'National Conference on Advancements in Computer Science and Engineering'
International Journal of Scientific Research in Computer Science,
Engineering and Information Technology | ISSN : 2456-3307 (www.ijsrcseit.com)

Big Data in Cloud Computing Security Issues

Prerana Pradeep, Prerna Ubana, Pruthvi Reddy V

Department of Computer Science and Engineering, New Horizon College of Engineering, Bangalore, Karnataka, India

ABSTRACT

Big Data and Cloud Computing are two of the most talked about topics over the past few years. They provide high efficiency and reliability along with strong security. Big Data is one such topic which researchers are trying their hand at to solve as it is complex as well as interesting. It gives us an idea as to how to make security perfect by safeguarding sensitive information and not allowing security breach to take place. As the adoption of cloud computing continues to grow, the management and analysis of large-scale data, commonly referred to as big data, have become critical for organizations. However, the integration of big data with cloud computing introduces significant security challenges that need to be addressed. This abstract provides an overview of the security issues associated with the utilization of big data in cloud computing environments. It explores the unique characteristics of big data and the vulnerabilities it introduces, including data breaches, unauthorized access, data privacy concerns, and the potential for insider threats. Furthermore, these paper highlights the implications of these security issues on cloud computing architectures, data storage, data processing, and data transmission. It also discusses current research efforts and best practices for mitigating security risks and protecting big data in cloud environments. By understanding and addressing these security challenges, organizations can confidently leverage the benefits of big data analytics while ensuring the confidentiality, integrity, and availability of their data assets in cloud computing environments.

I. INTRODUCTION

Based on research in the past, most of the structured or unstructured data and the size will double in every two years. By using standardization techniques, cloud computing can be commodified based on computing time and data storage. When we say "Big Data" we mean services such as Amazon.com, AT&T, GoGrid, Joyent, IBM etc. In present times, the most popular IaaS service provider is Amazon with its elastic cloud computing. Need for Private Cloud: As the volume of big data is continuously increasing along with the variety and velocity, the existing infrastructure must adapt to the requirements. To do so, private Cloud is introduced. This was done by Synnex Corporation powered by Nebula. As a result of this, a fully integrated private cloud system was engineered for MongoDB as well as Apache Cassandra. Private Clouds do not share resources whereas they are dedicated to a single organization, hence the name "Private". Need for Public Cloud: Unlike Private Clouds, Public Clouds do share resources for data transfer, processing and storage. However, Public Clouds are not preferred as there are issues regarding security. Need for Hybrid Cloud: This is a combination of both Private as well as Public Cloud. It has enhanced advantages of both clouds and minimized disadvantages of the same.

Security is another major and important issue which is the talk of the town. There are basically 3 V's of security in big data: Variety, Volume and Velocity. The security used by Big Data is as: Vormetric Encryption- It secures Bid Data at volume and file system level. Data Security Platform- This includes very strong encryption, fine- grained access control and security intelligence. Encryption and Key Management- Data breach requires encryption to safeguard data. Fine- Grained Access Controls- Vormetric provides fine grained access controls. It selectively allows permission to authorized users. The challenges faced by Big Data has evolved into: Modelling, Analysis and Implementation.

II. METHODS AND MATERIAL

Cloud computing: Cloud computing is a model of delivering on-demand computing services, including servers, storage, databases, networking, software, analytics, and intelligence, over the internet ("the cloud"). It allows individuals and organizations to access computing resources from anywhere with an internet connection without having to invest in costly hardware and infrastructure.

Cloud computing services are offered by cloud providers, such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud, who manage the underlying hardware, software, and security required to provide the service. Users can choose to use these services on a pay- per-use basis or on a subscription basis, which provides them with flexibility and cost savings compared to traditionalIT infrastructure.

There are three main types of cloud computing services: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). SaaS provides users with access to software applications over the internet, while PaaS provides users with a platform to develop, run, and manage their own software applications. IaaS provides users with access to virtualized computing resources, such as servers, storage, and networking, on a pay-per-use basis.

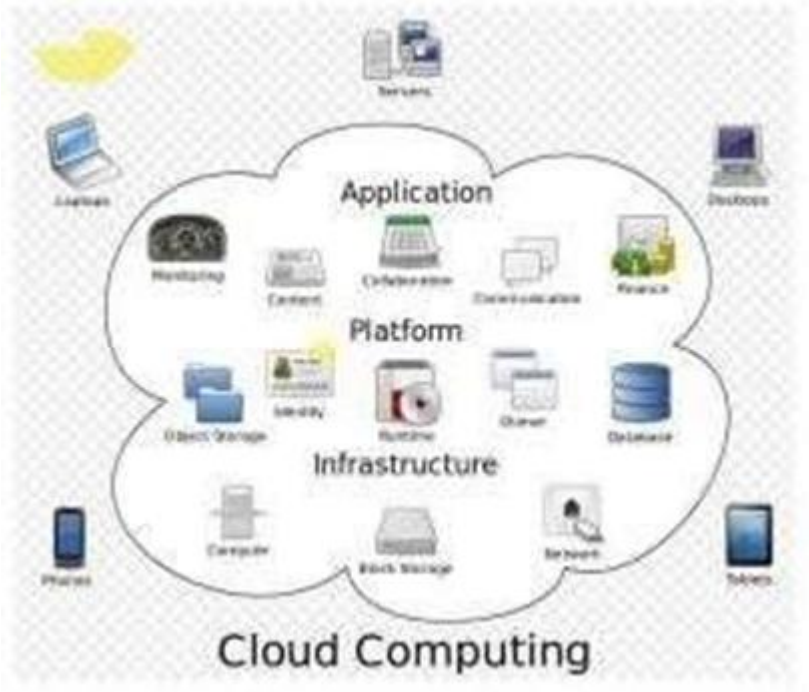

Figure 1: Cloud Computing

Big data: It refers to extremely large and complex data sets that are difficult to process using traditional data processing techniques. These data sets are characterized by their volume, velocity, variety, and veracity. Volume refers to the sheer size of the data, which can range from terabytes to petabytes and beyond.

Velocity refers to the speed at which data is generated and needs to be processed. Variety refers to the diverse formats and types of data, such as text, images, video, and sensor data. Veracity refers to the quality and reliability of the data.

Big data is generated by a wide range of sources, such as social media, sensors, transactions, and machine logs. It is often used in business, healthcare, finance, and other fields to gain insights and make data-driven decisions. To process big data, specialized tools and techniques such as Hadoop, Spark, and NoSQL databases are used.

Figure 2: Big data

1. MapReduce: is a programming model and software framework that is used to process and generate large amounts of data in a distributed and parallel computing environment. It was originally developed by Google, and is now widely used in big data processing applications.
   The MapReduce model consists of two main phases: the map phase and the reduces phase. In the map phase, the input data is divided into smaller chunks and processed by multiple nodes in parallel. Each node performs a specified operation on the data, which produces a set of key-value pairs. In the reduce phase, the key-value pairs generated by the map phase are aggregated based on the keys, and a final output is produced. The MapReduce framework provides automatic parallelization and fault tolerance, making it suitable for processing large datasets on distributed systems. It has been used in a wide range of applications, including web indexing, data mining, machine learning, and scientific computing.
2. Hadoop Distributed System: Hadoop is an open- source distributed processing framework that allows users to store and process large data sets across clusters of commodity hardware. It is designed to handle big data, which refers to data sets that are too large and complex to be processed using traditional data processing methods.

Together, these components allow Hadoop to process large amounts of data in parallel across clusters of commodity hardware, making it a popular tool for big data processing and analysis.

III. MATERIALS

1. Cloud Computing Infrastructure: It refers to the collection of hardware, software, and networking resources that are used to support cloud computing services. These resources are typically provided by cloud service providers, such as Amazon Web Services, Microsoft Azure, or Google Cloud Platform.

