Security Issues Associated with Big Data in Cloud Computing

K.Iswarya

Assistant Professor, Department of Computer Science
Idhaya College for Women, Kumbakonam, India

ABSTRACT

In this paper, we discuss security issues for cloud computing, Big data, Map Reduce and Hadoop environment. The main focus is on security issues in cloud computing that are associated with big data. Big data applications are a great benefit to organizations, business, companies and many large scale and small scale industries. We also discuss various possible solutions for the issues in cloud computing security and Hadoop. Cloud computing security is developing at a rapid pace which includes computer security, network security, information security, and data privacy. Cloud computing plays a very vital role in protecting data, applications and the related infrastructure with the help of policies, technologies, controls, and big data tools. Moreover, cloud computing, big data and its applications, advantages are likely to represent the most promising new frontiers in science.

KEYWORDS

I. *CLOUD COMPUTING, BIG DATA, HADOOP,MAP REDUCE, HDFS (HADOOP DISTRIBUTED FILE SYSTEM)*

II. INTRODUCTION

In order to analyze complex data and to identify patterns it is very important to securely store, manage and share large amounts of complex data.Cloud comes with an explicit security challenge, i.e. the data owner might not have any control of where the data is placed. The reason behind this control issue is that if one wants to get the benefits of cloud computing, he/she must also utilize the allocation of resources and also the scheduling given by the controls. Hence it is required to protect the data in the midst of untrustworthy processes. Since cloud involves extensive complexity, we believe that rather than providing a holistic solution to securing the cloud, it would be ideal to make noteworthy enhancements in securing the cloud that will ultimately provide us with a secure cloud.

Google has introduced MapReduce framework for processing large amounts of data on commodity hardware. Apache's Hadoop distributed file system (HDFS) is evolving as a superior software component for cloud computing combined along with integrated parts such as MapReduce. Hadoop, which is an open-source implementation of Google MapReduce, including a distributed file system, provides to the application programmer the abstraction of the map and the reduce. With Hadoop it is easier for organizations to get a grip on the large volumes of data being generated each day, but at the same time can also create problems related to security, data access, monitoring, high availability and business continuity.

In this paper, we come up with some approaches in providing security. We ought 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. 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.

.

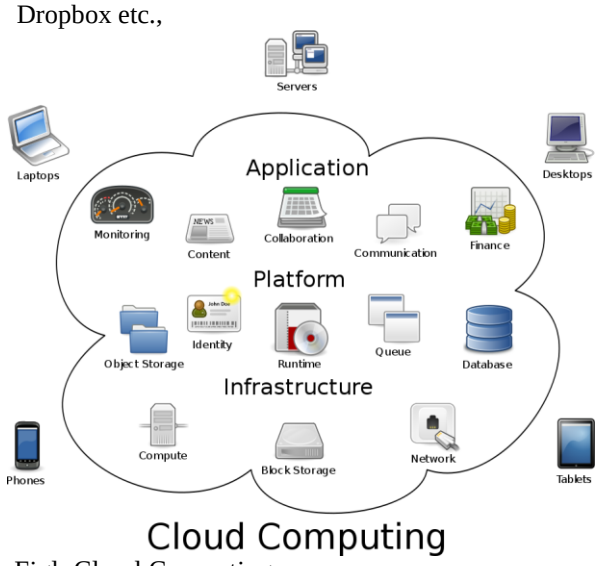
1. Cloud Computing

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. The goal of Cloud Computing is to make use of increasing computing power to execute millions of instructions per second. Cloud Computing uses networks of a large group of servers with specialized connections to distribute data processing among the servers. Instead of installing a software suite for each computer, this technology requires to install a single software in each computer that allows users to log into a Web-based service and which also hosts all the programs required by the user. There's a significant workload shift, in a cloud computing system.

Local computers no longer have to take the entire burden when it comes to running applications. Cloud computing technology is being used to

minimize the usage cost of computing resources The cloud network, consisting of a network of computers, handles the load instead. The cost of software and hardware on the user end decreases.

The only thing that must be done at the user's end is to run the cloud interface software to connect to the cloud. Cloud Computing consists of a front end and back end. The front end includes the user's computer and software required to access the cloud network. Back end consists of various computers, servers and database systems that create the cloud. The user can access applications in the cloud network from anywhere by connecting to the cloud using the Internet. Some of the real time applications which use Cloud Computing are Gmail, Google Calendar, Google Docs and Dropbox etc.,



Figl. Cloud Computing

1.2 Big Data

Big Data is the word used to describe massive volumes of structured and unstructured data that are so large that it is very difficult to process this data using traditional databases and software technologies. The term "Big Data" is believed to be originated from the Web search companies who had to query loosely structured very large distributed data. The three main terms that signify Big Data have the following properties:

a) Volume: Many factors contribute towards increasing Volume - storing transaction data, live streaming data and data collected from sensors etc.,

b) Variety: Today data comes in all types of formats — from traditional databases, text documents, emails, video, audio, transactions etc.,

c) Velocity: This means how fast the data is being produced and how fast the data needs to be processed to meet the demand.

The other two dimensions that need to consider with respect to Big Data are Variability and Complexity .

d) Variability: Along with the Velocity, the data flows can be highly inconsistent with periodic peaks.

e) Complexity: Complexity of the data also needs to be considered when the data is coming from multiple sources. The data must be linked, matched, cleansed and transformed into required formats before actual processing.

Technologies today not only support the collection of large amounts of data, but also help in utilizing such data effectively. Some of the real time examples of Big Data are Credit card transactions made all over the world with respect to a Bank, Walmart customer transactions, and Facebook users generating social interaction data.

Fig2. Big Data

When making an attempt to understand the concept of Big Data, the words such as "Map Reduce" and "Hadoop" cannot be avoided.

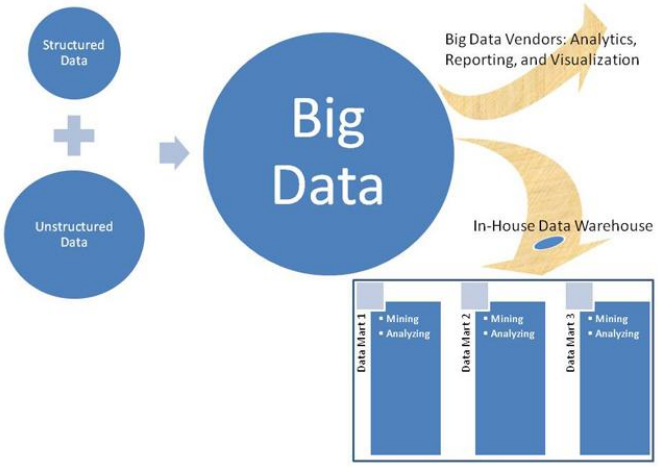


Fig3.Big data

1.3 Hadoop

Hadoop, which is a free, Java-based programming framework supports the processing of large sets of data in a distributed computing environment. It is a part of theApacheproject sponsored by the Apache Software Foundation. Hadoop cluster uses a Master/Slave structure . Using Hadoop, large data sets can be processed across a cluster of servers and applications can be run on systems with thousands of nodes involving thousands ofterabytes. Distributed file system in Hadoop helps in rapid data transfer rates and allows the system to continue its normal operation even in the case of some node failures. This approach lowers the risk of an entire system failure, even in the case of a significant number of node failures. Hadoop enables a computing solution that is scalable, cost effective, 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 projects — Map Reduce and Hadoop Distributed File System (HDFS).

1.4 Map Reduce

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

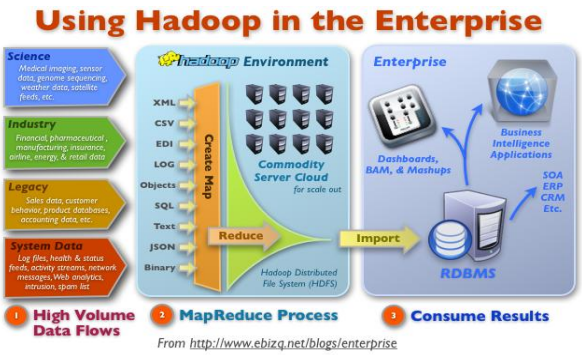


Fig4.Hadoop with Map reduce

1.5 Hadoop Distributed File System (HDFS)

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

1.6 Big data applications

The big data application refers to the large scale distributed applications which usually work with large data sets. Data exploration and analysis turned into a difficult problem in many sectors in the span of big data. With large and complex data, computation becomes difficult to be handled by the traditional data processing applications which triggers the development of big data applications . Google's map reduce framework and apache Hadoop are the defacto software systems for big data applications, in which these applications generates a huge amount of intermediate data. Manufacturing and Bioinformatics are the two major areas of big data applications.

Big data provide an infrastructure for transparency in manufacturing industry, which has the ability to unravel uncertainties such as inconsistent component performance and availability. In these big data applications, a conceptual framework of predictive manufacturing begins with data acquisition where there is a possibility to acquire different types of sensory data such as pressure, vibration, acoustics, voltage, current, and controller data. The combination of sensory data and historical data constructs the big data in manufacturing. This generated big data from the above combination acts as the input into predictive tools and preventive strategies such as prognostics and health management.

Another important application for Hadoop is Bioinformatics which covers the next generation sequencing and other biological domains. Bioinformatics which requires a large scale data analysis, uses Hadoop. Cloud computing gets the parallel distributed computing framework together with computer clusters and web interfaces.

1.7 Big data advantages

In Big data, the software packages provide a rich set of tools and options where an individual could map the entire data landscape across the company, thus allowing the individual to analyze the threats he/she faces internally. This is considered as one of the main advantages as big data keeps the data safe. With this an individual can be able to detect the potentially sensitive information that is not protected in an appropriate manner and makes sure it is stored according to the regulatory requirements.

There are some common characteristics of big data, such as

a) Big data integrates both structured and unstructured data.
b) Addresses speed and scalability, mobility and security, flexibility and stability.
c) In big data the realization time to information is critical to extract value from various data sources, including mobile devices, radio frequency identification, the web and a growing list of automated sensory technologies.

All the organizations and business would benefit from speed, capacity, and scalability of cloud storage. Moreover, end users can visualize the data and companies can find new business opportunities.Another notable advantage with big-data is, data analytics, which allow the individual to personalize the content or look and feel of the website in real time so that it suits the each customer entering the website. If big data are combined with predictive analytics, it produces a challenge for many industries. The combination results in the exploration of these four areas:

a) Calculate the risks on large portfolios
b) Detect, prevent, and re-audit financial fraud
c) Improve delinquent collections
d) Execute high value marketing campaigns

1.8 Need of security in big data

For marketing and research, many of the businesses uses big data, but may not have the fundamental assets particularly from a security perspective. If a security breach occurs to big data, it would result in even more serious legal repercussions and reputational damage than at present. In this new era, many companies are using the technology to store and analyze petabytes of data about their company, business and their customers. As a result, information classification becomes even more critical. For making big data secure, techniques such as encryption, logging, honeypot detection must be necessary. In many organizations, the deployment of big data for fraud detection is very attractive and useful.

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 analyzing multiple data sources.

Not only security but also data privacy challenges existing industries and federal organizations. With the increase in the use of big data in business, many companies are wrestling with privacy issues. Data privacy is a liability, thus companies must be on privacy defensive. But unlike security, privacy should be considered as an asset, therefore it becomes a selling point for both customers and other stakeholders. There should be a balance between data privacy and national security.

2. Motivation and Related Work

2.1. Motivation

Along with the increasing popularity of the Could Computing environments, the security issues introduced through adaptation of this technology are also increasing. Though Cloud Computing offers many benefits, it is vulnerable to attacks. Attackers are consistently trying to find loop holesto attack the cloud computing environment. The traditional security mechanisms which are used are reconsidered because of these cloud computing deployments. Ability to visualize, control and inspect the network links and ports is required to ensure security. Hence there is a need to invest in understanding the challenges, loop holes and components prone to attacks with respect to cloud computing, and come up with a platform and infrastructure which is less vulnerable to attacks.

2.2. Related Work

Hadoop (a cloud computing framework), a Java based distributed system, is a new framework in the market. Since Hadoop is new and still being developed to add more features, there are many security issues which need to be addressed. Researchers have identified some of the issues and started working on this. Some of the notable outcomes, which is related to our domain and helped us to explore, are presented below.

The World Wide Web consortium has identified the importance of SPARQL which can be used in diverse data sources. Later on, the idea of secured query was proposed in order to increase privacy in privacy/utility tradeoff. Here, Jelena, of the USC Information Science Institute, has explained that the queries can be processed according to the policy of the provider, rather than all query processing. Bertino et al published a paper on access control for XML Document. In the paper, cryptography and digital signature technique are explained, and techniques of access control to XML data document is stressed for secured environment. Later on, he published another paper on authentic third party XML document distribution which imposed another trusted layer of security to the paradigm.

Kevin Hamlen and et al proposed that data can be stored in a database encrypted rather than plain text. The advantage of storing data encrypted is that even though intruder can get into the database, he or she can't get the actual data. But, the disadvantage is that encryption requires a lot of overhead. Instead of processing the plain text, most of the operation will take place in cryptographic form. Hence the approach of processing in cryptographic form added extra to security layer.

IBM researchers also explained that the query processing should take place in a secured environment. Then, the use of Kerberos has been highly effective. Kerberos is nothing but a system of authentication that has been developed at MIT. Kerberos uses an encryption technology along with a trusted third party, an arbitrator, to be able to perform a secure authentication on an open network. To be more specific, Kerberos uses cryptographic tickets to avoid transmitting plain text passwords over the wire. Kerberos is based upon Needham-Schroeder protocol.

Airavat has shown us some significant advancement security in the Map Reduce environment. In the paper, Roy and et al have used the access control mechanism along with differential privacy. They have worked upon mathematical bound potential privacy violation which prevents information leak beyond data provider's policy. The above works have influenced us, and we are analyzing various approaches to make the cloud environment more secure for data transfer and computation.

3. ISSUES AND CHALLENGES

Cloud computing comes with numerous security issues because it encompasses many technologies including networks, databases, operating systems, virtualization, resource scheduling, transaction management, load balancing, concurrency control and memory management. Hence, security issues of these systems and technologies are applicable to cloud computing. For example,it is very important for the network which interconnects the systems in a cloud to be secure. Also, virtualization paradigm in cloud computing results in several security concerns. For example, mapping of the virtual machines to the physical machines has to be performed very securely.

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 data explosion is going to make life difficult in many industries, and the companies will gain considerable advantage which is capable to adapt well and gain the ability to analyze such data explosions over those other companies. Finally, data mining techniques can be used in the malware detection in clouds.

The challenges of security in cloud computing environments can be categorized into network level, use authentication level, data level, and generic issues.

Network level: The challenges that can be categorized under a network level deal with network protocols and network security, such as distributed nodes, distributed data, Internode communication.

Authentication level: The challenges that can be categorized under user authentication level deals with encryption/decryption techniques, authentication methods such as administrative rights for nodes, authentication of applications and nodes, and logging.

Data level :The challenges that can be categorized under data level deals with data integrity and availability such as data protection and distributed data.

Generic types: The challenges that can be categorized under general level are traditional security tools, and use of different technologies

3.1 Distributed Nodes

Distributed nodes are an architectural issue. The computation is done in any set of nodes. Basically, data is processed in those nodes which have the necessary resources. Since it can happen anywhere across the clusters, it is very difficult to find the exact location of computation. Because of this it is very difficult to ensure the security of the place where computation is done.

3.2 Distributed Data

In order to alleviate parallel computation,a large data set can be stored in many pieces across many machines. Also, redundant copies of data are made to ensure data reliability. In case a particular chunk is corrupted, the data can be retrieved from its copies. In the cloud environment, it is extremely difficult to find exactly where pieces of a file are stored. Also,

these pieces of data are copied to another node/machines based on availability and maintenance operations. In traditional centralized data security system, critical data is wrapped around various security tools.This cannot be applied to cloud environments since all related data are not presented in one place and it changes.

3.3 Internode Communication

Much Hadoop distributions use RPC over TCP/IP for user data/operational data transfer between nodes. This happens over a network, distributed around globe consisting of wireless and wired networks. Therefore, anyone can tap and modify the inter node communication for breaking into systems.

3.4 Data Protection

Many cloud environments like Hadoop store the data as it is without encryption to improve efficiency. If a hacker can access a set of machines, there is no way to stop him to steal the critical data present in those machines.

3.5 Administrative Rights for Nodes

A node has administrative rights and can access any data. This uncontrolled access to any data is very dangerous as a malicious node can steal or manipulate critical user data.

3.6 Authentication of Applications and Nodes

Nodes can join clusters to increase the parallel operations. In case of no authentication, third part nodes can join clusters to steal user data or disrupt the operations of the cluster.

3.7 Logging

In the absence of logging in a cloud environment, no activity is recorded which modify or delete user data. No information is stored like which nodes have joined cluster, which Map Reduce jobs have run, what changes are made because of these jobs. In the absence of these logs, it is very difficult to find if someone has breached the cluster if any, malicious altering of data is done which needs to be reverted. Also, in the absence of logs, internal users can do malicious data manipulations without getting caught.

3.8 Traditional Security Tools

Traditional security tools are designed for traditional systems where scalability is not huge as cloud environment. Because of this, traditional security tools which are developed over years cannot be directly applied to this distributed form of cloud computing and these tools do not scale as well as the cloud scales.

3.9 Use of Different Technologies

Cloud consists of various technologies which has many interacting complex components. Components include database, computing power, network, and many other stuff. Because of the wide use of technologies, a small security weakness in one component can bring down the whole system. Because of this diversity, maintaining security in the cloud is very challenging.

4. THE PROPOSED APPROACHES

We present various security measures which would improve the security of cloud computing environment. Since the cloud environment is a mixture of many different technologies, we propose various solutions which collectively will make the environment secure. The proposed solutions encourage the use of multiple technologies/ tools to mitigate the security problem specified in previous sections. Security recommendations are designed such that they do not decrease the efficiency and scaling of cloud systems.

Following security measures should be taken to ensure the security in a cloud environment.

4.1 File Encryption

Since the data is present in the machines in a cluster, a hacker can steal all the critical information. Therefore, all the data stored should be encrypted. Different encryption keys should be used on different machines and the key information should be stored centrally behind strong firewalls. This way, even if a hacker is able to get the data, he cannot extract meaningful information from it and misuse it. User data will be stored securely in an encrypted manner.

4.2 Network Encryption

All the network communication should be encrypted as per industry standards. The RPC procedure calls which take place should happen over SSL so that even if a hacker can tap into network communication packets, he cannot extract useful information or manipulate packets.

4.3 Logging

All the map reduce jobs which modify the data should be logged. Also, the information of users, which are responsible for those jobs should be logged. These logs should be audited regularly to find if any, malicious operations are performed or any malicious user is manipulating the data in the nodes.

4.4 Software Format and Node Maintenance

Nodes which run the software should be formatted regularly to eliminate any virus present. All the application softwares and Hadoop software should be updated to make the system more secure.

4.5 Nodes Authentication

Whenever a node joins a cluster, it should be authenticated. In case of a malicious node, it should not be allowed to join the cluster. Authentication techniques like Kerberos can be used to validate the authorized nodes from malicious ones.

4.6 Rigorous System Testing of Map Reduce Jobs

After a developer writes a map reduce job, it should be thoroughly tested in a distributed environment instead of a single machine to ensure the robustness and stability of the job.

4.7 Honeypot Nodes

Honey pot nodes should be present in the cluster, which appear like a regular node but is a trap. These honeypots trap the hackers and necessary actions would be taken to eliminate hackers.

4.8 Layered Framework for Assuring Cloud

A layered framework for assuring cloud computing as shown in Figure (1)=9 consists of the secure virtual machine layer, secure cloud storage layer, secure cloud data layer, and the secure virtual network monitor layer. Cross cutting services are rendered by the policy layer, the cloud monitoring layer, the reliability layer and the risk analysis layer.

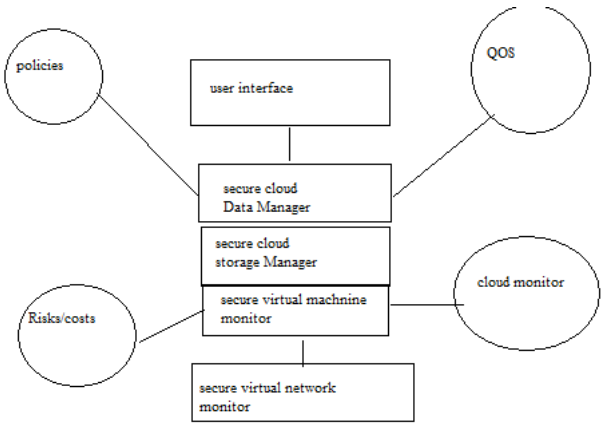


Fig 5.Layered framework for assuring cloud

4.9 Third Party Secure Data Publication to Cloud

Cloud computing helps in storing of data at a remote site inorder to maximize resource utilization. Therefore, it is very important for this data to be protected and access should be given only to authorized individuals. Hence this fundamentally amounts to secure third party publication of data that is required for data outsourcing, as well as for external publications. In the cloud environment, the machine serves the role of a third party publisher, which stores the sensitive data in the cloud. This data needs to be protected, and the above discussed techniques have to be applied to ensure the maintenance of authenticity and completeness.

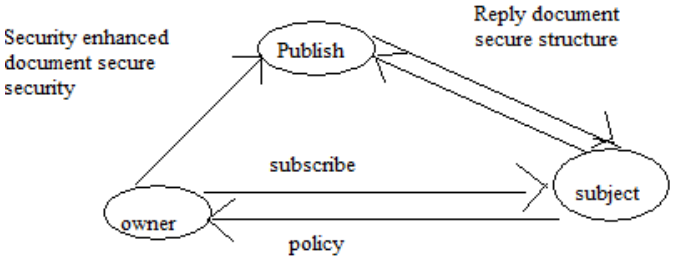


Fig6.Third party secure data publication applied to cloud

4.10 Access Control

Integration of mandatory access control and differential privacy in distributed environment will be a good security measure. Data providers will control the security policy of their sensitive data. They will also control the mathematical bound on privacy violation that could take place. In the above approach, users can perform data computation without any leakage of data. To prevent information leak, SELinux will be used. SELinux is nothing but Security-Enhanced Linux, which is a feature that provides the mechanism for supporting access control security policy through the use of Linux Security Modules (LSM) in the Linux Kernel.

Enforcement of differential privacy will be done using modification to Java Virtual Machine and the Map Reduce framework. It will have inbuilt applications which store the user identity pool for the whole cloud service. So the cloud service will not have to maintain each user's identity for each application. In addition to the above methodologies, cloud service will support third party authentication. The third party will be trusted by both the cloud service and accessing user. Third party authentication will add an additional security layer to the cloud service.

Real time access control will be a good security measure in the cloud environment. In addition to access control to the cloud environment,

operational control within a database in the cloud can be used to prevent configuration drift and unauthorized application changes. Multiple factors such as **IP** address, time of the day, and authentication method can be used in a flexible way to employ above measures. For example, access can be restricted to specific middle tier, creating a trusted path to the data. Keeping a security administrator separate from the database administrator will be a good idea. The label security method will be implemented to protect sensitive data by assigning data label or classifying data.

Data can be classified as public, confidential and sensitive. If the user label matches with the label of the data, then access is provided to the user. Examination of numerous data breaches has shown that auditing could have helped in early detection of problems and avoids them. Auditing of events and tracking of logs taking place in the cloud environment will enable possible attack. Fine grain auditing just like Oracle 9i enables conditional auditing on the specific application column.

5. Conclusion

Cloud environment is widely used in industry and research aspects; therefore security is an important aspect for organizations running on these cloud environments. Using proposed approaches, cloud environments can be secured for complex business operations

6. REFERENCES

[1] Ren, Yulong, and Wen Tang. "A Service Integrity Assurance Framework For Cloud Computing Based On *Mapreduce."Proceedings of IEEE CCIS2012*.. 30 2012-Nov. 1 2012

[2] 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.

[3] A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices.". Noida: 2013, pp. 404 — 409, 8-10 Aug. 2013.

[4] 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.

[5] 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.

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

[7] 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.

[8] Xu-bin, LI , JIANG Wen-rui, JIANG Yi, ZOU Quan "Hadoop Applications in Bioinformatics." *Open Cirrus Summit (OCS), 2012 Seventh*, Beijing, Jun 19-20, 2012, pp. 48 - 52.

[9] Bertino, Elisa, Silvana Castano, Elena Ferrari, and Marco Mesiti. "Specifying and enforcing access control policies for XML document sources." pp 139-151.

[10] Kilzer, Ann, Emmett Witchel, Indrajit Roy, Vitaly Shmatikov, and Srinath T.V. Setty. "Airavat: Security and Privacy for MapReduce."

[11] "Security-Enhanced Linux."Security-Enhanced Linux. N.p. Web. 13 Dec 2013.