Big Data Analysis: A Review

Sanyam Sareen
Department of Computer Engineering and Technology
Guru Nanak Dev University
Amritsar, India 143040

Shivangi Ahuja
Department of Computer Engineering and Technology
Guru Nanak Dev University
Amritsar, India 143040

ABSTRACT
The term Big Data accounts for analysis of already procured heterogeneous, structured, unstructured data to find connections between already existing links and predicting the future ones. Big Data finds its use in almost all aspects of society including healthcare, mining, telecom industries etc. It aims at quicker computation of all the humongous data collected from various sources. Big Data and decision making are concomitant so it is influencing IT sectors in present days too. Because Big Data is dependent upon the storage capacity, confidentiality and data complexity come as big loop holes. The sources of this mammoth volume of data include digital pictures and videos, online transactions, GPS signals, sensors etc. Currently hadoop handles big data change but the rate at which the data is increasing new techno logical developments need to made to buttress the already existing system.

Keywords
Hadoop, Big Data, analytics, Hadoop Ecosystem

1. INTRODUCTION
Big data has quickly developed into a hotspot in recent years, getting great consideration from industries, academia and even governments all over the world [1]. More or less, each part of the society is being steered by big data whether it be finance, automobile, healthcare or education. The governments and businesses are all gathering lots of data these days from movies, images, transactions etc. The American multinational retail corporation, Walmart recently came up with a retail link, a gizmo that presents its merchants with a view of the demand in its stores so merchants know exactly when it is time to refill the stores rather than waiting for an order from Walmart stores [2]. The data is incredibly valuable since analyzing it let us do things like detect fraud, consider enhanced data-driven choices while maintaining flexibility and swiftness [3]. Since, the data is so abundant, all of it wouldn't fit any more on a single processor or a single disk. So, there is an urgent need to circulate it across thousands of nodes. The main idea behind big data technology is that if the data is distributed and they are made to run in parallel, computations can be made much quicker and things could be done at a faster rate which couldn't possibly be done earlier. How big is big data? In the last two years, over 90 percent of data has been created [4]. In just 30 minutes, a jet engine can produce 10 terabytes of massive data [5]. The 2009 science fiction film Avatar produced nearly 1 petabyte of data for graphic rendering purposes [6]. Big data deals with the dataset in an old-fashioned technique by using the database approach. When it comes for business dealings, some eminent skills are practiced. The three key forms of data can be either structured or semi structured or unstructured. The one which is structured, is formatted type and hence, can be used directly. The unstructured data like the one from social media, is basically unformatted [7]. The semi structured data can be either former or latter type. The top firms like Google, Facebook, Amazon.com etc. engender tons of data from the very beginning. Amidst of statistical data analysis, a data scientist might want to calculate means, averages, correlations and all sorts of other operations. For example, an analyst might want to look at unemployment versus population versus income versus states. If he has all the data in hadoop, he is able to do it. Big data's power can also be experienced with machine learning and all sorts of other analysis. Once you've got the data in hadoop, there is almost no limit to what you can do. Hadoop is not just a single product or platform, it's a very rich ecosystem of tools, technologies and platforms almost all of which are open source and work together [8].

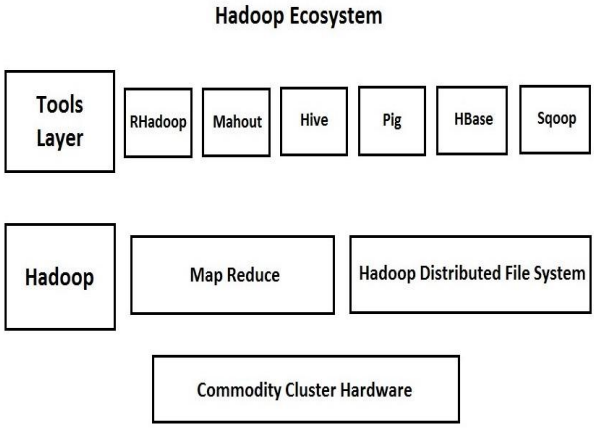

Figure 1: Hadoop ecosystem

The figure 1 describes the hadoop ecosystem. At the lowest level is the commodity hardware and software which runs on almost every operating system hence, eliminating the need to buy any special hardware. On top of that is the hadoop layer which consists of map reduce and hadoop distributed file system (HDFS). The top layer consists of the tools and utilities like RHadoop, Mahout, hive, pig, hbase, sqoop. The RHadoop is basically the statistical data processing using the R programming language. Mahout is the machine learning tool. The tools like hive and pig are NOSQL tools. Finally, sqoop is used for getting data in and out of the hadoop.

