Data Warehousing Process Modeling from Classical Approaches to New Trends: Main Features and Comparisons

Asma Dhaouadi, Khadija Bousselmi, Mohamed Mohsen Gammoudi, Sébastien Monnet, Slimane Hammoudi

data

MDPI

Article

Data Warehousing Process Modeling from Classical Approaches to New Trends: Main Features and Comparisons

Asma Dhaouadi [1,2,3,*], Khadija Bousselmi [2], Mohamed Mohsen Gammoudi [1,4] and Sébastien Monnet [2] and Slimane Hammoudi [5]

1 RIADI Laboratory, University of Manouba, Mannouba 2010, Tunisia
2 LISTIC Laboratory, University of Savoie Mont Blanc, France Annecy-Chambéry, 74940 Chambéry, France
3 Faculty of Sciences of Tunis, University of Tunis El Manar, Tunis 1068, Tunisia
4 Higher Institute of Arts and Multimedia Manouba, University of Manouba, Manouba 2010, Tunisia
5 ERIS, ESEO-Grande Ecole d'Ingénieurs Généralistes, 49100 Angers, France
* Correspondence: asma.dhaouadi@univ-smb.fr or asma.dhaouadi@fst.utm.tn

Abstract: The extract, transform, and load (ETL) process is at the core of data warehousing architectures. As such, the success of data warehouse (DW) projects is essentially based on the proper modeling of the ETL process. As there is no standard model for the representation and design of this process, several researchers have made efforts to propose modeling methods based on different formalisms, such as unified modeling language (UML), ontology, model-driven architecture (MDA), model-driven development (MDD), and graphical flow, which includes business process model notation (BPMN), colored Petri nets (CPN), Yet Another Workflow Language (YAWL), CommonCube, entity modeling diagram (EMD), and so on. With the emergence of Big Data, despite the multitude of relevant approaches proposed for modeling the ETL process in classical environments, part of the community has been motivated to provide new data warehousing methods that support Big Data specifications. In this paper, we present a summary of relevant works related to the modeling of data warehousing approaches, from classical ETL processes to ELT design approaches. A systematic literature review is conducted and a detailed set of comparison criteria are defined in order to allow the reader to better understand the evolution of these processes. Our study paints a complete picture of ETL modeling approaches, from their advent to the era of Big Data, while comparing their main characteristics. This study allows for the identification of the main challenges and issues related to the design of Big Data warehousing systems, mainly involving the lack of a generic design model for data collection, storage, processing, querying, and analysis.

Keywords: ETL process; data warehouse; ETL modeling; Big Data; UML; BPMN; ontology; MDA; graphical flow; systematic review

Citation: Dhaouadi, A.; Bousselmi, K.; Gammoudi, M.M.; Monnet, S.; Hammoudi, S. Data Warehousing Process Modeling from Classical Approaches to New Trends: Main Features and Comparisons. *Data* **2022**, *7*, 113. https://doi.org/10.3390/data7080113

Academic Editor: Kesheng Wu

Received: 5 May 2022
Accepted: 28 July 2022
Published: 12 August 2022

1. Introduction

The globalization and the spread of information technology, the strong concurrency between different companies, and the urge for quick and easy access to reliable and relevant information have incited business leaders to replace traditional business computing systems with other decision support systems. These business intelligence (BI) decision systems appeared with the introduction of the data warehouse (DW) by Bill Inmon in 1991. According to [1], "A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision-making process". A DW is a system used for integrating, storing, and processing data from often heterogeneous data sources, in order to provide decision-makers with a multi-dimensional view. The integration of these data is achieved through a three-phase process: extract, transform, and load (ETL). This process is responsible for extracting data from different data sources, transforming them (by preparation, conversion, clean, filter, conversion, join, aggregation, and so on), and loading them into a DW. Consequently, the general framework for ETL processes