RESEARCH ARTICLE OPEN ACCESS

A Research on Big Data Analytics Security and Privacy in Cloud, Data Mining, Hadoop and Mapreduce

Nandhini.P

PG Scholar, Department of Computer Science and Engineering, Jansons Institute of Technology, Coimbatore, Tamil Nadu, India.

ABSTARCT

Today, the Big Data and its analysis plays a major role in the world of Information Technology with the applications of Cloud Technology, Data Mining, Hadoop and MapReduce. Securing the valuable data from the intruders, viruses and worms are a challenge for the past several decades. So many researchers developed methods and technologies to protect the data. Since all traditional technologies are applicable for only their structured data, we required a new technology to secure and make privacy in the structured, semi-structured and unstructured data (Big Data). In this research paper we have studied various security and privacy methodologies proposed by the various researchers and analyse the merits and demerits of those methodologies.

Key Words: Big Data, Data Analysis, Cloud, Data Mining, Hadoop, Mapreduce, Security and Privacy Methodologies.

I. INTRODUCTION:

Big data is used to store data just like old method like MYSQL, SQL & many more. It is more fast & useful than previous language. Manipulation rate is fast and easy to manage [5]. The term "Big Data" is related with managing (manipulating) high amount of data exist in digitalized form that is collected by various companies or organization. As everyday data are being collected from applications, networks, social media and other sources Big Data is emerging. Studies have shown that by 2020 the world will have increased 50 times the amount of data it had in 2011, which was currently 1.8 zettabytes or 1.8 trillion gigabytes of data. The analysis of Big Data involves multiple distinct phases which include data acquisition and recording, information extraction and cleaning, data integration, aggregation and representation, query processing, data modelling and analysis and interpretation. Each of these phases introduces challenges. Heterogeneity, scale, timeliness, complexity and privacy are certain challenges of big data [7]. There are four different aspects of big data security:

- Infrastructure security,
- Data privacy,
- Data management,
- Integrity and reactive security

Figure 1:Big Data

Characteristics of Big Data is shown in the below figure:

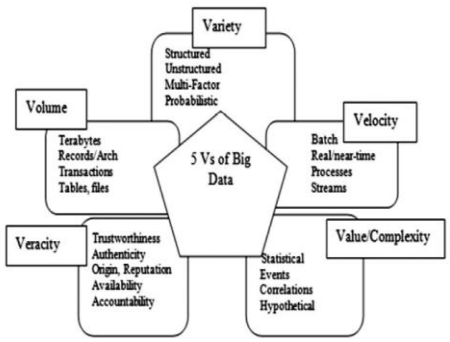

Figure 2: Five Vs of Big Data

Cloud Computing is a technology which depends on sharing of computing resources than having local servers or personal devices to handle the applications. In cloud computing, the word "Cloud" means "The Internet", so Cloud Computing means a type of computing in which services are delivered through the internet [10]. The goal of Cloud Computing is to make use of increasing computing power to execute millions of instructions per second. Clouds provide three types of services, as follows: (i) infrastructure-as-a-service, IaaS, provides infrastructure in terms of virtual machines, storage, and networks, (ii) platform-as-a-service, PaaS, provides a scalable software platform allowing the development of custom applications, and (iii) software-as-a-service, SaaS, provides software running in clouds as a service, for example, emails and databases. Clouds can be classified into three types, as follows: (i) public cloud: a cloud that provides services to many users and is not under the control of a single exclusive user, (ii) private cloud: a cloud that has its proprietary resources and is under the control of a single exclusive user, and (iii) hybrid cloud: a combination of public and private clouds.

One of the most common platform-as-a-service computational paradigms is MapReduce. Hadoop map reduce is a part of Hadoop framework. It will process large amounts of data in parallel on clusters of commodity hardware resources used to write applications that process large in reliable and fault tolerant manner. It 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 [10]. Basically MapReduce is a programming system for distributed processing large-scale data in an efficient and fault tolerant manner on a private, public, or hybrid cloud. Mapreduce is extensively used daily around the world as an efficient distributed computation tool for a large class of problems. Security and privacy of data and MapReduce computations are essential concerns when a MapReduce computation is executed in public or hybrid clouds. In order to execute a MapReduce job in public and hybrid clouds, authentication of mappers-reducers, confidentiality of data-computations, integrity of data-computations, and correctness-freshness of the outputs are required [12]. In this paper we also provide a review of existing security and privacy protocols for MapReduce and discuss their overhead issues and concentrated on providing security to big data which was stored on cloud.

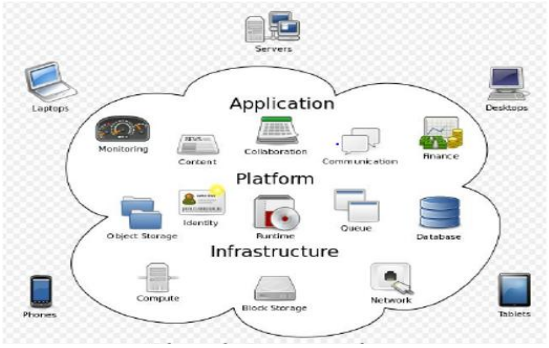

Figure 3: Cloud Computing

Cloud computing and big data are now become as twin technology. Big data provides users the ability to use commodity computing to process distributed queries across multiple datasets and return resultant sets in a timely manner. Cloud computing provides the underlying engine through the use of Hadoop, a class of distributed data-processing platforms [9].

It mainly deals with the security issues faced while using data mining technique from an expanded proportion and review different processes that can help to secure the information. The basic idea here is to identify various types of users who face security issues regarding data mining applications. Recent studies of PPDM primarily centre on how to minimize the security risk arise by data mining methods. In that the "information gaining" is regarded as an equivalent word for another term "Knowledge Discovery from Data" (KDD) which highlights the objective of the mining procedure [15].

