Business Intelligence Through Big Data Analytics, Data Mining and Machine Learning

Wael M. S. Yafooz, Zainab Binti Abu Bakar, S. K. Ahammad Fahad and Ahamed. M Mithun

Abstract There is a huge amount of data creating during the fourth industry revaluation and the data are generating explosively by various fields of the Internet of Things (IoT). The organizations are producing and storing the huge amount of data into the data servers every moment. This data comes from social media, sensors, tracking, website, and online news articles. The Google, Facebook, Walmart, and Taobao are the most remarkable organizations are generating most of the data in the web servers. Data comes into three forms as structured (text/numeric), semi structured (audio, video, and image) and unstructured (XML and RSS feeds). A business makes revenue from the analysis of 20% of such data, which is a structured form while 80% of data is unstructured. Therefore, unstructured data contains valuable information that can help the organization to improve the business productive, better decision-making, extract the insights, new products and services and understand the market conditions in various fields such as shopping, finance, education, manufacturing, and healthcare. The unstructured data are needed to be analyzed and distribute in a structured manner, that is required information's are to be gathered through the data mining techniques are used to mining the data. In this paper, expose the importance of data analytics and data management for beneficial usage of business intelligence,

The original version of this chapter was revised: The contributing author's name in this Chapter was corrected. The correction to this chapter is available at https://doi.org/10.1007/978-981-13-9364-8_36

W. M. S. Yafooz (✉) · Z. B. A. Bakar · S. K. A. Fahad · Ahamed. M Mithun
Faculty of Computer and Information Technology, Al-Madinah International University, Kuala Lumpur, Malaysia
e-mail: wael.mohamed@mediu.edu.my; waelmohammed@hotmail.com

Z. B. A. Bakar
e-mail: zainab.abubakar@mediu.edu.my

S. K. A. Fahad
e-mail: fahad.wasd@gmail.com

Ahamed. M Mithun
e-mail: mithun_lonedies@yahoo.com

N. Sharma et al. (eds.), *Data Management, Analytics and Innovation*,
Advances in Intelligent Systems and Computing 1016,
https://doi.org/10.1007/978-981-13-9364-8_17

big data, data mining and machine and data management. In addition, the different techniques that can be used to discover the knowledge and useful information from such data been analyzed. This can be beneficial for numerous users concern on text mining and convert complex data into meaningful information for researchers, analyst, data scientist, and business decision makers as well.

Keywords Data mining · Business intelligence · Unstructured data · Structured information

1 Introduction

There is a huge amount of data creating during the fourth industry revaluation. Such data generates by human through social media (facebook, twitter, linked-in or Instagram) or by computer machine such as sensors, GPS, website or application systems [1]. The organizations are producing and storing the huge amount of data into the data servers every moment. This data comes from social media, sensors, tracking, website, and online news articles. The Google, Facebook, Walmart, and Taobao are the most remarkable organizations are generating most of the data in the web servers. Data comes into three forms as structured (text/numeric), semi structured (audio, video, and image) and unstructured (XML and RSS feeds). A business makes revenue from the analysis of 20% of such data which is a structured form while 80% of data is unstructured. Therefore, unstructured data contains valuable information that can help the organization to improve the business productive as well as significant for security, education, manufacturing, and healthcare as well. This can be achieved through big data analytics and data management in order to achieve the business intelligence.

The Business intelligence (BI) plays a vital role to help the decision maker to see the insights to improve productive or fast and better decision. In addition, BI can assist enhance the effectiveness of operational rules and its impression on superintendence systems, corporate-level decision-making, budgeting, financial and administration recording, making strategic choices in a dynamic business environment [2]. BI is the technologies, applications, and systems for the compilation, combination, analysis, and exhibition of the business report to help immeasurable with active business decision executing way for enforced to gain, learn and control their data to further decision-making in a plan to develop business procedures [3].

On the other hand, the big data management from diverse data formats is the main competition in business and as well as for management. The data management consists of serious management problem where current tools are not adequate to manage such massive data volumes [4]. The importance of typical big data management related to storage, pre-processing, processing and security, where new challenges in terms of storage capacity, data integration complexity, analytical tools and lack of governance. The big data management is a complex process, particularly in abundant

data originated from heterogeneous sources that are to be used for BI and decision-making. A report stated on managing big data that 75% of organizations manage some form of big data. The aim of big data management is to ensure the effectiveness of big data storage, analytics applications and security [5].

