A Review on Big Data: Views, Categories and Aspects

Diwakar Shukla
HOD
Department of Copmuter Science and Applications
Dr. H.S Gour Vishwavidyalaya, Sagar, MP, India

Abdul Alim
Research Scholar
Department of Computer Science and Applications
Dr. H.S Gour Vishwavidyalaya, Sagar, MP, India

ABSTRACT
Now-a-days every organization is moving towards on web based application and cope up lots of data sets. Data may be structured, unstructured or semi-structure. These data need processing, analysis and storage in proper format using innovative techniques and methodologies. Big data parameterized into three basic categories Volume, Variety and Velocity. The social media like Facebook, Whatsapp, Twitter, hike. are generating lots of data in a day in the form of text-messages, audio-recording, images, videos etc. The problem which appears is how to manage this huge data in a systematic way because users want quick response on web search or on smart phone access using web apps. In this paper we have studied how big data invoke in different areas like social media, atmosphere, hospitals, research centers etc. We have suggested contributions in big data categories, applications in different area's like in Machine learning, Social Network, Bio-Informatics, Data Mining, and Clouds along with challenges, future scope and storage prospects.

Keywords
Big Data, Map Reduce and Hadoop, Storage, Optimization, 3Vs, Machine Learning, Cloud, Bio Informatics, Social Networking

1. INTRODUCTION
In 1965, the US Government had planned and implemented the world's first data center to store 742 million tax returns and 175 million sets of fingerprints on magnetic tape. It was a motivation which generated, in due course the idea of term big data in visually exploring Gigabyte datasets in real time. In 1991, the thought expanded and published by the association for computing machinery. *Peter Lyman and Hal* Varian (2000) attempted to quantify the amount of digital information in the world and evaluating rate of growth for the first time. The world's total yearly production of print, video, optical, voice and magnetic content would require roughly 1.5 million gigabytes of storing capacity. Moreover , around 5.5 million people of world are in touch of social sites, in deep practice of upload, download and share own data with friends. Up to the end of year 2010 , it was found that more and more people are using mobile devices to access digital data, than usual office or home computers. In continuation, 80% of business executives working in a company reported in a survey that big data analytics is a top priority issues for decision making their business [1]. Few years ago, a renowned group of business noticed this fact and proposed 3Vs of big data. After that, another renowned company pleaded for adding 4th Vs which was accepted by the most of scientist [2]. Figure 1 describe the same in diagrammatic form.

Fig 1: 4Vs in big data (Source: [2])

Big data out fitted the capability of traditional data processing technologies. It differentiated to traditional technologies in three ways: (a) the amount of data (Volume) ,(b) the rate of data generation (Velocity),(c) and the types of structured and unstructured data (Variety) (see [3]) .

In society, people constantly interact with each other causing the rapid development of social computing over time and space. An average Internet user consumes large amount of digital content every day through social sites such as Facebook, Twitter, YouTube, Instagram and on many other similar chat applications [4]. The Cisco Internet business Solution Group (IBSG) predicted that 25 billion devices would be connected to the Internet by 2015 which would reach to 50 billion by 2020. Exchange of such large amounts of data referred to as big data. The traditional distributed system of machines, and traditionally created databases are not able to capture effectively storage part, not able to manage and analyze this data. Old tools and techniques have limited scalability [5].

Big data derive radical changes too in traditional data analytics platform. To perform any kind of conclusive analysis on such voluminous and complex data, supportive hardware platforms become outfitted and choosing appropriate hardware , software platforms becomes a crucial decision , especially when user's requirements are to be satisfied in a reasonable amount of time. When users need to decide the right platforms to chose from, they have to investigate what their application or algorithm needs are ? Fundamental issues are how quickly do a person get the required results? How big the data to be processed? How does the model building require to cope up iterations [6].?

To extract value from big data, special tools and techniques are needed. New algorithms, scalable and high performance processing infrastructure, analytics skills need to developed to support . However, the big data creates new challenges for the validation and verification task like data selection and validation are critical to the effectiveness and performance of

such data analysis, but large volume and varieties of big data big data create a grand challenge for the selection and validation [7].

2. BIG DATA AND CLOUD COMPUTING

Targio Hashem et al. (2014) explored that addressing big data is a challenging, time demanding task that requires a large computational infrastructure to ensure successful data processing and analysis. The rise of same in cloud computing was reviewed by the many authors with definition, characteristics, and classification of big data along with fruitful discussion on cloud computing. The relationship between big data storage systems and Hadoop technology is the modern approach for dealing with such issue. A figure below describes the classification scheme with inner content [8].

