International Journal of Engineering & Technology, 7 (3.6) (2018) 234-236

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Authorization of Data In Hadoop Using Apache Sentry

N. Sirisha[1*], K.V.D. Kiran[2]

[1]Department Of Computer Science And Engineering, K L University, Vaddeswaram, Guntur District, Andhra Pradesh, India
Department Of Computer Science And Engineering, MLR Institute Of Technology, Dundigal, Hyderabad, India.
[2]Department Of Computer Science And Engineering, MLR Institute Of Technology, Dundigal, Hyderabad, India.
**Corresponding Author E-Mail:Grandhishirisha@Gmail.Com*

Abstract

Big Data has become more popular, as it can provide on-demand, reliable and flexible services to users such as storage and its processing. The data security has become a major issue in the Big data. The open source HDFS software is used to store huge amount of data with high throughput and fault tolerance and Map Reduce is used for its computations and processing. However, it is a significant target in the Hadoop system, security model was not designed and became the major drawback of Hadoop software. In terms of storage, meta data security, sensitive data and also the data security will be an serious issue in HDFS. With the importance of Hadoop in today's enterprises, there is also an increasing trend in providing a high security features in enterprises. Over recent years, only some level of security in Hadoop such as Kerberos and Transparent Data Encryption(TDE),Encryption techniques, hash techniques are shown for Hadoop. This paper, shows the efforts that are made to present Hadoop Authorization security issues using Apache Sentry in HDFS.

Keywords: *Hadoop, apache sentry, security, TDE, encryption zone, knox, ranger.*

1. Introduction

In today's world, Big Data is a prominent trend originated from an era of cloud computing. The cloud has ideal platforms for making use of Big Data. Data which is beyond the storage capacity and processing techniques is nothing but a Big Data. Sensors, CC cams, online shopping, Air lines, NCDC, hospitals data are the different data generated factors. As number of applications in cloud is increasing such as manufacturing, Health care, Insurance and retail, the data security is becoming an major era in the big data. In 1990's 1GB to 20 GB data is getting stored, as days are passing on in 2014 1TB to 100 TB of data is getting stored and getting processed, which leads to a storage problem. In Big data, data security is becoming a major problem especially when the number of individuals in the enterprises that manages sensitive data to process their private data such as healthcare and financial records.

Hadoop is one of the most popular framework for Big data analysis. Recent work towards data protection like Transparent Encryption for Hadoop [2] and Cloudera's "Security for Hadoop" [3] are all vital steps that focus on some of the security issues that stems when Hadoop framework is considered. Moreover, few of the techniques such as encryption-decryption is weirdly used, whereas the implications of high availability, data security, replication and are leaved unnoticed. This can be maintained by the enterprise itself [4].Major challenges that are noticed by the Hadoop investigator includes the location of the data leakage node among various hadoop nodes.

Highlighting these challenges, some measures are taken to provide security in Hadoop HDFS. Techniques like Kerberos for authentication, Outh for authorization, and ACLs for data protection can be used in Big data.

2. Related Work

Present research on Hadoop security shows that authentication, Authorization, data protection are the areas where security issues arises. In Big data, a secure Hadoop[3] architecture was proposed that adds encryption and decryption functions in the Hadoop distributed file system(HDFS). In this method, HDFS data cannot be readable even if it is accessed by third party because the attacker may not have the secretkey. Data is secured using the secret key with a concept of encryption and decryption. Even though it is a basic solution for securing Hadoop, performance is high.

In [5], the foremost review was to reduce the complexity and cost of hadoop cluster using Hadoop-as-a-service offerings as a public cloud service provider. To provide security this work shown a novel algorithm called SDFS design and implementation is shown. This work analyzed the performance of SDFS and minimized computational overhead. In[6],Chandni Grover used a kadmin. Local utility provided by MIT KDC to create admin user for KDC. Kerberos use this JCE to encrypt or decrypt the Kerberos ticket it generates. In this Admin user trying to get the initial credentials from KDC Database of Kerberos. Ticket generated by the TGS for admin user. Ranger Creating Policy for user and Groups for Different Files and Directories. In [7], the work shows that there is no effective mechanism for file privacy protection HDFS, so it is unsecure to apply it in real cloud environment. In this paper, a data encryption method based on HDFS is presented. We employed hybrid encryption scheme to protect file blocks and session keys, which can prevent datanode intruders from stealing user data. In contrast to the other similar works, we keep the advantage of light weightness for client. The experiments show that the proposed method introduces 43% overhead, but the architecture overhead is negligible. Therefore, the future work is to take advantage of GPUs or multicore technology for paralleling

the encryption/decryption modules to improve the overall performance.
Unlike [7], this work shows the authorization of data in hadoop using apache sentry. Apache Sentry works well in providing the authorization, which is a security drawback in Hadoop HDFS. All the security issues in hadoop cluster are mentioned and its solution using Apache sentry is shown with the help of Sentry. The configuration of sentry is finely explained in this paper.

