hossam maghraby

شبكة المعلومات الجامعية

بسم الله الرحمن الرحيم

hossam maghraby

شبكة المعلومات الجامعية

شبكة المعلومات الجامعية

التوثيق الالكتروني والميكروفيلم

hossam maghraby

شبكة المعلومات الجامعية

جامعة عين شمس

التوثيق الإلكتروني والميكروفيلم

قسم

نقسم بالله العظيم أن المادة التي تم توثيقها وتسجيلها علي هذه الأقراص المدمجة قد أعدت دون أية تغيرات

يجب أن

تحفظ هذه الأقراص المدمجة بعيدا عن الغبار

hossam maghraby

شبكة المعلومات الجامعية

بعض الوثائق الأصلية تالفة

hossam maghraby

شبكة المعلومات الجامعية

بالرسالة صفحات لم ترد بالأصل

B١٢٢٩٤

Cairo University
Institute of Statistical Studies And Research
Department of Computer and Information Science

A Proposed Technique For Metadata Management For Data Warehouse

By

Ahmed Abdou Ahmed Fouda

A thesis submitted to Institute of Statistical Studies And Research Cairo University In Partial Fulfillment of requirements for master degree in computer and information science.

Supervised By

Prof. Dr. Ibrahim Farag Abdel Rahman

Computer Science Department
Faculty Of Computers And Information-Cairo University

Dr. Reda Abdel Wahab Ahmed

Information Technology Department
Faculty Of Computers And Information-Cairo University

July / 2003

Cairo University
Institute of Statistical Studies
And Research
Department of computer and
Information Science

APPROVAL SHEET

A Proposed Technique For Meta Data Management For Data Warehouse

Submitted By

Ahmed Abdou Ahmed Fouda

This thesis for master degree in computer and information science, Department of Computer and Information Science, Institute of Statistical Studies and Research Cairo University, has been approved by:

Name	Signature
Prof. Dr.: Ibrahim Farag Abdel Rahman	.....
Prof. Dr.: Abdel Badie Mohamed Salem	.....
Dr.: Mervat Hassan Ghieth	...Mervat.....

Date: 05/07/2003

Acknowledgment

My sincere thanks for the effort and very useful tips and wise suggestions made by of Prof. Dr. Ibrahim Farag through out the work, his patience and encouragement is very much appreciation.

Also extended thanks and deep appreciation to Dr. Reda Abdel Wahab for his continuous guidance through out the work done, he suggested very useful ideas and gave me a lot of his valuable advises, encouragement and time.

My thanks also extended to my colleagues in Misr Insurance Co. and to my family who made life easy for me to be able to do this work.

I certify that this work has not been accepted in substance for any academic degree and is not being concurrently submitted in candidature for any other degree.

Any portions of this thesis for which I am indebted to other sources are mentioned and explicit reference are given.

Student
Ahmed Abdou Ahmed Fouda

ABSTRACT

Consistent metadata should be captured and stored in a repository to be shared by various users groups and software components. Users need the repository as a consistent documentation in order to effectively and efficiently achieve their particular tasks. On the other hand, software components may produce and consume the information of the repository. Thus, the repository has to provide a structure for fitting its utilization purposes. At the conceptual level, the model of metadata describes the structure of the repository. Each repository should provide a documented and user-extendable model, which easily allows the extension of the system for additional requirements.

Metadata has been identified as a key factor on data warehouse projects. It captures all kind of information necessary to analyze, design, build, use and interpret the data warehouse contents. Standardization of metadata representation and exchange is of great importance in order to spread the use of metadata to enable the interoperability between repositories, and tool integration within data warehousing architectures.

In this work

- We compared two standard models for metadata management, Open Information Model (OIM) and Common Warehouse Model (CWM). In this thesis we preferred to apply CWM due to:

1- OIM is much broader in scope than CWM, OIM is not specific for data warehouse only, but CWM deals more with metadata for data warehousing and provides a framework representing and exchanging metadata.

2- CWM is supported with the metadata object facility (MOF) standard for metadata exchange, whereas OIM is not (i.e. OIM is only descriptive model).

3- OIM supports relational-to-relational mappings only, while CWM is not tied to a particular model.

- We tried to simplify the standard model by using a subset of CWM components consisting of:

  1- Warehouse and online analytical processing (OLAP) component.
  2- Source data component.
  3- Warehouse deployment component.
  4- Transformation component.
  5- Business subject component.
  6- Extraction, transformation and loading (ETL) statistics component.
  7- Query statistics component.

  The proposed simple model keeps the main aspects of the basic model CWM which are integration, data lineage, query generation, supported considerable amount of business metadata and supported all models of transformation. The proposed simple model fulfills business organization requirements. It helps in eliminating redundancy, centralizing documentation, creating a management console and reducing resources required for system support.
- As a result of this study we applied the simple model on collective insurance data warehouse to develop a metadata system for an insurance company. This system focuses on the design and implementation of metadata repository as important part of a metadata management system.

