DATA MINING WITH BIG DATA ANALYTICS

K.SARANYA
Department of Computer Science and Engineering,
Nadar Saraswathi College of Engineering and Technology,India.
saraninnovator@gmail.com

***Abstract*— Big data is the term for a collection of data sets which are large and complex, it contain structured and unstructured both type of data. Data comes from everywhere, sensors used to gather climate information, posts to social media sites, digital pictures and videos etc .,This data is known as big data. Useful data can be extracted from this big data with the help of data mining. Data mining is a technique for discovering interesting patterns as well as descriptive, understandable models from large scale data. Big data analytics is the process of examining large data sets containing a variety of data types to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits. In this paper we overviewed types of big data and challenges in big data analytics for future.**

Keyword- component, formatting, style, styling, insert

I. INTRODUCTION

The term 'Big Data' appeared for first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the NextWave of InfraStress". Big Data mining was very relevant from the beginning, as the first book mentioning 'Big Data' is a data mining book that appeared also in 1998 by Weiss and Indrukya . However,the first academic paper with the words 'Big Data' in the title appeared a bit later in 2000 in a paper by Diebold .The origin of the term 'Big Data' is due to the fact that we are creating a huge amount of data every day. Usama Fayyad in his invited talk at the KDD BigMine 12 Workshop presented amazing data numbers about internet usage, among them the following: each day Google has more than 1 billion queries per day, Twitter has more than 250 milion tweets per day, Facebook has more than 800 million updates per day, and YouTube has more than 4 billion views per day. The data produced nowadays is estimated in the order of zettabytes, and it is growing around 40% every year.A new large source of data is going to be generated from mobile devices and big companies as Google,Apple,Facebook,Yahoo are starting to look carefully to this data to find useful patterns to improve user experience. “Big data” is pervasive, and yet still the notion engenders confusion. Big data has been used to convey all sorts of concepts, including: huge quantities of data,social media analytics, next generation data management capabilities, real-time data, and much more.

II. BIG DATA ANALYTICS

Analyzing Big Data is a challenging task as it involves large distributed file systems which should be fault tolerant, flexible and scalable. Map Reduce is widely been used for the efficient analysis of Big Data. Traditional DBMS techniques like Joins and Indexing and other techniques like graph search is used for classification and clustering of Big Data.

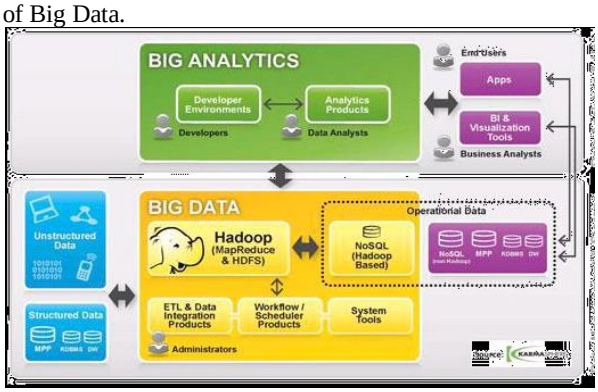

Fig 1. Over view of Big Data Analytics

These techniques are being adopted to be used in Map Reduce. In this research paper the authors suggest various methods for catering to the problems in hand through Map Reduce framework over Hadoop Distributed File System (HDFS). Map Reduce is a Minimization technique which makes use of file indexing with mapping, sorting, shuffling and finally reducing. Keyword-Big Data Analysis, Big Data Management, Map Reduce, HDFSI. Big data analytics must effectively mine massive datasets at different levels in realtime or near realtime - including modeling, visualization, prediction, and optimization - such that inherent promises can be revealed to improve decision making and acquire further advantages.

