Information Lifecycle Management for Business Data

An Oracle White Paper
June 2007

ORACLE

Information Lifecycle Management for Business Data

INTRODUCTION

Although most organizations have long regarded their stores of data as one of their most valuable corporate assets, how this data was managed and maintained varies enormously. Originally, data was used to help achieve operational goals, run the business and help identify the future direction and success of the company. However, new government regulations and guidelines are a key driving force in how and why data is being retained, as they are now requiring organizations to retain and control information for very long periods of time. So today there is two additional objectives IT managers are trying to satisfy: to store vast quantities of data, for the lowest possible cost; and to meet the new regulatory requirements for data retention and protection.

Regulatory Requirements

Previously organizations retained data because they wanted to, today, many organizations have to retain specific data for specified periods of time. Failure to comply with these regulations could result in organizations having to pay very heavy fine. Therefore, around the world, a number of regulatory requirements, such as Sarbanes-Oxley, HIPAA, DOD5015.2-STD in the US and the European Data Privacy Directive in the European Union are changing how organizations manage their data. These regulations specify what data must be retained, whether it can be changed, and for how long it must be retained, which could be for a period of 30 years or even longer.

These regulations frequently demand that electronic data is secure from unauthorized access and changes, and there is an audit trail of all changes to data and by whom. The Oracle Database 11*g* has already shown that it can retain huge quantities of data without impacting application performance. It also contains the features required to restrict access and prevent unauthorized changes to data. Therefore a user can be authorized to insert data, but can never change it. When data does have to be deleted, then it would be performed by a special job, which was given privileges just for that task. Oracle also provides cryptographic functions that can be used to demonstrate that a highly privileged user has not intentionally modified data and Oracle Audit vault automates the collection and analysis of audit data from multiple systems thus creating the industry's most secure and scalable audit warehouse.

What is ILM?

Information today comes in a wide variety of types, for example it could be an email message, a photograph or an order in an Online Transaction Processing System. Therefore, once you know the type of data and how it will be used, you already have an understanding of what its evolution and final destiny is likely to be.

The challenge now before all organizations, is to understand how their data evolves, determine how it grows, monitor how its usage change over time, and decide how long it should survive. Whilst adhering to all the rules and regulations that now apply to that data.

Information Lifecycle Management (ILM) is designed to address these issues, with a combination of processes, policies, software and hardware so that the appropriate technology can be used for each phase of the lifecycle of the data.

THE LIFECYCLE OF DATA

An analysis of your data will most likely reveal, that initially it is accessed and maybe updated on a very frequent basis. As the age of the data increases, its access frequency diminishes to almost negligible, if any. Therefore, most organizations find themselves in the situation where most of their users, are accessing all of the current data, and very few users, are accessing, all of the other data. Thus, data can be described as being; active, less active, historical or ready to be archived.

Data Lifecycle

With so much data being held, data, during its lifetime, the data will be moved to different physical locations. This is because depending on where it is in its lifecycle; it needs to be located on the most appropriate storage device.

WHY ORACLE DATABASE 11G FOR BUSINESS ILM

The Oracle Database 11*g* is capable of storing many different types of data, and storing all of your data in an Oracle database, means that it is much easier to manage, because the data is all in one place, instead of being stored using many different formats. Thus the Oracle database is the ideal platform to implement an Information Lifecycle Management policy, because it has a number of features, which makes it very easy to implement an ILM solution:

- Application Transparency; data classification is transparent
- Fine-grained; managing data at individual row level
- Low-Cost; uses low cost to reduce costs
- Enforceable Compliance Policies; define and enforce policies

Application Transparency is very important in ILM, because it means that there is no need to customize applications and it also allows various changes to be made to the data without any impact on the applications that are using that data. Therefore, data can easily be moved at the different stages of its lifecycle and access to the data can be optimized via the database. Another important benefit is that application transparency offers the flexibility required to quickly adapt to any new regulatory requirements, again without any impact on the existing applications.

Oracle is able to view the data at a very *fine-grained* level, because it can manage it as individual rows, and it also has the ability to group all related data together, where storage devices only see bytes and blocks.

With so much data to retain, using *low cost* storage is a key factor in implementing ILM. Since Oracle can take advantage of all the different types of storage devices, because it is hardware independent, the maximum amount of data can be held for the lowest possible cost.

When information is kept for *compliance* reasons, it is imperative to show to the regulatory bodies that data is being retained and managed in accordance with the regulations. Within Oracle it is possible to define security and audit policies, which enforce and log all access to data.

Oracle Database 11*g* Manages All Types of Data

Information Lifecycle Management is concerned with all data that is held within an organization. This means not just structured data, such as orders in an OLTP system or a history of sales in a data warehouse, but it is also concerned with unstructured data such as email, documents and images.

Oracle Content Database, which is one of the options of the Oracle Database Enterprise Edition, provides the capability to store unstructured data such as images and documents. It includes role based security to ensure content is only accessed by authorized personnel and policies which describe what happens to the content during its lifetime.

Therefore if all the information your organization cares about, is held inside an Oracle Database, you can take advantage of the features and functionality provided by the database, to manage and move the data as it evolves during its lifetime, without having to worry about managing multiple types of data stores.

IMPLEMENTING ILM IN 4 EASY STEPS

Building an Information Management Lifecycle solution, using Oracle Database 11*g* is quite straightforward, and it can be achieved by following these four simple steps.

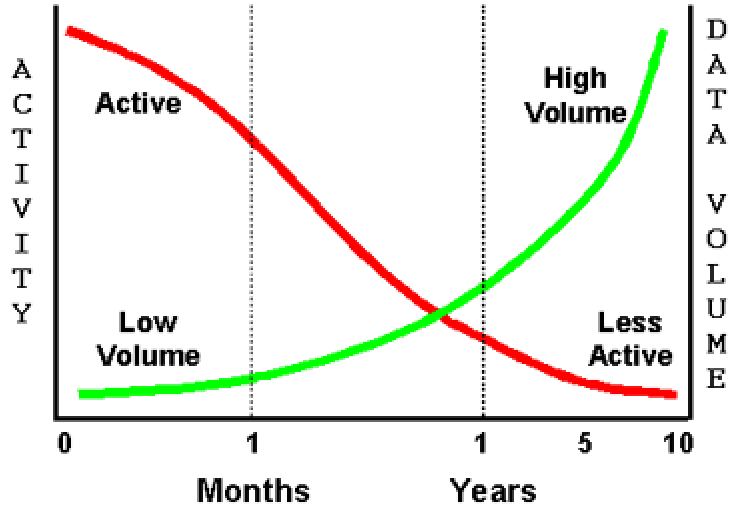
Step 1 - Define the Data Classes

In order to make effective use of Information Lifecycle Management, the first step is to look at all the data in your organization, what type of data is it, where is it stored? and determine:

- which data is important, where is it and what needs to be retained
- how this data flows within the organization
- what happens to this data over time and is it still needed
- the degree of data availability and protection that is needed
- data retention for legal and business requirements

Once there is an understanding of how the data will be used, it can then be classified on this basis. The most common type of classification is by age or date, but other types are possible, such as by product or privacy, or a hybrid classification could be used such as by privacy and age. Once the data has been classified, the policies that will be applied to that data, will depend upon its class.

In order to treat the data classes differently, the data needs to be physically separated. When information is first created it is often frequently accessed, but then over time it may be referenced very infrequently. For instance, when a customer places an order, they regularly look at it to see its status and that it has been shipped. But once it arrives, they may never reference that order again. This order would also be included in regular reports that are run to see what goods are being ordered, but, over time, it would not figure in any of the reports and may only be referenced in the future if someone does a detailed analysis that involves this data. Therefore, orders could be classified by the Financial Quarters; Q1, Q2, Q3, Q4 and as Historical Orders.



The advantage of using this approach, is that when the data is grouped at the row level by their class, which in this example would be the date of the order, all orders for Q1 can be managed as a self contained unit, where as the orders for Q2 would reside in a different class. In Oracle this can be achieved by using partitioning, and since partitions are completely transparent to the application, the data is physically separated, but the application still sees all of the orders.

Step 2 – Create Storage Tiers for the Data Classes

Since Oracle Database 11*g*, can take advantage of all the different storage options that are available, the next step is to establish the following storage tiers:

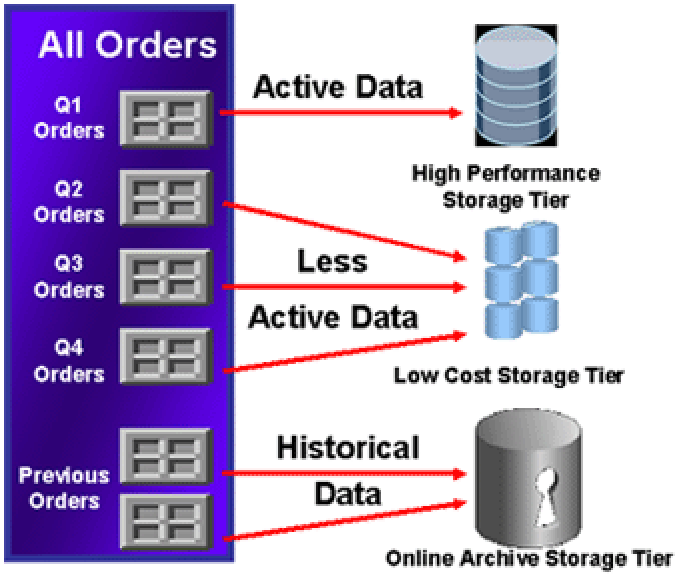
- High Performance
- Low Cost
- Online Archive
- Offline Archive (optional)

The *high performance* storage tier is where all the important and frequently accessed data would be stored, such as the partition holding our Q1 orders. This would utilize the smaller, faster disks on high performance storage devices.

The *low cost* storage tier is where the less frequently accessed data is stored, such as the partitions holding the orders for Q2, Q3 and Q4. This tier would be built using large capacity disks, such as those found in modular storage arrays or the low costs ATA disks, which offer the maximum storage inexpensively.

The *online archive* storage tier is where all the data that is never or hardly accessed would be stored. It is likely to be extremely large and to store the maximum quantity of data on the online archive storage tier, various techniques can be used to compress the data. This tier could be located in the database or it could be in another database, which serves as a central archive database for all information within the enterprise.

Stored on low cost storage devices like the ATA drives, the data would still be online and available, for a cost that is only slightly higher than storing this information on tape, without all the disadvantages that come with archiving data to tape. If the Online Archive storage tier is identified as read-only, then it would be impossible to change the data and database backups would not be required after it is backed up the first time.



The o*ffline archive* storage tier is optional, because it is only used when there is a requirement to remove data from the database and store it in some other format such as XML on a tape.

The Costs Savings of using Tiered Storage

One of the benefits of implementing an ILM strategy is the cost savings that can result from using multiple tiered storage. Let's assume that we have the following data to store; 50gb on High Performance, 500gb on Low Cost and 2Tb on the Online Archive, and that the cost per gb is $72 on the High Performance tier, $14 on the Low Cost and $7 on the Online Archive tier.

The table below illustrates the possible cost savings using tiered storage, rather than storing all data on one class of storage. As you can see, the costs savings can be quite significant and if the data is suitable for database compression, then even further cost savings are possible.

