New Trends in Big Data Profiling

Chapter · July 2022

DOI: 10.1007/978-3-031-10461-9_55

4 authors, including:

Julia Couto
Pontifícia Universidade Católica do Rio Grande do Sul
18 PUBLICATIONS **80** CITATIONS
SEE PROFILE

Duncan D. Ruiz
Pontifícia Universidade Católica do Rio Grande do Sul
111 PUBLICATIONS **942** CITATIONS
SEE PROFILE

New trends in big data profiling

Júlia Colleoni Couto, Juliana Damasio, Rafael Bordini, and Duncan Ruiz

School of Technology, PUCRS University, Porto Alegre, RS, Brazil,
julia.couto@edu.pucrs.br

Abstract. A known challenge related to big data is that data ingestion occurs continuously and at high speed, and the data profile can quickly vary because of this dynamism. Data profiling can range from simple summaries to more complex statistics, which is essential for understanding the data. For a data scientist, it is essential to know the profile of the data to be handled, and this information needs to be updated according to the new data that is continuously arriving. This paper reviews the literature about how data profiling is being used in big data ecosystems. We search in eight relevant web databases to map the papers that present big data profiling trends. We focus on categorizing and reviewing the current progress on big data profiling for the leading tools, scenarios, datasets, metadata, and information extracted. Finally, we explore some potential future issues and challenges in big data profiling research.

Keywords: big data, data profiling, data lakes

1 Introduction

One of the ways to present information about the data we have stored is by generating data profiles. Data profiling creates data summaries of varied complexities, from simple counts, such as the number of records [1], to more complex inferences, such as functional data dependencies [2]. Data profiling allows us to understand better the data we have, and it is essential to help us choose the tools and techniques we will use to process the data according to its characteristics. It is useful for query optimization, scientific data management, data analytics, cleansing, and integration [3]. Data profiling is also useful in conventional file systems (such as those used in Windows and Linux), but it is essential in big data environments, mainly due to volume, velocity, and variety.

In this paper, we review the literature about data profiling in big data. We aim to understand the big picture of how data profiling is being done in big data. We present the most used tools and techniques, the types of data, the areas of application, the type of information extracted, and the challenges related to the big data profiling research field. To the best of our knowledge, no previous studies have systematically addressed this issue.

To achieve our goal, we perform a Systematic Literature Review (SLR), based on eight electronic databases, containing papers published from 2013 to 2019. We started with 103 papers, and, using inclusion and exclusion criteria, we selected

20 papers for the final set. We use the PRISMA checklist [4] to help us improve the quality of our report, and we use the process suggested by Brereton et al. [5] to plan the steps to follow. We use the Kappa method [6] to enhance results quality and measure the level of agreement between the researchers. Two researchers worked on analyzing the papers to reduce bias, and two others were involved in case of disagreement.

Our main contribution is related to characterizing new trends in big data profiling. For instance, we found that R, Python, and Talend are the most used tools, and we identified seven areas of application, namely, automotive, business, city, health, industry, web, and others. We also mapped the datasets they use in those areas, mostly based on online repositories, real-world datasets, and data auto-generated. Our analysis also shows that most papers use data profiling to generate metadata rather than using metadata to generate data profiling. Furthermore, data type, origin, and temporal characteristics are among the most frequent metadata presented in the papers.

We also create a classification for the type of information extracted using data profiling (statistics, dependencies, quality, data characteristics, data classification, data patterns, timeliness, and business processes and rules). Finally, we present and discuss 15 challenges related to big data profiling: complexity, continuous profiling, incremental profiling, interpretation, lack of research, metadata, online profiling, poor data quality, profiling dynamic data, topical profiling, value, variability, variety, visualization, and volume. We believe that our findings can provide directions for people interested in researching the field of big data profiling.

2 Materials and methods

An SLR is a widely used scientific method for systematically surveying, identifying, evaluating, and interpreting existing papers on a topic of interest [7]. We performed an SLR using the protocol proposed by Brereton et al. [5]. This method has 3 phases, named *Plan, Conduct*, and *Document*. Also, we chose to develop and report our systematic review following the Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) [4] because the document helps us build the protocol and best arrange the items to report. In the following sections, we detail how we perform each phase.

2.1 Plan review

The planning phase introduces the processes and steps to ground the SLR, and it should be carefully done because it is the basis of all subsequent research. In this phase, we define research questions, develop, and assess the review protocol.

Specify research question Our main goal is to identify how data profiling is being used in the big data context. To do this, we create the Research Questions

(RQ) presented in Table 1, which are important to we can get an overview of big data profiling.

The RQ1 is useful to understand how to perform big data profiling, so people who would like to start working with big data profiling can start by exploring the most used tools. RQ2 helps us understand the main areas of applications and what kind of datasets are being used to report studies on data profiling, so beginners in big data profiling can focus on some areas or specifics datasets to start exploring profiling, for instance.

In RQ3, we are interested in understanding the type of metadata collected by the papers, and it can help people who will develop a big data profiling application to map the most important characteristics of the data to be presented. RQ4 explores the most commonly presented type of information. Finally, RQ5 maps the main challenges pointed by the selected papers, and thus we suggest future research directions for big data profiling.

We use the PICO (Population, Intervention, Comparison, and Outcome) and PICo (Population, Interest, and Context) to help in formulating our RQs. PICO and PICo are similar evidence-based models that can be combined and used to improve the research's delimitation, clarify the scope, and elaborate the research question. Table 2 presents our research scope.

