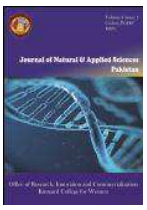
Contents lists available http://www.kinnaird.edu.pk/

Journal of Natural & Applied Sciences Pakistan

Journal homepage: http://jnasp.kinnaird.edu.pk/



SECURITY ISSUES OF BIG DATA ANALYTICS

Neha Khalid[1], Mishaal Sikander[1], Soha Bint Zarar Khan[1], Dr. Muhammad Rizwan[1*], Dr. Fahad Ahmad[1]
[1]Department of Computer Sciences,
Kinnaird College for Women, Lahore, Pakistan.

Article Info

*Corresponding Author
Tel: +92 333-4881501
Email Id:
muhammad.rizwan@kinnaird.edu.pk

Abstract

The world in which we are living today is full of technology and digitalization, due to this our daily activities have resulted in large amount of data, this huge amount of unprocessed information is mainly from social networking sites and collectively called Big Data. Big Data is used by many organizations through which marketing decisions are made and behavior of data is detected. This "Big Data" is controlled and maintained by Big Data Analytics. Organizations face challenges to maintain huge amount of data because of current inefficient methodologies. And this fact is emerging as a challenge and also risk to security. Considering these issues, strong techniques need to be proposed immediately. This paper highlights major security issues of big data analytics. Later a solution to overcome these big data analytics security issues is proposed.

Keywords

Big data, big data analytics, security issue, bloom filters, analytics variants, COUBF.

1. Introduction

Huge amount of content generated merely because of our digitalization i.e the use of social networking sites, computers, and phones has given rise to big data. This big data is increasing day by day because of our excessive usage. Big data can be good for some organizations as it helps them to get useful information out of it within no time. Methods that are currently used to maintain, process and analyze this big data are not that efficient. There is an urgent need to introduce alternative and more efficient techniques and tools that can handle today's big data. These techniques, methods, tools that are used to process and control big data are called Big Data Analytics [1].

Whenever we are dealing with information, the most important thing is the security of collected information. So, to secure collected records are of utmost importance.

Researchers have focused in order to propose new techniques that not only provide organizations with big data benefits but also enable them to use big data without risking security [2].

Following is the papers sectioning, in the first introduction section overview of big data and big data analytics is given. In the third section, major security issues are raised and discussed. Furthermore, a solution is proposed to overcome security challenges in big data analytics. In the fourth section, previous work done in overcoming such issues has been discussed. Lastly a proposed solution is given based on the outcome of a comparison of all previous techniques.

1.1 Big Data & Analytics Overview

1.1.1 Big Data

With the passage of time, people interaction with computing technology generates large amount of data, which is termed as Big Data. Big Data, as discussed, is large volume of information, This needs new processes and algorithms for better results. It involves gathering, storing and taking valuable information out of it as shown in figure 1.0 [3].

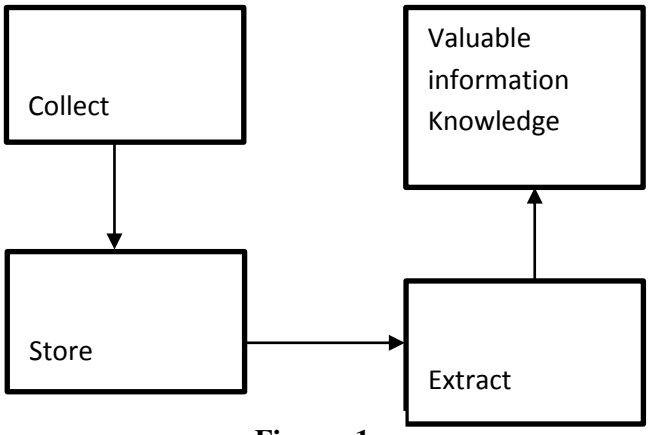


Figure: 1

1.1.2 Big Data Analytics

The process of inspecting large amount of big data to get useful information out of it than can benefit organization is called big data analytics. It involves controlling, maintaining, processing of Big Data. These are specific platforms comprising of systems placed over many different parallel nodes and clusters, allowing the performing of complex and huge computations on reduced infrastructure. In Table 1, we describe some of the famous analytics techniques. Out of all, Hadoop is the first analytic tool used for big data processing [3].

1.2 Security Issues

This paper puts major focus on the security challenges of big data analytics. These issues need to be addressed as soon as possible to secure the value of confidential information. Potential challenges are Network Forensics, Fraud Detection, Data Privacy and Data Origin.

1.2.1 Fraud Detection

The most recent vibrant challenge in society is the detection of fraud based on the use of big data analytics. Telephone, memberships and credit card companies, banks are mostly needy of fraud detecting system [2].

1.2.2 Network Forensics

To present handling of large amount of network traffic and large amount of data in inter-network an investigating process i.e. network forensics has been introduced. There are many problems in accessing the networking devices. These challenges can be solved by big data analytics [2].

1.2.3 Data Privacy

This issue depends on utilization of data by the respective organizations. Data should be used only for the purpose it was created and it should not be shared among others to gain any profit or business out of it. Many methods and techniques facilitate access to confidential data and make it easier to commit an illegal act. For all entities involved in this process security and privacy policies must be introduced [2].

1.2.4 Data Origin

This is a major issue because data can originate both from trustworthy and untrustworthy sources for example, old news about company posted through big data analytics which can cause down fall of company's share. In our work, only authentic and reliable data is considered when analyzing data using big data tools. Moreover, data integrity and authenticity are crucial components in our mind while analyzing data [2].

All these constraints can be resolved with big data analytics only but with the little change in existing techniques. In the later section, we have proposed a technique than can overcome all these big data analytics security issues.

Table 1

Analytics technique	Description
Hadoop	Open source software, deals with large amount of data
MAP- REDUCE	Handle data with a technique called "Map & Reduce" technique. It too deals with large volume of data.
Hadoop Distributed File System	It is the main part of Hadoop.
Hive	Handles data requisition and management over distributed systems.
HCatalog	Deals with data storage
NOSQL	Deals with retrieving any form of data

2. Literature Review

Previous work done to overcome network security issues of big data analytics: There are numerous techniques that are presented, which have an ability to secure issues related to network security in big data and cloud environment. These techniques have a framework based on minimizing the node authority by de- privileging them in a way that does not allow inspecting and tracing user activity by the cloud provider. And then CL-PR scheme certificate-less proxy re-encryption was introduced to enable the user to share data and also control the access of data. Authors then came up with idea of blind processing over outsourced data. It depends upon homophobic encryption and also leads the data blindly. without disclosing the data and its type when it is accessed. Another scheme is used to protect the privacy of data and improve query processing in outsourced database which is known as cryptographic-based transformation scheme. Pay-by-data model was introduced through which access to huge amount of data generated by users is possible and the data is protected by using authentication service [5].

Another possible solution for the mentioned problem is BAYESIAN CLASSIFICATION MINING. This technique consists of the extraction of the data from the system and it is getting through SCADA (Supervisory Control and Data Acquisition) systems. This technique consists of two combined techniques i.e. data mining and network forensics. It is one of the area where data mining and network forensics are combined. In this technique first the class level of network is predicted under the category of network attack and it predicts the probability of their occurrences. And then it is confirmed that all the attributes are independent from others, this step help to lead towards efficient and effective result [6].

Research has shown that ultra-high dimensional data models can help to overcome issues in network domain. Such models will predict and detect intrusions and attacks. With the help of Hadoop, a network information can be analyzed and regulated efficiently. By comparing datasets in databases, malicious data in the network can be easily controlled. A framework, MapReduce, has been proposed which is typically a framework of Hadoop. It orchestrates the processing by marshaling the servers over a distributed network. It manages all communications, transformation between various system parts with great efficiency [7].

Increasing network complexity and scalability has contributed to emerging network security challenges. Under this the common issue is Misuse detection and anomaly detection. It uses pattern recognition technique to recognize anomalies. It is not easy and to some extent impossible to match all patterns with immense increase in network traffic. To overcome this issue, researchers have worked on various techniques and have categorized methods into three classes namely supervised methods, unsupervised methods and hybrid methods. They are based on clustering, classification and outlier algorithms. However these techniques face problems with regard of performance and cost. Moreover, it is better to adopt a mixed hybrid approach rather than a single method. The mixed framework can yield the most advantages of methods and alleviate the shortages, so that they always can obtain a better performance than a single algorithm [8].

3. Problem Statement

This paper focuses mainly on the security issues of big data analytics more specifically on the most prevalent issue of BDSA i-e network forensics. This issue needs to be addressed as soon as possible to secure the value of confidential information. Unfortunately, previous security solutions did not yield the best results for big data network streams. Indeed, a great work has been done to overcome these issues but efficiency lacking in all these techniques made these issues emerging on an even bigger scale.

