North American Journal of Engineering and Research

Est. 2020

Comparative Study of Big Query, Redshift, and Snowflake

Venkata Soma
Email: binay.svrs@gmail.com

Abstract

In the present era of the intense utilisation of big data, effective data warehousing solutions are necessary for the management and analysis of larger volumes of structured and unstructured data. The rapid emergence of cloud computing has transformed the data warehouse inclusion techniques, cost-effectiveness and scalability over the traditional warehousing solutions. This research paper examines the evolution of data warehousing within cloud platforms by emphasizing the cost implications and performance metrics. The well-known cloud-based solutions such as "Amazon Redshift", "Snowflake" and "Google Big Query" are analysed within this paper focusing on their query performance, architecture and integration abilities. This paper also highlights the potential issues related to cloud computing. The findings of this research underline the transformative influence of cloud data warehousing on the decision-making process of an organisation

Keywords: Data warehousing, cloud computing, performance analysis, scalability, cost-efficiency, cloud platforms**.**

Introduction

Project specification

This research paper tries to focus on the critical evolution of data warehousing solutions in the cloud platforms. This emphasis is on the specification of the performance analysis for warehouse solutions and the cost implications. Data warehouse serves as repositories for structured as well as unstructured data which enables businesses to extract valuable insights which drive strategic decisions. The increase in cloud computing has transformed this landscape through providing “platform as a service” (PaaS) and “infrastructure as a service” (IaaS). This research paper aims to evaluate the performance and cost implications related to cloud platforms [1]. The proposed research seeks to identify these issues through the conduction of a holistic analysis of data warehousing solutions in cloud platforms.

Aim and objectives

Aim

This research aims to interpret the comparative analysis of Big Query, Redshift and Snowflake.

Objectives

- To provide a comparative study about the Big Query, Snowflake and Redshift
- To investigate the best warehouse solutions among these three warehouse solution tools

Research Questions

RQ 1: What are the main features of Big Query, Snowflake and Redshift?
RQ 2: Which one is suitable most as the best warehouse solution among Big Query, Snowflake and Redshift?

Research Rationale

Data warehousing solutions assist in storing and managing large volumes of structured along unstructured data [2]. This delivers a centralised repository for analytics, reporting and business intelligence. The emergence of cloud computing has altered the process of deployment and management of data warehouses. This offers scalability, cost-effectiveness and flexibility which is attainable with traditional assumption-based solutions.

Literature Review

Research background

Data warehouse plays a critical role in modern business intelligence and analytics strategies which serve as key repositories for structured and unstructured data. The rapid emergence of cloud computing has reshaped data warehousing settings by maintaining cost-effectiveness and flexibility compared to traditional solutions. Despite the widespread adoption of cloud-based data warehouses (DWH), various organisations face substantial issues in understanding and evaluating the performance.

Critical Assessment

Privacy: Sometimes the user data may be accessed through the organised company with or without their permission. The service providers can access the data of the cloud at any point in time which underlines the privacy issues [3]. This can lead to the accidental or deliberate alteration or removal of the information.

Security: The cloud-based services include a third party for securing and storing necessary data. The third-party incorporation leads to significant concerns about the sharing of confidential data or information of the users with third parties which underscores the significant security concerns in cloud-based computing.

Compliance: There exist various regulations associated with data security and hosting. In order to secure compliance with regulations such as the "Federal Information Security Management Act", "Insurance Portability and Accountability Act" and many more required the adoption of deployment modes which is expensive.

Abuse: While offering cloud services, it is necessary to ensure that the clients do not use cloud computing for diverse purposes. In the year 2009, a banking Trojan utilised Amazon's popular services for controlling the infected personal computers and problems regarding software updates.

Sustainability: This sustainability problem refers to the minimisation of the impact of cloud computing on the environment. Countries such as Sweden, Finland and Switzerland leverage the favourable climate and generous renewable energy to attract the data centres of cloud computing.

Recovery of lost data: It is necessary to review all the terms and conditions before signing up for any type of cloud service [4]. This can further assist in ensuring the meeting of user requirements and designing of well-maintained infrastructure with essential resources. After the subscription, it is essential to entrust the user data to any third party.

Higher level of cost: Utilisation of cloud services requires a powerful network with a higher level of data transfer rate than traditional internet networks [5]. A user can face issues in the utilisation of ordinary cloud services while working on complicated projects.

