Exploring Data Analysis Tools in Google Cloud Platform

Bala Keerthimai Akurathi

Master's in Information Technology, University of Cincinnati,akuratbi@mail.uc.edu

In this Modern age data has been evolving very rapidly. However, Humans cannot analyze this wide range of data. We need some machine to process analyze these kind of raw form of data and need to derive knowledge from it as it available in the market in unprocessed form. The data analysis tool will be more helpful to process this huge ocean of data. Various platforms like Amazon, Azure, Google offers different analytical tools The current study explores various tools that are necessary for data analysis in Google Cloud Platform by conducting semi-structured expert interviews from the sample made by my professor, gathered responses from the sample for deriving themes using thematic analysis by Braun and Clarke. The topics ended up being Big Query, Data Flow and Data Prep, with Big Query being the most popular data analysis tool provided by Google Cloud. However Google prioritize these tools in order to provide wide range of services for their end users and organizations. These data analysis tools will continue to be useful assets for all businesses as long as data remains essential and is used effectively for exploring and enhancing.

CCS CONCEPTS • Data • Analysis • Cloud• Tools• Challenges

Additional Keywords and Phrases: Big Query, Data Flow, Data Prep, Google Cloud Platform, Cloud Computing

1 INTRODUCTION

The data has become the most hazardous coinage in this era of New Media Age. There has been a huge lump of data generated by various platforms like social media, e-commerce applications, Healthcare etc. It is not the data in its unprocessed form itself but the knowledge that is gained from the raw data is responsible for its evolution and expansion. Analyzing this large amount of data plays an important role in every organization and this analysis is not limited to only IT but to all organizations [1]. In order to perform analysis on this huge ocean of data it's been a challenging task for all the industries. Traditional data analysis methods itself will not be sufficient to manage this massive amount of data. There come the modern tools that are required for data analysis especially offered by the Google cloud platform to drive through and extracting the knowledge from the data [2]. These data analysis techniques have a significant societal influence as they control some of the environmental conditions, aid in disaster management, and reduce the transmission of illnesses during times of public health crisis etc.

This part of study answers my research question what are different analysis tools offered by Google cloud platform. There are many data analysis tools in GCP like Big Query, Data Flow, and Data Prep etc., where Big Query is one of the data analysis tools with a significant social impact that is used to make quick business choices and also offers increased security and privacy [3]. It manages very large volumes of data and also offers strong economic development that supports innovation, enabling businesses to use these technologies to create new goods and services [4].In order to transform and move the data for the purpose of data processing there exists a tool called Data Flow in Cloud where it follow the concept of pipelines. Data engineers use these pipelines for analyzing the data to perform business applications such as OLAP [5]. The accurate and reliable data can be obtained after processing and our data may contain some errors, missing values and many more and so as to deal with and recognize such type of data there comes the tool in cloud called Data Prep that provides quality to our data in unstructured format [6].There are also other tools that are helpful for analysis like data proc, studio, looker and many more. These tools will be a valuable addition to all the organization for the better outcomes.

It will be a difficult undertaking for Google Cloud to analyze this enormous volume of data because of the various difficulties it faces. Big Query performs best when used with Google's infrastructure and is limited to a set number of requests with a maximum 10MB return size per request. It cannot be utilized with other infrastructures in this situation, and analysis will be challenging. Similar problems arise when data transformation occurs through pipelines, and when we apply the notion of data flow, there may be connection timeouts, which makes it challenging to fix and discover redundant data. The majority of businesses value data, and performance should be consistent even in the presence of vast volumes of data for analysis and they also need more protection and safety over their data [7].

Prior to the cloud era, the allocation of a wide variety of resources for data storage and analysis was both complex and expensive. After the dawn of the cloud era, services like cloud computing, which allows us to rent resources according to our usage, and services like Google Cloud Platform (GCP) are making a significant effect and are extremely accessible and secure, making it simple and scalable to analyze data [8][9]. Many Researchers have explored these services so far to provide better solutions for innovation and success of their business. GCP in data analysis ensures to fully utilize the data to any extent. In conclusion to introduction I have learnt various organizations use different data analysis tools not only offered by Google but also other cloud service providers like AWS and Azure for keeping their data more secure and there may be limitations on the size of data they need to be analyzed[10].