Big data is, however, not perfect. It does face ample of threats. As big data relies on widespread storing capability and the data is growing in bulk, it is not feasible for the existing data managing systems to fulfill the data requests. Moreover, the contrasting ability of big data makes it incapable of storing due to the prevailing algorithms. Confidentiality is the chief concern in outsourced data. There are data mining firms which look into an individual's data and process them for their own good without any authorization.

2. LITERATURE REVIEW

Since, the online databank is pretty diverse in nature, It makes it tough for the data scientists to dig up and process meaningful insights from the data which itself is a challenge. The data can vary in terms of data types, metadata information and the whereabouts of the database collection. The prominent hurdles in big data are none other than incompleteness and heterogeneity [9].

The conventional method of using a relational database management system (RDBMS) for the purpose of data computation is not of much assistance since the data is enormous. By considering this fact, RDBMS has agreed to add additional memory to the database. Other obstacles in big data as focused in literature tend to be data complexity, computational complexity and system complexity [10]. The 5 Vs – volume, veracity, variety, value and velocity are actually stated to refer to the shortcomings of the relational database management system (RDBMS) [11]. An overview of the 5 V dimensions of big data is presented in the following table.

Table 1. Big data dimensions

1. Volume	The quantity of data denotes volume which is growing exponentially from gigabytes to petabytes day after day.
2. Veracity	Veracity represents the range to which the data is ambiguous.
3. Variety	The variety of data discusses whether it is structured, semi structured or unstructured which may further be examined to produce significant insights.
4. Value	Working with Big Data is worthless unless humans are able to convert it into value.
5. Velocity	Velocity refers to the rapidity at which fresh data is created.

The various tools and platforms for analyzing large data sets include hadoop, map reduce, apache storm etc.

1. Hadoop: An open-source framework powered by Apache community – hadoopallowing distributed processing of huge data-sets across bundles of computers. Hadoop remains at the epicenter of a group of open source tasks, comprising of tools for machine learning, data organization, data gathering and data storage.

2. Map Reduce: Map reduce is a programming practice for dealing with parallelizable difficulties across huge datasets.Various programming languages including C#, R, python, ruby are consumed to create several map reduce libraries. Apache hadoop is an execution of none other than map reduce.

3. Apache Storm: Apache storm is a distributed real-time computing system, written in clojure and java programming, for processing rapidly enormous set of real-time data.

A company's economy is influenced by big data analytics [12]. Numerous models of big data analytics have been proposed in the literature as following -

1. Co-creation processes had been transformed into collaborative possessions [13].
2. Functioning of purchase order sizing processes can be computed [14].
3. Big data adjacent to visualization can be consumed to examine the competitiveness of submarkets [15].
4. The success in the e-commerce field can be boosted by joining and put into operation the big data and classical management models for the prescriptive upshot [16].

Big data has its impact on hospitality industry too. The web traffic data, search query and hotel habitation demand can be anticipated [17]. Whether mobile internet is available or not, the visitor flows to tourism destinations can be anticipated [18]. A customized recommendation arrangement on travel pack can be crafted [19]. Even a tourism advertising knowledge network can be established for analyzing the reviews [20].

3. APPLICATIONS OF BIG DATA

3.1 Big Data Analytics Applications:

There are a new type of software applications that have come up in the market, that analyze voluminous amount of data by using huge parallel processing frameworks like hadoop. Such applications are developed using a small bit of data in a pseudo-cloud environment. After that the applications are deployed in large-scale cloud environment with even more processing potential and larger input information.

Big data analytics applications are basically a new type of software applications that make maximum use of large-scale data that is too large to suitably fit in memory or even on hard drive [21]. Big data can have a wide variety of sources like the runtime statistics about traffic, tweets during some international events like the Olympic games, stock market updates, usage information of an online game [22], or information from any other fluctuating data-intensive software system.

3.2 Clustering: The users would be able to automatically find the groups within the stored data based on some specific data dimensions through mere point and click dialog using k-means algorithm (clustering). Clustering eases it to identify and groups based on customer type, surfing behavior, shopping patterns, medical records etc [23].

3.3 Data Mining: The big data processing framework for data mining includes three tiers, considerations on data accessing and computing (tier 1), data privacy and domain knowledge (tier 2), and big data mining algorithms (tier 3).

3.3.1 Tier1: Mining Platform of Big Data

For the mining procedures, data mining systems require intensive units to do analysis of data and comparisons. Thus the mining platform should have sufficient access to, data and computer resources.

3.3.2 Tier2: Semantics of Big Data and Application Knowledge

In big data semantics and application knowledge basically

refer to the number of aspects that are related to rules and regulations, the policies, knowledge of the user, and the information of the domain. Significant most issues of this tier are, data sharing and privacy; and domain application knowledge.

3.3.3 Tier3: Mining Algorithms

Big data applications have independent sources and decentralized controls so collecting them to a centralized space for mining procedure because of the huge transmission cost and private data concerns. Other option could be carrying out mining process at the decentralized sites only but it may lead to biased decisions and models. For this a big data mining system should be able to enable information exchange fusion mechanism so that all the distributed sites work unanimously to achieve a common goal. There are two steps to ensure this: model mining and correlations [24].

3.4 Healthcare: The volume of healthcare data is expected to increase drastically in the upcoming years [25]. There could be a wide range of benefits of big data in healthcare area including detection of diseases at an early stage, easy detection of healthcare frauds, and effective management of individual and population health. Certain verdicts can be predicted more efficiently based on the large amount of historical data available, like the patients who can undergo surgery, patients who are at a risk of medical complications, patients who are at a risk of hospital-acquired diseases [26].

3.5 Telecom: The service providers are competing in the cutthroat world of telecom services. As the number of subscribers are increasing, the providers are focusing on increasing the revenue, reduce the opex, chum and enhancing the customer experience, here is the point where big data comes into play [23].

4. FUTURE SCOPE OF BIG DATA

Every day, 2.5 quintillion bytes of data is created – so much that the 90% of the data today has been created in the last two years only [27]. The data is collected from everywhere: digital pictures and videos, purchase transaction records, and cell phone GPS signals, sensors used to gather climate information, posts to social media sites [27] such colossal amount of data can be referred as big data. The current techniques are becoming obsolete, due exponentially increase of data. Comprehensive coding skills, domain knowledge and statistics are required to deal with the big data.

Today, like few technologies have done before, big data is also influencing IT industries. Sensor-enabled machines, mobile devices, cloud computing, social media, satellites generates massive data that help different organizations improve their decision making and take their business to another level. Data is the biggest thing to hit the industry. Every day data is generated in such a rapid manner that, traditional database and other data storing system will gradually give up in storing, retrieving, and finding relationships among data. Through the use of commodity hardware and distribution, big data technologies have addressed the problems related to this new big data revolution.

Due to increasing demand, all companies are exploring big data strategies. The problem is lack of company's internal expertise and best practices. Wikibon's Jeff Kelly says, it's a perfect storm of product and services [28]. A Google cloud platform is being launched by Google, which will provide developers to develop a wide variety of products from simple websites to complex applications. It will give users the privilege of launching virtual machines, storing huge amount of data online, and many other things [29]. Where big data plays an immense role in data processing, all these required huge amount of data processing.

The predictions from the Future Scope of Big Data are:

1. From the rest of the business intelligence market, visual data discovery tools will be growing 2.5 times faster. Soon, investing in this enabler of end-user self-service will become a requirement for all enterprises [30].
2. There will be shortage of skilled staff. In the U.S., there will be five times that many positions requiring related skills in data management and interpretation and 181,000 deep analytics roles in 2018 [30].
3. Cloud-based big data and analytics (BDA) solutions will grow three times faster than spending for on-premise solutions, over the next five years [30].
4. The unified data platform architecture, by 2019, will become the foundation of BDA strategy. Across information management, analysis, and search technology, the unification will occur [30].
5. In 2018, growth in applications incorporating advanced and predictive analytics, including machine learning, will accelerate. These apps will grow 65% faster than apps without predictive functionality [30].
6. Through 2019, decision management platforms will expand at a CAGR of 60% in response to the need for greater consistency in decision - making and decision - making process knowledge retention [30].
7. External data has already been purchased by 70% of large organizations and 100% will do by 2019. By selling data or providing value-added content, more organizations will begin to monetize their data [30].
8. In 2018, adoption of technology to continuously analyze streams of events will accelerate as it is applied to Internet of Things (IoT) analytics, which is expected to grow at a five-year compound annual growth rate (CAGR) of 30% [30].
9. In 2018, rich media (video, audio, image) analytics will at least triple and for BDA technology investment it will emerge as a key driver [30].
10. Half of all consumers will interact with services based on cognitive computing on a regular basis by 2018 [30].

