Data Warehouse
Development Life Cycle

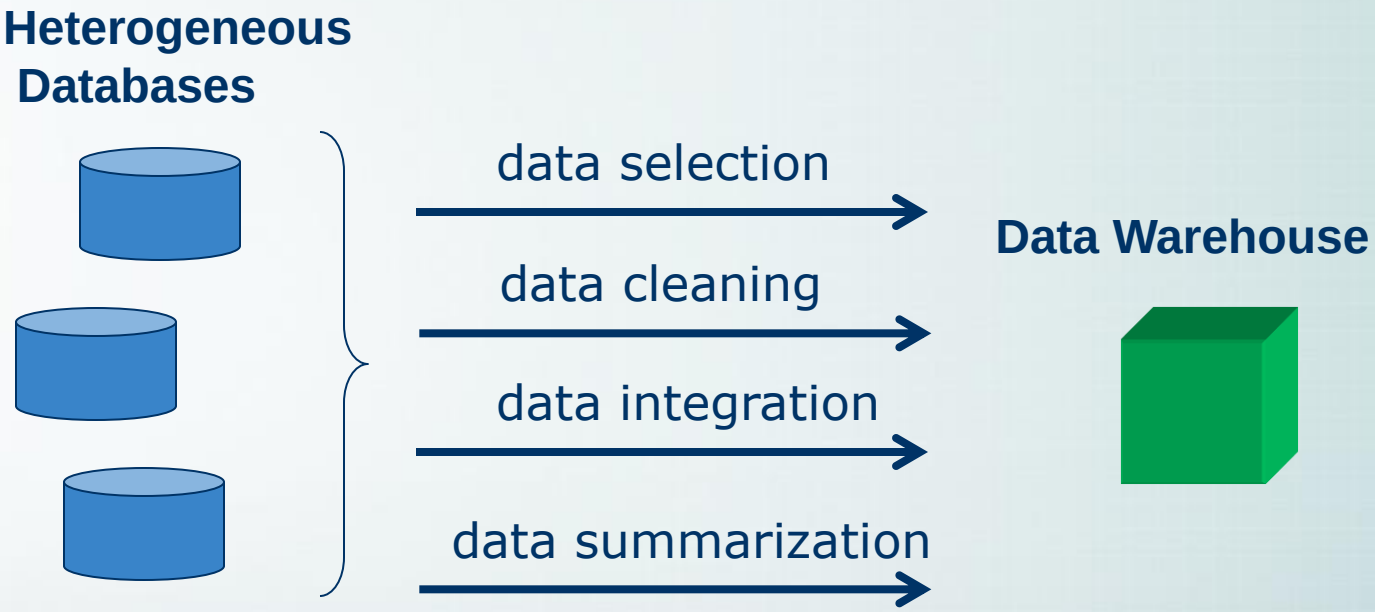
Munawar, Ir. MMSI. M.Com, PhD

Agenda

1. What is Data Warehouse
2. Data Warehouse Architecture
3. DW Development Life Cycle

Why Data Warehousing?

