Week 4 Summary

Dr Na Yao

XML

Extensible Markup Language (XML)

- A meta-language (a language for describing other languages) that enables designers to create their own customized tags to provide functionality not available with HTML.

- XML has a document format similar to HTML but…
  - Tags describe content instead of formatting

<Elements> and “attributes”

- First element must be a root element, which can contain other (sub)elements.
- Attributes are name-value pairs that contain descriptive information about an element.
- An element can have many attributes

Relational Model versus XML

	Relational	XML
Structure	Tables, columns, rows	Hierarchical, tree
Schema	Fixed in advance	Self-describing, flexible
Queries	SQL, simple language, standard	Not so simple
Ordering	None	Ordered
Implementation	Mature, native	Add-on

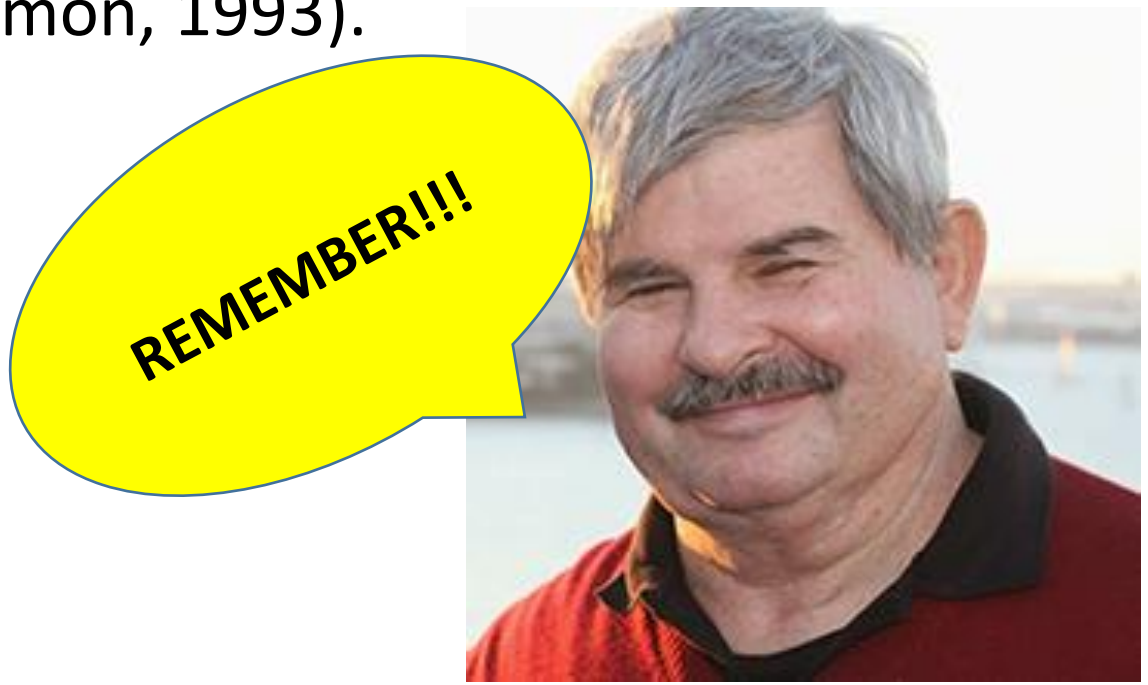
DTD, XSD…. Why we need them??

- XML is a **SEMI-STRUCTURED** data-model.
- The data structure of XML does not have to be strictly followed as in the relational model

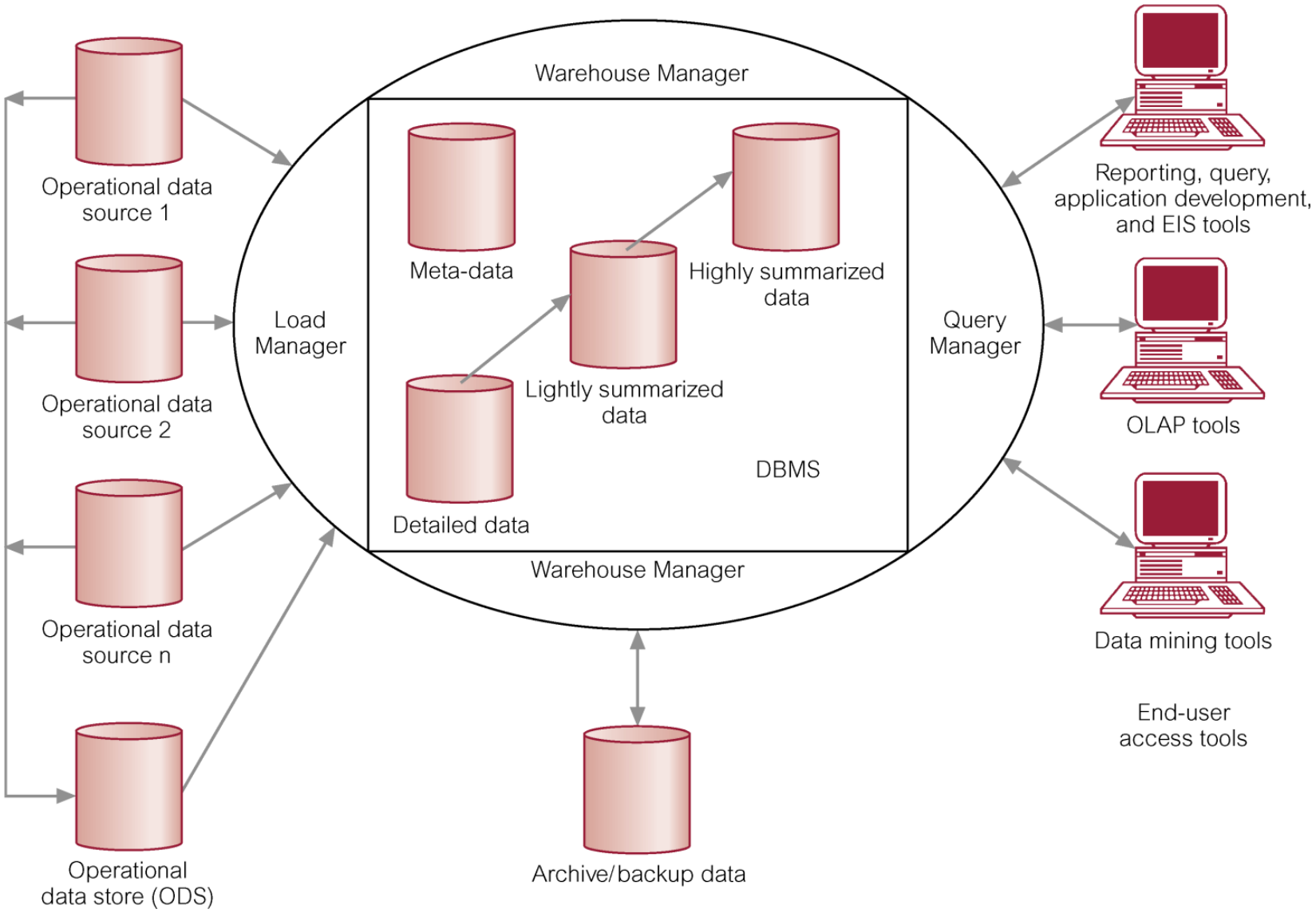
- DTD and XSD provides a schema (the valid syntax) for an XML document.

Data Warehousing

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



Example Data Warehouse Architecture



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)

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)

Data Mining, Definition

- The process of extracting **valid**, previously **unknown**, **comprehensible**, and **actionable** information from large databases and using it to make crucial **business decisions**.

- Involves the analysis of data and the use of software techniques for finding **hidden and unexpected** patterns and relationships in sets of data.

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

So is data mining = data ware house??

Data warehouse relates to the **storage** of the data used for data mining

NoSQL

alternative, non-traditional DB technology to be used in large scale environments where (ACID) transactions are not a priority

What NoSQL should NOT be used for

- Anything that requires frequent updates as well as reads, or that requires high integrity and atomicity (ACID properties)
- Examples are similar to transaction databases for inventory and financial records
- This is not just a question of massive data or distributed processing!
- There are large, distributed relational databases like Visa or Amazon that need more structured data with transaction semantics
- These applications are better suited to relational databases even at large scale