International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Security Issues of Big Data Hadoop

Rohit Sharma
P.G. Department of Computer Science & IT,
SGAD College
Khadur Sahib (Tarn Taran)

Abstract: In this we discuss security issues for big data Hadoop environment. Big data applications are a great benefit to organization, business and in many small and large scale industries. Security and privacy issues are magnified by velocity, variety and volume of big data. Hadoop projects security as top agenda which in turn represents classified as critical term. With the increasing acceptance of Hadoop, there is increasing trend to create a vast security feature. Therefore a traditional security mechanism, which are tailored to securing a small scale static data are in adequate. The important issues relating to Hadoop are authentication, authorization, editing and encryption within a cluster. In this paper we have highlighted different security aspects of big data Hadoop.

Index Terms: Big data, Hadoop ,HDFS, mapreduce

INTRODUCTION

Big data [1] is a word use to describe massive volumes of structured and unstructured data that are so large that it is very difficult to process. The main terms which represent big data have the following properties as shown in the Fig. 1.

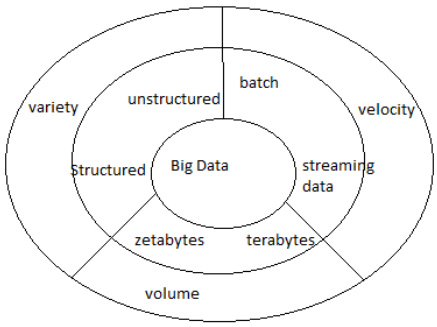

Fig.1. Characterization of big data-variety, velocity and volume

Volume

Many factor contributing towards increase volumes [9] .for example social networking sites, data collection from sensors, machines, networks etc.

Variety

Data comes in different formats. For examples traditional database, text documents, audio and video files. In past data was stored in the form of documents, spreadsheets, and databases. Now in this data comes from emails, pictures, PDF's etc.

Received: 5 March Revised: 13 March Accepted: 22 March
Index in Cosmos
April 2018 Volume 8 Number 4
UGC APPROVED

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Velocity

This means at a very high speed the data is being produced and how much speed we need to be processed to meet the demand.

Complexity

The data comes from multiple sources .It is unstructured so it is to be transformed into required format before actual processing [2].

The volumes of big data are on a very high node, which can be seen from the fact that in the year 2012 there were few dozen terabytes of data in a single dataset which have increase to petabytes today. To carter the demands of industry new manifestos of big data re commissioned. 5 Exabyte (1 exabyte =1.1529*1018bytes) of data were created by human until 2003.now days same amount of information is processed in two days. In 2012 data was expanded to 2.72 zettabyte. It is forecast that this figure is double every two years, reaching the number to 8 zettabyte by 2015[4].

The today technology not only support large amount of data, but also help in utilizing such data effectively. Some of the real time example of big data is transactions made from the credit cards, face book, Twitter and what's app are generating the social networking data. To process the large amount of data from different sources, the Hadoop is used.

Hadoop is a free, java based programming based framework that supports the processing of large data sets in distributed environment. Hadoop allows running applications on systems with thousands of nodes with thousands of terabytes of data. Hadoop consists of distributed file system, data storage and analytics platforms and a layer that handles parallel computation, rate of flow and configuration administration. HDFS runs across the nodes in a Hadoop cluster and connects the file system on many input and output data nodes to make them into one big file system. The Hadoop architecture consist of Hadoop kernel,mapreduce ,PIG, hive Hadoop Distributed File System (HDFS) and number of related components.

- Hdfs:
- Mapreduce
- Hadoop kernel
- PIG
- Hive

TRADITIONAL HADOOP SECURITY

When Hadoop was developed there was no security model to implement the security. There was no authentication of users and services. Although auditing and authorizations controls were used in earlier distributions, such controls can be evaded as any other user can impersonate any other user. The security features of authorization and authenticity was later added which have some weakness because all the programmers have given the same privileges to all data in the cluster, any job could access any of the data in cluster any user could read any data set. The user could lower the priories of the other to increase his priorities to complete his job faster as mapreduce has no concept of authorization or authentication. The Hadoop supports some security features of Kerberos implementation, the use of firewalls and the basic HDFS permissions. Kerberos is not a

Received: 5 March Revised: 13 March Accepted: 22 March
Index in Cosmos
April 2018 Volume 8 Number 4
UGC APPROVED

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

compulsory requirement for Hadoop and it is not easy to install and configure on the cluster. Hadoop security is not properly implemented by firewalls, as the firewall is breached; the cluster is wide open for attack. It also offers no protection from security failure which originates from firewalls. An attacker can steal the data from the centre as data is unencrypted and there is no authentication enforced or access [5]. There are different categories of security violations like unauthorized release of information, modification of information and denial of resources. Different types of threat are:

- An unauthorized user may gain access privileges and submit a job to change the priority of job.
- An unauthorized client may read/write the data block of file at a data node.
- An unauthorized user may access an HDFS file via http
- A user may submit an overflow to data centre as another user.
- An unauthorized user may access intermediate data of map job via its task trackers.

