J Supercomput (2016) 72:3073–3113
DOI 10.1007/s11227-015-1501-1

A general perspective of Big Data: applications, tools, challenges and trends

Lisbeth Rodríguez-Mazahua[1] · Cristian-Aarón Rodríguez-Enríquez[1] · José Luis Sánchez-Cervantes[1] · Jair Cervantes[2] · Jorge Luis García-Alcaraz[3] · Giner Alor-Hernández[1]

Published online: 20 August 2015

Abstract Big Data has become a very popular term. It refers to the enormous amount of structured, semi-structured and unstructured data that are exponentially generated by high-performance applications in many domains: biochemistry, genetics, molecular biology, physics, astronomy, business, to mention a few. Since the literature of Big Data has increased significantly in recent years, it becomes necessary to develop an overview of the state-of-the-art in Big Data. This paper aims to provide a comprehensive review of Big Data literature of the last 4 years, to identify the main challenges, areas of application, tools and emergent trends of Big Data. To meet this objective, we have analyzed and classified 457 papers concerning Big Data. This review gives relevant

✉ Giner Alor-Hernández
galor@itorizaba.edu.mx

Lisbeth Rodríguez-Mazahua
lrodriguez@itorizaba.edu.mx

Cristian-Aarón Rodríguez-Enríquez
crodriguezen@gmail.com

José Luis Sánchez-Cervantes
isc.jolu@gmail.com

Jair Cervantes
chazarra17@gmail.com

Jorge Luis García-Alcaraz
jorge.garcia@uacj.mx

1 Division of Research and Postgraduate Studies, Instituto Tecnológico de Orizaba, Av. Oriente 9 852. Col Emiliano Zapata, C.P. 94320 Orizaba, Mexico

2 Centro Universitario UAEM Texcoco, Universidad Autónoma del Estado de México, Av. Jardín Zumpango s/n, Fracc. El Tejocote, Texcoco, Estado de México, Mexico

3 Departamento de Ingeniera Industrial y Manufactura, Instituto de Ingeniería y Tecnología, Universidad Autónoma de Ciudad Juárez, Ciudad Juárez, Mexico

information to practitioners and researchers about the main trends in research and application of Big Data in different technical domains, as well as a reference overview of Big Data tools.

Keywords Application domains · Classification · Big Data · Literature review

1 Introduction

The term of Big Data is mainly used to describe massive, heterogeneous, and often unstructured digital content that is difficult to process using traditional data management tools and techniques [1]. Recently, companies, academia and government become interested in the high potential of Big Data. Nevertheless, a major challenge for Information Technology (IT) researches and practitioners is that this data growth rate is fast exceeding their ability to both: (1) design appropriate systems to handle the data effectively and (2) analyze it to extract relevant meaning for decision making. As a result, it is necessary to use tools and frameworks for the effective organization, management and analysis of such datasets. Big Data can be described using the 5V model [2] illustrated in Fig. 1. This model is an extension of the previously defined 3V model in [3], and includes:

– *Volume* (*the era of size*) With the generation and collection of masses of data, data scale becomes increasingly big. The data produced nowadays is estimated in the order of zettabytes, and it is growing around 40 % every year [4].
– *Velocity* (*the era of streaming data*) Means the timeliness of Big Data, specifically, data collection and analysis must be rapidly and timely conducted, so as to maximize the use of the commercial value of Big Data.

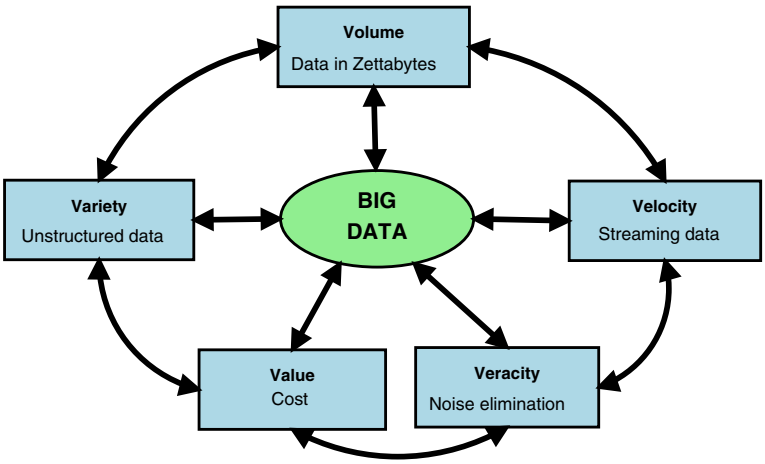

Fig. 1 The 5V model that currently defines Big Data

- *Variety* (*the era of unstructured data*) Indicates the various types of data, which include semi-structured and unstructured data such as audio, video, webpage, and text, as well as traditional structured data.
- *Value* (*the era of cost associated with data*) While the data are being generated, collected and analyzed from different quarters, it is important to state that today's data have some costs. The data itself can be a "commodity" that can be sold to third parties for revenue. Moreover, understanding the cost or value of the data can aid in budget decision making at estimating the storage cost of the data.
- *Veracity* (*the era of data pollution that needs cleansing*) There is the need to check the accuracy of the data by eliminating the noise through methodologies such as data pedigree and sanitization. This is to ensure data quality so that decisions that are made from the collected data are accurate and effective.

This work provides a review of several pieces of research concerning Big Data. This literature review was done to provide an evaluation perspective to determine the fields of application in where the use of Big Data has generated a major impact and the challenges addressed by Big Data research papers. Furthermore, this work has also identified the most commonly used tools, frameworks, and programming languages for Big Data. One of the most important goals of this work is to provide a reference guide to identify trends on Big Data development in terms of its practical application.

In literature, we have identified several survey papers on Big Data such as [5–12]. However, the main difference between this paper and other survey papers is that we provide a more comprehensive review of Big Data literature of the last 4 years, to identify the main challenges, areas of application, tools and emergent trends of Big Data. Some surveys, like [5,9,10], gave an overview of the state-of-the-art in Big Data analytics. Therefore, they only focused on one challenge of Big Data. Kaisler et al. [8] only identified some of the major challenges of Big Data. Other surveys [6,7,11,12] did not discuss the emergent trends of Big Data.

This paper is structured as follows. Section 2 describes the research methodology used in this study. Section 3 presents the criteria for the classification of the research papers reviewed. Section 4 gives a classification of Big Data papers considering four interesting criteria: (1) year of publication; (2) editorial; (3) domain of application, and (4) challenge addressed. Section 5 provides the analysis of the impact of Big Data on knowledge domains. Section 6 presents the classification of Big Data tools according to the kind of analysis. Section 7 describes some Big Data trends. Section 8 addresses future and open issues on Big Data. Finally Sect. 9 presents the conclusions and future directions.

2 Research methodology

The methodology is composed of three stages. The first stage was the research of works related to Big Data in several databases of scientific journals. The second concerned the classification of these works in the different fields of knowledge according to criteria established in Sect. 3 of this work. Finally, the third stage of the methodology involved the report of detailed literature review that identifies challenges addressed and technologies, tools, frameworks and programming languages used in the imple-

mentation of the proof of concept applications in some of the reported works. For this review we first searched in the major databases of electronic journals for a comprehensive bibliography of relevant research of Big Data. The digital libraries considered were: (1) ACM Digital Library, (2) IEEE Xplore Digital Library, (3) Science Direct (Elsevier) and (4) SpringerLink. Papers published by academic journals, workshops, and international conference proceedings are thought to be reliable and worthy.

Moreover, we also employed a keyword-based search to select the most relevant articles. The main keywords employed were: (1) Big Data, (2) Big Data Tools, (3) Big Data Applications, (4) Big Data Frameworks, (5) Big Data Mining, and (6) Large-Scale Datasets. The works not directly related to Big Data or not suitable for the study were discarded. The following statements describe the criteria considered for the omission of a research paper:

- Unpublished working papers, non-peer-reviewed papers, non-English papers, textbooks, Master and Doctoral dissertations.
- Because research of Big Data is relatively current, we have only searched recent articles published between 2010 and 2014. This 4-year period is considered to be representative for Big Data Research.

The selection process resulted in 457 research papers selected from four different digital libraries. Each paper was carefully reviewed and classified into one of the application field comprising Big Data. This review provides a reliable and comprehensible basis to understand the state of the art of Big Data and its applications. In following section, we describe the classification method to select the research papers considering five interesting criteria.

3 Classification method

The research papers selected were classified by considering the following criteria: (1) domains and fields of application of Big Data according to the "Research Trends", special issue on Big Data by Elsevier [13]; (2) specific domains or sub-domains; (3) challenge that they addressed; (4) tool or framework for Big Data management, and (5) programming language for Big Data applications.

3.1 Classification considering domains of application

Most of Big Data applications are designed for a specific domain, such as data mining [4], manufacturing [14], biomedicine [15], among others. Therefore, we have first classified the works according to the "Research Trends", special issue on Big Data by Elsevier, which groups the specialized domains (i.e., omics, genetics, agriculture, to mention but a few) as follows: (1) Computer Science, (2) Engineering, (3) Mathematics, (4) Business, Management and Accounting, (5) Physics and Astronomy, (6) Biochemistry, Genetics and Molecular Biology, (7) Social Sciences, (8) Materials Science, (9) Medicine, (10) Decision Sciences, (11) Multidisciplinary, (12) Arts and Humanities.

3.2 Classification considering Big Data challenges

We classified the works according to the difficulty that they addressed. According to [11,16] there are a lot of challenges when handle Big Data problems, difficulties lie in data capture, storage, searching, sharing, analysis and visualization.

3.2.1 Data capture

Big Data can be collected from several sources: transactions [17], metadata [18], social media [19–21], sensors [22,23], and experiments [24]. Due to the variety of disparate data sources and the sheer volume, it is difficult to collect and integrate data with scalability from distributed locations [25]. According to Brown et al. [17] future competitive benefits may accrue to companies that can not only capture more and better data but also use that data effectively at scale. Other problems related to this challenge are data transmission [25], automatic generation of metadata [18,26], and data pre-processing [25,26].

3.2.2 Data storage

Big Data not only requires a huge amount of storage, but also demands new data management on large distributed systems because conventional database systems have difficulty to manage Big Data [27]. NoSQL databases (i.e., non traditional databases), such as Apache Cassandra [28], Apache HBase [29] and Project Voldemort [30], are becoming the core technology for Big Data because of characteristics like being schema free, supporting easy replication, possessing a simple API, eventual consistency and supporting a huge amount of data [12,25,31]. MapReduce [32] permits automatic parallelization and scalable data distribution across many computers. The most popular implementation available as open-source software is Apache Hadoop [33]. Also, Cloud computing has been recognized as an alternative for massive data storage and computing in [1,18,34–36].

3.2.3 Data search

Since data are to be used to make accurate decisions in time, it becomes necessary that it should be available in accurate, complete and timely manner [7]. Query optimization has been crucial for efficiently answering a large class of complex analytic SQL queries. Even for newer platforms based on MapReduce and its invariants, the interest in leveraging higher level query languages (e.g., HiveQL [37], PigLatin [38], and SCOPE [39]) is extremely strong. In such highly parallel platforms, the cost of shuffling data across nodes is considerable and thus query optimization and physical design continue to be critical elements of the infrastructure [40].

3.2.4 Data sharing

Sharing data is now as important as producing it [41]. Professionals will continue to produce and consume information that is specific to their own business needs, but

it is now generated in a way that can be connected and shared to other aspects of an enterprise [42]. Researches in several scientific disciplines, such as ecology [43], medicine [44] and biology [35,45], are facing issues regarding data preservation and sharing. Some of these issues are data curation [11,43,45] and privacy preservation [44,46].

3.2.5 Data analysis

Timely and cost-effective analytics over Big Data is now a key ingredient for success in many businesses [17,18,47–49], scientific [15,19–21,34,35,50] and engineering [14,51,52] disciplines, and government endeavors [53,54]. The allure of hardware replication and system expandability as represented by Cloud computing [1,55] along with the MapReduce [27,56] and Message Passing Interface (MPI) [57] parallel program systems offers one solution to these challenges by utilizing a distributed approach [8]. Some problems related to this challenge are performance optimization of software frameworks for Big Data analytics [27,56,58,59], and scalable and distributed data management ecosystem for deep analytics (i.e. the implementation of various data mining [2,4,60], machine learning [61,62] and statistical [9,63] algorithms for analysis) [55,64,65].

3.2.6 Data visualization

The characteristics of Big Data make the data visualization a challenging task [62]. According to Fisher et al. [66], visual interfaces are extremely well suited for: (1) inspecting data at multiple scales in conjunction with statistical analysis; (2) providing a way to maintain context by showing data as a subset of a larger part of the data, showing correlated variables, and so on, and (3) helping to identify patterns over time in data streams. Visual analytics is an emerging field in which massive datasets are presented to users in visually compelling ways with the hope that users will be able to discover interesting relationships. Visual analytics requires generating many visualizations (often interactively [67,68]) across many datasets [54,69].

3.3 Classification considering the kind of analysis

A classification of programming languages, tools and frameworks used to implement Big Data was performed. Big Data tools are classified according to the kind of analysis in: (1) batch analysis, where data are firstly stored and then analyzed; (2) stream analysis, which is the extraction of information and knowledge discovery from continuous, rapidly generated streams of data, and (3) interactive analysis, which process the data in an interactive environment, allowing users to undertake their own analysis of information [5,11]. Also, this paper verified the support for Big Data applications provided by each of the languages presented in the classification. This way, professionals and developers can reuse this information as a reference guide to determine which technologies would be most suitable to implement in a Big Data application.

3.4 Classification process

The classification by domains of applications proposed in Sect. 3.1 was equally applied to all previously selected and reviewed research papers. The overall classification process consisted of the following stages:

1. Search over electronic database.
2. Classification by research trends application fields.
3. Classification by particular application field.

We apply the selection criteria shown in Fig. 2 getting the results that are described in the next section.

4 Classification of research papers

The results of this classification will offer the most important guidelines for future research on Big Data applications. In addition, results will also provide some insights about used technologies or a guide for practitioners and academics considering future directions. The details of the classification are described below.

