Asian Journal of Information Technology 16 (6): 561-568, 2017
ISSN: 1682-3915
© Medwell Journals, 2017

Comparative Study on Hadoop Distributed File System Based on Security Issues

[1]Hadeer Mahmoud, [1]Abdelfatah Hegazy and [2]Mohamed H. Khafagy
[1]Department of Computer Science, Faculty of Computers and Information,
Arab Academy for Science Technology and Maritime Transport, Cairo, Egypt
[2]Department of Computer Science, Faculty of Computers and Information,
Fayoum University, Faiyum, Egypt

Abstract: Cloud computing attracts a considerable attention in both academic and business areas in the meantime. In most markets, many cloud computing and BigData frameworks used a large scale of data processing. Hadoop is one of the best choices in open-source cloud computing and BigData framework used more and more in the business world. Recently, the weaknesses of security methods emerged as one of the main challenges interfering with development in general. Hadoop's writing and reading HDFS blocks also lack the support of security so, we need a solution to secure the Hadoop distributed file system. This study discusses the current security methods of Hadoop and provides a comparative view of BigData security mechanisms, its security property and performance; thinking carefully about some methods to improve its trustworthiness and security.

Key words: Security, BigData, Hadoop, HDFS, MapReduce, cloud

INTRODUCTION

Cloud computing is spreading through the IT world like wildfire. It continues to experience more widespread adoption. Cloud computing became a significant factor concerning the e-Commerce and e-Business processes. There are concerns among some CIOs concerning using cloud technologies intensely. Cloud computing offers significant cost savings by eliminating upfront expenses for hardware and software. Cloud computing offers significant cost savings by eliminating upfront expenses for hardware and software. Cloud computing is a general term for anything that involves delivering hosted services over the internet. Cloud computing is a model for enabling convenient on-demand network access to a shared pool of configurable computing resources, e.g., networks, servers, storage, applications and services. For users, cloud computing is a pay-per-use-on-demand mode that can conveniently access shared IT resources through the internet (Kumar and Goudar, 2012). Cloud computing services can be private, public or hybrid while cloud computing is yet in its initial stages but the advantages it has brought along are tremendous. Due to the lack of standards, users and computing industry is still reluctant to fully accept cloud computing concept. Let's discuss some of the challenges and issues such as; security, cost, performance, incompatibility, limited customization, uptime, vendor lock-in and compliance showed in Fig. 1. This study objective of highlighting the main security challenges that may affect BigData.

MATERIALS AND METHODS

Definition of BigData: BigData describes data sets with huge amount of size which makes it a tough challenge for any commonly used tools to manage, capture or process data within a short time. The size of BigData continuously increases as of 2012, it ranged from few terabytes to many beta bytes of data. Hence, it needs a set of techniques and technology to get insights from data sets that are complex and diverse (Snijders *et al*., 2012; Hashem *et al*., 2015). There are various explanations of BigData via. Vs. The 5 Vs are usually used to show of BigData as volume, velocity, variety, veracity and value (Chen and Zhang, 2014; Adluru *et al*., 2015). The volume describes the quantity of generated and stored data; velocity describes the speed at which the data is generated and processed to fulfill the demands; variety shows the type and nature of the data; veracity describes the quality of captured data that can vary greatly and value provides outputs for gains from large data sets. Identifying characteristics of the data are helpful in extracting its hidden patterns. BigData is classified into 10 categories (Fig. 2).

Corresponding Author: Hadeer Mahmoud, Department of Computer Science, Faculty of Computers and Information, Arab Academy for Science Technology and Maritime Transport, Cairo, Egypt

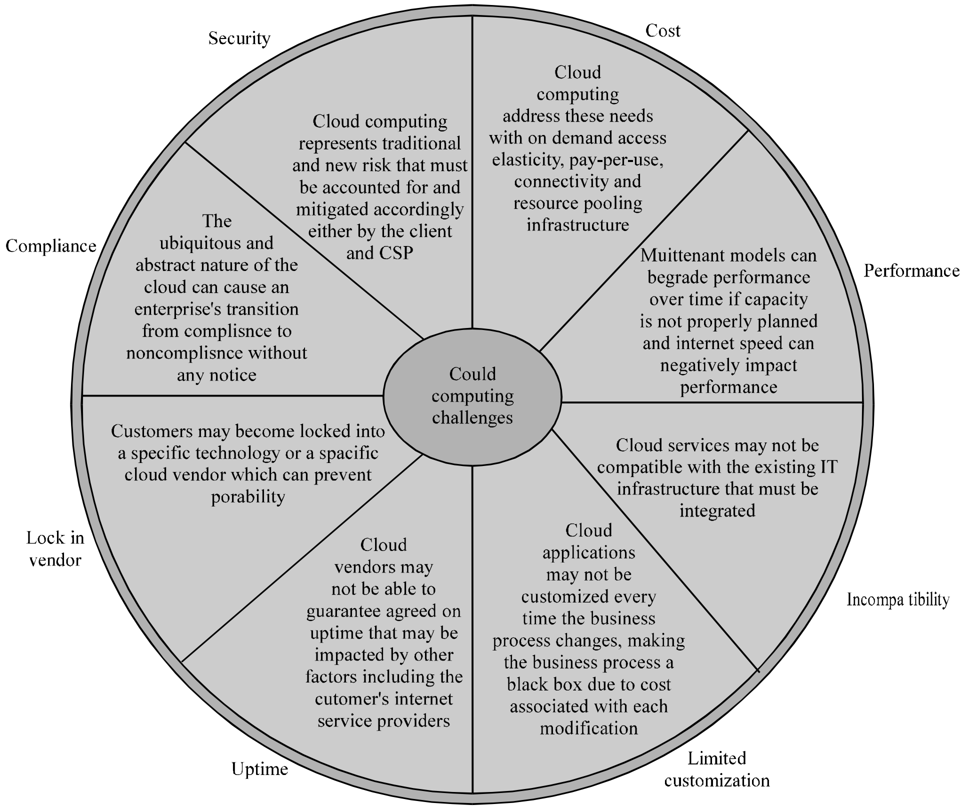
Security
Cost
Cloud computing represents traditional and new risk that must be accounted for and mitigated accordingly either by the client and CSP
Cloud computing address these needs with on demand access elasticity, pay-per-use, connectivity and resource pooling infrastructure
Compliance
The ubiquitous and abstract nature of the cloud can cause an enterprise's transition from complisnce to noncomplisnce without any notice
Muittenant models can begrade performance over time if capacity is not properly planned and internet speed can negatively impact performance
Performance
Could computing challenges
Customers may become locked into a specific technology or a spacific cloud vendor which can prevent porability
Cloud services may not be compatible with the existing IT infrastructure that must be integrated
Lock in vendor
Cloud vendors may not be able to guarantee agreed on uptime that may be impacted by other factors including the cutomer's internet service providers
Cloud applications may not be customized every time the business process changes, making the business process a black box due to cost associated with each modification
Incompa tibility
Uptime
Limited customization

Fig. 1: Cloud computing challenges

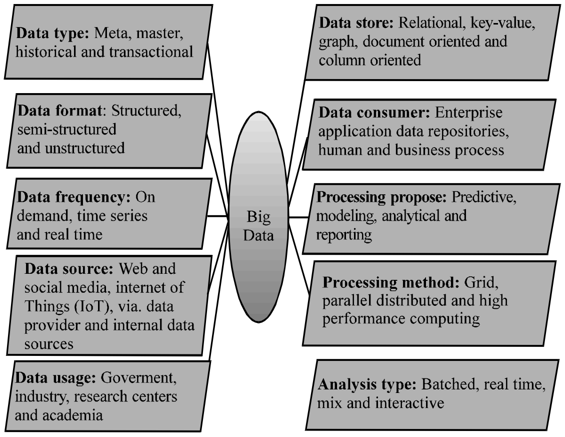
Data type: Meta, master, historical and transactional
Data format: Structured, semi-structured and unstructured
Data frequency: On demand, time series and real time
Data source: Web and social media, internet of Things (IoT), via. data provider and internal data sources
Data usage: Goverment, industry, research centers and academia
Big Data
Data store: Relational, key-value, graph, document oriented and column oriented
Data consumer: Enterprise application data repositories, human and business process
Processing propose: Predictive, modeling, analytical and reporting
Processing method: Grid, parallel distributed and high performance computing
Analysis type: Batched, real time, mix and interactive

Fig. 2: BigData's classification

APACHE Hadoop: Hadoop is an open-source software framework used for distributed storage and processing of the very large datasets. The Hadoop Map-reduce framework is responsible for sorting any outputs of the maps and then it is sent to the reduce tasks. Modules are the included in this project (TASF., 2013).

Hadoop common: Refers to the collection of common utility and libraries that support other modules.

Hadoop Distributed File System (HDFS^{TM}): A distributed this file system that the provides a very high through pu access to the application's data.

Hadoop YARN: This is a framework for jobs scheduling and cluster resource management.

Hadoop MapReduce: A YARN-based system responsible for parallel processing large data sets.

Hadoop distributed file system: HDFS has a master architecture (Fig. 3) that consists of a single master server named NameNode which stores Metadata that manages the file system namespace and controls the access to the files used by clients. There are a number of DataNodes; each cluster contains a single node which manages the storage of these nodes. HDFS creates a file system namespace and allows user's data to be stored in files. Every file consists of one or more blocks that are stored in a set of DataNodes. The NameNode runs the operations of file system namespace such as opening,

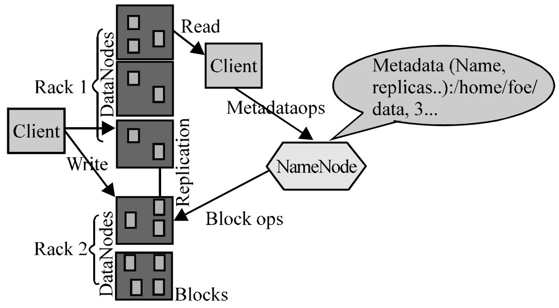

Fig. 3: Structure of HDFS files system

closing and the renaming files and directories. It maps also the blocks to DataNodes. The DataNodes are responsible for serving read and write requests from the file system's clients. The DataNodes are responsible for block creation, deletion and replication upon instruction from the NameNode.

The NameNode and DataNode are pieces of software designed to run on commodity machines. These machines typically run a GNU/Linux operating system. HDFS is built using the Java language in other words, any machine supports Java can run the NameNode or the DataNode Software (IDGCI., 2017).

Hadoop MapReduce: Hadoop MapReduce is a software framework for smooth writing applications which process large datasets with a parallel processing on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. The dataset splits into chunks by a MapReduce job, processed by the mapping procedure that performs filtering and sorting in a completely parallel manner. The framework sort's the outputs of the maps which are then inputted to the reduction procedure performs a summary operation.

Typically, both the input and the output of the job are stored in a file-system. The MapReduce framework consists of one master JobTracker and one slave Task Tracker for each cluster node. JobTracker is responsible for scheduling any job component tasks on these slaves. Usually, it monitors and re-executes any failed task (Zhao *et al*., 2014).

RESULTS AND DISCUSSION

BigData security: The aims of the study security of BigData to identify the security challenges that the companies and banks are suffering when implementing BigData solutions from infrastructures to analytics applications and how those are alleviated. A BigData security analytics solution consists of three basic characteristics (ENISA, 2015), analytical flexibility, scale and performance are discussed in Fig. 4.

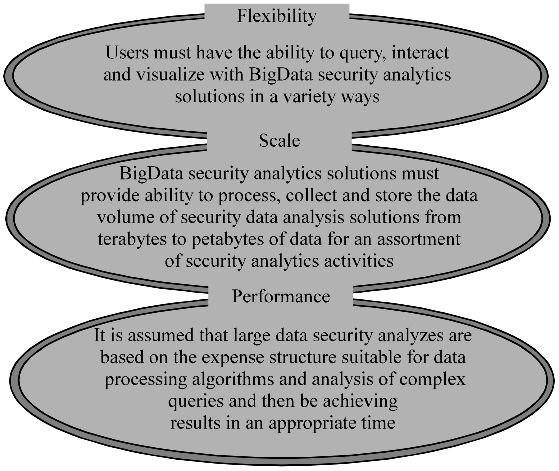
Flexibility
Users must have the ability to query, interact and visualize with BigData security analytics solutions in a variety ways
Scale
BigData security analytics solutions must provide ability to process, collect and store the data volume of security data analysis solutions from terabytes to petabytes of data for an assortment of security analytics activities
Performance
It is assumed that large data security analyzes are based on the expense structure suitable for data processing algorithms and analysis of complex queries and then be achieving results in an appropriate time

Fig. 4: A BigData security analytics solution

Security issues: There are some challenges for the management of large data where ineffective tools used to put them in an unsafe way, private and public databases have a lot of weaknesses and threats, volunteered and unexpected leakage of data makes hackers access to the resources they needed easy (Rein and Willis, 1976). In the framework of distributed programming, the security issues start when there is a large amount of private data stored in databases in a regular format or unencrypted. With a non-authorized people it is difficult to secure the data when it transferred from homogeneous data to heterogeneous data because some tools not developed with security certificates and policy of massive data.

Security property: Security mechanisms used to achieve the cryptography goals, there are 5 main objectives of cryptography. Every security system must provide a set of security functions guarantees the system's security. We refer to these functions as objectives. These objectives can be enrolled under the following 5 main categories in Fig. 5.

Authentication: The process of proving one's identity. This means that before the system sends and receives data, the receiver and sender identity should be verified.

Confidentiality: Ensures that only the intended receiver reads the message and no one else does, this is usually how most people identify a secure system.

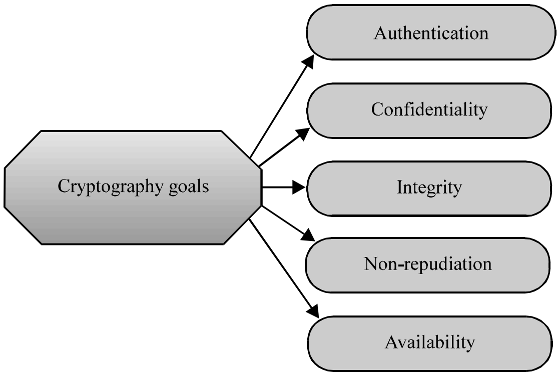

Fig. 5: Cryptography goals

Integrity: Ensures that the message received by the user was not altered or manipulated with. The basic form of the integrity is packet checksum in Ipv4 packets.

Non-repudiation: A way proving that the sender is really the one who sent the message; meaning that both the sender and the receiver cannot deny they have sent/received a certain message.

Service reliability/availability: Since, the main problem with most systems is getting hacked by an intruder, who can cause a downtime in availability such systems provide a way to provide their users with the quality of service they expect.

Hadoop security: Security by default Hadoop runs in a non-secure mode which has no actual authentication. By configuring Hadoop to run in secure mode each user and service needs to be authenticated by Kerberos in order to use Hadoop services. Security can be accomplished at different layers in a traditional data management software/hardware stack, e.g., file-system-level encryption; this option provides high performance and application transparency and it is typically easy to deploy but it is not able to model few application-level policies. For example, multi-tenant applications might seek encrypting based on the end user and a database might want to use a different type of encryption for each column stored within a single file.

Security techniques in HDFS: We should describe the latest technology and the mechanisms used to secure the Hadoop distributed file system it includes many security methods like authentication, encryption, decryption, compression to store BigData securely and to show efficiency and complexity affected of each algorithm, including some techniques that assure the security of the whole system.

Integrating hybrid encryption schemes and HDFS: Now a days, cloud computing became widespread because it's an open source framework for Hadoop and also became widely used for large-scale data processing. However, data confidentiality was enhanced by some little studies of Hadoop against storage servers. In this study, the data confidentiality issue will be addressed by integrating hybrid encryption schemes and the Hadoop Distributed File System (HDFS). The study applies integration of HDFS-RSA and HDFS-pairing implementation, as extensions of HDFS. The performance overhead of HDFS-RSA and HDFS-pairing has been measured by some experiments.

After this integration the experiments showed an acceptable overhead on reading operations and considerable overhead on writing operations. The implementation of this integration is appropriate for write-once-read-many applications.

This study modifies the reading and writing operations in the fuse-DFS module in detail described in two implementations:

- HDFS-RSA uses the open SSL toolkit
- HDFS-pairing uses the PBC library

This integration was planned to be a solid step towards guarantee data confidentiality in Hadoop.

Encrypted HDFS: In large data processing Hadoop is the most commonly used and distributed programming framework where file storage increases quickly and secure computing is needed. However, encryption of storing HDFS blocks is not supported in the current Hadoop, which is a fundamental solution for secure Hadoop. The study introduces an encryption and decryption techniques in HDFS to secure Hadoop architecture.

This study secures HDFS by implementing the AES encrypt/decrypt class and adds to the compression codec in Hadoop (Park and Lee, 2013).

Experiments on Hadoop showed that the representative MapReduce job on encrypted HDFS generates affordable computation overhead <7%.

