Speed of thought interactive analytics and other real life Use Cases

Richard Langlois
President IT Architecture & Strategy

September 2016
About the Speaker – Richard Langlois

Richard is the president of IT Architecture & Strategy which provides training and consulting services in Big Data, Analytics, BI, Enterprise Architecture and Data Governance.

Recently, Richard was the director of search and big data analytics / director of enterprise data management for Yellow Pages (Canada), where his team provided development of solutions, data architecture and governance, and metadata management for all operational and analytics needs of Yellow Pages.

Prior to his roles at Yellow Pages, Richard was enterprise architect adviser at National Bank and Desjardins Group, global chief architect at TataCommunications and led the Canadian BI practice at Capgemini.

He also worked directly or thru consulting mandates at: Air Canada, Bell Canada, CN, Canadian Tire, GM, Hydro-Quebec, Investors Group, Seer Technologies, Sikorsky Aircraft, Texas Instruments, Unisys and multiple government agencies.
Understanding Big Data and Analytics from a Business Perspective

DO WE HAVE ANY ACTIONABLE ANALYTICS FROM OUR BIG DATA IN THE CLOUD?

YES, THE DATA SHOWS THAT MY PRODUCTIVITY PLUNGES WHENEVER YOU LEARN NEW JARGON.

MAYBE IN-MEMORY COMPUTING WILL ACCELERATE YOUR APPLICATIONS.

PLUNGE, PLUNGE, PLUNGE.
## What Business Value are you Hoping to Achieve?

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Enhanced customer experience</td>
<td>51%</td>
<td>77%</td>
<td>44%</td>
<td>70%</td>
<td>66%</td>
<td>65%</td>
<td>80%</td>
<td>83%</td>
<td>89%</td>
<td>60%</td>
</tr>
<tr>
<td>More targeted marketing</td>
<td>46%</td>
<td>69%</td>
<td>13%</td>
<td>30%</td>
<td>77%</td>
<td>50%</td>
<td>25%</td>
<td>33%</td>
<td>44%</td>
<td>33%</td>
</tr>
<tr>
<td>Process efficiency</td>
<td>57%</td>
<td>38%</td>
<td>50%</td>
<td>60%</td>
<td>32%</td>
<td>35%</td>
<td>40%</td>
<td>83%</td>
<td>56%</td>
<td>73%</td>
</tr>
<tr>
<td>Cost reduction</td>
<td>54%</td>
<td>31%</td>
<td>50%</td>
<td>10%</td>
<td>32%</td>
<td>15%</td>
<td>45%</td>
<td>67%</td>
<td>56%</td>
<td>53%</td>
</tr>
<tr>
<td>New products</td>
<td>32%</td>
<td>62%</td>
<td>38%</td>
<td>10%</td>
<td>23%</td>
<td>25%</td>
<td>40%</td>
<td>-</td>
<td>33%</td>
<td>20%</td>
</tr>
<tr>
<td>Developing information products</td>
<td>41%</td>
<td>31%</td>
<td>44%</td>
<td>-</td>
<td>9%</td>
<td>25%</td>
<td>30%</td>
<td>17%</td>
<td>22%</td>
<td>27%</td>
</tr>
<tr>
<td>Improved risk management</td>
<td>27%</td>
<td>15%</td>
<td>31%</td>
<td>10%</td>
<td>14%</td>
<td>70%</td>
<td>55%</td>
<td>67%</td>
<td>22%</td>
<td>33%</td>
</tr>
<tr>
<td>Regulatory compliance</td>
<td>19%</td>
<td>15%</td>
<td>13%</td>
<td>10%</td>
<td>9%</td>
<td>50%</td>
<td>20%</td>
<td>67%</td>
<td>44%</td>
<td>27%</td>
</tr>
<tr>
<td>Enhanced security capabilities</td>
<td>5%</td>
<td>-</td>
<td>19%</td>
<td>30%</td>
<td>14%</td>
<td>35%</td>
<td>5%</td>
<td>17%</td>
<td>11%</td>
<td>13%</td>
</tr>
<tr>
<td>Other</td>
<td>3%</td>
<td>8%</td>
<td>-</td>
<td>10%</td>
<td>-</td>
<td>5%</td>
<td>-</td>
<td>17%</td>
<td>-</td>
<td>13%</td>
</tr>
<tr>
<td>Not yet defined</td>
<td>5%</td>
<td>-</td>
<td>6%</td>
<td>20%</td>
<td>9%</td>
<td>-</td>
<td>5%</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Multiple Responses Allowed

[Strata+Hadoop World](http://www.strata+hadoopworld.com)
What Is Analytics – The 4 types of Analytics?