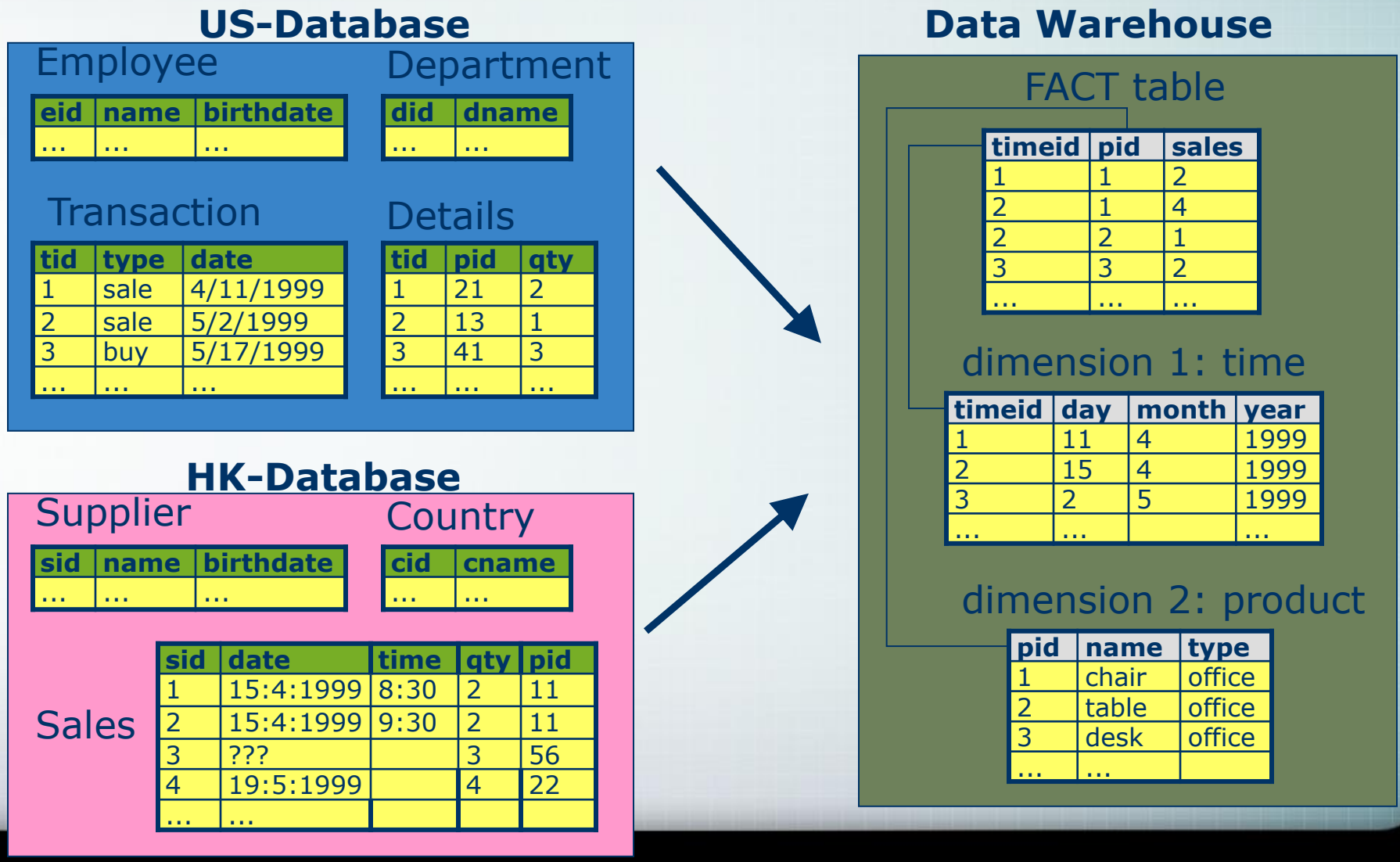
- Data warehousing can be considered as an important preprocessing step for data mining



- A data warehouse also provides on-line analytical processing (OLAP) tools for interactive multidimensional data analysis.

WHAT IS DATA WAREHOUSE

Example of a Data Warehouse (1)



Example of a Data Warehouse (2)

- **Data Selection**
  - Only data which are important for analysis are selected (e.g., information about employees, departments, etc. are not stored in the warehouse)
  - Therefore the data warehouse is subject-oriented
- **Data Integration**
  - Consistency of attribute names
  - Consistency of attribute data types. (e.g., dates are converted to a consistent format)
  - Consistency of values (e.g., product-ids are converted to correspond to the same products from both sources)
  - Integration of data (e.g, data from both sources are integrated into the warehouse)

Example of a Data Warehouse (3)

Data Cleaning

- Tuples which are incomplete or logically inconsistent are cleaned

Data Summarization

- Values are summarized according to the desired level of analysis
- For example, HK database records the daytime a sales transaction takes place, but the most detailed time unit we are interested for analysis is the day.

