

NOTE:

This document is for informational purposes. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described in this document remains at the sole discretion of Oracle.

This document in any form, software or printed matter, contains proprietary information that is the exclusive property of Oracle. This document and information contained herein may not be disclosed, copied, reproduced, or distributed to anyone outside Oracle without prior written consent of Oracle. This document is not part of your license agreement nor can it be incorporated into any contractual agreement with Oracle or its subsidiaries or affiliates.

Oracle's Platform for Life Sciences

Note:	2
Introduction	5
IT Challenges for Life Sciences	5
Access Distributed Data	5
Integrate a Variety of Data Types	6
Manage Vast Quantities of Data	6
Enable Secure Collaboration	7
Find Patterns and Insights	7
Overview of Oracle's Platform for Life Sciences	7
Advantages of Oracle's Platform for Life Sciences:	8
Functionality	9
Access Distributed Data	9
Dblinks	10
Oracle Generic Connectivity	10
Oracle Transparent Gateways	10
External Tables	11
Distributed Queries	11
SQL*Loader	11
MERGE Statement	12
Data Pump	12
Transportable Tablespaces	12
Oracle Streams	12
Oracle Warehouse Builder	13
Migration Toolkit	13
Integrate a Variety of Data Types	14
XML DB	14
Oracle Spatial RDF Data Model	14
LOB Data Type	14
Oracle Text	15
Oracle Spatial Network Data Model	16
Oracle <i>interMedia</i>	16
Oracle Files	17
User-Defined Data Types and Extensible Indexing	17
Manage Vast Quantities of Data	18
Real Application Clusters	19
Automated Storage Management	20
Partitioning	20

Oracle 10g Application Server	21
Oracle Data Guard.....	21
Oracle Scheduler.....	22
Enable Secure Collaboration	22
Virtual Private Database and Oracle Label Security	22
Auditing	23
Advanced Queuing.....	24
Oracle Workflow.....	24
Oracle 10g AS Web Services.....	24
Oracle 10g AS Portal.....	25
Oracle HTML DB.....	26
Find Patterns and Insights.....	26
Oracle Data Mining.....	26
Oracle OLAP	28
Oracle Discoverer	29
Statistics	29
Regular Expression Searches	30
Table Functions.....	30
IEEE Support.....	30
Conclusion.....	30

The mapping of the human genome signals that modern biology has evolved into a science of information.

Database infrastructure has become the critical component for competing in life sciences R&D.

INTRODUCTION

The mapping of the human genome signals that modern biology has evolved into a science of information. In recent years, there have been significant breakthroughs in genomics, proteomics, and clinical research, with many developments occurring not only inside laboratories *in-vivo* and *in-vitro*, but also *in-silico* with the aid of databases and high-performance computing platforms. These advancements generate huge amounts of data that, if harvested and analyzed effectively, promise to find novel treatments for diseases.

Significant discoveries and insights in life sciences often require the analysis of related data from multiple research disciplines. However, most of this data is in isolated silos on various computer systems, in many formats, and analyzed by different software applications. Life sciences organizations need an information technology (IT) solution that enables scientists to integrate, analyze and share genomic, proteomic, chemical, clinical trial, hospital, and other related life sciences data.

Oracle provides information technology to help life sciences organizations access, integrate, manage, query, mine and share the valuable information hidden in their data. Benefiting from a rich history of database enhancements, Oracle has emerged as the leading platform in life sciences with a 75 - 80% market share (IDC quote in *InfoWeek*, Dec. 12, 2002).

This white paper discusses key IT issues in life sciences, and presents features in the Oracle Database, the Oracle Application Server, and Collaboration Suite that address these challenges and that have enabled Oracle to become the “de facto” standard for the life sciences industry. These features deliver a powerful infrastructure that forms “Oracle’s Platform for Life Sciences” — an IT solution that under girds life sciences initiatives and helps increase Research and Development (R&D) productivity.

IT CHALLENGES FOR LIFE SCIENCES

Access Distributed Data

Life sciences companies need to be able to access a wide range of distributed data in order to be able to effectively discover new drugs. Companies need to be able to access the many public databanks, and integrate that data with in-house data that

may have been purchased or created internally (Figure 1). In addition, companies need to be able to share data that has been created by different departments or therapeutic areas across the whole organization, to enable companies to maximize the information that they generate from data.

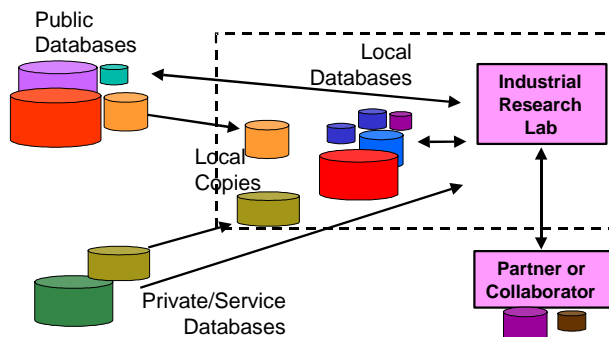


Figure 1. A Typical Life Sciences Research Environment

The Oracle Database is the leading platform in life sciences with a “75 - 80% market share”, per IDC.

Integrate a Variety of Data Types

Life sciences data are made available in many different formats. For example, XML formats are frequently used for storing literature, and sharing microarray results and sequence data; jpeg and tiff formats are regularly used for images, RDF is being explored for protein and pathways data, Adobe PDF format is used for clinical submissions, and Word and HTML are commonly used for documents and articles. Life Sciences companies need to be able to integrate all of these data types in order to be able to extract as much information as possible about any particular entity. The integration of all of these data types is made more complex by the lack of a consensus naming convention and evolving data standards.

Manage Vast Quantities of Data

Life sciences research is generating vast amounts of data due to the increase in automation and high throughput technologies. For example, genomic databases are estimated to double in size every six months, more than twice as fast as Moore’s Law predictions for microprocessor performance improvements. Life sciences data producers such as the Broad Institute and the Wellcome Trust Sanger Institute are already managing many terabytes (TBs) of data, and require high throughput processing capabilities and 24 x 7 availability to keep pace with the high volume of data. It is anticipated that data generated from systems biology and medical images will place still greater demands on organizations.

“Oracle is an excellent database. It’s been around for years, it’s been honed and developed, and it’s very good at handling large volumes of information-and that’s exactly what we need. ”

—Jennifer Allerton,
CIO, Roche

Enable Secure Collaboration

Collaboration efforts among biotech, pharmaceutical, and research institutions often place complex requirements for data access and security—often across teams and multiple applications, including legacy systems, inside and outside an organization. IT infrastructure needs to support collaboration and yet still enforce secure data access, authentication and integrity schemes.

In addition, intellectual property protection and regulatory compliance have become critical factor for any company getting a drug to market. A robust IT solution can facilitate a company in securely managing all R&D data, and meeting regulatory compliance.

Find Patterns and Insights

Once researchers are able to access and integrate the many data types, they strive to find patterns and insights in the data. These discoveries help scientists identify possible targets and leads, and help them prioritize the ones to progress. Many analytical techniques need to be embraced in order to extract information from structured and unstructured data, which include searching, querying, complex deductive and statistical analysis and inductive data and text mining. What exacerbates this transforming of “data” to “new information” is the traditional requirement of extracting the data to separate analytical or statistical engines. This increases information latency and unnecessarily exposes the data to possible security or data corruption risks.

OVERVIEW OF ORACLE’S PLATFORM FOR LIFE SCIENCES

Database infrastructure has become the critical component for competing in life sciences R&D. The explosion of data associated with R&D in life sciences requires that the data be stored in a system that is secure, highly available and scalable. In addition, it needs to enable scientists to be able to easily load, access, integrate, manage, query, analyze, and share data.

The bulk of life sciences data in companies begins as flat files generated by laboratory instruments or downloaded from public Web sites. Much life sciences data is published in flat file format because it is the lowest common denominator academic institutes use to communicate with others. This has led to scientists becoming familiar with flat file formats, and for this data type to be support by many of the academic and shareware bioinformatics tools.

However, as the amount of data in R&D increases dramatically, and as the analysis of data across multiple research disciplines becomes more important, researchers need a more sophisticated data management model.

The Oracle Database 10g, Oracle 10g Application Server and Oracle Collaboration Suite 10g present a set of features that address IT challenges in the life sciences:

- Access Distributed Data
- Integrate a Variety of Data Types

- Manage Vast Quantities of Data
- Find Patterns and Insights
- Enable Secure Collaboration

Together, these products deliver a powerful infrastructure that forms “Oracle’s Platform for Life Sciences.” As a result of the wide range of functionality in the Oracle Database, it is estimated to have an impressive “75 - 80% market share in life sciences” (IDC quote in *InfoWeek*, Dec. 12, 2002).

