DOI 10.29042/2018-3290-3293

Helix Vol. 8(2): 3290-3293

A Survey on Big Data Security Algorithms

[*1] K. Lakshmi Prasanna, [2] M. Thrilok Reddy, [3] S. Shiva Prakash

[1, 2, 3] SVEC, Tirupati

Email: [*] *prasannaengg321@gmail.com, m.thrilokreddy@gmail.com, shivasthaneekam@gmail.com*

Received: 10th December 2017, Accepted: 8th February 2018, Published: 28th February 2018

Abstract

Technology today has advanced to a level wherein assortment of data can be done for each and every granular facet of a business, in real time. Privacy and security is the most important challenges in the big data. To secure existing big data surroundings anticipated to increase risks of breaches and leakages from private data and increased adoption of cloud technology because of the ability of buying, processing ability and storage space on-demand. Revealing traditional and new data warehouses and repositories to the exterior world with the risk to be affected to hackers and harmful outsiders and insiders. Within this paper the existing big data circumstances has been summarized along with issues encountered and security conditions that need attention. Also some existing methods have been explained to demonstrate current and standard guidelines for solving the problems.

Keywords: Privacy, Security, Big Data

1. Introduction

The most frequently employed solution as regards securing data privacy in a Big Data system is cryptography. Cryptography has been used to protect data for a considerable amount of time. This tendency continues in the case of Big Data, but it has a few inherent characteristics that make the direct application of traditional cryptography techniques impossible. Big data refers to data that is so large in quantity or complexness that current technology struggles to store and process it effectively. Such a need has resulted in the arrival of modern software like Apache Hadoop which uses Map Reduce to process and analyze large amounts of data by parallelizing the process and using distributed hardware [1].Big data consists of both old and new systems to help assess huge amounts of data in a acceptable timeframe and to produce insights which businesses can take action [2]. Technology today has advanced to a level wherein assortment of data can be done for each and every granular facet of a business, instantly. Electronic devices, electric power grids and modern software all generate huge amounts of data, which is in terms of petabytes (1,000 terabytes or 1,000,000 gigabytes), exabytes (1000 petabytes or 1 million terabytes) and zettabytes (1 million petabytes). Together with the development of IOT (Internet of things), every modern electronic gadgets are now-a-days is linked to the internet and is also collecting, creating and saving data which is huge. Total data produced by corporate companies, the internet and devices are anticipated to double to any extent further before next ten years[3]. Software which companies have used for such a long time for handling and examining data are not capable of handling this huge amount of data and therefore needed advanced parallel processing and distributed technology like Map Reduce to get the information analyzed or refined.

2. Big Data Characteristics

Big data is mostly a combination of huge level of data which is of heterogeneous type, data resources and format, and which moves in and out of companies at a rate of which is not controllable with traditional technology. Hence it is essential that such large level of data needs significant storage, different resources of data imply that data different in framework which have to be included to derive value, and real-time processing, storage space and evaluation would be needed for data because of the high speed.

Big Data is mostly seen as the 4 Vs - Volume, Variety, velocity and Veracity. Table 1 below summarizes how these 4 factors affect Big Data.

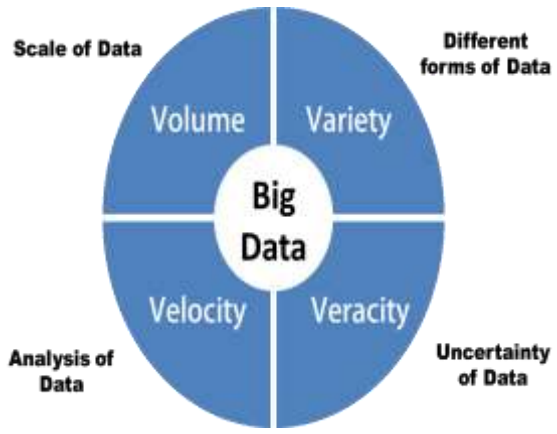

Fig 1: Big Data Characteristics

Big data has Issues and Difficulties that arise scheduled to increased needs for storage space, security and handling. Traditional and modern hardware are not capable of handling these difficulties alone. Listed here are some issues that big data systems must talk about:
Data-Storage: Data today is not simply created and produced by humans, but also by devices. Current development in disk technology capacities aren't capable of holding all the info produced together. Furthermore, being able to access this data all at the same time can clog communication sites.

