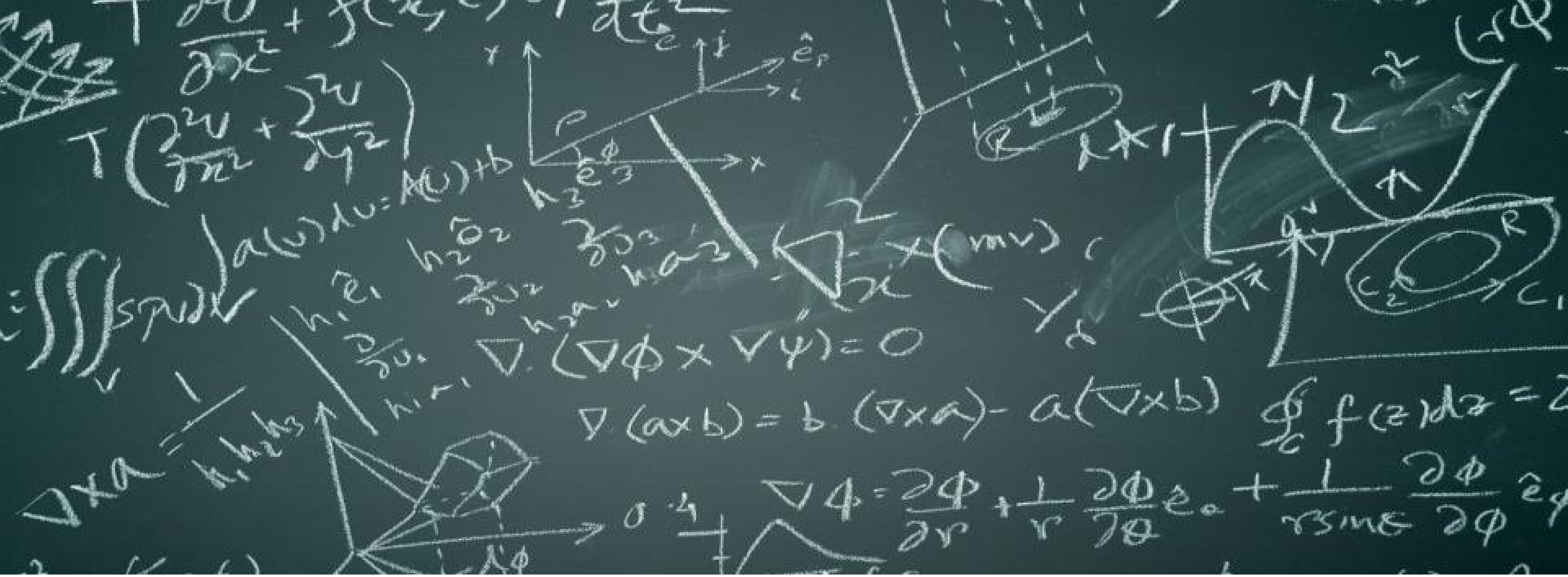


Overview of HPCC Systems and Case Studies

Presenter: Brian Bounds, Director of Software Engineering
November 2014

LexisNexis

Risk Solutions

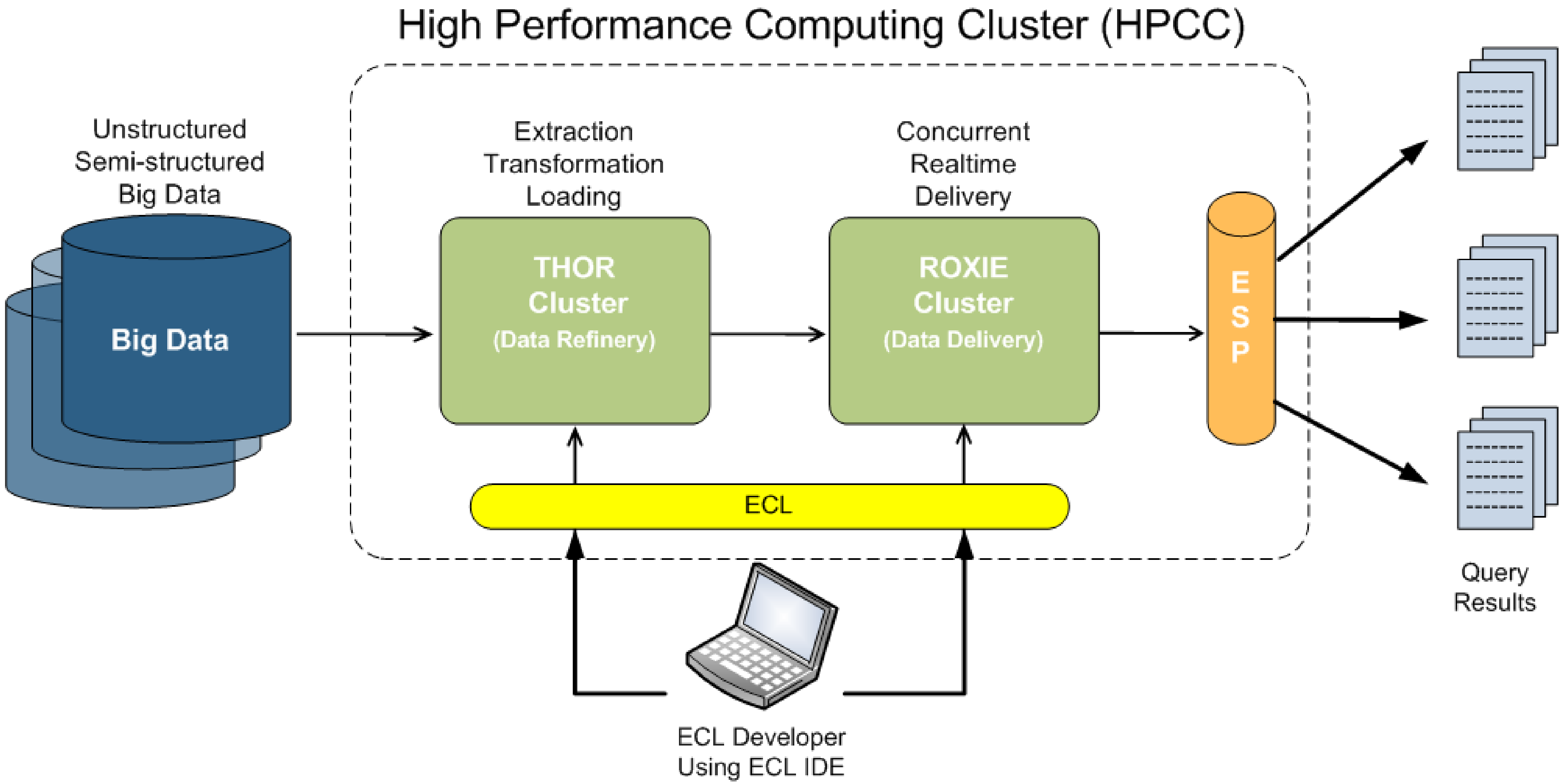
