International Journal of Scientific & Engineering Research Volume 9, Issue 4, April-2018
ISSN 2229-5518

A Study on Big data security issues and challenges

[1]A.Dhanush, [2]S.SenthilMurugan

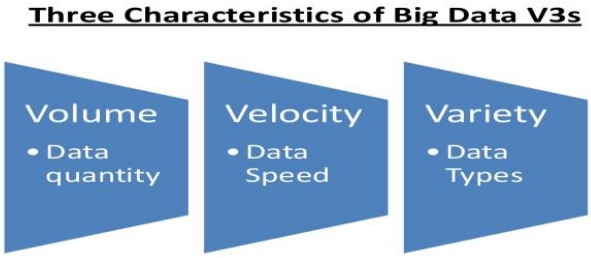
Abstract— The term of Big Data is now used almost everywhere in our daily life. The amount of data is world is growing day by day. Data is growing because of use of internet, smart phones and social networks. Big data is a collection of data sets which is very large in size as well as complex. Generally size of the data is pet byte and Exabyte. The big data change the way that data is managed and used. Some of the applications are in areas such as healthcare, banking, retail, traffic management, education and so on. This paper highlights important concept of big data and we discuss various aspect of big data. We define big data and discuss the parameters along which big data is defined. This includes the three v's of big data which are velocity, volume and variety.

Keywords— Big data, pet byte, Exabyte, velocity, volume, variety

1 INTRODUCTION

Big data is a collective term referring to data that is large and complex that is exceeds the processing capability of conventional data management systems and software techniques. However with big data come big values. With many big data analyzing technologies, insights can be derived to enable better decision making for critical development areas such as health care, economic productivity, energy, and natural disaster prediction. Security and privacy are great issues in big data due to its huge volume, high velocity, large variety like large scale cloud infrastructure, variety in data sources and formats, data acquisition of streaming data, inter cloud mediation and other. The use of large networks of computer increases the region of attack to an all new level of the entire system. The various challenges related to big data and cloud computing and its security issues and the reason why they crop up are explained later in detail.

1.1 Characteristics of big data



1.2 Volume

The world big in big data is due to the sheer size of big data that it actually means. It refers to the vast amounts of data is generated every seconds, minute, hour and day in our digitized world. Every minute 204 emails are sent, 200,000 photos are uploaded and 1.8 million likes are generated on Face book. On YouTube 1.3 million videos are viewed and 72 hours of video are uploaded. Its size is massive to the extent that they are measured by the like of pet bytes, Exabyte's and zetta bytes.

1.3 Variety

Today, data comes in all types of formats. Structured, numeric data in traditional databases. Information created form line-of-business application. Unstructured text documents, email, video, audio, stock ticker data and financial transaction. We need to find ways of governing, merging and managing these diverse forms of data.

1.4 Velocity

Velocity refers to the speed at which big data is created or moves from one point to another and the increasing pace at which it needs to be stored and analyzed. The processing of data in real time to match its production rate as it gets generated is the main goal of big data analytics. It allows personalization of advertisement on web pages one visits based on recent search, viewing and purchase history. Thus we can put it this way, if a business cannot take advantage of the data as it gets generated and analyze it as speed, it is missing opportunities.

2 ISSUES AND CHALLENGES IN BIG DATA

2.1 Big Data Issues and Challenges Related to Characteristics of Big Data

2.1.1 Data volume

When data volume is thought of the very first issue that occurs is storage. As data volume increases so the amount of space required to store data efficiently also increases. Not only that the huge volumes of data needs to be retrieved at a fast speed to extract results from them. The advent of social networking sites have led to production of data of the order of terabytes every day. Such volumes of data are difficult to be handled using existing traditional databases.

