REVIEW ARTICLE
DOI: 10.5433/1679-0375.2021v42n1p113

Review and comparison of works on heterogeneous data and semantic analysis in Big Data

Revisão e comparação de trabalhos sobre dados heterogêneos e análise semântica em Big Data

Vitor Valério de Souza Campos[1]; Jacques Duílio Brancher[2]; Francyelcyo Pussi Farias[3]; José Luiz Villela Marcondes Mioni[4]; Pedro Luiz Garbim Brahim[5]

Abstract

In integration approaches, heterogeneity is one of the main challenging factors on the task of providing integration among different data sources, whose solution lies in the search for equality among them. This work describes the state of the art and theoretical foundation involved in the structural and semantic analysis of heterogeneous data and information. The work aims to review methods and techniques used in data integration in Big Data, considering data heterogeneity, reviewing techniques that use the concepts of Semantic Web, Cloud Computing, Data Analysis, Big Data, Data Warehouse and other technologies to solve the problem of data heterogeneity. The research was divided into three stages. In the first stage, articles were selected from digital libraries according to their titles and keywords. In the second stage, the papers went through a second filter based on their summary, and, besides that, duplicate articles were also removed. The works' introduction and conclusion were analyzed in the third stage to select the articles belonging to this systematic review. Throughout the study, articles were analyzed, compared and categorized. At the end of each section, the interrelationships and possible areas for future work were shown.

Keywords: Data analysis. Heterogeneous data. Big data. Semantic heterogeneity. Structural heterogeneity.

Resumo

Nas abordagens de integração, a heterogeneidade é um dos principais fatores que dificulta a tarefa de prover integração entre diferentes fontes de dados, cuja solução encontra-se na busca de igualdades entre elas. Este trabalho descreve o estado da arte e a fundamentação teórica envolvida na análise estrutural e semântica de dados e informação heterogênea. O trabalho tem como objetivo revisar métodos e técnicas utilizados na integração de dados em Big Data, considerando a heterogeneidade dos dados, revisando técnicas que utilizam os conceitos da Web Semântica, Computação em nuvem, Análise de dados, Big Data, Data Warehouse e outras tecnologias para resolver o problema da heterogeneidade de dados. A pesquisa foi dividida em três etapas. Na primeira etapa foram selecionados artigos nas bibliotecas digitais com base no título e nas palavras-chave. Na segunda etapa passaram por um segundo filtro baseado em seu resumo, e, além disso, artigos duplicados também foram removidos. Na terceira etapa a introdução e conclusão dos trabalhos foram analisados para que se escolhesse os artigos pertencentes a esta revisão sistemática. Os artigos foram analisados, comparados, e categorizados ao longo do estudo. Ao final de cada seção, mostrou-se as interrelações e a possíveis áreas para trabalhos futuros

Palavras-chave: Análise de dados. Dados heterogêneos. Big data. Heterogeneidade semântica. Heterogeneidade estrutural.

[1] Prof. Dr., Dept. Computer Science, UEL, Londrina, PR, Brazil; E-mail: valerio@uel.br
[2] Prof. Dr., Dept. Computer Science, UEL, Londrina, PR, Brazil; E-mail: jacques@uel.br
[3] Master's Degree student, Dept. Computer Science, UEL, Londrina, PR, Brazil; E-mail: elcyof@gmail.com
[4] Master's Degree student, Dept. Computer Science, UEL, PR, Brazil; E-mail: luiz.vmm@gmail.com
[5] Graduation student, Dept. Computer Science, UEL, PR, Brazil; E-mail: plgbpedro2@gmail.com

Introduction

In building a Data Warehouse (DW), data transfer activities from the origin data source to its destination source are highly complicated and iterative, especially as new data sources are added. DW construction and data management approaches have been supported by integration tools through the Extract, Transformation, and Load (ETL) process (KIMBALL; ROSS, 2011). These tools are designed to handle large volumes of data and are not flexible for handling semi-structured or unstructured data. Another aspect that must be considered in the integration process is heterogeneity.

In traditional approaches for integration, heterogeneity is one of the main factors that increase complexity to provide integration between different data sources. This factor occurs at two levels. Technical or structural level: heterogeneity refers to the technological differences between the various components of hardware, software, communication systems, Database Management Systems (DBMS), and programming languages currently found in large corporations. Conceptual or semantic level: heterogeneity is a product of different interpretations of the meaning of certain terms from different data models (ARPUTHAMARY; AROCKIAM, 2015).

Therefore, as systems on different models are integrated, a solution to structural and semantic heterogeneity problems is needed. Several examples of structural conflicts exist (ARPUTHAMARY; AROCKIAM, 2015), such as:

- Conflicts between names, which derive from different ways of modeling a real-world problem, as it is common to find data with the same semantic content, yet with different names (synonyms) and the presence of homonyms that denote the use of the same name to refer to different concepts;
- Conflicts between different Database (DB) schemes that use distinct structures to represent the same information;
- Conflicts between different DB schemes that use similar structures to represent diverse information;
- Conflicts of inclusion of entities and attributes due to generalization abstraction, which happens when an entity from one database is logically included in another entity from another DB;
- Data type conflicts due to aggregation abstraction, which happens when the domain or type of the attribute is different for semantically equivalent attributes;
- Composition conflicts occur due to the abstraction of aggregation, which happens when similar concepts are represented in a DB as aggregation and another not.

Another fundamental problem in the integration process is semantic heterogeneity. Its resolution is essential to provide interoperability between multiple sources. Different conceptualizations and different DB schemes are typically used to represent such replicated data. According to Dong and Srivastava (2013), some issues must be resolved to manage semantic heterogeneity, among them:

- Provide an integrated view of the overlapping data sets of multiple DB;
- Provide support on updating against an integrated view;
- Identify and specify the relationships between two or more instantiations of replicated data; and
- Keep replicated data in sync.

The situation is entirely different in the Big Data environment, as traditional integration approaches prove to be inefficient as the problem is handled (DONG; SRIVASTAVA, 2013).

Big Data basically derives from a large or huge data volume, defined as a situation in which the volume, speed, and variety of data exceed the storage or computing capacity for making accurate and timely decisions. The storage of these large volumes of data can be done by introducing data centers, where data manipulation is complex and must be considered carefully (CUZZOCREA; SONG; DAVIS, 2011; LI *et al.*, 2012).

According to Cuzzocrea, Song and Davis (2011), Big Data refers to huge amounts of unstructured data produced by high-performance applications that fall into a wide and heterogeneous family of application scenarios, such as computing, scientific applications on social networks, e-government applications, medical information systems, among others. The data stored in the underlying layer of all these application scenarios have some specific common characteristics, including:

I. Large-scale data, which refers to the size and distribution of the data repositories;
II. Scalability issues, which refer to the ability of applications to run large-scale data repositories on a large scale, to scale over the rapid growth in size and data inputs;
III. Advanced support for the low-level ETL process, raw data with little structured information;

IV. Easy and interpretable conception and development of the science of analysis on large volume repositories to obtain intelligence and extract valuable knowledge.

