J Supercomput (2016) 72:1494–1516
DOI 10.1007/s11227-016-1677-z

Handling big data: research challenges and future directions

I. Anagnostopoulos[1] · S. Zeadally[2] · E. Exposito[3]

Published online: 25 February 2016

Abstract Today, an enormous amount of data is being continuously generated in all walks of life by all kinds of devices and systems every day. A significant portion of such data is being captured, stored, aggregated and analyzed in a systematic way without losing its "4V" (i.e., volume, velocity, variety, and veracity) characteristics. We review major drivers of big data today as well the recent trends and established platforms that offer valuable perspectives on the information stored in large and heterogeneous data sets. Then, we present a classification of some of the most important challenges when handling big data. Based on this classification, we recommend solutions that could address the identified challenges, and in addition we highlight cross-disciplinary research directions that need further investigation in the future.

Keywords Big data · Data curation · Data cleansing · Data analytics · Privacy · Trust

1 Introduction

The accelerated growth and pervasive development of Web, Internet and Cloud technologies are enabling the global interconnection of heterogeneous information systems and social networks, as well as the proliferation of Internet of Things (IoT) and web of objects embedded with advanced data capture technologies. These technological

✉ S. Zeadally
szeadally@uky.edu

[1] University of Thessaly, Nea Ionia, Greece

[2] University of Kentucky, Lexington, USA

[3] University of Toulouse, Toulouse, France

advances have led to the generation of tremendous structured and unstructured data from a wide range and multiple data sources that cannot be processed through conventional analytics tools. Today, research activities and industry efforts in different disciplines are exploring various efficient and intelligent techniques that provide the best interpretation of this large volume of data being generated from different types of heterogeneous sources.

Many definitions have been proposed for “Big Data”. In [1] Jacobs provided a meta-definition of big data which refers to “data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time”. The author in [2] also adopted a generic definition of big data which refers to “data that’s too big, too fast, or too hard for existing tools to process”. Both definitions remain abstract which leave the academic and the business communities to focus on more specific definitions related to data characteristics and processes. For instance, Wu et al. [3] proposed the *HACE* Theorem which defines big data as “large-volume, **h**eterogeneous, **a**utonomous sources with distributed and decentralized control, and seeks to explore **c**omplex and **e**volving relationships among data”. McKinsey Global Institute, a global consulting agency, defines the big data as “datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze”. International Data Cooperation (IDC) shares the same viewpoint and highlights the “4V” characteristic of big data by defining it 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 high-velocity capture, discovery, and/or analysis” [4].

According to the IDC,[1] the size of data is doubling size each year and it will reach about 44 zettabytes by 2020. For example, data from embedded systems will grow from 2 % in 2013 to 10 % in 2020. Big data emphasizes not only the huge volume of data, but also its diversity and the speed at which it must be managed as well as its correctness. Along with the tremendous growth of the digital universe, a large number of sources churn out heterogeneous data types (numeric, images, video, text, etc.) with different representation forms. Data can be stored in relational databases in a structured format or generated by connected devices and social networks in a heterogeneous format. Information systems, sensors, and devices are being designed by diverse manufacturers, each one defining their own algorithms and data representations. This heterogeneity hinders data analytics and processing. Moreover, the size of data is growing rapidly. Data streaming often needs to be processed in real time for various applications such as healthcare, e-commerce and finance, or offline by applying statistical and machine learning algorithms. These algorithms need to support data discovery capabilities on individual sources, as well as specialized and complex analytics results by aggregating different sources of data. From the huge volumes of data, we can extract valuable data and useful information that can provide the basis for intelligent services and assist many decision-making systems.

[1] IDC: http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm.

1.1 Drivers of big data

The main drivers of big data can be classified into four broad categories: technical-oriented, performance-oriented, productivity-oriented and cost-oriented.

From the technical perspective, the heterogeneity and the large volume of data represent the main information technology (IT) challenges for data mining and processing [5], especially for IoT data which is mostly unstructured. Traditional systems are based on Relational Databases Management Systems (RDBMS), which do not support unstructured data. RDBMS are known by their static schema model and their limitations to handle changing data structures. Initially, they were designed for a single server and have later been changed to accommodate grid/cloud computing environments, but they are inadequate for big data storage [6]. Big data technologies provide novel approaches, methods and techniques that exceed the storage and analysis capability of current or conventional systems and deal with large and complex data [7]. From the performance viewpoint, the complexity of RDBMS operations slows down applications' response time especially in cases where real-time data processing is required (e.g., to prevent or predict situations).

In the past, the focus was on small data for business intelligence and prediction, but today we have a deluge of data everywhere. The ability to correlate more data allows us to discover new and better information. From the huge volume of various types of data, we may predict the future, generate valuable hidden information and deduce preventive actions, which could increase productivity. For example, major manufacturers, employ sensor data directly from their production lines, thus creating self-tuning actions that will avoid time-consuming human interventions. Data are analyzed by clusters of high-performance computers, and their results are used in order to increase productivity based on time-sensitive information, environmental conditions or other external parameters that affect the production line. Similarly, an effective use of big data brings value to consumers and reduces service costs. Large retail companies are able to identify the digital footprint of thousands of individual customers, directly from their web and online social media behavior, and model their market behavior in real time. According to McKinsey [8], the potential value from data in the healthcare sector could reach more than \$300 billion in value every year. They also estimated that big data could save more than €100 billion in the developed economies of Europe.

1.2 Big data applications

Big data has changed the way that we manipulate data. It is being leveraged in many disciplines such as healthcare, space science, criminal justice and other domains with powerful outcomes. In healthcare, recently, at the University of Southern California in Los Angeles, computer scientists and medical experts collaborating together have created algorithms that use data generated by various sensors including body sensors to better treat Parkinson's disease [9]. By this way, medical experts can evaluate the treatment efficiency and notify patients. The University of Arizona and National Taiwan University collaborated and came up with big data platform for diabetes named

DiabeticLink[2] for patient empowerment and personalization [10]. DiabeticLink is an interactive platform that allows exchanging disease information and experience. It aggregates multiple sources of data such as forums, drug side effects, electronic health records and other sources. Another impressive use case in healthcare is the use of big data in genomics to discover novel genes and help on personalizing the treatment [11]. In this context, Cloudera have teamed up with the Institute for Genomics and Multiscale Biology, at the Mount Sinai School of Medicine, to assist researchers using big data technologies in order to predict and understand the process and treatment of diseases.

Getting more individualized data about humans, either from what they generated using the web and mobile device or from sensors such biometric sensors and deployed cameras, big data systems have proven efficiency for real-time prediction of urban crimes. For example, Microsoft and New York Police Department (NYPD) teamed up to fight crime by developing a big data surveillance system [12]. This system collects data from cameras, sensors on the streets, license plate detectors and law enforcement databases, aggregates these massive data sets and extracts relevant information to locate crimes which are taking place in real time. In the same discipline, the Department of Criminology at the University of Pennsylvania, in Philadelphia, built an algorithm that predicts who will be a victim of a homicide based on a variety of data, including reports from local police precincts [13]. Thus, police forces will be able to warn the victims, advise and protect them from the crime. From human data to data space, the National Aeronautics and Space Administration (NASA) [14] uses big data to collect climate data, predict the weather on earth, monitor ice caps on Mars and also search for distant galaxies. Moreover, the NASA offers a visualization tool that converts space data into images, graphs, videos that help astronomers and scientist to discover new space patterns.

