

考試日期：100 年 11 月 1 日 (星期二) 第 4 節	※請務必註記，准帶項目打「V」，否則打「X」。					1. 需加發計算紙(B5 白紙)或答案紙請在試題內封袋備註。 2. 本命題紙為 A4 大小，印刷格式統一由閱場判斷是否放大或合併印刷，如有特殊需求，請註記：
本試題共 1 頁，附件共 0 頁，印刷份數：20 份	計算機	課本	筆記	字典 電子辭典	其他	放大成 B4
備註：可另攜 A4 筆記 1 張(親筆手寫) (禁止攜帶影印及列印輸出版本筆記) 考試時間：100/11/01 14:20-15:50 教室：L310	X	X	V	X	X	

1. What is Data Warehouse? (10%)

**Ans:** “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

**Ref:** 1001DW04 pp.5-10

## 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

Source: Han & Kamber (2006)

5

## Data Warehouse

- Subject-oriented
- Integrated
- Time-variant
- Nonvolatile

Source: Han & Kamber (2006)

6

### 第 1 頁

注意：1、考試求公平及公正，請同學務必自律，維護學校與學生之榮譽。  
2、考試時不得有交談、攜卷出場、窺視、傳遞、代考、夾帶等違規行為，違者將受嚴重議處。

考試日期：100 年 11 月 1 日 (星期二) 第 4 節	※請務必註記，准帶項目打「V」，否則打「X」。					1. 需加發計算紙(B5 白紙)或答案紙請在試題內封袋備註。 2. 本命題紙為 A4 大小，印刷格式統一由閱場判斷是否放大或合併印刷，如有特殊需求，請註記：
本試題共 1 頁，附件共 0 頁，印刷份數：20 份	計算機	課本	筆記	字典 電子辭典	其他	放大成 B4
備註：可另攜 A4 筆記 1 張(親筆手寫) (禁止攜帶影印及列印輸出版本筆記) 考試時間：100/11/01 14:20-15:50 教室：L310	X	X	V	X	X	

### 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**

Source: Han & Kamber (2006)

7

### 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.

Source: Han & Kamber (2006)

8

### 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"

Source: Han & Kamber (2006)

9

### Data Warehouse— Nonvolatile

- 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**

Source: Han & Kamber (2006)

10

## 第 2 頁

- 注意：1、考試求公平及公正，請同學務必自律，維護學校與學生之榮譽。  
2、考試時不得有交談、攜卷出場、窺視、傳遞、代考、夾帶等違規行為，違者將受嚴重議處。

考試日期：100 年 11 月 1 日 (星期二) 第 4 節	※請務必註記，准帶項目打「V」，否則打「X」。					1. 需加發計算紙(B5 白紙)或答案紙請在試題內封袋備註。 2. 本命題紙為 A4 大小，印刷格式統一由閱場判斷是否放大或合併印刷，如有特殊需求，請註記：
本試題共 1 頁，附件共 0 頁，印刷份數：20 份	計算機	課本	筆記	字典 電子辭典	其他	放大成 B4
備註：可另攜 A4 筆記 1 張(親筆手寫) (禁止攜帶影印及列印輸出版本筆記) 考試時間：100/11/01 14:20-15:50 教室：L310	X	X	V	X	X	

2. How is a data warehouse different from a database? How are they similar? (10%)

Ans:

Ref: 1001DW02 pp.19-20

## Differences between a data warehouse and a database

- Data warehouse:
  - A data warehouse is a repository of information collected from **multiple sources** over a **history of time** stored under a **unified schema** and used for **data analysis** and **decision support**
  - There could be multiple heterogeneous databases where the schema of one database may not agree with the schema of another.
- Database:
  - A database is a collection of **interrelated data** that represents the current status of the stored data.
  - A database system supports ad-hoc query and on-line transaction processing.

Source: Han & Kamber (2006)

19

## Similarities between a data warehouse and a database

- Both are repositories of information storing huge amounts of persistent data.

Source: Han & Kamber (2006)

20

### 第 3 頁

注意：1、考試求公平及公正，請同學務必自律，維護學校與學生之榮譽。  
2、考試時不得有交談、攜卷出場、窺視、傳遞、代考、夾帶等違規行為，違者將受嚴重議處。

考試科目：資料倉儲

開課班別：資管 系 四年 P 班 命題教授：戴敏育

考試日期：100 年 11 月 1 日 (星期二) 第 4 節

※請務必註記，准帶項目打「V」，否則打「X」。

1. 需加發計算紙(B5 白紙)或答案紙請在試題內封袋備註。  
2. 本命題紙為A4 大小，印刷格式統一由閱場判斷是否放大或合併印刷，如有特殊需求，請註記：

本試題共 1 頁，附件共 0 頁，印刷份數：20 份

計算機	課本	筆記	字典 電子辭典	其他
X	X	V	X	X

放大成 B4

備註：可另攜 A4 筆記 1 張(親筆手寫)

(禁止攜帶影印及列印輸出版本筆記)

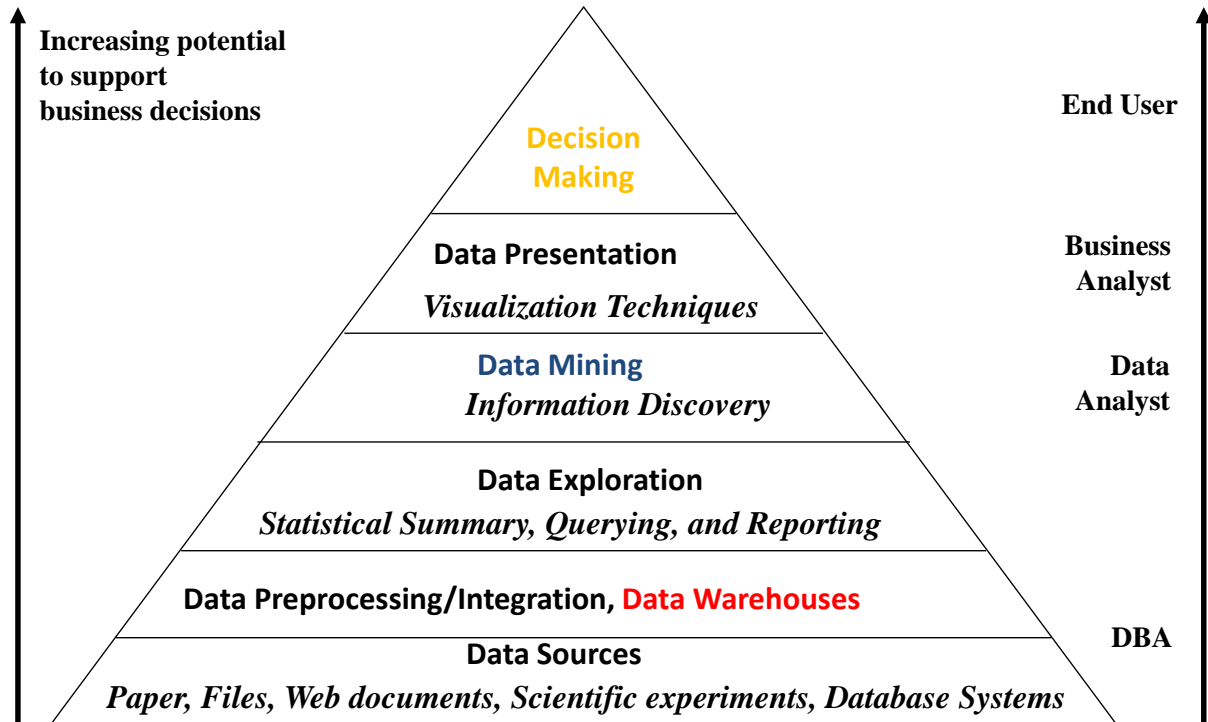
考試時間：100/11/01 14:20-15:50 教室：L310

3. What are the relationships between Data Warehouse, Data Mining, and Business Intelligence? (10%)

Ans:

Ref: 1001DW02 p.5 (1001DW01 p.6)

## Data Warehouse Data Mining and Business Intelligence



Source: Han & Kamber (2006)

5

考試日期：100 年 11 月 1 日 (星期二) 第 4 節  
 本試題共 1 頁，附件共 0 頁，印刷份數：20 份  
 備註：可另攜 A4 筆記 1 張(親筆手寫)  
 (禁止攜帶影印及列印輸出版本筆記)  
 考試時間：100/11/01 14:20-15:50 教室：L310

※請務必註記，准帶項目打「V」，否則打「X」。	計算機	課本	筆記	字典 電子辭典	其他
	X	X	V	X	X

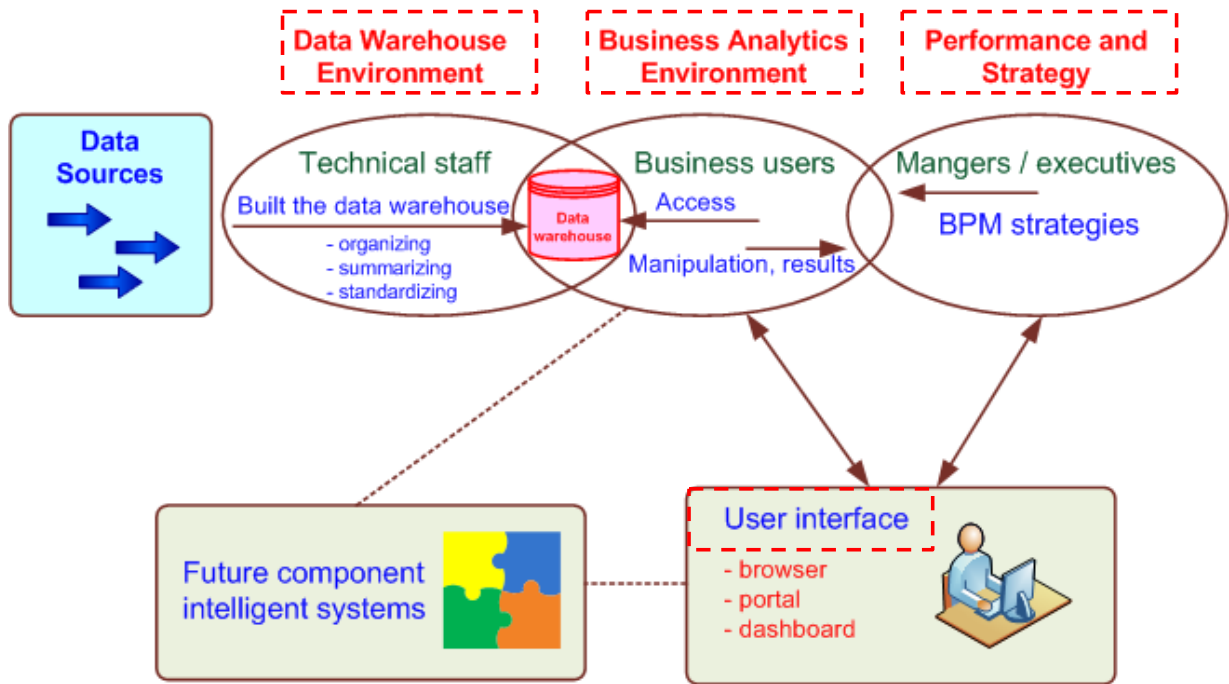
1. 需加發計算紙(B5 白紙)或答案紙請在試題內封袋備註。  
 2. 本命題紙為 A4 大小，印刷格式統一由閱場判斷是否放大或合併印刷，如有特殊需求，請註記：  
**放大成 B4**

4. What is the high-level architecture of Business Intelligence (BI)? (10%)

Ans:

Ref: 1001DW02 p.10

## A High-Level Architecture of BI



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

10

考試日期：100 年 11 月 1 日 (星期二) 第 4 節

※請務必註記，准帶項目打「V」，否則打「X」。

1. 需加發計算紙(B5 白紙)或答案紙請在試題內封袋備註。  
2. 本命題紙為A4 大小，印刷格式統一由閱場判斷是否放大或合併印刷，如有特殊需求，請註記：

本試題共 1 頁，附件共 0 頁，印刷份數：20 份

計算機	課本	筆記	字典 電子辭典	其他
X	X	V	X	X

備註：可另攜 A4 筆記 1 張(親筆手寫)

(禁止攜帶影印及列印輸出版本筆記)

考試時間：100/11/01 14:20-15:50 教室：L310

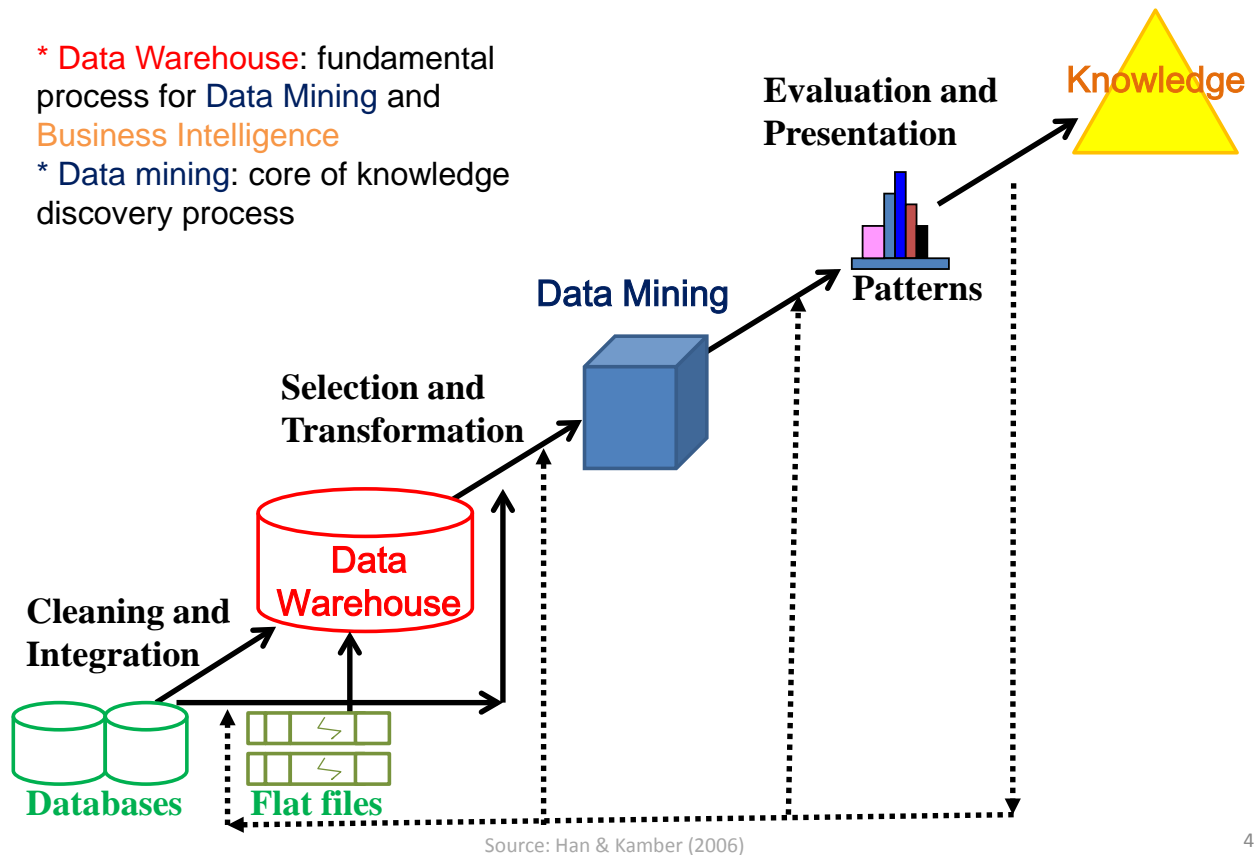
放大成 B4

5. What is the knowledge discovery (KDD) process? (10%)

Ans:

Ref: 1001DW02 p.4 (1001DW01 p.5)

## Knowledge Discovery (KDD) Process



考試日期：100 年 11 月 1 日 (星期二) 第 4 節

※請務必註記，准帶項目打「V」，否則打「X」。

1. 需加發計算紙(B5 白紙)或答案紙請在試題內封袋備註。  
2. 本命題紙為A4 大小，印刷格式統一由閱場判斷是否放大或合併印刷，如有特殊需求，請註記：

本試題共 1 頁，附件共 0 頁，印刷份數：20 份

計算機	課本	筆記	字典 電子辭典	其他
-----	----	----	------------	----

備註：可另攜 A4 筆記 1 張(親筆手寫)

X	X	V	X	X
---	---	---	---	---

(禁止攜帶影印及列印輸出版本筆記)

考試時間：100/11/01 14:20-15:50 教室：L310

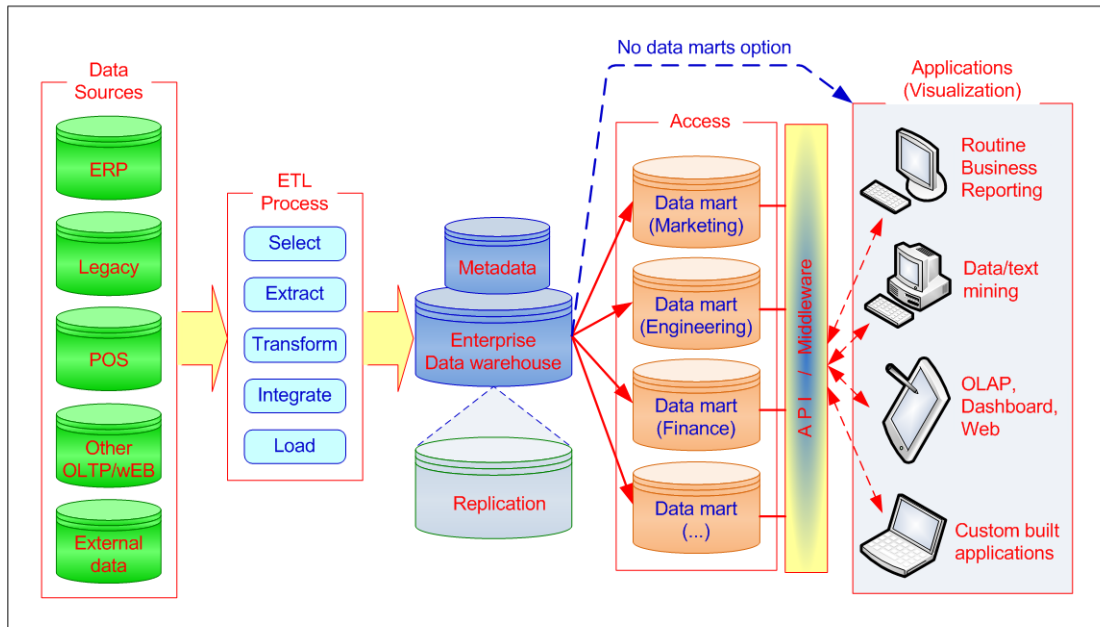
放大成 B4

6. What is the conceptual framework for data warehousing? (10%)

Ans:

Ref: 1001DW03 p.4

## A Conceptual Framework for Data Warehousing



source: Turban et al. (2011), Decision Support and Business Intelligence Systems



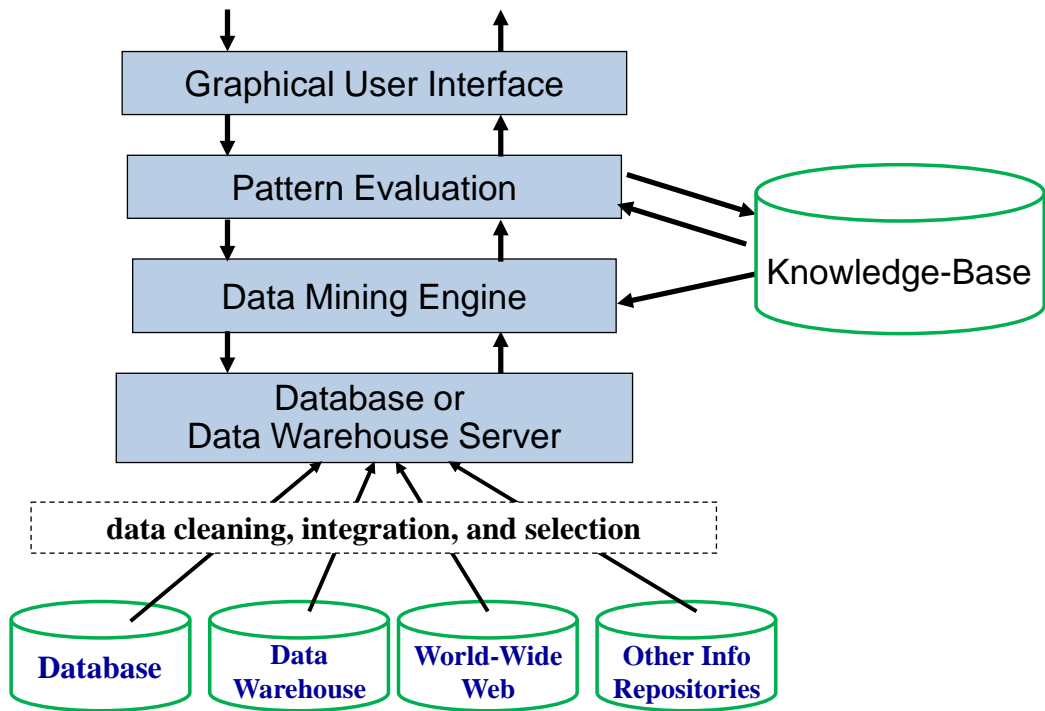
考試日期：100 年 11 月 1 日 (星期二) 第 4 節	※請務必註記，准帶項目打「V」，否則打「X」。					1. 需加發計算紙(B5 白紙)或答案紙請在試題內封袋備註。 2. 本命題紙為 A4 大小，印刷格式統一由閱場判斷是否放大或合併印刷，如有特殊需求，請註記：
本試題共 1 頁，附件共 0 頁，印刷份數：20 份	計算機	課本	筆記	字典 電子辭典	其他	放大成 B4
備註：可另攜 A4 筆記 1 張(親筆手寫) (禁止攜帶影印及列印輸出版本筆記) 考試時間：100/11/01 14:20-15:50 教室：L310	X	X	V	X	X	

