RESEARCHERID
THOMSON REUTERS
[INIT: January 2017]
ISSN 2348 – 8034
Impact Factor- 4.022
GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

CLOUD COMPUTING AND HADOOP SECURITY ANALYSIS

Sandeep Kumar Mohapatra*[1] & Anamika Upadhyay[2]
[*1]Senior Technical Lead, Aricent Technologies Pvt. Ltd., Bangalore, India
[2] Technical Lead, Aricent Technologies Pvt. Ltd., Bangalore, India

ABSTRACT

This paper carries out the approaches and implementation of analytics on cloud environment for big data application. Despite the popularity on analytics and big data putting them into practice is itself a big challenge as its very complex and time consuming. Cloud computing has been revolutionizing the IT industry by adding flexibility to the way IT is consumed enabling the industry to only pay for resources and services they use. In recent years with the popularization of concept and application of cloud computing and Hadoop, which is an open source cloud computing model is focused more by academia and industrial sectors. Hadoop is an open-source software platform for distributed computing dealing with a parallel processing of large data sets. As analytics thrives on data, and typically large volumes of it, it makes no sense for analytical platforms to exist outside of the cloud leading to inefficient, time consuming migration of this data from source to the analytics clusters. Running Hadoop clusters in the same cloud environment is an obvious solution to this problem. With the growing acceptance of Hadoop, there is an increasing trend to incorporate more enterprise security features. In this paper we will be dealing with running Hadoop on cloud environment along with considering some of the security measures which we will be incorporating to make the system faster as well as making the user data secure.

Keywords- *Big data, Hadoop, Cloud Computing, MapReduce, data security.*

I. INTRODUCTION

Today Hadoop is becoming standard platform for massive data analysis across industries, internet giants like yahoo and Facebook are processing petabytes of data using Hadoop for managing their daily operations. In this area new business is also flourishing, which are providing Hadoop distribution, which is easier to setup and use. Cloud service providers like amazon and Google are provisioning data analytics as service, which help the users to run their Hadoop jobs on managed facilities. The power of Hadoop comes from its two components, a distributed file system called HDFS and programming model called the MapReduce, which supports massive data storage and processing at the user computers [1]. HDFS breaks the files into multiple blocks and replicates that block on multiple data nodes, whereas the MapReduce divides the user jobs into multiple tasks and executes those tasks on multiple worker nodes in parallel depending upon the data locality.

This paper argues that data analytics enabled by MapReduce are particularly synergistic with the utility computing or pay-as-you-go model enabled by the cloud this model is beneficial for many companies (especially for start-ups and medium-sized businesses). An interesting artifact of this utility we can visualize as using 1000 machines for 1 hour, which will be equivalent to using one machine for 1000 hours. Thus, the MapReduce job can potentially improve its performance while incurring same cost by acquiring several machines and executing in parallel. A physical cluster on the other hand has a limitation on number of machines available hence influencing the performance. Further data analytics jobs are initiated periodically as background batch jobs, which can allocate resources on demand and release them after the job is done. We are focusing the most important technical issues on enabling cloud analytics, but also highlight some of important non-technical challenges faced by organizations that want to provide analytics as a service in the cloud [1,2]. The advantages of performing Hadoop jobs on top of a cloud platform one of the major key challenges are facing the end-user is efficiently provisioning such Hadoop jobs, which has the capacity to efficiently use and release resources on the go such that the resources are maximum utilized. Second key objective for performing Hadoop jobs on cloud is optimization, that we believe to be very relevant to end users is to minimize the cost at the same time maximizing the performance by decreasing the execution time.

ISSN 2348 – 8034
Impact Factor- 4.022

Apart from efficiency and performance, one of the major challenges of implementing Hadoop on cloud is data security protection of information from theft, corruption, and natural disaster at the same time allowing the data and information on the cloud remain accessible to the users. Private clouds are somewhat secure as they are deployed within the locales of the organization, which are protected by firewall, but on public clouds, data is not secure and is prone to all sorts of danger, as the data location is known publically. With the growing popularity of cloud, computing, big and small enterprises are taking it as an option to shift their workloads to the cloud. Before completely moving to issues needs to be considered for multi-tenancy, data security, data integration, etc. this if not taken seriously will be threat to the organization. There are issues around multi-tenancy, data security, software license, data integration, etc., that have to be considered before enterprises can deploy their Hadoop framework to the cloud.

II. SYSTEM SECURITY OVERVIEW

Securing data is not a problem for limited number of people but for everyone dealing with sensitive data, which varies from executives and business stakeholders to scientist and academic professionals. A data security strategy must be defined first as a roadmap for integrating Hadoop into the enterprise information system. As we move to the cloud-computing environment, both applications and resources are delivered over the Internet as services. In such an infrastructure where the resources are accessed over the internet, it becomes particularly serious for data security and data privacy because the data is located in different places even in the entire globe. We have to make the data secure in both cloud environment and Hadoop system for helping the enterprise for successfully implementing the HDFS over cloud.

HDFS Security overview

Hadoop is a distributed system which allows storing big data and also supports parallel processing which is helpful to process very large amounts of data in terabytes or petabytes, and provides high throughput access to this data with the help of Hadoop distributed file system. Files are replicated across more than one machine to ensure durability and high availability for parallel applications. Processing personal or sensitive data on distributed environment requires secure computing. Hadoop brings many benefits to enterprise, but it is also making data prone to cyber-attack. Attackers are constantly monitoring the unsafe system to target, and in such cases Hadoop becomes starting point with huge data stored in that [3].

Current version of Hadoop has very basic rudimentary implementation which makes the Hadoop cluster prone to following attacks [5].

1. Hadoop does not authenticate the client before allowing the uses to access HDFS.
2. One can get the datanode access bypassing the namenodes, which can lead to sniffing and eavesdropping of data packets sent from datanodes
3. Anyone can communicate with Datanode without communicating with the namenode and ask for data location.
4. Big data stacks were designed with little or no security in mind. Prevailing big data installations are built on the web services model, with few or no facilities for preventing common web threats making it highly susceptible.

Although auditing and authorization controls (HDFS file permissions and ACLs) were used in earlier distributions, such access control was easily evaded because any user could impersonate any other user. Because impersonation was frequent and done by most users, the security controls measures that did subsist were not very effective. Later authorization and authentication was added, but that to have some weakness in it[12].

All users and programmers had the same level of access privileges to all the data in the cluster, any job could access any of the data in the cluster, and any user could read any data set [4]. Because MapReduce had no concept of authentication or authorization, an impish user could lower the priorities of other Hadoop jobs in order to make his job complete faster or to be executed first – or worse, he could kill the other jobs.

ISSN 2348 – 8034
Impact Factor- 4.022

The Hadoop community supports some security features through the current Kerberos implementation, the use of firewalls, and basic HDFS permissions and ACLs [5]. Kerberos is not a compulsory requirement for a Hadoop cluster, making it possible to run entire clusters without deploying or implementing any security. Kerberos is also not very easy to install and configure on the cluster, and to integrate with Active Directory (AD) and Lightweight Directory Access Protocol, (LDAP) services. [6]

Hadoop security is not properly addressed by firewalls, once a firewall is breached; the cluster is wide-open for attack. Firewalls offer no protection for data at-rest or in-motion within the cluster. Firewalls also offer no protection from security failure which originates from within the firewall perimeter [6]. An attacker who can enter the data center either physically or electronically can steal the data they want, since the data is un-encrypted and there is no authentication enforced for access [6, 10].

Cloud computing security overview

Cloud computing has two basic functions of computing and data storage. In cloud computing environment user does not have to use any physical hardware they can just access their data and finish computing just through internet connectivity, in such scenario clients are even not aware of the location where the data is stored and which machines are executing their computing tasks. Coming to data storage, data protection and data security becomes two important factors for gaining users trust for making cloud successfully used with other techniques like Hadoop.

Figure

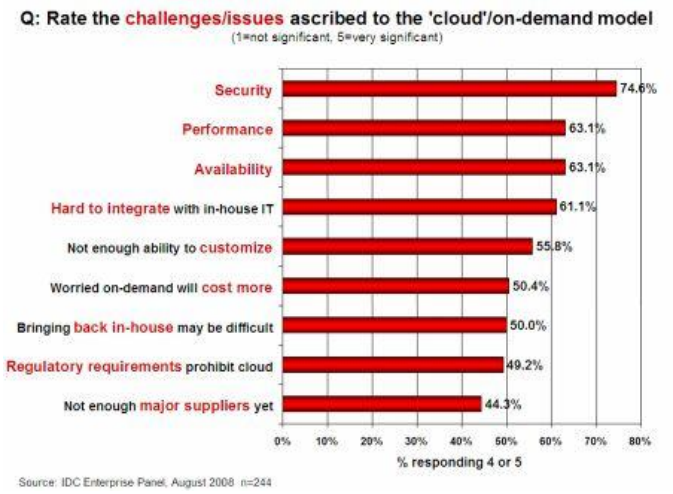

Fig 1- Major Security challenges.

Traditional system security mostly means keeping bad people out. The common threat either comes with compromise with the auth/access control system, or impersonate existing users, which is the major threat.
Cloud computing bring new threats in the form of:

- Data, applications, resources are located with provider
- User identity management is handled by the cloud
- User access control rules, security policies and enforcement are managed by the cloud provider
- Multiple customers are sharing same resource.
- Consumer relies on provider to ensure
    - Data security and privacy
    - Resource availability
    - Monitoring and repairing of services/resources

ISSN 2348 – 8034
Impact Factor- 4.022

III. SYSTEM SECURITY MODELS

The data storage and Data as a Service are the capabilities which are provided by Clouds are important component, but for analytics, it is equally important and relevant to use the data to build models that can be utilized for analysis of data to find meaningful information from the big data like forecasting and prescriptions. Moreover, as models are built based on the available data, those models need to be tested and verified against new data in order to evaluate their ability to forecast future predictions.

The key challenge in the area of Model Building and Scoring is the discovery of techniques that are able to explore the rapid elasticity and large scale of Cloud systems. Given that the amount of data available for Big Data analytics is increasing, timely processing of such data for building and scoring would give a relevant advantage for businesses able to explore such a capability. In the same direction, standards and interfaces for these activities are also required, as they would help to disseminate ‘‘prediction and analytics as services’’ providers that would compete for customers. If the use of such services does not incur vendor lock in customers can choose the service provider only based on cost and performance of services, enabling the emergence of a new competitive market and on top of that security.

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 [4].

IV. ISSUES AND CHALLENGES

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 [5, 6].

The challenges of security in cloud computing environments can be categorized into network level, user 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

ISSN 2348 – 8034
Impact Factor- 4.022

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.
Cloud consists of various technologies which has many interacting complex components. Components include database, computing power, network, and much 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.

V. SECURITY MEASURES FOR IMPLEMENTATION

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. Security recommendations are designed such that they do not decrease the efficiency and scaling of cloud systems.

Figure

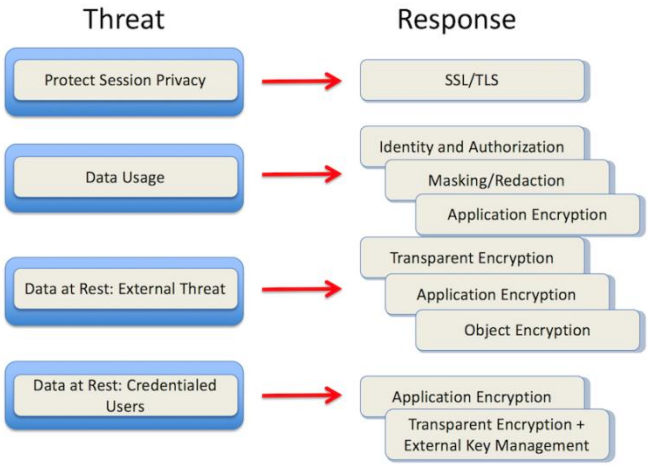

Fig 2 Security measures

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

1. ***Authentication:*** Ensuring the proper authentication of users who access Hadoop. In addition, that authorized Hadoop users can only access the data that they are entitled to access. It should be ensured that data access histories for all users are recorded in accordance with compliance regulations and for other important purposes. To address these kerbose protocol can be integrated with Hadoop, on top of the kerbose protocol we can add simple public key infrastructure to the system so that user can independently authenticate with secure processors and retrieve a secret keys which can be used for encrypting the secure data. Security of HDFS architecture can be further enhanced by using kerbose over SSL which can be used for mutual authentication and access control.
2. ***Encryption of data and file system:*** we must ensure the protection of data both at rest and in transit through advanced encryption algorithm. An application level encryption Mapreduce can be introduced that uses pre-uploaded plain text to the HDFS, to secure the file system. This encrypted HDFS will be little slower as compared generic HDFS, but this will ensure that the data of the enterprise is secured [9].
3. ***Network Encryption*:** Hadoop is a distributed system running on distinct machines, which means that data must be transmitted over the network on a regular basis. Authentication with SSH secure login is the first and most important line of defense in a system of trusted and open networks. 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. ***Node authentication:*** whenever a node whether it's a datanode or a namenode 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.
5. ***Layered Framework for Cloud:*** A layered framework for assuring cloud computing 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.
6. ***Third Party Secure Data Publication to Cloud****:* Cloud computing helps in storing of data at a remote site in order 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.
7. ***Access Control:*** HDFS ACLs give you the ability to specify fine-grained file permissions for specific named users or named groups, not just the file's owner and group. HDFS ACLs are modeled after POSIX ACLs. Best practice is to rely on traditional permission bits to implement most permission requirements, and define a smaller number of ACLs to augment the permission bits with a few exceptional rules. We can prevent outside hackers or malicious insiders from surreptitiously modifying access control lists (ACLs) and accessing sensitive data [7].
8. ***Transport Layer Security:*** if the goal is to protect the session privacy, then TLS is an option, which one can use. Transport encryption protects all communication from access or modification from the attackers. It is an important data security technique, as data moving "across the wire" is often most vulnerable to interception and theft, whether transferring data across clusters in a rack within the datacenter, nodes in the cloud, or especially across a public network over internet

VI. CONCLUSION

In Big Data World, where data is accumulated from various sources, which can be of any format, security is a major concern (critical requirement) as there is no fixed source of data. With the Hadoop gaining larger acceptance within the industry, a natural concern over the security has spread. A growing need to accept and assimilate these security solution and commercial security features has surfaced. In this paper, we have tried to cover all the security solution to secure the Hadoop ecosystem. Since cloud environment is widely used in industry and research aspects, therefore security is an important aspect for organizations, which are running on these cloud environments. Using proposed approaches, cloud environments and Hadoop deployment on cloud can be successfully achieved with proper security for complex business operations.

REFERENCES

1. *R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, I. Brandic, Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility, Future Gener. Comput. Syst. 25 (6) (2009) 599–616.*
2. *M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R.H. Katz, A. Konwinski, G. Lee, D.A. Patterson, A. Rabkin, I. Stoica, M. Zaharia, Above the Clouds: A Berkeley View of Cloud Computing, Technical report UCB/EECS-2009-28, Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, USA (February 2009).*
3. *Apache Hadoop, http://hadoop.apache.org.*
4. *Cloud Security Alliance "Top Ten big Data Security and Privacy Challenges"*
5. *Devaraj Das, Owen O'Malley, Sanjay Radia, and Kan Zhang "Adding Security to Apache Hadoop"*
6. *Zettaset "The Big Data Security Gap: Protecting the Hadoop Cluster"*

7. *Y. Reddy. Access control for sensitive data in hadoop distributed file systems. In INFOCOMP 2013, The Third International Conference on Advanced Communications and Computation , pages 72–78, 2013.*
8. *G. Sadasivam, K. Kumari, and S. Rubika. A novel authentication service for hadoop in cloud environment. In Cloud Computing in Emerging Markets (CCEM), 2012 IEEE International Conference on , pages 1–6. IEEE, 2012.*
9. *S. Park and Y. Lee. Secure hadoop with encrypted hdfs. In Grid and Pervasive Computing , pages 134–141. Springer, 2013.*
10. *Z. Shen and Q. Tong. The security of cloud computing system enabled by trusted computing technology. In Signal Processing Systems (ICSPS), 2010 2nd International Conference on , volume 2, pages V2–11. IEEE, 2010.*
11. *J. Dean, S. Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, Communications of the ACM 51(1).*
12. *Kevin T. Smith "Big Data Security : The Evolution of Hadoop's Security Model"*