# Innovative Methodology for Elevating Big Data Analysis and Security

*Sahel Aloune , Ismail Hababeh, Feras Al-Hawari, Tamer Alajrami*
School of Electrical Engineering and Information Technology- Computer Engineering Dept.
Amman, Jordan
{sahel.alouneh, ismail.hababeh, feras.alhawari, t.alajrami}@gju.edu.jo

***Abstract*— big amount of data and information transfer among, within and through organizations all over the globe. This big data contains a great deal of sensitive, confidential and restricted data, like financial, legal or private information. Any loss, threat, or leakage may cause high-risk effects on such data. Securing big data during analysis phase is still a challenge in cloud systems. This paper proposes a methodology to protect big data during analysis, by classifying data before any action such as moving, copying or processing take place. Based on big data classification, the security procedures will be activated according to data level criticality.**

***Keywords*— big data; data classification; encryption threats; risks; data loss; data Leakage**

## I. INTRODUCTION

The Big data [7] indicate that we deal with large amount of structured, unstructured and semi structured data produced in different types. We usually describe big data when data expands to petabytes, exabytes, or youtabytes of volume. Big data includes all fields of knowledge; i.e. science, business, art, etc. which increase big data complexity.

The Big data complexity makes it difficult to be managed, queried or analyzed through traditional tools and mechanisms. The complexity of big data is characterized by four main factors namely, volume, velocity, variety and veracity [8].

- Volume refers to huge amount of data we deal with like petabytes, exabytes, and youtabytes.
- Velocity refers to the rapid growing rate of data generation and processing, specifically in real time systems.
- Variety refers to different forms of data, like online transactions, emails, videos, audios, pictures, medical records, social networking interface, science applications, search interrogation, and many more.
- Veracity refers to the confidentiality, integrity and availability of the data which includes data security and privacy.

Big data is different from the data being stored in traditional warehouses. The warehouse data need to be cleansed, documented and also trusted. Furthermore, it should fit the basic structure of that warehouse to be stored. On the other hand, big data is not only handles the data being stored in traditional warehouses, but also deals with structured, unstructured, and semi-structured data that needs different techniques and tools to be processed and analyzed.

On the other hand, big data analysis is used to make accurate decisions in time the data becomes necessary and should be available in accurate, complete and timely manner. This make big data management and governance process bit complex, in addition to the necessity to make data open and available in standard forms and formats. Thus, finding solutions through new techniques and tools lead to better decision making, business intelligence and productivity improvements.

Sophisticated analytics can substantially improve decision making, minimize risks, and discover valuable insights. Big data decision making can replace human decision making with automated algorithms and machine learning.

The era of Big Data could yield new management principles [3]. In the early days of professionalized corporate management, leaders discovered that minimum efficient scale was a key determinant of competitive success. Similarly, future competitive benefits are likely to accrue to companies that can not only capture more and better data but also use that data effectively at scale [9].

Security issues are one of the key challenges facing big data where the organization store and process a huge amount of sensitive and confidential information about their customers and employees, trade secrets and financial information.

Storing and processing data in one place may make these data a precious target for attackers, which can leave large amount of information exposed and unprotected. If there is any manipulation or sabotage, this could cause loss of confidence in the organization and damage their image and reputation. Because all of these risks, the big data stores must properly controlled and protected through new innovative privacy and security methods.

The increasing trend of big data leads to new data threats, security concerns and risks. This could be happen especially when we deal with sensitive and critical data such as trade secrets and financial information [14]. As we look to gain value from such information, we also seek for protecting and saving sensitive and critical data.

This paper presented an integrated methodology for big data classification and security. Big data is classified based on

978-1-5090-4580-8/16/$31.00 

its criticality and sensitivity, and then the high critical and sensitive data go through protection and encryption technique before applying any transaction or action such as data migration, duplication and analysis.

## II. Related Work

In this section, we discuss some of the main big data techniques proposed in the literature. We focus here especially on the security aspects of big data related work.

Author in [1] describes big data advantages and how possible to capture, process, and share huge amounts of data. In addition, this research presents how to prevent crime, manage emergences and identify new business opportunities. This approach considered data security and privacy the critical requirements during analysis and transmitting all dimensions related to big data such as volume, verity velocity and value are presented and discussed. Big data security is addressed and implied that any comprehensive data security solution must meet the confidentiality, integrity and availability.

Authors in [2] describe the internet and cybersecurity, identify most cyber-attacks such as spamming, search poisoning botnet denial of service and fishing malware. In addition, this paper presents cybersecurity terms and define cybersecurity goals and solution, then describe the importance of cybersecurity in the presence of big data within computer networks which prevents hackers to enter any network.

