Data Integration

- Explosion of intranet and extranet information
- 80% of corporate information is unmanaged
- By 2004 30X more enterprise data than 1999
- The average company:
  - maintains 49 distinct enterprise applications
  - spends 50% of total IT budget on integration-related efforts

Source: Alon Halevy: Structures, Semantics and Statistics. Keynote at VLDB’04
The Integration Challenge

- Complex and heterogeneous environments
  - Many different types of systems
  - Many inter-related applications
- Escalating needs
  - Variety, velocity, volume
  - People are expensive

"DB World"

Physical Integration
(Data Warehousing)

- Meta data
- Analysis Tools
- Data Marts
- Data Warehouse
- Import (ETL)
- Operational Systems

Virtual Integration
(Query Mediators)

- Client 1
- Mediator
- Wrapper 1
- Source 1
- Wrappers
- Sources

Source: H. Ho: Model management tutorial. VLDB 2007
**General search engine architecture**

*Arasu et. al.: Searching the Web. ACM Trans. Internet Techn, 2001*

**Integration "Sextant"**

- **Data quality issues for many sources/instances**
- **High setup effort**

<table>
<thead>
<tr>
<th>Instance Integration</th>
<th>Query-Mediators</th>
<th>P2P Systems; Meta Entity Search</th>
<th>Meta Search Engines</th>
</tr>
</thead>
<tbody>
<tr>
<td>virtual / hybrid</td>
<td>Data Warehouses</td>
<td>Entity / XML Search Engines</td>
<td>Search Engines</td>
</tr>
<tr>
<td>physical</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>high</td>
<td>high</td>
<td>medium</td>
<td>low</td>
</tr>
</tbody>
</table>

**Semantic (Schema) Integration**

- **Scalability**
  - low
  - high
Dataspaces*

- Improved scalability compared to "schema first" integration
- Data co-existence approach
- Dataspace = set of participants + relationships (mappings)
  - Participants
    - data sources (RDB, XML, files, web services, ...)
    - differences w.r.t. structure, updates, queries
  - Relationships
    - "is view", "schema mapping", "created independently", ...
- Heterogeneous Services
  - catalog & browse, search & query, index, ...

* Franklin et.al.: From databases to dataspaces: a new abstraction for information management. SIGMOD Record, 2005

PayGo: Architecture

[Madhavan et.al., CIDR'07]

- Web-scale
- Multiple domains
- Exploit structures & mappings for query answering
  - Clustering
  - Approximated Mappings

Extracted structure, metadata, data

Data sources (Web, deep web, Google Base, annotations, etc.)
(Some) Problems of current data integration approaches

• **Setup time too high**: crawling; schema mapping / integration …

• Current data integration approaches are *query-focussed*: search engine queries, query mediators, warehouse access

• **Queries are not enough**: complex data integration problems cannot easily be solved in 1 query / search
  – What is the most cited XSym paper so far?
  – Which famous scientists lived close to the VLDB 2007 venue

• Data quality for heterogeneous/dirty web data and query results

• Execution time for dynamic fusion of larger data sets

Workflow-like data integration

• "You only have 1 day - how far can you go to solve a data integration problem?"

• Reuse + Combine existing (data) services within **data integration workflows**
  – Reuse existing services
  – Reuse existing data integration systems, e.g. search engines, query mediators, warehouses
  – Combine query/service results within a workflow
  – Perform data cleaning and data transformation
  – Perform data analysis

• Must be supported by a flexible data integration framework

• Workflow-like data integration complements query-based data integration
Workflow-based Integration

- **Examples**
  - *Offline*: ETL processes for Data Warehouses
  - *Online*: Workflows for analyzing biological data

- **New aspects:**
  - Combine ETL and analysis workflows (on-demand information extraction)
  - Share and reuse existing data services and tools
  - Reuse existing (entity) search engines
  - Easy development and use of workflows (-> Mashups)

---

Mashups - a light-weight data integration approach

- "A web mashup is a web page or application that combines data from two or more external online sources." (ProgrammableWeb)
- "A mashup is a web application that combines data from more than one source into an integrated experience." (Wikipedia)
- "Mashups are an exciting genre of interactive Web applications that draw upon content retrieved from external data sources to create entirely new and innovative services." (Merrill: Mashups: The new breed of Web app)
Mashup Example: Forbes List

Ranking by Forbes List of best paid celebrities

YouTube video

Hometown displayed by Google Maps

http://www.mibazaar.com/top100celebrities

Locations for conference

• Displays conference venue of VLDB 2007 as well as nearby hotels, train stations, airports, ...

