



# Overview



COMP323 Chapter 1

# About Me

---

- ▶ Dr Patrick Pang 彭祥佑博士
- ▶ Lecturer in the School of Applied Sciences
  - ▶ PhD University of Melbourne, Australia
- ▶ [patrickpang@ipm.edu.mo](mailto:patrickpang@ipm.edu.mo)
- ▶ [www.patrickpang.net](http://www.patrickpang.net)
  
- ▶ Office Hours:
  - ▶ Location: N46B
  - ▶ Thursdays 9:00-11:30, Fridays 9:30-13:00

# Learning Outcomes

---


- ▶ After completing the learning module, students will be able to:
  1. Contrast data warehouses with operational databases in the aspects of design and utilization
  2. Design a conceptual and logical model for a data mart that satisfies user requirements
  3. Analyze multidimensional data using OLAP technology
  4. Apply suitable data mining algorithms to discover patterns in data warehouses



# Key Topics

---

- ▶ Generic databases vs. data warehouses
- ▶ Architectures of data warehouses
- ▶ Conceptual modeling (star schema, snowflake schema)
- ▶ OLAP queries
- ▶ Metadata and Extract, Transform, Load (ETL)
- ▶ Data mining (e.g. decision trees)



**Attendance will  
be taken in  
every session**

# Assessments

---

## ► Assignments (40%)

- Assignment 1 (20%; Due Week 6\*) – Conceptual modeling of data mart
- Assignment 2 (20%; Due Week 12\*) – SQL queries for OLAP
- Assignments will be released and submitted through Canvas

## ► Test (20%; Week 8\*)

## ► Final Exam (40%)



*\* Dates are tentative and may be adjusted based on the teaching progress*

# How to survive this module?

---



All materials of this unit can be downloaded online from Canvas. Check regularly for updates.



You can contact your teacher via email. I aim to respond in 24-48 hours (exclude weekends and holidays).



Prepare before you come to the class. Read slides and other provided material before class.



Attend your classes on time. Submit your assessments on time.



Ask questions early. Don't wait until the last minute otherwise nobody will be able to help!



No plagiarism and academic misconduct! They are taken very seriously.

# Chapter 1 Outline

---

## ▶ A. Business Intelligence

- ▶ Operational databases vs. data warehouses
- ▶ Defining features of data warehouses

## ▶ B. Architecture

- ▶ Two layers. "why don't run analysis on operational data?"
- ▶ Data warehouse vs. data mart
- ▶ Hub-and-spoke vs. Bus
- ▶ Reconciled data layer

## ▶ C. Design Methodology

- ▶ ETL process
- ▶ Seven phases in data mart design
- ▶ Data mart design in future chapters

# Part A. Business Intelligence

---

- ▶ Operational systems and databases
- ▶ Decision-support systems
  - ▶ Different data requirements
- ▶ Business Intelligence
  - ▶ Data warehousing vs. Data analysis
  - ▶ Features of Data warehouse
  - ▶ Operational databases vs. Data Warehouses



# Operational Systems and Databases

- ▶ Operational systems "make the wheel of business turn"
  - ▶ E.g. Point-of-sales, inventory control, transportation
  - ▶ Also known as OLTP (Online Transactional Processing)
- ▶ .. supported by numerous operational databases.
  - ▶ Read/Write access. Update frequently
  - ▶ Integrity is important
  - ▶ Avoid data redundancy
  - ▶ Normalized data model
  - ▶ Pre-determined transactions in application programs.
  - ▶ Keep current data only
  - ▶ ...

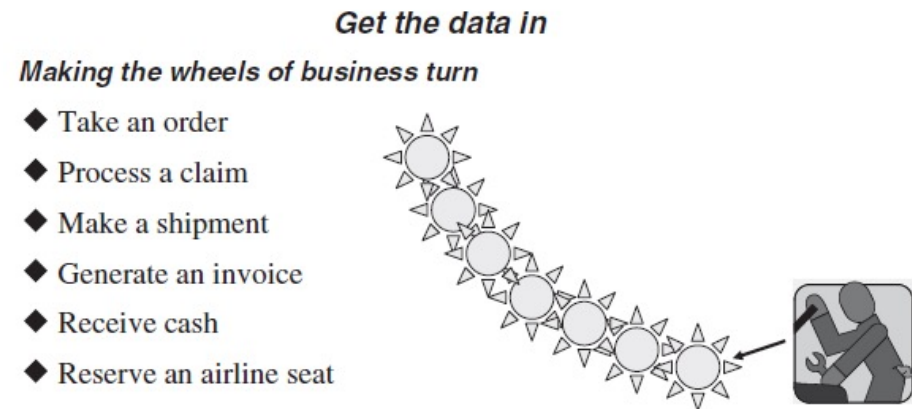
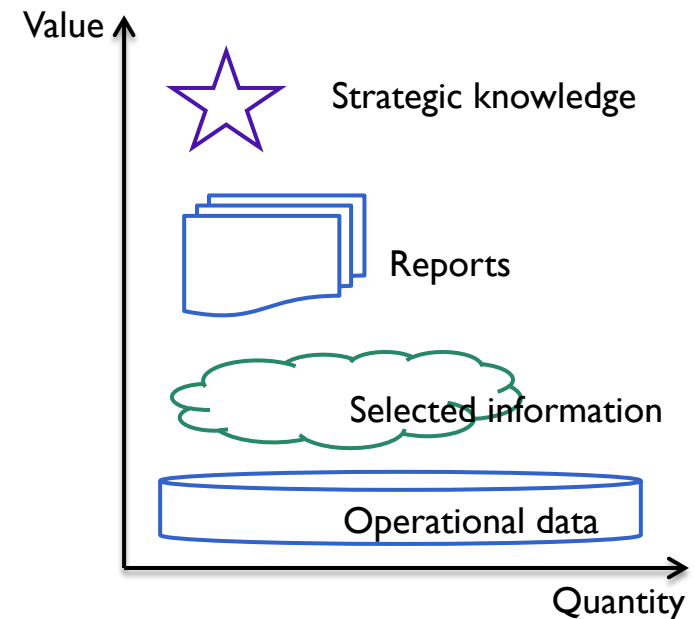


Figure 1-5 Operational systems.

# Data $\neq$ Information

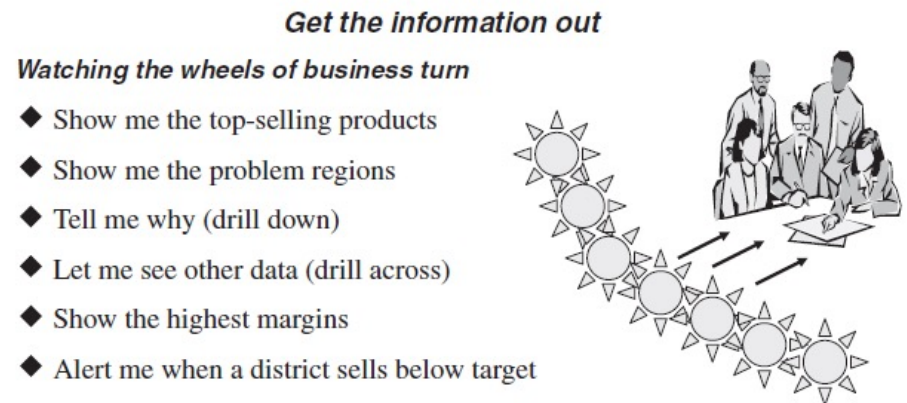
- ▶ Operational systems generate a huge amount of detail data
- ▶ The managers need strategic information to make proper decisions. Usually, huge amount of data is summarized with tailor-made queries and presented in reports for managers.
- ▶ Availability of too much data makes the extraction of the most important information difficult



# Decision-Support Systems

---

- ▶ Decision-support systems helps managers to make proper decisions when they establish objectives and monitor results
  - ▶ "Increase sales by 15% in the North East division over the next 5 years"
  - ▶ "Improve product quality levels in the top five product groups"
- ▶ Decision-support systems "watch the wheels of business turn"
  - ▶ A common function is OLAP (Online Analytic Processing)



**Figure 1-6** Decision-support systems.

# Example Queries to Decision-Support Systems

---

- ▶ Q1: Find the 5 top-selling products for each product subcategory that contributes more than 20% of the sales within its product category.
- ▶ Q2: As of Dec 15, 2020, determine shipping priority and potential gross revenue of the orders that have the 10 largest gross revenues among the orders that had not yet been shipped. Consider orders from the book market segment only.
- ▶ Regular database models and systems are not suitable for this type of queries.