Distribution by year of publication Big Data literature has increased considerably in the last 4 years; therefore, we analyzed Big Data papers from 2010 to 2014. The distribution of the analyzed papers according to this year of publication is depicted in Fig. 3. Analyzing the distribution of selected papers over the last years allows for inferring the increasing relevance of Big Data. A rise in the publication of research works related to Big Data over the past years can show the growing relevance of the Big Data research field.

Figure 3 shows a significant increase of Big Data papers for the period 2010–2014. Research works of 2010 were mainly focused on the use of Cloud computing to manage unstructured and large datasets in the fields of Biology and Business. These datasets were generated at astonishing scales by diverse technologies such as social networks, the Internet of Things, and genome sequencing [34,70,71]. From 2010 to 2011 we can notice almost an exponential growing. This is motivated by the high demand of Big Data analytics by industries and organizations to extract useful information from huge and chaotic data sets to support their core operations in many business and scientific applications [58]. There was an increase of more than 200 % from 2011 to 2012. This is due to a significant rise of works related to Big Data mining [4,72] and Big Data visualization [73,74]. Literature on Big Data had a growth of more than 150 % from 2012 to 2014. Research papers addressed the challenges of capturing [75], storing [76, 77], searching [78], sharing [23], analyzing [9,79] and visualizing [80,81] Big Data sets in several fields such as Computer Science [82,83]; Mathematics [63]; Business, Management and Accounting [41]; Engineering [84]; Physics and Astronomy [85,86]; Biochemistry, Molecular Biology and Genetics [87]; Medicine [88]; Social Sciences [89,90]; Materials Science [24]; Decision Sciences [91], and Arts and Humanities [92].

Distribution by Publisher Distribution of research papers by publisher is shown in Fig. 4. According to this figure, IEEE is the publisher with the greatest number of articles published about Big Data, since the work that this editorial publishes concerns

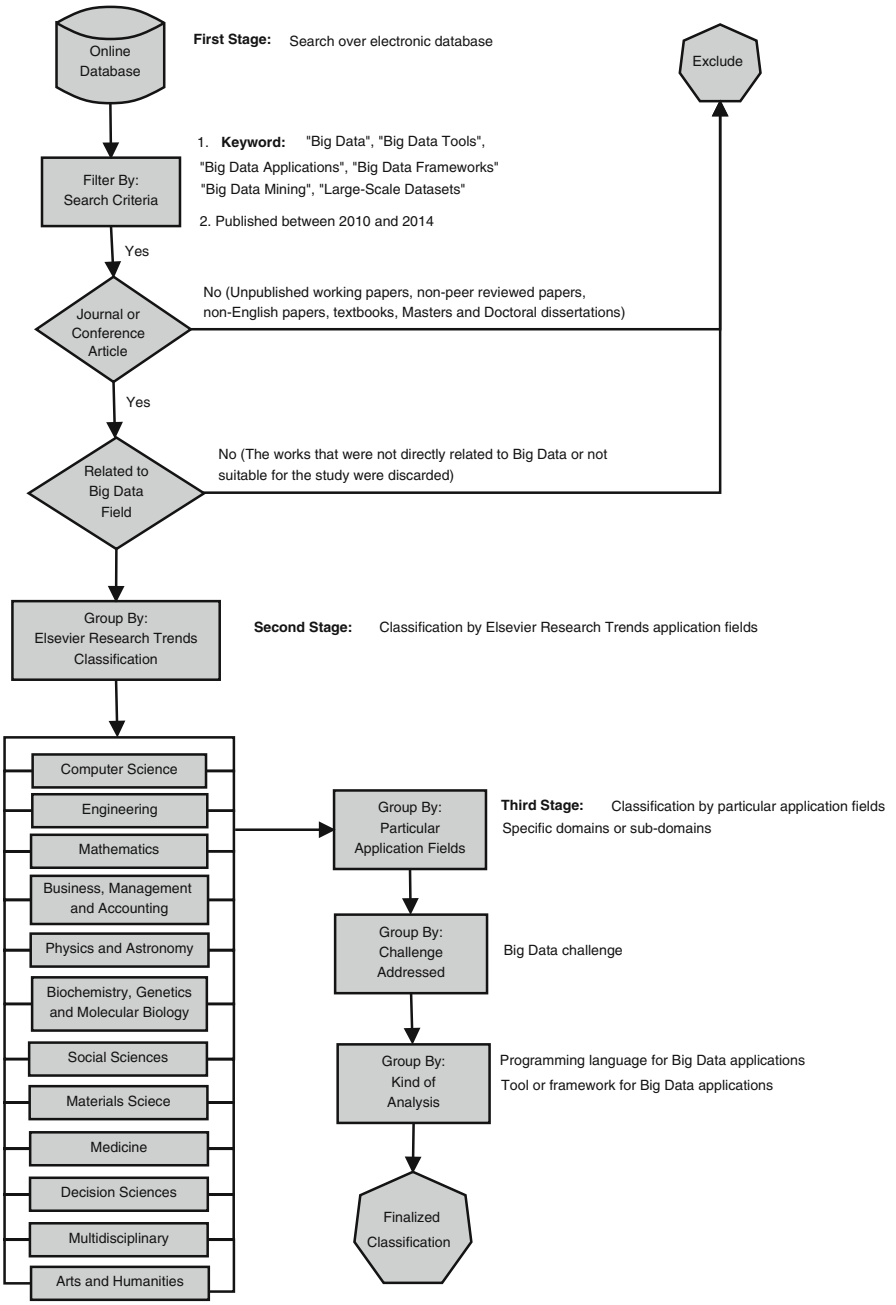

Fig. 2 Selection criteria flow diagram

practical applications of Big Data in all fields of knowledge, other editorials like McKinsey & Company focus most in only one domain, i.e., Business, Management and Accounting.

Fig. 3 Distribution of research papers by year of publication

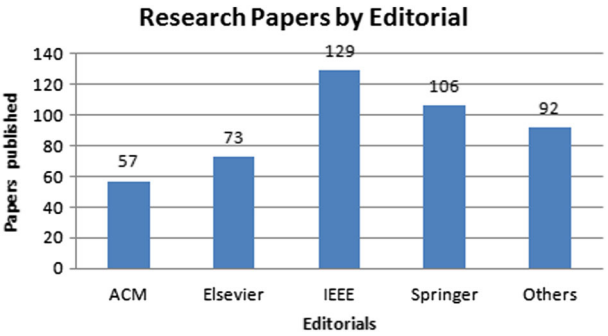

Fig. 4 Distribution of research papers by publisher

It is important to consider the distribution of papers by publisher to determine the paper focus: the challenge addressed by the works and the domain considered. Figure 5 shows the distribution of papers by the fields of application proposed in [13]. As Fig. 5 suggests, Computer Science field is the most important field in most of the editorials. This is due to the fact that one of the subdomains of Computer Science is Cloud computing, and most papers are related to this subdomain since Cloud computing is recognized as a scalable and cost-efficient solution to the Big Data challenge [93].

Figure 6 illustrates the distribution of research papers by challenge addressed. Figure 6 demonstrates that analysis is the most important challenge taken into account for Big Data researchers. This is because analysis drives nearly every aspect of our modern society, including mobile services, retail, manufacturing, financial services, life sciences and physical sciences [26]. In the next section, we describe the impact of Big Data on the twelve fields of application.

5 Impact of Big Data on knowledge domains

Halevi and Moed [13] searched research papers about Big Data from 1970 to 2010 on Scopus^{TM}. According to Halevi and Moed, subject areas researching Big Data is

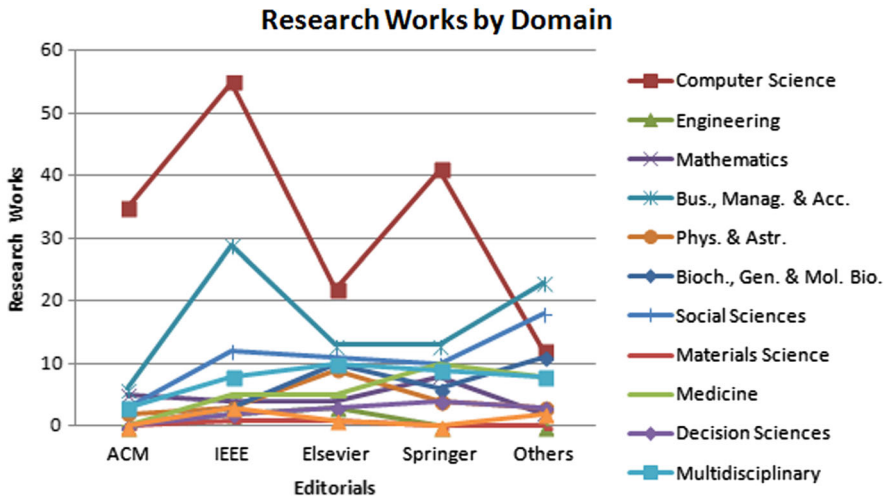

Fig. 5 Distribution of research works by "Research Trends" domains

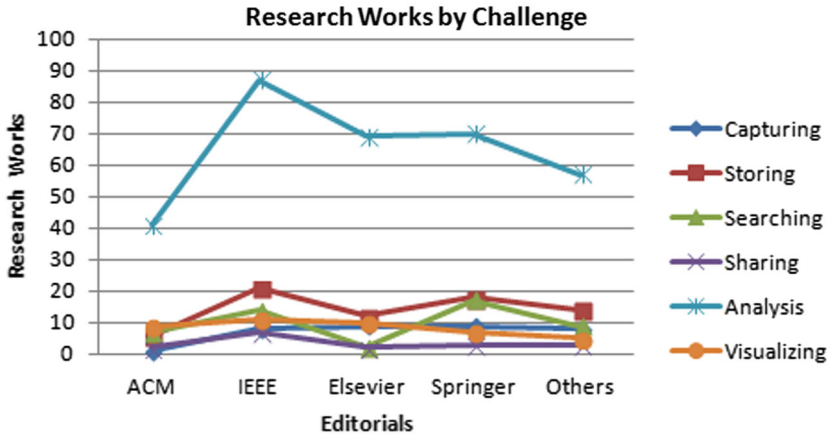

Fig. 6 Distribution of research works by challenge addressed

classified into twelve categories: (1) Computer Science; (2) Engineering; (3) Mathematics; (4) Business, Management and Accounting; (5) Physics and Astronomy; (6) Biochemistry, Genetics and Molecular Biology; (7) Social Sciences; (8) Materials Science; (9) Medicine; (10) Decision Sciences; (11) Multidisciplinary, and (12) Arts and Humanities. We analyzed the impact of Big Data literature in every area of application. Figure 7 shows the distribution of research papers by challenge addressed in every area. The most important challenge for all areas is Big Data analysis. This section discusses the most relevant works on these domains of knowledge, considering their importance in the literature review.

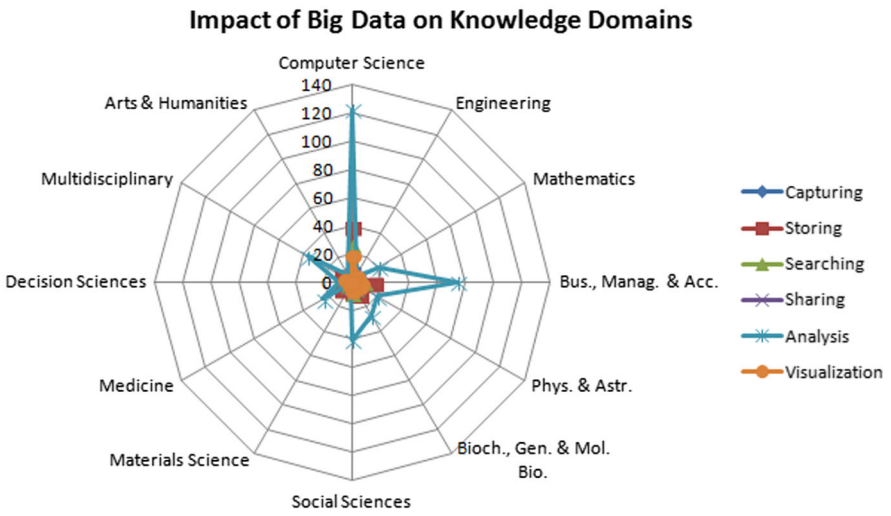

Fig. 7 Impact of Big Data on knowledge domains

5.1 Computer science

Many Big Data papers in computer science are related to databases and Cloud computing. In the database world Big Data problems arose when enterprises identified a need to create data warehouses to house their historical business data and to run large relational queries over the data for business analytics and reporting purposes [65]. Cloud computing is an extremely successful paradigm of service-oriented computing, and it has revolutionized the way on how the computing infrastructure is abstracted and used [55]. The National Institute for Standards and Technology (NIST) [94] defines Cloud computing as "a pay-per-use model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction".

Databases Literature related to databases is mainly focused on providing advanced analytics into a database environment [64,65,82,95], scalable data management [55,64] and performance optimization of the systems for managing Big Data [27,53,59].

Database Management systems (DBMSs) do a great job of helping a user maintain and curate a dataset, adding new data over time, maintaining derived views of the data, evolving its schema, and supporting backup, high availability, and consistency in various ways. The trouble comes when users want to take the accumulated data, collected over months or years, and learn something from it and they want the answer in seconds or minutes. The pathologies of Big Data are primarily those of analysis [95].

Existing DBMSs do not lend themselves to sophisticated analysis at the scale many users would like [64]. However, according to Madden [64], although databases do not solve velocity and variety aspects of the Big Data problem, several tools (some based on databases) get part-way there. What is missing is twofold: it is necessary to improve sta-

tistics and machine learning algorithms to be more robust and easier for unsophisticated users to apply. Second, it is required to develop a data management ecosystem around these algorithms so that users can manage and evolve their data, enforce consistency properties over it, and browse, visualize, and understand the results of their algorithms.

Analytics over Big Data play a relevant role in the context of data warehousing and OLAP research. In [82], authors provided an overview of the state-of-the-art research issues and achievements in the field of analytics over Big Data, and they extended the discussion to analytics over big multidimensional data as well, by highlighting open problems and trends.

Scalable and distributed data management has been the vision of the database research community for more than three decades. Much research has focused on designing scalable systems for both update intensive workloads as well as ad hoc analysis workloads. Changes in the data access patterns of applications and the need to scale out to thousands of commodity machines led to the birth of a new class of systems referred to as Key-Value stores [96,97] which are now being widely adopted by various enterprises. In the domain of data analysis, the MapReduce paradigm [32] and its open-source implementation Hadoop [33] has also seen widespread adoption in industry and academia alike [55].

There are many approaches for improving the performance of the Hadoop-based systems for Big Data analytics. Starfish [53] is a self-tuning system that gets good performance automatically, without any need for users to understand and manipulate the many tuning knobs in Hadoop. RCFile (Record Columnar File) [27] is a fast and space-efficient Big Data placement structure. Camdoop [59] optimizes the network to improve application performance.

In [65] authors reviewed the history of systems for managing Big Data in the database world and examined recent Big Data activities and architectures. Their focus was on architectural issues, particularly on the components and layers that have been developed in parallel databases and Hadoop and how they are being used to tackle the challenges posed by Big Data problems. Authors also presented the approach being taken in their ASTERIX project at UC Irvine, and hinted at their own answers to the questions regarding the "right" components and the "right" set of layers for taming the modern Big Data beast.

Cloud computing The development of Cloud computing provides solutions for the storage and processing of Big Data [1,12]. Some of the issues faced by Big Data researches related to the use of Cloud computing are data privacy [46,98,99], data management [40,55], and efficient processing and analysis of data [100,101].

The most popular cloud paradigms include: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). The concept, however, can also be extended to Database as a Service (DaaS) or Storage as a Service (SaaS). Elasticity, pay-per-use, low upfront investment, low time to market, and transfer of risks are some of the major enabling features that make Cloud computing a ubiquitous paradigm for deploying novel applications which were not economically feasible in a traditional enterprise infrastructure settings [55].

International Data Corporation (IDC) [102] estimated that by 2020, nearly 40 % of the information in the digital universe will be touched by Cloud computing (meaning

that a byte will be stored or processed in a cloud somewhere in its journey from originator to disposal). Perhaps as much as 15 % will be maintained in a cloud [36].

In [100], authors described a cloud-bursting based on maximally overlapped load-balancing algorithm to optimize the performance of Big Data analytics that can be run in loosely coupled and distributed computing environments such as federated clouds. To reduce the quantity and the processing time of Big Data sets encountered by the current typical Cloud Big Data processing techniques, in [101], authors proposed a spatiotemporal compression technique-based approach on cloud to deal with Big Data and big graph data from real-world applications.

One drawback associated with the utilization of Cloud computing, given the scale of data, is that transmitting such data over the Internet takes prolonged periods of time, sometimes even in the region of weeks. Thus, the bottleneck is the rate of data transfer, i.e., getting data into and out of the cloud [93].

5.2 Engineering

There are some papers related to Big Data in the Engineering Field, specifically in Manufacturing, Electrical Engineering and Chemical Engineering.

In Manufacturing, some important works are [14,47,103]. According to Lee et al. [14], to become more competitive, manufacturers need to integrate advanced analytics and cyber-physical systems to adapt to, as well as take advantage of, the current Big Data environment. In [103], authors affirmed that current supply chain professionals are inundated with data, motivating new ways of thinking about how data are produced, organized, and analyzed. This has provided an impetus for organizations to adopt and perfect data analytic functions (e.g. data science, predictive analytics, and Big Data) to enhance supply chain processes and, ultimately, performance. However, management decisions informed by the use of these data analytic methods are only as good as the data on which they are based. Therefore, Hazen et al. [103] introduced the data quality problem in the context of supply chain management (SCM) and propose methods for monitoring and controlling data quality.

One important paper related to Chemical Engineering is [52]. Belaud et al. [52] defined sustainability as a paradigm for thinking about the future in which environmental, societal and economic considerations are equitable in the pursuit of an improved lifestyle. Authors affirmed that Engineering domains have to develop according to this paradigm. Therefore, they discussed an open platform for collaborative simulation, scientific Big Data analysis and 3D visualizations for Computer-Aided Design/Engineering (CAD/E). Authors validated the platform using it as a support for sustainability in Chemical Engineering. Furthermore, an industrial sustainability-based application for managing natural resources and the environment was explained and discussed. The value brought to the scientific and industrial community is to make remote analysis and collaboration easily available and scalable.

Regarding Electrical Engineering, there are some important papers that analyzed the role of Big Data in the Smart Grid, destined to replace conventional electric grid [12,51]. In [104], authors defined System of Systems (SoS) as integrated, independent operating systems working in a cooperative mode to achieve a higher performance. Tannahill and Jamshidi [104] affirmed that a typical example of SoS is the Smart Grid,

and that a small-scale version of this SoS is a micro-grid designed to provide electric power to a local community. Thus, authors demonstrated how to construct a bridge between SoS and data analytics to develop reliable models for such systems. They used data analytics to generate a model to forecast produced photovoltaic energy to assist in the optimization of a micro-grid SoS. Tools like fuzzy inference, neural networks, Principal Component Analysis (PCA), and genetic algorithms were utilized.

5.3 Mathematics

Mathematics and, more specifically, statistics are some of the key fields for Big Data. Data are tamed and understood using computer and mathematical models. These models, like metaphors in the literature, are explanatory simplifications [105]. Statistical analysis is concerned with both summarizing large data sets (i.e., average, min, etc.) and in defining models for prediction. Such analysis is often the first step in understanding the data.

As massive data acquisition and storage becomes increasingly affordable, a wide variety of enterprises are employing statisticians to engage in sophisticated data analysis. These statisticians may have strong software skills but would typically rather focus on deep data analysis than database management [106]. To tackle this problem, Cohen et al. [106] developed a hierarchy of mathematical concepts in SQL, and encapsulate them in a way that allow analysts to work in relatively familiar statistical terminology, without having to develop statistical methods in SQL from first principles for each computation. They designed methods to convince a parallel database to behave like a massively scalable statistical package.

In [107], authors declared that many of the state-of-the-art Big Data analytics-driven systems, a. k. a. *trained systems*, are largely statistical and combine rich databases with software driven by statistical analysis and machine learning. Kumar et al. [107] presented the Hazy project to make these systems easier to build and maintain.

Big privacy (secure and confidential use of Big Data) perspectives from both computer science and statistical science are presented in [108]. Authors described research into how to define and measure the risks of confidentiality breaches. They also explained some approaches to data protection.

In the context of Big Data it is of high interest to find methods to reduce the size of the streaming data but keep its main characteristics according to these optimization problems. For this purpose, Feldman et al. [109] proposed to use coresets. A coreset is a semantic compression of a given dataset. Coresets are used for solving hard machine learning problems in parallel. In [110], authors explored and applied powerful methods and models developed in statistics to estimate results and the accuracy obtained from sampled data. They proposed a method and a system that optimize the workflow computation on massive data-sets to achieve the desired accuracy while minimizing the time and the resources required.

According to [111] theories and methods of mathematics and statistics are incorporated in the science about Data, i.e., Data Science. A practitioner of data science is called a data scientist who typically has a strong expertise in some scientific discipline, in addition to the ability of working with various elements of mathematics, statistics and computer science.

5.4 Business, management and accounting

According to the McKinsey Global Institute [47], companies embracing Big Data are able to outperform their peers. It estimates that a retailer that properly harnesses Big Data has the potential to increase its operating margins by more than 60 % by gaining market share over its competitors by taking advantage of detailed customer data.

Big Data represents both big opportunities and big challenges for CIOs (Chief Information Officers). Almost every CIO aspires to make IT a more valued asset to the organization. And IT is front and center in Big Data projects, which are typically at the boundaries of the business where any of the most significant business expansion or cost reduction opportunities lie [18].

Many companies are taking data use to new levels, using IT to support rigorous, constant business experimentation that guides decisions and to test new products, business models, and innovations in customer experience. In some cases the new approaches help companies make decisions in real time. This trend has the potential to drive a radical transformation in research, innovation, and marketing [71]. For example, TiMR [112] is a framework for temporal analytics on Big Data used for web advertising.

Because of Big Data, managers can measure, and hence know, radically more about their businesses, and directly translate that knowledge into improved decision making and performance. Using Big Data enables managers to decide on the basis of evidence rather than intuition. For that reason it has the potential to revolutionize management [48].

Business intelligence and analytics (BI&A) and the related field of Big Data analytics have become increasingly important in the academic and business communities over the past two decades. In [49], authors provided a framework that identifies the evolution, applications and emerging research areas of BI&A. According to [113], top-performing organizations use analytics five times more than lower performers. Russom [114] affirmed that using advanced analytics, business can study Big Data to understand the current state of the business and track still-evolving aspects such as customer behavior.

As the complexity of enterprise systems increases, the need for monitoring and analyzing such systems also grows. In [31], authors evaluated six modern (open-source) data stores (HBase, Cassandra, Voldemort, Redis, VoltDB, and MySQL) in the context of application performance monitoring. They evaluated these systems with data and workloads that can be found in application performance monitoring, as well as, online advertisement, power monitoring, and many other use cases.

5.5 Physics and astronomy

High-throughput technologies raise great expectations within the scientific community and far beyond, including discovery of new drugs, a better understanding of the earth's climate, and improved ability to examine history and culture. The growth of data in the "big sciences" such as astronomy, physics, and biology has lead not only to new models of science (collectively known as the "Fourth Paradigm") but also to the emergence of new fields of study such as astroinformatics and computational biology [115].

In the geospatial research area, Cloud computing has attracted increasing attention as a way of solving data-intensive, computing-intensive, and access-intensive geospatial problems [116]. The Land Transformation Model (LTM) is a Land Use Land Cover Change (LUCC) model which was originally developed to simulate local-scale LUCC patterns. The model uses a commercial windows-based GIS program to process and manage spatial data and an artificial neural network (ANN) program within a series of batch routines to learn about spatial patterns in data. In [117], authors provided an overview of a redesigned LTM capable of running at continental scales and a fine (30 m) resolution using a new architecture that employs a windows-based High-Performance Computing (HPC) cluster. Pijanowski et al. [117] provided an overview of the new architecture discussed within the context of modeling LUCC that requires: (1) using an HPC to run a modified version of their LTM; (2) managing large datasets in term of size and quantity of files; (3) integration of tools that are executed using different scripting languages; and (4) a large number of steps necessitating several aspects of job management.

Traditional gazetteers (dictionaries of georeferenced place names) are built and maintained by authoritative mapping agencies. In the age of Big Data, it is possible to construct gazetteers in a data-driven approach by mining rich volunteered geographic information (VGI) from the Web. In [86], authors built a scalable distributed platform and a high-performance geoprocessing workflow based on the Hadoop ecosystem to harvest crowd-sourced gazetteer entries. Using experiments based on geotagged datasets in Flickr, they found that the MapReduce-based platform running on the spatially enabled Hadoop cluster can reduce the processing time compared with traditional desktop-based operations by an order of magnitude. This work offered new insights on enriching future gazetteers with the use of Hadoop clusters, and makes contributions in connecting GIS to the cloud computing environment for the next frontier of Big Geo-Data analytics.

Rapid increases in high-performance computing are feeding the development of larger and more complex data sets in climate research. To understand the Big Data from climate observation in real time is critical for scientific research, agriculture, public transportation weather insurance company, aviation and military activities. This is because it can protect profits of everybody from bad weather. However, conventional climate analysis techniques are inadequate in dealing with the complexities of today's data. In [85], authors described and demonstrated a visual analytics system, called the Exploratory Data Analysis ENvironment (EDEN), with specific application to the analysis of complex earth system simulation data sets. EDEN represents the type of interactive visual analytics tools that are necessary to transform data into insight, thereby improving critical comprehension of earth system processes.

5.6 Biochemistry, genetics and molecular biology

Life Sciences have been highly affected by the generation of large data sets, specifically by overloads of omics information (genomes, epigenomes, transcriptomes and other omics data from cells, tissues and organisms) that necessitates the development of databases and methods for efficient storage, retrieval, integration and analysis of massive data. Omics biology, like most scientific disciplines, is in an era of accelerated

increase of data, so-called Big Data Biology (BDB) [118]. In less than 10 years, the time and cost of sequencing genomes was reduced by a factor of 1 million. Today, personal genomes can be sequenced and mapped faster for a few thousand dollars [119].

Personal genomics is a key enabler for predictive medicine, where a patient's genetic profile can be used to determine the most appropriate medical treatment. According to [120], the integration between hardware and software infrastructures tailored to deal with Big Data in life sciences will become more common in the years to come. Applying Big Data platforms and analytics in the realm of natural science not only has the potential to change lives, but also to save them. Medical/genomics research is thus the dream use case for Big Data technologies which, if unified, are likely to have a profoundly positive impact on mankind. In [120], author gave an overview of the challenges faced by Big Data production, transfer and analysis in genomics. According to the author, the revolutionary changes in Big Data generation and acquisition create profound challenges for storage, transfer and security of information.

Since the proliferation of the Human Genome project at the turn of the Century, there has been an unprecedented proliferation of genomic sequence data. Biology's Big Data sets are now more expensive to store, process and analyze than they are to generate. In [93], authors provided an overview of Cloud computing and Big Data technologies, and discuss how such expertise can be used to deal with biology's Big Data sets. In particular, Hadoop is discussed, together with an overview of the current usage of Hadoop within the bioinformatics community.

According to [44], the ability to protect medical and genomics data in the era of Big Data and a changing privacy landscape is a growing challenge. While Cloud computing is championed as a method for handling such Big Data sets, its perceived insecurity is viewed by many as a key inhibitor to its widespread adoption in the commercial life sciences sector. Indeed, this may explain why its employment has primarily been adopted by research and academic labs.

