Universitas Esa Unggul

Smart, Creative and Entrepreneurial

Data Warehouse

Munawar, PhD

Session 03

Data Warehouse Development Phase

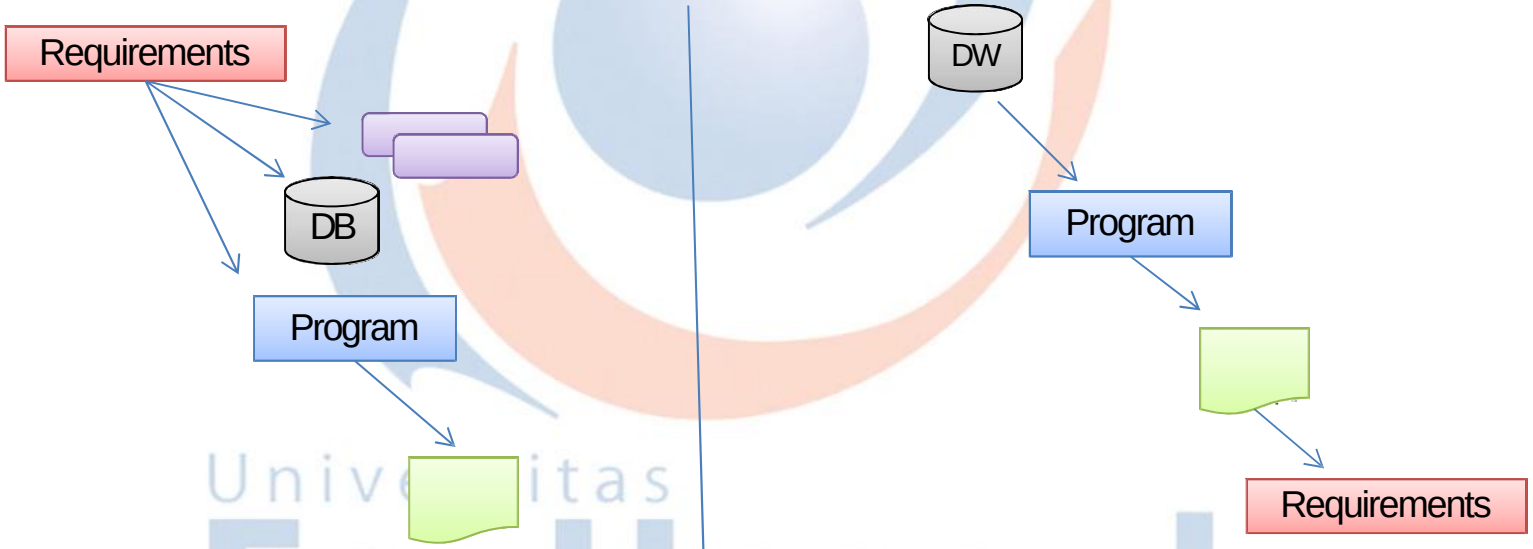
Agenda

- Data Warehouse Life Cycle
- Classical DW Dev Life Cycle
- Operating a DW
- Q/A ?

Data Warehouse Dev Life Cycle

Live Cycle of DWs

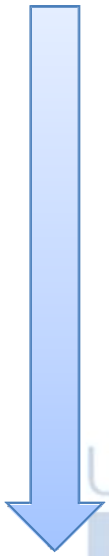
- **S**ystem **D**evelopment **L**ife **C**ycle (SDLC)
  - Classical SDLC vs. DW SDLC



  - DW SDLC is almost the **opposite** of classical SDLC, since requirements are not known from the beginning

Live Cycle of DWs …

- **Classical SDLC vs. DW SDLC**



Classical SDLC	DW SDLC
Requirements gathering	Implement warehouse
Analysis	Integrate data
Design	Test for bias
Programming	Program against data
Testing	Design DSSsystem
Integration	Analyze results
Implementation	Understand requirements

