



# **Data Flow of Motivated Data Using Heterogeneous Method for Complexity Reduction**

S.Banupriya <sup>1</sup>, V.Vijayadeepa <sup>2</sup>

Research Scholar, Dept. of CS, Muthayammal College of Arts & Science, Rasipuram, Namakkal, India<sup>1</sup>

HOD, Dept. of CS, Muthayammal College of Arts & Science, Rasipuram, Namakkal, India<sup>2</sup>

**ABSTRACT:** Big Data is huge data sets with sizes above the ability of commonly used software tools to manage, and process the data within a tolerable minimum time. The era of Big Data has arrived astonishingly in the past few years. Numerous data are produced in the form of documents, chatting messages, audio, video, and applications and they are spread in the web. 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.

**KEYWORDS:** Big data, HACE theorem, data revolution, data driven model.

## **I. INTRODUCTION**

The era of Big Data has arrived astonishingly in the past few years. Numerous data are produced in the form of documents, chatting messages, audio, video, and applications and they are spread in the web. It will be harder to analyze these enormous data and we need more intricate algorithms and applications for mining these heterogeneous data. Also Big data has property of autonomous sources with complex and evolving relationships. In the internet every day quintillion bytes of data are created and our capability for data generation has never been so powerful and enormous ever since last few centuries. Our capability for data generation has never been so powerful and enormous ever since the invention of the information technology in the early 19th century. As another example, on 4 October 2012, the first presidential debate between President Barack Obama and Governor Mitt Romney triggered more than 10 million tweets within 2 hours. Among all these tweets, the specific moments that generated the most discussions actually revealed the public interests, such as the discussions about medicare and vouchers. Such online discussions provide a new means to sense the public interests and generate feedback in realtime, and are mostly appealing compared to generic media, such as radio or TV broadcasting. Another example is Flickr, a public picture sharing site, which received 1.8 million photos per day, on average, from February to March 2012. Assuming the size of each photo is 2 megabytes (MB), this requires 3.6 terabytes (TB) storage every single day. Indeed, as an old saying states: "a picture is worth a thousand words," the billions of pictures on Flickr are a treasure tank for us to explore the human society, social events, public affairs, disasters, and so on, only if we have the power to harness the enormous amount of data. The above examples demonstrate the rise of Big Data applications where data collection has grown tremendously 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. For example, the square kilometer array (SKA) in radio astronomy consists of 1,000 to 1,500 15-meter dishes in a central 5-km area. It provides 100 times more sensitive vision than any existing radio telescopes, answering fundamental questions about the Universe. However, with a 40 gigabytes (GB)/second data volume, the data generated from the SKA are exceptionally large. Although researchers have confirmed that interesting patterns, such as transient radio anomalies can be discovered from



# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

the SKA data, existing methods can only work in an offline fashion and are incapable of handling this Big Data scenario in real time. As a result, the unprecedented data volumes require an effective data analysis and prediction platform to achieve fast response and real-time classification for such Big Data.

## II. RELATED WORK

### Big Data Mining Platforms (Tier I)

Due to the multisource, massive, heterogeneous, and dynamic characteristics of application data involved in a distributed environment, one of the most important characteristics of Big Data is to carry out computing on the petabyte (PB), even the exabyte (EB)-level data with a complex computing process. Therefore, utilizing a parallel computing infrastructure, its corresponding programming language support, and software models to efficiently analyze and mine the distributed data are the critical goals for Big Data processing to change from “quantity” to “quality.” Currently, Big Data processing mainly depends on parallel programming models like Map Reduce, as well as providing a cloud computing platform of Big Data services for the public. MapReduce is a batch-oriented parallel computing model. There is still a certain gap in performance with relational databases.