- [1]*A.Danush, Second Year Master of Computer Applications in Priyadarshini Engineering College, Vaniyambadi, E-mail: danushashokan97@mail.com*
- [2]*S.SenthilMurugan, Assistant Professor Master of Computer Applications in Priyadarshini Engineering College, Vaniyambadi, E-mail: mailmurugan.78@mail.com*
  (*This information is optional; change it according to your need.)*

2.1.2 Data velocity

Computer systems are creating more and more data, both operational and analytical at increasing speeds and the number of consumers of that data are growing. People want all of the data and they want it as soon as possible leading to what is trending as high-velocity data. High velocity data can mean millions of rows of data per second. Traditional database systems are not capable enough of performing analytics on such volumes of data and that is constantly in motion.

2.1.3 Data variety

Big data comes in many a form like messages, updates and images in social media sites, GPS signals from sensors and cell phones and a whole lot more. Smart phones and other mobiles devices can be bracketed in the same category. As these devices are ubiquitous the traditional databases that store most corporate information until recently are found to be ill suited to these data. Much of these data are unstructured and unwieldy and noisy which requires rigorous technique for decision making based on the data. Better algorithms to analyze them are an issue too.

2.1.4 Data value

Data are stored by different organizations to gain insights from them and use them for analytics for business intelligence. This storing produces a gap between the business leaders and the IT professionals. The business leaders are concerned with adding value to their business and obtaining profits from it. More the data more are the insights. This however doesn't go well with the IT professionals as they have to deal with the technicalities related to storing and processing the huge amounts of data.

2.2 Big Data Management, Human Resource and Man Power Issues and Challenges:

Big data management deals with organization, administration and governance of large volumes of structured and unstructured data. It aims to ensure a high level of data quality and accessibility for business intelligence and big data analytics applications.

2.3 Big Data Technical Issues and Challenges:

2.3.1 Fault Tolerance

Fault-tolerant computing is tedious and requires extremely complex algorithms. A foolproof, cent percent reliable fault tolerant machine or software is simply a far-fetched idea. Divide the entire computation to be done into tasks and assign these tasks to different nodes for computation.

2.3.2 Data Heterogeneity

80% of data in today's world are unstructured data. It encompassed almost every kind of data we produce on a daily basis like social media interaction, document sharing, fax transfers, emails, messages and a lot more. Working with unstructured data is inconvenient and expensive too. Converting these to structured data is unfeasible as well.

2.3.3 Data Quality

As has been mentioned earlier, storage of big data is very expensive and there is always a tiff between business leaders and IT professionals regarding the amount of data the company or the organization is storing

2.3.4 Scalability

The challenge in scalability of big data has led to cloud computing. It is capable of aggregating multiple different workloads with different performance goals into very large clusters. This needs high level of sharing of resources that is quite expensive and brings along with it various challenges like executing various jobs so that the goal of every workload is met successfully

2.4 Big Data Storage and Transport Issues and Challenges

Each time a new storage medium is invented the quantity of data becomes more and more. The capacity of current disks are about 4 terabytes per disk so 1 exabyte requires 25000 disks. Even if a single computer system is capable enough of processing 1 exabyte, to directly work with that many number of disks is well beyond its capacity. Accessing this surge of data overwhelms current communication networks. If 1 gigabyte per second network has an effective sustainable transfer rate of 80% its sustainable bandwidth is about 100 megabytes. This boils down to transferring 1 exabyte for 2800 hours, provided the sustainable transfer rate is maintained.

2.5 Big Data Processing Issues and Challenges: Effective

Processing of big data requires immense parallel processing and new analytics algorithms so as to provide rapid information. Often it may be unknown how to deal with a very large and varied volume of data and whether all of it needs to be analyzed. Challenges also include finding out data points that are really of importance and how to utilize the data to extract maximum benefit from it.

2.6 Big Data Privacy and Security Issues and Challenges

Often in big data analysis, the personal information of people from a database or from social networking sites needs to be combined with external large data sets. Thus facts about anyone which might have been confidential become open to the world. Often it leads to taking insights in people's lives of which they are unaware of. Often it happens that a more educated person having better knowledge and concepts about big data analysis takes advantage of predictive analysis over a person who is less educated than him.

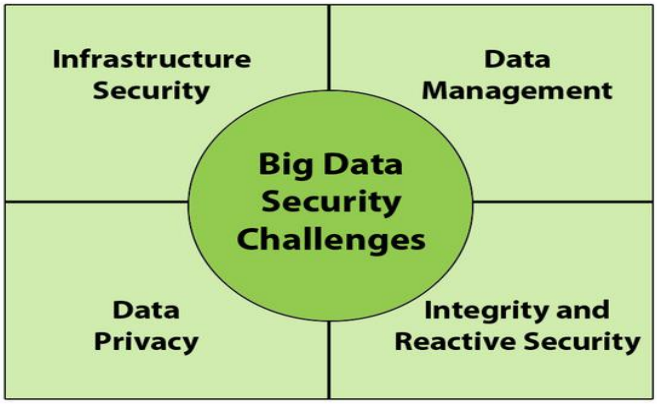
3 REASONS FOR SECURITY AND PRIVACY ISSUES AND CHALLENGES IN BIG DATA

Security and privacy are big concerns as far as big data are concerned and as big data grows by volume every day, every minute, every second so are these concerns on the rise. A prime reason for security and privacy concern in big data is because big data is now widely accessible. Data are shared on a large scale by scientists, doctors, business officials, government agencies and normal people. However the tools and technologies that have been developed till date to handle these huge volumes of data are not efficient enough to provide adequate security and privacy to data. The data security and privacy maintenance regarding big data lacks adequate policies that ensure agreement with current approaches to security and privacy. The present technologies have weak security and privacy maintenance capability so they are continuously being breached both accidentally and intentionally. Thus reassessing and updating current approaches to prevent data leakage has to be done on a continuous basis.

There is lack of spending on IT security to protect big data by the companies. About 10% of a company's IT budget should be spent on security but below 9% is spent on an average thus making it tougher for themselves to protect their data.

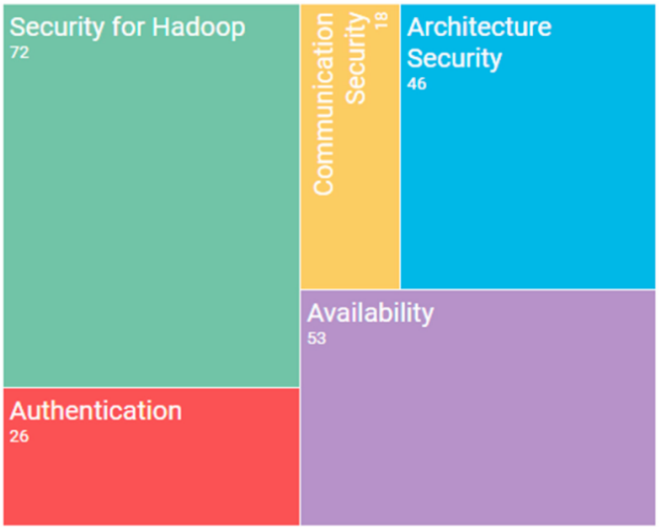
3.1 Main Topic Found

In order to show the main topics found during the research, we have decided to divide them into the four principal aspects employed in the division created by the Big Data Working Group at the Cloud Security Alliance organisation: infrastructure security, data privacy, data management, and integrity and reactive security. This classification has also been used by the NIST Big Data security group for the creation of a security standard for Big Data.



3.2 Infrastructure Security

When discussing infrastructure security, it is necessary to highlight the main technologies and frameworks found as regards securing the architecture of a Big Data system, and particularly those based on the Hadoop technology, since it is that most frequently used.



3.2.1 Security for Hadoop

The graphic shows that the main topic dealt with by these researching infrastructure security is security for Hadoop. As explained in previous sections, Hadoop can be considered as a de factostandard for implementing a Big Data environment in a company. The security problems related to this technology have, therefore, been widely discussed by researchers, who have also proposed various methods with which to improve the security of the Hadoop system.

3.2.2 Availability

Researchers have also dealt with the subject of availability in Big Data systems. One of the main characteristics of Big Data environments, and by extension of a Hadoop implementation, is the availability attained by the use of hundreds of computers in which the data are not only stored, but are also replicated along the cluster.

