Lecture 03

Data Warehouse Life Cycle & Basic Architecture

Summary – last week

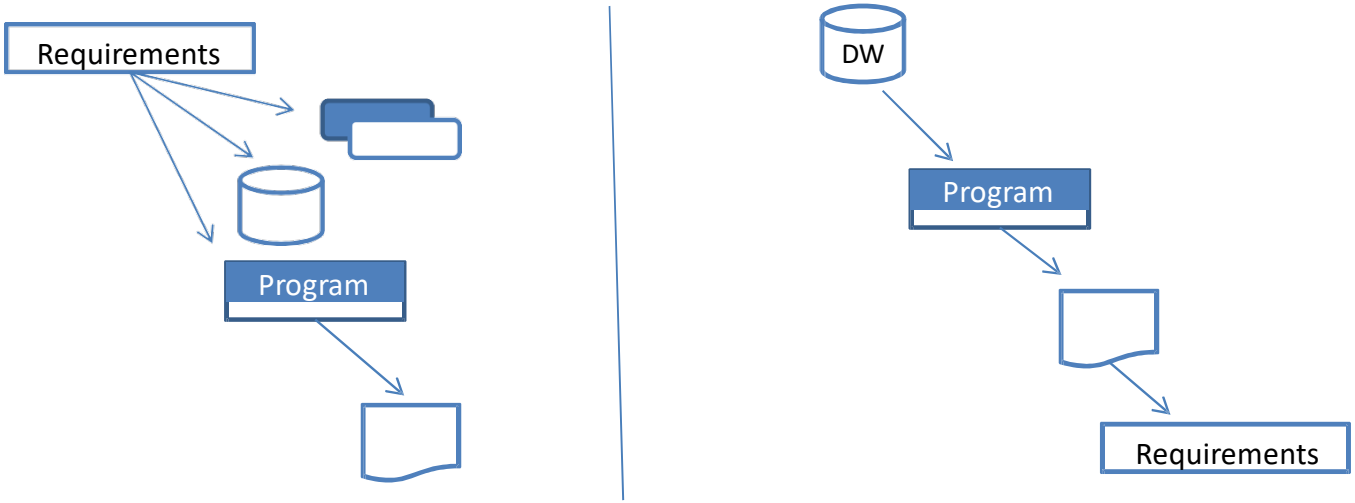
- Last week:
  - Introduction to DWH

- This week:
  - DW Lifecycle and Basic Architecture

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Life cycle of DW

- Classical SDLC vs. DW SDLC

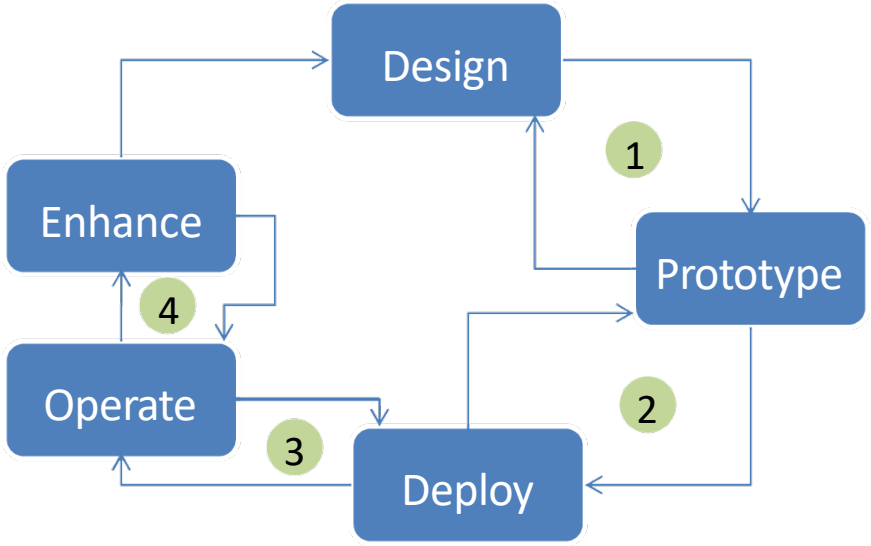


– DW SDLC is almost the opposite of classical SDLC

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Life cycle of DW (cont'd.)

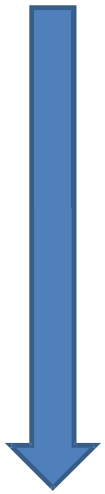
- Enhancement needs the modification of
  - HW - physical components
  - Operations and management processes
  - Logical schema designs



Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Life cycle of DW (cont'd.)

