Big Data – Is Not “Yet Another IT Project”

Krish Krishnan
President, Sixth Sense Advisors Inc
“Bridge to Big Data”
Oct 23rd 2012
Background

- Applications, OLTP Systems, Traditional Data Warehouse and Business Intelligence solutions have been driven, designed and managed by IT teams. While the existing solutions are providing some business value, they are not being considered as groundbreaking.

- The nuance of IT built and commissioned systems bode well in traditional data management worlds but not with Big Data.
  - What are the differences?
  - What is the critical success factor for a Big Data Project?
  - What is the role of IT in the future?
Tipping Points

- When the *Unexpected* becomes *Expected*
- Viral change inflected by a boiling point or a dramatic moment of critical mass

- Internet 2.0
  - Long Tail
- Social Media
  - Collaboration
  - Sharing
- Mobility
  - SmartPhone / iPhone
  - WiFi / 3G / 4G
  - Devices
  - Cloud
A Growing Trend

Expectations for BI are changing w/o anyone telling us

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Expectations</th>
<th>Reality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed</td>
<td>Speed of the Internet</td>
<td>Speed = Infra + Arch + Design</td>
</tr>
<tr>
<td>Accessibility</td>
<td>Accessibility of a Smartphone</td>
<td>BI Tool licenses &amp; security</td>
</tr>
<tr>
<td>Usability</td>
<td>IPAD - Mobility</td>
<td>Web Enabled BI Tool</td>
</tr>
<tr>
<td>Availability</td>
<td>Google Search</td>
<td>Data &amp; Report Metadata</td>
</tr>
<tr>
<td>Delivery</td>
<td>Speed of questions</td>
<td>Methodology &amp; Signoff</td>
</tr>
<tr>
<td>Data</td>
<td>Access to everything</td>
<td>Structured Data</td>
</tr>
<tr>
<td>Scalability</td>
<td>Cloud (Amazon)</td>
<td>Existing Infrastructure</td>
</tr>
<tr>
<td>Cost</td>
<td>Cell phone or Free WIFI</td>
<td>Millions</td>
</tr>
</tbody>
</table>
Forces Shaping Business

Potential New Data
- Unstructured Data
- GeoSpatial
- RFID
- Machine Logs
- Clickstream Logs
- Image Data
- Video Data

Technology
- Hadoop
- NoSQL
- Cassandra
- Cloud Deployments
- Mobility

Business Intelligence & Analytics
- Log Parsing
- 3rd Party Data
- Micro Targeting
- Micro Sourcing

Competition

Consumer
- Voice of Customer
- Social Media
- Multi Channel CRM
- SRM
Business Evolution – Product to Customer Focus Shifts
State of Data
Shaking The Foundations

Top management is responsible for setting strategy
Getting better, getting faster is the way to win
IT creates competitive advantage
Being revolutionary is high risk
We can merge our way to competitiveness
Innovation equals new products and new technology
Strategy is the easy part, implementation the hard part
Change starts at the top
Our real problem is execution
Big companies can’t innovate

Everyone is responsible for setting strategy
Rule-busting innovation is the way to win
Unconventional business concepts create competitive advantage
More of the same is high risk
There’s no correlation between size and competitiveness
Innovation equals entirely new business concepts
Strategy is the easy only if you’re content to be an imitator
Change starts with activists
Our real problem is innovation
Big companies can become gray-haired revolutionaries

Big Data

- Complex
- Ambiguous
- Structure-less
- Minimally Organized
Noise vs Value

- Truth or Dare -
  - Is Big Data Noise or Value?
  - Is Big Data a passing phase?
  - Is Big Data just hype?
  - Is Big Data not like other data?
  - Is Big Data having real nuggets?

