WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Problems Integrating Big Data on Cloud : A Review

Sugriv Maurya[1], Deepika Bansal[2]

Department of Information Technology, Maharaja Agrasen Institute of Technology, Rohini, Delhi, India

A B S T R A C T :

Since cloud computing and big data are designed to function together, their underlying concepts provide a number of integration obstacles. If these issues are handled, however, they can be extremely appropriate for a wide range of applications. The paper examines a number of distinct facets of the issue at hand as well as potential solutions

Keywords: Cloud Computing, Big Data, Data Security , Scalability, Data Integration

Introduction

In recent years, there has been a growing need to store and manage an expanding amount of data in industries including government, science, and finance. Big data handles data processing and storage, while cloud computing offers a more stable, available, fault-tolerant, and scalable environment in which big data systems are functional. In both the business and scientific domains, big data—and especially big data analytics—is viewed as a means of correlating data, identifying patterns, and projecting future trends. Because these two technologies can give businesses a competitive edge and give scientists a mechanism to gather and synthesize trial data, there is a lot of interest in merging them. Cloud computers provide access to both software and hardware resources, with the goal of upending the conventional computing paradigm. The internet is used to provide these services. It is well-liked due to its extensive schedule, affordability, and accessibility. The five main features of cloud computing—on-demand capabilities, broad network accessibility, source merging, rapid adaptability, and a well-balanced solution—set it apart from earlier computer paradigms. Users can access terabytes of storage, high processing power, and high availability in a pay-as-you-go model. Cloud computing offers limitless storage and computational power, enabling the extraction of vast volumes of data. The capacity of the cloud to virtualize resources, little communication with cloud service providers is necessary for this. Large data processing and storage necessitates availability, fault tolerance, and scalability. Through hardware virtualization, cloud computing offers all of these. Because the cloud makes large data accessible, scalable, and fault-tolerant, big data and cloud computing are therefore complementary ideas. Businesses view big data as a possible commercial opportunity. Consequently, several new companies have started to concentrate on offering Big Data as a Service, including Cloudera, Hortonworks, Teradata, and others Data Baseasa Service (DBaaS) or BDaaS. Through businesses like Google, IBM, Amazon, and Microsoft consumers. While the term "big data" primarily refers to the storing of massive amounts of data, it also refers to methods for processing and deriving knowledge from it. The five "V's"—volume, variety, velocity, value, and veracity—are the five distinct characteristics that are used to characterize big data.

This field is always changing, but it still has a lot of limitations. Because the amount of data is increasing so quickly, it is physically inefficient to keep all of it. Businesses must therefore be able to create policies that outline the lifecycle and expiration date of data (datago vernance).Additionally, they must to outline who can access customer data and why. Security and privacy are a growing worry that is being thoroughly studied as more data moves to the cloud. Big data enables users to conduct distributed queries over various data sets and produce result sets in a timely way using commodity computing. Hadoop, a type of distributed data processing platform, is used by cloud computing to supply the underlying engine. Huge data sources from the Web and cloud are handled using a programming paradigm for huge datasets with a parallel distributed algorithm in a cluster and saved in a distributed fault-tolerant database.

Instead than using local storage connected to a computer or other electronic device, big data makes use of distributed storage technologies based on cloud computing. Rapidly expanding cloud-based applications are the driving force behind big data evaluation. As a result, cloud computing functions as a service model in addition to offering resources for big data processing and calculation.

2. Literature Review

In order to supply enormous data with high demand, these numerous servers operate in tandem. Several servers are already used by cloud computing, which also permits resource allocations. Thus, it is a good idea to construct the large data on these cloud multi-servers and utilize the resource allocation capabilities offered by the cloud settings, which would improve big data analysis efficiency. Performance would be enhanced for both if massive data were stored on cloud systems. Cloud systems can manage large volumes of data at once because they are mostly built on remote multiple servers. Because of this feature, big data can handle enormous volumes of data thanks to modern analytics techniques. Costs would decrease as a result

of the integration of big data and cloud computing. However, in order to handle the enormous volume of data, big data necessitates server clusters and volumes. Instead of building new servers and volumes for big data, cloud computing systems can act as the framework for all of these. This offers greater flexibility and scalability and removes the significant financial outlays that would otherwise be required for big data computers and servers. Despite all of these benefits of the cloud computing and big data integration, there are a few hazards and difficulties to take into account when implementing big data in a cloud setting. The security of the big data cloud environment should be the primary concern. The combination of the two plus the development of a new, unknown platform results in some security flaws. Platform heterogeneity is among the most well-known security flaws in Big Data clouds. Numerous Big Data implementations exist that need to put a new platform in the cloud, and new security tools must be created to cope with these new Big Data platforms because the cloud's current security procedures and tools won't work for such platforms. Authentication, access control, encryption, intrusion detection, and event logging and monitoring are a few examples of these security tools. When integrating with the cloud environment, strategies for Big Data aggregation should be taken into account in addition to security standards. The location and nature of the data present another difficulty because, in the case of Big Data, the data may be spread over multiple places. These places might or might not be part of the cloud environment. The kind of processing that should be done on the data, how parallel the processing should be, and if the data should be transported to a processing environment or the processing should be done on the data's location.

