iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** XI **Month of publication:** November 2023

DOI: https://doi.org/10.22214/ijraset.2023.56926

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XI Nov 2023- Available at www.ijraset.com

Efficient Data Management Strategies for Cloud Computing Environments

Debayan Guha[1], Ronit Biswas[2], Anirban Bhar[3], Soumya Bhattacharyya[4]
[1, 2]*B.Tech student, Department of Information Technology, Narula Institute of Technology, Kolkata, India*
[3, 4]*Assistant Professor, Department of Information Technology, Narula Institute of Technology, Kolkata, India*

Abstract: ***Cloud computing is rapidly gaining traction in business. It offers businesses online services on demand (such as One drive, iCloud and Google One) and allows them to cut costs on hardware and IT support. Variety of data to be analysed continues to quickly rise as new applications like social network analysis, semantic Web analysis, and bioinformatics network analysis proliferate. The problem of efficiently managing and analysing massive amounts of data is both fascinating and important. There has been a lot of focus on big data lately, from researchers to business leaders to policymakers. From both a technical and practical perspective, this paper presents numerous methods for processing massive amounts of data. We begin by discussing the most important aspects of big data analytics and processing from the perspective of cloud data management and the mechanisms involved in processing large amounts of data. This covers the cloud computing platform, cloud architecture, cloud database, and data storage scheme. It examines the responsibilities of hosting businesses who own and run cloud computing datacentres (like Amazon) with regards to data security and privacy. As an example, it takes into account the fact that many cloud service providers (like Dropbox and Salesforce) really rent out space in the cloud' from larger hosting corporations. And it examines the business and private 'clouders' using these services. As we all familiar with cloud computing, it's not a latest technology, rather we can mention it as an emerging technology where most of the industry is trying to store not only its crucial data for redundancy but also looking for the service management. In that scenario first thing comes in mind is management of data in most efficient way possible. The new concepts that cloud introduce, such as multi-tenancy, resource sharing and outsourcing, create new challenges to the security community. Proposing new security rules, models, and protocols to handle the distinct cloud security concerns is necessary in addition to the capacity to grow and fine-tune the security mechanisms developed for traditional computing systems. The advent of cloud computing has revolutionized the information technology sector. It's like a dream come true, allowing developers with innovative ideas to create new Internet services without the hefty costs of buying and managing hardware. They no longer need to worry about having too much or too little capacity. Big companies can also benefit, running large tasks efficiently by scaling resources as needed, without incurring extra expenses for scale. This flexibility is truly a breakthrough in the world of IT.***
Keywords: Cloud Computing, IT Support, Big Data, Data analytics, Decision-making.

I. INTRODUCTION

The introduction of cloud computing has caused a revolution in the information technology sector. This revolution has made it possible for consumers and organizations to enjoy the benefits of on-demand internet services like as OneDrive, iCloud, and Google One while simultaneously decreasing the need for expensive hardware and IT support. This new technology lays enormous responsibility not only on hosting firms like Amazon that own and run cloud data centres, but also on cloud service providers like Dropbox and Salesforce that lease space from these hosting companies. In addition, it is tailored to meet the requirements of both commercial and individual customers who depend on the cloud services. Because cloud computing is still a relatively new technology, it has brought to the foreground a very important component of data management. Multi-tenancy, resource sharing, and outsourcing are all examples of concepts that present new issues for information security. In order to address these problems, not only will it be necessary to adapt the security mechanisms used in traditional computing systems, but it will also be necessary to build novel security policies, models, and protocols that are adapted to the specific security challenges that are posed by cloud computing. As a result of the cloud's adaptability and low operating costs, the landscape of the information technology industry has been completely revolutionized. This has made it possible for software developers to build innovative Internet services, and it has made it possible for major businesses to scale their resources in an effective manner. This represents a tremendous technological advance for the sector. This article investigates the incorporation of data management into the realm of cloud computing, providing light on the complexities of this game-changing technology in the process.

