ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 3, Issue 2, February 2018

Security Issues in Hadoop Framework: A Review

[1] K. Vishal Reddy, [2] Jayantrao B. Patil, [3] Ratnadeep R.Deshmukh
[1]Deogiri Institute of Engineering and Management Studies, Aurangabad, India. [2] R.C. Patel Institute of Technology, Shirpur, India. [3] Dept. of CS and IT, Dr. B. A. M. University, Aurangabad, India.

Abstract: **In this era of Big Data, organizations collect massive volumes of data in order to derive insights for making decisions. To handle and process big data, new architectures and technologies were evolved. Out of these technologies, Hadoop framework has been adopted by many organizations for storing and processing large complex data. Hadoop framework is an open source technology which has not prioritized security in its initial stages of development. Originally Hadoop has developed to work behind the firewall. But, due to widespread acceptance of Hadoop by many enterprises has provoked most of the cloud distributors for providing Hadoop-as-a-service. This paper firstly describes Hadoop framework, then Hadoop execution on OpenStack and finally focuses on the recent efforts taken by various researchers to provide security in Hadoop framework.**

Keywords: **Big Data, Cloud Computing, Hadoop, OpenStack.**

I. INTRODUCTION

In digital era, the explosion of data is observed from various sources in various formats. This explosion of data in large volumes is from NYSE (New York Stock Exchange), Government sectors like Aadhaar, E-Retails like ebay, Banks and Financial services like JPMorganChase, Social networking sites like facebook, Twitter, LinkedIN, Retails like Sears. All these organizations require large volumes of data to extract the hidden treasure. In order to handle and analyze such massive amounts of data new architectures and technologies are required. Some open source and commercial frameworks are developed to handle and process big data effectively. Hadoop [1], Spark, Strom and S4 are some open source framework technologies. Some of the commercial frameworks are Google's big query, Amazon's EMR (Elastic MapReduce), Microsoft's Windows Azure and so on.

Hadoop is an open source framework used to handle and process large volumes of datasets in a distributed environment. Hadoop fallows batch processing of the data. Spark is a fast, in-memory, distributed, large scale data processing engine. It comes with elegant development API's for data processing, SQL, streaming and machine learning. Strom is an open source engine for processing the real-time data in a distributed environment. S4 is an open source platform for processing and managing the data streams in real-time.

The commercial frameworks mentioned above are discussing the cloud platform which is done on a remote location over the internet. Now-a-days, open source platforms are available in the market which facilitates the types of cloud on local infrastructure. Such platforms are OpenStack, VMWare's cloud, Ubuntu cloud, Eucalyptus cloud, CloudStack, Mirantis and so on.

Along with above mentioned frameworks there are some NoSQL databases which are used to handle large volumes of data. These NoSQL databases are based on Brewer's CAP (Consistency, Availability and Partition Tolerance) theorem. NoSQL [2] databases must satisfy Partition Tolerance and one out of Consistency or Availability. Such NoSQL databases are DynamoDB, Redis, Voldemort, Cassandra, HBase, MongoDB, SimpleDB, CouchDB, BigTable. All the NoSQL technologies are broadly classified into four data stores a) Key-Value b) Document-oriented c) Column-oriented d) Graph based.

All the technologies and architectures discussed above are well suited to handle and process big data in an efficient manner. The biggest concern with these technologies revolves around the security and protection of sensitive information. Our focus in this paper is limited to present security related to Hadoop and OpenStack technologies. We have various security tools available to protect the sensitive data, but all these existing security tools haven't developed by keeping big data technologies in mind. The existing security tools may not be feasible for big data technologies, because these technologies are equipped with rich set of characteristics (reliable, flexible, economical and scalable). As both the technologies are highly distributed and scalable in nature, lacks in proper access control mechanism towards the data by users. It also lacks in maintaining confidentiality and integrity, while the data-in-transit or data-at-rest. Hadoop's popularity has made various vendors for releasing "security-enhanced" distributions of Hadoop and solutions that compliment Hadoop security. This is evidenced by such products as Cloudera Sentry, Zettaset Secure Data Warehouse, IBM InfoSphere Optim Data Masking and the list could go on. At the same time, Apache projects, such as Apache Accumulo provide mechanisms for adding additional security when using Hadoop. Finally, other open source projects, such as Project Rhino (contributed by Intel), Knox Gateway, Falcon and Ranger (contributed by HortonWorks) [3] promise that big changes are coming to Hadoop itself in security perspective.

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 3, Issue 2, February 2018

Our review basically focuses on security issues in big data technologies. This review is organized as fallows. First, we discuss related to big data. Secondly, we discuss Hadoop, security issues in Hadoop and proposed techniques to overcome some security issues. Then we discuss related to cloud computing, OpenStack, and security issues in OpenStack.

II. BIG DATA

The large volume of data explosion has popularized buzz word "Big Data". According to International Data Corporation (IDC) [4], defines big data as: "Big data technologies describe a new generation of technologies and architectures, designed to

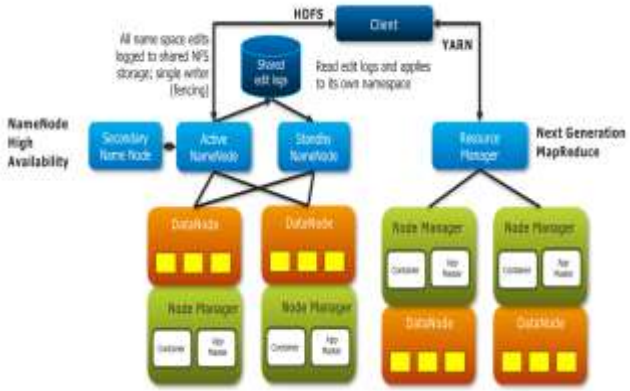


Figure 1: Hadoop Architecture

economically extract value from very large volumes of a wide variety of data, by enabling high velocity acquisition, curation, analyzing and searching". From this definition, we have characterized big data by 4 V's features i.e., volume, velocity, variety, Veracity. So in order to handle large volumes of data, an efficient storage is required for maintaining data for longer period and in sorted manner.

To extract information from large volume of raw data, an efficient processing mechanism is required. Along with storing and processing the technologies used for big data should support various challenges like scalability, availability, data integrity and quality assurance while transformation, privacy and security.

By 2020, it is estimated the digital universe will reach 44 zettabytes of data [5]. The world's information is doubling every two years. By 2020 the world will generate 50 times the amount of information and 75 times the number of information containers.

III. HADOOP

According to White [6], Hadoop is an open source Data Management with scale-out storage and distributed processing. Apache Hadoop is a framework that allows for the distributed processing of large data sets across clusters of community computers using a simple programming model. Hadoop framework as shown in Figure 1 has two major components, HDFS (Hadoop Distributed File Systems) for storage and YARN (Yet Another Resource Negotiator) for processing the data stored in HDFS. HDFS has three daemons i) Active NameNode, ii) Hot Standby NameNode and iii) DataNode. YARN consists of two daemons, i) Node Manager and ii) Resource Manager.

HDFS the base layer of Hadoop framework will create a disk on top of native file system (ext3, etx4, XFS) of Linux flavor. HDFS will divide a file into a number of blocks, by default the size of each block is 128 MB. HDFS provides redundant copies of a block, by default 3 replicas were created. NameNode will store namespaces and metadata of all the files, blocks, attributes and so on. These namespaces and metadata is stored in RAM, as well as at regular time intervals this is stored to an FSImage and edit log file which is placed on the disk to persist the data. The actual data in the form of blocks is stored on the DataNodes. DataNode will provide a heartbeat message to NameNode for every 3 seconds and for every tenth acknowledgement it will send entire status to NameNode. YARN is responsible for scheduling, assigning and monitoring of jobs submitted by clients. Resource Manager will allocate the resources for job execution and these resources are negotiated by application master. Node Manager is responsible to execute the jobs. While executing the jobs it monitors the utilization of resources for each job and reports it to the resource manager. Initially Hadoop has implemented without security in mind. Slowly Hadoop is accepted by many organizations for data analytics as an efficient platform.

IV. CLOUD COMPUTING

According to [10], Cloud Computing means delivery of IT resources as a service to the end users on demand over the network of networks i.e., Internet. The provision of services in a timely, on-demand manner, to allow scaling up and down of resources can occur in cloud. Cloud platforms are based on virtualization, network, storage, web services. These are treated as building blocks of cloud computing. Figure 3 describes different deployment models, service models and essential characteristics of cloud computing.

There are four deployment models:

1) Private cloud: This type of cloud which is solely dedicated to one organization.

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 3, Issue 2, February 2018

2) Community cloud: A group of organizations come together and make a service level agreement (SLA) to share a common cloud.
3) Public cloud: This type of cloud is available to the general public or a large industry group.
4) Hybrid cloud: This type of cloud is a combination of private and public cloud. This type of cloud is used to place sensitive data in private cloud and insensitive data can be placed on the public cloud.

Cloud has classified all the service models into three categories. These service models are designed according to the customer's interest.

1) Software as a Service (SaaS): The end or novice users can use SaaS model for example Salesforce.com
2) Platform as a Service (PaaS): developers can use PaaS for example Google Apps
3) Infrastructure as a Service (IaaS): operators or IT firms can go with IaaS for example OpenStack.

Cloud computing has some essential characteristics:

1) Broad network Access: As cloud provides its resources and services over the internet, one can access his/her data from anywhere at any time.
2) Measured Services: Cloud providers will bill the customer only to the amount of services they consume.
3) On-Demand Self-Service: As a user, you can use the service you require on-demand.
4) Resource Pooling: A collection of resources provided to the end users.

This study is related to the IaaS where in implementation of private cloud using OpenStack technology and observing the security issues.

Figure 3: Anatomy of Cloud Computing

A. OpenStack

In [11], researcher has mentioned OpenStack, is initiated in July 2010 by NASA and RackSpace. OpenStack is a Stack of software tools for building and managing cloud computing platforms for both public and private clouds. It has the facility to control large pool of hardware resources (compute, storage, and networking) throughout a datacentre. An administrator can control and empower their users with required resources through the dashboard. Predominantly OpenStack acts as an Infrastructure as a service (IaaS) platform, it is free and open-source software released under the terms of the Apache License. The project aims for simple implementation, massive scalability, and a rich set of features. OpenStack provides an IaaS solution through a variety of complementary services. OpenStack provides an IaaS solution through a variety of complementary services. Each service shown in Table 1 offers an Application Programming Interface (API) that facilitates this integration. OpenStack is a contribution from a community of organizations and it will not have any professional support for end users.

Table 1. Services in OpenStackService

Service	Project Name	Description
Dashboard	Horizon	Provides a web-based self-service portal to interact with underlying OpenStack services, such as launching an instance, assigning IP addresses and configuring access controls.
Compute	Nova	Manages the lifecycle of compute instances in an OpenStack environment. Responsibilities include spawning, scheduling and decommissioning of virtual machines on demand.
Networking	Neutron	Enables Network-Connectivity- as-a-Service for other OpenStack services, such as OpenStack Compute. Provides an API for users to define networks and the attachments into them. Has a pluggable architecture that supports many popular networking vendors and technologies.
Identity Service	Keystone	Provides an authentication and authorization service for other OpenStack services. Provides a catalog of endpoints for all OpenStack services.
Image Service	Glance	Stores and retrieves virtual machine disk images. OpenStack Compute makes use of this during

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 3, Issue 2, February 2018

		instance provisioning.
Hadoop Service	Sahara	This enables the data intensive applications to process using Hadoop on top of Openstack.

So Red Hat adopted OpenStack and there had tried to offer them with bundles, add-ons and support. Any developer can contribute to OpenStack community by using Python because entire OpenStack is developed in Python

B. Why OpenStack?

Many companies are contributing to OpenStack, under platinum members there are AT&T, Ubuntu, Hewlett Packard, IBM, Intel, Rackspace, Red Hat, SUSE. These are main players in OpenStack who are promoting this technology. There are 24 companies who are funding under gold members some of them are Yahoo, PayPal, Dell, Cisco, Fujitsu, Hitachi, EMC2 and so many. Almost all networking and telecom companies are using only OpenStack because this technology has plugin support whereas other cloud providers don't have this kind of support. Overall there are 500 companies who are working presently and contributing to OpenStack.

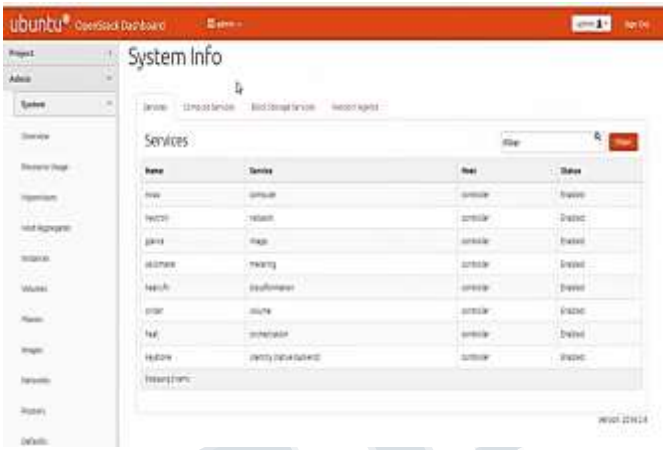
Figure 4 shows the dashboard of OpenStack with various services. As OpenStack is an open source, the entire code is available which can be downloaded and customized according to the usage of customers. The same is not possible with Amazon cloud because it is proprietary in which we can't know the internal operations of this cloud.

VMWare Cloud is proprietary software there will not reveal out implementation details like what sought of programming language used to develop. Technologies used in OpenStack are transparent and a high level of customization is also provided. Basically there is no comparison between OpenStack and Amazon because OpenStack is software but whereas Amazon is a service. As an end user we are never going to know the internals of Amazon but OpenStack is software where we have to deploy it on our own resources. Amazon is a pre-built service which can be used.

V. SECURITY MODEL PROPOSED IN OPENSTACK AND HADOOP

A. Security in Openstack

Keystone is an OpenStack project that provides Identity, Token, Catalog and Policy services for use specifically by projects in the OpenStack family. It implements OpenStack's Identity API. Keystone is organized as a group of internal services exposed on one or many endpoints. Many of these services are used in a combined fashion by the frontend, for example an authentication call will validate



User/Project credentials with the Identity service and upon success create and return a token with the token service. Keystone is based on service catalog, creation of projects, users, roles and mapping of users to projects via roles. In Keystone we have various services as follows:

Identity: The Identity service provides authentication credential validation ad data about Users, Groups.

Resource: The resource service provides data about projects and domains. Like the Identity service, this data may either be managed directly by the service or be pulled from another authoritative backend service, such as LDAP.

Assignment: The assignment service provides data about roles and role assignments to the entities managed by the Identity and Resource services.

Token: The token service validates and manages tokens used for authenticating requests once a user's credentials have already been verified.

Catalog: The catalog service provides an endpoint registry used for endpoint discovery.

Policy: The policy service provider a rule-based authorization engine and the associated rule management interface.

In the OpenStack, vendors should provide a proper security, privacy, integrity, and confidentiality to the sensitive data stored by customers. But some concerns are prevailing due to security service centric architecture of OpenStack. All services get authentication and authorization through keystone service only in OpenStack. The OpenStack Security Project (OSSP) publishes Security Notes to advise users of security related issues. Security notes are similar to advisories; they address vulnerabilities in 3rd party tools typically used within OpenStack deployments and provide guidance on common configuration mistakes that can result in an insecure operating environment.

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 3, Issue 2, February 2018

B. Security issues with Hadoop:

1. There are no Service-level authorization and web proxy capabilities in YARN.
2. Most security tools fail to scale and perform with big data environments.
3. Insufficient Authentication: Do not authenticate user's services.
4. No privacy and no Integrity: Insecure Network Transport and no Message level Security between nodes.
5. Arbitrary code Execution: No user verification for MapReduce code execution, malicious users could submit a job. Client can communicate and submit any MapReduce job to Job Tracker. Job Tracker will not verify the client's authenticity even the data may be owned by anybody else. For example, we have a sales and a financial data stored on Hadoop cluster. Even a sales team developer can write a MapReduce program to extract the financial data from HDFS storage.

C. Security Models Proposed for Hadoop:

In [7], researcher proposed a Network-coding and Multi-node reading technique to safeguard the data. Hadoop stores the data in the form of small blocks distributed across the Hadoop cluster. These blocks consists the data in the format of records. So these records are translated into matrix form by applying the Random Linear Network Coding technique. Firstly a coding coefficient factors are applied on each and every record and then the formed matrix is encoded twice. In this process there will be a reference link between two blocks which is stored at the end of the each block. But this technique will provide security to the data when the data is in transfer state over the network. Along with that it increases the overhead on Name Node.

In [8], author presented provision of security to Hadoop based on two techniques. One is to apply Kerberos technique which will provide only authentication and authorization permissions to the user. It is a security which is provided between a client and the Hadoop cluster. The other one is Bull Eye Approach which will allow the authorized person to secure the data in a right way. This Kerberos depends on three thumb rules which are mentioned as follows:

1. Principle (User)/Service (Name Node, Job Tracker, Data Node and Task Tracker)
2. Authentication
3. Authorization

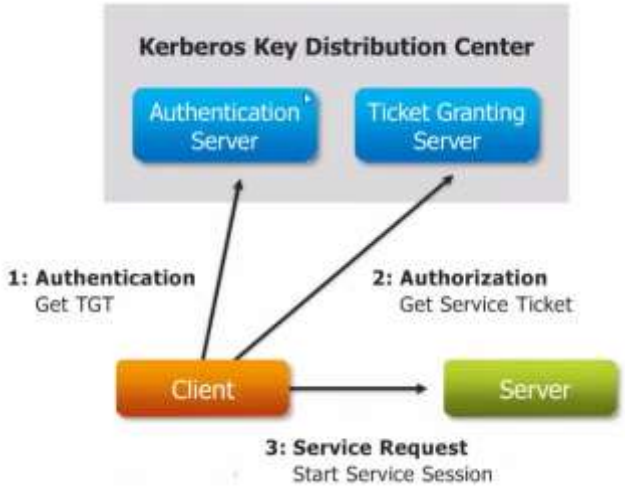


Figure 2: Kerberos workflow

As shown in Figure 2, User from a client machine contacts Key Distribution Center (KDC) where two processes are up and running one Authentication Server (AS) and the other Ticket Granting Server (TGT). AS Authenticates the user and it will compact a TGT, this will look out all authorizations provided for a particular user and provides a TGT Token. This token consist all permissions of a user to various files and folders stored on HDFS. Later, a user can request the Server to initiate a service.

But this Kerberos technique is not a Hadoop concept; it's an operating system concept which is available on Linux and windows operating system. Kerberos is installed at operating system layer and then it is integrated with the Hadoop cluster. By integrating Kerberos in Hadoop provides security to the entire cluster. But the tokens provided to clients over the network can be leaked to malicious attackers.

In [9], proposed a model which will authenticate the clients before accessing any service. In this paper the author has developed an Elliptic curve-based authentication token. A hash function keys are used to secure the tokens sent to users in order to have access to the services.

We have some more tools which will secure Hadoop cluster:

Apache Sentry: Developed by Cloudera which will provide Role-Based Authorization to both data stored on DataNodes as well as to the metadata stored on RAM by NameNode of HDFS components.

Project Rhino: this project was led by Intel for providing authorization framework across Hadoop and its sub-projects (Pig, Hive, Hbase, Sqoop, Oozie and so on…).

Apache Knox: this project is initiated by Hortonworks which will provide an REST API Gateway as single access point for all REST interactions with Hadoop cluster.

As earlier works describes, various models had developed for providing security to the big data which holds the data but not

to the data. In Apache Hadoop, the large volume of data splits into blocks and these blocks are saved across datanodes in a distributed environment. In a large Hadoop cluster data can be leaked if not encrypted properly. The most frequent security concern with the Hadoop framework is inability to integrate the enterprise level security. The security issues such as unencrypted data in-transit or at-rest. Hadoop will not check for data accuracy or provenance for the jobs executed by different clients. The scalability of Hadoop framework is the major concern to have a proper security mechanism.

VI. CONCLUSION

The already proposed works represents different approaches for providing security in Hadoop and cloud. But these are restricted to ACL (Access Control Lists) i.e., authentication and authorization is provided to open source technologies. We can improve the security to data by providing proper cryptographic techniques on the data.

REFERENCES

[1] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, S. U. Khan, " The rise of "big data" on cloud computing: Review and open research issues" Elsevier, pp. 99-115,2014.

[2] H. Hu, Y. Wen, T. Chua and X. li, "Toward scalable systems for big data analytics: A technology tutorial", IEEE Access, vol No. 2, pp. 652-687, 2014.

[3] http://wikibon.com/new-approaches-required-for-comprehensive-hadoop-security (15th Feb, 2018).

[4] J. Gantz and D. Reinsel, "Extracting value from chaos," in Proc. IDC iView, pp. 1-12, 2011.

[5] https://www.emc.com/en-us/big-data/index.htm

[6] T. White, Hadoop: The Definitive Guide, O'Reilly Media, Sebastapol, CA, 2009.

[7] Y. Ma, Y. Zhou, Y. Yu, C. Peng, Z. Wang, and S. Du, "A Novel Approach for Improving Security and Storage Efficiency on HDFS," 6th International Conference on Ambient Systems, Networks and Technologies, ScienceDirect, pp. 631-635, 2015.

[8] B. Saraladevi, N. Pazhaniraja, P. Victer Paul, M. S. Saleem Basha, P. Dhavachelvan, "Big Data and Hadoop- A study in security perspective," 2nd International Symposium on Big Data and Cloud Computing, ScienceDirect, pp. 596-601, 2015.

[9] Y. –S. Jeong, and Y. –T. Kim, "A token-based authentication security scheme for Hadoop distributed file system using elliptic curve cryptography," Springer, 2015.

[10] M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, et al., "A view of cloud computing," Communications of the ACM 53 (4) pp. 50–58, 2010.

[11] P. Miller, "Architecting OpenStack for enterprise reality", GIGAOM Research, 2014.