Journal of Advances and Scholarly Researches in Allied Education
Vol. 13, Issue No. 1, April-2017, ISSN 2230-7540

Study on the Issues and Challenges of Big Data

Raju G.[1]* Dr. Kampa Ratna Babu[2]

[1] Research Scholar, SVU

[2] Associate Professor

Abstract – Data mining is a procedure planned to survey logical data (commonly business or market related data - additionally recognized as "Big Data"). There are a few data mining techniques, for example, exception investigation, association, clustering, expectation and affiliation standard mining. In this examination a few applications and the significance of clustering is talked about. To analyze the gigantic volume of data, clustering algorithms help in giving a ground-breaking meta-learning device. Various clustering techniques (counting customary and the as of late created) in reference to huge data sets with their aces and cons are being talked about in this exploration. Clustering techniques are generally utilized in various field like picture preparing, data mining and so on for finding distinctive new designs in basic data. Bunch techniques are worried about creating calculation that is demonstrated to be valuable. Presently multi day's numerous old strategies for clustering have been adjusted and returned to mirror the reality of certain calculation and improved on the off chance that they give a favorable position.

----------X----------

INTRODUCTION

"Big Data and Internet of Things (IoT)" are the blessings of the pioneering technological transformation in the present era of modern consumers and organizations. IoT has its concept rooted in the idea of interconnecting electronics and physical devices using a preferred connecting medium. Usually, the wireless medium is used for such interconnections. The idea of Big Data which was incepted in the year 2005 stated that any large quantity of data which cannot be maintained and organized by native tools and technologies can be termed as Big Data. It may comprise of datasets whose size ranges from Gigabytes to Terabytes and in some cases from Petabytes to Exabyte depending on the scenario and the applications.

DATA

An introduction Data may be defined as a collection of values involving qualitative or quantitative variables. Data may constitute raw facts, figures, numbers etc. Broadly data can be divided into three prime categories. These are given below.

Types of Data

Any data can be characterized by the way it is organized in the storage systems. Primarily there are three broad categories of data.

1. Structured data a
2. Semi-Structured data
3. Un-Structured data

Out of these categories approximately a whopping 80% of the data at our disposal is unstructured while only 15% of the data is semi-structured and a mere 5% of the data is structured.

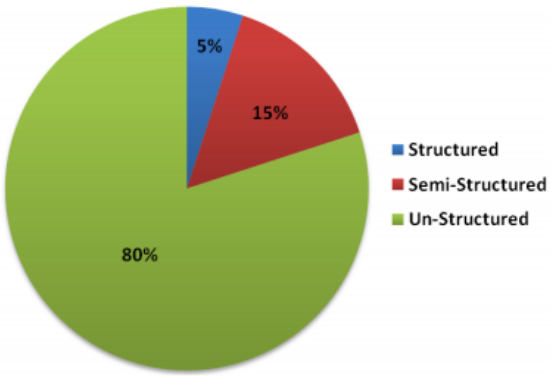

Fig 1: Approximate Percentage composition of different kinds of Data

Big Data:

The quickly changing innovation and the exponential development in data and information from fluctuated sources are making immense heaps of data which the conventional apparatuses and framework can't successfully store, process and break down. Such enormous data and

information sets are known as big data. The expression "Big Data" was first begat by "Roger Magoulas". One important point about big data is that it is a "relative" term, for example the meaning of big data changes from organization to organization and time to time. For instance, 500GB can be considered as big data for one organization while 20PB could be considered as big data for the other organization. The goliath measure of Big Data whenever oversaw appropriately can give significant bits of knowledge to the businesses to enter new verticals alongside fitting following of the data in heap ways. The big data has a few "properties" and "attributes" which recognizes it from the customary data sets.

BIG DATA TECHNOLOGIES

- **Column-oriented databases:** In section situated database stores data in segments instead of lines, which is utilized to packs enormous data and quick inquiries.

- **Schema-less databases:** Mapping less databases are generally called as NoSQL databases. Database gives a system to capacity and recovery of data that is demonstrated in methods other than the forbidden relations utilized in social databases. There are two sorts of database, for example, archive stores and key esteem stores that stores and recovers gigantic measure of organized, unstructured and semi organized data.

NEED FOR BIG DATA

The monstrous volume of data couldn't be immediately prepared by customary database strategies and instruments and it for the most part engaged and took care of organized data. At the season of progress of PCs the measure of data set away in the PCs are less because of its base stockpiling limit. After the creation of frameworks administration, the data set away in PCs are expanded in light of the fact that the improved advancements in the hardware parts. Next, the section of a web makes a blast to store huge accumulations of data and it may be used for different purposes. This circumstance raised worries about the presentation of new research related ideas like data mining, organizing, picture handling, network processing, distributed computing and so on are used for breaking down the various sorts of data which are used in different spaces. Various new techniques, calculations, ideas and strategies have been proposed by the analysts for dissecting the static data sets.

Characteristics of Big Data

The big data can be characterized by 7 V's. These are listed below:

1. Volume
2. Velocity
3. Variety
4. Veracity
5. Value
6. Variability
7. Visualization

Volume

It accounts to the huge amount of data that is being generated every day. It is a well-known fact that, about 80 % of the data that we have today at our disposal is created in the past few years. These data include text, audios, videos, images, seismic data, DNA related data, scientific data, data from sensors, data from medical records etc. Some examples of such data explosion include

- 200 hours of videos uploaded on YouTube every minute on a daily basis.
- 500 million tweets per day
- 200 million Whatsapp voice message are sent daily
- 34 billion text messages of Whatsapp are sent and received daily.

Velocity

Velocity may be defined as the pace at which the data is being generated. With the advent of IoT technology, the rate of generation of data has taken an exponential growth. Some of the examples include:

- 2.7 billion "Like" actions per day on Facebook.
- 300 million new photos uploaded daily on Facebook.
- Instagram receives 3.5 billion "Likes" per day and 80 million photos are shared per day.
- Google receives and processes 40,000 queries per second on a daily basis.

Variety

The data that is being generated consists of a mix of structured, semi-structured and unstructured data. However, more than 85% of the big data so

generated is largely unstructured. It comes in varied forms like text, images, audios, videos, logs, and data from sensors etc. Variety is one of the biggest challenges that the big data management systems have to handle. For processing the heterogeneous data from varied sources, it should first be converted into a standard format which in itself a big challenge.

Variability

It may be defined as the property of the data wherein the meaning of the data is rapidly evolving and changing For example, at a particular point in time, the data means one thing, which at some other time and it means something else.

Veracity

Veracity refers to the trustworthiness of data. With the increase in volumes and variety of data, its veracity tends to decrease because of the complexities attached with the generated data. Veracity is an important characteristic to be considered in order to perform the risk analysis on the basis of a given big data set.

Challenges &Opportunities

Applications, Issues and Challenges of Big Data The technological advancement around the globe has led to the rapid generation of huge amount of data which cannot be effectively handled by the classical tools, techniques, systems and sub-systems. Big data has certainly created a paradigm shift and transformed the way we live, work, analyze and visualize the things around us. These huge piles of data contain an ocean of information which is hidden within them. If we are able to extract those hidden patterns and information, we can have state-of-the-art insights about the data and the system as a whole which can be used to develop novel tools, techniques, technologies, methodologies, predictions and forecasting etc. for the betterment and wellbeing of the human beings and organizations. For example, by analyzing the data of a customer, companies can give better-customized products and services to the users.

1. **Data-driven decision making:** With the emergence of big data technologies, the design making process has taken a complete paradigm shift. Today, the decisions are primarily data-driven, which means that the decisions are being taken on the basis of the data generated/captured/stored/analysed by the system. This promotes a better understanding of the system and its working as a whole which is pivotal in taking informed decisions.

2. **In-depth and better insights about the data and the system:** The effective mining of big data can open doors for state-of-the-art and previously undiscovered patterns and thus can prove to be extremely useful for the organizations and enterprises.

3. **Unlocking new horizons of Information:** The value and visualization clarity that big data brings forth is surely capable of transforming the very core of information analysis and processing system. This information can be used for the economic and strategic growth of an individual, region or the country.

4. **Better training of the systems and individuals:** With new tools and technologies like Deep learning, Artificial Intelligence and Edge Computing, we are now capable of providing cutting edge training to the systems and individuals.

5. **State-of-the-art SWOT Analysis:** If we are able to analyze the strengths, weaknesses, opportunities and threats of the system, we can come up with techniques to develop best possible systems.

6. **Finding new relationships among data:** If we are able to find novel (previously undiscovered) relationships among the data, we can use those relationships to come up with better solutions and services to the users. For example, if we are analyzing a patient's data and we come across two data sample (which were previously unrelated) and somehow we process those individual data to deduce a new relationship among them, then this new relation can be useful in providing better diagnosis and treatment to the patient.

7. **Improve Operational Efficiency:** Big data can be used to improve the overall operational efficiency of an organization by identifying and analyzing patterns and relationship among various sections of operations.

