PANEVĖŽIO KOLEGIJA ISSN 2029-1280, eISSN 2669-0071. Taikomieji tyrimai studijose ir praktikoje – Applied Research in Studies and Practice, 2023, 19.

BIG DATA ANALYTICS IN CLOUD COMPUTING

Jurgita LIEPONIENĖ

Panevėžio kolegija/ State Higher Education Institution, Lithuania

Abstract. The modern world witnesses an exponential growth in digital data generation from various sources. Big data exhibit exceptional volume, velocity, and diversity, predominantly comprising unstructured or semi-structured formats. Preserving their accuracy and value necessitates the application of advanced Big Data technologies. Cloud Computing has revolutionized data management, offering organizations virtually unlimited storage and computational resources through models Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS). This paper explores the potential of Big Data analytics within the context of cloud computing and evaluates data analytics effectiveness via Google's BigQuery platform.

Keywords: cloud computing, big data, data processing, BigQuery

INTRODUCTION

In today's world, digital devices collect and generate an increasing amount of data every day. If 20 or 30 years ago only 1 percent of the information produced was digital, now over 94 percent of this information is digital and it comes from various sources such as our mobile phones, servers, sensor devices on the Internet of things, social networks, etc. (Berisha, Mëziu & Shabani, 2022). These data are distinguished not only by their remarkable volume and the speed at which they require storage and processing but also by their diversity, with the majority being unstructured or semi-structured. The preservation of their accuracy and value necessitates the application of advanced Big Data technologies and techniques.

Cloud Computing has emerged as a transformative force, significantly simplifying the storage, processing, and analysis of large-scale data. By harnessing the power of the Cloud, organizations gain access to virtually limitless storage capacity and computational resources, offered by a multitude of vendors. Notably, Cloud delivery models like IAAS (Infrastructure as a Service) and PAAS (Platform as a Service) have become invaluable assets for diverse enterprises in their pursuit of more efficient and expeditious Big Data management.

The objective of the research is to is to examine the potential of Big Data analytics within the context of cloud computing, assess the effectiveness of data analytics using Google's BigQuery platform.

Research methods: analysis of scientific literature, analysis of Google BigQuery platform.

The rest of this paper is structured as follows. Section 1 of this article explores concept of big data. Section 2 analyzes possibilities of big data analytics in cloud computing. Section 3 presents the results of analysis of Google BigQuery platform that assesses the effectiveness of data analytics using Google's BigQuery platform. Finally, Section 4 presents conclusions.

BIG DATA CONCEPT: LITERATURE REVIEW

Big data refers to large and complex datasets that cannot be easily managed, processed, and analyzed using traditional data processing methods (Alam, Singh & Shahin, 2023). Big data are characterized by three Vs: volume, velocity, and variety. These characteristics were introduced by Gartner to define the various challenges in big data (Sandhu, 2022). With new generation architecture, data are now stored in different types of formats; hence, the three Vs may be extended to five Vs, namely, volume, velocity, variety, value, and veracity (Sandhu, 2022). The main characteristics of big data can be described:

- **Volume**: the dataset that conforms to the big data standard is constantly changing and increasing over time. In big data, there is a large amount of data with sizes ranging from terabytes to zettabytes.
- **Velocity**: big data is characterized by the rapid generation of data, which, in turn, necessitates the rapid processing of that data in order to derive useful insights.
- **Variety**: data comes in various types, including structured data such as database data, semi-structured data such as XML data, and unstructured data such as sound, images, videos, web pages, text, etc.
- **Veracity:** veracity refers to the quality, correctness, and trustworthiness of data. Therefore, maintaining veracity in data is mandatory (Sandhu, 2022).
- **Value:** value is an important characteristic of big data. It relates to how data can be dealt with and converted into meaningful information (Alam, Singh & Shahin, 2023).

In the table below, traditional, and big data are compared. Traditional data typically refers to structured data that is organized in a predefined manner. Traditional data represents information collected and stored over time, following

specific formats and standards. Traditional data is characterized by its reliability, accuracy, and consistency, making it relatively easy to manage and process using relational databases and analytical tools.

Table 1. Comparison of traditional data and big data

No	Features	Traditional data	Big data
1.	Volume	The volume of data is limited	Extremely large volumes of data
2.	Variety	Structured data	Semi-structured and unstructured data
3.	Velocity	Low velocity: slow rate at which data is generated, collected, and processed.	High velocity: data is produced rapidly and continuously, often in real-time or near real-time
4.	Veracity	It is reliable and comes from trusted sources	It can be prone to noise, inaccuracies, and inconsistencies
5.	Value	Lower potential value	Provide valuable insights of the data

In summary, big data extends far beyond the constraints of traditional data, embracing vast volumes, high velocity, diverse variety, and complex veracity, making it a multifaceted and dynamic field with profound implications for numerous sectors. Big data analysis and interpretation require sophisticated technologies, algorithms, and expertise, leading to innovative solutions and transformative discoveries.

BIG DATA ANALYTICS IN CLOUD COMPUTING

Cloud computing refers to the delivery of various services, including software, storage, databases, networking, analytics, and intelligence, over the internet to offer faster innovation, flexible resources, and economies of scale. Instead of owning and maintaining physical servers or computers, individuals and organizations can access computing resources on-demand from a cloud service provider (Balachandran & Prasad, 2017). Cloud computing services are typically categorized into three main models:

- **Infrastructure as a Service (IaaS):** IaaS provides virtualized computing resources over the internet. Users can rent virtual machines, storage, and networking components on a pay-as-you-go basis. This model eliminates the need for physical hardware and allows users to scale resources based on their requirements.
- **Platform as a Service (PaaS):** PaaS offers a platform that allows developers to build, deploy, and manage applications without dealing with the complexities of infrastructure. PaaS provides a development environment with tools and services that facilitate the development process, making it easier and more efficient to create and deploy applications.
- **Software as a Service (SaaS):** SaaS delivers software applications over the internet on a subscription basis. Users can access these applications through a web browser, eliminating the need for installation and maintenance. SaaS providers handle software updates, security, and performance, allowing users to focus on using the software rather than managing it.

Cloud computing offers several advantages, including cost efficiency, scalability, flexibility, and accessibility. It enables businesses and individuals to access computing resources and services without the need for significant upfront investments in hardware and software. Cloud computing also supports collaborative work, data storage, and analysis, making it a fundamental technology for various industries and applications.

With the generation of an enormous amount of data, cloud computing is playing a significant role in the storage and management of that data (Islam & Reza, 2019). Big data analytics in cloud computing refers to the process of analyzing large and complex datasets using advanced analytics techniques within a cloud computing environment. Big Data and Cloud computing are a major trend that are rapidly growing and new challenges and solutions are being published every day (Samir et. al., 2020).

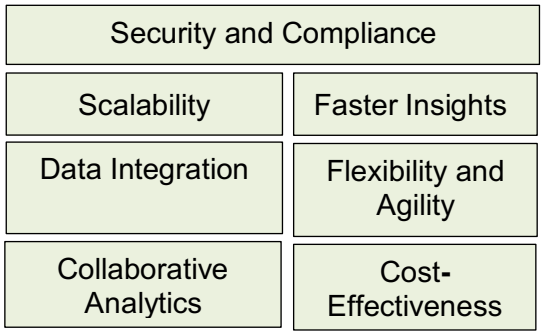

Figure 1. **Advantages of integrating big data analytics with cloud computing**

Because big data is now considered vital for many organizations and fields, service providers such as Amazon, Google and Microsoft are offering their own big data systems in a cost-efficient manner. These systems offer scalability for business of all sizes. This had led to the prominence of the term Analytics as a Service (AaaS) as a faster and efficient

way to integrate, transform and visualize different types of data (Berisha, Mëziu & Shabani, 2022). AaaS provides a faster and scalable way to integrate different types of structured, semi-structured and unstructured data, analyze them, transform and visualize them in real time (Islam & Reza, 2019). Big data analytics in cloud computing opens up a wide range of possibilities and opportunities for businesses and organizations (Figure 1).

Scalability: Cloud computing platforms offer virtually unlimited scalability, allowing businesses to store and process massive volumes of data. As data grows, cloud resources can be easily scaled up to handle increased computational and storage demands.

Cost-Effectiveness: Cloud computing operates on a pay-as-you-go model, where businesses pay only for the resources they use. This cost-effective approach is particularly beneficial for big data analytics projects, where processing requirements can vary over time.

Faster Insights: Cloud-based big data analytics solutions can rapidly process and analyze large datasets, providing businesses with real-time or near-real-time insights. This speed is crucial for making data-driven decisions and gaining a competitive edge.

Flexibility and Agility: Cloud platforms offer a variety of tools and services for big data analytics, allowing organizations to choose the most suitable technologies for their specific needs. This flexibility enables businesses to adapt their analytics strategies quickly in response to changing requirements or technological advancements.

Data Integration: Cloud-based big data analytics solutions can integrate with various data sources, including structured and unstructured data. Businesses can analyze data from diverse sources such as social media, IoT devices, customer interactions, and more, gaining comprehensive insights into their operations and customer behavior.

Collaborative Analytics: Cloud-based big data platforms facilitate collaboration among teams and departments. Multiple users can access and analyze data simultaneously, share insights, and collaborate on analytics projects in real-time, regardless of their physical locations.

Security and Compliance: Cloud providers invest heavily in security measures to protect data, including encryption, access controls, and compliance certifications. This ensures that sensitive data used in big data analytics is secure and meets regulatory requirements.

The integration of big data analytics with cloud computing provides businesses with powerful tools and capabilities to harness the potential of large datasets, gain valuable insights, and drive innovation and growth. The possibilities are vast and continually expanding as technology advances.

EVALUATION OF BIGQUERY PLATFORM

Google Cloud Platform provides a variety of services tailored for analyzing and processing large datasets. Among these services, BigQuery stands out as a fully managed, serverless data warehouse designed for scalable analysis over petabytes of data. Operating as a Platform as a Service (PaaS), it supports queries using ANSI SQL and boasts built-in machine learning capabilities. Rather than being kept in a line shape, the data is stored as columns, allowing for storage to be oriented (Sharma & Kumar, P, 2022). Only the columns that are required for data analysis are searched, resulting in a significant reduction in latency (Sharma & Kumar, P, 2022). Query processing and aggregation amongst different nodes with thousands of servers is done using a binary tree.

Since its inception in 2011, BigQuery has garnered widespread popularity, with numerous major companies leveraging its capabilities for their data analytics needs. From a user perspective, BigQuery has an intuitive user interface which can be accessed in several ways depending on user needs. The simplest way to inter-act with this tool is to use its graphical web interface. The most straightforward approach to interact with this tool is through its graphical web interface. Within the BigQuery web interface, users have the flexibility to add or select existing datasets, schedule and formulate queries, transfer data, and view results. This user-friendly platform empowers businesses and analysts to efficiently explore and analyze vast datasets, making data-driven decision-making processes streamlined and accessible. Evaluation of BigQuery platform according different criteria is presented in table 2.

Table 2. Evaluation of BigQuery platform

No	Criteria	Description
1.	Storage	Columnar storage: the data is organized into tables and stored in a columnar format, which allows for efficient query execution.
2.	Architecture	BigQuery uses a distributed architecture for data storage and processing.
3.	Permissions	Multiple permissions: it's possible to handle various access permissions, such as read-only, editing, and owner.
4.	Access methods	Multiple access methods: a BigQuery Browser, a BigQuery Command-line tool, a REST-based API.
5.	Security	SSL (Secure Sockets Layer) is used in the solution to assure security.
6.	SQL Compatibility	BigQuery supports standard SQL queries.
7.	Integration	BigQuery seamlessly integrates with other Google Cloud services, such as Google Cloud Storage and Google Data Studio.

Google BigQuery operates as an exceptionally rapid analytics database, providing customers with unparalleled performance and the freedom to explore vast datasets without limitations. The actual speed of BigQuery is a topic of interest, prompting an investigation into the factors influencing its performance. Google BigQuery has numerous Public Datasets. Bigquery-samples:wikipedia_benchmark tables were chosen for experiment (size of tables ranges from was 1k to 10B). This dataset consists of a single row of data for every Wikipedia revision/update, this includes the contributor_username who updated the article, and the article title itself, it also has language and contributor IP. The experiment showed that the execution of a simple SQL query with WHERE clause and GROUP BY statement in the largest table took about 10 seconds (Figure2). Query execution time depends on table size but increase slowly.

