Knowledge-Based Systems
IS430

Data Warehousing

Mostafa Z. Ali
mzali@just.edu.jo
Learning Objectives

• Understand the basic definitions and concepts of data warehouses
• Understand data warehousing architectures
• Describe the processes used in developing and managing data warehouses
• Explain data warehousing operations
• Explain the role of data warehouses in decision support
Learning Objectives

• Explain data integration and the extraction, transformation, and load (ETL) processes
• Describe real-time (active) data warehousing
• Understand data warehouse administration and security issues
Data Warehousing Definitions and Concepts

• Data warehouse

A physical repository where relational data are specially organized to provide enterprise-wide, cleansed data in a standardized format
Data Warehousing Definitions and Concepts

• Characteristics of data warehousing
  – Subject oriented
  – Integrated
  – Time variant (time series)
  – Nonvolatile
  – Web based
  – Relational/multidimensional
  – Client/server
  – Real-time
  – Include metadata
Data Warehousing Definitions and Concepts

• **Data mart**
  A departmental data warehouse that stores only relevant data

• **Dependent data mart**
  A subset that is created directly from a data warehouse

• **Independent data mart**
  A small data warehouse designed for a strategic business unit or a department
Data Warehousing Definitions and Concepts

• **Operational data stores (ODS)**
  A type of database often used as an interim area for a data warehouse, especially for customer information files

• **Oper marts**
  An operational data mart. An oper mart is a small-scale data mart typically used by a single department or functional area in an organization
Data Warehousing Definitions and Concepts

• **Enterprise data warehouse (EDW)**
  A technology that provides a vehicle for pushing data from source systems into a data warehouse

• **Metadata**
  Data about data. In a data warehouse, metadata describe the contents of a data warehouse and the manner of its use
Data Warehousing Process Overview

• Organizations continuously collect data, information, and knowledge at an increasingly accelerated rate and store them in computerized systems.

• The number of users needing to access the information continues to increase as a result of improved reliability and availability of network access, especially the Internet.
Data Warehousing
Process Overview

FIGURE 5.1 Data Warehouse Framework and Views
Data Warehousing Process Overview

• The major components of a data warehousing process
  – Data sources
  – Data extraction
  – Data loading
  – Comprehensive database
  – Metadata
  – Middleware tools
Data Warehousing Architectures

• Three parts of the data warehouse
  – The data warehouse that contains the data and associated software
  – Data acquisition (back-end) software that extracts data from legacy systems and external sources, consolidates and summarizes them, and loads them into the data warehouse
  – Client (front-end) software that allows users to access and analyze data from the warehouse
Data Warehousing Process Overview

FIGURE 5.2 Architecture of a Three-Tier Data Warehouse
Data Warehousing
Process Overview

FIGURE 5.3  Architecture of a Two-Tier Data Warehouse
Data Warehousing
Process Overview

FIGURE 5.4 Architecture of Web-Based Data Warehousing

- Web browser
- Web server
- Web pages
- Application server
- Data warehouse
- Client
- Internet, intranet, and/or extranet
Data Warehousing Architectures

• Issues to consider when deciding which architecture to use:
  – *Which database management system (DBMS) should be used?*
  – *Will parallel processing and/or partitioning be used?*
  – *Will data migration tools be used to load the data warehouse?*
  – *What tools will be used to support data retrieval and analysis?*
Data Warehousing
Process Overview

FIGURE 5.5 Alternative Data Warehouse Architectures

Source Systems

ET L

Central metadata

Central data warehouse

Data analysis

Data mart

Data analysis

RDBMS

5.5a Enterprise Data Warehousing Architecture
Data Warehousing
Process Overview

FIGURE 5.5b  Data Mart Architecture
Data Warehousing Process Overview

FIGURE 5.5c  Hub-and-Spoke Data Mart Architecture
Data Warehousing
Process Overview

FIGURE 5.5d  Enterprise Warehouse and Operational Data Store
Data Warehousing
Process Overview

FIGURE 5.5e  Distributed Data Warehouse Architecture
Data Warehousing Process Overview

![Data Warehousing Process Diagram](image)

<table>
<thead>
<tr>
<th>Independent Data Marts</th>
<th>Leave Data Where it Lies</th>
<th>Dependent Data Marts</th>
<th>Centralized Integrated Data With Direct Access</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pros</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Easy to Build Organizationally</td>
<td>No need for ETL</td>
<td>Allows easier customization of user interfaces &amp; reports</td>
<td>Business Enterprise view</td>
</tr>
<tr>
<td>Easy to Build Technically</td>
<td>No need for separate platform</td>
<td></td>
<td>Design consistency &amp; data quality</td>
</tr>
<tr>
<td><strong>Cons</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Business Enterprise view unavailable</td>
<td>Only viable for low volume</td>
<td>Business Enterprise view challenging</td>
<td>Requires corporate leadership and vision</td>
</tr>
<tr>
<td>Redundant data costs</td>
<td>Meta data issues</td>
<td>Redundant data costs</td>
<td></td>
</tr>
<tr>
<td>High ETL costs</td>
<td>Network bandwidth and join complexity issues</td>
<td>High DBA and operational costs</td>
<td></td>
</tr>
<tr>
<td>High App costs</td>
<td>Workload typically placed on workstation</td>
<td>Data latency</td>
<td></td>
</tr>
<tr>
<td>High DBA and operational costs</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**FIGURE 5.6** Alternative Architectures for Data Warehousing Efforts
Data Warehousing
Process Overview

FIGURE 5.7 Teradata Corp.’s Enterprise Data Warehouse
Ten factors that potentially affect the architecture selection decision:

1. Information interdependence between organizational units
2. Upper management’s information needs
3. Urgency of need for a data warehouse
4. Nature of end-user tasks
5. Constraints on resources
6. Strategic view of the data warehouse prior to implementation
7. Compatibility with existing systems
8. Perceived ability of the in-house IT staff
9. Technical issues
10. Social/political factors
Data Integration and the Extraction, Transformation, and Load (ETL) Process

• Data integration

Integration that comprises three major processes: data access, data federation, and change capture. When these three processes are correctly implemented, data can be accessed and made accessible to an array of ETL and analysis tools and data warehousing environments.
Data Integration and the Extraction, Transformation, and Load (ETL) Process

• **Enterprise application integration (EAI)**
  
  A technology that provides a vehicle for pushing data from source systems into a data warehouse
Data Integration and the Extraction, Transformation, and Load (ETL) Process

• Enterprise information integration (EII)

An evolving tool space that promises real-time data integration from a variety of sources, such as relational databases, Web services, and multidimensional databases
Data Integration and the Extraction, Transformation, and Load (ETL) Process

- **Extraction, transformation, and load (ETL)**
  A data warehousing process that consists of extraction (i.e., reading data from a database), transformation (i.e., converting the extracted data from its previous form into the form in which it needs to be so that it can be placed into a data warehouse or simply another database), and load (i.e., putting the data into the data warehouse)
Data Integration and the Extraction, Transformation, and Load (ETL) Process

FIGURE 5.8 The ETL Process
Data Integration and the Extraction, Transformation, and Load (ETL) Process

• Issues affect whether an organization will purchase data transformation tools or build the transformation process itself
  – Data transformation tools are expensive
  – Data transformation tools may have a long learning curve
  – It is difficult to measure how the IT organization is doing until it has learned to use the data transformation tools
Data Integration and the Extraction, Transformation, and Load (ETL) Process

• Important criteria in selecting an ETL tool
  – Ability to read from and write to an unlimited number of data source architectures
  – Automatic capturing and delivery of metadata
  – A history of conforming to open standards
  – An easy-to-use interface for the developer and the functional user
Data Warehouse Development

• Direct benefits of a data warehouse
  – Allows end users to perform extensive analysis
  – Allows a consolidated view of corporate data
  – Better and more timely information A
  – Enhanced system performance
  – Simplification of data access
Data Warehouse Development

• Indirect benefits result from end users using these direct benefits
  – Enhance business knowledge
  – Present competitive advantage
  – Enhance customer service and satisfaction
  – Facilitate decision making
  – Help in reforming business processes
Data Warehouse Development

• Data warehouse vendors
  – Six guidelines to considered when developing a vendor list:
    1. Financial strength
    2. ERP linkages
    3. Qualified consultants
    4. Market share
    5. Industry experience
    6. Established partnerships
Data Warehouse Development

• Data warehouse development approaches
  – Inmon Model: EDW approach
  – Kimball Model: Data mart approach

• Which model is best?
  – There is no one-size-fits-all strategy to data warehousing
  – One alternative is the hosted warehouse
Data Warehouse Development

• Data warehouse structure: The Star Schema
  – Dimensional modeling
    A retrieval-based system that supports high-volume query access
  – Dimension tables
    A table that address *how* data will be analyzed
Data Warehouse Development

Figure 5.9 Star Schema

Star Schema Example
Automobile Insurance Data Warehouse

Driver

Automotive

Claim Information

Location

Time

Dimension:
How data will be accessed (e.g. by location, time period, type of automobile or driver)

Facts:
Central table that contains summarized (usually) information; also contains foreign keys to access each dimension table
Data Warehouse Development

• **Grain**
  A definition of the highest level of detail that is supported in a data warehouse

• **Drill-down**
  The process of probing beyond a summarized value to investigate each of the detail transactions that comprise the summary
Data Warehouse Development

