Volume 5, Issue 2, February 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Mining High Dimensional Data Sets Using Big Data

G. Yogaraj[1], A. Arumuga Arun[2]

[1] PG Scholar, P.B.College of Engineering, Sriperumbudur, India

[2] Asst. Prof / CSE, Loyola Institute of Technology, Chennai, India

Abstract— ***Data usage has now become vast due to the rapid growth in applications oriented to computers and internet. Big data addresses many of these concerns. These data sets have multiple dimensions and sources. With a greater growth in the field of networking Big data has now spontaneously developed in many areas of science and technology. This research paper processes a Big Data model by presenting a HACE theorem. The proposed model involves many considerations based on various user attributes. We also monitor and analyse various implementation challenges involved in the revolution of Big Data.***

Keywords: Big Data, HACE, mining, security, data-driven model, demand-driven model etc…

I. INTRODUCTION

Big data is a buzzword, or catch-phrase, used to describe a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques. In most enterprise scenarios the data is too big or it moves too fast or it exceeds current processing capacity.While the term may seem to reference the volume of data, that isn't always the case. The term big data, especially when used by vendors, may refer to the technology (which includes tools and processes) that an organization requires to handle the large amounts of data and storage facilities.The term big data is believed to have originated with Web search companies who had to query very large distributed aggregations of loosely-structured data.

As of 2012, limits on the size of data sets that are feasible to process in a reasonable amount of time were on the order of exabytes of data. Scientists regularly encounter limitations due to large data sets in many areas, including meteorology, genomics, connectomics, complex physics simulations, and biological and environmental research. The limitations also affect Internet search, finance and business informatics. Data sets grow in size in part because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, radio-frequency identification readers, and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 Exabyte's (2.5×1018) of data were created. The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization.

Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers". What is considered "big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data set in its domain. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration."

II. RELATED WORK

Due to the multi-source, massive, heterogeneous and dynamiccharacteristics of application data involved in a distributed environment, one of the importantcharacteristics of Big Data is computing tasks on the petabytes (PB), even the exabyte (EB)-level datawith a complex computing process. Therefore, utilizing a parallel computer infrastructure, itscorresponding programming language support, and software models to efficiently analyze and mine thedistributed PB, even EB-level data are the critical goal for Big Data processing to change from “quantity”to “quality”.Currently, Big Data processing mainly depends on parallel programming models like MapReduce, aswell as providing a cloud computing platform of Big Data services for the public. MapReduce is a batchorientedparallel computing model. There is still a certain gap in performance with relational databases.

How to improve the performance of MapReduce and enhance the real-time nature of large-scale dataprocessing is a hot topic in research. The MapReduce parallel programming model has been applied inmany machine learning and data mining algorithms. Data mining algorithms usually need to scan throughthe training data for getting the statistics to solve or optimize model parameters. It calls for intensivecomputing to access the large-scale data frequently. In order to improve the efficiency of algorithms, Chuet al. proposed a general-purpose parallel programming method which is applicable to a large number ofmachine learning algorithms based on the simple MapReduce programming model on multi-coreprocessors. 10 classic data mining algorithms are realized in the framework, including locally weightedlinear regression, k-Means, logistic regression, naive Bayes, linear support vector machines, theindependent variable analysis, Gaussian discriminant analysis, expectation maximization and backpropagationneural networks [Chu et al., 2006].

With the analysis of these classical machine learningalgorithms, we argue that the computational operations in the algorithm learning process could betransformed into a summation operation on a number of training data sets. Summation operations couldbe performed on different subsets independently and achieve penalization executed easily on theMapReduce programming platform. Therefore, a large-scale data set could be divided into several subsetsand assigned to multiple Mapper nodes. Then various summation operations could be performed onthese Mapper nodes to get intermediate results. Finally, learning algorithms are parallel executed throughmerging summation of Reduce nodes.

Ranger et al. [2007] proposed a MapReduce-based applicationprogramming interface Phoenix, which supports parallel programming in the environment of multi-coreand multi-processor systems, and realized three data mining algorithms including k-Means, principalcomponent analysis, and linear regression. Gillick et al. [2006] improved the MapReduce'simplementation mechanism in Hadoop, evaluated the algorithms' performance of single-pass learning,iterative learning and query-based learning in the MapReduce framework, studied how to share databetween computing nodes involved in parallel learning algorithms, how to deal with distributed storagedata, and then showed that the MapReduce mechanisms suitable for large-scale data mining by testingseries of standard data mining tasks on medium-size clusters. Papadimitriou and Sun [2008] proposed adistributed collaborative aggregation (DisCo) framework using practical distributed data pre-processingand collaborative aggregation techniques. The implementation on Hadoop in an open source MapReduceproject showed that DisCo has perfect scalability and can process and analyze massive data sets (withhundreds of GB).

Knowledge evolution is a common phenomenon in real-world systems. For example, the clinician'streatment programs will constantly adjust with the conditions of the patient, such as family economicstatus, health insurance, the course of treatment, treatment effects, and distribution of cardiovascular andother chronic epidemiological changes with the passage of time. In the knowledge discovery process,concept drifting aims to analyze the phenomenon of implicit target concept changes or even fundamentalchanges triggered by context changes in data streams. According to different types of concept drifts,knowledge evolution can take forms of mutation drift, progressive drift, and data distribution drift, basedon single features, multiple feature, and streaming features [Wu et al., 2013].

III. HACE THEOREM

HACE Theorem: Big Data starts with large-volume, heterogeneous, autonomous sources withdistributed and decentralized control, and seeks to explore complex and evolving relationships amongdata.

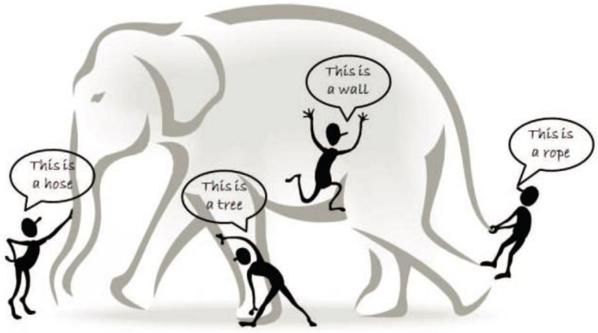

Figure 1 The blind men and the giant elephant: the localized (limited) view of each blind man leads to a biased conclusion.

These characteristics make it an extreme challenge for discovering useful knowledge from the BigData. In a naïve sense, we can imagine that a number of blind men are trying to size up a giant elephant(see Figure 1), 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 collected during the process.Because each person's view is limited to his local region, it is not surprising that the blind men will eachconclude independently that the elephant "feels" like a rope, a hose, or a wall, depending on the regioneach of them is limited to. To make the problem even more complicated, let's assume that (a) the elephantis growing rapidly and its pose also changes constantly, and (b) the blind men also learn from each otherwhile exchanging information on their respective feelings on the elephant. Exploring the Big Data in thisscenario is equivalent to aggregating heterogeneous information from different sources (blind men) tohelp 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 andthen getting an expert to draw one single picture with a combined view, concerning that each individualmay speak a different language (heterogeneous and diverse information sources) and they may even haveprivacy concerns about the messages they deliberate in the information exchange process.

A. Huge Data with Heterogeneous and Diverse Dimensionality

One of the fundamental characteristics of the Big Data is the huge volume of data represented byheterogeneous and diverse dimensionalities. This is because different information collectors use their ownschemata for data recording, and the nature of different applications also results in diverse representationsof the data. For example, each single human being in a bio-medical world can be represented by usingsimple demographic information such as gender, age, family disease history etc. For X-ray examinationand CT scan of each individual, images or videos are used to represent the results because they providevisual 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 becausethis is the way that our current techniques acquire the data. Under such circumstances, the heterogeneousfeatures refer to the different types of representations for the same individuals, and the diverse featuresrefer to the variety of the features involved to represent each single observation. Imagine that differentorganizations (or health practitioners) may have their own schemata to represent each patient, the dataheterogeneity and diverse dimensionality issues become major challenges if we are trying to enable dataaggregation by combining data from all sources.

B. Autonomous Sources with Distributed and Decentralized Control

Autonomous data sources with distributed and decentralized controls are a main characteristic of Big Dataapplications. Being autonomous, each data sources is able to generate and collect information withoutinvolving (or relying on) any centralized control. This is similar to the World Wide Web (WWW) settingwhere each web server provides a certain amount of information and each server is able to fully functionwithout necessarily relying on other servers. On the other hand, the enormous volumes of the data alsomake an application vulnerable to attacks or malfunctions, if the whole system has to rely on anycentralized control unit. For major Big Data related applications, such as Google, Flicker, Facebook, andWalmart, a large number of server farms are deployed all over the world to ensure nonstop services andquick responses for local markets. Such autonomous sources are not only the solutions of the technicaldesigns, but also the results of the legislation and the regulation rules in different countries/regions. Forexample, Asian markets of Walmart are inherently different from its North American markets in terms ofseasonal promotions, top sell items, and customer behaviors. More specifically, the local governmentregulations also impact on the wholesale management process and eventually result in datarepresentations and data warehouses for local markets.

C. Complex and Evolving Relationships

While the volume of the Big Data increases, so do the complexity and the relationships underneath thedata. In an early stage of data centralized information systems, the focus is on finding best feature valuesto represent each observation. This is similar to using a number of data fields, such as age, gender, income,education background etc., to characterize each individual. This type of sample-feature representationinherently treats each individual as an independent entity without considering their social connectionswhich is one of the most important factors of the human society. People form friend circles based on theircommon hobbies or connections by biological relationships. Such social connections commonly exist innot only our daily activities, but also are very popular in virtual worlds. For example, major socialnetwork sites, such as Facebook or Twitter, are mainly characterized by social functions such as friendconnectionsand followers (in Twitter). The correlations between individuals inherently complicate thewhole data representation and any reasoning process. In the sample-feature representation, individuals areregarded similar if they share similar feature values, whereas in the sample-feature-relationshiprepresentation, two individuals can be linked together (through their social connections) even though theymight share nothing in common in the feature domains at all. In a dynamic world, the features used torepresent the individuals and the social ties used to represent our connections may also evolve withrespect to temporal, spatial, and other factors. Such a complication is becoming part of the reality for BigData applications, where the key is to take the complex (non-linear, many-to-many) data relationships,along with the evolving changes, into consideration, to discover useful patterns from Big Data collections.

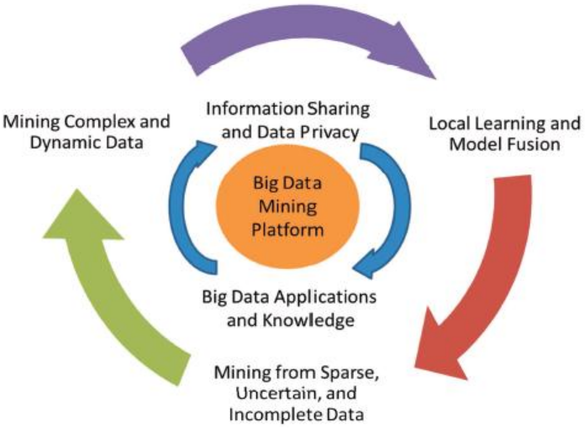
Information Sharing and Data Privacy
Mining Complex and Dynamic Data
Local Learning and Model Fusion
Big Data Mining Platform
Big Data Applications and Knowledge
Mining from Sparse, Uncertain, and Incomplete Data

Figure 2A Big Data processing framework

IV. CHALLENGES IN BIG DATA

For an intelligent learning database system (Wu 2000) to handle Big Data, the essential key is to scale upto the exceptionally large volume of data and provide treatments for the characteristics featured by theaforementioned HACE theorem. Figure 2 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 privacy and domain knowledge (Tier II), and Big Data mining algorithms (Tier III).The challenges at Tier I focus on data accessing and actual computing procedures. Because Big Dataare often stored at different locations and data volumes may continuously grow, an effective computingplatform will have to take distributed large-scale data storage into consideration for computing. Forexample, while typical data mining algorithms require all data to be loaded into the main memory, this isbecoming a clear technical barrier for Big Data because moving data across different locations isexpensive (e.g., subject to intensive network communication and other IO costs), even if we do have asuper large main memory to hold all data for computing.

The challenges at Tier II centeraround semantics and domain knowledge for different Big Dataapplications. Such information can provide additional benefits to the mining process, as well as addtechnical barriers to the Big Data access (Tier I) and mining algorithms (Tier III). For example, dependingon different domain applications, the data privacy and information sharing mechanisms between dataproducers and data consumers can be significantly different. Sharing sensor network data for applicationslike 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 aboveprivacy issues, the application domains can also provide additional information to benefit or guide BigData mining algorithm designs.

At Tier III, the data mining challenges concentrate on algorithm designs in tackling the difficultiesraised by the Big Data volumes, distributed data distributions, and by complex and dynamic datacharacteristics. The circle at Tier III contains three stages. Firstly, sparse, heterogeneous, uncertain,incomplete, and multi-source data are preprocessed by data fusion techniques. Secondly, complex anddynamic data are mined after pre-processing. Thirdly, the global knowledge that is obtained by locallearning and model fusion is tested and relevant information is fed back to the pre-processing stage. Thenthe model and parameters are adjusted according to the feedback. In the whole process, informationsharing is not only a promise of smooth development of each stage, but also a purpose of Big Dataprocessing.

A. Tier I: Big Data Mining Platform

In typical data mining systems, the mining procedures require computational intensive computing unitsfor 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, asingle desktop computer, which contains hard disk and CPU processors, is sufficient to fulfill the datamining goals. Indeed, many data mining algorithm are designed to handle this type of problem settings.For medium scale data mining tasks, data are typically large (and possibly distributed) and cannot be fitinto the main memory. Common solutions are to rely on parallel computing (Shafer et al. 1996; Luo et al.2012) or collective mining (Chen et al. 2004) to sample and aggregate data from different sources andthen use parallel computing programming (such as the Message Passing Interface) to carry out the miningprocess.

B. Tier II: Big Data Semantics and Application Knowledge

Semantics and application knowledge in Big Data refer to numerous aspects related to the regulations,policies, user knowledge, and domain information. The two most important issues at this tier include (1)data sharing and privacy; and (2) domain and application knowledge. The former provides answers toresolve concerns on how data are maintained, accessed, and shared; whereas the latter focuses onanswering questions like "what are the underlying applications ?" and "what are the knowledge orpatterns users intend to discover from the data ?".

C. Tier III: Big Data Mining Algorithms

As Big Data applications are featured with autonomous sources and decentralized controls, aggregatingdistributed data sources to a centralized site for mining is systematically prohibitive due to the potentialtransmission cost and privacy concerns. On the other hand, although we can always carry out miningactivities at each distributed site, the biased view of the data collected at each different site often leads tobiased decisions or models, just like the elephant and blind men case. Modelmining and correlations are the key steps to ensure that models or patterns discovered from multipleinformation sources can be consolidated to meet the global mining objective. At themodel or pattern level, each site can carry out local mining activities, with respect to the localized data, todiscover local patterns. By exchanging patterns between multiple sources, new global patterns can besynthetized by aggregating patterns across all sites (Wu and Zhang 2003). At the knowledge level, modelcorrelation analysis investigates the relevance between models generated from different data sources todetermine how relevant the data sources are correlated to each other, and how to form accurate decisionsbased on models built from autonomous sources.

V. CONCLUSION

Driven by real-world applications and key industrial stakeholders and initialized by national fundingagencies, managing and mining Big Data have shown to be a challenging yet very compelling task. Whilethe term Big Data literally concerns about data volumes, our HACE theorem suggests that the keycharacteristics of the Big Data are (1) huge with heterogeneous and diverse data sources, (2) autonomouswith distributed and decentralized control, and (3) complex and evolving in data and knowledgeassociations. Such combined characteristics suggest that Big Data requires a "big mind" to consolidatedata for maximum values (Jacobs 2009).

In order to explore Big Data, we have analyzed several challenges at the data, model, and systemlevels. To support Big Data mining, high performance computing platforms are required which imposesystematic designs to unleash the full power of the Big Data. At the data level, the autonomousinformation sources and the variety of the data collection environments, often result in data withcomplicated conditions, such as missing/uncertain values. In other situations, privacy concerns, noise anderrors can be introduced into the data, to produce altered data copies. Developing a safe and soundinformation sharing protocol is a major challenge. At the model level, the key challenge is to generateglobal models by combining locally discovered patterns to form a unifying view. This requires carefullydesigned algorithms to analyze model correlations between distributed sites, and fuse decisions frommultiple sources to gain a best model out of the Big Data. At the system level, the essential challenge isthat a Big Data mining framework needs to consider complex relationships between samples, models, anddata sources, along with their evolving changes with time and other possible factors. A system needs to becarefully designed so that unstructured data can be linked through their complex relationships to formuseful patterns, and the growth of data volumes and item relationships should help form legitimatepatterns 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 scienceand engineering domains. With Big Data technologies, we will hopefully be able to provide most relevantand most accurate social sensing feedback to better understand our society at real-time. We can furtherstimulate the participation of the public audiences in the data production circle for societal andeconomical events. The era of Big Data has arrived.

REFERENCES

[1] Ahmed and Karypis 2012, Rezwan Ahmed, George Karypis, Algorithms for mining the evolution ofconserved relational states in dynamic networks, Knowledge and Information Systems, December2012, Volume 33, Issue 3, pp 603-630

[2] Alam et al. 2012, Md. HijbulAlam, JongWoo Ha, SangKeun Lee, Novel approaches to crawlingimportant pages early, Knowledge and Information Systems, December 2012, Volume 33, Issue 3, pp707-734

[3] Aral S. and Walker D. 2012, Identifying influential and susceptible members of social networks,Science, vol.337, pp.337-341.

[4] Machanavajjhala and Reiter 2012, AshwinMachanavajjhala, Jerome P. Reiter: Big privacy: protectingconfidentiality in big data. ACM Crossroads, 19(1): 20-23, 2012.

[5] Banerjee and Agarwal 2012, Soumya Banerjee, Nitin Agarwal, Analyzing collective behavior fromblogs using swarm intelligence, Knowledge and Information Systems, December 2012, Volume 33,Issue 3, pp 523-547

[6] Birney E. 2012, The making of ENCODE: Lessons for big-data projects, Nature, vol.489, pp.49-51.

[7] Bollen et al. 2011, J. Bollen, H. Mao, and X. Zeng, Twitter Mood Predicts the Stock Market, Journalof Computational Science, 2(1):1-8, 2011.

[8] Borgatti S., Mehra A., Brass D., and Labianca G. 2009, Network analysis in the social sciences,Science, vol. 323, pp.892-895.

[9] Bughin et al. 2010, J Bughin, M Chui, J Manyika, Clouds, big data, and smart assets: Ten techenabledbusiness trends to watch, McKinSey Quarterly, 2010.

[10] Centola D. 2010, The spread of behavior in an online social network experiment, Science, vol.329,pp.1194-1197.

[11] Chang et al., 2009, Chang E.Y., Bai H., and Zhu K., Parallel algorithms for mining large-scale richmediadata, In: Proceedings of the 17th ACM International Conference on Multimedia (MM '09), NewYork, NY, USA, 2009, pp. 917-918.

[12] Chen et al. 2004, R. Chen, K. Sivakumar, and H. Kargupta, Collective Mining of Bayesian Networksfrom Distributed Heterogeneous Data, Knowledge and Information Systems, 6(2):164-187, 2004.

[13] Chen et al. 2012, Yi-Cheng Chen, Wen-Chih Peng, Suh-Yin Lee, Efficient algorithms for influencemaximization in social networks, Knowledge and Information Systems, December 2012, Volume 33,Issue 3, pp 577-601

[14] Chu et al., 2006, Chu C.T., Kim S.K., Lin Y.A., Yu Y., Bradski G.R., Ng A.Y., Olukotun K., Mapreducefor machine learning on multicore, In: Proceedings of the 20th Annual Conference on NeuralInformation Processing Systems (NIPS '06), MIT Press, 2006, pp. 281-288.

[15] Cormode G. and Srivastava D. 2009, Anonymized Data: Generation, Models, Usage, in Proc. OfSIGMOD, 2009. pp. 1015-1018.

[16] Das et al., 2010, Das S., Sismanis Y., Beyer K.S., Gemulla R., Haas P.J., McPherson J., Ricardo:Integrating R and Hadoop, In: Proceedings of the 2010 ACM SIGMOD International Conference onManagement of data (SIGMOD '10), 2010, pp. 987-998.

[17] Dewdney P., Hall P., Schilizzi R., and Lazio J. 2009, The square kilometre Array, Proc. of IEEE,vol.97, no.8.

[18] Domingos and Hulten, 2000, Domingos P. and Hulten G., Mining high-speed data streams, In:Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and DataMining (KDD□00), 2000, pp. 71-80.

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

[20] Efron B. 1994, Missing data, imputation, and the Bootstrap, Journal of the American StatisticalAssociation, vol.89, no.426, pp.463-475.

[21] Ghoting et al., 2009, Ghoting A., Pednault E., Hadoop-ML: An infrastructure for the rapidimplementation of parallel reusable analytics, In: Proceedinds of the Large-Scale Machine Learning:Parallelism and Massive Datasets Workshop (NIPS-2009).

[22] Gillick et al., 2006, Gillick D., Faria A., DeNero J., MapReduce: Distributed Computing for MachineLearning, Berkley, December 18, 2006.

[23] Helft M. 2008, Google uses searches to track Flu's spread, The New York Times,http://www.nytimes.com/2008/11/12/technology/internet/12flu.html.