International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Big Data using Cloud Computing: Security Issues and Approaches

Darshan.K.N[1] , Dr.Samitha.Khaiyum[2]

Post Graduate Student, Department of M.C.A, Dayananda Sagar College Of Engineering, Bangalore, India
Head of Department (HOD), Department of M.C.A, Dayananda Sagar College Of Engineering, Bangalore, India

---

Abstract - Big Data is a data analysis methodology made possible by current technological and architectural advancements. However, big data requires a significant investment in hardware and processing power, making it prohibitively expensive for small and medium-sized firms to implement. Small and medium-sized enterprises can benefit from cloud computing promise of big data implementation. The MapReduce programming paradigm is used to process large amounts of data. The MapReduce paradigm typically necessitates networked connected storage and parallel computation. The computing requirements of MapReduce programming are frequently beyond the capabilities of small and medium-sized businesses. Cloud computing is network access to computing resources on demand, given by a third party Hardware cost reduction, processing cost reduction, and the flexibility to verify the value of big data are three significant reasons for small and medium-sized organizations to embrace cloud computing for big data technology implementation. Security and loss of control are two main concerns with cloud computing.

Key Words: Big data, MapReduce, cloud computing, private cloud, public cloud, and hybrid cloud.

1.INTRODUCTION

Big Data is a data-analytics methodology made possible by a new generation of technologies and architecture that enable high-speed data capture, storage, and analysis. Email, mobile device output, sensor-generated data, and social media output are all examples of data sources that go beyond the standard corporate database. Data is no longer limited to organized database records, but also includes unstructured data, or information that does not follow a set of rules. Big Data needs a lots of storage space. While the cost of storage has decreased, the resources required to harness big data might still be prohibitively expensive for small and medium-sized organizations. Clustered network-attached storage will be the foundation of a typical big data storage and analysis architecture (NAS)[1]. The NAS devices are then linked together to allow for enormous data sharing and searching. For small to medium-sized organizations exploring Big Data analytic approaches, cloud computing data storage is a potential choice. Cloud computing refers to on-demand network access to computing resources that are typically provided by a third party and require little managerial effort on the part of the organization. Cloud computing has a variety of designs and deployment models, and these architectures and models can be combined with other technologies and design techniques[3]. Small and medium-sized business owners that can't afford clustered NAS equipment can look into a variety of cloud computing options to satisfy their big data needs. To be competitive and profitable, small and medium-sized businesses must choose the right cloud computing solution.

2. CLOUD COMPUTING

Cloud computing is a method of handling applications that relies on the sharing of computer resources rather than local servers or personal devices. Cloud Computing refers to a sort of computing in which services are provided through the Internet because "Cloud" stands for "Internet.". Cloud computing aims to take use of increasing computer capacity to handle millions of instructions per second[1]. Cloud computing distributes data processing among a large number of servers over networks with specialised connections. Instead of installing software on each computer, this method necessitates the installation of a single piece of software on each machine that lets users to log into a Web-based service and hosts all of the user's programs. In a cloud computing system, there is a major workload shift. When it comes to executing apps, local PCs no longer have to bear the entire strain. PaaS(Platform as a service),SaaS(software as a service), IaaS(infrastructure as a service), and HaaS(hardware as a service) are some of the most common cloud computing deployment methods. The public cloud, private cloud, and hybrid cloud are the three types of cloud computing. Pay-as-you-go services are available in the public cloud. A private cloud is an organization's internal data centre that is not available to the public but is built on cloud architecture. The hybrid cloud is a mix of both public and private cloud services[4]. Cloud computing technology is being used to reduce computing resource utilisation costs.

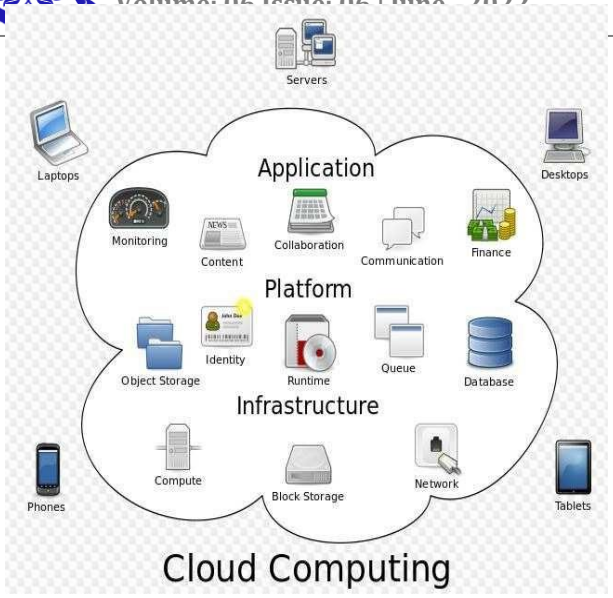

Fig -1: Cloud Computing.

Instead, the burden is handled by a cloud network, which is made up of a collection of computers. On the user end, the cost of software and hardware falls. To connect to the cloud platform, the only thing the user needs to do is to execute the cloud interface program. A front end and a back end make up cloud computing. The front end includes the user's computer and any programmes needed to connect to the cloud network. The cloud's back end is made up of multiple computers, servers, and database systems[4]. By connecting to the cloud over the Internet, the user can access apps in the cloud network from anywhere. Gmail, Google Calendar, Google Docs, and Dropbox are examples of real-time apps that leverage Cloud Computing.

3. BIG DATA

Big Data is the word used to describe huge volumes of unstructured and structured data that are so large that is very hard to process this data using traditional databases and software technologies[5]. The term Big Data is emerged from the Web search companies who had to query loosely structured very huge distributed data. The Big Data have the properties: volume, variety, velocity, variability and complexity.

3.1. HADOOP

Hadoop is a free Java-based programming platform that allows massive data sets to be processed in a distributed computing environment. It is part of the Apache Software Foundation-sponsored Apache project. A Master/Slave structure is used in a Hadoop cluster. Large data sets can be processed over a cluster of computers using Hadoop[10], and applications can be run on systems with thousands of nodes and terabytes of data. Hadoop's distributed file system aids in quick data transfer rates and allows the system to continue operating normally even if some nodes fail. Even when a large number of nodes fail, this strategy reduces the danger of the entire system failing. Hadoop allows for a scalable, cost-effective, adaptable, and fault-tolerant computing solution. Hadoop Framework is used by well-known companies like Google, Yahoo, Amazon, and IBM to support their data-intensive applications. Map Reduce and the Hadoop Distributed File System (HDFS) are Hadoop's two main modules.

3.2. MAPREDUCE

Hadoop Map Reduce is a framework for developing applications that process massive amounts of data in parallel on commodity hardware in a reliable[11] and fault-tolerant manner. A Map Reduce job splits the data into smaller distinct chunks, which are then processed in parallel by Map jobs. The outputs of the framework-sorted maps are then fed into the reduction tasks. In most of the cases, the job's input and output are both kept in a file system. The framework handles monitoring, scheduling and re-executing failed tasks.

3.3. HDFS

For data storage, Hadoop Distributed File System is a file system that covers all nodes in a Hadoop cluster. It connects file systems[11] on the local nodes to form a single big file system. To combat node failures, HDFS improves dependability by replicating data across numerous sources.

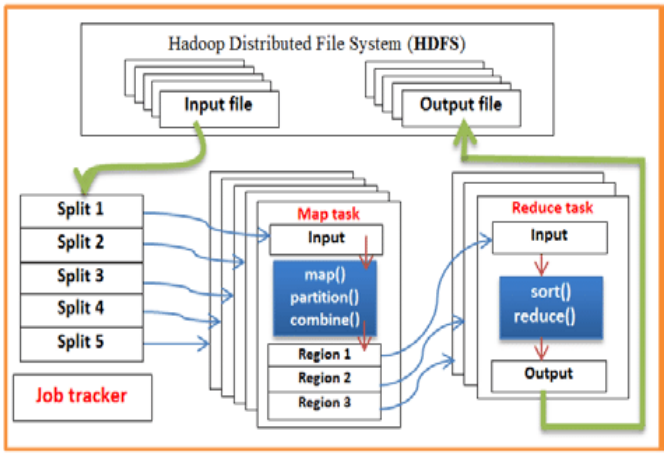

Fig -2: Hadoop Framework.