4. Proposed Solution

Meeting the rate of data growth to handle big data seems to be a challenge. The traditional solutions and techniques available for big data security cannot be applied while dealing with real time big data network streams. The reasons why these old techniques were not applicable was that the techniques are not able to efficiently hold space and time events. To increase the efficiency and security of big data analytics over the networking domain we have suggested adopting the technique of bloom filter. Bloom filter acts as a space and time efficient probabilistic structured data in BDSA. To overcome any issues in data analytics BF (bloom filter) and its variants play a vital role, which will be discussed in later section. Security analytics helps in monitoring the network and detect malicious patterns. In a later section we are going to use comparative and mathematical approach to show how it is better than other data structures or frameworks, and why it is a time and space efficient technique. Among all BF versions, CouBF yields the best result. It has reached recall rate which is higher than 98% with

efficient time and memory consumption. We have the least false positive rate and fastest average running time.

Table 2

Properties	Bloom Filter	Other data structures
Storage	Do not store data items at all	Store data items themselves
Insertion	Insertion time is a fixed constant	No other data structures possesses this property
FN	No false Negatives	False Negative chances are high
Deletion	No deletion allowed	Deletion Allowed
FP	Low false positive probability	High probability
Time	Time efficient, its size is independent of items already in list or set.	Less time efficient
Collision	No collision	Collisions are immense and handling is difficult
Hash functions	Multiple hash functions	Limited Hash functions
Control on size and FP rate	Can control size and FP rates	No efficient method to control it
Hash Area	Small hash area but still works perfectly	Large hash area

5. Result Discussion

5.1 Comaparative Analysis

In the previous section, solution to network security issues in big data analytics was proposed. In this section, a comparative analysis has been done to show how the proposed solution offers better way to overcome those issues. Later, two of the big issues under network forensics had been solved through above mentioned technique via mathematical approach.

As discussed earlier, underlying threats under networking involves different anomalies such as intrusion, unusual behavior, spam emails and other cyber-attack types. Bloom Filter, solution proposed, and its variants are widely spread in network security proved how CouBF has field. It is type of a data structure that makes efficient use of space in minimal amount of time. After improvements the bloomier filter (BLF) was made, it can make unrestrained functions. For economical usage of storage these functions are generalized. Functions remain unaffected because of automatic updates. To represent that capacity of data that can be transmitted over a network, compressed BF (ComBF) was introduced. To figure out spam email messages, counting BF (CouBF) was introduced. To tell the size of static set which will be represented and unchangeable over time dynamic BF(DBF) creates a dynamic bit matrix of a×b. Generalized BF (GBF), puts upper bound limit on the chances of false positives to eradicate security issues.

Table 2 shows comparison of Bloom Filter with other data structures such as self-balancing binary search trees, tries, hash tables, or simple arrays or linked lists. From above table we can conclude that, many advantages accrue from this approach, over a network, route of a packet can be determined by enabling IP trace backing that highlights source of the attacks. This technique makes the system more scalable than before. Another advantage is the capability to trace any attack long after it is over. The major distinguishing point is its size; number of bits is constant and set upon initialization.

5.2 Mathematical Solution

5.2.1 Hashing Approach

In networking, one of the issues is taking count of spam email messages under CouBF implementation is proposed. Coincidental bit occurs when a single call is used by two or more email messages which can increase to too many counters. [2] Suppose the number of email messages K= 4, then following setup is valid.

$A\,[h_1(e)]=1$
$A[h_2(e)]=0$
$A[h_3(e)]=0$
$A[h_4(e)]=1$

1	0	0	1

So when the counter will increase they will become as follows:

$A[h_1(e)]=2$
$A[h_2(e)]=1$
$A[h_3(e)]=1$
$A[h_4(e)]=2$

2	1	1	2

For making a hashing approach H_1, only those mail will be increased which have a minimum counter. So, the above will become as follows.

$$A[h1(e)]=1$$
$$A[h2(e)]=1$$
$$A[h3(e)]=1$$
$$A[h_4(e)]=1$$

1	1	1	1

The second heuristic approach H2 is that if any of two or more h1 (e)hk (e) hits a same counter then it will increase once. So, H_1 represents hits due to multiple messages and H_2 represents hits due to one single message.

Table 3: Tabular Representation of how Bf Variants Contribute To Overcome Network Security Problems

Bloom Filter variants	False Positives	False Negatives	Security Usage	Network Security domains
Standard (SBF)	✓	--	✓	Authentication Firewalling Anomaly Detection Misbehavior Detection Node Replication Privacy Preserving Email Protection
Adaptive ABF)	✓	--	--	---
Bloomier Filter(BF)	✓	--	✓	String matching
Compressed (ComBF)	✓	--	✓	Authentication IP trace backing
Counting(CouBF)	✓	--	--	Firewalling Email Protection String Matching
Dynamic(DBF)	✓	--	--	Node Replication detection
Space Code(SBF)	✓	--	--	IP trace backing
Generalized(GBF)	✓	✓	✓	IP trace backing
Hierarchical (HBF)	✓	--	--	○ IP tracebacking

The question is now arrises how this Cou BF is made by Universal Hashing. A class of universal hashing function is used for its construction[2].

