Array 1-2 (2019) 100002

Contents lists available at ScienceDirect

Array

journal homepage: www.elsevier.com/journals/array/2590-0056/open-access-journal

Big Data: Hadoop framework vulnerabilities, security issues and attacks

Gurjit Singh Bhathal *, Amardeep Singh

[a] Department of Computer Science and Engineering, Punjabi University Patiala, Punjab 147002, India

ARTICLE INFO

Keywords:
Big Data
Hadoop
MapReduce
Security threats
Vulnerability

ABSTRACT

Big Data is a collection of different hardware and software technologies, which have heterogeneous infrastructure. Hadoop framework plays a leading role in storing and processing Big Data. It offers fast and cost-effective solution for Big Data and is used in different sectors like healthcare, insurance and social media. Hadoop is an open source framework based on a distributed computing model and is applied for processing and storing data on a cluster of commodity computers. Due to the flexibility of framework, some vulnerabilities arise. These vulnerabilities are threats to the data and lead to attacks. In this paper, different types of vulnerabilities are discussed and possible solutions are provided to reduce or eliminate these vulnerabilities. The experiment setup used to perform common attacks to understand the concept and implementation of a solution to avoid those attacks is presented. The results show the effect of attacks on the performance. According to results, there is need to protect data using defense-in-depth to security.

1. Introduction

Big Data is a collection of huge amount of historical and important data, which is the most valuable asset of every organization and being utilized intelligently for business can support decisions based on real facts rather than perceptions. The term Big Data was coined by Charles Tilly in Oxford dictionary in 1980 [1]. At present, considerable data is generated from multiple sources including social media sites, different remote sensors, cell-phone GPS signals, transaction records, and log files. Owing to the socialization of the Internet, terabytes of structured, unstructured, and semi-structured data are produced online every day and much of this information has inherent business values. Thus, if it's not captured and analyzed properly, then considerable vital data will be getting lost. Big Data is to archive a collection of historical data, using Hadoop framework with the help of different tools for analytics, which are faster than previous traditional analytic tools.

Hadoop is a synonym of Big data. Big Data comprises of four V's or characteristics: volume, velocity, variety, and veracity [2,3]. The significant growth of data has led to issues related to not only volume, velocity, variety and veracity of data but also to data security and privacy. Recently, the addition of another "V", i.e., vulnerability, has been proposed to be related to it (See Fig. 1) [4].

Big Data technology extracts interesting value from a data lake and many countries have launched important projects based on this technology. USA is one of the leaders to seize the Big Data opportunity. In March 2012, under Obama's Administration, USA launched the Big Data Research and Development Initiative with a budget of $200 millions [5]. Due to this Big Data project initiated globally with new technologies, frameworks, many new models have been developed. Infrastructure has been created to provide for more storage capacity, parallel processing, and real-time analysis in a heterogeneous environment [6]. Big Data technology offers improved flexibility, scalability, and performance in a cost-effective manner with commodity computers. The cost of most storage and processing solutions is continuously dropping due to the use of the advanced sustainable technology [7].

This new technology has been developed to ensure data privacy and security in contrast of traditional technologies. However, this technology can also be used for negative purposes. As a growing number of businesses and individuals store and process their private data using this technology, it has become a significant target of data attacks.

2. Hadoop framework

Hadoop is an open source framework by Apache Software Foundation primarily based on distributed computing and parallel processing concepts for batch operations [33]. Hadoop has combination of components - HDFS for storage, MapReduce for data processing and YARN for resource management in cluster.

Hadoop was originally developed in 2005 and first released in 2011 to support distribution for the Nutch search engine project at Yahoo [8].

* Corresponding author.
E-mail addresses: gurjit.bhathal@gmail.com (G.S. Bhathal), amardeep_dhiman@yahoo.com (A. Singh).

https://doi.org/10.1016/j.array.2019.100002
Received 2 March 2019; Received in revised form 29 June 2019; Accepted 10 July 2019
Available online 20 July 2019

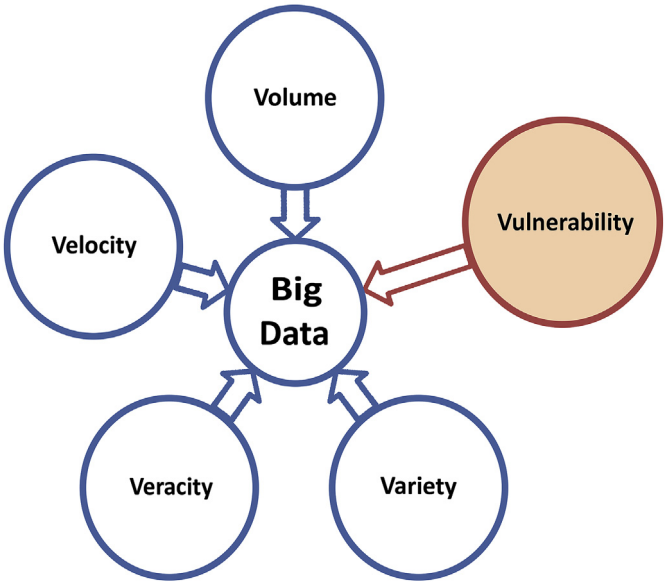

Fig. 1. The V's are represent different characteristics of Big Data, due the system vulnerabilities one more V is included.

This release was with minimum security support, being run on the trusted environment. Hadoop has since emerged as one of the most cutting-edge technologies to store, process and analyze big data through use of a cluster scale-out environment [9]. Hadoop users are spread all over the world, mostly large companies including eBay (Using Apache Hadoop of 532 nodes cluster), Facebook (using 1100-machine cluster with 8800 cores and about 12 PB raw storage) and Yahoo (biggest Hadoop cluster: with 4500 nodes each 2*4cpu boxes w 4*1TB disk & 16GB RAM) [10]. Forrester predicts the global Big Data software market will increase from $42B in 2018 to $103B in 2027 [11].

