CSPP 53017: Data Warehousing

Winter 2013

Lecture 2

Svetlozar Nestorov

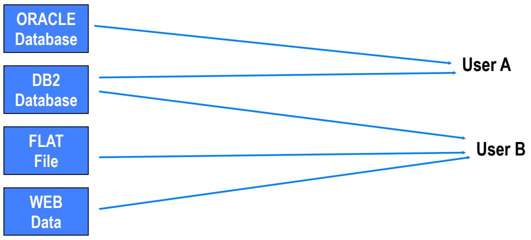
Class News

- Class web page: http://bit.ly/WTWXV9
- Subscribe to the mailing list
- Homework 1 is out now; due by 1:59am on Tue, Jan 29.
  - Project draft proposal
  - Aggregates, duplicates, and NULLs on Gradiance
- 15 minute in-class quiz next week
  - Covers the first two lectures and the Gradiance homework.

Basic Elements of the Data Warehouse

- Source Systems
  - Operational systems whose function is to capture the transactions of the business
- ETL System
  - Used for **ETL – Extraction, Transformation, and Load**
  - ETL includes a set of processes used to clean, transform, combine, de-duplicate, archive, and prepare source data for use in the data warehouse
- Target System
  - Data warehouse
- Presentation Server
  - Physical machine on which the data warehouse data is organized and stored

Example Scenario



ORACLE Database

DB2 Database

FLAT File

WEB Data

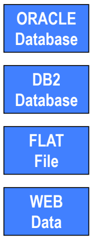
User A

User B

Source Systems

Possible Scenario Within
an Enterprise (Blue – Operational, Red – Analytical)

Example Scenario



ORACLE Database

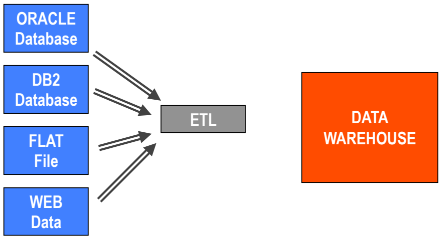
DB2 Database

FLAT File

WEB Data

Source Systems

Example Scenario



ORACLE Database

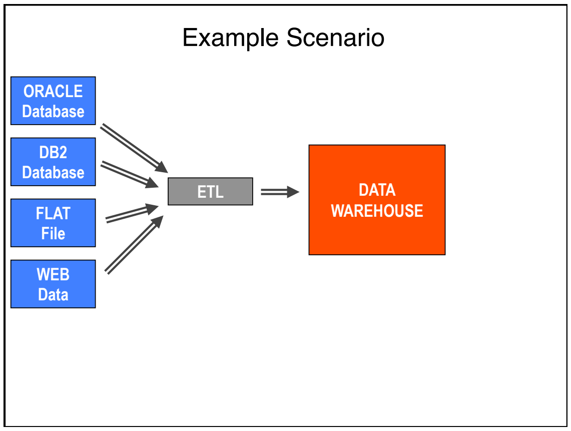
DB2 Database

FLAT File

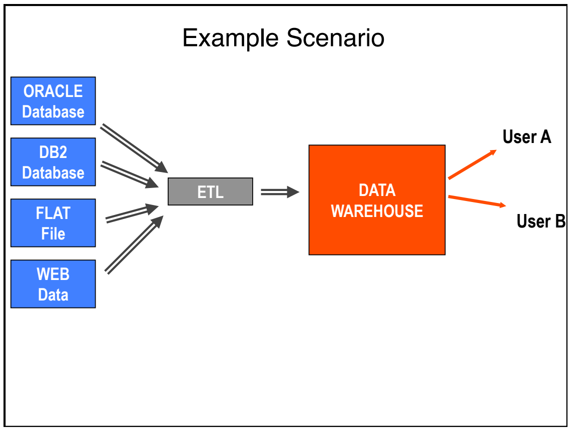
WEB Data

ETL

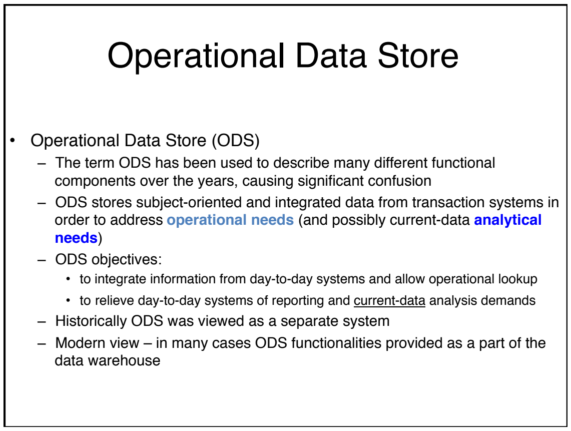
DATA WAREHOUSE



Example Scenario



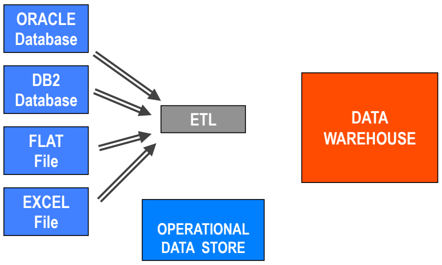
Example Scenario



Operational Data Store

- Operational Data Store (ODS)
  - The term ODS has been used to describe many different functional components over the years, causing significant confusion
  - ODS stores subject-oriented and integrated data from transaction systems in order to address **operational needs** (and possibly current-data **analytical needs**)
  - ODS objectives:
    - to integrate information from day-to-day systems and allow operational lookup
    - to relieve day-to-day systems of reporting and current-data analysis demands
  - Historically ODS was viewed as a separate system
  - Modern view – in many cases ODS functionalities provided as a part of the data warehouse

