Security with Big Data in Cloud Computing

M. Ramya M.Sc., M.Phil., P. Sonali, N. Naga Dharshini

1 Assistant Professor, Department of Computer Science, Mangayarkarasi College of Arts and Science For Women
2 Department of Computer Science, Mangayarkarasi College of Arts and Science For Women
3 Department of Computer Science, Mangayarkarasi College of Arts and Science For Women

Abstract

In recent years, big data and cloud computing are the major issues in an organization. It enables computing resources to be provided as information technology service with high efficiency and effectiveness. Today’s big data is the one of the major problem that researchers try to solve it. The main focus is on security issues in cloud computing that are associated with big data. Big data applications are a great benefit to organizations, business, companies and many large scale and small scale industries. The another important issue is how to gain a perfect security for big data in cloud computing. The possible solutions for the issues in cloud computing security and Hadoop. Cloud computing plays a very vital role in protecting data, applications and the related infrastructure with the help of policies, technologies, controls, and big data tools. Cloud computing, big data and its applications, advantages are likely to represent the most promising new frontiers in science[1].

Keywords: Cloud computing, Big data, Hadoop, Map reduce, HDFS, Big data security, CSA.

Introduction

Big data is the term for data sets, so large and complicated that it becomes difficult to process using traditional data management tools or processing applications. The data and to identify patterns it is very important to securely store, manage and share large amounts of complex data. Cloud comes with an explicit security challenge, i.e. the data owner might not have any control of where the data is placed. Hadoop is an open-source implementation of Google MapReduce, including a distributed file system, provides to the application programmer
the abstraction of the map and the reduce. With Hadoop it is easier for organizations to get a grip on the large volumes of data being generated each day, but at the same time can also create problems related to security, data access, monitoring, high availability and business continuity. Google has introduced MapReduce framework for processing large amounts of data on commodity hardware. Apache’s Hadoop distributed file system (HDFS) is evolving as a superior software component for cloud computing combined along with integrated parts such as Map Reduce[2]. Cloud Security Alliance (CSA) model outlines where the data is being processed and stored, and includes the big data sources, processing clusters and endpoint consumers of the data (systems, mobile devices, etc.), along with the cloud environments. Security and privacy issues are magnified by velocity, volume, and verity of big data, such as large scale cloud infrastructures, diversity of data sources and formats, streaming nature of data acquisition and high volume inter-cloud migration.

[2] BIG DATA

Big Data is the word used to describe massive volumes of structured and unstructured data that are so large that it is very difficult to process this data using traditional databases and software technologies. The term “Big Data” is believed to be originated from the Web search companies who had to query loosely structured very large distributed data. The three main terms that signify Big Data have the following properties:

### Figure 1 5 Vs Of Big data

![5 Vs Of Big data](image)

- **Volume**: Many factors contribute towards increasing Volume streaming data and data collected from sensors etc.

- **Variety**: Today data comes in all types of formats emails, video, audio, transactions etc.

- **Velocity**: This means how fast the data is being produced and how fast the data needs to be processed to meet the demand.
The other three dimensions that need to consider with respect to Big Data are Veracity, Variability and Complexity.

d) Variability: Along with the Velocity, the d peaks. This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

e) Veracity: The quality of the data being captured can vary greatly. Accuracy of analysis depends on the veracity of the source data.

f) Complexity: Complexity of the data also needs to be considered when the data is coming from multiple sources. The data must be linked, matched, cleansed and transformed into required formats before actual processing.

When making an attempt to understand the concept of Big Data, the words and “Hadoop” cannot be avoided.

[3] HADOOP

Hadoop, which is a free, Java-based programming framework supports the processing of large sets of data in a distributed computing environment. It is a part of the Apache project sponsored by the Apache Software Foundation. Hadoop cluster uses a Master/Slave structure. Using Hadoop, large data sets can be processed across a cluster of servers and applications can be run on systems with thousands of nodes involving thousands of terabytes. Distributed file system in Hadoop helps in rapid data transfer rates and allows the system to continue its normal operation even in the case of some node failures. This approach lowers the risk of a system failure, even in the case of a significant number of node failures. Hadoop enables a computing solution that is scalable, cost effective, flexible and fault tolerant. Hadoop Framework is used by popular companies like Google, Yahoo, Amazon and IBM etc., to support their applications involving huge amounts of data. Hadoop has two main sub projects – Map Reduce and Hadoop Distributed File System (HDFS).

[4] CLOUD ISSUES AND CHALLENGES

Cloud computing comes with numerous security issues because it encompasses many technologies including networks, databases, operating systems, virtualization, resource scheduling, transaction management, load balancing, concurrency control and memory management. Hence, security issues of these systems and technologies are applicable to cloud computing. For example, it is very important for the network which interconnects the systems in a cloud to be secure. Also, virtualization paradigm in cloud computing results in several security concerns. Data security not only involves the encryption of the data, but also ensures that appropriate policies are enforced for data sharing. In addition, resource allocation and memory
management algorithms also have to be secure. The big data issues are most acutely felt in certain industries, such as telecoms, web marketing and advertising, retail and financial services, and certain government activities. The data explosion is going to make life difficult in many industries, and the companies will gain considerable advantage which is capable to adapt well and gain the ability to analyze such data explosions over those other companies. Finally, data mining techniques can be used in the malware detection in clouds. The challenges of security in cloud computing environments can be categorized into network level, user authentication level, data level, and generic issues.

