ILM Whitepaper V1.0

IBM Global Services

December 14, 2007

IBM CORPORATION, © 2005

© 2005 IBM Corporation

IBM Global Services

Prepared by:
Tim Waters – Senior Storage Consultant and Architect, IGS, (404)592-3394

Table of Contents

IBM © 2005

Table of Figures

IBM © 2005

Chapter

1

Data Rationalization

Data rationalization is the technique recommended by IBM to help uncover how much of a client's allocated and used storage capacity on a host-by-host basis is being utilized. This utilization, typically 50% industry wide according to Gartner, can be operated on in several ways. The two preferred methods to reclaim storage capacity are discussed in this chapter. Firstly, data cleanup is performed to remove, delete, archive, migrate, etc. files on affected servers where the value has dwindled over time. This actually lowers the effective utilization, perhaps to 40% on average. However, this offsets future growth as additional growth in data can reuse this recently freed up 10%.

Secondly, with 40% effective utilization from a host perspective, the storage capacity that was allocated to the hosts can be reclaimed or de-allocated which drives effective utilization up, perhaps to 80% at the host level. This de-allocated storage capacity can then be re-allocated to other hosts that are growing more rapidly. This free storage capacity offsets the need to purchase additional storage capacity and provides for cost avoidance in a business case scenario.

IBM TotalStorage Productivity Center for Data is the tool used to start the data rationalization process.

Common Agent Services

IBM TotalStorage Productivity Center for Data and IBM TotalStorage Productivity Center for Fabric run on a common agent services infrastructure. The infrastructure provides an agent manager for authentication, authorization of services and maintains a registry of configuration information about its common agents (managed computer) for each of the applications.

IBM TotalStorage Productivity Center for Data and IBM TotalStorage Productivity Center for Fabric will leverage the common agent services infrastructure. The common agent services registry database will reside on a local IBM DB2 database, located on the IBM TotalStorage Productivity Center for Data management server.

TPC for Data

IBM TotalStorage Productivity Center for Data is used to define storage-based policy management, automate file system extensions and report on database capacity. It is capable of reporting and managing various storage systems, including direct access storage, network attached storage and intelligent disk subsystems.

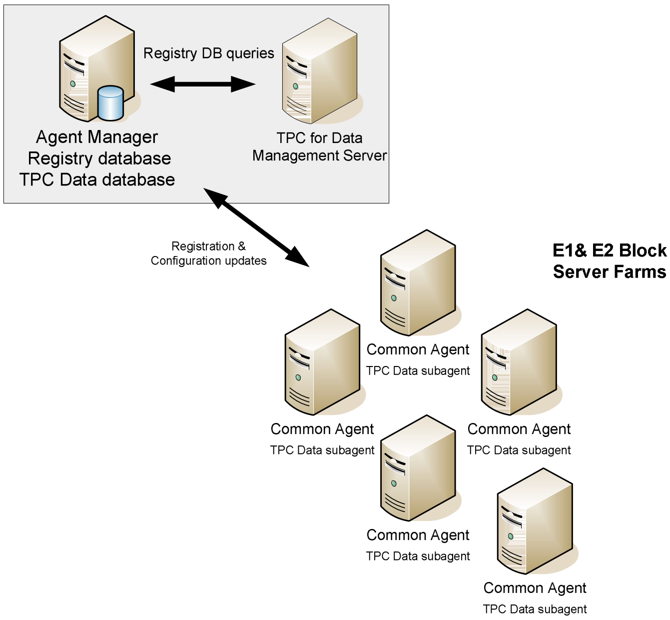
- The IBM TotalStorage Productivity Center for Data management server will be located in the Ford FMCC data center.

- The IBM TotalStorage Productivity Center for Data management server will reside on a dedicated server in the production environment. All agents will be

IBM © 2005

deployed in phased roll-out. Fifty (50) agents initially in test/development servers. Followed by two hundred a week until all are deployed.

Figure 1 shows common agent services supporting IBM TotalStorage Productivity Center for Data.



• Figure 1 Common Agent Services with IBM TotalStorage Productivity Center for Data

IBM TotalStorage Productivity Center for Data agents for Windows and UNIX will be deployed to all servers in the FMCC, ECC, and PDC data centers.

The IBM TotalStorage Productivity Center for Data database should be deployed to manage the following databases: SQL, Oracle and Sybase based on these criteria:

- Creating groups for databases will allow for easier monitoring.
- Defining a probe will assist in creating an inventory of files, logs and tables within the monitored database.
- Scans should be defined to gather extensive database statistics that will be leveraged for reporting.
- Alerts should be configured to forward SNMP traps to the CA Unicenter SNMP Manager console.
- Scan jobs maybe created to assist in monitoring database consumption and storage

The IBM TotalStorage Productivity Center for Data backend repository will reside on a local DB2 Enterprise Edition installation which can be the same database instance on which the common agent services registry resides.

IBM © 2005

Once the IBM TotalStorage Productivity Center for Data environment is deployed, probes will be configured and run to collect inventory, file system, database and storage subsystem information.

TPC for Data Sizing Recommendations

It is advisable not to manage more than one thousand five-hundred (1,500) IBM TotalStorage Productivity Center for Data agents per IBM TotalStorage Productivity Center for Data management server. If the Ford management environment grows to more than one thousand five-hundred (1,500) agents, a second IBM TotalStorage Productivity Center for Data management server will be required. The design described in this document allows for the possibility of adding additional IBM TotalStorage Productivity Center for Data management servers by leveraging the current common agent services infrastructure.

Host Agent Considerations

Agents should be deployed to all platforms types identified in the previous figure – Location and Machine Counts. This will provide Ford with a thorough data collection throughout the enterprise.

- Agent deployment could leverage the native Windows and Unix automated remote install functionality.

CIM Agent Considerations

IBM TotalStorage Productivity Center for Data uses CIM agents to gather and report on information from storage subsystem controllers. CIM agents will need to be deployed to support the full product roll-out.

Useful Reports

To support the Information Lifecycle Management (ILM) strategy at Ford, IBM TotalStorage Productivity Center for Data will be used as a tool for taking an inventory of enterprise data and storage assets. In support of this strategy, IBM will work with Ford to define and create customer reports to meet there reporting needs. In Addition, the following reports should also be used for Ford's information classification project:

- Perform an inventory of assets in the environment. This should be performed by running the following report. Reporting > Asset> System Wide > By Capacity. This report shows physical disk capacity as it relates to servers.

- Reports that are focused around file system utilization. For example: Reporting > Capacity > File system Capacity > By File system. This report shows the percent utilization of the file systems that is currently being monitored. This report could be sorted to only show file systems that have poor utilization. The report will assist in determining what might be a more appropriate capacity for the file system.

- Reports that are focus on file duplication. This report is useful for identifying files that are duplicate files. They may be wasting space or they are no longer needed. For example: Reporting > Capacity > Usage > Files > Duplicate Files > Network Wide.

IBM © 2005

- Reports on Obsolete and Orphan Files will also be useful. Generate reports that are focus on file duplication. This report is useful for identifying files that are duplicate files. They may be wasting space or they are no longer needed. For example: Reporting > Capacity > Usage > Files > Duplicate Files > Network Wide.

- Reports focused around file system activity. Reporting > Usage > Access Load > Access Time > By File system. This report will show a file system view and how much data has been accessed by time period. Ford should use this report to make decisions on which tier the file system belongs based on the frequency of data being accessed.

Information Types

An information type is defined by IBM to be the level of granularity of focus for the development of information classes. Ideally, every file in an I/T environment should be classified, but quantity limitations preclude that the level of focus in determining the information type should move up higher to the level of DB tables filesystems, volumes, database containers, logical disks, physical disks, servers, and applications. It is therefore sufficient to classify information using a server view for the level of granularity. Thus, servers (hosts) will be classified at Ford into information classes comprised logical groupings of functional and non-functional characteristics for valid data.

In-valid Data Cleanup

Since it is only the valid data that is classified, the in-valid data must be operated upon in a sufficient manner to free up valuable storage capacity and offset future data growth. This is accomplished by using TPC for Data to gather empirical information about an application or server and analyze this information. During analysis, certain thresholds are chosen to determine if data is stale, orphan, and duplicate. Once those thresholds are chosen, further analysis on the files and databases themselves will reveal chances to reclaim this used storage capacity by deleting or archiving these files and databases. After cleanup, the end result from a cost perspective is overall lower filesystem level or volume level utilization, but storage capacity to offset future growth and identifiable opportunities to reclaim unused storage capacity.

