

Data Warehousing

Lecture 1 Introduction to Data Warehousing

CITS3401
CITS5504

Dr. Sirui Li

Computer Science and
Software Engineering

School of Maths, Physics
and Computing

Acknowledgement: The lecture slides are adapted from online sources. Please see references at the end of the lecture.

2

Outline of Today's Lecture

- Unit Structure and Overview
- A Brief History of Data(base) Technologies
- Storing Data in Databases and Data Warehouses
- What is a Data Warehouse
- OLTP vs. OLAP
- Data Mining

2

Teaching team for CITS3401&5504

- **Lecturer and Unit Coordinator**
 - Dr. Sirui Li. Preferred name Siri.
 - **Email:** sirui.li@uwa.edu.au
 - **Office:** Maths Room 2.04
 - **Consultation:** Wednesday 15:30-16:30 in office and on Teams (no appointments needed)

3

Teaching team for CITS3401&5504



Jiachuan Liu (Master of DS) - CITS3401



Jichunyang Li (MBA Graduate with Distinction) – CITS5504



Pascal Sun (PhD candidate) – CITS5504

Unit	Day	Time	Venue	Lab Facilitator
CITS3401_SEM-1_CR - Data Warehousing	Monday	10:00 - 11:00	CSSE: [201] Computer Lab	Jiachuan Liu
CITS3401_SEM-1_CR - Data Warehousing	Monday	11:00 - 12:00	CSSE: [201] Computer Lab	Jiachuan Liu
CITS5504_SEM-1_CR - Data Warehousing	Monday	14:00 - 15:00	CSSE: [205] Computer Lab	Pascal Sun
CITS5504_SEM-1_CR - Data Warehousing	Monday	14:00 - 15:00	CSSE: [201] Computer Lab	Jichunyang Li
CITS5504_SEM-1_CR - Data Warehousing	Monday	15:00 - 16:00	CSSE: [201] Computer Lab	Jichunyang Li
CITS5504_SEM-1_CR - Data Warehousing	Tuesday	13:00 - 14:00	CSSE: [205] Computer Lab	Jichunyang Li
CITS5504_SEM-1_CR - Data Warehousing	Wednesday	08:00 - 09:00	CSSE: [201] Computer Lab	Jichunyang Li
CITS5504_SEM-1_CR - Data Warehousing	Wednesday	09:00 - 10:00	CSSE: [201] Computer Lab	Jichunyang Li
CITS5504_SEM-1_CR - Data Warehousing	Wednesday	17:00 - 18:00	ENCM: [2078] South Civil Computer Room B	Jichunyang Li
CITS3401_SEM-1_CR - Data Warehousing	Thursday	08:00 - 09:00	CSSE: [201] Computer Lab	Jiachuan Liu
CITS3401_SEM-1_CR - Data Warehousing	Thursday	09:00 - 10:00	CSSE: [201] Computer Lab	Jiachuan Liu
CITS5504_SEM-1_CR - Data Warehousing	Thursday	16:00 - 17:00	CSSE: [201] Computer Lab	Jichunyang Li
CITS5504_SEM-1_CR - Data Warehousing	Thursday	17:00 - 18:00	CSSE: [201] Computer Lab	Jichunyang Li
CITS3401_SEM-1_CR - Data Warehousing	Friday	10:00 - 11:00	CSSE: [205] Computer Lab	Jiachuan Liu
CITS3401_SEM-1_CR - Data Warehousing	Friday	11:00 - 12:00	CSSE: [205] Computer Lab	Jiachuan Liu
CITS3401_SEM-1_CR - Data Warehousing	Friday	12:00 - 13:00	CSSE: [201] Computer Lab	Jiachuan Liu
CITS3401_SEM-1_CR - Data Warehousing	Friday	13:00 - 14:00	CSSE: [201] Computer Lab	Jiachuan Liu

4

CITS3401 and CITS5504



- Same Lecture Hours and Venue:
 - Time: Wednesdays 13:00 – 15:00
 - Venue: PHYS [G41]
 - All the lectures will be recorded.
- Same Consultation Hour:
 - Wednesdays 15:30 – 16:30 in person and on Teams (no appointment needed).
 - Other times can be arranged too (send an email to book).
 - Sirui is in Maths Room 2.04 and will also be on Microsoft Teams.
- Similar Teaching Material
 - Same lecture slides and lab sheets
 - **Similar** projects (e.g. extra challenging tasks for CITS5504 students)
- Same Web Pages on LMS
- Same Teams Area
 - Unit Specific Channels
- Same Assessment Structures