Example Scenario



ORACLE Database

DB2 Database

FLAT File

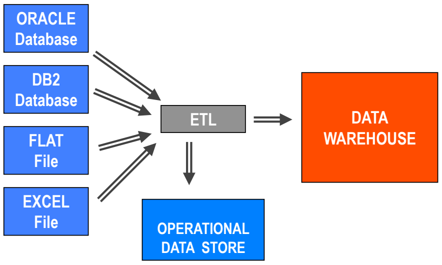
EXCEL File

ETL

DATA WAREHOUSE

OPERATIONAL DATA STORE

Example Scenario



ORACLE Database

DB2 Database

FLAT File

EXCEL File

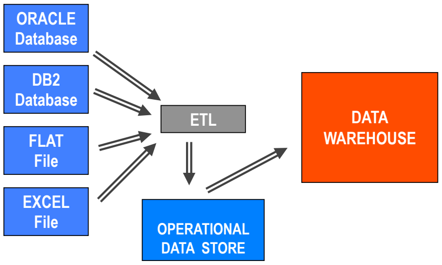
ETL

DATA WAREHOUSE

OPERATIONAL DATA STORE

Possible Scenario Within
an Enterprise

Example Scenario



ORACLE Database

DB2 Database

FLAT File

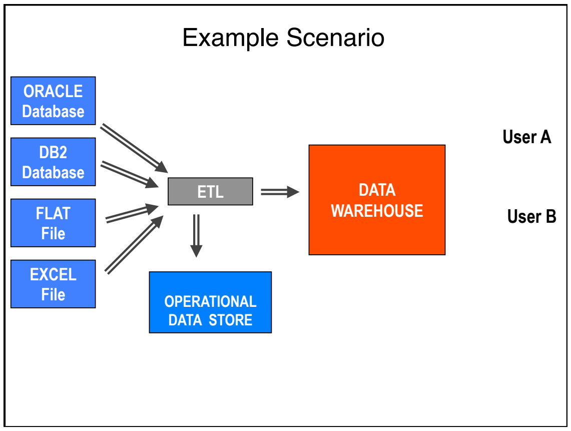
EXCEL File

ETL

DATA WAREHOUSE

OPERATIONAL DATA STORE

OR this
(it really depends if the scope of the
ODS and DWH is the same)



Example Scenario

ORACLE Database

DB2 Database

FLAT File

EXCEL File

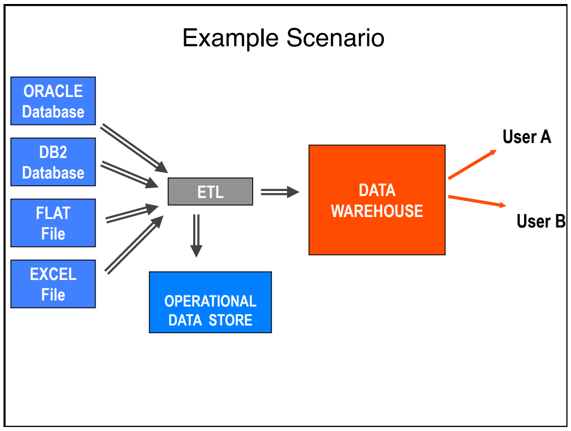
ETL

DATA WAREHOUSE

OPERATIONAL DATA STORE

User A

User B



Example Scenario

ORACLE Database

DB2 Database

FLAT File

EXCEL File

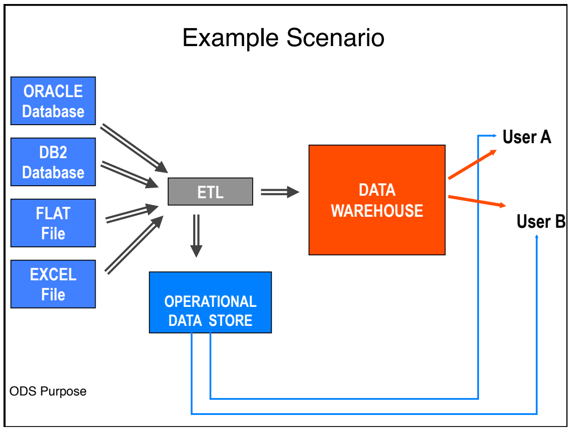
ETL

DATA WAREHOUSE

OPERATIONAL DATA STORE

User A

User B



Example Scenario

ORACLE Database

DB2 Database

FLAT File

EXCEL File

ETL

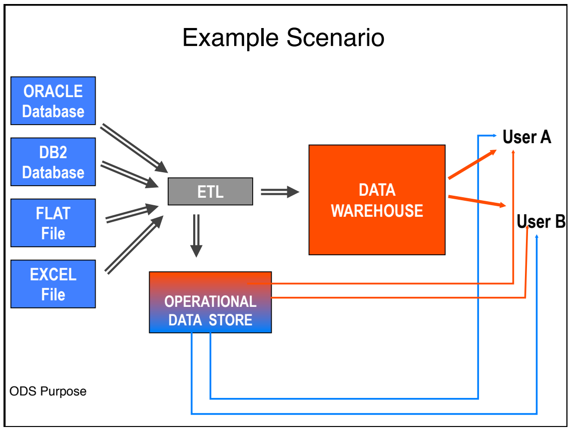
DATA WAREHOUSE

OPERATIONAL DATA STORE

User A

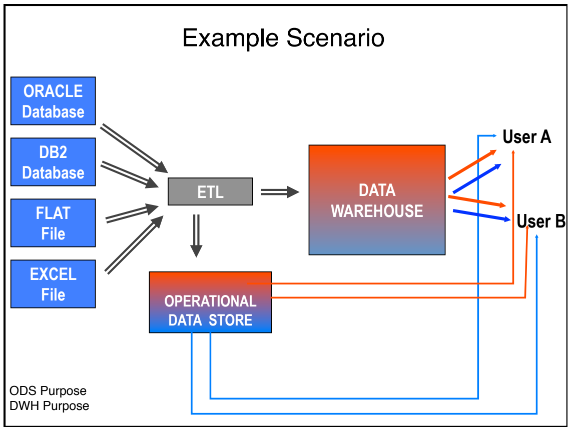
User B

ODS Purpose



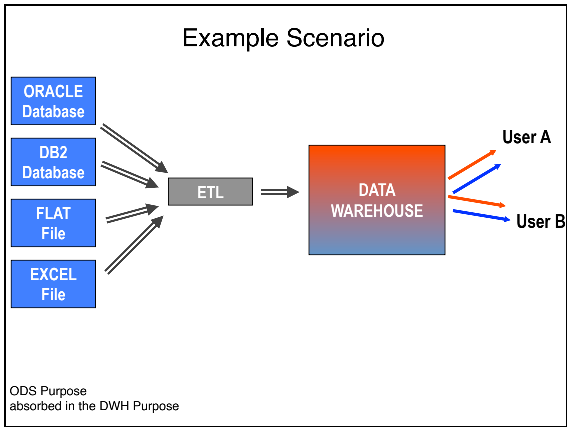
Example Scenario

ODS Purpose



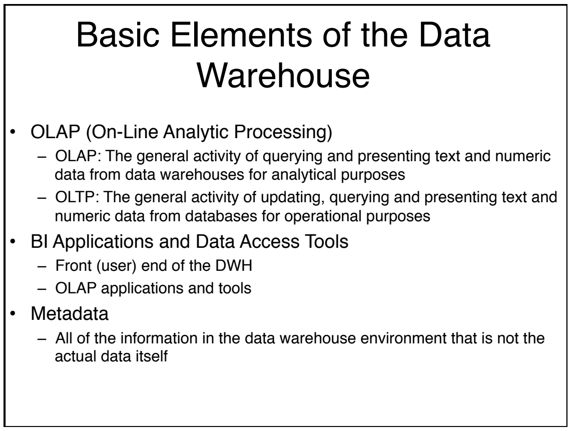
Example Scenario

ODS Purpose
DWH Purpose



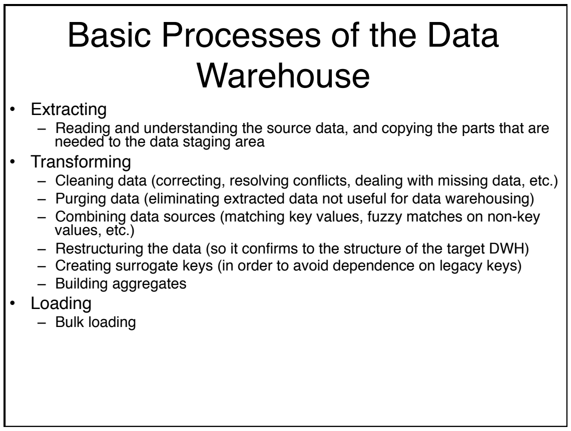
Example Scenario

ODS Purpose
absorbed in the DWH Purpose



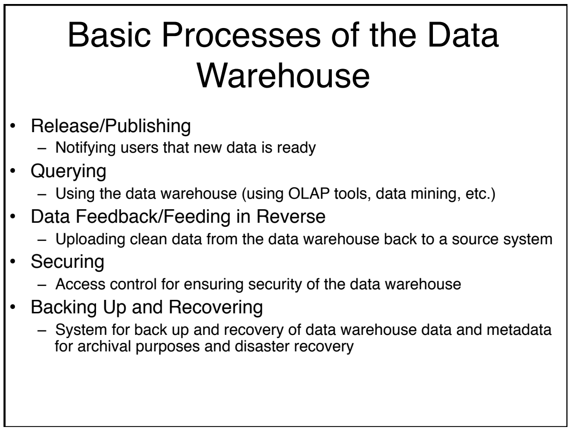
Basic Elements of the Data Warehouse

- OLAP (On-Line Analytic Processing)
  - OLAP: The general activity of querying and presenting text and numeric data from data warehouses for analytical purposes
  - OLTP: The general activity of updating, querying and presenting text and numeric data from databases for operational purposes
- BI Applications and Data Access Tools
  - Front (user) end of the DWH
  - OLAP applications and tools
- Metadata
  - All of the information in the data warehouse environment that is not the actual data itself



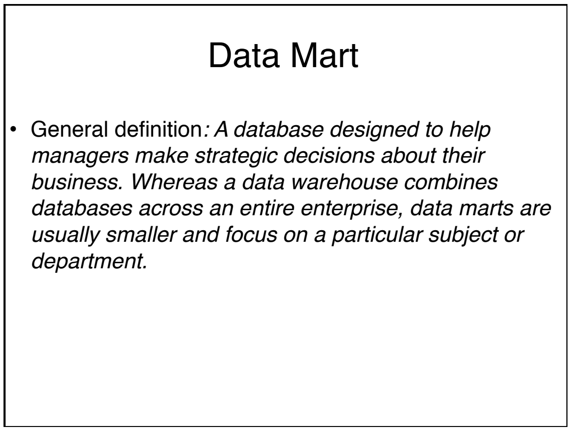
Basic Processes of the Data Warehouse

- Extracting
  - Reading and understanding the source data, and copying the parts that are needed to the data staging area
- Transforming
  - Cleaning data (correcting, resolving conflicts, dealing with missing data, etc.)
  - Purging data (eliminating extracted data not useful for data warehousing)
  - Combining data sources (matching key values, fuzzy matches on non-key values, etc.)
  - Restructuring the data (so it confirms to the structure of the target DWH)
  - Creating surrogate keys (in order to avoid dependence on legacy keys)
  - Building aggregates
- Loading
  - Bulk loading



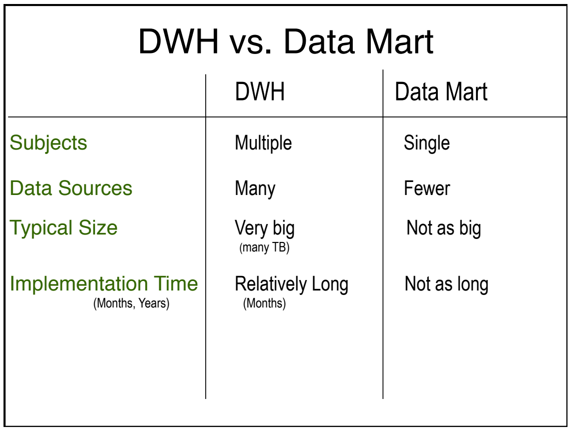
Basic Processes of the Data Warehouse

- Release/Publishing
  - Notifying users that new data is ready
- Querying
  - Using the data warehouse (using OLAP tools, data mining, etc.)
- Data Feedback/Feeding in Reverse
  - Uploading clean data from the data warehouse back to a source system
- Securing
  - Access control for ensuring security of the data warehouse
- Backing Up and Recovering
  - System for back up and recovery of data warehouse data and metadata for archival purposes and disaster recovery



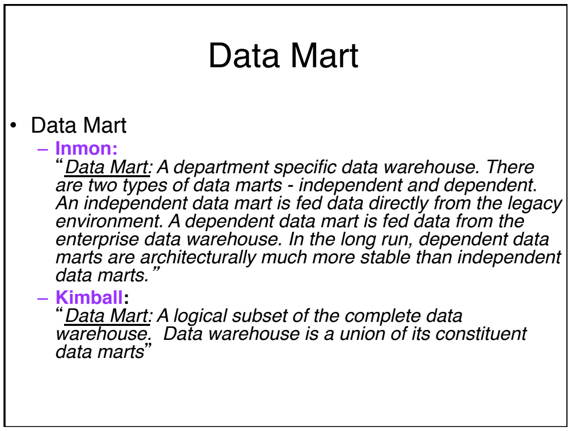
Data Mart

- General definition*: A database designed to help managers make strategic decisions about their business. Whereas a data warehouse combines databases across an entire enterprise, data marts are usually smaller and focus on a particular subject or department.*



DWH vs. Data Mart

	DWH	Data Mart
Subjects	Multiple	Single
Data Sources	Many	Fewer
Typical Size	Very big (many TB)	Not as big
Implementation Time (Months, Years)	Relatively Long (Months)	Not as long



Data Mart

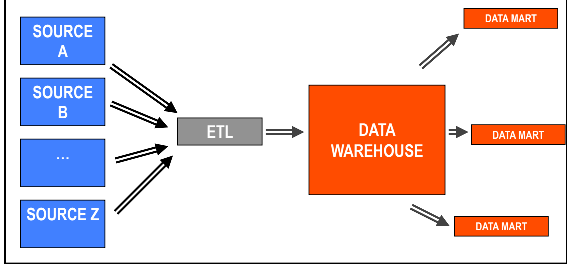
- Data Mart
  - **Inmon:**
    "*Data Mart: A department specific data warehouse. There are two types of data marts - independent and dependent. An independent data mart is fed data directly from the legacy environment. A dependent data mart is fed data from the enterprise data warehouse. In the long run, dependent data marts are architecturally much more stable than independent data marts.*"
  - **Kimball:**
    "*Data Mart: A logical subset of the complete data warehouse. Data warehouse is a union of its constituent data marts*"

Data Mart

- Data Mart
  - **Inmon:**
    "*Data Mart: A department specific data warehouse. There are two types of data marts - independent and dependent. An independent data mart is fed data directly from the legacy environment. A dependent* ***data mart is fed data from the enterprise data warehouse.*** *In the long run, dependent data marts are architecturally much more stable than independent data marts.*"
  - **Kimball:**
    "*Data Mart: A logical subset of the complete data warehouse.* ***Data warehouse is a union of its constituent data marts***"

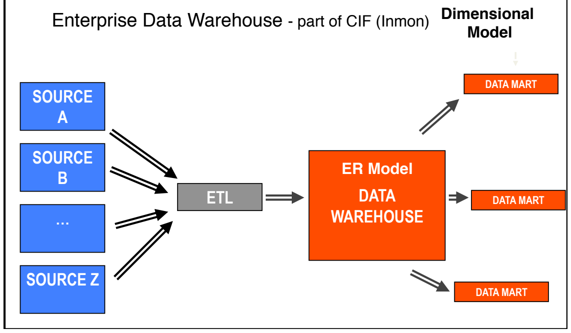
Data Warehouse Architecture Choices

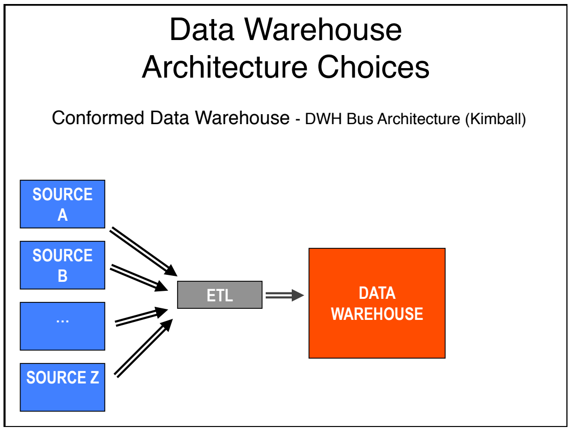
Enterprise Data Warehouse - part of CIF (Inmon)



Data Warehouse Architecture Choices

Enterprise Data Warehouse - part of CIF (Inmon)





Data Warehouse Architecture Choices

Conformed Data Warehouse - DWH Bus Architecture (Kimball)

SOURCE A

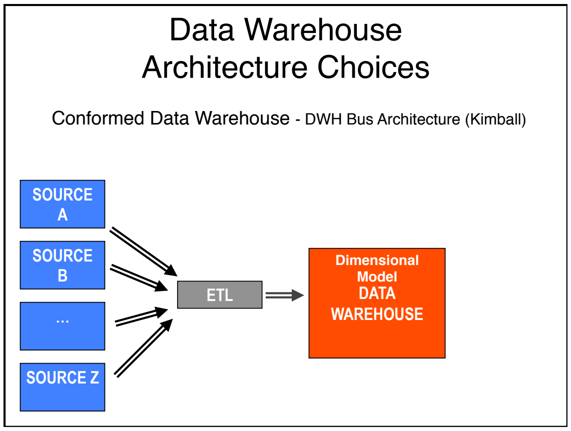
SOURCE B

...

SOURCE Z

ETL

DATA WAREHOUSE



Data Warehouse Architecture Choices

Conformed Data Warehouse - DWH Bus Architecture (Kimball)

SOURCE A

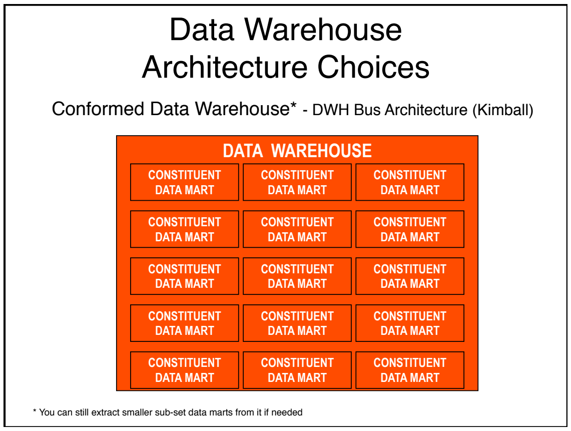
SOURCE B

...

SOURCE Z

ETL

Dimensional Model DATA WAREHOUSE



Data Warehouse Architecture Choices

Conformed Data Warehouse* - DWH Bus Architecture (Kimball)

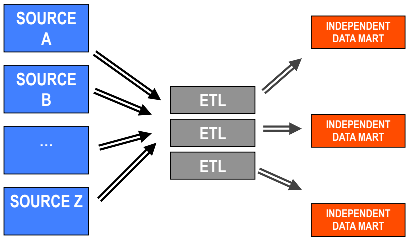
DATA WAREHOUSE

CONSTITUENT DATA MART	CONSTITUENT DATA MART	CONSTITUENT DATA MART
CONSTITUENT DATA MART	CONSTITUENT DATA MART	CONSTITUENT DATA MART
CONSTITUENT DATA MART	CONSTITUENT DATA MART	CONSTITUENT DATA MART
CONSTITUENT DATA MART	CONSTITUENT DATA MART	CONSTITUENT DATA MART
CONSTITUENT DATA MART	CONSTITUENT DATA MART	CONSTITUENT DATA MART

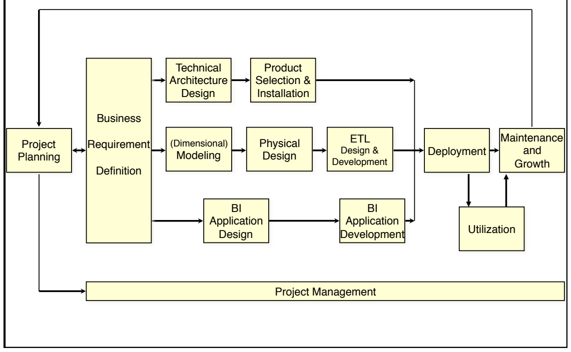
* You can still extract smaller sub-set data marts from it if needed

Data Warehouse Architecture Choices

Independent Data Marts ("bad" choice, but very common)