Molecular biological data have rapidly increased with the recent progress of the Omics field. This situation is also a feature of the ethnomedicinal survey as the number of medicinal plants is estimated to be 40,000–70,000 around the world [121] and many countries utilize these plants as blended herbal medicines, e.g., China (traditional Chinese medicine), Japan (Kampo medicine), India (Ayurveda, Siddha and Unani) and Indonesia (Jamu). In [87], authors reviewed the usage of KNApSAcK Family DB in metabolomics and related area, discussed several statistical methods for handling multivariate data and showed their application on Indonesian blended herbal medicines (Jamu) as a case study. Exploration using Biplot (a multivariate exploration tool) revealed that many plants are rarely utilized while some plants are highly utilized toward specific efficacy. Furthermore, the ingredients of Jamu formulas were modeled using Partial Least Squares Discriminant Analysis (PLS-DA) to predict their efficacy. The plants used in each Jamu medicine served as the predictors, whereas the efficacy of each Jamu provided the responses. This model produced 71.6 % correct classification in predicting efficacy. Permutation test then was used to determine plants that serve as main ingredients in Jamu formula by evaluating the significance of the PLS-DA coefficients. Next, to explain the role of plants that serve as main ingredients in Jamu medicines, information of pharmacological activity of the plants was added

to the predictor block. Then N-PLS-DA model, multiway version of PLS-DA, was utilized to handle the three-dimensional array of the predictor block. The resulting N-PLS-DA model revealed that the effects of some pharmacological activities are specific for certain efficacy and the other activities are diverse toward many efficacies. Mathematical modeling introduced in the study can be utilized in global analysis of Big Data targeting to reveal the underlying biology.

5.7 Social sciences

Computer scientists, physicists, economists, mathematicians, political scientists, bioinformaticists, sociologists, and many others are clamoring for access to massive quantities of information produced by and about people, things and their interactions [122]. The rise of social media along with the progress in computational tools that can process massive amounts of data makes possibly a fundamentally new approach for the study of human beings and society [19].

In [20,21,122], authors focused on Big Data in social media context. Boyd and Crawford [122] declared that historically speaking, collecting data in Social Science has been hard, time consuming, and resource intensive. Therefore, much of the enthusiasm surrounding Big Data stems from the perception that it offers easy access to massive amounts of data. Burgess and Bruns [20] described and reflected on some of the sociotechnical, political and economic aspects of the lists of tweets that may be gathered using the Twitter API. Mahrt and Scharkow [21] gave an overview of methodological issues related to Internet research across different disciplines and communities in the social sciences. A deeper discussion is focused on two aspects: the current debate around the concept of Big Data and the question of finding "meaning" in digital media data.

Social scientists rely on surveys to explain political behavior. Survey research provides the foundation for scientific understanding of voting. According to [123], a new era characterized by high-quality public registration records, national commercial voter lists, and new technologies for Big Data management create an opportunity to revisit survey validation. Through validation, it is possible to learn more about the nature of misreporting and, more importantly, about the nature of political participation. Thus, Ansolabehere and Hersh [123] conducted a fifty-state vote validation and tested that the new Big Data tools facilitate a less costly and more reliable match between survey responses and public records.

Big Data is already transforming the study of how social networks function. Today, social-network research involves mining huge digital data sets of collective behavior online [105]. Nevertheless, the availability of large data sets and the use of analytics clearly implicate privacy concerns. The tasks of ensuring data security and protecting privacy become harder as information is multiplied and shared even more around the world. In [124], authors called for the development of a model where the benefits of data for businesses and researches were balanced against individual privacy rights. Such a model would help determine whether processing can be justified based on legitimate business interest or only subject to individual consent, and whether consent must be structured as opt-in or opt-out.

In [90], authors presented a novel and efficient way of analyzing Big Data sets used in social science research. Authors provided and demonstrated a way to deal with such datasets without the need for high-performance distributed computational facilities. Using an Internet census dataset and with the help of freely available tools and programming libraries (JHDF5 and GIS tools, such as R and Qgis) authors visualized global IP activity in a spatial and time dimension. Authors observed a considerable reduction in storage size of their dataset coupled with a faster processing time.

5.8 Materials science

In materials science, advances in data analysis have initiated a revolution in how researches conduct their work, analyze properties and trend their data, and even discover new materials. Materials science data tend to be particularly heterogeneous in terms of their type and source compared with data encountered in other fields. Materials scientists can easily generate large sets of data from experiments or from large simulations. For example, The Spalation Neutron Source (SNS) [125] at Oak Ridge National Laboratory (ORNL) in Tennessee, a user facility that carries out hundreds of material science experiments each year, is capable of creating hundreds of gigabytes of data in a single experiment [126].

It is critical to experimentalists to reduce these data to something manageable and to access such data in a fast and easy way. ORNL address these issues through ADARA (the Accelerating Data Acquisition, Reduction and Analysis) collaboration project [127]. ADARA provides near-real-time access to result data sets (both raw event data and reduced data) so that instrument scientists and users now obtain live feedback from their experiments. Real-time access to experimental data means scientists can make immediate decisions as to how to steer their experiments, resulting in less wasted time and better data. Experts in high-performance file systems, parallel processing, cluster configuration and management, data management, and Neutron Science have worked together in the ADARA project. The close collaboration between computational experts and experimentalists, combined with Big Data, is the key idea behind the US government's Materials Genome Initiative. The goal of this initiative is to use this collaboration to significantly reduce the time and cost to bring new materials from the laboratory to the marketplace.

Another approach used by researches is the application of artificial intelligence to analyze large data sets, for example, the application of machine learning to atomistic simulation [128]. According to von Lilienfeld [128], the most sophisticated way of analyzing large data sets is using artificial intelligence to detect trends in the data, then to quantity those trends and use them as models that implicitly make use of all the data in the set. The resulting models can then be used to design better materials. With recent advances in experiment, computation, and data analytics, Big Data has the potential to result in significant material advances as they did for genomics [126].

5.9 Medicine

Recently, a report [47] from McKinsey estimated that if US health care could use Big Data creatively and effectively to drive efficiency and quality, then the potential

value from data in the sector could be more than $300 billion in value every year, two-thirds of which would be in the form of reducing national health care expenditures by about 8 %. The predictive power of Big Data is explored in public health. For example, researchers have found a spike in Google search requests for terms like "flu symptoms" and "flu treatments" a couple of weeks before there is an increase in flu patients coming to hospital in a region (and emergency room reports usually lags behind visits by 2 weeks or so) [105].

Big Data analytics is already impacting health decisions and patient care. Healthcare stakeholders such as pharmaceutical-industry experts, payors, and providers are now beginning to analyze Big Data to obtain insights. Although these efforts are in their early stages, they could collectively help the industry address problems related to variability in healthcare quality and escalating healthcare spend. For instance, researches can mine the data to see what treatments are most effective for particular conditions, identify patterns related to drug side effects or hospital readmissions, and gain other important information that can help patients and reduce cost [129].

According to [130], Big Data now enables faster identification of high-risk patients, more effective interventions, and closer monitoring. Also, authors believe the Big Data revolution will uncover many new learning opportunities to fully understand subpopulation efficacy of cancer therapies and the predictive indicators of relapse.

Drug discovery is related to Big Data analytics as the process may require the collection, processing and analysis of extremely large volume of structured and unstructured biomedical data stemming from a wide range of experiments and surveys collected by hospitals, laboratories, pharmaceutical companies or even social media. To analyze such diversity of data types in large volumes for the purpose of drug discovery, it is necessary to develop algorithms that are simple, effective, efficient, and scalable. In [131], authors recognized the importance of Big Data analytics to improve the drug discovery process, contributing to better drug efficacy and safety for pharmaceutical companies and regulators.

A series of breakthroughs in medical science and IT are triggering a convergence between the healthcare industry and the life sciences industry that will quickly lead to more intimate and interactive relations among patients, their doctors and biopharmaceutical companies [132]. The increasing availability and growth rate of biomedical information provides an opportunity for future personalized medicine programs that will significantly improve patient care. The main benefits of applying Big Data analytics in personalized medicine include saving time while improving the overall quality and efficacy of treating disease. In [15], author discussed the major improvements in combining omics and clinical health data in terms of their application to personalized medicine.

5.10 Decision sciences

One of the most critical aspects of using Big Data is its impact on how decisions are made and who gets to make them. When data are scarce, expensive to obtain, or not available in digital form, it makes sense to let well-placed people make decisions, which they do on the basis of experience they have, built-up patterns and relationships

they have observed and internalized. "Intuition" is the label given to this style of inference and decision making. People state their opinions about what the future holds—what is going to happen, how well something will work, and so on—and then plan accordingly. For particular important decisions, these people are typically high up in the organization, or they are expensive outsiders brought in because of their expertise and track records. Many in the Big Data community maintain that companies often make most of their important decisions by relying on "HIPPO"—the highest-paid person's opinion [48].

The primary reason organizations are investing in Big Data is to improve analytic capabilities and make smarter business decisions [133]. Having efficient and effective decision making processes with right data that is transformed to be meaningful information with data-driven discoveries (e.g., analytics) are becoming mainstream processes for companies to run smarter, more agile and efficient businesses [134].

A decision support system (DSS), used in e-commerce, is a term used to describe any computer application that enhances the ability of the user to make decisions [135]. In [134], authors declared that using service-oriented decision support systems (DSS in cloud) is one of the major trends for many organization in hopes of becoming more agile. Therefore, authors proposed a conceptual framework for DSS in cloud.

5.11 Multidisciplinary

According to [111] Data Science is the study and practice of extracting additional knowledge and deriving valuable insights from data. It calls for multidisciplinary approaches that incorporate theories and methods from many fields including mathematics, statistics, pattern recognition, knowledge engineering, machine learning, high-performance computing, etc. It brings about a number of new research topics on data itself such as data life cycle management, data privacy solution, elastic data computing, spatial-temporal nature and social aspects of Big Data. From a multidisciplinary perspective, Big Data is a newly emerging field that encompasses a number of disciplines. It depends on inter-disciplinary study and practice that can help data science gain a competitive edge.

In this subsection, we include the papers related to Smart Cities and Knowledge Discovery from Data (KDD) because they involve the application of many fields.

Smart City There are several papers related to Smart Cities. An Smart City is defined as a city connecting the physical infrastructure, the IT infrastructure, the social infrastructure, and the business infrastructure to leverage the collective intelligence of the city [136]. From the ICT perspective, the possibility of realization of Smart Cities is being enabled by smarter hardware (smart phones, sensor nets, smart household appliances, etc.), which can organize in an 'Internet of Things' (IoT) and thus becomes a major source of user environment-specific data [137]. In [138], authors showed that Big Data is a promising field for Smart City business exploitation. According to [139], the smart cities need to have a robust and scalable video surveillance infrastructure. Therefore, Dey et al. [139] showed that large scale video surveillance backend can be integrated to the open source cloud-based data stores available in the Big Data trend. Jara et al. [140] provided some data-driven models as a proof of concept of the potential of the

Big Data Analytic tool for Smart Cities. Khan et al. [137] proposed an architecture for Cloud-based Big Data analytics focused towards the smart city use cases. In [141], authors proposed a system architecture for Smart City applications which employs ontology reasoning and distributed stream processing framework on the cloud.

Knowledge Discovery from Data (KDD) KDD [142] refers to a set of activities designed to extract new knowledge from complex datasets. The KDD process is often interdisciplinary and spans computer science, statistics, visualization, and domain expertise. It is composed of multiple stages, covering from data analytics to data mining. KDD has become strategically important for large business enterprises, government organizations, and research institutions. Effective KDD requires effective organizational and technological practices to be in place. Therefore, in [54], authors outlined empirically derived principles for the establishment of effective architectures for knowledge discovery over Big Data.

The most fundamental challenge for Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions [143]. In [4], Big Data mining is defined as the capability of extracting useful information from these large datasets or streams of data. Fan and Bifet [4] also presented a broad overview of the topic, its current status, controversy, and a forecast to the future. They introduce four articles, written by influential scientists in the field, covering the most interesting and state-of-the-art topics on Big Data mining.

Efficient clusterization constitutes one of the important components of data mining. The objective of [144] is to bring forward a simple and efficacious tool for clustering of diverse information items in a data stream mode. Furthermore, Berkovich and Liao [144] revealed a parallel between the computational model integrating Big Data streams and the organization of information processing in the brain. In [72], authors presented implementation details for doing both correlation analysis and association rule mining over streams. Specifically, they implemented Pearson-Product Moment Correlation for analytics and Apriori and FP Growth algorithms for stream mining inside a popular event stream processing engine called Esper. Moens et al. [145] investigated the applicability of Frequent Itemset Mining (FIM) techniques on the MapReduce platform.

5.12 Arts and humanities

The emergence of social media in the middle of 2000s created opportunities to study social and cultural processes and dynamics in new ways. Manovich [19] addressed some of the theoretical aspects and practical issues raised by the possibility of using massive amounts of social and cultural data in humanities sciences.

Humanities communities and projects are characterized by multi-lateral and often global collaborations between researches from all over the world that need to be engaged into collaborative groups/communities and supported by collaborative infrastructure to share data, discovery/research results and cooperatively evaluate results. The current trend to digitalize all currently collected physical artifacts will create in the near future a huge amount of data that must be widely and openly accessible [24].

According to this literature review, Computer Science is the most important area of application for Big Data because most of the works use technologies like databases and Cloud computing to solve the problems of storing and processing Big Data. Also Big Data is taking a key role in all the areas, where the most important challenge addressed is data analysis. The main subfield of Mathematics for Big Data research is Statistics because it provides several techniques for data mining and analytics. In areas like Physics and Astronomy; Materials Science, and Biochemistry, Genetics and Molecular Biology using Big Data analytics it is possible to extract insights of massive and heterogenous experiment data in real time. In Social Sciences and Humanities, now it is common to mine social media to discover human behavior and relationships. Organizations are able to make data-driven decisions in a timely manner if they have efficient Big Data analytics tools and data scientists.

