IIT School of Applied Technology
ILLINOIS INSTITUTE OF TECHNOLOGY
information technology & management

526 Data Warehousing

February 5, 2019
Week 4 Presentation

Week 4 Topic

Introduction to Data Warehousing

Attendance

- In-Class
  - Roster call
- On-Line
  - Proof of Online Attendance
  - 3 screenshots with http://www.clocktab.com/
  - Email the screenshots to daniel.lee@iit.edu by end of Saturday with the title "Proof of Attendance"
  - Submission after the due (Sunday) will NOT be taken into account
- Up to 4 absences will not negatively impact

Attendance for Online Participants

A sample screenshot:

Grading Assignments

- Late Submission
  - Start early, ask questions early
  - First day after due: 15%
  - %5 per day afterward
  - Maximum penalty: 40%
- Maximum Deduction
  - 50% (includes late submission)

Tableau and PowerBI

- Power BI Self Guide Learning
  - https://docs.microsoft.com/en-us/power-bi/guided-learning/
- Tableau for Students (Download and Installation)
  - https://www.tableau.com/academic/students
- Tableau Desktop Free Training Videos
  - https://www.tableau.com/learn/training

Visualization References

- Visual Vocabularly
  - ft.com/vocabulary
- The Visual Reference
  - https://www.sqlbi.com/ref/power-bi-visuals-reference/
- Story Telling
  - https://www.analyticsvidhya.com/blog/2017/10/art-story-telling-data-science/

What is Data Warehouse?

- A decision support database that is maintained separately from the organization's operational database
  - Provides subject-oriented, integrated, time-variant, and nonvolatile collection of data for analysis
- Data warehousing:
  - The process of constructing and using data warehouses

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

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: information from a historical perspective (e.g., past 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"

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
  - Requires only two operations in data accessing:
    - *loading (destructive or incremental)*
    - *access of data*

OLTP vs. OLAP

- OLTP (on-line transaction processing)
  - Major task of traditional relational DBMS
  - Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
- OLAP (on-line analytical processing)
  - Major task of data warehouse system
  - Data analysis and decision making

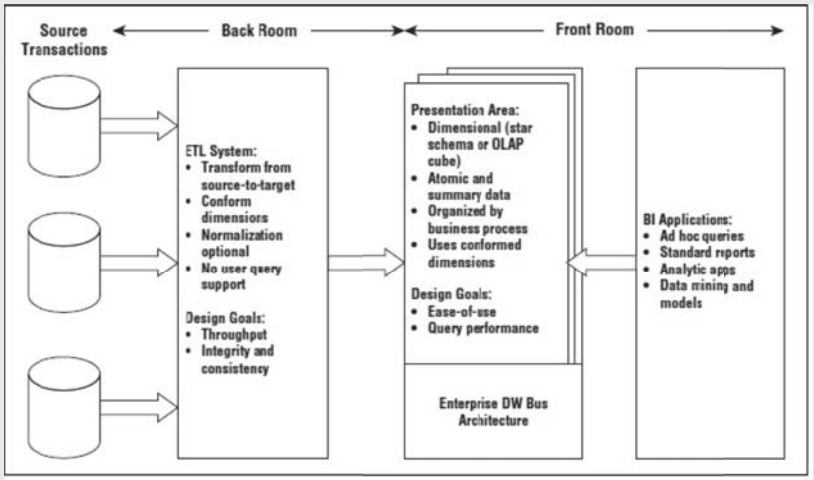
OLTP vs. OLAP

	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 Separate Data Warehouse?

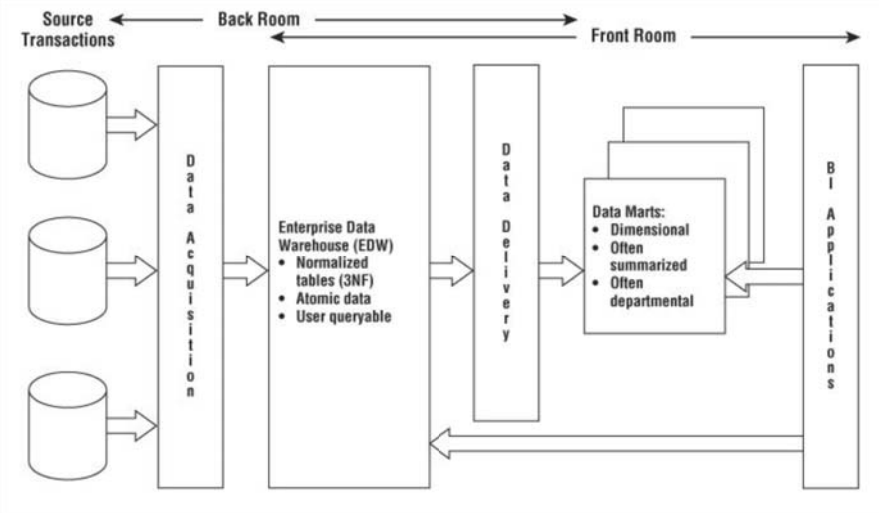
- High performance for both systems
  - OLTP: Tuned for OLTP - indexing, concurrency control, recovery
  - DW: Tuned for OLAP - complex OLAP queries, multidimensional view, consolidation
- Different Needs
  - Historical Data: DW requires historical data which operational systems do not typically maintain
  - Integration: DW requires consolidation (aggregation, summarization) of data from heterogeneous sources
  - Data Quality: Different sources typically use inconsistent data representations, codes and formats which have to be reconciled

Data Warehouse: Kimball's Architecture



Core Elements Kimball DW/BI Architecture.

Data Warehouse: William Inman's Architecture



ILLINOIS INSTITUTE OF TECHNOLOGY School of Applied Technology

ITMD - 526

Data Warehouse Backroom Tools

- Data Extraction
  - get data from multiple, heterogeneous, and external sources
- Data Transformation/cleaning
  - convert data from legacy or host format to warehouse format
- Load
  - sort, summarize, consolidate, computed views, check integrity, and build indicies and partitions
- Refresh

ILLINOIS INSTITUTE OF TECHNOLOGY School of Applied Technology

ITMD - 526

Data Warehouse Usage

- Information processing
  - supports querying, basic statistical analysis, reporting, and visualization
- Analytical processing
  - multidimensional analysis of data warehouse data such as OLAP cubes for slice-dice, drilling, pivoting
- Data mining
  - knowledge discovery from hidden patterns
  - supports associations, constructing analytical models, performing classification and prediction

ILLINOIS INSTITUTE OF TECHNOLOGY School of Applied Technology

ITMD - 526

Week 4 Class Exercises

- Pentaho Data Integration
- Importing AdventureWorks
- Reverse Engineering via MySQL Workbench
- (Optional) Creating an ERD via LucidChart