www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 6 June 2015, Page No. 12610-12613

Securing Hadoop Using Layered Approach

Nawghare Pushpmala S.

Savitribai Phule Pune University, Pune, Maharashtra
Pushpa.nawghare@email.com

Abstract: ***Hadoop is a distributed system that provides a distributed filesystem and MapReduce batch job processing on large clusters using commodity servers. Although Hadoop is used on private clusters behind an organization's firewalls, Hadoop is often provided as a shared multi-tenant service and is used to store sensitive data; as a result, strong authentication and authorization is necessary to protect private data.The biggest challenge for big data from a security point of view is the protection of user's privacy. Hadoop Big Data frequently contains huge amounts of personal identifiable information and therefore privacy of users is a huge concern. Security and privacy issues are magnified by velocity, volume, and variety of big data, such as large-scale cloud infrastructures, diversity of data sources and formats, streaming nature of data acquisition, and high volume inter cloud migration. Therefore, traditional security mechanisms, which are tailored to securing small scale static (as opposed to streaming) data, are inadequate. As an increasing number of enterprises move towards production deployments of Hadoop, security continues to be an important topic. In this paper I am adding how Layered Approach: Secure Protocol complies with current and future security implementation standards providing authentication and authorization and integrating additional levels such as data encryption support.***

Keywords: Big Data, Hadoop, Security, Secure Protocol.

1. Introduction

One of the biggest concerns in our present age revolves around the security and protection of sensitive information. In our current era of Big Data, our organizations are collecting, analyzing, and making decisions based on analysis of massive amounts of data sets from various sources, and security in this process is becoming increasingly more important. At the same time, more and more organizations are being required to enforce access control and privacy restrictions on these data sets to meet regulatory requirements such as privacy protection laws. Network security breaches from internal and external attackers are on the rise, often taking months to be detected, and those affected are paying the price. Organizations that have not properly controlled access to their data sets are facing lawsuits, negative publicity, and regulatory fines. Hadoop is one of the main technologies powering Big Data implementations. This Paper, cover some of the ways in which data security can be ensured while implementing Big Data solutions using Hadoop.

1.1 Evolution of Hadoop security

During the initial development 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 has become mainstream, organizations are putting a lot of data from varied sources onto a Hadoop cluster, creating a possible data security situation. The Hadoop community has realized that more robust security controls are needed and has decided to focus on the security aspect and new security features are being developed. Hadoop is a distributed system for storing large amounts of data and processing the data in parallel. This paper describes the security issues that

arise in Hadoop and how we address them. Hadoop is a complex distributed system that poses a unique set of challenges for adding security.

2. Related Study

Consider the following eye-opening statistics: A study released this year by Symantec and the Ponemon Institute found that the average organizational cost of *one* security breach in the United States is 5.4 million dollars. Another recent study shows that the cost of cybercrime in the U.S. economy alone is 140 billion dollars per year.

- One of the largest breaches in recent history involved Sony's Playstation Network in 2011, and experts estimate Sony's costs related to the breach to be somewhere between 2.7 and 24 billion dollars (a wide range, but the breach was so large, it is almost impossible to quantify).
- Netflix and AOL have already faced (and in some cases, settled) millions of dollars in lawsuits over their management of large sets of data and their protection of personal information – even data that they had "anonymized" and released for research.
- Beyond quantifiable costs related to security breaches (loss of customers and business partners, lawsuits, regulatory fines), organizations that have experienced such incidents report that the fallout from a data breach results in a diminished trust of the organization and a damaged reputation that could put a company out of business

These days, we see widespread adoption of Hadoop. Hadoop has grown beyond a series of open source projects for programmers, and, now, organizations have matured in their understanding of Big Data technologies and their expectations on the benefits of Hadoop. Acknowledging the added value that can be generated by applying analytics on Big Data in Hadoop in a cost-effective way, many organizations have successfully passed the proof of concept stage and moved on to setting up production clusters. With that, new aspects of deploying Hadoop gain the focus Among these aspects, data security is the one we see coming up most often. Though requirements differ depending on the type of organization and level of

regulations typically applied within an industry sector, most organizations actively consider and implement security as an integral part of a productive Hadoop environment. The challenge is to deploy solutions that bring analytics to Hadoop while seamlessly integrating with data security policies and platforms that make security transparent and easily applicable for users in order to facilitate frictionless building of modern analytics.

