Big Data: Strategy & Execution

Ajay Bhargava
Director – Analytics & Big Data
Insurance & Healthcare, TCS

ajay.bhargava@tcs.com
1. **Verb** (Analysis - what you **do** with Big Data) & the “other” **Adjective-Noun** (Business Insights/Value) is far more valuable than the **Adjective-Noun** (Big Data) itself

2. **A critical prerequisite** for Big Data Initiatives is a good **Use Case**

3. “**C**”- **level support** for sponsoring the business use case is critical to get your relationship with Big Data off the ground

4. **Data as a Strategic Asset. Is P&C Insurer == Retailer in disguise?**

5. **Good Management of Big Data (Quality, Architecture, Governance, Security, Breadth & Depth)** $\Rightarrow$ **Quality of Analysis** $\Rightarrow$ **Better Value**

6. **SENSE & RESPOND** are in a vicious, faster cycle than ever before
Initial Interests from our Customers (sample set)

1. **Customer-Centricity: CMO, CIO**
   - Cross-Channel (360° view across Website (Direct), Agent, Call Center, Social, MobileApp, eMail, Fax, Mail etc.)
   - Marketing Experimentation - New Products & Services
   - Multi-Dimensional Segmentation of 1
   - Recommend Next Best Offer
   - Listening on Social - Sentiment & Weblog Analysis
   - Cycle efficiencies from Prospect to Customer
   - Call Center: Cost to Investment Center
   - Gaming Data - Social Behavior - [www.vwobservatory.com](http://www.vwobservatory.com)

2. **Operations: COO, CFO, CCO**
   - Customer Service - Claims Cycle business process optimization
   - Reduction in Business SLAs: ETL performance

3. **Risk Management: CIO, CCO, CFO**
   - Fraud: Business Rules, Predictive Analytics, Text Mining, AR Mining
   - Search: Structured & Unstructured
POC Objective
Sentiment Analysis of in-house product (Kaching) using Facebook, Twitter and Gizmodo user comments

HADOOP Stack used for Implementation
• Data extraction from source data using Hadoop common libraries into HDFS
• Load data into HIVE from HDFS for performing sentiment analysis
• HIVE UDFs and Mahout libraries for sentiment analysis
• Apply business rules using Pig Scripts for web log analysis
• Pentaho Report Designer for reporting the results
### Use Case 1
Get the Top 10 accessed websites by reading through the performance log

<table>
<thead>
<tr>
<th>Job #</th>
<th>Data (GB)</th>
<th># Records</th>
<th># Mappers</th>
<th># Reducers</th>
<th>Processing Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>285</td>
<td>423,181,518</td>
<td>1061</td>
<td>285</td>
<td>01 hr 24 min 00 sec</td>
</tr>
</tbody>
</table>

### Use Case 2
Reading through the performance logs, find the typical number of transactions (count of websites accessed) in a single user session

<table>
<thead>
<tr>
<th>Job #</th>
<th>Data (GB)</th>
<th># Records</th>
<th># Mappers</th>
<th># Reducers</th>
<th>Processing Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>285</td>
<td>423,181,518</td>
<td>1061</td>
<td>285</td>
<td>01 hr 42 min 00 sec</td>
</tr>
<tr>
<td>2</td>
<td>1.08</td>
<td>75,624,607</td>
<td>4</td>
<td>2</td>
<td>01 hr 01 min 43 sec</td>
</tr>
</tbody>
</table>

### Use Case 3
Group customers based on their residential address, postal codes and identify the types of interaction

<table>
<thead>
<tr>
<th>Job #</th>
<th>Data (GB)</th>
<th># Records</th>
<th># Mappers</th>
<th># Reducers</th>
<th>Processing Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.34</td>
<td>57,414,131</td>
<td>17</td>
<td>5</td>
<td>02 min 42 sec</td>
</tr>
<tr>
<td>2</td>
<td>4.89</td>
<td>90,986,110</td>
<td>19</td>
<td>5</td>
<td>03 min 14 sec</td>
</tr>
<tr>
<td>3</td>
<td>0.74</td>
<td>27,419,136</td>
<td>5</td>
<td>1</td>
<td>01 min 59 sec</td>
</tr>
<tr>
<td>4</td>
<td>0.0007</td>
<td>31,112</td>
<td>14</td>
<td>0</td>
<td>00 min 18 sec</td>
</tr>
</tbody>
</table>
“Customer Value Management” Application receives customer data from upstream systems in flat files. This data is extracted and loaded into Data Warehouse (Teradata) using ETL (Informatica) job.

The job includes:
- **Referential Integrity check**
- **Technical Validation**
- **Delta identification**
- **Business Validation**
- **Data Cleansing**
- **Data Transformation**
- **Data Enrichment**

Perform single customer view (SCV) process against the huge data in Teradata and finally a Master qualification Process.

**TCS Hadoop Solution**
- **Scoop/Java Frameworks** - Load the Data to HDFS
- **Map Reduce Jobs** - Data Validation & Transformation
- **TCS Workflow framework on Oozie** - Workflow Management
- **TCS Rules framework** - Configuring Transformation & Validation Rules
- 0.7 GB (600K Records) of incoming records (Customer and Policy) were validated against 14 GB (64 Million) Records and transformed.
- Hadoop Infrastructure can horizontally scale by addition of any number of nodes.
- Traditional ETL will have limitation (Not possible) to scale in processing TBs & PBs.
- The data was extracted to HDFS from App Database for optimal performance.

Actually executed scenarios and the results are consistent
Use Case: e.g. Evaluate the impact of Wal-Mart bribery allegations on our Portfolio

1. Start with (Proof of Concept/Deployment)
   - **Big:** Business Value Use Case: Simple Unstr Search
   - **Small:** Scope: Small set of PDF,DOC
   - **Big:** Data Platform/Architecture: EC2 9-node, Lucerne
   - **Small:** Investment: Miniscule
   - **Big:** Support/backing: CTO of 1 major LOB
   - **Small:** Team: 6
   - **Big:** Collaboration - IT & Business: Use case dictated
   - **Small:** Duration: ~ 3 months
   - **Big:** Data Management emphasis: No Priv./Qual. Issues
   - **Small:** Increments of Success: Move to Str Search

2. Reevaluate/Deploy

This capability is **NOT POSSIBLE** in the current environment. Unstructured Data is rarely & inconsistently looked at, and people dependent.
Thank You

ajay.bhargava@tcs.com