* **Email to teachers:** 
  + [**lejin@ebay.com**](mailto:lejin@ebay.com) **(Orlando, general questions)**
  + [**dagu@ebay.com**](mailto:dagu@ebay.com) **(About first homework)**
* **Lecture 1**
  + Involve software and hardware vendors.
    - Teradata, Oracle, Micro strategy (platform may vary).
  + Use the University’s infrastructure
    - Dedicate a small Data Warehousing lab for the students to experiment and develop projects
* **Schedule:**
  + 29/9 Introduction
  + 10/10 Data integration
  + 17/10 Data modeling
  + 24/10 ETL Design
  + 31/10 BI & OLAP
  + 07/11 Architecture
  + 14/11 Lab work
  + 21/11 Presentation Preparation
  + 28/11 Final Presentation
* **My responsibilities**
  + Research related topics before and after the classes
  + Attend all classes and participate actively in discussions
  + Submit all homework within time limits and requested format
  + Participate in final discussion.
* **Todays topic** **- Data Warehouse Basics**
  + What is datawarehouse? Look up Russel Ackoff.
    - Data: Symbols
      * Ex: we are starting to lose our customers.
    - Information: data that are processed to be useful; provides answers to “who”, ”what”, “where”, and “when” questions.
      * Ex: Economy is down and then we stared losing customers
    - Knowledge: application of data and information; answers “how” questions.
      * Ex: Prices are going up, saleries are still the same.
    - Wisdom: evaluate understanding.
    - Understanding: appreciation of “why”
  + Definitions of Intelligence
    - What is?
  + Business Intelligence, definition:
    - A broad category of technologies, applications and practices for the collection, integration, analysis, and presentation of business information. The purpose of business intelligence is to support enterprise users make better business decision.
  + How a company uses BI
    - Customer Relationship department needs info on its customers to retain them.
    - Sales Dep needs info on which products are selling to which demographic population in which markets, by which channel.
    - Product planning dept needs info on sales rates in order to optimize plan for production.
    - Marketing dept needs info on who is buying which products in order to create intelligent marketing and advertising programs
    - The finance dept needs information on profitability in order to adapt cost structures for maximum values.
* What is Data Warehouse?
  + Subject-oriented – Data that gives information about a particular subject instead of about a company’s on-going operations.
    - Ex: improve trust
  + Integrated – Data that is gathered into the data warehouse form a variety of sources and wered (Merged?) into a coherent whole.
  + Time-variant
  + Non-volatile
* Why business needs a data warehouse?
  + Competitive strategy
    - Information is instrumental as a competitive differentiator
    - Not doing so may put you at a competitive disadvantage
    - Business is requiring more complex and real time information
  + Leverage and potential
    - Strategic knowledge is embedded in your data
    - You already own and have access to 95% of the data you would need
    - Already have the foundation of an enterprise wide, scalable infrastructure built with the best of breed technologies
  + Efficiencies and cost reduction
    - Reduce multiple redundant data sources, reducing maintenance and overhead
    - Single version of the truth reduces complexities and reconciliation efforts
    - Greater visibility to all data allows for better analysis and control of our cost drivers.
* **Terms I really need to know, look up on my own:**
  + **OLTP (Old) – Online transaction processing**
    - **OLTP (online transaction processing) is a class of program that facilitates and manages transaction-oriented applications, typically for data entry and retrieval transactions in a number of industries, including banking, airlines, mail-order, supermarkets, and manufacturers. Probably the most widely installed OLTP product is IBM's CICS (Customer Information Control System).**
  + **DSS (Old) – Decision support system**
    - **A decision support system (DSS) is a computer program application that analyzes business data and presents it so that users can make business decisions more easily. It is an "informational application" (to distinguish it from an "operational application" that collects the data in the course of normal business operation).Typical information that a decision support application might gather and present would be:**
      * **Comparative sales figures between one week and the next**
      * **Projected revenue figures based on new product sales assumptions**
      * **The consequences of different decision alternatives, given past experience in a context that is described**
    - **A decision support system may present information graphically and may include an expert system or artificial intelligence (AI). It may be aimed at business executives or some other group of knowledge workers.**
  + **OLCP (Today)**
    - **Online complex processing (OLCP) is a class of real-time data processing involving complex queries, lengthy queries and/or simultaneous reads and writes to the same records.**
    - **Processing complex queries, long transactions and simultaneous reads and writes to the same record. Contrast with OLTP, in which records are updated in a more predictable manner**
  + **OLAP (Today)**
    - **OLAP (online analytical processing) is computer processing that enables a user to easily and selectively extract and view data from different points of view. For example, a user can request that data be analyzed to display a spreadsheet showing all of a company's beach ball products sold in Florida in the month of July, compare revenue figures with those for the same products in September, and then see a comparison of other product sales in Florida in the same time period. To facilitate this kind of analysis, OLAP data is stored in a multidimensional database. Whereas a relational database can be thought of as two-dimensional, a multidimensional database considers each data attribute (such as product, geographic sales region, and time period) as a separate "dimension." OLAP software can locate the intersection of dimensions (all products sold in the Eastern region above a certain price during a certain time period) and display them. Attributes such as time periods can be broken down into sub attributes.**
    - **OLAP can be used for data mining or the discovery of previously undiscerned relationships between data items. An OLAP database does not need to be as large as a data warehouse, since not all transactional data is needed for trend analysis. Using Open Database Connectivity (ODBC), data can be imported from existing relational databases to create a multidimensional database for OLAP.**
    - **Two leading OLAP products are Hyperion Solution's Essbase and Oracle's Express Server. OLAP products are typically designed for multiple-user environments, with the cost of the software based on the number of users.**
* Data warehouse maturity model
  + Stage 1: Reporting WHAT happened?
    - Primarily batch
  + Stage 2: Analyzing WHY did it happen?
    - Increase in ad hoc queries
  + Stage 3: Predicting WHY will it happen?
    - Analytical modeling grows
  + Stage 4: Operationalizing WHAT IS happening?
    - Continuous update and time sensitive queries become important
  + Stage 3: Active warehousing MAKING it happen!
    - Event based something…
* Source Systems
  + Transaction Databases
  + Special Purpose applications (e.g ERP, CRM etc)
  + Legacy data
  + Data from External parties
  + Excel spreadsheets
  + Web traffic data
* Operational Data Store(ODS)
  + Two major interpretations
    1. ODS serves as a point of integration for operational systems
    2. ODS supplies current, detailed data for decision support.
  + The ODS should be a subject oriented, integrated, frequently updated store af detailed data to support transaction systems with integrated data.
* Data marts
  + A data mart is a subset of an organizational data store, usually oriented to a specific purpose or major data subject, that may be distributed to support…
* **ETL – Extract, Transform, Load**
  + **Extract** reads data from a specified database and writes a subset of data to a file
  + **Transform** changes the new data set using rules or lookup tables
  + **Load** writes the data set to a database intended to support reporting, queries, drill down and other decision support solutions
* ETL Evolution
  + First gen: Legacy code generators
    - Performance was very good because of the native complied code, hard to maintain.
  + Second gen: The proprietary ETL Engines
    - One ETL language resolves different languages on different platforms
    - Bottleneck from row by row processing
    - Slower
  + Third gen: The E-L-T Architecture (Extract, then just load, THEN transform)
    - Reduces network traffic to improve performance
    - Distributes the load for better scalability
    - Delegates the load
    - Supports all types of sources
    - Takes advantage of the architecture of the database itself
* **Dimensions**
  + One of the perspectives that can be used to analyze the data
  + When you are browsing the data, you can view the data from the perspective of different combinations of dimensions
  + For a sales database, the dimensions could include Product, Time, Store, and Promotion.
  + Dimensions contain one ore more hierarchies, which have levels for drilling up and drilling down in the data.
* **Facts (a.k.a Measures)**
  + In a star schema, a fact table is the central table which contains the individual facts being stored in the database
  + There are two types of fields in a fact table
    1. The fields storing the foreign keys which connect each particular fact to the appropriate value in each dimension
    2. The fields storing the individual facts (or measures) – such as number, amount, or price
  + The granularity of the fact table is one of the most significant design decisions in creating a data warehouse. The facts should be as detailed as possible to allow for the data to be viewed from the greatest number of perspectives.
* **Star schema**
* **Snowflake Schema**
* To avoid joins, redundancies in warehouses can be accepted, denormalize.
* **Data Warhouse Architecture**
  + A sound data warehouse architecture enables:
    - Extensibility
    - Reusability
    - Improved productivity
  + Data architecture plan
    - Data is the “What” that the data warehouse is all about
    - The data area of the architecture includes the contents of the data warehouse:
      * The list of data that is important to the business
      * The data stores that make up the overall data warehouse environment
      * The data sources that feed into the data warehouse
      * Design of the logical and physical data models, aggregations, hierarchies, and so on, based on business needs
      * The granularity, volume and timing of the data at various points in the data warehouse
* OLAP – online analytical processing
* Gather reporting requirements for the library data warehouse BI reports
* Get familiar with the library database
* Think about what information the library management team would need to run the library more effectively. Write up a summary (1-2 pages) of your thoughts.
* Print you homework and hand over to the teacher (one per group)
* Research on the internet for 1 successful and 1 unsuccessful data warehouse project. Prepare 1 paragraph story on why they were successful or why they where not.