Privacy and Security issues in Big Data : A Case Study of Characteristics, Challenges, and Solution

Parth Sabhadiya
Department of Computer Engineering
A D Patel Institute of Technology
Anand, India

Nayankumar Sorathiya
Department of Computer Engineering
A D Patel Institute of Technology
Anand, India

Vaikunth Desai
Department of Computer Engineering
A D Patel Institute of Technology
Anand, India

ABSTRACT
Nowadays, Data is one of the most important recommendations for research in industry and academia. The continuous rapid growth in the volume of data like people's lifestyles, daily habits, and intended to save data from textual to images, videos, etc. have created a new problem and it's not handled by any traditional technologies. Solving this problem through the creation of a new paradigm: Big Data. But big data is a double-edged sword means it brings solution volume of data but also brings certain risks also in terms of privacy and security.it is difficult to handle the security and privacy of the data in this paper we are discussing the Big data 10V's characteristics, challenges of the big data and it's a solution with research in all perspective area.

General Terms
Your general terms must be any term which can be used for general classification of the submitted material such as Pattern Recognition, Security, Algorithms et. al.

Keywords
10V's , Hadoop Security, Cloud Security, Data Management, Data Privacy, Integrity and Reactive Security

1. INTRODUCTION
Over 4.57 billion people (59%) out of 7.77 billion population of the world(at the end of April 2020) are connected with the internet. According to the GSMA real-time intelligence data, there now 9.82 billion mobile connections worldwide, which surpass the current world population with 2.05 billion. In 2010, our global created over 1ZB(Zettabytes) of data and data generation was rose to 33ZB by 2018.IDC predicts that the global datasphere will grow to 175ZB by 2025. It is evident with the introduction of the Internet of Things(IoT), Cloud Computing, and Big Data [1].

Human beings have always intended together data and they also moving from textual data to richer data including photos, videos, songs, and interactive maps associated with metadata such as location information and time/date stamps[1]. Rapidly rise in technology consequently has led to overflow data. As a result of this technological revolution, big data is becoming rapidly an important issue in the future. Big Data is a data set, which is difficult to capture, store, filter, share, analyze, and visualize on it with current technologies. Big data is gaining more attention since the number of devices connected to the called "Internet of Things" (IoT)is still increasing to unpredicted levels, providing large amounts of data that needs to be transformed into valuable information[2]. Big data analysis proficiency can be used to perceive and prevent advanced threats and spiteful intruders. Along with data security, data privacy issues are also considered as assets for business organizations and companies.

When we are looking at security and privacy issues in the Big Data domain we have to see it's applications like astronomy and other e-sciences usually operate on non-personal details and as such generally do not have meaningful privacy issues[3].

Although Big Data provides many things, it brings many challenges and issues of security, privacy, and demands many requirements. In this paper, a wide range of these challenges and opportunities have been described. The rest of the paper is organized as; Section II has 10V's Characteristics for Big Data. Section III has explored the Challenges of Big Data. Section IV has focused on the solution of Big Data security and privacy issue in this paper.

2. CHARACTERISTICS OF BIG DATA
Nowadays, with the proliferation of devices connected with the internet and each other therefore the volume of data processed, stored, and collected is rapidly increasing every day, which also brings new challenges in data security. As a fact, the currently use security mechanisms such as DMZs(Demilitarized Zone) and Firewalls cannot be used in Big Data infrastructure it does not fulfil the requirements and policies of Bring Your Own Device(BYOD)[2]. There are various explanations of Big Data via its characteristics.10V's are typically used to characterize of Big Data as Volume, Velocity, Value, Veracity , Viscosity, Variability, Validity Volatility, Viability and Variety describe in figure 1. Volume is the Data Scale; Velocity is Data Processing; Value is Data Usefulness in decision making; Veracity is the Data Quality and Accuracy; Viscosity is the Data Complexity; Variability is the Data Flow Inconsistency; Validity is the Data properly understand; Volatility is the Data Durability; Viability is the Data Activeness and Variety is the Data Heterogeneity: Structured, Semi-Structured and Unstructured[4].

