Some Data Integration and Database Issues in E-Commerce (and world peace)

Ashish Gupta
Amazon.com

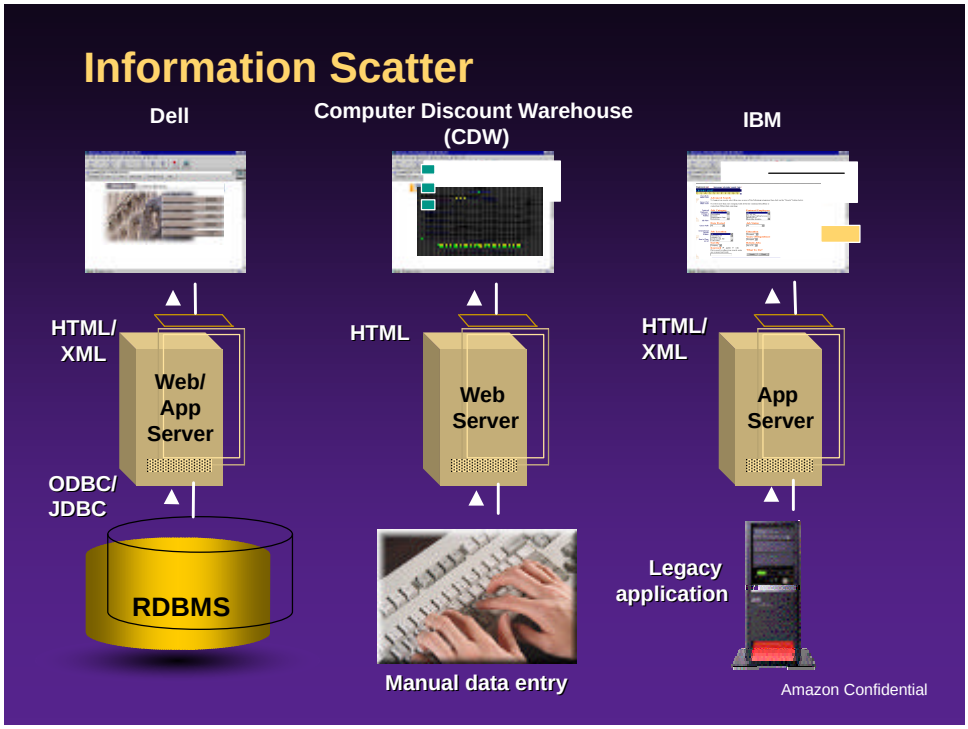
Overview

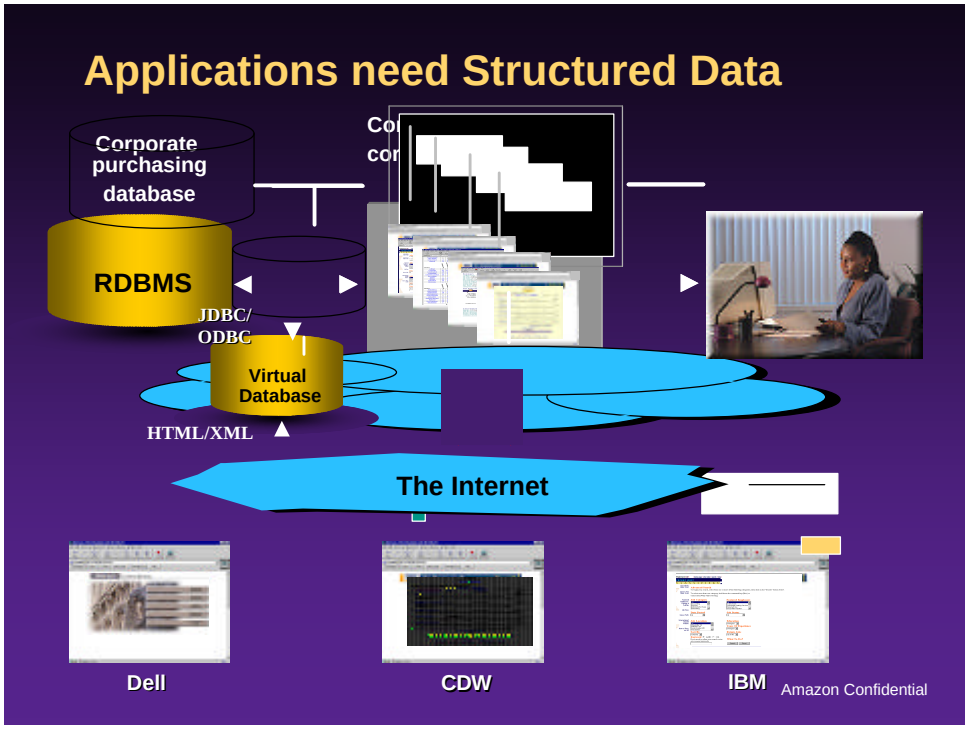
- Background (Junglee and Amazon)
- Example driven issues
  - objects - not relations
  - availability not consistency
  - query not transaction
  - “like” not “equal to”
  - limited query set not “adhoc querying”
  - ranked searching
  - incomplete query processing
  - Caching and materialized views

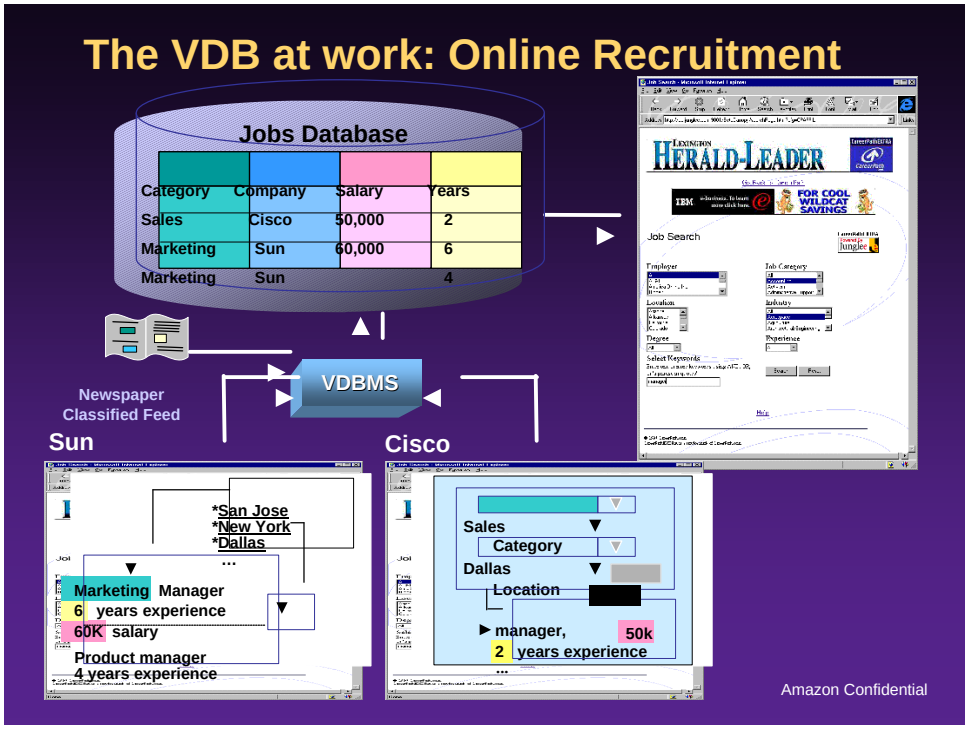
Rules of engagement

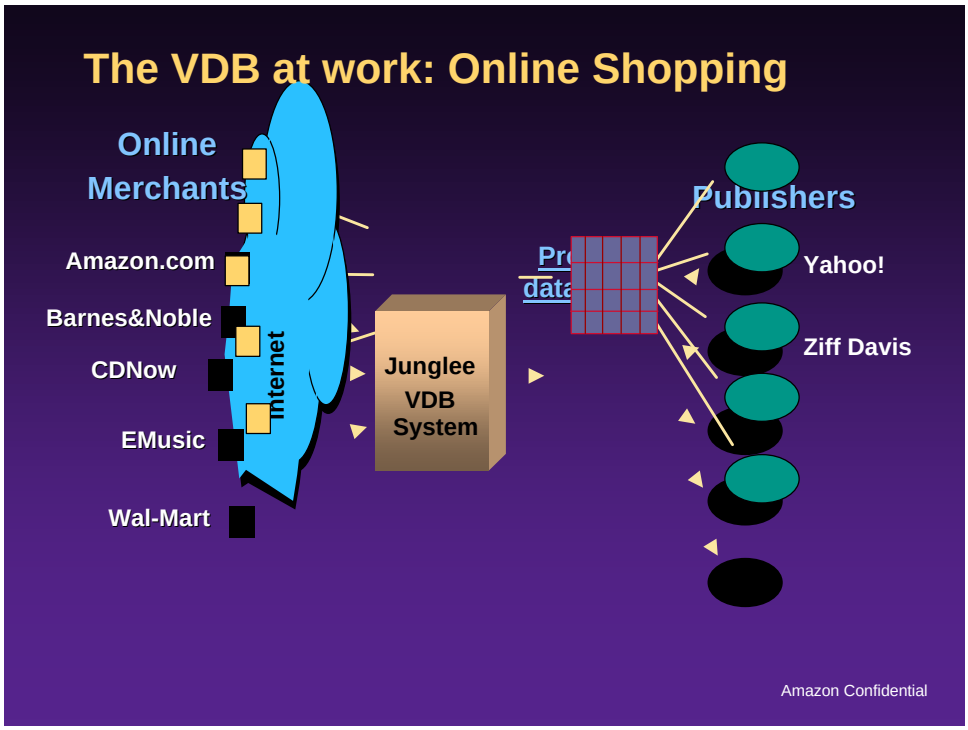
- I promise not to wake you up if you promise not to wake me up
- Please ask questions anytime
- This is one perspective - please add others

Junglee Background


Information Scatter
Dell
Computer Discount Warehouse (CDW)
IBM
HTML/ XML
HTML
HTML/ XML
Web/ App Server
Web Server
App Server
ODBC/ JDBC
RDBMS
Manual data entry
Legacy application
Amazon Confidential


Applications need Structured Data
Corporate purchasing database
RDBMS
JDBC/ ODBC
Virtual Database
HTML/XML
The Internet
Dell
CDW
IBM
Amazon Confidential


The VDB at work: Online Recruitment
Jobs Database
Category Company Salary Years
Sales Cisco 50,000 2
Marketing Sun 60,000 6
Marketing Sun 4
Newspaper Classified Feed
VDBMS
Sun
Cisco
*San Jose
*New York
*Dallas
Marketing Manager
6 years experience
60K salary
Product manager
4 years experience
Sales
Category
Dallas
Location
manager, 50k
2 years experience
Amazon Confidential


The VDB at work: Online Shopping
Online Merchants
Amazon.com
Barnes&Noble
CDNow
EMusic
Wal-Mart
Internet
Junglee VDB System
Publishers
Yahoo!
Ziff Davis
Amazon Confidential

Technical Challenges

- Organization
  - Web sites on a single topic vary widely in how they organize their data.
- Unstructured text
  - Web sites embed key data in unstructured text.
- Vocabulary differences
  - Web sites use several terms to denote the same data item.
- Reliability
  - Web sites are often unreachable
- Autonomy
  - Web sites change without notice
- Instance (not schema) level application development