This paper exposes the importance of data analytics and data management for beneficial usage of big data, and data mining and machine learning for BI and decision-making of management. In addition, the different techniques that can be used to discover the knowledge and useful information from such data also been analyzed.

This paper organized as Sect. 2 demonstrates the big data architecture, while importance of BI highlighted in Sect. 3. Section 4 explains the big data analytics. The data mining techniques are described in Sects. 5 and 6 shows the steps of data mining. The conclusion of this paper in last section.

2 Big Data Architecture

Big data is used to describe the exponential growth of structured and unstructured data. The greatest big data challenge is that a large portion of it is not structured, often in the form of unstructured text. Therefore, there are several steps in order to handle such data. Big data architecture is the overarching system that a business uses to steer its data analytics work. The big data architecture is shown in Fig. 1. There are four logical layers that exist in big data architecture [6] as follows:

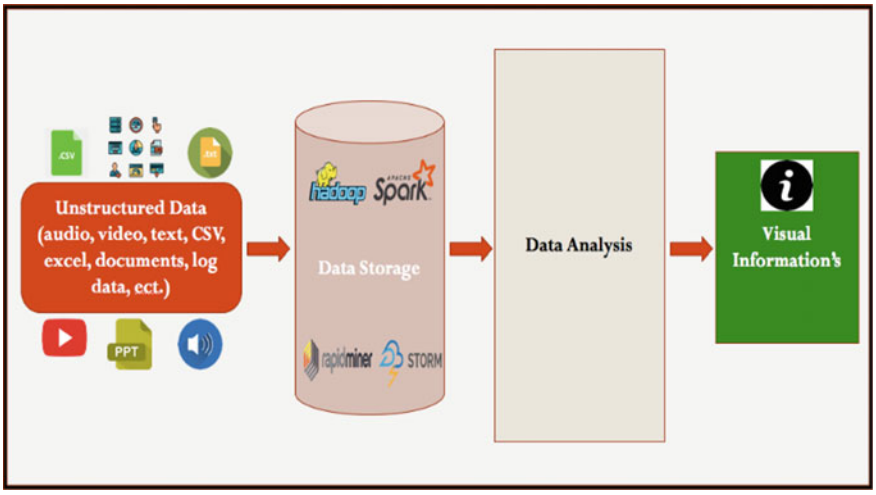

Fig. 1 Big Data architecture

Big Data Sources Layer: This is the initial layer where data comes across organization records, social media, customer records, servers, sensors, internet logs, mobile devices etc. This accepts all structured and unstructured data types.

Data Storage Layer: From gathered sources the data are lived in this layer and converts to specific formats for tools accessibility. The huge amount of data can be accessible where the structured data stored in RDBMS and the unstructured in HDFS or no SQL database.

Data Analysis/Processing Layer: This layer includes the analysis or processing to get the data to be useful which interacts for the BI. Few tools are used in this layer to analyze them into a format such as MapReduce.

Consumption/Output Layer: After the analysis or processing data are prepared to visualize its information. The output can be charts, reports and figures as well depends on the requirement.

3 Importance of Business Intelligence

To manage and develop business from earlier data, each business requires to receive remarkable judgment from anything they have done previously. If management takes critique by BI than it will support the managerial group to consider relevant declarations. BI is the technologies, applications, and systems for the compilation, combination, analysis, and exhibition of the business report to help immeasurable with active business decision executing way for enforced to gain, learn and control their data to further decision-making in a plan to develop business procedures [3]. Therefore, BI can assist enhance the effectiveness of operational rules and its impression on superintendence systems, corporate-level decision-making, budgeting, financial and administration recording, making strategic choices in a dynamic business environment [2].

Modern businesses continue to use a strategy to leverage data (especially Big Data) and achieve a sustainable contentious advantage. By transforming raw data into presentable information and understandable knowledge through the utilization of the latest information technology that can be applied at a managerial level in the decision building. Businesses produce huge investments in BI systems to accomplish goal-oriented, modern, and sustainable competing for advantage and take possibly huge advantages as a result of certain expenses [7]. BI successfully ruling Retail Industry, Insurance, Banking, Finance & securities, Telecommunications, Manufacturing industry for appropriate data mining operation on remain data on different companies. Most important features are Analysis that supports cross selling and up selling, Customer segmentation and profiling, Analysis of Parameters Importance, Survival time analysis, Analysis of consumer loyalty and consumer switching to competition, Credit scoring, Fraud detection, Fraud Detection, Web-Farming (investigation of the Internet content) [8]. The business intelligence can be archive it's objectives using the big data analytics techniques.