SECURITY ISSUES AND CHALLENGES

Hadoop presents some different sets of security issues for data centers managers and security professionals. The various security issues and challenges are:

Fragmented data

Big data clusters contain data that portray the quality of fluidity,allowing multiple copies moving from one node to another which ensures redundancy and resiliency [7].the data is available for fragmentation which is shared among multiple servers which results in more complexity as there is no security model to handle this issue.

Node to node communication

The main issue with Hadoop is they don't implement secure communication; they bring into the use of RPC over TCP/IP[7].

Distributed computing

Since the available of resources increases with distributed computing as the data is processes at any instant where it is available. This results in high risks of attacks then in the centralized computing.

Interaction with client

Communication with client takes place with resource manager, nodes. Even communication is efficient but it is difficult to shield nodes from clients and name server from nodes.

Controlling data access

The available database security schema provides role base access. Big data was designed with very little security in mind. The installation of big data is based on web services model with very little security for preventing web threat making it a highly susceptible.

Received: 5 March Revised: 13 March Accepted: 22 March
Index in Cosmos
April 2018 Volume 8 Number 4
UGC APPROVED

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

SECURITY SOLUTIONS FOR HADOOP

Security is an important concern for enterprise software. Hadoop is a distributed system which allows us to store and process large amount of data in parallel. Hadoop is used as an multitenant service which stote the sensitive data such as financial and the personal data. To protect the sensitive data we need a strong authentication and authorization [11].

The Hadoop ecosystem consists of various components and all these components are need to be secure. Each component has its own security complication issues and should need to be configured properly based on the architecture[10]. So in this part we are providing the different security solutions with the technologies for the securing the big data Hadoop.

Authentication

Authentication is used to verifying user or system identity accessing the system.Hadoop provides Kerberos as a main authentication .initially SASL/GSSAPI was used to implement Kerberos and mutually authenticate users, applications and services over the RPC connections [12]. Hadoop also supports Pluggable authentication for HTTP web consoles which means web applications and consoles could implements their own authentication mechanism for HTTP connections..The Hadoop components supports SASL framework that is the RPC layer can be changed to support the SASL based mutual authentication. Mapreduce supports Kerberos authentication. In hdfs communication between the namenode and data node is over RPC connection and the mutual Kerberos is performed between them.pig uses the user credentials to submit the job to Hadoop[7,8].so there is no need of any additional requirement of Kerberos security authentication. In pig user should authenticate with KDC and get a valid Kerberos ticket.

Oozie provides user authentication to the oozie web services. It also provides Kerberos HTTP simple and protected GSSAPI NEGOATION mechanism (SPENGO) authentication for web clients. SPENGO is used when user want to communicate to remote server, but is not sure of the authentication protocol to use. There are number of data flows involved in Hadoop authentication like Kerberos RPC is used FOR THE USER AUTHENTICATION, applications and services. SPNEGO is used for web consoles and used for delegation tokens.[12,13] Delegation token is a two party authentication protocol used between user and name node for authenticating users.map reduce, oozie and hdfs supports delegation token. The technology used for the authentication and access for various Hadoop services in cluster is:

1) Apache Knox

This is system that provides a single point of authentication and access for various Hadoop servies in a cluster.Hadoop supports various authentication and token verification scenarios. It manages security across multiple clusters versions of Hadoop. The goal of the Knox Gateway is to provide a single point of secure access for Hadoop clusters. The solution is implemented as a gateway (or small cluster of gateways) that exposes access to Hadoop clusters through a Representational State Transfer (REST)- full API. The gateway provides a firewall between users and Hadoop clusters and can manage access to clusters that run different versions of Hadoop shown as Fig. 2.

Received: 5 March Revised: 13 March Accepted: 22 March
Index in Cosmos
April 2018 Volume 8 Number 4
UGC APPROVED

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

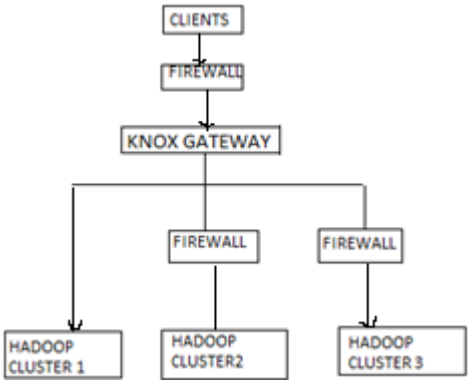

Fig. 2. Perimeter security with the Apache Knox Gateway

Authorization

It is process in which privileges for user or system is provided. In Hadoop access control is implemented by using file based permission which follows UNIX permission model. Hadoop offers authorization using file permission in HDFS and resource level access control using acls for mapreduce access control at service level.Haddop base offers user authorization on tables ,column .the user authorization is implemented using coprocessors ,which are like database triggers in Hadoop base[10].They intercept any request to the atble before and after. In Hive, authorization is implemented using Apache sentry.pig provides authorization using ACL's for job queues. Zookeper also offers authorization using node ACL's. Hue provides access control via file system permission.

Now a day's many organizations used more flexible and dynamic access control policies based on XACML and attribute based access control. Hadoop cannot be configured to support RBAC, ABAC ACCESS CONTROL USING THIRD party framework or tool. Zettaset orchestration provides a role based access control support and enables Kerberos to be seamlessly integrated into Hadoop system [10]. The technology used for the authorization module for Hadoop is:

2) Apache Sentry

Apache Sentry (incubating) is a system for enforcing the fine grained role based authorization to data and metadata stored on a Hadoop cluster. Apache Sentry is an effort undergoing incubation at The Apache Software Foundation (ASF). Incubation is required of all newly accepted projects until a further review indicates that the infrastructure, communications, and decision making process have stabilized in a manner consistent with other successful ASF projects. While incubation status is not necessarily a reflection of the completeness or stability of the code, it does indicate that the project has yet to be fully endorsed by the ASF[8].

Encryption

Encryption is a method to ensure the confidentiality and privacy of user information. It secures the sensitive data in Hadoop.Hadoop is a distributed system in which data is to be transmitted over the network, so there is a need of demand to move the sensitive information with some special kind of protection and should be secure both at rest and in motion. This data should be protected during the transfer. The Simple Authentication and Security Layer (SASL) is used for encrypting the data in

Received: 5 March Revised: 13 March Accepted: 22 March
Index in Cosmos
April 2018 Volume 8 Number 4
UGC APPROVED

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

motion in Hadoop system. SASL technique gives the guarantee of the data being exchanged between client and server and make sure that it is not encrypted by the middle man. The two ways by which the data at rest is protected are firstly: when the file is stored in Hadoop, the complete file can be encrypted first and then stored in Hadoop[10]. Secondly to apply encryption t data blocks once they are loaded into Hadoop system. Hadoop provides various encryption for various channels like RPC,HTTP,and data transfer protocol for data in motion. Hadoop crypto codec framework and crypto codec implementation have been to support data at rest encryption. HDFS supports AES,OS Level encryption for data at rest. ZOOKEEPER, OOZIE, HIVE, HADOOPBASE, PIG Doesn't offer data at rest encryption, for this the encryption can be implemented via custom encryption techniques. To protect data in motion and at rest, encryption making techniques can be implemented. Intel distribution offers encryption and compression of files.

3) Project Rhino

It provides block level encryption for the data stored in the Hadoop.It supports key distribution and management so that MR can decrypt data block and execute the program as per requirement. It also enhances the security of HBASE by offering cell level authentication and transparent encryption for table stored in Hadoop. It also provides token based authentication.

CONCLUSION

During the initial days of Big Data implementations using Hadoop, the main motivation was to get data into the Hadoop cluster and perform analytics on it. With the Hadoop gaining larger acceptance within the industry, a natural concern over the security has increased. As organizations have matured their understanding of Big Data, the data security and privacy policies of such implementations are being questioned. Though Hadoop lacks a robust security and privacy framework, the increasing interest in this area is ensuring that appropriate solutions are developed. However suppliers are now emerging with new products which secure data in ways which are almost transparent to user.While security and privacy issues can be addressed to an extent using existing Hadoop mechanisms, more robust tools and techniques are needed.

REFERENCES

1. Ethics of Big Data Kord Davis, Doug Patterson , O'Reilly Media.
2. A,Katal Wazid M ,and Goudar R.H. "Big Data :Issues,challenges,tools and good practices."Noida;2013 PP 404-409,8-10 AUG 2013.
3. Securing Big Data:Security Recommendations for Hadoop and nosql Environments."Securosisblog version 1.0(2012).
4. Zettaset "the Big data security Gap :Protecting the hadoop cluster".
5. K.Chitranjan, and kala karun"A Review on Hadoop-HDFS infrastructure extension"jeju Island Pp132-137, 2013.
6. Kevin t.smith "Big data security:the evolution of Hadoop Security model".
7. M.Tim Jones"Hadoop security and sentry".
8. Vinay shukla:Hadoop Security-today and tomorrow.
9. Douglas,Laney."The impotance of Big Data:A Defination".Gartner Retrieved june 21 2012.

Received: 5 March Revised: 13 March Accepted: 22 March
Index in Cosmos
April 2018 Volume 8 Number 4
UGC APPROVED

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

10. Sudesh nararyan,"securing Hadoop implement robust end to end security for your hadoop ecosystem".
11. Data Security for Hadoop – Add-on Choices Proliferating, Merv Adrian, Gartner, 2014.
12. Hadoop security: A jungle of options, Michael Steinhart, AllAnalytics.com, 2014.

Received: 5 March Revised: 13 March Accepted: 22 March
Index in Cosmos
April 2018 Volume 8 Number 4
UGC APPROVED