INFORMATION SOCIETY TECHNOLOGIES
(IST)
PROGRAMME

OpenMolGRID

DESCRIPTION OF DATA WAREHOUSING

Contract Reference:	**IST-2001-37238**
Document identifier:	**OpenMolGRID-1-D1.4e-0112-2-1-DescDataWarehousingData**
Date:	**08/01/2004**
Work package:	**WP1: Grid Data Warehousing of Molecular Structure – Property (Activity) Information**
Partner:	**UU**
Lead Partner:	**UU**
Document status:	**APPROVED**
Classification:	**PUBLIC**
Deliverable identifier:	**D1.4e**

Abstract: A general description of the data warehousing process

Delivery Slip

	Name	Partner	Date
From	Damian McCourt	UU	15/09/03
Verified by	WPM	All	02/10/03
Approved by	G.H.F.Diercksen (TC) R.Ferenczi (QE)	OMC CGX	20/10/03 07/04/04

Document Log

Issue	Date	Comment	Author
0-0	06/09/03	First Version	Damian McCourt
1-0	09/09/03	Submitted for review	Damian McCourt
2-0	15/09/03	Submitted for Authorisation	Damian McCourt
2-1	07/01/04	Updated due to the change of the document template (version1.3)	Jean Jing

Document Change Record

Issue	Item	Reason for Change
2-0	Document Status	Document Authorised
2-1	change of document template	The standard template of the document is changed

Files

Files in this section relate to actual storage locations on the BSCW server located at https://hermes.chem.ut.ee/bscw/bscw.cgi. The URL below describes the location on BSCW from the root OpenMolGRID directory

Software Products	User files / URL
Word 2000/XP	OpenMolGRID/Workpackage 1/Deliverables/ OpenMolGRID-1-D1.4e-0112-2-1-DescDataWarehousingD

Project information

Project acronym:	OpenMolGRID
Project full title:	Open Computing GRID for Molecular Science and Engineering
Proposal/Contract no.:	IST-2001-37238
European Commission:	
Project Officer:	Annalisa BOGLIOLO
Address:	European Commission - DG Information Society F2 - Grids for Complex Problem Solving B-1049 Brussels Belgium
Office	BU31 4/79
Phone:	+32 2 295 8131
Fax:	+32 2 299 1749
E-mail	annalisa.bogliolo@cec.eu.int
Project Coordinator:	Mathilde ROMBERG
Address:	Forschungszentrum Jülich GmbH ZAM D-52425 Jülich Germany
Phone:	+49 2461 61 3703
Fax:	+49 2461 61 6656
E-mail	m.romberg@fz-juelich.de

Contents

1. Introduction

1.1. Purpose and scope

The purpose of this document is to provide a general description of the data warehousing process in order to give a better understanding to other partners

1.2. Document Structure

In addition to this section the document contains the following sections:

- Section 2 – a description of data warehousing
- Section 3 – data warehousing in Biology
- Section 4 – references

2. Data Warehousing

A data warehouse [1,2] is generally accepted as a large database that is populated with significant amounts of what is considered as transactional data. The purpose of the data is to support decision-making processes with the aid of data analysis techniques. Data warehouses play a major role in the corporate sector where companies wish to observe trends relating to their business data. A typical example is this of a supermarket. In order to maximise profits, it is important to know which products sell well during which periods of the year and whether or not placing promotions on certain products will be beneficial. In order to help the supermarket executives to reach such decisions, data warehouses are employed. Their role is to periodically take snapshots of data from transactional databases and load them into the data warehouse. Data mining techniques are typically employed to derive information from the warehouse that will ultimately help in the decision making process.

A major difference [3] between a data warehouse and a typical transactional database relates to the frequency in which information is updated. A transactional database is one in which updates are made on a transactional basis. Each new transaction in a transactional database results in the database being updated. This may result in several thousand transactions being carried out every minute depending on the nature of the data. This differs greatly from the data warehousing scenario where data is updated at regular time intervals. Typical time periods for data warehouse updates are daily, weekly and monthly. Fundamental to the role of data warehousing in this context is the time dimensionality or historic nature of the data contained in the warehouse. The data can be viewed as a three-dimensional. Each time *t* represents the state of the business at that time and this can be compared with any other time (or snapshot) that is stored in the warehouse (e.g. *t-1*). In addition to the set time intervals, the data from previous data warehouse updates remains in its current state. It is considered to be non-volatile. Once data is entered into a data warehouse, it can be considered as read only.

