International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 3, March 2015
Available at http://internationaljournalofresearch.org

Importance of HACE and Hadoop among Big data Applications

[1]G. Bramhaiah Achary; [2]Dr. P. Venkateswarlu & [3]B.V. Srikanth

[1]M.Tech (CSE), Department of Computer Science & Engineering Nagole Institute of Technology & Science, Kuntloor (V), Hayathnagar (M), RR District, Hyderabad, India.
E-mail id: bramhi.chary@gmail.com
[2]Professor & HOD, Department of Computer Science & Engineering.
E-mail id: venkat123.pedakolmi@gmail.com.
[3]Associate Professor, Department of Computer Science & Engineering.
E-mail id: bvsrikanth123@gmail.com

Abstract:

Big data is a collection of massive and complex data sets that include the huge quantities of data, social media analytics, data management capabilities, real-time data. Big data analytics is the process of examining large amounts of data. Big Data is characterized by the dimensions volume, variety, and velocity, while there are some well-established methods for big data processing such as Hadoop and HACE theorem which uses the map-reduce paradigm and Extract useful information from unstructured big data. Using MapReduce programming paradigm the big data is processed.

Keywords:

Big Data; Parameters; Evolution; Hadoop; HACE

1. INTRODUCTION

With the growth of technologies and services, the large amount of data is produced that can be structured and unstructured from the different sources. Such type of data is very difficult to process that contains the billions records of millions people information that includes the web sales, social media, audios, images and so on. [6] The need of big data comes from the Big Companies like yahoo, Google, facebook etc for the purpose of analysis of big amount of data which is in unstructured form. Google contains the large amount of information .So; there is the need of Big Data Analytics that is the processing of the complex and massive datasets [5]. Big data analytics analyze the large amount of information used to uncover the hidden patterns and the other information which is useful and important information for the use [2].

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 Next Wave 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 Big Mine‟ 12Workshop presented amazing data numbers about internet usage, among them the following: each day Google has more than 1

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 3, March 2015
Available at http://internationaljournalofresearch.org

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, Face book, 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. Whatever the label, organizations are starting to understand and explore how to process and analyze a vast array of information in new ways. In doing so, a small, but growing group of pioneers is achieving breakthrough business outcomes. In industries throughout the world, executives recognize the need to learn more about how to exploit big data. But despite what seems like unrelenting media attention, it can be hard to find in-depth information on what organizations are really doing. So, we sought to better understand how organizations view big data – and to what extent they are currently using it to benefit their businesses.

Big Data Parameters As the data is too big from various sources in different form; it is characterized by the 3 Vs. The three Vs of Big Data are: Variety, Volume and Velocity [15]. Fig: Parameters Variety makes the data too big. Data comes from the various sources that can be of structured, unstructured and semi structured type. Different variety of data include the text, audio, video, log files, sensor data etc. Volume represent the size of the data how the data is large. The size of the data is represented in terabytes and pet bytes. Velocity Define the motion of the data and the analysis of streaming of the data [16].

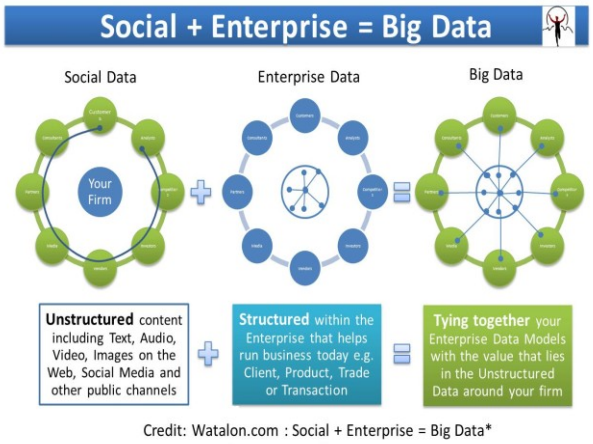

Fig 1. Theme of Bigdata

II. 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, smart phones, 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 cannot 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.

