Data Warehouse

Dr Na Yao

Objectives

- Understand and be able to explain main concepts and benefits associated with data warehousing.
- Understand and be able to explain difference between Online Transaction Processing (OLTP) and data warehouse.
- Understand and be able to explain OnLine Analytical Processing (OLAP).

The Emergence of Data Warehousing

- Problems:
  - Accumulation of growing amounts of data in operational databases.
  - Organizations wanted to use data to support decision-making.
  - Operational systems were never designed to support such business activities.

- Solution:
  - Data warehouse (DW): to meet the requirements of a system capable of supporting decision-making, receiving data from multiple operational data sources.

Data Warehousing Concepts

- A *subject-oriented, integrated, time-variant,* and *non-volatile* collection of data in support of management's decision-making process (Inmon, 1993).

How could you organize data???

- Depending on how you obtain it

Application-oriented

- Depending on what do you want to use it for

Subject-oriented

Subject-oriented Data

How is the warehouse organized?

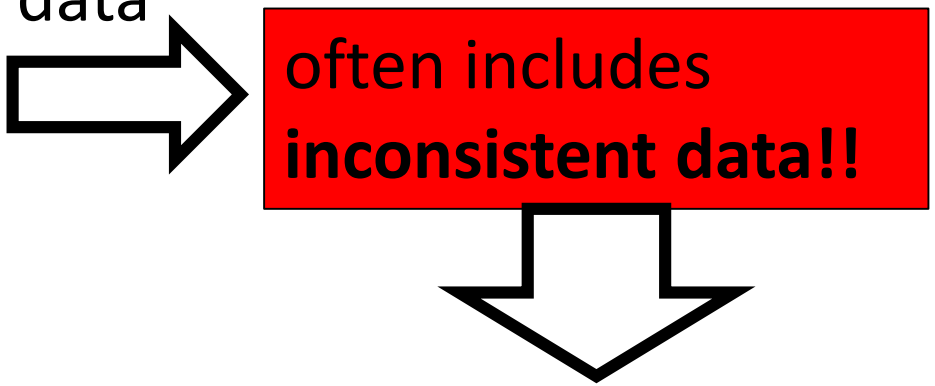
- Around the major **subjects** of the enterprise (e.g. customers, products, and sales) rather than the major **application** areas (e.g. customer invoicing, stock control, and product sales).

Why?

- Reflects the need to store **decision-support** data rather than **application-oriented data**.

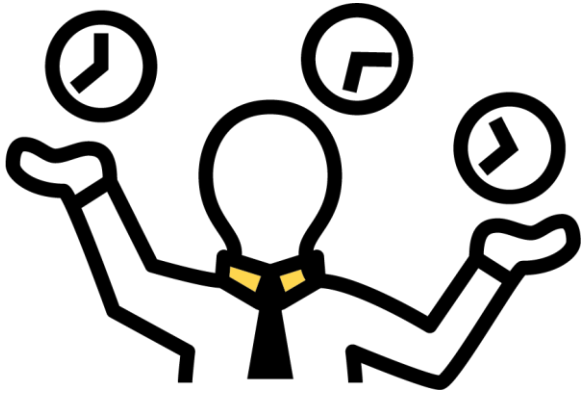
Integrated Data

- The data warehouse integrates
  - corporate application-oriented data
  - from different source systems



- The integrated data source **must be made consistent** to present a **unified view** of the data to the users.

Time-variant Data



- Data in the warehouse is **only accurate and valid at some point** in time or over some time interval.

- Time-variance is also shown in the extended time that the data is held, the implicit or explicit association of time with all data, and the fact that the data represents a series of snapshots.

Non-volatile Data

- Data in the warehouse is **not normally updated in real-time (RT)** but is **refreshed from operational systems on a regular basis**. (However, emerging trend is towards RT or near RT DWs)

- New data is always added as a **supplement** to the database, rather than a **replacement**.