In 2008 Google processed 20 petabyte of data on a single day. This was the starting phase of new technology which can process a large amount of data in a short time span. The size of the cluster is scaling horizontally and vertically with increased demand [9]. Due to expanded (scaled out) clusters, the network threats and vulnerabilities have also increased. To fulfill the market demands of security and centralize administration, some vendor specific Hadoop distributions improvised the official release of Apache Hadoop. The popular Hadoop distributions are Cloudera CDH-5.13 (Cloudera Distribution Including Apache Hadoop), Hortonworks HDP-3.1 (Hortonworks Data Platform) and MapR [12,13]. It is considered to to scale up to a cluster of 4000 nodes, to recognize server crashes as “normal” in large organizations, and to put together data storage and analytics robust against failures. The functional parts of Hadoop are Name Node for metadata, Data Node to store actual blocks (64 MB or 128 MB) of data, Job Tracker and TaskTracker. Hadoop has multiple interfaces, default settings, large volume, and lack of internal cluster security, and hence, it is under threat [14].

2.1. Hadoop distributed file system (HDFS)

Hadoop distributed file system is a fault tolerant storage system. It distributes large files across the cluster consisting of many Data Nodes with local storage. During this process Name Node partitions, the original file into a block, size of 64 MB by default and replicates it to the different Data Nodes based on pre-defined rules. Name Node also maintains the metadata for this replication and allocation. Each data block is replicated three times for high availability, two on same rack Data Nodes and one from different rack Data Node. Data Node of the cluster stores a small fragment of the entire file. Name Node is always aware of which data block belongs to which file,where the data blocks are placed, and where storage capacities are occupied. Using periodically transmitted signals, Name Node always knows which Data Nodes are still alive. If a signal (heartbeat) is missing, Name Node recognizes the failure of a Data Node, removes the failed Data Node from the Hadoop cluster, and tries to distribute the data load equally across the existing Data Nodes. Furthermore, the Name Node ensures that the defined number of data copies are always maintained for high availability.

2.2. Map reduce

MapReduce is a parallel processing framework work based on the master-slave principle, similar to HDFS. It is a combination of one master (JobTracker) daemon and 3 slaves (TaskTracker), daemons per slave in a cluster. Map Readuce processing data parallel based on different algorithms for a map and reduce. This works in two steps, map task and reduce task. This JobTracker divides the dataset into multiple chunks called tasks (Map tasks) and distribute them to three Data Nodes (TaskTracker) by default across the distributed connected commodity computrers on a network for parallel processing.

Typically, the map tasks run on the same cluster Data Nodes where data resides (Data locality). If a node is already heavily loaded, another node that is close to the data, i.e., preferably a node in the same rack, is selected. Intermediate results are unavailable to the user and are exchanged among the nodes(Shuffling), and thereafter, merged by the reduced tasks to obtain the results.

Intermediate results of map phases can be aggregated by keeping the data volume as low as possible during trasnfer from map tasks to reduce tasks. The intermediate results are deposited in the local file system of the Data Nodes. The JobTracker is responsible for monitoring to reschedule any task in case of disruption.If a task does not notify any progress in a pre-determined time, or if a Data Node fails completely, all tasks are restarted on another server including tasks but, they are not finished. If a task runs extremely slowly, the JobTracker also restarts the task on another server in order to execute the overall job in appropriate time (speculative execution).

The only weak spot here is the JobTracker itself because it represents a single point of failure. Hence, to overcome this problem, secondary Name Node is introduced in the system to make as many redundant components as possible, in order to keep the probability of failure low. MapReduce can be directly applied for executing business analytics, a frequent use case is to transform data into an optimized shape for analytics. The security issues with the MapReduce framework include lack of authentication within Hadoop, communication between Hadoop daemons being unsecured, and the fact that Hadoop daemons do not authenticate each other.

2.3. Yet another resource negotiator (YARN)

YARN is the sub-project of Apache Hadoop. In 2012, MapReduce was divided into two parts: MapReduce and YARN [8]. The basic principle of YARN is to divide resource management and job scheduling functionalities into separate daemons. The function of a resource manager is to arbitrate resources between all system applications, and with the help of a node manager, to monitor feedback of the resource usage of CPU, memory, disk, and network on each node. The resource manager has two main components: scheduler and application manager. The scheduler allocates resources to the various running applications, and performs scheduling based on the resource requirements of the applications. The resource manager accepts job submissions, and each job is allocated to application manager. The application manager allocates container for executing the application, and to restart the application master container on failure, each application has application-specific master.

3. Vulnerability

Vulnerability is a flaw or weakness in system security procedures, design, implementation, or internal controls. Vulnerability can be

accidently triggered or intentionally exploited, causing security breaches [15].

All Hadoop environment components like Sentry, Flink, and Storm are prone to attacks caused by various vulnerabilities; it can be in software, web interface, or network. The Hadoop framework is a complex collection of application programs, distributed computing software, hardware, and policies for assessing these resources. Vulnerabilities are divided into the following three categories:

- Technology/Software Vulnerabilities; for example, the Hadoop framework is written entirely in Java, which is heavily exploited by cybercriminals and implicated in various security breaches.
- Configuration/Web Interface Vulnerabilities: Hadoop has a weak configuration because numerous default settings such as default ports and IP addresses are vulnerable and have been recently exploited. Mostly Hadoop web interfaces are vulnerable to XSS scripting attack, such as Hue.
- Network/Security Policy Vulnerabilities: the Hadoop framework is a mixture of various databases; different types of users in these policies are not configured properly, and thus, fine-grained policies are required at both the service level and the data level.

3.1. Litrature review

Vulnerabilities are divided into three major categories: infrastructure security, data privacy, and data management [16]. These are further categoriesd in three dimensions: architecture dimension, life cycle of data dimension, and data value chain dimension. According to the author, infrastructure involves the hardware and software vulnerabilities which are in architecture dimension. Data privacy involves the data at rest and data in transit which involve the life cycle of the data.