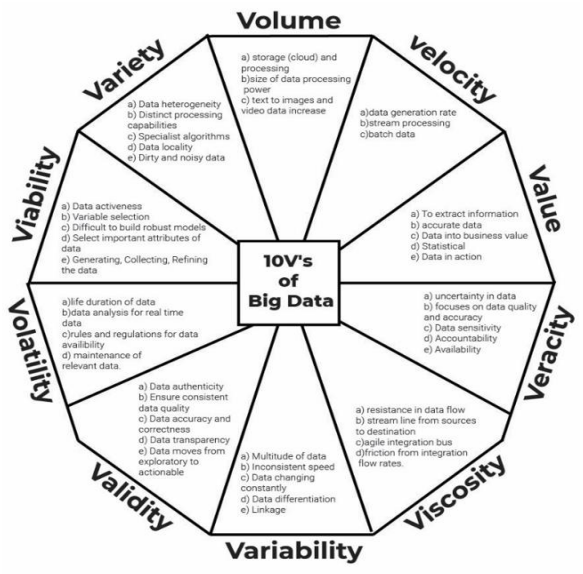

Figure 1. Characteristics of Big Data

3. CHALLENGES OF BIG DATA SECURITY AND PRIVACY

Solutions do not fulfill the requirements when dealing with big data to make sure security and privacy. Big data solutions often rely on traditional firewalls or execution at the application layer to restrict access to the information but firewalls transport security layer ; source of data can be unknown and yes anonymized data can be re-identified[5].

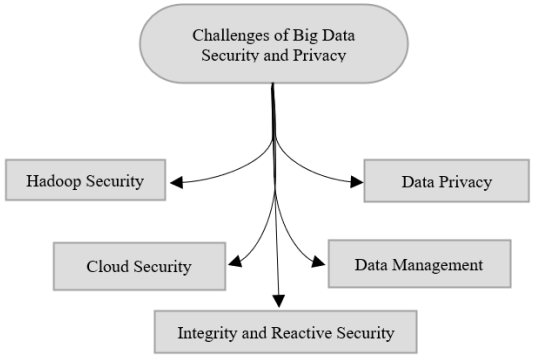

Figure 2. Challenges of Big Data Security and privacy

For these causes, advanced technologies are developed to authentication, protect, and auditing big data in terms of infrastructure, privacy, and management. Considering these approaches, this paper has categorized privacy and security challenges for big data under 5 titles as Hadoop security, Cloud security, Data management, Integrity and Reactive Security, and Data Privacy (Fig. 2).

3.1 Hadoop Security

As per survey, Hadoop is main topic deal with by those researching infrastructure security. Hadoop is the software framework for storing and processing a vast amount of data. It has a combination of components- HDFS(Hadoop Distributed File System) for storage, MapReduce for data processing and YARN(Yet Another Resource Negotiator) for resource management in a cluster[5].

All Hadoop environment components like entry, Flink, and Storm are prone to attacks caused by various vulnerabilities; it can be in software, web interface, or network. The Hadoop framework is a complicated body of application programs, distributed computing software, hardware, and policies for evaluating these resources. Here there are different vulnerabilities year-wise data which are detected and reported in the Hadoop framework (Fig.3)[5].

