IR & DB: Toward Controversy and Philosophy

Joe Hellerstein

UC Berkeley &
Intel Research

Controversial Statements

Controversial Statements

- Surajit was right

Controversial Statements

- Surajit was right
- I am an expert on this stuff

Controversial Statements

- Surajit was right
- I am an expert on this stuff
  - Led integration of AltaVista and Cohera federated DB
  - Keyword search, Q-grams, auto-taxonomization, visual data cleaning & integration, etc.
    - With consulting from Marti Hearst (IR meets DB!)
  - [H & S, SIGMOD '01 industrial track]

Controversial Statements II

- This is mostly busyness, some re-search

Controversial Statements II

- This is mostly busyness, some re-search
  - Identify BKMs
  - Perform SI
  - Micro$oft will work on TCO

Controversial Statements II

- This is mostly busyness, some re-search
  - Identify BKMs
  - Perform SI
  - Micro$oft will work on TCO
- N$F?

- SO: Let's "Uplevel the Dialogue"

Let's Get Philosophical

- We have a narrow definition of structure
- Structure is not about Relational Databases vs. Text Databases vs. "semi-structured"
  - In fact, even text is very structured!

Let's Get Philosophical

- We have a narrow definition of structure
- Structure is not about Relational Databases vs. Text Databases vs. "semi-structured"
  - In fact, even text is very structured!

- Human discourse based on "deep structure"
  - Ferdinand De Saussure, the father of Structural Linguistics
  - Extended by Chomsky
  - Also speak of "relational" nature of language

So what's the philosophical diff?

- Twofold
  - The SOURCE of the structure
  - The USE of the structure

Relational Databases

- ENGINEERED STRUCTURE
  - The design of structure is at the heart of the discussion
  - Codd's lessons: simple design for robust evolution
- LANGUAGE AS PROGRAMMATIC INTERFACE
  - Semantically strict queries provide predictable results
  - Suitable for computer interpretation
    - Programmer can reason about invariants
  - Hence good for embedding into application code
    - Relatively few users interface directly to a Database!

Information Retrieval

- "FOUND" STRUCTURE
  - Take a pile of information, and extract structure therefrom
- HUMAN-CENTRIC EXPLOITATION
  - Rough understanding of query intent
  - Interface more important than query language
    - User can browse/filter/interpret some results
  - Requires a human in the loop
    - Relatively few programs embed IR techniques invisibly

The Synergy

- This is not about semi-structured data!!!

It's not that semi-structured is bad...

It's not that semi-structured is bad...

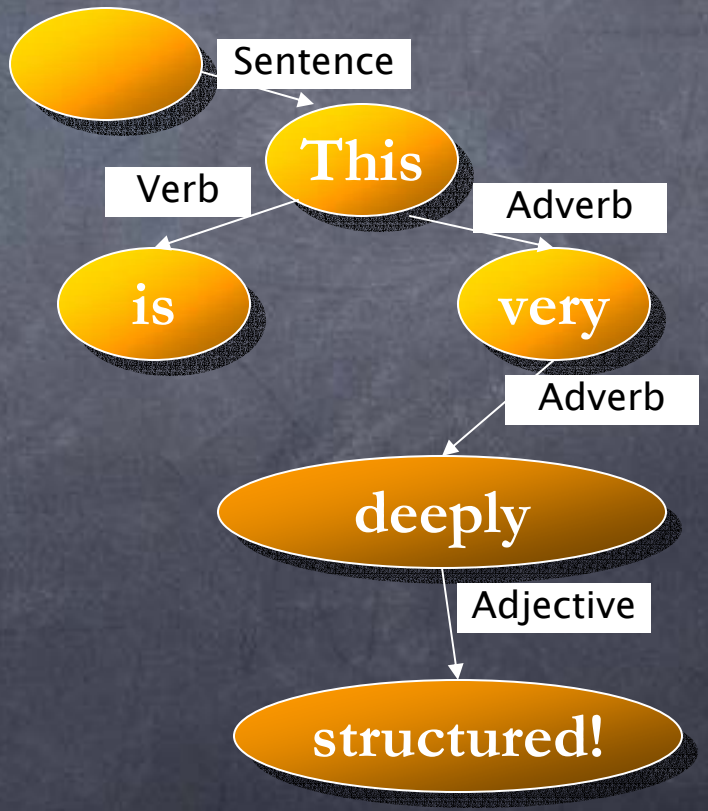
It's just that semi-structured is not semi-structured