According to Dong and Srivastava (2013), integration in Big Data differs from traditional data integration, which includes virtual integration and materialized, several dimensions storage:

- Volume: not only each data source contains a huge volume of data, but also the number of data sources, even for a single domain, grows to tens of thousands;
- Speed: as a direct consequence of the rate at which data is being collected and continuously made available, many of the data sources are very dynamic;
- Variety: Data sources (even in the same domain) are highly heterogeneous, both at the schema level in relation to how they structure their data and at the instance level in terms of how they describe the same real-world entity, exhibiting considerable variety, even for substantially similar entities;
- Veracity: data sources (even in the same domain) have very different qualities, with significant differences in the coverage, accuracy, and freshness of data provided.

For a variety of scheme levels, the work by Li *et al.* (2012) shows that many attributes from different sources in the stock market, totaling up 333 attributes, among them attributes with the same semantics, but with other names, that were manually matched resulting in 153 attributes, called global attributes. These attributes are attributed to the Zipf Law distribution, which says that only a tiny portion of attributes has high coverage, and most have low coverage. In fact, 21 attributes (13.7%) are provided by at least a third of the sources, and more than 86% of attributes are defined by less than 25% of the sources.

Guo *et al.* (2010) show an experiment with business data on a single website that aggregated data obtained from other sources for the variety at the instance level. It considers only the name, telephone, and address attributes for two zip codes to present a solution for linking records in the presence of uniqueness restrictions and wrong values using a linking and merging process applying them globally.

Thus, to integrate different data sources, it is necessary to solve the problems of heterogeneity, whose solution lies in the search for equality between various data sources. This work aims to present a systematic review of the subject's literature to show structural and semantic heterogeneity in the process of integrating different data sources in Big Data.

Materials and methods

The work reviews the methods used in data integration in Big Data, considering the heterogeneity of data, analyzing techniques that use the Semantic Web concepts, Cloud Computing, data analysis, Big Data, DW, and other technologies covering the period from 2011 to 2020.

The bibliographical research was carried out in English and Portuguese. The vast majority of publications and conferences related to this subject are in English, so this language was included in search terms. However, as the need to research content on national scientific production is considered, Portuguese was included.

The articles reviewed in this research were taken from the following digital libraries:

- IEEE publications – IEEExplore;
- Science Direct;
- ACM Digital Library;
- Google Scholar;
- Scopus.

Research questions

Q1. What initiatives were presented for the problem of structural heterogeneity in the process of integrating different data sources in Big Data?

Q2. What initiatives were presented for the problem of semantic heterogeneity in the process of integrating different data sources in Big Data?

Terms and synonyms used in this research

Terms and synonyms used in this work are shown in Table 1.

Search expressions

Search expressions proposed for this works' systematic review are presented in Table 2. They were subdivided into four search expressions to obtain the most significant possible number of articles for analysis.

Table 1 – Terms and synonyms used in this work

Big Data	Structural technical	Analytics
Data Integration	Algorithms	Autonomous sources
Heterogeneity	Data Warehousing	Complex
Semantics	OLAP	Evolving associations

Source: The authors.

Table 2 – Search expressions used in the systematic review

Search expression	Search words
1	(“big data”) AND (“data integration” OR “integração de dados”) AND (“heterogeneity” OR “heterogeneidade”) AND (“semantics” OR “semântica”) OR (“algorithms” OR “algoritmos”) OR (“analytics” OR “ciência da análise” OR “ciência dos dados”)
2	(“big data”) AND (“data integration” OR “integração de dados”) AND (“heterogeneity” OR “heterogeneidade”) AND (“structural” OR “estrutural” OR “tecnical” OR “técnica”) OR (“analytics” OR “ciência da análise” OR “ciência dos dados”)
3	(“big data”) AND (“data integration” OR “integração de dados”) AND (“autonomous sources” OR “fontes autônomas”) AND (“complex” OR “complexa”) OR (“evolving associations” OR “associações”)
4	(“big data”) AND (“data integration” OR “integração de dados”) AND (“heterogeneity” OR “heterogeneidade”) OR (“structural” OR “estrutural” OR “tecnical” OR “técnica”) OR (“semantics” OR “semântica”) OR (“analytics” OR “ciência da análise” OR “ciência dos dados”) AND (“data warehousing”) OR (“algorithms” OR “algoritmos”) OR (“OLAP”)

Source: The authors.

In Table 2, search expression 1, was created to select articles that involved data integration, heterogeneity, and semantics, focusing on algorithms and data science. Search expression 2, was created to select articles that involved data integration, heterogeneity, and focusing on structural conflicts and data science. Search expression 3, was created to select articles that involved data integration and that additionally focused on autonomous sources. Search expression 4, was created to select articles involving data integration, semantic and structural heterogeneity, and that also focused on data warehousing.

Search words were run on respective search engines with adaptations for correct document retrieval. Search engines used only advanced search mode descriptors.

Document selection and tool used

The Mendeley Desktop tool was the reference manager used to manipulate publications retrieved by search engines. Mendeley identified repetitions, retrieved complete documents, and provided a field for notes on articles.

Therefore, sources for systematic review were selected and evaluated according to the inclusion and exclusion criteria, dividing the research into three stages. In the first stage, articles were searched in the digital libraries mentioned above and selected based on their title and keywords. In the second stage, articles went through a second filter based on their summary, and in addition, duplicate articles were removed. In the third and last stage, the works' introduction and conclusion were analyzed to select articles for systematic review.

Search results in digital libraries

Table 3 shows the number of references retrieved according to the search engine used.

Table 3 – Items obtained in search engines results

Search engines	Items returned
IEEE	203
ACM Digital Library	204
Science Direct	41
Google Scholar	1120
Scopus	49

Source: The authors.

As seen in Table 4, after reading and analyzing the relevance of titles, keywords, and abstract to the proposal of this literature review, 1583 were discarded.

Table 4 – Items discarded from search engines results

Search engines	Items discarded
IEEE	187
ACM Digital Library	197
Science Direct	34
Google Scholar	1117
Scopus	48

Source: The authors.

As seen in Table 5, after the deletion process, 34 articles remained.

Table 5 – Selected articles from search engines results

Search engines	Selected articles
IEEE	16
ACM Digital Library	07
Science Direct	07
Google Scholar	03
Scopus	01

Source: The authors.

Proposed criteria for article analysis

Articles were classified to the following criteria:

- Big Data: informs if the term Big Data is used;
- Business Intelligence (BI): informs if the BI is used;
- Data integration: tells if the term data integration is used;
- Cloud computing: reports if the term cloud computing is used;
- Semantic heterogeneity: informs if the term semantic heterogeneity is used;
- Ontology: Informs if the term ontology is used.

