

COMP37332  
Data Integration and Analysis

---

Goran Nenadic, Sandra Sampaio  
School of Computer Science

---

---

---


---

---

---

---

---



Context

- Previous database courses focused on:
  - **Database technologies:** infrastructure for managing and querying data.
  - **Database design:** techniques for working out what to store and how.
  - **Database programming:** developing applications over databases.
- This course unit focuses principally on making the most of data within an organisation:
  - **Data integration:** getting the data into a form that supports and facilitates aggregation, exploration and mining.
  - **Data analysis:** techniques for learning new lessons from the data.

---

---

---


---

---

---

---

---



Module contents

**Data integration**

- Distributed DBs
  - Concepts and issues
  - Top-Down and Bottom-Up Design
  - Data distribution in Oracle
- Data Warehousing
  - Modelling, design, architectures
  - ETL (extract, transform, load) process

---

---

---


---

---

---

---

---



## Module contents

---

### **Data analysis**

- On-line Analytical Processing (OLAP)
  - Exploration, analysis and integration of data
  - OLAP operations and SQL extensions
- Data mining
  - Role of data mining
  - Mining techniques (classification, clustering, association rules, etc)

---

---

---


---

---

---

---

---



## Staff

---

- Sandra Sampaio (ssampaio@cs.man.ac.uk)
  - Distributed databases
- Goran Nenadic (g.nenadic@manchester.ac.uk)
  - Data Warehousing
  - OLAP
  - Data mining

---

---

---


---

---

---

---

---



## Organisation

---

- Lectures and guest lectures
  - introducing main concepts
- Tutorials
  - understanding and practical work
- Labs
  - some topics will have practical labs (e.g. data analysis using software products)
- Materials
  - all materials are on Blackboard (lectures, tutorials, labs)

---

---

---

---

---

---

---

---



## Reading list

- Title: Principles of Distributed Database Systems  
Author: M. Ozsu and P. Valduriez  
ISBN: 0130412120  
Publisher: Prentice-Hal
- Title: Database systems: a Practical Approach to Design, Implementation, and Management  
Author: TM. Connolly, CE. Begg  
ISBN: 0130412120  
Publisher: Pearson Education Limited
- Title: Fundamentals of Database Systems  
Author: Elmasri, R., Navathe, S.  
Publisher: 5<sup>th</sup> Edition, Benjamin/Cummings
- Many online materials, including ORACLE documentation, WEKA, RapidMiner, tutorials etc.

---

---

---

---

---

---

---

---



## Assessment

- Exam: 85%
  - 2 hours, calculators allowed
  - 3 out of 5 questions
- Laboratory: 15%
  - two assessed labs
- Pre-requisites
  - Good knowledge of SQL
  - Basic of maths

---

---

---

---

---

---

---

---



---

---

---

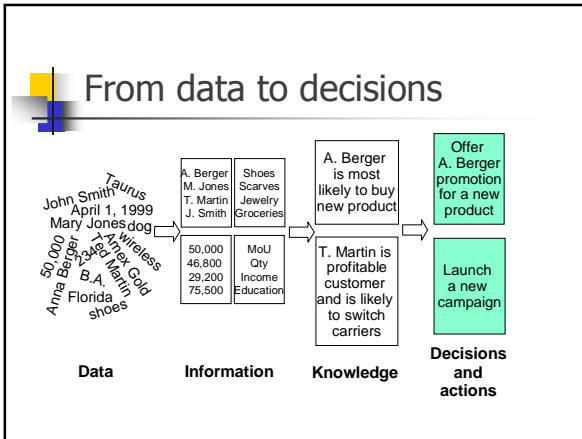
---

---

---

---

---




---

---

---

---

---

---

---

---

- ### Data/information management
- Store useful data and information
    - day-to-day operational data (e.g. transactions)
    - external data (e.g. market data)
    - human resources data
  - Provide effective information storage and access to support data analysis and integration
    - distributed/federated databases:
      - leave the data where it is
    - data warehouses
      - construct a new database for analysis purposes

---

---

---

---

---

---

---

---

- ### The kinds of data we have
- Traditional **"transactional"** information, i.e. operational data that documents everyday life in an enterprise/organisation
    - retail (e.g. sales in supermarket stores)
    - financial services (e.g. ATM withdrawals)
    - transport (e.g. flight bookings)
    - telecommunications (e.g. mobile billing, Internet)
    - healthcare (e.g. drug prescriptions)
  - Recording and processing this type of data is known as **"online transaction processing" (OLTP)**

---

---

---

---

---

---

---

---

## Online transaction processing

- OLTP: processing and recording transactions that create *new* data and/or update existing information in operational DBs:
  - insertions, updates, deletions.
- Typically a small number of rows are affected in each transaction.
- Traditional DBMS optimised to perform well in OLTP, but not in comprehensive exploration, aggregation and decision making.

