Data Classification

The Foundation for Intelligent Information Management

David Hill    Wayne Kernochan
Principal     President

Infostructure Associates
Leveraging Information for Organizational Success
• The material contained in this tutorial is copyrighted by Storage World Conference (SWC).
• This presentation is a project of the SWC Education Committee.
• Individuals or companies wishing to use portions of this presentation should contact Lindsey Mitobe lindseym@storageworldconference.com
  – This information cannot be used outside SWC without the express permission of SWC management.
Agenda

- The benefits from and the problems for establishing a data classification process
- Understanding the basics of data classification
- General tool strategies for helping to classify data
- Applying different lenses to help in the data classification process
- How storage data classification relates to other technologies
- Conclusions
Data classification is the process of separating data into separate piles (i.e., categories) to which different policies apply.
Data classification organizes data so that IT can manage it better
  - Use of tiered storage (with resulting cost savings)
  - Better use of existing storage by being able to get rid of unnecessary data

Data classification is necessary to isolate compliance data so that the data can be managed effectively
• Information classification means faster, broader, more relevant query results for the “real-time enterprise”
SNIA’s Data Management Forum has a powerful vision for ILM.

Although infrastructure and services can exist without ILM, the starting step for ILM is data classification.
• Do you really need to drag all your tapes back from the warehouse for Sarbanes-Oxley legal discovery?
• Business management sees no compelling reasons for data classification
  – Better IT efficiency and cost savings are nice, but does not affect them directly
  – Compliance may be a necessity, but it only a targeted — not universal — application of data classification
• Data classification has to involve the business user and, even if willing, the process may be hard
  – Defining business rules, especially where terms such as customer and product are involved, across business groups may be difficult
  – Legacy business rules that have already been incorporated into applications may be poorly documented and crossing them may create a peril
• Data classification is “below the business level”
This is a start for building a compelling case for data classification.

<table>
<thead>
<tr>
<th>Problems</th>
<th>Results</th>
</tr>
</thead>
</table>
| **Business** | • Not globalized  
                     • Inaccessible (doesn’t exist)  
                     • Incomprehensible  
                     • CEO doesn’t know how he or she is doing  
                     • Enterprise reacts slowly  
                     • No one on same page |
| **IT** | • Uncoordinated  
                     • Low quality (mistakes)  
                     • Low performance  
                     • High admin costs  
                     • Bad decisions  
                     • No decisions |
| **Storage** | • Not available across vendors  
                     • Not classified by storage uses  
                     • Non-existent  
                     • Inefficient administration  
                     • No ILM |
Some Benefits from Extending Data Classification

- Developing greater knowledge of information enables querying across broader sets of data
- Identification of new relationships between data may be useful for competitive advantage
- Easier programming at a higher (business metadata level)
- Enhanced data quality (in conjunction with ETL [Extract, Transform, Load] tools)
Two Problems to Solve

- How to go about the process of data classification
- How that understanding helps in making a comprehensive case for data classification as the foundation for intelligent information management
Data Classification Playbook
Definitions

- **Data** — ordered (i.e., non-random) bits that are packaged into recognizable structures, such as files and database records
- **Content** — packages of data that contain useful information
- **Information** — human or computer interpretable data that is usable by business users
- **Metadata** — data about data
- **Intelligent Information Management** — enterprise-wide administration at the metadata business level across all vendors and data types.
What is data classification?

- Recall, data classification is the process of separating data into separate piles (i.e. categories) to which different policies apply.
  - Policies are the implementation of business rules
  - The business is responsible for setting the business rules
- Different categories are treated differently
  - Different service levels — such as data protection, data security, and compliance
  - May be on different tiers of storage
  - May be migrated when data classification status changes
• Treating data categories differently implies action based up Service Level Objectives (SLOs)
  – If no difference exists between two groups of data, they can be collapsed into one group

• Data classification is an on-going process
  – Data does not necessarily stay in the same category. Movement from one category to another is based upon change in status, due to conditions, events, or time.
More Thoughts on Data Classification

- Data classification is a means to an end and not an end in and of itself.
- Data classification is both a manual and an automated process:
  - **Manual** — end users develop business rules which are incorporated as policies that can be executed by software.
  - **Automated** — data can be identified or tagged as being in different classes based on policies.
  - **Iterative manual/automated** — classic chicken or egg — development of business rules may depend upon knowing what data is available and that may depend on using data discovery and data search capabilities.
General Tool Strategies for Helping to Classify Data