Developing the review protocol We selected eight relevant computer science electronic databases to develop and apply our search protocol: Scopus, IEEE Xplorer, Springer, Google Scholar, Science Direct, ACM, Web of Science, and arXiv. We included papers published in English, regardless of the year of publication. We do not specify a start date because we aim to map data profiling's evolution in big data since its beginning.

Afterward, we identify the most important keywords related to our research question, such as `"data profiling" AND ("big data" OR "data lake")`. We combine these terms to create the search expression according to each electronic database's mechanism (see Table 3). For example, in ArXiv and ACM, we joined two search strings since the results obtained using both were more aligned to what we expected. We performed the searches in the abstract, title, and keywords fields.

We defined a control study to validate the search expression. A control study is a primary study resulting from a non-systematized web-search, which is known

Table 1: **Research questions.**

Nº	**Question**
RQ1	*What are the tools for big data profiling?*
RQ2	*What are the areas of application and datasets reported to be profiled?*
RQ3	*What type of metadata did the papers collect?*
RQ4	*Which information is extracted using data profiling?*
RQ5	*What are the challenges in big data profiling?*

Table 2: **PICO and PICo definitions.**

PICO	PICo
Population: Big data systems **Intervention**: Data profiling **Comparison:** Data warehouses **Outcome:** Tools and challenges	**Population:** Big data systems **Interest:** Tools and challenges **Context:** Data profiling

to answer our research questions. We use it to check if the search strings are adequate. If this paper were in the electronic database, it had to come up in the search with the search string that we previously defined. If the search did not return the control study, the search string needed to be adjusted until they did so. We chose the following control study: Juddoo, Suraj. "Overview of data quality challenges in the context of Big Data." *2015 International Conference on Computing, Communication and Security (ICCCS)*. IEEE, 2015. [3]. We choose this paper because it is highly related to our research, because it presents a related literature review and answers some of our research questions.

Assessing quality of the studies We followed the selection criteria for the inclusion and exclusion of papers to get only results related to our research topic. The papers we accepted met all the following criteria:

- Be qualitative or quantitative research about data profiling in big data.
- Be available on the internet for downloading.
- Present a complete study in electronic format.
- Be a paper, review, or journal, published on the selected electronic databases.

The papers we rejected met at least one of the following criteria:

- Incomplete or short paper (less than four pages).
- Unavailable for download.
- Duplicated paper.
- Written in a language other than English.
- Paper is not about data profiling in big data.
- Literature review or mapping (this criteria was only used for the review about data integration in data lakes).
- Ph.D., M.Sc., or Undergraduate theses.

Validating review protocol One researcher (JC) developed the review protocol and made several trials changing the search string to obtain results relevant and aligned to the research question. Then, another researcher (JD) performed the second review. They made new adjustments together, based on their reviews. Based on this validation, we agreed to develop the SLR using the protocol we present here.

Table 3: **Search strings for each electronic database.**

Electronic Database	Search String
Scopus	`(TITLE-ABS-KEY ("data profil*") AND TITLE-ABS-KEY ("big data" OR "data lake*"))`
IEEE Xplore	`("All Metadata":"data profiling" AND ("big data" OR "data lake"))`
Springer	https://link.springer.com/search?dc.title=%22data+profiling%22+%28%22big+data%22+OR+%22data+lake%22%29&date-facet-mode=between&showAll=true
Google Scholar	`allintitle: "big data" OR "data lake" OR"data profiling"`
Science Direct	`Title, abstract, keywords: "data profiling" AND ("big data" OR "data lake")`
ACM	`(Searched for acmdlTitle:(+"data profiling" +"big data") OR recordAbstract:(+"data profiling" +"big data") OR keywords.author.keyword:(+"data profiling" +"big data")) JOIN(Searched for acmdlTitle:(+"data profiling" +"data lake") OR recordAbstract:(+"data profiling" +"data lake") OR keywords.author.keyword:(+"data profiling" +"data lake"))`
Web of Science	`(from all databases): TOPIC: ("data profiling") AND TOPIC: ("big data" OR "data lake") Timespan: All years. Databases: WOS, DIIDW, KJD, RSCI, SCIELO. Search language=Auto`
arXiv	`(Query: order: -announced_date_first; size: 50; include_cross_list: True; terms: AND all="data profiling"; AND all="big data") JOIN (Query: order: -announced_date_first; size: 50; include_cross_list: True; terms: AND all="data profiling"; AND all="data lake")`

2.2 Conducting the review

In this phase, we start applying the protocol we previously defined. To do so, we apply the search string to each electronic database and extract the results in a BibTeX file format. Only Springer and arXiv do not facilitate this process, so we have to select each register, copy its BibTeX, and then consolidate it into a single file. Also, Google Scholar has a slightly different process, where we have to log-in to a Google account, run the search, mark each result as favorite and export the results, 20 at a time, and then we also have to consolidate it in a single file. Of course, this process refers to the available version of the web searchers we used when conducting the review phase (Dec. 2019), and for each one, the process can evolve or change in future versions.

Table 4: **Kappa results through each iteration — Table based on Landis & Koch [8].**

Kappa values	Strength of agreement	Value
<0	Poor	
0 – 0.20	Slight	
0,21 – 0,40	Fair	
0,41 – 0,60	Moderate	
0,61 – 0,80	Substantial	
0,81 – 1	Almost perfect	0.84

Identifying relevant research To certify that our research questions were not already answered in previous work, we started to search the literature for related work. Using the search string, we found 103 papers to be analyzed.

Extracting the required data We used the StArt[1] tool to help us organize and classify the papers. We register our search protocol at StArt, and then we import all results extracted from each electronic database. StArt has an execution process having 3 phases:

- Identification: we register the databases, create search sessions, and import the BibTeX files for each database.
- Selection: we read the title, abstract, and keywords for all papers and apply the selection criteria.
- Extraction: we find and download all papers we accepted in the selection to check if they answer our research questions. Only the ones that match all inclusion criteria and none of the exclusion criteria are accepted.

To reduce bias, we split the work between two researchers. First, in the selection phase, one author (JC) applied the papers' selection criteria and defined each paper as accepted or rejected. Then, a second researcher (JD) individually reviewed the accepted and excluded papers. When the authors (JC and JD) disagreed, they discussed to reach a consensus. If there is still no consensus, we contact the other two authors were contacted to help decide.

We used the Kappa Method to measure interrater reliability to measure the level of agreement between the researchers [6]. According to Landis & Koch [8], we can interpret the results we obtain using Kappa according to the scale we present in Table 4. We can see in that Table that we achieved an "almost perfect" agreement. That happened because the researchers discussed the main objectives before starting the SLR, so they were aligned.

2.3 Document review

Synthesizing data We accepted a total of 20 papers. Table 5 presents the number of papers per electronic database. In this Table, we can see that most of

[1] http://lapes.dc.ufscar.br/tools/start_tool

the papers came from Scopus and IEEE Xplore. Although Google Scholar and Scopus indexes most sources, when one paper is found in more than one engine, we kept only the paper available on the original database.

Regarding the 83 papers we rejected, most rejected papers presented a duplicated study (42 papers), or they were not papers about data profiling in big data (32 papers). We also rejected two incomplete or short papers, two unavailable papers, two papers written in another language than English (one in Japanese and one in Spanish), and two theses.

Another interesting aspect we can see in Table 5 is that 1/3 of the papers we accepted are from Scopus. that happened because Scopus [28] is the largest database of abstracts and scientific citations, compiling more than 71 million records, 23 million titles, and 5,000 publishers, among them ACM, Elsevier, IEEE, Springer, etc. Figure 1 presents the PRISMA flow diagram with the number of papers accepted in each step of the research.

Study characteristics: We found results between the years 2013 to 2019, being 2018 and 2019 the years with most of the papers (Figure 2). This result demonstrates the growing interest of researchers on this topic over the years.

Also, we analyze papers by country based on the first author's institutional affiliation. The review included studies from eleven countries. Figure 3 shows that the highest concentration of published research is in the United States (4 papers) and Germany (4 papers).

3 Results

In this section, we present an outline of the approaches reported in the papers. Among the papers, three of them reviewed the literature on data profiling tasks and tools classification [12,25,26]; two papers are about general classification [17,20]; two papers are about data wrangling [27,13]; three are about data lakes [11,9,24]; five papers are about data quality [16,19,23,3,18]; and five present varied approaches [14,22,10,15,21].

In [12,25,26], the authors reviewed the literature for classifying data profiling tasks and tools. Dai et al. [26] review the literature about data profiling, and

Table 5: **Papers per electronic database.**

Electronic Database	Initial	Accepted papers
ACM	5	2 papers: [9,10]
arXiv	2	0 paper
Google Scholar	17	1 papers: [11]
IEEE Xplore	11	5 papers: [12,13,14,15,3]
Science Direct	3	2 papers: [16,17]
Scopus	38	7 papers: [18,19,20,21,22,23,24]
Springer	8	2 papers: [25,26]
Web of Science	19	1 papers: [27]

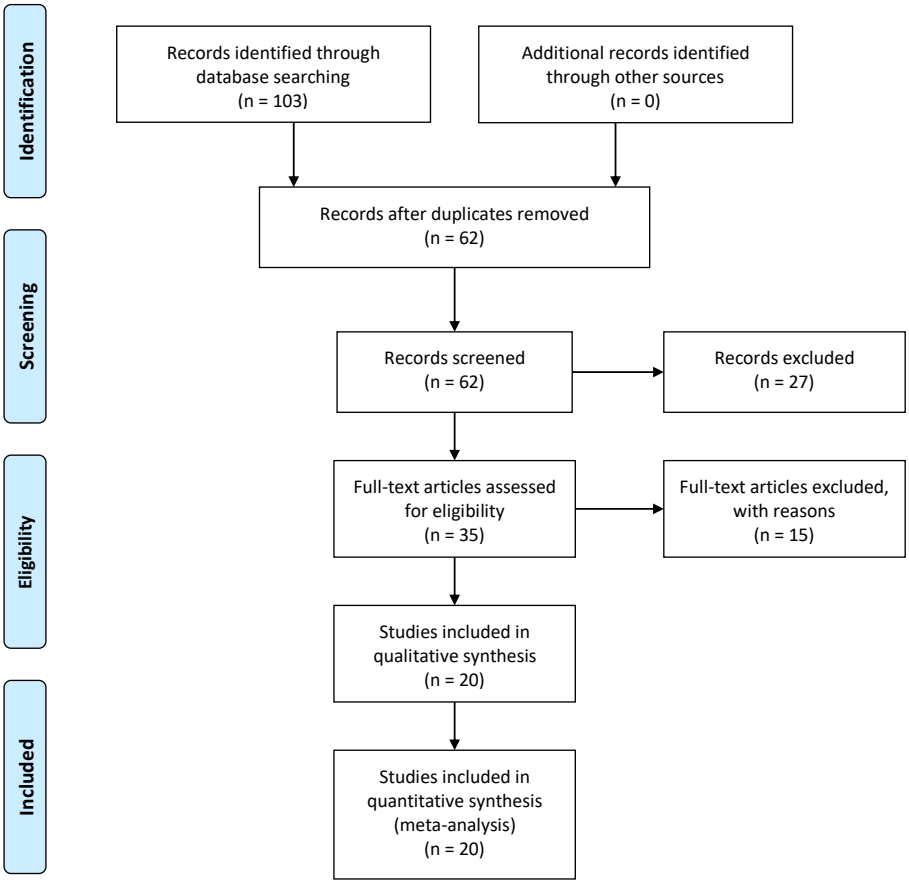

Fig. 1: **PRISMA flow diagram.**

then they propose a new definition and classification for data profiling tasks and existing tools. Then, they present data quality metrics and score calculation and present a framework for data profiling in big data. Abedjan et al. [12] and Abedjan [25] classify data profiling tasks and review the state-of-the-art about data profiling systems and techniques. Unlike our work, they do not follow a structured method to perform systematic reviews. Besides the classification of the tools, we answer different research questions, including areas of application, datasets, metadata, information extracted, and the main challenges of big data profiling.

Also, Vieira et al. [17] and Sun et al. [20] perform classification tasks, although the former uses classification to quantify the impact in the volume of data while the latter uses classification for data prediction.

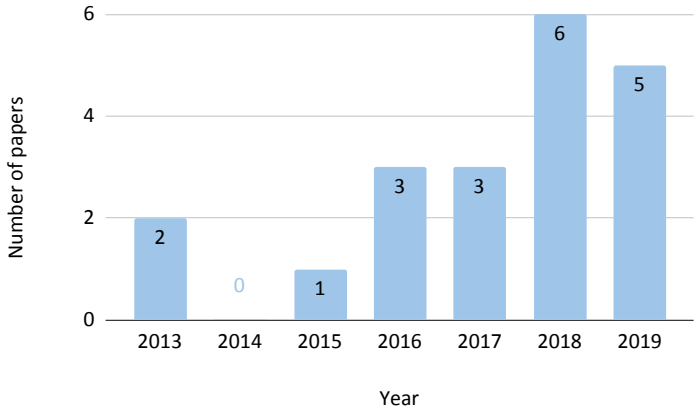

Fig. 2: **Papers per year.**

Regarding data wrangling, Sampaio et al. [27] develop a conceptual approach based on a domain-specific language for data wrangling, which is a process for data quality improvement that includes data profiling. In this case, they use data profiling to identify quality issues. In its turn, Koehler et al. [13] presents a wrangling process to use data context for data wrangling automation.

Papers [16,19,23,3,18] addresses data quality. Ardagna et al. [16] propose a data quality service to analyze big data. They present a methodology to help the user tuning the parameters to fit its intentions: budget minimization, time minimization, or confidence maximization. To do so, they create a model named CCT (Confidence/Cost/Time) that captures the interrelationships between non-functional requirements. Taleb et al. [19] propose a Big Data Quality Profiling Model that involves several modules such as sampling, profiling, exploratory quality profiling, quality profile repository, and data quality profile. Jang et al. [18] propose a data profiling model using statistical analysis techniques to derive attributes for big data quality diagnosis. Chrimes and Zamani [23] establish an interactive big data analytics platform with simulated patient data. They used open-source software technologies that and they built a platform based on HDFS (Hadoop Distributed File System) and HBase (a key-value NoSQL database). Furthermore, Juddoo [3] also presents a literature review, focusing on an overview of data quality challenges in the context of Big Data.

Some papers also report outcomes related to data lakes [11,9,24]. Alserafi et al. [11] present a framework for data governance in data lakes. They propose techniques for the automated analytical discovery of cross-data lake content relationships (information profiles). Thus, they perform metadata annotation, extraction, management, and exploitation to identify duplicate datasets, relations among datasets, and outlier datasets. Further, Maccioni and Torlone [9] proposes Kayak to expedite data preparation in a data lake. Kayak is a data

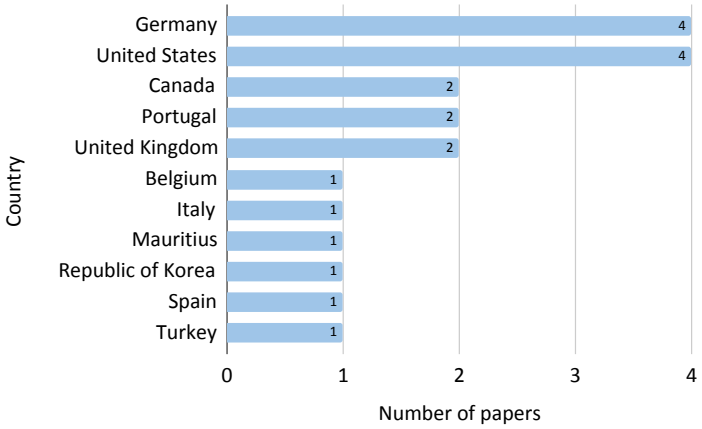

Fig. 3: **Papers per country.**

management framework that implements Adhoc primitives and executes them with an efficient strategy. Khalid and Zimányi [24] developed goal-based and rule-based agents to generate metadata profiles in a data lake. The rule-based agent operates on rules to categorize metadata files, and the goal-based agents work on goals to differentiate the types of metadata and add sections to the metadata profiles.

Among the papers that present varied approaches, Liu et al. [14] present an integrated method to address the heterogeneity issue in modeling big time-series data, while Heise et al. [22] presented a strategy to find all the unique and non-unique combinations of columns in a dataset; Shaabani and Meinel [10] propose a system for inclusion dependency injection discovery; Canbek et al. [15] created a profiling approach to gain insights about a group of datasets in different dimensions, using four data profiling techniques: basic profiling, timeline profiling, duplicate samples profiling, and Density/sparsity profiling. Finally, Santos et al. [21] reports work on Big Data Warehousing (BDW) and claims that data profiling is a part of BDW entities' resolution, which is a component that addresses the integration of data and business processes in a BDW.

Now we will present the answers to the research questions, using the analysis we performed on the papers we have just briefly described.

3.1 RQ1: What are the most used tools for big data profiling?

We found 15 papers that mention the tool they use to generate data profiling in big data. The paper that presents most tools are [26] (nine tools). We also found six papers that present tools developed by the authors: [14,22,11,10,24,9]. Among the most cited tools, we found R and Python programming languages, and Talend, briefly described below.

- R programming language[2]: reported by 4 papers. ([18,15,19,27]). R is a free software environment, mainly composed of an interpreter for the R programming language and often used with RStudio IDE[3]. R is mostly used for statistical computing and graphics generation.
- Python programming language[4]: presented in 3 papers ([15,19,24]). According to [29], Python is the language most people want to work with and is among the most loved by the developers. Python is also used for statistical computing and often for developing machine learning models.
- Talend[5]: cited by 3 papers ([17,27,26]). Talend is an open-source tool for data integration that provides services for data collection, government, transformation, and sharing.

Other tools are also mentioned once in the papers. Some have a free version, and others are commercial. Among the tools that presents at least one free version, we found DataCleaner tool, Aggregate Profiler Tool, and Talend Open Studio for Data quality [26], Hadoop, Kafka, and Zookeeper [22], OpenRefine, and Apache Taverna [27], Apache Spark [19], MongoDB [15], and HBase [23].

As for commercial tools, papers reported using Informatic Data Profiling, SAP Information Steward, Oracle Enterprise Data Quality, Collibra Data Stewardship Manage, IBM InfosSphere Information Analyzer, and SAS DataFlux Data Management Studio [26], and Trifacta Wrangler [27].

3.2 RQ2: What are the areas of application and datasets reported to be profiled?

We identified 13 papers that describe the areas of application where they perform big data profiling. Among these, 4 presented more than one area [11,10,17,13]. We group these areas of application in seven categories as follows, and we describe the corresponding datasets:

- Automotive: taxi and Uber [20], urban traffic [27], cars, and crashes datasets [11].
- Business: papers that used stock and strikes datasets [11], a business decision-support database benchmark (TPC-H) [22], and financial data [13].
- City: data related to smart city [16], weather [18], and road safety [24], real-estate domain and the United Kingdom open government data portal [13].
- Health: includes papers that used breast-cancer and diabetes datasets [11], hospital data [23], and biological databases (H-Genome, Mb, Pdb) [10].
- Industry: data related to power plant [14], supply chain, and automotive electronics industry [17].
- Web: includes Wikipedia data, linked open data about famous people, anonymized web-log data, and open music encyclopedia data [10].

[2] Available at https://www.r-project.org Accessed in November, 2019.
[3] Available at https://rstudio.com Accessed in November, 2019.
[4] Available at https://www.python.org Accessed in November, 2019.
[5] Available at https://www.talend.com Accessed in November, 2019.

Table 6: **Number of papers per type of information extracted using data profiling in big data.**

Nº of papers	Type of information	Papers
14	statistics	[25,11,15,26,27,19,12,16,22,18,3,24,9,21]
9	dependencies	[25,12,23,26,22,3,13,9,10]
6	quality	[25,18,27,21,19,17]
5	data characteristics	[15,23,26,13,9]
3	data classification	[25,24,20]
3	data patterns	[25,12,26]
3	timeliness	[15,26,14]
2	business processes and rules	[26,21]

– Others: includes papers that used basketball, tae, and tic-tac-toe games, a dataset about flowers (Iris) [11], and Android mobile malware datasets [15].

The papers present three types of datasets: online repositories [11,18,24,10], real-world datasets generated by the researchers [14,15,16,20,27,17,13], and generated data [23,22]. Online repositories include OpenML, Kaggle, biological databases, and Wikipedia. The authors who generated their datasets used data from medical records, GPS sensors, fare collection systems that collect data, and an open government data portal, for example. Finally, [22] also uses a database benchmark to generate data.

