The 5 Components of a Data Strategy

Evan Levy
SAS
Vice President, Data Management Services

The Reality of Applications and Data

Applications
- Rigorous development process
- Economies-of-scale methods

Data
- Data is a “from scratch” activity
- Data standards are often point solutions

IT strategy defines tools, platforms, development and approach

Most organizations have no data strategy
What is a Strategy?

“…a method or plan chosen to bring about a desired future, such as achievement of a goal or solution to a problem…”

“...a careful plan or method for achieving a particular goal usually over a long period of time…”

“...strategy is different from vision, mission, goals... It is the result of choices on where to play and how to win, to maximize long-term value…”

What is a Data Strategy?

A plan designed to improve all of the ways you acquire, store, manage, share, and use data

Data Strategy

**The 5 Essential Components**

- **Identify**
- **Govern**
- **Provision**
- **Integrate**
- **Store**

THE CORE COMPONENTS
Data Strategy

The Core Components

- **Identify**: The ability to identify data and understand its meaning regardless of its structure, origin, or location
- **Provision**: Enabling data to be packaged and made available while respecting all rules and access guidelines
- **Store**: Persisting data in a structure and location that supports access and processing across the enterprise
- **Integrate**: Moving and combining data residing in multiple locations and providing a unified view of the data
- **Govern**: Establishing and communicating information policies and mechanisms to ensure effective data usage.

Data Strategy is not “One Size Fits All”

- Each component is independent and can evolve as needed
- Each component is specific to an individual set of skills and capabilities
- Not all organizations need an “enterprise-class” data strategy
- Complexity (and maturity) increase with organizational scope
- The Data Strategy should establish goals for each component
The Core Components

**Identify**

The ability to identify data and understand its meaning regardless of its structure, origin, or location

**Challenges**

- Data Sources aren’t documented; existence is based on tribal knowledge
- There is no list of data elements, their names, and their meanings
- Terminology and meanings can vary across applications (and business areas)
- Groupings and hierarchy varies across business areas
- The “best source” is based on current views, opinions, and convenience

**Strategy Options**

<table>
<thead>
<tr>
<th>Terminology &amp; Meaning</th>
<th>Application Aligned</th>
<th>Business Defined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metadata</td>
<td>Static Document</td>
<td>Actively Managed</td>
</tr>
<tr>
<td>Organization Scope</td>
<td>Single Organization</td>
<td>Enterprise</td>
</tr>
<tr>
<td>System Adoption</td>
<td>Single System</td>
<td>All Systems</td>
</tr>
<tr>
<td>Grouping / Hierarchy Support</td>
<td>Application Aligned</td>
<td>Business Defined</td>
</tr>
</tbody>
</table>

**Provision**

Enabling data to be packaged and made available while respecting all rules and access guidelines

**Challenges**

- Once created, data is typically required by 10+ systems
- Sharing data is a courtesy (at the convenience of the source not the consumer)
- Source selection is based on availability without attention to accuracy
- Every data source has a unique set of methods for sharing and distribution
- No one is responsible for sharing data

**Strategy Options**

<table>
<thead>
<tr>
<th>Packaging</th>
<th>Source Formatted</th>
<th>Consumer Packaged</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platform Access</td>
<td>Online &amp; Bulk Extract</td>
<td>Tracking</td>
</tr>
<tr>
<td>Internal storage</td>
<td>Delivery Tracking</td>
<td>Acceptance Checking</td>
</tr>
<tr>
<td>None</td>
<td>Actively Monitored</td>
<td>Data Audience</td>
</tr>
<tr>
<td>Developers</td>
<td>Users &amp; Partners</td>
<td></td>
</tr>
</tbody>
</table>
The Core Components

Store

Persisting data in a structure and location that supports access and processing across the enterprise

Challenges

- Everyone has their own copy of data
- There’s no onboarding process or means to identify available data
- Data is structured based on the source, instead of access and usage
- Manipulating data over a time is difficult (due to changes)
- A “one size fits all” approach isn’t appropriate for needs (online, BI, etc.)