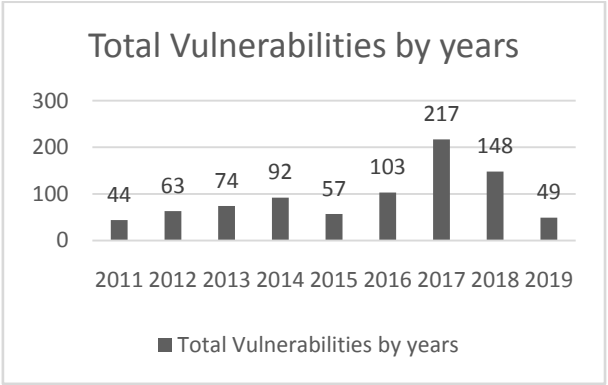

Figure 3. Different Vulnerabilities Reported in Hadoop

Security feature added in Hadoop with the two fundamental goals: Preventing unauthorized access to the files stored in HDFS and Not exceeding high cost while achieving authorization. Hadoop, by default, does not do any authentication, which can have severe effects on the corporate data centers. To overcome this limitation, Kerberos which provides a secure way to authenticate users was introduced in the Hadoop Ecosystem. Kerberos is the network authentication protocol developed at MIT, which uses "tickets" to allow nodes to identify themselves. For data protection, Hadoop HDFS implements transparent encryption. This encryption is end-to-end encryption, which means that only the client will encrypt or decrypt the data. Hadoop HDFS will never store or have access to unencrypted data or unencrypted data encryption keys, satisfying at-rest encryption, and in-transit encryption[6].

There is a proposing for a security replica for G-Hadoop (an extension of the MapReduce framework) that clarify users authentication and some security procedure in order to save the system from conventional effects.

3.2 Cloud Security

As per National Institute of Standards and Technology(NIST), Cloud computing is a prototype for sanction everywhere, suitable, whenever required network access to a allocate pool of configurable computing assets (e.g., networks, servers, storage, applications, and services) that can be quickly supplied and let go with small management attempt or service supplier association. This cloud model is composed of five essential characteristics Fig.4 [12].

Cloud computing is related to processing or computing data at the cloud. Which is shared computing resources rather than having local servers or own personal devices Cloud has great resource capability like software, hardware, application, and that offering in a single system view. Also cloud has a powerful architecture to perform a large scale and complex computing. Here, our goal is to create an understanding of cloud computing as a solution for tackling big data, like high dimensional data sets, large size, and multi-media. After Big Data came into a world, there are many services that emerge like AaaS(Analysis as a Service), DBaaS(Big data as a Service), and DaaS(Database as a Service)[7].

