International Journal of Advanced Technology in Engineering and Science
Vol. No.5, Issue No. 06, June 2017
www.ijates.com
ijates
ISSN 2348 - 7550

Visualization, Security and Privacy Challenges of Big Data

Asia Mashkoor, Mohd Vasim Ahamad

Aligarh Muslim University (India)

ABSTRACT

Visualisation of information helps in obtaining the patterns from Data. Data may be structured, semi-structures or unstructured. There are various sources of data causing variety in data and also increase in the volume of data at very high velocity. Big data is become too massive to store and too dense to visualize. Now the Big data has become huge data. But ensuring the reliability in big data, as big data have 5Ws as its dimensions, is difficult. In this paper, we have discussed about the challenges and security issues with Big data as it have multiple dimensions to grow.

Keywords: Big Data, Visualization, Security and Privacy, Big Data Analytics, Information Security

I. INTRODUCTION

Exponentially increase in data lead to the word Big data. The data is created by everything at alarming velocity in enormous amount and various varieties. Now data is become so Big that Big can be replace by word Huge and can considered as Huge data. Gartner in 2011[2] proposed 3Vs characteristics of Big data that are Volume Velocity and Variety. Volume describe, the enormous amount of data are present. It becomes so massive that obtaining the affective pattern is becoming difficult. It is an open source, considering reliability in data also not possible. With enormous amount of data, there is not only the storage problem but also how to analyze the data in order to obtain the meaning full pattern with reliability.

Velocity- 2.5 quintillion bytes of data create every day. It is so much that 90% of the data in the world today has been created in the last two years alone[3].Velocity of data means both rate of producing data and velocity of processing data in order to meet the demand. Variety describes structured and non-structured data. It deal with the complexity of data and it includes hierarchical data, stock ticker data, tabular data (databases), e-mail, video, documents, metering data, still images, audio, financial transactions and more. Mostly data comes from social media website and mobiles phones. There are varieties of information regarding the particular topic which lead to the problem in analyzing the information hence not able to obtain the definite or exact pattern. Other 5Vs are Veracity, Variability, Value, Visibility, and Visualisation that are added to these 3Vs.

II. DIMENSIONS OF BIG DATA

Each incident data contain 5Ws dimension means where the data has created, when (at what time) it was created, what data has created, why it has create, how the data was created and for whom it was created. In [1] Zhang and co-author has proposed the 5WS density model. They proposed a function for incident data as:

$$f(t, x, y, z, p, q)$$

where "t" is a time stamp for each incident data, "x" represents types of data, "y" represents channel used for transmission, "z" show reason for incident of data, "p" represents creator of data and "q" represents who received the data.

Data comes from everywhere. An increasing amount of data is becoming available on the internet. Each and every one constantly producing the data. But there may be a chance that the data has created by un- authorize person by un-authorized source. The information created by data may be wrong, incomplete or created at multiple times. It also causes the problem of trustworthiness. Because of ambiguity in data, extracting the exact or accurate pattern from data is difficult. It is necessary to correlate the data or maintain in some hierarchy in order to obtain some hidden pattern.

Achieving the integrity, confidentiality and availability with dimension of big data is becoming difficult.

III. CHALLENGES OF BIG DATA

3.1 Noise, outlier, incomplete and inconsistency in data

As Big data have high dimensionality aggregation of data from multiple sources lead to accumulation of noise, incomplete, and inconsistency in data. This creates issues of heterogeneity, experimental variations and statistical biases, and requires us to develop more adaptive and robust procedures. High dimensionality creates large amount of data which emerges issues such as heavy computational cost and algorithmic instability.

Map Reduce is a frame work used for writing the applications for handling the massive, structured, semi-structured and un-structured data [4]. It has four actors, Job client, Job tracker, Task tracker, Distributed file system which perform their own specific operation on each item. Map reduce split the enormous amount of data into large datasets and arrange them into some logical order for parallel processing. Parallel processing speeding the multiple transactions at a time, improve the reliability in cluster and help in returning the solution more quickly with greater reliability.

Map reduces also used for secure computation in Distributed database.