Linkage with aim

The proposed research explores emerging trends and technologies in cloud data warehousing such as integration with advanced analytics instruments and serverless architects. Consideration of the performance such as data ingestion rate, query response times and system reliability altered across distinct cloud platforms.

Encapsulation of application

Cloud solutions assist in maintaining scalability along with rapidity which is crucial for organisations to navigate the complications of modern data management. The cloud solutions introduce various considerations which include query performance, data latency and diversified cost structures depending on the usage patterns. In the present realm of evolved cloud data warehouses, challenges and innovation arise continuously. In order to observe the market demands, the main data warehouse platforms play a critical role in maintaining data competencies within the cloud platforms.

Methodology

Research Philosophy

The research is associated with the comparative analysis of the Big Query, Amazon Redshift and Snowflake, especially within the sports industry. The research will be utilised the interpretivism philosophy to emphasize the perspectives of researchers on this topic. It will explore the opinions of the developers, users, and administrators about the significance of the different warehouse solutions. Interpretivism encompasses social theories and perspectives that include a view of reality as socially constructed.

Research Approach

The research is associated with the different warehouse solutions in the management within the sports sectors. This research includes the deductive approach for the investigation of the efficiency of the integration of these three warehouse solutions in cloud computing. Through the incorporation of the deductive method, this project will provide the insights of previously working individuals through data collection along with data analysis.

Research design

To collect and analyse the data about the performance of Big Query, Saltstack and Amazon Redshift within the sports industry, the secondary qualitative method has been utilised. This will assist in providing an overview of the development and deployment of Redshift, Saltstack and Big Query within cloud computing.

Data collection method

The data collection process will be performed through the peer review of previously published scholarly articles, and journals accessed through Google Scholar and other sites.

The gathered information will be analysed using thematic analysis.

Ethical considerations

The maintenance of the ethical perspectives is crucial for this study. Privacy and authorisation laws must be followed at the time of using confidential information about sports events. The privacy of the sensitive data within cloud computing must be safeguarded to maintain ethical resilience.

Results

Critical Analysis

"Amazon Redshift"

Amazon Redshift is a data warehouse service offered as part of "Amazon Web Services" (AWS). This delivers a straightforward and cost-effective solution for the analysis of user data across both data lakes and on-promise data warehousing. This poses the ability to provide performance at a ten times faster rate than the traditional solutions. The Amazon Redshift uses machine learning and columnar storage on high-performing disks [6]. Users can be able to set up and deploy the new adapt warehouse quickly with the help of Amazon Redshift. It fosters querying across the exabytes of data stored in Amazon S3 data lakes and petabytes of the data stored in its data warehouse.

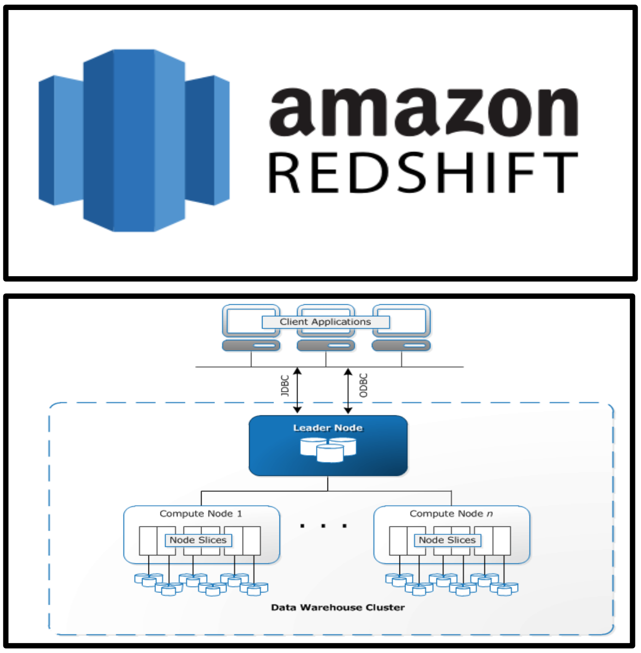

Figure 1: "Amazon Redshift Architecture" [6]

The architecture of Amazon Redshift includes the following structure:

Redshift Cluster: In the Amazon Redshift, a cluster is constructed through several nodes which create its core infrastructure. The cluster includes multiple computer nodes and one leader node.