As per another work [17], research on security and privacy issues of Big Data is divided into five heads: Hadoop security, monitoring and auditing, key management and anonymization. This paper specifically discusses security and privacy issues of data in cloud environment of Hadoop. Author also discusses various encryption mechanisms to impalement security in Hadoop HDFS (Hadoop Distributed File System) to achieve authentication in Hadoop, Kerberos network authentication protocol used. Other methods and algorithms discussed for monitoring and security of sensitive information are Bull eye algorithm and the NNSE (Name Node Security Enhance) method, used between the master node and data nodes in Hadoop.

Verizon released a white paper [18] on cloud security. In that paper, the cloud infrastructure layered security model is proposed. That model is divided into four parts: basic security, logical security, value-added security, and governance. In that paper Verizon also explained the sticky policy framework architecture. This is proposed for securing the Big Data applications on cloud infrastructure.

In another paper [19], author has discussed the different types of attacks such as impersonation, denial of service, replay, eavesdropping, man in middle, and repudiation. According to the author, MapReduce computation is distributed in nature and open opportunities for a wide range of attacks. Secure MapReduce computation requires proper authentication, authorization, access control, and availability of data for the mapper and reducer, and confidentiality of data. The Kerberos protocol is recommended for authentication. It uses various tokens such as delegation, block access, and job tokens. Other security tools such as Apache Knox, Apache Sentry, and Apache Ranger are discussed and recommended. In the paper, data privacy protection from advisory and multiuser cloud providers on a single public cloud are also discussed.

3.2. Vulnerability databases

Numerous online databases are available on internet and these are exposing countless vulnerabilities in different types of products including hardware and software. National Vulnerability Database (NVD) is a vulnerability management repository containing security-related software flaws and impact metrics. The Computer Emergency Readiness Team (CERT) is another vulnerability database providing information about software vulnerabilties. Microsoft Security Bulletins is also related to security issues discovered in Microsoft software, published by the Microsoft Security Response Center (MSRC). Common Vulnerabilities and Exposures (CVE) is another database list of vulnerabilities with an identification number. Open Source Vulnerability Database (OSVDB) provides accurate and unbiased information about security vulnerabilities. CVE contains numerous vulnerabilities of cyber security products and services from around the world.

CVE determines vulnerabilities unambiguously by CVE numbers like CVE –2017–3161. The central set in these numbers indicates the year of discovery, whereas the last set indicates AN number ranging from one to any, started from one every year and incremented by one for each reportable vulnerability. The CWD ID–79 (Common Weakness Enumeration ID) uniquely identifies this vulnerability type, such as cross-site scripting (XSS). The newly reported vulnerability is added into CVE Database after following some steps of procedure before displaying publicly, which is controlled by an organization called MITRE. Hadoop is also a software framework that contains several vulnerabilities, which will be discussed in this paper. The list of Hadoop system vulnerabilities taken from the CVE database is given below [20]. As per the CVE database records of last nine years, the Apache Software Foundation detected/reported several vulnerabilities in their products, but in current year 2019 till date total 49 vulnerabilities have been detected out of those 14 are related to XSS and one DoS. According to the data available in the CVE database, web interfaces are the most vulnerable. Almost all detected vulnerabilities are patched, but some of them are exploited by the attackers with DoS, XSS, and gain access of system (see Table 1) list the vulnerabilities categories [21]. In table (See Table 2) data shows details of vulnerability. Some of the vulnerabilities have been detected 2018 and were reported in 2019.

3.3. Patch management

Vulnerability of patch management is the process of rooting out and eliminating these weaknesses before these are abused. Efficiency of patch management system depends on how fast vulnerability is detected, corrected, and patched through periodic penetration (pen testing) and code review using vulnerability scanning. As per the 2017 report, usage of vulnerability scanning has increased by 41.7% globally [22]. In patch management, newly discovered vulnerabilities are patched using vendor-provided patches. Input validation/sanitization should be conducted by the filtering and verification of incoming traffic by using a web application firewall, which blocks attacks before they can exploit vulnerabilities and is a substitute for fully sanitizing an application code. Vulnerability scanners automate security auditing by scanning your network and websites for different security risks and also possible for some to even automate the patching process. The most popular tools are Nexpose is an open source developed by Rapid7 carrying out a wide

Table 1
Different Vulnerabilities year wise data which are detected and reported in Hadoop framework.

Year	Total Vulnerabilities	DoS	XSS
2011	44	15	7
2012	63	19	6
2013	74	25	9
2014	92	23	6
2015	57	19	5
2016	103	15	17
2017	217	29	22
2018	148	15	9
2019	49	1	14

Table 2
Detailed CVE of Apache Hadoop and Hadoop distributions already reported.

CVE ID	Description
CVE-2018-11767	Hadoop 2.7.5 to 2.9.1, KMS blocking users or granting access to users incorrectly, if the system uses non-default groups mapping mechanisms
CVE-2018-11766	Apache Hadoop 2.7.4 to 2.7.6 privilege escalation vulnerability. A user who can escalate to yarn user can possibly run arbitrary commands as root user
CVE-2017-7669	In Apache Hadoop 2.8.0 the LinuxContainerExecutor runs docker commands as root with insufficient input validation. When the docker feature is enabled, authenticated users can run commands as root.
CVE-2017-3162	HDFS clients interact with a servlet on the Data Node to browse the HDFS namespace. The Name Node is provided as a query parameter that is not validated in Apache Hadoop before 2.7.0.
CVE-2017-3161	The HDFS web UI in Apache Hadoop before 2.7.0 is vulnerable to a cross-site scripting (XSS) attack through an unescaped query parameter.
CVE-2017-15713	Vulnerability in Apache Hadoop 3.0.0 allows a cluster user to expose private files owned by the user running the MapReduce job history server process. The malicious user can construct a configuration file containing XML directives that reference sensitive files on the MapReduce job history server host.

range of network checks, others are openVAS, Nmap and wireshark etc.

