Introduction to Data Warehouses

Helena Galhardas
DEI/IST

References

- J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2001 (chpts. 1 and 4)
- A. Wichert, H. Galhardas, Sistemas de Apoio à Decisão (slides), MEIC/IST
Recap: Building a Data Integration System

Create a middleware mediator or data integration system over the sources

– Can be warehoused (a data warehouse) or virtual
– Presents a uniform query interface and schema
– Abstracts away multitude of sources; consults them for relevant data
  • Unifies different source data formats (and possibly schemas)
  • Sources are generally autonomous, not designed to be integrated
– Sources may be local DBs or remote web sources/services
– Sources may require certain input to return output (e.g., web forms)
  • binding patterns describe these

Outline

• Introduction
  – Motivation for data warehousing
  – Definition of data warehouse
  – Normalized versus non-normalized data
  – Multidimensional modeling
• Typical data warehouse architecture
Introduction

• Organizations face increasingly complex challenges to achieve operational goals so need analysis tools for decision support
• Business intelligence (BI): Methodologies, processes, architectures, and technologies to transform raw data into useful information for decision making
  – Collect and summarize vast amounts of data
• Extraction, transformation, integration, and cleansing processes take data from sources, and store them in a common repository called:
  – Data warehouse (DW): integral part of decision-support systems. Provides an infrastructure that enables users to get efficient, accurate responses to complex queries

Knowledge Discovery (KDD) Process
Exploiting data in a DW

- A wide variety of systems and tools to exploit the data in a warehouse
- **Online Analytical Processing (OLAP)**
  - Allows users to interactively query and aggregate data in a warehouse
  - Decision makers can analyze information at various levels of detail
- **Data mining** extracts interesting knowledge hidden in data warehouses
- Typical techniques that exploit a data warehouse:
  - Reporting: dashboards, alerts
  - Performance management: metrics, key performance indicators (KPIs), dashboards
  - Analytics: OLAP, data mining, time series analysis, text mining, web analytics, data visualization

Motivation

- Traditional **operational or transactional databases** do not satisfy the requirements for data analysis
  - Designed/optimized to support **daily business operations**
    - primary concern: concurrent access and recovery techniques to guarantee data consistency
  - Contain detailed data, do not include historical data, and perform poorly for complex queries that involve many tables or aggregate large volumes of data
- To analyze the behavior of an organization, data from several operational systems must be **integrated**
  - Difficult to accomplish due to many differences in data definition and content
Definition of Data Warehouse (DW)

• Collection of subject-oriented, integrated, nonvolatile, and time-varying data to support management decisions (Inmon definition)

Data Warehouse: Subject-Oriented

• Organized around major subjects, such as customer, product, sales
• Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing
• Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process
Data Warehouse: Integrated

- Constructed by integrating multiple, heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, breakfast covered, etc.
  - When data is moved to the warehouse, it is converted.

Data Warehouse: Nonvolatile

- A physically separate store of data transformed from the operational environment
- Operational update of data does not occur in the data warehouse environment
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only three operations in data accessing:
    - initial loading of data, access of data, and periodic data refreshment
Data Warehouse: Time Variant

• The time horizon for the data warehouse is significantly longer than that of operational systems
  – Operational database: current value data
  – Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
• Every key structure in the data warehouse
  – Contains an element of time, explicitly or implicitly

Outline

• Introduction
  – Motivation for data warehousing
  – Definition of data warehouse
  ➢ Normalized versus non-normalized data
    – Multidimensional modeling
• Typical data warehouse architecture
Normalized vs non-normalized data

- **Relational databases**: highly normalized to guarantee consistency under frequent updates
  - Usually achieved at a higher cost of querying (normalization partitions data into multiple tables)
  - This is not appropriate for data warehouses
- **Data warehouses** must deliver good performance for the complex queries needed for analysis tasks
  - Examples of complex queries:
    - How much did the unit sale A sell in January?
    - What is the total sales amount of this unit in the first semester?
  - Less degree of normalization required => multidimensional modeling

Store: (Normalized) Relational Model
Outline

• Introduction
  – Motivation for data warehousing
  – Definition of data warehouse
  – Normalized versus non-normalized data
  ➢ Multidimensional modeling
• Typical data warehouse architecture
Multidimensional modeling

- Views data as consisting of facts linked to dimensions.
  - Facts represent the focus of analysis (e.g., analysis of sales in stores)
    - usually numeric values, e.g., units_sold or dollars sold
  - Dimensions used to analyze measures from several perspectives, e.g.:
    - Time dimension to analyze changes in sales over various periods of time
    - Location dimension to analyze sales according to the geographic distribution of stores
- Dimensions include attributes that form hierarchies which enable decision-making users to explore measures at various levels of detail, e.g.:
  - month → quarter → year in the time dimension
  - city → state → country in the location dimension
- Aggregation of measures occurs when a hierarchy is traversed, e.g., moving from month to year yields aggregated values of sales for the various years.
• **Query Language**
  – Once a data warehouse has been implemented, analytical queries can be submitted
  – **MDX (MultiDimensional eXpressions):** de facto standard language for querying a multidimensional database

• **Physical level:** concerned with implementation issues

• Three techniques are normally used for improving system performance:
  – Materialized views
  – Indexing
  – Data partitioning

Outline

• **Introduction**
  – Motivation for data warehousing
  – Definition of data warehouse
  – Normalized versus non-normalized data
  – Multidimensional modeling

  ➢ **Typical data warehouse architecture**
**Components of the DW architecture**

- **Back-end tier:**
  - Extraction, Transformation, and Loading (ETL) process: feeds data into the data warehouse from operational databases and other data sources
  - Data Staging Area (DSA): intermediate database where all the data integration and transformation processes are run prior to the loading of the data into the data warehouse

- **Data warehouse tier:**
  - Enterprise data warehouse and/or several data marts
  - Metadata repository storing information about the data warehouse and its contents

- **OLAP tier** composed of:
  - OLAP server which provides a multidimensional view of the data, regardless the actual way in which data are stored

- **Front-end tier** is used for data analysis and visualization
  - Contains client tools such as OLAP tools, reporting tools, statistical tools, and data-mining tools
Data Warehouse Architecture

Back-End Tier

3-step process: Extraction, Transformation, and Loading (ETL)

1. **Extraction** gathers data from multiple, heterogeneous data sources internal or external to the organization

2. **Transformation** modifies the data from the format of the data sources to the warehouse format; this includes:
   - **Cleaning**: Removes errors and inconsistencies in the data and converts it into a standardized format
   - **Integration**: Reconciles data from different data sources, both at the schema and at the data level
   - **Aggregation**: Summarizes the data obtained from data sources according to the granularity of the data warehouse

3. **Loading** feeds the data warehouse with the transformed data, including refreshing the data warehouse, that is, propagating updates from the data sources to the data warehouse at a specified frequency

- **Data staging area** (also called operational data store): A database where data extracted from the sources undergoes successive modifications before being loaded into the data warehouse
Data Warehouse Architecture

**DW Tier**

- **Enterprise data warehouse**, centralized and encompassing an entire organization
- Several **data marts**: specialized departmental data warehouses
- **Metadata**
  - **Business metadata** describes the semantics of the data, organizational rules, policies, and constraints related to the data
  - **Technical metadata** describes how data are structured and stored in a computer system, and the applications and processes that manipulate the data
- **Metadata repository** may contain information such as:
  - Metadata describing the structure of the data warehouse and the data marts, at the conceptual/logical level (facts, dimensions, hierarchies, ...) and at the physical level (indexes, partitions, ...)
  - Security information (user authorization and access control), and monitoring information (usage statistics, error reports, audit trails)
  - Metadata describing data sources: schemas, ownership, update frequencies, legal limitations, access methods
  - Metadata describing the ETL process: data lineage, data extraction, cleaning, transformation rules, etc.
OLAP Tier

- **OLAP server**, which presents business users with multidimensional data from data warehouses or data marts
  - Products include OLAP extensions and tools allowing building, querying, and navigating cubes, analysis, and reporting
- There is not yet a standardized language for defining and manipulating data cubes
  - **MDX** (MultiDimensional eXpressions): query language for OLAP databases, a de facto standard for querying OLAP systems
  - SQL extended for providing analytical capabilities: **SQL/OLAP**
OLAP Server Architectures

- **Relational OLAP (ROLAP)**
  - Use relational or extended-relational DBMS to store and manage warehouse data and OLAP middleware
  - Include optimization of DBMS backend, implementation of aggregation navigation logic, and additional tools and services
  - Greater scalability

- **Multidimensional OLAP (MOLAP)**
  - Sparse array-based multidimensional storage engine
  - Fast indexing to pre-computed summarized data

- **Hybrid OLAP (HOLAP)** (e.g., Microsoft SQLServer)
  - Flexibility, e.g., low level: relational, high-level: array

- **Specialized SQL servers** (e.g., Redbricks)
  - Specialized support for SQL queries over star/snowflake schemas

Data Warehouse Architecture
Front-End Tier

- Client tools that allow users to exploit the content of the data warehouse
  - **OLAP tools**: allow interactive exploration and manipulation of the warehouse data and formulation of complex ad-hoc queries
  - **Reporting tools**: enable the production, delivery, and management of reports, which can be paper-based, interactive, or web-based
    - Reports use predefined queries asking for specific information in a specific format, performed on a regular basis
  - **Statistical tools**: used to analyze and visualize the cube data using statistical methods
  - **Data mining tools**: allow users to analyze data in order to discover valuable knowledge such as patterns and trends, and also allow to make predictions based on current data

Variations of the architecture

1. Only an enterprise data warehouse without data marts or, alternatively, an enterprise data warehouse does not exist
2. An OLAP server does not exist and/or the client tools directly access the data warehouse
3. Neither a data warehouse nor an OLAP server - virtual data warehouse (virtual data integration) - which defines a set of views over operational databases that are materialized for efficient access
   - does not contain historical data, centralized metadata, etc.
4. Data staging area may not be needed when the data in the source systems conforms very closely to the data in the warehouse
Next Lecture

• OLAP and Multidimensional Model