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 11, November 2016

A Study on Securing Hadoop with Kerberos

Prof. M.M. Siddiqui, Hitesh Rathi, Akansh Jain, Rupal Dahite, Neha Nimbhore

Dept. of Computer, MES College of Engineering, Savitribai Phule Pune University, Pune, India

ABSTRACT: Hadoop, as an open-source cloud computing and big data framework, is increasingly used in the business world, hence data security now becomes one of the main problems. This paper first describes the hadoop project and its security mechanisms, its security problems and risks,how to implement the token authentication based on the Kerberos pre-authentication framework, kerberos pitfalls and their possible solutions, finally based on previous descriptions, concludes Hadoop's security challenges.

I.INTRODUCTION

Apache Hadoop builds an open-source software ecosystem for storage and large-scale processing of unstructured data and evolves to be best big data platform.It consists of Hadoop-common,HDFS,MapReduce,Yarn at its core.It has various components like HBase, Hive, Spark,etc. which are built on top of the core in the ecosystem.The ecosystem now contains more than 20 components and is still increasing day by day.

Hadoop Distributed File System (HDFS) is a distributed, scalable, and portable file system written in Java for the Hadoop framework, and actually it is cloud storage the most widely used tool.Financial organizations using Hadoop started to store their confidential sensitive data on Hadoop clusters. So, a need for a strong authentication and authorization mechanism to protect the sensitive data is observed and also there is a need for a highly secure authentication system to restrict the access to the confidential business data that are processed and stored in an open framework like Hadoop.

Initially,Hadoop services do not authenticate users or other services it considers that the entire cluster, user and the environment were trusted.As a result, Hadoop is subject to some security risks. A user can access data of any other user.This makes it impossible to enforce access control in an uncooperative environment. For example, file permission checking on HDFS can be easily circumvented.

An attacker can masquerade as Hadoop services. For example, user code running on a MapReduce cluster can register itself as a new TaskTracker. The malicious user can read or modify the data in the other's cluster and suspend or kill the other job to execute his job earlier than the other to complete, because the data node does not enforce access control policies.

II.APACHE HADOOP PROJECT

The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing. It uses programming models that permits for the distributed parallel processing of big data sets of large size across clusters of computers so that a cluster of hadoop can easily scale up and also scale down from single servers to many machines in which each of them offer local computation and storage. Many companies like amazon, facebook, yahoo, etc. store and process their data on hadoop.This proves its reliability and robustness.At its core it consists of Hadoop-common as common libraries and facilities like security aspect, HDFS as the distributed storage system, MapReduce as a programming model for large scale data processing and YARN as the cluster-wise resource scheduling and management.The library is designed to detect and handle failures at the application layer,delivers high-available service above the cluster of computers, each of which may be prone to failures.

Enterprises wants to protect sensitive data,while the lack of security mechanism now becomes one of the main problems in hadoop's development and use. Because of the lack of a valid user authentication and data security defense measures, Hadoop is now facing many security problems in the data storage.

III.CURRENT HADOOP SECURITY

Hadoop considers network and its users as trusted, hadoop client uses local username. This is known as Hadoop default. There is no Encryption provided between hadoop cluster and client,all data is stored without any encryption

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 11, November 2016

and managed by central server called namenode.As Hadoop and HDFS have no strong security model,in paticular communication between clients and datanode is not encrypted some mechanisms are added to hadoop.Hence Hadoop is secured with kerberos and it provides mutual authentication and protect against attacks.Every user and service has a kerberos principal and credentials are by Service:Keytab and User: password which RPC Encryption should be enabled.
There are different Layers of security implemented.

1.Apache Knox Gateway:
It provides single point of authentication and access for apache Hadoop services.It uses HTTP for accessing hadoop cluster and provides:

a. Single REST API Access Point
b. Centralized authentication, authorization
c. LDAP/AD Authentication,service authorization and Audit
d.Eliminates SSH edge node risks.
e.Hide Network topology.

2.Authentication:
This process collects credentials from client and identify them.By using this security untrusted users do not gain access to the cluster network and trusted network.for strong authentication Hadoop uses :

a. kerberos
b. LDAP/AD

a. **Kerberos** is a network authentication protocol. It is designed to provide strong authentication for client/server applications.It works on the basis of Tickets.It is a robust authentication system which verifies the identities of principals for users and servers in a distributed system based on symmetric encryption cryptography. Key Distribution Center (KDC) has two main components, Authentication Service (AS) to provide authentication and Ticket Granting Service (TGS) to provide issue ticket. Typically, a user authenticates to AS providing a password and if successful TGS issues a Ticket Granting Ticket (TGT). TGT is stored in cache and used when user communicates with a network service. To do so the client sends the TGT to TGS specifying the targeted service principal and gets issued a Service Ticket. The client then sends the service ticket to the service along with its service request. The service authenticates the client via the service ticket and authorizes the access appropriately.

b. **LDAP** (Lightweight Directory Access Protocol) is a software protocol for enabling anyone to locate organizations, individuals, and other resources such as files and devices in a network, whether on the public Internet or on a corporate intranet. Kerberos can be connected to corporate LDAP environments to centrally provision user information. perimeter authentication can also be provided throug Apache Knox for REST APIs and Web services .

3. Authorization
Authorization is the process of ensuring that users have access only to data as per policies. Hadoop already provides fine-grained authorization via file permissions in HDFS, resource-level access control for MapReduce and YARN, and coarser-grained access control at a service level.
The authorization role is used by providers that make access decisions for the requested resources based on the effective user identity context. This identity context is determined by the authentication provider and the identity assertion provider mapping rules. Evaluation of the identity contexts user and group principals against a set of access policies is done by the authorization provider in order to determine whether access should be granted to the effective user for the requested resource.
Knox Gateway provides an ACL based authorization provider that evaluates rules that comprise of username, groups and ip addresses. These ACLs are bound to and protect resources at the service level. That is, they protect access to the Hadoop services themselves based on user, group and remote ip address. For common authorization framework security administrator is provided with single console to manage all authorization policies for hadoop components.

4. OS Security and Data Protection:
Data protection involves protecting data when stored and transferred , including encryption and masking. Encryption provides an added layer of security by protecting data when it is transferred and when it is stored. Masking capabilities

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 11, November 2016

enable security administrators to desensitize for display or temporary storage. Existing capabilities are used for encrypting data at flight and using new solutions for encryption at rest,data discovery, and data masking by hadoop.

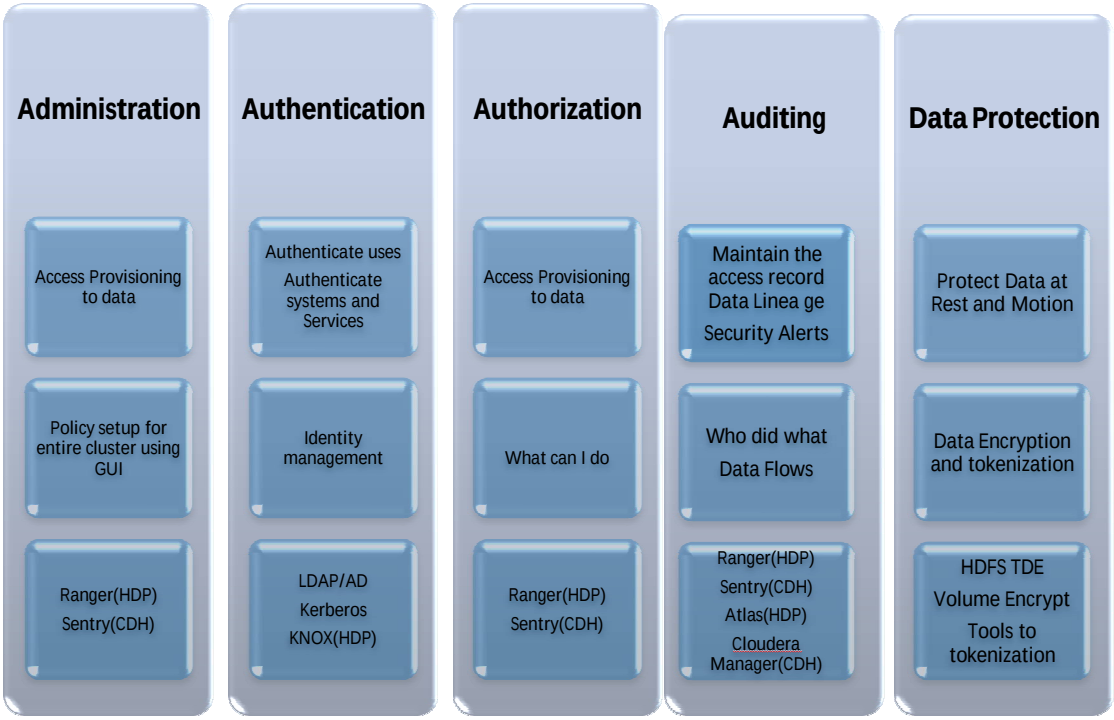

Fig. Security Pillars

IV. KERBEROSE PITFALLS AND THEIR POSSIBLE SOLUTION

Though Kerberos is very suitable solution for securing Hadoop system, it has some disadvantages listed below.This problem can be overcome by implementing various policies to make it more secure.

1. Relying completely on Automated Kerberos Wizards for enabling Kerberos may not configure it properly. Automated Kerberos Wizards are not enough ,hence it requires additional manual setups. Components which has already kerberos enabled are not directly managed by Cloudera. Based on services running system can be palnned and designed. Serivices should be factored for security upgrades as it can be error-prone or time-consuming.

2.Managing Kerberos Principals for users without security policies will lead to vunlerabilities. For all developers new users are created in kerberos,Hence users should be created with security policies.These Policies include-strong password, password expiry(few days), ticket lifetime(few hours),as end user accounts are most prone to the attacks

3.Data migration between clusters is very comman task,but kerberos donot provide any stratergy for secure clusters data migration. ThereFore strategies are required for moving data from secure to non-secure cluster and vice versa.This may involve tweaking Kerberos rules and may have a performance impact

4.Generating long duration Kerberos tickets without frequent renewal for your application accounts makes cluster vulnerable for hackers to analyse the cluster and a possible attacks. If Any code uses application accounts to authenticate to Kerberos and generate tickets, generated ticket's default validity is 24 hours but in reality, code doesn't run for 24 hours. Hence set Kerberos ticket lifetime to minimum and renew it when necessary

5.Using default "max_retries" and "kdc_timeout" Kerberos configuration leads to increase in downtime. Kerberos has Default max_tries=3 and kdc_timeout=30 secs , if Kdc is down, Hadoop services would try to connect to KDC server

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 11, November 2016

for 3 times with a time lag of 30secs. This Making the downtime as 90 seconds before switching to slave KDC, resulting in breaking most of the jobs. This problem can be avoided using Recommended settings as max_tries=1 or 0 and kdc_timeout=as minimum as possible.

6.Hadoop cluster has codes deployed for running specific business logic. Deployed code need to get Kerberos ticket and authenticate, hence deployed code should be complaint to Kerberos. Security upgrades should have plans for such code development and deployment.Every ecosystem has different ways of using Kerberos hence it need proper testing and planning.

7.Having only one Kerberos Key Distribution Center creates single point of failure.It affects the High Availability. None of the services like HDFS, M-R would work if KDC is down therefore it is recommended to have Master-Slave KDC to support KDC failure.

8.All users of a clusters must be provisioned on all servers in the cluster because, not provisioning all users of your cluster on all your cluster nodes will give malicious user access to system. Hadoop lets you submit and execute arbitrary code and every individual task on the cluster use the username and UID of user ,as Hadoop assumes that administration would restrict user if required. User management can be done through /etc/passwd or Open LDAP or Active Directory

9.As Kerberos DB files are encrypted files containing username & password,Backing up Kerberos Database files should not be a part of standard backup process. If it is done as standard process ,it will create multiple orders, making system vulnerable by creating multiple access points for hackers.

10.Kerberos has no stratergy for identifying potential authentication breaches. Kerberos logs contain useful information. For eg.Failed login attempt and tickets reuse are stored in logs. Hence it needs to integrate Kerberos logs centralized security and monitoring tools and it provides a way to actively monitor hacking attempts.

V. LOGGING AND AUDITING

Systems and network device reporting is important to the overall health and security of Hadoop systems. Every cluster,client-host or Applications provide logging features. Log provides a clear view of who owns the process, what action was initiated, when it was initiated, where the action occurred and why the process ran? Thus, it is utmost important that most information should be logged in log files.
Log is a record of actions and events that takes place on a computer system. Logs are the primary record keepers of system and network activity. When security controls experience failures, logs would be particularly helpful.
Auditing is the formal examination and review of actions taken by hadoop users. Event auditing allows the reliable, fine-grained, and configurable logging of a variety of security-relevant system events, including logins, configuration changes and file & service access. These log records can be invaluable for live system monitoring, intrusion detection, and postmortem analysis.
Logging and Auditig helps administrator to detect malicious activities across hadoop cluster.
Any Suspected actions are ten reported to adminstrator, this makes early detection of attack easier.

VI. CONCLUSION

In this paper, we reviewed security of Apache Hadoop platform including its present security situation, threats and some methods enhancing its security level.
Some challenges in designing security mechanism for Hadoop and improving its security seems to be the following factors:
i. Scale of the system is large.
ii. Hadoop is a distributed file system, so file is partitioned and distributed through the cluster.
iii. Next job execution may be done on a different node from which the user has been authenticated and the job has been submitted.
iv. Tasks from different users may be executed on a single node.

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 11, November 2016

v. Users can access the system through some workflow system.

Altogether, identifying Data type which is being stored in cluster play vital role in securing hadoop cluster. how users will access the data, for example through a middleware application or directly, What are the controls placed in the middleware and whether these controls are sufficient are considered and then deciding about choosing one or some or all of security approaches described above. Enabling Kerberos and putting a firewall around Hadoop cluster can be useful if controls aren't sufficient. Communication between Client and hadoop server needs to be secured. Then having applications accessing Hadoop services over REST, Apache Knox can be very appreciate to put it between the application and the Hadoop cluster. Currently there are authorization controls at various layers in Hadoop, From ACL in MR to HDFS permission, and more access controls improvements are coming. Enabling wire encryption to protect data as it moves in Hadoop or using custom and other solutions for encrypting data at rest (as it sits in HDFS) can be considered. Token based authentication becomes increasingly important and interested for Apache Hadoop to meet data access security requirement for better integration with existing authentication providers. Kerberos avoids deployment overhead and risk as found in other solutions.If Above discussed problems are considered and their solution are implemented a secure system can be designed.

REFERENCES

1. A token authentication solution for hadoop based on kerberos pre-authentication2014 International Conference on Data Science and Advanced Analytics (DSAA)
2. A Survey on Security of Hadoop 2014 4th International Conference on Computer and Knowledge Engineering(ICCKE)
3. MIT Kerberos Consortium, Best Practices for Integrating Kerberos into Your Application, 2008.
4. Token Based Authentication and Single Sign On JIRA. https://issues.apache.org/jira/browse/HADOOP-9392.
5. ApacheTM Hadoop®! Available: http://hadoop.apache.org/
6. Hadoop with Kerberos – Deployment Considerations Global Architecture and Technology Enablement Practice
7. Hadoop Security Design Owen O'Malley, Kan Zhang, Sanjay Radia,Ram Marti, and Christopher Harrell Yahoo!
8. M. Yuan, "Study of Security Mechanism based on Hadoop,"
9. Information Security and Communications Privacy, vol. 6, p. 042, 2012.
10. Securing your Hadoop Infrastructure with Apache Knox. Available: http://hortonworks.com/hadoop- tutorial/securing-hadoop-infrastructure-apache-knox/ 2014.
11. X. Zhang, "Secure Your Hadoop Cluster With Apache Sentry," edition: Cloudera, April 07, 2014.

BIOGRAPHY

Prof. M. M. Siddiqui, Assistant Professor, Dept. of Computer Engg., MES College of Engineering, Pune, India
Hitesh Rathi, B.E Student, Computer Engineering, MES College of Engineering, Pune, India
Akansh Jain, B.E Student, Computer Engineering, MES College of Engineering, Pune, India
Rupal Dahite, B.E Student, Computer Engineering, MES College of Engineering, Pune, India
Neha Nimbhore, B.E Student, Computer Engineering, MES College of Engineering, Pune, India