IIT School of Applied Technology

ILLINOIS INSTITUTE OF TECHNOLOGY

information technology & management

526 Data Warehousing

January 13, 2016

Week 1 Presentation

Data Warehousing 526
Course Objectives

- Clearly understand the top down (William Inmon) and bottom-up (Ralph Kimball) approaches for building data warehouses.
- Correctly use data warehousing and business intelligence terminology
- Explain and apply the Business Dimensional Lifecycle
- Perform multidimensional analysis
- Describe data warehouse infrastructure issues

Data Warehousing 526
Course Objectives (cont'd)

- Determine the need for the management of meta data
- Explain the components of a data warehouse technical architecture.
- Describe techniques for data extraction transformation and loading into a data warehouse (ETL)
- Explain the techniques used for data presentation by analytical applications
- Demonstrate the techniques for building a dimensional data mart/warehouse.

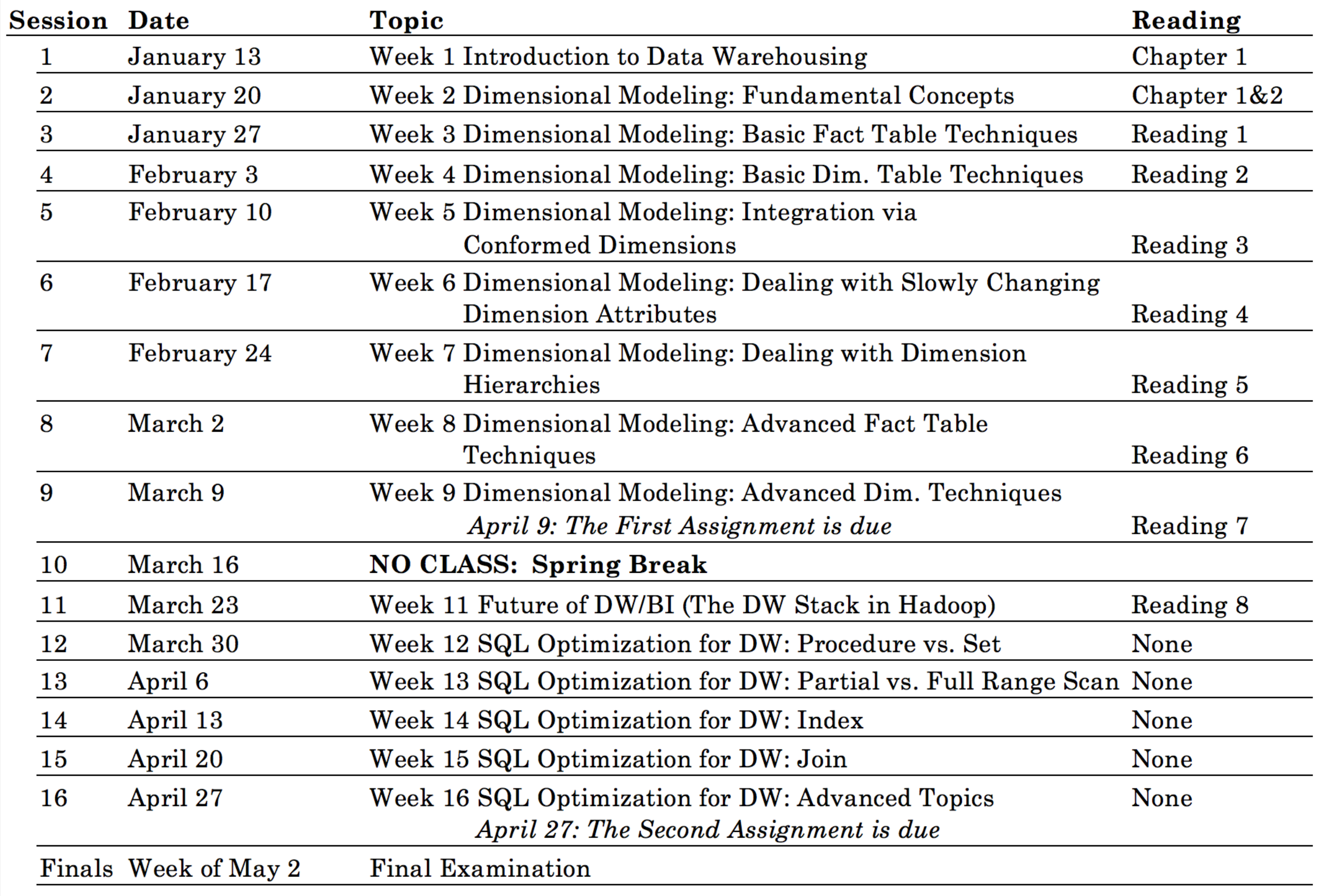
Course Outline

- Introduction to Data Warehouse
- Dimensional Modeling Fundamentals
- Implementation with ETL and Reporting tools
- Advanced SQL Optimization Principles and Best Practices

Books and Course Materials

- Main Textbook
  - Kimball, R. (2013). The Data Warehouse Lifecycle Toolkit. ISBN 0-471-25547-5. John Wiley & Sons Inc.
  - Articles and Kimble Design Tips www.kimballgroup.com (Resource tab)
- Reference Books
  - Inmon, W.H. (2001). Corporate Information Factory. ISBN 0-471-39961-2. John Wiley & Sons Inc.
  - Meade, K. (2014). Oracle SQL Performance Tuning and Optimization: It's all about the Cardinalities. ISBN 1-501-02269-5. CreateSpace Independent Publishing Platform.

Schedule of Topics



Session	Date	Topic	Reading
1	January 13	Week 1 Introduction to Data Warehousing	Chapter 1
2	January 20	Week 2 Dimensional Modeling: Fundamental Concepts	Chapter 1&2
3	January 27	Week 3 Dimensional Modeling: Basic Fact Table Techniques	Reading 1
4	February 3	Week 4 Dimensional Modeling: Basic Dim. Table Techniques	Reading 2
5	February 10	Week 5 Dimensional Modeling: Integration via Conformed Dimensions	Reading 3
6	February 17	Week 6 Dimensional Modeling: Dealing with Slowly Changing Dimension Attributes	Reading 4
7	February 24	Week 7 Dimensional Modeling: Dealing with Dimension Hierarchies	Reading 5
8	March 2	Week 8 Dimensional Modeling: Advanced Fact Table Techniques	Reading 6
9	March 9	Week 9 Dimensional Modeling: Advanced Dim. Techniques *April 9: The First Assignment is due*	Reading 7
10	March 16	**NO CLASS: Spring Break**	
11	March 23	Week 11 Future of DW/BI (The DW Stack in Hadoop)	Reading 8
12	March 30	Week 12 SQL Optimization for DW: Procedure vs. Set	None
13	April 6	Week 13 SQL Optimization for DW: Partial vs. Full Range Scan	None
14	April 13	Week 14 SQL Optimization for DW: Index	None
15	April 20	Week 15 SQL Optimization for DW: Join	None
16	April 27	Week 16 SQL Optimization for DW: Advanced Topics *April 27: The Second Assignment is due*	None
Finals	Week of May 2	Final Examination	

Course Assignments, Exercises

- Class Exercises/Homework
  - Dimensional Modeling
  - ETL implementation: Each student is expected to set up an ETL environment on his/her own computer
  - Advanced SQL Optimization (for Extra Credit)
- 2 Course Assignments
- A Final

More details on the assignments will be available according to the course schedule.

Class exercises are due one week from assigned date. Class exercises will be handed out at the end of class, as necessary. If a class exercise is assigned, due date for it will be the following Wednesday.

ITMD - 526

Course Grading

Grading criteria for ITMD 526 students will be as follows:

- **A** *Outstanding work reflecting substantial effort:* 90-100%
- **B** *Adequate work fully meeting that expected of a graduate student:* 80-89.99%
- **C** *Weak but marginally satisfactory work not fully meeting expectations:* 65-79.99%
- **E** *Unsatisfactory work:* 0-64.99%
- *No Exceptions!*

The final grade for the class will be calculated as follows:

➢ Assignment 1	**25%**
➢ Assignment 2	**25%**
➢ Final Exam	**25%**
➢ Class Exercises & Participation	**25%**

Other Class Logistics

- Class time: 6:25 PM ~ 9:05 PM. Break time will be 7:30 ~ 7:40 PM.
- Blackboard will the main hub for course materials distribution, assignment submission, communications, etc.
  - Other means of contact:
    Email: Best means of communication is email. I will respond within 24 hours
  - Phone: (773) 312-5342 (For a long message, please send me an email instead)
- Bring your laptop for class exercises
- The first part of each class will focus on concepts and principles while the second part will focus on implementation aspects such as hands on practices

Week 1 Topic

Introduction to Data Warehousing

What is 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.
- "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management' s decision-making process."—W. H. Inmon
- 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
  - When data is moved to the warehouse, it is converted.

