

MASTERS DEGREE DEFENSE

COMPUTER SCIENCE DEPARTMENT

## **The Conceptual Modeling for ETL Process**

By

Jijun Wang

A data warehouse is a copy of transaction data specifically structured for query, reporting, and analysis. Data warehouse makes use of an ETL process, which stands for Extraction – Transformation – Loading, to deal with data cleaning and loading. ETL is a consolidation process that involves retrieving data from a variety of sources, transforming it to meet business needs, and ultimately loading into a data warehouse. In our practical example, ETL process extracts data from human resource system, financial system, and other text or Excel data sources, transform it, and populates the data to Manpower Data Warehouse. Building ETL is the most labor-intensive and lengthy procedure in data warehouse implementation, covering up to eighty percent of effort and expenses [Vassiliadis et al. 2002]. Though ETL is a software system and plays such an important role in data warehouse construction, it is often treated just as a special ad hoc process set running on an ETL tool, or a group of custom programs developed without serious plan and design. Nowadays, these problems are realized and some formal methodologies for ETL have been proposed. However, these theories or methods are still in laboratory phase, without being popularly applied in practice.

Since ETL process plays such an important role in data warehouse building, a well-modeled and robustly designed ETL process is required. In this research, we explored the conceptual model of ETL process proposed by Panos Vassiliadis and applied the conceptual model to an actual ETL process in the Manpower Data Warehouse project. Upon the implementation experience, we strongly underwent the advantages brought by the formal modeling methodology such as maintainability, reuse, and extendibility. However, we also discovered that existing notation symbols are awkward in denoting some practical situations or logics from practical projects. Based on these discoveries we suggested some extended notation symbols for the ETL conceptual model. These notation extensions help to express graphically the timing order or sequence logic in ETL conceptual model and to remove ambiguity in diagram notes. Using the extended notation, we presented the ETL conceptual model of our practical example more clearly and concisely.

Date: Monday, November 26, 2007

Time: 11:00 am - 12:00 pm

Location: 8-48 Computer Science Conference Room

Advisor / Committee Chair: Dr. Daisy Sang  
Committee Members: Dr. H. K. Liu, Dr. Gilbert Young