### Data Integration: From the Enterprise into Your Kitchen

### Alon Halevy C.E.O. Recruit Institute of Technology

May 14, 2017

Joint work with: Wang-Chiew Tan, George Mihaila, Behzad Golshan

# Outline

- Throwback to the 90's:
  - Answering queries using views: what and why
- Broader view: data integration

– Still, mostly 90's

Data integration today

The Web and data lakes

• Two data integration challenges

- Might sound familiar

not a survey

### The View Rewriting Problem

Given a query Q and a set of view definitions  $V_1, ..., V_n$ :

Is it possible to find a rewriting of Q using only the V's?

V<sub>1</sub>(A,B) :- cites(A,B), cites(B,A) V<sub>2</sub>(C,D) :- sameTopic(C,D), cites(C,C1), cites(D,D1)

### Query: q(x,y) :- sameTopic(X,Y), cites(X,Y), cites(Y,X)

Query rewriting:  $q'(X,Y) := V_1(X,Y), V_2(X,Y)$ 

### The View Rewriting Problem

Given a query Q and a set of view definitions  $V_1, ..., V_n$ :

Is it possible to find a rewriting of Q using only the V's?

V<sub>1</sub>(A,B) :- follows(A,B), follows(B,A) V<sub>2</sub>(C,D) :- FBfriends(C,D), follows(C1,C), follows(D1,D)

Query: q(X,Y) :- FBfriends(X,Y), follows(X,Y), follows(Y,X)

Query rewriting:  $q'(X,Y) := V_1(X,Y), V_2(X,Y)$ 

### An Equivalent Query Rewriting

### Query: q(X,Y) :- FBfriends(X,Y), follows(X,Y), follows(Y,X)

Query rewriting:  $q'(X,Y) := V_1(X,Y), V_2(X,Y)$ 

### Unfolding of the rewriting: q''(X,Y) :- follows(X,Y), follows(Y,X), FBfriends(X,Y), follows (Z,X), follows(W,Y)

The unfolding is *equivalent* to the original query.

# Why Do We Care?

- Query optimization
- Data warehouse design

– Semantic data caching

- Bridging between storage schema and logical schema
- Data citation

### Data integration





### Local As View

- Describe each source as a view
  - Answering a query *requires* view rewriting
  - Unlike GAV, the rewriting algorithm does the hard work of combining sources
- Rewriting may not be equivalent to the query
   Data sources may not be available or complete
- Goal: find *maximally-contained* rewriting

Q' is a maximally-contained rewriting of Q w.r.t. *L* using the V's if there is no other Q'' such that: Q'' strictly contains Q', and Q'' is contained in Q.

### A Basic Decidability Result



Q is a query, V is a set of views

- If Q has built-in predicates and V doesn't, then deciding if there is a rewriting of Q using V is NP-complete.
- If V also has built-in predicates, then the problem is ∏<sup>P</sup><sub>2</sub>-complete.

Incomplete Info/ Certain Answers **View extensions represent partial information:** Extensions of the views v<sub>1</sub>,...v<sub>n</sub> define a set of databases **D** that are **consistent** v<sub>1</sub>,...v<sub>n</sub>.

The tuple *t* is a *certain answer* to a query Q if it would be an answer in every database in **D**. [Abiteboul & Duschka]

- An equivalent rewriting provides *all* certain answers
- A maximally-contained rewriting provides *only* certain answers

# A Fertile Ground for Exploration

- View definition language & query language
- Equivalent or maximally contained rewriting
- Semantic constraints (e.g., FD's, inclusions)
- Completeness/soundness of the views
- Binding patterns restrictions
- Data exchange

# Outline

- ✓ Throwback to the 90's:
  - Answering queries using views: what and why
- >Broader view: data integration
  - ≻Still, mostly 90's
- Data integration today
  - The Web and data lakes
- Two data integration challenges
  - Might sound familiar

### The Data Integration Universe

Creating semantic mappings [Clio, LSD, model management] Language for semantic mappings and reformulation algorithms

Reference reconciliation: "Alon Halevy" = "Alon Levy"?