IBM © 2005

Chapter 2

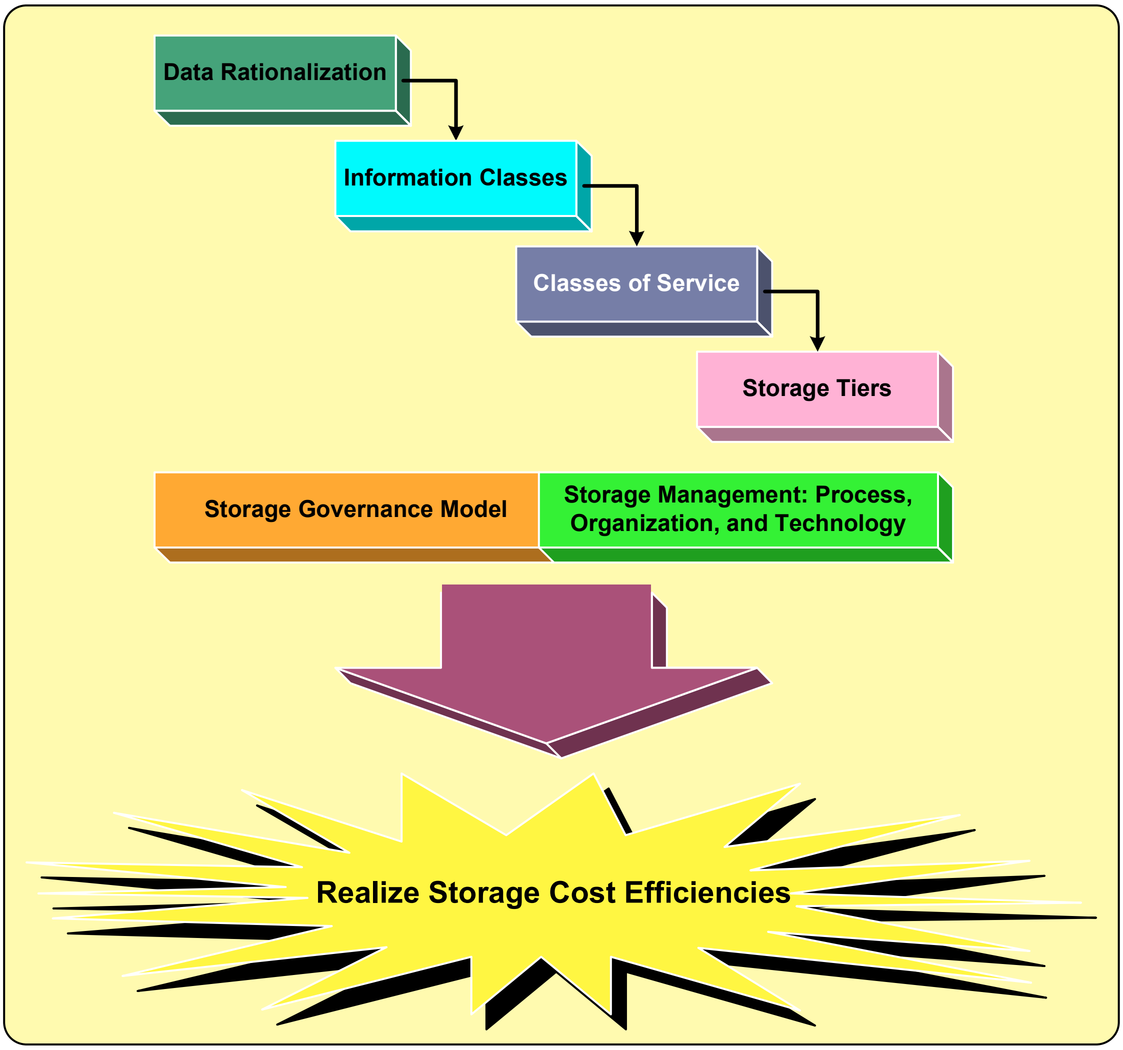
Information Lifecycle Management

At the highest level, clients describe Business Requirements that are used by a Business Process. These Applications that support these business requirements provide input into the definition of the policies used in the management of Information. This management of information describes the non-functional requirements, which are leveraged by the storage infrastructure layer to define Classes of Service and tiered storage architecture.

Information Value Management provides the abstraction and the transformation of business goals into business value for which ILM policies are implemented through Content Management Services and Information Management Services, within the Storage Infrastructure. Information Value Management provides closed loop feedback to the business regarding the costs, risks, and status.

The high level goal of implementing an ILM framework is to optimize the storage cost efficiency of the storage architecture.

IBM © 2005



• Figure 2 ILM Framework

Figure 2 teaches that data rationalization is the first step to realizing storage cost efficiencies. Once completed, the valid data is grouped into information classes that are used to build classes of services and ILM policies. Technology requirements are then grouped to form storage technology tiers. At the same time, storage governance and storage management practices are enhanced and continually matured to sustain the cost efficiencies for the long term.

Achieving storage cost efficiencies starts with gaining initial savings, then maximizing those savings and finally ensuring the elements are in place to sustain the savings. Examples are but not limited to:

- Activities for gaining initial savings:

  1. Reducing the amount of used storage as a result of initial clean-up

  2. Validate SAN requirements and reclaim used switches and switch ports

IBM © 2005

3. Validating data replication requirements in order to reclaim used storage space and offset future growth requirements

4. Developing and documenting information classification

5. Developing and documenting classes of service

6. Design and implement the tiered storage architecture

7. Migrating existing information to lower cost storage using a tiered storage architecture

- Activities for maximizing savings:

  1. Reconfiguring the current storage environment effectively improving the available to raw utilization

  2. Reclaiming available storage that has been over allocated

  3. Enhancing the information classification, classes of service and tiered storage architecture

- Activities for sustaining savings:

  1. Developing an ESA governance model

  2. Implementing changes to existing storage management processes like capacity planning and provisioning that effectively improve capacity utilization on an ongoing basis

Information Taxonomies

First, the “valid” data group must initially be classified objectively and empirically to form the information taxonomies. Information can be considered on the file level, file system level, data group level, application level, etc. The information type must be decided upon in advance. Initially, it is sufficient to look at the application or server level for assessment purposes. This level of granularity of classification is agreed upon in advance (e.g. application level) and then further refined if required. Information taxonomies are created by looking at numerous key information management characteristics that are scored considering empirical information about the in-scope environment. Examples of these characteristics are:

- Residency Cycle
- Backup Cycle
- Archival Cycle
- Performance Cycle
- Security
- Recoverability

IBM © 2005

- Availability
- Growth
- Criticality

Business Value and Information Classes

The information taxonomies in conjunction with the agreed upon business value is used to create several information classes. Since the business value of information is subjective, the business owner and enterprise storage infrastructure support must collaborate to define the business value. By adding the business value to the information taxonomy, the business becomes a “voting” stakeholder in the classification of their information. In order to prevent the subjective business value from becoming the singularly critical factor in classification of information, it is also assigned a weighting factor. The business is asked to provide feedback into the weighting of the business value vs. the key information management characteristics.

Classes of Service

The information classes combined with key managed service requirements are used to form classes of service. Initially, the classes or service are comprised only of the key information management characteristics. In order to form the classes of service, key managed service requirements are defined. Once defined, the key managed service requirements are added to the information classes to form classes of service. Classes of service provide that basis for the services portfolio of the storage architecture.

ILM Policies

The backup, archive, and performance cycles for information are used to determine the ILM policies for the management of that information. In actuality, some of this is already known in a storage environment since and ILM policy is combined from various policies including:

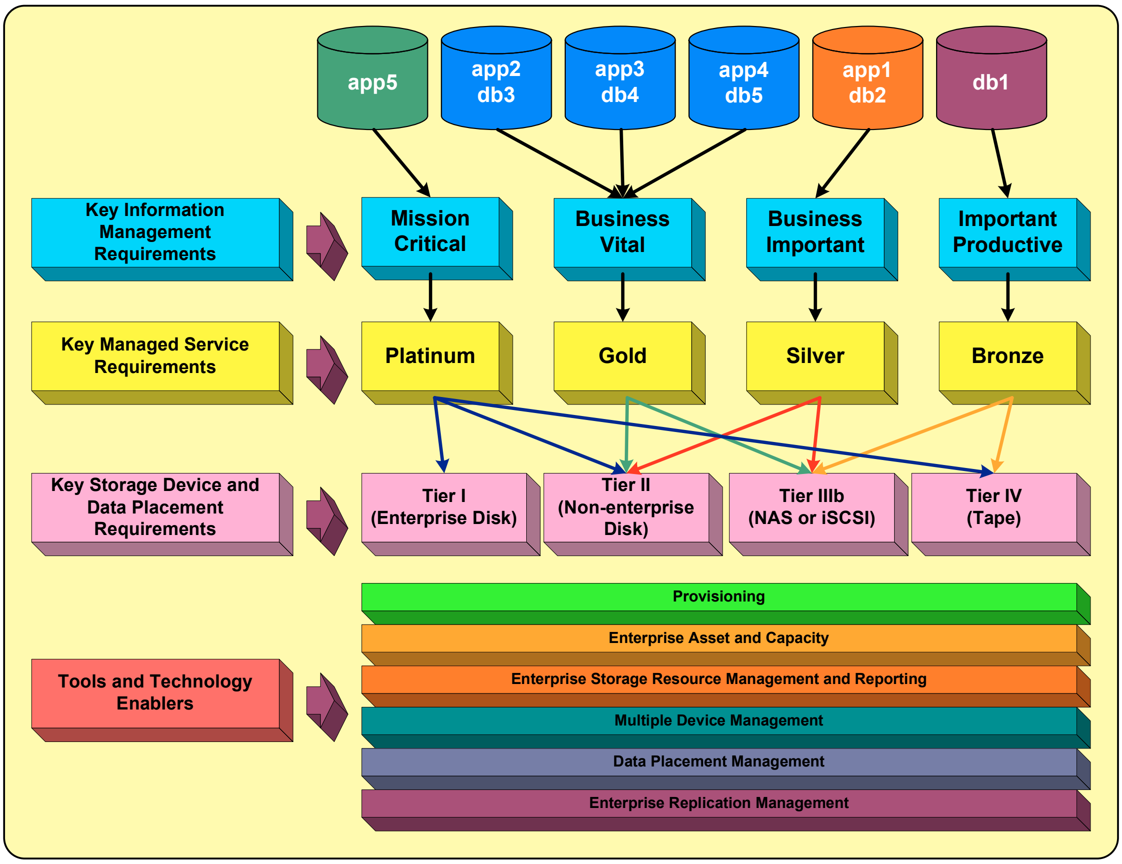
- Initial data placement policies
- Data lifecycle management policies
- Backup management policies
- Archive management policies
- Vaulting management policies
- Records/Content management policies

Tiered Storage Architecture

Class of service definitions combined with key storage device requirements help to define the storage tiers. Key storage device requirements are used to select the storage technologies of the storage architecture and create the storage design. Storage technologies are logically grouped into storage tiers. Initial used capacity and growth

IBM © 2005

requirement for each information type are aggregated to determine initial used capacity and growth requirement for each storage technology.



• Figure 3 Example Storage Architecture

Figure 3 shows an example of overall storage architecture with the information classes, classes of service, storage technologies, and supporting tools and technology enablers.