6 Big Data tools: techniques, frameworks, tools and programming languages support

Several big data tools have been developed to manage and analyze Big Data. Big Data requires exceptional tools to efficiently process large quantities of data within tolerable elapsed times. The Big Data phenomenon is intrinsically related to the open-source software revolution. Large companies such as Facebook, Yahoo!, Twitter, LinkedIn benefit and contribute to open-source projects [4]. In this section, we present the most used techniques, frameworks, tools and programming languages to deal with Big Data.

6.1 Big Data techniques

There are many techniques that can be used to analyze Big Data. In the following, we detail some specific techniques that are frequently used in Big Data projects.

1. *Statistics* The science of the collection, organization, and interpretation of data, including the design of surveys and experiments. Statistical techniques are often used to make judgments about what relationships between variables could have occurred by chance (the "null hypothesis"), and what relationships between variables likely result from some kind of underlying causal relationships (i.e., that are "statistically significant"). Statistical techniques are also used to reduce the likelihood of Type I errors ("false positives") and Type II errors ("false negatives"). An example of an application is A/B testing to determine what types of marketing material will most increase revenue [47].
2. *Machine learning* Its goal is to turn observational data into a *model* that can be used to predict for or explain yet unseen data. Increasingly, machine learning is at the core of data analysis for actionable business insights and optimizations. Today, machine learning is deployed widely: recommender systems drive the sales of most online shops; classifiers help keep spam out of email accounts; computational advertising systems drive revenues; content recommenders provide targeted user experiences. Machine learning is also enabling scientists to interpret and draw

new insights from massive datasets in many domains, including such fields as astronomy, high-energy physics, and computational biology [61].

3. *Data mining* It explores and analyzes large quantities of data to discover meaningful patterns. The field of data mining is interdisciplinary. Data mining uses a combination of pattern-recognition rules, statistical rules, as well as rules drawn from machine learning. Data mining has wide applicability in intelligence and security analysis, genetics, the social and natural sciences, and business. Some of data mining techniques include association rule learning, cluster analysis, classification and regression [146].
4. *Signal processing* Big Data challenges offer ample opportunities for signal processing research, where data-driven statistical learning algorithms are envisioned to facilitate distributed and real-time analytics. Both classical and modern signal processing techniques, such as principal component analysis, dictionary learning and compressive sampling, have already placed significant emphasis on time/data adaptivity, robustness, as well as compression and dimensionality reduction. While the principal role of computer science in Big Data research is undeniable, the nature and scope of the emerging data science field is certainly multidisciplinary and welcomes signal processing expertise and its recent advances [147].
5. *Visualization techniques* Multidimensional visualization techniques are widely used as effective information abstraction tools for analysis of high-dimensional databases in knowledge discovery, information awareness and decision making process [148]. Displaying the distribution of the high-dimensional data through low-dimension visual space, researches can quickly become aware of the information like feature, relationship, cluster and trend. The targeted visualization can reflect the nature of data effectively [74].

6.2 Big Data tools

Big Data tools are categorized into three alternative paradigms according to the kind of analysis: (1) batch analysis where data are firstly stored and then analyzed; (2) stream processing which analyzes data as soon as possible to derive its results, and (3) interactive analysis which processes the data allowing users to undertake their own analysis of information [11,24]. Table 1 presents the classification of the frameworks, tools and programming languages for Big Data applications.

6.2.1 Big Data tools based on batch analysis

Google MapReduce [32] is a Java-based framework designed to run over a cluster of machines in a distributed way. It is a batch-oriented parallel computing model. Figure 8 presented in [149] depicts the overall strategy. One node is elected to be the master responsible for assigning the work, while the rest are workers. The input data are divided into splits and the master assigns splits to *Map* workers. Each worker processes the corresponding input split, generates key/value pairs and writes them to intermediate files (on disk or in memory). The master notifies the *Reduce* workers

Table 1 Frameworks, tools and programming languages for Big Data applications

Name	Specified use	Advantages	Programming language
Batch analysis			
Google mapreduce	Data processing on large clusters	Simple, scalable, fault tolerant	Java
Apache Hadoop	Infrastructure and platform	Scalable, reliability, completeness, extensibility	Java, Python, R, HiveQL, Pig Latin
Microsoft Dryad	Infrastructure and platform	Good programmability, fault-tolerant	DryadLINQ, SCOPE
Apache Mahout	Machine learning algorithms	Scalable, good maturity	Java
Stream analysis			
Apache Storm	Real-time computation system	Fault-tolerant, simple, scalable, efficient, easy to use and operate	All
Apache S4	Stream computing platform	Proven, scalable, fault-tolerant, extensible	Java
Apache Spark	Engine for large-scale data processing	Fast, general, easy of use	Scala, Java, Python
MOA	Framework for data stream mining	Extensible, scalable	Java
Interactive analysis			
Apache Drill	SQL query engine for Hadoop and NoSQL	Agile, flexible, familiar	SQL
SpagoBI	Business Intelligence	Agile, real time BI on Big Data streaming	Java
D3.js	Interactive	Scalable	JavaScript

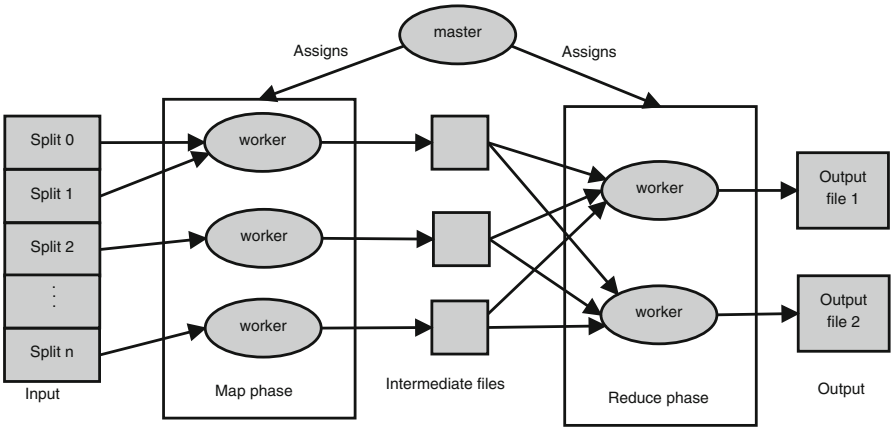

Fig. 8 MapReduce strategy

about the location of the intermediate files and the *Reduce* workers read data, process it according to the *Reduce* function, and finally, write data to output files.

MapReduce has been widely adopted and received extensive attention from both academia and industry due to its promising capability. Simplicity, scalability, and fault tolerance are three main salient features of the *MapReduce* framework [99].

MapReduce is a good example of Big Data processing in a cloud environment. It allows an unexperienced programmer to develop parallel programs and create a program capable of using computers in a cloud [150]. MapReduce accelerates the processing of large amounts of data in a cloud; thus, MapReduce, is the preferred computation model of cloud providers [151].

Despite such a great success and benefits, MapReduce exhibits several limitations, making it unsuitable for the overall spectrum of needs for large-scale data processing. In particular, the *MapReduce* paradigm is affected by several performance limitations, introducing high latency in data access and making it not suitable for interactive use. This makes the *MapReduce* paradigm unsuitable for event-based online Big Data processing architectures, and motivates the need of investigating other different paradigms and novel platforms for large-scale event stream-driven analytics solutions [152]. Also, iterative algorithms are not able to obtain a good performance as they need to launch a *MapReduce* job for each iteration notably increasing the computation time due to the overhead [153].

Apache Hadoop [33] is a collection of Java-based open-source software inspired by *Google BigTable* [96], *Google File System* [154] and *MapReduce*. It includes a *MapReduce* component (for distributed computation) and a scalable storage component, *Hadoop File System (HDFS)*, that can often replace costly SAN devices [54]. While *HDFS* is used predominantly as the distributed file system on *Hadoop*, others file systems like *Amazon S3* are also supported. Figure 9 illustrates the layers found in the software architecture of Hadoop stack.

Big Data provides users the ability to use commodity computing to process distributed queries across multiple datasets and return resultant sets in a timely manner. Cloud computing provides the underlying engine through the use of Hadoop [150].

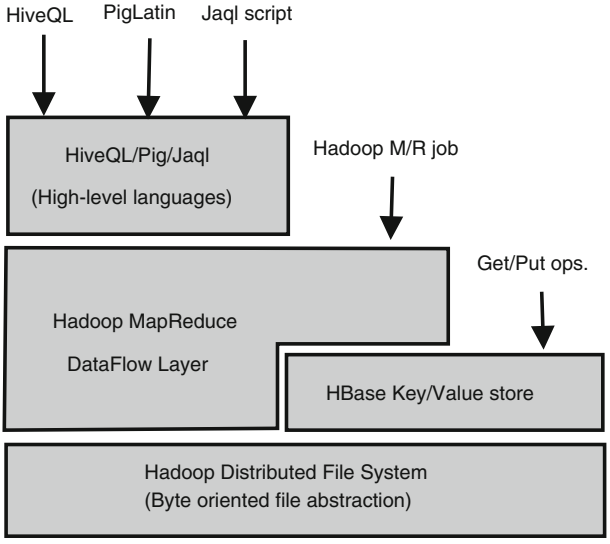

Fig. 9 Hadoop software stack [65]

Analytics with *Hadoop* involves loading data as files into the distributed file system, and then running parallel *MapReduce* computations on the data. *MapReduce* computations in *Hadoop* can be expressed directly in general-purpose programming languages like *Java* or *Python*, domain-specific languages like *R*, or generated automatically from SQL-like declarative languages like *HiveQL* and *Pig Latin*. This coverage of the language spectrum makes *Hadoop* well suited for deep analytics.

Also, an unheralded aspect of *Hadoop* is its extensibility, i.e., the ease with which many of *Hadoop*'s core components like the scheduler, storage subsystem, input/output data formats, data partitioner, compression algorithms, caching layer, and monitoring can be customized and replaced [53]. *Hadoop* sub-projects such as *Hive* and *HBase* offer additional data management solutions for storing unstructured and semi-structured data sets [54].

The *Hadoop* system has quickly become a "gold-standard" in industry as a highly scalable data-intensive *MapReduce* platform, and it is now widely used for use cases including Web indexing, clickstream and log analysis, and certain large-scale information extraction and machine learning tasks [65]. *Hadoop* is used by leading technology companies such as Amazon, Facebook, LinkedIn, and Twitter [93,155].

Nevertheless, *Hadoop* also has some disadvantages. One drawback is that programming *Hadoop* is not a trivial task; requiring significant expertise in *Java* to develop parallelized programs. Efforts have been made to simplify this process, even in the technology sector, with developed software libraries such as Hive to add an "SQL" like interface that will generate parallelized *Hadoop* jobs in the background. *Python* streaming has also been made available to circumvent complex *Java* programming by wrapping in *Python*, a more lightweight scripting language. Another problem is that raw *Hadoop*-based systems usually lack powerful statistics and visualization tools

[64]. Finally, *Hadoop* is powerful only if in the right hands and still difficult to set up, use and maintain [93].

Microsoft Dryad [156] includes a parallel runtime system called *Dryad* and two higher level programming models, *DryadLINQ* [157] and the SQL-like *SCOPE* language [39], which utilize *Dryad* under the covers [65]. *Dryad* is an infrastructure which allows a programmer to use the resources of a computer cluster or data center for running data-parallel programs. A *Dryad* programmer can use thousands of machines, each of them with multiples processors or cores, without knowing anything about concurrent programming.

A *Dryad* programmer writes several sequential programs and connects them using one-way channels. The computation is structured as a directed graph: programs are graph vertices, while the channels are graph edges. A *Dryad* job is a graph generator which can synthesize any directed acyclic graph. These graphs can change during execution, in response to important events in the computation.

Dryad is quite expressive. It completely subsumes other computation frameworks, such as *Google MapReduce*, or the relational algebra. Moreover, *Dryad* handles job creation and management, resource management, job monitoring and visualization, fault-tolerance, re-execution, scheduling and accounting.

Apache Mahout [158] is a scalable data mining and machine learning open-source software based mainly in *Hadoop*. Currently *Mahout* supports mainly three use cases: recommendation, clustering and classification. The *Apache Mahout* project aims to make building intelligent applications easier and faster.

6.2.2 Big Data tools for stream analysis

Data streams are widely used in financial analysis, online trading, medical testing, and so on. Static knowledge discovery methods cannot adapt to the characteristics of dynamic data streams, such as continuity, variability, rapidity, and infinity, and can easily lead to the loss of useful information. Therefore, effective theoretical and technical frameworks are needed to support data stream mining [60].

Real-time streaming platforms are tools in Big Data that are able to handle large volumes of data, arriving to the system at high velocities, using compute clusters to balance the workload. Those systems inherit some properties of Message Passing Interface (MPI) clusters but add scalability to the feature set. They are able to rebalance the workload if too many messages need to be processed in a certain compute node. There are four systems that can be considered enough for productive environments: *Project Storm* (developed at Twitter), *S4* (developed at Yahoo), *project Spark*, and *MOA* (Massive Online Analysis).

Apache Storm [159] is a free and open source distributed real-time computation system. *Storm* makes it easy to reliably process unbounded streams of data, doing for real-time processing what *Hadoop* did for batch processing. *Storm* is simple and it can be used with any programming language. *Storm* has many use cases: real-time analytics, online machine learning, continuous computation, and more. *Storm* is fast: a benchmark clocked it at over a million tuples processed per second per node. It is scalable, fault tolerant and easy to set up and operate.

Apache S4 [160] is a general-purpose, distributed, scalable, fault-tolerant, pluggable platform that allows programmers to easily develop applications for processing continuous unbounded streams of data. In *S4*, computation is performed by Processing Elements (PEs) and messages are transmitted between them in the form of data events. The state of each PE is inaccessible to other PEs; event emission and consumption is the only mode of interaction between PEs. *S4* provides the capability to route events to appropriate PEs and to create new instances of PEs. These aspects of the design provide the properties of encapsulation and location transparency. Developers can write PEs in the *Java* programming language.

Apache Spark [161] is an in-memory cluster computing engine that provides support for a variety of workloads, including batch, streaming, and iterative computations. In a relative short time, *Spark* has become the most active Big Data project in the open-source community, and is already being used by over one hundreds of companies and research institutions.

Spark was developed in the UC Berkeley AMPLab and it is used to run large-scale applications such as spam filtering and traffic prediction. *Spark* provides primitives for in-memory cluster computing and APIs in *Scala*, *Java* and *Python*.

MOA [162] is a stream data mining open-source software to perform data mining in real time. It has implementations of classification, regression, clustering and frequent item set mining and frequent graph mining. *SAMOA* is a software project for distributed stream mining that will combine *S4* and *Storm* with *MOA* [4].

6.2.3 Big Data tools based on interactive analysis

Apache Drill [163] is a framework that supports data-intensive distributed applications for interactive analysis of large-scale datasets. *Drill* is a version of *Google's Dremel* System, which is scalable, interactive ad hoc query system for analysis of read-only nested data. Furthermore, its goal is to be able to scale 10,000 servers or more and to be able to process petabytes of data and trillions of records in seconds.

SpagoBI [164] manages large volumes of heterogeneous structured and unstructured data, to perform real-time Business Intelligence on Big Data streaming and give meaning to data through the semantic analysis. *SpagoBI* supplies meaningful data insights through the main concept of persistable and schedulable datasets, and using tools such as self-service BI, ad hoc reporting, interactive dashboard and exploratory analysis. An agile and interactive way to perform insights over data and to refine one's own queries on Big Data can be obtained through the self-service BI feature provided by *SpagoBI* suite. In fact, *SpagoBI* provides a Query By Example engine (QbE) which allows power users to create their own queries in a graphical and smart way over their datasets. Several filter conditions and aggregations can be applied on the datasets, so that results can instantly been used by ad hoc reporting engine for further graphical analysis. Newly created datasets can be saved and used by the other *SpagoBI* engines, as well as scheduled and persisted.

D3 [165] is a *JavaScript*-based library for manipulating data contained in various document types, but this simply enables it to create qualitative, interactive visualiza-

tions for a large variety of datasets. *D3* is built on familiar web standards like *HTML* and *CSS*, making it browser friendly for end users and enabling interface designers to style data with a standard like *CSS* rather than a proprietary format or expensive graphic design tool like *Adobe Photoshop*. These familiar standards also enable technical communications to focus on choosing appropriate visualization(s) for the data and audience involved, and enabling users to access it through commonly available tools (like a browser), rather than learning how to use proprietary tool skills that do not extend to other software.

7 Challenges and trends on Big Data

After the analysis of the importance of Big Data and the classification of research papers in the domains of application, we identify three trends on Big Data which represents many challenges for researchers. Big Data is conceived as the powerful tool to exploit all the potential of the Smart manufacturing, the Internet of Things and the Smart Cities. In this section, we describe these trends and challenges in detail.

7.1 Smart manufacturing

Industry stands to gain many benefits from Big Data as more sophisticated and automated data analytics technologies are developed. For example, sensors in each Ford Focus Electric car produce streams of data while the car is driven and when it is parked. Not only is the driver informed, but Ford and others can analyze the information to improve the vehicle and to deal with issues like placing charging stations or avoiding grid overload. Smart manufacturing has the potential of fundamentally changing how products are invented, manufactured, shipped, and sold.

A number of developments under way now aim to realize the greater potential of Big Data in smart manufacturing. The Smart Manufacturing Leadership Coalition [166] aims to overcome the barriers to the development and deployment of smart manufacturing systems. The members of the coalition include stakeholders from industry, academia, government, and manufacturing. Teams of members are developing a shared, open-architecture infrastructure called the Smart Manufacturing Platform that allows manufacturing to assemble a combination of controls, models, and productivity metrics to customize modeling and control systems, making use of Big Data flows from fully instrumented plants in real-time. The development of smart manufacturing platform is intended to help management across the manufacturing ecosystem, and to forecast activities that, for example, can be used to slash cycle times and reduce the risks of product development. The infrastructure will enable third parties to develop networked industrial applications that provide tools for resource metering, automated data generation, intuitive user interfaces, and Big Data analytics. Future products can be outfitted with sensors that connect to the cloud, enabling after-sales service offerings. There could be an option in cars, for instance, that will alert drivers or service centers when maintenance is needed. Data showing how customer use products can suggest improvement in design and manufacturing.

Smart manufacturing is likely to evolve into the new paradigm of cognitive manufacturing, in which machining and measurements are merged to form more flexible and controlled environments. When unforeseen changes or significant alterations happen, machining process planning systems receive online measurement results, make decisions, and adjust machining operations accordingly in real time.

It is conceivable that one day machines and process factory will be equipped with capabilities that allow them to assess and increase their scope of operation autonomously. This changes the manufacturing system from a deterministic one, where all planning is carried out off-line, to a dynamic one that can determine and reason about processes, plans, and operations [167].

7.2 The internet of things

The IoT has gained significant attention over the last few years. With the advances in sensor hardware technology and cheap materials, sensors are expected to be attached to all the objects around us, so these can communicate with each other with minimum human intervention. Understanding sensor data is one of the main challenges that the IoT would face. This vision has been supported and heavily invested by governments, interest groups, companies and researches institutes. For example, context awareness has been identified as an important IoT research need by the Cluster of European Research Projects on the IoT (CERP-IoT) [168] funded by the European Union (EU). The EU has allocated a time frame for research and development into context-aware computing focused on IoT to be carried out during 2015–2020 [169].

IoT is generating prodigious amount of data, increasing sophisticated analytic mechanisms and tools, that are providing insight that allow us to operate the machines in entirely new ways, more effectively, and in a collaborative way. The power of data provided by all the resources that are being connected to Internet will bring a new conception of the world, where the Big Data analysis is required to take advantage of its potential for high-level modeling and knowledge engineering [140].

7.3 Smart cities

Although several papers have discussed the importance of Big Data in Smart Cities and Cloud computing is recognized as a great opportunity to manage, analyze and process the Big Data generated by cities, there is a need of new tools and services to process and analyze city data effectively [137]. The amount of sensors in Smart Grids, combined with those in Smart Buildings, will vastly increase the data influx in the near future [170].

Smart cities of tomorrow will rely not only on sensors within the city infrastructure, but also on a large number of devices that will willingly sense and integrate their data into technological platforms used for introspection into the habits and situations of individuals and city-large communities. Predictions say that cities will generate over 4.1 terabytes per day per square kilometer of urbanized land area by 2016. Handling efficiently such amounts of data is already a challenge [22].

With the majority of the civilization of today living in cities, problems are prone to increase, and thus solutions are urgently needed. Compelled by this necessity, city halls and political decision maker have becoming very alert, with small and large technology companies alike hoping to jump on the smart city wagon of the twenty-first century. The smart city market is estimated to be of hundreds of billions of dollars by 2020 [138].

In the future Big Data will have higher impact on several aspects of our everyday life and our behavior. More people will live in Smart Cities, where all the devices will communicate with each other without human intervention. Fabs and products will be more "intelligent". IoT, Smart Cities and Smart manufacturing will generate data in unseen proportions, which must be analyzed in real time. Therefore, it is necessary to start to develop the tools and abilities that will allow to get value from Big Data in this context.

8 Future and open issues on Big Data

Although Big Data is a hot topic, it is necessary to solve many issues with respect to evaluation of Big Data systems, data privacy, data cleaning and the future development of predictive algorithms.

8.1 Evaluation of Big Data systems

A Big Data benchmark suite is needed eagerly by customers, industry and academia recently [171]. Benchmarks are important tools for evaluating an information system. While the performance of traditional database systems is well understood and measured by long-established institutions such as Transaction Processing Performance Council (TCP), there is neither a clear definition of the performance of Big Data systems nor a generally agreed upon metric for comparing these systems [172].

Benchmarking Big Data systems are much more challenging than ever before. First, Big Data systems are still in their infant stage and consequently they are not well understood [173]. Second, the complexity and diversity of Big Data systems and their rapid evolution give rise to various challenges about benchmark design to test such systems efficiently and successfully [174]. Considering the broad use of Big Data systems, for the sake of fairness, big data benchmarks must include diversity of data and workloads, which is the prerequisite for evaluating Big Data systems and architecture. Most of the state-of-the art Big Data benchmarking efforts target specific types of applications or system software stacks [175]. It is, therefore, unclear whether these benchmarks can be used to precisely evaluate the performance of Big Data systems.

Also, current benchmarks for spatial computing remain limited to small data sizes and only a portion of current popular data types. New benchmarks need to be built around Spatial Big Datasets, incorporating all four data types (raster, vector, network, spatio-temporal), while covering a wide variety of use-cases from emergency management, location-based services, advanced routing services, to mention a few. New performance metrics, both functional (e.g., mobile interactions per second) and non-functional (e.g., disaster resilience footprint), will facilitate comparison between new systems being created and promoted by various spatial computing vendors [176].

8.2 Data privacy

In Big Data applications, data privacy is one of the most concerned issues because processing large-scale privacy sensitive data sets often requires computation resources provisioned by public cloud services. Sub-tree data anonymization is a widely adopted scheme to anonymize data sets for privacy preservation. Top-Down Specialization [46] and Bottom-Up Generalization are two ways to fulfill sub-tree anonymization. However, existing approaches for sub-tree anonymization fall short of parallelization capability, thereby lacking scalability in handling Big Data in cloud [99].

According to [177], the biggest challenge that will need to be addressed is the balance between data privacy and reproducible research. How to balance user privacy and reproducible research results is a decision that will impact all of us both as consumers and also as data users. It requires us as a society to define what our right to privacy actually includes. This balance of competing interests includes questions of legality as well as technology. Regardless of the outcome, there will be significant impacts on organizations in these ways: to comply with research standards, adhere to legislation, and fund enforcement of the laws.

8.3 Big Data cleaning

Data cleaning is, in fact, a lively subject that has played an important part in the history of data management and data analytics, and it is still undergoing rapid development. Moreover, data cleaning is considered as a main challenge in the era of Big Data due to the increasing volume, velocity and variety of data in many applications. Some of the issues to be addressed or improved to meet practical needs are: (1) tool selection. Given a database and a wide range of data cleaning tools, the first challenge question is which tool to pick for the given specific task; (2) rule discovery. Rules discovered by automatic algorithms are far for clean themselves. Therefore, often times, manually selecting/cleaning thousands of discovered rules is a must, yet a difficult process; (3) usability. In fact, usability has been identified as an important feature of data management, since it is challenging for humans to interact with machines. This problem is harder when comes to the specific topic of data cleaning, since given detected errors, there is normally no evidence that which values are correct and which are wrong, even for humans. Hence, more efforts should be put to usability of data cleaning systems so as to effectively involve users as first-class citizens [178].

8.4 Future development of predictive algorithms

The science of modeling will march forward at an even faster rate than it has already. New predictive algorithms will be developed that in special situations will perform better than stable algorithms that have been around for many years. With the exception of the occasional breakthrough, Big Data practitioners will need to be well versed in the traditional methods and be capable of evaluating new challengers in objective ways. One thing is certain: the number of algorithms claiming huge advances will increase much faster than the ones that actually move the industry forward. The newest algo-

rithm inventions will build from these great ideas and show incremental improvement for the next 5–10 years. The development of algorithms will continue as more success is found in data mining and leveraging analytics within organizations [177].

9 Conclusions and future work

Big Data is taking a relevant role in several fields and this trend is expected to increase in the future. In this paper, we reviewed research papers of Big Data from 2010 to 2014 and classified the works into twelve areas of application: (1) Computer Science; (2) Engineering; (3) Mathematics; (4) Business, Management and Accounting; (5) Physics and Astronomy; (6) Biochemistry, Genetics and Molecular Biology; (7) Social Sciences; (8) Material Sciences; (9) Medicine; (10) Decision Sciences; (11) Multidisciplinary; and (12) Arts and Humanities. In addition, we identify sub domains of key importance for Big Data research. We conclude that Computer Science is the most important area for Big Data research.

Moreover, we provided a classification of Big Data papers into six challenges addressed: (1) Capture; (2) Store; (3) Search; (4) Share; (5) Analysis; and (6) Visualization. Our conclusion was that Analysis is the most important challenge for Big Data research because it is being applied in all the areas of knowledge to gain insights of the value of Big Data. Also, we determined the most used frameworks (i.e., Hadoop, Mahout, Storm, S4, Spark, and Drill from Apache; MapReduce from Google; Dryad from Microsoft; MOA from the University of Waikato, New Zealand; SpagoBI from OW2 Consortium, and D3.js from Michael Bostock and community) and programming languages for Big Data applications. The importance of this paper resides in that it can provide researches and practitioners an overview about the state-of-the-art in Big Data applications, and give them a reference of the new trends and most significant areas of opportunity in which Big Data can be applied.

In the future we expect to improve this research in three aspects: (1) providing a more specialized review of frameworks and programming languages for Big Data analysis to extend the information about the features, benefits and limitations of each framework and language to provide a set of guidelines, and proofs of concept to expose the functionality of each programming language or framework; (2) including a review of works published in workshops, technical reports and white papers, and (3) considering papers from other scientific databases.

Acknowledgments The authors are very grateful to National Technological of Mexico for supporting this work. Also, this research paper was sponsored by the National Council of Science and Technology (CONACYT), as well as by the Public Education Secretary (SEP) through PRODEP.

References

1. Talia D (2013) Clouds for scalable big data analytics. Computer 46(5):98–101
2. Lomotey RK, Deters R (2014) Towards knowledge discovery in big data. In: Proceeding of the 8th international symposium on service oriented system engineering. IEEE Computer Society, pp 181–191

3. Laney D (2001) 3-D management: controlling data volume, velocity, and variety. Application Delivery Strategies. META Group Original Research Note 949, pp 1–4. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf. Accessed 11 Aug 2015
4. Fan W, Bifet A (2012) Mining big data: current status, and forecast to the future. SIGKDD Explor 14(2):1–5
5. Begoli E (2012) A short survey on the state of the art in architectures and platforms for large scale data analysis and knowledge discovery from data. In: Proceeding of the joint working IEEE/IFIP Conference on software architecture (WICSA) and European conference on software architecture (ECSA), pp 177–183
6. Sagiroglu S, Sinanc D (2013) Big data: a review. In: Proceeding of the 2013 international conference on collaboration technologies and systems (CTS). IEEE Computer Society, pp 42–47
7. Katal A, Wazid M, Goudar RH (2013) Big data: issues, challenges, tools and good practices, In: Sixth international conference on contemporary computing (IC3), pp 404–409
8. Kaisler S, Armour F, Espinosa JA, Money W (2013) Big data: issues and challenges moving forward. In: Proceeding of the 46th Hawaii international conference on system sciences, pp 995–1004
9. Louridas P, Ebert C (2013) Embedded Analytics and Statistics for Big Data. IEEE Softw 30(6):33–39
10. Kambatla K, Kollias G, Kumar V, Grama A (2014) Trends in big data analytics. J Parallel Distrib Comput 74(7):2561–2573
11. Chen PCL, Zhang C-Y (2014) Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inf Sci Elsevier 275:314–347
12. Chen M, Mao S, Liu Y (2014) Big data: a survey. Mob Netw Appl 19:171–209
13. Halevi G, Moed H (2012) The evolution of big data as a research and scientific topic: overview of the literature. Res Trends 30:3–6
14. Lee J, Lapira E, Bagheri B, Kao H (2013) Recent advances and trends in predictive manufacturing systems in big data environment. Manufact Lett 1(1):38–41
15. Costa FF (2014) Big data in biomedicine. Drug Discov Today Elsevier 19(4):433–440
16. Patel AB, Birla M, Nair U (2012) Addressing big data problem using Hadoop and MapReduce. In: NIRMA University international conference on engineering, NuiCONE, pp 1–5
17. Brown B, Chui M, Manyika J (2011) Are you Ready for the Era of 'Big Data'? McKinsey Q 4:24–35
18. Gantz J, Reinsel D (2011) Extracting value from chaos. IDC IVIEW: IDC Analyze the Future 1142:1–12
19. Manovich L (2012) Trending: the promises and the challenges of big social data. In: Gold MK (ed) Debates in the digital humanities. University of Minessota Press, Minneapolis, pp 460–475
20. Burgess J, Bruns A (2012) Twitter archives and the challenges of "Big Social Data" for media and communication research. M/C J 15(5):1–7
21. Mahrt M, Scharkow M (2013) The value of big data in digital media research. J Broadcast Electron Media 57(1):20–33
22. Dobre C, Xhafa F (2014) Intelligent services for big data science. Future Gener Comput Syst 37:267–281
23. Laurila JK, Gatica-Perez D, Aad I et al (2013) From big smartphone data to worldwide research: the mobile data challenge. Pervasive Mob Comput 9(6):752–771
24. Demchenko Y, Grosso P, de Laat C, Membrey P (2013) Addressing Big Data Issues in Scientific Data Infrastructure. In: International Conference on Collaboration Technologies and Systems (CTS). IEEE Computer Society
25. Hu H, Wen Y, Chua T-S, Li X (2014) Toward scalable systems for big data analytics: a technology tutorial. IEEE Access 2:652–687
26. Agrawal D, Bernstein P, Bertino E et al (2011) Challenges and Opportunities with Big Data 2011-1. Cyber Center Technical Reports, (Paper 1). Retrieved from http://dpcs.lib.purdue.edu/cctech/1
27. He Y, Lee R, Huai Y et al. (2011) RCFile: a fast and space-efficient data placement structure in mapreduce-based warehouse systems. In: Proceeding of the IEEE international conference on data engineering (ICDE), pp 1199–1208
28. Lakshman A, Malik P (2010) Cassandra: a decentralized structured storage system. ACM SIGOPS Oper Syst Rev 44(2):35–40
29. The Apache Software Foundation. Apache HBase. http://hbase.apache.org
30. Voldemort. Project Voldemort. http://project-voldemort.com

31. Rabl T, Sadoghi M, Jacobsen H-A et al (2012) Solving big data challenges for enterprise application performance management. J VLDB Endow 5(12):1724–1735
32. Dean J, Ghemawat S (2008) MapReduce: Simplified Data Processing on Large Clusters. Commun ACM 51(1):107–113
33. White T (2009) Hadoop: the definite guide, 1st edn. OReilly Media Inc, Sebastopol
34. Schadt E, Linderman MD, Sorenson J et al (2010) Computational Solutions to Large-Scale Data Management and Analysis. Nat Rev Genet 11:647–657
35. Marx V (2013) Biology: The Big Challenges of Big Data. Nature 498:255–260
36. Gantz J, Reinsel D (2012) The digital Universe in 2020: big data, bigger digital shadows, and biggest growth in the far east. IDC IVIEW: IDC Analyze the Future 1414_v3:1–16
37. Thusoo A, Sarma JS, Jain N et al (2010) Hive-A petabyte scale data Warehouse using Hadoop. In: Proceeding of ICDE. IEEE, pp 996–1005
38. Olston C, Reed B, Srivastava U et al (2008) Pig Latin: a not-so-foreign language for data processing. In: Proceeding of the SIGMOD conference, pp 1099–1110
39. Chaiken R, Jenkins B, Larson PA et al (2008) SCOPE: easy and efficient parallel processing of massive data sets. Proc VLDB Endow 1(2):1265–1276
40. Chaudhuri S (2012) What next? A Half-Dozen data management research goals for big data and the cloud. In: Proceeding of the symposium on principles of database systems (PODS). ACM, pp 1–4
41. Naseer A, Laera L, Matsutsuka T (2013) Enterprise BigGraph. In: 46th Hawaii international conference on system sciences. IEEE Computer Society, pp 1005–1014
42. Wood D (2012) Linking enterprise data. Springer, New York
43. Hampton SE, Strasser CA et al (2013) Big data and the future of ecology. Front Ecol Environ 11(3):156–162
44. Schadt E (2012) The changing privacy landscape in the Era of big data. Mol Syst Biol 8(612):1–3
45. Ranganathan S, Schönbach C, Kelso J et al (2011) Towards big data science in the decade ahead from 10 years of InCoB and the 1st ISCB-Asia joint conference. BMC Inf 12(13):1–4
46. Zhang X, Yang LT, Liu C, Chen J (2014) A scalable two-phase top-down specialization approach for data anonymization using mapreduce on cloud. IEEE Trans Parallel Distrib Syst 25(2):363–373
47. Manyika J, Chui M, Brown B et al (2011) Big data: the next frontier for innovation, competition and productivity. McKinsey Global Institute, New York
48. McAfee A, Brynjolfsson E (2012) Big data: the management revolution. Harv Bus Rev 90(10):60–68
49. Chen H, Chiang RHL, Storey VC (2012) Business intelligence and analytics: from big data to big impact. Manag Inf Syst Q (MIS) Q 36(4):1165–1188
50. Boyd D, Crawford K (2012) Critical questions for big data provocations for a cultural, technological, and scholarly phenomenon. Inf Commun Soc 15(5):662–679
51. Kezunovic M, Xie L, Grijalva S (2013) The role of big data in improving power system operation and protection. In: IREP symposium bulk power system dynamics and control -ix optimization, security and control of the emerging power grid. IEEE computer society
52. Belaud J-P, Negny S, Dupros F et al (2014) Collaborative simulation and scientific big data analysis: illustration for sustainability in natural hazards management and chemical process engineering. Comput Ind 65:521–535
53. Herodotou H, Lim H, Luo G et al (2011) Starfish: a self-tuning system for big data analytics. In: Proceeding of the 5th biennial conference on innovative data systems research (CIDR 11), pp 261–272
54. Begoli E, Horey J (2012) Design principles for effective knowledge discovery from big data. In: Proceeding of the joint working IEEE/IFIP conference on software architecture (WICSA) and European conference on software architecture (ECSA), pp 215–218
55. Agrawal D, Das S, Abbadi AE (2011) Big data and cloud computing: current state and future opportunities. In: Proceeding of the 14th international conference on extending database technology (EDBT/ICDT). ACM, pp 530–533
56. Chen Y, Alspaugh S, Katz R (2012) Interactive analytical processing in big data systems: a cross-industry study of mapreduce workloads. J VLDB Endow 5(12):1802–1813
57. Walker DW, Dongarra JJ (1996) MPI: a standard message passing interface. Supercomputer 12:56–68
58. Huai Y, Lee R, Zhang S et al (2011) DOT: a matrix model for analyzing, optimizing and deploying software for big data analytics in distributed systems. In: Proceeding of the ACM symposium on cloud computing

59. Costa P, Donnelly A, Rowstron A, OShea G (2012) Camdoop: exploiting in-network aggregation for big data applications. In: Proceeding of the USENIX symposium on networked systems design and implementation (NSDI). ACM
60. Wu X, Zhu X, Wu G-Q, Ding W (2014) Data mining with big data. IEEE Trans Knowl Data Eng 26(1):97–107
61. Bu Y, Brokar V, Carey MJ et al (2012) Scaling datalog for machine learning on big data. Computer research repository (CoRR) Cornell University Library, pp 1–14. http://arxiv.org/pdf/1203.0160v2.pdf. Accessed 11 Aug 2015
62. Suthaharan S (2014) Big data classification: problems and challenges in network intrusion prediction with machine learning. ACM SIGMETRICS Perform Eval Rev 41(4):70–73
63. Wang W, Lu D, Zhou X et al (2013) Statistical wavelet-based anomaly detection in big data with compressive sensing. EURASIP J Wirel Commun Netw 2013(269):1–6
64. Madden S (2012) From databases to big data. IEEE Internet Comput 16(3):4–6
65. Borkar V, Carey MJ, Li C (2012) Inside "Big Data Management": ogres, onions, or parfaits? In: Proceeding of EDBT/ICDT joint conference. ACM
66. Fisher D, DeLine R, Czerwinsk M, Drucker S (2012) Interactions with big data analytics. Interactions 19(3):50–59
67. Shen Z, Wei J, Sundaresan N, Ma K-L (2012) Visual analysis of massive web session data. In: IEEE symposium on large data analysis and visualization (LDAV), pp 65–72
68. Light RP, Polley DE, Börner K (2014) Open data and open code for big science studies. Scientometrics 101(2):1535–1551
69. Camacho J (2014) Visualizing big data with compressed score plots: approach and research challenges. Chemometr Intell Lab Syst 135:110–125
70. Aronova E, Baker KS, Oreskes N (2010) Big science and big data in biology. Hist Stud Nat Sci 40(2):183–224
71. Bughin J, Chui M, Maniya J (2010) Clouds, big data, and smart assets: ten tech-enabled business trends to watch. McKinsey Q 56(1):75–86
72. Ari I, Olmezogullari E, Celebi OF (2012) Data stream analytics and mining in the cloud. In: IEEE international conference on cloud computing technology and science. IEEE Computer Society, pp 857–862
73. Takeda S, Kobayashi A, Kobayashi H et al (2012) Irregular trend finder: visualization tool for analyzing time-series big data. In: IEEE international conference on visual analytics science and technology (VAST). IEEE Computer Society, pp 305–306
74. Ma C-L, Shang X-F, Yuan Y-B (2012) A three-dimensional display for big data sets. In: International conference on machine learning and cybernetics (ICMLC). IEEE Computer Society, pp 1541–1545
75. Xu X, Yang Z, Xiu J, Liu C (2013) A big data acquisition engine based on rule engine. J Chin Univ Post Telecommun 20(1):45–49
76. Uehara M (2013) Split file model for big data in low throughput storage. In: IEEE International conference on complex, intelligent, and software intensive systems, pp 250–256
77. Khalid A, Afzal H, Aftab S (2014) Balancing scalability, performance and fault tolerance for structured data (BSPF). In: IEEE international conference on advanced communication technology (ICACT), pp 725–732
78. Xu Z, Mei L, Liu Y, Hu C (2013) Video structural description: a semantic based model for representing and organizing video surveillance big data. In: IEEE international conference on computational science and engineering, pp 802–809
79. Wang Y, Li B, Luo R, Chen Y (2014) Energy efficient neural networks for big data analytics. In: Design, automation and test in Europe conference and exhibition (DATE), pp 1–2
80. Bi C, Ono K, Ma K-L et al (2013) Proper orthogonal decomposition based parallel compression for visualizing big data on the K computer. In: IEEE symposium on large data analysis and visualization, pp 121–122
81. Bao F, Chen J (2014) Visual framework for big data in d3.js. In: Proceeding of the 2014 IEEE workshop on electronics, computer and applications, pp 47–50
82. Cuzzocrea A, Moussa R, Xu G (2013) OLAP*: effectively and efficiently supporting parallel OLAP over big data. Model Data Eng 8216:38–49
83. Czarnul P (2014) A workflow application for parallel processing of big data from an internet portal. Proc Comput Sci 29:499–508

