Preventive Measures for Security Threat In Big Data Environment

Sheeba Praveen[1], Sumaiya[2]

Dept. CSE, Integral University,

Lucknow, India

***Abstract*— The advent of Big Data along with phenomenal benefits and possibilities it also bring few worrisome challenges. It lays out many questions when it's come to security and privacy of data.There is a heap of problems and issues with big data, out of many the big problem for scientist and engineers is security and privacy of data. Of late it has been a big concern among experts that the bulk of data which is flowing all around must be protected. The challenges faced today is securing a huge amount of data and being able to avoid any sort of breach or data leakage .Governments around the world are pushing themselves and private companies to make data transparent and accessible. All sorts of data is going around and putting sensitive or personal data in public domain could invite risk and could have catastrophic effect. We must think carefully about the role of technology and how we design and engineered next generation systems to appropriately protect and manage privacy, in particular within the context of how policy and laws are developed to protect personal privacy. Big Data touches across so many aspect of life like banking, insurance, medical, health, government hence issue of data privacy is of great importance.**

What we are proposing is an accountability approach to privacy, when security approaches are insufficient. The accountability approach is a supplement to, and not a replacement for upfront prevention.

***Keywords*—Hadoop, Map Reduce Hadoop, YARN, Strategy Policy, Tactical Policy HDFS, Volume, Variety, Velocity, Variability**

I. INTRODUCTION

David Parker, vice president for Big Data at SAP, said "Data privacy is the biggest big-ticket issue, and Big Data sharing can be undertaken for the greater good, or with wrong intentions."

Big data originates from multiple sources including sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals, to name a few. Thanks to cloud computing and the socialization of the Internet, petabytes of unstructured data are created daily online and much of this information has an intrinsic business value if it can be captured and analyzed. For example, mobile communications companies collect data from cell towers; oil and gas companies collect data from refinery sensors and seismic exploration; electric power utilities collect data from power plants and distribution systems. Businesses collect large amounts of user-generated data from prospects and customers including credit card numbers, social security numbers, data on buying habits and patterns of usage. The influx of big data and the need to move this information throughout an organization has created a massive new target for hackers and other cyber criminals. This data, which was previously unusable by organizations is now highly valuable, is subject to privacy laws and compliance regulations, and must be protected [1].

When was your data security plan established? For most enterprises the plans in place were designed in a pre-cloud world of in-house, on-premise databases. There were no thoughts of unstructured data, no thoughts of cloud providers that could place data outside the enterprise firewall and in multiple external data centers that are tied to the enterprise only in a contractual manner. With most data security plans designed in a prior millennium, it's probably time for a new data security plan that takes into account today's BYOD, mobility and cloud risk factors.

II. REVIEW REPORT

Big Data: "*Data is the new science. Big Data holds the answers." said Pat Gelsinger, the Chief Executive Officer of VMware, Inc. and former Chief Operating Officer of EMC Corporation.*

Big Data describes massive volumes of structured and unstructured data that are so large that it is very difficult to process this data using traditional databases and software technologies. The main components of Big Data are:

- **Volume**: Many factors contribute towards increasing Volume streaming data and data collected from sensors etc.,
- **Variety**: Today data comes in all types of formats emails, video, audio, transactions etc.,
- **Velocity**: This means how fast the data is being produced and how fast the data needs to be processed to meet the demand.
- **Variability**: Along with the Velocity, the d peaks.
- **Complexity**: Complexity of the data also needs to be considered when the data is coming from multiple sources. The data must be linked, matched, cleansed and transformed into required formats before actual processing.

Technologies today not only support the collection of large amounts of utilizing such data effectively. Transactions made all over the world with respect to a Bank, Walmart customer

transactions, and Facebook users generating social interaction data. When making an attempt to understand the concept of Big Data, the words and "Hadoop" cannot be avoided.