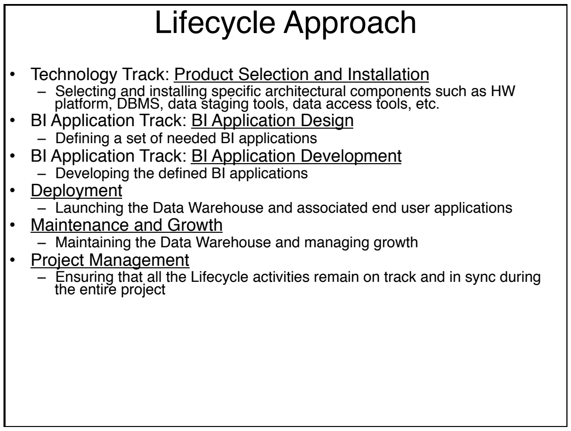


The DWH/BI Lifecycle



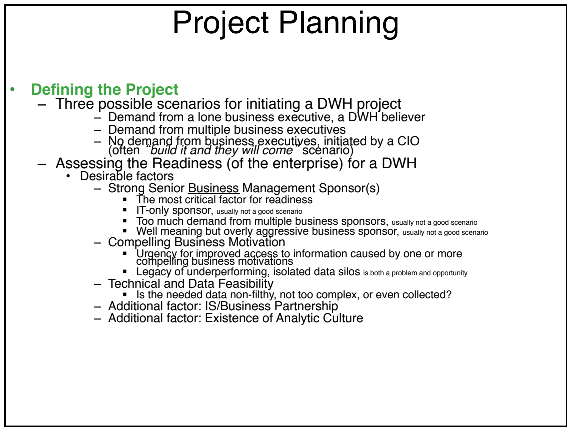
Lifecycle Approach

- Project Planning
  - Assessing and planning the project
- Business Requirements Definition
  - Defining and collecting the requirements **(the most critical step)**
- Dimensional (and/or ER) Modeling
  - Modeling the Data Warehouse
- Data Track: Physical Design
  - Defining the physical structures for supporting the Data Warehouse (e.g. indexing and partitioning)
- Data Track: ETL Design and Development
  - Designing and developing extraction, transformation, and load processes
- Technology Track: Technical Architecture Design
  - Defining and/or designing the custom code, home grown utilities (specific programs for managing computer resources) and of-the-shelf tools necessary for data acquisition and data access



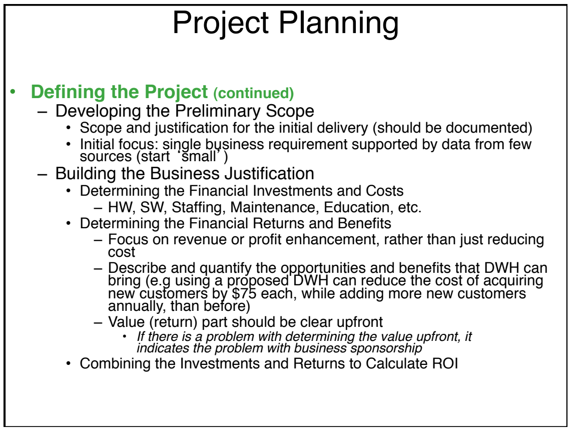
Lifecycle Approach

- Technology Track: Product Selection and Installation
  - Selecting and installing specific architectural components such as HW platform, DBMS, data staging tools, data access tools, etc.
- BI Application Track: BI Application Design
  - Defining a set of needed BI applications
- BI Application Track: BI Application Development
  - Developing the defined BI applications
- Deployment
  - Launching the Data Warehouse and associated end user applications
- Maintenance and Growth
  - Maintaining the Data Warehouse and managing growth
- Project Management
  - Ensuring that all the Lifecycle activities remain on track and in sync during the entire project



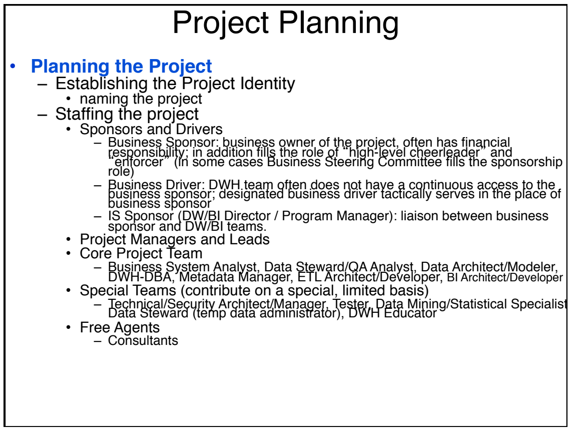
Project Planning

- **Defining the Project**
  - Three possible scenarios for initiating a DWH project
    - Demand from a lone business executive, a DWH believer
    - Demand from multiple business executives
    - No demand from business executives, initiated by a CIO (often *“build it and they will come”* scenario)
  - Assessing the Readiness (of the enterprise) for a DWH
    - Desirable factors
      - Strong Senior Business Management Sponsor(s)
        - The most critical factor for readiness
        - IT-only sponsor, usually not a good scenario
        - Too much demand from multiple business sponsors, usually not a good scenario
        - Well meaning but overly aggressive business sponsor, usually not a good scenario
      - Compelling Business Motivation
        - Urgency for improved access to information caused by one or more compelling business motivations
        - Legacy of underperforming, isolated data silos is both a problem and opportunity
      - Technical and Data Feasibility
        - Is the needed data non-filthy, not too complex, or even collected?
      - Additional factor: IS/Business Partnership
      - Additional factor: Existence of Analytic Culture



