I J R B A T, Issue (XI) Vol (II) May 2023 : 306-311
A Double-Blind Peer Reviewed & Refereed Journal
OPEN ACCESS
e-ISSN 2347 – 517X
Original Article

INTERNATIONAL JOURNAL OF RESEARCHES IN BIOSCIENCES, AGRICULTURE AND TECHNOLOGY
 www.ijrbat.in

REVIEW OF BIG DATA CHALLENGES AND RELATED ISSUES

Shital P. Adkine[1] and Dr. Keshao D. Kalaskar[2]

[1]Department Of Computer Science,
Sardar Patel Mahavidyalaya, Chandrapur, Gondwana University Gadchiroli, Maharashtra
[2]Department Of Computer Science, Dr. Ambedkar College, Chandrapur

Communicated : 23.01.2023 | Revision : 28.02.2023 & 08.03.2023 Accepted : 03.04.2023 | Published : 30.05.2023

ABSTRACT:
In today's world the rapid rise of Internet-based technologies likes social media platforms and mobile devices enormous amount of data available. Many companies and market leader unprepared for handling such a large, random and high velocity data with volume. . The way in which collection of large and complex datasets become difficult to handling using traditional database management tools and data processing applications. This paper focus on challenges in different aspects of big data, such as data sources, content format, data processing, data staging etc. Issues and challenges related to big data, specifically privacy attacks, security and counter-techniques are discussed. Tools and techniques adopted by various organizations to store different types of big data are also highlighted. This study recognized different research areas such as a lack of unidentified techniques for unstructured big data, data traffic pattern determination and mechanisms for high velocity data.

Keywords: - Component; Privacy; Unstructured Big Data, Big Data Classification, Big Data Tools.

INTRODUCTION :

Big data the term itself indicating its meaning a massive pool of data. Now a days valuable asset means data. The usage of big data spread due to commercialization and digitization in each area.

Mainly big data lies on 4 pillars of v's, volume variety, velocity, and veracity

Volume refers to the enormous amount of data that are being generated every day in sets, tables and files whereas velocity is the rate of growth and how fast the data are gathered for being analysis. Variety deals with information about the types of data such as structured, unstructured, semi-structured etc. The fourth V deals with veracity that includes availability and accountability. for example, banking system, transportation. While mobile applications that collect huge amount data. Other devices, such as healthcare equipment maintained data every minute.

All of these examples are produce big data.

This survey focused on issues and challenges in different areas of big data, and keeping eye on privacy problem in unstructured big data. This survey also focuses on different privacy attacks.

BIG DATA CLASSIFICATION

The pillars of big data, such as volume, velocity, veracity and variety, are dominated by valuable information that created promote the large-scale marketing efforts of software and hardware companies trying to advertise their product 'big data solutions'. Many companies focus on big data solution for structured data ignoring on unstructured data like such as text messages, videos and audio files captured from mobile devices which is much harder to analyzed. Company faces many difficulties to finding solution on this.

A recent study shows that the largest portion of big data consists of unstructured data, while structured data is only a small subset of such data [1]. Big data can be classified based on its categories, such as data storage, contents

http://doi.org/10.29369/ijrbat.2023.02.1.0045

formats and data staging [2]. Every categories have their own characteristics. big data is classified as follows:

Data Sources

Data sources are mainly deals with production of data. Some of the real time example of big data is internet-based applications , transactions made from the credit cards, face book, Twitter and what's app are generating the social networking data Instagram, Flicker, YouTube. These websites allow people to communicate people virtually and shared different types of data and information, while sharing this personal and interpersonal data a greater chance of misused of data, thus various security issues is a most important factor. The today's era technology not only support large amount of data, but also help in utilizing such data effectively.

One of main source of big data is the Internet of Things (IoT), which is based on numbers of sensors that collectively operate to generate enormous amounts of data. Data is generated from sensing devices including mobile devices, satellites and other sensors related to healthcare and weather stations [3]. The connectivity of large number of heterogeneous devices produce huge data [4], which includes features such as heterogeneity, variety, unstructured feature, noise, and high redundancy.