3.2.3 Architecture Security

Another different approach is that of describing a new Big Data architecture, or modifying the typical one, in order to improve the security of the environment. The new architecture based on the Hadoop file system which, when combined with network coding and multi-node reading, makes it possible to improve the security of the system.

3.2.4 Authentication

The value of the data obtained after executing a Big Data process can, to a great extent, be determined by its authenticity. A few papers deal with this problem by proposing solutions related to authentication.

3.2.5 Communication Security

The security as regards communications between different parts of the Big Data ecosystem is a topic that often is ignored, and only a small number of papers therefore deal with this problem.

3.3 Data privacy

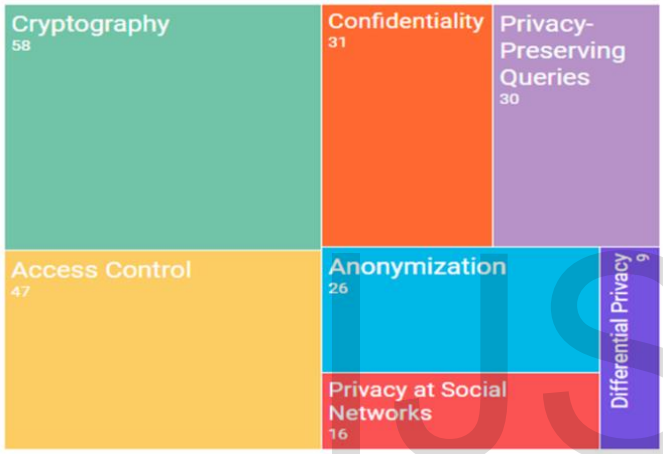
Data privacy is probably the topic about which ordinary people are most concerned, but it should also be one of the greatest concerns for the organizations that use Big Data techniques.

A Big Data system usually contains an enormous amount of personal information that organizations use in order to obtain a benefit from that data.

The below Figure contains a graphic that shows the main ways in which this problem is dealt with, and the quantity of papers found for each septic topic.

3.3.1 Cryptography

The most frequently employed solution as regards securi ng data privacy in a Big Data system is cryptography.Crypt ography has been used to protect data for a considerable a mount of time.



3.3.2 Access control

Access control is on the basic traditional techniques used to achieve the security of a system .its main objective is to restrict non-desirable user's access to the system. In the case of big data, the access control problem is related to the fact that there are only basic forms of access control. In order to solve this problem, some authors propose a framework that supports the integration of access control features.

3.3.3 Confidentiality

Although privacy is traditionally treated as a part of confidentiality, we decided to change the order owing to the tremendous impact that privacy has on the general public's perception of Big Data technology. The authors that approach this problem often propose new techniques such as computing on masked data (CMD), which improves data confidentiality and integrity by allowing direct computations to be made on masked data, or new schemes, such as Trusted Scheme for Hadoop Cluster (TSHC) which creates a new architecture framework for Hadoop in order to improve the confidentiality and security of the data

3.3.4 Privacy-Preserving Queries

The main purpose of a Big Data system is to analyze the data in order to obtain valuable information. However, while we manipulate that data we should not forget its privacy. A few papers pay attention to the problem of how to make queries whilst simultaneously not violating the privacy of the data. One way in which to achieve this protection is by encrypting the data, as discussed previously, but this adds a new problem: how do we analyze the encrypted data? Some authors propose that this problem can be solved by means of a secure keyword search mechanism over that encrypted data.

3.3.5 Anonymisation

One of the most extended ways in which to protect the privacy of data is by anonymising it. This consists of applying some kind of technique or mechanism to the data in order to remove the sensitive information from it or to hide it. Big Data usually implies a large amount of data, and this problem, therefore, increases in Big Data environments. The authors propose a hybrid method that combines the two most frequently used anonymisation schemes: top-down specialization (TDS) and bottom-up generalization (bug).

3.3.6 Privacy in Social Networks

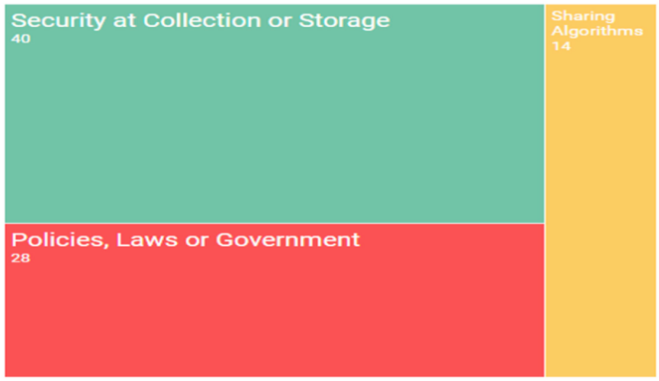
Social networks are all around us. The popularity of Social networks is currently huge, and almost everybody with access to the Internet has at least one account with them. People share a lot of personal information in these networks without actually worrying about what the organization behind them will do with their data. This data, along with the strong analysis capability of Big Data, is a huge threat to our personal privacy. Addressing this problem is not an easy task, and some authors suggest new legislation with which to increase the protection of data privacy. Another paper, meanwhile, proposes a technique that can be used to increase the control that users have over their own data in social networks.

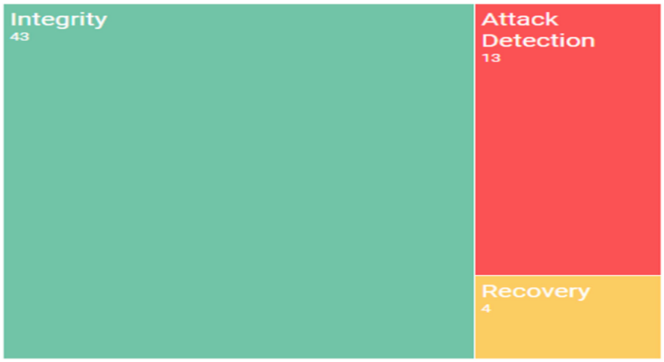
3.3.7 Differential Privacy

The objective of differential privacy is to provide a method with which to maximize the value of analysis of a set of data while minimizing the chances of identifying users' identities. A few papers focus on achieving privacy in Big Data by applying differential privacy techniques.

3.4 Data Management

This section focuses on what to do once the data is contained in the Big Data environment. It not only shows how to secure the data that is stored in the Big Data system, but also how to share that data. We shall also discuss the different policies and legislation that authors suggest in order to use Big Data techniques safely. The below Figure contains a graphic that shows the topics that will be discussed in this section, along with the quantity of papers found for each specific topic.





3.4.1 Security at Collection or Storage

As mentioned previously, Big Data usually implies a huge amount of data, it is therefore, important not only to find a means to protect data when it's stored in a big data environment, but also to know how to initially collect that data.

3.4.2 Policies, laws, or government

Every disruptive technology brings new problem with it, and big data is no exception. The problems related to big are mostly related to the increase in the use of this technique to obtain value from a large amount of data by using it powerful analysis characteristics. This could imply a threat to people's privacy. In order to reduce that risk, many authors propose the creation of new legislation and laws that will allow these new problems to be confronted in an effective manner.

3.4.3 Sharing Algorithms

In order to obtain the maximum possible value from data, it is necessary to share that data among the cluster in which Big Data is running or to share those results for collaboration. However, again, we have the problem of how to guarantee security and privacy when that sharing process is taking place. Some authors approach this problem by increasing the surveillance of the user taking part in data sharing, while others propose securing the transmission itself by creating a new technique based on nested sparse sampling and coprime sampling.

3.5 Integrity and Reactive Security

One of the bases on which Big Data is supported is the capacity to receive streams of data from many different origins and with distinct formats: either structural data or non-structural data. This increases the importance of checking that the data's integrity is good so that it can be used properly. This topic also covers the use case of applying Big Data in order to monitor security so as to detect whether a system is being attacked. Figure 6 contains a bar chart that shows the main subtopics found during the systematic mapping study, and the quantity of papers for each specific topic.