Query optimization and execution: Eddies, adaptive execution

### Enterprise Data Integration (\$30M Slide)



Walmart's banana problem

### **Integration of Scientific Data**



### Data Integration Assumptions: 90's

- Mediated schema is of reasonable size
- Data sources are mostly structured
- We *need* to integrate all the sources
- Data is mostly correct and consistent (modulo cleaning)

All these assumptions are reasonable when the number of sources is in the 10's.

# Outline

- ✓ Throwback to the 90's:
- ✓ Answering queries using views: what and why
   ✓ Broader view: data integration
   ✓ Still, mostly 90's
- Data integration today
  - The Web and data lakes
- Two data integration challenges
  - Might sound familiar

### Today: Data Integration Powers Personal Assistants

Google	when does my flight leave tomorrow	<b>\$</b> Q
	All News Maps Shopping Videos	More Settings Tools
	About 28,500,000 results (1.03 seconds)	
	Flights on Saturday, May 13 Only you can see this result	
	From unitedairlines@united.com United Flight 220	
	SFO+	• ORD
	San Francisco · Sat, May 13 Scheduled departure Terminal Gate	Chicago · Sat, May 13 Scheduled arrival Terminal Gate
	Passenger Information	7:15 PM
	HALEVY/ALON	A4LJKV Seat

View email for this flight

### And It's In Your Kitchen & Bedroom



# The data integration vision of the 90's



### More "views" Are Being Added



### Check out: doneList on YouTube.

# Speech recognition and natural language understanding

Creating semantic mappings [Clio, LSD, model management] Language for semantic mappings and reformulation algorithms

Reference reconciliation: "Alon Halevy" = "Alon Levy"?

Query optimization and execution: Eddies, adaptive execution

### The Landscape Changed

- The Web (& personal agents)
- Enterprise data lakes
- Data science

100's of millions of data sources:

- *Mediated schema* is merely an approximation
- Data sources are of all kinds
- We *cannot* integrate all the sources
- Data is mostly correct and consistent
  - Still valid! Everything on the Web is true, right?

### Integration in Search Engines

# Views map from query intent to queries over the Google Knowledge Graph

what is the at	omic weigh	Į	, Q			
All Images	News	Videos	Shopping	More	Settings	Tools
About 6,290,000	results (0.87	seconds)				
Oxygen / Ato	mic mass					
15.999	4 u ±	0.00	04 u		8 Oxyg 15,99	en 94

### But the KG Has to be Very Broad

when was madonna born	ę	<b>, Q</b>	
All News Images Shopping V	ideos More Settings	Tools	
About 19,500,000 results (1.99 seconds)			
Madonna / Born	chicago populatio	n	
August 16, 1050 /			
years), Bay City, M	All News Ma	aps Images Shopping	
		who is the king of denmark	<b>y</b> d
Jennifer Lopez July 24, 1969	About 87,900,000 re	All News Images Shopping Videos More	Settings Tools
	Chicago / Popul	About 53,500,000 results (1.13 seconds)	
		Denmark / Monarch	
	2.719 mi	Margrethe II of Denmark Since 1972	
		More about Margrethe II of Denmark	

### Repeat x 100 Languages



Commentaires

### Knowledge Graph Rough Edges

rio o	rio olympics medal count							
All	News	Videos	Shopping	Images	More	Settings	Tools	

About 9,280,000 results (0.85 seconds)

Medal table		
Rank	NOC	Total
1	United States (USA)	121
2	Great Britain (GBR)	67
3	China (CHN)	70
4	Russia (RUS)	55

84 more rows

2016 Summer Olympics medal table - Wikipedia https://en.wikipedia.org/wiki/2016\_Summer\_Olympics\_medal\_table

### Long Nose Queries



### Recruit Holdings: A Lifestyle Company



Over 200 online services Notable subsidiaries: indeed.com, Treatwell

### Beauty and Massage



LIST YOUR BUSINESS LOG IN

HAIR I HAIR REMOVAL I MASSAGE I NAILS I FACE I BODY I SPA DAYS & BREAKS I VALENTINE'S DAY I EXPLORE SALONS



### **Relevant to Most PODS Authors**



# treatwell

### **Book Yourself Fabulous**

### **Relevant to Most PODS Authors**



⋇	United Kingdom	
=	Nederland Neth	erlands
-	Deutschland	
•	België Belgique	Belgium
••	France	
=	Österreich	
<u>.</u>	España	
•	Ireland	
•	Italia	
	Schweiz	
=	Lietuva	