Image source - internet
Tweet - @jdoe – very disappointed with @united @checkin lousy svc, bad mgmt, long lines #fail

sentiment
brand
process

Actual Insights

IT Perspective – Text of 140 chars will be stored as string and source is Twitter, with datetime stamp, username.
Big Data

- IT Perspective
  - Another class and type of data
  - Needs new infrastructure that is scalable and new tools and technologies
  - Needs custom programming and development
  - High maintenance
  - Complex and Noisy

- Business Perspective
  - New data, different from anything seen or available to date
  - Lots of new patterns, insights and trends
  - Highly useful both in short cycle times and deep mining
  - Noise has lots of hidden value (relative & perceived until proven)
  - Competitive and Customer insights
Big Data Differentiators

- Is Data definitely, but
- It is not like anything you have dealt with before
- Has no finite structure
- Heavy on volume, ambiguous in metadata, loosely connected in structure
- If data is linguistic oriented, processing will be heavily contextual
- If data is numeric oriented, processing needs more granular data or has more number crunching
- If semi structured data, it needs more expansion and can have a mixed workload
Business Drives Big Data

- Big Data needs investigation and inspection prior to processing
- Big Data needs more human interface and emotional tie-in
- Big Data needs more understanding of context
  - Facebook / Twitter / Flickr
  - LinkedIn
  - Forums / Collaboration Suites
  - Textual / Video / Image Data
- Big Data needs more hypothesis prior to standardization
- Big Data needs to be treated as a non-IT project
Challenges

- Acquisition of data
- Quality of data
- Metadata
- Value of data
- Organizational readiness
- Adoption
- Visualization
- Time To Delivery
## Analysis

<table>
<thead>
<tr>
<th>Business</th>
<th>Acquire</th>
<th>Search</th>
<th>Process (Inspect / Analyze)</th>
<th>Metrics</th>
<th>Visualize</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3&lt;sup&gt;rd&lt;/sup&gt; Party Data</td>
<td>Patterns</td>
<td>Categorize</td>
<td>Score</td>
<td>Trends</td>
</tr>
<tr>
<td></td>
<td>RSS Feeds</td>
<td>Phrases</td>
<td>Filter</td>
<td>Aggregate</td>
<td>Sentiments</td>
</tr>
<tr>
<td></td>
<td>Content Platforms</td>
<td>Keys</td>
<td>Enrich</td>
<td>Domain</td>
<td>Alerts</td>
</tr>
<tr>
<td></td>
<td>Internet</td>
<td>Tag</td>
<td>Contextualize</td>
<td></td>
<td>Competitive Intel</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Geo-Tag</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Meta-Tag</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Integrate</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IT</th>
<th>Acquire</th>
<th>Search</th>
<th>Process (Inspect / Analyze)</th>
<th>Metrics</th>
<th>Visualize</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Search Platforms</td>
<td>Metadata</td>
<td>Platform Setup &amp;</td>
<td>Platform Setup &amp;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Crawlers</td>
<td>GeoSpatial Data</td>
<td>Integration</td>
<td>Workflow support</td>
<td></td>
</tr>
<tr>
<td></td>
<td>API Development</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

@copyright Sixth Sense Advisors Inc
Processing – The Details

- Tag
- Categorize
- Classify
- Cleanse (Data Quality Rules)
- Semantic Integration
- Measure
- Visualize

No IT involvement seen in this cycle
Why Does Big Data Differ

- IT does not know the data
- Business knows the **intelligence** to be applied to the data to derive **value**
- Business knows how to discover data **patterns** (manual and automated)
- Business understands the **semantics** better
- Business can perform data **interrogation** in an experiment and associate rules of engagement early on for data usefulness
- Business can sift the data to curate the **context**
- Big Data needs to be **curated** to be useful
The Paradigm Shift

Data Warehouse

IT
- Driver
- Program Owner
- Budget Sponsor
- Maintain
- Support

Business
- User

Big Data

IT
- Facilitate
- Maintain
- Support
- Manage

Business
- Driver
- Budget Sponsors
- Program Owner
- Define & Consume
Perspective – Food For Thought
Integrated View

Data Analytics (RDBMS)

Big Data Infrastructure

The New Data Warehouse

Taxonomy
Metadata
Search
Semantics
Visualization

You Still Need The RDBMS – But the New Data Warehouse is beyond just the RDBMS
Critical Success Factors

- Business needs **to own and execute** the Big Data program
- Data collection and discovery is the **most critical** step
- **Metadata** is needed to process the data **prior and post** Data Warehouse integration
- **Data quality** can be **processed** by integrating **Taxonomies**
- **Data visualization** is needed to discover data
- **Metrics** and **Metadata** will be the bridge to integrate to the Data Warehouse
Questions
Thank You

Contact

rkrish1124@yahoo.com
Twitter: @datagenius

Images Source - Internet