IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact factor: 4.295

(Volume 3, Issue 6)

Available online at www.ijariit.com

A Survey on Security and Privacy Methods of Bigdata in Cloud Computing Environment

Soumyashree
Assistant Professor
HKBK College of Engineering, Bengaluru, Karnataka
soumya4041@gmail.com

Nithyashree K
Assistant Professor
HKBK College of Engineering, Bengaluru, Karnataka
nithyak08@gmail.com

Abstract - Big Data and cloud computing are two important issues in the recent years, that enable computing resources to be provided as Information Technology services with high efficiency and effectiveness. 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. Cloud computing plays a very vital role in protecting data. Hadoop Framework is used by popular companies like Google, Yahoo, Amazon and IBM etc. Now-a-days big data is one of the most serious problems that researchers have focused their research on it to understand the problem of how big data could be handled in the current systems and managed with cloud computing. One of the most important issues is how to gain perfect security for big data in cloud computing. Our paper makes a Survey of big data with cloud computing, security and the mechanisms that are used to protect and secure with available clouds.

Keywords: Cloud Computing, Big Data, Cloud Providers, NAS, Big Data Security, big data privacy. Hadoop, Distributed file System

1. INTRODUCTION

In order to analyze complex 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. The reason behind this control issue is that if one wants to get the benefits of cloud computing, he/she must also utilize the allocation of resources and also the scheduling given by the controls. Hence it is required to protect the data in the midst of untrustworthy processes. Since cloud involves extensive complexity, we believe that rather than providing a holistic solution to securing the cloud, it would be ideal to make noteworthy enhancements in securing the cloud that will ultimately provide us with maximum security. Google has introduced MapReduce framework for processing large amounts of data on commodity hardware like Apache’s Hadoop distributed file system (HDFS) in Big data.

1.1 Introduction to Cloud Computing

Cloud computing is the delivery of computing services—servers, storage, databases, networking, software, analytics and much more over the Internet (“the cloud”). Companies offering these computing services are called cloud providers and typically charge for cloud computing services based on usage, similar to how you are billed for water or electricity at home.

A. Benefits of Cloud Computing

Cloud computing offers a number of benefits, including the potential for:

- Rapid scalability and deployment capabilities (providing just-in-time computing power and infrastructure).
- Decreased maintenance/upgrades.
- Improved resource utilization—elasticity, flexibility, efficiencies.
- Improved economies of scale • Improved collaboration capabilities.
- Ability to engage in usage-based pricing, making computing a variable expense, rather than a fixed capital cost with high overhead.

- Reduced information technology (IT) infrastructure needs—both up-front and support costs.

B. Types of cloud deployments

There are three different ways to deploy cloud computing resources: public cloud, private cloud and hybrid cloud shown in figure1.1 below.

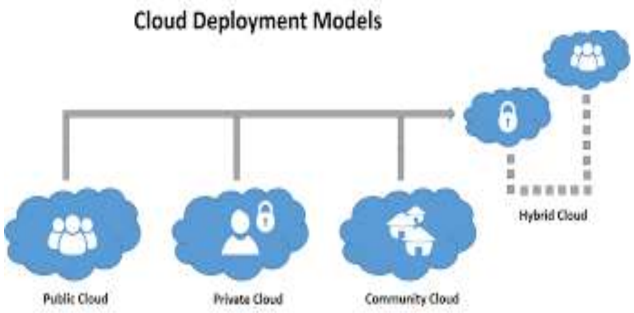

Fig 1.1: Deployments of Cloud

- **Public Cloud**

Public clouds are owned and operated by a third-party cloud service provider, which deliver their computing resources like servers and storage over the Internet. Microsoft Azure is an example of a public cloud. With a public cloud, all hardware, software and other supporting infrastructure is owned and managed by the cloud provider. These services can be accessed and account can be managed using a web browser.

- **Private Cloud**

A private cloud refers to cloud computing resources used exclusively by a single business or organization. A private cloud can be physically located on the company's on-site datacenter. Some companies also pay third-party service providers to host their private cloud. A private cloud is one in which the services and infrastructure are maintained on a private network.