7. What is the architecture of a typical data mining system? (10%)

Ans:

Ref: 1001DW02 p.15

## Architecture of a typical data mining system



Source: Han & Kamber (2006)

15



考試科目：資料倉儲

開課班別：資管 系 四年 P 班 命題教授：戴敏育

考試日期：100 年 11 月 1 日 (星期二) 第 4 節	※請務必註記，准帶項目打「V」，否則打「X」。					1. 需加發計算紙(B5 白紙)或答案紙請在試題內封袋備註。 2. 本命題紙為 A4 大小，印刷格式統一由閱場判斷是否放大或合併印刷，如有特殊需求，請註記：
本試題共 1 頁，附件共 0 頁，印刷份數：20 份	計算機	課本	筆記	字典 電子辭典	其他	放大成 B4
備註：可另攜 A4 筆記 1 張(親筆手寫) (禁止攜帶影印及列印輸出版本筆記) 考試時間：100/11/01 14:20-15:50 教室：L310	X	X	V	X	X	

8. What are OLTP and OLAP? What are the differences between OLTP and OLAP? (10%)

Ans:

Ref: 1001DW04 pp.12-13

### Data Warehouse vs. Operational DBMS

- OLTP (on-line transaction processing)
  - Major task of traditional relational DBMS
  - Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
- OLAP (on-line analytical processing)
  - Major task of data warehouse system
  - Data analysis and decision making
- Distinct features (OLTP vs. OLAP):
  - User and system orientation: customer vs. market
  - Data contents: current, detailed vs. historical, consolidated
  - Database design: ER + application vs. star + subject
  - View: current, local vs. evolutionary, integrated
  - Access patterns: update vs. read-only but complex queries

Source: Han & Kamber (2006)

12

### OLTP vs. OLAP

	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

Source: Han & Kamber (2006)

13

### 第 9 頁

- 注意：1、考試求公平及公正，請同學務必自律，維護學校與學生之榮譽。  
2、考試時不得有交談、攜卷出場、窺視、傳遞、代考、夾帶等違規行為，違者將受嚴重議處。

考試日期：100 年 11 月 1 日 (星期二) 第 4 節	※請務必註記，准帶項目打「V」，否則打「X」。					1. 需加發計算紙(B5 白紙)或答案紙請在試題內封袋備註。 2. 本命題紙為 A4 大小，印刷格式統一由閱場判斷是否放大或合併印刷，如有特殊需求，請註記：
本試題共 1 頁，附件共 0 頁，印刷份數：20 份	計算機	課本	筆記	字典 電子辭典	其他	放大成 B4
備註：可另攜 A4 筆記 1 張(親筆手寫) (禁止攜帶影印及列印輸出版本筆記) 考試時間：100/11/01 14:20-15:50 教室：L310	X	X	V	X	X	

9. Why is data preprocessing important? What are the major tasks in data preprocessing? (10%)

Ans:

Ref: 1001DW03 p.12, 14 (Option: pp.10-11)

### Why Is Data Preprocessing Important?

- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
  - Data warehouse needs consistent integration of quality data
- Data **extraction, cleaning, and transformation** comprises the majority of the work of building a data warehouse

Source: Han & Kamber (2006)

12

### Major Tasks in Data Preprocessing

- Data cleaning
  - Fill in **missing values**, smooth **noisy data**, identify or remove **outliers**, and resolve **inconsistencies**
- Data integration
  - Integration of multiple **databases**, **data cubes**, or **files**
- Data transformation
  - **Normalization** and **aggregation**
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
  - Part of data reduction but with particular importance, especially for numerical data

Source: Han & Kamber (2006)

14

### 第 10 頁

- 注意：1、考試求公平及公正，請同學務必自律，維護學校與學生之榮譽。  
2、考試時不得有交談、攜卷出場、窺視、傳遞、代考、夾帶等違規行為，違者將受嚴重議處。

考試日期：100 年 11 月 1 日 (星期二) 第 4 節	※請務必註記，准帶項目打「V」，否則打「X」。					1. 需加發計算紙(B5 白紙)或答案紙請在試題內封袋備註。 2. 本命題紙為 A4 大小，印刷格式統一由閱場判斷是否放大或合併印刷，如有特殊需求，請註記：
本試題共 1 頁，附件共 0 頁，印刷份數：20 份	計算機	課本	筆記	字典 電子辭典	其他	放大成 B4
備註：可另攜 A4 筆記 1 張(親筆手寫) (禁止攜帶影印及列印輸出版本筆記) 考試時間：100/11/01 14:20-15:50 教室：L310	X	X	V	X	X	