Context

- About Brian
- HPCC Systems in the Context of LexisNexis
- Isn’t Big Data just Data?

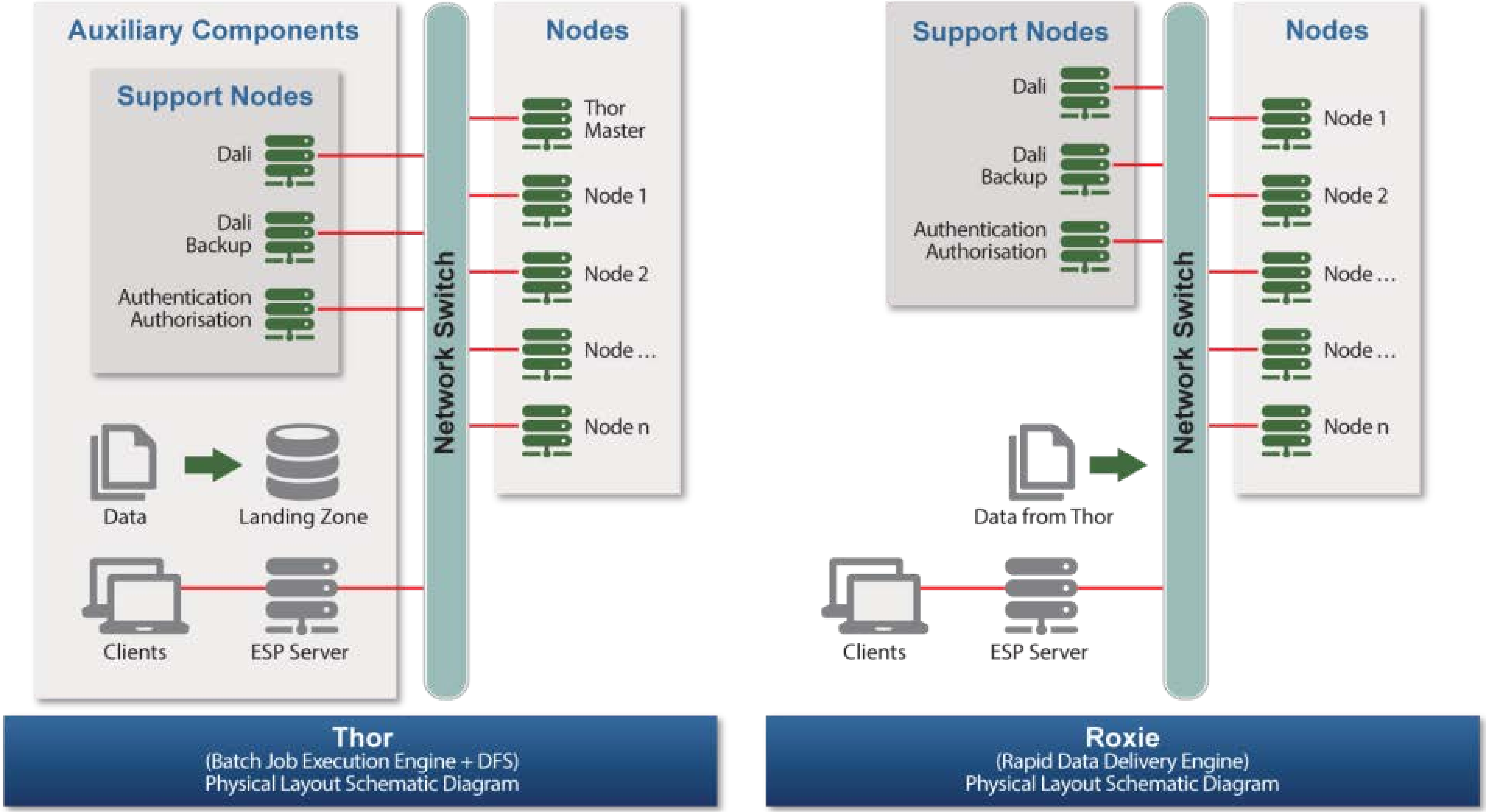
Components of the Open Source HPCC Systems platform

