Asian Journal of Computer Science and Technology
ISSN: 2249-0701 Vol.7 No.2, 2018, pp. 128-133

Big Data Security and Privacy Issues

Gayatri Kapil[1], Alka Agrawal[2] and R. A. Khan[3]
Department of Information Technology, Babasaheb Bhimrao Ambedkar University,
Lucknow, Uttar Pradesh, India
E-Mail: gayatri1258@gmail.com, alka_csjmu@yahoo.co.in, khanraees@yahoo.com

Abstract **- Big data gradually become a hot topic of research and business and has been growing at exponential rate. It is a combination of structured, semi-structured & unstructured data which is generated constantly through various sources from different platforms like web servers, mobile devices, social network, private and public cloud etc. Big data is used in many organisations and enterprises, big data security and privacy have been increasingly concerned. However, there is a clear contradiction between the large data security and privacy and the widespread use of big data. In this paper, we have indicated challenges of security and privacy in big data. Then, we have presented some possible methods and techniques to ensure big data security and privacy.**
Keywords**: Big Data, Big Data Dimensions, Security and Privacy**

I. INTRODUCTION

Big data continues to expand including ever increasing high volume data sets, often used software tools employed to capture, create and process data sets, are being tested to their maximum potential with respect to acceptable elapse time rates. The last decade has been witnessing an exponential growth in 'big data' as it is being produced by and for everyone from mobile devices, BPO centers, web servers social media avenues and etc. shown in the figure-1. What needs to be acknowledged here is that since a large amount of data is being created every second across the world, the quantum of data is only going to increase in future. In Wikipedia, big data is an all-inclusive word for any set of data collection so very large and complex that is difficult to use traditional data processing applications. A widely recognized definition relates to IDC: "large data technologies describes a new generation of technologies and architecture, which is high-speed capture, discovery and / or analysis" [1].

A comparative analysis shows that only half a decade ago a personal computer could store a data up to few hundreds of gigabytes while present scenario and upcoming trends show that digital information will shoot up by a staggering figure of 44.1 from .8 ZB to 35 ZB [2]. With this trend, it has become pertinent for all the stakeholders to gear up to the challenges thrown at by the sheer size of 'big data' and to take requisite measures for storage handling and processing of the same. Another aspect that needs appreciation is that big data must be processed and analyzed in a time-bound manner to ensure optimum value extraction and the outcomes must be presented in such a way that it can influence business decisions. However, to derive maximum gains, it is imperative that the organizations should work with the desired amalgamation of human resources, processes, and technology.

Some very large organizations, which are using big data, their biggest concern are the security of the cloud-based system, confidential information and storages data system. Fatal attacks on IT systems are becoming increasingly more difficult and a new malware is being made now and later, unfortunately, despite the enterprises working with Big Data, every problem is resolved but less concern on data security though security one really is important to worry and security is the extremely highly ranked priority for any enterprise. However, the discovery and utilization of extraordinary value of Big Data will increase the risk of security and privacy.

For example, 'Amazon monitors our shopping priorities and Google is our browsing habit of learning, while Twitter knows what's on our minds. Facebook joins all our social connections as well as all the information. Mobile operators not only know who we talk to, but what big data is next to those that analyze those people, to combat valuable insights for all signals, deposits, storage, and more, Which indicates growth by reusing personal data'[3]. If the Internet age has threatened security and privacy, then the age of Big Data puts them at greater risk. Large data security usually involves the use of Big Data, which implement solutions to increase the security, reliability and security of a distributed system. Big data privacy is centered on the protection of large data from unauthorized access and unwanted estimates [4]. It is well known that large data is a valuable source of information based on a strong and accurate security solution. However, Big Data often contains sensitive information that needs to be protected from unauthorized access and release. Of course, there is no challenge to security and privacy, if we cannot remove the value from Big Data.

Big data protection and privacy principles should be balanced against the additional social value of Big Data. In this paper, we have first reviewed the security and privacy in Big Data. Then, we have presented some potential methods and techniques to ensure security and privacy in Big Data.

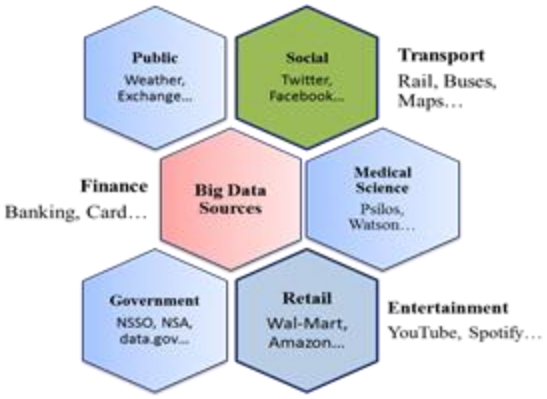

Fig. 1 Big Data Sources

Since large data consist of increasingly high-volume datasets, software tools used to capture, create and process the data sets are often used, tested their maximum capacity in relation to the time of acceptable time rates. A comparative analysis shows that only half a decade ago, a private computer could save up to a few hundred gigabytes of data, while the current scenario and the upcoming trends show that digital information ranges from .8 ZB to 35 ZB [2]. With this trend, it has become appropriate for all stakeholders to face the challenges raised by the large size of 'big figures' and to take the necessary measures for storage and processing. Thus, big data is a moving target and requires more attention to capture, curate, handle and process it. Though, initially, it was expected that the data was less and can be easily handled by RDBMS but now RDMS tools have failed to manage big data. However, there are many tools to handle big data; Hadoop is highly used among all. Hadoop is an open source framework developed by Apache Software Foundation. It is used for distribution, processing and running application for a large amount of datasets. The research in big data revolves around the techniques in order to manage and use big data efficiently & effectively.

A. Dimensions of Big Data

Big data become a most happening research topic in many areas, such as scientific research, finance and business since the era of computing began, the subject of large data has been intrinsic to the world of digital communication and information science. It is often understood as an assembly of the data set. It has got many definitions as different concerns from researchers, technical practitioners, and individuals. In Wikipedia, 'big data is defined as a set of data collection that is very huge and complicated in which it is difficult to use traditional data processing applications'. In the report of McKinsey & Company 'Big Data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyse' [5]. In both definitions, datasets can develop over time as large data in the form of technology development and can vary by region. It can be seen that the volume of data is not the only criterion for Big Data. In addition to the large scale of data, Big Data has some other features that define differences between themselves and big data. In a 2001 research report, the challenges and opportunities brought by the increased data are defined as 3V models, i.e., diversity, velocity and measurement [6]. Gartner analyst Doug Lane and many other enterprises and researchers used the "3V" model to describe the big data [7]. In 2012, Gartner updated his definition as Big data is high volume, high velocity and/or high diverse information assets that demand cost-effective, new forms of information processing, which include advanced insights, decision making, and Automation enables the process" [8]. From Gartner' definition, it can easily be inferred that big data is an information asset having high volume, velocity, and variety which require specific technologies and analytical methods to extract valuable information from it [9]. In the "3Vs" model, the word Volume is the size of the data set; Velocity indicates the speed of the data between the source and the destination, and Variety describes the range of diversity of data types and sources. These features that Big Data handles large amounts of data and uses different types of data, including unstructured data and attributes that were never used in the data database prior to Big Data. IDC has determined big data in 2011 that Big Data has provided us a new type of techniques and architectures to draw out the value from very large volumes of a wide range of data economically by enabling the high-velocity capture, discovery and analysis [1]. Researchers have explored big data by adding further vs. according to their needs. SAS (Statistical Analysis System) added two extra dimensions i.e. Convertibility and Complexity [10]. In addition, Oracle has defined the big data in terms of 4 V's, Volume, Velocity, Variety, and Value [11]. The characteristics of large data can be summarized as 4V, are similar to 3V models, and value means that large Data has great social value. The 4V model was widely recognized because it indicates the most important problem, which is a great, different type, and fast data datasets generated in large data. The Author also explored big data by adding future 3V's, Verbosity, Voluntariness, and Versatility that provide the simple and effective management of big data which used in value added applications and research environment.

