RL Consulting

People • Process • Technology
Organization • Integration

IT Services Management

Data Management Process

White Paper

Prepared by:
Rick Leopoldi
May 25, 2002

This paper discusses the processes and methods to define, characterize, and orient data within the storage technology hierarchy. In addition, it addresses various and necessary data usage and profile characteristics to effectively manage data and storage with an IT service management perspective.

Introduction

Managing physical and logical storage devices is a necessary part of storage management. That is, understanding the performance, capacity, and utilization of storage devices is critical to any storage management effort. However within a larger scope of IT service management, it is not the only part. Managing the data that resides on the storage devices is an essential part of an ITSM effective data and storage management effort. Data management implies an understanding of the data usage requirements, characteristics and life cycle and employing them to manage data on physical and logical storage devices.

Analyzing 3 separate areas of data can accomplish this; data definition, data usage life cycle, and data profile characteristics. Once completed, a more effective ITSM data and storage management effort can be achieved allowing a more definitive understanding of where data can be placed physically and logically within the storage technology hierarchy.

Data Definition Process

This initial data definition process provides a high-level overview to characterize data and determine usage requirements by mapping it to organizational views.

Corporate Data

Typically, corporate indicates data that is vital to the corporation such as data that is used for establishing and managing corporate policies (i.e., Human Resource data) or data that has to be highly secure for reasons of trade or national secrecy. In most organizations there is typically several such databases such as the corporate books of account, maybe the database the client is using for employee policy management or trade secret files. A common thread in evaluating data in this category will be the "tightness" or "security" type rules of access associated with the data.

Public Data

This is reference data of use to the whole community that has increased value if it is current and correct. Examples would be legal documents, government regulations, technical reference data, document template data and so on. If the databases are large, and/or volatile, this data would typically reside on a SAN.

EMAIL databases would also fall into this category, allowing a wide community to access it. In addition, there is usually a need for some security at the individual or group level even though the whole database is still "Public". A shared Repository/Dictionary would be another example of "Public" data.

Divisional "Corporate" Data
This is data that is vital to the corporation but accessed and used by a smaller divisional community within the organization. This data could be files and/or databases that support corporate business functions and have corporate data characteristics such as restricted access requirements for human resource information, payroll, security, and trade secrecy.

Departmental Data
This is data that is of interest to a more localized community, literally a department, or maybe a geographical region such as a regional office. The data, at least in summary form, may well be part of a feeder system to databases required at the organization's business unit level or a corporate wide database, but in its detailed form it can typically reside on a departmental system.

Private Data
This is data that is private and/or personal to an individual, or a small homogeneous group of individuals. It may consist of working files that are a preamble to supplying data to departmental or corporate databases, but at the "working" stage are private. Security needs are typically less. The data may be downloaded copies of departmental or corporate data, where direct updates to the original file are not being done.

As with departmental data, there may be other factors that cause private data to be located on departmental shared LAN or possibly a SAN but this is atypical.

It is natural to associate private data type with workstations and this is an area where the platform overlap may occur. Typically, when LAN's behave as departmental systems, there may be issues of security or transaction rates that can impact placing this data on a department shared LAN or SAN.

The process for deciding where a given database should reside should begin with this initial data requirement and characterization analysis. This is a recommended process for managing data to a lesser number of standard platforms.

Data Usage Lifecycle

Throughout the useful life of any particular piece of data (i.e., file, dataset, or database), data usage characteristics such as data activity, performance, capacity, recently reference information, aging and backup requirements, and migration requirements are critical in understanding the nature of the data and how it is being used. Understanding these usage characteristics indicate the lifecycle attributes and are a necessary step to determine how to best manage the data given its particular requirements and usage characteristics. These make up a significant portion of a data profile.

Additional Data Profile Characteristics

In addition to data usage lifecycle characteristics, the following data characteristics are necessary to determine an effective data profile.

Security - this refers to the level of security needed to control access to the data.

Response Time - this is not the normal use of response time. In this case, it indicates the amount of time the user can wait for the data trying to be accessed.

Concurrency - the number of users needing potential concurrent access to the data

Updating - this gives an indication of the number of data accesses that are for updating. This may have implications on security needs and concurrency of data access.

Currency - this refers to the "up to date-ness" need of the user trying to access the data.

Commonality - different from concurrency, this is the number of concurrent users given a certain physical and/or logical grouping. Commonality suggests a measure of how many groups across the organization need access to the data.

Volatility - somewhat related to Updating, this is a measure of how rapidly the data changes from day to day.

Frequency of Use - again, somewhat related to other measures like Updating, Volatility and Concurrency, this is a measurement of the "rhythm" of access. For example, is it all day and every day, spasmodic, or infrequent?

Integrity Responsibility - this is basically to do with data ownership, which is critical to any data management effort. The owner is the organization that promises to keep the data as good and as current as the other users require. In this area there is a tendency to put the data near to where that organization resides, or on the platform that they normally use.

Management Overhead - this refers to system routines such as reorganization of data, journaling, migration, backup and recovery, extracts, and archiving.

Cost per Gigabyte - this refers to the cost of storage on a per gigabyte basis, but more specifically to some targeted, managed cost, not the actual cost. This is done because some technologies might "skew" the cost paradigm.

Size – typically, large databases will not fit on certain technology platforms. This is an area where the overlap between platforms can change over time.

Source - this refers to the natural source of the data as determined by the initial data definition process. That is, is the data corporate, public, divisional "corporate", departmental, or private?

Data Location Process

Presented here is one possible method to determine where data could be placed within the physical and logical storage technology hierarchy and where data could reside, physically. This high-level process uses input from the data definition, data usage lifecycle, and data profile characteristics.

In every organization, there are some databases where there are clear and compelling reasons why the data is held in one place (not necessarily on a SAN but usually). A typical example of an overall hierarchy of data would be:

Nature of Data	Possible Location
Corporate	SAN
Public	Shared LAN or SAN
Divisional "Corporate"	Shared LAN or SAN
Departmental	Shared LAN
Private	Workstation, Shared LAN, SAN

Beginning with this initial high-level data definition process, the basic idea is to determine, for each file, dataset, or database, the data usage characteristics using the data profile characteristics as a base set of criteria.

The approach is to take criteria such as security, response time, concurrency, updating, currency, etc. and use some metric to determine whether it is mostly true for this data, partially true, or not true at all (High, Medium or Low could be used).

This allows for an organization's unique requirements to qualify and quantify the importance of the particular file, dataset, or database for each criterion. As an example:

1) If any criteria is determined as high, the data should be located on a SAN, or
2) If the majority of criteria is determined as high, the data should be located on a SAN, or
3) There may be a case for weighting some of the criteria so that some items are more important than others.

Another possible perspective might be:

1) Common, high usage, lots-of-users file, with a high degree of currency, would be a SAN
2) A low update, local data, low-number-of-users file may be on a department and/or shared LAN
3) A read only, single user file could be on a workstation

The potential benefits of approaching data management in this fashion is that it:

- Allows the organization to qualify and quantify the importance (metric - High, Medium, Low) of each data profile characteristic and assign a "relative" priority for what is the most appropriate platform for the data
- Allows the organization to determine the benefits and drawbacks of each platform as it relates to how well that platform satisfies each of the data profile characteristics
- Could be used as a determination of where any given application should reside because typically the platform on which an application should reside is best determined by how well the requirements of the data are satisfied by existing on that platform

It should be noted that once a determination is made where a particular file, dataset, or database should reside, other data may have to be there too, if they are closely linked. It is likely that the application most responsible for updating the data will exist on the same platform, but not necessarily applications that only access it.