Data Warehouse
NETTAB workshop 2011

Carlo Combi
Elena Gaspari
Alberto Sabaini

Department of Computer Science, University of Verona, Verona, Italy
Outline

1. Introduction
2. Data Warehousing
3. ETL Tools
4. The Multidimensional Data Model
5. Data analysis techniques
6. DW Conceptual Design
7. Data Warehouses and Clinical Domains
8. Summary
Outline

1 Introduction
2 Data Warehousing
3 ETL Tools
4 The Multidimensional Data Model
5 Data analysis techniques
6 DW Conceptual Design
7 Data Warehouses and Clinical Domains
8 Summary
Introduction

Definition

Information = value-increasing asset, needed to effectively plan and control decision-based activities, as diagnosis, therapy planning, monitoring, health care management.

Unfortunately data ≠ information.

Having a huge amount of data makes it difficult to extract useful information.
Data Warehousing process was born to handle this huge amount of data that increased in this last decade.

Mixing together analytical and transactional queries leads to inevitable delays.

**Basic Idea:** to separate *On-Line Analytical Processing* (OLAP) from *On-Line Transactional Processing* (OLTP), building a new collector of information that integrates data from different sources i.e., the Data Warehouse.
Outline

1 Introduction

2 Data Warehousing

3 ETL Tools

4 The Multidimensional Data Model

5 Data analysis techniques

6 DW Conceptual Design

7 Data Warehouses and Clinical Domains

8 Summary
Data Warehousing

Definition

Decision Support System: set of techniques and software tools to extract information from a set of data stored in different sources.

Among the Decision Support Systems, Data Warehouse Systems are those that are more established in the industrial world and could be suitably used also for biomedical data.

Definition

Data Warehousing: a collection of methods, technologies and tools to assist the knowledge worker (clinician, manager, nurse, epidemiologist, technician) to perform data analysis aimed at improving decision making and information assets.
Complaints

- We have a huge amount of data but we can not access it!
- Why people doing the same role are showing significantly different results?
- We want to select, combine and manipulate data in every possible way!
- Show me only what is important!
- Everyone knows that some data are not correct!
Characteristics of the Warehousing process

- Accessibility to users with limited knowledge of computing and data structures.
- Data integration based on a standard model.
- Flexible query to take full advantage of the wealth of information.
- Synthesis to allow targeted and effective analysis.
- Multidimensional representation to provide an intuitive view of information.
- Correctness and completeness of integrated data.
Definition

A Data Warehouse is a collection of support data for decision processes, which is:

- subject-oriented;
- integrated and sound;
- representative of the temporal evolution;
- non-volatile.
Subject-Oriented
Integrated and Sound

The DW relies on multiple sources of heterogeneous data ⇒ the goal is to return a unified vision.
Representative of the temporal evolution

Operational DB
- limited historical content;
- time is not part of keys;
- data updates.

Data Warehouse
- rich historical content;
- time is part of keys;
- data cannot be updated/modified.
Data Warehousing

Non-volatile

In principle data are never deleted from the DW and updates are performed off-line ⇒ read-only.
## Summarizing

<table>
<thead>
<tr>
<th></th>
<th>Operational DB</th>
<th>Data Warehouse</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Users</strong></td>
<td>thousand</td>
<td>hundreds</td>
</tr>
<tr>
<td><strong>Workload</strong></td>
<td>default transactions</td>
<td>ad-hoc queries</td>
</tr>
<tr>
<td><strong>Access</strong></td>
<td>hundreds of records in reading and writing</td>
<td>millions of records especially in reading</td>
</tr>
<tr>
<td><strong>Purpose</strong></td>
<td>depends on the application</td>
<td>decision support</td>
</tr>
<tr>
<td><strong>Data</strong></td>
<td>basic, numeric and alphanumerical</td>
<td>synthesis, numeric</td>
</tr>
<tr>
<td><strong>Integration</strong></td>
<td>by application</td>
<td>by subject</td>
</tr>
<tr>
<td><strong>Quality</strong></td>
<td>integrity</td>
<td>consistency</td>
</tr>
<tr>
<td><strong>Temporal coverage</strong></td>
<td>only current data</td>
<td>current and historical data</td>
</tr>
<tr>
<td><strong>Updates</strong></td>
<td>continue</td>
<td>periodic</td>
</tr>
<tr>
<td><strong>Model</strong></td>
<td>normalized</td>
<td>denormalized and multidimensional</td>
</tr>
<tr>
<td><strong>Development</strong></td>
<td>cascade</td>
<td>iterative</td>
</tr>
</tbody>
</table>
Outline