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
  - 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
  - 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*

Data Warehouse vs. Operational DBMS

- 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
- Distinct features (OLTP vs. OLAP):
  - User and system orientation: customer vs. market
  - Data contents: current, detailed vs. historical, consolidated
  - Database design: ER + application vs. star + subject
  - View: current, local vs. evolutionary, integrated
  - Access patterns: update vs. read-only but complex queries

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
  - 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:
  - missing data: Decision support requires historical data which operational DBs do not typically maintain
  - data consolidation: DS 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
- Note: There are more and more systems which perform OLAP analysis directly on relational databases (Agile DW)

Design of Data Warehouse: A Business Analysis Framework

- Four views regarding the design of a data warehouse
  - Top-down view
    - allows selection of the relevant information necessary for the data warehouse
  - Data source view
    - exposes the information being captured, stored, and managed by operational systems
  - Data warehouse view
    - consists of fact tables and dimension tables
  - Business query view
    - sees the perspectives of data in the warehouse from the view of end-user

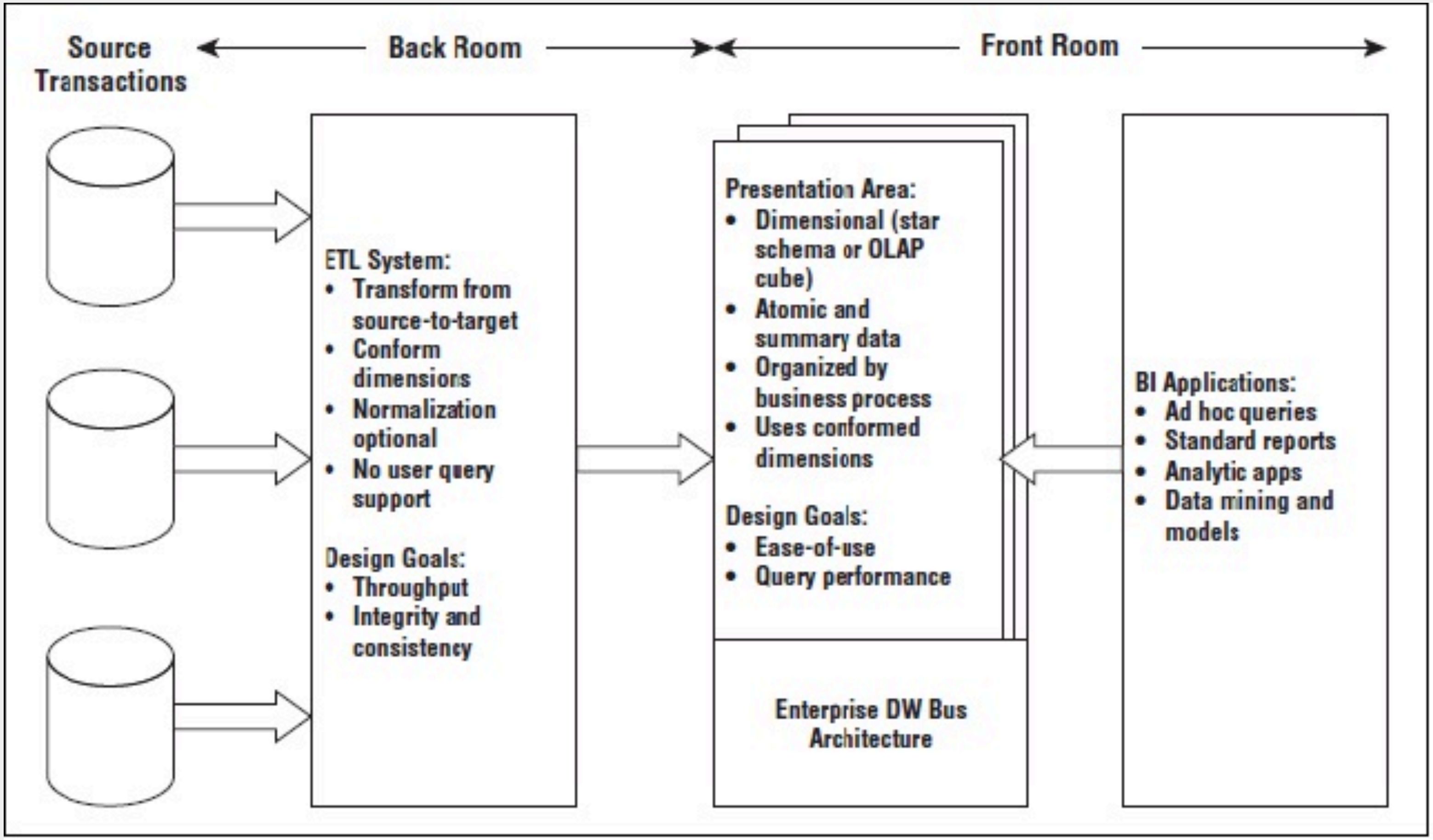
Data Warehouse Design Process

- Top-down, bottom-up approaches or a combination of both
  - Top-down: Starts with overall design and planning (mature)
  - Bottom-up: Starts with experiments and prototypes (rapid)
- From software engineering point of view
  - Waterfall: structured and systematic analysis at each step before proceeding to the next
  - Spiral: rapid generation of increasingly functional systems, short turn around time, quick turn around

Data Warehouse Design Process (Cont'd)

- Typical data warehouse design process
  - Choose a business process to model, e.g., orders, invoices, etc.
  - Choose the *grain* (*atomic level of data*) of the business process
  - Choose the dimensions that will apply to each fact table record
  - Choose the measure that will populate each fact table record

Data Warehouse: Kimball's Architecture



Source Transactions ← Back Room → ← Front Room →

ETL System:
- Transform from source-to-target
- Conform dimensions
- Normalization optional
- No user query support

Design Goals:
- Throughput
- Integrity and consistency

Presentation Area:
- Dimensional (star schema or OLAP cube)
- Atomic and summary data
- Organized by business process
- Uses conformed dimensions

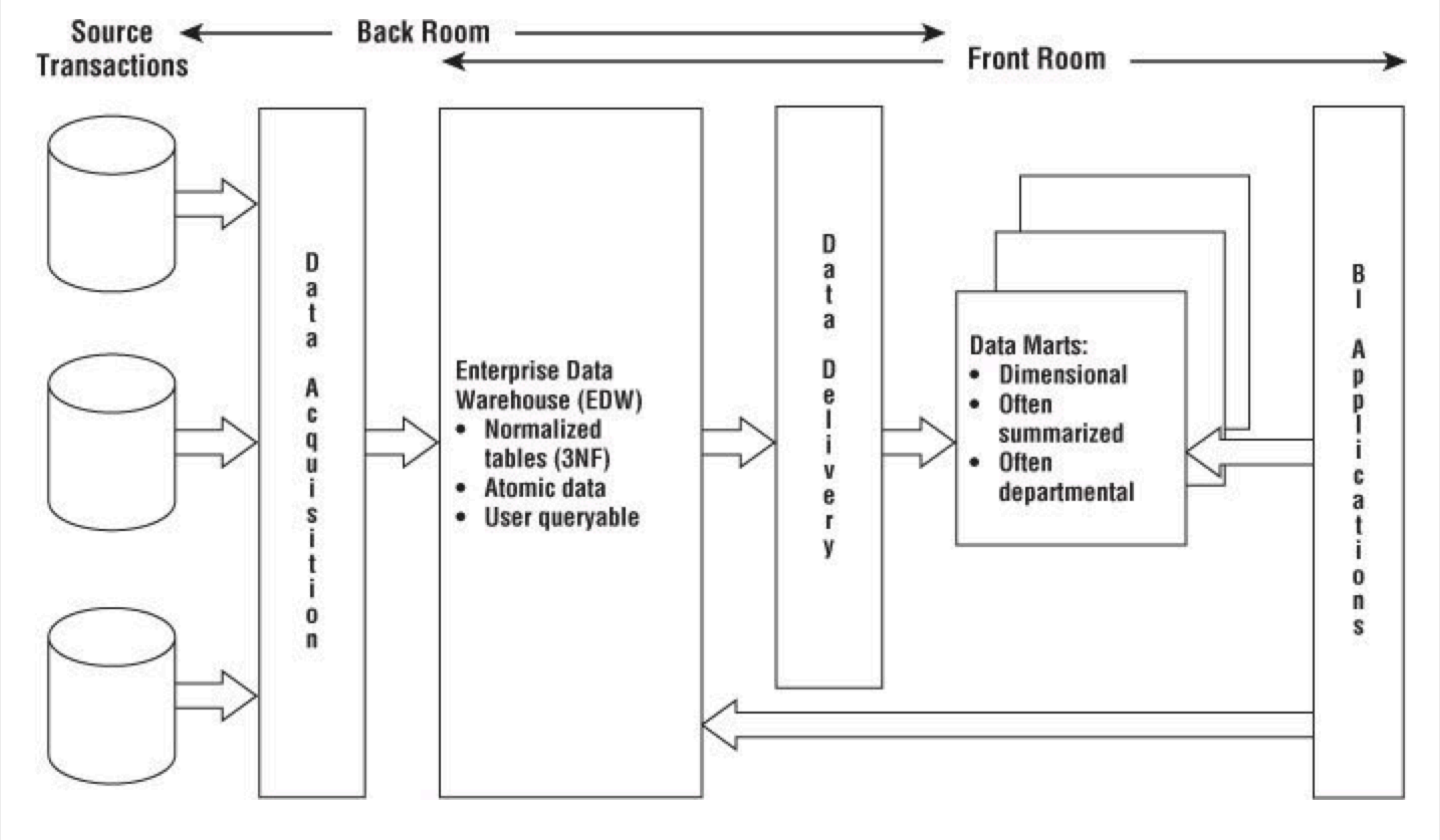
Design Goals:
- Ease-of-use
- Query performance