Agadir Spa

Jewellery Quarter, Birmingham, UK



Agadir Spa is a an authentic Moroccan beauty centre based in the Jewellery Quarter in Birmingham, where we offer full spa treatment services.

If we could understand that Morocco is not location but a type of beauty treatments we could answer queries such as "Moroccan massage" or "Moroccan spa"

### **Reformulate Data Integration**

- It's a data space (or data lake), not a data integration system
  - Basic operation to support: search
  - See Goods @ Google (SIGMOD, 2016)

- Pay-as-you-go data integration:
  - For deeper integration, invest where you see the most value (typically measured by query traffic)

### **Curate Head Content**



About 154,000 results (0.36 seconds)

Michael Stonebraker - Wikipedia, the free encyclopedia

https://en.wikipedia.org/wiki/Michael\_Stonebraker • Wikipedia ~ Michael Ralph Stonebraker (born October 11, 1943) is a computer scientist specializing in database research. Through a series of academic prototypes and ... Major projects - Students - Selected works - References

#### Michael Stonebraker - A.M. Turing Award Winner

amturing.acm.org/award.../stonebraker\_1172121.cfm 

Turing Award 
Michael Stonebraker has made fundamental contributions to database systems, which are one of the critical applications of computers today and contain much ...

#### Michael Stonebraker | MIT CSAIL

https://www.csa... 
 MIT Computer Science and Artificial Intelligence Labo... 
 Michael Stonebraker has been a pioneer of data base research and technology for more than a guarter of a century. He was the main architect of the INGRES ...

#### Michael Stonebraker wins \$1 million Turing Award | MIT News

newsoffice.mit.edu/2015/michael-stonebraker-wins-turing-award-0325 = Mar 25, 2015 - Michael Stonebraker, a researcher at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) who has revolutionized the field of ...

Michael Stonebraker | VoltDB



#### Michael Stonebraker

Computer Scientist

Michael Ralph Stonebraker is a computer scientist specializing in database research. Through a series of academic prototypes and commercial startups, Stonebraker's research and products are central to ... Wikipedia

Born: October 11, 1943 (age 71), Milton, NH

Awards: Turing Award, IEEE John von Neumann Medal

Organizations founded: Vertica, StreamBase Systems, Cohera

Books: Object-relational DBMSs: The Next Great Wave

Notable students: Joseph M. Hellerstein

Education: University of Michigan (1971), Princeton University (1985)



### **Attribute Queries**



### **Multi-Valued Attributes Supported**



### **Complicated Queries**

pegoogle	t-mot	bile store	downtown s	eattle su	nday open	ing hours		<mark>୍</mark> ୟ ବ	
	All	Maps	Shopping	News	Images	More +	Search tools		
	About	19,600,000	0 results (0.64	seconds)					
	<b>1</b> ′ ⊤-№	1AM Iobile Se	<b>–6PM</b> attle, Sunda	y hours				Г	·M

Feedback

### Structured Data Can Be Funny

how tall is stephen colbert

Web

News Images Shopping

Videos

More -Search tools J,

About 391,000 results (0.33 seconds)

### 5' 10.5" (1.79m -ish)

Stephen Colbert, Height









Jon Stewart Shorter



Conan O'Brien

Feedback

### **Enable Long-Tail Curators**



About 185,000 results (0.82 seconds)

#### Port 2011 Rankings - World Barista Championship

www.worldbaristachampionship.org/wp-content/.../2011-WBC-Ranking-Order.pdf = Crafted Coffee Company. New Zealand. 540.5. 15. AnneStine Bae. The Coffee Collective. Denmark. 538. 2011 WORLD BARISTA CHAMPIONSHIP RANKING ...