In a situation involving cloud-based data management, institutions rent storage and processing capacity in order to make data management apps function. This is done rather than making significant capital in-house investments for infrastructure. Even though there has been a substantial amount of marketing push from major public cloud providers, the adoption of cloud-based systems has been rather low (Geczy et al, 2013). The principal supporters of public cloud services are the public cloud providers themselves. According to what Mann and Morton outlined in their respective studies, data management on the cloud first appears to be highly promising from an economic point of view, but it becomes expensive in the long term. According to some estimates (Mann, 2011; Morton and Alford, 2009), the typical payback period for public cloud computing services in the present day is less than two years. According to Anthes (2010) and Lanois (2010), the legal protection of data and services, as well as the control of those characteristics, are among the most crucial issues for businesses. Shi provides a description of the cloud in an essay from 2010. (Shi et al, 2010) Although there has not been a standardized definition of cloud computing, he identifies the substantial qualities of it as scalability, fault tolerance, high performance cost, pay-as-you-go, and other similar characteristics. Cloud-based data management systems offer a solution that is both flexible and cost-effective for expanding horizontally using commodity hardware. Furthermore, the scaling of the server resources is completely transparent to the applications. Since the beginning of this decade, it has been seen that an increasing number of businesses are migrating their data management applications from expensive and high-end servers to less expensive solutions that are hosted in the cloud. In the field of cloud computing, one of the most vitally significant topics for research is data management. There are currently various cloud-based data management systems that are operational. Some examples of these systems include BigTable in Google, Cassandra and Hive in Facebook, HBase in Streamy, and PNUTS in Yahoo! Computing in the cloud has emerged as a significant force in the field of data management research and now plays an important part. This research study based on mentioned technological literatures indicate that the research in cloud-based data management systems is still in the early stages, and there is considerable potential for IS scholars to contribute to this area of study. The findings of this work make it abundantly evident that the researchers have comparable research interests in the following fields: data management architecture, data security, and privacy in the cloud.

II. LITERATURE REVIEW

The exponential growth of computational capacity over the past two decades has generated an enormous volume of data, necessitating a paradigm shift in the architecture of computing and mechanisms for processing large-scale data. A few weeks prior to his January 2007 disappearance at sea off the coast of California, Jim Gray, a Microsoft researcher and pioneer in database software, referred to the transition as the "fourth paradigm" [1]. Recently, computational science has replaced experimental, theoretical, and the previous three paradigms. Gray argued that developing a new generation of computing tools to manage, visualize, and analyse the data deluge is the only way to deal with this paradigm. Overall, contemporary computer architectures are becoming more unbalanced, as the latency disparity between mechanical hard disks and multi-core CPUs widens annually [2]. This imbalance exacerbates the difficulties associated with data-intensive computing. However, this is not the only issue that requires attention or resolution. Contemporary applications that manipulate petabytes and terabytes of distributed data [3] require guaranteed Quality of Service (QoS) when accessing networked environments. Neglecting the network mechanisms will result in applications being provided with a best effort service, which falls short of meeting their requirements [4]. In light of this, an architecture that can scale for the foreseeable future is imperatively required to address these issues in a methodical and general fashion. Gray countered that the current trend should prioritize the affordability of clusters of computers for data management and processing, rather than the acquisition of the largest and quickest individual computer. In contrast to their predecessors, which were confined to local and dedicated networks, interconnection technologies are crucial in today's Internet-connected clusters [5]. This is because these clusters must operate in parallel, irrespective of their distances, in order to manipulate the data sets required by the applications. Over the past few months, cloud computing has generated considerable excitement. According to Gartner, cloud computing ranks first among the ten most disruptive technologies that will emerge in the coming years [6]. The provision of computational infrastructure is characterized by a paradigm shift in the cloud computing industry. By relocating this infrastructure to the network, this paradigm reduces the expenses associated with the administration of hardware and software resources. As a result, organizations and individuals are granted immediate access to application services from any location across the globe. As a result, it embodies the long-held aspiration to view computing as a utility [7], in which the principles of economies of scale effectively reduce the cost of computing infrastructure. Prominent corporations including Amazon, Google, IBM, Microsoft, and Sun Microsystems have initiated the process of constructing additional data centres across the globe to host Cloud computing applications. This strategic move aims to augment redundancy and guarantee dependability in the event of site malfunctions. Indeed, the definition of cloud computing has been the subject of considerable debate in scholarship and industry [8,9,10].

The functional definition of cloud computing, as established by the National Institute of Standards and Technology (NIST) of the United States, encompasses the elements that are generally accepted. The NIST working definition [11] is frequently cited in documents and initiatives of the United States government.

