International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 21 (2018) pp. 15386-15396

Big Data Analytics in Data Mining – A Review

[1]M.Uma, [2] **Dr. V. Baby Deepa**,
[1]Ph.D. Research Scholar(P.T.), PG and Research Department of Computer Science, Government Arts College (Autonomous), Karur-639 005, Tamilnadu, India.

[2]Assistant Professor, PG and Research Department of Computer, Science, Government Arts College (Autonomous), Karur-639 005, Tamilnadu, India.

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 the extract value and knowledge from these data sets. 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 presents an overview of big data content, classification, analytics, algorithm, open issues and challenges.

Keywords: Big Data, Data Mining, Analytics, Data Input, Analysis, Framework

INTRODUCTION

As the information technology spreads fast, most of the data were born digital as well as exchanged on internet today. According to the estimation of Lyman and Varian [1], the new data stored in digital media devices have already been more than 92 % in 2002, while the size of these new data was also more than five Exabyte's. In fact, the problems of analyzing the large scale data were not suddenly occurred but have been there for several years because the creation of data is usually much easier than finding useful things from the data. Even though computer systems today are much faster than those in the 1930s, the large scale data is a strain to analyze by the computers we have today.

In response to the problems of analyzing large-scale data, quite a few efficient methods [2], such as sampling, data condensation, density-based approaches, grid-based approaches, divide and conquer, incremental learning, and distributed computing, have been presented. Of course, these methods are constantly used to improve the performance of the operators of data analytics process. The results of these methods illustrate that with the efficient methods at hand, we may be able to analyze the large-scale data in a reasonable time. The dimensional reduction method (e.g., principal components analysis; PCA [3]) is a typical example that is aimed at reducing the input data volume to accelerate the process of data analytics. Another reduction method that reduces the data computations of data clustering is sampling [4], which can also be used to speed up the computation time of data analytics.

Although the advances of computer systems and internet technologies have witnessed the development of computing hardware following the Moore's law for several decades, the problems of handling the large-scale data still exist when we are entering the age of big data. That is why Fisher et al. [5] pointed out that big data means that the data is unable to be handled and processed by most current information systems or methods because data in the big data era will not only become too big to be loaded into a single machine, it also implies that most traditional data mining methods or data analytics developed for a centralized data analysis process may not be able to be applied directly to big data. In addition to the issues of data size, Laney [6] presented a well-known definition (also called 3Vs) to explain what the "big" data is: volume, velocity, and variety. The definition of 3Vs implies that the data size is large, the data will be created rapidly, and the data will be existed in multiple types and captured from different sources, respectively. Later studies [7] pointed out that the definition of 3Vs is insufficient to explain the big data we face now. Thus, veracity, validity, value, variability, venue, vocabulary, and vagueness were added to make some complement explanation of big data [8].

A numerous researches are therefore focusing on developing effective technologies to analyze the big data. To discuss in deep the big data analytics, this paper gives not only a systematic description of traditional large-scale data analytics but also a detailed discussion about the differences between data and big data analytics framework for the data scientists or researchers to focus on the big data analytics.

Moreover, although several data analytics and frameworks have been presented in recent years, with their pros and cons being discussed in different studies, a complete discussion from the perspective of data mining and knowledge discovery in databases still is needed. As a result, this paper is aimed at providing a brief review for the researchers on the data mining and distributed computing domains to have a basic idea to use or develop data analytics for big data.

DEFINITION AND CHARACTERISTICS OF BIG DATA

Big data is a term utilized to refer to the increase in the volume of data that are difficult to store, process and analyze through traditional database technologies. The nature of big data is indistinct and involves considerable processes to identify and translate the data into new insight. The tern "big data" is relatively new in IT and business. However, several researchers and practitioners have utilized the term in previous literature. For instance, [9] referred to big data as a large volume of scientific data for visualization. Several definitions of big data currently exist. For instance, [10] defined big data as "the amount of data just beyond technology's capability to store, manage, and process efficiently." Meanwhile, [11] and [12] defined big data as characterized by three Vs: volume, variety, and velocity. The terms volume, variety, and velocity were originally introduced by Gartner to describe the elements of big data challenges. IDC also defined big data technologies as "a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high velocity capture, discovery, and/or analysis." [13] specified that big data is not only characterized by the three V's mentioned above but may also extend to four V's, namely, Volume, Variety, Velocity, and Value (Figure 1 and 2). This 4V definition is widely recognized because it highlights the meaning and necessity of big data.

The following definition is proposed based on the above-mentioned definitions and our observation and analysis of the essence of big data. Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale.

1) ***Volume*** refers to the amount of all types of data generated from different sources and continues to expand. The benefit of gathering large amounts of data includes the creation of hidden information and patterns through data analysis. Laurila et al. [14] provided a unique collection of longitudinal data from smart mobile devices and made this collection available to the research community. The aforesaid initiative is called mobile data challenge motivated by Nokia [14]. Collecting longitudinal data requires considerable effort and underlying investments. Nevertheless, such mobile data challenge produced an interesting result similar to that in the examination of the predictability of human behavior patterns or means to share data based on human mobility and visualization techniques for complex data.

2) ***Variety*** refers to the different types of data collected via sensors, smart phones, or social networks. Such data types include video, image, text, audio, and data logs, in either structured or unstructured format. Most of the data generated from mobile applications are in unstructured format. For example, text messages, online games, blogs, and social media generate different types of unstructured data through mobile devices and sensors. Internet users also generate an extremely diverse set of structured and unstructured data [15].

