Tackling The Challenges of Big Data

Big Data Collection

Michael Stonebraker

Professor

Massachusetts Institute of Technology

Tackling The Challenges of Big Data

Big Data Collection

Data Cleaning & Integration

Introduction

Michael Stonebraker

Professor

Massachusetts Institute of Technology

Data Curation

- Ingest
- Validate
- Transform
- Correct
- Consolidate (dedup)
- And visualize information to be integrated

Data Warehouse Roots

- Retail sector started integrating sales data into a data warehouse in the early 1990's
- Average system was 2X budget and 2X late
- Because of data integration headaches
- However, warehouse paid for itself within 6 months with smarter buying decisions!

Issues

- sold $100K of widgets to IBM, Inc.
- sold 800K Euros of m-widgets to IBM, SA

- Translate currencies
- Is IBM, SA the same as IBM, Inc?
- Are m-widgets the same as widgets?

The Pile-On

- Essentially all enterprises followed suit and built warehouses of customer facing data
- Serviced by so-called Extract-Transform-and Load (ETL) tools

Architecture

- http://en.wikipedia.org/wiki/Data_integration

Traditional Wisdom - ETL

- **Human defines a global schema**
- **Assign a programmer to each data source to:**
  - Understand it
  - Write local to global mapping (in a scripting language)
  - Write cleaning routine
  - Run the ETL
- **Scales to (maybe) 25 data sources**

Traditional ETL Methodology – Schema Mapping

http://www.google.com/imgres?imgurl=http://www.xmlschema.info/images/shots/map_xml_thumb.gif&imgrefurl=http://www.xmlschema.info/xml_schema_mapping.html&h=469&w=600&sz=63&tbnid=70oECAvg0TMwkM:&tbnh=102&tbnw=131&zoom=1&usg=__AsYa-CEcZeR6MX9IV8JqvdpX9RM=&docid=MbKKOGXt3LED1M&sa=X&ei=XDGJUpi9OYevsQT5lYGAAg&ved=0CDQQ9QEwAg

Traditional ETL Methodology – Data Transformation

http://www.informatica.com/us/products/enterprise-data-integration/powercenter/

Current Situation

- **Enterprises want to integrate more and more data sources**
  - Miller beer example
  - Novartis example
  - Goby example (we will see this again)
- **Traditional ETL won't scale!!!!**
- **Point-projects in departments -- Staffed by a data scientist**
  - Brand manager deciding marketing spend
  - Augmenting demographic customer data for department use
- **Traditional ETL way too heavy-weight**

The Rest of This Module

- Curation example
- Low end (individual data scientist) support
- Enterprise support

Tackling The Challenges of Big Data
Big Data Collection
Data Cleaning & Integration
Introduction

THANK YOU

Tackling The Challenges of Big Data
Big Data Collection

Michael Stonebraker
Professor
Massachusetts Institute of Technology

Tackling The Challenges of Big Data
Big Data Collection
Data Cleaning & Integration
Issues

Michael Stonebraker
Professor
Massachusetts Institute of Technology

The Problem

Demo of Goby.com

In Summary

- Data is dirty!!!!!
- Sometimes not clear how to clean it
  - 2 restaurants at the same address: food court or one went out of business??
- Transformations may be a big problem

Tackling The Challenges of Big Data

Big Data Collection

Data Cleaning & Integration

Issues

THANK YOU

Tackling The Challenges of Big Data

Big Data Collection

Michael Stonebraker

Professor

Massachusetts Institute of Technology

Tackling The Challenges of Big Data

Big Data Collection

Data Cleaning & Integration

New Ideas - 1

Michael Stonebraker

Professor

Massachusetts Institute of Technology

Startups in This Space (probably a bunch more)

- Paxata
- Trifacta (commercial Data Wrangler)
- Cambridge Semantics
- Data Tamer
- ClearStory
- Attivio

