Industrial Big Data Analytics
Lessons from the trenches
Flavio Villanustre
LexisNexis Risk Solutions
**Big Data funnel**

- **Public Records**
- **Proprietary Data**
- **News Articles**
- **Unstructured Records**
- **Structured Records**

**High Performance Computing Cluster Platform (HPCC)**

- Grid Computing
- Data-centric language (ECL)
- Integrated Delivery System that offers data plus analytics

**Open Source Components**

**Analysis Applications**
- Multi-bureau/multi-source models and bureau roll-over support
- Extensive experience leveraging atomic level data, combing and leveraging disparate data sources
- Approximately 400 models deployed (custom and flagship)

**Key Capabilities**
- Data and Analytics
- Identity Verification and Authentication
- Fraud Detection and Prevention
- Investigation
- Screening
- Receivables Management

**Decreasing content volume**

**Increasing information density**
Open source distributed data processing and storage architecture

Splits problems into pieces to be worked in parallel by commodity servers

Data-centric dataflow language (ECL) and higher level abstractions (KEL/SALT/DSP)

Big Data” language brings the computing to the data

Scalable data processing and analytics for subject matter experts

A model providing hierarchical abstraction for efficiency and high performance

SprayResult := RECORD, MAXLENGTH(2048)
STRING dfuWUID;
STRING externalName;
STRING internalName;
INTEGER size;
END;
SprayResult doSpray(dirAlias l) := TRANSFORM
STRING fname := l.name[1..LENGTH(l.name)-4];
STRING dsname := '~THOR::Patents:' + fname + ':XML';
SELF.externalName := l.name;
SELF.size := l.size;
SELF.internalName := dsname;
SELF.dfuWUID := SprayFunction(l.name, dsname);
END;
ds1 := dirList;
OUTPUT(COUNT(ds1), NAMED('Files_2_Spray'));
ds3 := NOTHOR(PROJECT(ds1, doSpray(LEFT)));
OUTPUT(ds3, NAMED('THOR::Patents:SPRAY_Log::' + WORKUNIT));

http://hpccsystems.lexisnexis.com/
LexisNexis Technology: Our HPCC Architecture

Big Data Technology + Vest Data Resources + Linking & Analytics + Industry-Specific Expertise & Delivery = Customer-Focused Solutions

HPCC Platform

Data Refinery (Thor)  Data Delivery Engine (Roxie)  ESP  Analytical Reporting

Enterprise Control Language (ECL)
LexisNexis Technology: Our HPCC Architecture

Thor

- Massively Parallel Extract Transform and Load (ETL) engine
  - Designed for data parallel execution. Leverages inexpensive locally attached storage and moves compute functions to the data. Programmable using a dataflow model where nodes are activities and edges represent datasets flowing.

- Enables data integration on a scale not previously available:
  - Current LexisNexis core data linking process generates hundreds of billions of intermediate results at peak

- Ideal for:
  - Massive joins/merges
  - Large-scale sorts & transformations
  - Model training over very large high dimensional corpus
LexisNexis Technology: Our HPCC Architecture

Roxie

- A massively parallel, high throughput, structured query response engine
- Low latency queries with high fault tolerance
- Leverages distributed data and keys for scalable storage and retrieval
- Automatically exposes queries through SOAP, RESTful, JSON and JDBC interfaces
- Ideal for
  - Highly concurrent multi-user applications
  - Full text ranked Boolean search
  - On-the-fly regression models
  - Multi-dimensional and hierarchical real-time grouping and filtering of large datasets
LexisNexis Technology: Our HPCC Architecture

ECL

• An easy to use, data-centric programming language optimized for large-scale data management and query processing
• Provides a dataflow programming model: activities represented as nodes and datasets flow through the edges
• Highly efficient: automatically distributes workload across all nodes.
• Compiles to C++.
• Abstracts out the underlying details of parallelism, data distribution, redundancy, etc.
• Large library of efficient modules to handle common data manipulation tasks
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.
Task

• After adding the Link ID to the carrier Data, LexisNexis HPCC technology then added 2 additional degrees of relative

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 months of network traffic to identify patterns and suspicious behavior
Application of HPCC: Beacon Analysis

**HPCC powers sophisticated near real time analytics to address rapidly evolving threats.**

**Chart key**
- **Horizontal axis:** time on a logarithmic scale
- **Vertical axis:** standard deviation (in hundredths)
- **Bubble size:** number of observed transmissions

**Reading the chart**
Each bubble represents a group of communications between 2 points that occurred more than once.

Suspicious activity would be to the far right with low standard deviation, i.e., highly regular at long intervals to reduce likelihood of detection.

As threats get more sophisticated, the standard deviation will likely increase and other attributes will change as well.

**Design and Benchmark Testing**
- Sort 90 Billion Records in ~2 hours on 400 node systems
- Supported 14 dimensional queries with ~2 second response
- Provided scalable design to accommodate 300TB (90 days) of netflow data w/70 billion new records per day
Application of HPCC: Network Detection

Show me all the traffic over last 6 months leaving the network going to systems known to be part of BotNets...
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.
A Highly Efficient Programming Process
- Domain Specific Language compiles to ECL
- Provides automated data profiling, parsing, cleansing, normalization and standardization
- Sophisticated specificity based linking and clustering

Scalable Automated Linking Technology (SALT)

- 42 Lines of SALT
- 3,980 Lines of ECL
- 482,410 Lines of C++
• Calculates record matching field weights based on term specificity and matching weights
  o What is the chance that two records for “John Smith” refer to the same person? How about “Flavio Villanustre”?

• It also takes into account transitive relationships
  o What if these two records for “John Smith” were already linked to “Flavio Villanustre”? 
  o How many “John Smiths” does “Flavio Villanustre” know? 
  o Are these two “John Smith” records referring to the same person now?
We’ve patented a unique, multi-dimensional linking model for businesses

1. A business entity at an address (e.g., the red dot at one building/address)
2. The legal entity (combination of all three red dots across all three addresses)
3. A collection of separate legal entities that appear to be related that are located at an address.
4. The combination of the related legal entities (all red, blue, and gray dots across all addresses)
5. The roll-up of combined org entities into a larger entity
Large Scale Aggregated & Structured Group Behavior Analysis

Measuring the context of a single data-point within the ever changing stream of economies, geographies, social group, crisis, time and more.

- Grouping Mechanisms
  Social, Business, Stereotypical, Time/Crisis

- Aggregated Group Behavior
  average age, gender, income within a social group
  # houses owned in a social group
  # customers within a social group
  # home owners in a neighbourhood in default
  # derogatory risk within a social group

- Structured Group Behavior
  # properties sold between people in the same social group
  # sold from a social group where the next owner resulted in 1st payment default
  # prescriptions within a social group where prescriber and patient are in the same social group
  # people who have moved between the same neighborhoods within the same time frame
  # change over time of derogatory risk within a social group

- Graph Analysis Overview
- Address Risk Analysis
- Geo-Social Analysis
Graph Analysis Overview – Social Graph

- Shared Historical Addresses
  - 12 Billion +

- Shared Assets
  - (Property, Vehicles, Watercraft, etc.)
  - 700 million + Property Deeds

- Shared Business Ownership
  - 1.5 Billion +

4 Billion Relationships

Risk Solutions
Graph Analysis – Financial Services Customer
Example #1

Background
Financial Services company hypothesized that organized groups were targeting them and desired to tackle bust-out fraud at a social level. Although some of the risk was not individually large at the account level, the company wanted to ascertain where that risk was growing socially without them realizing that those accounts were connected to each other.

Data and Analysis
Provided LexisNexis Risk with a 5 million accounts flagged with a combination of active, known fraud, charge offs and preemptively closed tags.

For all 5 million accounts, investigate every historical address, every asset and every business for the person attached to the account to find all the people they connect to; then for those people do the same to find the people that they connect to and count the number of those consumers that are also on the original customer list and are either active accounts or known fraud.