4. Architectural security issues in hadoop

Hadoop, as we recognize, is an open-source venture that includes various modules, which are separately developed over time to add different types of functionalities to its core capabilities. Security was a late addition, and thus, Hadoop lacks a consistent security model. By default, Hadoop assumes a trusted environment. Hadoop has focused on improving its efficiency. Researchers are gradually paying attention to Hadoop security concerns and building security modules for it. However, currently, there is no existing evaluation for these Hadoop security modules.

Due to huge volume, rapid growth, and diversity of data, these are unstoppable and existing security solutions are not adequate, which were not designed and build with Big Data in consideration. The Hadoop ecosystem is a mixture of different applications including Pig, Hive, Flume, Oozie, HBase, Spark, and Strom. Each of these applications require hardening to add security capabilities to a Big Data environment and functions to be scaled with the data. Bolt-on security doesn't scale well and easy. The security tools vendor have customizable offerings and applying a one point entry (gateway/perimeter) so that commands and data are entered into the cluster from single entry.

Hadoop is distinguished by its fundamentally different deployment model, which exhibits highly distributed, redundant, and elastic data repositories [34]. However, the architecture of distributed computing present a unique set of following vulnerabilities and security threats for data center managers and security professionals.

- Distributed computing and fragmented data
- Node-to-Node communication and access to data
- Multiple interfaces

5. Security threats and possible attacks

A threat is a potential danger to an information system. A threat is that someone, or something, will identify a specific vulnerability and use it against a company or individual to attack the system [15]. The basic principle of security is CIA, which refers to confidentiality, integrity, and availability. Confidentiality is only possible if only an authenticated user can assess the system; it is also important to determine who should have access to the system and what resources a user can access. Hadoop security is divided into two levels. With respect to the time at level-1, the Hadoop system starts with a minimum security computing in only a trusted environment. The Hadoop system adds Kerberos for authentication as a perimeter security but not inside the cluster in level-2. Different security projects such as Apache Sentry, Apache KNOX, Apache Ranger, and Rhino are included in Hadoop at security phase Level-3. Kerberos also integrates with AD and LDAP. Each security product provides a discrete security solution due to the different purposes and functionality of each project. However, the Hadoop system is still required to move to security level-4 as a concrete solution for centralized security in one project.

How to investigate data leakage attacks in Hadoop is important but a long-neglected issue as per McAfee's report, which states that different threats to data have increased recently [23]. The following section explains security threats that can hamper the operation of MapReduce and all framework components in the absence of a protected MapReduce environment. A dispersed and replicated data processing of MapReduce can unlock an opportunity for a large range of attacks.

- Authentication and authorization: Identity and authentication are central to any security effort. Without it we cannot determine who should have access to data and who should not. This is done at user level by full integration of Active directory (AD) and Lightweight Directory Access Protocol (LDAP) with Kerberos, and at service level with Role-Based Access Control (RBAC) at the node level.
- Administrative data access should be ACL's, File Permission and Segregation of administrative roles of OS administrators and Hadoop administrators.
- Hadoop stacks have many different components which come with default setting and no security. Configuration and patch management is required when running different configurations and patch levels at one time.
- There are no built-in monitoring tools to detect misuse or block malicious queries. Logs should configure to capture both the correct event types and sufficient information to determine user actions. Audit and Monitoring tools are important when volume and velocity of data are high.
- Applications are built on web service models, especially in social networks like Facebook, Yahoo and Snap Chat. Hadoop Web Applications may be vulnerable to well-known attacks due to insecure API's. Big Data cluster APIs need to be protected from usual web service attacks command injection, buffer overflow attacks, and all the other.

5.1. Impersonation attacks

Impersonation attacks occur when an attacker tries to access resources by impersonating as an authorized identity. The attacker steals the credentials of an authorized user by using different methods and attacks Hadoop cluster resources, services, and jobs. Impersonation can be performed by replaying the tickets issued by the authentication server (Kerberos) for different purposes. Once the attacker gets access to the cluster, they can perform any type of data leak and slow down the processing of MapReduce.

5.2. DENIAL-OF-SERVICE (DoS)

A DoS attack [24] occurs when the resources are unavailable to authorized users. As per the Verizon 2017 Data Breaches Report, 11246 attack incidents have been reported and out of those 5 breaches have been successful [25]. DoS attacks are accomplished by flooding the target with traffic or causes a crash of data. In each instances, the DoS attack deprives legitimate users of the service or resource they expect. There are two general ways of DoS attacks: flooding services or crashing services. Flood attacks occur once the system receives an excessive amount of traffic for the server to buffer, inflicting it to abate and eventually stop.

Well-known flood attacks embody distributed denial of service (DDoS), buffer overflow, SYN flood, Ping to Death, and HTTP flood. In Hadoop, the Name Node and authentication server are vulnerable to DoS attacks. The Name Node has a master daemon that is responsible for scheduling and coordinating the execution of MapReduce applications on the data nodes. A DoS attack on the Name Node can halt all computations of MapReduce and read-write operations of HDFS.

5.3. CROSS-SITE scripting (XSS)

XSS [26] is a common attack which injects a malicious code into a vulnerable web application. It differs from SQL injection because in that it does not directly target the application itself. But instead, the web application users are the ones at risk. XSS attacks can be broken down into two types: stored and reflected. Stored XSS, also known as persistent XSS, is more damaging than reflected XSS and occurs when a malicious script is injected directly into a vulnerable web application. Reflected XSS involves reflecting a malicious script of a web application onto a user's browser. The script is embedded into a link and is only activated once that link is clicked on. Hadoop web UI's are vulnerable to attacks, and recently, different database installations have been attacked.