Example of a Data Warehouse (4)

- **Example of an OLAP query (collects counts)**
  - Summarize all company sales according to product and year, and further aggregate on each of these dimensions.

year

product	1999	2000	2001	2002	ALL
chairs	25	37	89	21	172
tables	10	30	0	45	85
desks	56	84	9	35	184
shelves	19	20	0	71	110
boards	5	16	11	15	47
ALL	115	187	109	187	598

Data cube

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
    - E.g., Hotel price: currency, tax, breakfast covered, etc.
  - When data is moved to the warehouse, it is converted.

Data Warehouse—Time Variant

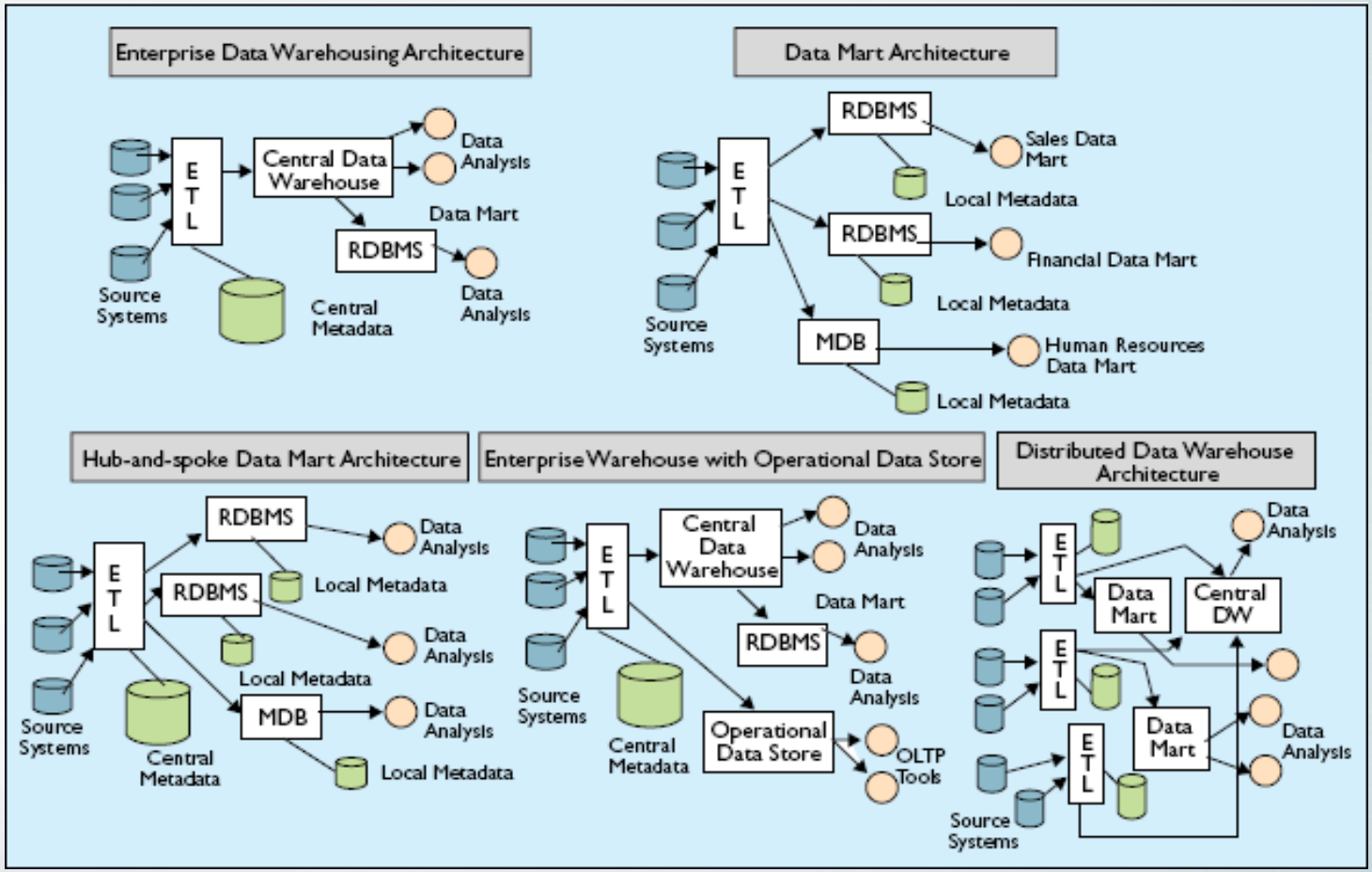
- **The time horizon for the data warehouse is significantly longer than that of operational systems.**
  - Operational database: current value data.
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- **Every key structure in the data warehouse**
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain “time element” (the time elements could be extracted from log files of transactions)

