International Journal of Electrical and Computer Engineering (IJECE)
Vol. 7, No. 5, October 2017, pp. 2873~2882
ISSN: 2088-8708, DOI: 10.11591/ijece.v7i5.pp2873-2882

Content an Insight to Security Paradigm for BigData on Cloud: Current Trend and Research

Chhaya S Dule[1], Girijamma H. A.[2]
[1]Dept.of CSE, RNSIT, Dept. of CSE, Jyothy Institute of Technology, Bangalore, Visvesvaraya Technological University (VTU), Belgaum, and Karnataka, India
[2]Dept of CSE, RNS Institute of Technology Bangalore, India, Visvesvaraya Technological University (VTU), Belgaum, Karnataka India

Article Info

Article history:

Received Jan 7, 2017
Revised Jun 12, 2017
Accepted Sept 11, 2017

Keywords:

Big data
Cloud
Privacy
Security
Tools

ABSTRACT

The sucesssive growth of collabrative applications producing Bigdata on timeline leads new opprutinity to setup commodities on cloud infrastructure. Mnay organizations will have demand of an efficient data storage mechanism and also the efficient data analysis. The Big Data (BD) also faces some of the security issues for the important data or information which is shared or transferred over the cloud. These issues include the tampering, losing control over the data, etc. This survey work offers some of the interesting, important aspects of big data including the high security and privacy issue. In this, the survey of existing research works for the preservation of privacy and security mechanism and also the existing tools for it are stated. The discussions for upcoming tools which are needed to be focused on performance improvement are discussed. With the survey analysis, a research gap is illustrated, and a future research idea is presented

Corresponding Author:

Ms Chhaya S Dule,
Research Scholar,
Dept. of CSE, RNSIT, Asst. Prof, Dept. of CSE,
Institute of Technology, Bangalore, Visvesvaraya Technological University (VTU), Belgaum, Karnataka, India
Email: chhayaresearch15@gmail.com

1. INTRODUCTION

The expanded technology in various areas has made interest towards the big data and has become a trending research subject offering various applications in different areas like social, climate, government, etc. The research topic big data falls under the category of big data with machine learning [1-2]. This top research scenario will not be completed without the networking, as the real-time applicability demands complex, huge data processing. Still, it is observed that the BD is mysteriously challenging research subject. In this still there is a need of addressing various problems under various scenarios and also better algorithms are needed to be developed to solve these ongoing issues. The major issues include privacy and security in data operation over the cloud under real time scenario. The collected data in BD are easily accessible when the heterogeneous data is transferred over the cloud. The confidential data which includes clinical data, research data, government data, and military data and when these data is transferred over the internet it can be accessed and malfunctioned by the intruders. The tools which are used or developed to handle the tremendous amount of data are successful in handling but fail to maintain or preserve the security & privacy [3]. The existing tools are falls below the security protection zone which come for large scale of data [1-2], [4]. The expose of the tremendous amount of data in BD leads a better analytical response but will fail to offer better security that means will give more security issues. The exposed data can be very useful for the hacker & owner prospective, and the hacker can change it accordingly as per his needs. Hence there is a need

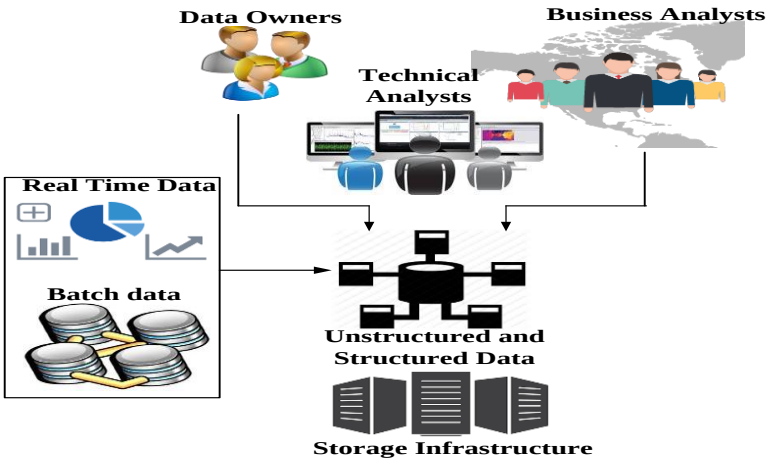
of proper protection for entire communication and analysis process. As the study analysis says that the conventional method or security techniques are failed to handle loads of huge data having huge volume, velocity, complexity, etc. [5]. The privacy in all the prospects is the overwhelmed and unresolved concern. For big data security and privacy preservation there exist three different kinds of methods, i.e., K-Anonymity, T-closeness, and L-Diversity. Also, the consent and notice are used for consumer privacy preservation; in this, the user information can be exchanged only after getting the consent from the user using the notice. The method can be used when the user uses a new service [6]. Also, the differential method can also be used for privacy preservation in big data. But in real time application, the privacy is unresolved. This paper gives the survey for the privacy and security preservation in Big Data.

