International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 04 | Apr-2018 www.irjet.net p-ISSN: 2395-0072

Big Data Storage in Hadoop: A Review of Security Issues and Threats

Kavita K. Kanyan[1], Er. Ritika Mehra[2]

[1] *PG Student, Department of Computer* Science & Engineering, *R.P. Inderaprastha Institute of Technology, Haryana, India*
[2]*Assistant Professor, Department of Computer Science & Engineering, R.P. Inderaparastha Institute of Technology, Haryana, India*

Abstract –*The data in current era is increasing rapidly every year it is difficult to handle such huge amount of data using traditional applications/tools. This complex and huge amount of data also termed as Big Data. Hadoop is a framework of choice for the storage of Big Data. Data in Hadoop environment comes from different sources which makes it vulnerable to attacks. Aggregating data into one environment also increases the risk of data theft and accidental disclosure. In this paper different security issues related to storage of Big Data in Hadoop has been discussed.*

Key Words*: *Big Data, Hadoop, Security, Threats, HDFS

1. INTRODUCTION

1.1 BIG DATA

The data in the world is growing rapidly every year it is difficult to process this complex and huge amount of data using traditional applications/tools. Big Data is a term used to describe such data. This data could be structured or unstructured. Because of the variety of data, Big Data always brings number of challenges relating to its volume and complexity. Big Data enable an organization to data creation, collection, retrieval, manage, analyze and making decision that is remarkable in terms of volume, velocity, and variety.

Big Data 3 V's are [1].

Volume: At present the data existing is in petabytes and is supposed to increase to zettabytes in nearby future. The social media, financial institution, medical institution, government, Sensors, Logs producing data in order of terabytes every day and this amount of data is definitely difficult to be handled using the existing traditional systems.

Velocity: At present data change rapidly through the archived data, legacy collections and from streamed data that comes from multiple resources sensors, traditional file records, cellular technology, social media and many more.

Variety: At present data comes in different forms including data-streams, text, picture, audio, video, structured, semi structured, unstructured. Unstructured data is difficult to handle with traditional tools and techniques. Thus our traditional systems are not capable enough on performing the analytics on the data which is constantly in motion.

The importance of big data doesn't revolve around how much data you have, but what you do with it. You can take data from any source and analyze it to find answers that enable 1) cost reductions, 2) time reductions, 3) new product development and optimized offerings, and 4) smart decision making.

1. 2 HADOOP

Hadoop is the platform of choice for working with extremely large volumes of data or we can say big data.Hadoop is a java based free framework that can effectively store large amount of data in a cluster. This framework works in parallel on a cluster and has an ability to allow us to process data across all nodes.

1.2.1 HDFS (Hadoop Distributed File System)

Hadoop Distributed File System (HDFS) is the storage system of Hadoop which splits big data and distribute across many nodes in a cluster. This also replicates data in a cluster thus providing high availability.

HDFS has master/slave architecture. An HDFS cluster consists of a single NameNode, a master server that manages the file system namespace and regulates access to files by clients. In addition, there are a number of DataNodes, usually one per node in the cluster, which manage storage attached to the nodes that they run on [2]

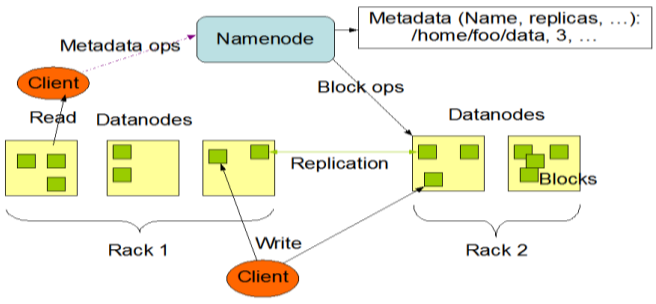

Fig-1: HDFS Architecture

HDFS exposes a file system namespace and allows user data to be stored in files. Internally, a file is split into one or more blocks and these blocks are stored in a set of DataNodes. The NameNode executes file system namespace operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to DataNodes. The

DataNodes are responsible for serving read and write requests from the file system's clients. The DataNodes also perform block creation, deletion, and replication upon instruction from the NameNode.

2. SECURITY ISSUES IN HADOOP

Hadoop was originally designed without security in mind. It was a straightforward tool for small groups to run mapreduce jobs on large amount of data with no security. But in recent years Hadoop evolved into an enterprise data platform, users from different business groups use Hadoop to access data taken from multiple sources such as weblogs, online transaction and social media interactions etc. of security sensitivities providing the need for security.

Hadoop is a dynamic framework of features, functions, and access points which makes security more difficult. Each option brings its own security options and issues. Each module runs a specific version of code, has its own configuration, and may require independent authentication to work in the cluster.

In a Hadoop cluster environment, data is processed wherever resources are available, supported by massively parallel computation. This is quite different from the centralized architecture of a traditional relational datastore. Hadoop's distributed architecture creates an environment that is highly vulnerable to attack at multiple points, as opposed to the centralized repositories that are monolithic and easier to secure. Data within Hadoop clusters is fluid, with multiple copies moving to and from different nodes to ensure redundancy and resiliency. Data can also be sliced into fragments that are shared across multiple servers. These characteristics add new complexity, and demand a different approach to data security

2.1 TYPES OF THREATS

A key component to arriving at robust security architecture for a distributed system is to understand the threats that are likely to be present, and to be able to categorize them to better understand what security mechanisms need to be in place to help mitigate those threats [3].