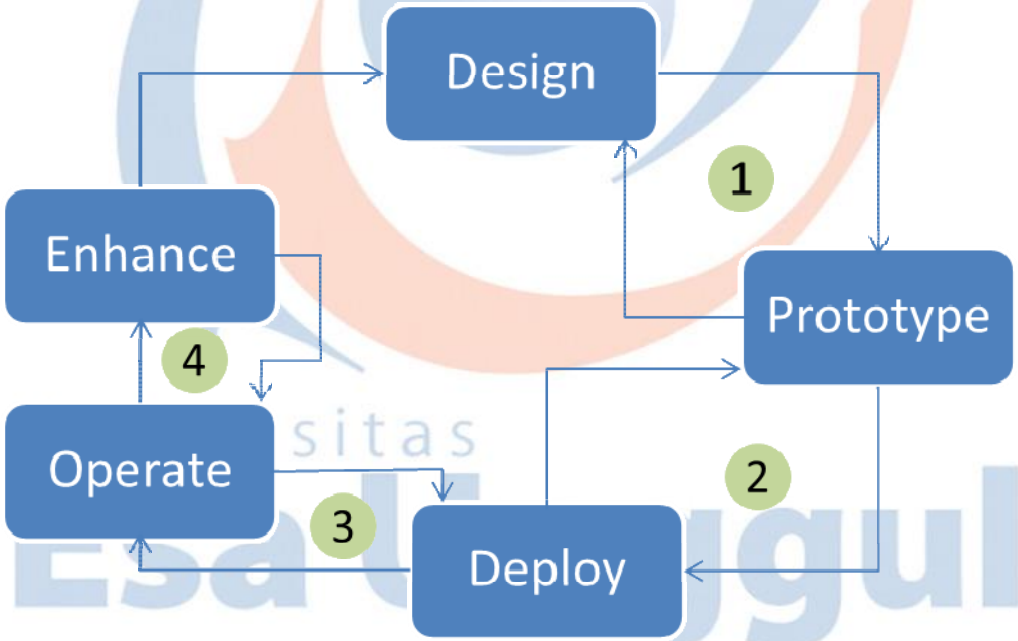
  - Because it is the opposite of SDLC, DW SDLC is also called CLDS

Live Cycle of DWs …

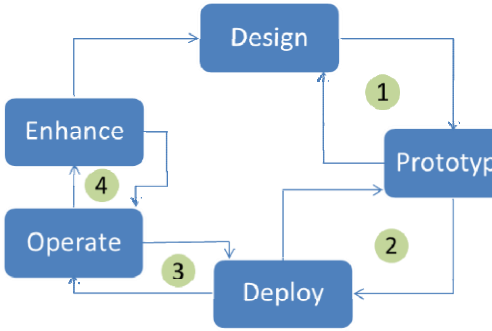
- CLDS is a **data driven** development life cycle
  - It starts with data
    - Once data is at hand it is integrated and tested against bias
    - Programs are written against the data and the results are analyzed and finally the requirements of the system are understood
    - Once requirements are understood, adjustments are made to the design and the cycle starts all over
  - **"spiral development methodology"**

Live Cycle of DWs …

- Lifecycle phases



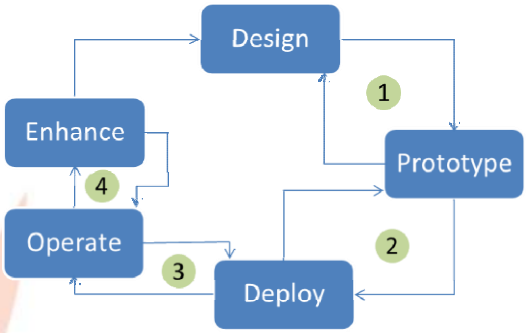
Live Cycle of DWs …



- **Design**
  - Interviewing the end-users in cycles
  - Analyzing the data source system (ODS)
  - Defining the key performance indicators
  - Mapping the decision-making processes to the underlying information needs
  - Logical and physical schema design
- **Prototype**
  - Objective is to **constrain** and in some cases **reframe** end-user requirements

Live Cycle of DWs …

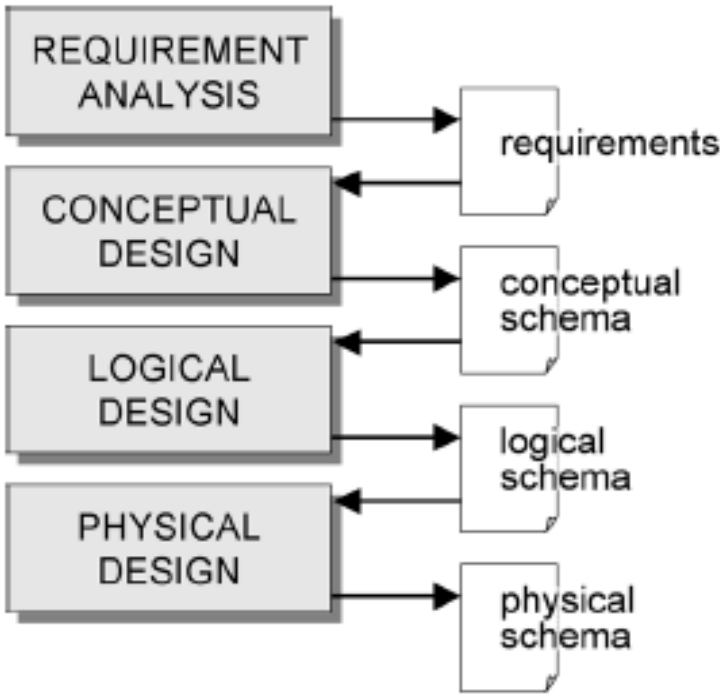
- **Deployment**
  - Development of documentation
  - Personal training
  - Operations and management processes



