



Data Life Cycle Services

Archiving and Storage

Audience

Discovery Informatics IT management and business sponsors: with an interest in managing the data life cycle of biological results.

Abstract

This brochure outlines our service offerings for effective data life cycle management:

- Archival: Data archiving strategies
- Warehousing: Results Marts
- Partitioning: Effective Data Partitioning

An effective data life cycle management strategy can improve operational efficiency and increase productivity.

We can help...

Archiving and Storage

→ Introduction

An operational database deployed used across multiple sites for high throughput screening can quickly become large and raise performance and management challenges. Operations such as recalculation of database statistics, import and export of data can become difficult to achieve within non-operational time slices. Querying of the database can become adversely affected by the data volumes even when most of the data is not relevant to the search results. Common approaches to this problem often include development of an in-house mart or warehouse for results data, optimised for query performance rather than operational efficiency. This extracts and transforms the data from the OLTP operational data model into a query optimised OLAP warehouse model, improving query performance. However, whilst this addresses query performance issues it does not address problems associated with the operational data.

This document presents a number of service offerings which improve the long term storage and maintenance of data from leading data management solutions. Our approach focuses of the archiving of results data and the creation of a separate results star schema as a long term query mart for results data.

We are firm believers in a structured approach to any project and use a consultative approach to our services. The success of any project starts with a clear understanding of the requirements. Our business analysis services are designed to gather a clear understanding of the key user goals from any solution. Using a combination of user interviews and use-case capture and documentation using UML our approach captures not only the key requirements but also an understanding of the purposes and users goals the solution must fulfil.

→ Data Archiving

We skilled in development data archiving tools, customised to your needs, to move all related data for a set of tests into a separate set of mirror tables, ready for archiving/restoration using a standard Oracle dump utilities. This can include

- Study
- Dictionaries
- Protocol
- Test

This provides a mechanism of archiving and removing data from your operational database without recourse to logical deletion and purging using standard mechanisms which do not address archival of data.

Our archive tools move the data to separate schema for export, then deletes it from the live database schema. The archival area can then be saved to an external format. This provides better data security. Results are saved to a separate location where the standard Oracle export utilities can be used for archival.

The Archiving tools will include:

- **Archive Selection:** We will develop to your requirements an archival selection tool that allows administrators to select data for archival, generating a set of archive tasks.
- **Archive Export:** A tool will be developed to support execution of archiving tasks. A number of parameters are passed to allow selection of records for archival. All results data associated with a test, study or protocol including XL templates, are archived during this procedure.
- **Data Purge:** Archived data is automatically marked as logically deleted. This utility purges this data from the operational database.
- **Archive import:** This tool supports the import of an archived data set.

Work Plan

Example work plan for the archive tool project:

Operation	Man/Days	Location
Initial requirements analysis	2	Onsite
Design	3	Offsite
Creation of Selector tool	10	Offsite
Creation of export Tool	10	Offsite
Creation of import Tool	10	Offsite
Documentation	5	Offsite
Testing	20	75/25 off/on site
Rollout	5	Online
Total	65	

This delivers a user driven tool which flags and exports study level blocks of data out of the database in a controlled manner. It also supports import of archived data into the database.

Advantages

- Simple to understand
- User driven
- Does not require enterprise license for Oracle

→ Results Mart

An effective archival strategy provides a method of removing data from the operational database improving performance and throughput. Part of the data life cycle however is to ensure that the important biological results remain online data and can be effectively accessed to facilitate an understanding of biological activity. To address the query performance and long term access to results data, a results mart will be developed to provide a second copy of results data transformed into a star schema optimised for query performance. This can be implemented as a simple data mart or part of a larger data consolidation and warehousing project.

This solution is based on more than 10 years experience of deploying leading data management solutions and the design and development of several results marts. It consists of a set of views onto the source database schema and triggers on a selection of high level tables identifying when a test has changed. This provides a logical and consistent view of the data for transfer into the results mart. Data consistency problems such as units of measure can be corrected by a set of data transformation procedures used to mode the data. Results are then presented in a result centric mart for storage and querying.

Part of our approach includes analysis of the key requirements for the solution allowing us to adjust our results mart design to accommodate specific requirements. Our standard results mart consists of:

Central Fact tables for Numeric Results and Descriptive Results.

Dimensions for Plates, Sample, Compound, Result Type, Result Context (linked conditions), Study, Protocol, and Time

Work Plan

Example work plan:

Operation	Man/Days	Location
Initial requirements analysis	5	Onsite
Design	10	offsite
Source Scripts	5	offsite
Mart Scripts	5	offsite
Transformation Scripts	20	offsite
Documentation	10	offsite
Testing	20	75/25 off/on site
Rollout	5	online
Total	80	

Advantages

- Moves data from a propriety schema to an easy to query star schema
- Allow conversion of numeric data to common base dimensions
- Provides a method to consolidate of multiple schema

➔ Effective data partitioning

The standard Oracle approach to handling very large data sets is to physically divide data into smaller units which are easier to manage. These partitions must follow the logical structure of the data so that most operations are scoped to only affect a single partition at a time.

Partitioning

For most data models this involves partitioning of test tables by study or protocol. This separates the data into the key logical unit of a study. This limits all insert/update operations to a study and allows the easy export/import and deletion of study level data. This helps to reduce the size of the operational database, archiving out data that has been moved to a warehouse.

We provide a set of specialist triggers on a number of tables from the data model. Provided that you use system generated Test Set and Test Occasion Id values it is possible to partition all work in the database by study, protocol or test. We will analyse your database and advice on appropriate strategies for partitioning.

Work Plan

An example work plan:

Operation	Man/Days	Location
Initial requirements analysis	2	Onsite
Design	3	Offsite
Create partition, generation and addition scripts	3	Offsite
Create Trigger scripts	2	Offsite
Documentation	2	Offsite
Testing	10	50/50
Rollout	3	Online
Total	25	

This delivers a logically structured database where study level blocks of data can be moved off line quickly by the database administrators.

Advantages

- Data in smaller and more manageable partitions
- Reduced data recovery time
- Import / Export can be done at the "Partition Level".
- Faster access of data

➔ Conclusion

Our approach can be applied to multiple data management solutions including in-house systems. Driven from over 10 years experience of managing discovery data our services provide effective practical approaches to managing data.

Contact

Telephone +44 (0) 2380 411098 FAX +44 (0) 20 7871 0317 Mobile +44 (0) 7914 896943