590q: The Database Seminar
Fall 2013

Query Languages for Nested Data Models
Dan Suciu
1\textsuperscript{st} Normal Form

- 1\textsuperscript{st} normal form says that all tables are flat
- Why? Data independence principles

- “Normalization” means taking your data and shredding it into flat tables

- Consequence: databases have joins galore...
Dremmel paper (google):

• The data used in Web and scientific computing are often non-relational. [...] Data structures [...] lend themselves naturally to a nested representation. Normalizing and recombining such data at Web scale is usually prohibitive. A nested data model underlies most of the structured data processing at Google and reportedly at other major Web companies.

In other words: “de-normalize the data to avoid joins”
**NFNF (or NF²)**

<table>
<thead>
<tr>
<th>DocID</th>
<th>Date</th>
<th>FullText</th>
</tr>
</thead>
<tbody>
<tr>
<td>222</td>
<td>2007</td>
<td>In this paper we showed ...</td>
</tr>
<tr>
<td>123</td>
<td>2012</td>
<td>In the previous paper we showed</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**1NF**

<table>
<thead>
<tr>
<th>DocID</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>222</td>
<td>In</td>
</tr>
<tr>
<td>222</td>
<td>this</td>
</tr>
<tr>
<td>222</td>
<td>paper</td>
</tr>
<tr>
<td>222</td>
<td>we</td>
</tr>
<tr>
<td>222</td>
<td>show</td>
</tr>
<tr>
<td>123</td>
<td>In</td>
</tr>
<tr>
<td>123</td>
<td>the</td>
</tr>
<tr>
<td>123</td>
<td>previous</td>
</tr>
<tr>
<td>123</td>
<td>paper</td>
</tr>
<tr>
<td>123</td>
<td>we</td>
</tr>
<tr>
<td>123</td>
<td>showed</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>
Ancient History of NF²

• VERSO project at INRIA, circa 1982

• S. Abiteboul and N. Bidoit, Non-first normal form relations: an algebra allowing data restructuring, 1986
  – COURSE(STUDENT)*(BOOK)*

• Schek, Scholl: The relational model with relation-valued attributes. 1986
  – NF²

• S.J. Thomas and P.C. Fischer, Nested relational structures, 1986
Ancient History of NF$^2$

Early papers searched for a query language

- 1NF: query language is either FOL or RA
- NF$^2$: what is a natural query language?

- FOL is ill suited for nested relations
- RA is better:
  - flat RA: $\times$, $\sigma$, $\Pi$, $\cup$, -
  - nest: $\{A \times B\} \rightarrow \{A \times \{B\}\}$
  - unnest: $\{A \times \{B\}\} \rightarrow \{A \times B\}$
1-2. Principles

Early 90’s at Penn:
- Buneman, Tannen, Wong redesigned a query language from first principles – category theory
- Main construct:
  
  \[
  \text{from } f : A \rightarrow B \text{ to } \text{map}(f) : \{A\} \rightarrow \{B\}
  \]
  
  variation: \( f : A \rightarrow \{B\} \text{ to } \text{ext}(f) : \{A\} \rightarrow \{B\} \)

Example: nested_join

\[
\{A \times \{B \times C\} \times \{B \times D\}\} \rightarrow \{A \times \{B \times C \times D\}\}
\]

Papers: Naturally embedded; Comprehension; Wadler’s Comprehending Monads

Discussion: design principles; minimal set of operators; ...
3. Case Studies

- Dremmel/Big-Query, AQL, Jaql, ...

- Some treats nested relations as second-class citizens. E.g. Big-Query:
  - Group-aggregation v.s. scoped aggregation
  - Can join main tables, but not nested tables

Papers: Dremmel, Asterix QL (AQL), Pig Latin

Discussion: how natural can they express queries on nested tables?
4. Implementation

- Most systems “flatten” nested collections
- Naturally leads to column-oriented storage
- ...and compression

Papers: Dremmel (2) C-Store, XMill
5. Conservativity

• Back to theory. Recall that FOL has limited expressive power (no transitive closure, no parity). Do we get more expressive power if we use nested relations?

\[
\{A \times B \times C\} \rightarrow \{A \times C \times \{B \times C\}\} \rightarrow \{B \times C\}
\]

• Answer: no! [Paredaens&Van Gucht]

Nested Relational Algebra is a conservative extension of the Relational Algebra

Paper: Wong’s proof using rewriting

Discussion: practical implications
6. Nested Relations and Iteration

• Q: What if we combine nested relations + iteration?
• A: you can compute powerset!
  powerset: \{A\} \rightarrow \{\{A\}\}
• Also: conservativity theorem no longer holds
• Lesson: you don’t don’t want to do that

• However, if you add bounded iteration then the conservativity theorem still holds
• Question: is this the right language design?

Paper: Bounded fixpoint
Discussion: alternate proof of conservativity
7. Parallelism

• Writing a user defined aggregate:
  \[ \text{agg} : \{A\} \rightarrow B \]

• Two ways:
  combine: \[ B \times A \rightarrow B \]
  or merge: \[ B \times B \rightarrow B \]

• It turns out that the former captures PTIME, the latter captures NC

Paper: A Query Language for NC
Discussion: automatic rewriting combine to merge?
8. While Languages

- GraphLab, Pregel consists of a while-loop plus (a-)synchronous updates
- “Updates” are key constraints:
  If $R(A, B)$ has key $A$, then:
  $R(x,y) :-$ some expression
  could mean “replace $y$ with new values”
- Conflicts? Asynchronous, non-deterministic
- In logic this is captured by the $W$ operator
<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Main:</th>
<th>Optional:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday, 10/7</td>
<td>Comprehension</td>
<td>Buneman et al: Comprehension Syntax. SIGMOD Record, 1994</td>
<td>A modern nested data model: Protobuf</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Wadler: Comprehending Monads, 1992 (Sec. 2, 3 only)</td>
</tr>
<tr>
<td>Monday, 10/14</td>
<td>Principles</td>
<td>Tannen et al: Naturally Embedded Query Languages. ICDT 1992 (Sec 1-4 only)</td>
<td></td>
</tr>
<tr>
<td>Monday, 10/21</td>
<td>Case Studies</td>
<td>Melnik et al., Dremel: interactive analysis of web-scale datasets. CACM 2011</td>
<td>The Asterix Query Language</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Olston et al: Pig latin: a not-so-foreign language for data processing, SIGMOD 2008</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Jaql</td>
</tr>
<tr>
<td>Monday, 10/28</td>
<td>Implementation</td>
<td>Melnik et al, Dremel: Interactive Analysis of Web-Scale Datasets. PVLDB 2011</td>
<td>Abadi et al, Column-stores vs. row-stores: howdifferent are they really? SIGMOD 2008</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Liefke, Suciu: XMILL: An Efficient Compressor for XML Data. SIGMOD 2000</td>
</tr>
<tr>
<td>Monday, 11/4</td>
<td>Conservativity (1)</td>
<td>Wong: Normal Forms and Conservative Properties for Query Languages over Collection Types. PODS 1993</td>
<td></td>
</tr>
<tr>
<td>Monday, 11/11</td>
<td></td>
<td>-- external speakers</td>
<td></td>
</tr>
<tr>
<td>Monday, 12/2</td>
<td>While-Language</td>
<td>Abiteboul, Vianu: Fixpoint Extensions of First-Order Logic and Datalog-Like Languages. LICS 1989</td>
<td></td>
</tr>
</tbody>
</table>