Improving the performance of Map Reduce and enhancing the real-time nature of large-scale data processing have received a significant amount of attention, with Map Reduce parallel programming being applied to many machine learning and data mining algorithms. Data mining algorithms usually need to scan through the training data for obtaining the statistics to solve or optimize model parameters. It calls for intensive computing to access the large-scale data frequently. To improve the efficiency of algorithms, Chu et al. proposed a general-purpose parallel programming method, which is applicable to a large number of machine learning algorithms based on the simple Map Reduce programming model on multicore processors. Ten classical data mining algorithms are realized in the framework, including locally weighted linear regression, k-Means, logistic regression, naive Bayes, linear support vector machines, the independent variable analysis, Gaussian discriminant analysis, expectation maximization, and back-propagation neural networks. With the analysis of these classical machine learning algorithms, we argue that the computational operations in the algorithm learning process could be transformed into a summation operation on a number of training data sets. Summation operations could be performed on different subsets independently and achieve penalization executed easily on the Map Reduce programming platform. Therefore, a large-scale data set could be divided into several subsets and assigned to multiple Mapper nodes. Then, various summation operations could be performed on the Mapper nodes to collect intermediate results. Finally, learning algorithms are executed in parallel through merging summation on Reduce nodes. Ranger et al. [39] proposed a Map Reduce-based application programming interface Phoenix, which supports parallel programming in the environment of multicore and multiprocessor systems, and realized three data mining algorithms including k-Means, principal component analysis, and linear regression. Gillick et al. improved the Map Reduce’s implementation mechanism in Hadoop, evaluated the algorithms’ performance of single-pass learning, iterative learning, and query-based learning in the Map Reduce framework, studied data sharing between computing nodes involved in parallel learning algorithms, distributed data storage, and then showed that the Map Reduce mechanisms suitable for large-scale data mining by testing series of standard data mining tasks on medium-size clusters. Papadimitriou and Sun [38] proposed a distributed collaborative aggregation (DisCo) framework using practical distributed data preprocessing and collaborative aggregation techniques.

The implementation on Hadoop in an open source Map Reduce project showed that DisCo has perfect scalability and can process and analyze massive data sets (with hundreds of GB). To improve the weak scalability of traditional analysis software and poor analysis capabilities of Hadoop systems, Das et al. [16] conducted a study of the integration of R (open source statistical analysis software) and Hadoop. The in-depth integration pushes data computation to parallel processing, which enables powerful deep analysis capabilities for Hadoop. Wegener et al. [47] achieved the integration of Weka (an open-source machine learning and data mining software tool) and Map Reduce. Standard Weka tools can only run on a single machine, with a limitation of 1-GB memory. After algorithm parallelization, Weka breaks through the limitations and improves performance by taking the advantage of parallel computing to handle more than 100-GB data on Map Reduce clusters. Ghoting et al. proposed Hadoop-ML, on which developers can easily build task-parallel or data-parallel machine learning and data mining algorithms on program blocks under the language runtime environment. 5.2 Big Data Semantics and Application Knowledge (Tier II)



# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

## 2.1 DISADVANTAGES OF EXISTING SYSTEM:

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.

## III. PROPOSED SYSTEM

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.

## ADVANTAGES OF PROPOSED SYSTEM:

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

## IV. BIG DATA CHARACTERISTICS: HACE THEOREM

**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 naive sense, we can imagine that a number of blind men are trying to size up a giant elephant, which will be the Big Data in this context. The goal of each blind man is to draw a picture (or conclusion) of the elephant 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 elephant "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 1) the elephant is growing rapidly and its pose changes constantly, and 2) each blind man may have his own (possible unreliable and inaccurate) information sources that tell him about biased 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 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 elephant in a real-time fashion. Indeed, this task is not as simple as asking each blind man to describe his feelings about the elephant 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.

### Huge Data with Heterogeneous and Diverse Dimensionality

One of the fundamental characteristics of the Big Data is the huge volume of data represented by heterogeneous and diverse dimensionalities. This is because different information collectors prefer their own schemata or protocols for data recording, and the nature of different applications also results in diverse data representations. For example, each single human being in a biomedical world can be represented by using simple demographic information such as gender, age, family disease history, and so on. For X-ray examination and CT scan of each individual, images or videos are used to represent the results because they provide visual information for doctors to carry detailed examinations. For a DNA or genomic-related test, microarray expression images and sequences are used to represent the genetic code information because this is the way that our current techniques acquire the data. Under such circumstances, the

# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

heterogeneous features refer to the different types of representations for the same individuals, and the diverse features refer to the variety of the features involved to represent each single observation. Imagine that different organizations (or health practitioners) may have their own schemata to represent each patient, the data heterogeneity and diverse dimensionality issues become major challenges if we are trying to enable data aggregation by combining data from all sources.

## Autonomous Sources with Distributed and 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. On the other hand, the enormous volumes of the data also make an application vulnerable to attacks or malfunctions, if the whole system has to rely on any centralized control unit. For major Big Data-related applications, such as Google, Flickr, Facebook, and Walmart, a large number of server farms are deployed all over the world to ensure nonstop services and quick responses for local markets. Such autonomous sources are not only the solutions of the technical designs, but also the results of the legislation and the regulation rules in different countries/ regions. For example, Asian markets of Walmart 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 restructured data representations and data warehouses for local markets.

## Complex and Evolving Relationships

While the volume of the Big Data increases, so do the complexity and the relationships underneath the data. In an early stage of data centralized information systems, the focus is on finding best feature values to represent each observation. This is similar to using a number of data fields, such as age, gender, income, education background, and so on, to characterize each individual. 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.

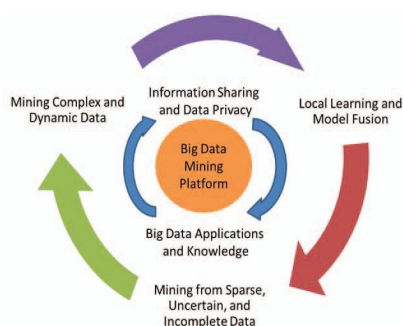


Fig 1. Big data structure

Our friend circles may be formed based on the common hobbies or people are connected by biological relationships. Such social connections commonly exist not only in our daily activities, but also are very popular in cyberworlds. For example, major social network sites, such as Facebook or Twitter, are mainly characterized by social functions such as friend-connections and followers (in Twitter). The correlations between individuals inherently complicate the whole data representation and any reasoning process on the data. In the sample-feature representation, individuals are regarded similar if they share similar feature values, whereas in the sample-feature-relationship representation, two individuals can be linked together (through their social connections) even though they might share nothing in common in the feature domains at all. 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. Such a complication is becoming part of the reality for Big Data applications, where the key is to take the complex (nonlinear, many-to-many) data relationships, along with the evolving changes, into consideration, to discover useful patterns from Big Data collections.



# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

## V. DATA MINING CHALLENGES WITH BIG DATA

For an intelligent learning database system to handle Big Data, the essential key is to scale up to the exceptionally large volume of data and provide treatments for the characteristics featured by the aforementioned HACE theorem 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 privacy and domain knowledge (Tier II), and Big Data mining algorithms (Tier III). 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. For example, typical data mining algorithms require all data to be loaded into the main memory, this, however, is becoming a clear technical barrier for Big Data because moving data across different locations is expensive (e.g., subject to intensive network communication and other IO costs), even if we do have a super large main memory to hold all data 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). For example, depending on different domain applications, the data privacy and information sharing mechanisms between data producers and data consumers can be significantly different. Sharing sensor network data for applications like water quality monitoring may not be discouraged, whereas releasing and sharing mobile users' location information is clearly not acceptable for majority, if not all, applications. In addition to the above privacy issues, the application domains can also provide additional information to benefit or guide Big Data mining algorithm designs. For example, in market basket transactions data, each transaction is considered independent and the discovered knowledge is typically represented by finding highly correlated items, possibly with respect to different temporal and/or spatial restrictions. In a social network, on the other hand, users are linked and share dependency structures. The knowledge is then represented by user communities, leaders in each group, and social influence modeling, and so on. Therefore, understanding semantics and application knowledge is important for both low-level data access and for high-level mining algorithm designs.

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. The circle at Tier III contains three stages. First, sparse, heterogeneous, uncertain, incomplete, and multisource data are preprocessed by data fusion techniques. Second, complex and dynamic data are mined after preprocessing. Third, the global knowledge obtained by local learning and model fusion is tested and relevant information is feedback to the preprocessing stage. Then, the model and parameters are adjusted according to the feedback. In the whole process, information sharing is not only a promise of smooth development of each stage, but also a purpose of Big Data processing. In the following, we elaborate challenges with respect to the three tier framework Tier I: Big Data Mining Platform In typical data mining systems, the mining procedures require computational intensive computing units for data analysis and comparisons. A computing platform is, therefore, needed to have efficient access to, at least, two types of resources: data and computing processors. For small scale data mining tasks, a single desktop computer, which contains hard disk and CPU processors, is sufficient to fulfill the data mining goals. Indeed, many data mining algorithm are designed for this type of problem settings. For medium scale data mining tasks, data are typically large (and possibly distributed) and cannot be fit into the main memory. Common solutions are to rely on parallel computing or collective mining to sample and aggregate data from different sources and then use parallel computing programming (such as the Message Passing Interface) to carry out the mining process.

## VI. CONCLUSION AND FUTURE WORK

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





# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

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. 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.

## REFERENCES

1. Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining," J. Cryptology, vol. 15, no. 3, pp. 177-206, 2002.
2. 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.
3. J. Lorch, B. Parno, J. Mickens, M. Raykova, and J. Schiffman, "Shoroud: Ensuring Private Access to Large-Scale Data in the Data Center," Proc. 11th USENIX Conf. File and Storage Technologies (FAST '13), 2013.
4. D. Luo, C. Ding, and H. Huang, "Parallelization with Multiplicative Algorithms for Big Data Mining," Proc. IEEE 12th Int'l Conf. Data Mining, pp. 489-498, 2012.
5. J. Mervis, "U.S. Science Policy: Agencies Rally to Tackle Big Data," Science, vol. 336, no. 6077, p. 22, 2012.
6. F. Michel, "How Many Photos Are Uploaded to Flickr Every Day and Month?" <http://www.flickr.com/photos/franckmichel/6855169886/>, 2012.
7. T. Mitchell, "Mining our Reality," Science, vol. 326, pp. 1644-1645, 2009.
8. Nature Editorial, "Community Cleverness Required," Nature, vol. 455, no. 7209, p. 1, Sept. 2008.
9. S. Papadimitriou and J. Sun, "Disco: Distributed Co-Clustering with Map-Reduce: A Case Study Towards Petabyte-Scale End-to-End Mining," Proc. IEEE Eighth Int'l Conf. Data Mining (ICDM '08), pp. 512-521, 2008.
10. C. Ranger, R. Raghuraman, A. Penmetsa, G. Bradski, and C. Kozyrakis, "Evaluating Map Reduce for Multi-Core and Multiprocessor Systems," Proc. IEEE 13th Int'l Symp. High Performance Computer Architecture (HPCA '07), pp. 13-24, 2007.
11. A. Rajaraman and J. Ullman, Mining of Massive Data Sets. Cambridge Univ. Press, 2011.
12. A. Reed, D. Thompson, W. Majid, and K. Wagstaff, "Real Time Machine Learning to Find Fast Transient Radio Anomalies: A Semi-Supervised Approach Combining Detection and RFI Excision," Proc. Int'l Astronomical Union Symp. Time Domain Astronomy, Sept. 2011.
13. E. Schadt, "The Changing Privacy Landscape in the Era of Big Data," Molecular Systems, vol. 8, article 612, 2012.
14. J. Shafer, R. Agrawal, and M. Mehta, "SPRINT: A Scalable Parallel Classifier for Data Mining," Proc. 22nd VLDB Conf., 1996. 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.
15. K. Su, H. Huang, X. Wu, and S. Zhang, "A Logical Framework for Identifying Quality Knowledge from Different Data Sources," Decision Support Systems, vol. 42, no. 3, pp. 1673-1683, 2006.

## BIOGRAPHY



Mrs.S.Banupriya has born on 02.04.1990 in Tamilnadu, India. She received BCA degree from Vivekanandha College of Arts and Science for Women, Affiliated to Periyar University, Salem (2007-2010) Tamilnadu (India). She received MCA in 2012 from Vivekanantha College of Arts and Science for Women, Affiliated to Periyar University Tamilnadu, India., She is Pursuing M.phil (full time) from Muthayammal College of Arts & Science in Periyar University, Salem, Tamilnadu, India. Her interested area is Data Mining.



ISSN(Online): 2320-9801  
ISSN (Print) : 2320-9798

## International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Issue 9, September 2015**



V.Vijayadeepa, received her B.Sc. degree from university of Madras and M.Sc. degree and MCA degree from Periyar University .She has completed her M.Phil. at Bharathidasan University .She is having 12 years' experience in collegiate teaching and she is the associate professor and Head of the department of computer science and applications in Muthayammal college of arts and science , Rasipuram affiliated by Periyar university .Her main research interests include personalized web search ,web information retrieval ,data mining, data warehousing and information systems.