Content Format

Content formats of big data can also be used for classification. Different types of big data based on content format are as follows:

Structured Data

The data stored in relational databases table in the format of row and column. Structured data include numbers, text, and dates; in terms of database it is called strings. Data have fixed structures and these structures used for organizations to creating a perfect model. Data model permission to store, process and operate on data. Analysis and storing of structured data is very easy. Because of high cost, limited storage space and techniques used for processing, causes RDBMS the only path to store and process the data effectively. Programming language called Structured Query Language (SQL) is used for managing this type of data.

Unstructured Data

Without any specific structure and due to this could not be stored in a row and column format is unstructured data. The data is contradictory to that of structured data. It cannot be stored in a databank. Volume of this data is growing extremely fast which is very tough to manage and analyze it completely. To analyze the unstructured data advanced technology knowledge is needed.

Semi-structured Data

Data which is in the form of structured data but it does not fit the data model is semi-structured data. It cannot be stored in the form of data table, but it can be stored in some particular types of files which hold some specific marker or tags. These markers are distinguished by some specific rule and the data is enforced to be stored with a ranking. This form of data increased rapidly after the introduction of the World Wide Web where various form of data need medium for interchanging the information like XML and JSON.

Data Staging

Data which is not in a valid format, cannot be directly used for analysis. For example, consider data collected from social media is unstructured consisting of audio, video, and images. All of the data is in a format that requires processing to clean and convert into a structured format so that it can be easily analyzed [5].Remaining of data consider as waste ,process of cleaning is perform to identified the data. This staging process removes anomalies in data and is called normalization [6].

http://doi.org/10.29369/ijrbat.2023.02.1.0045

Data Processing

Data can be classified based on the type of processing that generates the data .The data processing given as follows: Real time processing and batch processing Map reduce is an example of batch processing which run by many organization. MapReduce [7], introduced by Google, is the programming model that provides abstraction from underlying hardware and facilitates parallel programming and execution on multiple clusters. Real time processing like Storm, S4, sparks streaming that deals with frequently changing dynamic data Synchronization and results composition are issues that require further research effort in this domain [8]

Data Stores

Big data analytics require clusters of data storage for effective output. As for very large-scale datasets traditional relational database models are not designed therefore in big data analytics performance issues arises. As a solution, No-SQL databases are preferred over SQL databases for processing due to the ability of horizontal partitioning of data, extensive processing capability and better performance [9]. Databases with NoSQL can be classified into following three different data store formats:

1. Document-oriented: Documents such as PDF or MS Word and several different formats such as Java Script Object Notation (JSON) and Extensible Markup Language (XML) are stored in document-oriented data stores [2]. One document stored in data is equivalent to a row in a relational database and the related query is applied to the document contents. Data can be stored in different domains.

2. Column-oriented: The database stored the data in column format .These databases store data in columns along with attributes rather than in rows 10]. Examples are BigTable, HBase and yperTable. One of the challenges in Column-oriented databases is the difficultly in data profiling, which needs further investigation [11].

3. Graphs database: This database is based on graph theory. The nodes and edges in graphs represented relations and their link to represent and store data in database. Dryad is an example of Graph database. However, selection of appropriate graph platform is still a challenge for researcher.

4. Key-value: Very large datasets are handled in Key-value data stores format. The data is accessed via a key using different algorithm [12]. One of the examples of such systems is Dynamo.

ISSUES AND CHALLENGES WITH BIG DATA

Big Data refers to huge volumes complex data structures and non- uniform distribution of data that cannot be processed effectively with the traditional applications. These requirements require designs of computing frameworks, system architecture. The definition of Big Data, given by Gartner is, “Big Data is high-volume, and high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation”.

A more focused on the processing time of large scale data, which is significantly higher than small scale data sets. This lead to delayed analysis in time critical applications such as robotics, space science and healthcare [13]

Privacy is one of the major concerns issue in big data. In areas such as the credit card industry and traffic

management systems require quick response from the analytics to take appropriate action [14].