1.1 Research Question

What are the different data analysis tools used in Google Cloud Platform?

2 METHODOLOGY

I would like to know about the different kinds of analysis tools that can be used mostly by Google cloud platform and to answer my research question, I used expert interview research methodology which is a both semi structured and quantitative to the study where the professor has made the group of 6 members randomly along with me and I have interviewed the students of my University of the same class. To get the results for my question I interviewed them where they also have some sound knowledge about the cloud and tools used for data analysis as they also come from Information technology background. For this procedure, I have chosen random sampling as my sample technique. As ACM and IEEE publications offer high levels of reliability, I used them as references for a deeper knowledge of the content of my data analytic tools that can be used in cloud by Google.

2.1 Sample

The sample size for the study is n=5 samples that was made by my professor randomly with fellow students of my University from the class in SoIT. In order to find out the answer to my research question and for further study from the perspective of the experts, the six samples, including me, were linked through teams from October 4 - October 6 in-person as well as through teams call and all the members are thoughtfully included and responded well in all the interviews. Responses of each sample is saved and compiled in workbooks. The study has been done carefully by not gathering any personal information from the sample and ensured that their responses are kept safe. Every sample has been utilized to the maximum extent possible.

2.2 Measure

For the purpose of my study for different analysis tool in Google cloud platform I have designed a set of 9 questions to do interview for my group allocated by my professor where 2 questions out of them are to understand about the background related to the study and remaining questions aims to derive the response for my study. The goal of this interview is to get the responses from the experts that are necessary for my study.

2.3 Design

I have designed all the interview questions that are open-ended and wrote them in the last section of the paper .They are taken from the existing literature where I took reference from the ACM and IEEE articles after going through the papers. I preferred these papers because they provide good quality and accurate results that are very essential for my study. The experts whom I interviewed have a sound knowledge about my study and they will provide the required information based on their expertise in the field of exploring tools in data analysis in cloud. The questions are written such that they include all the essential information required for the study, and the answers are meant to be brief and straightforward.

2.4 Procedure

In order to get responses for my study I have reached my experts for the interview through teams and scheduled the in person interview at our own free times. Out of 5 samples 2 samples are connected by teams call and three of them met in-person for face-to face interview. I wrote few responses in notes and recorded few of them with their permission and saved in workbook. I have ensured all the privacy concerns and kept the information safe and secure. After collecting all the responses from the experts I will be using thematic analysis by Braun and Clarke that will be helpful for my study for exploring the analysis tools in GCP and illustrated my responses in Table 1 [11].

Table 1: Thematic Analysis for exploring data analysis Tools in GCP

Themes	Code	Illustrative Quotations
Big Query (80%,N=4)	Data Storage (60%,N=3)	“Big Query is one of the most widely used fully qualified data analysis tool that is highly scalable. “ ,”It is server-less machine and it is incorporated with built-in engine of the data analysis.”, ”Big Query is an analysis tool and server-less and it has the capacity to handle and perform analysis on large amounts of data.”
	Increases security and privacy (20%,N=1)	”Big Query is one of top analysis tool that provides highly secure environment to the end users.”
Data Flow (60%,N=3)	Pipeline (40%,N=2)	”Data Flow Uses Pipelines in order to process the real time data and streamline data.”, ”The concept of pipelines is used for the dataflow and it also uses a framework Apache Beam in order to represent the data in the form of pipeline.”
	Data Source (20%,N=1)	“Data Flow is one of the most commonly used data processing tool. It handles both the actual and streamline data.“
Data Prep (40%,N=2)	Data Cleaning (20%,N=1)	“Data prep is used as a data visualization tool that is used for knowing about the data and used to clean the data and make it ready for analysis.”
	Data Transformation (20%,N=1)	”Data Prep is a widely used tool for the data that has to be readily prepared and converted to other forms it accepts both structured and unstructured form of the data.”

3 RESULTS

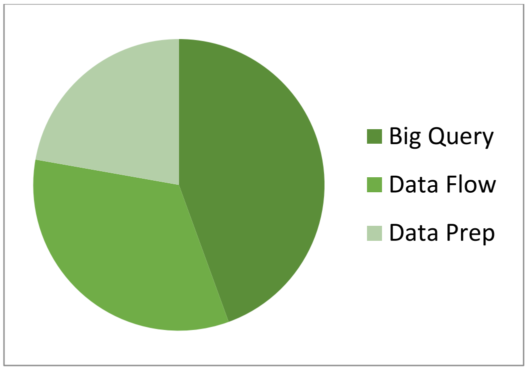
I saved all the responses that I got from interviewing the experts and made sure that the responses are secure and the data is not missing. I have cross checked all my responses before compiling them into results. In order to take my study forward I will be applying thematic analysis to my study of exploring the data analysis tools in Google Cloud Platform to make it more efficient and identify themes from the data. Here themes are derived in terms of frequency where frequency of a theme is determined by how frequently it has been addressed by experts and they are represented in the Table 2.

Table 2: Themes

Theme	Frequency
Big Query	4
Data Flow	3
Data Prep	2

After following from thematic analysis, I identified three themes they are Big Query, Data Flow, Data Prep where frequency of Big Query is more where 80% of the sample talked about it and followed by Data Flow 60% and then Data Prep 40%. As one of the most widely used data analysis tools on the Google Cloud Platform that provides additional privacy and safety, Big Query was frequently mentioned in conversation. Data Flow is recognized for its data processing nature through pipelines and ranks second after Big Query in data analysis tools in Google Cloud Platform that are mentioned by few people. Data Prep makes the data clean and helps in visualization purposes and holds third place in data analysis tools in Google Cloud Platform and is least used. In order to provide a clear explanation these themes and frequencies are illustrated in the form of pie chart in the below Fig 1.

Fig 1 Theme and their Frequencies



4 DISCUSSIONS

As per the results I have got the highest frequency for Big Query because this is one of the server-less and analytical services used by Google cloud that mainly includes querying large data sets and going through ETL process on large data sets[1]. The reasons why the experts that I interviewed have inclined towards this are because from an analytical point of view it also provides security as high priority [3] [4]. It also takes the efficient management of cost into account and makes the infrastructure more cost efficient. Secondly, it also offers a service of paying for resources for what we use, and also delivers analysis of actual data and helps to consolidate with various types of tools used in data analytics.

Accordingly, the next highest frequency I got is for Cloud Data Flow where it is one of the most fully qualified services under cloud in Google that is used for processing the data and also permits for building and developing the pipelines for both static and dynamic data [5]. Moreover it makes the task easier for the data having complex problems, automatically scales resources based on demand and clear the way for analyzing the data. Other than this it also provides a platform for the users in order to transform the data in the form of workflow using one of the models in data processing like Apache Beam.

Finally, I got the least frequency for Cloud Data Prep where it is a data conversion and preparation tool in the cloud that is offered by Google and it provides a platform for visually representing the data without the need of coding in order to purify the data and enhance it .Besides Big Query and Data Flow it stands out in third place because of some drawbacks where it can't prioritize the security and privacy concerns like big query and also cannot solve complex data problems like dataflow. It is only used for preprocessing of stream data [6] [7].

In my study the sample size consists of only five people out of which one sample has no knowledge on study and yet able to put some abilities to know about the study and four of them have the basic knowledge of the study and put their efforts to respond and provide answers about their views of exploring analysis tools in cloud by Google .So there is drawback in selection of the sample process for semi structured expert interview methodology. The consequences of my study may miss in exploring the other data tools for analysis in Google cloud that offers good services to the end users. Therefore, if my study includes more samples, I may have discovered more themes that will be beneficial to my research and it may not have an impact on the results.

I have followed all the ethics and policies that are provided by my university and I did not face any ethical issues while I am doing this study.

REFERENCES

[1] N. Naik, "Connecting google cloud system with organizational systems for effortless data analysis by anyone, anytime, anywhere," 2016 IEEE International Symposium on Systems Engineering (ISSE), Edinburgh, UK, 2016, pp. 1-6, doi: 10.1109/SysEng.2016.7753150.

[2] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. The KDD process for extracting useful knowledge from volumes of data. Commun. ACM 39, 11 (Nov. 1996), 27–34. https://doi.org/10.1145/240455.240464.

[3] B. Kotecha and H. Joshiyara, "Handling Non-Relational Databases on Big Query with Scheduling Approach and Performance Analysis," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697561.

[4] Z. Trencheva, P. Latkoski, M. Porjazoski and B. Popovski, "Cloud-based big data analysis for OTT video services," 2023 30th International Conference on Systems, Signals and Image Processing (IWSSIP), Ohrid, North Macedonia, 2023, pp. 1-5, doi: 10.1109/IWSSIP58668.2023.10180262.

[5] Haotian Gao, Cong Yue, Tien Tuan Anh Dinh, Zhiyong Huang, and Beng Chin Ooi. 2023. Enabling Secure and Efficient Data Analytics Pipeline Evolution with Trusted Execution Environment. Proc. VLDB Endow. 16, 10 (June 2023), 2485–2498. https://doi.org/10.14778/3603581.3603589.

[6] Kun-Ta Chuang, Hung-Leng Chen, and Ming-Syan Chen. 2009. Feature-preserved sampling over streaming data. ACM Trans. Knowl. Discov. Data 2, 4, Article 15 (January 2009), 45 pages. https://doi.org/10.1145/1460797.1460798

[7] Bhavani Thuraisingham. 2015. Big Data Security and Privacy. In Proceedings of the 5th ACM Conference on Data and Application Security and Privacy (CODASPY '15). Association for Computing Machinery, New York, NY, USA, 279–280. https://doi.org/10.1145/2699026.2699136.

[8] Yunqi Kan. 2021. A Cloud Computing Resource Optimal Allocation Scheme Based on Data Correlation Analysis. In Proceedings of the 4th International Conference on Electronics, Communications and Control Engineering (ICECC '21). Association for Computing Machinery, New York, NY, USA, 26–31. https://doi.org/10.1145/3462676.3462681.

[9] Nicholas J. Mitchell and Kazi Zunnurhain. 2019. Google cloud platform security. In Proceedings of the 4th ACM/IEEE Symposium on Edge Computing (SEC '19). Association for Computing Machinery, New York, NY, USA, 319–322. https://doi.org/10.1145/3318216.3363371.

[10] Gabriel Costa Silva, Reginaldo Ré, and Marco Aurélio Graciotto Silva. 2018. Evaluating efficiency, effectiveness and satisfaction of AWS and azure from the perspective of cloud beginners. In Proceedings of the 28th Annual International Conference on Computer Science and Software Engineering (CASCON '18). IBM Corp., USA, 114–125.

[11] Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101.

A APPENDICES

I have some of the interview questions below for conducting expert interviews:

Hope everyone is doing well. My name is Bala Keerthimai Akurathi. I have been doing expert interviews for my study of the different analysis tools used by the Google Cloud Platform, and here are some interview questions I have been asking them. Having a conversation about these questions will assist me with my work and allow me to utilize the responses for my study.

A.1 Interview Questions

1. Can you tell me something about Google cloud platform?
2. Can you briefly explain whether you have any prior knowledge in Google cloud platform for data analysis?
3. List some data analysis tools that you know in GCP?
4. Explain briefly some of the advantages of any two analysis tools in GCP?
5. Why Big Query tops the list among all the data analysis tools in GCP?
6. Give some details about the type of data that can be used for analysis?
7. Tell me something about pipelines in GCP?
8. How Data flow is useful in data transformation and ETL process?
9. What is the role Data Prep in GCP?