File management in generic Hadoop has weakness in its security. Though encryption is the essential file protection method, its real implementation has not been fully examined.

The study shows that it encrypts and decrypts the file before it is written or read in the secure HDFS. A read or write request from a client triggers a decryption or encryption function to HDFS blocks at each DataNode by using 128 bit AES with ECB mode that is compatible with HDFS blocks. The study revealed that they planned to secure MapReduce job in HDFS with marginal performance degradation <7%, through implementing AES Codec into Hadoop.

Triple encryption scheme for Hadoop-data security: In past years, cloud computing has been growing because it allows services to scale on demand to perform automatic failure recovery and it can provide users with reliable, flexible and low-cost services. Data security is the main issue to protect the cloud applications because it is now available for any user. In order to ensure data security protection in the cloud a novel triple encryption scheme is proposed in this study which combines HDFS file encryption using DEA (Data Encryption Algorithm) and using RSA to encrypt data key and then using IDEA (International Data Encryption Algorithm) to encrypt the user's RSA private key.

In the study implementation of triple encryption scheme using the hybrid encryption to encrypt HDFS files used RSA and DES based on IDEA to encrypt user's RSA private key then it integrated into Hadoop based cloud data storage (Yang *et al*., 2013).

The principle of data hybrid encryption is that hybrid encryption method encrypted HDFS files; a HDFS file is symmetrically encrypted with a unique key k and the key k is then asymmetrically encrypted by owner's public key. Symmetrical encryption is secure and more expensive cost than asymmetrical encryption. Hybrid encryption is a compromising choice against the two forms of encryption above. Hybrid encryption uses DES algorithm to encrypt files and get the data key and then uses RSA algorithm to encrypt the data key. The user keeps the private key in order to decrypt the data key.

They planned to improve the performance of data encryption and decryption using MapReduce to make encryption and decryption parallel.

BigData security and privacy: BigData security and privacy became very important because of all technologies started to use BigData. In this study discuss the challenges in maintaining the privacy and security of BigData.

Random encryption techniques and authentication of the distributed data access and storage of BigData are implemented to achieve privacy and security. That makes the system strong while maintaining the performance standards. Securing the system to the highest level is the priority because access to these huge volumes of private data might be very damaging when misused. Proposed mechanisms; there are privacy and security risks for the Hadoop eco system as the NameNode and the DataNode have the entire control over the data. The user has no control over the data; there is a need for establishing a trust (Xu *et al*., 2012) between the user and the NameNode.

To make the system more secure paper needs to implement the randomized encryption techniques on the data that are implemented in this system; namely; RSA, the Rijndael, AES and the RC6.

They planned to achieve privacy and security of the information stored in the cloud by using the Hadoop system.

BigData security perspective: BigData is the analysis of large sets of data generated from user data, sensor data, medical and enterprise data. The hadoop is responsible for managing, store and distribute BigData across several server nodes. This study discussed the security is important of BigData and focused more on security issue arising in a hadoop architecture base layer that is called Hadoop Distributed File System (HDFS) .

The HDFS security is enhanced by using three approaches like Kerberos Mechanism, Bull Eye algorithm and NameNode approach.

This study shows the BigData, information and focused on increasing security of BigData by implementing three approaches (Saraladevi *et al*., 2015) occurs in the NameNode and DataNode.

They planned to secure the BigData by applying any/all these approaches in Hadoop Distributed File System (HDFS) which is the base layer in Hadoop where it contains a large number of blocks.

Security frame work in G-Hadoop: G-Hadoop is an extension of the Hadoop MapReduce framework with the functionality of allowing the MapReduce tasks to run on multiple clusters in a grid environment. However, G-Hadoop uses job submission mechanism and user authentication as Hadoop. G-Hadoop does not support the grid environment because it is designed for a single cluster. The Secure Shell (SSH) protocol used in G-Hadoop prototype to secure the connection between the user and the target cluster. This study designed a novel security framework for G-Hadoop to secure the connection between the user and each participating cluster.

The study proposes a new security model for G-Hadoop. This model consists of several security solutions like the SSL (Secure Sockets Layer) protocol and public key cryptography or GSI (Globus Security Infrastructure). A single-sign-on the approach is used to simplify the process of user authentication and Job submission of the current G-Hadoop.