IV. RESULTS AND DISCUSSION

A. Big Data in Cloud Computing:

Cloud computing and big data are two complementary technologies that have revolutionized the way organizations store, process, and analyze large amounts of data. Cloud computing refers to the delivery of computing resources, including hardware, software, and storage, over the internet. It allows users to access computing resources on-demand and pay only for what they use, making it a cost- effective solution for managing big data. Big data refers to the large and complex data sets that are difficult to process using traditional data processing methods. It includes both structured and unstructured data from various sources, including social media, IoT devices, and sensors. In the context of cloud computing, big data can be stored and processed using cloud-based services such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP). These services provide scalable and flexible storage and processing capabilities, allowing organizations to store and process large amounts of data without having to invest in expensive hardware andinfrastructure.

Cloud-based big data services also offer advanced analytics capabilities, including machine learning and artificial intelligence, which enable organizations to gain insights and make data- driven decisions.

In summary, cloud computing provides a scalable and cost-effective platform for storing and processing big data, while big data analytics enables organizations to extract valuable insights and make data-driven decisions.

B. Future of Big Data in Cloud Computing:

The future of big data in cloud computing is promising, as both technologies continue to evolve and innovate, enabling organizations to store, process, and analyze increasingly large and complex data sets. Some of the key trends that are expected to shape the future of big data in cloud computing include:

a). Edge computing: As the number of IoT devices and sensors continues to grow, there is a growing need for processing data at the edge of the network, closer to where the data is generated. Cloud providers are expected to offer more edge computing services to support this trend.

b). Hybrid cloud: Many organizations are adopting a hybrid cloud approach, which combines public and private cloud services, to balance the benefits of public cloud with the security and control of private cloud. Big data solutions that can span both public and private clouds are likely to become more popular.

c). AI and machine learning: Big data analytics powered by AI and machine learning will continue to grow in importance, as organizations seek to extract valuable insights from their data. Cloud providers will continue to invest in AI and machine learning services to support this trend.

d). Data governance and privacy: As data privacy and governance regulations become more stringent, cloud providers will need to offer more advanced data governance and privacy features to support compliance.

e). Cloud-native big data: Cloud-native architectures, which are designed to take full advantage of cloud computing infrastructure, are likely to become more prevalent in big data solutions. This will enable faster and more scalable processing of big data in the cloud.

Overall, the future of big data in cloud computing is bright, as organizations continue to rely on these technologies to gain insights and make data-driven decisions. Cloud providers will continue to innovate and improve their big data services to meet the evolving needs of their customers.

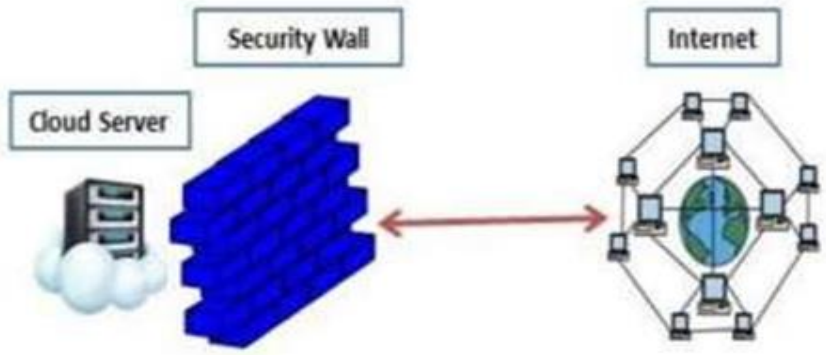

Figure 3: Security wall for cloud server

C. Cloud Computing Security:

It is not accurate to say that security in cloud computing is inherently bad, as cloud providers invest significant resources in security to protect their infrastructure and their customers' data. However, there are several reasons why security breaches can occur in cloud computing:

Here, we classified the challenges we face in big data in cloud computing and the challenges that we face in the infrastructure and in general.