1. Introduction
2. Data Warehousing
3. ETL Tools
4. The Multidimensional Data Model
5. Data analysis techniques
6. DW Conceptual Design
7. Data Warehouses and Clinical Domains
8. Summary
The role of Extraction, Transformation and Loading tools is to feed a single data source, detailed, comprehensive, and of high quality, which may in turn feed the DW (Reconciliation).

During the feeding process of the DW, reconciliation takes place:
- when the DW is populated for the first time;
- when the DW is periodically updated.

**Stages of the reconciliation process:**
1. extraction
2. cleaning
3. transformation
4. loading
Extraction

The relevant data are extracted from the sources. The choice on what data to extract is based on their quality.

- **Static extraction**: when the DW is populated for the first time (snapshot of operational data).
- **Incremental extraction**: when the DW is periodically updated (captures the changes in the sources since the last update).
Cleaning

Improving the quality of the extracted data:

- duplicate data
- inconsistency between values
- missing data
- misuse of a field
- impossible or incorrect values
- inconsistent values due to different conventions adopted
- inconsistent values due to typing errors
Transformation

Converts the data into a uniform format.

**Feeding of reconciled data:**
- conversion and normalization: modify the format and the unit of measure to standardize data;
- matching: establishes correspondences between equivalent fields from different sources;
- selection: reduces the number of fields and records compared to the sources.

**Feeding of DW:**
- denormalization: replaces the normalization;
- aggregation: makes appropriate summary of data.
Data loading on DW:

- **Refresh**: data are completely rewritten, previous data are replaced;
- **Update**: data are added to DW only when a change occurred in sources.
Outline

1. Introduction
2. Data Warehousing
3. ETL Tools
4. The Multidimensional Data Model
5. Data analysis techniques
6. DW Conceptual Design
7. Data Warehouses and Clinical Domains
8. Summary
Multidimensional Model

The model allows one to represent and query data stored in the Data Warehouse.

- A Data Warehouse is usually built incrementally and is composed of one or more data marts.
- A Data Mart may be composed of several Cubes.

Facts of interest are represented in cubes, where:

- each cell of the cube contains numerical measures that quantify the fact;
- each axis of the cube represents a dimension of interest for the analysis;
- each dimension can be the root of a hierarchy of attributes used to aggregate data.
Figure: On 05/07/2009, 10 patients affected by ischemic heart disease were admitted to the cardiology department.
Hierarchies

Department
- Diagnostic Imaging
- Gastroenterology
- General Surgery
- Haematology
- Neurology
- Cardiology
- Critical Care

Base
- Base1
- Base2

All departments

Pathology
- ... (omitted)
- Heart failure
- Hypertensive heart disease
- Ischaemic heart disease
- Inflammatory heart disease
- Valvular heart disease

Category
- Infectious and parasitic diseases
- Cancer
- Respiratory Diseases
- Mental disorders
- ... (omitted)
- Heart diseases

All pathologies
Slicing and Dicing

Diagram of multidimensional data model with dimensions for Department, Pathology, Date, and Year. Examples include slicing on 05-07-2009 and dicing on Base=Base2 and Category=Heart diseases.
Roll-up and Drill-Down
Outline

1. Introduction
2. Data Warehousing
3. ETL Tools
4. The Multidimensional Data Model
5. Data analysis techniques
6. DW Conceptual Design
7. Data Warehouses and Clinical Domains
8. Summary
Data analysis techniques: Reporting

For users who periodically need to access to information with a fixed structure.
Data analysis techniques: OLAP

- OLAP users are able to actively build a complex analysis session in which each step is a consequence of previous results.
- Flexible interface.
- Easy to use and effective.
Definition

**OLAP session:** navigation path in the analysis cube. It consists of an analysis of one or more facts of interest at different levels of detail. This path is composed by a sequence of queries often formulated by difference with respect to the previous query.