At Least Two Foci

- Support for the individual data scientist
- Enterprise data integration

Support for the Data Scientist

- Wrangler video
- http://vis.stanford.edu/wrangler/

Summary

- Expect more systems in this space
- At low prices
- Market will be gated by the availability of data scientists
  – Insurance example

Tackling The Challenges of Big Data
Big Data Collection
Data Cleaning & Integration
New Ideas - 1

THANK YOU

Tackling The Challenges of Big Data
Big Data Collection

Michael Stonebraker
Professor
Massachusetts Institute of Technology

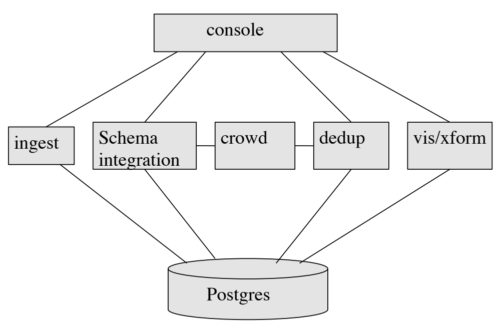
Tackling The Challenges of Big Data
Big Data Collection
Data Cleaning & Integration
New Ideas - 2

Michael Stonebraker
Professor
Massachusetts Institute of Technology

Data Tamer Goals

- Do the "long tail"
  - Better/cheaper/faster than the ad-hoc techniques being used currently
- By inverting the normal ETL architecture
  - Machine learning and statistics
  - Ask for human help only when automatic algorithms are unsure

Data Tamer Architecture



Data Tamer -- Ingest

- Assumes (for now) a data source is a collection of records, each a collection of (attribute-name, value) pairs.
- Loaded into Postgres

Data Tamer – Schema Integration

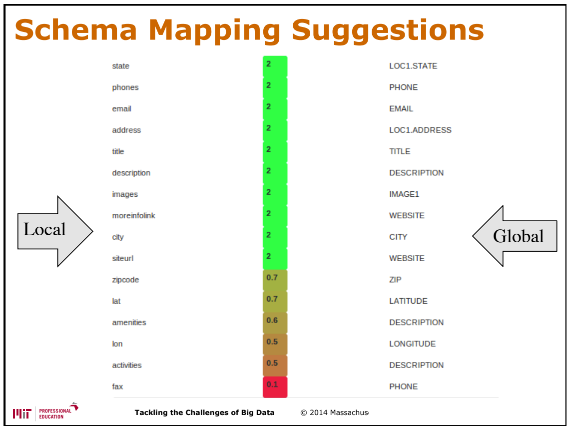
- Must be told whether there is a predefined partial or complete global schema or nothing
- Starts integrating data sources
  - Using synonyms, templates, and authoritative tables for help
  - 1st couple of sources require asking the crowd for answers
  - System gets better and better over time

