THOMSON REUTERS
Big Data at a Big Company

University of Minnesota – Digital Technology Center

Mike Edwards – Principal Architect
Data Center Operations

2012/12/12
Thomson Reuters

$12.7 billion 2011 revenues
What do we mean by Big Data?
What do we mean by *Big Data*?

- Volume
- Velocity
- Variety
- Visualization
- Value
Volume (Facebook)

500 MM people *

Color ~ distance, # of pairings

Largest cluster > 100PB of data

DW adds > .5 PB / day

* 2010
Volume & Variety (Thomson Reuters - Legal)

100+ years of data

- Courts (600+)
  - 200 million+ documents
    - opinions, briefs, motions, orders, filings, dockets

- Legislatures (52)
  - 500 million+ documents
    - laws, regulations, codes

- Authors (3000+)
  - 4 million+ documents
    - analytical materials

- 3rd Party Providers (1500+)
  - 700 million+ documents
    - news stories, 3rd party

- 6.5 billion+ relationships
- 600,000+ case classifications *
- 400,000+ law firms
- 1,000,000+ legal professionals
- ~100,000 Key Numbers
- 700 million+ KeyCite refs

* 2010
Velocity (LHC)

- 17 mile circumference
- 7 TeV / particle max
- 5 GB / sec data
- 1 DVD / sec data
- 86,000 DVD / day data
- Level 1 (hw): 100K / sec
- Level 2 (sw): 100 / sec

Velocity (Thomson Reuters – Financial)

- Tick History product
- Aggregated each market day and served back
- We add 100 GB of compressed data per day
- 30% - 40% growth / year
Visualization (Thomson Reuters – IP & Science)

MOBILE PATENT SUITS

Patent-related suits between mobile device/component manufacturers

Amazon  Apple  Barnes & Noble  Ericsson  Foxconn

Qualcomm  RIM  Samsung  Sony  ZTE

Google  HTC  Huawei  Inventec  Kodak

Microsoft  Motorola  Nokia  Oracle  LG

*Utilises Google's Android OS
*Cases resolved


Source: Reuters, news reports

*Kodak’s separate suit against Apple will be decided on Aug 30.
Visualization (Flowing Data)

Value (C.R.U.S.H.)

- Crime
- Reduction
- Using
- Statistical
- History

Beat date / time / place

- Sexual assault #1, #2
- 5000 rapes
- Moved inside
Value (Thomson Reuters – CLEAR)

Documents  
- Company Authority
- People Authority

Entity Extraction & Resolution  
- People
- Authority
- Documents

Relationships & Attributes
- People Warehouse
- Company Warehouse
- Consolidated
- Lead
- Evaluation
- And
- Reporting
Semantic Interoperability

$12.7 billion 2011 revenues\(^1\)
The Practice (Data Science)

geeks (passion)

data architects (cross-stack)

data scientists (visualization, exploratory data analysis, data mining, machine learning)

DBAs (internals)

information architects (modeling)
Hot Topics

• hardware – memory and performance
• hardware – storage and ELT
• system software – streaming data
• system software – graph-based data stores
• analysis – open-ended pattern searching
• analysis – modeling
• visualization – views
• visualization – interactivity
• bringing it all together – collective intelligence
• bringing it all together – intelligence amplification