Network level: The challenges that can be categorized under a network level deal with network protocols and network security, such as distributed nodes, distributed data, inter node communication.

Authentication level: The challenges that can be categorized under user authentication level deals with encryption/decryption techniques, authentication methods such as administrative rights for nodes, authentication of applications and nodes, and logging.

Data level: The challenges that can be categorized under data level deals with data integrity and availability such as data protection and distributed data.

Generic types: The challenges that can be categorized under general level are traditional security tools, and use of different technologies.

[5] BIG DATA PRIVACY AND SECURITY

Big Data is one of the most talked about technology. But lost among all the excitement about the potential of Big Data are the very real security and privacy challenges that threaten to slow this momentum. Security and privacy issues are magnified by the three V’s of big data: Velocity, Volume, and Variety. These factors include variables such as large-scale cloud infrastructures, diversity of data sources and formats, streaming nature of data acquisition and the increasingly high volume of inter cloud migrations. Consequently, traditional security mechanisms, which are tailored to securing small-scale static data, often fall short[6]. The CSA’s Big Data Working Group followed a three step process to arrive at top security and privacy challenges presented by Big Data; Interviewed CSA members and surveyed security practitioner oriented trade journals to draft an initial list of high priority security and privacy problems studied published solutions.

Characterized a problem as a challenge if the proposed solution does not cover the problem scenarios. The Working Group researchers compiled their list of the Top 10 challenges as shown in figure 1.5 below. The Expanded Top 10 Big Data challenges have evolved from the initial list of challenges presented at CSA Congress to an expanded version that addresses three new distinct issues:
Modeling: formalizing a threat model that covers most of the cyber-attack or data-leakage scenarios.

Analysis: finding tractable solutions based on the threat model.

Implementation: implanting the solution in existing infrastructures.

Figure 2 Top 10 Challenges of CSA’s Big Data Working Group

The information security practitioners at the Cloud Security Alliance know that big data and analytics systems [7]. The group’s latest 10 major security and privacy challenges facing infrastructure providers and customers. By outlining the issues involved, along with analysis of internal and external threats and summaries of current approaches to mitigating those risks, the alliance’s members hope to prod technology vendors, academic researchers and practitioners to collaborate on computing techniques and business practices that reduce the risks associated with analyzing massive datasets using innovative data analytics.

Real-time system monitoring techniques that works well on smaller volumes of data but not very large datasets. The growing number of devices, from smartphones to sensors, producing data for analysis. General confusion “surrounding the diverse legal and policy restrictions that lead to ad hoc approaches for ensuring security and privacy[8]. Given the very large data sets that contribute to a Big Data implementations, there is a virtual certainty that either protected information or critical Intellectual Property (IP) will be present. This information is distributed throughout the Big Data implementation as needed with the result that the entire data storage layer needs security protection. The are many types of protection and security used such as.

Vormetric Encryption: seamlessly protects Big Data environments at the file system and volume level. This Big Data analytics security solution allows organizations to gain the benefits of the
intelligence gleaned from Big Data analytics while maintaining the security of their data – with no changes to operation of the application or to system operation or administration.

Data Security Platform: The Vormetric Data Security Platform secures critical data – placing the safeguards and access controls for your data with your data. The data security platform includes strong encryption, key management, fine-grained access controls and the security intelligence information needed to identify the latest in advanced persistent threats (APTs) and other security attacks on your data.

Encryption and Key Management: Data breach mitigation and compliance regimes require encryption to safeguard data. Vormetric provides the strong, centrally managed, encryption and key management that enables compliance and is transparent to processes, applications and users.

Fine-grained Access Controls: Vormetric provides the fine-grained, policy based access controls that restrict access to data that has been encrypted allowing only approved access to data by processes and users as required to meet strict compliance requirements. Privileged users of all types (including system, network and even cloud administrators) can see plaintext information only if specifically enabled to do so. System update and administrative processes continue to work freely – but see only encrypted data, not the plaintext source.

Security Intelligence: Vormetric logs capture all access attempts to protected data providing high value, security intelligence information that can be used with a Security Information and Event Management solution to identify compromised accounts and malicious insiders as well as finding access patterns by processes and users that may represent and APT attack in process. Use the Vormetric Toolkit to easily deploy, integrate and manage your Vormetric Data Security implementation with the rest of your big data implementation[10].

![Vormetric Concept Automation](image)

**Figure 1.6. Vormetric Concept Automation:**
CONCLUSION

Recently, researchers focusing their efforts in how to manage, handling and also processing the huge amount of data as known a Big data deals with three concepts volume, Variety and velocity which requires a new mechanisms to manage, processing, storing, analyzing and securing the big data. Its managing and processing of big data have many problems and required more efforts to handle these requirements when deal with big data, security is one of the challenges that arise when systems try to handle the concept of big data. More researches required to overcome the security of big data instead of current security algorithms and methods. Cloud environment is widely used in industry and research aspects; therefore security is an important aspect for organizations running on these cloud environments. Using proposed approaches, cloud environments can be secured for complex business operations. By leveraging the work of the CSA working group for big data and focusing explicitly on the key controls that should be in place, enterprise consumers can help to properly evaluate the state of big data infrastructure and applications in their service providers' environments.

REFERENCE