5.4. Recent attacks

Hadoop is reportedly easily identifiable to hackers worldwide by simply sniffing to open instances, and around 5307 Hadoop clusters are exposed to the web with their security settings off, exploit them with receptive attack by hackers [27]. The online search engine Shodan2 shows details about all servers and devices connected to the Internet. The merit of this is security recommendation, while the demerit is that it is used by hackers to see all details of any network, server location, and other settings. How to protect data attacks in Hadoop is an important issue. To counter these attacks and stop data theft, it is first important to thoroughly understand the vulnerabilities as well as threats, and then, it can be worked toward devising a strategy using defense-in-depth to secure data.

6. Attack implementation setup

In our experiment, Hadoop cluster was setup in a virtual environment with a quick-start cloudera cluster in virtual box. The environment was configured on Intel processor i3 with 16 GB RAM and 1 TB hard disk. To perform the experiment, latest version of CentOS 7 was used and Hadoop 3.1.1 was included in CHD-5.13. In cluster, one Name Node and three Data Nodes are configured. The secure terminal was inbuilt and was running on port 4200 and Web UI also known as cloudera manager at port 7180 to perform the attack, some information gathered like IP addresses and open ports of the Name Node using zenmap and wireshark to capture the traffic. To run MapReduce job the wordcount example taken from the link includes mapper, reducer, combiner and main java program in one single file. https://archive.cloudera.com/cdh5/cdh/5/hadoop/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html.

6.1. Denial-of-service attack

To perform Denial of service (DoS) attack, need to focus on the communication between a Name Nodes and Data Nodes, this communication is totally based on the TCP connection, which performs three-way handshaking to establish a connection first send the SYK packet then receive SYK ACK packet back and then ACK to complete the connection. SYK-flood attack was performed which is suitable for Denial of service (DoS) attack against the TCP connection and this attack was performed from a compromised Data Node. In this attack Name Node din't shutdown or crashed due to the high availabiltyof architecture and fault tolerant but decreases the performance as shown in the experiment.

To perform a Denial-of-Service attack against Name Node, the penetration testing tool hping3 is used. Previously, this tool was used as a security tool for a firewall security check, port scanning and trace route. It allows many different types of packets (like: TCP, UDP, ICMP and RAW-IP Protocols) for attack to the destination address Port 50070, the open port for the web interface of the Name Node, was targeted with a SYK-flood attack on Name Node IP address. At the same time it is needed to perform a MapReduce job to check any effect of attack on system performance. A large file was chosen to perform the experiments. This text file is downloaded from https://www.hdfstutorial.com/blog/datasets-for-hadoop-practice/and the block size was 128 MB in order to simulate Hadoop cluster. Which would have many blocks per Node. The specific MapReduce word count program runs in the Hadoop cluster (See Fig. 2). It was observed DoS attacks slows down the execution of running jobs. The job was run during the attack using hping3 and same job was also run without any attack against the cluster so as to know the effect on

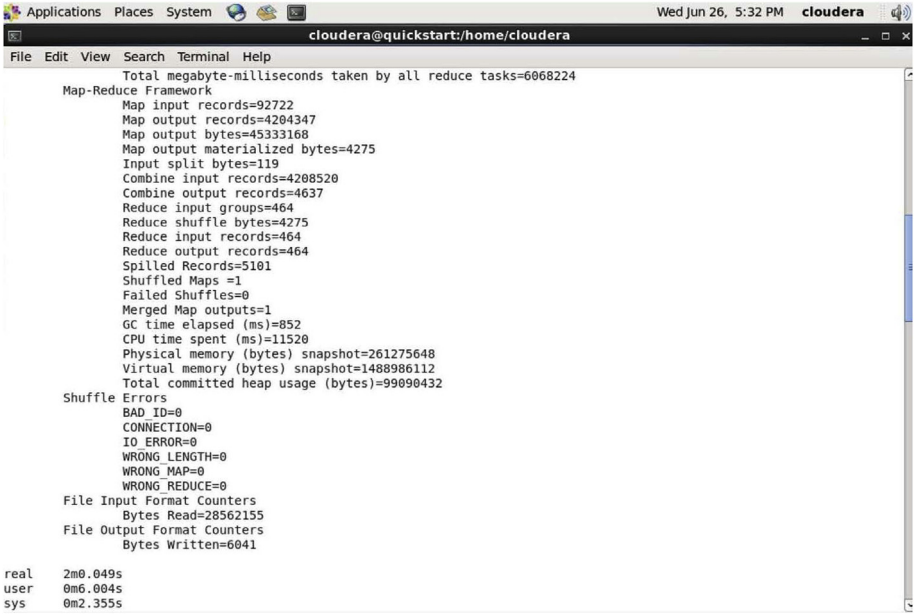

Fig. 2. Screen with details after Hadoop MapReduce completion.

Table 3
Different processing time of MapReduce job without attack and with attack.

Experiment (Same data set)	Time (s) taken to process MapReduce job		
	Without Attack	Ports Block Attack	SYK Flood Attacks
1	142.635	851.917	268.257
2	117.952	701.605	253.318
3	127.164	980.829	273.765
Average	129.25	844.783	265.113

completion time, which could be compared in table (See Table 3) to check DoS attack affect on MapReduce processing in Hadoop cluster.

6.2. Block port communication

To block the communication between a Name Node and Data Nodes, the strategy made was little bit different because Hadoop system is resilient in the case of hardware failure, so there is no way to stop any hardware of Data Node or Name Node. For implementing block communication, the different types of available ports and services were observed between a Name Node and Data Node. Hadoop system detects failure dynamically and allocates data to the other Data Node. The failure detects with the help of the heartbeat signal sent by Data Nodes to the Name Node, sowe targeted heartbeat signal meansif a Name Node receives heartbeat signal regularly Name Node sending other data to the Data Node even Data Node is compromised.

To implement blocks communication attack, the network traffic was captured and analyzed using the wireshark tool (See Fig. 3). In analysis, constant traffic was found between the Name Node and the Data Nodes on the Name Node ports of 7255 (a value which depends upon each individual configuration of Hadoop) and 8031 (default resource tracker port).After examining the contents of packets on both of these ports found that both of them have been involved in the heartbeat messages exchanged between Name Node and Data Node. To block the other communication firewall rule added to block other ports traffic except these two ports.

7. Security tools for hadoop cluster

The below section briefly explains some existing reviewed security tools for Hadoop cluster security. Diffrent features of these tools are compared in detail in table (See Table 4).

7.1. Apache KNOX

Apache Knox Gateway [28] is a single access point to a single or multiple Hadoop clusters based on the concept of the stateless reverse proxy framework. It also provides authentication to a group of authenticated users, authorization, auditing, and system monitoring. Knox hides data and details of Hadoop cluster installations, simplifies the amount of services that user have to be compelled to move with, and limits the numbers of access points with single entrance URL. Knox has a REST API -based perimeter security entrance system that authenticates user credentials against AD/LDAP and solely a successfully authenticated user is allowed access to the Hadoop cluster.

7.2. Apache Sentry

Apache Sentry [29] is a granular, role-based authorization and a multi-tenant administration module for Hadoop. Sentry can manage access to data and metadata by enforcing an accurate level of privileges to authenticated users and applications in a Hadoop cluster. It is extraordinarily standard and might support authorization for a broad vary of data models in Hadoop. It is versatile and permits the definition of

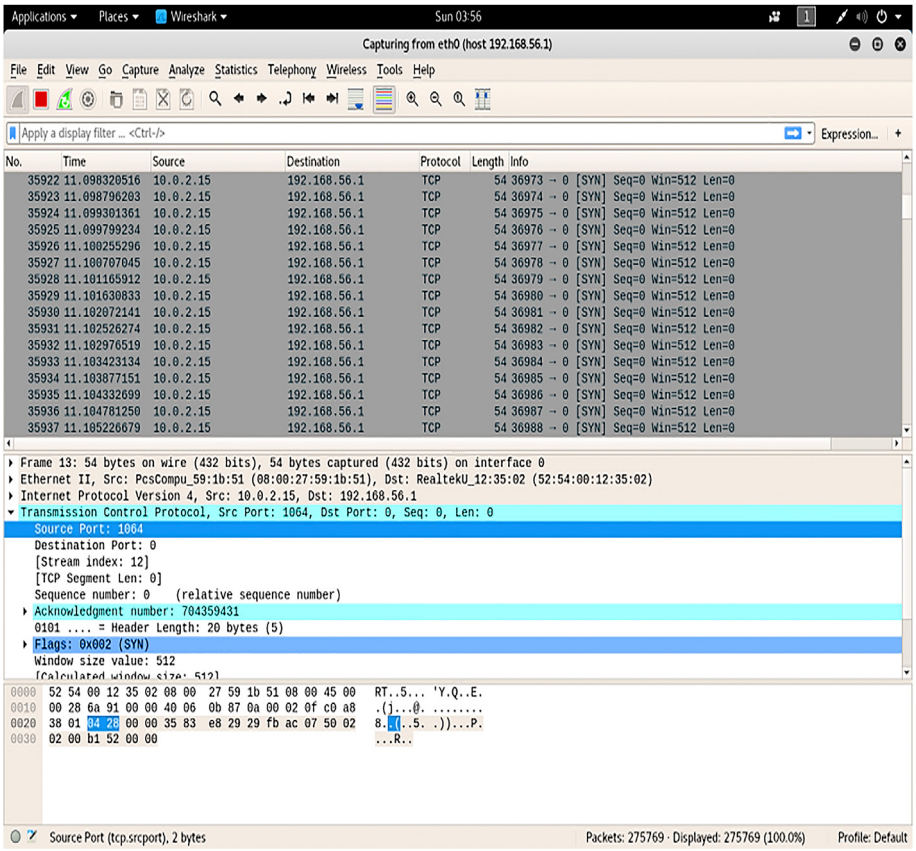

Fig. 3. Wireshark Screen short captured traffic during hping3 SYN attack.

Table 4
Comparison detailed feature of security solutions used in Apache Hadoop Framework and in vendor specific Hadoop distributions.

Functions	Apache Knox Gateway	Apache Sentry	Apache Ranger	Project Rhino	Kerberos
Introduced By	Hortonworks	Cloudera	Hortonworks	Intel	MIT
Administration/ Authentication	Authentication Single Access point HTTP, Perimeter Security using REST API, Integration with user Credential AD/LDAP	Integration with Apache Knox	Centralized security administration using REST APIs and authorization	Single Sign-on and token-based authentication, common authorization framework for Hadoop	Encrypted Token-based Authentication to All Hadoop Resources
Authorization	Support Service-Level Audit	Fine- Grained Role Based but Granularity for Columns not supported, At File level in HDFS	Centralized Authorization using RBAC, attribute-based access control. Granularity for Columns in Hive	Cell-Level support for HBase	Not Supported
Audit	Support Service-Level Audit	Not Supported	Centralize auditing of user access for all the components of Hadoop	Support Logging audit	Not Supported
Data Protection	Not Supported	Not Supported	Not Supported	Encryption and Key Management	Not Supported

authorization rules to validate a user's or application's access requests for Hadoop resources. Sentry is meant to be a pluggable authorization engine for Hadoop elements like Apache Hive, Apache Solr, Impala, and HDFS.

