JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

INTRODUCTION TO BIG DATA AND SECURITY ISSUES

Dr. Manish Kumar*[1],Mr. Prabhat Kumar*[2]
*[1]Assistant Professor,
University Department of Computer Applications,
Vinoba Bhave University, Hazaribag, Jharkhand, India
*[2]Assistant Professor,
University Department of Computer Applications,
Vinoba Bhave University, Hazaribag, Jharkhand, India

ABSTRACT

Big data, which refers to datasets too large to be managed with existing database management tools, is emerging in many major applications including web search, business computing, social media, genomics and the weather. Big data is a large volume, high velocity, and variety of information assets that require cost-effective ways of processing information that enable deeper understanding, decision making, and process automation. Big data presents a huge challenge for database research and data analysis. Big data poses a major challenge to design highly scalable algorithms and systems to integrate data and discover great hidden values from complex, diverse and large-scale datasets. This research paper intends to explore the security and privacy concerns occurring while securing the Big Data.

Keywords: - Big Data, Security, Privacy, Encryption, Big Data, Volume, Velocity, Variety, Storage etc.

I. INTRODUCTION

Big data is one of the hottest topics of the past two decades. This is due to the enormous amount of data that has been produced and consumed around the world. The great evolution of the Internet in recent years has led to this drastic generation of data.

Big data refers to large sets of raw data, and this data can be structured, semi-structured or unstructured. We cannot specify a single source where the data comes from, it is collected from many sources ranging from business transactions, photos, videos, search engines, social media, websites, apps and many more. This information is collected, recorded, stored and analyzed to gain meaningful insights that will help the organization grow [1].

Big data is much more than a large amount of data. It's a way to provide opportunities to use new and existing data and discover new ways to capture future data to make a real difference to industry players and make them more agile. With the advent of the Internet of Things (IoT), more and more objects and devices connect to the Internet and collect data on customer usage patterns and product performance.

Big data is characterized by 3 V's [3].

Data Volume-The volume of data refers to the enormous amount of data generated and collected by organizations, which can be in the order of petabytes or even exabytes. This large volume of data can be difficult to store and manage, especially if it is growing rapidly.

Data Velocity-Data velocity refers to the speed at which data is generated and processed. In many cases, data is generated and collected in real time, which means it must be processed and analyzed quickly to be useful. This can be challenging because it requires fast and scalable data processing systems.

Data Variety-Data variety refers to the many different forms that data can take, such as structured data in databases, unstructured data in text documents or social media posts, and semi-structured data in log files. This variety of data can make integration and analysis difficult with traditional methods and requires specialized tools and techniques.

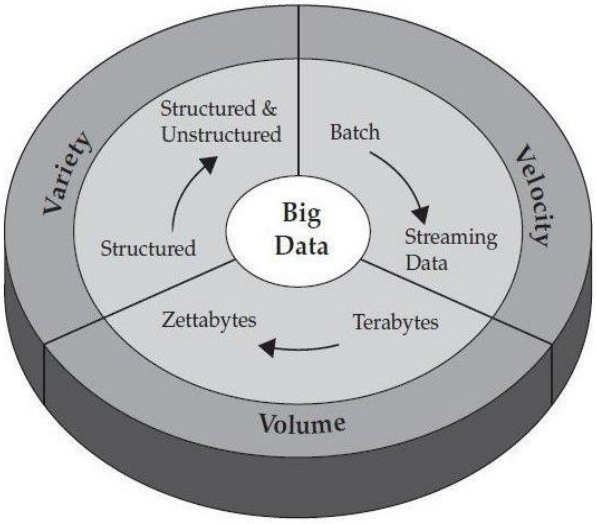

Figure-1 V's of Big Data

Key enablers for the appearance and growth of 'Big-Data' are:

- Increasing storage capabilities
- Increasing processing power
- Availability of data

A. Tools used in Big Data [2].

1. NoSQL :-Databases Mongo DB, Couch DB, Cassandra, Redis, BigTable, Hbase, Hypertable, Voldemort, Riak, ZooKeeper.
2. MapReduce:- Hadoop, Hive, Pig, Cascading, Cascalog, Caffeine, S4, MapR, Acunu, Flume, Kafka, Azkaban, Oozie, Greenplum.
3. Storage:-S3, Hadoop Distributed File System.
4. Servers:- EC2, Google App Engine, Elastic, Beanstalk, Heroku.
5. Processing:- R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, ElasticSearch, Datameer, BigSheets.