- Each step of the analysis session is made up by the application of an **OLAP** operator (Slice, Dice, Roll-up and Drill-down).
- The result is a multidimensional (sub)cube.
**Data analysis techniques: Data Mining**

**Definition**

Data mining is the process of extracting new knowledge from large data sets by combining methods from statistics and artificial intelligence with database management: clustering, time series analysis, what-if analysis, association rules, pattern recognition, probabilistic reasoning, classification.
Outline

1. Introduction
2. Data Warehousing
3. ETL Tools
4. The Multidimensional Data Model
5. Data analysis techniques
6. DW Conceptual Design
7. Data Warehouses and Clinical Domains
8. Summary
DFM is a graphical conceptual model for data marts, designed to:

- efficiently support the conceptual design;
- create an environment where the user can specify queries in an easy way;
- enable the dialogue between designer and end user to refine the specified requirements;
- create a stable platform to derive the design at the logical level;
- return expressive and unambiguous documentation.

The conceptual representation generated by DFM is a set of fact schemas. The basic elements modeled by fact schemas are: facts, measures, dimensions and hierarchies.
DFM: Basic Constructs

- **Fact**: concept of interest for decision making. It represents a set of events that occur in the considered (clinical/healthcare) domain and evolve over time.
- **Measure**: is a numeric property of a fact and describes a quantitative aspect of interest for the analysis.
- **Dimension**: is a property of a fact with a finite domain.

![Diagram of HOSPITALIZATION concept]
DFM: Basic Constructs

- Dimensional attributes: dimensions and any other related attribute.
- Hierarchy: directed tree whose nodes are dimensional attributes and whose edges represent associations many-to-one between pairs of attributes.
Events

Definition

A **primary event** is a particular occurrence of a fact, identified by a tuple consisting of a value for each dimension and each measure.

A **secondary event** is the aggregation of all corresponding primary events, according to the values of some dimensional attributes not being root of the hierarchy they belong to. Each secondary event is associated with a value for each measure.
Additivity

Aggregation requires to define a suitable operator to compose measure values of the primary events to calculate measure values to be associated with secondary events.

There are three categories of measures:

- **flow measurements**: refer to a period after which they are assessed cumulatively;
- **level measurements**: are evaluated at particular time instants;
- **unit measurements**: are evaluated at particular time instants, but they are expressed in relative terms.

<table>
<thead>
<tr>
<th>measure type</th>
<th>Temporal hierarchies</th>
<th>Non-Temporal hierarchies</th>
</tr>
</thead>
<tbody>
<tr>
<td>flow measurements</td>
<td>SUM, AVG, MIN, MAX</td>
<td>SUM, AVG, MIN, MAX</td>
</tr>
<tr>
<td>level measurements</td>
<td>AVG, MIN, MAX</td>
<td>SUM, AVG, MIN, MAX</td>
</tr>
<tr>
<td>unit measurements</td>
<td>AVG, MIN, MAX</td>
<td>AVG, MIN, MAX</td>
</tr>
</tbody>
</table>
Outline

1 Introduction
2 Data Warehousing
3 ETL Tools
4 The Multidimensional Data Model
5 Data analysis techniques
6 DW Conceptual Design
7 Data Warehouses and Clinical Domains
8 Summary
Clinical studies before a drug is put on the market are not able to guarantee the absence of side effects of commercial drugs.

The surveillance through spontaneous reporting for products already on the market can detect the majority of side effects of drugs.

The reporting of suspected adverse drug reactions feed the database of the World Health Vigibase.

The European equivalent, EudraVigilance, was instituted in 2001.
EudraVigilance and Vigibase, however, collect only a portion of the Italian reporting forms and their content.

The Italian report forms were initially stored in the database of the GIF group and later in the National Network of Pharmacovigilance (RNF), with a partial overlap.
There was the need to develop an integrated database for storing data previously described, along with a procedure for its feeding (continuous and periodic) and the tools to efficiently extract derived aggregated values.

A data warehouse system has been developed for the analysis of adverse drug reactions reported in Italy.