One of the biggest problems with the big data security is that ultimately the big data are stored in warehouse which still uses conventional authentication mechanism. As analytical process is an expensive process that not only requires specialized tools but also requires different forms of infrastructure. Majority of the existing mining tools (e.g. Apache Hive) is based on open source programs which is unfortunately the same tool used by the intruder to initiate an attack. Not only has this majority of the storage also used the concept of multi-tenancy in order to minimize the operation cost of the user. Usage of multi-tenancy is another cause of problems toward majority of the security threats mainly the privacy problems. Although, there are various security challenges in cloud environment e.g. integrity, non-repudiation, confidentiality, privacy, availability, etc, the service provider is yet to ensure resiliency against potential threats over cloud environment apart from the technical challenges in processing big data. This paper reviews such problems and techniques. The sectional organization of the paper is done as follows: Section 2 discusses about a conceptual description of big data followed by discussion of privacy issues in big data in Section 3. Discussion about existing security tools is elaborated in Section 4. The advantage points on Big Data are discussed in Section 5. The briefing of an elaborated research work towards security of big data is carried out in Section 6 followed by highlights of research gap in Section 7. Finally, Section 8 outlines the conclusion of the paper.

2. ESSENSIALS OF BIG DATA

2.1. Architecture of Big Data

The term big data can be said to be a large voluminous data that is characterized by more complex form of data with sophisticated relationship existing among such data sets. The main advantage of big data is that it performs the better analysis of huge data than conventional analysis methods. Due to this reason the big data has gained very much interest in the present generation, which has advancement in the data collection, data storage as well as data interpretation. From last few decades, the use of digital media is being increased in many areas which generates the tremendous amount of data, e.g. hospital data, bank data, social networking data, etc. The data storage cost is decreasing day-by-day by which we can store the entire data rather than discarding it. In addition to this, many of the data analyzing techniques are developed but very few of them have succeeded in efficient data analysis [1], [7]. The big data in the real world is like the collection of huge resources which can be used regularly. The architecture (Figure 1) of the big data consists of i) data generation point (unstructured/structured), ii) data owners, iii) business analysts, and iv) data batch.



Figurem 1. Architecture of big data

2.2. Big Data Charecteristics

There are five different big data characteristics (Figure 2) [8], i.e., i) Data volume, ii) Data velocity iii) Data Variety, iv) Data Value and v) Data Complexity.

- *Data Volume:* -The tremendous amount of data itself form high volume of big data. The existing data volume size is 1015 terabytes, and it has been predicted that the data volume size is predicted as 1021 zettabytes in future. Owing to large volume of data, it is quite a challenging task to apply any form of conventional data processing or analysis on the top of big data.
- *Data Velocity:* -The data volume is a problem that deals with the data speed for various sources. Owing to higher speed of data generation, it is quite a difficult task to capture the live data and apply analysis on it.
- *Data Variety:* -The data variety is a problem arising from different forms of the data over the distributed and large network. The data is in the form of video, audio, text, etc. measures the data representation. Owing to different forms, it quite a hard task to write a dynamic query system considering such heterogeneity.
- *Data Value:* -The data value can measure the data usefulness in decision making. The user can compile the data stored and which can offer the filtered data. However, capturing data value considering above mentioned problem in big data is really a difficult one.
- *Data Complexity:* -The data complexity measures interconnection and independence of big data structures that needs large data changed.

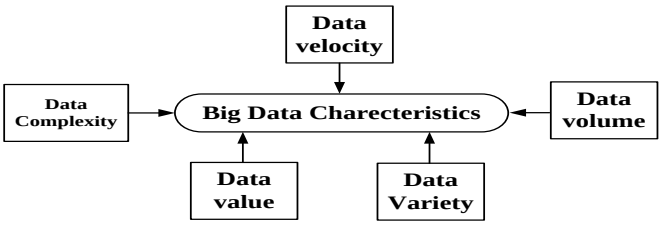


Figure 2. Big data characteristics

2.3. Issues in Big Data

There are few conceptual points which define the big data issues that need to be analyzed by the organization and need to adopt the technology efficiently.The issues and problems of the big data need to be handled separately.