Figure 4: Security issues in big data and cloud computing.

- Challenge in network: As there are so many nodes connected, distributed node connected this cause threats in keeping track which device is connected to which one and is it secure.
- Challenge in authentication: It is hard to keep track of every user that uses these services as there are so many users using those services and sometimes, we give some suspicious users more rights.
- Challenge in data: Data security and integration and availability is the most crucial part so we have to more concern about.

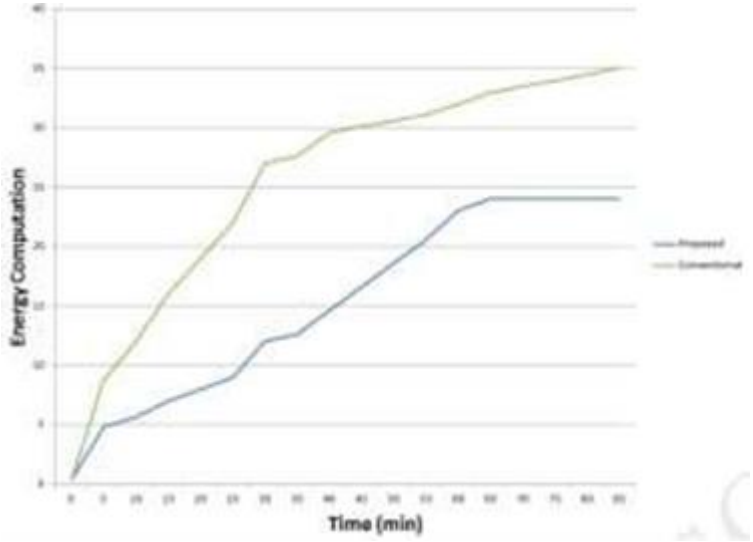

Figure 5: Energy Efficiency Comparison.

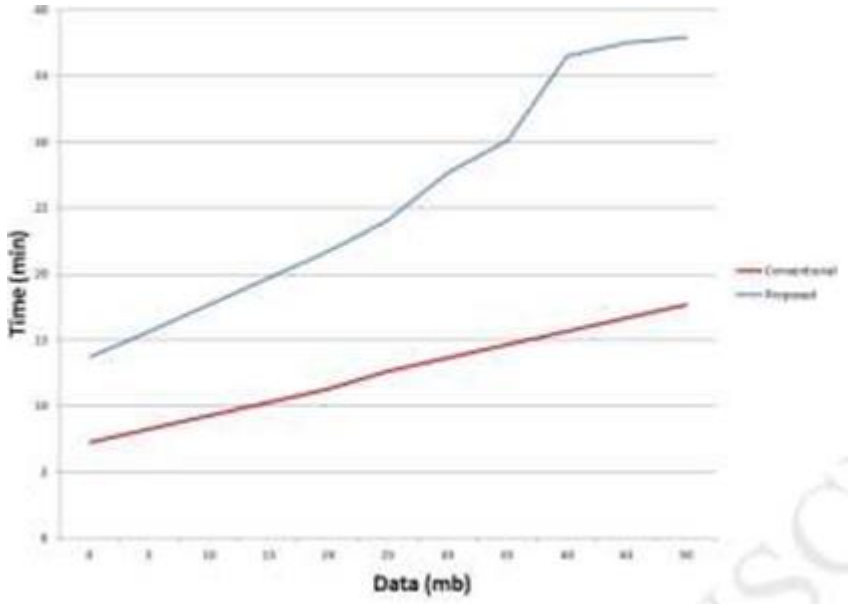

Figure 6: Data Transmission through Time.

The above discussed points are the border view of seeing the need of advancing the security so, now we should be discussed in more details about the security related issues. In the above points we mentioned about network so as in distributed network data is transferred through these nodes to various ways and to various users so it is quite hard to find the exact location that where it is getting used and where it is going.

In distributed data users' data is stored on various machine and reductant copies of it so in time of failure of system we can access it easily but it is very hard to find that where that particular file is located in which

system as it keeps on changing. In transferring the information from one node to another Hadoop distributions use RPC over TCP/IP so that no one can tap or modify the inner node communication. Data protection is must in it but in Hadoop doesn't encrypt the data to store it as it improves efficiencybut hackers can easily misuse it if they get access of it. The uncontrolled access if we give it to someone it is very dangerous as anyone can steal or manipulate the data so we must be careful with administrative rights. If we don't have proper authentication for accessing any third party can enter and steal or manipulate the data. Logging is the major issue when there is no activity in the node still, we can modify the data or delete it and we don't have a proper log system to keep the track of the people who made changes in the node then it will be very difficult to find who made changes in the node. So, if any malicious activity happened then it is very hard to track who did that. The old traditional methods we can't implement directly here in advance data managing system we have to use different technologies to handle this vast area of cloud computing infrastructure.

Big data applications	Security	Connectivity	Performance	Latency (Delay)	Privacy
Predictable & efficient	X		X	X	X
Holistic network	X	X	X		X
Network partitioning	X		X	X	X
Scale out	X	X			X
Unified ethernet fabrics	X	X	X		X

Possible ways to handle security challenges. As we discussed in the above section the challenges, we are facing in the security there are quite a few ways we can increase the security To mitigate these risks, it is essential to implement proper security controls, such as access controls, encryption, vulnerability management, and incident response planning. Cloud providers and customers must work together to ensure that security is a top priority in cloud computing. Since data is present in cluster a hacker can steal it so in order to keep our data safe, we must keep our data in encrypted way and the key for decrypting should be stored behind a strong wall of firebase so if a hacker access the file it won't be useful to them so file encryption is must. The network through which the communication is happening between the nodes must be secure and encrypted as per industry standards should happen over SSL so no one can tap it. Node maintained should be there means that means on a regular basis the software must be get reset so that if previously any virus is present, it will get remove after the reset.

Map reduce based work must be logged in order to keep track of the users which accessed the resources and modify or change it so it will be easy to track back if any malicious activity happen. Node Authentication is must as data is in form of clusters so if any malicious activity is detected though the node it can block it there only it won't allow to communicate or enter.

After we are done developing the map reduce, we have to test it thoroughly and we should make honeypot nodes so if any hackers try to access, they will get trap in this node and through this node wecan get to know about it and we can block those users. Third party security is needed so that it will be easy for doing the authentication of the users who tried to access the data and the cloud computing will be easy to maintain and be more secure about it as it clouds computing stores the data in remote location so we need third party to manage the authentication which store sensitive data. We should keep security administrator separate from the database administrator should support third party authentication to prevent the data leak and corrupting of

data these days we use SELinux [17] will be used. SELinux is nothing but Security- Enhanced Linux, it provides access control security policy and we should keep auditing so we can detect the error in early stage only.

V. CONCLUSION

Cloud computing and big data is a very vast and a growing area as many areas are becoming digital so more data is generated on a regular bases so it is hard to keep the data secure without reducing the performance. So, here we discussed about the different challenges we are facing in securing data in cloud computing and then we discussed about what are possible solution we can give for the discussed problem and making cloud environment more secure for complex operations.

VI. REFERENCES

[1]. Intel IT center, "Peer Research Big Data Analytics ", Intel's IT Manager Survey on How Organizations Are Using Big Data, AUGUST 2012.

[2]. http://www.informationweek.com/big-data/bigdataanalytics/big-databrings-big-security-problems/d/did/1252747.

[3]. Vahid Ashktorab1 , Seyed Reza Taghizadeh2 and Dr. Kamran Zamanifar3 ," A Survey on Cloud Computing and Current Solution Providers", International Journal of Application or Innovation in Engineering & Management (IJAIEM), October 2012.

[4]. 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.

[5]. 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.