5

Assessment Structures



Assessment #	Assessment Task	Weight	Assessment period/date/s	Submission Procedure
1	Lab 1	2%	29/3 11:59 PM	LMS
	Lab 4	3%	29/3 11:59 PM	during lab sessions
2	Project 1	25%	14/4 11:59 PM	cssubmit
	Project 2	20%	26/5 11:59 PM	cssubmit
3	Exam	50%	03/6 08:00 AM - 18/6 05:00 PM	Exam Venues

There is no marks allocated to participation. However, active engagement in lectures and lab activities are strong indicators of achieving success in this unit.

6

Lab sessions



- Weekly supervised lab session (1 hour) starting from Week 2.
- Encouraged to post questions on Teams.
- No lab sessions during public holidays and study break.
- Important Dates:
 - <https://www.uwa.edu.au/students/My-course/Important-dates>

7

Projects



Two projects worth 45%.

- Project 1 submission (25%)
 - individual effort
 - final submission at the end of Week 6 (14/04/2024)
- Project 2 submission (20%)
 - group of 1-2 students
 - final submission at the end of Week 12 (26/05/2024)
- Project 1: Analysis of a business scenario through OLAP (Online Analytical Processing) tools:
 - Docker, Python
 - PostgreSQL/SQL server
 - Atoti
- Project 2: Implementing a Graph Database for Data Warehousing Queries.
 - Neo4J <https://neo4j.com/>
 - Neo4J is the most popular graph data platform



8

Setting up Software Environment

There are two approaches

1. Install all the software in your **own computer**

- **Recommended approach**
- **Pros:** fast and accessible at any time
- **Cons:** time to install and fix some tech issues

2. Use lab computers

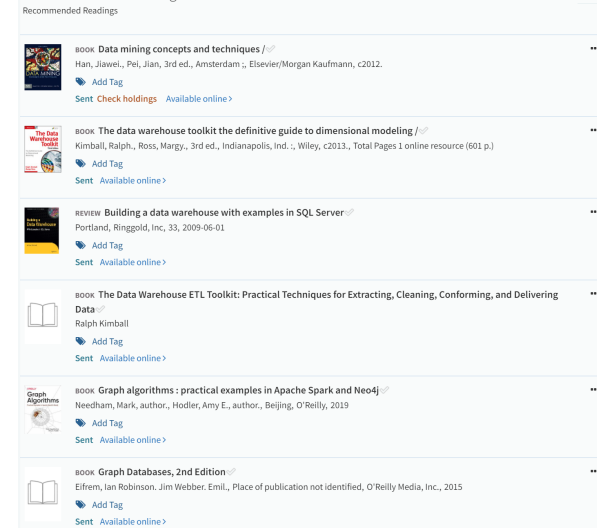
- **Pros:** easier to setup; no need for your own computer
- **Cons:** less flexibly, and it might be challenging to install packages due to permission issues.

- Always **backup your work or important data to OneDrive.**

9

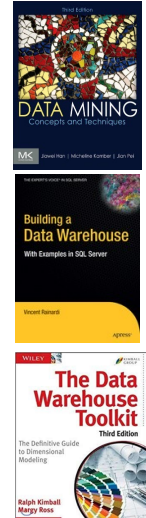
Text Books and Recommended Readings

Recommended Readings (6)