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 3, March 2015
Available at http://internationaljournalofresearch.org

Fig 2. Big data Sources

III. TECHNIQUES AND TECHNOLOGY

For the purpose of processing the large amount of data, the big data requires exceptional technologies. The various techniques and technologies have been introduced for manipulating, analyzing and visualizing the big data [20]. There are many solutions to handle the Big Data, but the Hadoop is one of the most widely used technologies.

A. Hadoop: Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. Hadoop was developed by Google's Mapreduce that is a software framework where an application break down into various parts. The Current Appache Hadoop ecosystem consists of the Hadoop Kernel, Mapreduce, HDFS and numbers of various components like Apache Hive, Base and Zookeeper [17]. MapReduce is a programming framework for distributed computing which is created by the Google in which divide and conquer method is used to break the large complex data into small units and process them. Map Reduce have two stages which are [18]: Map ():- The master node takes the input, divide into smaller subparts and distribute into worker nodes. A worker node further do this again that leads to the multi-level tree structure. The worker node process the m=smaller problem and passes the answer back to the master Node. Reduce ():- The, Master node collects the answers from all the sub problems and combines them together to form the output

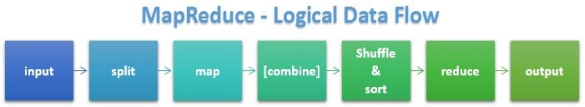

Fig 3. MapReduce logical dataflow

IV. HACE Theorem.

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 extreme challenge for discovering useful knowledge from the Big Data. In a naïve sense, we can imagine that a number of blind men are trying to size up a giant Camel, which will be the Big Data in this context. The goal of each blind man is to draw a picture (or conclusion) of the Camel according to the part of information he collects during the process. Because each person's view is limited to his local region, it is not surprising that the blind men will each conclude independently that the camel "feels" like a rope, a hose, or a wall, depending on the region each of them is limited to. To make the problem even more complicated, let us assume that the camel is growing rapidly and its pose changes constantly, and each blind man may have his own (possible unreliable and inaccurate) information sources that tell him about biased 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 inherently biased).

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 3, March 2015
Available at http://internationaljournalofresearch.org

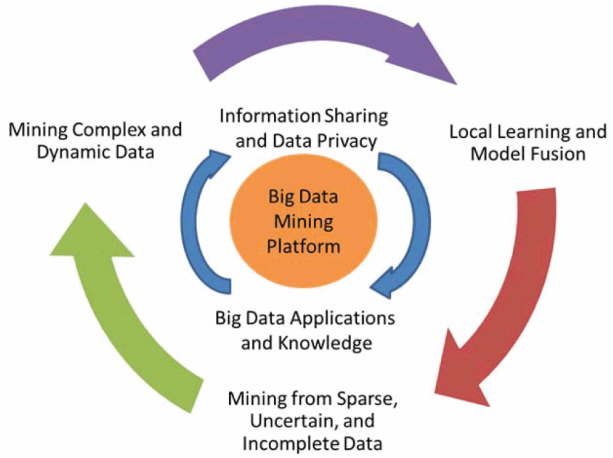

Fig. 4. A Big Data processing framework:

Exploring the Big Data in this scenario is equivalent to aggregating heterogeneous information from different sources (blind men) to help draw a best possible picture to reveal the genuine gesture of the camel in a real-time fashion. Indeed, this task is not as simple as asking each blind man to describe his feelings about the camel and then getting an expert to draw one single picture with a combined view, concerning that each individual may speak a different language (heterogeneous and diverse information sources) and they may even have privacy concerns about the messages they deliberate in the information exchange process. The term Big Data literally concerns about data volumes, HACE theorem suggests that the key characteristics of the Big Data are A. Huge with heterogeneous and diverse data sources:- One of the fundamental characteristics of the Big Data is the huge volume of data represented by heterogeneous and diverse dimensionalities. This huge volume of data comes from various sites like Twitter, Myspace, Orkut and LinkedIn etc. B. Decentralized control:- Autonomous data sources with distributed and decentralized controls are a main characteristic of Big Data applications. Being autonomous, each data source is able to generate and collect information without involving (or relying on) any centralized control. This is similar to the World Wide Web (WWW) setting where each web server provides a certain amount of information and each server is able to fully function without necessarily relying on other servers C. Complex data and knowledge associations:- Multi structure, multisource data is complex data, Examples of complex data types are bills of materials, word processing documents, maps, time-series, images and video. Such combined characteristics suggest that Big Data require a "big mind" to consolidate data for maximum values.

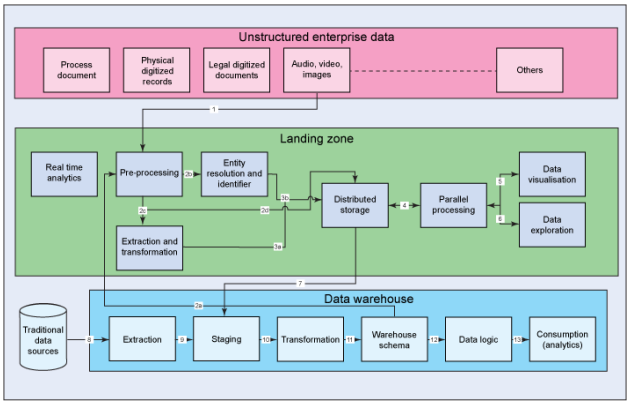

Fig 5.Datamining Process on Unstructured data

V. DATA MINING FOR BIG DATA

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: 1. Classification 2. Estimation 3. Prediction 4. Association rules 5. Clustering 6. Description A. Classification is a process of generalizing the data according to different instances. Several major kinds of classification algorithms in data mining are Decision tree, k-nearest neighbor classifier, Naive Bayes, Apriori and AdaBoost. Classification consists

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 3, March 2015
Available at http://internationaljournalofresearch.org

of examining the features of a newly presented object and assigning to it a predefined class. The classification task is characterized by the well-defined classes, and a training set consisting of reclassified examples. B. Estimation Estimation deals with continuously valued outcomes. Given some input data, we use estimation to come up with a value for some unknown continuous variables such as income, height or credit card balance. C. Prediction It‟s a statement about the way things will happen in the future , often but not always based on experience or knowledge. Prediction may be a statement in which some outcome is expected. D.Association Rules An association rule is a rule which implies certain association relationships among a set of objects (such as “occur together” or “one implies the other”) in a database. E.Clustering Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data.

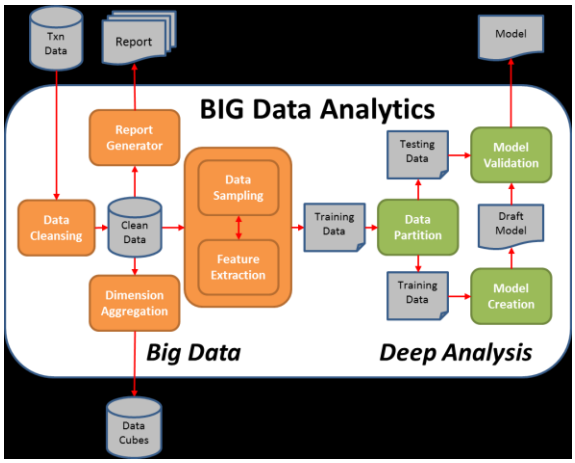

Fig 6. Big data Analytics

V1. CHALLENGES IN BIG DATA

Meeting the challenges presented by big data will be difficult. The volume of data is already enormous and increasing every day. The velocity of its generation and growth is increasing, driven in part by the proliferation of internet connected devices. Furthermore, the variety of data being generated is also expanding, and organization’s capability to capture and process this data is limited. Current technology, architecture, management and analysis approaches are unable to cope with the flood of data, and organizations will need to change the way they think about, plan, govern, manage, process and report on data to realize the potential of big data.

A. Privacy, security and trust:

The Australian Government is committed to protecting the privacy rights of its citizens and has recently strengthened the Privacy Act (through the passing of the Privacy Amendment (Enhancing Privacy Protection) Bill 2012) to enhance the protection of and set clearer boundaries for usage of personal information. Government agencies, when collecting or managing citizens data, are subject to a range of legislative controls, and must comply with the a number of acts and regulations such as the Freedom of Information Act (1982), the Archives Act (1983), the Telecommunications Act (1997) ,the Electronic Transactions Act (1999), and the Intelligence Services Act (2001). These legislative instruments are designed to maintain public confidence in the government as an effective and secure repository and steward of citizen information. The use of big data by government agencies will not change this; rather it may add an additional layer of complexity in terms of managing information security risks. Big data sources, the transport and delivery systems within and across agencies, and the end points for this data will all become targets of interest for hackers, both local and international and will need to be protected. The public release of large machine-readable data sets, as part of the open government policy, could potentially provide an opportunity for unfriendly state and non-state actors to glean sensitive information, or create a mosaic of exploitable information from apparently innocuous data. This threat will

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 3, March 2015
Available at http://internationaljournalofresearch.org

need to be understood and carefully managed. The potential value of big data is a function of the number of relevant, disparate datasets that can be linked and analysed to reveal new patterns, trends and insights. Public trust in government agencies is required before citizens will be able to understand that such linking and analysis can take place while preserving the privacy rights of individuals.

B. Data management and sharing

Accessible information is the lifeblood of a robust democracy and a productive economy to Government agencies realize that for data to have any value it needs to be discoverable, accessible and usable, and the significance of these requirements only increases as the discussion turns towards big data. Government agencies must achieve these requirements whilst still adhering to privacy laws. The processes surrounding the way data is collected, handled, utilized and managed by agencies will need to be aligned with all relevant legislative and regulatory instruments with a focus on making the data available for analysis in a lawful, controlled and meaningful way. Data also needs to be accurate, complete and timely if it is to be used to support complex analysis and decision making. For these reasons, management and governance focus needs to be on making data open and available across government via standardized APIs, formats and metadata. Improved quality of data will produce tangible benefits in terms of business intelligence, decision making, sustainable cost-savings and productivity improvements. The current trend towards open data and open government has seen a focus on making data sets available to the public, however these „open" initiatives need to also put focus on making data open, available and standardized within and between agencies in such a way that allows inter-governmental agency use and collaboration to the extent made possible by the privacy laws.

C.Technology and analytical systems:

The emergence of big data and the potential to undertake complex analysis of very large data sets is, essentially, a consequence of recent advances in the technology that allow this. If big data analytics is to be adopted by agencies, a large amount of stress may be placed upon current ICT systems and solutions which presently carry the burden of processing, analyzing and archiving data. Government agencies will need to manage these new requirements efficiently in order to deliver net benefits through the adoption of new technologies.

VII.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 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 Big data mining 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.] Bakshi, K.,(2012)," Considerations for big data: Architecture and approach"

[2.] Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W., (18-22 Dec.,2012) , "Shared disk big data analytics with Apache Hadoop"

[3.] Aditya B. Patel, Manashvi Birla, Ushma Nair ,(6-8 Dec. 2012),"Addressing Big Data Problem Using Hadoop and Map Reduce"

[4.] Wei Fan and Albert Bifet " Mining Big Data:Current Status and Forecast to the Future",Vol 14,Issue 2,2013

[5.] Algorithm and approaches to handle large Data-A Survey,IJCSN Vol 2,Issue 3,2013

[6.] 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

[7.] Xu Y etal, balancing reducer workload for skewed data using sampling based partioning 2013.

[8.] X. Niuniu and L. Yuxun, "Review of Decision Trees," IEEE, 2010 .

[9.] 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

ABOUT THE AUTHOR

G Bramhaiah Achary pursuing M.Tech (CSE) from Department of Computer Science & Engineering, Nagole Institute of Technology & Science, Hyderabad, India and received B.Tech Degree in Information Technology from Jawaharlal Nehru Technological University, Hyderabad, India.