III.FRAMEWORK TO MANAGE BIG-DATA IN CLOUD COMPUTING ENVIRONMENT

Numerous researchers have opined that commercial DBMSs are inadequate for the processing of data on an exceedingly large scale. The database server is a potential bottleneck for traditional architectures during periods of high throughput. Two essential objectives of big data processing—cost and scalability—are constrained by the capabilities of a single database server. D. Kossmann et al. introduced four distinct architectures, namely partitioning, replication, distributed control, and cache architecture, which are derived from the classic multi-tier database application architecture. These architectures were designed to accommodate a variety of large data processing models [3]. Clearly, the alternative providers target distinct application types and have distinct business models: Google appears to prioritize small applications that handle light duties, whereas Azure is the most cost-effective option for medium to large-scale services at present. A significant number of contemporary cloud service providers are implementing hybrid architectures that effectively meet their operational service demands. This section focuses on three fundamental components of big data architecture: an open-source cloud infrastructure, a distributed file system, and non-structural and semi-structured data storage.

A. Distributed File System

An example of a chunk-based distributed file system that ensures fault tolerance through data partitioning and replication is Google File System (GFS) [4]. It functions as the foundational storage layer of Google's cloud computing platform, facilitating the storage and retrieval of MapReduce output [5]. Hadoop, in a similar fashion, utilizes the Hadoop Distributed File System (HDFS) [6], an open-source alternative to GFS, as its data storage layer. GFS and HDFS are user-level filesystems that are designed to handle large files (measured in gigabytes) while extensively optimizing for performance without implementing POSIX semantics.[7]. Amazon Simple Storage utility (S3) [8] is a web utility provided by Amazon Web Services for online public storage. Clusters maintained on the server-on-demand infrastructure of Amazon Elastic Compute Cloud are the intended users of this file system. S3 endeavours to deliver low latency, scalability, and high availability at commodity prices. ES2 [9] is an epiC6 elastic storage system that is specifically engineered to accommodate both functions within a single storage medium. Efficient data input from various sources, a flexible data partitioning scheme, an index, and parallel sequential scan are all features of the system. Furthermore, there exist general filesystems, including Kosmos Distributed Filesystem (KFS)8 and Moose File System (MFS7), which have yet to be discussed.

B. Storage of Non-structural and Semi-structured Data

As an increasing number of IT companies experience the benefits of Web 2.0, their requirements for storing and analysing data from a variety of web services, including search logs, crawled web content, and click streams, typically in the petabyte range, increase. Nevertheless, web data sets are frequently non-relational or less structured, and the processing of these semi-structured data sets on a large scale presents an additional obstacle. Furthermore, the aforementioned basic distributed file systems are inadequate to meet the requirements of major service providers such as Google, Yahoo!, Microsoft, and Amazon. Each service provider's goal is to assist prospective users, and they all possess the most recent advancements in big data administration systems for cloud environments. Bigtable [10] is a Google distributed storage system designed to scale to the storage of petabytes of structured data across thousands of commodity servers for the purpose of data management. Bigtable lacks the capability to accommodate a comprehensive relational data model. On the contrary, it furnishes customers with a straightforward data model that facilitates dynamic manipulation of data format and layout. PNUTS [11] is a hosted database system of enormous scope that was specifically engineered to provide support for the web applications of Yahoo! the system's primary emphasis is on providing data for web applications, as opposed to handling intricate queries. On the PNUTS platform, the development and maintenance of new applications is a straightforward process with minimal associated costs. Constructed to provide support for internal Amazon applications, Dynamo is a scalable and highly available distributed key/value data store [12]. To fulfill the demands of these applications, it offers a straightforward primary-key interface. It is, nevertheless, distinct from key-value storage systems. A hybrid data management system, Llama [13], was proposed by Facebook as a novel cluster-based data warehouse system. This system integrates the advantageous aspects of both row-wise and column-wise database systems. Additionally, they delineate CFile, a novel column-wise file format designed for Hadoop that exhibits superior data analysis performance compared to alternative file formats.

