Developing "Best Practices" in Data Management, Data quality, and Data Integration of Big Data

John Barry Taylor
LURES: License Utilization and Revenue Enhancement System

- Envisioned as a SAS solution for Customer Relationship Management

- **Goal:** Integrate all databases at TPWD that track customer interactions, especially revenue
Problems:

- Disparate database platforms
- No previous consideration of integration
- Data quality issues common
Addressing the 4 “V”s of Big Data

- Volume
- Velocity
- Variety
- Veracity
Initial phase

- Purchased SAS Enterprise BI, single DataFlux license, SAS Access for OLEDB
- OLEDB, BI Metadata Server allowed us to integrate data from Oracle, Sybase, MS SQL Server, and also MS Access and Excel data
- Used SAS Enterprise Guide to build Extract, Transform, and Load (ETL) jobs to create SAS data tables optimized for reporting
Developing a solid plan for security and a complementary metadata setup essential for producing useful content

Generally based on functional roles (based on need for access vs. arbitrary organizational hierarchy)

To protect users from misuse of data, access to raw data limited to “Subject Matter Expert” report developers
Consumers vs. Power Users

- Power Users *(report builders, data stewards)* can access raw data
- Consumers can only access data via Information Maps, Stored Processes, or SAS Tables, all specifically designed and formatted for ease of use
<table>
<thead>
<tr>
<th>Permissions</th>
<th>Consumers</th>
<th>Power Users</th>
<th>Admins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read Metadata</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Write Metadata</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Check In Metadata</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Write</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Create</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delete</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Administer</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>
• Great deal of effort spent on establishing “Best Practices” for organization of, and access to, data products
Metadata Organization

[Image of a file explorer window showing various folders and files organized in a hierarchical structure.]
Metadata Organization
Great deal of effort spent on establishing “Best Practices” for organization and access of data products

Settled on a naming convention that helps consumers find the products they seek
Metadata Organization
Phase 2

- Upgraded to more robust BI platform
- Purchased SAS Data Integration Studio
- SAS DataFlux integrated with DI Studio
EG vs DI Studio as an ETL tool

- Both can be used for scheduling ETL jobs
- Data validation and cleansing can be done within a single DI Studio job
- In DI Studio, jobs are Metadata objects
  - Enables scheduling of dependencies within jobs
  - Enables impact analysis for metadata changes
  - Enables Audit and Performance tools to be used
Addressing the 4 “V”s of Big Data

- Volume
- Velocity
- Variety
- Veracity
Volume

- Pass-thru SQL
- ETLs for pre-summarization of complex data
- E-mail alerts that report on ETL success or problems
- OLAP Cubes
Velocity

- Scheduling of ETL jobs to run overnight
- Indexing for increased query speeds
- Hash
- Removal of data that adds no value
- Appropriate infrastructure sizing can’t be ignored
Variety

- “Subject Matter Expert” involvement in building data products
- OLEDB as a universal translator across database vendors *(ODBC another option)*
- Data standardization in ETLs
- DataFlux for probabilistic matching on customer information *(names, address, birthdates, etc.)*
Veracity

• Data explorations to identify quality issues
• Right fielding
• Spell correction
• Address standardization
• Geocoding to fix incorrect location attributes
• Data preparation specifically for analytics