- **Use Descriptive analytics** to understand historical performance, alert you to events, spot trends, and monitor the performance of models and actions
  - Scorecards
  - Dashboards
  - Reports

- **Use Diagnostic analytics** to visualize and interact with the data, discover relationships, and explain why you are observing outcomes, events or trends
  - Query & Analysis (OLAP)
  - Data Discovery

- **Use Predictive analytics** to answer questions about the future or to determine the likelihood of unknown outcomes
  - Statistics
  - R and others

- **Use Prescriptive analytics** to optimize decisions, efficiently allocate resources or find the best customer treatments
  - Optimization
  - What-if
  - Statistics
  - Scoring Systems
Understanding the Spectrum of Analytics and Big Data

- Broader Usage
  - Ubiquitous
  - Self-Service
  - Shared

- Greater Empowerment
  - Corporate
  - Social / 3rd Party
  - Structured

- Broader Distribution
  - Un- & Semi-Structured
  - Silo

- Wider Input
  - Real-Time
  - Batch

- More Data
  - Embedded
  - Process
  - Integration

- Analytic Location
  - On-Premise
  - Cloud

- More Detail
  - Granular

- More Relevance
  - Decisions

- Greater Sophistication
  - Prescriptive

- Lower Latency
Challenges facing Big Data & Analytics Projects
Many Organizations are still in this Mode

- Highly complex
- Lots of people & Skillsets
- Multiple copies of data
- Stale data
- Rigid schema
- Tough to change

<table>
<thead>
<tr>
<th>VISUALIZATION TOOLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>QUERY ENGINE</td>
</tr>
<tr>
<td>MART</td>
</tr>
<tr>
<td>MART</td>
</tr>
<tr>
<td>MART</td>
</tr>
<tr>
<td>ETL</td>
</tr>
<tr>
<td>DATA WAREHOUSE</td>
</tr>
<tr>
<td>ETL</td>
</tr>
<tr>
<td>DATA SOURCES</td>
</tr>
</tbody>
</table>
Big Data and Analytics Challenges and Priorities

Respondents indicate establishing a roadmap and creating a culture of data-based decision making are their top challenges for 2016.

Note: The values in the graph do not add up to 100%, because it was a “select top three” question.
Top Concerns with Big Data Analytics Projects

- Skill set
- Management
- Security
- Performance
- Governance
- Accessibility

Source: AtScale Hadoop Maturity Survey (Oct 2015)  
In Partnership with Cloudera, Hortonworks, MapR & Tableau
Data Governance

How do you govern Data given:

- Popularity of Data Lakes
- Co-existence of DW and Data Lakes
- Data in the cloud (or in premise) or both
- Bi-Modal organizations
- Proliferation of Analytics Solutions
- …
About Yellow Pages (Canada)
About Yellow Pages (YP) - Canada

Yellow Pages’ Mission
We aim to champion the digital economy by offering consumers and merchants media and marketing solutions that help them interact and transact.

- Yellow Pages is a leading digital media and marketing solutions company in Canada.
- Serving approximately 244,000 (Q2, 2016) local businesses through our nationwide sales force of media consultants.
# The Building Blocks of Relevance

<table>
<thead>
<tr>
<th>What is <strong>Content</strong>?</th>
<th>What is <strong>User Data</strong>?</th>
<th>What is <strong>Knowledge</strong>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advertiser Content</td>
<td>User Characteristics</td>
<td>Linguistic</td>
</tr>
<tr>
<td>(merchants, products, deals …)</td>
<td>(demographics, preferences, …)</td>
<td>(synonyms, acronyms, multilingual,…)</td>
</tr>
<tr>
<td>Editorial Content</td>
<td>User Context</td>
<td>Geo Data</td>
</tr>
<tr>
<td>(smart tips, top ten lists, …)</td>
<td>(location, time, device, …)</td>
<td>(location names, polygons, …)</td>
</tr>
<tr>
<td>User-Generated Content</td>
<td>User Behavior</td>
<td>Role:</td>
</tr>
<tr>
<td>(reviews, ratings …)</td>
<td>(origin, session actions, …)</td>
<td>Provide the data that the user is looking for</td>
</tr>
</tbody>
</table>

- **Role:**
  - Advertisers: Looking for Qualified Consumers
  - Consumers: Looking for the “Best Merchant & Product”

YP’s Business Model is a Multi-Sided Platform
# The Building Blocks of Relevance

<table>
<thead>
<tr>
<th>What is Search?</th>
<th>What is Presentation?</th>
<th>What is Optimization?</th>
</tr>
</thead>
<tbody>
<tr>
<td>An engine that: 1) interprets user queries</td>
<td>An engine that: builds the presentation layout based on the semantic value of content</td>
<td>Scenario Management, Tracking and Analytics</td>
</tr>
<tr>
<td>2) extracts the proper result set</td>
<td>Example: “Movie Theater” vs “Restaurant” Layouts</td>
<td>Provides the ability to test and measures the results for variations applied in the other building blocks</td>
</tr>
<tr>
<td>3) and rank the results</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Role:**
- **Search:** It is at the heart of the user search process
- **Presentation:** Ensures that user quickly gets the most out of the content presented and enhances conversion
- **Optimization:** Provides the “empirical / scientific” tools to optimize relevance, user engagement and conversion
Content is syndicated on other platforms

Digital revenue represented approximately 67.7% of consolidated revenues (Q2, 2016)
Big Data & Analytics Use Cases
Guiding Principles

Single Source of Truth for

• Consumers | Searches | Advertising

Complete Source of Truth

• Assets based | Rich Context | Granular | All products/channels are collected | Veracity / Governed

Efficient Data Collection

• One Time Collection | No dependency on analytics solution supplier

Data Accessibility

• Volume | Variety | Good response time

Real Time

• Appropriate Velocity
Guiding Principles

• 1 single TMS for all YP
  • TMS needs to be ‘Analytics solution neutral’
  • Changing TMS is costly (even with Standard Data Layer) due to need to retag and time it takes for mobile users to upgrade (+ 6 months)

• It is acceptable, even recommended, to keep a hybrid environment where analytics based applications coexist with the Digital Analytics Database (DAD). Benefits:
  • Allows selection of best of breed solutions
  • Speedier time to market
  • Raw events need to be captured even when using COTS Analytics Solutions

• Real-time correlation of the client-side data and the server-side data may help in creating 1 uniquely identified event
Guiding Principles | Supporting a Bi-Modal Organization

Characteristics:
- Reliable
- Robust
- Monitored
- Single version of the Truth
- Slower to Implement
- Faster to query
- IT Centric
- Plan-driven
- Internal Platform (but does not have to)

Characteristics:
- Flexibility
- Multiple Prototypes
- Faster to Implement
- Slower to query
- Business Centric
- Empirical, iterative
- Internal (the Industrial Analytic)
& Multiple External platforms
Guiding Principles | Breaking the Analytics Silos

Digital Data

Consumer Analytics
- Total Digital Population (Direct & Owned, Paid, SEO)
- Engagement (incl. Profiling, Growth Hacker)
- Interactions (leads)
- Behaviour
- Social Media & Personalisation

Advertiser Analytics
- Advertiser ROI (incl. Lead box)
- Life-cycle
- Market Potential
- Campaign Management (incl. MPN / YP Lead) / Placement Performance
- Data Monetization

Search Relevancy
- Content
- Search Experience
- Consumer Profile Based Search

Business Intelligence
- Sales
- Financial
- Corporate Performance
- Segmentation & Pricing

⚠️ It is much easier now to create silos with Cloud based solutions…
## Guiding Principles | Define a Big Data & Analytics Services Offering

<table>
<thead>
<tr>
<th>Strategy, Planning &amp; Architecture Services</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Big Data Analytics roadmaps</strong></td>
<td>In a context of going faster, new silos will be created and will need to be industrialized back in the Analytics Warehouse (AW)</td>
</tr>
<tr>
<td><strong>Data Architecture</strong></td>
<td>New data sources will require Data Architecture services to maintain integration and correlation of all digital events in AW</td>
</tr>
<tr>
<td><strong>Semantic Data Layer</strong></td>
<td>To support self-serve analysis via OLAP</td>
</tr>
<tr>
<td><strong>Data Platform Capacity Planning</strong></td>
<td>To plan growth of clusters based on planned projects and actual usage</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Development Services</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Collection</strong></td>
<td>Allowing optimized (low impacts on digital properties) collection of digital events (standardized and contextually rich)</td>
</tr>
<tr>
<td><strong>Robust ETL development</strong></td>
<td>To maintain data freshness with proper SLA</td>
</tr>
<tr>
<td><strong>Development of operational analytics app</strong></td>
<td>To build operational platforms such as YP Analytics</td>
</tr>
<tr>
<td><strong>Development of data services</strong></td>
<td>Provide data services used by applications such as Compass and YP Analytics to ensure data consistency</td>
</tr>
<tr>
<td><strong>Development of reports and dashboards</strong></td>
<td>For complex reports requiring more advance expertise and production of corporate dashboards</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Enablement &amp; Data Services</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Aggr. mgmt &amp; query rewrite capability</strong></td>
<td>To shield analysts &amp; data scientists from cumbersome data manipulation work</td>
</tr>
<tr>
<td><strong>Provide profiling/discovery/light ETL tools</strong></td>
<td>To use raw data sources (data lakes)</td>
</tr>
<tr>
<td><strong>Data sciences</strong></td>
<td>For business units without data scientists</td>
</tr>
<tr>
<td><strong>Data Quality &amp; Governance</strong></td>
<td>To ensure right level of data quality</td>
</tr>
</tbody>
</table>
1st Use Case | Yellow Pages Analytics (YPA)