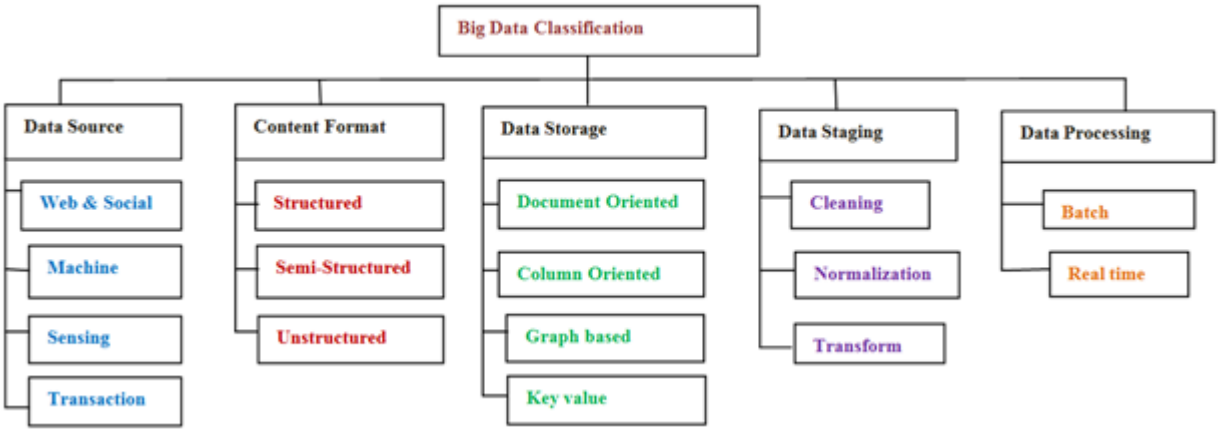

Fig 2: big data classification (Source: [8])

2.1 Relationship between Cloud Computing and Big Data

Big data provides users the ability to use commodity computing to process distributed queries across multiple data sets and return to resultant sets in a timely manner. On other hand, cloud computing provides the underlying engine through the use of Hadoop technology, which is a class of distributed data processing platforms [8].It appeals that cloud computing and big data are inter-related.

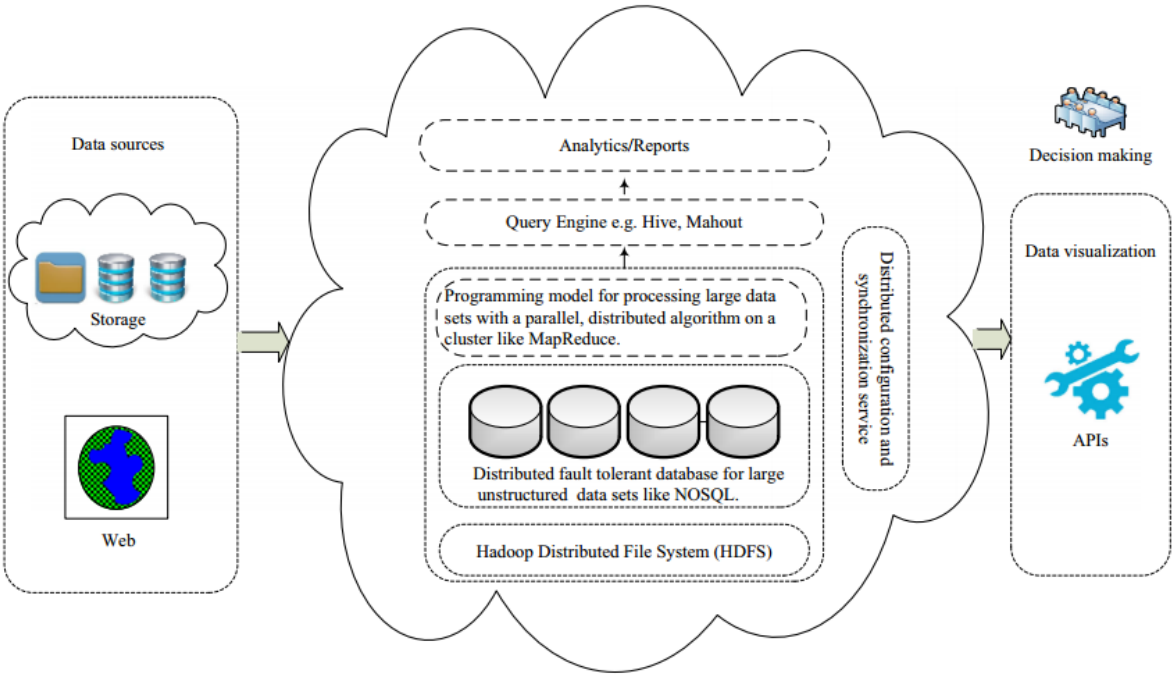

Fig 3: Relationship between cloud computing and big data (Source: [8])

Inukollu et al. (2014) discussed security issues for cloud computing, in big data, Map Reduce approach under Hadoop environment. Authors placed various possible solutions in view of such data for problems in cloud computing security [9]. Figure 3 explains the relationships.

3. BIG DATA AND DOCKER CONTAINER

In entire world, at present the decision system being driven by data with more and more information about indivisuals, companies and goverments than ever before. Every business is powered by information and communication technology generating big data set up. Future business Intellegance, prospects and forecast can be extracted from big data. Tools like NoSQL and MapReduce technologies can find an efficient way to store, orgnize and process the big data with the help of Virtualization and Linux container technologies.

Provisioning containers on top of virtual machine is a better model for resource utilization. As most of containers share the same CPU, the context switch latency for each container increase singnificantly. Such increase leads to a negative impact on the network I/O throughput and creates a bottlenack in big data environments.

Some authors studied container networking in view to factors of context switch latency. They evaluated Hadoop benchmarks against the number of containers, virtual machines and observed a bottleneck where Hadoop cluster throughput is non linear over number of nodes sharing the same CPU. This bottleneck is due to virtual network layers which adds a significant delay to Round Trip Time (RTT) of data packets. Present idea could be extended to analize the practical implications of virtual network layers and supose to lead lead to a solution to improve upon the performance of big data environments in view to containers.

