BIG DATA CHALLENGES AND ISSUES: REVIEW ON ANALYTIC TECHNIQUES

Saritha K

School of Computer Sciences, Mahatma Gandhi University,
Kottayam, Kerala, India
sarithakris@gmail.com

Sajimon Abraham

School of Management and Business Studies
Mahatma Gandhi University,
Kottayam, Kerala,India
sajimabraham@rediffmail.com

Abstract

The cumulative growth of data from various sources has led to the era of big data. Big Data analytics give rise opportunities in designing of competitive offer packages for customers to provide reliable services, but analysis must be accurate and timely for successful decision making. To understand the different fields of Big Data its storage, processing, analytics and finally the visualization has to be described. As data storage in Big Data is not coping with the traditional databases, distributed file systems and NoSQL databases are normally used. MapReduce programming model is commonly used for processing such type of unstructured data. For finding useful patterns and information from Big Data statistics as well as machine learning techniques are employed. Here we reviewed different steps involved in predictive modeling with partitioning for statistical analysis. Artificial Neural Network is used for analytics and it is projected that this will produce a strong training phase for future analytics.

Keywords: Data processing, Hadoop Distributed File system, NoSQL, Data analytics, Predictive Modeling, Linear Regression, Neural Network.

1. Introduction

The extensive use of digital equipments such as mobile phones, sensors, purchase transactions and social media networks has led to the exponential increase of data [8]. The large volume of complex and growing data generated from many distinct sources has led to the era of Big Data. This huge collection of data may contain hidden insights or intelligence. When examine properly, Big Data can bring new business insights, open new market and create competitive advantages. We can derive the hidden knowledge from this large volume of data through the careful analysis. For this a new scientific model has been bourn as Data Intensive Scientific Discovery (DISD) also known as Big Data Analytics [19].

The International Data Corporation IDC [1] reports that the marketing of Big Data is about $16.1 billion in 2014. Another report of IDC [2] forecasts that it will grow up to $32.4 billion by 2017. The report of [3] and [4] pointed out that the marketing of Big Data shows an annual growth of 26% that is $46.34 billion and $114 billion by 2018 respectively The research firm Markets and Markets predicts that the global big data market will show a 26 percent compound annual growth rate from 2013 to 2018. Even though the marketing values of Big data in these researches and technology reports [1- 5] are different, these forecasts usually indicate that the scope of Big Data will grown rapidly in the forthcoming future.

The characteristics of Big Data can be summarized in terms of different V's. *Volume*- Volume focuses on the size of the data set. The volume of Big Data may reach the level of terabytes or even petabytes which is far beyond the usual limits of megabytes or gigabytes. *Velocity*- It indicates the speed of the data in and out. It refers to the rate of recurrence of the data generation, the dynamic features of data and the importance of generating results in real time. For deal with these challenging tasks we need new technologies that have the capacity of extracting meaningful information from large and diversified data is arriving continuously. *Variety*- Big Data is the combination of varieties of data such as structured, unstructured and semi structured data. So it is a challenging task to deal with different varieties of data captured from various sources that arriving. *Veracity*- Combining dissimilar data from different sources provide valuable insights rather than isolated data. *Variability*-

Variability refers to the data whose meaning is constantly changing. *Value*- The possible value of Big Data is huge. Actually Big Data starts with large volume, heterogeneous, autonomous sources with distributed and decentralized control are very complex and it is a tedious task to generate relationship among data.

Hai Wang et al. [19] presents in their work the challenges and new trends faced by Big Data while taking decision making. C L Philip Chen and Chu-yang Zhang [8] in their paper has discussed about various methods to manage the flooded data, modular, cloud, bio-inspired and quantum computing. Muhammad Bilal et al. [6] give detailed review on the topic Big Data as Big Data Engineering and Big Data Analytics. Here the authors focus on Statistics, Mining and Artificial Neural Network technologies. Schneider, A et al. [15] in their review article discuss the performance and interpretation of linear regression analysis. Regression models such as simple multivariable and multivariate are detailed here.