2.1 Hadoop Background

Hadoop has been under development at Yahoo! And a few other organization as an Apache open source project over the last 5 years. It is gaining wide use in the industry. Yahoo!, for example, has deployed tens of Hadoop clusters, each typically with 4,000 nodes and 15 petabytes. Hadoop contains two main components. The first component, HDFS, is a distributed file system similar to GFS. HDFS contains a metadata server called the NameNode that stores the hierarchical file and directory name space and the corresponding metadata, and a set of DataNodes that stores the individual blocks of each files. Each block, identified by a block id, is replicated at multiple DataNodes. Client perform file metadata operations such as create file and open file, at the NameNode over an RPC protocol and read/write the data of a file directly to DataNodes using a streaming socket protocol called the data-transfer protocol.

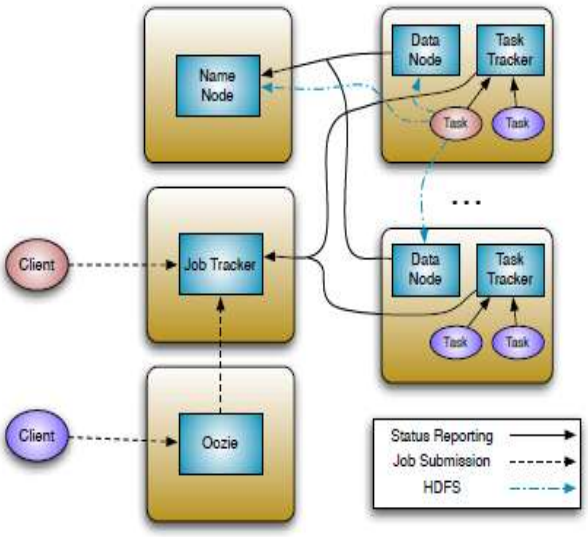

Figure 1: Hadoop High Level Architecture

The second component is a framework for processing large amounts of data in parallel using the MapReduce paradigm . HDFS DataNodes also serve as compute nodes for MapReduce to allow computation to be performed close to the data being processed. A user submits a MapReduce job to the JobTracker which schedules the job to be executed on the compute nodes. Each compute node has a small daemon called the TaskTracker that launches map and reduce tasks of a job; the task tracker also serves intermediate data created by map tasks to reduce tasks. There are additional services deployed with Hadoop, one of which are relevant to the security concerns discussed in this paper. Ooozie is a workflow system that provides a way to schedule and submit a DAG of MapReduce jobs that are triggered for execution by data availability or time.

2.2 Challenges in Adding Security to Hadoop

There are number of security challenges for organizations securing Hadoop. Big data originates from multiple sources including sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals, to name a few. Thanks to cloud computing and the socialization of the Internet, peta bytes of unstructured data are created daily online and much of this information has an intrinsic business value if it can be captured and analyzed. For example, 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 credit card numbers, social security numbers, data on buying habits and patterns of usage. 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 was previously unusable by organizations is now highly valuable, is subject to privacy laws and compliance regulations, and must be protected.

2.3 Hadoop's Architecture Presents Unique Security Issues

Big data is distinguished by its fundamentally different deployment model: highly distributed, redundant, and elastic data repositories enabled by the Hadoop File System. Rather than being a siloed, centralized data repository, such as a solitary Oracle Database, a Hadoop cluster may consist of anywhere from tens to thousands of nodes. This group of machines works in tandem to appear as a single entity, much like a mainframe, but with much lower capital expense and operating cost. But the characteristics of Hadoop's distributed computing architecture present a unique set of challenges for datacenter managers and security professionals.

- Distributed computing - Data is processed anywhere resources are available, enabling massively parallel computation. This creates complicated environments that are highly vulnerable to attack, as opposed to the centralized repositories that are monolithic and easier to secure.
- Fragmented data - Data within big data clusters is fluid, with multiple copies moving to and from different nodes to ensure redundancy and resiliency. Data can become sliced into fragments that are shared across multiple servers. This fragmentation adds more complexity to the security challenge.
- Access to data - Role-Based Access Control (RBAC) is central to most database security frameworks, but most big data environments only offer access control at the schema level, with no finer granularity to address users by role and related access.
- Node-to-node communication - Hadoop and the vast majority of distributions don't communicate securely; they use RPC over TCP/IP.
- Virtually no security - Big data stacks build in almost no security. Aside from service-level authorization and web proxy capabilities from YARN, no facilities are available to protect data stores, applications, or core Hadoop features. All big data installations are built on the web services model, with few or no facilities for countering common web threats.