3. Security Issues in Hadoop Cluster

- Unauthorized clients can act like an authorized users and access the cluster.
- Retrieve the blocks directly from the data nodes by bypassing the name node.
- Unauthorized accessof data packets by the third party being sent by data nodes to client.
- Not all users should have access to sensitive data
- No user verification for Map Reduce code Execution, malicious users could submit a job
- Insecure network transport
- No message level security.

Hadoop security considerations- Reasons for security in HADOOP[11]

- Hadoop has sensitive data-As hadoop is growing, different data organizations look to store. Often the data is proprietary of personal and it must be protected.
- Hadoop is subject to compliance adherence-It should follow some government regulations, compliance like HIPPA, PII, FISMA.

4. Hadoop Security Primer

Hadoop security primarily depends on i) Authentication ii) Authorization iii)Data Protection iv) Governance and Auditing.

Authentication

Authentication is identifying the user. Trusted users doesn't have access to the cluster network. In a trusted network, who you are is determined by a client host. Strong Authentication is provided by few techniques like Kerberos, LDAP Active Directory, LDAP, AD integrated with Kerberos, establishing a single point of truth and single sign On.

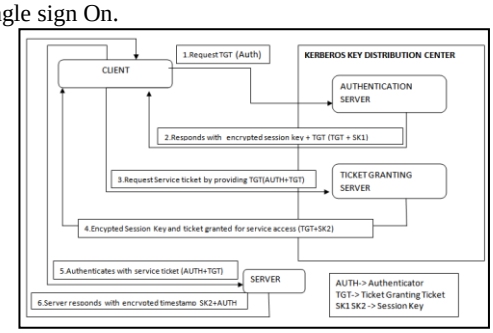

Fig. 1: Kerberos

Authorization

Authorization[12] determines if you can access. Permissions such as X/W/R for U/G/O are permitted by HDFS POSIX. Other components like MR JOB QUEUE, HBASE ACLS have authorization on tables & Column family. Accumulo provides cell - level access control and impersonation [12]. Authorization using Apache sentry is shown in next section.

Data Protection

Data protection is must when the data is at REST and when the data is at Transit. When the machines are in OFF state, the data will be available at rest. Some of them includes data on hard drives, flash drives and USB. Encryption on data is provided by Hadoop in Transit using Hypertext Transfer Protocol (HTTP), Java Database Connectivity/Oracle Database Connectivity (JDBC/ ODBC), Distributed Transaction Processing(DTP), Remote Procedure Call (RPC). Basically, Hadoop does not have native encryption on data at rest(HDFS-6134).

Governance and Auditing

Distributed file system and Map Reduce provides basic audit support.

5. Hadoop Authorization Using Apache Sentry

There are many traditional IT security controls that can be used for securing a Hadoop environment. The standard protection controls include SIEM (security information and event management), network firewalls, IDPS (intrusion detection and protection systems), vulnerability management, configuration control, etc. All these are general security control levels. For the next advanced level of security, the open source community has been investing heavily in developing Hadoop best practices and specific tools to provide enterprise grade security. The pillars of Hadoop security are: audit, authentication, data protection and authorization. Hadoop authorization is considered as the limited ability to deny or grant access rights on a granular basis. Due to the availability of SQL-style authorization to Hive and HDFS ACLs in Hadoop, this is improving. With the use of the above tools, the capability to track and audit each user accesses to particular services and HDFS data components is also improving. Current authorization is fragmented, coarse - grained and manual [13].

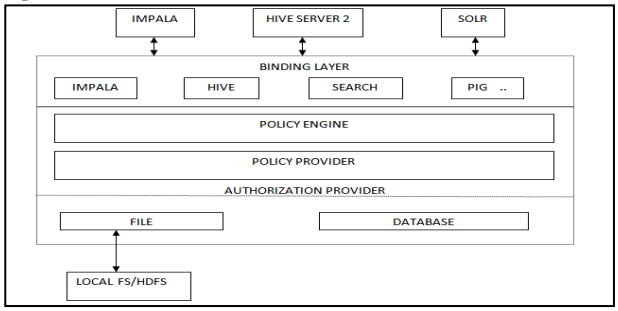

Fig. 2: Sentry Architecture

Key Benefits of Apache Sentry

It stores sensitive data in Hadoop and extended Hdfs to more users has a issue with regulations. Servers, Databases, Tables & views, Indexes, Collections are all fine grained. Privileges such as SELECT, INSERT are all role-based. Apache sentry[13] is capable of multi-tenant administrations where different policies for each database schema can be maintained by different admins.

Working of Sentry

Apache sentry[14] first validates SQL grammar and constructs tree to validate statement objects to check Authorization and forwards to execution planner. Apache sentry has actors to define Authorization policies.
Actors in Apache sentry are user, user group, resources, privileges, role. User actor is to authenticate user where the

identity of user can be obtained from session context. User group actor is defined beyond sentry policy which is obtained from user directory (LDAP,AD, HDFS) and also It can be available from session context. Actor-Resources are to protect data in files, directory on HDFS, in tables or views in Hive, URI, Resource can be hierarchical. Privilege actoris the action or operation associated with a resource. Actor-Roles is a collection of privileges defined in sentry policy.

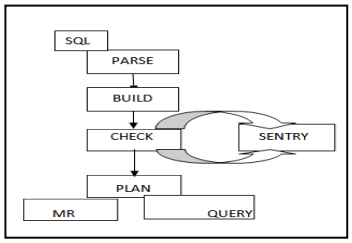

Fig. 3: Working of apache sentry

Working of apache sentry shows that it

- validates SQL grammar
- construct statement tree
- validate statement objects
- forward to execution planner

This section also shows you the installation of Apache Sentry in Hadoop Cluster using cloud era manager. Cloud era manager is a tool where hadoop[15] environment can be easily accessed.

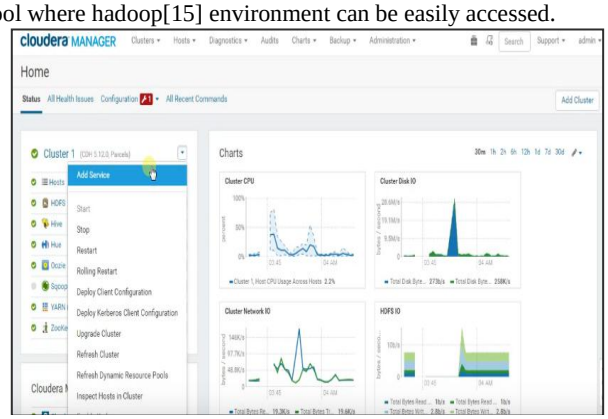

Fig. 4: Adding apache sentry service to cloudera manager

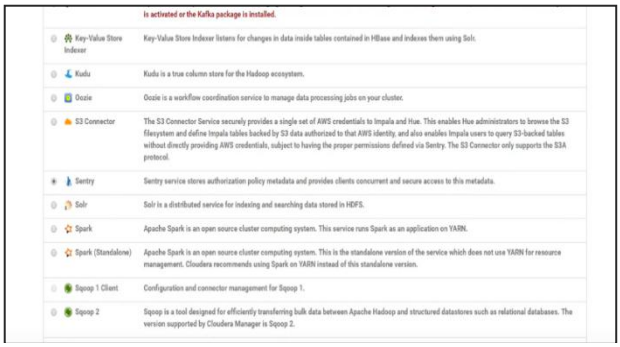

Fig. 5: Adding sentry service to cluster1

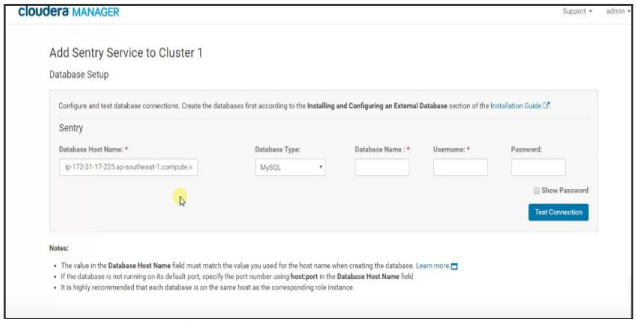

Fig. 6: Addition of database hostname, type, database name and credentials to test the connection

