Week 1 part 2: Database Applications and Technologies

…Data everywhere…
SQL Databases, Packaged applications
Data warehouses, Groupware
Internet databases, Data mining
Object-relational databases,
Scientific databases

Based on a Colloquium by Philip A. Bernstein presented at the University of Toronto, on October 19, 1999

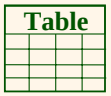
The Data Scalability Problem

- **Very large data sets are everywhere ...**
- **Terabyte data warehouses**
- **Scientific databases (GB's/day)**
- **Thousands of databases per enterprise**
- **Thousands of tables per database**
- **The Internet … all of the world's documents at your fingertips!**
- **1B machines with 8GB = 8PB (Pentabytes, that is)**
- **What's limiting our ability to fully exploit on-line data is not just size but *complexity*.**

Where do these data come from?
What technologies do they use??

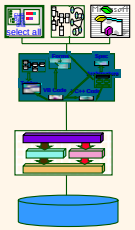
Whatever they use, they need models (schemas, metadata,…)

SQL Database Systems



- ~ $5.4B in new SQL DB licenses in 1998
  - Plus another ~ $1B in DB tools
  - 2/3 transaction processing, 1/3 decision support
- Transaction processing
  - 135K txns/min (tpmC), $97/tpmC, $13.1M system cost
  - 20K txns/min (tpmC), $15/tpmC, $305K system cost
    - 7x txn rate, for 43x the price! (Scale out vs. scale up)
  - TP monitors are becoming Internet application servers
  - They're at the core of all big e-commerce sites.
- Query optimization is excellent, but there's still much room for improvement

Information Resource Management



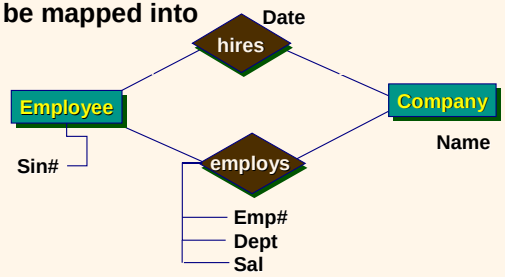
- Managing descriptions of DB's and applications
  - Reverse engineer & import code and DB schema
  - Catalogue management and browser tools
- Applications
  - Metadata reporting- find relevant databases
  - Year 2000 - find date fields
  - Database translation - move to a different DBMS
- Technology – entity-relationship model on SQL DBMS
  - Platinum (CA), Viasoft (R&O)
  - No automatic integration or categorization

Example of DB Reengineering

The relation

`Employee(emp#,cname,dept,hiredt,sin#,sal)`

might be mapped into

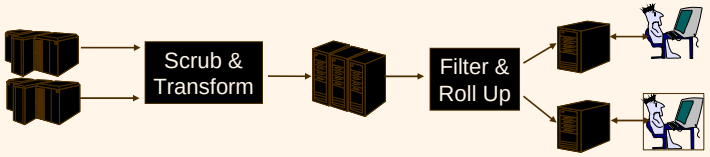


Packaged Applications

- Packaged applications (Enterprise Resource Planning systems) drive a lot of DB business -- there is little custom application development at large companies
- Commercial applications - finance, manufacturing, distribution, human resources, order processing ...
- Vendors include SAP ($5.1B), Peoplesoft ($1.3B), Baan ($700M), Oracle apps ($700M), growing at 20%
- They all have big built-in repositories of models to manage DB mapping, customization, & application upgrades.
- Models drive package integration
  - Load models from legacy apps; then write adapters
  - Interfaces on which to build custom front ends

Data Warehousing

- Business goal - decision support for everyone
- Problem - Ad hoc query on production DBs doesn't work
- Approach - Create snapshot DB for decision support
  = a data warehouse
- One of main growth areas of the DB business
  - $2B revenue in 1995, grew to $8B in 1998 (h/w + s/w)



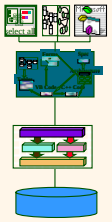
The Killer Model Mgmt App

- Creating and maintaining a Data Warehouse is hard. You need tools, which require lots of models.

Inconsistent data formats	Data quality & timeliness
Missing or invalid data	Relate technical and business models
Semantic inconsistencies	Tracing data lineage (instance-level)

Model-driven data transformation tools

- Generate code for loading a data warehouse
- Version schemas and transformations for lineage
- Use tool-specific repository engines on SQL DBMS
- More powerful and general-purpose tools are needed



Groupware Applications

- E-Mail/BBoards aren't yet on relational DBMSs
  - They're huge… 100MB/person x 10K persons = 1TB
  - Heavily used for document management
  - $2B/yr + 15%/yr, a major and growing part of IT.
- Web-based portals are essential to core business, combine data warehouse and document DB (finance, marketing)
- Very weak model management today
  - Ad hoc extensibility of models in today's mail systems
  - Yahoo!-like categorization and its data sources e.g., cost centers, organization structure, sales docs…
  - Web site management (dependencies and configurations)

Internet Databases

- The other big growth area of today's DB business
- Internet content-oriented meta-data is big business -- indexing, categorization hierarchies, comparison shopping, …
- Structural metadata is just starting to get attention
  - XML interoperation requires libraries of DTD's
  - Need tools to map to related DB schemas, forms, …
- E-commerce application deployment requires configuration definition and requirements -- TP monitor and ORB technology is weak.
- Vision - Plug in a DB or application and it's accessible and manageable; Improve hot links, search engines, and categorizations

Data Mining

- Extract patterns and models from the data
  - Automatically, from large data sets
  - Use statistics, pattern recognition, machine learning, visualization
- Applications - finance, fraud detection, astronomy, marketing analysis, medical diagnosis, biology …
- A small market, with much projected growth
- Must define and manage data subsets to mine
  - And retain measures, derived data, and lineage
  - Like scientific experiments, with irregular analysis steps
  - Versioning is important.

Object-Relational Databases

- Goal - Capture more of the world's data, not just scalars - text, video, audio, time series, graphs, digital photos, medical records, insurance claims, etc.
- Richer operators - specialized to data types -- content-based retrieval, image similarity, path search,
- Richer constraints & triggers, e.g., to support workflow
- Adding a data type affects every DBMS component, components must be designed for extensibility
- All DB vendors are working on it.
  - Today's systems are incomplete and hard to extend
  - 3rd parties will supply data blades / extenders / cartridges

Managing Object-Relational Models

OR DB requires richer models in its catalogues

- Connections between types, such as inheritance, nesting, and relationships
- Abstract data types, including operations (code becomes a more intrinsic part of DBMS)
- That is, DBMS catalog's information model is extended and becomes more extensible
- OR DB could be used for extended types that are especially suitable for models
  - Labeled directed graphs, with closure functionality
  - Versions and configurations

Scientific Databases

- CERN laboratory planning a new generation of experiments in particle physics which will generate 1-10PB per year starting in 2005.
- These data will be distributed world-wide to be analyzed by scientists (public.web.cern.ch/Public/
- To deal with the data, there is an international Data Grid project e.g., www.cacr.caltech.edu/ppdg/, grid.web.cern.ch/grid/

What Was Left Out

- Digital libraries
- Document management
- Directory services
- Advanced file systems
- Software databases and repositories
- Configuration management systems
- . . .

Asilomar Report Grand Challenge

The Information Utility

"...Make it easy for everyone to store, organize, access, and analyze the majority of human information online…"