7.3. Apache Ranger

Apache Ranger [30] provides a centralized framework for securing Hadoop. Ranger is associate authorization system that allows/denies access to Hadoop cluster resources (HDFS files, Hive tables, etc.) supported pre-defined Ranger policies to authenticated users. When a user request comes to Ranger, it is assumed to have been already authenticated. Apache Ranger uses Kerberos for authentication and Apache Knox for authorization: role -based access control (RBAC). Apache Knox also supports auditing of HDFS, Hive, and HBase, and Apache Ranger uses wire encryption for data protection.

7.4. Project Rhino

Project Rhino [31] provides data protection to Hadoop stack with the single-sign-on (SSO) concept and supports encryption, a common authentication –authorization module key management. Rhino enhances cell-level encryption and fine-grained access control to HBase 0.98 and encryption to data-at-rest in Apache Hadoop. Data encryption in Hadoop requires both data-at-rest and data-in-transit; however, most Hadoop components provide encryption for data-in-transit only.

7.5. Kerberos

Kerberos [32] is an authentication protocol developed by MIT. It is used in Hadoop to provide authentication to the user to access a Hadoop cluster. The Kerberos protocol uses secret-key cryptography to provide secure communications over a non -secure network. Kerberos is an SSO ticket-based system that relies on KDC. The Kerberos protocol generates three types of tickets for authentication: delegation token is a secret key between a user and Name Node for authentication, block access token is used to access a file from HDFS authenticated by Name Node and Data Node jointly to access a data block on the Data Node, and job token is generated by JobTracker to authenticate tasks at TaskTrackers. Kerberos KDC comprises of three components: an authentication server, a ticket -granting server (TGS), and a database.

8. Discussion

To protect the Hadoop environment, authentication, authorization, and accounting policies for accessing and using data stored in the Hadoop clusters must be implemented. If one node of Hadoop cluster is compromised then there are chances of any type of attack and data theft is possible. Securing the data in transit as well as data at rest is required. Securing Hadoop not only involves securing access to Hadoop and securing the stored data (data at rest) but also the whole gamut of security that all IT operations use, such as network security and operating system security.

Various open source communities, IT development organizations and research institutes are working together to improve the Hadoop infrastructure, tools, and services. Big Data innovations are shared through open-source platforms which are helpful to promote Big Data technologies. However, users facing difficulties due various versions of modules from different sources combined into a Hadoop framework. Because each Hadoop module has its own curve of maturity, there is a risk of version incompatibility inside the Hadoop framework. In addition, the integration of different modules of different vendors with different technologies on a single platform increases security risks. However, most of the time, the combination of technologies from different sources may bring hidden risks that are neither fully investigated nor tested. To overcome this issue, many IT products/solution vendors such as IBM, Cloudera, MapR, and Hortonworks have developed their own modules and packaged them into core Apache Hadoop API and delievered improvised Hadoop distributios to the market. One of the goals is to ensure compatibility, security, and performance of all combined modules. Currently, the Hadoop system and different distributions use more than one security solutions for securing data and computation, as mentioned in the previous section, such as Apache Ranger, Apache Sentry, and Apache Knox. Each new module has its own set of vulnerabilities. If the system is needed to be flexible and interpretable, then automatically system will be vulnerable. Hadoop requires single security solution so that vulnerability can reduce that leads to attacks due to third party softwares and due to different modules used in Hadoop. The results of our the experiment shows the impact of attacks on performance in Table-III, as well as security breach possible due to vulnerabilities. Before implementing any module in Hadoop environment, risk management study of system and defense in depth is recommended.

9. Conclusion

Traditional enterprise security products implement security controls, but are still insufficient for holistically addressing the security challenges introduced by Big Data. To address the problems posed by data aggregation, organizations must find new ways to safeguard their critical tools, techniques, and procedures used to acquire, maintain, and analyze data of particular concern to improve the robustness of its security infrastructure, as most commercially available security software have access to all real and derived data available on the network. The add-on security features provided by the third party are not useful in the Big Data environment. Furthermore, analysts and other users of information technology need more effective security training that will help them understand the specific threats they face, that how their security choices will impact

the outcomes, and how to reconcile security objectives with mission objectives. Finally, system designers need to consider the security implications affecting user-facing tools. Users would benefit from features, giving them security-relevant feedback throughout their workflows, rather than using a separate security tool for forensic validation. Measuring the impact of these "little" security mechanisms may prove challenging and problematic. However, field studies such as those cited in this paper suggest that they can have a large impact on scaling security in a manner that paces with the scale of data that they protect. Thus, there is a need to upgrade the complete Hadoop system released with all security features without installing and to configure it separately.

Declaration of Competing Interest

The authors declare no conflict of interest.

