Enhancing Big Data Warehousing for Efficient, Integrated and Advanced Analytics

Visionary Paper

Maribel Yasmina Santos[1(✉)], Carlos Costa[1,2], João Galvão[1], Carina Andrade[1], Oscar Pastor [3], and Ana Cristina Marcén[3]

[1] ALGORITMI Research Centre, University of Minho, Guimarães, Portugal
{maribel, carlos.costa, joao.galvao, carina.andrade}@dsi.uminho.pt

[2] Centre for Computer Graphics - CCG, Guimarães, Portugal

[3] Research Center on Software Production Methods (PROS), Universitat Politècnica de València, Valencia, Spain
{opastor, acmarcen}@pros.upv.es

Abstract. The existing capacity to collect, store, process and analyze huge amounts of data that is rapidly generated, i.e., Big Data, is characterized by fast technological developments and by a limited set of conceptual advances that guide researchers and practitioners in the implementation of Big Data systems. New data stores or processing tools frequently appear, proposing new (and usually more efficient) ways to store and query data (like SQL-on-Hadoop). Although very relevant, the lack of common methodological guidelines or practices has motivated the implementation of Big Data systems based on use-case driven approaches. This is also the case for one of the most valuable organizational data assets, the Data Warehouse, which needs to be rethought in the way it is designed, modeled, implemented, managed and monitored. This paper addresses some of the research challenges in Big Data Warehousing systems, proposing a vision that looks into: (i) the integration of new business processes and data sources; (ii) the proper way to achieve this integration; (iii) the management of these complex data systems and the enhancement of their performance; (iv) the automation of some of their analytical capabilities with Complex Event Processing and Machine Learning; and, (v) the flexible and highly customizable visualization of their data, providing an advanced decision-making support environment.

Keywords: Big data warehouse · Data governance · Data profiling · Event processing · Performance

1 Introduction

Current advancements in Information Technologies motivated organizations to look towards increased business value and more efficient ways to perform their daily activities. This is usually achieved with data-driven decision-making processes that are

C. Cappiello and M. Ruiz (Eds.): CAiSE Forum 2019, LNBIP 350, pp. 215–226, 2019.
https://doi.org/10.1007/978-3-030-21297-1_19

based in the collection, storage, processing and analysis of huge amounts of data [1]. Also, it is usually associated to characteristics like volume, velocity, variety, variability, veracity and value [2–4], among others, trying to call our attention to the complexity of this area and the difficulties in the integration of such diverse set of data sources, as well as the multiple technologies needed to handle them. Big Data as a research topic is facing several challenges, from the ambiguity and lack of common approaches to the need of significant organizational changes [5], despite some existing efforts of standardizing constructs and logical components of general Big Data systems (e.g., NIST Big Data Reference Architecture [6]). In particular, the research community is looking into the role of a Data Warehouse (DW) in Big Data environments [7], as this data system is usually based on strict relational data models, costly scalability and, in some cases, inefficient performance, opening several opportunities for emerging theories, methodologies, models, or methods for designing and implementing a Big Data Warehouse (BDW) [8]. This can be seen as a flexible, scalable and highly performant system that uses Big Data techniques and technologies to support mixed and complex analytical workloads (e.g., streaming analysis, ad hoc querying, data visualization, data mining, simulations) in several emerging contexts [8]. Although its relevance for supporting advanced analytical processes, research in this area is yet at an early stage, with increased ambiguity in the constructs that can be used, and lacking common approaches.

With the goal of advancing the state-of-the-art and tackle the lack of conceptual guidelines, some of our previous work [8–10] addressed the proposal of models (representations of logical and technological components, data flows, and data structures), methods (structured practices), and instantiations (with demonstration cases based on prototyping and benchmarking) on how to design and implement BDWs. Although filling a major scientific and technical gap, these works were focused on the BDW itself and on its main architectural components, technologies and design patterns, not considering all the additional components, processes and frameworks that must interact with, feed, support (for data analysis and visualization), manage and evaluate this data asset. For advancing data analytics and enhancing the role of the BDW in organizations, this paper presents an overview of the current challenges and some research directions in Big Data Warehousing (BDWing) systems, looking for a continuous practice that allows:

- The integration of new business processes and data sources (*how this integration should be done to provide an integrated view of the organizational business processes and data?*);
- The proper way to achieve this integration, based on adequate data models (*how existing data models should evolve to seamlessly integrate new data sources avoiding uncoordinated data silos?*);
- The management of these complex data systems and the enhancement of their performance (*how can BDWs be monitored in terms of their evolution - business processes and data - and in terms of their performance?*);
- The automation of some of their analytical capabilities (*how can Complex Event Processing and Machine Learning automate and enhance BDWs with advanced real-time events processing?*); and,

- The flexible and highly customizable visualization of their data, providing an advanced decision-making support environment (*how can visualization tools be extended to allow the development of highly interactive and customized data visualizations?*).

Although some of these challenges and open issues may also be relevant for the traditional DW, the complexity associated with the volume, variety and velocity of the data, the variability in data collection and processing, the veracity of the data sources, the complexity of integrating diverse sets of data, the types of analytical workloads (batch, streaming, and interactive), and the diverse and complex technological landscape, impose addressing them with the specificities and needs of Big Data contexts.

This paper is organized as follows. Section 2 presents some background concepts and related work. Section 3 describes the overall framework for advancing BDW research. Section 4 concludes with some highlights of the presented research topics.

2 Background and Related Work

The concept of Data Warehousing (DWing) has a long history, mainly associated to the need to access, analyze and present data to support fact-based analytics [11]. Its aim is to access multiple records at a time. A DW structure is optimized for processing analytical tasks, such as repeatedly queries, reports, Online Analytical Processing – OLAP, data mining or other data science approaches. In a Big Data context, some challenges arise such as inadequate governance of data, lack of skills, cost of implementing new technologies, and difficulties in addressing a modern solution that can ingest and process the ever-increasing amount or types of data [8]. The concept of BDW has been constrained by the fast technological evolution around Big Data, giving short time for developing and maturing research contributions in this area [3]. The BDW can be implemented using two main strategies: (i) the "lift and shift" strategy [12], amplifying the capabilities of relational DWs with Big Data technologies, such as Hadoop or NoSQL databases, proposing particular solutions for specific use cases that may lead to possible uncoordinated data silos [12]; (ii) the "rip and replace" strategy, in which a traditional DW is completely replaced by Big Data technologies [13]. These non-structured practices and guidelines are not sufficient [8], as practitioners and researchers need well-established approaches or guidelines, based on rigorously evaluated models and methods to design and build BDWs [14].

Some existing works explore implementations of DWs on top of NoSQL databases, such as document-oriented [15], column-oriented [16] and graph databases [17], despite the fact that these databases are mainly oriented towards Online Transaction Processing (OLTP) applications [18]. Other works look into storage and processing technologies, discussing SQL-on-Hadoop systems like Hive [19] and Impala [20], or improving these technologies with the use of new storage and processing mechanisms [21]. Moreover, advancements in analytical and integration mechanisms suitable for BDWs are also available [22–24]. In [25], the authors present a framework for evaluating methodologies to design BDWs, defining a set of criteria like application, agility, ontological approach, paradigm, and logical modeling. The authors also divide the methodologies into classes (e.g., automatic, incremental, and non-relational) and define the characteristics being addressed by the methodologies (e.g., value, variety, and velocity).

Taking this into consideration, we have been proposing several prescriptive BDWing contributions grounded on a "rip and replace" strategy and other relevant general contributions in the area of Big Data (e.g., NIST Big Data architecture), to fulfill an emerging gap within the literature, namely the lack of a prescriptive approach to guide practitioners in the design and implementation of BDWs, wherein they can follow rigorous models and methods in real-world projects. Supported by our previous work with: (i) methodological guidelines on the design and implementation of BDWing systems, with proof-of-concepts in the context of Industry 4.0 at Bosch Car Multimedia Braga [9, 10] or Smart Cities [26]; (ii) the extensive evaluation of SQL-on-Hadoop systems for data processing [8, 27–29]; and, (iii) the relevance of extracting value from data, moving from Big Data to Smart Data [30, 31], the next section presents a set of research directions in this field.

3 Big Data Warehousing Systems

Researching in BDWing lacks from reference frameworks and methodological guidelines that help researchers and practitioners in the process of enhancing this valuable data system. Figure 1 depicts the vision proposed in this paper for an integrated BDWing environment supporting the decision-making process. For the current challenges and research directions identified in the introduction of this paper, five main components are here proposed to address them: (i) BDW Entities Resolution, including tasks such as data collection, preparation, enrichment, profiling, and lineage; (ii) BDW Modelling and Implementation; (iii) BDW Management; (iv) BDW Intelligent Event Broker; and, (v) BDW Visualization. Currently, the design and implementation of a BDW is mostly based on use-case driven approaches, preventing a long-term view of the BDW evolution and performance, reason why this proposal looks into a data lifecycle that continuously assists the integration of new data sources, the monitoring of the BDW as a valuable organizational asset, and the enhancement of the decision-making process throughout an innovative and interconnected approach.

3.1 BDW Entities Resolution

The BDW Entities Resolution component addresses the adequate integration of new business processes and data in the BDW, providing a unified and relevant view of the organizational data for decision-making, see Fig. 1. This includes tasks for data sources identification, data understanding, data cleansing, data fusion, data transformation, among many others, with the aim of understanding how new business processes and data can be integrated in the BDW. This is seen as a complex process that is able to deal with the variety of data sources usually available in Big Data contexts, providing adequate and efficient processes to give structure to unstructured or semi-structured data sources (using Data Science techniques and technologies), and to identify relevant entities for analysis, with the Collection, Preparation and Enrichment (CPE) Pipeline, which can be seen as part of a general approach of Big Data management, such as SILE (Search, Identify, Load and Exploitation) proposed in [30], aiming to systematize the search and identification of relevant data to be loaded, analyzed and exploited by a Genomic Information Systems. Although in this case the method is proposed and applied to a specific domain, genomics, the principles are transversal to any application domain: search for relevant data sources; identify relevant datasets; load the relevant data; and, exploit the value of data. From the identification of new business processes

