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. 3, Issue 10, October 2015

Hadoop Security- A Review

[1]Prachi R. Gawali, [2] Roshani Talmale, [3] Rajesh Babu

[1] M.Tech Student, Dept of CSE, Tulsiramji Gaikwad-Patil College of Engg and Tech, Nagpur, India.

[2]Head of the Department, Dept of CSE, Tulsiramji Gaikwad-Patil College of Engg and Tech, Nagpur, India.

[3]Assistant Professor, Dept of CSE, Tulsiramji Gaikwad-Patil College of Engg and Tech, Nagpur, India.

ABSTRACT: In this new era of Big Data, with cheap data storage devices and cheap processing power becoming available, organizations are collecting massive volumes of data, with the intent of deriving insights and making decisions. While most of the focus is on collecting data, also the security and privacy issues are magnified by the volume, variety, and velocity of Big Data. The diversity of data sources, formats, and data flows, combined with the streaming nature of data collection and high volume create unique security risks, having all data at one place increases the risk of data security and any kind of data breach can lead to negative publicity and a loss of customer confidence. Hadoop is one of the main technologies powering Big Data implementations. In this paper, we cover some of the ways in which data security can be ensured while implementing Big Data solutions using Hadoop.

KEYWORDS: Hadoop; Big Data; enterprise; defense; risk; Security and Privacy

I. INTRODUCTION

The term "Big Data" refers to the large amounts of digital information that various orgamnisations collect. Industry estimates on the growth rate of data is roughly double every two years, from 2500 Exabytes in 2012 to 40,000 Exabytes in 2020. Big data is not a particular technology. It is a collection of attributes and capabilities. Big Data has been collected and utilized by big organisations for several decades. Software infrastructures such as Hadoop enable developers and analysts to easily leverage hundreds of computing nodes to perform data-parallel computing which was not there before.

Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. Hadoop allows running applications on systems with thousands of nodes with thousands of terabytes of data. Its distributed file system supports fast data transfer rates among nodes and allows the system to continue operating uninterrupted at times of node failure .Hadoop consists of distributed file system, data storage and analytics platforms and a layer that handles parallel computation, rate of flow (workflow) and configuration administration. HDFS runs across the nodes in a Hadoop cluster and together connects the file systems on many input and output data nodes to make them into one big file system . The present Hadoop ecosystem consists of the Hadoop kernel, Map-Reduce, the Hadoop distributed file system (HDFS) and a number of related components such as Apache Hive, HBase, Oozie, Pig and Zookeeper and these components are explained as below:

- *HDFS*: A highly faults tolerant distributed file system that is responsible for storing data on the clusters.
- *MapReduce:* A powerful parallel programming technique for distributed processing of vast amount of data on clusters.
- *HBase*: A column oriented distributed NoSQL database for random read/write access.
- *Pig*: A high level data programming language for analyzing data of Hadoop computation.
- *Hive*: A data warehousing application that provides a SQL like access and relational model.
- *Sqoop*: A project for transferring/importing data between relational databases and Hadoop.
- *Oozie*: An orchestration and workflow management for dependent Hadoop jobs.

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. 3, Issue 10, October 2015

The architectural diagram of Hadoop framework is described in figure 1, as below we can get the brief idea of the Hadoop components and how they are arranged in levels.

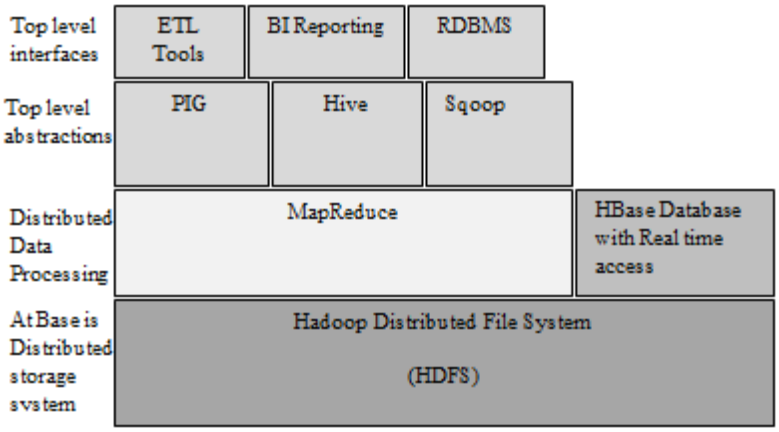

Fig.1 Hadoop Architecture

During the initial development period of Hadoop, security was not a prime focus area. In most of the cases, the Hadoop platform was being developed using data sets where security was not a prime concern because the data was publicly available. However, As Hadoop became a more popular platform for data analytics and processing, organizations are putting a lot of data from varied sources onto a Hadoop cluster, creating a possible data security situation also the security professionals began to express concerns about the insider threat of malicious users in a Hadoop cluster. A malicious developer could easily write code to impersonate other users' Hadoop services (e.g. writing a new TaskTracker and registering itself as a Hadoop service, or impersonating the hdfs or mapped users, deleting everything in HDFS, etc.). Because DataNodes enforced no access control, a malicious user could read arbitrary data blocks from DataNodes, bypassing access control restrictions, or writing garbage data to DataNodes, undermining the integrity of the data to be analyzed. Anyone could submit a job to a JobTracker and it could be arbitrarily executed.

II. HADOOP SECURITY ISSUES

Fragmented Data Sets: Big Data clusters contain data that allow multiple copies moving to-and-fro various nodes ensuring redundancy and resiliency. The data that is available for fragmentation and can be shared across multiple servers more complexity is added as a result of the fragmentation which poses a security issue due to the absence of a security model.

Distributed Computing: the data source is not fixed resources are processed where available, these lead to large levels of parallel computation. Complicated environments are created that are at high risks of attacks than their counterparts of repositories that are centrally managed and monolithic.

Controlling Data Access: big data only provides access control at schema level. There is no finer granularity in addressing proposed users in terms of roles and access related scenarios.

Node-to-node communication: Hadoop don't implement secure communication; they use the RPC (Remote Procedure Call) over TCP/IP.

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. 3, Issue 10, October 2015

Client Interaction: Communication of client takes place with resource manager, data nodes. Clients that have been compromised tend to propagate malicious data or links to either service.

Virtually no security: big data stacks where designed with no security in mind. There is no security for common web threats too.

III. RELATED WORK

In [2]author explain the brief overview of the security risks of Hadoop environment, its security requirements, and design considerations. Also gave data flow diagrams. He also elaborates the requirements and design of RPC, HDFS, MapReduce, delegation token, block access token, these all are used for security purpose of Hadoop framework. In [8] author gives the information about Apache sentry project. Apache sentry an open source project by Cloudera is an authorization module for Hadoop that offers the granular, role-based authorization required to provide precise levels of access to the right users and applications. It support for role-based authorization, fine-grained authorization, and multi-tenant administration .The Apache Knox Gateway is a system that provides a single point of authentication and access for various Hadoop services in a cluster. It provides a perimeter security solution for Hadoop. The second advantage is it supports various authentication and token verification scenarios. It manages security across multiple clusters and versions of Hadoop. It also provides SSO solutions, and allows integrating other identity management solutions such as LDAP, Active Directory (AD), and SAML based SSO and other SSO systems [6]. In [11] author gives the brief details of present work in the hadoop security work, Project Rhino provides an integrated end-to-end data security solution to the Hadoop ecosystem. It provides a token based authentication and SSO solution. It offers Hadoop crypto codec framework and crypto codec implementation to provide block level encryption for the data stored in 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 supports audit logging framework for easy audit trails.

IV. PROPOSED WORK

In proposed work we are trying to implement the agent based security module for Hadoop environment. We consider any large data set for processing it in Hadoop cluster. In our propose architecture we will perform data pre-processing , also perform load balancing of work by using the concept of master and slave to increase the performance within the cluster nodes. For data pre-processing we will use the combine approach i.e. we use Neive Bayes algorithm and support vector machine. And while transferring the data from master to slave we take help of agents to perform encryption and decryption of data for providing security to our data. For this we will use the HMAC-SHA2 algorithm.

V. CONCLUSION

As we know the big data is major constrain in new era of technology, and providing security to this big data is big issue. Every large organisation is maintaining its big data, and they are willing to use Hadoop for processing their data. In this paper we have tried to cover all the security solution to secure the Hadoop ecosystem. Also we suggest the agent based security model for Hadoop environment.

REFERENCES

1. Tom White O'Reilly |Yahoo! Press "Hadoop The definitive guide"4th edition,March 2015.
2. Owen O'Malley, Kan Zhang, Sanjay Radia, Ram Marti, and Christopher Harrell "Hadoop Security Design",Yahoo,2009.
3. Vormetric "Securing Big Data: Security Recommendations for Hadoop and NoSQL Environments, October 12, 2012".
4. Zettaset "The Big Data Security Gap: Protecting the Hadoop Cluster",April 2014.
5. Devaraj Das, Owen O'Malley, Sanjay Radia, and Kan Zhang "Adding Security to Apache Hadoop", Hortonworks, IBM .
6. Horton works "Technical Preview for Apache Knox Gateway", Hortonworks, November 2013.

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. 3, Issue 10, October 2015

7. Kevin T. Smith "Big Data Security: The Evolution of Hadoop's Security Model", infoq.com, August 2013.
8. M. Tim Jones "Hadoop Security and Sentry", IBM, January 2014.
9. Vinay Shukla's "Hadoop Security: Today and Tomorrow", Hortonworks, December 2013.
10. Sudheesh Narayana, Packt Publishing "Securing Hadoop- Implement robust end-to-end security for your Hadoop ecosystem".

BIOGRAPHY

Prachi Ramesh Gawali is a student of Mtech 2nd year in the Computer Science and Engineering Department, Tulsiramji Gaikwad-Patil College of Engineering and Technology, Rashtrasant Tukdoji Maharaj Nagpur University, Nagpur. Doing her Mtech Project under the guidance of Prof. Roshani Talmale, HOD of Computer Science and Engineering, and Co-guidance Prof. Rajesh Babu, of Tulsiramji Gaikwad-Patil College of Engineering and Technology.