Reinventing Business Intelligence through Big Data

Dr. Flavio Villanustre

VP, Technology and lead of the Open Source HPCC Systems initiative
LexisNexis Risk Solutions
Reed Elsevier
The term BI is coined in 1958.

1960’s
Beginnings as Decision Support Systems

1980’s
DSS, Data Warehouses, Executive Information Systems, OLAP

2000’s
Big Data brings new challenges (integration, cleansing, semantic analysis, new analytics)

2010’s
BI is evolving once again

A brief history of business intelligence

“From Analytics to Action” Symposium – March 26th, 2014
LexisNexis Risk Solutions: About us

- Part of Reed Elsevier: worldwide digital information company
- Provider of data and analytics-based solutions
- Helps clients across all industries assess, predict and manage risk
- Leading positions in insurance, financial, legal and collections
- Products delivered using state of the art technology and analytics
- Designed and developed its own Big Data analytical platform (the HPCC Systems Big Data Platform) over the past 15 years
- Released the HPCC Systems Big Data Platform as an Open Source initiative in 2011
- Continues to innovate through higher level abstractions, in areas of data integration and linkage, graph analytics, machine learning, data exploration and data visualization
The modern BI

Customer data scattered among different business units
- Products: airlines, hotels, bank, insurance
- Channels: captive, independent or cross-sale agents, call center, internet
- Functions: marketing, sales, pricing, services

Integrating customer data based on Big Data technologies
Customer Identity Resolution
Customer Relationship Identification

Enrich customer profiles with external data
- Consumption
- Wealth
- Credit
- Social media

• Cross-sales
• Customer 360 view
  o Current customers from other Divisions
  o Past customers
  o Known fraudsters
  o Black-listed persons
  o Family members of existing customers
  o Customers of discounted affinity groups

• Fraud detection
• Total customer value
• Influencers
• Disrupters
• Attrition models
• Up-sale predictors
• Pricing estimators
Data Warehousing:
- Sources are no longer just internal
- Data is dirty, incomplete, untrustworthy and structure doesn’t reflect semantics anymore
- Data volumes grow exponentially
- Real time analytics are still required for operational BI
- Exploration drives new applications

Data Analysis
- Feature extraction in unstructured data requires novel treatments which may not be easy to generalize
- Moving the data from the warehouse to the analytical system is no longer an option: in-place analytics

Graph Analytics:
- Certain problems are better represented as graphs, but graph engines can quickly show scalability limitations

Big Data, big challenges
Must leverage Big Data and offer in-place analytics
Must be able to manage unstructured data (NLP, for example)
Must support advanced data mining and analytics
Must offer near real time analysis
Must be scalable and flexible

Based on the TDWI checklist report: THE MODERN DATA WAREHOUSE
Data profiling and exploratory data analysis (EDA) are required to assess the semantics and value.

Quality controls, cleansing, parsing and normalization/standardization are key.

Rules based data integration show significant limitations: probabilistic record linkage is becoming the standard.

It is no longer just about Big Data; it’s about Linked Data.

Linking goes beyond connecting attributes with entities: it must also identify links among entities.

Iterative exploration is paramount.
Linked data from Big Data

Data Sources

- Profiling
- Parsing
- Cleansing
- Normalization
- Standardization

Data Preparation Processes (ETL)

- Matching Weights and Threshold Computation
- Blocking/Searching
- Weight Assignment and Record Comparison
- Record Match Decision

Linked Data File

Record Linkage Processes

- Additional Data Ingest

Linking Iterations
### Probabilistic record linkage

**Others**
- **Rules-based matching:**
  - Based on logic (IF/ELSE or SWITCH statements)
  - Example: If field values 1, 2 and 5 from source ‘a’ are equivalent to values 3, 6 and 7 in source ‘b’, respectively, then declare a match.

**SALT (Scalable Automated Linking Technology)**
- **Probability-based matching:**
  - Based on computation of weights and thresholds; a match is declared only when the sum of all weights surpasses a certain threshold
  - Example:

```
<table>
<thead>
<tr>
<th>Source A</th>
<th>Source B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
</tr>
</tbody>
</table>
```

**Threshold = 29**

- **Sum of Individual Field Specificity Values**

```
<table>
<thead>
<tr>
<th>Source A</th>
<th>Source B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
</tr>
</tbody>
</table>
```

**Threshold = 29**

- **Sum of Individual Field Specificity Values**
Advanced analytics and machine learning

- OLAP alone is not sufficient anymore
- Machine learning is becoming prevalent
  - Exploratory data analysis
  - Correlation analysis
  - Unsupervised and semi-supervised learning (clustering, distributions)
  - Supervised learning (regression and classification models)
- Feature extraction can be challenging (think binary data)
  - Naïve approaches to feature extraction lend themselves to high dimensional data: Manifold hypothesis
  - Automated, generalizable and hierarchical feature extraction through deep learning
- NLP is integral part of a number of applications (sentiment analysis, entity/relationship extraction, etc.)
Many real world business problems can be modeled as graphs

Influence/disruption graphs are important to BI: 360 customer view applications, business risk management, etc.

Social Network Analytics is useful for fraud detection

Social graphs can help with entity disambiguation (using the topology of the local graph as another attribute)

Recommendation systems leverage multiple intertwined graphs (web traversal, customer/object, customer/customer and object/object relationships, etc.)
A social graph example

Relationships are from public records (non-obvious in the healthcare data domain)

MIKE JONES MD
- Is the prescribing doctor who prescribed Vicodin to patients in the target social cluster (James Anderson)
- He is a member of the same social cluster
- Also personally filled a vicodin prescription for himself.

Question: Is it normal for you and 15 of your associates to all receive a prescription for vicodin within the same short timespan?
When referring to graphs, you may think about triple stores and just-in-time traversal and analytics.
These triple stores tend to require linear and continuous memory spaces.
But graphs can be also expressed as adjacency matrices.
Adjacency matrices can be pre-processed to speed up query time: turn $O(n^2)$ or $O(n^3)$ into $O(1)$ or $O(\log n)$.
And similar operations can be done through more traditional data centric approaches (data join operations, for example).

\[ A^2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \]

Exponentiation of an adjacency matrix returning a new matrix with all paths of that degree.
Putting it all together at LexisNexis

- Linked data at the core on HPCC
- Thor handles data profiling, parsing, cleansing, standardization and linkage, while Roxie handles real-time linking, analytics and delivery
- Data integration and entity linkage handled by SALT
- Regression, classification, clustering using ECL-ML
- Graph analytics with KEL (Knowledge Engineering Language)
- Visualization layer for interactive exploration and reporting
BI is undergoing a significant change, fuelled by Big Data

Horizontal scalability and in-place analytics are increasingly important, particularly for iterative exploration

New disciplines are required to make sense of the data:
  • More sophisticated data integration
  • Machine Learning
  • Graph theory

OLAP is not going away but proactive analytics are key

While the challenge can be significant, the possibilities are endless
The Open Source HPCC Systems platform: [http://hpccsystems.com](http://hpccsystems.com)


Machine Learning portal: [http://hpccsystems.com/ml](http://hpccsystems.com/ml)

The HPCC Systems blog: [http://hpccsystems.com/blog](http://hpccsystems.com/blog)

Community Forums: [http://hpccsystems.com/bb](http://hpccsystems.com/bb)

Our GitHub portal: [https://github.com/hpcc-systems](https://github.com/hpcc-systems)