North American Journal of Engineering and Research

Est. 2020

Implementing data warehousing solution in Google Cloud Using Big Query

Venkata Soma
Email: binay.svrs@gmail.com

Abstract

This study focused on assessing the effect of implementing a data warehousing solution in Google Cloud using BigQuery. This study shows the BigQuery architecture, which is based on the separation of storage and processing. The data is kept in a replicated, dependable, and distributed storage system, with elastic distributed compute nodes handling data input and processing. BigQuery proved to be a dependable, user-friendly, and customisable platform for extracting structured data from blockchains. The analysis yielded results in minutes to hours. BigQuery offers significant SQL capabilities, which will help the NY Mets or any other sports business to keep their secret information.

Keywords: BigQuery, Google Cloud, SQL, warehousing solution, data management, data security

Introduction

Project specification

Data warehousing is the process of gathering, storing, and managing massive amounts of organised and unstructured data to provide organisations with a centralised repository for all critical information. This allows firms to do complicated data analytics and get useful insights that lead to more informed decision-making. Organisations may optimise data management, improve internal operations, and respond more effectively to market needs by implementing a complete data warehouse solution. There are various data warehouse systems on the market, each with its own set of features and capabilities. These systems differ in terms of design, performance, scalability, usability, data source support, data integration, and analytics capabilities. Businesses must examine their size, industry, data requirements, and available resources while looking for the best data warehouse solution.

Aim and Objectives

This study aims at assessing the effectiveness of implementing data warehouse solution in Google Cloud, by using BigQuery.

Objectives

The objectives are following:

- To understand BigQuery
- To assess effectiveness of BigQuery in increase data visualisation
- To assess difference between BigQuery than other technologies

Research question

- What is BigQuery?
- How does BigQuery is effective to increase data visualisation?
- How does BigQuery is different from other technologies?

Research Rationale

The data is stored in a replicated, reliable, and distributed storage system, with elastic distributed computing nodes handling data input and processing. Furthermore, BigQuery has a distinct shuffle service built on top of "disaggregated distributed memory", which improves interaction among compute nodes and allows query checkpointing for dynamic preoptimization [1]. These services connect the system to

deliver a business data warehouse solution. Such a solution will be beneficial for the sports industry; all organisations in this industry will be able to improve their data management services through a control plane and foster task management.

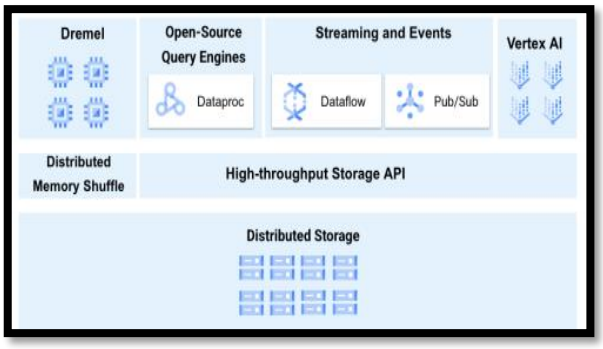

Figure 1: High-Level BigQuery Architecture [1]

Literature review

Research background

Google BigQuery stands out for its real-time analytics and ease of usage. BigQuery is a fully managed, serverless computing data warehouse that frees enterprises from infrastructure administration, allowing them to focus on obtaining data insights. Furthermore, BigQuery is notable for its inbuilt machine learning (ML) capabilities [2]. Subscription model, monitoring expenditures can be difficult for organisations that conduct extensive data analytics activities. Users may utilise "BigQuery Analytics Hub" to securely communicate data assets inside and between organisations, establish and manage data storage spaces, and improve analysis using commercial, publicly traded, and Google datasets.

Critical Assessment

Google Cloud Platform provides BigQuery machine learning, a cloud-based service. Data analysts and scientists may use SQL queries to construct and deploy machine learning models on large datasets, eliminating the need for a considerable understanding of enables data analysts to develop, train, assess, and forecast ML models using their existing SQL expertise [3]. It allows a sports organisation to develop powerful and extensible ML models while remaining in an individual Google Cloud service. The service includes a number of pre-built machine learning models, including "logistic regression", "linear regression", "k-means clustering", and "time-series forecasting", that may be applied to data using simple SQL commands. Users may also use "TensorFlow" or "Keras" to build bespoke models and incorporate them into "SQL queries" using BigQuery ML.

BigQuery is a distributed metadata management solution that solves scaling issues. By approaching metadata management similarly to data management, an organisation created a system capable of storing rich metadata, scaling to huge tables, and giving fast access from the query engine. BigQuery allows for scalable analytics over a petabyte of data. The BigQuery design relies on separating storage and computation. Data is stored in a distributed storage system, while elastic computing nodes handle intake and processing. BigQuery offers a shuffle service based on distributed memory. The Shuffle service allows for communication between computing nodes. Horizontal services like APIs, metadata, and security bind the system together. BigQuery allows for dynamic physical query strategies. The Query Coordinator creates an initial plan, but as the query executes, the plan adjusts based on data statistics. Statistics include data flow between stages, table row count, distribution, and skew. Statistics impact stage parallelism and physical operator choices, such as shuffled vs. broadcast join.

Linkage to aim

The emphasis of the study [4] is on data warehouses tailored for the cloud. Its analysis focused on cutting computational power for complicated analytical queries. GCP's BigQuery is a server-free data warehouse that specialises in quickly processing massive datasets. Its strengths include machine learning and simplicity of use for sophisticated data searches. With such discussion, this review section is meeting the aim and objectives of this study.

Theoretical background

Google BigQuery is chosen as the CDW solution due to its cost-effectiveness, server lessness, machine learning capabilities, and ability to expand across many clouds [5]. “Information Asymmetry theory” is maintained maintain a symmetry in the information management process. BigQuery follows this theory to enhance data management processes for efficient stakeholder management. The BigQuery platform allows for the creation of cloud storage buckets in the Google Cloud Platform, where data can be uploaded directly to AdventureWorksDW for extraction. However, due to license limitations, this feature has not been tested. Additionally, it allows for the creation of tables from Google Big Table, Amazon S3, and Azure Blob Storage. SQL may be used for transformation in BigQuery projects. Additionally, the BigQuery API may be used with Python on Google Colab or

other platforms to import data into BigQuery. Hence, it can be said that the NY Mets' use of the Google Cloud system is crucial for developing its data security management.

Literature gap

This review mentions what BigQuery is; however, it does not have adequate information on the way BigQuery can serve the sport industry and each sports organisation. Hence, urgent research is required to assess the effectiveness and efficiency of data warehouse management in Google Cloud by using BigQuery for the sports industry and information security management.

Methodology

Research philosophy

This research process followed interpretivism philosophy to interpret the effectiveness of data warehouse solution in Google cloud through BigQuery. This philosophy interprets all information and assesses those to explore an in-depth insight regarding the subject matter.

Research approach

The current research process followed an inductive approach in which moves from specific information to general information. Hence, this approach analysed from the effectiveness of BigQuery application to its efficiency in the overall and general data management.

Research design

This study used a secondary qualitative design; the gathered secondary data is analysed with qualitative and thematic processes. Qualitative design promotes descriptive analysis which interpret gathered information to generate useful insight about the effectiveness of BigQuery in the data warehouse solution.

Data collection method

The current study gathered secondary data from electric database, such as Google Scholar and ProQuest. This process maintained a 5-year time frame for data collection, that is from 2013 to 2018. Hence, most current information is gathered.

Ethical consideration

This research paper maintained ethical approaches as this mentioned all the names of authors and publishing year properly. This avoided plagiarism issue and is used only for academic purpose.

Results

Critical approach

Without adequate technology and infrastructure, storing and analysing massive data volumes may be time-consuming and costly. BigQuery, an enterprise data warehouse, addresses this issue by utilising Google's infrastructure to speed up SQL queries. Transfer the data to BigQuery for automatic handling. Access to projects and data may be regulated depending on necessity, allowing others to see and query data in the data warehouse. BigQuery may be accessed by the "Cloud Console", the "bq command-line tool", or the "BigQuery REST API" via Java, .NET, or "Python client libraries" [6]. BigQuery may be integrated with third-party solutions for data visualisation and loading. To begin using BigQuery, no resources, including discs and virtual computers, are required. There are five key steps: "Collect NoSQL documents", "prepare them", "store them in Cloud Storage", "extract and convert schemas", and "Load data" from "App Engine" to "BigQuery data warehouse".

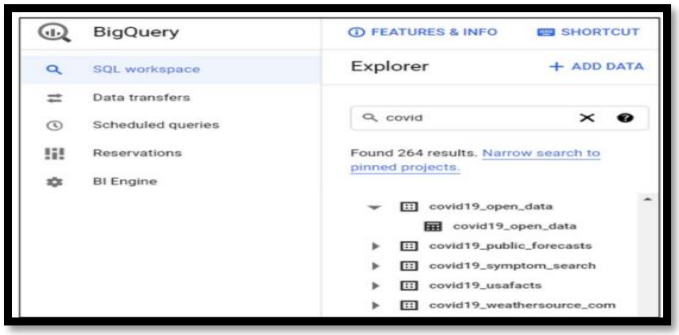

Figure 2: Interface of BigQuery [6]

Findings and result

Theme 1: Cloud management through BigQuery

Google Cloud Platform offers services for large data analytics and processing. This article covers the architecture and components of Biguery, a popular large. This PaaS allows for ANSI SQL querying. Machine learning capabilities are also included. Since its inception in 2011, it has grown in popularity and is widely used by major corporations for data analytics. BigQuery has an easy-to-use interface that may be accessed in several ways based on user needs. To utilise this tool, just use the graphical online interface (Figure 2). The use of a cloud console or Bigquery APIs is a somewhat more complex but quicker solution. The Bigquery online interface (Figure 2) allows users to add or choose datasets, schedule and perform searches, transfer data, and view results.