## Why Data Preprocessing?

- Data in the real world is dirty
  - **incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation=" "
  - **noisy**: containing errors or outliers
    - e.g., Salary="-10"
  - **inconsistent**: containing discrepancies in codes or names
    - e.g., Age="42" Birthday="03/07/1997"
    - e.g., Was rating "1,2,3", now rating "A, B, C"
    - e.g., discrepancy between duplicate records

Source: Han &amp; Kamber (2006)

10

## Why Is Data Dirty?

- Incomplete data may come from
  - "Not applicable" data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)
- Duplicate records also need data cleaning

Source: Han &amp; Kamber (2006)

11

考試日期：100 年 11 月 1 日 (星期二) 第 4 節	※請務必註記，准帶項目打「V」，否則打「X」。					1. 需加發計算紙(B5 白紙)或答案紙請在試題內封袋備註。 2. 本命題紙為 A4 大小，印刷格式統一由閱場判斷是否放大或合併印刷，如有特殊需求，請註記：
本試題共 1 頁，附件共 0 頁，印刷份數：20 份	計算機	課本	筆記	字典 電子辭典	其他	放大成 B4
備註：可另攜 A4 筆記 1 張(親筆手寫) (禁止攜帶影印及列印輸出版本筆記) 考試時間：100/11/01 14:20-15:50 教室：L310	X	X	V	X	X	

10. What is the conceptual modeling of data warehouses? Please illustrate the major schemas used in data warehouse. (10%)

Ans:

Ref: 1001DW04 pp.17-20

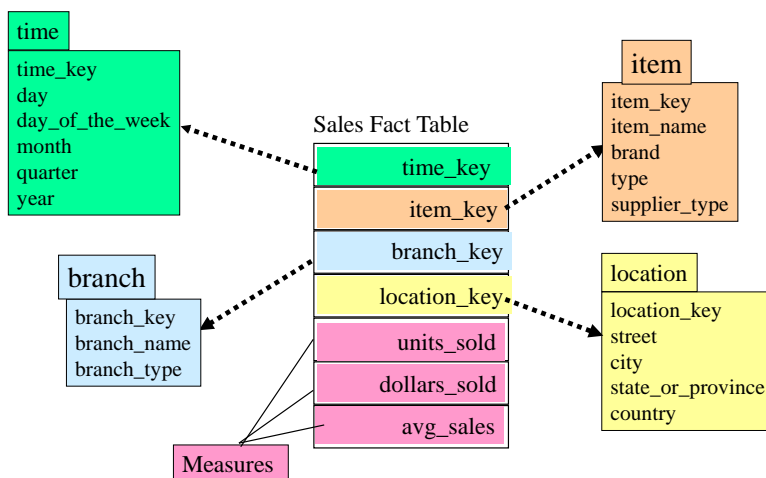
### Conceptual Modeling of Data Warehouses

- Modeling data warehouses: dimensions & measures
  - Star schema: A fact table in the middle connected to a set of dimension tables
  - Snowflake schema: A refinement of star schema where some dimensional hierarchy is **normalized** into a set of smaller dimension tables, forming a shape similar to snowflake
  - Fact constellations: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called **galaxy schema** or fact constellation

Source: Han & Kamber (2006)

17

### Example of Star Schema



Source: Han & Kamber (2006)