When deploying the Big Data to a cloud system environment, all of these issues should be taken into account. Furthermore, optimizing the Big Data cloud topology presents additional difficulty since it outlines the configuration, cloud size, clusters, and nodes that must be present in order to achieve the best Big Data cloud model. Data heterogeneity is the most significant open research question in data staging. The format of data collected from various sources is not organized. For example, blogs, social networking, and mobile cloud-based apps are not properly formatted, much like text messages, videos, and pictures. It can be difficult to transform and clean such unstructured data before putting it in the warehouse for analysis tasks. To store and retrieve vast volumes of data, a variety of techniques have been put forth.

A cloud computing environment has seen the implementation of several of these ideas. However, a number of problems prevent such solutions from being implemented successfully, such as the inability of existing cloud technologies to deliver the high performance and capacity required to handle enormous volumes of data, optimization of current file systems to meet data mining volume requirements applications, and the ways in which information can be kept to facilitate easy retrieval and server migration. Choosing the right model is essential when analyzing big amounts of data. Scalable analysis techniques are necessary to extract meaningful information from massive data sets in a timely manner. Nevertheless, when it comes to massive data analysis, existing algorithms are ineffective. Thus, to process such data, effective data analysis methods and technologies are needed. The performance of each algorithm starts to rise linearly as processing resources increase. Even though cloud computing has revolutionized contemporary ICT technology, a number of Several threats and issues, including privacy, confidentiality, integrity, and availability of data, are present in big data using cloud computing platforms. As a result, data security must be measured once data is outsourced to cloud service providers, and the cloud must be evaluated on a regular basis to protect it from threats. These security threats are exacerbated by the volume, velocity, and variety of big data.

3. Problem Overview on various characteristics of Big Data

Finding knowledge in data sources is accomplished through the use of data mining techniques. It is used to both extract useful information from data and evaluate data from various sources. Data mining is also used to find links and trends in the dataset, forecast patterns or values, and categorize and categorize data. It is required in fields such as business, science, marketing, advertising, and medical, among others. In addition to providing a form of understanding discovery system composed of numerous decentralized data analysis solutions, an integrated data Mining approach and cloud computer are used to gain rapid access to innovation. Despite the benefits of big data, there are also many drawbacks. Of these drawbacks, protecting data privacy is the most crucial issue in big data mining applications since processing large amounts of sensitive data, like medical records and bank transaction records, requires that the private information not be disclosed to unauthorized parties.

3.1 Volume

Exaflop Computing : Although modern supercomputers and clouds can handle petabyte datasets, handling Exabyte-sized datasets still poses a number of issues since processing and transferring such massive amounts of data over a network requires high performance and capacity. Data Governance: This broad term refers to businesses with sizable datasets and describes procedures for managing data accesses over the course of their life cycle and preserving valuable data. It is a matter that requires cautious consideration. When it comes to hardware prices, the notion that storage is inexpensive and likely to decrease further is accurate. But a big data DBMS also involves other costs like software licenses, energy, and infrastructure upkeep. The total cost of ownership (TCO), which is expected to be seven times greater than the costs of purchasing hardware, is the sum of all these expenses.

3.2 Velocity

Since the current systems hardly manage data peaks automatically, scalability and elasticity in cloud computing, particularly specifically with relation to large data management systems, are areas that require more investigation. Based on a number of factors, including security, workload rebalancing (i.e., the necessity to rebalance burden), and redundancy (which would enable fault tolerance and availability), a properly scalable system would permit both manual and automatic reactive and proactive scalability. Distributed database storage system: Large volumes of data are stored and retrieved by reusing a variety of technologies. One crucial component of big data is cloud computing. Every day, a variety of devices produce big data. Data processing and migration between multiple servers, as well as straightforward data storage, are now the primary challenges in distributed frameworks.

Since both the data and the queries are encrypted, data access is not an issue. However, the cost of encryption often results in longer query processing times

3.3 Value

Data processing and cleaning: Preprocessing and cleaning, which includes data merging, data filtering, data consistency, and data optimization, are necessary for data storage and acquisition. The large number of data sources makes it challenging to process and clean data. Furthermore, data sources might be incomplete or contain mistakes and noise. The task demonstrates how to clean up vast amounts of data and how to assess the reliability of such data. Disaster Recovery: Being able to react effectively to dangerous situations is crucial because losing data could lead to financial loss. Disaster recovery procedures may be crucial to the successful deployment of large data database management systems (DBMSs) in the cloud and to maintaining their availability and fault tolerance. Security: When businesses are thinking about transferring data to the cloud, this issue becomes problematic. It's challenging to explain issues like who actually owns the data, where it is, who can access it, and what kind of permissions they have the only issue is that less sensitive data that can be useful for big data analysis is likewise encrypted and inaccessible. Loss of information could be another issue

3.4 Veracity

Privacy: Analytical software can identify users' private information, including their energy usage, online activities, medical records, grocery store records, and more. The scrutiny of this data raises questions around discrimination, exclusion, profiling, and loss of control. These days, privacy policies appear to be founded on user consent and information that people voluntarily submit. When big data is employed with cloud computing frameworks, problems and risks such data availability, confidentiality, real-time monitoring, identity and access authorization control, integrity, and privacy arise. Others: Businesses frequently decide to physically deliver hard drives to the data centers so that data may be transferred because moving data to the cloud is a very slow procedure. However, uploading data to the cloud using this method is not the most practical or secure option.

3.5 Variety

Data Visualisation: This method involves presenting complex data in a graphical format for easy comprehension. The conventional graphical method of representing data is simple if it is structured. Unstructured or Semi structured data makes it challenging to depict in real time with a high degree of variation.

Heterogeneity: handling such a large diversity of data and varying velocity rates is a challenging issue for big data systems. The fact that new file formats are always being developed without any sort of standardization makes this work even more difficult. Because big data is collected from various devices in various formats, including text, audio, video, and image, it is inherently heterogeneous. Data must be cleansed and converted before being loaded into a warehouse, and large data operations can be difficult. It is very challenging to combine all unstructured data and reconcile it for use in creating reports in real time.

4. Case Studies

4.1 Health Care Sector

In the healthcare industry, managing data is an enormous responsibility. Millions of patients are involved, and the data must be carefully preserved for future use in addressing public and general concerns. In order to address the dilemma, the healthcare business aims to improve the implementation of their system by integrating it with cloud computing. Maintaining the security and quality of data in the healthcare sector requires a systematic approach to data management. One administrative method that can be useful for storing data is data management. Verifying and processing the essential data to guarantee its dependability. The need for data management has grown in the modern day as various businesses must deal with massive volumes of data. It is crucial to keep data security in mind. Since many departments and sectors handle sensitive data, data security is a crucial component in data protection. Cloud computing technology adaptation can help the healthcare sector avoid physical servers by enabling efficient data storage at low cost. Nonetheless, there are still several security-related problems that may prevent the industry from embracing cloud computing technologies.

Real Estate Sector

Cloud computing Data security is a significant issue, according to opponents. The apparent drawback of cloud computing to the uninformed eye is that the client no longer has control over security. Any business's lifeblood is its data, therefore when management is asked to transfer it to an external entity, that could potentially expose it to competitors, hackers, and other risks, which causes red flags and worry to rise rapidly. Although data security is a major concern, the majority of cloud service providers outperform many small and midsize real estate companies. These businesses employ a centralized system that is watched over by a group of experts, allowing them to swiftly strengthen security and react to threats. As cybercriminals become more skilled, protecting data has become too difficult for an in-house IT professional or an IT service managing many, distinct client systems.

Although data security is a major concern, the majority of cloud service providers outperform many small and midsize real estate companies. These businesses employ a centralized system that is watched over by a group of experts, allowing them to swiftly strengthen security and react to threats. As

cybercriminals become more skilled, protecting data has become too difficult for an in-house IT professional or an IT service managing many, distinct client systems. Real estate firms need to choose their contractors wisely when it comes to security. The cloud provider should be able to provide certifications and references from clients that are publicly traded or have security concerns. Additionally, the service must adhere to government-mandated data security regulations or industry-specific standards.

Performance as well as Availability The continuous availability and performance of cloud computing services, especially in times of crisis and disaster, is another significant concern among real estate professionals. Realtors were flooded with requests from clients who had suffered storm-related home damage when Hurricane Ike struck the Gulf Coast in September 2008. When their clients needed them the most, those real estate agents who lost Among the remedies for the problems mentioned above are base, Pig, Hive, Sqoop, Oozie, Hadoop, MapReduce, NoSQL, and Apache Spark .Hadoop is the most important of these systems for managing large amounts of data

5. Conclusion

Cloud computing and big data are closely related. As technology advanced, big data models offered real-time analysis of heterogeneous datasets, parallel technologies, distributed computing, and massive storage capacity. Big data models also take privacy and data security into account. Big data necessitates the usage of cloud computing since it demands a lot of storage space. Scalability and cost savings are two benefits of cloud computing. Additionally, it offers enormous amounts of processing power and storage space. Since most firms are unable to defend against various threats, the enormous volume, velocity, and variety of data is causing issues. It is impossible to conduct huge data mining operations without jeopardizing privacy.

As the amount of data grows daily, big data systems and especially analytical tools have emerged as a significant source of innovation that offer methods for storing, processing, and retrieving information from peta byte datasets. Cloud environments make extensive use of big data solutions by offering big data systems scalable, available, and fault-tolerant environments. Cloud computing seems to be a great way to store large amounts of data. On the other side, juggling two conflicting design ideas is another challenge when working with huge data on the cloud. Combining big data and cloud computing technology can help businesses and educational institutions have a better future. Large volumes of data in a variety of forms may be stored and processed quickly, producing data that will help businesses and academic institutions grow quickly. The amount of data has increased over the past few decades and is still growing daily. Multiple sources generate data in a variety of formats. As a result, the range of data is likewise growing.

REFERENCES

[1] Neves, Pedro & Schmerl, Bradley & Cámara, Javier & Bernardino, Jorge. (2016). Big Data in Cloud Computing: Features and Issues 307-314.10.5220/000584630307031.

[2] Pushpa Mannava, "An Overview of Cloud Computing and Deployment of Big Data Analytics in the Cloud", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 1 Issue 1, pp. 209-215, 2014. 10.32628/IJSRSET207278

[3] Kumar Sharma, D., Sreenivasa Chakravarthi, D., Ara Shaikh, A.,Al Ayub Ahmed, A., Jaiswal, S., & Naved, M. (2021). The aspect of vast data management problem in healthcare sector and implementation of cloud computing technique. Materials Today: Proceedings. doi:10.1016/j.matpr.2021.07.388

[4] Sharma, Anil & Singh, Gurwinder & Rehman, Shabnum. (2020). A Review of Big Data Challenges and Preserving Privacy in Big Data.10.1007/978-981-15-0694-9_7.

[5] Amanpreet Kaur Sandhu. Big Data with Cloud Computing: Discussions and Challenges. Big Data Mining and Analytics, 2022, 5(1): 32-40. BIG DATA MINING AND ANALYTICS ISSN 2096 -0654 03/06 pp 32 – 40 Volume 5, Number 1, March 2022 DOI: 10.26599/BDMA.2021.9020016.

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

[7]El-Seoud, S. A., El-Sofany, H. F., Abdelfattah, M. A. F., & Mohamed, R. (2017). Big Data and Cloud Computing: Trends and Challenges. International Journal of Interactive Mobile Technologies (iJIM), 11(2), 34. doi:10.3991/ijim.v11i2.6561

[8]Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S. (2015). The rise of "big data" on cloud computing: Review and open research issues. Information Systems, 47, 98–115. doi:10.1016/j.is.2014.07.006

[9]Manoj Muniswamaiah, Tilak Agerwala, Charles Tappert , "Big Data in Cloud Computing Review and Opportunities", eprint arXiv:1912.10821,International Journal of Computer Science & Information Technology (IJCSIT) Vol 11, No 4, August 2019

[10]N. Deepa, Quoc-Viet Pham, Dinh C. Nguyen, Sweta Bhattacharya, B. Prabadevi, Thippa Reddy Gadekallu, Praveen Kumar Reddy Maddikunta, Fang Fang, Pubudu N. Pathirana, A survey on blockchain for big data: Approaches, opportunities, and future directions, Future Generation Computer Systems,Volume 131,2022.

[11] Hariri, R.H., Fredericks, E.M. & Bowers, K.M. Uncertainty in big data analytics: survey, opportunities, and challenges. J Big Data 6, 44 (2019). https://doi.org/10.1186/s40537-019-0206-3

[12] Jin Wang, Yaqiong Yang, Tian Wang, R. Simon Sherratt, Jingyu Zhang, "Big Data Service Architecture: A Survey," Journal of Internet Technology, vol. 21, no. 2 , pp. 393-405, Mar. 2020.