The project has been realized in collaboration with the Pharmacovigilance Regional Centre of the Veneto.
To develop such database, specific steps must performed (In this example we will focus only on the realization of the source level and the feeding level):

- Data Analysis
- Designing the system architecture
- Integration and feeding strategy
- Data Reconciliation
- ETL Phase
Available databases

- drug database Codifa2000
- the databases for the MedDRA, WHOART, and ATC terminology
- GIF and RNF databases

![Diagram showing data warehouse connections and databases](image-url)
The system architecture
Data Reconciliation

During the merge of the data from the GIF and RNF we found a problem in the choice of a global identifier for the report:

- The paper reporting forms are identified by a serial number and the initials of the region
- The RNF reporting forms are identified by a different sequence number managed by the Ministry of Health

We decided to create a third identifier of 13 characters structured as follows:

<table>
<thead>
<tr>
<th>Local schema</th>
<th>Global schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region</td>
<td>Number</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GIF</th>
<th>RNF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local schema</td>
<td>Global schema</td>
</tr>
<tr>
<td>Region</td>
<td>Number</td>
</tr>
<tr>
<td>VEN</td>
<td>26174</td>
</tr>
<tr>
<td></td>
<td>GIF-VEN-26174</td>
</tr>
<tr>
<td></td>
<td>Number</td>
</tr>
<tr>
<td></td>
<td>92143</td>
</tr>
<tr>
<td></td>
<td>AIFA-00092143</td>
</tr>
</tbody>
</table>
Data Reconciliation

Since the two archives have collected reporting forms filled in overlapping periods, it was also necessary to distinguish duplicates resulting from their merge. The reconciled level was then populated with all the reporting forms that are unique in the database of RNF as well as with those present in both databases, applying a selection to avoid duplicated forms.

![Diagram showing the reconciliation process between GIF and RNF archives from 1988 to 2007. The diagram illustrates the collection of reports and the assignment of dual coding.](Image)
The Regional Centre for Neonatal Metabolic Diseases (CRMMN) of Verona performs newborn screening for major hereditary metabolic and endocrine diseases:

- galactosemia
- PKU (Phenylketonuria syndrome)
- Biotinidase deficiency
- deficiency of glucose-6-phosphate-dehydrogenase (G6PD)
- congenital hypothyroidism (IC)
- Leucine or maple syrup urine disease (MSUD)
- congenital adrenal hyperplasia (CAH)

The screening is carried out on samples of blood taken from the heel of the newborn after 48 hours of birth.
Qualitative assessment Fact schema

- The fact represented in this schema is the outcome of qualitative assessments, which records the result of the clinical validation of each quality examination carried out.
- The measures of interest are the number of times each outcome occurred of qualitative assessments (number of NO, VP, FN, DUB, etc.).
Doubt outcomes (DUB) grouped by weight

In order to identify differences in weight groups, possibly highlighting different risk levels, the data mart represents the number of tests that indicate suspect values, according to the weight of the newborn.
Total outcome number

We can see the total number of results by selecting only the dimensions date of birth, gestational age, type of examination, number of control and outcome. A total of 808,625 screening examinations has been performed.
Total DUB recall for patient born in 2010

Selecting only the DUB results and focusing only on patients born in 2010, it can be noticed that there were 1,495 recalled.

<table>
<thead>
<tr>
<th></th>
<th>Birth Date</th>
<th>Gestational Age</th>
<th>Exam Type</th>
<th>Control Number</th>
<th>Outcome</th>
<th>Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>All Patients.GestAges</td>
<td>All ExamTypes</td>
<td>All ControlNumbers</td>
<td>All Outcomes</td>
<td>358.073</td>
<td>DUB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.495</td>
</tr>
</tbody>
</table>
Drill-down on the type of examination

By performing a drill down peration on the type of examination, it is possible to highlight how the quality tests carried out are grouped when the outcome is DUB.

<table>
<thead>
<tr>
<th>Birth Date</th>
<th>Gestational Age</th>
<th>Exam Type</th>
<th>Control Number</th>
<th>Outcome</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>All Patiente.GestAges</td>
<td>All ExamTypes</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>1.495</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BIOT</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G6PD</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>188</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LEU</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N170HPQUAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>234</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NT4QUAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>477</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NTSHSQual</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>549</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PKU</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>35</td>
</tr>
</tbody>
</table>
Focusing on the congenital adrenal hyperplasia

By selecting the exam type N170HPQUAL (qualitative exam for revealing the presence of CAH) and performing a drill-replace on the two members of the gestational age class level, we can see how the number of DUB for congenital adrenal hyperplasia (CAH) is related to the gestational age.

<table>
<thead>
<tr>
<th>Birth Date</th>
<th>Gestational Age</th>
<th>Exam Type</th>
<th>Control Number</th>
<th>Outcome</th>
<th>Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>26</td>
<td>N170HPQUAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>27</td>
<td>N170HPQUAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>N170HPQUAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>N170HPQUAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>N170HPQUAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>31</td>
<td>N170HPQUAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>N170HPQUAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>33</td>
<td>N170HPQUAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>34</td>
<td>N170HPQUAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>N170HPQUAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>N170HPQUAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>37</td>
<td>N170HPQUAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>38</td>
<td>N170HPQUAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>39</td>
<td>N170HPQUAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>N170HPQUAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>41</td>
<td>N170HPQUAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>42</td>
<td>N170HPQUAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>1</td>
</tr>
</tbody>
</table>
Focusing on the congenital adrenal hyperplasia

The peak of DUB for CAH coincides with 36 weeks of gestation, but fell considerably with higher gestational ages.
Using weight instead of age

By choosing to display the hierarchy of the patient’s weight, instead of the gestational age, we can see the change in the number of DUB for CAH according to the weight category.

<table>
<thead>
<tr>
<th>Birth Date</th>
<th>Weight</th>
<th>Exam Type</th>
<th>Control Number</th>
<th>Outcome</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>empty</td>
<td>N170HPQUAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>12</td>
</tr>
<tr>
<td>&lt;1800</td>
<td></td>
<td>N170HPQUAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>102</td>
</tr>
<tr>
<td>1800-2500</td>
<td></td>
<td>N170HPQUAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>50</td>
</tr>
<tr>
<td>&gt;2500</td>
<td></td>
<td>N170HPQUAL</td>
<td>All ControlNumbers</td>
<td>DUB</td>
<td>70</td>
</tr>
</tbody>
</table>
Using weight instead of age

- 2010. <1800 (102)
- 2010. 1800-2500 (50)
- 2010. >2500 (70)
- 2010. empty (12)
We want to observe the pattern of three consecutive purchases of drugs, even the same product, for a patient;
we consider a time span of 30 days as the maximum delay between the purchase of drug and the next one, and of 1 day as the minimum delay;
The analysis focuses on the three most requested drugs for all patients in 2005.
Drug patterns cube

This cube represents patterns of subsequent prescriptions for a given patient, within thirty days between a purchase of a drug and the next one.
Most requested drug triplets

Viewing the 15 consecutive prescriptions at the second level ATC that verifies the most requested pattern, we note that the one that contains three times “antacids, peptic ulcer and antimeteorics”, exceeds any other triple with 20,384 requests.
Most requested drug triplets

Selecting “antacids, peptic ulcer and antimeteorics” and moving along the hierarchy to the fourth ATC level, we note that the most popular triplet consists of three identical descriptions of the “acid pump inhibitors”, with more than 12,000 instances.
Most requested drug triplets

T1.1 (omeprazol) is the triplet most purchased, at the description ATC level. We select only this triplet to reach the level description.
Most requested drug triplets

The drug “ANTRA 20*14 CPS R.M. 20 MG” at the Description level is in the most requested triplet, which has been requested 1,117 times.
Most requested drug triplets

The triplet was requested from 476 different patients, with an average duration of the pattern of 39.493 days, and the pharmaceutical company earned 170,523.72 euros.

<table>
<thead>
<tr>
<th>Minsan Cod Drugs 1 atc</th>
<th>Minsan Cod Drugs 2 atc</th>
<th>Minsan Cod Drugs 3 atc</th>
<th>Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>028245090</td>
<td>028245090</td>
<td>028245090</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cardinality</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.117</td>
</tr>
</tbody>
</table>
Outline

1 Introduction

2 Data Warehousing

3 ETL Tools

4 The Multidimensional Data Model

5 Data analysis techniques

6 DW Conceptual Design

7 Data Warehouses and Clinical Domains

8 Summary
Summary

- Having a huge amount of data makes it difficult to extract useful information.
- Data Warehouse System are the Decision Support System more established in the industrial world.
- As we have seen, this technology can be successfully used in the medical domain and allows:
  - Integrating different data sources, for a more complete and extended analysis
  - Finding unknown recurring pattern in the patients data
  - Discovering time patterns and improving the resource management