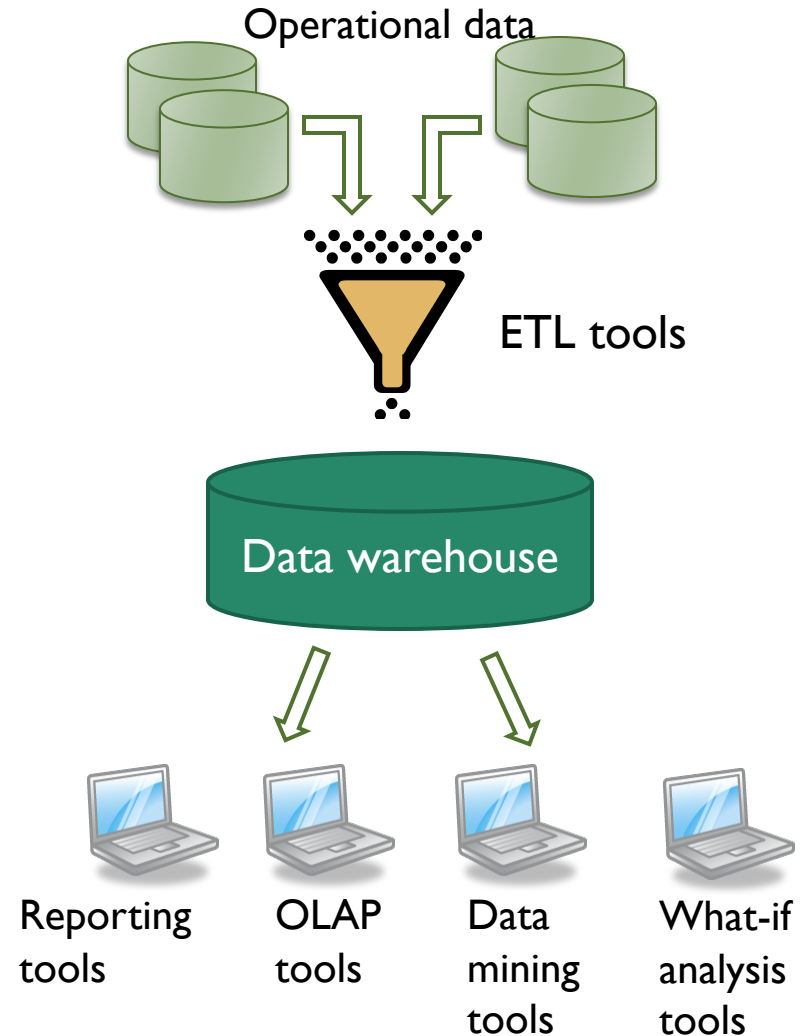
# Key problems of running analysis on operational databases

---

- ▶ **Complex and unusable models**
  - ▶ Many DB models are difficult to understand
  - ▶ DB models do not focus on a single clear business purpose
- ▶ **Same data found in many different systems**
  - ▶ The same concept is defined differently
  - ▶ Do not support analysis across business functions
- ▶ **Data quality is bad**
  - ▶ Missing data, imprecise data, different use of systems
- ▶ **Data are "volatile"**
  - ▶ Data deleted in operational systems (e.g. 6 months)
  - ▶ Data change over time – no historical information

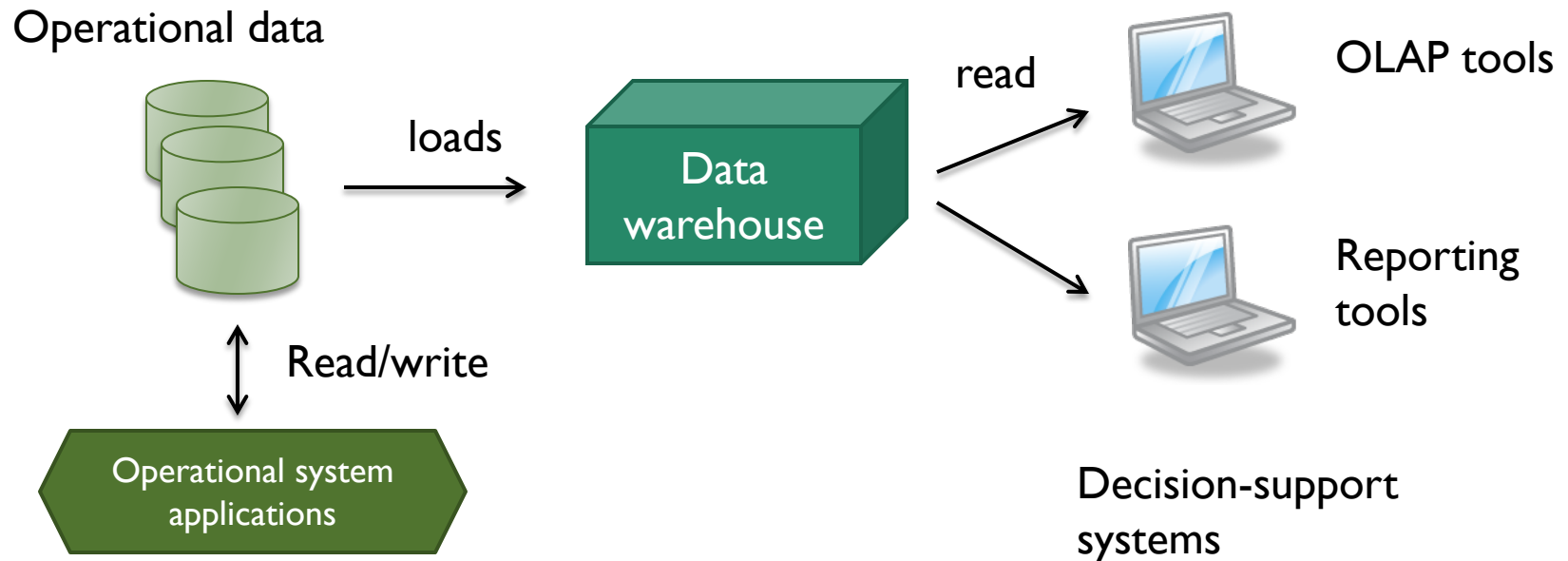
# Data Warehouse

- ▶ Data is integrated in advance, and stored in Data Warehouse for query and analysis
  - ▶ Barry Devlin, IBM Consultant: "A data warehouse is simply a single, complete, and consistent store of data obtained from a variety of sources and made available to end users in a way they can understand and use it in a business context."
- ▶ The data warehouse **separates** operational data from analysis applications.



# Data Warehouse

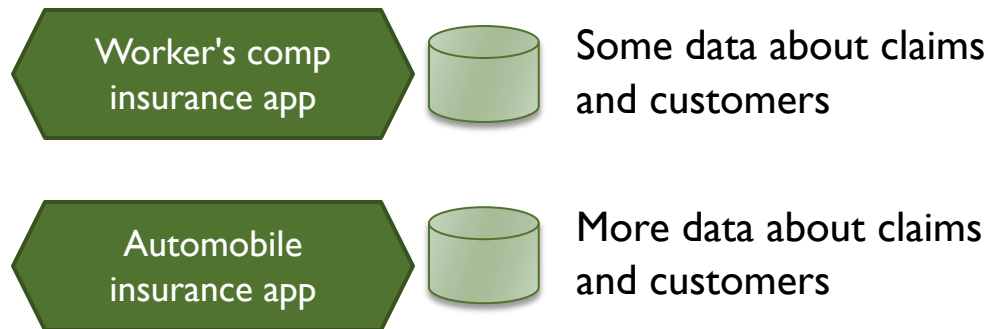
- ▶ A **data warehouse** is a **subject-oriented, integrated, nonvolatile, and time variant** collection of data in support of management's decisions. (by Bill Inmon, "Father of data warehouse")