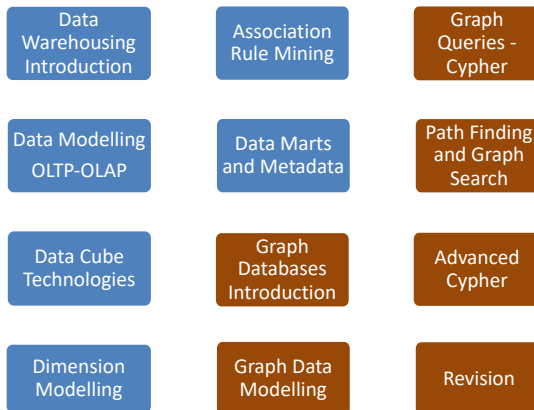
Recommended Readings

- book Data mining concepts and techniques / Han, Jiawei., Pei, Jian, 3rd ed., Amsterdam : Elsevier/Morgan Kaufmann, c2012.
Add Tag
Sent Check holdings Available online >
- book The data warehouse toolkit the definitive guide to dimensional modeling / Kimball, Ralph., Ross, Margy., 3rd ed., Indianapolis, Ind. : Wiley, c2013., Total Pages 1 online resource (801 p.)
Add Tag
Sent Available online >
- REVIEW Building a data warehouse with examples in SQL Server / Portland, Ringgold, Inc. 33, 2009-06-01.
Add Tag
Sent Available online >
- book The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data / Kimball, Ralph
Add Tag
Sent Available online >
- book Graph algorithms : practical examples in Apache Spark and Neo4j / Needham, Mark, author., Hodler, Amy E., author., Beijing, O'Reilly, 2019
Add Tag
Sent Available online >
- book Graph Databases, 2nd Edition / Eifrem, Ian Robinson. Jim Webber, Emil., Place of publication not identified, O'Reilly Media, Inc., 2015
Add Tag
Sent Available online >



10

Unit Overview



11

Outline of Today's Lecture

- Unit Structure and Overview
- **A Brief History of Data(base) Technologies**
- Storing Data in Databases and Data Warehouses
- What is Data Warehouse
- OLTP vs. OLAP
- Data Mining

12

The Evolution of Data Technologies



- The “dark age”: paper forms in file cabinets
- Computerised systems emerged
 - Initially for big projects like Social Security
 - Same functionality as old paper-based systems
- The “golden age”: databases are everywhere
 - Most activities tracked electronically
 - Stored data provides detailed history of activity
- The next step: use data for decision-making
 - Knowledge discovery from data (a.k.a. Data Mining)
 - One of the enabling technologies: **data warehousing**

Cambridge Analytica Ltd was founded in 2013.

13

Evolution of Database Technology



- **1960s:**
 - (Electronic) Data collection, database creation
 - IMS (database system by IBM) and network DBMS
 - IMS introduced **application code should be separated from data.**
- **1970s:**
 - Relational data model, relational DBMS implementation
 - The term “relational database” was invented by E. F. Codd at IBM.
 - E. F. Codd won Turing Award in 1981.
- **1980s:** **Microsoft SQL Server was first released in 1989; MySQL in 1995**
 - RDBMS, advanced data models (Object Oriented, deductive, etc.)
 - Application-oriented DBMS (spatial, scientific, engineering, etc.)
 - SQL standard adopted by ISO and ANSI.
- **1990s:**
 - Data mining, data warehousing, multimedia DBs, and Web DBs

Google was founded in 1998; Yahoo! in 1994; Baidu and Yandex in 2000

14

Evolution of Database Technology (cont.)



- **2000s**
 - Stream data management and mining
 - Data mining and its applications
 - Web tech. (XML, data integration) and geographic info. systems
 - Frequently asked question in job interviews:
 - **Did you work on XML related research?**
- **2010s** **Apache Spark was first released in 2014.**
 - NoSQL (Graph Databases, Document Stores)
 - Mining from Varieties of Data:
 - Multimedia Data (Images, Audios, and Videos)
 - Natural Language Processing, Social Network Data
 - Machine (Deep) Learning on various data types (e.g. text, videos)
 - Self-Driving Databases, Autonomous Databases
 - Frequently asked question in job interviews:
 - **Did you work on ML related research?**

15

Outline of Today's Lecture



- Unit Structure and Overview
- A Brief History of Data(base) Technologies
- **Storing Data in Databases and Data Warehouses**
- What is Data Warehouse
- OLTP vs. OLAP
- Data Mining

16

Where to Store Data

Relational Databases

- Support “delete, insert, update, and query”
- Consistency/integrity is crucial
- Queries are often simple
- Data from single department/organisation



Data Warehouses

- Mainly support “query”
- Queries are more complex



Other distributed file systems

- Hadoop Distributed File Systems (HDFS)



17

Storing Data in Relational Database

- A relational database is a **collection of tables**, each of which is assigned a unique name.
 - Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples (records or rows).
 - Each tuple in a relational table represents an object identified by unique key and described by a set of attribute values.

Student_id	Name	Unit_id	School_id	Score
2212201	Jan Smith	CITS3401	006	98
...

18

Storing Data in Relational Database (cont.)