Table Of Content

CHAPTER 1: INTRODUCTION 10

1.1 RESEARCH OBJECTIVE 11

1.2 ORGANIZATION OF THE THESIS 12

CHAPTER 2: DATA WAREHOUSE OVERVIEW 13

2.1 INTRODUCTION AND BACKGROUND 14

2.2 WHAT IS A DATA WAREHOUSE? 15

2.3 MAIN CHARACTERISTICS OF DATA WAREHOUSE DATA 16

2.4 DATA WAREHOUSE BENEFITS 17

2.5 DATA WAREHOUSE COMPONENTS 18

2.6 DEVELOPING DATA WAREHOUSES 21

2.7 MANAGING DATA WAREHOUSES 23

CHAPTER 3: DATA WAREHOUSE MAINTENANCE 25

3.1. EXTRACTION, TRANSFORMATION AND LOADING TOOLS 26

3.2 DATA EXTRACTION FROM ALL SOURCES 28

3.3 DATA MOVEMENT 28

3.4 DATE CLEANING 29

3.5 DATA TRANSFORMATIONS 29

3.6 INITIAL LOAD 30

CHAPTER 4: METADATA AND METADATA MANAGEMENT 31

4.1 METADATA DEFINITION 32

4.2 THE FRAMEWORK OF METADATA IN THE DATA WAREHOUSE 32

4.3 IMPORTANCE OF METADATA 35

4.4 WHY DO WE NEED TO MANAGE METADATA? 35

4.5 DIFFERENT VIEWS OF METADATA 37

4.5.1 BUSINESS USER'S VIEW 37

4.5.2 DATA WAREHOUSE ADMINISTRATOR'S VIEW 38

4.5.3 DSS DEVELOPER'S VIEW 39

4.6 SOURCES OF METADATA 40

4.7 METADATA MANAGEMENT: HISTORICAL APPROACHES 41

4.8 CLASSIFICATION OF DATA WAREHOUSE METADATA 44
4.8.1 TECHNICAL METADATA 46
4.8.2 SEMANTIC METADATA (BUSINESS METADATA) 47
4.9 STRUCTURING THE METADATA ARCHITECTURE 48
4.9.1 CENTRALIZED METADATA REPOSITORY ARCHITECTURE APPROACH 49
4.9.2 DECENTRALIZED METADATA REPOSITORY ARCHITECTURE APPROACH 50
4.9.3 SHARED METADATA REPOSITORY ARCHITECTURE APPROACH 51

CHAPTER 5: METADATA MODELS 53

5.1 METADATA STANDARDS FOR DATA WAREHOUSING: 54
5.1.1 OPEN INFORMATION MODEL (OIM) 55
5.1.2 COMMON WAREHOUSE METAMODEL (CWM) 58
5.1.3 COMPARISON 60
5.1.4 COMPARISON SUMMARY 64
5.2 SIMPLE METADATA MODEL 67

CHAPTER 6: CASE STUDY 88

6.1 INTRODUCTION 89
6.2 EXAMPLE 89

CHAPTER 7: CONCLUSION AND FUTURE RESEARCH WORK .. 121

7.1 CONCLUION 122
7.2 FUTURE RESEARCH WORK 122

REFERENCES 121

LIST OF TABLES

Table 5.1:	Repository Metadata Content	72
Table 5.2:	Key Entity Listing From Traditional ER Model	73
Table 5.3:	Associative Entity Listing from Traditional ER Model	75
Table 5.4:	Subject Area Entity Attributes And Purpose	76
Table 5.5:	Target Table Entity Attributes And Purpose	77
Table 5.6:	Target Column Entity Attributes And Purpose	79
Table 5.7:	Source Column Entity Attributes And Purpose	80
Table 5.8:	Target Domain Entity Attributes And Purpose	81
Table 5.9:	Source Domain Entity Attributes And Purpose	82
Table 5.10:	Source Entity Attributes And Purpose	82
Table 5.11:	Machine Entity Attributes And Purpose	84
Table 5.12:	ETL Process Entity Attributes And Purpose	84
Table 5.13:	Program And OLAP Entity Attributes And Purpose	85
Table 5.14:	ETL Process Statistics Entity Attributes And Purpose	86
Table 5.15:	Query Process Statistics Entity Attributes And Purpose	87
Table 6.1:	Repository Metadata Content	93
Table 6.2:	Repository Entities	96
Table 6.3:	Subject Area Entity Attributes	97
Table 6.4:	Subject Area Map Entity Attributes	97
Table 6.5:	Target Table Entity Attributes	97
Table 6.6:	Target Column Entity Attributes	98
Table 6.7:	Target Table Column Map Entity Attributes	98
Table 6.8:	Source Entity Attributes (files or table)	99
Table 6.9:	Source To Target Column Map Entity Attributes	99
Table 6.10:	Source Column Entity Attributes	100
Table 6.11:	Target Domain Entity Attributes	100
Table 6.12:	Source Domain Entity Attributes	101
Table 6.13:	Source To Target Domain Map Entity Attributes	101
Table 6.14:	Machine Entity Attributes	101
Table 6.15:	ETL Process Entity Attributes	102

Table 6.16: Machine ETL Process Map Entity Attributes 102
Table 6.17: ETL Process Statistics Entity Attributes 102
Table 6.18: ETL Process Source To Target Column Map Entity Attributes 103
Table 6.19: Program And OLAP Entity Attributes 103
Table 6.20: Program Target Table Map Entity Attributes 103
Table 6.21: Query Statistics Target Table Map Entity 104
Table 6.22: Query Statistics Entity Attributes 104