References

[1] Dontha R. The origins of big data. Available: https://www.kdnuggets.com/2017/02/origins-big-data.html. [Accessed 2 January 2019].
[2] Hashem IAT, et al. The rise of "big data" on cloud computing: review and open research issues. Inf Syst 2015:98–115.
[3] Hurwitz JS, Nugent A, Halper F, Kaufman M. Big data for dummies. USA: wiley; 2013.
[4] Sharma S. Vulnerability – introducing V of big data. Available: https://www.datasciencecentral.com/profiles/blogs/vulnerability-introducing-10th-v-of-big-data. [Accessed 2 January 2019].
[5] Weiss R, Zgorski LJ. Obama administration unveils "big data" initiative: announces $200 million in new R&D investments. Washington, DC: The White House; 2012.
[6] Brauna TD, Siegel HJ. A comparison of eleven static heuristics for mapping a class of independent tasks onto heterogeneous distributed computing systems. J Parallel Distrib Comput 2001:810–37.
[7] Purcell BM. Big Data using cloud computing. Holy Family University Journal of Technology; 2013.
[8] Rouse M. Apache Hadoop YARN. Available: https://searchdatamanagement.techtarget.com/definition/Apache-Hadoop-YARN-Yet-Another-Resource-Negotiator. [Accessed 4 January 2019].
[9] Scaling. Available: https://stackoverflow.com/questions/11707879/difference-between-scaling-horizontally-and-vertically-for-databases/12349220. [Accessed 4 January 2019].
[10] XingWang. Powered by Apache Hadoop. Available: https://wiki.apache.org/hadoop/PoweredBy. [Accessed 4 January 2019].
[11] Columbus L. Forecast of Big Data market size, based on revenue. may 2018. p. 23. Available: https://www.forbes.com/sites/louiscolumbus/2018/05/23/10-charts-that-will-change-your-perspective-of-big-datas-growth/#1a9db88e2926. [Accessed 5 january 2019].
[12] SteveLoughran. Products that include Apache Hadoop or derivative works and commercial support. Available: https://wiki.apache.org/hadoop/Distributions%20and%20Commercial%20Support. [Accessed 5 January 2019].
[13] Gurjit singh Bhathal ASD. Big data solution: improvised distributions. In: Proceedings of the second international conference on intelligent computing and control systems. Madurai, India: ICICCS 2018; 2018.
[14] Fujitsu. FUJITSU technology solutions. Munich: Fujitsu; 2017.
[15] Amal Dahbur BMABT. A survey of risks, threats and vulnerabilities in cloud computing. Int J Cloud Appl Comput (IJCAC) 2011:11–5.
[16] Ye H, Cheng X, Yuan M, Xu L, Gao J, Cheng C. A survey of security and privacy in big data. In: 16th international symposium on communications and information technologies. ISCIT); 2016.
[17] Terzi DS, Terzi R, Sagiroglu S. A survey on security and privacy issues in big data. In: 10th international conference for internet technology and secured transactions. ICITST); 2015.
[18] Sharif A, Cooney S, Gong S. Current security threats and prevention measures relating to cloud services, Hadoop concurrent processing, and big data. In: IEEE international conference on big data; 2015. Washington, DC, USA.
[19] Derbeko P, et al. Security and privacy aspects in MapReduce on clouds: a survey. Comput Sci Rev 2016:1–28.
[20] MITRE Corporation. Apache vulnerability statistics. Available: https://www.cvedetails.com/vendor/45/Apache.html. [Accessed 7 January 2019].
[21] Yoder M. Cloudera's process for handling security vulnerabilities. Available: https://blog.cloudera.com/blog/2016/05/clouderas-process-for-handling-security-vulnerabilities/. [Accessed 8 January 2019].
[22] Bekker G. 2019 thales data threat report – global edition. Available: https://www.thalesesecurity.com/2019/data-threat-report. [Accessed 8 January 2019].
[23] Raj Samani CB. McAfee threats report 2018. Available: https://www.mcafee.com/enterprise/en-us/assets/reports/rp-quarterly-threats-dec-2018.pdf. [Accessed 2 February 2019].
[24] CyberPedia. An overview of DoS attacks. Available: https://www.paloaltonetworks.com/cyberpedia/what-is-a-denial-of-service-attack-dos. [Accessed 20 January 2019].
[25] Verizon. Data breach digest: perspective is reality. Available: https://enterprise.verizon.com/resources/reports/data-breach-digest/. [Accessed 2 February 2019].
[26] Incapsula. CROSS SITE SCRIPTING (XSS) ATTACK. Available: https://www.incapsula.com/web-application-security/cross-site-scripting-xss- attacks.html. [Accessed 4 February 2019].
[27] Millman R. Thousands of Hadoop clusters still not being secured against attacks. Available: https://www.scmagazineuk.com/thousands-of-hadoop-clusters-still-not-being-secured-against- attacks/article/637389/. [Accessed 5 February 2019].
[28] Apache. Apache Knox gateway 0.14.x user's guide. Available: https://knox.apache.org/books/knox-0-14-0/user-guide.html. [Accessed 6 February 2019].
[29] Sun D. Sentry tutorial. Available: https://cwiki.apache.org/confluence/display/SENTRY/Sentry+Tutorial. [Accessed 5 February 2019].
[30] A. S. Foundation. Apache ranger. Available: http://ranger.apache.org/index.html. [Accessed 5 February 2019].
[31] Andrew Wang CL. Project Rhino goal: at-rest encryption for Apache Hadoop. Available: https://blog.cloudera.com/blog/2014/06/project-rhino-goal-at-rest-encryption/. [Accessed 1 March 2019].
[32] Tom Yu e a. Kerberos: the network authentication protocol. Available: https://web.mit.edu/kerberos/. [Accessed March 2019].
[33] Bhathal GS, Singh A. Big data computing with distributed computing frameworks. In: Saini H, Singh R, Kumar G, Rather G, Santhi K, editors. Innovations in Electronics and Communication Engineering. Lecture Notes in Networks and Systems, vol. 65. Singapore: Springer; 2019. https://doi.org/10.1007/978-981-13-3765-9_49.
[34] Parmar RR, et al. Large-Scale Encryption in the Hadoop Environment: Challenges and Solutions. IEEE Access 2017;5:7156–63. https://doi.org/10.1109/ACCESS.2017.2700228.

Gurjit Singh Bhathal. Mr. Gurjit Singh Bhathal is an Assistant Professor in Department of Computer Science and Engineering in Punjabi University Patiala. He has over 18 years of experience in India and abroad in the field of Information Technology. His research interests in the area of security including, Cloud security, Big Data security and Cyber Security. He has completed his B.Tech in Computer Science and Engineering from SLIET Longowal and M.Tech from Punjabi University. He is a Microsoft Certified Engineer (MCSE) and IBM certified of Big Data and Data Privacy Fundamentals. He has more than 60 Publications including International journals, National and International conferences and three books in his credit. He has served as a system engineer in Toronto (Canada). He also served as a principal in college before joining the Punjabi University. He was also invited for Expert talk on Big Data in different colleges and universities in National level workshops. Mr. Gurjit Singh Bhathal recently awarded Outstanding Scientist in Computer Science and Engineering in 4th Annual Research Meet - ARM 2018