- By Application
- By Metadata
  - File Metadata
  - Extended Metadata
  - Content Analysis
• Application is classified as mission-critical, business-critical, task-critical, etc.
  - SLOs are set for the application
  - All data for the application goes along for the same ride with a recovery point objective (RPO) being set for the allowable data loss for the application

• **Advantage:** Simplicity

• **Disadvantage:** Wrong approach
  - This is application classification, not data classification
  - Business may benefit from application classification, but still needs data classification
Why not just application classification?

- Data does not belong over its lifecycle to one application
  - Classifying data is necessary because that determines which application should have control
  - Not classifying data would imply that data never changes value throughout its lifecycle if all data in the application is treated the same
- Two or more classes should not exist in the same application pool
  - Data which no longer serve the primary purpose of the application should be physically separated into another pool
  - Logical separation is not enough. Trying to handle exceptions within an application would overburden the application
- **File Metadata** — classify data by filename extension, pathname, size, owner, creation date, etc.
  - **Advantage**: Automatically available without extra effort
  - **Disadvantage**: Limited capabilities

- **Extended Metadata** — additional metadata — such as security, detailed ownership (e.g. department), and workflow routing
  - **Advantage**: Improved granularity for data classification
  - **Disadvantage**: Limited primary to process control and not content management
Adding in Content Analysis

- **Indexing** — organizes content (such as by keyword) so that they can be used by full-text Boolean search engines
  - **Advantage:** This is really metadata and is the heart of the global metadata repository
  - **Disadvantage:** This is content analysis, but the next level of the game is context analysis; check to see how this helps

- **Full Content Analysis** — this requires examining the contents of the data packet to understand how this data is related to other data
  - **Advantage:** Deeper and richer analysis
  - **Disadvantage:** Time-consuming and not scalable
• **Storage management** — discovers, monitors, and controls physical storage assets.

• **Data management** — the non-data-path control and use of the data itself from creation to deletion, such as migration, replication, and backup/restore processes.

• **Information management** — manages the content and decision-making relationships of information as it moves through the lifecycle of a business process, such as records management and content management.
The management lens frames the type of benefits that can be derived from data classification for ILM

**Storage management**
- **Accomplishes:** Tiering of storage
- **Key Benefit:** Cost savings

**Data management**
- **Accomplishes:** data migration, data protection, and compliance
- **Key Benefits:** Higher service levels, IT manageability

**Information management**
- **Accomplishes:** better use of information assets, eDiscovery
- **Key Benefit:** Increased value (gross margin, revenues, productivity) to the enterprise

Determine how your data classification strategy will affect each of the types of management.
## The Data Lens

<table>
<thead>
<tr>
<th>Type</th>
<th>Structured</th>
<th>Semi-Structured</th>
<th>Unstructured</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Common Forms</strong></td>
<td>Database</td>
<td>“Text” documents such as e-mail, word processing, presentations, spreadsheets</td>
<td>Natively bitmapped data, such as video, audio, pictures, and MRI scans</td>
</tr>
<tr>
<td><strong>Key Differentiator</strong></td>
<td>Sort</td>
<td>Search</td>
<td>Sense</td>
</tr>
</tbody>
</table>
| **Examples**       | • OLTP systems, such as CRM and ERP  
• Data warehousing | • Personal productivity, such as e-mail and word processing  
• Web sites using HTTP | • Entertainment, such as video and audio  
• Imaging, such as digital photography and bitmapped medical tests |
Distinguishing Between Unstructured and Semi-Structured Data

- Unstructured data are typically BLOBs (Binary Large Objects)
  - Native (i.e. without enhanced metadata) unstructured data cannot be used by content-aware applications
  - Managed through file metadata
  - Administration and classification different from other data types
- Note that business documents (word processing, presentations, spreadsheets) are often incorrectly classified as unstructured data
  - Business documents are semi-structured data
  - *Semi-structured data* can be managed through not only file data, but also *through content-aware applications*
• Metadata all over the place

- Enterprise Applications
- Business-Process software
- Business Intelligence

Business Level

Databases

Global metadata repository

Content Management

Database Level

Storage Area Networks

Data-level metadata sub-repository

Systems Management

Hardware Level

Disk, Tape

Hardware-level metadata sub-repository

Servers

Cross-database, data-copy, business-process metadata

Structured, semi-structured, unstructured metadata

File, block metadata (age, striping)
Don’t reinvent the wheel - 2

- Tools to generate, extract, and combine metadata
  - Example: Enterprise Information Integration (EII)
How Accurate Will The Classification Be?