As privacy is one of important issue not much solutions that exist in this regard. And industry solutions to this privacy problem in big data will depends on research outputs in this area. For that purpose, it is important to review existing methods and solutions from both application

http://doi.org/10.29369/ijrbat.2023.02.1.0045

and theoretical point of view. Social networking sites like Facebook and Twitter, what's app etc are leading producers of big data. Therefore, a review of existing solutions, frameworks, measurements perspective is provided. While there exists many challenges in big data, this paper mainly focuses on survey on the privacy problem in it.

The Privacy Problem in Big Data

Privacy is the primary concerns when we talk about effective big data management. In previous studies concentrate on cryptography, communication and information theory. the classical privacy preserving techniques like cryptography cannot be directly applied to big data sets. Considering the very large size of big data, it is difficult to use existing cryptographic solutions effectively. Another limitation is imposed by limited processing and storage capacity of mobile devices, which make encryption and decryption a non-feasible solution [15]. Therefore, conventional cryptographic solutions are not useful for big data. privacy mainly based on as follows.

Data Clustering

Data clustering is one popular method for data processing where data divided without its label to form different groups. One of major issues with these clustering algorithms is dependency. It is based on one data format, and big data deals with unstructured data. Following are the different data clustering techniques.

a) K-anonymity

Protecting data using k-anonymity is quite simple and easy to understand. In k-anonymity method, different algorithms, models and frameworks are proposed to solve different privacy attacks. K-anonymity can be defined as the property which distinguishes each record from k-1 other records based on a quasi-identifier. In this method at least k records are required in each equivalence class to achieve anonymity. For example, if all records in a table satisfy k- anonymity condition, then for some value of k, a record can be identified with 1/k confidence if quasi identifiers are known. A method of two-level vertex anonymization against a neighborhood attack is proposed in [16].

b) L-diversity

k-anonymity technique is effective against identity disclosure and neighborhood attack but limited to provide safety against attribute disclosure. To overcome this new concept l-diversity was proposed for privacy technique. In this l-diversity attributes are divided into sensitive and non-sensitive attribute Machanavajjhala et al. [17] showed that an attacker can obtain the value of a sensitive attribute for a record if the diversity in that attribute is low (Homogeneity attack). When dealing with relational database two factor influences a privacy of any record one is uniformity in the key attributes of a table other is knowledge about a particular record in the table. If attacker identified record correctly then it be positive disclosure otherwise negative disclosure. This privacy limitation is not sufficient to prevent attribute disclosure.

c) T-closeness

To tackle with limitation in l-diversity technique, the idea of t-closeness is proposed. T-closeness on an attribute in equivalence class and table have a close distribution that is not more than a threshold t, therefore measurement of distance between two distribution like M and N is computed by Earth Mover's Distance (EMD). Li et al. [18] propose t-closeness as an improved model for privacy protection. This model requires the difference between a global distribution and local distributions Two distribution M and N with set of elements in each and ground distance between two element of each set is

$$WORK(M,N,F) = \sum_{i=1} \sum_{j=1} mdij$$

http://doi.org/10.29369/ijrbat.2023.02.1.0045

CONCLUSION AND FUTURE WORKS

This paper focus on the big data information and characteristics used in all over the world. Security and privacy preservation is more challenging task in big data. Mainly, this paper focused on the privacy problem and the techniques used to handle the anonymity of users. The security issue is being highlighted more. It is observed that while all existing techniques in big data work effectively for small-scale structured and uniform data, but unable to work on unstructured, disturbed, non-uniform and very large volume data set.

REFERENCES:

A. Gandomi and M. Haider, “Beyond the hype: Big data concepts, methods, and analytics,” International Journal of Information Management, vol. 35, no. 2, pp. 137–144, 2015.

A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. Ullah Khan, “The rise of ‘big data’ on cloud computing: Review and open research issues,” Information Systems, vol. 47, pp. 98–115, 2015.