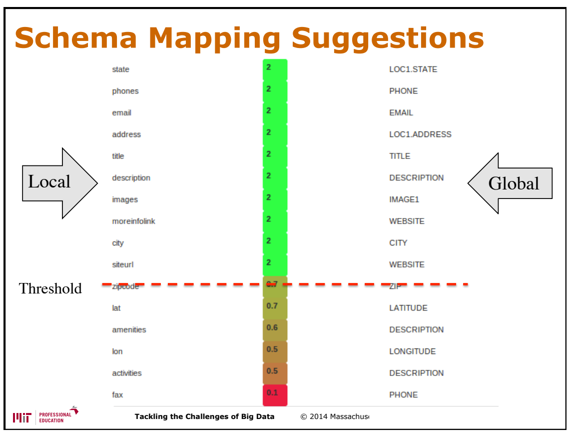
Data Tamer – Schema Integration

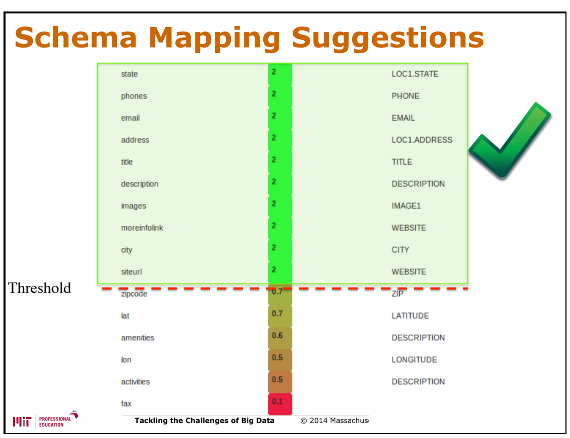
- Inner loop is a collection of experts
  - T-test on the data
  - Cosine similarity on attribute names
  - Cosine similarity on the data
- Scores combined heuristically
- After modest training, get 90% of the matching attributes on Goby and Novartis automatically
  - Cuts human cost dramatically

Data Tamer – Crowd Sourcing

- Hierarchy of experts
- With specializations
- With algorithms to adjust the “expertness” of experts
- And a marketplace to perform load balancing
- Currently doing a large scale evaluation at Novartis
  - Late flash: it works!!!!


Schema Mapping Suggestions
Local
Global
state 2 LOC1.STATE
phones 2 PHONE
email 2 EMAIL
address 2 LOC1.ADDRESS
title 2 TITLE
description 2 DESCRIPTION
images 2 IMAGE1
moreinfolink 2 WEBSITE
city 2 CITY
siteurl 2 WEBSITE
zipcode 0.7 ZIP
lat 0.7 LATITUDE
amenities 0.6 DESCRIPTION
lon 0.5 LONGITUDE
activities 0.5 DESCRIPTION
fax 0.1 PHONE
Tackling the Challenges of Big Data
© 2014 Massachus


Schema Mapping Suggestions
Local
Global
Threshold
state 2 LOC1.STATE
phones 2 PHONE
email 2 EMAIL
address 2 LOC1.ADDRESS
title 2 TITLE
description 2 DESCRIPTION
images 2 IMAGE1
moreinfolink 2 WEBSITE
city 2 CITY
siteurl 2 WEBSITE
zipcode 0.7 ZIP
lat 0.7 LATITUDE
amenities 0.6 DESCRIPTION
lon 0.5 LONGITUDE
activities 0.5 DESCRIPTION
fax 0.1 PHONE
Tackling the Challenges of Big Data
© 2014 Massachus


Schema Mapping Suggestions
Threshold
state 2 LOC1.STATE
phones 2 PHONE
email 2 EMAIL
address 2 LOC1.ADDRESS
title 2 TITLE
description 2 DESCRIPTION
images 2 IMAGE1
moreinfolink 2 WEBSITE
city 2 CITY
siteurl 2 WEBSITE
zipcode 0.7 ZIP
lat 0.7 LATITUDE
amenities 0.6 DESCRIPTION
lon 0.5 LONGITUDE
activities 0.5 DESCRIPTION
fax 0.1
Tackling the Challenges of Big Data
© 2014 Massachus

Schema Mapping Suggestions

state	2	LOC1.STATE
phones	2	PHONE
email	2	EMAIL
address	2	LOC1.ADDRESS
title	2	TITLE
description	2	DESCRIPTION
images	2	IMAGE1
moreinfolink	2	WEBSITE
city	2	CITY
siteurl	2	WEBSITE
zipcode	0.7	ZIP
lat	0.7	LATITUDE
amenities	0.6	DESCRIPTION
lon	0.5	LONGITUDE
activities	0.5	DESCRIPTION
fax	0.1	