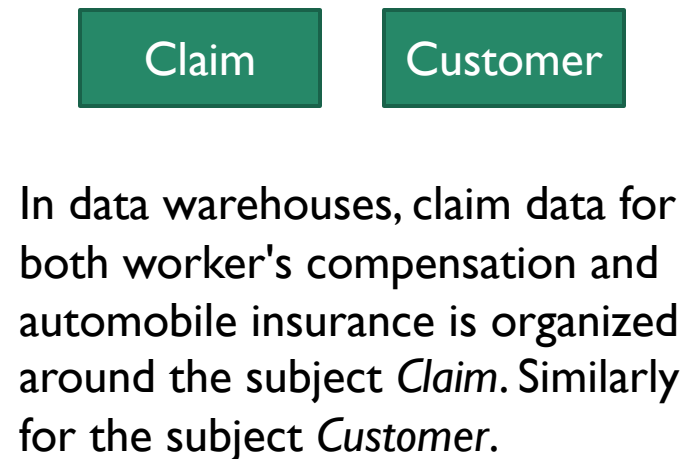
# Subject-Oriented Data

---

- ▶ Operational data sets are organized around individual applications
- ▶ In a data warehouse, all data sets about the same real-world business subjects or event are tied together
  - ▶ E.g. sales, inventory, shipment



In operational systems, each application stores data relating to its functions.

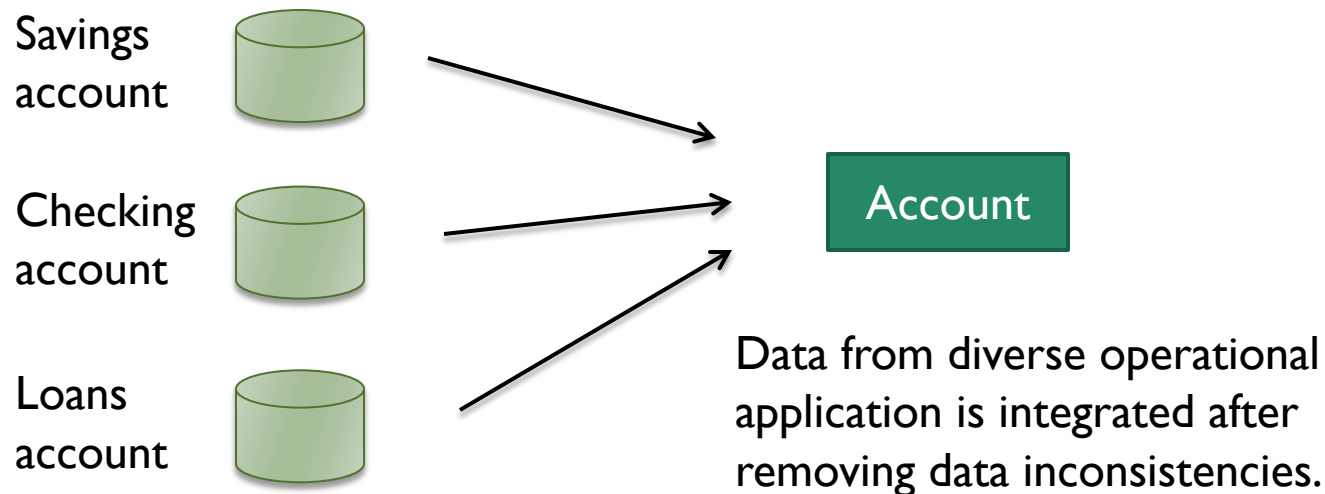




# Integrated Data

---

- ▶ Data about a subject in a data warehouse is pulled together from various applications
- ▶ These data are reconciled to remove inconsistency



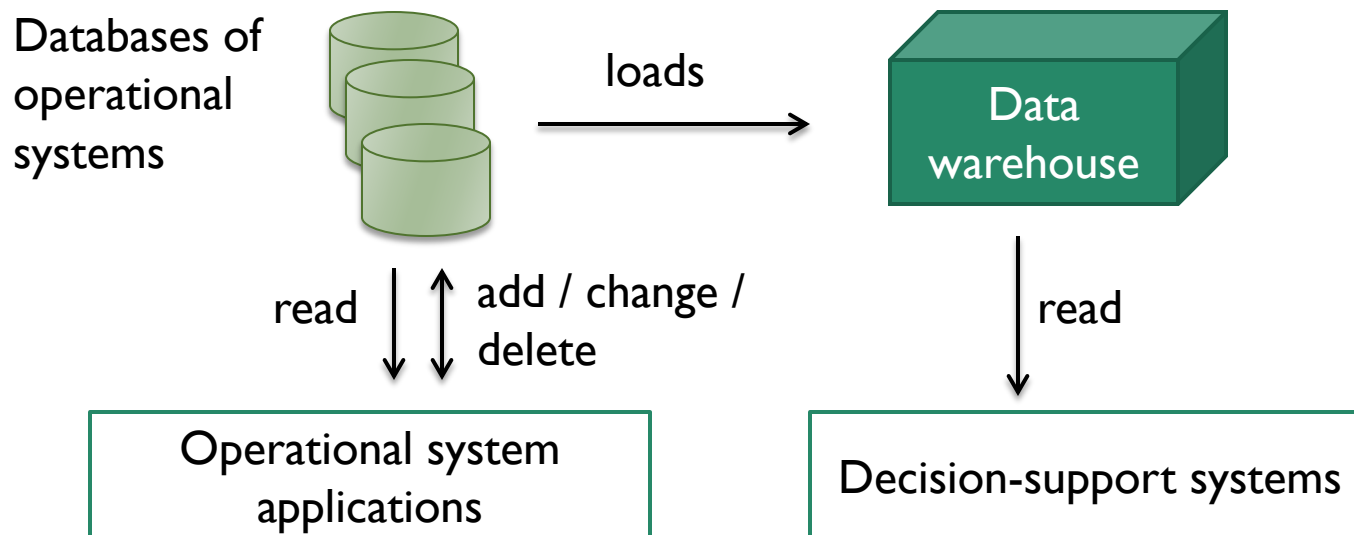
# Time-Variant Data

---

- ▶ Stored data in operational systems contains *current* values.
- ▶ A data warehouse has to contain *historical* data for analysis and decision making:
  - ▶ Snapshots over past and current periods (e.g. inventory level at different time)
  - ▶ Changes to data are tracked and recorded (e.g. both current and past purchases of customers)
- ▶ Every data structure in the data warehouse contains the time element.
  - ▶ E.g. sales quantity in a day, week, month, quarter, or year.

# Nonvolatile Data

- ▶ After loading from operational systems to the data warehouse at specific intervals, the data are usually not updated or deleted.



# Granularity

---

- ▶ Granularity in DW refers to level of details.
  - ▶ The lower the level of detail, the finer is the data granularity
- ▶ Operational systems usually keep data at the lowest level of detail
  - ▶ E.g. In a point-of-sale system for a grocery store, the units of sale are captured and stored at the level of units of a product per transaction at the check-out counter.
- ▶ When a user queries DW for analysis, he/she usually starts by looking at summary data.
  - ▶ E.g. The user may start with total sale units of a product in an entire region. Then the user may want to look at the breakdown by cities in the region. The next step may be the examination of sale units by the next level of individual stores.
  - ▶ DW may keep data at both low and high levels of detail

# Example: granularity

---

## THREE DATA LEVELS IN A BANKING DATA WAREHOUSE

<u>Daily Detail</u>	<u>Monthly Summary</u>	<u>Quarterly Summary</u>
Account	Account	Account
Activity Date	Month	Quarter
Amount	Number of transactions	Number of transactions
Deposit/Withdrawal	Withdrawals	Withdrawals
	Deposits	Deposits
	Beginning Balance	Beginning Balance
	Ending Balance	Ending Balance

Data granularity refers to the level of detail. Depending on the requirements, multiple levels of detail may be present. Many data warehouses have at least dual levels of granularity.

