Dataspaces: A New Abstraction for Data Management

Mike Franklin, Alon Halevy, David Maier, Jennifer Widom
Today’s Agenda

• Why databases are great.
• What problems people really have
  ✷ Why databases are not great.
• Data integration and sharing:
  ✷ Nice, but doesn’t address all the problem.
• Dataspaces:
  ✷ Initial concepts, a note on politics
  ✷ Research challenges
Databases Are Great

- Very clean abstraction for data management.
- High-level querying with efficient query processing.
- Strong guarantees. Your data will survive anything.
- *Put your data in the database*, and your worries will go away.
Today’s DM Challenges

• A set of inter-related data sources:
  - The enterprise
  - Large science projects
  - Government agencies
  - The battlefield
  - The desktop (and its extensions)
  - A library
  - The ‘smart’ home

• We’ve heard this before. What’s new?
A Quick History of Data Integration

- Until late 90’s:
  - Integration by warehousing
  - Integration by custom code
- Late 90’s (boom years):
  - Virtual data integration (data stays at the source, queried on the fly)
  - Nimble, Cohera and others.
  - EII (Enterprise Information Integration): new buzzword. Still buzzing now too.
Virtual Data Integration

Mediated Schema

Independence of:
- source & location
- data model, syntax
- semantic variations

Semantic Mappings

Query

SSN Name Category
123-45-6789 Charles undergrad
234-56-7890 Dan grad

CID Name Quarter
CSE444 Databases fall
CSE541 Operating systems winter

<title> The best of ... </title>
<artist> Carreras </artist>
<artist> Pavarotti </artist>
<artist> Domingo </artist>
<price> 19.95 </price>
Peer Data Management Systems

Q1 UW → The other UW
Q3 The other UW → Stanford
Q4 Stanford → Q4
Q5 Stanford → DBLP
Q6 DBLP → CiteSeer
Q7 CiteSeer → Berkeley
Q8 Berkeley → U. Toronto
Q9 U. Toronto → Q2
Q10 Q2 → LAV, GLAV
Q11 LAV, GLAV → Q1
Q12 Q1 → UW
DI: Nice but Limited

• Still thinking about it like DB people.
• You can only manage data if it is:
  ❖ Explicitly put in the database (or some source)
  ❖ Fully mapped to the mediated schema.
• Upfront cost is too high:
  ❖ Benefits not always clear at the outset.
Mike’s First Figure

% Functional

Time (or cost)

Dataspaces

Schema First
Mike’s Second Figure

<table>
<thead>
<tr>
<th>Administrative Proximity</th>
<th>High Semantic Integration</th>
<th>Low Semantic Integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Far</td>
<td>DBMS, Federated DBMS</td>
<td>Web Search</td>
</tr>
<tr>
<td>Near</td>
<td>Virtual Organization</td>
<td>Desktop Search</td>
</tr>
</tbody>
</table>
Bernstein’s Story
List my CSE 444 students from last year

Find the budget for my NSF SEIII Grant
Find the experiments run an hour before the SIGMOD deadline. What *were* we thinking?
Alon’s First Figure

A Dataspace
Participants: Examples

- Structured databases (relational, XML)
- Files of various applications
- Code collections
- Web services, software packages
- Sensors

- Different query capabilities
- Some updateable, others not
- Some more structured than others
- May stream
Relationships: Examples

- Full schema mappings
  - E.g., views of each other, replicas
- A was manually created from B and C
- A is a snapshot of B on a certain date
- A and B reflect the same underlying physical entity (but are different)
- A was sent to me at the same time as B.
Dataspace Services

- Search & query: on data, schema, meta-anything.
  - Query lineage, hypothetical queries, ...
- Mining.
- Set up workflows.
- Monitoring for special events.
- Soft constraints, recovery, consistency, ...
The Dataspace System (DSS)

- Participant and relationship discovery
- Search & Update
- Dataspace admin:
  - recovery
  - replication, ...
- Catalog:
  - participants
  - relationships
- DSS local store and index
A Note on Politics

• RDBMS have been a great identity
  ✷ But has it served its purpose?
  ✷ We’ve moved on, but the external perception hasn’t.
  ✷ Too much alcohol served at CIDR.

• Dataspaces could be a new identity
  ✷ 80% of our work is already on it anyway
  ✷ Some exciting new problems (next)
  ✷ “Because that’s the size of the problem”
Challenges: Search/Query

• What does search mean over a heterogeneous collection? Ranking?
• Answer queries despite schema heterogeneity and with no mappings.
• Support spectrum of search to query
  ◦ Given keywords, identify what db may be relevant.
• No single data model, not even mediated.
Challenges: Lineage and Uncertainty

• When everything is fluffy, life is uncertain.
• Need to model:
  - Uncertainty and lineage \textit{and} the relationship between them.
  - Hypothetical queries.
  - Different types of uncertainty:
    - Is it in the data?
    - Is it a result of approximate integration and translations?
Indexing a Dataspace

• Build a heterogeneous index on *everything*.
• Think: Google desktop, but with clever indexing of (semi)-structured sources.
• Resolve multiple references to objects in the dataspace.
• Materialize some of the data for faster access.
Dataspace Discovery

• What do I have in my enterprise??
• Tasks:
  ✷ Find the sources and classify them.
  ✷ Suggest mappings between sources.
  ✷ Suggest which sources may be related.
  ✷ Maintain this over time.
  ✷ Create associations between data items.
Consistency and Recovery

- Mike?
Reuse, Reuse and Reuse

• Reuse any human effort related to a dataspace.
• First example:
  - Reuse schema mappings
  - E.g., everyclassified.com includes 4500 mappings. Reuse was key.
• Next steps:
  - Reuse other human annotations
  - Reuse for more removed tasks.
Summary

Dataspaces -- because:

- That’s the size of the problem
- The field needs funding
- There is a ton of exciting stuff to do