It's not that semi-structured is bad…

It's just that semi-structured is not semi-structured

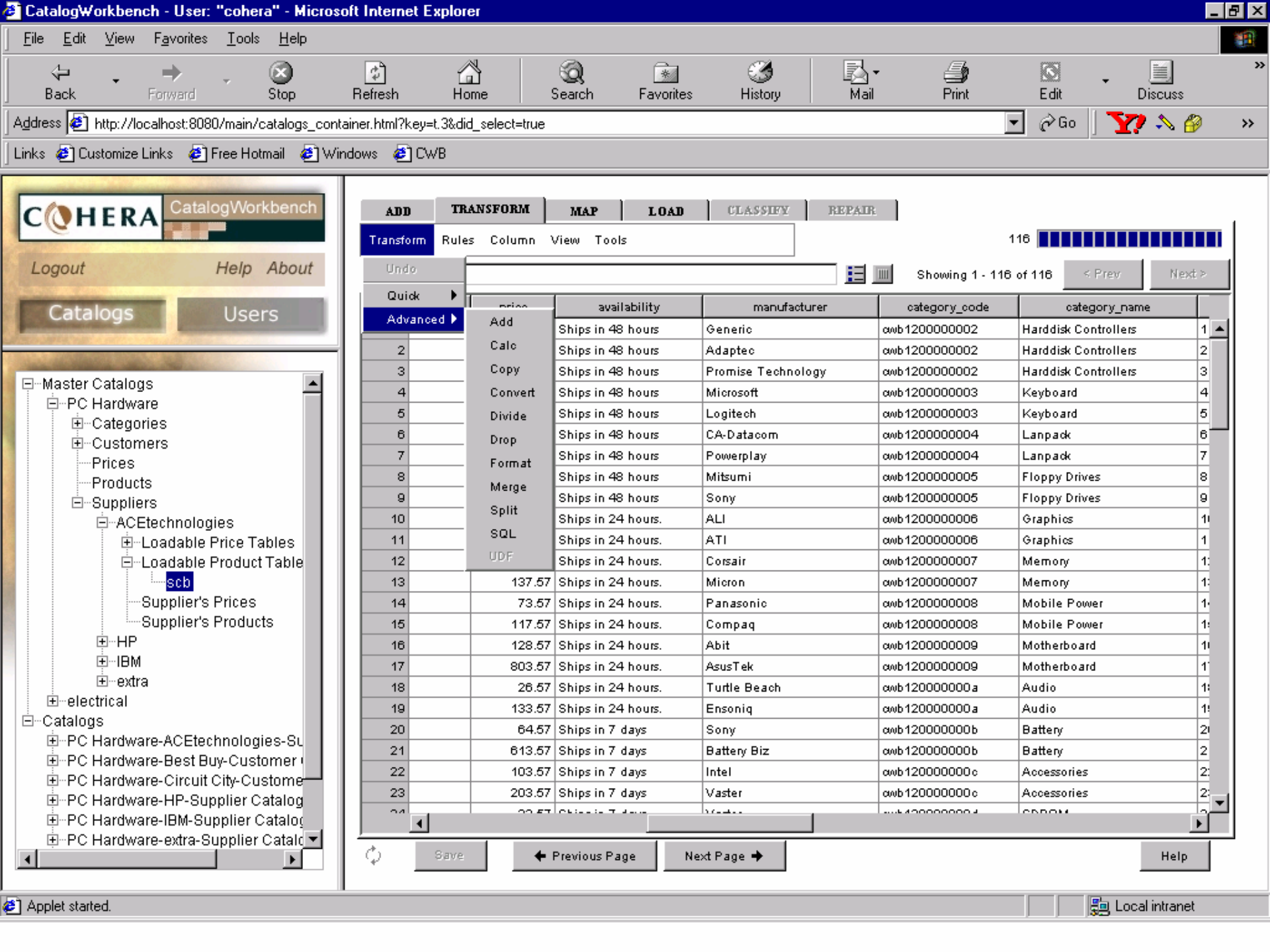
It's not that semi-structured is bad...

