BIG DATA ANALYSIS USING HACE THEOREM

Deepak S. Tamhane, Sultana N. Sayyad

***Abstract*- Big Data consists of huge modules, difficult, growing data sets with numerous and , independent sources. With the fast development of networking, storage of data, and the data gathering capacity, Big Data are now quickly increasing in all science and engineering domains, as well as animal, genetic and biomedical sciences. This paper elaborates a HACE theorem that states the characteristics of the Big Data revolution, and proposes a Big Data processing model from the data mining view. This data-oriented model contains demand-driven aggregation of data sources, mining and study, user knowledge modeling, and security and privacy issues. We examine the difficult issues in the data-oriented model and also in the Big Data revolution.**

***Index Terms* :- Big Data, Data mining, Hace theorem, 5V's,Privacy**

1. Introduction

Every day 3 billion kilobytes of data are produced and today 90 percent of the data in the web were created within the last two years. Our ability for data making has never been so dominant and massive since the creation of the information technology in the early 19th century. One example like Prime Minister Narendra Modi has discussed with the Pakistan's last Prime Minister Nawaz Sharif about two nation development and interrelated cooperation against terrorism such online debate offer a new resources to logic the public happiness and make feedback in real-time, and are mostly engaging compared to media, such as radio as well as TV broadcasting. In another instance, a public picture distribution site, flickr, which achieved 2.5 million photos per day.Each photo is assumed the size of 2 megabyte, this needs 5 terabytes storage every single day and as an old axiom elaborates that a single picture has value of lacks of words. The pictures on Flicker are a huge tank for us to search the human civilization, social proceedings, public relationships, disasters, and so on, only if we have the power to attach the massive amount of data.

These instances shows the rise of BIG DATA applications where data gathering has grown extremely and is beyond the capability of usually used software tools to catch, control, and make the procedure. The most essential challenge for BIG DATA applications is to discover the huge volume of data and mine useful information for future events. In many occurrences, the information mining process has to be very capable and close to real time because storing all practical data is nearly in flexible. For a instance, the square kilometer array (SKA) used in radio astronomy contains of 1,000 to 1,500 15-meter dishes in a central 5-km area. It offers 100 times more responsive image than any existing radio telescopes. However, with a 50 gigabytes (GB) second data volume, the data delivers from the SKA are specially large. Although scientist have confirmed that attractive patterns, such as temporary radio anomalies can be exposed from the SKA data, existing processes can boosts in an offline manner and are incapable of handling this Big Data scenario in real time. As a result, the unique data storages require an useful data study and calculation stage to get fast reply and real-time categorization for Big Data.

In this paper we suggest a HACE theorem to form Big Data features.

2. Big Data

Big Data is a comprehensive term for any collection of data sets so large and multifarious that it becomes difficult to process them using conventional data processing applications. The challenges include analysis, capture, search, sharing, storage, transfer, revelation, and privacy violations. The tendency to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, prevent diseases, combat crime and so on.

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 multifarious information, such as customer reviews from feasible websites, photos and other multimedia, and comments on social networking sites. These data can not be separated into categorized or analyzed numerically.

ISSN: 2278 – 1323

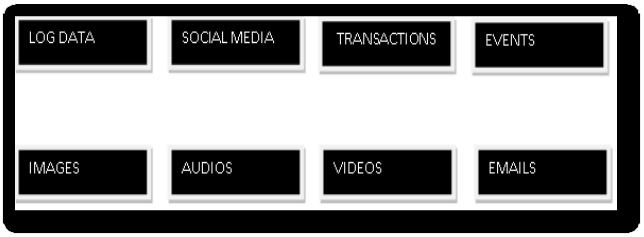

Figure 1. Sources of BIG DATA

3. Big Data Characteristic(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

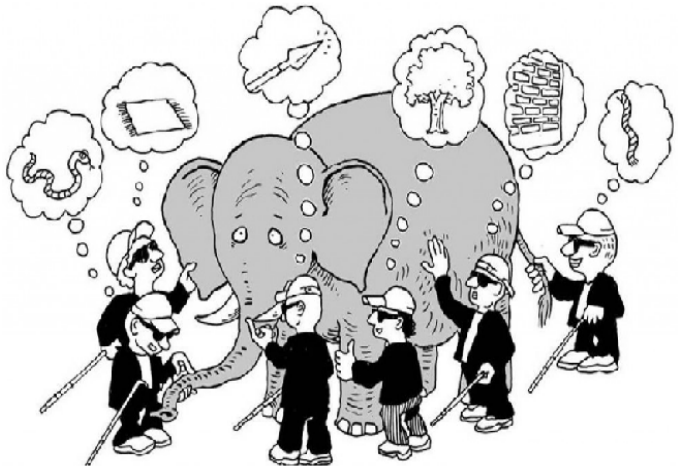

Figure 2. The blind men and the enormous elephant: the restricted view of each blind man leads to a biased conclusion.

These characteristics make it an intense challenge fordiscovering useful knowledge from the Big Data. In a native sense, we can imagine that a number of blind menare trying to size up a giant elephant (see Fig. 2), whichwill be the Big Data in this context. The goal of each blindman is to extract conclusion of the elephantaccording 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 elephant "feels" like a rope, a wall, a tree, a mat, or a snake depending on the part each of them is limited to. To make the problem even more complex, let us accept that 1) the elephant is increasing quickly and its posture varies continually, and 2) each blind man may have his own information sources that tell him about subjective knowledge about the elephant (e.g., one blind man may exchange his feeling about the elephant 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 elephant in a actual way. Certainly, this job is not as simple as enquiring each blind man to designate his spirits about the camel 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 term Big Data literally concerns about data volumes, HACE theorem suggests that the key characteristics of the Big Data are

A. **Huge with various and miscellaneous data sources**:-One of the fundamental characteristics of the Big Data is the huge volume of data represented by various and miscellaneous dimensionalities. This huge volume of data comes from various sites like Twitter, MySpace, Orkut and LinkedIn etc. This is because different information collectors prefer their own representation or procedure for data recording, and the nature of different applications also results in various data representations

B. **Autonomous Sources with circulated & disperse Control: -** Autonomous Sources with circulated & disperse Control are a main characteristic of Big Data applications. Being autonomous, each data source is able to produce and collect information without connecting 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 depending on other servers. On the other hand, the massive volumes of the data also make an application susceptible to attacks or failure, if the whole system has to depend on any centralized control unit. For example, Asian markets of Wal-Mart are inherently different from its North American markets in terms of seasonal promotions, top sell items, and customer behaviors. More specifically, the local government regulations also impact on the wholesale management process and result in

ISSN: 2278 – 1323

restructured data representations and data warehouses for local markets.

C. **Complex and Evolving associations:-** In an early stage of data centralized information systems, the focus is on finding best feature values to represent each observation. This type of sample feature representation inherently treats each individual as an independent entity without considering their social connections, which is one of the most important factors of the human society. The correlations between individuals inherently complicate the whole data representation and any reasoning process on the data. In a dynamic world, the features used to represent the individuals and the social ties used to represent our connections may also evolve with respect to temporal, spatial, and other factors. 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.

4. The 5 V's of Big Data

In the past, the term "Big Data" has served as a catch all phrase for the huge amounts of information available and collected in the digital world. Today, Big Data is being called the rising power of the 21st century and is helping as much more than a buzzword, acting as a huge d High Performance Inter-Thread Messaging Library in the technology channel. With a growth rate of 50% a year, harnessing all of the components of Big Data presents a real challenge. (See Figure 3) It shows 5 V's in Big Data.

In a 2001 metaGroup publication, Gartner analyst Doug Laney introduced the 3 V's of data management, defining the 3 main components of data as volume, velocity and variety.

Volume –Volume refers to the vast amounts of data that is generated every second. With 90 percent of the world's data created in the last 2 years , the volume of data that is being collected daily is what presents immediate challenges for businesses.

Velocity –Velocity refers to the speed at which new data is generated and the speed at which it moves around. For example, The New York Stock Exchange captures about 1 terabyte of trade information daily. Reacting fast enough and analyzing the streaming data is troubling to businesses, with speeds and peak periods often inconsistent.

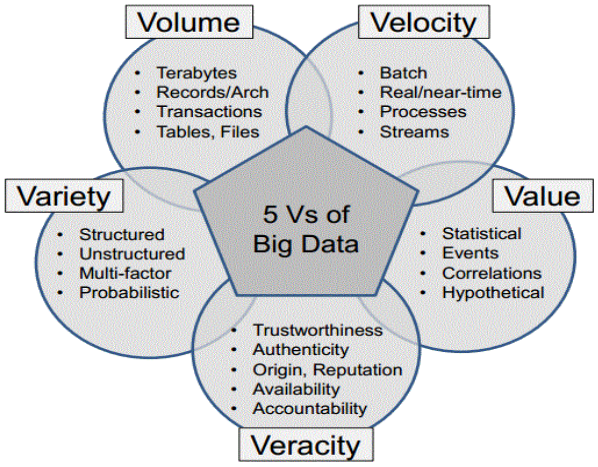

Figure 3. Five Vs of BIG DATA

Variety – Refers to the different forms of data that we collect and use. Data comes in different formats, such as structured and unstructured. To make matters even more challenging, because of the explosion of data generated by social media sources, 80 to 85 percent of the entire world's data is now unstructured (text, audio, video, click streams, log files and so on).

As the years have continued and the amount of data produced significantly increases, we now know much more about what defines Big Data, and IBM has introduced a fourth V, Veracity, as outlined in their infographic.

Veracity – The average billion dollar company is losing $130 million a year due to poor data management. Veracity refers to the uncertainty surrounding data, which is due to data inconsistency and incompleteness, which leads to another challenge, keeping Big Data organized.

The volume, velocity, variety and veracity of data that is being generated today goes beyond what traditional analytics systems can handle in a timely and efficient manner. This leads to the fifth V that organizations are struggling with, finding the Value within their data.

Value – Through effective data mining and analytics, the massive amount of data that we collect throughout the normal course of doing business can be put to good use and yield value and business opportunities. By applying data mining and analytics to expose valuable business information embedded in structured, unstructured, and streaming data and data

ISSN: 2278 – 1323

warehouses, this insight can be used to help revamp supply chains, improve program planning, track sales and marketing activities, measure performance across channels, and transform into an on-demand business. A Big Data strategy gives businesses the capability to better analyze this data with a goal of accelerating profitable growth.

5. Data Mining Challenges With Big Data

For an intelligent knowledge database system [13] to handle Big Data, the essential key is to scale up to the extremely large volume of data and provide actions for the characteristics featured by the HACE theorem. Figure. 4 shows a conceptual view of the Big Data processing framework, which includes three tiers from inside out with considerations on data accessing and computing (Tier I), data isolation and domain knowledge (Tier II), and Big Data mining algorithms (Tier III).

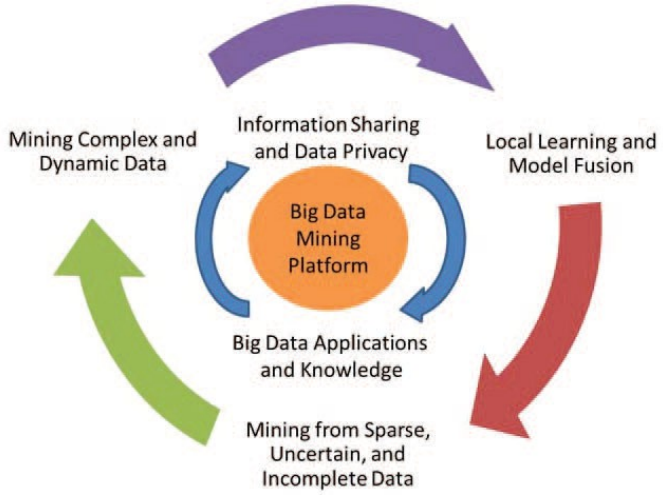

Figure 4. a conceptual view of the Big Data processing framework

5.1. Tier I: Big Data Mining Platform (Data Accessing & Computing):

In typical data mining systems, the mining procedures require computational thorough computing units for data analysis and comparisons. For Big Data mining, because amount of data is massive so that a single personal computer (PC) cannot handle, a typical Big Data processing framework will depend on cluster computers with a high-performance computing platform, with a data mining task being executed by running some parallel Computing tools, such as MapReduce or Enterprise Control Language (ECL), on a large number of clusters. The function of the software module is to make sure that a single data mining task, such as finding the best match of a query from a database with billions of records, is divided into many small tasks each of which is running on one or multiple cluster.

5.2 Tier II: data isolation and domain knowledge

In Big Data, Semantic & Application knowledge refer to several aspect related to the rules, policies, user information & application information. The most important aspect in this tier contain 1) Information sharing and its confidentiality; and 2) domain and application knowledge.

5.2.1 Information Sharing and its confidentiality

Information sharing is an crucial goal for all systems relating multiple parties [7]. While the Goal of sharing is clear, a real-world concern is that Big Data applications are related to sensitive information, such as banking transactions and medical records. Simple data interactions do not resolve privacy concerns [6], [8], [11]., but public revelation of an individual's personal locations/movements over time can have serious repercussion for privacy. To protect privacy, two common approaches are to
1) limit access to the data, such as adding certification or access control to the data entries, so sensitive information is accessible by a limited group of users only, and
2) Remove data fields such that sensitive information cannot be pinpointed to an individual record [5].

5.2.2 Domain and Application Knowledge

Domain and application knowledge [9] provides necessary information for designing Big Data mining algorithms and systems. In a simple case, Application knowledge can help to identify right features for modeling the essential data. The domain and application knowledge can also help design feasible business objectives by using Big Data analytical techniques.

5.3 Tier III: Big Data Mining Algorithms

5.3.1 Local knowledge and Model synthesis for Multiple Information Sources

As Big Data applications are featured with independent sources and decentralized controls, collecting all distributed data sources to a centralized site for mining is thoroughly excessive due to the

ISSN: 2278 – 1323

possible transmission cost and privacy concerns. More specifically, the global mining can be featured with a two-step process, at data, model, and at knowledge levels. At the data level, each local site can calculate the data statistics based on the local data sources and exchange the statistics between sites to achieve a global data distribution view. At the model or pattern level, each site can carry out local mining activities, with respect to the localized data, to discover local patterns. By exchanging patterns between multiple sources, new global patterns can be synthetized by aggregating patterns across all sites [2]. At the knowledge level, model correlation analysis finds out the importance between models generated from different data sources to determine how relevant the data sources are correlated with each other, and how to form accurate decisions based on models built from autonomous sources.

5.3.2 Mining from meager, tentative, and partial Data

Meager, tentative, and partial data are defining features for Big Data applications. Being meager, the number of data points is too few for deriving consistent conclusions. Tentative data are a special type of data reality where each data set is no longer deterministic but is subject to some casual/inaccurate distributions. The absent values can be caused by different realities, such as the failure of a sensor node, or some regular policies to intentionally skip some values. While most modern data mining algorithms have in-built solutions to handle absent values, data attribution is an established research field that seeks to attribute absent values to produce enhanced models

5.3.3 Mining Complex and Dynamic Data

The growth of Big Data is driven by the fast growing of complex data and their changes in volumes and in nature [6]. Documents posted on WWW servers, Internet backbones, social networks, communication networks, and transportation networks, and so on are all featured with complex data. While complex dependency structures based on the data elevate the difficulty for our knowledge systems, However, Big Data complexity is presented in many aspects, including complex diverse data types, complex essential semantic relations in data, and complex association networks among data. In Big Data, data types include structured data, unstructured data, and semistructured data, and so on. particularly, there are relational databases, text, hyper-text, image, audio and video data, and so on.

Complex intrinsic semantic associations in data. News on the web, comments on facebook, pictures on piccassa, and video clips on YouTube may discuss about an academic awardwinning event at the same time. There is no doubt that there are strong semantic associations in these data. Mining complex semantic associations from "text-image-video" data will significantly help improve application system performance such as search engines or recommendation systems.

Complex relationship networks in data. In the context of Big Data, there exist relationships between individuals. On the Internet, individuals are webpages and the pages linking to each other via hyperlinks form a complex network. There also exist social relationships between individuals forming complex social networks, such as big relationship data from Facebook, Twitter, LinkedIn, and other social media [3], [4], including call detail records (CDR), devices and sensors information [1], [12], GPS and geocoded map data, massive image files transferred by the Manage File Transfer protocol, web text and click-stream data [2], scientific information, e-mail [10], and so on. To deal with complex relationship networks, emerging research efforts have begun to address the issues of structure-and-evolution, crowds-and-interaction, and information-and-communication.

6 Conclusions

While the term Big Data exactly related to data volumes, our HACE theorem applies the key characteristics of the Big Data are 1) huge with various and diverse data sources, 2) independent with scattered and decentralized control, and 3) difficult and developing in data and knowledge associations. Such mutual characteristics propose that Big Data require a "big mind" to merge data for maximum values [9]. To discover Big Data, we have analyzed some challenges at the data, model, and system levels. To maintain Big Data mining, high-performance computing platforms are necessary, which enforce organized designs to set free the full power of the Big Data. At the data level, the independent information sources and the range of the data collection environments, often result in data with complex conditions, such as uncertain values. In other situations, isolation concerns, noise, and errors can be introduced into the data, to construct distorted data copies. Mounting a secure and sound information sharing procedure is a main challenge. At the model level, the key challenge is to produce global models by joining locally searched patterns to form a unifying view. At the system level, the necessary challenge is that a Big Data mining framework desires to think difficult interaction between samples, models, and data sources, along with their sprouting changes with time and other possible factors. A system requests to be carefully designed so

ISSN: 2278 – 1323

that formless data can be linked through their difficult relationships to make useful patterns, and the growth of data volumes and item relationships should help form legal patterns to guess the trend and future.

References

[1] R. Ahmed and G. Karypis, "Algorithms for Mining the Evolutionof Conserved Relational States in Dynamic Networks," Knowledgeand Information Systems, vol. 33, no. 3, pp. 603-630, Dec. 2012.
[2] M.H. Alam, J.W. Ha, and S.K. Lee, "Novel Approaches to Crawling Important Pages Early," Knowledge and Information Systems, vol. 33, no. 3, pp 707-734, Dec. 2012.
[3] Y.-C. Chen, W.-C. Peng, and S.-Y. Lee, "Efficient Algorithms forInfluence Maximization in Social Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 577-601, Dec. 2012.
[4] G. Cormode and D. Srivastava, "Anonymized Data: Generation,Models, Usage," Proc. ACM SIGMOD Int'l Conf. Management Data,pp. 1015-1018, 2009.
[5] G. Duncan, "Privacy by Design," Science, vol. 317, pp. 1178-1179, 2007.
[6] D. Howe et al., "Big Data: The Future of Biocuration," Nature, vol. 455, pp. 47-50, Sept. 2008.
[7] B. Huberman, "Sociology of Science: Big Data Deserve a Bigger Audience," Nature, vol. 482, p. 308, 2012.
[8] I. Kopanas, N. Avouris, and S. Daskalaki, "The Role of Domain Knowledge in a Large Scale Data Mining Project," Proc. Second Hellenic Conf. AI: Methods and Applications of Artificial Intelligence, I.P. Vlahavas, C.D. Spyropoulos, eds., pp. 288-299, 2002.
[9] W. Liu and T. Wang, "Online Active Multi-Field Learning for Efficient Email Spam Filtering," Knowledge and Information Systems, vol. 33, no. 1, pp. 117-136, Oct. 2012.
[10] E. Schadt, "The Changing Privacy Landscape in the Era of Big Data," Molecular Systems, vol. 8, article 612, 2012.
[11] A. da Silva, R. Chiky, and G. He´brail, "A Clustering Approach for Sampling Data Streams in Sensor Networks," Knowledge and Information Systems, vol. 32, no. 1, pp. 1-23, July 2012.
[12] X. Wu, "Building Intelligent Learning Database Systems," AI Magazine, vol. 21, no. 3, pp. 61-67, 2000

Author Profile

Deepak S. Tamhane, Received the Bachelor degree (B.E.) in Information Technology in 2009 from SVPM COE, Malegaon (Bk). He is now pursuing Master's degree in Computer Science Engineering at MLR, Institute of Technology, Hyderabad. He is currently working as a Lecturer in PGM College of Engineering, Wagholi, Pune
Email-Id: deepaktamhane18@gmail.com

Sultana N. Sayyad, Received the Bachelor degree (B.E.) in Information Technology in 2009 from SVPM COE, Malegaon (Bk). She is now pursuing Master's degree in Computer Science Engineering at MLR, Institute of Technology, Hyderabad. She is currently working as a Lecturer in Al-Ameen College of Engineering, Koregaon Bhima, Pune
Email-Id: sayyadsultana75@gmail.com

.

ISSN: 2278 – 1323