4 Big Data Analytics

The big data analytics which is the machine learning techniques are needed due to datasets are often distributed and their size and privacy considerations evidence distributed techniques. The data resides on platforms with varying computational and network capabilities. The benefits of big-data analytics and the diversity of application pose challenges. For example, Walmart servers handle every hour more than one million customer transactions, and these information's are inserted into databases with more than 2.5 petabytes of data, and this is the equivalent of 167 times the number of books in the Library of Congress. Herein, the Large Hadron Collider at CERN produces around 15 petabytes of data annually, this is enough to fill more than 1.7 million dual-layer DVDs per year [7]. The big data analytics are used for education, healthcare, media, insurance, manufacturing, and the government. Big data analytics has been evolved from business intelligence and decision support systems that enable healthcare organizations to analyze an immense volume, variety, and velocity of data across a wide range of healthcare networks to support evidence-based decision-making and action taking [9]. Therefore, from the discussion it's evident, the big data analytics and data management [10] is important in business intelligence for four reasons are:

First, **better decision-making**: Big data analytics can analyze past data to make predictions about the future. Thus, businesses can not only make better present decisions but also prepare for the future. Second, **Cost reduction**: Big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data. In addition, provide insights on the impact of different variables. Third, **new products and services**: With the ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want. Big data analytics, more companies are creating new products to meet customers' needs. Fourth, **understand the market conditions**: By analyzing big data, you can get a better understanding of current market conditions.

In order to retrieve the significant information's, there are few challenges and features to be considered in big data analytics tools and techniques, and they are including, scalability, and fault tolerance as well [11–13]. Table 1 represented few widely used tools with the features they provide for big data analytics.

5 Data Mining Techniques

Picking the suitable data mining technique is one of the most significant responsibilities in the Data Mining Process. Nature of business and the kind of object or difficulty suffered in business provides relevant direction to determine the fittest technique [14]. Applying data mining techniques there is some generalized approach, it can be referring to enhance the efficiency and cost-effectiveness. Several core techniques that are performing in the data mining process, specify the character of the

Table 1 Most common techniques of Big Data

Features/tools	Scalability	Fault tolerance	Visualization	Policy	Citations
Hadoop	Yes	Yes	Graphical format	Processing of big data	[29, 22]
Apache Spark	No	Yes	Through charts	Filters large scale data	[30]
Cassandra	Yes	Yes	Used tableau	NoSQL techniques	[31]
Apache SAMOA	Yes	No	Snapshots views	Mines repositories and presents	[32]
RapidMiner	Yes	Yes	Various output formats	Supports all mining process	[24]
NodeXL	Yes	No	Graph	Maps networks using excel	[33]
Apache Storm	Yes	Yes	Graph, Maps, Charts	Real-time computation	[34]
Hive	Yes	Yes	Graphical	Static data analyzes	[23]

mining operation and reclamation option of data. The mining technique is highly productive on the result [15]. There are lot of techniques but among them, Association Rule, Classification, Clustering, Decisions Tree, and the Neural Networks are profoundly practical and successful.

Association Rule: Association (relation) is the usual simple and straightforward data mining procedure. It is too powerful and well-researched systems of data mining. By extraction of interesting similarities, connections, common formations within collections of items in Database. Association is also particular to dependently associated variables. Because or running by determining the correlation among items in the related transaction, the association is more comprehended as relationship procedure [16]. By using association technique, retailers can perform analysis on customer's habits. Retailers might obtain an explanation from history of sales data, consumers who order drinks while they purchase fast-food. They can put them beside each other to conserve time for the client and increase sales. Association rule depends on two significant information. Those are; Help and Confidence. "How frequently is the rule implemented?" Is represent support and Confidence is "How frequently the rule is true?" [17].

$$\text{Support}(\text{FASTFOOD} \rightrightarrows \text{DRINKS}) = \frac{\text{Number of time order DRINKS when purchase FASTFOOD}}{\text{Total number of Transection.}}$$

If Support (FASTFOOD $\rightrightarrows$ DRINKS) is 5%, it indicates that Consumer demands fast-food and drinks together 5% of total purchase in the history.

$$\text{Confidence}(\text{FASTFOOD} \rightrightarrows \text{DRINKS}) \frac{\text{Support}(\text{FASTFOOD} \cup \text{DRINKS})}{\text{Support}(\text{FASTFOOD})}$$

If in association rule Confidence (FASTFOOD $\rightrightarrows$ DRINKS) = 75%. That means, 75% customer order drinks when they purchase fast-food.