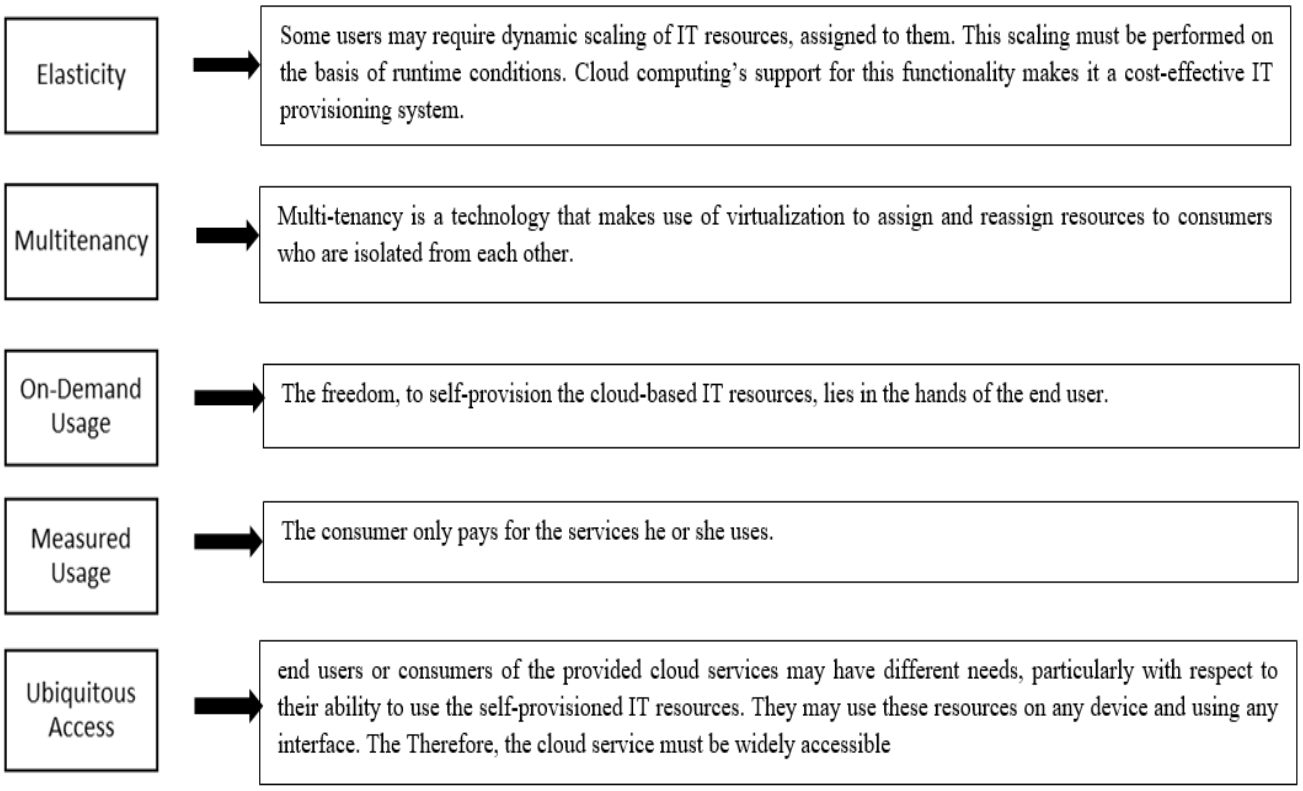

Figure 4. Part of cloud computing model

3.3 Integrity and Reactive Security

There are many bases in big data analysis, One of the main bases on which big data is supported is strength to receive streams of data from serval different origins and with distinct formats: both Structural data or non-structural data. The outcome of this is to increases the importance of checking that the data's integrity is good so that it can be used properly[7]. This issue also covers the use case of applying big data in order to monitor security so as to determine whether a system is being attacked.Fig.5 includes a design that contains main subtopics found during organized mapping study and the number of papers for each specific topic.

Figure 5. Main Topics Regarding Integrity And Reactive Security

1) Integrity

Data Integrity is the all-embracing accuracy, consistency and trustworthiness of data. It protects anyone's personal data or information from unauthorized organizations during its lifecycle to regulatory compliance - such as GDPR compliance[8]. This all maintained by a collection of procedure, rules and standards generated during the graphic phase. When the integrity of data is safe , the information stored in database will remain complete and correct no matter how rapidly it's retrieved or how long it is stored. There are six types(Fig.6) of data integrity: 1) Physical Integrity, 2) Logical Integrity, 3) Entity Integrity, 4) Referential Integrity, 5) Domain Integrity, 6) User-Defined Integrity.

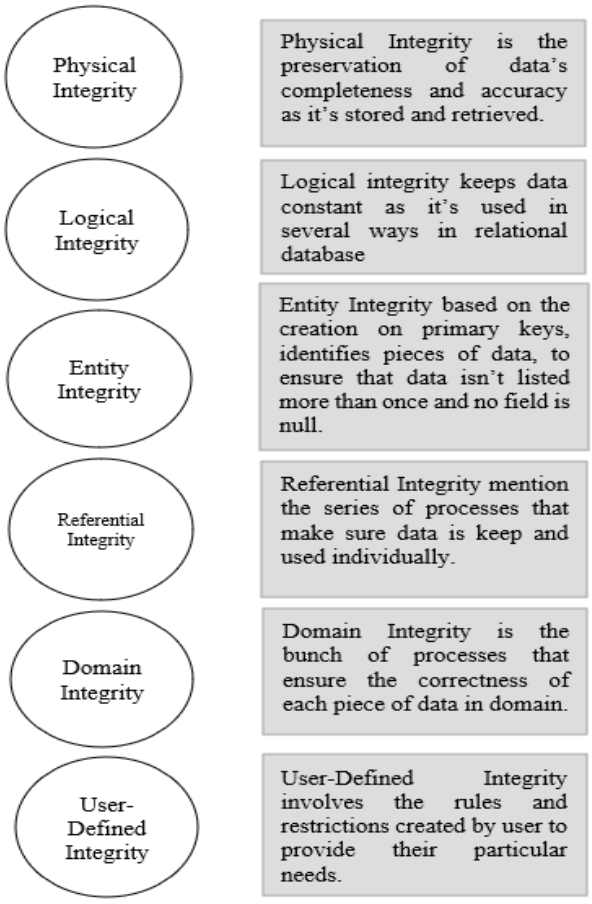

Figure 6. Six types of Data Integrity

2) Attack Detection
As happened with all systems, Big data may be an ambush by malicious users. Big data analysis strategy can take out information from a variety of sources to detect future attacks. To prevent the attacks, Big data analysis framework applies Map Reduced intrusion detection system based on the clustering algorithm.

3) Recovery
The main purpose of this topic is to develop particular policies or controls in order to ensure that the system retrieves as soon as possible when a disaster occurs. Many companies nowadays store their data in big data systems, denote that if a disaster occurs the whole company could be in danger.

3.4 Data Management

Under the Data management section, the focus is on what to do when the data is contained in the Big Data environment. It not only describes how to secure the data that is stored in the Big Data system, but also shows how to share that data. Here we will discuss three main topics related to Data management. Figure 7 contains a graphic that shows the topics that will be discussed in this section. Monitoring and ensuring the availability of all big data resources through a centralized interface/dashboard[12].

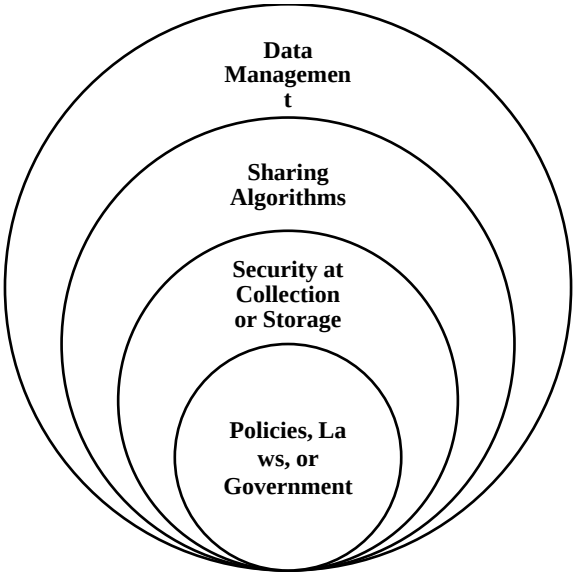

Figure 7. Main topics regarding infrastructure security.

1) Policies, Laws, or Government
Every disruptive technology in market brings new problems with it, and Big Data is no exception. In May 25,2018 the European Union's General Data Protection Regulation(GDPR) comes into effect by replacing the 1995 Data Protective Directive (DPD)[8]. The data regulation by the GDPR is only that related to individuals; it does not apply organizations where data is maintained. Suspicious problems related to Big Data are related to the increase in the use of this technique to obtain value from a large amount of data by using its powerful analysis characteristics. This can be a threat to people's privacy. In order to maintain 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. For example, first best propose from authors is a legal framework in order to protect students' privacy. The purpose of this content was also to find some frameworks or initiatives that attempt to establish a robust government of the security of the data in a Big Data system.

2). Security at Collection or Storage
Big Data usually suggest a huge amount of data. Collection of data plays major role in Data management. , important not only to find a means to protect data when it is stored in a Big Data environment, but also to know how to collect data in initial stage. In order to solve these problems, some authors suggest a mechanism with which to protect data owners' privacy by creating the acceptable level of privacy. Another methodology found in , suggests that security storage can be protected by dividing the data stored in the Big Data system into sequenced parts and storing them in different cloud storage service providers.

3). Sharing Algorithms
In order to obtain the maximum possible output from data, it is best way to share that data among the cluster in which Big Data is running or to share those results for collaboration. Since, we have the problem related to guarantee security and privacy when that sharing process is taking place. To approach this problem we need to increase 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[6].

3.5 Data Privacy

Data privacy is the topic where common people are most concerned. Big data system usually stores vast amount of personal data or information that organizations use in order to obtain a benefit from the information. Therefore, we should ask ourselves where the limit regarding the use of that information is. So, use data securely and protect privacy, we need to understand privacy risks associated with big data[9].

While big data provides many benefits to organizations of all shapes and sizes, it also comes with several significant privacy risks including: (1). Data Breaches, (2). Data Brokerage, (3). Data Discrimination.

1) Data Breaches:
Data breaches happened when data is accessed without authorization. Data breaches are outcome of weak password, Too many permissions, Physical attacks, out-of-date software, Malware attacks and many more. Therefore keeping software up-to-date, Changing password frequently, Multi-factor authentication, and educating employees on best security practices can help prevent data breaches.

2) Data Brokerage:
The sale of unsafe and incorrect data is contemplated as data brokerage. Some organizations gather and sell customer profile and personal data, which contains false information that leads to faulty algorithms. Before buying data companies should do their research and make sure they are receiving data from a trustworthy provider that offers correct data.

3) Data Discrimination:
After all data can consist of customer private and demographic information, organizations may develop algorithms that punish individuals based on age, gender, or ethnicity. Organizations should always have a accurate representation of customers, account for biases, and put morality above analytics.

Several techniques and mechanisms with which to protect the privacy of the data, and also allow organizations to still make a profit from it have therefore been developed, and attempt to solve this problem in many different ways. Figure(8) contains a design that shows the main ways in which this problem is dealt with[7].

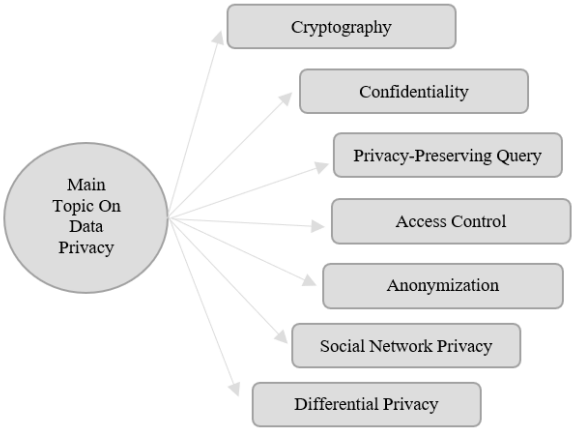

Figure 8. Main Topic On Data Privacy

1) Cryptography
The most rapidly employed solution as regards securing data privacy in a big data system is cryptography. Cryptography is the study and practice of techniques for safe communication in the existence of third parties called adversaries. Safe communication refers to framework where the message shared between two parties can not be accessed by an adversary. In Cryptography, an adversary is a venomous entity, which goal to precious information thereby undermining the principles of data security[7].

2) Confidentiality
Privacy is a traditionally used as a part of confidentiality. Confidentiality refers to some rules and guidelines usually executed under confidentiality agreements which confirms that the information is restricted to some people.

3) Privacy-Preserving Query
The main purpose of big data system is to examine the data in order to obtain important information. However, while handle that data we should not forget its privacy[6]. If the data is public then it is a threat to separate privacy as the data is held by information holder.

4) Access Control
Access control main objective is to restrict non-desirable users access to the system. HDFS is used on individual clusters behind firewalls and used strong authentication and access control to protect the diplomatic private and public data.