B. Technology and Fundamental Tools for Big Data

Various framework and file systems have been developed for the use of storage, management and analysis of big data as shown in figure-2. Data cleaning, Data mining, Data collection, Data integration and Data visualization are the fundamental tools for big data computing and provide a fast engine for big data computing. Hadoop, Cassandra, MongoDB, Apache Hive, Hbase, Cloudera are the tools used for big data storage and management [15]. Cassandra is used for fast processing during very heavy writes and reads the environment and stored data which is very large to fit on the server, but still want a friendly familiar interface to it. HBase is used for real-time big data applications which contain billions of rows and millions of columns in tables built for low latency operations. Apache Hive is used for

analysis of large datasets stored in HDFS. Also, used for data summarization, query and ad-hoc analysis to process structured and semi-structured data in Hadoop. MongoDB is used for dynamic queries, defining indexes for good performance on a big database which makes applications faster and more efficiently at scale.

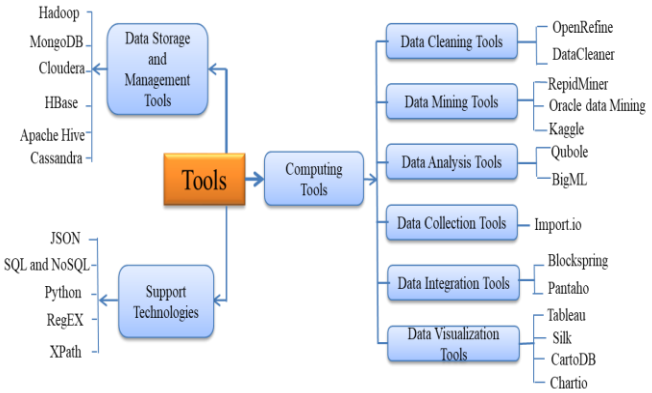

Fig. 2 Big Data Tools

Cloudera is used for administration management. On the other hand, Hadoop is highly used to handle all tools in big data. It is an open source framework developed by Apache Software Foundation and used for distribution, processing and running application for a large amount of datasets. It works on a master/slave architecture in which master is called as Name Node, and slave is called as Data Node. Name Node controls the access to the data by clients and data node manages the storage of continuously coming data on the nodes. Hadoop splits the file into blocks and then stores in the Data Nodes. Each data block is replicated to 3 different data nodes to provide high availability of the Hadoop system [12]. Hadoop Distributed File System (HDFS) is a core component of Hadoop and used to store input and output data. Hadoop Map Reduce is a central module which is used to collect the data according to a query [13]. It provides high-speed access within the application and is being used by big industries like Google, Yahoo, Facebook, and etc. [2]. Other Support technologies like JSON (Java Script Object Notation), SQL (Structure Query Language) & NoSQL, Python, RegEx and XPath give the main contribution in big data process. NoSQL is a technique for dealing data that is difficult to handle with traditional SQL [14]. Python is a high-level programming language and used widely for general purpose programming. RegEx is used for string pattern matching and XPath is a query language which is used in data extraction and XML nodes selection.

With Hadoop and other big data technologies coming out as key IT components for most of the organizations and business environments, there's a growing focus on finding the business benefits of big data analytics applications to help and justify investments in them. This raises a logical question; what are the biggest benefits that companies are getting or hoping to get from their big data initiatives? Many organizations, entrepreneurs, and IT professionals are replacing several traditional technologies with Hadoop framework and use it as the process, storage, database, business intelligence (BI) & IT, data exploratory analytic and data warehouse solution shown in figure-3. Hadoop framework facilitates urban planning, environment modeling, visualization, analysis, quality classification, securing the environment, computational analysis, biological understanding, designing and manufacturing process required by organizations and cost-effective models as well as the elegant exploration of the result. It was assisted with customer preferences, urban planning by visualization of environment modelling and traffic patterns, optimize their productivity, improve of equipment, increased production and improving efficiency, healthcare's professionals to prevent diseases and improving patient health, research organizations to obtain quality of research and physical science, medical science and scientific research to identify and prevent fraud, government agencies.