**Figure 2-4** Data granularity.

# Comparison

---

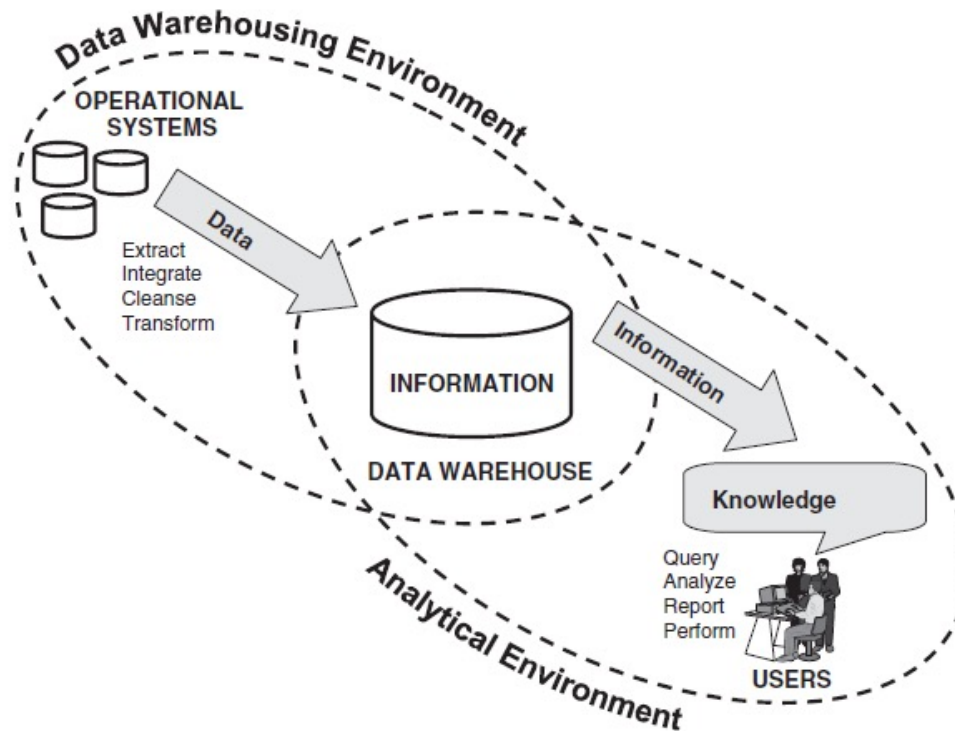
Feature	Operational Databases	Data Warehouses
Users	Thousands	Hundreds
Workload	Preset transactions	Specific analysis queries
Access	To hundreds of records, write and read mode	To millions of records, mainly read-only mode
Goal	Depends on applications	Decision-making support
Data integration	Application-based	Subject-based
Quality	In terms of integrity	In terms of consistency
Time coverage	Current data only	Current and historical data
Updates	Continuous	Periodical
Model	Normalized	Denormalized, multidimensional

# Business Intelligence (BI)

---

- ▶ Business Intelligence is a combination of technologies to improve business decision making by fact-based support systems
- ▶ The data warehouse (DW) stores the data. Analysis tools **use the DW**
  - ▶ Reports, which are defined by a query and a layout. (e.g. monthly receipts during the last quarter for every product category)
  - ▶ OLAP tools, which enables users to interactively analyze multidimensional data from multiple perspectives.
  - ▶ What-if analysis (sensitivity analysis)
  - ▶ Data mining, which discovers patterns and relationship in data (e.g. market basket analysis)
- ▶ A DW is a means rather than a goal...it is only a success if it is heavily used

# Business Intelligence (BI)



**Figure 1-10** BI: data warehousing and analytical environments.



# Review questions

---

- ▶ A typical retail store collects huge amount of data through its operational systems. Name three types of transaction data likely to be collected by a retail store in large volumes during its daily operation.
  - ▶ Examine opportunities that can be provided by strategic information
  - ▶ Other industries to consider: bank, medical center
- ▶ Describe five difference between operational data and strategic information.
- ▶ What are the defining features of data warehouses? Explain each feature and contrast it with operational databases.
- ▶ Illustrate the idea of granularity with an example of different data levels in a data warehouse for a retail store.

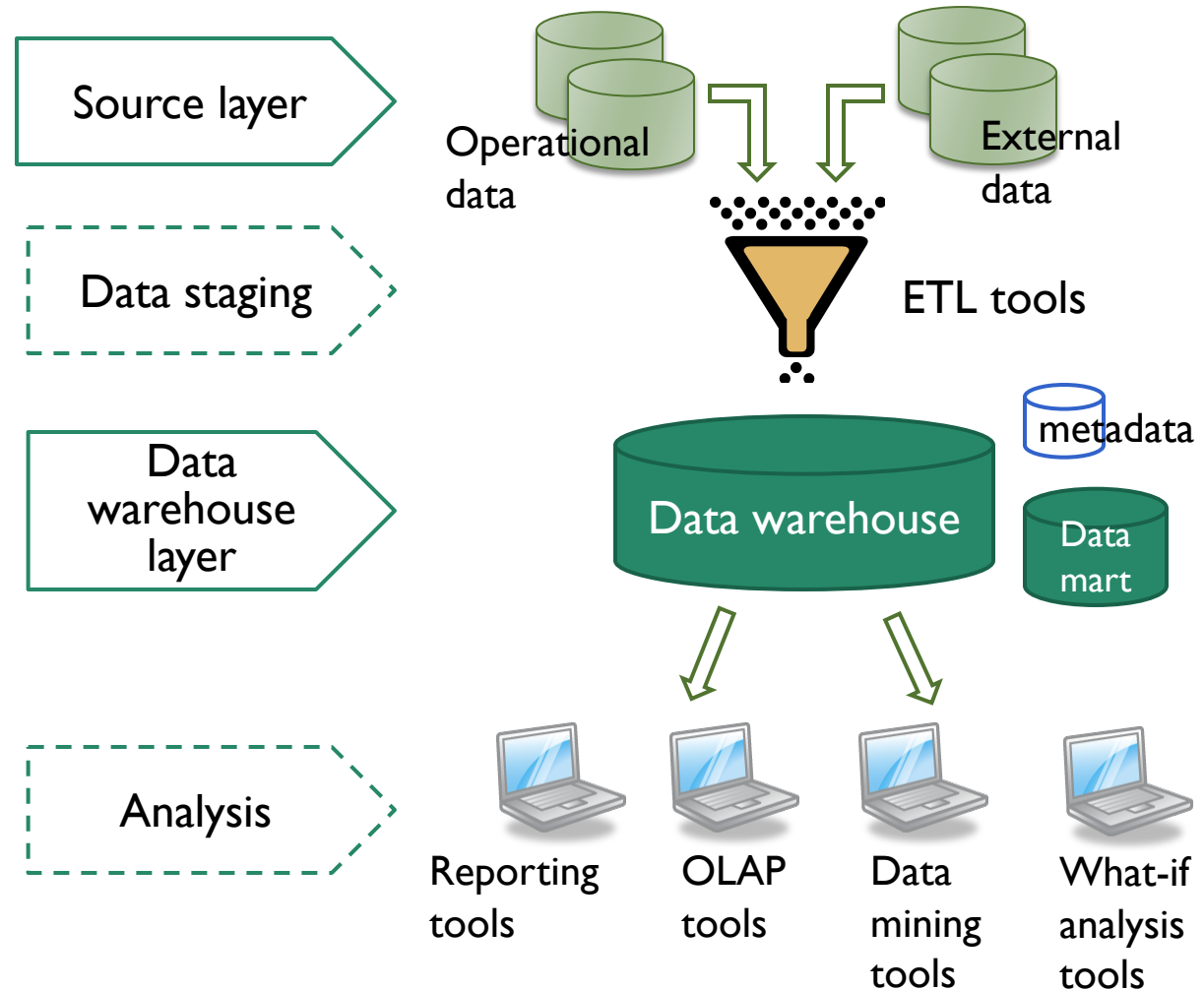
# Part B. Architecture

---

- ▶ Layer architecture
  - ▶ Data integration
- ▶ Inside the data warehouse layer
  - ▶ Data marts vs. Data warehouses
  - ▶ Metadata
- ▶ Data warehouse design approaches
  - ▶ Hub-and-spoke
  - ▶ Bus

# Two-Layer Architecture

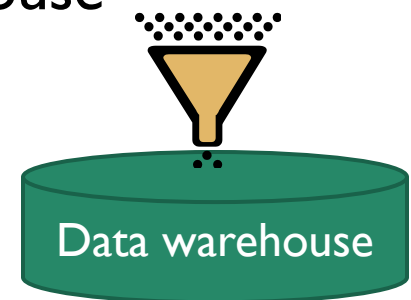
In the two-layer architecture, a data warehouse **separates** sources from analysis applications.



