Big Data Analytics: A Literature Review Paper

Article *in* Lecture Notes in Computer Science · August 2014

DOI: 10.1007/978-3-319-08976-8_16

2 authors:

Nada Elgendy
The German University in Cairo
4 PUBLICATIONS **25** CITATIONS
SEE PROFILE

Ahmed Elragal
Luleå University of Technology
52 PUBLICATIONS **259** CITATIONS
SEE PROFILE

Big Data Analytics: A Literature Review Paper

Nada Elgendy and Ahmed Elragal

Department of Business Informatics & Operations,
German University in Cairo (GUC), Cairo, Egypt
{nada.el-gendy,ahmed.elragal}@guc.edu.eg

Abstract. In the information era, enormous amounts of data have become available on hand to decision makers. Big data refers to datasets that are not only big, but also high in variety and velocity, which makes them difficult to handle using traditional tools and techniques. Due to the rapid growth of such data, solutions need to be studied and provided in order to handle and extract value and knowledge from these datasets. Furthermore, decision makers need to be able to gain valuable insights from such varied and rapidly changing data, ranging from daily transactions to customer interactions and social network data. Such value can be provided using big data analytics, which is the application of advanced analytics techniques on big data. This paper aims to analyze some of the different analytics methods and tools which can be applied to big data, as well as the opportunities provided by the application of big data analytics in various decision domains.

Keywords: big data, data mining, analytics, decision making.

1 Introduction

Imagine a world without data storage; a place where every detail about a person or organization, every transaction performed, or every aspect which can be documented is lost directly after use. Organizations would thus lose the ability to extract valuable information and knowledge, perform detailed analyses, as well as provide new opportunities and advantages. Anything ranging from customer names and addresses, to products available, to purchases made, to employees hired, etc. has become essential for day-to-day continuity. Data is the building block upon which any organization thrives.

Now think of the extent of details and the surge of data and information provided nowadays through the advancements in technologies and the internet. With the increase in storage capabilities and methods of data collection, huge amounts of data have become easily available. Every second, more and more data is being created and needs to be stored and analyzed in order to extract value. Furthermore, data has become cheaper to store, so organizations need to get as much value as possible from the huge amounts of stored data.

The size, variety, and rapid change of such data require a new type of big data analytics, as well as different storage and analysis methods. Such sheer amounts of big data need to be properly analyzed, and pertaining information should be extracted.

P. Perner (Ed.): ICDM 2014, LNAI 8557, pp. 214–227, 2014.

The contribution of this paper is to provide an analysis of the available literature on big data analytics. Accordingly, some of the various big data tools, methods, and technologies which can be applied are discussed, and their applications and opportunities provided in several decision domains are portrayed.

The literature was selected based on its novelty and discussion of important topics related to big data, in order to serve the purpose of our research. The publication years range from 2008-2013, with most of the literature focusing on big data ranging from 2011-2013. This is due to big data being a recently focused upon topic. Furthermore, our corpus mostly includes research from some of the top journals, conferences, and white papers by leading corporations in the industry. Due to long review process of journals, most of the papers discussing big data analytics, its tools and methods, and its applications were found to be conference papers, and white papers. While big data analytics is being researched in academia, several of the industrial advancements and new technologies provided were mostly discussed in industry papers.

2 Big Data Analytics

The term "Big Data" has recently been applied to datasets that grow so large that they become awkward to work with using traditional database management systems. They are data sets whose size is beyond the ability of commonly used software tools and storage systems to capture, store, manage, as well as process the data within a tolerable elapsed time [12].

Big data sizes are constantly increasing, currently ranging from a few dozen terabytes (TB) to many petabytes (PB) of data in a single data set. Consequently, some of the difficulties related to big data include capture, storage, search, sharing, analytics, and visualizing. Today, enterprises are exploring large volumes of highly detailed data so as to discover facts they didn't know before [17].

Hence, big data analytics is where advanced analytic techniques are applied on big data sets. Analytics based on large data samples reveals and leverages business change. However, the larger the set of data, the more difficult it becomes to manage [17].

In this section, we will start by discussing the characteristics of big data, as well as its importance. Naturally, business benefit can commonly be derived from analyzing larger and more complex data sets that require real time or near-real time capabilities; however, this leads to a need for new data architectures, analytical methods, and tools. Therefore the successive section will elaborate the big data analytics tools and methods, in particular, starting with the big data storage and management, then moving on to the big data analytic processing. It then concludes with some of the various big data analyses which have grown in usage with big data.

