Big Data-9Vs, Challenges and Solutions

K. Khurshid[1], A. A. Khan[2], H. Siddiqi[3], I. Rashid[4]

[1,2,3,4]*Military College of Signals (MCS), National University of Science & Technology (NUST).*
[1]kiran.phd@students.mcs.edu.pk

Abstract-Big data refers to the data whose large volume, value, complexity, exponential arrival and growth rate makes it very difficult to be captured, efficiently managed, processed and analyzed within a certain time frame by conventional techniques. This paper provides a comprehensive review on the background of big data, its importance and related challenges. It gives an insight into fundamentals of security and risks associated with big data along with certain workable solutions. Another input in this work is the revision of existing IBM's 3V model of big data consisting of Value, Velocity and Volume. A more realistic model compromising of 9Vs has been devised for future modelling and classification of big data.

Keywords-Big Data, Data Modelling, 3V Model, Data Mining, Data Security, Data Breaches, Information Management.

I. INTRODUCTION

A. Origin of Big data

The concept of big data has been there in history since long however the terminology "big data" was devised in 2005 by Roger Mougalas from an American media company, "O'Reilly Media". In olden times, some several years ago, tracks of using data to record quantities exist. The antiquity of accounting dates to around seven hundred years. In the book, "Themes in the history of book keeping" it is written that primary development of accounting epochs to ancient Mesopotamia. Accounting was familiarized in Mesopotamia for enumerating the growth of herds and crops, hence data recording was done. The credit of developing the first data processing machine named "Colossu" goes to the British nation, who invented it in 1943 to decipher Nazi codes. Colossus appeared during the World War II. It was used for searching patterns in the intercepted messages and had a search rate of 5,000 characters per second. It reduced the human effort from weeks to merely hours. Over the past 20 years, with the advancements in technologies, expanse of data has increased by tremendous amounts in almost all the fields. Information age has started in the 20th century. International Data Corporation (IDC), a subsidiary of International Data Group (IDG) in its report stated that in 2011 the created and copied data around the world had a size 1.8 zetta bytes which within five years increased by almost nine times. Internet of things (IoT) which provides connectivity of devices is also a leading contribution to data explosion.

Nowadays data generation is so speedy, and to date about 2.5 quintillion bytes of data are created daily [i].

If a personal computer has around 500 GB of data, then all the data in the world would require storage capacity of approximately 20 billion personal computers and due to the information technology era processing of such tremendous amount of data will not be that much tedious as it was decades earlier. An example is the human genome decryption process which earlier took approximately 10 years of processing and now within week [ii]

Data generation is so easy and quick, for example, on average, 72 hours of videos are uploaded to YouTube in every minute [iii]. "The Human Face of Big Data" has derived certain statistics on data generation rate. As per the media project there are approximately nine hundred fifty-five million monthly active accounts on Facebook with seventy languages, daily one hundred forty billion photos are uploaded, friend connections made per day is one twenty-five billion, every day approximately three billion likes and comments and thirty billion pieces of content have been posted [iv].48 hours of video are uploaded every minute on YouTube and number of views per day is approximately four billion. Every minute five hundred seventy-one new websites are created. Google processes twenty peta bytes of data and monitors 7.2 billion pages daily with translation facility of sixty-six languages. As far as twitter is concerned one billion Tweets are posted every three days from more than one forty million active users on Twitter [v]. It is expected that information will increase by fifty times in the next decade however number of information technology specialists will increase by 1.5 times only [vi].

B. Vs' of Big data

Doug Laney, an expert of META (currently Gartner), an American Research and Advisory Firm, in a research report in 2001has defined a 3Vs model to describe big data and the trials and prospects associated with it. The 3Vs show Volume, Velocity, and Variety. Although such a model was not originally used to define big data, Gartner and many other enterprises, including IBM and some research departments of Microsoft still used the "3Vs" model to describe big data [vii]. Recently researchers have added one more V to the 3V model that is Value. In the "3Vs" model, Volume is the size of the data. Next is Velocity, which is the timeline of big data per its arrival rate. Variety is the categorization of data aimed for maximally utilizing the commercial value of big data, its gathering, sorting and analysis. Variety has three subcategories

1) Structured Data
2) Semi-structured Data
3) Unstructured Data

IBM has focused on Velocity, Variety and Volume for characterizing and tackling with big data.

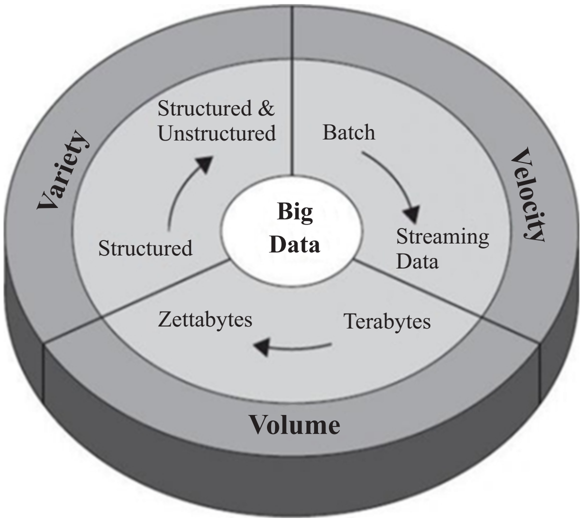

Fig. 1. IBM's characterization of big data [viii]

1) Volume

Volume means large amount of data. As mentioned above in the history section, big data is growing rapidly, it is generated continuously at an unimaginable rate.

- Per the internet live stats, on average over 40,000 search queries are processed by Google every second
- In Facebook's Q4 2016 earnings results report, there are 1.86 Billion active users on Facebook, with 400 million using voice and video chat per month. Number of posts that are liked on Facebook each day are around 6 Billion. Many new accounts are created within a minute on Facebook.
- Baidu which is a Chinese search engine, processes around 100 PB of data per day.
- Taobao which is linked with Alibaba, a Chinese e commerce company for online trading, produces data of tens of TB per day.
- 2.5 Million posts are liked by Instagram users per minute.

The Wall Street Journal reports that every day more than one billion hours of YouTube videos are watched around the world which is more than the regular TV viewership. Around 400 hours of videos are uploaded to YouTube every minute. CTIA, the US wireless communication industry, states that SMS messages sent in US each day are 6 billion, each month more than 180 billion and 2.27 trillion each year.

As per IBM, approximately 2.7 Zeta-bytes of data exist in the digital universe today. This volume is tremendous and processing and storage of extremely high volumes of data whose value is not known and is yet to be determined requires intelligent algorithms and efficient servers.

Big Data needs to be analyzed in such a way as to change the value of data into a high value data, that is, the data that has minimal density to a data that has high density.

2) Velocity

Velocity is the speed of new data generation, and the rate at which data travels from one place to another. Generally, this speed is equated to the speed of light due to very high velocity of capturing, generating and movement of data. The high velocity can be seen from the example that social media messages go viral within seconds, the news which is just out, spreads like fire in the jungle and within seconds it gets twitted, becomes a Facebook status of many and gets discussed on social tools. Similarly, the speed with which credit card transactions are done is so fast that within seconds the data from the card goes into Bank and information gets updated. To track fraudulent activities done using credit card can take only milliseconds. It takes very little time for trading systems to evaluate social media networks and to assess and chose elements that prompt decisions to sell or buy segments. The highest velocity data normally streams directly in the memory versus being written to the disk. It is analyzed directly without even putting it into data bases. Internet of Things (IoT), that is the data from the sensors, mostly required in applications related to health and safety monitoring, requires real time analysis, estimation and action on data.

3) Variety

In data types, there is a huge diversity, for example there are text messages, voice chat, video chat, images, data from social media discussions, sensors' data, audio/video recordings etc. In the past, few decades', big data could have easily been structured, as it was mostly tabular type of data related to finances etc. Now the data is coming from different streams all the time so almost 80% of the total data in the world is not structured, and because of this unordered sequence it cannot be easily arranged into a tabular form for example audio data, images, videos or data from social media.

We can classify data as structured, semi-structured and un structured. Unstructured and semi-structured data types' need more sorting and processing levels to go through as compared to the other data type to derive meaning. Once the meaning of data is extracted then unstructured data and structured data become the same in the sense that unstructured data require the same level of auditability as structured data, similarly summarization details and privacy requirements are same. Complication increase when without any notice, data from a known source changes. These types of frequent and real-time changes are ahuge burden as it involves analyzing the data again and updating the transactional environment.

Initially IBM in their "Understanding Big Data-eBook" addressed Volume, Velocity and Variety only for defining big data but later IBM scientists added one more V of "Value" to the 3V model.

4) Value

The fourth V of big data model is more important now as data without value is meaningless.

To get advantage from data sets, value needs to be determined. There is a need of skill set involving analysis of data along with quantitative and investigative methods. Mostly value is derived by looking at certain patterns from user's list of activities in the data, scanning from his preferences, behavior or sentiments. If meaning is derived then a lot of work can be done on the data for example offers can be made by location using advertising etc., risk analysis can be done, fraudulent activities can be caught, value added data can be sold etc. For finding the value of data along with software tools and artificial intelligence techniques, insightful analysts are also required.

The original 3Vs are now becoming old school as they are more IT or infrastructure oriented. The model lacks data discovery, identification of new patterns, quality, reliability and development process. The new 6Vs, veracity, visualization, value, viscosity, virality and validity are more important now. We have defined a new model for big data consisting of 9Vs, 6Vs on the top of the old 3Vs.

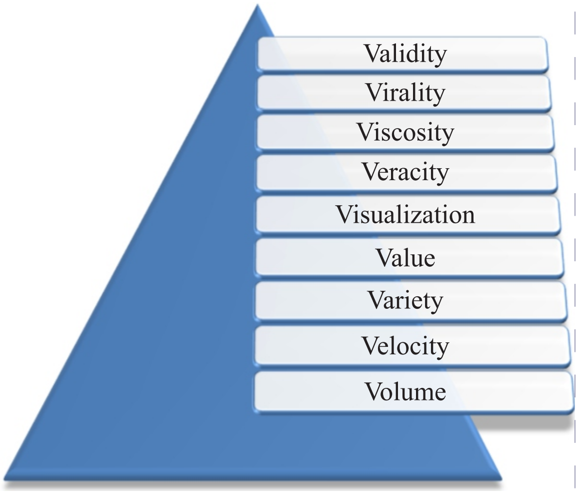

Fig. 2. 9V model of Big Data

5) Visualization

Big data generally appears like a black box, to get new direction into analysis visualization is the key concept. Visualization helps in making one see multiple dimensions of big data. Although new software and technologies have enabled analysts to understand and get a deeper insight to the tremendous amount of data however still unprecedented data visualization is needed totruly understand the crux of data.

6) Veracity

Veracity is the accuracy and reliability of data. It addresses the issue whether the data stored and arranged for mining is meaningful to the problem at hand or not. The data which is not of interest appears as noise and abnormality in the system and must be kept away from the data of interest. Veracity is one of the biggest challenges in data analysis unlike the other Vs.

7) Viscosity

It is related to the resistance when going through certain data sets and the complexity of processing it. When navigating through big data, because of different variety of data, complexity of processing each set is different. Faster complex event processing is required for removing the resistances associated with data sets.

8) Virality

Virality is spread rate of data in the network. Prediction of the virality content can be useful in many applications and for viral markets. It can even help in optimizing the network performance.

9) Validity

Validity is the rationality and strength of data. If accurate data is available for analysis, then speed of processing and analyzing the data is fast and right decisions can be made.

All these Vs bind the discovering of data process, quality measuring, model development and collaboration, and above all hook into organizational benefit. But theoretically and in all the discussions old 3V's are given importance as they form the baseline for big data.

C. Technologies

For the efficient processing of tremendous amount of data, we need exceptional technologies so that data may be processed within tolerable limit of time. McKinsey report suggests suitable technologies include [viii] A/B testing, genetic algorithms, data fusion and integration, simulation, machine learning, crowd sourcing, natural language processing, signal processing, time series analysis and visualization. Apart from these, other tools which are applied to big data include massively parallel processing (MPP) databases, search-based applications, distributed file systems (DFS), data mining, distributed databases, cloud-based infrastructure and the Internet. Most of MPP relational databases can store and process data up to petabytes.

II. Importance of Big Data

Companies can gain much better understanding of their business, customers, products' value, competitors, etc., if big data is efficiently captured, effectively processed and analyzed. They can find out their flaws which can be looked upon for improvements. This can lead to efficiency developments, increased sales, better products and services offered and efficient customer service.

Big data has high potential in many areas of life. McKinsey Global Institute has specified five areas which can be improved by the knowledge of big data. These areas are healthcare, Public sector, retail, manufacturing and personal location data [ix].

1) Healthcare: Big data can be used in clinical decision support systems, it can provide analytical information of patient's profile for giving personalized medicine prescription. The information of patients can be useful for analyzing disease patterns to improve public health.

2) Retail: By observing demands patterns of users' big data can be used in product placement designs and improvement in performance. It can help in optimization of price, variety, store behavior analysis and web based markets.

3) Public sector: It can discover needs of people and make customized actions for products and services. This can help in improving performance, decision making and innovation of new products and services.

4) Manufacturing: Big data can improve sales by demand forecasting with the help of web search based applications. It can improve supply chain planning and sales support.

5) Personal location data: Big data can help in advertising targeted locations. It can provide emergency response and planning of urban locations and new business models.

A. Usefulness of Big Data

Big data plays a major part in the improvement of economic functions, gives a boost to the productivity, develops healthy competition among enterprises and organizations of public sectors and creates huge benefits for consumers.

During 2009 flu epidemic, Google played a key role in informing the world about prevention and cure techniques. The company obtained timely information by analyzing big data. Even the disease prevention centers did not provide such valuable information as Google did. To prevent the spread of disease, almost all countries required the prevention and cure centers to quickly spread the information about the ways to prevent and cure the epidemic influenza. Mostly patients who got infected usually did not immediately visit doctors. Secondly information transfer rate from hospitals to the disease prevention centers was not very quick and delay was involved then the disease prevention centers also took time to analyze the information sent from the hospitals and to obtain summarized results. Google found this out by noting that the entries which were sought repeatedly at its search engines during the spread of disease were different from those at which were searched for at ordinary times. It was exposed that there was correlation between the search of influenza spread and frequency of searching in both location and time. There were 45 search entry groups that were linked to the spreading of influenza. Google used this information in certain mathematic models to deduce results. They used the results and forecasted the spread of influenza and predicted even the places where influenza started spreading from. The research results related to this have been published in Nature [x].

Another case is that of Farecast. In 2008, Microsoft purchased Farecast, an online sci-tech venture company in the U. S. Farecast forecasts the travelling trends and increase or decrease in air ticket price ranges of different airlines. In Bing search engine, the system has also been incorporated by Microsoft. By 2012, the system has saved nearly USD 50 per ticket per passenger, with the forecasted accuracy as high as seventy five percent [xi].

More than seventy percent of state IT officials think big data is extremely useful in public safety. For instance, police departments use big data knowledge to create analytical and predictive models for knowing where crimes may occur and at what instance of time they are most probable to occur. This helps them in deploying officers to prevent crimes, thereby helping to reduce the overall crime rate in specific locations.

These were very few examples, if big data databases are analyzed properly they can give benefits in any vertical.

B. Investment in Big Data

Obama regime in 2012 announced investment of more than 200 million dollars in research of big data. The reason of investment was to develop instruments and methods for accessing, organizing and collecting findings from tremendous masses of digital data [xii].
A recent survey of IT and business leaders by Gartner states that more than three quarters of companies are either investing or they are planning to invest in future, in big data which is a three percent increase over 2014.
As per “big data public private forum” the European Commission is funding the 2 years long Big Data Public Private Forum through their Seventh Framework Program to engage companies, academics and other stakeholders in discussing big data issues. The project aims to define a strategy to improve the big data economy. Outcomes of this project will be used for Horizon 2020 for their next framework program [xiii].Frobes Insight's survey of 2017 states that 90% of businesses are investing on big data on medium to large scale with the following adoption rates:

Investment Banking :70%, Telecom: 60%, Health care: 55%, Professional services and manufacturing: 50%.

It can be seen clearly that all the sectors are making investments on big data to gain key insights for making smarter decisions.

III. Similar Technologies

There are few technologies which are closely linked with big data. These include cloud computing, Internet of Things (IoT) and data centers.

A. Cloud Computing

Cloud computing and big data have very close association. Cloud computing uses efficient and well-equipped storage capacity and computing resources to provide solutions for processing and storage of big data applications under concentrated management. Both the technologies have mutual beneficial relation, the emergence of big data has geared up the growth of cloud computing.

B. IoT

Big data and the IoT work in unification. IoT is linked with electronics, software, sensors and networking for flow of information between different entities. For collecting data of several types different kind of networking sensors are embedded in machines, devices and even humans. The data can be record of environmental changes, security, geographical data, astronomical, medical and logistic data. IoT can be considered as the major contributor to big data. It can be taken as a sub class of big data.

C. Data Center

Data center is a big repository of data, used for storing and management of data related to a business.

The data centers can be very useful for big data as they not only provide concentrated storage but can also be used for acquiring data, managing it, organizing it per the requirements and generating the data values. As big data is emerging at an enormous rate, it is bringing great development opportunities for data centers but on a negative side increasing number of challenges as well. Just like IoT, data centers and big data go side by side as emergence of big data is promoting the emergence and creation of efficient software and new infrastructure of data center.

IV. Challenges

A. Heterogeneity

Data is not of the same type, it is generated in my different forms like pictures, texts, and audios etc. Plus, data is different in structure, semantics and organization. If data was homogeneous processing would have been easier, however due to the heterogeneous nature of data, complexity of processing arises. To cater this problem first efficient representation techniques must exist to increase the value of data. Secondly redundancy must be removed, unnecessary components must be filtered so that compressed and better organized data is available for extracting information.

B. Life Cycle

The problem of current storage systems is that they cannot support massive data. Every system has a limit to store data. Scalability of databases is a very big issue. There must be criteria for deciding which data needs to be kept in storage and the bulk which must be wasted. Generally, more the data, more value it has.

A realistic life cycle to organize data comprises of collection, processing, storing and securing, using, sharing, archiving and finally destroying the data. Data privacy, high-level prioritization, liability, value and policies may change the life cycle procedure for Big data.

C. Security

The two main problems faced by almost all the companies working with big data are:

- The struggle of management and timely retrieval of enormous and ever-increasing amount of data.
- The security and privacy concerns.

As per Cyber Edge [xiv] more than 60 percent of 763 security practitioners survey reported successful cyber-attacks on their midsize-to-large companies. The Verizon 2015 Data Breach Investigations Report (DBIR) tallied nearly 80,000 security incidents, including 2,122 confirmed data breaches [xv].

Few years back, a single data breach affected records from 1 million to 10 million of victimized company but nowadays a single breach can compromise over 200 million records causing multi-million-dollar loss, damage to the name of brand along with facing regulatory consequences.

An example is the case of Ashley Madison, a website for extramarital affairs. It suffered from data breach in July 2015. 25gigabyte of the company's data was leaked. This resulted not only in destroying the repute of victim but also resulted in disturbing their customers with two suicide cases reported.

Another case is of Target, United States based discount retailer, the data breach happened in 2013 holiday shopping season. In the data breach, which occurred at the point of sale system, cyber criminals stole the information of at least 40 million credit and debit cards. They also hacked data of about 70 million additional customers which included their names, addresses email addresses and phone numbers.

As reported by Naked Security, “In total, the breach has cost Target $290m so far, of which insurance should cover $90m, the company. However, there are still shareholder lawsuits to come, as well as probes by the Federal Trade Commission and state attorneys general, which could well push the total costs of the incident to over $300m.”

One more data breach case is of Horizon Blue Cross Blue Shield of New Jersey. The company has faced two such cases, one in 2013 and other in 2016. In January 2008, an employ's laptop was stolen after which the corporate policy of the company was changed. The company stated to encrypt all the laptops, desktops and mobile devices. However, in 2013's data breach, investigation concluded that many laptops were not encrypted. Again after 2016 data breach it was found out that many of the laptops were password protected and not encrypted which resulted in exposure of the names of customers, their birth dates along with insurance identification and social security numbers and in some cases customer's clinical information. Per a press statement the company had to pay 1.1million dollars for providing poor security solutions.

Similar case is of Yahoo. The company stated a breach of around 500 million email addresses in September 2016. In the same year, 117 million users' data on LinkedIn, a professional networking site, was compromised and black market purchased massive amount of emails and passwords.

From few of the breaches mentioned above it is clear that big data is brining big security headaches. It is through these harsh experiences that business and IT experts are learning to protect and secure big data in a better way.

V. Security Solutions

Big data despite of all its advantages, opportunities and prosperity is also bringing up more challenges with information security [xvi].In a report by The Identity Theft Resource Center, the number of data breaches has grown by 40% percent in 2016. Big data is a big problem if proper security policy, periodic technical and non-technical evaluation and data encryption solutions are not in place.

A security model consisting of the following points has been given [xvii, iv]

- First characteristic of the security model is that both the internal and external data sources matter for innovation. They multiply in value and create a cooperative learning effect.
- Second characteristic of model is automated tools for gathering and normalizing various data types.
- Third characteristic focuses onanalytical engines for processing of massive volumes of data in real time.
- Fourth characteristic is about advanced monitoring systems that continuously monitor systems and resources and then make decisions based on behavior and risk models.
- Fifth characteristic of model is Active controls for decision making such as additional authentications and blocking of data transfers.
- Sixth characteristic is to have a centralized repositoryfor security analysts to query all security related information.
- Next characteristic discusses multi-layer infrastructures that allow scalability and are capable to process long and complicated searches and queries.
- Last characteristic is to have high enough integration through security and risk management tools for simplifying the process of investigation pertinent to probable problems.

Along with these characteristics, big data analytics need new machine learning theory and better artificial intelligence algorithms as the process is becoming outdated and output lack intuitive physical interpretation [xviii].

Better interpretation techniques are required for deriving information from the machine learning algorithms [xix].

Map reduce is generally the technique used for sorting and analyzing data. It involves two steps. First, one is Map step that involves breaking the big data into smaller chunks. The data chunks are then processed by worker nodes under the supervision of job tracker node. In the second step, which is Reduced step, data analyzing and merging is done. Hadoop, inspired from Google's Big Table, is framework that is based on map reduce strategy. One big problem associated Hadoop is that it was not developed for securing data rather it was created to store and process big data in a faster way, using a distributed fashion approach. Many IT companies are using Hadoop despite of the security concerns. A Better software framework is needed which must be up-to-date, scalable and security equipped for better management and security of big data.

One reason for security breaches is that, all the data is kept collectively at one place. For attackers keeping all the important data in one single place makes it very easy to hack and sabotage the organization.

Another security concern is weak authentication system. Strong cryptography is needed for secure authentication system. Intrusion Prevention System (IPS) and Intrusion Detection System (IDS) must be employed to prevent malicious attacks on the networks. Moreover, firewalls need to be monitored and updated continuously to avoid imminent treats.

Authority to access data must be in a hierarchal way i.e. privileges should be for certain people only to access databases. Only administrators must have physical access to data [xx]. Better encryption and hashing techniques are also needed for end to end data protection.

All the data at different nodes must be protected using highest security standards. Consistent maintenance, monitoring, and analysis of audit logs files is also important.

Despite of all the security measures discussed above preventing big data from breaches and making security-equipped frameworks is still an open research area where exploration needs to be done. Methods and cohesive efforts are needed for preserving and protecting confidential information in big data systems. [xxi]

VI. Conclusion

In this article, we have presented an overview of big data's content, its scope, related technologies, usefulness, challenges and security concerns. For future referencing we have devised a 9V model for big data based on the IBM's 3V model. The 9V model gives a more realistic and practical approximation of big data than the old 3V model. We have also provided a snapshot of current privacy and security state for big data to understand and discover further research areas. Even though description, methods, tools and techniques linked with big data are available in the literature, there are still many points left to be analyzed. In this work, we have not been able to resolve the entire matter about this important topic but we have tried to summarize all the most important aspects, hopefully it has provided some useful discussions and points to ponder on.

References

[i] Big Data Analytics [Online]. Available:http://www-01.ibm.com/software/data/bigdata/

[ii] P. Bhardwaj, A. Gupta, M. Sharma, M. Gupta and S. Singhal, "A Survey on Comparative Analysis of Big Data Tools,"*International Journal of Computer Science and Mobile Computing*, Vol.5 Issue.5, pg. 789-793, May- 2016.

[iii] V. M. Schonberger and K. Cukier. Big data: a revolution that will transform how we live, work, and think.2013.

[iv] S.Seref and D. Sinanc, "Big data: A review," *2013 International Conference on Collaboration Technologies and Systems.*

[v] B. Marr. Big Data**:** Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance. John Wiley & Sons. 2015.

[vi] C. Tankard, "Big Data Security,"*Network Security Newsletter*, Elsevier, ISSN 1353-4858, July 2012

[vii] Laney D. 3-d data management: controlling data volume, velocity and variety. META Group Research Note, 6 February 2001.

[viii] P. Zikopoulos and C. Eaton, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, 1st ed. McGraw-Hill Osborne Media (IBM), 2011.

[ix] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A.H. Byers, "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, 2011.

[x] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski and L. Brilliant, "Detecting influenza epidemics using search engine query data," Nature 457(7232): 1012–1014,2008

[xi] M. Chen, S. Mao and Y. Liu, "Big Data-a survey," *Mobile Network Appl*, Springer 2014.

[xii] R. Weiss and L. J. Zgorski, "Obama Administration Unveils Big Data Initiative Announces $200 Million in new R&D Investments," Office of Science and Technology Policy Executive Office of the President, March 2012.

[xiii] Horizon 2020 Framework Programme policy on open science, [online] Available: http://ec.europa.eu/programmes/horizon2020/en/h2020-section/open-science-open-access.

[xiv] "2014 Cyberthreat Defense Report," Cyber Edge Group, 2014.

[xv] "2015 Data Breach Investigations Report," Verizon, 2015.

[xvi] Y. Mengke, Z. Xiaoguang, Z.Jianqiu and X. Jianjian, "Challenges and solutions of information security issues in the age of big data," *China Communications*, Volume: 13, Issue: 3, March 2016.

[xvii] Curry, E. Kirda, E. Schwartz, W.H. Stewart and A. Yoran," Big Data Fuels Intelligence Driven Security," RSA Security Briefs available online; http://www.emc.com/collateral/industry-overview//big-data fuels-intelligence-driven-security-io.pdf

[xviii] K. L. Wagstaff, "Machine learning that matters,"*29th Int. Conf. Mach. Learn. (ICML), Pasadena, CA, USA,* 2012, pp. 529–536.

[xix] J. Hu and A. V. Vasilakos, "Energy Big Data Analytics and Security:Challenges and Opportunities," *IEEE Transactions on Smart Grid,* Vol. 7, No. 5, September 2016.

[xx] Big data by Duygu Sinanc. [online] Available: https://prezi.com/hyavodg6h0nb/big-data/

[xxi] J. Shamsi and M. Khojaye, "Understanding Privacy Violations in Big Data Systems," IT Professional, Vol. 20, Issue 3, pp. 73-81, June 2018.