Impact Factor (SJIF): 3.632

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-244

Volume 3, Issue 7, July-2016

Security Threat Model and Security Framework for Big Data Security

Bavya S[1], Manasa N[2], B N Veerappa[3]

[1]*Department of Studies in Computer science and Engineering, UBDTCE, Davangere*

[2]*Department of Studies in Computer science and Engineering, UBDTCE, Davangere*

[3]*Department of Studies in Computer science and Engineering, UBDTCE, Davangere*

Abstract- *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 acquisition and high volume create unique security risks. In this paper, we are going to discuss particularly security challenges when organisations started to move sensitive data to a Big Data repository like hadoop. It identifies different security threat models and security frameworks. Its main goal is to propose a hadoop system which can face the security issues by using suitable security mechanisms.*

Keywords—Hadoop,security threat models, security framework, Privacy & Security,Security issues.

I. INTRODUCTION

The Big Data means massive amount of crucial data that companies collect. Recent technological advances and novel applications, such as sensors, cyber-physical systems, smart mobile devices, cloud systems, data analytics, and social networks, are making possible to capture, process, and share huge amounts of data –referred to as big data - and to extract useful knowledge, such as patterns, from this data and predict trends and events. Big Data may contain more than peta bytes of data. For every two years growth rate of data doubles. Big data is making possible tasks that before were impossible, like preventing disease spreading and crime, personalizing healthcare, quickly identifying business opportunities, managing emergencies, protecting the homeland, and so on.

NIST defines Big Data as the following:

Big Data consists of extensive datasets, primarily in the characteristics of volume, velocity, and/or variety that require a scalable architecture for efficient storage, manipulation, and analysis.

Big data is relavant for all components of our society. Industry is using big data for shifting business intelligence from reporting and decision support to prediction and next move decisions. This use of big data emphasizes that big data is critical for obtaining actionable knowledge. Governments are also interested in using big data and predictive analytics to improve decision making and transparency, to engage citizens in public affairs, to improve national security. Healthcare Represents another major area to which big data may offer novel opportunities. Learning health systems are currently focusing on turning health care data into knowledge, translating that knowledge into practice, and creating new data by means of advanced information technology.

Characteristics of Big Data are:

1. It handles a petabyte of data or more.
2. It has distributed redundant data storage.
3. It does parallel task processing.
4. Can provide data processing capabilities.
5. It has extremely fast data insertion.
6. Has central management and orchestration.

This paper details the security challenges when organisation start moving sensitive data to a Big Data repository like hadoop. It provides the different threat models and the security control framework to address and mitigate the risks due to the identified security threats. In the following sections, the paper describes the architecture of hadoop eco system and weakness of such system. Then we can identify the suitable security mechanisms.

II. RELATED WORK

Ajit Gaddam et.al [2] details the security challenges when organizations start moving sensitive data to a Big Data repository like Hadoop. It identifies the different threat models and the security control framework to address and mitigate security risks due to the identified threat conditions and usage models. The framework outlined in this paper is also meant to be distribution agnostic.

Priya P. Sharma et.al [3] Hadoop projects treat Security as a top agenda item which in turn represents which is again classified as a critical item. Be it financial applications that are deemed sensitive, to healthcare initiatives, Hadoop is traversing new territories which demand security-subtle environments. With the growing acceptance of Hadoop, there is an increasing trend to incorporate more and more enterprise security features. In due course of time, we have seen Hadoop gradually develop to label important issues pertaining to, what we summarize as 3ADE (authentication, authorization, auditing, and encryption) within a cluster. There is no dearth of Production environments that support Hadoop Clusters.In this paper, we aim at studying "Big Data" security at the environmental level, along with the probing of built-in protections and the Achilles heel of these systems, and also embarking on a journey to assess a few issues that we are dealing with today in procuring contemporary Big Data and proceeds to propose security solutions and commercially accessible techniques to address the same.

Vinod Sharma [4] introduces the big data technology along with its importance in the modern world and existing projects like hadoop which are effective and important in changing the concept of science into big science. Hadoop, Map Reduce and No SQL are the major big data technology. This paper also throws some light on other challenges and issues. The various challenges and issues in adapting and accepting Big data security and suggest some more security standards and concept that make robust hadoop ecosystem without any processing overhead.

III. COMPONENTS

3.1 Hadoop Security weakness:

Traditional Relational Database Management Systems (RDBMS) security has evolved over the years and with many 'eyeballs' assessing the security through various security evaluations. Unlike such solutions, Hadoop security has not undergone the same level of rigor or evaluation for that matter and thus can claim little assurance of the implemented security. Another big challenge is that today, there is no standardization or portability of security controls between the different Open-Source Software (OSS) projects and the different Hadoop or Big Data vendors. Hadoop security is completely fragmented. This is true even when the above parties implement the same security feature for the same Hadoop component. Vendors and OSS parties' force-fit security into the Apache Hadoop framework.

Top 10 Security & Privacy Challenges

The Cloud Security Alliance Big Data Security Working Group has compiled the following as the Top 10 security and privacy challenges to overcome in Big Data [4].

1. Secure computations in distributed programming frameworks
2. Security best practices for non-relational data stores
3. Secure data storage and transactions logs
4. End-point input validation/filtering
5. Real-time security monitoring
6. Scalable privacy-preserving data mining and analytics
7. Cryptographically enforced data centric security

8. Granular access control

9. Granular audits

10. Data provenance

3.2 Proposed System:

The following section provides the target security architecture framework for Big Data platform security. The core components of the proposed Big Data Security Framework are the following: 1. Data Management 2. Identity & Access Management 3. Data Protection & Privacy 4. Network Security 5. Infrastructure Security & Integrity

The above '5 pillars' of Big Data Security Framework are further decomposed into 21 sub-components, each of which are critical to ensuring the security and mitigating the security risk and threat vectors to the Big Data stack.

3.3 Mechanisms:

We have to follow different mechanisms to improve security threat models and security framework. They are

3.3.1.Data Management

Data Management component is decomposed into three core sub-components. They are Data Classification, Data Discovery, and Data Tagging.

3.3.2.Identity & Access Management

POSIX-style permissions in secure HDFS are the basis for many access controls across the Hadoop stack. User Entitlement + Data Metering Provide users access to data by centrally managing access policies. It is important to tie policy to data and not to the access method. RBAC Authorization Deliver fine-grained authorization through Role Based Access Control (RBAC). Manage data access by role (and not user). Determine relationships between users & roles through groups. Leverage AD/LDAP group membership and enforce rules across all data access paths

3.3.3.Data Protection & Privacy

The majority of the Hadoop distributions and vendor add-ons package either data-at-rest encryption at a block or (whole) file level. Application level cryptographic protection (like field-level/column-level encryption, data tokenization, and data redaction/masking provide the next level of security needed.

Application Level Cryptography (Tokenization, field-level encryption) While encryption at the field/element level can offer security granularity and audit tracking capabilities, it comes at the expense of requiring manual intervention to determine the fields that require encryption and where and how to enable authorized decryption.

Transparent Encryption (disk / HDFS layer) Full Disk Encryption (FDE) prevents access via the storage medium. File encryption can also guard against (privileged) access at the node's operating-system level.

Data Masking/ Data Redaction Data masking or data redaction before load in the typical ETL process de-identifies personally identifiable information (PII) data before load. Therefore, no sensitive data is stored in Hadoop, keeping the Hadoop Cluster potentially out of (audit) scope

3.3.4.Network Security

The Network Security layer is decomposed into four sub-components. They are data protection in-transit and network zoning + authorization components.

Data Protection In-Transit Secure communications are required for HDFS to protect data-in-transit. There are multiple threat scenarios that in turn mandate the necessity for https and prevent information disclosure or elevation of privilege threat categories. Using the TLS protocol (which is now available in all Hadoop distributions) to authenticate and ensure privacy of communications between nodes, name servers, and applications. An attacker can gain unauthorized access to data by intercepting communications to Hadoop consoles. This could include communication between NameNodes and DataNodes that are in the clear back to the Hadoop clients and in turn can result in credentials/data to be sniffed. Tokens that are granted to the user post- Kerberos authentication can also be sniffed and can be used to impersonate users on the NameNode.

Following are the controls that when implemented in a Big Data cluster can ensure properties of data confidentiality.

1. Packet level encryption using TLS from the client to Hadoop cluster 2. Packet level encryption using TLS within the cluster itself. This includes using https between NameMode to Job Tracker to DataNode. 3. Packet level encryption using TLS in the cluster (e.g. mapper-reducer) 4. Use LDAP over SSL (LDAPS) rather than LDAP when communicating with the corporate enterprise directories to prevent sniffing attacks. 5. Allow your admins to configure and enable encrypted shuffle and TLS/https for HDFS, MapReduce, YARN, HBase UIs etc.

Network Security Zoning The Hadoop clusters must be segmented into points of delivery (PODs) with chokepoints such as Top of Rack (ToR) switches where network Access Control Lists (ACLs) limit the allowed traffic to approved levels.

End users must not be able to connect to the individual data nodes, but to the name nodes only. The Apache Knox gateway for example, provides the capability to control traffic in and out of Hadoop at the per-service-level granularity. A basic firewall that should allow access only to the Hadoop NameNode, or, where sufficient, to an Apache Knox gateway. Clients will never need to communicate directly with, for example, a DataNode.

3.3.5.Infrastructure Security & Integrity

The Infrastructure Security & Integrity layer is decomposed into four core sub-components. They are Logging/Audit, Secure Enhanced Linux (SELinux), File Integrity + Data Tamper Monitoring, and Privileged User and Activity Monitoring.

Logging / Audit All system/ecosystem changes unique to Hadoop clusters need to be audited with the audit logs being protected. Examples include: Addition/deletion of data and management nodes Changes in management node states including job tracker nodes, name nodes

Pre-shared secrets or certificates that are rolled out when the initial package of the Hadoop distribution or of the security solution is pushed to the node prevent the addition of unauthorized cluster nodes.

When data is not limited to one of the core Hadoop components, Hadoop data security ends up having many moving parts and high percentage of fragmentation. Consequently, there results a sprawl of metadata and audit logs across all fragments.

In a typical enterprise, the DBAs are typically leveraged to place the security responsibility at the table, row, column, or cell level and while the configuration of file systems and system administrators, and the Security Access Control team is usually accountable for the more granular file level permissions.

Yet, in Hadoop, POSIX-style HDFS permissions are frequently important for data security or are at times the only means to enforce data security at all. This leads to questions concerning the manageability of Hadoop security. Technologies recommendations to address data fragmentation: Apache Falcon is an incubating Apache OSS project that focuses on data management. It provides graphical data lineage and actively controls the data life cycle. Metadata is retrieved and mashed up from wherever the Hadoop application stores it. Cloudera Navigator is a proprietary tool and GUI that is part of Cloudera's Distribution Including Apache Hadoop (CDH) distribution. CDH Navigator is a tool to address log sprawl, lineage and some aspects of data discovery. Metadata is retrieved and mashed up from wherever the Hadoop application stores it. Zettaset Orchestrator is a product for harnessing the overall fragmentation of Hadoop security with a proprietary combined GUI and workflow. Zettaset has its own metadata repository where metadata from all Hadoop components is collected and stored.

Secure Enhanced Linux (SELinux) SELinux was created by the United States National Security Agency (NSA) as a set of patches to the Linux Kernel using Linux Security Modules (LSM). It was eventually released by the NSA under the GPL license and has been adopted by the upstream Linux kernel.

SELinux is an example of a Mandatory Access Control (MAC) for Linux. Historically Hadoop and other Big Data platforms built on top of Linux and UNIX systems have had discretionary access control. What this means for example is that a privileged user like root is omnipotent. By enforcing and configuring SELinux on your Big Data environment, through MAC, there is policy which is administratively set and fixed. Even if a user changes any settings on their home directory, the policy prevents another user or process from accessing it. A sample policy for example that can be implemented is to make library files executable but not writable or vice-versa. Jobs can write to /tmp location but not be able to execute anything in there. This is a great way to prevent command injection attacks among others. With policies configured, even if someone who is a sysadmin or a malicious user is able to gain access to root using SSH or some other attack vector, they may be able to read and write a lot of stuff. However, they won't be able to execute anything incl. potentially any data exfiltration methods.

The general recommendation is to run SELinux is permissive mode with regular workloads on your cluster, reflecting typical usage, including using any tools. The warnings generated can then be used to define the SELinux policy which after tuning can be deployed in a 'targeted enforcement' mode.

IV. CONCLUSION

Hadoop and big data are no longer buzz words in large enterprises. Whether for the correct reasons or not, enterprise data warehouses are moving to Hadoop and along with it come petabytes of data. In this paper we have laid the groundwork for conducting future security assessments on the Big Data ecosystem and securing it. This is to ensure that Big Data in Hadoop does not become a big problem or a big target. Vendors pitch their technologies as the magical silver bullet. However, there are many challenges when it comes to deploying security controls in your Big Data environment. This paper also provides the Big Data threat model which the reader can further expand and customize to their organizational environment. It also provides a target reference architecture around Big Data security and covers the entire control stack. Hadoop and big data represent a green field opportunity for security practitioners. It provides a chance to get ahead of the curve, test and deploy your tools, processes, patterns, and techniques before big data becomes a big problem.

REFERENCES

[1] EMC Big Data 2020 Projects http://www.emc.com/leadership/digital- universe/iview/big-data-2020.htm l

[2] NIST Special Publication 1500-1 NIST Big Data Interoperability Framework: Volume 1, Definitions http://bigdatawg.nist.gov/_uploadfiles/M0392_v1 _3022325181.pdf

[3] Securosis – Securing Big Data Security issues with Hadoop environments https://securosis.com/blog/securing-big-data- security-issues-with-hadoop-environments

[4] Top 10 Big Data Security and Privacy Challenges, Cloud Security Alliance, 2012 https://downloads.cloudsecurityalliance.org/initi atives/bdwg/Big_Data_Top_Ten_v1.pdf

[5] Hadoop in Action, Second Edition by Manning Publications. ISBN: 9781617291227 h t t p : / / w w w . m a n n i n g . c o m / l a m 2 /