84. Hui K, Mou J (2013) Case of small-data analysis for ion implanters in the era of big-data FDC. In: IEEE annual SEMI advanced semiconductor manufacturing conference (ASMC), pp 315–319
85. Steed CA, Ricciuto DM, Shipman G et al (2013) Big data visual analytics for exploratory earth system simulation analysis. Comput Geosci 61:71–82
86. Gao S, Li L, Li W et al (2014) Constructing Gazetteers from volunteered big geo-data based on Hadoop. Comput Environ Urban Syst. doi:10.1016/j.compenvurbsys.2014.02.004
87. Afendi FM, Ono N, Nakamura Y et al (2013) Data mining methods for OMICS and knowledge of crude medicinal plants toward big data biology. Comput Struct Biotechnol J 4(5):1–14
88. Levy V (2013) A predictive tool for nonattendance at a speciality clinic: an application of multivariate probabilistic big data analytics. In: Proceeding of the IEEE international conference and expo on emerging technologies for a smarter world (CEWIT), pp 1–4
89. Park HW, Leydesdorff L (2013) Decomposing social and semantic networks in emerging "Big Data" research. J Inf 7(3):756–765
90. Ackermann K, Angus SD (2014) A resource efficient big data analysis method for the social sciences: the case of global IP activity. Proc Comput Sci 29(2014):2360–2369
91. Provost F, Fawcett T (2013) Data science and its relationship to big data and data-driven decision making. Big Data 1(1):51–59
92. Rybicki J, von St Vieth B, Mallmann D (2013) A concept of generic workspace for big data processing in humanities. In: IEEE international conference on big data, pp 63–70
93. O'Driscoll A, Daugelaite J, Sleator RD (2013) "Big Data", Hadoop and cloud computing in genomics. J Biomed Inform 46(6):774–781
94. NIST: http://www.nist.gov
95. Jacobs A (2009) The pathologies of big data. Commun ACM 52(8):36–44
96. Chang F, Dean J, Ghemawat S et al (2008) BigTable: a distributed storage system for structured data. ACM Trans Comput Syst 26(2):1–26
97. DeCandia G, Hastorum D, Jampani M et al (2007) Dynamo: Amazons highly available key-value store. In: Proceeding of the 21st ACM SIGOPS symposium on operating systems principles, pp 205–220
98. Dou W, Zhang X, Liu J et al (2013) HireSome-II: towards privacy-aware cross-cloud service composition for big data applications. IEEE Trans Parallel Distrib Syst TPDS 26(2):455–466
99. Zhang X, Liu C, Nepal S et al (2014) A hybrid approach for scalable sub-tree anonymization over big data using mapreduce on cloud. J Comput Syst Sci 80(5):1008–1020
100. Jung G, Gnanasambandam N, Mukherjee T (2012) Synchronous parallel processing of big-data analytics services to optimize performance in federated clouds. In: Proceeding of the 2012 IEEE 5th international conference on cloud computing, pp 811–818
101. Yang C, Zhang X, Zhong C et al (2014) A Spatiotemporal compression based approach for efficient big data processing on cloud. J Comput Syst Sci 80(8):1563–1583
102. IDC: http://www.idc.com
103. Hazen BT, Boone CA, Ezell JD et al (2014) Data Quality for data science, predictive analysis, and big data in supply chain management: an introduction to the problem and suggestions for research and applications. Int J Prod Econ 154:72–80
104. Tannahill BK, Jamshidi M (2014) System of systems and big data analytics -bridging the gap. Comput Electr Eng 40:2–15
105. Lohr S (2012) The age of big data. The New York Times, New York
106. Cohen J, Dolan B, Dunlap M et al (2009) MAD skills: new analysis practices for big data. In: Proceeding of the VLDB 09. VLDB endowment
107. Kumar A, Niu F, Ré C (2013) Hazy: make it easier to build and maintain big-data analytics. Commun ACM 56(3):40–49
108. Machanavajjgala A, Reiter JP (2012) Big privacy: protecting confidentiality in big data. Magazine XRDS: crossroads. ACM Mag Stud Big Data 19(1):20–23
109. Feldman D, Schmidt M, Sohler C (2013) Turning big data into tiny data: constant-size coresets for k-means, PCA and projective clustering. In: Proceeding of the annual ACM-SIAM symposium on discrete algorithms (SODA), pp 1434–1453
110. Laptev N, Zeng K, Zaniolo C (2013) Very fast estimation for result and accuracy of big data analytics: the EARL system. In: Proceeding of the IEEE international conference on data engineering (ICDE), pp 1296–1299

111. Wu Z, Chin OB (2014) From big data to data science: a multi-disciplinary perspective. Big Data Res 1:1
112. Chandramouli B, Goldstein J, Duan S (2012) Temporal analytics on big data for web advertising. In: Proceeding of the IEEE 28th international conference on data engineering (ICDE), pp 90–101
113. LaValle S, Lesser E, Shockley R et al (2011) Big data, analytics, and the path from insights to value. Hum Cap Rev Focus Hum Cap Anal 1(1)
114. Russom P (2011) Big data analytics. TDWI Best Practices Report, Fourth Quarter, pp 1–37. ftp://ftp.software.ibm.com/software/tw/Defining_Big_Data_through_3V_v.pdf. Accessed 11 Aug 2015
115. Borgman CL (2010) Research data: who will share what, with whom, when, and why? Working Paper No. 161, German Data Forum (RatSWD). Retrieved from www.germandataforum.de
116. Yang C, Goodchild M, Huang Q et al (2011) Spatial cloud computing: how can the geospatial sciences use and help shape cloud computing? Int J Digit Earth 4(4):305–329
117. Pijanowski BC, Tayyebi A, Doucette J et al (2014) A big data urban growth simulation at a national scale: configuring the GIS and neural network based land transformation model to run in a high performance computing (HPC) environment. Environ Model Softw 51:250–268
118. Callebaut W (2012) Scientific perspectivism: a philosopher of sciences response to the challenge of big data biology. Stud Hist Philos Biol Biomed Sci 43(1):69–80
119. Vanacek J (2012) How cloud and big data are impacting the human genome: touching 7 billion lives. Forbes. http://www.forbes.com/sites/sap/2012/04/16/how-cloud-and-big-data-are-impacting-the-human-genome-touching-7-billion-lives/. Accessed 11 Aug 2015
120. Costa FF (2012) Big data in genomics: challenges and solutions. GIT Lab J 11–12:1–4
121. Varpoorte R, Kim H, Choi Y (2006) Plants as source of medicines:new perspectives. In: Bogers RJ, Craker LE, Lange D (eds) Medicinal and aromatic plants. Springer, Netherlands, pp 261–273
122. Boyd D, Crawford K (2011) Six provocations for big data. In: A decade in internet time: symposium on the dynamics of the internet and society. doi:10.2139/ssrn.1926431. Accessed 11 Aug 2015
123. Ansolabehere S, Hersh E (2012) Validation: what big data reveal about survey misreporting and the real electorate. Polit Anal 20(4):437–459
124. Tene O, Polonetsky J (2012) Privacy in the age of big data: a time for big decisions. Standf Law Rev 63:63–69
125. Spalation Neutron Source (SNS). http://neutrons.ornl.gov/sns
126. White AA (2013) Big data are shaping the future of materials science. MRS Bull 38:594–595
127. ADARA. http://www.csm.ornl.gov/newsite/adara.html
128. Von Lilienfeld OA (2013) First principles view on chemical compound space: gaining rigorous atomistic control of molecular properties. Int J Quantum Chem 113(12):1676–1689
129. Groves P, Kayyali B, Knott D et al (2013) The big-data revolution in US health care: accelerating value and innovation. McKinsey & Company, New York
130. Kayyali B, Knott D, Van Kauiken S (2013) The big-data revolution in US health care: accelerating value and innovation. McKinsey & Company, New York
131. Lusher SJ, McGuire R, van Schaik RC et al (2014) Data-driven medicinal chemistry in the Era of big data. Drug Discov Today 19(7):859–868
132. Costa FF (2013) Social networks, web-based tools and diseases: implication for biomedical research. Drug Discov Today Elsevier 18(5–6):272–281
133. New Vantage Partners (2012) Big data executive survey 2012. Consolidated summary report. http://newvantage.com/wp-content/uploads/2012/12/NVP-Big-Data-Survey-Themes-Trends.pdf. Accessed 11 Aug 2015
134. Demirkan H, Delen D (2013) Leveraging the capabilities of service-oriented decision support systems: putting analytics and big data in cloud. Decis Support Syst 558(1):412–421
135. Roman S, Katerina S (2012) The usability of agent-based simulation in decision support system of e-commerce architecture. Int J Inf Eng Electron Bus 4(1):10–17
136. Harrison C, Eckman B, Hamilton R et al (2010) Foundations for smarter cities. IBM J Res Dev 54(4):1–16
137. Khan Z, Anjum A, Liaquat Kiani S (2013) Cloud based big data analytics for smart future cities. In: Proceeding of the IEE/ACM 6th international conference on utility and cloud computing, pp 381–386
138. Vilajosana I, Llosa J, Martinez B et al (2013) Bootstrapping smart cities through a self-sustainable model based on big data flows. IEEE Commun Mag 51(6):128–134

139. Dey S, Chakravorty A, Naskar S, Misra P (2012) Smart city surveillance: leveraging benefits of cloud data stores. In: Proceeding of the first IEEE international workshop on global trends in smart cities, pp 868–876
140. Jara AJ, Genoud D, Bocchi Y (2014) Big data in smart cities: from poisson to human dynamics. In: Proceeding of the IEEE 28th international conference on advanced information networking and applications workshops (WAINA). IEEE computer society, pp 785–790
141. Girtelschmid S, Steinbauer M, Kumar V et al (2013) Big data in large scale intelligent smart city installations. In: Proceeding of the international conference on information integration and web-based applications and services (IIWAS). ACM
142. Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (1996) Advances in knowledge discovery and data mining. American Association for Artificial Intelligence, New York
143. Rajaraman A, Ullman J (2011) Mining of massive data sets. Cambridge Univercity Press, Cambridge
144. Berkovich S, Liao D (2012) On clusterization of big data streams. In. Proceeding of the 3rd international conference on computing for geospatial research and applications (COM.Geo). ACM
145. Moens S, Aksehirli E, Goethals B (2013) Frequent itemset mining for big data. In: Proceeding of the IEEE international conference on big data, pp 111–118
146. Ledolter J (2013) Data mining and business analytics with R. John Wiley & Sons, New York
147. Slavakis K, Giannakis GB, Mateos G (2014) Modeling and optimization for big data analytics. IEEE Signal Process Mag 31(5):18–31
148. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323
149. Grolinger K, Hayes M, Higashino WA et al (2014) Challenges for MapReduce in big data. In: Proceeding of the 2014 IEEE world congress on services (SERVICES), pp 182–189
150. Hashem IAT, Yaqoob I, Badrul Anuar N et al (2015) The rise of "Big Data" on cloud computing: review and open research issues. Inf Syst 47:98–115
151. Zhifeng X, Yang X (2013) Security and privacy in cloud computing. IEEE Commun Surv Tutor 15(2):843–859
152. Esposito C, Ficco M, Palmieri F et al (2014) A knowledge-based platform for big data analytics based on publish/subscribe services and stream processing. Knowl Based Syst 79:3–17
153. López V, del Río S, Benítez JM et al (2014) Cost-sensitive Linguistic fuzzy rule based classification systems under the mapreduce framework for imbalanced big data. Fuzzy Sets Syst 258:5–38
154. Ghemawat S, Gobioff H, Leung S-T (2003) The Google file system. In: Proceeding of the 19th ACM symposium on operating systems principles SOSP 03, pp 29–43
155. Lin J, Ryaboy D (2012) Scaling big data mining infrastructure: the twitter experience. SIGKDD Explor 14(2):6–19
156. Isard M, Budiu M, Yu Y et al (2007) Dryad: distributed data-parallel programs from sequential building blocks In: Proceeding of the 2nd ACM SIGOPS/EuroSys European conference on computer systems, pp 59–72
157. Yu Y, Isard M, Fetterly D et al (2008) DryadLINQ: a system for general-purpose distributed data-parallel computing using a high-level language. In: Proceeding of the 8th USENIX conference on operating systems design and implementation, pp 1–14
158. Owen S, Anil R, Dunning T et al (2011) Mahout in action. Manning Publications Co. Greenwich, CT, USA
159. Apache Storm. https://storm.apache.org/
160. Neumeyer L, Robbins B, Nair A et al (2010) S4: distributed stream computing platform. In: Proceeding of the 2010 international conference on data mining workshops (ICDMW). IEEE
161. Stoica I (2014) Conquering big data with spark and BDAS. In: Proceeding of the ACM international conference on measurement and modeling of computer systems
162. Bifet A, Holmes G, Kirkby R et al (2010) MOA: massive online analysis. J Mach Learn Res (JMLR) 11:1601–1604
163. Apache Drill. http://drill.apache.org/
164. Franceschini M (2013) How to maximize the value of big data with the open source SpagoBI suite through a comprehensive approach. In: Proceeding of the VLDB endowment, vol 6, pp 1170–1171
165. Bostock M, Ogievetsky V, Heer J (2011) D3 data-driven documents. IEEE Trans Vis Comput Graph 17(12):2301–2309
166. SMLC: Smart Manufacturing Leadership Coalition. https://smartmanufacturingcoalition.org/
167. Ahmed KN (2013) Putting big data to work. Mech Eng 135:32–37

168. Guillemin P, Friess P (2009) Internet of things: strategic research roadmap. The cluster of European research projects. Tech. Rep. http://www.internet-of-things-research.eu/pdf/IoT_Cluster_Strategic_Research_Agenda_2009.pdf. Accessed 11 Aug 2015
169. Perera C, Zaslavsky A, Christen P et al (2014) Context aware computing for the internet of things: a survey. IEEE Commun Surv Tutor 16(1):414–454
170. Stimmel CL, Gohn B (2012) Smart grid data analytics: smart meter, grid operations, asset management, and renewable energy integration data analytics: global market analysis and forecasts. Research Report (Executive Summary), 3Q, pp 1–16
171. Qin X, Zhou X (2013) A survey on benchmarks for big data and some more considerations. In: Yin H, Tang K, Gao Y et al (eds) Intelligent data engineering and automated learning-IDEAL 2013. LNCS, vol 8206. Springer, Berlin, Heidelberg, pp 619–627
172. Baru C, Bhandarkar M, Nambiar E et al (2013) Benchmarking big data systems and the big data top100 list. Big Data 1(1):60–64
173. Xiong W, Yu Z, Bei Z et al (2013) A characterization of big data benchmarks. In: 2013 IEEE international conference on big data, pp 118–125
174. Ming Z, Luo C, Gao W et al (2014) BDGS: a scalable big data generator suite in big data benchmarking. Adv Big Data Benchmark LNCS 8585:138–154
175. Wang L, Zhan J, Luo C et al (2014) BigDataBench: A Big Data Benchmark Suite from Internet Services. In: Proceeding of the IEEE 20th international symposium on high performance computer architecture (HPCA), pp 488–499
176. Shekhar S, Evans MR, Gunturi V (2014) Benchmarking spatial big data. Specif Big Data Bechmark LNCS 8163:81–93
177. Dean J (2014) Big data, data mining and machine learning: value creation for business leaders and practitioners. Wiley, New York
178. Tang N (2014) Big data cleaning. Web Technol Appl LNCS 8709:13–24