4. RELATIONSHIP BETWEEN THE CLOUD COMPUTING AND BIG DATA

Cloud computing and Big data both are linked with each other. Big data enables its users to conduct distributed queries over various different datasets using commodity computing and then provide result sets in a timely way. Hadoop is one of the type of distributed data-processing platform, that offers the fundamental engine in cloud computing. Large data sources from the web and the cloud are stored in a distributed fault-tolerant database and processed in a cluster using a large datasets programming and parallel distributed algorithms. The basic goal of data visualisation is to have the analytical results displayed visually through various graphs in order to make relevant decisions.

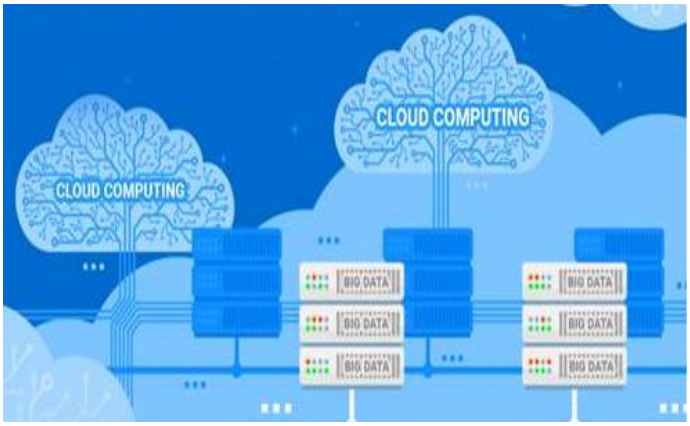

Fig -2: Relationship between Cloud Computing and Big Data.

5. ISSUES AND CHALLENGES

Data security entails not only the encryption of data, but also the implementation of proper data sharing protocols. Algorithms for memory management and resource allocation must be secure. Certain businesses, such as telecommunications, web marketing and advertising, retail and financial services, and government activities, are particularly affected by challenges of the big data.

1. **Distributed Nodes**
   Distributed nodes are a design problem. Any number of nodes can be used to perform the computation. In general, data is processed in nodes that have the requisite resources.[9] It's tough to pinpoint the actual place of computation because it can happen anywhere across the clusters. As a result, ensuring the security of the computing environment is quite complex.

2. **Internode Communication**
   For user data/operational data transfer between nodes, many Hadoop distributions employ RPC via TCP/IP. This takes place over a global network [9] that includes both wireless and wired networks. As a result, anyone can tap into and manipulate inter-node traffic in order to hack into systems
3. **Data Protection**
   To boost efficiency, many cloud environments, such as Hadoop, store data without encryption. If a hacker gains access to a group of machines, there is no way to prevent him from stealing the sensitive information stored on those devices.
4. **Administrative Rights for Nodes**
   A node can access any data and has administrative rights[9]. Because a rogue node can steal or manipulate important user data, having unrestricted access to any data is extremely risky.
5. **Authentication of Applications and Nodes**
   To boost the number of parallel operations, nodes might join clusters. Third-party nodes can join the clusters without any authentication and steal the secred user data or disrupt cluster operations.

6. PROPOSED APPROACHES

Various security measures are presented that would increase the security of the cloud computing environment. Because the cloud environment is made up of a variety of technologies, we present a number of solutions that, when combined, will make the environment secure. To address the security issue mentioned in the preceding sections, the

recommended solutions advocate the employment of different technologies/tools.