VLDB 2007 Locations
Important Locations for the VLDB 2007 Conference hosted at University of Vienna, Austria
View: Public
Created by mas on Jun 27 - Updated 10:59 hours ago

Conference Venue
The conference venue will host all
Arbel Belzmann
Erlzmannasse 8, A-1030 Vienna Booking

Hotel Astenia
Kammer Strasse 32-34, A-1010 Wien Booking

Austria Trend Hotel Aranay
Kemho Munchide 99-99, A-1060 Wien

Austria Trend Hotel Europa
Kammer Strasse 16, A-1010 Wien Booking

Austria Trend Parkhotel Schmirgern
Hollinger Hauptstraße 15-30, A-1130 Wien

Search Results My Maps
Mashups: Driving forces

- AJAX (Asynchronous Javascript and XML)
  - Desktop-like look-and-feel of Web applications
- Development tools, e.g. Google Web toolkit
- Visual development tools without programming need
- Increasing number of Web services (APIs)
  - Easy access to "interesting" content and services
  - 50% of mashups use Google Maps

ProgrammableWeb

#Mashups 2300
#Mashups/Day 3
#APIs 509
09/09/2007

The Big Picture: Mashup Fabric*

* Jhingran: Enterprise Information Mashups: Integrating Information, Simply. Keynote at VLDB'06
Mashup Tools: Overview

Mashup Builders

Managing mashup components, e.g., maps, feeds

Data Transformation/
Data Aggregation

Data transformation workflows

Source Wrappers

Information Extraction

Dapper

- "Get any content from the web"
  - Information extraction from any website
  - Transform content into XML, RSS, ...
- Easy-to-use 5-step-approach

1. User defines basis URL, e.g., news.google.com
2. User generates few example pages, e.g., different search results
   (search for "iphone", "microsoft", ...)
   Dapper performs a comparative analysis for identifying parts of the same structure,
   e.g., several news entries, news headline, ...
3. User selects relevant content, e.g., headlines
4. User defines output structure, e.g., one item for each news entry
5. Save "dapp" to make it accessible via URL
Dapper: Google News as RSS feed

- "Get any content from the web"
  - Information extraction from any website into XML, RSS ...
- Automatic detection of parts ("blocks") of the same structure
  - e.g., several news entries, news headline, ...
- → User selects "blocks" for extraction

Yahoo Pipes

- Composition tool to aggregate, manipulate, and mashup web content, especially RSS feeds
  - Pipe = data transformation workflow
  - Visual specification
- Example: Aggregated News Alert
  - User input: keyword(s)
  - Parallel search at
    - Yahoo! News
    - MSN Live News ...
  - Merge
  - Sort by date
  - Deduplication (unique title)
  - Output
Yahoo Pipes: Aggregated News Alert (2)

Google News: New zero-day flows found in AOL, Yahoo IM - Tech Republic
New zero-day flows found in AOL, Yahoo IM Tech Republic, KY - 2 hours ago According to ZDNet Blogs, this makes it the third major security hiccup found in Yahoo Messenger over the last few months. Exploit code has been released...

Y! News: Thursday | 20 September, 2007 (ARNnet)
Attack code that targets Yahoo Messenger has been published on the Internet, a security researcher warned Wednesday, marking the ninth exploit aimed at the popular instant messaging software so far this year.

I'm Online! A Beginner's Guide to Instant Messaging American Chronicle, CA - 21 hours ago However you need to remember that usually not all features of Windows Live Messenger, AIM, or Yahoo messenger will be available for you if you are using...

Google News: Yahoo! Mash too Little too Late? - Uber-Review
Conads.com Yahoo! Mash too Little too Late? Uber-Review - 19 Sep 2007 Yahoo has the platform to build the dominant social network. It has an established client base with Yahoo Groups. The chat rooms of Yahoo Messenger are... Can Yahoo Mash Cut It? Search Newz Yahoo Mash takes on social...