"Google BigQuery" automatically encrypts data with Google-managed keys. Customer-managed encryption keys allow organisations to safeguard their data, although this was not

tested owing to license limits. "Google BigQuery" offers various data protection and governance features, such as "Identity and Access Management" (IAM) for resource security, column and row level data security, cloud data loss prevention and cataloguing, encryption, inspection, auditing, transaction processing, and logging [7]. Hence, it will support the documents of the NY Mets; the customer-managed encryption keys will benefit this organisation in understanding the requirements of its target customers while organising the sports. Row-level data security will also improve the data security system for all information and credentials of the athletes, coaches and other stakeholders.

Theme 2: Data storage and internal data management through BigQuery

End the era of digitalisation, with data created from multiple online and offline sources every second. Big data refers to large amounts of data with diverse qualities. Traditional methods for storing, processing, analysing, visualising, and extracting valuable information from Big Data on local devices are problematic. A cloud computing platform is used to tackle this problem. Cloud computing provides powerful processing, storage, and apps that are independent of user device performance. Many users can access resources and services from the cloud remotely on a pay-as-you-go basis [8]. This eliminates the need for users to purchase and install costly resources locally. BigQuery is a cloud-based platform for completely regulated data. The warehouse enables fast processing of large volumes of data, comparable to Google's performance.

Figure 3: Different sources of data, integrated into BigQuery [8]

Using Google's affordable price world-class scalability and security architecture gives robust business analytics. BigQuery is a rapid and petabyte-scale data warehouse solution for BigData analysis. BigQery recommends using SQL to query, represent, and analyse BigData without the need for infrastructure or database administrators [9]. Most institutions and corporate organisations, including startups and Fortune 500 businesses, utilise it. Figure 3 depicts the data sources incorporated into BigQuery.

Theme 3: Data visualisation through BigQuery Data is visualised utilising BigQuery platforms. BigQuery systems provide for easier data processing and visualisation. BigQuery may help with the evaluation of sales data for commercial items. Furthermore, big data and data mining can give precise insights into health data. The data undergoes processing using BigQuery systems. Using BigQuery platforms may make data analysis and visualisation easier [9]. To get an idea of the volume of information, descriptive analysis is used on the data. After processing the data, BigQuery is used to visualise it. Data processing and visualisation make it easier to make scholarly decisions. A bar graph and an area map are two data visualisation types utilised in studies. Hence, it can be said that BigQuery can enhance data visuality about the players' performances. To provide sales insight, BigQuery can visualise the sale of tickets and the number of audiences. In fact, it can help the NY Mets assess the contribution of the investors as well.

GCP's BigQuery environment was utilised for data processing and analysis in the next two stages, while Data Studio was used for data visualisation. BigQuery is a cloud database capable of processing enormous volumes of data from diverse sources, addressing database administration issues and facilitating large-scale data analysis. This system uses SQL-like queries to examine enormous volumes of data in real time. A new project is established to begin processing data using BigQuery, which generates a dataset. To create a new dataset, pick the specified territory as the data array's location. To upload data to BigQuery, save it.

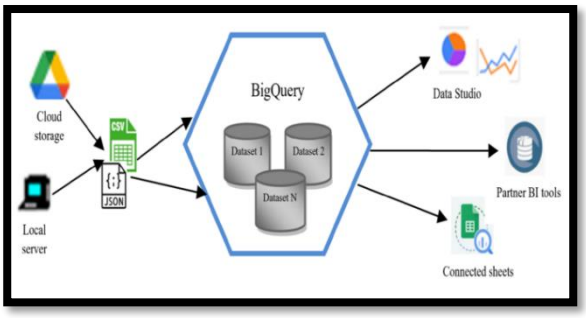

Figure 4: The process of big data processing in BigQuery [10]

Theme 4: Differences between BigQuery and other technologies

Blockchain technologies have grown in popularity since they offer autonomous asset transactions and smart contract capability. Bitcoin blocks have been used to store various information, including words and pictures, since their inception. Onchain's data storage features are beneficial for a range of valid uses. It does, however, carry a systemic danger. Abuse of data storage capability on public blockchains, such as publishing illicit information, might result in legal implications for operators and users. This can jeopardise the operational integrity of whole ecosystems [10]. BigQuery was shown to be a reliable, user-friendly, and adaptable platform for extracting structured data from blockchains. Results were quickly produced, ranging from minutes to hours, based on the analysis.

