MIS 400/500: Session # 3

Purpose
- Understand the basic definitions, concepts, and role of data warehouses in decision support
- Understand data warehousing architectures
- Comprehend the processes used in developing, operating, and managing data warehouses
- Understand data integration and the extraction, transformation, and load (ETL) processes
- Understand data warehouse administration and security issues

Common BI Architecture


Data Warehousing: Single Version of the Truth

“The truth, the whole truth, and nothing but the truth …”


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 Warehouse - Definitions

- A data warehouse is a subject oriented, integrated, time-variant, non-volatile collection of data in support of management’s decision-making process. (W. H. Inmon, 1994)

- A copy of transaction data specifically structured for query and analysis (R. Kimball, “The Data Warehouse Toolkit” 1996)
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 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

- 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

- 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

- 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 Architecture Alternatives**


**Standard Practices for Architecture Design**


**Standard Practices for Data Modeling**


**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 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?
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.

- Enterprise application integration (EAI)
  A technology that provides a vehicle for pushing data from source systems into a data warehouse.

- 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.

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).

![The ETL Process](image-url)
Data Integration and 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 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
  - 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 consider 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 addresses how data will be analyzed

- 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

  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
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

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

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

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

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

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

Data Warehousing Conceptual Layers

Persistent Staging Area (PSA)
- Storage area for data extracted from sources
- Requested data is saved directly from its source (without changes)
- First step in loading data into the operational data store (ODS) or data warehouse

Operational Data Store (ODS)
- Operational reporting
- Granular data
- Volatile, near real time
- May feed Data Warehouse layer at set intervals

Data Warehouse
- Historical foundation for BI
- Granular data
- Integrated
- Nonvolatile
- Application neutral

(Architected) Data Marts
- Focus: Information needs of a business unit or function
- Often aggregated, may be granular
- Often dimensional data models
Real-Time Data Warehousing

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

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

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

An EAI with real-time data collection can reduce or eliminate the nightly batch processes
**Data Warehouse Administrator**

- 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