A Survey Paper on Security Issue with Big Data on Association Rule Mining

Prof. Asha Patel
Assistant Professor
Department of Computer Engineering
SAL Institute of Technology & Engineering Research Gujarat India

Abstract— Big data implies performing computation and database operations for massive amounts of data, remotely from the data owner's enterprise. Since a key value proposition of big data is access to data from multiple and diverse domains, security and privacy will play a very important role in big data research and technology [1] Data mining is defined as the process of extracting useful knowledge or information from large data repository. To provide privacy to the data is the major issue so that the third party is not able to access the sensitive information. In this paper I clearly define the privacy prevention for Association rule mining in Big data.

Key words: Big Data, Security, Privacy Preserving on Big Data Five V's

I. INTRODUCTION

Big Data is a wide range of large data sets almost impossible to manage and process using traditional data management tools – due to their size, but also their complexity. 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 behaviour and market behaviour. Association Rule mining is the technique of extracting frequent mining from correlation for database transaction.[5].

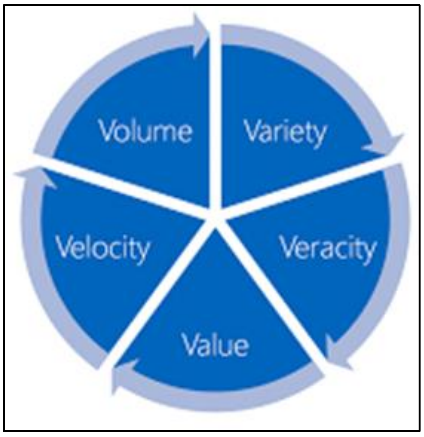

Fig. 1: Five v's of big data

Five v's of Big Data are [2]

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. There are two other metrics of defining Big Data.
4) Veracity 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.[2].
5) Value: 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 [2].

II. TYPES OF BIG DATA: [4]

There are two types of big data.

- Structured Data
- Unstructured Data. [4].

A. Structured Data

Structured Data are numbers and words that can be easily categorized and analyzed. These data are generated by things like network sensors embedded in electronic devices, smart phones, and global positioning system (GPS) devices. Structured data also include things like sales figures, account balances, and transaction data.

B. Unstructured Data

Unstructured Data include more complex information, such as customer reviews from commercial websites, photos and other multimedia, and comments on social networking sites. These data cannot easily be separated into categories or analyzed numerically. The explosive growth of the Internet in recent years means that the variety and amount of big data continue to grow. Much of that growth comes from unstructured data. When making an attempt to understand the concept of Big Data, the words such as ―maps reduceand ―Hadoop‖ cannot be avoided.[4].

III. PRIVACY PRESERVING IN DATA MINING USING IN ASSOCIATION RULE MINING[3]

In data mining the process of privacy preserving has played a vital role. It helps in providing the security to the sensitive information or knowledge and protecting information from unauthorized access without affecting the security of the data. Now a day's people are aware of the privacy intrusions on their personal data and they do not share their sensitive information to unauthorized people. Lack of privacy may generate the unintentional results. Several methods have been proposed in privacy but still it has its significance. The results of privacy preserving data mining algorithms is explained in terms of its data utility, performance, or level of uncertainty to data mining algorithms etc. There is no privacy preserving algorithms exists that exceed other algorithms on all possible criteria like utility, cost, complexity, performance, tolerance against data mining algorithms etc. In case of horizontally partitioned dataset the security is not provided for distributed privacy preserving association rule mining. Apriori and FP Growth algorithm are applied to analyze the performance and security. The results produced by the FP Growth algorithm are better than the Apriori algorithm. The combination of the horizontal and vertical partitioning of the dataset is known as the hybrid partitioning. When privacy is provided to both horizontal and vertical partitioned dataset in distributed and centralized scenario can improve the accuracy which overcomes the accuracy problem in the vertical partitioning. Association rule mining is used to group the related items and preserving the individual data privacy without compromise the accuracy of global data mining task and global association patterns were driven from the distributed data. Global rules are generated after the vertical partitioning of the dataset and percentage of missed rules and percentage of spurious rules were calculated. When two party algorithm is used with minimum support level, it will efficiently discover frequent itemsets without revealing individual transaction values. It will achieve good individual security.

IV. CONCLUSION

From the above survey paper one can get the idea about how the privacy and preserving can be needed in Big Data specially in association rule mining. Big data have various challenges related to security like-computation in distributed programming, security of data storage and transaction log, input filtering from client, scalable data mining and analytics, access control and secure communication.

REFERENCES

[1] Kalyani Shirudkar, Dilip Motwani Big-Data Security International Journal of Advanced Research in Computer Science and Software Engineering

[2] Raghav Toshniwal* Kanishka Ghosh Dastidar Asoke Nath Big Data Security Issues and Challenges International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Issue 2, Volume 2 (February 2015)

[3] Harpreet Kaur Shaveta Angurala Privacy Preserving in Data Mining using FP Growth Algorithm on Hybrid Partitioned Dataset International Journal of Computer Applications (0975 – 8887) Volume 147 – No.3, August 2016

[4] Subaira.A.S*, 2Gayathri.R, 3Sindhujaa.N Security Issues and Challenges in Big Data Analysis International Journal of Advanced Research in Computer Science and Software Engineering.