The work flow of the authentication procedure is shown in Fig. 6 at which authentication procedure and the

Table 1: Classification of BigDatasecurity studies

				Security property			
Work/Year	Methods	Purpose/Strengths	Limitations/Weakness	Confidentiality	Integrity	Availability	Authentication
Data Security in HDFS (Shetty and Manjaiah, 2016)	Implement secure HDFS by implemented method in which OAuth	Security and enhance file encryption upload time in HDFS	Low security techniques for securing data		√		√
Saraladevi *et al.* (2015) BigData and Hadoop-A study in security perspective	Kerberos mechanism, Bull Eye algorithm and NameNode approach	Enhances HDFS security by using three approaches	Increases the security at only Hadoop base layer (HDFS) so, three approaches didn't use in other layers of Hadoop technology			√	√
Zhao *et al.* (2014) a security framework in G-Hadoop for big data computing a cross distributed Cloud data centers (IDGCI, 2017)	SSL and public key cryptography algorithm or the concepts of other security solutions such as GSI	New security model for G-Hadoop	Applies only a public key cryptographic algorithm in this work SSL cannot guarantee the absolute security		√	√	√
Park and Lee (2013) secure Hadoop with encrypted HDFS	Implement secure HDFS by adding the AES algorithm by using ECB mode to encrypt/decrypt class to compression codec in Hadoop	Encrypt HDFS generates affordable computation overhead <7%	Using ECB, this mode data are divided into 128-bit blocks and each block is encrypted one at a time. Separate encryptions with different blocks are totally independent of each other	√		√	
Yang *et al* (2013) a novel triple encryption scheme for Hadoop-based cloud data security	This combines HDFS file encryption using DEA and the data key encryption with RSA and then encrypts the user's RSA private key using IDEA	Secure data in cloud data storage by using a novel triple encryption scheme	Limited performance because it doesn't use the parallel processing of the encryption and decryption using MapReduce	√	√		√
Lin *et al.* (2012) toward data confidentiality dentiality via. integrating hybrid encryption schemes and HDFS	Implement two integrations HDFS-RSA and HDFS-pairing	Achieves data confidentiality for Hadoop by using integrations providing alternatives	HDFS-RSA and HDFS-pairing (asymmetric) are slower in writing operation than the reading operation Not suitable for applications with many-write and few-read operations	√			

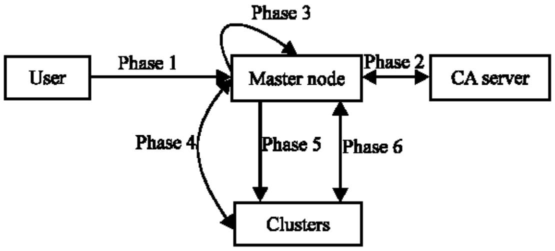

Fig. 6: The work flow of the authentication procedure

interaction between the components of the designed security framework. This workflow includes some main phases: user authentication; proxy credential assignment; preparing authentication information on the master node; authentication of the master node and slave nodes job execution; termination and disconnection.

With these security mechanisms the designed security framework has the ability to prevent the most common attacks such as MITM attack, replay attacks and delays attack and to ensure a secure communication of G-Hadoop over public networks.

They planned to protect the G-Hadoop system from any attacks using security mechanisms based on SSL and Cryptographic algorithm or concepts of security solutions such as GSI. The security provides a complete solution and trust full of the single-sign-on process for the user to access G-Hadoop.

Data security in HDFS: Hadoop is become the most popularly used distributed programming framework for processing large data with Hadoop Distributed File System.

The study proposes to secure Hadoop Distributed File System is implemented using encryption/decryption of data which is to be stored in HDFS. Some algorithms used for encryption/decryption in real time. AES algorithm results using encryption into file size increased to double of original file and hence file upload time also increases. The technique used removes this drawback in this project.

The technique (Shetty and Manjaiah, 2016) used to enhance file uploads time, decrease file encryption size using OAuth is an Open Authentication Protocol that is achieve authentication. The authorization token used for secure data on HDFS by a Key that generated by it. They planned to Secure data and provide unique authorization token for each user. this project enhanced file uploaded time using OAuth does achieve the authentication.

Finally, the summary of the comparative study of the method, strengths, weaknesses and security property of the techniques used in securing the HDFS are described in Table 1.

CONCLUSION

This survey addresses the latest situation in the security of APACHE Hadoop and its performance. It revealed some challenges in implementing security mechanism for Hadoop and the way of enhancing the security level and achieving the performance of different types of the mechanisms. At last, it compared between methods of BigData security and privacy.

Security of BigData is the most important matter to be addressed more in the future so we need to develop new techniques using different algorithms such as RC6, Blowfish, 3DES and ETEA. This system requires improving existing technologies to get accurate results and maintaining security of BigData handled on a different platform. We hoped that this study would help to understand the latest security level of BigData and its ecosystem better in order to enhance it in the future.

AKNOWLEDGEMENTS

The researchers would like to thank all those who contributed to this study. Further to this, suggestions that have improved the presentation, correctness and quality of this study.

REFERENCES

Adluru, P., S.S. Datla and X. Zhang, 2015. Hadoop eco system for BigData security and privacy. Proceedings of the Conference on Systems, Applications and Technology Conference (LISAT), May 1-1, 2015, IEEE, Farmingdale, New York, USA., ISBN:978-1-4799-8643-9, pp: 1-6.

Chen, C.L.P. and C.Y. Zhang, 2014. Data-intensive applications, challenges, techniques and technologies: A survey on BigData. Inform. Sci., 275: 314-347.

ENISA, 2015. BigData Security Good Practices and Recommendations on the Security of BigData Systems. ENISA Publisher, Heraklion, Greece, ISBN:9789292041427.

Hashem, I.A.T., I. Yaqoob, N.B. Anuar, S. Mokhtar, A. Gani and S.U. Khan, 2015. The rise of BigData on cloud computing: Review and open research issues. Inform. Syst., 47: 98-115.

IDGCI., 2017. Cisco Nexus 9516 data center switch aces a grueling high-density stress test. IDG Communications Inc, New York, USA. http://www.networkworld.com/article/2224394/cisco-subnet/defining-big-data-security analytics.html.

Kumar, S. and R.H. Goudar, 2012. Cloud computing: Research issues, challenges, architecture, platforms and applications; A survey. Intl. J. Future Comput. Commun., 1: 356-360.

Lin, H.Y., S.T. Shen, W.G. Tzeng and B.S.P. Lin, 2012. Toward data confidentiality via. integrating hybrid encryption schemes and Hadoop distributed file system. Proceedings of the IEEE 26th International Conference on Advanced Information Networking and Applications (AINA), March 26-29, 2012, IEEE, Fukuoka, Japan, ISBN:978-1-4673-0714-7, pp: 740-747.

Park, S. and Y. Lee, 2013. Secure hadoop with encrypted HDFS. Proceedings of the 8th International Conference on Grid and Pervasive Computing (GPC 2013), May 9-11, 2013, Springer, Seoul, Korea, pp: 134-141.

Rein, T. and H.W. Willis, 1976. Privacy and Security Issues in Information System. The Rand Corporation Publisher, Santa Monica, California.

Saraladevi, B., N. Pazhaniraja, P.V. Paul, M.S. Basha and P. Dhavachelvan, 2015. Big data and Hadoop: A study in security perspective. Procedia Comput. Sci., 50: 596-601.

Shetty, M.M. and D.H. Manjaiah, 2016. Data security in Hadoop distributed file system. Proceedings of the International Conference on Emerging Technological Trends (ICETT), October 21-22, 2016, IEEE, Kollam, India, ISBN: 978-1-5090-3752-0, pp: 1-5.

Snijders, C., U. Matzat and U.D. Reips, 2012. Big Data: Big gaps of knowledge in the field of internet science. Int. J. Internet Sci., 7: 1-5.

TASF., 2013. HDFS architecture guide. The Apache Software Foundation, Forest Hill, Maryland. https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html#Introduction.

Xu, L., X. Wu and X. Zhang, 2012. CL-PRE: A certificateless proxy re-encryption scheme for secure data sharing with public cloud. Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security, May 02-04, 2012, ACM, New York, USA., ISBN:978- 1-4503-1648-4, pp: 87-88.

Yang, C., W. Lin and M. Liu, 2013. A novel triple encryption scheme for hadoop-based cloud data security. Proceedings of the 4th International Conference on Emerging Intelligent Data and Web Technologies (EIDWT), September 9-11, 2013, IEEE, Xi'an, China, ISBN:978-1-4799-2141-6, pp: 437-442.

Zhao, J., L. Wang, J. Tao, J. Chen and W. Sun *et al*., 2014. A security framework in g-hadoop for big data computing across distributed cloud data centres. J. Comput. Syst. Sci., 80: 994-1007.