2.1.1 Authentication: How do we know the users who access the Hadoop framework are authenticated users, they are who they say they are and not impersonating some other user.

2.1.2 Access control: Another security challenge in Hadoop is to ensure that user can only access the data that they are entitled to access. Protection from unauthorized access to data and information is notable challenge within data transmission process. Because of the dynamic scalability of Hadoop cluster illegal user can disguise as a data node server and join to the cluster to receive data information from name node. Illegal user disguises as authorized user by altering data package to request service sources.

2.1.3 Data Protection: How the data is encrypted or otherwise protected while it is in storage (at rest) and when it is moving across the network (in motion)? So that it cannot be accessed by malicious user. Hadoop and the vast majority of distributions (Cassandra, MongoDB, Couchbase, etc.) don't communicate securely by default — they use unencrypted RPC over TCP/IP. TLS and SSL capabilities are bundled in big data distributions, but not always used between client applications and the cluster resource manager, and seldom for inter-node communication. This leaves data in transit, along with application queries, accessible for inspection and tampering.

2.1.4 Monitoring, filtering and blocking: There are no built-in monitoring tools to detect misuse or block malicious queries. There isn't even a consensus yet on what a malicious big data query looks like — aside from crappy queries written by bad programmers.

2.1.5 Multi-tenancy: Hadoop is commonly used to serve multiple applications and tenants, each of which may be from m different groups with one firm, or altogether different companies. The security control should be implemented to ensure privacy so that data of one tenant should not be shared with another tenant.

2.1.6 Auditing: Storing the data access history for compliance and other purposes such as if someone has breached the cluster, can the user detect it, or trace back to the root cause.

2.1.7 Client interaction: Clients interact with the resource manager and nodes. Gateway services can be created to load data, but clients communicate directly with both resource managers and individual data nodes. Compromised clients may send malicious data or link to services. This facilitates efficient communication but makes it difficult to protect nodes from clients, clients from nodes, and even name servers from nodes. Worse, the distribution of self-organizing nodes is a poor fit for security tools such as gateways, firewalls, and monitors

2.1.8 Distributed nodes: One of the key advantages of big data is we can perform computation on data without moving the data. Data is processed wherever resources are available, enabling massively parallel computation. Unfortunately this produces complicated environments with lots of attack surface. With so many moving parts it is difficult to verify consistency or security across a highly distributed cluster of (possibly heterogeneous) platforms. Patching, configuration management, node identity, and data at rest protection — and consistent deployment of each — are all issues.

2.2 THREATS AND RISK ASSESSMENT

It's important that in addition to understanding the types of threats we should also assess the risk to particular distributed system**.** Assessing the threats to a distributed

system involves taking a closer look at two key components: the users and the environment [4].

2.2.1 User Assessment

It is important to understand users to whom different services of Hadoop will be exposed to. There can be mainly two types of user one who needs direct access to the distributed system such as developers or business analysts or the users with indirect access to the system. If the users of the distributed system are all developers, several assumptions can be made about the need for shell access to nodes in the system, logfiles to debug jobs, and developer tools. On the other hand, business intelligence analysts might not need any of those things and will instead require a suite of analytical tools that interact with the distributed system on the user's behalf. On the other hand users with indirect access to the system won't need access to data or processing resources of the system. These types of users need to be accounted for in the overall security model.

2.2.2 Environment Assessment

To assess the risk for our distributed system, we'll also need to understand the environment it resides in. One of the key criteria for assessing the environment is to look at whether the distributed system is accessible to the Internet. If so, a whole host of threats are far more likely to be realized, such as DoS attacks, vulnerability exploits, and viruses. Distributed systems that are indeed connected to the Internet will require constant monitoring and alerting, as well as a regular cadence for applying software patches and updating various security software definitions.

Another criterion to evaluate the environment is to understand where the servers that comprise the distributed system are physically located. Are they located in your company data center? Are they in a third-party–managed data center? Are they in a public cloud infrastructure? Understanding the answer to these questions will start to frame the problem of providing a security assessment. For example, if the distributed system is hosted in a public cloud

It's hard to definitively know who has direct access to the machines. Threats to communications that occur across an open network to a shared public cloud have a much higher risk of happening than those that are within private company data center.

3. CONCLUSION

In the era of Big Data where data comes from multiple sources security is of main concern. Hadoop framework is used for storing and processing huge amount of data. Since the first priority of Hadoop is efficient voluminous computation and then comes security, Hadoop has many security issues. Some of the security issues are figuring out what data users need to protect, who can access what pieces of data, and what actual mechanisms to use.

REFERENCES

[1] Bermen, Jules J. "Principle of Big Data", Morgan Kaufmann, Waltham, 2013.

[2] Hadoop Architecture

https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html.

[3] https://securosis.com/assets/library/reports/Securing_Hadoop_Final_V2.pdf.

[4] Ben Spivey, Joey Echeverria "Hadoop Security" , O'Rielly Media ,Inc.,2015.