LexisNexis Approach
- Address Standardization and LexID the input.
- Join the linked input to the LexisNexis Social Graph (5 Million input to 4 Billion relationships).
- Calculate grouped social aggregates (how many of each type of flagged account is within each account’s network neighborhood).
- Score and rank order the result.
Graph Analysis – Results Overview

**Known Fraud**

<table>
<thead>
<tr>
<th># within 1 degree</th>
<th>#accounts</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4055099</td>
</tr>
<tr>
<td>1</td>
<td>39266</td>
</tr>
<tr>
<td>2</td>
<td>638</td>
</tr>
<tr>
<td>3</td>
<td>31</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

31 Accounts associated with 3 known fraud accounts (one of which could be the known fraud itself)

**Charge Off**

<table>
<thead>
<tr>
<th># within 2 degrees</th>
<th>#accounts</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4079930</td>
</tr>
<tr>
<td>1</td>
<td>14894</td>
</tr>
<tr>
<td>2</td>
<td>212</td>
</tr>
</tbody>
</table>

212 Accounts associated with 2 Charge off accounts

**Preemptive close**

<table>
<thead>
<tr>
<th># within 1 degree</th>
<th>#accounts</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3834840</td>
</tr>
<tr>
<td>1</td>
<td>239472</td>
</tr>
<tr>
<td>2</td>
<td>19026</td>
</tr>
<tr>
<td>3</td>
<td>1322</td>
</tr>
<tr>
<td>4</td>
<td>194</td>
</tr>
<tr>
<td>5</td>
<td>71</td>
</tr>
<tr>
<td>6</td>
<td>21</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>14</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
</tr>
</tbody>
</table>

3 Accounts associated with 15 Preemptive Closes (one of which could be a preemptive close itself)
Graph Analysis – Account Overview

Aggregated Group Behavior variables for a single account

- Flatten the graph
- Calculate Aggregate Group Behavior measurements
- Drive predictive analytics at a granular level using graph variables

Small Network
- Fairly tightly connected to each other
- Active account with 3 active accounts within the network
- 9 accounts within that network that are known fraud
- 1 account within the network that has been preemptively closed
Graph Analysis – Account Network Visualization
Geo-Social Analysis
Use Case #1

Background
- Consumer applies for store credit at a store 600 miles from their home.
- Consumer ships to an address not associated with their identity and it is a significant distance from their current address.
- Patient fills a prescription outside of their normal geographic location.
- Tax payer files a refund request from an address not associated with the identity.

These events tend to generate high levels of false positives.

Geo-Social Filtering
- Create a complex distance calculation from the event location to the location of the relatives and associates within 2 degrees of the consumer.
- Filter out any events that are within a threshold distance of the consumer or one of their relatives and associates.
- Powerful method of reducing false positives when calculating risk based on geographic distance of an event to the consumer.
Geo-Social Analysis

Use Case #2

Background

- Organized network commits series of crimes across the country remaining largely untraceable to traditional detection approaches.
- Syndicates using family members to steal identity information across the United States.
- Tax Fraud syndicate ships fraudulent tax refunds to addresses near their residences

Geo-Social Matching

- Create a complex distance calculation from the crime location to the location of every recent address for 270 million active identities.
- Calculate Aggregate Group Behavior variables to rank order which social groups are closely tied geo-socially to the crime locations.
- Powerful method of leveraging large data and graph analysis to tackle organized crime.
Resources

- Open Source HPCC Systems Platform: http://hpccsystems.com
- Free Online Training: http://learn.lexisnexis.com/hpcc
- SALT: http://hpccsystems.com/salt
- KEL: https://hpccsystems.com/download/free-modules/kel-lite
- Machine Learning portal: http://hpccsystems.com/ml
- The HPCC Systems blog: http://hpccsystems.com/blog
- Community Forums: http://hpccsystems.com/bb
- Source Code: https://github.com/bcc-systems
- WIKI: https://wiki.hpccsystems.com