Based on these criteria, articles were then analyzed considering the following subsections:

- Resolution of semantic heterogeneity through data analysis;
- Resolution of heterogeneity in Big Data;
- Resolution of heterogeneity with a focus on BI;
- Resolution of heterogeneity using DW and Middleware;
- Resolution of heterogeneity using ontologies and Web Semantic technologies.

Selected articles' analysis

In this section, according to the criteria presented above, the articles will be grouped and analyzed considering the proposals presented for the resolution of structural and semantic heterogeneity.

Resolution of heterogeneity using ontologies and technologies of the Web Semantics

The work by Nugraheni, Akbar and Saptawati (2016) proposes a framework for developing a semantic DW that can handle incomplete data and data heterogeneity problems (format, syntax, and structure) through the use of ontologies. The framework also provides tools for dealing with an incomplete data source. The framework provides tools to transform instances of objects relevant to a class of external ontologies that generate the required data. A hybrid methodology identifies the multidimensional elements and the conceptual design of the multidimensional scheme.

The approach created by Fathy, Gad and Badr (2019) aims to translate queries sent in SPARQL Protocol and RDF Query Language (SPARQL) to queries in NoSQL graphs. In the first step, an extension of RDB-to-RDF mapping language (xR2RML) mapping file that includes the triple maps of the semantic model in question for the NoSQL graph is used as an input, creating an intermediate representation between SPARQL and NoSQL graph queries. Then, a SPARQL query is sent, which will be translated into a NoSQL graph query based on the mapping created in the first step so that the generated query can then be executed in the NoSQL graph DB.

Nadal *et al*. (2019) developed a query rewrite algorithm. The user searches on a global ontology (Global Graph) created based on the data sources in their project. The algorithm then rewrites the query to perform the search on the original data source. Each data source has a local ontology (Source Graph) with a wrapper responsible for performing queries on the sources. Each of these wrappers is related to a fragment of the global ontology. A semi-automatic algorithm was also created that updates the global ontology as the data source is changed.

In Ostrowski *et al*. (2016), a risk management system in the supply chain in an automotive sector is presented with the objective of integrating data. It aims to combine data from automakers and suppliers of an automotive factory with meteorological data. It has used the Allegro Graph software that works with geospatial ontologies already on the Web, allowing the assignment of latitude and longitude to its suppliers and assemblers based on the street address. Data on a climate ontology was also added to that of the automakers/suppliers. This final ontology can predict weather events that can delay sales and manufacture parts consulted by SPARQL queries. For his future work, the author identifies several challenges that need to be solved. Problems such as incompatible data support real-time data flow, and incomplete data are some of the challenges.

Ninmagadda and Dreher (2013) reports how the oil ontology describes the semantics of all the sources of data conceptually analyzed. It does not depend on keywords or similarity metrics. The conceptual structure of oil ontology promotes the reuse of concepts and algebraic operators' reuse to consult instances of oil ontology. This refinement is based on ontology and structuring of multidimensional data adapts to the DW. The data integration process facilitates metadata models derived for sedimentary basins in Indonesia and is helpful for data mining and subsequent data for interpretation, including mapping geological knowledge.

Keller *et al*. (2016) describe a system implemented to combine heterogeneous data from air traffic management systems using semantic integration techniques. The system transforms the different source data formats into a unified semantic representation using ontologies, linked data, semantic integration techniques, and analyzing a subset of government flights, weather, and part of the United States air traffic management system. These systems track planes as they follow their flight path and maintain data on aircraft routing and weather conditions that can affect air traffic.

Considerations on the resolution of semantic heterogeneity through the use of ontologies technologies of Web Semantics

Ontology deals with the nature of reality, exploring similarities, differences, and relationships between existing data types. It deals with content related to the types and relationships between objects used in a given domain of knowledge, as they provide terms to express a body of knowledge about a domain (MCDANIEL; STOREY, 2019). Thus, the discussed articles used ontologies to integrate the different data sources. Three methods can perform this integration: single ontology method, multiple ontologies, and hybrid (ALKHAMISI; SALEH, 2020).

In the approach of a single ontology, the source schemes are made available through a shared global ontology, which offers a uniform interface for each user.

In the approach of multiple ontologies as local sources, they are mapped to local ontologies, and the connection between them is semantically established. Finally, the hybrid ontology approach explores the combination of the first two. In this approach, the sources are described by local ontologies and it is performs the mapping with a shared global ontology. Another important point is about which architecture the proposal of each article was based on, for example, DW, Middleware, or other technology to access data sources with the purpose to integrate and make information available, and finally, additional comments are presented to the submitted proposal. Table 6 compares the semantic integration approaches based on Ontologies and Technologies of the Web Semantics.

It can be mentioned as limitations presented in the work by Fathy, Gad and Badr (2019), the data sources that define the same domain are often incomplete and may be incompatible with each other, making the mapping need adjustments. For this problem Nadal *et al*. (2019) created a semi-automatic algorithm for updating the ontology; if a data source is updated, the algorithm updates the ontology. The work that is still needed is to improve the semantics of the data sets, even though some industrial ontologies already exist.

Based on the analysis of the articles, possible research directions are suggested:

- Automation the comparison of ontologies in real-time.
- Creation of more complete industrial ontologies.

Resolution of semantic heterogeneity through data analysis

Pelekis, Theodoridis and Janssens (2014) present a unified structure for management and analysis of data objects that include both trajectories and their semantic counterpart. Solutions are provided to develop semantic knowledge systems in the real world of mobile object databases and trajectory data storage systems. The respective query processing algorithms have been created.

The work by Assaf *et al*. (2012) presents a framework that allows to combine data from different data sources in a semi-automatic way. The method adopted by RUBIX explores linked data to improve the process of schema equality through statistical algorithms and vector algebra. It uses a set of Google Refine Server extensions and a plug-in for the user interface.

Gao and Xiao (2013) proposes a data integration model based on an extension of Open Grid Service Architecture-Data Access and Integration (OGSA-DAI) to provide uniform access to data while protecting the structural and semantic heterogeneity of the data. It uses the grid service to preserve the schema mapping, thus obtaining transparent access to heterogeneous distributed data and the associated query.

Madkour, Aref and Basalamah (2013) presents the proposal of knowledge cubes as a semantically guided data management architecture, where data management is influenced by data semantics and not by a predefined scheme. Knowledge cubes use Resource Description Framework (RDF) to store data, which allows keeping linked data from the Web. The cube is smart at identifying when to update its data based on the query predicates and the frequency of requested items. Knowledge cubes support the five pillars of Big Data, also known as the five V's: Volume, Speed, Veracity, Variety, and Value.