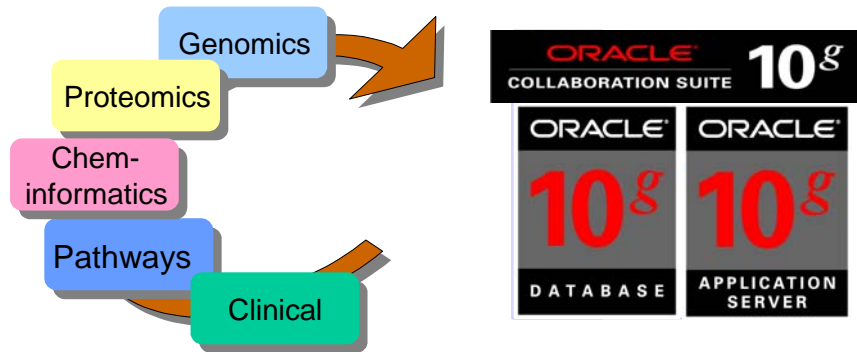


Figure 2. The Oracle Life Sciences Platform Consists of the Oracle Database, Oracle Application Server and Oracle Collaboration Suite

Oracle helps organizations to access distributed data, integrate a variety of data types, manage vast quantities of data, enable secure collaboration and find patterns and insights.

Advantages of Oracle's Platform for Life Sciences:

Oracle's Platform for Life Sciences provides functionality to solve the following major IT problems.

- The Oracle Database provides many features for accessing distributed data, enabling scientists to manage data—flat files, relational tables, distributed databases, and “graph databases”—in a federated architecture. The Oracle Database also has several tools for data loading, which simplifies centralized data management.
- The Oracle Database manages the data types that are found in the life sciences, including both structured (relational) and unstructured data (Microsoft Office, HTML, tiff, jpeg, etc.), allowing researchers to easily integrate and query information across data types.
- The Oracle Database provides the industry leading scalability, reliability and high availability that is required for managing life sciences data as it grows.
- A powerful set of analytical capabilities is provided with the Oracle Database. They include Statistics, Regular Expression Searches, BLAST, OLAP, Data Mining, Text Mining, and the ability to create custom Table Functions. Analytical capabilities embedded within the database eliminate the data

management overhead of needing to take data out of the database in order to analyze it, and helps to keep data secure. It is possible to build analytical pipelines in the database by using the analytical capabilities in combination with Oracle's workflow tools.

- Many collaboration and security features are provided by Oracle's Life Sciences Platform, including auditing, authentication, encryption, and access control, to help protect intellectual property and to aid regulatory compliance.
- Desktop applications on Windows, Mac, and UNIX operating systems can work with data in the Oracle Database through SMB, WebDAV, NFS, and AFP protocols, making the Oracle Database appear similar to any file folder on the desktop, but with the advantage that data is held in a highly available and secure environment.
- The Oracle Life Sciences Platform makes it easy to build new desktop or Web applications that access data in the database and invoke its analytical functionality, as it provides tools and APIs that conform to the industry standards (Java, XML, J2EE, JDBC, JSPs, Web services, etc.).
- Oracle BPEL and the Oracle BPEL Process Manager facilitate development of SOA based applications by composing a set of synchronous and asynchronous services into an end-to-end BPEL process flow to create, for example, bioanalytical pipelines or custom workflows.
- Many Laboratory Information Management Systems (LIMS), Enterprise Resource Management (ERP) and Customer Relationship Management (CRM) applications use the Oracle database for the data repository. By using Oracle, life sciences applications can work seamlessly with other enterprise applications. This enables R&D departments to share data along the value chain to development, manufacturing and sales and marketing.

The Oracle Database provides many features for accessing distributed data, enabling scientists to manage data in a federated architecture.

FUNCTIONALITY OF ORACLE'S PLATFORM FOR LIFE SCIENCES

In the following sections, we briefly present the features of Oracle's Life Sciences Platform that address the major IT challenges in life sciences. All the features we will discuss are provided in three products - Oracle Database, Oracle Application Server and Oracle Collaboration Suite. By leveraging and adopting Oracle as the IT platform, life sciences organizations can greatly increase R&D productivity, and thereby reduce costs.

Access Distributed Data

To analyze life sciences data, scientists must first be able to access and integrate data from the many different heterogeneous data sources. The Oracle database provides a rich set of functionality that allows life scientists to quickly, reliably, and easily access and integrate data.

Dblinks

Dblinks enables users of an Oracle Database to be able to query data that is held in other Oracle Databases. This functionality is especially useful in large companies that have many instances of Oracle. Examples of using Dblinks include joining the output of high throughput screening experiments with related literature and annotation database searches.