Data Warehouse—Non-Volatile

- 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 ARCHITECTURE

Data Warehouse Architecture



Source: Sen and Sinha, 2005

DW Architecture (cont'd...)

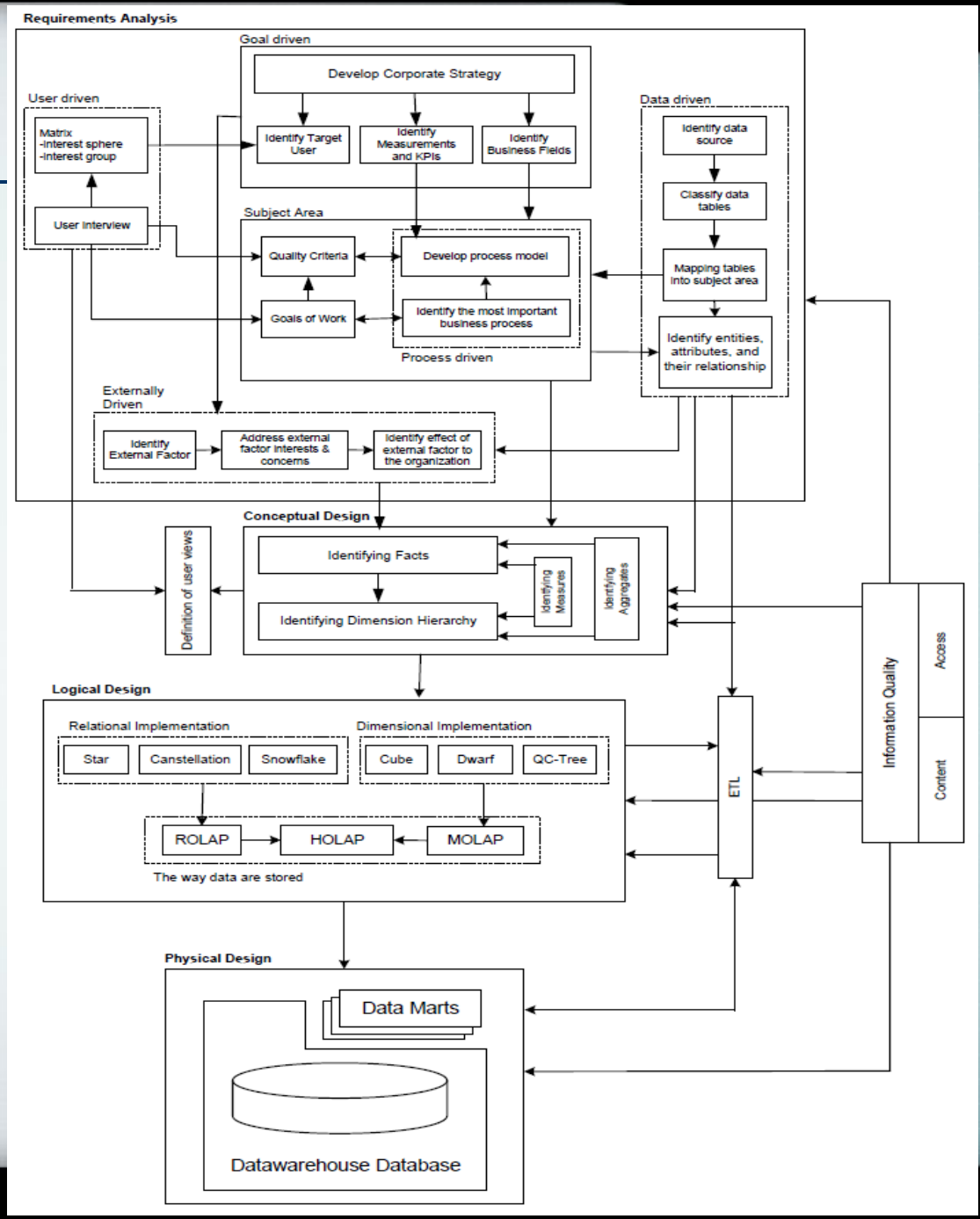
- **Top Down Design** → Centralized DW
  - supplies the needs of multiple departments
  - feedback from departments or user groups can be used to customise requirements
  - The failure rate of a centralised DW is much higher than that of a DM
- **Bottom Up Design** → Constuction of Data Mart
  - a DW can be regarded as the integration of different DMs
  - DM is intended for examining a single subject area of corporate operations
  - less risky than a top–down approach
  - it may create redundancies and is difficult to integrate because DMs tend to provide a narrow perspective of corporate data