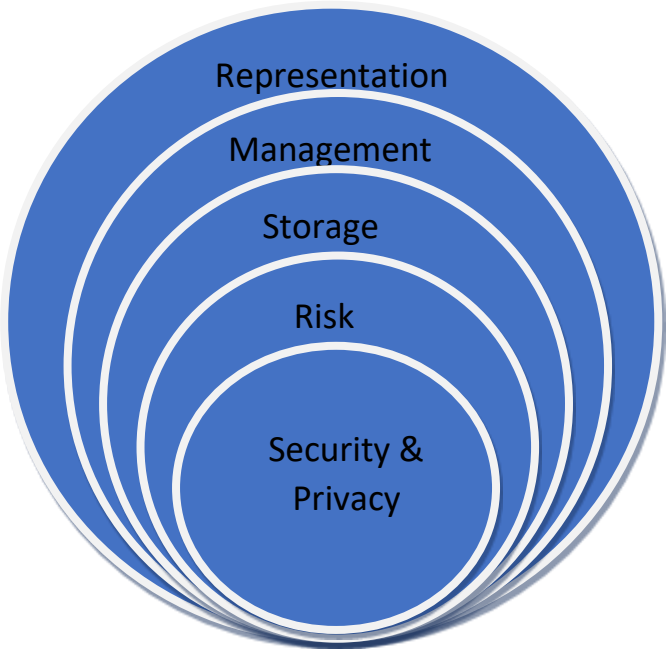

Fig 2: Big Data Challenges

Storage space must be saved by detaching redundant data. Discovering redundant data from unprocessed datasets and eliminating them is a requirement of reducing storage space and processing over head.
Data Management: Data can be allocated geographically, be possessed or managed and handled by multiple entities. Data may be in several formats and could be without proper information on source metadata and provenance. Issues of gain access to, metadata, recommendations and integration etc. have to be incorporated each and every level for satisfactorily controlling the data. That is difficult anticipated to huge amount and variety. High speed of data consumption helps it be difficult to validate and process in real-time

Data Representation: Datasets can be gathered from different resources and may differ in type and composition. Integrating these data units to give a uniform framework across datasets is a need.

Risk Recognition: Risks are associated with breaches and leakages from private data. Protecting private data is a fundamental requirement when managing health care and financial data. Breaches and leakages can result in lack of business if company secrets are exposed and lawsuits and financial reduction if user private data is exposed to the exterior world. Hence risk recognition must be a proactive process - this means breaches and leakages should be avoided and recognized before they have a potential to occur.

Data Security and Level of privacy: Data today is going through increased adoption of cloud technology because of the ability of buying processing ability and storage space on-demand. That is exposing business data to the exterior world and therefore novel big data security methods have to be taken up to secure data from dropping into the wrong hands. With this paper we give attention to security challenges encountered by Big Data and explore research guidelines.

3. Security and privacy challenges with big data
The below listed are the top level of privacy and security issues as reported by Big Data Working Group in their Apr 2013 paper[4]. While describing the challenges, improvement and research guidelines along these lines will be explored.

System Security
1) Secure Computations in Distributed Development Frameworks: Here parallel encoding methodologies are being used to process large amounts of data. The map reduce platform is one particular example. The info being processed must be guaranteed from untrusted mappers.

2) Protecting Non-Relational Data Stores: Non-relational data stores like MongoDB were designed with the intension to be used in analytics and security must be with-in middleware. Obtaining these non-relational data stores is important as they contain critical business data used for analytical handling. By hacking into such data stores an adversary may obtain critical business hypersensitive data or individually identifiable information. A report by Heyens and Petryka noticed that MongoDB databases commercially hosted were accessible openly via the Internet[5]. In MongoDB, encryption support is not built-in. 3rd party collection for encryption (field level) e.g. mongoid-encrypted-fields can be used[8].

Machine level security using Active Directory/LDAP should be integrated.

Methods to System Security

Xiao et al. Presented the idea of Accountable Mapreduce wherein each node is organized in charge of their behavior. Here Auditor nodes perform "accountability checks" (A-tests) to check on each node and identify nodes with distrustful behavior. This platform does apply to Map Reduce alternatives on the cloud. Liu et al. mentioned about an algorithm which is designed to reduce visibility of very sensitive data for organizations in storage space and also during communication. It is designed to discover a data drip recognition system using MapReduce [6]. The algorithm picks up plaintext or unencrypted very sensitive data in storage space and communication. While doing this it generally does not really know what the information is - thus preserving the level of privacy of very sensitive data.

Data privacy

1) Privacy - Preserving Data Mining and Analytics: Analytical control done on data shouldn't expose personal data for users or granular details that could lead experts to trace back again the data to the initial details. Data mining and predictive analytics techniques need to apply privacy conserving techniques e.g. masking very sensitive data and other anonymization techniques.

2) Cryptographically Enforced Data Centric Security: Data must be encrypted and usage of data must be handled by restricting permissions and enforcing gain access to restrictions on the data and end user level.

3) Granular Access - Control: Gain access to control can be coarse grained wherein gain access to is provided at a broader level and gain access to has lesser degrees of categorization. The restriction here's that to enforce security at a broader level, security must be higher and constraints to data tend to be more. Fine grained gain access to control provides access control at a much granular level, thus promoting legal and plan conformity requirements.

Methods to Ensuring Data Personal privacy

Agrawal et al recommended a strategy related to Privacy Preserving Data mining for building classifiers using training data. The data which classifiers are designed is not identical to the initial, while being not the same as the initial in circulation and beliefs. A new reconstruction method was suggested to recognize the initial data circulation after analyzing training dataset. Classifiers for data mining can be built by using this and the correctness of the classifiers were been shown to be much like classifiers built using the real data. Because the real data is not used to build classifiers in cases like this, privacy of very sensitive end user information can be maintained.

