI commands, or functions/Code snippets in Java, Go, C#, Python, Node.js etc. The gives the developer the liberty to modify the functionality in any way he/she desires. It is a very powerful tool for Industries as there's almost anything you can do with a large data store with high efficiency.

6. Automating the Migration Process and Running transformations in Google Big Query

As discussed, there are numerous advantages of automating services in GCP for in-house use. Analytic companies have subject matter experts who spend a lot of cost and effort in dealing with these underlying steps. It also gives superior control to build an in-house application that's used for automating migrations and running transformations.

One important thing to consider here is the authentication with GCP. Every application that works or uses resources in the GCP needs to authenticate with GCP using one of the methods. The methods that can be used are :

1. Setting the Classpath variables: Setting the Google Environment Variables in the bashrc file. These variables are automatically detected by the Google Cloud Platform Client and there's no intervention required in this method.
2. Using a Service account: A service account is a way by which an application can authenticate by specifying the path to the JSON key that can be downloaded from GCP console in the IAM section.
3. Oauth: The most widely used authentication method. In this an access key and a secret key is used that can be found in the GCP console. These keys are then explicitly used to authenticate an access to GCP resources.

Another thing to consider is the need to synchronize between AWS S3 and GCP. The data from AWS S3 can be scheduled to transfer atomically or can be a repetitive task. There are again ways by which we can synchronise AWS S3 and GCP. For example, there's something called rclone or another thing called rsync. On running the command, the updated data in AWS S3 is transferred into GCS automatically and the latest versions of the data is maintained.

6. PHASES OF AUTOMATION

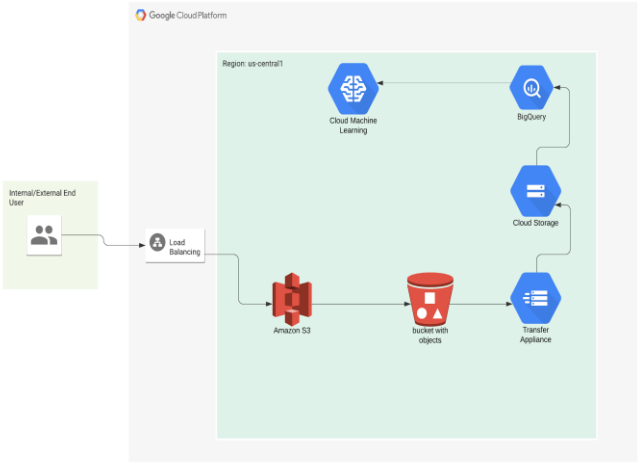

Fig. 1 Phases of Automation

1. **Build API Endpoints:** There are various frameworks that can be used to build API Endpoints. One widely used and lightweight framework is Dropwizard. Using Dropwizard, APIs can be created using Java and the Users will be able to use the application using these APIs.
2. **Load files from AWS S3 into GCS:** The transfer of data from AWS S3 to GCS is done using the transfer appliance of the GCP. The data can be transferred immediately or can be scheduled for any time of the day.
3. **Create Tables and Define Schema:** Create a table in GBQ for the data that has to be loaded from GCS into GBQ. This table can be created using the console, CLI or the client libraries in various programming languages. Each table is associated with a schema definition that needs to be specified while creating it. Tables can be created using code but you can use any other method. In case of Avro data, the schema need not be defined explicitly, the GBQ detects the schema automatically.
4. **Load GCS files into Tables in GBQ:** If you're doing it using code you'll need to create a BigQuery instance. Load files and keep the source as GCS. If the file schema adheres to the one specified, the file will be loaded successfully else it'll not be loaded. One additional thing that can be done here is that this can be an event-based action that'll be only triggered only when there's an update in the GCS data.
5. **Run Queries in GBQ:** The queries can be run using code, or the console and other methods can also be used. There are two types of SQLs supported in GCP - Legacy SQL and standard SQL. The queries can be run on the tables to bring about the transformations and make meaning out of the data.
6. **Build a model using AutoML:** Now that the data is available in the right form and the data is clean, well transformed and structured, it is well suitable for building a ML model on the data and leveraging the ML capabilities of GCP. AutoML proves to be very accurate in predictive models.
7. **Visualize:** Data analysis is nothing without visualization. Presenting the data in a from that is easily understandable and tells a story rather than just numbers is very important in the field of Analytics. Tools such as Tableau and PowerBI can be used and integrated with GCP to visualize the results.

7. CONCLUSIONS

Data-driven digital transformation strategies leverage analytics tools, and the power of reliable and timely data to provide new insights, creativity, and quicker decision-making. The effect is greater efficiency for the company and an improved competitive advantage. New cloud-based analytics technologies allow questions involving a broad range of data sources to be raised economically in forms which were never conceivable before.

Since the stack of GCP technology offers low-cost and high-performance services, it is well designed for a wide variety of use cases. It offers options for hosting your IICS Stable Agents on Google Compute Engine, use Google Networking to connect to your on-site network, store Google BigQuery data about products and service offerings In Google Cloud Storage, archive your files.

8. REFERENCES

[1] Sosinsky, Barrie. Cloud computing bible. Vol. 762. John Wiley & Sons, 2010.
[2] Armbrust, Michael, et al. "A view of cloud computing." Communications of the ACM 53.4 (2010): 50-58.
[3] Buyya, Rajkumar, Chee Shin Yeo, and Srikumar Venugopal. "Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities." High Performance Computing and Communications, 2008. HPCC'08. 10th IEEE International Conference on. Ieee, 2008
[4] Wang, Lizhe, et al. "Cloud computing: a perspective study." New Generation Computing 28.2 (2010):137-146.