590q: The Database Seminar
Fall 2013

Query Languages for Nested Data Models

Dan Suciu

1st Normal Form

- 1st 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…

Modern Big Data Systems

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 NF^2)

DocID	Date	FullText
222	2007	In this paper we show …
123	2012	In the previous paper we showed …
…	…	

1NF

DocID	Date
222	2007
123	2012
…	…

DocID	Text
222	In
222	this
222	paper
222	we
222	show
	…
123	In
123	the
123	previous
123	paper
123	we
123	showed
	…

Ancient History of NF^2

- 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
  - NF^2
- S.J. Thomas and P.C. Fischer, Nested relational structures, 1986

Stretch the relational query language to deal with nested relations

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:

  from $f : A \rightarrow B$ to $map(f) : \{A\} \rightarrow \{B\}$

  variation: $f : A \rightarrow \{B\}$ to $ext(f) : \{A\} \rightarrow \{B\}$

  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 treat 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

DocID	Date	FullText
222	2007	In this paper we show …
123	2012	In the previous paper we showed …
…	…	

DocID
222
123
…

Date
2007
2012
…

FullText
In
this
paper
…

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} → {{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:

  $$\text{agg} : \{A\} \to B$$

- Two ways:

  $$\text{combine}: B \times A \to B$$

  or $$\text{merge}: B \times B \to B$$

- 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:

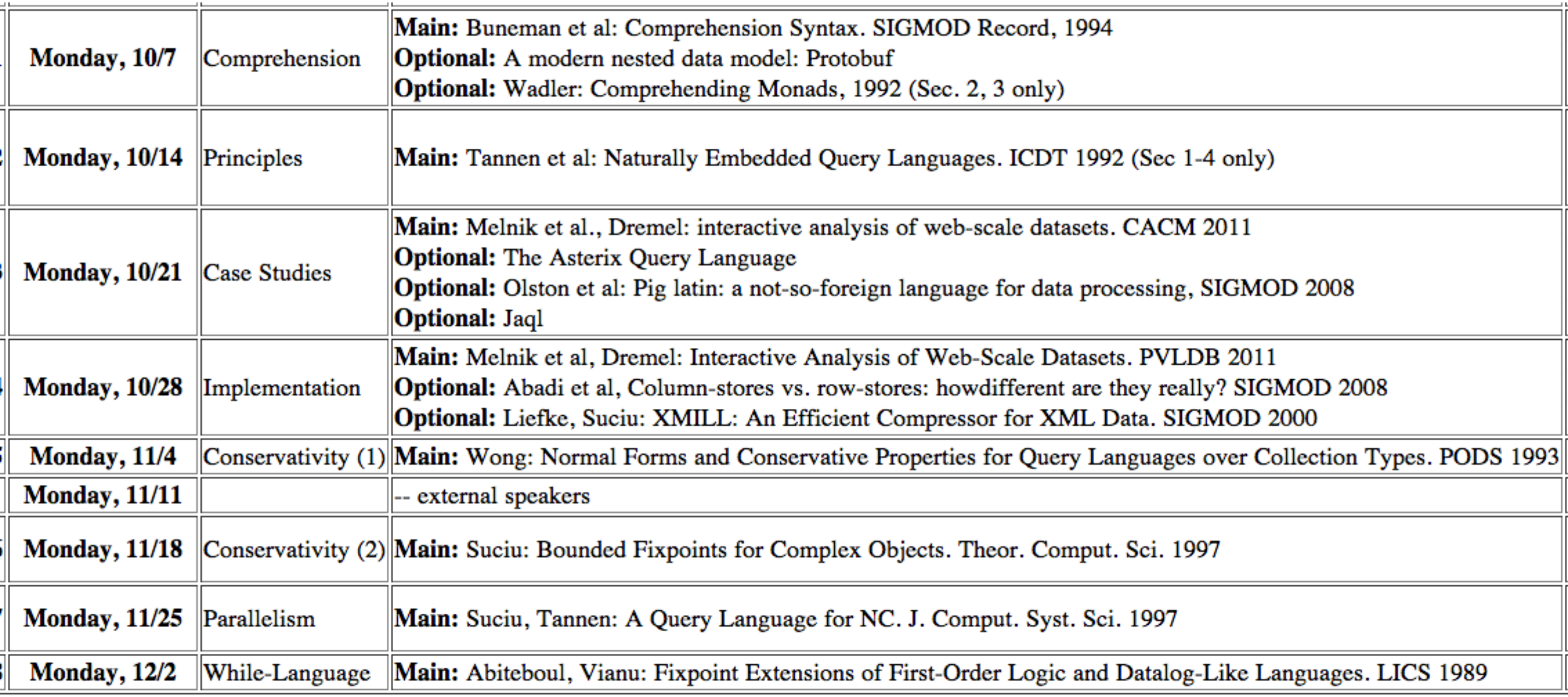
  R(x,y) :- some expression

  could mean "replace y with new values"
- Conflicts? Asynchronous, non-determinstic
- In logic this is captured by the W operator

Paper: May want to change…

Discussion: GraphLab and/or pregel

Outline of 590q



Monday, 10/7	Comprehension	**Main:** Buneman et al: Comprehension Syntax. SIGMOD Record, 1994 **Optional:** A modern nested data model: Protobuf **Optional:** Wadler: Comprehending Monads, 1992 (Sec. 2, 3 only)
Monday, 10/14	Principles	**Main:** Tannen et al: Naturally Embedded Query Languages. ICDT 1992 (Sec 1-4 only)
Monday, 10/21	Case Studies	**Main:** Melnik et al., Dremel: interactive analysis of web-scale datasets. CACM 2011 **Optional:** The Asterix Query Language **Optional:** Olston et al: Pig latin: a not-so-foreign language for data processing, SIGMOD 2008 **Optional:** Jaql
Monday, 10/28	Implementation	**Main:** Melnik et al, Dremel: Interactive Analysis of Web-Scale Datasets. PVLDB 2011 **Optional:** Abadi et al, Column-stores vs. row-stores: howdifferent are they really? SIGMOD 2008 **Optional:** Liefke, Suciu: XMILL: An Efficient Compressor for XML Data. SIGMOD 2000
Monday, 11/4	Conservativity (1)	**Main:** Wong: Normal Forms and Conservative Properties for Query Languages over Collection Types. PODS 1993
Monday, 11/11		-- external speakers
Monday, 11/18	Conservativity (2)	**Main:** Suciu: Bounded Fixpoints for Complex Objects. Theor. Comput. Sci. 1997
Monday, 11/25	Parallelism	**Main:** Suciu, Tannen: A Query Language for NC. J. Comput. Syst. Sci. 1997
Monday, 12/2	While-Language	**Main:** Abiteboul, Vianu: Fixpoint Extensions of First-Order Logic and Datalog-Like Languages. LICS 1989