IBM © 2005

ILM Whitepaper V1.0

IBM Global Services

December 14, 2007

IBM CORPORATION, © 2005

© 2005 IBM Corporation

IBM Global Services

Prepared by:
Tim Waters – Senior Storage Consultant and Architect, IGS, (404)592-3394

Table of Contents

IBM © 2005

Table of Figures

IBM © 2005

Chapter

1

Data Rationalization

Data rationalization is the technique recommended by IBM to help uncover how much of a client's allocated and used storage capacity on a host-by-host basis is being utilized. This utilization, typically 50% industry wide according to Gartner, can be operated on in several ways. The two preferred methods to reclaim storage capacity are discussed in this chapter. Firstly, data cleanup is performed to remove, delete, archive, migrate, etc. files on affected servers where the value has dwindled over time. This actually lowers the effective utilization, perhaps to 40% on average. However, this offsets future growth as additional growth in data can reuse this recently freed up 10%.

Secondly, with 40% effective utilization from a host perspective, the storage capacity that was allocated to the hosts can be reclaimed or de-allocated which drives effective utilization up, perhaps to 80% at the host level. This de-allocated storage capacity can then be re-allocated to other hosts that are growing more rapidly. This free storage capacity offsets the need to purchase additional storage capacity and provides for cost avoidance in a business case scenario.

IBM TotalStorage Productivity Center for Data is the tool used to start the data rationalization process.

Common Agent Services

IBM TotalStorage Productivity Center for Data and IBM TotalStorage Productivity Center for Fabric run on a common agent services infrastructure. The infrastructure provides an agent manager for authentication, authorization of services and maintains a registry of configuration information about its common agents (managed computer) for each of the applications.

IBM TotalStorage Productivity Center for Data and IBM TotalStorage Productivity Center for Fabric will leverage the common agent services infrastructure. The common agent services registry database will reside on a local IBM DB2 database, located on the IBM TotalStorage Productivity Center for Data management server.

TPC for Data

IBM TotalStorage Productivity Center for Data is used to define storage-based policy management, automate file system extensions and report on database capacity. It is capable of reporting and managing various storage systems, including direct access storage, network attached storage and intelligent disk subsystems.

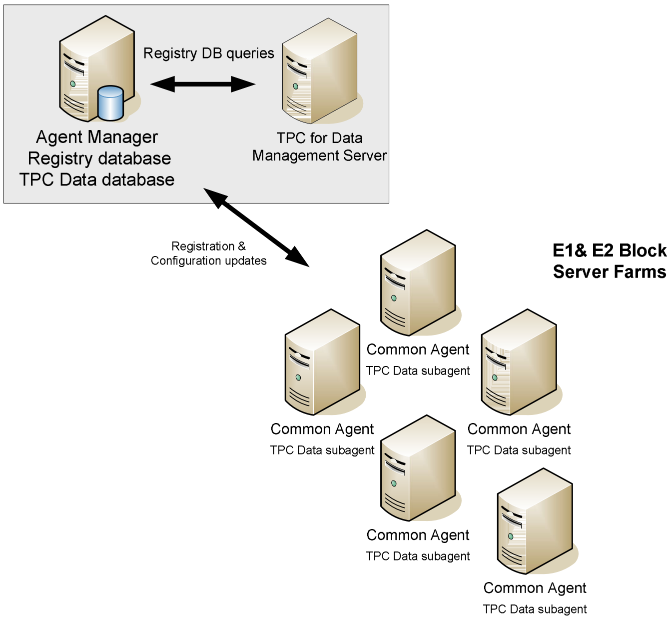
- The IBM TotalStorage Productivity Center for Data management server will be located in the Ford FMCC data center.

- The IBM TotalStorage Productivity Center for Data management server will reside on a dedicated server in the production environment. All agents will be

IBM © 2005

deployed in phased roll-out. Fifty (50) agents initially in test/development servers. Followed by two hundred a week until all are deployed.

Figure 1 shows common agent services supporting IBM TotalStorage Productivity Center for Data.



• Figure 1 Common Agent Services with IBM TotalStorage Productivity Center for Data

IBM TotalStorage Productivity Center for Data agents for Windows and UNIX will be deployed to all servers in the FMCC, ECC, and PDC data centers.

The IBM TotalStorage Productivity Center for Data database should be deployed to manage the following databases: SQL, Oracle and Sybase based on these criteria:

- Creating groups for databases will allow for easier monitoring.
- Defining a probe will assist in creating an inventory of files, logs and tables within the monitored database.
- Scans should be defined to gather extensive database statistics that will be leveraged for reporting.
- Alerts should be configured to forward SNMP traps to the CA Unicenter SNMP Manager console.
- Scan jobs maybe created to assist in monitoring database consumption and storage

The IBM TotalStorage Productivity Center for Data backend repository will reside on a local DB2 Enterprise Edition installation which can be the same database instance on which the common agent services registry resides.

IBM © 2005

Once the IBM TotalStorage Productivity Center for Data environment is deployed, probes will be configured and run to collect inventory, file system, database and storage subsystem information.