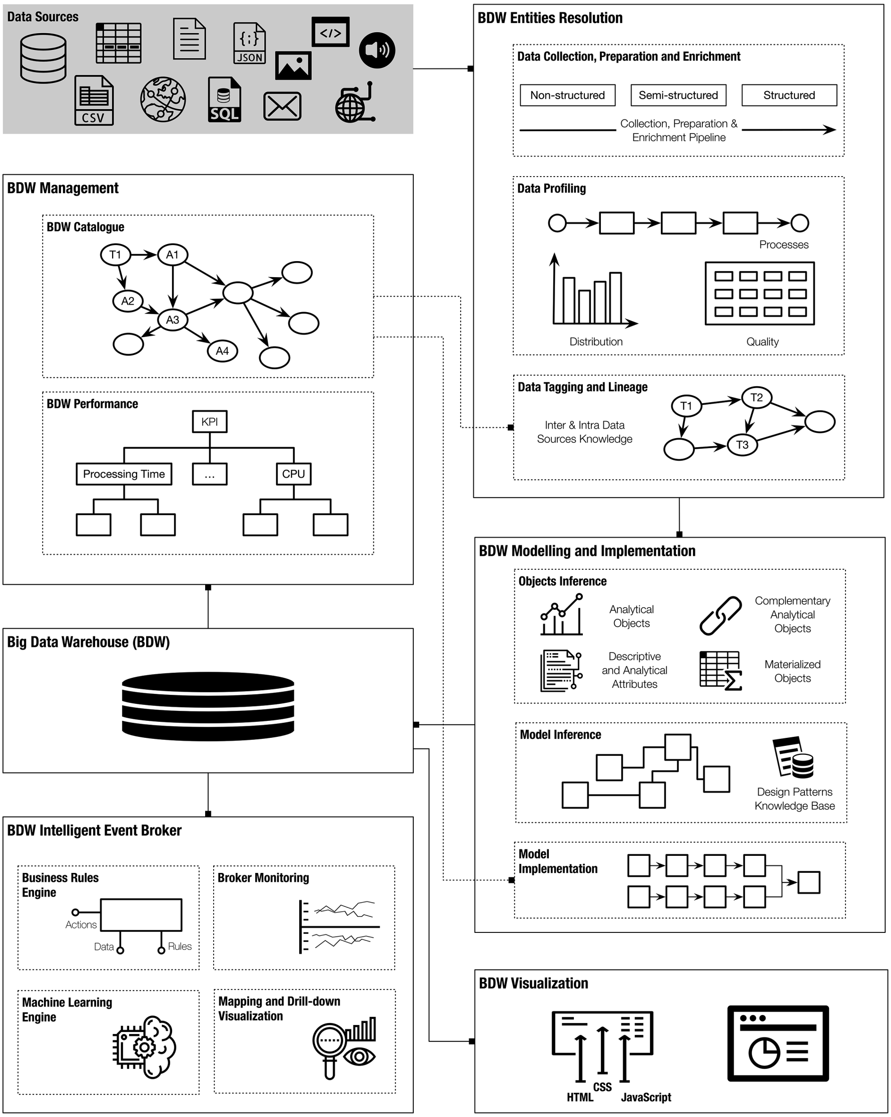

Fig. 1. BDW for efficient, integrated and advanced analytics

and data to their adequate integration in the BDW, there is the need to devise automatic or semi-automatic approaches that are able to take them as input and evaluate how they can be integrated in an already existing data structure, which is highly complex and that needs to comply with demanding workloads and performance issues.

With structured data, Data Profiling characterizes the new business processes (helping in the integration of the new data sources [32] with the already existing data in the BDW), the new data sources, and the attributes that define the events associated to those business processes, addressing their possible values, distribution, and quality.

To seamlessly integrate the arriving data (either in batches or in streaming), Data Tagging and Lineage is complemented with the computation of a set of semantic indicators that verify the affinity and joinability of the attributes [33], showing how they relate to each other and how their integration in the BDW is possible. This is represented by the Inter & Intra Data Sources Knowledge. The Inter Knowledge helps in integrating new data sources with the ones already available in the BDW, whereas the Intra Knowledge gives a conceptual overview of the new data sources. Information for Data Tagging and Lineage can be automatically collected from CPE workloads, Structured Query Language (SQL) scripts, databases' metadata, among others. As a result, for the identified business processes and data, a graph data structure makes available the Inter & Intra Data Sources Knowledge with their characterization and the semantic information that guides their integration in the BDW.

3.2 BDW Modelling and Implementation

In the vision proposed in this paper, BDW Modelling and Implementation must be guided by appropriate methodological guidelines, and not by ad hoc or use case-driven approaches, identifying the suitable data model to integrate the new data in the BDW, ensuring efficient query processing, mixed complex workloads, and an adequate decision support environment (see Fig. 1). As seen in [8], different data modeling approaches can be followed for designing and implementing BDWs, such as completely flat (denormalized) data tables, star schema models, or hybrid approaches that use flat tables and star schemas depending on the data cardinality and distribution. For these different design patterns, which can optimize query processing [27], and for the new business processes and data, the data modelling constructs need to be inferred using the information available from the BDW Entities Resolution. Here, the characteristics of the data and the data modelling constructs are mapped, identifying:

- Analytical objects (such as sales, inventory management, purchases, among others);
- Complementary analytical objects (similar to conformed dimensions in the Kimball approach [11]);
- Descriptive and analytical attributes, where descriptive attributes add a meaning to the analytical attributes, like product descriptions for the value of sales;
- Materialized objects (views) that increase efficiency for complex and long-running queries.

With these, a Design Patterns Knowledge Base is used to derive a data model and to later implement it using (semi)-automated procedures, adding the new physical structures and data to the BDW. This knowledge base stores information about the data modeling design patterns, as well as their performance attending to the characteristics of the data. This way, a data model can be derived, suggesting its implementation by following a specific design pattern. Afterwards, as will be seen in the next subsection, if the volume of data increases, if the data distribution changes, or if performance in query processing is not satisfactory, the BDW Management component can recommend changes to the data model, suggesting the adoption of different design patterns. The model implementation, having the data model and the CPE workloads, can be optimized leveraging agile and performant BDW's updates.