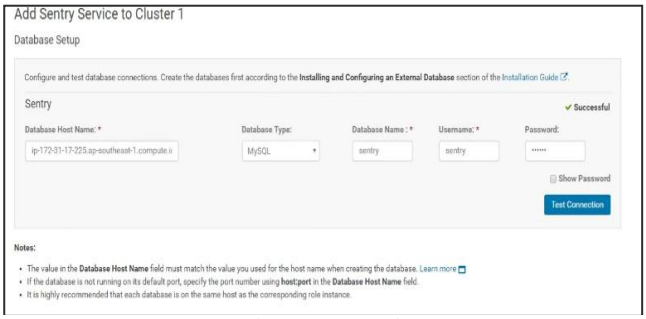

Fig. 7: Successful installation of apache sentry

6. Conclusion

Presently, security in Big Data is a major area, where all the information is mined from different sources of data ware house to a single distributed environment. So, the security is a primary issue. This work, presents the security in terms of data authorization at an HDFS storage level which is not achieved by Kerberos. This paper discusses the way sensitive data is secured and how Apache sentry is used to protect sensitive data in the HDFS and how it provides the authorization of data in Hadoop environment. In Future, later versions of Hadoop with a high variety of security mechanisms for securing data is necessary.

References

[1] Sirisha N & Kiran KVD, "Protection Of Encroachment On Bigdata Aspects", *International Journal of Mechanical Engineering and Technology (IJMET)*, Vol.8, No.7, (2017), pp.550–558.

[2] Park S & Lee Y, "Secure Hadoopwith Encrypted HDFS", *Springer-Verlag Berlin Heidelberg*, (2013), pp.134–141.

[3] Dean J & Ghemawat S, "MapReduce: simplified data processing on large clusters", *CACM*, Vol.51, No.1, (2008), pp.107-113.

[4] Park S & Lee Y, "Secure hadoop with encrypted HDFS", *International Conference on Grid and Pervasive Computing*, (2013), pp.134-141.

[5] Zerfos P, Yeo H, Paulovicks BD & Sheinin V, "SDFS: Secure distributed file system for data-at-rest security for Hadoop-as-a-service", *IEEE International Conference on Big Data (Big Data)*, (2015), pp.1262-1271.

[6] Grover C & Aulakh MK, "Big Data Authentication and Authorization in HDP (Hadoop Distributed platform) using Kerberos and Ranger", *2nd International Conference on Recent Innovations in Management and Engineering*, (2017), pp.44-51.

[7] Cheng Z, Zhang D, Huang H & Qian Z, "Design and Implementation of Data Encryptionin Cloud based on HDFS", *International Workshop on Cloud Computing and Information Security*, (2013), pp.274-277.

[8] Shehzad D, Khan Z, Dag H & Bozkus Z, "A novel hybrid encryption scheme to ensure Hadoop based cloud data security", *International Journal of Computer Science and Information Security*, Vol.14,No.4,(2016).

[9] Rabin MO, "Efficient Dispersal of Information for Security, Load Balancing, and Fault Tolerance", *Journal of the Association for Computing Machinery*, Vol.36, No.2, (1989), pp.335-348.

[10] "Transparent Encryption in HDFS. https://hadoop.apache.org/docs/r2.7.2/hadoop-project dist/hadoophdfs/TransparentEncryption.html.

[11] Byers J, Luby M, Mitzenmacher M & Reg A e, "A Digital Foundation Approach to Reliable Distribution of Bulk Data", *Proc.ACM SIGCOMM*'98, Vol.28, No.4, (1998), pp.56-67.

[12] Darade SA & Kamble K, "Network Level Security in Hadoop Using Wire Encryption", *International journal of Advanced research in science management and technology*,Vol.1, No.6, (2015).

[13] Cloudera Inc., "HDFS Data At Rest Encryption", http://www.cloudera.com/content/cloudera/en/documentation/core/latest/topics/cdh_sg_hdfs_encryption.html#xd_583c10bfdbd326ba--5a52cca-1476e7473cd--7f85, 2015.

[14] IBM BigInsights on Cloud, IBM, 2016. http://www-03.ibm.com/software/products/en/ibm-biginsights-oncloud.

[15] Vivekanand & Vidyavathi BM, "Security Challenges in Big Data: Review", *International Journal of Advanced Research in Computer Science*, Vol.6, No.6, (2015).