III. TYPES OF BIG DATA AND SOURCES

There are two types of big data: structured and unstructured. Structured data are numbers and words that can be easily categorized and analyzed. These data are generated by things like network sensors embedded in electronic devices, smartphones, and global positioning system (GPS) devices. Structured data also include things like sales figures, account balances, and transaction data. Unstructured data include more complex information, such as customer reviews from commercial websites, photos and other multimedia, and comments on social networking sites. These data can not easily be separated into categories or analyzed numerically."Unstructured big data is the things that humans are saying,"says big data consulting firm vice president Tony Jewitt of Plano,Texas."It uses natural language."Analysis of unstructured data relies on keywords, which allow users to filter the data based on searchable terms. The explosive growth of the Internet in recent years means that the variety and amount of big data continue to grow. Much of that growth comes from unstructured data.

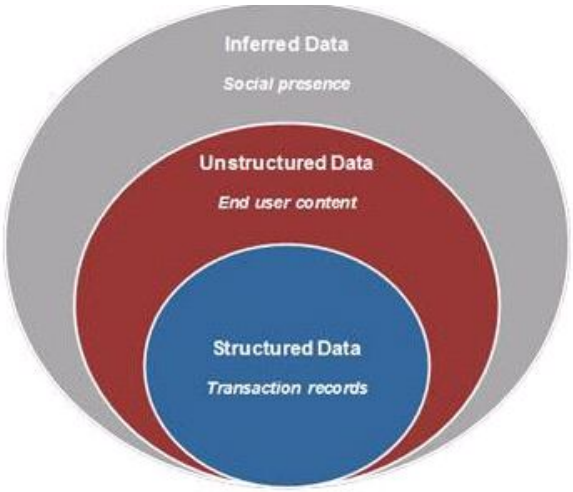

Fig.2. Sources of Big Data

IV. HACE THEOREM

HACE theorem is theorem to model the big data characteristics. Big Data starts with large-volume, Heterogeneous; Autonomous sources with distributed and decentralized control, and seeks to explore Complex and Evolving relationships among data.These characteristics make it an intense challenge for discovering useful knowledge from the Big Data. In a native sense, we can imagine that a number of blind men are trying to size up a giant camel (see Fig. 2), which will be the Big Data in this context. The aim of each blind man is to extract conclusion of the camel according to the part of information he collects during the procedure. Because each individual's opinion is restricted to his native area, it is expected that the blind men will each conclude independently that the camel "feels" like a rope, a stone, a stick, depending on the part each of them is limited.To make the problem even more complex, let us accept that 1) the camel is increasing quickly and its posture varies frequently, and 2) each blind man may have his own information sources that tell him about subjective knowledge about the camel (e.g., one blind man may exchange his feeling about the camel with another blind man, where the exchanged knowledge is intrinsically subjected). Exploring the Big Data in this scenario is equivalent to form various information from different sources (blind men) to help to draw a best possible illustration to uncover the actual sign of the camel in a actual way. Certainly, this job is not as simple as enquiring each blind man to designate his spirits about the elephant and then getting an skilled to draw one single picture with a joint opinion, regarding that each separate may express a different language (varied and diverse information sources) and they may even have confidentiality concerns about the messages they measured in the information exchange procedure.

The key characteristics of BIG DATA analytics are

Big data analysis should be viewed from two perspectives:

a) Decision-oriented

Decision-oriented analysis is more akin to traditional business intelligence. Look at selective subsets and representations of larger data sources and try to apply the results to the process of making business decisions. Certainly these decisions might result in some kind of action or process change, but the purpose of the analysis is to augment decision making.

b) Action-oriented

Action-oriented analysis is used for rapid response, when a pattern emerges or specific kinds of data are detected and action is required. Taking advantage of big data through analysis and causing proactive or reactive behavior changes offer great potential for early adopters.

V. THREE V'S IN BIG DATA VOLUME VARIETY VELOCITY

Big data can be described by the following characteristics

a) Volume

The quantity of generated data is important in this context. The size of the data determines the value and potential of the data under consideration, and whether it can actually be considered big data or not. The name 'big data' itself contains a term related to size, and hence the characteristic.

b) *Variety*

The type of content, and an essential fact that data analysts must know. This helps people who are associated with and analyze the data to effectively use the data to their advantage and thus uphold its importance.

c) *Velocity*

In this context, the speed at which the data is generated and processed to meet the demands and the challenges that lie in the path of growth and development.

d) ***Variability***

The inconsistency the data can show at times—which can hamper the process of handling and managing the data effectively.

e) *Veracity*

The quality of captured data, which can vary greatly. Accurate analysis depends on the veracity of source data.

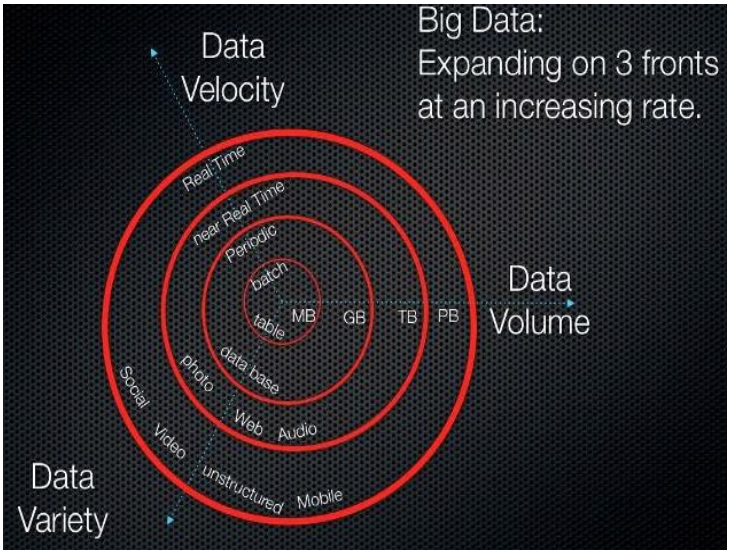

Fig.3. Three V's in Big Data Management

VI. DATA MINING WITH BIG DATA ANALYTICS

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational database.

Data mining as a term used for the specific classes of six activities or tasks as follows:

a) ***Classification***

A classifier is a tool in data mining that takes a bunch of data representing things we want to classify and attempts to predict which class the new data belongs to.

b) ***Estimation***

Estimation deals with a model/mechanism is formed, then data analysis is performed in that model which was actually formed from that data itself.

c) ***Prediction***

Prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest.

d) ***Association***

An association rule is a rule which implies certain association relationships among a certain set of objects in a database.

e) Clustering

Clustering is the task of segmenting a diverse group in to a number of similar subgroup or cluster. In, clustering there are no predefined classes.

TABLE 1

Difference between Big data analytics and Data mining

Big data analytics	Data mining
Big data is a term for large data set.	Data mining refers to the activity of going through big data set to look for relevant information.
Big data is the asset.	Data mining is the handler which provide beneficial result**.**
Big data varies depending on the capabilities of the organization and on the capabilities of the applications that are traditionally used to process and analyze the data.	Data mining refers to the operation that involve relatively sophisticated search operation.

VII. CHALLENGES IN BIG DATA ANALYTICS

Often Big Data Analytics has been talked about as a “problem” because it couldn’t be easily processed with traditional systems based on relational databases, but it really is a tremendous opportunity to enhance and even transform how you run your business. The value of Big Data Analytics can be significant. It can lead to innovations, such as new pricing models, new ways to engage with your customers and partners, new product ideas, or new market opportunities. For example, at a recent conference, one large financial institution estimated that Big Data Analytics could help them reduce by 30- 65% the time to market and cost of their strategic innovation projects.

The top four Big Data Analytics challenges as:

Data integration – The ability to combine data that is not similar in structure or source and to do so quickly and at reasonable cost. With such variety, a related challenge is how to manage and control data quality so that you can meaningfully connect wellunderstood data from your data warehouse with data that is less well understood.

Data volume – The ability to process the volume at an acceptable speed so that the information is available to decision makers when they need it.

Skills availability – Big Data Analytics is being harnessed with new tools and is being looked at in different ways. There a shortage of people with the skills to bring together the data, analyze it and publish the results or conclusions.

Solution cost – Since Big Data Analytics has opened up a world of possible business improvements, there is a great deal of experimentation and discovery taking place to determine the patterns that matter and the insights that turn to value. To ensure a positive ROI on a Big Data project, therefore, it is crucial to reduce the cost of the solutions used to value.

VIII. **FORECAST TO THE FUTURE**

There are many future important challenges in Big Data management and analytics, that arise from the nature of data: large, diverse, and evolving. These are some of the challenges that researchers and practitioners will have to deal during the next years:

a) Analytics Architecture

It is not clear yet how an optimal architecture of an analytics systems should be to deal with historic data and with real-time data at the same time. An interesting proposal is the Lambda architecture of Nathan Marz.The Lambda Architecture solves the problem of computing arbitrary functions on arbitrary data in real time by decomposing the problem into three layers: the batch layer, the serving layer, and the speed layer. It combines in the same system Hadoop for the batch layer, and Storm for the speed layer. The properties of the system are: robust and fault tolerant, scalable, general, extensible, allows ad hoc queries, minimal maintenance, and debuggable.

b) Statistical significance

It is important to achieve significant statistical results, and not be fooled by randomness. AsEfron explains in his book about Large Scale Inference it is easy to go wrong with huge data sets and thousands of questions to answer at once.

c) Distributed mining

Many data mining techniques are not trivial to paralyze. To have distributed versions of some methods, a lot of research is needed with practical and theoretical analysis to provide new methods.

d) Hidden Big Data

Large quantities of useful data are getting lost since new data is largely untagged file based and unstructured data. The 2012 IDC study on Big Data explains that in 2012, 23% (643 exabytes) of the digital universe would be useful for Big Data if tagged and analyzed. However, currently only 3% of the potentially useful data is tagged, and even less is analysed.

IX.CONCLUSION

Big data is the term for a collection of complex data sets, Data mining is an analytic process designed to explore data(usually large amount of data-typically business or market related-also known as “big data”)in search of consistent patterns and then to validate the findings by applying the detected patterns to new subsets of data. To support Big data analytics and data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data. We regard Big data as an emerging trend and the need for Data mining with Big data analytics is rising in all science and engineering domains. With Big data technologies, we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at real time.

REFERENCES

1) Alex Berson and Stephen J.Smith Data Warehousing,Data Mining and OLAP edition 2010
2) Department of Finance and Deregulation Australian Government Big Data Strategy-Issue Paper March 2013
3) Alex Berson and Stephen J. Smith, “ Data Warehousing, Data Mining & OLAP”, Tata McGraw – Hill Edition, Tenth Reprint 2007
4) NASSCOM Big Data Report 2012
5) Jiawei Han and Micheline Kamber, “Data Mining Concepts and Techniques”, Second Edition, Elsevier, 2007
6) Wei Fan and Albert Bifet “ Mining Big Data:Current Status and Forecast to the Future”,Vol 14,Issue 2,2013
7) Algorithm and approaches to handle large Data-A Survey,IJCSN Vol 2,Issue 3,2013
8) K.P. Soman, Shyam Diwakar and V. Ajay “, Insight into Data mining Theory and Practice”, Easter Economy Edition, Prentice Hall of India, 2006
9) Daniel T.Larose, “Data Mining Methods and Models”, Wile-Interscience, 2006
10) Xindong Wu , Gong-Quing Wu and Wei Ding “ Data Mining with Big data “, IEEE Transactions on Knoweledge and Data Enginnering Vol 26 No1 Jan 2014

11) Xu Y etal, balancing reducer workload for skewed data using sampling based partioning 2013.
12) Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner "Decision Trees-What Are They?" 10. Weiss, S.H. and Indurkhya, N. (1998), Predictive Data Mining: A Practical Guide, Morgan Kaufmann Publishers, San Francisco, CA.

AUTHOR PROFILE

I have completed B.Tech Information Technology at Odaiyappa college of Engineering and Technology under the control Anna university Trichirapalli, India. And I did my M.E Computer science at Sethu Institute of Technology under the control of Anna university Chennai, India. At Present I am working as an Assistant Professor in the Department of Computer Science and Engineering at Nadar saraswathi College of Engineering and Technology, Theni

.