It's just that semi-structured is not semi-structured



The Synergy

- DB folk working on FOUND STRUCTURE
  - E.g. text queries on "structured" data
  - Information extraction
  - User-centric data mining
  - Schema corpi
- DB folk working on HUMAN INTERFACES for handling uncertainty
  - E.g. keyword search of structure data
  - E.g. online aggregation, visual data cleaning cube navigation, other user-centric mining tasks


CatalogWorkbench - User: "cohera" - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss
Address http://localhost:8080/main/catalogs_container.html?key=t.3&did_select=true
Go
Links Customize Links Free Hotmail Windows CWB
COHERA CatalogWorkbench
Logout Help About
Catalogs Users
Master Catalogs
PC Hardware
Categories
Customers
Prices
Products
Suppliers
ACEtechnologies
Loadable Price Tables
Loadable Product Table
scb
Supplier's Prices
Supplier's Products
HP
IBM
extra
electrical
Catalogs
PC Hardware-ACEtechnologies-Su
PC Hardware-Best Buy-Customer
PC Hardware-Circuit City-Custome
PC Hardware-HP-Supplier Catalog
PC Hardware-IBM-Supplier Catalo
PC Hardware-extra-Supplier Catalo
ADD TRANSFORM MAP LOAD CLASSIFY REPAIR
Transform Rules Column View Tools
116
Undo
Showing 1 - 116 of 116
< Prev Next >
Quick
Advanced
Add
Calc
Copy
Convert
Divide
Drop
Format
Merge
Split
SQL
UDF
price availability manufacturer category_code category_name
1 Ships in 48 hours Generic cwb1200000002 Harddisk Controllers 1
2 Ships in 48 hours Adaptec cwb1200000002 Harddisk Controllers 2
3 Ships in 48 hours Promise Technology cwb1200000002 Harddisk Controllers 3
4 Ships in 48 hours Microsoft cwb1200000003 Keyboard 4
5 Ships in 48 hours Logitech cwb1200000003 Keyboard 5
6 Ships in 48 hours CA-Datacom cwb1200000004 Lanpack 6
7 Ships in 48 hours Powerplay cwb1200000004 Lanpack 7
8 Ships in 48 hours Mitsumi cwb1200000005 Floppy Drives 8
9 Ships in 48 hours Sony cwb1200000005 Floppy Drives 9
10 Ships in 24 hours. ALI cwb1200000006 Graphics 10
11 Ships in 24 hours. ATI cwb1200000006 Graphics 1
12 Ships in 24 hours. Corsair cwb1200000007 Memory 1
13 137.57 Ships in 24 hours. Micron cwb1200000007 Memory 1
14 73.57 Ships in 24 hours. Panasonic cwb1200000008 Mobile Power 1
15 117.57 Ships in 24 hours. Compaq cwb1200000008 Mobile Power 1
16 128.57 Ships in 24 hours. Abit cwb1200000009 Motherboard 1
17 803.57 Ships in 24 hours. AsusTek cwb1200000009 Motherboard 1
18 26.57 Ships in 24 hours. Turtle Beach cwb120000000a Audio 1
19 133.57 Ships in 24 hours. Ensoniq cwb120000000a Audio 1
20 64.57 Ships in 7 days Sony cwb120000000b Battery 2
21 613.57 Ships in 7 days Battery Biz cwb120000000b Battery 2
22 103.57 Ships in 7 days Intel cwb120000000c Accessories 2
23 203.57 Ships in 7 days Vaster cwb120000000c Accessories 2
Save
Previous Page
Next Page
Help
Applet started.
Local intranet

The Synergy

- IR folk working on ENGINEERED STRUCTURE
  - E.g. document design
  - But the simpler, the better (Codd)!!
- IR folk working on embedded systems
  - Google SOAP interface clients (Googlisms)
  - ??
- Frankly, I'm out of my comfort zone

For More Rambling on this Topic

- See the WebDB keynote on my home page
  - http://www.cs.berkeley.edu/~jmh

The Inktomi Search Engine