Wrap-Up: Data Sharing and the Web

Zachary G. Ives
University of Pennsylvania

April 16, 2003

Administrivia I

- Reminder: Monday is your project presentation
  - About 5-7 minutes each
  - Slides are allowed (but not required); demos welcome (but also not required)
    - What did you do?
    - What were the hard problems?
    - How are you solving them?
    - How are you evaluating your work?

Take-home final exam also distributed Monday

- Will likely be 3-4 essay questions; open-book, open-notes
- Due Friday May 2 before 6PM EST

Administrivia II

- Project reports due Friday May 2
  - 5-10 pages for implementation project; 10-15 pages for survey paper
  - Structured like a conference paper
  - Remember to answer:
    - What problem you're addressing
    - Why it's a problem
    - How you're attacking it
    - How you're going to show you succeeded

Where Have We Been? I

- Our themes:
  - Data sharing using semantics
  - Web and Internet
- The central characters:
  - Data integration
    - What's does it look like?
    - What does it do?
    - vs. warehousing?
  - Decentralized variants
    - Peer data management systems

Where Have We Been? II

- Processing queries
  - Query optimization
    - What were the two systems?
  - Query execution
    - Basic principles
  - Adaptive variants
    - Inter-query: SITs
    - Intra-query: mid-query re-optimization, eddies
  - Query processing in a distributed context
    - Mariposa, Tukwila
  - Answering queries using views

Where Have We Been? III

- Other ways of sharing data:
  - Publish-subscribe: XFilter
  - IR-style querying
  - IR/DB hybrids: WHIRL, extended XML-QL
  - Difference-based models:
    - Change detection in semistructured data
    - Heraclitus
    - Harmony
  - Semantic Web
  - Groupware
- Matching schemas and data

How Much Real-World Impact? I

- Data integration:
  - Systems available from Nimble Technology, Enosys Markets, IBM, BEA, Progressive Software, …
- Query optimization:
  - Every DBMS uses System-R or Volcano-derived optimizer
- Query execution:
  - Basic techniques common to every DBMS
- Adaptive query processing:
  - Only a few techniques (e.g., SITs) currently available
- Answering queries using views:
  - Data integration systems; DB2, Oracle, likely SQL Server

How Much Real-World Impact? II

- Publish-subscribe:
  - Not too much use (exception: RSS, available from Netscape/AOL, CNET, etc.)
- IR querying:
  - Google is partly based on this!
  - DB hybrids: DB2 Text Extender, SQL Server text ext.
- Synchronization/change detection:
  - CVS, diff3, Unison
- Groupware:
  - SMTP, Exchange server, Domino, Groove, …
- Semantic Web:
  - The jury is still out…

What's Likely to Be Adopted Soon?

- Adaptive query processing
  - "Autonomic computing" is the big buzzword
  - Much like JIT compiling…
- Industry may actually be driving most of the adaptive work in the future:
  - Adapting to new configurations, new hardware
  - Adapting to workloads
  - Self-healing (failover)

What Are Today's Hottest Research Topics? I

- XML!!
  - Everyone does papers on XML processing, XML indices, XML validation, XML constraints, XML compression, and so on
- Data streams!!!!
  - The new hot topic, though it has many meanings
  - Sensor data; stock data; astronomy data
  - Mining streams; statistics for streams; querying streams over sliding windows; etc.
- Traditional stuff: optimization, execution, indexing

What Are Today's Hottest Research Topics? II

- Peer-to-peer architectures
  - Using P2P for scalability reasons
- Semantic translation
  - Mapping schemas; manipulating schemas
  - Querying in this context
  - Semantic Web
- Architecture-aware DBs
  - Cache-aware algorithms; scheduler-aware algorithms

In Summary…

- I hope you found this course to be “less dull than average” ☺
  - Databases doesn’t have to mean payroll systems or web back-ends
- I hope you learned something that’s useful regardless of whether you want to be a DB researcher
  - How to read and analyze technical papers
  - How to present
  - How to write
  - ... And a little about what goes on inside the DBMS
- Thanks for being a fabulous class!