• Data warehousing implementation issues
  – Implementing a data warehouse is generally a massive effort that must be planned and executed according to established methods
  – There are many facets to the project lifecycle, and no single person can be an expert in each area
Data Warehouse Development

Eleven major tasks that could be performed in parallel for successful implementation of a data warehouse (Solomon, 2005):

1. Establishment of service-level agreements and data-refresh requirements
2. Identification of data sources and their governance policies
3. Data quality planning
4. Data model design
5. ETL tool selection
6. Relational database software and platform selection
7. Data transport
8. Data conversion
9. Reconciliation process
10. Purge and archive planning
11. End-user support
Data Warehouse Development

• Some best practices for implementing a data warehouse (Weir, 2002):
  – Project must fit with corporate strategy and business objectives
  – There must be complete buy-in to the project by executives, managers, and users
  – It is important to manage user expectations about the completed project
  – The data warehouse must be built incrementally
  – Build in adaptability
Data Warehouse Development

• Some best practices for implementing a data warehouse (Weir, 2002):
  – The project must be managed by both IT and business professionals
  – Develop a business/supplier relationship
  – Only load data that have been cleansed and are of a quality understood by the organization
  – Do not overlook training requirements
  – Be politically aware
Data Warehouse Development

• Failure factors in data warehouse projects:
  – Cultural issues being ignored
  – Inappropriate architecture
  – Unclear business objectives
  – Missing information
  – Unrealistic expectations
  – Low levels of data summarization
  – Low data quality
Data Warehouse Development

- Issues to consider to build a successful data warehouse:
  - Starting with the wrong sponsorship chain
  - Setting expectations that you cannot meet and frustrating executives at the moment of truth
  - Engaging in politically naive behavior
  - Loading the warehouse with information just because it is available
Data Warehouse Development

• Issues to consider to build a successful data warehouse:
  – Believing that data warehousing database design is the same as transactional database design
  – Choosing a data warehouse manager who is technology oriented rather than user oriented
  – Focusing on traditional internal record-oriented data and ignoring the value of external data and of text, images, and, perhaps, sound and video
Data Warehouse Development

• Issues to consider to build a successful data warehouse:
  – Delivering data with overlapping and confusing definitions
  – Believing promises of performance, capacity, and scalability
  – Believing that your problems are over when the data warehouse is up and running
  – Focusing on ad hoc data mining and periodic reporting instead of alerts
Data Warehouse Development

– Implementation factors that can be categorized into three criteria
  • Organizational issues
  • Project issues
  • Technical issues

– User participation in the development of data and access modeling is a critical success factor in data warehouse development
Data Warehouse Development

• Massive data warehouses and scalability
  – The main issues pertaining to scalability:
    • The amount of data in the warehouse
    • How quickly the warehouse is expected to grow
    • The number of concurrent users
    • The complexity of user queries
  – Good scalability means that queries and other data-access functions will grow linearly with the size of the warehouse
Real-Time Data Warehousing

- Real-time (active) data warehousing
  The process of loading and providing data via a data warehouse as they become available
Real-Time Data Warehousing

• Levels of data warehouses:
  1. Reports what happened
  2. Some analysis occurs
  3. Provides prediction capabilities,
  4. Operationalization
  5. Becomes capable of making events happen
Real-Time Data Warehousing

**FIGURE 5.10** Enterprise Decision Evolution
Real-Time Data Warehousing

**FIGURE 5.11** The Teradata Active EDW
Real-Time Data Warehousing

• The need for real-time data
  – A business often cannot afford to wait a whole day for its operational data to load into the data warehouse for analysis
  – Provides incremental real-time data showing every state change and almost analogous patterns over time
  – Maintaining metadata in sync is possible
  – Less costly to develop, maintain, and secure one huge data warehouse so that data are centralized for BI/BA tools
  – An EAI with real-time data collection can reduce or eliminate the nightly batch processes
Real-Time Data Warehousing

• The need for real-time data
  – A business often cannot afford to wait a whole day for its operational data to load into the data warehouse for analysis
  – Provides incremental real-time data showing every state change and almost analogous patterns over time
  – Maintaining metadata in sync is possible
  – Less costly to develop, maintain, and secure one huge data warehouse so that data are centralized for BI/BA tools
  – An EAI with real-time data collection can reduce or eliminate the nightly batch processes
Data Warehouse
Administration and Security Issues

• Data warehouse administrator (DWA)

A person responsible for the administration and management of a data warehouse
Data Warehouse
Administration and Security Issues

• Effective security in a data warehouse should focus on four main areas:
  – Establishing effective corporate and security policies and procedures
  – Implementing logical security procedures and techniques to restrict access
  – Limiting physical access to the data center environment
  – Establishing an effective internal control review process with an emphasis on security and privacy