II. LITERATURE SURVEY

Vishal Joshi 2020 [4] presents some key concepts of privacy and security challenges specific to big data to bring renewed attention to hardening big data infrastructure. The document Big Data Security/Privacy Issues and Challenges provides an explanation of how many big data users face problems in dealing with day-to-day big data operations at various stages of the big data ecosystem. This article explains the research conducted to address the main problems and challenges of big data security and sheds light on what to consider when working with big data. R. Sumithra 2018 [5] presents a broad survey of security, privacy, confidentiality issues and challenges in Big Data and cloud computing. Big data provides good decision making, offers the benefits of a data warehouse and additional data analysis capabilities of distributed file systems. Big data is profitable. This document also resolved legal issues related to intellectual property rights, data privacy and integrity, cyber security, and the big data code of conduct. Jose Mura 2014 [1] Big Data security and privacy issues. This document discusses current security and privacy issues. There are many sources of unstructured data, including social media, sensors, scientific applications, surveillance, video and image files, Internet search indexing, medical records, business transactions, and system logs. P. Kamakshi Dec 2014 [6] Research on big data and privacy issues. In this article, the strength and applications of big data, as well as various privacy issues, are discussed. Lo'ai A. Tawalbeh and Gokay Saldamli In this article [7], existing multi-tier cloud architectures are discussed and a solution for big data storage, use of P2P Cloud System (P2PCS) for big data and hybrid model of mobile cloud based computing on the concept of cloudlets and apply this model to healthcare systems such as processing and analyzing case studies. Minit Arora and Dr. Himanshu Bahuguna (2016) [8] present research work that organizations used various data anonymization methods to ensure data security and privacy. The most common method of ensuring security and privacy is through verbal and written commitments. However, our past history has shown that this method is flawed. Passwords, controlled access, and two-factor authentication are a low-level

but routinely used technical solution to strengthen security and privacy when sharing and aggregating data in dynamic, distributed data systems. K.P.Maheswari, P.Ramya and S.Nirmala Devi (2017) [9] present a study and analysis of security threats levels in big data and cloud security. Problems related to big data are more acute in some sectors and in some government activities. The security issues of big data systems and technologies are also applicable to cloud computing because it is important for the network that connects the systems. JL Jonesston Dhas, S. Maria Celestin Vigila and C.Ezhil Star (2017) [10] designed a security and privacy protection framework for archiving health information using Big Data. Storing medical records as big data has many real-time problems. Among them is how to protect data in the cloud. Here's how to identify the record and how to protect health information from unauthorized users.

III. METHODOLOGY

In general, for any survey, there are a number of methods and techniques that are considered and used to conduct the survey. These methods mainly refer to qualitative methods. In fact, qualitative and quantitative methods are recognized as two primary research methods available for use by researchers. Both methods and techniques are quite different from each other. In this case, qualitative methods in the form of literature reviews were considered. Literature reviews are recognized as a common form of qualitative technique and can be said to be regarded primarily for their efficiency and simplicity. Indeed, it would not be wrong to say that literature reviews allow a researcher to ensure that the topic or concept in question is being explored effectively. For example, it presents and provides the researcher with an authentic and systematic method for exploring different investigations and studies. Through this in-depth exploration, the concept can be effectively detailed. In this case, a literature review was selected and conducted because it appears to fit the nature of the research. Indeed, if the time scale of the survey had been long, quantitative methods in the form of questionnaires could have been considered. Furthermore, it would have been considered if the nature of the investigation had been different. However, in this case, the use of questionnaires to carry out surveys was not considered appropriate. In fact, it can be said that its use has not been found to be effective. It would not have produced the necessary results. In fact, it can be said that if the questionnaires had been taken into account, this would have produced quantitative data. However, it would not have produced qualitative and conceptual information. Therefore, quantitative methods were not considered in this case. Generally, such methods and techniques are considered and used when there is a need to ensure that quantitative information is obtained on the subject in question. In this investigation, in fact, there was no need to obtain quantitative information. Instead, there was a need to evaluate the concept and, for this, the bibliographic reviews were recognized and considered the most appropriate technique to carry out the search. Indeed, it would not be wrong to say that the use of literature reviews made it possible to explore the topic effectively. Numerous articles from reputable journals were considered and carefully reviewed to perform the search to obtain all necessary information to be included in this study.

IV. BIG DATA SAFETY

Every day, the security of confidential information gains more and more attention. With this, we can understand that security occupies a very high place in any company's consideration. However, with the ease of adoption of web-based, Smartphone and cloud-based applications, it has become easy to access secret information across different platforms. These platforms are very vulnerable to hackers, especially if they cannot be managed properly.

Unlike before, companies now collect and use a lot of customer data. Lack of data security can lead to serious security issues and your organization's reputation will be at risk. When it comes to Big Data, financial and reputational things about a company can change overnight.

A. Big Data Security Problems [11] [12]

Usually, for any given data, the security problem arises if proper security measures are not taken, either by firewall or antivirus software, or both. These measures work effectively if the data is within the system. But what if the data goes beyond your system, could it be the cloud?

In many cases, computations in distributed systems have less protection, say one or two layers. This kind of security level is not recommended at all. At some point, connection security and access control encryption will be ineffective and inaccessible to the department. of IT who depend solely on it. Automated data transformation requires additional security standards, which are often not available. The suggested detailed audits are not performed regularly on big data due to the large amount of data involved. Non-relational databases often continue to evolve, making it difficult for security solutions to maintain what is needed. Conducting

unethical searches such as IT experts who practice data mining can also reveal a lot of personal information, without the knowledge of the users. Whenever a system requires large amounts of information, it absolutely needs to be validated to keep it reliable and accurate. Due to the size of Big Data, its origin is not always tracked or verified.

1. Fake data: One of the most important security issues in big data is the generation of false data. False data makes it impossible to detect other security issues in the system and can be the cause of loss of customer data. False signals from simulated data can complicate fraud detection and disrupt all business processes.

2. Distributed frameworks: Most big data implementations actually spread large processing jobs across many systems for faster analysis. Hadoop is a well-known example of an open source technology involved in this and originally had no form of security. Distributed computing could mean less data processed by everyone system, but it means many more systems where security issues can arise.

3. Non-relational data stores: Think of NoSQL databases, which often lack security (which is instead provided somewhat through middleware).

4. Storage: In the big data architecture, data is typically stored at multiple tiers depending on the business needs in terms of performance versus cost. For example, high priority "hot" data is usually stored on flash media. So locking down storage means creating a level-aware strategy.

5. Endpoints: Security solutions that extract logs from endpoints will need to validate the authenticity of those endpoints or the analysis won't be very useful.

6. Real-time security/compliance tools: Generate massive amounts of information; the key is to find a way to ignore false positives so that human talent can focus on the real violations.

7. Data mining solutions: Is the heart of many big data environments; they find models that suggest business strategies. For this reason, it is particularly important to ensure that they are protected not only from external threats, but also from internal users abusing network privileges to obtain sensitive information, adding another layer of big data security concerns.

8. Access Controls: As with corporate IT as a whole, it is extremely important to provide a system where encrypted authentication/validation verifies that users are who they say they are and determines who can see what.

B. Best Practices for Securing Big Data [13] [14]

Here are some ways to strengthen Big Data security:

- **Protect the data itself, not just the perimeter:** Focusing on securing the walls around data seems to be the goal of many organizations, with nearly 90% of security budgets spent on firewall technology. However, there are hundreds of possible ways to bypass a firewall; also through customers, suppliers and employees. All these people have the ability to bypass external cyber security and misuse sensitive data. For this reason, you need to make sure your security efforts are focused on the data itself, not just the perimeter.
- **Big Data Cryptography:** Another common technique used for data protection is data encryption. It is used to ensure the confidentiality of Big Data. Unlike typical encryption techniques, it should be noted that homographic encryption also allows computation of the data to be encrypted. As a result, this method guarantees the confidentiality of information, which allows extracting information through calculations and analysis.
- **Granular Access Control:** Granular access control is a computer concept that refers to the practice of granting different levels of access to a certain resource to certain users. Access determines what a user is authorized to do on a system. Control levels are used to prevent insider information from being tampered with, lost, or used maliciously in an organization. It is good practice to follow the principle of minimal access where users have access only to the parts of the system they need.
- **Security Surveillance:** There is no doubt that you need to ensure constant vigilance or detect security incidents and related issues and problems in real time. To ensure Big Data security surveillance, there are several solutions that companies can consider and use. These solutions include dynamic security threat analysis, security information and event management, and data loss prevention. In fact, these solutions are based on correlation and consolidation methods between different data sources. There is also an important need to perform regular audits for the enforcement and verification of different security policies and best practices for employees and users.
- **Data Access Monitoring:** In fact, there are more and more security issues and threats due to the increasing rate of data exchange in the cloud and between distributed systems. To address these challenges, it has been proposed to integrate controls at the data stage. However, this integration has also been found to be insufficient to combat security threats. Additionally, access controls must be granular so that access is limited by responsibilities and roles. There are often multiple methods to ensure data privacy and access control, including federated identity management, smart cards, and certificates. Constant monitoring of security threats can be efficient in the sense that threats and problems can be quickly identified and dealt with without major difficulties or problems.
- **Centralized Security Management:** Several companies use the cloud for data storage. The underlying goal is to leverage a

centralized security mechanism and standard compliance infrastructure [26]. However, achieving zero risk status is very difficult. The cloud tends to attract the attention of hackers and malefactors because it represents a base or center of confidential information.

V. BIG DATA APPLICATION IN DIFFERENT DOMAINS

A. Big Data in Education

1) Improve the grading system: Big Data allows educational institutions to track student performance in various subjects individually and collectively. They can develop solutions to help students achieve their goals. Analyzing student scores in different subjects helps institutions create curricula tailored to their students' learning abilities. Through data analysis, they can uncover the factors that affect student performance and provide effective solutions.

2) Career Guidance for Students: Students can use big data to determine which career path is right for them. Based on the student's area of interest and performance in the respective subjects, professors can guide the student on the right career path to choose. Career options are expanded for students through the use of big data. Schools can offer career guidance using big data.

3) Propose new learning plans: students have different learning abilities. Some can learn simply by reading while others can learn by writing. Students can also learn by watching videos or through other methods. Students are often forced to adapt to a defined learning structure, which can hinder their growth. This negatively affects students' academic performance. Learning plans can be created based on students' abilities using Big Data Analytics. Data analytics can help students identify their strengths and weaknesses and provide relevant study plans and course materials [15].

B. Uses of Big Data in Healthcare

1) Predictive Analytics in Healthcare: In healthcare, Big Data plays a crucial role in predictive analytics. With the help of predictive analytics, doctors can provide excellent treatment to their patients. Ensures patient safety. By identifying which patients are at risk for which diseases, doctors can make informed decisions that will improve patient health.

2) Electronic Health Records (FCE): In the medical industry, Big Data is mainly used to create electronic health records. Healthcare industries were facing challenges in managing the growing number of past medical records. Currently, every patient has their own medical history, such as a list of medications, medical reports, lab test results, etc. Healthcare industries can easily maintain and access patient data with the use of electronic health records. There is a separate file for each patient. It can be easily changed by the doctor and can be shared safely.

3) Real-Time Monitoring: Patients now receive excellent treatment from healthcare systems that monitor their health in real-time. There are many tools available that analyze patient data and suggest actions that clinicians should take. There are many wearable sensors that track patients' health, such as blood pressure, pulse, heart rate, etc., which can be monitored by doctors. Reduce unnecessary patient visits to hospitals [16].

C. Application of Big Data in the banking and financial sector

1) Real-Time Stock Market Insights: Machine Insights powered by Big Data allow financial industries to analyze stock rates by thinking about social and political trends that may affect the stock market. It allows monitoring of traces in real time, thus presenting the functionality to analysts to consider big statistics and make smart decisions.

2) Fraud detection and prevention: money laundering fueled by big data is significantly responsible for fraud detection and prevention. The security risk posed by the credit card is mitigated with analytics. When any secure and valuable credit card information is stolen, banks immediately block the card and transactions. Then, inform the customer about the security threats.

3) Trading Analysis: Banks, hedge funds and others in economic markets use Big Data for analysis of changes. Trading analytics helps the financial and banking industries support pre-trade screening, high-frequency trading, predictive analytics, and sentiment measurement [17].

VI. CONCLUSION

In this paper, we emphasize on the basics of Big Data. We also explain how various organizations approach big data issues. In general it can be said that while big data offers some significant benefits to firms, it also poses some major issues and challenges. There is also some research presented in this paper regarding these challenges, but it does not provide a complete and clear solution. There's some information and technologies that can add to the most relevant and challenging Big Data privacy and security issues. Safety and privacy concerns are largely related to the fact that a vast amount of personal information is freely available in digital form. Using customers' personal information, many organizations use Big Data for their own benefit, profit and for achieve their

goals. As part of the Big Data Code of Conduct, we also need to resolve legal issues related to intellectual property, Property rights, data integrity and cyber security. In this paper we also discuss the best practices for securing big data application in different domains. There are many surveys conducted in recent years based on which it will not be wrong to tell that Big Data technology is gaining ground in almost every sector.

REFERENCES

[1] J. Moura, "Security and Privacy Issues of Big Data," Handbook ofresearch on trends and future directions in big data and web intelligence., no. 20-52, 2015.
[2] https://hevodata.com/learn/big-data-security/.
[3] "https://www.analytixlabs.co.in/blog/characteristics-of-big-data/".
[4] M. V. Joshi, "Security/Privacy Issues and Challenges in Big Data,"International Research Journal of Engineering and Technology (IRJET),vol. 07, no. 06, 2020.
[5] R. Sumithra, "Security, Privacy Issues and Challenges in Big Data and Cloud," Special Issue based on Proceedings of 4th International Conference on Cyber Security (ICCS), 2018.
[6] P.Kamakshi, "SURVEY ON BIG DATA AND RELATED PRIVACY ISSUES," International Journal of Research in Engineering and Tech- nology, vol. 03, no. 12, Dec 2014.
[7] L. A. T. a. G. Saldamli, "Reconsidering big data security and privacyin cloud and mobile cloud systems," Journal of King Saud University – Computer and Information Science, 2019.
[8] M. A. a. D. H. Bahuguna, "Big Data Security – The Big Challenge," International Journal of Scientific Engineering Research, vol. 7, no. 12, Dec 2016.
[9] P. a. S. D. K.P.Maheswari, "STUDY AND ANALYSES OF SECURITY LEVELS IN BIG DATA AND CLOUD COMPUTING," International Journal of Innvative Research in Science and Engineering, vol. 3, no.02, 2017.
[10] S. M. C. V. a. C. E. S. J.L. Joneston Dhas, "A Framework on Security and PrivacyPreserving for Storage of Health Information Using Big Data," International Science Press, 2017.
[11] J. C. Ogbonna, F. O. Nwokoma and A. Ejem, "Database Security Issues: A Review," International Journal of Science and Research, vol. 6, no. 8, 2015.
[12]U. Sivarajah, M. M. Kamal, Z. Irani and V. Weerakkody, "Critical analysis of Big Data challenges and analytical methods," Journal of Business Research, vol. 70, pp. 263-286, 2017.
[13]Dona Sarkar, Asoke Nath, "Big Data – A Pilot Study on Scope and Challenges", International Journal of Advance Research in Computer Science and Management Studies (IJARCSMS, ISSN: 2371-7782), Volume 2, Issue 12, Dec 31, Page: 9-19(2014).
[14]Kogge, P.M.,(20-24 May,2013), "Big data, deep data, and the effect of system architectures on performance" Szczuka, Marcin,(24-28 June,2013)," How deep data becomes big data".
[15] http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf
[16] http://www.nessi-europe.com/Files/Private/NESSI_WhitePaper_BigData.pdf
[17] http://sites.amd.com/sa/Documents/IDC_AMD_Big_Data_Whitepaper.pdf