HPCC Systems

Open Source, Big Data Processing and Analytics

LexisNexis®

Risk Solutions

LexisNexis

LexisNexis leverages data about people, businesses, and assets to assess risk and opportunity associated with industry-specific problems

Insurance	Assess underwriting risk and verify applicant data; prevent/ investigate fraudulent claims; optimize policy administration
Financial Services	Prevent/investigate money laundering and comply with laws Prevent/investigate identity fraud
HPCC Systems technology	Open Source and commercial offerings of our leading Big Data platform
Receivables Management	Assist collections by locating delinquent debtors Assess collectability of debts and prioritize collection efforts
Legal	Locate/vet witnesses, assess assets/associations of parties in legal actions; Perform diligence on prospective clients (KYC)
Government	Locate missing children/suspects; research/document cases; reduce entitlement fraud; accelerate revenue collection
Health Care	Verify patient identity, eligibility, and ability to pay Validate provider credentials; prevent/investigate fraud

LexisNexis Risk Solutions works with Fortune 1000 and mid-market clients across all industries and both federal and state governments.

- Customers in over **139** countries
- **6** of the world's top 10 banks
- **100%** of the top 50 U.S. banks
- **80%** of the Fortune 500 companies
- **100%** of U.S. P&C insurance carriers
- **All 50** U.S. states
- **70%** of local governments
- **80%** of U.S. Federal agencies
- **97** of AM Law 100 firms

HPCC Systems

LexisNexis Big Data Value Chain

Collection – Structured, unstructured and semi-structured data from multiple sources

Ingestion – loading vast amounts of data onto a single data store

Discovery & Cleansing – understanding format and content; clean up and formatting

Integration – linking, entity extraction, entity resolution, indexing and data fusion

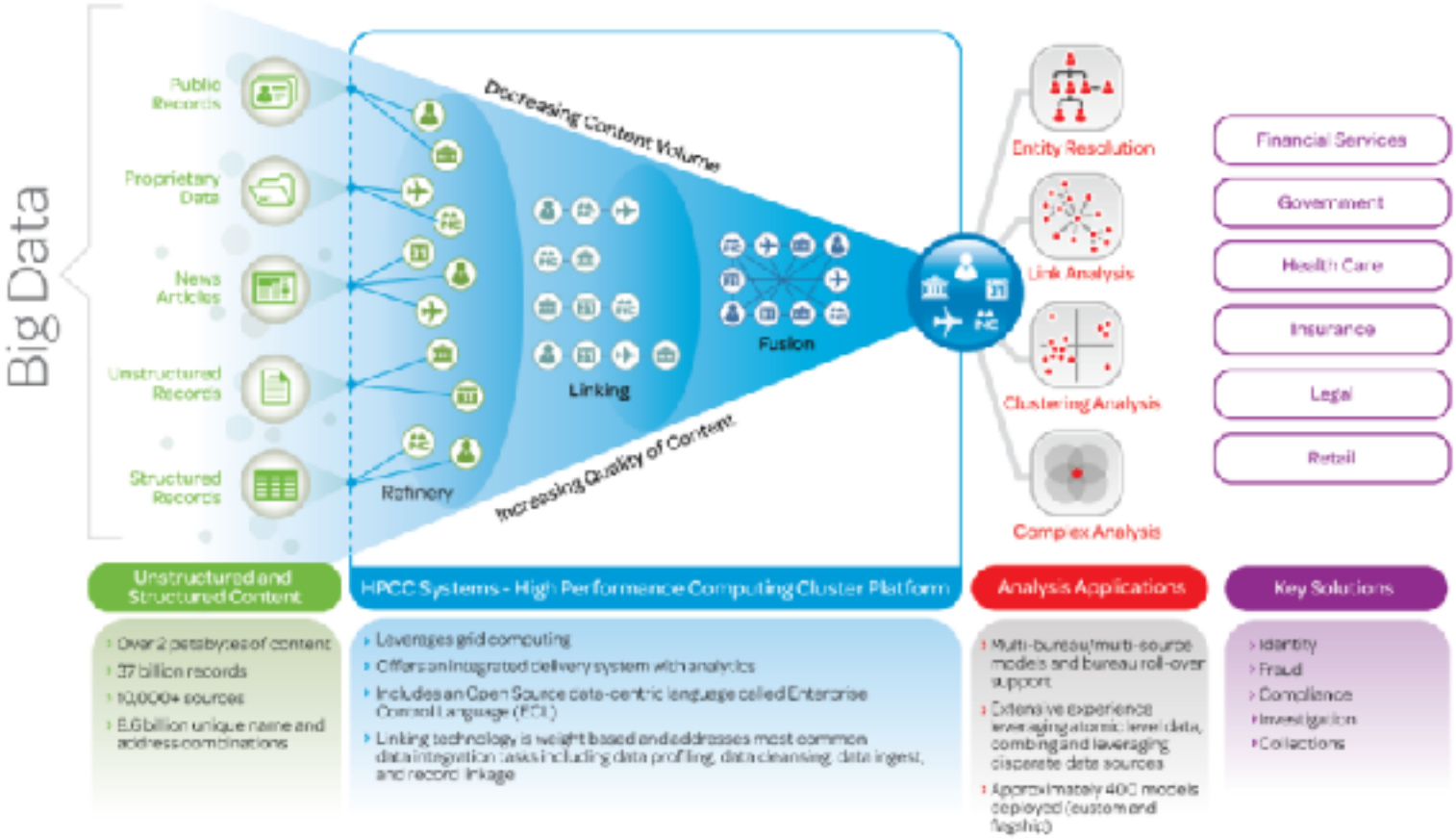
Analysis – Intelligence, statistics, predictive and text analytics, machine learning

