Available online at www.ijiere.com
International Journal of Innovative and Emerging Research in Engineering
e-ISSN: 2394 - 3343 p-ISSN: 2394 - 5494

Security Issues in Big Data: In Context with Hadoop

Ms. Singh Arpita Jitendrakumar
Student of Third Year of Computer Science & Engineering, 2014-2015
Shri Sant Gadge Baba College of Engineering and Technology, Bhusawal.
North Maharashtra University, Jalgaon, Maharashtra – 425203, India

ABSTRACT:
Big data came into existence when the traditional relational database systems were not able to handle the unstructured data (weblogs, videos, photos, social updates, human behavior) generated today by organization, social media, or from any other data generating source. Data is increasing in size day by day and Hadoop is used to process such large amount of data. In is paper, I made a study of various security issues associated with big data in context with the Hadoop environment and the various solution techniques and technologies involve in securing the big data Hadoop.

Keywords: Big Data, SASL, delegation, cell level, variety, unauthorized.

I. INTRODUCTION

Big data means data which is large in size, volume, variety. Nowadays the size a data is increasing rapidly, use of social media, Smartphone's, online shopping's etc. The volumes of Big data are on a roll, which can be inferred from the fact that as far back in the year 2012, there were a few dozen terabytes of data in a single dataset, which has interestingly been catapulted to many petabytes today. Such large amount of data is used for commercial purpose by enterprise to increase their business profit and many other applications, and therefore there is a need to secure such large amount of data and its processing. Big data has the following characteristics:

- Volume: In Big data the word big it-self define the size the data. Volume is associated with the size of big data. At present the data is supposed to be petabytes with could increase to zettabytes in near future.
- Velocity: Velocity in Big data deals with the speed of the data coming from various sources. Velocity characteristic is not limited to the speed of incoming data but also speed at which the data flows and aggregated.
- Variety: Data variety is a measure of the richness of the data representation – text, images video, audio, etc. the data processed is not of a single type it consists of semi structured data and unstructured data.
- Value: Data value measures the usefulness of data for making decisions. The data science is useful in getting to know the data, but "analytic science" encompasses the predictive power of big data. Various users can run certain queries against the data stored and thus can deduct important results from the filtered data obtained and also rank it according to the dimensions they require. These reports help people to find the business trends according to which they can make change in their strategies.
- Complexity: Complexity measures the degree of a interdependence in big data structures such that a small change (or combination of small changes) in one or a few elements can yield very large changes or a small change that ripple across or cascade through the system and substantially affect its behavior, or no change at all (Katal, Wazid, & Goudar, 2013) and interconnectedness (possibly very large).

II. BIG DATA PRESENTS A NEW SECURITY CHALLENGE

Big data originates from multiple sources including sensors used to gather digital pictures and videos, climate information, posts to social media sites, purchase transaction records, and cell phone GPS signals. With cloud computing and the socialization of the Internet, unstructured data petabytes of are created online daily and much of this information has an intrinsic business value if it can be captured and analyzed.

Examples, mobile communications companies collect data from cell towers; oil and gas companies collect data from refinery sensors and seismic exploration; electric power utilities collect data from power plants and distribution systems. Businesses collect large amounts of user-generated data from prospects and customers including social security numbers, credit card numbers, and data on patterns of usage and buying habits.

The influx of big data and the need to move this information throughout an organization has created a massive new target for hackers and other cybercriminals. This data, which previously was unusable by various organizations is now highly valuable and is subject to privacy laws and compliance regulations, and must always be protected.

III. PROCESSING BIG DATA

Hadoop allows running applications on systems with thousands of nodes with thousands of terabytes of data [2]. The distributed file system supports fast data transfer rates among nodes allowing the system to continue operating uninterrupted at times of node failure. Hadoop has of distributed file system, analytics platforms and data storage and a layer handling parallel computation, rate of flow (workflow) and configuration administration [8].the HDFS runs across the nodes in a Hadoop cluster with together connects the file systems on many input and output data nodes and make one big file system [2]. Hadoop ecosystem have Hadoop kernel, Map- Reduce, the Hadoop distributed file system (HDFS) and a number of related components such as Apache, Oozie, Hive, HBase ,Pig and Zookeeper and these components that are explained as below[7,8]:

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

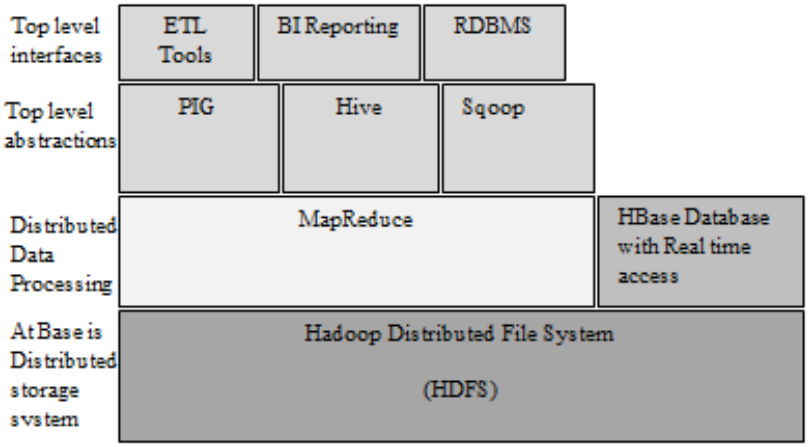

Figure 1: Hadoop architecture

IV. BIG DATA HADOOP'S TRADITIONAL SECURITY

A. OVERVIEW

Originally Hadoop was developed without any security in mind, no security model, no authentication of users and services and no data privacy, so anybody could submit arbitrary code to be executed. Auditing and authorization controls (HDFS file permissions and ACLs) used during earlier distributions were easily evade because any user could impersonate other user. So various security controls measures that did subsist were not very effective. Later authorization and authentication was added, but had some weakness. All programmers users and had the same level of access privileges to all the data in the cluster, any one could access any of the data in the cluster, and any user could read any data set [4]. MapReduce had no concept of authentication or authorization, user could lower the priorities of other Hadoop jobs to make his job complete faster or to be executed first – or worse, he could kill the other jobs. Supplementary security cannot keep up. The Hadoop supports some security features with current Kerberos implementation, firewalls, and HDFS permissions and ACLs [5].

B. THREATS TO SECURITY

The related threats associated with processing data in Hadoop ecosystem are as follows:

1. Unauthorized client: An unauthorized client may write/read a data block of a file at a Data Node using the pipeline streaming Data-transfer protocol. And if gained access privileges and can submit a job to a queue or delete or change priority of the job. And can access intermediate data of Map job via its task trackers HTTP shuffle protocol.

2. Task: A task in execution may make use of host OS interfaces to access other tasks, or would access local data which include intermediate Map output or the local storage of the Data Node that runs on the same physical node. Similarly, A task or node may masquerade as a Hadoop service component such as a Name Node, job tracker, Data Node, task tracker etc.

3. Unauthorized user: An unauthorized user could execute arbitrary code or carry out further attacks by accessing an HDFS file via the RPC or via HTTP protocols. Similarly, he may sniff/ eavesdrop to data packets being sent by Data nodes to client and can submit a workflow to Oozie as another user. Data Nodes does not impose access control, he could bypass the various access control mechanism or restrictions and read arbitrary data blocks from Data Nodes or writing garbage data to Data Node to corrupt it. [10]

V. SECURITY ISSUES

Hadoop present security issues for data centre managers and security professionals. The security issues are as below [5, 6, 18]:

1. Fragmented Data: 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.
2. 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.
3. 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.
4. Node-to-node communication: Hadoop don't implement secure communication; they use the RPC (Remote Procedure Call) over TCP/IP.
5. 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.
6. Virtually no security: big data stacks where designed with no security in mind. There is no security for common web threats too.

VI. SOLUTION FOR BIG DATA SECURITY IN HADOOP

Analyzing the security issues associated with big data Hadoop. Here in this paper I have mentioned the solutions that help in ensuring the security of data. [18]

A. AUTHENTICATION

Hadoop have Kerberos as a primary authentication. Initially SASL/GSSAPI was used for implementing Kerberos and mutually authenticates users, their applications, and other Hadoop services over the RPC connections [7]. Hadoop supports "Pluggable" Authentication for HTTP Web Consoles, meaning implementers of web applications and web consoles can implement their own authentication mechanism for HTTP connections. HDFS communications that is between the Name Node and Data Nodes is over RPC connection and mutual Kerberos authentication is performed between them [15]. HBase supports SASL Kerberos secure client authentication via RPC, HTTP. Delegation token which is a two party authentication protocol used between user and the Name Node for authenticating users, it is very simple and more effective than three party protocol used by Kerberos [7, 15].Oozie and HDFS, MapReduce supports delegation token.

B. ACLs AND AUTHORIZATION

In Hadoop, the access controls is implemented using file-based permissions following the UNIX permissions model. In HDFS, Access control to files could be enforced by the Name Node through file permissions and ACLs of users and groups. MapReduce gives ACLs for job queues; defining which users or groups can submit jobs to a queue or change queue properties. Hadoop gives fine-grained authorization using file permissions in HDFS and resource level access control using ACLs for MapReduce and coarser grained access control at a service level [13].

C. ENCRYPTION

The data needs protection during the transfer to and from the Hadoop system. The simple authentication and security layer (SASL) authentication framework is used to encrypt the data in motion in Hadoop ecosystem. SASL security provide guarantee of the data being exchanged between client and servers and ensures that, the data is not readable by "man-in middle". [15]. Hadoop also supports encryption capability to various channels like RPC, HTTP, and Data Transfer Protocol for data in motion.

D. AUDIT TRAILS

To meet the security compliance requirements, auditing the entire Hadoop ecosystem on a periodic basis and deploy or implement a system that does log monitoring is necessary. HDFS and MapReduce have base audit support. Apache Hive megastore does maintain audit information for Hive interactions [13, 15]. Apache Oozie, the workflow engine, provides audit trail for services, Hue Supports audit logs. For that Hadoop component which does not have built-in audit logging, audit logs monitoring tools can be used. [6][15]

VII. ZETTASET ORCHESTRATOR

A. PERIMETER SECURITY FAILS

Vendors of data security think that traditional parameter security solutions such as intrusion detection/prevention technologies and firewalls can properly address the Hadoop and distributed cluster security. But all security solutions that rely on perimeter security fail to provide effective security to the Hadoop cluster. Firewalls that attempt to map IP to actual AD credentials are problematic in Hadoop environment. Firewalls only restrict access on basis of IP/ports and do not know anything about the Hadoop file system. [6]

B. MOVE SECURITY CLOSER TO THE DATA

Zettaset Orchestrator gives a security solution for big data embedded in the data cluster itself which moves security as close to the data as possible, and protection that perimeter security devices such as firewalls fails to deliver. Orchestrator address the security gaps that open-source solutions ignore, with big data management solution that is hardened to address policy, access control, compliance, and risk management in the Hadoop cluster environment. Orchestrator takes into account RBAC that strengthens user authentication process. Orchestrator makes simple the integration of Hadoop clusters into an existing security policy framework, and support for AD, LDAP. For organizations with compliance reporting requirements, Orchestrator provides extensive logging, search, and auditing capabilities.

Zettaset Orchestrator solution is specifically designed to meet the security requirements of the distributed architectures that predominate in big data and Hadoop environments. Orchestrator creates security wrapper around Hadoop distribution and distributed computing environment which make it enterprise-ready. With Orchestrator, organizations deploy Hadoop in data center environments [6]

VIII. CONCLUSION

Today the size of data is increasing rapidly, in this generation of big data where source of data is not fixed there is a need to secure data coming from various sources. As Hadoop is used to process such data, in this paper I have made a study of various security issues associated with big data in Hadoop environment and the possible solutions implemented. I think that if we focus on the data that are stored and processed, so that personal privacy is not lost, security can effectively work.

ACKNOWLEDGMENT

I feel great pleasure in submitting this Paper on “Security Issues in Big Data: In Context with Hadoop”. I wish to express true sense of gratitude towards my Principal Dr. R. P. Singh. And a special thanks to my H.O.D and Guide Prof. D. D. Patil who at very discrete step in preparation of this Paper contributed his valuable guidance and help to solve every problem that arose. Also, most likely I would like to express my sincere gratitude towards my parents for always being there when I needed them the most. With all respect and gratitude, I would like to thank all the people, who have helped me directly or indirectly. I owe my all success to them.

REFERENCES

[1] Cloud Security Alliance “Top Ten big Data Security and Privacy Challenges”.
[2] Tom White O’Reilly |Yahoo! Press “Hadoop The definitive guide”.
[3] Owen O’Malley, Kan Zhang, Sanjay Radia, Ram Marti, and Christopher Harrell “Hadoop Security Design”.
[4] Mike Ferguson “Enterprise Information Protection - The Impact of Big Data”.
[5] Volumetric “Securing Big Data: Security Recommendations for Hadoop and NoSQL Environments”, October 12, 2012.
[6] Zettaset “The Big Data Security Gap: Protecting the Hadoop Cluster”.
[7] Devaraj Das, Owen O’Malley, Sanjay Radia, and Kan Zhang “Adding Security to Apache Hadoop”.
[8] Seref SAGIROGLU and Duygu SINANC “Big Data: A Review Collaboration Technologies and Systems (CTS)”, 2013 International Conference, May 2013.
[9] Horton works “Technical Preview for Apache Knox Gateway”.
[10] Kevin T. Smith “Big Data Security: The Evolution of Hadoop’s Security Model”.
[11] M. Tim Jones “Hadoop Security and Sentry”.
[12] Victor L. Voydock and Stephen T. Kent “Security mechanisms in high-level network protocols”. ACM Comput. Surv.1983.
[13] Vinay Shukla s “Hadoop Security: Today and Tomorrow”.
[14] MahadevSatyanarayanan “Integrating security in a large distributed system”. ACM Trans. Comput. Syst. 1989.
[15] Sudheesh Narayana, Packt Publishing “Securing Hadoop- Implement robust end-to-end security for your Hadoop ecosystem”.
[16] S. Singh and N. Singh, "Big Data Analytics", 2012 International Conference on Communication, Information & Computing Technology Mumbai India, IEEE, October 2011.
[17] Jeffhurtblog.com “three-vs.-of-big-data-as-applied-conferences”, July 7, 2012.
[18] 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.