Oracle Generic Connectivity

Oracle Generic Connectivity is a generic solution that is available as part of the Oracle Database, which relies on industry standard ODBC or OLEDB to access non-Oracle system. In the life sciences, this functionality is primarily used to access the MySQL or Postgres databases that some scientists use on their personal desktops.

Oracle Transparent Gateways

Oracle Transparent Gateways are an option for the Oracle Database and provide end-to-end certified, tailored solutions, specifically coded for non-Oracle database systems and access these systems using their native interface. These powerful solutions provide scientists both location and operation transparency in accessing heterogeneous data sources. This enables scientists to quickly, efficiently, and economically deploy data that may exist on many disparate systems through a single application, providing a comprehensive view of the data, regardless of the data sources.

The Oracle Transparent Gateway solution provides the ability to translate SQL dialect, data dictionary, and data types between the Oracle and non-Oracle databases and has the ability to securely manage the transactions with a non-Oracle system. Remote objects can also appear as if they were local by defining a VIEW or SYNONYM.

For example, a user might execute the following SELECT statement:

```
SELECT geneId, geneName, description FROM  
genome@non_Oraclesystem;
```

After the SYNONYM genome is created on the local database, the user would access the remote object by executing the following SELECT statement:

```
SELECT geneID, geneName, description FROM genome;
```

LION Bioscience Inc. has developed and sell a product called the SRS Gateway for Oracle™, that enables Oracle users or Oracle based applications to incorporate data that is indexed by SRS into the result of an Oracle Database query.

The SRS Gateway for Oracle™ from LION Bioscience enables Oracle users or Oracle based applications to incorporate SRS indexed data into the result of an Oracle Database query.

External Tables

A lot of bioinformatics data is made available in flat files, such as the sequence data in GenBank, and the protein data in UniProt. The External Tables feature allows users to access and query data in external flat files as if they were in the Oracle Database. Once the metadata for an External Table is created, it is possible to query and manipulate the external flat file data directly using SQL, PL/SQL, and Java. The External Tables are read-only and do not allow index creation, however it is possible to use the External Tables feature to bring data into Oracle and to then index the data as part of a Materialized View.

Distributed Queries

Distributed Queries enables users to perform optimized queries across the distributed Oracle and non-Oracle data sources that are highlighted above. The cost-based optimizer is able to capture statistics for remote tables, and is able to consider network bandwidth and latency in deciding what parts of the query plan should be remotely mapped.

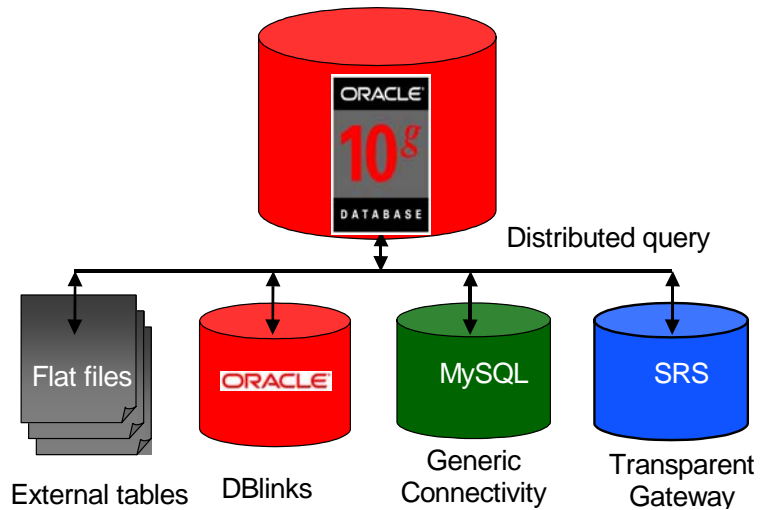


Figure 3. Oracle Database Features for Accessing Federated Data.

The Oracle Database also has several utilities for data loading that simplify centralized data management.

SQL*Loader

SQL*Loader allows users to load data into Oracle tables from operating system files. Oracle's SQL*Loader can load data from multiple data files during the same session, as well as having the capability to load data into multiple tables. The data can be loaded from disk, tape, or named pipe.

SQL*Loader contains a powerful data parsing engine which puts little limitation on the format of the data in the data file. For example, it can load arbitrarily complex object-relational data. The data can be manipulated using SQL functions before it is

loaded. SQL*Loader also has the capability to selectively load data, which includes filtering on the value of a record. The feature is character set aware, allowing users to specify the character set of the data.

Many academic groups and institutes have written SQL*Loader scripts for importing publicly available flat file data sources into the Oracle Database, and are prepared to share the files within the life sciences community.