- 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 DSS system
Integration	Analyze results
Implementation	Understand requirements

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

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

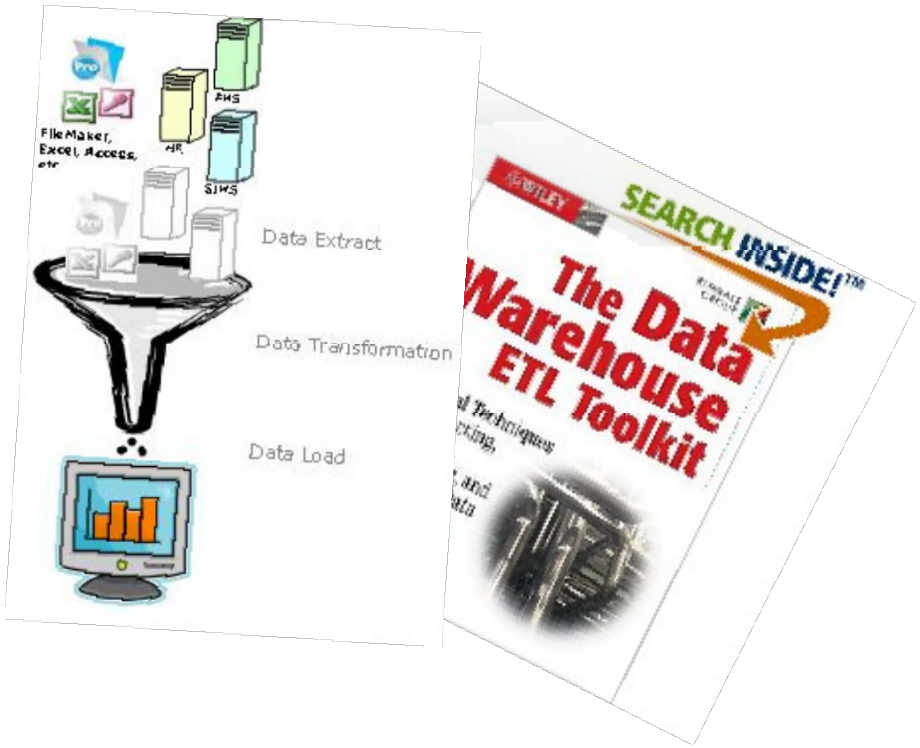
Life cycle of DW (cont'd.)

- 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"

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Operating a DW

- In Operating a DW the following phases can be identified
  - Monitoring
  - Extraction
  - Transforming
  - Loading
  - Analyzing



Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

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 the next steps will be applied on

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Monitoring (cont'd.)

- Monitoring techniques
  - Active mechanisms – Event Condition Action (ECA)

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

  - Replication mechanisms
    - Snapshot:
      - Local copy of data, similar to aView
      - Used by Oracle 9i
    - Data replication
      - Replicates and maintains data in destination tables through data propagation processes
      - Used by IBM

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Monitoring (cont'd.)

- Protocol based mechanisms
  - Since DBMS write protocol data for transaction management, the protocol can be used also for monitoring
  - Difficult due to the fact that the protocol format is proprietary and subject to change
- Application managed mechanisms
  - Hard to implement for legacy systems
  - Based on *time stamping* or *data comparison*

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Extraction

- Extraction
  - Reads the data which was selected throughout the monitoring phase and inserts it in the data structures of the workplace
  - Due to large data volume, compression can be used
  - 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

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Extraction (cont'd.)

- On request:
  - For example when a new item is added to a product group
- Event driven:
  - Event driven extraction can be helpful in scenarios where time, or the number of modifications over passing a specified threshold triggers the extraction. For example each night at 03:00 or each time 50 new modifications took place, an extraction is performed
- Immediate:
  - In some special cases like the stock market it can be necessary that the changes propagate immediately to the warehouse

– The extraction largely depends on hardware and the software used for the DW and the data source

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

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
    - Correcting of wrong information

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Transforming (cont'd.)

- Normalization:
  - "Michael Hill" → "Michael, Hill" vs.

    "Hill Michael" → "Michael, Hill"
- Date handling:
  - "MM-DD-YYYY" → "MM.DD.YYYY"
- Measurement units and scaling:
  - 10 inch → 25,4 cm
  - 30 mph → 48,279 km/h
- Save calculated values
  - Price_excl_GST = Price_excl_GST * 1.15
- Aggregation
  - Daily sums can be added into weekly ones
  - Different levels of granularity can be used

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Transforming (cont'd.)

