Abstract

Information Technology has been one of the drivers of the revolution that currently is happening in today’s management decisions in most organizations. The amount of data gathered and processed through the use of computing devices has been growing every day, providing a valuable source of information for decision makers that are managing every type of organization, public or private. Gathering the right amount of data in a centralized and unified repository like a data warehouse is similar to build the foundations for a system that will act as a base to support decision making processes requiring factual information. Nevertheless, the complexity of building such a repository is very challenging, as well as developing all the components of a data warehousing system. One of the most critical components of a data warehousing system is the Extract-Transform-Load component, ETL for short, which is responsible for gathering data from information sources, clean, transform and conform it in order to store it in a data warehouse. Several designing methodologies for the ETL components have been presented in the last few years with very little impact in ETL commercial tools. Basically, this was due to an existing gap between the conceptual design of an ETL system and its correspondent physical implementation. The methodologies proposed ranged from new approaches, with novel notation and diagrams, to the adoption and expansion of current standard modeling notations, like UML or BPMN. However, all these proposals do not contain enough detail to be translated automatically into a specific execution platform. The use of a standard well-known notation like Relational Algebra might bridge the gap between the conceptual design and the physical design of an ETL component, mainly due to its formal approach that is based on a limited set of operators and also due to its functional characteristics like being a procedural language operating over data stored in relational format. The abstraction that Relational Algebra provides over the technological infrastructure might also be an advantage for uncommon execution platforms, like computing grids that provide an exceptional amount of processing power that is very critical for ETL systems. Additionally, partitioning data and task distribution over computing nodes works quite well with a Relational Algebra approach. An extensive research over the use of Relational Algebra in the ETL context was conducted to validate its usage. To complement this, a set of Relational Algebra patterns were also developed to support the most common ETL tasks, like changing data capture, data quality enforcement, data conciliation and integration, slowly changing dimensions and surrogate key pipelining. All these patterns provide a formal approach to the referred ETL tasks by specifying all the operations needed to accomplish them in a series of Relational Algebra operations. To evaluate the feasibility of the work done in this thesis, we used a real ETL application scenario for the extraction of data in two different social networks operational systems, storing hashtag usage information in a specific data mart. The ability to analyze trends in social network usage is a hot topic in today’s media and information coverage. A complete design of the ETL component using the patterns developed previously is also provided, as well as a critical evaluation of its usage.

Keywords: Data Warehousing Systems, Extract-Transform-Load Processes, Relational Algebra, ETL Conceptual and Logical Modeling, ETL Patterns, BPMN.