3.2 Visualisation

Visualisation of Big data, with having lots of inconsistency is a challenging task. Visualisation helps in acquiring more knowledge and retrieving the pattern from the given data sets. Big data is too dense to visualise. We need to carefully select the dimensions of big data for the visualisation in order to get interesting patterns because if we minimize the dimensions, we may lost the some interesting patterns. But if use all dimensions it would become too voluminous to retrieve correct or exact patterns. Tableau, Gephi, Plotly, Excel 2016, Microsoft Power BI are some tools used for visualisation [5]. Other tools are [6] Microsoft Power BI, Dygraphs, Chart JS, Raw, and Leaflet which have their own limitations. They are discussed as follows:

i. **Tableau:** It provides the interactive platform for visualisation. It support all the type data formats and it is also publically available for free of cost but it have demerit that it provides its service online with 1GB of storage only.

ii. **Gephi:** It is an open-source network analysis tool used for handling the large and complex datasets. The network analysis includes

- Social Network Analysis

• Link Analysis

• Biological Network Analysis

But it only specified to graph visualisation.

iii. **Plotly:** It creates charts and dashboards online. But it also provides limited online uploading storage. No official offline client for Plotly is available.

iv. **Data wrapper:** It used for generating the graph and it is very easy to grasp the knowledge from the raw data [7].

v. **Microsoft Power BI:** It is cloud base powerful business analytics service. Power BI can integrate the 60 types of sources and start creating the visualisation within a minute. It also combine MS Office, Share Point and SQL Server [6]. We don't need programming skills for queries, it can support natural language.

vi. **Chart JS:** Visualisation performs in the form of charts. And it uses D3.js library at fronted code.

vii. **Dygraphs:** This is used for representing the large volume of data. It requires prior knowledge of programming skills.

viii. **Raw:** It is very simplest tool that allow the user to simple paste the data. Creating graph require very simple step. It uses the D3.js library [6].

ix. **Leaflet:** It can support to work on with mobile and desktop also. It used for visualising the data generated by conversation and high traffic. It include JavaScript library that helps the user for developing the interactive maps.

3.3 Visualization Techniques

Treemap, Circle packing, Sunburst, Parallel coordinate, Stream graph and Circular network diagram are some of the many visualization Techniques. Table I shows where these techniques are used.

TABLE I The Comparison of Different Techniques

Techniques	Large Data Volume	Data Variety	Data Dynamics
Tree Map	Y	N	N
Circle packing	Y	N	N
Sunburst	Y	N	Y
Parallel Coordinates	Y	Y	Y
Steam Graph	Y	N	Y
Circular Network Diagram	Y	Y	N

Other techniques are Network Technique, Temporal Technique, Multi-Dimensional, Three Dimensional (3-D), Two Dimensional (2-D), One Dimensional (1-D), and Geometric transformation. A single type of visualisation technique cannot be applicable everywhere. We have to choose wisely which technique to use when.

3.4 Security and Privacy Challenges

Big data goes on emerging with its characteristics in exponential way. With its exponentially increasing behaviour, it gives rise to many challenges which include visualisation, analysis, updating, querying, storage and

giving security to the data. Giving security to big data is equally challenging as visualisation is. Data goes on gathering from multiple sources, to protect it, is becoming an open issue. And to handle the collected data without security we cannot design a reliable system for big data analytics. Before big data gather more data from everywhere we have to take some effective measure for tightening the security to the big data.

Traditional techniques are not capable enough in dealing with big data for ensuring security and privacy. By providing encryption schemes, access control, firewall transport layer security, IDS (Intrusion Detection System), IPS (Intrusion Prevention System), and antivirus technology are not satisfying the demand of security. They all are can be broken. Four Big data security challenges are [8]

- Infrastructure security
- Data privacy
- Data management
- Integrity and Reactive security

3.4.1 Infrastructure Security

In dealing with infrastructure security following point should be consider:

- Architecture security
- Authentication
- Security for Hadoop
- Availability

i. **Architecture security:** In [9] Hadoop file system combined with network coding and multi-node reading, for securing the system. And another method for authentication issue in HDFS (Hadoop Distributed File System) is Kerberos mechanism based on Ticket Granting [10].

ii. **Authentication:** Provenance of data leads to the problem of authenticity. Many researchers have involved in tightening the security towards origin of data.

iii. **Security for Hadoop:** For securing the data from hackers who try to access the information in cloud, the author in [11] proposed control access mechanism established among user and name node which is the component of HDFS and manages data node. A hash function is generated from user side and name node side also. Both hash function are compared, if compared result are correct than access are provided to the user. SHA-256 used for generating hash function. RSA, Rijndael, AES and RC6 are used to restricting the hacker from accessing the whole data.

iv. **Availability:** By the extension of Hadoop implementation, it's not only increase the availability of data, it also replicate the data among clusters. In [12] author purposed a system help in achieving the high availability with having multiple active Name Node at same time.

3.4.2 Data Privacy

An enormous amount of personal information contain by big data. Social networking sites are one of the examples. Many e-commerce companies and hacker retrieve the information from there and utilise it for their benefits without. The methods for giving privacy to the Big Data are:

- Cryptography
- Anonymization
- Privacy Preserving Queries

- Privacy at social networks

i. **Cryptography:** For securing the data on cloud platform encryption, decryption, compression and authentication can secure the data up to the good extend but encryption capabilities are confined in providing extensive effort to secure the data. In [13] author proposed bitmap encryption scheme that guarantees users' privacy. In [14] author established new schema which use the cryptographic virtual mapping to create data path.

ii. **Anonymization:** In Big data era, Protecting Personally Identifiable information (PII) is increasingly difficult. Personal data must be anonymized (de-identified) and transferred into secure channels [15]. Data anonymization is the process of hiding and protecting the information. In this information get sanitize by either encryption or removing personal identifiable information from data sets. In [16] author proposed Adaptive Utility based Anonymization for providing privacy which base on association mining.

iii. **Privacy Preserving Queries:** It concern with analysing the data in the secure manner without violating the privacy of data. One of the methods is encryption for secure analysis and another is secure keyword search mechanism over that encrypted data [17].

iv. **Privacy at social networks:** Social networks are the pool of personal information. For hacker it is the source of stealing the information used for their benefits and they may also get success up to the great extent by utilising these information. Many social networking sites providing the privacy control mechanisms for restricting the other unauthorized users. Much legislation proposed by author and government that may use for providing the protection of personal information privacy.

3.4.3 Data Management

Data management is focuses on security at

- Storage level and
- Algorithm uses for sharing the data

i. **Storage level:** In [18] author proposed the method for protecting the data owner's privacy by introducing the parameter used for measuring the acceptable level of privacy. IoT based devices storing the data in clouds. Hadoop system assure the security of information in clouds. Another method used for security in HDFS (Hadoop distributed file system) are[10]:

- Kerberos mechanism based on Ticket
- Granting Ticket or Service Ticket.
- Bull Eye Algorithm used for monitoring the data in 360^{0} .

ii. **Algorithm uses for sharing the data:** Author in [19] proposed the system SafeShare which restricted the unauthorized access and allowing high confidentiality in data sharing.

3.4.4 Integrity and Reactive Security

This focuses on

- Development of methods to ensure confidentiality.
- Development of mechanisms to control big data access.
- Development of methods and means of remote integrity control for data.
- Development of data provenance methods for tracing and recording the origins of data.

- Development of methods for secure collection, processing and storage.

i. **Integrity:** Integrity is one of important dimension of security along with confidentiality and availability and achieving proper integrity with dynamic behaviour of big data is not possible In Big data era, it is important to integrate big data and Cloud computing together for safety application. Many authentication measures are used for securing the data in clouds like encryption, establishing the hash function between user and name node side and etc. In [20] author proposed the model for managing the integrity in Big data.

ii. **Attack Detection:** Fraud Detection and Network forensics are example of attack detection. Banks, credit cards and phone companies etc are mostly affected by fraud. For fraud detection the technologies Hadoop map reduce that includes Pig, Hive, Mahout, and RHadoop and other are MapReduce that is used by the WINE and Bot-Cloud2, are used for securing in Big data.[21] Forensic investigator continuously involves in visualizing and monitoring the traffic and it is the challenging task for them. In [21] author proposed the Bayesian Classification mining technique for detecting the level of attacks.

iii. **Recovery:** There are not such researchers found that effectively recovery when a disaster occurs. Many system has design for providing the full security but if they get failed, having very less quality to get recover with time as disaster occur. In [22] some technologies are recommended by author for recovery from disaster.

IV. CONCLUSIONS

In this paper we have discussed about the Challenges and issues of security and privacy in big data that can be exploited by hacker. Because of traffic in data provenance, big data still need more special kind of concern on security. We have also compared and discussed about some visualization tool and technology. This paper can help people to choose more accurate tools and technologies for visualizing the data.

REFERENCES

[1] J. Zhang, Z. Meng, and M. L. Huang, "BigData visualization: Parallel coordinates using density approach," 2014 2nd Int. Conf. Syst. Informatics, ICSAI 2014, no. Icsai, pp. 1056–1063, 2015.

[2] Stamford, "Gartner Says Solving 'Big Data' Challenge Involcves MoreThan Just Managing Volumes of Data", posted on June 27, 2011, http://www.gartner.com /newsroom/id/1731916.

[3] https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html.

[4] C. Iii, "S Ection Vi : E Merging I Ssues," no. July, pp. 131–138, 2005.

[5] R. S. Raghav, S. Pothula, T. Vengattaraman, and D. Ponnurangam, "A Survey Of Data Visualization Tools For Analyzing Large Volume Of Data In Big Data Platform."

[6] S. M. Ali, N. Gupta, G. K. Nayak, and R. K. Lenka, "Big Data Visualization : Tools and Challenges," pp. 656–660, 2016.

[7] A. S. Syed Fiaz, N. Asha, D. Sumathi, and A. S. Syed Navaz, "Visualization: Enhancing big data more adaptable and valuable," Int. J. Appl. Eng. Res., vol. 11, no. 4, pp. 2801–2804, 2016.

[8] J. Moreno, M. Serrano, and E. Fernández-Medina, "Main Issues in Big Data Security," Futur. Internet, vol. 8, no. 3, p. 44, 2016.

[9] Y. Ma, Y. Zhou, Y. Yu, C. Peng, Z. Wang, and S. Du, "A Novel Approach for Improving Security and Storage Efficiency on HDFS," Procedia Comput. Sci., vol. 52, pp. 631–635, 2015.

[10] B. Saraladevi, N. Pazhaniraja, P. V. Paul, M. S. S. Basha, and P. Dhavachelvan, "Big Data and Hadoop-a Study in Security Perspective," Procedia Comput. Sci., vol. 50, pp. 596–601, 2015.

[11] P. Adluru, S. S. Datla, and X. Zhang, "Hadoop eco system for big data security and privacy," in 2015 Long Island Systems, Applications and Technology, 2015, pp. 1–6.

[12] P. Colombo and E. Ferrari, "Privacy Aware Access Control for Big Data: A Research Roadmap," Big Data Res., vol. 2, no. 4, pp. 145–154, 2015.

[13] M. Yoon, A. Cho, M. Jang, and J.-W. Chang, "A data encryption scheme and GPU-based query processing algorithm for spatial data outsourcing," in 2015 International Conference on Big Data and Smart Computing (BIGCOMP), 2015, pp. 202–209.

[14] H. Cheng, C. Rong, K. Hwang, W. Wang, and Y. Li, "Secure big data storage and sharing scheme for cloud tenants," China Commun., vol. 12, no. 6, pp. 106–115, Jun. 2015.

[15] M. R. Islam and M. E. Islam, "An approach to provide security to unstructured Big Data," in The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014), 2014, pp. 1–5.

[16] J. J. Panackal and A. S. Pillai, "Adaptive Utility-based Anonymization Model: Performance Evaluation on Big Data Sets," Procedia Comput. Sci., vol. 50, pp. 347–352, 2015.

[17] M. Kuzu, M. S. Islam, and M. Kantarcioglu, "Distributed Search over Encrypted Big Data," in Proceedings of the 5th ACM Conference on Data and Application Security and Privacy - CODASPY '15, 2015, pp. 271–278.

[18] Lei Xu, Chunxiao Jiang, Yan Chen, Yong Ren, and K. J. R. Liu, "Privacy or Utility in Data Collection? A Contract Theoretic Approach," IEEE J. Sel. Top. Signal Process. vol. 9, no. 7, pp. 1256–1269, Oct. 2015.

[19] D. Thilakanathan, R. Calvo, S. Chen, and S. Nepal, "Secure and Controlled Sharing of Data in Distributed Computing," Comput. Sci. Eng. (CSE), 2013 IEEE 16th Int. Conf., pp. 825–832, 2013.

[20] I. Lebdaoui, S. El Hajji, and G. Orhanou, "Managing big data integrity," in 2016 International Conference on Engineering & MIS (ICEMIS), 2016, pp. 1–6.

[21] B. B. Jayasingh, M. R. Patra, and D. B. Mahesh, "Security Issues and Challenges of Big Data Analytics," pp. 204–208, 2016.

[22] Chang, V. towards a Big Data system disaster recovery in a Private Cloud. Ad Hoc Netw. 2015,35, 65–82..