5) Anonymization
Data Anonymization is a type of data sanitization whose aim is privacy protection. One of the main extended ways in which to protect the privacy of data is by anonymizing it. Anonymization techniques allow companies to in such a way that the privacy of separates within the data set remains safe at least in some way.

6) Social Network Privacy
Big data is of great importance specially in wireless communication field like social networking. Users uploaded contents like video, blogging, social chats, forms and so on, Big data can be used as attacking on people using these contents.

7) Differential Privacy
Main purpose of differential privacy is to provide method with which to increase the value of analysis of a set of data while decreasing the chances of identifying users specification.

4. SOLUTION OF DATA SECURITY AND PRIVACY

As we discussed earlier the challenges of big data security and privacy here is some solution of the data security and privacy it contains all challenges brief solution work and some future assumptions.

Hadoop security solution around in the four main security pillars of it: Authentication, Authorization, Auditing, and Data protection. This four solution is to make strong Hadoop security also it has two technologies which are making strong it: Apache sentry and Knox. In cloud security, Big data solutions have two fundamental requirements. Because of the size of the data is 'big'. To store this huge big data, it requires a large and scalable storage space. Moreover, the standard analytics algorithms are computing-intensive. Infrastructural solution is the best solution that can support this level of computation. The cloud meets both these requirements well[12].

In current market, there are low-cost storage solution available with the cloud. In spite of this, the user pays for the services he or she uses, which makes the solution all the more cost effective. Secondly, cloud solutions gives hardware free environment to the user, which allows effective and efficient processing of large datasets. In data management, Make a Plan to develop and write a data management plan(DMP). This will help in terms of charts estimated data usage, ownership, accessibility guidelines, archiving approaches etc. Data management plan consists preferred file formats, naming conventions, Access parameter for various stakeholders, backup and archiving processes[4].

Initially determining if your storage needs best suit a data warehouse or a data lake. After doing this make a frame for consistent and enforced, agreement for naming files, folders, directories, users, and more. This kind of foundational piece of data management, as these parameters will determine how to store all future data, and inconsistencies will result in errors and incomplete intelligence.

Data storage changes frequently as fast as the technology demanding it, so any approach should be flexible and have a reasonable archiving approach to keep costs manageable. In Data privacy solution we can Prevent unauthorized access, declaration, and modification of data stored across your undertaking-on premises or in the cloud. Back up your information or data with confidence using workable deployment options and rapid, powdery recovery- access your hybrid cloud. A firewall is one of the best defense for a network because it secluded one network from another. Firewalls eliminate undesirable traffic from entering the network. In addition, you can only open some ports, which give hackers less chances to download your information or data. Antivirus solution help to find and remove trojans, rootkits, and viruses that can break, update or damage your sensitive data[7].

Data encryption is very main solution when you have top secret files that you don't want to be read even if they are stolen. Although personal data can be keep safe by cryptographic algorithms. To prevent your sensitive data properly, you also need to audit changes in your system and attempts to access critical data. In Integrity and Reactive Security Solution Promoting a culture of integrity decreases data integrity risk. Quality control measures contain particular people and processes to verify employees are working with data in line with data integrity[11].

5. CONCLUSION

Big data is a trending topic because no one application can be imagined without it producing new forms of data, operating on data-driven algorithms, and consuming the specified amount of data. In this paper we discussed the characteristic of the big data contain all V's also we find challenges that affect big data security and privacy and provide its solution. we cover all solution possibility that does not mean that it is no longer study on this paradigm, in fact from now studies should focus on a more specific solution for preventing the tradition problem in the future and maybe Big data can be useful for development of the new technologies.

Every year there are many research paper writing on big data security and privacy, in future also researchers more paper write on it because without big data security and privacy it produces difficult in industries and academia.

