A REVIEW ON SECURITY AND PRIVACY ISSUES OF BIGDATA

[1]Mr. Shreyas K.S, [2] Mrs. Swasthika Jain T.J, [3]Mrs. Hema K
[1]Student, [2]Assistant Professor, [3]Assistant Professor
[1]Dept of CSE,
[1]GITAM, Bengaluru, India

ABSTRACT

As the world is growing larger and faster, the need of communication around the globe is must needed and it is the most happening thing. The interaction among the people is happening through social media, search engines, blogs, websites etc. This is generating enormous amount of data named as "Big Data". The Big Data has been adopted and implemented in various fields like banking, finance, health care applications, social media, IT sectors and much more. The Big Data gained its importance but it is very much essential that it should overcome the challenges and it should make sure that it overcoming security and privacy issues. This paper describes the overall representation of security and privacy issues of Big Data.

INTRODUCTION

The term Big Data is now used almost everywhere in our daily life. The virtual connections are more happening than the physical connections among the people resulting in the generation of enormous amount of data. Big Data can be seen in the finance and business where enormous amount of stock exchange, banking, online and onsite purchasing data flows through computerized systems every day and are then captured and stored for inventory monitoring, customer behavior and market behavior. It can also be seen in the life sciences where big sets of data such as genome sequencing, clinical data and patient data are analyzed and used to advance breakthroughs in science in research. Other areas of research where Big Data is of central importance are astronomy, oceanography, and engineering among many others. The leap in computational and storage power enables the collection, storage and analysis of these Big Data sets and companies introducing innovative technological solutions to Big Data analytics are flourishing. Big Data are high-volume, high-velocity and high-variety information, demands new forms of processing to enhance decision making, insight discovery and process optimization[1]. In general big data means collecting storing and extracting the valuable information from enormous amount of data. This process is also called as knowledge discovering process .

Knowledge discovering process

1. **Data Cleaning**: Data cleaning is defined as removal of noisy and irrelevant data from collection.
- Cleaning in case of Missing values.
- Cleaning ***noisy*** data, where noise is a random or variance error.
- Cleaning with Data discrepancy detection and Data transformation tools.

2. **Data Integration**: Data integration is defined as heterogeneous data from multiple sources combined in a common source(DataWarehouse).
- Data integration using Data Migration tools.
- Data integration using Data Synchronization tools.
- Data integration using ETL(Extract-Load-Transformation) process.

3. **Data Selection**: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.
- Data selection using Neural network.
- Data selection using Decision Trees.
- Data selection using Naive bayes.
- Data selection using Clustering, Regression, etc.

4. **Data Transformation**: Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure.

Data Transformation is a two step process:

- **Data Mapping**: Assigning elements from source base to destination to capture transformations.

- **Code generation**: Creation of the actual transformation program.

5. **Data Mining**: Data mining is defined as clever techniques that are applied to extract patterns potentially useful.

- Transforms task relevant data into patterns.
- Decides purpose of model using classification or characterization.

6. **Pattern Evaluation**: Pattern Evaluation is defined as as identifying strictly increasing patterns representing knowledge based on given measures.

- Find interestingness score of each pattern.
- Uses summarization and Visualization to make data understandable by user.

7. **Knowledge representation**: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.

- Generate reports.
- Generate tables.[2]