The structure of this paper is as follows. In section 2 gives the details of the sub field of Big Data section 3 describes the predictive modeling and partitioning the Big Data. Section 4 contains the nonproductive modeling with artificial neural network. We conclude with a brief discussion in the last section.

2. Subtleties of Big Data

Analytics of Big Data involves a lot of challenges due to its various characteristics. A group of limitations faced in large data management includes scalability, unstructured data, accessibility, fault tolerance etc. It is noticeable that Big Data can bear intelligent and fortunate decisions for organizations. From the perspective of decision making, the process of data intensive science can be interpreted by the Fig.1. To understand the different fields of data intensive science we have a clear idea about each aspect [8].

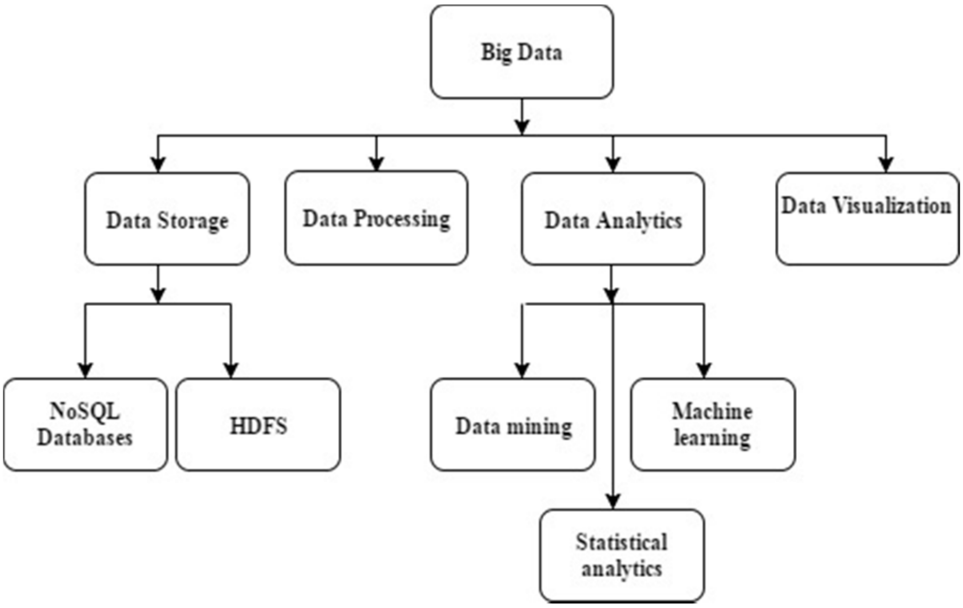

Fig 1. Different fields of Big Data

2.1. Data Storage

One of the main aspects is Storage, which is provided either by Distributed File Systems or NoSQL databases [2]. Big Data has changed the way of capture and store, including data storage devices, architecture of the storage devices and data access mechanisms.

- *Hadoop Distributed File System (HDFS)*

HDFS is mainly used for managing larger datasets with cluster of commodity hardware with streaming access patterns. Streaming access patterns means “write ones, read any number of times, but no change in the content of the file”. The chances of hardware failure are higher in such a settings, it provides greater fault tolerance for hard ware failures. In HDFs large files are broken into chunks and stored across multiple data nodes as local files [10]. The key traits of HDFS to achieve fault tolerance and high availability are data distribution and replication. There is however situations where the usage of HDFS degrades performance particularly in application requiring low latency data access. Similarly it is also not ideal for storing a large number of small files due to the associated overhead for managing their metadata.

HDF is based on typical master slave architecture [13]. An HDFS cluster is made up of a single name node, which is the master node, this node holds and manage whole metadata of HDFS. This name node the central control point and many redistribute replicas as needed. The computing systems in each cluster are called Data nodes [4] and these nodes act as the slave nodes. Data node holds the blocks /units of a file, which runs at the clients in parallel.

- NoSQL Databases

Relational databases served the IT industry for the past couple of decades. But recently emerged applications demands more scalability, performance and flexibility [7]. Relational databases are found unsuitable for these applications due to their specialized storage and processing needs. Now a new system came into being called NoSQL – Not Only SQL – which is widely used to handle unstructured data. It basically provides flexibility and scalability. There is no schema used in NoSQL databases which is very helpful for dealing with huge amount of unstructured data. Different data models used in NoSQL are [12]
Key- Value pair :- this is the simplest model. There is no schema used in the key – value pair store like in RDBMS which provides great flexibility. There is one unique key and a particular value corresponding to that. Key – Value pair is used in MapReduce by Map function.
Document oriented :- It is one of the famous NoSQL database. Here data are stored as document with no schema and also without the concept of normalization.
Columnar :- This model favors the storage of sparse data sets, grouped sub-columns and aggregated columns.
Graph :-This model used to store the data in the form of nodes and edges. It is very easy and better approach than RDBMS due to its processing convenience. It is very helpful in performing graph like queries.

Prominent NoSQL databases and their features are described in various literatures and a consolidated form of the features are detailed in Table-1 [6]

Table 1. Prominent NOSQL systems and there features.

Product Name	Product Description	Data Model	Concurrency	Storage	Key features
Cassandra	Apache Cassandra is a scalable database that provides fault tolerance and consistency on of commodity servers	Columnar Key-value	MVCC	Disk, Hadoop	High availability, Tolerance of partition
Base	HBase is a distributed data store that extended Google Big table to scale on HDFS. Its novelty lies in storing and accessing data with random access. It does not restrict the kind of data being stored.	Columnar Key-value	Locks	Hadoop	Consistent, Tolerance of partition
MongoDB	MongoDB is a document-oriented data base. It facilitates storage of documents with variable schemas and is suitable for applications that storing complex type data.	Columnar Document	Locks	Disk, GFS	Consistent, Tolerance of partition
CouchDB	CouchDB is suitable for large scale web and mobile applications. It facilitate data storage that are queried through web browsers via HTTP. JavaScript is used to index, integrate, and transforms database.	Document Key-value	MVCC	Disk	High availability, Tolerance of partition
BarkeleyDB	BarkeleyDB is an embedded database for key-value data set.	Key-value	ACID	RDF	High availability, Consistent, Tolerance of partition
Riak	It is a distributed database that provides scalability and high availability. It achieves performance and fault tolerance through built in distribution and replications.	Key-value	ACID	Disk	High availability, Tolerance of partition

2.2 Data Processing

Parallel and distributed computations are the core of data processing. A large number of processing models are developed for this purpose and one of the best algorithms for this is MapReduce. Traditional systems have centralized servers to store and process data. This system creates too much bottleneck while processing file simultaneously. Google solved this issue by using an algorithm called MapReduce [13]. MR is based on the divide and conquers methods, and works by recursively breaking down a complex problem in to many sub problems, until this sub problem is scalable for solving directly. After that this sub problems are assigned to a cluster of working nodes and solved in separate and parallel ways. Finally, the solutions to the sub problems are combined to given a solution to the original problem. The divide and conquer method is implemented by two steps, Map step and Reduce step , which are submitted to separate processes called Mappers and Reducers [11]. The Fig. 2 shows different phases of MapReduce programming model. The Map task takes a set of data and converts it into another set of data, where individual elements are broken down into key-value pair. The Reduce task takes the output from the Map task and combines them and it requires a wide range of processing. Once the execution is over, it gives zero or more key-value pairs to the final step.

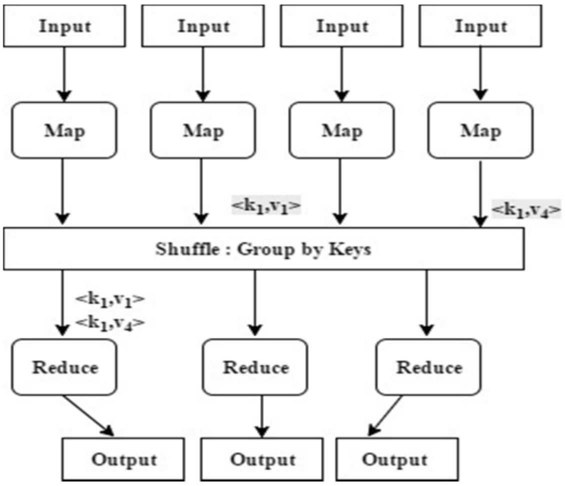

Fig 2. MapReduce architecture.

2.3 Data Analytics

The first impression of big data is its volume, so the biggest and most important challenges is its scalability [8].There have been traditionally many related disciplines that have essentially the same core focus, finding useful patterns in data. Big data analytics is broadening of the field of data analytics and in cooperates many of the techniques have already been performed. The data analysis has to be employed with new technologies and techniques to capture correct value from it. Extra ordinary techniques are needed for efficiently process big volume of data within time. Different data analytics techniques are data mining, statistics and machine learning. Table 2 gives the details of programming tools used for Big data Analytics [6, 8].

2.3.1 Data Mining

Various technologies are used to dig out precious knowledge form the data. These technologies include clustering, classification, regression model and association rule mining. Traditional mining algorithms have many challenges in the case of Big Data mining [11]. The present algorithms are not competent to get the needed information from Big Data and hence some improvisation is needed. Parallel computing [42] and reduction of data dimension [43] are to be considered. In [43] the authors first propose an extreme learning machine tree (ELM-Tree) model based on the heuristics of uncertainty reduction. In the ELM-Tree model, information entropy and ambiguity are used as the uncertainty measures for splitting decision tree (DT) nodes Lot of classification and clustering algorithm has developed for Big Data samples, such as fuzzy, clustering and so on. The efficiency of data mining algorithms can be improved by properly manipulating the data base activities such as data access, query optimization, ordering and grouping.

2.3.2 Statistics

Statistics is the study of collection, analysis and finalization of conclusion with main focus on the correct tool and methods. The present statistical techniques are not match to manipulate Big Data and therefore many researchers have developed new techniques [12]. The core issue for taking decisions by using statistical technique is the management of fundamental relationship and co-relationship among samples. New techniques such as parallel statistics, statistical computing and learning are developed for overcoming the difficulties of the traditional techniques. Map-Reduce programming model is designed for applying parallel statistics in Big Data Analytics. Parallel computing platform can be created with the help of Hadoop. For exploiting patterns from huge data machine learning is combined with statistical techniques and this process can be treated as statistical learning.

2.3.3 Machine learning

Machine learning (ML) is an essential subfield of Artificial intelligence which is used to design algorithms that allow computer to develop behaviors based on experimental data. ML tasks can be categorized into Supervised learning, Unsupervised learning, and Association model [7].

- *Unsupervised learning* is also referred to as Clustering. Here to find clusters that has similarity in their characteristics. Data in one cluster are similar between each other while there is no similarity between different clusters. Traditional clustering algorithms are insufficient in the case of Big Data analytics because they typically require all the data is in the same format and be loaded into the same machine so as to find some useful things from the whole data [11]. In [9], Chiang M C et. al suggest that analyzing large scale and high dimensionality dataset using k-means clustering algorithms. This paper presents an efficient algorithm, called pattern reduction (PR), for reducing the computation time of k-means and k-means-based clustering algorithms. The characteristics of Big Data have several new challenges for data clustering issues. In [17], Shirkhorshidi et al. divide the big data clustering in to two categories as single-machine clustering and multiple-machine clustering with the help of parallel and MapReduce algorithms. Xu H et al [21] presents that CloudVista system is used to address the problems with big data caused by using existing data reduction approaches. Here clustering is performing parallel and RandGen algorithm is used to achieve an efficient balance between interactivity and batch processing. Feldman et al. [22] proposed in their work how convert the traditional clustering algorithm to big data clustering problems with the help of tree structure of merge and reduce approach. Here the authors propose to divide the data set with the merge-and-reduce approach and use parallel streaming algorithms for problems such as *k*-means, PCA and projective clustering.

Table 2 Big Data Analytics tools

Tool name	Description	Supported languages	ML at scale	Supported algorithms
Apache Mahout	Mahout is an open source machine learning frame work for quickly writing scalable and high performance ML applications.	Java	Yes	Collaborative filtering Classification Clustering Regression
R	R is an open source programming language for statistical analysis. R is extremely extensible. With huge developer base, thousands of R packages are available to provide verity of functionalities. The graphics supported by R are highly polished and very powerful.	Many languages	Yes	Collaborative filtering Classification Clustering Regression
MLbase	Spark has constituted a novel ML platform called MLbase., which has brought together highly robest ML component., such as ML optimizer, MLlib and MLI to support the full life cycle activities required to implement as well as use ML algorithms.	Java Python	Yes	Collaborative filtering Classification Clustering Regression
Oryx	Oryx is an open source ML library that has evolved over time out of the libraries and toolkits developed by Cloudera. An interesting feature of Oryx is its ability to keep the model updated under emerging streams of data from Hadoop.	Java	Yes	Collaborative filtering Classification Clustering Regression

- *Supervised learning* is also known as Classification, take decisions based on the trained decisions taken previously. As the classification algorithm mainly learned by training, correct trained decisions are to be set aside to learn accurately. Naïve Bayes Classifier is a simple and popular algorithm for supervised classification algorithm. Jiang et al. [23] presented Bayesian classification methodology for detecting the damages in building structural design of construction industry. The effectiveness of the proposed technique is reporting the damaged goods in building a multi-storey building. Artificial neural network algorithms are good for classification. Most of the currently employed ANNs for artificial intelligence are base on statistical estimation [24], classification optimization [25] and control theory [26].

- *Association model* is used for social network analysis like sentimental analysis, web mining and text mining etc. Because social network is part of the daily life of people and its data is also a kind of Big Data. Analysis of a social network can be used to predict the behavior of a user. Social Network Analysis includes social system design [27], human behavior modeling [28], and graph query mining [29]. To enhance the efficiency of these algorithms with the help of Big data platforms like HDFS and MapReduce.

3. Predictive Modeling and Partitioning the Data Set

Actually statistics is an essential part of Big Data analytics. Hence different statistical techniques are used for the same. In statistical modeling of dataset describes the causal relationship among two attributes or along with several attributes. The attribute which is predictable is called the dependent variable or the response variable. While the attribute that is used for predicting the cause of response variable is called independent variable or predictor variable. Statistical dependences between the attributes give the measure of association among the attributes. If all the attributes are quantitative, that is continuous or discrete, and then a correlation matrix is calculated as a measure of the correlation between the relationships among them. But in the case of qualitative attributes the number of occurrences of a particular class can be obtained. Many kinds of Statistical models are used in the different disciples for the purpose of model building and testing, one finds a collection of perceptions about the relationship between causal explanation and observed prediction [10]. The process of predictive model consists of different steps such as Define the goal, Data collection and management, Data preprocessing, Exploratory data analysis, Build the model and Interpreting the result [22,23].

In the era of Big Data we need efficient statistical methods are needed for predicting the outcomes [30]. The industry using statistical methods in various applications such as causes of construction delay [31], decision support for construction litigation [32], identifying action of workers and heavy machinery [33] etc. In [15] Schneider A et al. the authors give a brief overview of different linear regression models, illustrative examples are given and the results are interpreted. In [34], Jun S et al. propose a new analytical methodology for Big Data analysis in regression problems for reducing the computing burden. In [35], Elizabeth D s et al. presents statistical methods for online analytical processing. Here the authors develop iterative estimating algorithms and statistical inference for linear model and estimating the equations that update as new data arrive. The dataset used for this is the airline on-time statistics.

4. Nonparametric Modeling with Artificial Neural Networks (ANN)

ANN algorithms are well suited to problems of classification or function estimation. Since the invention of ANN algorithms, these are widely used in the field of solving complex industrial problems. Multilayer Perceptron (MLP) is the most commonly used type of ANN [12]. ANNs is typically made up of three layers: input layer, hidden layer and output layer. Data samples in MLP Neural Networks are normalized and are fed into the input layer. This data moves from the input layer to one or two hidden layers and is finally passed on the output layer, producing an output of the given ANN algorithms. During training phase, the weight values between the connections are adjusted. Back propagation is a commonly used algorithm for training the ANNs. Recently ANNs have been applied for scales forecasting, because they have very promising performance in the area of control, prediction and pattern recognition. However, there is big challenge if it is employed for analyzing Big Data because of the natural contradiction between the necessity of more hidden layers and nodes for higher performance and the memory and time consuming in a NN [3]. Two approaches have been adopted to ease the contradiction. The first one is to reduce the size of the data by sampling techniques, and the second one is to put NNs in parallel and distributed settings [1]. ANN algorithms have recently brought revolution in machine learning through deep learning. New algorithms of ANN are designed to learn from high dimensionality data, which seek special attention in all the supply chain industry applications where ANN is employed.

In [37], Moselhi et al. propose the usefulness of ANN over the conventional expert-based systems. Here the authors employed for developing various applications for the construction industry. In [37], Frank C et al. propose the statistical time series model and ANN based model were investigated for forecasting women’s clothing sales. Sun Z L et al. [38] give a detailed study of novel ANN techniques called extreme learning machine (ELM) to investigate the relationship between sales amount and some significant factors which affect demand. In [39] Kuo and Xue reported that ANNs are better than many conventional statistical forecasting methods. Dursun Delen and others [40] suggest to combining two popular data mining algorithms such as artificial neural networks and decision trees along with a statistical method, logistic regression, to develop a prediction models using a large dataset. Jose M et al [41], employed that artificial neural networks combined with decision trees to search through large datasets seeking subtle patterns in prognostic factors, and that may further assist the selection of appropriate treatments for the individual patient.

5. Conclusion

In this review paper, we give a brief overview on Big Data problems and analysis techniques by using statistical as well as machine learning methods. Big Data analytics is still in the in initial stage of developments. Existing Big Data techniques and tools are very limited to solve the real problems completely. In this paper the various steps in statistical modeling process is described. Here we propose to partition the large data set, unlike traditional statistical analysis. Though various statistical methods are developed for analysis, the results obtained are casual relationship and correlation ship among data. So we propose to introduce a strong training phase for analysis using Back propagation algorithms in ANN and we think the most beneficial will be to use ANN for classification algorithms.

References

[1] Big data market:2014 prediction from IDC and IIA Forbers Tech report 2013 [online] Avilable: http://www.forbes.com/sites/gilperss/2013/12/12/12/16-1-billion-big-data-market-2014-prediction-from-idc-and-iia/.

[2] **Big** Data and analytics-[online] Available:http://www.idc.com/promo/thirdplatform/fourpillars/bigdataanalytics

[3] Big Data Market to Reach $46.34 Billion by 2018 Available: http://www.eweek.com/database/big-data-market-to-reach-46.34-billion-by-2018

[4] Big Data Spending to Reach $114 Billionin 2018; Look For Machine Learning to Drive Anaytics, ABI research, Available: https://www.abiresearch.com/press/big-data-spending-to-reach-114-billion-in-2018-loo/

[5] Big Data Market Size and Vendor Revenues *By Jeff Kelly with David Vellante and David Floye Available: http://wikibon.org/wiki/v/Big_Data_Market_Size_and_Vendor_Revenues*

[6] Bilal, M., Oyedele, L. O., Qadir, J., *et al*. (2016). *Big Data in the construction industry: A review of present status, opportunities, and future trends*. Advanced Engineering Informatics, 30(3), 500-521

[7] Bansal, S., & Rana, D. A. (2014). *Transitioning from Relational Databases to Big Data*. International Journal of Advanced Research in Computer Science and Software Engineering, 4(1).

[8] Chen, C. P., & Zhang, C. Y. (2014). *Data-intensive applications, challenges, techniques and technologies: A survey on Big Data*. Information Sciences, *275*, 314-347.

[9] Chiang, M. C., Tsai, C. W., & Yang, C. S. (2011). A time-efficient pattern reduction algorithm for k-means clustering. *Information Sciences*, *181*(4), 716-731.

[10] Kumar, P., & Rathore, D. V. S. (2014). Efficient capabilities of processing of big data using hadoop map reduce. *International Journal of Advanced Research in Computer and Communication Engineering*, *3*(6), 7123-6..

[11] Mary, A. J. J., & Arockiam, L. (2015). A Study on Basic Concepts of Big Data. *International Journal*, *1*(3).

[12] Singh, J., & Singla, V. (2015). Big Data: Tools and Technologies in Big Data. *International Journal of Computer Applications*, *112*(15).

[13] Maitrey, S., & Jha, C. K. (2015). MapReduce: Simplified data analysis of Big data. *Procedia Computer Science*, *57*, 563-571.

[14] Padhy, N., & Mishra, D. P. (2012). 2, and RasmitaPanigrahi3 "The Survey of Data Mining Applications And Feature Scope" International Journal of Computer Science. *Engineering and Information Technology (IJCSEIT)*, *2*(3).

[15] Schneider, A., Hommel, G., & Blettner, M. (2010). Linear Regression Analysis. *Dtsch Ä Rztebl Int*, *107*(44), 776-82.

[16] Shmueil, G.(2010), "To Explin or Predict?", Statistical science, vol25 © Institute of Mathematical Science.

[17] Shirkhorshidi, A. S., Aghabozorgi, S., Wah, T. Y., & Herawan, T. (2014, June). Big data clustering: a review. In *International Conference on Computational Science and Its Applications* (pp. 707-720). Springer International Publishing.

[18] Tsai, C. W., Lai, C. F., Chao, H. C., & Vasilakos, A. V. (2015). Big data analytics: a survey. *Journal of Big Data*, *2*(1), 21.

[19] Wang, H., Xu, Z., Fujita, H., & Liu, S. (2016). Towards felicitous decision making: An overview on challenges and trends of Big Data. *Information Sciences*, *367*, 747-765.

[20] Yıldırım, A. A., Özdoğan, C., & Watson, D. (2014). Parallel data reduction techniques for big datasets. *Big data management, technologies, and applications*, 72-93.

[21] Xu, H., Li, Z., Guo, S., & Chen, K. (2012). Cloudvista: interactive and economical visual cluster analysis for big data in the cloud. *Proceedings of the VLDB Endowment*, *5*(12), 1886-1889.

[22] Feldman, D., Schmidt, M., & Sohler, C. (2013, January). Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 1434-1453). Society for Industrial and Applied Mathematics.

[23] Jiang, X., & Mahadevan, S. (2008). Bayesian probabilistic inference for nonparametric damage detection of structures. *Journal of engineering mechanics*, *134*(10), 820-831.

[24] Mansour, Y., Chang, A. Y., Tamby, J., Vaahedi, E., Corns, B. R., & El-Sharkawi, M. A. (1997). Large scale dynamic security screening and ranking using neural networks. *IEEE Transactions on Power Systems*, *12*(2), 954-960.

[25] Oh, C., & Zak, S. H. (2010). Large-scale pattern storage and retrieval using generalized brain-state-in-a-box neural networks. *IEEE transactions on neural networks*, *21*(4), 633-643.

[26] Liu, Y. J., Chen, C. P., Wen, G. X., & Tong, S. (2011). Adaptive neural output feedback tracking control for a class of uncertain discrete-time nonlinear systems. *IEEE Transactions on Neural Networks*, *22*(7), 1162-1167.

[27] Zhang, Y., & van der Schaar, M. (2012). Information production and link formation in social computing systems. *IEEE Journal on Selected Areas in Communications*, *30*(11), 2136-2145.

[28] Lane, Nicholas D., Ye Xu, Hong Lu, Andrew T. Campbell, Tanzeem Choudhury, and Shane B. Eisenman. "Exploiting social networks for large-scale human behavior modeling." *IEEE Pervasive Computing* 10, no. 4 (2011): 45-53.

[29] Ma, H., King, I., & Lyu, M. R. (2012). Mining web graphs for recommendations. *IEEE Transactions on Knowledge and Data Engineering*, *24*(6), 1051-1064.

[30] Ha, S., Lee, S., & Lee, K. (2014). Standardization Requirements Analysis on Big Data in Public Sector based on Potential Business Models. *International Journal of Software Engineering and Its Applications*, *8*(11), 165-172.

[31] Kim, H., Soibelman, L., & Grobler, F. (2008). Factor selection for delay analysis using knowledge discovery in databases. *Automation in Construction*, *17*(5), 550-560.

[32] Mahfouz, T. S. (2009). Construction legal support for differing site conditions (DSC) through statistical modeling and machine learning (ML).

[33] Huang, Y., & Beck, J. L. (2013). Novel sparse Bayesian learning for structural health monitoring using incomplete modal data. In *Computing in Civil Engineering (2013)* (pp. 121-128).

[34] Jun, S., Lee, S. J., & Ryu, J. B. (2015). A Divided Regression Analysis for Big Data. *International Journal of Software Engineering and Its Applications*, *9*(5), 21-32.

[35] Schifano, E. D., Wu, J., Wang, C., Yan, J., & Chen, M. H. (2016). Online updating of statistical inference in the big data setting. *Technometrics*, *58*(3), 393-403.

[36] Moselhi, O., Hegazy, T., & Fazio, P. (1991). Neural networks as tools in construction. *Journal of Construction Engineering and Management*, *117*(4), 606-625.

[37] Frank, C., Garg, A., Sztandera, L., & Raheja, A. (2003). Forecasting women's apparel sales using mathematical modeling. *International Journal of Clothing Science and Technology*, *15*(2), 107-125.

[38] Sun, Z. L., Choi, T. M., Au, K. F., & Yu, Y. (2008). Sales forecasting using extreme learning machine with applications in fashion retailing. *Decision Support Systems*, *46*(1), 411-419.

[39] Kuo, R. J., & Xue, K. C. (1999). Fuzzy neural networks with application to sales forecasting. *Fuzzy Sets and Systems*, *108*(2), 123-143.

[40] Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, *34*(2), 113-127.

[41] Jerez-Aragonés, J. M., Gómez-Ruiz, J. A., Ramos-Jiménez, G., Muñoz-Pérez, J., & Alba-Conejo, E. (2003). A combined neural network and decision trees model for prognosis of breast cancer relapse. *Artificial intelligence in medicine*, *27*(1), 45-63.

[42] He, Q., Wang, H., Zhuang, F., Shang, T., & Shi, Z. (2015). Parallel sampling from big data with uncertainty distribution. *Fuzzy Sets and Systems*, *258*, 117-133.

[43] Wang, R., He, Y. L., Chow, C. Y., Ou, F. F., & Zhang, J. (2015). Learning ELM-tree from big data based on uncertainty reduction. *Fuzzy Sets and Systems*, *258*, 79-100.

[44] Wang, G., Gunasekaran, A., Ngai, E. W., & Papadopoulos, T. (2016). Big data analytics in logistics and supply chain management: Certain investigations for research and applications. *International Journal of Production Economics*, *176*, 98-110.

[45] Yıldırım, A. A., Özdoğan, C., & Watson, D. (2014). Parallel data reduction techniques for big datasets. *Big data management, technologies, and applications*, 72-93.

[46] Li, B., Ch’ng, E., Chong, A. Y. L., & Bao, H. (2016). Predicting online e-marketplace sales performances: A big data approach. *Computers & Industrial Engineering*, *101*, 565-571.