

# Databases Illuminated

## Chapter 13

### Databases and the Internet

# Databases and the WWW

- WWW is a loosely organized information resource
- Many websites use static linked HTML files
  - can become inconsistent and outdated
- Many organizations provide dynamic access to databases directly from the Web
- Dynamic database access from Web introduces new problems for designers and DBAs

# Uses for Web-based DB Applications

- e-commerce has pushed organizations to develop Web-based database applications
  - To create world-wide markets
  - To deliver information
  - To provide better customer service
  - To communicate with their suppliers
  - To provide training for employees
  - To expand the workplace
  - ...Many other innovative activities

# Origins of The Internet

- Developed from **Arpanet**, communications network created in the 1960s by DARPA, US agency, for linking government and academic research institutions
- Used a common protocol, **TCP/IP**
- US National Science Foundation took over management of the network, then referred to as the **Internet**
- Navigating and using the Internet required considerable sophistication

# World Wide Web

- Tim Berners-Lee proposed a method of simplifying access to Internet resources in 1989
- Led to the development of the **World Wide Web**
- included notions of **URL, HTTP, HTML, hypertext, graphical browsers with links**
- Automated finding, downloading, and displaying files on the Internet

# URL

- Specific type of **Uniform Resource Identifier (URI)**
  - String giving the location of any type of resource on the Internet-Web pages, mailboxes, downloadable files, etc.

# HTTP

- Communications protocol
  - Standard for structure of messages
- HTTP is a **stateless** protocol
  - No facility for remembering previous interactions
  - Creates a problem for e-commerce, which requires a continuous session with the user

# HTML

- Data format used for presenting content on the Internet
- A **markup** language because HTML documents contain tags that provide formatting information for the text
- HTML document can contain applets, audio files, images, video files, content

# XML

- Extensible Markup Language - standard for document storage, exchange, and retrieval
- Created in 1996 by the World Wide Web Consortium (W3) XML Special Interest Group
- Users can define their own markup language, including their own tags that describe data items in documents, including databases
- Can define the structure of heterogeneous databases and support translation of data between different databases

# Components of XML Documents

- **Element** is the basic component of an XML document
- Document contains one or more XML elements, each of which has a **start tag** showing the name of the element, some **character data**, and an **end tag**
- Can be **sub-elements** of other elements- must be properly nested
- Can have **attributes** whose names and values are shown inside the element's start tag
- Attributes occur only once within each element, while sub-elements can occur any number of times
- Document can contain **entity references**-refer to external files, common text, Unicode characters, or reserved symbols

# Well-Formed XML Document

- Obey rules of XML
  - Starts with XML declaration
  - Root element contains all other elements
  - All elements properly nested

# DTD and XML Schema

- Users can define their own markup language by writing either
  - A **Document Type Declaration (DTD)**
    - A specification for a set of rules for the elements, attributes, and entities of a document
    - A document that obeys the rules of its associated DTD is **type-valid**
  - An **XML Schema**
    - New, more powerful way to describe the structure of documents
    - A document that conforms to an XML schema is **schema-valid**

# DTD Rules

- DTD is enclosed in `<!DOCTYPE name[DTDdeclaration]>`
- each **element** is declared using a type declaration with structure `<!ELEMENT (content type)>`
- In an element declaration, the name of any sub-element can be followed by one of the symbols \*, + or ?, to indicate the number of times the sub-element occurs
- **Attribute** list declarations for elements are declared outside the element

# XML Schema

- Permits more complex structure than DTD
- Additional fundamental datatypes, UDTs
- User-created domain vocabulary
- Supports uniqueness and foreign key constraints
- Schema lists elements and attributes
  - Elements may be complex, which means they have sub-elements, or simple
  - elements can occur multiple times
  - Attributes or elements can be used to store data values
  - Attributes used for simple values that are not repeated

# Three-tier Architecture

- Three major functions required in an Internet environment: presentation, application logic, data management
- Placement of functions depends on architecture of system
- **Three tier architectures** completely separate application logic from data management
  - Client handles the user interface, the **presentation layer** or first tier
  - **Application server** executes the application logic -the **middle tier**
  - **Database server** forms the third tier
- Communications network connects each tier to the next

# Advantages of 3-tier Architecture

- Allows support for **thin clients** that only handle the presentation layer
- Independence of tiers; may use different platforms
- Easier application maintenance on the application server
- Integrated transparent data access to heterogeneous data sources
- Scalability

# Presentation Layer

- **HTML forms** often used at the presentation layer
- Scripting languages such as Perl, JavaScript, JScript, VBScript, may be embedded in HTML to provide some client-side processing
- **Style sheets** specify how data is presented on specific devices

# Application Server

- Middle tier - responsible for executing applications
  - Determines the flow of control
  - Acquires input data from presentation layer
  - Makes data requests to database server
  - Accepts query results from database layer
  - Uses them to assemble dynamically generated HTML pages
- Server-side processing can use different technologies such as Java Servlets, Java Server pages, etc.
- **CGI**, Common Gateway Interface, can be used to connect HTML forms with application programs
- To maintain state during a session, servers may use cookies, hidden fields in HTML forms, and URI extensions.
  - Cookies generated at the middle tier using Java's Cookie class, sent to the client, where they are stored in the browser cache

# Data Layer

- Third layer is standard database or other data source
- Ideally on separate server

# XML and Semi-structured Data Model

- Semi-structured data model uses a tree structure
- Nodes represent **complex objects** or **atomic values**
- An **edge** represents either relationship between an object and its sub-object, or between an object and its value
- Leaf nodes, with no sub-objects, represent values
- Nodes of the graph for a structured XML document are ordered using **pre-order traversal**, depth-first, left-to-right order
- There is no separate schema, since the graph is self-describing

# Queries

- **XQuery** is W3C standard query language for XML data
  - Uses the abstract logical structure of a document as it is encoded in XML
  - Queries use a path expression, which comes from an earlier language, **XPath**
  - Consists of the document name and specification of the elements to be retrieved, using a path relationship
  - Can add conditions to any nodes in a path expression
  - Evaluated by reading forward in the document until a node of the specified type and condition is encountered

# FLOWR Expressions

- XQuery uses a **FLOWR** expression::FOR, LET, WHERE, ORDER BY, and RETURN clauses
- Ex

```
FOR $C IN doc("CustomerList.xml")//Customer
WHERE $C/Type="Individual"
ORDER BY Name
RETURN <Result> $N/Name, $N/Status </Result>
```
- Allows for binding of variables to results
- Allows for iterating through the nodes of a document
- Allows joins to be performed
- Allows data to be restructured
- XQuery provides many predefined functions, including count, avg, max, min, and sum, which can be used in FLOWR expressions.

# XML and Relational Databases

- Relational DBMSs extended their native datatypes to allow storage of XML documents
- Also possible to use SQL with XPath expressions to retrieve values from the database
- Existing heterogeneous databases can be queried using standard languages such as SQL, and query results can be placed into an XML instance document
- Query language has to have facilities that can tag and structure relational data into XML format