Cuzzocrea *et al*. (2014) describe a composite methodology that combines techniques based on semantics and multidimensional analysis paradigms to improve the knowledge discovery phase. It effectively and efficiently supports the discovery of collaboration processes from log data from heterogeneous data sources, creating a multidimensional taxonomy to analyze such logs according to a high level of abstraction.

The approach created by Saes (2019) uses Artificial Intelligence technologies to automatically integrate large amounts of data from several sources, whether structured or unstructured. The proposed framework aims to evaluate the data based on its metadata to verify the similarity between data and its possible level of integration. Its structure is flexible, as it uses integration modules, easing maintenance, implementation, and integration of new data models if necessary.

Considerations on the resolution of semantic heterogeneity through data analysis

The discussed works present a process of integrating data sources seeking to address structural and semantic heterogeneity using layered frameworks or architectures, presenting a unified view for data queries. The works by Assaf *et al*. (2012), Gao, Xiao (2013) and Saes (2019) present a semi-automatic integration process while the works by Pelekis, Theodoridis and Janssens (2014) and Madkour, Aref and Basalamah (2013) automatically and respectively present solutions for these through the use of ontologies, ETL of semantic data, semantic cubes and frameworks.

Table 6 – Comparison of semantic integration approaches based on Ontologies and Technologies of the Web Semantics

Article	Approach	Structure	Comments
NUGRAHENI; AKBAR; SAPTAWATI, 2016	Hybrid ontology.	Framework for development semantic DW using ontologies.	Framework to builds semantic DW that integrates multidimensional data and ontology in generated elements of the dimensions.
FATHY; GAD; BADR, 2019	Shared global ontology.	Create a mapping so that the search queries can be translated and used in different sources.	Ontology-Based Data Access (OBDA) method and query mapping are done through the local-as-view approach.
NADAL *et al.*, 2019	Shared global ontology.	Create a mapping so that the search queries can be translated and used in different sources.	OBDA method and query mapping are done through the local-as-view approach.
OSTROWSKI *et al.*, 2016	Hybrid ontology.	Use Semantic Web Technologies as a means of integration and development of Big Data applications.	Support federated ontologies in Allegro Graph software, and SPARQL query.
NINMAGADDA; DREHER, 2013	Shared global ontology.	Integration of approaches for building semantic, syntactic, and schematic relationships among multidimensional and heterogeneous data sources	BI ontology-based multidimensional data warehousing and mining technologies.
KELLER, *et al.*, 2016	Shared global ontology.	An ontological model using a data translation system from the source to the ontological triple.	Semantic integration.

Source: The authors.

The main differences between the works by Madkour, Aref and Basalamah (2013) and Pelekis, Theodoridis and Janssens (2014), is that the first uses a location-based semantic system while the second forms a semantic cube by subject through linked data, which can be by topics (Sports), contextual (University Library), spatial (country), temporal (the 1950s), or a combination of them. Assaf *et al.* (2012) work in the same way that Madkour, Aref and Basalamah (2013) uses linked data. The works by Gao, Xiao (2013) and Saes (2019) adopt a middleware architecture, the first proposing an architecture for data management based on Grid, which differentiates it from the others. Table 7 compares the semantic integration approaches based on data analysis.

Based on the analysis of the articles, possible research directions are suggested:

- Automation of the process of locating instances of objects with semantic conflicts through the use of ontologies;
- Creation of an automatic process for capturing data from the Semantic Mobility Network for storage in the Semantic Mobility Database;
- Use of linked data to integrate data sources with the use of instance-based ontologies;
- Application of Composite Methodology for Supporting Collaboration Pattern Discovery techniques via Semantic Enrichment in Cloud Computing and Big Data.
- Creation of more complete industrial ontologies.

Resolution of heterogeneity in Big Data

Deb Nath, Hose and Pedersen (2015) describes a programmable semantic ETL framework to process and integrate data semantically by crossing Web Semantics and DW technologies to overcome the limitations of traditional ETL tools. The Web Semantics Technology was introduced to convert Web documents to Web data in RDF format, so it could be in a machine-readable format.

Table 7 – Comparison of the integration approaches based on data analysis.

Article	Approach	Structure	Application
PELEKIS; THEODORIDIS; JANSSENS, 2014	Framework for semantic trajectory modeling.	Semantic Mobility Database (SMD) and Semantic Mobility Cubes.	Location-based social networks.
ASSAF *et al.*, 2012	Framework for semi-automatic data integration.	Two layers: Google Refine Server and the modular web application called Butterfly.	Integration of business analysis systems and external data sources through linked data.
GAO; XIAO, 2013	Shared global view.	OGSA-DAI (Middleware) with the data layer, business logic, presentation layer, and Client layer.	Resolution of structural and semantic heterogeneity using a shared global view.
CUZZOCREA *et al.*, 2014	Semantics-based techniques and multi-dimensional analysis paradigms.	Framework for creating a common abstract model for knowledge extraction.	Integration of log-based schemas.
MADKOUR; AREF; BASALAMAH, 2013	Middleware architecture.	Semantic cube created through linked data.	Integration of semi-structured data sources through RDF regardless of data sources.
SAES, 2019	Federated Databases.	Integration architecture with Artificial Intelligence.	Integration of semi-structured and structured data sources.

Source: The authors.

Therefore, RDF Schema (RDFS) and Web Ontology Language (OWL) were used. This data is integrated with each other semantically. Thus, it supports data sources with semantic knowledge, semantic integration, and the creation of a semantic DW, composed of an ontology and its instances.

Bortoli *et al*. (2016) show a use case. The combination of semantics and Big Data technologies are used to define a semantic ETL to effectively and efficiently support infraction activities in tackling consumption tax evasion in the Valle d'Aosta region. A domain ontology, Open Refine, and the Okkam Entity Name System are used for the scalable data integration process, leading to a knowledge base for tax assessment. In addition, the concept of Entiton was presented as flexible and efficient, suitable for large-scale data inferences and analytical tasks.

Chen *et al*. (2016), an ETL framework for Big Data integration called Big Data ETL (BDETL), is presented to integrate a vast amount of heterogeneous data from several different sources in a dispatch control system for electrical networks. The data extraction process consists of extracting data from sources and uploading it to the Hadoop Distributed File System (HDFS). Data is then extracted from HDFS according to rules predefined by the user and executed by map and reduce services during the transformation process that will remove duplicate data, normalize it, check its integrity and consistency, and other processes. During the transformation process, the BDETL task manager sends tasks to the Hadoop Job-Tracker in sequential order. Then the data is transformed according to the rules defined by a uniform data model. The loading process stores the data in a Hive DW that uses HDFS for data storage but presents data from HDFS files as Dispatching and Control System logic tables.

Mountasser, Storey and Frikh (2015) proposes a multi-layered prototype for Big Data management processes based on semantic integration and dynamic extraction of knowledge in a large-scale environment. The efficient use of semantic data integration can reduce storage dimension and improve performance analysis, offering adequate flexibility and extensibility for real-time data analysis. In addition, the use of semantics can enhance the quality of data and guarantee credibility, which in turn improves research and knowledge discovery.

Considerations on the resolution of semantic heterogeneity through Big Data

In the work by Deb Nath, Hose and Pedersen (2015), the data sources can be structured, semi-structured, RDF and SPARQL. All ETL phases make use of the ontology definition, and the database generated can be consulted by SPARQL. For Mountasser, Storey and Frikh (2015), the unstructured sources, geographic location, and log files are also considered. In the work by Bortoli *et al*. (2016), the semantic ETL uses Open Refine and its storage in the HDFS system.

Using the Hadoop ecosystem, Chen *et al*. (2016) present a proposal to integrate heterogeneous data sources to the DCS environment, use parallel computing power together with the MapReduce algorithm, and use a uniform data model specific to DCS.

As can be seen in Big Data, according to the works presented, the use of Ontologies and the semantic ETL process are fundamental for making data available to users. Table 8 compares the semantic integration approaches based in Big Data.

Resolution of heterogeneity focused on business intelligence

Aufaure *et al*. (2016) presents recent work on BI in real-time combined with the management of semantic data flows. It also offers underlying approaches, such as continuous queries, summary and data matching, and reasoning about data flow.

Bondarev and Zakirov (2015) shows the technologies involved in building and using a DW for data integration. In addition to the process of integrating structured data, it made use of the Hadoop platform to integrate unstructured and semi-structured data.

Ghosh, Halder and Sen (2015) present a methodology for developing DW in a BI environment. In this methodology, analytical processing has a significant role in business analysis. It is an integral part of BI in business applications through DW, which is the methodology used to design and consult Online Analytical Processing (OLAP). The analytical environment incorporates DW, Data Mining, Data Mart, and Virtual DW. The environment architecture consists of modules with specific integrated and interconnected modules to perform analytical processing-based BI.

Shafiee, Barker and Rasekh (2018) aims to automate and improve a hydraulic data system, which receives information from the weather forecast database and data from various sensors. In their approach, the authors use parallel computing for data to be processed and stored. A data lake called "Water Data Lake" will store the data throughout the process. Hadoop-based technologies were used in its creation. The data analysis process, responsible for cleaning data, filling in necessary values, and filtering out irrelevant information, was automated by the authors. The analysis process is also done in a distributed way using the Apache Spark software, which is responsible for distributing the data transformation tasks in a computational cluster. After being processed, the data is then again stored in the "Water Data Lake". A Middleware layer performs search queries on the data lake as soon as newly processed data is available. This step is responsible for standardizing and validating the data and then sending it to the wrappers responsible for receiving data streaming, modeling it, and sending the results back to the data lake. The results can then be used for analysis and decision-making.

The purpose of the article by Bala, Boussaid and Alimazighi (2014) is to work on the impact of Big Data in a decision support environment and, more particularly, in the data integration phase. For this, a platform called Parallel Extract Transform and Load (P-ETL) was developed to extract, transform and load vast amounts of data in a DW.

ABBAS *et al*. (2015) offers a cloud-based framework that offers personalized recommendations on health plans, using Multi-attribute Utility Theory (MAUT) to help users compare different health plans. Health insurance is based on coverage and cost criteria. The plan information for each of the providers is retrieved using Data as a Service (DaaS). Software as a Service (SaaS) is implemented to offer personalized recommendations, applying a classification technique to the plans identified according to user-specified criteria.

Considerations on the resolution of semantic heterogeneity through business intelligence

It is observed that the discussed works adopt different approaches, as shown in Table 9. It shows the different choices for building a database to support decision-making. The works Bondarev and Zakirov (2015), Ghosh, Halder and Sen (2015) and Aufaure *et al*. (2016) use BI architecture, the last being in real-time. A limitation presented by the work by Bala, Boussaid and Alimazighi (2014) due to the parallelism of the ETL process is the response time that depends on the number of tasks and to improve the performance of the system it

Table 8 – Comparison of the integration approaches based in Big Data.

Article	Approach	Structure	Application
DEB NATH; HOSE; PEDERSEN, 2015	Shared global ontology.	Framework for ETL implemented in Python.	Semantic DW.
BORTOLI *et al.*, 2016	Shared global ontology.	Semantic technologies with Big Data tools.	Semantic ETL.
CHEN *et al.*, 2016	Model of heterogeneous data integration for Big Data, using a global ontology for DCS.	A Big Data ETL architecture.	Integration of heterogeneous data for Big Data.
MOUNTASSER *et al.*, 2015	Shared global view.	Multi-Layered Knowledge Extraction Prototype.	Semantic data integration.

Source: The authors.

is suggested the addition of new nodes and task numbers. The work by Abbas *et al*. (2015) differs because it uses the cloud-based structure to provide access to information.

According to Choi, Chan and Yue (2016), it is observed that there was an evolution of BI systems for data collection in real-time with, for example, airline companies, automotive industries, due to the use of technology such as Radio-frequency iDentIFication (RDIF). Internet of Things (IoT) also generates data in real-time and can be applied by food supply chains, stock management, transportation, among others. However, they are subject to an intrinsic restriction created by the sensors' heterogeneous nature. With the design of a specific system to address heterogeneity, IoT can be integrated with BI systems.

Resolution of heterogeneity using data Warehouses and Middleware

Le-Phuoc *et al*. (2016) provides an integrated unified view to query and explore heterogeneous data connected to the IoT in real-time using linked data, called the Graph of Things (GoT). GoT is supported by a scalable and resilient software stack to handle billions of historical and static data records. Millions of data instances are being fetched and enriched to connect to GoT in real-time.

Alqarni and Pardede (2012) proposes a multi-layer scheme to map structured data stored in a DW and unstructured data in business-related documents. Linguistically correlated data is identified using WordNet to allow integration between both data sources. It uses an Extensible Markup Language (XML) schema for unstructured documents to assist the mapping process.

Faeldon, España and Sabido (2014) proposes to improve the integration process by optimizing the workflow used in numerical weather forecasting from the perspective of how data is acquired, processed, and analyzed in the high-performance Computing system.

Malviya, Udhani and Soni (2016) uses a framework based on the R language to analyze Big Data in cloud computing. It analyzes structured and unstructured data from social networking sites. Analyses can be performed on different parameters. The focus of the article is to analyze data using R, which is a language for modeling data and statistics.

Komamizu, Amagasa and Kitagawa (2015) proposes a framework called SPARQL Olap Over Linked Data (SPOOL), which works to reduce the effort to access the most recent data from the numerical records of the SPOOL data set through SPARQL without downloading the entire data record. SPOOL provides a series of SPARQL queries that extract objects and attributes from linked data datasets and convert them into star/snowflake schema and materialize relevant triples such as fact tables and dimensions for OLAP.