# Data Staging

---

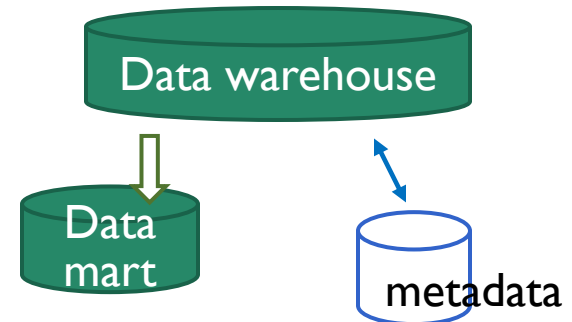
- ▶ A data warehouse system uses **heterogeneous** sources of data.
  - ▶ Data from operational systems include relational databases and files (e.g. Excel spreadsheets, XML files)
  - ▶ Data outside the corporate walls include the Web and other external data sources like reports and data sets
- ▶ Source data have **inconsistencies** and are in different data model and format!
- ▶ **ETL tools** (Extract, Transform, and Load) can merge heterogeneous schemata, extract, transform, cleanse, validate, filter and load source data into a data warehouse



# Inside the Data Warehouse Layer

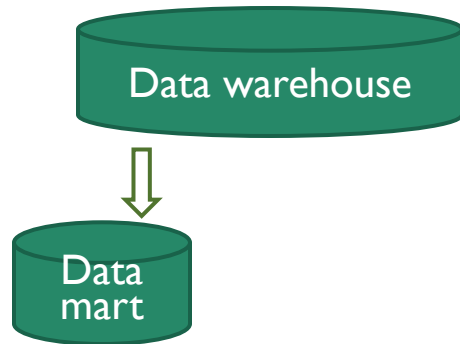
---

- ▶ Information is stored at one logically centralized single repository called **data warehouse**.
- ▶ The data warehouse for an enterprise is usually complex and huge
  - ▶ Too complex to build in one step
  - ▶ Not all information are necessary for a certain analysis task
- ▶ A **data mart** is a subset or an aggregation of the data stored to a primary data warehouse. It includes a set of information pieces relevant to a specific business area, corporate department, or category of users.
- ▶ A **metadata** repository keeps metadata. These metadata specifies data sources, transformation processes, population policies, logical and physical schemata, and user profiles.



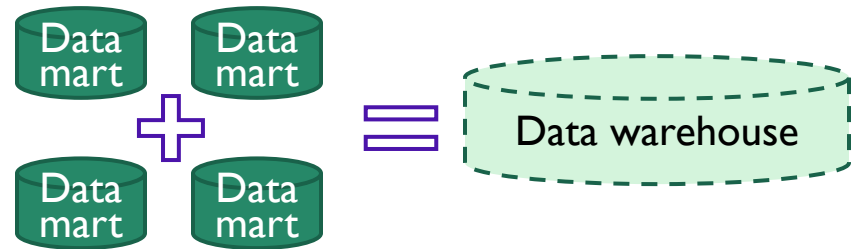
# Approaches to Structure DW

---



## Hub-and-spoke

Extract required data into a data mart for running analysis

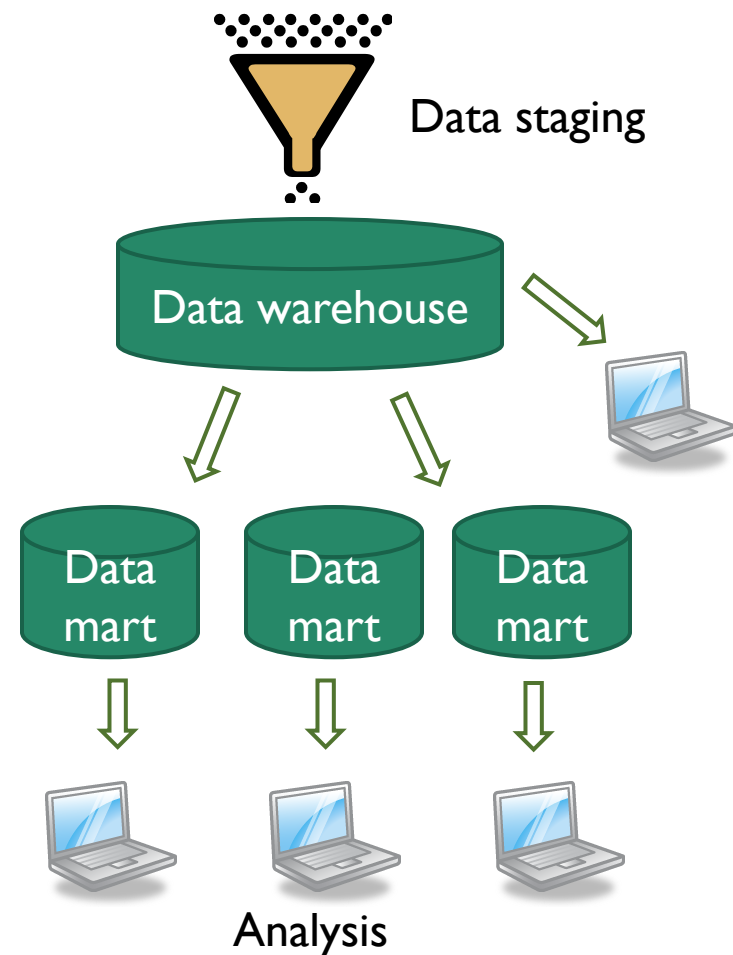


## Bus

Data warehouse as a collection of data marts

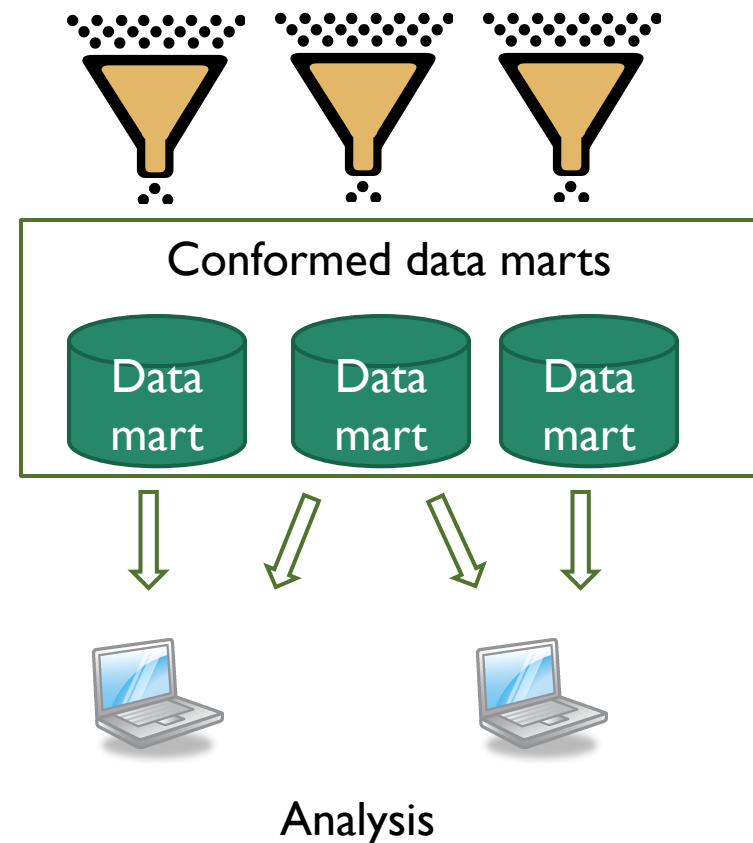
# Hub-and-Spoke Architecture

- ▶ Top-down approach. Recommended by **Bill Inmon**.
- ▶ A data warehouse as a centralized repository for the entire enterprise
  - ▶ stores data at the lowest level of detail. Usually relational model.
- ▶ Data marts are created on demand for departmental analytical needs
  - ▶ May filter and aggregate DW data
  - ▶ Better performance than running query directly on DW
- ▶ May also run analysis directly on DW



