International Journal of Advanced Engineering and Management Research

Vol. 8, No. 02; 2023

ISSN: 2456-3676

The Concept of Big Data and Solutions of Cloud Computing

Mustapha Malami Idina[1]
[1]University of Debrecen, Department of Mechatronics Engineering,
4028, Debrecen, Otemeto Strt 2-4
Mmustey.

[1]Kebbi State University of Science & Technology, Department of Computer Science
PMB 1144, Aliero, Kebbi State, Nigeria

doi: 10.51505/ijaemr.2023.8210 URL: http://dx.doi.org/10.51505/ijaemr.2023.8210

Abstract
The terms "big data" and "cloud computing" are often used interchangeably since many public cloud services analyze large amounts of data. Every day, a massive amount of data is produced from abundant of derivation, and this data needs to be evaluated, categorized, and stored with the assistance of cloud computing. The quick progress that has taken place in a variety of IT fields is to blame for this massive increase in the amount of data that is being produced. Conventional data processing methods are unable to handle this data due to its extremely large amount, the rapidity with which it changes, and the multitude of formats in which it is stored. Cloud computing makes available on-demand computing resources and services, making it possible to handle and analyze the vast amounts of data in a manner that is both effective and simplified. This research will discuss how on the general concept of big data and how Cloud computing provides solutions to big data by processing analyzing this huge data. We will start by Introduction of Big data, big data Impact on Today's world, big data architecture, cloud computing, Categories & offerings of Cloud computing then further discuss the solutions offered by cloud computing to big data and limitations of big data on the cloud.

Keywords: big data, cloud, azure, data, storage, artificial inteligence, machine learning

1. Introduction

The process that we used to characterize the ever-increasing volumes and varieties of data that are being collected is referred to as Big Data.

However, Big Data takes into account the quantity of data, the rate at which new information is generated and compiled, as well as the range or breadth of the data points that are being examined. The examination of large amounts of data raises difficulties in sampling, which had previously been restricted to merely allowing for observations and sample. As a result, the veracity of the data can be understood to refer to either its quality or its insightfulness. The volume and variety of data can bring costs and dangers that surpass an organization's capacity to create and exploit value from big data if significant investment in expertise for big data veracity is not made (Francesco, 2021)

Vol. 8, No. 02; 2023

ISSN: 2456-3676

Big data is an umbrella term that refers to the non-traditional tactics and technologies that are required to collect, organize, and process huge datasets in order to get insights from them. Although working with data that exceeds the processing capability or storage capacity of a single computer is not a novel challenge, the prevalence, scale, and utility of this form of computing have significantly expanded during the past few years (Hashem et al., 2014)

1.1 Big data Impact in today's world

Big Data allows businesses to adopt new tactics, which in turn improves decision-making, performance, and the ability to explore new prospects. Big data allows businesses to watch and analyze buying patterns, feedbacks, purchasing behavior, and a variety of other elements that affect sales, which enables these businesses to examine the decision-making processes of their clients. Big data may help firms enhance their marketing, advertising, and promotional tactics to increase client engagement and sales. Analyzing historical and real-time data can reveal consumer or business buyer preferences. This helps organizations meet client needs.

Access to real-time data gives businesses the ability to strengthen their intelligence gathering and security analysis, which is another way that big data contributes to the prevention and detection of cybercrime and fraud.

In the health sector, big data helps doctors and medical scientists identify diseases. Healthcare companies and government agencies receive the latest infectious disease outbreak and risk data through electronic health records, social media, the internet, and other sources using big data (Perez et al., 2015).

1.2 Big data architectures

Data sets that are too large or complex to be processed by traditional database systems are the major focus of a Big Data architecture's ingestion, processing, and analysis functions. The landscape of data has evolved significantly throughout the years. The ways in which you can use data and the ways in which you are expected to use it have both evolved. The cost of data storage has decreased drastically in recent years, but the number of data collection methods continues to expand.

Vol. 8, No. 02; 2023

ISSN: 2456-3676

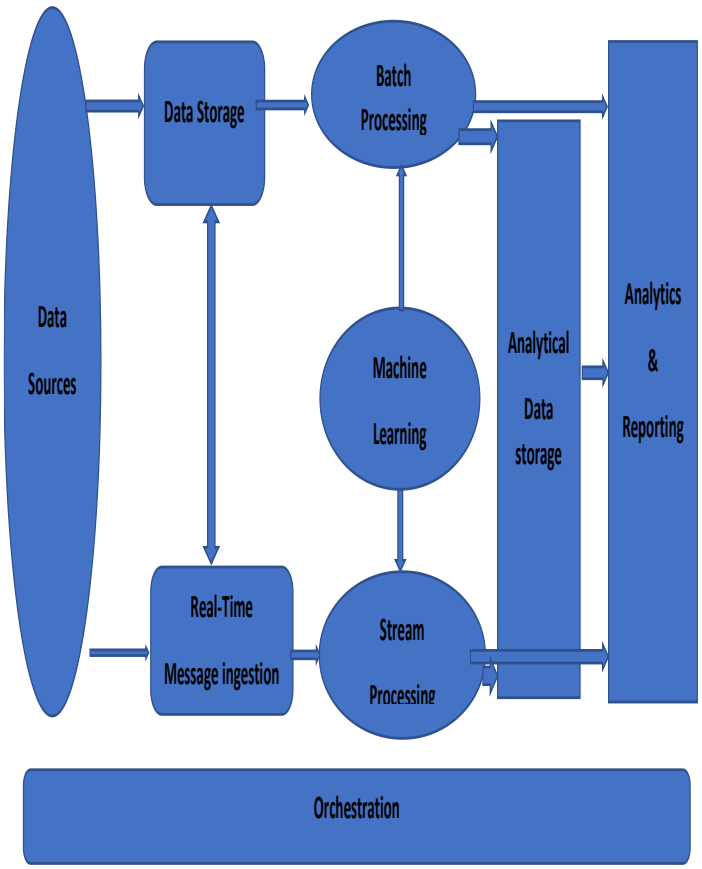

Figure 1: Big data Architecture (Tejada, 2022).

1.3 Components of Big Data Architecture

A. Data Sources: The obvious place to begin with any solution involving big data is with the data sources. Web server log files, relational databases, and real-time data sources are data sources.

B. Data Storage: There are file stores that are dispersed in nature and can store a range of format-based large files that include data. The data lake may also store a large number of massive files in a variety of formats. This is the data that is maintained for batch-built processes and stored in the file storage. Among other blob containers, they support HDFS, Azure, AWS, and GCP storage.

C. Batch Processing: Long-running jobs, which filter and aggregate data while also preparing it for analysis, are used to divide each chunk of data into its own set of distinct categories. A big data solution must frequently employ lengthy batch operations to process data files for filtering, aggregating, and preparing data for analysis. This involves using hive, Scala, java or U-SQL jobs in Azure Data Lake Analytics.

Vol. 8, No. 02; 2023

ISSN: 2456-3676

D. Real-time message ingestion**:** This element of the big data architecture incorporates a mechanism for capturing and storing messages derived from real-time sources in preparation for stream processing (Tejada, 2022).

E. Stream Processing: There is a difference between real-time message ingest and stream processing. The first method considers all the data that is being ingested from the beginning and then uses it as a publish-subscribe tool, whereas the second method uses the data that is being ingested as a publish-subscribe tool. After real-time messages are captured, the data is then processed through filters and aggregated using stream processing.

F. Machine Learning: it is responsible for the process that takes place between stream processing and batch processing. ML is a mechanics that is highly important, and now, thanks to the proliferation of big data, it has become more effective for the gathering of data, the analysis of data, and the integration of data.

G. Analytical data store: Analytical tools make advantage of the data store that is based on HBase or any other NoSQL data warehouse technology to perform analysis and processing on data that has already been processed. It's possible that a relational Kimball-style data warehouse is the analytical data store that's responsible for providing answers to these queries. This type of data warehouse is typical of most seen in business intelligence (BI) systems.

H. Analysis and reporting**:** Many big data solutions have as their primary objective the generation of actionable insights from the data using analysis and reporting. Also, the created insights must be processed, the processing is performed by using reporting and analysis tools that integrate embedded technology and a solution to provide meaningful graphs, analysis, and business-beneficial insights.

I. Orchestration: Orchestration refers to the process of automatically configuring, coordinating, and managing several software applications, computer systems, and other types of software. The workflows that are involved in repeated data processing operations can be automated with the help of a technology called orchestration.

1.4 Vs of Big Data

Big data consist of different set of data from various sources, and it is normally characterized by phenomenon knowns as Vs of Big data. The Vs of Big Data started from 3Vs, which are Volume, Velocity and Variety. However, overtime another 3Vs were added to describe Big Data, making them 6 Vs of Big Data

1. Volume: One of the many distinctive features of big data is volume, which places special emphasis on the correlation between data quantity and processing speed. This is rapidly evolving as more and more information is gathered. Similarly, to how information technology can store and process vast amounts of data.

Vol. 8, No. 02; 2023

ISSN: 2456-3676

2. Variety: Having the ability to capitalize on the potential offered by a variety of data sources and types, including structured and unstructured data, is essential. Integrating various types of data into a structure that can be easily managed is essential to developing a successful big data opportunity.
3. Velocity: Velocity implies that massive amounts of data should be processed rapidly, in a stream-like manner, because the data keeps coming. For example, during 30 minutes of flight, a single Jet engine creates more than 10 gigabytes of data.
4. Veracity: It is a term that refers to the anomalies, noises, and biases that can be found in data. Being able to determine whether or not a piece of data is relevant to a particular purpose, whether or not it is accurate, and using that data effectively.
5. Value: This refers to the objective, situation, or business result that the analytical solution needs to solve.
6. Variability: This pertains to the process of determining whether or not the contextualizing structure of the data stream is regular and reliable, regardless of the degree of unpredictability present in the environment. It establishes the necessity of gathering useful data while taking into account every conceivable scenario.

2.0 Cloud Computing

Cloud refers to internet-accessible computer servers and their programs and databases.

Cloud computing refers to the practice of storing and retrieving data through a network of remote servers that are housed on the internet. Cloud computing may be described in layman's terms as a virtual platform that does not impose any restrictions on the amount of data you can store or how easily you can access that data over the internet (Jarosz, 2022).

Cloud computing reduces the user's hardware and software requirements. We have all encountered cloud computing at some point, and some of the most prominent cloud services we have used or continue to use include email services such as Gmail, Hotmail, Yahoo, etc.

Cloud computing is one of the most imprecisely defined technical terms in the annals of technology's long and illustrious history. One of the reasons for this is because cloud computing may be implemented in a wide variety of application settings. Another reason is that numerous businesses are promoting cloud computing as a means of increasing revenue (Ling et al., 2020)

Popular examples of Cloud computing services are Google cloud, AWS, Azure etc.

However, we have three (3) major cloud computing offerings are

1. IAAS – Infrastructure as a Service
2. PAAS – Platform as a Service
3. SAAS – Software as a Service

Vol. 8, No. 02; 2023

ISSN: 2456-3676

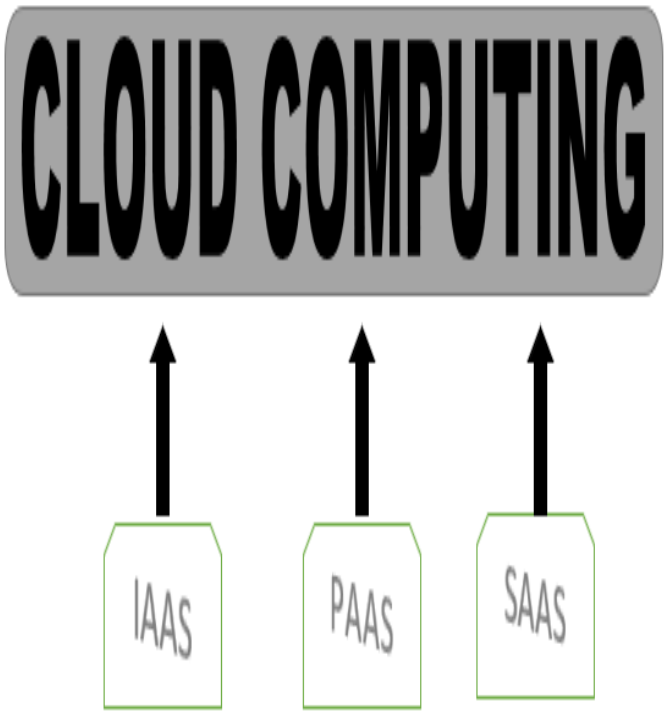

Figure 2: Major Cloud Computing offerings

2.1 Cloud Computing To Big Data

A. Big data Analysis: Because cloud technology employs a method of data integration from a variety of sources, conducting big data analysis there is now simpler and more enhanced. This is due to the fact that big data analysis in the cloud produces results that are both more accurate and more useful.

B. Security: When large amounts of data are kept in the cloud, they are shielded from intrusions into their privacy and are better able to withstand attacks from external forces. This is because cloud data is processed and kept in a centralized place known as a cloud storage server. Big data makes use of third parties in order to have scalable and elastic cloud solutions. This is accomplished through the signing of a Service Level Agreement (SLA) between customers and service providers in order to establish a foundation of trust between the parties involved. The cloud storage server is an open environment and an open-source application.

C. Perfect Match: The main reason for the existence of cloud computing services is the advent of big data, and the only reason we generate big data is because we are aware that we have a technology that will assist us in easily analyzing, processing, and categorizing the data that we generate. Applications that run in the cloud frequently store and handle big data; hence, cloud technology will be less useful, if not completely useless, in the absence of big data. Therefore, it is safe to state that both can co-exist with one other and that both needs each other.

Vol. 8, No. 02; 2023

ISSN: 2456-3676

D. Flexible Infrastructure: Traditional infrastructures are often overwhelmed by the volume, velocity, and variety of Big Data, making analysis of this data a substantial burden. Since the Cloud's underlying infrastructure can be easily scaled to meet fluctuating demands, managing workloads is a breeze.
E. Speed and Feasibility: The processing and analysis of massive amounts of data, which would previously have required weeks or months to complete using more traditional methods, may now be completed in only a few minutes. The traditional method of data processing requires a large number of servers and a substantial amount of physical work to increase the processing speed and storage capacity, whereas cloud-based technology enables increased processing speed and storage capacity to be very easily adjusted to meet the needs of individual users.

2.2 Utilization of Cloud Computing For Big Data

A fault-tolerant distributed database is used to store massive amounts of data collected from the web and the cloud. This data is then put through a programming model designed specifically for processing large datasets, which uses a parallel distributed algorithm.

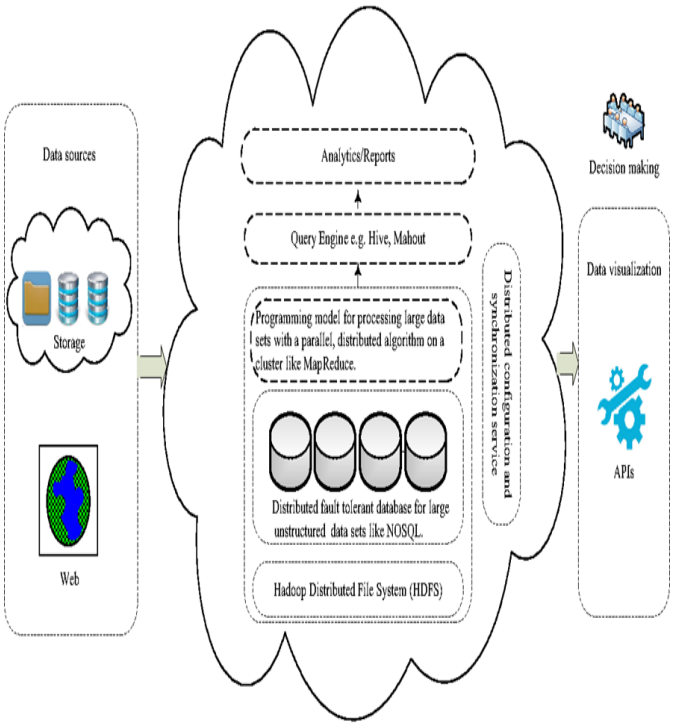

Figure 3: Cloud computing usage in big data (Yi et ak., 2014).

2.3 Disadvantage of Big Data In The Cloud

There are a few restrictions to large data in the cloud despite its many advantages, which are.

A. Cost: Long-term storage of large amounts of data in the cloud is prohibitively expensive, as is the migration process required for classification of the data. As we have seen, cloud

Vol. 8, No. 02; 2023

ISSN: 2456-3676

computing provides a relatively streamlined method of data processing, but it is also very expensive. Big data processing on the cloud must be managed in a highly cost-effective manner.

B. Constant Attention and Care: Due to the lack of standardization that exists for processing large data on the cloud, big data on the cloud needs to have constant care and maintenance performed on it.

4. Conclusion

Big data is a large issue that is rapidly evolving. Identifying formerly undetectable patterns and gaining understanding of previously unknown behaviors are two of the primary uses for big data platforms. The exponential expansion of Big data, AI, and ML necessitates the development of efficient methods for analyzing the huge volumes of data generated daily. Big data and cloud computing play an essential part in our society, particularly in areas like the medical, business, and aviation industries, amongst others, which generate massive amounts of data that need to be refined in order to address the many challenges that are present in these sectors. Cloud environments provide big data systems with environments that are fault-tolerant, scalable, and available, which is a significant boost to the effectiveness of big data solutions.

References

Francesco, C., Raffaele, O., Enzo, P., & Ian, M. (2021). Big Data for Creating and Capturing Value in the Digitalized Environment: Unpacking the Effects of Volume, Variety, and Veracity on Firm Performance. Journal of Product Innovation Management, 38(1), 49–67. https://doi.org/10.1111/jpim.12545

Hashem, I. A. T., et al. (2014). The rise of "big data" on cloud computing: Review and open research issues. Information Systems, 47, 98–115.

Jarosz, J. (2022). Big data and cloud computing: roles and relationships, techniques and tools. JDA, 1(1), 33–41.

Ling, Q., Zhigou, L., Yujian, D., & Leitao, G. (2020). Cloud Computing: An Overview. ResearchGate. https://www.researchgate.net/publication/344087649_Cloud_Computing_An_Overview

Perez, A. J., Poon, C., Merrifield, R., Wong, S., Yang, G. Z., & Zhong, G. (2015). Big Data for Health. IEEE Journal of Biomedical and Health Informatics, 19, 10.1109/JBHI.2015.2450362.

Singh, A., & Saxena, A. (2022, June 03). Big Data Architecture - Detailed Explanation. InterviewBit. https://www.interviewbit.com/blog/big-data-architecture/

Tejada, Z. (2022). Big data architectures. Microsoft Learn. https://learn.microsoft.com/en-us/azure/architecture/data-guide/big-data/

Yi, X., Liu, F., Liu, J., & Jin, H. (2014). Building a network highway for big data: Architecture and challenges. IEEE Network, 28(4), 5–13. https://doi.org/10.1109/MNET.2014.6863125Postman, N. (1979). *Teaching as a conserving activity*. New York, NY: Delacorte Press.