- *Inter-Related Issues with Characteristics:* There are some issues which are inter-related with the big data characteristics. When the volume of data rises, the value of various kinds of data will fall. Today, the social sites are generating terabytes of data, and it has become more difficult to handle the data using traditional techniques [4]. At present, the uses of e-commerce-based applications are increased as well as various transactional data. The conventional data handling techniques are failed to maintain the data velocity management in a particular bandwidth. The data type may be in the form of structured, semi-structured and unstructured data and such form cannot be managed with the traditional techniques of analytics. The data of different organizations will be different and they will adopt different data analytic methods. This will lead to the business gap between the IT organization and business leaders and also makes storage issue. Also, the big data complexity is an issue of big data that requires thousands of parallel running software's with many servers.
- *Inter-Related Issues with Transport and Storage:* Today, data created with many social networking sites are generating the huge volume of data, i.e., the data is generating with various kinds of devices like mobiles, computers, etc. The data quantity is more than Exabyte. The current or traditional data techniques limit terabytes of data. With the existing techniques to transfer the Exabyte of data, it takes nearly 3k hours. If the data transfer is sustained, it needs some more time.
- *Inter-Related Issues with Data Management:* The management of data is one of the biggest issues in big data. The data with different size, format, etc., and validation of these complex data is impractical in nature. The representation of present digital data in a rich manner is acceptable with collection of methods, but there is no efficient mechanism for data management.

3. PRIVACY ISSUES IN BIG DATA

The big data provides the vast application advantages but the conventional data analysis methods fail to provide the proper privacy mechanism. The privacy concern of the big data includes the private data disclosure to the world.The open-end issue with the big data is privacy and security.

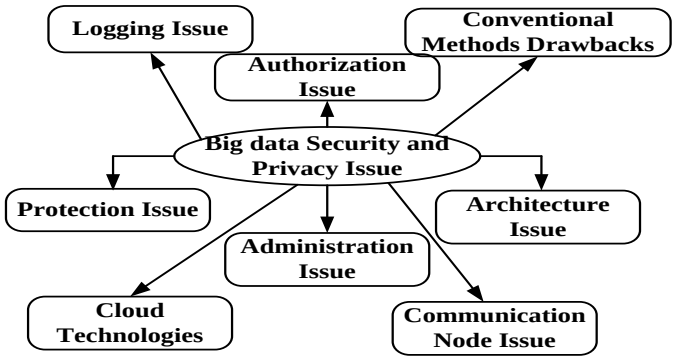


Figure 3. Big data security & privacy

The big data may also include some of the personal data which is shared in the social networking sites. These data can be combined with the other data sets in the real world leading to the exposure of personal data to others. The shared personal data in the social media can be compromised or used by others for the purpose of business in illegitimate manner without any knowledge of the user. Some of the personal data can misuse for criminal activities hence it needs a better focus [9].Some of the security and privacy challenges (Figure 3) are stated below:

- *Protection Issue:* Some of the data which we store in the cloud will not be encrypted for efficiency purpose which will be compromised or encounter protection issue for the important data.
- *Administration Issue:* Every administrative node has an accessibility of any data which may introduce a malicious code and by which the data can be manipulated or compromised easily.
- *Architecture Issue:* In architecture of distributed node, the data can be processed by many resources. When the node gets distributed anywhere in the cluster, it will be very tough to identify the computation and by which the security assurance, in such case, will be very difficult.
- *Communication Issue:* Many of the Hadoop-based data communication may perform over the wired or wireless connection by which anyone can tap or hack the network node and collect the important data.
- *Cloud Technologies:* The existing technologies of the cloud are not much efficient in offering the personal data security.
- *Logging Issue:* As we all know that the access of the cloud for a user is logging independent by which user will not be having any kind of data modification.
- *Conventional Methods Drawbacks:* In previous data management, tools are not much efficient in handing the huge volume of data, by which data gets leaked in the real world.
- *Authorization Issue:* The joining of the third party service provides make security and privacy issue for users in any network activity.

