Survey on Big Data Analytics for Digital World

MANMEET SINGH (manmeet_96@ymail.com)

Abstract

Every day quintillion bytes of data is generated,90% of which has been created in the last two years alone. From this it can predict the amount of data that will be generated in future. It is necessary, to introduce some techniques for analyzing such a huge amount data to face the challenges and to deal with limitation of 'Big data Analysis'. Such Big Data analysis can be used nearly in every aspect of our modern society, including mobile services, retail and consumer services, manufacturing, business intelligence, financial services and robotics. The motive of this paper is to understand how big data is generated and the necessity of analyzing such data. This paper also gives a short glimpse of Big Data analysis implications in the real world and its role in every field along with challenges and advantages. This paper also explores various techniques, algorithms, systems of big data analytics in various sectors of digital world.

Keywords: Informatics, Big Data, Mining, Big Analysis, Security

I. Introduction

In today's digital world, the tremendous amount of data generated from various sources due to the increased use of internet, mobile phones, digital devices, social networking sites, satellite images, sensors, etc. This led to generation of Big Data. These data can be in the form of numbers, digital pictures, videos, blogs, sensor information, signals, call logs, etc. In developing country any information has to spread like wild fire in every aspect like education, healthcare, etc. Also in today's world involvement of people in the social networking sites like Facebook, Twitter and Google+ has increased tremendously. Big Data helps to improve decision making, fast transformation, policy making, providing solution and mechanism for the development of society in many eras. For this purposed big data present in the world has to be stored and analyzed which is done using various techniques, algorithms, systems of big data analytics.

A. Big Data and Big Data Analytics

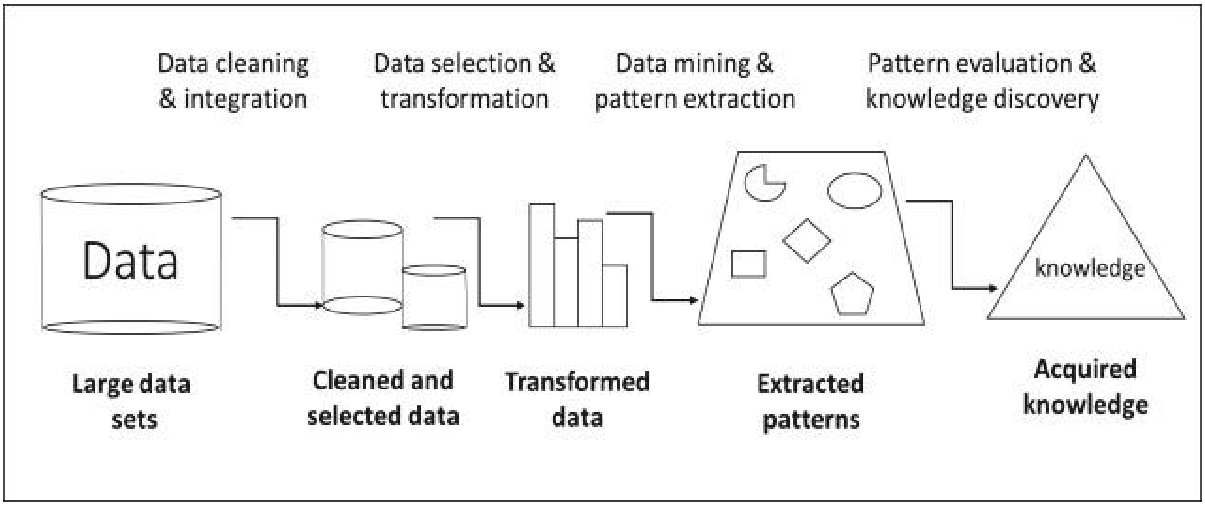
The term big data is being increasingly used almost everywhere on the planet i.e. online and offline. It doesn't completely relate to or depend on computers only. It comes under an absolute/powerful term called Information technologies and fields of studies and business.

Big data is essentially the data on which analysis is done and acquired result is used for prediction and other purposes.

The Big data is a phrase used to mean a massive size of structured and unstructured data. The term structured data refers to data stored in database and unstructured data term refers to data that doesn't reside in a traditional row-column database. Using traditional processing techniques, it is difficult to handle such large and complex data which comes from different sources such as: posts on social media, cell phone, videos, communities, online transaction etc. This data keeps changing day to day in many fields.

When using traditional approach, the data is stored in files which are maintained by file systems under the operating system's control. Data is stored as records. Problems with traditional approach are Data security, Data redundancy, Data isolation, Data dependence, Lack of flexibility, Concurrent access anomalies. Big data will be used by next generations of IT industries which are working on domains like cloud computing, IoT and social businesses. In recent years, big data was helpful in many of areas like healthcare, traffic analysis, business intelligence, data storages, which offers challenging opportunities in future (Debi & Kauser, 2016). In addition to this, study on big data will help to understand essential elements, characteristics and to recognize the complex patterns, to find better information and then knowledge and guide the design of computational methods and algorithms on big data. Generally, Data warehouses have been used to manage large datasets. In this process extraction of knowledgeable data from available datasets is a main issue. It is observed that all the data available in the big data form is not actually useful in every aspect for doing analysis or any decision making process. Hence the biggest challenge is to extract useful data from vast amount of data and utilize it efficiently i.e. Big Data Analytics. In order to extract some kind of patterns and useful information for gaining knowledge, KDD process can be used. Knowledge Discovery is a process of identifying patterns which are valid, understandable and which are strongly useful when retrieved from the large data sets (Francesco, 2015).