Leader Node: The leader node manages all the communications with the application of the client. This collaborates with the computer nodes to undertake tasks such as query parsing, distribution of the compiled code to the compute nodes and development of execution plans.

Compute Node: Each of the computer nodes has its central processing units, storage disk and memory. The applications of the users directly access only access leader node, not the computer nodes.

The Amazon Redshift is designed to provide rapid query performance across datasets ranging between gigabytes to exabytes. This includes data compression, columnar storage and zone maps for the minimisation of the I/O operations during these queries [7]. The redshift correspondents to the SQL operations through using a massively parallel processing architecture. The users can use a new data warehouse with the minimum effort through the AWS management process. In the query process, Redshift provides flexibility through which users can be able to execute SQL queries directly within the AWS processing. This warehouse solution ensures fault tolerance with the features which continuously handle the cluster health.

"Google Big Query"

Google Big Query is an "enterprise-grade cloud-native data warehouse" which first launched as a service in the year of 2010. The Big Query evolved into a more managed and economic data warehouse which processed ad-hoc queries and blazing-fast interactive [8].

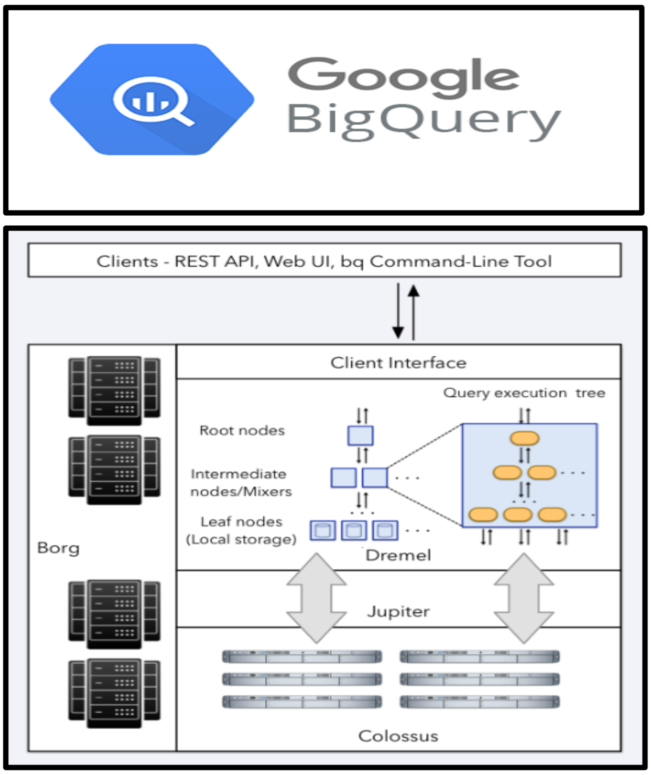
Google
BigQuery
Clients - REST API, Web UI, bq Command-Line Tool
Client Interface
Query execution tree
Root nodes
Intermediate nodes/Mixers
Leaf nodes (Local storage)
Borg
Dremel
Jupiter
Colossus

Figure 2: "Google Big Query Architecture" [9]

The architecture of the Google Big Query is as follows:

Dremel: The Dremel assist in excelling the scalability for the processing of petabytes of data rapidly. It includes the columnar data layouts and the tree architecture to manage a larger amount of data.

Colossus: This serves as the distributed file system of Google across its data centre which ensures comprehensive data storage.

Jupiter network: This network connects the Dremel and Colossus which facilitates the high speed and bi-directional data traffic within the data centre of Google. This network entitles the uninterrupted data movement which is required for the effective communication between the query execution and storage system.

Big Query offers various powerful features for machine learning and data analysis. The BI engines enhance the speed and validity of the Big Query which enables the rapid analysis of the complicated data sets within fractions. The Big Query ML assist scientists in constructing and including the ML models through the utilisation of simple SQL queries.

"Snowflake"

The Snowflake company, which started its operations in 2012 provides cloud-based data storage along with analytics services. This allows the corporate users to store and analyse the data that uses cloud-based software and hardware. The Snowflake warehouse solution utilises a new SQL database engine with a unique architecture structured for the cloud.

