ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
*A Peer Reviewed Open Access International Journal*

# Mining and Detection HACE Theorem That Characterizes the Features of the Big Data

**Hima Sampathi Rao**
M.Tech,
Department of CSE,
AVNIET, JNTUH,
Hyderabad.

**SK Abdul Nabi**
Professor and HOD,
Department of CSE,
AVNIET, JNTUH,
Hyderabad.

## Abstract:

Big Data is a new term used to identify the datasets that due to their large size and complexity, we can not manage them with our current methodologies or data mining software tools. Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it.

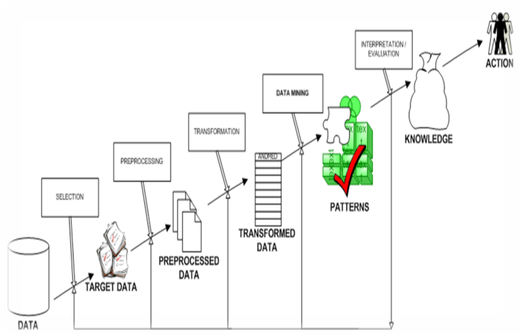
This paper presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data revolution.

## Index terms:

Big Data, autonomous sources, complex and evolving associations, data mining, heterogeneity.

## INTRODUCTION:

What is Data Mining?




## Structure of Data Mining:

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. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

## BIG DATA:

Streaming data analysis in real time is becoming the fastest and most efficient way to obtain useful knowledge from what is happening now, allowing organizations to react quickly when problems appear or to detect new trends helping to improve their performance. Evolving data streams are contributing to the growth of data created over the last few years. We are creating the same quantity of data every two days, as we created from the dawn of time up until 2003. Evolving data streams methods are becoming a low-cost, green methodology for real time online prediction and analysis. We discuss the current and future trends of mining evolving data streams, and the challenges that the field will have to overcome during the next years. Nowadays, the quantity of data that is created every two days is estimated to be 5 exabytes. This amount of data is similar to the amount of data created from the dawn of time up until 2003. Moreover, it was estimated that 2007 was the first year in which it was not possible to store all the data that we are producing. This massive amount of data opens new challenging discovery tasks.

Data stream real time analytics are needed to manage the data currently generated, at an ever increasing rate, from such applications as: sensor networks, measurements in network monitoring and traffic management, log records or click-streams in web exploring, manufacturing processes, call detail records, email, blogging, twitter posts and others. In fact, all data generated can be considered as streaming data or as a snapshot of streaming data, since it is obtained from an interval of time.In the data stream model, data arrive at high speed, and algorithms that process them must do so under very strict constraints of space and time. Consequently, data streams pose several challenges for data mining algorithm design. First, algorithms must make use of limited resources (time and memory). Second, they must deal with data whose nature or distribution changes over time.

## 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 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 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.2 Government agencies realise 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, utilised 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 standardised 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 standardised 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, analysing 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.

## LITERATURE SURVEY:

Dynamic networks have recently being recognized as a powerful abstraction to model and represent the temporal changes and dynamic aspects of the data underlying many complex systems. Significant insights regarding the stable relational patterns among the entities can be gained by analyzing temporal evolution of the complex entity relations. This can help identify the transitions from one conserved state to the next and may provide evidence to the existence of external factors that are responsible for changing the stable relational patterns in these networks.

This paper presents a new data mining method that analyzes the time-persistent relations or states between the entities of the dynamic networks and captures all maximal non-redundant evolution paths of the stable relational states. Experimental results based on multiple datasets from real-world applications show that the method is efficient and scalable.

Web crawlers are essential to many Web applications, such as Web search engines, Web archives, and Web directories, which maintain Web pages in their local repositories. In this paper, we study the problem of crawl scheduling that biases crawl ordering toward important pages. We propose a set of crawling algorithms for effective and efficient crawl ordering by prioritizing important pages with the well-known PageRank as the importance metric.

In order to score URLs, the proposed algorithms utilize various features, including partial link structure, inter-host links, page titles, and topic relevance. We conduct a large-scale experiment using publicly available data sets to examine the effect of each feature on crawl ordering and evaluate the performance of many algorithms. The experimental results verify the efficacy of our schemes. In particular, compared with the representative Rank Mass crawler, the FPR-title-host algorithm reduces computational overhead by a factor as great as three in running time while improving effectiveness by 5 % in cumulative PageRank.

## PROBLEMS IN CURRENT SYSTEM:

The rise of Big Data applications where data collection has grown tremen dously and is beyond the ability of commonly used software tools to capture, manage, and process within a “tolerable elapsed time.” The most fundamental challenge for Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions. In many situations, the knowledge extraction process has to be very efficient and close to real time because storing all observed data is nearly infeasible. The unprecedented data volumes require an effective data analysis and prediction platform to achieve fast response and real-time classification for such Big Data.