# Bus Architecture

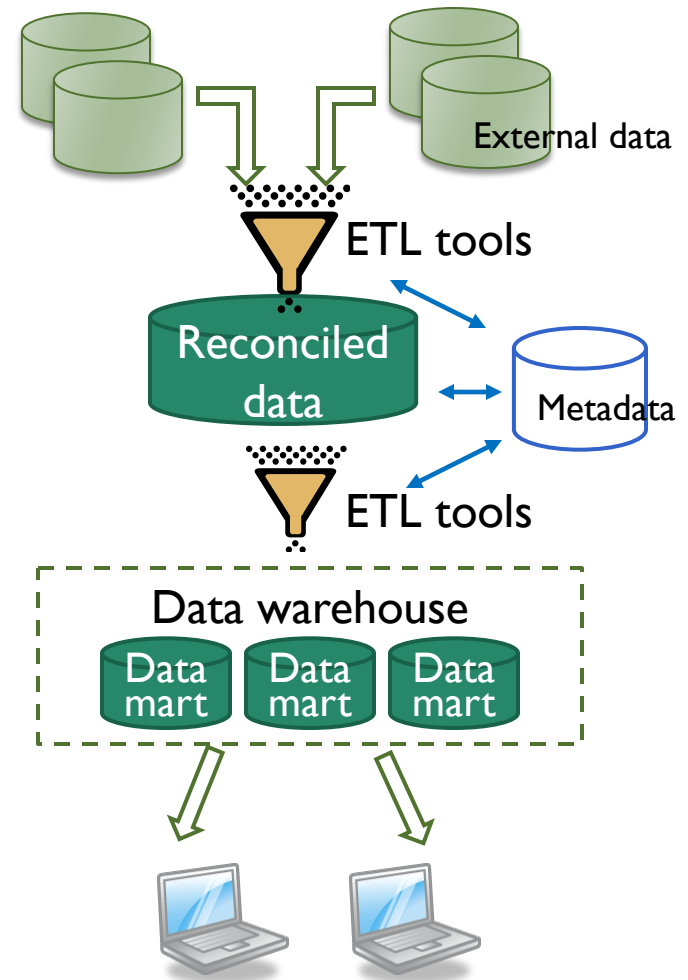
- ▶ Bottom-up approach.  
Recommended by **Ralph Kimball**.
- ▶ A data warehouse as a collection of data marts with *conformed analysis dimensions*.
- ▶ Data marts are independent, and used as building blocks while incrementally developing DW
  - ▶ Store data at the lowest level of detail and also as summaries. Dimensional model.
- ▶ Data marts are joined or 'unioned' together by dimensions





# Hybrid approach

- ▶ To handle the complexity of ETL process from operational data to data mart, some designer adds a **reconciled data layer**
  - ▶ Usually uses relational data model
  - ▶ Operational data are **integrated** and **cleansed** and the obtained data are saved in reconciled data layer
  - ▶ i.e. reconciled data layer is a consistent model of all corporate data, but its data model may not be suitable for data analysis
- ▶ Some ETL tool handles differences in data models when it loads data from reconciled data layer to data marts



# Benefits of data warehouses

---

- ▶ The ETL tools integrate and clean the data from operational databases when they load the DW
  - ▶ But they also duplicate the data ...
- ▶ Benefits to separate the DW from operational databases:
  - ▶ High availability of good quality data for analysis
  - ▶ Special-purpose data model for DW
  - ▶ Optimized data stores and indexing for DW

# Benefits, 1

---

- ▶ High availability of good quality data for analysis
  - ▶ Good quality information is always available, even when access to sources is denied
  - ▶ Analysis query processing typically involve a large amount of disk I/O and CPU time. It is important to ensure that data analysis does not affect the reliable processing of transactions in operational systems.

# Benefits, 2

---

- ▶ Special-purpose data model for DW
  - ▶ Operational databases are generally based on **relational** model. This model is suitable for data of fine granularity and linked by complex relationship
  - ▶ DWs are logically structured according to the **multi-dimensional** model
    - ▶ easier to understand
    - ▶ keep both detail and summarized data

# Benefits, 3

---

- ▶ Data warehouses can use specific design solutions aimed at performance optimization of analysis and report applications
  - ▶ E.g. column-oriented data stores, bitmap indexes
  - ▶ Ref: <https://docs.oracle.com/database/121/DWHSG/schemas.htm#DWHSG019>

	Operational databases	Data warehouses
Concurrency control and recovery	Required to ensure the consistency and robustness of transactions	Not required. OLAP uses mainly read-only access
Access methods	Typical SQL queries need to join several tables, but involve a smaller number of records	OLAP queries involve computation of large groups of data at summarized levels.
Time coverage	Current data only	Current and historical data. Data loading are done in batches

# Difference in Concurrency Control

---

- ▶ Typical transactions in operational databases involve both reads and writes
  - ▶ Concurrency control is essential to ensure consistency and robustness of transactions
  - ▶ (consider ACID properties)
  - ▶ In addition, need to recover from transaction failure.
- ▶ OLAP mainly uses read-only access. No concurrency control is required.
  - ▶ Data loading is done in batches

# Difference in Access Methods

---

- ▶ Typical transactions in operational databases need to join several tables, but involve a smaller number of records
  - ▶ In typical SQL queries, we need to
    - ▶ join several tables using primary and foreign keys
    - ▶ select a limited number of records from a large table using filter conditions
- ▶ OLAP queries involve computation of large groups of data at summarized levels.
  - ▶ We need to
    - ▶ Select a large portion of records from a large table
    - ▶ Summarize some columns, e.g. find the total sales amount

# Difference in Time Coverage

---

- ▶ Operational databases usually only keep current data, which may change frequently
  - ▶ E.g. purchase orders in an online shopping mall
- ▶ Data warehouses keep both current and historical data
  - ▶ E.g. purchase records of last 5 years
  - ▶ The data seldom changes after loading
  - ▶ Data loading are done in batches



# Review questions

---

- ▶ Describe the Two-layer architecture of data warehouse.
- ▶ What are the benefits of this architecture?
  - ▶ Describe the difference in data model between the data warehouse and operational databases.
  - ▶ Explain the different requirements in concurrency control and access method in OLAP and OLTP.
- ▶ Trace the flow of data through the data warehouse from beginning (source data) to end (analysis result).
- ▶ What are the functions of data staging?
- ▶ Compare data warehouses with data marts.
- ▶ Describe the bus architecture and the hub-and-spoke architecture.

