Data Quality and Data Analytics

Data Summit
May 2017
<table>
<thead>
<tr>
<th>Bright House Networks</th>
<th>Verizon Communications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bright House</td>
<td>Verizon Florida Inc.</td>
</tr>
<tr>
<td>Bright House Network</td>
<td>Verizon Florida Inc.</td>
</tr>
<tr>
<td>Bright House Networks</td>
<td>Verizon Florida, Inc.</td>
</tr>
<tr>
<td>Bright House</td>
<td>Verizon North</td>
</tr>
<tr>
<td>Bright House</td>
<td>Verizon Residential</td>
</tr>
<tr>
<td>Florida Light &amp; Power</td>
<td>Verizon South</td>
</tr>
<tr>
<td>Florida Power</td>
<td>Verizon Wireless</td>
</tr>
<tr>
<td>Florida Power &amp; Light</td>
<td>Verizonwireless</td>
</tr>
<tr>
<td>Florida Power &amp; Light</td>
<td>Verizon</td>
</tr>
<tr>
<td>Florida Power &amp; Light Company</td>
<td>Verizon (formerly GTE)</td>
</tr>
<tr>
<td>Florida Power and Light</td>
<td>Verizon Communications</td>
</tr>
<tr>
<td>Florida Power and Light</td>
<td>Verizon Florida Inc.</td>
</tr>
<tr>
<td>Florida Power and Light</td>
<td>Verizon Florida, Inc.</td>
</tr>
<tr>
<td>Florida Power and Light</td>
<td>Verizon North</td>
</tr>
<tr>
<td>Florida Power and Light</td>
<td>Verizon Residential</td>
</tr>
<tr>
<td>Florida Power and Light</td>
<td>Verizon South</td>
</tr>
<tr>
<td>Florida Power and Light</td>
<td>Verizon Wireless</td>
</tr>
<tr>
<td>Florida Power and Light</td>
<td>Verizonwireless</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>American Express</th>
<th>American Express Trs Co Inc</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Express</td>
<td>American Express Trs Co Inc</td>
</tr>
<tr>
<td>American Expressss</td>
<td>American Expressss</td>
</tr>
</tbody>
</table>
About Jeff Fried

Longtime Search Nerd
- CTO, BA Insight
- Senior PM, Microsoft
- VP, FAST
- SVP, LingoMotors

Passionate About
- Search
- SharePoint
- Search-driven applications
- Information Strategy

Blog:
BAinsight.com/blog

Technet Column
“A View from the Crawlspace”
@jefffried
jeff.fried@bainsight.com

Focused on Search and SharePoint since 2004
About BA Insight

- Our software quickly connects SharePoint users to needed information
- Software portfolio includes:
  - **Connectivity** - Secure connectors and federation to over 50 content systems
  - **Applications** - Improve the user experience to boost productivity
  - **Classification** - Autotagging, metadata generation, and text analytics
  - **Analytics** - Insight into how SharePoint portals, search, & content are being used
- Hundreds of successful implementations including:
This session

Data Quality

Data Quality Solutions

How Data Analytics Improves Data Quality

Exploration and People in the Loop
Conventional Definition of Data Quality

- **Accuracy**
  - The data was recorded correctly.

- **Completeness**
  - All relevant data was recorded.

- **Uniqueness**
  - Entities are recorded once.

- **Timeliness**
  - The data is kept up to date.
    - Special problems in federated data: time consistency.

- **Consistency**
  - The data agrees with itself.
Identify & assess the degree to which poor data quality impedes business objectives

Define business-related data quality rules, perform measurements, set performance targets

Inspect, monitor, and remediate when quality of data is not acceptable

Design quality improvement processes that remediate process flaws

Move data quality improvement methods and processes into production
Source: Experian Global Data Benchmark Report 2017
Planned data management projects over the next 12 months

- Data cleansing: 33%
- Data integration: 31%
- Data migration: 28%
- Data preparation: 26%
- Data enrichment: 26%
- Data analytics: 23%
- Holistic data quality: 22%
- Master data management: 22%
- Single customer view: 22%
- Data governance: 19%
- We have no data management projects planned: 5%

Source: Experian Global Data Benchmark Report 2017
Requirements for Data Quality Solutions

**Monitoring**
Tracking and monitoring the state of Quality activities and Quality of Data

**Cleansing**
Amend, remove or enrich data that is incorrect or incomplete. This includes correction, standardization and enrichment.

**Profiling**
Analysis of the data source to provide insight into the quality of the data and help to identify data quality issues.

**Matching**
Identifying, linking or merging related entries within or across sets of data.
N-Grams for name/address resolution

Improves Data Quality through de-duplication
How Data Analytics Improves Data Quality
“Hot Research Topic” Examples

Entity Resolution ↔ MDM

“Self-Healing” Customer Data Management
Example: Major Communications Company

- Address data is critical
  - Used by 240 applications
  - Many point solutions for address matching

- Human Error as primary issue

- Existing solution had very basic matching / cleansing functionality
Common Issues

- Suite 501 L 5 shell house 170 north tc adl
  - Issues: No separating commas; no state (SA), no postcode, adl abbreviated
  - Correct Result: Suite 501, Level 5, Shell House, 170 North Terrace, Adelaide, SA

- 6 Prinya place Burlee waters QLD
  - Issues: Prinya and Burlee are misspelt
  - Correct Result: 6 Prinia Place, Burleigh Waters, QLD

- Jefferey Rd, Mt Lofti, QLD
  - Issues: road should be Geoffrey, Locality misspelt, no postcode
  - Correct Result: Geoffrey Rd, Mt Lofty, 4000, QLD

- 64- Industrial Av, Wilsonton, 4340, QLD
  - Issues: 64 is in the from field instead of the to field, post code is incorrect
  - Correct Result: -64 Industrial Av, Wilsonton, 4350, QLD
Resulting Solution

- National database for confirmed addresses
- 14 million address records, 600 thousand streets
- Used for emergency services and investigative agencies
- Many address types and relationships
- Domain-specific business rules and search logic
- Client format translation (input and output)
- Structured and unstructured input
- Confirmed address write-back
Example: Leading Insurance Vendor

Challenge:
- Client information is in several legacy systems with no standard definition for clients across the systems
- Client data is erroneous and duplicated

Solution:
- A consolidated, single view of an individual client across multiple systems
- Automated process that can match clients based on relevancy, statistics and business defined matching rules.
- Capability of learning from incorrectly matched entries that a user corrects
- Batch processing of data matching, merging and a results and exception report
- Drives changes to legacy systems and master file
Data match and merge based on rules to create golden records
Data Quality Dashboard

Customers by Business Value

DQ Exceptions

Address Verification Status

Profile Tags

Courtesy of Reltio
Exploration and People in the Loop
Modes

- Data Input
- Duplicate & Dirt Detection
- Outlier Detection
- Exploratory Discovery
- Process Improvement
Example: Feedback Loops

- Customer
- Sales Order
- Billing
- Customer Account Information
- Provisioning
- Customer Care

Existing Data Flow
Missing Data Flow
If it’s Big Data, it’s machine generated

So...I don’t need to worry about data quality, right?

Chart courtesy of xkcd
This session

Data Quality

Data Quality Solutions

How Data Analytics Improves Data Quality

Exploration and People in the Loop
www.BAinsight.com

Jeff.Fried@BAinsight.com

@jefffried

“Now! That should clear up a few things around here!”