2 Big data trends

Big data technologies need to support the “4V” characteristics and provide scalable data storage, distributed analytics and real-time processing and data visualization. This has motivated many vendors to develop powerful platforms that satisfy these requirements and offer new vision for data analytics to extract and identify hidden information based on the large data sets.

2.1 Big data platforms

Today, unstructured data, which accounts for more than 90 % of the digital universe [4], do not fit the strict relational model. For real-time applications, fast read and write operations are required. This has contributed to the development of Not only SQL (NoSQL) databases for managing and designing large and distributed databases [15]. Three data model approaches are often used in NoSQL: key values, document base and

[2] http://www.diabeticlink.org/.

column base [16]. These stores are mostly used by web applications such as Facebook which uses the Cassandra column store and Google with its BigTable store. Big data is not just about data volume, but is very strongly related to an important processing phase which is data analytics.

While the development of NoSQL databases provides scalable and distributed storage of data, new query tools and analytic platforms have been proposed to explore and smoothly process the large volumes of data currently being generated. The most popular platform is Hadoop,[3] an Apache open source platform. Hadoop has the potential to process and aggregate both structured and unstructured large-scale data stored in distributed server nodes using its Hadoop Distributed File System (HDFS). It implements the mapReduce framework for parallel data processing. Pentaho[4] is an example of big data analytic platform that uses Hadoop and NoSQL stores in business analytics. Yahoo! uses Hadoop running over 42,000 servers for searching and spam filtering [17]. Many vendors such as Cloudera, IBM, MapR, Hortonworks offer Hadoop as a cloud-based distribution platform for scalable data processing.

Other platforms have been developed on the top of Hadoop to operate more complex algorithms, such as data mining and machine learning, which detect and identify patterns able to discover and predict new information from the large-scale data. Mahout[5] is an open package dedicated to distributed and scalable data mining and machine learning. It supports big data analytics on the Hadoop platform [18]. It offers a set of clustering and classification algorithms such as Logistic Regression, Bayesian models, Support Vector Machines, and Random Forest among others. ML-Hadoop[6] is another package based on Hadoop that offers machine learning algorithms such as Markov model, multiple linear regression, k-mean clustering and Naïve Bayes classifier. Other projects were developed for expressing large data set analysis in MapReduce. These projects include Apache Hive[7] and Apache Pig.[8] Facebook initially developed hive for analyzing its generated data and for generating reports [19]. Yahoo! Research developed Pig Latin, a textual language for Apache Pig that offers a high-level programming language for encoding data analysis tasks in MapReduce programs [20].

Another important aspect of big data is the ability to support real-time data processing which conventional systems are inadequate to perform because of the large amounts of data involved. In this context, Apache Storm,[9] an open source distributed fault-tolerant real-time computation system, has been developed. It is a scalable system that allows processing of one million tuples per second per node. It can be used with any programming language. Storm is not the only Hadoop project for streaming data, Spark[10] is another Apache project for large-scale data processing which is one hun-

[3] http://hadoop.apache.org/.

[4] http://www.pentaho.com/product/big-data-analytics.

[5] https://mahout.apache.org/.

[6] https://code.google.com/p/ml-hadoop/.

[7] https://hive.apache.org/.

[8] http://pig.apache.org/.

[9] http://storm.apache.org/.

[10] https://spark.apache.org/.

Table 1 Hadoop-based platforms according to their main characteristics and their big data application domain

Platforms	Characteristics			
	Real-time analytics	Data integration	Open source	Application domain
Pentaho	$\surd$	$\surd$		Business analytics
Mahoot			$\surd$	Data mining, machine learning algorithms
ML-Hadoop			$\surd$	Data mining, machine learning algorithms
Hive		$\surd$	$\surd$	MapReduce-based large data set analysis
Pig		$\surd$	$\surd$	MapReduce-based large-scale parallel implementations
Storm	$\surd$		$\surd$	Real-time analytics, online machine learning, continuous computation, distributed extraction–transformation–loading (ETL) processes
Spark	$\surd$	$\surd$	$\surd$	Large-scale real-time data processing, scalable machine learning applications
Sqoop		$\surd$	$\surd$	Transfers bulk data between Apache Hadoop and structured datastores (RDBMS)
Flume		$\surd$	$\surd$	Efficient collection, aggregation, and transfers of large amounts of log data
Zookeeper		$\surd$	$\surd$	Reliable and efficient distributed coordination in distributed applications

The "$\surd$" sign highlights the provision of a specific characteristic

dred times faster than the Hadoop MapReduce in memory. Spark supports the MLlib[11] library for scalable machine learning. Other projects such as Sqoop[12] for importing relational databases such as MySQL to the HDFS, Flume[13] for moving large amounts of data log for analysis, ZooKeeper[14] for coordinating parallel processing with high-performance, are developed to leverage from Hadoop distributed processing. Table 1 summarizes the main characteristics of all the above platforms in terms of real-time analytics and data integration provision, as well as in terms of their big data application domain. Apart from Pehtaho, which is a proprietary business-oriented big data

[11] http://spark.apache.org/docs/latest/mllib-guide.html.

[12] http://sqoop.apache.org/.

[13] http://flume.apache.org/.

[14] http://zookeeper.apache.org/.

integration and analytics platform, all the other platforms are Apache Hadoop-based open source initiatives supported by their respective communities.

New big data platforms continue to emerge and are being developed to handle the various characteristics of the huge amounts of various types of data that are being continuously generated at high rates by multiple heterogeneous sources.

2.2 Big data analytics and visualization

Big data analytics have revolutionized basic data analytics by introducing new technologies that extend conventional data mining, statistical and machine learning methods to support parallel processing of distributed, heterogeneous and huge amounts of data [19,21]. Big data analytics contribute to predicting and extracting useful values from large data sets, and more importantly help to retrieve hidden information and accurate insights, to accelerate the decision-making. Social networks such as Facebook, LinkedIn and Twitter, involve millions of users who have contributed to the generation of big data. Exploring these data sets helps on predicting behaviors that allow enhancing marketing, sales, online commerce and user experience, and preventing frauds in different domains [22]. Twitter implements predictive analytics based on supervised machine learning algorithms in order to predict future events based on tweets and user profiles. To deal with large data sets, the Twitter analytic stack, which is based on a large Hadoop cluster, integrates different open source components such as HBase, Elephant Bird,[15] and more importantly Pig, which is extended to encode machine learning algorithms as scripts [23]. Based on the extended Pig, the authors of [23] proposed a sentiment analysis technique that can be applied to large number of tweets using optimization algorithms (i.e., stochastic gradient descent) in order to predict positive or negative opinions of tweets.