3.3 RQ3: What type of metadata did the papers collect?

We found four studies that reported the metadata they collected. [13] presented structural properties, column name tokens, column names, data types, schema paths, and parent and leaf relationships in the schema. [24] reported the collection of column count, data types, number of rows, data domain, date of data publishing, dataset origin, labeling definitions, data previews, column descriptions, creation date, data labels, data variables, attribute counter, attribute lists, number of missing data values, version management, metadata identifier, and metadata type. [10] collected dataset name, size, number of non-empty relations (tables), number of attributes (columns), number of tuples (rows), minimum, maximum, and average number of tuples per relation, and number of unary inclusion dependencies in the dataset. [23] reported that the metadata they collected was: admin source, admin type, and encounter type (the type of attendance in a hospital).

Most papers do not report the type of metadata they collect to generate data profiling, since most use data profiling to generate metadata, unlike the above papers that used metadata to create data profiling. Thus, next, we will show that all accepted papers answer RQ4, about which information was generated from the big data profiling.

3.4 RQ4: Which information is extracted using data profiling?

Based on the selected papers' analysis, we create a classification with the eight types of information most commonly extracted using big data profiling. We cluster the papers based on the type of information informed to be extracted from the data profiling. Table 6 shows a summary, with the number of papers per type of information. Below, we detail each class and present the type of information extracted for each paper, sorted by the number of papers in descending order.

- Statistics: papers in this group reported using data profiling for presenting a numerical analysis of the data. The papers mentioned discovering data correlation and association rules [25,11,26], checking data distribution [15,25,12,24,21,26], check data cardinality, or number of distinct values [25,15,26,22,12], check the number of null values [12,3,26,16], check mean, standard deviations, minimum and maximum values, [26,16,11], missing values [18], and general data summaries [25,12,24,13,19].
- Dependencies: papers in this group presented the use of data profiling for finding relationships between different datasets, or data attributes or columns in the same dataset, such as discovering foreign keys [10,3,26], detecting functional dependencies and inclusion dependencies [25,12,9,13,23,22], and computing joinability and affinity between two datasets [9].
- Quality: this group of papers reported using data profiling to discover data issues [27,21], such as outliers [25,18], syntactic errors [17], and data quality details, such as missing data and data problems [19].
- Data characteristics: these papers reported presenting basic profiling [15,9], data structures and data creator [26], descriptive information about data sources [13], and data characteristics, character lengths, and data sources [23].
- Data classification: in this group, they perform data categorization and clusterization, creating groups according to the data profile [25,24,20].
- Data patterns: these papers reported using data profiling to find interesting data patterns and behavior [25,12,26].
- Timeliness: they present temporal data using profiling techniques: age and freshness of the dataset [15], time of creation, and time patterns [26], and the trajectory of feature values along the time [14].
- Business processes and rules: they use data profiling to understand business rules [26] and business processes [21].

3.5 RQ5: What are the challenges in big data profiling?

During our analysis, we found ten papers that report 15 challenges the authors face when performing big data profiling. We list the challenges alongside their descriptions below.

1. Complexity: data profiling is a complex operation that belongs to the data preparation process [9]. Variety and volume create challenges related to

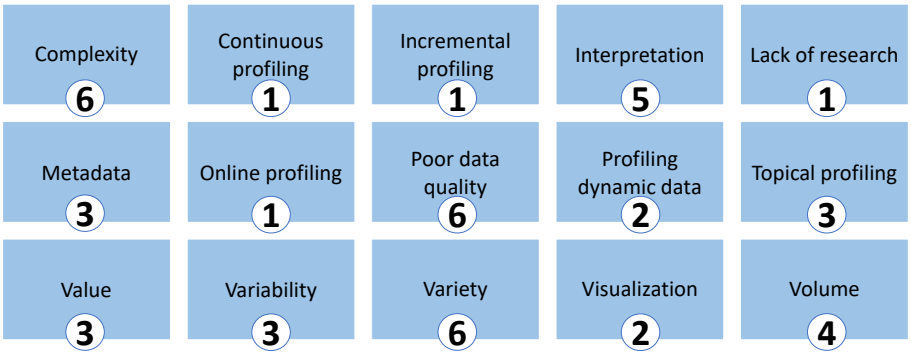

Fig. 4: **Number of papers per challenge.**

computational complexity, such as memory requirements [3]. Complexity related to the environment increases the challenges in big data profiling. [25,12,15,26].

2. Continuous profiling: automatically updating data profiling on the fly, while more data is entering the system is challenging because it requires the data profiling algorithms being always running, and it spends resources that could be used for other tasks. [3].
3. Incremental profiling: updating data profiling according to a predetermined amount of time [3].
4. Interpretation: being able to understand and interpret data profiling results [19,25,12,15,3].
5. Lack of research: the authors state there is not much research on the big data profiling research topic [19].
6. Metadata: creating metadata is the biggest challenge, according to [24], since metadata can be created manually or through data profiling. Selecting the proper metadata to generate data profiling is another challenge [12,25].
7. Online profiling: present intermediate results to the user, with a predefined confidence level [3].
8. Poor data quality: data quality impacts on data profiling results veracity [12,25,16,26,3,19].
9. Profiling dynamic data: profiling dynamic data, such as streams, is an open challenge because these types of data often change, making previous profiling obsolete [12,26].
10. Topical profiling: data profiling traditional techniques usually do not consider the whole context, such as data semantics [17], data values, structures and standards, business rules, and characteristics [18], or the use of specific datasets, such as social media [3].
11. Value: it is challenging to use data profiling for transforming the proper data in decision support so that we can generate value [16,15,3].
12. Variability: data that vary regarding size, content, and other aspects, and it requires the use of different algorithms at the same time to be able to profile different data [14,11,15]

13. Variety: profiling of heterogeneous data (audio, text, video, etc) [12,11,15,26,3,17].
14. Visualization: generate visualizations to help understanding data profiling [12,19].
15. Volume: challenges related to the size of the datasets [3,12,15,26].

Figure 4 shows the number of papers that mention each challenge. We can see that the most frequently mentioned challenges are related to complexity, poor data quality, and variety (6 papers), followed by interpretation issues (5 papers). We also performed an analysis to check which challenges are already addressed by which papers (see Table 7). To do so, we disregard the papers that are literature reviews ([25,26,3], to compare only the papers that present solutions (frameworks, algorithms, or software) for big data profiling. We also remove the challenge *lack of research* from the Table since all published papers help address this challenge. To perform this investigation, we read the papers searching for the challenges keywords. Then we performed an overall reading to check if the papers really did not address the challenges. For instance, [22] does not talk about data visualization, but they present visualizations with the profiling results, so we understand that they address the visualization challenge. Table 7 shows that [15] addresses most of the challenges, followed by [16] and [23].

4 Discussion and conclusions

In this systematic literature review we identified the new trends in big data profiling, mapping the tools, application areas, datasets, metadata, information, and related challenges. By applying the SLR process, we selected 20 papers that answer at least one of our research questions, published from 2013 to 2019. Thus, we conclude that big data profiling is a reasonably new research topic and presents growing interest from the research community. During paper analysis, we found that the R and Python programming languages are among the most used tools for big data profiling, alongside the Hadoop ecosystem and other commercial tools.

We also classified into seven groups the areas of application the papers presented: automotive, business, city, health, industry, web, and others. We also mapped the datasets reported in the papers. When we search for the metadata that the papers reported using, we found only four papers that followed the approach of using metadata to create data profiling. On the other hand, all the accepted papers presented the information they extracted using data profiling. We group the information into eight types: statistics, dependencies, quality, data characteristics, data classification, data patterns, timeliness, and business processes and rules. Most of the papers use data profiling for presenting statistics (70%), dependencies (45%), and information about data quality (30%).

Most importantly, we map 15 interesting challenges related to big data profiling, which can open new research directions related to theory and practice. The challenges are related to complexity, continuous profiling, incremental profiling, interpretation, lack of research, metadata, online profiling, poor data quality,

Table 7: **Papers versus challenges they address** (*Lack of research* was omitted on purpose).

Papers	**Challenges**	Complexity	Continuous profiling	Incremental profiling	Interpretation	Metadata	Online profiling	Poor data quality	Profiling dynamic data	Topical profiling	Value	Variability	Variety	Visualization	Volume
Canbek et al., 2018 [15]	9				✓	✓		✓		✓	✓	✓	✓	✓	✓
Ardagna et al., 2018 [16]	7							✓	✓	✓	✓	✓	✓		✓
Chrimes and Zamani, 2017 [23]	7	✓			✓	✓				✓	✓			✓	✓
Koehler et al., 2019 [13]	6	✓				✓		✓			✓		✓		✓
Sampaio et al., 2019 [27]	6	✓						✓		✓			✓	✓	✓
Vieira et al., 2020 [17]	6	✓						✓		✓			✓	✓	✓
Santos et al., 2019 [21]	5	✓									✓		✓	✓	✓
Alserafi et al., 2016 [11]	4					✓		✓				✓	✓		
Liu et al., 2013 [14]	4	✓					✓		✓	✓					
Maccioni and Torlone, 2017 [9]	4			✓		✓	✓							✓	
Taleb et al., 2019 [19]	4					✓					✓			✓	✓
Khalid and Zimányi,2019 [24]	3	✓				✓								✓	
Jang et al., 2018 [18]	2							✓			✓				
Sun et al., 2018 [20]	2									✓				✓	
Shaabani and Meinel, 2018 [10]	1					✓									
Heise et al., 2013 [22]	1													✓	

profiling dynamic data, topical profiling, value, variability, variety, visualization, and volume. The challenges that have been least explored are continuous profiling, incremental profiling, interpretation, online profiling, profiling dynamic data, and variability. In fact, we found no paper that has addressed the challenge of continuous profiling, perhaps because it is the most challenging since it requires that the data profiling algorithm is always running, requiring many resources.

Thus, we expect that our paper can also be used by researchers and in the industry by providing beginners with relevant aspects concerning big data profiling.

In future work, we want to develop a model that uses data profiling in Hadoop-based data lakes, aiming to address the challenges we pointed out in this paper.

Acknowledgments

This study was financed in part by SAP SE and by the *Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior – Brasil (CAPES)* – Finance Code 001.

References

1. Johnson, T.: Data profiling, chap. D, pp. 604–608. Springer, Boston, US (2009)
2. Caruccio, L., Deufemia, V., Naumann, F., Polese, G.: Discovering relaxed functional dependencies based on multi-attribute dominance. IEEE Transactions on Knowledge and Data Engineering pp. 1–1 (2020). DOI 10.1109/TKDE.2020.2967722
3. Juddoo, S.: Overview of data quality challenges in the context of big data. In: International Conference on Computing, Communication and Security, pp. 1–9. IEEE, Pamplemousses, MU (2015)
4. Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L.A.: Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Systematic reviews **4**, 1–9 (Jan, 2015)
5. Brereton, P., Kitchenham, B.A., Budgen, D., Turner, M., Khalil, M.: Lessons from applying the systematic literature review process within the software engineering domain. Journal of Systems and Software **80**, 571–583 (Apr, 2007)
6. McHugh, M.L.: Interrater reliability: The Kappa statistic. Biochemia medica **22**, 276–282 (Oct, 2012)
7. Kitchenham, B.: Procedures for performing systematic reviews. Keele, UK, Keele University **33**(2004), 1–26 (2004)
8. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**, 159–174 (Mar, 1977)
9. Maccioni, A., Torlone, R.: Crossing the finish line faster when paddling the data lake with KAYAK. Proceedings of the Very Large Databases Endowment **10**, 1853–1856 (Aug, 2017)
10. Shaabani, N., Meinel, C.: Improving the efficiency of inclusion dependency detection. In: International Conference on Information and Knowledge Management, pp. 207–216. ACM, Torino, IT (2018)
11. Alserafi, A., Abelló, A., Romero, O., Calders, T.: Towards information profiling: Data lake content metadata management. In: International Conference on Data Mining Workshops, pp. 178–185. IEEE, Barcelona, ES (2016)
12. Abedjan, Z., Golab, L., Naumann, F.: Data profiling. In: International Conference on Data Engineering, pp. 1432–1435. IEEE, Helsinki, FI (2016)
13. Koehler, M., Abel, E., Bogatu, A., Civili, C., Mazilu, L., Konstantinou, N., Fernandes, A., Keane, J., Libkin, L., Paton, N.W.: Incorporating data context to cost-effectively automate end-to-end data wrangling. IEEE Transactions on Big Data **X**, 1–18 (Apr, 2019)
14. Liu, B., Chen, H., Sharma, A., Jiang, G., Xiong, H.: Modeling heterogeneous time series dynamics to profile big sensor data in complex physical systems. In: International Conference on Big Data, pp. 631–638. IEEE, Santa Clara, US (2013)

15. Canbek, G., Sagiroglu, S., Taskaya Temizel, T.: New techniques in profiling big datasets for machine learning with a concise review of Android mobile malware datasets. In: International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism, pp. 117–121. IEEE, Ankara, TR (2018)
16. Ardagna, D., Cappiello, C., Samá, W., Vitali, M.: Context-aware data quality assessment for big data. Future Generation Computer Systems **89**, 548 – 562 (Dec, 2018)
17. Vieira, A.A., Dias, L.M., Santos, M.Y., Pereira, G.A., Oliveira, J.A.: On the use of simulation as a big data semantic validator for supply chain management. Simulation Modelling Practice and Theory **98**, 1–13 (Jan, 2020)
18. Jang, W.J., Kim, J.Y., Lim, B.T., Gim, G.Y.: A study on data profiling based on the statistical analysis for big data quality diagnosis. International Journal of Advanced Science and Technology **117**, 77–88 (Mar, 2018)
19. Taleb, I., Serhani, M., Dssouli, R.: Big data quality: A data quality profiling model. Lecture Notes in Computer Science **11517**, 61–77 (Jun, 2019)
20. Sun, H., Hu, S., McIntosh, S., Cao, Y.: Big data trip classification on the New York City taxi and Uber sensor network. Journal of Internet Technology **19**, 591–598 (Feb, 2018)
21. Santos, M., Costa, C., Galvão, J., Andrade, C., Pastor, O., Marcén, A.: Enhancing big data warehousing for efficient, integrated and advanced analytics: visionary paper. Lecture Notes in Business Information Processing **350**, 215–226 (Jun, 2019)
22. Heise, A., Quiané-Ruiz, J., Abedjan, Z., Jentzsch, A., Naumann, F.: Scalable discovery of unique column combinations. Proceedings of the VLDB Endowment **7**, 301–312 (Dec, 2013)
23. Chrimes, D., Zamani, H.: Using distributed data over HBase in big data analytics platform for clinical services. Computational and Mathematical Methods in Medicine **2017**, 1–16 (Dec, 2017)
24. Khalid, H., Zimányi, E.: Using rule and goal based agents to create metadata profiles. Communications in Computer and Information Science **1064**, 365–377 (Sep, 2019)
25. Abedjan, Z.: An introduction to data profiling. In: Business Intelligence and Big Data, pp. 1–20. Springer, Cham, DE (2018)
26. Dai, W., Wardlaw, I., Cui, Y., Mehdi, K., Li, Y., Long, J.: Data profiling technology of data governance regarding big data: Review and rethinking. In: Information Technology: New Generations, pp. 439–450. Springer, Cham, DE (2016)
27. Sampaio, S., Aljubairah, M., Permana, H.A., Sampaio, P.: A conceptual approach for supporting traffic data wrangling tasks. The Computer Journal **62**, 461–480 (Mar, 2019)
28. Elsevier: Scopus. https://www.elsevier.com/solutions/scopus (2021). November, 2021
29. StackOverflow: Annual developer survey results. Retrieved from https://insights.stackoverflow.com/survey/2019 (2021). November, 2021