Chapter 4&5. Data Cube: Data Warehousing and OLAP

Meng Jiang

CSE 40647/60647 Data Science Fall 2017

Introduction to Data Mining

Data Warehouse

- Defined in many different ways, but not rigorously
  - A decision support database that is maintained separately from the organization's operational database

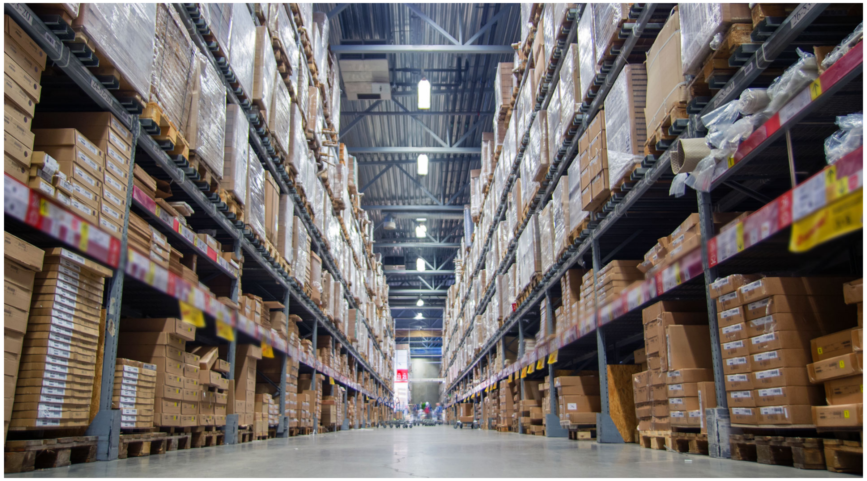
Operational Databases

(Data) Marts

(Data) Warehouse

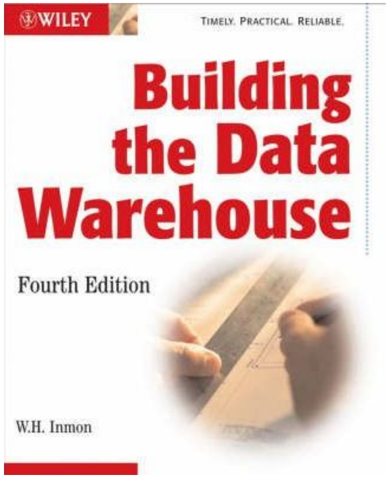
Data Warehouse

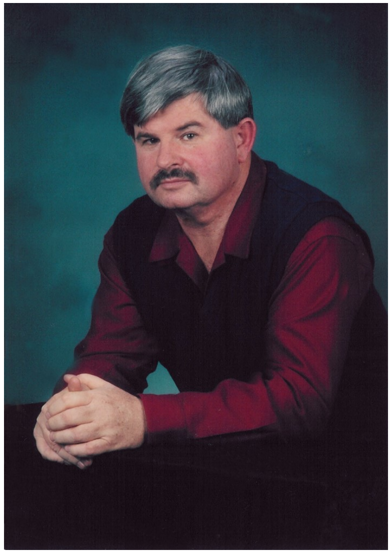
- Defined in many different ways, but not rigorously
  - A decision support database that is maintained separately from the organization's operational database
  - Support information processing by providing a solid platform of consolidated, historical data for analysis



Data Warehouse

- “A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process.”—William H. (Bill) Inmon





- Data warehousing:
  - The process of constructing and using data warehouses

(1) 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

(2) 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
    - Ex. Hotel price: differences on currency, tax, breakfast covered, and parking

(3) 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
  - But the key of operational data may or may not contain "time element"

(4) Nonvolatile

- Independence
  - A physically separate store of data transformed from the operational environment
- Static: Operational update of data does NOT occur in the data warehouse environment
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing:
    - initial loading of data and access of data

OL**T**P vs OL**A**P

- OLTP: **Online** transactional processing
  - DBMS operations
  - Query and transactional processing
- OLAP: **Online** analytical processing
  - Data warehouse operations (drilling, slicing, dicing, etc.)
  - Data analysis to support decision making

