


Review

Modeling and Management Big Data in Databases—A Systematic Literature Review

Diana Martinez-Mosquera ^{1,*} , Rosa Navarrete ¹ and Sergio Lujan-Mora ² 

¹ Department of Informatics and Computer Science, Escuela Politécnica Nacional, 170525 Quito, Ecuador; rosa.navarrete@epn.edu.ec

² Department of Software and Computing Systems, University of Alicante, 03690 Alicante, Spain; sergio.lujan@ua.es

* Correspondence: diana.martinez@epn.edu.ec; Tel.: +593-02-2976300

Received: 25 November 2019; Accepted: 9 January 2020; Published: 15 January 2020



Abstract: The work presented in this paper is motivated by the acknowledgement that a complete and updated systematic literature review (SLR) that consolidates all the research efforts for Big Data modeling and management is missing. This study answers three research questions. The first question is how the number of published papers about Big Data modeling and management has evolved over time. The second question is whether the research is focused on semi-structured and/or unstructured data and what techniques are applied. Finally, the third question determines what trends and gaps exist according to three key concepts: the data source, the modeling and the database. As result, 36 studies, collected from the most important scientific digital libraries and covering the period between 2010 and 2019, were deemed relevant. Moreover, we present a complete bibliometric analysis in order to provide detailed information about the authors and the publication data in a single document. This SLR reveal very interesting facts. For instance, Entity Relationship and document-oriented are the most researched models at the conceptual and logical abstraction level respectively and MongoDB is the most frequent implementation at the physical. Furthermore, 2.78% studies have proposed approaches oriented to hybrid databases with a real case for structured, semi-structured and unstructured data.

Keywords: big data; management; modeling; literature review

1. Introduction

The Big Data modeling term became widespread in 2011, as is visible in Figure 1. This figure shows searches in Google Trends related to Big Data modeling, which are intensified from 2011 onwards. Searches before 2004 are not presented, since Google Trends does not store earlier data. In recent years, researchers have consolidated their efforts to study new paradigms to deal with Big Data. Thus, novel Big Data modeling and management in databases approaches have emerged, in line with the new requirements. In consequence, new techniques in the database context have evolved towards Not Only SQL (NoSQL).

The work presented in this paper is motivated by the acknowledgement that a complete systematic literature review (SLR) that consolidates all the research efforts for Big Data modeling and management in databases is missing. An SLR is the best way to collect, summarize and evaluate all scientific evidence about a topic [1]. It allows for the description of research areas shown the greatest and least interest by researchers. Considering the exposed issues, the SLR conducted in this work can contribute to solving this lack by collecting and analyzing details about the research published from 2010 to 2019. As a basis for our SLR, we adhered to the guidelines proposed by Kitchenham [1]. Moreover, this paper presents a complete bibliometric analysis and summarizes existing evidence about research

on Big Data modeling and management in databases. With the information attained from the analysis performed, we will identify trends and gaps in the published research to provide a background for new research. As result, 1376 papers were obtained from scientific libraries and 36 studies were selected as relevant. All the research efforts were mapped in order to respond the three research questions defined in this research. Our main goal is to consolidate the main works to provide an awareness of the trends and the gaps related to Big Data modeling.

The remainder of this paper is organized as follows. First is the Introduction section, containing the meaning of the different terms discussed in this study. Second, a Method section presents the process used to perform the planning, conducting and reporting of the SLR. The planning phase describes the identification of the need for a SLR study and the development of a review protocol, objectives and justification, research questions and strategy. The conducting phase presents the inclusion and selection criteria for the final corpus of selected studies.

Third, the Results section answers the research questions in three subsections. The first subsection, the Bibliometric Analysis, comprises the reporting stage, including authors' information, such as affiliation and country and relevant data about their works; for instance, publication information, number of citations, funding source, year, digital library, impact factor, ranking. The second subsection, the Systematic Literature Review, presents a mapping of the selected studies according to three key concepts in a concept matrix regarding to the dataset source, modeling and database. The third subsection, the Discussion, highlights relevant findings in the SLR study in order to identify existing trends and gaps. Finally, conclusions and future works are presented.

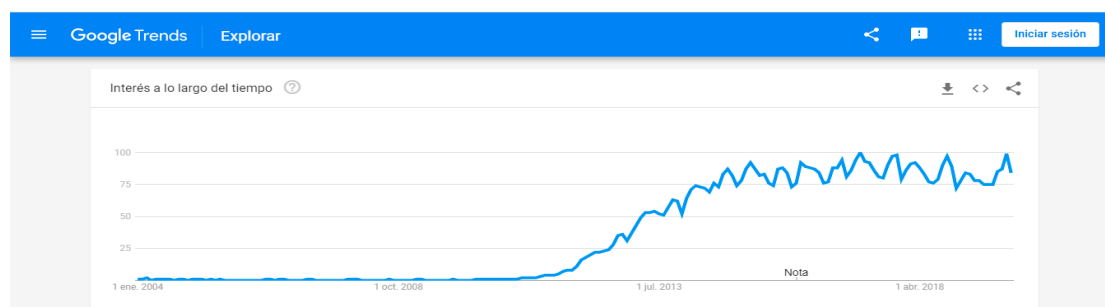


Figure 1. Trend of use of term Big Data [2].

1.1. Big Data Concepts

In this part, we describe the main concepts related to Big Data, in order to provide a general overview for the reader and a background of the terms discussed later.

1.1.1. A Brief History of Big Data

The production and processing of large volumes of data began to be of interest to researchers many years ago. By 1944, estimations for the size of libraries, which increased rapidly every year, were made in American universities [3]. In 1997, at the Institute of Electrical and Electronics Engineers (IEEE) Conference on Visualization, the term “Big Data” was used for the first time during the presentation of a study about large datasets' visualization [4].

Big Data is the buzzword of recent years, that is, a fashionable expression in information systems. The general population relates the term Big Data to its literal meaning of large volumes of data. However, Big Data is a generic term used to refer to large and complex datasets that arise from the combination of famous Big Data V's that characterize it [5].

1.1.2. Big Data Characterization

As mentioned before, Big Data does not refer only to high volumes of data to be processed. At the beginning of the Big Data studies, their volume, velocity and variety were considered as fundamental

characteristics, which were known as the three Vs of Big Data. After advances in the research, new Vs, such as value and veracity, were established. Currently, there are authors who propose up to 42 characteristics needed to consider data as Big Data, therefore, they define 42 Vs for Big Data [6]. For the purposes of our study, we will mention only ten Vs of Big Data, that are presented in a scientific study [7]. Table 1 summarizes each characteristic, along with a brief description.

Table 1. Ten Vs Big Data.

Characteristic	Brief Description
Volume	Large data sets
Velocity	High data generation rate
Variety	Different type of data formats
Variability	Consistent data
Viscosity	Data velocity variations
Virality	Data transmission rate
Veracity	Accuracy of data
Validity	Assessment of data
Visualization	Data symbolization
Value	Useful data to retrieve info

1.1.3. Volume and Velocity

To deal with the Volume and Velocity characteristics of Big Data, ecosystems and architectural solutions, such as lambda and kappa, have been created. Both architectures propose a structure of layers to process Big Data; the main difference between them is that lambda proposes a layer for batch data processing and another for streaming data, while kappa proposes a single layer for both batch and streaming processing [8]. This SLR focuses on data modeling, a concept related to the Variety characteristic, which is explained next.

1.1.4. Variety

Variety is a characteristic referring to the different types of data and the categories and management of a big data repository. As per this characteristic, Big Data has been classified into structured, semi-structured and unstructured data [9,10]. The next subsections explain in detail each data type.

Structured Data

In Big Data, structured data are represented in tabular form, in spreadsheets or relational databases [10]. To deal with this type of data, widely developed and known technologies and techniques are used. However, according to the report presented by the CISCO company, this type of data only constituted 10% of all existing data in 2014 [11]. Therefore, it is very important to analyze the 90% of the remaining data, corresponding to the semi-structured data and unstructured data that will be described below.

Semi-Structured Data

Semi-structured data are considered to be data that do not obey a formal structure, such as a relational database model. However, they present an internal organization that facilitates its processing; for instance, servers' logs in comma-separated values (csv) format, documents in eXtensible Markup Language (XML) format, JavaScript Object Notation (JSON) and Binary JSON (BSON) and so forth. Some authors may consider XML and JSON as structured [10].

Unstructured Data

Unstructured data are considered those that have either no predefined schema or no organization in their structure [12]. Within this type of data are text documents, emails, sensor data, audio files, images files, video files, data from websites, chats, electronic health records, social media data and spatio-temporal data, among others [9]. According to CISCO, the volume of unstructured data between 2017 and 2022 is expected to increase up to twelvefold [13].

To support the Variety, Volume and Velocity of Big Data, non-relational, distributed and open source data storage systems have been created. These systems include horizontal scalability, linearization, high availability and fault tolerance. Usually, these databases are known as NoSQL.

1.2. NoSQL

The existing paradigms for dealing with regular data are neither enough nor suitable to deal with Big Data requirements. For that reason, at the data storage level, the introduction of novel approaches, such as the NoSQL databases, is required. NoSQL refers to Not Only SQL, the term used for all the non-relational databases [14]. NoSQL databases are considered schema-less, as they are designed to work without structure [14]; however, in practice, there is a need for a self-sufficient model to define how data will be organized and retrieved from the database. To solve this requirement, some diverse NoSQL data models are proposed.

Data Models

A data model is a representation of the structure of the data for processing and organization [15]. A data model is considered a primary element for storage, analysis and processing in storage systems.

Currently, storage systems are classified into two large groups, relational and non-relational. Within the relational, the well-known models are the Entity–Relationship (ER), Extended Entity Relationship (EER), Key-Cube and Multidimensional, among others. The objective of this article is not to present a deep study of the models considered as classic: they are well-known and do not need to be explained. We only develop a study of the models that are a novelty for Big Data.

For non-relational systems, there are the NoSQL databases; for them, the data models are classified into four main categories [9]:

1. Column-oriented
2. Document-oriented
3. Graph
4. Key-value

Column-Oriented

In this model, data are represented in tabular form by columns and rows. The columns are identifiable by a partition key that is unique and mandatory and the rows by an optional clustering key. The primary key is the combination of the partition and clustering key. Basically, the schema of the tables consists of a set of columns, a primary key and a data type [16]. For Database Management Systems (DBMS) that use the column-oriented data model, we can mention Accumulo, Amazon SimpleDB, Cassandra, Cloudata, Druid, Elassandra, Flink, HBase, Hortonworks, HPCC, Hypertable, IBM Informix, Kudu, MonetDB, Scylla and Splice Machine, among others [17].

Document-Oriented

In this model, data are stored in key-value pairs, value documents in XML, JSON or BSON formats. Each of the documents can have nested subdocuments, indexes, fields and attributes [15]. As examples of DBMS that use the document-oriented data model, we can mention ArangoDB, Azure, BagriDB,

Cloud Datastore, CouchDB, DocumentDB, Elastic, IBM Cloudant, MongoDB, NosDB, RavenDB, RethinkDB, SequoiaDB, ToroDB and UnQlite, among others [17].

Graph

This model consists of a graph that contains nodes and edges. A node represents an entity and an edge represents the relationship between entities. There are several graph structures: Undirected/directed, Labeled graphs, Attributed graphs, Bigdata, Multigraphs, Hypergraphs and Nested graphs, among others [10]. Some examples of DBMS that use the graph data model are AllegroGraph, ArangoDB, Infinite Graph, GraphBase, HyperGraphDB, InfoGrid, Meronymy, Neo4j, Onyx Database, Titan, Trinity, Virtuoso OpenLink, Sparksee and WhiteDB [17].

Key-value

In this model, the data are represented by a key-value tuple. The key represents a unique identifier indexed to a value that represents data of arbitrary type, structure and size [10]. Secondary keys and indexes are not supported [15]. Aerospike, Azure Table Storage, BangDB, Berkeley DB, DynamoDB, GenieDB, KeyDB, Redis, Riak, Scalaris, Voldemort, among others [17] are examples of DBMS that use the key-value data model.

Table 2 summarizes the main characteristics of NoSQL data models, such as its main concept, structure, techniques to create the data model, advantages and disadvantages.

Table 2. NoSQL characteristics.

Characteristic/Data Model	Column-Oriented	Document-Oriented	Graph-Oriented	Key-Value
Concept	A model that allows representing data in columns	A model that allows representing data via structured text	A model that allows representing data and their connections	A model that allows representing the data in a simple format (key and values)
Structure	Data are stored in tables	Nesting of key-value pairs	Set of data objects (nodes)	Tuple of two strings (key and value)
		Each document identified by a unique identifier	Set of links between the objects (edges)	A key represents any entity's attribute
		Any value can be a structured document		Values can be of any data type
	Values in a column are stored consecutively	Key and value are separated by a colon ":"		
		Key-value pairs are separated by commas ","		
		Data enclosed in curly braces denotes documents		
Techniques	With compression: Lightweight encoding Bit-vector encoding Dictionary encoding Frame of reference encoding Differential encoding	Denormalized flat model	Simple direct graph Undirected multigraph Directed multigraph	NA
		Denormalized model with more structure (metadata)	Weighted graph	
	With join algorithm	Shattered, equivalent to normalization (https://pdfs.semanticscholar.org/ea15/945ce9ec0c12b92794b8ace69ce44ebe40cc.pdf)	Hypergraph	
	With late materialization		Nested graph	
	Tuple at a time			
Applications	Consumer data Inventory data	JSON documents XML documents	Social networks Supply-chain Medical records IT operations Transports	User profiles and their attributes
Advantages	High performance in loading and querying operations Efficient data compression and partitioning (both horizontally and vertically) Scalability Support for massive parallel processing Well-suited for Online Analytical Processing and OnLine Transaction Processing workloads	Support for multiple document types Support for atomicity, consistency, isolation and durability transactions Scalability Suitable for complex data, nested documents and arrays	Easy modeling Fast and simple querying Scalability	Easy design and implementation Fault tolerance Redundancy Scalability High speed
Disadvantages	Difficult to use wide-columns Delays in querying specific data	Information duplication across multiple documents Inconsistencies in complex designs	Lack of a standard declarative language Support to limited concurrency and parallelism	Very basic query language Some queries can only depend on the primary key

1.3. Data Abstraction Levels

Generally, in the design of both relational and NoSQL databases, three levels of abstraction are used: conceptual, logical and physical. Data modeling is understood as the technique that records the features of data elements in a map that describes the data used in a process. Data modeling illustrates how the data elements are organized and related [18]. Relational modeling methodologies have well established procedures, as a result of decades of research [16]; however, for NoSQL databases the modeling methodologies, specifically for Big Data, are a novel topic that continues to be studied.

Data modeling at the conceptual level is closely related to the scope of the business process. Therefore, the conceptual model is technologic-agnostic and independent of the database to be used. Thus, already-known models for relational databases can be used in non-relational databases. At the logical level, the modeling is focused on the data model to be used. For NoSQL databases, the modeling is aimed at representing the data structure of the column-oriented, document-oriented, graph or key-value models, as described previously. On the physical level, the modeling will represent the own schema of the selected database; that is, the specific implementation of the NoSQL database, such as Cassandra or MongoDB.

2. Method

This SLR study was undertaken based on the guidelines proposed by Kitchenham [1], resulting in a three-phase division: (1) planning the SLR study, (2) conducting the SLR study and (3) reporting the SLR study. Figure 2 summarizes the phases in our SLR study.

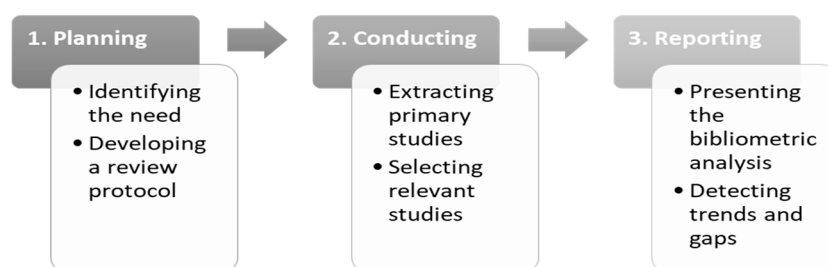


Figure 2. Systematic Literature Review phases.

The activities included in the first phase were checking the existence of other SLR studies about the topic of interest; defining the review protocol to be applied in order to mitigate possible biases from involved researchers; conceiving objectives, justifications and research questions; and, finally, defining the strategy to be pursued.

In the second phase, the activities were to collect the primary studies from the sources defined by the planning step and subsequently apply the criteria for inclusion to select only those studies related to Big Data modeling.

At last, in the third phase, the corresponding activities were to perform a full bibliometric analysis to provide the information about the authors and publication data in a single document. Furthermore, each selected relevant study was mapped to several key concepts. The analysis results allowed us to answer the research questions.

Each phase is further elaborated in the next subsections.

2.1. Planning the SLR Study

The main goals of this phase are the identification of the need for a SLR study and the development of a review protocol.

2.1.1. Identification of need for a SLR study

Following the suggestions provided by some works [1,19], we searched for SLRs and similar publications related to Big Data modeling and management, to verify if there was a gap that could be covered with the SLR proposed in this work. We found four works dating 2015 [5], 2016 [20], 2017 [15] and 2018 [10], which are described below.

Ribeiro, Silva and Rodrigues da Silva [5] completed a survey in 2015 focused on data modeling and data analytics. Although not an SLR study, the work describes some concepts that are relevant to the Big Data models. The authors identified the four main data models for Big Data—key-value, document, wide-column and graph—also described in our work. They also presented a brief summary of the abstraction levels, concepts, languages, modeling tools and database tools support. However, their study is not as detailed as ours, nor do they present a bibliometric analysis. For instance, there is a lack of information about the data models used at the conceptual, logical and physical levels, the techniques used for transforming towards the different abstraction levels, the research trends, which data set sources, types and models for Big Data are the most studied and so forth. Furthermore, our SLR is up to date on August 2019. Nevertheless and similarly to us, the study pays special attention to the fact that Big Data modeling and management in databases must be considered for research, documentation and development, as they demonstrate the data modeling necessity as a means to improve the development process in Big Data. However, they do not cover the criteria that we have mentioned before.

Sousa and Val Cura [20] cover the 2012 to 2016 timeframe. They present an SLR study about logical modeling for NoSQL databases. The authors nominate 12 articles and classify them under conceptual, logical and physical levels. They also identify modeling proposals for NoSQL databases, for NoSQL databases' migration and layers' proposals. We do not consider it as a complete work, since in our research we examined 1376 articles about Big Data modeling and management. Furthermore, they conclude that no research about data model conversion from conceptual to logical level existed at this time, even though our findings revealed the existence of several studies related to it.

Davoudian, Chen and Liu [10] present a thorough study of all the concepts and techniques used in NoSQL databases; the data models used in Big Data are described but in our work we also present a deep study on Big Data modeling methods. This is considered as a relevant work but it does not show a bibliometric analysis of all authors, conferences and journals, among other relevant information to know the trends and gaps in this topic of research. Additionally, our work focuses on examining each of the studies conducted in research to give researchers a guide to current approaches and future directions.

Wu, Sark and Zhu [15] identify some NoSQL databases and focuses to compare them according to their data model and the theorem, which indicates that a distributed system can only guarantee two of the following three properties simultaneously: Consistency, Availability and Partition tolerance (CAP) [21]. This work also does not consist of an SLR like the one presented in this work.

Furthermore, other recent surveys related to Big Data have been published; for instance, one describing the state-of-the-art about methodologies developed for multimedia Big Data analytics and the challenges, techniques, applications, strategies and future outlook [22]. Another study presents and analyzes in detail the current stage of Big Data environments and platforms and available garbage collection algorithms [23]. These works neither cover the scope of our research questions for Big Data modeling and management, nor achieve the same level of detail and precision.

In the next subsection, we detail the development of our review protocol, asserting our objectives and justification and the research questions.

2.1.2. Development of a Review Protocol

A review protocol is essential in order to mitigate any possible bias from researchers and it must be defined before conducting the SLR [1]. Thus, during this stage, we formed the applied method. First, we proposed specific development goals and the respective justification for our work. Then,

we formulated three research questions with the intent of summarizing the existing evidence about Big Data modeling and management. Finally, we elaborated a strategy to conduct this SLR study effectively.

Objectives and Justification

The first objective is to present information about the most relevant research about Big Data modeling and management in a comprehensive bibliometric analysis. This study contains a number of studies from the digital libraries and details the authors, their institutional affiliations, countries and publication details, such as the publication year and their impact factors in the Journal Citation Reports (JCR) and the Scimago Journal Rank (SJR) for journals and in the CORE Ranking for Conferences.

Based on our findings, the second objective was to detect the different approaches for Big Data modeling used in the different studies in order to determine trends and gaps within the three key concepts, source, modeling and database. The SLR study conducted in this research can focus all the research related to Big Data modeling into a single document, to benefit the industry, the academy and the community.

Research Questions

This stage comprises the most important phase of the protocol development [1]. Hence, we took particular care while following Kitchenham's suggestions. Firstly, we identified three actors within the population: (1) researchers, (2) information analysts and (3) software developers who research, document and implement solutions for Big Data modeling and management in databases. Secondly, we considered collecting all the approaches related to data modeling oriented to Big Data. Thirdly, as outcomes, we intend to summarize the findings and determine the trends and gaps in the studied topic. This study raised the following research questions:

Research Question (RQ1): *How has the number of published papers about Big Data modeling and management changed over time?*

Rationale: Our interest is to consolidate, through a bibliometric analysis, all the research efforts for the topic, providing researchers with the ability to know all the information about the authors and the publication data in a single document. Thus, the reader can know how the studies, in our topic of interest, have grown over time, who were the authors who provided significant contributions towards the subject, which are the most cited studies and which countries are most interested in this research topic, as well as which journals and conferences are involved in this topic and which scientific libraries have the major share of studies about Big Data modeling and management. In addition, we wanted to know whether these researches were mostly funded by industry or the academy.

Research Question (RQ2): *Are there any research studies that focus on approaches for semi-structured and unstructured data and what techniques to apply?*

Rationale: Our goal is to find out whether the studies are focused on semi-structured and unstructured data, which, according to the data specified in the Big Data Concepts subsection, comprised most of the available data. In addition, we intend to present what models the researchers propose at each modeling abstraction level and to determine three key concepts: source, modeling and database:

- For source: The dataset sources and data types;
- For modeling: The data abstraction levels, the data model proposed at conceptual, logical and physical levels, the techniques used for transformations between abstraction levels, the applied modeling language, the modeling methodology and the proposed tools for automatic model transformation;
- For database: The database type and the evaluation and performance comparison between models.

Research Question (RQ3): *What are the trends and gaps in Big Data modeling and management?*

Rationale: Based on the data obtained in RQ2, our main interest is to present the solutions proposed by researchers in this topic in a consolidated work. The objective is to allow researchers to focus their efforts on the gaps and solutions that allow for standardization over the currently existent or novel methods.

Strategy

The strategy to conduct an exhaustive compilation of studies on the topic of interest included four actions:

1. Finding primary studies from scientific digital libraries, mainly considering whether: (1) they contain indexed research documents, (2) there is a high frequency of databases update and (3) they publish related research about our topic of interest. The sources listed below comply with the desired requirements, in order to focus our systematic review of relevant research:

- IEEE Xplore
- ScienceDirect
- Scopus
- Web of Science (WoS)

Moreover, according to a comprehensive study, which evaluated the quality of 28 scientific search systems, Google Scholar is inappropriate as principal search system, while ScienceDirect, Scopus and WoS are suitable to evidence synthesis in an SLR [24].

2. Applying the inclusion criteria to the primary studies, in order to select those studies related to Big Data modeling and management, we conducted a search of a specific terms-matching process within the articles' titles, abstracts and keywords. Based on our research questions, two major search terms were derived: big data and model. The terms were selected after combining several options, in order to get a significant number of articles and these terms covered the majority of studies that addressed our research subject.

Due to the fact that the selected digital libraries do not share a common search syntax, we enumerated all the search strings applied in each one. The word "model" has been used because some studies use this term when referring to modeling:

- IEEE Xplore—(((“Document Title”:“big data” and “data model”) OR “Abstract”:“big data” and “data model”) OR “Index Terms”:“big data” and “data model”)
- ScienceDirect—Title, abstract, keywords: “big data” and “data model”
- Scopus—TITLE-ABS-KEY (“big data” AND “data model”)
- WoS—TS = (“big data” AND “data model”). TS regards to Topic fields that include titles, abstracts and keywords.

For the inclusion criteria, only studies written in English and published in conferences or journals were considered. Although no date-limiting factor was defined in our search criteria, it was observed that no results prior to 2010 were returned by any selected scientific library. These results match with Figure 1, where a report from Google Trends demonstrates that the term Big Data started to become popular in 2011;

3. Reviewing the studies for a second time through a reading of the papers' content allowed us to discard the ones not relevant to the context of our topic of interest;

4. The snowballing technique was applied to locate additional relevant articles according to existing references from within the already-reviewed studies.

2.2. Conducting the SLR Study

The search of the SLR study was conducted in August 2019. Figure 3 provides a representation of the selection process applied to the studies. Duplicated studies were discarded from the potentially relevant studies stage.

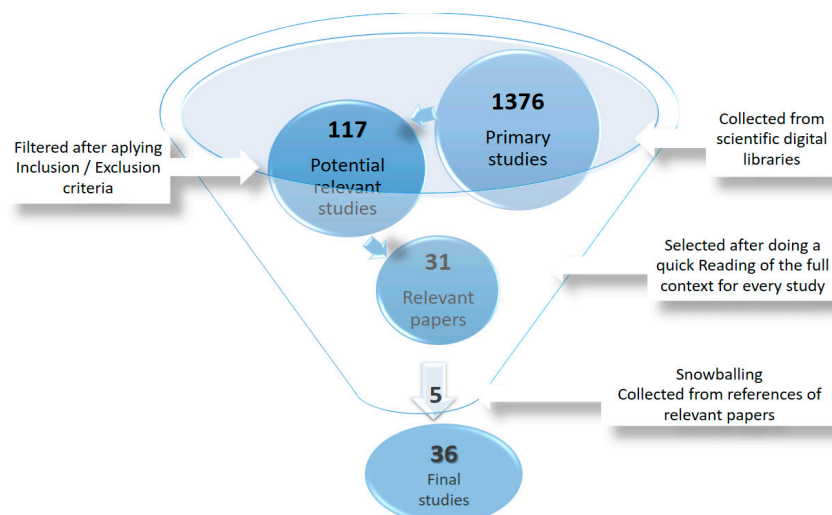


Figure 3. Selection Process.

2.2.1. Inclusion Stage

In this stage, we selected studies related to Big Data modeling and checked their titles, abstracts and keywords according to the previously planned strategy. We considered only English studies that addressed our research questions and published in conferences or journals. Our main objective is to identify the different approaches to data modeling and management in data stores in a general way for the different types of data at the three abstraction levels. As a result, 1 Chinese article and 27 articles corresponding to books, book chapters, letters, notes or editorials, were discarded. Additionally, we also discarded 1259 articles that, although mentioning data models, referred to specific applications or not related to data persistence but to data ingestion, data lakes or data analytics. From this stage, 117 studies were accepted.

2.2.2. Selection Stage

At this stage, a quick review of the full content of every study allowed us to select only those studies related to Big Data modeling. This resulted in the acceptance of 31 studies, the rejection of 70 papers and the filtering of 16 duplicated works. At this phase, we eliminated the duplicated papers.

After scanning the whole content of these selected 31 studies, we also included five new papers after the snowballing review. Finally, 36 studies made up our final corpus to report the SLR study.

2.3. Reporting the SLR Study

The objective of this step is to answer the research questions raised in the review protocol. For this purpose, this study is divided into three parts. In the first part, we perform a bibliometric analysis to answer RQ1; in the second part, we present the literary review with the most relevant data of the approaches in a concept matrix to answer RQ2; and in the third part, we discuss the trends and gaps to answer RQ3. Section 3 presents in detail the results collected from the activities described in this phase.

3. Results

In this section, we answer the research questions via the below activities:

1. A bibliometric analysis, to gather information about the authors and the publication data, the authors and countries with more contributions in the subject, the impact of the selected studies and how the research has grown throughout time, as well as the journals and conferences proceedings where the studies were published;
2. A literature review to map the studies according to three key concepts—source, modeling and database—in a concept matrix. In the source concept, we analyze the dataset sources and data types.

In the modeling concept, we analyze the data abstraction levels, the data models proposed at the conceptual, logical and physical levels, the techniques used to perform transformations between abstraction levels, the applied modeling language, the modeling methodology and the proposed modeling tools. At the database concept, we analyze the type and conduct an evaluation and performance comparison between models;

3. A discussion to identify trends and gaps in Big Data modeling and management.

3.1. Bibliometric Analysis

The objective of this analysis is to answer RQ1. To answer the first part of the rationale of this question, we analyze the results of the inclusion and selection stages. In Figure 4, we summarize the results of the inclusion stage and highlight some findings:

- The average annual growth rate of published articles follows Equation (1)

$$y = 30.309x - 29.2 \quad (1)$$

- Prior to 2010, no relevant studies about Big Data modeling are published
- Since 2015, the number of studies has increased significantly and, in 2018, there were 318 published articles. In 2019, there were already 106 publications before August
- Scopus ranked the highest of all considered sources, with 760 collected works, followed by WoS with 321 works, IEEE Xplore with 200 and ScienceDirect with 95

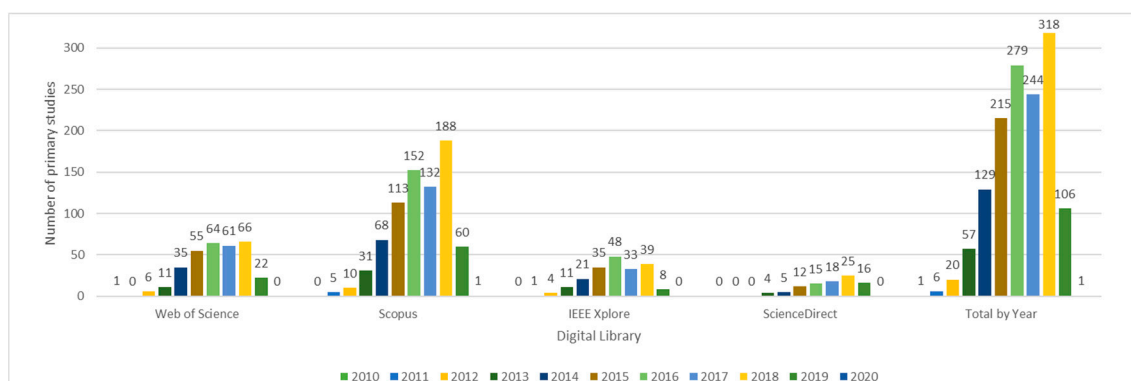


Figure 4. Number of primary studies by year and source.

The results of the selection stage are presented in Figure 5, organized by source and year. We can highlight the following findings:

1. Prior to 2013, no relevant studies were found;
2. The year in which we found the most quantity of studies about Big Data modeling is 2018. However, it is important to highlight that 2019 is ongoing and could ultimately have more studies than 2018;
3. With 27 papers, Scopus is the source holding the highest number of relevant studies, followed by WoS and IEEE Xplore with two papers each. ScienceDirect does not report any relevant paper about the topic.

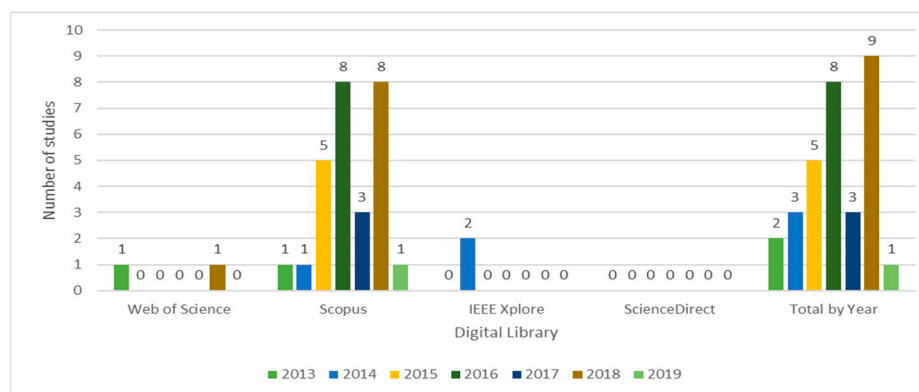


Figure 5. Results of selection stage.

Table 3 summarizes the main data of each of the selected articles, among them, the reference, the first author name and affiliation, the country where the research was done, the identification of the journal or conference, the digital library, the publication year, the number of citations in Scopus, the knowledge application and the existence of funding.

Table 3. Bibliometric Analysis.

Reference	First Author's Name	First Author's Affiliation	Country	Journal/Conference ID	Digital Library	Publication Year	Citations in Scopus	Knowledge Application	Funding
[25]	Jie Song	Software College, Northeastern	China	J1	Scopus	2019	0	Academy	Yes
[26]	Laurent Thiry	University of Haute Alsace	France	J2	Scopus	2018	0	Academy	NA
[27]	Victor Martins de Sousa	UNIFACCAMP	Brazil	C1	Scopus	2018	1	Academy	Yes
[28]	Igor Zečević	University of Novi Sad	Serbia	J3	Scopus	2018	2	Academy	Yes
[29]	Antonio M. Rinaldi	University of Naples Federico II	Italy	C2	Scopus	2018	1	Academy	NA
[30]	Shady Hamouda	Emirates College of Technology	United Arab Emirates	C3	Scopus	2018	1	Academy	NA
[31]	Dippy Aggarwal	University of Cincinnati	United States of America	J4	Scopus	2018	0	Academy	NA
[32]	Alfonso de la Vega	University of Cantabria	Spain	C4	Scopus	2018	0	Academy	Yes
[33]	Xu Chen	North Minzu University	China	J5	Scopus	2018	0	Academy	NA
[18]	Maribel Yasmína Santos	University of Minho	Portugal	J6	Scopus	2017	10	Academy	Yes
[34]	Kwangchu Shin	Kook Min University	South Korea	J7	Scopus	2017	7	Academy	Yes
[35]	Fatma Abdelhedi	Toulouse Capitole University	France	C5	Scopus	2017	1	Academy	NA
[36]	Aravind Mohan	Wayne State University	United States of America	C6	Scopus	2016	7	Academy	Yes

Table 3. Cont.

Reference	First Author's Name	First Author's Affiliation	Country	Journal/Conference ID	Digital Library	Publication Year	Citations in Scopus	Knowledge Application	Funding
[37]	Massimo Villari	University of Messina	Italy	C7	Scopus	2016	2	Academy	NA
[38]	Maribel Yasmina Santos	University of Minho	Portugal	C8	Scopus	2016	10	Academy	Yes
[39]	Maribel Yasmina Santos	University of Minho	Portugal	J8	Scopus	2016	8	Academy	Yes
[40]	Ganesh B. Solanke	PCCoE, Nigdi	India	C9	Scopus	2018	0	Academy	NA
[41]	Vincent Reniers	KU Leuven	Belgium	C10	Scopus	2018	0	Academy	Yes
[42]	Fatma Abdelhedi	Toulouse Capitole University	France	C11	Scopus	2016	4	Academy	NA
[43]	Max Chevalier	University of Toulouse	France	C12	Scopus	2016	6	Academy	ANRT
[44]	Shreya Banerjee	National Institute of Technology	India	C13	Scopus	2015	6	Academy	NA
[16]	Artem Chebotko	DataStax Inc.	United States of America	C14	Scopus	2015	43	Industry	Yes
[45]	Wenduo Feng	Guangxi University	China	C15	Scopus	2015	2	Academy	Yes
[46]	Ling Chen	Zhejiang University	China	C16	Scopus	2015	1	Academy	Yes
[47]	Max Chevalier	University of Toulouse	France	C17	Scopus	2015	14	Academy	NA
[48]	Dewi W. Wardani	Sebelas Maret University	Indonesia	C18	Scopus	2014	7	Academy	NA
[49]	Ming Zhe	Hubei University of Technology	China	C19	Scopus	2013	0	Academy	NA
[50]	Mohamed Nadjib Mami	University of Bonn	Germany	C20	Scopus	2016	5	Academy	Yes
[51]	Dan Han	University of Alberta	Canada	C21	WoS	2013	0	Academy	Yes
[52]	Zhiyun Zheng	Zhengzhou University	China	C22	IEEE	2014	1	Academy	Yes
[53]	Dongqi Wei	University of Geosciences	China	C23	IEEE	2014	3	Academy	NA
[14]	Karamjit Kaur	Thapar University	India	C24	IEEE	2013	59	Academy	NA
[54]	Michael J. Mior	University of Waterloo	Canada	J1	IEEE	2017	12	Academy	Yes
[55]	Max Chevalier	University of Toulouse	France	C25	Scopus	2016	4	Academy	Yes
[56]	Harley Vera	University of Brasilia	Brazil	C26	Scopus	2015	8	Academy	NA
[57]	Robert T. Mason	Regis University	United States of America	C27	NA	2015	0	Academy	NA

3.1.1. Authors

In Figure 6, it is possible to verify the names of the first authors who have made major contributions to the subject. Thus, Maribel Yasmina Santos from Portugal and Max Chevalier from France occupy first place with three articles each and Fatma Abdelhedi from France is in second place with two articles. According to the observed data, two of Santos' studies were published in 2016 and another in 2017. Their research was performed in collaboration with University of Minho and it is the only one from Portugal presented in the final corpus of studies.

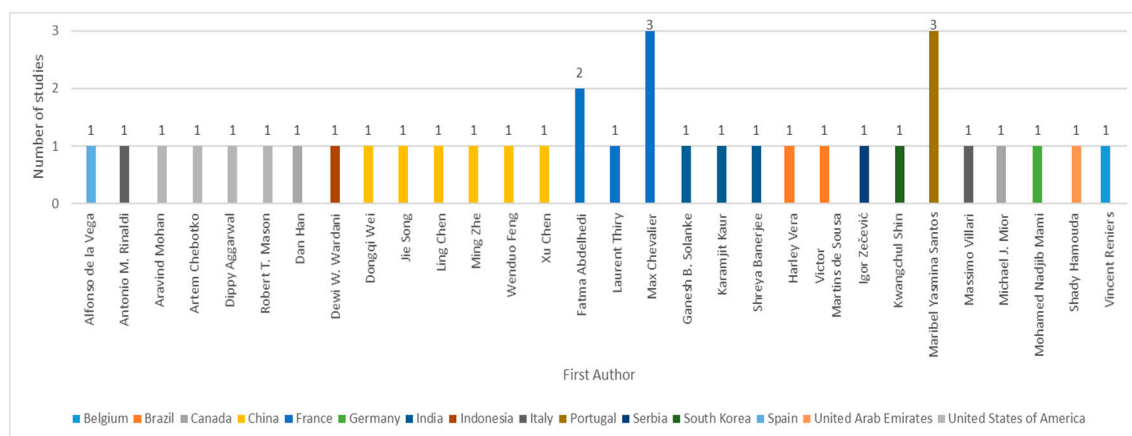


Figure 6. Contribution by author.

Regarding France, the authors Chevalier, Abdelhedi and Laurent Thiry investigated the topic and added six contributions in total, one published in 2015, 3 in 2016 and one in 2017 and 2018. Their research is linked to University of Haute Alsace and to University of Toulouse.

Countries such as the USA, China and India have made several contributions from different authors. For the USA there are four articles, in 2015 Robert Mason of the Regis University and Artem Chebotko from DataStax Inc. presented an article each, in 2016 Aravind Mohan from Wayne State University and in 2018, Dippy Aggarwal of the University of Cincinnati, also presented their approaches. In China, the six authors and institutions that have made contributions were, in 2013 Ming Zhe Hubei of the University of Technology, in 2014 Dongqi Wei from the University of Geosciences, in 2015 Ling Chen from Zhejiang University and Wenduo Feng of the Guangxi University, on 2018 Xu Chen University of the North Minzu and in 2019, Jie Song from the Software College, Northeastern. There are three studies from India published by Karamjit Kaur from Thapar University in 2013, Shreya Banerjee of the National Institute of Technology in 2015 and Ganesh Soanke from PCCoE, Nigdi in 2018.

In total, 15 different institutions, one from the industry and 14 from the academy have presented relevant works and it can be observed that even in 2018 and 2019 the subject is still being actively investigated.

3.1.2. Countries and Years

In this part, after discarding 16 duplicated results, 101 of the 117 articles collected after the inclusion stage were taken as sample. These articles contain research pertaining to our area of interest and allow us to analyze a greater number of articles.

Figure 7 shows that the leading countries in the topic of interest are the USA and China, with 17 and 16 articles, respectively. The country where Chevalier performed his research, France, takes third place with eight articles and, with seven publications each, Italy and Spain take the fourth spot. Finally, Germany takes fifth place with six studies.

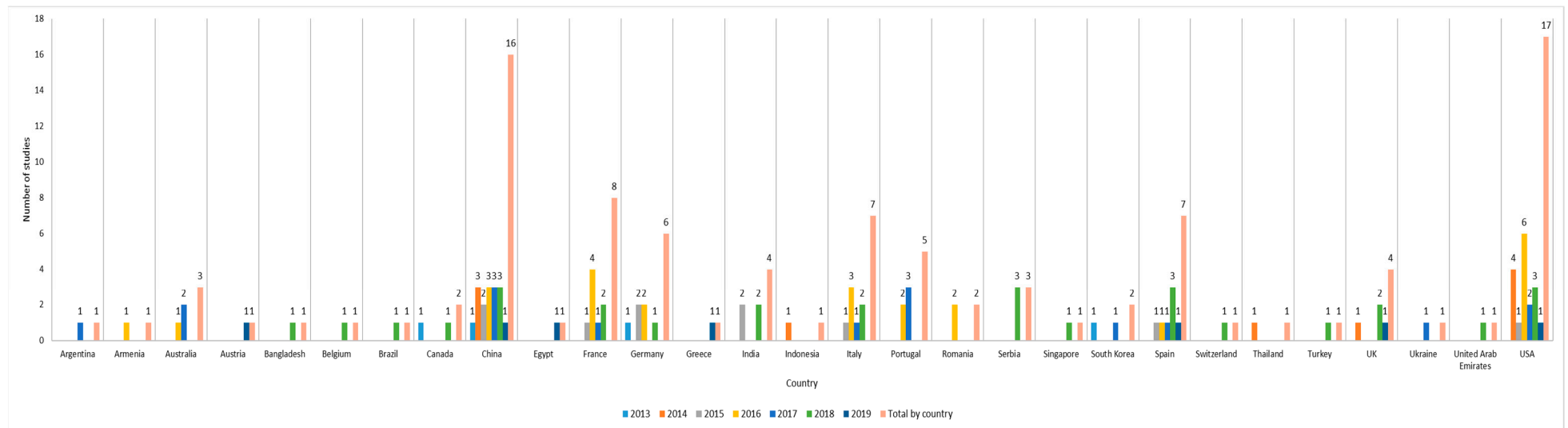


Figure 7. Contribution by year and country.