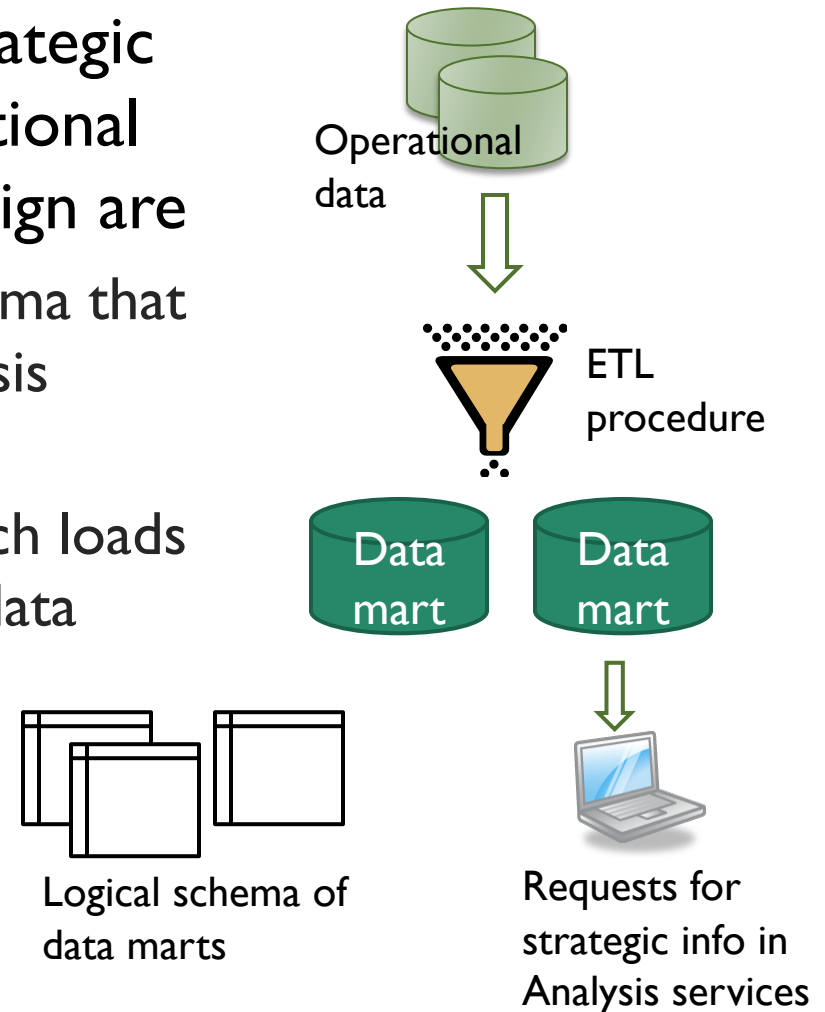
# Part C. Design Methodology

---

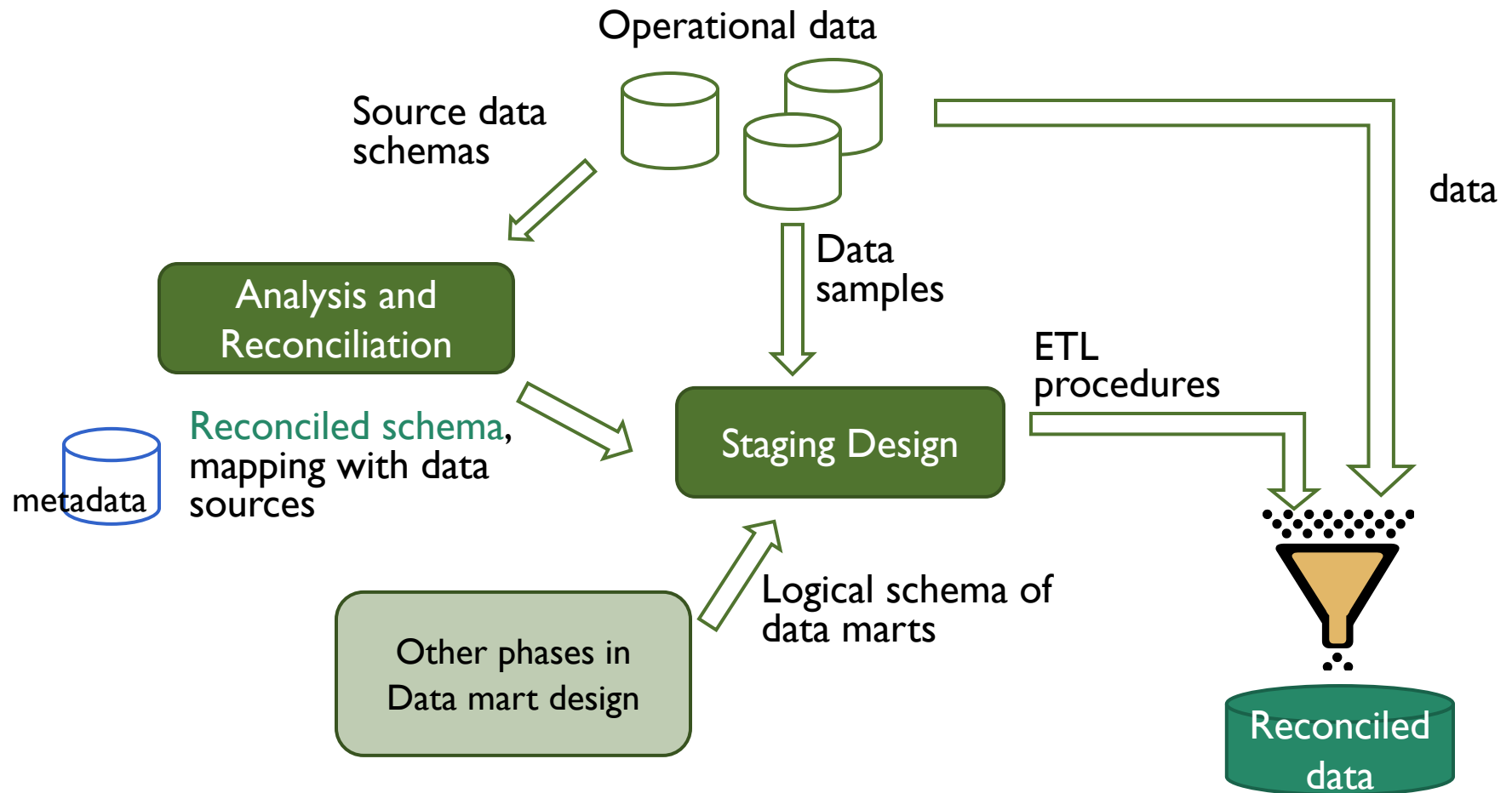
- ▶ We will use the Bus architecture and design data marts with conformed analysis dimensions in this module.
- ▶ Goals of data mart design
- ▶ The reconciled data layer
  - ▶ How to design
  - ▶ How to populate
- ▶ Seven phases in the Design Methodology

# Data Mart Design

- ▶ Given user requirements of strategic information and existing operational data, the goals of data mart design are
  - ▶ To design a logical / physical schema that is suitable to answer users' analysis requirements
  - ▶ To design an ETL procedure which loads the data mart from operational data



# Defining and populating the reconciled layer



# Analysis and Reconciliation

---

- ▶ This phase defines and documents the reconciled schema
- ▶ (Analysis) Understand the meaning and structure of source data
  - ▶ Analyze and understand available source schemata
  - ▶ Select which groups of data can be useful for the purposes of decision-making for the data mart
  - ▶ Assess data quality
- ▶ (Reconciliation) remove inconsistencies and integrate the data
  - ▶ Cleansing
  - ▶ Transformation
  - ▶ If multiple data sources are to be used, integrate their schemata to determine common features and remove every inconsistency

# Cleansing

---

- ▶ Improves data quality by rectifying data values
  - ▶ Duplicate data (e.g. a patient is recorded many times)
  - ▶ Inconsistent values that are logically associated (e.g. addresses and ZIP codes)
  - ▶ Missing data
  - ▶ Unexpected use of fields (e.g. email field to store tel no.)
  - ▶ Impossible or wrong values (e.g. 2/30/2009)
  - ▶ Inconsistent values for a single entity (e.g. Macau vs. Macao, abbreviation)
  - ▶ Typing mistakes

# Transformation

---

- ▶ **Convert data from its source format into the DW format**
  - ▶ Unit of measures (e.g. length in meter vs. feet)
  - ▶ Different data format (e.g. 1/2/2012, Feb 1, 2012)
  - ▶ Separation (e.g. split a name into first name and last name, split an address into building, street, zip code, city and country)
  - ▶ Matching equivalent fields in different sources
  - ▶ Different data model (e.g. XML to relational)

# Exercise

---

```
// DB1. Mobile service
Customer (contractNo, firstName, lastName, idcard, mobileNum, address,
contactTel, emailAddr)
VoiceUsage (callerMobileNum, calleeMobileNum, startTime, duration)
DataUsage (mobileNum, hourInDay, kbyteSent, kbyteReceived)
```

```
// DB2. Broadband service
Customer (userid, fullname, idcard, telephone, address)
Usage (userid, hourInDay, mbSent, mbReceived)
```

A telecommunication company provides both mobile phone and broadband Internet services. There are separate operational databases for the two services. The management wants to analyze the data usage of customers in different hours in a day. You are requested to design a data mart to integrate the data usage data from both databases. Discuss various issues involved in the design of the reconciled data layer.



