2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)

A Survey on Deep Learning in Big Data

Mehdi Gheisari†, Guojun Wang*†, Md Zakirul Alam Bhuiyan†‡
†School of Computer Science and Educational Software, Guangzhou University, Guangzhou, China, 510006
‡Department of Computer and Information Sciences, Fordham University, New York, NY, 10458
*Correspondence to: csgjwang@gzhu.edu.cn

***Abstract*—Big Data means extremely huge large data sets that can be analyzed to find patterns, trends. One technique that can be used for data analysis so that able to help us find abstract patterns in Big Data is Deep Learning. If we apply Deep Learning to Big Data, we can find unknown and useful patterns that were impossible so far. With the help of Deep Learning, AI is getting smart. There is a hypothesis in this regard, the more data, the more abstract knowledge. So a handy survey of Big Data, Deep Learning and its application in Big Data is necessary. In this paper, we provide a comprehensive survey on what is Big Data, comparing methods, its research problems, and trends. Then a survey of Deep Learning, its methods, comparison of frameworks, and algorithms is presented. And at last, application of Deep Learning in Big Data, its challenges, open research problems and future trends are presented.**

***Index Terms*—Big Data; Deep learning; Deep Learning Challenges; Machine Learning; Deep Learning Methods; Big Data Challenges.**

I. Introduction

If we want to have a glance at the data generation history from 1960, we can see this trend in overall: 1960-1990, relational databases; 1990-2000, OLAP technology; 2000- 2010, column based data storages and cloud computing; and 2010-2016, Big Data applications. These days, Knowledge plays a key role to get success. Many companies need more abstract knowledge. This need can be satisfied by a combination of two major domains: Big Data and Deep Learning. Each device can generate data. This situation will become worse if each device can be connected to other devices to use their information. In other words, with the emergence of Internet of Things, we are facing with huge amount of data that needed to be stored and managed, one Example of Big Data. in brief, with the advances in digital devices such as digital sensors, large amounts of data have been generated at a fast speed that resulted in an area named Big Data. Big Data is not only about producing data from sensors; It can be provided by humans, texts, images and so on. Big Data has a great impact on technologies and computing. In other words, we have more data these days that current methods cannot deal with these data. In simple word, the term of Big Data means collecting, processing and presenting the results of huge amounts of data that come at high speed in a variety of formats. Traditional Machine Learning tools have shortcoming when they face with Big Data and want to solve Big Data area problems [1]. The following figure shows comparisons of ML techniques and their drawbacks.

For example, we can apply Deep Learning that is a tool for understanding higher abstract knowledge in most steps of Big Data area problems. But preferably it needs high volumes of data. If we want to become more successful in this competitive area, we need to find abstract patterns. The more pattern, the more success. In this regard, we need to know the application of Deep Learning in Big Data, how to use it that is the aim of this paper. Authors contribution is:

1) A handy introduction to Big Data and comparing its methods.
2) A handy introduction to Deep Learning, comparing algorithms and frameworks.
3) The authors describe the application of Deep Learning in Big Data.

Section 2 presents Big Data steps, challenges, and future trends. Section 3 shows Deep Learning and Machine Learning tools, frameworks. Section 4 describes the application of Deep Learning in Big Data, future trends, and open research problems. Conclusion and future work will be presented in section 5.

II. Big Data

The rise of Big Data has been caused by increase of data storage capability, increase of computational power, and more data volume accessibility. Most of the current technologies that are used to handle Big Data challenges are focusing on six main issues of that called Volume, Velocity, Variety, Veracity, Validity, and Volatility. The first one is Volume that means we are facing with huge amounts of data that most of traditional algorithms are not able to deal with this challenge. For example, Each minute 15h of videos are uploaded to Facebook so that collects more than 50 TB per day. With respect to the amounts of data generating each day, we can predict the growth rate of data in next years [2]. The data growth is 40 percent per year. Each year around 1.2 ZB data are produced. Huge companies such as Twitter, Facebook, and Yahoo have recently begun tapping into large volume data benefits. The definition of high volume is not specified in predefined term and it is a relative measure depends on the current situation of the enterprise [3]. The second challenge is Variety that in brief means we are facing with variety types of file formats and even unstructured ones such as PDFs, emails, audios and so on. These data should be unified for further processes [4]. The third V is Velocity that means data are coming in a very fast manner, the rate at which data are coming is striking, that may hang the system easily. It shows the need for real-time algorithms. The next two Vs (Veracity and Validity) have major similarities with each other, mean

978-1-5386-3221-5/17 $31.00 © 2017 IEEE
DOI 10.1109/CSE-EUC.2017.215

Algorithm	Type	Class	Restriction bias	Preference bias
K-Nearest Neighbor	Supervised learning	Instance based	Generally speaking, KNN is good for measuring distance-based approximations, but it suffers from the curse of dimensionality	Prefers problems that are distance based
Naive Bayesian Classification	Supervised learning	Probabilistic	Works on problems where the inputs are independent from each other	Prefers problems where the probability will always be greater than zero for each class
Hidden Markov Models	Supervised/ unsupervised	Markovian	Generally works well for system information where the Markov assumption holds	Prefers time series data and memory-less information
Support Vector Machine	Supervised learning	Decision boundary	Works where there is a definite distinction between two classifications	Prefers binary classification problems
Neural Networks	Supervised learning	Nonlinear functional approximation	Has little restriction bias	Prefers binary inputs
Clustering	Unsupervised	Clustering	No restriction	Prefers data that is in groupings given some form of distance (Euclidean, Manhattan, or others)
(Kernel) Ridge Regression	Supervised	Regression	Has low restriction on problems it can solve	Prefers continuous variables
Filtering	Unsupervised	Feature transformation	No restriction	Prefer data to have lots of variables on which to filter

Fig. 1. comparison between machine learning techniques

data must be as clean, trustworthy, usefulness, result data should be valid, as possible for later processing phases. The more data sources and types, the more difficult sustaining trust [5]. And the last V is the Volatility that means how much time data should remain in the system so that they are useful for the system. McKinsey added Value as the seventh V that means the amount of hidden knowledge inside Big Data [6]. We also can consider open research problems from another viewpoint as follows, six parameters: Availability, Scalability, Integrity, Heterogeneity, Resource Optimization, and Velocity (related to stream processing). Labrinidis and Jagadish in [7] described some challenges and research problems with respect to Scalability, Heterogeneity aspects of Big Data management. Other parameters such as availability and integrity are covered in [8]. These parameters are defined as follows: -Availability: Means data should be accessible and available whenever and wherever user requests data even in the case of failure occurrence. Data analysis methods should provide availability to support large amounts of data along with a high-speed stream of data [9].

-Scalability: refers if a system supports large amounts of increasing data efficiently or not. Scalability is an important issue mostly from 2011 for industrial applications to scale well in limited memory.

-Data Integrity: points to data accuracy. The situation becomes worse when different users with different privileges change data in the cloud. Cloud is in charge of managing databases. Therefore, users have to obey cloud policy for data integrity [10].

-Heterogeneity: refers to different types of data such as structured, unstructured and semi-structured [11].

-Resource Optimization: means using existing resources efficiently. A precise policy for resource optimization is needed for guaranteeing distributed access to Big Data.

-Velocity: means the speed of data creation and data analysis. The increased amount of digital devices like smart phones, tablets caused the increase of speed of data generation. Thus, the need for real-time analyses is obligatory. These are very application dependent that means can differ for each application to another application. And from steps point of view, Big Data area can be divided into three main Phases: **Big Data preprocessing**, means doing some preliminary actions toward data with the aim of data preparation such as data cleansing and so on. **Big Data storage** means how data should be stored. **Big Data management** means how we should manage data in order to get best achievement such as clustering, classification and so on [12].

A. Preprocessing

For better decision-making, we should provide quality data for data analyzing step. In other words, the quality of data is critical to quality decision. We should also verify data before decision. Preprocessing data means transforming, inconsistency, incomplete data that have many errors into an appropriate format for further analyses. In other words, data must be structured prior to analysis stage [13]. For example, in one database we may have STUEDNTID and in the other, we may have Student Identifier. It prepares data for further processing and analysis. There are some steps for achieving preprocessing section goal as described as follows:

1. Data cleansing: Removing inaccuracies, incompleteness, and inconsistencies of data.

2. Data transformation: Means doing additional processes like aggregation, or transformation. This step has a striking influence on future steps.

3. Data integration: It provides a single view over distributed data from different sources.

4. Data transmission: Defines a method for transferring raw data to storage system such as object storage, data center or distributed cloud storage.

5. Data reduction: reducing the size of large databases for real-time applications [14].

6. Data discretization: It is a notable step for decision tree learning process. It refers to attribute intervals so that obtained values will be reduced [10] .

The following sub-sections present more detail about some preprocessing steps:

1) Data Transmission: Data transmission is one step of preprocessing phase. It means sending raw data to data storage. One example of proposed method in this area is sending data through a high-capacity pipe from data source to data center. This type of transmission needs to know networks architecture along with transportation protocol.

2) Data Cleansing: In simple word means detecting incomplete, and irrational data. We can modify or delete these kinds of data in order to achieve quality improvement for further processing steps. Maletic and Marcus took into consideration five stages in order to achieve clean data: 1) recognizing types of errors 2) finding error instances 3) correct error instances and error types 4) update data input procedure in order to reduce further errors that may occur 5) checking data affairs like limitations, formats, and rationalities. Data cleansing is an indispensable and principal part of data analysis step. In brief, there are two main problems in data cleansing step: i)Data are imprecise ii) Data are incomplete(there are missing parts in the dataset) and we should address these problems as much as we can.

3) Stream Processing: Processing of stream data is a challenge that researchers have faced in Big Data area. The stream requirements are completely different with traditional batch processing. In more detail, there are some emerging applications producing large amounts of dedicated data to servers in order to real-time processing. One example is stock trading that we should use real-time processing in order to achieve an enhanced decision. While large volumes of data are received by servers for processing, we are not able to use traditional centralized techniques. There are some applications in this regard called Distributed Stream Processing Systems (DSPS) [15]. But most of the people use traditional centralized databases in order to analyze such huge amounts of data due to lack of tools. As mentioned earlier, there are many open research topics in the stream processing part that are described as follows:

1-Data Mobility: It means that the number of steps that are required to get the final result.

2-Data Division or Partitioning: The algorithms are used for partitioning data. In the brief, partitioning strategies should be used in order to achieve better data parallelism.

3-Data Availability: We should propose a technique that guarantees data availability in case of failures occurrence.

4-Query Processing: We should propose a query processor for distributed data processing efficiently with considering data streams. One possibility of this is doing deterministic processing (always get the same answer) and another one is non-deterministic (the output depends on the current situation) one.

5-Data Storage: Another open research problem in Big Data is how to store data for future usage.

6-Stream Imperfections: Techniques dealing with data stream imperfections like delayed messages or out-of-order messages.

B. Data Storage

Storing data in petabyte scale is a challenge not only for researchers but also for internet organizations. These days we can hardly adapt existing databases to Big Data usage. Although Cloud Computing reveals a shift to a new computing paradigm, it cannot assure consistency easily when storing Big Data in cloud storage. It is not a good way to waste data since it may contribute to better decision-making. So it is critical to have a storage management system in order to provide enough data storage, and optimized information retrieval [13].

1) Replication: Replication is a big activity that makes data available and accessible whenever user asks. When data are variable, the accuracy of each replicated copy is much more challenging. The two factors that we should consider in replication are replication sites and consistency. These two factors play more important role in Big Data environment as managing these huge amounts of data are more difficult than usual form [16].

2) Indexing: For large databases, it is not wise to retrieve stored data and searching data in sequential form like an unordered array [17]. Indexing data improves the performance of storage manager. So proposing a suitable indexing mechanism is challenging. There are three challenges in indexing area 1) multi-variable and multi-site searching 2) performing different types of queries 3) data search when they are numerical. Authors in [18] proposed a new method for keyword searching in data stream environment. It uses a tree based index structure and adopts sharing of a single list of event indices to speed up query responses. Index load time is a challenge now same as space consumption [19]. A Support Vector Machine indexing algorithm was introduced in [20] for video data with the aim of modeling human behavior. It changes transition probability calculation mechanism and applies different states to determine the score of input data. While it produces a relatively accurate query result with minimum time, it is time-consuming in learning process. A fuzzy-based method can be used for indexing of moving objects where indexing images are captured during object's movements. It provides a trade-off between query response time and index regeneration. The index supports data variables and it is scalable. And as experiments show it has better performance than the previous algorithms of other moving index techniques.

C. Big Data Management and Processing

There are four types of data models we have faced in Big Data area: 1- data that we can store them in relational 2- semi-structured data same as XML 3- graph data such as those we use for social media and the last one is unstructured data such as text data, hand-written articles [21]. One important question here is why are not we able to use traditional databases such as

Relational Databases in Big Data? One of the basic answers is that most of the relational databases are not designed to scale to thousands of loosely coupled machines [22]. Because of two reasons, companies tended to leave traditional databases: the first one is traditional databases are not scalable and the second one is that it is very expensive if we want to use non-distributed traditional databases along with adding layers on top. So companies decided to implement their own file system (HDFS), distributed storage systems(Google Bigtable [23]), distributed programming frameworks (MapReduce), and even distributed database management systems(Apache Cassandra) [24]. Furthermore, Big Data management is a complex process especially when data are gathered from heterogeneous sources to be used for decision-making and scoping out a strategy. Authors in [25] noted that about 75 percent of organizations apply at least one form of Big Data. Big Data management area brought new challenges in terms of data fusion complexity, storage of data, analytical tools and shortage of governance. We also can categorize Big Data management processes into two main categories as authors [26] reported (1) Big Data science and (2) Big Data infrastructure. Science means studying techniques regarding data acquisition, conditioning, and evaluation. Its infrastructure is focused on the improvement of existing technologies same as managing, analyzing, visualizing of data [27].

Table 1 describes well-known methods with regard to storage, pre-processing and processing steps of Big Data and compare them from six features as described above.

For example, in an application that heterogeneity is not as important as velocity, we can use SOHAC algorithm, the first row of the table, as part of our method.

1) Classification and Clustering:

a) Classification: Unstructured data will be stored in a distributed database such as SimpleDB, Cassandra, or Hbase. After storing data, these data are processed by using data mining techniques. Mathematical methods can also be involved in analysis step such as classification, classifying objects into different predefined groups, using decision trees, statistics, Linear programming, Neural Networks [28].

b) Clustering: Another method is Clustering that means creating groups of objects based on their meaningful attributes so that large amounts of data sets are able to be represented in a few data sets, summarizing gathered data into groups where data with similar features are near to each other. It reduces the needs for high storage resources by accommodating large amounts of data in limited storage space that is still a challenging work [29]. One proposed solution for this challenge is Storage-Optimizing Hierarchical Agglomerative Clustering (SOHAC)that is proposed by [30]. They introduced a new storage structure that requires less storage space than usual time. Basic type of this algorithm was previously. But it was limited in computing high-dimensional data. A single matrix contained data is decomposed into sub-matrices. Based on the features, the new matrices reduce value in each row [27]. Useful data will be recorded since redundant data will be neglected in order to save more space. The algorithm takes into consideration hierarchical agglomerative strategy. In abstract level, at first, it defines clusters for each object. Then clusters will be merged with each other with the aim of forming K clusters as defined in the initialization of process [31].

III. Machine Learning and Deep Learning

a) Machine Learning: In general, we have two types of learning: 1- Shallow learning(Machine Learning, learning without explicitly programming) such as decision trees, Support Vector Machines(SVMs) that it is likely to fall short when we want to extract useful information from huge amounts of data and even if they would not fall short, they will not have satisfied accuracy.

An important question here is with all the different algorithms in the ML, how can we choose the best one for our purpose? If we want to predict or forecast a target value, then we should use supervised learning techniques such as Neural Networks (NN) that we know the correct answers previously. In other words, supervised learning problems are categorized into ”regression” and ”classification” problems. In a regression problem, we are trying to predict output of continuous values, meaning that we are trying to map input variables to some continuous functions. In a classification problem, we are instead trying to predict results into discrete outputs [32]. In other words, we are trying to map input variables into discrete categories [33]. In brief, First, we need to consider our goal. What are we trying to get out of that? (Do you want a probability that it might rain tomorrow, or you want to find groups of voters with similar interests?) What data do you have or can you collect? If you are trying to predict or forecast a target value, then you need to look into supervised learning. If not, then unsupervised learning is the place you want to be. If you have chosen supervised learning, what is your target value? Is it a discrete value like Yes/No, 1/2/3, A/B/C, or Red/Yellow/Black? If so, then you may look into classification. If the target value can take on a number of values, say any value from 0.00 to 100.00, or -999 to 999, or + to -, then you need to look into regression. If you are not trying to predict a target value, then you need to look into unsupervised learning. Are you trying to fit your data into some discrete groups? If so and that is all you need, you should look into clustering. Do you need to have some numerical estimate of how strong the fit is into each group? If you answer yes, then you probably should look into a density estimation algorithm [4].

b) Deep Learning: We need new insight from data, not only for top-level executives but also can be used for providing better services to customers. One tool for reaching this aim is Deep Learning (DL). Deep Learning is a promising avenue of research into automated complex feature extraction at a high level of abstraction. Deep Learning is about learning multiple levels of representations and abstractions that help to make sense of data such as images, sound, and text. One of the unique characteristics of deep learning algorithms is its ability to utilize unlabeled data during training [34]. We are able to discover intermediate or abstract representations which

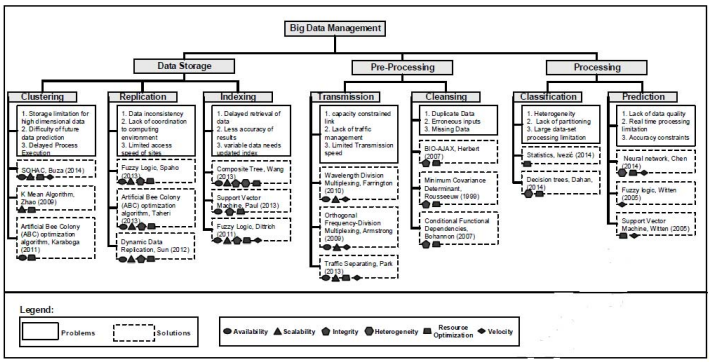

Fig. 2. Categorizing Big Data management problems and current researches

TABLE I
A COMPARISON OF BIG DATA METHODS

BDM Activity		Relative Method	Authors	Availability	Scalability	Integrity	Heterogeneity	Resource Optimization	Velocity
Data Storage	Clustering	Storage-Optimizing Hierarchical Agglomerative Clustering- SOHAC	Buza, Nagy et al. (2014)	Yes	Yes	N/A	No	Yes	Yes
		K-Mean	Zhao, Ma et al. (2009)	No	Yes	N/A	No	Yes	No
		Artificial Bee Colony (ABC) optimization algorithm	Karaboga and Ozturk (2011)	Yes	No	N/A	No	Yes	No
	Replication	Fuzzy Logic	Spaho, Barolli et al. (2013)	Yes	Yes	Yes	No	Yes	No
		Artificial Bee Colony (ABC) optimization algorithm	Taheri, Choon Lee et al. (2013)	Yes	Yes	Yes	No	Yes	No
		Dynamic Data Replication	Sun, Chang et al. (2012)	Yes	Yes	Yes	No	Yes	No
	Indexing	Composite Tree	Wang, Holub et al. (2013)	Yes	Yes	Yes	Yes	Yes	Yes
		Support Vector Machine	Paul, Chen et al. (2013)	Yes	No	Yes	No	Yes	No
		Fuzzy Logic	Ditrich, Blunschi et al. (2011)	Yes	Yes	Yes	No	Yes	Yes
Pre-Processing	Transmission	Wavelength Division Multiplexing (WDM)	Farrington, Porter et al. (2010)	Yes	Yes	N/A	N/A	N/A	Yes
		Orthogonal Frequency-Division Multiplexing (OFDM)	Armstrong (2009)	Yes	Yes	N/A	N/A	N/A	Yes
		Traffic Separating	Park, Yeo et al. (2013)	Yes	Yes	N/A	N/A	Yes	Yes
	Cleansing	BIO-AJAX	Herbert and Wang (2007)	N/A	N/A	Yes	No	Yes	N/A
		Minimum Covariance Determinant (MCD)	Rousseeuw and Driessen (1999)	N/A	N/A	Yes	No	Yes	N/A
		Conditional Functional Dependencies (CFD)	Bohannon, Fan et al. (2007)	N/A	N/A	Yes	No	Yes	N/A
Process ing	Classification	Statistics	Ivezić, Connolly et al. (2014)	N/A	N/A	N/A	No	Yes	No
		Decision trees	Dahan, Cohen et al. (2014)	N/A	N/A	N/A	Yes	Yes	No
	Prediction	Neural network	Chen, Xu et al. (2014)	N/A	N/A	N/A	Yes	Yes	Yes
		Fuzzy logic	Witten and Frank (2005)	N/A	N/A	N/A	No	No	Yes
		Support vector machine	Steinwart and Christmann (2008)	N/A	N/A	N/A	No	Yes	Yes

are carried out using unsupervised learning in a hierarchical fashion, One level each time then higher-level features are defined based on lower-level features. It can improve classification modeling results and it has a major capability for generalization of learning. One example of DL is extracting invariant features of a person from an image. In simple word, it produces more sensible knowledge from raw data and it is called our observation variety. It generally learns data features in a greedy layer-wise manner. In addition, It implements a layered, hierarchical architecture of learning leads to richer data representation. It stacks up non-linear feature extractors with the aim of getting better machine learning results such as a better classification model, invariant property of data representation. It has outstanding result in a variety of applications like speech recognition, computer vision, and Natural language processing (NLP), debate winner prediction in elections based on public opinion, enables the ability to analyze and predict traffic jams faster in congestion, finding a new mechanism that effects complex traffic systems. Most of traditional machine learning algorithms cannot extract non-linear patterns. DL generates learning patterns and also generates relationships beyond neighbors. DL not only provides complex representations of data, but it also makes machines independent from human [35]. It extracts useful information (representation, features) from unsupervised data without human intervention. In simple word, DL consists of consecutive layers that each layer provides a local abstract in its output. Each layer poses a nonlinear transformation on its input; we have a complicated abstract representation of data in the last layer output. The more layers data go through, the more complicated and abstract representation we get. The final representation is a high non-linear transformation of data. Dl does not try to extract predefined representations; in reverse, it tries to disentangle factors of variation in data to find invariant patterns. For learning compact representations, Deep Learning models are better than shallow learning models. The compact representations are efficient because they need less computation. It makes it possible to learn nonlinear representations of huge amounts of data [11].

IV. Application of Deep learning in Big Data

If we want to have a look of application of Deep Learning in Big Data, DL deals mainly with two V's of Big Data characteristics: Volume and Variety. It means that DL are suited for analyzing and extracting useful knowledge from both large huge amounts of data and data collected from different sources [36]. One example of application of Deep Learning in Big Data is Microsoft speech recognition (MAVIS) that is using DL enables searching of audios and video files through human voices and speeches [20] [37]. Another usage of DL on Big Data environment is used by Google company for Image search service. They used DL for understanding images so that can be used for image annotation and tagging that is useful for image search engines and image retrieval or even image indexing. When we want to apply Dl, we face some challenges that we need to address them same as:

1) Deep Learning for High Volumes of Data

1.1. The first one is whether we should use all entire Big Data input or not. In general, we apply DL algorithms in a portion of available Big Data for training goal and we use the rest of data for extracting abstract representations and from another point of view, question is that how much volume of data is needed for training data.

1.2. Another open problem is domain adaptation, in applications which training data is different from the distribution of test data. If we want to look at this problem from another viewpoint, we can point to how we can increase the generalization capacity of DL; generalizing learnt patterns where there is a change between input domain and target domain.

1.3. Another problem is defining criteria for allowing data representations to provide useful future semantic meanings. In simple word, each extracted data representation should not be allowed to provide useful meaning. We must have some criteria to obtain better data representations.

1.4. Another one is that most of the DL algorithms need a specified loss and we should know what is our aim to extract, sometimes it is very difficult to understand them in the Big Data environment.

1.5. The other problem is that most of them do not provide analytical results that can be understandable easily. In other words, because of its complexity, you cannot analyze the procedure easily. This situation becomes worse in a Big Data environment.

1.6. Deep Learning seems suitable for the integration of heterogeneous data with multiple modalities due to its capability of learning abstract representations.

1.7. The last but not the least major problem is that they need labeled data. If we can not provide labeled data, they will have bad performance. One possible solution for this is that we can use reinforcement learning, the system gathers data by itself, and the only need for us is giving rewards to the system.

2) Deep Learning for High Variety of Data

These days, data come in all types of formats from a variety sources, probably with different distributions. For example, the rapidly growing multimedia data coming from the web and mobile devices include a huge collection of images, videos and audio streams, graphics and animations, and unstructured text, each with different characteristics. There are open questions in this regard that need to be addressed as some of them presented as follows:

2.1. given that different sources may offer conflicting information, how can we resolve the conflicts and fuse the data from different sources effectively and efficiently?

2.2. if the system performance benefits from significantly enlarged modalities?

2.3. in which level deep learning architectures are appro-

priate for feature fusion of heterogeneous data?

3) Deep Learning for High Velocity of Data
   Data are generating at extremely high speed and need to be processed at fast speed. One solution for learning from such high-velocity data is online learning approaches that can be done by deep learning. Only limited progress in online deep learning has been made in recent years. There are some challenges in this matter such as:

3.1. Data are often non-stationary, data distributions are changing during the time.

3.2. the big question is whether we can benefit from Big Data along with deep architectures for transfer learning or not.

V. Future Trends and Open Research Problems

Some future research topics may be categorized as following:

A. Big Data Preprocessing

One challenge is data integrity that means sharing data among users efficiently. Even though, data integration definition is not much clear in most of the applications. For example, authors in tried to state that with the aim of using a single system for two different companies with different products, we need to find out how the combined data can operate in a single environment and integrated system. They believed that data integration is much harder than Artificial intelligence. So the two challenging research topics in this field are generating and facilitating integration tools. The quality of data is not predetermined. After using data, we are able to find its quality. The more quality data, the better results. Data providers demand error-free data and it is relatively impossible to use only one method of data cleaning to achieve the best quality data is a challenge. In order to achieve quality data, we must combine different cleansing methods to meet the organization's need. With the increasing speed of data volume and transformation, the reputation of collected data depends on the quality and availability of information they provided [38]. But traditional methods were proposed in order to provide equal access to resources. For instance, in the traditional era, network administrators should investigate network traffic. But with the emerging of Big Data, data analysts must analyze data not go through many details.

B. Big Data Analytics

It relates to database searching, mining, and analysis. With the usage of data mining in the big data area, a business can enhance its services. Big Data Mining is a challenge because of data complexity and scalability. The two salient goals of data analyses are: first detecting relationships between data features and second predicting future instances. In other words, it means searching in a vast area to offer guidelines to users. Steed, Ricchiuto proposed a new visual analytical system for the earth that analyze complex earth system simulation which named Exploratory Data Analysis Environment (EDAE) . Previously, data had always been analyzed by the trial-and-error methods which were very difficult in complicated situations, with vast amounts of data and data heterogeneity [39]. Authors in [40] discussed that obtaining useful information from large amounts of data need scalable algorithms. Additional applications and Cloud infrastructures are needed to deal with data parallelism. Algorithm orders increase exponentially with the increase of data size [41]. We have four type of analyses in simple words:

- Descriptive Analysis: What is happening in data now.
- Predictive Analysis: What will happen in the future.
- Discovery Analysis: Discovering an existing rule from existing data
- Perceptive Analysis: What should we do in future based on current data.

C. Semantic Indexing

Another usage of DL and open challenge is using it for semantic indexing with the aim of better information retrieval. It means we should store semantic indexes rather than storing as raw data bytes due to massive amounts of data and low storage capacity. DL generates high-level data representations. We can use this abstract data representation to provide better indexing method [42].

D. Data Governance

It is another important core of Big Data Management and it means defining rules, laws and controlling over data. One example is that if Big Data should be stored in the cloud, we must take some policies like which type of data needs to be stored, how quickly data should be accessed, rules for data such as transparency, integrity, check and balances, and last but not the least change management. There are many open research topics in this field like best decision-making mechanism, reduction of operational friction [43].

E. Big Data Integration

It means collecting data from multiple sources and storing them with the aim of providing a unified view. Integrating different types of data is a complex issue that can be even worse when we have different applications [39]. Many open research topics are associated with data integration like real-time data access, the performance of the system, and overlapping of the same data [44].

VI. Conclusion

Nowadays, it is necessary to grapple with Big Data with the aim of extracting better abstract knowledge. One technique that is applicable for this aim is Deep Learning (Hierarchical Learning) that provides higher-level data abstraction. Deep Learning is a useful technique that can also be used in the Big Data environment and has its own advantages and disadvantages. In general, the more data, the higher level abstract data, but we face many challenges. This paper surveys at first Big Data steps, then Machine learning and Deep Learning and at last application of Deep Learning in Big Data, future trends, and open research problems. In the future, we

have a plan to pay attention to above areas in more detail and also investigating Big Data problems in the industry. We are going to also have a survey on Big Data security and privacy issue. Then we want to address other problems such as semantic indexing, data tagging and so on.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant Numbers 61632009, 61472451, and 61402543 and in part by the High Level Talents Program of Higher Education in Guangdong Province under Grant 2016ZJ01.

REFERENCES

[1] Yoshua Bengio et al. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[2] Jinchuan Chen, Yueguo Chen, Xiaoyong Du, Cuiping Li, Jiaheng Lu, Suyun Zhao, and Xuan Zhou. Big data challenge: a data management perspective. *Frontiers of Computer Science*, 7(2):157–164, 2013.

[3] Dylan Maltby. Big data analytics. In *74th Annual Meeting of the Association for Information Science and Technology (ASIST)*, pages 1–6, 2011.

[4] Han Hu, Yonggang Wen, Tat-Seng Chua, and Xuelong Li. Toward scalable systems for big data analytics: A technology tutorial. *IEEE Access*, 2:652–687, 2014.

[5] Aisha Siddiqa, Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Mohsen Marjani, Shahabuddin Shamshirband, Abdullah Gani, and Fariza Nasaruddin. A survey of big data management: taxonomy and state-of-the-art. *Journal of Network and Computer Applications*, 71:151–166, 2016.

[6] Min Chen, Shiwen Mao, and Yunhao Liu. Big data: A survey. *Mobile Networks and Applications*, 19(2):171–209, 2014.

[7] Alexandros Labrinidis and Hosagrahar V Jagadish. Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, 5(12):2032–2033, 2012.

[8] Chang Liu, Chi Yang, Xuyun Zhang, and Jinjun Chen. External integrity verification for outsourced big data in cloud and iot: A big picture. *Future Generation Computer Systems*, 49:58–67, 2015.

[9] Katina Michael and Keith W Miller. Big data: New opportunities and new challenges [guest editors' introduction]. *Computer*, 46(6):22–24, 2013.

[10] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[11] Xue-Wen Chen and Xiaotong Lin. Big data deep learning: challenges and perspectives. *IEEE Access*, 2:514–525, 2014.

[12] CL Philip Chen and Chun-Yang Zhang. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, 275:314–347, 2014.

[13] Andrew McAfee, Erik Brynjolfsson, Thomas H Davenport, DJ Patil, and Dominic Barton. Big data. *The management revolution. Harvard Bus Rev*, 90(10):61–67, 2012.

[14] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[15] Anton Riabov and Zhen Liu. Scalable planning for distributed stream processing systems. In *ICAPS*, pages 31–41, 2006.

[16] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[17] Jens Dittrich, Lukas Blunschi, and Marcos Antonio Vaz Salles. Movies: indexing moving objects by shooting index images. *Geoinformatica*, 15(4):727–767, 2011.

[18] Guoliang Li, Beng Chin Ooi, Jianhua Feng, Jianyong Wang, and Lizhu Zhou. Ease: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 903–914. ACM, 2008.

[19] Li Deng, Dong Yu, et al. Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014.

[20] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008.

[21] Divyakant Agrawal, Amr El Abbadi, Shyam Antony, and Sudipto Das. Data management challenges in cloud computing infrastructures. In *International Workshop on Databases in Networked Information Systems*, pages 1–10. Springer, 2010.

[22] GNU Octave. Gnu octave. *línea]. Available: http://www. gnu. org/software/octave*, 2012.

[23] Xiao Chen. Google big table. *2010*.

[24] Da-Wei Sun, Gui-Ran Chang, Shang Gao, Li-Zhong Jin, and Xing-Wei Wang. Modeling a dynamic data replication strategy to increase system availability in cloud computing environments. *Journal of computer science and technology*, 27(2):256–272, 2012.

[25] Daniel E O'Leary. Artificial intelligence and big data. *IEEE Intelligent Systems*, 28(2):96–99, 2013.

[26] Vivien Marx. Biology: The big challenges of big data. *Nature*, 498(7453):255–260, 2013.

[27] P. Porkar. Sensor networks challenges. In *11th international conference on data networks, DNCOCO '12,*, 7-9 September 2012.

[28] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 513–520, 2011.

[29] Shahabi Amir. *clustering algorithm in Wireless Sensor network*, chapter Sustainable Interdependent Networks: From Theory to Application. Springer, accepted for publication (2018).

[30] Krisztian Buza, Gábor I. Nagy, and Alexandros Nanopoulos. Storage-optimizing clustering algorithms for high-dimensional tick data. *Expert Syst. Appl.*, 41:4148–4157, 2014.

[31] Mehdi Jafari, Jing Wang, Yongrui Qin, Mehdi Gheisari, Amir Shahab Shahabi, and Xiaohui Tao. Automatic text summarization using fuzzy inference. In *Automation and Computing (ICAC), 2016 22nd International Conference on*, pages 256–260. IEEE, 2016.

[32] Dervis Karaboga and Celal Ozturk. A novel clustering approach: Artificial bee colony (abc) algorithm. *Applied soft computing*, 11(1):652–657, 2011.

[33] Mehdi Gheisari, Ali Akbar Movassagh, Yongrui Qin, Jianming Yong, Xiaohui Tao, Ji Zhang, and Haifeng Shen. Nsssd: A new semantic hierarchical storage for sensor data. In *Computer Supported Cooperative Work in Design (CSCWD), 2016 IEEE 20th International Conference on*, pages 174–179. IEEE, 2016.

[34] Surajit Chaudhuri, Umeshwar Dayal, and Vivek Narasayya. An overview of business intelligence technology. *Communications of the ACM*, 54(8):88–98, 2011.

[35] T. Tran, M. Rahman, M. Z. A. Bhuiyan, A. Kubota, S. Kiyomoto, and K. Omote. Optimizing share size in efficient and robust secret sharing scheme for big data. *IEEE Transactions on Big Data*, PP(99):1–1, 2017.

[36] Richard S Sutton and Andrew G Barto. *Introduction to reinforcement learning*, volume 135. MIT Press Cambridge, 1998.

[37] Steve Lohr. The age of big data. *New York Times*, 11(2012), 2012.

[38] M. Z. A. Bhuiyan and J. Wu. Event detection through differential pattern mining in cyber-physical systems. Jun 2017.

[39] W. Yu, J. Li, M. Z. A. Bhuiyan, R. Zhang, and J. Huai. Ring: Real-time emerging anomaly monitoring system over text streams. *IEEE Transactions on Big Data*, PP(99):1–1, 2017.

[40] Katherine G Herbert and Jason TL Wang. Biological data cleaning: a case study. *International Journal of Information Quality*, 1(1):60–82, 2007.

[41] Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald, and Edin Muharemagic. Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1):1, 2015.

[42] Todd A Letsche and Michael W Berry. Large-scale information retrieval with latent semantic indexing. *Information sciences*, 100(1-4):105–137, 1997.

[43] Vijay Khatri and Carol V Brown. Designing data governance. *Communications of the ACM*, 53(1):148–152, 2010.

[44] Xin Luna Dong and Divesh Srivastava. Big data integration. In *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*, pages 1245–1248. IEEE, 2013.