Indonesian Journal of Electrical Engineering and Computer Science
Vol. 12, No. 1, October 2018, pp. 107~116
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v12.i1.pp107-116

Hadoop Security Challenges and Its Solution Using KNOX

Sirisha. N[1], Kiran K.V.D[2], R. Karthik[3]
[1,2]K L University, India
[3]MLR Institute of Technology, India

Article Info

Article history:

Received Jun 1, 2018
Revised Jul 10, 2018
Accepted Jul 25, 2018

Keywords:

Kerberos
HDFS
Knox
SSL
LDAP
REST API

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 logics, 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.

Corresponding Author:

Sirisha. N,
K. L. university,
India.
Email: rayam16@gmail.com

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.

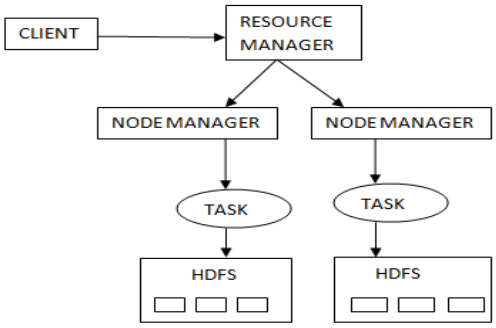


Figure 1. Overview of Hadoop Distributed File System

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 1. Survey of a security and its solution in Big Data

	Map-Reduce [2]	HDFS [2]	Hbase [3]	Hive [4]	Hue [5]	Oozie [5]	Zookeeper [6]	Pig [7]
Authentication [8]	MD5-Digest, GSSAPI (Kerberos), Delegation tokens	SASL framework , Delegation tokens	Kerberos, SASL (secure client authentication)	Apache Knox, LDAP authentication	Kerberos (Pluggable)	Delegation tokens, Kerberos	Kerberos authentication at RPC layer	User level permissions
Authorization [8]	Job & Queue ACL (resource level)	POSIX permissions , ABAC	HBase access control list on tables, columns.	Apache Knox, Lightweight Directory Access Protocol authentication	ACLs and FS permissions	ACLs [10] and FS permissions	ACLs	ACLs, Apache Sentry
Encryption of data at rest [8]	---	Advanced encryption standard, OS level	Arbiter solution	Arbiter solution	Arbiter solution	Arbiter solution	N/A	Arbiter solution
Encryption of data at Transit [8]	RPC – Simple Authentication and Security Layer, HTTPS	RPC – Simple Authentication and Security Layer, Data transfer protocol	Simple Authentication and Security Layer (secure RPC)	Third party solution	HTTPS	Secure Socket Layer/Transport layer security	Arbiter solution	Simple Authentication and Security Layer
Audit Trials [9]	Yes	Yes	No (But Arbiter solution can be used)	Yes	Yes	Yes	Arbiter solution	Arbiter solution

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 2. Security in Hadoop today

S.NO.	SECURITY PILLARS	CURRENT SOLUTIONS
1.	Administrator • Central Management & Consistent security	Apache Knox
2.	Authentication • Authenticate users and systems	Apache Knox, Native Kerberos
3.	Authorization • Provision access to data	Apache Knox
4.	Audit • Maintain a record of data access	Apache Knox, Hadoop native audit
5.	Data Protection • Protect data at rest & in motion	HDFS transparent, HBase encryption, Vendor solutions

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. Bydefault 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 Knox'ified Hadoop is shown in Figure 2.