OLTP vs OLAP

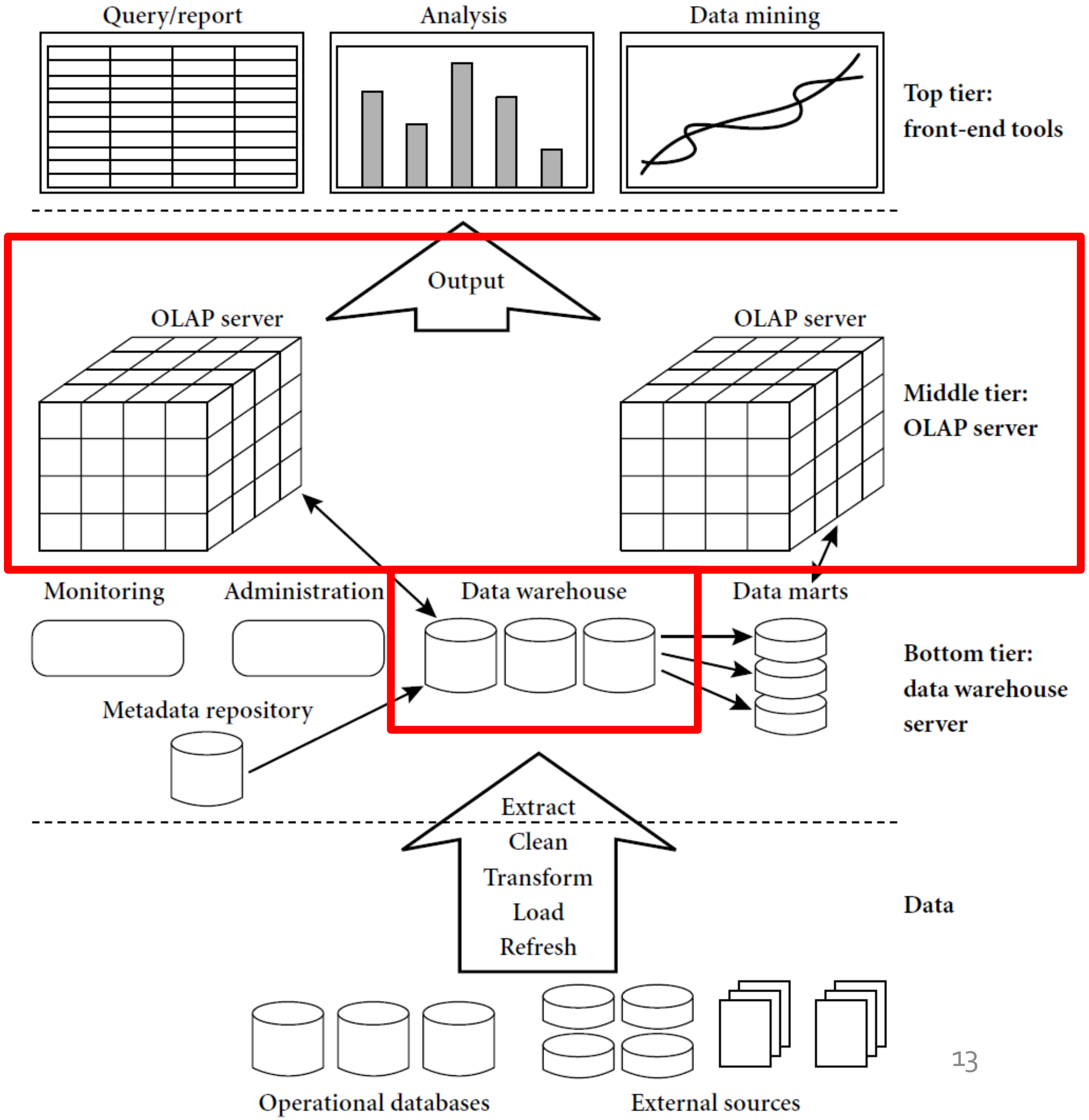
	OLTP	**OLAP**
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

Why a Separate Data Warehouse?

- High performance for both systems
  - DBMS— tuned for **OLTP**: access methods, indexing, concurrency control, recovery
  - Warehouse—tuned for **OLAP**: complex OLAP queries, multidimensional view, consolidation
- Different functions and different data:
  - Decision support requires historical data which operational DBs do not typically maintain
  - DS requires consolidation (aggregation, summarization) of data from heterogeneous sources
  - Different sources typically use inconsistent data representations, codes and formats which have to be reconciled

Data Warehouse: A Multi-Tiered Architecture

- Top Tier: Front-End Tools
- **Middle Tier: OLAP Server**
- **Bottom Tier: Data Warehouse Server**
- Data



From Data to Data Warehouse: Extraction, Transformation, and Loading (ETL)

- **Data extraction**
  - get data from multiple, heterogeneous, and external sources
- **Data cleaning**
  - detect errors in the data and rectify them when possible
- **Data transformation**
  - convert data from legacy or host format to warehouse format
- **Load**
  - sort, summarize, consolidate, compute views, check integrity, and build indices and partitions
- **Refresh**
  - propagate the updates from the data sources to the warehouse

From Tables and Spreadsheets to Data Cubes

- A **data warehouse** is based on a multidimensional data model which views data in the form of a **data cube**
- A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions
  - **Dimension tables**, such as item (item_name, brand, type), or time (day, week, month, quarter, year)
  - **Fact table** contains **measures** (such as dollars_sold) and keys to each of the related dimension tables
- **Data cube**: A lattice of cuboids
  - In data warehousing literature, an **n-D base cube** is called a **base cuboid**
  - The top most **0-D cuboid**, which holds the highest-level of summarization, is called the **apex cuboid**
  - The lattice of cuboids forms a **data cube**

Data Warehouse Usage

- Three kinds of data warehouse applications
  - Information processing
    - supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts and graphs
  - Analytical processing
    - multidimensional analysis of data warehouse data
    - supports basic OLAP operations, slice-dice, drilling, pivoting
  - Data mining
    - knowledge discovery from hidden patterns
    - supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools

Efficient Processing OLAP Queries

- **Determine which operations** should be performed on the available cuboids
  - Transform drill, roll, etc. into corresponding SQL and/or OLAP operations, e.g., dice = selection + projection
- **Determine which materialized cuboid(s)** should be selected for OLAP op.
  - Let the query to be processed be on {*brand, province_or_state*} with the condition "*year = 2004*", and there are 4 materialized cuboids available:

    1) {*year, item_name, city*}

    2) {*year, brand, country*}

    3) {*year, brand, province_or_state*} √

    4) {*item_name, province_or_state*} where *year = 2004*

    Which should be selected to process the query?

https://career4.successfactors.com/career?career_ns=job_listing&company=FCMA&navBarLevel=JOB_SEARCH&rcm_site_locale=en_US&career_job_req_id=9661

Discussion

- **Career Opportunities: Data Warehouse Architect (9661)** - Posted **07/10/2017** - **Information Technology** - **KY - Louisville** - **Kentucky**
- The primary responsibility of the Data Warehouse Architect is the implementation and management of data standards and procedures surrounding the data warehouse. This would primarily include the design and development of logical and physical data models, databases, distributed data management, and information management functions. Additionally, the Data Warehouse Architect will have the responsibility for maintaining the enterprise data architecture vision, strategy, principles, and standards.
- **MAJOR RESPONSIBILITIES:**
- Design, develop and maintain an enterprise, business centric data model and data dictionary (logical model) incorporating both internal and external information systems, providing relevant data elements to enable both ad-hoc and strategic reporting, as well as non-reporting functions.
- Be responsible and accountable for crafting the overall architectural direction of the enterprise data warehouse strategy that aligns with the stated objectives of the business' multidimensional design.

MAJOR RESPONSIBILITIES:

- Architect the overall data warehouse design - conceptual, logical, and physical representations.
- Develop and use business knowledge to critically evaluate information gathered from multiple sources, reconcile conflicts, and develop detailed requirements from high-level information.
- Assist in Data Quality research and User Acceptance Testing.
- Maintain enterprise data management strategies, guiding principles, governance documentation, along with processes and standards.
- Maintain a structure for business information, understand current and emerging technologies, and align applications with business priorities.
- Define standards for the data warehouse, the integration/migration strategy for data, and data structure conventions.
- Define standards, structures, and techniques for capturing data from sources, cleansing, and integrating data.
- Lead the design of robust, scalable, and maintainable data integration processes.

