

COMP37332

# Introduction to Data Warehousing

Goran Nenadic  
School of Computer Science

1

---

---

---

---

---

---

---

---

## Aims

- Understand the need for data warehousing
- Learn basic principles of data warehousing
- Understand data warehousing models and architectures

2

---

---

---

---

---

---

---

---

## Plan

- Data integration, analysis and business intelligence
- Data warehousing
  - definition
  - modelling
  - architectures
  - trends and open issues
- Summary
- **Lab 1**: 09 March 2010, 13-14 (3<sup>rd</sup> year Lab)

3

---

---

---

---

---

---

---

---

## Need for data analysis

- Modern business/science environment
  - markets evolve faster than ever
  - competition is more intense than ever
  - quantity of information is increasing
- In order to succeed, an organisation must
  - have a comprehensive view of all of its aspects
    - > data integration
  - make informed and reliable decisions
    - > data analysis
  - take timely actions and accurate predictions

4

---

---

---

---

---

---

---

---

## “Business intelligence”

- Business intelligence
  - an ongoing process of monitoring the competitive environment in order to identify opportunities to act on, and/or threats to business to be avoided.
- It is analytical analysis of available business data (internal and external)
- It is **NOT** about spying, sleuthing, espionage
  - it is estimated that 80% of business intelligence of an organisation is hidden inside its own business

5

---

---

---

---

---

---

---

---

## Business intelligence examples

- Customer data and patterns
  - What are the characteristics of our customers?*
  - What are their buying patterns?*
  - Who are the customers likely to move away?*
  - Who are the most loyal customers?*

6

---

---

---

---

---

---

---

---

continued

## Business intelligence examples

- Sales analysis and identification of trends
  - Which products sell the most at specific time periods?
  - What are the products that are selling best as combinations?
  - What are the products sold during the highest profitability transactions?
  - How many visas were issued country-by-country for the three most busy months in the last 12 months?

7

---

---

---

---

---

---

---

---

continued

## Business intelligence examples

- Business targets and promotion effectiveness
  - Who are the customers most likely to respond to an advertising campaign by post?
  - Which day of the week a new advertising campaign should be launched?
  - How promotional campaigns are linked with other leading brands over time?
  - How a certain campaign has affected the sales in a region?
  - How were visa applications affected by implementing a new on-line access system?
- etc.

8

---

---

---

---

---

---

---

---

recall from the Intro lecture

## Data/information management

- Store and integrate all useful data
  - day-to-day operational data (e.g. transactions)
  - external data (e.g. market data)
  - human resources data
- Provide effective information storage for integrated data, and easy information access to support data analysis

9

---

---

---

---

---

---

---

---

recall from the Intro lecture

## 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)

10

---

---

---

---

---

---

---

---

recall from the Intro lecture

## 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

11

---

---

---

---

---

---

---

---

recall from the Intro lecture

12

---

---

---

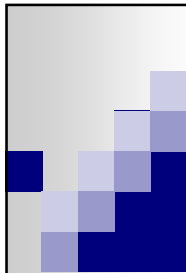
---

---

---

---

---



# Data warehousing

13

---

---

---

---

---

---

---

---

## Data warehouse (DW)

- DW: an integrated database designed to support data analysis, business intelligence, and better and faster decision making
- DWs integrate and aggregate data from various operational and external DBs maintained by different units
- DW needs to provide
  - more complex aggregation and analysis of data
  - mining "new" data (e.g. spending trends)

14

---

---

---

---

---

---

---

---

## Data warehouse - definition

*A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of the management's decision-making process.*

*Bill Inmon  
(~1990)*

15

---

---

---

---

---

---

---

---

### Data warehouse - definition

- **Subject-oriented:** the warehouse is organised around the major subject(s) of the enterprise.
- **Integrated:** it merges various source data from different applications and systems.
- **Time-variant** - data in the warehouse is only accurate and valid at some point in time or over some time interval.
- **Non-volatile** - data is not updated in real time but is refreshed from operational systems on a regular basis.

