The Value of Large Scale Entity Analysis for National Security

Mrs. Mary Galvin, LexisNexis Special Services
Dr. Flavio Villanustre, LexisNexis Risk Solutions
March 27th, 2013
LexisNexis

- Over 35 years of really large data experience
- Information solutions to enterprise customers
- More than $6 billion in revenue (2010)
- Acquired by Reed Elsevier in 1994
- RE Stock Symbols: valued at $12 billion in 2010; [NYSE: ENL; NYSE: RUK]; 34,000 employees
Government Customers

- **All 50** U.S. states
- **70%** of local governments
- **80%** of U.S. Federal agencies
LexisNexis Helps Transform Data Into Insights

**Health and Human Services**
Our program integrity solutions assist agencies in improving their overall operational processes and proactively combating fraud, waste and abuse across the continuum.

**Law Enforcement**
Our investigative solutions assist law enforcement agencies with keeping their officers and community safe by providing investigative insight built on accurate data, unique sources and collaboration tools.

**Tax and Revenue**
Our revenue discovery and recovery solutions assist tax and revenue agencies in enhancing collections processes and proactively detecting and preventing fraud to help uncover and recover new revenue streams.

**Legal**
Our legal solutions assist agencies with locating and vetting witnesses, assess asset and associations of parties in legal actions, and performing diligence on perspective clients.
1. Rapid Prototyping Projects
   • The Open Source Nature of the HPCC Combined with its Immediate Availability on AWS Enable **Big Data Analysis to Keep Pace with Mission**

2. Fraud Detection Pertaining to:
   • Counterfeit Parts/Supply Chain
   • Visas & Passports

3. Identity Resolution & Graph Analysis
Precise entity extraction and resolution are normally critical to the success of many missions. Usually involves:
- Linking all attributes for an entity together
- Identifying relationships between entities to create a complete social graph
- Categorizing the strength of the relationships
- Having definable levels of precision in this process
- Handling fuzziness (phonetics, string distance, etc.)
Challenges

• Correctly identifying entities in structured and unstructured data sources, can be challenging as:
  • There are normally no unique identifiers to be leveraged
  • Content can be dirty, incomplete and have errors
  • Heuristic rules based systems require high maintenance and are not portable across problem domains, geographies and languages
  • Complete pre-existing dictionaries aren’t available
  • Handling the resulting massive graphs creates other problems, around partitioning and distribution
The LexisNexis Open Source Platform: HPCC Systems

High Performance Computing Cluster (HPCC)

Unstructured Semi-structured Big Data

Big Data

Extraction Transformation Loading

THOR Cluster (Data Refinery)

Concurrent Realtime Delivery

ROXIE Cluster (Data Delivery)

ESP

ECL

ECL Developer Using ECL IDE

Query Results
Probabilistic Record Linkage

• Identifying and linking all records referring to the same entity across all data sources
• Also known as “specificity based linking”
• It usually involves:
  • Building specificity tables for all values present in the data
  • Value and field specificity as one of the main metrics for the linkage process
  • Iterative process to progressively identify these clusters of attributes until convergence
  • Use of phonetics, string distance and other metrics to allow for fuzzy matching
How It Works

• What is the probability that two records with identical first and last name, and city are referring to the same entity?
  • For the “John Smith, NYC” case
  • For the “Flavio Villanustre, Atlanta” case
• What if there are typos in the data? How do these probabilities change?
• How do we build relationships across entities?
  • Cohabitation
  • Shared ownership of assets
  • Transactional information
  • Correlation
Scalable Automated Linking Technology (SALT)

- The acronym stands for “Scalable Automated Linking Technology”
- Template based ECL code generator
- Provides for automated data profiling, QA/QC, parsing, cleansing, normalization and standardization
- Sophisticated specificity and relatives based linking and clustering
Beyond SALT: Smart View

- 0 Lines of Code!
- 42 Lines of SALT
- 3,980 Lines of ECL
- 482,410 Lines of C++
Questions

Email:  Mary.Galvin@LexisNexis.com
       Flavio.Villanustre@LexisNexis.com

Website:  http://hpccsystems.com
Links for More Information

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