MAJOR RESPONSIBILITIES:

- Recommend hardware and software products; participate in the acquisition, evaluation, and testing of hardware and software products and establish standards and provide guidance for the use of those products.
- Develop and maintain effective teams as well as organizational working relationships and partnerships to include user training and engagement.
- Manage Data Warehouse related projects.
- Manage the design and operation of interfaces and data updates to the Data Warehouse.
- Design, develop and maintain the necessary data repositories (physical model), adapting industry “best practices” for continuous improvement
- Work with Enterprise Architects, Database Architects, and the Application Development teams to establish agreed data acquisition strategies, service level agreements and disaster recovery procedures concerning the data warehouse environments. Lead the strategy definition for overall data acquisition processes and methodologies. Be responsible and accountable for data warehouse capabilities achieved through sound, well architected designs.

MAJOR RESPONSIBILITIES:

- Ensure the Data Warehouse is robust, with minimal downtime and is refreshed to agreed timescales.
- Acts as liaison between Business Technology and our business units.
- Deliver a project scope and data strategy that directly supports the key business drivers.
- Design and direct the implementation of security requirements for the data warehouse.
- **MINIMUM REQUIREMENTS**
- **Education and Experience**
- Bachelor's degree in computer science, business analytics or related field, and a minimum of five years relevant work experience building, deploying, and supporting Enterprise Data Warehouse capabilities. Good project management skills, oral and written communication skills, and analytical skills necessary.
- **KNOWLEDGE, SKILLS AND ABILITIES:**
- **Knowledge of:** Microsoft BI Stack and data management tools including SharePoint, Performance Point, InfoPath, Excel Services for SharePoint, Power BI, Power Map, Power Query, PowerPivot and other Excel extensions. Exposure other desktop analytics or BI tools such as Tableau, Microstrategy, Cognos, and others are considered a plus.
- **Skills in: SQL, MDX, Multidimensional modeling, Dashboard reporting, KPI/Metric analysis and presentation,**
- **Ability to: Coordinate end users and analysts and convert user requirements into effective visualizations and interactive dashboards.**

Summary

- Data warehousing: A multi-dimensional model of a data warehouse
  - A data cube consists of *dimensions* & *measures*
  - Star schema, snowflake schema, fact constellations
  - OLAP operations: drilling, rolling, slicing, dicing and pivoting
- Implementation: Efficient computation of data cubes
  - Partial vs. full vs. no materialization
  - OLAP query processing

References

- S. Agarwal, R. Agrawal, P. M. Deshpande, A. Gupta, J. F. Naughton, R. Ramakrishnan, and S. Sarawagi. On the computation of multidimensional aggregates. VLDB'96
- D. Agrawal, A. E. Abbadi, A. Singh, and T. Yurek. Efficient view maintenance in data warehouses. SIGMOD'97
- R. Agrawal, A. Gupta, and S. Sarawagi. Modeling multidimensional databases. ICDE'97
- **S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. ACM SIGMOD Record, 26:65-74, 1997**
- J. Gray, et al. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. Data Mining and Knowledge Discovery, 1:29-54, 1997.
- A. Gupta and I. S. Mumick. Materialized Views: Techniques, Implementations, and Applications. MIT Press, 1999
- J. Han. Towards on-line analytical mining in large databases. *SIGMOD Record*, 1998
- V. Harinarayan, A. Rajaraman, and J. D. Ullman. Implementing data cubes efficiently. SIGMOD'96

References (cont.)

- C. Imhoff, N. Galemmo, and J. G. Geiger. Mastering Data Warehouse Design: Relational and Dimensional Techniques. John Wiley, 2003
- W. H. Inmon. Building the Data Warehouse. John Wiley, 1996
- R. Kimball and M. Ross. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. 2ed. John Wiley, 2002
- P. O'Neil and D. Quass. Improved query performance with variant indexes. SIGMOD'97
- S. Sarawagi and M. Stonebraker. Efficient organization of large multidimensional arrays. ICDE'94
- P. Valduriez. Join indices. ACM Trans. Database Systems, 12:218-246, 1987.
- J. Widom. Research problems in data warehousing. CIKM’95.
- K. Wu, E. Otoo, and A. Shoshani, Optimal Bitmap Indices with Efficient Compression, ACM Trans. on Database Systems (TODS), 31(1), 2006, pp. 1-38.