From Cost Center to Profit Center – Data Management Best Approaches
July 13, 2016
SPENDING ALL DAY WRANGLING DATA?

Reduce Data Prep Time & Drive Better Analytics
Overview of RedPoint Global

Launched in 2006

Founded and staffed by industry veterans

Headquarters: Wellesley, Massachusetts

Offices in US, UK, Australia, Philippines

Global customer base

Serves most major industries
RedPoint Data Management Ranks High in Gartner Critical Capabilities Report

### Product or Service Scores for Operational/Transaction Data Quality

<table>
<thead>
<tr>
<th>RedPoint</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIOsoft</td>
<td></td>
<td></td>
<td></td>
<td>4.19</td>
<td></td>
</tr>
<tr>
<td>IBM</td>
<td></td>
<td></td>
<td></td>
<td>4.14</td>
<td></td>
</tr>
<tr>
<td>Informatica</td>
<td></td>
<td></td>
<td></td>
<td>4.04</td>
<td></td>
</tr>
<tr>
<td>Information Builders</td>
<td></td>
<td></td>
<td></td>
<td>3.92</td>
<td></td>
</tr>
<tr>
<td>SAS</td>
<td></td>
<td></td>
<td></td>
<td>3.84</td>
<td></td>
</tr>
<tr>
<td>Trillium</td>
<td></td>
<td></td>
<td></td>
<td>3.84</td>
<td></td>
</tr>
<tr>
<td>Ataccama</td>
<td></td>
<td></td>
<td></td>
<td>3.79</td>
<td></td>
</tr>
<tr>
<td>Uniserv</td>
<td></td>
<td></td>
<td></td>
<td>3.77</td>
<td></td>
</tr>
<tr>
<td>DataMentors</td>
<td></td>
<td></td>
<td></td>
<td>3.70</td>
<td></td>
</tr>
<tr>
<td>SAP</td>
<td></td>
<td></td>
<td></td>
<td>3.68</td>
<td></td>
</tr>
<tr>
<td>Experian</td>
<td></td>
<td></td>
<td></td>
<td>3.67</td>
<td></td>
</tr>
<tr>
<td>Neopost</td>
<td></td>
<td></td>
<td></td>
<td>3.64</td>
<td></td>
</tr>
<tr>
<td>Innovative</td>
<td></td>
<td></td>
<td></td>
<td>3.60</td>
<td></td>
</tr>
<tr>
<td>Talend</td>
<td></td>
<td></td>
<td></td>
<td>3.58</td>
<td></td>
</tr>
<tr>
<td>BackOffice</td>
<td></td>
<td></td>
<td></td>
<td>3.55</td>
<td></td>
</tr>
<tr>
<td>Pitney Bowes</td>
<td></td>
<td></td>
<td></td>
<td>3.52</td>
<td></td>
</tr>
<tr>
<td>Oracle</td>
<td></td>
<td></td>
<td></td>
<td>3.49</td>
<td></td>
</tr>
</tbody>
</table>

### Product or Service Scores for Data Integration

<table>
<thead>
<tr>
<th>RedPoint</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIOsoft</td>
<td></td>
<td></td>
<td></td>
<td>4.37</td>
<td></td>
</tr>
<tr>
<td>IBM</td>
<td></td>
<td></td>
<td></td>
<td>4.21</td>
<td></td>
</tr>
<tr>
<td>Informatica</td>
<td></td>
<td></td>
<td></td>
<td>4.13</td>
<td></td>
</tr>
<tr>
<td>Information Builders</td>
<td></td>
<td></td>
<td></td>
<td>4.10</td>
<td></td>
</tr>
<tr>
<td>DataMentors</td>
<td></td>
<td></td>
<td></td>
<td>4.02</td>
<td></td>
</tr>
<tr>
<td>SAS</td>
<td></td>
<td></td>
<td></td>
<td>3.99</td>
<td></td>
</tr>
<tr>
<td>Trillium</td>
<td></td>
<td></td>
<td></td>
<td>3.99</td>
<td></td>
</tr>
<tr>
<td>Uniserv</td>
<td></td>
<td></td>
<td></td>
<td>3.97</td>
<td></td>
</tr>
<tr>
<td>Neopost</td>
<td></td>
<td></td>
<td></td>
<td>3.95</td>
<td></td>
</tr>
<tr>
<td>Ataccama</td>
<td></td>
<td></td>
<td></td>
<td>3.94</td>
<td></td>
</tr>
<tr>
<td>Experian</td>
<td></td>
<td></td>
<td></td>
<td>3.91</td>
<td></td>
</tr>
<tr>
<td>Innovative</td>
<td></td>
<td></td>
<td></td>
<td>3.86</td>
<td></td>
</tr>
<tr>
<td>SAP</td>
<td></td>
<td></td>
<td></td>
<td>3.80</td>
<td></td>
</tr>
<tr>
<td>Pitney Bowes</td>
<td></td>
<td></td>
<td></td>
<td>3.75</td>
<td></td>
</tr>
<tr>
<td>Talend</td>
<td></td>
<td></td>
<td></td>
<td>3.73</td>
<td></td>
</tr>
<tr>
<td>Oracle</td>
<td></td>
<td></td>
<td></td>
<td>3.64</td>
<td></td>
</tr>
<tr>
<td>BackOffice</td>
<td></td>
<td></td>
<td></td>
<td>3.62</td>
<td></td>
</tr>
</tbody>
</table>
Big Data Can Become Big Information
What Needs to Happen?

What question needs to be answered?

Structures defined to support entities

Assembly of relevant bits of data

Potential business value that can be realized through execution
## Attributes of Information

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RELEVANT</strong></td>
<td>Information must pertain to a specific problem. General data must be connected to reveal relevance of the information.</td>
</tr>
<tr>
<td><strong>COMPLETE</strong></td>
<td>Partial information is often worse than no information. Partial information frequently leads to worse conclusions than if no data had been used at all.</td>
</tr>
<tr>
<td><strong>ACCURATE</strong></td>
<td>This one is obvious. In a context like health care, inaccurate data can be fatal. Precision is required across all applications of information.</td>
</tr>
<tr>
<td><strong>CURRENT</strong></td>
<td>As data ages, it becomes less accurate. Multiple research studies by Google and others show the decay in the accuracy of analytics as data becomes stale.</td>
</tr>
<tr>
<td><strong>ECONOMICAL</strong></td>
<td>There has to be a clear cost benefit. This requires work to identify the realizable benefit of information but this is also what drives the use if successful</td>
</tr>
</tbody>
</table>
Current State of Data Preparation

Time spent preparing data: 80%

DATA PREP

- Denormalizing analytic data
- Recoding attributes
- Data exploration, univariate analysis, data profiling
- Normalizing value

IN PRODUCTION MODE

Time spent tuning algorithm: 20%
When Data Prep is Reduced

Time spent preparing data: 20%

Latency between data extraction and application is inversely correlated to the lift.

No coding increases productivity!

Fast modeling means greater agility and precision

Time spent tuning algorithm: 80%
The Elephant in the Room

Skills Gap

- Severe shortage of MR or Spark skilled resources
- Very expensive resources and hard to retain
- Inconsistent skills lead to inconsistent results
- Under utilizes existing resources
- Prevents broad leverage of investments across enterprise

Maturity & Governance

- A nascent technology ecosystem around Hadoop
- Emerging technologies only address narrow slivers of functionality
- New applications are not enterprise class
- Legacy applications have built short term capabilities

Data Into Information

- Data is not useful in its raw state, it must be turned into information
- Benefit of Hadoop is that same data can be used from many perspectives
- Analysts must now do the structuring of the data based on intended use of the data
Key Data Mastering Functionality Needed for Fast Data Prep

ETL & ELT
- Profiling, reads/writes, transformations
- Single project for all jobs

Data Quality
- Cleanse data
- Parsing, correction
- Geo-spatial analysis

Integration & Matching
- Grouping
- Fuzzy match

Master Key Management
- Create keys
- Track changes
- Maintain matches over time

Web Services Integration
- Consume and publish
- HTTP/HTTPS protocols
- XML/JSON/SOAP formats

Process Automation & Operations
- Job scheduling, monitoring, notifications
- Central point of control
- Meta Data Management

Hadoop Integration
- Pure YARN integration into Hadoop
- No coding data quality

Java SDK Layer
- Java SDK for rapid development
- Public project incubator for project sharing
Benchmarks – Project Gutenberg

**MapReduce**

- public static class MapClass extends Mapper<WordOffset, Text, Text, IntWritable> {
  - private final static String delimiters = "",./<>?;\-_+()&*%^#$!@`~|«»¡¢£€¥¦®¯±¶·¿";
  - private final static IntWritable one = new IntWritable(1);
  - private Text word = new Text();
  - public void map(WordOffset key, Text value, Context context) throws IOException, InterruptedException {
  - String line = value.toString();
  - StringTokenizer itr = new StringTokenizer(line, delimiters);
  - while (itr.hasMoreTokens()) {
  - word.set(itr.nextToken());
  - context.write(word, one);
  - }
  - }
- }

- >150 Lines of MR code
- 6 hours of development
- 6 minutes runtime
- Needs extensive optimization

**Pig**

- SET pig.maxCombinedSplitSize 67
- SET pig.splitCombination true
- A = LOAD '"testdata/pg/*/*/*';
- B = FOREACH A GENERATE FLATTEN(TOKENIZE((chararray)$0)) AS word;
- C = FOREACH B GENERATE UPPER(word) AS word;
- D = GROUP C BY word;
- E = FOREACH D GENERATE COUNT(C) AS occurrences, group;
- F = ORDER E BY occurrences DESC;
- STORE F INTO '/user/cleonardi/pg-count';

- ~50 Lines of script code
- 3 hours of development
- 15 minutes runtime
- User-defined functions needed before running script

**RedPoint**

- 0 Lines of code
- 15 minutes of development
- 3 minutes runtime
- No tuning or optimization required
Intel’s POV on Data Quality Outside Hadoop

Data Cleansing, Data Lakes, and BI

Inputs and Sources
Data is collected from various sources

Data Processing
Data is ingested, cleansed, stored, and conformed for use

Self-Service Data Access
Data is available in different timeframes depending on user’s needs

- Conventional wisdom moves data out of the cluster for Data Quality and MDM
- Costly process in time and money
- Does not take advantage of scalable cluster processing
RedPoint’s Marketing Data Lake

Data Sources
- CRM
- ERP
- Billing
- Subscriber
- Product
- Network
- Weather
- Compete
- Manuf.
- Clickstream
- Online Chat
- Sensor Data
- Social Media
- Call Detail Records
- Fabrication Logs
- Sales Feedback
- Field Feedback

Data Ingestion
- YARN
- In cluster Data Quality Processing,
- In cluster Persistent Matching,
- In cluster MDM
- Process native document or tabular data

Data Lake
- Persistent Entity Resolution, Linkage and Keying

Specialized Analytic Databases & Caches
- Teradata
- IBM
- Redshift
- Oracle
- MySQL

Production RDBMS Databases
- Oracle
- IBM
- Teradata

Planning
Analytics
Interactions
Real-Time

© RedPoint Global Inc. 2016
Visit RedPoint at www.redpoint.net where you can:

- Access the latest Gartner Critical Capabilities for Data Quality report
- View Customer Case studies
- Request a Free Trial

Todd Hinton
Vice President of Product Strategy
RedPoint Global

todd.hinton@redpoint.net
@toddondata