Conclusion

Data warehousing is the collection, storage, and management of enormous volumes of organised and unstructured data in order to provide businesses with a single repository for all vital information. This enables businesses to do complex data analytics and get relevant insights, resulting in better-informed decision-making. Google BigQuery distinguishes itself with its real-time analytics and simplicity of use. BigQuery is a fully managed, serverless computing data warehouse that relieves organisations of infrastructure management, allowing them to concentrate on data analysis. High-level BigQuery architecture will assist sports organisations in maintaining their confidential information of their athletes, their stats, coaches, and other stakeholders.

Future work

This study provides an opportunity for all organisations in the sports industry to assess the benefits of BigQuery. Further study can be conducted on the result of using BigQuery in sports, as well as other industry.

Research recommendation

BigQuery provides extensive SQL functionality; such functionality will allow NY Mets or any other sports organisation to maintain their confidential information. However, further approaches are required to handle and analyse the data, particularly for the file outcomes. The study also has reflected problems with subqueries and regular expressions. BigQuery confines regular expression patterns to only one capturing group, making pattern matching problematic for URLs. It provides effective data filtering for validation and processing by other applications in the sports industry by increased data visualisation.

References

[1] S. Saif, and S., Wazir. “Performance analysis of big data and cloud computing techniques: a survey.” Procedia computer science, 132, pp.118-127. 2018. https://www.sciencedirect.com/science/article/pii/S1877050918309062

[2] A., Sebaa, F., Chikh, A. Nouicer, and A., Tari. “Research in big data warehousing using Hadoop.” Journal of Information Systems Engineering & Management, Vol 2 No. 2, p.10.2017. https://elk.adalidda.com/2017/06/research-in-big-data-warehousing-using-hadoop.pdf

[3] T., Dokeroglu, S.A. Sert, and M.S. “Cinar. Evolutionary multiobjective query workload optimization of Cloud data warehouses.” The Scientific World Journal, 2014(1), p.435254. 2014. https://onlinelibrary.wiley.com/doi/abs/10.1155/2014/435254

[4] A., Gupta, F., Yang, J., Govig, A., Kirsch, K., Chan, K., Lai, S., Wu, S., Dhoot, A.R., Kumar, A. Agiwal, and S., Bhansali. “Mesa: A geo-replicated online data warehouse for Google's advertising system.” Communications of the ACM, Vol 59 No. 7, pp.117-125. 2016. https://link.springer.com/chapter/10.1007/978-3-319-65930-5_1

[5] L., Antova, D., Bryant, T., Cao, M., Duller, M.A. Soliman, and F.M., Waas. “Rapid adoption of cloud data warehouse technology using Datometry Hyper-Q. In Proceedings of the 2018 International Conference on Management of Data (pp. 825-839). 2018, May. https://dl.acm.org/doi/abs/10.1145/3183713.3190652

[6] M. Bahrami, and M., Singhal. “The role of cloud computing architecture in big data.” Information granularity, big data, and computational intelligence, pp.275-295. 2015. https://link.springer.com/chapter/10.1007/978-3-319-08254-7_13

[7] T.N., Hewage, M.N., Halgamuge, A. Syed, and G., Ekici. “Big data techniques of Google, Amazon, Facebook and Twitter.” J. Commun., Vol 13 No 2, pp.94-100. 2018. https://www.researchgate.net/profile/Malka-Halgamuge/publication/323588192_Review_Big_Data_Techniques_of_Google_Amazon_Facebook_and_Twitter/links/5b89eddf4585151fd1403fa3/Review-Big-Data-Techniques-of-Google-Amazon-Facebook-and-Twitter.pdf

[8] R., Sahal, M., Nihad, M.H. Khafagy, and F.A., Omara. “iHOME: index-based JOIN query optimization for limited big data storage.” Journal of Grid Computing, 16, pp.345-380. 2018. https://link.springer.com/article/10.1007/s10723-018-9431-9

[9] V.R., Vadiyala, P.R. Baddam, and S., Kaluvakuri. "Demystifying Google Cloud: A Comprehensive Review of Cloud Computing Services." Asian Journal of Applied Science and Engineering, 5(1), pp.207-218. 2016. https://www.researchgate.net/profile/Swathi-Kaluvakuri/publication/376841685_Demystifying_Google_Cloud_A_Comprehensive_Review_of_Cloud_Computing_Services/links/6596d6290bb2c7472b32b95c/Demystifying-Google-Cloud-A-Comprehensive-Review-of-Cloud-Computing-Services.pdf

[10] S.A., El-Seoud, H.F., El-Sofany, M. Abdelfattah, and R., Mohamed. "Big Data and Cloud Computing: Trends and Challenges." International Journal of Interactive Mobile Technologies, Vol 11 No 2. 2017. https://www.academia.edu/download/97190135/4356.pdf