Computer Standards & Interfaces 54 (2017) 105–115

Contents lists available at ScienceDirect

Computer Standards & Interfaces

journal homepage: www.elsevier.com/locate/csi

Research on Big Data – A systematic mapping study

Jacky Akoka[a,b], Isabelle Comyn-Wattiau[c,*], Nabil Laoufi[a]

[a] CEDRIC-CNAM, Paris, France
[b] TEM-Institut Mines Telecom, Evry, France
[c] ESSEC Business School, Cergy-Pontoise, France

ARTICLE INFO

Keywords:
Big Data
Systematic mapping study
Framework
Artefact
Usage
Analytics

ABSTRACT

Big Data has emerged as a significant area of study for both practitioners and researchers. Big Data is a term for massive data sets with large structure. In 2012, Big Data passed the top of the Gartner Hype Cycle, attesting the maturity level of this technology and its applications. The aim of this paper is to examine how do researchers grasp the big data concept? We will answer the following questions: How many research papers are produced? What is the annual trend of publications? What are the hot topics in big data research? What are the most investigated big data topics? Why the research is performed? What are the most frequently obtained research artefacts? What does big data research produces? Who are the active authors? Which journals include papers on Big Data? What are the active disciplines? For this purpose, we provide a framework identifying existing and emerging research areas of Big Data. This framework is based on eight dimensions, including the SMACIT (Social Mobile Analytics Cloud Internet of Things) perspective. Current and past research in Big Data are analyzed using a systematic mapping study of publications based on more than a decade of related academic publications. The results have shown that significant contributions have been made by the research community, attested by a continuous increase in the number of scientific publications that address Big Data. We found that researchers are increasingly involved in research combining Big Data and Analytics, Cloud, Internet of things, mobility or social media. As for quality objectives, besides an interest in performance, other topics as scalability is emerging. Moreover, security and quality aspects become important. Researchers on Big Data provide more algorithms, frameworks, and architectures than other artifacts. Finally, application domains such as earth, energy, medicine, ecology, marketing, and health attract more attention from researchers on big data. A complementary content analysis on a subset of papers sheds some light on the evolving field of big data research.

1. Introduction

Nowadays, organizations and individuals generate large amounts of data at a very high rate. With an impressing amount of data arriving at an exabyte scale, new insights can be obtained from their contents. The latter will help organizations to gain richer insights and improve their competitive position. Moreover, it is generally accepted that relevant information obtained using Big Data technologies will enhance enterprises efficiency and competitiveness.

International Data Corporation (IDC) found that the created and copied data volume in the world was 1.8 zettabites (ZB). It is estimated that this figure will double every other two years in the near future [1]. [2] asserts that Big Data can improve the potential value of the US medical industry estimated at USD 300 billion. It considers that retailers that fully utilize Big Data may increase their profit by more than 60%. Finally, according to McKinsey, Big Data may also be utilized to improve the efficiency of government operations. Let us remind that 5 exabytes (EB) of data were created by human until 2003. Today this amount of information is created in two days. In 2012, digital world was expanded to 2.72 ZB. It is predicted to double every two years, reaching about 8 ZB by 2015 [3]. IBM indicates that every day 2.5 EB of data are created. CISCO predicts that, by 2020, 50 billion devices will be connected to networks and to the Internet. The investment in spending on Information Technology (IT) infrastructure of the digital universe and telecommunications will grow by 40% between 2012 and 2020. Big Data will account for 40%. Moreover IDC expects that 23% of the information in the digital universe (or 643 EB) would be useful for Big Data. It includes data originated from surveillance footage, embedded and medical devices, entertainment, social media, as well as consumer images.

Companies are learning to take advantage of Big Data. They use real-time information from sensors, radio frequency identification to

* Corresponding author.
E-mail addresses: akoka@cnam.fr (J. Akoka), wattiau@essec.edu (I. Comyn-Wattiau), laouf_na@auditeur.cnam.fr (N. Laoufi).

http://dx.doi.org/10.1016/j.csi.2017.01.004
Received 1 April 2016; Received in revised form 2 January 2017; Accepted 15 January 2017
Available online 17 January 2017
0920-5489/

understand their business environments and to create new products and services. Organizations capitalize on Big Data in three ways: (i) they pay attention to data flows, (ii) they rely on data scientists, (iii) they are moving analytics away from the IT function [4]. Huge investments are being made by companies with great expectations for the gains to be made. Big Data is considered to be the new engine to sustain the high growth of the information industry. Enterprises' competitiveness is increasingly determined by their abilities to leverage the technologies associated with Big Data. However, several questions remain to be answered. What does the future hold? Will the changes be transformative? What domains are likely to benefit the most? Even though it is difficult to bring answers to these questions, companies had great expectations as attested by Gartner. In 2012, Big Data passed the top of the Gartner Hype Cycle. In 2014, Big Data moved to Trough Disillusionment phase, attesting a maturity level of companies which invested in this technology. Successful applications of Big Data in industry are reported in many publications. Thus, it seems that industry is ahead of academia [5].

In [6], we analyzed Big Data research using five different perspectives, namely: the timeline, the context, the objectives, the artefacts created, and the applications (usages). In this paper we updated the data set of papers and performed a complementary ascending analysis allowing us to confirm and enrich these perspectives as well as adding a new one characterizing: who are the researchers publishing on big data? In which journals they disseminate their research? Which disciplines are involved? Moreover, we enlarged the source by including all scientific domains as referenced by ScienceDirect. In Big Data Research Journal, we conducted a specific analysis in order to detect the false negative papers. Finally, we screened all relevant papers in order to check the potential presence of false positive elements. We explore the term Big Data using the peer reviewed literature. Our research question may be defined as follows: how do researchers grasp the big data concept? The rest of the paper is worded as follows. Section 2 presents a synthetic literature review on Big Data. In Section 3, we define our framework as a multidimensional model. Section 4 presents our approach to evaluate the evolution of Big Data in terms of contributions by the research community. To this end, we position Big Data in the framework. In addition, we perform a content analysis on a subset of papers, providing some insight on big data research. Section 5 concludes the work.

2. Big Data – a literature review

Big Data has been a buzzword in the last decade. The term, coined by Roger Magoulas[1], refers to large data sets almost impossible to manage and process using traditional data management tools. It refers to various forms of large information sets requiring complex computational platforms in order to be analyzed.

There have been some discussions on the definition of Big Data [7,8]. [9] defined it using the 3Vs model (Volume, Velocity, and Variety). [1] defined Big Data as "a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high velocity capture, discovery, and/or analysis". Based on the 3Vs model, [10] presents an extensive review of Big Data research issues, mentioning security as a main one. [11] synthesizes the definitions proposed by companies (Google, Oracle, Gartner, Microsoft, Intel, etc.) concluding by "Big data is a term describing the storage and analysis of large and or complex data sets using a series of techniques including, but not limited to: NoSQL, MapReduce and machine learning" since all definitions mention size, complexity, and technologies as constituting the specificity of the Big Data concept. [12] analyzes previous definitions according to four axes: technology, method, information, and impact, and finally proposes the following definition: "Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value." [13] define Big Data aiming at the characteristics of the generated data, containing both the amount and structure of the data. [14] enrich the data characteristics by additional attributes, such as the scope, target, and structure of the data. [15] focus on the amount of data and include the aspect of method. [16] include the aspect of methods and IT infrastructure topics. [17] incorporate data characteristics and infrastructure. [2] incorporate the method aspect.

[16] focus on Business analytics resulting from Big Data. However, published in 2012, it only reports early results in big data and BI & A (Business Intelligence and Analytics). [18] is a comprehensive state-of-the-art on Big Data Research. The authors introduce the general background of Big Data and review related technologies such as Cloud computing, Internet of Things (IoT), and Hadoop. They also emphasize the four phases of the value chain of Big Data, i.e., data generation, data acquisition, data storage, and data analysis. They examine applications of Big Data, namely enterprise management, IoT, social networks, medical applications, collective intelligence, and smart grids. They review the main techniques that Big Data Research implements.

Emerging Big Data opportunities can be classified into several topic areas. Jeanne Ross (MIT) proposed the five key areas of Social media, Mobile systems, Analytics, Cloud, and IoT (SMACIT) as significant drivers for enterprise digital transformation. Such a classification aims to emphasize the relations of these technologies with Big Data characteristics. IoT, Mobile and Social network are major sources of Big Data.

[19] present a literature survey on Big Data analytics. The specificity of analytics involving Big Data is due to the hard deadlines, and to data quality. Methods need to be scaled. The challenges range from building storage systems to collecting data from distributed sources in order to run a diverse set of computations on data. This paper highlights three aspects of Big Data analytics: hardware specificities, software platforms, and a few application domains. There is a multi-step pipeline required to extract value from data: heterogeneity, incompleteness, scale, timeliness, privacy, and process complexity [20]. [21] present a description of Big Data focusing on the analytic methods used specifically for Big Data. They emphasize analytics related to unstructured data, which constitute 95% of Big Data, such as text, audio, video, and social media data. Some potential applications from different fields are: evolution of commercial applications, and evolution of scientific applications. [22] present a survey of the big data area, listing the challenges and solutions in industries and academics from the perspectives of engineers, computer scientists, and statisticians.

There are many solutions for Big Data related to cloud computing. Depending on the level of volume, variety, velocity, it is important to choose appropriate Big Data tools. Thanks to the cloud, we move to Big Data as a Service or Analytics as a Service [23]. Thus, customer and provider's staff are much more involved in the loop. [24] proposed a classification for big data, a conceptual view of big data, and a cloud services model. They also investigate some innovative research issues such as privacy, legal and regulatory issues, and governance.

Although the current IoT data is not the dominant part of Big Data, by 2030, the quantity of sensors will reach one trillion. Then the IoT data will be the most important part of Big Data [25]. Big Data in IoT has three features that conform to the Big Data paradigm: (i) abundant terminals generating masses of data; (ii) data generated by IoT is usually semi-structured or unstructured; (iii) data of IoT is useful only when analyzed.

Finally, [16] mentioned the following applications: e-commerce, e-government, science and technology, smart health, security and public safety. One example is sales planning allowing organizations to optimize their commodity prices. They can also improve their opera-

[1] http://strata.oreilly.com/2010/01/roger-magoulas-on-big-data.html.

tion efficiency and optimize the labor force, and use Big Data to conduct inventory and logistic optimization. Big Data enables enterprises to predict the consumer behavior. In e-commerce numerous transactions can be conducted and recorded every day.

Despite the success of Big Data applications, some obstacles and challenges in the development of Big Data applications remain. Among them let us mention: data representation, redundancy reduction and data compression, data life cycle management, analytical mechanism, data confidentiality, energy management, expendability and scalability, and cooperation [20,26].

[27] analyze the merging of a Big Data architecture in an already existing information system It also tackles semantics aspects (reasoning, coreference resolution, entity linking, information extraction, consolidation, paraphrase resolution, ontology alignment) in the context of Big Data. [28] survey programming models developed for cluster cloud and grid supporting big data analytics. [29] discuss several underlying methodologies to handle the data deluge, such as granular computing, cloud computing, bio-inspired computing, and quantum computing. [30] analyze the main issues and challenges according to three steps of the process which are data gathering, data processing, and data mining. [31] compare six different big data models (BigTable, Cassandra, HBase, MongoDB, CouchDB, CrowdDB) as six alternative ways to implement big data sets.

[32] focus on security aspects. It structures big data security and privacy concerns in five categories, that are respectively Hadoop security, cloud security, anonymization, monitoring and auditing, and key management. [33] compare nine big data systems on security criteria (authentication, encryption, auditing, communication protocol, etc.). [34] exploresspecific data quality problems appearing or worsening in big data areas, e.g. inauthentic data collection, information incompleteness, noise, representativeness, consistency, reliability.

[35] is dedicated to in-memory big data management. Performance problems belong to seven families: index, data layout, parallelism, transaction management, query optimization, fault tolerance, and data overflow. Eighteen in-memory systems are compared. [36] focus on visualization tools for big data. Ten commercial systems (including Tableau, Qlik View, Spotfire, JMP) are compared. [37] compare hardware platforms for big data according to six characteristics, i.e. scalability, data I/O rate, fault tolerance, real-time processing, data size supported, and iterative task support. [38] review the real-time big data systems. [39] survey the clustering algorithms and evaluates their ability to deal with volume, velocity, and variety aspects. [40] illustrate how these clustering algorithms are currently evolving to meet these different requirements.

[41] listed and sorted the challenges pertaining to big data as follows: data growth, data infrastructure, data governance/policy, data integration, data velocity, data variety, data compliance/regulation, data visualization.

[42] elicit different dilemmas regarding big data and provide an interesting in-depth analysis of them. The authors categorize dilemmas as follows: 1) epistemological dilemmas such as the relationship between data and knowledge, the distinction between causal relation and statistical correlation; 2) methodological dilemmas rendering obsolete the dichotomy between quantitative and qualitative research, 3) aesthetic dilemmas, in data visualization where accuracy and aesthetics may be contradictory; 4) technological dilemmas; 5) legal and ethical dilemmas, e.g. regarding privacy; 6) political economy dilemmas.

The literature also contains many reviews exploring how big data impacts some domains. These reviews are generally published in the journals specific to the corresponding domains. They help researchers to identify how big data solutions may help them in developing new contributions in their fields. As an example [43] is dedicated to a survey on big data analytics in healthcare and government.

[44] see three types of academic units discussing Big Data research. The first one is the hard core science units which include research in astronomy, climate science, and genomics. The second one is the information sciences units contributing to the Big Data paradigm but moderately. The last group is composed of "all units that have realized the potential of working with diverse and large datasets", such as units related to health care, public health, education, public policy, government studies, marketing and retail, and finance. [45] found a continuous increase in the number of publications that address Big Data in scientific databases, such as Scopus. [46] review the current literature on Big Data and reveal a focus on the technical perspective.

The authors of [47] conducted a systematic literature review on big data and identified different relevant dimensions: 1) type of value creation, including creating transparency, enabling experimentation, segmenting populations, supporting human decisions, innovating new business models, products or services; 2) enabled business value, through data policies, technologies, organizational change, access to data, industry structure; 3) industry (retail, healthcare, ecology, education, government, manufacturing, services, technology); 4) research approach; 5) journal. They also report a case study providing qualitative information on big data specificities.

As a conclusion, the literature is prolific on different aspects: 1) big data definitions, 2) big data impacts, 3) classical solutions (algorithms, models, systems) revisited in the context of big data, 4) big data issues and challenges. Thus the researchers address the big data topic according to numerous dimensions. We collected these dimensions and organized them to measure how each one describes big data research:

- The time dimension: many papers show the impressive curve of big data research publications;
- the application dimension: several papers list the application domains of big data research;
- the objective dimension: some papers review how research addresses security aspects, others deal with quality and several analyze the performance and optimization aspects.

In our study, we added several other dimensions aiming at gaining a more comprehensive view of the field. In order to obtain a deep understanding of big data research, we decided to conduct a systematic mapping study. In the following, we explain why this type of study is relevant in such a case and then list the research questions that this mapping study aims to answer.

3. Research method: a systematic mapping process

The objective of this paper is to structure the existing literature of the field of big data. To this end, we performed a systematic mapping study to categorize and summarize the existing information concerning the research questions. Systematic mapping is a relevant method for structuring a research field such as big data. It can be considered as a methodology that offers a visual summary map of the field. It is used to describe big data research undertaken so far. It gives an overview of big data research activity. It aims at collecting papers on big data, performing a classification of these papers, and obtaining an overview on the current state of research. Systematic mapping [48] provides mechanisms to identify research evidence. Unlike systematic review [49] which provides an evaluation of the state of the art related to a topic of interest, systematic mapping is a more open form of systematic review. The latter reviews published papers and analyzes their methodology and results. Although it has several benefits, it suffers from the fact that it requires a considerable effort. On the other hand, systematic mapping requires less effort while summarizing the existing literature. Given the fact that the big data area can be characterized by its novelty, a systematic mapping study is recommended for this research area where there is a lack of relevant primary studies [49]. Systematic mapping allows us to obtain an overview of big data research area and describes how far it is covered in research. It helps in building a

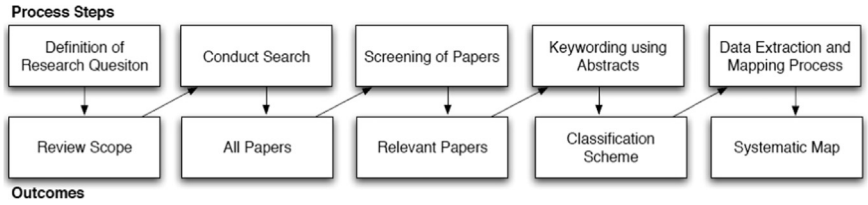

Fig. 1. The systematic mapping process [48].

classification scheme. One result is the frequencies of publications for the categories obtained. To a certain extent it describes the coverage of the big data field. Moreover, it enables to answer the research questions. Finally it provides a visual summary by mapping the results.

As a summary let us stress that systematic mapping and systematic review are different in terms of goals, breadth, validity issues and implications. However systematic mapping, unlike systematic review, is appropriate here since it focuses on classification of research. It allows us to identify research gaps in big data. It enables thematic analysis of big data research. It reflects based on search strings and inclusion criteria. It is visually more appealing.

We have adapted and applied the systematic mapping approach described in [48] to Big Data research. The process used includes the definition of research questions, conducting the search for relevant papers, screening of papers, keywording using abstracts, keywords, and titles, and finally data extraction and mapping (Fig. 1). As it can be seen, each step has an outcome. The final result of the process is the systematic map.

3.1. Definition of research questions

The first step consists in defining the research question. It covers the review scope. It allows us to define precisely what we want to accomplish and where we want to search for the information. In other terms, it enables to identify forums for our research area. Our main objective in this phase is to provide an outline of the Big Data research area. This amounts to identify mainly the actors, the quantity, the frequency, the motivation, the forums, the main approaches, the results, and the type of research performed. Our research questions (RQs) as shown in Table 1.

In [6], we addressed the five following dimensions:

- the timeline of publications (When),
- the objectives (Why),
- the artifacts produced (How),
- the related hot topics (What),
- the application domains (Where).

We then performed a brainstorming process to have a more complete viewpoint which led us to study the authors, the journals, and the active disciplines. Thus, in this paper, we enriched the set of research questions addressed. Moreover, for each dimension, we elicited its different aggregation levels.

Table 1
Research questions.

RQ 1 How many research papers are produced?
RQ 2 What is the annual trend of publications?
RQ 3 What are the hot topics in big data research? What are the most investigated big data topics?
RQ 4 Why the research is performed?
RQ 5 What are the most frequently obtained research artefacts?
RQ 6 What does big data research produces?
RQ 7 Who are the active authors?
RQ 8 Which journals include papers on Big Data?
RQ 9 What are the active disciplines?

3.2. Conduct search for all papers

The second step is devoted to conducting a search for all papers. It consists in using search strings in ScienceDirect databases. It is obvious that the search-strings are driven from our research questions. We used search strings in order to identify all papers related to Big data research. The scientific database that was intensively used is ScienceDirect. We browsed through journal publications and looked specifically for publications defined by ScienceDirect as “original research”. The search strings used were successively applied to the abstracts or keywords or titles. As a second search, we looked for articles in the abstracts only, then in the keywords only, and finally in the titles only. Our search was applied to all scientific domains as they are defined by ScienceDirect.

More precisely, the main steps used are: (i) We checked keyword, title, and abstract fields within the database. (ii) We derived the terms from the research questions to create the search strings. (iii) To this end, we first derived the main search terms. Then we checked the keywords for relevant papers already known and looked for alternative forms of the terms such as synonyms and relevant keywords. (iv) Finally, we used Boolean operators OR and AND to incorporate them into the search strings. We show below an example of a final search string.

```
TITLE-ABSTR-KEY("Big data") and TITLE-ABSTR-KEY
("social" or "mobil" OR "analytics" OR "cloud" OR "inter-
net of things" OR "System design" OR "Language" OR
"Prototype" OR "Metric" OR "Algorithm" OR "Guideline"
OR "Methodology" OR "Architecture" OR "Framework" OR
"Taxonomy" OR "Ontology" OR "Healthcare" OR "Marketing"
OR "Tourism" OR "Finance" OR "Government" OR "Education"
OR "health" or "hospitality" OR "Traceability" OR
"Integrity" OR "Confidentiality" OR "availability" or "
privacy" or " reliability " or " scalability " or "
performance " or" usability " or" quality" or "secur-
ity").
```

As a result of this step, we obtained (1) different lists of papers describing Big Data research as a whole, (2) different lists of papers linking Big Data research to one or several dimensions, (3) a set of abstracts to be further analyzed.

The search is limited to the papers using the “big data” string either in the title and/or in the keywords and/or in the abstract. Some may consider that this produces a bias in the results. However, given the fact that the “big data” buzzword reaches a high penetration rate in firms as well as in the whole society, our initial research question was to identify the current landscape of research riding the “big data” wave. We are aware of the fact that it may ignore a facet of research addressing big data issues without using the related buzzword. This aspect is discussed in Section 4.4.

3.3. Screening for relevant papers

The notion of relevant papers is defined according to a set of inclusion and exclusion criteria. Exclusion criteria were used to exclude papers that are false positive and therefore not relevant to answer the research questions. The screening process selects relevant papers verifying the inclusion criteria such as the ones related to the

dimensions of our framework described in Section 3.4.

More precisely, we performed a screening process of the papers, considering only relevant papers, defined in ScienceDirect as < < original research > > . At this step, it is needed to define inclusion and exclusion criteria, derived from the research questions. Inclusion and exclusion criteria allow us to select the appropriate papers from literature. The inclusion criteria adopted are: (1) Only research papers published in the journals referenced in ScienceDirect; (2) Only studies described as "original research" in ScienceDirect; (3) Only papers related to themes of big data in their title or abstract or keywords; (4) Only studies restricted to publication date ranging from 2006 to February 2016. The exclusion criteria were: (1) Papers mentioning "big data" in their abstract but that cannot be considered as describing research on big data. (2) Papers containing keywords related to our dimensions (e.g. algorithm which belongs to the artifact dimension) but discovered as false positives (e.g. the paper does not describe a new algorithm but only references an existing algorithm) according to these dimensions. Let's note that all papers describing big data research are considered as relevant. However, the main objective of this step is to refine this relevance according to each dimension analyzed.

As a result of this step, we obtained the same lists of papers and abstracts as in Section 3.2 but refined thanks to exclusion and inclusion criteria.

3.4. Keywording using abstracts, keywords, and titles

The fourth step aims at obtaining a classification scheme. The literature review on Big Data revealed many dimensions for analyzing contributions. In order to enrich this state-of-the-art, we compiled the most significant dimensions found in the literature and enriched them, resulting in a new framework described as a multidimensional model (Fig. 2). Thus the classification scheme is based on our multidimensional model. Thanks to the keywording, we completed the set of values for each dimension. Moreover we added three new dimensions related to the sources of research (authors, journals, communities/academic disciplines).

First, we considered the context dimension which is linked to the SMACIT perspective. The latter is not often referenced in academic publications. However it deserved much attention from companies as a way to structure their digital transformation. Thus, we propose to analyze how Big Data and each of the SMACIT components, i.e. Social media, Mobility, Analytics, Cloud, and Internet of Things are correlated in academic publications. Each of these components is itself a hot topic in computer science research. Studying the overlapping between them and Big Data is what we call the context.

The state-of-the-art on Big Data (Section 2) illustrated the new challenges linked to volume, velocity, and variety. Hence research in Big Data aims to produce concepts, methods, and tools to build high quality IT solutions. Thus, we propose the Objective dimension to analyze which are the main targets of researchers in Big Data. Using mainly ISO software quality criteria, we propose to study the correlation of Big Data research with the following quality objectives: integrity, confidentiality, privacy, traceability, reliability, scalability, performance, usability, quality, security. We particularly detailed the security criterion (e.g. integrity, confidentiality, privacy, traceability) since it is a main issue mentioned in Big Data research.