- Free solution available for all Canadian local businesses
- User-friendly dashboards to make digital analytics accessible
- Usual digital metrics (visitors, visits, page views, interactions/clicks)
- KPI focused on ROI and revenue (calls, in-shop walk-ins, digital contacts)
- Market Potential tool
- Uses Market Data
- Reports covering all product assets

Won Big Data & Analytics Octas in 2015!
1st Use Case | Behind YP Analytics

BENEFITS:
- Enables a complete view of digital performance from consumer behaviour to advertiser ROI
- Reduce Data Manipulations and allow better and faster analytics
- Data can be correlated without gymnastics
- Increase Data Quality and Data Governance (1 version of the truth)
- Data easily accessible through visualization tools
- Data Science and Machine learning using same data
- Since existing Operational Analytics platform perform well, lets keep them to free up Hadoop resources

Previously: Most Data Sources were going directly to Infobright DB

Currently: All new Data Analytics sources go thru Hadoop
All Data Services optimized for Merchant / Account

**Operational Analytics**
- YP Analytics
- CAA Services
- Infobright DB

**Integration Services**
- Tibco ActiveMatrix

**Data Services**
- XML

**Data Abstraction / Metadata**
- Business Views

**Physical Storage / Advertiser Analytics DB**
- Tables
- Views
- Stored Procedures

**Applications**
- Compass (Sales Tool)
- YP Analytics
2nd Use Case | Speed of Thought Interactive Analytics

What happen when queries are not predetermined and not for specific merchant / account?

Query performance was tested in the YP’s Production Hadoop environment
- AtScale allocated at 30% of Hadoop cluster capacity on 26 Worker/Data nodes
- Queries written and executed manually through HUE
- Same queries executed via visualisation engine with no AtScale aggregates
- Same queries executed again (cleaned cache). AtScale engine hits the aggregate table

<table>
<thead>
<tr>
<th>'Virtual Cube'</th>
<th>Query Response (s)</th>
<th>Aggregate Building</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hue</td>
<td>AtScale No Aggregate (via Tableau)</td>
</tr>
<tr>
<td>RAM</td>
<td>60</td>
<td>126</td>
</tr>
<tr>
<td>RAM</td>
<td>534</td>
<td>606</td>
</tr>
<tr>
<td>Phone Leads Online &amp; Print</td>
<td>8.4</td>
<td>3.71</td>
</tr>
<tr>
<td>Phone Leads Online &amp; Print</td>
<td>7.7</td>
<td>6.8</td>
</tr>
</tbody>
</table>
2nd Use Case | Behind Speed of Thought Interactive Analytics

Industrial Analytics

YP Hadoop Data Platform

AtScale Platform

SQL/MDX dialects + ODBC, JDBC or OLE/DB protocols

Metadata/OLAP + Aggregate mgt.

Hadoop Data Platform (Cloudera)

DB definition (Hive)  SQL Engine (Impala)

AtScale Aggregates

DATA (HDFS, Parquet)

Data Ingestion / ETL Jobs

Discovery/Data Lakes

Need to explore new data sources to support new initiatives

• It should be exception based and not the norm
• The new data sources to explore should be used in conjunction with industrialized data (blend)

New Data Sources

Light ETL (data lakes)
2nd Use Case | Behind Speed of Thought Interactive Analytics

1. Tableau reports publishing
2. Scheduled Tableau Data Extracts
3. Direct data consumption
4. Data Ingestion / ETL
5. Impala Queries

YP Hadoop Data Platform
- AtScale Platform
  - SQL/MDX dialects + ODBC, JDBC or OLE/DB protocols
  - Metadata/OLAP + Aggregate mgt.
- Hadoop Data Platform (Cloudera)
  - DB definition (Hive)
  - SQL Engine (Impala)
  - DATA (HDFS, Parquet)

Tableau Creators
- Tableau Desktop

Tableau Online
- Tableau Server

Tableau Consumers
- Google BigQuery
- CRM (cloud)
- Google Analytics

 Meditative data sources

YP data sources
Questions?