Hadoop framework provides processing and storage of the huge volume of data. But, there is no security against the external threats. All Industries pays more attention to the benefits of big data analytics tools and least on the concern of privacy and security issues. However, privacy and security protection is a set of concern, which should be considered before creating big data environments. Because all companies and businesses want same security features in their big data as they get in non-large data information system. There are the following important challenges that should be considered when dealing with the big data in respect of security perspective.

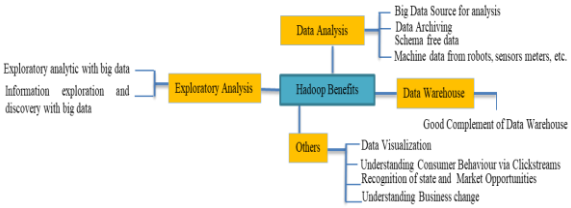

Fig. 3 Hadoop Applications

II. CHALLENGES OF BIG DATA

The challenges which are face are very large and complicated to handle. The disproportionate growth of the data and inabilities of organisations to execute or process it, have resulted in outbreak of challenges in big data. So, there are many issues arises in the big data environment. They are Processing Challenges, Data Challenges, Human Resources and Man Power Challenges, Technical Challenges, and Security & Privacy Challenges. Even a single issue encompasses a group of technical research problems and has its own task of surviving in big data and mainly focusing on security and privacy issues.

A. Processing Challenges

Managing data in large and fast-growing volumes has been a daunting problem for many decades. Big data's storage and transmission issues have become significant and are

magnified because of its characteristics like velocity, volume, and variety. When large datasets stored on the same infrastructure, controlling ownership and accessing are the important issues. The majority of generated data i.e., 80% of the world's total data is unstructured, and their sources are increasing rapidly, therefore, their alignment and sorting are creating more difficulties.

B. Data Challenges

Data Challenges in terms of handling are often classified regarding four Vs i.e. volume (data at rest), Velocity, Veracity & Variety. Velocity (data in Motion); means how fast the data is being generated and how fast it must be processed to meet the demand. Veracity (data in doubt); hence the challenge is to check its "accuracy" or "truthfulness" whether data which is being processed and later used is trustworthy or not. Variety (Data in many forms); Data comes from various sources and such information from every source cannot be analysed and processed in the same way.

C. Human Resources and Man Power Challenges

Human skills utilize information, but 60% of the total companies say that they don't have the possible skills. These skills should be extended to research, analytical, interpretative and creative. These skills are strongly needed to be developed in individuals with necessary training programs organized by different organizations.

D. Technical Challenges

The quality data will only lead to the better results, other irrelevant or useless data will end up taking more storages & time and will result in no positive output. The structured data is always highly manageable. When a broad range of application data is being collected, in such case it is very difficult to find out the valuable information and high-quality data.

E. Security and Privacy Challenges

The privacy of information has increased in the context of big data. For example, in case of electronic health records, there are strict laws made by the government in which they have clearly mentioned the gravity of different types of data which can be disclosed in different contexts [5, 16]. Security and Privacy management is a major problem in technical and social network. For example, consider data gathered from location-based services, for which the user needs to share the location with their service provider. An attacker guesses a source of his location from the query source. They can steal the data by copying it and sometimes the attackers keep it in hard disk or in laptops. There are so many methods to snatch the data by snooping attack, and Brute Force attack. Nowadays in the era of digitalization big data will play an extremely crucial role in paving the next level of generation in the world. Because we are observing that mobile has become a child's play and people has collected so many data to secure their memories and the next reason for using social media is that they can pay their full attention on balancing the security and privacy protection on social networking. To secure the data we can store data increase from single tier to multi store tier. Therefore, we have highlighted some challenges of big data security and privacy [17].