3) ***Velocity*** refers to the speed of data transfer. The contents of data constantly change because of the absorption of complementary data collections, introduction of previously archived data or legacy collections, and streamed data arriving from multiple sources [12].

4) ***Value*** is the most important aspect of big data; it refers to the process of discovering huge hidden values from large datasets with various types and rapid generation [16].

CLASSIFICATION OF BIG DATA

Big data are classified into different categories to better understand their characteristics. Figure 2 shows the numerous categories of big data. The classification is important because of large-scale data in the cloud. The classification is based on five aspects: data sources, content format, data stores, data staging, and data processing.

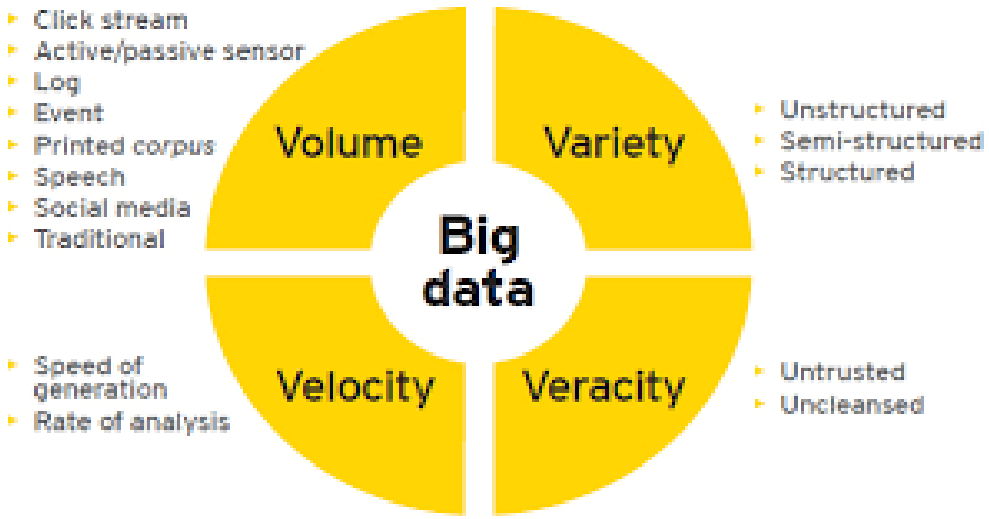

Figure 1: Four V's of big data

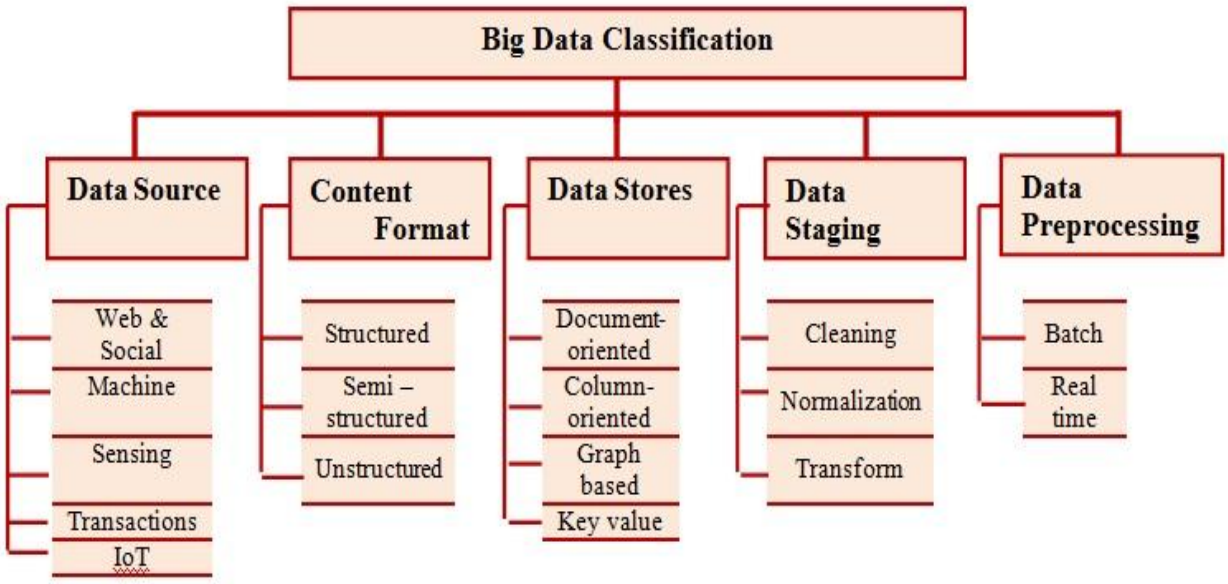

Figure 2: Big data classification

Each of these categories has its own characteristics and complexities as described in Table 2. Data sources include internet data, sensing and all stores of transnational information, ranges from unstructured to highly structured are stored in various formats. Most popular is the relational database that comes in a large number of varieties [17]. As the result of the wide variety of data sources, the captured data differ in size with respect to redundancy, consistency and noise, etc.

Table 1: Various categories of Big Data

Data Sources	
Social media	Social media is the source of information generated via URL to share or exchange information and ideas in virtual communities and networks, such as collaborative projects, blogs and microblogs, Facebook, and Twitter.
Machine-generated data	Machine data are information automatically generated from a hardware or software, such as computers, medical devices, or other machines, without human intervention.
Sensing	Several sensing devices exist to measure physical quantities and change them into signals.
Transactions	Transaction data, such as financial and work data, comprise an event that involves a time dimension to describe the data.
IoT	IoT represents a set of objects that are uniquely identifiable as a part of the Internet. These objects include smart phones, digital cameras, and tablets. When these devices connect with one another over the Internet, they enable more smart processes and services that support basic, economic, environmental, and health needs. A large number of devices connected to the Internet provides many types of services and produces huge amounts of data and information [54].
Content Format	
Structured	Structured data are often managed SQL, a programming language created for managing and querying data in RDBMS. Structured data are easy to input, query, store, and analyze. Examples of structured data include numbers, words, and dates.
Semi-structured	Semi-structured data are data that do not follow a conventional database system. Semi-structured data may be in the form of structured data that are not organized in relational database models, such as tables. Capturing semi-structured data for analysis is different from capturing a fixed file format. Therefore, capturing semi-structured data requires the use of complex rules that dynamically decide the next process after capturing the data [55].

Table 1: Various categories of Big Data

Data Sources	
Unstructured	Unstructured data, such as text messages, location information, videos, and social media data, are data that do not follow a specified format. Considering that the size of this type of data continues to increase through the use of smart phones, the need to analyze and understand such data has become a challenge.
Data Stores	
Document-oriented	Document-oriented data stores are mainly designed to store and retrieve collections of documents or information and support complex data forms in several standard formats, such as JSON, XML, and binary forms (e.g., PDF and MS Word). A document oriented data store is similar to a record or row in a relational database but is more flexible and can retrieve documents based on their contents (e.g., MongoDB, SimpleDB, and CouchDB).
Column-oriented	A column-oriented database stores its content in columns aside from rows, with attribute values belonging to the same column stored contiguously. Column-oriented is different from classical database systems that store entire rows one after the other, such as BigTable [56].
Graph database	A graph database, such as Neo4j, is designed to store and represent data that utilize a graph model with nodes, edges, and properties related to one another through relations [57].
Key-value	Key-value is an alternative relational database system that stores and accesses data designed to scale to a very large size [58]. Dynamo [59] is a good example of a highly available key-value storage system; it is used by amazon.com in some of its services. Similarly, [60] proposed a scalable key-value store to support transactional multi-key access using a single key access supported by key-value for use in G-store designs. A scalable clustering method to perform a large task in datasets. Other examples of key-value stores are Apache Hbase [61], Apache Cassandra [62], and Voldemort. Hbase uses HDFS, an open-source version of Google's BigTable built on Cassandra. Hbase stores data into tables, rows, and cells. Rows are sorted by row key, and each cell in a table is specified by a row key, a column key, and a version, with the content contained as an un-interpreted array of bytes.
Data Staging	
Cleaning	Cleaning is the process of identifying incomplete and unreasonable data [63].
Transform	Transform is the process of transforming data into a form suitable for analysis.
Normalization	Normalization is the method of structuring database schema to minimize redundancy [64].
Data Processing	
Batch	Map Reduce-based systems have been adopted by many organizations in the past few years for long-running batch jobs [65]. Such system allows for the scaling of applications across large clusters of machines comprising thousands of nodes.
Real time	One of the most famous and powerful real time process-based big data tools is simple scalable streaming system (S4) [66].S4 is a distributed computing platform that allows programmers to conveniently develop applications for processing continuous unbounded streams of data. S4 is a scalable, partially fault tolerant, general purpose, and pluggable platform.

BIG DATA ANALYTICS

Nowadays, the data that need to be analyzed are not just large, but they are composed of various data types, and even including streaming data [18]. Since big data has the unique features of "massive, high dimensional, heterogeneous, complex, unstructured, incomplete, noisy, and erroneous," which may change the statistical and data analysis approaches [19]. Although it seems that big data makes it possible for us to collect more data to find more useful information, the truth is that more data do not necessarily mean more useful information. It may contain more ambiguous or abnormal data. For instance, a user may have multiple accounts, or an account may be used by multiple users, which may degrade the accuracy of the mining results [20]. Therefore, several new issues for data analytics come up, such as privacy, security, storage, fault tolerance, and quality of data [21].

The big data may be created by handheld device, social network, and internet of things, multimedia, and many other new applications that all have the characteristics of volume, velocity, and variety. As a result, the whole data analytics has

to be re-examined from the following perspectives:

- From the volume perspective, the deluge of input data is the very first thing that we need to face because it may paralyze the data analytics. Different from traditional data analytics, for the wireless sensor network data analysis, Baraniuk [22] pointed out that the bottleneck of big data analytics will be shifted from sensor to processing, communications, storage of sensing data. This is because sensors can gather much more data, but when uploading such large data to upper layer system, it may create bottlenecks everywhere.
- In addition, from the velocity perspective, real-time or streaming data bring up the problem of large quantity of data coming into the data analytics within a short duration but the device and system may not be able to handle these input data. This situation is similar to that of the network flow analysis for which we typically cannot mirror and analyze everything we can gather.
- From the variety perspective, because the incoming data may use different types or have incomplete data, how to handle them also bring up another issue for the input operators of data analytics.

DATA ANALYTICS

To make the whole process of Knowledge Discovery in Databases (KDD) more clear, Fayyad and his colleagues summarized the KDD process by a few operations in [23], which are selection, preprocessing, transformation, data mining, and interpretation/evaluation. As shown in figure 3, with these operators at hand we will be able to build a complete data analytics system to gather data first and then find information from the data and display the knowledge to the user. According to our observation, the number of research articles and technical reports that focus on data mining is typically more than the number focusing on other operators, but it does not mean that the other operators of KDD are unimportant.

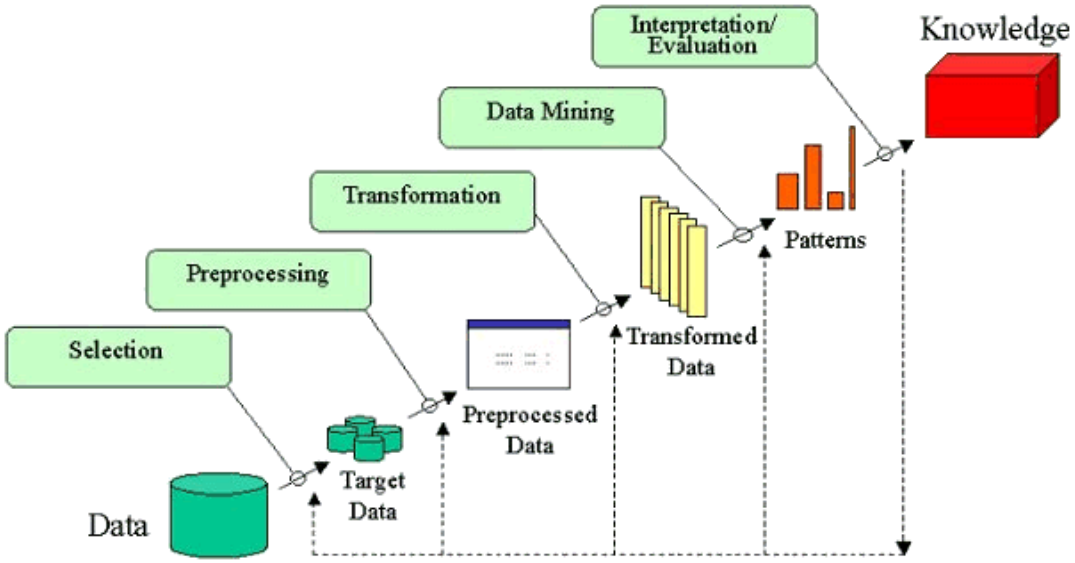

Figure 3: The process of knowledge discovery in databases

Data Input

As shown in figure 3, the gathering, selection, preprocessing, and transformation operators are in the input part. The selection operator usually plays the role of knowing which kind of data was required for data analysis and selects the relevant information from the gathered data or databases; thus, these gathered data from different data resources will need to be integrated to the target data. The preprocessing operator plays a different role in dealing with the input data which is aimed at detecting, cleaning, and filtering the unnecessary, inconsistent, and incomplete data to make them the useful data. After the selection and preprocessing operators, the characteristics of the secondary data still may be in a number of different data formats; therefore, the KDD process needs to transform them into a data-mining-capable format which is performed by the transformation operator. The methods for reducing the complexity and downsizing the data scale to make the data useful for data analysis part are usually employed in the transformation, such as dimensional reduction, sampling, coding, or transformation. The data extraction, data cleaning, data integration, data transformation, and data reduction operators can be regarded as the preprocessing processes of data analysis [24] which attempts to extract useful data from the raw data (also called the primary data) and refine them so that they can be used by the following data analyses. If the data are a duplicate copy, incomplete, inconsistent, noisy, or outliers, then these operators have to clean them up. If the data are too complex or too large to be handled, these operators will also try to reduce them. If the raw data have errors or omissions, the roles of these operators are to identify them and make them consistent.

Data Analysis

Since the data analysis (as shown in figure 3) in KDD is responsible for finding the hidden patterns/rules/information from the data, most researchers in this field use the term data mining to describe how they refine the "ground" (i.e., raw data) into "gold nugget" (i.e., information or knowledge). The data mining methods [24] are not limited to data problem specific methods. In fact, other technologies (e.g., statistical or machine learning technologies) have also been used to analyze the data for many years. In the early stages of data analysis, the statistical methods were used for analyzing the data to help us understand the situation we are facing, such as public opinion poll or TV programme rating. Like the statistical analysis, the problem specific methods for data mining also attempted to understand the meaning from the collected data.

After the data mining problem was presented, some of the domain specific algorithms are also developed. An example is the Apriori algorithm [25] which is one of the useful algorithms designed for the association rules problem. Although most definitions of data mining problems are simple, the computation costs are quite high. To speed up the response time of a data mining operator, machine learning [26], metaheuristic algorithms [27], and distributed computing [28] were used alone or combined with the traditional data mining algorithms to provide more efficient ways for solving the data mining problem.

1: ***Input*** Data D
2: ***Initialize*** candidate solutions r
3: **While** the termination criterion is not met
4: *d=Scan (D)*
5: *v=Construct (d,r,o)*
6: *r= Update (v)*
7: **End**
8: ***Output*** rules r

Figure 4: Data mining algorithm

As figure 4 shows, most data mining algorithms contain the initialization, data input and output, data scan, rules construction, and rules update operators [29]. In figure 4, D represents the raw data, d the data from the scan operator, r the rules, o the predefined measurement, and v the candidate rules. The scan, construct, and update operators will be performed repeatedly until the termination criterion is met. The timing to employ the scan operator depends on the design of the data mining algorithm; thus, it can be considered as an optional operator. Most of the data algorithms can be described by figure 4 in which it also shows that the representative algorithms—clustering, classification, association rules, and sequential patterns—will apply these operators to find the hidden information from the raw data. Thus, modifying these operators will be one of the possible ways for enhancing the performance of the data analysis.

BIG DATA ANALYSIS FRAMEWORKS AND PLATFORMS

Various solutions have been presented for the big data analytics which can be divided [30] into (1) Processing/Compute: Hadoop [31], Nvidia CUDA [32], or Twitter Storm [33], (2) Storage: Titan or HDFS, and (3) Analytics: MLPACK [34] or Mahout [53]. Although there exist commercial products for data analysis [34], most of the studies on the traditional data analysis are focused on the design and development of efficient and/or effective "ways" to find the useful things from the data. But when we enter the age of big data, most of the current computer systems will not be able to handle the whole dataset all at once; thus, how to design a good data analytics framework or platform3 and how to design analysis methods are both important things for the data analysis process.

BIG DATA ANALYSIS ALGORITHMS

Because the big data issues have appeared for nearly 10 years, in [35], Fan and Bifet pointed out that the terms "big data" [36] and "big data mining" [37] were first presented in 1998, respectively. The big data and big data mining almost appearing at the same time explained that finding something from big data will be one of the major tasks in this research domain. Data mining algorithms for data analysis also play the vital role in the big data analysis, in terms of the computation cost, memory requirement, and accuracy of the end results. In this section, we will give a brief discussion from the perspective of analysis and search algorithms to explain its importance for big data analytics.

Clustering algorithms In the big data age, traditional clustering algorithms will become even more limited than before because they typically require that all the data be in the same format and be loaded into the same machine so as to find some useful things from the whole data. Although the problem [38] of analyzing large-scale and high-dimensional dataset has attracted many researchers from various disciplines in the last century, and several solutions [39] have been presented in recent years, the characteristics of big data still brought up several new challenges for the data clustering issues. Among them, how to reduce the data complexity is one of the important issues for big data clustering. In [40], Shirkhorshidi et al. divided the big data clustering into two categories: single-machine clustering (i.e., sampling and dimension reduction solutions), and multiple-machine clustering (parallel and Map Reduce solutions). This means that traditional reduction solutions can also be used in the big data age because the complexity and memory space needed for the process of data analysis will be decreased by using sampling and dimension reduction methods. More precisely, sampling can be regarded as reducing the "amount of data" entered into a data analyzing process while dimension reduction can be regarded as "downsizing the whole dataset" because irrelevant dimensions will be discarded before the data analyzing process is carried out.

CloudVista [41] is a representative solution for clustering big data which used cloud computing to perform the clustering

process in parallel. BIRCH [42] and sampling method were used in CloudVista to show that it is able to handle large-scale data, e.g., 25 million census records. Using GPU to enhance the performance of a clustering algorithm is another promising solution for big data mining. The multiple species flocking (MSF) [43] was applied to the CUDA platform from NVIDIA to reduce the computation time of clustering algorithm in [44]. The simulation results show that the speedup factor can be increased from 30 up to 60 by using GPU for data clustering. Since most traditional clustering algorithms (e.g., k-means) require a computation that is centralized, how to make them capable of handling big data clustering problems is the major concern of Feldman et al. [45] who use a tree construction for generating the core sets in parallel which is called the "merge-and-reduce" approach. Moreover, Feldman et al. pointed out that by using this solution for clustering, the update time per datum and memory of the traditional clustering algorithms can be significantly reduced.

Classification algorithms Similar to the clustering algorithm for big data mining, several studies also attempted to modify the traditional classification algorithms to make them work on a parallel computing environment or to develop new classification algorithms which work naturally on a parallel computing environment. In [46], the design of classification algorithm took into account the input data that are gathered by distributed data sources and they will be processed by a heterogeneous set of learners. In this study, Tekin et al. presented a novel classification algorithm called "classify or send for classification" (CoS). They assumed that each learner can be used to process the input data in two different ways in a distributed data classification system. One is to perform a classification function by itself while the other is to forward the input data to another learner to have them labeled. The information will be exchanged between different learners. In brief, this kind of solutions can be regarded as a cooperative learning to improve the accuracy in solving the big data classification problem. An interesting solution uses the quantum computing to reduce the memory space and computing cost of a classification algorithm. For example, in [47], Rebentrost et al. presented a quantum based support vector machine for big data classification and argued that the classification algorithm they proposed can be implemented with a time complexity O(log NM) where N is the number of dimensions and M is the number of training data. There are bright prospects for big data mining by using quantum-based search algorithm when the hardware of quantum computing has become mature.

Frequent pattern mining algorithms Most of the researches on frequent pattern mining (i.e., association rules and sequential pattern mining) were focused on handling large-scale dataset at the very beginning because some early approaches of them were attempted to analyze the data from the transaction data of large shopping mall. Because the number of transactions usually is more than "tens of thousands", the issues about how to handle the large scale data were studied for several years, such as FP-tree [48] using the tree structure to include the frequent patterns to further reduce the computation time of association rule mining. In addition to the traditional frequent pattern mining algorithms, of course, parallel computing and cloud computing technologies have also attracted researchers in this research domain. Among them, the map-reduce solution was used for the studies [49] to enhance the performance of the frequent pattern mining algorithm. By using the map-reduce model for frequent pattern mining algorithm, it can be easily expected that its application to "cloud platform" [50, 51] will definitely become a popular trend in the forthcoming future. The study of [52] not only used the map-reduce model, it also allowed users to express their specific interest constraints in the process of frequent pattern mining. The performance of these methods by using map-reduce model for big data analysis is, no doubt, better than the traditional frequent pattern mining algorithms running on a single machine.

OPEN ISSUES

Although the data analytics today may be inefficient for big data caused by the environment, devices, systems, and even problems that are quite different from traditional mining problems, because several characteristics of big data also exist in the traditional data analytics. Several open issues caused by the big data will be addressed as the platform/framework and data mining perspectives in this section to explain what dilemmas we may confront because of big data. Here are some of the open issues:

Input and Output Ratio of Platform

A large number of reports and researches mentioned that we will enter the big data age in the near future. Some of them insinuated to us that these fruitful results of big data will lead us to a whole new world where "everything" is possible; therefore, the big data analytics will be an omniscient and omnipotent system. From the pragmatic perspective, the big data analytics is indeed useful and has many possibilities which can help us more accurately understand the so-called "things." However, the situation in most studies of big data analytics is that they argued that the results of big data are valuable, but the business models of most big data analytics are not clear. The fact is that assuming we have infinite computing resources for big data analytics is a thoroughly impracticable plan, the input and output ratio (e.g., return on investment) will need to be taken into account before an organization constructs the big data analytics center.

Communication Between Systems

Since most big data analytics systems will be designed for parallel computing, and they typically will work on other systems (e.g., cloud platform) or work with other systems (e.g., search engine or knowledge base), the communication between the big data analytics and other systems will strongly impact the performance of the whole process of KDD. The first research issue for the communication is that the communication cost will incur between systems of data analytics. How to reduce the communication cost will be the

very first thing that the data scientists need to care. Another research issue for the communication is how the big data analytics communicates with other systems. The consistency of data between different systems, modules, and operators is also an important open issue on the communication between systems. Because the communication will appear more frequently between systems of big data analytics, how to reduce the cost of communication and how to make the communication between these systems as reliable as possible will be the two important open issues for big data analytics.

Bottlenecks on Data Analytics System

The bottlenecks will be appeared in different places of the data analytics for big data because the environments, systems, and input data have changed which are different from the traditional data analytics. The data deluge of big data will fill up the "input" system of data analytics, and it will also increase the computation load of the data "analysis" system. This situation is just like the torrent of water (i.e., data deluge) rushed down the mountain (i.e., data analytics), how to split it and how to avoid it flowing into a narrow place (e.g., the operator is not able to handle the input data) will be the most important things to avoid the bottlenecks in data analytics system. One of the current solutions to the avoidance of bottlenecks on a data analytics system is to add more computation resources while the other is to split the analysis works to different computation nodes. A complete consideration for the whole data analytics to avoid the bottlenecks of that kind of analytics system is still needed for big data.

Security Issues

Since much more environment data and human behavior will be gathered to the big data analytics, how to protect them will also be an open issue because without a security way to handle the collected data, the big data analytics cannot be a reliable system. In spite of the security that we have to tighten for big data analytics before it can gather more data from everywhere, the fact is that until now, there are still not many studies focusing on the security issues of the big data analytics. According to our observation, the security issues of big data analytics can be divided into fourfold: input, data analysis, output, and communication with other systems. For the input, it can be regarded as the data gathering which is relevant to the sensor, the handheld devices, and even the devices of internet of things. One of the important security issues on the input part of big data analytics is to make sure that the sensors will not be compromised by the attacks. For the analysis and input, it can be regarded as the security problem of such a system. For communication with other system, the security problem is on the communications between big data analytics and other external systems. Because of these latent problems, security has become one of the open issues of big data analytics.

CHALLENGES IN BIG DATA

i. Heterogeneity and Incompleteness

When humans consume information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step in (or prior to) data analysis. Computer systems work most efficiently if they can store multiple items that are all identical in size and structure. Efficient representation, access, and analysis of semi-structured data require further work.

ii. Scale

Of course, the first thing anyone thinks of with big data is its size. After all, the word "big" is there in the very name. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by processors getting faster, following Moore's law, to provide us with the resources needed to cope with increasing volumes of data. But, there is a fundamental shift underway now: data volume is scaling faster than compute resources, and CPU speeds are static.

iii. ***Timeliness***

The flip side of size is speed. The larger the data set to be processed, the longer it will take to analyze. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster. However, it is not just this speed that is usually meant when one speaks of Velocity in the context of big data. Rather, there is an acquisition rate challenge

iv. ***Privacy***

The privacy of data is another huge concern, and one that increases in the context of big data. For electronic health records, there are strict laws governing what can and cannot be done. For other data, regulations, particularly in the US, are less forceful. However, there is great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources. Managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of big data.

v. ***Human Collaboration***

In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a hard time finding. Ideally, analytics for big data will not be all computational rather it will be designed explicitly to have a human in the loop. The new sub-field of visual analytics is attempting to do this, at least with respect to the modeling and analysis phase in the pipeline. In today's complex world, it often takes multiple experts from different domains to really understand what is going on. A big data analysis system must support input from multiple human experts, and shared exploration of results. These multiple experts may be separated in space and time when it is too expensive to assemble an entire team

together in one room. The data system has to accept this distributed expert input, and support their collaboration.

CONCLUSION

Big data analytics is trying to take advantage of the excess of information to use it productively. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. It must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of big data.

REFERENCES

[1]. Lyman P, Varian H. How much information 2003? Tech. Rep, 2004. [Online]. http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/ printable_ report. pdf.

[2]. Xu R, Wunsch D. Clustering. Hoboken: Wiley-IEEE Press; 2009.

[3]. Ding C, He X. K-means clustering via principal component analysis. In: Proceedings of the twenty-first international conference on machine learning; 2004. pp. 1–9.

[4]. Kollios G, Gunopulos D, Koudas N, Berchtold S. Efficient biased sampling for approximate clustering and outlier detection in large data sets. IEEE Trans Knowl Data Eng. 2003;15 (5):1170–87.

[5]. Fisher D, DeLine R, Czerwinski M, Drucker S. Interactions with big data analytics. Interactions. 2012;19(3):50–9.

[6]. Laney D. 3D data management: controlling data volume, velocity, and variety. META Group, Tech. Rep. 2001. [Online]. http://blogs.gartner.com/doug-laney/files/2012/01/ad9493D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[7]. van Rijmenam M. Why the 3v's are not sufficient to describe big data. BigData Startups, Tech. Rep. 2013. [Online]. http://www.bigdata-startups.com/3vs-sufficient-describe-big-data/ .

[8]. Borne K. Top 10 big data challenges a serious look at 10 big data v's. Tech. Rep. 2014. [Online]. https://www.mapr.com/blog/top-10-big-data-challenges-look-10-big-data-v.

[9]. M. Cox, D. Ellsworth, Managing Big Data For Scientific Visualization, ACM Siggraph, MRJ/NASA Ames Research Center, 1997.

[10]. J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A.H.Byers, Big data: The next frontier for innovation, competition, and productivity, (2011).

[11]. P. Zikopoulos, K. Parasuraman, T. Deutsch, J. Giles, D. Corrigan, Harness the Power of Big Data The IBM Big Data Platform, McGraw Hill Professional, 2012.

[12]. J.J. Berman, Introduction, in: Principles of Big Data, Morgan Kaufmann, Boston, 2013, xix–xxvi (pp).

[13]. J. Gantz, D. Reinsel, Extracting value from chaos, IDC iView (2011) 1–12.

[14]. J.K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, M. Miettinen, The mobile data challenge: Big data for mobile computing research, Workshop on the Nokia Mobile Data Challenge, in: Proceedings of the Conjunction with the 10th International Conference on Pervasive Computing, 2012, pp. 1–8.

[15]. D.E. O'Leary, Artificial intelligence and big data, IEEE Intell. Syst. 28 (2013) 96–99.

[16]. M. Chen, S. Mao, Y. Liu, Big data: a survey, Mob. Netw. Appl. 19 (2) (2014) 1–39.

[17]. J. Hurwitz, A. Nugent, F. Halper, M. Kaufman, Big data for dummies, For Dummies (2013).

[18]. Russom P. Big data analytics. TDWI: Tech. Rep; 2011.

[19]. Ma C, Zhang HH, Wang X. Machine learning for big data analytics in plants. Trends Plant Sci. 2014;19(12):798–808.

[20]. Boyd D, Crawford K. Critical questions for big data. Inform Commun Soc. 2012;15(5):662– 79.

[21]. Katal A, Wazid M, Goudar R. Big data: issues, challenges, tools and good practices. In: Proceedings of the international conference on contemporary computing; 2013. pp. 404–409.

[22]. Baraniuk RG. More is less: signal processing and the data deluge. Science. 2011;331 (6018):717–9.

[23]. Fayyad UM, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag. 1996;17(3):37–54.

[24]. Han J. Data mining: concepts and techniques. San Francisco: Morgan Kaufmann Publishers Inc.; 2005.

[25]. Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. Proc ACM SIGMOD Int Conf Manag Data. 1993;22(2):207–16.

[26]. Witten IH, Frank E. Data mining: practical machine learning tools and techniques. San Francisco: Morgan Kaufmann Publishers Inc.; 2005.

[27]. Abbass H, Newton C, Sarker R. Data mining: a

heuristic approach. Hershey: IGI Global; 2002.

[28]. Cannataro M, Congiusta A, Pugliese A, Talia D, Trunfio P. Distributed data mining on grids: services, tools, and applications. IEEE Trans Syst Man Cyber Part B Cyber. 2004;34 (6):2451–65.

[29]. Tsai C-W, Lai C-F, Chiang M-C, Yang L. Data mining for internet of things: a survey. IEEE Commun Surv Tutor. 2014;16(1):77–97.

[30]. Pospiech M, Felden C. Big data—a state-of-the-art. In: Proceedings of the Americas conference on information systems; 2012. pp. 1–23. [Online]. http://aisel.aisnet.org/ amcis2012/proceedings/DecisionSupport/22.

[31]. Apache Hadoop, February 2, 2015. [Online]. http://hadoop.apache.org.

[32]. Cuda, February 2, 2015. [Online]. http://www.nvidia.com/object /cuda_home_new . html.

[33]. Apache Storm, February 2, 2015. [Online]. http://storm.apache.org/.

[34]. Curtin RR, Cline JR, Slagle NP, March WB, Ram P, Mehta NA, Gray AG. MLPACK: a scalable C++ machine learning library. J Mach Learn Res. 2013;14:801–5.

[35]. Fan W, Bifet A. Mining big data: current status, and forecast to the future. ACM SIGKDD Explor Newslett. 2013;14(2):1–5.

[36]. Diebold FX. On the origin(s) and development of the term "big data". Penn Institute for Economic Research, Department of Economics, University of Pennsylvania, Tech. Rep. 2012. [Online]. http://economics.sas. upenn.edu/sites/economics.sas.upenn. edu/files/12-037.pdf.

[37]. Weiss SM, Indurkhya N. Predictive data mining: a practical guide. San Francisco: Morgan Kaufmann Publishers Inc.; 1998.

[38]. Chiang M-C, Tsai C-W, Yang C-S. A time-efficient pattern reduction algorithm for k-means clustering. Inform Sci. 2011;181(4):716–31.

[39]. Fahad A, Alshatri N, Tari Z, Alamri A, Khalil I, Zomaya A, Foufou S, Bouras A. A survey of clustering algorithms for big data: taxonomy and empirical analysis. IEEE Trans Emerg Topics Comp. 2014;2(3):267–79.

[40]. Shirkhorshidi AS, Aghabozorgi SR, Teh YW, Herawan T. Big data clustering: a review. In: Proceedings of the international conference on computational science and its applications; 2014. pp. 707–720.

[41]. Xu H, Li Z, Guo S, Chen K. Cloudvista: interactive and economical visual cluster analysis for big data in the cloud. Proc VLDB Endow. 2012;5(12):1886–9.

[42]. Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases. In: Proceedings of the ACM SIGMOD international conference on management of data; 1996. pp. 103–114.

[43]. Cui X, Gao J, Potok TE. A flocking based algorithm for document clustering analysis. J Syst Archit. 2006;52(89):505–15.

[44]. Cui X, Charles JS, Potok T. GPU enhanced parallel computing for large scale data clustering. Future Gener Comp Syst. 2013;29(7):1736–41.

[45]. Feldman D, Schmidt M, Sohler C. Turning big data into tiny data: constant-size coresets for k-means, pca and projective clustering. In: Proceedings of the ACM-SIAM symposium on discrete algorithms; 2013. pp. 1434–1453.

[46]. Tekin C, van der Schaar M. Distributed online big data classification using context information. In: Proceedings of the Allerton conference on communication, control, and computing; 2013. pp. 1435–1442.

[47]. Rebentrost P, Mohseni M, Lloyd S. Quantum support vector machine for big feature and big data classification. CoRR, vol. abs/1307.0471; 2014. [Online]. http://dblp.uni-trier.de/db/ journals/corr/corr1307.html#RebentrostML13.

[48]. Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In: Proceedings of the ACM SIGMOD international conference on management of data; 2000. pp. 1–12.

[49]. Lin MY, Lee PY, Hsueh SC. Apriori-based frequent itemset mining algorithms on mapreduce. In: Proceedings of the international conference on ubiquitous information management and communication; 2012. pp. 76:1–76:8.

[50]. Yang L, Shi Z, Xu L, Liang F, Kirsh I. DH-TRIE frequent pattern mining on hadoop using JPA. In: Proceedings of the international conference on granular computing; 2011. pp. 875– 878.

[51]. Huang JW, Lin SC, Chen MS. DPSP: Distributed progressive sequential pattern mining on the cloud. In: Proceedings of the advances in knowledge discovery and data mining, vol. 6119; 2010. pp. 27–34.

[52]. Leung CS, MacKinnon R, Jiang F. Reducing the search space for big data mining for interesting patterns from uncertain data. In: Proceedings of the international congress on big data; 2014. pp. 315–322.

[53]. Apache Mahout, February 2, 2015. [Online]. http://mahout.apache.org/.

[54]. B.P. Rao, P. Saluia, N. Sharma, A. Mittal, S.V. Sharma, Cloud computing for Internet of Things & sensing based applications, In Proceedings of the

Sensing Technology (ICST), 2012 SixthbInternational Conference on, IEEE, 2012, pp. 374–380.

[55]. B. Franks, Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics, Wiley. com John Wiley Sons Inc, 2012.

[56]. F. Chang, J. Dean, S. Ghemawat, W.C. Hsieh, D.A. Wallach, M. Burrows, T. Chandra, A. Fikes, R.E. Gruber, Bigtable: a distributed storage system for structured data, ACM Trans. Comput. Syst. (TOCS) 26 (2008) 4.

[57]. P. Neubauer, Graph databases, NOSQL and Neo4j, in, 2010.

[58]. M. Seeger, S. Ultra-Large-Sites, Key-Value stores: a practical over-view, Comput. Sci. Media (2009).

[59]. G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, W. Vogels, Dynamo: amazon's highly available key-value store, SOSP 41 (6) (2007) 205–220.

[60]. S. Das, D. Agrawal, A. El Abbadi, G-store: a scalable data store for transactional multi key access in the cloud, in: Proceedingsof the 1st ACM symposium on Cloud computing, ACM, 2010, pp. 163–174.

[61]. R.C. Taylor, An overview of the Hadoop/MapReduce/Hbase framework and its current applications in bioinformatics, BMC Bioinf. 11 (2010) S1.

[62]. A. Lakshman, P. Malik, The Apache cassandra project, in, 2011.

[63]. E. Rahm, H.H. Do, Data cleaning: problems and current approaches, IEEE Data Eng. Bull. 23 (2000) 3–13.

[64]. J. Quackenbush, Microarray data normalization and transformation, Nat. Genet. 32 (2002) 496–501.

[65]. Y. Chen, S. Alspaugh, R. Katz, Interactive analytical processing in big data systems: a cross-industry study of MapReduce workloads, Proc. VLDB Endow. 5 (2012) 1802–1813.

[66]. L. Neumeyer, B. Robbins, A. Nair, A. Kesari, S4: Distributed Stream Computing Platform, Data Mining Workshops (ICDMW), 2010 IEEE International Conference on, 2010, pp. 170–177.