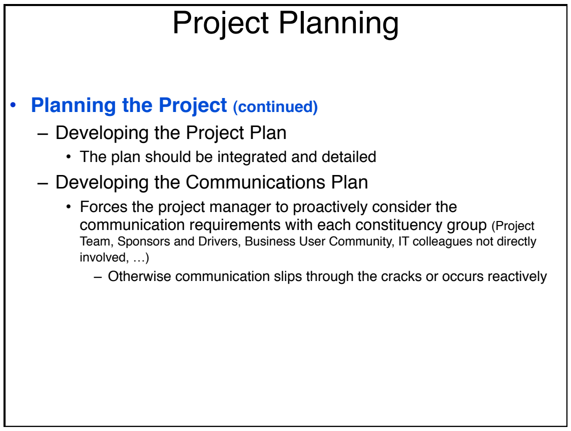
Project Planning

- **Defining the Project (continued)**
  - Developing the Preliminary Scope
    - Scope and justification for the initial delivery (should be documented)
    - Initial focus: single business requirement supported by data from few sources (start “small”)
  - Building the Business Justification
    - Determining the Financial Investments and Costs
      - HW, SW, Staffing, Maintenance, Education, etc.
    - Determining the Financial Returns and Benefits
      - Focus on revenue or profit enhancement, rather than just reducing cost
      - Describe and quantify the opportunities and benefits that DWH can bring (e.g using a proposed DWH can reduce the cost of acquiring new customers by $75 each, while adding more new customers annually, than before)
      - Value (return) part should be clear upfront
        - *If there is a problem with determining the value upfront, it indicates the problem with business sponsorship*
    - Combining the Investments and Returns to Calculate ROI



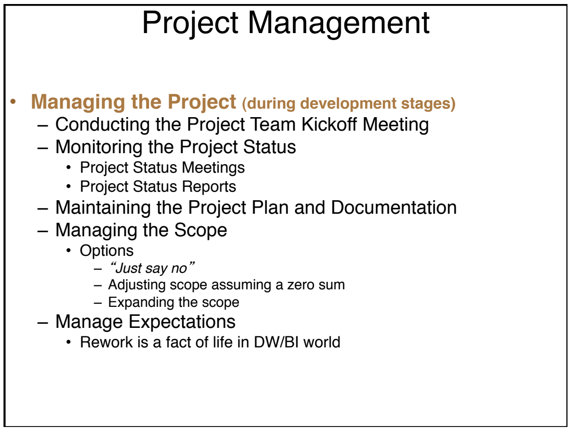
Project Planning

- **Planning the Project**
  - Establishing the Project Identity
    - naming the project
  - Staffing the project
    - Sponsors and Drivers
      - Business Sponsor: business owner of the project, often has financial responsibility; in addition fills the role of "high-level cheerleader" and "enforcer" (in some cases Business Steering Committee fills the sponsorship role)
      - Business Driver: DWH team often does not have a continuous access to the business sponsor; designated business driver tactically serves in the place of business sponsor
      - IS Sponsor (DW/BI Director / Program Manager): liaison between business sponsor and DW/BI teams.
    - Project Managers and Leads
    - Core Project Team
      - Business System Analyst, Data Steward/QA Analyst, Data Architect/Modeler, DWH-DBA, Metadata Manager, ETL Architect/Developer, BI Architect/Developer
    - Special Teams (contribute on a special, limited basis)
      - Technical/Security Architect/Manager, Tester, Data Mining/Statistical Specialist Data Steward (temp data administrator), DWH Educator
    - Free Agents
      - Consultants



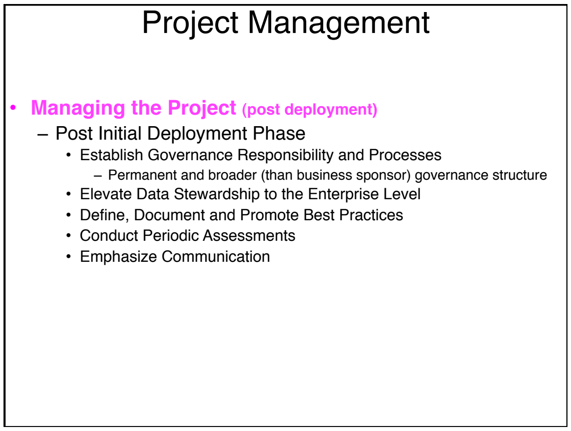
Project Planning

- **Planning the Project (continued)**
  - Developing the Project Plan
    - The plan should be integrated and detailed
  - Developing the Communications Plan
    - Forces the project manager to proactively consider the communication requirements with each constituency group (Project Team, Sponsors and Drivers, Business User Community, IT colleagues not directly involved, …)
      - Otherwise communication slips through the cracks or occurs reactively



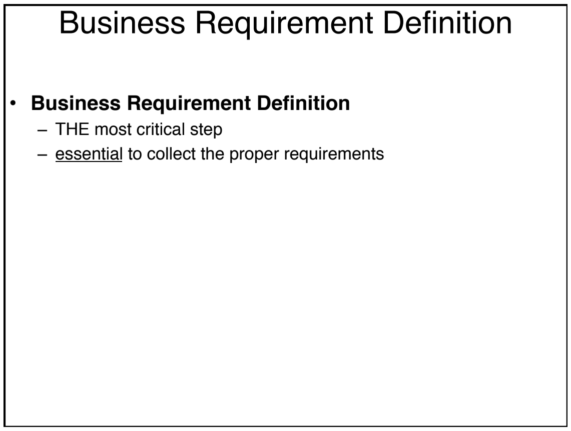
Project Management

- **Managing the Project (during development stages)**
  - Conducting the Project Team Kickoff Meeting
  - Monitoring the Project Status
    - Project Status Meetings
    - Project Status Reports
  - Maintaining the Project Plan and Documentation
  - Managing the Scope
    - Options
      - *"Just say no"*
      - Adjusting scope assuming a zero sum
      - Expanding the scope
  - Manage Expectations
    - Rework is a fact of life in DW/BI world