Delivery – querying, visualization, real time delivery on enterprise-class

Sample Use Cases for HPCC Systems

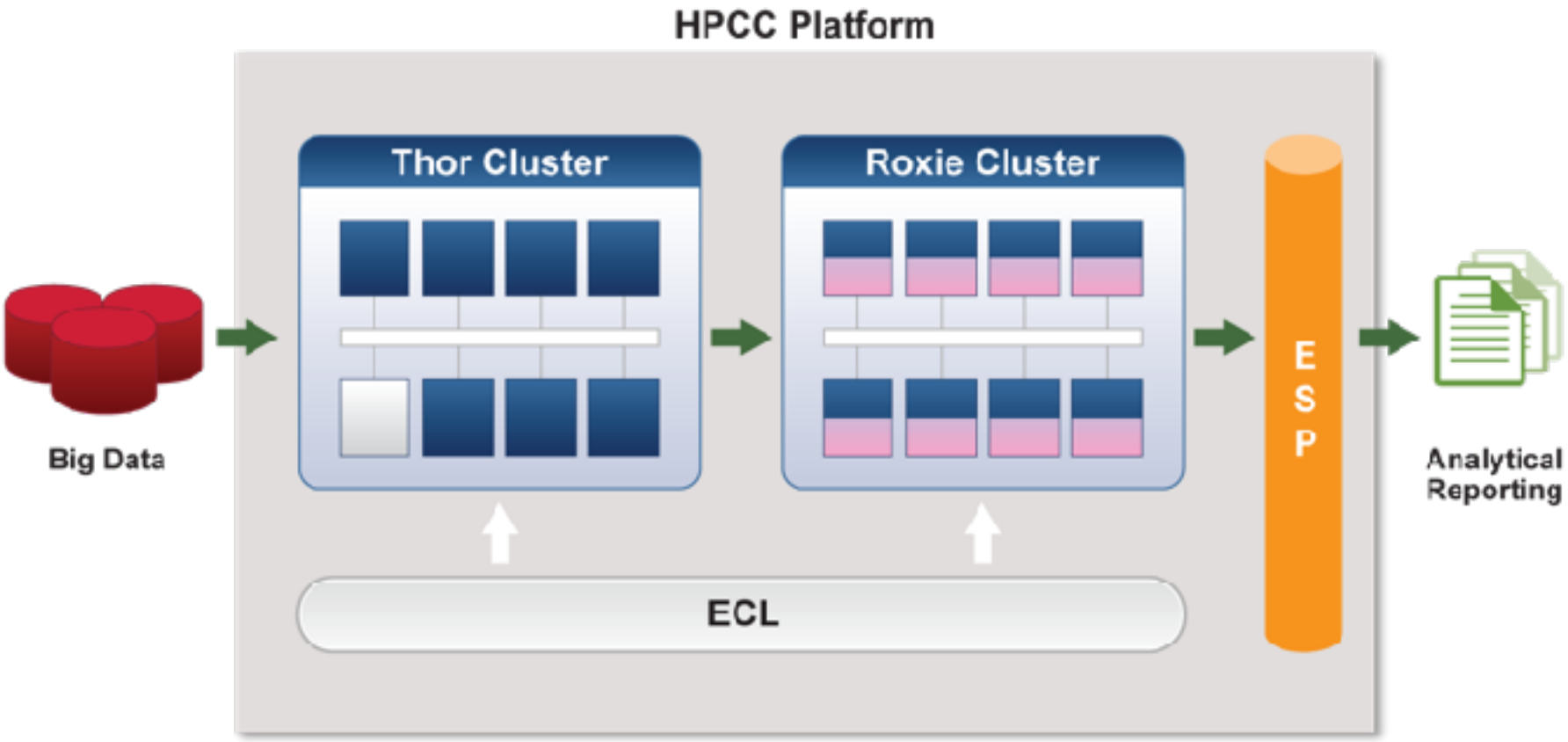
Vertical	Example
Automate Identification	Automate Identity Identification and disambiguation for people, businesses, assets
Complex Queries	Run complex queries on transaction data (years, months, decades) to see past and present behavior
Predict Behavior	Predict behavior and patterns leveraging current and historical data from a variety of data sources
Financial Services	Analyze current and historical consumer transaction data from various sources to see fraud patterns or opportunities during a lifetime cycle of a customer.
Government	Manage and analyze the flow of people, businesses and assets from various data sources to monitor transactions and fraud rings.
Health Care/Insurance	Manage customer/patient transaction data on electronic medical records / claims including medical information, medical procedures, prescriptions, etc.
Internet	Monitor social media outlets for trends and patterns. Monitor all the data on your network to guard against cyber security attacks.
Retail	Capture and analyze information from all branches of stores, locations, departments to manage pricing, inventory, distribution.
Telecommunications	Processing customer data such as billions of call records, texts, streaming media and GPS history. Analyze customer churn, usage behavior patterns, failure.

HPCC Systems is an Open Source Platform for Big Data Processing



- **High Performance Computing Cluster (HPCC Systems)** enables data integration on a scale not previously available and real-time answers to millions of users. Built for Big Data and proven for 10 years with enterprise customers.
- **ECL Parallel Programming Language** optimized for business differentiating data intensive applications
- **Single architecture** offers two data platforms (query and refinery) and a consistent data-intensive programming language (ECL)

HPCC Systems Architecture

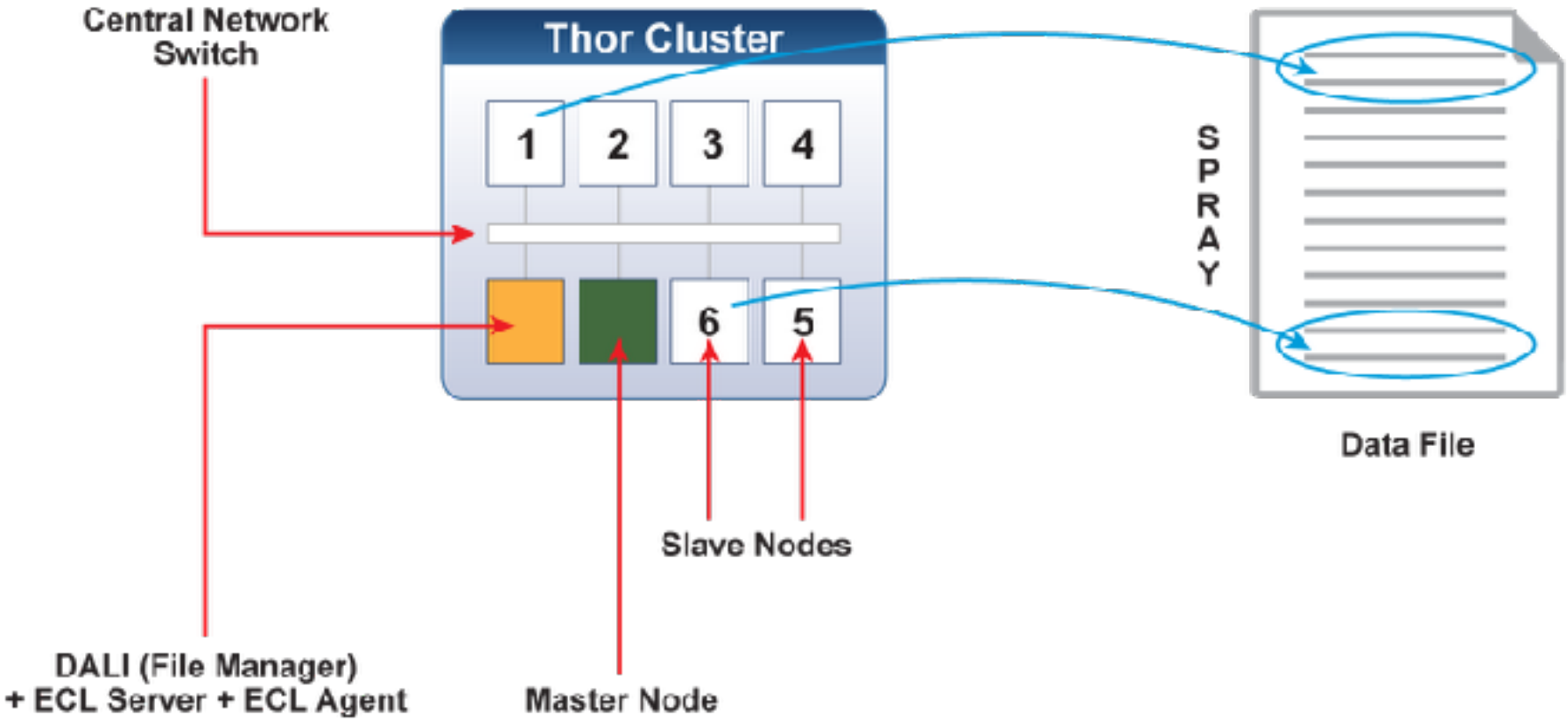


How HPCC Systems Helps Address the Talent Crunch

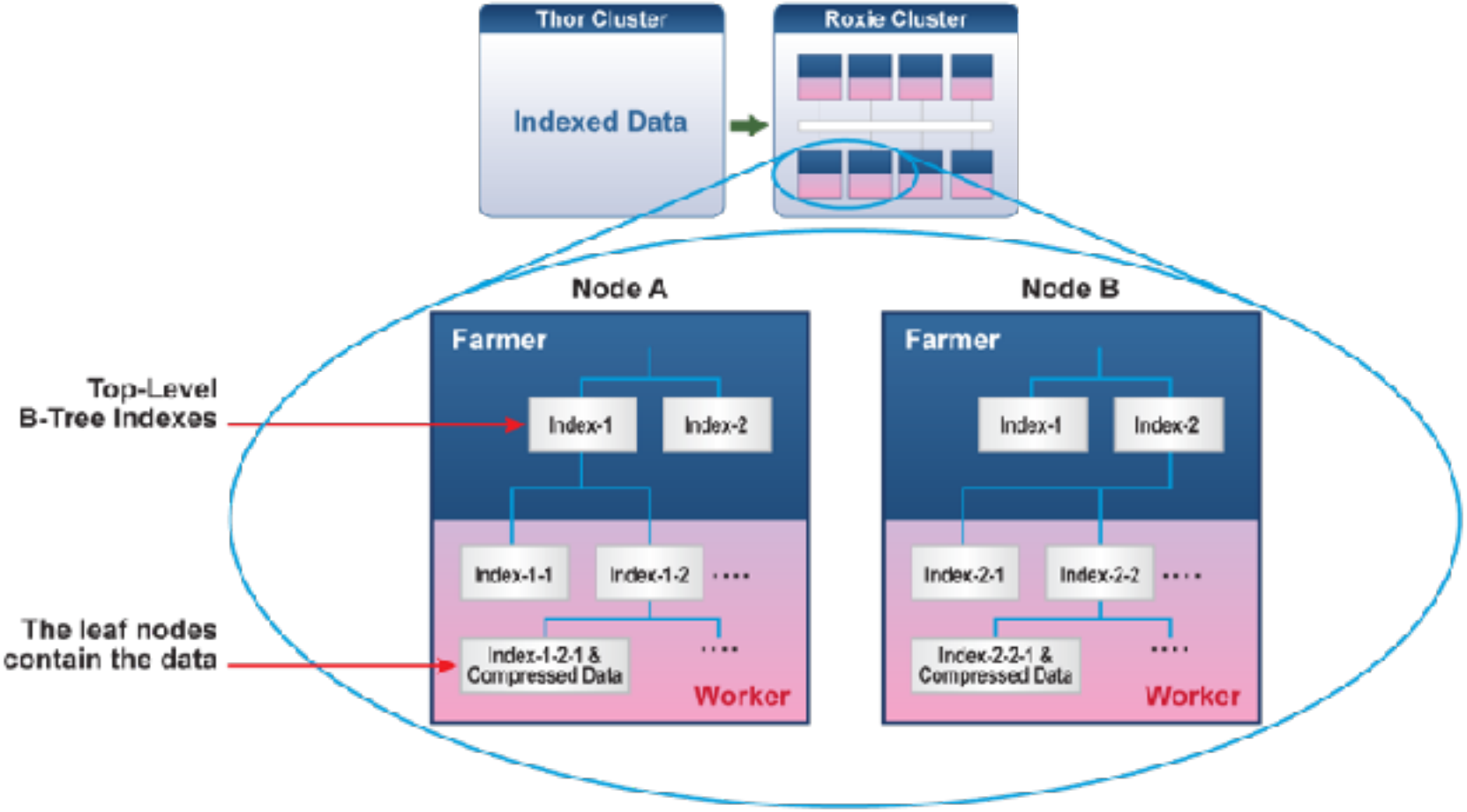
Types of Users:	Developers	Linkers	Modelers	Data Scientists	BI Analysts	Business Managers
What they do	Ingest data Write queries Create products	Link datasets	Discover attributes, create models & scoring logic	Discover patterns Write algorithms	Understand patterns Discover anomalies	Monitor business performance and customer needs, dynamics
Talent & Work Crunch Issue	Developers want a faster, better way to solve problems. Hadoop requires many developers, which is expensive for firms	Hard for companies to find and hire. They need a better, faster way to understand how to link data for analysis.	Hard for companies to find and hire. They need a better, faster way to model data so that are spending their time on analysis, not programing models	Hard and expensive for companies to find and hire. Need more time on analysis not low-level data manipulation tasks.	Hard and expensive for companies to find and hire. Need more time on analysis not low-level data manipulation tasks.	They struggle with IT and data solutions to get the data and insight they need to make business decisions. Most IT and data solutions are not connected.
High-level tools:	HPCC Flow, KEL	Smart View, SALT, KEL	HPCC Flow, EDA	HPCC Flow, Circuits, RAMPS	Scored Search, Dashboard	Dashboard
Low-level tools:	ECL	N/A	N/A	KEL, SALT, ECL	N/A	N/A