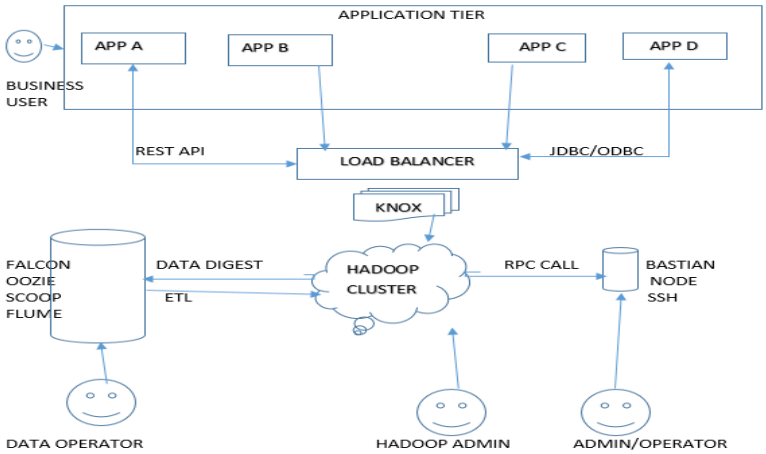


Figure 3. Extend Hadoop API Reach with KNOX

3.2. Goals of Knox [12]

Knox provides,

a) perimeter security[17] for Hadoop REST API's to make Hadoop security easier to set up and use.i.e., curl –I –k –u guest:guest –password –x GET\'https://localhost:8443/gateway/sandbox/webhdfs/vs/tmp/LISCENCE?op=OPEN '.
b) Authentication [18] and token verification at the perimeter by enabling authentication integration with enterprise and cloud identity management systems.
c) Service level authorization at the perimeter.
d) It exposes a single URL hierarchy that aggregates REST APIs of a Hadoop cluster.
e) Knox securely extends the reach of Hadoop[19] APIS to anyone on any device.
f) Serves as a gateway for Hadoop's REST API. Different Rest APIs varying levels of authentication, authorization, SSL and SSO capabilities.
g) It avoids exposing the cluster port and host names to all users.

New Apache Knox Features in HDP 2.2:

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

3.3. Knox-Rest Hierarchies

It provides a single REST hierarchy for all Hadoop services. Normal HADOOP[21] has different HOSTS, different PORTS and exposes the details about the cluster topology viz., http://namenode:50070/webhdfs/,"http://namenode:50070/webhdfs/..,http://hivenode:10001/cliservice"..,"http

://namenode:50070/webhdfs/..,http://hivenode:10001/cliservice"http://hivenode:10001/cliservice, http://localhost:11000/oozie. Whereas Knox has one HOST, one PORT, Consistent Structure viz., https://knox:8443/webhdfs, https://knox:8443/hive, https://knox:8443/oozie. Knox is only effective with proper perimeter security [22] configured. Knox'ified Hadoop cluster is in Figure 3.

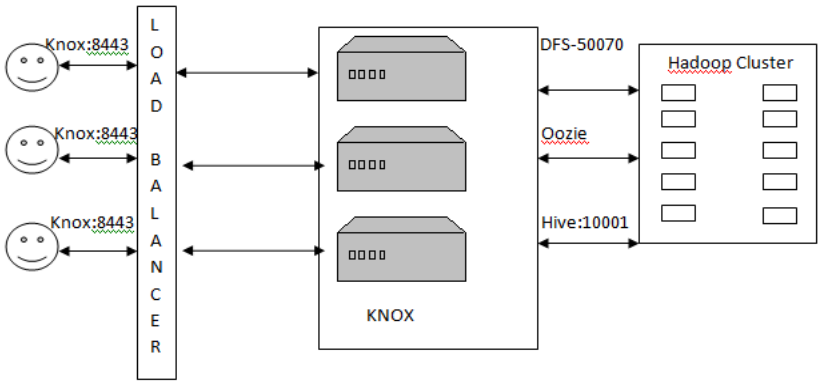


Figure 4. Knox Architecture showing REST hierarchies

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.

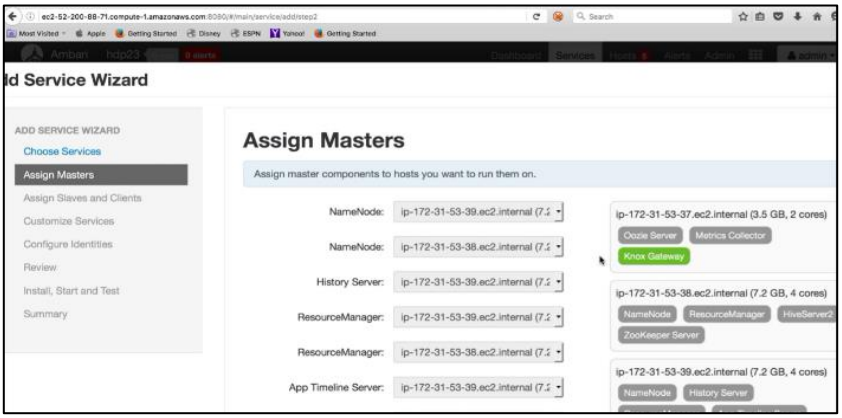


Figure 5. Starting Knox Gateway

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.