3. Security for Hadoop

For implementing Hadoop security, there is a common understanding of the respective measures to be implemented

among leading Hadoop vendors. All Hadoop distribution providers promote a 4-layer security model for Hadoop as shown in Figure 2. Sometimes they use different names for the security layers but the underlying concepts are typically similar.

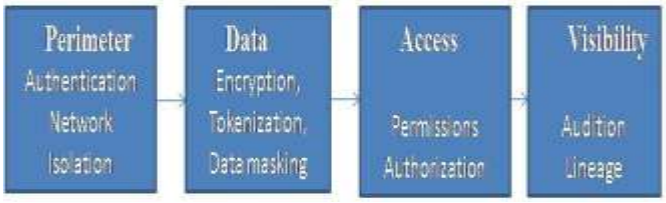

Figure 2: Data Security Implementation Model

- **Guard the Perimeter**

One of the benefits of Hadoop is that it easily allows for multiple entry points both for data flow and user access. This means to secure cluster access, you can't simply put up a firewall around your cluster with a single access gate. Cloudera provides comprehensive perimeter security that preserves the agility of multiple entry points while providing strong authentication that's easy to manage. Manager makes it easy to secure your cluster using industry standard Kerberos, LDAP/AD, and SAML [1].

- **Protect Data-At-Rest and Data In Motion**

Through our acquisition of Gazzang, Cloudera is able to protect data-at-rest and data in motion through encryption and powerful key management - all integrated into Cloudera Navigator. Navigator Encrypt provides massively scalable, high performance encryption for critical Hadoop data. Navigator Key Trustee is a "virtual safe-deposit box" for managing encryption keys, certificates, and passwords. Both are completely integrated into Navigator and are available today.

- **Control Access with Project Rhino and Sentry**

Project Rhino is an open source initiative dedicated to enhancing security in Hadoop. In collaboration with Intel and the community, As part of the initiative, Apache Sentry as a common authorization framework. Sentry integrates with the Hadoop platform and allows you to store sensitive data while meeting regulatory requirements. Sentry allows for fine-grained authorization and role-based access controls across the Hadoop platform, all through a single, unified system. Multi-tenant administration also allows you to extend Hadoop to more users, securely.

- **Gain Visibility**

Navigator is the only native end-to-end governance solution for Hadoop. It provides full visibility into where data came from and how it's being used to verify authenticity and easily comply with regulatory requirements. Key features of Navigator include: comprehensive auditing, fine-grained access controls, discovery and exploration, and data lineage. Navigator also includes Navigator Encrypt and Navigator Key Trustee for high-performance data-at-rest encryption and enterprise key management.

3.1 Big Data Security – A Three Tier Approach

Hadoop security can be considered to be a multi-layered approach. Each layer has different set of security approaches and techniques, as depicted in Figure 3.

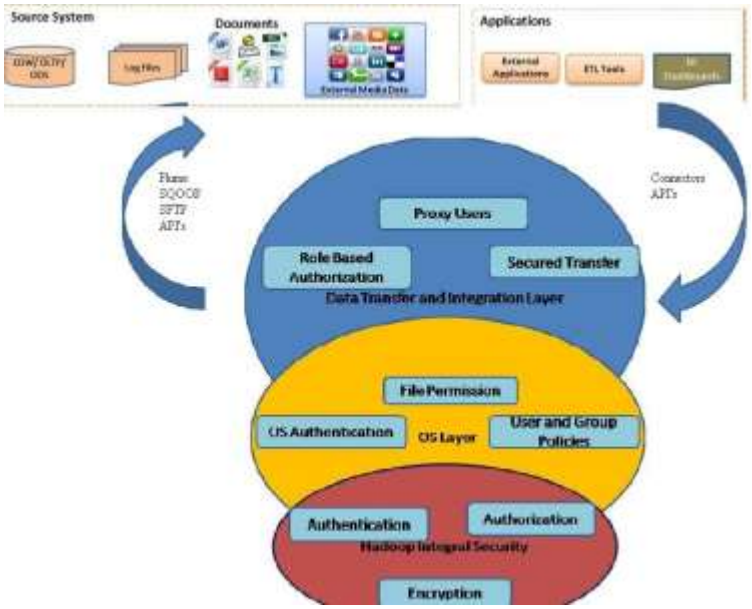

Figure 3: Three-tier Security Approach for Hadoop

- **Data Transfer & Integration Layer**

The first layer of security is at the integration cusp between the different source systems and Hadoop ecosystem. For data ingestion into and dissemination out of Hadoop, there are different methods and techniques which can transfer data back and forth from source systems. Security aspects of some of the tools/techniques for data transfer are listed below:

Apache Flume – Flume can be used for collecting, aggregating, and moving large amounts of data from multiple sources into Hadoop Distributed File System (HDFS). If multiple users need to transfer the data using Flume agent to HDFS, proxy users can be created and mapped to a single principal user. Alternately, Kerberos principal can be used to access Hadoop directly.

Apache Sqoop – Apache Sqoop can be used to transfer data to and from relational databases to Hadoop. It provides role-based access and execution restrictions using 'Admin' and 'Operator' roles. This enforces restrictions on execution of activities like import and export of data by end users.

External Tools – Extract, Transform and Load (ETL) tools or custom built applications can connect to Hadoop data stores like HBase or Hive. These data stores support Kerberos, Lightweight Directory Access Protocol (LDAP) & custom pluggable authentication. The external applications can access Hadoop as itself or by impersonating the connected user using proxy privileges which can be configured in Hadoop.

File Transfer – Secured File Transfer Protocol (SFTP) is a good option for data transfer. Also if an FTP server is to be used, then it will be better to use single user access of FTP server or use proxy user credentials with required permissions.

OS Layer - Authorization & Authentication

The Hadoop file system is similar to a Portable Operating System Interface for uniX (POSIX) file system and gives administrators and users the ability to apply file permissions and control read and write access. The interconnect of the base Operating System (OS) and Hadoop cluster is another layer which has to be secured. Big Data applications are typically deployed on Hadoop infrastructure that resides on top of the OS. It is important to consider OS users, group policies and the file permissions at the OS layer, while securing the Hadoop cluster.

For overcoming the OS related concerns, Hadoop should be configured using a user id, which is not the root user or is not part of the root users group. This user can act as a super-user for Hadoop Name Node and can have the rights to start and stop Hadoop processes. In a Hadoop ecosystem, several users, namely 'hdfs', 'mapred', 'yarn' are created during installation. Typically, a common Unix group is created to provide access to these Hadoop internal users. But, for end users who need to access HDFS, it is best to use proxy users for the same instead of giving group access. In order to further enhance the security of Hadoop cluster, security features integral to Hadoop must be fully utilized in addition to OS users and file permissions. **Proposa"Title of the RFI/RFP Response" Confidente consent of pune Global Solutions and <CLIENT> is prohibited. Pune Sensitive A Quick Look at Hadoop Security Page 5 of 7** pune Sensitive .

4. Conclusion

During the initial days of Big Data implementations using Hadoop, the prime motivation was to get data into the Hadoop cluster and perform analytics on it. 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 using layered approach. While security and privacy issues can be addressed to an extent using existing Hadoop mechanisms, more robust tools and techniques are needed.

References

[1] Jason Shih Etu, "Hadoop Security Overview- From Security Infrastucture Deployment toHigh level Servies," Hadoop and Big Data Technology Conference, 30 Nov. 2012

[2] Devaraj das,Owen O'Malley,Sanjay Radia,Khan Zhang, "Adding Security to Apache Hadoop", @Hortonworks, www.hortonworks.com

[3] Venkata Narasimha Inukollu, Sailaja Arsi, and Srinivasa Rao Ravuri, "Security Issues Associated with Big Data In Cloud Computing," International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014

[4] Hao, Chen, and Ying Qiao. "Research of Cloud Computing based on the Hadoop platform.". Chengdu, China: 2011, pp. 181 – 184, 21-23 Oct 2011.

[5] Cloud Security Alliance-Top Ten Big Data Security and Privacy Challenges. www.isaca.org

Author Profile

Nawghare Pushpmala received the B.E. and M.E. degrees in Computer Science and Engineering from SRTMU Nanded, in 2008 and Computer Networking and Engineering from BAMU Aurangabad Maharashtra in 2014, respectively. Currently working as a Assistant Professor in Zeal Education Society's Dyanganga College of Engineering and Research, Pune, Maharashtra.