---

---

---

---

---

---

---

---

### Data warehouse characteristics

- DW typically integrates several resources
  - e.g. sales DBs from various regions/states/years
- Requires more historical data than generally maintained in operational DBs
- Must be optimised for access to very large amounts of data
- Mostly read-accessed, rarely write-accessed
- DW data may be more coarse grained than in operational DBs;
  - also de-personalised

---

---

---

---

---

---

---

---

### Data warehouse characteristics

- DWs are maintained *separately* from operational data
  - functional reasons
    - "historical" data (e.g. sales over long periods of time)
    - consolidated data (aggregated, summarised from various sources, including both internal and external)
  - performance reasons
    - avoid degradation of services (operational DBs are not optimised for "non-transactional" applications)
    - DW updates may not be so frequent

---

---

---

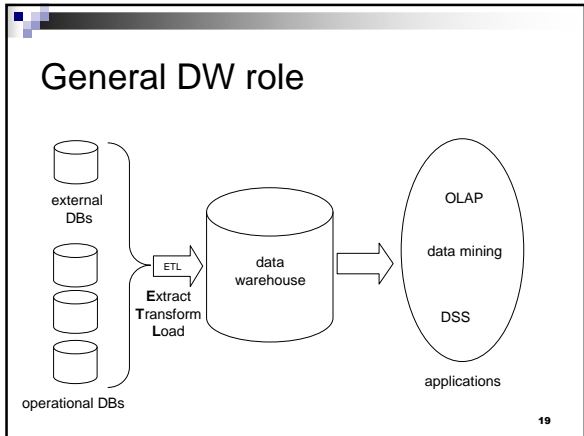
---

---

---

---

---




---

---

---

---

---

---

---

---

- ### Data warehouse applications
- Online analytical processing (OLAP)
    - complex analysis of data from DW
    - e.g. trend analysis, time series, etc.
  - Decision support systems (DSS)
    - high level data processing for management
    - executive information systems (EIS)
  - Data mining (DM)
    - support for "knowledge discovery"
    - search for unanticipated knowledge
- 20

---

---

---

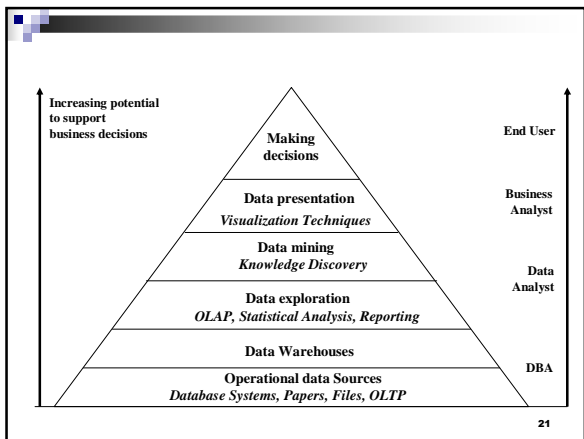
---

---

---

---

---




---

---

---

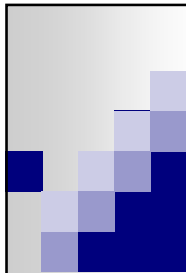
---

---

---

---

---



## Data warehouse modelling and design

22

---

---

---

---

---

---

---

---

### DW modelling and design

- Modelling questions
  - What are the user requirements (types of analysis needed)?
  - Which data should be considered (measures)?
  - Which data is available (OLTP, integration)?
  - What are the data inter-dependences (integration)?
  - What is the scale of the project?
- Define data model and design the schema
  - traditional ER (entity-relationship) modelling techniques may not be appropriate
  - multidimensional model

23

---

---

---

---

---

---

---

---

### Multidimensional data model

- Popular data model for DW
- Dimensionality modelling
  - logical design technique that aims to present the data in a standard, intuitive form that allows for high-performance access.
- Uses the concepts of entity-relationship (ER) modelling with some restrictions
- Focused on **key performance indicators** that need to be analysed

24

---

---

---

---

---

---

---

---

## Multidimensional data model

- Objects of analysis (*key performance indicators*) are numeric **measures**
  - e.g. sales, budget, number of applications, ...
- DW: a set of points in a multi-dimensional space
  - e.g. (store, product, amount, time)
  - time is typically one of the dimensions, as it is of particular significance to decision support
- *Dimensions* describe characteristics of the measures, i.e. their "context"