In the network domain, Liu et al. [24] proposed a Hadoop-based platform for network traffic analysis. The monitored cellular network data is stored in HDFS and HBase for managing both structured and unstructured data. Based on different indices such as IP addresses, transport-layer protocols, and TCP/IP five-tuple, the authors proposed different data mining algorithms encoded with Hadoop mapReduce tasks to reveal network traffic content and user behavior phenomena not shown before with the goal of improving the user experience. In the same network context and HTTP packet analysis, Marchal et al. [25] proposed the PhishStorm system that automatically detects phishing by performing real-time analytic of any URL based on phishing and legitimate data sets. PhishStorm is based on Storm, for real-time big data processing, and statistical Bloom filters.

From network traffic to plant science, Ma et al. [26] studied the integration of machine learning and data mining tools in a big data platform pertaining to plant sciences. The framework integrates different big data technologies such as Sqoop, Hive and Pig. It relies on Mahout and RHadoop for data analysis. More precisely, the author used the Random Forest (RF) algorithm based on trained data sets to discover genes

[15] https://github.com/twitter/elephant-bird/.

responsive to salt stress and recommend high-priority candidate genes for phenotypic screening experiments.

In healthcare, Chandola et al. [27] presented three case studies that prove the efficiency of big data analytics to identify fraudulent healthcare providers. The first use case, which detects frauds, relies on text mining operated over healthcare insurance claims that have been collected by various health insurance agencies, and over a list of fraudulent providers that have been sanctioned for fraudulent behavior in the State of Texas. The second use case, which prevents frauds, is based on analyzing data from social networks by using a predictive model to assess healthcare fraud risk at the time of enrollment. Finally, the third case implements a statistical model that focuses on analyzing time-ordered sequence of claims. The authors used Hadoop and Hive platforms, and Mahout was used for mining the large data sets and the R framework was used for implementing statistical models.

The main purpose of big data analytics and stream data processing is to retrieve more accurate results and help scientist and domain experts to discover and extract information not previously known. One of the best ways to achieve this objective is to provide visuals via graphs, images or diagrams by using different types of visualization and graphical tools. According to 3M cooperation's behavioral research, the human brain processes images 60,000 times faster than text [28]. The presentation of big data analytic results using interactive graphs accelerates the experts' interpretation and assists them in making the right decision(s). Some libraries have been developed for visualization and they include D3.js,[16] Polymaps,[17] NodeXL,[18] etc. Moebio Labs[19] developed advanced interactive visualization projects that can work with very large data sets. They developed the Newk[20] application to visualize twitter conversations. Chandola et al. [27] also provided big data visualization feature using Python Networkx[21] to understand the relationships of providers in the healthcare system and visualize patterns of fraudulent behaviors.

Reda et al. [29] proposed the hybrid-reality (HR) environments that support scalable visualization of complex and large heterogeneous amounts of data. The authors combined virtual reality and high-resolution tiled liquid crystal display (LCD) walls, to create hybrid 2D–3D information. CAVE2 is a real-world application of the HR environment. CAVE2 was used by genomic researchers to visualize hundreds and thousands of sequenced genomes. An example of a molecular-dynamics simulation of a glass fissure comprising approximately five million atoms was analyzed in this work.

As the amount of data keeps increasing, big data visualization will become even more challenging. It is not just about displaying the visuals representing the data but also allowing the user to take full advantage of the richness of the visual analytic

[16] https://github.com/mbostock/d3/wiki/Gallery.

[17] http://polymaps.org/.

[18] http://www.nodexlgraphgallery.org/Pages/Default.aspx.

[19] http://moebio.com/.

[20] http://moebio.com/newk/twitter/.

[21] https://networkx.github.io/.

results. Novel, high-performance visualization tools will need to be developed in the future to handle big data analytics [30].

3 Handling big data: challenges and issues

There are several challenges and issues that need to be addressed in the area of big data handling. These challenges should be addressed holistically drawing expertise from different computer science areas. Some of the big data handling challenges scientists should address include *Data Cleansing/Acquisition/Capture*, *Data Storage/Sharing/Transfer*, *Data Analysis*, as well as some *ethical issues* that arise from the exposure and processing of big data.

In an effort to classify the challenges from a technical point of view, the authors of [31] distinguish six major technical steps in conjunction with five challenges that need to be addressed during these steps. In this paper, we distinguish the challenges as well as their correlation, according to a bottom-up layer approach that corresponds to three fundamental (big) data handling processes, namely "Get", "Save", and "Analyze". We then classify the challenges in four types and we further highlight the problems that need to be addressed, which are either technical or not.

As depicted in Fig. 1, the challenges of the first layer ("Get") demand clean, integrated and re-usable data from various sources. Data curation and standardization when capturing data from multiple sources are the "keys" here. In the second layer ("Save"), challenges deal with solutions for the storage and migration of the data among different data warehouses (transferring/sharing issues). Enhanced network infrastructures are essential, yet the proper encounter of the first layer challenges (e.g., data cleansing) reduces the need for infrastructure resources. Finally, the challenges that belong to third layer ("Analyse") produce meaningful results to the big data stakeholders. Elastic software platforms for parallel and massive process are important in this layer.

The fourth and last type of challenges according to our classification has to do with the sensitive and human-centric dimension of handling big data. Ethical problems and related issues (personal information process, privacy violations, intellectual properties, etc.) appear in all three handling layers and they persist if not properly dealt from the

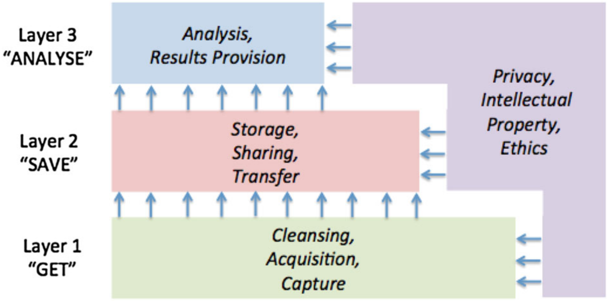

Fig. 1 Big data handling layers and their challenges according to our classification

lower layer (as shown Fig. 1). Further details about each challenge type and their distinct limitations are analyzed in the following.

3.1 Data cleansing, acquisition and capture

As with all data, big data needs to satisfy some fundamental quality criteria. These "quality" criteria mean that they have to be easy-to-reach, easy-to-identify and easy-to-read, by being in some uniform format for further processing. Data cleansing (also often called data scrubbing) deals with inconsistent data, data derived from heterogeneous sources, and data that are not up-to-date or valid over time. After cleansing, data will be cured (ready for contemporary use and available for discovery and reuse). Without cleansing there is significant "noise" when data are acquired. Even in cases where big data satisfies the quality criteria (clean and consistent data), some sources may generate a huge amount of data. Thus, another challenge is to develop filtering mechanisms capable of keeping useful information [32]. Furthermore, metadata schemes need to be employed for describing the "robust" part of the data that is left after the filtering process, thereby facilitating their reproducibility (data provenance which refers to the process of determining and storing the sources of the data [33]).

Data acquisition can be defined as a twofold procedure. First, it samples a physical phenomenon from its real-world conditions (sources), and second, it converts the resulting samples into finite digital numeric values as input to an information system. In the big data area, data sources vary from personal smartphone devices that produce "small data", as a result of targeted data acquisition and data mining, to extremely large sets of data from highly demanding infrastructures (e.g., for high-energy physics measurements, such as the Large Hadron Collider at CERN or as in the case of a large set of mobile data sources).

