Asian Journal of Computer Science and Technology
ISSN: 2249-0701 Vol.7 No.1, 2018, pp. 33-39

An Improved Security Threat Model for Big Data Life Cycle

Kanika[1], Alka[2] and R.A. Khan[3]
[1]Research Scholar, [2]Assistant Professor, [3]Professor,
[1,2&3]Department of Information Technology, Babasaheb Bhimrao Ambedkar University, Uttar Pradesh, India
E-Mail: Sharma.kanika247@gmail.com, alka_csjmu@yahoo.co.in, khanraees@yahoo.com

Abstract **- Big data is a huge amount of data created by individuals related to their medical, internet activity, social networking sites, energy usage communication patterns etc. From these sources, data is being collected and processed by various survey organizations, national statistical agencies, medical centres, and other companies etc. There are many security challenges which occur during data transactions, such as un-authentication, phishing, Vishing, data mining based attacks, etc. From a security point of view the biggest challenge for big data is the protection of user's privacy. Yazan et.al, have presented big data lifecycle threat model. This paper does a critical review of the work. An Improved Security Threat Model for Big Data Life Cycle has been proposed as a main contribution of the paper. A new phase i.e. data creation phase has been added to the life cycle and it is claimed that the phase is very important one with respect to security and privacy. To justify the claim theoretical and statistical evidences have been provided.**

Keywords: **Big Data, Big Data Security, Security Threat Model for Big Data Life cycle.**

I. INTRODUCTION

In today's era, data and information are created and processed at very high speed producing large volume of data added to the database. This large volume of data is called 'Big Data,'. This data comes from several sources including social media, YouTube, online transactions, etc. Big data has five characteristics volume, velocity, value, variety and complexity. It used in a variety of ways in various areas including the health sector, public sector, social networking sites, government sector, etc [1]. According to IBM, 80% of the data generated by various organizations is unstructured, and this is in a variety of formats such as text, video, audio, diagrams, images and combinations of any two or more formats [10]. From last two years, users are facing the various challenges due to growing size of the data. Because of its complexity and size, it cannot be handled through traditional techniques available to handle ordinary database [1].

With the increase in the use and demands of big data, software industries are facing privacy issues. Individual's privacy is still a major problem and providing security has become very crucial [2]. There is various privacy challenges including a privacy breach occurred in 2006. 20 million search queries were released by America Online (AOL). These queries were posed by users over a three month period to facilitate research on information retrieval. On behalf of this information, two reporters from New York Times were able to find the identity of user No. 4417749 based on just search history [17]. CISCO has estimated that at the end of 2016 the annual global data traffic will reach 6.6 zettabytes. So there is a need to develop such approaches that not only support the collection of a large amount of data but also effectively handle or operate vast data requests with minimum time and maximum privacy [3, 4]. While protecting the big data is a big question which needs to be answered whether a particular data is in the category of `public and private. [4].

For providing privacy and security, big data should be examined from diverse angles. A careful thinking should be there for the protection of data itself. To maintain the security of big data Yazan et.al proposed a Big Data Lifecycle threat model. This threat model is based on the work of Xu et.al [7]. According to Yazan et.al, the life span of big data can be divided into four phases. The four phases of big data lifecycle threat model include data collection, data storage, data analytics, and knowledge creation. For each phase, they have described security threats and attacks. The threat model proposed by the author has many drawbacks. The researcher critically reviews the model and a modified threat model has been proposed.

This paper as organized as Section III of this paper briefly describes yazan's security life cycle. In sub-section III, the bright side of the life cycle, and the limitations of under the life cycle. Section IV presents some suggestions to improve the cycle, and a pictorial representation of the improved security process has been depicted and explained. In section V, statistical analysis and results has been explained and finally the paper concluded in section VI.