C. Platform for Open-Source Clouds

The fundamental concept underlying data centres is to optimize the utilization of compute resources through the implementation of virtualization technology. As a result, it offers fundamental components including storage, CPUs, and network bandwidth at a low unit cost as commodities provided by specialized service providers. Virtualization is incorporated into cloud architectures by the majority of research institutions and businesses in order to achieve their big data management objectives. The leading cloud management platforms for infrastructure as a service (IaaS) include Amazon Web Services (AWS), Eucalptus, Opennebula, Cloudstack, and Openstack. Although AWS9 is not free, it is widely utilized in elastic platforms. It is extremely user-friendly and pays only on a pay-as-you-go basis. The Eucalyptus [14] operates as an open source in IaaS. A virtual machine is employed to manage and control resources. As the first IaaS cloud management platform, Eucalyptus enters into an API-compatible agreement with AWS. It holds a dominant market share in the private cloud sector within the AWS ecological environment. OpenNebula [15] supports a multitude of integration environments. It can provide superior interoperability, the most comprehensive features, and adaptable methods for constructing private, public, or hybrid clouds. OpenNebula lacks support for Service Oriented Architectures (SOAs) and its decoupling between components that are independent of computation, storage, and the network is inadequate. CloudStack10 is an open-source cloud operating system that utilizes user-supplied hardware to provide public cloud computation comparable to Amazon EC2. Users of CloudStack can maximize the benefits of cloud computing, including increased end-user efficiency, scalability, and rapid deployment of new services and systems. CloudStack is presently one of the open-source initiatives developed by Apache. It has mature functions already. Nevertheless, additional improvement is required in the lax coupling and component design. OpenStack11 is a collection of open-source software initiatives that collaborate with enterprises, developers, and researchers to foster an open-source community. Individuals within this community are united by the objective of developing a cloud infrastructure that is straightforward to implement, extraordinarily scalable, and loaded with advantageous functionalities. OpenStack's architecture and components are stable and uncomplicated, making it an excellent option for developing enterprise-specific applications. At the moment, OpenStack is characterized by a thriving community and ecological milieu. Nonetheless, it has some deficiencies, including unfinished features and a dearth of commercial support.

IV. CHALLENGES OF CLOUD DATA MANAGEMENT

The process of deploying data-intensive applications in a cloud environment is neither simple nor straightforward. Abadi [30] and Armbrust et al. [7] constituted a list of impediments to the expansion of cloud computing applications.

1) *Availability of a Service:* A distributed system, at its core, is a system that functions reliably across an extensive network. One notable characteristic of network computing is the potential for network interconnections to vanish. Organizations are concerned about the availability of cloud computing services. Achieving high availability is among the most difficult objectives due to the fact that even a minimal disruption can result in substantial financial ramifications and undermine consumer confidence.
2) *Data Confidentiality:* Migrating data off-site generally amplifies the quantity of potential security vulnerabilities, necessitating the implementation of suitable precautions. Typically, transactional databases encompass the entirety of the operational data required to facilitate mission-critical business operations. This dataset comprises information at the most fundamental level of specificity, frequently encompassing confidential data like credit card numbers or customer information. Hence, should such confidential information not be encrypted utilizing a key not stored on the host, a third party could gain unauthorized access to it without notifying the client.
3) *Data Lock-In:* Active standardization of APIs for cloud computing has not yet occurred, resulting in data lock-in. Customers are therefore unable to effortlessly transfer their data and programs from one website to another. The apprehensions regarding the challenges associated with data extraction from the cloud are impeding the implementation of cloud computing by certain organizations. While some cloud computing providers may find customer lockin appealing, users of cloud computing remain susceptible to various risks, including price hikes, reliability issues, and even provider insolvency.
4) *Data Transfer Bottlenecks:* To reduce expenses, cloud customers and providers must consider the ramifications of placement and traffic at each tier of the infrastructure.
5) *Application Parallelization:* Elastic computing capacity is contingent upon the parallelizability of the workload. Obtaining additional computational resources does not require an immediate upgrade to a larger, more powerful machine. On the contrary, acquiring the supplementary resources generally involves assigning additional server instances to a given task.

6) *Shared-Nothing Architecture:* In order to operate effectively on a cloud environment, data management applications ought to adhere to a shared-nothing architecture [31]. This configuration ensures that every node is autonomous and self-sufficient, and that no single point of contention exists throughout the system. It is not common for transactional data management systems to employ a shared-nothing architecture.
7) *Performance Unpredictability:* Numerous high-performance computing (HPC) applications rely on the concurrent execution of all program threads. Nevertheless, virtual platforms and operating systems of the present day do not offer this functionality.
8) *Application Debugging in Large-Scale Distributed Systems:* Error elimination in these extremely large-scale distributed systems is a challenging aspect of programming in cloud computing. It is not uncommon for these flaws to be undetectable in smaller configurations; therefore, debugging must take place at the same magnitude as in the production datacentres.