Fig.1. KDD Process (Source)


Data cleaning & integration
Data selection & transformation
Data mining & pattern extraction
Pattern evaluation & knowledge discovery
Data
knowledge
Large data sets
Cleaned and selected data
Transformed data
Extracted patterns
Acquired knowledge

The steps involved in Knowledge discovery process are:

- **Data cleaning:** In this step, the unwanted noise and data which is not in consistent manner are removed.
- **Data integration:** In this step, multiple data sources from where the data comes are combined.
- **Data selection:** In this step, the data relevant to the analysis is retrieved from the database i.e., selecting a subset of database or data samples, on which discovery has to be performed.
- **Data transformation:** In this step, by performing summary or aggregate functions the data is transformed coherently for appropriate mining which means reducing and projecting the data, in order to derive a representation suitable for the specific task to be performed.
- **Data mining:** In this step, some intelligent methods for example, summarization, classification, clustering, regression, and proper algorithms are used to extract the appropriate patterns and represent the output results. x Pattern evaluation: Evaluation done by user to identify and extract knowledge from mined data.
- **Knowledge representation:** Acquired knowledge is represented.

Data mining has been a very powerful Big data analysis tool used in KDD process. Big data analytics are used in order to do analysis or process big data so as to disclose the hidden patterns, make decisions, identify the relationships or others. It reduces efforts of government by bringing better services to their citizens by improving healthcare, transport, education and other areas of life cycle. For example: Transport or traffic problems can be solved by doing analysis of traffic records (Rahat *et.al.*, 2016). Also remains helpful in business era by identifying what users prefers, what leaves them unsatisfied so as to provide better services or create improved products Benefits that big data analytics brings are speed and efficiency.

B. Big Data Characteristics

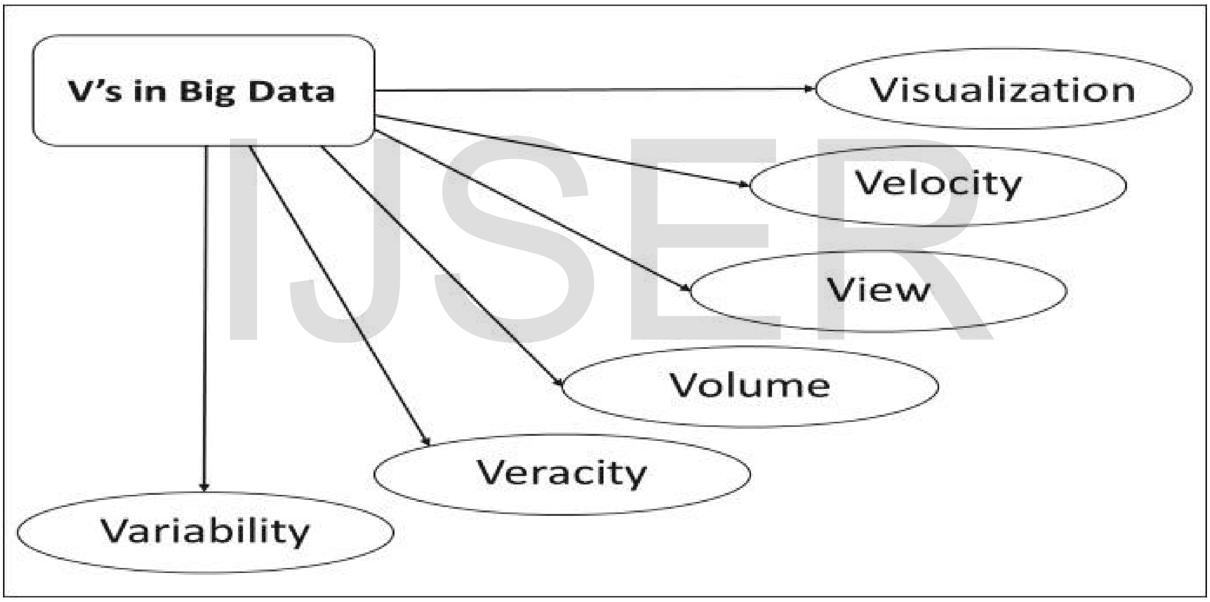
Big data is mainly characterized by Volume, Variety and Velocity.

A. **Volume:** Volume describes the size of data. Today data generated exists in petabytes and most likely to increase up to zettabytes in future. Twitter and Facebook generate around TB's of data every day respectively (Huddar & Ramannavar, Year)

B. **Variety:** Variety refers to the different types of data which includes raw, structured, semi-structured and unstructured data from different sources in the form of pictures (.TIFF, .JPEG, .PNG) or documents (.DOCX, PDF) or videos (.AVI, .MNG) (Quinlan, 1986).

C. **Velocity:** Velocity deals with the speed of data coming from various sources (13) It refers to in what time data is generated and stored and its respective rates of processing and retrieving.

Fig.2. V's of Big Data Source



Big data also additionally characterised by:

- **Veracity:** Veracity refers to the degree in which a person trusts information which is in use to take decisions [12]. In short accuracy of data is being checked for further processing (Quinlan, 1996)
- **Variability:** It shows variation of data in particular variety (Quinlan, 1996)
- It also considers inconsistencies of data flow [13].
- **Value:** User runs queries against stored data and collects important results from filtered data. Also can rank according to requirements [13].
- **Visualization:** Finding a way to represent information that makes findings clear. It is one of the challenges of big data.

II. Challenges Involved in Big Data Analytics

a. Storage: Initially, large amount of data was stored by spending much cost for storage purpose. In the end nobody needs the whole data and they get deleted finally because there is no such large memory space to store them, hence the first challenge for big data analysis is to make available storage mediums with good input/output speed.

b. Knowledge Discovery and Computational Complexities: Knowledge discovery and representation is a prime issue in big data. It includes a number of sub fields such as authentication, archiving, management, preservation, information retrieval, and representation. Since the size of big data keeps increasing exponentially, the available tools may be inefficient to process these data for obtaining meaningful information. It fails to handle computational complexities, uncertainty and inconsistencies which leads to a great challenge to develop techniques and technologies that can deal computational complexity, uncertainty and inconsistencies in an effective manner.

c. Scalability: One more important challenge for big data analysis technique is its scalability. Early years the focus was only on data analysis and its speed up processors followed by Moore's Law. For the former, it is necessary to develop sampling, on-line, and multiresolution analysis techniques.

d. Heterogeneity: It means diversity. When humans consume information, there are great chances of heterogeneity which is handled comfortably. However, computers work efficiently if they are able to store multiple items that are identical in size and structure. Hence to collect such information is one of the challenge in big data analytics [1].

e. Privacy: Concern about privacy and information security regulations is another huge challenge in the context of big data. There is great public fear about inappropriate use of personal data. Managing privacy is effectively both a technical and sociological problem which must be addressed jointly from both perspectives to realize the promise of big data [1].

f. Timeliness: The other side of size is speed. Larger the dataset to be processed, the time needed to analyse it is also large. There exist many situations that requires quick analysis e.g. Fraudulent credit card transaction [1].

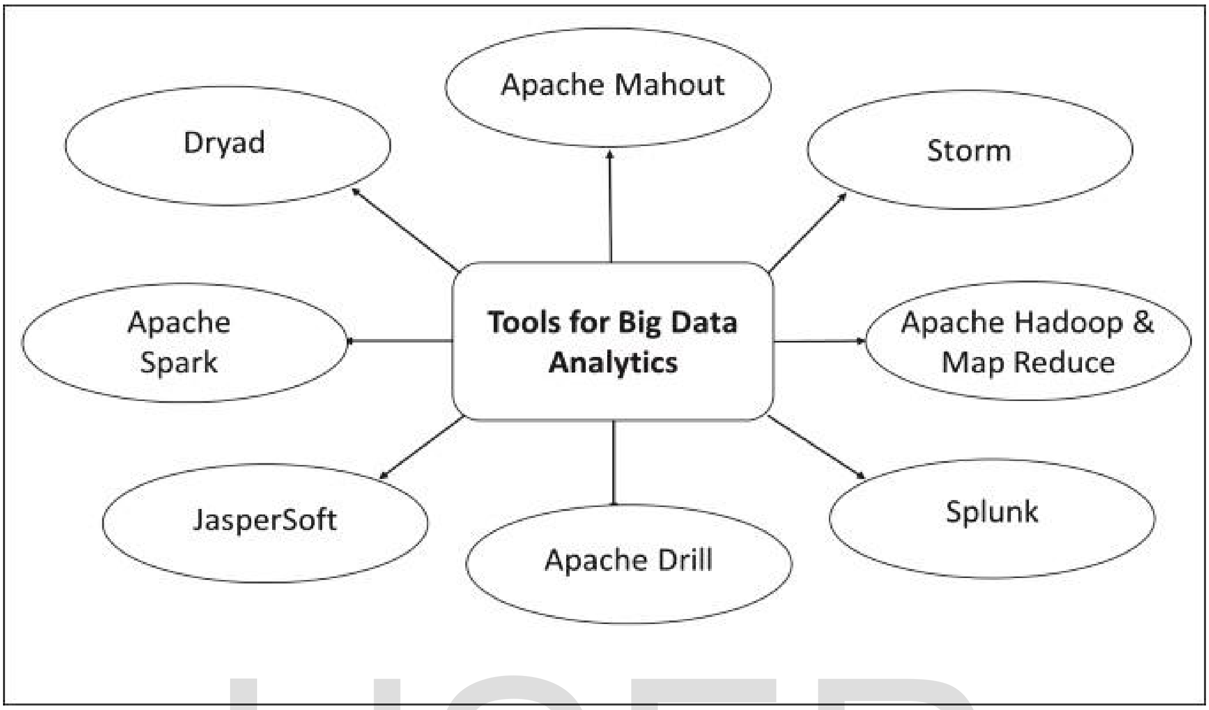
g. Scale: First thing to be taken into consideration is size. Managing large and rapidly increasing volume of data has been challenging issue for many decades [1].

