Tackling The Challenges of Big Data

Big Data Collection

Michael Stonebraker

Professor

Massachusetts Institute of Technology

MIT PROFESSIONAL EDUCATION

CSAIL
MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY

Tackling The Challenges of Big Data

Data Cleaning and Integration:

Three Years Later

Michael Stonebraker

Professor

Massachusetts Institute of Technology

Approaches

Traditional ETL vendors

- Won't scale to large number of sources
- As noted earlier in this module

Data preparation tools

Enterprise curation tools

Data lakes

The future

Supporting data science at scale

Data Preparation Tools – Lots of Choices

Trifacta

Paxata

Alteryx

Cambridge Semantics

Clear Story

Informatica Springbok/Rev

…

Enterprise Data Curation

Tamr

Deep Dive

Tamr Experience

Data cleaning

- No esperanto
- Need to run a collection of tools
- Need domain specific tools

Data transformation

- Always present

Tamr Experience

Important to Verticalize

- Field training
- Auxiliary data sets (enrichment)
- Specialized algorithms
- Word of mouth viral spread

Verticals

Human trafficking (Deep Dive)

- Predict which web advertisements entail trafficking, using domain-specific rules
- Works like a charm
- In use at several big city police departments

Verticals

Procurement (Tamr)

- Categorize spend transactions for analysis
- “Supplier mastering” to get “most favored nation” status
- “Parts mastering” to buy from the cheapest source
- In use at GE (325 procurement systems)
- Estimated to save $300M in 2016

Enterprise Data Curation

A bunch of verticals where a product supplemented by professional services will work

Every customer needs help

- Suitable expertise rarely exists in-house

Competition is from the much-more-expensive consulting firms

- Palentir, MuSigma, …

Data Curation Entails

Ingest (into a common place)

Clean (-9999 means null)

Transform (Euros to $)

Schema match (your salary is my wages)

De-duplicate (Mike Stonebraker, M.R. Stonebraker)

Export (to a downstream system of record)

Data Lakes

Solve only the ingest problem

Which is at most 5% of the problem

- Leaving the remaining 95% unsolved

Generates a data swamp not a data lake

- Enterprise junk drawer

Supporting Data Science at Scale (Mark Schrieber – Merck)

Merck has 4K Oracle data bases

Countless other data sets inside the firewall

A Merck scientist has a hypothesis (e.g. is diabetes correlated with ritalen consumption?)

- Mark estimates 98% of scientist time is spent finding and curating data sets of interest)
- Nobody estimates less than 80%

Goal: knock this number down

Outline of Data Civilizer

M.I.T./Waterloo/QCRI prototype

Working with Merck, Novartis and the M.I.T. data warehouse project

Outline of Data Civilizer

Discovery system
Basically a super catalog of data resources

Data stitching system

- Scientist generates items of interest (name and salary of Chem employees)
- Goal is to create the definition of a view or views that contains the desired information
- User can then query this view

Outline of Data Civilizer

Since the raw data sources are in a variety of storage systems

- Need a data federation system to be able to extract data from the variety of sources

Want to clean/xform data on demand

- Since this entails human checking
- Cleaning and transformation must be integrated into federation system
- Need accuracy or cost goals

Data Civilizer Optimization

Usually want to keep (materialize) views that have been constructed

Need to perform incremental curation as sources get updated

Need to decide whether to use an MV or go back to the raw data

Watch this space for user reaction to Civilizer

Summary

Curation tools are improving

Currently this is the “800 pound gorilla in the corner”

Optimistic that we can tame the gorilla

Tackling The Challenges of Big Data

Big Data Collection

THANK YOU

MIT PROFESSIONAL EDUCATION

CSAIL
MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY