Cairo University
Institute of Statistical Studies and Research
Computer sciences Department

Data Integration in Data Warehousing

By:
Kareem Kamal Abd Elghany

Computer sciences Department - ISSR – Cairo University

A Thesis submitted in partial fulfillment of the
Requirements for the degree of Master of
Science in Computer Sciences

Under supervision of:
Prof. Dr. Osman Hegazy Mohamed Osman Hegazy

Information Systems Dept. Faculty of computers and information
Cairo University
Prof. Dr. Bahaa El-Din Helmy

Computer Sciences Dept. Institute of Statistical Studies and Research
Cairo University

2009

Table of contents

List of Figures

Acknowledgments

I wish to express my gratitude to ALLAH whose great help is the first factor in every thing I can do in my life.

I would like to express my sincere thanks to my supervisor Prof. Dr.Osman Hegazy, for his guidance and kind supervision.

Special thanks go to the soul of Prof. Dr. Bhaa Eldin Helmy and Prof. Dr. Kamal Abd Elghany (my father) God bless them.

I would like to thank all of my family especially my wife and my mother for their kind and praying for me.

Finally, I would like to send my deepest thanks to all the people who helped me and stood by me to be able to accomplish this work.

Abstract

A Data Warehouse is a collection of technologies aimed to enable the decision maker to make faster and better decisions. The data warehousing has some main problems need to be solved to increase the quality of the data warehouse and its data.

One of the popular problems of the data warehousing is Data Integration, which is the making of query across multiple autonomous and heterogeneous data sources with respect of quality factors.

There are many integration systems (Software) each of which has its own features and benefits which will be stated in details in this research, then the researcher choose one of them to apply a case study.

Keywords: Data Warehouse, Data Integration, Data Quality.

Chapter I

INTRODUCTION

Introduction

A Data Warehouse is a large database that gathers data coming from different sources of an organization, in order to make the retrieval and the querying processes easier and faster.

1.1 ETL Processes in Data Warehousing

ETL stands for Extraction, Transformation, and Loading. This process is either more complex and takes a lot of time than any other parts of data warehouse implementation. It contains the extraction of data from different sources, data cleansing, data customization, data reformatting, data integration and finally data insertion into a data warehouse.

1.2 Data Integration

Data integration is one of the most important aspects of a data warehouse. It is the capture and movement of data from one database on a source system to another database on a target system. It can be found in applications that need to query across multiple autonomous and heterogeneous data sources.

1.3 Data Quality in Data Integration

Data quality has been defined and measured using stringent constraints that need to be satisfied, for example, the concepts of accuracy, completeness, timeliness and consistent figure frequency.

1.4 Problem statement

The quality of the data stored at different sites can be different and also varies over time, so it requires dynamic data integration methods to resolve data conflicts.

1.5 Objective

The research aims at investigating the problems of optimizing data integration during building a data warehouse through making a survey about data integration systems defining the quality of data in the integration process.

1.6 Action plan

The researcher followed the following procedures. They were as follows:

- Surveying some of the integration systems and identifying the features and benefits of each.
- choosing one system to clarify the research objective.
- Applying his approach to the problems of data warehousing and data integration.
- Designing an integrated system using data with different time and application.
- Stating conclusions and future work.

1.7 Thesis outline

This thesis is organized as follows:

A background about data warehousing is presented in Chapter 2 in which the components, characteristics and importance of the data warehousing will be discussed. The researcher briefly describes the ETL process and Meta-data, and the problems of data warehousing.

Chapter 3 presents the approaches, the types, the steps and the importance of data integration. The researcher also describes the measurement factors of data quality.

Chapter 4 presents a survey of some of the popular integration systems describing the features and the benefits of each, and then the researcher will choose one of them to present the research idea as a case study which will be stated in Chapter 5.

Finally, the researcher concludes by summarizing the contributions and presenting future work in Chapter 6.

Chapter II

DATA WAREHOUSING

Data Warehousing

2.1 Data Warehouse definitions

The researcher finds that there are many definitions of Data Warehouse and finds that the following one is the closest one to this research.

A Data Warehouse is a set of materialized views over the operational information sources of an organization, designed to provide support for data analysis and management's decisions. (2)

From the researcher's point of view, a Data Warehouse is a large database that gathers data coming from different sources of an organization in order to make the retrieval and the querying processes easier and faster.

2.2 Data warehousing characteristics

Data warehousing has four characteristics (6) that define its data there are as follows:

2.2.1 Subject-Oriented

Data warehouses are designed to analyze data; this ability to define a data warehouse by subject matter makes the data warehouse subject-oriented and multidimensional.

2.2.2 Integrated

Integration is closely related to subject orientation. Data warehouses should put data from disparate sources into a consistent format, and they must resolve such problems to be integrated.

2.2.3 Nonvolatile

Nonvolatile means that once entered into the warehouse data should not change. This is logical because the purpose of a warehouse is analyzing what has occurred**.**

2.2.4 Time Variant

In order to discover trends in business, analysts need large amounts of data so data warehouses should focus on changing over time which is meant that they are time variant**.**

2.3 Data Warehousing Architectures

A Data Warehouse Architecture is a way of representing the overall structure of data, communication, processing and presentation, which exists for end-user computing within the enterprise. There are three common architectures:

2.3.1 Basic Data Warehouse Architecture

In a simple architecture for a data warehouse, end users directly access data derived from several source systems through the data warehouse.

2.3.2 Data Warehouse Architecture with a Staging Area

In this architecture, the idea is to clean and process the operational data before putting it into the warehouse by using a staging area.

2.3.3 Data Warehouse Architecture with a Staging Area and Data Marts

It is used to customize the warehouse for different groups by adding some data marts.