Data Mining with Big Data using Hace Theorm

U. Dinesh[1] Ms. Jayanthi[2]
[1]Research Scholar (M Phil) [2]Associate Professor
[1]Department of Computer Science & Engineering
[1,2]RVS College of Arts and Science, Sulur

Abstract— Big Data concerns with a large-volume, complex data, growing data sets with multiple, independent sources. With the fast enhancement of networking, data storage, and the data collection capacity, Big Data is now fast expanding in all science and engineering fields, as well as physical, biological and bio-medical sciences too. We move on with HACE theorem that characterizes and describe the features of the Big Data revolutions, and recommends a Big Data processing model, from the data mining perspective view. This data-driven model involves data-driven aggregation of information sources, mining and analysis, handler interest modeling, security and privacy thoughts. We inspect the challenging subjects in the data-driven model and in the Big Data revolution.

Key words: Big Data, Hace theorem, Data mining

I. INTRODUCTION

Every Day 3.5 billion kilobytes of data are produced and nowadays 90 % of the data in the universe produced with in the last two years our capability for data creation has never been so powerful and massive since the establishment of the information in the early 19th Century. Example like on April, 23 2015 Net Neutrality debates in India While online activists and even big Internet companies have come out to support Net Neutrality, the debate really not simple when it's in India and its triggered more than 3 million comments within a day in social Network. Such online arguments provide a new way to make feedback in real-time than compare with media like radio, Television broadcasting. Another example is the social networking giant twitter The Complete volume of data being stored today is exploding. In the year 200,800,000 petabytes (PB) of data were stored in the world. We assume this number to reach 40 zettabytes (ZB) by 2020.Twitter alone generates more than 7 terabytes (TB) of data every day, another great social Media Facebook 10 TB of data every day. The billions of text, image, videos, and sounds on twitter are a huge container for us to enhance human society social occasions, public activities, and disasters so on only if we have the power to prefix the large volume of the data.

The above instance illustrate the growth of BIG DATA applications where (data) information collection has developed tremendously and away from the commonly used software to capture, control and process with in a target time. The most important challenge for BIG DATA applications is to explore the huge volume of data and extract the useful hidden information for future process In many condition, the data mining process has to be very effective and close to real time because storing all the data is in flexible. For example, the Square Kilo meter Array (SKMA) in Radio astronomy on 2,000 to 2,500 20-Meter dishes in a central 10-Km zone. It give gives 100 times more clear vision than any pre-existing radio telescope However, with a 30 gigabytes(GB), the data produced from the SKA is very high. Although scientists have confirmed that stimulating patterns, such as temporary radio can be innovate from the SKA data available methods are not able to handle this BIG DATA As an output, the unparalleled information volumes want an effective data analysis. In this we recommend a HACE theorem to form Big Data features

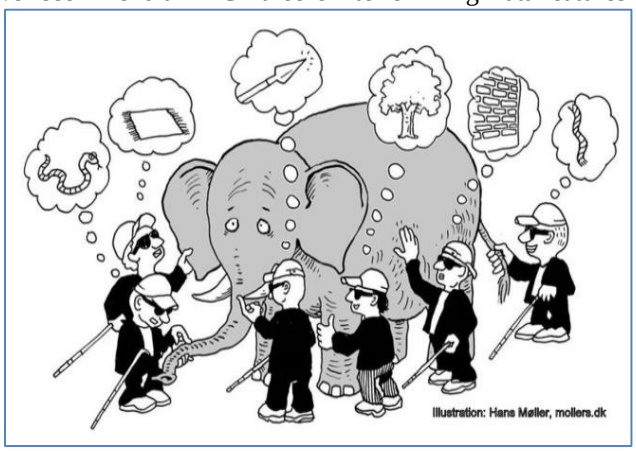

Fig. 1: Blind Man and Giant Elephant

II. BIG DATA CHARACTERISTIC'S

A. *HACE Theorem*

Big Data begin with huge-volume; Mixed, Independent sources with distributed and decentralized control, and seeks to explore Complex and Evolving relationships among data (information)These characteristics make it a great challenge for discovering useful information from the Big Data. In a raw sense, we can visualize that a number of blind men are trying to find the size of giant elephant the target of each blind man is to drawn conclusion of the elephant according to the part of information he collected during the time of the process. They each will conclude independent that the elephant "feels" like a wall, a rope, a mat, a tree, or a snake depending on the portion each of them is limited to. Make the problem even more complex let's consider that (a) the elephant is increasing rapidly and it's changing position rapidly and (b) each blind man may have his own data source that sounds about the subjective information about the elephant (e.g., one blind man may exchange his sense about the elephant with another one, where the exchanged information is basically subjected).