---

---

---

---

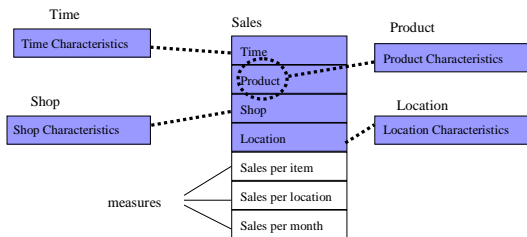
---

---

---

---

## Example



---

---

---

---

---

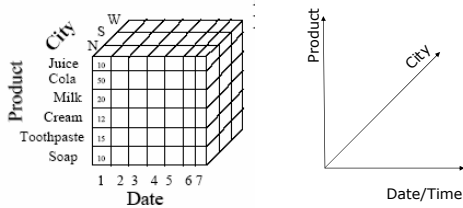
---

---

---

## Multidimensional data example (1)

- Dimensions: *Product, City, Date*
- Hyper-cube representation



---

---

---

---

---

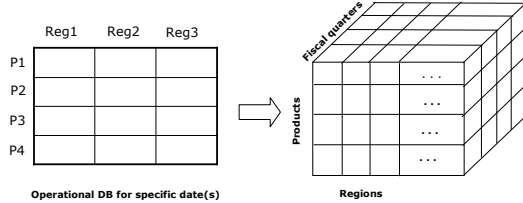
---

---

---

## Multidimensional data example (2)

- Operational database: *Product, Region, Date*
- DW dimensions: *Product, Region, Fiscal quarter*



28

---

---

---

---

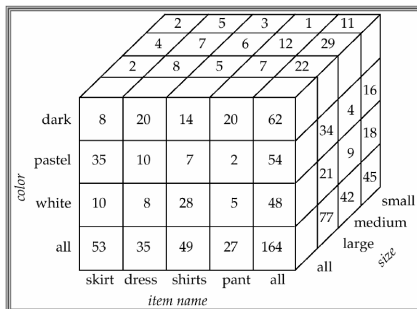
---

---

---

---

## Example – cube



29

---

---

---

---

---

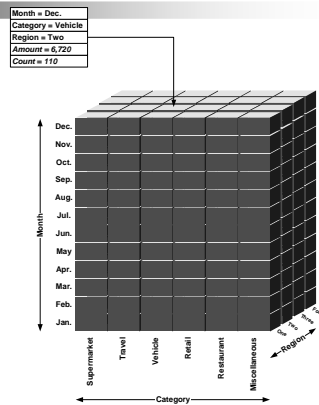
---

---

---

## Example

multidimensional cube for credit card purchases



30

---

---

---

---

---

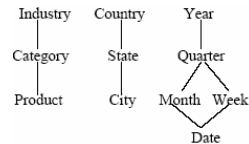
---

---

---

## Attributes

- Each dimension has a set of attributes
  - e.g. *product*: category, year, manufacturer
  - e.g. *store*: town, county, manager, size
- Attributes can be related
  - hierarchically
  - lattice



31

---

---

---

---

---

---

---

---

## DW data models

For multidimensional DB implementation, we use two types of tables:

- **dimension tables**
  - tuples of attributes of a dimension
  - one table for each of the dimensions
- **fact table**
  - contains the data – “facts”
  - typically only one fact table

32

---

---

---

---

---

---

---

---

continued

## DW data models

Example:

- dimension table for *product*:  
product ID, name, manufacturer, type, etc.
- dimension table for *fiscal quarter*:  
quarter ID, quarter number, year, BEG/END dates
- fact table  
(*product\_ID, quarter\_ID, region\_ID, amount*)

33

---

---

---

---

---

---

---

---

## DW data models

- Each dimension table has a simple primary key  
*product\_ID* or *quarter\_ID*
- Fact table uses a composite primary key
  - the primary key of the fact table is made up of two or more foreign keys linking to dimension tables.
- Such simple model structure is called a **star schema**

---

---

---

---

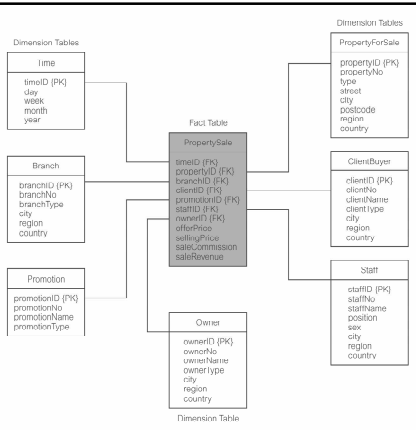
---

---

---

---

## Star schema example



---

---

---

---

---

---

---

---

## Data model schemas

- **Star schema:** a logical structure that has a fact table containing *factual data* in the centre, surrounded by dimension tables (for each dimension) containing *reference data*.
- **Snowflake schema:** a variant of the star schema where dimension tables can further have dimension tables, a kind of fractal structure.
- **Starflake schema:** a hybrid structure that contains a mixture of star and snowflake schemas.

