Building Data Warehouses in the Era of Big Data

An Approach for Scalable and Flexible Big Data Warehouses

Carlos Costa[1] and Maribel Yasmina Santos[2](✉)

[1] Centre for Computer Graphics – CCG, Guimarães, Portugal
carlos.costa@dsi.uminho.pt

[2] ALGORITMI Research Centre, Department of Information Systems, University of Minho, Guimarães, Portugal
maribel@dsi.uminho.pt

Abstract. During the last few years, the concept of Big Data Warehousing gained significant attention from the scientific community, highlighting the need to make design changes to the traditional Data Warehouse (DW) due to its limitations, in order to achieve new characteristics relevant in Big Data contexts (e.g., scalability on commodity hardware, real-time performance, and flexible storage). The state-of-the-art in Big Data Warehousing reflects the young age of the concept, as well as ambiguity and the lack of common approaches to build Big Data Warehouses (BDWs). Consequently, an approach to design and implement these complex systems is of major relevance to business analytics researchers and practitioners. In this tutorial, the design and implementation of BDWs is targeted, in order to present a general approach that researchers and practitioners can follow in their Big Data Warehousing projects, exploring several demonstration cases focusing on system design and data modelling examples in areas like smart cities, retail, finance, manufacturing, among others.

Keywords: Big Data · Data Warehousing · Big Data Warehousing · Analytics

1 Topic Relevance and Novelty

Nowadays, the community is studying the role of the DW in Big Data environments [1], thus the concept of BDW is emerging, with new characteristics and design changes. Currently, research on this topic is scarce and the state-of-the-art shows that the design of BDWs should focus both on the physical layer (infrastructure) and on the logical layer (data models and interoperability between components) [2]. Moreover, in general terms, a BDW can be implemented by leveraging the capabilities of Hadoop, NoSQL, or NewSQL to either complement or fully replace traditional relational databases. The existing non-structured practices and guidelines are not enough. The lack of prescriptive research on the topic of Big Data Warehousing is alarming, as there is no common approach to design and implement BDWs, as formerly existed in the

P. Giorgini and B. Weber (Eds.): CAiSE 2019, LNCS 11483, pp. 693–695, 2019.
https://doi.org/10.1007/978-3-030-21290-2

realm of traditional DWs, and the community needs a rigorously evaluated approach to design and build BDWs, according to recent and improved characteristics and data structures [3, 4].

The shift to a use case driven approach and the young age of Big Data as a research topic result in ambiguity regarding BDWs, but, as [5] claims, it would be a mistake to discard decades of architectural best practices based on the assumption that storage for Big Data is not relational nor driven by data modelling. The SQL-on-Hadoop movement [6] proves that data structures known for many years are more relevant than ever, although modified and optimized, as will be seen in this tutorial.

2 Goal and Objectives

The main goal of this tutorial is to disseminate an approach that can be prescribed for BDW design and implementation, providing to practitioners and researchers a structured and evaluated way of building these complex systems. This approach intends to avoid the risk of uncoordinated data silos frequently seen in today's environments, due to a "lift and shift" strategy and an excessive focus on trying to find the best technology to meet the demands. Considering this context, this tutorial will help the audience understand the purpose and characteristics of BDWs, and it will also demonstrate how to use a general approach for the design and implementation of BDW that can be replicated for several real-world applications. The tutorial objectives are as follows:

1. Understand the role of the BDW in the vast Big Data landscape, and clearly identify the technologies suitable for this context;
2. Define the logical components and data flows of the Big Data Warehousing system, using appropriate and replicable design philosophies and constructs;
3. Plan adequate data pipelines to collect, prepare, and enrich batch and streaming data;
4. Learn how to apply a data modelling method for BDWs, in order to avoid different ad hoc approaches applied in each project or use case;
5. Learn how to design a system for a real-world Big Data Warehousing application domain, focusing on data modelling and data visualization capabilities;
6. Apply the data modelling method in several real-world applications (e.g., smart cities, retail, finance, and manufacturing), exercising different modelling guidelines.

References

1. Krishnan, K.: Data Warehousing in the Age of Big Data. Morgan Kaufmann Publishers Inc., San Francisco (2013)
2. Russom, P.: Evolving Data Warehouse Architectures in the Age of Big Data. The Data Warehouse Institute (2014)
3. Costa, C., Santos, M.Y.: Evaluating several design patterns and trends in big data warehousing systems. In: Krogstie, J., Reijers, H. (eds.) CAiSE 2018. LNCS, vol. 10816, pp. 459–473. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91563-0_28

4. Costa, C., Andrade, C., Santos, M.Y.: Big data warehouses for smart industries. In: Sakr, S., Zomaya, A. (eds.) Encyclopedia of Big Data Technologies. Springer, Cham (2018)
5. Clegg, D.: Evolving data warehouse and BI architectures: the big data challenge. TDWI Bus. Intell. J. **20**, 19–24 (2015)
6. Floratou, A., Minhas, U.F., Özcan, F.: SQL-on-Hadoop: full circle back to shared-nothing database architectures. Proc. VLDB Endow. **7**, 1295–1306 (2014). https://doi.org/10.14778/2732977.2733002