Benefits of Data Warehousing

- Potential high returns on investment
- Competitive advantage
- Increased productivity of corporate decision-makers

Data Warehouse Queries (some examples)

- What was the total revenue for Scotland in the third quarter of 2001?
- What was the total revenue for property sales for each type of property in the UK in 2018?
- What are the three most popular areas in each city for the renting of property in 2019 and how does this compare with the figures for the previous two years?
- What is the monthly revenue for property sales at each branch office, compared with rolling 12-monthly prior figures?
- Which type of property sells for more than the average selling price, for properties in the main cities of UK and how does this correlate to demographic data?

Data Warehouse Queries

- Queries for Data Wharehouse range from the relatively simple to the highly complex ones.
- Dependent on the type of end-user access tools used.

- End-user access tools include:
  - Traditional reporting and query
  - OLAP
  - Data mining

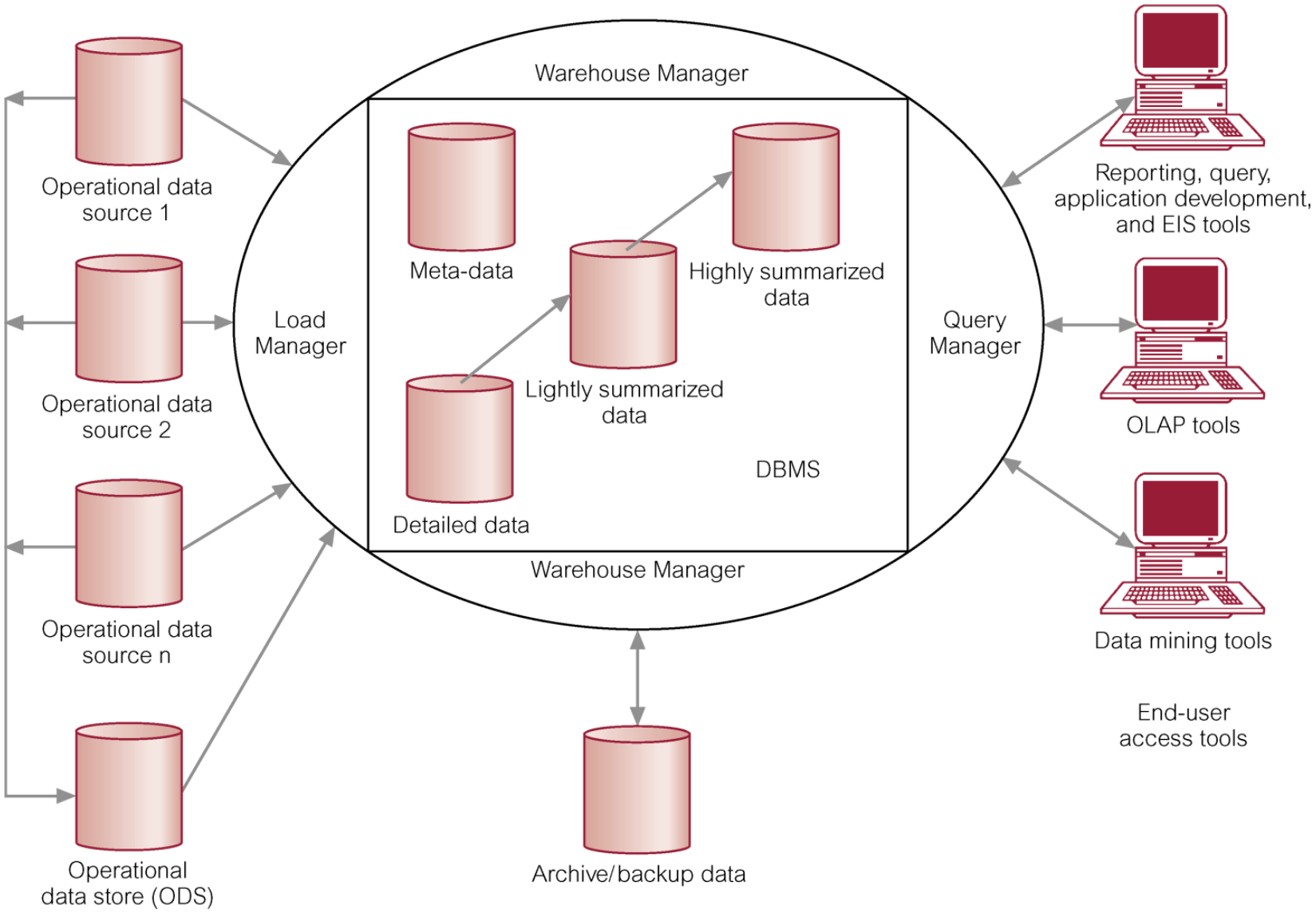
OLTP and OLAP

- OLTP: OnLine Transaction Processing
  - Transactional
  - provide source data to data warehouses

- OLAP: OnLine Analytical Processing
  - Analytical
  - Analysis of warehouse data

  (we will see more later)

Example Data Warehouse Architecture



DW Architecture components

- An *operational data store (ODS)* is a repository of current and integrated operational data used for analysis.
- The *load manager* performs all operations associated with the extraction and loading of data into the warehouse.
- The *warehouse manager* performs all the operations associated with the management of the data, such as the transformation and merging of source data; creation of indexes and views on base tables; and backing up and archiving data.
- The *query manager* performs all the operations associated with the management of user queries.
- *Metadata* (data about data) definitions are used by all the processes in the warehouse.

End-User Access Tools

- Main purpose of DW is to support decision makers and this is achieved through the provision of a range of access tools including:
  - reporting and querying,
  - application and development,
  - OLAP,
  - data mining.

Business Intelligence Technologies

- **Why?** Ever-increasing demand by users for more powerful access tools that provide advanced analytical capabilities.
- Main **types** of access tools:
  - Online Analytical Processing (OLAP)
  - Data mining.
- **What?** An environment that includes a data warehouse (or more commonly one or more data marts) together with tools such as OLAP and/or data mining → Business Intelligence (BI) technologies

Online Analytical Processing (OLAP)

- Original definition - The dynamic synthesis, analysis, and consolidation of large volumes of multi-dimensional data, Codd (1993).

- Describes a technology that is designed to optimize the storing and querying of large volumes of multi-dimensional data that is aggregated (summarized) to various levels of detail to support the analysis of this data.

Online Analytical Processing (OLAP)

- Can easily answer ‘who?’ and ‘what?’ questions, however, ability to answer ‘why?’ type questions distinguishes OLAP from general-purpose query tools.

- Types of analysis ranges from basic navigation and browsing (slicing and dicing) to calculations, to more complex analyses such as time series and complex modeling.

Examples of OLAP applications in various functional areas

Functional area	Examples of OLAP applications
Finance	Budgeting, activity-based costing, financial performance analysis, and financial modeling
Sales	Sales analysis and sales forecasting
Marketing	Market research analysis, sales forecasting, promotions analysis, customer analysis, and market/customer segmentation
Manufacturing	Production planning and defect analysis

OLAP Applications

- OLAP applications in very different functional areas
- Common key features:
  - multi-dimensional views of data
  - support for complex calculations
  - time intelligence

Multi-dimensional Data

- **Data:** facts (numeric measurements),

  example: property sales revenue data

- **Relationships** between data: the association of those facts with dimensions,

  example: location (of the property) and time (of the property sale).

How to represent multi-dimensional data

- relational table,
- matrix
- data cube

Multi-dimensional Data as 3-field Table versus 2-D Matrix

City	Time	Total Revenue
Glasgow	Q1	29726
Glasgow	Q2	30443
Glasgow	Q3	30582
Glasgow	Q4	31390
London	Q1	43555
London	Q2	48244
London	Q3	56222
London	Q4	45632
Aberdeen	Q1	53210
Aberdeen	Q2	34567
Aberdeen	Q3	45677
Aberdeen	Q4	50056
.....	.....	.....
.....	.....	.....

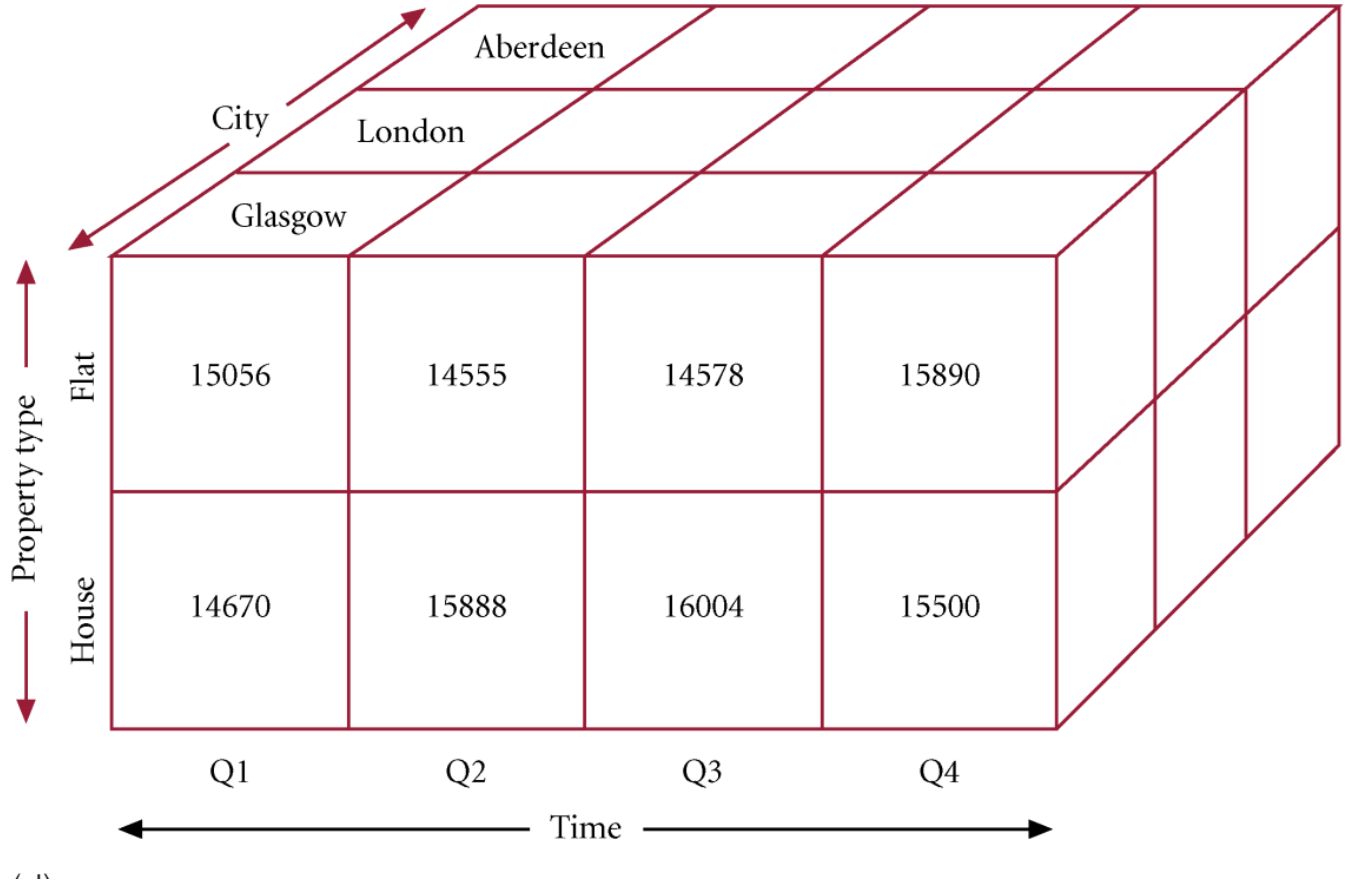
City

Time

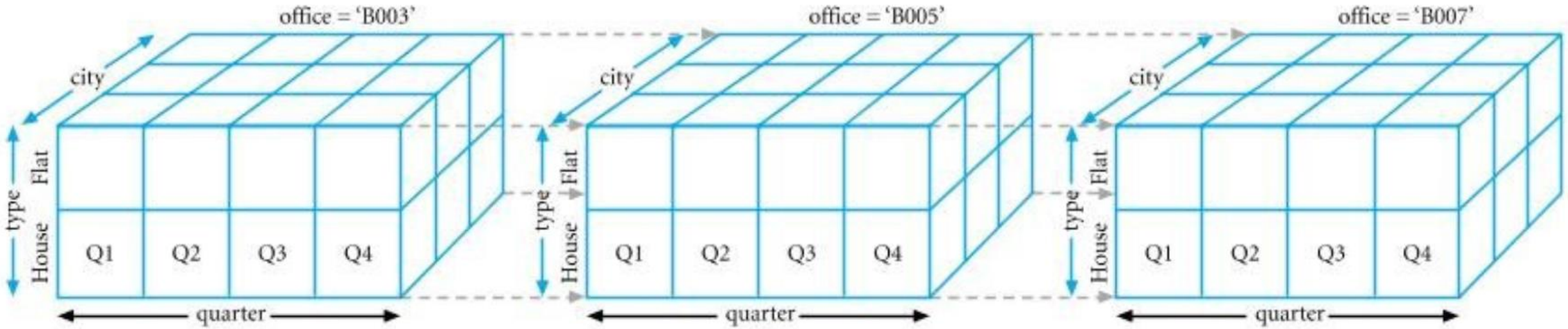
City / Quarter	Glasgow	London	Aberdeen	.....
Q1	29726	43555	53210	.....
Q2	30443	48244	34567	.....
Q3	30582	56222	45677	.....
Q4	31390	45632	50056	.....

Multi-dimensional Data as 4-field Table versus 3-D Cube

Property Type	City	Time	Total Revenue
Flat	Glasgow	Q1	15056
House	Glasgow	Q1	14670
Flat	Glasgow	Q2	14555
House	Glasgow	Q2	15888
Flat	Glasgow	Q3	14578
House	Glasgow	Q3	16004
Flat	Glasgow	Q4	15890
House	Glasgow	Q4	15500
Flat	London	Q1	19678
House	London	Q1	23877
Flat	London	Q2	19567
House	London	Q2	28677
.....	.....	.....	.....
.....	.....	.....	.....



Multi-dimensional Data as series of 3-D Cubes



Multi-dimensional data and OLAP cubes

- We consider cubes as solid 3-D structures with equal sides. However, the OLAP cube is n-dimensional structure (with sides that need not be equal).

Dimensional Hierarchy

- A *dimensional hierarchy* defines mappings from a set of lower-level concepts to higher level concepts.
- For example, for the sales revenue data, the lowest level for the location dimension is at the level of zipCode, which maps to area (of a city), which maps to city, which maps to region (of a country), which at the highest level maps to country.
- {zipCode --> area --> city --> region --> country}
- {day --> month --> quarter --> year}

Dimensional Operations

- The analytical operations that can be performed on data cubes include:
  - Roll-up
  - Drill-down
  - Slice and Dice
  - Pivot

Dimensional Operations

- *Roll-up* performs aggregations on the data by moving up the dimensional hierarchy or by dimensional reduction e.g. 4-D sales data to 3-D sales data.
- *Drill-down* is the reverse of roll-up and involves revealing the detailed data that forms the aggregated data.

Dimensional Operations

- *Slice and dice* - ability to look at data from different viewpoints. The slice operation performs a selection on one dimension of the data whereas dice uses two or more dimensions.
- E.g. a *slice* of sales revenue (type = ‘Flat’) and a *dice* (type = ‘Flat’ and time = ‘Q1’).

Dimensional Operations

- *Pivot* - ability to rotate the data to provide an alternative view of the same data e.g. sales revenue data displayed using the location (city) as x-axis against time (quarter) as the y-axis can be rotated so that time (quarter) is the x-axis against location (city) is the y-axis.

Comparison of OLTP Systems and Data Warehousing

Characteristic	OLTP Systems	Data Warehousing Systems
Main Purpose	Supports operational processing	Supports analytical processing
Data age	Current	Historic (but trend is towards also including current data)
Data latency	Real-time	Depends on length of cycle for data supplements to warehouse
Data granularity	Detailed data	Detailed data, lightly and highly summarised data
Data processing	Predictable pattern of data insertions, deletions, updates and queries. High level of transaction throughput.	Less predictable pattern of data queries. Medium to low level of transaction throughput.
Reporting	Predictable, one-dimensional, relatively static fixed reporting	Unpredictable, multi-dimensional, dynamic reporting
Users	Serves large number of operational users	Serves lower number of managerial users (but trend is towards also supporting analytical requirements of operational users)