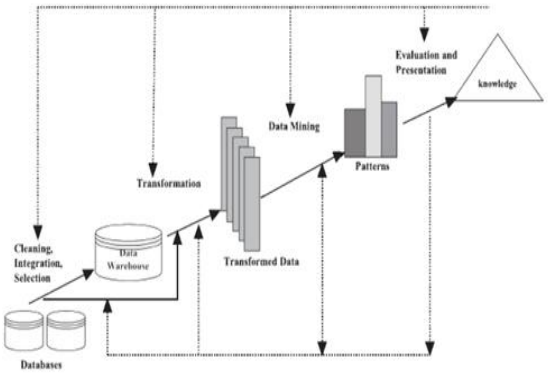

Figure 4: Knowledge-Discovery Process

Step 1: Data pre-processing: Essential operations incorporate information determination (to recover information important to the KDD undertaking from the database), information cleaning (to expel noise and conflicting information, to handle the missing information fields, and so forth.) and

information joining (to consolidate information from numerous sources).
Step 2: Data transformation: The objective is to change information into structures fitting for the mining undertaking, that is, to discover helpful elements to speak to the information. Highlight determination and highlight change are essential operations.
Step 3: Data mining: This is a key procedure where wise techniques are utilized to concentrate information designs.
Step 4: Pattern assessment and presentation: Fundamental operations consolidate recognizing the genuinely intriguing examples which speak to information, and exhibiting the mined learning in a straightforward manner.

Hadoop was produced from GFS (Google File System) and MapReduce papers distributed by google in 2003 and 2004 individually. Hadoop is a system of devices which underpins running application on enormous information and it is actualized in java. It furnishes MapReduce programming engineering with a Hadoop circulated record system(HDFS), which has gigantic information handling capacity with a huge number of item equipment's by utilizing essentially its guide and diminish capacities [11].

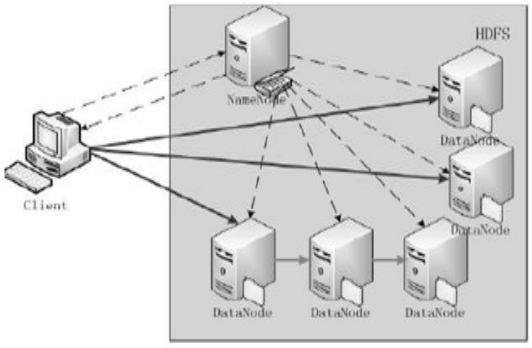

Figure 5: Basic setup for HDFS architecture

In a Big data, the processing of large sets of data in a distributed computing environment is supported by Hadoop, which is a free, Java-based programming framework. Hadoop cluster uses a master/slave structure, which can process a large set of data across a cluster of servers and applications can be run on systems with thousands of nodes involving thousands of terabytes. For handling those network from the failure it uses distributed file system, which supports rapid data transfer rates and allows the system to continue its normal operation and the risk of system failure gets reduced. Hadoop provides 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 tasks – Map Reduce and Hadoop Distributed File System [HDFS].

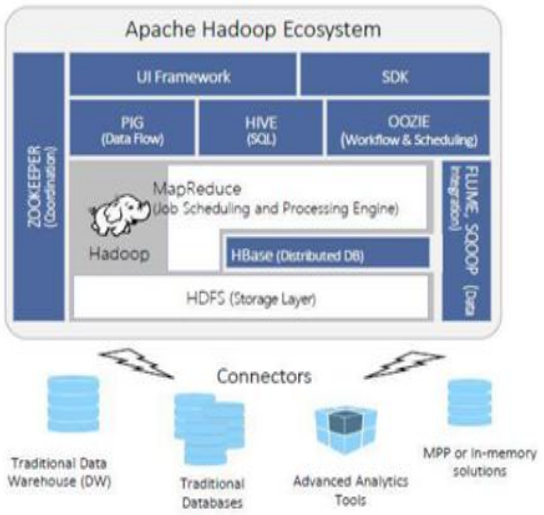

Figure 6: HDFS Architecture

HDFS is a file system used for data storage and it covers all the nodes in a Hadoop cluster. 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 [10].

II. LITERATURE SURVEY

A REVIEW ON BIG DATA SECURITY AND HANDLING

Ninny Bhogal and Shaveta Jain(2017)[1] presents a memory-and-time efficient probabilistic method for viably dealing with the massive facts to fulfil the price of records development. In addition, the Blossom channel device (Bloom filter out-BF) and its versions are compressed in regard of their commitments to the machine safety area. Also, the large data protection exam is classified against the BF and its variations. By directing a discover one of a kind avenues concerning a massive extent of records, the provided machine along a calculation for disposing of the data is tried. The results established that the BF can be utilized to overcome the effectiveness ill inside the space and-time of both ordering and breaking down tremendous information. One of our destiny works is to actualize the Dynamic Bloom filter (DBF) and examine its execution towards the Counting Bloom clear out (CouBF) with respect to the substantial records ordering and recovery. In this paper they proposed a talented and first-class-grained fact get to govern plot for big facts, where they get to method which won't launch any protection

statistics. Not the same as the contemporary techniques which simply mostly hide the characteristic values inside the get to processes, their method can shroud the entire property (as opposed to simply its features) within the get to preparations.

ALGORITHM/TECHNIQUES

- Application Software Security
- Maintenance, Monitoring, and Analysis of Audit Logs
- Secure Configurations for Hardware and Software
- Account Monitoring and Control

SECURITY AND PRIVACY IN BIG DATA

Mohammed S.Al-Kahtani(2017)[2] presents a comprehensive survey on big data network security. This work starts by introducing the distributed architecture of big data networks which focuses on the network security technologies and classifying threats related to security and privacy issues and what type of defence mechanism can be implemented to help mitigate the network vulnerabilities from Big Data and SDN. Here explored specific security areas which include big data network intrusion detection, network threat monitoring systems based on MapReduce machine-learning methods, and flow-based anomaly detection.

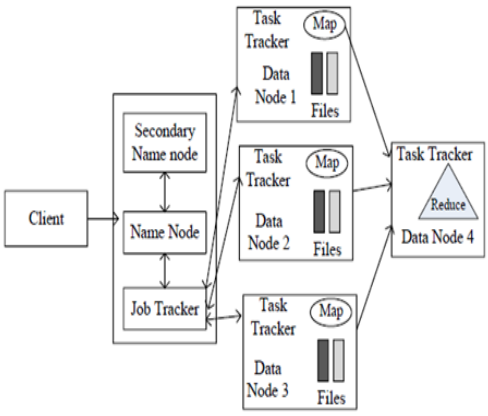

Figure 7: Hadoop based on MapReduce paradigm

ALGORITHM/TECHNIQUES

The current state of the art research in Big Data includes designing network topology, distributed algorithms, integration of software defined networks (SDN), scheduling, optimizations and load balancing among different commodity computers.

DATA ANALYTICS APPLICATION USED IN THE FIELD OF BIG DATA FOR SECURITY INTELLIGENCE

Prof. Amar Nath Singh, Er. Anurag Pattanayak, Er. Gyanachanda Samantaray(2016)[3] observed that, now a day we need a high speed processing environment. Hence we need a proper concrete solution for processing of such data. So they have go for the analysis of big data first before the processing. The technological advances in storage, processing, and analysis of Big Data include

(a) the rapidly decreasing cost of storage and CPU power in recent years

(b) the flexibility and cost-effectiveness of datacentres and cloud computing for elastic computation and storage

(c) the development of new frameworks such as Hadoop,

which allow users to take advantage of these, distributed computing systems storing large quantities of data through flexible parallel processing. Hence, by using this approach, the traditional approach is now a day's no longer used.

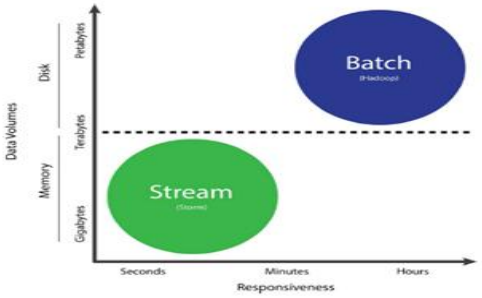

Figure 8: Batch and Stream Processing

Big Data analytics used for security purposes:

- Network Security
- Enterprise Events Analytics
- Advanced Persistent Threats Detection

ALGORITHM/TECHNIQUES

Data-driven information security dates back to bank fraud detection and anomaly-based intrusion detection systems. Fraud detection is one of the most visible uses for big data analytics. Credit card companies have conducted fraud detection for decades. However, the custom-built infrastructure to mine Big Data for fraud detection was not economical to adapt for other fraud detection uses.

BIG DATA SECURITY – THE BIG CHALLENGE

Minit Arora and Dr Himanshu Bahuguna(2016)[4] presents a survey that

organizations used various methods of de-identification to ensure security and privacy. The most common solution to ensure security and privacy may be oral and written pledges. However, history has shown that this method is flawed. Passwords, controlled access, and two factor authentication is low-level, but routinely used, technical solution to enforce security and privacy when sharing and aggregating data across dynamic, distributed data systems. Access permissions such as these can potentially be broken by both the intentional sharing of permissions and the continuation of permissions after they are no longer required or permitted. More advanced technological solution is cryptography. The famous encryption schemes have AES and RSA. Recent revelations show that the National Security Administration (NSA) may have already found ways to break or circumvent existing internet encryption schemes. Virtual barriers such as firewalls, secure sockets layer and transport layer security are designed to restrict access to data. Each of these technologies can be broken, however, and thus need to be constantly monitored, with fixes applied as needed. Tracking, monitoring or auditing software is developed to provide a history of data flow and network access by an individual user in order to ensure compliance with security related. The limitation of this technology is that it is difficult and costly to implement on a large scale or with distributed data systems and users because it requires dedicated staff to read and interpret the findings, and the software can be exploited to monitor individual behaviour rather than protecting data. Thus the traditional de-identification techniques are not applicable in the era of Big Data since the de-identification technique widespread uses. The tasks of ensuring Big Data security and privacy become more difficult as information is increased. Computer scientists have repeatedly shown that even anonymized data can often be re-identified and attributed to specific individuals.

ALGORITHM/TECHNIQUES

Privacy-preserving techniques, including privacy-preserving aggregation, operations over encrypted data, and de-identification techniques.

BIG DATA: SECURITY ISSUES AND CHALLENGES

Naveen Rishishwar, Vartika and Mr. Kapil Tomar(2017)[5] presents an overview about the Big Data security issues and challenges. Big data handles a petabyte of data or more. It has distributed redundant data storage. Can leverage parallel task processing, provide data processing (MapReduce or equivalent) capabilities and has extremely fast data insertion. Has central management and orchestration. Is hardware agnostic. Is extensible where its basic capabilities can be augmented and altered. Nothing is perfect each and every thing have their own merit and demerit (pros and cons) so big data also have their own. Some of them are given below with their possible solution.

- Storage issues
- Security
- Processing issue in Big Data
- Privacy in Big Data
- Redundancy

BIG DATA – SECURITY WITH PRIVACY

Bhavani Thuraisingham(2014)[6] presents an overview about the big data and privacy along with its security. Many privacy enhancing techniques have been proposed over the last fifteen years, ranging from cryptographic techniques such as oblivious data structures that hide data access patterns to data anonymization techniques that transform the data to make more difficult to link specific data records to specific individuals. The Privacy-Enhancing Symposium (PET) series, and journals, such as Transactions on Data Privacy. However, many such techniques either do not scale to very large data sets and/or do not specifically address the problem of reconciling security with privacy. At the same time, there are a few approaches that focus on efficiently reconciling security with privacy and they discuss them in what follows.

- Privacy-preserving data matching
- Privacy-preserving collaborative data mining
- Privacy-preserving biometric authentication

The computational, storage and communication costs of given protocols need to be considered. These costs could be especially significant for privacy-preserving protocols that involve cryptography. Given these three dimensions, one can imagine a multi-objective framework where different dimensions could be emphasized:

- Maximize utility, given risk and costs constraints
- Minimize privacy risks, given the utility and cost constraints
- Minimize cost, given the utility and risk constraints

Comprehensive solutions to the problem of security with privacy for big data require addressing many research challenges and multidisciplinary approaches. They outline significant directions in what follows:

Data Confidentiality: Several data confidentiality techniques and mechanisms exist – the most notable being access control systems and

encryptions. Both techniques have been widely investigated. However, for access control systems for big data we need approaches for:

- Merging large numbers of access control policies
- Automatically administering authorizations for big data and in particular for granting permissions
- Enforcing access control policies on heterogeneous multi-media data
- Enforcing access control policies in big data stores
- Automatically designing, evolving, and managing access control policies

SURVEY ON SECURITY ISSUES OF GROWING TECHNOLOGY: BIG DATA

Trupti V. Pathrabe(2017)[7] presents a comprehensive survey on security issues of growing technology which is related to Big Data. The analysis of Big Data involves multiple distinct phases which include data acquisition and recording, information extraction and cleaning, data integration, aggregation and representation, query processing, data modelling and analysis and interpretation. Each of these phases introduces challenges. Heterogeneity, scale, timeliness, complexity and privacy are certain challenges of big data.

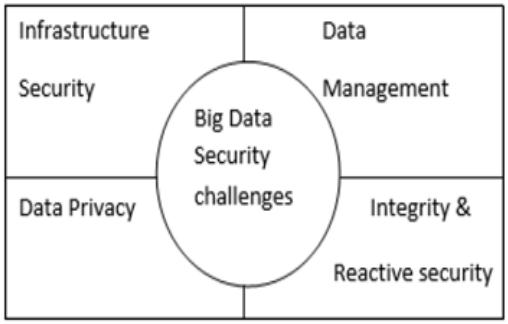

Figure 9: Challenges in Big Data Security

There are four different aspects of Big Data security:

- Infrastructure Security,
  - Security for Hadoop
  - Availability
  - Architecture Security
  - Group Communication
  - Communication Security
  - Authentication
- Data Privacy,
  - Cryptography
  - Access Control
  - Confidentiality
  - Privacy-Preserving Queries
  - Privacy in Social Networks
  - Anonymization
  - Differential Privacy
- Data Management,
  - Security at Collection or Storage
  - Policies, Laws, or Government
  - Sharing Algorithms
- Integrity and Reactive Security
  - Integrity
  - Attack Detection
  - Recovery

SECURITY ISSUES ASSOCIATED WITH BIG DATA IN CLOUD COMPUTING

Venkata Narasimha Inukollu, Sailaja Arsi and Srinivasa Rao Ravuri(2014)[8] presents a paper on security issues associated with big data in cloud computing. They design a system that can scale to handle a large number of sites and also be able to process large and massive amounts of data. However, state of the art systems utilizing HDFS and MapReduce are not quite enough/sufficient because of the fact that they do not provide required security measures to protect sensitive data. 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 challenges of security in cloud computing environments can be categorized into network level, user authentication level, data level, and generic issues.

Here, they come up with some approaches in providing security. Moreover, Hadoop framework is used to solve problems and manage data conveniently by using different techniques such as combining the k-means with data mining technology. They present various security measures which would improve the security of cloud computing environment. Since the cloud environment is a mixture of many different technologies, they propose various solutions which collectively will make the environment secure. The proposed solutions encourage the use of multiple technologies/ tools to mitigate the security problems. Security recommendations are designed such that they do not decrease the efficiency and scaling of cloud systems.

ALGORITHM/TECHNIQUES

Following security measures should be taken to ensure the security in a cloud environment: File Encryption, Network Encryption, Logging,

Software Format and Node Maintenance, Nodes Authentication, Rigorous System Testing of Map Reduce Jobs, Honeypot Nodes, Layered Framework for Assuring Cloud, Third Party Secure Data Publication to Cloud, Access Control.

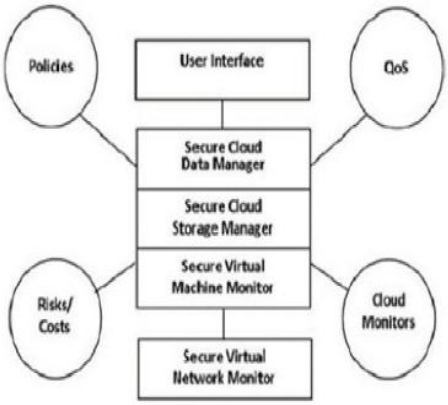

Figure 10: Layered Framework for Assuring Cloud

SECURITY PERSPECTIVES ON DEPLOYMENT OF BIG DATA USING CLOUD: A SURVEY

R.Kalaivani(2017)[9] presents a survey about the security perspectives on deployment of big data using cloud. Cloud computing not only provides facilities for the computation and processing of big data but also serves as a service model. Cloud computing infrastructure can serve as an effective platform to address the data storage required to perform big data analysis. Cloud computing is correlated with a new pattern for the provision of computing infrastructure and big data processing method for all types of resources available in the cloud through data analysis. Several cloud-based technologies had to cope with this new environment because dealing with big data for concurrent processing had become increasingly complicated. The security issues associated with cloud computing devices and environments can be categorized into the following: network level, user authentication level, data level, and generic issues.

Since cloud computing technologies was combination of various technologies, here listed few of the security measures that would protect big data in cloud environment: file encryption, network encryption, logging, software format and node maintenance, nodes authentication, rigorous system testing of map reduce jobs, honeypot nodes.

STUDY AND ANALYSES OF SECURITY LEVELS IN BIG DATA AND CLOUD COMPUTING

K.P.Maheswari, P.Ramya and S.Nirmala Devi(2017)[10] provides a study and analyses of security levels in big data and cloud computing. The big data issues are most acutely felt in certain industries and in certain government activities. The security issues of big data systems and technologies are also applicable to cloud computing because it is very important for the network which interconnects the systems. In addition, resource allocation and memory management algorithms also have to be secure. Data mining techniques can be used in the malware detection in clouds. The challenges of security in cloud computing environments can be categorized into four levels

(I) Network Level
(Ii) User Authentication Level
(Iii) Data Level
(Iv) Generic issues

There is always a possibility of occurrence of security violations by unintended, unauthorized access or inappropriate access by privileged users. Some of the ways to avail authentication are

(i)Use authentication methods such as Kerberos etc.
(ii) Encrypt the file, which provides CIA Traids of security (Confidentiality, Integrity, Availability)
(iii)Access controls implementation: by providing privileges for user or system to enhance security.
(iv)Use key management, to distribute keys and certificates and manage different keys for each group, application, and user.
(v) Logging helps to detect attacks, diagnose failures, or investigate unusual behaviour and activities can be recorded.

ALGORITHM/TECHNIQUES

Data mining techniques can be used in the malware detection in clouds. Security of big data can be enhanced by using the techniques of authentication, authorization, encryption and audit trails.

III. REVIEW ON BIG DATA SECURITY IN HADOOP

Mr. Shrikant Rangrao Kadam and Vijaykumar Patil(2017)[11] provides on review on big data security in Hadoop. Following are the areas where threat identify in Hadoop-Hadoop does enforce authenticate any user or service, data node can't have any access control mechanism to protect data block, an attacker can presence as Hadoop service, super-user of system does anything without checking, an executing MapReduce may use the host operating system interfaces.

ALGORITHM/TECHNIQUES

Hadoop engineering comprises of an ace and all others are slaves. Ace contains NameNode that oversees metadata and get to control to record framework for mapping, DataNode and square of

document, slaves are DataNode which store information. The HDFS contains information in piece of settled size, of course square size is 64 MB. Every square is recreated three circumstances in various DataNode, even in the wake of preparing or each time Hadoop keeps up replication figure three. Hadoop give MapReduce programming model which split employment into different assignments (guide or lessen) to process more than one HDFS information hinders in parallel. HDFS underpins a compose once-read-many model.

Secure Hadoop encode each record before written in HDFS. It is accounted for that each data node or slave is an item server which perform encryption or decoding at nearby site utilizing its CPUs. Propelled Encryption Standard (AES) is most well-known calculation that bolster square figure, henceforth it is appropriate for HDFS pieces. AES accessible with 128- piece AES, 192-piece AES and 256-piece AES, 128-piece AES is utilized the greater part of times in light of its straightforwardness. There are distinctive methods of operations of AES: ECB, OFB, CTR, XTS and CBC. It is accounted for that AES: ECB is great decision of encryption or unscrambling calculation since its simultaneously played out a calculation in a disseminated domain.

- Encryption in HDFS.

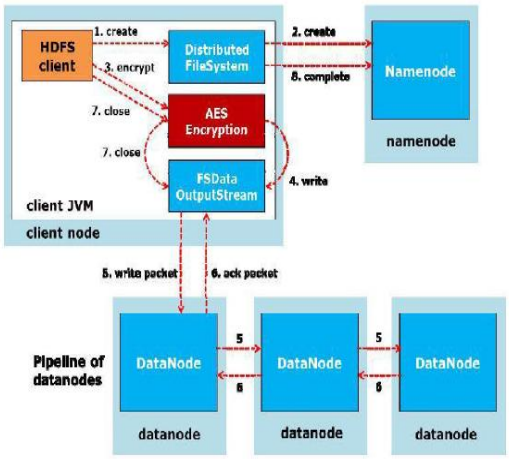

Figure 11: Encryption Process

Above figure indicates operation that spare each piece into HDFS, customer split every record into settled size square and scrambles it before transfer to Hadoop document framework. It is accounted for that encryption and unscrambling can be actualized essentially by utilizing Java class. Customers, itself perform encryption utilizing AES calculation on the CPU and exchange encoded piece to HDFS (DataNode). At that point collector DataNode (First DataNode where piece store) reproduce hinder into two different DataNodes.

- Decryption in HDFS.

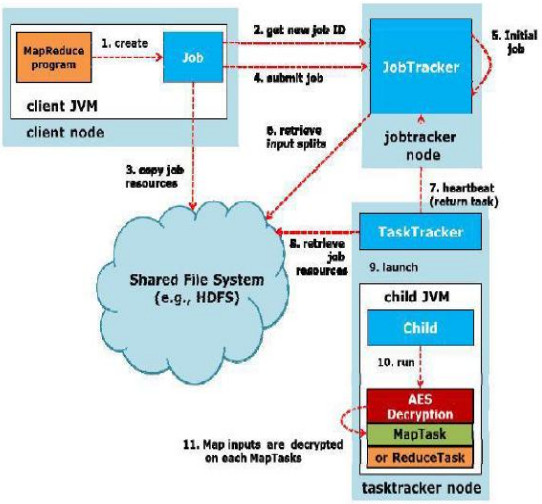

Figure 12: Decryption Process

Information pieces are composed by customer to DataNode successively, however amid execution of MapReduce occupation numerous squares are perused (decoded) parallel at TaskTracker. Above figure demonstrates that MapTask read and encode information obstructs at TaskTracker utilizing AES encryption strategy. It is accounted for that various MapTasks are executing in Hadoop at specialist destinations. HDFS bolsters compose once-read-many model, it is accounted for that simultaneous decoding of HDFS square well reasonable for some MapReduce employments.

SECURITY AND PRIVACY ASPECTS IN MAPREDUCE ON CLOUDS: A SURVEY

Philip Derbeko, Shlomi Dolev, Ehud Gudes and Shantanu Sharma(2016)[12] provides a survey on security and privacy aspects in MapReduce on clouds. The existing and proposed systems in their survey are discussed below.

Existing security algorithms for MapReduce:

- ➢ Security threats in MapReduce
  - o Impersonation attack
  - o Denial-of-Service (DOS) attack
  - o Replay attack
  - o Eavesdropping.
  - o Man-in-the-Middle (MiM) attacks
  - o Repudiation
- ➢ Security requirements in MapReduce
  - o Authentication, authorization, and access control of mappers and reducers
  - o Availability of data, mappers, and reducers
  - o Confidentiality of computations and data
  - o Integrity of computations and data
  - o Verification of outputs
  - o Accounting and auditing of computations and data

- Adversarial models for MapReduce security
  - Honest-but-Curious adversary
  - Malicious adversary
  - Knowledgeable adversary
  - Network and nodes access adversary

Privacy challenges in MapReduce computing:

- Data privacy protection from adversarial cloud providers
- Protection of data from adversarial users
- Multiusers on a single public cloud

Privacy requirements in MapReduce:

- Protection of data providers
- Untrusted cloud providers
- Utilization and privacy tradeoff
- Efficiency

Adversarial models for MapReduce privacy:

- Honest-but-curious adversary
- Malicious adversary
- Knowledgeable adversary
- Network and node adversary

Proposed solutions for privacy in mapreduce:

Some existing solutions for privacy in MapReduce. They categorize privacy algorithms in MapReduce into three types, as follows:

i. Algorithms for ensuring privacy in hybrid clouds,
ii. Algorithms ensuring data privacy in the presence of adversarial users,
iii. Algorithms for ensuring privacy in the presence of adversarial cloud providers

Proposed solutions for securing mapreduce

- Authentication, authorization, and access control based approaches
- Apache Knox
- Apache Sentry
- Apache Ranger
- Project Rhino
- Apache Accumulo
- Airavat
- Vigiles
- Guardmr

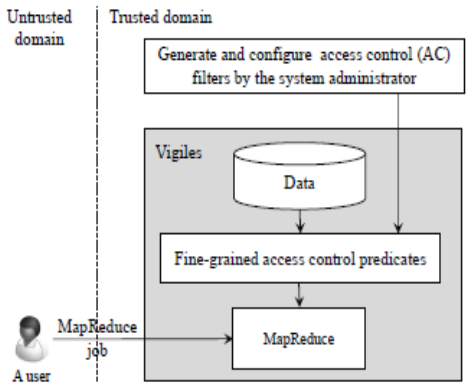

Figure 13: Vigiles access control mechanism.

- An encryption-decryption based approach for data transmission
- Secure Data Migration (SecDM)
  - Approaches for security and integrity of storage
- IBigTable: an enhancement of BigTable, called iBigTable
- HDFS-RSA and HDFS-Pairing
- Security Architecture of Private Storage Cloud based on HDFS(SAPSC)

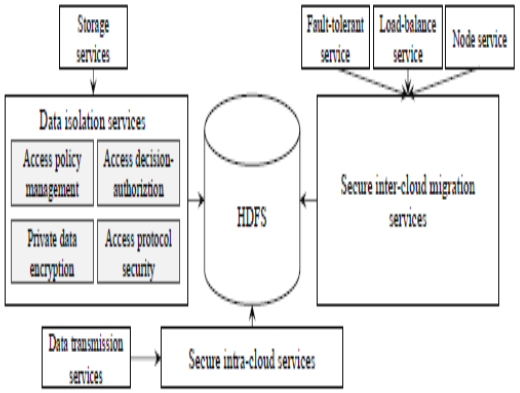

Figure 14: SAPSC architecture for data security.

