

Secured Big Data Computing in Cloud Environments

G. Kishore Kumar

Research Scholar

Department of Computer Science

Chikkanna Government Arts College, Tirupur, TN, India

Dr. M. Gobi

Assistant Professor

Department of Computer Science

Chikkanna Government Arts College, Tirupur, TN, India

Abstract

"Big Data" refers to data sets that are huge in size and/or complex. They may be evaluated computationally to reveal trends, associations and patterns, particularly connecting to human behavior and interactions wherein outdated data processing application software's are inadequate to deal with them. There are various security challenges linked in this area and if any security breach occurs then it would result in great loss in data and reputation, as the massive volume of data is involved. In addition, traditional security mechanisms and tailored mechanisms for big data would not be sufficient. In this paper, we highlight the data security and privacy challenges proposed by various sources and proposed encryption technique in our future research. For making big data secure, techniques such as encryption must be necessary.

Keywords- Big Data, Challenges, Security, Privacy, Cloud Computing

I. INTRODUCTION

A. Cloud Computing

Cloud Computing refers to the exercise of using a network, which comprises remote servers, hosted on the Internet to process, manage and store data, as opposed to a personal computer or a local server. It is a type of Internet-based computing, which offers shared computer handling resources and data to computers and/or other devices required on demand. It is a prototype for facilitating on-demand and global access to a shared pool of configurable computing resources (e.g., computer networks, servers, storage, applications and services). Cloud computing and storage solutions offer various provisions to the users and/or enterprises for storing and processing their own data. This can be either done in a privately owned data center or owned by the third parties whereas these data centers might be located anywhere in the world. [As per US National Institute of Standards and Technology (NIST)]

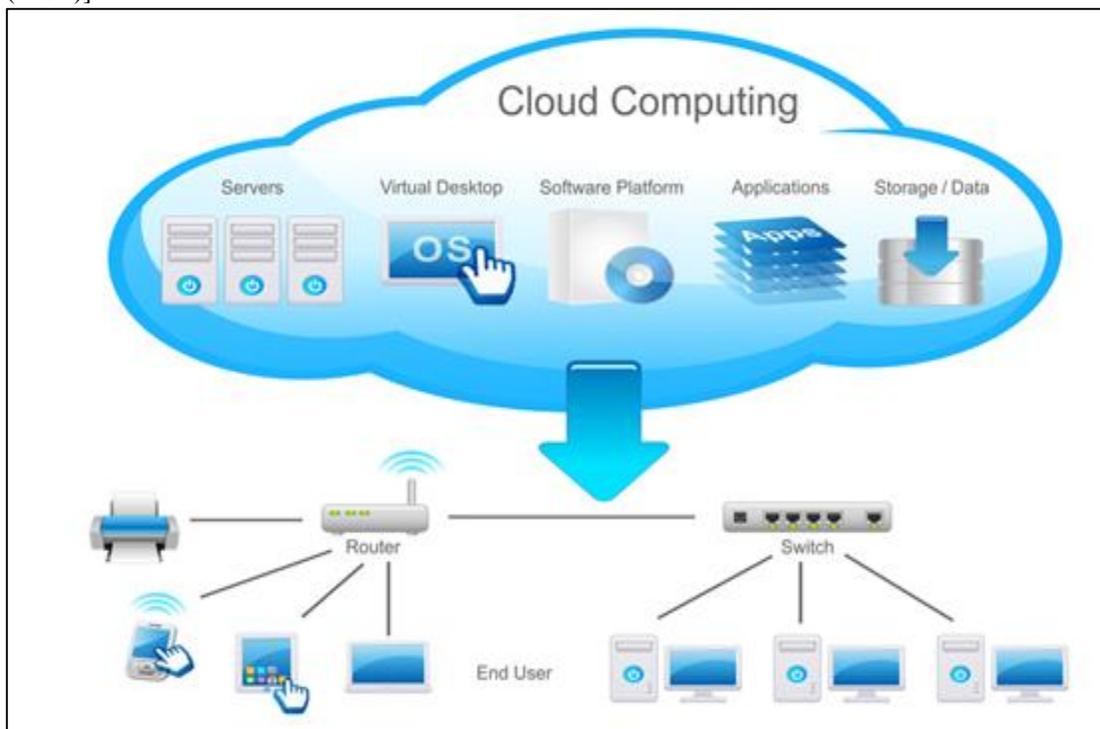


Fig. 1: Cloud Computing Overview

B. Big Data

Big Data refers to defining enormous volumes of structured and unstructured data that are so huge. It is very tough to process these data using outdated databases and software technologies. The term “Big Data” is companies who have to query loosely structured very large distributed data.