#### World Barista Championship

#### www.worldbaristachampionship.org/ +

May 14, 2014 - Picture by Jake Olson for World Coffee Events. The World Barista Championship (WBC) is the preeminent international coffee competition ... You've visited this page 5 times. Last visit: 1/5/15

#### 2011 World Barista Championship - live streaming video powered by ... https://livestream.com/worldbaristachampionshiplive •

2-5 June, Live from Bogota, Colombia, the 2011 World Barlata Championship! (note: embedding on other websites is disabled. Please point your browsers t.

### World Barista Championship 2011

<\*

Barista competition

#### Number of participants: 54

Location: Bogotá

Winner: Alejandro Mendez, El Salvador

2nd place: Pete Licata, United States

3rd place: Matt Perger, Australia

4th place: Javier Garcia, Spain

5th place: Miki Suzuki, Japan

### Fall Back on Unintegrated Content

Google	top coffee exporters								
	Web Shopping News Images Maps More - Search t						Search tools		
	About 935,000 results (0.44 seconds)								
	Main exporters by country in 2014								
	Count	try	6	0 kilogram	bags	Metric	Tons		
	Brazil		4	5,342,000		2,720,520			
	Vietna	am	2	27,500,000		1,650,0	000		
	Colombia         11,600,000         696,000           Indonesia         6,850,000         411,000								
	49 more rows, 1 more column								

https://en.wikipedia.org/.../List\_of\_countries\_by\_coffee\_producti...

Wikipedia -

Feedback

# Why Is the Sky Blue?

Google	why i	s the sky	ļ	, Q				
	All	Books	Videos	Images	Shopping	More	Settings	Tools

About 170,000,000 results (0.66 seconds)

A clear cloudless day-time **sky** is **blue** because molecules in the air scatter **blue** light from the sun more than they scatter red light. When we look towards the sun at sunset, we see red and orange colours because the **blue** light has been scattered out and away from the line of sight.



Why is the sky Blue? math.ucr.edu/home/baez/physics/General/BlueSky/blue\_sky.html

About this result · Feedback

# Outline

- ✓ Throwback to the 90's:
- ✓ Answering queries using views: what and why
   ✓ Broader view: data integration
   ✓ Still, mostly 90's
- $\checkmark$  Data integration today
  - The Web and data lakes
- > Two data integration challenges
  - Might sound familiar

### Many Algorithms to Answer Queries



### who is the speaker of the house?





### Stand with Speaker Ryan - speakerryan.com

bute \$10 to help usher in a positive, conservative agenda!

### John Boehner

Speaker of the U.S. House of Representatives



58 more rows, 3 more columns

beaker of the U.S. House of Representatives | United States ... https://www.britannica.com/.../Speaker-of-the-US-House-of-Re... Encyclopædia Britannica -



Challenge #1: Combining Structured and Unstructured Data

- We have methods for extracting facts from unstructured data
  - But then we don't know how to reconcile the extractions with structured data
- Two observations about the solution:
  - Combination should be specified declaratively
    - Right now it's hardwired somewhere in code
  - Need to keep the original data (context) around.
     We might need it later.

# Challenge #2: Open-Source Tools

- The field has been around for a while and made impressive progress
- But...
  - No (free) tools available
  - Hard to prototype a data integration application
  - Each research effort needs to start from scratch
  - Missing opportunities for impact on the world (e.g., in data science)

### We Need Integration Operators



# BigGorilla.org

- Goal: open source data integration and preparation eco-system.
- In Python (for easy prototyping)
- Currently: string matching, data matching, schema matching, scraping.
  - See tutorials, code and demos on the site
- Collaboration with U. Wisconsin; looking for additional contributors.



# Conclusions

- Data integration is everywhere:
  - Even in my kitchen!
  - Sources described by views
  - AI (NLP & Speech) are enabling new data integration scenarios
  - And enterprises have not solved the problem yet
- But we face two age-old problems:
  - The world is messy and our models are not adequate
  - Data integration is still too hard and labor intensive





### Q & A