Amazon.com

Some facts

- Information scatter - many catalogs come together
- 50 times more queries than orders
- Users have personalities and personas
- sub-second response times
- catalogs come in at all times of day

Lessons??

Objects not relations

- Multi-valued attributes
  - a “shirt” can be in multiple “sizes”
  - relational model is not the best
- Need for ID
  - differentiate one “shirt” from another
  - enable alert services - as new products become available, inform customers
  - Correlate information from other sources
- Hierarchical, not flat, access patterns

Query not transaction

- Query data is frequently separate from transaction data
- Query response times are much more critical than update times given query volumes are much higher than transactions

Limited queries - not “adhoc” queries

- Applications expose queries
- Little flexibility to go outside of fixed navigation paths
- queries driven by “consumer experience” - unsophisticated customers
- Queries need to be optimized a-priori

Incomplete queries

- Users do not necessarily want all answers
- Users want to see first few answers quickly
- Users hit “stop” and further manipulate the result set

“like” not “equal to”

- Web applications need “inexact” matches
- Users seldom know the exact format to enter search terms (case, spaces, spelling, etc.)
- If user asks for shirts that cost less than $50, they will also accept those that are $51
- What if three of four “conjunctions” are satisfied?
- Null values may sometimes need to be false and other times true - depending on the applicatio

Caching and materialized views

- Because of huge query volume, queries are repeated
- limited query sets - subset of DB relevant
- Based on user profiles, different views exposed
- Quick response times crucial

Tightly integrated keyword searching

- Most web data has text component
- Users are un-sophisticated and need simple interfaces - often keyword box
- Drill down often needs relational search

Ranking is important

- Results are not equally important
- Ranking is not just algorithmic
  - other users' input
  - same users historical behaviour
  - input from experts in the area
  - input from vendors

Unclear separation of schema and instance

- Applications are aware of values in the database
  - example: pull down menu values tightly controlled by “editors”
- Tolerance for “garbage” values is very low
- Separation between “display” values and “database instance”
  - example: display value may be “MD” whereas database has “maryland”

Incremental updates to Databases

- Products are added to “catalog” incrementally
- Indexes need near-real-time incremental updates
  - example, some new products became available
  - quantity of product changed
  - price of product changed due to discount

Attribute values are more complex

- Some attributes are functions of others
  - discount price is a function of “buyer”, “original price”, “order volume”
- An attribute may be visible under specific condition
- Many attribute values may be null

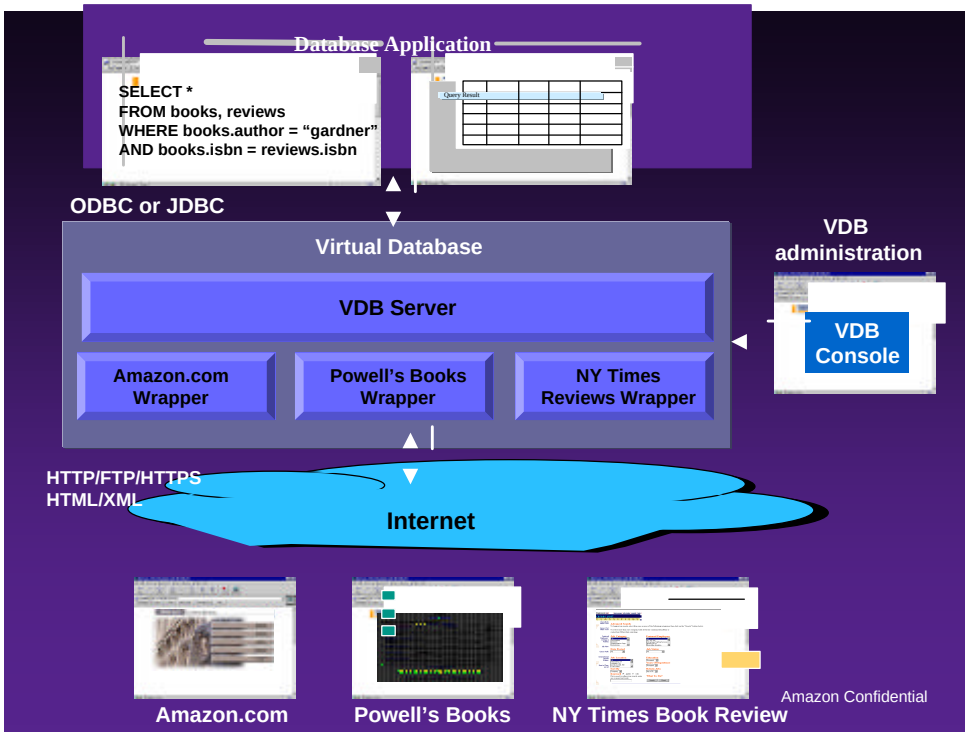
Availability not consistency

- Web applications need to be available 24 hours
- Information need not be complete or consistent

Data Cleansing needs

- Data comes from non-transaction sources
- Terminology varies between sources

Need a new database system


Database Application
SELECT *
FROM books, reviews
WHERE books.author = "gardner"
AND books.isbn = reviews.isbn
ODBC or JDBC
Virtual Database
VDB Server
Amazon.com Wrapper
Powell's Books Wrapper
NY Times Reviews Wrapper
VDB administration
VDB Console
HTTP/FTP/HTTPS
HTML/XML
Internet
Amazon.com
Powell's Books
NY Times Book Review
Amazon Confidential