In the Internet of Things, we have highly heterogeneous data sources where well-defined data is acquired and captured from different sensors for various types of events. The diversity of events span from environmental conditions (e.g., temperature, humidity, light conditions, air pressure), object conditions (e.g., tire-pressure monitoring, obstacle warning in vehicles), to human-centric conditions (e.g., driving behavior that indicates fatigue, distraction, etc.).

Latency issues make data acquisition and capture even more challenging. In engineering, latency reflects the time a system needs to produce an output after an input stimulus. In a system that monitors and fuses data from several different sensors, latency can be accumulated or even reproduced if not properly identified. Minimizing latency is important for real-time applications that are time-sensitive. For some applications, the time difference between the input and the output is not critical, as is the case with sentiment analysis in Twitter users, or opinion analysis in election campaigns through online social networks (OSNs). Nevertheless, for such applications, latency cannot be ignored when mass user behavior with respect to real-life events dynamically changes a desired output. For example, online betting platforms try to define a dynamic break-even point that maximizes the gain with respect to the real-time behavior of thousands of players that constantly changes over time. In the latter example, the basic challenge is to enable low latency responses for calculating this point of balance (break-even) between making either a profit or a loss, as the number and the behavior of players change.

3.2 Storage, sharing, transfer

The aforementioned challenges are strongly related to scalability issues. Another major issue is the choice of the location where to store the big data. New storage models have been introduced, mainly in the form of elastic web services that offer petabyte-scale data warehouses. Furthermore, enhanced RDBMS with Hadoop Map/Reduce integration schemes, as well as NoSQL DBMS, not only face the problem of storage, but also face the constraint of keeping low input/output latency from large data repositories. One of the benefits from the use of alternative DBMS schemas (e.g., NoSQL) is that they allow changes to take place without costly re-organization at the storage layer because their properties are not tied to a specific data model but can be stored in any necessary structure or format [34]. Thus, the challenge in big data storage is about how to avoid increased latency while handling high-volume data acquisitions and supporting a variety of mixed data structures. As far as sharing and the transfer of big data are concerned, both contribute to the so-called "Internet plumping problem". This is because the growth of wired/optical and wireless infrastructures does not keep up with the growth of data. For example, it is estimated that between 2012 and 2017 the connection speed in mobile networks will be increased seven times, while for the same period the global mobile data traffic is expected to increase by thirteen times.[22] During this time period, apart from delivering content to users, mobile devices (e.g., smartphones, tablets) will be "sensors" for data acquisition and capture. In addition, we also expect that a significant amount of machine-to-machine traffic data to be generated thereby loading networking infrastructures further. All the above issues contribute to the ultimate challenge of how so much data will "flow" through the network and be shared and stored.

3.3 Analysis and collection of results

The analysis of massive amounts of data to return useful results back to consumers is also one of the biggest challenges of big data. Data consumers (researchers, citizens, policy makers, etc.) are usually unaware or they simply do not care about some the difficulties confronted by the previous challenges mentioned earlier. They only want high-quality results in a format that they can easily understand and use. The high volume of data along with its heterogeneity and varied data structures make the challenge of analysis and results provision even harder. Academic and research communities worldwide are leveraging classic data mining algorithms in the big data era [35]. However, an important issue that should be addressed in the future is the dispersion of error that occurs in each separate analysis among mixed data structures. In big data analysis, where a large number of operational functions (e.g., Hadoop Map/Reduce

[22] http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf.

integration schemes) are executed, errors may be introduced at each distinct level of analysis, thus reducing the inference effectiveness of traditional algorithms that work well only with ordinary-sized data sets. Furthermore, data analysis and the results obtained may be also time-dependent, while the desired accuracy and precision of results from big data analysis depend on the target audience making use of these results. For instance, a target audience may be more interested in having precise trend analytics (e.g., the influence of a political blog and its virality among different election regions) while another audience may be keen on having the break-even profit threshold quickly calculated (as discussed previously) during which many players bet simultaneously on the same real-time event.

3.4 Ethical considerations

There are several ethical issues that arise when handling big data. According to Davis, ethical considerations include issues of identity, privacy, ownership, and reputation [36]. Identity has to do with the relationship between our real and online lives, while privacy involves actions and rules that define who has access and to what information? Ownership or accountability issues define who owns the data and who is the stakeholder responsible for its management, distribution and analysis? Intellectual property right issues also arise during the collection, storage, sharing and processing of big data. Finally, veracity is related to trust issues.

All the above ethical considerations are related to each other. For example, the identification of a user through his/her online activity leads to the loss of being free from public attention (privacy loss). Personal preferences and habits may be captured, stored, shared and analyzed for other peoples' uses. There is an ongoing debate whether emerging online applications that facilitate the upload of multimedia content through wearable devices, should or should not be allowed because they trespass the private sphere of an individual upon capturing, storing or sharing an event. The digital footprints left behind, may reveal unique features about the individual similar to the "digital DNA"—that would otherwise go unnoticed [37,38]. The situation is even more sensitive when it has to do with data from medical records, bio-signals, genetics and other confidential health-related information. From big data cleansing up to analysis, the privacy of patients may be threatened due to the exposure of health information to unauthorized stakeholders. For example, even if key descriptive fields (e.g., name, date of birth) from electronic health records (EHRs) are hidden, it has been proven that when the rest of the data are properly combined with other data sources and their properties, it is still possible to identify the patient with a high probability [39].

Table 2 shows a classification of the challenges versus the limitations scientists should deal with when handling big data.

4 Future research directions

Based on the above classification of challenges and related issues that big data scientists need to address, we present some future research directions and we discuss some possible solutions that can be applied to the handling of big data.

Table 2 Challenges versus issues to deal within big data

Challenges	Issues							
	Scalability	Source heterogeneity	Inconsistent data	Latency	Timeliness	Processing complexity	Data provenance	Trust, privacy
Cleansing/acquisition/capture	√	√	√	√	√	√	√	√
Storage/sharing/transfer	√			√	√		√	√
Analysis/results	√	√		√	√	√	√	√
Ethical considerations		√					√	√

The "√" sign highlights the presence of a limitation with respect to the challenges

4.1 Data cleansing and data curation

As we mentioned previously, inconsistent data from many different sources is a significant challenge when handling big data. To address this challenge, one possible solution is to set strict rules from the very beginning of data creation (data entry, submission, acquisition, etc.) in all big data applications. This means that the big data research community should adapt some of the traditional solutions for data cleansing and data curation to take into consideration the 4Vs of big data. For example, in the work described in [40], the authors introduced a continuous data-cleansing framework that can be applied to constrained environments and big data. They employ traditional machine learning approaches to predict the type of the repair needed in order to resolve data inconsistencies. Similarly, by modifying a semantic keyword-matching algorithm, the authors in [41,42] introduced a generic semantic-based framework for achieving effective big data cleansing. Finally, in [43] the authors clean, curate and publish data from multiple open linked sources in the context of the semantic web. In addition, the data providers/creators should be motivated or even rewarded to refine their data in re-useable and integrated formats with the help of data curation platforms. Some open tools for data curation include Open Refine,[23] Data Tamer [42], Data Wrangler,[24] and Plyr.[25] Other proprietary tools include IBM InfoSphere,[26] DataCleaner,[27] Paxata[28] and Xplenty.[29]

