BIG DATA: PRIVACY AND INCONSISTENCY ISSUES

ArulMurugan R[1], Anguraj S[2], Boopathi R[3]

[1]Student, Information Technology, K.S.R College of Engineering, Tamilnadu, India
[2]Assistent Professor, Information Technology, K.S.R College of Engineering, Tamilnadu, India
[3]Student, Information Technology, K.S.R College of Engineering, Tamilnadu, India

Abstract

In the current Big Data world we are literally drowning in data that has been generated every day. Big data refers to data volumes in the range of exabytes (10^{18}) and beyond. Such volumes exceed the capacity of current on-line storage systems and processing systems. Data, information, and knowledge are being created and collected at a rate that is rapidly approaching the exabyte/year range. But, its creation and aggregation are accelerating and will approach the zettabyte/year range within a few years. Such massive volumes of data generated are new to storage industry. Many concerns have been raised to formulate the data issues. Big Data will be the only key to overcome all these data issues. In this paper we are going to provide a closer look towards various privacy issues that has been haunting various Big Data security later we provide an analysis about the various inconsistencies that has an impact on Big Data analytics.

Keywords: *Big Data, privacy in Big Data, Big Data security, Inconsistencies in Big Data, security issues in Bid Data.*

--***--

1. INTRODUCTION

The Big Data has become a catch phrase in the data field ever since the traditional databases have failed to handle Big Data. Even though Big Data is in its early days, strive for Big Data has already begun as we are getting ready to process petabytes of data. Meanwhile Big Data analytics is also reaching a different dimension as it is said to be the next important phenomena in the data industry.

The term Big Data refers to massive collection of structured and unstructured data that are not suitable for processing in a traditional database. Due to which Big Data analytics has been given affinitive importance. The Big Data is not just a large volume of data; it is actually a data treasure with valuable insights that are hidden inside. These insights are the current need for different purpose starting from decision making to scientific research. Hence the Big Data is the only solution to address all their concerns as the current growth of data industry has grown to an unexpected level in recent years.

2. CHARACTERISTICS OF BIG DATA

The Granter's Dough Laney [1] describes that Big Data has three major characteristics namely volume, variety and velocity. Another important characteristic that has been under lot of speculation is the veracity. The characteristics of Big Data are summarized in table 1.

Table -1: Characteristics of Big Data

Characteristics	Description
Volume	The volume refers to the amount of data. It's been increasing exponentially. They are generally data at rest.
Velocity	Velocity is the speed in which the data has been generated and processed. They are generally data in motion.
Variety	Variety refers to the various data formats and structures like text, numerical data, multimedia data etc.,

3. ORIGIN OF BIG DATA

The Big Data is generated from the various sources like social media and networks, many scientific instruments and business process which generate data continuous data in different forms. Moreover temperature devices, mobile devices, sensors track data with respect to time.

The logs such as transmission logs, server logs, and event logs are sources of Big Data that are currently employed in Big Data analytics.

4. EMPLOYING BIG DAT. FIELDS A

The reason behind providing more emphasis on Big Data is that they have become an integrated factor for future enhancement in many domains. The various sectors such as life and physical sciences, healthcare, transportation, banking, finance, government [5] has started employing Big Data to extract the necessary insights.

The business areas such as innovation in business, education, decision making must consider Big Data analytics to stay among the competitors. The Geographical Information Systems are widely used to identify the crime partners [6] and the public health partner [7] employs Big Data analytics.

The data that has been collected from various formats from various sources are cleansed, analyzed, summarized and finally visualized before it can employed for Big Data analytics to obtain insights about the particular domain.

5. ISSUES IN BIG DATA:

The Big Data comprises of several issues such as storage issues [1], management issues, privacy issues and inconsistency issues.

Among these issues the privacy and inconsistency issues are the major rising up issues which are going to address.

5.1 Privacy Issues

Nowadays we share our personal information such as our likes, dislikes and our whereabouts more frequently over the internet. These footprints may lead to various privacy issues like location disclosure, personal information disclosure.

5.1.1 Need for Privacy

Due to invariable increase in the internet usage the privacy remains an unsolved mystery still now. These privacy concerns generally occur due to illiteracy in privacy.

5.1.2 Challenges of Privacy

1) Individual privacy*:* When the information about a particular person is merged together with a very large datasets, there exists a fair possibility of identifying the various new hidden insights about the particular individual which he/she is unaware of [3].

2) Geo Information Privacy*:* While dealing with the GIS information the location disclosure through various sources can become a privacy threat.

3) Business Privacy*:* The various data and the logs such as transaction logs, server logs can reveal confidential information that would provide an opportunity to misuse the information. For example if company A and B are business competitors then both A and B must not reveal any data between each other as this might result in loss of business privacy.