The term Big Data exactly concerns about information volumes, HACE theorem advises that the fundamental characteristics of the Big Data are

1) Huge Date with Different Data Sources

One of the fundamental characteristics of the Big Data is the huge volume of information represented by with several and Different perspective view. This large volume of data comes from various sites like Facebook, Twitter, MySpace, Orkut and LinkedIn, Printrest etc. This is because different date collectors prefer their own representation for data record, and each different application also produces output in the form of different data representation

2) Autonomous Sources with Circulated & Decentralized Control

Autonomous Sources with circulated & Decentralized Control are a main characteristic of Big Data applications. Being independent, each data source is able to create and collect data without involving any centralized control. This is same as like the World Wide Web (WWW) setting where each web server provides a certain amount of data and each server (webhost) is able to fully function without necessarily depending on other servers(webhost). On the other hand, the large volumes of the data also make an application vulnerable to hack or failure, if the whole system has to depend on any centralized control unit. Major Big Data associated Application such as Google, Yahoo, Facebook, and Wal-Mart they are servers are placed all over the world For example, American markets of Amazon are inherently different from its Indian markets in terms of seasonal promotions, top sell items, and customer activity sites .

3) Difficult and Evolving Relationship

While the size of the Big Data rises, so do the complexity and the relations underneath the data. At initial stage of data centralized information systems, the target is discover the best feature values to represent each reflexion. This is like using a number of data entity's, such as age, gender, education, country etc., to characterize each as separate This type of sample –feature represent inherently treats an each individual as a separate entity without think about their social linking which is one of the major elements of the human society. People from the friend circle connected based on common things like hobbies. Such social interaction are exist not only in day today life also its play major role in this world. For scenario so major social networking sites such as Facebook, Twitter or Instagram are mainly considered by social activities such as friends-suggestion and followers (Twitter and Instagram) the connection between each and individuals inherently complicate the whole data visualization and any intellectual process. In the model –feature representation, individuals are regarded similar if they share same feature values, where as in the model-feature relation representation, two independent can be connected together even though they may not share anything in common feature area at all. In a world which is keep on moving, the feature used to visualize the individual. Such a complex thing and it's becoming a part of the reality.

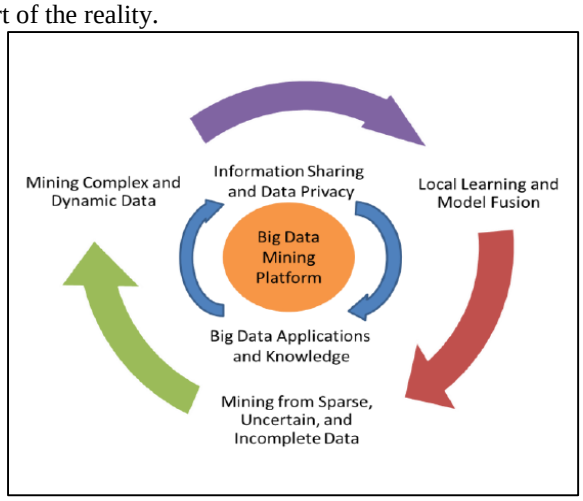

Fig. 2: A Big Data Processing Structure

Data application is hints to take complex (many-to-many, non-Linear) data relation, along with the developing change based on constraints to discover useful information from the Big Data repository.