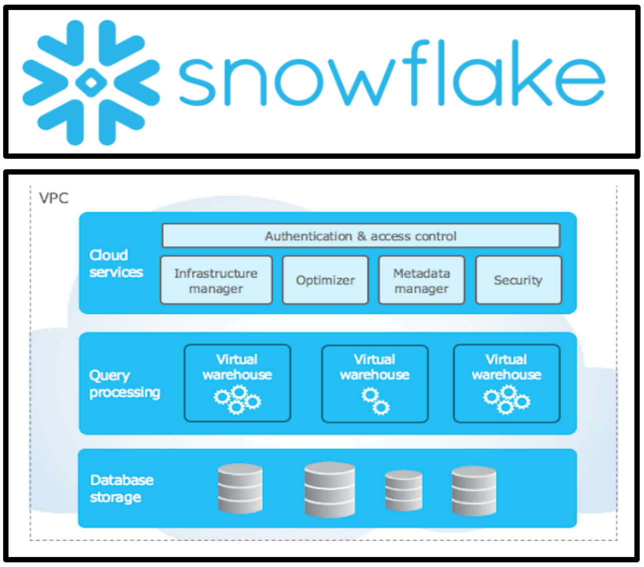

Figure 3: "Snowflake Architecture" [10]

The Snowflake architecture includes cloud services, database storage and query processing.

Cloud Services: The cloud services layer in Snowflake encompasses the synchronisation of the activities across its elements. It assists in handling the user demands from login through query execution.

Database Storage: The snowflake reorganises the data stored into its storage system to optimised, compressed and columnar format. Snowflake handles the organisation by optimising the file size, compression, structure, statistics and metadata.

These warehouse solutions allow users to share structured and semi-structured data within the organisation [10]. The multi-cluster shared data architecture expedites the scalability which enables the users to reshape the resources for the extreme data processing within any kind of disruption. It ensures holistic security from the user's access mechanisms for data storage, authentication and access control.

Findings and Discussion

Theme 1: Assessment of Performance Metrics

Amazon Redshift is an emerging cloud-based data warehousing solution which is renowned for its scalability and performance. The AWS ecosystem of Amazon Redshift offers a vigorous platform for businesses to manage large-scale data analytics and warehousing requirements. The columnar storage architecture enhances the query performance by minimising I/O operations and optimisation of the data retrieval for analytical queries. This assists in handling complicated analytical workloads including the aggregations and data-intensive operations which is found in data warehousing scenarios. The Google Big Query is a powerful cloud-based data warehouse which is constructed to handle large-scale data analytics with simplicity and speed.

Theme 2: Analysis of Cost Implications

This warehouse solution poses the capability to execute SQL queries swiftly over petabytes of the data stored on the Google Cloud Storage. It can be achieved through a distributed architecture which automatically lines up the resources depending on the complications in query and the overall data size. Additionally, this warehouse solution correlated with the other Google Cloud Services such as Google Data Studio for visualisation which facilitates the holistic data analytics ecosystem. In the study, the Snowflake is characterised as a significant cloud data warehousing platform which is popular for its innovative architecture to manage different data workloads. This warehouse solution offers various distinct features in the operations within the cloud environment to cater to modern data management requirements.

Theme 3: The integration of the abilities

The cloud platforms leverage the parallel processing and the distribution of the computing architectures for delivering high-performance query execution along with data processing. It further enables a faster pace of analytics management and real-time reporting for the enhancement of the informed decision-making process. For instance, the well-structured architecture of Snowflake segregates the compute and the storage layers which allows the independent scaling process for each of the components. This architecture is advantageous for the optimisation of performance and cost efficiency as compute resources for scaling up the workload demand within impacting the stored data. Cloud data warehouse solutions offer unparalleled scalability compared to traditional solutions. These warehouse techniques allow organisations to enlarge their storage capability and compute the available resources. This scalability allows the researchers and academic organisations to stare and analyse the data without the constraints of the warehouse infrastructure.

Evaluation

The research on data warehousing solutions in the cloud platforms, particularly focusing on the performance and cost analysis is essential within the evolving terrain of cloud technologies. In this research paper, various cloud-based data warehousing solutions such as Google Big Query, Amazon Redshift and many others have been mentioned. This includes the assessment of various factors such as scalability and query execution times under different workloads. The cloud data warehouse includes other cloud services and instruments such as data lakes, machine learning, business intelligence and ELT pipelines which involve the extraction, transformation and loading of the stored data. Integration with these areas fosters the symmetric data ecosystem in which the data can be processed, analysed and used across a diversified range of applications. Analysing the influence of the data volume and the complications on the performance, this research paper extends the reliability of validity of this research for the utilisation in the other research.