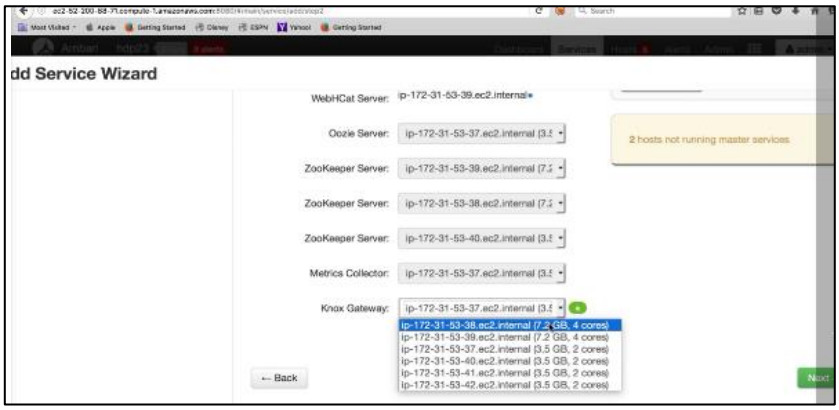


Figure 6. Assigning Knox Gateway

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

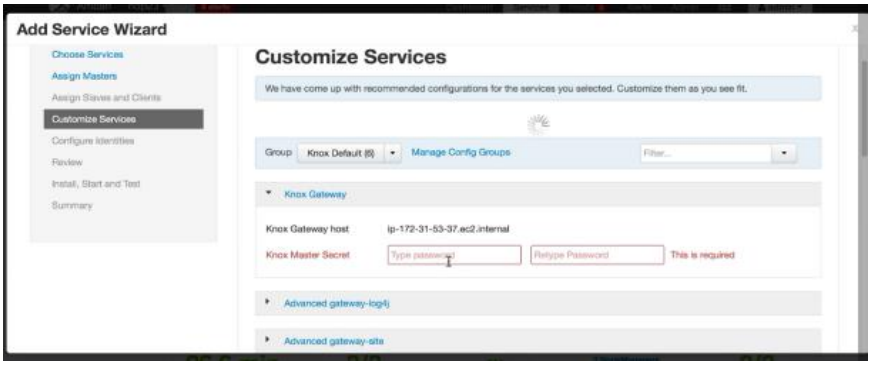


Figure 7. Knox Master secret

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

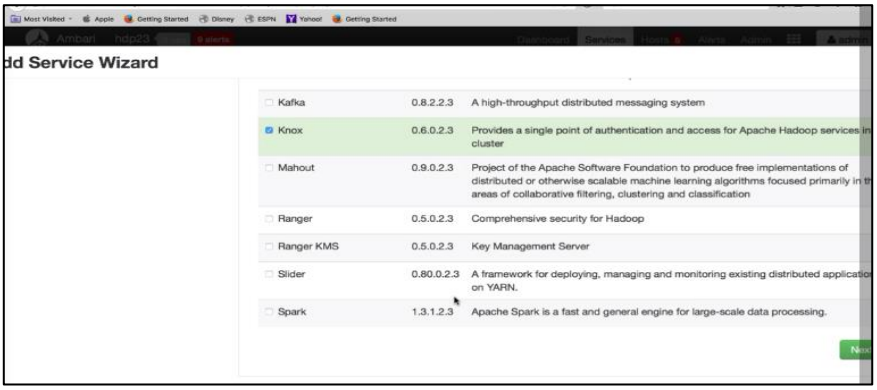


Figure 8. Knox service wizard

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

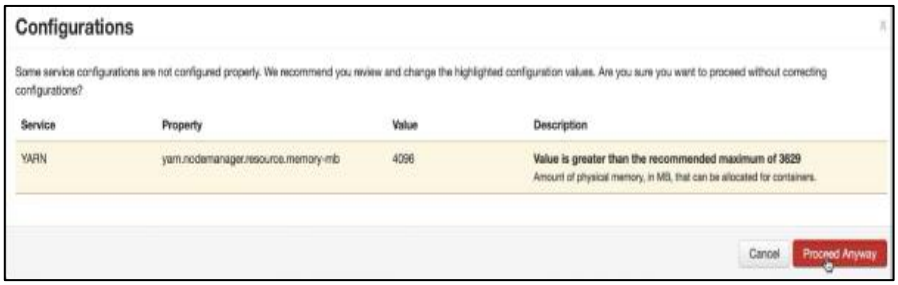


Figure 9. Configuration of different services

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

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.

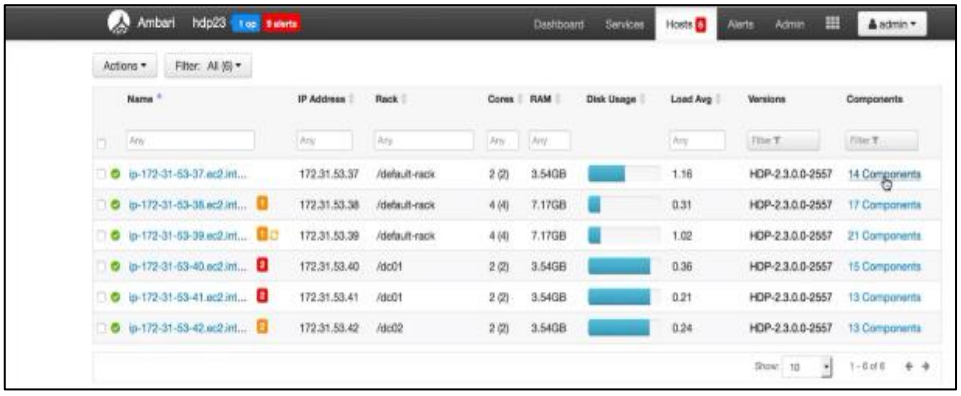


Figure 11. Knox components

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.

