A Comprehensive Survey on Security Challenges and Techniques in Big Data

[1]V. Vijayaganth, [2]P. Purusothaman,[3]M. Krishnamoorthi

ABSTRACT--*Data plays a major role in most of the business today. Due to smarter life style, data is generated at high volume, variety and velocity. This led to new problems in the current systems. Big data has evolved to address the problems in conventional system which cannot hold large volume of data. However, there exist problems associated to data security and privacy in the big data system. In the literature there are several solutions to address these problems. This paper presents an overview on the basic notion of big data and data analytics, some techniques and also presents some of security problems associated to big data.*

Keywords-- *Big data, Hadoop, MapReduce, Data Privacy.*

I. INTRODUCTION

Nowadays data has become important property for many industries including healthcare, education, government and engineering. The goal of organization can be achieved through extracting information through data on daily basis [1]. It is understood from the literature that more than 90% of data are created in recent years and it is observed that in just two days 5EB (Exabyte) of data are created by humans [2]. The huge volumes of data are created from social networks, Internet of Things (IoT) and multimedia [3] which is not capable of processing through existing traditional structural database. These lead to creating variety of data, like unstructured, semi structured, and structured data [4]. Data captured from various sources involve public data and private data as well [5].

The four major elements of big data security are “data privacy, data management and integrity, infrastructure security and reactive security” [6]. Big data technology is evolved due to volume or variety of data the security, privacy and quality of data is also an impact.

II. FUNDAMENTALS OF BIG DATA

Big data allows management of massiveamount of data and analyze it [7]. Big data changed the traditional system in terms of volume, velocity, variety, veracity and value. The challenging aspect of big data is to mine useful information from large quantity of data generated from different sources and in different formats. Most organization changed their storage of data [8], which allows understanding the business process with the new technology.

[1] *Department of CSE KPR Institute of Engineering and Technology Coimbatore, India, kv.vijayaganth@gmail.com*

[2] *Department of IT Bannari Amman Institute of Technology Sathyamangalam, India, purusothaman@bitsathy.ac.in*

[3] *Department of CSE Dr. N.G.P. Institute of,Technology Coimbatore, India, drkrishnamoorthim@gmail.co m*

Received: 22 Feb 2020 | Revised: 13 Mar 2020 | Accepted: 05 Apr 2020

Apache Hadoop is a framework to distribute large dataset across clusters using programming model [9]. Hadoop distributed file system (HDFS) is utilized to store data in server by NameNode and DataNode in Hadoop Environment. NameNode is used to store metadata and DataNode to store application data [10]. Big data uses MapReduce program to process and generate large amount of dataset. MapReduce use map function to create set of key/value pair and merges intermediate key/value pair to produce a solution [11]. MapReduce framework is one of the ways to process the data. There are several tools that consist of the Hadoop ecosystem, like Mahout, Pig, Zookeeper, Hive, Sqoop, Spark, or HBase.

A. Big Data Technology

Big Data is categorized into two main groups namely operational and analytics. In operational system, the data is captured and stored. The analytical system provides analytical capabilities for complex analytics of data that has been stored. Big Data Technology is a Software-Utility that is designed to analyze process and extract the information from an extremely complex and large data sets which cannot be processed by the Traditional Data Processing Software. The necessity of the Big Data Processing technologies is to analyze the huge volume of Real-time data and come up with Conclusions and Predictions to reduce the risks in the future.

The two major categorizes of big data technologies are namely operational and analytical big data technologies. The data that we generate in normal day to day activities will comes under operational big data. It can be online transactions, online shopping, social media, or the data from any specific compan etc. Analytical big data is little complex than the operational big data and it is an advanced edition of big data technologies. Few examples of analytical big data technologies are as namely medical records, weather forecast information and stock marketing.

B. Big Data Analytics

Big data analytics focuses on variety and quantity of data to reveal uncovered information. It process based on the market needs. Big data facilitates the organization to understand the importance of information.

C. Stages in Big Data Analytics

Before processing the data, data has to undergo the following steps to make efficient use of data. The stages are

Data Acquisition:

Data acquisition is the process of cleaning the data before it's been processed. Cleaning the data happens after gathering and filtering the data needed for the application. An unclean data will lead to huge deviation in the prediction in the results. Smart devices generate continuous data with the help of sensors. Since the unstructured nature of data, most data are discarded because it may contain irrelevant data [12].

Data Extraction:

Big data provides data extraction to analyze the data from a large dataset which is complex in a traditional system. Extracting data from a large data set for processing may require additional information called metadata. Metadata is data about data, for example data may include its timestamp or geo- location data. When these data is stored in a location then the process is known as ETL (Extraction, Transformation and Loading). The challenge here is to maintain security of the sensitive data like Personal Identification Information (PII). The possible chance to maintain security is to eliminate more secure data from extraction.

Data Collection:

Data collection is a process of gathering data for analytic purpose. Prediction of information from a single source may lead to fault prediction. So, data has to be collected from various sources to make better prediction. Few examples of industries that use big data analytics for prediction include hospitality, healthcare companies, public service agencies and retail business. EHR in medicine uses digital record of every patient to maintain data like demographics, medical history, and laboratory scan result. Weather prediction based on temperature is important to agriculture and commodity markets. The study of weather evolvement over time helps in prediction of climatic conditions, so that life’s can be saved from any natural disasters.

