Big Data Analytics Part 2 (Unit 4)

MBI304- Data Mining & Data Analytics

Mamta Sagar
Department of Bioinformatics
University Institute of Engineering & Technology, CSJM University, Kanpur

Big data analytics (BDA): tools and methods

Big data storage and management The most difficult problem that needs to be solved to handle big data effectively is storage; it is not necessarily easy to deal with large quantities and varieties of data (Elgendy and Elragal, 2014; Zhong, et al., 2016; Lv, Z. et al., 2017). There are many big data storage and analysis models. Where the large amount of data is caused by the sheer variety of users and devices, a data centre may be necessary for storing and processing the data. Establishing network infrastructure is necessary to help gather this rapidly generated data, which is then sent to the data centre before being accessed by users (Lv et al., 2017). Research by Yi et al. (2014) identifies the components of the network that must be established, such as an original data network, the bridges used for connecting and transmitting to data centres, and at least one data centre. Another study (H. Eszter, 2015) highlighted the issues in using big data through specific locations and showed that the users could not select data through the data network. For storage models, the most important challenge is how to deal with the sheer amount of data, as ultra-scalable solutions can block the processing of certain data sources, causing inefficiency. Building more scalable big data technology is a challenge, and any new technology must offer data gathering and distribution among nodes spread through the world (Lv et al., 2017). Structured data storage and retrieval methods include "relational databases, data marts, and data warehouses" (Elgendy, N. and Elragal, A., 2014). Data is extracted from outside sources, then transformed to fit operational needs, and finally loaded into the database. The data is then uploaded from the operational data store to longer-term storage using Extract, Transform, Load (ETL) or Extract, Load, Transform (ELT) tools. The data is then cleaned, transformed, and catalogued before use (Bakshi, 2012; Elgendy and Elragal, 2014). A big data environment requires analysis skills, unlike the Enterprise Data Warehouse (EDW) traditional environment (Hartmann, T. et al., 2019).

➢ The big data environment accepts and demands all possible data sources. On the other hand, EDW approaches data sources with caution, as it is more streamlined towards supporting structured data (Elgendy and Elragal, 2014; Hartmann et al., 2019)

. ➢ Due to increasing number of data sources and data analyses possible, big data storage requires agile databases to give analysts the opportunity to produce and adapt to data easily and quickly (Elgendy and Elragal, 2014; Hartmann et al., 2019).

➢ A big data repository must be deep, allowing analysts to analyse the datasets deeply by using complex statistical methods (Elgendy and Elragal, 2014; Hartmann, T. et al., 2019).

Hadoop is a popular big data analytics framework. Hadoop "provides reliability, scalability, and manageability by providing an implementation for the MapReduce paradigm as well as gluing the storage and analytics together" (Elgendy, N. and Elragal, A., 2014). Hadoop includes HDFS which

is for the big data storage and MapReduce for big data analytics, and it can process extremely large amount of data by dividing the data into smaller blocks, then specifying datasets to be distributed across cluster nodes (Raghupathi and Raghupathi, 2014; Elgendy and Elragal, 2014). Hadoop incorporates several technologies: "Hive is a data warehouse implementation for Hadoop, MapReduce is a programming model in Hadoop, and Pig is a querying language for Hadoop which has similarities to the SQL language for relational databases" (Zuech et al., 2015). First-generation technology generated the Apache Spark project in software terms (Watson, 2019), but Hadoop has a great deal more power, which offers advantages to analytics in terms of memory. It can work with both batch and real-time workloads, is easy to program with Java code, and can connect to Apache projects and other software within a closed ecosystem. Hadoop's components are shown in Figure 11 (Watson, 2019): 1. Spark SQL runs SQL-like queries on structured data. 2. Spark streaming provides real-time data processing. 3. MLib provides a machine learning library of algorithms and utilities. 4. Graph X provides application algorithms.

