A Review on Big Data Concepts and various Analytic Techniques

Kawale S. M. [#1], Dr. Holambe A. N.[*2], Bokefode J. D.[#3]

[#]Lecturer & Department of computer science, college of engineering (poly), Pandharpur Solapur, Maharashtra, India.
[*]HOD & Department of computer science, college of engineering, Osmanabad Osmanabad, Maharashtra, India.
[#3]Asst. Prof. & Department of computer science, college of engineering, Pandharpur Solapur, Maharashtra, India

Abstract
The 'Big Data' is the rapidly growing and modern technique to collect, persist, share, supervise and examine large sized datasets which comes with high speed and having different structures. Big data datasets are those that exceed the capacity of simple kind of database and data management architecture used in earlier days. Data may be structured; unstructured or semi-structured which needs more computing power to gather and analyze data collected from different sources. Big data can manage variety of data such as structured, semi-structured and unstructured data. Structured data means those data that formatted in straightforward manner according to the database management system. Semi-structured and unstructured data contains all type of unformatted data such as multimedia and social media content. Big Data require new architecture to manage data, new techniques and algorithms to retrieve data and analytics to discover hidden knowledge from it because large data sets having wide range, variety, and difficulty. This paper clarifies the big data and their related terms such as big data analytics, explore the possibilities about future research and present the in progress research and related findings that could help research scholars', businesses and data service providers to study and develop big data analytics projects. Now a days, most of the enterprises are investigate big data to improve the organization position in current market trends.

Keywords: *Big Data, Analytics, MapReduce, HDFS.*

I. INTRODUCTION

Today, every field is based on digitization and it grows exponentially. Due to the high growth in digitization huge amount of structured as well as unstructured data was generated and process is going on continuously. The data is being generated and collected from different sources such as, transactions, social media, sensors, retails, audios, videos, government sectors etc. For example, in facebook every month 40 billion contents are being shared. It is necessary for organizations to mine this data continuously to survive in current market trends and become a good competitor. When data is analyzed properly it helps the organizations to define current and future strategies. The conventional data processing techniques gives degraded performance while creating, managing and analyzing big data. Hadoop provides platform for structuring and managing Big Data, and making it useful for analytics purposes. Big data analytics is important and advanced analytic techniques which operate on big data for examining large amounts of data. In analytics, data divided into different sectors to assess it according to time, and compare one sector to another. With the help of big data enterprises can develop a more systematic and perceptive understanding of their business, which helps to increase the productivity and innovation.

A. Definition

Big data can be referred as data sets or collection of data sets whose have high velocity, size and intricacy, which make them difficult to manage and process with traditional technologies and tool and also difficult to capture it with high data rate and difficult to perform analytics using relational databases and statistical or visualization techniques[1]. Depending on the size of data sets a particular data set is considered as big data, data sets having size from 40-50 terabytes to multiple petabytes. Big Data System having layered architecture and it having three layers. These layers are Infrastructure Layer, Computing Layer, and Application Layer it can be shown in figure 2.

B. Describing Big Data via the Three Vs

Volume of data:

Volume can be referred as the size of data. Large amount of data is collected from different sources such as, transactions, social media, sensors, retails, audios, videos, government sectors etc. It ranges from terabytes to petabytes.

Variety of data:

Variety of data means type of data to which Big data support. Big data supports different types of data such as structured, unstructured and semi structured.

Fig.1 Three Vs of Big Data

Velocity of data: Velocity considered as the speed of data capturing, processing and visualizing it. Many time-sensitive areas big data plays very important role[4].

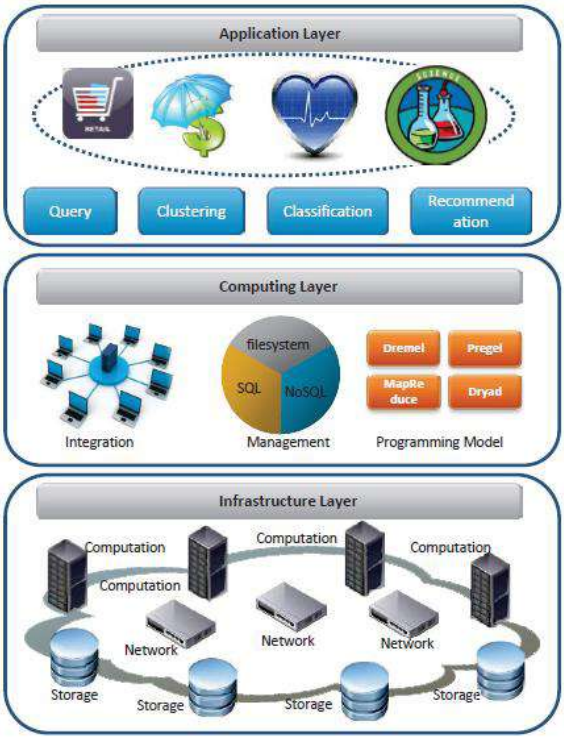

Fig. 2 Layered Architecture of Big Data

II. CHALLENGES IN BIG DATA

A. *Heterogeneity and Incompleteness*

Data collected from different sources are heterogeneous. Data analysis techniques require same type (structured) and complete data for visualize it in understandable manner. There for data must structure carefully before data analysis. First challenge is efficient representation and collection of heterogeneous data. Hadoop give support for processing heterogeneous data and analysis of this data[2][5].

B. *Scale*

Data collected for analysis is very big; managing these data requires scalable computing power, high speed sensors, strong network, and huge storage capability.[2][5]

C. *Timeliness*

The information is being generated from various sources needs to process before analyse it. Therefore it requires more time to analyse it [5].

D. *Privacy*

The confidentiality issue of data is more important in the context of Big Data. Managing privacy is must be addressed while managing big data [5].

III. CHALLENGES IN BIG DATA

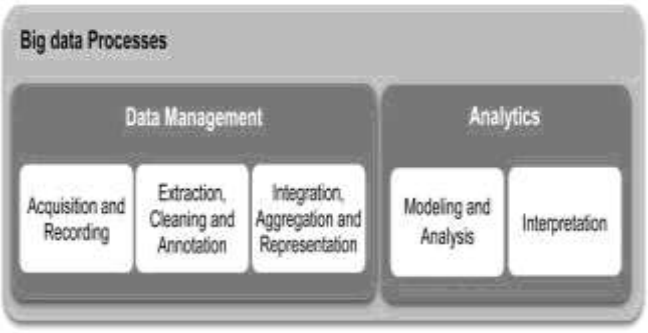

Fig. 3 Processes for extracting insights from big data.

a. *Data Acquisition and Recording*

Big Data has been generated from different data capturing sources. For example, simulation and different scientific experiments easily generates peta bytes. Most of these data is not useful; it needs to be filtered. The first challenge is, data needs to be filtered in such way that important data will not be loose. The second challenge is, generate the correct metadata for stored data.

b. *Information Extraction and Cleaning*

Whatever information was collected from different sources is not in required format for analysis. This information needs to be cleaned and arrange into proper format for analysis. For that require an information extraction tools, that fetch the required data from the essential sources.

c. *Data Integration, Aggregation, and Representation*

Big data is heterogeneous in nature, it is not easy to store it and lob it into a repository. This data needs to structure carefully so that it will be useful for data analysis. Effective management, representation, data access policies must be considered.

d. Query Processing, Data Modelling, and Analysis

Different methods are available for mining knowledgeable data from big data. Querying Big Data is different from traditional techniques because it is heterogeneous, dynamic and inter-related. Querying or Mining Big data requires integrated and efficiently accessible data techniques and scalable mining algorithms.

e. Interpretation

Analysing Big Data is not having value if analytical information is not presented in user-friendly manner. This information must be interpreted with proper visualization and clear specification. With this interpretation additional guidelines or information must be provided for better understanding. This additional information is considered as provenance of the data.

IV. BIG DATA ANALYTICS

Big data have large volume of data. Organizations require efficient algorithms or techniques to processes high volume of data to turn it into useful information. Data is high velocity and unstructured it has to be analysed. The process for extracting meaningful data from big data is performed in five stages, shown in Fig. 3. These five stages again categorized in to two sub processes, first is data management and second are analytics. The data management entails procedure and different techniques to be required for storing data and to arrange it in proper way to perform analysis on it. The Analytics involves different techniques used for analysing data and to retrieve meaningful insights and intelligence from big data. Subsequent sections illustrate the different analytical techniques used for structured as well as unstructured data[6].

a. Text analytics

Text analytics is a technique used for retrieving meaningful data from textual data. Textual data held by businesses, Social network sides, email logs, online application and forums, educational documents, news channels, and call centre logs these are sources of textual data. In text analytics, contains three main stages statistical analysis, computational linguistic, and machine learning. Text analytics facilitate businesses and organization to retrieve meaningful summaries from large volumes of generated text, which helps for decision-making.

b. Sentiment Analysis

Sentiment analysis helps businesses to determine the sentiments from their customers regarding the products. Sentiment analysis techniques can be categorized into three subgroups namely document-level, sentence-level, and aspect-based. In Document-level techniques, documents are inspected for positive or negative sentiment [3]. In Sentence-level techniques, collected sentences are scanned for the polarity to known the present entity. In Aspect-based techniques, documents are determined for sentiments and entities aspects are identified to clarify to which each sentiment refers.

c. Audio analytics

Audio analytics are applied to speeches or spoken audios. These techniques also called as speech analytics. Now days, audio analytics plays vital role in call centres and healthcare system. All these techniques help to evaluate agent performance, to improve sales rate, to understand customer behaviour and to identify and solve product related issue [4].

d. Video analytics

In Video analytics, video streams are analysed for meaningful information .This analysis also known as video content analysis (VCA).To provide security and surveillance over the place video analytics is used. For instance, in YouTube daily countless videos are uploaded and viewed. To understand the user behaviour and retrieve meaningful insights from video different analytics techniques are used [4].

e. Social media analytics

Social media analytics firstly used in telecommunications industry, and then it is adopted by sociologists to understand interpersonal relationships. This analytics is used for analysing relationship between peoples working in many fields. Social media is gathered from different social sites such as Facebook, reedit, and blogs [7]. To extract information from the structure of social network different techniques are introduced such as, Community detection, Social influence analysis, Link prediction [5].

f. Predictive analytics

Predictive analytics is nothing but forecasting about future outcomes from current and historical data. These probabilities used to plan businesses and according to it work has been done [9]. Predictive analytics is used to understand the customer's future requirement, to design products according to the market trends and to identify probable risk and scope for organizations. Predictive Analytics performed through different techniques but one of the famous techniques are Machine learning, fuzzy logic, data mining and regression analysis which help analysts to make hypothesis regarding business to improve the position of an organizations [6].

V. BIG DATA ANALYTICS SOFTWARE

Apache Hadoop is one of the best and famous platform used for big data processing[8].It is "an open source software project that based on

distributed processing over the big data stored in servers"[11][12]. Design of Hadoop is flexible and scalable which scales according to the requirement and scales up to thousands of server. It provides high degree of fault tolerance. Apache software foundation takes initiative to design software which handles large volume of data. It handles all type of data. Hadoop platform mainly categorized in to two projects MapReduce and HDFS [10]. MapReduce framework assign works to different node clusters whereas HDFS (Hadoop Distributed File System) link node clusters to local nodes to make single Big File system[13].

VI. CONCLUSION

This paper, describes the Big data and its related basic concepts and identified challenges related to it and if organizations need to meet current market trends. They must have to collect huge amount of data and need to implement high processing capabilities to process these data then these data can be refined using different analytics techniques for correct decision making and strategic planning. Hadoop provide flexible platform to process and analyze Big data.

REFERENCES

[1] Advancing Discovery in Science and Engineering. Computing Community Consortium. Spring 2011.

[2] Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data.Proceedings of the VLDB Endowment, 5(12), 2032–2033.

[3] Feldman, R. (2013). Techniques and applications for sentiment analysis. Communi-cations of the ACM, 56(4), 82–89.

[4] Amir Gandomi, Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics.International Journal of Information Management,ScienceDirect.

[5] M.M. Anwar, M.F. Zafar, Z. Ahmed. A proposed Preventive Information Security System. IEEE International Conference on Electrical Engineering, April, 2007.

[6] MacDonald, Neil, 2012, Information Security is Becoming a Big Data Analytic Problem, Gartner, (23 March 2012), DOI= http://www.gartner.com/id=1960615

[7] Larry Barrett, "Big data analytics: the enterprise's next great security weapon?" February 2014. [14] http://www.edupristine

[8] G. Noseworthy, Infographic: Managing the Big Flood of Big Data in Digital Marketing, 2012 http://analyzingmedia.com/2012/infograp hic-big-flood-of-big-data-in-digitalmarketing.

[9] H. Moed, The Evolution of Big Data as a Research and Scientific Topic: Overview of the Literature, 2012, Research Trends, http://www.researchtrends.com .

[10] A Navint Partners White Paper, "Why is BIG Data Important?" May 2012, http://www.navint.com/images/Big.Data.pdf

[11] Greenplum. A unified engine for RDBMS and Map Reduce,2009. http://www.greenplum.com/resources/mapreduce/.

[12] Oracle Information Architecture: An Architect's Guide to Big Data, An Oracle White Paper in Enterprise Architecture August 2012

[13] http://bigdataarchitecture.com/

[14] http://www.informationweek.com/softw Database Systems Journal vol. III, no. 4/2012 13 are/business-intelligence/sas-gets-hip-tohadoop-for-big-ata/240009035-pgno=2

[15] http://en.wikipedia.org/wiki/Apache_Hadoop.