**LITERATURE SURVEY**

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

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

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. The ETL process is used to extract data from different sources; transform them to meet specific analytical needs; and, finally, load the processed data into a dedicated storage system to support them, called a data warehouse. As the success of the project and the ease of its maintenance are strongly linked to the modeling stage, all DW development projects should rely on the well-designed modeling of the data warehousing process, as there is no standard model for the representation and design of this process at present. In the early 2000s, the research community worked towards proposing different methods for conceptual, logical, and physical modeling for the ETL process. As a result, many studies have been published in this field, where each proposed contribution has its specific advantages and suffers from limitations. However, with the emergence of Big Data, the community has been faced with new challenges. Hence, considering the importance of this topic, our main objective in this paper was to review relevant research conducted from the introduction of ETLs to the present day. In this paper, we defined a set of comparison criteria to simplify understanding ETL/ELT process characteristics. We categorized the proposed research works into six major classes, UML, ontology, MDA and MDD, graphical flow, ad hoc formalisms, and, finally, contributions in the context of Big Data. Then, a comparative study of the different contributions was presented and discussed. Our synthetic study browsed the related review papers in this field and we discussed other findings from our survey, thus proving the usefulness of our literature review. We proposed some general recommendations and a case study using the comparative study. Finally, we found that, to date, no synthetic study in the field of ETL process modeling considering the characteristics of Big Data has been carried out. Consequently, ETL process modeling, in its different phases, must evolve to support the new generation of technologies that have emerged in the era of Big Data, particularly in terms of data collection, storage, processing, querying, and analysis.

**Title:** **Integration of Data Warehouse and Unstructured Business Documents**

**Author**: Ahmad Abdullah Alqarni; Eric Pardede

The profusion of unstructured data forced organizations to manage and take advantage of such data especially in the decision making process. The feasibility of integrating or mapping unstructured data to a data warehouse is becoming significant to bridge this gap and take the full potential of these data. In this paper, we propose a multi-layer schema for mapping structured data stored in a data warehouse and unstructured data in business-related documents. The multi-layer schema facilitates the mapping between the two different data. Linguistically correlated data is identified using Word Net to enable the integration between both data sources. We also propose a generic XML schema for business-related unstructured documents to assist the mapping. The use Word Net to identify the matching result is promising in the absence of schema-instance and without the need to domain specific knowledge. The recent development of analytical information systems shows that the necessary integration of structured and unstructured data sources in data warehousing is possible. The usage of the market information system shows that the database improves the analytical power of decision makers, in order to recognize tendencies in the energy market promptly. Nevertheless the respective model and the system must grant high flexibility to adjust them to changing conditions in the energy market. Furthermore the activities on the energy market and the work of the analysts will enhance the system. Market information systems have to be optimized by better evaluation of external information and automatization of process integration. Only documents of decision relevance should be delivered to the management. The ROI of data warehouse projects can be increased if event-based and accepted information improves the decision quality significantly. The information flow alignment in MAIS is equivalent to a classification problem. We assure this by using role profiles and embedded recommendation systems with a document trigger mechanism. Furthermore the use of a simulation method is tightly linked to this process by matching simulation variables to trigger conditions. The integration of metadata from a data warehouse, personalized search patterns and simulation variables give a powerful repository for active data warehousing. The theoretical approach and the benefit of creating interfaces for the meta models are part of further research. Nevertheless, decision makers gain individualized decision support and early insight into future developments.

The quality of classification algorithms must be examined in regular time intervals to guarantee best results. Therefore it is necessary to optimize the structure of the test environment which has to support intersubjective and intertemporal comparability of the test results. Classification evaluations are often accomplished; however these results are only important in the context of the selected data set and evaluation environment. In order to acquire concrete statements for MAIS, such an evaluation environment and the results are described in this paper. In order to find the perfect search terms, the most relevant documents are to be found so that not just the classification itself has to be optimized, but the Internet retrieval as well.

**Title:** **The History, Present, and Future of ETL Technology**

**Author:** Alkis Simitsis, Spiros Skiadopoulos, Panos Vassiliadis

There is an abundance of data, but a large volume of it is unusable. Data may be noisy, unstructured, stored in incompatible for direct analysis medium or format, and often expensive to access. In most practical cases, the data needs to be processed before it can be used to extract valuable business insights. We refer to the nontrivial, end-to-end operation of extracting intelligence from raw data as an ETL process. In this paper, we review how the ETL technology has been evolved in the last 25 years, from a rather neglected engineering challenge to a first-class citizen in analytics and data processing. We present a brief historical overview of ETL, discuss its various applications and incarnations in modern data processing environments, and argue about exciting, feasible or wishful, potential future directions. The ETL technology and data integration in general has been the cornerstone of business intelligence, decision making, and data analytics for over 25 years. ETL thrives while at the same time it evolves along with shifting business needs and data technology advancements. As researchers and practitioners alike are exploring ways to extract value from large collections of raw data, ETL is the connecting glue to make this happen. In this paper, we presented a brief overview of the ETL history, described recent trends in the end-to-end data stack, and discussed some interesting, in our opinion, future directions that will most likely impact the next generation of ETL and data integration technology. The past 20+ years have been educating, enjoyable, and productive in devising and realizing efficient and effective ways to tame data intricacies and peculiarities blending a multiplicity of technologies and applying them in the real world. We look forward to the next 20 that will be even more exciting and fruitful.

**Title:** **An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management**

**Author:** Athira Nambiar \* and Divyansh Mundra

Data is the lifeblood of any organization. In today’s world, organizations recognize the vital role of data in modern business intelligence systems for making meaningful decisions and staying competitive in the field. Efficient and optimal data analytics provides a competitive edge to its performance and services. Major organizations generate, collect and process vast amounts of data, falling under the category of big data. Managing and analyzing the sheer volume and variety of big data is a cumbersome process. At the same time, proper utilization of the vast collection of an organization’s information can generate meaningful insights into business tactics. In this regard, two of the popular data management systems in the area of big data analytics (i.e., data warehouse and data lake) act as platforms to accumulate the big data generated and used by organizations. Although seemingly similar, both of them differ in terms of their characteristics and applications. This article presents a detailed overview of the roles of data warehouses and data lakes in modern enterprise data management. We detail the definitions, characteristics and related works for the respective data management frameworks. Furthermore, we explain the architecture and design considerations of the current state of the art. Finally, we provide a perspective on the challenges and promising research directions for the future. Enterprises and business organizations exploit a huge volume of data to understand their customers and to make informed business decisions to stay competitive in the field. However, big data come in a variety of formats and types (e.g., structured, semi-structured and unstructured data), making it difficult for businesses to manage and use them effectively. Based on the structure of the data, typically, two types of data storage are utilized in enterprise data management: the data warehouse (DW) and data lake (DL). Although being used as interchangeable terms, they are two distinct storage forms with unique characteristics that serve different purposes. In this review, a comparative analysis of data warehouses and data lakes by highlighting the key differences between the two data management approaches was envisaged. In particular, the definitions of the data warehouse and data lake, highlighting their characteristics and key differences, were detailed. Furthermore, the architecture and design aspects of both DWs and DLs are clearly discussed. In addition, a detailed overview of the popular DW and DL tools and services was also provided. The key challenges of big data analytics in general, as well as the challenges of implementation of DWs and DLs, were also critically analyzed in this survey. Finally, the opportunities and future research directions were contemplated. We hope that the thorough comparison of existing data warehouses vs. data lakes and the discussion of open research challenges in this survey will motivate the future development of enterprise data management and benefit the research community significantly.

**Title:** **An Efficient Spark-Based Hybrid Frequent Itemset Mining Algorithm for Big Data**

**Author**: Mohamed Reda Al-Bana, Marwa Salah Farhan and Nermin Abdelhakim Othman

Frequent itemset mining (FIM) is a common approach for discovering hidden frequent patterns from transactional databases used in prediction, association rules, classification, etc. Apriori is an FIM elementary algorithm with iterative nature used to find the frequent itemsets. Apriori is used to scan the dataset multiple times to generate big frequent itemsets with different cardinalities. Apriori performance descends when data gets bigger due to the multiple dataset scan to extract the frequent itemsets. Eclat is a scalable version of the Apriori algorithm that utilizes a vertical layout. The vertical layout has many advantages; it helps to solve the problem of multiple datasets scanning and has information that helps to find each itemset support. In a vertical layout, itemset support can be achieved by intersecting transaction ids (tidset/tids) and pruning irrelevant itemsets. However, when tids become too big for memory, it affects algorithms efficiency. In this paper, we introduce SHFIM (spark-based hybrid frequent itemset mining), which is a three-phase algorithm that utilizes both horizontal and vertical layout diffset instead of tidset to keep track of the differences between transaction ids rather than the intersections. Moreover, some improvements are developed to decrease the number of candidate itemsets. SHFIM is implemented and tested over the Spark framework, which utilizes the RDD (resilient distributed datasets) concept and in-memory processing that tackles MapReduce framework problem. We compared the SHFIM performance with Spark-based Eclat and dEclat algorithms for the four benchmark datasets. Experimental results proved that SHFIM outperforms Eclat and dEclat Spark-based algorithms in both dense and sparse datasets in terms of execution time. FIM is the most common technique used in discovering frequent patterns from transactional datasets. Frequent patterns have a wide effect in many applications including classifications, market basket analysis, mobile computing, web mining, etc. Apriori is computing intensive algorithm; therefore, lots of resources (Memory and processing) are required. Moreover, Apriori uses horizontal data representation and has some challenges such as multiple dataset scans and candidate generating in each iteration, which makes Apriori suffer from big data. Vertical data representation is smaller than horizontal representation in size and carries information through tidsets about each itemset support. Eclat uses vertical data representation (tidset) and achieved observed performance in sparse datasets, but in dense datasets, it suffers when intermediate results of tidsets become too large for memory. In this paper, we proposed SHFIM (Spark-based Hybrid Frequent Itemset Mining) a novel algorithm that utilizes both the horizontal and vertical layouts to solve the drawbacks in both Apriori and Eclat. SHFIM is a three phases algorithm, which works perfectly in a distributed environment. Phases one and two use the horizontal layout to extract the first and second frequent itemset. Phase three is an iterative process to extract k frequent itemset in k iterations. This phase uses mainly diffset to enhance execution time and memory consumption because it shrinks itemsets into smaller sets that will be mined in less time and consume less space. The support of an itemset is calculated by exploiting the vertical layout in every worker node. As the vertical layout size is smaller than the horizontal layout, therefore it requires less memory and less execution time. We developed SHFIM on Spark framework due to its ability to deal with the iterative problem better than Hadoop MapReduce. Spark is 100 times quicker than Hadoop in data processing and has lots of features such as in-memory processing, RDD data structure, broadcasting variables, partitioning of data, and applied Spark best practices to reduce data shuffling between nodes. These features make the Spark the best choice for us that help SHFIM to deal with big data efficiently and increase its execution time performance. Extensive experiments have been conducted between SHFIM, Eclat, and dEclat over Spark clusters for dense and sparse datasets. The Experimental results proved that SHFIM can compete well in both dense and sparse datasets and shows noticeably better performance in either lower or higher min\_sup in terms of execution time than others in datasets that have lots of variable-length transactions which is the nature of big data. In the future work, we are planning to enhance the SHFIM be more efficient. The results proved that the use of tidset, diffset, and Bloom Filter accelerate the speed of FIM in big datasets. We plan to choose between tidset and diffset on the itemset itself rather than the whole dataset instead of applying the diffset and continue using diffset from the third iteration in the whole dataset.