1. **File Encryption**
   A hacker can take all of the vital information because the data is stored on the machines in a cluster. As a result, all stored data should be encrypted. On separate machines, different encryption keys should be used, and the key information should be held centrally behind robust firewalls. Even if a hacker obtains the data, he will be unable to extract useful information from it and misuse it. User information will be encrypted and is stored safely from attackers.
2. **Network Encryption**
   According to industry requirements, every network communication should be encrypted[9]. Even if a hacker can tap into network communication packets, he won't be able to extract useful information or change packets if RPC procedure calls are made via SSL.
3. **Logging**
   All map reduction jobs that change the data should be tracked. Users who are responsible for certain jobs should also have their information logged. These logs should be inspected on a regular basis to see if any malicious operations or users are changing the data in the nodes.
4. **Software Format and Node Maintenance**
   Nodes that run the software must be formatted on a regular basis to remove any kind viruses. To make the system more safe from viruses[9], the application software and the Hadoop software should be upgraded time to time.
5. **Nodes Authentication**
   A node should be authenticated whenever it joins a cluster. It should not be permitted to join the cluster if it is malicious. Authentication mechanisms such as Kerberos can be used to distinguish between legitimate and malicious nodes.

7. CONCLUSION

Small and medium-sized businesses can use cloud computing to utilise big data technology while committing fewer resources. The big data model's processing capabilities could bring fresh insights to the business in terms of performance improvement, decision support, and business model, product, and service innovation. Cost savings in hardware and resources processing, as well as the freedom to experiment with big data technology before committing significant company resources, are all the advantages of integrating big data technology through cloud computing. Businesses can choose from several cloud computing service models, each with trade-offs between the benefits of cost reductions and worries about data security and loss of control.

REFERENCES

[1] Ren, Yulong, and Wen Tang. "A SERVICE INTEGRITY ASSURANCE FRAMEWORK FOR CLOUD COMPUTING BASED ON MAPREDUCE."*Proceedings of IEEE CCIS2012*. Hangzhou: 2012, pp 240 – 244, Oct. 30 2012-Nov. 1 2012.

[2] N, Gonzalez, Miers C, Redigolo F, Carvalho T, Simplicio M, de Sousa G.T, and Pourzandi M. "A Quantitative Analysis of Current Security Concerns and Solutions for Cloud Computing.". Athens: 2011., pp 231 – 238, Nov. 29 2011- Dec. 1 2011.

[3] Carraro, G., & Chong, F. (2006, October). Software as a service: An enterprise perspective. Retrieved from http://msdn.microsoft.com/en-us/library/aa905332.aspx #enterprisertw_topic3

[4] Aslam, U., Ullah, I, & Ansara, S. (2010, November). Open source private cloud computing. Interdisciplinary Journal of Contemporary Research in Business. 2(7), 399-407.

[5] Hao, Chen, and Ying Qiao. "Research of Cloud Computing based on the Hadoop platform.". Chengdu, China: 2011, pp. 181 – 184, 21-23 Oct 2011.

[6] Y, Amanatullah, Ipung H.P., Juliandri A, and Lim C. "Toward cloud computing reference architecture: Cloud service management perspective.". Jakarta: 2013, pp. 1-4, 13-14 Jun. 2013.

[7] A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices.". Noida: 2013, pp. 404 – 409, 8-10 Aug. 2013.

[8] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G…Zaharia, M. (2010, April). A view of cloud computing. Communications of the ACM, 53(4), 50-58. DOI: 10.1145/1721654.1721672

[9] "Securing Big Data: Security Recommendations for Hadoop and NoSQL Environments."Securosis blog, version 1.0 (2012)

[10] Lu, Huang, Ting-tin Hu, and Hai-shan Chen. "Research on Hadoop Cloud Computing Model and its Applications.". Hangzhou, China: 2012, pp. 59 – 63, 21-24 Oct. 2012.

[11] Wie, Jiang , Ravi V.T, and Agrawal G. "A Map-Reduce System with an Alternate API for Multi-core Environments.". Melbourne, VIC: 2010, pp. 84-93, 17-20 May. 2010.