TPC for Data Sizing Recommendations

It is advisable not to manage more than one thousand five-hundred (1,500) IBM TotalStorage Productivity Center for Data agents per IBM TotalStorage Productivity Center for Data management server. If the Ford management environment grows to more than one thousand five-hundred (1,500) agents, a second IBM TotalStorage Productivity Center for Data management server will be required. The design described in this document allows for the possibility of adding additional IBM TotalStorage Productivity Center for Data management servers by leveraging the current common agent services infrastructure.

Host Agent Considerations

Agents should be deployed to all platforms types identified in the previous figure – Location and Machine Counts. This will provide Ford with a thorough data collection throughout the enterprise.

- Agent deployment could leverage the native Windows and Unix automated remote install functionality.

CIM Agent Considerations

IBM TotalStorage Productivity Center for Data uses CIM agents to gather and report on information from storage subsystem controllers. CIM agents will need to be deployed to support the full product roll-out.

Useful Reports

To support the Information Lifecycle Management (ILM) strategy at Ford, IBM TotalStorage Productivity Center for Data will be used as a tool for taking an inventory of enterprise data and storage assets. In support of this strategy, IBM will work with Ford to define and create customer reports to meet there reporting needs. In Addition, the following reports should also be used for Ford's information classification project:

- Perform an inventory of assets in the environment. This should be performed by running the following report. Reporting > Asset> System Wide > By Capacity. This report shows physical disk capacity as it relates to servers.

- Reports that are focused around file system utilization. For example: Reporting > Capacity > File system Capacity > By File system. This report shows the percent utilization of the file systems that is currently being monitored. This report could be sorted to only show file systems that have poor utilization. The report will assist in determining what might be a more appropriate capacity for the file system.

- Reports that are focus on file duplication. This report is useful for identifying files that are duplicate files. They may be wasting space or they are no longer needed. For example: Reporting > Capacity > Usage > Files > Duplicate Files > Network Wide.

IBM © 2005

- Reports on Obsolete and Orphan Files will also be useful. Generate reports that are focus on file duplication. This report is useful for identifying files that are duplicate files. They may be wasting space or they are no longer needed. For example: Reporting > Capacity > Usage > Files > Duplicate Files > Network Wide.

- Reports focused around file system activity. Reporting > Usage > Access Load > Access Time > By File system. This report will show a file system view and how much data has been accessed by time period. Ford should use this report to make decisions on which tier the file system belongs based on the frequency of data being accessed.

Information Types

An information type is defined by IBM to be the level of granularity of focus for the development of information classes. Ideally, every file in an I/T environment should be classified, but quantity limitations preclude that the level of focus in determining the information type should move up higher to the level of DB tables filesystems, volumes, database containers, logical disks, physical disks, servers, and applications. It is therefore sufficient to classify information using a server view for the level of granularity. Thus, servers (hosts) will be classified at Ford into information classes comprised logical groupings of functional and non-functional characteristics for valid data.

In-valid Data Cleanup

Since it is only the valid data that is classified, the in-valid data must be operated upon in a sufficient manner to free up valuable storage capacity and offset future data growth. This is accomplished by using TPC for Data to gather empirical information about an application or server and analyze this information. During analysis, certain thresholds are chosen to determine if data is stale, orphan, and duplicate. Once those thresholds are chosen, further analysis on the files and databases themselves will reveal chances to reclaim this used storage capacity by deleting or archiving these files and databases. After cleanup, the end result from a cost perspective is overall lower filesystem level or volume level utilization, but storage capacity to offset future growth and identifiable opportunities to reclaim unused storage capacity.

IBM © 2005

Chapter

2

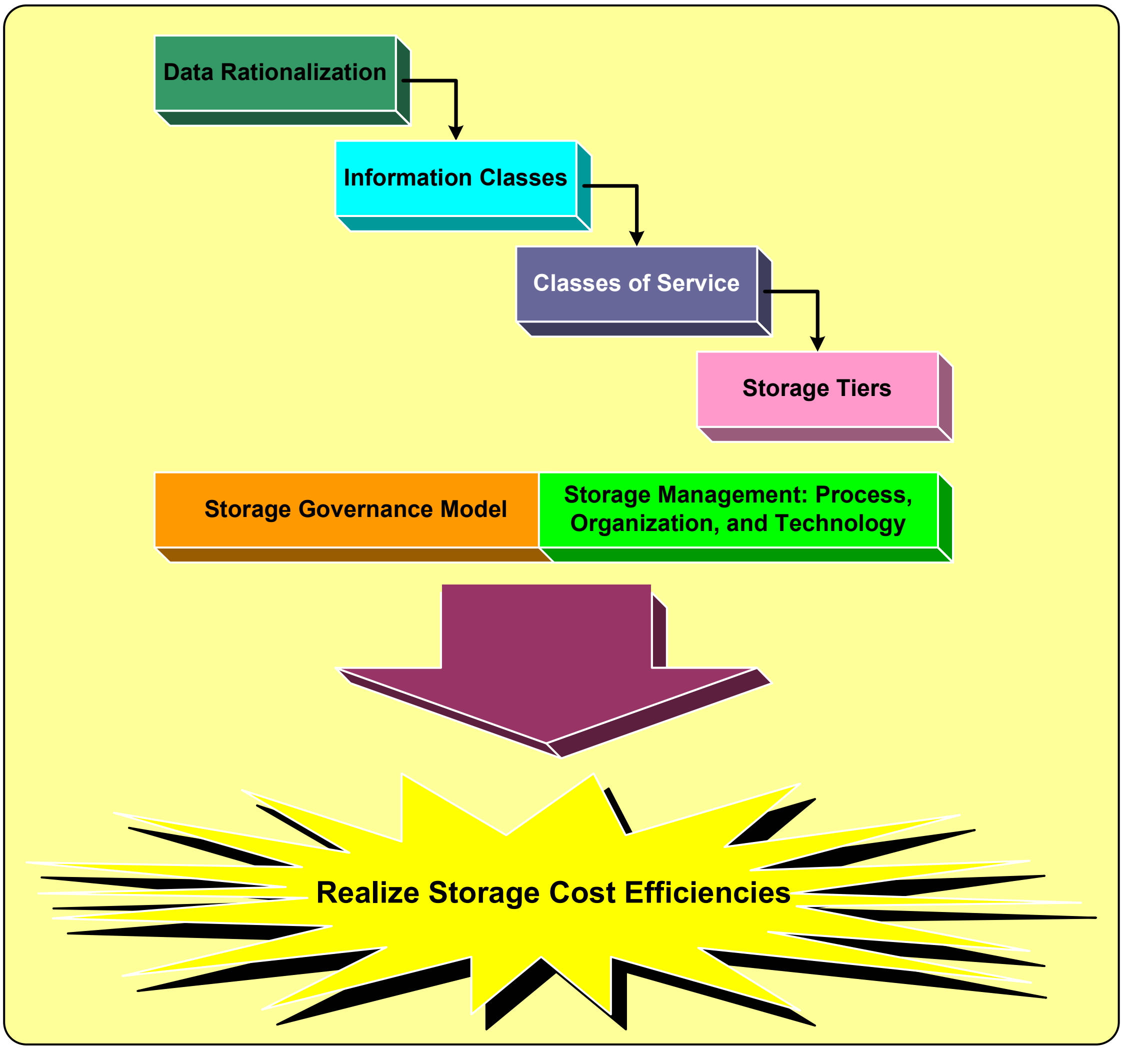
Information Lifecycle Management

At the highest level, clients describe Business Requirements that are used by a Business Process. These Applications that support these business requirements provide input into the definition of the policies used in the management of Information. This management of information describes the non-functional requirements, which are leveraged by the storage infrastructure layer to define Classes of Service and tiered storage architecture.

Information Value Management provides the abstraction and the transformation of business goals into business value for which ILM policies are implemented through Content Management Services and Information Management Services, within the Storage Infrastructure. Information Value Management provides closed loop feedback to the business regarding the costs, risks, and status.

The high level goal of implementing an ILM framework is to optimize the storage cost efficiency of the storage architecture.

IBM © 2005



• Figure 2 ILM Framework

Figure 2 teaches that data rationalization is the first step to realizing storage cost efficiencies. Once completed, the valid data is grouped into information classes that are used to build classes of services and ILM policies. Technology requirements are then grouped to form storage technology tiers. At the same time, storage governance and storage management practices are enhanced and continually matured to sustain the cost efficiencies for the long term.

Achieving storage cost efficiencies starts with gaining initial savings, then maximizing those savings and finally ensuring the elements are in place to sustain the savings. Examples are but not limited to:

- Activities for gaining initial savings:

  1. Reducing the amount of used storage as a result of initial clean-up

  2. Validate SAN requirements and reclaim used switches and switch ports

IBM © 2005

3. Validating data replication requirements in order to reclaim used storage space and offset future growth requirements

4. Developing and documenting information classification

5. Developing and documenting classes of service

6. Design and implement the tiered storage architecture

7. Migrating existing information to lower cost storage using a tiered storage architecture

- Activities for maximizing savings:

1. Reconfiguring the current storage environment effectively improving the available to raw utilization

2. Reclaiming available storage that has been over allocated

3. Enhancing the information classification, classes of service and tiered storage architecture

- Activities for sustaining savings:

1. Developing an ESA governance model

2. Implementing changes to existing storage management processes like capacity planning and provisioning that effectively improve capacity utilization on an ongoing basis

Information Taxonomies

First, the "valid" data group must initially be classified objectively and empirically to form the information taxonomies. Information can be considered on the file level, file system level, data group level, application level, etc. The information type must be decided upon in advance. Initially, it is sufficient to look at the application or server level for assessment purposes. This level of granularity of classification is agreed upon in advance (e.g. application level) and then further refined if required. Information taxonomies are created by looking at numerous key information management characteristics that are scored considering empirical information about the in-scope environment. Examples of these characteristics are:

- Residency Cycle
- Backup Cycle
- Archival Cycle
- Performance Cycle
- Security
- Recoverability

IBM © 2005

- Availability
- Growth
- Criticality

Business Value and Information Classes

The information taxonomies in conjunction with the agreed upon business value is used to create several information classes. Since the business value of information is subjective, the business owner and enterprise storage infrastructure support must collaborate to define the business value. By adding the business value to the information taxonomy, the business becomes a "voting" stakeholder in the classification of their information. In order to prevent the subjective business value from becoming the singularly critical factor in classification of information, it is also assigned a weighting factor. The business is asked to provide feedback into the weighting of the business value vs. the key information management characteristics.