- **Operation**
  - Day-to-day maintenance of the DW needs a good management of ongoing **E**xtraction, **T**ransformation and **L**oading (ETL) process
- **Enhancement** requires the modification of
  - HW - physical components
  - Operations and management processes
  - Logical schema designs

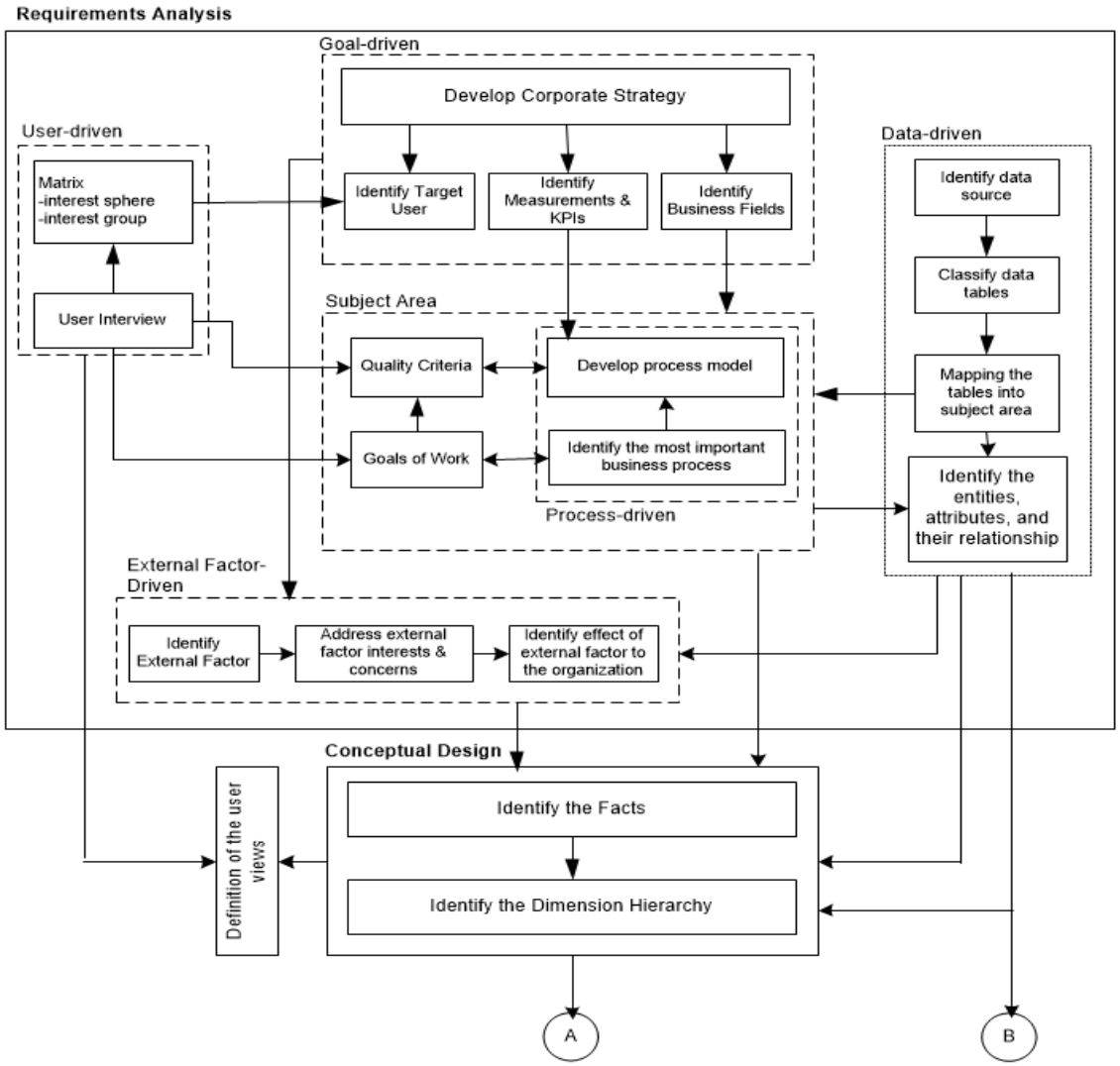
Classical DW Dev Life Cycle

Core of DW Dev Life Cycle



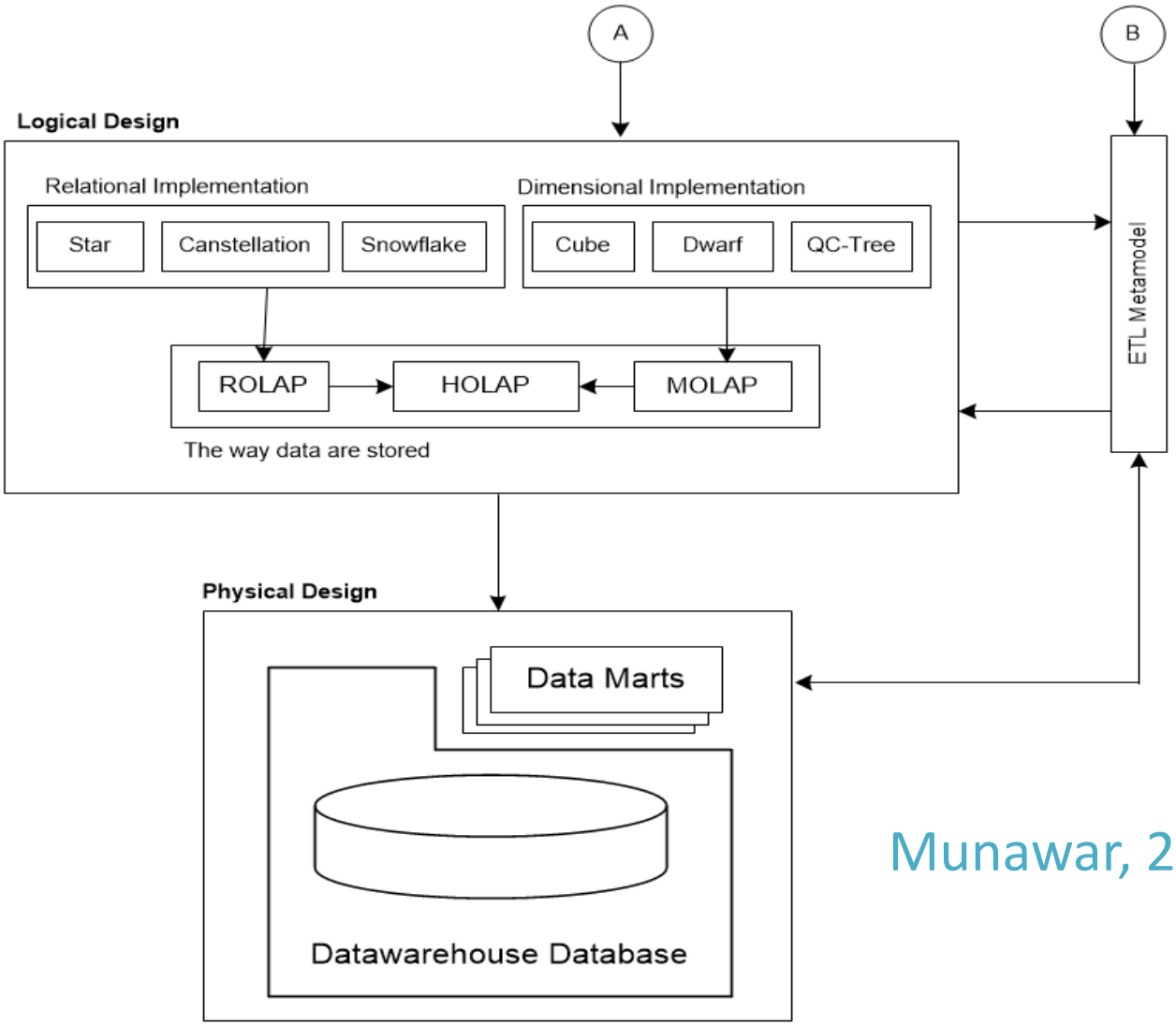
Rizzi, 2009

Common Practices DW Dev Life Cycle



Munawar, 2016

Common Practices DW Dev Life Cycle…



Munawar, 2016

Requirements Analysis

Requirements Analysis	Strengths	Weaknesses
User-driven	• Involvement of end users is essentials in DW projects to ensure the successful use of DW (Niedrite et al, 2009)	• Unclear of users' understanding of DW, business strategies or organizational processes, make the degree of obsolescence of resulting schemata is high (Niedrite et al, 2009) • It takes time expensive to achieve consensus on requirements with many different point of view (Niedrite et al, 2009;Lujan-Mora, 2002)
Data-driven	• The fastest way to define a DW model (Niedrite et al, 2009) • Simpler (Winter & Strauch, 2003) • Very stable (Winter & Strauch, 2003)	• Such models perhaps do not reflect all of the facts that are needed in analysis business goals (Niedrite et al, 2009)] • User involvement is limited (Winter & Strauch, 2003) • Multidimensional schemata produced could not match with user requirements if information is not actually present in the data source (Winter & Strauch, 2003)