- The HPCC Systems platform includes:
  - Thor: batch oriented data manipulation, linking and analytics engine
  - Roxie: real-time data delivery and analytics engine
- A high level declarative dataflow language: ECL
  - Implicitly parallel
  - No side effects
  - Code/data encapsulation
  - Extensible
  - Highly optimized
  - Builds graphical execution plans
  - Compiles into C++ and native machine code
  - Common to Thor and Roxie
- An extensive library of ECL modules, including data profiling, linking, graph analytics, and Machine Learning

The LexisNexis Open Source HPCC Systems platform



Detailed HPCC Systems Platform Architecture



HPCC Systems Overview Video

Drea’s HPCC Overview

Features to Care About

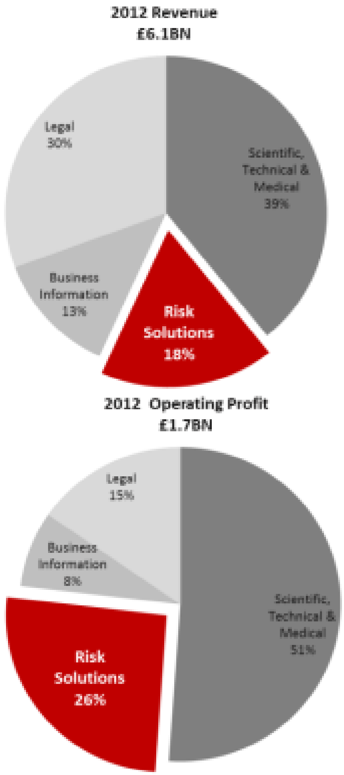
- The Programming Language (ECL)
- The Delivery Engine (ROXIE)
- Enterprise Readiness
- Big Data … becomes Data (that might be Big)

Case Study #1 (Enterprise) – LexisNexis Risk Solutions

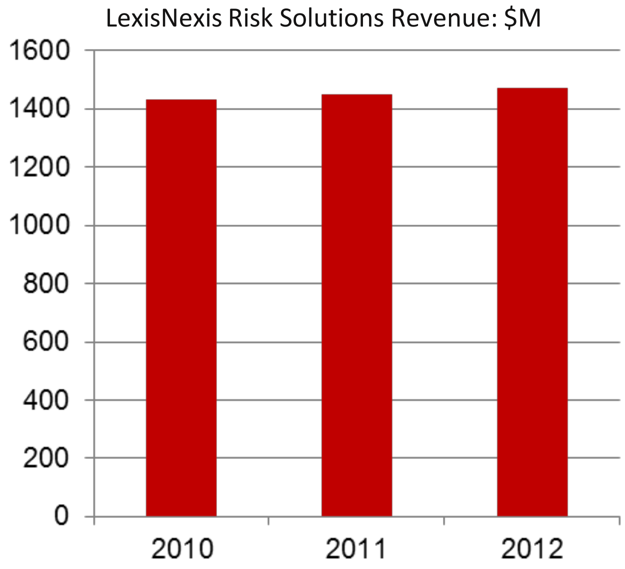
Case Study #1 (Enterprise) – LexisNexis Risk Solutions

We are among the largest providers of risk solutions in the market today

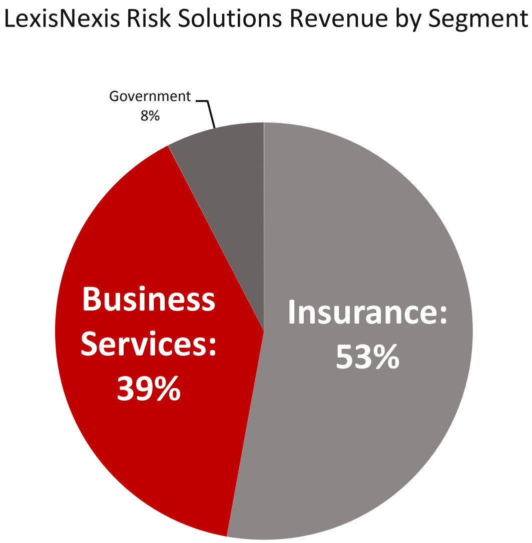
Reed Elsevier is a world leading provider of information solutions.



LexisNexis Risk Solutions has seen sustained revenue and profit growth



LexisNexis Risk Solutions is a leading provider in the U.S. across Business Services, Insurance and Government segments.



Case Study #1 (Enterprise) – LexisNexis Risk Solutions

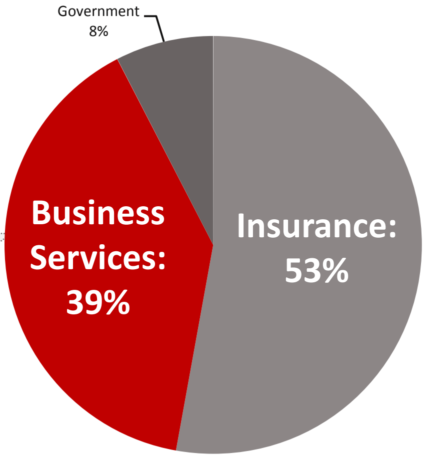
We are among the largest providers of risk solutions in the market today

Our customers include:

- 99 of the top 100 US banks
- 90% of the Fortune 500
- 100% of US P&C insurance carriers
- All 50 US states, 70% of local governments and 80% of US federal agencies
- 97 of Am Law 100 firms

LexisNexis Risk Solutions is a leading provider in the U.S. across Business Services, Insurance and Government segments.