3.3 BDW Management

As the number of business processes and data sources starts to increase in the BDW, there is the need to know which tables and attributes were stored, to which business processes they are related, when they were created or added, and how they are evolving over time, such as how many rows were added, and when were they added. This gives real-time information about the BDW and its evolution. For the adequate BDW Management, the BDW Catalogue (see Fig. 1), a graph-based structure with the BDW's metadata, includes information about the business processes, tables, attributes, loading processes, among others. This structure also complements the semantic knowledge needed for Data Tagging and Lineage in the BDW Entities Resolution, establishing the Inter & Intra Data Sources Knowledge, as it complements the knowledge of the existing data with the one obtained from the new business processes and data sources.

Besides cataloguing the BDW, supporting its governance, it is also relevant to monitor its performance, verifying its efficiency in query processing. This challenge is not seen here as a process of adopting more performant processing tools [28], but as an architectural change in terms of the data models, adjusting the design patterns attending to the characteristics of the available data [8]. This is important to devise strategies for improving the BDW's efficiency. In this case, a recommendation system can analyze the current state of the BDW, using descriptive statistics, affinity measures, joinability measures, metadata, query performance and query repetition, for instance, and propose changes to the BDW data model, through the implementation of additional analytical objects or structures like complementary analytical objects or materialized objects [10], thus increasing the overall efficiency of the system. With this, data models can evolve, changing the previously adopted design patterns, if that is advantageous for the BDW's efficiency. Moreover, the BDW Performance must use a Key Performance Indicators (KPIs) tree that assists this monitoring task, providing a list of objective metrics and the corresponding targets. As the data models evolve, the KPIs can show the impact of those changes in the overall performance of the BDW.

3.4 BDW Intelligent Event Broker

For processing real-time data in the BDW, as a relevant functionality of an analytical system in a Big Data context, there is the need for automated decision-making processes through Complex Event Processing and Machine Learning, adopting innovative ways to process complex events in a streamlined, scalable, analytical and integrated way [34, 35]. The BDW Intelligent Event Broker, see Fig. 1, is a just-in-time data dissemination system using highly flexible business rules and Machine Learning models to handle the event data that is available mainly due to the proliferation of IoT devices. Therefore, there are several contexts in which this system can be used (industry, smart cities, logistics, agriculture, among others), preventing possible problems by using the data produced by several sources and processing it in real-time. The monitoring of a production line is a relevant example where the verification of the rules in a defect product can result in the application of Machine Learning models to predict if the next products will also be defective, and then activating the needed actions. Consequently, for this system, several components are needed, such as the following:

- Business Rules Engine for defining a set of business rules to be applied to the data/events, as well as the actions that must be taken as a consequence of triggering a specific rule. A repository of business rules for managerial actions at different organizational levels (mainly tactical and operational) feeds the Intelligent Event Broker and uses data that arrives to the BDW (in streaming or batch), combining real-time and historical data in the decision-making process;
- Machine Learning Engine for importing previously trained Machine Learning models from a Models Lake, using a Machine Learning as a Service approach, in order to predict future events and, if needed, provide corrective or optimal actions regarding the event;
- Broker Monitoring to automatically track the functioning of the Intelligent Event Broker, by collecting metadata regarding rules, triggers, KPIs, among others. This component is used to monitor the performance of the broker and devise strategies to improve it; and,
- Mapping and Drill-down Visualization to: (i) inspect the rules that have been activated and drill-down into the data that activated those rules; and (ii) visualize KPIs about the broker itself and drill-down into their relationship with the rules, the triggers and the corresponding data.

Considering the analyzed related work [36, 37], some concepts and components here mentioned are widely recognized for this type of system (e.g., rules and triggers). However, these works do not consider: (i) the inclusion of concepts similar to the Machine Learning Models Lake component that can be helpful for patterns discovery in Complex Event Processing systems for Big Data contexts; and, (ii) the relevance of the system monitoring through an innovative visualization platform, as its evolution can quickly become untraceable in Big Data contexts.

3.5 BDW Visualization

Visualization is one of the key components to take advantage of the data made available through the BDW, enhancing decision making with appropriate visual analytics tools. Technological developments in Big Data contexts are mainly driven by open source initiatives, but as the open source Big Data Visualization landscape is still very limited when compared to commercial solutions, practitioners have mainly two alternatives, namely, use open source solutions or custom-made Web visualizations.

In both cases, existing applications usually provide an environment for static and/or more dynamic analyses, with classical or more advanced visualization methods [38], or with the identification of interaction patterns for designing user interfaces oriented towards extracting knowledge from Big Data [39]. In open source tools like Superset (https://superset.incubator.apache.org), base functionalities are provided but improvements are still needed regarding customization. When using commercial solutions, usually including a wide variety of visualization methods and functionalities, there is still the lack of customization and interaction that can be achieved with custom-made Web platforms, like real-time access to data, highly customized events and interactions, or calculations involving multiple sources. Taking this into consideration, there is the need for BDW Visualization (see Fig. 1) platforms oriented towards dashboard development by advanced data analysts and data scientists, providing a way to create custom-made and interactive dashboards using small portions of reusable code that can be easily integrated (like HTML, CSS and JavaScript code), including rich, highly flexible, customizable and interactive charts or other visualization components.

4 Conclusions

This paper highlighted the research topics associated with current challenges and open issues in BDWs as highly flexible, scalable and performant systems for supporting decision-making processes. In this work, some proposals were refined and structured to become a roadmap for the research community for the next years. This vision tries to highlight were value can be added to a BDW, by approaching a fast-changing world that needs to deal with the constant integration of new business processes and data sources, and by understanding the proper way to adjust the BDW and its data model as this evolution occurs. Also, it is crucial to deal with the management of these complex data systems to enhance their performance, as well as addressing real-time analytical capabilities through the use of Complex Event Processing and Machine Learning.

In this paper, all these challenges were instantiated with research areas. For *the integration of new business processes and data sources*, approaches from research areas like Entities Resolution, Data Profiling, Data Tagging, and Data Lineage can be applied to provide information for *the proper way to achieve this integration*, based on appropriate data models, with the attempt to provide a data model in a semi-automated way, based on a Design Patterns Knowledge Base. This type of contribution will help organizations that deal with huge amounts of data arriving from several sources and will help them *to manage these complex data systems and to enhance their performance*, reducing the time needed for tasks such as the BDW management and modeling, allowing their users to focus on retrieving value from data. Moreover, the capability to deal with other contexts, like events and streaming processing, *automating the analytical capabilities of a BDW*, is another way to enhance the BDW and its value.

Currently, streaming data is constantly being produced in different contexts by the several interconnected devices and people within the organizations. Its efficient processing and usage are relevant to promote better decisions for decision makers, or, sometimes, take decisions in an automated way. To accomplish this goal, the Intelligent Event Broker is responsible for the real-time application of business rules and Complex

Event Processing, allowing the identification of events and problems that can be dispatched by several triggers that take semi-automated actions. The Machine Learning component is a core component of the Broker, making available the problems identification and recommendations even before these problems occur, based on the data that arrives to the system. For this kind of system, its complexity needs to be managed, being a monitoring and visualization component proposed to understand and track what happens in the system.

In addition, *the flexible and highly customizable visualization of data*, for extracting value from BDWs, is tightly-coupled with an adequate data visualization mechanism, reason why this work discusses flexible, highly customizable, and interactive visualization mechanisms based on portions of reusable code, which will provide an advanced decision-making support environment based on rich Web-based user interfaces.

Therefore, reference frameworks and methodological guidelines are strictly required in this domain to provide effective solutions intended to manage adequately the studied problem. After analyzing relevant research directions, the paper proposes and introduces a framework that takes into account the most significant aspects of the domain, and that can be used as a starting point to characterize BDWs for efficient, integrated and advanced analytics as expressed in the work title. It is our firm intention to apply, improve and extend it (where necessary) using challenging examples as the Genome Data Science domain and the Industry 4.0 environment with the Bosch Car Multimedia case, in which we already have at the moment some initially, encouraging results.

Acknowledgements. This work has been supported by FCT – *Fundação para a Ciência e Tecnologia*, Projects Scope UID/CEC/00319/2019 and PDE/00040/2013, and the Doctoral scholarships PD/BDE/135100/2017 and PD/BDE/135101/2017. We also thank both the Spanish State Research Agency and the Generalitat Valenciana under the projects DataME TIN2016-80811-P, ACIF/2018/171, and PROMETEO/2018/176. This paper uses icons made by Freepik, from www.flaticon.com.

References

1. Madden, S.: From databases to big data. IEEE Internet Comput. **16**(3), 4–6 (2012)
2. Dumbill, E.: Making sense of big data. Big Data **1**, 1–2 (2013)
3. Gandomi, A., Haider, M.: Beyond the hype: Big data concepts, methods, and analytics. Int. J. Inf. Manag. **35**, 137–144 (2015)
4. Philip Chen, C.L., Zhang, C.-Y.: Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inf. Sci. **275**, 314–347 (2014)
5. Costa, C., Santos, M.Y.: Big data: state-of-the-art concepts, techniques, technologies, modeling approaches and research challenges. IAENG Int. J. Comput. Sci. **44**, 285–301 (2017)
6. NBD-PWG: NIST Big Data Interoperability Framework (2015)
7. Krishnan, K.: Data Warehousing in the Age of Big Data. Elsevier, Burlington (2013)

8. Costa, C., Santos, M.Y.: Evaluating several design patterns and trends in big data warehousing systems. In: Krogstie, J., Reijers, H.A. (eds.) CAiSE 2018. LNCS, vol. 10816, pp. 459–473. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91563-0_28
9. Santos, M.Y., et al.: A Big Data system supporting Bosch Braga Industry 4.0 strategy. Int. J. Inf. Manag. **37**, 750–760 (2017)
10. Costa, C., Andrade, C., Santos, M.Y.: Big data warehouses for smart industries. In: Sakr, S., Zomaya, A. (eds.) Encyclopedia of Big Data Technologies, pp. 1–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-63962-8_204-1
11. Kimball, R., Ross, M.: The Data Warehouse Toolkit: The definitive Guide to Dimensional Modeling. Wiley, Indianapolis (2013)
12. Clegg, D.: Evolving data warehouse and BI architectures: the big data challenge. TDWI Bus. Intell. J. **20**, 19–24 (2015)
13. Russom, P.: Data Warehouse Modernization in the Age of Big Data Analytics (2016)
14. Russom, P.: Evolving Data Warehouse Architectures in the Age of Big Data (2014)
15. 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, Rome, Italy, pp. 142–149 (2016). https://doi.org/10.5220/0005830801420149
16. Chevalier, M., El Malki, M., Kopliku, A., Teste, O., Tournier, R.: Implementing multidimensional data warehouses into NoSQL. In: 17th International Conference on Enterprise Information Systems (ICEIS), Barcelona, Spain (2015)
17. Gröger, C., Schwarz, H., Mitschang, B.: The deep data warehouse: link-based integration and enrichment of warehouse data and unstructured content. In: IEEE 18th International Enterprise Distributed Object Computing Conference (EDOC), pp. 210–217 (2014)
18. Cattell, R.: Scalable SQL and NoSQL data stores. ACM SIGMOD Record. **39**, 12 (2011)
19. Thusoo, A., et al.: Hive-a petabyte scale data warehouse using hadoop. In: 2010 IEEE 26th International Conference on Data Engineering (ICDE), pp. 996–1005. IEEE (2010)
20. Pandis, I.: Impala: a modern, open-source SQL engine for hadoop. In: 7th Biennial Conference on Innovative Data Systems Research (CIDR), p. 10 (2015)
21. Huai, Y., et al.: Major technical advancements in apache hive. In: Proceedings of the 2014 ACM SIGMOD international conference on Management of data - SIGMOD 2014, pp. 1235–1246. ACM Press, Snowbird (2014). https://doi.org/10.1145/2588555.2595630
22. Li, X., Mao, Y.: Real-Time data ETL framework for big real-time data analysis. In: 2015 IEEE International Conference on Information and Automation, pp. 1289–1294. IEEE, Lijiang (2015). https://doi.org/10.1109/ICInfA.2015.7279485
23. Song, J., Guo, C., Wang, Z., Zhang, Y., Yu, G., Pierson, J.-M.: HaoLap: a hadoop based OLAP system for big data. J. Syst. Softw. **102**, 167–181 (2015)
24. Wang, H., et al.: Efficient query processing framework for big data warehouse: an almost join-free approach. Front. Comput. Sci. **9**, 224–236 (2015)
25. Tria, F.D., Lefons, E., Tangorra, F.: A framework for evaluating design methodologies for big data warehouses: measurement of the design process. Int. J. Data Warehouse. Min. **14**(1), 15–39 (2018)
26. Costa, C., Santos, M.Y.: The SusCity big data warehousing approach for smart cities. In: Proceedings of International Database Engineering & Applications Symposium. Bristol, United Kingdom (2017). https://doi.org/10.1145/3105831.3105841
27. Costa, E., Costa, C., Santos, M.Y.: Efficient big data modelling and organization for hadoop hive-based data warehouses. In: Themistocleous, M., Morabito, V. (eds.) EMCIS 2017. LNBIP, vol. 299, pp. 3–16. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65930-5_1