4. EXISTING SECURITY TOOLS

The emerging and growing Big Data need an extraordinary technology that can process the data with greater volume within a shortest time. The mechanism adopted for big data such as massively parallel processing database, distributed database, cloud computing, scalable data storage units, etc. In real time applications, big data analytics plays a major role. In recent past, there are various tools or techniques are developed or examined to store, aggregate, manipulate, visualize and analyze the data. These all the above methods have considered by computer science, economics, mathematics and also even the statistics. This gives an idea that the organizational units are adopted or interested to adopt value from BD in such a way that it can achieve flexibility, discipline in the method [10]. From the study analysis presented by the International Data Corporation (IDC) infers that the data in coming half decade will grow more than 40% from now. In these data growing scenario can be adopted in the embedded units having applications in building

construction, medical area, etc. The unstructured data includes emails, videos and files can grow more than 75% in next half decade. The main issues are that the professionals of IT who are handling the growing data can offer only one and a half percent of data management for the faster-growing world. Today millions of digital data are generating and its size in every year is increasing. If we turn back and look at in last decade, the data has grown bigger and bigger and has the challenge that how these data can be analyzed. The big data technique is giving an idea how this data can realize.A big data technique should meet the following performance factors:

- The technique should be able to define the issues like variety, velocity, volume, veracity, etc.
- The technique should have the capability of enhancing the data performance, reliability, security, etc.
- It should have the ability to get connected with databases, warehouses, etc.
- The technology should be open source and need to have functionality integration ability.
- The technology must have low latency, robustness and also the fault tolerant ability.
- The technology must have the scalability and extensibility.
- The technology must allow the ad-hoc queries along with minimal maintenance.

Currently, the Apache Foundation based Hadoop tool is an open source composed of numerous small sub systems of infrastructures facilitating distributed cloud computing. The subsystems in Hadoop can be given as i) Hadoop Distributed File (HDF) system is also considered as file system and ii) MapReduce may also noted as programming paradigm. The other subsystems face issues while working with the huge volume of data. In recent past, the works have given significant results to store the tremendous amount of data but failed to give the accuracy in reading back these data. In these methods, reading of data takes more time than the time taken to read the small voluminous data. This issue can be attained by reading the number of data disks at a time, but it does not make any sense as it leads to the use of a number of disks. The process may lead to hardware failures, and it can be overcome by replication.

The above-mentioned issues can be resolved with the help of HDF system and the data combining issues by Map Reduce. The advantage of Map Reduce is that it will minimize the computation risks dealing with reading and writing. Thus we can say that the Hadoop-based system can offer a reliable solution for data storage and present an efficient analysis mechanism. The HDF system and Map Reduce can offer storage.

4.1. HDF System

The HDF [11] system is a file system and is designed to store a large volume of data and streaming large volume of data access, run the clusters, etc. Normally the block size in it is 64 MB, and it helps to reduce the required disk for storage. The cluster of the HDF system consists of two nodes master or name node, workers or data nodes. The function of the master node is to manage the file system name-space, Meta data and File System (FS) in a tree. The function of the data node is to store and retrieve the blocks addressed by the master node. Once the data is retrieved, it will be reported back to the master node along with the list of stored block. This means in the absence of master node it will be hard to access a file.

4.2. MapReduce System

This is a programming paradigm allows huge scalability. This paradigm performs Map and reduces task [12]. The Map tasks are the inputs taken from the distributed FS, which generates a key value sequence pair as per written code for map function. The generated sequence pair will be collected by the master controller and are separated with the reduced task after sorting by key. In sorting a key having same value will end up with same reduce task. The function of reduce task is to merge the entire working key values with a key in same time.

Some of the recent techniques are addressed below:

- IBM Infosphere (IBMIs) Insights: This is an open source which composed of IBM big sheet along with Apache Hadoop platform offering better data analysis without imposing the schematic in its format and does the speedy analysis.
- Kognitio platform: This is an analytical format offers faster scalable database analysis.
- ParAccel: This is a parallel processing database analysis platform offers strong compilation, optimization, etc.
- SAND: This is an analytical platform that will give the linear data scalability via parallel processing.

Recently, there is some research performed to improve the Hadoop security by some of the industries [13] .i.e., Apache Ranger, Apache Rhino, Apache Sentry and Apache Knox (Figure 4).

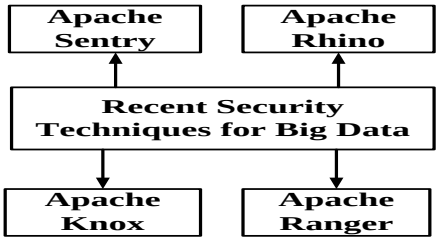


Figure 4. Recent security techniques

The brief discussion of the existing tools is as follows:

- *Apache Ranger*: This is a data security framework for the Hadoop platform that enables the enterprises to run the different workloads in the multi-tenant environment. The ARr function is to give the centralized administration security to control each and every security tasks. Also, it aims to provide better authorization.
- *Apache Rhino*: This security method has been initiated and developed by the Intel Corporation in the year 2013 and significantly it obtained Hadoop ecosystem (HE) security. This method can offer many management, logging, authorization security of Hadoop.
- *Apache Knox*: This is also a recent effort for security preservation for the secure and authorized access to Hadoop cluster via organizational policies. This is an enhanced version of job execution and cluster data execution. This offers easy web service integration between Kerberos authentication and existing authentication providers.
- *Apache Santry*: This enforces the fine-grained authorization for metadata and data of Hadoop cluster. This helps to define access control.