1. Most distributed system computations have only one level of protection, which is not recommended.
2. Non-Relational Database (NOSQL) is evolving actively, which makes it difficult to continue the demand for security solutions.
3. Automatic data transfer requires additional security measures, which are not often available.
4. When a system receives a large amount of information, it should be valid to be reliable and accurate; although this practice is not always the same.
5. Practicing information mining unethical IT expert can collect personal data without asking permission or without informing them.
6. Recommended detailed audit is not regularly included in big data because it contains large amounts of information.
7. Due to the size of Big Data, its origin is constantly monitored and not tracked.
8. Big data is stored in various nodes belonging to many clusters which are distributed around the all over world. All infrastructures between clusters and nodes are assured through ordinary public and private networks. Though, if someone can modify the inter-node communication it would be easy to extract valuable information. Therefore, it is a good challenge for big data tools to adopt new secure network protocols in order to protect interactions between different parties.

Henceforth in order to increase the security of data for future some cryptographic framework techniques and robust algorithm must be developed. Similarly some tools are developed like Hadoop, NoSQL and other technologies can be used for big data storages, management and data analysis. In our proposed work some ideas are given to overcome security and privacy issues in big data environment.

III. BIG DATA SECURITY APPROACHES

A. Zero Trust Data

In big data, the sensitive data is rapidly growing and moving throughout the whole world and being accessed by most of the people everywhere in organizations. Many of the organizations are looking to implement "zero trust data" to secure their sensitive data in data environment from being accessed by some malicious people. Recently this approach is using in companies like Dell, Deloite, Sap Hana, Hortanworks, Cloudera and Phemi Central. The Phemi central which is a new class of data warehouse for providing privacy and advanced features for data management and

governance that uses big data technologies to handle any volume and variety of data. This new approach can help to gain the customer's confidence in business as it tightens the security controls under the organization [18-19]. Data security is the key of a successful architecture of big data. Zero trust segments the networks based on data sensitivity which can help to understand surely about the sensitive data location. Nevertheless, organisations who face major struggle know about the residence of sensitive data related to these systems. Furthermore, new compliance mandates across the globe require security professionals to know where this data is stored and provide an audit trail to prove that is it protected.

B. Guardium Data Encryption v3.0

In increasing growth of big data, Artificial Intelligence (AI) and Machine learning that constantly provide new innovation for large organizations and business or business with ancient techniques which can help companies to analyse unstructured data more efficient and accurately to efforts and adopted to adopt new and more secure data encryption technologies. The right set of data encryption techniques and capabilities will be available for new and emerging threats can secure data against it. It can serve as a strategic tool set for all kinds of business environments whether it is adaptable, new technical platform or older, more harsh heritage technologies. Because different types of encryption help in protecting various data types, new approaches are introduced to gain sensitive information to increase the security of large data which is called "Guardium Data Encryption v3.0". This approach assists in protecting file and database data from abuse and compliance with industry and regulatory requirements. For this there is no coding or modification required for applications or databases and protects both structured and unstructured data. Solving solutions of large and complex environments it provides extensible protection for logging files, configuration files and other database outputs. Recently from this perspective, IBM has expanded its data encryption portfolio beyond the Guardian for file and database encryption to include IBM Gardium for file and data Encryption, Teradata encryption and tokening. Additionally, it has been extended to provide live data transformation that helps users encrypt files and increase and decrease efficiency in overheads without ever taking them offline and increases efficiency and decreases overhead, allowing organizations to protect more data more quickly [20-21].

IV. CONCLUSION

While the most important requirements for security and privacy in big data. This paper shows the use and characteristics of the world's extensive information. Some potential methods and techniques have been discussed to ensure security and privacy. Apart from this, we have mentioned that correlation of large data is the basis of data on key data protection and privacy issues. To increase security of big data by adding these methods using any approach or by combining these two approaches in big data environment. These approaches are introduced to overcome certain issues occurs in big data location. In future these approaches can also be implemented in other layers of big data Technology.

REFERENCES

[1] D. Laney, "3-D Data Management: Controlling Data Volume, Velocity and Variety", META Group Original Research Note, 2001.

[2] S. Vikram Phaneendra and E. Madhusudhan Reddy, "Big Data-solutions for RDBMS problems- A survey", in 12th *IEEE/IFIP Network Operations & Management Symposium* (NOMS 2010) Osaka, Japan, April 19-23, 2013.

[3] J. Shafer, S. Rixner and L. Alan Cox, "The Hadoop Distributed File system: Balancing Portability and Performance Analysis of Systems & Software (ISPASS)", *IEEE International Symposium*, pp. 122-133, March 2010.

[4] G. J, D. Reinsel, "Extracting value from chaos", IDC *iview*, pp. 1-12, 2011.

[5] M. Beyer, "Gartner says solving big data challenge involves more than just managing volumes of data", *Gartner*, 2011.

[6] L. Bonnet, A. Laurent and M. Sala, "Reduce, you say: What nosql can do for Data Aggregation and bi in Large Repositories[C]// Database and Expert Systems Applications (DEXA)", *22nd International Workshop on. IEEE*, pp. 483-488, 2011.

[7] L. Richard, Villars, W. Carl, Olofson and M. Eastwood, "Big Data: What is it and why you should care?", [Online] Available. www.idc.com, 2011.

[8] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A. Hung Byers, "Big data: The Next Frontier for Innovation, Competition, and Productivity", McKinsey Global Institute, 2012.

[9] M. A. Beyer and D. Laney, "The Importance of 'Big Data': A Definition", Stamford, *CT: Gartner*, 2012.

[10] V. Mayer-Schönberger and K. Cukier, "Big data: A Revolution that will Transform How We Live, Work, and Think", *Houghton Mifflin Harcourt*, 2013.

[11] M. Troester (2013), "Big Data Meets Big Data Analytics", [Online] Available: www.sas.com/resources/.../ WR46345.pdf, retrieved 10/02/14.

[12] Oracle, "Information Management and Big Data: A Reference Architecture", [Online] Available. www.oracle.com/.../ infomgmt-big-data-r, 2013.

[13] Oguntimilehin and E. O. Ademola, "A Review of Big Data Management, Benefits and Challenges", *Journal of Emerging Trends in Computing and Information Sciences*, Vol. 5, pp. 433-437, June, 2014.

[14] C.L. Philip Chen and C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data", *Information Sciences*, Vol. 275, pp. 314-347, 2014.

[15] Y. Gahi, M. Guennoun and H. T. Mouftah, "Big Data Analytics: Security and Privacy Challenges", *IEEE Symposium on Computers and Communication (ISCC)*, 2016.

[16] http://www.datacenterknowledge.com/archives/2016/01/19/nine-main-challenges-big-data-security

[17] G. Kapil, A. Agrawal and R. A. Khan, "A Study of Big Data Characteristics",in *Proc. of the Int. Conf. on Communication and Electronics Systems, IEEE Xplore*, pp.1-4, 2016.

[18] C. A. Ardagna and E. Damiani, "Business Intelligence meets Big Data: An Overview on Security and Privacy".

[19] https://www.businesswire.com/news/home/20150922006046/en/PHEMI%C2%A0Enables%C2%A0Data-Driven-Enterprises-Zero-Trust%C2%A0Data

[20] http://www.informationsecuritybuzz.com/articles/securing-big-data-starts-with-zero-trust/

[21] https://securityintelligence.com/news/ibm-announces-new-guardium-data-encryption-v3-0-portfolio/

[22] https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=an&subtype=ca&appname=gpateam&supplier=897&letternum=ENUS217-428.