## Drawbacks:

The challenges at Tier I focus on data accessing and arithmetic computing procedures. Because Big Data are often stored at different locations and data volumes may continuously grow, an effective computing platform will have to take distributed large-scale data storage into consideration for computing.The challenges at Tier II center around semantics and domain knowledge for different Big Data applications. Such information can provide additional benefits to the mining process, as well as add technical barriers to the Big Data access (Tier I) and mining algorithms (Tier III). At Tier III, the data mining challenges concentrate on algorithm designs in tackling the difficulties raised by the Big Data volumes, distributed data distributions, and by complex and dynamic data characteristics.We propose a HACE theorem to model Big Data characteristics. The characteristics of HACH make it an extreme challenge for discovering useful knowledge from the Big Data. The HACE theorem suggests that the key characteristics of the Big Data are 1) huge with heterogeneous and diverse data sources, 2) autonomous with distributed and decentralized control, and 3) complex and evolving in data and knowledge associations. To support Big Data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data.

Provide most relevant and most accurate social sensing feedback to better understand our society at real time.

## HACE:

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 persons 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).

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:

Multistructure, 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.

## IMPLEMENTATION:

### Integrating and mining biodata:

We have integrated and mined biodata from multiple sources to decipher and utilize the structure of biological networks to shed new insights on the functions of biological systems. We address the theoretical underpinnings and current and future enabling technologies for integrating and mining biological networks. We have expanded and integrated the techniques and methods in information acquisition, transmission, and processing for information networks. We have developed methods for semantic-based data integra-tion, automated hypothesis generation from mined data, and automated scalable analytical tools to evaluate simulation results and refine models.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

## Big Data Fast Response:

We propose to build a stream-based Big Data analytic framework for fast response and real-time decision making.

•Designing Big Data sampling mechanisms to reduce Big Data volumes to a manageable size for processing

•Building prediction models from Big Data streams. Such models can adaptively adjust to the dynamic changing of the data, as well as accurately predict the trend of the data in the future; and

• A knowledge indexing framework to ensure real-time data monitoring and classification for Big Data applications.

## Pattern matching and mining:

We perform a systematic investigation on pattern matching, pattern mining with wildcards, and application problems as follows:

* Exploration of the NP-hard complexity of the matching and mining problems,

• Multiple patterns matching with wildcards,

• Approximate pattern matching and mining, and

• Application of our research onto ubiquitous personalized information processing and bioinformatics.

## Key technologies for integration and mining:

We have performed an investigation on the availability and statistical regularities of multisource, massive and dynamic information, including cross-media search based on information extraction, sampling, uncertain informa-tion querying, and cross-domain and cross-platform information polymerization. To break through the limitations of traditional data mining methods, we have studied heterogeneous information discovery and mining in complex inline data, mining in data streams, multigranularity knowledge discovery from massive multisource data, distribution regula-rities of massive knowledge, quality fusion of massive knowledge.

## Group influence and interactions:

» Employing group influence and information diffusion models, and deliberating group interaction rules in social networks using dynamic game theory.

» Studying interactive individual selection and effect evaluations under social networks affected by group emotion, and analyzing emotional interactions and influence among individuals and groups, and

» Establishing an interactive influence model and its computing methods for social network groups, to reveal the interactive influence effects and evolution of social networks.

## CONCLUSION:

Driven by real-world applications and key industrial stakeholders and initialized by national funding agencies, managing and mining Big Data have shown to be a challenging yet very compelling task. While the term Big Data literally concerns about data volumes, our HACE theorem suggests that the key characteristics of the Big Data are 1) huge with heterogeneous and diverse data sources, 2) autonomous with distributed and decentralized control, and 3) complex and evolving in data and knowledge associations. Such combined characteristics suggest that Big Data require a "big mind" to consolidate data for maximum values .To explore Big Data, we have analyzed several challenges at the data, model, and system levels.

To support Big Data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data. At the data level, the autonomous information sources and the variety of the data collection environments, often result in data with complicated conditions, such as missing/uncertain values. In other situations, privacy concerns, noise, and errors can be introduced into the data, to produce altered data copies. Developing a safe and sound information sharing protocol is a major challenge. At the model level, the key challenge is to generate global models by combining locally discovered patterns to form a unifying view. This requires carefully designed algorithms to analyze model correlations between distributed sites, and fuse decisions from multiple sources to gain a best model out of the Big Data.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

At the system level, the essential challenge is that a Big Data mining framework needs to consider complex relationships between samples, models, and data sources, along with their evolving changes with time and other possible factors. A system needs to be carefully designed so that unstructured data can be linked through their complex relationships to form useful patterns, and the growth of data volumes and item relationships should help form legitimate patterns to predict the trend and future.We regard Big Data as an emerging trend and the need for Big Data mining is arising 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. We can further stimulate the participation of the public audiences in the data production circle for societal and economical events. The era of Big Data has arrived.

## FUTURE ENHANCEMENT:

Open nature of Big Data systems exposes them to malicious activity. Building trust relationships among peers can mitigate attacks of malicious peers. In future we can enhance the system with distributed algorithms that enable a peer to reason about trustworthiness of other peers based on past interactions and recommendations. Peers create their own trust network in their proximity by using local information available and do not try to learn global trust information. Two contexts of trust, service, and recommendation contexts are defined to measure trustworthiness in providing services and giving recommendations. Interactions and recommendations are evaluated based on importance, recentness, and peer satisfaction parameters. Additionally, recommender's trustworthiness and confidence about a recommendation are considered while evaluating recommendations. Simulation experiments on a file sharing application show that the proposed model can mitigate attacks on 16 different malicious behavior models. In the experiments, good peers were able to form trust relationships in their proximity and isolate malicious peers.

## REFERENCES:

[1] R. Ahmed and G. Karypis, "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks," Knowledge and 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] S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," Science, vol. 337, pp. 337-341, 2012.

[4] A. Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data," ACM Crossroads, vol. 19, no. 1,
pp. 20-23, 2012.

[5] S. Banerjee and N. Agarwal, "Analyzing Collective Behavior from Blogs Using Swarm Intelligence," Knowledge and Information Systems, vol. 33, no. 3, pp. 523-547, Dec. 2012.

[6] E. Birney, "The Making of ENCODE: Lessons for Big-Data Projects," Nature, vol. 489, pp. 49-51, 2012.

[7] J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," J. Computational Science, vol. 2, no. 1, pp. 1-8, 2011.

[8] S. Borgatti, A. Mehra, D. Brass, and G. Labianca, "Network
Analysis in the Social Sciences," Science, vol. 323, pp. 892-895, 2009.

[9] J. Bughin, M. Chui, and J. Manyika, Clouds, Big Data, and Smart Assets: Ten Tech-Enabled Business Trends to Watch. McKinSey Quarterly, 2010.

[10] D. Centola, "The Spread of Behavior in an Online Social Network Experiment," Science, vol. 329, pp. 1194-1197, 2010.

[11] E.Y. Chang, H. Bai, and K. Zhu, "Parallel Algorithms for Mining Large-Scale Rich-Media Data," Proc. 17th ACM Int'l Conf. Multimedia, (MM '09,) pp. 917-918, 2009.

[12] R. Chen, K. Sivakumar, and H. Kargupta, "Collective Mining of Bayesian Networks from Distributed Heterogeneous Data," Knowledge and Information Systems, vol. 6, no. 2, pp. 164-187, 2004.

[13] Y.-C. Chen, W.-C. Peng, and S.-Y. Lee, "Efficient Algorithms for Influence Maximization in Social Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 577-601, Dec. 2012.

[14] C.T. Chu, S.K. Kim, Y.A. Lin, Y. Yu, G.R. Bradski, A.Y. Ng, and K. Olukotun, "Map-Reduce for Machine Learning on Multicore," Proc. 20th Ann. Conf. Neural Information Processing Systems (NIPS '06), pp. 281-288, 2006.

[15] G. Cormode and D. Srivastava, "Anonymized Data: Generation, Models, Usage," Proc. ACM SIGMOD Int'l Conf. Management Data, pp. 1015-1018, 2009.

[16] S. Das, Y. Sismanis, K.S. Beyer, R. Gemulla, P.J. Haas, and J. McPherson, "Ricardo: Integrating R and Hadoop," Proc. ACM SIGMOD Int'l Conf. Management Data (SIGMOD '10), pp. 987-998. 2010.

[17] P. Dewdney, P. Hall, R. Schilizzi, and J. Lazio, "The Square Kilometre Array," Proc. IEEE, vol. 97, no. 8, pp. 1482-1496, Aug. 2009.

[18] P. Domingos and G. Hulten, "Mining High-Speed Data Streams," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '00), pp. 71-80, 2000.

[19] G. Duncan, "Privacy by Design," Science, vol. 317, pp. 1178-1179, 2007.

[20] B. Efron, "Missing Data, Imputation, and the Bootstrap," J. Am. Statistical Assoc., vol. 89, no. 426, pp. 463-475, 1994.

[21] A. Ghoting and E. Pednault, "Hadoop-ML: An Infrastructure for the Rapid Implementation of Parallel Reusable Analytics," Proc. Large-Scale Machine Learning: Parallelism and Massive Data Sets Workshop (NIPS '09), 2009.

[22] D. Gillick, A. Faria, and J. DeNero, MapReduce: Distributed Computing for Machine Learning, Berkley, Dec. 2006.

[23] M. Helft, "Google Uses Searches to Track Flu's Spread," the new York Times, http://www.nytimes.com/2008/11/12/technology/internet/12flu.html 2008.