

Data Management Maturation Process

Early Stages

- “Buy more” approach to data management
 - ▣ Ignore the problem, buy more disk, tape, etc. and keep everything
 - ▣ "Tiered Storage" look for cheaper ways to implement more "cheaper" disks
 - ▣ Clean up consists of sending out emails about targeted data, backup target data, delete targeted data and hope no one complains
 - ▣ Implement search and classification technology -> depending on indexing scheme, index alone can grow to 28% of source data size
- Many at this stage have no visibility into the problem they are about to encounter. Hindsight visibility is 20/20, begin to appreciate scope of problem after encounter problems associated with large scale data management.
- Establish a common namespace (Network File Switch, Clustered Filesystem, etc.)
- Implement search engine technology
- Use of engine recognizes a underlying problem exists -> not managing data in a repeatable, consistent, disciplined manner
- Realize search engine is part of the solution but limitations exist:
 - ▣ non-ASCII binary & image files
 - ▣ "a priori" knowledge required. Inability to infer semantics from syntax via textual analysis.
 - ▣ Ignores physical resource aspect of data management, often promotes inefficient use of physical resources (redundant data, poor data and storage management)
 - ▣ non-public or non-published information
 - ▣ no information “end of life” of disposal facilities to cull data -> managing less data is integral component of data management



Data Management Maturation Process

Middle Stages

- Manage data through the use of application "silos," data that needs to be managed, controlled, organized needs to be in the silo.
- Realize "silos" are part of the solution but limitations exist:
 - Difficulty managing data outside the application as actual file names are changed to SKU numbered names (application is holding the data hostage)
 - Difficulty predicting and adopting next "useful" application
- Recognize benefit of predicable, consistent, file system organization. Organize data then point applications to it .
- Recognize the importance and role of naming conventions
- Manually implement naming conventions, usually effective to about 3 levels deep in hierarchy
- Data Wrangler (Data Management or Data I/O) staff continues to increase
- Data Management team manually classify & interact with user community to obtain key information required to bring some level of precision to the process.
- Constant "finger wagging" at non-conforming users
- Begin to automate naming conventions through the use of internally developed tools & scripts (make shot, make project, etc.
- Use of internal development resources on non "core competency" development
- Begin to encounter scaling, cost, support, turnover, and flexibility issues with in house tools as they often evolve with one script layered onto another, to another, etc.
- Complexity begins to require increasingly higher level of technical expertise.
- Many aspects of internal solution (i.e. Naming conventions, file system partitions) are "hard coded" to facility, tools require constant developer maintenance

Data Management Maturation Process

Advanced Stages

- ⦿ Extend knowledgebase for data management decisions to incorporate key information that resides within the end user community
- ⦿ Recognize data management is very complex, distributed problem set. Requires automated processes along with distributed responsibilities.
- ⦿ Reach stage of enlightenment - the collection of distributed file systems comprise the largest possible "data silo," thereby removing limitations on file types, formats that can be managed.
- ⦿ Second stage of enlightenment - New file system perspective -> files in a directory considered to be a hierarchical database with files as atomic data elements, file names are semantic tags, and the directory structure is the schema.
- ⦿ Exploring mechanisms to cross-reference or cross-index views of file system (analogous to iPod)
- ⦿ Exploring mechanisms whereby incoming data is supplied to organization with structure that enables an immediately usable format, eliminating manual steps.
- ⦿ Exploring mechanisms to accommodate virtual workplace, with "pack & go" and "join & sync" file system capabilities
- ⦿ Exploring ability to maintain file system organization through application software's "save as" file system entry point
- ⦿ Exploring automated naming conventions within actual software applications