- **Hybrid Cloud**

Hybrid clouds combine public and private clouds, bound together by technology that allows data and applications to be shared between them. By allowing data and applications to move between private and public clouds, hybrid cloud gives businesses greater flexibility and more deployment options.

C. Data Security Aspects on Cloud

Cloud computing security or, more simply, cloud security refers to a broad set of policies, technologies, and controls deployed to protect data, applications, and the associated infrastructure of cloud computing. It is a sub-domain of computer security, Network security and more broadly, information Security [1].

- **Confidentiality**

Data confidentiality is the property that data contents are not made available or disclosed to illegal users. Outsourced data is stored in a cloud and out of the owners' direct control. Only authorized users can access the sensitive data while others, including CSPs, should not gain any information of the data.

- **Access Controllability**

Access controllability means that a data owner can perform the selective restriction of access to his data outsourced to cloud. Legal users can be authorized by the owner to access the data, while others cannot access it without permissions.

- **Integrity**

Data integrity demands maintaining and assuring the accuracy and completeness of data. A data owner always expects that his data in a cloud can be stored correctly and trustworthily. It means that the data should not be illegally tampered, improperly modified, deliberately deleted, or maliciously fabricated. If any undesirable operations corrupt or delete the data, the owner should be able to detect the corruption or loss. Further, when a portion of the outsourced data is corrupted or lost, it can still be retrieved by the data users.

1.2 Big Data Introduction

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 [5]" is used by companies who handle query of loosely structured very large amount of distributed data.
The three main terms that generally signify Big Data are shown in fig 1.2.
i. Volume: This has to do with the amount of data generated on a daily basis which is so large and keeps increasing with time
ii. Variety: Today data is created in different type, form and formats such as emails, video, audio, transactions etc.
iii. Velocity: This has to do with the speed it takes to produce data and how fast this data produced needs to be processed on time to meet individual demand.

The other two properties that need to be critically considered when talking about Big Data are Variability and Complexity as depicted by [3].
i. Variability: This goes along with velocity, and it has to do with how inconsistent the flow of data can be with respect to time and how far it can go.
ii. Complexity: The complexity of the data must be considered especially when we have multiple sources of data. The data must be rearranged in a format that will be suitable for processing.

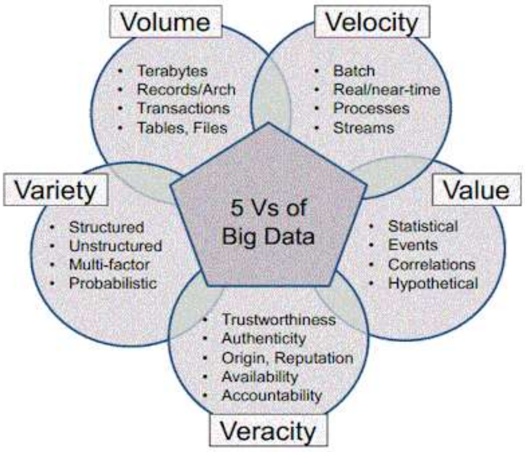

Fig 1.2: Five properties of data

1.3 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 [6]. 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 [2] 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 an entire system failure, even in the case of a significant number of node failures. Hadoop enables a computing solution that is scalable, cost effective, and flexible and fault tolerant. Hadoop Framework is used by popular companies like Google, Yahoo, Amazon, 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).

- **MapReduce**

Hadoop MapReduce is a framework [4] used to write applications that process large amounts of data in parallel on clusters of commodity hardware resources in a reliable, fault-tolerant manner. A Map Reduce job first divides the data into individual chunks which are processed by Map jobs in parallel. The outputs of the maps sorted by the framework are then input to the reduce tasks. Generally the input and the output of the job are both stored in a file-system. Scheduling, Monitoring and re-executing failed tasks are taken care by the framework.