Storage Tier	Single Tier using High Performance Disks	Multiple Storage Tiers	Multiple Tiers with Database Compression
High Performance (50gb)	$3,600	$3,600	$3,600
Low Cost (500gb)	$36,000	$7,000	$7,000
Online Archive (2Tb)	$144,000	$14,000	$8,500
	$180,300	**$24,600**	**$19,100**

Assigning Classes to Storage Tiers

Once the storage tiers have been defined, the data classes (partitions) identified in step one can be assigned to the appropriate storage tiers. This provides an easy way to distribute the data across the appropriate storage devices depending on its usage, whilst still keeping the data online and readable available and stored on the most cost-effective device. Since partitioning is completely transparent to anyone accessing the data, no applications changes are required, thus it can be implemented at any time at minimal cost.

Step 3 – Create Data Access and Migration Policies

The next step is to specify who can access the data and the operations they may perform and how to move the data during its lifetime.

Managing Access to Data

Regulatory requirements are beginning to place exacting demands on how data can be accessed. Using Oracle Database 11g, extremely effective methods for controlling what authorized users of the database may see is to use database *views* or implement security policies using *virtual private database*.

A *security policy* implemented via *virtual private database* determines exactly which data can be seen, therefore authorized users could see the information for Q1, Q2, Q3 and Q4, but only special users could view the historical data. Using this approach, the data is still available to those who need access to it, but for the vast of majority of users, it is now invisible and therefore is not included or accessed by any of their queries.

A security policy is defined at the database level and it is transparently applied to all database users. The benefit of this approach is that it provides a secure and controlled environment for accessing the data, which cannot be overridden and can be implemented without requiring any application changes. In addition read-only tablespaces can be defined which ensures that the data cannot change.

Oracle Database Vault can prevent highly privileged users, including powerful application DBAs and others, from accessing sensitive applications and data in Oracle databases outside their authorized responsibilities. It can also protect existing applications quickly and easily and requires no changes to your applications.

Migrate Data between Classes

During the lifecycle of the data it will be necessary to move it at various times and this occurs for a variety of reasons, such as:

- for performance, only a limited number of orders are held on the high performance disks
- data is no longer frequently accessed and is using valuable high performance storage and needs to be moved to a low-cost storage device
- legal requirements demand that the information is always available for a given period of time, and it needs to be held safely at the lowest possible cost

Within Oracle Database 11*g*, there are a number of ways that data can be physically moved within the database to take advantage of the different storage tiers. If the data is partitioned then, the partition containing the orders for Q2, can be moved from the high performance storage tier to the low cost storage tier online. Since the data is being moved within the database, it can be physically moved, without

affecting the applications that require it, or cause disruption to regular users as the partition can be moved online.

Sometimes individual data items must be moved rather than a group of data. For example, suppose data was classified according to a level of privacy, and a report, which was once secret, is now to be made available to the public. If the classification changed from secret to public, and the data was partitioned on its privacy classification, the row would automatically move to the partition containing public data.

Whenever data is moved from its original source, then it is very important to ensure that the process selected adheres to any regulatory requirements, such as, the data cannot be altered, is secure from unauthorized access, easily readable and stored in an approved location.

Regulatory Compliance

The new regulatory requirements such as Sarbanes-Oxley, HIPAA, DOD5015.2-STD in the US and the European Data Privacy Directive in the European Union are playing a key role in the long-term retention of data because they are imposing strict rules on how data is held. Now organizations have to protect against unauthorized changes and possibly show details of every change ever made to a record.

Oracle Database 11*g*, already contains a number of features which will enable an organization to comply with the new regulations, and these features are described in the Oracle white paper, Applying Oracle Security Technologies for Regulatory Compliance. The white paper, Best Practices for California SB1386, illustrates how the features in the Oracle Database can be used to comply with a specific regulation and further information is availab0le in the white paper, Privacy Protection in Oracle Database 10g.