Classification: Classification is the most deliberate and generally practiced supervised learning data mining task. Given an object, allowing it to one of the predefined target sections and classify individually in a set of data inside a predefined set of categories or collections described as classification. Classification is a complicated data mining procedure that overcomes to assemble multiple properties mutually into discernable divisions, which can apply to carry additional outcomes to accurately predict the target class for individual state. Classification algorithms attempt to define relations among properties in a training set to incorporate new observations and there have two principal rules in this method [14]. Learning—Data are examined by the analysis algorithm and model is created from the practice examples. Classification—In this rule, the data is applied to estimate the accuracy of the classification precepts and allow a label to an unlabeled analysis situation [18].

Naive Bayes classifier, Random Forest and AdaBoost are successful and rapidly utilized in classification data mining technique. Naive Bayes classifier is a great example of a classifier that estimates unknown conditional probabilities to recognize classes. It delivers successful results in medical diagnosis, banking, document categorization, marketing, and pattern recognition. Random Forest is an excellent training model including the goal of decreasing variance and building efficiency of performance. By ensemble learning approach with several training sets to classify input parameters for every tree in the forest. AdaBoost is the most suitable binary classification solution that connects a number of soft learners to perform better detachment between classes [19].

Clustering: Clustering is the concept to unite objects in clusters according to their similarity. Clustering is very similar to classification, but clustering is an unsupervised learning technique that grouping chunks of data together in meaningful clusters based on their similarities [20, 21]. As like classification, clustering groups those are named cluster and those are not described previously. It was majored by the specialty of data points and the relationships between the individual data points based on their properties. Clustering is a blinded unsupervised learning rule of the data mining process that can determine the correlation within data points based on the qualities it can understand. Sometimes, clustering called segmentation and it helps to understand, the changes happened in the database. Clustering algorithms are divide into meaningful groups.

There are several kinds of clustering methods including thousands of algorithm for different object. Most significant are; Partitioning, Hierarchical, Density-Based, Grid-Based, and Model-Based Methods. Clustering Algorithm is divided base on different types of clustering method [22]. (a) Partitioning Based: K-means, K-modes, K-medoids, PAM, CLARANS, CLARA, and FCM. (b) Hierarchical Based:BIRCh, CURE, ROCK, Chameleon, and Echidna. (c) Density-Based: DBSCAN, OPTICS,

DBLASD, and DENCLUE. (d) Grid-Based: Wave-cluster, STING, CLIQUE, and OptiGrid (e) Model-Based: EM, COBWEB, CLASSIT, andSOMs.

Decisions Tree: Decision tree technique model is simple to learn for users. The decision tree can be utilized both as a component of the adoption patterns to establish the suitability and preference of particular data begin with a simple question that has two and more replies [23]. Each solution guides to an additional question to improve recognizing the data. This prediction can be performed based on any response determine the data, that can obtain the terminal determination. Several predictions might be based on the historical practice that supports the structure of the decision tree frequently practiced with classification systems to associate standard information, and including predictive methods [14].

Decision trees produce a hierarchical partitioning of the data. Those several partitions at the leaf level to the several classes that produced by the application of a split basis. The separation principle stated on an individual attribute, or on multiple attributes. Once it connected as a univariate split and applied as multivariate split. The approach is to recursively break training data to maximize the difference in various classes over several nodes. The perception is many classes are maximized on various classes when the delivered node is maximized. The tree-shaped formation that describes collections of arrangements, tree nodes describe property value, the branch describes the result of the test and subsequently, tree leaves represent class relationships. The group instance begins at the source node and, depending on the results, regarding the proper parts till the leaf [23].

Neural Networks: Neural Network is an extensive technique applied in the starting stages of the data mining technology. Neural networks are automated to a remarkable extent and because the user is not required to have much knowledge regarding the database. Node and the Link are the two principal elements of the Neural Network technique [24]. Node, which coordinates to the neuron in the human brain and the Link, suits the connections among the neurons in the human brain. To execute neural network efficiently, three factors need to reflect. Wherewith the nodes are correlated? When should the training rule be suspended? And Number of processing units to be applied? [14].

The formation of neurons and their interconnections have described the architecture of the network and those interconnected are in the single or multiple layers. Every neural network model has distinctive architectures and those architectures use separate learning methods with their individual benefits and limitations. Neural networks is a forbidding modeling technique and some complicated models are impossible to understand completely. Therefore, to know the Neural network technique, there have two explications is recommended. Neural network must pack up and let to be practiced for a single application and bonded with skillful advising co-operation [25].

Forward and Backpropagation, Neural Networks hold by these two states. Continuously ultimate output activation function is frequently applied to produce the inputs to meet the class description in the forwarding condition. Final output at the output layer produces an error value and backpropagation states stat operation to updating of the weights in the prior layers are determined as a function of the errors and weights in the course before of it.

6 Steps of Data Mining

Data Mining is regarding interpreting the immense volume of data and extracting of knowledge from its several objects. Fundamentally, the chance of losing the rich and influential message carried by the massive databases was standing and this demands the adoption of sufficient systems to gain beneficial data so that the scope of data mining had been developed in 1980s and is still advancing. The individual approach leads and produces the data mining assignments and its utilization [26]. For some businesses, the purposes of data mining recognize developing marketing abilities, identifying different trims, and predicting the prospect based on earlier observations and modern inclinations. As databases become extensive, it turns more challenging to maintain enterprise preparation. There is an audible demand to examine the data to sustain devising and additional purposes of an entrepreneur. Data mining could further continue practiced to recognize unusual performance. An intelligence agency could define a strange behavior of its representatives practicing some aforementioned technology [27].

There are extensive amounts of current and historical data existing and stored. Different standard models for data mining are proposed and some are established. All those models are described in subsequent steps. These steps support performing the data mining responsibilities. Three models are mostly followed by the data mining experts and researchers for data mining process and these types are; Knowledge Discovery Databases (KDD) process model, CRISP-DM and SEMMA. The Knowledge Discovery Databases (KDD) model to gaining knowledge in data and emphasizes the important level of particular data with nine steps. Cross Industry Standard Process for Data Mining (CRISP-DM) launched by Daimler with six steps or phases and improves over the years [28]. With five distinct phases identified as Sample, Explore, Modify, Model, Assess (SEMMA), this model was developed by SAS Institute Inc. Data Mining sometimes named to knowledge discovery database (KDD) due to, it is the method of examining data from different sources and comprehensions, and condensing it into knowledge that can be presentable, that knowledge can minimize loss and improve return or both [17]. See Table 2.

7 Conclusion

Business Intelligence is the technologies, applications, and systems for the compilation, combination, analysis, and exhibition of the business report. BI help immeasurable with active business decision executing way for enforced to gain, learn and control their data to further decision-making in a plan to develop business procedures. A business makes revenue from the analysis of 20% of such data which is a structured form while 80% of data is unstructured. Therefore, unstructured data contains valuable information that can help the organization to improve the business

Table 2 Steps and key features of KDD

Steps	Key features		
	KDD	CRISP-DM	SEMMA
1. Learning and understanding of the application domain	• Defined objects based on the client's point • State and assuming the purpose and principles	• Uncovers factors like success patterns, enterprise, and data mining aspirations • Learn the fundamentals of business terminologies and technical phases	• Sampling data • Portion took from a huge dataset to obtain meaningful knowledge • Small enough to handle instantly
2. Creating a target dataset	• Create a target dataset including the subset of data units which process will do performed	• Data gathering, monitoring quality and examining of information form hypotheses for unexplained information	
3. Data cleaning and pre-processing	• Targeted data cleansing and pre-processing for data externally any noise and inequalities • DBMS points are selected such as data schema, type of data, and mapping of dropping and unfamiliar values in the database		• Exploration of data • Expanding the understanding and conceptions • Improving the development rule by combing for trends and irregularities
4. Data reduction and projection or data transformation	• Determining valuable characteristics and images to describe the data • Prepossessed and transformed into a conventional format • Transformation of data from one form to another so that data mining algorithms can be performed efficiently	• Collection and development of the ultimate dataset • Records, table and attributes assortment • Cleaning and transformation of data	• Modification of data by creating, selecting and transformation of variables • Focus on the model selection process • Seems for outliers and decreasing the number of variables

(continued)

Table 2 (continued)

Steps	Key features		
	KDD	CRISP-DM	SEMMA
5. Choosing the function suitable data mining task	• Appropriate data mining task is decided based on distinct intentions • Determining the scope of the model assumed by the data mining algorithm	• Determination and utilization of different modeling procedures • Various models are constructed for the same data mining predicament by valuing separate parameters	• Automatically explores for a sequence of data • Various modeling methods are present and several types of the model have its individual concentration • Relevant for the particular condition on the data for data mining
6. Choosing the suitable data mining algorithm(s)	• Appropriate data mining algorithms are picked for exploring various patterns from data • Fitting algorithms are decided based on balancing the overall standards for data mining		
7. Employing data mining algorithm	• Decided algorithms are performed on preprocessed and transformed data • Exploring patterns of interest in a set of records or a particular representable structure		

(continued)

Table 2 (continued)

Steps	Key features		
	KDD	CRISP-DM	SEMMA
8. Interpreting mined patterns	• Interpretation and evaluation of mining patterns • Associate in selected patterns visualization • Eliminating redundant or irrelevant patterns, and transposing the useful ones into terms comprehensible by users	• Evaluation of recovered models by determining the application of results • Explanation of each models depends on the algorithm going to implement	• Evaluation of the reliability and application of findings • Evaluates the performance
9. Using discovered knowledge	• Discovered knowledge is applied to different goals • Discovered knowledge can apply involved individuals or can be integrated with another system for additional progress • It to affected individuals, as well as monitoring for and determining possible conflicts with previously obtained knowledge	• Determining the technique of gathering knowledge and decisions • Organizing, reporting and presenting the obtained knowledge when required	

productive as well as significant for security, education, manufacturing, and healthcare as well. So, the data management, data mining and machine learning techniques are required in order to extract the insights from huge amount of data. By using such techniques, business intelligence gets better decision-making, Cost reduction, new products and services and understand the market conditions.

References

1. Yafooz, W. M. S., Abidin, S. Z., & Omar, N. (2011, November). Challenges and issues on online news management. In *2011 IEEE International Conference on Control System, Computing and Engineering (ICCSCE)* (pp. 482–487). IEEE.
2. Richards, G., Yeoh, W., Chong, A. Y. L., & Popovič, A. (2017). Business intelligence effectiveness and corporate performance management. An empirical analysis. *Journal of Computer Information Systems*, 1–9.
3. Balachandran, B. M., & Prasad, S. (2017). Challenges and benefits of deploying big data analytics in the cloud for business intelligence. *Procedia Computer Science, 112,* 1112–1122.
4. Yafooz, W. M. S., Abidin, S. Z., Omar, N. & Hilles, S. (2016, September). Interactive big data visualization model based on hot issues (online news articles). In *International Conference on Soft Computing in Data Science* (pp. 89–99). Singapore: Springer.
5. Siddiqa, A., Hashem, I. A. T., Yaqoob, I., Marjani, M., Shamshirband, S., Gani, A., et al. (2016). A survey of big data management: Taxonomy and state-of-the-art. *Journal of Network and Computer Applications, 71,* 151–166.
6. Borodo, S. M., Shamsuddin, S. M., & Hasan, S. (2016). Big data platforms and techniques. *Indonesian Journal of Electrical Engineering and Computer Science, 1*(1), 191–200.
7. Sparks, B. H., & McCann, J. T. (2015). Factors influencing business intelligence system use in decision making and organisational performance. *International Journal of Sustainable Strategic Management, 5*(1), 31–54.
8. Qureshi, N. A., Khan, B. A., & Saif, J. A. (2017). Business intelligence systems in the holistic infrastructure development supporting decision-making in organisations.
9. Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change, 126,* 3–13.
10. Yafooz, W. M. S., Abidin, S. Z., Omar, N., & Idrus, Z. (2013, December). Managing unstructured data in relational databases. In *2013 IEEE Conference on Systems, Process & Control (ICSPC)* (pp. 198–203). IEEE.
11. Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013, January). Big data: Issues and challenges moving forward. In 2013 *46th Hawaii International Conference on System Sciences* (HICSS) (pp. 995–1004). IEEE.
12. Zhou, Z. H., Chawla, N. V., Jin, Y., & Williams, G. J. (2014). Big data opportunities and challenges: Discussions from data analytics perspectives [discussion forum]. *IEEE Computational Intelligence Magazine, 9*(4), 62–74.
13. Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research, 70,* 263–286.
14. Fahad, S. A., & Alam, M. M. (2016). A modified K-means algorithm for big data clustering. *International Journal of Computer Science Engineering and Technology, 6*(4), 129–132.
15. Fahad, S. A., & Yafooz, W. M. (2017). Design and develop semantic textual document clustering model. *Journal of Computer Science and information technology, 5*(2), 26–39. https://doi.org/10.15640/jcsit.v5n2a4.
16. Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A. V., & Rong, X. (2015). Data mining for the internet of things: Literature review and challenges. *International Journal of Distributed Sensor Networks, 11*(8), 431047.

17. Birant, D., & Yıldırım, P. (2016). A Framework for data mining and knowledge discovery in cloud computing. In Z. Mahmood (Ed.), *Data science and big data computing: frameworks and methodologies* (pp. 245–267). Cham: Springer. https://doi.org/10.1007/978-3-319-31861-5_11.
18. Aggarwal, C. C. (2015). *Data classification: Algorithms and applications*. Boca Raton: CRC Press, Taylor & Francis Group.
19. Yafooz, W. M. S., Abidin, S. Z., & Omar, N. (2011, November). Towards automatic column-based data object clustering for multilingual databases. In *2011 IEEE International Conference on Control System, Computing and Engineering (ICCSCE)* (pp. 415–420). IEEE
20. Yafooz, W. M. S., Abidin, S. Z., Omar, N., & Halim, R. A. (2014). Model for automatic textual data clustering in relational databases schema. In *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)* (pp. 31–40). Singapore: Springer.
21. Ali, S. M., & Tuteja, M. R. (2014). Data mining techniques. *International Journal of Computer Science and Mobile Computing, 3*(4), 879–883.
22. Kotwal, A., Fulari, P., Jadhav, D., & Kad, R. (2016). Improvement in sentiment analysis of twitter data using hadoop. *Imperial Journal of Interdisciplinary Research*, *2*(7).
23. Bhawnani, D., Sanwlani, A., Ahuja, H., & Bohra, D. (2015). Big Data analytics on cab company's customer dataset using Hive and Tableau. *International Journal of Scientific Engineering and Applied Science (IJSEAS)*, *1*(2), ISSN, 2395-3470.
24. Rangra, K., & Bansal, K. L. (2014). Comparative study of data mining tools. *International Journal of Advanced Research in Computer Science and Software Engineering*, *4*(6).
25. Alton, L. (2017, December 22). The 7 most important data mining techniques. Retrieved October 2, 2018, from https://www.datasciencecentral.com/profiles/blogs/the-7-most-important-data-mining-techniques.
26. Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research, 12*(1), 217–222.
27. Thuraisingham, B. (2014). *Data mining technologies, techniques, tools, and trends*. CRC press.
28. Alhendawi, K. M., & Baharudin, A. S. (2014). A classification model for predicting web users satisfaction with information systems success using data mining techniques. *Journal of Software Engineering*.
29. Park, D., Wang, J., & Kee, Y. S. (2016). In-storage computing for Hadoop MapReduce framework: Challenges and possibilities. *IEEE Transactions on Computers*.
30. Pirozzi, D., Scarano, V., Begg, S., De Sercey, G., Fish, A., & Harvey, A. (2016, July). Filter large-scale engine data using apache spark. In *2016 IEEE 14th International Conference on Industrial Informatics (INDIN)* (pp. 1300–1305). IEEE.
31. Santos, M. Y., e Sá, J. O., Andrade, C., Lima, F. V., Costa, E., Costa, C., … & Galvão, J. (2017). A Big Data system supporting Bosch Braga Industry 4.0 strategy. *International Journal of Information Management*, *37*(6), 750–760.
32. Minelli, R., & Lanza, M. (2013, September). SAMOA—A visual software analytics platform for mobile applications. In 2013 29th IEEE *International Conference on Software Maintenance (ICSM)* (pp. 476–479). IEEE.
33. Yep, J., & Shulman, J. (2014). Analyzing the library's Twitter network: using NodeXL to visualize impact. *College & Research Libraries News, 75*(4), 177–186.
34. Raina, I., Gujar, S., Shah, P., Desai, A., & Bodkhe, B. (2014). Twitter sentiment analysis using apache storm. *International Journal of Recent Technology and Engineering, 3*(5), 23–26.