The basic procedure is to let users provide data with arbitrary noise put into it. 2 options for modifying values are considered - by discretizing values into mutually exclusive classes and by changing values by using a function of arbitrary values with standard or Gaussian distribution.

Lindell et al suggested a strategy on Privacy Preserving Data Mining considering 2 people trying to run data mining with a union of the respective databases[7]. The objective was not to reveal unneeded information. They used decision tree learning and the Iterative Dichotomiser 3 (ID3) algorithm to propose a best approach. No one involved here discovered more than the result itself.

Quasi-identifiers are attributes you can use to distinctively identify individuals by linking to exterior data. To counter this security concern the idea of k-anonimity[8] was presented. In this technique data is generalized or suppressed to lessen granularity. If a record k in a dataset is indistinguishable from at least k - 1 other record regarding every group of quasi-identifier attributes, the dataset can be called k-anonymous. It had been shown however that k-anonimity is not full evidence as it pertains to data personal privacy as possibly subjected to problems like "Homogeneity attack" when there exists Homogeneity of very sensitive characteristics and "Background knowledge attack" wherein background knowledge on individual is effective in identification.

IBE or Identity Based Encryption [9] is a kind of public key cryptography. Any information about the identification of an individual e.g. email can be utilized as the public key. IBE was expansion of the ID-based encryption [10] recommended by Shamir and in the 2003 paper published by Boneh et. al[9].

Data Management

1) Secure Data storage area and Transaction Logs: Data and logs are stored and maintained by automatic tiering systems which by default provide minimal security to reduce tiers and much more security to raised tiers. Data not used for period of time may be automatically drawn into lower tiers by the storage space system and therefore conclude having lower security. This data, even though old may have highly confidential information and having lower security may be unsafe.

2) Granular Audits: Granular audits must support real-time security monitoring. A security attack might not continually be detected using real-time security monitoring. Audits of logs at a granular level have to

be done to recognize potential security occurrences that could have occurred.

3) **Data Provenance**: Provenance is in charge of storing possession information. For digital documents provenance is very important to regulatory and conformity purposes. It's important to prevent destructive users from adding or eliminating provenance data. For instance, a malicious consumer may be enthusiastic about eliminating provenance data after tampering with a file to avoid traceability.

Methods for Secure Data Management

Li et.al. in their paper have suggested solutions for secure auditing and data deduplication for cloud centered environments[11]. They may have suggested 2 systems - SecCloud which helps audit integrity of data and generate data tags. This reduces your time and effort of auditing. SecCloud also permits secure deduplication. An increased version of SecCloud called SecCloud+ includes an integral server for increased security by encryption.

Liu et.al. recommended a way of integrity confirmation of cloud centered big data when the auditing is performed with a third party. The auditing plan is dependent on BLS signature.

Conclusion

While Big Data technology is enhancing daily this mean that the quantity of data combined with the rate of which data is streaming into businesses today is increasing. Private data must be shielded from adversaries and destructive software - both to keep up integrity of the info and personal privacy of very sensitive information. Business decisions produced from data is very important as this what drives future guidelines, hence maintaining accuracy and reliability of data is vital. The security issues must therefore be handled and new novel security methods need to appear that may be modified to Big Data. While software security moves long back computation history, not absolutely all techniques are suited to Big Data. The success of security techniques in safeguarding data and the power of sharing data without privacy concerns will determine the probable of Big Data version to cloud centered environments in future.

References

[1] J. Dean and S. Ghemawat, "MapReduce: Simplied Data Processing on Large lusters,", pp. 137–149, 2004.

[2] J. Hurwitz, A. Nugent, F. Halper, and M. Kaufman, Big Data For Dummies. John Wiley & Sons, Inc., 2013.

[3] J. Gantz and D. Reinsel, "THE DIGITAL UNIVERSE IN 2020 : Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East," IDC IVIEW, no. December 2012, pp. 1–16,2012.

[4] BigDataWorkingGroup, "Expanded Top Ten Big Data Security and Privacy hallenges,"2013.

[5] J. Heyens, K. Greshake, and E. Petryka, "MongoDB databases at risk - Several thousand MongoDBs without access control on the Internet," 2015

[6] F. Liu, X. Shu, D. Yao, and A. R. Butt, "Privacy-Preserving Scanning of Big Content for Sensitive Data Exposure with MapReduce,", 2015, pp. 195–206.

[7] Y. Lindell and B. Pinkas, "Privacy-preserving Data Mining," Crypto '00, vol. 29, pp. 36–54, 2000.

[8] P. Samarati and L. Sweeney, "Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppresion.,", pp. 384–393, 1998.

[9] D. Boneh and M. Franklin, "Identity-Based Encryption from the Weil Pairing," SIAM Journal on Computing, vol. 32, no. 3, pp. 586–615, 2003.

[10] A. Shamir, "Identity-Based Cryptosystems and Signature Schemes," Advances in Cryptology, vol. 196, pp. 47–53, 1984

[11] J. Li, J. Li, D. Xie, and Z. Cai, "Secure Auditing and Deduplicating Data in Cloud," IEEE Transactions on Computers, vol. 9340, no. c, pp. 1–1, 2015.