Big Data research aims at providing professionals with models, methods, and tools to deal with Big Data applications. Thus we propose a third dimension called Artifact, allowing us to analyze which are the main contributions of Big Data research. In order to obtain a detailed viewpoint of the artifacts proposed in academic publications, we used the typology of artifacts proposed in [50]. Following [51] proposing four categories of artifacts, i.e. constructs, models, methods, and instantiations, [50] refines the categories, distinguishing between language (construct), meta-model, system design, ontology, taxonomy, framework, architecture, metric (models), methodology, guideline, algorithm (methods), and prototype (instantiation). Even if research on Big Data is very dynamic, not all artifacts are actually produced. This analysis will provide researchers with future research avenues.

The maturity of research is also linked to the validation of contributions through real life applications. Thus we also listed the main application domains and analyzed how academics address these domains. We built the list of domains by screening literature and eliciting the main fields where Big Data could bring new solutions for companies. The resulting list constitutes the Usage dimension. In [6], we elicited the following domains: healthcare, education, government, finance, marketing, and tourism since we limited our analysis to papers published in Computer Science and Decision Science domains. By enlarging the analysis to all other domains, we obtained the following additional domains: earth, energy, medicine, ecology, chemistry, agriculture. As an example, the application domain energy takes into account papers mentioning the keywords energy, petrol or gas.

The dimensions Author, Journal and Discipline have been added to the original model proposed in [6]. The definition of Author and Journal dimensions are straightforward. The Discipline dimension refers to ScienceDirect classification scheme (Table 2).

Finally, the last dimension considered is time (Year dimension). For several dimensions (Artifact and Objective notably), we could have refined them as multilevel dimensions.

Thus, the multidimensional model, including its instances, may be considered as the result of this step. This model is described in more details below.

3.5. Data extraction and mapping process

The last step is devoted to data extracting and mapping process.

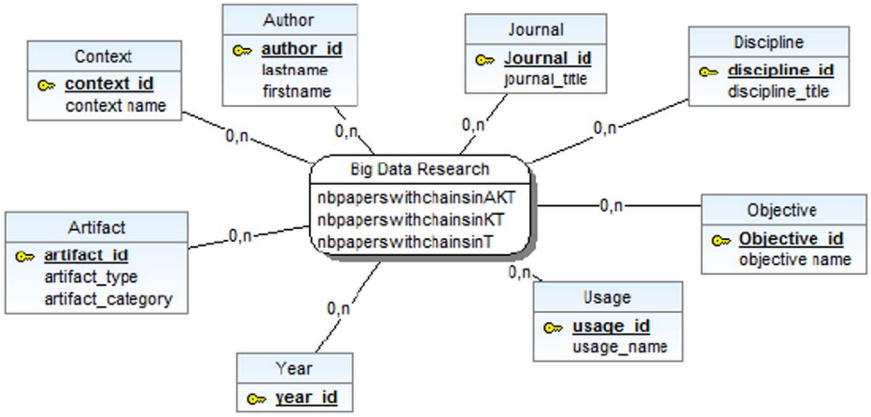

Fig. 2. Multidimensional model.

Table 2
ScienceDirect classification scheme.

Agricultural and Biological Sciences
Arts and Humanities
Biochemistry, Genetics and Molecular Biology
Business, Management and Accounting
Chemical Engineering
Chemistry
Computer Science
Decision Sciences
Earth and Planetary Sciences
Economics, Econometrics and Finance
Energy
Engineering
Environmental Science
Immunology and Microbiology
Materials Science
Mathematics
Medicine and Dentistry
Neuroscience
Nursing and Health Professions
Pharmacology, Toxicology and Pharmaceutical Science
Physics and Astronomy
Psychology
Social Sciences
Veterinary Science and Veterinary Medicine

Using the multidimensional model as a framework, we first sorted all the relevant papers. We then extracted the data. The results obtained are described in Section 4. Let us mention that we first extracted the data related to all papers containing the term "big data". The search took into account only original articles as considered by ScienceDirect. It encompasses all the journals indexed by ScienceDirect. The second step involved the extraction of relevant papers using at least one dimension of the framework. Frequencies have been derived from a final classification table.

Section 4 describes the results of our bibliometric study along the eight dimensions. Screening the abstract (A), keywords (K), and title (T) of the papers, we defined three measures which are respectively the number of original research papers published with the topic contained either in abstracts (A) or in keywords (K) or in titles (T), the number of original research papers with the topic mentioned in K, and finally the number of original research papers with the topic in T.

4. Big Data: an emerging research field

Our objective is to consider the term Big Data and to characterize the ways in which the research community used it in the research literature. This effort to understand and characterize the current state of Big Data related research was performed along the eight dimensions of the model, all of them being based on the ScienceDirect corpus. The latter enables different aggregated views of the results based on year, abstract, title, keywords, and specific search chains related to the context of the research, its objectives, the artefacts produced, as well as its usages. In addition to the overall characteristics of the publications on Big Data, a thematic contextual analysis of the abstracts, the titles, and the keywords, was performed.

In the following, we first provide answers to the research questions. We then discuss the main findings obtained. Going beyond the first results, we perform a content analysis in order to gain more insights. Finally, we present some limits of our study.

4.1. Answers to the research questions

Table 3 presents aggregate measures of big data research along the different search spaces (A, K or T) for the time horizon considered (2013–2016).

In Table 3, the first four dimensions (context, artifact, usage, and objective) allow us to select the correlation between big data research and one given context (resp. one given artifact, one given usage, and one given objective). Hence, the value 334 in the first cell means that 334 papers mention both "big data" and at least one chain among ("social", "mobile", "analytics","cloud", "internet of things") in their abstracts or keywords or titles. The following four dimensions (year, author, journal, and discipline) allow us to analyze the distribution of big data papers per year (resp. per author, per journal, per discipline). The last row summarizes the number of papers (original research) which include the chain « big data » respectively in {A and/or K and/or T}, in {K and/or T}, and in T.

Table 3
Big Data research – number of papers per dimension.

Nb. of papers *Dimensions*	With chains in A, K or T	With chains in K or T	With chains in T
Context	334	138	65
Artifact	419	52	36
Usage	113	32	9
Objective	373	54	32
Year	693	398	240
Author	693	398	240
Journal	693	398	240
Discipline	693	398	240
Total	693	398	240

RQ1: How many research papers are produced?

As of March 2016, the total number of research papers containing the chain "big data" in A or K or T is 693. In K or T this number is 398. It becomes equal to 240 in T only. Note that in [6] corresponding to July 2015, those numbers were respectively 486, 415, and 187. These numbers testify to a significant increase between the two dates considered. This is especially clear when one considers the time horizon, as described below.

RQ2: What is the annual trend of publications?

We show in Fig. 3 the annual trend of big data papers. The first dimension analyzed is time. On the graphical representation, we omitted the year 2016 (since it is not over) in order not to distort the curve. This figure illustrates the recent but explosive emergence of big data research. Note that the activity of the emerging big data community was not visible until 2012. There is a surge starting in 2013.

Such increase in a new topic is rare, especially in computer science. In order to verify this claim, in [6], we reported our findings following a comparison of "big data" buzzword with other computer science new topics which appeared in the previous decades such as object-orientation, XML, and open data. Thus, to the best of our knowledge, no comparable explosive phenomenon was observed previously.

RQ3: What are the hot topics in big data research?

In order to be able to answer the question related to the most investigated big data topics, we consider the third dimension, namely Context. The latter takes into account the focus of research on both big data and one topic of SMACIT (Cloud, Analytics, Social, Mobility, and

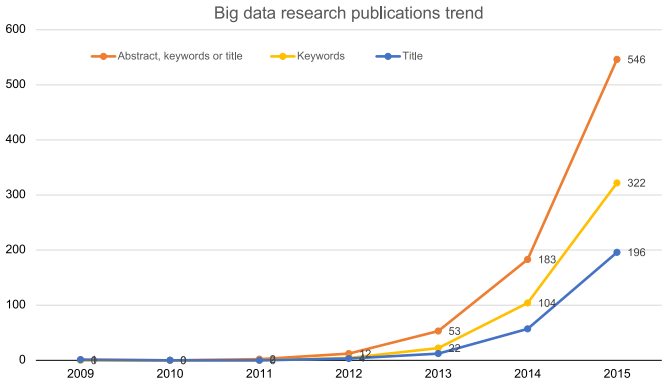

Fig. 3. Number of papers on big data per year.

Table 4
Number of papers dedicated to big data and to one of hot topics (SMACIT).

Nb of papers *Dimensions*	With chains in A, K or T	With chains in K or T	With chains in T
Cloud	154	66	25
Analytics	136	49	36
Social	99	33	11
Mobility	78	15	5
Internet of things	37	15	4

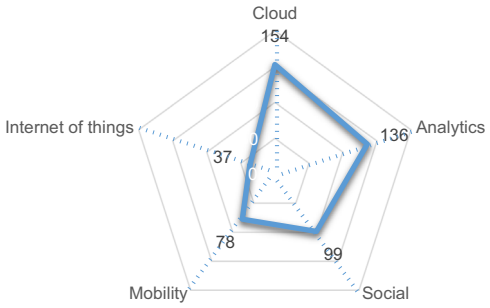

Fig. 4. Number of papers on big data addressing one of SMACIT topics.

Internet of things).

We can see in Table 4 that the most popular topics are Cloud and Analytics, while Internet of things seems to be lagging behind. Compared with the previous work presented in [6], the ranking of the topics has changed. Cloud exceeds Analytics, except when considered only the titles. In this case, papers containing both "Analytics" and "Big data" in their titles exceed those containing both "Cloud" and "Big data". However the numbers of papers related to all topics increase. In terms of evolution, we expect a real increase in the number of papers dealing with IoT topic since it is a more recent subject.

In addition, Fig. 4 provides a visual summary of publications by topics. Although Internet of things is one of the main markets of big data applications, it is not well reflected in research publications. Besides, it seems that the research on mobile data analytics is just starting. Finally, although many social networks have become popular over the years, the research on social big data is in its beginnings.

RQ4: **Why the research is produced?**

To answer this question, we have to consider the third dimension related to the objectives of the research. We searched for the term Big Data and each objective corresponding to the terms: security, quality, performance, privacy, confidentiality, integrity, traceability, reliability, scalability, and usability (Table 5).

As in the previous study [6], the performance objective dominates. Followed by quality and security. When considering columns 2 and 3, security combined with privacy is leading. Thus, security (with privacy) seems to be considered as very important in the context of big data

Table 5
Papers on BIG Data per objective.

Nb of papers *Dimensions*	With chains in A, K or T	With chains in K or T	With chains in T
Performance	242	18	13
Quality	66	7	3
Security	61	13	8
Scalability	48	3	0
Privacy	42	18	8
Reliability	18	1	0
Availability	18	0	0
Integrity	7	2	1
Usability	7	0	0
Confidentiality	3	1	1
Traceability	2	0	0

Table 6
Big Data research produces artifacts.

Nb of papers *Dimensions*	With chains in A, K or T	With chains in K or T	With chains in T
Algorithm	192	13	9
Framework	171	9	14
Architecture	99	12	5
Methodology	53	0	4
Language	32	11	3
Metric	30	1	0
Prototype	27	1	1
Ontology	19	7	1
System design	10	0	1
Guideline	5	0	0
Taxonomy	5	2	0

whereas performance remains a strong issue for computer scientists in charge of improving their algorithms. Scalability is emerging as an objective. Let us notice that, in the first column, if privacy, integrity, confidentiality, availability, and traceability are grouped with security, quality will be demoted to third place. However, one should be aware of the fact that "quality" is a portmanteau word encompassing many aspects. This is confirmed by its weak numbers in the second and third columns. Finally, let us remark that traceability and usability are relatively unexplored in big data research. That means that these objectives have not yet been revisited in the context of big data.

RQ5: What are the most frequently obtained research artifacts?

Concentrating on design science research, the fourth dimension takes into account one of the artifacts produced by the research, namely: language, meta-model, system design, ontology, taxonomy, framework, architecture, methodology, guideline, algorithm, method fragment, metric, and prototype (Table 6 and Fig. 5).

The three main artifacts produced by the research on big data are: algorithms, frameworks, and architectures. Note that the level of publications is higher compared to the results obtained in the previous study [6], evidencing a significant progress. System design, guideline and taxonomy seem not to be the main concerns of researchers in the context of big data. When considering only chains in K or T, ontology emerges as an issue confronted by big data researchers. Analyzing the last column, one can see that only two artifacts emerge, namely framework and algorithm. However, framework can be considered as a portmanteau word.

As it can be seen in Fig. 5, the grouping of several artifacts in a meta artifact, for example guideline, methodology, algorithm, and metric, as Method, allows us to exhibit a balance between three meta artifacts

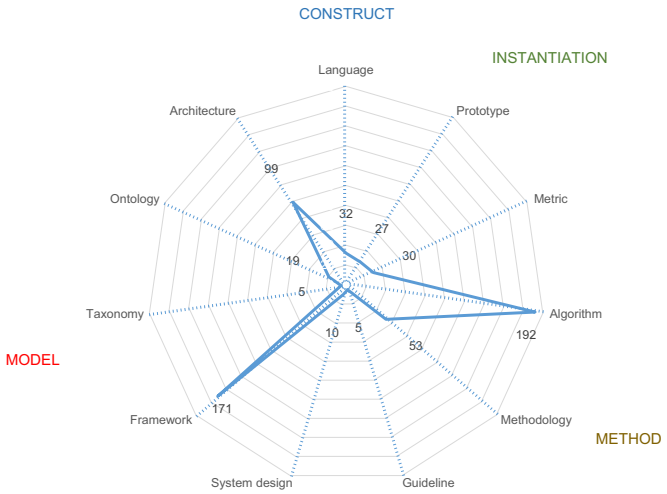

Fig. 5. Big Data research and artifacts.

Table 7
Papers on Big Data application domains.

Nb of papers *Dimensions*	With chains in A, K or T	With chains in K or T	With chains in T
Earth	83	16	8
Energy	58	13	9
Medicine	51	10	3
Ecology	48	10	2
Marketing	46	8	1
Health	44	11	2
Finance	35	4	4
Government	35	5	3
Education	22	2	2
Chemistry	10	2	2
Tourism	7	4	1
Agriculture	2	1	1

(method, model, and construct) in terms of papers published. The fourth meta artifact (instantiation) is clearly lagging behind.

RQ6: What does big data research produces?

The fifth dimension considers usage or domain applications. We searched for the number of articles containing the term Big Data and one of the terms characterizing the usage, namely: healthcare or public health, education, public sector or government, banking or finance, tourism or hospitality management, marketing or retail, earth, energy, medicine, ecology, chemistry, agriculture.

As it can be seen (Table 7), the domain earth has the highest score. It is predominant in relation to other application domains. Energy is in the second position followed by domains such as medicine, ecology, marketing, and health. Surprisingly, area such as agriculture, tourism, and chemistry have not been the object of intensive research and publications. When considering column 3, it appears that finance emerges as the third application domain in the context of big data. In terms of evolution, we can expect the government application domain to be more attractive in the near future.

Fig. 6 allows us to notice the lack of discontinuity in terms of published papers in the different application domains.

RQ7: Who are the active authors?

Table 8 shows the number of researchers working on the theme of big data. This number is fairly high (726). Note that only four researchers (Lizhe Wang, Rajiv Ranjan, Francisco Herrera, Jinjun Chen) have written 4 papers (second row), whereas 682 authors have published only one paper (fifth row). The dispersion is very large. Up to date, there are not many specialists dominating the big data research area.

RQ8: Which journals include papers on Big Data?