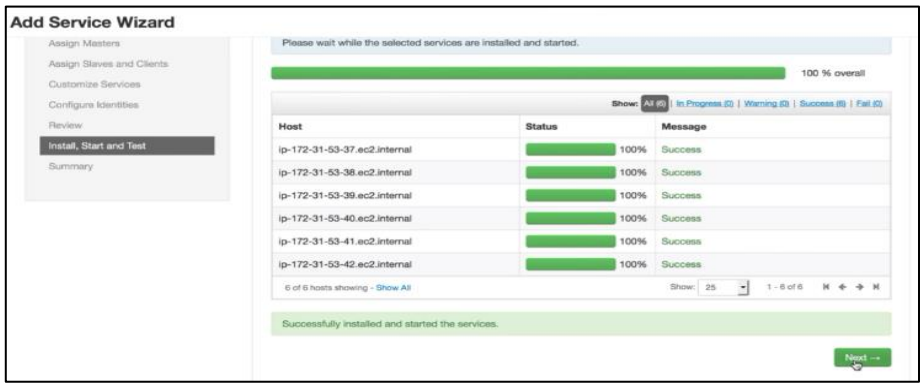


Figure 12. Installation success

3.5. Restarting services in Knox using ambari

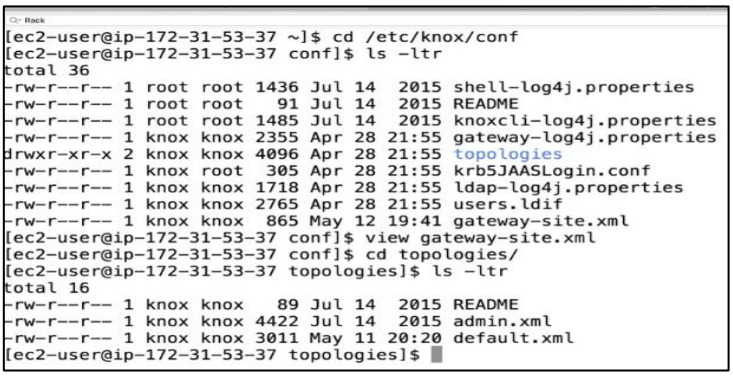
Unless all the services are restarted, desired results cannot be obtained. If it shows an orange cycles near the services menu, then ensure that all the services has to be restarted. To do that click on orange cycles->restart->restart all Affected->confirm restart All. So that, all the services will 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

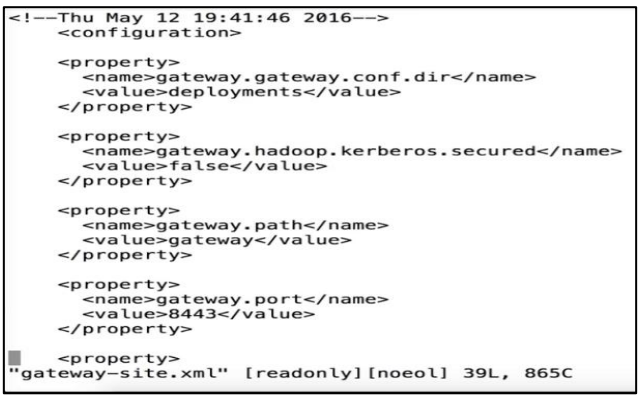
These commands shows gateway-site.xml, which is represented in Figure.13,14.

Knox topologies



```
[ec2-user@ip-172-31-53-37 ~]$ cd /etc/knox/conf
[ec2-user@ip-172-31-53-37 conf]$ ls -ltr
total 36
-rw-r--r-- 1 root root 1436 Jul 14  2015 shell-log4j.properties
-rw-r--r-- 1 root root   91 Jul 14  2015 README
-rw-r--r-- 1 root root 1485 Jul 14  2015 knoxcli-log4j.properties
-rw-r--r-- 1 knox knox 2355 Apr 28 21:55 gateway-log4j.properties
drwxr-xr-x 2 knox knox 4096 Apr 28 21:55 topologies
-rw-r--r-- 1 knox root  305 Apr 28 21:55 krb5JAASLogin.conf
-rw-r--r-- 1 knox knox 1718 Apr 28 21:55 ldap-log4j.properties
-rw-r--r-- 1 knox knox 2765 Apr 28 21:55 users.ldif
-rw-r--r-- 1 knox knox  865 May 12 19:41 gateway-site.xml
[ec2-user@ip-172-31-53-37 conf]$ view gateway-site.xml
[ec2-user@ip-172-31-53-37 conf]$ cd topologies/
[ec2-user@ip-172-31-53-37 topologies]$ ls -ltr
total 16
-rw-r--r-- 1 knox knox   89 Jul 14  2015 README
-rw-r--r-- 1 knox knox 4422 Jul 14  2015 admin.xml
-rw-r--r-- 1 knox knox 3011 May 11 20:20 default.xml
[ec2-user@ip-172-31-53-37 topologies]$
```

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



```
<!--Thu May 12 19:41:46 2016-->
    <configuration>

    <property>
      <name>gateway.gateway.conf.dir</name>
      <value>deployments</value>
    </property>

    <property>
      <name>gateway.hadoop.kerberos.secured</name>
      <value>false</value>
    </property>

    <property>
      <name>gateway.path</name>
      <value>gateway</value>
    </property>

    <property>
      <name>gateway.port</name>
      <value>8443</value>
    </property>

    <property>
"gateway-site.xml" [readonly][noeol] 39L, 865C
```

