MAKE BIG DATA AGILE: IMPROVE RELIABILITY AND TRUSTWORTHINESS OF BIG DATA BY ALIGNING PROCESS AND DATA

The Data Warehouse is More Urgently Needed Than Ever

- Operational data increasingly dominant
- Low latency data trending toward zero delay
- Customer behavior insights being monetized
- “Big data” introduces many new challenges
  - Huge volumes, many types of sources
  - Structured ➔ semi-structured ➔ unstructured
  - Non-relational processing
  - Raw data, often dirty, often sparse
Threats to the Data Warehouse Mission

- Waterfall approach
  - Compartmentalized responsibilities
  - Formal specifications, handoffs, change bureaucracy
  - Excessive documentation
  - Resistance to course changes
  - Infrequent deliverables

- Integration overwhelm

- Untrustworthy, unreliable data

- “Starting Over” because of Big Data ??

Big Data Extends the Data Warehouse

- Use cases that drive the bottom line:
  - Targeted ads in real time based on web site activity and/or geo-tracking
  - “Satisfaction crash” intervention based on voice-to-text analysis of support calls
  - Customer sentiment change response based on social media
  - Emergency response planning based on satellite image comparison
  - Fraud detection based on cross-market brokerage transaction comparisons
Adopt the Best Aspects of Agile Development

- **Business driven**
  - Sophisticated business sponsor
  - All IT members involved in KPI identification

- **Small teams** with end to end responsibility
  - Eliminate non value-added tasks

- **Decompose** the project

- **Frequent** deliverables

- **Reasonable** documentation

- **Expect and encourage** mid-course corrections

Use the Value Stream Map to Identify Best Agile Opportunities
Kimball Lifecycle Approach is Inherently Agile

Prepare for an Agile EDW with Data Profiling

- Data profiling is a necessary foundation of the EDW
- Qualify or disqualify new data sources at the earliest opportunity
- Implement data quality triage:
  - Demand source system fix, or
  - Correct data in DW ETL, or
  - Tag data and pass through
- Embed data profiling throughout pipeline:
  - Source ➔ ETL back room ➔ BI front room
- Integrate data profiling tightly with data flow development: discovery spawns in line module
Agile Data Governance Fits Rhythm of EDW Development

- Recognize, protect data as a **corporate asset**
- **Continuously manage** data quality at all levels
  - Operational sources
  - Data integration pipelines
  - BI delivery
- **Continuously extend** EDW integration with conformed dimensions driven by enterprise MDM
- **Continuously add** new sources, facts, dimensions, and dimension attributes
  - Dimensional approach minimizes rework, impact

Agile Approach: Four Active Steps

- Manage/plan with the data warehouse **bus matrix**
- Develop with **enterprise MDM**
- Integrate incrementally
- Manage data quality incrementally
Agile Approach: Use the Data Warehouse Bus Matrix

- Decompose and manage data warehouse implementation projects
- Implement rows of matrix one at a time

<table>
<thead>
<tr>
<th>Business Processes</th>
<th>Date</th>
<th>Product</th>
<th>Customer</th>
<th>Store</th>
<th>Website</th>
<th>Promotion</th>
<th>Terms</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web Sales</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Store Sales</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corporate Sales</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Promotion Tracking</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Credit Operations</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social Media</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Call Center Voice</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Agile Approach: Leverage Development with Enterprise MDM

- Move entity creation and admin to operations
  - Move away from downstream MDM
  - Reduce repeated ETL demands for same entities
- Establish single source for entity content
  - Operational systems, social media tracking, and EDW subscribe to MDM
  - Roll out MDM incrementally
Agile Approach:
3. Integrate Incrementally

- Data warehouse integration is **drilling across**:

<table>
<thead>
<tr>
<th>Customer Category</th>
<th>Web Sales ($M)</th>
<th>Store Sales ($M)</th>
<th>Corporate Sales ($M)</th>
<th>Call Satisfaction (%)</th>
<th>Sentiment Change (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bronze</td>
<td>$214</td>
<td>$383</td>
<td>$199</td>
<td>34%</td>
<td>+15%</td>
</tr>
<tr>
<td>Gold</td>
<td>$563</td>
<td>$887</td>
<td>$447</td>
<td>42%</td>
<td>+17%</td>
</tr>
<tr>
<td>Platinum</td>
<td>$1,023</td>
<td>$2,248</td>
<td>$850</td>
<td>83%</td>
<td>+35%</td>
</tr>
</tbody>
</table>

- Agile advantages of **conformed dimensions**
  - Install conformed attributes **one source at a time**
  - Add new conformed attributes **one at a time**
  - Deliver value **with each iteration**

---

Agile Approach:
4. Manage Data Quality Incrementally

- Use data profiler to **build data quality screens**
  - **Column** screens
  - **Structure** screens
  - **Business rule** screens

- Manage data quality through **event database**

- Agile advantages of this DQ architecture
  - Build screens **one at a time**
  - Add or subtract screens
Agile Approach:
Special Considerations for Big Data

- Big data fits comfortably with agile approach
  - Agile approach tolerates uncertainty, back tracking, and midcourse corrections
  - Big data often must be analyzed before structure is understood

- Extend familiar RDBMS experience to non-relational data with Hadoop resources
  - MapReduce, Hbase, Cassandra
  - Data profiling extended to non-relational data
  - Hadoop as a preprocessor for creating structured data

Next Steps

- Integration is your fate!
  - Establish continuous agile process for identifying, vetting, and implementing new data sources

- Use data virtualization for rapid prototyping and proofs of concept

- Choose a Big Data source and a Hadoop configuration as a pilot project
  - Build IT skills
  - Introduce new KPIs to business community

- Use Kimball/Informatica Big Data white paper as guide to growing EDW to support Big Data
The Kimball Group Resource

- [www.kimballgroup.com](http://www.kimballgroup.com)
- **Best selling** data warehouse books
  - NEW BOOK! The Kimball Group Reader
- In depth **data warehouse classes** taught by **primary authors**
  - Dimensional modeling (Ralph/Margy)
  - Data warehouse lifecycle (Margy/Warren)
  - ETL architecture (Ralph/Bob)
- **Dimensional design reviews and consulting** by Kimball Group principals