In general, the Big Data hardware/software innovations are found & led by various companies such as large-scale advertising, social networking and internet search. For example, Google does an analysis and delivering the search results & personalized advertising in milliseconds on the clicks, links and content available on 1.5 trillion page views per day. This is remarkable and great achievement in computer science & Engineering.[16]

The open source community, broader commercial and public sector have the contributed technologies from Google, Yahoo, Oracle and a few other companies for taking the challenge in making the Big Data work for their own. As opposed to the pioneers, this broader community views the big data in a bit different way. Instead of interpreting the data independently, they add the new data to the existing operational or analytical systems.

Where do the data can be secure? : In memory, discovery lab, Data Reservoir, Data Warehouse, Access through tools and Networks.



Fig. 2: Big Data Overview

Big Data is defined by the following four “V”s:

- Volume
- Variety
- Velocity
- Veracity

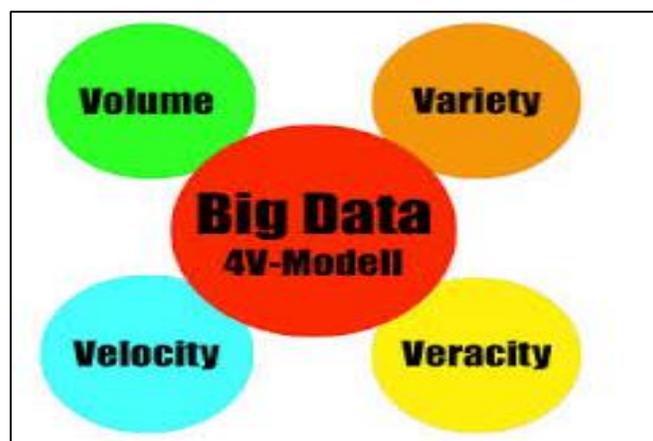


Fig. 3: 4V's of Big Data

C. Hadoop

Hadoop is a Java-based programming framework, which is an open source. Hadoop supports the storage and processing of very large data sets in a distributed computing environment. It belongs to the Apache Software Foundation and integrated in Apache projects. The Hadoop framework is mostly formed in the Java programming language, with partial native code in C and command line utilities written as shell scripts. Though this MapReduce-Java code is common, any programming language along

with “Hadoop Streaming” for “map” and “reduce” parts implementation of the user’s program can use this. Other projects in the Hadoop ecosystem expose richer user interfaces.

D. Map Reduce

A MapReduce job functions in such a way that the input data set gets split into autonomous chunks that are processed by map tasks in parallel method. The MapReduce framework inputs to the reduce tasks after sorting the outputs of the maps. After processing both inputs & output of the job are stored in a file system.

E. Big Data & Cloud

Cloud computing plays a key role in Big Data; not only because of infrastructure and tools provided, but also because of a business model that can be followed by Big Data Analytics (e.g. Analytics as a Service (AaaS) or Big Data as a Service (BDaaS)). However, AaaS/BDaaS gets various challenges due to customer and provider’s staff are much more in the loop than in outdated Cloud providers offering infrastructure/platform/software as a service.

Cloud computing comprises various security concerns due to the various technologies incorporated in it, like databases, networks, operating systems, load balancing, memory management, etc. Hence, the security issues associated with these systems are affecting cloud environments as well.

II. NEED OF SECURITY IN BIG DATA

Big Data requires the same kind of principles and practices as like mentioned earlier in the Cloud. The enterprise security management requires centralized access, resource authorization and manage through complete audit practices. The Big data strategy would be aligned with the standard enterprise practices & policies that are already established to avoid multiple/duplicate implementations and having a centralized environment management.

