



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

<u>Chapter 1: Introduction</u>	1
1.1 ETL processes in data warehousing	2
1.2 Data Integration	2
1.3 Data quality in Data Integration	2
1.4 Problem statement	3
1.5 Objective	3
1.6 Action plan	3
1.7 Thesis outline	3
<u>Chapter 2: Data Warehousing</u>	5
2.1 Data Warehouse definitions	6
2.2 Data warehousing characteristics	6
2.3 Data warehousing Architectures	7
2.4 Data warehouse Stages	12
2.4.1 Data capture	12
2.4.2 Transformation & cleansing	13
2.4.3 Aggregation & analysis	13
2.4.4 Presentation	14
2.5 The importance of data warehousing	14
2.6 Data warehousing schemas	16
2.6.1 Star schema	16
2.6.2 Snowflake schemas	17
2.7 Data warehousing components	17
2.7.1 Load Management	18
2.7.2 Warehouse Management	18
2.7.3 Query Management	18
2.8 Data warehouse Frameworks	19
2.9 Data warehouse Methodologies	21
2.10 ETL processes in data warehousing	24
2.10.1 ETL tasks	24
2.10.2 ETL steps	25
2.10.2.1 Extraction	25
2.10.2.2 Transformation	26
2.10.2.3 Loading	26
2.11 Meta data in data warehousing	27
2.12 Data warehousing problems	28

2.12.1 Wrapper/Monitors	29
2.12.2 Warehouse specifications	29
2.12.3 Data Integration	29
Chapter 3: Data Integration	30
3.1 Data Integration definitions	31
3.2 History of data integration	31
3.3 Data Integration Importance	32
3.4 Levels of Data Integration	33
3.4.1 Hand coding of data interfaces	34
3.4.2 Source-to-target mapping, translation tools	34
3.4.3 Integration hubs and brokers	35
3.4.4 Full model-based integration	35
3.5 Data Integration framework	36
3.6 Data Integration approaches	38
3.6.1 Develop and build your own solution	38
3.6.2 Acquire a commercial offering	39
3.7 Types of integration	41
3.7.1 Bi-directional Integration	41
3.7.2 Synchronous vs. Asynchronous Integration	41
3.7.3 Data Staging vs. Peer-to-Peer Integration	42
3.7.4 Cascading Integration	43
3.7.5 Global-as-view integration	43
3.7.6 Local-as-view integration	43
3.7.7 View cooperation integration	44
3.8 Steps to develop an integrated schema	44
3.8.1 The Pre-integration Step	44
3.8.2 The Correspondence Identification Step	46
3.8.3 The Integration Step	47
3.9 Integration strategies	48
3.9.1 Consolidated	48
3.9.2 Federated	49
3.9.3 Shared	49
3.10 Enterprise information integration	49
3.10.1 EII Factors	50
3.10.2 EII scenario	50
3.10.3 EII challenges	51
3.11 Data quality in Data Integration	52
3.11.1 Data quality definition	52

3.11.2 Data quality framework	53
3.11.3 Evaluating Data quality framework	55
3.12 Adopted Model of Integration	55
Chapter 4: Survey and comparison between Data Integration Systems	57
4.1 Qxchange	59
4.2 SAS	60
4.3 SYBASE	64
4.4 CONNX	66
4.5 ORACLE	67
4.6 Pentaho	68
4.7 SQL Server	70
<u>Chapter 5: Implementation (Case study).</u>	74
<u>Chapter 6: Conclusion and Future Work</u>	88
Appendix A: Pentaho Screen shots	91
Appendix B: Screen shots of Data Integration Systems	112
References	125

List of Figures

Figure (2.1) Data Warehousing architectures summary	8
Figure (2.2) Data Warehousing Layers	12
Figure (2.3) The layout of a Star Schema	17
Figure (2.4) Data Warehouse components	19
Figure (2.5) A Data Warehouse Framework	21
Figure (3.1) Data Integration Framework	37
Figure (3.2) Integrated Schema steps	48
Figure (3.3) Integration Model	56
Figure (5.1) Example of a Transformation	78
Figure (5.2) Example of a Job	78
Figure (5.3) Pentaho toolbar icons	79
Figure (5.4) Collecting the data	80
Figure (5.5) Using Pentaho to convert the inputs to different types of DB	81
Figure (5.6 -a) Scenario 1 of Integration	84
Figure (5.6 -b) Scenario 2 of Integration	85
Figure (A.1) ODBC Connection	93
Figure (A.2) Create new repository	94
Figure (A.3) Create new transformation	95
Figure (A.4) Set connection information For Access database	96
Figure (A.5) Set connection information For Oracle database	97
Figure (A.6) Connection Report	98
Figure (A.7) Select the input Excel file	99
Figure (A.8) Final design of transformation 1	100
Figure (A.9) Final design of transformation 2	101

Figure (A.10) Select attributes from table in the Access DB	102
Figure (A.11) Select attributes from table in the Oracle DB	103
Figure (A.12) Define the meta data	104
Figure (A.13) SQL script part 1	105
Figure (A.14) SQL script part 2	106
Figure (A.15) Run the transformation	107
Figure (A.16) No. of Errors during the integration process	108
Figure (A.17) Time of the integration process	109
Figure (A.18) Adding a select value step in the transformation	110
Figure (A.19) Time of the integration process after adding the selection step	111

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.