Virtualization technology become the facto standard for all public and private cloud requirements. It has consolidated all the hardware components creating software redundency layers for elastic work loads. Docker framework introduced manageable Linux containers called Dockers Containers which is an open platform for dovelopers and system administrators to build, ship and run distributed applications. Hadoop is an open source software framework for storing and processing big data in a distributed fashion on large clusters [10].

Zheeg and Thain (2015) explored work flows pattern that are a widely used abstraction for representing large scientific applications and executing on distributed systems such as clusters, clouds, and grids. However, workflow systems have been largely silent on the question of precisely what environment each task in the workflow is expected to run in. As a result, a workflow may run correctly in the environment in which it was designed, when workflow move to another machine, it is more likely to fail due to deference in the operating system, installed applications, available data, and so forth. Lightweight container technology has recently appeared as a potential solution to this problem, by providing well-defined execution environments at the operating system level. Computer Scientist considered how to container technology in best way into an existing workflow system, using Makeflow, Work Queue. A brief performance study of Docker shows very little overhead in CPU and I/O performance, but significant costs in creating and deleting containers. Academic contributors have suggested different methods of connecting containers over different points of the infrastructure, and explained several methods of managing the container images that must be distributed to execute tasks. It was also discussed the performance of a large bioinformatics workload on a Docker-enabled cluster, and distributed systems. It was observed that the best configuration to be locally-managed containers could be shared among multiple tasks [11].

4. BIG DATA MINING AND FORECASTING

Sawant and Desai (2015) discussed as big data defining it as collections of large, complex data which is difficult to process using current methodologies, standard database management practices or analytical solutions. Mining of such is a kind of capability of discovering knowledge and patterns from the large sets of information, streams of data. Mobile phones and Social Medias are major data exhaust producers. Mining this is a challenge or a new frontier as a next step. Authors represented overview, its sources, challenges and opportunities to forecast the future prospects. There are various concepts needs to summarized in the form of articles, research monographs written by scientists, and scholars in the relevant field [12].

Harsoor and Patil (2015) explored viewpoint regarding big data of household products sold by various subsidiaries of the retail store network which are geographically scattered at various locations. Supply chain inefficiencies occur, at different places and locations, when the market potential may not be evaluated by the retailers. Many times, it is not easy for the retailers to understand the market conditions at scattered geographical locations. Organization of retail store network have to understand the market conditions to intensify its goods to be bought and sold so that more and more customers get attracted in that direction.

Business forecast helps retailers to view and exhaustive picture of forecasting the sales and make up mind for general idea of coming years. It changes are needed then those are to be done in the retail store's objective so that success is achieved more profitably .It also helps the customers to be happy by providing the products desired by them in desired time. When customers are happy then they prefer the store resources they need to their satisfaction. By this sale in specific store in which the customers purchase more items causing more profit. The forecasting of sales pattern helps to know the retailers the demand of the product. Big data mining and forecasting together are complementary. Authors have made an attempt of understanding the retail store business driving factors by analyzing the sales data of a renowned store that is geographically located scattered with and the forecast of sales for coming 39 weeks as a solution. By sales forecasting the retail networks are supported so that the resources can be managed efficiently [13, 14]. Fan and Bifet (2013) suggested that big data mining is capability of extracting useful information from these large datasets or streams of data, as it was not possible before to do it [14].

5. BIG DATA AND STORAGE

Liu and Zhou **(**2015) suggested that there is huge demand for data storage. It is because the traditional data storage procedure is based on relational database that cannot meet the needs, many application systems tend to use NoSQL to resolve the problem of big data storage. However, NoSQL is based on the relationship between operation supports, so that part of existing application system is difficult to use in simple way of transformation. Authors have given a big data comprehension in academic contribution compatible with

relational storage model, storage scheme, which can not only meet with the big data storage requirements, but also can support the most of relational operations, so that the original system based on relational database can easily be ported to new storage schemes [15].

Li and Gou (2012) discussed recent escalations in Internet development and problem aspects appeared due to Volume of data that have created a growing demand for high capacity storage solutions. Although Cloud storage yielded new ways of storing, accessing and managing data, there is still a need for an inexpensive, effective and efficient storage solution, especially suited to big data management and analysis. Contribution presented an in- depth analysis of the key features of big data storage services for both unstructured and semi-structured data, and discussed how such services be constructed and deployed. Description incorporates how different technologies can be combined to provide a single, highly scalable, efficient and performance-aware big data storage system. The focus is on issues of data de-duplication for enterprises and private organizations. It is particularly valuable for inexperienced solution providers suggestion them to swiftly set up their own big data storage services [16].

6. BIG DATA AND OPTIMIZATION

Li (2013) pointed out by the survey that, performance optimization is a classic and important topic in cloud computing because appropriate optimization techniques provide better application output even with less system resource consumption, compared to usual cases.

Dataflow based performance analysis a tool for big data cloud. Hitune is shown to be effective in assisting users doing Hadoop performance analysis and system parameter tuning. Limitations of existing approaches, such as Hadoop logs and metrics were also compared and discussed by many authors. Few interesting case studies on big data processing in cloud computing environment were found in literature content. Efforts of the Fijitsu laboratory are based on data storage and complex event processing, as well as workflow description in distributed data processing.

A recent online cost-minimization algorithm was depicted focusing on real time cost minimizations for uploading massive and dynamic data onto the cloud. Two online algorithms have been suggested who achieved competitive cost reduction ratios. However, methods are evaluated in a limited scale. The suggested algorithms need to be further evaluated at larger and more competitive scales like data streaming applications including larger complex topologies [17].

7. BIG DATA AND PRIVACY

Harvais Simo (2015) discussed recent advances in data collection and computational statistics together which increases computer processing power, along with the plunging costs of storage. It leads to technologies to effectively analyze large sets of heterogeneous data ubiquitous. Applying such technologies (often referred to as big data technologies) to an ever growing number and variety of internal and external data sources, businesses and institutions can discover hidden correlations between data items, and extract actionable insights for innovation and economic growth. While on one hand, big data technologies yield strong promises, they raise critical security, privacy, and ethical issues, which if left unaddressed may become significant barriers to the fulfillment of expected opportunities and long-term success of big data. Authors discussed the benefits of big data to individuals and society at large, focusing on seven key use cases like big data for business optimization and customer analytics, big data and science, big data and health care, big data and finance, big data and the emerging energy distribution systems, big/open data as enablers of openness and efficiency in government, and big data in view to security. In addition to benefits and opportunities, the security, privacy, and ethical issues are also of prime importance [18].

*Zeng (*2015) delivered about protection in big data environment which is an effective mode of thinking and working. There are many security risks in data collection, storage and use. Privacy leakage causes serious problems to the users; false data lead to in analysis. Author introduced the security problems faced by big data uses and analyzes the causes of privacy problems along with principle to resolve the problems [19].

In widely noted speech on January 17, 2014, President Barack Obama charged his counselor, John Podesta, with leading a comprehensive review of big data and privacy. The content incorporate views of privacy experts, technologists and business leaders and focused at how the challenges inherent in big data are being confronted by both the public and private sectors, whether it can forge International norms on how to manage this and how we can continue to promote the free flow of information in ways that are consistent with both privacy and security. The President and counselor Podesta asked the President's council of advisor on Science and Technology (PCAST) to assist to technology dimension.

It was resolved that, PCAST will study the technological aspects of the intersection of big data with individual privacy, in relation to both the current state and possible future states. Relevant big data include data and metadata collected, or potentially collectable, from or about individuals by entities that include the government, the private sector and other individuals [20].

8. BIG DATA AND SOCIAL ENVIRONMENT

European commission explored one of the most commonly recognized applications of big data as social media data analysis, to know social network impact on customer behavior. It would be easy to say that big data means social media data, but this assumption would miss to capture both existing applications as well as the potential of this paradigm. The concept is generated from an increasing plurality of sources, including internet clicks, mobile transactions, user-generated content, and social media; moreover, enforced generated content through sensor networks or business transactions like sales queries and purchase is also a massive source. In addition, genomics, health care, engineering, operations management, the industrial internet and finance are all add to the big data pervasiveness.

The huge amount of novel data being generated has important contributions in epidemiology, specifically in temporal public health surveillance. In the era of satellite sensors, a diversity of epidemiologically relevant environmental information can be sourced globally at daily intervals. Big data allows a closer temporal matching of disease outbreaks with covariates which may improve the accuracy of mapping models.

The electronic incorporation among people producing vast quantities of social, psychological, and organizational data that social workers can harness to address society's most difficult problems. The computerized service, education, health records, social media posts, web searches GPS devices,

network of sensors help to illuminate problems. Technological enhancement makes it possible to manage and analyze in real time.

In digital scenario vast quantities of data exist, a large part of which remains un-analyzed for social well-begins, social policy, and social action. A great deal of useful information remains trapped in the silos of legacy information systems. Despite these existing troves of data, the social sector lags behind in data driven strategies. Efforts to make social spending more evidence-based or to make social systems more effective often falter because of the high cost of gathering timely and complete data. [21].

Coulton et al. (2015) discussed that big data technology and services expected to grow worldwide at a compound annual growth rate of 40%. Moreover, in the UK alone, the number of big data staff specialist working in large firms will increase by more than 240% over the next five years. Progress in the IT environment of public sector (availability of broadband and big data tools, cloud services, HPC) will lead to cost reduction of operations, increase of efficiency and personalized services for citizens [22].

9. BIG DATA AND SOCIAL NETWORK

In 2010, further European commission has given exciting description as social networks have drastic growth in recent years. Such provide extremely suitable space to instantly share multimedia information among individuals. This network is powerful reflection of the structure and dynamics of the society and one gets platform for interaction of the Internet generation with both technology and people. Drastic growth of social multimedia and user generated content is revolutionizing all around to content value chain including production, processing, distribution and consumption. It has also originated and brought to the multimedia sector a new underestimated and new critical aspect of science and technology, social interaction and networking. Online content sharing services among communities, multimedia communication over the Internet, social multimedia search, interactive services, entertainment, health-care and security applications have brought change in life style and generated a new research area called social multi- media computing; in which well established computing and multimedia networking technologies clubbed together with emerging social media research.

Social Networking services have converted communication with others to entertain and actually be live. Social Networking is such primary reasons that many people proud to become Internet users. This is robust indicator of what is really happening online. Now-a-days, users produce and consume significant quantities of multimedia content. Moreover, such behavior when combined with Social Networking forms a new Internet era of multimedia content sharing through Social Networking Sites. More than 200 SNSs of worldwide impact are known today and this number is growing fast. Many of the existing top web sites are either pure SNSs or offer some social networking capabilities.

Tan et al. (2013) explored methods of handling large data sets originated from many domains. Deriving knowledge is more difficult than ever before as when we must do it by intricately processing big data. Social network paradigm could enable a level of for solving big data processing challenges. Recent and popular social networking websites such as Twitter, Facebook, LinkedIn, YouTube, and Wikipedia have captured Exabyte of information associated with their daily interactions connected people. Social networking conceptualized for social scientists in the context of human social networks, mathematicians and physicists in the context of complex network computer scientists in the examination of information or Internet-enabled services.

Social Networks as big data understanding has generated a problem when business, management, or information systems specialists hope to predict consumer behavior to enhance marketing, sales, and e-commerce. Many social networking sites have between 10 and 200 million users, therefore sampling too is a central focus to most of studies. Although significantly time- consuming, gaining insight from the entire dataset might provide the most optimal solutions. In terms of volume, also a report at the end of 2011, Facebook had 721 million individuals and 68.7 billion friendship edges (see http://arxiv.org/abs/1111.4503). In terms of velocity, Twitter and Face- book respectively generate 7 tera-bytes and 10 tera-bytes of data daily. These data also need to be processed at the speed of thought [23].

10. CLASSIFICATION OF TECHNOLOGIES OF BIG DATA

Moniruzzaman and Hossain (2013) discussed critical issue storing that digital world is growing very fast and becoming more complex in the volume (terabyte to petabyte), variety (structured and un-structured and hybrid), velocity (high speed in growth) in nature. This is typically considered to be a data collection that has grown so large it cannot be effectively managed or exploited using conventional data management tools: e.g., classic relational database management systems (RDBMS) or conventional search engines. To handle this problem, traditional RDBMS are complemented by specifically designed a rich set of alternative DBMS; such as - NoSQL, NewSQL and Search-based systems. Authors contribution motivates is to provide - classification, characteristics and evaluation of NoSQL databases in big data analytics. It is intended to help users, especially to the organizations to obtain an independent understanding of the strengths and weaknesses of various NoSQL database approaches to supporting applications that process huge volumes of data [24].

Tsai et al. (2015) described through survey how the age of big data is now coming.? But the traditional data analytics may not be able to handle such large quantities of data. The question arises how to develop a high performance platform to efficiently analyze big data and how to design an appropriate mining algorithm to find the useful things from source of big data. To discuss this deeply issue, authors contribution begins with a brief introduction to data analytics, followed by the discussions of big data analytics. Some important open issues and further research directions are in [25].

11. GENERAL COMPUTING AND BIG DATA

Zhou et al. (2014) have discussed that, big data as a term has been among the biggest trends of the last three years, leading to an upsurge of research, as well as industry and government applications. It is deemed a powerful raw material that can impact multidisciplinary research endeavors as well as government and business performance. One of goal to share the data analytics opinions and perspectives relating to the new opportunities and challenges brought forth by the big data movement. Authors bring together diverse perspectives, coming from different geographical locations with different

core research expertise and different affiliations and work experiences. Researchers must evoke for discussion rather than to provide a comprehensive survey of big data research [26].

Najafabadi et al. (2015) presented view point that big data Analytics and Deep Learning are two high-focus of data science. It has become important as many organizations both public and private have been collecting massive amounts of domain-specific information, which can contain useful information about problems such as national intelligence, cyber security, fraud detection, marketing, and medical informatics. Companies such as Google and Microsoft are analyzing large volumes of data for business analysis and decisions, impacting existing and future technology. Deep Learning algorithms extract high-level, complex abstractions as data representations through a hierarchical learning process. Complex abstractions are learnt at a given level based on relatively simpler abstractions formulated in the preceding level in the hierarchy. A key benefit of Deep Learning is the analysis and learning of massive amounts of unsupervised data, making it a valuable tool for big data Analytics where raw data is largely unlabeled and un-categorized.

Many authors explored how Deep Learning can be utilized for addressing some important problems in big data analytics, including extracting complex patterns from massive volumes of data, semantic indexing, data tagging, fast information retrieval and simplifying discriminative tasks. It is to investigate some aspects of Deep Learning research that need further exploration to incorporate specific challenges including streaming data, high-dimensional data, scalability of models, and distributed computing. By presenting insights into relevant future works by posing some questions, including defining data sampling criteria, domain adaptation modeling, defining criteria for obtaining useful data abstractions, improving semantic indexing, semi-supervised learning, and active learning that can optimized goals [27].

12. MACHINE LEARNING AND BIG DATA

Tarwani et al. (2015) have given an idea how big data analytics plays an important role in making sense of the data and exploiting its value. It is a significant challenge to learn and develop new types of machine learning algorithms. Scaling up big data to proper dimensionality is an challenge that can encounter in machine learning algorithms plus there are challenges of dealing with Velocity, Volume and many more for all types of machine learning algorithms. Group of authors have given first exploring big data concept, bringing with an urgent need for advanced data acquisition, management, and analysis mechanisms. Authors have presented the concept of big data and highlighted the four phases of big data that are generating data, acquisition of data, storing this large data, and then analyzing data. Major focus is on dealing with big data using machine learning (ML) and highlighted the three ML methods: supervised learning, unsupervised learning and reinforcement learning and its impact on big data analytics [28].

Kashyap et al. (2015) addressed that Bioinformatics research is characterized by voluminous and incremental datasets and complex data analytics methods. The machine learning methods used in bioinformatics are iterative and parallel. These methods can be scaled to handle big data using the distributed and parallel computing technologies.

Usually big data tools perform computation in batch mode that is not optimized for iterative processing and there remains high data dependency among operations. In recent years, parallel, incremental, and multi-view machine learning algorithms have been proposed. Similarly, graph-based architectures and in-memory big data tools have been developed to minimize I/O cost and to optimize iterative processing.

13. BIO-INFORMATICS AND BIG DATA

However, there is lack of standard big data architectures and tools for many important bioinformatics problems, such as fast construction of co-expression and regulatory networks and salient module identification, detection of complexes over growing protein-protein interaction data, fast analysis of massive DNA, RNA, and protein sequence data, and fast querying on incremental and heterogeneous disease networks. Many contributions addressed the issues and challenges posed by several big data problems in bioinformatics and given an overview of the state of the art and the future research opportunities [29].

Chen and Zhang (2014) discussed that, there are so much potential and highly useful values hidden in the huge volume of data. A new scientific paradigm is born as a data intensive scientific discovery (DISD), relating to big data problems. A large number of fields and sectors, ranging from economic and business activities to public administration, from national security to scientific researches in many areas which involve big data problems. On the other hand, it is extremely valuable to produce productivity in business and evolutionary breakthroughs in scientific disciplines, which give us a lot of opportunities to make progresses in many fields. There are no doubts the future competitions in business productivity and technologies will surely convert in to big data explorations.

On other hand, big data arise with many challenges, such as difficulties in data capture, data storage, data analysis and visualization. Authors aimed to demonstrate a close view about big data, including its applications, opportunities and challenges, as well as the state-of-the-art technologies. The authors also discussed several underlying methodologies to handle the data deluge, for example granular computing, cloud computing, bio-inspired computing and quantum computing [30].

14. CHALLENGES WITH BIG DATA

Kaisler et al. (2013) analyzed that, the volumes exceed the capacity of current on-line storage systems and processing systems. Data, information, and knowledge are being created and collected at a rate that is rapidly approaching the exabyte/year range. But, its creation and aggregation are accelerating and will approach the zettabyte/year range within a few years. The Volume is only one aspect of big data; other attributes are variety, velocity, value, and complexity. Storage and data transport are technology issues which seem to be solvable in the near-term, but represent long- term challenges that require research and new paradigms. Authors analyzed the issues and challenges and suggested collaborative research program into methodologies for big data analysis and design [31].

One useful contribution n suggests about recent big data issue which open several new avenues. Given the unprecedented amount of data that has been produced, collected and stored

likely in the coming years, one of the technology industry's great challenges is how to derive benefit from it. While big data can be definitely perceived as a big blessing, big challenges also arise with large-scale datasets. The sheer volume of data makes it often impossible to run analytics using a central processor and storage and distributed processing with parallelized multi-processors is preferred while the data themselves are stored in the cloud. In addition, as the size of data grows exponentially, current algorithms are not efficient or scalable enough to deal with such large volumes of data. Designing more accurate intelligent models so as to satisfy the market needs hence bring huge opportunities as well as challenges to these communities. [32]. Figure 4 explains the global challenges and table 1 big data.

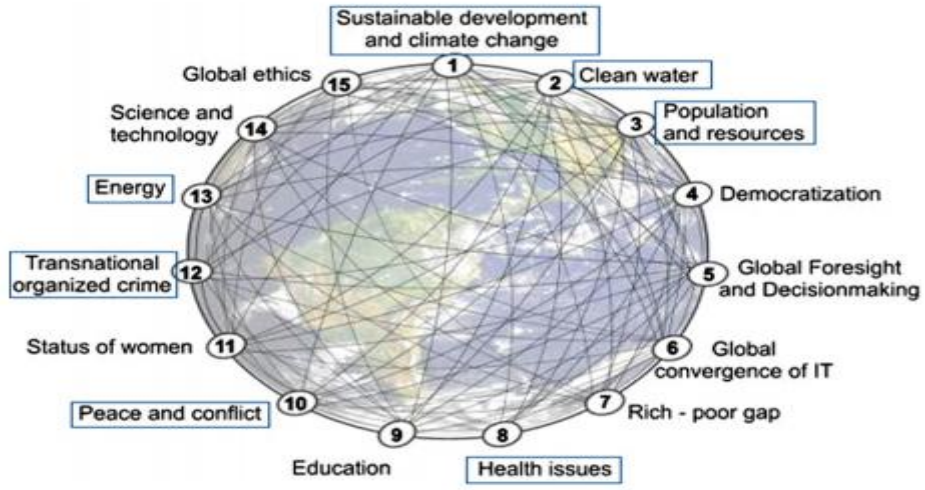

Fig 4: Global challenges facing humanity (Source: [33])

Table 1. Challenges in big data (Source: [30])

Challenges	Description
Data Representation	Heterogeneity in type, structure, semantics, organizations, granularity and accessibility.
Redundancy reduction and data comparison	There is high level redundancy in datasets, most data generated by the sensor networks are highly redundant.
Data life cycle management	A data importance principle related to the analytical value should be developed to decide which data shall be store and which data shall be discarded.
Analytical mechanism	Traditional RDBMSs are strictly designed with a lack of scalability and expandability which could not meet the performance requirements.
Data confidentiality	Big data service providers at present could not effectively maintain and analyze such huge data sets because of their limited capacity. They must rely on professionals or tools to analyze such data which increase the potential safety risk.
Energy management	Processing, storage and transmission of big data will consume more and more electronic energy.
Expandability and scalability	The analytical system of big data must support present and future datasets
Cooperation	Big data network architecture must be established to helps scientist and engineers in various fields access different kinds of data and fully utilize their expertise.