LexisNexis Risk Solutions Revenue by Segment



Case Study #1 (Enterprise) – LexisNexis Risk Solutions

We have a unique set of capabilities: Data, Linking, Analytics, and Product Development

Data Technology		Vast Data Resources		Linking & Analytics		Industry-Specific Expertise & Delivery		Customer-Focused Solutions
	+		+		+		=	
• Speed • Capacity • Cost savings		• Process • Sources • Coverage		• Advanced linking & analytics • Accuracy & efficiency • Protect private information		• Aligned with our customers' industries • Deep industry expertise		• Predict, manage and assess risk across many industries.

Case Study #1 (Enterprise) – LexisNexis Risk Solutions

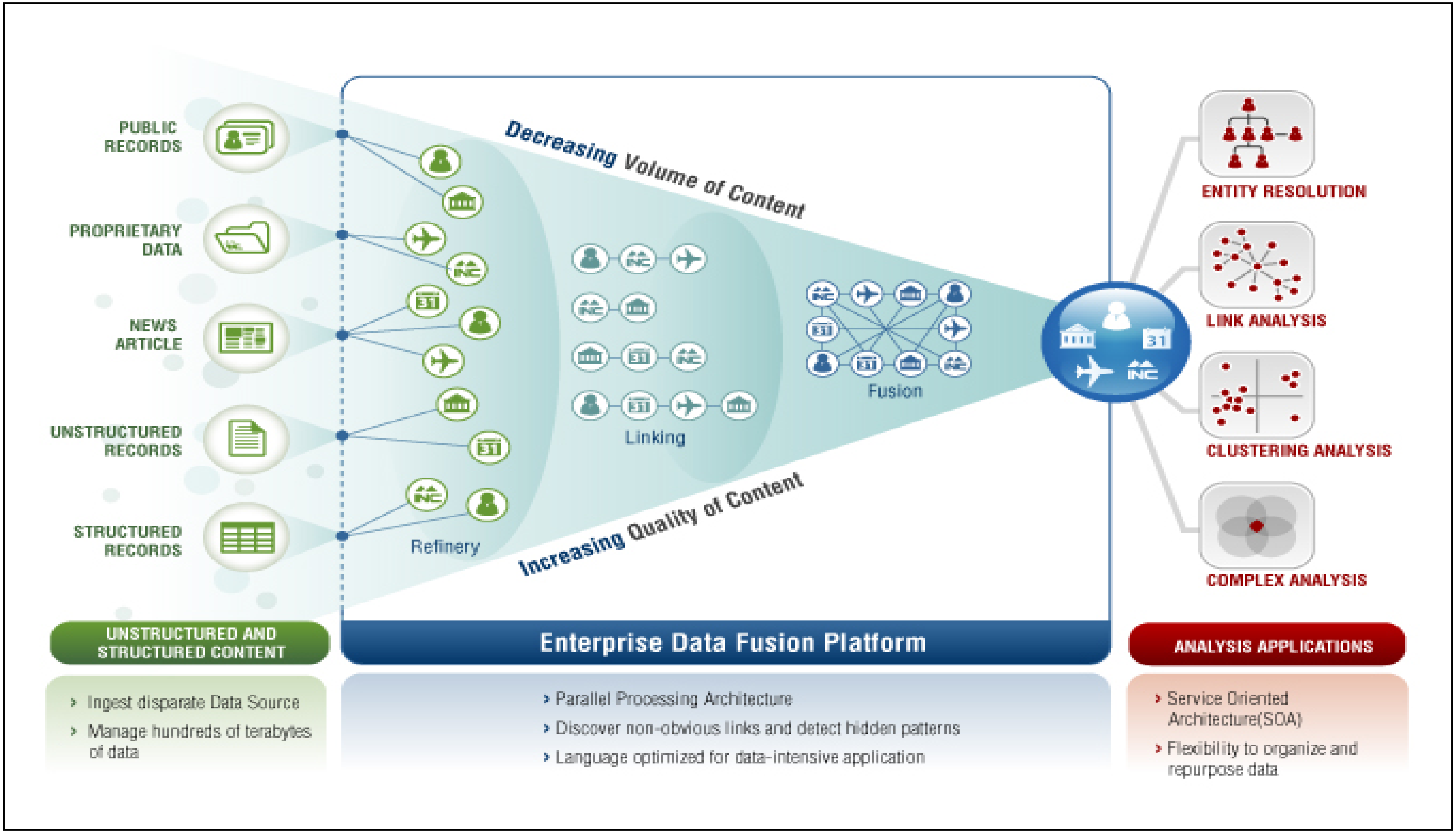
Access to more than 25 billion public record filings

Break-down of record counts for the more popular data sets:

Data Source	# of records
Associates/Relatives	1.8 Billion
Bankruptcy	23 Million
Business BDID's	283 Million
Business People Links	959 Million
Canadian Phones	62 Million
Consumer Header	10.8 Billion
Criminal	216 Million
Date of Birth	5.2 Billion
Death	98 Million
Drivers Licenses	397 Million
EDA Phones	124 Million
FEINs	10.4 Million
Historical Phones	800 Million
Hunting and Fishing Licenses	67 Million
Liens and Judgments	244 Million

Data Source	# of records
People at Work	1.5 Billion
Private Phones	172 Million
Professional Licenses	94 Million
Property	2.5 Billion
Sex Offenders	550,000
SSN's	7.2 Billion
Student Records	38 Million
TIN	2.9 Million
Unique ADLs - active	257 Million
Utility	645 Million
Vehicle Titles	635 Million
Vehicle Registrations	2.5 Billion
White Pages	116 Million
Wireless Phones	101 Million
Yellow Pages	14 Million

Case Study #1 (Enterprise) – LexisNexis Risk Solutions



- Ingest disparate Data Source
- Manage hundreds of terabytes of data

- Parallel Processing Architecture
- Discover non-obvious links and detect hidden patterns
- Language optimized for data-intensive application

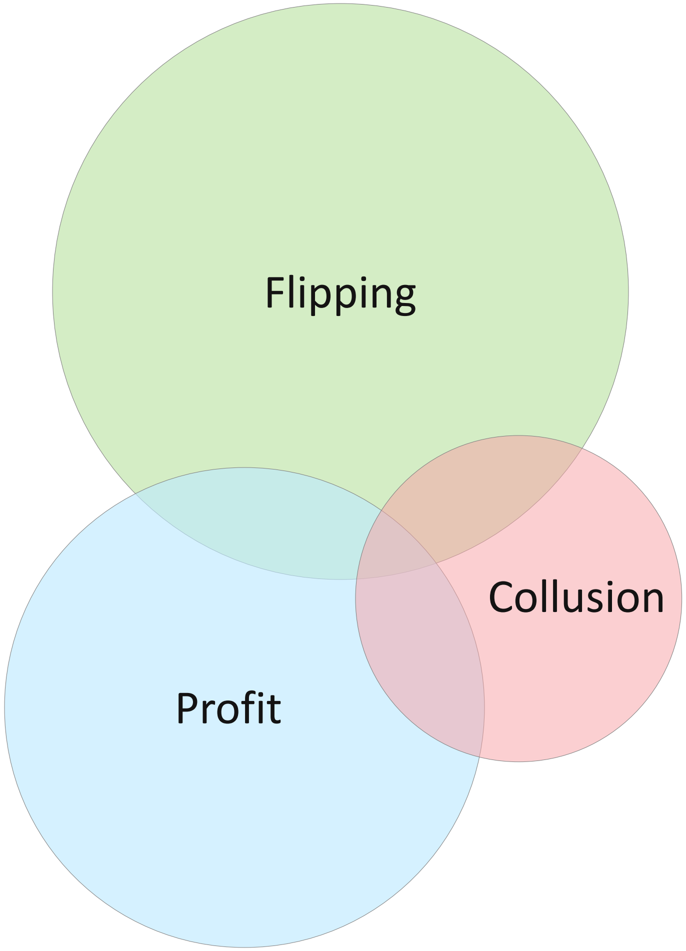
- Service Oriented Architecture(SOA)
- Flexibility to organize and repurpose data

Case Study #2 (Boil the Ocean) – Property Transaction Risk

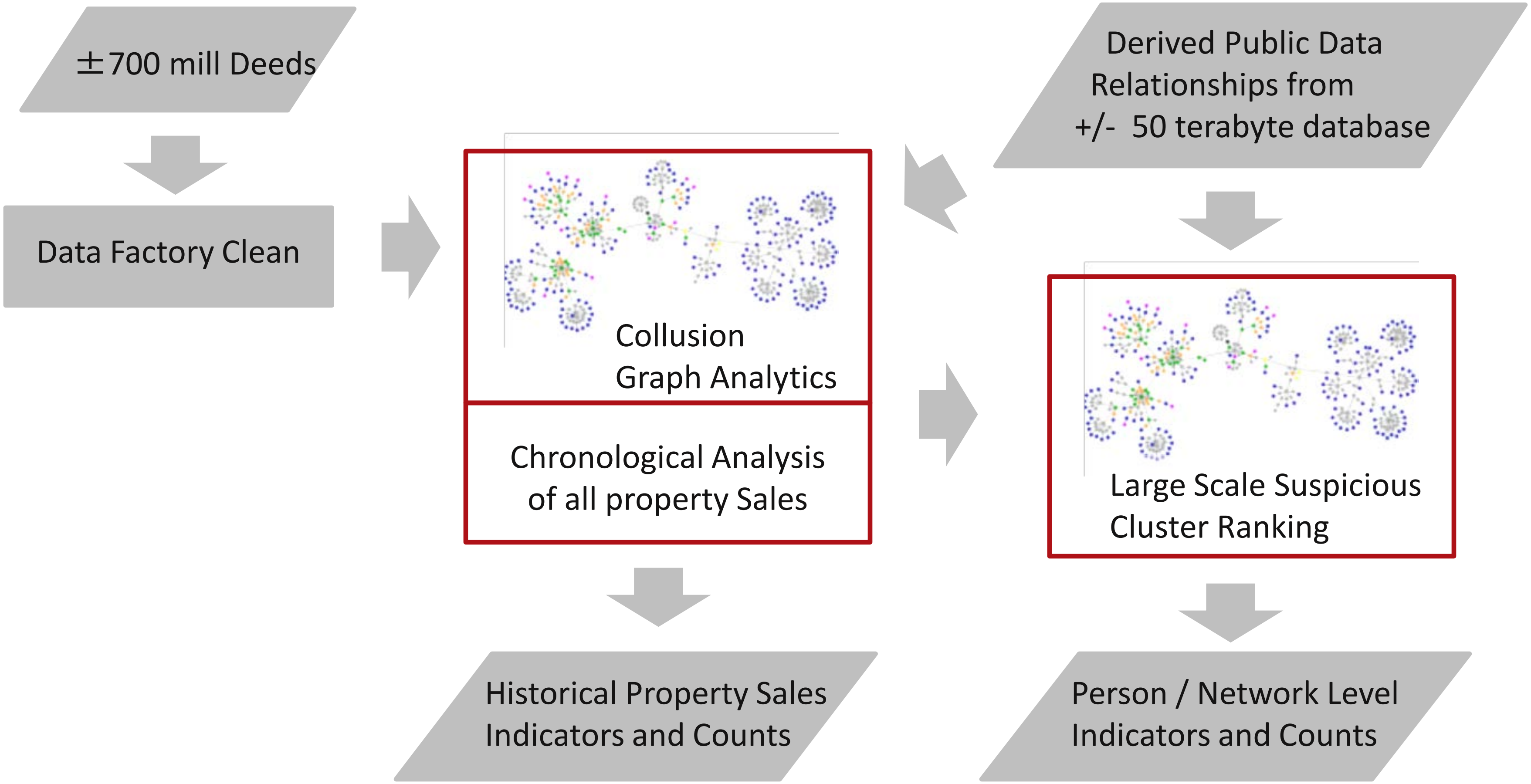
Case Study #2 (Boil the Ocean) – Property Transaction Risk

Three core transaction variables measured

- Velocity
- Profit (or not)
- Buyer to Seller Relationship Distance (Potential of Collusion)



Case Study #2 (Boil the Ocean) – Property Transaction Risk



±700 mill Deeds

Data Factory Clean

Collusion
Graph Analytics

Chronological Analysis
of all property Sales

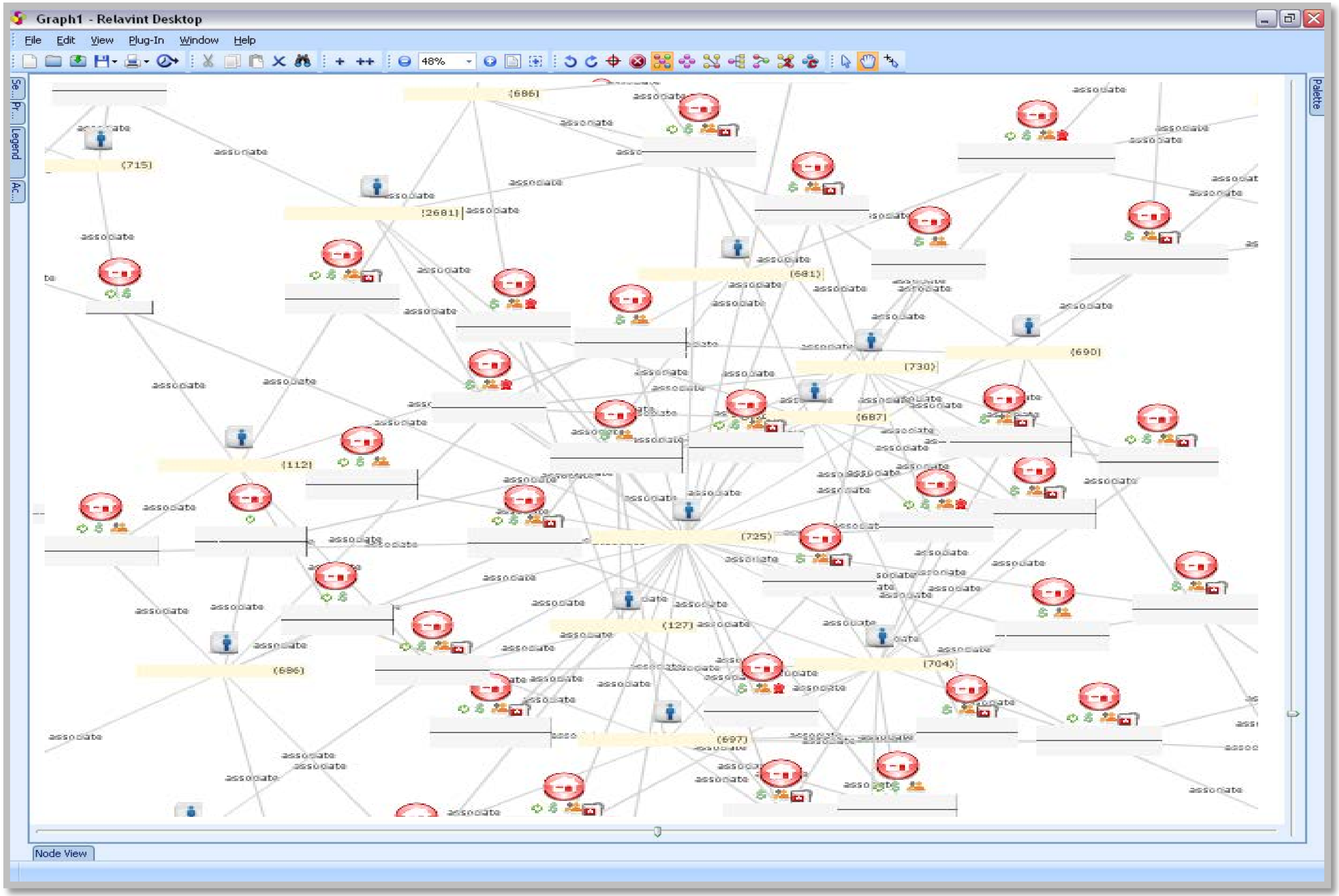
Historical Property Sales
Indicators and Counts

Derived Public Data
Relationships from
+/- 50 terabyte database

Large Scale Suspicious
Cluster Ranking

Person / Network Level
Indicators and Counts

Case Study #2 (Boil the Ocean) – Property Transaction Risk



Case Study #2 (Boil the Ocean) – Property Transaction Risk

Large scale measurement of influencers strategically placed to potentially direct suspicious transactions.

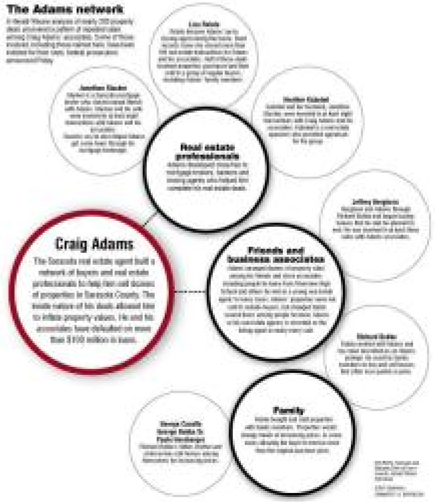
- All data on one supercomputer measuring over a decade of property transfers nationwide.

- Data Products to turn other Data into compelling intelligence.

- Large Scale Graph Analytics allow for identifying known unknowns.

- Florida Proof of Concept
  - Highest ranked influencers
    - Identified known ringleaders in flipping and equity stripping schemes.
    - Typically not connected directly to suspicious transactions.
  - Known ringleaders not the Highest Ranking.

- Clusters with high levels of potential collusion.
- Clusters offloading property, generating defaults.
- Agile Framework able to keep step with emerging schemes in real estate.


The Adams network
Craig Adams
The Sarasota real estate agent built a network of buyers and real estate professionals to help him sell dozens of properties in Sarasota County. The inside nature of his deals allowed him to inflate property values. He and his associates have defaulted on more than $100 million in loans.
Real estate professionals
Friends and business associates
Family

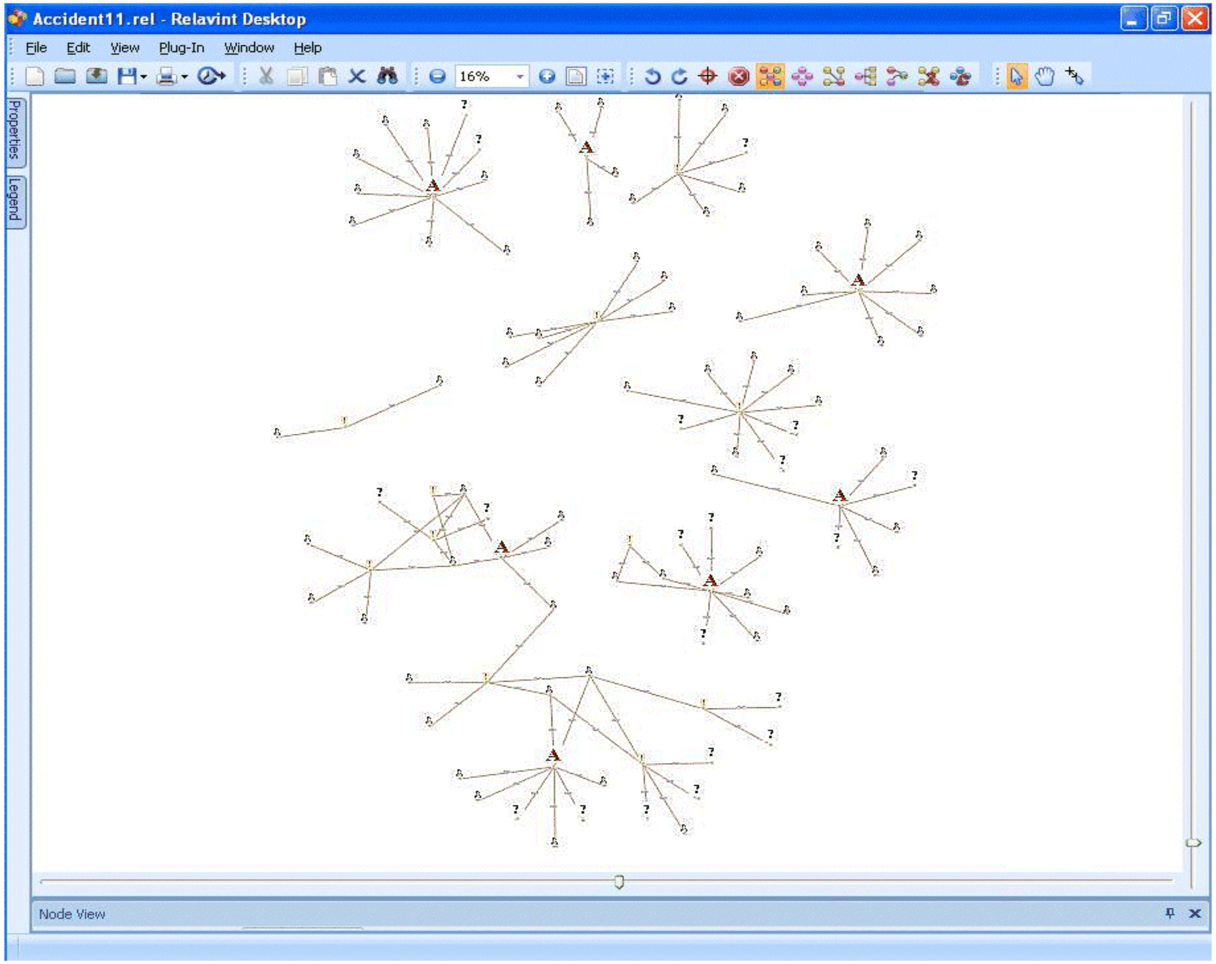
Case Study #3 (Serendipity) – Family Ties

Case Study #3 (Serendipity) – Family Ties between Claims

Scenario

This view of carrier data shows seven known fraud claims and an additional linked claim.

The Insurance company data **only finds a connection between two of the seven claims**, and only identified one other claim as being weakly connected.



