International Journal of Trend in Scientific Research and Development, Volume 1(2), ISSN: 2456-6470
www.ijtsrd.com

A Study on Cloud Based Big Data Analytics and Storage Solutions

Mohamed Ziyad Ta
Lecturer in Computer Engineering, SSM Polytechnic College, Tirur, Kerala, India

ABSTRACT
The advent of the digital age has led to a rise in different types of data with every passing day. In fact, it is expected that half of the total data will be on the cloud by 2016. This data is complex and must be stored, processed, and analysed for information that organisations can use. Given the storage and computing requirements of big data analytics, cloud computing is an ideal platform. This establishes cloud-based analytics as a viable research area. However, several issues and risks must be addressed before practical applications of this synergistic model can be widely used. This paper investigates current research, challenges, open issues, and future research directions in this field of study.

Keywords: *Cloud-based Big Data Analytics, Big Data, Big Data Analytics, Big Data Cloud Computing, Storage Solutions*

Introduction
The cloud provides exceptional flexibility, allowing organisations to expand their capabilities by incorporating big data analytics. Big data and analytics investments can be critical in driving efficient and cost-effective infrastructure.

Cloud computing models have the potential to accelerate the development of scalable big data solutions. The cloud allows for greater flexibility in accessing data, delivering insights, and driving value. Cloud-enabled big data analytics, on the other hand, is not a one-size-fits-all solution.

This is why it is critical to work with a partner who offers a variety of cloud options to support your big data initiatives. To select the best cloud platform for your analytics service, we consider a variety of factors such as cost, security, and workload.

Among the various cloud delivery models, a Private Cloud can provide a more cost-effective model for in-house big data analysis while also preparing internal resources with public cloud services. The Hybrid Cloud model enables organisations to use public services for on-demand storage and computing for certain analytics initiatives, while also providing additional capacity and scale as needed.

Analytics as a service (AaaS) framework
You can address user needs across various analytics requirements with our cloud-based AaaS. We assist our clients in developing a comprehensive cloud-based big data strategy, defining an AaaS framework, and optimising the value of their enterprise data.

Our AaaS framework comprises of the following capabilities:

- Capturing structured and unstructured data from a variety of trusted sources- identifying and prioritising the most critical data and deciding what to retain and for how long-
- Data management and data control with governance and policy guidelines- in compliance with specific industry requirements across the global enterprise
- Performing data integration, analysis, transformation, and visualisation in order to deliver the right information to the right place, at the right time, to the right people.

Depending on your usage and requirements, e-Zest provides various cloud models for your big data initiatives and assists you in making the best use of your IT budget by utilising analytics as a service (AaaS) supported by private, public, and hybrid cloud.

Characteristics of Big Data
One question that researchers have struggled with is what constitutes "big data." As a result, Gartner analyst Doug Laney introduced the 3 V model in 2001, which consists of three features that must be present in order for data to be considered "big data": volume, velocity, and variety. Volume is a property or feature that determines the size of data, which is typically reported in Terabytes or Petabytes. For

example, social networks such as Facebook store photos of users, among other things. Because of the large number of users, it is estimated that Facebook stores approximately 250 billion photos and over 2.5 trillion posts. This is a massive amount of data that must be stored and processed. The most defining feature of 'big data' is its volume. [1-2]

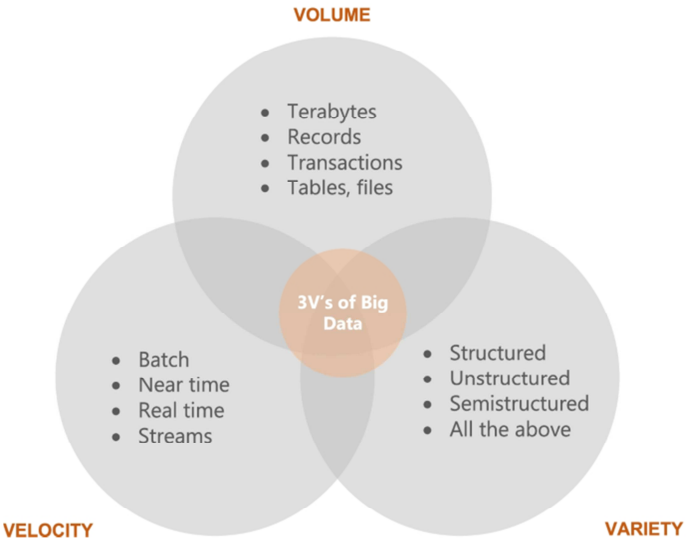

Figure 1: The Three Vs of Big Data [2]

The second property or feature is velocity. This refers to the amount of data generated or the rate at which it must be processed and analyzed.

Big data analytics in cloud computing

Cloud computing refers to the delivery of computing services such as servers, storage, databases, networking, software, analytics, and so on via the Internet ("the cloud") in order to provide flexible resources, faster innovation, and economies of scale. Cloud computing has transformed how computing infrastructure is abstracted and utilised. Cloud paradigms have been broadened to include anything that can be categorised as a service (hence x a service). Cloud computing's numerous advantages, such as elasticity, a pay-as-you-go or pay-per-use model, low upfront investment, and so on, have made it a viable and desirable option for big data storage, management, and analytics. Because big data is now considered critical for many organisations and fields, service providers such as Amazon, Google, and Microsoft are offering their own cost-effective big data systems. Scalability is provided by these systems for businesses of all sizes. As a result, the term *Analytics as a Service (AaaS)* has gained prominence as a faster and more efficient way to integrate, transform, and visualise various types of data [3].

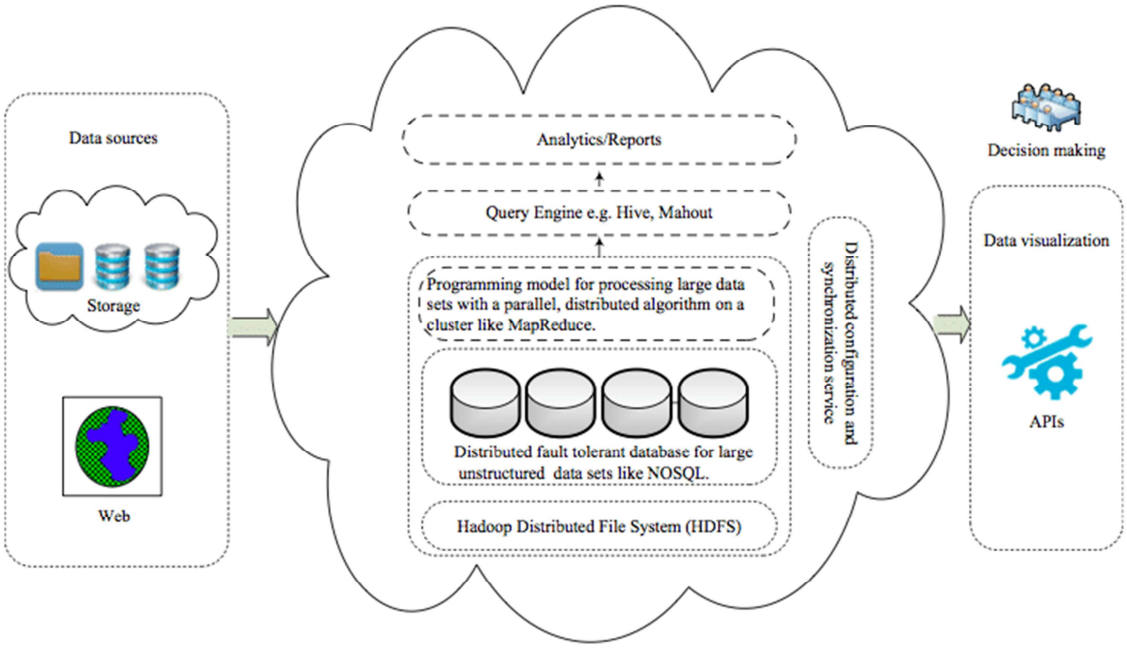

Figure 2: Application of Cloud Computing in Big Data

Lack of interactivity has recently been identified as a major issue, and several efforts have been made in this area. Borthakur, Gray, Sarma, Muthukkaruppan, Spiegelberg, Kuang, Ranganathan, Molkov, Menon, Rash, Schmidt, and Aiyer [4] improve the responsiveness of the HBase and HDFS implementations. Strambei assesses the viability of OLAP Web Services for cloud-based architectures, with the goal of enabling open and widespread access to web analytical technologies. [5]

Review of Literature:

Many authors and organisations have attempted to define 'Big Data.' According to Wikipedia, "Big Data refers to data volumes in the exabyte range and beyond." According to Wikipedia, big data is an accumulation of datasets that are so large and complex that they are difficult to process using database management tools or traditional data processing applications, with challenges such as capture, storage, search, sharing, transfer, analysis, and visualisation.

According to Sam Madden of the Massachusetts Institute of Technology (MIT), "Big Data" is data that is too large, too fast, or too difficult for existing tools to process. Too big data refers to data in the petabyte range originating from various sources. 'Too fast' refers to data growth that is rapid and must be processed quickly. Too difficult refers to the difficulty that arises as a result of the data failing to adapt to the existing processing tools. Big data is defined as "massive amounts of data collected over time that are difficult to analyse and handle using common database management tools," according to PCMag (one of the most popular journals on technological trends). There are numerous other definitions for Big Data, but we believe that these are sufficient to gain an understanding of the concept [6].

Artificial intelligence-based algorithms for data mining were developed in the 1980s. Wu, Kumar, Quinlan, Ghosh, Yang, Motoda, McLachlan, Ng, Liu, Yu, Zhou, Steinbach, Hand, and Steinberg [7] mention k-means, C4.5, Apriori, Expectation Maximization (EM), PageRank, SVM (support vector machine), AdaBoost, CART, a ve Bayes, and kNN as the ten most influential data mining algorithms (k-nearest neighbors). The majority of these algorithms have also been used commercially. Alam and Shakil [8] propose a data management architecture based on cloud techniques.

MapReduce is a popular model for data processing on a cluster of computers. Jackson, Vijayakumar, Quadir, and Bharathi [9] conduct a survey of big data analytics programming models. Although it identifies MapReduce/Hadoop as the most productive model for Big Data Analytics, it also mentions that languages and extensions such as HiveQL, Latin, and Pig have significant advantages for this use.

Hadoop is simply an open-source implementation of the MapReduce framework, which was designed to be a distributed file system in the first place. According to Neaga and Hao [10], Hadoop has evolved into a complete ecosystem or infrastructure that works alongside MapReduce components and includes a variety of software systems such as the Hive and Pig languages, a coordination service known as Zookeeper, and a distributed table store known as HBase.

Objectives:

- A research paper on cloud-based big data analytics and storage solutions.
- Dependence of average compute time on dataset size
- Big Query performance tests
- Using data studio to visualise data

Research Methodology:

The methodical, theoretical examination of the techniques used in a specific discipline is referred to as research methodology. It entails a comprehensive theoretical examination of the canonical practises and rules associated with a particular field of study. It typically includes terms like "paradigm," "theoretical model," "phases," as well as "quantitative" and "qualitative" methods. The data for this descriptive study was gathered from a variety of secondary sources, including education and development books, journals, scholarly papers, government publications, and printed and online reference materials.

Result and Discussion:

One benefit of using imported data in the cloud is the ability to manage its access and visibility within the scope of the cloud project and cloud members. Depending on how the data is used, it can be saved directly to the local computer using the "save results" option, which offers a variety of formats and data extension settings to choose from, or it can be explored in various configurations using the "explore data" option. You can also save built queries for later use or schedule query execution intervals for more precise data transmutation via API endpoints. Figure 4 depicts how the average compute time changes/increases as the size of the dataset used grows [11].

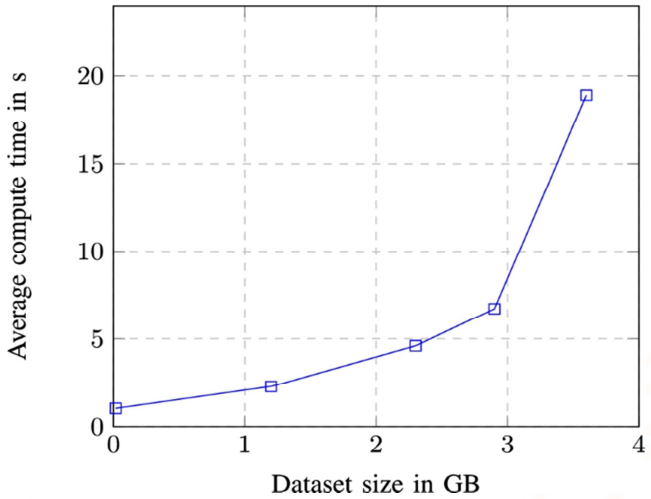

Figure 4 depicts the average compute time dependence on dataset size.

Before proceeding to data exploration, let us examine BigQuery performance results in simple queries with variable dataset sizes. The query execution details of five simple select queries run on five different datasets are shown in Table 1. The data shows a correlation between the size of the dataset and its average read, write, and compute performance, which is displayed against six different performance categories. [12]

Table 1: Table 1 BigQuery performance tests

Elapsed time (S)	Slot time Consumed (S)	Average Read (ms)	Average Compute (s)	Average Write (ms)	Number of Rows	Size of the Dataset
03	0.043	22	1.048	2	61,900	0.0175 GB
72	828.547	355	2.3	28,599	41,340	1.2 GB
3.3	3.663	945	4.6	109	100,000	2.3 GB
2.1	2.424	118	6.7	77	30,646	2.9 GB
1.6	1,506	237	18.9	145	100,000	3.6 GB

The graph shows that the relationship between dataset size and average compute size is exponential, which means that as data size increases, so does average compute time. [13]

Data returned from constructed queries can be transferred to data studio, which is an integrated tool for better displaying and visualising gathered information, in addition to being displayed in a simple tabular form or as a JSON object. [14]

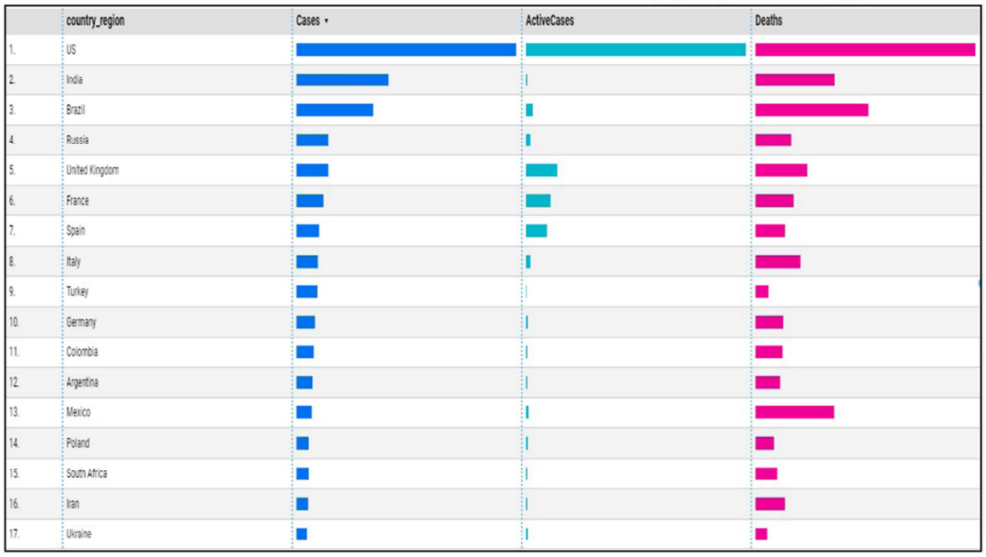

Figure 5: Data visualisation with data studio

Conclusion:

This is a big data age, and the emergence of this field of study has piqued the interest of many practitioners and researchers. Given the rate at which data is being generated in the digital world, big data analytics and analysis have become increasingly important. Furthermore, the majority of this data is already in the cloud. As a result, moving big data analytics to the cloud is a viable option.

According to our findings, big data is growing at a rapid pace, resulting in both benefits and challenges. Cloud computing is widely regarded as the best solution for storing, processing, and analysing large amounts of data. Companies such as Amazon, Google, and Microsoft provide public services to help with the Big Data process. According to our findings, Big Data analytics offers numerous advantages in a variety of fields and sectors, including healthcare, education, and business.

References:

[1] Weathington J (2012) Big Data Defined. Tech Republic. https://www.techrepublic.com/article/big-data-defined/

[2] Kaisler S, Armour F, Espinosa J (2013) Big data: issues and challenges moving forward, Wailea, Maui, HI, s.n, pp 995–1004

[3] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A. and Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. Information Systems 47 (2015) 98-115.

[4] Borthakur, D., Gray, J., Sarma, J. S., Muthukkaruppan, K., Spiegelberg, N., Kuang, H., Ranganathan, K., Molkov, D., Menon, A., Rash, S., Schmidt, R. and Aiyer, A. (2011). Apache Hadoop Goes Real-time at Facebook, in: Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2011), ACM, New York, USA, 2011, pp. 1071–1080. Retrieved from: http://cloud.pubs.dbs.unileipzig.de/sites/cloud.pubs.dbs.unileipzig.de/files/RealtimeHadoopSigmod2011.pdf

[5] Strambei, C. (2012). OLAP Services on Cloud Architecture. IBIMA Publishing. Journal of Software and Systems Development. Vol. 2012 (2012). DOI: 10.5171/2012.840273. Retrieved from: http://www.ibimapublishing.com/journals/JSSD/2012/840273/840273.pdf

[6] Weathington J (2012) Big Data Defined. Tech Republic. https://www.techrepublic.com/article/big-data-defined/R

[7] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A, Liu, B., Yu, P. S., Zhou, Z. H., Steinbach, M., Hand, D. J. and Steinberg, D. (2008). Top 10 algorithms in Data Mining. Knowl Inf Syst (2008) 14:1–37. DOI: 10.1007/s10115-007-0114-2. Retrieved from: http://www.cs.umd.edu/~samir/498/10Algorithms-08.pdf

[8] Alam, M., & Shakil, K. A. (2013). Cloud Database Management System Architecture. UACEE International Journal of Computer Science and its Applications, 3(1), 27-31.

[9] Jackson, J. C., Vijayakumar, V., Quadir, M. A. and Bharathi, C. (2015). Survey on Programming Models and Environments for Cluster, Cloud and Grid Computing that defends Big Data. 2nd International Symposium on Big Data and Cloud Computing (ISBCC '15). Procedia Computer Science 50 (2015) 517-523.

[10] Neaga, I. and Hao, Y. (2014). A Holistic Analysis of Cloud Based Big Data Mining. International Journal of Knowledge, Innovation and Entrepreneurship. Volume 2 No. 2, 2014, pp. 56—64. Retrieved from: http://ijkie.org/IJKIE_December2014_IRINA&HAO.pdf

[11] Google Cloud Platform. (n.d.). Big Query. Retrieved from: https://cloud.google.com/bigquery/

[12] GigaSpaces. (2012). Big Data Survey. Retrieved from: http://www.gigaspaces.com/sites/default/files/product/BigDataSurvey_Report.pdf

[13] Amazon Kinesis. (n.d.). Developer Resources. Retrieved from: http://aws.amazon.com/kinesis/developer-resources/

[14] Apache S4. (n.d.). Distributed Stream Computing Platform. Retrieved from: http://incubator.apache.org/s4/