Hc,d(x)=((cx+d)modp)modm

Where p is for prime number, c,m and d are integers and x is the hashed key on which hashing is supposed to be performed.

5.3 False Positive Solution

When using CouBF technique under Networks, another problem like false positive can occur probably when email message happens rarely and its least count becomes greater than the actual number of entries, resulting in a false positive rate that manifests itself as a counting error rate. This FP counter can be calculated and solved, by setting the counter to minimum and it will increase when email is encountered. For any message a, if the least count of Ah1(a)]……[hk(a)]≠ 20, then it has an incorrect count and the false positive will be incremented by 1 and then false positive/number of e-mail messages will be calculated in each round [2].

Following are the results of different techniques:

5.3.1 Recall Rate Testing

The ratio of found and inserted spam email messages is called recall rate. For example if 50 spam email messages are used and then inserted, so the pseudo spam will become a specific number like 10,20,30,40,50 in this experiment, and after the experiment is completed, a 98% of spam emails were found and 99% of messages were inserted multiple times which created error [2].

5.3.2 False Positive Testing

When each spam e-mail/error message are inserted multiple/100%, 99% of inserted ones error message were found which make false positive rate becomes low [2].

5.3.3 Memory Consumption Testing

For consumption of the memory the following formula is used:

m*size of cell

For 2,000,000 e-mail messages when m = 2,000,000 with 6 hash functions, then it will consume 12 megabytes.[2]

5.3.4 Time Consumption Testing

The time which will be taken by CPU to check the emails CouBF was 921 seconds which is equal to 15.3 minutes [2].

6. Conclusion

Nowadays the task to handle big data properly and efficiently has become quite a challenge because of the high rate of growth of data. Due to this challenge the threat frequency of security has much increased, so an adequate effective technique is proposed which are advantageous in terms of cost, storage, time and speed. Bloom filter (BF) is a data structure that is space efficient which supports the membership queries. [1] BF constantly balances the downside when false positive probabilities take place and this is done by space saving and locating time feature of BF. This same approach is also used for many security applications. Specifically, Bloom filters (BFs) are a general aid for network processing and improving the performance and scalability of distributed systems. Major security issues of big data analytics has been raised in the paper and a solution to overcome these big data analytics security issues has been proposed. A contribution of BF and its variants to network security field has been explained. One of our future works is to determine which system/tool would be the best replacement for standard bloom filter and CouBF in the network domain having the least false positive rate and fastest average running time.

References

A. Gani, A. S. (2015). A survey on indexing techniques for big data: taxonomy and performance evaluation. *vol.46*, pp. 241-284. Knowledge and Information Systems.

Alsuhibany, S. A. (n.d.). A space-and-time efficient technique for big data security analytics., (pp. 1--6).

Alvaro A. Cárdenas, P. K. (2016). Big Data Analytics for Security. *IEEE Security & Privacy 2013 IEEE*, (pp. 74-76).

Cuzzocrea, I.-Y. S. (2011J). Analytics over large-scale multidimensional data: the big data revolution! *In Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP*, (pp. 101-104).

Dopazo, J. (2014). Genomics and transcriptomics in drug discovery,. *vol.19*, pp. 126-132. Drug Discovery Today.

H. Dubey, J. Y. (2015). Fog Data: Enhancing Telehealth Big Data Through Fog Computing. *ACM ASE BigData & SocialInformatics*, (pp. 1-6).

J.P. Nivash, E. D. (2014). Analysis on enhancing storm to efficiently process big data in real time. *Proceedings of the 2014 International Conference on Computing, Communication* (pp. pp.1-5). Networking Technologies.

M. Gärtner, A. R. (2013). Bridging structured and unstructured data via hybrid semantic search and interactive ontologyenhanced query formulation. *vol. 41*, pp. 761-792. Knowledge and Information Systems.

M. Weidner, J. D. (2013). Fast OLAP Query Execution in Main Memory on Large Data in a Cluster. *IEEE International Conference on Big Data*, (pp. 518-524).

R. Buyya, K. R. (2015). Big Data Analytics-Enhanced Cloud Computing: Challenges, Architectural Elements, and Future Directions. *arXiv:1510.06486.*

S. Manegold, P. B. (2002). Optimizing mainmemory join on modern hardware. *TKDE*, *vol. 14*, pp. 709–730.

X. Wu, X. Z.-Q. (2014). Data mining with big data," In IEEE Transactions on Knowledge and Data Engineering., *Vol.26*, pp. 97-107.

X. Zhang, T. K. (2004). Strategies for using additional resources in parallel hash-based join algorithms. *HPDC*, (pp. 4–13).