III. Data Mining Challenges with Big Data

For an intellectual learning database system to handle Big Data .the needed hint is to scale up to the very large volume of data and provide solution for the characteristics featured by the above-mentioned HACE theorem. Figure 2 shows a theoretical vision of the Big Data processing structure which includes three tiers from inside out with respects on data retrieving and executing (Tier I), data privacy and domain information (Tier II), and Big Data mining procedures (Tier III). The Challenges at Tier 1(Process 1) focus on data receiving and actual execution procedures. Because Big Data are frequently stored at different area and data size and may continuously develop and effective computing will have to take scattered large-scale data storage into attention for computing For Scenario, While typical data mining procedures require all the information to be load in to the main memory, this is become a clear technical block for Big Data because relocating data across different location is huge expensive(e.g., Network Communication and its peripheral cost even if have large amount of main memory to store all the data for processing. The challenge at II Tier centre on semantic and domain data of various Big Data applications, such data can give additional credits to the mining process as well as add technical region to the big data access (Tier I) and mining procedures (Tier III). For scenario, Based on different domain application, the data terms and information sharing procedure between data producers and data consumer can be expressively different Sharing sensor network information for application .Sharing mobile user location. In additional above privacy issues, the application domain also can provide extra information to guide Big Data mining procedure design.For Scenario, in market basket operation data, each operation is consider individual and discover the information is typically represent the correlated items, possibly with respect to different temporal. In a social Networking, on the other hand , User are connected and share structure .the information is then represent by user groups, leader in each communities , and social effects etc., therefore understanding structure and application knowledge is important for high and low level mining algorithm design The circle at Tier III contains three phases. First, spare, mixed, uncertain, unfinished, and multi-source data are pre-processed by data fusion techniques. Second, Complex and dynamic information are mined after the pre-processing .third, the global information that is obtained by local learning and typical mixture is tested and relevant data is feedback to the pre-processing stage. Then the model and parameter are change according to the commands (feedbacks).In the whole development state, information sharing is not only for smooth development at each stage also for big data function.

A. Tier I: Big Data Mining Platform

In traditional data mining systems, the mining procedures require computational serious computing units for data analyse and comparing. A computing platform is there need

to have easy access to, at least, two types of things: knowledge and another one is act. For small Gage data mining tasks, a single PC (personal computer) which contain secondary memory disk and Processors, CPU is enough to full fill the data mining target. Many data mining procedures are designed to handle this type of issues. For medium Gage data mining process, data are typically huge and it can't able dump in main memory. General Solution are parallel computing or collective mining , aggregate from various data sources and use parallel computing programming to carry out the mining procedure.

For Big Data mining, because data Gauge is far outside the volume that is single PC (personal computer) can process, a typical Big Data proceeding framework will rely on cluster machines with huge performance computing zone, where a data mining job is deployed by executing some parallel program tools, Such as ECL (Enterprise Control Language), on large number of computing cluster The play of the software tools is to make sure that individual data mining job, such as finding the better match of a query form the storage(database) with millions of sample, is divided in to number of small units each of which is executing on one or more number if computing clusters(node) Such a Big Data System, Which combined both hardware and software components, is hardly alive without any key industrial stock holders idea.

Big Data mining offer opportunity to go above their relational database to rely on less structured information like email, sensor, social media and images that can be mined for useful data. Most business intelligence companies, such as Oracle, IBM etc., have all the featured they are own to support customer. These diverse data source and coordinate with vendors existing data to find new insight and capitalized on hidden associations

B. Tier II: Big Data Semantics and Application Information

Semantic and application information in Big Data refer to number of aspects related to the rule and policies, user information and domain knowledge. The two major issues at tier include (a) data sharing and privacy and (b) domain and application information .The prior provides answer to solution on how data are maintained, accessed and shared.

1) Knowledge Sharing and Data Privacy

Knowledge sharing is the ultimate target for all the system involving multiple groups. While the inspiration for share clear, a real-world fear is that Big Data application are related to sensitive knowledge. Such as hospital records, banking transactions and so easy data exchange or transfer do not over rule the privacy concerns

For example, People's location and their preference and their choice, one can enable a verity of useful location based on services, but public revelation of an individual's activities over time can have serious significance for privacy. To secure privacy, two common ways are to (a) secure access to the data, such an adding certification for privacy control to the data entry, so sensitive data is accessible by a specific group of users only and (b) anonymous data fields such as sensitive information cannot be pointed to an each record. For The first approach common challenge are to pattern privacy certification or access limit control mechanism, such that no sensitive data can be misconducted by unauthorized .the main aim is to inject randomness in the data to ensure number of privacy goals .For scenario the most common k-nearest measure is to ensure that each and every database must be in distinct from k-1 others.

One of the major profits of the data ammonization based data sharing approach that once anonym zed, data can be easily shared across different groups without involving any privacy access controls. This leads to another bunch of research area namely privacy data mining, where multiple party each hold some important data are try to achieve data mining target. Without any important data sharing this privacy protective mining goal two type of approach 1.communication protocol 2.some special data mining method for achieve knowledge

2) Domain and Application Information

Domain and application information gives needed information for design Big Data mining procedures and system. In a simple case, domain information can help analyses the right enhancement for model that underlying data. The domain and application information can also help pattern achievable business target by using big data analytic techniques For scenario, share market data are a typical area which is constantly generate huge amount of information, such as buys, bids, open market, close market in every single second. The market continuous evolves in different factor. An appealing big data mining task is to design a system of Big Data to analyses the movement of the stock market in next minutes its display and significant business value to the developer without right domain knowledge cleared metric and measurements to analyses the market movement

C. Tier III: Big Data Mining Procedures

1) Local Learning and Model Synthesis for Multiple data Sources

Big Data application are featured with individual source and distributed controls, grouping decentralized data source to be centralized for mining is methodically prohibitive due to the potential transformation cost and security concerns .oh the another view, although we can always carry out mining process at each decentralized sites the based view of the data collected at each different sites based decisions or model, like elephant and blind man conceptual. Big Data mining system has make data changes and synthesis mechanism to ensure that all circulated sites (or data sources) can work together to achieve a global optimization target. Model mining and correlation are the basic steps to certify that patterns produced from multiple data source can be aggregate to meet the global mining objective. More exactly can be featured with a two-step for process, at data, model and at information levels At the information level, each site can calculate the information statics based on the information sources and change the statics between local sites to achieve the target data distribution perspective at the model or pattern level, each site can out site mining content activities, with respective to the localized data.By exchanging the design between multiple sources and its aggregated by pattern across all the sites. Data source to determine how data source is relevant and correlated to each other, and how to form correct decision based on the pattern make form the various individual sources

2) Mining from Sparse, Indeterminate, and Imperfect Data

Sparse, Indeterminate, and Imperfect Data are defining feature of Big Data applications. Being sparse the number of data location is too few drawing conclusion. This is normal data dimensional issue, where data in a high dimensional space. Common approaches are to dimensional reductions to reduce the data view.

Indeterminate data are a special type of the reality with data field is not much longer deterministic but its subject to some other error distribute. This mainly linked to main domain specified applications with incorrect data fetching .For data security related application, user may inject errors in to information in order to remain autonomous. For uncertain data, the main challenges are that every data item is visualized as some sample distribute. Common solutions are data circulations in to consideration for model parameter similar method have also been applied for decision tree or queries such as general unsupervised learning approaches in data mining Undefined data are a special type of data reality where each data entity is no longer deterministic but is subject to some error distributions. This is mainly linked to area specific applications with incorrect data analyses and collections. For example, data formed from GPS equipment is inherently undefined, mainly because the technology block of the device bounds the precision of the data to certain levels (such as 1 meter). As an outcome, each recording location is represented by a mean value plus an alteration to indicate expected errors. For data confidentiality related applications, users may intentionally inject randomness into the data in order to remain nameless.

For undefined data, the main challenge is that each data item is represented as some sample deliveries but not as a single value, so most prevailing data mining procedures cannot be directly applied. Common solutions are to take the data circulations into thought to estimate model limits.

3) Mining Dynamic and Complex Data

The growth of Big Data is driven by the fast increasing of complex data and their changes in sizes and in nature. Documents posted on WWW web servers, Internet, social site, transportation networks and communication networks etc. are all contained with complex data. While complex dependence structures underneath the data raise the trouble for our learning systems, they also offer exciting opportunities that simple data demonstrations are incapable of accomplishing. For scenario, researchers have successfully used face book, a well-known social networking facility, to detect events such as earthquakes and major social activities, with nearly online speed and very high accurateness Making use of composite data is a major task for Big Data applications, because any two parties in a difficult network are possibly involved to each other with a social linking. Such a linking is quadratic with respect to the number of clusters in the network, so a million clusters network may be subject to trillion connections. For a large social network site, like Facebook, the number of users has already reached 10 billion, and analyzing such a huge network is a big challenge for Big Data mining. If we take daily user actions into consideration, the scale of difficulty will be even more amazing. Stimulated by the above challenges, many data mining methods have been developed to find interesting information from Big Data with complex relations and dynamically changing sizes.

D. Complex Heterogeneous Data Types

In Big Data, data types include unstructured data, semi-structured data and structured data, etc. Specifically, there are relational databases, text, hyper-text, audio, image and video etc. The existing data models include key-values stores, big table data clones, graph database and document databases, which are listed in an upward of the difficulty of these data models. Traditional data models are incapable of handling difficult data in the context of Big Data. Currently, there is no recognized effective and well-organized data model to handle Big Data real-time handling for complex data is a very challenging task.

IV. BIG DATA ASSOCIATED WORK

A. Data Mining Platforms

Due to the multi-system, huge, mixed and dynamic characteristics of application data involved in a distributed environment, Big Data is computing tasks on the Petabytes (PB), Exabyte (EB)-level data with a complex computing process. Therefore, utilizing a parallel computer structure, its resultant programming language support, and software models to efficiently analyses and data mine the distributed PB, EB-level data are the critical aim for Big Data handling to change from "quantity" to "quality". Currently, Big Data treating mainly based on parallel programming models like MapReduce, and cloud computing platform of Big Data services for the community. The MapReduce parallel programming model has been applied in machine learning and data mining procedures. We argue that the computational processes in the procedure learning process could be transformed into an abstract operation on a number of training data sets.

1) Big Data Semantics and Application Information

In secure protection of huge data, proposed a multi-layer irregular set model, which can accurately describe the granularity change produced by different levels of simplification and provide a theoretic foundation for measuring the data efficiency criteria in the anonymization process, and planned a dynamic mechanism for balancing confidentiality and data utility, to solve the optimal refinement order for classification.

For applications involving Big Data and incredible data sizes, it is often the case that data are physically spread at different locations, which means that consumers no longer physically possess they are storage data.

2) Big Data Mining Procedures / Big Data Mining Algorithms

The main aim for discovering information from huge data is improving the efficiency of single-source mining methods. On the basis of gradual improvement of computer peripheral functions, researcher continues to explore ways to improve the effectiveness of knowledge discovery algorithms to make them better for huge data. Due to huge data typically coming from different data sources, the knowledge discovery of the huge data must be accomplished using a multi-source mining mechanism. Information evolution is a common phenomenon in real-world.

V. CONCLUSIONS

Driven by practical applications and important industrial stakeholders and mining Big Data have shown to be a challenging yet very exciting task. While the term Big Data accurately concerns about data sizes, our HACE theorem suggests that the key features of the Big Data are:

1) Huge with heterogeneous and dissimilar data sources
2) Autonomous with decentralized control
3) Complex and evolving in data and information associations.

We honour Big Data as an emerging trend and the need for Big Data mining is rising in all science and manufacturing domains. We can further motivate the participation of the public audiences in the data creation circle for economic events. The time period of Big Data has arrived.

REFERENCES

[1] Ahmed and Karypis 2012, Rezwan Ahmed, George Karypis, Algorithms for mining the evolution of conserved relational states in dynamic networks, Knowledge and Information Systems, December 2012, Volume 33, Issue 3, pp 603-630
[2] Alam et al. 2012, Md. Hijbul Alam, JongWoo Ha, SangKeun Lee, Novel approaches to crawling important pages early, Knowledge and Information Systems, December 2012, Volume 33, Issue 3, pp 707-734
[3] Aral S. and Walker D. 2012, Identifying influential and susceptible members of social networks, Science, vol.337, pp.337-341.
[4] Liu and Wang 2012, Wuying Liu, Ting Wang, Online active multi-field learning for efficient email spam filtering, Knowledge and Information Systems, October 2012, Volume 33, Issue 1, pp 117-136
[5] Dr. Gary Parker, vol 7, 2004, Data Mining: Modules in emerging fields, CD-ROM.
[6] 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.
[7] 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.
[8] E. Schadt, "The Changing Privacy Landscape in the Era of Big Data," Molecular Systems, vol. 8, article 612, 2012.
[9] 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.
[10] X. Wu, "Building Intelligent Learning Database Systems," AI Magazine, vol. 21, no. 3, pp. 61-67, 2000.