Deep Dive on HPCC Systems Components

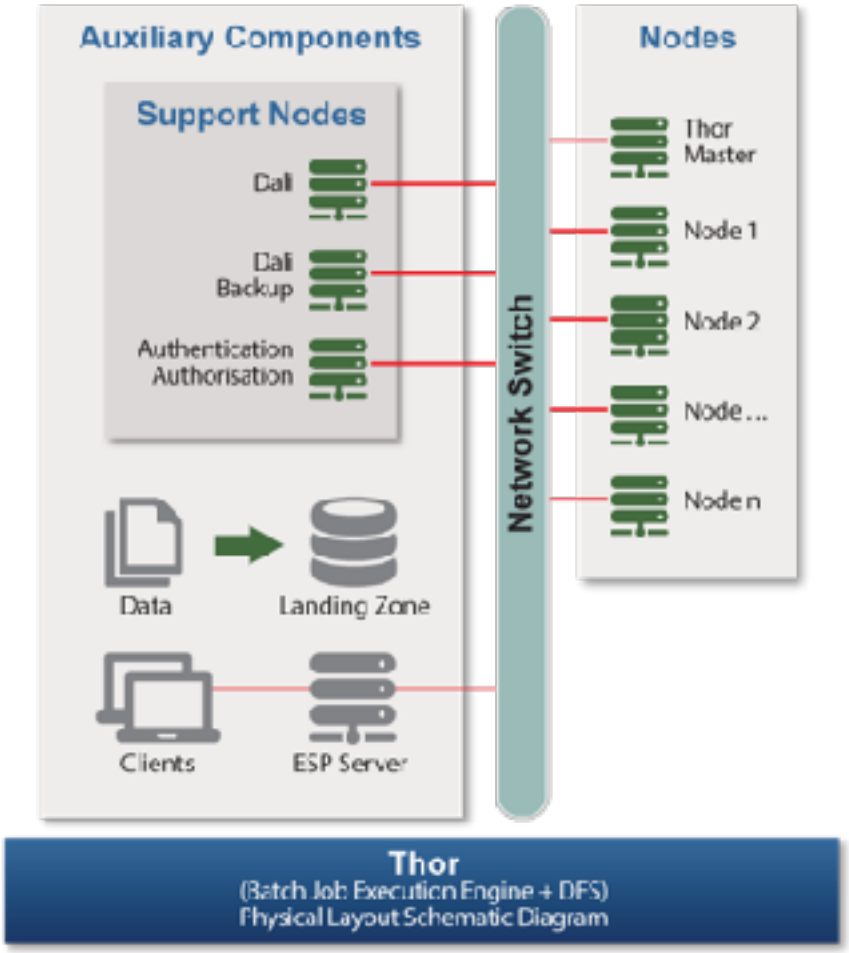
Thor Logical Architecture

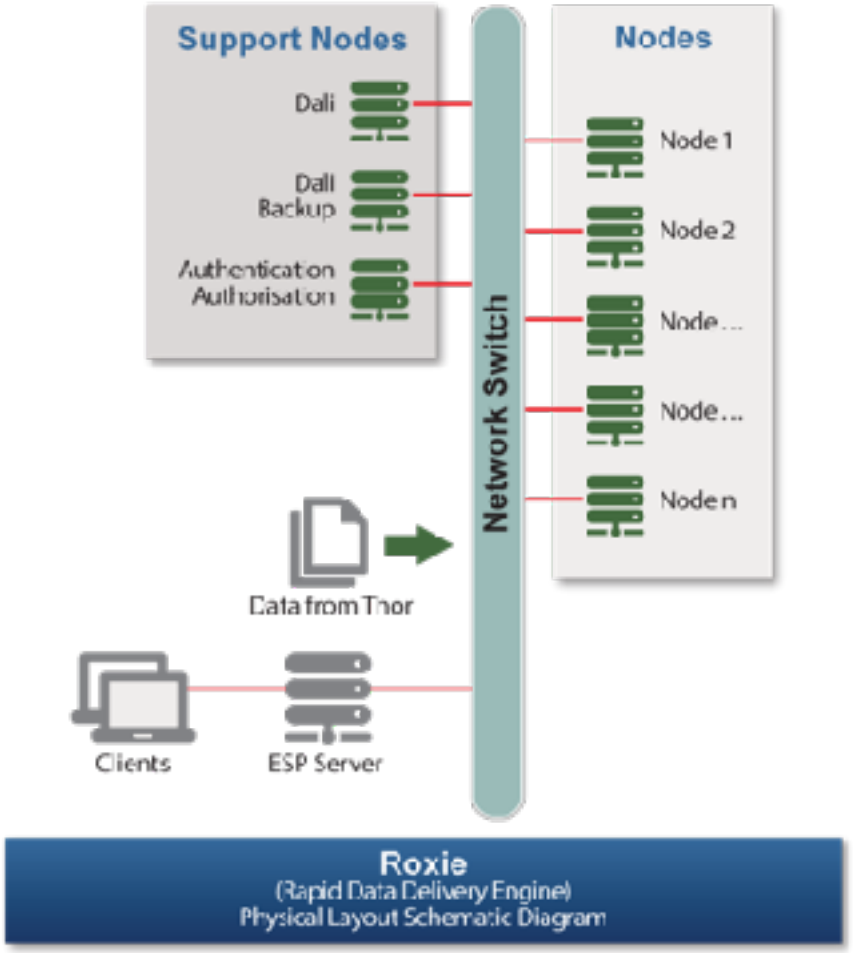


Roxie Logical Architecture



Thor Physical Architecture





Data Refinery Process

Programming Language is called Enterprise Control Language (ECL)

Declarative programming language: Describe **what** needs to be done and **not how** to do it

Powerful: Unlike Java, high level primitives as JOIN, TRANSFORM, PROJECT, SORT, DISTRIBUTE, MAP, etc. are available. Higher level code means fewer programmers & shortens time to delivery

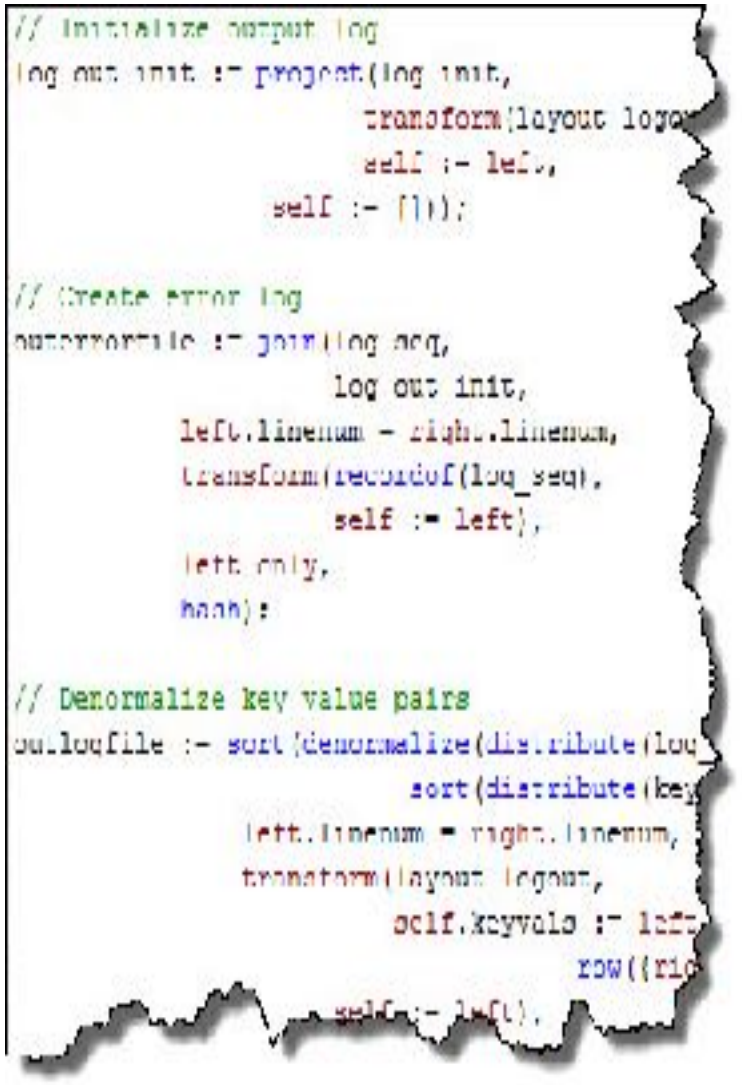
Extensible: As new attributes are defined, they become primitives that other programmers can use

Implicitly parallel: Parallelism is built into the underlying platform. The programmer needs not be concerned with it

Maintainable: A high level programming language, no side effects and attribute encapsulation provide for more succinct, reliable and easier to troubleshoot code

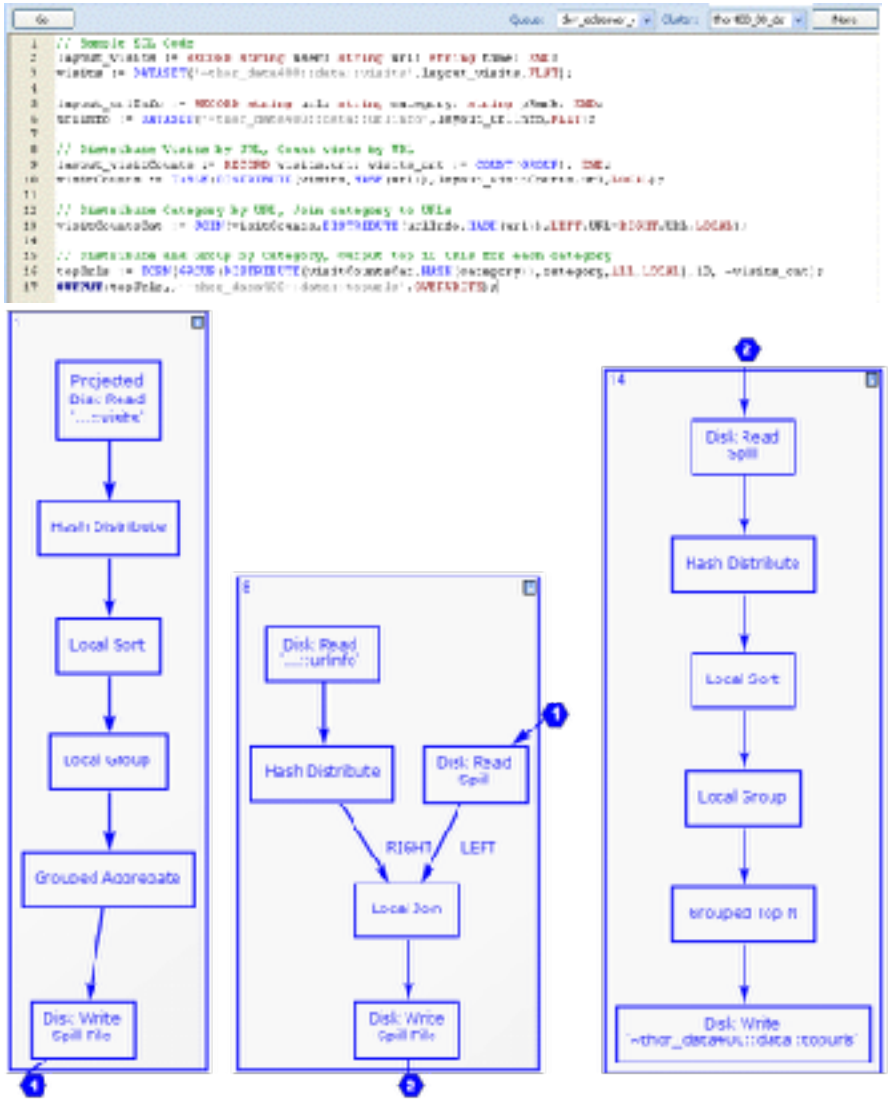
Complete: ECL provides for a complete data programming paradigm

Homogeneous: One language to express data algorithms across the entire HPCC platform, including data ETL and high speed data delivery



Programming Language is called Enterprise Control Language (ECL)

- ECL is a declarative, data-centric, programming language which can expressed concisely, parallelizes naturally, is free from side effects, and results in highly-optimized executable code.
- ECL is designed for a specific problem domain (data-intensive computing), which makes resulting programs clearer, more compact, and more expressive. ECL provides a more natural way to think about data processing problems for large distributed datasets.
- Since ECL is declarative, execution is not determined by the order of the language statements, but from the sequence of dataflows and transformations represented by the language statements. The ECL compiler determines the optimum execution strategy and graph.
- ECL incorporates transparent and implicit parallelism regardless of the size of the computing cluster and reduces the complexity of parallel programming increasing the productivity of application developers.
- The ECL compiler generates highly optimized C++ for execution.
- ECL provides a comprehensive IDE and programming tools including an Eclipse plugin.
- ECL is provided with a large library of efficient modules to handle common data manipulation tasks.