h. Human collaboration: Even though computational analysis has been advanced, there still remains some patterns that humans can easily detect but computer algorithms fails or takes hard time for it. Ideally analytics for big data will not be all computational-rather it will be designed explicitly to have a human in the process [1].

i. **Change in the functionality is biggest challenge.** For e.g. Hadoop is advancing all of the time.

III. Tools For Big Data Analytics

Fig.3. Big Data Analytical Tools (Source)



For processing of Big Data many tools are available. In this section, some current big data analysing tools are discussed:

a. Apache Hadoop and Map Reduce:

Apache Hadoop and Map Reduce is a powerful and well established software framework for solving big data problems, also use for fault tolerant storage that support data-intensive distributed applications. In which Map Reduce work as programming model based on divide and conquer method for large processing datasets. Hadoop also use for easily writing application to process large data in parallel on large cluster with property focuses on fault tolerance and reliability. Hadoop works on two kind of nodes one is master node (JobTracker) and other is slave node (TaskTracker) out of which master node divides input into sub nodes and distributes those to slave nodes in map reduce that means it provides job scheduling and task distribution for slave node which perform all task as assigned by master. Master node perform a task of monitoring the nodes so that if one node fail to execute/perform its task it go for other node to perform that task. In real, Hadoop job client submits the job and configuration to JobTracker which performs a task distribution to TaskTracker, task scheduling, monitoring and provide information to job client. Hadoop having coordination function use to improve performs of Hadoop job (Debi & Kauser , 2016) [15] [32].

b. Apache Mahout

Apache Mahout is open source project by Apache Software Foundation (ASF) used in producing scalable machine learning algorithm. Algorithms of mahout is welldesigned and optimize algorithm including clustering, classification, pattern mining, regression, dimensionality reduction, evolutionary algorithms, and batch based collaborative filtering run on top of Hadoop platform through map reduce and it has designed to establish a vibrant, responsive, diverse community to facilitate discussions on the project and potential use cases. A person has to purchase license of apache software to use it. Apache Mahout provide scalable and commercial machine learning techniques for large scale and intelligent data analysis applications such as Facebook, Google, and Twitter [22] (Debi & Kauser , 2016) [15].

c. Apache Spark

Apache spark is open source cluster computing and data processing framework mainly built for speed, ease of use and sophisticated analytics originally developed in 2009 at UC Berkeley's. With its rapid selection across a wide range of industries it has become a largest open source community in Big data, for example used by Netflix, yahoo, etc. Apache Spark runs on the top of Hadoop distributed file system (HDFS) infrastructure to provide enhanced and additional functionality. It mainly consists of driver program, cluster manager and worker nodes where driver program is as starting point of execution, cluster manager allocates resources and worker nodes do the data processing in the form of tasks. Advantages of Apache sparks are it provide fault tolerance without replication, supports MapReduce as well as streaming data, machine learning, and graph algorithms, can be run on different languages, helps to run an application in Hadoop cluster, up to 100 times faster in memory, and 10 times faster when running on disk or storing data on disk has become record holder for large scale on disk sorting (Debi & Kauser , 2016) [15].

d. Dryad

Dryad is another popular investigating programming models for writing parallel as well as distributed program to scale from small cluster to a large one. In other word dryad is an infrastructure allows user to use computer cluster or data centre cluster for parallel and distributed programming. Without knowing anything about concurrent programming Dryad users use thousands of machines, each of them with multiple processors or cores. Dryad provides a large number of functionality including generating of job graph also has capability to synthesize any direct acyclic graph, scheduling of the machines for the available processes, transition failure shandling in the cluster, collection of performance metrics, visualizing the job, invoking user defined policies and dynamically updating the job graph in response to these policy decisions without knowing the semantics of the vertices (Li *et.al.*, 2012; Debi & Kauser , 2016). [15].

e. Storm

Storm (event processor) is a real time distributed and fault tolerant computation processing system for large volume high velocity data in contrasts with Hadoop which is for batch processing. Storm real time processing system developed by BackType. Storm has advantages like is used for large streaming data, distributed and fault tolerant real time computation system for processing large streaming data, it is east to operate, horizontally scalable, runs by any programming language, provides guaranteed message processing (Debi & Kauser , 2016). Storm cluster consists of two kinds of node master and slave which implement two type of roles such as nimbus and supervisor respectively has a same functionality of MapReduce framework with JobTracker and TaskTracker (Debi & Kauser , 2016).

f. Apache Drill

Apache drill is distributed system for interactive analysis which has more flexibility to support many types of query languages, data formats, exploit nested data and data sources. It uses HDFS for storage and for batch analysis uses MapReduce (Debi & Kauser , 2016).