Table 9 shows that < < Procedia Computer Science > > is the dominant journal (200 papers) among those referenced by ScienceDirect, in terms of publications on Big Data. Far behind is < < Future Generation Computer Systems > > (55 articles). The remaining papers (471) are scattered among several journals. Note that the journal "Big Data research" has published only 19 original research papers. The reason is that the journal is too recent in order to be visible.

Even though the number of published papers in some journals is weak, we have to admit that several non computer science journals show a real interest in the big data topic.

RQ9: What are the active disciplines?

Our aim is to determine the most active disciplines in terms of research and publications on Big Data. We used ScienceDirect dis-

Table 8
Main Big Data authors.

Nb of papers	Nb of authors
4	4
3	13
2	27
1	682
Total	726

Table 9
Main Big Data journals.

Journal	# Papers
Procedia Computer Science	200
Future Generation Computer Systems	55
Neurocomputing	33
Information sciences	20
Knowledge-based systems	19
Big Data Research	19
Expert Systems with Applications	16
Transportation Research Part C: Emerging Technology	14
Technological forecasting and social change	12
Computer Law & Security Review	11
Information Systems	10
Journal of Parallel and Distributed Computing	10
Decision Support Systems	10
Journal of Systems and Software	10
Computers, Environment and Urban Systems	9
Ad Hoc Networks	9
Digital Investigation	9
Neural networks	9
Environmental Modelling & Software	9
Network Security	8
International Journal of Production Economics	8
Parallel Computing	8
Procedia Technology	7
Others	178

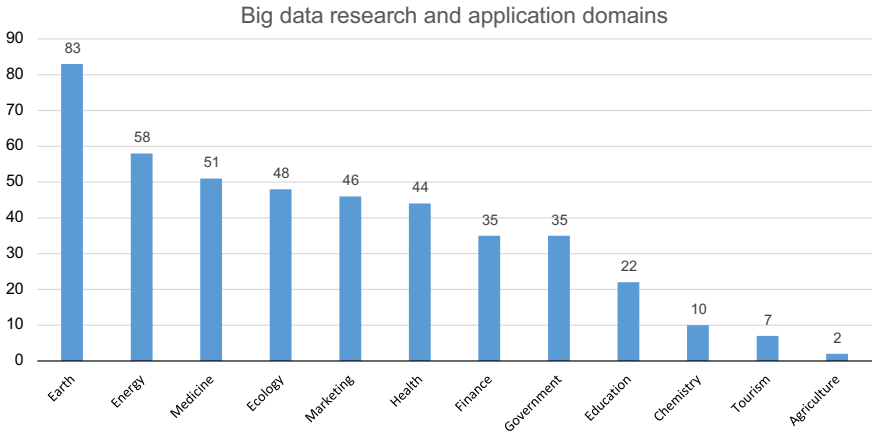

Fig. 6. Number of papers describing applications of big data in different domains.

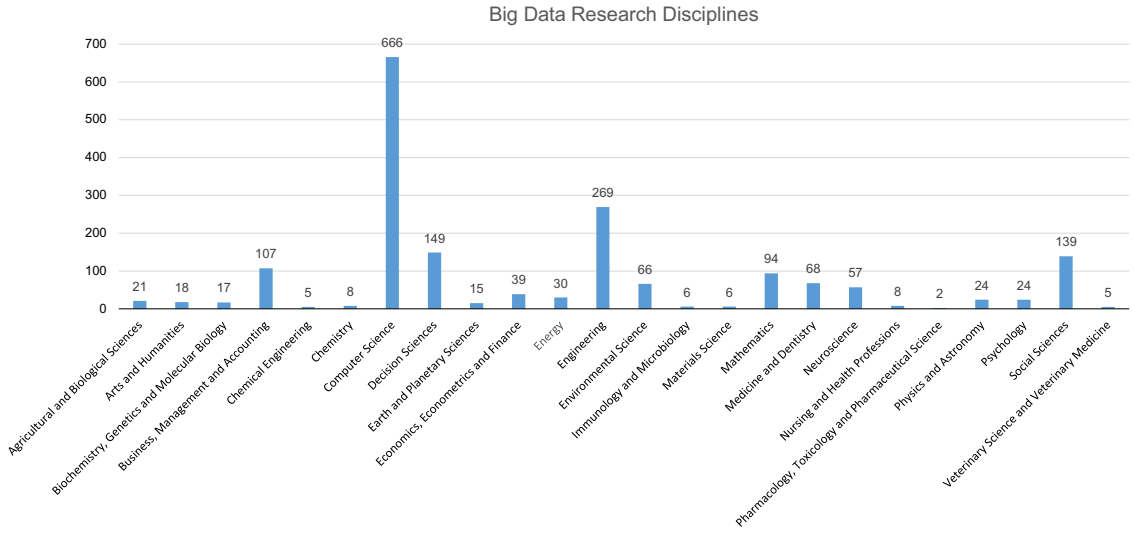

Fig. 7. Number of big data papers per discipline.

ciplines classification scheme. It is not surprising that Computer Science is by far the most active discipline (666 published papers). Fig. 7 shows that Computer Science is followed by the Engineering discipline (269 papers). Aside from the scientific disciplines (computer science: 666, engineering: 269, decision science: 149, mathematics: 94), those who publish the most are Social Sciences (139) and Business, Management and Accounting (107). Let us remind that the large number of publications is due to the fact that some journals are multidisciplinary.

4.2. Discussion

The discussion will focus on a comparison between the previous study and the one presented in this article. Overall, Big Data has a large coverage but a short history. In our previous study [6], we found a growth of research articles about big data starting mainly from 2011. This trend intensified. The number of articles has increased from 486 to 693 with a significant surge in 2013. Publications combining Big Data and SMACIT numbered only 264 and started to appear in 2011. It reaches 504 articles as of 2015. Another interesting finding is related to the focus level. In our previous study, this focus decreased from 486 to 187 publications. In our current study, this focus decreases from 693 to 240 articles. The decay rate is twice. In other words, the existence of the term Big Data in an abstract does not imply that the main subject of the research is on Big Data. This is not the case when the term is present in the title meaning that Big Data is the core of the paper. In our previous study, the association of the term Big Data with Analytics or Cloud is much stronger than with the other terms of SMACIT. We found a change in our current study. Cloud exceeds Analytics. As a consequence, it seems that researchers were not fully involved in research combining Big Data and IoT or mobility or social media. This is not the case in 2015. The interest in these topics is bigger than before. As for quality objectives, the dominant topic area was performance in the previous study. This is still the case as of 2015. However, scalability is emerging as an objective. Moreover, if privacy, integrity, confidentiality, availability, and traceability are grouped with security, quality will be demoted to third place. This is a real change since our previous study. In terms of artefacts produced by research on Big Data, two topics have emerged in our previous study: framework and algorithm. This seemed to indicate that researchers on Big Data provided more frameworks and algorithms than taxonomies or ontologies or other artifacts. This is not the case as of 2015 since the three main artifacts produced by the research on big data are: algorithms, frameworks, and architectures. Our previous results showed that Usages of Big Data did not attract researchers except for two application domains: Marketing and retail, and Healthcare. The situation today (beginning of 2016) is different. The domain earth has the highest score and seems to attract more researchers. Energy, medicine, ecology, marketing, and health are good candidates for researchers. As a result of our systematic mapping, Fig. 8 sketches the main characteristics of Big Data Research by focusing on the three main values of each dimension.

4.3. Beyond the mapping study: a content analysis

We tackled an important question related to the sense of what our research and analysis yields. Our aim is to shed more light on the evolving field of big data research and to provide more insights. We selected the subset of papers containing the “big data” string in their title and published between 2013 and 2015, which means 393 papers. Moreover, among those 393 papers, we extracted papers containing at least one of SMACIT strings in their titles. The number of papers obtained is 64. We performed a content analysis on this set of papers.

To gain more insight in the big data research, we mapped the 64

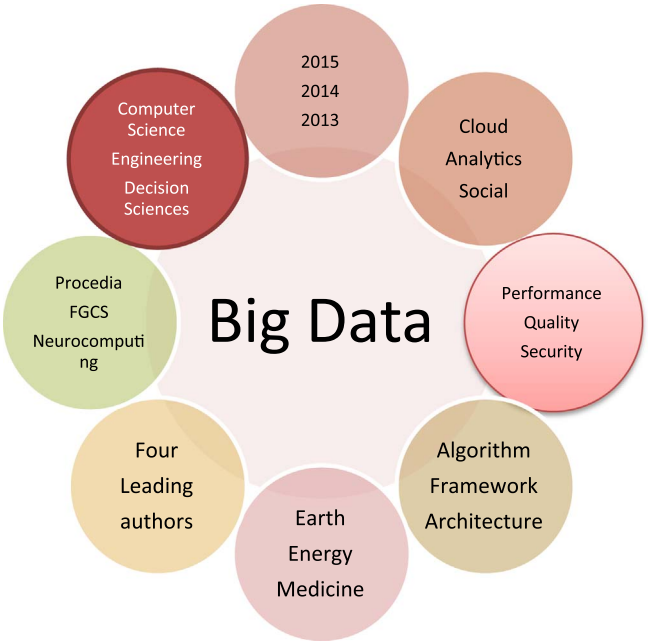

Fig. 8. A rough mapping of Big Data research.

Table 10
Content analysis.

Big Data		Social	Mobile	Analytics	Cloud	IoT	Total+
Data storage	Clustering	0	0	6	7	0	**13**
	Replication	0	0	0	1	0	**1**
	Indexing	0	0	1	0	0	**1**
	Other storage	0	1	2	5	0	**8**
Pre-processing	Transmission	0	2	3	2	1	**8**
	Cleansing	0	0	1	1	1	**3**
Processing	Classification	4	1	8	3	1	**17**
	Prediction	2	0	10	1	0	**13**
Others	Societal aspects	0	0	4	0	0	**4**
	Big data adoption	0	0	1	0	0	**1**
	Project management	0	0	2	1	0	**3**
	Research avenues	0	0	0	1	0	**1**
Total+		**6**	**4**	**38**	**22**	**3**	**73**

papers within the seven branches of Siddiqa's Big Data Management techniques taxonomy [52]. Siddiqa et al. first decomposed Big Data Management in three axes: data storage, pre-processing, and processing. The next decomposition level splits data storage into clustering, replication, and indexing. Pre-processing is composed of transmission and cleansing. Finally processing consists of classification and prediction. In order to take into account the whole set of 64 papers, we had to extend Siddiqa's taxonomy adding 1) a data storage subcategory (other storage) for other data storage techniques dealing mainly with security aspects; and 2) a complementary category (others) for papers dealing with other aspects, such as societal subjects, adoption models, project management, and research avenues. The mapping was achieved through a screening process of the abstract and, when necessary, the whole paper. The result of our mapping is given below (Table 10). Each cell corresponds to the number of papers addressing big data techniques and dealing with one of the SMACIT topics. For example, 6 papers containing both Analytics and Big data in their title address clustering issues.

+ Some papers may be counted several times when they deal with more than one SMACIT topics.

As it can be seen at Table 10, a majority of papers focus mainly on three big data techniques, i.e. classification, clustering, and prediction. On the contrary, techniques such as replication, indexing, and cleansing don't attract many research papers. In terms of SMACIT axes, the dominant topic is analytics distributed along the different techniques, followed by the cloud topic. Crossing the dimensions, analytics papers are dedicated mainly to processing techniques (classification as well as prediction), which is not surprising. Regarding cloud topic, papers focus naturally on data storage techniques.

4.4. Threats to validity

Several limitations can characterize our study, such as study design, impact, and data limitations. In the study design category, let us mention the choice of ScienceDirect library. However, the latter offers a very adequate search engine and a wide variety of publications, both in computer science and in other research disciplines. The second limitation, also related to study design, resides in the choice of only journal papers. This results in analyzing only mature research and not research in progress. We argue that it is a more reliable measure tool since journal papers provide us with a more comprehensive overview of a given field even if it can generate a time lag.

As for impact limitation, we can mention the choice of sticking to "big data" keyword as a mandatory component of the selected papers. Many authors work on big data without using this term neither in their abstracts, nor in their keywords and titles. We choose not to consider these papers since we were interested mainly on research surfing on the "big data" buzzword. In future research we plan to extend our scope by including all the keywords used in Big Data Research Journal. Let us mention that we conducted a specific study to estimate the resulting bias. Based on Big Data Research Journal articles, published since 2014, we checked how many papers did not contain the "big data" string neither in the title nor in the abstract nor in the keywords. We found only six out of thirty-two papers. If we generalize this estimation to the whole set of journals, the maximum bias will be less than 19%.

The last category of limitation is related to data limitations. We did not perform a deep analysis of selected papers allowing us to determine whether they tackle new big data issues characterized by the three Vs (Volume, Velocity, Variety) or improve past techniques without taking into account at least two Vs. Another limitation is due to our choice not to take into account the data life-cycle. It would require a different bibliometric study without using keywords.

5. Conclusion and future research

Research on Big Data has seen an explosion of publications since 2013. In order to characterize the emergence of Big Data as a research topic, this paper looks at this topic from eight different perspectives: the timeline, the context, the objectives, the artifacts created, the applications (usages), the authors, the journals, and the academic disciplines. We explore the term Big Data using the peer reviewed literature defining three focus levels. A systematic mapping study was performed to identify and analyze research on big data, covering publications between 2006 and 2016. It included all publications of journals indexed in ScienceDirect database. This digital library was selected because they are the most important repositories for research in computer science and more precisely in big data. We considered only journals articles defined as "original research". The numbers presented in this article reflect the indexed publications of ScienceDirect in March 2016. The main results obtained are: (i) There is a significant growth of research articles about Big Data since 2013. (ii) There is a diversity of interest by researchers on issues such as the objectives, the artefacts produced, the quality criteria used, and the usages and applications of Big Data. (iii) Big Data Research focuses mainly on three techniques, i.e. clustering, classification, and prediction. It is a multiple viewpoint for practitioners who want to understand what Big Data research produces. Each aspect is also a guideline for researchers in the field helping in the elicitation of correct dimensions (e.g. which application domain should be developed). One limitation of this research is related to the database searched. A similar search on other databases might result in slightly different findings. The other limitations were discussed above and represent future research avenues. Future research also includes a systematic literature review including a deep content analysis of the complete articles. Our multidimensional model contains eight flat dimensions. Another future research will be to enrich it with dimension hierarchies.