Scalable Automated Linking Technology (SALT) and LexID

SALT - Algorithms for Data Analysis and Linking

Data Ingest
- Transform a base file into a rolling historic record of incoming records.

Data Profiling
- Generate statistics for field lengths, word counts, population rate, etc.

Data Hygiene
- Format fields correctly at the character level (i.e., capitals, punctuation, spelling errors, etc.)

Entity Resolution
- Create 360^0 profiles on individuals, organizations, etc. through iteratively linking data from multiple sources

Relationship Extraction
- Uncover hidden patterns within massive datasets to determine how entities relate to one another

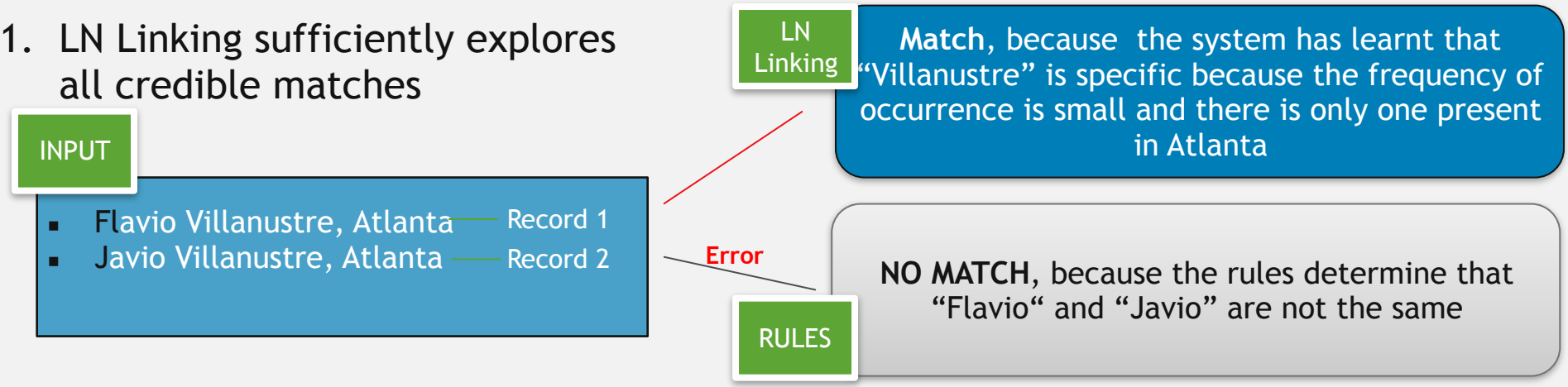
SALT

Scalable Automated Linking Technology

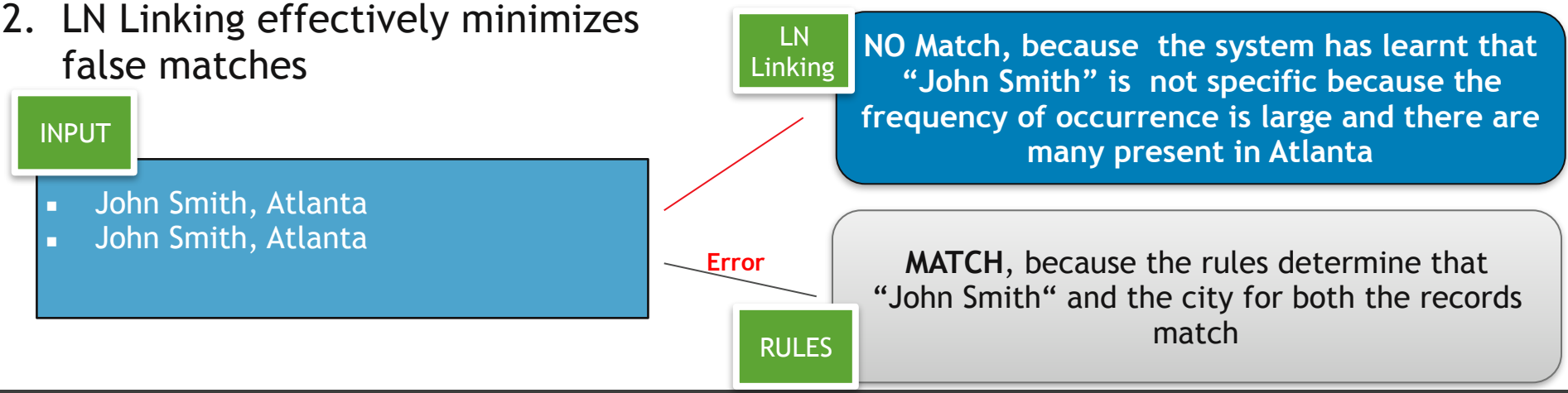
LexisNexis linking is statistically-based (not rules-based), using the LexisNexis master databases for reference

Two examples where linkage strictly based on rules doesn't work:

1. LN Linking sufficiently explores all credible matches



2. LN Linking effectively minimizes false matches



The secret sauce in our portfolio is LexID℠

LexID℠

The fastest linking technology platform available with results that help you make intelligent information connections.

LexID℠ is the ingredient behind our products that turns disparate information into meaningful insights. This technology enables customers using our products to identify, link and organize information quickly with a high degree of accuracy.

Get a more complete picture

Make intelligent information connections beyond the obvious by drawing insights from both traditional and new sources of data.

Better results, faster

Use the fastest technology for processing large amounts of data to help you solve cases more quickly and confidently.

Protect private information

Keep customer SSNs and FEINs secure and enjoy peace of mind knowing you are taking steps to observe the highest levels of privacy and compliance.

Case Studies

Case Study: Fraud in Medicaid

Scenario

Proof of concept for Office of the Medicaid Inspector Generation (OMIG) of large Northeastern state. Social groups game the Medicaid system which results in fraud and improper payments.

Task

Given a large list of names and addresses, identify social clusters of Medicaid recipients living in expensive houses, driving expensive cars.

Result

Interesting recipients were identified using asset variables, revealing hundreds of high-end automobiles and properties.

Leveraging the Public Data Social Graph, large social groups of interesting recipients were identified along with links to provider networks.