At present, big data is being handled by the software named hadoop. Extensive research needs to be carried out and revolutionary technologies need to be developed, to harness the potential of big data completely in the future.

As people digitize their lives, big data has reached critical mass. The project manager at NASA within the Human Adaptation and Countermeasures Division of the Space Life Sciences Directorate Nicholas Skytland said, "People are walking sensors."

It can be concluded by taking an average of all the figures suggested by leading big data market analyst and research firms that approximately 15 percent of all IT organizations will move to cloud-based service platforms, and between 2018 and 2021, this service market is expected to grow about 35 percent [31].

5. CONCLUSION

Every day the data is created in quintillion bytes. The 90% of the data has been created in the last two years .The data is so abundant; all of it wouldn't fit on a single processor or a single

disk. The governments and businesses are accumulating a lot of data nowadays from movies, images, transactions etc. The data is very valuable since researchers analyze things like detect fraud, consider enhanced data-driven choices to maintain flexibility and swiftness.

Big Data is used for this purpose for storing this abundant data. Each part of the society is being run by big data whether it is finance, automobile, healthcare or education. There are three forms of data structured, which is formatted type and is used directly or unstructured, which is unformatted like social media or semi structured. The main idea behind big data technology is that if the distribution of data and running it parallel can make computations quicker and things can be completed at faster rate. Big data deals with the dataset in an old-fashioned technique by using the database approach. The power of Big Data can be seen with machine learning and all the analysis.

The various tools and platforms for analyzing large data sets include hadoop, map reduce, apache storm etc. Hadoop is an open-source framework powered by Apache community – hadoop allows distributed processing of large data-sets. Hadoop is an epicenter of a group of open source tasks, comprising of tools for machine learning, data organization, data gathering and data storage. The hadoop ecosystem contains the lowest level as commodity hardware and software which runs on almost every operating system, the hadoop layer which consists of map reduce and hadoop distributed file system (HDFS) and The top layer consists of the tools and utilities like RHadoop, Mahout, hive, pig, hbase, sqoop.

Big data relies on large storing capability and as the data is increasing a big number, it is not good for the existing data managing systems to fulfill the data requests. The Relational Database Management System (RDBMS) is the traditional method for data computation is n more suitable. The shortcomings of the relational database management system (RDBMS) are the 5 Vs – volume, veracity, variety, value and velocity. Map reduce is a programming practice for dealing with parallelizable difficulties across large data. to create several map reduce libraries programming languages like C#, R, python, ruby are used. Apache hadoop is an execution of none other than map reduce. Applications of big data are in BDA(Big Data Analytics Applications),clustering, data mining, healthcare and telecom.

The current techniques are becoming weak, due to continuously increase in data. Better coding skills, proper knowledge and statistics are required to work with the big data. Through the use of commodity hardware and distribution, big data technologies have addressed the problems related to this new big data revolution. As the demands are increasing, companies are trying to explore big data strategies. At present, big data is being handled by the software named hadoop. Extensive research needs to be carried out and revolutionary technologies need to be developed, to harness the potential of big data completely in the future.

6. REFERENCES

[1] Huang, Y., Porter, A. L., Cunningham, S. W., Robinson, D. K., Liu, J., & Zhu, D. (2017). A technology delivery system for characterizing the supply side of technology emergence: Illustrated for Big Data & Analytics. *Technological Forecasting and Social Change*.

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

[3] Sushil (2017). Multi-criteria valuation of flexibility initiatives using integrated TISM – IRP with a big data framework. *Production Planning & Control*

[4] Science Daily, https://www.sciencedaily.com/releases/2013/05/130522085217.htm. Big Data, for better or worse: 90% of world's data generated over last two years. Accessed on: 23/06/2018

[5] Zaslavsky, A., Perera, C., Georgakopoulos, D. (2013). Sensing as a service and big data. Retrieved from https://arxiv.org/abs/1301.0159.fAccessedfon:f23/06/2018.

[6] ThegNextgWeb, https://thenextweb.com/space/2010/01/01/avatar-takes-1-petabyte-storage-space-equivalent-32-year-long-mp3/ Believe it or not: Avatar takes 1 petabyte of storage space, equivalent to a 32 YEAR long MP3. Accessed on: 24/06/2018

[7] Abbass , H. A., Leu, G., and Merrick, K. (2016). A Review of Theoretical and Practical Challenges of Trusted Autonomy in Big Data. Theoretical Foundations for Big Data Applications: Challenges and Opportunities.

[8] Bhosale, Harshawardhan S., and Devendra P. Gadekar. "A review paper on Big Data and Hadoop." International Journal of Scientific and Research Publications 4, no. 10 (2014): 1-7

[9] Ammu, N., & Irfanuddin, M. (2013). Big data challenges. International Journal of Advanced Trends in Computer Science and Engineering, 2(1), 613-615.

[10] Jin, J., Liu, Y., Ji, P., & Liu, H. (2016). Understanding big consumer opinion data for market-driven product design. International Journal of Production Research.

[11] Jin, X., Wah, B. W., Cheng, X., & Wang, Y. (2015). Significance and challenges of big data research. Big Data Research.

[12] Rust, R. T., & Huang, M. H. (2014). The service revolution and the transformation of marketing science. Marketing Science.

[13] Xie, K., Wu, Y., Xiao, J., & Hu, Q. (2016). Value co-creation between firms and customers: The role of big data-based cooperative assets. Information & Management.

[14] Bock, S., &Isik, F. (2015). A new two-dimensional performance measure in purchase order sizing. International Journal of Production Research.

[15] France, S. L., & Ghose, S. (2016). An analysis and visualization methodology for identifying and testing market structure. Marketing Science, 35(1), 182-197.

[16] Qi, J., Zhang, Z., Jeon, S., & Zhou, Y. (2016). Mining customer requirements from online reviews: A product

improvement perspective. Information & Management.

[17] Yang, Y., Pan, B., & Song, H. (2014). Predicting hotel demand using destination marketing organization's web traffic data. Journal of Travel Research.

[18] Raun, J., Ahas, R., &Tiru, M. (2016). Measuring tourism destinations using mobile tracking data. Tourism Management.

[19] He, J., Liu, H., &Xiong, H. (2016). SocoTraveler: Travel-package recommendations leveraging social influence of different relationship types. Information & Management.

[20] Dolnicar, S., & Ring, A. (2014). Tourism marketing research: Past, present and future. Annals of Tourism Research.

[21] Fisher, D., DeLine, R., Czerwinski, M., &Drucker, S. (2012). Interactions with big data analytics.

[22] Wingfield, N. Virtual product, real profits: Players spend on zynga's games, but quality turns some off. Wall Street Journal

[23] Kuchipudi, S., &Reddy, T. (2015). Application of Big data in Various Fields.

[24] Xindong, W., Xingquan, Z., Gong-Qing, W., &Wei, D. (2014). Data Mining with Big Data.

[25] IgHealthgTran, http://ihealthtran.com/wordpress/2013/03/iht%C2%B2-releases-big-data-research-report-download-today/ Healthcare data. Accessed on: 26/06/2018

[26] Raghupathi, W., &Raghupathi,V. (2014). Big data analytics in healthcare: promise and potential.

[27] Apache Hive, http://hive.apache.org Accessed on: ggggg23/06/2018

[28] Mukherjee, S., &Shaw R. (2016). Big Data - Concepts, Applications, Challenges and Future Scope.

[29] Wikipedia, https://en.wikipedia.org/wiki/Google_Cloud_Platform Google Cloud Platform. Accessed on: 25/06/2018

[30] Idc, http://www.idc.com/getdoc.jsp?containerId=prUS25329114 Big Data Scope. Accessed on: 25/06/2018

[31] Shahriar, A., &Samuel, F**.(**2016**) .**Big data analytics in E-commerce: a systematic review and agenda for future research.