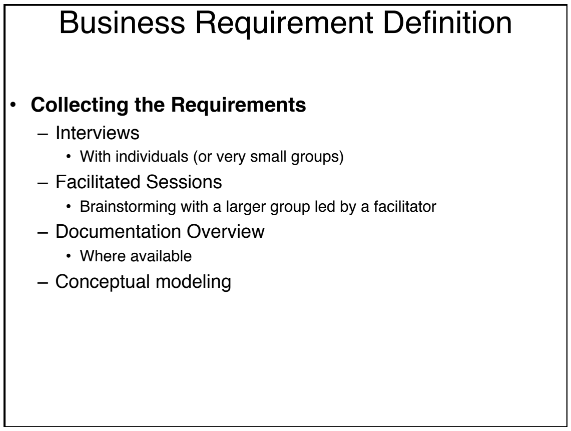
Project Management

- **Managing the Project** **(post deployment)**
  - Post Initial Deployment Phase
    - Establish Governance Responsibility and Processes
      - Permanent and broader (than business sponsor) governance structure
    - Elevate Data Stewardship to the Enterprise Level
    - Define, Document and Promote Best Practices
    - Conduct Periodic Assessments
    - Emphasize Communication



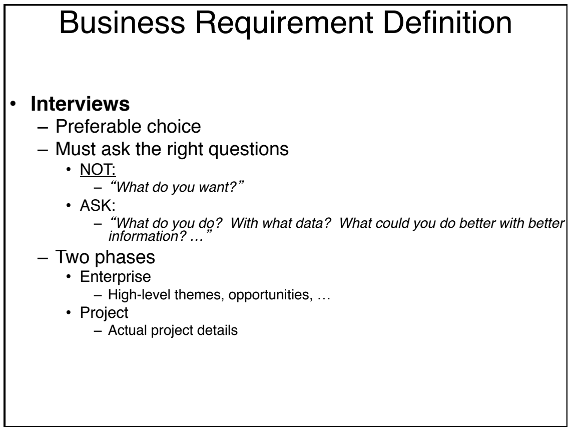
Business Requirement Definition

- **Business Requirement Definition**
  - THE most critical step
  - essential to collect the proper requirements



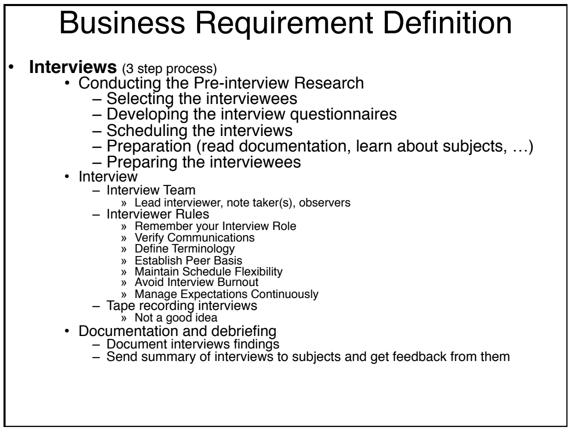
Business Requirement Definition

- **Collecting the Requirements**
  - Interviews
    - With individuals (or very small groups)
  - Facilitated Sessions
    - Brainstorming with a larger group led by a facilitator
  - Documentation Overview
    - Where available
  - Conceptual modeling



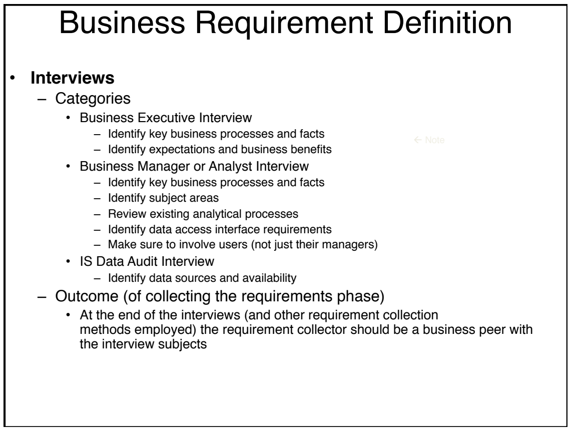
Business Requirement Definition

- **Interviews**
  - – Preferable choice
  - – Must ask the right questions
    - NOT:
      - – *"What do you want?"*
    - ASK:
      - – *"What do you do? With what data? What could you do better with better information? …"*
  - – Two phases
    - Enterprise
      - – High-level themes, opportunities, …
    - Project
      - – Actual project details



Business Requirement Definition

- **Interviews** (3 step process)
  - Conducting the Pre-interview Research
    - – Selecting the interviewees
    - – Developing the interview questionnaires
    - – Scheduling the interviews
    - – Preparation (read documentation, learn about subjects, …)
    - – Preparing the interviewees
  - Interview
    - – Interview Team
      - » Lead interviewer, note taker(s), observers
    - – Interviewer Rules
      - » Remember your Interview Role
      - » Verify Communications
      - » Define Terminology
      - » Establish Peer Basis
      - » Maintain Schedule Flexibility
      - » Avoid Interview Burnout
      - » Manage Expectations Continuously
    - – Tape recording interviews
      - » Not a good idea
  - Documentation and debriefing
    - – Document interviews findings
    - – Send summary of interviews to subjects and get feedback from them



Business Requirement Definition

- **Interviews**
  - – Categories
    - Business Executive Interview
      - – Identify key business processes and facts
      - – Identify expectations and business benefits

← Note

    - Business Manager or Analyst Interview
      - – Identify key business processes and facts
      - – Identify subject areas
      - – Review existing analytical processes
      - – Identify data access interface requirements
      - – Make sure to involve users (not just their managers)
    - IS Data Audit Interview
      - – Identify data sources and availability
  - – Outcome (of collecting the requirements phase)
    - At the end of the interviews (and other requirement collection methods employed) the requirement collector should be a business peer with the interview subjects