18

考試日期：100 年 11 月 1 日 (星期二) 第 4 節 ※請務必註記，准帶項目打「V」，否則打「X」。

1. 需加發計算紙(B5 白紙)或答案紙請在試題內封袋備註。  
2. 本命題紙為 A4 大小，印刷格式統一由閱場判斷是否放大或合併印刷，如有特殊需求，請註記：

本試題共 1 頁，附件共 0 頁，印刷份數：20 份

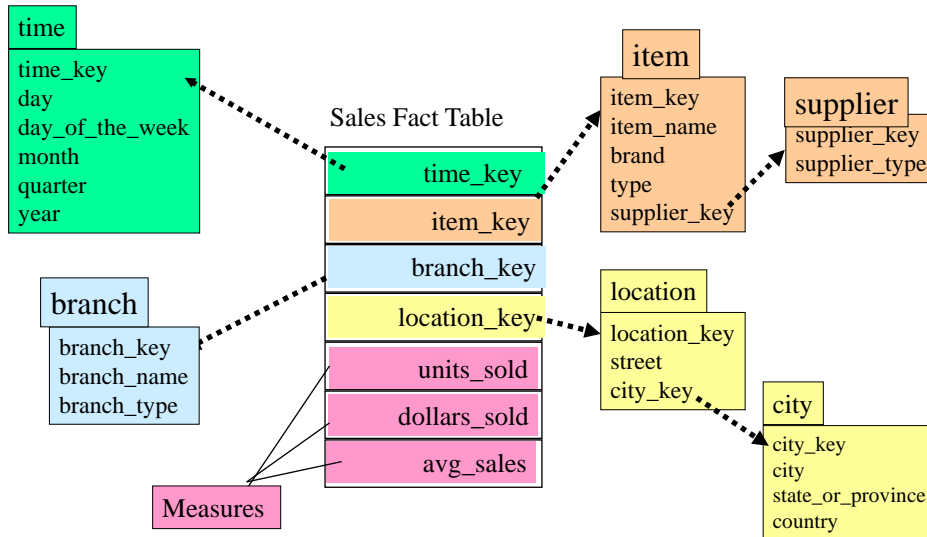
計算機	課本	筆記	字典 電子辭典	其他
X	X	V	X	X

備註：可另攜 A4 筆記 1 張(親筆手寫)  
(禁止攜帶影印及列印輸出版本筆記)

考試時間：100/11/01 14:20-15:50 教室：L310

放大成 B4

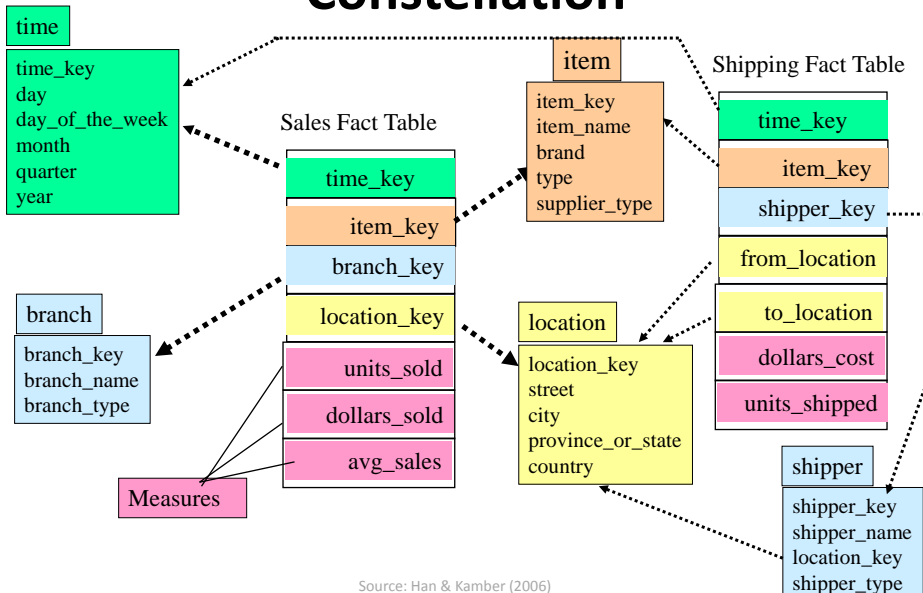
## Example of Snowflake Schema



Source: Han & Kamber (2006)

19

## Example of Fact Constellation



Source: Han & Kamber (2006)

考試日期：100 年 11 月 1 日 (星期二) 第 4 節  
 本試題共 1 頁，附件共 0 頁，印刷份數：20 份  
 備註：可另攜 A4 筆記 1 張 (親筆手寫)  
 (禁止攜帶影印及列印輸出版本筆記)  
 考試時間：100/11/01 14:20-15:50 教室：L310

※請務必註記，准帶項目打「V」，否則打「X」。	計算機	課本	筆記	字典 電子辭典	其他
	X	X	V	X	X

1. 需加發計算紙(B5 白紙)或答案紙請在試題內封袋備註。  
 2. 本命題紙為 A4 大小，印刷格式統一由閱場判斷是否放大或合併印刷，如有特殊需求，請註記：  
**放大成 B4**

11. What are the typical OLAP operations in data warehouse? (10%)

Ans:

Ref: 1001DW04 pp.32-33

## Typical OLAP Operations

- Roll up (drill-up): summarize data
  - by climbing up hierarchy or by dimension reduction
- Drill down (roll down): reverse of roll-up
  - from higher level summary to lower level summary or detailed data, or introducing new dimensions
- Slice and dice: project and select
- Pivot (rotate):
  - reorient the cube, visualization, 3D to series of 2D planes
- Other operations
  - drill across: involving (across) more than one fact table
  - drill through: through the bottom level of the cube to its back-end relational tables (using SQL)