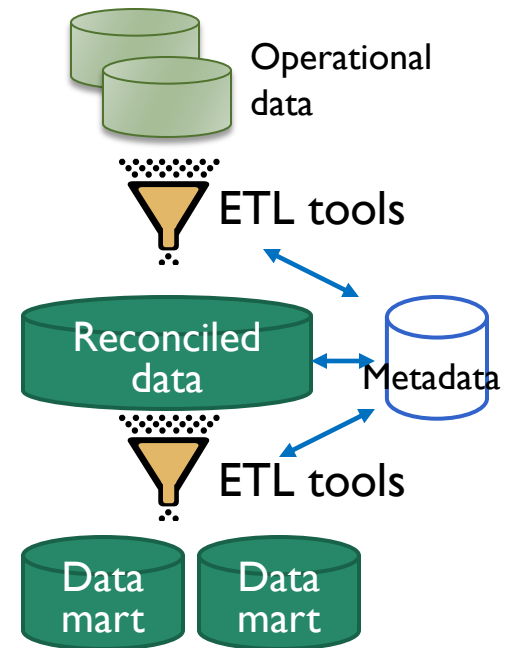
# Partial solution

---

- ▶ **Integration:**
  - ▶ How to identify the same person who is a patron of both services?
  - ▶ How to map the attribute of the two tables?
  - ▶ Primary key of the customer table in the reconciled schema?
  - ▶ How to merge the data usage in mobile and broadband networks?
- ▶ **Cleansing and transforming:**
  - ▶ Different unit?
  - ▶ Missing data? Separation
- ▶ **Also try to write the schema for the reconciled data layer.**

# Staging Design

- ▶ Define ETL procedures in order to load the data coming from operational sources into data marts
  - ▶ Operational data  $\Rightarrow$  reconciled database:
    - ▶ most complex
    - ▶ handle cleansing, transformation and integration
  - ▶ Reconciled database  $\Rightarrow$  data marts:
    - ▶ adjust the data in reconciled data to the star schema used for multi-dimensional analyses
    - ▶ Calculation of derived data
    - ▶ Denormalization, surrogate key, aggregate data
  - ▶ Usually implemented in data integration tools, e.g. 'Microsoft SQL Server Integration Services'

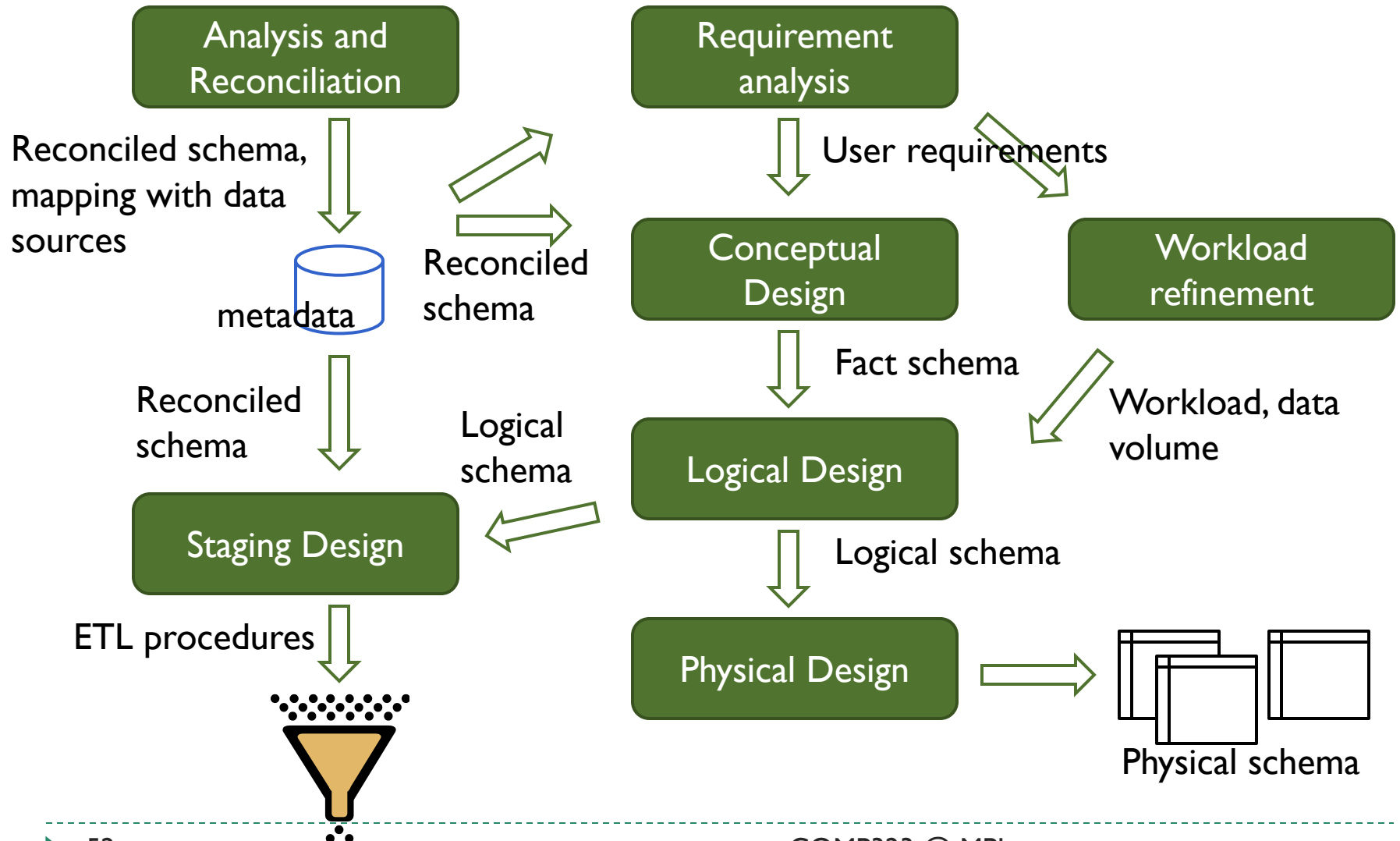


# Design Methodology of Data Mart

---

- ▶ We will use a seven-phase methodology to design a data mart
  - ▶ **Analysis and reconciliation** – design and document the reconciled schema
  - ▶ **Requirement analysis** – designers collect, filter, and document end-user requirements to select relevant information to achieve strategic goals
  - ▶ **Conceptual design** – create conceptual schema based on user requirements and reconciled schema
  - ▶ **Workload refinement** – validate the conceptual schema by checking the query workload against the schema
  - ▶ **Logical design** – select a logical model (e.g. ROLAP) and design logical schema
  - ▶ **Data staging design** – design the population process (ETL)
  - ▶ **Physical design** – select indexes to optimize performance. Also select the specific DBMS

# Design Methodology



Phases	Input	Output	People involved
Analysis and reconciliation	Operational source schemas	Reconciled schema	Designer, data processing center staff
Requirement analysis	Strategic goals	Requirement spec, preliminary workload	Designer, end users <span>Chap 2</span>
Conceptual design	Reconciled schema, requirement spec	Fact schemas	Designer, end users <span>Chap 2</span>
Workload refinement	Fact schemas, preliminary workload	Workload, data volume, validated fact schemas	Designer, end users <span>Chap 2</span>
Logical design	Fact schemas, target logical model, workload	Logical data mart schema	Designer <span>Chap 3</span>
Staging design	Source schemas, reconciled schema, logical data mart schema	ETL procedures	Designer, database administrators
Physical design	Logical data mart schema, target DBMS, workload	Physical data mart schema	Designer

# Review questions

---

- ▶ What are the goals of data mart design?
- ▶ What are usually done in the 'Analysis and reconciliation' phase?
- ▶ List four examples of data cleansing and data transforming in data staging.
- ▶ Try to design the schema of the reconciled database in the telecommunication company example.
- ▶ Describe the seven phases in the design methodology for data mart. Draw a diagram to illustrate the input and output of each phase.
- ▶ Compare the design methodology of data marts with the design methodology of operational databases.

# Outline for the rest of the course

---

## Data Warehousing

- ▶ Chap 2. Requirement analysis, Conceptual design
- ▶ Chap 3. Logical design

## Data Analysis

- ▶ Chap 3. Relational OLAP and SQL queries
- ▶ Chap 5. Data mining