3.5.1 Integrity

Integrity has traditionally been defined as the maintenance of the consistency, accuracy, and trustworthiness of data. It protects data from unauthorized alterations during its lifecycle. Integrity is considered to be one of the three basic dimensions of security (along with confidentiality and availability). Ensuring integrity is critical in a Big Data environment, and authors agree as to the difficulty of achieving the proper integrity of data when attempting to manage this problem. For example, they propose an external integrity verification of the data or a framework to ensure it during a MapReduce process.

3.5.2 Attack Detection

As occurs with all systems, Big Data may be attacked by malicious users. Some authors, therefore, take advantage of the inherit characteristics of Big Data and suggest certain indicators that may be a sign that the Big Data environment is under attack. For instance, in the authors develop a computational system that captures the provenance data related to a MapReduce process. There are also researchers who propose an intrusion detection system especially intended for the specific characteristics of a Big Data environment.

3.5.3 Recovery

The main purpose of this topic is to create particular policies or controls in order to ensure that the system recovers as soon as possible when a disaster occurs. Many organizations currently store their data in Big Data systems, signifying that if a disaster occurs the entire company could be in danger. We have found only a few papers that cover this problem. For example, in there are some recommendations regarding what can be done to recover from a desperate situation.

4 Conclusion

This paper provides an explanation of the research carried out in order to discover the main problem and challenges related to security in big data, and how researchers are dealing with these problems. This objectives was achieved by following the systematic mapping study methodology, which allowed us to find the papers related to our main goal. In conclusion, the big data technology seem to be reaching a mature stage, and that is the reason

why there have been a number of studies created the last year. However, that does not mean that is no longer necessary to study this paradigm, in fact, the studies created from now should focus on more specific problem. Furthermore, big data can be useful as a base for the development of the future technologies that will change the world as we see it, like the internet of things (IOT), or on-demand services, and that is the reason why big data is, after all, the future.

REFERENCES

[1] www.coursera.org, Introduction to Big Data, University of California, sanDiego. https://www.coursera.org/learn/big-data-introduction

[2] http://www.slideshare.net/HarshMishra3/harsh-big-data-seminar-report. Published: 4th January 2014 in Technology, Education Harsh Kishore Mishra.Center for Computer Science and Technology. School of Engineering andTechnology, Central University of Punjab, Bhatinda.

[3] Schmitt, C., Shoffner, M., Owen P., Wang, X., Lamm, B., Mostafa, J., Barker,M., Krishnamurthy, A., Wilhelmsen, K., Ahalt, S., & Fecho, K. (2013): Security and Privacy in the Era of Big Data: The SMW, a Technological Solution to theChallenge of Data Leakage. RENCI, University of North Carolina at Chapel Hill. Text: http://dx.doi.org/10.7921/G0WD3XHT Vol. 1, No. 2 in the RENCIWhite Paper Series, November 2013.Created in collaboration with the NationalConsortium for Data Science. (www.data2discovery.org)

[4] Big Data: Issues and Challenges Moving Forward. 2013 46th HawaiiInternational Conference on System Sciences Stephen Kaisler, i_SWCorporation. Frank Armour, American University. J. Alberto Espinosa, American University. William Money, George Washington University

[5] Big Data: The Management Revolution. Andrew McAfee and Erik BrynjolfssonOctober 2012. Harvard Business Review

[6]http://www.dataversity.net/common-big-data-management-issues-solutions/ TheMost Common Big Data Management Issues (And Their Solutions). By: A.R.Guess. July 15 2014.

[7] TDWI Research. TDWI Best Practices Report. Managing Big Data. FourthQuarter 2013. By Philip Russom.

[8] Expanded Top Ten Big Data Security and Privacy Challenges Big Data WorkingGroup. April 2013. © 2013 Cloud Security Alliance – All Rights Reserved

[9] Challenges and Security Issues in Big Data Analysis. Reena Singh. Kunver ArifAli. IJIRSET. Volume: 5. Issue: 1. January 2016.