Step 4 – Define and Enforce Compliance Policies

The fourth step in defining our ILM environment is the creation of compliance policies, which when data is decentralized and fragmented, have to be defined and enforced in every data location, which could easily result in a compliance policy being overlooked. However, using the Oracle Database to provide a central location for storing data, it is then very easy to enforce compliance policies as they are all managed and enforced from one central location. When defining compliance policies there are five areas to consider:

- Retention
- Immutability
- Privacy
- Auditing
- Expiration

The *retention* policy will describe how the data is to be retained, for how long it must be kept and what happens at life end. Therefore, an example of a retention policy is that a record must be stored in its original form, no modifications are allowed, it must be kept for 7 years and then it may be deleted. Using Oracle security it is possible to ensure that data remains unchanged and only using authorized processes, can data be removed at the appropriate time.

Immutability is concerned with proving to an external party that data is complete and has not been modified. Cryptographic signatures can be created and held either inside or outside of the database, to show that data has not been altered or tampered in any way.

With so much data being retained today, it is extremely important to maintain *privacy* of data at all times and the Oracle database provides several ways to ensure data privacy. Access to data can be strictly controlled using security policies defined using Virtual Private Database (VPD), which define exactly which information a user may see. Maintained at the database level, these policies cannot be violated by anyone. In addition from Oracle Database 10*g* Release 2, individual columns can also be encrypted, and from Oracle Database 11g at the tablespace level, so that anyone looking at the raw data cannot see its contents.

The Oracle Database also has its own *auditing* capability to track all access and changes to data. These can be defined at the table level or via fine-grained auditing which specifies the criteria for when an audit record should be generated, such as, someone tried to change a salary or attempted to alter a processed order.

Oracle Database Vault ensures that data, is only ever accessed by authorized personnel, and Oracle Audit Vault provides an audit warehouse of all activity upon the database.

Ultimately, data may *expire* for business or regulatory reasons and its need to be removed from the database. Since that can involve removing vast quantities, such as all the orders for 1999, the Oracle database can remove data very quickly and efficiently by simply dropping the partition, which contains the information identified for removal.

Oracle ILM Assistant

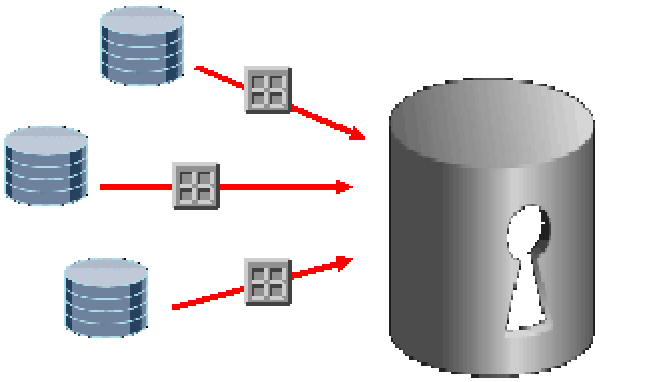
The Oracle ILM Assistant is a GUI based tool for managing your ILM environment. It provides the ability to create lifecycle definitions, which are assigned to tables in the database. Then based on the lifecycle policy, the ILM Assistant advises when it is time to move, archive or delete data. It will also illustrate the storage requirements and cost savings associated with moving the data. Other capabilities include the ability to show how to partition a table based on a lifecycle definition and to simulate the events on a table if it were partitioned.

To assist with managing compliance issues, the ILM Assistant will show all VPD and FGA policies that have been defined. It can also display and query all the audit records and generate and compare digital signatures.

The Oracle ILM Assistant requires Oracle Database 9i or greater and that Oracle Application Express must be installed in the database where the tables to be managed by the ILM Assistant reside. Initially, only tables partitioned on a date are eligible and the ILM Assistant does not make any physical changes to the database, instead it generates scripts so that you can perform the tasks when you are ready.

The Benefits of an Online Archive

There usually comes a point during the lifecycle of the data, when it is no longer being regularly accessed and is considered eligible for archiving. Traditionally, at this time, the data would have been removed from the database and stored on tape, because it is very cheap and is capable of storing vast quantities of information for a very low cost. Today it is no longer necessary to archive that data to tape, instead it can remain in the database, or transferred to a central online archive database. All this information would be stored using low-cost storage devices whose cost per gigabyte is very close to that of tape.



An old school of thought was that data needed to be archived out of the database in order to maintain the performance of the database. This is in fact not true. Databases can perform extremely well with huge quantities of data. In fact, many customers in the financial, web and telco industries have extremely large quantities of data (see the Winter Corporation 2005 TopTen Program Survey at http://www.wintercorp.com/VLDB/2005_TopTen_Survey/TopTenWinners_2005.asp) and yet the database performance is excellent. The only case where database performance is significantly impacted by having large quantities of data is table scans. However, partitioning, views, and virtual private database can be used to hide historical data, thereby eliminating any performance penalty caused by scanning the extra data retained in the database.

There are a number of benefits of keeping all of the data in a database for archival. The most important is that the data will always be instantly available, therefore time is not wasted trying to locate the tapes where the data was archived to and worrying

as to whether the tape is readable and still in a format that can be loaded into the database.

If the data has been archived for many years, then development time also may be needed to write a program to reload the data into the database from the tape archive. This could prove expensive, and time-consuming, especially if the data is extremely old. Therefore, if the data is retained in the database, then this is not a problem, because it is already online, and in the latest database format.

Holding all the historical data in the database no longer impacts the time required to backup the database and the size of the backup. When RMAN is used to backup the database, it will only include in the backup, the data that has changed. Since the historical data is no longer changing, once the data has been backed, it will not be backed up again. Therefore the backup time and storage requirements are the same when compared to as if the data had been archived from the database. i.e. the data is written to tape once and the data is stored on one set of tapes in both cases

Another important factor to consider is how the data is to be physically removed from the database, especially if it is to be transferred from a production system to a central database archive. Oracle provides the capability to move this data rapidly between databases by using transportable tablespaces, which moves the data as a complete unit.

Historically, there has been reluctance in some organizations to remove data from the database. However, attitudes are now changing, because data is now being requested in legal matters and the cost of e-discovery to supply that data can be quite expensive. Hence, if there is no legal requirement to retain the data, why keep it when it is no longer required. Therefore, when it is time to remove that data from the database, the fastest way, is to remove a set of data from the database. This is achieved by keeping the data in its own partition and then the partitioned can be dropped, which is a very fast operation. However, if this approach is not possible and a conventional SQL delete statement is used, one should not under-estimate the time this will take.

If there is a requirement to remove data from the database and there is some possibility that it may have to be returned to the database at some time in the future. Considering removing the data in a database format such as a transportable tablespace, or use the XML capability built-in to the Oracle Database, to extract the information in an open format.

Consider an online archive of your data into an Oracle Database. The reasons are clear.

- The cost of disk is approaching that of tape, eliminate the time to find the tape that contains the data
- Data remains online when needed, faster access to meet business requirements

- Use the current application to access the data, no need to waste resources to build a new tool

CONCLUSION

Oracle Database 11g provides an ideal platform for implementing ILM because it's simple to use since there are no specialized data stores to manage and it operates independent of any hardware. It has proven fast performance, therefore all information can be quickly retrieved. The security features in the database ensure that data is secure from unauthorized access and data is always transactionally consistent.

The Oracle Database provides total flexibility, therefore it can rapidly adapt to any change in requirements, which is extremely important due to the continuing evolution of the emerging regulations. With the ILM Assistant it is possible to define what should happen to data during its lifetime and be reminded when it is time to perform those actions. Finally, the Oracle Database has already been around for over 20 years, therefore you know that Oracle will be supported for many years, which is very reassuring when your data may have to be retained for 30 years or more.

Information Lifecycle Management enables us to understand our data, which is an extremely valuable business asset, which must be managed properly, to ensure business success and regulatory compliance. Using Oracle Database 11g, a comprehensive ILM solution can be implemented for the lowest possible cost.

Information LifeCycle Management for Business Data
June 2007
Author: Lilian Hobbs
Contributing Authors:

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200
www.oracle.com