---

---

---

---

---

---

---

---

## Why analyse data?

### Business viewpoint

- Lots of data is being collected and stored:
  - web data, e-commerce
  - purchases at department/grocery stores
  - bank/credit card transactions
- Competitive pressure is strong:
  - provide business intelligence and/or customised services



---

---

---

---

---

---

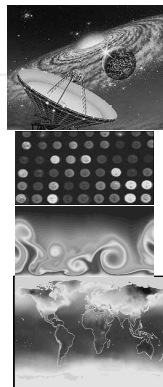
---

---

## Why analyse data?

### Scientific viewpoint

- Data collected and stored at enormous speeds (GB/hour):
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Data analysis may help scientists in
  - classifying and segmenting data
  - hypothesis formation



---

---

---


---

---

---

---

---



## Why analyse data?

- Business-related applications dominant (?)
  - until recently largest data providers (10-100 Tbytes)
  - retail and financial sectors in particular
- Comparisons
 

■ a novel	1 Mbytes
■ digital movie	10 Gbytes
■ scientific journals per year	1 Tbyte
■ Library of Congress	20 Tbytes
■ production of information	1500 Pbytes/year

---

---

---

---


---

---

---

---

continued



## Why analyse data?

- Scientific databases more and more important
  - biomedicine/bioinformatics/genetics
    - in the range of Pbytes per year (gene expressions)
  - astronomy
    - already in hundreds of Tbytes/year
  - environmental science
    - already in hundreds of Tbytes/year; predictions: 15 Pbytes
  - medicine and health care; electronic patient records
    - ~Pbytes (mostly images)
  - social sciences
    - census data, government data, social networks

---

---

---

---


---

---

---

---

continued



## Why analyse data?

- WWW as a data source
  - 150 mil web domains, ~12 billion pages (2006)
  - various data types (numerical, text, images, videos)
- Governmental and security applications
  - integration of several databases
    - personal information (shopping habits, bank accounts, travelling, web-logs, communications – phone, emails)
    - GPS (spatial data)
    - social networks

---

---

---

---

---

---

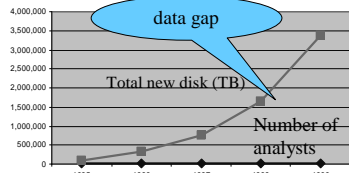
---

---

continued

## Why analyse data?

- There is often information "hidden" in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analysed at all



---

---

---

---

---

---

---

---

## Data analysis/mining examples

- Identify/profile customers that are most likely to respond to a new product/promotion.
- Identify trends/shifts in behaviour:
  - customers that buy video equipment, usually do not buy fitness equipment
  - within three months after buying a camera, customers are also likely to buy an accessory item
  - proteins with such expression array data that might be related to specific function or disease
- Identify *unusual* credit card activity that is likely to be fraudulent.

---

---

---

---

---

---

---

---

## Plan

- Week 1: Introduction to distributed DBs
- Week 2: Distributed DB
- Week 3: Distributed DBs + tutorial
- Week 4: Data warehousing
- Week 5: OLAP
- Week 6: Lab test 1 – DW + OLAP (9 March)  
Introduction to data mining
- Week 7: Association rule mining
- Week 8: Data classification

---

---

---

---

---

---

---

---

continued

## Plan

- Easter break
- Week 9: Data clustering
- Week 10: Guest lecture (IBM)
- Week 11: Lab test 2 – data mining (4 May)
- Week 12: Enterprise Resource Planning  
Revision

---

---

---

---

---

---

---

## Summary

- This course unit aims to introduce:
  - Architectures for integrating and organising data in a way that supports further analyses:
    - Distributed databases
    - Data warehouses
  - Techniques for learning new lessons from the data once integrated:
    - OLAP
    - Data mining

---

---

---

---

---

---

---

## Course Web page

Blackboard – COMP37332

<http://personalpages.manchester.ac.uk/staff/G.Nenadic/COMP37332>

---

---

---

---

---

---

---