Journal of Advances and Scholarly Researches in Allied Education
Vol. 16, Issue No. 6, May-2019, ISSN 2230-7540

Conceptual Study on Utilization of Big Data Processing Techniques

Upendra Singh Aswal[1]* Sanjeev Kukreti[2] Pareshwar Parsad Barmola[3]

[1,2,3] Graphic Era (Deemed to be University) Dehradun Uttrakhand

Abstract – With the advent of the digital world, there data is growing day by day due to the usage of internet and many devices such as smart phones, laptops, personal and machines increasing at a brisk pace. Big data is the buzzword that would improve and make the data available in three different formats such as structured format, unstructured format and semi-structured format. In this digital period of big data, there are huge chucks of data made available at work, life and study and even for the development of the nation's economy. Big data has enormous data that is gathered from various sources such as social networking sites like Facebook and the conventional data is gathered from the online shopping sites. The big data would store this data in the distributed architecture of the framework. Hadoop is an open source framework that would allow you to create distributed applications to process huge chucks of data. In the recent times, there is quite developed in the hotspot which is grabbing the attention of educational institutions, industries and governments globally. The nature and science have come up with various challenges to find out the opportunities that are put before by the big data. The research study would majorly focus on the big data, techniques and its concepts clearly.

Keywords: Big Data, Processing Techniques, Technology, Applications etc.

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

I. INTRODUCTION

Big data is the widely used term, since this is transforming the way data is stored and generated. Basically, big data is not about storing huge data, but it also offers many features that are different from the concepts related to huge data and too large data. There are many definitions that are given by experts and researchers about bid data in the literature study that was conducted by them. These definitions will help one to perceive the concept of big data [1].

The present global population would go beyond 7.2 billion and out of which 2 billion people are using the Internet. In addition, around 5 billion people are using the mobile devices. With the progression of the technology, many people who are generating huge chucks of data by using their devices. This is true with remote sensors which would produce heterogeneous data that is either in the structured or unstructured format. This type of data is called as big data. This data is categorized into three different aspects. There include data that is big, the data that is not categorized as the regular relational database and the third is the data that is generated, captured and is processed briskly. Big data is the best business application that is growing at a brisk pace in the IT industry. This has generated a lot of interest in different fields such as manufacturing of the healthcare products, banking transactions, social media and satellite imaging. Conventionally, data that is stored in the structured format would increase the informational contents. The volume of the data would be in the semi structured and unstructured formats. The processing of end to end data would gain the momentum by translating the structured data in the relational systems of the database management and the analytics of the unstructured data.

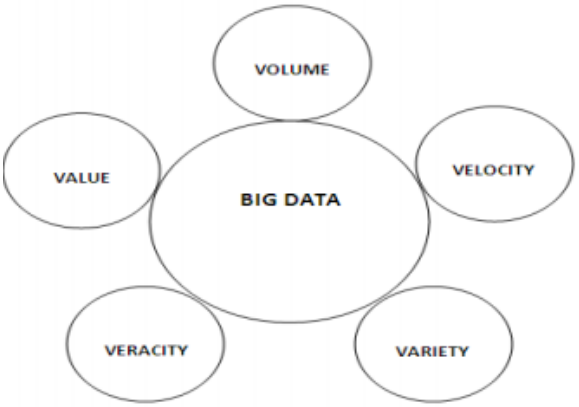

Figure 1: Big Data Structure

Big data have gained a lot of attention in the education industry and IT sector. In this digital and computing world, information is collected at a brisk pace and would exceed the boundary. The present 2 billion people across the globe are connected

with the Internet connection and around 5 billion people are using mobile phones. By the end of 2020, around 50 billion devices are expected to use the Internet connection. Big data, in short BD will its high potential and insights would boost the decision making process. This has grabbed the attention of interest of academics and practitioners. Big Data Analytics (BDA) would increase and have become a new trend where many companies are embracing this technology to extract valuable information with the help of big data. The analytical process would deploy and use the BDA tools where the organizations would use it as a key tool to boost the operational efficiency despite it to be a strategic potential to generate new revenues and gain key advantages over the competitors in the market. There is a wide range of analytical applications that are used. Before you use expensive BD tools, it is important to learn the landscape of BDA and pick the right one that is a perfect fit for your business needs.