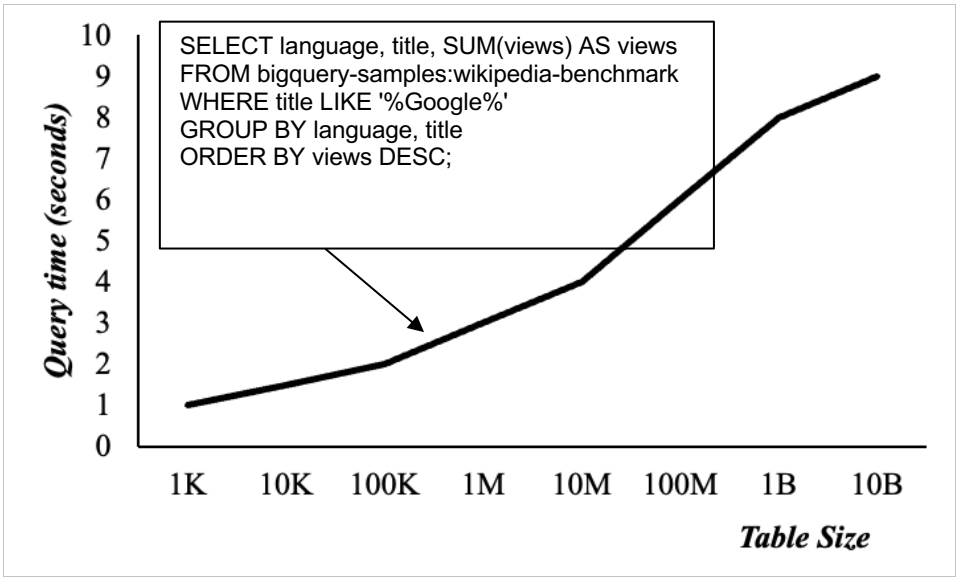

Figure 2. **Query execution in different size tables**

Google Cloud Platform offers a range of services for processing large datasets, with BigQuery standing out as a serverless data warehouse that supports scalable analysis over vast amounts of data using ANSI SQL and built-in machine learning features. Utilizing column-oriented storage and a binary tree-based processing system, BigQuery delivers rapid query performance, enabling users to explore extensive datasets efficiently, as demonstrated performed experiment.

CONCLUSIONS

Big data's scope exceeds traditional data limitations, embracing extensive volume, high speed, diverse types, and intricate complexities, leading to profound implications across various sectors. Analysis and interpretation of big data demand advanced technologies and expertise, fostering innovation and transformative discoveries.

The fusion of big data analytics with cloud computing equips businesses with potent tools to leverage large datasets, extract valuable insights, and fuel innovation and expansion. The potential is vast and continuously expanding with technological advancements.

Within the Google Cloud Platform, BigQuery stands out as a serverless data warehouse, supporting scalable analysis over vast datasets using ANSI SQL and integrated machine learning features. Its efficient column-oriented storage and binary tree-based processing system enable rapid query performance, showcasing its effectiveness for exploring extensive datasets.

REFERENCES

Alam, N., Singh, V. & Shahin, K. (2023). Big Data: An Overview with Legal Aspects and Future Prospects. *Journal of Emerging Technologies and Innovative Research*, 10(5), 476-485.

Balachandran, B.; Prasad, S. (2017). Challenges and Benefits of Deploying Big Data Analytics in the Cloud for Business Intelligence. *Procedia Computer Science*, 112, p.1112–1122.

Berisha, B., Mëziu, E. & Shabani, I. (2022). Big data analytics in Cloud computing: an overview. *Journal of Cloud Computing: Advances, Systems and Applications*, 11(24), 1-10. https://doi.org/10.1186/s13677-022-00301-w

Islam, M.; Reza, S. (2019). The Rise of Big Data and Cloud Computing. *Internet of Things and Cloud Computing*, 7(2), p. 45–53.

Sandhu, A (2022). Big Data with Cloud Computing: Discussions and Challenges. *Big Data Mining and Analytics*, 5(1), 32-40.

Samir, A.; Hosam, F.; Mohamed, A.; Reham, M. (2020). Big Data and Cloud Computing: Trends and Challenges. *International Journal of Interactive Mobile Technologies*, 11(2), p. 34–52.

Sharma, V.; Kumar, P. (2022). Big Query at a Glance. *International Research Journal of Modernization in Engineering Technology and Science*, p. 2684–2689.

Zanjani, M., Kabalci, Y. & Shahinzadeh, I. (2023). An Overview of Big Data Concepts, Methods, and Analytics: Challenges, Issues, and Opportunities. *2023 5th Global Power, Energy and Communication Conference (GPECOM) Papers*, 1-10.