Many view a data warehouse as just another database, but this is a misconception. Data warehouses are typically built to support information access, rather than efficiency of storage. Many modern databases are built around the relational model or other models and the main aim of such models is to provide efficient storage while still maintaining reasonable retrieval rates. However, modern information analysis approaches require substantial amounts of pre-processing before they can derive information due to the way in which the information is stored. The pre-processing is typically required because of the approach adopted during the database's creation, for example normalisation in the case of the relational model. Traditionally the efficient use of space has been the deciding factor on the type of model to adopt and often resulted in the relational model being chosen. This ultimately resulted in high levels of normalisation being carried out, making the retrieval of information difficult. With the relatively cheap hardware rates today, large storage capacities are possible at the fraction of the cost when the relational and other approaches were devised and so storage restrictions are less of a factor during database design today. Data Warehousing has emerged from the fact that storage is less restricted and addresses the needs of the data rather than the physical limits.

Data warehouses are typically built from a traditional database management system (DBMS). However, the design of the warehouse does not consider storage efficiency as a limiting problem. Instead it redundantly stores information in a format that more closely relates to the way in which it will be used during information analysis.

Data warehousing therefore not only includes the physical storage of the data, but also the processes surrounding its generation and access.

3. Data Warehousing in Biology

Data warehousing has been heavily used in business areas such as banking, insurance and retail. However, its application in biology is far behind due to several reasons including [4,5]:

1) Data in biology is massive, complex and fast-growing
2) A great amount of biological data is dynamic
3) The concepts and requirements for data warehouse in biology is different from those the traditional application which is market driven

Data warehousing in biology is supposed to provide data that is subject-oriented, integrated, non-volatile, expert-interpreted collection of biological data. Its purpose is to support the analysis of biological data and discovery of knowledge. The type of data can be historical data and metadata that integrate from heterogonous systems and extracted from literature, experiments and various biological databases. The objectives of biological data warehouse are easy access, effective integration and full of quality of data (data required to be clean, accurate, consistent, relevant to data mining and knowledge discovery processes). The operations involved in a biological data warehouse consist of data clearance, extraction, integration and transformation.

The key issues exist in design of data warehouse, which are data modelling, data transformation, integration of data resources and various tools, and maintenance. No exclusion is from biological data warehouse, what is more, it is more challenging than the traditional data warehouse because the biological data is massive, heterogonous, complex, incomplete, noise, errors-trended.

4. References

[1] Chaudhuri, S.; Dayal, U.; An Overview of Data Warehousing and OLAP Technology; http://www.cs.sfu.ca/CourseCentral/459/han/papers/chaudhuri97.pdf

[2] Kimball, R.; The Data Warehouse Toolkit; John Wiley and Sons, Inc., 1999

[3] Jarke,M.; Lenzerini,M.; Vassiliou,Y.; Vassiliadis,P; Fundamentals of Data Warehouses; Springer-Verlag, 2003, ISBN: 3-540-42089-4

[4] Schonbach,C.; Kowalski-Saunders,P.; Brusic,V.; Data Warehousing in Molecular Biology; Briefings in Bioinformatics 1(1): 190-198(2000).

[5] Dubitzky,W.; Krebs,O.; Eils,R.; Minding, OLAPing, and Mining Biological Data: Towards a Data Warehousing Concept in Biology; in press.

[6] Golgarelli,M.; Stefano,R.; A Methodological Framework for Data Warehouse Design; DOLAP 1998

[7] Adamson,C.; Venerable,M.; Data Warehouse Design Solutions. J. Wiley & Sons, Inc. 1998

[8] Inmon,W.H.; Building the Data Warehouse; John Wiley, 1992

[9] Harinarayan,V.; Rajaraman,A; Ullman,J.D.; Implementing Data Cubes Efficiently; Proc. of SIGMOD Conf., 1996

[10] Wu, M.C.; Buchmann,A.P.; Research Issues in Data Warehousing; BTW German Database Conference, 1997