After the data are cleaned, we need to deal with the heterogeneity of data from various sources. To achieve this goal, we need to develop middleware protocol stacks that can support different data formats. Such middleware protocol stacks should have their own internal "translation" map based on some common language format for translating different data formats. The middleware layer should be able to produce an ecosystem of interoperable data sources according to the necessary metadata and integration schemes created by the data curation tools. Such a solution is illustrated in Fig. 2, where big data from different sources are captured, curated and transformed in an integrated, uniform format. In this case, the common language for translating the different data formats is the eXtensible Resource Protocol (XRI)[30] XRI describes physical sources by employing different structured identifiers from other domains, such as Internationalized Resource Identifiers (IRIs), Domain Name Service (DNS), International Standard Book Number (ISBN), Telephone Uniform Resource Identifiers (Tel URIs), etc. By using XRI, the data discovery process is independent of the application, the domain, the location, and the delivery scheme the data follows. Thus, data can be shared across different data warehouses and can use different communi-

[23] http://openrefine.org/.

[24] http://vis.stanford.edu/wrangler/.

[25] http://cran.r-project.org/web/packages/plyr/index.html.

[26] http://www-01.ibm.com/software/data/infosphere/.

[27] http://datacleaner.org/.

[28] http://www.paxata.com/.

[29] https://www.xplenty.com/.

[30] http://docs.oasis-open.org/xri/2.0/specs/xri-resolution-V2.0.html.

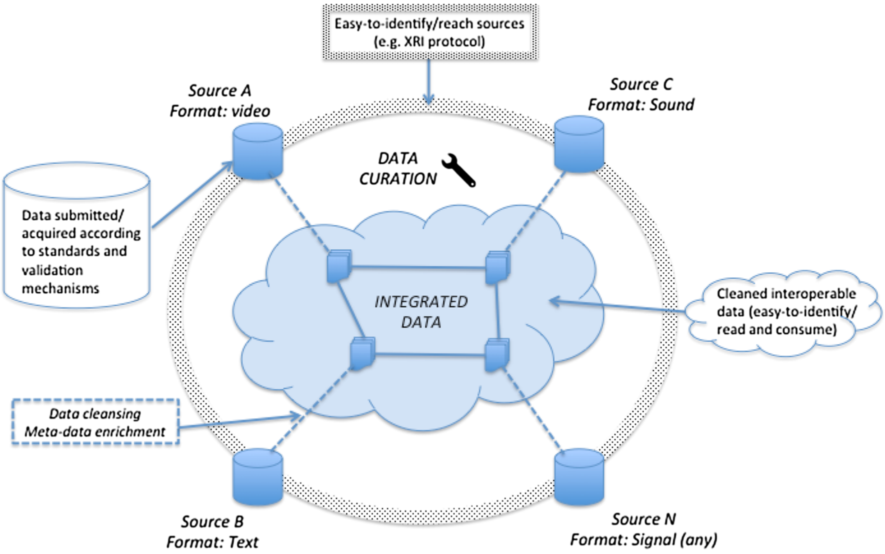

Fig. 2 Big data capture from heterogeneous sources and their integration in a uniform format

cation protocols. Similar solutions are highlighted in [44,45], where the authors focus mainly on the challenges of data integration and data interoperability in big data, while they explore the progress that has been made by the data integration community on the topics of source mapping, record linkage and data fusion.

Finally, effective data cleansing can be achieved over large data sets by using crowdsourcing techniques. Such techniques work by gathering specific opinions, ideas, answers or other simple interactions coming out from a large community, rather than from the knowledge of a few domain experts. A good example of a data-cleansing technique based on crowdsourcing is the “reCAPTCHA” tool. On one hand, this tool offers a free anti-bot service to the community, while on the other hand it helps with the massive digitization of old printed material for the cases where classical optical character recognition (OCR) techniques fail to perform adequately. By making use of crowdsourcing, the authors of the work described in [46] present the architecture of a cloud platform where a community of disaster response researchers could use the platform to curate big data. Similarly, collective intelligence (crowdsourcing) methods along with traditional artificial intelligence techniques for big data curation and integration are also discussed in [47].

4.2 Storage/sharing/transfer of big data

Another major challenge with big data storage is the handling of high-volume data acquisitions from highly diverse data structures with minimum latency. In an attempt to make use of simple parallelization models for processing big data, Google introduced MapReduce in 2005. MapReduce, which runs over the Google File System (GFS), revolutionized the way researchers deal with massive amounts of data that is stored over thousands of possible nodes [48]. This programming model was supported by a distributed storage system designed for handling structured data called BigTable [49]. Such big data computing tools were made available to the research community as Hadoop MapReduce (running over the Hadoop Distributed File System - HDFS)[31] and HBase.[32] Amazon, as one of the key-players in storage and data management, also adopted a similar structured data handling approach named DynamoDB,[33] which is a high performance, no relational database that offers web services over the Internet.

On the issue of “where to put my big data?”, cloud-based services have become a promising solution. During the early ages of the World Wide Web at 1993, 1 GByte of storage used to cost nearly 2000 US dollars, while today many online storage providers offer tens of GBytes for free. An increasing number of researchers, organizations, large companies and other stakeholders are adopting cloud-based solutions for storing their big data instead of the classical DBMS. With the cloud-based approach, there is no need for maintenance, backups or continuous updates in both hardware/software at the stakeholder’s side. As the needs scale up, we pay for them, transferring all issues such as quality-of-service or quality-of-experience to the big data storage provider’s side.

Performance issues such as latency and throughput will be the responsibility of the provider’s side, where all storage infrastructures should utilize cutting edge technologies [such as solid-state drives (SSDs)] to optimize data transfers. However, since SSDs are still expensive, a good solution for the big data stakeholders’ community is to have their data stored and replicated in a cloud environment, capable of supporting an open distributed storage system (e.g., HDFS/HBase) over a large number of commodity servers that keep data on hybrid drives. Hybrid drives combine the benefits of SSDs, along with the cheap storage capacity of classical magnetic disks (HDDs). Moreover, they provide higher throughput and present lower latency compared to HDDs, by employing solid-state storage for advanced caching of big data.

Finally, we need to transfer and share big data. To address the “Internet plumping problem” that we mentioned earlier, one solution is the dynamic data placement in content delivery networks (CDNs) around the world. In the work described in [50] the authors proposed a workload-aware framework called SWORD for solving the problems of data placement and replication in CDNs in the context of graph partitioning. By modeling the expected workload as a hypergraph, they introduced partitioning techniques that minimize the average number of nodes involved in the delivery transactions. In addition, intelligence can be seen here as a mechanism for allocating a fair bandwidth and data placement among CDN nodes. Such intelligence can be derived from the context of the delivered data, the users that share this data, as well as the network infrastructure (intelligent networks). The authors in [51,52] examined the behavior of real online social networks that deliver multimedia content and then they designed propagation predictors to enable a social-aware replication strategy for delivering multimedia content to users. These predictors are continuously calculated from

[31] https://hadoop.apache.org/.

[32] https://hbase.apache.org/.