2.1 Characteristics of Big Data

Big data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures, analytics, and tools in order to enable insights that unlock new sources of business value. Three main features characterize big data: volume, variety, and velocity, or the three V's. The volume of the data is its size, and

how enormous it is. Velocity refers to the rate with which data is changing, or how often it is created. Finally, variety includes the different formats and types of data, as well as the different kinds of uses and ways of analyzing the data [9].

Data volume is the primary attribute of big data. Big data can be quantified by size in TBs or PBs, as well as even the number of records, transactions, tables, or files. Additionally, one of the things that make big data really big is that it's coming from a greater variety of sources than ever before, including logs, clickstreams, and social media. Using these sources for analytics means that common structured data is now joined by unstructured data, such as text and human language, and semi-structured data, such as eXtensible Markup Language (XML) or Rich Site Summary (RSS) feeds. There's also data, which is hard to categorize since it comes from audio, video, and other devices. Furthermore, multi-dimensional data can be drawn from a data warehouse to add historic context to big data. Thus, with big data, variety is just as big as volume.

Moreover, big data can be described by its velocity or speed. This is basically the frequency of data generation or the frequency of data delivery. The leading edge of big data is streaming data, which is collected in real-time from the websites [17]. Some researchers and organizations have discussed the addition of a fourth V, or veracity. Veracity focuses on the quality of the data. This characterizes big data quality as good, bad, or undefined due to data inconsistency, incompleteness, ambiguity, latency, deception, and approximations [22].

2.2 Big Data Analytics Tools and Methods

With the evolution of technology and the increased multitudes of data flowing in and out of organizations daily, there has become a need for faster and more efficient ways of analyzing such data. Having piles of data on hand is no longer enough to make efficient decisions at the right time.

Such data sets can no longer be easily analyzed with traditional data management and analysis techniques and infrastructures. Therefore, there arises a need for new tools and methods specialized for big data analytics, as well as the required architectures for storing and managing such data. Accordingly, the emergence of big data has an effect on everything from the data itself and its collection, to the processing, to the final extracted decisions.

Consequently, [8] proposed the Big – Data, Analytics, and Decisions (B-DAD) framework which incorporates the big data analytics tools and methods into the decision making process [8]. The framework maps the different big data storage, management, and processing tools, analytics tools and methods, and visualization and evaluation tools to the different phases of the decision making process. Hence, the changes associated with big data analytics are reflected in three main areas: big data storage and architecture, data and analytics processing, and, finally, the big data analyses which can be applied for knowledge discovery and informed decision making. Each area will be further discussed in this section. However, since big data is still evolving as an important field of research, and new findings and tools are constantly developing, this section is not exhaustive of all the possibilities, and focuses on providing a general idea, rather than a list of all potential opportunities and technologies.

Big Data Storage and Management

One of the first things organizations have to manage when dealing with big data, is where and how this data will be stored once it is acquired. The traditional methods of structured data storage and retrieval include relational databases, data marts, and data warehouses. The data is uploaded to the storage from operational data stores using Extract, Transform, Load (ETL), or Extract, Load, Transform (ELT), tools which extract the data from outside sources, transform the data to fit operational needs, and finally load the data into the database or data warehouse. Thus, the data is cleaned, transformed, and catalogued before being made available for data mining and online analytical functions [3].

However, the big data environment calls for Magnetic, Agile, Deep (MAD) analysis skills, which differ from the aspects of a traditional Enterprise Data Warehouse (EDW) environment. First of all, traditional EDW approaches discourage the incorporation of new data sources until they are cleansed and integrated. Due to the ubiquity of data nowadays, big data environments need to be magnetic, thus attracting all the data sources, regardless of the data quality [5]. Furthermore, given the growing numbers of data sources, as well as the sophistication of the data analyses, big data storage should allow analysts to easily produce and adapt data rapidly. This requires an agile database, whose logical and physical contents can adapt in sync with rapid data evolution [11]. Finally, since current data analyses use complex statistical methods, and analysts need to be able to study enormous datasets by drilling up and down, a big data repository also needs to be deep, and serve as a sophisticated algorithmic runtime engine [5].

Accordingly, several solutions, ranging from distributed systems and Massive Parallel Processing (MPP) databases for providing high query performance and platform scalability, to non-relational or in-memory databases, have been used for big data.

Non-relational databases, such as Not Only SQL (NoSQL), were developed for storing and managing unstructured, or non-relational, data. NoSQL databases aim for massive scaling, data model flexibility, and simplified application development and deployment. Contrary to relational databases, NoSQL databases separate data management and data storage. Such databases rather focus on the high-performance scalable data storage, and allow data management tasks to be written in the application layer instead of having it written in databases specific languages [3].