The key challenge in this area is discovering techniques that are able to explore the rapid elasticity and large scale of Cloud environment. As the data available for Big Data analytics are increasing drastically, timely processing of such data would be a relevant advantage to be able to explore such a capability for businesses. In the same way, standards and interfaces for these activities are also required, as they would help to disseminate “prediction and analytics as services” providers that would compete for customers.

In the area of Big Data, many associated research challenges are available. First of all the data processing techniques needs to more efficient/effective to have real-time work. The major area in computer networks and software analytics, which has major significance in large-scale infrastructure management like Cloud with associated software/open source development and its quality developments.

III. SECURITY ISSUES ASSOCIATED WITH BIG DATA IN CLOUD COMPUTING

The challenges of security in cloud computing environments can be categorized into the following major areas viz. Network level, User authentication level, Data level and generic issues.[5]

Network level: The network security and network protocols, like distributed data, nodes, internode communication etc. fall under this category.

- Authentication level: The various authentication methods related to node level management, user authentication and encryption techniques fall under this category. The authentication methods involve logging, application and nodes authentication and admin rights for nodes.
- Data level: The data availability and integrity like data protection fall under this category.
- Generic types: The usage of various technologies and traditional security tools fall under this category.

Below given are fewer security issues identified in the area of big data in cloud computing:

- The network that does the interconnection to the systems in the cloud must be very secure.
- The virtualization paradigm in cloud computing has various security issues.
- The mapping between virtual machines and physical machines must be accomplished very securely.
- The data security, which comprises the data encryption and relevant policies, must be imposed for data sharing.

Moreover, memory management and resource allocation algorithms also must be secure

IV. RELATED WORK, ISSUES & CHALLENGES

There are six major research areas identified in the big data security in the Cloud are given below:[8]

- Applied ontology
- Security
- Storage and Transport
- Accessibility

- Inconsistencies
- Mobility

The following are the top 10 challenges identified by CSA[1] in the area of Big Data in Cloud Computing:

- Secure computations in distributed programming frameworks
- Security best practices for non-relational data stores
- Secure data storage and transactions logs
- End-point input validation/filtering
- Real-time security/compliance monitoring
- Scalable and compassable privacy-preserving data mining and analytics
- Cryptographically enforced access control and secure communication
- Granular access control
- Granular audits
- Data provenance

The following is the security classification/framework identified by the current researchers:

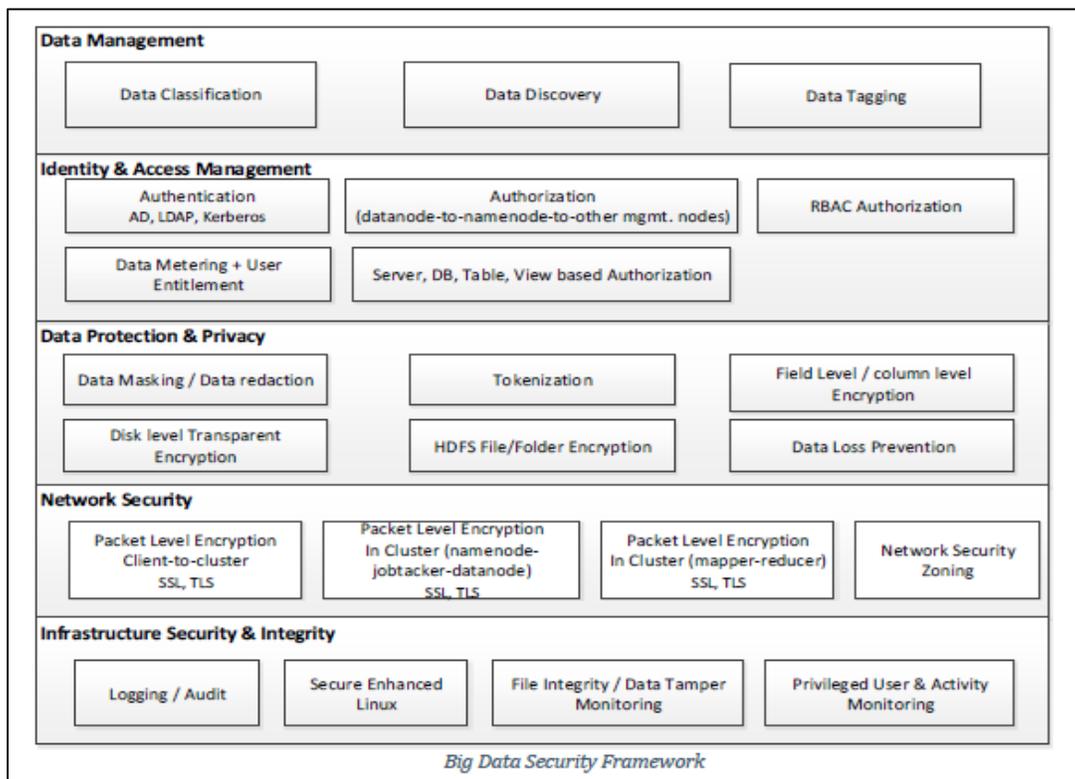
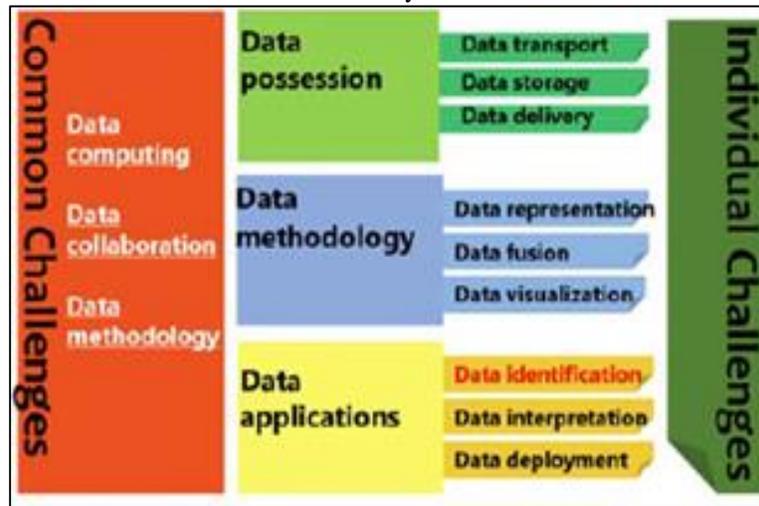


Fig. 4: Security classification/framework

The following are the few research areas identified by the current researchers wherein any of these areas can be utilized for our further research:

- Logging: The logging needs to be enabled for all map reduce jobs that modify the available data along with the user information who are responsible for those. These logs can be used for auditing purposes to find out whether any malicious operations are executed in the data available in the nodes.
- Software Format and Node Maintenance: The nodes, which executes the software, must be formatted on a regular basis to remove the virus if available. In addition, all the software’s including Hadoop must be updated to have the complete secured system.
- Nodes Authentication: The authentication must be done when any node is joined with a cluster. No malicious node must be allowed to join the cluster. Various authentication techniques like Kerberos can be used for validation between authorized and malicious nodes.
- Rigorous System Testing of Map Reduce Jobs: Once a map reduce job is created by any developer, the testing must be done in a distributed environment thoroughly instead of a single/stand-alone machine. This will ensure the stability and robustness of the job created.
- Honeytrap Nodes: Honey trap nodes in a cluster can be used to find and eliminate hackers. These nodes will appear like normal nodes, but they are the traps to find out the hackers.

Table 1: Abbreviations and Acronyms

Sr. No.	Abbreviation	Meaning
1	CSA	Cloud Security Alliance
2	HDFS	Hadoop distributed file system
3	BDA	Big Data Analytics
4	GFS	Google File System

V. CONCLUSION / FUTURE RESEARCH DIRECTIONS

Security and privacy plays major role in the big data processing and storage. This involves the massive use of huge infrastructures and third-party services that used for hosting various services like data hosting, critical operations performance etc. As this grows exponentially, massive amount of challenges is bringing in the data storage, monitoring and security. In addition, the prevalence of cloud brings more challenges into the picture along with the available big data challenges. As the cloud, being an open network the focus must be on the data security & storage, which is an essential and effective privacy protection in this vast area. Our future research would be focusing on these three major areas Network level, User Authentication Level and Data Level wherein the security features can be implemented or enhanced on the data security framework, that are already available.[3]

REFERENCES

- [1] CSA, "Top Ten Big Data Security and Privacy Challenges", November 2012
- [2] CSA, "Big Data Analytics for Security Intelligence", September 2013 Satoshi Tsuchiya, Yoshinori Sakamoto, Yuichi Tsuchimoto & Vivian Lee, "Big Data Processing in Cloud Environments"
- [3] Sophia Yakoubov, Vijay Gadepally, Nabil Schear, Emily Shen, Arkady Yerukhimovich, "A Survey of Cryptographic Approaches to Securing Big-Data Analytics in the Cloud", MIT Lincoln Laboratory
- [4] B. Hari Krishna, Dr.S. Kiran, G. Muralia, R. Pradeep Kumar Reddy,
- [5] "Security Issues In Service Model Of Cloud Computing Environment", 2016 International Conference on Computational Science
- [6] Marcos D. Assunção, Rodrigo N. Calheiros, Silvia Bianchi, Marco A.S. Netto, Rajkumar Buyya, "Big Data computing and clouds: Trends and future directions", J. Parallel Distrib. Comput. 79–80 (2015) 3–15
- [7] Venkata Narasimha Inukollu, Sailaja Arsi and Srinivasa Rao Ravuri, "SECURITY ISSUES ASSOCIATED WITH BIG DATA IN CLOUD COMPUTING", International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014
- [8] Raghav Toshniwal, Kanishka Ghosh Dastidar, Asoke Nath, "Big Data Security Issues and Challenges", International Journal of Innovative Research in Advanced Engineering (IJRAE) ISSN: 2349-2163 Issue 2, Volume 2 (February 2015)
- [9] S. Justin Samuel, Koundinya RVP, Kotha Sashidhar and C.R. Bharathi, "A SURVEY ON BIG DATA AND ITS RESEARCH CHALLENGES", VOL. 10, NO. 8, MAY 2015 ISSN 1819-6608 ARPN Journal of Engineering and Applied Sciences
- [10] Mingmin Chi, Antonio Plaza, Jon Atli Benediktsson, Zhongyi Sun, Jinsheng Shen, and Yangyong Zhu, "Big Data for Remote Sensing: Challenges and Opportunities", 2015
- [11] Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, Keqiu Li, "Big Data Processing in Cloud Computing Environments", 2012 International Symposium on Pervasive Systems, Algorithms and Networks
- [12] Zan Mo, Yanfei Li, "Research of Big Data Based on the Views of Technology and Application", School of Management, Guangdong University of Technology, Guangzhou, China
- [13] Dr. Jangala. Sasi Kiran, M.Sravanthi, K.Preethi, M.Anusha, "Recent Issues and Challenges on Big Data in Cloud Computing", IJCST Vol. 6, Issue 2, April - June 2015
- [14] Raghavendra Kune, Pramod Kumar Konugurthi, Arun Agarwal, Raghavendra Rao Chillarige and Rajkumar Buyya, "The anatomy of big data computing", Softw. Pract. Exper. 2016; 46:79–105
- [15] Bo Li, "Survey of Recent Research Progress and Issues in Big Data"
- [16] Research Trends, "Special Issue on Big Data"
- [17] Zan Mo, Yanfei Li, "Research of Big Data Based on the Views of Technology and Application", American Journal of Industrial and Business Management, 2015, 5, 192-197

- [18] Zhang Hongjun, Hao Wenning, He Dengchao, Mao Yuxing, "Survey of Research on Information Security in Big Data", XXXIV - CSBC 2014
- [19] Elmustafa Sayed Ali Ahmed and Rashid A.Saeed, "A Survey of Big Data Cloud Computing Security", International Journal of Computer Science and Software Engineering (IJCSSE), Volume 3, Issue 1, December 2014
- [20] "An Enterprise Architect's Guide to Big Data", ORACLE ENTERPRISE ARCHITECTURE WHITE PAPER | MARCH 2016
- [21] Fran Howarth, "Harnessing big data for security", A White Paper by Bloor Research-February 2013
- [22] Julio Moreno, Manuel A. Serrano and Eduardo Fernández-Medina, "Main Issues in Big Data Security", Alarcos Research Group, University of Castilla-La Mancha
- [23] "Challenges and Opportunities with Big Data", A community white paper developed by leading researchers across the United States
- [24] Kalyani Shirudkar, Dilip Motwani, "Big-Data Security", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 3, March 2015