Case Study #3 (Serendipity) – Family Ties between Claims

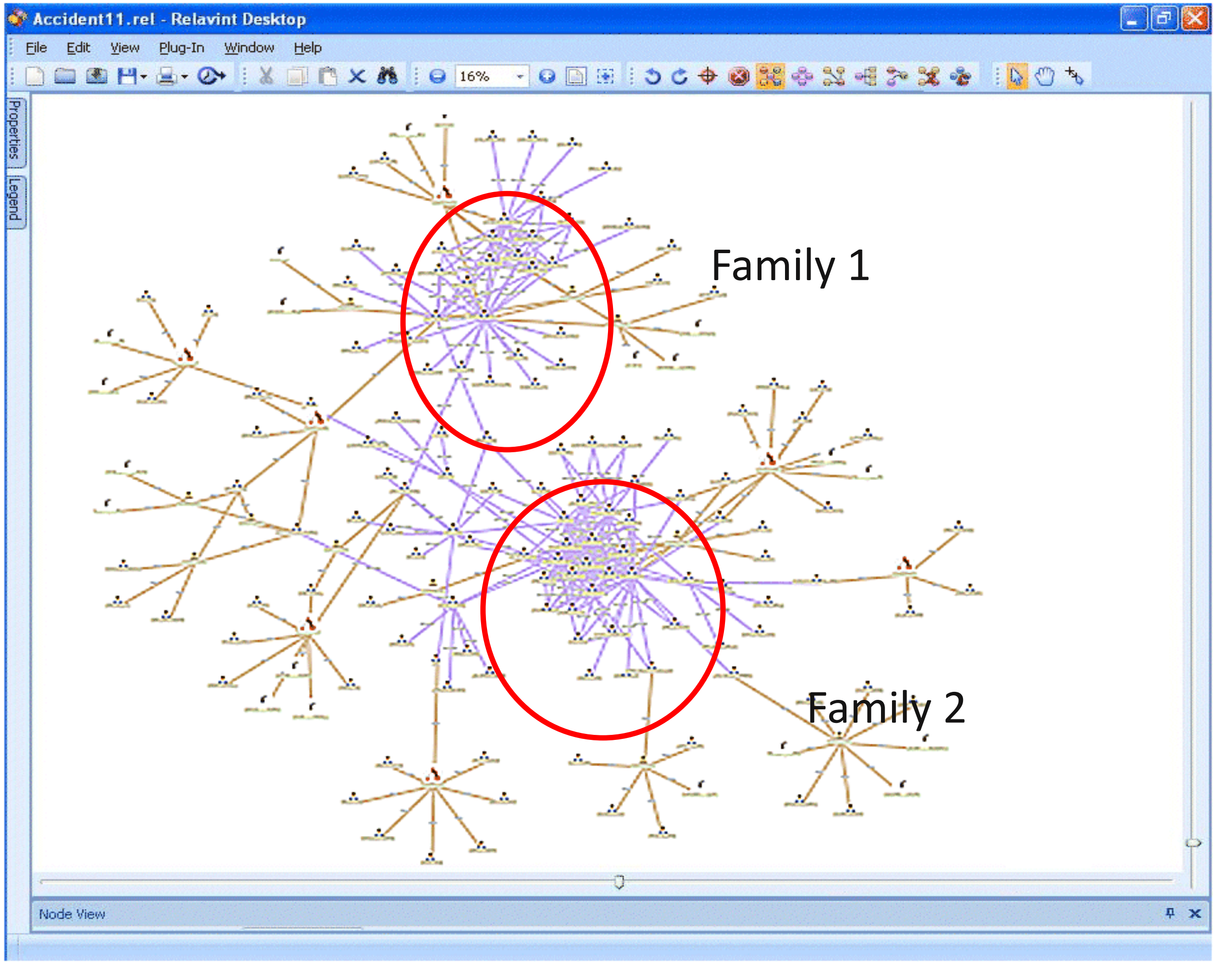
Task

After adding the LexID to the carrier Data, LexisNexis HPCC technology then explored 2 additional degrees of relative separation

Result

The results showed **two family groups interconnected on all of these seven claims**.

The links were much stronger than the carrier data previously supported.



Case Study #4 (Tactical) – 30 Hour Job

Case Study #4 (Tactical) – 30 Hour Job

Objective:

- Re-engineer long-running legacy process (proof-of-concept)
- 3m+ rows in … 500m rows out … 30+ hours
- Use similar hardware to maximize comparability

Pre-Coding Set Up:

- Legacy developers … completed on-line ECL training
- 1 ECL developer … exposed to legacy process and data
- Dump of input data and known result files
- Create HPCC hardware environment comparable to legacy environment
    - Amazon AWS 4 x m1.xlarge total (3 x m1.xlarge Thor Slaves)
    - 12-slave CPUs

Coding:

- Meet-up for 1 week coding session

Case Study #4 (Tactical) – 30 Hour Job

Legacy Environment:

- Oracle on Intel SMP
- 16 cores

HPCC Environment:

- Amazon AWS
- 1 x m1.xlarge (support node)
- 3 x m1.xlarge (Thor Slaves)
- 12-slave Cores

Results:

- 3 Days of Coding
- 450-ish Lines of ECL
- Legacy Run-Time: 30+ hours
- HPCC Run-Time: 1.5 hours

Additional Infomation

- LexisNexis Open Source HPCC Systems Platform: http://hpccsystems.com
- Free Online Training: http://learn.lexisnexis.com/hpcc
- SALT: http://hpccsystems.com/products-and-services/products/modules/SALT
- Machine Learning portal: http://hpccsystems.com/ml
- The HPCC Systems blog: http://hpccsystems.com/blog
- Community Forums: http://hpccsystems.com/bb
- Our GitHub portal: https://github.com/hpcc-systems
- JIRA: https://track.hpccsystems.com

Thank you!