V. CONCLUSION AND FUTURE SCOPE

The process of generating new explicit or tacit knowledge through the synthesis of antecedent information and data, or knowledge creation itself [Becerra-Fernandez et al., 2004]. This is significant because it facilitates the exploration of uncharted territories in research. Despite the growing attention towards research in the field of data management, there remains a necessity for individuals to share their insights and findings.

In addition, this review attempts to highlight the most recent developments in cloud data management research. Environmental concerns, security and privacy issues, and problems of scale (including the storage of petabytes of data, the provision of facilities for analytical processing, online query processing, and the execution of queries in vast parallel) are all challenges associated with cloud data management. The pragmatic implementation of public, private, and hybrid cloud architectures is also examined in this study. The accessibility, possession, and geographical placement of cloud-based environments vary between the two. Organizations benefit the most from private clouds since they allow for unfettered management of their data, applications, and infrastructure (Orakwue, 2010).

However, research in this field is not particularly prominent. A hybrid framework exists to address the challenges and risks associated with public clouds, which are deemed the most detrimental for organizations (Hofmann and Woods, 2010). Control over an organization's valuable infrastructure, services, and data is relinquished. Sotomayor et al. (2009) define hybrid clouds as "a combination of public and private cloud technologies."

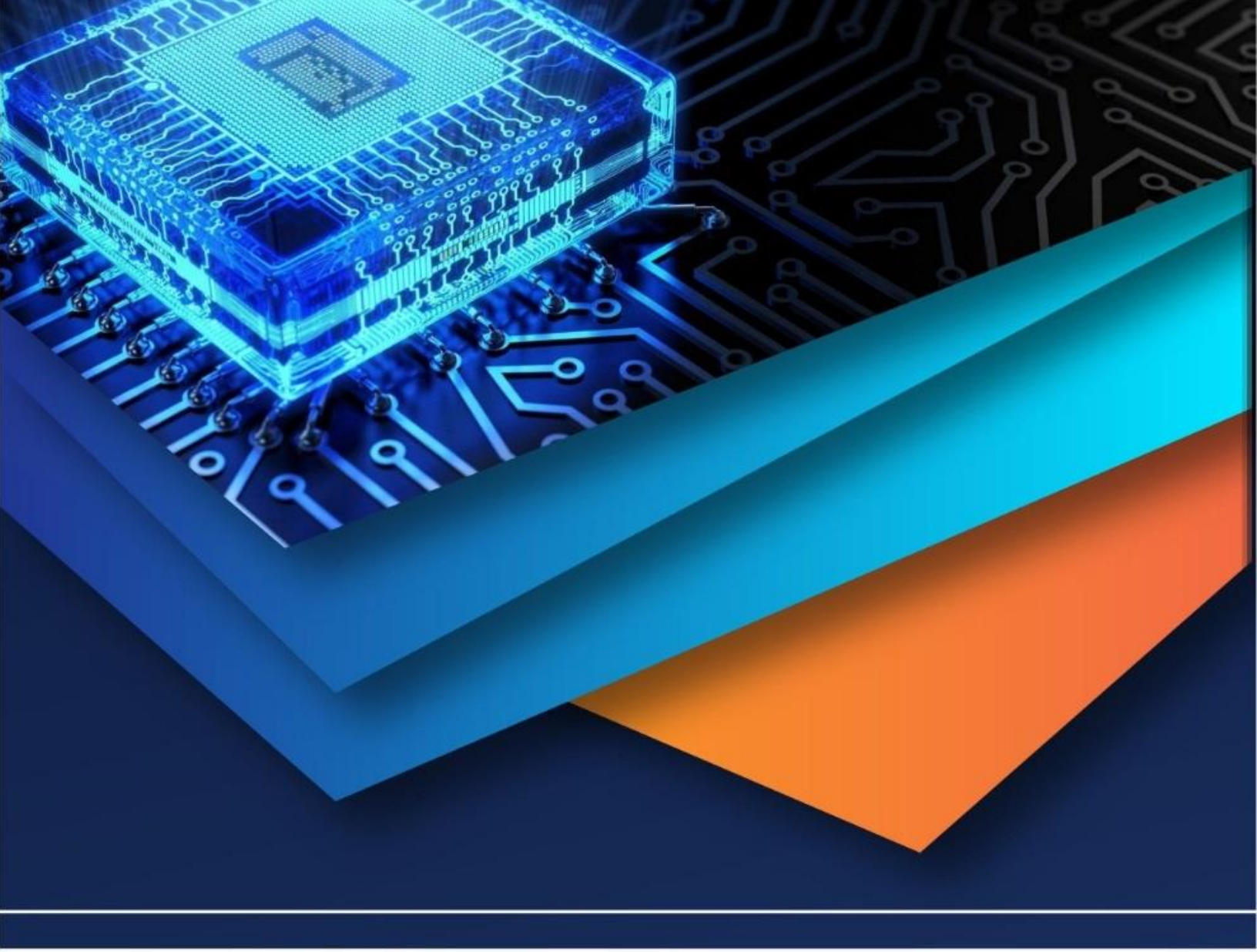
While there has been some progress made in this area of study, more work needs to be done to improve accessibility, analytical processing, and query processing. There is still a need for research into optimal methods for administering hybrid clouds. One drawback associated with grouping, swapping, or merging is that they diminish the integrity of the data. Limited scholarly inquiry has been devoted to this matter, suggesting that additional research may be warranted.