The analysis identified key individuals not in the data supplied along with connections to suspicious volumes of “property flipping” potentially indicative of mortgage fraud and money laundering

Make Description	#	Make Description	#
Mercedes-Benz	46	Chevrolet	2
Lexus	41	Hummer	2
BMW	27	Jeep	2
Infiniti	17	Nissan	2
Acura	9	Toyota	2
Lincoln	8	Aston Martin	1
Audi	7	Bentley	1
Land Rover	7	Cadillac	1
Porsche	6	GMC	1
Jaguar	5	Honda	1
Mercedes Benz	3	Volkswagen	1
Saab	3	Volvo	1

Case Study: Fraud in Prescription Drugs

Scenario

Healthcare insurers need better analytics to identify drug seeking behavior and schemes that recruit members to use their membership fraudulently.

Groups of people collude to source schedule drugs through multiple members to avoid being detected by rules based systems.

Providers recruit members to provide and escalate services that are not rendered.

Task

Given a large set of prescriptions. Calculate normal social distributions of each brand and detect where there is an unusual socialization of prescriptions and services.

Result

The analysis detected social groups that are sourcing Vicodin and other schedule drugs. Identifies prescribers and pharmacies involved to help the insurer focus investigations and intervene strategically to mitigate risk.

Case Study: Insurance Score From 100 Days To 30 Minutes

Scenario

One of the top 3 insurance providers using Oracle analytics products on multiple statistical model platforms with disparate technologies

- Insurer issues request to re-run all past reports for a customer: 11.5M reports since 2004
- Using their current technology infrastructure it would take 100 days to run these 11.5M reports

Task

Migration of 75 different models, 74,000 lines of code and approximately 700 definitions to the HPCC

- Models were migrated in 1 man month.
- Using a small development system (and only one developer), we ran 11.5 million reports in 66 minutes
- Performance on a production-size system: 30 minutes
- Testing demonstrated our ability to work in batch or online

Result

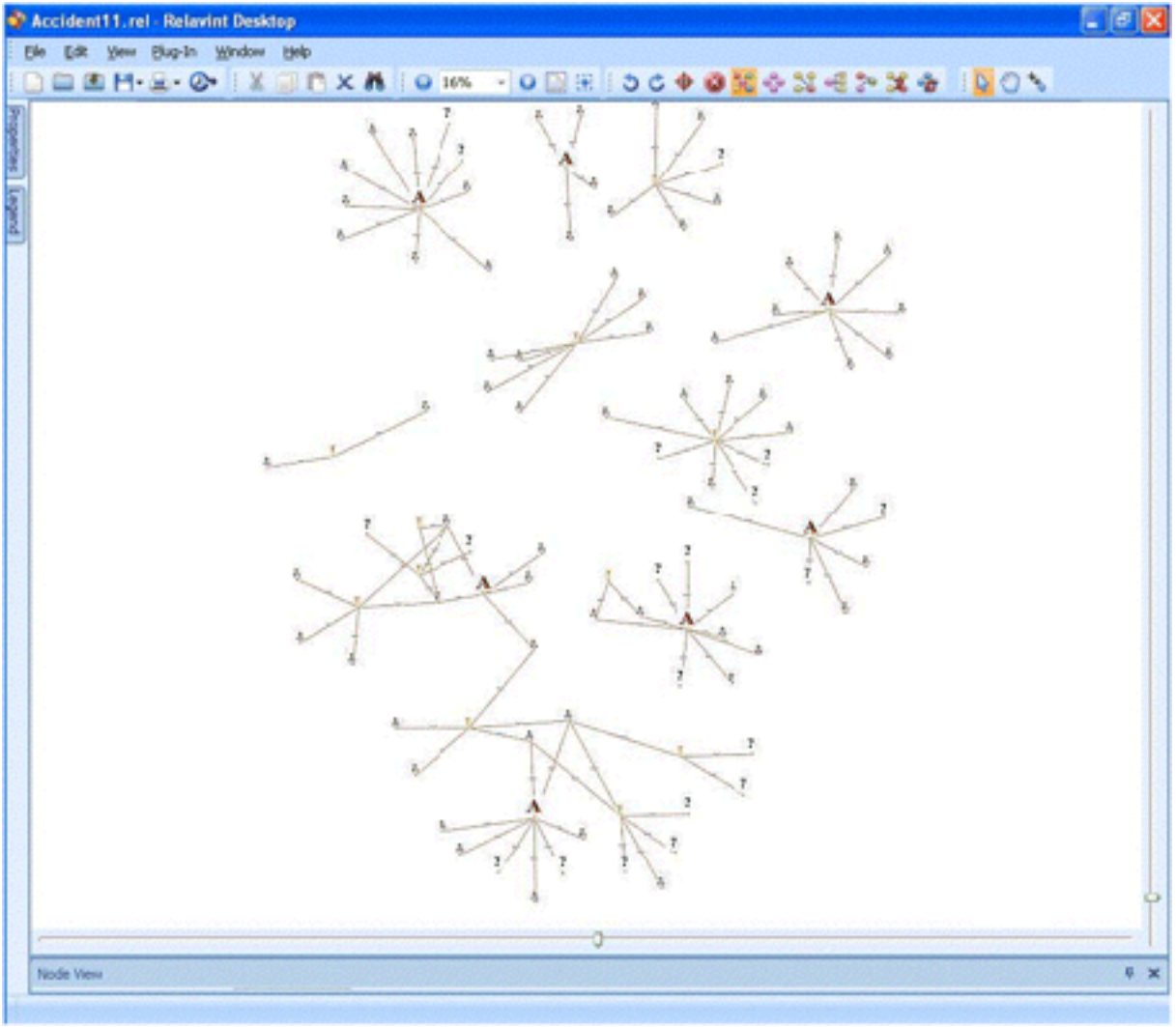
- Reduced manual work, increases reliability, and created capability to do new scores
- Decreased development time from 1 year to several weeks; decreased run time from 100 days to 30 minutes
- One HPCC (one infrastructure) translates into less people maintaining systems.

Case Study: Fraud & Collusion in Auto Insurance 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: Fraud & Collusion in Auto Insurance Claims

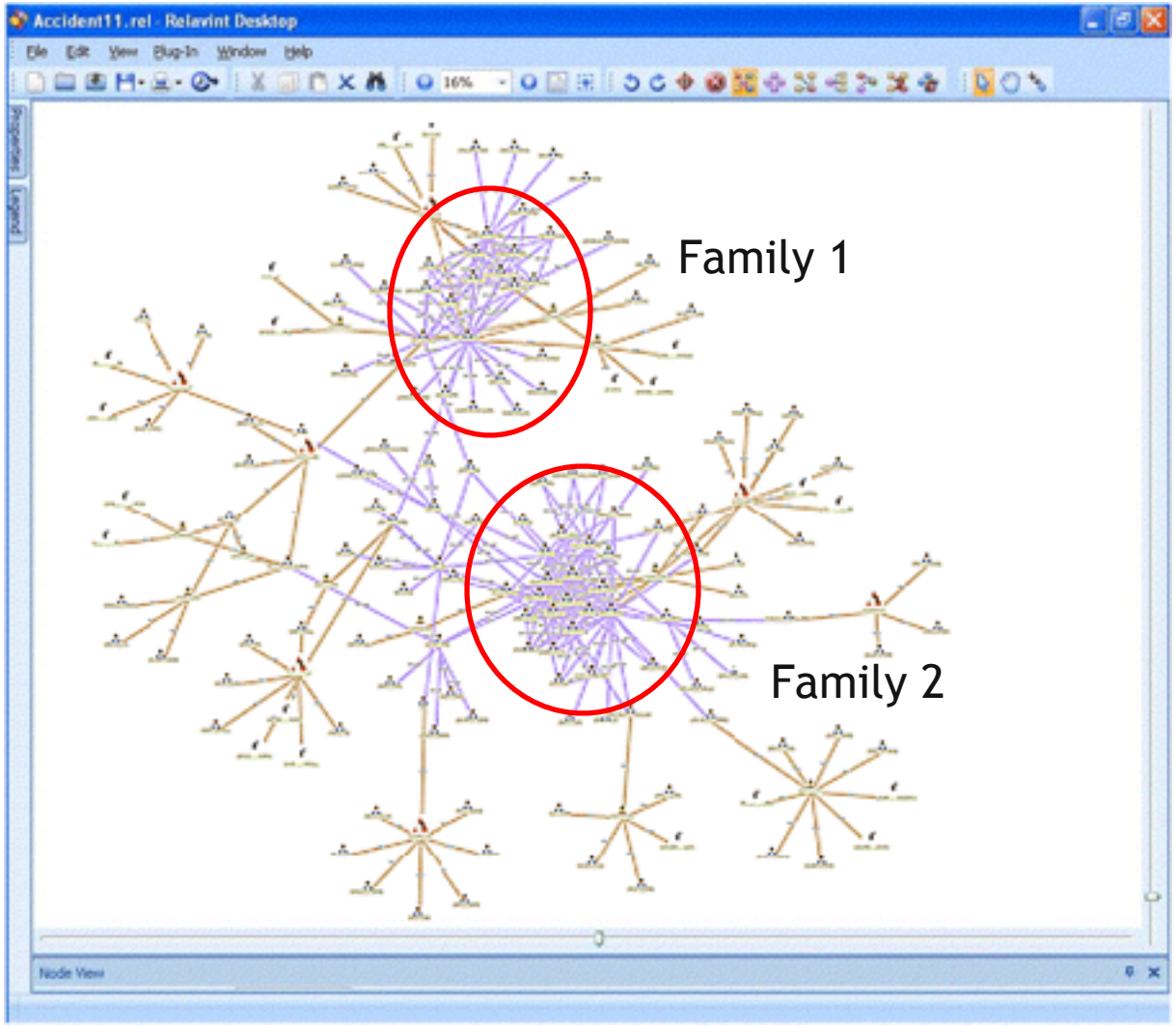
Task

After adding the LexID to the carrier Data, LexisNexis HPCC technology then added 2 additional degrees of associations

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: Network Traffic Analysis in Seconds

Scenario

Conventional network sensor and monitoring solutions are constrained by inability to quickly ingest massive data volumes for analysis

- 15 minutes of network traffic can generate 4 Terabytes of data, which can take 6 hours to process
- 90 days of network traffic can add up to 300 Terabytes

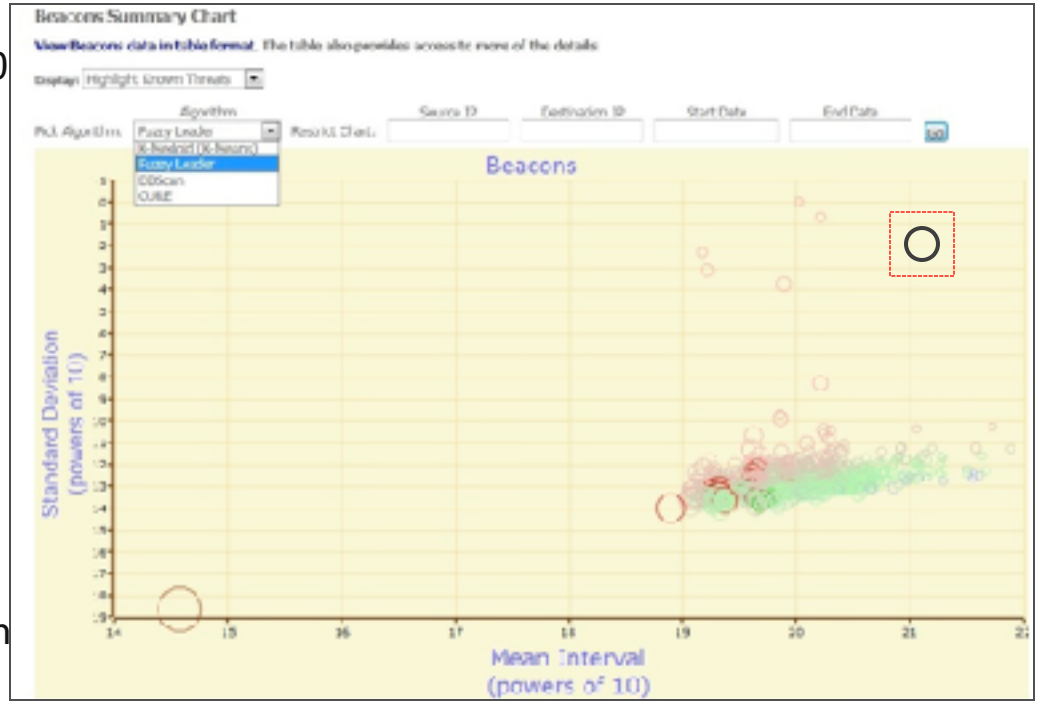
Task

Drill into all the data to see if any US government systems have communicated with any suspect systems of foreign organizations in the last 6 months

- In this scenario, we look specifically for traffic occurring at unusual hours of the day

Result

In seconds, HPCC Systems sorted through month of network traffic to identify patterns and suspicious behavior



Contact Us

info@hpccsystems.com

US: 1.877.316.9669

Intl: 1.678.694.2200