4) Log Storage Privacy*:* The corporate companies usually have to maintain huge volume of logs for future references. If these logs accessed in improper manner, it could result in access to denied systems.

5) Under Privileged Privacy*:* By analyzing the large datasets it is easy to identify the under privileged ones. The current social media and other resources make this much easier. Hence under privileged privacy can be at stake.

1) Individual Attacks*:* The various forms of the individual attacks are discussed below.

The major privacy issue is getting the acknowledgement from the user whose information is taken for analytics. Here the problem occurs when the user initially believes that his/her personal information is secure but later if he/she feels that no longer privacy, then various privacy laws and data collector must guarantee the information is removed.

Next, issues that if several datasets when merged together can hold some unique information about the individual (user). The more focused attack is the identification of the single entry from the dataset in such a way that the particular individual information is obtained with higher confidence.

In the targeted attack one or more details of a given individual can be identified. This attack is said to have higher impact on privacy.

2) Industry Attacks*:* The most common industrial attack in privacy is creating the confusion and distraction in the information which in turn makes the available datasets to less useful for analyzing the business strategies.

The correlated information in the datasets may sometimes be able to reveal the self-calculated information.

The data sets containing confusing and distracting content can lead to improper insights that result in decreased decision making capability. For example if company A and B are competitors that follow same datasets. Suppose if A passes the datasets which contains confusing and distracting information to B. Then obviously company B’s decision making capability will automatically decrease.

5.2 Inconsistency Issues

When several large datasets are combined together which has a different origin the inconsistencies are bound to occur. Moreover the real world is prone to have inconsistent data. Generally the inconsistency in data occurs due to conflicting information. If these inconsistencies are not handled properly then the data sets would not have any integrity.

The inconsistencies can be a major problem in various fields such as heuristics, problem solving, research analysis and business prediction analysis.

In order to get a clear picture about where the inconsistencies creep in we must first analyze the level of the contents in big data.

5.2.1 Various Levels of Big Data

The following are the three broad levels of big data.
1) Data*:* The data is the preliminary level in big data. They are basic and less structured form of the content in big data. Generally the data are text, graphical contents, video, audio, symbols and numeric.

For instance consider the data “1986” it may refer to year or registration id or amount or anything that can be numeric. Due to this character the inconsistencies are hard to identify.

The inconsistencies at this level are very hard to identify. The only remedy is to make sure that data is collected by considering the integrity factor.

2) Information*:* The information is partially structured hence it is to an extent easier to identify the inconsistency. The information usually provides an extra value to the data.

Eg1: “Year: 1986” identifies the data to be a year. Hence year can only be numeric due to which inconsistencies can be eliminated to an extent.

3) Knowledge*:* They are well structured and more enhanced representation among the contents in the data sets. Mostly the inconsistencies are eliminated at this level but at rare scenario the inconsistencies may occur. Hence care must be taken while converting the content from data to knowledge.

5.2.2 Types of Inconsistencies

In the process of converting the data into knowledge the inconsistencies must be eliminated to perform an effective analysis. Among the various inconsistencies that are available the most common inconsistencies are discussed below.

1) Text Inconsistencies*:* The Text inconsistencies usually originate from various sources like emails, blogs, social media etc. These inconsistencies can highly alter the integrity of the data. Whenever two texts are referring to the same event or entity, then they are said to be co-reference. The co-reference is a mandatory condition for text inconsistencies. Various types of inconsistencies are summarized in table 2.

Table -2: List of text inconsistencies [16]

Type	Example
Synonyms	Bob is a brilliant student. Bob is a bright student.
Antonyms	Bob is a bright student. Bob is dull student.
Disagreeing	Bob has $100. Bob has $1000.
Contradictory	Beta was developed in 2014. Beta was deployed in 2013.
Complementary	Bob will graduate this year. Bob will not graduate this year.

The other texts in consistencies are inheritance, asymmetric, anti inverse, mismatching, etc.

2) Functional Dependency Inconsistency*:* The functional inconsistency may develop if the condition for an operation that requires the integrity constraints is not satisfied. The functional dependencies may arise in the following three conditions. The table 3 describes the functional dependencies.

Table -3: Functional Dependency Inconsistencies [13], [17]

Single FD	Violation of single functional dependency
Multiple FD	Violation of multiple functional dependency
Conditional FD	Violation of conditional functional dependency

3) Temporal Inconsistencies*:* Whenever the datasets contains temporal attribute then the temporal inconsistencies may occur. Since the temporal data deals with reference to time, the conflict can occur easily. The time interval relationship between conflicting data items can result in partial temporal inconsistency (or) complete inconsistency.

3.1 Partial Inconsistencies*:* The time intervals of two inconsistent events are partially overlapping. In a temperature time series data, a temperature recording of -4ºc in April in South India is inconsistent with the context.