- **Hadoop Distributed File System (HDFS)**

HDFS [8] is a file system that spans all the nodes in a Hadoop cluster for data storage. It links together file systems on local nodes to make it into one large file system. HDFS improves reliability by replicating data across multiple sources to overcome node failures

1.4 Big Data Applications

The big data application refers to the large scale distributed applications which usually work with large data sets. Data exploration and analysis turned into a difficult problem in many sectors in the span of big data. With large and complex data, computation becomes difficult to be handled by the traditional data processing applications which triggers the development of big data applications

- Google's map reduce framework and apache Hadoop are the software systems for big data applications, in which these applications generates a huge amount of intermediate data. Manufacturing and Bioinformatics are the two major areas of big data applications.
- Big data provide an infrastructure for transparency in manufacturing industry, which has the ability to unravel uncertainties such as inconsistent component performance and availability.
- In these big data applications, a conceptual framework of predictive manufacturing begins with data acquisition where there is a possibility to acquire different types of sensory data such as pressure, vibration, acoustics, voltage, current, and controller data
- The combination of sensory data and historical data constructs the big data in manufacturing. This generated big data from the above combination acts as the input into predictive tools and preventive strategies like health management can be incorporated.
- Another important application for Hadoop is Bioinformatics which covers the next generation sequencing and other biological domains. Bioinformatics which requires a large scale data analysis, uses Hadoop. Cloud computing gets the parallel distributed computing framework together with computer clusters and web interface.

2. RELATED WORK

Much work has been done on big data and cloud computing issues and challenges. To mention a few, [8] conducted a study on the security issues and challenges in the cloud and they looked at the advantages of using cloud services which include scalability, resilience, flexibility, efficiency and outsourcing non-core activities. Furthermore, they highlighted that Cloud computing offers an innovative business model for organizations to adopt IT services without upfront investment. Despite the potential gains of using cloud services, organizations are slow in accepting it due to security issues and challenges associated with it. Security is one of the major issues which hamper the growth of cloud and the main issue is lack of trust from both party because the idea of handing over important data to another company is worrisome to the subscribers of cloud services; such that the consumers need to be vigilant in understanding the risks of data breaches in this new environment.

[9] Looked at the security issues for cloud computing, big data, Map Reduce and Hadoop environment. The main focus is on security issues in cloud computing looking at how important data is for effective running of any business, therefore cloud computing security is developing at a rapid pace which includes computer security, network security, information security, and data privacy. 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. Moreover, in cloud computing, big data and its applications, advantages are likely to represent the most promising new frontiers in science; it is therefore necessary to protect such important infrastructure.

[10] carried out a study on the awareness and adoption of cloud computing in Nigeria by small and medium scale enterprises in Lagos State and the found out that of all the challenges facing cloud computing adoption are security, privacy, lack of liability of providers in case of security incidents, and difficulty of migrating to the cloud (legacy software) that ranked high. [15] Carried out a study on big data and current cloud computing issues and challenges by looking at a detailed analysis of between big data and cloud computing security issues and challenges focusing on the cloud computing types and the service delivery types. However, the researcher highlighted that big data entails a huge commitment of hardware and processing resources, thereby making adoption costs of big data technology unaffordable to small and medium sized businesses. Cloud computing on the other hand offers a lot of advantages to small and medium business especially but yet the rate of its adoption is still low due to security threat associated with cloud technology [17]. From all the related works reviewed, one paramount problem is the issue of security associated with big data technology and cloud computing which cannot be overemphasized. Hence this paper will critically examine this issue with other related issues associated with big data and cloud computing.

3. BIG DATA SECURITY CHALLENGES AND ISSUES.

1. Most distributed systems' computations have only a single level of protection, which is not recommended.
2. Non-relational databases (NoSQL) are actively evolving, making it difficult for security solutions to keep up with demand.
3. Automated data transfer requires additional security measures, which are often not available.
4. When a system receives a large amount of information, it should be validated to remain trustworthy and accurate; this practice doesn't always occur, however.
5. Unethical IT specialists practicing information mining can gather personal data without asking users for permission or notifying them.
6. Access control encryption and connections security can become dated and inaccessible to the IT specialists who rely on it.
7. Some organizations cannot – or do not – institute access controls to divide the level of confidentiality within the company. Recommended detailed audits are not routinely performed on Big Data due to the huge amount of information involved.
8. Due to the size of Big Data, its origins are not consistently monitored and tracked.

The challenge of detecting and preventing [3] advanced threats and malicious intruders, must be solved using big data style analysis. These techniques help in detecting the threats in the early stages using more sophisticated pattern analysis and analyzing multiple data sources.

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, Internodes communication.

Authentication level: The challenges that can be categorized under user authentication level deals with encryption and 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.

4. PRIVACY ENCRYPTION STRATEGY

We concentrate on privacy issues and propose a novel data encryption approach, named as *Dynamic Data Encryption Strategy* (D2ES). High Level Architecture of Dynamic Data Encryption model which is designed to protect data owners' privacy at the highest level when using the applicable devices and networking facilities. Fig.4 shows the high level architecture of D2ES model, which illustrates the main procedures and techniques [11].

Two major techniques used in D2ES are:

(1) Classifying data packages according to privacy level and (2) determine whether data packages can be encrypted under the timing constraints.

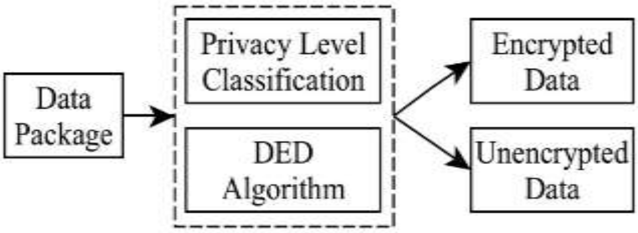

Fig 4.1: High Level Architecture of Dynamic Data Encryption Strategy (D2ES) Model

Our research is significant for generating an adoptive solution to protecting data owners' privacy. Main contributions of this work are twofold:
1) Encrypts data packages to maximize the privacy protection level under certain constraints.
2) The proposed algorithm offers an optimal solution providing the maximum value of total privacy weights.
3) The findings of this research provide big data-based solutions with an adaptive approach of protecting privacy. The proposed method can be also implemented in distributed storages in cloud computing.

4.1 Dynamic Data Encryption Strategy (D2ES) Model

Phases of Dynamic Data Encryption Strategy
(D2ES) Model based on the definitions given in section, we present our D2ES model in this section. The crucial goal of D2ES model is solving the problem defined in Definition II.1.There are mainly three phases forming the solution. Fig.4.2 illustrates three crucial phases of D2ES model [12].

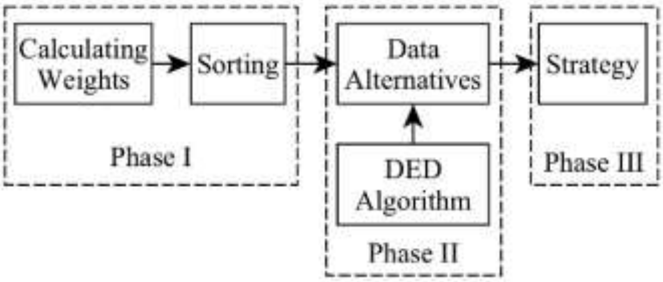

Fig 4.2: Dynamic Data Encryption Strategy (D2ES) Model

Phase I: Sorting by Weights: This is a preparation phase of the model. All data package types are sorted at this phase. The sorting operations consider both execution time and privacy protections; thus two variables are involved, which are PWVs and the corresponding encryption execution time.

Phase II: Data Alternatives: This phase is the crucial step of selecting data packages for encryption operations.

5. CONCLUSION

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 and privacy issues of big data and considered the practical implementations in cloud computing. Proposed approach, D2ES, was designed to maximize the efficiency of privacy protections. Main algorithm supporting D2ES model was DED algorithm that was developed to dynamically alternative data packages for encryptions under different timing constraints.

6. FUTURE WORK

Cloud computing experts believe that the most reasonable way to improve the security of Big Data is through the continual expansion of the antivirus industry. A multitude of antivirus vendors, offering a variety of solutions, provides a better defense against Big Data security threats. Here are some additional recommendations to strengthen Big Data security: Focus on application security, rather than device security. Isolate devices and servers containing critical data. Introduce real-time security information and event management. Provide reactive and proactive protection. For the next step, we will continue to design the detailed process and approach of the secure storage and sharing for cloud big data.

REFERENCES

[1] Ren, Yulong, and Wen Tang. "A Service Integrity Assurance Framework for Cloud Computing Based On Mapreduce."*Proceedings OfIeee Ccis2012*. Hangzhou: 2012, Pp- 240 –244, Oct. 30 2012-Nov. 1 2012.

[2] Hao, Chen, and Ying Qiao. "Research Of Cloud Computing Based On The Hadoop Platform.". Chengdu, China: 2011, Pp. 181 – 184, 21-23 Oct 2011.

[3] 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.

[4] 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. International Journal Of Network Security & Its Applications (Ijnsa), Vol.6, No.3, May 2014.

[5] K, Chitharanjan, And Kala Karun A. "A Review On Hadoop — Hdfs Infrastructure Extensions.".JejuIsland: 2013, Pp. 132-137, 11-12 Apr. 2013.

[6] 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.

[7] 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. International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014 56.

[8] K, Chitharanjan, and Kala Karun A. "A review on Hadoop — HDFS infrastructure extensions.".JeJu Island: 2013, pp. 132-137, 11-12 Apr. 2013.

[9] F.C.P, Muhtaroglu, Demir S, Obali M, and Girgin C. "Business model canvas perspective on big data applications." Big Data, 2013 IEEE International Conference, Silicon Valley, CA, Oct 6-9, 2013, pp. 32 - 37. [10] Zhao, Yaxiong , and Jie Wu. "Dache: A data aware caching for big-data applications using the MapReduce framework." INFOCOM, 2013 Proceedings IEEE, Turin, Apr 14-19, 2013, pp. 35 - 39.

[11] Y. Yu, M. Au, G. Ateniese, X. Huang, W. Susilo, Y. Dai, and G. Min. Identity-based remote data integrity checking with perfect data privacy preserving for cloud storage. IEEE Transactions on Information Forensics and Security, PP (99):1, 2016.

[12] L. Weng, L. Amsaleg, A. Morton, and S. Marchand-Maillet. A privacy-preserving framework for large-scale content-based information retrieval. IEEE Transactions on Information Forensics and Security, 10(1):152–167, 2015.

[13] K. Gai, M. Qiu, H. Zhao, and J. Xiong. Privacy-aware adaptive data encryption strategy of big data in cloud computing. In The 2nd IEEE International Conference of Scalable and Smart Cloud (SSC 2016), pages 273–278, Beijing, China, 2016. IEEE.

[14] Y. Zhang, C. Xu, S. Yu, H. Li, and X. Zhang. SCLPV: Secure certificate less public verification for cloud-based cyber-physical social systems against malicious auditors. IEEE Transactions on Computational Social Systems, 2(4):159–170, 2015.

[15] C. Wang, S. Chow, Q. Wang, K. Ren, and W. Lou. Privacy-preserving public auditing for secure cloud storage. IEEE Transactions on Computers, 62(2):362–375, 2013.

[16] Y. Li, W. Dai, Z. Ming, and M. Qiu. Privacy protection for preventing data over-collection in smart city. *IEEE Transactions on Computers*, PP:1, 2015.

[17] K. Gai, M. Qiu, L. Chen, and M. Liu. Electronic health record error prevention approach using ontology in big data. In *17thIEEE International Conference on High Performance Computing and Communications*, pages 752–757, New York, USA, 2015.

[18] M. Qiu, K. Gai, B. Thuraisingham, L. Tao, and H. Zhao. Proactive user-centric secure data scheme using attribute-based semantic access controls for mobile clouds in financial industry. *Future Generation Computer Systems*, PP:1, 2016.

[19] K. Gai, M. Qiu, B. Thuraisingham, and L. Tao. Proactive attribute based secure data schema for mobile cloud in financial industry. In *The IEEE International Symposium on Big Data Security on Cloud;17th IEEE International Conference on High Performance Computing and Communications*, pages 1332–1337, New York, USA, 2015.