Enterprise DW Bus Architecture

BI Applications:
- Ad hoc queries
- Standard reports
- Analytic apps
- Data mining and models

Core Elements Kimball DW/BI Architecture.

Data Warehouse: William Inman's Architecture



Three Data Warehouse Models

- Enterprise warehouse
  - collects all of the information about subjects spanning the entire organization
- Data Mart
  - a subset of corporate-wide data that is of value to a specific groups of users. Its scope is confined to specific, selected groups, such as marketing data mart
    - Independent (landing zone) vs. dependent (directly from warehouse) data mart
- Virtual warehouse
  - A set of views over operational databases
  - Only some of the possible summary views may be materialized

Data Warehouse Back-End Tools and Utilities

- 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 indicies and partitions
- Refresh
  - propagate the updates from the data sources to the warehouse (Disruptive, Incremental, Real-time, etc)

Metadata Repository

- Meta data is the data defining warehouse objects. It stores:
- Description of the structure of the data warehouse (Technical spec)
  - schema, view, dimensions, hierarchies, derived data definition, data mart locations and contents
- Operational meta-data
  - data lineage (history of migrated data and transformation path), currency of data (active, archived, or purged), monitoring information (warehouse usage statistics, error reports, audit trails)

Metadata Repository (cont'd)

- The algorithms used for summarization
- The mapping from operational environment to the data warehouse
- Business data
  - business terms and definitions, ownership of data, charging policies

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

Week 1 Class Exercise & Homework

Introduction to ETL

Pentaho Data Integration (PDI)

- Open source ETL tool
- Large user community
- Support almost all data formats and sources
- Easy to integrate into Pentaho BI suite
- Introduction: http://www.slideshare.net/mattcasters/pentaho-data-integration-introduction

Lab: Setting Up ETL Environment

1. MySQL Community Server 5.x
2. MySQL Client Tools
   - SQLYog Community: https://github.com/webyog/sqlyog-community
   - MySQL Workbench via MySQL Installer
3. Pentaho Data Integration (Kettle) 5.4
   - http://sourceforge.net/projects/pentaho/files/Data%20Integration/5.4/
   - Documentation: https://help.pentaho.com/Documentation/5.4/0F0/0H0 (Select "DI only")
4. MySQL Jave Connector / Driver
   - http://dev.mysql.com/downloads/connector/j/

Week 1 Class Exercise (homework):

- Set up your own ETL environment following the same steps as the demo today.
- You can use your choice of operating system and/or virtualization software.
- Record the video of the the working PDI, upload it to your choice of video website (YouTube, Screencast, Vimoe, etc), and submit the link to the Blackboard
- Screen recording software
  - Camstudio for Windows: https://www.youtube.com/watch?v=WQ5_6szOf48
  - QuickTime Player for Mac: http://osxdaily.com/2010/11/16/screen-recorder-mac/
  - Etc.