Data Structuring:

Data from various sources are collected and stored in a structured format for future use. Data structure is a process of organizing, storing and retrieving data in a structured way. Data structure type includes the array, files, record, table and trees. Relationaldatabase use structured way to store data. Common relational database application system with structure data includes online reservation system, retail, and any transactions. In relational database SQL queries are used to access the structured data. Unstructured data uses NoSQL which has no schema defined to store and retrieve the data. Text files, social media data, mobile data and sensor data are examples of human generated unstructured data [13].

Data Visualization:

In data structuring the data is converted to structured format and then query is passed to retrieve the data and made available to present data in visual format. A good visualization highlights useful information by telling a story and removing noise from data. The Human brain can much more easily process the most common types of visualizations like charts, tables, graphs, maps and dashboards visualization turns abstract data into visual patterns.

Data Interpretation:

Data Interpretation is a process of inferring insights from the processed data. The interpretation reflects on the quality and value of data. The accuracy of prediction highly depends on quality of data. To make data as quality data, data must be processed with advanced tools (analytics and algorithms). Descriptive analysis provides information of what happened by summarizing the past event. Diagnostic analysis provides the cause of the event. Predictive analysis gives insight what is the probability to happen in future. Prescriptive analysis provides the action to be taken to avoid an event occurring in future by analyzing the past.

III. SECURITY IN BIG DATA

Big Data security is to safeguard data and analytics processes from all the factors that could affect their confidentiality, both in the cloud and on-premise. In big data, the amount of data extends to terabytes or yottabyte. There are several ways to organize the implementation ofsecurity measures to protect their big data analysis.

Data privacy is most common concern in data preservation for any organization. A big data system contain large amount of personal information, organization using these data must be careful and restrict data only to authorized persons. There are several techniques to secure the confidentiality of data. Encryption technique is worldwide for security [14]. Encryption technique is used to translate a original plain text to cipher text. Text in the cipher text format is difficult for a human to read and infer the content in it. This technique uses key for symmetric and asymmetric key to encrypt and decrypt data. Hence, it makes difficult for hackers to read the message. The other possible way to secure the system is by installing firewall on big data server.

The security of the system can be achieved through access control. Access control in big data is a problem that provides only basic forms of access. A framework that integrates access control features is proposed to solve the above problem [15]. From the literature it is observed that few other researchers' focuses on MapReduce process to overcome the above issues [16].

The biggest dispute in big data is processing query on encrypted data. The figure 1 shows the current encryption scheme used in big data [12].

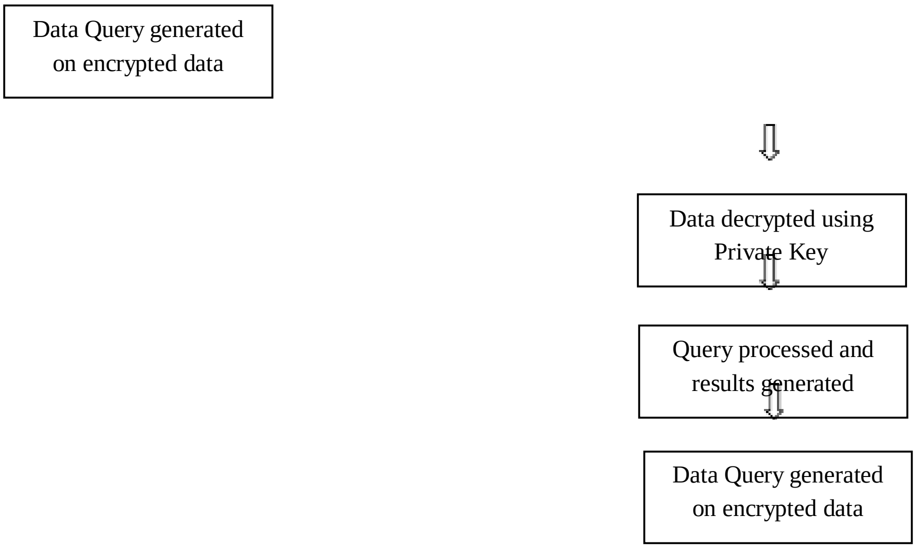
Data Query generated on encrypted data
Data decrypted using Private Key
Query processed and results generated
Data Query generated on encrypted data

FIGURE 1: Data Encryption System

multipartycomputation)[19]usesasecure

IV. CHALLENGES IN BIG DATA

scheme with*n* □ 4*t* □1parties, where *t*represents

The dispute in big data are listed as follows,

- How to select relevant and important data.
- Connectivity to various transaction points is another challenge.
- Big data need to work across multi disciplinary departments such as IT, Manufacturing, Marketing and Finance.
- Security issues related to big data collection. Various other challenges are
- Handling the increase in growth rate of the data
- creating an accurate and clear understanding of data in a timely approach
- Recruiting and retaining big data skilled talent
- Blending of dissimilar data source
- Data validation
- Securing the big data
- Organizational animosity

V. TECHNIQUES

Techniques used to solve big data security and privacy involve cryptography [17].

1. Homomorphic encryption

Homomorphic encryption techniques [18] can operate directly without decrypting cipher text. It uses SHE (Semi- Homomorphic Encryption) and FHE (Full- Homomorphic Encryption) schemes. SHE supports basic operation like addition, multiplication, where FHE supports calculation of arbitrary polynomial.

2. Secure Multiparty Computation

This scheme allows several parties to manipulate a function *f* in distributed environment using private keys. It checks the correctness of the result and ensures no fraudulence happened. AMPC (asynchronouthe number of honest parties.

3. Attribute-based encryption

ABE is one of the most commonly used encryption technique to access data securely in cloud environment [20]. In this technique both the key and cipher text with set of attribute is set by data owner and users will be given legal authority to use these attributes.

VI. CONCLUSION

A huge volume of data is generated from various sources like social media, IoT devices, banking data and human generated data. These data are generated at high velocity and data is of various kinds. Big data uses Hadoop

Received: 22 Feb 2020 | Revised: 13 Mar 2020 | Accepted: 05 Apr 2020

framework to distribute large dataset across the system. The stored data that comes from various sources are encrypted using cryptography algorithms and data enters the server through firewall to provide security at the network entry level. As data is stored the organization must ensure only authenticated users to access the personal information of data to provide data privacy. Access control can be used as security techniques to access the system and basic information from the protected data. Data privacy is not only the security issues in big data but there are also other security issues on which many researchers are working.

REFERENCES

1. Mayer-Schönberger, V.; Cukier, K. Big Data: “A Revolution that Will Transform How We Live, Work, and Think”; Houghton Mifflin Harcourt: Boston, MA, USA, 2013.
2. Sagiroglu, S.; Sinanc, D. “Big data: A review”. In Proceedings of the 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA, USA, 20–24 May 2013; pp. 42–47.
3. Hashem, I.A.T.; Yaqoob, I.; Anuar, N.B.; Mokhtar, S.; Gani, A.; Ullah Khan, S. “The rise of big data on cloud computing: Review and open research issues”. Inf. Syst. 2015, 47, 98–115.
4. Sharma, S. “Rise of Big Data and related issues”. In Proceedings of the 2015 Annual IEEE India Conference (INDICON), New Delhi, India, 17–20 December 2015; pp. 1–6.
5. Eynon, R. “The rise of Big Data: What does it mean for education, technology, and media research? Learn. Media Technol”. 2013, 38, 237– 240.
6. “Big Data Working Group; Cloud Security Alliance (CSA). Expanded Top Ten Big Data Security and Privacy”. April 2013.
7. Meng, X.; Ci, X. “Big data management: Concepts, techniques and challenges”. Comput. Res. Dev. 2013, 50,146–169.
8. Cumbley, R.; Church, P. Is “Big Data” creepy Comput. Law Secur. Rev. 2013, 29, 601–609.
9. Apache Hadoop. Available online: http://hadoop.apache.org/
10. Shvachko, K.; Kuang, H.; Radia, S.; Chansler, R. “The Hadoop distributed file system”. In Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST2010), Incline Village, NV, USA, 3–7 May 2010.
11. Dean, J.; Ghemawat, S. MapReduce: “Simplified Data Processing on Large Clusters”. Commun. ACM 2004, 51, 107–113.
12. RaghavToshniwal, et. al, “Big Data Security Issues and Challenges”, International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Issue 2, Volume 2 (February 2015).
13. T. Prasanth, M. Gunasekaran “A mutual refinement technique for big data retrieval using hash tag graph” Springer, Cross mark, November 2017

14. Yoon, M.; Cho, A.; Jang, M.; Chang, J.W. "A data encryption scheme and GPU-based query processing algorithm for spatial data outsourcing". In Proceedings of the 2015 International Conference on Big Data and Smart Computing (BIGCOMP), Jeju, Korea, 9–12 February 2015; pp. 202– 209.
15. Colombo, P.; Ferrari, E. "Privacy Aware Access Control for Big Data: A Research Roadmap. Big Data" Res. 2015, 2, 145–154.
16. Ulusoy, H.; Colombo, P.; Ferrari, E.; Kantarcioglu, M.; Pattuk, E. GuardMR: "Fine-grained Security Policy Enforcement for MapReduce Systems". In Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security, Singapore, 14–17 April 2015; pp. 285–296.
17. RongxinBao, Zhikui Chen, Mohammad S. Obaidat: "Challenges and techniques in Big data security and privacy: A review". Wiley, March, 2018, pp-1-8.
18. Pulse Global, " Big Data for Development: Challenges and Opportunities", NacionesUnidas, Nueva York, mayo: Global Pulse 2012.
19. Patra A, Choudhury A, Rangan CP, "Efficient asynchronous verifiable secret sharing and multiparty computation", J Cryptol. 2015;28(1):49-109
20. Sahai A, Waters B, "Fuzzy identity-based encryption" International conference on Theory and Application of Cryptographic Techniques. Nice, French Riviera, France. Springer-Verlag. 2005:457-473