References

[1] J. Gantz, D. Reinsel, Extracting value from chaos, IDC iView (2011) 1–12.
[2] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, et al., Big Data: the next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute, San Francisco, 2011.
[3] I. Intel, Center: Planning Guide: Getting Started with Hadoop, Steps IT Managers Can Take to Move Forward with Big Data Analytics, 2012.
[4] T. Davenport, P. Barth, R. Bean, How Big Data is Different. MIT Sloan Mgt Review, 2012.
[5] X. Jin, B.W. Wah, X. Cheng, Y. Wang, Significance and challenges of big data research, J. Big Data Res. 2 (2) (2015) 59–64.
[6] J. Akoka, I. Comyn-Wattiau, N. Laoufi, Research on Big Data – Characterizing the Field and its Dimensions, Advances in Conceptual Modeling, MOBID Workshop, Proceedings of ER2015 Workshops., October 2015, Vol. 9382, pp. 173–183, Series Lecture Notes in Computer Science (LNCS), Stockholm, Sweden, ⟨http://doi.org/10.1007/978-3-319-25747-1_18⟩.
[7] O.R. Team, Big Data Now: Current Perspectives from O'Reilly Radar Sebastopol, USA: O'Reilly Media, CA, 2011.
[8] M. Grobelnik, Big Data Tutorial, ⟨http://videolectures.net/eswc2012_grobelnik_big_data/⟩.
[9] D. Laney, 3-d data management: controlling data volume, velocity and variety. META Group Research Note, 2001.
[10] S. Sagiroglu, D. Sinanc, Big Data: a review, in: IEEE Int. Conf. on CTS, 2013.
[11] J.S. Ward, A. Barker, Undefined by data: a survey of big data definitions, - arXiv preprint arXiv:1309.5821- arxiv.org, 2013.
[12] A. De Mauro, M. Greco, M. Grimaldi, What is big data? A consensual definition and a review of key research topics, in: AIP Conference Proceedings- cloudtribes.com, 2015.
[13] A. Cuzzocrea, I.Y. Song, K. Davis, Analytics over large-scale multidimensional data: the big data revolution!, in: Proceedings of the 14th international workshop on Data Warehousing and OLAP. New York, New York, USA: ACM, 2011, pp. 101–103.
[14] C. Bizer, P. Boncz, M.L. Brodie, O. Erling, The meaningful use of Big Data: four Perspectives, SIGMOD 40 (4) (2011) 56–60.
[15] A. Jacobs, The pathologies of Big Data, Commun. ACM 52 (8) (2009) 36.
[16] H. Chen, R.H.L. Chiang, V.C. Storey, Business intelligence and analytics: from big data to big impact, MIS Q. 36 (4) (2012) 1165–1188.
[17] S. Madden, From databases to big data, IEEE Comput. 16 (3) (2012) 4–6.
[18] M. Chen, S. Mao, Y. Liu, Big Data: a survey (171–20)Mobile Netw. Appl. 19 (2014) (171–20).
[19] K. Kambatla, G. Kollias, V. Kumar, A. Grama, Trends in Big Data analytics, J. Parallel Distrib. Comput. 74 (2014) 2561–2573.
[20] A. Labrinidis, H.V. Jagadish, Challenges and opportunities with Big Data, Proc. VLDB Endow. 5 (12) (2012) 2032–2033.
[21] A. Gandomi, M. Haider, Beyond the hype: big Data concepts, methods, and, analytics, Int. J. Inf. Manag. 35 (2015) 137–144.
[22] H. Fang, Z. Zhang, C.J. Wang, M. Daneshmand, A survey of big data research, IEEE Netw. 29 (5) (2015) 6–9.
[23] M.D. Assunçao, R.N. Calheiros, S. Bianchi, M.A.S. Netto, R. Buyya, Big Data computing and clouds: trends and future directions, J. Parallel Distrib. Comput. 79-80 (2015) 3–15.
[24] I.A.T. Hashem, I. Yaqoob, N.B. Anuar, S. Mokhtar, A. Gani, The rise of "big data" on cloud computing: review and open research issues, Inf. Syst. 47 (2015) 98–115 (Elsevier).
[25] M. Chen, S. Mao, Y. Zhang, V.C.M. Leung, Big Data: Related Technologies Challenges and Future Prospects, Springer, Cham, Heidelberg, New York, Dordrecht, London, 2014.
[26] D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, L. Haas, A. Halevy, J. Han et al., Challenges and Opportunities with Big Data. A Community White Paper Developed by Researches Across the United States.
[27] C.K. Emani, N. Cullot, C. Nicolle, Understandable Big Data: a survey, Computer Sci. Rev. 17 (2015) 70–81.
[28] J.C. Jackson, V. Vijayakumar, M.A. Quadir, C. Bharathi, Survey on programming models and environments for cluster, cloud, and grid computing that defends big data, Procedia Comp. Sci. 50 (2015) 517–523 (ISSN 1877-0509).
[29] C.L. Philip Chen, C.Y. Zhang, Data-intensive applications, challenges, techniques and technologies: a survey on Big Data, Inf. Sci. 275 (2014) 314–347.
[30] J. Zhang, X. Yao, G. Han, Y. Gui, A survey of recent technologies and challenges in big data utilizations, in: 2015 International Conference on Information and Communication Technology Convergence (ICTC), Oct, 2015.
[31] S. Sharma, U.S. Tim, J. Wong, S. Gadia, S. Sharma, A brief review on leading big data models, Data Science Journal- jlc.jst.go.jp, 2014.
[32] D.S. Terzi, R. Terzi, S. Sagiroglu, A survey on security and privacy issues in big data, in: Proceedings of the 10th International Conference for Internet Technology and Secured Transactions (ICITST), IEEE, 2015, pp. 202–207.
[33] E. Sahafizadeh, M.A. Nematbakhsh, A survey on security issues in Big Data and NoSQL, ACSIJ Adv. Comput. Sci.: Int. J. 4 (4) (2015) No.16.
[34] J. Liu, J. Li, W. Li, J. Wu, Rethinking big data: a review on the data quality and usage issues, ISPRS J. Photogramm. Remote Sens. (2015).
[35] H. Zhang, G. Chen, B.C. Ooi, K.L. Tan, In-memory big data management and processing: a survey, IEEE Trans. Knowl. Data Eng. 27 (7) (2015) 1920–1948.
[36] L. Zhang, A. Stoffel, M. Behrisch, et al., Visual analytics for the big data era—a comparative review of state-of-the-art commercial systems, in: Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on. IEEE, 2012, pp. 173–182.
[37] D. Singh, C.K. Reddy, A survey on platforms for big data analytics. D Singh, CK Reddy, Journal of big data, Springer, 2015.
[38] X. Liu, N. Iftikhar, X. Xie, Survey of real-time processing systems for big data, in: Proceedings of the 18th International Database Engineering and Applications Symposium, ACM, 2014, pp. 356–361.
[39] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, A survey of clustering algorithms for big data: taxonomy and empirical analysis, Emerg. Top. Comput. IEEE Trans. 2 (3) (2014) 267–279.
[40] A.S. Shirkhorshidi, S. Aghabozorgi, T.Y. Wah, T. Herawan, Big Data Clustering: A Review – In Computational Science and Its Applications–ICCSA, Springer International Publishing, Cham, Heidelberg, New York, Dordrecht, London, 2014, pp. 707–720.
[41] N. Khan, I. Yaqoob, I.A.T. Hashem, Z. Inayat, et al.: Big data: survey, technologies, opportunities, and challenges, The Scientific World Journal, 2014.
[42] H. Ekbia, M. Mattioli, I. Kouper, G. Arave, Big data, bigger dilemmas: a critical review, J. Assoc. Inf. Sci. Technol. 66 (8) (2015) 1523–1545.
[43] J. Archenaa, J. Mary Anita, E.A.: a survey of big data analytics in healthcare and government, Procedia Comput. Sci. 50 (2015) 408–413 (ISSN 1877-0509).
[44] P.B. Goes, Big Data and IS research methods, MIS Quarterly 38(3), 2014, pp. Iii viii.
[45] T. Hansmann, P. Niemeyer, Big Data – Characterizing an Emerging Research Field using Topic ModelsIEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014.
[46] M. Pospiech, C. Felden, Big Data: A State-of-the-Art, in: Americas Conference on Information Systems, 2012.
[47] S.F. Wamba, S. Akter, A. Edwards, G. Chopin, How 'big data' can make big impact: findings from a systematic review and a longitudinal case study, Int. J. Prod. Econ. 165 (2015) 234–246.
[48] K. Petersen, R. Feldt, S. Mujtaba, M. Mattsson, Systematic mapping studies in software engineering, in: 12th International Conference on Evaluation and Assessment in Software Engineering, Vol. 17(1), 2008, pp. 1–10.
[49] B.A. Kitchenham, S. Charters, Guidelines for performing systematic literature reviews in software engineering, Technical Report EBSE-2007-01, School of Computer Science and Mathematics, Keele University, 2007.
[50] N. Prat, I. Comyn-Wattiau, J. Akoka, Artifact evaluation in information systems design science research – a holistic view, in: PACIS 2014 Proceedings, Paper 23, 2014.
[51] S.T. March, G.F. Smith, Design and natural science research on information technology, Decis. Support Syst. 15 (4) (1995) 251 (66).
[52] A. Siddiqa, I.A.T. Hashem, I. Yaqoob, M. Marjani, S. Shamshirband, A. Gani, F. Nasaruddin, A survey of big data management: taxonomy and state-of-the-art, J. Netw. Comput. Appl. 71 (2016) 151–166.