  - Approaches for result verification and accounting
- Redundancy based approach
  - ClusterBFT(Cluster Byzantine Failure Tolerant)
  - SecureMR
  - Overhead issues
- Redundancy with trust based approaches
  - Accountable MapReduce
  - Verification-based Integrity Assurance Framework (VIAF)
  - Cross Cloud MapReduce (CCMR)
  - IntegrityMR
  - Verification-Based Anti-Collusive Worker Scheduling (VAWS).
  - Hatman
  - TrustMR
  - Trusted Sampling-based Third-Party Result Verification (TS-TRV) Overhead issues.
- Log analysis and Watermarking-based approaches
  - Restrictive issues

A SURVEY PAPER ON SECURITY ISSUE WITH BIG DATA ON ASSOCIATION RULE MINING

Asha Patel(2017)[13] Presents a survey paper on security issue with big data on association Rule Mining. Association Rule Mining is the

technique of extracting frequent mining from correlation for database transaction. In data mining, the process of privacy preserving has played a vital role. It helps in providing the security to the sensitive information or knowledge and protecting information from unauthorized access without affecting the security of the data. Now a day's people are aware of the privacy intrusions on their personal data and they do not share their sensitive information to unauthorized people. Lack of privacy may generate the unintentional results. Several methods have been proposed in privacy but still it has its significance. The results of privacy preserving data mining algorithms is explained in terms of its data utility, performance, or level of uncertainty to data mining algorithms etc. There is no privacy preserving algorithms exists that exceed other algorithms on all possible criteria like utility, cost, complexity, performance, tolerance against data mining algorithms etc. In case of horizontally partitioned dataset the security is not provided for distributed privacy preserving association rule mining.

ALGORITHM/TECHNIQUES

Apriori and FP Growth algorithm are applied to analyse the performance and security. The results produced by the FP growth algorithm are better than the Apriori algorithm. The combination of the horizontal and vertical partitioning of the dataset is known as the hybrid partitioning. When privacy is provided to both horizontal and vertical partitioned datasets in distributed and centralized scenario can improve the accuracy which overcomes the accuracy problem in the vertical partitioning. Association rule mining is used to group the related items and preserving the individual data privacy without compromise the accuracy of global data mining task and global association patterns were driven from the distributed data. Global rules are generated after the vertical partitioning of the dataset and percentage of missed rules and percentage of spurious rules were calculated. When two party algorithm is used with minimum support level, it will efficiently discover frequent item sets without revealing individual transaction values. It will achieve good individual security.

BIG DATA AND DATABASE SECURITY

Vinod B. Bharat, Pramod B. Deshmukh, Laxmikant S. Malphedwar, P. Malathi and Nilesh N. Wani(2017)[14] presents an idea about Big Data and Database Security. In this paper they concentrated on the huge information security and protection challenges. They concentrated on survival security professional oriental exchange diaries to centre an underlying rundown of high-need security and protection issues and landed at the accompanying main ten difficulties.

1. Secure calculations in disseminated programming structures
2. Security best practices for non-social information stores
3. Secure information stockpiling and exchanges logs
4. End-point info acceptance/sifting
5. Ongoing security observing
6. Adaptable and compostable security saving information mining and examination
7. Cryptographically upheld information driven security
8. Granular access control
9. Granular reviews
10. Information provenance

ALGORITHM/TECHNIQUES

This paper has uncovered the real security issues that should be tended to in Big Data handling and capacity. A few specialists have realized the utilization of encryption together with Kerberos convention keeping in mind the end goal to make the information more secure. In any case, these security and protection issues come in various structures such that Kerberos won't not be sufficient to completely secure the information. Amid Map-Reduce system in Hadoop, mapper hubs handle a given arrangement of information and recovery the middle person information inside their nearby documents. The reducer hubs will then duplicate this information from the mapper hubs and later on total it to create the general result. We might want to present an extra focal hub which interfaces with both the mapper and the reducer hubs. The delegate information will then be put away in this hub rather than the mapper hubs' nearby record framework. An edge safeguard component will then be utilized to screen all the movement going into and out of the hub to secure the information.

INFORMATION SECURITY ISSUES IN BIG DATA: SOLUTION USING PPDM (PRIVACY PRESERVING DATA MINING)

Sanchita Gupta, Akashkataria, Shubham Rathore and Dharmendra Singh Rajput(2016)[15] discussed about information security issues in big data and provides a solution using PPDM (Privacy Preserving Data Mining). The expression "information gaining" is regarded as an equivalent word for another term "Knowledge Discovery from Data" (KDD) which highlights the objective of the mining procedure.

Step 1: Data pre-processing
Step 2: Data transformation
Step 3: Data mining

Step 4: Pattern assessment and presentation

- Privacy Preserving Data Mining

The target of PPDM is to defend delicate data from spontaneous or unsanctioned revelation, and in the meantime, save the utility of the information.

- Privacy Preserving in Big Data Analytics
- Big mobile data
- Health care data
- Web-based networking data
- Information supplier
- Information authority
- Data miner
- Decision maker

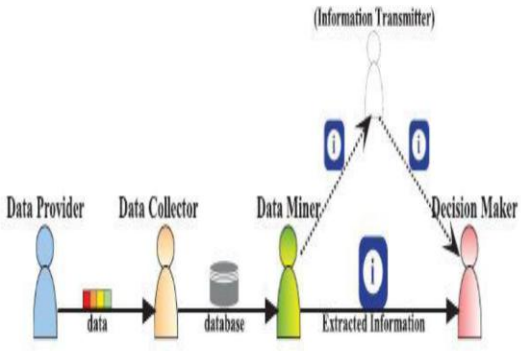

Figure 15: Application Scenario.

In the proposed system, they specified the following,

1. For information supplier,

The security safeguarding goal is to viably control the measure of touchy information uncovered to others. To accomplish this objective, he can use security instruments to utmost other's entrance to his information, offer his information at sale to get enough pay for protection misfortune, or distort his information to shroud his actual character.

2. For information gatherer

The security protecting target is to discharge helpful information to information diggers without unveiling information suppliers' characters and delicate data about them. To accomplish this objective, he needs to create appropriate security models to measure the conceivable loss of protection under various assaults, and apply anonymization systems to the information.

3. For data miner

The security protecting target is to get right information mining comes about while keep delicate data undisclosed either during the time spent information mining or in the mining comes about. To accomplish this objective, he can pick a legitimate technique to alter the information before certain mining calculations are connected to, or use secure calculation conventions to guarantee the wellbeing of private information and delicate data contained in the scholarly model.

4. For decision maker

The protection saving goal is to make a right judgment about the believability of the information mining results that has found. To accomplish this objective, he can use provenance strategies to follow back the historical backdrop of the got data, or manufacture classifier to separate genuine data from false data.

ALGORITHM/TECHNIQUES

Data Mining, Big Data, PPDM-Privacy Preserving Data Mining, Security, Pattern Analysis, Business Intelligence, Knowledge Discovery Techniques.

SECURITY AND PRIVACY – A BIG CONCERN IN BIG DATA A CASE STUDY ON TRACKING AND MONITORING SYSTEM

Tilwani Mashook, Patel Malay and Pooja Mehta(2017)[16] presents a case study on Tracking and Monitoring System based on Security and Privacy in Big Data. The main goals of Employee Tracking Systems are to monitor the employees or the field labourers and help them analyse themselves also to let the organizations to analyse their performance. The raw data obtained from the servers is processed online or offline for detailed analysis at the remote server according to the application requirements. Nine Big Data Security Challenges for tracking and monitoring applications:

- Most distributed systems' computations have only a single level of protection, which is not recommended.
- Non-relational databases (NoSQL) are actively evolving, making it difficult for security solutions to keep up with demand.
- Automated data transfer requires additional security measures, which are often not available.
- 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.
- Unethical IT specialists practicing information mining can gather personal data without asking users for permission or notifying them.
- Access control encryption and connections security can become dated and inaccessible to the IT specialists who rely on it.
- 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.
- Due to the size of Big Data, its origins are not consistently monitored and tracked.

Some of the latest challenges observed in the Big Data by the Tracking and Monitoring Application's Organization: user data privacy, granular access, monitoring in real-time, granular audits, preserve the privacy in data mining and analytics, encrypted data-centric security, data provenance and verification, integrity and reactive security.

ALGORITHM/TECHNIQUES

During the transition phase, the EHR vendor must work closely with the healthcare provider for a smooth and secure transition. The company should provide some type of comprehensive user guide for the users in the provider's practice for implementing some of the software's or applications in the device of the human who is supposed to be tracked or monitored.

A FRAMEWORK ON SECURITY AND PRIVACY-PRESERVING FOR STORAGE OF HEALTH INFORMATION USING BIG DATA

J.L. Joneston Dhas, S. Maria Celestin Vigila and C. Ezhil Star(2017)[17] designed a framework on Security and Privacy-Preserving for Storage of Health Information using Big Data. There are many real time problems when we store the health record as a big data. The first is how a user will protect the information in the cloud. The next one is how to identify the record and how to protect the health information from the unauthorised user. The size of the data is the main challenge for big data. Other challenges faced by the health information are speed, variety and heterogeneity of data. The system must mine, process the data and change to make decision making from that data. The data is coming from different sources and there are different types of people use the cloud. Some are trusted and some are untrusted. So privacy preservation, data auditing and data protection should be achieved for electronic health information. So a public auditing should be done periodically and the integrity of the data is verified. An efficient access control mechanism should be provided to control the unauthorised user.

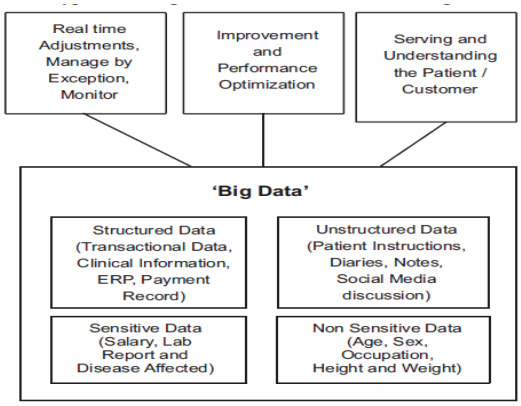

Figure 16: Basic task done in big data which is apply in health data

In the proposed system a secured system model is shown in the below figure. It consists of the following entities: Owner, Users (Hospital), Authorities (AA), Data Receiver, Public Key (PU), Private Key (PR), Cloud Server, Access Control, Identify Receiver, Result Verification.

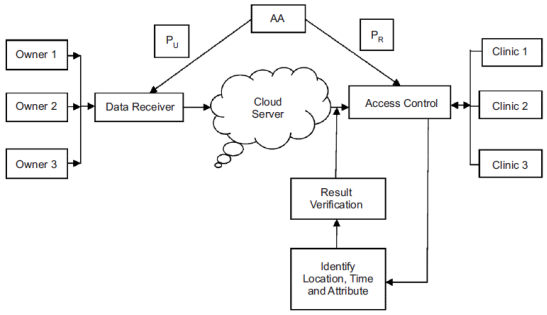

Figure 17: Secured Cloud Server Framework for Health Record

In this framework the owner will be a patient or the hospital where the patient takes the treatment. All the patient information will be taken regularly and it will be processed by the data receiver and it will do the several processes like encryption, compression, analysis etc. Authorization agent generates a public key and private key and distribute to data receiver and access control. After processing the data, it will be stored in the cloud server. Data retrieval will be done in the clinic. It will have a strong access control and only the authorized person will be able to access the information. It will have a location attribute and the user will be able to access the information only in the particular time and location. So the unauthorized person will be able to access the information. Once the person wants to access the information first he will be work in a particular location and it will be identified through GPS (Global Positioning System) and some other object will also be taken as the attributes. All the location

and the time will be identified by result verification. Once all the attributes are verified as correct the decryption key will be given to the user to decrypt the needed information by the data receiver. This framework ensures an efficient process of big data and provides a good relationship between the patient and the doctor. Since all the patient information are available up to date. So the doctor can easily predict the patient disease and provide the treatment quickly and efficiently. In this framework all the patient records are securely stored and retrieved by an authorized person only.

ALGORITHM/TECHNIQUES

To secure the health data, different techniques such as authentication, digital ware marking and MPEG encryption schemes are used.

IV. CONCLUSION

The study of various methodologies by many researchers are making the data secured and provide privacy which made clear about the various methods, its merits and demerits and inabilities for providing security and privacy in Big Data. With this, we can come to conclude that we required some new technologies or the considerable modifications in the available technology.

V. FUTURE ENHANCEMENT

The following are some of the future enhancements which I have found while referring these papers. To reinforce big data security- focus on software protection, in location of tool safety. Isolate gadgets and servers containing important facts. Introduce real-time security data and event control. Provide reactive and proactive protection [1]. Traffic management, monitoring, additional anomaly-detection security strategies such as MapReduce machine learning must be used to help mitigate collective threats such as botnet and DDoS attacks. In order to keep the system packet flows consistent flow-based intrusions detection is sought after [2]. Another major thing will be privacy requirements in big data collection, storage and processing [4]. Major big data security challenges are: In Big Data most distributed systems computations have only a single level of protection, which is not recommended. Non-relational databases (NoSQL) are actively evolving, making it difficult for security solutions to keep up with demand. Automated data transfer requires additional security measures, which are often not available. When a system receives a large amount of information, it should be validated to remain trustworthy and accurate [5].

A major issue arising from big data is that correlating many (big) data sets one can extract unanticipated information i.e. privacy-preserving data correlation techniques. Relevant issues and research directions that need to be investigated include

- Techniques to control what is extracted and to check that what is extracted can be used and/or shared
- Support for both personal privacy and population privacy
- Efficient and scalable privacy enhancing techniques
- Usability of data privacy policies
- Approaches for data services monetization
- Data publication
- Privacy implication on data quality
- Risk models
- Data ownership
- Human factors
- Data lifecycle framework

The challenge of detecting and preventing 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 analysing multiple data sources. Not only security but also data privacy challenges are existing in industries and federal organizations. There should be a balance between data privacy and national security [8].

Based on this survey, it is also identified several important issues and challenges that require further research, as follows:

- Extending the authorization framework (security of MapReduce).
- Integrating with a trust infrastructure (security of MapReduce). There are several domains of trust that must be made explicit and verified for MapReduce framework.
- Processing on encrypted data (security and privacy of MapReduce).
- Supporting multiple geographically distributed clusters for executing a single job (security and privacy of MapReduce).
- Extending MapReduce algorithms with privacy preserving support (privacy of MapReduce).

Another lacking field of the research is holistic frameworks i.e., frameworks that solve more than a single problem, especially solving both the security and privacy aspects, and integrating some of the mentioned algorithms and frameworks, which provide computational security and privacy of data for MapReduce computations. It is believed that in the future, it will have MapReduce frameworks that provide multiple types of computations in information secure manner [12].

REFERENCES

[1] Ninny Bhogal, Shaveta Jain," A Review on Big Data Security and Handling", International Research Based Journal, Vol(6)-Issue(1) , ISSN 2348-1943, March,11, 2017.

[2] Mohammed S.Al-Kahtani," Security and Privacy in Big Data", International Journal of Computer Engineering and Information Technology, VOL. 9, NO. 2, E-ISSN 2412-8856, February 2017.

[3] Prof. Amar Nath Singh, Er. Anurag Pattanayak, Er. Gyanachanda Samantaray," Data Analytics Application used in the field of Big Data for Security Intelligence", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 6, Issue 1, ISSN 2278-6856, January - February 2017.

[4] Minit Arora, Dr Himanshu Bahuguna, "Big Data Security – The Big Challenge", International Journal of Scientific & Engineering Research, Volume 7, Issue 12, ISSN 2229-5518, December-2016.

[5] Naveen Rishishwar, Vartika, Mr. Kapil Tomar," Big Data: Security Issues and Challenges", International Journal of Technical Research and Applications, e-ISSN: 2320-8163, Special Issue 42 (AMBALIKA), PP. 21-25, March 2017.

[6] Bhavani Thuraisingham, "Big Data – Security with Privacy", NSF Workshop, September 16-17, 2014.

[7] Trupti V. Pathrabe, "Survey on Security Issues of Growing Technology: Big Data", IJIRST, National Conference on Latest Trends in Networking and Cyber Security, March 2017.

[8] 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.

[9] R.Kalaivani," Security Perspectives on Deployment of Big Data using Cloud: A Survey", International Journal of Advanced Networking & Applications (IJANA), Volume: 08, Issue: 05 Pages: 5-9, Special Issue, 2017.

[10] K.P.Maheswari, P.Ramya, S.Nirmala Devi," Study and Analyses of Security Levels in Big Data and Cloud Computing", International on Recent Trends in Engineering Science, Humanities and Management, February 2017.

[11] Mr. Shrikant Rangrao Kadam, Vijaykumar Patil, "Review on Big Data Security in Hadoop", International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395 -0056, Volume: 04 Issue: 01, p-ISSN: 2395-0072, Jan -2017.

[12] Philip Derbeko, Shlomi Dolev, Ehud Gudes, Shantanu Sharma, "Security and Privacy Aspects in MapReduce on Clouds: A Survey", Elsevier Computer Science Review, arXiv:1605.00677v1 [cs.DB] 2 May 2016.

[13] Asha Patel, "A Survey Paper on Security Issue with Big Data on Association Rule Mining", IJIRST, National Conference on Latest Trends in Networking and Cyber Security, March 2017.

[14] Vinod B. Bharat, Pramod B. Deshmukh, Laxmikant S. Malphedwar, P. Malathi and Nilesh N. Wani, "Big Data and Database Security", IJCTA, 10(8), pp. 517-528 ISSN: 0974-5572, International Science Press, 2017.

[15] Sanchita Gupta, Akashkataria, Shubham Rathore, Dharmendra Singh Rajput," Information Security Issues in Big Data: Solution using PPDM (Privacy Preserving Data Mining)", International Journal of Pharmacy & Technology, ISSN: 0975-766X, Vol. 8, Issue No.4, November 2016.

[16] Tilwani Mashook, Patel Malay, Pooja Mehta," Security and Privacy – A Big Concern in Big Data a case study on Tracking and Monitoring System", IJIRST, National Conference on Latest Trends in Networking and Cyber Security, March 2017.

[17] J.L. Joneston Dhas, S. Maria Celestin Vigila and C. Ezhil Star," A Framework on Security and Privacy-Preserving for Storage of Health Information Using Big Data", IJCTA, 10(03), Pp. 91-100, International Science Press, 2017.

P. Nandhini "A Research on Big Data Analytics Security and Privacy in Cloud, Data Mining, Hadoop and Mapreduce "International Journal of Engineering Research and Applications (IJERA) , vol. 8, no. 4, 2018, pp. 65-78