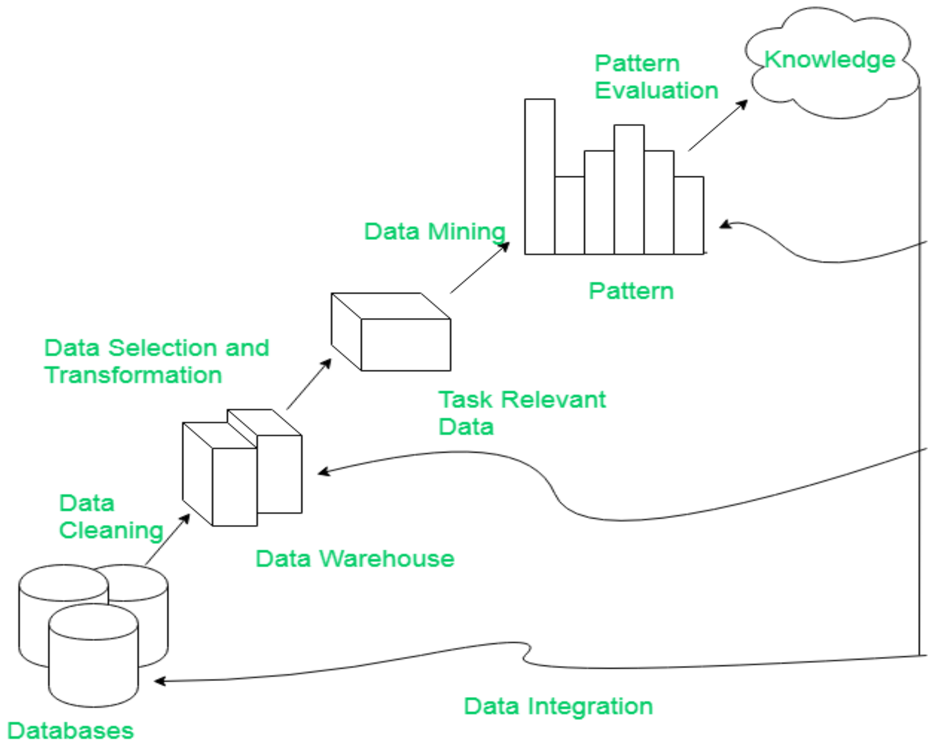

Figure 1:- Knowledge discovery process.

The V's defining big data.

1) **Volume:** There has been an exponential growth in the volume of data that is being dealt with. Data is not just in the form of text data, but also in the form of videos, music and large image files. Data is now stored in terms of Terabytes and even Petabytes in different enterprises. With the growth of the database, we need to re-evaluate the architecture and applications built to handle the data.

2) **Velocity:** Data is streaming in at unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. Reacting quickly enough to deal with data velocity is a challenge for most organizations.

3) **Variety:** Today, data comes in all types of formats. Structured, numeric data in traditional databases. Information created from line-of-business applications. Unstructured text documents, email, video, audio, stock ticker data and financial transactions. We need to find ways of governing, merging and managing these diverse forms of data.

4) **Variability:** Variability. In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Daily, seasonal and event-triggered peak data loads can be challenging to manage. Even more so with unstructured data involved.[3]

5) **Complexity:** Complexity. Today's data comes from multiple sources. And it is still an undertaking to link, match, cleanse and transform data across systems. However, it is necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control. A data environment can lie along the extremes on any one of the following parameters, or a combination of them, or even all of them together.

Security challenges in Big Data

Various big data analytics for security issues and privacy challenges are discussed here. In recent research threat environment big data can be differentiated from traditional technologies in four types: volume, variety, velocity and value. Security issues challenges are Amplified by velocity, variety and volume of big data such as very large scale cloud framework, distinction of data source and pattern, cascading nature of data acquisition .These are amplifying at a rapid rate and has required a shift in how protected venders manage threat, attacks. Following are few big data security challenges[12][13].

TABLE 1: Comparison of Various Security Challenges

Security challenges	Description
Protected database storage and transaction log file	With fast growth of database volume, availability and scalability have required auto tiring for the big data management. Auto tiring solution do not keep track of where the database is actually stored, which act as new demanding to protected database storage.
Secure computations in distributed frameworks	Parallelism is used in computations and physical storage to process very large data. MapReduce framework is an example. Protecting the mappers and protecting the data in the presence of untrusted mapper are two major attack prevention measures.
Privacy issues for non-relational data stores	Example is NoSQL databases. Usually NoSQL database embedded protection in the middleware. It does not provide any type of support for enforcing it explicitly in the database. However, gathering aspect of NoSQL databases impose additional demanding to the strength of such privacy practice.
End-point input validation/filtering	This approach is used to identify the trusted data .this method verify that source of data input details are not despiteful. If malicious input found then what is the method to filter malicious input from our collection?
Real-time security and compliance monitoring	This approach gives the number of alerts generated by privacy devices. These alerts lead to many false positives, which are mostly ignored or simply “clicked away,” as humans cannot cope with the shear amount. It is used to provide, for instance, real-time problem detection based on scalable privacy analysis
Granular audits	This security challenges are used for Compliance, regulation and forensics reasons. Example is Missed attacks. Granular audit is used to deal with the data object, which probably are allocated.
Cryptographically enforced access control and secure communication	This approach is used to ensure fairness, authentication, and agreement among the distributed entities.
Granulated access control	Secrecy-secures access to data by people that should not have access. Granulated access control give data manager a skiver in place of a blade to share data as much as possible without agree privacy.
Information security	Security of data has become a big data problem in itself. How to tackle big data from security point of view is a tough task. For example financial services, commercial website , Social networking sites, health sector, networking, anomaly detection. So information security is one of the big data issues.
Metadata provenance	Data will increase in complexity due to large provenance graphs generated from provenance enabled programming environments in big data applications. Analysis of such large provenance graphs to detect metadata dependencies for security and confidentiality purpose is computationally intensive

Stages involved in Big Data

1) **Data Acquisition:** The first step in Big Data is acquiring the data itself. With the growing medium the rate of data generation is rising exponentially. With the introduction of smart devices which are used with a wide array of sensors continuously generate data. The Large Haudron Collider in Switzerland produces petabytes of data. Most of this data is not useful and can be discarded, however due to its unstructured form; selectively discarding the data presents a challenge. This data becomes more potent in nature when it’s merged with other valuable data and superimposed.