28. Rodrigues, M., Santos, M.Y., Bernardino, J.: Big data processing tools: an experimental performance evaluation. WIREs Data Min. Knowl. Discov. **9**(2), e1297 (2019)
29. Santos, M.Y., et al.: Evaluating SQL-on-hadoop for big data warehousing on not-so-good hardware. In: Proceedings of International Database Engineering & Applications Symposium (IDEAS 2017), pp. 242–252. ACM Press (2017). https://doi.org/10.1145/3105831.3105842
30. León Palacio, A., Pastor López, Ó.: Smart data for genomic information systems: the SILE method. Complex Syst. Inf. Model. Q. 1–23 (2018). https://doi.org/10.7250/csimq.2018-17.01
31. Palacio, A.L., López, Ó.P., Ródenas, J.C.C.: A method to identify relevant genome data: conceptual modeling for the medicine of precision. In: Trujillo, J.C., Davis, K.C., Du, X., Li, Z., Ling, T.W., Li, G., Lee, M.L. (eds.) ER 2018. LNCS, vol. 11157, pp. 597–609. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00847-5_44
32. Hui, J., Li, L., Zhang, Z.: Integration of big data: a survey. In: Zhou, Q., Gan, Y., Jing, W., Song, X., Wang, Y., Lu, Z. (eds.) ICPCSEE 2018. CCIS, vol. 901, pp. 101–121. Springer, Singapore (2018). https://doi.org/10.1007/978-981-13-2203-7_9
33. Maccioni, A., Torlone, R.: KAYAK: a framework for just-in-time data preparation in a data lake. In: Krogstie, J., Reijers, H.A. (eds.) CAiSE 2018. LNCS, vol. 10816, pp. 474–489. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91563-0_29
34. Flouris, I., Giatrakos, N., Deligiannakis, A., Garofalakis, M., Kamp, M., Mock, M.: Issues in complex event processing: status and prospects in the Big Data era. J. Syst. Softw. **127**, 217–236 (2017). https://doi.org/10.1016/j.jss.2016.06.011
35. Zhang, P., Shi, X., Khan, S.U.: QuantCloud: enabling big data complex event processing for quantitative finance through a data-driven execution. IEEE Trans. Big Data (2018). https://doi.org/10.1109/TBDATA.2018.2847629
36. Hadar, E.: BIDCEP: a vision of big data complex event processing for near real-time data streaming: position paper, a practitioner view. In: CAiSE 2016 Industry Track, CEUR Workshop Proceedings (2016)
37. Flouris, I., et al.: FERARI: a prototype for complex event processing over streaming multi-cloud platforms. In: Proceedings of the 2016 International Conference on Management of Data - SIGMOD 2016, pp. 2093–2096. ACM Press, San Francisco (2016). https://doi.org/10.1145/2882903.2899395
38. Bikakis, N.: Big data visualization tools. In: Sakr, S., Zomaya, A. (eds.) Encyclopedia of Big Data Technologies. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-63962-8_109-1
39. Iñiguez-Jarrín, C., Panach, J.I., Pastor López, O.: Defining interaction design patterns to extract knowledge from big data. In: Krogstie, J., Reijers, H.A. (eds.) CAiSE 2018. LNCS, vol. 10816, pp. 490–504. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91563-0_30