- “When we first collected the data, there were 83,000 customer records, 54,000 supplier records, and almost 100,000 products in various different systems. And the key goal here was a new repository where there should be only one record for each product, only one record for each customer, and one record for each supplier.”
- “Many suppliers ... supplier records, had a supplier number and a supplier name...maybe a country code. Not even an address existed.”
- “There’s no validation to make sure that you entered the right piece of information in the right field. And, hence, you have this mixed up master data for address. This might be okay for sending out letters to your business partners, but if you want to do any form of reporting on this... trying to do some even geographical segmentation, it’s very and completely impossible.”
- “15,000 records had four zeros as the postal code ... even though four zeros is not a real postal code.”
  - Thomas Ravn, Arla Foods, 2005, on their Master Data Management effort
Steps to Take

- Determine the scope of the data classification effort
  - *Universal* — include all data types and all applications
  - *Targeted* — select one data type and one or a few applications
  - Targeted vs. universal — easier approval and execution for targeted, but, if universal is the long-term goal, may be better to seek approval and plan up front

- Match up what types of data need to be classified with what the data classification tool that you choose does
  - *Structured* — databases
  - *Semi-structured* — e-mail and other documents
  - *Unstructured* — BLOBs

- Map the data classification process with the management (storage, data, and information) benefits that can be derived.

- Use pre-existing tools and metadata to populate the classified-data repository where possible
• **Active Changeable** — live production pool of data, which is presumed to be changeable (i.e. read/write)
  - Actually, data that is unlikely to change (such as closed transactions or old e-mails) are commingled with changeable data

• **Active Archive** — fixed content data that still serves an active (e.g. online) business need

• **Deep Archive** — offline fixed content data that is recalled only in special circumstances

• **Data Destruction** — the process of destroying all copies of selected data
Where Should Your Data Be Today?

Production Data

Active Changeable Pool  →  Active Archive Pool  →  Deep Archive Pool

Endpoint Analysis: What characteristics should data have to be in each pool?

<table>
<thead>
<tr>
<th>Availability</th>
<th>Online</th>
<th>Online</th>
<th>Offline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of Data</td>
<td>Changeable data</td>
<td>Fixed data</td>
<td>Fixed data</td>
</tr>
<tr>
<td>Purpose</td>
<td>•OLTP</td>
<td>•Business intelligence</td>
<td>Ad hoc recall, e.g. regulatory requirements</td>
</tr>
<tr>
<td></td>
<td>•Collaboration</td>
<td>•Reference</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>•Compliance</td>
<td></td>
</tr>
</tbody>
</table>
Applying Data Classification for Intelligent Information Management

- Data classification applies the discovery of data, the application of policy-drive search capability, and the assignment of data to the proper storage pool for its current intended purpose.
- Now the application(s) that serves each class of data can serve the purpose for which each was intended
  - Use is made of the global metadata repository
  - Data classification has created the foundation for intelligent information management.
Conclusions

- Data classification is both a manual and an automated task.
  - To help classify, different perspectives can be used, e.g., endpoint analysis uses the current purpose to help determine where the data should.
  - Tools that use metadata and content analysis can automate the assignment once business rules are implemented into policies.

- A compelling case for data classification can be made
  - Especially when the case for it extends beyond storage to embrace IT and the business as a whole.
  - Especially when classified data is the foundation of a global metadata repository
    - But you’re already talking prose, even if you didn’t know it! (Master Data Management)
Thank You! – Questions please.

David Hill
(781) 326-0038
www.mesabigroup.com
davidhill@mesabigroup.com

Wayne Kernochan
(781) 862-5236
www.valleyviewventures.com
wayne.kernochan@valleyviewventures.com

Infostructure Associates
Leveraging Information for Organizational Success