Due to the interconnectedness of devices over the World Wide Web, data is increasingly being collated and stored in the cloud.

2) **Data Extraction**: All of the data generated and acquired is not of use. It contains a large amount of redundant or unimportant data. For instance, a simple CCTV camera, constantly polls sensor to gather information of the user's movements. However, when the user is in a state of inactivity, the data generated by the activity sensor is redundant and of no use. The challenges presented in data extraction are twofold: firstly, due to nature of data generated, deciding which data to keep and which to discard increasingly depends on the context in which the data was initially generated. For instance, footage of a security camera with the same frames may be discarded however it is important not to discard similar data in a case where it is being generated by a heart-rate sensor. Secondly, a lack of a common platform presents its own set of challenges. Due to wide variety of data that exists, bringing them under a common platform to standardize data extraction is a major challenge.

3) **Data Collation**: Data from a singular source often is not enough for analysis or prediction. More than one data sources are often combined to give a bigger picture to analyze. For example a health monitor application often collects data from the heart-rate sensor, pedometer, etc. to summarize the health information of the user. Likewise, weather prediction software take in data from many sources which reveal the daily humidity, temperature, precipitation, etc. In the scheme of Big Data convergence of data to form a bigger picture is often considered a very important part of processing.

4) **Data Structuring**: Once all the data is aggregated, it is very important to present and store data for further use in a structured format. The structuring is important so queries can be made on the data. Data structuring employs methods of organizing the data in a particular schema. Various new platforms, such as NoSQL, can query even on unstructured data and are being increasingly used for Big Data Analysis. A major issue with big data is providing real time results and therefore structuring of aggregated data needs to be done at a rapid pace.

5) **Data Visualization:** Once the data is structured, queries are made on the data and the data is presented in a visual format. Data Analysis involves targeting areas of interest and providing results based on the data that has been structured. For instance, data containing average temperatures are shown alongside water consumption rates to calculate a relation in between them. This analysis and presentation of data makes it ready for consumption for users. Raw data cannot be used to gain insights or for judging patterns, therefore "humanizing" the data becomes all the more important.

6) **Data Interpretation:** The ultimate step in Big Data processing includes interpretation and gaining valuable information from the data that is processed. The information gained can be of two types: Retrospective Analysis includes gaining insights about events and actions that have already taken place. For instance, data about the television viewership for a show in different areas can help us judge the popularity of the show in those areas. Prospective Analysis includes judging patterns and discerning trends for future from data that is already been generated. Weather Prediction using big data analysis is an example of prospective analysis. Problems accruing from such interpretations pertain to fallacious and misleading trends being predicted. This is particularly dangerous due to an increasing reliance on data for key decisions. For example, if a particular symptom is plotted against the likelihood of being diagnosed with a particular disease, it might lead to misinformation about the symptom being caused due to the particular disease itself. Insights gained from data interpretation are therefore very important and the primary reason for processing big data as well. All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

Security and Privacy issues of Big Data

The data is said to be secured only when it satisfies confidentiality, integrity and availability .privacy and authentication also plays a key role in Big Data. We should ensure that we are having an good security.

>Confidentiality

Confidentiality relates to applying some rules and restrictions to data against illegal disclosure. Limiting access to the data and various cryptographic techniques are widely used to sustain confidentiality. One of following way is used to ensure confidentiality in typical security mechanism [4]:

- Data is encrypted during transition and stored as plaintext.

- Authentication is used on stored plain text data to grant access.

- Data is encrypted when stored and decrypted when in use. Confidentiality is assured using Authentication, Authorization and Access control (AAA) [4].

Authentication is referred to as user identity establishment. Authorization is used to grant resource access to the authenticated user. Access control refers to enforcing resource access permission for authorized use to authenticated users.

>Integrity

Data integrity provides protection against altering of data by an unauthorized user in an unauthorized manner. Hardware errors, user errors, software errors or intruders are main reasons for data integrity issues [5]. Salami attacks, data diddling attacks, trust relationship attacks, man in the middle attack, and session hijacking attacks are most well-known attacks on data integrity [5]. Integrity can be maintained using data provenance, data trustworthiness, data loss and data reduplication. Data provenance is related to information about creation process as well as sources of data through which it is transformed. It is the process to check all states of data from initial state to current state. Debugging, security and trust models are various applications of data provenance [6-7]. Without data provenance information, user never comes to know from where the data came, what and which transformations have applied on data. This affects the value of data [8].