- A relational database is a **collection of tables**, each of which is assigned a unique name.
 - Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples (records or rows).
 - Each tuple in a relational table represents an object identified by unique key and described by a set of attribute values.
- Relational data can be accessed by database queries written in a relational language such as **SQL**.
- A given query is transformed into a set of relational operations such as *join*, *selection* and *projection*, and is then optimised for efficient processing.
- Efficiency of **retrieval**, efficiency of **update** and **integrity** are the key requirements of a good relational database.

19

An Example - AllElectronics

- Four relational tables: *customer*, *item*, *employee* and *branch*.

customer

<u>cust_ID</u>	name	address	age	income	credit_info	category	...
C1	Smith, Sandy	1223 Lake Ave., Chicago, IL	31	\$78000	1	3	...
...

item

<u>item_ID</u>	name	brand	category	type	price	place_made	supplier	cost
I3	hi-res-TV	Toshiba	high resolution	TV	\$988.00	Japan	NikoX	\$600.00
I8	Laptop	Dell	laptop	computer	\$1369.00	USA	Dell	\$983.00
...

employee

<u>empl_ID</u>	name	category	group	salary	commission
E55	Jones, Jane	home entertainment	manager	\$118,000	2%
...

branch

<u>branch_ID</u>	name	address
B1	City Square	396 Michigan Ave., Chicago, IL
...

20

An Example – AllElectronics (cont.)

- Four relational tables: *customer*, *item*, *employee* and *branch*.
- Each relation consists of a set of attributes.

cust_ID	name	address	age	income	credit_rft	category	...
C1	Smith, Sandy	1223 Lake Ave., Chicago, IL	31	\$7800	1	3	...
...

item_ID	name	brand	category	type	price	place_made	supplier	cost
I3	hi-res-TV	Toshiba	high resolution	TV	\$988.00	Japan	NikoX	\$600.00
I8	Laptop	Dell	laptop	computer	\$1369.00	USA	Dell	\$983.00
...

empl_ID	name	category	group	salary	commission
E55	Jones, Jane	home entertainment	manager	\$118,000	2%
...

branch_ID	name	address
B1	City Square	396 Michigan Ave., Chicago, IL
...

purchases

trans_ID	cust_ID	empl_ID	date	time	method_paid	amount
T100	C1	E55	03/21/2005	15:45	Visa	\$1357.00
...

items_sold

trans_ID	item_ID	qty
T100	I3	1
T100	I8	2
...

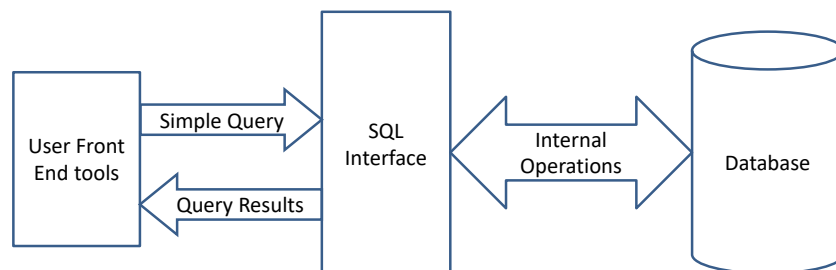
works_at

empl_ID	branch_ID
E55	B1
...	...

Purpose of Relational Database

- The main purpose of a relational database is to store data **correctly** and retrieve data **on demand**.
- This type of data processing is sometime called Online Transaction Processing (OLTP).
- Relational databases are **passive data repositories** in the sense that a query only shows you what is stored in the database, but cannot tell you much about the **meaning or trend** of the data.

Query Answering in Relational Database

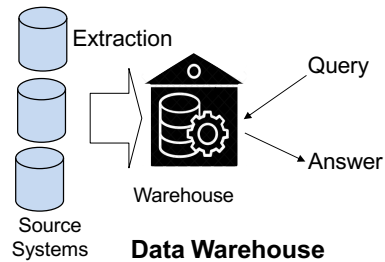


Transactional Database

- A transactional database consists of a file where each record represents a transaction.
- Supports nested relation
- Transaction id: Items, customer name, date...
- Sample Queries:
 - Show me all the items purchased by 'X'
 - How many transactions include item number 'Y'?
 - market basket data analysis: Which items sold well together? (**Frequent item set**)

Storing Data in Data Warehouses

- Data are from multiple data sources
 - Relational DB systems, flat files, .csv files, ...
- Data are integrated into a data warehouse
- Data are stored in DBMS using tables.
- Query answering process is more complex.