U. S. Pushpa, “A Review of Big Data and Anonymization“http://www.bbc.co.uk/news/technology-43465968

P. R. B. B, P. Saluja, N. Sharma, A. Mittal, and S. V. Sharma, “Cloud Computing for Internet of Things & Sensing Based Applications,” Sensing Technology (ICST), pp. 374–380, 2012.

S. Srivastava, “Appraising a Decade of Research in the Field of Big Data ‘The Next Big Thing,’” Computing for Sustainable Global Development (INDIACom), no. 2014, pp. 2171–2175, 2016.

J. Quackenbush, “Microarray data normalization and transformation,” Nature Genetics, vol. 32, no. december, pp. 496– 501, 2002.

J. Dean and S. Ghemawat, "Mapreduce: Simplified Data Processing on Large Clusters," Communications of the ACM, vol. 51, no. 1, pp. 107-113, 2008.

R. Casado and M. Younas, “Emerging trends and technologies in big data processing,” 2014.

S. Venkatraman, S. Kaspi, Kiran Fahd, and R. Venkatraman, “SQL Versus NoSQL Movement with Big Data Analytics,” International Journal of Information Technology and Computer Science, vol. 8, no. 12, pp. 59–66, 2016.

D. J. Abadi, P. A. Boncz, and S. Harizopoulos, “Column-oriented Database Systems,” Proceedings of the VLDB Endowment, vol. 2, no. 2, pp. 1664–1665, 2009.

F. Naumann, “Data Profiling Revisited,” vol. 42, no. 4, 2013.

M. Seeger, “Key-Value stores: a practical overview,” pp. 1–21, 2009.

B. Baesens, R. Bapna, J. R. Marsden, J. Vanthienen, and J. L. Zhao, “Transformational issues of big data and analytics in networked business,” MIS quarterly, vol. 38, no. 2, pp. 629–631, 2014.

S. Amini and C. Prehofer, “Big Data Analytics Architecture for Real- Time Traffic Control,” Models and Technologies for Intelligent Transportation Systems (MT-ITS), 2017.

S. Yu, “Big Privacy: Challenges and Opportunities of Privacy Study in the Age of Big Data,” IEEE Access, vol. 4, pp. 2751–2763, 2016.

B. Zhou and J. Pei, “The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks,” Knowledge and Information Systems, vol. 28, no. 1, pp. 47– 77, 2011.

http://doi.org/10.29369/ijrbat.2023.02.1.0045

A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 1, no. 1, 3–es, 2007.

N. Li, T. Li, and S. Venkatasubramanian, "T-closeness: Privacy beyond k-anonymity and l-diversity," in2007 IEEE 23rd International Conference on Data Engineering, IEEE, 2007, pp. 106–115.

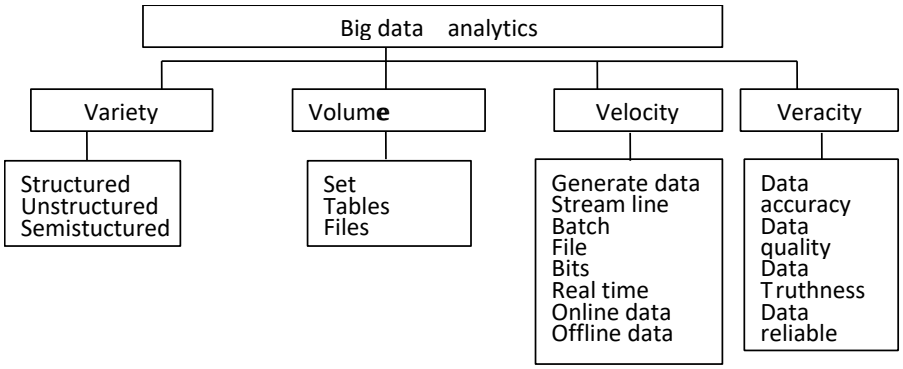

Fig1.Big Data Charecteristics

http://doi.org/10.29369/ijrbat.2023.02.1.0045