g. Jaspersoft

It is a scalable big data analytical platform and has a capability of fast data visualization on popular storage platforms, including MongoDB, Cassandra, Redis etc. and also Jaspersoft is that it can quickly explore big data without extraction, transformation, and loading (ETL), have an ability to build powerful hypertext mark-up language and generated reports can be shared with anyone inside or outside user's organization. (Debi & Kauser , 2016)

g) Splunk

Splunk is a real-time and intelligent platform developed for exploiting machine generated big data. It combines the upto the moment cloud technologies and big data. It is different from other because it includes indexing structured, unstructured machine generated data, real-time searching, reporting analytical results, and dashboards (Debi & Kauser , 2016)

IV. Applications of Big Data Analytics

1. Student's behavior monitoring:

Terrorism is considered as major threat to the society in today's times. Major issue like Security threat from senseless terrorist attacks on innocent/unarmed civilian should be taken into consideration. Big data analytics helps to deal with such security related problems. The basics of system is proposed based on big data technologies that can be used to examine/monitor the students from any particular university and with the predictive analysis conclusion or prediction is done whether some of the students are becoming habituated to unorthodox ideologies that may lead to very sensitive issue like terrorism. In a university setting, a huge amount of personal and academic data is available. Various tools are available for monitoring and analysis of data. The basic idea is analysis of student's behavior aims to monitor a student to observe if he/she is deviating from normal behavior. If ideological deviation is not checked, this may lead the student to fall into some illegal or unauthorized activities. Hence the monitoring and prediction provides the early warning system to prevent loss of life with backed up information, guidance, advice, motivation and feedback to particular student and also respective guardian. This can help in improving student's behavior and also realization about further consequences is done. This can save the innocent lives and removes the negative fallouts of terrorist activities. The sources of big data for such behavior analytics: traditional databases, personal data, web digital trail, outdoor activities, surveillance videos, parking sensors. Hadoop Platform is the most popular trend to deal with big data in recent days. Hence enough data is available in a university environment that can be used with the help of Big Data model and accompanying technologies to monitor and predict deviant behavior in students [2].

2. Social Media:

The amount of information now available to crunch and parse in the service of analysing absolutely anything is massive and growing every second. In understanding big data's impact on social media marketing strategies is that social media is a part of big data. The study gives research on implications of the use of big data analytics for business intelligence purpose. This is implemented on the data collected from Social media channels in China. BI plays an important role in improving organizational performance by identifying new opportunities, highlighting potential threats, revealing new business insights. Increase in use of social media like twitter, Facebook, weibo gives birth to the big data. This study helps for BI to improve decision making capabilities, faster decision making, understanding of customer needs, developing strategies for launching new products and services, exploring new markets, improving inventory turnovers, reducing customer complaints, and enhancing staff productivity and efficiency in Chinese business [6 Considering the social media challenges for big data analysis, the design of twitter data analysis is implemented using MongoDB DBMS in server-side for static or dynamic data sources and for APIs on the client-side HTML, PHP, JavaScript are used (Rekha & Parvathi, 2015). Big data analytics in social media era is a big challenge as all of the data generated by social sites is in large amount as well as in unstructured form (Rekha & Parvathi, 2015).

3. Mobile Networking:

Big data has been a catch phrase in the several sectors for many years, but now a day's network sector is also realizing its potential. Big data is important to mobile operators as it promises to provide growth in several ways. Big data analytics helps to search for better understanding of their networks, operations and customers. Also improves the performance of mobile cellular networks also focuses on new revenue streams (Liu *et.al.*, 2015). Paper introduces a unified data model based on the random matrix theory and machine learning. Also represents architectural framework for applying the big data analytics in the mobile cellular networks. Big data analytics efficiently extracts more insightful information than traditional data analytics. Matrix neural networks takes matrices directly as inputs.

4. Business analytics:

Customer behavior analytics is an upcoming and unexplored market that has greater Potential for better advancements. Knowing which customers are most valuable buyers is important because it helps for further business. Big data comes into picture here that have capability to take the business organization at higher level by analyzing customer behavior and transform it into valuable insights. At this point big data analytics is necessary. For analysing data decision tree can be used efficiently. The survey provides Map reduce implementation of well-known statistical classifier C4.5 decision tree algorithm (Quinlan, 1993 & 1996) Also system aims to customer data visualization using D3 (Quinlan, 1986) i.e. Data Driven Documents that allows to build customized graphics. Customer analytics is incomplete without data visualization.

Key concepts for customer analytics are: Venn diagram, Data profiling, Forecasting, Mapping, Association rules, Decision tree [6]. Tools for data visualization are Polymaps, Flot, D3.js, SAS Visual Analytics [2]. Big data analytics can be applied to predict the risks in software projects at early stages to increase its productivity and profit (Quinlan, 1996).