Cloud computing optimizes the utilization of distributed resources, integrating them to attain increased throughput and address challenges associated with large-scale computation. Concurrently, cloud computing enables the reduction of expenses, the enhancement of flexibility and responsiveness, and the improvement of service quality. In addition to IBM, Microsoft, and Google, Amazon, and Yahoo, numerous IT companies have put forth their own cloud computing strategies. Concurrently, numerous telecommunications operators have placed significant emphasis on cloud computing. This paper provides an overview of fundamental concepts, including various cloud models, deployment models, and cloud storage techniques. In addition, a number of the risks associated with cloud-based data storage and their respective remedies are briefly described.

REFERENCES

[1] "Big data: science in the petabyte era," Nature 455 (7209): 1, 2008.
[2] Douglas and Laney, "The importance of 'big data': A definition," 2008.
[3] D. Kossmann, T. Kraska, and S. Loesing, "An evaluation of alternative architectures for transaction processing in the cloud," in Proceedings of the 2010 international conference on Management of data. ACM, 2010, pp. 579–590.
[4] S. Ghemawat, H. Gobioff, and S. Leung, "The google file system," in ACM SIGOPS Operating Systems Review, vol. 37, no. 5. ACM, 2003, pp. 29–43.
[5] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," Communications of the ACM, vol. 51, no. 1, pp. 107–113, 2008.
[6] D. Borthakur, "The hadoop distributed file system: Architecture and design," Hadoop Project Website, vol. 11, 2007.
[7] A. Rabkin and R. Katz, "Chukwa: A system for reliable large-scale log collection," in USENIX Conference on Large Installation System Administration, 2010, pp. 1–15.
[8] S. Sakr, A. Liu, D. Batista, and M. Alomari, "A survey of large scale data management approaches in cloud environments," Communications Surveys & Tutorials, IEEE, vol. 13, no. 3, pp. 311–336, 2011.
[9] Y. Cao, C. Chen, F. Guo, D. Jiang, Y. Lin, B. Ooi, H. Vo, S. Wu, and Q. Xu, "Es2: A cloud data storage system for supporting both oltp and olap," in Data Engineering (ICDE), 2011 IEEE 27th International Conference on. IEEE, 2011, pp. 291–302.

[10] F. Chang, J. Dean, S. Ghemawat, W. Hsieh, D. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. Gruber, "Bigtable: A distributed structured data storage system," in 7th OSDI, 2006, pp. 305–314.
[11] B. Cooper, R. Ramakrishnan, U. Srivastava, A. Silberstein, P. Bohannon, H. Jacobsen, N. Puz, D. Weaver, and R. Yerneni, "Pnuts: Yahoo!'s hosted data serving platform," Proceedings of the VLDB Endowment, vol. 1, no. 2, pp. 1277–1288, 2008.
[12] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels, "Dynamo: amazon's highly available keyvalue store," in ACM SIGOPS Operating Systems Review, vol. 41, no. 6. ACM, 2007, pp. 205–220.
[13] Y. Lin, D. Agrawal, C. Chen, B. Ooi, and S. Wu, "Llama: leveraging columnar storage for scalable join processing in the mapreduce framework," in Proceedings of the 2011 international conference on Management of data. ACM, 2011, pp. 961–972.
[14] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L. Youseff, and D. Zagorodnov, "The eucalyptus open-source cloud-computing system," in Cluster Computing and the Grid, 2009. CCGRID'09. 9th IEEE/ACM International Symposium on. IEEE, 2009, pp. 124–131.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY