Hadoop Security Challenges and Its Solution Using KNOX

Sirisha. N¹, Kiran K.V.D², R. Karthik³
¹,²K L University, India
³MLR Institute of Technology, India

ABSTRACT

Big Data is a new technology and architecture. It can work on a very large volume of a variety of data with high-velocity, discovery, and/or analysis. Big Data is about the fast-growing sources of data such as web logs, Sensor networks, Social media, Internet text and documents, Internet pages, Search Index data, scientific research. Big data also formally introduces a complex range of analysis. Big data can evaluate mixed data (structured and unstructured) from multiple sources. As there are some security issues in big data which are no longer solved using the hashing techniques on large amount of data, this paper shows an idea of new approach of designing a Knox’ified Hadoop cluster.

Copyright © 2018 Institute of Advanced Engineering and Science. All rights reserved.

1. INTRODUCTION

An extensive variety of techniques and technologies has been developed to seize data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating and information privacy.

In this paper, the big data issues are more focused in terms of security issues that raised in Hadoop Architecture [1] base layer called Hadoop Distributed File System (HDFS) represented in Figure 1. The new Hadoop security design relies on the use of Knox [2] and Ranger.

Figure 1. Overview of Hadoop Distributed File System
2. SECURITY ROADMAP

The Hadoop [1] supports few of the security features using Kerberos, firewalls, ACLS, LDAP etc. As Hadoop cluster [1] installation, Kerberos installations are very tough enough, providing security to Hadoop is also a major problem in the current situation.

Security Roadmap shows the details of different technologies that are emerged with Hadoop today and are represented in Table 1.

<table>
<thead>
<tr>
<th>Authentication [8]</th>
<th>MD5, Digest, GISSAPI (Kerberos), Delegation tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authorization [8]</td>
<td>Job &amp; Queue ACL (resource level)</td>
</tr>
<tr>
<td>Encryption of data at rest [8]</td>
<td>---</td>
</tr>
<tr>
<td>Audit Trials [9]</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 2 shows the Security in hadoop today with five security pillars Administrator, Authentication, Authorization, Audit, Data protection. The current solutions are Apache Knox, Native Kerberos, Audit, Encryptions are the few solutions currently under work. From these solutions Knox is described in next section.

<table>
<thead>
<tr>
<th>S.NO.</th>
<th>SECURITY PILLARS</th>
<th>CURRENT SOLUTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Administrator</td>
<td>Apache Knox</td>
</tr>
<tr>
<td>2.</td>
<td>Authentication</td>
<td>Apache Knox, Native Kerberos</td>
</tr>
<tr>
<td>3.</td>
<td>Authorization</td>
<td>Apache Knox</td>
</tr>
<tr>
<td>4.</td>
<td>Audit</td>
<td>Apache Knox, Hadoop native audit</td>
</tr>
<tr>
<td>5.</td>
<td>Data Protection</td>
<td>HDFS transparent, HBase encryption, Vendor solutions</td>
</tr>
</tbody>
</table>

3. KNOX

KNOX is developed by HortonWorks. Knox is a REST Representational State Transfer (It is sometimes spelled “ReST”) API gateway for interacting with hadoop services [11]. Apache Knox Gateway is a system that provides a single point of authentication and access for apache Hadoop services in a...
cluster [12]. The aim is to simplify Hadoop security for both users and operators. The gateway runs as a server (or cluster of servers) that provide centralized access to one or more Hadoop clusters. It is designed to obscure hadoop cluster topology from outside world. Plugins for hadoop services includes WebHDFS, Oozie, Hive, Hbase, HCatalog. Knox [13] supports LDAP/Active Directory integration. It audits all Knox-managed gateway traffic. It also provides Service – level authorization to hadoop services. It has an End-to-End wire encryption via SSL. By default Knox offers SSL encryption from the client to the Knox gateway [14]. A SSL setup is also possible between Knox and hadoop services [15].

3.1. Knox-Architechture
This section in the paper shows the architecture of Knox which consists of one or more servers that sit outside the hadoop cluster. It is designed to replace SSH “edge-node” for accessing hadoop. It provides a single port to access Hadoop [16] services with a default port: 8443. It is designed to integrate with Kerberos & LDAP (Lightweight Directory Access Protocol) to handle authentication services [25-27]. A Knoxified Hadoop is shown in Figure 2.

![Figure 3. Extend Hadoop API Reach with KNOX](image)

3.2. Goals of Knox [12]
Knox provides,

- perimeter security[17] for Hadoop REST API’s to make Hadoop security easier to set up and use.i.e.,
  

- Authentication [18] and token verification at the perimeter by enabling authentication integration with enterprise and cloud identity management systems.

