The Future of Data: A Smorgasbord

Guy M. Lohman
IBM Almaden Research Center

Myth #1: XML will "solve" the data format problem

- Heterogeneity will always reign!
- Not everything will be XMLized! Legacy systems, flat files, the next "great thing",...
- Who's going to control the semantics of all those XML tags? Remember, the "X" stands for "extensible"! Everyone and his mother will be coming up with new tags, and who knows what they mean when you're searching the web.
XML is NOT a Panacea

- EXAMPLE 1: What does tag `<salary>` mean?
  - What currency?
  - What frequency? (annual, monthly, hourly, ...?)
- EXAMPLE 2: What does value "order" mean when its tag is `<type>`?
  - Type of what?
  - "Order" of what? Purchase? Sequence?
- Gives *some* increased context...
- But only a *slight* improvement over Google search!

Myth #2: Relational is Dead --
Native XML repositories are the future

- Relational DBMSs are hugely successful, with a complete array of utilities, features, and performance honing.
- Evolutionary rather than revolutionary changes are the only way that change will happen.
- Remember how object-oriented systems, which surely subsumed relational systems, were going to replace relational?
**Myth #3: Just shred everything into relational tables!**

- Boy, that's a LOT of work for all documents, few of which will ever be retrieved by queries
- Many documents won't even be searched!
- This won't exploit the nesting structure that XML provides -- a lost opportunity

**Myth #4: Everything's off the Web as Data Streams**

- SOMEONE has to store the stuff!
- Companies won't store their corporate jewels on the Web, except possibly in an Intranet inside the firewall
- Cacheing will become even more commonplace, for performance
Myth #5: There's just one copy of the data I'm interested in

- Multiple levels of cacheing is now commonplace
  - Edge servers
  - Mobile clients that are periodically detached
  - Multiple tiers
  - Multiple components within a server
- Different degrees of synchronization
- Synchronizing is a major headache!

Cache Write-Through Dilemma

Guido = 'jerk'

Guido = 'smart'

Guido = 'smart'

Guido = 'nice'

Replica 1

Replica 2

Master
Cache Write-Through Dilemma

Replica 1
Guido = 'jerk'

Replica 2
Guido = 'nice'

Master
Guido = 'smart'

Cache Write-Through Dilemma

Replica 1
Guido = 'jerk'

Replica 2
Guido = 'nice'

Master
Guido = ???

Myth #6: Don't need to integrate data -- use Web Services

- Back to the future!
- Return to the "Balkanization" of data silos!
- Encapsulating data within an app
  - makes sense for security
  - but not within an enterprise!

App Silos vs. Integration
Who’s REALLY Doing These?

- Stock quotes
- Searching Shakespeare's plays
- Most XPath examples

More Realistic Examples

- Everything on IBM stock: price +
  - Analysts' opinions
  - News items
- A great statistic I saw a while ago (when?)...
  - In an article on the web?
  - In an e-mail from someone? Who? Folder?
  - In my Palm? Where?
  - In a presentation someone sent me?
  - In a paper I read?
  - In a file (which directory?) on my
    - development machine?
    - laptop?
My Position

- Heterogeneity will always reign
  - Format (structured, semi-structured, unstructured)
  - Schema chaos, even for structured data!
  - Schema and data are interchangeable
- A "Data Smorgasbord"
- Deal with it!
- Databases (not apps) are still the best hope for integrating data (richer modeling)

Consequences

- Will see:
  - Ad hoc "communities" for standardizing semantics of tags (like e-marketplaces)
  - Products promising integration
- Need:
  - Richer semantic models (yes, even for XML!)
  - More robust/adaptive query processing
  - Better tools for managing diversity