5. Affective Humanoid Service Robots:

The increasing demand for automation in all aspects of life has majorly contributed to the growth of Humanoid service robots to serve in highly complicated and intelligence demanding applications. Application such as smart home/school/campus environments, smart care, healthcare, children education. The highly complicated and intelligence demanding applications by using a Big Data Analytics as a Service approach, with which a novel Distributed Collaboration and Continuous Learning (DCCL) middleware platform is developed to support

collaborative humanoid service robots. Big data analytics also utilizes scalable data processing platforms such as Cloud computing with customized Data Mining or Machine Learning techniques which is useful in humanoid service robot era [30].

6. Healthcare and government agencies:

Data generated by the healthcare and government agencies is very large and that too in unstructured form. Without proper processing and analysis all of this data cannot be made useful. Using Big Data analytics Hadoop performing real-time processing on large datasets is helping in improving healthcare. The data in healthcare organizations is generated from records in hospitals, clinics and patient’s data. Big data helps in uncovering decision making by identifying data patterns and relationships between these patterns using machine learning techniques. Big data sources in healthcare era are clinical reports, X-Rays, history of patients in hospitals, diet followed by patients, lists of doctors and nurses in particular hospitals, health register data, medicines and their expiry dates. Based on these data an improved healthcare is provided for patients. By predicting the basic needs of citizens through analysis done on survey conducted among citizens can be implemented using Hadoop along with enabled security using access control schemes [25].

Big data processing includes security challenges like providing network level security, authorized users are involved in systems only, maintaining logs for identifying hackers. Where these issues can be solved. For providing network level security communication is done using RPC in systems. And for authentication of users a two-way authentication is provided. For providing security to data encryption algorithms are used where this encrypted data is transmitted with attribute-basedencryption method and so that malicious users cannot access data. In such cases, map-reduce job helps in identifying that which user is responsible for the leakage of sensitive data [25].

7. Smart and Connected Communities:

The main focus of smart and connected communities SCC for smart cities is to live in the present, plan for the future, and remember the past. The main focus is to improve livability, preservation, revitalization, and attainability of a community. By using Big data and IoT to SCC will help in many ways for smart city. As IoT has ability to provide a present network of connected devices and smart sensors for SCC and Big Data has potential to provide real-time control for IoT. The main opportunities of IoT in SCC are mobile crowd sensing (MCS) and cyber-physical cloud computing. This is helpful in various SCC application like healthcare, disaster management, transportation at smart level. This leads to generation of large data which is mainly handled by Big data analytics. These Big data analytics provides smart decision making, analysing, collection etc. for SCC data. But this both era IoT and Big data in SCC has some challenges. Big data in SCC is fighting with data heterogeneity due to different data from different sources, sensors and for different purpose or operation. It also has problem of decision making in under uncertainty which may improve by understanding, representing, processing, optimal sequential decision making [31].

V.Conclusion

In recent years, the amount of data generated is vast. In this paper, we survey the big data analytics, its various challenges and issues and tools to analyze the big data. Through better analysis of the large volumes of data that are becoming available, there is potential for making faster advances in many scientific disciplines and improving the profitability and success. By effectively applying big data analytics, nearly for every department involving sales and marketing, customer support, business intelligence, operation and maintenance, network construction, etc. can achieve significant benefits. We hope the content discussed in this paper, can be helpful for future analytics.

References

[1] "https://courses.bigdatauniversity.com/courses/course-v1:BigDataUniversity+BD0101EN+2016/info," [Online]. Available: https://bigdatauniversity.com/. [Accessed 7 Feburary 2017].

[2] J. Zhu, E. Zhuang, J. Fu, J. Baranowski, A. Ford and J. Shen, "A Framework-Based Approach to Utility Big Data Analytics," *IEEE TRANSACTIONS ON POWER SYSTEMS,* vol. 31, no. 3, pp. 2455-2462, 2015.

[3] C. Zhang, X. Shen, X. Pei and Y. Yao, "Applying Big Data Analytics Into Network Security: Challenges, Techniques and Outlooks," in *IEEE International Conference on Smart Cloud*, New York, NY, USA, 2016.

[4] Y. Wang and N. Hajli, "Exploring the path to big data analytics success in healthcare," *Journal of Business Research Elsevier,* pp. 287-299, 15 August 2016.

[5] Y. Wang and N. Hajli, "Exploring the path to big data analytics successin health care," *Elsevier,* pp. 287-299, 2017.

[6] G. Wang, H. Chen and J. Xu, "Automatically detecting criminal identity deception: an adaptive detection algorithm," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans,* vol. 36, no. 5, pp. 988-999, 21 August 2006.

[7] C. Verma and D. R. Pandey, "Big data representation forgrade analysis through hadoop framework," in *IEEE 6th InternationalConference - Cloud System and Big Data Engineering*, India, 2016.

[8] P. Vashisht and V. Gupta, "Big Data Analytics Techniques: A Survey," in *IEEE InternationalConference on Green Computing and Internet of Things (ICGCIoT)*, Noida, India, 8-10 October,2015.

[9] R. van der Hulst, "Introduction to Social Network Analysis (SNA) as an investigative tool," *Trends inOrganisedCrime,* vol. 12, no. 2, pp. 101-121, june 2009.

[10] M. Utmal and R. K. Pandey, "Taxanomy on the integration of hadoop and rapid miner for big data analytics," in *International Conference on Computational Intelligence and Communication Networks*, India, 2015.

[11] Y. Tseng, Z. P. Ho, K. S. Yeng and C. C. Chen, "Mining term networks from text collections for crime investigation," *Expert systems with applications,* vol. 39, no. 11, pp. 10082-10090, 2012.

[12] C. W. Tsai, F. C. Lai, C. H. Chao and A. V. Vasilakos, "Big Data Analytics: a survey," *Springer,* pp. 1-32, 2015.

[13] C. W. Tsai, C. F. Lai, H. C. Chao and A. V. Vasilakos, "Big Data analytics: a survey," *Journalof big data, a springer open journal,* pp. 1-32, 2015.

[14] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff and R. Murthy, "Hive: a warehousing solution over a map-reduce framework," *Proceedings of the VLDB Endowment,* vol. 2, no. 2, pp. 1626-1629, August 2009.

[15] M. Starno, "M. A neural network applied to criminal psychologica lprofiling: an Italian initiative.," *Int J Offender Ther Comp Criminol,* vol. 48, no. 4, pp. 495-503, 2004.

[16] S. Sivaraman and D. R. Manickachezian, "High performance and fault tolerant distributed file system for big data storage and processing using hadoop," in *International Conference on Intelligent Computing Applications*, 2014.

[17] E. Sivaraman and D. Manichachezian, "High performance and fault tolerant distributed file system for big data storage andprocessing using hadoop," in *International Conference on Intelligent Computing Applications*, Coimbatore, India, 2014.

[18] K. Singh and R. Kaur, "Hadoop: Addressing challanges of big data," *IEEE,* pp. 686-689, 2014.

[19] K. Singh and R. Kaur, "Hadoop: Addressing challanges of Big Data," in *International Advance computing conference IAAC*, India, 2014.

[20] j. Schreoder, J. Xu, H. chen and m. Chau, "Automated criminal link analysis based on domain knowledge," *Journal of the association for Information scienceand Technology,* vol. 58, no. 6, pp. 842-855, 21 February 2007.

[21] D. Schmidt, W. C. Chen and G. Ostrouchov, "Introducing a New Client/Server Framework for Big Data Analytics with the R Language," in *XSEDE16*, Miami, USA, 12-21 July, 2016.

[22] M. M. Rathore, A. Ahmad, A. Paul and A. Daniel, "Hadoop based real time big data architecture for remote sensing earth observatory system," in *6th ICCCNT*, Denton, USA, july 13-15 2015.

[23] V. Rajaraman, "Big data analytics," *Resonance,* pp. 695-7166, 2016.

[24] P. J. A. Patel and D. P. Sharma, "Big data/or Better Health Planning," in *IEEE International Conference on Advances in Engineering & Technology Research (ICAETR)* , Unnao, India , August 01-02, 2014.

[25] A. Pal, P. Agrawal, K. Jain and S. Aggarwal, "A performance analysis of map reduce task with large number of files dataset in big data using hadoop," in *Fourth International Conference on Communication Systems*

and Network Technologies, 2014.

[26] A. Pal, K. Jain, P. Agrawal and S. Agrawal, "A performance analysis of map reduce task with large number of file dataset inbig data using hadoop," in *Fourth International confreence on Communication systems and Network Technologies*, India, 2014.

[27] A. Mukherjee, J. Datta, R. Jorapur, R. Singhvi, S. Haloi and W. Akram, "Shared disk big data analytics with apache hadoop," *IEEE,* p. 26, 2012.

[28] R. Mennour and M. Batouche , "Drug discovery for breast cancer based on big data analytics techniques," in *5th internationalconference on Information and Communication Technology and Accessibility*, Marrakech, Morocco, 21-23 December,2015.

[29] D. T. Mahmood and U. Afzal, "Security Analytics: Big Data Analytics for Cybersecurity," in *IEEE 2nd National Conference on Information Assurance (NCIA)*, Rawalpindi, Pakistan, 2013.

[30] Y. Lu, X. Lou, M. Polgar and Y. Cao, "Social network analysis," *Journal of Computer information systems,* vol. 51, no. 2, pp. 31-41, 2010.

[31] J. Lluis and B. Garcia, "A Quick View on Current Techniques and Machine Learning Algorithms for Big Data Analytics," in *IEEE ICTON 2016*, Europe, 206.

[32] J. W. Lee, J. S. jeong, M. Kim and K. H. Yoo, "Safe-Return-Home Service based on Big Data Analytics," in *ACM Proceedings of the 2015 International Conference on Big Data Applications and Services*, Jeju Island, Republic of Korea, October 20 - 23, 2015.

[33] A. Lakshman and P. Mlaik, "Cassandra - A Decentralized Structured Storage System," ACM sigops, 2010.

[34] A. Jain and V. Bhatnagar, "Crime Data Analysis Using Pig with Hadoop," in *International Conference on Information Security & Privacy (ICISP2015)*, Nagpur, INDIA, 11-12 December 2015.

