

Information Models: Applications and Technology

Application Areas
Model Management Technology



Colloquium by Philip A. Bernstein
in DCS on October 19, 1999

The Data Scalability Problem

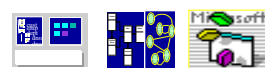
- Very large data sets are everywhere ...
- Terabyte data warehouses
- Scientific databases (GB's/day)
- Thousands of databases per enterprise
- Thousands of tables per database
- The Internet ... all of the world's documents at your fingertips!
- 1B machines with 8GB = 8PB (Pentabytes, that is)
- What's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

Main Themes

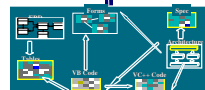


- Use structural metadata (models) to manage data complexity.
- These may include database schemas, DTDs, conceptual models (UML/ER diagrams) and more.
- Model management is a large and growing part of data management, not just for databases though.
- It's a messy area, in need of unification
 - ✓ I'll (...Phil, that is...) suggest how.

Model Management Architecture



**Model-Driven
Tools**



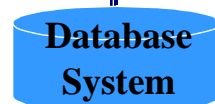
**Information
Model**

- What do products do today?



**Repository
Engine**

- What's easy?
What's hard?



**Database
Engine**

Outline

✓ Introduction

■ Model Management in Today's Databases

- SQL Databases
- Internet Databases
- Packaged Applications
- Data Mining
- Data Warehouses
- Object-Relational Databases
- Groupware

■ Model Management Technology

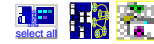
■ Conclusions

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



Packaged Applications

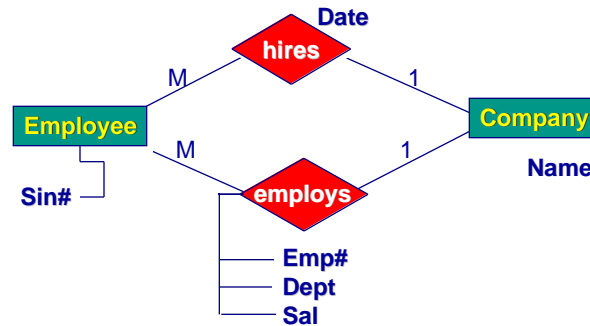
- Packaged applications (Enterprise Resource Planning systems) drive a lot of DB business -- there is little custom application development at large companies
- Commercial applications - finance, manufacturing, distribution, human resources, order processing ...
- Vendors include SAP (\$5.1B), Peoplesoft (\$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

Example of DB Reengineering

The relation

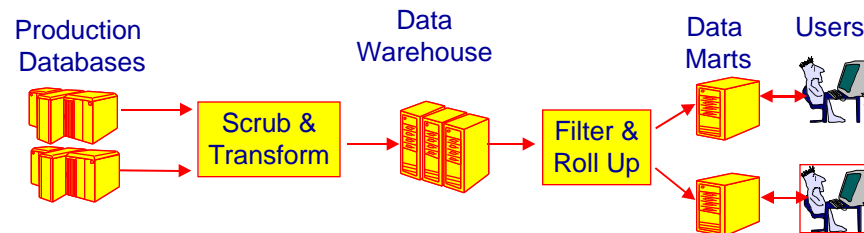
Employee(emp#, cname, dept, hiredt, sin#, sal)

might be mapped into



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

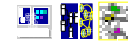


- Creating and maintaining a Data Warehouse is hard. You need tools, which require lots of models.

- Inconsistent data formats
- Missing or invalid data
- Semantic inconsistencies
- Data quality & timeliness
- Relate technical and business models
- 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

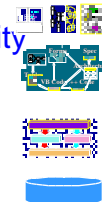


- Goal - Capture more of the world's data, not just scalars - text, video, audio, time series, graphs, digital photos, medical records, insurance claims, etc.
- Richer operators - specialized to data types -- content-based retrieval, image similarity, path search,
- Richer constraints & triggers, e.g., 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



What Was Left Out

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

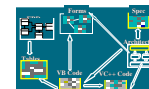
Scientific Databases

- 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/

Outline

- ✓ Introduction
- ✓ Model Management in Today's Databases
- Model Management Technology
- Conclusions

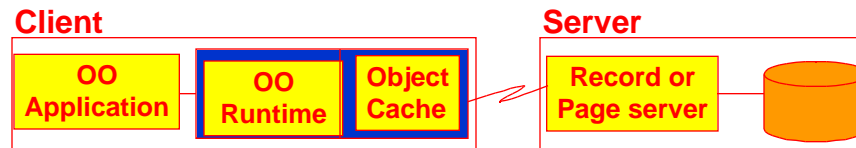
Model Management Technology



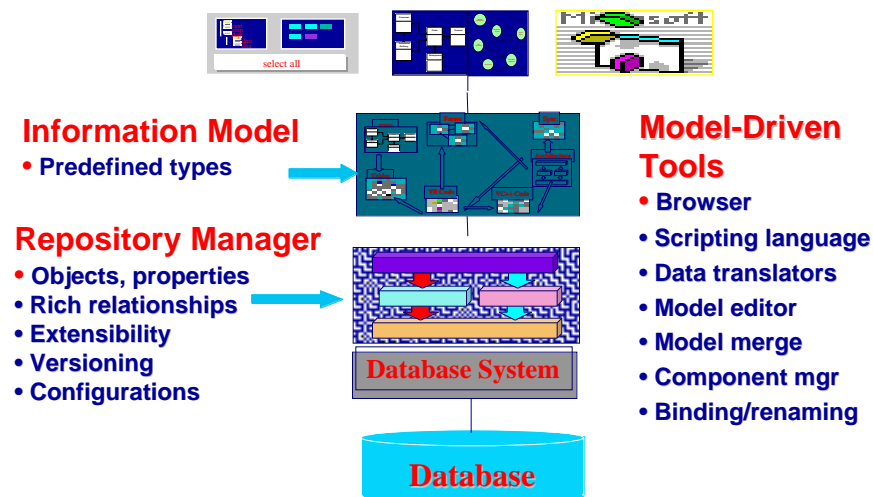
- Repository – a DB system for managing models of engineered artifacts, global across tools
- Today's repository products
 - ✓ Many are hard-coded meta-models (commodity tools)
 - ✓ Most run on RDBMSs (Platinum, SAP, Oracle, MS, ...)
 - ✓ A few implemented their own DBMS (Softlab, Viasoft)
 - ✓ A few run on OODBs (IBM, Unisys)

Use an Object-Oriented DB to Manage Models?

- Cures mismatch between OO applications and data
- Especially for design apps: CAD, CASE, CAM
- Like OR, it encapsulates behavior with objects
- Not a commercial success (2% market share)
 - ✓ Doesn't replace a SQL DB (weak TP, SQL, catalog)
 - ✓ Not a complete solution for design information



A Repository Product

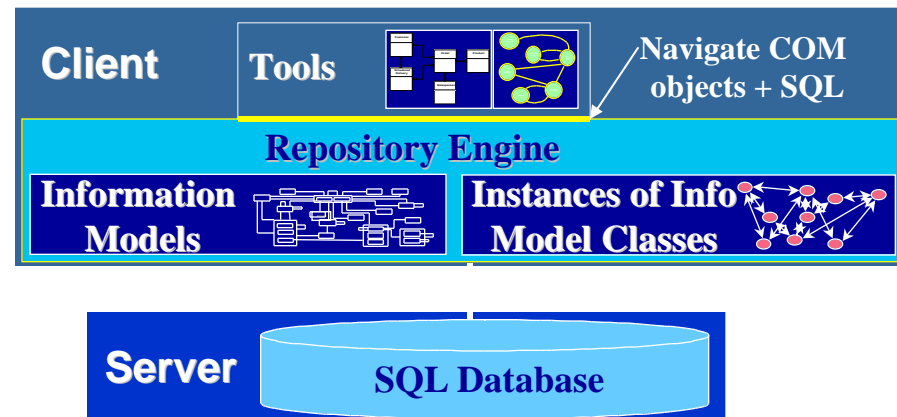


Major Technical Issues

- Engine functions are highly interdependent
 - ✓ Versioned objects and relationship with extensibility
- High Performance - to drive GUI tools
 - ✓ Client caching, object clustering, clever prefetching
 - ✓ Scalable to hundreds of on-line users and large DBs
- Generic model-driven tools
 - ✓ Browsing, merging, impact analysis, translation, ...
- All on the customer's favorite SQL DBMS
 - ✓ Which is very hard to do efficiently

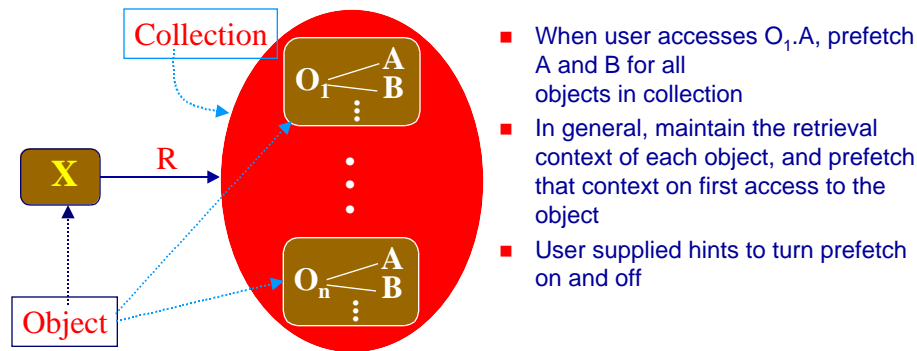
Microsoft Repository

- Contains COM objects, properties, relationships
- In SQL Server and Visual Studio (since 1997)
- Supports data warehouse and component mgmt



Performance: Context-based Prefetch

- When translating navigation to SQL, you must batch
 - ✓ Over 5x performance gain
 - ✓ But the application accesses an object-at-a-time ...



- When user accesses O₁.A, prefetch A and B for all objects in collection
- In general, maintain the retrieval context of each object, and prefetch that context on first access to the object
- User supplied hints to turn prefetch on and off

Versioning

- Versions with branching/merging are required
- To avoid blowup in database size, use delta storage
 - ✓ Rows contain a version range
 - ✓ Updated content causes range to split

ObjID	Branch ID	Version Start	Version End	Attr1	Attr2	...
17	0	0	∞	"A"	1	
22	0	0	2	"B"	4	
22	0	3	∞	"C"	4	

- Relationships are harder -- bidirectional, sequenced, pinned

Other Features

- Seamless integration with COM
- Dynamic type extensibility
- Flexible object-to-table mapping
- SQL queries that return COM objects
- Workspaces and checkout-checkin
- Large and growing information model

Outline

- ✓ **Introduction**
- ✓ **Model Management in Today's Databases**
- ✓ **Model Management Technology**
- **Conclusions**

Asilomar Report Grand Challenge

The Information Utility

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

Improved model management is a major part of the solution.

What's Next for Model Management Research?

- **Better repository functionality & performance**
- **Automated heterogeneous database integration**
- **Automated integration of web content**
- **Auto-everything DB systems**
- **More unification of data and process**

Conclusions

- **Model management is a field which requires more coordinated attention.**
- **Model management repositories are an important kind of DB system -- need better implementation technology for them on SQL DBMSs.**
- **Model management is at the core of the most important open problems in the DB field.**

To Find Out More...

<http://www.research.microsoft.com/~philbe>

philbe@microsoft.com