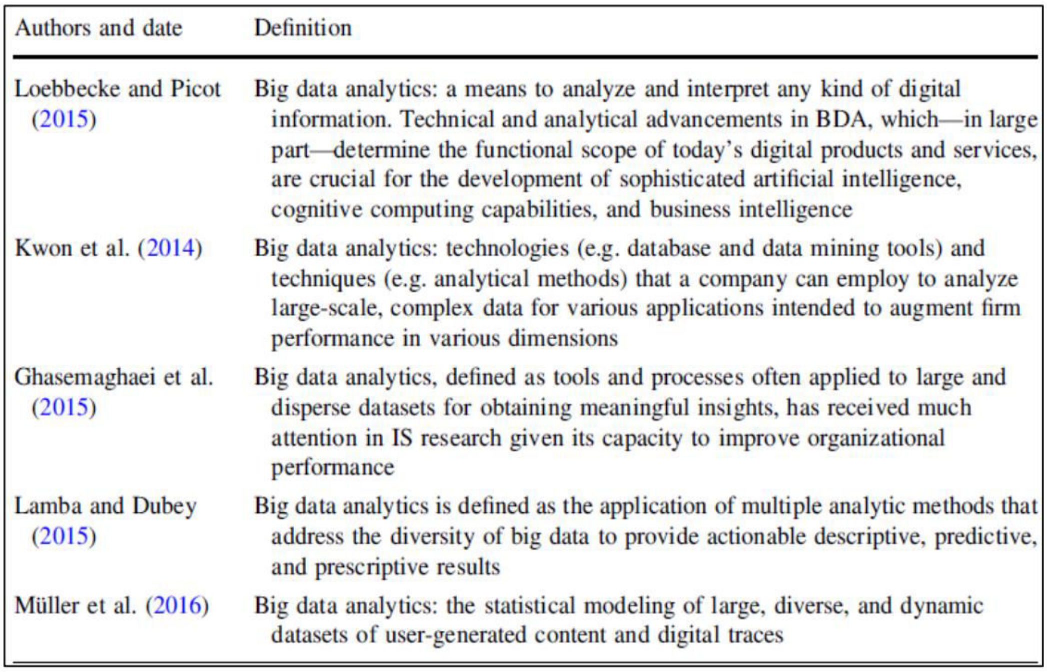
7.2. Big data analytics processing

Analytics processing is the next issue after big data storage. According to He et al. (2011), big data analytics processing has four critical requirements: a) Fast data loading: limited interference between disk and network, to speed up query execution. b) Fast query processing: workloads are heavy, therefore real-time requests should be processed as quickly as possible to satisfy user requirements. The data placement structure should also have the ability process multiple queries as query volumes increase.

c) Highly efficient utilization of storage space: as user activities grow rapidly, they need scalable storage capacity and computing power. As disk space is limited, it is necessary to manage data storage during processing and address the space issues adaptively. d) Strong adaptivity to highly dynamic workload patterns: the underlying system should be highly adaptive, as data processes have different workload patterns and the analysing of big datasets has many different applications and users, with different purposes and methods (Elgendy, N. and Elragal, A., 2014). The work presented by García et al. (2016) shows that using big data frameworks for storing, processing, and analysing data has changed the context of knowledge discovery from data, mainly in terms of data mining processes and pre-processing, with a particular focus on the rise of data pre-processing in cloud computing. The presented solution covered various data pre-processing technique families with factors such as maximum size supported examined in terms of big data and data pre-processing throughout all of the families of methods. Moreover, various big data framework such as Hadoop, Spark, and Flink were discussed.

Big data analytics Big data growth continues apace, and many organisations are now interested in managing and analysing data. Organisations trying to benefit from big data are adopting big data analytics to facilitate faster and better decisions, as it is not easy to analyse datasets with analysis techniques and infrastructure based on traditional data management (Constantiou et al., 2015). The need for new tools and methods specialised for big data analytics is thus also growing. The emergence of big data is affecting everything from data itself to its collection and processing, and, finally, the extracted decisions. Providing big data tools and technologies can help in managing the growth of network-produced data, which is otherwise exponential, as well as in increasing the capability of organisations to scale and capture the required data to reduce database performance problems (Elgendy, N. and Elragal, A., 2014). Further big data analytics definitions are clarified in Table 4. Opening any popular scientific or business publication today, whether online or in the physical world, generally involves running into a reference to data science, analytics, big data, or some combination of these terms (Agarwal and Dhar, 2014). Some researchers are focusing on big data definitions (Akter et al., 2016; Mikalef et al., 2018), while others analyse the tools, techniques, and procedures required for analysis (Russom, 2011), and others seek to explain big data analytics' impact on business value (Mikalef et al., 2018)

Table 4: Sample definitions of big data analytics, adopted from (Mikalef et al., 2018)



Authors and date	Definition
Loebbecke and Picot (2015)	Big data analytics: a means to analyze and interpret any kind of digital information. Technical and analytical advancements in BDA, which—in large part—determine the functional scope of today's digital products and services, are crucial for the development of sophisticated artificial intelligence, cognitive computing capabilities, and business intelligence
Kwon et al. (2014)	Big data analytics: technologies (e.g. database and data mining tools) and techniques (e.g. analytical methods) that a company can employ to analyze large-scale, complex data for various applications intended to augment firm performance in various dimensions
Ghasemaghaei et al. (2015)	Big data analytics, defined as tools and processes often applied to large and disperse datasets for obtaining meaningful insights, has received much attention in IS research given its capacity to improve organizational performance
Lamba and Dubey (2015)	Big data analytics is defined as the application of multiple analytic methods that address the diversity of big data to provide actionable descriptive, predictive, and prescriptive results
Müller et al. (2016)	Big data analytics: the statistical modeling of large, diverse, and dynamic datasets of user-generated content and digital traces

People now aim to both to collect data and understand its importance and meaning for use in making decisions. The data to be analysed is large in volume and consists of various types. "Massive, high dimensional, heterogeneous, complex, unstructured, incomplete, noisy, and erroneous" (Ma et al., 2014), are features of big data that require changes in statistical and data analysis approaches. It is also important to understand the content of big data. The process of applying algorithms to analyse the content of big data is part of data analytics, which is used for 1) analysing sets of data information and their relationships, 2) extracting previously unknown valid patterns, and 3) for detecting important relationships between stored variables. In this section, various big data analyses will be discussed, beginning with the data analysis techniques available and some of the common big data analytics suites, finally discussing several big data platforms and tools. Data analysis techniques can be characterised into four types, as shown in Figure 12

7.1.1. Supervised techniques

A supervised technique refers to where data are trained and tested, and the training data is labelled. Labelled means that the full history of what has happened to the data is known, and thus the history for the data variables is known. Supervised learning involves training a system based on labelled data and this requires a supervisor with the ability to expect the output from each input that can train the system according to its expectations. When the system is trained, it can give predictions within "many applications of classification and fault detection and channel coding and decoding" (Kotsiantis, et al., 2007; Cui, et al., 2019). This technique is used for approximating a function between the input and output. The idea is for the system to learn the training dataset's classifiers (the labelled documents) then to automatically apply this classification to an unknown dataset's un-labelled documents. This learning technology thus involves learning from example (Boyd-Graber et al., 2014; Müller et al., 2016; Breed and Verster, 2019). Regression is an example of the supervised learning algorithm, as are Linear Regression, Decision Trees (DT), Support Vector Machine (SVM), K Nearest Neighbour (K-NN), Naive Bayes Classifier (NBC), Random Forest, and neural networks (NN). However, many of these supervised techniques cannot be used with wireless networks, and as the learning techniques are dependent on the data training, the results are also restricted (Cui et al., 2019)

Regression Analysis: is mathematical tool used to discover correlations between several variables based on experimental or observed data. Where analysis defines the relationships between variables as non-random, such analysis may make the correlations between variables appear simpler and more regular (Lei et al., 2016), as shown in Figure 13.

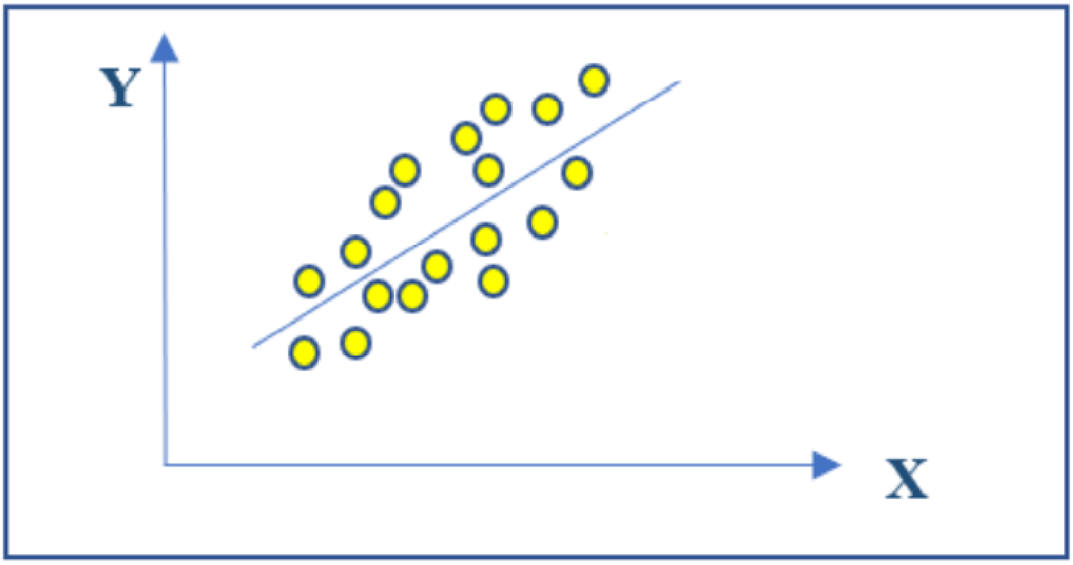


Figure 13: Regression analysis

Structured data mostly utilises predictive analytics, and this overshadows other analytics forms for 95% of big data (Gandomi and Haider, 2015). However, new statistical techniques for big data have emerged which clarify the differentiation of big data from smaller data sets. In practice, however, most statistical methods were designed for smaller datasets, in particular, samples. Usually, scientists make predictions based on theories in the prediction domain. However, big data analytics can deliver predictions that depend on the sequence of data processing and execution. According to Kitchin (2014) and Müller et al. (2016), • big data brings new challenge as it is generated from different system sources. The data retrieved from each source system should thus be sent to a central repository; • the relationship between operations should be defined to allow reconstruction of datasets from multiple sources; • the knowledge discovery process should be automated from data or datasets to make predictions; • generating new theories is required to create and improve models. Predicted target theory generates a set of predictors; however, some theories explain the relationships between independent and dependent predictors more effectively; • there is a shift from theory-driven to process-driven prediction based on analysing the BDA steps and identifying the challenges, theoretically informing future BDA needs throughout data acquisition, pre-processing analysis, and interpretation.

7.1.2. Un-supervised techniques Here, the training data is unlabelled. Unlabelled means that the history of the data is missing, there is no history available for data variables, and the data have not been trained and tested. Thus, unsupervised techniques require separate training data (Boyd-Graber et al., 2014; Müller et al., 2016; Breed and Verster, 2019). Unsupervised learning requires deducing functions for presenting unknown structures from unlabelled data. This technique does not require a supervisor, which means that the system must have the ability to proceed independently with training based on unlabelled data input (Cui et al., 2019). Examples of unsupervised learning algorithms include clustering algorithms, combinatorial algorithms, A priori algorithms, Self-Organizing Maps (SOM), and applications of game theory. These techniques are used for classifying the input data into different clusters or classes based on the data distribution (Jiang et al., 2017; Cui et al., 2019). Cluster Analysis: This method is based on grouping objects and classifying them depending on shared features. It is used for differentiation between objects to allow division into clusters. Thus, data which are related to each other or have the same features will be placed in a cluster or a group and unrelated data will be in other groups (Wu et al., 2018; Cui et al., 2019), as shown in Figure 14. F

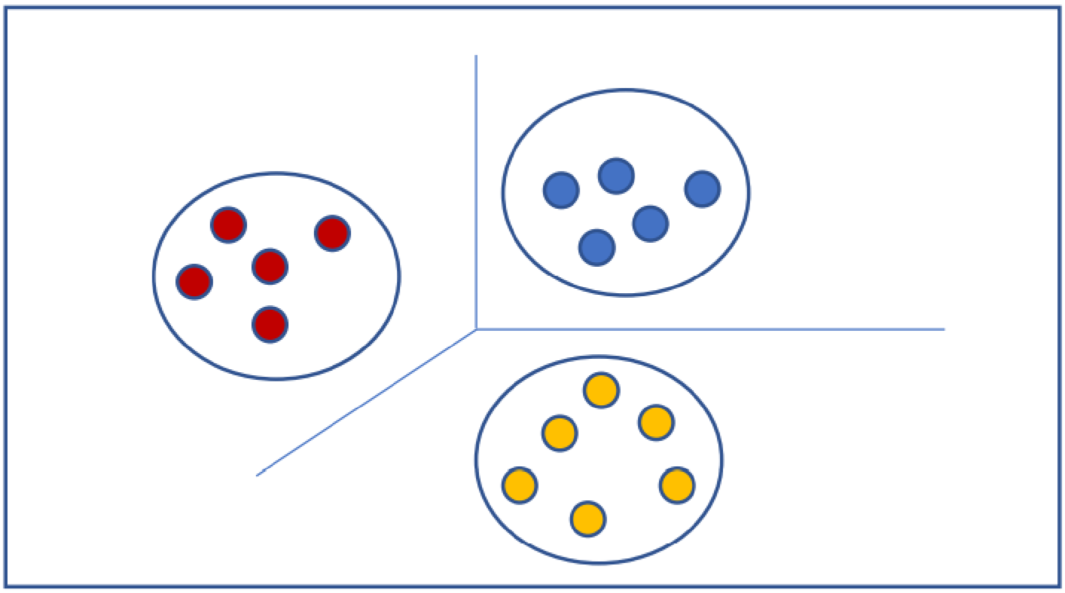


Figure 14: Cluster analysis.

7.1.3. Semi-supervised techniques Where some of the data is labelled and some is unlabelled, supervised and unsupervised techniques can also be mixed. Algorithms are applied for both labelled and unlabelled data, and even with incomplete information or missing training sets, some of the dataset's classifiers can be learned.

Both supervised and unsupervised techniques focus on one aspect (target separation or independent variable distribution, respectively), and using them together may thus give better results (Breed, D.G. and Verster, T., 2019).

7.1.4. Reinforcement learning (RL)

Reinforcement learning involves setting and classifying real-time data changesin a way that allows the learning framework to adapt based on those changes (Wu et al., 2018; Cui et al., 2019). The components of an RL algorithm are the agent; the environment; and the actions. The actions are taken by the algorithm based on the environment, and depending on the feedback from the environment, it determines whether the action is positive, thus using it again in future, or negative, thus discarding it. An example of reinforcement learning is Markov Chains (Markov Decision Process) (Müller et al., 2016). The difference between RL and supervised or unsupervised learning is that RL works based on the feedback which is either good or not depending on the situation and is hence dynamic, while supervised and unsupervised learning give static solutions (Cui, et al., 2019). The RL process includes an actor which acts in the environment with its own copy of the data; the data can thus be stored in a separate replay memory and sampled by the learner to be computed within the policy parameters. The actor learners then receive the updated policy parameters (Mnih et al., 2015; Mnih et al., 2016). The Map-Reduce framework was utilized by Li and Schuurmans (2011) for parallelising batch reinforcement learning methods with linear function approximation (Mnih et al., 2016). Applying parallelism helped speed up large matrix operations but did not assist the collection of experience or stabilise learning. The reinforcement learning goal is to develop policies that help in decision making. An example, is Q-learning, where the algorithm has no knowledge of the data but has the ability to find out about the data in an automated way (Wu et al., 2018; Cui et al., 2019). Q-learning is one of the most popular reinforcement learning algorithms, though it learns unrealistically high action values as it includes" a maximization step of overestimated action values, which tends to prefer overestimated to underestimated values" (Hester et al., 2018). The Q-learning algorithm is thus best used for overestimating action values in specific conditions. Recently, Q-learning has been combined with deep neural networks to produce Double Q-learning (DQN); that combination also suffers from overestimations (Mnih et al.,2015). Deep neural networks are artificial neural networks with multiple layers between the input and output layer, which help RL algorithms to provide effective performance. However, it was previously thought was that combining simple online RL algorithms with deep neural networks was unstable (Mnih et al., 2013; Mnih et al., 2015; Schulman et al., 2015; Mnih et al., 2016; Van Hasselt et al., 2016). The common idea arising from early studies was that the data sequences observed by online RL agents were not stable, and had no strong correlations to RL updates. However, data can be batched if the agent's data are stored in an experience replay memory (Schulman et al., 2015) or sampled from different time steps randomly (Mnih et al., 2013; Mnih et al. , 2016; Van Hasselt et al., 2016), and the Double Q-learning algorithm can work with large-scale function approximation (Hasselt,H.V., 2010). Thus, a new algorithm known as Double DQN (a combination of Double Q-learning with neural networks) has been constructed which offers higher scores on several games; however, this algorithm has not displayed more accurate value estimation (Hester et al., 2018).

7.4. Analytics techniques Correlation Analysis:

this is an analytical method used to determine the relationships such as "correlation, correlative dependence, and mutual restriction, among observed phenomena and accordingly conducting forecast and control" (Chen, M., Mao, S. and Liu, Y., 2014), as shown in Figure 15. Positive correlation, on the left means while one variable increases so does the other. No linear correlation on the middle means there is no visible relationship between the variables. Negative correlation on the right means as one variable increases, the other decreases (Chen, M., Mao, S. and Liu, Y., 2014).

Text Mining: This converts the content from unstructured text to structured text in order to help uncover the meaning and the information contained. Factor Analysis: This groups several related variables into a single factor, which means that fewer factors are used in analysis, which is thus simpler. The research presented in Schelén et al. (2015) examines the state-of-the-art in big data at that time and discusses research agendas. In addition, it defines the basic technology and toolsets used. It is not easy to analyse datasets with traditional data management techniques (Constantiou and Kallinikos, 2015); therefore, new methods and tools have been developed for big data analytics, as well as for storing and managing such data. These solutions thus need to be studied in terms of handling datasets and extracting knowledge and value. In addition, the rapid changes in data volume, variety, velocity, and value require decision makers to know how to obtain valuable insights.

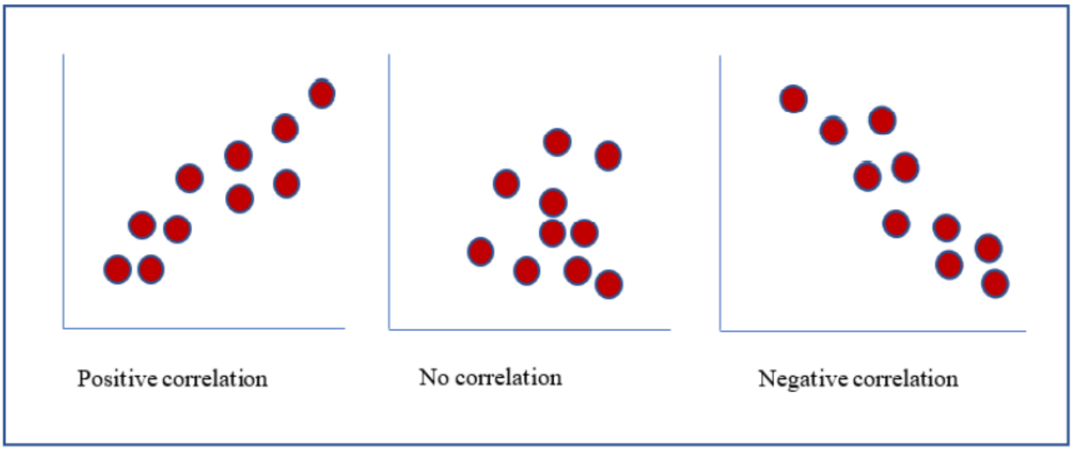


Figure 15: Correlation Analysis.

Traditional data analysis uses formal statistical methods to analyse data, constructing, extracting, and refining useful data, and identifying subject matter relationships in order to maximise the value of data. It can now be regarded as an analysis technique to be used for special kinds of data, though many traditional data analysis methods are still be used for big data analysis where analysts have backgrounds in statistics and computer science. Association rules, clustering, classification, decision trees, and regression are the most common data analytics methods; however, some additional analyses have become common in terms of big data, especially in terms of social media, which relies on social networking and content sharing. Social network analysis is thus dependent on the relationships between social entities. Text mining used to analyse the contents of documents and to develop an understanding of the information therein. Sentiment analysis is then used to analyse the emotions underlying that content, and this more important form of analysis uses language processing to identify such information.

Finally, advanced data visualisation is becoming an important analysis tool, as this enables faster and better decision making (Russom, 2011; Elgendy and Elragal, 2016). Some of the more common models and analyses are explained further below, and shown in Figure 16:

• **Text analytics: ➢ Sentiment Analysis:** This is based on understanding the subjects' emotions from their text patterns to help in organising viewpoints into good or bad, positive or negative (Mouthami et al., 2013). This analysis helps firms by alerting them where customers are dissatisfied or seeking to shift to other products, allowing preventative actions to be taken (Elgendy, N. and Elragal, A., 2014). • Audio analytics or speech analytics using technical approaches:

➢ LVCSR: large-vocabulary continuous speech recognition, indexing and searching. ➢ Phonetic-based systems: work with sounds or phonemes (Gandomi and Haider, 2015).

• **Social media and social network analysis (SNA):** Social media depends on multiple tools and frameworks for collecting, monitoring, summarising, analysing, and visualising social media data, and SNA depends on social entities' relationships with each other to measure the knowledge linking parties, including who shares information, what information, and with whom. SNA tries to get develop network patterns, while social media tries to uncover useful patterns and user information using text mining or sentiment analysis (Elgendy and Elragal, 2014; Gandomi and Haider, 2015).

• **Data Visualisation:** This can be used even by decision makers with little knowledge about the data, as it presents the information visually prior to deep analysis. Advanced Data visualisation (ADV) offersstrong potential growth to big data analytics as it allows analysis of data at several levels by taking advantage of human perceptual and reasoning abilities (Manyika et al., 2011; Russom, 2011; Elragal, and Klischewski, 2017)

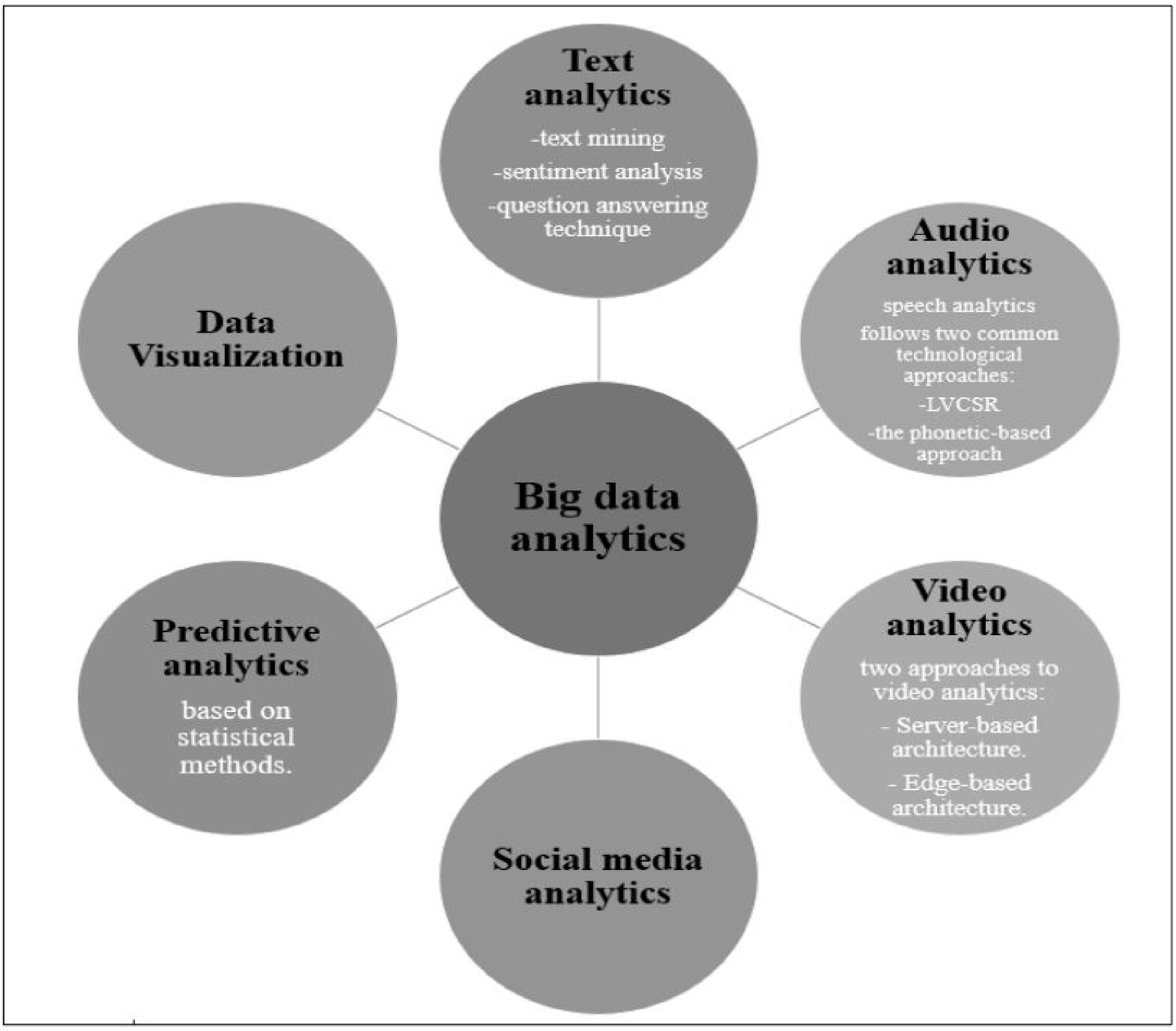

Text analytics
-text mining
-sentiment analysis
-question answering technique
Audio analytics
speech analytics follows two common technological approaches:
-LVCSR
-the phonetic-based approach
Data Visualization
Big data analytics
Predictive analytics
based on statistical methods.
Video analytics
two approaches to video analytics:
- Server-based architecture.
- Edge-based architecture.
Social media analytics

Figure 16: Common big data analytic methods.

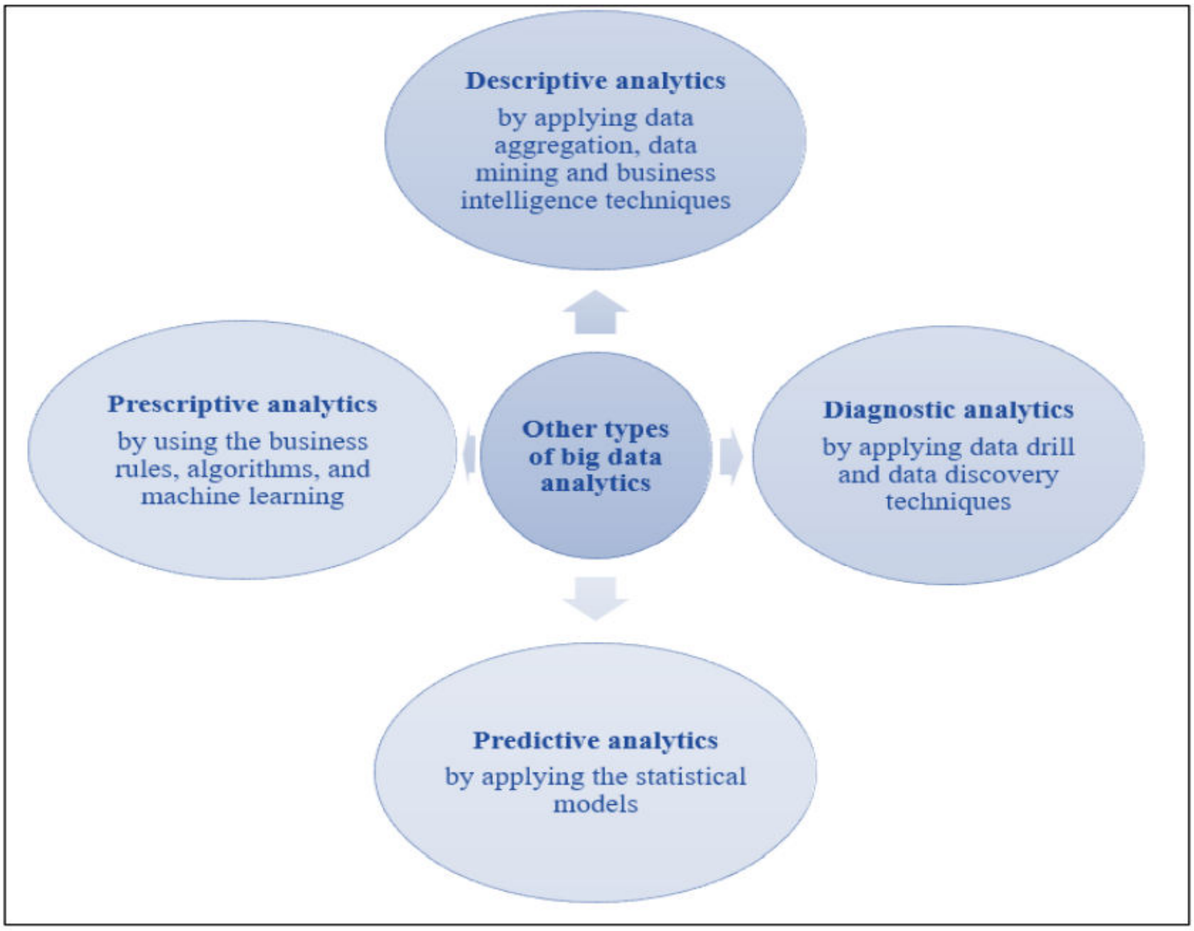

Descriptive analytics
by applying data aggregation, data mining and business intelligence techniques
Prescriptive analytics
by using the business rules, algorithms, and machine learning
Other types of big data analytics
Diagnostic analytics
by applying data drill and data discovery techniques
Predictive analytics
by applying the statistical models

Figure 17: Other types of big data analytics[2]

The other types of big data analytics used for systematic review are presented by Grover and Kar (2017), and these include descriptive analytics, diagnostic analytics, predictive analytics, and prescriptive analytics, as shown in Figure 17. Organisations and individual rend to use statistical models for predictive purposes, as most predictive models are built with statistical criteria. Artificial intelligence modelling is also becoming more popular. Machine learning algorithms can combine statistical and artificial intelligence methods in order to analyse large amounts of data with high-performance (Watson, 2019).

Descriptive analytics describes either what has happened or what is going to happen, while diagnostic analytics estimates the reason for something having happened, which requires techniques for discovering a problem’s root causes. Predictive analytics attempts to determine the most likely future outcomes by applying statistical models (Waller and Fawcett, 2013), while prescriptive analytics explains and predicts the future and describes outcomes using tools such as optimisation, simulation, business rules, algorithms, and machine learning (Banerjee et al., 2013; Grover and Kar, 2017).

SOFTWARE TOOLS FOR HANDLING BIG DATA There are many tools that help in achieving these goals and help data scientists to process data for analyzing them. Many new languages, frameworks and data storage technologies have emerged that supports handling of big data.

R: is an open-source statistical computing language that provides a wide variety of statistical and graphical techniques to derive insights from the data. It has an effective data handling and storage facility and supports vector operations with a suite of operators for faster processing. It has all the features of a standard programming language and supports conditional arguments, loops, and user-defined functions. R is supported by a huge number of packages through Comprehensive R Archive Network (CRAN). It is available on Windows, Linux, and Mac platforms. It has a strong documentation for each package. It has a strong support for data munging, data mining and machine learning algorithms along with a good support for reading and writing in distributed environment, which makes it appropriate for handling big data. However, the memory management, speed, and efficiency are probably the biggest challenge faced by R. R Studio is an Integrated Development Environment that is developed for programming in R language. It is distributed for standalone Desktop machines as well as it supports client-server architecture, which can be accessed from any browser.

Python: is yet another popular programming language, which is open source and is supported by Windows, Linux and Mac platforms. It hosts thousands of packages from third-party or community contributed modules. NumPy, Scikit, and Pandas support some of the popular packages for machine learning and data mining for data preprocessing, computing and modeling. NumPy is the base package for scientific computing. It adds support for large, multi-dimensional arrays and matrices with Python. Scikit supports classification, regression, clustering, dimensionality reduction, feature selection, and preprocessing and model selection algorithms. Pandas help in data munging and preparation for data analysis and modeling. It has strong support for graph analysis with its NetworkX library and nltk for text analytics and Natural language processing. Python is very user-friendly and great for quick and dirty analysis on a problem. It also integrates well with spark through the pyspark library.

Scala: is an object-oriented language and has an acronym for “Scalable Language”. The object and every operation in Scala is a method-call, just like any object-oriented language. It requires java virtual machine environment. Spark, an in-memory cluster computing framework is written in Scala. Scala is becoming popular programming tool for handling big data problems.

Apache Spark: is an in-memory cluster computing technology designed for fast computation, which is implemented in Scala. It uses Hadoop for storage purpose as it has its own cluster management capability. It provides built-in APIs for Java, Scala, and Python. Recently, it has also started supporting R. It comes with 80 high-level operators for interactive querying. The inmemory computation is supported with its Resilient Distributed Data (RDD) framework, which distributes the data frame into smaller chunks on different machines for faster computation. It also supports Map and Reduce for data processing. It supports SQL, data streaming, graph processing algorithms and machine learning algorithms. Though Spark can be accessed with Python, Java, and R, it has a strong support for Scala and is more stable at this point of time. It supports deep learning with sparkling water in H2O.

Apache Hive: is an open-source platform that provides facilities for querying and managing large datasets residing in distributed storage (For example, HDFS). It is similar to SQL and it is called as HiveQL. It uses Map Reduce for processing the queries and also supports developers to plug in their custom mapper and reducer codes when HiveQL lacks in expressing the desired logic. Apache Pig: is a platform that allows analysts to analyzing large data sets. It is a high-level programming language, called as Pig Latin for creating MapReduce programs that requires Hadoop for data storage. The Pig Latin code is extended with the help of User-Defined Functions that can be written in Java, Python and few other languages. It is amenable to substantial parallelization, which in turns enables them to handle very large data sets. Amazon Elastic Compute Cloud (EC2): is a web service that provides compute capacity over the cloud. It gives full control of the computing resources and allows developers to run their computation in the desired computing environment. It is one of the most successful cloud computing platforms. It works on the principle of the pay-as-you-go model. Few other frameworks that support big data are MongoDB, BlinkDB, Tachyon, Cassandra, CouchDB, Clojure, Tableau, Splunk and others

Sarah Al-Shiakhli , Big Data Analytics: A Literature Review Perspective