5. BIG DATA ADVANTAGES

Today in every sector we find the applications of BD it may be social, technology, science, etc. The prime application of the Big Data is to perform analysis of the large and massive streams of data which cannot be analyzed using conventional analytical algorithms.

Figure 5. Big data advantages

The applicability depends on how the human can use it according to his necessity. In following some of the BD, advantage is addressed.

- *Customer Targeting and Understanding*: The above-titled advantages are uncovered in different sectors. In this scenario, the BD can be used to target the customers by providing some of the offers and understanding them with customer opinions about their products. The organizations are trying to spread their current data in social Medias to figure out the customer's opinion and improve its product or service.
- Optimization of Business Process and Understanding: The technology of BD is also applied in much business process analysis. With BD a retailer can be able to understand the stock rates by which it gives a route how the delivery can be optimized. For example in human resources, the BDA is adopted that includes talent acquisition and optimization.

- In Advancement of Research: The use of BDA in the research area is making a lot of buzzes as it is offering new ideas. The availability of huge data storage can help to analyze any research logic in-depth.
- *In Heathcare Industry*: With the help of BDA, it is possible to decode complete DNA string in less time and which impacts on disease finding. Clinically it helps to maintain the patients past and present health data.
- *In Machine Performance Optimization*: With the BDA improvement, many of the machines are become automated and smarter. Today, we can find the automated care by Google. A car from Toyota has many cameras with different sensors; GPS interconnected to a computer and can give a safer driving without human presence.
- *In Security Enforcement*: The BDA is also adopted in the prevention of cyber crimes, unauthorized banking transactions and also in identifying the terrorist attacks.

6. RESEARCHES PERFORMED ON BIG DATA SECURITY

In this section, the total numbers of researches which are presented in IEEE Xplore are formulated below and are represented in following plot Table 1.

Table 1. Survey in BD Security

Publications	Total numbers
Conference Publications	1728
Journals & Magazines	190
Early Access Articles	19
Standards	9
Books & eBooks	7
Courses	1

From the table, it is observed that there exist number of conference publications than journals and magazines, which means there are least studies in it.The section discusses the some of the existing recent research works towards privacy and security preservation in BD. Following are the selected works from IEEE Xplore to idealize the research gap.

The concept which is demonstrated in Lei et al. [14] gives privacy and security based data mining for big data information. Also, it is mentioned that some of the sensitive data may get disclosure to the unauthorized access in various ways during the data processing, data collection, etc. In the authors, work has found four different kinds of data mining tools which can offer better data mining like Data Miner, Decision Maker, Data Provider and Data Collector. The study gives some of the privacy concerns and idealizes some of the interesting solutions for it. The work described in Hu et al. [15] gives the prominent energy efficient mechanisms in which the current power issues with technological issues are addressed. In the work authors have considered one of the privacy and security issue of BD i.e. architectural issue is addressed. Yan et al. [16] introduced a unique concept of Encrypted BD duplication over the cloud computing. The data which can be stored in the cloud by encrypting it but it will emit the data duplication in the cloud. This issue will offer data storage and processing issue, and in this case, the traditional duplication schemes will not perform efficiently, or it will not work. In that sense, other has offered a mechanism to duplicate the stored data. The mechanism works on the concept of re-encryption. The performance analysis of simulation results from a prominent solution for data duplication. A significant framework of categorization and application of privacy preservation mechanism in BD mining stated in Xu et al. [17]. The simulation results of this privacy preservation framework are matching with the game theory analysis concepts. The frequently used framework scenario [18] is presented in Figure 6. In which the data publishing, data mining, and data collection process are involved. In this, the data collector will collect the data from various data providers and sell it to the users, who can do some mining processes. To give some compensation for the provider's privacy losses, the data collector will offer some incentives. The data records in this consist of various attributes.