>Availability

Data availability ensures that data must be available for use when authorized users want to use. High Availabilitysystems are the solution for satisfying data availability [4]. Backup servers, replication and alternative communication links are widely used to design HA system. However, emergence of cloud computing has narrow downed issues of data availability for Big Data due to high uptime of cloud. Although Denial of Service (DoS) attack, Distributed Denial of Service (DDoS) attack and SYN flood attack are known attacks to breach data availability that need revised solutions [5].

>Monitoring and auditing

Auditing and monitoring provides active security to detect abnormal activity or intrusion from Big Data network security system. Monitoring of suspicious and non-suspicious data behavior is essential in Big Data. With enormous amount of data, conventional approach of blocking data on server and verifying data on client for data auditing is not feasible for Big Data. Kim et el. [9] has discussed abnormal behavior detection technique based on cyber targeted attack response system to detect abnormal behavior proactively. Dynamic nature of Big Data makes auditing complicated. However, healthy attempt to audit dynamic data storage using novel Merkle Hash Tree (MHT) based public auditing approach is elaborated in [10]. Proposed auditing method solves the communication overhead problem and authentication problem.

>Key management

Key Management and key sharing between users, servers and data centers are emerging security issues for Big Data. Wen et el. [11] has explained various approaches for key management such as secret sharing, server aided approach, and encryption with signature. In key management with Ramp secret sharing scheme (RSSS), users don't need to maintain any key on their own but instead of they have to share secrets among multiple servers.

>Data privacy

Recently we capture high amount of traceable data that were never taken and stored in the past. Data privacy issues crop up as a result of this traceable data that was anonymous. Data privacy intents to assure Personally Identifiable Information (PII) should not be shared without informed assent to related data owner. Even after sharing, use of PII is often restrained to specific reason. For example, to develop effective treatment or medicine, study of patient’s medical record is essential. Hence, PII of patient must be analyzed to protect privacy.

CONCLUSION

In this paper we have discussed discovering process, The importance of V’s, stages involved in Big Data and security and privacy issues of Big Data such as confidentiality, integrity, availability, monitoring and auditing, key management, data privacy.

The researchers are been continuously done to find the best way to overcome security and privacy challenges in Big Data. However there is a need in the future to find the best way to enhance the security and to take care of security and privacy.

As the Big Data is in its early stage, we hope that our exhaustive survey will help to develop better security and privacy solutions.

REFERENCES

[1] C. Philip Chen and C. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data", Information Sciences, vol. 275, 2014, pp. 314-347.

[2] https://www.geeksforgeeks.org/data-mining-kdd-process/

[3] http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf

[4] S. Sudarsan, R. Jetley and S. Ramaswamy, "Security and Privacy of Big Data", Studies in Big Data, 2015, pp. 121-136.

[5] "Types of Network Attacks against Confidentiality, Integrity and Availability", Omnisecu.com, 2017. [Online]. Available: http://www.omnisecu.com/ccna-security/types-of-network-attacks.php. [Accessed: 23- Jan- 2017].

[6] Z. Azmi, "Opportunities and Security Challenges of Big Data", Current and Emerging Trends in Cyber Operations, 2015, pp. 181-197.

[7]B. Glavic, "Big Data Provenance: Challenges and Implications for Benchmarking", Specifying Big Data Benchmarks, 2014, pp. 72-80.

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

[9] H. Kim, I. Kim and T. Chung, "Abnormal Behavior Detection Technique Based on Big Data", Lecture Notes in Electrical Engineering, 2014, pp. 553-563.

[10]C.LIU, R.Ranjan, C.Yang.DPA: Top-Down Levelled Multi-Replica Merkle Hash Tree Based Secure Public Auditing for Dynamic Big Data Storage on Cloud," in IEEE Transactions on Computers, vol. 64, no. 9, 2015, pp. 26092622.

[11] Y. Jeong and S. Shin, "An Efficient Authentication Scheme to Protect User Privacy in Seamless Big Data Services", Wireless Personal Communications, vol. 86, no. 1, 2015, pp. 7-19.

[12]Top Ten Big Data Security and Privacy Challenges: Cloud Security Alliance 2012

[13]Jainendra Singh, “Real Time Big Data Analytic: Security Concern and Challenges with Machine Learning Algorithm” IEEE 2014