THE QUEST FOR PRIVACY AND SECURITY IN VARIOUS BIG DATA APPLICATIONS: A STUDY

Alice Joseph[1], Prof. Mathew Cherian[2]
[1]Department Of Computer Science And Engineering, [2]Department Of Mechanical Engineering,
[1, 2] Cochin University College of Engineering Kuttanad, Pulincunnoo P.O, Alappuzha, Kerala, India

Abstract
This paper provides a literature review on the need of security and privacy issues of various big data applications. The first section gives a brief description of big data. The second section reviews the various big data applications and, hence, explains the importance of privacy and security of Big Data in the third section.

Index Terms: Big Data, Application, Security, Privacy

1. INTRODUCTION

As per the researchers of Big data, Big data is described as the large volume of data, now popularly characterized by five *V*'s (initially it was three, but two have added to emphasize the need for data authenticity and business value) which are, volume, velocity, variety, veracity and value, as shown in Fig. 1:

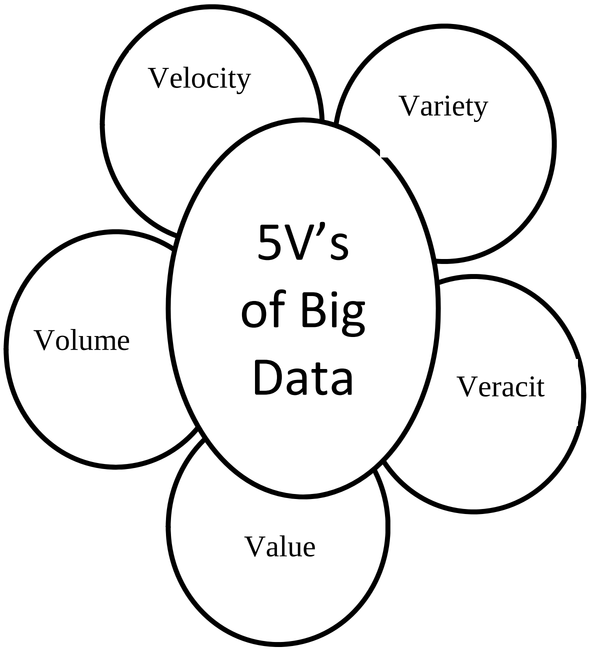

Fig. 1 Five V's of Big data

Volume refers to the amount of data generated and measured in terabytes (240) or petabytes (250), and is rapidly heading toward exabytes (260). The rate at which new data are generated is often characterized as velocity, and data production occurs at very high rates. Data are heterogeneous and can be highly structured, semi-structured, or totally unstructured, and is denoted by variety. Veracity is due to the diversity among data sources, data evolution and raises concerns about security, privacy, trust, accountability, and creating a need to verify secure data provenance; and value measures the usefulness of data in making decisions [1, 2].

Traditional data processing applications are not sufficient to process big data with so large or complex data sets. Recent years the amount of data generated by internet, social networking sites, sensor networks, healthcare applications, vast research works and organizations, is drastically increasing day by day. Typical problems encountered when dealing with Big Data include capture, storage, dissemination, search, analytics and visualization.

A major challenge for senior executives, IT researchers and practitioners due to the exponential growth of data are: (1) design appropriate systems to handle the data securely to provide privacy and (2) analyze it to extract the intended data only to provide relevant meaning for decision making to the targeted people.

1. BIG DATA APPLICATIONS

Some of the big data applications are batch processing types and requires complex calculations and multiple iterations over entire data. Some of them are real time in nature which needs low latency requirements and process incoming stream of data. In this section, examples of such different big data analytics applications are discussed.

2.1 Social Media

Social media, internet-based media, helps the individuals to interact with one another, exchanging personal details such as biographical data, professional information, personal photos, texts, messages and instant thoughts, question papers, assignments, and workshops in Education domain, online survey, marketing, and targeting customers in Business domain, and jokes, music, and videos in Entertainment domain.

Most of the social media started to interact with friends and relatives and for entertainment, but now it is used in businesses to reach out to customers, for marketing their products. So new distributed data analysis frame works have to be introduced to access the subsets of the "big data" datasets. Moreover, the amount of photos uploaded to the social media is growing rapidly and a majority of these photos have no privacy relevance.

Data reside in more dispersed data stores bridging users of a larger network. New comprehensive cross-network, cross cloud data models must be designed to optimize performance based on the distribution of information and users. Then conventional security and access control systems, like active directory, are based on the tree-structured organization of users. A need will exist for highly self-configurable security policies to guard users' security and privacy while also preserving privacy fixed within the data.

Most of the data both originates and resides in the Internet, one open challenge is determining how Internet computing technology should evolve to let us access, assemble, analyse, and act on big data. Big data analytics can accrue the insight of crowds, disclose patterns, and produce best practices as shown in Fig.2.

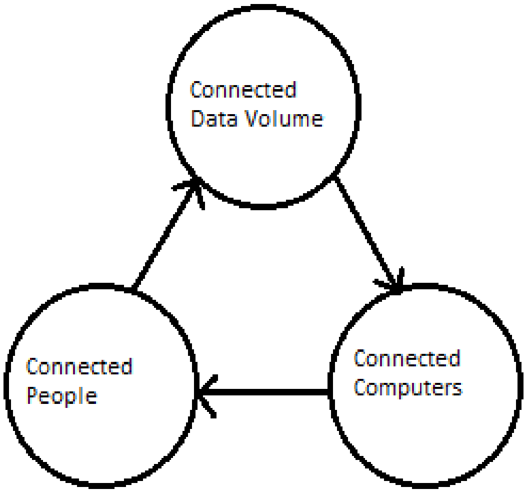

Fig. 2 People in the network produce a data stream that's analyzed by networked computers, and to produce the intelligence that thrives back to networked people.

The amazing growth and diversity in connected data continues to profoundly affect how people make sense of this data. Connected people in the network produce an incessant data stream that is deposited into a repository of connected data. Big data analytics should be done on these data by clouds or connected computers; and these computers generate intelligence that subsequently proliferates back to connected people [3]. The traditional Big Data applications work on non-personal information and generally do not have significant privacy issues. The privacy critical Big Data applications lie in the domains of the social web, consumer and business analytics and governmental surveillance.

The amount of user-generated data uploaded to the web is expanding rapidly and it is beyond the capabilities to see which media impacts the user privacy. When privacy is concerned, there have been a lot of questions like how can users control the data, who can access to this and what they publish themselves. But more focus will be given to the area of what the controlling companies do with this information. This issue is addressed by calls for regulatory intervention. This can be seen in a social context, i.e. what happens if friends or associates see this data and also what happens when other companies with Big Data analytics gather this information.

The privacy of users' data in social networks and photo-sharing sites mainly focus on access control. Another promising trend is the competence of many present devices to implant geo-data and other metadata into the created content. [4]

It is hard to develop an inbuilt technique to protect sensitive information because we live in the era of big data characterized by unique opportunities [5] to sense, store and analyze social data describing human activities in great extent.

2.2 Modern Industry

Sensors embedded in modern machines produce 1000 Exabyte of data annually and is expected to increase many fold in the next ten years [6]. The bad decision which is taken by the organizations disrupts the customers and everyone in between. Big data helps the management to take best decisions by the number of correlations and statistics.

Developing a way for people to correctly value their data, privacy and information security [16] would be a major additional step forward in developing financially viable, private and secure alternatives. This leads to an information age where people can maintain their privacy and retain ownership and control over their digital assets they produce and maintain.

The potential benefits and challenges of big data will naturally differ from sector to sector. Various areas in public and private sectors such as product and market development, operational efficiency, market demand predictions, decision making, and customer experience and loyalty are benefited by the use of big data.

It is seen from Fig. 3, the most important outcomes expected from the use of big data are the customer-centric ones. Recently modern industries use information gathered in various ways and forms for customer analytics; to understand customer needs and predict future behaviors and thus provide better service to them. Customer behavior can be analyzed with the help of data retrieved from sensors embedded in smart products.

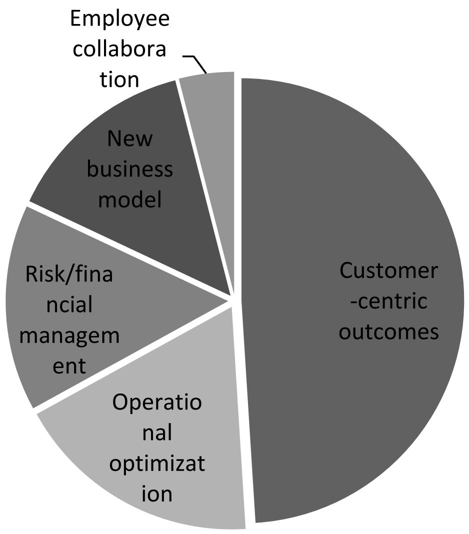

Fig 3: The use of big data by the respondents

Preventive measures can be ensured prior to the failure of the product. In this way, big data can be used to improve the development of the next generation of products and services.

Vulnerability of big data is far higher because of its size, distribution and broad range of access. In addition, many sophisticated software components do not take security seriously enough, including parts of companies' big data infrastructure. This opens a further avenue of potential attack.

Hadoop software allows programmers to process a large amount of data in a distributed computing infrastructure. Many big companies have adopted Hadoop as their corporate data platform, even though its access control mechanism wasn't designed for large-scale adoption of big data [6].

2.2.1 Industrial Big Data in internet

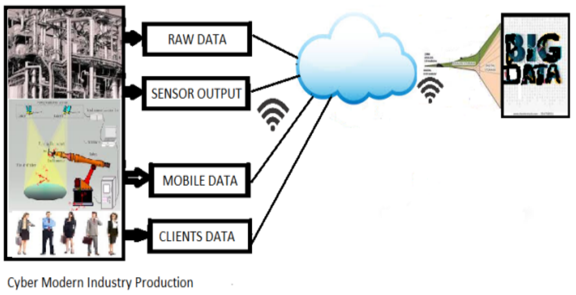

Fig 4. Industrial Big Data connected to the internet

In smart factories, the machines and resources of an enterprise is connected to the internet directly or through external adapters. The machine tools are enriched with knowledge provided by the big data analytics. Business management tools will be connected to the machines, capturing heterogeneous data from different sources. The smart enterprise will collect, transmit and analyse huge amounts of data (big data) to produce intelligent products (smart products) that know how they have been produced. The operators can communicate with these network using mobile phones. The data produced from machine tools and the human operators are in huge size. The analysis of these data produces significant result which is used for management decision making.

Therefore, specific focus should be given in transforming the basis of the production systems into cyber-physical production systems as shown in Fig.4. A main challenge towards the transformation to cyber-physical production systems is the design and development of standard and secure communication protocols capable of interfacing existing systems and collecting and exchanging manufacturing data.

Given the growth of business competition worldwide, manufacturers these days should expect to be the target of theft of trade secrets and intellectual property via corporate espionage. Companies in manufacturing are most likely to face security threats such as cyber espionage, denial of service and web applications attacks. Companies will need to address challenges such as ensuring data privacy and security, to safeguard customer information as well as meet regulatory compliance requirements [7].

2.3 Finance

The financial analyst wants to correlate company’s proprietary data with data on the network storage to know about some kinds of investment or business opportunities. He also wants to correlate public and private data to study about market trends and new opportunities. These types of network storage access leads to security vulnerabilities such as man-in-the-middle attacks. The sensitive information should be intercepted or disrupted due to its high value and huge traffic. The usual procedure to protect data using firewall or to guard it by providing information security such as encryption or hashing techniques cannot be used for big data due to its size and distributive organization.

The need for new cyber security tools or frameworks to be developed to prevent the emerging security vulnerabilities in financial networks. These tools or framework should support big data analytics in real-time networks supporting rapid trading, credit card verification, and other financial services related to consumer banking, investment banking, and their supporting data infrastructures. The essential data sharing between companies can be encouraged by protecting proprietary information.

Attackers are searching for new methods to accomplish financial gain. The money is stolen not only by selling the stolen data but also through sabotage and fabrication of data records or transactions. Financial/banking industry has to make sure it is on top of such threat path.

IBM's X-Force Research Team released a report on Security Trends in the Financial Services Sector found that this industry is attacked 65 percent more often than any other, resulting in more than 200 million records being breached in 2016, a 937 percent increase year over year. [11].

2.4 Transportation

Real-time big data analytics is very much needed in Transportation system which provides various kinds of information to travellers within a very short time. Real time applications guarantee "deadlines" such that it requires all the needed resources available while processing. Real-time big data analytics in transportation helps vehicle tracking, route selection for the destination, estimate time to reach to the destination, change route due to some incidents, find the fastest route for emergency delivery of items and dynamic route identification for emergency vehicles for the quick arrival to the destination. Sensor technology is used for Monitoring Traffic conditions and is connected to the communication technology. Real-time security monitoring is always challenging, given the number of alerts generated by security devices [9]. The volume and velocity of data streams are increasing abundantly leads to the need of big data. Moreover, big data technologies allow fast processing and analytics of different types of data and in turn can be used to deliver real-time inconsistency detection related to scalable security analytics. [10]

2.5 Stock market

A stock market is the mass of buyers and sellers to buy or sale shares or stocks of registered companies. The data generated every day in stock market is big in volume and dynamic. By analyzing these data in real-time helps to detect fraud and illegal activities and improves the performance of the stock market. Real-time data analytics of stock market helps to predict the share prices before actual changes occurs in share prices, so that timely selling or buying of shares can be done for higher profit margin. Earlier decision making ability for buying or selling shares, automated trading of shares and threads detection system, which could increase number of buyer and seller in the market. Financial threads detection in quick time and detection of illegal activities in market which helps to improve market performance. All these merits of stock market can be achieved if real-time data analytics of stock market is possible.

Due to temporal nature of the data required in the stock market, big data architecture is being implemented in such a way that all data must be updated and integrated on timely basis and reports should be generated based on the current information. These information will reach to all partakes of the market to accomplish efficient market guess. Small investors too have equal opportunity to access relevant data due to shared resource and other above discussed benefits of big data technologies. This will increase participation, investments, more stable market, more accurate prediction and finally it leads to stable and mature market.

Big data architecture with latest technologies will provide easy access of information flow and it is relatively affordable to all investors. Cloud based big data engine is used to integrate different heterogeneous and unstructured data. Cloud based technology will provide all analytical services to the clients at reasonable costs and processing of data with quick response [13].

2.6 Defense

In defense sector or in intelligent service make right decision in time for saving human lives. For example, winning in war, power or strength is not only the concern but also making the right decision in time. And for that lot of information needed to analyze like information about different vehicles used in war, opposition strengths or any movement, current situation or historical information related to the war, number of soldiers, and other related information or resources needed to collect and analyzed in time to take right movement or decision in the war. Also in war data generation is very big in volume and very dynamic.

The data should be collected and analyzed dynamically along with other static information to plan for the next step or to make decision on the fly in the war. These sorts of various action or decision are needed to make in defense or military or national security center for the sake of life of human being. Hence, real-time data analytics application or system of such kind of big data is very important to deal with these sort of situation arise in security sector. The aggregation of big data across multiple sources to form a virtual repository has already given rise to questions of security (in terms of confidentiality), legality and ethics [14].

2.7 Natural Disasters

The world faced significant number of natural disasters like earthquakes, floods, tsunami, cyclones, volcanoes etc. which costs huge number of human lives, health, economies and various resources. Early prediction and warnings of such natural disasters can save the lives of thousands of people and resources. But these early warning systems for natural disaster involve real-time processing of huge amount of distributed data that are collected in real-time. For this type of data collection various sensors and GPS or Satellite technology can be used. However the main challenging task is to analyze those data in timely fashion to provide early warnings.

An early natural disaster system which is capable of dealing with huge distributed data and time-critical processing of that information. These warnings from the natural disaster system can save thousands of human lives, economies, and resources and help them to take required action/initiative at a time.

2. PRIVACY AND SECURITY CONCERNS

The more information is collected up by powerful computers as giant sets of data, big data, the more likely that it includes personal or sensitive information. Accordingly, new and improved security tools and mechanisms are needed in order to provide effective protection towards data. Sources of information and types of data diverge greatly, allowing multiple opportunities for access. The advantages of Social media are so many but they are posing threat to Internal Security in various forms like Cyber Terrorism, Fraud, crime, spreading violence, etc. Most of the organizations are familiar with the security mechanisms that are relevant in protecting structured data, but the way of protecting unstructured data, the knowledge may be lacking. The rate of generation, transmission and analysis of data may also pose some security threats [15].

And finally, distributed computing, which is the only way to process the huge quantity of "big data", opens up additional opportunities for data breaches [16]. The smart enterprise will collect, transmit and analyse huge amounts of data (big data) to produce intelligent products (smart products). Big data companies like Amazon heavily rely on distributed computing, which typically involves data centers geographically dispersed across the whole world. Amazon divides its global operations into 12 regions each containing multiple data centers and being potentially subject to both physical attacks and persistent cyber-attacks against the tens of thousands of individual servers housed inside. [17]

It is important to remember that cyber security is to be as strong as the weakest link in its chain. If one of these persons lacked adequate security controls and suffered a data breach, the data they're trying to protect between them would now be vulnerable.

Cybercriminals are focusing privileged users to access financial data. Therefore the strict enforcement of access policies and continuous monitoring of activities to detect anomalous activities are very important. Strong encryption is vital to prevent data theft, modification, disclosure or destruction in the entire life cycle of data, especially during storage and transmission. Employees should be aware of various attacks like spear phishing attack that target to open malicious attachments or click infected links. Kaspersky Lab published a massive cyber-attack targeting the financial/banking industry. With a spear phishing attack they gained access to the targeted institutions, forged the accounting systems, inflating account balances and siphoning off money. It is estimated that around $1 billion was stolen from some 100 financial institutions worldwide [12]. Now the insiders are responsible for more financial sector attacks than outsiders. Security events to be identified by correlation and analytical tools as a malicious attack that will gather, interrupt, repudiate, damage or terminate information system resources or information itself.

Protecting Big Data privacy is also an important area to consider. Individuals and institutions who entrust their data to an organization expect their private information to be protected. In a survey commissioned by Big Brother Watch, over 80% of the respondents across Australia, India, Japan and the Republic of Korea are concerned about online privacy. The survey also reveals that 59% of the respondents believe that their privacy is not sufficiently protected when they use the Internet. However, recent surveys of the American public suggest that there is low confidence in the ability of organizations to guarantee privacy of their data. Notably, only 38% of the respondents stated that have confidence that their credit card companies will ensure privacy and security of their personal credit card activity records. This proportion of confident respondents drops to 16% and 11%, respectively, when asked about their trust of search engine providers and social media sites [8]. As personal information becomes increasingly shared, bought and sold as a commodity, incidents of unauthorized disclosure are likely to grow. Such incidents will further erode public confidence in privacy of their information.

While a major concern, identity theft by cyber-intrusion is only one of the privacy challenges faced by the public in the era of Big Data. This era raises new privacy risks that have been collectively denoted the "4 R challenges of Big Data" in [8]: Reuse, Repurposing, Recombination, and Reanalysis. Reuse, repurposing and reanalysis risks arise when data that is collected with prior consent for one purpose are analyzed for another purpose that may cause harm to the individual. Recombination denotes the risk of personal re-identification achieved by combining collected data with information from other data sources.

By launching an "inference attack" on the data, recombination can succeed, at least partially, even when the data is summarized, anonymized and encrypted." Recent research on new methods of privacy protection, e.g., differential, shows that an individual's information can be protected from recombination with some loss in data fidelity. However, associated losses in data quality may decrease data utility for the organization and its data-sharing partners.

This raises several questions. 1) What are the "privacy-at-risk" data sharing practices in the investment, commercial banking and financial services sectors? 2) How can these risks to be reduced and consumer confidence to be improved? 3) Should the banking industry consider adopting IRB-style broad consent regulations governing the use and sharing of personal data? 4) Should the consumer have more control on how their financial data will be anonymized, used and shared? 5) What new categories of personal data are emerging, e.g., personal accessory data (IoT) and health data (wearable sensors), that could be properly used or shared? [8]

4. CONCLUSION

This paper has presented some applications which required real time and batch processing of big data. The amount of data has been increasing day by day and as a result, data analyzing becomes more challenging. This situation becomes more and more complex when real-time data analytics take place. The real time big data analytics generate fast response. Many of these applications are directly related to human lives. Hence, successful implementations of natural disaster applications can save significant amount of human lives and also can minimize the risks of human lives. Besides, some of these applications can enhance the quality of human lives such as transportation.

It is critical to demand a sensitive level of security in many ways such as terms and conditions, service level agreements, and security trust seals from organizations collecting and using big data. Security measures such as encryption, access control, intrusion detection, backups, auditing and corporate procedures can prevent data from being breached and falling into the wrong hands. As such, Privacy can be promoted by security. At the same time, sensitive security can also hurt privacy, it can provide legitimate excuses to collect more private information such as employees' web surfing history on work computers.

REFERENCES:

[1] Feng Xia, Senior Member, IEEE,Wei Wang, TeshomeMegersaBekele, and Huan Liu, Fellow, IEEE, "Big Scholarly Data: A Survey", IEEE Transactionson Big Data, Vol. 3, No. 1, January-March 2017.

[2] DeepankarBharadwaj, Research Scholar, Dr. ArvindShukla, IFTM University. HOD, Department of Computer Applications, IFTM University, Moradabad(UP), "Text Mining Technique on Big Data Using Genetic Algorithm (A Review)" , International Journal of Computer Engineering and Applications, Volume X, Issue IX, Sep. 16, Issn 2321-3469.

[3] Wei Tan, M. Brian Blake and Iman Saleh, Schahram Dustdar, "Social-Network Sourced Big Data Analytics", IEEE INTERNET COMPUTING, 1089-7801/13/$31.00 © 2013 IEEE

[4] Matthew Smith, Benjamin Henne, , "Big Data Privacy Issues in Public Social Media", 978-1-4673-1703-0/12/$31.00 ©2013 IEEE

[5] Monreale et al., "Privacy-by-design in big data analytics and social mining", EPJ Data Science 2014, 2014:10

[6] Shen Yin, Okyay Kaynak, "Big Data for Modern Industry:Challenges and Trends", Proceedings of the IEEE, Vol. 103, No. 2, February 2015.

[7] D. Mourtzis, E. Vlachou, N. Milas, "Industrial Big Data as a result of IoT adoption in Manufacturing", 5th CIRP Global Web Conference Research and Innovation for Future Production, Procedia CIRP 55 (2016) 290 – 295

[8] Alfred O. Hero, John H. Holland Distinguished University Professor and R. Jamison and Betty Williams Professor of Engineering, University of Michigan; Co-Director, Michigan Institute for Data Science (MIDAS) , Data Privacy and Security, Big Data on Finance, Center on Finance, Law and Policy, University of Michigan Law School, October 27, 2016

[9] Akinul Islam Jony, Department of Computer Science, American International University – Bangladesh, Applications of Real-Time Big Data Analytics, International Journal of Computer Applications (0975 - 8887), Volume 144 - No.5, June 2016.

[10] Khantil Choksi,Niriksha Dalal, Mr.Kshitij Gupte and Dr.Anjali Jivani, "Security And Privacy Challenges In Big Data", International Journal of Latest Trends in Engineering and Technology Vol.(7) Issue(3), pp. 313-318.

[11] Doug Olenick, Online Editor, SC Media US , Cybercrime ,Financial services sector most attacked in 2016: IBM, https:// www. scmagazine.com/ financial-services-sector-most-attacked-in-2016-ibm/article/ 653706/, April 28, 2017

[12] Fran Howarth, Sabotage: The Latest Threat to the Financial/Banking Industry, "https://securityintelligence.com/sabotage-the-latest-threat-to-the-financialbanking-industry/", August 30, 2016

[13] Krishna Kumar Singh, Research Scholar, Priti Dimri, Ph.D, Associate Professor and Head, Krishna Nand Rastogi Research Scholar, Dept. of Computer Science G.B.P.E.C, Pauri Garhwal Uttarakhand, "The Implications of Big Data in Indian Stock Market", International Journal of Computer Applications (0975 – 8887) Volume 99– No.5, August 2014

[14] Neil Couch and Bill Robins, "Big Data for Defence and Security", Royal United Services Institute, Occasional Paper, September 2013

[15] Khairulliza Ahmad Salleh, Lech Janczewski, "Technological, organizational and environmental security and privacy issues of big data: A literature review", Science Direct, Procedia Computer Science 100 (2016) 19– 28.

[16] Jason Parms, https: // www. business. com / articles /privacy – and – security -issues – in – the – age – of – big - data, February 22, 2017

[17] Jungwoo Ryoo, Big data security problems threaten consumers' privacy, March 23, 2016 9.08pm AEDT