---

IBM DAMIA

- Similar to Yahoo! Pipes
- Modules
  - Sources: URLs, Excel files
  - Operators (see right)
- Example: Aggregated News
Mashups: Characteristics

- Easy and fast development
  - Visual programming (drag & drop) or integration with power development environments (e.g., Eclipse)
- Service-oriented paradigm
  - Sharing and reuse of web services
- Web2.0 interfaces
- Standardized XML-based data formats, e.g. RSS, SOAP/REST (data exchange)
- Simple processing workflows
- Simple *instance-based data integration*
  - Geographical coordinates
  - Keywords (e.g. names)
- Simple keyword queries dominate (no query transformation)
- Limited result postprocessing (primarily merge instead of match)

Mashups: Query example

- Keyword queries
- "Merge instead of match"
Mashups: Open Problems

• More complex queries, e.g. for heterogeneous entity search engines (-> query transformation)
• Data quality
  – Precision and recall depend on developer’s choices (source selection, query formulation)
  – Typos, missing/wrong attribute values (e.g., due to extraction errors)
  – Duplicates, i.e., sources contain multiple instances for the same (real world) object
• Performance for large data volume (automatic optimization)
• Semantic repository of services
  – Service description & service discovery
• Support for business applications, e.g. security restrictions

Entity search engine (GS): quality problems

Duplicates due to
• Extraction errors (title, authors)
• Different titles
• Typos (author name)
• Heterogeneous venue names
• Missing / additional authors (!)

Heterogeneous venue names
• How to query for "VLDB '97"?
Information Fusion with iFuice [RTA+05]

• Generic data integration platform for structured and unstructured data sources
  – Query / search / id-based data access
• Workflow-like data integration with operator-based programming model
  – Generic high-level operators for use within script programs
  – Example: query traverse, map, union, aggregate,
• Utilization of instance-level mappings
  – Correspondences between object instances
  – Represent semantic relationship ("is same", "is associated to")
• Metadata repository for data sources and services
  – Semantic object (e.g., Author, Publication) and mapping types
• Iterative query strategies
• On-the-fly object matching

Mashup Framework: Architecture
On-the-fly object matching

- Object matching is an important part of data integration
  - prerequisite for information fusion
  - Example: group together multiple Google Scholar entries

- Goals
  - seamless integration in data integration workflows
  - effective & efficient

Many object matching approaches ...

Object matching approaches

- Value-based
  - Single attribute
    - unsupervised
      - Aggregation function with threshold
      - User-specified Rules:
        - Hernandez et al. (SIGMOD 1995)
      - Clustering
        - Monge, Elkan (DMKD 1997)
        - Mc Callum et al. (SIGKDD 2000)
        - Cohen, Richman (SIGKDD 2002)
    - supervised
  - Multiple attributes
    - unsupervised
    - supervised

- Context-based
  - Hierarchies:
    - Ananthakrishna et al. (VDLB 2002)
  - Graphs:
    - Bhattacharya, Getoor (DMKD 2004)
  - Dong et al. (SIGMOD 2005)
  - Ontologies

- Decision trees
  - Verykios et al. (Information Sciences 2000)
  - Tejada et al. (Information Systems 2001)
  - Support Vector Machine
    - Bilenko, Mooney (SIGKDD 2003)
    - Minton et al. (2005)
Many data cleaning frameworks ...

• Research prototypes
  – AJAX (Galhardas et al., VLDB 2001)
  – IntelliClean (Lee et al., SIGKDD 2000)
  – Potter's Wheel (Raman et al., VLDB 2001)
  – Febrl (Christen, Churches, PAKDD 2004)
  – TAILOR (Elfeky et al., Data Eng. 2002)
  – MOMA (Thor, Rahm, CIDR 2007)
  – ...

• Commercial solutions
  – DataCleanser (EDD), Merge/Purge Library (Sagent/QM Software), MasterMerge (Pitnew Bowes) ...
  – MS SQL Server 2005: Data Cleaning Operators (Fuzzy Join / Lookup)
  – ...

Seminar
Seminarziele

- Beschäftigung mit einem praxis- und wissenschaftlich relevanten Thema
- Erarbeitung und Durchführung eines Vortrags zu einem Thema unter Verwendung wissenschaftlicher (englischer) Literatur
- Diskussion
- Schriftliche Ausarbeitung zu dem Thema
- Hilfe und Feedback durch Betreuer / Seminarteilnehmer

Seminarbedingungen

- **Scheinvergabe** / Prüfungsleistungsnachweis erfordert
  - Selbständiger Vortrag mit Diskussion
  - Schriftliche Ausarbeitung (ca. 15-25 Seiten)
    - Ausarbeitung vom Betreuer abzunehmen
    - Ausarbeitung soll zum Vortragstermin vorliegen
  - Teilnahme an allen Vortragsterminen
  - Teilnahme an Diskussion
- **Themenrückgabe**
  - In Ausnahmefällen, jedoch spätestens bis 31.10.2007
  - Ansonsten: erfolglose Teilnahme (Note 5)