II. LITERATURE REVIEW

Hakan Ozkose et.al [15] has described the process of big data and explained how big data comes in the picture? They gave the detail literature review of big data and explained, yesterday, today, and future of big data. They explained that storing and processing of data become difficult and classical approaches remain incapable of doing such transactions. Katal Avita, *et.al* [8] explained that big data requires new technologies and architectures to extract valuable information. They have discussed some issues, challenges regarding big data, some tools and techniques such as Hadoop; map reduces, etc. Chanchal Yadav et.al [9]

described architecture for big data. They have presented a review of various algorithms from the year 1994 to a year for 2013 handling large data sets. While pointing out the various security issues, like data integrity, the drawback of the hash function, they have listed out various tools for analyzing the big data. According to IDC [6], there are five levels of security: privacy, compliance, custodial, confidential, and lockdown. If the data is continuously growing with this exponential speed, the expected volume of data would be 40 trillion gigabytes by 2020. Nawsher Khan, *et.al* [13] has described that growth rate of the amount of collected data generates numerous critical issues and challenges such as rapid data growth transfer speed and security issues. They provided study surveys and classified the various attributes of big data.

Sung Hawn Kim [2] have explained two important elements for big data security. First attribute relevance in big data is a key element for fetching information. Second, they defined it is impossible to protect all big data attributes. Colin Tankard [14] have explained some recommends providing better control over big data sets, such as archiving, access control and data release prevention should be brought together. He has described that big data centralized storage is so sensitive. It creates new security challenges. Edith Ramirez [16] has talked about the security of big data. He explained that big data brings big benefits. In this paper, he explained some privacy challenges had been discussed such as unauthorized access, data provenance, etc. he gave some case studies that how an attacker can steal user's private information. He has described four steps of big data life cycle including infrastructure security, data privacy, data management, and reactive security.

III. BIG DATA SECURITY LIFE CYCLE

Big data is very huge, and its size is increasing day by day [5]. According to IDC report, it is assumed that till 2020 data will grow from 130 exabytes to 40,000 exabytes, or 40 trillion gigabytes [6]. With this huge amount of data, the security and privacy breaches are expected to rise exponentially [6]. The researcher must be cared out to protect the data provided by the user. Big data lifecycle threat model in fig 1, presented by yazan alshboul et.al, have following four phases, i.e. data collection, data storage, data analytics, and knowledge creation. The aim of the authors is to secure data in the heterogeneous environment.

A. Big data security life cycle has following phases

1. Data Collection Phase

The first phase of the big data lifecycle threat model is data collection phase. In this phase, data comes from different sources and with different formats. The author has explained that to be secure big data collection phase must be secured and protected. They have emphasized that data should be collected from trusted sources. Some security measures can be used in this data collection phase including limited access control and encrypting some data fields [1].

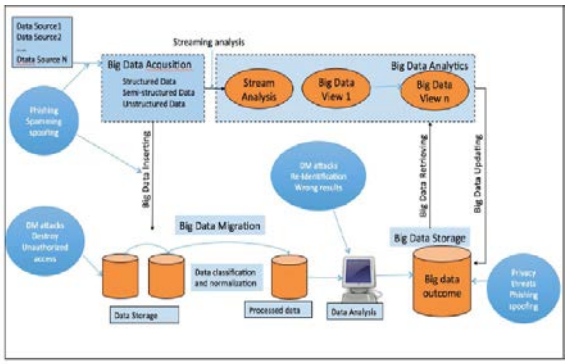

Figure 1: Big Data Lifecycle Threat Model

2. Data storage phase

The second phase of the life cycle is data storage phase. In this phase, the data is stored and prepared to be used in the next phase. As data is sensitive, it is essential to take sufficient precautions during data storage for the safety of data in this phase, the author have described some security measures such as permutation, data partitioning, and data anonymization [1].

3. Data Analytics Phase

After secure collection and storage of data, the next phase comes i.e. Data analytics phase. This phase is used to create knowledge. For this phase, the authors have described data mining methods such as clustering, association rule mining and classification. The authors described that the data miners also used data mining for extracting sensitive data. To secure data, in this phase, data mining process and its output must be secured against data mining based attacks [1].

4. Knowledge Creation Phase

The last phase of this life cycle is knowledge creation phase. Decision makers use this phase. The knowledge generated in this phase is considered as sensitive information. The authors have emphasized that the organizations should be

very careful with data protection. It should not be displayed in public [1].

B. The Bright Side

With the increased used and the creation of big data, several issues are cropping up. These include processing issues, storage issues, transport issues and security issues [3]. Yazan's big data lifecycle threat model has given the idea about security threats and attacks that come in each phase. There are many techniques and approaches available for securing big data, but yazan et.al big data life cycle is new and one of its kinds. This type of study has never been in the light before. The key benefits of threats and security model of big data life cycle, proposed by yazan et.al includes following:

1. The threat and security model addresses the security attacks and threats in every phase of big data life cycle.
2. This type of life cycle is unique and can be adapted easily.
3. The big data lifecycle threat model provides depth insight into threats at each phase of big data life cycle.
4. It also provides countermeasures to mitigate the threats in each phase of big data life cycle.

C. The Dark Side

As almost everything has brighter as well as darker aspects so with this life cycle given by Yazan et.al. Despite having so many reasons for favorable industrial adaptation of the process, there are negative aspects also. Some pertinent ones are listed as follows:
The big data lifecycle threat model is incomplete because it does not provide the source of data clearly. They do not explain the role of data creator. Does it ignore the user who provides or creates data?

1. The researchers have failed to discuss step by step explanation of the life cycle.
2. The life cycle discussed, seems vague and incomplete.
3. In the context of this life cycle, they have not been explained the terms streaming analysis, big data migration, and big data analytics.

IV. IMPROVED SECURITY THREAT MODEL FOR BIG DATA LIFE CYCLE

In the year 2014 Xu et.al [7] have identified four different types of users i.e. data provider, data collector, data miner, and decision maker. Yazan's et.al extended the model from Xu, et.al. They have described the four user's role in big data environment: data provider, data collector, and data miner and decision maker. In the context of big data, they have discussed user's privacy concern and countermeasures that can be adopted to protect sensitive data. Corresponding to the four user's roles in Xu model, Yazan et.al has extended their model and introduced four phases in their model. These four phases of lifecycle including data collection, data storage, and data analytics and knowledge creation phase. The first phase of this threat model is data collection phase. In this phase, data comes from different sources and with different formats. The second phase is data storage. In this phase, the data is stored and prepared for the use in the next phase. The third phase is data analytics phase. This phase is used to create knowledge. The last phase of his threat model is knowledge creation phase. However, the authors have completely ignored the role of data creator and have omitted the most important phase in their model related to data creation phase. Data creator must be the main phase of this threats and security model for big data life cycle. Also, in their threat model, they have not discussed anything about the origin of the data. So it is highly desirable to introduce another phase namely data creator. In fig 2 authors described the new security threat model for big data life cycle which includes the following phases:

1. Data Creation Phase

The most important phase of improved security threat model for big data life cycle is data creation phase. A data creator is a person who provides data to the data collector. Keeping in view the security, it is a very important phase. Data creator is a person who creates data and post/submit to the data collector. He/she is the person who reveals information with his/her responsibility. Once a creator hands over his/her information to others then, there is always a possibility of attacks. Attackers can misuse his/her information provided to the data collector. There are many cases reported, where a hacker have stolen sensitive data of the user.

So a creator should aware of such type of attacks on his/her data. A creator must take into account the following suggestion to avoid an attack on his/her data. If a data creator considers his/her data to be sensitive, i.e. any theft of data may damage his/her privacy or reputation, then while providing the data, creator must keep in mind the following suggestions:

1. The creator must keep in his/her mind that once any information revealed from his side, he loses control over it, irrespective of the sensitivity of data.
2. A creator should provide only relevant data to the data collector.
3. The creator must be sure about the authenticity of the data collector.
4. The creator must not provide his sensitive data until required.

Once the data have been handed over to the collector, privacy of the same cannot be ensured [7]. Now the privacy of the data is depended on the data collection phase as well as subsequent phases. It is necessary for the creator to reveal only relevant information to others. As an end user (creator) is most prone to the security attacks including phishing, authenticity, etc. Security countermeasures must be applied in this phase.

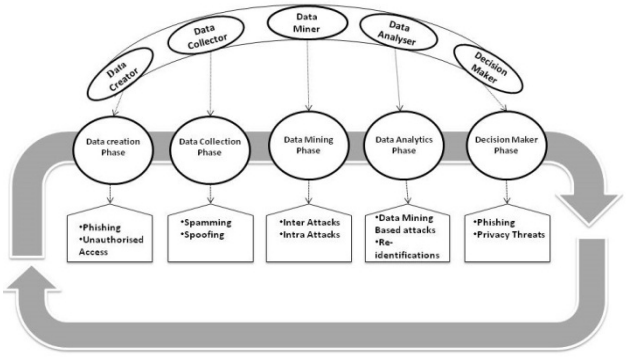

Fig. 2 The new security threat model for big data life cycle

2. Data Collection Phase

The second phase of the proposed security threat model is data collection phase. In this phase data collector plays two different roles: one is data collector as well as a data provider. This phase plays a role of a data collector phase and data creator phase plays a role of a data provider for the data mining phase. Data collection phase is vulnerable to several attacks like phishing, spoofing attacks [1]. To prevent from these attacks, a data collector must provide limited data to data miner. When data collector/provider provides data to data miner, then he/she always concerned about the privacy of the data [7].

3. Data Mining Phase

The data collector provides modified data to the data miner. The primary concern of data miner is to prevent sensitive data from rogue. Data miner needs to modify data, what he/she got from the data collector. In this phase, there are some attacks [7]. Data mining phase is vulnerable to several attacks like novel networks attacks, content-based attacks [11], distributed/denial of service attacks [12], etc.

4. Data Analytics Phase

The fourth phase of this proposed security threat model is data analytics phase. In this phase, the data analyzed the data obtained from data mining phase. For the data analysis, the analyzer examines huge amounts of data to discover hidden patterns, to extract sensitive data. The information obtained from this phase is used for decision-making [7].

5. Decision Maker Phase

The last phase of this proposed security threat model is decision maker phase. This phase utilizes verified and valued information obtained from the previous phase. The exact aim of data mining is to provide the useful information to the decision maker so, that decision maker can select a better way to achieve organization's goals. A decision maker has no responsibility for the protection of data [7].

V. IMPORTANCE OF THE DATA CREATİON PHASE

In big data lifecycle threat model authors have added a new phase 'Data Creation Phase' to the threat model presented by Yazan et.al. In this phase the creator creates information and this information is collected by the data collection phase. A creator may create information including name, Email, Mobile etc. Once the information is reached to the collection phase, the creator loses control over this. It is now choice of the collector whether to maintain privacy of the collected information or not. If creator beforehand chosen not to provide the data, disclosure of which may create problem for him/her later them several problems related to privacy can be avoided. This justifies the importance of "data creation phase". To strengthen the need and importance of the data creation phase, the researchers have collected data of 165 major cases of privacy breach from various sources. These case studies are available at [18-46]. The data is analyzed keeping in view the importance of data creation phase. The results reflect that if the creators would have taken care in advance about what data should be provided and what not, then theses fraud could have been avoided.

VI. STATISTICAL ANALYSIS AND RESULTS

For the purpose of the research Social Networking Sites (facebook, twitter, instagram, linkedin, wechat, snapchat) and Commercial Sites (jeevansathi.com, shaadi.com, policybazar.com, and lifepartner.com) have been chosen. To investigate it is found that these sites frequently ask for the information including email id, mobile No, Date of Birth, gender, address etc. These are the basic information required by social sites and business sites. Fig. 3 and fig. 4 show the percentage of basic information asked by the social networking sites and commercial sites respectively. It is found that mobile no; email id and date of birth are the most frequently asked information.

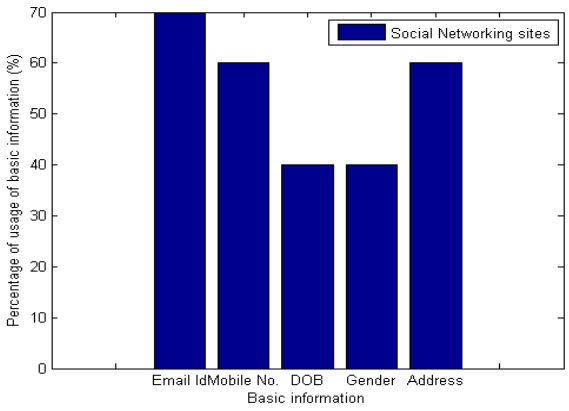

Fig.3 Basic Information asked by Social Networking sites

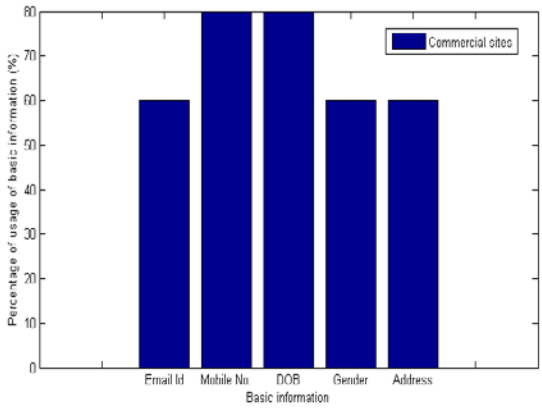

Fig.4 Basic Information asked by Commercial sites

Though this information does not look very significant with respect to privacy and security but exposure of this has given birth to several securities related incidents. For the purpose of research, the major incident during the year 2006-2016 has been taken for analysis. The analysis (refer fig 5) shows that with the reveal of only mobile number and email id, several incidents are taking place. Increase in the security related incidents with the year shows the importance of the data creation phase.

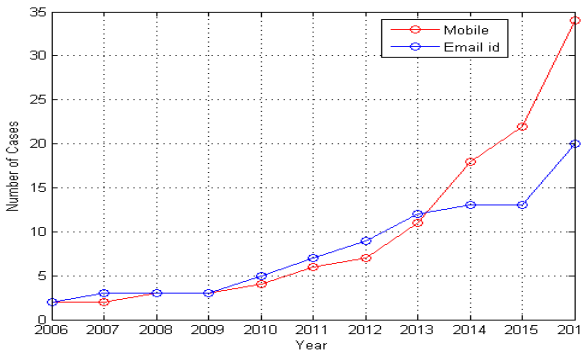

Fig. 5 Number of cyber cases happened due to reveal of mobile number and email id

The research has also calculated the amount of money lost in these incidents. Fig. 6 and Fig. 7 show the year wise loss in terms of money due to reveal of email id and mobile number respectively. This again proves the importance of data creation phase in the life cycle.

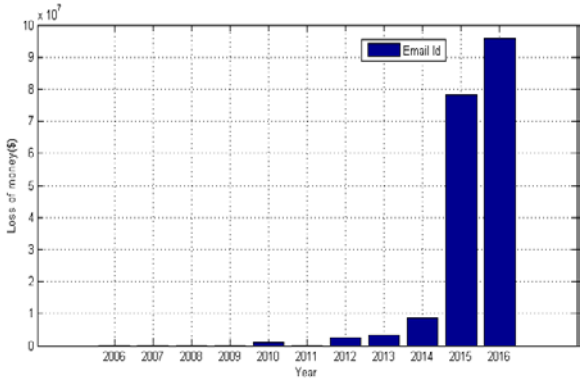

Fig.6 Loss of money by misusing email id

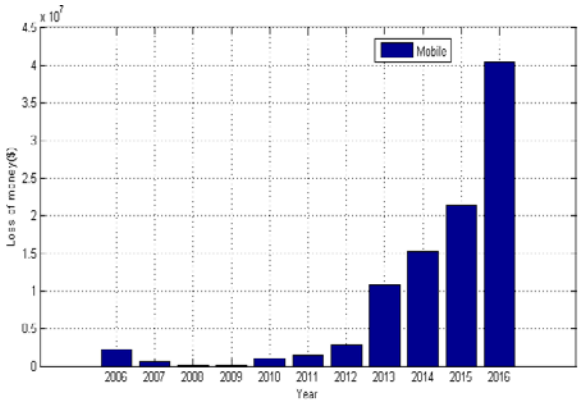

Fig.7 Loss of money by misusing Mobile Number

VII. CONCLUSION

Big data is becoming a huge boon for the IT industry. It is used in various sectors such as health sector, social media, government sector and public sector, etc. because of IT technologies and the internet, a large amount of digital information created and distributed each day. In today's heterogeneous environments data is collected from several sources such as social networking, through sensors, online transactions, etc. It is very vast in size that cannot be easily handled through traditional database handling techniques. With the increased use of big data, there arise many issues; especially security issues which may badly impact a person's or an organization's privacy. In this paper we have presented an improved security threat model for big data life cycle, it consists five phases including data creation phase, data collection phase, data mining phase, data analytics phase and decision-making phase. In this paper authors have shown the vulnerability of basic information that is asked by the social sites and commercial sites. Researchers have collected 165 major cases for the theoretical and statistical analysis. This analysis shows that with the reveal of only mobile number and email id, several incidents are taking place. Increase in the security related incidents and money loss with the year shows the importance of the data creation phase. The understanding of the behaviour of security related incidents and impacts of these incidents on society will help to find out the enough means to overcome the situation.

REFERENCES

[1] L. Ertoz , E. Eilertson ,A. Lazarevic ,P. Tan *et al*. "Detection and Summarization of Novel Network Attacks Using Data Mining", *Technical Report*, 2003

[2] B. Muthukumaran, "Cyber crime scenario in India", *Criminal Investigations Department Review*, pp. 17-23, 2008.

[3] Nigerian national held for lottery fraud, available at: http://newsaboutfrauds.blogspot.in/2009/05/nigerian-national-held-for-lottery.html?m=0

[4] Shri rajesh aggarwal, available at: https://it.maharashtra.gov.in/Site/Upload/ACT/DIT_Adjudication_PoonaAuto_Vs_PNB-22022013.pdf, November 2011.

[5] J. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the Far East", *IDC Analyze the future*, pp. 1-16, year 2012.

[6] Colin Tankard, "Big data security", *Network Security*, pp. 5-8, year 2012.

[7] Wu Xindong, zhu Xingquan, qing Wu Gong ,Ding Wei, "Big privacy: protecting confidentiality in big data." *XRDS: Crossroads, the ACM Magazine for Students*, pp. 20-23, year 2012.

[8] Kim Hwan Sung, kim nam and chung tai, "Attribute relationship evaluation methodology for big data security." *IT Convergence and Security (ICITCS), 2013 International Conference on. IEEE*, 2013.

[9] Tole Alexandru Adrian, "Big data challenges", *Database Syst*, pp. 31-40, 2013.

[10] Katal Avita, Wazid Mohammad, Goudar R H, "Big Data: Issues, Challenges, Tools and Good Practices", *Contemporary Computing (IC3), 2013 Sixth International Conference on. IEEE*, pp. 404-409, 2013.

[11] Yadav chanchal, Wang Shuliang and Kumar Manoj, "Algorithm and approaches to handle large Data-A Survey", *IJCSN International Journal of Computer Science and Network*, 2013.

[12] Tommie Singleton, available at: https://www.aicpa.org/InterestAreas/ForensicAndValuation/Resources/ElectronicDataAnalysis/DownloadableDocuments/Top-5-CyberCrimes.pdf, October 2013.

[13] Inukollu, et.al, "Security issues associated with big data in cloud computing", *International Journal of Network Security & Its Applications*, pp 45-55, 2014.
[14] Xu, J Chunxiao, J Wang, Y Jian, R Yong, "Information Security in Big Data: Privacy and Data Mining", *IEEE*, pp. 1149-1176, 2014.
[15] Bhaya, Wesam, and Mehdi Ebady Manaa. "A proactive DDoS attack detection approach using data mining cluster analysis", *Journal of Next Generation Information Technology*, Vol. 5. No. 4, pp. 36-47, 2014.
[16] N. Khan, I. Yaqoob, I.A.T. Hashem, Z. Inayat, W.K, Mahmoud Ali, M. Alam, M. Shiraz and A. Gani, "Big data: survey, technologies, opportunities, and challenges", *The Scientific World Journal*, pp. 1-18, 2014.
[17] 12 arrested in vishing case, available at: https://www.scmagazine.com/massive-uptick-in-tax-scam-phishing-emails-records-cost-50-on-the-dark-web/article/648452/ , may 2014
[18] Alshboul, Yazan, Wang Yong and Nepali Kumar Raj, "Big Data Life Cycle: Threats and Security Model.", *Americas Conference on Information Systems*, pp 1-7, the year 2015.
[19] Shirudkar kalyani and Motwani Dilip, "Big Data Security", *International Journal of Advanced Research in Computer Science and Software Engineering*, pp. 1100-1109, 2015.
[20] Daniel Ben, "Big Data and analytics in higher education: Opportunities and challenges", *British Journal of Educational Technology*, pp. 904-920, 2015.
[21] Özköse, Hakan, Emin Sertaç Arı, and Cevriye Gencer, "Yesterday, Today and Tomorrow of Big Data", *Procedia- social and Behavioral Sciences*, pp. 1042-1050, 2015.
[22] Edith Ramirez," Securing the Big Data Life Cycle", *MIT Technology review custom + ORACLE*, pp. 1-8, 2015.
[23] Calling time on telephone fraud, available at: http://www.financial-ombudsman.org.uk/assets/pdf/vishing-insight- report2015.pdf
[24] 2016: A reflection of the year in cybercrime, available at: http://www.csoonline.com/article/3146807/compliance/2016-a-reflection-of-the-year-in-cybercrime.html
[25] Vishing,available at:https://www.techopedia.com/definition/4159/vishing.
[26] http://www.actionfraud.police.uk/
[27] Cyber Law of India,available at: http:// www. cyberlawsindia. net/index.html
[28] Spear phishing, available at: http:// searchsecurity. techtarget.com/definition/spear-phishing
[29] Austia's facc, hit by cyber fraud, fires ceo, available at: http://www.reuters.com/article/us-facc-ceo-idUSKCN0YG0ZF
[30] Private bank's manager arrested for phishing scam, available at: http://www.thehindu.com/news/cities/bangalore/Private-bank%E2%80%99s-deputy-manager-arrested-for-phishing-scam/article13996057.ece
[31] http://www.bankinfosecurity.in/
[32] Not easy to crack cyber crime, available at: http:// www.thehindu.com/news/cities/bangalore/Not-easy-to-crack-cyber-crime/article14006807.ece
[33] Massive uptick in tax scam phishing emails, records cost $50 on the Dark Web, available at: https://www.scmagazine.com/massive-uptick-in-tax-scam-phishing-emails-records-cost-50-on-the-dark web/article/648452/
[34] Phone banking fraud hits thousands; tricksters deal Rs 12,000-cr blow, available at: http://www.hindustantimes.com/india/phone-banking-fraud-hits-thousands-tricksters-deal-rs-12-000-cr-blow/story-AP7DdpYdoHwX0UADrQnxRM.html
[35] Phishing and vishing cases on the rise in city , available at: http://www.thehindu.com/news/cities/Visakhapatnam/phising-and-vishing-cases-on-the-rise-in city/article7680053.ece
[36] Cyber crimes on the rise in State , available at: http://www.thehindu.com/todays-paper/tp-national/tp-kerala/cyber-crimes-on-the-rise-in-state/article3068112.ece
[37] Phishing: DG&IGP's case solved but 57 more wait, available at: http://www.thehindu.com/news/cities/bangalore/phishing-dgigps-case-solved-but-57-more-wait/article7347371.ece
[38] Cabbie falls for phishing call, loses Rs. 65,000, available at: http://www.thehindu.com/news/national/karnataka/cabbie-falls-for-phishing-call-loses-rs-65000/article8386845.ece
[39] Cyber crimes on the rise in State, available at: http://www.thehindu.com/todays-paper/tp-national/tp-kerala/cyber-crimes-on-the-rise-in-state/article3068112.ece
[40] Youth in jail for sending e-mail threat, available at: http://www.thehindu.com/todays-paper/Youth-in-jail-for-sending-e-mail-threat/article14813210.ece
[41] Mumbai: Cyber fraudsters cheat bizman of Rs 17L using duplicate SIM, available at: http://www.hindustantimes.com/mumbai/mumbai-cyber-fraudsters-cheat-bizman-of-rs-17l-using-duplicate-sim/story-ox88HEeD6LAwoC8QOMpRqJ.html
[42] Phone scams: 'A criminal convinced me he was from TalkTalk and stole £13,600, available at: http://www.telegraph.co.uk/money/consumer-affairs/phone-scams-a-criminal-convinced-me-he-was-from-talktalk-and-sto/
[43] Top 5 biggest phishing scams, available at: http:// www.theinquirer.net/inquirer/feature/2460065/top-5-biggest-phishing- scams/page/3
[44] Online lottery frauds on the rise, say cops, available at: http://indianexpress.com/article/cities/pune/online-lottery-frauds-on-the-rise-say-cops/
[45] Available at: http://www.jagran.com/bihar/shiekhpura-cyber-crime-15212116.html?src=Search-ART-cyber
[46] Online fraud: Top Nigerian scammer arrested, available at: http://www.bbc.com/new/world-africa-36939751