Figure 14. Knox-gatewaysite.xml

```
        The name parameter specifies the external host names in a comma separated list.
        The value parameter specifies corresponding internal host names in a comma separated

        Note that when you are using Sandbox, the external hostname needs to be localhost, as
out
        of box sandbox.xml.  This is because Sandbox uses port mapping to allow clients to co
the
        Hadoop services using localhost.  In real clusters, external host names would almost
localhost.
        -->
        <provider>
            <role>hostmap</role>
            <name>static</name>
            <enabled>true</enabled>
            <param><name>localhost</name><value>sandbox,sandbox.hortonworks.com</value></para
        </provider>

    </gateway>

    <service>
        <role>KNOX</role>
    </service>

</topology>
```

Figure 15. Knox-admin.xml

```
        <topology>

            <gateway>

                <provider>
                    <role>authentication</role>
                    <name>ShiroProvider</name>
                    <enabled>true</enabled>
                    <param>
                        <name>sessionTimeout</name>
                        <value>30</value>
                    </param>
                    <param>
                        <name>main.ldapRealm</name>
                        <value>org.apache.hadoop.gateway.shirorealm.KnoxLdapRealm</value>
                    </param>
                    <param>
                        <name>main.ldapRealm.userDnTemplate</name>
                        <value>uid={0},ou=people,dc=hadoop,dc=apache,dc=org</value>
                    </param>
                    <param>
                        <name>main.ldapRealm.contextFactory.url</name>
                        <value>ldap://ip-172-31-53-37.ec2.internal:33389</value>
"default.xml" [readonly][noeol] 89L, 3011C
```

Figure 16. Default.xml

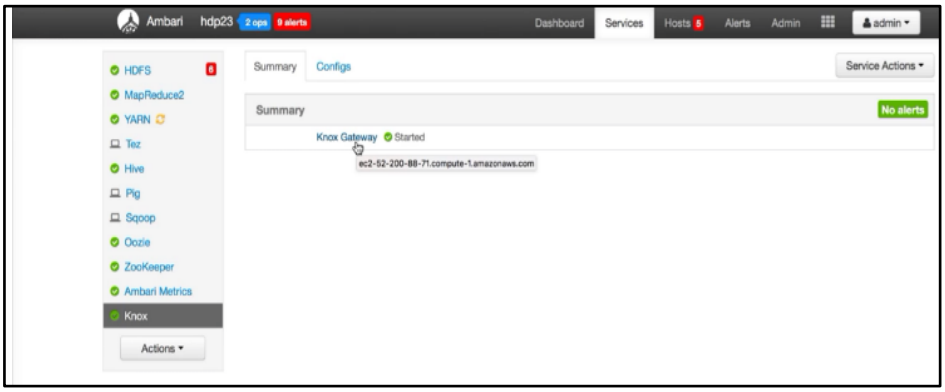


Figure 17. Knox-Ambari

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

Curl –I –v –k admin:admin –password –x GET 'https://ip-172-31-53-37.ec2.internal:8443/gateway/hdp23/OOZIE/v1/admin/build-version'.

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

[1] N.Sirisha,K.V.D.Kiran, "Protection Of Encroachment On Bigdata Aspects", International Journal of Mechanical Engineering and Technology (IJMET), Volume 8, Issue 7, July 2017, pp. 550–558.

[2] Priya P. Sharma, Chandrakant P. Navdeti , "Securing Big Data Hadoop: A Review of Security Issues, Threats and Solution" ,(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2126-2131, ISSN:0975-9646.

[3] T.K.Das ,P. Mohan Kumar," BIG Data Analytics: A Framework for Unstructured Data Analysis", International Journal of Engineering and Technology (IJET), ISSN : 0975-4024 Vol 5 No 1 Feb-Mar 2013.

[4] Sanjay Rathee," Big Data and Hadoop with components like Flume, Pig, Hive and Jaql", International Conference on Cloud, Big Data and Trust 2013, Nov 13-15.

[5] Priya P. Sharma, Chandrakant P. Navdeti, " Securing Big Data Hadoop: A Review of Security Issues, Threats and Solution", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2126-2131.