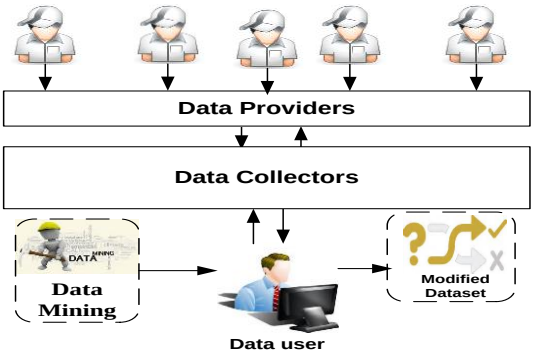


Figure 6. Framework of [17] work

Liang et al. [19] discussed a cipher text multi-sharing data control mechanism is presented to preserve the privacy in BD storage. This mechanism combines both the anonymous method with the proxy re-encryption method, by which a cipher text can be conditionally shared multiple times without disclosure of data. Perera et al. [20] addressed the necessity of BD privacy in the Internet-of-Things (IoT) world. It is admitted that the IoT can help to collect the large set of data on a single platform, but will face privacy issues for the user's data. Nedelcu et al. [20] have addressed the BD challenges and the advantages of BD in manufacturing field. This study analysis includes some of the significant criteria's like the scope of BD and advantages of BD in companies prospective. The privacy and BD concept was also emphasized by Brian et al. [21] where it has mentioned that only providing the security with encryption mechanism is not enough but it is a privacy solution for the data. Some future recommendations are also addressed which may give the better idea for a security solution. The study of Cardenas et al. [22] gave the consequences of big data analytics for security purpose. In this, it is addressed that there is need of certain researches to be kept exploring the various intrusion mechanisms evolving with technology advancement. Yu et al. [23] has expressed the networking concept in big data. In this work several research articles presented for the networking issues in BD, survey overview is described. The work carried out by Baek et al. [24] has discussed about a security technique towards safeguarding the enterprise applications towards smart grids. A hierarchical structure was developed that offers extensive analytical service in most secured manner using digital signature, identity-based encryption mechanism, and proxy re-encryption policies. Problem of data sharing in most secured manner was discussed in the work carried out by Dong et al. [25]. The authors have used the technique of re-encryption that explicitly uses transformation operation over the ciphertext over the virtual machine in order to secure the sensitive data. The security is achieved using identity-based cryptography mainly. There are also various studies towards data privacy. One of such study was discussed by Fu et al. [26] who addressed the problem of sensitive data privacy. An encryption scheme for existing cloud search services is formulated in this work in order to further strengthen the cloud-based semantic system. A typical adversarial model was build using semantic-based approach on text mining was introduced by the author. The study outcome was found to possess better search time with reduced complexity. Hameed and Ali [27] have presented a technique for thwarting the security threats of distributed denial-of-service using Hadoop. A mitigation technique was introduced for identification of possible flooding threat in MapReduce framework. Similar category of the work was also carried out by Jose and Binu [28] considering the case study of intrusion over DNS servers. A counter measure was constructed using Hadoop and MapReduce. Hongbing et al. [29] have introduced the mechanism to secure the storage system of big data using cloud indexing services itself as well as it also uses trapdoor function to further strength the encryption standard. Usage of MapReduce was also seen for evaluating the strength and effectiveness of existing validation techniques over cloud environment. The study produces a good literature support of existing authentication techniques on cloud using MapReduce. Just like MapReduce, there are studies where only Hadoop platform was emphasized for security. The work carried out by Li et al. [30] has used Advanced Encryption Standard (AES) for Hadoop clusters in order to secure multimedia files. The study outcome was finally witnessed with lowered encryption time. Usage of AES was also seen in the work carried out by Wang et al. [31] adopting nanowires. Different forms of memory-based array were deployed for this purpose using external interface of input-output. Schuster et al. [32] have presented a technique that secured analytical framework that uses MapReduce framework. Yang et al. [33] have introduced the mechanism of attribute-based encryption for addressing the problem of secured access control. A unique technique of policy updating scheme was introduced by the author that also balances the computation burden. Zhang et al. [34]

have presented a back propagation technique of higher order in order to perform efficient encryption mechanism over cloud environment.

7. RESEARCH GAP

This section discusses about the problem of existing research work that is still left unsolved. The survey analysis of recent works and some good study for BD privacy and security suggests than the issue is very critical and it needs a proper solution in real time applications scenario. An exhaustive study towards all the recently explored research contribution shows that they provide an efficient technique to solve certain security problems over big data approach in cloud. However, the area of big data is such a vast that it will be quite early to conclude effectiveness of any existing study as robust and resilient towards potential security threats. The problems that are left unsolved are as briefed below:

1. *Less emphasis on authentication mechanism*: There is a less number of improvement being carried out towards ensure a robust and fault-free authentication policy for big data users. Although, there are various novel ideas towards authentication system in cloud but they were never testified for their applicability over big data analytic-based application.
2. *Few studies focusing on privacy*: At present majority of the studies towards big data security have addressed the problem of data security and access control mainly. There is a very less emphasis on the design security that is mainly due to massive data size.
3. *Less focus towards data brokers*: At present, many of the cloud service providers are using multi-tenancy. They also have a practice of sharing certain segment of the data to the third party that tremendously increases potential risk. Privacy is the first thing to get compromised. The existing solution doesn't address such problems.
4. *Storage insecurity in big data*: It is well known fact that NoSQL database is still evolving and quite problematic to retain optimal security as per the demands. Normally, the big data are stored in multiple tiers where existing system doesn't really focus on how such existing encryption strategy is compliant of tier-based storage strategy.

8. CONCLUSION

Ensuring security towards cloud is never an easy task for any service provider especially in the presence of such malicious activities over internet. With the technology of cloud modernizing, the attackers are also becoming smart. From storage viewpoint, storing the operational data takes the storage cost but storing the analyzed data takes the cost of both storage and processing carried out to transform it. Hence, big data is basically an expensive affair toward cloud storage system in order to store it. The problem might turn worst, if such expensive data is subjected to common or potential threats. Hence, this paper significantly discusses the major topics related to big data security and privacy. With this paper, we provided the recent works that are given IEEE Xplore. From the analysis of research gap analysis, it is pointed than no much studies offered efficient security. There are some recent methods can be executed properly to make better security system form BD. Still, these recent methods need continuous research to make them more efficient with increasing real-time data. Our future work will be focused towards ensuring modeling towards strengthening the privacy problems in big data security. Possibilities of using optimization theory will be quite higher as we don't want to increase the resource cost over big data infrastructure while at the same time we want to achieve optimal privacy protection and resiliency from major potential threats.

REFERENCES

[1] S. Lohr, "The Age of Big Data," New York Times, Vol. 11, 2012
[2] M. Swan, "The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery," Big Data, Vol. 1.2, pp. 85-99, 2013
[3] F. Provost, and T. Fawcett, "Data Science and its Relationship to Big Data and Data-driven Decision Making," Big Data, Vol. 1.1, pp.51-59, 2013
[4] T Sutikno, D Stiawan, IMI Subroto, "Fortifying big data infrastructures to face security and privacy issues," *TELKOMNIKA Telecommunication Computing Electronics and Control*., vol. 12, no. 4, pp. 751-752, 2014.
[5] C. Wang, Cong, "Privacy-preserving Public Auditing for Data Storage Security in Cloud Computing," Infocom, 2010 Proceedings IEEE, 2010
[6] O. Tene and J. Polonetsky, "Big data for all: Privacy and User Control in the Age of Analytics," Nw. J. Tech. & Intell. Prop, Vol. 11, 2012
[7] R. L. Villars, C.W. Olofson, and M. Eastwood, "Big data: What it is and Why You Should Care." White Paper, IDC, 2011