For the USA, four articles were published in 2014, one article in 2015, six articles in 2016, two articles in 2017, three articles in 2018 and one article in 2019. Therefore, the research in that country started in 2014, had the most contributions in 2016 and continues through 2019. Regarding China, their first article was published in 2013, followed by three articles in 2014, two articles in 2015, during 2016, 2017 and 2018 three articles in every year and one article in 2019. It can be seen that China initiated its research in 2013 and still continues to investigate the topic. It is also worth mentioning the constant article publications observed between 2014 and 2018. France started in 2015 with one article, four articles in 2016, one article in 2017 and two articles in 2018. This country started the research in 2015 and 2016 was the year with more contributions. Italy and Spain also started the research in 2015. Italy presented more articles during 2016 and Spain in 2018. Regarding 2019, only Spain has published one article. In 2013, Germany started the research with one article and its last published article was found in 2018.

As conclusion, from 2015 onwards, more countries start contributing to the scientific production on this topic, doubling the number of published articles in 2016. In 2018 and 2019, the trend remains. However, the year 2019 is still ongoing; therefore, it is likely that many studies will be published before the end of the year.

3.1.3. Citations

Table 3 presents the number of citations of the studies in Scopus. Figure 8 presents the article with the greatest impact, which has 59 citations and was published by Karamjit Kaur from India, followed by one by Artem Chebotko from the USA, with 43 citations. It is important to highlight that both authors also belong to the countries with more contributions.

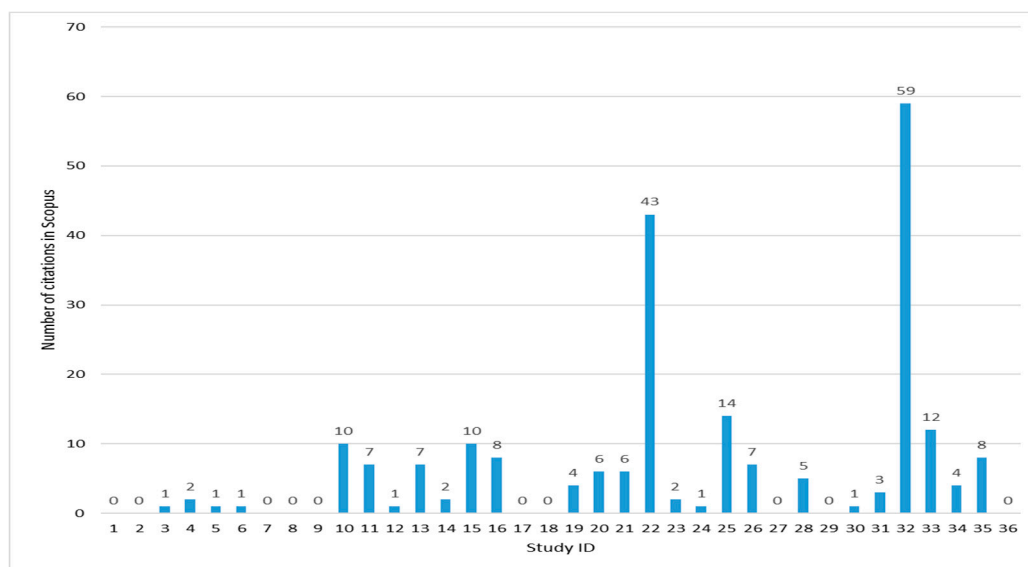


Figure 8. Number of citations in Scopus.

The most cited article has “Modeling and querying data in NoSQL databases” as a title and was published in 2013. The second most cited article is titled “A Big Data Modeling Methodology for Apache Cassandra” and was published in 2015. Further details about these publications are presented in the SLR section.

It can also be noted, according to Table 3, that 97.22% of the articles belong to the academy and that 52.78% of the articles were funded. According to our criteria, this topic is considered of high relevance because funds are allocated in projects for research.

Tables 4 and 5 provide information to the reader about the journals and conferences where the studies are published; their impact factor is also presented in the JCR and SJR, and, for the conferences,

their ranking. It is important to highlight that 75% of the studies were presented in conferences, thus we can anticipate that for the current year there are studies still under progress, that have not reached their final stage.

Table 4. Information of journals where the relevant studies were presented.

Journal ID	Journal Name	Country	JCR IF	SJR	Study ID
J1	IEEE Transaction on Knowledge and Data Engineering	United States of America	3.86	1.1	1, 33
J2	Journal of Big Data	United Kingdom	NA	1.1	2
J3	Enterprise Information Systems	United Kingdom	2.12	0.7	4
J4	Advances in Intelligent Systems and Computing	Germany	NA	0.2	7
J5	Filomat	Serbia	0.79	0.4	9
J6	Journal of Management Analytics	United Kingdom	NA	NA	10
J7	International Journal of Applied Engineering Research	India	NA	0.1	11
J8	Lecture Notes in Computer Science	Germany	0.4	0.3	16

Table 5. Information of conferences where the relevant studies were presented.

Conference ID	Conference Name	CORE 2018 Ranking	Study ID
C1	20th International Conference on Information Integration and Web-Based Applications and Services	C	3
C2	10th International Conference on Management of Digital EcoSystems	Not ranked	5
C3	2017 International Conference on Big Data Innovations and Applications	Not ranked	6
C4	8th International Conference on Model and Data Engineering	Not ranked	8
C5	19th International Conference on Enterprise Information Systems	C	12
C6	5th IEEE International Congress on Big Data	Not ranked	13
C7	2016 IEEE Symposium on Computers and Communication	B	14
C8	9th International C* Conference on Computer Science and Software Engineering	Not ranked	15
C9	2017 International Conference on Computing, Communication, Control and Automation	Not ranked	17
C10	2017 IEEE International Conference on Big Data	Not ranked	18
C11	8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management	C	19
C12	18th International Conference on Enterprise Information Systems	C	20
C13	2015 IEEE International Conference on Industrial Informatics	Not ranked	21
C14	4th IEEE International Congress on Big Data	Not ranked	22

Table 5. Cont.

Conference ID	Conference Name	CORE 2018 Ranking	Study ID
C15	2015 IEEE International Conference on Smart City/SocialCom/SustainCom together with DataCom	Not ranked	23
C16	2015 IEEE International Conference on Multimedia Big Data	Not ranked	24
C17	17th International Conference on Big Data Analytics and Knowledge Discovery	Not ranked	25
C18	2014 International Conference on Computer, Control, Informatics and Its Applications	Not ranked	26
C19	2013 International Conference on Computer Sciences and Applications	Not ranked	27
C20	18th International Conference on Big Data Analytics and Knowledge Discovery	Not ranked	28
C21	2013 IEEE Sixth International Conference on Cloud Computing	B	29
C22	3rd IEEE International Congress on Big Data	Not ranked	30
C23	2014 Fifth International Conference on Computing for Geospatial Research and Application	Not ranked	31
C24	2013 IEEE International Conference on Big Data	Not ranked	32
C25	IEEE Tenth International Conference on Research Challenges in Information Science	B	34
C26	2nd Annual International Symposium on Information Management and Big Data	B	35
C27	Informing Science & IT Education Conference	C	36

3.1.4. Journals

We present in Table 4 the list of journals where the selected relevant studies were published. The table contains the assigned journal identifier, the journal name, the journal's country, the impact factor (IF) in the JCR and SJR and the related study ID. We considered it important to display the JCR IF and the SJR, since they are indicators related to the quality of the research according to the number of citations of the published studies and their importance in the scientific research.

3.1.5. Conferences

We present in Table 5 the details of the conferences where some relevant studies were presented. The assigned conference identifier, the conference name, the core ranking and the respective studies identifiers are listed. We used the conference ranking Computing Research and Education Association of Australasia (CORE), 2018. This ranking was created by an association of computer science departments from universities in Australia and New Zealand. This Association provides conference rankings in the computing disciplines based on a mix of indicators, including citation rates, paper submission and acceptance rates. The rankings range are represented by the letters A*, A, B and C—A* being the best and C the worst.

Through the performed analysis, research question RQ1 is answered in significant detail. In order to answer the next two research questions, each of the selected articles deemed as relevant were analyzed, after a full reading of each of them.

3.2. Systematic Literature Review

The objective of this section is to answer the second research question, RQ2. To comply with this goal, we rely on the concept matrix [58] compiled in Appendix A. There, we synthesize the literature about each one of the 36 articles that comprise the final research corpus. Next, each of the key concepts that we have covered in this SLR will be described. Mainly, three domains are analyzed:

1. Source
2. Modeling
3. Database

3.2.1. Source

At this section, we analyze the dataset sources and data types. The dataset sources enable us to know whether the research was carried out in a real-world environment or in a test environment with simulated data. The use of real-world datasets is important to verify compliance with the volume, velocity, veracity and value that characterizes Big Data. As mentioned in Section 1.1.4, according to a study [11], 90% of the existent data in the world corresponds to semi-structured and unstructured data. For this reason, this concept allows us to validate if the research is oriented to these types of data.

Data Set Sources

After analyzing the 36 selected articles, it was determined that 22 articles used sample datasets for their proposals, 10 articles used real-world datasets and four did not present any example of their solutions—for this reason they do not mention any type of dataset. Therefore, it was concluded that more than 50% of the relevant studies did not present their verified proposals with real-world datasets.

By not using real-world datasets, the behavior of the solutions in a production environment cannot be verified. The main real-world datasets used in the studies were sensor data, image metadata, websites publications and electronic documents, as Figure 9 presents. As we can see, those datasets are categorized as unstructured data that we analyze in the next concept. In addition, batch processing is used by most of these approaches, while real-time processing is proposed by one study about data modeling for commercial flights in the USA [18].

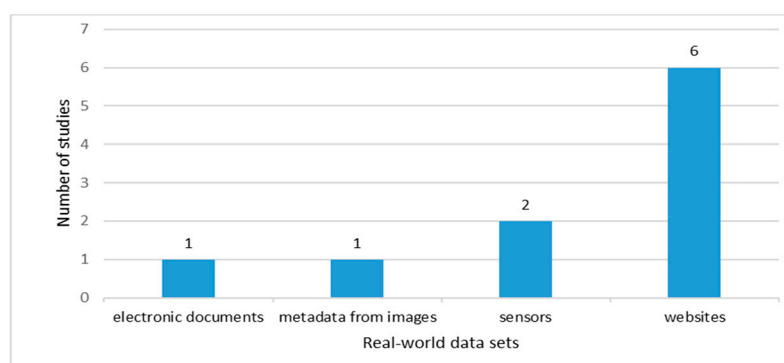


Figure 9. Types of used real-world data sets.

In Table 6, we summarize the ten studies with real-world data sets presented on Figure 9, in order to know under which application the studies were elaborated and whether they comply with the volume, velocity, veracity and value characteristics. The variety characteristic is analyzed in the next subsection Data Types. From Table 6, we can see that 90% of the studies do not justify the velocity characteristic.

Table 6. Analysis of real-world data sets used in the relevant studies.

Real-World Data Sets	Study ID	Domain	Volume	Velocity	Veracity	Value
Electronic documents	27	e-government electronic documents	A large number of e-government electronic documents	Not available	Documents are laws and regulations	Managing housing transfer process
Images metadata	5	Images from a web server	Network with 8000 relationships and 5990 nodes	Not available	Knowledge base of famous painters	Obtaining images related to the famous painters
Sensors	9	Fuzzy spatio-temporal data	Not available	Not available	Data from the real movement of the tropical cyclones Saomei and Bopha under the influence of subtropical high	Analyzing meteorological phenomena
	13	Vehicles into OpenXC	14 exabytes per year	Up to 5GB/hour	Data from devices that are installed in the vehicles	Providing insights on the risk level based on the drivers driving behavior
Websites	4	Web-based agriculture application	Not available	Not available	Data from a Precision Agriculture Based Information Service (PAIS) application	Providing an online service to the farmers, with 24/7 access to the images of the crops
	10	Commercial flights in the USA	More than 123 million records	Not available	Domestic flights in the USA obtained from RITA-BTS	Presenting the behavior of the companies regarding to the accomplished and cancelled flights
	18	Review site Epinions	Not available	Not available	Data from online consumers	Identifying the user preferences
	30	Microblog SINA	1.75 GB	Not available	Data from 1500 user profiles and their microblogs	Not available
	32	Slashdot	Not available	Not available	User posts	Finding useful information about the user posts
	33	EasyAntiCheat	Large volumes from real-time data of players behavior	Not available	Partial workload extracted from multiplayer games	Determine patterns for cheating detection

Data Types

In this respect, 18 studies from the relevant articles present solutions for unstructured data and 12 articles for semi-structured data. According to those data, it is possible to verify that the research about the modeling of unstructured and semi-structured data follows current trends 83.33% of the time.

Another interesting fact is that there are also studies that propose modeling approaches for structured data. This is because, for the data to be considered as Big Data, they must also comply with the variety characteristic. These structured data are analyzed in eight studies and come from relational databases.

3.2.2. Modeling

In this section, from the final corpus selected, we analyzed the proposed data abstraction levels, the models presented at the conceptual, logical and physical levels, the proposed approaches for transformation between abstraction levels, the modeling language, methodology and tools.

Data Abstraction Levels

For this concept, we intend to determine what levels of data abstraction have been covered by the studies for data modeling solutions. As mentioned in the Big Data Concepts subsection, there are three levels used for relational databases that are also used in NoSQL stores: conceptual, logical and physical. According to Figure 10, which summarizes the data obtained in Appendix A from the 36 studies, 25 present approaches for data modeling at the three levels of abstraction, therefore, those studies can be considered as complete works that reached the physical implementation of their proposals in a NoSQL storage. The other studies only cover one or two levels although it is possible that, in the future, their works will demonstrate their approaches at all three levels.

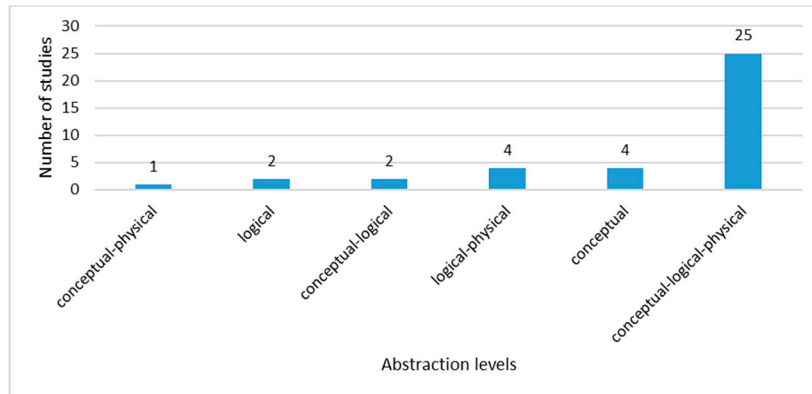


Figure 10. Data Abstraction Levels for modeling Big Data.

The next concept presents the data model proposed by authors for each data abstraction level.

Data Model at Conceptual Level

This concept comprises the models presented in each study from the final corpus at the conceptual abstraction level. As we mentioned before, this level is technology-agnostic and there is no restriction regarding the use of well-known models applied to relational databases.

Figure 11 presents 19 works using the ER model at the conceptual abstraction level. Within these 19 works, one proposes the use of Extended Binary Entity Relationship (EBER) [27], an ER-based model that adopts different types of attributes and a dominant role. Another study from the 19 works, proposes Enriched Entity Relationship (EER) [37] with graphic notation for the representation of Big Data.

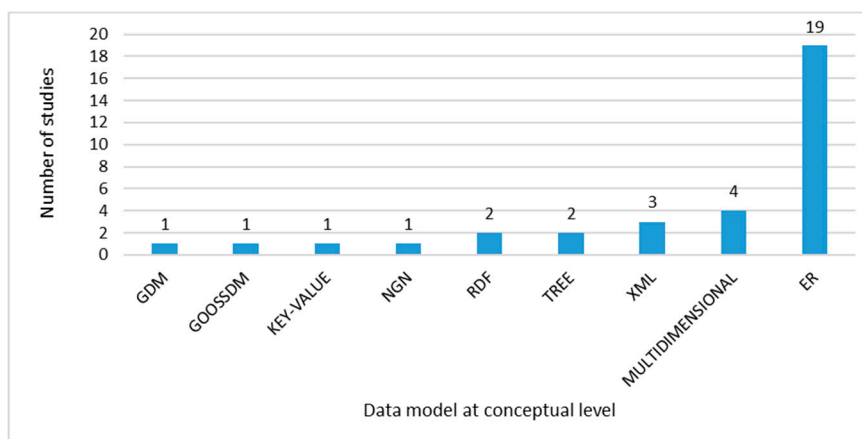


Figure 11. Data Model at conceptual abstraction level.

Furthermore, the use of the multidimensional data model is observed in four studies. It is assumed that this is derived from the increasing interest in DataWarehouses and DataMarts for Online Analytical Processing (OLAP), where the usage of ad-hoc queries is common. In addition, three papers propose the use of the XML model, which corresponds to an abstract representation of XML fragments. The other eight remaining works propose independent models, such as the Generic Data Model (GDM), the Graph Object Oriented Semi-Structured Data Model (GOOSSDM), Key-value, Novel Graphical Notation (NGN), Resource Description Framework (RDF), Tree and there are two works that do not propose any model.

Data Model at Logical Level

At the level of logical abstraction, according to the data obtained in Figure 12, the trend model is document-oriented with 12 studies, followed by graph-oriented and column-oriented, with seven studies each. As detailed in the Big Data Concepts subsection, there are four widely used models in NoSQL key-value: column-oriented, document-oriented and graph; however, key-value has been studied at this level of abstraction in just one proposal.

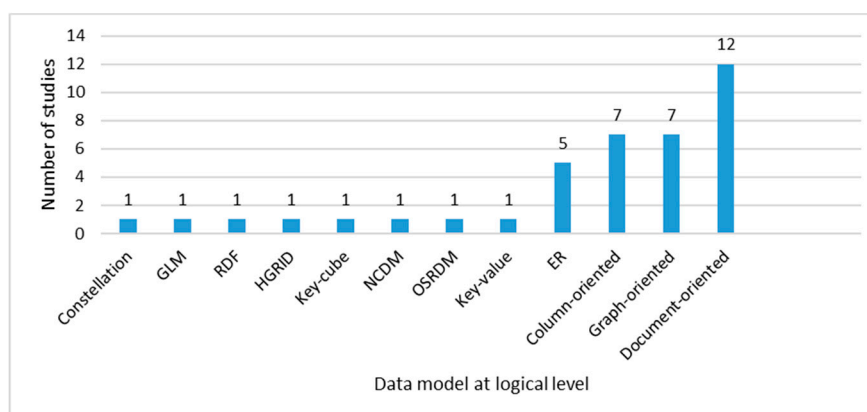


Figure 12. Data Model at logical abstraction level.

ER has also been proposed as a logical level model in three studies and the eight remaining studies have proposed independent solutions such as Constellation, Generic Logical Model (GLM), RDF, HGrid, NoSQL Collectional Data Model (NCDM), Open Scalable Relational Data Mode (OSRDM) and Key-cube. In addition, five studies do not propose any model at this level.

The data obtained in this section will be compared with the data from the following one, which determines the most studied data stores' implementations from the selected relevant articles.

Data Model at Physical Level

At this level, the physical implementations of the models in a specific DBMS are determined. According to the results obtained in Figure 13, the trend is towards the implementation in MongoDB with 13 proposals, followed by Neo4j with seven studies and, finally, Cassandra with six studies. These data match with the data presented in Figure 12, where the trend at the logical level is towards document-oriented, column-oriented and graph-oriented models.

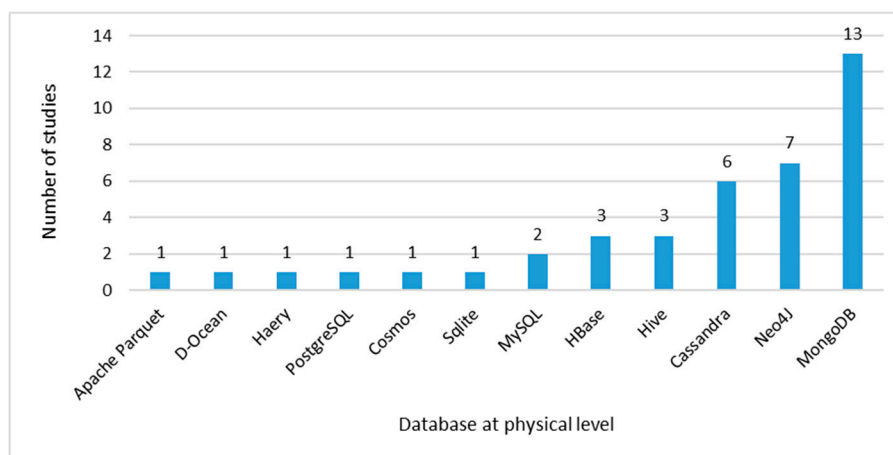


Figure 13. Data Model at physical abstraction level.

MongoDB is a document-oriented NoSQL DBMS that stores data in JSON. Each document has its own unique identifier, which is used as a primary key [40]. This DBMS is used by FourSquare, SourceForge, CERN and the European Organization for Nuclear Research, among other companies [59].

Neo4j is a graph-oriented NoSQL DBMS that organizes its data via labels for grouping nodes and edges, also called relationships and both nodes and edges can have properties in the form of key-value pairs [31]. This DBMS is especially used by Infojobs, a private company for job searches [59].

Cassandra is a column-oriented NoSQL DBMS that represents the data in a tabular form by columns and rows [16]. Big companies, such as Facebook and Twitter, use this DBMS [59].

We perceive that MongoDB is the most studied DBMS because large companies use it, probably because of its characteristics of support for aggregation and secondary indexes query operations and the consistency and partitioning tolerance mentioned in the Big Data Concepts subsection. Furthermore, these are open source databases with highly scalable, flexible and best performance compared with relational databases. These results give us the idea of a trend in each of the known data models—document-oriented—with MongoDB, column-oriented with Cassandra and graph with Neo4j.

There are also implementations of NoSQL HBase and Hive DBMS on a smaller scale and relational databases, of which PostgreSQL and MySQL are among the best-known. It is worth mentioning that there are studies that propose hybrid solutions with implementations that include different databases. More details about these studies will be presented in the Database section.

In summary, most studies propose the use of the ER model at the conceptual level, a document-oriented model at the logical level and the implementation of MongoDB at the physical level.

Transformation between Abstraction Levels

According to Figure 10, there are 25 selected studies that present their approaches at the three abstraction levels and eight studies at two levels differentiated from logical to physical, conceptual to logical and conceptual to physical. According to this concept, the proposed approaches for the transformation between the data abstraction levels are presented.

As presented in Figure 14, there are 19 studies where the authors propose their own novel mapping rules, which demonstrates the separate research that exists on this topic. Thus, it is difficult to decide which is the most appropriate when selecting any of them. Another interesting aspect is that 12 studies do not define transformation rules and there are six studies that propose transformations based on other techniques, such as the Linearization Algorithm (LA), ATL Transformation Language (ATL), Hoberman Heuristic (HH), Algorithm Cardinality (AC), Category Theory (CT) and Workload Space Constraint (WSC).

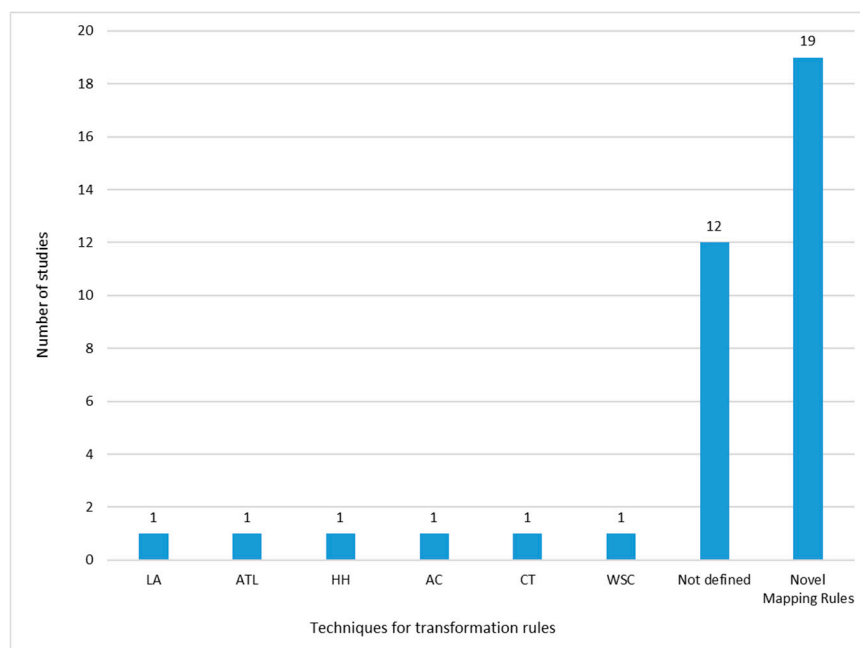


Figure 14. Transformation between data abstraction levels.

In general, the authors propose the below algorithm that takes a model as input, apply their own transformation rules and produce another model as output:

Input1: Conceptual Level: $C = \{C_i\}$, where C_i belongs to each i element from conceptual model.

Transformation rules: $R = \{r_j\}$, where r_j belongs to each j rule or constraint from mapping rules defined by the authors.

Output1: Logical Level: $L = C \cup R = \{l_k\}$, where l_k belongs to each k element from logical model.

Input2 = Output 1

Transformation rules': $R' = \{r'_j\}$, where r'_j belongs to each j rule or constraint from mapping rules defined by the authors.

Output2: Physical Level: $P = L \cup R' = \{p_k\}$, where p_k belongs to each k element from physical model.

Modeling Language

In this concept, it is important to clarify that a data model describes the characteristics of the data, its structure, its operations and its constraints; meanwhile, data modeling is the process of generating these models. The purpose of data modeling is for models to be used and understood by all modelers, developers and other persons working in the software development/engineering area in a standardized way. Thus, Figure 15 presents the results obtained from the mapping performed in Appendix A from the selected 36 relevant papers.

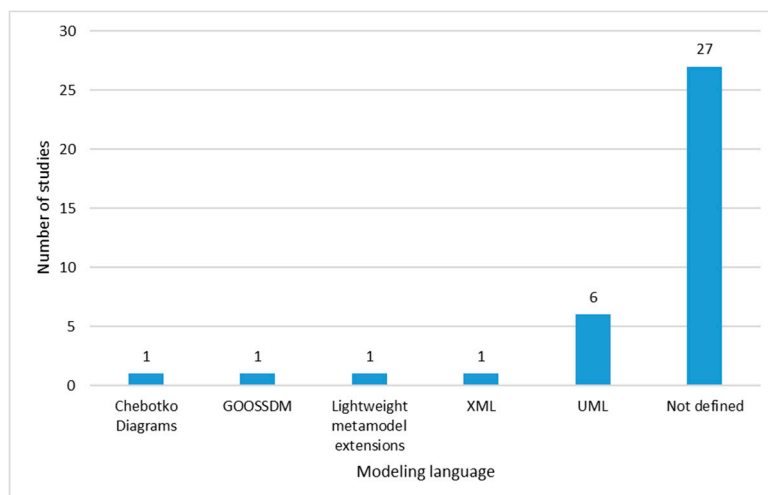


Figure 15. Data Modeling Language.

According to Figure 15, there is not a trend of data modeling language. There are 27 studies that do not define a standardized language used for the visualization of the models. Only six studies are adjusted to a standard such as the Unified Modeling Language (UML) and the other four propose the use of their own modeling language, like Chebotko diagrams, Graph Object Oriented Semi-Structured Data Model (GOOSSDM), lightweight metamodel extensions and XML. Of the six studies that present their models with the use of the UML, two of them use it in the conceptual level model [35,45], one uses it in the conceptual and logical models [14] and three use it in all conceptual, logical and physical levels [32,34,42].

According to several authors [16,41] and several implementation experiences, an important difference between relational databases and NoSQL databases is that the latter do not require normalization; that is, they support duplicated data. In this situation, data modeling in NoSQL databases generally begins with the formulation of questions about how the database data are to be consulted. These questions will define the entities and the relationships between those entities. This new paradigm moves from a data-driven modeling process to a query-driven modeling process. Thus, in the following concept, the methodologies for modeling data proposed in the final corpus of selected articles will be analyzed.

Modeling Methodology

As mentioned in the previous concept, this section aims to reveal the data modeling methodology proposed by the studies. According to the attained results, the trend of the proposals is to use the data-driven methodology presented in 33 studies. Data-driven modeling is a technique that, based on how the data are organized within the dataset and how they are derived from external systems, generates all the components to represent a model [60]. Only five studies propose query-driven modeling; it should be mentioned that the studies that propose workload-driven modeling, that are also based on query-driven, have been considered within these five studies.

Modeling Tool

We consider it important to know whether the selected studies also propose a computer tool for aiding in the model elaboration from scratch, validating the elaborated models and assisting in the automatic transformation between abstraction levels. Of the results obtained in the concept matrix, as shown in Figure 16, 29 studies do not propose any tool. Only two studies propose the use of the Eclipse Modeling Framework (EMF). Similarly, there are five studies that propose separate tools such as Kashlev Data Modeler (KDM) [16], scripts in Haskell [26], Mortadelo [32], Neoclipse [14],

NoSQL Schema Evaluator (NoSE) [54]. It is worth mentioning that all these tools allow modeling at the three levels of data abstraction.

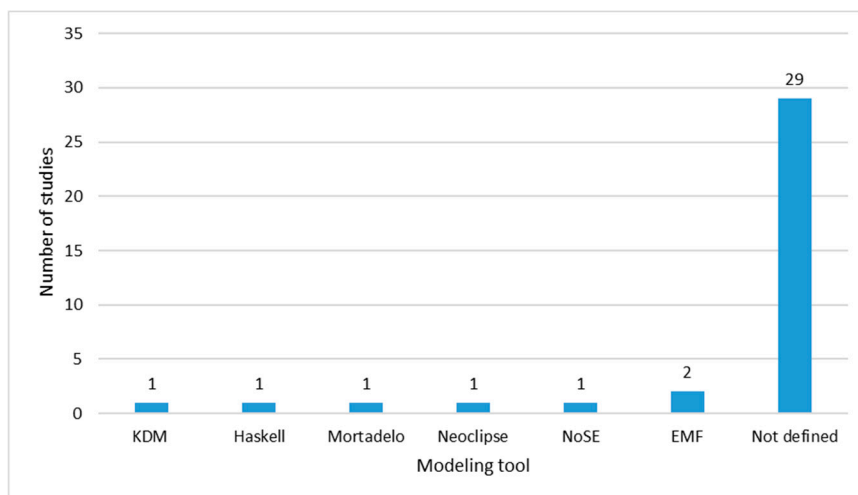


Figure 16. Data modeling tool (Eclipse Modeling Framework (EMF), Kashlev Data Modeler (KDM), NoSQL Schema Evaluator (NoSE)).

3.2.3. Database

In this section, we identified the proposed database types and the evaluation and performance comparisons carried out by the studies.

Database Type

At this aspect, we present the database types that the selected relevant studies proposed. They have been classified into two main groups, homogeneous and hybrids. By homogeneous, we mean those databases where the data are implemented in a single database. By hybrids, we mean systems where there are several databases implemented that can be relational and/or NoSQL. According to several studies [14,61,62], due to the variety characteristics of Big Data, the design and management of a database has become complex, so the systems are oriented towards a Polyglot Persistence System (PPS). This means that, when Big Data is stored, it is better to use different storage technologies, that are hybrid databases, so that applications can select the most appropriate one depending on the data they need [61]. Polyglot Persistence is the term used when an application is able to query data from different NoSQL databases [14].

According to the results of our SLR, 20 studies propose homogeneous solutions—that is, they focus on a single type of database—and only eight propose hybrid solutions. It is worth mentioning that, of these eight studies, none presents a solution that implements the following data models: E-R, document-oriented, column-oriented and graph. Likewise, eight studies do not define any type of physical implementation.

Among the studies that present solutions for at least three types of different DBMS are one that proposes implementations in SQLite, MongoDB, MySQL and Neo4j [26]; another one that proposes implementations in MySQL, MongoDB and Cosmos [28]; and another one that proposes implementations in Cassandra, MongoDB and Neo4j [35].

Evaluation and Performance Comparison

In this concept, the studies that have made an evaluation and performance comparison of their data models are presented. We consider this topic important, due to the results obtained in the Transformation between Abstraction Levels subsection, where the results found that nine studies present individual proposals, 12 undefined and six with different techniques.

Similarly, in the Modeling subsection, 27 studies do not present a standard modeling and, according to the Modeling Tool subsection, 29 studies do not define an automatic modeling tool. Based only on this information, it is difficult to select any of the proposals. For this reason, some of the works have carried out an evaluation of their proposed approaches, based on the data load time and the query execution time. From the requested results in Appendix A, eight studies submitted an evaluation of their proposals regarding query execution times [25,26,40,48,51,52,54,55], one study evaluates model transformation times [47] and another study compares data loading times [43]. Finally, some articles have mentioned the usefulness of reverse engineering to verify the validity of a proposed model [63]. Only one study is focused on this aspect but it does not test it [31].

According to the data of the key concepts analyzed previously, the trends and gaps found in our SLR study are presented in the discussion.

3.3. Discussion

In this section, we answer RQ3. Firstly, we refer to the relevant information attained from the bibliometric analysis. From this analysis, it is possible to verify the growth of this research topic. Since 2015, the number of studies has increased around 100 times and 2018 is the year with the most publications. Scopus is the source holding the highest number of relevant studies. The countries with the greatest contribution in the topic of Big Data modeling are the USA and China and more than 50% of the studies are funded.

Those data suggest existent interest in the topic in the research community. Although many researches refer to the Big Data analytics side, the engineering that takes care of the platforms, extraction techniques and loading and transforming data until reaching the required storage, is of no lesser importance, since these comprise the starting point required by analytics processes in order to produce value from the data.

Next, we will take the information attained from the SLR as a reference to discover the trends and gaps in the topic of interest.

3.3.1. Trends

To summarize the observed trends, Figure 17 helps as a reference to present the three main concepts that this analysis is focused on: source, modeling and database.

Source

For the Source aspect, the most researched data sources are unstructured data, specifically website information.

Modeling

For the Modeling aspect, seven main trends were identified within the analyzed corpus:

1. Most studies present their proposals at the conceptual, logical and physical abstraction levels;
2. ER model is the most used in the approaches at the conceptual abstraction level, followed by the multidimensional model and, thirdly, XML;
3. At the logical abstraction level, the most researched model is document-oriented, followed by column-oriented and graph-oriented;
4. At the physical abstraction level, implementations focus more on the MongoDB DBMS, followed by Neo4j and Cassandra. Thus, the following de facto standards have emerged, MongoDB for the document-oriented data model, Cassandra for column-oriented and Neo4j for graph data model. These data are supported by statistical information from DB-Engines Ranking - Trend Popularity of the Solid IT company as of December 2019 [64];
5. The most proposed modeling methodology is data-driven;

6. There is not a clear tendency towards a data modeling approach but there are proposals with UML and XML;
7. No data modeling tool is defined as a trend but some studies used EMF;
8. Regarding the different fields of application, ER is commonly used at the conceptual abstraction level in the different case studies. At the logical level, approaches for the migration from relational to document-oriented, graph and column-oriented models are proposed. Specifically, for spatio-temporal and transmission data, solutions for the graph model are proposed. And, for XML and JSON formats from web servers, solutions for all NoSQL data models are suggested.

Database

For the Database aspect, 55.55% of the studies focus on homogeneous database types, so the trend towards a persistent polyglot system is not observed.

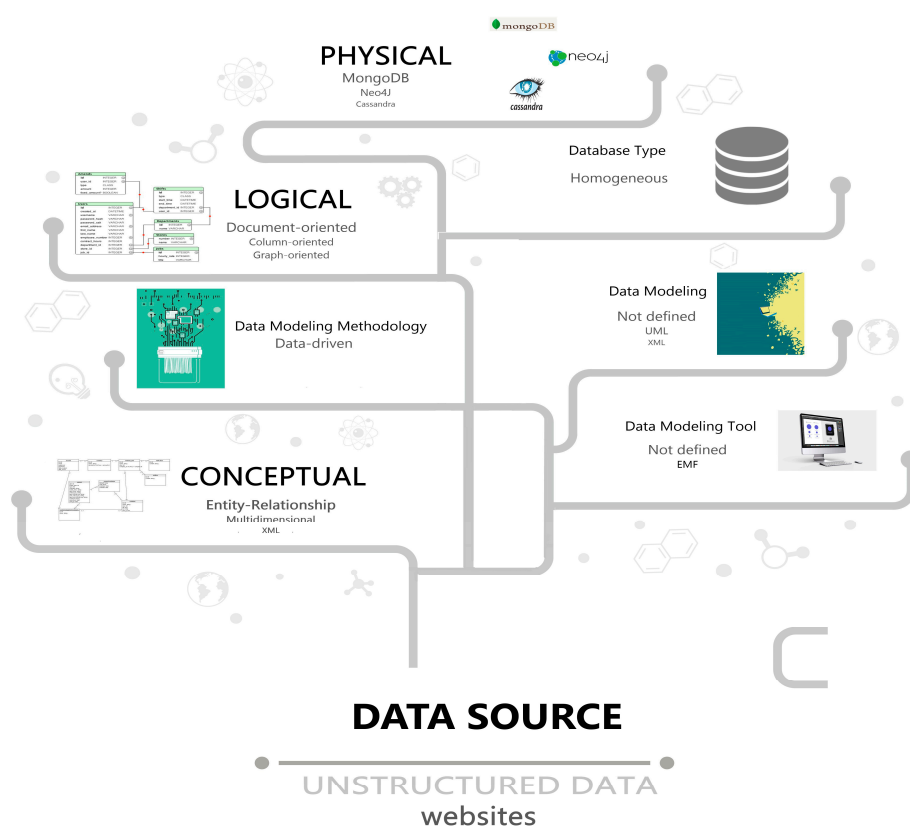


Figure 17. Trends for Big Data Modeling and Management.

3.3.2. Gaps

Figure 18 allowed us to identify also three main concepts for analyzing the gaps: source, modeling and database.

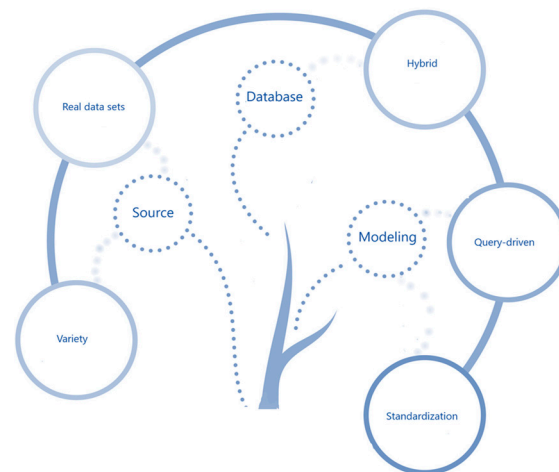


Figure 18. Gaps for Big Data Modeling and Management.

Source

For the Source aspect, two main gaps were identified within the analyzed corpus:

1. One of the main objectives of Big Data is to use the data to generate value; this value will depend on the needs of the business. For this reason, it is considered very important that studies are validated with real use cases, since only with the use of real datasets can whether value is being generated be verified. As result of our SLR, only ten studies, corresponding to 27.78%, use real datasets; 16.67% of these works present their case studies with data from websites, 5.56% use data from sensors and 2.78% involve electronic documents' data and images' metadata;
2. Additionally, other Big Data main features must be guaranteed, such as volume, velocity and veracity. According to the results of Table 6, not all of these features are justified in the approaches;
3. To comply with the variety, studies should consider that data can come in any format: structured, semi-structured or unstructured. Therefore, the proposed approaches should address any of these types. Only 2.78% have proposed a solution for all three types of data. The remaining 47.22% of the studies only present approaches for unstructured data, 30.56% only for structured data, 13.89% for semi-structured data, 2.78% combine structured and semi-structured data and 2.78% do not specify any type.

Modeling

For the Modeling aspect, the five main gaps found in our study correspond to standardization and data modeling methodology. Standards are considered important, since they guarantee a universal, uniform interpretation, readability of the data model, portability between database engines, platforms and applications, among other facilities for project managers, analysts, designers, developers and end-users of the databases. In summary, the results are as follows:

1. There is no standardization regarding the definition of mapping rules for the transformation between models at the conceptual, logical and physical data abstraction levels. Thus, the 36 studies propose different approaches for the transformation, making it difficult for users to choose the most appropriate one;
2. No NoSQL system has emerged as a standard or as a de facto standard yet;
3. There is no clearly defined use of some standardized language or method for modeling data at the logical and physical levels. According to the results of our SLR, only at the conceptual level can ER be considered as a trend, maybe because the conceptual level is technologic-agnostic;
4. As mentioned in the Modeling Language subsection, for NoSQL databases the new paradigm for the modeling process is query-driven. However, only five studies are focused on this methodology.

5. For Big Data analysis, the implementation of efficient systems is required. In the data-driven methodology, the models are designed before assessing what queries will be performed. The limitation of this solution is that all data will be stored, when only a limited fraction are needed to answer the required queries. On the contrary, in query-driven methodologies, a set of queries must be expressed, evaluated and integrated before modeling, so that planning can focus on the answers to the necessary queries. For solutions where real-time data are used, data-driven models have the risk of being impractical because of the diverse nature of data streams [65].
6. A research [66] demonstrates that the treatment of data, in terms of time–cost, with the use of query-driven requires less processing time than data-driven. This, in turn, leads to smaller amounts of consumed energy and, therefore, longer life of equipment.
7. According to another study [67], the use of a query-driven methodology allowed users to make queries in a natural user language that focuses on relevant regions of interest;

Techniques such as reverse engineering are not taken into account; thus, this detail could complement the studies. This method is widely used in relational databases on projects that work with existent databases and when no documented models exist (or there is a risk of these models being outdated).

Database

For the Database aspect, the main gap found is related to the variety characteristics of Big Data. To manage this feature correctly, NoSQL databases must be targeted to PPS. A PPS can be obtained through storage in hybrid databases; that is, it supports different storage technologies. Only eight studies have proposed solutions oriented to hybrid databases and, out of these, only one study demonstrated this with a real case that models and manages both structured, semi-structured and unstructured data and stores them in different databases [28].

4. Conclusions

As a limitation, we know that we were not able to collect all the existent research about the topic, due to the vast number of synonyms and constraints in the major search terms. However, the high quantity of primary studies—a set of 1376 papers was obtained—provides a trustworthy and complete data set.

The SLR study on Big Data modeling and management conducted in this paper has identified 36 relevant studies. These works have been selected based on the inclusion and selection process of research papers from digital scientific libraries until August 2019; there are no published papers prior to 2010.

We have raised three research questions that were answered through bibliometric analysis, an SLR and a discussion. The results of the bibliometric study provide a lot of relevant information. For instance, from 2015 onwards, the number of studies has increased significantly as this was the year when many countries started to contribute to the topic of interest. The leading countries in this topic are the USA and China. The USA have one of the most impactful studies.

The results from the SLR and from the Discussion also reveal some very interesting facts. For instance, more than 50% of the studies do not verify their proposals with real-world datasets. Big Data's velocity characteristic is not justified by 90% of the studies. ER is the most used model at the conceptual abstraction level, document-oriented is the most researched model at the logical abstraction level and MongoDB DBMS is the most frequent implementation at the physical level. Moreover, as the main gaps, we identified the lack of proposal evaluations and a few studies focused on query-driven methodology and hybrid database solutions.

The contribution of this SLR study, by clarifying the knowledge on the specific topic of Big Data modeling, can support researchers and practitioners in improving their Big Data projects. The relevant works collected can be a useful starting point to new studies into Big Data modeling.

Finally, we know that, due to the need to analyze large volumes of data with a variety of structures, which arrive in high frequency, database research became more focused towards NoSQL. However, NoSQL DBMSs have not been able to acquire some strengths that are already present in relational databases, such as the ability to support Consistency and Availability at the same time; for this reason, NewSQL has emerged and may pose a solution to the problems faced by both DBMSs. As a future work, we also expect to study this new database system.

We hope to keep the SLR up to date and present results for other concepts. Moreover, based on the gaps, we will drive our research in these aspects.

Author Contributions: D.M.-M. contributed to the conception and design of the work, acquisition of data, analysis and interpretation of data; S.L.-M. and R.N., critical reviewing of the accuracy of research, intellectual content and final approval of the version to be submitted. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: We thanks to the industry contribution to know the needs in their Big Data projects at store side.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Concept Matrix

ID	Authors	Title	Data Set				Model			Modeling				Database Type	Evaluation and Performance Comparison
			Source	Structured	Semi-Structured	Unstructured	Conceptual	Logical	Physical	Transformation between Abstraction Levels	Modeling Language	Modeling Tool	Modeling Methodology		
1	Jie SONG, Hongyan HE, Richard THOMAS, Yubin BAO, Ge YU	Haery: a Hadoop based Query System on Accumulative and High-dimensional Data Model for Big Data	sample data	-	-	text file	key-value	key-cube	Haery	Query Algorithm Linearization Algorithm	Not defined	NA	model-driven query-driven	Homogeneous	queries
2	Laurent Thiry, Heng Zhao and Michel Hassenforder	Categories for (Big) Data models and optimization	sample data	-	csv	-	Entity–Relationship (ER)	ER document graph	Sqlite Mongo MySQL Neo4J	Category Theory (CT)	Not defined	Haskell script	model-driven	Hybrid	queries
3	Victor Martins de Sousa, Luis Mariano del Val Cura	Logical Design of Graph Databases from an Entity Relationship Conceptual Model	sample data	-	-	website	Extended Binary ER (EB-ER)	graph	Neo4j	Mapping Algorithm Cardinality Constraints Algorithm Vertex Constraints Algorithm	Not defined	NA	model-driven	Homogeneous	NA
4	Igor Zečević, Petar Bjeljac, Branko Perišić, Stevan Stankovski, Danijel Venus & Gordana Ostojić	Model-driven development of hybrid databases using lightweight metamodel extensions	real data	management data	JSON files	websites	ER	ER document key-value graph	MySQL MongoDB Cosmos	Mapping rules	lightweight metamodel extensions	NA	model-driven	Hybrid	NA
5	Antonio M. Rinaldi, Cristiano Russo	A Semantic-based Model to represent Multimedia Big Data	real data	-	-	metadata from images	-	graph	Neo4j	NA	Not defined	NA	model-driven query-driven	Homogeneous	NA
6	Shady Hamoud, Zurinahni Zainol	Document-Oriented Data Schema for Relational Database Migration to NoSQL	sample data	relational database data	-	-	ER	document	MongoDB	Mapping rules Normalization & Denormalization process	Not defined	NA	model-driven	Homogeneous	NA

7	Dippy Aggarwal, Karen C. Davis	Employing Graph Databases as a Standardization Model for Addressing Heterogeneity and Integration	sample data	relational database data RDF	csv	-	ER Resource Description Framework (RDF)	graph	Neo4j	Mapping rules	Not defined	NA	model-driven	Homogeneous	NA
8	Alfonso de la Vega, Diego García-Saiz, Carlos Blanco, Marta Zorrilla, and Pablo Sánchez	Mortadelo: A Model-Driven Framework for NoSQL Database Design	sample data	relational database	-	-	Generic Data Model (GDM)	column document	Cassandra MongoDB	Mapping rules	Unified Modeling Language (UML)	Mortadelo	model-driven	Hybrid	NA
9	Xu Chena, Li Yanb, Weijun Lia, Fu Zhangc	Fuzzy Spatio-temporal Data Modeling Based on XML Schema	real data	-	-	sensors	XML	-	-	NA	Tree	NA	model-driven	NA	NA
10	Maribel Yasmina Santos, Bruno Martinho & Carlos Costa	Modeling and implementing big data warehouses for decision support	real data	multidimensional	-	websites	ER	ER	Hive	Mapping rules	Not defined	NA	model-driven	Homogeneous	NA
11	Kwangchul Shin, Chulhyun Hwang, Hoekyung Jung	NoSQL Database Design Using UML Conceptual Data Model Based on Peter Chen's Framework	sample data	-	csv	-	ER	ER	-	Mapping rules	UML	NA	model-driven	NA	NA
12	Fatma Abdelhedi, Amal Ait Brahim, Faten Atigui, Gilles Zurfluh	Logical Unified Modeling For NoSQL DataBases	NA	a class diagram from a relational database	-	-	ER	GLM	Cassandra MongoDB Neo4J	Mapping rules	UML at conceptual	EMF	model-driven	Hybrid	NA
13	Victor Martins de Sousa, Luis Mariano del Val Cura	A NoSQL Data Model For Scalable Big Data Workflow Execution	real data	-	-	sensors	NA	NCDM	NA	NA	NA	NA	NA	NA	NA
14	Massimo Villari, Antonio Celesti, Maurizio Giacobbe and Maria Fazio	Enriched E-R Model to Design Hybrid Database for Big Data Solutions	NA	-	-	e-health medical records	EE-R	-	-	NA	NA	NA	model-driven	NA	NA

15	Maribel Yasmina Santos, Carlos Costa	Data Warehousing in Big Data From Multidimensional to Tabular Data Models	sample data	multidimensional	-	-	multidimensional	constellation	Hive	Mapping rules	NA	NA	model-driven	Homogeneous	NA
16	Maribel Yasmina Santos, Carlos Costa	Data Models in NoSQL Databases for Big Data Contexts	sample data	relational database data	-	-	ER	column ER	Hbase Hive	Mapping rules	NA	NA	model-driven	Hybrid	NA
17	Ganesh B. Solanke, K. Rajeswari	Migration of Relational Database to MongoDB and Data Analytics using Naïve Bayes Classifier based on Mapreduce Approach	sample data	relational database data	-	-	ER	document	MongoDB	Mapping rules	NA	SQL scripts	model-driven	Homogeneous	queries
18	Vincent Reniers, Dimitri Van Landuyt, Ansar Rafique, Wouter Joosen	Schema Design Support for Semi-Structured Data: Finding the Sweet Spot between NF and De-NF	real data	-	-	websites	ER	document	-	Mapping rules	NA	NA	model-driven workload-driven	Homogeneous	NA
19	Fatma Abdelhedi, Amal Ait Brahim, Faten Atigui and Gilles Zurfluh	Big Data and Knowledge Management: How to Implement Conceptual Models in NoSQL Systems?	sample data	-	-	sensors	ER	column	Hbase Cassandra	Mapping rules	UML XML	EMF	model-driven	Hybrid	NA
20	Max Chevalier, Mohammed El Malki, Arlind Kopliku, Olivier Teste and Ronan Tournier	Document-oriented Models for Data Warehouses	sample data	-	JSON files	-	multidimensional	document	MongoDB	Mappings rules	NA	NA	model-driven	Homogeneous	data load
21	Shreya Banerjee, Renuka Shaw, Anirban Sarkar, Narayan C Debnath	Towards Logical Level Design of Big Data	sample data	relational database data	-	-	GOOSSDM	document	MongoDB	Mapping rules	GOOSSDM	NA	model-driven	Homogeneous	NA

22	Artem Chebotko, Andrey Kashlev, Shiyong Lu	A Big Data Modeling Methodology for Apache Cassandra	sample data	-	-	website	ER	column	Cassandra	Mapping rules Application workflow Mapping patterns	Chebotko Diagrams at logical level	KDM	query-driven	Homogeneous	NA
23	Wenduo Feng*, Ping Gu, Chao Zhang, Kai Zhou	Transforming UML Class Diagram into Cassandra Data Model with Annotations	sample data	-	JSON files	-	ER	column	Cassandra	ATL Transformation Language (ATL) Mapping rules	UML at conceptual	NA	model-driven	Homogeneous	NA
24	Ling Chen, Jian Shao, Zhou Yu, Jianling Sun, Fei Wu, Yueting Zhuang	RAISE: A Whole Process Modeling Method for Unstructured Data Management	NA	-	-	website	XML	-	D-Ocean	NA	NA	NA	model-driven	Homogeneous	NA
25	M. Chevalier, M. El Malki, A. Kopliku, O. Teste and R. Tournier	Implementation of Multidimensional Databases with Document-Oriented NoSQL	sample data	-	JSON files	-	multidimensional	document	MongoDB	Mapping rules	NA	NA	model-driven	Homogeneous	transformation
26	Dewi W. Wardani, Josef Küng	Semantic Mapping Relational to Graph Model	sample data	relational database data	-	-	ER	graph RDF linked data	Neo4j	Mapping rules	NA	NA	model-driven	Homogeneous	queries
27	Ming Zhe, Kang Ruihua	A Data Modeling Approach for Electronic Document based on Metamodel	real data	-	JSON files	electronic documents	Tree	-	-	NA	NA	NA	model-driven	NA	NA
28	Mohamed Nadjib Mami, Simon Scerri, Sören Auer and Maria-Esther Vidal	Towards Semantification of Big Data Technology	sample data	-	-	website	RDF RDF/XML	column	Apache Parquet	NA	NA	NA	NA	NA	NA
29	Dan Han, Eleni Stroulia	HGrid: A Data Model for Large Geospatial Data Sets in Hbase	sample data	-	-	sensors	-	Hgrid	Hbase	NA	NA	NA	model-driven	Homogeneous	queries
30	Zhiyun Zheng, Zhimeng Du, Lun Li, Yike Guo	Big Data-Oriented Open Scalable Relational Data Model	real data	-	-	websites	-	OSRDM	-	NA	NA	NA	model-driven	NA	queries

31	Dongqi Wei, Chaoling Li, Wumuti Naheman, Jianxin Wei, Junlu Yang	Organizing and Storing Method for Large-scale Unstructured Data Set with Complex Content	NA	-	-	geosciences data	Tree	-	-	NA	NA	NA	model-driven	NA	NA
32	Karamjit Kaur, Rinkle Rani	Modeling and Querying Data in NoSQL Databases	real data	-	-	websites	ER	document graph	MongoDB Neo4j	NA	UML at conceptual and document	Neclipse	model-driven	Hybrid	NA
33	Michael J. Mior, Kenneth Salem, Ashraf Aboulhaga and Rui Liu	NoSE: Schema Design for NoSQL Applications	real data	-	-	websites	ER	column	Cassandra	Workload Space Constraint	NA	NoSE	query-driven	Homogeneous	queries
34	Max Chevalier, Mohammed El Malki, Arlind Kopliku, Olivier Teste, Ronan Tournier	Document-Oriented Data Warehouses: Models and Extended Cuboids	sample data	relational database	-	-	multidimensional	document	MongoDB PostgreSQL	Mapping rules	NA	NA	model-driven	Hybrid	queries
35	Harley Vera, Wagner Boaventura, Maristela Holanda, Valeria Guimarães, Fernanda Hondo	Data Modeling for NoSQL Document-Oriented Databases	sample data	-	-	sensors	NGN	document	MongoDB	NA	NA	NA	model-driven	Homogeneous	NA
36	Robert T. Mason	NoSQL Databases and Data Modeling Techniques for a Document-oriented NoSQL Database	sample data	-	-	-	ER	document	MongoDB	Hoberman heuristic	NA	NA	model-driven	Homogeneous	NA

References

1. Kitchenham, B. *Procedures for Performing Systematic Reviews*; Keele University: Keele, UK, 2004; Volume 33, pp. 1–26.
2. Google. Google Trends. Available online: <https://trends.google.es/trends/explore?date=all&q=%22big%20data%22> (accessed on 23 August 2019).
3. Rider, F. *The Scholar and the Future of the Research Library: A Problem and Its Solution*; Hadham Press: New York, NY, USA, 1944; pp. 98–100.
4. Cox, M.; Ellsworth, D. Application-controlled demand paging for out-of-core visualization. In Proceedings of the 8th IEEE Conference on Visualization, Phoenix, AZ, USA, 24 October 1997; pp. 235–244.
5. Ribeiro, A.; Rodrigues da Silva, A. Data Modeling and Data Analytics: A Survey from a Big Data Perspective. *J. Softw. Eng. Appl.* **2015**, *8*, 617–634. [CrossRef]
6. Shafer, T. The 42 V's of Big Data and Data Science. Available online: <https://www.elderresearch.com/blog/42-v-of-big-data> (accessed on 23 August 2019).
7. Manogaran, G.; Thota, C.; Lopez, D.; Vijayakumar, V.; Abbas, K.M.; Sundarsekar, R. Big Data Knowledge System in Healthcare. In *Internet of Things and Big Data Technologies for Next Generation Healthcare*; Springer International Publishing: Cham, Switzerland, 2017; Volume 23, pp. 133–157.
8. Persico, V.; Pescapé, A.; Picariello, A.; Sperli, G. Benchmarking big data architectures for social networks data processing using public cloud platforms. *Future Gener. Comput. Syst.* **2018**, *89*, 98–109. [CrossRef]
9. Costa, C.; Santos, Y. Big Data: State-of-the-art Concepts, Techniques, Technologies, Modeling Approaches and Research Challenges. *Int. J. Comput. Sci.* **2017**, *44*, 1–17.
10. Davoudian, A.; Chen, L.; Liu, M. A Survey on NoSQL Stores. *ACM Comput. Surv.* **2018**, *51*, 1–43. [CrossRef]
11. CISCO. Big Data: Not Just Big, but Different—Part 2. Available online: https://www.cisco.com/c/dam/en_us/about/ciscoitnetwork/enterprise-networks/docs/i-bd-04212014-not-just-big-different.pdf (accessed on 10 September 2019).
12. O'Sullivan, P.; Thompson, G.; Clifford, A. Applying data models to big data architectures. *IBM J. Res. Dev.* **2014**, *58*, 18:1–18:11. [CrossRef]
13. CISCO VNI. Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper. Available online: <https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.htm> (accessed on 10 September 2019).
14. Karamjit, K.; Rinkle, R. Modeling and querying data in NoSQL databases. In Proceedings of the 1st IEEE International Conference on Big Data, Silicon Valley, CA, USA, 6–9 October 2013; pp. 1–7.
15. Wu, D.; Sakr, S.; Zhu, L. Big Data Storage and Data Models. In *Handbook of Big Data Technologies*; Springer International Publishing: Cham, Switzerland, 2017; pp. 3–29. [CrossRef]
16. Chebotko, A.; Kashlev, A.; Lu, S. A Big Data Modeling Methodology for Apache Cassandra. In Proceedings of the 4th IEEE International Congress on Big Data, New York, NY, USA, 27 June–2 July 2015; pp. 238–245.
17. Edlich, S. List of NoSQL Database Management Systems. Available online: <http://nosql-database.org/> (accessed on 15 September 2019).
18. Santos, M.Y.; Martinho, B.; Costa, C. Modelling and implementing big data warehouses for decision support. *J. Manag. Anal.* **2017**, *4*, 111–129. [CrossRef]
19. Centre for Reviews and Dissemination; (University of York: Centre for Reviews and Dissemination, York, UK). Undertaking systematic reviews of research on effectiveness: CRD's guidance for carrying out or commissioning reviews. Personal communication, 2001.
20. Martins de Sousa, V.; del Val Cura, L.M. Modelagem Lógica para Bancos de Dados NoSQL: Uma revisão sistemática. *Anais WCF* **2016**, *3*, 32–39.
21. Brewer, E.A. Towards robust distributed systems. In Proceedings of the ACM Symposium on Principles of Distributed Computing, Portland, Oregon, 16–19 July 2000; Volume 7.
22. Pouyanfar, S.; Yang, Y.; Chen, S.; Shyu, M.L.; Iyengar, S. Multimedia big data analytics: A survey. *ACM Comput. Surv.* **2018**, *51*, 10. [CrossRef]
23. Bruno, R.; Ferreira, P. A Study on Garbage Collection Algorithms for Big Data Environments. *ACM Comput. Surv.* **2018**, *51*, 20. [CrossRef]

24. Gusenbauer, M.; Haddaway, N. Which Academic Search Systems are Suitable for Systematic Reviews or Meta-Analyses? Evaluating Retrieval Qualities of Google Scholar, PubMed and 26 other Resources. *Res. Synth. Methods* **2019**. [[CrossRef](#)] [[PubMed](#)]
25. Song, J.; He, H.; Thomas, R.; Bao, Y.; Yu, G. Haery: A Hadoop based Query System on Accumulative and High-dimensional Data Model for Big Data. *IEEE Trans. Knowl. Data Eng.* **2019**. [[CrossRef](#)]
26. Thiry, L.; Zhao, H.; Hassenforder, M. Categories for (Big) Data models and optimization. *J. Big Data* **2018**, *5*, 1–20. [[CrossRef](#)]
27. Martins de Sousa, V.; del Val Cura, L.M. Logical Design of Graph Databases from an Entity-Relationship Conceptual Model. In Proceedings of the 20th International Conference on Information Integration and Web-Based Applications and Services, Yogyakarta, Indonesia, 19–21 November 2018; pp. 183–189.
28. Zečević, I.; Bjeljac, P.; Perišić, B.; Stankovski, S.; Venus, D.; Ostojić, G. Model driven development of hybrid databases using lightweight metamodel extensions. *Enterp. Inf. Syst.* **2018**, *12*, 1221–1238. [[CrossRef](#)]
29. Rinaldi, A.; Russo, C. A Semantic-based Model to represent Multimedia Big Data. In Proceedings of the 10th International Conference on Management of Digital EcoSystems, Tokyo, Japan, 25–28 September 2018; pp. 31–38.
30. Hamouda, S.; Zainol, Z. Document-Oriented Data Schema for Relational Database Migration to NoSQL. In Proceedings of the 2017 International Conference on Big Data Innovations and Applications, Prague, Czech Republic, 21–23 August 2018; pp. 43–50.
31. Aggarwal, D.; Davis, K. Employing Graph Databases as a Standardization Model for Addressing Heterogeneity and Integration. *Adv. Intell. Syst. Comput.* **2018**, *561*, 109–138. [[CrossRef](#)]
32. De la Vega, A.; García-Saiz, D.; Blanco, C.; Zorrilla, M.; Sánchez, P. Mortadelo: A Model-Driven Framework for NoSQL Database Design. In Proceedings of the 8th International Conference on Model and Data Engineering, Marrakesh, Morocco, 24–26 October 2018; pp. 41–57.
33. Chen, X.; Yan, L.; Li, W.; Zhang, F. Fuzzy spatio-temporal data modeling based on XML schema. *Filomat* **2018**, *32*, 1663–1677. [[CrossRef](#)]
34. Shin, K.; Hwang, C.; Jung, H. NoSQL Database Design Using UML Conceptual Data Model Based on Peter Chen's Framework. *Int. J. Appl. Eng. Res.* **2017**, *12*, 632–636.
35. Abdelhedi, F.; Brahim, A.A.; Atigui, F. Logical unified modeling for NoSQL databases. In Proceedings of the 19th International Conference on Enterprise Information Systems, Porto, Portugal, 26–29 April 2017; pp. 249–256.
36. Mohan, A.; Ebrahimi, M.; Lu, S.; Kotov, A. A NoSQL Data Model for Scalable Big Data Workflow Execution. In Proceedings of the 2016 IEEE International Congress on Big Data, San Francisco, CA, USA, 27 June–2 July 2016; pp. 52–59.
37. Villari, M.; Celesti, A.; Giacobbe, M.; Fazio, M. Enriched E-R Model to Design Hybrid Database for Big Data Solutions. In Proceedings of the 2016 IEEE Symposium on Computers and Communication, Messina, Italy, 27–30 June 2016; pp. 163–166.
38. Santos, M.Y.; Costa, C. Data Warehousing in Big Data: From Multidimensional to Tabular Data Models. In Proceedings of the 9th International C* Conference on Computer Science and Software Engineering, Porto, Portugal, 20–22 July 2016; Volume 20, pp. 51–60.
39. Santos, M.Y.; Costa, C. Data Models in NoSQL Databases for Big Data Contexts. In Proceedings of the International Conference on Data Mining and Big Data, Bali, Indonesia, 25–30 June 2016; Volume 9714, pp. 475–485.
40. Solanke, G.B.; Rajeswari, K. Migration of Relational Database to MongoDB and Data Analytics using Naïve Bayes Classifier based on Mapreduce Approach. In Proceedings of the 2017 International Conference on Computing, Communication, Control and Automation, Maharashtra, India, 17–18 July 2017; pp. 1–6.
41. Reniers, V.; Van Landuyt, D.; Rafique, A.; Joosen, W. Schema Design Support for Semi-Structured Data: Finding the Sweet Spot between NF and De-NF. In Proceedings of the 2017 IEEE International Conference on Big Data, Boston, MA, USA, 11–14 December 2017; pp. 2921–2930.
42. Abdelhedi, F.; Ait Brahim, A.; Atigui, F.; Zurfluh, G. Big Data and Knowledge Management: How to Implement Conceptual Models in NoSQL Systems. In Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016), Porto, Portugal, 9–11 November 2016; pp. 235–240.

43. Chevalier, M.; El Malki, M.; Kopliku, A.; Teste, O.; Tournier, R. Document-Oriented Models for Data Warehouses: NoSQL Document-Oriented for Data Warehouses. In Proceedings of the 18th International Conference on Enterprise Information Systems (ICEIS 2016), Rome, Italy, 25–28 April 2016; pp. 142–149. [\[CrossRef\]](#)
44. Banerjee, S.; Shaw, R.; Sarkar, A.; Debnath, N.C. Towards Logical Level Design of Big Data. In Proceedings of the 13th IEEE International Conference on Industrial Informatics, Cambridge, UK, 22–24 July 2015; pp. 1665–1671.
45. Feng, W.; Gu, P.; Zhang, C.; Zhou, K. Transforming UML Class Diagram into Cassandra Data Model with Annotations. In Proceedings of the IEEE International Conference on Smart City/SocialCom/SustainCom, Chengdu, China, 19–21 December 2015; pp. 798–805.
46. Chen, L.; Shao, J.; Yu, Z.; Sun, J.; Wu, F.; Zhuang, Y. RAISE: A Whole Process Modeling Method for Unstructured Data Management. In Proceedings of the 2015 IEEE International Conference on Multimedia Big Data, Beijing, China, 20–22 April 2015; pp. 9–12.
47. Chevalier, M.; Malki, M.; Kopliku, A.; Teste, O.; Tournier, R. Implementation of Multidimensional Databases with Document-Oriented NoSQL. *Lect. Notes Comput. Sci.* **2015**, *9263*, 379–390. [\[CrossRef\]](#)
48. Wardani, D.; Küng, J. Semantic Mapping Relational to Graph Model. In Proceedings of the 2014 International Conference on Computer, Control, Informatics and Its Applications, Bandung, Indonesia, 21–23 October 2014; pp. 160–165.
49. Zhe, M.; Ruihua, K. A Data Modeling Approach for Electronic Document Based on Metamodel. In Proceedings of the 2013 International Conference on Computer Sciences and Applications, Wuhan, China, 14–15 December 2013; pp. 829–832.
50. Mami, M.N.; Scerri, S.; Auer, S.; Vidal, M.E. Towards Semantification of Big Data Technology. In Proceedings of the 18th International Conference on Big Data Analytics and Knowledge Discovery, Porto, Portugal, 6–8 September 2016; pp. 376–390.
51. Han, D.; Stroulia, E. HGrid: A Data Model for Large Geospatial Data Sets in HBase. In Proceedings of the 6th International Conference on Cloud Computing, Shanghai, China, 9–11 November 2013; pp. 910–917.
52. Zheng, Z.; Du, L.; Guo, Y. BigData oriented open scalable relational data model. In Proceedings of the 3rd IEEE International Congress on Big Data, Washington, DC, USA, 27–30 October 2014; pp. 398–405.
53. Wei, D.; Li, C.; Naheman, W.; Wei, J.; Yang, J. Organizing and Storing Method for Large-scale Unstructured Data Set with Complex Content. In Proceedings of the 5th International Conference on Computing for Geospatial Research and Application, Washington, DC, USA, 4–6 August 2014; pp. 70–76.
54. Mior, M.; Salem, K.; Aboulnga, A.; Liu, R. NoSE: Schema design for NoSQL applications. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 2275–2289. [\[CrossRef\]](#)
55. Chavalier, M.; El Malki, M.; Kopliku, A.; Teste, O.; Tournier, R. Document-Oriented Data Warehouses: Models and Extended Cuboids. In Proceedings of the 10th IEEE International Conference on Research Challenges in Information Science, Grenoble, France, 1–3 June 2016; pp. 1–11.
56. Vera, H.; Boaventura, W.; Holanda, M.; Guimaraes, V.; Hondo, F. Data modeling for NoSQL document-oriented databases. In Proceedings of the CEUR Workshop, Turin, Italy, 28–29 September 2015; pp. 129–135.
57. Mason, R.T. NoSQL Databases and Data Modeling Techniques for a Document-oriented NoSQL Database. In Proceedings of the Informing Science & IT Education Conference, Tampa, FL, USA, 2–5 July 2015; pp. 259–268.
58. Webster, J.; Watson, R. Analyzing the Past to Prepare for the Future: Writing a Literature Review. *Manag. Inf. Syst.* **2002**, *26*, 13–23.
59. ACENS. Bases de Datos NoSQL. Qué son y Tipos que nos Podemos Encontrar. Available online: <https://www.acens.com/wp-content/images/2014/02/bbdd-nosql-wp-acens.pdf> (accessed on 20 September 2019).
60. IBM and IBM Knowledge Center. Data Driven Modeling. Available online: https://www.ibm.com/support/knowledgecenter/en/SSGTJF/com.ibm.help.omcloud.omniconfig.doc/productconcepts/c_OC_DDMIntro.html (accessed on 20 September 2019).
61. Abelló, A. Big Data Design. In Proceedings of the 18th International Workshop on Data Warehousing and OLAP, Melbourne, Australia, 23 October 2015; pp. 35–38.
62. Schaarschmidt, M.; Gessert, F.; Ritter, N. Towards automated polyglot persistence. In Proceedings of the Datenbanksysteme für Business, Technologie und Web, Stuttgart, Germany, 6–7 March 2015.
63. Sevilla Ruiz, D.; Morales, S.F.; Garcia Molina, J. Inferring Versioned Schemas from NoSQL Databases and Its Applications. *Lect. Notes Comput. Sci.* **2015**, *9381*, 467–480. [\[CrossRef\]](#)

64. Solid, I.T. DB-Engines Ranking—Trend Popularity. Available online: https://db-engines.com/en/ranking_trend (accessed on 1 January 2020).
65. Dell’Aglio, D.; Balduini, M.; Della Valle, E. Applying semantic interoperability principles to data stream management. In *Data Management in Pervasive Systems*; Springer: Cham, Switzerland, 2015; pp. 135–166. [[CrossRef](#)]
66. Haghighi, M. Market-based resource allocation for energy-efficient execution of multiple concurrent applications in wireless sensor networks. In *Mobile, Ubiquitous, and Intelligent Computing*; Springer: Berlin, Germany, 2014; pp. 173–178. [[CrossRef](#)]
67. Bajcsy, R.; Joshi, A.; Krotkov, E.; Zwarico, A. Landscan: A natural language and computer vision system for analyzing aerial images. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, Los Angeles, CA, USA, 18–23 August 1985; Volume 2, pp. 919–921.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).