Strategy Options

<table>
<thead>
<tr>
<th>Store</th>
<th>Onboarding Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application based</td>
<td>Raw, Cooked, &amp; Archived</td>
</tr>
<tr>
<td>Unmanaged</td>
<td>Versioned</td>
</tr>
</tbody>
</table>

| Application Specific   | Centrally Managed |

Integrate

Moving and combining data residing in multiple locations and providing a unified view of the data

Challenges

- Every project team (or user) integrates content based on their own logic
- Even with MDM folks still build their own match logic
- The DW team has ETL standards; other teams have their own standards
- Recording business rules in ETL metadata doesn’t help “self-service” data users
- There no centralized mechanism for tracking where the data has gone

Strategy Options

<table>
<thead>
<tr>
<th>Identification / Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central Rules</td>
</tr>
<tr>
<td>Governance aligned</td>
</tr>
<tr>
<td>Rule Identification</td>
</tr>
<tr>
<td>Centrally Managed</td>
</tr>
<tr>
<td>Source Specific</td>
</tr>
<tr>
<td>Tax &amp; Analytic Systems</td>
</tr>
<tr>
<td>Multi-Subjects</td>
</tr>
<tr>
<td>Actively Managed</td>
</tr>
</tbody>
</table>
The Core Components

**Govern**

- Establishing and communicating information policies and mechanisms to ensure effective data usage

### Challenges

- DG often focuses on metadata not policy or process definition
- Committees forget their focus is decision and delegation (they become a working team)
- Data is assumed to be accurate; there is no discussion about acceptance
- Mature governance should reduce committee activities not increase them
- Measuring policy/rule adoption is critical – and often ignored

### Strategy Options

<table>
<thead>
<tr>
<th>Policy and Rule Definition</th>
<th>Org Aligned</th>
<th>Cross Functional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Security</td>
<td>Application Specific</td>
<td>Business Defined</td>
</tr>
<tr>
<td>Data Quality</td>
<td>Application-centric</td>
<td>Stakeholder Defined</td>
</tr>
<tr>
<td>Data Management Oversight</td>
<td>None</td>
<td>Centrally Audited</td>
</tr>
<tr>
<td>Application-centric Mechanisms</td>
<td>Process Defined</td>
<td></td>
</tr>
</tbody>
</table>

---

### Real World Data Challenges

#### Users discussing a sales report

- There are 200 reports in the Green Book distributed to 800 users...
  - Numbers are interpreted differently across teams
  - Data accuracy is challenged
  - Report titles don’t match current terminology
  - Data source is unknown
  - No one is responsible for data correction

#### Profiling and mapping sales data

- Sales transactions are collected nightly...
  - There are 5 POS models (and 14 data layouts)
  - Reconciling logic is contained in ETL code
  - 4 systems require sales data (4 different code sets)
  - There’s no data stewardship process

#### Company has too many product descriptions

- There are 23 product descriptions...
  - The details vary from web to store to catalog
  - Vendors submit changes w/o our knowledge
  - No tracking of which descriptions are used
  - “The company looks bad”
  - Staff members waste time searching for the details
What’s Missing / What Can Help?

What details would simplify or improve the deployment of these reports?

- Use of standard terms within all reports (column names, titles, etc.)
- Documented definition of terms
- Identification of data sources within the report (or a link to those details)
- Identification of the data owner / steward of report
- A defined data acceptance process for those attributes included in report

Building the Data Strategy
Review the Data Strategy Components

- Establish standard business terms
- Construct data dictionary
- Identify source data stewardship of all shared data
- Require lineage details for all shared data
- Identify Systems of Record (by subject area)
- Implement DQ and correction
- Create process to identify new terms
- Identify role/responsibility for data correction
**Benefits of a Data Strategy**

- Identifies common goals across projects
- Supports creation of common methods and practices
- A framework for addressing common data needs
- Coordinate data initiatives and activities
- Mechanism to approve or deny new requests

**Moving Forward...**

*Identify the scope of your data strategy initiative*
- Identify the organizations that own/influence the various component areas
- Start small to limit initial effort; start large enough to ensure visibility

*Review your existing strengths and weaknesses*
- Consider expertise within each individual component area
- Focus on “good enough”, not perfection

*Consider future development efforts and ongoing obligations*
- Look at major application changes and upgrades
- Start with data sharing, access, and movement issues; use the components to categorize needs
Thanks!

Evan Levy
VP, Data Management Services
Evan.Levy@sas.com
@EvanJayLevy