25

Examples of More Complex Queries

- Show a list of all items that were sold in the last quarter
- Show the total sales of the last month, grouped by branch
- Which sales person has the highest amount of sales?
- How many sales transactions occurred in September?

26

Outline of Today's Lecture

- Unit Structure and Overview
- A Brief History of Data(base) Technologies
- Storing Data in Databases and Data Warehouses
- What is Data Warehouse
- OLTP vs. OLAP
- Data Mining

27

What is Data Warehouse?

- A data warehouse is a
 - subject-oriented,
 - integrated,
 - time-variant, and
 - nonvolatilecollection of data in support of management's decision-making process.

28

Data Warehouse – Subject Oriented



- Organised around major subjects, such as **customer, supplier, product, sales, time**.
- 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 making process**.

29

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

30

Data Warehouse – Integrated (cont.)



Data Integration is hard.

- Data warehouses combine data from multiple sources
- Data must be translated into a consistent format
- **Data integration represents ~80% of effort for a typical data warehouse project!**
- Some reasons why it's hard:
 - Metadata is poor or non-existent
 - Data quality is often bad
 - Missing or default values
 - Multiple spellings of the same thing (UWA vs. Uni. of WA. vs. University of Western Australia)
 - Inconsistent semantics
 - Marks across different universities?

31

Data Warehouse – Time Variant



- The time horizon for the data warehouse is **significantly longer than that of operational DB 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

32

- 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 :
 - *initial loading of data* and *access of data*.

- Unit Structure and Overview
- A Brief History of Data(base) Technologies
- Storing Data in Databases and Data Warehouses
- What is Data Warehouse
- **OLTP vs. OLAP**
- Data Mining

Key operations of Data Warehouses

Key operations of RDBMS

Data Warehouse (OLAP) vs. Operational DBMS (OLTP)

- OLTP (on-line transaction processing) and OLAP (on-line analytical processing) are two types of data processing systems.
- OLAP (on-line analytical processing) uses data to gain valuable insights, while the other is purely operational.

Data Warehouse (OLAP) vs. Operational DBMS (OLTP)

- OLTP (on-line transaction processing)
 - Major task of traditional/operational 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
 - Can organise and present data in various forms and combinations

Data Warehouse (OLAP) vs. Operational DBMS (OLTP)

On-Line Transaction Processing (OLTP)

- Many short transactions (queries + updates)
- Examples:
 - Update account balance
 - Enroll in course
 - Add book to shopping cart
- Queries touch small amounts of data (e.g. a few records)
- Updates are frequent
- Concurrency is biggest performance concern

On-Line Analytical Processing (OLAP)

- Long transactions, complex queries
- Examples:
 - Report total sales for each department in each month
 - Identify top-selling books
 - Count classes with < 10 students
- Queries touch large amounts of data
- Updates are infrequent
- Individual queries can require lots of resources

37

Comparison of OLTP and OLAP

Table 4.1 Comparison of OLTP and OLAP Systems

Feature	OLTP	OLAP
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
User	clerk, DBA, database professional	knowledge worker (e.g., manager, executive, analyst)
Function	day-to-day operations	long-term informational requirements decision support
DB design	ER-based, application-oriented	star/snowflake, subject-oriented
Data	current, guaranteed up-to-date	historic, accuracy maintained over time
Summarization	primitive, highly detailed	summarized, consolidated
View	detailed, flat relational	summarized, multidimensional
Unit of work	short, simple transaction	complex query
Access	read/write	mostly read
Focus	data in	information out
Operations	index/hash on primary key	lots of scans
Number of records accessed	tens	millions
Number of users	thousands	hundreds
DB size	GB to high-order GB	≥ TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

Note: Table is partially based on Chaudhuri and Dayal [CD97].

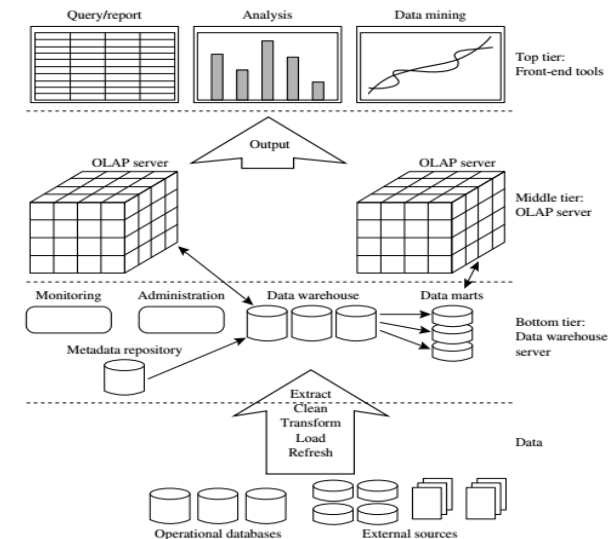
38

Building a Data Warehouse along with OLTP Systems

- Solution: Build a “data warehouse”
 - Copy data from various OLTP systems
 - Optimise data organisation, system tuning for OLAP
 - Transactions aren’t slowed by big analysis queries
 - Periodically refresh the data in the warehouse
- High performance for both systems
 - DBMS— tuned for OLTP: access methods, indexing, concurrency control, recovery
 - Warehouse—tuned for OLAP: complex OLAP queries, multidimensional view, consolidation.

39

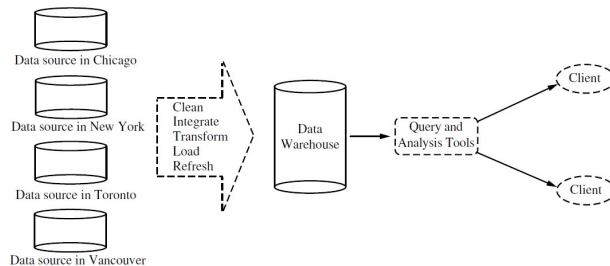
A three-tier data warehousing architecture



40

Data Warehouse of AllElectronics

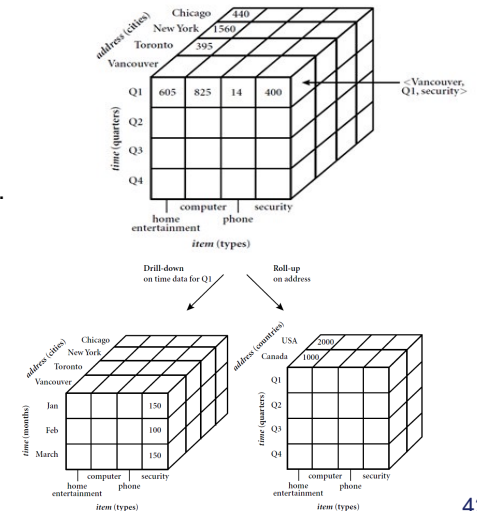
- A data warehouse is a **repository** of information collected from multiple sources, stored under a **unified schema**, and that usually resides at a single site.
- The need is to provide an analysis of the company's sales per item type per branch for a specified period.



40

Data Warehouse

- The data warehouse may store a summary of the transactions per item type for each store or, summarised to a higher level, for each sales region.



42

3 kinds of data warehouse applications

- Information processing**
 - supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts and graphs
- Analytical processing**
 - multidimensional analysis of data warehouse data
 - supports basic OLAP operations, slice-dice, drilling, pivoting
- Data mining**
 - knowledge discovery from hidden patterns
 - supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualisation tools.

43

Outline of Today's Lecture

- Unit Structure and Overview
- A Brief History of Data(base) Technologies
- Storing Data in Databases and Data Warehouses
- What is Data Warehouse
- OLTP vs. OLAP
- Data Mining**
 - An important application of Data Warehouses

44

Why Data Mining?

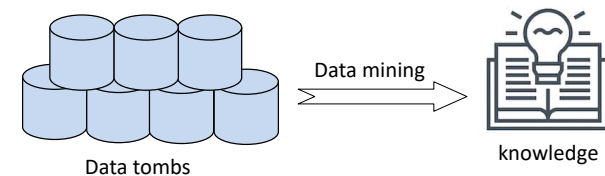
- The Explosive Growth of Data: from terabytes to petabytes
 - Data Explosion
 - Capability of generating, collecting, storing and managing data has grown tremendously in the last 50 years.
 - Data collection and data availability
 - Automated data collection tools, database systems, Web, computerised society
 - Major sources of abundant data
 - Business: Web, e-commerce, transactions, stocks, ...
 - Science: Remote sensing, bioinformatics, scientific simulation, ...
 - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- “Necessity is the mother of invention”—Data mining—**Automated** and **scalable** analysis of massive data sets

45

Why Data Mining?

Key points:

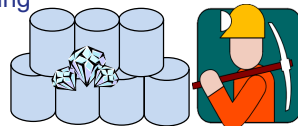
- Abundance of data and data archives are seldom visited.
- Far exceeded human ability for comprehension.
- Manual knowledge extraction is prone to biases and errors, and is extremely costly and time consuming.
- Data mining tools perform data analysis and uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research.



46

What is Data Mining?

- Data mining (knowledge discovery from data)
 - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
 - Data mining: **a misnomer?** (Knowledge Mining from data)
- Alternative names
 - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything “data mining”?
 - Simple search and query processing
 - (Deductive) expert systems

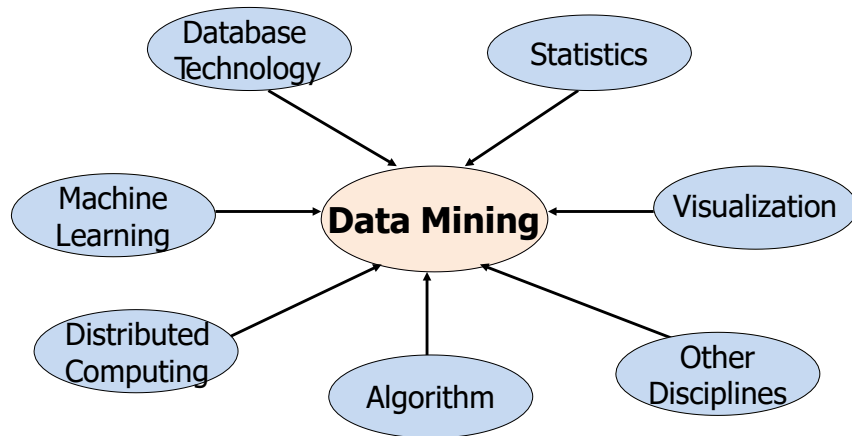


47

Data Mining in Different Data Sources

- Structured and semi-structured data
 - Relational database/ Object-relational data
 - Data Warehouse,
 - Transactional Database
- Unstructured data
 - Data streams and sensor data
 - Text data
 - Time-series data, temporal data, sequence data (incl. bio- sequences)
 - Graphs, social networks and information networks
 - Spatial data, spatiotemporal data and multimedia data