DM Integration (Chhabra and Pahwa, 2014)

- **Integration with dimensions sharing**
  - Multiple DMs can be joined over common dimensions (Kimball, 2002).
- **Integration with dimensions compatibility**
  - Two dimensions in different DM can be said compatible when their common information is consistent or their contents can be combined in a meaningful way (Chhabra and Pahwa, 2014; Cabibbo and Torlone, 2004)
- **Integration with generalization**
  - It still possible to integrate the DM via drill-across query even though the dimensions are not same (Abello, Samos, and Saltor, 2002).
  - Dimensions in the different DM can be connected by generalization

DATA WAREHOUSE DEVELOPMENT LIFE CYCLE

DW Development



Requirements Analysis

- Successful DW development is based on sound requirements analysis.
- The construction of a good-quality DW can be stimulated by excellent requirements analysis.
- Unfortunately, this type of examination is often unclear and characterised by uncertainty because of changes to requirements that can occur even in the short term; such changes are typically essential for organisations to keep pace with the rapid evolution of business conditions.
- Consequently, poorly performed requirements analysis is a major failure of many software projects

Conceptual Design

- Conceptual design is one of the most essential phases in the overall DW development process.
- A good-quality multidimensional model can be produced when the steps in designing conceptual schema are correctly followed.
- The establishment of a multidimensional model is influenced primarily by multi-driven requirements analysis and the existence and structure of data in operational systems

Logical Design

- Logical design is the most attractive step given that it presents tremendous benefits to system performance.
- It is intended to obtain conceptual schemata based on the data structure that will be applied by a DM or DW, with consideration for a number of constraints, particularly those related to disk space or query retrieval

Physical Design

- A DW environment can be represented in the form of a DW or DM
- Two major techniques for designing DWs are top down approach (focuses on the construction of a centralised DW) and bottom up (Constructing individual DMs)
- For most corporations, a top–down approach is a significantly favourable strategy in terms of cost and duration of implementation

ETL

- ETL is a process of integrating data from many sources (usually heterogeneous) into a DW database
- Considering the huge amount of data that should be processed and the number of data sources that should be integrated, designing ETL processes is very complicated and considerably time consuming

Thank You !

Munawar, PhD