Yahya *et al*. (2019), the old Malaysian lake data system (MyLake), describes how companies need to send data manually to the government, concluding that the old system has its limitations, among them, the fact that the information is in a data silo and is not integrated with each other. If any agency needs the data of another, it has to integrate itself according to its needs. This process is highly laborious and manual. The solution proposed is to create a unified system that integrates the data received from agencies, not integrating data every time information is needed.

Table 9 – Comparison of the integration approaches based on Business Intelligence.

Article	Approach	Structure	Application
AUFAURE *et al.*, 2016	Real-time BI using ontologies.	Real-time BI architecture with generic semantic integration.	Real-time data integration using semantic technologies.
BONDAREV; ZAKIROV, 2015	Integration of unstructured data on DW technology using Hadoop.	BI Architecture.	Integration of structured and unstructured data using Hadoop.
GHOSH; HALDER; SEN, 2015	Approach to DW development in BI environments.	BI Architecture.	Traditional DW construction methodology for BI environments.
SHAFIEE; BARKER; RASEKH, 2018	Middleware.	Framework with Wrapper, Middleware, and application database.	Decision support.
BALA; BOUSSAID; ALIMAZIGHI, 2014	Uses the Hadoop framework with the MapReduce paradigm to implement the ETL process in the data integration phase.	P-ETL architecture is organized into five modules: Extraction, Partitioning, Transformation, Reduction, and Load.	Provide DW load time speed for timely delivery for decision making.
ABBAS *et al.*, 2015	Shared global ontology.	Cloud-based system architecture.	Cloud-based health insurance recommendation system.

Source: The authors.

This information, once worked, is sent back to the agencies. To achieve its objectives, the authors propose creating a middleware to mediate data interaction with the application. The Middleware will act as an integration layer, which will be like a data exchange center. The agencies would send the data to the Middleware, which integrates the data that can be accessed through a user-friendly portal.

There are multiple issues yet to be worked on in DW that is associated with Big Data, which are: OLAP on Big Data, Big Data posting, and Big Data Privacy (CUZZOCREA; SACCÀ; ULLMAN, 2013). Other subjects that can generate new research are analysis of business requirements, data analysis, data modeling, data movements, data quality, data transfer, and data presentation (QIN; QIAN; ZHAO, 2015).

Finally, there are challenges in designing a Big Data integration architecture, such as defining the scope of the data, inconsistency of the data, optimization of queries, adequacy of resources for investment in Big Data, scalability, and the ETL process (KADADI *et al.*, 2014).

Considerations on the resolution of semantic heterogeneity through data Warehouses and Middleware

The works by Alqarni and Pardede (2012), and Malviya, Udhani and Soni (2016) consider structured and unstructured sources for integration and proposals for integrating different schemes as shown in Table 10. In the article by Malviya, Udhani and Soni (2016), one of the R Tools disadvantage points the memory management and the security topics and in Alqarni and Pardede (2012) there is the problem of high processing time by considering only Wordnet to detect similarity between data sources.

In the works developed by Komamizu, Amagasa, and Kitagawa (2015) and Faeldon, España and Sabido (2014), the focus is on numerical data, with the different applications shown in Table 10. The article by Komamizu, Amagasa and Kitagawa (2015) presents the problem of handling the missing data and the lack of definition of data types in the sources used. The article by Faeldon, España and Sabido (2014) is limited by the amount of data collected in real time for scientific and engineering applications, thus, it needs to improve the performance of managing data collected in real-time in scientific and engineering applications.

Table 10 – Comparison of the semantic integration approaches based on Data Warehouses and Middleware.

Article	Approach	Structure	Application
LE-PHUOC *et al.*, 2016	Linked Stream Middleware.	Layered architecture using ontologies and SPARQL engine.	Integration of the heterogeneous data connected in the IoT.
ALQARNI; PARDEDE, 2012	Uses XML Schema as a logical model to unify data stored in the DW with unstructured documents.	Architecture with multi-layer schemes for mapping DW data to XML Schema and an XML Schema document generator for unstructured documents.	Methodology for integrating DW with unstructured documents.
FAELDON; ESPAÑA; SABIDO, 2014	Integration of data analysis solutions in an High Performance Computing (HPC) workflow.	Scheme of a data-centric HPC for climate modeling.	HPC application workflow.
MALVIYA; UDHANI; SONI, 2016	Analysis of structured and unstructured data from social networking sites with the R tool.	R-Tools architecture.	Data analysis for decision making.
KOMAMIZU; AMAGASA; KITAGAWA, 2015	DW creation of linked data for OLAP queries. The Type-Partitioned Triple Store (TPTS) methodology was proposed.	Framework SPOOL.	Linked data containing numeric values extracted using SPARQL to create OLAP cubes.
YAHYA *et al.*, 2019	Middleware.	Centralized integration platform.	Integration of different repositories through middleware.

Source: The authors.

Conclusion

This research aimed to analyze the problem of structural and semantic heterogeneity in the process of integrating different data sources in Big Data, as well as proposed solutions, based on the literature found from 2011 to 2020.

We sought to answer the research question: What initiatives were presented for the problem of structural and semantic heterogeneity in integrating different data sources in Big Data?

The sources used for this research were: IEEE, Science Direct, ACM Digital Library, Google Scholar, and Scopus, and the Mendeley Desktop tool was the reference manager used to manipulate the publications retrieved by the search engines. With the Mendeley Desktop tool, it was possible to identify repetitions and recover complete documents.

In conducting the systematic review, sources were selected and evaluated according to the inclusion and exclusion criteria defined in the review protocol, and 34 articles were selected and analyzed by this research. They were classified according to the aggregating attributes: Big Data, BI, Data integration, Cloud computing, Semantic heterogeneity, and Ontologies.

Based on these aggregating attributes, the articles were distributed on the following topics: Data Analysis; Big data; BI; DW and Middleware; and Ontologies and technologies of the Web Semantic.

In this systematic review, several data integration techniques could be analyzed. Some use ETL methods to extract data from its sources, treat and homogenize that data, and store it. In contrast, others create global ontologies from databases and then use mapping to translate queries sent to these ontologies to execute at the source of origin. In the scope of semantic heterogeneity, semantic ETL is noted to perform data extraction and treatment. When addressing structural heterogeneity content and the use of ontologies in the solution of integrations free of dependence on keywords or similarity metrics, techniques such as cloud computing processing and multi-layer data prototypes are used in the processing and management of Big Data.

Some obstacles in Big Data still need to be overcome. Data sources can be incomplete and incompatible and, although reduced, there is still a need for manual work. However, as seen in this review, several researchers are managing to overcome some of these challenges using Web Semantics, parallel computing, and AI concepts.

As expected, each article has its line of study. However, it is possible to note that even studies with different characteristics use similar solutions, which mainly involve the use of semantic ETL to treat the heterogeneity of the data, before it can be stored in the DW. It is also observed that it uses semantics to improve the quality and reliability of different data objects.

This research revealed the growing use of algorithms in query processing between different data sources and logs from heterogeneous data sources. In addition, growing interest in the capacity for privacy, reliability, and OLAP in large volumes of information was noted.

Future research may involve the points mentioned above and a need to develop frameworks and semantic cubes to create a unified structure to combine and manage data.

Acknowledgments

To the State University of Londrina, on supporting this research.

References

ABBAS, A.; BILIAL, K.; ZHANG, L.; KHAN, S. U. A cloud based health insurance plan recommendation system: A user centered approach. *Future Generation Computer Systems*, [London], v. 43, p. 99-109, 2015.

ALKHAMISI, A. O.; SALEH, M. Ontology opportunities and challenges: discussions from semantic data integration perspectives. *In*: CONFERENCE ON DATA SCIENCE AND MACHINE LEARNING APPLICATIONS (CDMA). 6., 2020, Riyadh. *Proceedings* [...]. Riyadh: IEEE, 2020. p. 134-140.

ALQARNI, A. A.; PARDEDE, E. Integration of data warehouse and unstructured business documents. *In*: INTERNATIONAL CONFERENCE ON NETWORK-BASED INFORMATION SYSTEMS, 15., 2012, Melbourne. *Proceedings* [...]. Melbourne: IEEE, 2012. p. 32-37.

ARPUTHAMARY, B.; AROCKIAM, L. Data Integration in big data environment. *Bonfring International Journal of Data Mining*, Tamilnadu, v. 5, n. 1, p. 1-5, 2015.

ASSAF, A.; LOUW, E.; SENART, A.; FOLLENFANT, C.; TRONCY, R.; TRASTOUR, D. RUBIX: a framework for improving data integration with linked data. *In*: WOD 12: INTERNATIONAL WORKSHOP ON OPEN DATA, 1., 2012, Nantes. *Proceedings* [...]. New York: Association for Computing Machinery, 2012. p. 13-21.

AUFAURE, M.-A.; CHIKY, R.; CURÉ, O.; KHROUF, H.; KEPEKLIAN, G. From business intelligence to semantic data stream management. *Future Generation Computer Systems*, London, v. 63, p. 100-107, 2016.

BALA, M. BOUSSAID, O.; ALIMAZIGHI, Z. P-ETL: Parallel-ETL based on the MapReduce paradigm. *In*: INTERNATIONAL CONFERENCE ON COMPUTER SYSTEMS AND APPLICATIONS (AICCSA), 11., 2014, Doha. *Proceedigns* [...]. Doha: IEEE, 2014. p. 42-49.

BONDAREV, A.; ZAKIROV, D. Data warehouse on Hadoop platform for decision support systems in education. *In*: INTERNATIONAL CONFERENCE ON ELECTRONICS COMPUTER AND COMPUTATION (ICECCO), 12., 2015, Almaty. *Proceedings* [...].Almaty: Suleyman Demirel University, 2015. p. 1-4.

BORTOLI, S.; BOUQUET, P.; POMPERMAIER, F.; MOLINARI, A. Semantic big data for tax assessment. *In*: SBD 16: INTERNATIONAL WORKSHOP ON SEMANTIC BIG DATA, 16., 2016, San Francisco Califórnia. *Proceedings* [...]. New York: Association for Computing Machinery, 2016. p. 1-6.

CHEN, W.; WANG, R.; WU, R.; TANG, L.; FAN, J. Multi-source and heterogeneous data integration model for big data analytics in power DCS. *In*: INTERNATIONAL CONFERENCE ON CYBER-ENABLED DISTRIBUTED COMPUTING AND KNOWLEDGE DISCOVERY (CYBERC), 2016, Chengdu. *Proceedings* [...]. Chengdu: IEEE, 2016. p. 238-242.

CHOI, T.-M.; CHAN, H. K.; YUE, X. Recent development in big data analytics for business operations and risk management. *IEEE transactions on cybernetics*, New York, v. 47, n. 1, p. 81-92, 2016.

CUZZOCREA, A.; DIAMANTINI, C.; GENGA, L.; POTENA, D.; STORTI, E. A composite methodology for supporting collaboration pattern discovery via semantic enrichment and multidimensional analysis. *In*: INTERNATIONAL CONFERENCE OF SOFT COMPUTING AND PATTERN RECOGNITION (SOCPAR), 6., 2014, Tunis. *Proceedings* [...]. Tunis: IEEE, 2014. p. 459-464.

CUZZOCREA, A.; SACCÀ, D.; ULLMAN, J. D. Big data: a research agenda. *In*: IDEAS 13: INTERNATIONAL DATABASE ENGINEERING & APPLICATIONS SYMPOSIUM, 17., 2013, Barcelona. *Proceedings* [...]. New York: Association for Computing Machinery, 2013. p. 198-203.

CUZZOCREA, A.; SONG, I-Y.; DAVIS, K. C. Analytics over large-scale multidimensional data: the big data revolution!. *In*: DOLAP 11: INTERNATIONAL WORKSHOP ON DATA WAREHOUSING AND OLAP, 14., 2011, Glasgow Scotland. *Proceedings* [...]. New York: Association for Computing Machinery, 2011. p. 101-104.

DEB NATH, R. P.; HOSE, K.; PEDERSEN, T. B. Towards a programmable semantic extract-transform-load framework for semantic data warehouses. *In*: PROCEEDINGS OF THE ACM EIGHTEENTH INTERNATIONAL WORKSHOP ON DATA WAREHOUSING AND OLAP, 24., 2015, Melbourne. *Proceedings* [...]. New York: Association for Computing Machinery, 2015. p. 15-24.

DONG, X. L.; SRIVASTAVA, D. Big data integration. *In*: INTERNATIONAL CONFERENCE ON DATA ENGINEERING (ICDE), 29., 2013, Brisbane. *Proceedings* [...].Brisbane: IEEE, 2013. p. 1245-1248.

FAELDON, J.; ESPANA, K.; SABIDO, D. J. Data-centric HPC for numerical weather forecasting. *In*: INTERNATIONAL CONFERENCE ON PARALLEL PROCESSING WORKSHOPS, 43., 2014, Minneapolis. *Proceedings* [...].Minneapolis: IEEE, 2014. p. 79-84.

FATHY, N.; GAD, W.; BADR, N. A Unified Access to Heterogeneous big data through ontology-based semantic integration. *In*: INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTING AND INFORMATION SYSTEMS (ICICIS), 9., 2019, Cairo. *Proceedings* [...]. Cairo: IEEE, 2019. p. 387-392.

GAO, J.; XIAO, J. Research on heterogeneous data access and integration model based on OGSA-DAI. *In*: INTERNATIONAL CONFERENCE ON COMPUTATIONAL AND INFORMATION SCIENCES, 2013, Shiyang. *Proceedings* [...].Shiyang: IEEE, 2013. p. 1690-1693.

GHOSH, R.; HAIDER, S.; SEN, S. An integrated approach to deploy data warehouse in business intelligence environment. *In*: INTERNATIONAL CONFERENCE ON COMPUTER, COMMUNICATION, CONTROL AND INFORMATION TECHNOLOGY (C3IT). 13., 2015, Hooghly. *Proceedings* [...].Hooghly: IEEE, 2015. p. 1-4.

GUO, S.; DONG, X. L.; SRIVASTAVA, D.. Record linkage with uniqueness constraints and erroneous values. *Proceedings of the VLDB Endowment*, [S. l.], v. 3, n. 1/2, p. 417-428, 2010.

KADADI, A; AGRAWAL, R.; NYAMFUL, C.; ATIQ, R. Challenges of data integration and interoperability in big data. *In*: IEEE INTERNATIONAL CONFERENCE ON BIG DATA (BIG DATA), 2014, Washington. *Proceedings* [...].Washington: IEEE, 2014. p. 38-40.

KELLER, R.; RANJAN, S.; WEI, M. Y; ESHOW, M. M. Semantic representation and scale-up of integrated air traffic management data. *In*: SBD 16: INTERNATIONAL WORKSHOP ON SEMANTIC BIG DATA, 16., 2016, San Francisco. *Proceedings* [...]. New York: Association for Computing Machinery , 2016. p. 1-6.

KIMBALL, R.; ROSS, M. The data warehouse toolkit: the complete guide to dimensional modeling. 2nd. ed. New York: John Wiley & Sons, 2011.

KOMAMIZU, T.; AMAGASA, T.; KITAGAWA, H. SPOOL: a SPARQL-based ETL framework for OLAP over linked data. *In*: INTERNATIONAL CONFERENCE ON INFORMATION INTEGRATION AND WEB-BASED APPLICATIONS & SERVICES, 15., 2015, Brussels. *Proceedings* [...]. New York: Association for Computing Machinery, 2015. p. 1-10.

LE-PHUOC, D.; QUOC, H. N. M.; QUOC, H. N.; NHAT, T. T.; HAUSWIRTH, M. The graph of things: a step towards the live knowledge graph of connected things. Journal of Web Semantics, London, v. 37, p. 25-35, 2016.

LI, X., DONG, L.; LYONS, K.; MENG, W.; SRIVASTAVA, D. Truth finding on the deep Web: Is the problem solved?. *Proceedings of the VLDB Endowment*, [S. l.], v. 6, n. 2, p. 97-108, 2012.

MADKOUR, A.; AREF, W. G.; BASALAMAH, S. Knowledge cubes: a proposal for scalable and semantically-guided management of Big Data. *In*: IEEE INTERNATIONAL CONFERENCE ON BIG DATA, 2013, Silicon Valley. *Proceedings* [...].Silicon Valley: IEEE, 2013. p. 1-7.

MALVIYA, Ayushi; UDHANI, Amit; SONI, Suryakant. R-tool: Data analytic framework for big data. *In*: SYMPOSIUM ON COLOSSAL DATA ANALYSIS AND NETWORKING (CDAN), 2016, Indore. *Proceedings* [...]. Indore: IEEE, 2016. p. 1-5.

MCDANIEL, M.; STOREY, V. C. Evaluating domain ontologies: clarification, classification, and challenges. *ACM Computing Surveys*, New York, v. 52, n. 4, p. 1-44, 2019.

MOUNTASSER, I.; OUHBI, B.; FRIKH, B. From data to wisdom: a new multi-layer prototype for Big Data management process. *In*: INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEMS DESIGN AND APPLICATIONS (ISDA), 15., 2015, Marrakech. *Proceedings* [...].Marrakech: IEEE, 2015. p. 104-109.

NADAL, S.; ROMERO, O.; ABELLÓ, A.; VASSILIADIS, P.; VANSUMMEREN, S. An integration-oriented ontology to govern evolution in big data ecosystems. *Information Systems*, Elmsford, v. 79, p. 3-19, 2019.

NIMMAGADDA, S. L.; DREHER, H. V. Big-data integration methodologies for effective management and data mining of petroleum digital ecosystems. *In*: IEEE INTERNATIONAL CONFERENCE ON DIGITAL ECOSYSTEMS AND TECHNOLOGIES (DEST), 7, 2013, Menlo Park. *Proceedings* [...].Menlo Park: IEEE, 2013. p. 148-153.

NUGRAHENI, E.; AKBAR, S.; SAPTAWATI, G. Ayu Putri. Framework of semantic data warehouse for heterogeneous and incomplete data. *In*: IEEE REGION 10 SYMPOSIUM (TENSYMP), 2016, Bali. *Proceedings* [...]. Bali: IEEE, 2016. p. 161-166.

OSTROWSKI, D.; RYCHTYCKYJ, N.; MACNEILLE, P.; KIM, M. integration of big data using semantic web technologies. *In*: IEEE TENTH INTERNATIONAL CONFERENCE ON SEMANTIC COMPUTING (ICSC), 2016, Laguna Hills. *Proceedings* [...].Laguna Hills: IEEE, 2016. p. 382-385.

PELEKIS, N.; THEODORIDIS, Y.; JANSSENS, D. On the management and analysis of our lifesteps. *ACM SIGKDD Explorations Newsletter*, [S. l.], v. 15, n. 1, p. 23-32, 2014.

QIN, H.-F.; QIAN, Z.-M.; ZHAO, Y.-C. On the research of data warehouse in big data. *In*: INTERNATIONAL CONFERENCE ON NETWORK AND INFORMATION SYSTEMS FOR COMPUTERS, 2015, Wuhan. *Proceedings* [...].Wuhan: IEEE, 2015. p. 354-357.

SAES, K. R. *Abordagem para integração automática de dados estruturados e não estruturados em um contexto Big Data*. 2019. Tese (Doutorado) - Universidade de São Paulo, São Paulo, 2019.

SHAFIEE, M. E.; BARKER, Z.; RASEKH, A. Enhancing water system models by integrating big data. *Sustainable cities and Society*, London, v. 37, p. 485-491, 2018.

YAHYA, F.; FAZLI, B. M.; ABDULLAH, M. F.; ZULKIFLI, H. Extending the national lake database of malaysia (mylake) as a central data exchange using big data integration. *In*: INTERNATIONAL CONFERENCE ON DATA SCIENCE AND INFORMATION TECHNOLOGY, 2., 2019, Seoul. *Proceedings* [...]. New York: Association for Computing Machinery, 2019. p. 30-35.

Received: Mar. 15, 2021
Accepted: June 8, 2021
Published: June 30, 2021