- Data cleaning:
  - Consistency check
    - Delivery_date should be after Order_date
  - Completeness
    - Management of missing values as well as NULL values

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 performed by partitioning, parallelization and incremental actualization

Analyzing

- Analyze
  - Data access
    - Useful for extracting goal oriented information:
      - How many iPhones 7 were sold by Apple stores in Islamabad in the last 3 calendar weeks of 2017?
      - Although it is a common OLTP query, it might be to complex for the operational environment to handle
  - OLAP
    - Used to analyze data contained in DW
    - Used to answer requests like:
      - In which district does a product group register the highest profit
      - How did the profit change in comparison to the previous month?

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Analyzing (cont'd.)

- Mostly known as organized on a multidimensional data model
- Common operations for analyze are:
  - » Pivoting/Rotation
  - » Roll-up, Drill-down and Drill-across
  - » Slice and Dice

– Data mining

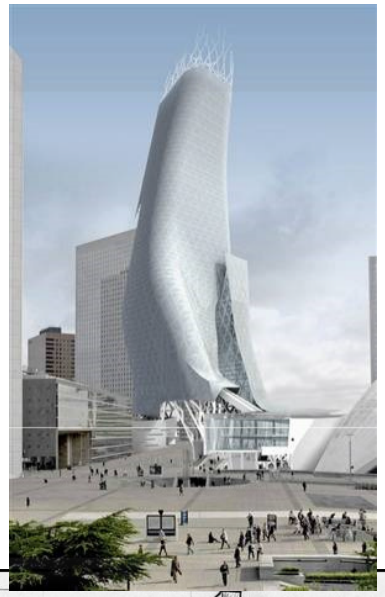
- Useful for identifying hidden patterns
- Refers to two separate processes:
  - KDD (Knowledge Discovery in Databases)
  - Prediction
- Useful for answering questions like:
  - How did the sales of this product group evolve?
- Methods and procedures for *data mining*
  - Clustering, Classification, Regression, Association rule learning

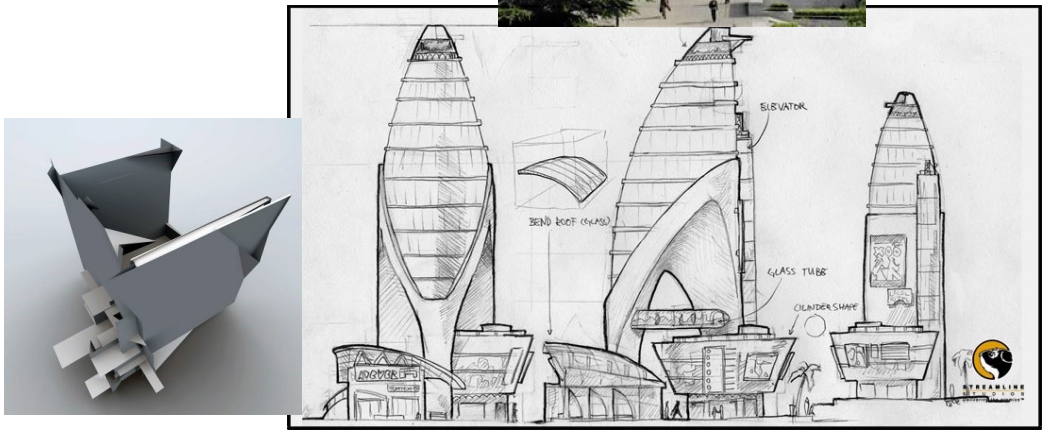
Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Architecture

Architecture

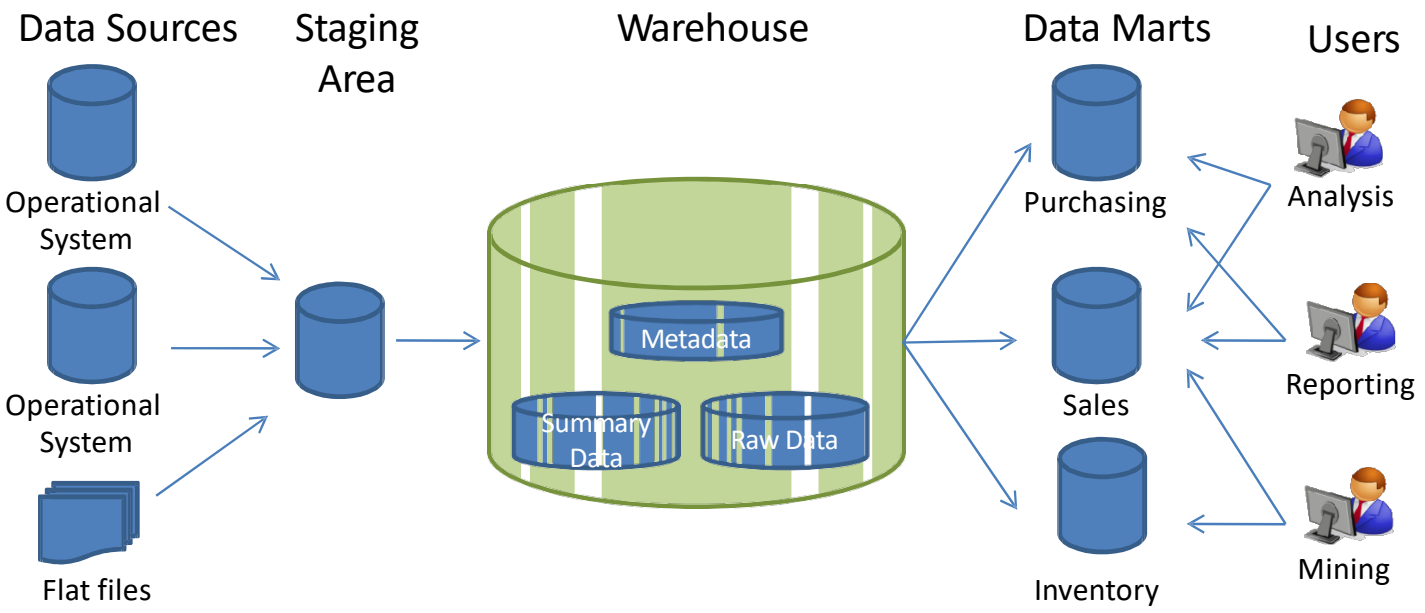
1. Basic Architecture
2. Storage Structures
3. Tier Architectures
4. Distributed DW
5. DW Data Modeling





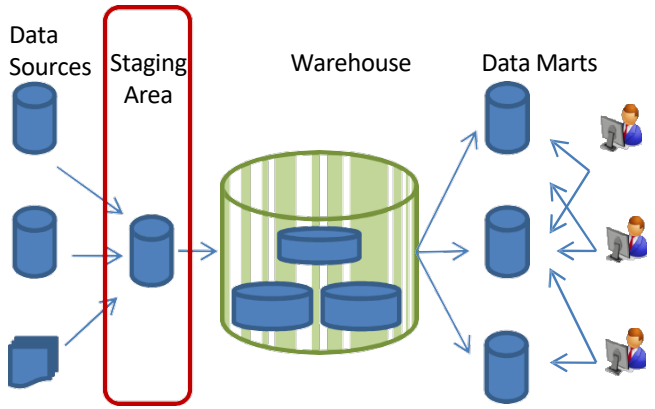
Basic Architecture

- Architecture of a DW



Basic Architecture (cont'd.)

- The Data Staging Area
  - Is both a storage and process area (the ETL process)
  - It represents everything that happens between the operational source system and the data presentation area
  - The key architectural requirement for data staging area is that it is off-limits to business users and does not provide query and presentation services



Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

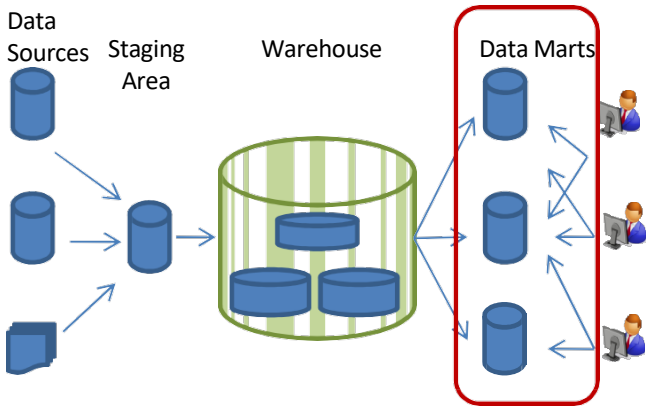
Basic Architecture (cont'd.)

- Customers aren't invited to visit the kitchen…
  - Similar to a restaurant's kitchen, the data staging area should be accessible only to skilled professionals

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Basic Architecture (cont'd.)

- The Data Presentation Area
  - Is where data is organized, stored and made available for queries, report writers, and other analytical processing
  - This area is the Warehouse as far as the business community is concerned



Summary

- DW life cycle:
  - DW life cycle
  - Operating a DW
    - Monitoring
    - Extraction
    - Transforming
    - Loading
    - Analyzing
  - Basic Architecture

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Next lecture

- DW Storage Architecture

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides