

Database Technology and Challenge of Emerging Applications

Shamkant B. Navathe
Georgia Institute of Technology
Atlanta, GA 30332
sham@cc.gatech.edu
<http://www.cc.gatech.edu/~sham>



©Shamkant B. Navathe

1

DATABASE TECHNOLOGIES

©Shamkant B. Navathe

2

The Science

Database Concepts-

- Models
- Design Methodologies
- Query languages
- Transaction Models

Database Theory:

- Data Dependency Theory
- Database Design and Normalization theory
- Theory of Concurrency Control, Transaction Management and Recovery

The Technology

Engineering of large scale robust products for reliable use in industry, government and commerce

The Science

Database Concepts-

- Models
- Design Methodologies
- Query languages
- Transaction Models

Database Theory:

- Data Dependency Theory
- Database Design and Normalization theory
- Theory of Concurrency Control, Transaction Management and Recovery

The Technology

Engineering of large scale robust products for reliable use in industry, government and commerce

History of Database Management Technology

1965-1980 Dominance of Hierarchical Model (IBM's IMS) and Network Based Systems

- IDS (Honeywell -> Bull)
- IDMS (Goodrich --> Cullinet --> CA)
- DMS 1100 (Univac --> Unisys)
- TOTAL/SUPRA (Cincom)
- VAX DBMS(Digital --> Compaq)

1980's -- Commercial Entry of the Relational Model

- IBM's SQL/DS, DB2, DB2/2
- INGRES, ORACLE, INFORMIX, SYBASE, ALLBASE(H-P)

1990's Increasing acceptance of O-O and other DBMSs

Late 90's: The Object Relational Approach

2000 and beyond: ??

Proliferation of Non-standard data formats and standards

MODELS OF DATA - continued

- **1964/1965**
 - HIERARCHICAL DATA MODEL
 - NETWORK (CODASYL/ DBTG) DATA MODEL
- **1970**
 - PROPOSAL OF RELATIONAL DATA MODEL
- **1976**
 - ENTITY RELATIONSHIP DATA MODEL
- **1982**
 - ADVENT OF RELATIONAL DBMSs (IBM)
- **Early 90's**
 - Object Oriented Database Model
 - 1993 – WWW (with HTML)
- **Late 90's**
 - XML and related family of markup languages

SOURCES OF DATA

- File Systems Based Data
 - Record-based
- Database Systems Based Data
 - Structured
 - Formatted
 - Largely Numeric
 - Passive and Static
- Newer Forms of Data
 - Web-based
 - Unstructured
 - Mostly text and images
 - “Dynamic Data”
 - “Standard” metadata (e.g., RDF)

TYPES AND FORMATS OF DATA

- Levels of Abstraction
 - Conceptual
 - Logical
 - Physical
- Three Level ANSI-SPARC architecture (1971) is still valid.
- Formats from Application Domains
 - Net CDF : multidimensional array data
 - ASN.1: networks, genome data management
 - AceDB : Genetic Database for Unstructured Data
 - Terrain 3-D data : DEM (digital elevation model)
TIN (triangular irregular network)

DBMS Product Families

Current Categories

- Client/Server Application Products
- Relational DBMSs and their variants
 - SQL Servers; Oracle, DB2, Sybase, Informix, Microsoft
 - Parallel DB Servers -- Teradata(AT&T), Oracle, Informix
- Object-Oriented Databases
 - Objectstore, Versant, O2, Gemstone, ONTOS, POET, Jasmine...
- Case/Design and Prototyping Tools
 - Conceptual/Logical Modelling - ERWin, S- Designer
 - Information Flow/Process Modelling - BPWin, Flowmark
 - Prototyping/Interface Design – Powerbuilder and beyond

Database Product Scene – The 90's

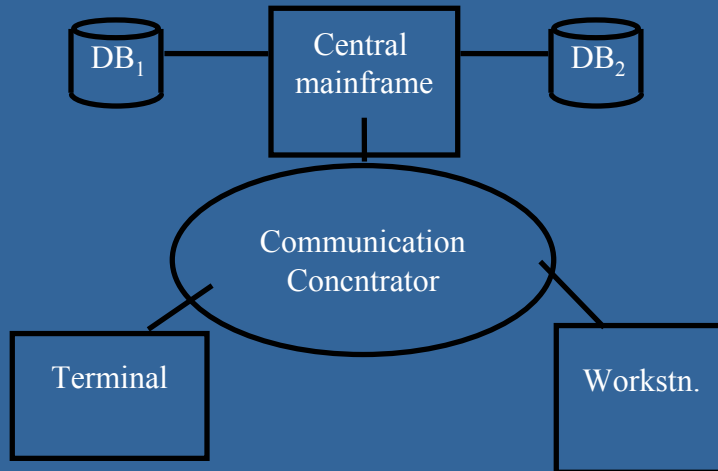
- Data Warehousing Products
 - Most relational vendors + Redbrick, Arbor Software, HP
 - DW market \$6 to 8 Bills /year
- On Line Analytical Processing (OLAP) Tools
 - Relational OLAP
 - Multi-dimensional OLAP
- Extraction, Transformation, Loading Tools
 - Prism
 - Extract (ETI)
- Search Engines, Document Managers, Text Processors
- Object Relational Databases
 - Universal Servers: Illustra+Informix, UniSQL
 - Oracle 8.x, IBM's UDB, Sybase Adaptive Server
- OOAD Tools , e.g., Rational Rose
- Open Source Databases: MySQL , Postgres

Merging Of Technologies

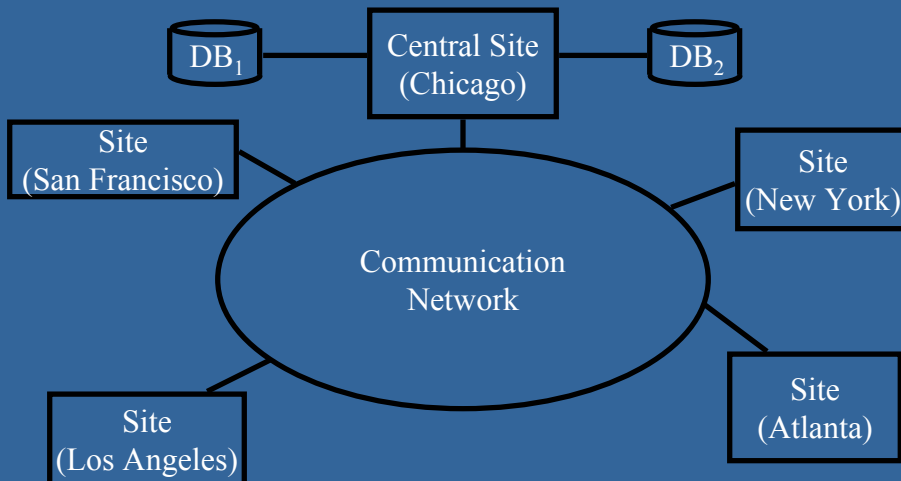
- **DB TECHNOLOGY** provides
 - Data Models
 - Design Theory
 - Integrity and Consistency Maintenance
 - Robust Transactions / Storage System Support
- **WEB TECHNOLOGY** provides
 - Global Infrastructure and a set of standards to support document exchange
 - A presentation format for Hypertext
 - User Interfaces and links for Document Retrieval
 - XML as a structured data format
 - Interactivity and ability to ship code with data

DBMS ARCHITECTURES

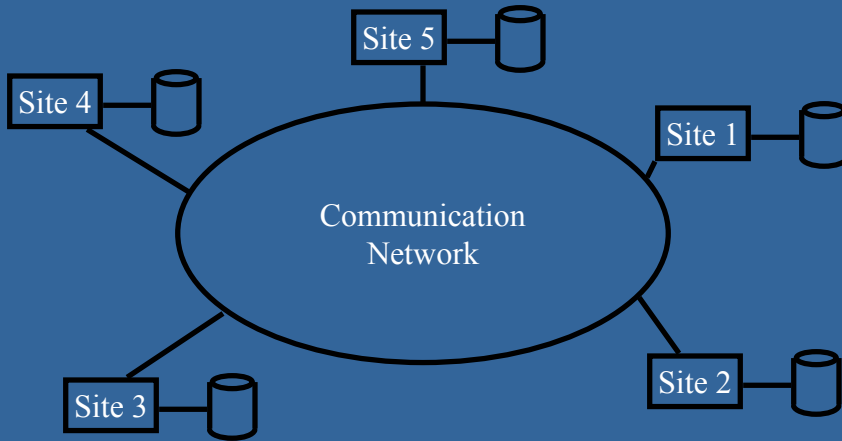
Centralized Databases with Terminals And workstations (from late 70's)



Centralized Databases on a Network



A Truly Distributed Database on a Network

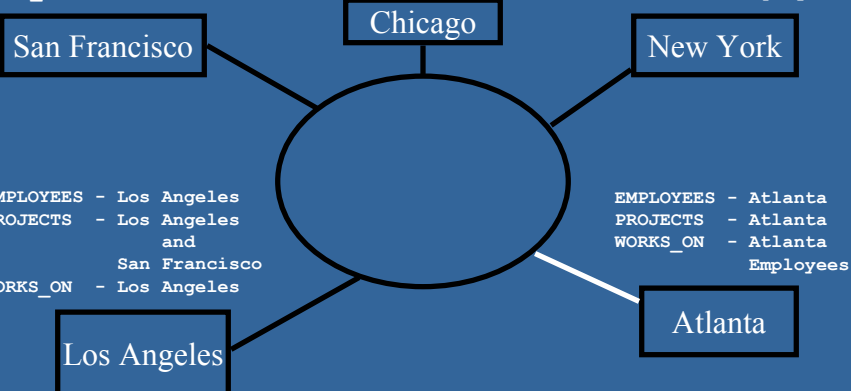


Data Distribution and Replication Among Distributed Databases

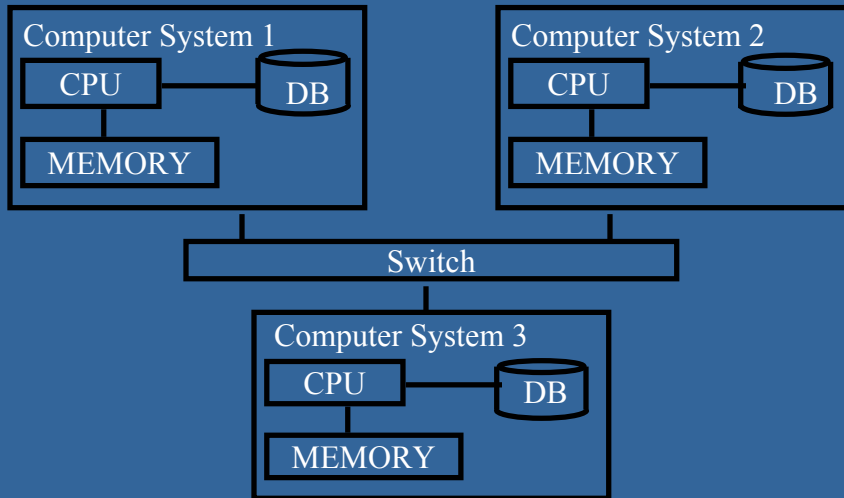
EMPLOYEES - San Francisco
and
Los Angeles
PROJECTS - San Francisco
WORKS_ON - San Francisco

EMPLOYEES - All
PROJECTS - All
WORKS_ON - All

EMPLOYEES - New York
PROJECTS - All
WORKS_ON - New York
Employees



Parallel Database Architecture



Shared Nothing Architecture (Multiprocessor System)

©Shamkant B. Navathe

17

DATABASE APPLICATIONS

©Shamkant B. Navathe

18

Traditional Applications

- Financial - Insurance, Accounting (GL, A/P, A/R), Banking
- Manufacturing - Inventory Management, Bill of Materials
- Government - Records Management
- HealthCare - Hospital Administration, Patient Mgmt.

Newer Application variations

- Functional categories of apps: manuf. scheduling, order-processing, materials management, shop-floor mgmt., sales, marketing, distribution, HR, project management
- ERP: Enterprise Resource Planning
- CRM: Customer Relationship Management
- Supply Chain Management
- Electronic Commerce and electronic funds transfer, payment and billing systems

Emerging Applications

- World Wide Web - The dominant force for the next n years (n=?)
- Engineering -- CAD/CAM/CIM/CAE
- Scientific Applications -- Weather Forecasting, Genome Databases, Earth Sciences (GIS), Environmental Applications
- Telecommunications + Databases
 - Network management, Telemedicine, Info Brokering
- Hypermedia Packaged Databases /Digital Libraries
 - Encyclopedias, Manuals, Part catalogs (CD-ROM Consumables), CD and Audio Libraries
- Multimedia, Entertainment, Visualization
 - Interactive, Virtual Reality based systems, Image repositories, Video on Demand
- Software engineering + Databases : treating code as data and managing versions

DB Technology at large

Systems Designed to apply across domains with possible database backends:

- Data Mining
 - IBM Intelligent Miner
 - SGI Mineset
- OLAP and Decision Support
- Document and Workflow Management
 - EDI, Scanning & Storage and Archiving
- ERP Systems (Enterprise Resource Planning) with DB backends
- Search and indexing services
- Information Brokering Services: multi-source information consolidation and delivery

Growing Complexity of Information

Data Management =
Handling Typical Business Data (Numeric)



Numeric and Non-numeric Data in Formatted
Records



Management of Complex Objects in General --
Design Objects, Biological Objects,
Documents, Physical Systems, Information
Flows



Multimedia Unstructured Data

EVER INCREASING USERS' EXPECTATIONS

WHAT DO USERS WANT?

Better interfaces

More visualization

More animation

Interactive Involvement

Different paradigms for search

querying, browsing, navigation,
exploration

***ALL THIS WITH RESPONSE TIME
CLOSE TO 2-3 SECONDS***

General Application Challenges

Multiple Dimensions of Information

Structural and Behavioral Data

Content + Relationship and Links among data

Incomplete, illdefined, illstructured, illformed information

Missing and erroneous information

From “Raw Data” to “Meaningful Information”

Extraction, Selection

Derivation, Deduction

Exploration and Discovery (data mining)

Real-time response with high volume databases

Automatic Monitoring and Control Apps in Robotics, Manufacturing

Users in the New Millenium

- **Hundreds of Millions PC's Worldwide**
- **Uninitiated, untrained users getting access and direct manipulation capability**
- **Internet access (particularly publicly accessed networks like the Internet) open up vast databases for public consumption**
- **Explosion of Data (3+ B web pages), Devices and Communication Modes**

Standards

Interoperability Standards like ODBC, JDBC, SOM, DCOM, OLE for interoperability

Document / Text

- EDI standards
- HTML
- XML/XSL/.....

Engineering

- PDES Products Data Exchange
- STEP ISO
- EXPRESS and EXPRESS-X

Object management standards

- Models - ODMG (Object Data Management Group)
- Definition - IDL
- Querying - ODL

Lack of Standards?

Security

- Modeling
- Language
- Implementation

GUI / Scripting Languages

- CGI
- PERL
- Java script

Object / Relational Database Management

- SQL3?
- XML/ XSL / XQL as DDLs and DMLs

OVERALL CHANGE OF FOCUS

- ROUTINE DATABASE MANAGEMENT FOR TRANSACTION SYSTEMS CONTINUES
- SOPHISTICATED DATA MANAGEMENT FOR PLANNING, DECISION MAKING
- APPLYING DATABASE AND RELATED WEB TECHNOLOGY FOR DOMAINS THAT HAVE BEEN LEFT OUT.

OLTP vs. OLAP

OLTP

OLAP

User	<ul style="list-style-type: none">• Clerk, IT Professional	<ul style="list-style-type: none">• Knowledge worker
Function	<ul style="list-style-type: none">• Day to day operations	<ul style="list-style-type: none">• Decision support
DB Design	<ul style="list-style-type: none">• Application-oriented (E-R based)	<ul style="list-style-type: none">• Subject-oriented (Star, snowflake)
Data	<ul style="list-style-type: none">• Current, Isolated	<ul style="list-style-type: none">• Historical, Consolidated
View	<ul style="list-style-type: none">• Detailed, Flat relational	<ul style="list-style-type: none">• Summarized, Multidimensional
Usage	<ul style="list-style-type: none">• Structured, Repetitive	<ul style="list-style-type: none">• Ad hoc
Unit of work	<ul style="list-style-type: none">• Short, Simple transaction	<ul style="list-style-type: none">• Complex query
Access	<ul style="list-style-type: none">• Read/write	<ul style="list-style-type: none">• Read Mostly
Operations	<ul style="list-style-type: none">• Index/hash on prim. Key	<ul style="list-style-type: none">• Lots of Scans
# Records accessed	<ul style="list-style-type: none">• Tens	<ul style="list-style-type: none">• Millions
#Users	<ul style="list-style-type: none">• Thousands	<ul style="list-style-type: none">• Hundreds
Db size	<ul style="list-style-type: none">• 100 MB-GB	<ul style="list-style-type: none">• 100GB-TB

Data Warehouse: a new name for a database

- A decision support database that is maintained separately from the organization's operational databases.
- A data warehouse is a
 - subject-oriented,
 - integrated,
 - time-varying,
 - non-volatilecollection of data that is used primarily in organizational decision making

Emerging Technologies tied to Data Management

- Communicating Databases
 - ubiquitous databases, mobile applications, mobile data
- More Intelligence in Data Management
 - reasoning, learning along many dimensions (model based, case based, analogy based, explanation based)
 - knowledge based systems with scaleable rule bases and databases
- Multimedia and User Interfaces- Display, Visualization, Animation
- Large Scale Software Development with Reuse / Component based Software Engineering
 - software parts, "glue" to hold pieces together
- Computer Supported Co-operative Work
 - working at home, working in teams

Internal DB Functionality Challenges

- **Functionality Trends**
 - More “behavioral” and “dynamic” specifications as a part of DB definition
 - More reliance on metadata
 - Migrate Application semantics into a database
 - Use of parallel processing combined with “declustering”
 - Rule processing as an integral function
- **Rule and Transaction Interaction**
 - Integrity control, security and authorization, active application control, derivation and deduction

Future Scenarios for databases

- Databases that refresh themselves by linking up with multiple sites and systems
- Databases that migrate with the users and are a part of different federations
- Databases that adapt to users’ needs and information request profile and keep info extracts
- Databases that prompt the users when new and relevant information arrives and deliver information in appropriate form (e.g. Point cast)

Current Research Thrusts

- New data models for new data types and relationships (image, video, sound, hypertext)
- Data Mining and Knowledge Discovery
- New retrieval and processing models for Browsing, Animation, Visualization
- Distributed, Parallel, Mobile, Replicated Database processing
- Methodologies for designing large scale applications -- corresponding tools (e.g.. reverse engineering).
- Heterogeneous database integration
 - practical issues -- databases on the WWW, application domain specific integration of data and tools

Open Issues

- **Push Vs. Pull Technologies**
 - How to achieve the right balance?
- **Content based Retrieval**
 - On all types of data including image, text, audio, video
- **Coping with Information Overload**
 - How to answer a query intelligently on the web?
- **Human Vs. Machine information processing: impedance mismatch?**

Challenges for the Database Professionals

- **Learning the application**
 - the Jargon
 - the “Process Model” of the environment
 - complexities, typical scenarios, rules, constraints
- **Apply DB Techniques to help in application**
 - conceptual modeling
 - views, transactions
 - specification, normalization, query optimization
- **Apply techniques outside DB area to DB management**
 - From AI, Information Retrieval, Software Engineering, UI

©Shamkant B. Navathe

37

Future

- **More variety of applications (and data)**
 - Cax where x = D (Design), E(Engineering), P(Publishing), E(Education)
 - Scientific (Biology, Weather forecasting, Space Image data)
 - Entertainment, Worldwide project teams, Electronic Commerce
- **More variety of users -- housewives, K-12 children, occasional and naïve users**
- **A great need for domain experts who can also understand db modeling and design**
- **More demands on performance**
 - scaling to larger databases
 - staying within decent response time

©Shamkant B. Navathe

38

TOWARD

- **Multilingual**
- **Multiplatform**
- **Multiprocessor**
- **Integrated**
- **Intelligent**
- **Interoperable**
- **Adaptive**

**Database and Knowledge Base Systems
and Personalized User Environments**

SOME REFERENCES

- **DATA ON THE WEB: From Relations to Semistructured Data and XML**
by Serge Abiteboul, Peter Buneman, Dan Suciu
Morgan Kaufmann, 2000.
- **Data management: Past, Present, Future**
by Jim Gray, IEEE Computer, 29:10, 1996,
pp. 38-46.
- **The Web in 2010: Challenges and Opportunities
for Database Research”,** by Gerhard Weikum,
University of Saarland, Saarbruecken,
Germany.
www-dbs.cs.uni-sb.de

Research in GT database group by former students



- **Mobile Intermittently Synchronized Databases- problems of consistency, scalability (1998-2003, Wai Gen yee)**
- **Conceptual Modeling of Secure data and risk assessment of security measures (1995-2004, Yong-chul Oh, Fariborz Farahmand)**
- **Transaction Models for web services and recovery and efficiency issues of web transactions (1999-2003 Deb Vander Meer)**

Research in GT database group



- **Genome Databases - mitochondrial Genome databas MITOMAP (with Emory, now at UC-Irvine): www.mitomap.org (1995-1998, Andy Kogelnik)**
- **Centralized and Parallel mining algorithms for association rules (1994-98, Ashok Savasere)**
- **Visual User Interfaces for Interactive Information retrieval (1993-97, Aravindan Veerasamy)**
- **Efficient management of parallel databases and indexes (1993-95, Kiran Achyutuni)**

SAMPLE PUBLICATIONS (Navathe with Students)

- “Grouping Techniques for Update Propagation in Intermittently Connected Databases” by Mahajan, Donahoo, Navathe, Ammar and Malik, *IEEE International Conference on Data Engineering*, February, 1998.
- “A Framework for Designing Update Objects to improve Server Scalability in Intermittently Synchronized Databases” by Yee, Donahoo, Navathe, *Conference on Information and Knowledge Management (CIKM)*, Wash. D.C., November 2000.

PUBLICATIONS (contd.)

- “Enabling Scalable Online personalization on the Web” by Vandermeer, Dutta, Datta, Ramamritham, Navathe, *ACM SIGECOMM International Conference*, October 2000.
- Waigen Yee, Michael J. Donahoo, Edward Omiecinski, and S. B. Navathe, “Scaling Replica Maintenance in Intermittently Synchronized Databases, *Proc. ACM Conference on Information and Knowledge Management (CIKM)*, Atlanta, November 2001.
- Waigen Yee, S. B. Navathe, Edward Omiecinski, and Christopher Germaine, “Bridging the Gap between Response time and Energy-efficiency in Broadcast Schedule Design,” *Proc. Eighth Extensions of Database Technology (EDBT) Conf.*, Prague, Czech Republic, Springer Verlag, March 2002.

PUBLICATIONS (contd.)

- “Enabling Scalable Online personalization on the Web” by Vandermeer, Dutta, Datta, Ramamritham, Navathe, *ACM SIGECOMM International Conference*, October 2000.
- “Discovery of Multiple-Level Association Rules from Large Databases” by Savasere, Omiecinski, Navathe, *21st International VLDB Conference*, Zurich, September 1995.
- “Mining for Strong Negative Associations in a Large Database of Customer Transactions,” by Savasere, Omiecinski, Navathe, *14th International Conference on Data Engineering*, Orlando, FL., February 1998.

PUBLICATIONS (contd.)

- A, • Veerasamy, S. Hudson and S. B. Navathe, "Visual Interfaces for Text Information Retrieval Systems," Proc. of Visual Database Systems -3, an IFIP WG 2.6 Workshop, Lausanne, Switzerland, March 1995.
 - K. Achyutuni, E. Omiecinski and S. Navathe. "Two Techniques for On-line IndexModification in Shared Nothing Parallel Databases," *In Proceedings of the ACM SIGMOD International Conference on Management of Data*, June 1996.

PUBLICATIONS (contd.)

“ Mining for Strong Negative Associations in a Large Database of Customer Transactions,” by Savasere, Omiecinski, Navathe, *14th International Conference on Data Engineering*, Orlando, FL., February 1998.