8. **Identifying new market:** The enterprises can harness big data to identify potential customers and new market places for their product and services. One example of these kinds of techniques is currently being used by online shopping giants like Flip kart, Amazon, wherein if a person selects one product then the website also displays other items which are related to the product that the customer has selected with a tagline "Users who buy this also

buys this" or "items frequently brought together".

9. **Improve customer satisfaction:** With the proper processing of big data, the organizations are able to track the usage of their products by the customers (in the form of taking feedbacks, promotional giveaways etc.). If the customers are not satisfied with the product or service, the option to return or refund works wonders in satisfying the customers and win their trust**.**

10. **Informed strategic decision making:** If the organizations and policy makers are able to identify the core needs and demands of the customers or citizens, they can bring up revolutionized strategic changes in the system in order to satisfy the users in a much comprehensive manner.

ISSUES AND CHALLENGES OF BIG DATA

1. Heterogeneous Data:
2. Massive Scale of Data:
3. Unstructured and Complex nature of Data:
4. Privacy and Security:
5. Data Ownership and provisioning:
6. Data Redundancy:
7. Data cleaning and identification of Data outliers and Noises in Data:
8. Data integration and linkage at various levels:
9. Data validity:
10. Data veracity:
11. Data authenticity:
12. Computational complexities:
13. System complexities:
14. Quality of Service (QoS):
15. Data Analysis:

BIG DATA AND ITS TOOLS

This introduces evolution of Big Data and its various tools, Big Data analytics in cloud computing, requirement for Big Data analytics, data mining & its process, data mining methods and roles of Big Data in various sectors.

- **Evolution of Big Data**

In Big Data from digital data to health data are included. It has evolved from various stages i.e. from primitive and structured data to complex relational data. Now a days very complex and unstructured data are also included. The concept of Big Data came into the light when the growth rate in volume of data was known as information explosion (about 70 years ago). In 1944 Fremont Rider, a librarian estimated that size of American Universities libraries is getting doubled every sixteen years.

- ***Tools Used with Big Data***

For the development of data, machine learning is used. The main aim to develop algorithms that permits computers to form a pattern based on observed data. Generally supervised and unsupervised learning algorithms are used with Big Data. To clarify the data the most fundamental algorithm is used and it is known as support vector machine (SVM). However in present time recently parallel support vector machine overcomes the drawback of SVM such as scalability, time, and memory problems. Neural Networks and Artificial Neural Networks ANN are mature techniques that are used in adaptive control, image analysis, and pattern recognition.

- ***Apache Hadoop and MapReduce***

Hadoop Hadoop provides an open-source software framework for distributed storage and processing applications on very large datasets, written in java. Hadoop platform includes higher level declarative languages for writing queries and data analysis pipelines. Hadoop is used by approximately 63% of organizations to manage and analyze huge number of unstructured logs and events (Sys. Con Media, 2011). Hadoop is composed of many components but in Big Data two mostly components Hadoop Distributed File System (HDFS) and MapReduce() are used. The other components provide complementary services and higher level of abstraction.

The following are the features supported by HDFS:

- Scalability: HDFS is scalable to petabytes of data and is flexible to add and/or remove data nodes in order to store and process huge data effectively.
- Reliability: When data is stored on HDFS, it divides the data into data blocks and are stored in data nodes in the Hadoop cluster. Block replication is maintained to provide reliability of data such that even if a particular data node goes down due to

power or hardware failures, the data is accessible.

- Fault tolerant: One of the primary reason in using Hadoop platform is high degree of fault tolerance. There might be chances of failures at either Name Node, Data Node or network components. Detection of faults and automatic recovery is the goal of HDFS. Replication of data makes HDFS reliable and fault-tolerant. By default, the replication factor used in Hadoop is three. Due to this replication, the Hadoop clusters are highly fault-tolerant.

- Computation cost: HDFS is a distributed file system capable of running on commodity hardware. HDFS moves the computational 10 work to the data. Data localization is an important concept of Hadoop that brings computations work closer to the node where the data resides and thus making the data processing much faster.

- Flexible: Hadoop is a Java-based platform that can run on any operating system. Hadoop enables to easily access various data sources and different formats which is suitable for structured, unstructured and semi-structured data formats. Hadoop can be used for a variety of purposes such as data warehousing, fraud detection and trade analysis.

- High throughput: The amount of data moved successfully from one node to another in a given time period is throughput. Parallel processing of data makes it possible in reducing the time taken to send the data from one node to another and which achieves high throughput.

MapReduce system is the main part in Hadoop framework that is used for processing and generating large datasets on a cluster with distributed or parallel algorithm. It is a programming paradigm used to process large volume of data by dividing the work into various independent nodes. A MapReduceprogram corresponds to two jobs, A Map method which include obtaining, filtering & sorting datasets and A Reduce method which include finding out summaries and generate final result. MapReduce system arranges distributed servers, manage all communications, parallel data transfers, also provide redundancy and fault tolerance.

MapReduce is a programming model for parallel data processing and comprises of two functions namely: Map function and reduce function. Map function operates on set of key, value pairs. Map is applied in parallel on input dataset. Reduce function operates on set of key, value pairs, which is applied in parallel to each group. The reduce function can iterate through the values which takes the output of map phase. The reduce function accepts the output of the map function and operates on the array of grouping of key, value pairs. The output of reduce function produces a collection of key, values.

RESOURCES IN CLOUD FOR BIG DATA

An ideal computing environment can be created using cloud and offering many different products for Big Data users. IaaS requires more investment of IT resources in implementing Big Data analytics with installation of software like: Hadoop framework, NoSQL database as Cassendra, MongoDB etc. Some examples of such type of cloud providers with IaaS for Big Data include: Amazon.com, AT&T and IBM. Some examples of using Cloud with Big Data:

- ***IaaS in Public Cloud:*** Using IaaS provider can be capable to create on-demand virtual machines with unlimited storage and large processing power. The infrastructure of public cloud provider would be used for Big Data services as anybody doesn't want to use their infrastructure. An example: Amazon Elastic Compute Cloud (Amazon EC2) service to run real-time predictive model, requires parallelprocessing of massively distributed data in a scalable manner.

- ***PaaS in a Private Cloud:*** Paas provides tools and libraries to its developers in cloud to fast develop, run and deploy applications in a private or public cloud without worry about maintaining complexities of Hadoop like implementation environment. PaaS integrated with Big Data is a fully packaged infrastructure that includes Big Data software, infrastructure, tools and managed services. Using PaaS enterprises can rapidly develop secure tools and techniques to Big Data analytics applications. PaaS developers are moving to enhance capabilities of Hadoop and MapReduce() like Big Data analytics applications [51]. An example: Google Cloud Engine offers cloud based capabilities for virtual machine computing with secure and flexible environment.

DESIGN PRINCIPLES FOR BIG DATA SYSTEMS

The seven principles in designing Big Data systems are as follows:

1. Good and proper architectures and frameworks are required for proper functioning of Big Data systems. The use

of Lambda architecture can putrefy the Big Data problem into three layers such as serving layer, batch layer, and speed layer.

2. Big Data applications must support a variety of analytical methods, making the application perform complex tasks.

3. Having tools that can perform different task for any given data sets.

4. Getting the right data to analyse

5. The Big Data processing must be distributed across multiple clusters.

6. The data sets must be distributable for in-memory storage.

A proper coordination needs to happen between the processing and the data sets used for processing.

APPLICATIONS OF CLUSTERING

Clustering has a large number of applications that can be applied in various domains. Some of the most popular applications of clustering are as follows:

- Uses of clustering in Banking: It is possible to understand the financial transactions based on historical account transactions. Clustering helps 14 in predicting the customer grouping of behavior and can predict the type of customers seeking for loan or deposits.

- Uses of clustering in Gene Analysis: Gene expression clustering is one of the most useful technique in gene analysis. It is observed that a human contains approximately 20000 genes that can be stored with the in 3.3 GB of memory [38]. Similarly Marbled lungfish, the largest vertebrate known contains genes that can fit in the disk sized 130 GB [39]. With such vast amount of data, there is a strong need of grouping of similarly related genes within the living organism and also to cluster related genes among the living things to group common functionality.

- Uses of clustering in data mining: Clustering in data mining is the grouping of a particular set of entities based on the characteristics separating them according to their similarities. Clustering plays a key role in extracting the information from web and system logs. Clustering documents have become a challenging issue due to its unstructured data format.

- Uses of clustering in weather forecasting: Weather forecasting is the prediction of the atmosphere at a particular place using scientific knowledge to make weather observations. Clustering weather stations by historical temperature data can be useful in discovering and predicting the climate changes and thereby reducing huge loss due to drastic climate changes.

- Uses of clustering in Machine learning: cluster analysis can be explained as set of observations into subsets, such that the observations within the same cluster are closer to each other based on some predefined criteria. Machine learning combined with big data has already started to gain attraction among several researchers. Clustering related to these datasets can play a vital role in identifying the patterns and grouping them according to some similarity.

- Uses of clustering in Image segmentation: Clustering of pixels in images can be useful in identifying similar and different patterns, which can be helpful in medical imaging. Face recognition, video surveillance and fingerprint recognition are some of the applications of clustering images.

- ***Association Rule Learning:*** In data mining, purpose of association rule learning is to discover or extract interesting links between data items from large databases. The common 44 algorithm is Apriori algorithm. Let us take example of supermarket where data is gathered about the purchasing habits of customer. Association rule as the name suggests is the process of determining the rules within the observed data result for any classification. These rules are derived on the bases of the predictions and classification values. They describe the relation between the several attributes and their complementing nature. The rules are used to make the target more specific for achieving more desired results with enhanced efficiency and improved output. The rule is defined with two parts one antecedent and other consequent. The consequent is dependent on the antecedent of the rule and is formed on its basis. These rules are then can be used to derive more patterns and facts and some more rules from the new data.

- ***Stream Data Mining:*** The stream data mining is the concept in which the dynamic data is to be analyzed and is used for business and marketing

purposes. The data collected from the internet, clicking data etc. is being generated day by day thus to analyze this data is one of the difficult task as the rate of flowing of data is different. There are very few techniques to analyze online streaming data.

- **Memory Allocation Techniques:** Memory allocation may be defined as a process of assigning memory blocks to the processes on their request. The operating system is responsible for allocating the memory to the process allocator. The allocator on receiving the memory blocks (which are typically very large in size) divides it into several small size partitions so as to fulfill as many requests as possible. The allocator is also responsible for returning the unused blocks for future usage. There are several types of memory allocation techniques. Some of them are given below:

CONCLUSION

As the innovative progressions are at its apex, an ever increasing number of associations are endeavoring hard towards accomplishing a future evidence demographic. Then again, the cutting edge buyers are misusing the advantages of this innovative change for their prosperity. Because of this innovative change in outlook, monstrous data is persistently getting generated. So as to deal with this data for examination and drawing surmisings, it is basic to comprehend the need to adequately store the data. There are a few issues examined here in this proposal business related to this big data. First and the most significant is unstructured nature of the data. On the off chance that the data has some predefined organization related with it, it ends up simpler to deal with such data. Be that as it may, over 80% of the data that is generated or created today is to a great extent unstructured and the most noticeably awful part is that there are no regular answers for handle such unstructured datasets. The expense brought about in mapping and changing these unstructured datasets to the organized datasets is past satisfactory breaking points. Thusly, new instruments must be contrived to deal with such huge datasets. Since the data is generated from a few random sources, the authenticity of the generated data gangs significantly bigger dangers.

REFERENCES

1. Agrawal, D., Das, S., & El Abbadi, A. (2011, March). Big data and cloud computing: current state and future opportunities. In *Proceedings of the 14th International Conference on Extending Database Technology* (pp. 530-533).ACM.

2. Barnaghi, P., Sheth, A., & Henson, C. (2013). From data to actionable knowledge: Big data challenges in the web of things [Guest Editors' Introduction]. *IEEE Intelligent Systems*, *28*(6), pp. 6-11.

3. Cai, H., Xu, B., Jiang, L., & Vasilakos, A. V. (2017). IoT-based big data storage systems in cloud computing: perspectives and challenges. *IEEE Internet of Things Journal*, *4*(1),pp. 75-87.

4. Davenport, T. H., Barth, P., & Bean, R. (2012). How big data is different. *MIT Sloan Management Review*, *54*(1), pp. 43.

5. Faraway, J. J., & Augustin, N. H. (2018). When small data beats big data. *Statistics & Probability Letters*, *136*, pp. 142-145.

6. Feinleib, D. (2013). *Big Data Demystified: How Big Data is Changing the Way We Live, Love, and Learn*. Big Data Group.

7. Gu, M., Li, X., & Cao, Y. (2014). Optical storage arrays: a perspective for future big data storage. *Light: Science & Applications*, *3*(5), pp. e177.

8. Jara, A. J., Ladid, L., & Gómez-Skarmeta, A. F. (2013). The Internet of Everything through IPv6: An Analysis of Challenges, Solutions and Opportunities. *JoWua*, *4*(3), pp. 97-118.

9. Cai, H., Xu, B., Jiang, L., & Vasilakos, A. V. (2017). IoT-based big data storage systems in cloud computing: perspectives and challenges. *IEEE Internet of Things Journal*, *4*(1), pp. 75-87.

Corresponding Author

Raju G.*

Research Scholar, SVU

graju.mtech@gmail.com