[35] P. Hunt, M. Konar, F. P. Junqueira and b. Reed, "ZooKeeper: Wait-free coordination for Internet-scale systems," in *USENIX annual Technical Conference,*, 2010.

[36] S. A. Hossain, "Big Data Analytics in Education: Prospects and Challenges," in *4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)*, Noida, India, 2-4 Sept. 2015.

[37] G. G. Helmer, J. Wong, V. Honavar and L. Miller, "Intelligent agents for intrusion detection," in *Intelligent agents for intrusion detection IEEE*, Syracuse, NY, USA, USA, 1998.

[38] M. Ghesmoune, H. Azzag, S. Benbernou, M. Lebbah, T. Duong and M. Ouziri, "Big Data: from collection to visualization," *Springer,* pp. 1-26, 2017.

[39] M. Ghesmoune, H. Azzag, S. Benbernou, M. Lebbah, T. Duong and M. Ouziri, "Big data: from collection to visualization," *Springer,* pp. 1-26, 2017.

[40] S. Gao and D. Xu, "Conceptual modeling and development of an intelligent agent-assisted decision support system for anti-money laundering," *Expert Systems with Applications,* vol. 36, no. 2, pp. 1493-1504, March 2009.

[41] Y. Gahi, M. Guennoun and H. T. Mouftah, "Big Data Analytics: Security and Privacy Challenges," in *IEEE Symposium on Computers and Communication (ISCC)*, Messina, Italy, 27-30 June 2016.

[42] C. M. Fuller, D. P. Biros and D. Delen, "An investigation of data and text mining methods for real world deception detection," *Expert Systems with Applications,* vol. 38, no. 7, pp. 8392-8398, July 2011.

[43] A. Floratou, J. M. Patel, E. J. Shekita and S. Tata, "Column-oriented storage techniques for MapReduce," *Proceedings of the VLDB Endowment,* vol. 4, no. 7, pp. 419-429, April 2011.

[44] A. Fernandez, S. D. Rio, V. Lopez, A. Bawakid, M. J. Jesus, J. M. Benitez and F. Herrera, "Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks," *data mining and knowledge discovery,* vol. 4, no. 5, pp. 380-409, 29 September 2014.

[45] S. Early, "Really, Really Big Data NASA at the Forefront of Analytics," *IT Professional,* vol. 18, no. 1, pp. 58-61, 2016.

[46] P. Dhaka and R. Johri, "Big Data Application: Study and Archival of Mental Health Data, using MongoDB," in *International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, Chennai, India, 3-5 March,2016.

[47] E. Dede, M. Govindaraju, R. S. Canon and L. Ramakrishnan, "Performance Evaluation of a MongoDB and Hadoop Platform for Scientific Data Analysis," in *Proceedings of the 4th ACM workshop on Scientific cloud computing*, New York, USA, June 17 - 17, 2013.

[48] K. Dahbur and T. Muscarello, "Classification system for serial criminalpatterns," *Artificial intelligence and law,* vol. 11, no. 4, pp. 251-269, December 2003.

[49] C. Chibelushi, . B. Sharp and H. Shah, "Digital Crime Forensic Sci Cyberspace," in *A Crime Text Mining Approach*, 2006, p. 20.

[50] M. Chau, J. J. Xu and H. Chen, "Extracting meaningful entities," in *national conference on Digital government research*, Los Angeles, California, USA, 2002.

[51] L. Cen, D. Ruta and J. Ng, "Big Education: Opportunities for Big Data Analytics," in *IEEE InternationalConference on Digital Signal Processing*, Singapore, 21-24 July 2015.

[52] D. Borthakur, "HDFS Architecture Guide," The Apache Software Foundation, 2008.

[53] G. Bordogna, L. Fringerio, A. Cuzzocrea and G. Psaila, "Clustering Geo-Tagged Tweets for Advanced Big Data Analytics," in *IEEE International Congress on Big Data*, Italy, 2016.

[54] M. Bendre, R. Thool and V. Thool, "Big Data in Precision Agriculture : Weather Forecasting for Future Farming," in *1st International Conference on Next Generation Computing Technologies (NGCT)*, Dehradun, India, 4-5 September, 2015.

[55] F. Armour, S. Kaisler and A. Espinosa, "Introduction to Big Data Analytics: Concepts, Methods, Techniques and Applications," in *IEEE 48th Hawaii International Conference on System Sciences*, Hawaii, 2015.

[56] R. Al-Zaidy, B. C. M. Fung and A. M. Youssef, "Towards Discovering Criminal Communities," in *Proceedings of the 2011 ACM Symposium on Applied Computing*, New York, USA, 2011.

[57] A. U. Abdullahi, R. Ahmad and N. M. Zakaria, "Big data:Performance profiling of metrological and ocenographic data on HIVE.," in *IEEE 3rd International Conference On Computer And Information Sciences (ICCOINS)*, India, 2016.

[58] I. "Hadoop as a service," IBM Corporation, Somers, New York,USA, 2015.