Classes of Service

The information classes combined with key managed service requirements are used to form classes of service. Initially, the classes or service are comprised only of the key information management characteristics. In order to form the classes of service, key managed service requirements are defined. Once defined, the key managed service requirements are added to the information classes to form classes of service. Classes of service provide that basis for the services portfolio of the storage architecture.

ILM Policies

The backup, archive, and performance cycles for information are used to determine the ILM policies for the management of that information. In actuality, some of this is already known in a storage environment since and ILM policy is combined from various policies including:

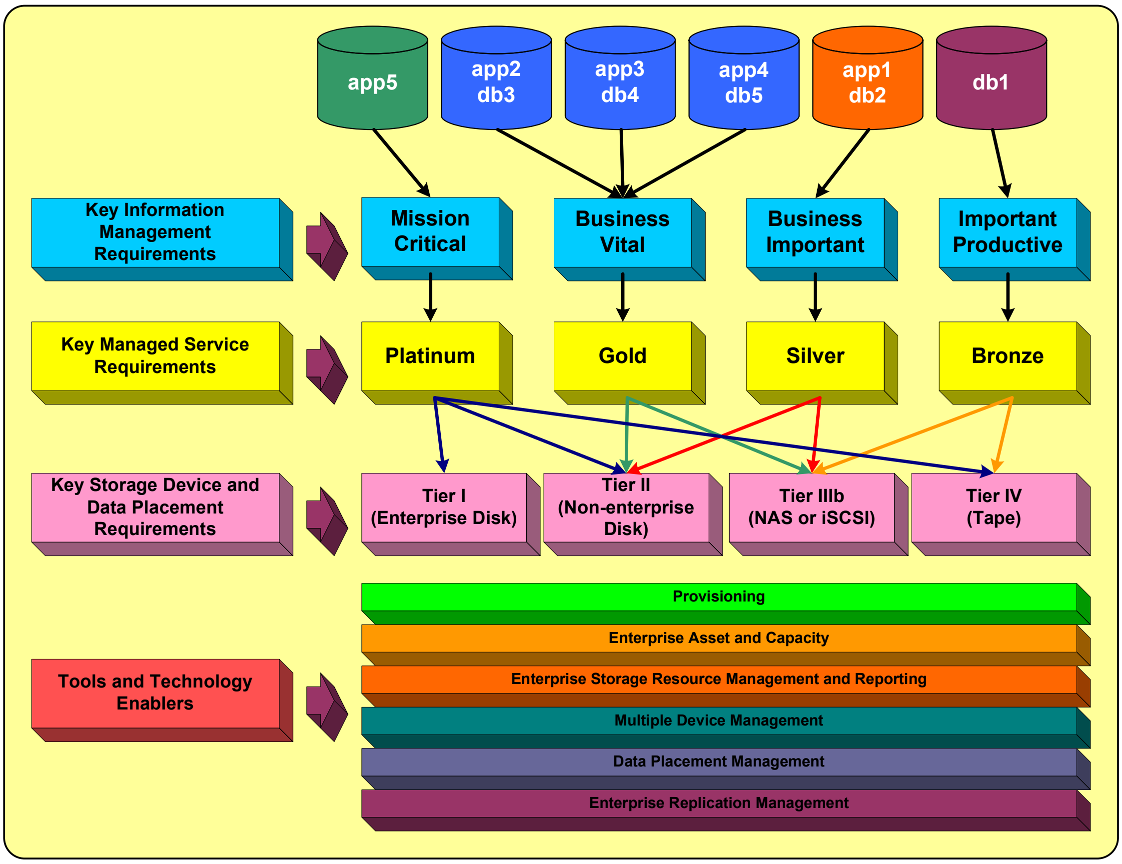
- Initial data placement policies
- Data lifecycle management policies
- Backup management policies
- Archive management policies
- Vaulting management policies
- Records/Content management policies

Tiered Storage Architecture

Class of service definitions combined with key storage device requirements help to define the storage tiers. Key storage device requirements are used to select the storage technologies of the storage architecture and create the storage design. Storage technologies are logically grouped into storage tiers. Initial used capacity and growth

IBM © 2005

requirement for each information type are aggregated to determine initial used capacity and growth requirement for each storage technology.



• Figure 3 Example Storage Architecture

Figure 3 shows an example of overall storage architecture with the information classes, classes of service, storage technologies, and supporting tools and technology enablers.

IBM © 2005