- Service level authorization at the perimeter.

- It exposes a single URL hierarchy that aggregates REST APIs of a Hadoop cluster.

- Knox securely extends the reach of Hadoop[19] APIs to anyone on any device.

- Serves as a gateway for Hadoop’s REST API. Different Rest APIs varying levels of authentication, authorization, SSL and SSO capabilities.

- It avoids exposing the cluster port and host names to all users.

  New Apache Knox Features in HDP 2.2:
  
  - Knox can be installed by using Ambari. It can start and stop a configuration.
  
  - It provides a new support for: YARN REST API, HDFS HA, SSL to HADOOP[20] cluster services (WEBHDFS, HBASE, HIVE, OOZIE).
  
  - It has Knox Management REST API.
  
  - Integrates with Apache Ranger for service level Authorization.

3.3. Knox-Rest Hierarchies
Steps to have Apache Knox up and running against a Hadoop cluster:
1. Verify system requirements.
2. Download a virtual machine (VM) with Hadoop.
3. Download apache Knox gateway.
4. Start the virtual machine with Hadoop.
5. Install Knox.
6. Start the LDAP embedded within Knox.
7. Start the Knox gateway.
8. Do Hadoop with Knox.

To get a file in HDFS via KNOX we use:
Curl -I -k -u guest:guest -password -x GET \
'https://localhost:8443/gateway/sandbox/webhdfs/v1/tmp/LICENSE op=OPEN'. When curl command is
used Kerberos [23], LDAP services [24] can be integrated with KNOX.

3.4. Knox Configuration Using Ambari
Go to Ambari, click on Add service and setup Knox by selecting Knox & click on next as shown in
Figure 5.
There will be a centralized master server i.e., Knox Gateway, select it and more gateways can also be selected if required by selecting the drop down list as shown in the Figure 6.

Figure 6. Assigning Knox Gateway

Now, goto customized services where user has to give a Knox Master secret input as shown in Figure 7.

Figure 7. Knox Master secret

In Add services Wizard, select the configure identities where we have to configure Knox by selecting the checkbox Knox as shown in Figure 8.

Figure 8. Knox service wizard
In this configurations window just select proceed anyway to deploy Knox as shown in Figure 9.

![Configuration of different services](image1.png)

**Figure 9. Configuration of different services**

Now, Knox will be deployed once clicking on deploy button as shown in Figure 10.

![Deploying the services](image2.png)

**Figure 10. Deploying the services**

Once deployed now, it takes some time to install all the services and its components on the cluster as shown in Figure 11.

![Knox components](image3.png)

**Figure 11. Knox components**

After deploying all the services are now configured using Ambari. Installation is success after doing all the above said process as shown in Figure 12.
Hadoop Security Challenges And Its Solution Using KNOX (Sirisha. N)

3.5. Restarting services in Knox using ambari

Unless all the services are restarted, desired results cannot be obtained. If it shows an orange cycles near the services menu, then ensure that all the services has to be restarted. To do that click on orange cycles->restart->restart all Affected->confirm restart All. So that, all the services will get restarted. Once Knox is installed, login to the host where Knox is been setup. It shows the $ i.e., [ec2-user@ip-172-31-53-37 ~]$.

After this use the following commands to see the defined properties for the Knox Gateway.

$Cd /etc/knox/conf
$ls --ltr

These commands shows gateway-site.xml, which is represented in Figure.13,14.

Knox topologies

---

Figure 12. Installation success

Figure 13. Knox configurations gatewaysite.xml, admin.xml, default.xml

Figure 14. Knox-gatewaysite.xml
3.6. Validating the Services that are Supported by Knox

Validation can be done using `Cd /etc/knox/conf` for gateway.site.xml and topologies has to be used for validating admin.xml and default.xml. Then execute the command

View users.ldif, and then


By using curl command all the services of Knox has to be validated.
4. CONCLUSION

This paper shows that the security risks such as insufficient authentication, no privacy, no integrity, arbitrary code execution are all the part of Kerberos. So, Knox is introduced in this paper to overcome these security risks. Software’s such as Ambari, Rsingh, Puppet, Chef are the automated software’s for working with 150 nodes or more. 4000 to 6000 name node clusters can be formed using these software’s and 10000 name nodes can be formed using puppet. Installation of ambari is shown in very detail in this paper and working will be shown in future work.

REFERENCES