Hadoop*: Hadoop is an Open Source (Java based), "Scalable", "fault tolerant" platform for large amount of unstructured data storage & processing, distributed across* machines [2]. Distributed file system in Hadoop helps in rapid data transfer rates and allows the system to continue its normal operation even in the case of some node failures. This approach lowers the risk of an entire system failure, even in the case of a significant number of node failures. Hadoop gives scalable, cost effective, and flexible and fault tolerant computing solution. Hadoop has three main sub components that are:

- **Map Reduce Hadoop:** Map Reduce is a programming framework, has the ability to take a query over a dataset, divide it, and run it in parallel over multiple nodes.
- **Hadoop Distributed File System (HDFS):** HDFS [8] is a file system that provides a distributed data storage system to store data in smaller blocks failsafe manner. It also links together file systems on local nodes to make it into one large file system. HDFS improves reliability by replicating data across multiple sources to overcome node failures
- **YARN(Yet Another Resource Negotiator)**: YARN[3] is an Apache Hadoop NextGen Map Reduce also call, Map Reduce 2.0 (MRv2) or YARN. It is used to split up the two major functionalities of the Job Tracker, resource management and job scheduling/monitoring, into separate daemons. The idea is to have a global Resource Manager (*RM*) and per-application Application Master (*AM*)[4].

III. BIG DATA PRIVACY CHALLENGES

The risks associated with Big Data impact every sector, every type of organization and every element of the IT infrastructure so here we are categorizing the type of risk:

A. Big Data Technological Risk

- Being a new technology it will introduce new vulnerability.
- Big Data implementations typically include open source code, with the potential for unrecognized back doors and default credentials.
- User authentication and access to data from multiple locations may not be sufficiently controlled [5].
- Hadoop, like many open source technologies such as UNIX and TCP/IP, was not created with security in mind
- Distributed computing is not secure because Data is processed anywhere resources are available, enabling massively parallel computation. This creates complicated environments that are highly vulnerable to attack, as opposed to the centralized repositories that are monolithic and easier to secure.
- Fragmented data is not secure because Data can become sliced into fragments that are shared across multiple servers. This fragmentation adds more complexity to the security challenge.
- Role-Based Access Control (RBAC) is central to most database security frameworks, but most big data environments only offer access control at the schema level, with no finer granularity to address users by role and related access.
- Node-to-node communication is not secure because Hadoop and the vast majority of distributions don't communicate securely; they use RPC over TCP/IP[6].

B. Big Data Personal Risk

As consumers, our lives and identities become more and more digitized every day. The more time we spend on the Internet, the more data we produce. In 2014, Facebook users shared 2,460,000 pieces of content, YouTube users upload 72 hours of new video, and Yelp users posted 26,380 reviews. These statistics are not per week or even per day**.[7]**
So the biggest challenge for big data from a security point of view is the protection of user's privacy. Big data frequently contains huge amounts of personal identifiable information and therefore privacy of users is a huge concern. The public sector is not immune either. In the UK, the Government has already identified the need to raise the bar in IT security following high profile breaches.

C. Big Data Organizational Risk

Organizations should run a risk assessment over the data they are collecting. They should consider whether they are collecting any customer information that should be kept private and establish adequate policies that protect the data and the right to privacy of their clients. If the data is shared with other organizations then it should be considered how this is done. Deliberately released data that turns out to infringe on privacy can have a huge impact on an organization from a reputational and economic point of view. What organizations need to do is to identify what information is of value for the business. If they capture all the information available they risk wasting time and resources processing data that will add little or no value to the business.[8]

D. Big Data Government Risk

The main problem from a governance point of view is that big data is a relatively new concept and therefore no one has created procedures and policies.

The challenge with big data is that the unstructured nature of the information makes it difficult to categorize, model and map the data when it is captured and stored. The problem is made worse by the fact that the data normally comes from external sources, often making it complicated to confirm its accuracy.

IV. SCALE OF DESTRUCTION FROM BIG DATA

A. *Cases of Data security breaches*

There had been many cases of huge security breaches which rock everyone and made people think. CSO [9] in 2012 published *"The 15 Worst Data Security Breaches of the 21st Century"* ,it list out some of the huge security breaches happened in recent past which left the company red faced and caused great loss. These thefts put a big question on credibility of the company and their competence in keeping data secure.

Company	Date	Impact
Monster.com	August 2007	Confidential information of 1.3 million job seekers stolen and used in a phishing scam. Hackers broke into the U.S. online recruitment site's password-protected resume library using credentials that Monster Worldwide Inc. said were stolen from its clients. Reuters reported that the attack was launched using two servers at a Web-hosting company in Ukraine and a group of personal computers that the hackers controlled after infecting them with a malicious software program.
.Fidelity National Information Services	July 2007	An employee of FIS subsidiary Certegy Check Services stole 3.2 million customer records including credit card, banking and personal information.
Stuxnet	Sometime in 2010, but origins date to 2007	Meant to attack Iran's nuclear power program, but will also serve as a template for real-world intrusion and service disruption of power grids, water supplies or public transportation systems.
Google/other Silicon Valley companies		Stolen intellectual property In an act of industrial espionage, the Chinese government launched a massive and unprecedented attack on Google, Yahoo, and dozens of other Silicon Valley companies. The Chinese hackers exploited a weakness in an old version of Internet Explorer to gain access to Google's internal network. It was first announced that China was trying to gather information on Chinese human rights activists.

B. *Loss through security breaches Organizational*

Recent spate on some high profile companies ,security breaches damage is cataphoric it not only damage the brand and reputation of companies but also the privacy and secrecy consumer and financial data is also at stake. The range of companies is vast, from financial and retail services organization to health care industry everybody is at stake.

According to Information Security Breaches Survey conducted by pwc who has cyber security practices in Europe, for many years reports that in latest survey of 2015 that security breaches has been rises to new high "There has been an increase in the number of both large and small organizations experiencing breaches, reversing the slight decrease found in last year's report. 90% of large organizations reported that they had suffered a security breach, up from 81% in 2014."[10]

Big data is now used in sectors as diverse as energy, medicine, advertising, and telecommunications. IBM and Ponemon Institute bring annual benchmark study on 2015" Cost of Data Breach Study: India" on cost of data breaches incident for companies based in India. Study state the "the cost of data breach in India increased significantly from 3,098 INR (India Rupees) in 2014 to 3,396 INR for one compromised record2 increased from 83.1 million INR to 88.5 million INR in 2015".(2015 Cost of Data Breach Study: India Ponemon Institute© Research Report Benchmark research sponsored by IBM Independently conducted by Ponemon Institute LLC May 2015)

Government -Security and privacy of data is foremost important in government sector, because of sensitive nature of information. Public sector generates lot of information, sector like security, defence, and finance consume lots of data and rely on big data across most of the functional operation. In recent times government is being subjected to huge security breaches it being described as biggest cyber security breaches occur in US as nearly 4 million US Federal employees personal data have been breached [11]. Described as one of the largest thefts of government data ever seen

According to US Government Accountability Office number of security incident at federal agencies has skyrocketed. Another theft which shook the US government is IRS security breach that compromised attack on tax information from more than 100,000 U.S. households The agency say that criminals had stolen Social Security numbers and other data to gain unauthorized access to the taxpayers accounts.[12].The Washington Post reported the hack of the Executive Office of the President's unclassified network, after cyber teams working to mitigate the malicious activity were forced to take some services offline, including [13]

Consumer - Giants like Facebook, Google gather huge amount of data not just about the world but of consumers themselves, thereby reshaping a range of markets based on empowering a narrow set of corporate advertisers and others to prey on consumer. In 2013 IAB internet advertising

revenue reports that Internet advertising surpassed broadcast advertising revenues in the United States for the first time.(*5 Interactive Advertising Bureau (IAB), IAB internet advertising revenue report 2013 full year results, April 2014;*).In March 2012, the Federal Communication Commission (FTC) issued a report, Protecting Consumer Privacy in an Era of Rapid Change, that sought to outline a framework for privacy protection for both businesses to adopt voluntarily and, where necessary, policymakers could mandate as part of general consumer protection. Online activity can be hidden from advertisers, data portability to allow users to switch easily between email and social networking services and take their data with them, and greater transparency and choice by consumers on where and how they share their data with companies. (How Big Data Enables Economic Harm to Consumers, Especially to Low-Income and Other Vulnerable Sectors of the Population [14])

World Privacy forum released a report that data brokers, analytics firms and retailers are secretly scoring consumer and financial institutions, wireless phone service providers, law enforcement agencies and others use these scores to do everything from promoting new products to investigating crimes, while these consumer scores are pervasive, most consumers don't know they exist.[17].

The Pew Research Center has published a new privacy poll on Americans 'Views about Data Collection and Security. According to the Pew survey, 74% of Americans believe control over personal information is "very important," yet only 9% believe they have such control.[14]

V. PROPOSED PREVENTIVE SOLUTION

Big data is a relatively new concept and therefore there is not a list of best practices yet that are widely recognized by the security community. However there are a number of general security recommendations that can be applied to big data:

A. Technological solution for preventing big data in an organization

- If you are storing your big data in the cloud, you must ensure that your provider has adequate protection mechanisms in place. Make sure that the provider carries out periodic security audits and agree penalties in case those adequate security standards are not met.
- Create an adequate access control policy that allows access to authorized users only.
- Both the raw data and the outcome from analytics should be adequately protected. Encryption should be used accordingly to ensure no sensitive data is leaked.
- Protect communications: Data in transit should be adequately protected to ensure its confidentiality and integrity.
- Use real-time security monitoring: Access to the data should be monitored. Threat intelligence should be used to prevent unauthorized access to the data.
- Anonymizing the data is also important to ensure that privacy concerns are addressed. It should be ensured that all sensitive information is removed from the set of records collected.
- Real-time security monitoring is also a key security component for a big data project. It is important that organizations monitor access to ensure that there is no unauthorized access. It is also important that threat intelligence is in place to ensure that more sophisticated attacks are detected and that the organizations can react to threats accordingly.

B. Strategic and tactical policy approaches for preventing big data in an organization

- The main challenge introduced by big data is how to identify sensitive pieces of information that are stored within the unstructured data set. Organizations must make sure that they isolate sensitive information and they should be able to prove that they have adequate processes in place to achieve it. Some vendors are starting to offer compliance toolkits designed to work in a big data environment.
- Anyone using third party cloud providers to store or process data will need to ensure that the providers are complying with regulations.
- Organizations should run a risk assessment over the data they are collecting. They should consider whether they are collecting any customer information that should be kept private and establish adequate policies that protect the data and the right to privacy of their clients.
- If the data is shared with other organizations then it should be considered how this is done. Deliberately released data that turns out to infringe on privacy can have a huge impact on an organization from a reputational and economic point of view.
- Organizations should also carefully consider regional laws around handling customer data, such as the EU Data Directive.

C. Individual efforts to prevention security

- Need to take a strong measure of responsibility and consciously decide with whom they will share their information and for what purposes."That, she said, means consumers need to force themselves to do what almost nobody does: Read privacy policies and terms of service agreements.
- Quit sharing so much on social media. If you only have a few people you want to see photos or videos, then send directly to them instead of posting where many can access them.
- Don't provide information to businesses or other organizations that are not necessary for the purposes for which you're doing business with them. Unless they really need your address and phone number, don't give it to them.

- Use an anonymous browser, like Hotspot Shield or Tor (The Onion Router) when visiting sites that might yield
- information that could cause people to draw inaccurate conclusions about you.[15]

D. Governance frameworks be adapted to handle big data security issues and risk

- What organizations need to do is to identify what information is of value for the business. If they capture all the information available they risk wasting time and resources processing data that will add little or no value to the business.[16]

VI. CONCLUSION

This paper explains various preventive measures for security and privacy challenges in Big Data. Since it is new phenomena and still evolving thus it is more vulnerable to attack. Security breaches are penetrated in different layers so prevention is also need accordingly. If proposed preventive measures are applied it could lessen the harm to a great extent.

REFERENCES

[1] file:///C:/Users/iu/Downloads/zettaset_wp_security_0413.pdf
[2] http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html
[3] http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html
[4] International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014
[5] http://www.computerweekly.com/feature/How-to-tackle-big-data-from-a-security-point-of-view
[6] file:///C:/Users/iu/Downloads/zettaset_wp_security_0413.pdf
[7] http://www.gigya.com/blog/big-datas-big-organizational-problems/
[8] https://www.mwrinfosecurity.com/articles/big-data-security---challenges-solutions/
[9] http://www.csoonline.com/
[10] https://www.pwc.co.uk/assets/pdf/cyber-security-2014-exec-summary.pdf
[11] http://www.foxnews.com/politics/2015/06/05/us-officials-massive-breach-federal-personnel-data/ June 5 2015
[12] Fox New Published May 27, 2015
[13] http://www.nextgov.com/cybersecurity/2014/12/year-breach-10-federal-agency-data-breaches-2014/102066/https://epic.org/privacy/big-data/
[14] https://www.ftc.gov/system/files/documents/public_comments/2014/08/00015-92370.pdf
[15] http://www.csoonline.com/article/2855641/big-data-security/the-5-worst-big-data-privacy-risks-and-how-to-guard-against-them.html
[16] https://www.mwrinfosecurity.com/articles/big-data-security---challenges-solutions/
[17] http://money.cnn.com/2014/04/02/pf/consumer-scores/