3.2 Complete Inconsistencies*:* The time intervals of two inconsistent events coincide or satisfy containment. In ECG a prolonged period of low value output in ECG is inconsistent with the normal heart rhythm pattern [11].

4) Spatial Inconsistencies*:* The spatial data consist of geometrical properties and have various spatial relations. The violation in the spatial constraints results in the inconsistencies. The spatial inconsistency may occur in reference to spatial relations such as topological, directional and distance [14].

5.2.3 Handling Inconsistencies:

As mentioned in earlier, the inconsistencies are difficult to handle in spite of the form of the content. The datasets must be cleansed before it is handled so as to eliminate maximum possibility of inconsistency occurrence.

The various artificial visualization tools, machine learning methodologies, pattern analysis should be performed to eradicate inconsistence in data.

6. CONCLUSIONS

In this paper we have stated various privacy and inconsistency issues that continue to haunt Big Data analytics. The privacy issues should be address by strictly enforcing the privacy laws and by also creating a secure system while handling user data.

Similarly the inconsistencies must be eliminated as early as possible to improve data quality there by providing efficient data analytics.

We have planned to provide solutions for various privacy and inconsistency issues as occur future work.

REFERENCES

[1]. Stephen Kaisler, Frank Armour, J. Albero Espinosa and William Money, “Big Data: Issues and challenges moving forward”, 2013, 46th Hawaii International conference on system services, pp.995- 1004.
[2]. Du Zhang, “Inconsistencies in big data”, Proc. 12th IEEE International conference on cognitive informatics & cognitive computing [ICCI*CC’13], pp.61-67.
[3]. Meiko Jenson, “Challenges of Privacy Protection in Big data Analytics”, Big Data (Big Data Congress), 2013 IEEE International Congress on IEEE, pp. 235-239.
[4]. Linna Li, Michael Good Child, “Is Privacy still an issue in the era of big data - Location disclosure in spatial foot prints”
[5]. J. Manyika, M.Chui, B. Brown, J. Bughin, R. Dobbs, C.Roxburgh, and A. H. Byers, Big Data: the next frontier for innovation, competition, and productivity, McKinsey Global Institute, June 2011.
[6]. Cromley, E. K., &McLafferty, S. L. (2011), GIS and public health, Guilford Press.
[7]. Chainey. S, & Ratcliffe. J (2005), GIS and crime mapping Wiley.
[8]. Du Zhang and M. Lu, Inconsistency-induced learning for perpetual learners, International Journal of Software Science and Computational Intelligence, Vol.3, No.4, 2011, pp.33-51.
[9]. D. Zhang, i2Learning: perpetual learning through bias shifting, in Proc. of the 24th International Conference on Software Engineering and Knowledge Engineering, July 2012, pp. 249-255.
[10]. D. Zhang and M. Lu, Learning through Overcoming Inheritance Inconsistencies, in Proc. of the 13th IEEE International Conference on Information Reuse and Integration, August 2012, pp.201-206.
[11]. V. Chandola, A. Banerjee, and V. Kumar, Anomaly detection: a survey, ACM Computing Surveys 41 (3) 2009, pp.1-58.
[12]. D. Zhang, On Temporal Properties of Knowledge Base Inconsistency, Springer Transactions on Computational Science V, LNCS 5540, 2009, pp.20-37.
[13]. M. V. Martinez, A. Pugliese, G. I. Simari, V. S. Subrahmanian, and H. Prade, How dirty is your relational database? -An axiomatic approach, in Proc. 9th EuropeanConference on Symbolic and Quantitative Approaches toReasoning with Uncertainty, Hammamet, Tunisia, LNAI 4724, 2007, pp.103-114.
[14]. A. Rodriguez, Inconsistency issues in spatial databases, in L. Bertossi et al (eds.) Inconsistency Tolerance, LNCS 3300, Springer-Verlag, 2004, pp.237-269.
[15]. Stonebraker, M. and J. Hong. 2012. “Researchers' Big Data Crisis; Understanding Design and Functionality”, Communications of the ACM, 55(2):10-11
[16]. M-C de Marneffe, A. N. Rafferty and C. D. Manning, Finding Contradictions in Text, Proc. of 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2008, pp.1039-1047.
[17]. W. Fan, F. Geerts, X. Jia, and A. Kementsietsidis, Conditional functional dependencies for capturing data inconsistencies, ACM Transactions on Database Systems, Vol. 33, Issue 2, June 2008.
[18]. Matthew Smith, Christian Szongott, Benjamin Henne, Gabriele von Voigt, “Big Data Privacy Issues in Public Social Media”, IEEE, 6th International Conference on Digital Ecosystems Technologies (DEST), 18-20 June 2012.