The authors in [3] present big data, data mining, compare between big data features, and propose attribute selection methodologies for protecting big data values. They conclude that extracting valuable information is the main goal of analyzing big data which need to be protected. The authors believe that the big data and cloud computing are major trends of modern computer technology. In addition, they define the big data as a technique to extract the value from huge data beyond the processing capabilities of existing databases. According to their conclusion, the authors claim that it is not possible to protect all data values inside the big data. Therefore, a methodology of selecting protected attributes big data security is created and discussed.

The authors in [4] proposed a security compliance model that presents security and access control features which are employed at the time of origin of Big Data. They believe that big data contain enormous potential and lucrative information that should be secured, audited and monitored. The main challenge issue in this research is the security and privacy as the size of big data continues to grow exponentially without any access control mechanism. However, the security that provided for big data is limited due to its exponential increasing trend.

The Authors in the [5] present an overview of big data substance include variety, volume, procedures, security challenges and privacy issues. They describe a big data as any large amount of structure and unstructured data. This data is produced from online transactions, emails, videos, audios, picture, posts, search interrogation, medical records, social networking interface, and science applications.

Authors in [6] present big data and big data characteristics "3V" that add a new characteristic value, veracity and the relation between big data and information security. In addition, information security and privacy are among the most challenge issues of big data in which big data analytics provide significant opportunities for solving different information security problems. They introduce Hadoop technologies, components and methodology as a tool for big data analysis, and discuss big data security threats and conclude that traditional security tools are not adequate for big data.

## III. Our Proposed Methodology

In this section, the methodology software and hardware requirements are introduced. Hadoop 2.6 [10] is installed on Linux (Ubuntu 16.04) [11] for single and multi-node cluster. The data work flow through hadoop is described in the following subsections.

### *A. Operating System*

A Debian-based Linux operating system (Ubuntu 16.04) is installed on each name or data node which requires 4 GB RAM ,100 GB HD and dual core processor.

### *B. Maintaining the Integrity of the Specifications*

Hadoop is an open source java-based programing framework that supports process and store huge amount of data in a distributed computing environment. Hadoop it is one of apache project products that uses a distributed file system which facilitates rapid data transfer rates among nodes and maintain system fault tolerant. Hadoop was inspired by Google's MapReduce [12]; a software framework in which an application is broken down into numerous small parts. Any of these parts can be run on any cluster-node in the cloud system.

Hadoop have two main components; MapReduce (Resource management, Processing Data) and Hadoop Distributed file system HDFS (Store Data).

### *C. MapReduce*

Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable fault-tolerant manner [12]. A MapReduce job is usually splits the input data-set into independent chunks which are processed in a completely parallel manner. The MapReduce framework consists of a single master JobTracker and one slave TaskTracker per cluster-node.

### *D. Hadoop Distributed File System*

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware [13]. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant (add more details and examples). HDFS is highly fault-tolerant and designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets.

978-1-5090-4580-8/16/$31.00 ©2016 IEEE

### E. HDFS Architecture

HDFS builds on a 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. 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 open, close, and rename 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

Hadoop initially launched in MR1 model that is described in Figure 1.

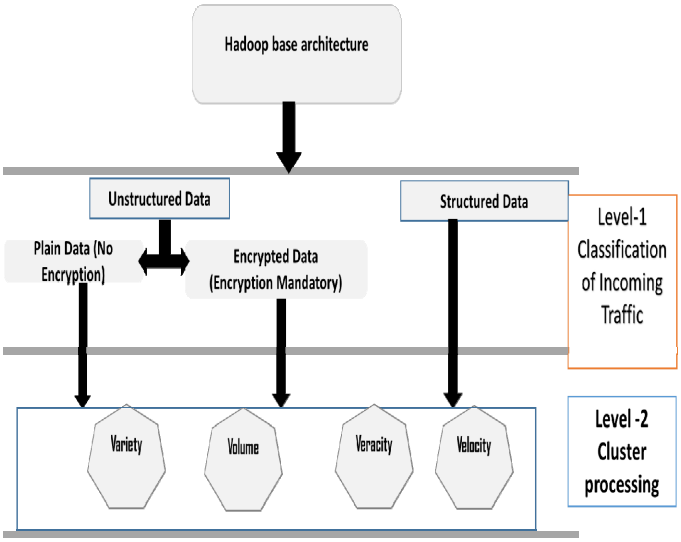


Fig.1. Hadoop Generation base platform model

## IV. CLUSTERING DATA NODES

In the case of a complex data, where zetabytes have to be processed, a huge communication time is required to perform the queries, which will increase the system load and decrease the system performance. However, grouping big data into clusters reduces the communication between nodes and then enhances the cloud system performance.

The clusters average communication is used to represent the communication value for all nodes in each cluster that is required for big data analysis processes and data storage.

The proposed clustering technique creates disjoint data clusters according to the least average communication cost between the big data nodes. Moreover, this technique helps to speed up the processes of big data analysis, thus minimizing the communication costs and improving the cloud performance.

### A. Clustering Parameters

The parameters that will be considered for the proposed clustering technique are described as follows:

• Cluster: Logical place that used to group big data nodes together.

• Big Data Nodes: Set of cloud data nodes.

• Clustering Range: The communication time needed for grouping two or more data nodes.

• Data Nodes Communication: The time needed for two data nodes to be communicated with each other.

Our clustering method categorizes the cloud data nodes according to the communication range and the communication time between data nodes.

### B. Clustering Algorithm

The communication between two data nodes is computed as the cost of creation the data packet in addition to the cost of transmitting the data packet from one node to the other. Each data node is included in one and only one cluster that shows the minimum average communication with other nodes.

*Repeat*
*For I = 1 to the number of big data nodes in the cloud*
*For J = 1 to the number of big data nodes in the cloud*
*If I ≠ J and communication between node I and node J <= Clustering Range Then*
*Node I and Node J are grouped together (at the same cluster)*
*EndIf*
*EndFor*
*EndFor*
*Until all cloud nodes system have been processed*

The proposed clustering algorithm is detailed as follows:

### C. Case Study

For simplicity, a connected cloud topology consists of 12 nodes, supported by database system are proposed to simulate and test our clustering method performance. 4 nodes out of the 12 are used as name nodes for executing and developing database applications, and the remaining 8 nodes are used as data nodes for facilities which produce administrational and financial information. Table I shows an example of the proposed communication between the nodes.

978-1-5090-4580-8/16/$31.00 ©2016 IEEE

TABEL I. Communication between 12 cloud nodes

| node # | node 1 | node 2 | node 3 | node 4 | node 5 | node 6 | node 7 | node 8 | node 9 | node 10 | node 11 | node 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| node 1 | 0 | 6 | 11 | 10 | 7 | 8 | 9 | 9 | 12 | 12 | 11 | 12 |
| node 2 | 6 | 0 | 10 | 10 | 2 | 3 | 4 | 2 | 7 | 8 | 6 | 7 |
| node 3 | 11 | 10 | 0 | 6 | 6 | 7 | 8 | 7 | 2 | 3 | 10 | 11 |
| node 4 | 10 | 10 | 6 | 0 | 8 | 7 | 6 | 7 | 9 | 6 | 1 | 1 |
| node 5 | 7 | 2 | 6 | 8 | 0 | 1 | 2 | 2 | 7 | 6 | 9 | 12 |
| node 6 | 8 | 3 | 7 | 7 | 1 | 0 | 1 | 2 | 8 | 8 | 6 | 8 |
| node 7 | 9 | 4 | 8 | 6 | 2 | 1 | 0 | 3 | 6 | 6 | 7 | 6 |
| node 8 | 9 | 2 | 7 | 7 | 2 | 2 | 3 | 0 | 8 | 7 | 6 | 6 |
| node 9 | 12 | 7 | 2 | 9 | 7 | 8 | 6 | 8 | 0 | 1 | 6 | 7 |
| node 10 | 12 | 8 | 3 | 6 | 6 | 8 | 6 | 7 | 1 | 0 | 7 | 6 |
| node 11 | 11 | 6 | 10 | 1 | 9 | 6 | 7 | 6 | 6 | 7 | 0 | 2 |
| node 12 | 12 | 7 | 11 | 1 | 12 | 8 | 6 | 6 | 7 | 6 | 2 | 0 |

Clusters can be generated by performing the proposed clustering algorithm on the given communication between nodes at clustering range, for example, equals to 5 ms/byte. Four clusters are generated to hold the 12 nodes, each cluster has different set of nodes, and each node belongs to only one cluster.

Applying the algorithm on the given nodes, the clusters that are generated are determined in Table II.

TABLE II. Generated Clusters

| Cluster # | Node # |
|---|---|
| C1 | N1 |
| C2 | N2, N5, N6, N7, N8 |
| C3 | N3, N9, N10 |
| C4 | N4, N11, N12 |

## V. PROTECTING BIG DATA

The big data administrators should be able to allocate their data freely and easily into cloud data nodes. However, this data allocation may include threats to the security of the big data. In this research, we propose an integrated method that stores the sensitive data in an encrypted form known only to the big data administrators and secures sensitive data allocation processes. Encryption techniques that are used to secure big data traffic are based on well-known techniques such as SHA-1 for encrypting big data traffic in the clusters. Key management technique used for asymmetric keys exchange is built upon Public Key Infrastructure (PKI) algorithm.

Our method protects against attacks that try to employ the data allocation process for stealing sensitive data. The proposed method builds on the Hadoop Distributed File System HDFS and addresses the security concerns of the sensitive data allocation processes.

We assume the trust between the source nodes cluster SNC and the target nodes cluster TNC. This can be done by

*1. The SNC sends to the TNC the identification data, such as data block IDs and Cluster Data Node addresses in an encryption form.*
*2. The TNC generates block access token and encrypts them by using the same encryption technique used in the SNC*
*3. Shares these block access token with its data nodes.*
*4. The TNC requests for reading the sensitive data from the SNC and sends it the respective block access token.*
*5. The SNC in turn*
- *Receives the request.*
- *Decrypts the block access token to verify authenticity of the request.*
- *Sends the requested data to the TNC along with the computed hash value of data encrypted by the access token.*
- *Starts a timer and waits for acknowledgment.*
- *If acknowledgment is received in time, the packet is deleted.*
- *If acknowledgment is not received in time, the packet is retransmitted until either a successful acknowledgment is received or a limit number of retransmissions are reached.*

*6. The TNC receives the data and*
- *Verifies its hash value.*
- *If the hash value is correctly verified, the TNC sends acknowledgment back to the SNC encrypted by the access token.*
- *If the acknowledgment is not received by SNC, the TNC may receive more than one copy of the same packet because of retransmissions. In this case the duplicated copies are deleted.*

978-1-5090-4580-8/16/$31.00 ©2016 IEEE

establishing user accounts with both the source nodes cluster and the destination nodes cluster. We assume that SNC and TNC are trusted by the user, but SNC and TNC may not trust each other. In the following algorithm, we present the detailed securing big data steps:

## VI. Conclusion and Future Work

In this paper, we have discussed the big data aspects with focus on security. This paper proposes a methodology to protect big data information during analysis, by classifying data before any action to be performed such as moving, copying or processing. Based on big data classification, the security procedures will be activated according to data criticality.

Future work will focus on testing more benchmarks for big data with focus on security of real time big data information.

## References


[1] E. Bertino, "Big Data - Security and Privacy," 2015 IEEE International Congress on Big Data, New York, NY, 2015, pp. 757-761.

[2] T. Mahmood and U. Afzal, "Security Analytics: Big Data Analytics for cybersecurity: A review of trends, techniques and tools," *Information Assurance (NCIA), 2013 2nd National Conference on*, Rawalpindi, 2013, pp. 129-134

[3] S. H. Kim, N. U. Kim and T. M. Chung, "Attribute Relationship Evaluation Methodology for Big Data Security," *IT Convergence and Security (ICITCS), 2013 International Conference on*, Macao, 2013, pp. 1-4.

[4] A. Gupta, A. Verma, P. Kalra and L. Kumar, "Big Data: A security compliance model," *IT in Business, Industry and Government (CSIBIG), 2014 Conference on*, Indore, 2014, pp. 1-5.

[5] A. Kumar, T. Hemlata, S. Yadav, "A Review on Big Data and Its Security", *2015 International Conferenceon Innovations in Information, Embedded and Communication Systems (ICIIECS 2015)*, India, 2015, pp. 178-183

[6] R. Alguliyev and Y. Imamverdiyev, "Big Data: Big Promises for Information Security," *Application of Information and Communication Technologies (AICT), 2014 IEEE 8th International Conference on*, Astana, 2014, pp. 1-4.

[7] R.Akerkar, "Big Data Computing ", International Standard Book Number 13: 978-1-4665-7838-8.

[8] F. Tekiner, J. Kaene, "Big Data Framework", *2013 IEEE International Conference on Systems, Man, and Cybernetics*.

[9] Aziz Nasridinov, Young-Ho Park, "Visual Analytics for Big Data Using R", *CGC '13 Proceedings of the 2013 International Conference on Cloud and Green Computing*, pp. 564-565, 2013.

[10] http://mirrors.sonic.net/apache/hadoop/common/hadoop-2.6.0 (last time Accessed 9/9/2016)

[11] https://www.ubuntu.com/ (last time Accessed 9/9/2016)

[12] Alex Holmes. Hadoop in Practice. Manning Publications Co. Greenwich, CT, USA. 2012.

[13] [13] Shvachko, Konstantin, et al. "The hadoop distributed file system." *2010 IEEE 26th symposium on mass storage systems and technologies (MSST)IEEE*, 2010.

[14] [14] Zikopoulos, Paul, and Chris Eaton. "Understanding big data: Analytics for enterprise class hadoop and streaming data". McGraw-Hill Osborne Media, 2011.


978-1-5090-4580-8/16/$31.00 ©2016 IEEE