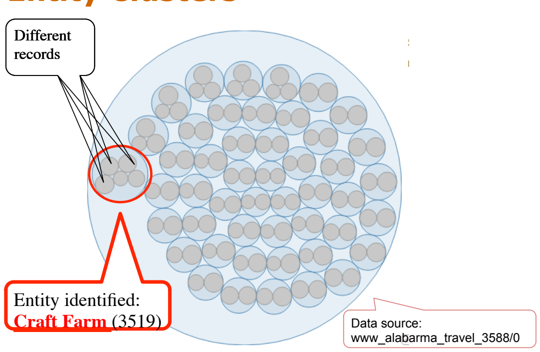
Threshold

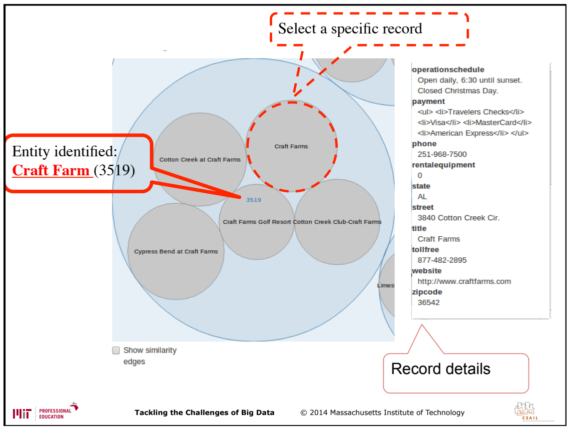
Crowd

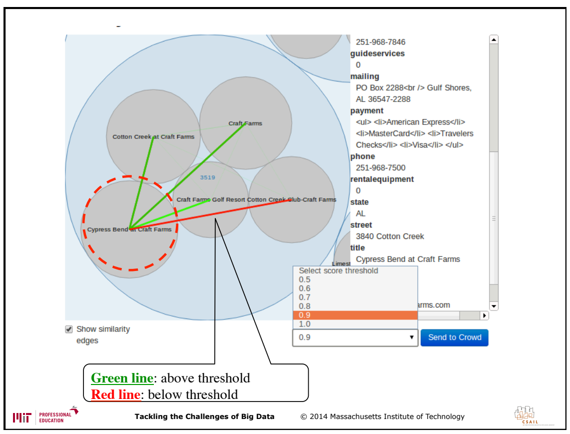
Data Tamer – Entity Consolidation

- On tables defined by schema integration module
- Entity matching on all attributes, weighted by value presence and distribution
- Basically a data clustering problem
- With a first pass to try to identify "blocks" of records
  - Otherwise N ** 2 in the number of records
- Wildly better than Goby; a bit better than domain-specific Verisk module

Entity Clusters







Data Tamer Future

- Text
- Relationships
- Hierarchical data (maybe)
- Adaptors
- Better algorithms
- User-defined operations

Tackling the Challenges of Big Data

Tackling The Challenges of Big Data
Big Data Collection
Data Cleaning & Integration
New Ideas - 2

THANK YOU

Tackling The Challenges of Big Data
Big Data Collection

Michael Stonebraker
Professor
Massachusetts Institute of Technology

Tackling The Challenges of Big Data
Big Data Collection
Data Cleaning & Integration
Summary

Michael Stonebraker
Professor
Massachusetts Institute of Technology

The Way Forward

- Enterprises will want to integrate more and more data sources
  - This is the number one headache of most of them
  - Too expensive to do manually with a programmer
- Remains to be seen what fraction of the market can be aided by Data Tamer-style tools
  - Initial results are encouraging

The Way Forward

- Cleaning will be a big issue forever
  - How clean does your data need to be?
- I imagine a big database of transformations
  - Pick the one that you need
- Data scientists will have to familiar with this stuff
  - Not just stat and data management

Tackling The Challenges of Big Data
Big Data Collection
Data Cleaning & Integration
Summary

THANK YOU

Tackling The Challenges of Big Data

Big Data Collection

Data Cleaning & Integration

Michael Stonebraker

Professor

Massachusetts Institute of Technology