There is also some other challenges- RDBMS apply only to structure data it is not capable to handle the huge volume and heterogeneity of big data. The research community has proposed solution from different perspectives [34]. Table 1

15. FUTURE SCOPE OF BIG DATA

Mukherjjee and Ravi (2016) discussed that big data has been coined to refer to the gargantuan bulk of data that cannot be dealt with by traditional data-handling techniques. It is a novel concept, and it is intend to elaborate in a palpable fashion. It commences with the concept of the subject in itself along with its properties and the two general approaches of dealing. Comprehensive study further goes on to elucidate applications of big data in all diverse aspects of economy. The utilization of big data analytics after integrating it with digital capabilities to secure business growth and its visualization to make it comprehensible to the technically apprenticed business analyzers. Aside this, the incorporation of big data in order to improve health, for the betterment of finance, telecom industry, food industry, for fraud detection and sentiment analysis have been delineated. Challenges that are hindering the growth of big data analytics are need to be accounted. This topic has been segregated into two arenas- one being the practical challenges while other being the theoretical challenges. Hurdles of securing the data and democratizing it have been elaborated amongst several others such as inability in finding sound data professionals and relevant software's that possess ability to process data at a high velocity[34].

Table 2. Explain the future research scope of big data (Source: [35])

1	Selection of appropriate data sources for particular goals. The amount of available data is increasing; due to current techniques do not allow us to process all dada available in timely manner.
2	Selection of appropriate data analysis methods. There is various type of methods that can be use to process data but on given a particular datasets many methods may be applicable.
3	Integration of different data sources, to study complicated marketing problems because some complicated business problems require combining data from different sources.
4	Investigation, to deal with the heterogeneity among data sources. Data collection and analysis methods may be different due to different structure, quality, granularity and objective.
5	Examination of balance investments in marketing intelligence techniques because big data enabled marketing intelligence will become come a competitive sources for consumer behavior and product planning.
6	Big data of a variety of formats and qualities will continue to grow and be digitized then the need to refine the framework as the big data technology.

16. CONCLUSION

In this exhaustive review authors have provided a comprehensive outlook of the current state of research contributions in big data. It contains views, categories and aspects of big data technologies and describe about big data analysis, big data view point using cloud computing. Mainly the big data problem could be characterized by Volume, Variety and Velocity and has immense amount of applications. Value and complexity are the other two aspects. It has been discussed how to view difficult problems in a simple way under big data technology. Focus has been made on analysis and computing, classification of technology, big data and cloud computing, Docker Container, storage, big data optimization and privacy, machine learning concept of social environment, social networking, forecasting and some special issues.

17. REFERENCES

[1] Marr, Bernard, 2015. "A brief history of big data everyone should read", available at: https:// www .linkedin.com/pulse/brief-history-big-data-everyone-should-read-bernard-marr.

[2] Jagdish,H.V. , "Big Data and Science:Myths and Reality", Journal of Big Data Research, 2015, Elsevier, pp. 49-52.

[3] Cárdenas, Alvaro A., 2013. "Big Data analytics for security intelligence" Cloud Security Alliance, available at: www.cloudsecurity alliance.org/ research/big-data.

[4] Singh, Dilpreet and Reddy, Chandan K. , "A survey on platforms for big data analytics", Journal of Big Data, 2014, Springer Open, 2:8, pp. 1-20.

[5] Ding Junhua, Hu Xin-Hua, and Gudivada Venkat , "A Machine Learning Based Framework for Verification and Validation of Massive Scale Image Data", IEEE Transaction on Big Data , 2017, Issue: 99, pp. 1-18.

[6] Olshannikova Ekaterina , Olsson Thomas, Huhtamäki Jukka and Hannu Kärkkäinen, "Conceptualizing Big Social Data", Journal of Big Data, 2017, Springer Open, 4:3,pp. 1-19.

[7] Barrachina Arantxa Duque and O'Driscoll Aisling , "A big data methodology for categorising technical support requests using Hadoop and Mahout", Journal of Big Data, 2014, Springer Open.1:1, pp. 1-11.

[8] TargioHashem,IbrahimAbaker , Yaqoob,Ibrar , Anuar ,Nor Badrul , Mokhtar ,Salimah, Gani ,Abdullah,and Khan ,SameeUllah , "the rise of big data on cloud computing: Review and open research issues", journal of Information System, 2015, Elsevier, vol 47, pp. 98-115.

[9] Inukollu, Venkata Narasimha, Arsi ,Sailaja and Rao Ravuri ,Srinivasa , "Security Issues Associated With Big Data In Cloud Computing", International Journal of Network Security & Its Applications (IJNSA), 2014, Vol.6, No.3, pp. 45-56.

[10] Varma,P.China Venkanna , "Analysis of a Network IO Botleneck in Big Data Environments Based on Docker Containers",Journal of Big Data Research, 2016, Elsevier, Vol 3, pp. 24-28.

[11] Zheng, Charles, and Thain, Douglas 2015. Integrating Containers into Workflows: A Case Study Using Makeflow, Work Queue, and Docker. In Proceeding of the 8th International Workshop on Virtualization Technologies in Distributed Computing, pp. 31-38.

[12] Sawant,Poonam G.,and Desai, B.L., "Big Data Mining: Challenges and Opportunities to Forecast Future Scenario", International Journal of Innovative Research in Computer and Communication Engineering, 2015, Vol. 3, Issue 6, pp. 5228-5232

[13] Harsoor,Anita S.,and Patil,Anushree , "Forecast Of Sales Of Walmart Store Using Big Data Applications", International Journal of Research in Engineering and Technology, 2015, Volume 04, Issue 06, pp. 51-59

[14] Fan,Wei, and Bifet,Albert 2013. "Mining BigData:Current Status, and Forecast to the Future", Volume 14, Issue 2, pp. 1-5.

[15] Liu, Na, and Zhou, Jianfei , "The Research and Application of a Big Data Storage Model", International Journal of Database Theory and Application, 2015, Vol.8, No.4 , pp. 319-330.

[16] Li, yang, Guo, Li and Guo, Yike 2012, An Efficient and Performance- Aware Big Data Storage System. In Proceeding of the CLOSER 2012 International Conference on Cloud Computing and Services science, pp. 102-116.

[17] Li, Bo 2013. Survey of Recent Research Progress and Issues in Big Data, available at: http :// www.cs.wustl.edu/~jain/cse570-13/ftp/bigdata2/

[18] Simo, Hervais 2015. Big Data: Opportunities and Privacy Challenges, available at: https://arxiv.org /ftp/arxiv/papers/1502/1502.00823.pdf

[19] Zeng,Gang , "Research on Privacy Protection in Big Data Environment", Journal of Engineering Research and Applications, 2015, Vol. 5, Issue 5, ISSN : 2248-9622.

[20] Report To The President Big Data and Privacy: A Technological Perspective 2014. available at: https://www.whitehouse.gov/blog/2014/05/01/pcast-releases-report-big-data-and-privacy

[21] European Commission- Directorate-General for Health and Consumers Unit D3 eHealth and Health Technology Assessment 2014. The Use of Big Data in Public Health Policy and Research, available at: http://ec.europa.eu/health/ehealth/docs/ev_201411 18_co07b_en.pdf

[22] Coulton,Claudia J. , Goerge ,Robert, Putnam-Hornstein ,Emily and Haan ,Benjamin de 2015. Harnessing Big Data for Social Good: A Grand Challenge for Social Work ,American Academy of Social Work and Social Welfare, available at: http://aaswsw.org/wp-content/uploads/2015/12 / W P11-with-cover.pdf

[23] Tan,Wei, Blake ,M. Brian , Saleh, Iman and Dustdar ,Schahram , " Social-Network-Sourced Big Data Analytics", IEEE Internet Computing, IEEE Computer Society, 2013, volume 17, issue 5, 62-69.

[24] Moniruzzaman,A. B. M. and Hossain,Syed Akhter , " NoSQL Database: New Era of Databases for Big data Analytics - Classification, Characteristics and Comparison", International Journal of Database, Theory and Application, 2013, Vol. 6, No. 4, pp. 1-13.

[25] Tsai, Chun, Wei, Lai ,Chin‐Feng, Chao ,Han‐Chieh and Vasilakos, Athanasios V. , " Big data analytics: a survey", Journal of Big Data, Springer Open, 2015, volume 2, issue 21, pp. 1-32.

[26] Zhou, Zhi-Hua , "Big Data Opportunities and Challenges: Discussions from Data Analytics Perspectives", IEEE Computational Intelligence Magazine, 2014, available at: http://ieeexplore .ieee .org/document/6920114/

[27] Najafabadi,Maryam M, Villanustre ,Flavio, Khoshgoftaar ,Taghi M, Seliya ,Naeem, Wald ,Randall and Muharemagic, Edin , " Deep learning applications and challenges in big data analytics", Journal of Big Data, 2015, Springer Open Journal, 2:1, pp. 1-21.

[28] Tarwani,Kanchan M., Saudagar ,Saleha S. and Misalkar ,Harshal D. , "Machine Learning in Big Data Analytics: An Overview",International Journal of Advanced Research in Computer Science and Software Engineering, 2015, Volume 5, Issue 4, pp. 270-274.

[29] Kashyap,Hirak, Ahmed ,Hasin Afzal, Hoque, Nazrul, Roy,Swarup and Kumar Bhattacharyya, Dhruba 2015. Big Data Analytics in Bioinformatics: A Machine Learning Perspective, available at: http://arxiv.org/pdf/1506.05101.pdf

[30] Chen, Min, and Mao, Shiwen, Liu, Yunhao , "Big Data: A survey", Mobile Network and Applications, Springer Link, 2014, volume 19, issue 2, pp. 171-209.

[31] Kaisler, Stephen, 2013. Big Data: Issues and Challenges Moving Forward", 46th Hawaii International Conference on System Sciences, IEEE, dio 10.1109/HICSS.2013.645

[32] Special Issue on Big Data, Analytics and High Performance Computing, available at: http:/ / www .journals.elsevier.com/big-data-research/call-for-papers/special-issue-on-big-data-analytics-and-high-performance-com

[33] Lee, Jae-Gil and Kang Minseo , "Geospatial Big Data: Challenges and Opportunities", Journal of Big Data Research, 2015, Elsevier, pp. 74-81.

[34] Mukherjee, Samiddha, and Shaw, Ravi , "Big Data – Concepts, Applications, Challenges and Future Scope", International Journal of Advanced Research in Computer and Communication Engineering, 2016, volume 5, issue 2, pp. 66-74.

[35] Fan, Shaokun, Lau Raymond Y.K. and Zhao J. Leon, "Demystifying big data analytics for business intelligence through the lens of marketing mix", Journal of Big Data Research, 2015, Elsevier, volume 2, issue 1, pp. 28-32.