Conclusion

Analysing the above research, it can be concluded that the holistic exploration of data warehousing solutions in the cloud platforms. Warehouse technologies have transformed the management and analysis techniques of organisational data. Various cloud data warehouse solutions including Amazon Redshift, Snowflake and Big Query offer unique scalability, cost-efficiency and performance. This expedites the uninterrupted integration with other cloud services such as data lakes, business intelligence instruments and machine learning which constructs a vigorous ecosystem for the data processing along with the data analysis.

Research Recommendations

The evolution of the data warehousing solutions within the cloud platforms showcases the necessary shifts in the data-leveraging decision-making process for the organisation. The ongoing upgradation in cloud technologies coupled with the performance and scalability of cloud data warehouses highlights the crucial implication of these solutions in the modernisation of data management practices.

Future Work

Further research in this area should focus on the exploration of the emerging trends of warehouse solutions. The identification of security issues and the optimisation of cost-effectiveness further enhance the value proposition of cloud-based data warehousing solutions across a diverse range of organisations.

References

[1] B., Dageville, T., Cruanes, M., Zukowski, V., Antonov, A., Avanes, J., Bock, J., Claybaugh, D., Engovatov, M., Hentschel, J. Huang, and A.W., Lee. “The snowflake elastic data warehouse.” In Proceedings of the 2016 International Conference on Management of Data (pp. 215-226). 2016, June. https://dl.acm.org/doi/abs/10.1145/2882903.2903741

[2] P.J., Ferreira, A. de Almeida, and J., Bernardino. “Data Warehousing in the Cloud: Amazon Redshift vs Microsoft Azure SQL.” In KDIR (pp. 318-325). 2017. https://www.scitepress.org/papers/2017/65871/65871.pdf

[3] S. Fernandes, and J., Bernardino. “Cloud Data Warehousing for SMEs.” In ICSOFT-EA (pp. 276-282). 2016. https://www.scitepress.org/papers/2016/59965/59965.pdf

[4] Dageville, B., Cruanes, T., Zukowski, M., Antonov, V., Avanes, A., Bock, J., Claybaugh, J., Engovatov, D., Hentschel, M., Huang, J. and Lee, A.W., 2016, June. “The snowflake elastic data warehouse.” In Proceedings of the 2016 International Conference on Management of Data (pp. 215-226). https://dl.acm.org/doi/abs/10.1145/2882903.2903741

[5] B. Stantic, and J., Pokorný. “Opportunities in big data management and processing.” In Databases and information systems VIII (pp. 15-26). IOS Press. 2014. https://ebooks.iospress.nl/volumearticle/38184

[6] M., Ziauddin, A., Witkowski, Y.J., Kim, D., Potapov, J. Lahorani, and M., Krishna. "Dimensions-based data clustering and zone maps." Proceedings of the VLDB Endowment, 10(12), pp.1622-1633. 2017. https://dl.acm.org/doi/abs/10.14778/3137765.3137769

[7] R., Nadipalli. “Effective business intelligence with QuickSight.” Packt Publishing Ltd. 2017. https://books.google.com/books?hl=en&lr=&id=mrkrDwAAQBAJ&oi=fnd&pg=PP1&dq=Big+Query,+Redshift,+and+Snowflake+in+big+data&ots=QMOpkohjrR&sig=u6oi3lM-5kpmXJ7hImqFaWXdBRw

[8] S., Jain, J., Yan, T. Cruane, and B., Howe. “Database-agnostic workload management.” arXiv preprint arXiv:1808.08355. 2018. https://arxiv.org/abs/1808.08355

[9] B., Dageville, T., Cruanes, M., Zukowski, V., Antonov, A., Avanes, J., Bock, J., Claybaugh, D., Engovatov, M., Hentschel, J. Huang, and A.W., Lee. “The snowflake elastic data warehouse.” In Proceedings of the 2016 International Conference on Management of Data (pp. 215-226). 2016, June. https://dl.acm.org/doi/abs/10.1145/2882903.2903741

[10] M. Abourezq, and A., Idrissi. “Database-as-a-service for big data: An overview.” International Journal of Advanced Computer Science and Applications, 7(1). 2016. https://pdfs.semanticscholar.org/a0e7/e17762d0297f9a4cb7c714aa1ee86962c512.pdf