Data Warehousing

01

Data Warehouse Architecture Overview

Notice

- Author
  - João Moura Pires (jmp@di.fct.unl.pt)

- This material can be freely used for personal or academic purposes without any previous authorization from the author, only if this notice is maintained with.

- For commercial purposes the use of any part of this material requires the previous authorization from the author.
Bibliography

- Many examples are extracted and adapted from

Table of Contents

- Decision Support Systems

- Historic perspective

- The need for a new approach

- DW Reference Model
According to Keen and Scott Morton (1978), the concept of decision support has evolved from two main areas of research:

- the theoretical studies of organizational decision making done at the Carnegie Institute of Technology during the late 1950s and early 1960s;
- the technical work on interactive computer systems, mainly carried out at the Massachusetts Institute of Technology in the 1960s.

It is considered that the concept of DSS became an area of research of its own in the middle of the 1970s, before gaining in intensity during the 1980s.

In the middle and late 1980s, executive information systems (EIS), group decision support systems (GDSS), and organizational decision support systems (ODSS) evolved from the single user and model-oriented DSS.

Beginning in about 1990, data warehousing and on-line analytical processing (OLAP) began broadening the realm of DSS.
Decision Support Systems - Fields

- It is clear that DSS belong to an environment with multidisciplinary foundations, including (but not exclusively):
  - Database research;
  - Artificial intelligence;
  - Human-computer interaction;
  - Simulation methods;
  - Software engineering;
  - Telecommunications.

Decision Support Systems - Many approaches

- At the conceptual level, Power (2002) differentiates DSSs:
  - A **model-driven DSS** emphasizes access to and manipulation of a statistical, financial, optimization, or simulation model. Model-driven DSS use data and parameters provided by DSS users to aid decision makers in analyzing a situation, but they are not necessarily data intensive.
  - A **communication-driven DSS** supports more than one person working on a shared task; (collaborative tools)
  - A **data-driven DSS** or data-oriented DSS emphasizes access to and manipulation of a time series of internal company data and, sometimes, external data.
  - A **document-driven DSS** manages, retrieves and manipulates unstructured information in a variety of electronic formats.
  - A **knowledge-driven DSS** provides specialized problem solving expertise stored as facts, rules, procedures, or in similar structures.

[Wikipedia - DSS]
Hättenschwiler (1999) identifies the following DSS users with different roles or functions in the decision making process:

- Decision maker
- Advisors
- Domain experts
- System experts
- Data collectors

**What is about this course?**

- **Data Warehouse (DW)**
  
  For now, let's think about DW as the Warehouse where all the important data is integrated and stored, including historical data, for future support of Data Analysis and Decision making.

- **On-line analytical processing (OLAP)**
  
  It is an approach to quickly provide the answer to analytical queries that are dimensional in nature. The data comes from the DW.
Decisions in the context of an organization?

- **Strategic decisions (long term)**
  - **Examples**
    - Analyzing the actual pattern buying to develop a new product;
    - Deciding the creation of a new university course.

- **Short term decisions - tactical decisions**
  - **Examples**
    - Changing the volumes of components to buy to our suppliers;
    - Analyzing the factors affecting the unsuccessful results of so many students.

Some analysis patterns used by OLAP users

- **Summarizing and aggregation of large amount of data**
- **Filtering, sorting, ranking**
- **Comparisons of different sets of data**
- **Search for outliers**
- **Analysis and discovery of patterns**
- **Analysis of tendencies in the data**
Who is DW and OLAP users?

- DSS analyst is a businessperson first and foremost, and a technician second. The primary job of the DSS analyst is to define and discover information used in corporate decision-making.

- The DSS analyst has a mindset of “Give me what I say I want, then I can tell you what I really want.”. In other words, the DSS analyst operates in a mode of discovery.

- This has a profound effect on the way the data warehouse is developed and on how systems using the data warehouse are developed.

Data Warehouse Architecture Overview - Historical perspective
Evolution

Historical perspective

Two different needs

- Running the organization
  - Operational Data
  - Transactional Data

- Analyzing the organization performance
  - Aggregating Data
  - Comparing Data
The “Extract” Program

Historical perspective

The “Extract” Program became very popular

- Because extract processing can move data out of the way of high-performance online processing, there is no conflict in terms of performance when the data needs to be analyzed en masse.

- When data is moved out of the operational, transaction-processing domain with an extract program, a shift in control of the data occurs. The end user then owns the data once he or she takes control of it.

[Inmon,2002]
The “Extract” Program became so popular that ...

Historical perspective

Problems with this pattern of many extracting programs

- Lack of Data credibility
- Productivity
- Inability to transform data into information

[Inmon, 2002]
Problems with this pattern of many extracting programs

- **Lack of Data credibility**
  - No time basis of data
    - Different moments for data extraction
  - The algorithmic differential of data
    - Analysis based on different data sets
  - The levels of extraction
    - extract and extract and extract....
  - The problem of external data
    - External data come into the system loosing their sources
  - No common source of data from the beginning
  - Different original sources

---

[Immon, 2002]
Problems with this pattern of many extracting programs

- Problems with productivity
  - Locate and analyze the data for the report
  - Compile data for the project
  - Get resources to accomplish these two tasks

- Different technical skills to access data across the organization
- Naming and concepts problems (ambiguity, etc)
- Data has to be *normalized* and integrated
- The process may be repeated for each new report need

Historical perspective

[Inmon, 2002]
Problems with this pattern of many extracting programs

- Problems with productivity

Locating the data requires looking at *lots* of files.

[Inmon, 2002]

---

Problems with this pattern of many extracting programs

- Problems with productivity

Lots of extract programs, each customized, have to cross many technological barriers.

[Inmon, 2002]
Problems with this pattern of many extracting programs

- Inability to transform data into information

The applications were never constructed with integration in mind, and they are no easier for the DSS analyst to decipher than they are for anyone else.

- There is not enough historical data stored in the applications to meet the needs of the DSS request.

How has account activity been different this year from each of the past five years for the financial institution?
The need for a different approach

Operational Systems (most - OLTP)

- OLTP – On Line Transaction Processing
  - Systems that support the running activities of the organization
  - Examples:
    - Point of sale in stores;
    - ATM and Bank operations
    - e-commerce (amazon, iTunes, etc)
  - Some characteristics:
    - Thousand of operations per second
    - Repeated operations dealing with small amounts of data (insert, update, remove)
    - Real Time
DW and OLAP systems

- **OLAP – On Line Analytical Processing**
  - Systems that provide the users the necessary capabilities to analyze many and different aspects of organization activities and its performance.

- **Examples**
  - How well certain product is selling in different regions? How well is the evolution in the market from its introduction?
  - Which are the top ten selling product in each region? and globally?

- **Some characteristics:**
  - Small number of queries (per day), when compared with OLTP systems
  - Large amount of data processed in each query, in order to obtain a small output.
  - It is hard to predict the queries and in general they are much more diverse, when compared with OLTP systems

---

Analytic versus Operational

The users of an OLTP system are *running* the wheels of the organization.

The users of a data warehouse are *watching* the wheels of the organization

[Kimball,2002]
Analytic versus Operational - Primitive and Derived Data

<table>
<thead>
<tr>
<th>PRIMITIVE DATA/OPERATIONAL DATA</th>
<th>DERIVED DATA/DSS DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>• application oriented</td>
<td>• subject oriented</td>
</tr>
<tr>
<td>• detailed</td>
<td>• summarized, otherwise refined</td>
</tr>
<tr>
<td>• accurate, as of the moment of access</td>
<td>• represents values over time, snapshots</td>
</tr>
<tr>
<td>• serves the clerical community</td>
<td>• serves the managerial community</td>
</tr>
<tr>
<td>• can be updated</td>
<td>• is not updated</td>
</tr>
<tr>
<td>• run repetitively</td>
<td>• run heuristically</td>
</tr>
<tr>
<td>• requirements for processing understood a priori</td>
<td>• requirements for processing not understood a priori</td>
</tr>
<tr>
<td>• compatible with the SDLC</td>
<td>• completely different life cycle</td>
</tr>
<tr>
<td>• performance sensitive</td>
<td>• performance relaxed</td>
</tr>
<tr>
<td>• accessed a unit at a time</td>
<td>• accessed a set at a time</td>
</tr>
<tr>
<td>• transaction driven</td>
<td>• analysis driven</td>
</tr>
<tr>
<td>• control of update a major concern in terms of ownership</td>
<td>• control of update no issue</td>
</tr>
<tr>
<td>• high availability</td>
<td>• relaxed availability</td>
</tr>
<tr>
<td>• managed in its entirety</td>
<td>• managed by subsets</td>
</tr>
<tr>
<td>• nonredundancy</td>
<td>• redundancy is a fact of life</td>
</tr>
<tr>
<td>• static structure; variable contents</td>
<td>• flexible structure</td>
</tr>
<tr>
<td>• small amount of data used in a process</td>
<td>• large amount of data used in a process</td>
</tr>
<tr>
<td>• supports day-to-day operations</td>
<td>• supports managerial needs</td>
</tr>
<tr>
<td>• high probability of access</td>
<td>• low, modest probability of access</td>
</tr>
</tbody>
</table>

The need for a different approach

Analytic versus Operational - Patterns of utilization

The different patterns of hardware utilization in the different environments.
Analytic versus Operational - Separated Environments

The need for a different approach

Data Warehouse Architecture Overview

DW Reference Model
A data warehouse is an analytical database that is used as the foundation of a decision support system. It is designed for large volumes of read-only data, providing intuitive access to information that will be used in making decisions.

A data warehouse is created as ongoing commitment by the organization to ensure the appropriate data is available to the appropriate end user at the appropriate time”

The Data Environment

**The Data Environment**

![Diagram](image1)

**The Data Environment - Data Integration**

![Diagram](image2)
The Data Environment - Reference Model

Data Warehouse Architecture Overview

Further Reading and Summary
Further Reading and Summary

- **Readings**
  - **Wikipedia**: Look for Data Warehouse, OLAP, OLTP, DSS
  - [Inmon, 2002] - from page 1 to 30.

- **What you should know:**
  - The concept of Decision Support System, its evolution, the different types of DSS and the related Scientific areas.
  - DW and OLAP viewed as Data-Driven DSS. The justification to the actual importance of DW and OLAP in the DSS world.
  - Basic understand of the DW reference architecture
  - Fundamental differences from OLTP and OLAP systems, models, use, and users
  - Some analysis patterns used by OLAP users.