Goal-driven	• Correct identification of the relevant indicators can be obtained (Niedrite et al, 2009;Lujan-Mora, 2002)	• Very dependable to the participation of top management in requirements analysis process [Lujan-Mora, 2002) • Need high capable staff in translation process from the collected high level requirements into quantifiable KPIs (Niedrite et al, 2009;Lujan-Mora, 2002) • It is hard to predict the needs of all senior managements (Niedrite et al, 2009)

All these approaches are complementary and should be used in parallel to achieve optimal design (Guo et al, 2006, Olive

Requirements Analysis …

Requirements Analysis	Strengths	Weaknesses
Process-driven	• Essential business processes and indicators to measure these processes are identified (Niedrite et al, 2009;Lujan-Mora, 2002) • Close to the need of business and adaptive with the business environment (Guo et all, 2006)	• The model reflect business processes not process of decision making (Niedrite et al, 2009)
External-driven	• Comply with governmental regulations or others external pressure require a disclosure about business operations (Frolick & Ariyachandra, 2006)	• Sometimes external data is needed to generate required information

All these approaches are complementary and should be used in parallel to achieve optimal design (Guo et al, 2006, Oliveira et al, 2012)

Conceptual Design

- The conceptual design allows having closer idea about the ways that a user can perceive an application domain (Saxena & Agarval, 2014) using multidimensional schemata through facts & their properties and distinguish dimensions & categorize them into hierarchies
- Existing approach to multidimensional modelling
  - M E/R (Sapia et al, 1999)
  - UML class diagram (Lujan-Mora et al, 2002)
  - Fact Schema (Husemann et al, 2000)
  - Dimensional Fact Model (Golfarelli, 2010)

Logical Design

Relational Implementation for multidimensional modeling (Mishra et al, 2008)

	Star Schema	Fact Constellation Schema	Snowflake Schema
Efficiency	High	High	Moderate
Usability	High	Moderate	Moderate
Reusability	Low	Low	High
Flexibility	High	High	Moderate
Redundancy	High	High	Low
Complexity	Low	Moderate	Moderate

Dimensional Implementation for multidimensional modeling

	Condensed Cube (Feng et al, 2004)	Dwarf (Sismanis et al, 2003)	QC-Tree (Laksmanan, 2003)
Size	Much smaller size of non-condensed cube	Highly compressed and clustered data cubes	Very compact data structure
Compression	• Fully pre-computed cube without compression • Neither decompression or further aggregation is required when answering queries	• Complete architecture that support queries, updates and roll-up data • A tunable granularity parameter that controls the amount of materialization performed	• It is elegant and lean in that only information it keeps on classes are their upper bound and measure(s)

Logical Design…

Data warehouse schema comparison

DW Schema	Advantages	Drawbacks
Star Schema	• The simplest structure (Moody & Kortink, 2008) • Reduces the number of tables (Basaran, 2005) • The number of relationships between the tables can be reduced (Basaran, 2005) • It reduces the number of joins required in user queries (Basaran, 2005) • It speed up query performance	• Very inflexible (Teklitz, 2000) • For every gigabyte of row data a schema will require at least an additional gigabytes for aggregations (Teklitz, 2000) • The amount of development maintenance effort needed to manage schema oriented DW (Teklitz, 2000)
Fact Canstellation Schema	• It reuses the dimension tables to save storage space (Levene & Loizou, 2003)	• It may not be useful for small organization because of its complexity (Feng et al, 2004)
Snowflake Schema	• It shows explicitly the hierarchical structures of each dimension (Teklitz, 2000) • It is intuitive and easy to understand (Arfaoui & Akaichi, 2012) • It can accommodate for aggregate data Arfaoui & Akaichi, 2012) • It is easily extensible by adding new attributes without inferring with existing database programs Arfaoui & Akaichi, 2012)	• It adds unnecessary complexity (Teklitz, 2000) • It reduces query performance (Teklitz, 2000)

ETL

- ETL is used to integrate heterogeneous systems

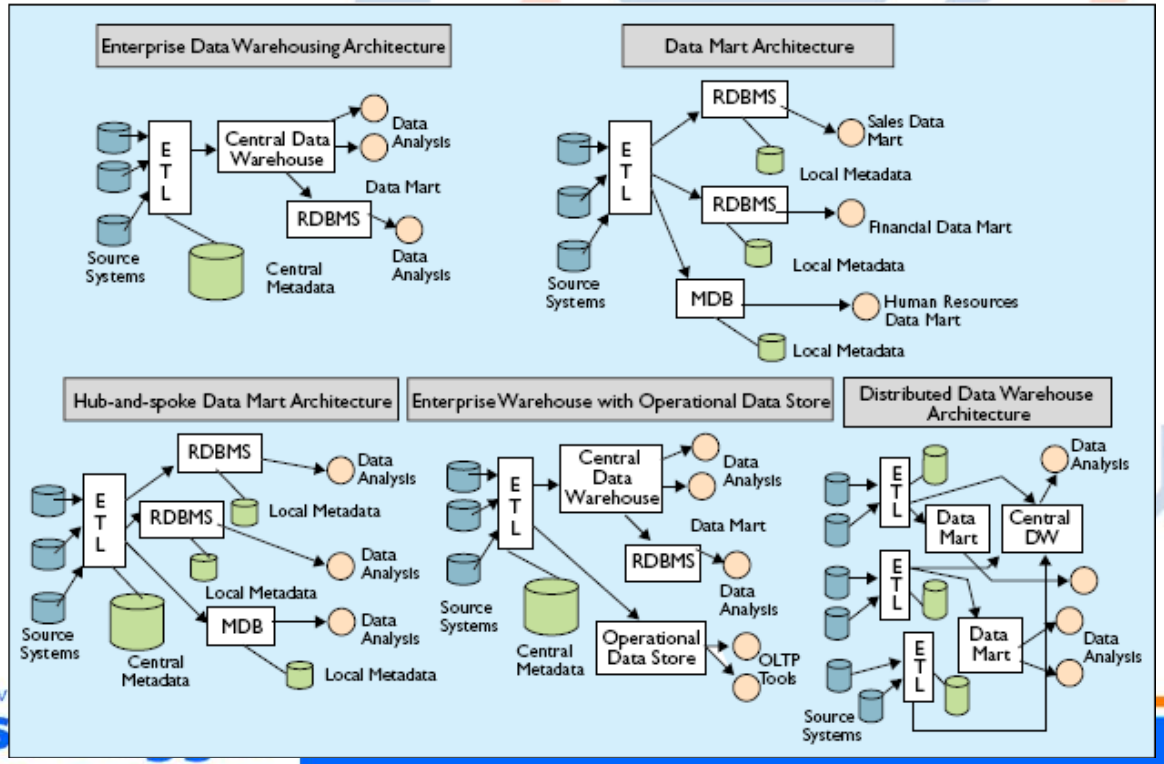
  With different DBMS, operating system, hardware, communication protocols

- ETL challenges
  - ✓ Getting the data from the source to target as fast as possible
  - ✓ Allow recovery from failure without restarting the whole process
- This leads to balance between **writing data** to staging tables or **keeping it in memory**

Physical Design

There are two major methods for architecture design of a datawarehouse (Kimball et al, 2008)

- Top down → centralized DW
- Bottom up → union of individual data mart



DW architecture (Sen & Sinha, 2005)

DW Dev Stages

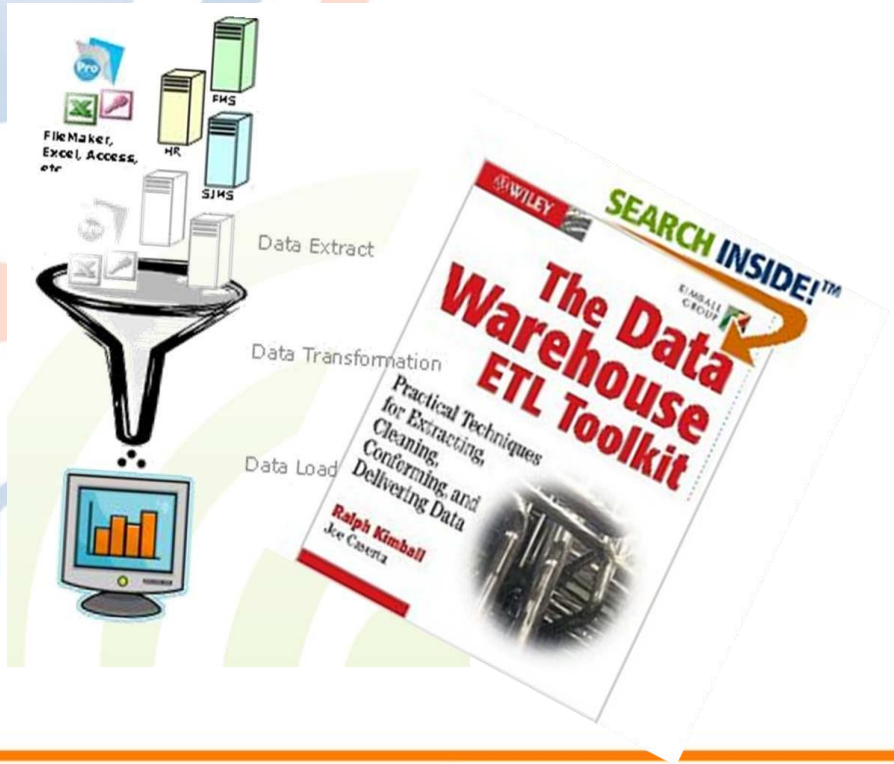
Task	Tools/ Techniques
* Requirements analysis	- Interview
- Data driven	- Subject area
- User driven	- Prototype
- Goal driven	- Scenario
- Process driven	- JAD
- Externally driven	
* Conceptual design	- E/R
- Multidimensional modelling	- Object oriented
	- Ad hoc model
* Logical design	
- Relational implementation	- M E/R
* Star, constellation, snowflake schemata	- Fact schema
- Multidimensional implementation	- UML class diagram
* Cubes, dwarfs and QC-Trees	- DFM
- Data storage	
* ROLAP, MOLAP, HOLAP	

* ETL	
- Extraction	- Immediate
	• Transaction log • Database trigger • Source file
	- Deferred
	• Time stamp based
	• File comparison
- Transformation	- Manual
	- Automated
	- Manual/Automated
- Loading	- Incremental
	- Full refresh
* Physical design	
- DW architecture design	
* Bottom up	- Independent data mart
* Top Down	- Centralized DW with dependent data mart
	- Centralized DW without dependent data mart
	- Virtual DW
	- Enterprise DW with operational data store

Operating a Data Warehouse

Operating a DW

- When **operating** a DW the following phases can be identified
  - Monitoring
  - Extraction
  - Transforming
  - Loading
  - Analyzing



Monitoring

- **Monitoring**
  - Surveillance of the data sources
  - Identification of data modification which is relevant to the DW

- Monitoring has an important role over the whole process deciding on which data to load, and when to load it into the DW

Monitoring…

- **Monitoring techniques**
  - **Active** mechanisms - Event Condition Action (ECA) rules:

EVENT	Payment
CONDITION	Account sum > 10 000 €
ACTION	Transfer to economy account

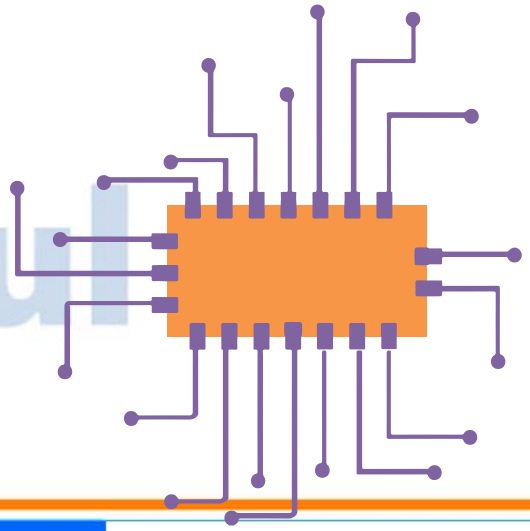
  - **Replication** mechanisms:
    - Oracle 9i – Snapshots are local copies of data (similar to a view): a snapshot is replaced completely on change
    - IBM DB2 – Data replication maintains and replicates data in destination tables through a data propagation processes (data is incrementally updated)

Monitoring…

- **Protocol** based mechanisms:
  - Since DBMSs write protocol data for transaction management, the protocol can also be used for monitoring
  - Problematic since protocol formats are proprietary and subject to change
- **Application** managed mechanisms:
  - Hard to implement for legacy systems
  - Based on *time stamping* or *data comparison*

Extraction

- **Extraction**
  - Reads the data selected during the monitoring phase and inserts it in the intermediate data structures of the workplace (“staging area”)
  - Due to large data volume, compression can be used

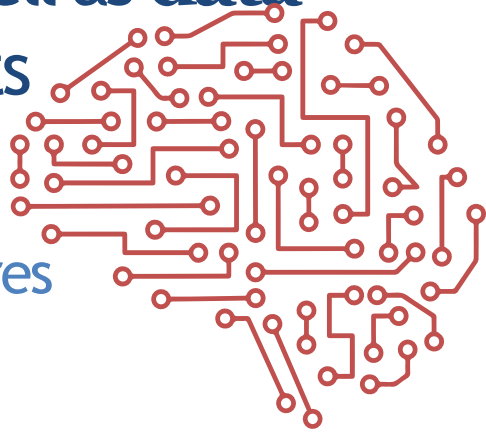


Extraction…

- The time-point for performing extraction can be
  - **Periodical**: weather or stock market information can be actualized more times in a day, while product specification can be actualized in a longer period of time
  - **On request:** e.g. when a new item is added to a product group
  - **Event driven:** event driven e.g. number of modifications over passing a specified threshold triggers the extraction
  - **Immediate:** in some special cases like the stock market it can be necessary that the changes propagate immediately to the warehouse
- Extraction largely depends on **hardware** and the **software** used for the D W and the data source

Transforming

- Transforming
  - Implies **adapting data, schema** as well as **data quality** to the application requirements
  - Data integration:
    - Transformation in de-normalized data structures
    - Handling of key attributes
    - Adaptation of different types of the same data
    - Conversion of encoding: "Buy","Sell" 1,2 vs. B,S 1,2
    - Date handling: "MM-DD-YYYY" "MM.DD.YYYY"



Transforming …

- String normalization
  - "Michael Schumacher" → "Michael, Schumacher" vs. "Schumacher Michael" → "Michael, Schumacher"
- Measurement units and scaling
  - 10 inch → 25,4 cm
  - 30 mph → 48,279 km/h
- Save calculated values
  - Price including tax = Price without tax * 1.19
- Aggregation
  - Daily sums can be added into weekly ones
  - Different levels of granularity can be used

Transforming …

- **Data cleansing (or data cleaning)**
  - Consistency check: Delivery Date < Order Date
  - Completeness: management of missing values as well as NULL values
  - Dictionary approaches for city or person names
  - Regular expressions for phone numbers or email addresses
  - Duplicate detection for redundancy elimination
  - Outlier detection as a warning system for possible errors

Loading

- Loading
  - Loading usually takes place during weekends or nights when the system is not under user stress
  - Split between initial load to initialize the DW and the periodical load to keep the DW updated
  - **Initial loading**
    - Implies big volumes of data and for this reason a **bulk loader** is used
  - Usually optimized by means of **parallelization** and **incremental actualization**

Analyzing

- Analysis phase
  - Data access - useful for extracting goal oriented information
    - How many iPhones 3G were sold in the Braunschweig stores of T-Mobile in the last 3 calendar weeks of 2010?
    - Although it's a common OLTP query, it might be to complex for the operational environment to handle
  - OLAP - the class of analytical operations running on the DW
    - In which district does a product group register the highest profit? And how did the profit change in comparison to the previous month?

Data Mining

- Data mining
  - Useful for identifying hidden patterns, e.g. customers buying wine also buy cheese
  - Useful for answering questions like: How does the typical iPad buyer look like? (for a targeted marketing campaign)
  - Methods and procedures for data mining: association rule mining, sequence pattern mining, classification, clustering, etc.

Thank You…