A. Big Data Applications

Big data is turning out to be omnipresent. This is used in every area of the business, be it in health or in the living standards, the big data analytics are used widely. To keep it simple, big data is used as a key field which is used as a zone where huge chucks of data is harnessed to take its advantage. The key applications of big data are as follows:

- **Integration**- In this 21st century, when you integrate digital capabilities to take the decision of the company, it would transform the way enterprises are operating. The process of transformation of companies would develop agility, flexibility and precision that would promote new growth. Gartner found that the confluence of the mobile devices, social networks, big data analytics and social networks would be used as a nexus of forces. The mobile and social technologies would change the way people would connect and communication with the companies and incorporate big data analytics in the whole process to prove it to be an advantage for the companies using it. This concept helps the companies to find different ways to leverage the data for boosting the revenue or cutting down the expenses through many are focused on the customer centric outcomes. These objectives are the key concern for many companies. However, this concern has been sorted out by switching to the big data technologies and integrating these technologies to the background operations and internal processes.

- **The Third Eye-** Data visualization: The organizations across the globe are slowly getting to learn the significance of using big data analytics. This helps them to predict the buying behavior patterns that would influence to buy things and detect frauds, which was actually a challenging task until understanding for many companies. However, big data analytics are found to be a solution for it. There are many business experts who get an opportunity to question and analyze the data as per the requirements regardless of the volume and complexity of the data involved. To attain the requirements, many data scientists are visualizing and presenting the data in a comprehensible manner. The many top companies such as Google, Twitter, Facebook, eBay, Wal-Mart and so on embracing data visualization techniques to reduce the complexities involved in handling huge chucks of data. Data visualization is giving many positive outcomes to the businesses. By using data analytics and data visualization, enterprises can top the potential market where there are huge chucks of data to generate huge revenue and attain stability of the business [3].

- **Big data and the food industry:** The use of big data in the food industry is gaining a huge popularity. This is used to track the product quality or to give the recommendations to the consumers based on their preferences or to come up with new marketing strategies that would leave a better experience for the customers. The usage of big data analytics in the food industry is turning out to be omnipresent. The IBM has worked along with the Cheesecake factory to thoroughly analyze the structured data such as location of the restaurant and the unstructured data such as flavors to attain customer satisfaction. The news article states that N2N would be working with IBM to help the Cheesecake factory with the right technology to have a good communication with the supply chain data quickly so that the companies do not need to recall or test the food items. According to Nardone, Starbucks, Dominos and Subway, which are using big data analytics to come up with personalized offers to the consumers. This is helping them to increase the customer base and attain a high level of customer satisfaction.

- **Big data and the finance world:** Big data is a powerful tool that is used to analyze the complicated stock market and its move and help the businesses to take a right business and financial decision. For instance, by doing the intelligent and comprehensive analysis of the big data that you find on the Google trends will help you to predict the stock market in no time.

This is not an accurate technique, but this is a great advancement in the field of finance. The research study that is carried out by the Warwick Business School found from the records of Google, Wikipedia and Amazon Mechanical Trunk in this duration between 2004 to 2012 to analyze the link that exist between the internet searches that is done on politics, business and the moves of the stock market. The paper that is published by the author considered the data extracted from Wikipedia and Amazon Mechanical Trunk. The findings of the study would give great possibilities that would bring changes in gathering the online information related to the politics and business. These are also related to the moves of the stock market. Big data is also used in the Quantitative Investing field where many of the data scientists would ignore the financial training and use computing power to anticipate the securities prices by getting ideas from various sources such as newswires, reports, weather bulletins, Facebook and Twitter. The findings of the study would give concrete proof about the complicated events like the big financial market moves, valuable data that would comprise of the search engine information with keywords that do not have semantic connections with the event. There are many searches that would increase in the information related to the political issues and the businesses would follow the fall of the stock market.

- **Big data in healthcare:** Healthcare is the key area where big data has a huge social impact. From diagnosing the patients with the ailment to the complicated medical research, big data would be used in different aspects. The devices such as Fitbit, Jawbone [5] and Samsung Gear fit [6] would be used by the users to track and then upload the information. The data would be put together and is made available for the doctors, which further allow them to diagnose the patients. There are many partnerships that are available like Pittsburgh Health data alliance which has been started. The Pittsburgh Health data alliance will work along with three of the universities such as Carnegie Mellon University, University of Pittsburgh and the UPMC. The website stated that the healthcare sector would produce a lot of data every day. The big data is the best way to change the landscape of the economy and finance. There are many financial institutions that are using big data policies to give a competitive edge to the rivals in the market. There are many complicated algorithms that are crafted to start doing the trades with the help of structured and unstructured data collected from different sources. The methods that were used till date are not really effective. However, comprehensive research would make sure to make stock market, financial companies and economies use big data analytics.

- **Big data for Telecom industry:** To gain clear cut insights with the help of the machine language that is run on Apache Hadoop will help the operators to take advantages of the datasets and provide quality service and great customer experience besides increasing the customer base by developing the ads that will motivate the target audience and promote the products, which reduces the operational expenses. To boost the customer service and attain high satisfaction, there are various concepts like machine learning and big data are widely implemented. The call detail records, customer service records, emails, web records and geospatial and weather information are key examples of the data that is easily accessible by the telecom operators. This huge chuck of data is hard to handle. The use of technologies will give ample benefits. The predictive maintenance would make sure that the operational disruptions are easy to predict, prevent and retrieve. The real-time processed information would be dynamic and need to have bandwidth cut down the outages as well as congestions.

- **Big Data in Fraud Detection:** Forensic data analytics or FDA is the area that is gaining a lot of prominence since the last decade. There are many companies who are using FDA to dig out the data. The key reasons would be unforeseen circumstances which would change based on the lack of expertise and awareness that would develop key tools to delve for big data due to lack of technological usage or not having the ability to handle huge quantities of data. The usage of big data analytics will help you to fight against various fraud activities. There are many companies which mine big data to keep frauds at bay. However, this has many limitations. There are companies which have silos of the data and limit the data analysis to be carried out. They also take the structured data into consideration to give the subset information. There is a holistic approach to implement and use big data analytics. There are companies like Pactera that will come up with different solutions to produce huge chucks of unstructured and structured data. In addition, they will have different models and algorithms to find patterns for

detecting frauds and anomalies and anticipating the behavior of customers.

II. BIG DATA PROCESSING

There are many different types of big data technologies that have come into the market and these are classified based on the data processing concepts. To process big data, there is a treasure trove of information used and analyzed from the gathered data to meet the requirements of the businesses, political parties and the departments conducting scientific research. The process is started by recovering the data, which would be gathered from different sources such as database, website, documents and CMS (Content Management System). Hadoop is used to store huge chucks of information. Prior to gathering huge information, it is important to record this from the sources which create information. In addition, it has to be filtered and optimized. The relevant information has to be recorded for different channels that would ignore the unwanted information [7]. There are work specific instruments that are made use, for instance, ETL. This is a method that would gather data from various systems and this gathered data would be uploaded into the data warehouse.

Table 1: ETL Stages in Approach

Stage	Stage Description
Transformation	At this level, the guidelines or philosophies those are implemented to the removed data loads into the end goal. If any kind of data will not end any modification whatever that kind of data is identified as “immediate move” or “go through information”
Loading	The loading stage loads huge volume of data loaded in a short period and should be optimized for better performance.
Extraction	Extraction is the initial segment of an ETL procedure which includes extracting the information from the source framework. In separating information accurately sets the phase for the achievement of ensuing procedures. The majority of the undertakings are to consolidate information with a few distinctive source frameworks.

B. Big Data Processing Technologies

- Schema-less databases: Scheme-reduced kind of databases are also termed as NoSQL databases. Database offers an approach for storing and extraction of information which is exhibited in other scenarios other than that of tabular information that is described in the relational databases. There exist two kinds of database one is like the document store and the other is key value storing which store and extracts huge amount of information of unstructured, structured or semi-structured information.

- Column- oriented databases: In this kind, databases stores information in the scenario of columns other than that in rows that is utilized to easily compress huge information and quick queries too.

- Map Reduce: This is a kind of programming pattern that enables performance enhancement opposing to hundreds and thousands of servers and server classifications for huge tasks. This technology involves of two tasks where the one is map task and the other is reducing task. In the first one, the input dataset is transformed as various crucial either pairs or tuples, while in the other multiple forms of output is assimilated to form a decreased array of tuples.

- Chukwa: This procedure observes extensive distributed system and it consolidates necessary semantics for log assortments and it makes use of end to end delivery approach.

- Hadoop: It is the most prevalent open source tool for managing large information and enforced in Map Reduce. It is a java dependant programming system that handles extensive information in allocating computing. Clusters those are in Hadoop make use of either slave or master structure. Distributed file scenario if this approach assists to distribute information in rapid circumstances. When there happens a situation of node failure, then a distribute file system then enables the whole system to operation on normal condition. Hadoop has the other two crucial sub-projects where those are Map Reduce and Hadoop Distributed File System (HDFS).

- Hive: It is like a data warehousing system that is developing using Hadoop. It has various storage kinds like plain content, RC file, ORC, Hbase and many other types. Existing user-defined operations are in utilization to manage dates, data mining tools and strings. It is like a SQL bridge that enables BI application to process various queries and Hadoop classifications.

- HDFS: It is the file system that encompasses the whole nodes in Hadoop collections for extensive storage of data. It connects the whole file systems as one on local node and let it appear as big file system. To stay back from the node disconnections, HDFS augments the protection by representing information all across various sources.

- Hbase: It is an ascendable distributive database system that makes use of HDFS for data storage. It assists both the structured information and column-oriented database.

♦ Storage technologies: To store extensive information, effective and well-organized methods are needed. The important observation of these methodologies is on data compression and storage virtualization.

III. CHALLENGES FACED IN BIG DATA PROCESSING

There exist multiple contests in connecting the possibility of extensive information these days, initiating from the outline of processing networks at the lower phase to investigation means at the uppermost level, together with sequence of open issues in scientific investigation. Of all these contests, few of them are because of the features of extensive information, few of the other by present assessment approaches and procedures and few of the other by the restrictions of present information processing systems. In this segment, we shortly explain the important subjects and challenges.

♦ Computational complexity: The three important characteristics are of extensive information, especially various sources, masa volume and quick altering will make it more complicated for conventional computing approaches (like machine learning, data extraction and information mining) to efficiently assist the procedure, investigation and analysis of huge information. Those kinds of measures are not easily dependant on previous statistics, inspectional aspects and recurrent procedures those are utilized in conventional procedures for managing just minimal amount of information. Modern tendencies are required to stay back from all the expectations those are in conventional computations those are dependent on the self-governing and similar allocation of information and suitable specimen for producing dependable statistics. When answering for these issues consisting of mass information, we are supposed to re-assess and know about its computability, processing ability and various procedures. To resolve the computational intricacy of mass informative applications, we are supposed to show more concentration on the entire life cycle of huge data applications so as to learn data prominent computing patterns those are dependent on the features of mass data. We are required to away from the conventional computing-dependant patterns and develop data centric push-system computing methods and know about feeble CAP system that is of shared and its arithmetical computing approach. We also require to establish systems for both distributed and running computing and establish an extensive information adapted computing outline where storage, communication all are better assimilated and adjusted.

♦ Processing issues: Processing extensive amount of information requires long periods. To know about the suitable and exact technology, the entire information is supposed to be analysed that might not be conceivable. The development of an index while receiving the information might decrease the processing period. For instance, think that Exabyte of information is required to be handled with. The processor enlarges 100 instructions in a single block at 5GB and this process will be for a period of 20 nanoseconds. An Exabyte of information will require modern approach techniques so as to offer periodically and operative data.

♦ Management Issues: This topic is more focused on the construction of a database for handling with comprehensive and mass amount of information. Management issue is the main obstacle in the process of handling mass amount of information. Information was allocated geographically and handled by various organizations. The sources of information might be different in both the aspects of spatially and non-permanently. The manual approaches are utilized as demanding protocol so as to make sure of the enhanced precision and rationality. The ignorance of any complicated open source dais offers best solution for the problem in an independent manner.

♦ Privacy and Security: Initially, the primary issues those are in mass information are of privacy and protection. Huge amount of data will be collected from multiple sources, where it involves of secrecy like credit card information, personal info and other sensible information. So the data has to be more protected for association and harmonization to allocate roles among various enterprises to ensure that

answerability is obligatory for the using information. Most of the online sources and communal systems need collaborative private data which is more exterior to the record level of accessibility.

- Data complexity: The appearance of huge information has shown us with unparalleled huge-level of instances when in discussion about computational issues, even we are facing more complicated data objects. As stated earlier, the distinctive features of massive information are of many kinds and forms, intricated inter-connections and extensively assorted information quality too. The intrinsic difficulty of massive data (consisting of difficult types, complicated systems and complicated outlines) shows its discernment, illustration, awareness and processing for more contesting and outcomes in strident enhances the processing difficulty than that of conventional processing approaches those are dependent on the entire information. Conventional data examinations and mining approaches like extraction, discovery of subjects, semantic investigations and emotional assessments, appear as extreme complicated when utilizing huge information. The basic issue is how to articulate or quantitively define the required features of the complication involved in massive information.

- The essential goal of the information illustration is to represent the data in an efficient approach for the computer assessment or any other kind of user understanding. However, the datasets seem to be inappropriate; the exact information looks to be futile in data analysis.

- Storage and Transport issues: Information is developed by any person at any location. As because the data detonation a modern storage medium is necessary. The information that is more than terabytes, when processed through present scenarios, then this will create an engulf in the network communication.

IV. BIG DATA AND INTERNET OF THINGS

IoT is not just the crucial source of massive information, but even the essential markets of massive data applications also. IoT is in lieu of the upcoming extensive wave in the progression of internet. IoT correspondent technologies are already into various kinds of domains. For instance, agricultural related enterprises are looking after their harvests in real time to enhance the quality of production and to protect the resources those are essential for farming, consisting of insecticides, manures and water too. Utility related organizations have enforced smart meters to observe the utilization of gas, water and energy and municipalities are developing “smart city” projects so as to assist for traffic crowding, enhance waste management, look over extraction of radiation from cell phones, and regulate traffic lights [11].

V. GROWTH AND DEVELOPMENT IN BIG DATA PROCESSING

The progressive technology of massive information broadcasted in “2012 Hadoop and Big Data Technology Conference” demonstrated mainly three topics and those are of information resources, massive info security concerns and assimilation of massive information and cloud computing [12]. The journal publishing supervisor Wired, Chris has declared that the information has developed the conventional scientific approach outdated. Even this declaration seems to be somewhat risky, but massive information has altered the lives of many people, and their thoughts too. Big data is utilized to the most extent. In the view of financial progress, most of the huge organizations concentrate on big data extremely. IDC’s statement stated that international information will enhance almost 50 times for the next 10 years. Oracle’s Chief, Mark Kurd, claimed that currently it is the epoch of massive information explosion and data raised at a frightening rate. Currently, the entire world information of almost million trillion. Information has enhanced 8 times from the period of 2005- 2017 and the anticipated data rate might reach to 35 million trillion. These days, most of the massive organizations utilize big data to ease many procedures and develop competences like Microsoft, Apple, Oracle, Amazon, Google, Facebook and Twitter.

VI. CONCLUSION

Big data is a developing sector where most of the analysis and investigation need to be done with. In the latest financial years, information is developed at a dramatic scenario. Investigation of this information seems to be challenging for the common person. At the final section of this article, we investigate on multiple challenges, complications and equipment those are utilized to investigate massive information. Every big data platform holds corresponding operation. Various methods those are utilized for the inspection consist of statistical methodologies, machine kind of learning, information mining and intelligent assessment and information streaming techniques. We have a faith that in the upcoming day’s investigators will pay more consideration for these procedures in order to solve complication those are

involved in big data in an efficient and effective scenario.

REFERENCES

1. Church, A.H. and Dutta, S. (2013). The Promise of Big Data for OD: Old Wine in New Bottles or the Next Generation of Data-Driven Methods for Change. OD Practitioner, 45, pp. 23-31.

2. Mayer-Schönberger, V. and Cukier, K. (2013). Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt, Boston.

3. Hoffmann, L. (2013). Looking Back at Big Data. Communications of the ACM, 56, pp. 21-23. http://dx.doi.org/10.1145/2436256.2436263

4. http://www.forbes.com/sites/bryanpearson/2015/04/10/exercise-inservice-fitbit-omni-channel-begs-for-omni-prescience/

5. http://www.engadget.com/2015/04/10/jawbone-up3-shipping-april20th/

6. http://www.samsung.com/uk/consumer/mobiledevices/wearables/gear/SM-R3500ZKABTU

7. Kaki, Gowtham, et. al. (2016). "Safe Memory Regions for Big Data Processing." transfer (successor ID, t, out List) pp. 17 -18.

8. Big Data Black Book: Covers Hadoop 2, Map Reduce, Hive, YARN, Pig, R and Data Visualization by DT Editorial Services Paperback, 2016.

9. Puneet Singh Duggal, Sanchita Paul (2013). Big Data Analysis: Challenges and Solutions‖, International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV

10. Bhandari, Renu, Vaibhav Hans, and Neelu Jyothi Ahuja (2016). "Big Data Security–Challenges and Recommendations." IJCSE pp. 93- 98.

11. Gandomi and M. Haider (2015). Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, 35(2), pp. 137-144.

12. K. Kambatla, G. Kollias, V. Kumar and A. Gram (2014). Trends in big data analytics, Journal of Parallel and Distributed Computing, 74(7), pp. 2561-2573.

Corresponding Author

Upendra Singh Aswal*

Graphic Era (Deemed to be University) Dehradun Uttrakhand

aswal.upendra2010@gmail.com