48

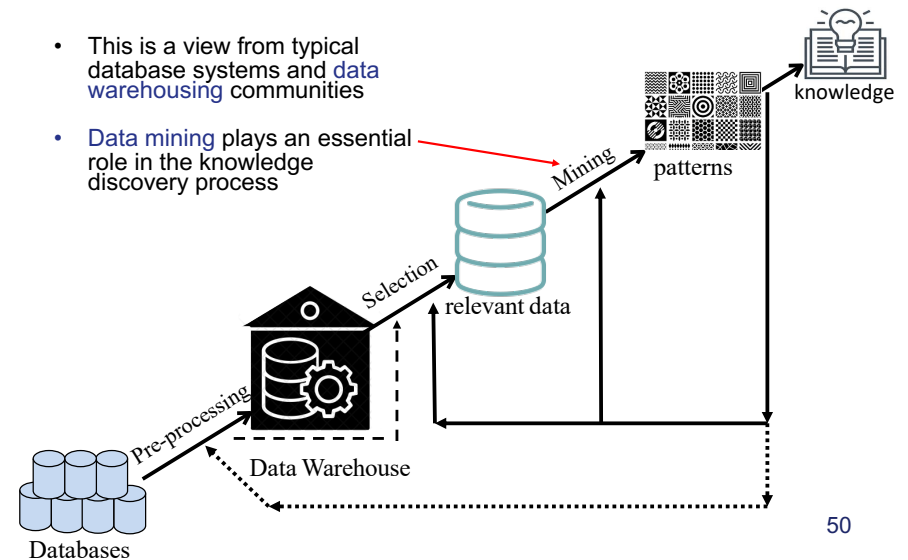


49

Steps of Knowledge Discovery from Data (KDD) Process

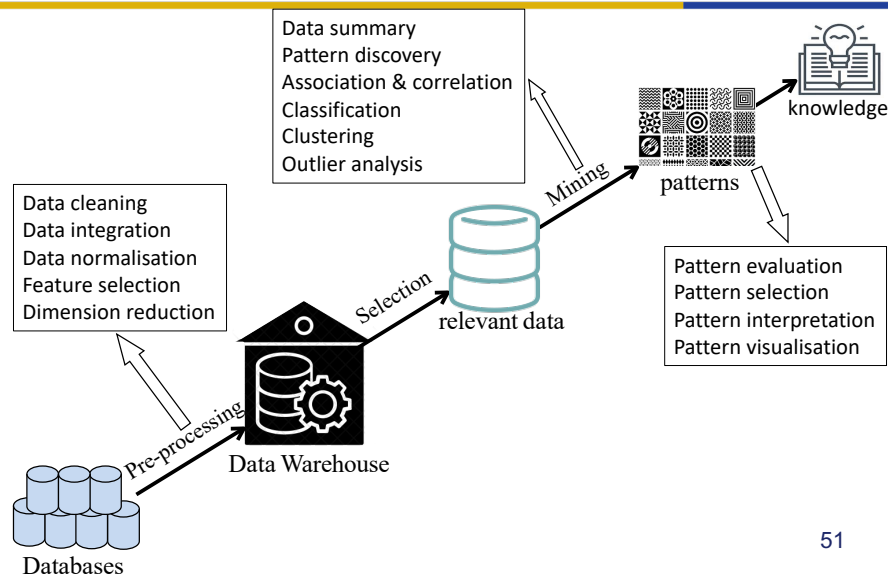
- This is a view from typical database systems and data warehousing communities

- Data mining plays an essential role in the knowledge discovery process



50

Techniques in the Process



51

References

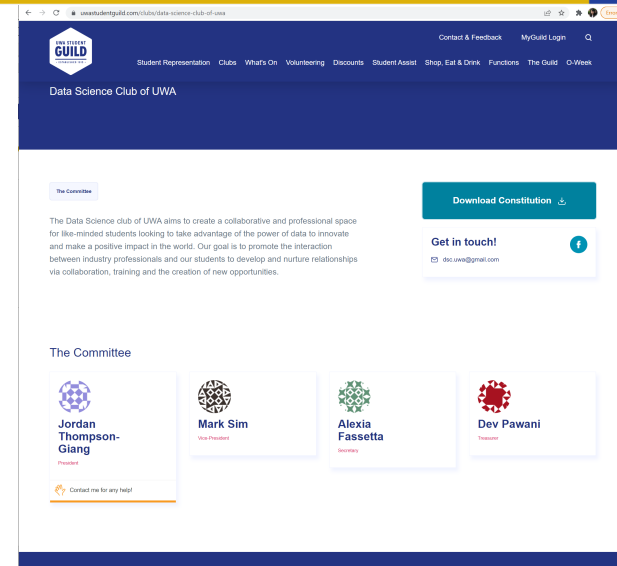
- Some slides are adapted from
 - <http://web.stanford.edu/class/cs345/>
 - https://hanj.cs.illinois.edu/bk3/bk3_slidesindex.htm
- Readings
 - Chapter 1.1, 1.2 and 4.1 of Han et al.'s book
 - [Evolution of Sciences \(page 4 to 8\).](#)
 - [Relational DB v.s. Transactional DB](#)

52



FOUNDING SPONSORS

We offer sponsorship packages for industry partners to engage in student scholarships and research collaborations.



DATA SCIENCE WITH DANIEL

BUILD. SHARE. LEARN.



We seek to build a community of Data Scientists, so that we can share our passion and learn together. We do this by bringing everything and everyone together in one place; Data Science with Daniel.

BUILD.



We create open source projects that anyone can contribute to, helping people to take the leap in starting to code in a collaborative environment.

SHARE.



We support anyone on their Data Science journey by providing an environment where they can ask questions, find answers and connect with others.

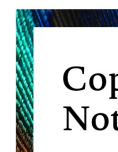
LEARN.



We help anyone to learn through \n, which is the newline character. This reflects the beginning of something new for those who are just getting started, a place where anything is possible.

Find us at:
www.datasciencewithdaniel.com.au

Copyright Notice



Copyright Notice

Material used in this recording may have been reproduced and communicated to you by or on behalf of **The University of Western Australia** in accordance with section 113P of the *Copyright Act 1968*.

Unless stated otherwise, all teaching and learning materials provided to you by the University are protected under the Copyright Act and is for your personal use only. This material must not be shared or distributed without the permission of the University and the copyright owner/s.