On the other hand, in-memory databases manage the data in server memory, thus eliminating disk input/output (I/O) and enabling real-time responses from the database. Instead of using mechanical disk drives, it is possible to store the primary database in silicon-based main memory. This results in orders of magnitude of improvement in the performance, and allows entirely new applications to be developed [16]. Furthermore, in-memory databases are now being used for advanced analytics on big data, especially to speed the access to and scoring of analytic models for analysis. This provides scalability for big data, and speed for discovery analytics [17].

Alternatively, Hadoop is a framework for performing big data analytics which provides reliability, scalability, and manageability by providing an implementation for the MapReduce paradigm, which is discussed in the following section, as well as gluing the storage and analytics together. Hadoop consists of two main components: the HDFS for the big data storage, and MapReduce for big data analytics [9]. The HDFS storage function provides a redundant and reliable distributed file system, which is optimized for large files, where a single file is split into blocks and distributed across

cluster nodes. Additionally, the data is protected among the nodes by a replication mechanism, which ensures availability and reliability despite any node failures [3]. There are two types of HDFS nodes: the Data Nodes and the Name Nodes. Data is stored in replicated file blocks across the multiple Data Nodes, and the Name Node acts as a regulator between the client and the Data Node, directing the client to the particular Data Node which contains the requested data [3].

Big Data Analytic Processing

After the big data storage, comes the analytic processing. According to [10], there are four critical requirements for big data processing. The first requirement is fast data loading. Since the disk and network traffic interferes with the query executions during data loading, it is necessary to reduce the data loading time. The second requirement is fast query processing. In order to satisfy the requirements of heavy workloads and real-time requests, many queries are response-time critical. Thus, the data placement structure must be capable of retaining high query processing speeds as the amounts of queries rapidly increase. Additionally, the third requirement for big data processing is the highly efficient utilization of storage space. Since the rapid growth in user activities can demand scalable storage capacity and computing power, limited disk space necessitates that data storage be well managed during processing, and issues on how to store the data so that space utilization is maximized be addressed. Finally, the fourth requirement is the strong adaptivity to highly dynamic workload patterns. As big data sets are analyzed by different applications and users, for different purposes, and in various ways, the underlying system should be highly adaptive to unexpected dynamics in data processing, and not specific to certain workload patterns [10].

Map Reduce is a parallel programming model, inspired by the "Map" and "Reduce" of functional languages, which is suitable for big data processing. It is the core of Hadoop, and performs the data processing and analytics functions [6]. According to EMC, the MapReduce paradigm is based on adding more computers or resources, rather than increasing the power or storage capacity of a single computer; in other words, scaling out rather than scaling up [9]. The fundamental idea of MapReduce is breaking a task down into stages and executing the stages in parallel in order to reduce the time needed to complete the task [6].

The first phase of the MapReduce job is to map input values to a set of key/value pairs as output. The "Map" function accordingly partitions large computational tasks into smaller tasks, and assigns them to the appropriate key/value pairs [6]. Thus, unstructured data, such as text, can be mapped to a structured key/value pair, where, for example, the key could be the word in the text and the value is the number of occurrences of the word. This output is then the input to the "Reduce" function [9]. Reduce then performs the collection and combination of this output, by combining all values which share the same key value, to provide the final result of the computational task [6].

The MapReduce function within Hadoop depends on two different nodes: the Job Tracker and the Task Tracker nodes. The Job Tracker nodes are the ones which are responsible for distributing the mapper and reducer functions to the available Task Trackers, as well as monitoring the results [9]. The MapReduce job starts by the JobTracker assigning a portion of an input file on the HDFS to a map task, running on a node [13]. On the other hand, the Task Tracker nodes actually run the jobs and communicate results back to the Job Tracker. That communication between nodes is often through files and directories in HDFS, so inter-node communication is minimized [9].

Figure 1 shows how the MapReduce nodes and the HDFS work together. At step 1, there is a very large dataset including log files, sensor data, or anything of the sorts. The HDFS stores replicas of the data, represented by the blue, yellow, beige, and pink icons, across the Data Nodes. In step 2, the client defines and executes a map job and a reduce job on a particular data set, and sends them both to the Job Tracker. The Job Tracker then distributes the jobs across the Task Trackers in step 3. The Task Tracker runs the mapper, and the mapper produces output that is then stored in the HDFS file system. Finally, in step 4, the reduce job runs across the mapped data in order to produce the result.

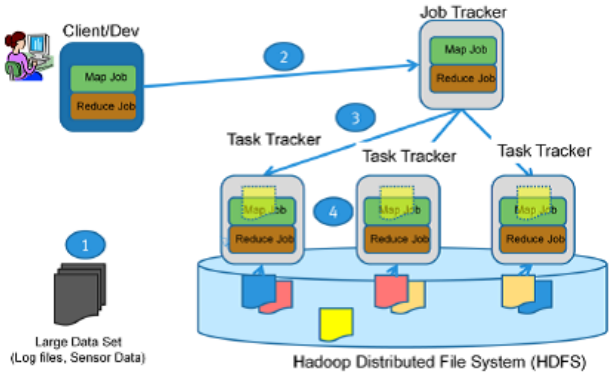

Fig. 1. MapReduce and HDFS

Hadoop is a MAD system, thus making it popular for big data analytics by loading data as files into the distributed file system, and running parallel MapReduce computations on the data. Hadoop gets its magnetism and agility from the fact that data is loaded into Hadoop simply by copying files into the distributed file system, and MapReduce interprets the data at processing time rather than loading time [11]. Thus, it is capable of attracting all data sources, as well as adapting its engines to any evolutions that may occur in such big data sources [6].

After big data is stored, managed, and processed, decision makers need to extract useful insights by performing big data analyses. In the subsections below, various big data analyses will be discussed, starting with selected traditional advanced data analytics methods, and followed by examples of some of the additional, applicable big data analyses.

Big Data Analytics

Nowadays, people don't just want to collect data, they want to understand the meaning and importance of the data, and use it to aid them in making decisions. Data analytics is the process of applying algorithms in order to analyze sets of data and extract useful and unknown patterns, relationships, and information [1]. Furthermore, data analytics are used to extract previously unknown, useful, valid, and hidden patterns and information from large data sets, as well as to detect important relationships among the stored variables. Therefore, analytics have had a significant impact on

research and technologies, since decision makers have become more and more interested in learning from previous data, thus gaining competitive advantage [21].

Along with some of the most common advanced data analytics methods, such as association rules, clustering, classification and decision trees, and regression some additional analyses have become common with big data.

For example, social media has recently become important for social networking and content sharing. Yet, the content that is generated from social media websites is enormous and remains largely unexploited. However, social media analytics can be used to analyze such data and extract useful information and predictions [2]. Social media analytics is based on developing and evaluating informatics frameworks and tools in order to collect, monitor, summarize, analyze, as well as visualize social media data. Furthermore, social media analytics facilitates understanding the reactions and conversations between people in online communities, as well as extracting useful patterns and intelligence from their interactions, in addition to what they share on social media websites [24].

On the other hand, Social Network Analysis (SNA) focuses on the relationships among social entities, as well as the patterns and implications of such relationships [23]. An SNA maps and measures both formal and informal relationships in order to comprehend what facilitates the flow of knowledge between interacting parties, such as who knows who, and who shares what knowledge or information with who and using what [19].

However, SNA differs from social media analysis, in that SNA tries to capture the social relationships and patterns between networks of people. On the other hand, social media analysis aims to analyze what social media users are saying in order to uncover useful patterns, information about the users, and sentiments. This is traditionally done using text mining or sentiment analysis, which are discussed below.

On the other hand, text mining is used to analyze a document or set of documents in order to understand the content within and the meaning of the information contained. Text mining has become very important nowadays since most of the information stored, not including audio, video, and images, consists of text. While data mining deals with structured data, text presents special characteristics which basically follow a non-relational form [18].

Moreover, sentiment analysis, or opinion mining, is becoming more and more important as online opinion data, such as blogs, product reviews, forums, and social data from social media sites like Twitter and Facebook, grow tremendously. Sentiment analysis focuses on analyzing and understanding emotions from subjective text patterns, and is enabled through text mining. It identifies opinions and attitudes of individuals towards certain topics, and is useful in classifying viewpoints as positive or negative. Sentiment analysis uses natural language processing and text analytics in order to identify and extract information by finding words that are indicative of a sentiment, as well as relationships between words, so that sentiments can be accurately identified [15].

Finally, from the strongest potential growths among big data analytics options is Advanced Data Visualization (ADV) and visual discovery [17]. Presenting information so that people can consume it effectively is a key challenge that needs to be met, in order for decision makers to be able to properly analyze data in a way to lead to concrete actions [14].

ADV has emerged as a powerful technique to discover knowledge from data. ADV combines data analysis methods with interactive visualization to enable comprehensive data exploration. It is a data driven exploratory approach that fits well in situations where analysts have little knowledge about the data [20]. With the generation of more and more data of high volume and complexity, an increasing demand has arisen for ADV solutions from many application domains [25]. Additionally, such visualization analyses take advantage of human perceptual and reasoning abilities, which enables them to thoroughly analyze data at both the overview and the detailed levels. Along with the size and complexity of big data, intuitive visual representation and interaction is needed to facilitate the analyst's perception and reasoning [20].

ADV can enable faster analysis, better decision making, and more effective presentation and comprehension of results by providing interactive statistical graphics and a point-and-click interface [4]. Furthermore, ADV is a natural fit for big data since it can scale its visualizations to represent thousands or millions of data points, unlike standard pie, bar, and line charts. Moreover, it can handle diverse data types, as well as present analytic data structures that aren't easily flattened onto a computer screen, such as hierarchies and neural nets. Additionally, most ADV tools and functions can support interfaces to all the leading data sources, thus enabling business analysts to explore data widely across a variety of sources in search of the right analytics dataset, usually in real-time [17].

3 Big Data Analytics and Decision Making

From the decision maker's perspective, the significance of big data lies in its ability to provide information and knowledge of value, upon which to base decisions. The managerial decision making process has been an important and thoroughly covered topic in research throughout the years.

Big data is becoming an increasingly important asset for decision makers. Large volumes of highly detailed data from various sources such as scanners, mobile phones, loyalty cards, the web, and social media platforms provide the opportunity to deliver significant benefits to organizations. This is possible only if the data is properly analyzed to reveal valuable insights, allowing for decision makers to capitalize upon the resulting opportunities from the wealth of historic and real-time data generated through supply chains, production processes, customer behaviors, etc. [4].

Moreover, organizations are currently accustomed to analyzing internal data, such as sales, shipments, and inventory. However, the need for analyzing external data, such as customer markets and supply chains, has arisen, and the use of big data can provide cumulative value and knowledge. With the increasing sizes and types of unstructured data on hand, it becomes necessary to make more informed decisions based on drawing meaningful inferences from the data [7].

Accordingly, [8] developed the B-DAD framework which maps big data tools and techniques, into the decision making process [8]. Such a framework is intended to enhance the quality of the decision making process in regards to dealing with big data. The first phase of the decision making process is the intelligence phase, where data which can be used to identify problems and opportunities is collected from internal and external data sources. In this phase, the sources of big data need to be identified,

and the data needs to be gathered from different sources, processed, stored, and migrated to the end user. Such big data needs to be treated accordingly, so after the data sources and types of data required for the analysis are defined, the chosen data is acquired and stored in any of the big data storage and management tools previously discussed After the big data is acquired and stored, it is then organized, prepared, and processed, This is achieved across a high-speed network using ETL/ELT or big data processing tools, which have been covered in the previous sections.

The next phase in the decision making process is the design phase, where possible courses of action are developed and analyzed through a conceptualization, or a representative model of the problem. The framework divides this phase into three steps, model planning, data analytics, and analyzing. Here, a model for data analytics, such as those previously discussed, is selected and planned, and then applied, and finally analyzed.

Consequently, the following phase in the decision making process is the choice phase, where methods are used to evaluate the impacts of the proposed solutions, or courses of action, from the design phase. Finally, the last phase in the decision making process is the implementation phase, where the proposed solution from the previous phase is implemented [8].

As the amount of big data continues to exponentially grow, organizations throughout the different sectors are becoming more interested in how to manage and analyze such data. Thus, they are rushing to seize the opportunities offered by big data, and gain the most benefit and insight possible, consequently adopting big data analytics in order to unlock economic value and make better and faster decisions. Therefore, organizations are turning towards big data analytics in order to analyze huge amounts of data faster, and reveal previously unseen patterns, sentiments, and customer intelligence. This section focuses on some of the different applications, both proposed and implemented, of big data analytics, and how these applications can aid organizations across different sectors to gain valuable insights and enhance decision making.

According to Manyika et al.'s research, big data can enable companies to create new products and services, enhance existing ones, as well as invent entirely new business models. Such benefits can be gained by applying big data analytics in different areas, such as customer intelligence, supply chain intelligence, performance, quality and risk management and fraud detection [14]. Furthermore, Cebr's study highlighted the main industries that can benefit from big data analytics, such as the manufacturing, retail, central government, healthcare, telecom, and banking industries [4].

3.1 Customer Intelligence

Big data analytics holds much potential for customer intelligence, and can highly benefit industries such as retail, banking, and telecommunications. Big data can create transparency, and make relevant data more easily accessible to stakeholders in a timely manner [14]. Big data analytics can provide organizations with the ability to profile and segment customers based on different socioeconomic characteristics, as well as increase levels of customer satisfaction and retention [4]. This can allow them to make more informed marketing decisions, and market to different segments based on their preferences along with the recognition of sales and marketing opportunities [17]. Moreover, social media can be used to inform companies what their customers like, as

well as what they don't like. By performing sentiment analysis on this data, firms can be alerted beforehand when customers are turning against them or shifting to different products, and accordingly take action [7].

Additionally, using SNAs to monitor customer sentiments towards brands, and identify influential individuals, can help organizations react to trends and perform direct marketing. Big data analytics can also enable the construction of predictive models for customer behavior and purchase patterns, therefore raising overall profitability [4]. Even organizations which have used segmentation for many years are beginning to deploy more sophisticated big data techniques, such as real-time micro-segmentation of customers, in order to target promotions and advertising [14]. Consequently, big data analytics can benefit organizations by enabling better targeted social influencer marketing, defining and predicting trends from market sentiments, as well as analyzing and understanding churn and other customer behaviors [17].

3.2 Supply Chain and Performance Management

As for supply chain management, big data analytics can be used to forecast demand changes, and accordingly match their supply. This can increasingly benefit the manufacturing, retail, as well as transport and logistics industries. By analyzing stock utilization and geospatial data on deliveries, organizations can automate replenishment decisions, which will reduce lead times and minimize costs and delays, as well as process interruptions. Additionally, decisions on changing suppliers, based on quality or price competitiveness, can be taken by analyzing supplier data to monitor performance. Furthermore, alternate pricing scenarios can be run instantly, which can enable a reduction in inventories and an increase in profit margins [4]. Accordingly, big data can lead to the identification of the root causes of cost, and provide for better planning and forecasting [17].

Another area where big data analytics can be of value is performance management, where the governmental and healthcare industries can easily benefit. With the increasing need to improve productivity, staff performance information can be monitored and forecasted by using predictive analytics tools. This can allow departments to link their strategic objectives with the service or user outcomes, thus leading to increased efficiencies. Additionally, with the availability of big data and performance information, as well as its accessibility to operations managers, the use of predictive KPIs, balanced scorecards, and dashboards within the organization can introduce operational benefits by enabling the monitoring of performance, as well as improving transparency, objectives setting, and planning and management functions [4].

3.3 Quality Management and Improvement

Especially for the manufacturing, energy and utilities, and telecommunications industries, big data can be used for quality management, in order to increase profitability and reduce costs by improving the quality of goods and services provided. For example, in the manufacturing process, predictive analytics on big data can be used to minimize the performance variability, as well as prevent quality issues by providing early warning alerts. This can reduce scrap rates, and decrease the time to market, since identifying any disruptions to the production process before they occur can save

significant expenditures [4]. Additionally, big data analytics can result in manufacturing lead improvements [17]. Furthermore, real-time data analyses and monitoring of machine logs can enable managers to make swifter decisions for quality management. Also, big data analytics can allow for the real-time monitoring of network demand, in addition to the forecasting of bandwidth in response to customer behavior.

Moreover, healthcare IT systems can improve the efficiency and quality of care, by communicating and integrating patient data across different departments and institutions, while retaining privacy controls [4]. Analyzing electronic health records can improve the continuity of care for individuals, as well as creating a massive dataset through which treatments and outcomes can be predicted and compared. Therefore, with the increasing use of electronic health records, along with the advancements in analytics tools, there arises an opportunity to mine the available de-identified patient information for assessing the quality of healthcare, as well as managing diseases and health services [22].

Additionally, the quality of citizens' lives can be improved through the utilization of big data. For healthcare, sensors can be used in hospitals and homes to provide the continuous monitoring of patients, and perform real-time analyses on the patient data streaming in. This can be used to alert individuals and their health care providers if any health anomalies are detected in the analysis, requiring the patient to seek medical help [22]. Patients can also be monitored remotely to analyze their adherence to their prescriptions, and improve drug and treatment options [14].

Moreover, by analyzing information from distributed sensors on handheld devices, roads, and vehicles, which provide real-time traffic information, transportation can be transformed and improved. Traffic jams can be predicted and prevented, and drivers can operate more safely and with less disruption to the traffic flow. Such a new type of traffic ecosystem, with "intelligent" connected cars, can potentially renovate transportation and how roadways are used [22]. Accordingly, big data applications can provide smart routing, according to real-time traffic information based on personal location data. Furthermore, such applications can automatically call for help when trouble is detected by the sensors, and inform users about accidents, scheduled roadwork, and congested areas in real-time [14].

Furthermore, big data can be used for better understanding changes in the location, frequency, and intensity of weather and climate. This can benefit citizens and businesses that rely upon weather, such as farmers, as well as tourism and transportation companies. Also, with new sensors and analysis techniques for developing long term climate models and nearer weather forecasts, weather related natural disasters can be predicted, and preventive or adaptive measures can be taken beforehand [22].

3.4 Risk Management and Fraud Detection

Industries such as investment or retail banking, as well as insurance, can benefit from big data analytics in the area of risk management. Since the evaluation and bearing of risk is a critical aspect for the financial services sector, big data analytics can help in selecting investments by analyzing the likelihood of gains against the likelihood of losses. Additionally, internal and external big data can be analyzed for the full and dynamic appraisal of risk exposures [4]. Accordingly, big data can benefit organizations by enabling the quantification of risks [17]. High-performance analytics can also

be used to integrate the risk profiles managed in isolation across separate departments, into enterprise wide risk profiles. This can aid in risk mitigation, since a comprehensive view of the different risk types and their interrelations is provided to decision makers [4].

Furthermore, new big data tools and technologies can provide for managing the exponential growth in network produced data, as well reduce database performance problems by increasing the ability to scale and capture the required data. Along with the enhancement in cyber analytics and data intensive computing solutions, organizations can incorporate multiple streams of data and automated analyses to protect themselves against cyber and network attacks [22].

As for fraud detection, especially in the government, banking, and insurance industries, big data analytics can be used to detect and prevent fraud [17]. Analytics are already commonly used in automated fraud detection, but organizations and sectors are looking towards harnessing the potentials of big data in order to improve their systems. Big data can allow them to match electronic data across several sources, between both public and private sectors, and perform faster analytics [4].

In addition, customer intelligence can be used to model normal customer behavior, and detect suspicious or divergent activities through the accurate flagging of outlier occurrences. Furthermore, providing systems with big data about prevailing fraud patterns can allow these systems to learn the new types of frauds and act accordingly, as the fraudsters adapt to the old systems designed to detect them. Also, SNAs can be used to identify the networks of collaborating fraudsters, as well as discover evidence of fraudulent insurance or benefits claims, which will lead to less fraudulent activity going undiscovered [4]. Thus, big data tools, techniques, and governance processes can increase the prevention and recovery of fraudulent transactions by dramatically increasing the speed of identification and detection of compliance patterns within all available data sets [22].

4 Conclusion

In this research, we have examined the innovative topic of big data, which has recently gained lots of interest due to its perceived unprecedented opportunities and benefits. In the information era we are currently living in, voluminous varieties of high velocity data are being produced daily, and within them lay intrinsic details and patterns of hidden knowledge which should be extracted and utilized. Hence, big data analytics can be applied to leverage business change and enhance decision making, by applying advanced analytic techniques on big data, and revealing hidden insights and valuable knowledge.

Accordingly, the literature was reviewed in order to provide an analysis of the big data analytics concepts which are being researched, as well as their importance to decision making. Consequently, big data was discussed, as well as its characteristics and importance. Moreover, some of the big data analytics tools and methods in particular were examined. Thus, big data storage and management, as well as big data analytics processing were detailed. In addition, some of the different advanced data analytics techniques were further discussed.

By applying such analytics to big data, valuable information can be extracted and exploited to enhance decision making and support informed decisions. Consequently, some of the different areas where big data analytics can support and aid in decision making were examined. It was found that big data analytics can provide vast horizons of opportunities in various applications and areas, such as customer intelligence, fraud detection, and supply chain management. Additionally, its benefits can serve different sectors and industries, such as healthcare, retail, telecom, manufacturing, etc.

Accordingly, this research has provided the people and the organizations with examples of the various big data tools, methods, and technologies which can be applied. This gives users an idea of the necessary technologies required, as well as developers an idea of what they can do to provide more enhanced solutions for big data analytics in support of decision making. Thus, the support of big data analytics to decision making was depicted.

Finally, any new technology, if applied correctly can bring with it several potential benefits and innovations, let alone big data, which is a remarkable field with a bright future, if approached correctly. However, big data is very difficult to deal with. It requires proper storage, management, integration, federation, cleansing, processing, analyzing, etc. With all the problems faced with traditional data management, big data exponentially increases these difficulties due to additional volumes, velocities, and varieties of data and sources which have to be dealt with. Therefore, future research can focus on providing a roadmap or framework for big data management which can encompass the previously stated difficulties.

We believe that big data analytics is of great significance in this era of data overflow, and can provide unforeseen insights and benefits to decision makers in various areas. If properly exploited and applied, big data analytics has the potential to provide a basis for advancements, on the scientific, technological, and humanitarian levels.

References

1. Adams, M.N.: Perspectives on Data Mining. International Journal of Market Research 52(1), 11–19 (2010)
2. Asur, S., Huberman, B.A.: Predicting the Future with Social Media. In: ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 492–499 (2010)
3. Bakshi, K.: Considerations for Big Data: Architecture and Approaches. In: Proceedings of the IEEE Aerospace Conference, pp. 1–7 (2012)
4. Cebr: Data equity, Unlocking the value of big data. in: SAS Reports, pp. 1–44 (2012)
5. Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., Welton, C.: MAD Skills: New Analysis Practices for Big Data. Proceedings of the ACM VLDB Endowment 2(2), 1481–1492 (2009)
6. Cuzzocrea, A., Song, I., Davis, K.C.: Analytics over Large-Scale Multidimensional Data: The Big Data Revolution! In: Proceedings of the ACM International Workshop on Data Warehousing and OLAP, pp. 101–104 (2011)
7. Economist Intelligence Unit: The Deciding Factor: Big Data & Decision Making. In: Capgemini Reports, pp. 1–24 (2012)

8. Elgendy, N.: Big Data Analytics in Support of the Decision Making Process. MSc Thesis, German University in Cairo, p. 164 (2013)
9. EMC: Data Science and Big Data Analytics. In: EMC Education Services, pp. 1–508 (2012)
10. He, Y., Lee, R., Huai, Y., Shao, Z., Jain, N., Zhang, X., Xu, Z.: RCFile: A Fast and Space-efficient Data Placement Structure in MapReduce-based Warehouse Systems. In: IEEE International Conference on Data Engineering (ICDE), pp. 1199–1208 (2011)
11. Herodotou, H., Lim, H., Luo, G., Borisov, N., Dong, L., Cetin, F.B., Babu, S.: Starfish: A Self-tuning System for Big Data Analytics. In: Proceedings of the Conference on Innovative Data Systems Research, pp. 261–272 (2011)
12. Kubick, W.R.: Big Data, Information and Meaning. In: Clinical Trial Insights, pp. 26–28 (2012)
13. Lee, R., Luo, T., Huai, Y., Wang, F., He, Y., Zhang, X.: Ysmart: Yet Another SQL-to-MapReduce Translator. In: IEEE International Conference on Distributed Computing Systems (ICDCS), pp. 25–36 (2011)
14. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big Data: The Next Frontier for Innovation, Competition, and Productivity. In: McKinsey Global Institute Reports, pp. 1–156 (2011)
15. Mouthami, K., Devi, K.N., Bhaskaran, V.M.: Sentiment Analysis and Classification Based on Textual Reviews. In: International Conference on Information Communication and Embedded Systems (ICICES), pp. 271–276 (2013)
16. Plattner, H., Zeier, A.: In-Memory Data Management: An Inflection Point for Enterprise Applications. Springer, Heidelberg (2011)
17. Russom, P.: Big Data Analytics. In: TDWI Best Practices Report, pp. 1–40 (2011)
18. Sanchez, D., Martin-Bautista, M.J., Blanco, I., Torre, C.: Text Knowledge Mining: An Alternative to Text Data Mining. In: IEEE International Conference on Data Mining Workshops, pp. 664–672 (2008)
19. Serrat, O.: Social Network Analysis. Knowledge Network Solutions 28, 1–4 (2009)
20. Shen, Z., Wei, J., Sundaresan, N., Ma, K.L.: Visual Analysis of Massive Web Session Data. In: Large Data Analysis and Visualization (LDAV), pp. 65–72 (2012)
21. Song, Z., Kusiak, A.: Optimizing Product Configurations with a Data Mining Approach. International Journal of Production Research 47(7), 1733–1751 (2009)
22. TechAmerica: Demystifying Big Data: A Practical Guide to Transforming the Business of Government. In: TechAmerica Reports, pp. 1–40 (2012)
23. Van der Valk, T., Gijsbers, G.: The Use of Social Network Analysis in Innovation Studies: Mapping Actors and Technologies. Innovation: Management, Policy & Practice 12(1), 5–17 (2010)
24. Zeng, D., Hsinchun, C., Lusch, R., Li, S.H.: Social Media Analytics and Intelligence. IEEE Intelligent Systems 25(6), 13–16 (2010)
25. Zhang, L., Stoffel, A., Behrisch, M., Mittelstadt, S., Schreck, T., Pompl, R., Weber, S., Last, H., Keim, D.: Visual Analytics for the Big Data Era—A Comparative Review of State-of-the-Art Commercial Systems. In: IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 173–182 (2012)