ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 3, March 2016**

# Survey on Security Primitives in Big Data on Cloud Computing

Prof. Chetan Andhare, Prof.Shubhangi Sonone

Assistant Professor, Dept. of IT, D.Y.Patil College of Engineering, Pune, India

**ABSTRACT:** This paper sheds a light on security primitives needed for big data on cloud computing environment. Primary concern is security in cloud computing that uses big data. Big data is large in size and so complex that conventional system applications fall short for challenges imposed by big data for its capturing, sharing, sharing,transferring, storing, privacyand security.Big data provides enormous opportunity for enterprises by giving access into new volumes and varieties of data .But without proper security primitives in place it turn into big problem. Cloud computing security is new emerging area in computer security that refers to set of policies, controls and encryption primitives to protect online data, system application and infrastructure for cloud computing. Security issues in cloud computing includes application level security, network level security, information security and data privacy. Also cloud security has to implement many different type of control such as deterrent control, preventive control, detective control and corrective controls for safeguard of its security architecture. Security for big data on cloud is most exciting domain that posing so many new security issues that we need to tackle to avail big data on cloud.

**KEYWORDS:** Cloud Computing, Big Data, Hadoop, Map Reduce, HDFS (Hadoop Distributed File System)

## 1. INTRODUCTION

It is the vital facet to firmly store, manage and share giant amounts of complicated knowledge to research and establish patterns .Cloud has associate degree external security challenge, that's the info owner won't apprehend wherever the info is precisely placed. For obtaining the advantages of cloud computing, he/she should additionally utilize the resource allocation and additionally the programming provided by the controls. Therefore it's needed to shield the info from the untruthful processes. As Cloud includes high quality, we have a tendency to believe that instead of providing a full approach we'll give future enhancements to the cloud security.For process giant amounts of information on goods hardware Google has introduced MapReduce [1] framework. Apache's Hadoop distributed classification system (HDFS) is evolving as a superior software package part for cloud computing combined at the side of integrated components like MapReduce. Hadoop, that is associate degree ASCII text file implementation of Google MapReduce, together with a distributed classification system, provides to the applying software engineer the abstraction of the map and also the cut back.With Hadoop it's easier for organizations to induce a foothold on the massive volumes of information being generated daily, however at constant time may also produce issues associated with security, knowledge access, monitoring, high accessibility and business continuity.
Some security approaches area unit being provided during this paper. we have a tendency to ought a system that may scale to handle variety of web sites and even be able to method large and large amounts of information. However, state of the art systems utilizing HDFS and MapReduce don't seem to be quite enough/sufficient due to the very fact that they are doing not give needed security measures to shield sensitive knowledge. Moreover, Hadoop framework is employed to unravel issues and manage knowledge handily by victimization totally different techniques like combining the k-means with data processing technology [3].

### 1.1 Cloud Computing

Sharing of computing resources means Cloud computing. It does not require local servers or personal devices to handle the applications. The services that are delivered through internet are known as cloud computing, as the word 'cloud' means 'internet'. Millions of instructions per second can be executed through cloud computing. A large group of servers with specialized connections to distribute data processing among the servers is used by cloud computing network. This technology requires installation of single software in each computer that allows users to log into a

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

service that is web-based and which will show all the programs required by the user. In a cloud computing system there's a significant workload shift.When it comes to running applications local computers no longer have to take the entire burden. To minimize the usage cost of computing resources [4] cloud computing technology is being used. A network of computers, the cloud network consists of handles the load instead. On the user end the cost of software and hardware decreases.The user must run the cloud interface software to connect to the cloud. Front end and back end services are there in cloud computing. The user's computer and software required to access the cloud network are there in front end services. Different computers, servers and database systems that create the cloud are there in the back end services. The cloudapplications can be accessed anywhere, anytime by using the Internet. Cloud Computing is excessively used by Gmail, Google Calendar, Google Docs and Dropbox etc.

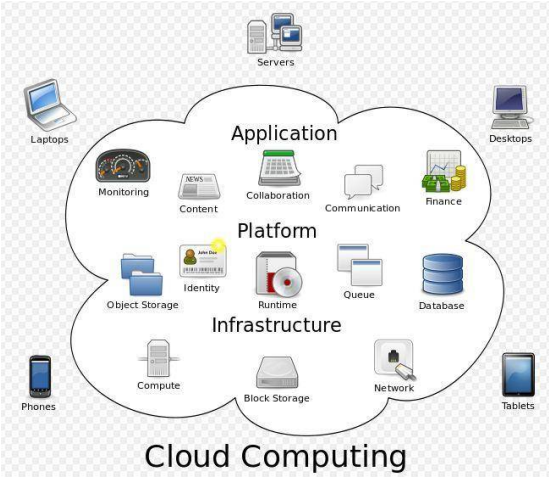


Fig1. Cloud Computing

**1.2 Big Data**

The term ‘Big Data’ means huge volume, high velocity and a variety of data. This big data is increasing tremendously day by day. Traditional data management systems and existing tools are facing difficulties to process such a big data.

The three main pillars that of Big Data are:

a) Volume: Huge amount of data is generated during big data applications.

b) Variety: The data may be structured, semi structured or unstructured such as documents, video, audio, transactions etc.,

c) Velocity: For time critical applications the faster processing is very important.
Example-share marketing, video streaming etc.

The other important aspects related to Big Data are Variability and Complexity [5].

d) Variability: The data flow can be highly inconsistent with periodic time being along with the Velocity.

Complexity: When the data is coming from multiple sources complexity of the data also needs to be considered. Before actual processing the data must be connected, matched, cleansed and transformed into required formats that is given. Large amounts of data collection are supported by today technologies. It also helps in utilizing such data efficiently and effectively. The real time examples of Big Data are Credit card transactions made all over the world with respect to a

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Bank, social interaction data generated by Facebook users.


Fig2. Big Data

The words such as “Hadoop” and “MapReduce” cannot be avoided while considering importance about Big Data.

**1.3 Hadoop**

Framework that supports the process of enormous sets of knowledge in a very distributed computing setting is understood asHadoop that could be a free, Java-based programming framework. it’s sponsored by the Apache software package Foundation. Master/Slave structure [6] is employed by Hadoop cluster. Victimization Hadoop, massive knowledge sets that may be processed across a cluster of servers and applications may be run on systems with thousands of nodes involving thousands of terabytes. In Hadoop fast knowledge transfer is completed and permits the system to continue its traditional operation even within the case of some node failures victimization Distributed filing system. Within the case of a big range of node failures this approach lowers the chance of a whole system failure. A computing resolution that's ascendable, price effective, versatile and fault tolerant is enabled by Hadoop. Well-liked firms like Google, Yahoo, Amazon and IBM etc., to support their applications involving large amounts of knowledge uses Hadoop Framework. Map scale back and Hadoop Distributed filing system (HDFS) area unit the 2 main comes in Hadoop Framework.

**1.4 Map Reduce**

Hadoop Map scale back may be a framework [7] accustomed write applications that method massive amounts of knowledge in parallel on clusters of goods hardware resources in a very reliable, fault-tolerant manner. A Map scale back job initial divides the info into individual chunks that are processed by Map jobs in parallel. The outputs of the maps sorted by the framework are then input to the scale back tasks. Typically the input and also the output of the task are each hold on in a very file-system. Scheduling, observance and re-executing failing tasks are taken care by the framework.

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

HDFS [8] is basically designed for storage of very big datasets consistently and to make available those respective datasets to end users at very high bandwidth. HDFS offers a structure for alteration and investigation of huge datasets with the support of MapReduce model. It is basically designed for implementation on service hardware platforms.

**1.6 Big data applications**

The big knowledge application refers to massive themassive the big} scale distributed applications that typically work with large knowledge sets. Knowledge exploration and analysis was a troublesome drawback in several sectors within the span of huge knowledge. With giant and sophisticated knowledge, computation becomes troublesome to be handled

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 3, March 2016**

by the ancient processing applications which trigger the event of huge knowledge applications [9]. Google's map cut back framework and apache Hadoop are the defector package systems [10] for giant knowledge applications, during which these applications generates an enormous quantity of intermediate knowledge. Producing and Bioinformatics are the 2 major areas of huge knowledge applications.
Big knowledge offer associate degree infrastructure for transparency in producing business that has the power to unravel uncertainties like inconsistent part performance and handiness. In these massive knowledge applications, a abstract framework of prognostic producing begins with knowledge acquisition wherever there's a clear stage to amass differing types of sensory knowledge like pressure, vibration, acoustics, voltage, current, and controller knowledge. the mixture of sensory knowledge and historical knowledge constructs the massive knowledge in producing. This generated massive knowledge from the on top of combination acts because the input into prognostic tools and preventive methods like prognostics and health management.Another necessary application for Hadoop is Bioinformatics that covers subsequent generation sequencing and alternative biological domains. Bioinformatics [11] which needs an outsized scale knowledge analysis uses Hadoop. Cloud computing gets the parallel distributed computing framework at the side of laptop clusters and internet interfaces.

**1.7 Big data advantages**

In huge knowledge, the code packages give a chic set of tools and choices wherever a personal might map the whole knowledge landscape across the corporate, so permitting the individual to investigate the threats he/she faces internally. this can be thought-about joined of the most benefits as huge knowledge keeps the information safe. With this a personal will be able to sight the possibly sensitive data that's not protected in Associate in nursing acceptable manner and makes positive it's keep in keeping with the regulative needs.

There are some common characteristics of massive knowledge, such as
a) Unstructured and structured knowledge each are integrated by huge knowledge.
b) Addresses speed and measurability, quality and security, flexibility and stability.
c)In huge knowledge the belief time to data is vital to extract worth from numerous knowledge sources, as well as mobile devices, oftenest identification, the net and a growing list of machine-driven sensory technologies.

All the organizations and business would have the benefit of speed, capacity, and quantifiability of cloud storage. Moreover, finish users will visualize the information and firms will notice new business opportunities. Another notable advantage with big-data is, information analytics, which permit the individual to individualize the content or look and feel of the web site in real time so it suits the every client coming into the web site. If massive information area unit combined with prophetical analytics, it produces a challenge for several industries. the mix ends up in the exploration of those four areas:

a) Calculate the risks on massive portfolios
b) Detect, prevent, and re-audit money fraud
c) Improve delinquent collections
d) Execute high price promoting campaigns

**1.8 Necessity of security in huge information**

For selling and analysis, several of the companies uses huge information, however might not have the basic assets significantly from a security perspective. If a security breach happens to huge information, it wouldresulting even additional serious legal repercussions and reputational harm than at the present. during this new era, several corporations square measure victimization the technology to store and analyze petabytes of information concerning their company, business and their customers. As a result, data classification becomes even additional vital. for creating huge information secure, techniques like cryptography, logging, Proteacynaroides detection should be necessary. In several organizations, the preparation of massive information for fraud detection is extremely engaging and helpful.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 3, March 2016**

The challenge of sleuthing and preventing advanced threats and malicious intruders should be solved victimization huge information vogue analysis. These techniques facilitate in sleuthing the threats within the early stages victimisation additional refined pattern analysis and analyzing multiple information sources.

Not solely security however conjointly information privacy challenges existing industries and federal organizations. With the rise within the use of massive information in business, several corporations square measure wrestling with privacy problems. Information privacy could be a liability; therefore corporations should get on privacy defensive. However not like security, privacy ought to be thought of as associate plus, so it becomes a point for each customers and different stakeholders. There ought to be a balance between information privacy and national security.

## II. MOTIVATION AND RELATED WORK

### 2.1. Motivation

Along with the increasing quality of the cloud computing environments, the protection problems introduced through adaptation of this technology also are increasing. although Cloud Computing offers several advantages, it's at risk of attacks. Attackersar systematically attempting to search out loopholes to attack the cloud computing atmosphere. the standard security mechanisms that ar used ar reconsidered as a result of these cloud computing deployments. Ability to ascertain, management and examine the network links and ports is needed to confirm security. thus there's a necessity to speculate in understanding the challenges, loop holes and elements vulnerable to attacks with relevancy cloud computing, and are available up with a platform and infrastructure that is a smaller amount at risk of attacks.

### 2.2. Related Work

Hadoop (a cloud computing framework), a Java based mostly distributed system, could be a new framework within the market. Since Hadoop is new and still being developed to feature additional options, there square measure several security problems which require to be self-addressed. Researchers have known a number of the problems and began engaged on this. a number of the notable outcomes, that is said to our domain and helped United States to explore, square measure conferred below.

The World Wide net syndicate has known the importance of SPARQL which may be utilized in numerous information sources. Later on, the concept of secured question was planned so as to extend privacy in privacy/utility trade-off. Here, Jelena, of the USC information processing Institute, has explained that the queries may be processed in keeping with the policy of the supplier, instead of all question process. Bertino et al revealed a paper on access management for XML Documents [12]. within the paper, cryptography and digital signature technique square measure explained, and techniques of access management to XML information document is stressed for secured setting. Later on, he revealed another paper on authentic third party XML document distribution [13] that obligatory another trusty layer of security to the paradigm.Kevin Hamlen and et al planned that information may be keep during a information encrypted instead of plain text. The advantage of storing information encrypted is that albeit persona non grata will get into the information, he or she can't get the particular information. But, the disadvantage is that encoding needs plenty of overhead. rather than process the plain text, most of the operation can happen in scientific discipline kind. thus the approach of process in scientific discipline kind additional to security layer.Airavat [14] has shown North American nation some important advancement security within the Map reduce setting. Within the paper, Roy and et al have used the access management mechanism along side differential privacy. they need worked upon mathematical sure potential privacy violation that prevents data leak on the far side information provider's policy. The on top of works have influenced North American nation, and that we area unit analyzing varied approaches to create the cloud setting safer for information transfer and computation..

## III. THE PROPOSED APPROACHES

We gift varied security measures which might improve the protection of cloud computing setting. Since the cloud atmosphere could be a mixture of the many totally different technologies, we tend to propose varied solutions that put together can build the surroundings secure. The projected solutions encourage the utilization of multiple technologies/

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

tools to mitigate the protection drawback laid out in previous sections. Security recommendations area unit designed such they are doing not decrease the potency and scaling of cloud systems.

Following security measures should be taken to confirm the safety during a cloud surroundings

**3.1 File cryptography**

Since the information is gift within the machines during a cluster, a hacker will steal all the essential data. Therefore, all the information keep ought to be encrypted. totally {different completely different} cryptography keys ought to be used on different machines and therefore the key data ought to be keep centrally behind robust firewalls. This way, notwithstanding a hacker is ready to induce the information, he cannot extract meaning data from it and misuse it. User knowledge are keep firmly in associate degree encrypted manner.

**3.2Network encoding**

All the network communication should be encrypted as per trade standards. The RPC procedure calls that happen ought to happen over SSL so even though a hacker will faucet into network communication packets, he cannot extract essential data or manipulate packets.

**3.3 Logging**

All the map scale back jobs that modify the information ought to be logged. Also, the data of users, that area unit to blame for those jobs ought to be logged. These logs should be audited frequently to search out if any, malicious operations area unit performed or any malicious user is manipulating the information within the nodes.

**3.4 Software Format and Node Maintenance**

Nodes that run the software system ought to be formatted often to eliminate any virus gift. All the applying software systems and Hadoop software ought to be updated to create the system safer.

**3.5 Nodes Authentication**

Every time cluster is formed by a node, it ought to be genuine .just in case of a malicious node, it shouldn't be allowed to hitch the cluster. Authentication techniques like Kerberos is accustomed to validate the approved nodes from malicious ones.

**3.6 Rigorous System Testing of MapReduce Jobs**

After a developer writes a map scale back job, it ought to be totally tested in an exceedingly distributed setting rather than one machine to make sure the hardiness and stability of the task.

**3.7 honeypot Nodes**

Honey pot nodes ought to be present within the cluster, that appear as if an everyday node however may be a entice. These honeypots lure the hackers and necessary actions would be taken to eliminate hackers.

**3.8 Layered Framework for reassuring Cloud**

A superimposed framework for reassuring cloud computing [16] as shown in Figure (1) has the secure virtual machine layer, secure cloud storage layer, highly secure cloud information layer, and also the most secure virtual network monitor layer. Cross cutting services are rendered by the policy layer, the cloud watching layer, the irresponsibleness layer and also the risk analysis layer.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

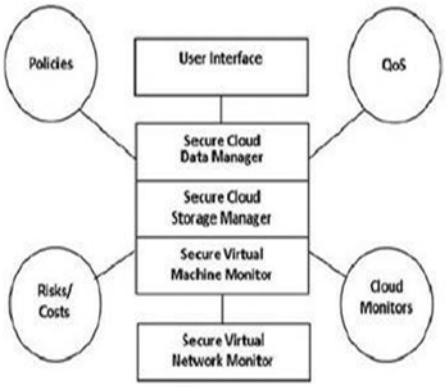


Fig3: Layered framework for assuring cloud [16]

### 3.9 Third Party Secure information Publication to Cloud

Cloud computing helps in storing of knowledge at an overseas web site in order to maximize resource allocation. Therefore, it's vital for this information to be protected and access ought to incline solely to approved people. thence this essentially amounts to secure third party publication {of information ofknowledge of information} that's needed for data outsourcing, likewise as for external publications. within the cloud surroundings, the machine serves the role of a 3rd party publisher, that stores the sensitive information within the cloud. This information has to be protected, and therefore the on top of mentioned techniques ought to be applied to confirm the upkeep of credibility and completeness.

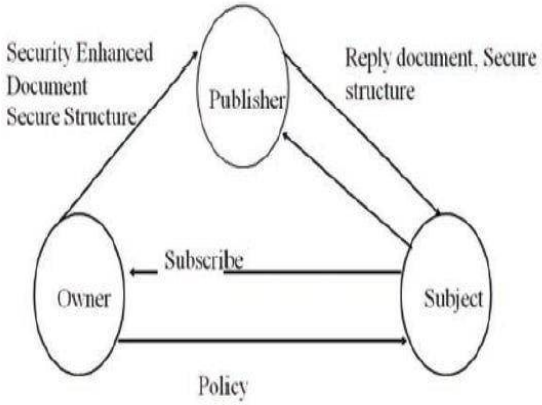


Fig4: Third party secure data publication applied to cloud [16].

### 3.10 Access management

Integration of obligatory access management and differential privacy in distributed surroundings are going to be an honest security live. knowledge suppliers can management the safety policy of their sensitive knowledge. they're going to additionally management the mathematical certain on privacy violation that might happen. within the on top of approach, users will perform knowledge computation with none escape of knowledge. to stop data leak, SELinux [17] are going to be used. SELinux is nothing however Security-Enhanced UNIX system which could be a feature that has the mechanism for supporting access management security policy through the employment of UNIX system Security

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 3, March 2016**

Modules (LSM) within the UNIX system Kernel.

Enforcement of differential privacy are going to be done victimisation modification to Java Virtual Machine and therefore the Map cut back framework. it'll have constitutional applications that store the user identity pool for the entire cloud service. that the cloud service won't ought to maintain every user's identity for every application. Additionally to the on top of methodologies, cloud service can support third party authentication. The third party are going to be sure by each the cloud service and accessing user. Third party authentication can add an extra security layer to the cloud service.

Real time access management are going to be an honest security live within the cloud surroundings. Additionally to access management to the cloud surroundings, operational management inside a information within the cloud will be wont to stop configuration drift and unauthorized application changes. Multiple factors like science address, time of the day, and authentication technique will be employed in a versatile thanks to use on top of measures. as an example, access will be restricted to specific middle tier, making a sure path to the information. Keeping a security administrator break free the information administrator are going to be an honest plan. The label security technique are going to be enforced to guard sensitive knowledge by distribution knowledge label or classifying knowledge.

Data will be classified as public, confidential and sensitive. If the user label matches with the label of the information, then access is provided to the user. Examination of diverse knowledge breaches has shown that auditing may have helped in early detection of issues and avoids them. Auditing of events and chase of logs happening within the cloud surroundings can alter potential attack. Fine grain auditing rather like Oracle 9i allows conditional auditing on the precise application column.

## IV. CONCLUSION

Cloud atmosphere is wide utilized in trade and analysis sides; so security is a vital aspect for organizations running on these cloud environments. victimization planned approaches, cloud environments will be secured for advanced business operations.

## REFERENCES

[1] Ren, Yulong, and Wen Tang. "A SERVICE INTEGRITY ASSURANCE FRAMEWORK FOR CLOUDCOMPUTING BASED ON MAPREDUCE."Proceedings of IEEE CCIS2012. Hangzhou: 2012, pp 240 –244, Oct. 30 2012-Nov. 1 2012
[2] N, Gonzalez, Miers C, Redigolo F, Carvalho T, Simplicio M, de Sousa G.T, and Pourzandi M. "A Quantitative Analysis of Current Security Concerns and Solutions for Cloud Computing.". Athens: 2011., pp 231 – 238, Nov. 29 2011- Dec. 1 2011
[3] [3] Hao, Chen, and Ying Qiao. "Research of Cloud Computing based on the Hadoop platform.". Chengdu, China: 2011, pp. 181 – 184, 21-23 Oct 2011.
[4] Y, Amanatullah, Ipung H.P., Juliandri A, and Lim C. "Toward cloud computing reference architecture: Cloud service management perspective.". Jakarta: 2013, pp. 1-4, 13-14 Jun. 2013.
[5] A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices.". Noida: 2013, pp. 404 – 409, 8-10 Aug. 2013.
[6] Lu, Huang, Ting-tin Hu, and Hai-shan Chen. "Research on Hadoop Cloud Computing Model and its Applications.". Hangzhou, China: 2012, pp. 59 – 63, 21-24 Oct. 2012.
[7] Wie, Jiang , Ravi V.T, and Agrawal G. "A Map-Reduce System with an Alternate API for Multi-core Environments.". Melbourne, VIC: 2010, pp. 84-93, 17-20 May. 2010.
[8] K, Chitharanjan, and Kala Karun A. "A review on hadoop — HDFS infrastructure extensions.".JeJu Island: 2013, pp. 132-137, 11-12 Apr. 2013.
[9] F.C.P, Muhtaroglu, Demir S, Obali M, and Girgin C. "Business model canvas perspective on big data applications." Big Data, 2013 IEEE International Conference, Silicon Valley, CA, Oct 6-9, 2013, pp. 32 - 37.
[10] Zhao, Yaxiong , and Jie Wu. "Dache: A data aware caching for big-data applications using the MapReduce framework." INFOCOM, 2013 Proceedings IEEE, Turin, Apr 14-19, 2013, pp. 35 - 39.
[11] Xu-bin, LI , JIANG Wen-rui, JIANG Yi, ZOU Quan "Hadoop Applications in Bioinformatics." OpenCirrus Summit (OCS), 2012 Seventh, Beijing, Jun 19-20, 2012, pp. 48 - 52.
[12] Bertino, Elisa, SilvanaCastano, Elena Ferrari, and Marco Mesiti. "Specifying and enforcing access control policies for XML document sources." pp 139-151.
[13] E, Bertino, Carminati B, Ferrari E, Gupta A , and Thuraisingham B. "Selective and Authentic Third-Party Distribution of XML Documents."2004, pp. 1263 - 1278.
[14] Kilzer, Ann, Emmett Witchel, Indrajit Roy, VitalyShmatikov, and Srinath T.V. Setty. "Airavat: Security and Privacy for MapReduce."

[15] "Securing Big Data: Security Recommendations for Hadoop and NoSQL Environments."Securosisblog, version 1.0 (2012)
[16] P.R , Anisha, Kishor Kumar Reddy C, Srinivasulu Reddy K, and Surender Reddy S. "Third Party Data Protection Applied To Cloud and Xacml Implementation in the Hadoop Environment With .