---

---

---

---

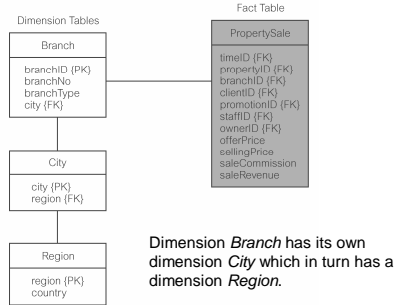
---

---

---

---

## Snowflake schema example



Dimension *Branch* has its own dimension *City* which in turn has a dimension *Region*.

Dimension tables are organised in a hierarchy by normalisation

37

---

---

---

---

---

---

---

---

---

---

---

---

## Data warehouse architectures

38

---

---

---

---

---

---

---

---

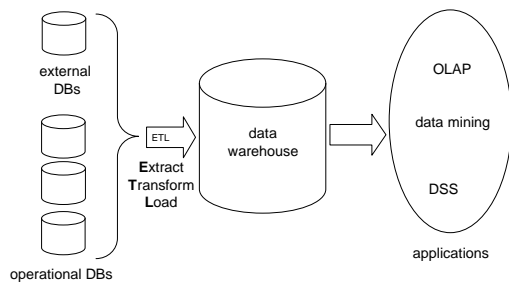
---

---

---

---

## General DW schema



39

---

---

---

---

---

---

---

---

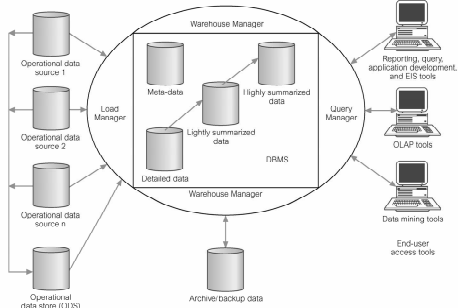
---

---

---

---

## Data warehouse architecture




---



---



---



---



---



---

## Architectural components - data

- **Operational data**
  - source data (operational DBs) for the DW
- **Operational data-store**
  - a repository of operational data
- **Meta-data**
  - description of data (data about data)
  - used for acquisition of data, DW management, and querying

41

---



---



---



---



---



---

continued

## Architectural components - data

- **Detailed data**
  - aggregation of other data
- **Lightly and highly summarised data**
  - generated by the warehouse manager
  - e.g. "summary" tables that hold aggregations such as sums, averages, counts of values in other tables
  - other derived data ("calculated columns")
- **Archive/backup data**

42

---



---



---



---



---



---

## Architectural components - tools

- **Load manager**
  - extraction, acquisition and loading of data into the DW
- **Warehouse manager**
  - management of the DW, e.g. indexing, backup
- **Query manager**
  - management of user queries
- **End-user access tools**
  - reporting, querying, application development tools
  - executive information system (EIS) tools
  - online analytical processing (OLAP) tools
  - data mining tools

43

---

---

---

---

---

---

---

---

## Load manager tasks

- **Data extraction**
  - acquisition of data from operational DBs
  - design changes in operational DBs require adjustments
- **Data cleansing**
  - detect data anomalies and rectify them
- **Data transformation/reformatting**
  - transforming and linking data
- **Loading**
  - loading, building indexes, etc.
- **Refreshing**
  - propagate updates from operational DBs to DW

44

---

---

---

---

---

---

---

---

## Data transformation/reformatting

- **Link and consolidate data from different sources**
  - handle possibly different DBs schemes
  - parsing and fuzzy matching may be needed
  - avoid un-necessary redundancy
- **Reformatting/mapping as appropriate**
  - town → region → state*
  - date → month → fiscal quarter → year*
- **Reconcile semantic mismatches**
  - transform into uniform measures (e.g. currency)
  - e.g. "gender" and "sex" should be the same

Recall: Manchester/UMIST examples (DDB lectures)

45

---

---

---

---

---

---

---

---

## Loading data

- “Materialise” prepared data and store it in DW
- Large volumes of data to be loaded
  - may take several days to load
  - can be performed in parallel
- Building
  - summary tables (various pre-calculated aggregations)
  - indexes (supporting most-likely queries)
  - meta-data (various information about the data)

46

---

---

---

---

---

---

---

---

## Meta-data generation

- Meta-data: data about data
- Build a meta-data repository
- Document the following
  - source DBs
  - DW schema
  - loading process description (applied transformation rules, cleaning methods, etc.)
  - refreshing policy
  - security issues, user profiles/groups
  - statistics
  - etc.

47

---

---

---

---

---

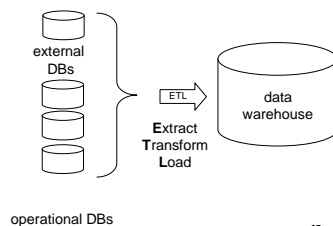
---

---

---

## ETL process

- Previous tasks also known as ETL
  - extract
  - transform
  - load



48

---

---

---

---

---

---

---

---

## Refreshing data in DW

- Refreshing intervals
  - e.g. every night, week, or quarterly
  - typically not so frequent (depending on the subject area; e.g. for stock DWs, up-the-minute data may be required)
- Possibly different refreshing periods for different operational DBs
- Approaches
  - full extraction and loading (expensive, not efficient)
  - incremental
    - detect and propagate only changes in operational DBs
    - logical and transactional correctness and integrity
    - purging out-of-date data

49

---

---

---

---

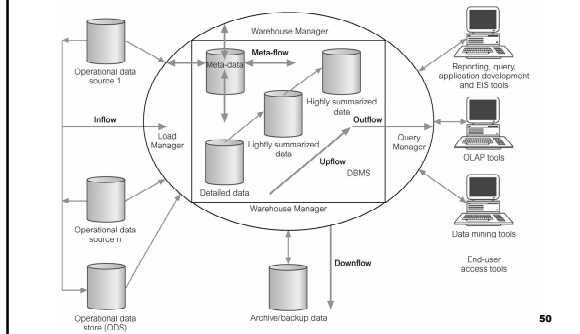
---

---

---

---

## Data warehouse data flows



50

---

---

---

---

---

---

---

---

continued

## Data warehouse data flows

- *In-flow*
  - extraction, cleansing and loading of the source data
- *Up-flow*
  - adding value to the data in the warehouse through summarising, packaging and distribution of the data
- *Down-flow*
  - archiving and backing-up data in the warehouse
- *Out-flow*
  - making the data available to end-users
- *Meta-flow*
  - managing the meta-data

51

---

---

---

---

---

---

---

---

## Data marts

- DW is generally meant to provide a single source of data
- Data marts: special-purpose warehouses
- A subset of a data warehouse that supports the requirements of a **particular department or business function**.
- Can be standalone or linked centrally to the corporate data warehouse

52

---

---

---

---

---

---

---

---

continued

## Data marts

- Do not normally contain detailed operational data, unlike data warehouse
  - granularity issue
- Good to control/allow access to data
- Sometimes used for building a DW
  - loading and consolidation from data marts

53

---

---

---

---

---

---

---

---

## DW architecture types

- Central warehouse
- Distributed warehouse
  - needs replication, partitioning, communication
  - replicated metadata repository on each site
  - (recall Distributed databases lectures)
- Federated warehouse
  - decentralised autonomous warehouses (e.g. containing data marts)
- Data web-house
  - a distributed data warehouse that is implemented over the Web with no central data repository

54

---

---

---

---

---

---

---

---



## Data warehousing: trends and open issues

55

---

---

---

---

---

---

---

---

### Typical problems

- Data homogenisation
- Complexity of integration
- Hidden problems with source systems
- Required data not captured
- Data ownership/access/privacy policies
  
- Underestimation of resources for data loading
- High demand for resources
- Increased end-user demands
- High maintenance costs

56

---

---

---

---

---

---

---

---

### Challenges

- Design and management of the DW project
  - managing design, construction, implementation
  - long-term and large-scale project
  - monitoring utilisation patterns and design solutions
- Administration of DW
  - an intensive, major activity
  - includes possible evolutions of data models
- Quality of data
  - merging data from heterogenous sources
  - integration issues (different namings, etc.)

57

---

---

---

---

---

---

---

---

## Open issues

- Automating acquisition and loading of operational data
- Self-maintainability
- Intelligent meta-data extraction/generation
- Performance optimisation
  - quite large DBs: many terabytes, growing by as much as 50% a year
- Privacy and security
  - depersonalisation of data, user authentication and authorization, role-based access control
  - see [Elmasri & Navathe; sec. 23.3.2]

58

---

---

---

---

---

---

---

---

## Examples of DW systems

- **IBM DB2 Universal Data Warehouse Edition**
  - a business intelligence platform that includes DB2, federated data access, data partitioning, enhanced extract, transform, and load (ETL), workload management etc.
  - see a case study on EDEKA
- **Microsoft SQL Server 2005 Data Transformation Services**
  - ETL data from source(s) to the data ware house
  - see a case study on Barnes & Noble
- **Oracle Business Intelligence Warehouse Builder**
  - dimensional design, ETL, deployment Manager
  - see tutorial/demo/documents on the web

59

---

---

---

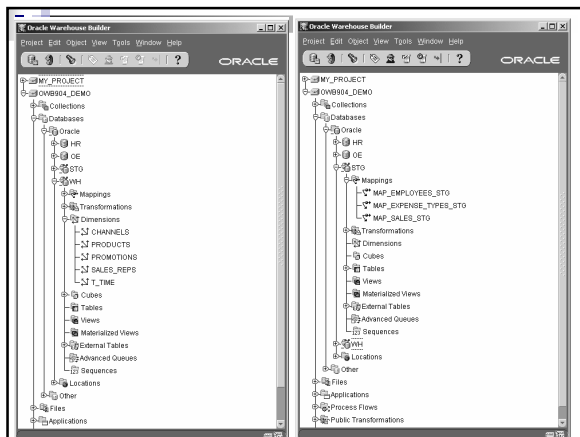
---

---

---

---

---



---

---

---

---

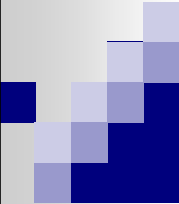
---

---

---

---





## Data warehousing: wrapping up

64

---

---

---

---

---

---

---

---

### Data warehouse design steps

1. Choose relevant business processes
2. Choose the grain (granulation) of DW
3. Identify the dimensions and model the schema
4. Choose, extract, transform and load the facts
5. Store pre-calculations in the fact table
6. Round out the dimension tables
7. Choose the duration/refresh of the database
8. Track slowly changing dimensions
9. Decide the query priorities and modes

65

---

---

---

---

---

---

---

---

### DW: expected benefits

- **Competitive advantage**
  - increased productivity of decision making
  - allow decision-makers to access data that reveal unknown, unavailable and untapped information (e.g. customer profiles, trends and demands)
- **Potential high returns on investment**
  - average return on business intelligence investments (hardware, data, tools, human resources) is over 400% over a period of 2-3 years (International Data Corporation data)
  - ◆ market growth: 1995: \$2 billion; 2004: \$10-15 billion

66

---

---

---

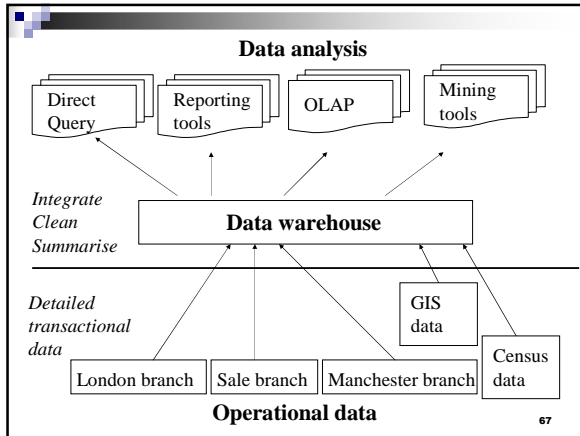
---

---

---

---

---




---

---

---

---

---

---

---

---

- ### Summary
- DW: an integrated database for supporting higher-level analysis of data
    - analysis using multi-dimensional and pre-aggregated data
    - data model: multidimensional
  - Integrates and aggregates operational DBs
  - Access to huge amount of data
    - optimised querying, indexing is needed
  - Huge benefits
    - supports competitive business intelligence
    - also, scientific data warehouses

---

---

---

---

---

---

---

---

- ### Reading for this lecture
- Chapters 31 and 32 in [Connolly & Begg]
  - Also, chapter 28 (28.1–28.4, 28.7) in [Elmasri & Navathe]
  - On-line materials: Blackboard and <http://personalpages.manchester.ac.uk/staff/G.Nenadic/COMP37332/>
- Additional reading (on the web)
- Chaudhuri, Dayal: *An overview of data warehousing and OLAP technology* (<http://portal.acm.org/citation.cfm?id=248616>)
  - W.H. Inmon: *Building the data warehouse*, Wiley, 2002 ISBN: 0471081302
- **Solve the relevant tutorial questions and prepare for the lab**
  - **Try Palo OLAP and Dundas OLAP services.**

---

---

---

---

---

---

---

---