6. REFERENCES

[1] Nawsher Khan, Mohammed Alsaqer, Habib Shah , Gran Badsha , Aftab Ahmad Abbasi and Soulmaz Salehian, "The 10 Vs, Issues and Challenges of Big Data" ,ICBDE '18: Proceedings of the 2018 International Conference on Big Data and Education,pp.52-56,March 2018. Available:https://dl.acm.org/doi/abs/10.1145/3206157.3206166 [Accessed March 2018]

[2] Jose Mora and Carlos Serrao ,"Security and Privacy Issues in Big Data",ARXIV Org. Available: https://arxiv.org/ftp/arxiv/papers/1601/1601.06206.pdf

[3] Matthew Smith,Christian Szongott,Benjamin Henne and Gabriele von Voigt," Big Data Privacy Issues in Public Social Media",6th IEEE International Conference on Digital Ecosystems and Technologies(DEST),July 2012.Available:https://ieeexplore.ieee.org/abstract/document/6227909[Accessed 02 July 2012]

[4] Duygu Sinanc Terzi, Ramazan Terzi and Seref Sagiroglu," A Survey on Security and Privacy Issues in Big Data ", The 10th International Conference for Internet Technology and Secured Transactions (ICITST-2015),December 2015. Available:https://ieeexplore.ieee.org/abstract/document/7412089 [Accessed 25 February 2016] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[5] Gurjit Singh Bhathal and Amardeep Singh," Big Data: Hadoop framework vulnerabilities, security issues and attacks",ELSEVIER,Vol. 1-2,April 2019 Available:https://www.sciencedirect.com/science/article/pii/S2590005619300025

[6] Julio Moreno,Manuel A. Serrano and Eduardo Fernandez-Medina , "Main Issues in Big Data Security",Future Internet,Vol. 8 no. 3,,August 2016.Available:https://www.mdpi.com/1999-5903/8/3/44 [Accessed 29 August,2016]

[7] Ranjan Kumar Behera , Kshira Sagar Sahoo, Sambit Mahapatra , Santanu Kumar Rath and Bibhudatta Sahoo," Security Issues in Distributed Computation for Big Data Analytics ",Research Gate,2018.Available :https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Security+Issues+in+Distributed+Computation+for+Big+Data+Analytics&btnG=

[8] Nils Gruschka, Vasileios Mavroeidis, Kamer Vishi and Meiko Jensen," Privacy Issues and Data Protection in Big Data: A Case Study Analysis under GDPR ", 2018 IEEE International Conference on Big Data (Big Data) , December 2018. Available : https://ieeexplore.ieee.org/abstract/document/8622621[Accessed 24 January 2019]

[9] Dongpo Zhang, " Big Data Security and Privacy Protection ", 8th International Conference on Management and Computer Science (ICMCS 2018),Vol. 77, October 2018.Available:https://www.atlantis-press.com/proceedings/icmcs-18/25904185[Accessed October 2018]

[10] Renu Bhandari , Vaibhav Hans and Neelu Jyothi Ahuja, "Big Data Security – Challenges and Recommendations ", International Journal of Computer Sciences and Engineering ,Vol. 4,pp. 93-98, Jan 2016.Available:https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Big+Data+Security+%E2%80%93+Challenges+and+Recommendations&btnG= [Accessed Jan 2016]

[11] Sung-Hwan Kim , Nam-Uk Kim and Tai-Myoung Chung ," Attribute Relationship Evaluation Methodology for Big Data Security ",2013 International Conference on IT Convergence and Security (ICITCS) ,December 2013. Available:https://ieeexplore.ieee.org/abstract/document/6717808 [Accessed 23 January 2014]

[12] Priya P. Sharma and Chandrakant P. Navdeti ," Securing Big Data Hadoop: A Review of Security Issues, Threats and Solution ", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2),2014.Available:https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Securing+Big+Data+Hadoop%3A+A+Review+of+Security+Issues%2C+Threats+and+Solution+&btnG=