I J C T A, 9(34) 2016, pp. 235-243

Significance and Challenges in Big Data: A Survey

B. Jothi[*], M. Pushpalatha[**] and S. Krishnaveni[*]

ABSTRACT

In upcoming time the Internet, mobile networks, cloud computing and wireless sensor devices have been a reason for huge amounts of data in industry, business and government areas. This paperwill briefabout the features and the importance of big data to identify the need for big data in present world .Then we will discuss about the various challenges, significance and complexities faced by big data ,as well as possible solutions to address these challenges. Finally, we conclude the paper by presenting several suggestions on carrying out big data projects

Keywords: Big Data, 3 V's, Hadoop, Pig, Hive, Hbase, DryadLINQ, SCOPE, Jaql.

1. INTRODUCTION

In the future, by virtue of big data, the competence under the second economy will no longer be that of labor productivity but of knowledge productivity generallypeople and systems areoverload with an massive amount of data generated from different web sources,which have reached a level from Exabyte's (1018) and zettabytes (1021) this is due to the fastest growth of advancement in technologies in digital sensors, communication, computation and storage devices which is forecasted to exceed the brain capacity of everyone living in the whole world by 2025 by Roger Magoulas.

1.1. What Is Big Data

The term "big data" can be defined as collection of data sets which are so huge and complex that processing using traditional data processing applications is very difficult.

Big data finds its usage in almost every area of today's industry and business functions and is surely an important factor in production especially by McKinsey. Manyika et al. [6] defines big data whose data sets is beyond the ability of traditional database tools to collect manage and analyze data. Likewise, Davis and Patterson [1] say "data is too big to handle and analyze by traditional database tools lie SQL. Edd Dumbill in [1] conveys the multidimensionality of big data by adding "data is too big, movesfast, doesn't fit with proper structures into architecture".

Big data is classified into various categories namely, based on data collected from physical world like sensors devices, scientific experiments like social networks, Internet, health, finance, economics, and transportation.

1.2. Features of Big Data

Based on traditional data, the properties of big data can be characterized by 5V, namely, huge Volume, high Velocity, high Variety, low Veracity, and high Value [8].

[*] Assistant Professor, SRM University, *E-mail: jothi.b@ktr.srmuniv.ac.in*

[**] Professor, SRM University, *E-mail: pushpalatha.m@ktr.srmuniv.ac.in; krishnaveni.s@ktr.srmuniv.ac.in*

Volume (Data in rest). In field of business analytics large amount of data is needed to produce more accurate results. Larger the data set ,results in better models. As a result many organization stores massive data sets like health care data ,financial, biochemistry and genetic data [8]

Variety (Data in many forms)

Nowadays data are generated in different forms, there is no a constant structure which can be processed in an ordered form .data categorized to highly structured (relational database data), semi structured (web logs, social media) or unstructured (video, image audios), which makes the generation another V namely variability added to variety to draw attention on semantics [1].

Velocity (Data in motion). It means the speed or transfer data rate for streaming data and make it available for access and delivery. In need it is just velocity of incoming data but the ability to stream fast moving data into bulk storage for batch processing.

Value (Data in highlight). Big data architectures are designed to economically generate a useful knowledge from large amount of data by knowledge discovery or analysis (replacing/supportinghuman decision, discovering needs, segmenting populations.

To customize actions) needed for business operation

Veracity (Data in doubt). It refers to the accuracy of data, truth or fact .sometimes uncertainties results due to ambiguities, incompleteness, spam and latency. So always veracity is not proven but they can be assigned as probability

2. BIG DATA MANAGEMENT

Data processing is a process of collecting,transforming,management of data information for end users. over time and years the key challenges faced are storage, transportation and processing of high throughput data. The Karmasphere [1] have divided big data analytics into Acquisition ,Organization, Analyze or Decision steps and similarly Computing Community Consortium divides into an Extraction/Cleaning step and an Integration step.

2.1. Acquisition

The architecture of big data is to acquire of high speed data from variety of different sources with different protocols where filters are needed to process raw data to reduce uncertainity.In some application the conditions of generation of data ,these metadata are important for further analysis.

2.2. Organization

Various forms of data like text, compressed video and audio files are generated from different sources which are parsed to extract valuable information like entities and relationships between them. so at this point data is either structured or semi structured have to be put in computable node to store it in right locations like datawarehouse, data marts, complex event processing engine or NOSQL database. ETL (extract, transform, load) has to be done to cleanse data.

2.3. Analyze

Werun queries, modeling and building algorithms to find new insight. Data mining is a computational process of discovering patterns from large sets using artificial intelligence ,machine learning, statistics and database systems by understanding the semantics.

2.4. Decision

After analysis the results are interpreted to originate valuable decisions from which user have understood about the particular data sets [14] thus the data are aggregated by collecting, cleansing and analyzing the data.

3. BIG DATA TECHNOLOGIES

Hadoop

There are various tools used in big data management to do analysis. Hadoop is apache open source framework developed for processing large amount of data with low cost common hardware which is reliable and scalable to support hundred to thousand of nodes in distributed way and is highly fault tolerant hadoop a java program supports complex data, conversion from instructed data to structured data ,parallel processing, machine learning and pattern identification etc .Regardless of the structure hadoop allows the ability to process large amount of data ,and it is made up of two core projects namely HDFS and Mapreduce [1].

3.1. HDFS

HDFS is based on Google file system which is designed in such way to run on large cluster of commodity hardware. HDFS strengths in its definition by saying it store very large datasets reliably and to stream those datasets at high bandwidth to user applicationsmean from 10 to 100 GB. And also it performs the batch processing rather than interactive use by users .In HDFS applications, files are written once but accessed many times and consequently data coherency is ensured, data are accessed in high throughput [4]. With HDFS file system metadata are stored in a dedicated server, theName Node, and the application data in other servers called Data Nodes. Except for processing large datasets, HDFS has many other goals whose major role is to detect and handle failure set the application layer. This objective is realized through a well-organized mechanism of replication where files are divided into blocks. Each block is replicated on a number of data nodes; all the data nodes containing a replica of a blockare not located in the same rack.

3.2. Map Reduce

It is used to solve the web search index creation problem and now days the main programming model is associated with implementation for processing and generating large datasets. The input data format work as framework for Map Reduce as a application specific is specified by the user and it is suitable for semi structured or unstructured data. The Map Reduce output is aset of <key, value>pairs. The name “Map Reduce” specifies that users have to specify an algorithm using two kernel functions: “Map” and “Reduce” by using intermediate key links. In a Hadoop cluster, a job is executed by subsequently breaking it down into pieces called tasks. When a node in Hadoop cluster receives a job, it has the ability to divide and run parallel over other nodes [12]. Here the data location problem is solved by the Job Tracker which communicates with the Name Node to help data nodes to send tasks to near-data data nodes. From this we can conclude <key, value>pairs are not a limitation to processing which does not seem, at first glance. It is feasible in map-reduce manner. Indeed, Map Reduce has been successfully used in RDF/RDFS and OWL reasoning and in structured data querying [23]. Around HDFS and Map Reduce there are tens of projects which cannot be presented in detail here. Those projects have been segregated as per their capabilities:

3.4. Storage and Management Capability

– **Cloud era Manager** It is a process of an end-to-end management application for Cloud era’s Distribution of RC File (Record Columnar File) a data placement structure for structured data. Here, tables are vertically, horizontally partitioned and compressed. It is an efficientstorage structure which allows fast data loading and query processing.

- **Database Capability** Oracle NoSQL is a very high performance <key, value>pair database which is convenient for non-predictive and dynamic data systems
- **Apache Cassandra** isa database which combines the convenience of column-indexes and log-structured updates high performance
- **Apache Hive** can be seen as a distributed data warehouse [15]. It enables easy data ETL from HDFS or other data store. It has the advantage of using a SQL. It work as "an open-source, in-memory, distributed NoSQL database which used for coordinating and naming services for managing distributed applications.

3.5. Processing Capability

- **Pig** which is intended to allow people using Hadoop to emphasize more on analyzing large datasets and spend less time to write mapper and reducer programs [11].
- **Chukwa** It is a data collection system which monitors distributed systems that are large;
- **Oozie** which is a open-source tool for handling complex pipelines of data processing. Using Oozie, users can define actions and dependencies between them and it will schedule them without any intervention [11].

3.6. Data Integration Capability

- **Apache Sqoop** a tool designed for transferring data from relational database. It automatically generates classes needed to import data into HDFS after analyzing the schema 'stables; then the reading of tables' contents is a parallel Map Reduce job.
- **Flume** is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data.

3.7. Visualization techniques

To make an ultimate goal for Big Data analysis and to achieve the goal requires good visualization of Big Data content. Thus it is considered the real interest vested in the area of visualization i.e. "techniques and technologies used for creating images, diagrams, or animations to communicate, understand, and thereby improving the results of big data" [10]. Big Data visualizations context is very static. Data are not being stored in a relational format and real-time updates which requires processing volumes of data; but nowadays this problem has started to be addressed [3]

- **Tag Cloud** it is a method for visualizing and linking concepts of a precise domain or web site. These concepts are always written using text properties like font size, weight, or color.
- **Clustergram** a technique displays how datasets of individual members are assigned to clusters for processing and partitioning results
- **History Flow .**F.B. Viégas, M. Wattenberg and K. Dave [14] presented the visualization technique designed to show document evolution efficiently.

  Where horizontal axis represents time and vertical axis represents author details.
- **Spatial information flow**. It is mostly represented as a lighting graph where edges gets connected to sites located on a map.

3.8. Data analytics

It is defined as a advanced analytic techniques in big data [16]. Big data and analytics can be put together. The prior conditions are present for the development of big data Analytics. Tools and storage capabilities

can handle big data. Its size big data provides large statistical samples and enhanced results of experiments. Due to the characteristics of big data, mainly variety, there are many techniques used for analytics on big data.

- **Association rule learning**to find relationships among entities (mainly used in recommendation systems).
- **Machine learning**It is used to learn complex patterns and make intelligent decisions based on computer.
- **Data mining** It is a combination of machine learning and statistics with database management.
- **Cluster analysis:**It is used as unsupervised machine learning. It aims to divide data into smaller clusters having the same set of characteristics not known in advance.

Crowd sourcing used to collect data and/or features and metadata to enhance the semantics of data.

- **Text analytics**aims for analyzing large collections of text (email, web pages, etc.) to extract information. Text analytics is used for topics modeling, Due to their design it results in inability to capture the subtleties of the processes which provide these data .Moreover, these techniques sometimes behave badly with very large datasets. It is the case for example of learning-based techniques. There, size of training data can exceed memory or the fast growing number of features can lead to a high execution time. Sengamedu presents some scalable methods that are applicable for machine learning. It mainly includes visualizations of multi-form, multi-source and real-time data.

4. SIGNIFICANCE OF BIG DATA

Because of its monstrous worth, huge information has been basically advancing the way we live, work, and think [9]. It gives detail the centrality of huge information in different viewpoints.

4.1. Significance to national development

Distributed computing, and other developing IT advances has made different information sources increment at an exponential rate, while at the same time making the structures and sorts of information progressively mind boggling. This usage will assume a vital part in advancing maintained financial development of nations and upgrade the aggressiveness of companies. In the up and coming time huge information will turn into another purpose of monetary development. With the assistance of huge information, organizations will have the capacity to overhaul and change to the method of Analysis as a Service, thereby adjusting the nature of IT businesses.. At the national level, the limit of aggregating, handling, and using tremendous measures of information will turn into another historic point of a nation's quality. In China, an administration report has plainly recommended that the internet, and also Remote Ocean and profound space, are key zones of the national center interests. The fall behind in the field of huge information exploration and applications not just means the loss of its modern key favorable position, additionally recommends provisos in its national security the internet. It is reported by the United States [10] in March 2012, is not just a key arrangement that elevates the US to persistently lead in the cutting edge fields, however plans to ensure its national security furthermore make propels in its financial improvement. By and large, the Western nations are moving under their national motivation towards a modernization of their national quality through enormous information examination and applications. It is normal that future monetary and political rivalries among nations will lay on misusing the capability of enormous information, among other conventional perspectives.

4.2. Significance to industrial upgrades

Enormous information is right now a typical issue confronted by numerous businesses, and tosses difficulties to these ventures' digitization. This implies that information is no more a mechanical part result, yet has turned into a key nexus of all viewpoints. In this sense, the investigation of normal issues and center advances of enormous information will be the attention on IT and its applications. It won't just be the new motor to maintain the high development of the data business, additionally the new device for ventures to enhance their intensity. Numerous huge organizations, including Google, Microsoft, Amazon, Face book, Alibaba, 2Baidu, 3Tencent, 4 and other IT mammoths, are taking a shot at distributed computing innovations and cloud-based figuring administrations. Huge information and distributed computing are seen as two features of a coin: enormous information is an utilization of distributed computing The firmly coupled huge information and distributed computing are anticipated to change the Internet environment and even influence the example of the whole data industry.

4.3. Significance to scientific research

It has set off an unrest in [11] scientific speculation and techniques. It is notable that the most punctual exploratory examination in mankind's history depended on trials. A long time after hypothetical science developed which was portrayed by the investigation of different laws and hypotheses. Be that as it may, on the grounds that hypothetical investigation is excessively mind boggling and not attainable for taking care of down to earth issues individuals started to look for reproduction based strategies, which prompted computational science. The development of enormous information has brought forth another examination that is with huge information analysts may just need to discover or mine from it the required data, learning and knowledge. They even does not have to straightforwardly get to the items to be study it isolates information concentrated science from computational science. Dark trusted that the fourth worldview might be the main systemic path for explaining a portion of the hardest worldwide difficulties we confront today. Basically, the fourth worldview is not just an adjustment in the method for investigative examination, additionally an adjustment in the way that individuals think [9].

4.4. Significance to emerging interdisciplinary research

Big data innovation and crucial exploration have turned into an examination center in the scholarly world. An interdisciplinary control called information science has been bit by bit ascended to zenith. This takes huge information as its exploration question and goes for summing up the extraction of learning from information. It is comprising of numerous orders comprising of data science, arithmetic, sociology, system science and so on Various strategies and hypotheses from numerous fields, including signal handling, likelihood hypothesis, machine learning, factual learning, PC programming, information product lodging, and superior processing have been utilized . Numerous examination focuses/organizations on enormous information have been set up as of late in various colleges all through the world

4.5. Significance to helping people better perceive the present

Huge Data has particularly enormous arranged information, contains an abundance of societal data and can in this manner be seen as a system mapped to society. To this end, examining huge information and further compressing and discovering signs and laws it certainly contains can help us better see the present.

Case in point, two illustration lists of interest created in China[12] make incredible utilization of information freely accessible from the Internet. Since 2007, China Survey and Assessment Center are partnered to Renmin University of China, has issued yearly "China Development Index." This record, with four individual files on wellbeing, training, expectation for everyday comforts, and social environment, plans to gauge business as usual and unscramble the issues of China's improvement. As another exertion,

since 2010, Xinhua News Agency, together with Dow Jones Newswires, distributed twice per year "Xinhua-Dow Jones International Financial Centers Development Index." By looking at and breaking down different subjective and target markers and by joining subjective and quantitative examination, this record uncovers the present improvement status and laws of worldwide money related focuses. Profound mining data contained in huge information can likewise individuals settle on better choices. In the eighteen months before Election Day, Obama's information investigation group made an enormous information handling framework. Through constant information gathering and examination not just might it be able to advise the crusade group how to discover voters and to stand out enough to be noticed, yet it likewise dissected the inclination for voters to vote. Consistently, the information investigation group directed reproduction on the race and exhibited reenactment brings about the following day to comprehend the likelihood that Obama may win in a few territories, taking into account which the group can dispense assets more pre-cicely. Later truths exhibited that the information investigation group assumed a urgent part in Obama's re-race, a long ways past individuals' creative energy. Investigating and mining huge information can likewise viably shield open security and battle criminal and monetary wrongdoings.

5. GRAND CHALLENGES OF BIG DATA

There are numerous difficulties in saddling the capability of huge information today, going from the configuration of preparing frameworks at the lower layer to examination implies at the higher layer and in addition a progression of open issues in exploratory exploration. Among these difficulties, some are created by the attributes of huge information, a few, by its present investigation models and strategies, and a few, by the confinements of current information handling frameworks.

5.1. Data complexity

The development of huge information has given us extraordinary substantial scale tests when managing computational issues, face significantly more mind boggling information objects. As a fore said, the run of the mill attributes of huge information are enhanced sorts and examples, convoluted between connections, and significantly changed information quality. Customary information examination and mining undertakings is, for example, recovery, subject disclosure, semantic investigation, and conclusion investigation, turn out to be to a great degree troublesome when utilizing huge information. The absence of information with respect to the laws of appropriation and affiliation relationship of huge data. The profound comprehension on the inalienable relationship between information multifaceted nature and computational intricacy of enormous information, and in addition space situated huge information handling strategies. All these enormously restrict our ability to plan exceedingly proficient computational models and techniques for taking care of issues utilizing huge information. A key issue is the way to detail or quantitatively portray the vital attributes of the multifaceted nature of huge information. The hypothesis and models of information dispersion under multi-modular interrelation To deal with inborn associations between information multifaceted nature and spatiotemporal computational many-sided quality. In addition, by demonstrating and examining the inherent instruments of information many-sided quality, we will have the capacity to explain the standards and components for preparing huge information into a strong establishment for enormous information figuring.

5.2. Computational complexity

Three of the key components of huge information, to be specific, multi-sources, enormous volume, and quick changing, make it troublesome for customary figuring techniques like machine learning, data retrieval, and information mining to successfully bolster the handling, examination and calculation of huge information. Such calculations can't just depend on past measurements, investigation apparatuses, and iterative calculations utilized as a part of conventional methodologies for taking care of little measures of information.

New methodologies should split far from suspicions made in customary calculations in light of autonomous and indistinguishable circulation of information and sufficient testing for creating dependable measurements. Rethink and explore its process ability, computational unpredictability, and calculations. New components in enormous information handling, for example, inadequate specimens, open and indeterminate information connections, and uneven dissemination of significant worth thickness, give awesome open doors, as well as stance fabulous difficulties, to considering the process ability of huge information and the improvement of new registering ideal models. To address the computational multifaceted nature of enormous information applications. It is essential to split far from customary processing driven standards and build up information driven push-style registering ideal models and investigate powerless CAP system shared-information framework model and its arithmetical computational hypothesis. it is important to create calculations for disseminated and gushing registering and shape a major information arranged figuring structure where correspondence, stockpiling, and processing are very much coordinated and improvement. To investigate existing diminishment based figuring strategies where huge information is decreased on interest from being sufficiently extensive to being simply enough and to being sufficiently profitable. At last, it is expected to create bootstrapping and testing based nearby calculation and guess techniques and propose novel hypothetical premise for enormous information calculations that are adaptable to taking care of a lot of information.

5.3. System complexity

Big data is suitable for handling diverse data sets to support scientific research. Because of enormous and complex data sets with sparse structure getting generated it is difficult to handle as it results in increase in computational time and long duty cycle. This results in design of new architecture, framework and processing systems for handling large data management challenge.

6. CONCLUSION

Enormous information has had a solid effect in verging on each part and industry today. In this paper, we have quickly explored the open doors and noteworthiness of huge information, and additionally some fabulous difficulties that enormous information brings us. Its a well known fact that in enormous information examination and applications, industry is in front of the educated community. For instance, as indicated by the figure Alibaba unveiled in March 2014, their server farm has put away more than 100 PB of handled information, which adds up to 100 million high-determination films. Amid the simply past “Singles’ Day” (otherwise called “Twofold 11 Day”), Alibaba pulled in CNY 9.3 billion in deals from this shopping occasion, which related to around 278 million or-ders. For this yearly shopping occasion, Alibaba built up a continuous information preparing stage called Galaxy, which can deal with 5 million exchanges for every second. The aggregate sum of information that Galaxy can prepare each day is around 2 PB. Industry is more effective in this admiration since it has two key main impetuses: they truly need to have huge information progressively and they have the necessities on improving utilization of the information. Firstly, there must be clear prerequisites, paying little respect to whether they are specialized, social, or financial. Furthermore, to productively work with huge information, we should investigate and discover the part structure or piece information to be handled. Discovering piece information and structures, which are sufficiently little but can portray the conduct and properties of the fundamental enormous information, is non-unimportant in light of the fact that it is extremely area particular. Thirdly, a top-down administration model ought to be received. In spite of the fact that a base up methodology may permit us to take care of some corner issues, the detached arrangements frequently can’t be assembled into a complete arrangement. At long last, the objective ought to be to take care of the whole issue by an incorporated arrangement, as opposed to making progress toward separated accomplishments in a couple of angles. To put it plainly, a coordinated building methodology ought to be utilized in dealing with a major information venture.

REFERENCES

[1] The Pre Big Data Matching Redundancy Avoidance Algorithm with Mapreduce byG. Somasekhar and K. Karthikeyan in Indian Journal of Science and Technology, Vol 8(33), DOI: 10.17485/ijst/2015/v8i33/77477, December 2015.

[2] MapReduce: A Technical Review T. Y. J. Naga Malleswari and G. Vadivu, Indian Journal of Science and Technology, Vol 9(1), DOI:10.17485/ijst/2016/v9i1/78964, January 2016.

[3] K. Davis, D. Patterson, Ethics of Big Data: Balancing Risk andInnovation, O'Reilly Media, 2012.

[4] T. Hey, S. Tansley, K. Tolle (Eds.), The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Corporation, 2009.

[5] G. Halevi, H. Moed, The evolution of big data as a researchand scientific topic: Overview of the literature, Res. Trends(2012) 3–6.

[6] Bigdata, http://en.wikipedia.org/wiki/Big_data, 2014.

[7] R. Thomson, C. Lebiere, S. Bennati, Human, model and machine: a complemen-tary approach to big data, in: Proceedings of the 2014 Workshop on Human Centered Big Data Research, HCBDR '14, 2014.

[8] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. Hung, Big data: the next frontier for innovation, competition, and productivity, Tech. rep., McKinsey Global Institute, 2011, available at: *http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation*

[9] G. Li, X. Cheng, Research status and scientific thinking of big data, Bull. Chin. Acad. Sci. 27(6) (2012) 647–657.

[10] I. O'Reilly Media, Big Data Now: 2014 Edition, O'Reilly Media, 2014.

[11] V. Mayer-Schonberger, K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think, Houghton Mifflin Harcourt, 2013.

[12] T. Kalil, Big data is a big deal, available at:http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal, 2012.

[13] T. Hey, S. Tansley, K. Tolle (Eds.), The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Corporation, 2009.

[14] I.M. Easton, L.R. Hsiao, The Chinesepeople's liberation army's unmanned aerial vehicle project: organizational capacities and operational capabilities, Tech. rep., 2049 Project Institute, March 2013.

[15] M. Schonlau, The clustergram: A graph for visualizinghierarchical and nonhierarchical cluster analyses, Stata J. 2 (4) (2002) 391–402. 12.

[16] F.B. Viégas, M.Wattenberg, K. Dave, Studying cooperation andconflict between authors with history flow visualizations,in: Proceedings of the SIGCHI Conference on Human Factorsin Computing Systems, CHI '04, ACM, New York, NY, USA,2004, pp. 575–582.

[17] D. Keim, H. Qu, K.-L. Ma, Big-data visualization, computergraphics and applications, IEEE 33 (4) (2013) 20–21.

[18] P. Russom, et al. Big data analytics, TDWI Best PracticesReport, Fourth Quarter.