[6] Varsha B.Bobade,"Survey Paper on Big Data and Hadoop**",** International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 03 Issue: 01 | Jan-2016

[7] Sanjay Rathee, "Big Data and Hadoop with components like Flume, Pig, Hive and Jaql", International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV.

[8] Saranya, S.," Dynamic Preclusion of Encroachment in Hadoop Distributed File System", Procedia Computer Science 50 (2015): 531-536.

[9] Dr. Md. Tabrez Quasim and Mohammad. Meraj, "Big Data Security and Privacy: A Short Review", International Journal of Mechanical Engineering and Technology, 8(4), 2017, pp. 408-412.

[10] Vinay Shukla, "Hadoop Security: Today and Tomorrow", https://hortonworks.com/blog/hadoop-security-today-and-tomorrow,December 09, 2013.

[11] Prachi R. Gawali, Roshani Talmale, Rajesh Babu,"Hadoop Security- A Review", International Journal of Innovative Research in Computer and Communication Engineering ,Vol. 3, Issue 10, October 2015.

[12] Horton works "Technical Preview for Apache Knox Gateway"

[13] "KNOX", Apache Knox Gateway 0.9.x User's Guide, https://knox.apache.org/books/knox-0-9-0/user-guide.html.

[14] "HDP Documentation", https://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk-know- Gateway-Admin-Guide/content.

[15] IbrahimAbaker, TargioHashemaIbrar, Yaqoob, NorBadrulAnuar, SalimahMokhtar, AbdullahGania, SameeUllah Khan, "The rise of "big data" on cloud computing: Review and open research issues" Information Systems 47 (2015) 98–115.

[16] Seref sagiroglu and Duygu SINANC, "Big Data: A Review Collaboration Technologies and Systems (CTS), 2013 International Conference ,May 2013"

[17] Victor L. Voydock and Stephen T. Kent "Security mechanisms in high-level network protocols. ACM Comput. Surv.1983".

[18] Sujitha, G., et al. Improving security of parallel algorithm using key encryption technique. Information Technology Journal 12.12 (2013): 2398.

[19] Cohen, Jason C., and Subrata Acharya." Towards a trusted HDFS storage platform: Mitigating threats to Hadoop infrastructures using hardware-accelerated encryption with TPM-rooted key protection". Journal of Information Security and Applications 19.3 (2014): 224-244.

[20] Hashem, Ibrahim Abaker Targio, et al. The rise of "big data" on cloud computing: Review and open research issues. Information Systems 47 (2015): 98-115.

[21] Jeong, Yoon-Su, and Yong-Tae Kim. A token-based authentication security scheme for Hadoop distributed file system using elliptic curve cryptography. Journal of Computer Virology and Hacking Techniques 11.3 (2015): 137-142.

[22] Jeong, Yoon-Su, Seung-Soo Shin, and Kun-Hee Han. High-dimentional data authentication protocol based on hash chain for Hadoop systems. Cluster Computing 19.1 (2016): 475- 484.

[23] Saraladevi, B., et al. Big Data and Hadoop-A study in security perspective. Procedia computer science 50 (2015): 596-601.

[24] Dr. Md. Tabrez Quasim and Mohammad. Meraj, Big Data Security and Privacy: A Short Review, International Journal of Mechanical Engineering and Technology, 8(4), 2017, pp. 408-412.

[25] Subashini, M.M., Das, S., Heble, S., Raj, U., Karthik, R., "Internet of things based wireless plant sensor for smart farming", Indonesian Journal of Electrical Engineering and Computer Science, Vol. 10, Issue 2, pp. 456-468, (2018).

[26] Nagaraju, J., Jyothi, K., Karthik, R., Bhaskara Reddy, P., Vucha, M., "Distributed optimal relay selection in wireless sensor networks", Indonesian Journal of Electrical Engineering and Computer Science, Vol. 7, Issue 1, pp. 71-74, (2017).
[27] Ranjith, S., Shreyas, Pradeep Kumar, K., Karthik, R., "Automatic border alert system for fishermen using GPS and GSM techniques", Indonesian Journal of Electrical Engineering and Computer Science, Vol. 7, Issue 1, pp. 84-89, (2017).