[8] I. Rubinstein, “Big Data: the end of Privacy or a New Beginning?,” International Data Privacy Law (Forthcoming), 12-56, 2013
[9] B.D. JAMES, “Security and Privacy Challenges in Cloud Computing Environments,” 2010
[10] A. Katal, M. Wazid, and R. H. Goudar, "Big Data: Issues, Challenges, Tools and Good Practices," Contemporary Computing (IC3), Sixth International Conference. IEEE, 2013.
[11] L. Sweeney, "K-Anonymity: A Model for Protecting Privacy,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 10.05, pp. 557-570, 2002
[12] I. Hashem, A. Targio, "The Rise of “Big Data” on Cloud Computing: Review and Open Research Issues," Information Systems, Vol. 47, pp. 98-115, 2015
[13] “5 Hadoop Security Projects”, https://www.xplenty.com/blog/2014/11/5-hadoop-security-projects/
[14] L. Xu, C. Jiang, J. Wang, J. Yuan and Y. Ren, "Information Security in Big Data: Privacy and Data Mining," in *IEEE Access*, vol. 2, no. , pp. 1149-1176, 2014.
[15] J. Hu and A. V. Vasilakos, "Energy Big Data Analytics and Security: Challenges and Opportunities," in *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2423-2436, Sept. 2016.
[16] Z. Yan, W. Ding, X. Yu, H. Zhu and R. H. Deng, "Deduplication on Encrypted Big Data in Cloud," in *IEEE Transactions on Big Data*, vol. 2, no. 2, pp. 138-150, June 1 2016.
[17] L. Xu, C. Jiang, Y. Chen, J. Wang and Y. Ren, "A Framework for Categorizing and Applying Privacy-Preservation Techniques in Big Data Mining," in *Computer*, vol. 49, no. 2, pp. 54-62, Feb. 2016.
[18] K. Liang, W. Susilo and J. K. Liu, "Privacy-Preserving Ciphertext Multi-Sharing Control for Big Data Storage," in *IEEE Transactions Information Forensics and Security*, vol. 10, no. 8, pp. 1578-1589, Aug. 2015.
[19] C. Perera, R. Ranjan, L. Wang, S. U. Khan and A. Y. Zomaya, "Big Data Privacy in the Internet of Things Era," in *IT Professional*, vol. 17, no. 3, pp. 32-39, May-June 2015.
[20] B. Nedelcu, "About Big Data and its Challenges and Benefits in Manufacturing." Database Systems Journal 4, no. 3, pp. 10-19, 2013
[21] B.M. Gaff, H. E. Sussman, and J. Geetter, "Privacy and Big Data." Computer 47, no. 6, pp.7-9, 2014
[22] A. A. Cárdenas, P. K. Manadhata and S. P. Rajan, "Big Data Analytics for Security," in *IEEE Security & Privacy*, vol. 11, no. 6, pp. 74-76, Nov.-Dec. 2013.
[23] S. Yu; M. Liu; W. Dou; X. Liu; S. Zhou, "Networking for Big Data: A Survey," in IEEE Communications Surveys & Tutorials , vol.PP, no.99, pp.1-1, 2016
[24] J. Baek, Q. H. Vu, J. K. Liu, X. Huang and Y. Xiang, "A Secure Cloud Computing Based Framework for Big Data Information Management of Smart Grid," in *IEEE Transactions on Cloud Computing*, vol. 3, no. 2, pp. 233-244, April-June 1 2015.
[25] Xinhua Dong, Ruixuan Li, Heng He, Wanwan Zhou, Zhengyuan Xue and Hao Wu, "Secure Sensitive Data Sharing on a Big Data Platform," in *Tsinghua Science and Technology*, vol. 20, no. 1, pp. 72-80, Feb. 2015.
[26] Z. Fu, J. Shu, X. Sun and N. Linge, "Smart Cloud Search Services: Verifiable Keyword-based Semantic Search Over Encrypted Cloud Data," in *IEEE Transactions on Consumer Electronics*, vol. 60, no. 4, pp. 762-770, Nov. 2014
[27] S. Hameed and U. Ali, "Efficacy of Live DDoS Detection with Hadoop," *NOMS 2016 - 2016 IEEE/IFIP Network Operations and Management Symposium*, Istanbul, 2016, pp. 488-494.
[28] A. S. Jose and B. A., "Automatic Detection and Rectification of DNS Reflection Amplification Attacks with Hadoop MapReduce and Chukwa," *2014 Fourth International Conference on Advances in Computing and Communications*, Cochin, 2014, pp. 195-198.
[29] H. Cheng, C. Rong, K. Hwang, W. Wang and Y. Li, "Secure Big Data Storage and Sharing Scheme for Cloud Tenants," in *China Communications*, vol. 12, no. 6, pp. 106-115, June 2015.
[30] M. Li, C. Yang and J. Tian, "Video Selective Encryption Based on Hadoop Platform," *2015 IEEE International Conference on Computational Intelligence & Communication Technology*, Ghaziabad, 2015, pp. 208-212.
[31] Y. Wang, L. Ni, C. H. Chang and H. Yu, "DW-AES: A Domain-Wall Nanowire-Based AES for High Throughput and Energy-Efficient Data Encryption in Non-Volatile Memory," in *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 11, pp. 2426-2440, Nov. 2016.
[32] F. Schuster *et al*., "VC3: Trustworthy Data Analytics in the Cloud Using SGX," *2015 IEEE Symposium on Security and Privacy*, San Jose, CA, 2015, pp. 38-54.
[33] K. Yang, X. Jia and K. Ren, "Secure and Verifiable Policy Update Outsourcing for Big Data Access Control in the Cloud," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 12, pp. 3461-3470, Dec. 1 2015
[34] Q. Zhang, L. T. Yang and Z. Chen, "Privacy Preserving Deep Computation Model on Cloud for Big Data Feature Learning," in *IEEE Transactions on Computers*, vol. 65, no. 5, pp. 1351-1362, May 1 2016.