[IDSTM: January 2017]
ISSN 2348 – 8034
Impact Factor- 4.022

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

BIG DATA: SECURITY ISSUES AND SOLUTIONS

Minakshi Mehra
Department Of CSE, Ganga Technical Campus, Bahadurgarh, India

ABSTRACT

In recent years, Big Data has rapidly developed in the fields of computer science and information management. It is used to solve complex problems in different areas of academia, engineering, industry, governments, social networks business and telecommunications. Apart from the extensive attention and growth of big data there are several security and privacy issues related to it. In this paper, we first describe the concept of big data, then we identify the possible security threats and then we describe the possible solutions to address these security threats.

Keywords-. *Big Data, Hadoop, MapReduce, Fully Homomorphic Encryption, and Attribute based encryption*

Introduction

"Big data" is large and complex set of data that is difficult to manage and process through traditional software tools. In Wikipedia, big data is defined as "The size of data set in Big Data is so large that it is difficult to capture, curate, manage, and process data by commonly used software tools within a tolerable elapsed time." [1].

"At IBM, big data is known as 'the art of the possible'. The company is certainly a leader in this space" [2].
Big data is a technology that is used to manage large structured or unstructured data set, as the size of data is constantly changing from terabytes to many petabytes, so there is a need of some technique to control and manage these data sets. In business organization the data is too large or moves too fast beyond its processing capacity. Big data have this capability to help these organizations to make their operations faster and can make intelligent decision. Big data have following characteristics which are usually known as 7 V's [1]:

- Volume – "Is the size of data is considered to be called as big data".
- Variety – "To which category the data belongs".
- Velocity – "How fast is data processed and generated".
- Variability – "Is data inconsistence".
- Veracity – "Is data meaningful and correct".
- Visualization – "Is data presentable (readable and accessible)".
- Value – "Is data worthless".

Big data uses cloud computing, as it requires a platform like Hadoop to analyze and store large set of data across distributed cluster and MapReduce to coordinate, combine and process data from various sources.

RELATED RESEARCH WORK

Various researches ongoing in this field are reviewed [3]. Madden open up challenges and opportunities in databases of big data [4]. Girola establishes virtualization planning and cloud computing methods in IBM data networking centre [5]. Keim portrays methodologies in big data virtualization [6]. Dittrich introduces contributions on optimizing big data processing efficiency in Hadoop and MapReduce [7]. Herodotou put forward a self-tuning system for big data analytics [8]. Yongqiang proposed RCFile as a fast and space efficient data placement structure in MapReduce warehouse [9]. Costa advocates an efficient in-network aggregation technique for big data applications [10], which considerably reduces the size of data transportation. Brunet comes up with their technique of Gaia Hadoop solution focusing on identifying potential challenges in Big Data [11]. Efforts on optimizing interactions with big data analytics were reported by Fisher, Danyel [12]. Begoli presented their design principles for efficient knowledge discovery [13]. Radoop: based on RapidMiner and Hadoop, has attracted attention in data analytics [14]. Agrawal describes current states and potential future opportunities for big data and cloud computing [15]. Lakew give an account of Resource management and allocation in multi-cluster clouds [16]. "Hitune: dataflow-based performance analysis for big data cloud." was presented by Jinquan [17]. Gripping anamnesis on big data processing in cloud computing environment was introduced by Changqing [18].

ISSN 2348 – 8034
Impact Factor- 4.022

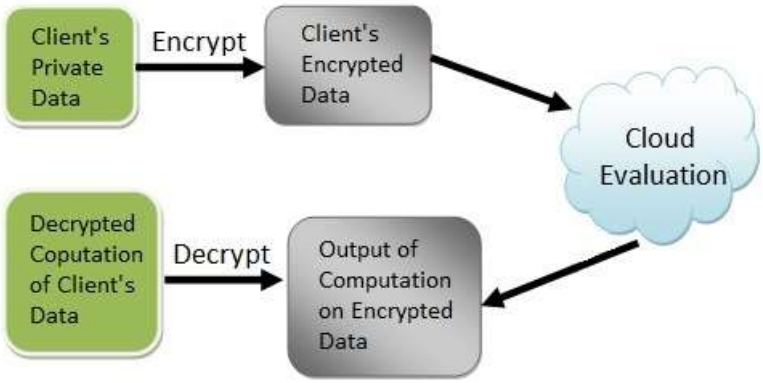
SECURITY ISSUES AND SOLUTIONS

Issues

Protection:

Protecting user's private information and credential information is a major concern as Big data contains large amount of these information, any breach can affect a much larger number of people's data. Whenever information is gathered for big data, the organization must ensure that they have right balance between utility of data and privacy. Before storing the data, it should be anonymized, by removing unique identifier for user. But this too will be security challenge as doing so will not guarantee that data will remain anonymous [19].

Figure:



Fully Homomorphic Encryption

Storage:

Storing data will lead to encryption problem. If a user sends data in encrypted form than it will create problem if cloud needs to perform some operation on that data. The solution to this is use of "Fully Homomorphic Encryption" [20] it allows the data stored in cloud to perform operations over encrypted data which in turn will generate new encrypted data. When the later encrypted data is decrypted the result will be same as they were carried out over plain text data. Therefore, the cloud will be able to perform operation over the encrypted data without knowing the plain text data [19].If the data is stored on cloud then the ownership of information should be maintained between the data owners and data storage owners. Access control mechanisms can be used for protecting the data.

Use:

Technology used to store big data like Hadoop doesn't provide any user authentication. This makes the problem of access control worse, as the software will leave the information open for unauthorized user. Firewalls can be used to restrict the access of information [19].

Solutions

Managing big data in an organization: Management of Big Data in an organization depends on the following aspects:

- Critical examination of cloud providers: Cloud provider must provide sufficient protections mechanism if data is stored on cloud.
- Access control policy: There must be access control policy that ensures authorized access only.
- Protection of data: Stored data must have encryption techniques to protect data from hacking.
- Protect Transaction: Ongoing transactions must be protected to ensure confidentiality and integrity
- Real time security: Threat intelligence systems must be used to prevent unauthorized access to the data.

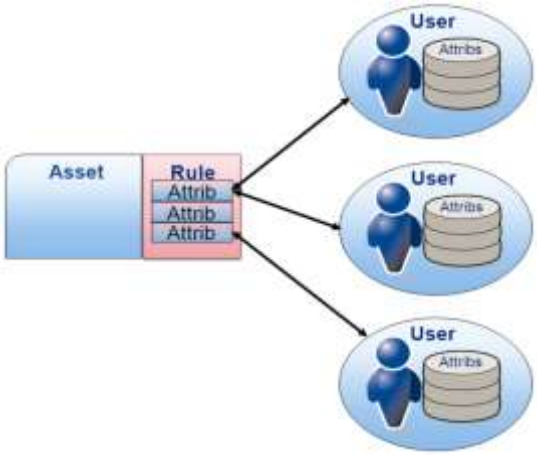
Technological Solutions:

- Encryption: Attribute based encryption [21] can help in providing fine- grained access control of encrypted data. In this type of encryption the secret key of the user and the ciphertext depends upon attributes. The

decryption of ciphertext is only possible of the set of attributes of the user key matches the attributes of ciphertext.

- Real Time Security monitoring: Organizations must monitor access to ensure that there is no unauthorized access. Threat intelligence systems must be up to date so that if there are any attacks than the system must react accordingly.
- Anonymization: The sensitive data must be removed from the records collected.

Figure:



Attribute based encryption

CONCLUSION

Security is the main aspect in big data. Therefore organization must take adequate steps to secure big data effectively and efficiently. Appropriate solutions must be used to secure the data. The organization must consider the 7 V's of big data and adjust their information lifecycle management accordingly. The main issue with big data is its unstructured nature of information which makes it difficult to model, map and categorize the data when it is stored. Therefore organizations must identify what information is of value for the business if they do this then risk wasting time and resources processing data will add little or no value to the business.

REFERENCES

[1] "Big data," in Wikipedia, Wikimedia Foundation, 2016. [Online]. Available: http://en.wikipedia.org/wiki/Big_data. Accessed: Jul. 28, 2016.
[2] "What is big data?," IBM, 2016. [Online]. Available: http://www-01.ibm.com/software/in/data/bigdata/. Accessed: Jul. 28, 2016.
[3] Bo Li, "Survey of recent research progress and issues in big data,". [Online]. Available: http://www.cse.wustl.edu/~jain/cse570-13/ftp/bigdata2/index.html. Accessed: Jul. 28, 2016.
[4] Madden, Sam. "From Databases to Big Data." IEEE Internet Computing 16.3 (2012).
[5] Girola, Michele, et al. IBM Data Center Networking: Planning for virtualization and cloud computing. IBM Redbooks, 2011.
[6] Keim, Daniel, Huamin Qu, and Kwan-Liu Ma. "Big-data visualization." IEEE Computer Graphics and Applications 33.4 (2013): 20-21.
[7] Dittrich, Jens, and Jorge-Arnulfo Quiané-Ruiz. "Efficient big data processing in Hadoop MapReduce." Proceedings of the VLDB Endowment 5.12 (2012): 2014-2015.
[8] Herodotou, Herodotos, et al. "Starfish: A Self-tuning System for Big Data Analytics." CIDR. Vol. 11.2011.
[9] He, Yongqiang, et al. "RCFile: A fast and space-efficient data placement structure in MapReduce-based warehouse systems." Data Engineering (ICDE), 2011 IEEE 27th International Conference on. IEEE, 2011.
[10] Costa, Paolo, et al. "Camdoop: exploiting in-network aggregation for big data applications." Presented as part of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12). 2012.

[11] Brunet, Pierre-Marie, Alain Montmorry, and Benoît Frezouls. "Big data challenges, an insight into the Gaia Hadoop solution." SpaceOps 2012. 2012. 1275512.
[12] Fisher, Danyel, et al. "Interactions with big data analytics." interactions 19.3 (2012): 50-59.
[13] Begoli, Edmon, and James Horey. "Design principles for effective knowledge discovery from big data." Software Architecture (WICSA) and European Conference on Software Architecture (ECSA), 2012 joint working IEEE/IFIP conference on. IEEE, 2012.
[14] Prekopcsák, Zoltán, et al. "Radoop: Analyzing big data with rapidminer and hadoop." Proceedings of the 2nd RapidMiner Community Meeting and Conference (RCOMM 2011). 2011.
[15] [15] Agrawal, Divyakant, et al.. "Big data and cloud computing: current state and future opportunities." Proceedings of the 14th International Conference on Extending Database Technology. ACM, 2011.
[16] Lakew, Ewnetu Bayuh. "Managing Resource Usage and Allocations in Multi-Cluster Clouds." (2013).
[17] Dai, Jinquan, et al. "HiTune: dataflow-based performance analysis for big data cloud." Proc. of the 2011 USENIX ATC (2011): 87-100.
[18] Ji, Changqing, et al. "Big data processing in cloud computing environments." Pervasive Systems, Algorithms and Networks (ISPAN), 2012 12th International Symposium on. IEEE, 2012.
[19] G. Lafuente, "Big data security - challenges & solutions," in MWR, MWR InfoSecurity, 2016. [Online]. Available: https://www.mwrinfosecurity.com/articles/big-data-security---challenges-solutions/. Accessed: Jul. 28, 2016.
[20] Gentry, Craig. A fully homomorphic encryption scheme. Diss. Stanford University, 2009.
[21] Goyal, Vipul, et al. "Attribute-based encryption for fine-grained access control of encrypted data." Proceedings of the 13th ACM conference on Computer and communications security. Acm, 2006