[33] http://aws.amazon.com/dynamodb/.

the users' behaviors, their physical locations and their influences across their followers in the social network, thus guiding the region selection, bandwidth reservation, and data placement in the replication framework proposed. Another possible solution in sharing big data could be the adaptation of the BitTorrent protocol, in a sense that big data volumes can be shared/transferred similar to the operation of this classic protocol. In other words, a standard (big) data unit (similar to a BitTorrent piece that consists of several segments) will flow over different paths, while storage and replication nodes will act as seeders (nodes that have all big data replicated) and leechers (nodes that have the big data partially replicated) that want to become seeders. In this way, big data can reach their destination without consuming the full capacity and resources along the paths of a CDN. Moreover, a good solution for big data transfers can be achieved through techniques where hosting services are not required to own the network infrastructure. Thus, cloud-based CDNs can create additional overlay networks as relievers when resource availability decreases. In the works described in [53] and [54] the authors identified the challenges of resource provisioning and replica placements in cloud-based CDNs such as dealing with dynamic demand patterns that create overlay networks that adjust to the placement of contents and route maps. Key content providers (e.g., Akamai, Amazon) should focus on these aforementioned issues because solving them will increase the end-user perception of quality [55]. One possible solution is the direct shipping of containers with physical storage devices (such as magnetic disks, hybrid drives, etc.) to the storage nodes. During the last few years, Amazon has been working on a project for delivering goods through unmanned aerial vehicles.[34] So, an interesting proposal is to have this project extended and use such drones for delivering physical storage devices (and thus their saved data) across storage nodes (known as buckets/edges), thus minimizing the bandwidth consumption in their network infrastructures.

4.3 Analysis and collection of results

Big data analysis is mostly data driven. This is in contrast to the classical approach where a problem is identified and a solution needs to be found to solve it. To solve big data problems theoretical computer scientists argue that the whole scientific approach should be adapted around the new big data challenges [35,56]. Thus, new theoretical computing models should be introduced. Computer scientists should work on new theoretical principles that deal with scaling inferences, as well as new artificial intelligence models that learn from large-scale data [57]. The authors in [58] proposed an auto-scaling strategy inspired by classical artificial intelligence models and we should be able to determine the right point in time to scale in or to scale out for workload of real data that is continuously changing. Besides new theoretical principles, researchers also need to work in order to improve MapReduce itself. This is because MapReduce is limited to the kind of applications that every input key-value pair is independent of each other. A new modeling paradigm is suggested in [59] which extends the general applicability of MapReduce by allowing the dependence within a set of input key-

[34] http://www.amazon.com/b?node=8037720011.

value pairs, while preserving independence among all sets. The authors introduce a two-phase data processing not only to guarantee dependence between key-value pairs, but also to exploit the cooperation between the mappers and reducers. Finally, the Compute Unified Device Architecture (CUDA) programming model, introduced by Nvidia, opens up a promising solution for big data analytics, by using the massively parallel and highly multi-threaded graphics processing units (GPUs). However, past experience indicates that it is still difficult to analyze big data streams through CUDA GPU programming techniques because of the complexities between the underlying GPU hardware architecture and classical programming models, as well as because of the limited bandwidth available between GPUs and CPUs [60]. The authors in [61] introduced a scheme that provides pseudo-virtual memory to GPU applications (Big Kernel) to enable programming on the GPU hardware architecture. According to their evaluation on six data-intensive benchmarks, they managed to achieve an average speedup between 1.7 and 3 over different buffering and multi-threaded CPU implementations. Similarly, the authors in [62] performed a large-scale graph analysis based on real online social network, and managed to analyze their results up to ten times faster as compared to a classical big data technique.

As far as the collection of results is concerned, a significant goal should be to display meaningful and easy-to-consume results back to the stakeholders. This requires that the results are adapted according to the stakeholders' needs and their ability to interpret the information presented to them. This would also be beneficial to other actors (data cleaners, data integrators, theoretical scientists, analysts) involved with the big data. Furthermore, the data format should be easy to understand by the stakeholders. Grouping of data should be another goal for effective results provision. By grouping the data (data clustering), we can provide easy-to-consume and visible results to the stakeholders who can take this added value and transfer it to their research community, their business or their society. One may argue that this process is similar to the first learning processes performed in our early childhood. Abstraction in knowledge comes through a continuous clustering of many and different stimuli our body and brain sense and analyze [63]. Finally, an important step during the output of results is the separation of valuable information from the outliers during the analysis of the big data set. The separation between beneficial and erroneous or redundant information should be clear, not only within the internal algorithmic analysis, but also to the audience of big data as well. This can help in the data cleansing, the data provenance and relieve hardware and software sources from the overhead of outliers and redundant data [64–67].

4.4 Ethical considerations

We identified earlier that the first ethical consideration for big data is about intellectual property (IP) and its related issues. In other words, organizations, big data providers and their stakeholders have to address the issue of acknowledgement of IP rights for the massive unstructured data they collect, store, process, share and serve. This is a very sensitive and complex problem because a big data set may consist of a variety of intellectual properties, such as research results, copyrights, trademarks, patents and other intangible assets. Moreover, the problem is further exacerbated because in most of the cases IP rights are practically "static" legal procedures in a digital world where the relationship and interaction between humans and machines constantly evolves. One possible solution is the adoption of a Creative Commons[35] (CC) license and its application during data capturing/collection. In addition to solving IP issues, these licenses provide the opportunity for big data users to build upon the distributed information, thus adding value to the big data. Some examples of big data applications (from different disciplines such as culture, education, transparency in government and research) that use CC licenses include Europeana,[36] the MIT OpenCourseWare,[37] the "Where public money goes worldwide"[38] [68] initiative and the "Personal Genome Project".[39] However, big data is expected to expand the broader issue of data ownership regarding intellectual property rights through CC licenses. The role of such licenses is to provide a framework for individual negotiations on IP rights, between the one who owns the copyright and one that uses it. Yet, in the case of big data where we have multiple sources with probably different licenses, a proof of ownership cannot be easily provided [69].

As far as privacy is concerned, it is commonly accepted that sometimes we do exchange some space from our privacy sphere in order to reap some benefits from the new digital world. A lot of efforts should be made at the very beginning of our interaction as simple users of this new digital realm. As mentioned previously, we always leave a digital footprint or a "digital DNA" that reveals preferences, geo-location information, interests or even dislikes. Nevertheless, it is beyond the human capabilities to control what part of information impacts our privacy. The authors of [70] proposed a modeling approach under which users of social media can gain awareness of how some specific personal information (e.g., geo-location, likes, personal browsing activity) can cause loss of their privacy, when properly combined with other social media. The study was conducted on large data sets from Flickr, Locr, Facebook and Google+. Furthermore, a global solution for privacy protection in big data platforms and silos includes the creation of ethical data sharing protocols and techniques, especially in research involving human subjects.[40] However, most existing privacy preserving approaches are tailored to small-scale data sets, making them unable to handle big data because of scalability issues. An approach that uses MapReduce techniques and deals with anonymization in big data is proposed in [71]. The authors demonstrated how, by employing semantic proximity in data values, they were able to identify the attributes that contain privacy-sensitive information, and performed massive anonymization as a proximity-aware clustering problem. Evaluation assessments in terms of scalability and time-efficiency (in the case of sharing massive electronic health record values) revealed that the proposed approach outperformed other approaches, by avoiding privacy leaks in big data local-recoding anonymization.

[35] http://creativecommons.org/.

[36] http://www.europeana.eu/.

[37] http://ocw.mit.edu/index.html.

[38] http://publicspending.net/.

[39] http://www.personalgenomes.org/.

[40] http://www.ands.org.au/guides/ethics-working-level.html.

5 Conclusion

The importance of big data will continue to grow in the future. The amount of data being produced is growing exponentially as digital systems become cheaper, application environments become more and more networked, and analytics platforms are shared through web 2.0 applications (e.g., social networks, cloud services, powerful elastic capabilities, etc.). Research in big data is attracting the attention of researchers from several fields spanning from computer science and informatics to networking, artificial intelligence, statistics, data warehouses, as well as social, ethical and legal issues.

In this work, we discussed some of the drivers behind big data proliferation as well some of the main platforms that satisfy the “4V” characteristics of big data. We proposed a classification of the challenges when handling big data based on bottom-up layer approach that classifies the problems and limitations we face when we acquire, store and analyze big data. Based on this classification, we present some future research directions and we discuss some possible solutions that can be applied to the handling of big data. The main aim of this work is to enhance the awareness in the diverse group of cross-disciplinary big data stakeholders with respect to the challenges and solutions associated with big data growth.

Acknowledgments We express our gratitude to Emna Mezghani for her contributions to this work. The authors thank the anonymous reviewers for their valuable comments and suggestions.

References

1. Jacobs A (2009) The pathologies of big data. Commun ACM 52(8):36–44
2. Madden S (2012) From databases to big data. IEEE Internet Comput 16(3):4–6
3. Wu X, Zhu X, Wu GQ, Ding W (2014) Data mining with big data. IEEE Trans Knowl Data Eng 26(1):97–107
4. Gantz J, Reinsel D (2011) Extracting value from chaos. IDC iView, pp 1–12
5. Banaee H, Ahmed MU, Loutfi A (2013) Data mining for wearable sensors in health monitoring systems: a review of recent trends and challenges. Sensors 13(12):17472–17500
6. Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Khan SU (2015) The rise of ‘big data’ on cloud computing: review and open research issues. Inf Syst 47:98–115
7. Kwon O, Lee N, Shin B (2014) Data quality management, data usage experience and acquisition intention of big data analytics. Int J Inf Manag 34(3):387–394
8. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH (2016) Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute, 2011. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation. Accessed 12 January 2016
9. Pretz K (2016) Better health care through data: how health analytics could contain costs and improve care. The IEEE Institute, New York. http://theinstitute.ieee.org/technology-focus/technology-topic/better-health-care-through-data. Accessed 12 January 2016
10. Chen H, Compton S, Hsiao O (2013) DiabeticLink: a health big data system for patient empowerment and personalized healthcare, vol 8040. In: Smart health. Springer, Berlin, pp 71–83
11. O’Driscoll A, Daugelaite J, Sleator RD (2013) Big data. Hadoop and cloud computing in genomics. J Biomed Inf 46(5):774–781
12. Big Data Insight Group. http://www.thebigdatainsightgroup.com/site/article/nypd-make-big-apple-safer-big-data. Accessed 12 January 2016
13. Rozenfeld M (2016) The future of crime prevention. IEEE Institute, New York. http://theinstitute.ieee.org/technology-focus/technology-topic/the-future-of-crime-prevention. Accessed 12 January 2016

14. NASA Jet Propulsion Laboratory, Managing the deluge of 'Big Data' from space. http://solarsystem.nasa.gov/news/display.cfm?News_ID=45192. Accessed 12 January 2016
15. Kambatla K, Kollias G, Kumar V, Grama A (2014) Trends in big data analytics. J Parallel Distrib Comput 74(7):2561–2573 ISSN 0743–7315
16. Atzeni P, Bugiotti F, Rossi L (2014) Uniform access to NoSQL systems. Inf Syst 43:117–133 ISSN 0306–4379
17. Chen M, Mao S, Liu Y (2014) Big data: a survey. Mobile Netw Appl 19(2):171–209
18. Owen S, Anil R, Dunning T, Friedman E (2011) Mahout in action. Manning Publications Co, USA ISBN: 9781935182689
19. Prakashbhai PA, Pandey HM (2014) Inference patterns from Big Data using aggregation, filtering and tagging—a survey. In: 5th international conference The next generation information technology summit (confluence), September 2014, pp 66–71
20. Hu H, Wen Y, Chua TS, Li X (2014) Toward scalable systems for big data analytics: a technology tutorial. IEEE Access 2:652–687
21. Che D, Safran M, Peng Z (2013) From big data to big data mining: challenges, issues, and opportunities. In: Lecture notes in computer science, vol 7827, pp 1–15
22. Tan W, Blake MB, Saleh I, Dustdar S (2013) Social-network-sourced big data analytics. IEEE Internet Comput 7(5):62–69
23. Lin J, Kolcz A (2012) Large-scale machine learning at twitter. In: Proceedings of the 2012 ACM SIGMOD international conference on management of data (SIGMOD '12). ACM, New York, pp 793–804
24. Liu J, Liu F, Ansari N (2014) Monitoring and analyzing big traffic data of a large-scale cellular network with Hadoop. IEEE Netw 28(4):32–39
25. Marchal S, Francois J, State R, Engel T (2014) Phishstorm: detecting phishing with streaming analytics. IEEE Trans Netw Serv Manag 11(4):458–471
26. Ma C, Zhang HH, Wang X (2014) Machine learning for Big Data analytics in plants. Trends Plant Sci 19(12):798–808
27. Chandola V, Sukumar SR, Schryver JC (2013) Knowledge discovery from massive healthcare claims data. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '13). ACM, New York, pp 1312–1320
28. 3M Meeting Network. http://www.3rd-force.org/meetingnetwork/files/meetingguide_pres.pdf. Accessed 12 January 2016
29. Reda K, Febretti A, Knoll A, Aurisano J, Leigh J, Johnson AE, Papka ME, Hereld M (2013) Visualizing large, heterogeneous data in hybrid-reality environments. IEEE Comput Graph Appl 33(4):38–48
30. Philip Chen CL, Zhang CY (2014) Data-intensive applications, challenges, techniques and technologies: a survey on Big Data. Inf Sci 275:314–347
31. Jagadish HV, Gehrke J, Labrinidis A, Papakonstantinou Y, Patel JM, Ramakrishnan R, Shahabi C (2014) Big data and its technical challenges. Commun ACM 57(7):86–94
32. Labrinidis A, Jagadish HV (2012) Challenges and opportunities with big data. Proc VLDB Endow 5(12):2032–2033
33. Buneman P, Khanna S, Tan W (2000) Data provenance: some basic issues. In: Proceedings of foundations of software technology and theoretical computer science (FST TCS 2000). LNCS, vol 1974, pp 87–93
34. Price S, Flach PA (2013) A Higher-order data flow model for heterogeneous Big Data. In: 2013 IEEE international conference on big data, October 2013, pp 569–574
35. Xindong W, Xingquan Z, Gong-Qing W, Wei D (2014) Data mining with big data. IEEE Trans Knowl Data Eng 26(1):97–107
36. Davis K, Patterson D (2012) Ethics of big data, O'Reilly. ISBN 978-1-4493-1179-7
37. Mann S (2012) Through the glass. Light IEEE Technol Soc Mag 31(3):10–14
38. Michael K, Miller KW (2013) Big data: new opportunities and new challenges. IEEE Comput 46(6):22–24
39. Kupwade PH, Seshadri R (2014) Big data security and privacy issues in healthcare. In: 2014 IEEE international congress on big data, pp 762–765
40. Volkovs M, Fei C, Szlichta J, Miller RJ (2014) Continuous data cleaning. In: 2014 IEEE 30th international conference on data engineering (ICDE), pp 244–255

41. Wang J, Song Z, Li Q, Yu J, Chen F (2014) Semantic-based intelligent data clean framework for big data. In: 2014 international conference on security, pattern analysis, and cybernetics (SPAC), pp 448–453
42. Stonebraker M, Bruckner D, Ilyas I, Beskales G, Cherniack M, Zdonik S, Pagan A, Xu S (2013) Data curation at scale: the data tamer system. In: Proceedings of biennial ACM conference on innovative data systems research (CIDR'13), Alisomar
43. Bansal SK (2014) Towards a semantic extract-transform-load (ETL) framework for big data integration. In: 2014 IEEE international congress on big data (BigData Congress), pp 522–529
44. Kadadi A, Agrawal R, Nyamful C, Atiq R (2014) Challenges of data integration and interoperability in big data. In: 2014 IEEE international conference on big data (Big Data), pp 38–40
45. Dong XL, Srivastava D (2013) Big data integration. In: 2013 IEEE 29th international conference on data engineering (ICDE), pp 1245–1248
46. Sowe SK, Zettsu K (2013) The architecture and design of a community-based cloud platform for curating big data. In: 2013 international conference on cyber-enabled distributed computing and knowledge discovery (CyberC), pp 171–178
47. O'Leary DE (2014) Embedding AI and crowdsourcing in the big data lake. IEEE Intell Syst 29(5):70–73
48. Dean J, Ghemawat S (2008) MapReduce: simplified data processing on large clusters. ACM Commun 51(1):107–113
49. Chang F, Dean J, Ghemawat S, Hsieh WC, Wallach DA, Burrows M, Chandra T, Fikes A, Gruber RE (2008) Bigtable: a distributed storage system for structured data. ACM Trans Comput Syst 26(2):1–26
50. Kumar KA, Quamar A, Deshpande A, Khuller S (2014) SWORD: workload-aware data placement and replica selection for cloud data management systems. VLDB J 23(6):845–870
51. Wang Z, Zhu W, Chen X, Sun L, Liu J, Chen M, Cui P, Yang S (2013) Propagation-based social-aware multimedia content distribution. ACM Trans Multimed Comput Commun Appl (TOMM) 9(1):52:1–52:20
52. Wang Z, Zhu W, Chen M, Sun L, Yang S (2015) CPCDN: content delivery powered by context and user intelligence. IEEE Trans Multimed 17(1):92–103
53. Menglan H, Jun L, Yang W, Veeravalli B (2014) Practical resource provisioning and caching with dynamic resilience for cloud-based content distribution networks. IEEE Trans Parall Distrib Syst 25(8):2169–2179
54. Suto K, Nishiyama H, Kato N, Nakachi T, Fujii T, Takahara A (2014) Toward integrating overlay and physical networks for robust parallel processing architecture. IEEE Netw 28(4):40–45
55. Jiayi L, Rosenberg C, Simon G, Texier G (2014) Optimal delivery of rate-adaptive streams in underprovisioned networks. IEEE J Select Areas Commun 32(4):706–718
56. Fiore S, D'Anca A, Elia D, Palazzo C, Foster I, Williams D, Aloisio G (2014) Ophidia: a full software stack for scientific data analytics. In: 2014 international conference on high performance computing & simulation (HPCS), pp 343–350
57. Bhandarkar SM, Arabnia HR, Smith JW (1995) A reconfigurable architecture for image processing and computer vision. Int J Pattern Recognit Artif Intell (IJPRAI) 9(2):201–229. **(Special issue on VLSI Algorithms and Architectures for Computer Vision. Image Processing, Pattern Recognition and AI)**
58. Heinze T, Pappalardo V, Jerzak Z, Fetzer C (2014) Auto-scaling techniques for elastic data stream processing. In: 2014 IEEE 30th international conference on data engineering workshops (ICDEW), pp 296–302
59. Hsiang HW, Tse CY, Chien MW (2014) Multiple two-phase data processing with mapreduce. In: 2014 IEEE 7th international conference on cloud computing (CLOUD), pp 352–359
60. Arif Wani M, Arabnia HR (2003) Parallel edge-region-based segmentation algorithm targeted at reconfigurable multi-ring network. J Supercomput 25(1):43–63
61. Mokhtari R, Stumm M (2014) BigKernel—high performance CPU-GPU communication pipelining for big data-style applications. In: 2014 IEEE 28th international parallel and distributed processing symposium, pp 819–828
62. Chatterjee A, Radhakrishnan S, Sekharan CN (2014) Connecting the dots: triangle completion and related problems on large data sets using GPUs. In: 2014 IEEE international conference on big data (Big Data), pp 1–8
63. Shahar Y (1997) A framework for knowledge-based temporal abstraction. Elsevier Artif Intell 90(1–2):79–133

64. Tajer A, Veeravalli VV, Poor HV (2014) Outlying sequence detection in large data sets: a data-driven approach. IEEE Signal Process Mag 31(5):44–56
65. Bhandarkar SM, Arabnia HR (1995) The REFINE multiprocessor: theoretical properties and algorithms. Elsevier Parall Comput 21(11):1783–1806
66. Bhandarkar SM, Arabnia HR (1995) The Hough transform on a reconfigurable multi-ring network. J Parall Distrib Comput 24(1):107–114
67. Arabnia HR, Bhandarkar SM (1996) Parallel stereocorrelation on a reconfigurable multi-ring network. J Supercomput 10(3):243–270
68. Vafopoulos M, Meimaris M, Anagnostopoulos I, Papantoniou A, Xidias I, Alexiou G, Vafeiadis G, Klonaras M, Loumos V (2015) Public spending as LOD: the case of Greece. Seman Web Interoperabil Usabil Applicabil Seman Web 6(2):155–164
69. Ekbia H, Mattioli M, Kouper I, Arave G, Ghazinejad A, Bowman T, Suri VR, Tsou A, Weingart S, Sugimoto CR (2014) Big data, bigger dilemmas: a critical review. J Assoc Inf Sci Technol. Wiley, New York
70. Smith M, Szongott C, Henne B, von Voigt G (2012) Big data privacy issues in public social media. In: 6th IEEE international conference on digital ecosystems technologies (DEST), pp 1–6
71. Zhang X, Dou W, Pei J, Nepal S, Yang C, Liu C, Chen J (2015) Proximity-aware local-recoding anonymization with mapreduce for scalable big data privacy preservation in cloud. IEEE Trans Comput 64(8):2293–2307