MERGE Statement

Life sciences data is constantly being updated with new sequence data, expression data, pathway information, and annotations. The MERGE statement supports fast update of records or insertion of new records, all in one step. This feature greatly enhances performance when merging two data sources and always outperforms the procedure logic of an “IF/ELSE” statement.

The MERGE Statement command supports loading to multiple tables at once, provides a mechanism to skip updating if a certain user specified condition is met, and provides a way to perform INSERT or UPDATE if the WHEN clause always evaluates to TRUE or FALSE. For example, the MERGE statement could be used to maintain an internal copy of a gene database, where an individual sequence entry would only be updated if the checksum of the sequence and annotation indicated that an external update had occurred.

Data Pump

Data Pump is a high speed, parallel infrastructure that enables quick movement of data and metadata from one database to another. This technology is the basis for Oracle’s current data movement utilities – Data Pump Import and Data Pump Export. By using the parallel parameter, the maximum number of threads of active execution servers operating on behalf of the job can be specified, resulting in better performance.

Transportable Tablespaces

The Transportable Tablespaces feature is a powerful mechanism that moves tablespaces between Oracle Databases. It is the most efficient and fastest way to move bulk data between Oracle Databases. For example, it allows researchers to publish research data from a private database to a public database or to download a subset of a recently updated public database. The Transportable Tablespaces feature is platform independent, which means that it is possible to move a tablespace that is running on one operating system to a database that is running on another operating system.

Oracle Streams

Researchers in life sciences often need to share data and information among multiple databases and applications. A typical bioinformatics IT system often needs to provide its users with new data that has been published in multiple data sources.

“The Oracle Data Warehouse is a key component of our IT platform for proteomics analysis. The massive amount of information we produce every day requires a system with proven performance to effectively capture our biological data”.

—Bernard Gagnon,
IT Director, Caprion

Oracle Streams provides an information sharing solution that allows data sources to publish events, and consumers to subscribe to those events that meet their subscription criteria. For example, the data source can be set up to publish an event whenever a new protein structure is introduced, the consumer subscribing to the event can then update their own system accordingly.

Oracle Streams can capture events of database changes such as DML and DDL statements and application generated messages. The captured events are published in a staging area.

The staging area is a queue that provides a service to store and manage captured events. It provides security, auditing, and tracking features for the messages and the changes that have been captured. Subscribers can examine the contents of the staging area and determine if the contents meet the criteria in their subscription. If so, they can then consume the messages. Transformations can be performed in the staging area to change data formats, column names, and data types, etc.

Messages in a staging area are consumed by an apply engine, or they are explicitly dequeued and consumed by an application. Oracle Streams provides a flexible apply engine. Users can define custom apply functions in applying the updates or use a standard apply function that directly apply the changes to the destination database.

Oracle Streams is an open information sharing solution that supports heterogeneous data movement in both directions among Oracle and non-Oracle systems. Each element of the product supports industry standards. Complex distributed environments that are common in the life sciences will benefit from a single solution to simplify information sharing.

Oracle Streams provides a unified framework for information sharing. It is able to combine message queuing, replication, events, data warehouse loading, notifications, and publish/subscribe into a single technology. Streams can also be used directly by applications as a message queuing feature, enabling communications between applications.

Oracle Warehouse Builder

Oracle Warehouse Builder is a tool to enable the design and deployment of data warehouses and data marts. Warehouse Builder enables dimensional design, ETL process design, extraction of data from disparate systems, and extensive metadata reporting.

Migration Toolkit

Oracle has a series of migration toolkits that can be used to rapidly migrate data in a non-Oracle database into an Oracle database. For example, there is a toolkit available that assists users in migrating from MySQL to Oracle.

The integration of many different data types is critical for an organization to maximize its scientific and business knowledge.

Integrate a Variety of Data Types

Life Sciences organizations need to integrate a variety of different data types that result from different scientific activities, for example, images from gels and scans, text from literature and in-house projects, XML from sequence and gene expression files, etc. The integration of all of these different data types is critical for an organization to maximize its scientific and business knowledge.

XML DB

XML is heavily used by the life sciences community for the exchange and annotation of biological information, for example sequence and expression data.

XML DB unites XML content and relational data into a single repository, allowing users to access the same data from both SQL and XML. Users can perform XML operations over table data and SQL operations over XML documents. It brings the power of the database, such as scalability, performance, security, and manageability, to XML data. It also eliminates the requirement of maintaining an extra XML processing layer or a separate XML repository outside of a relational database.

XML DB fully supports the standard W3C data model. The XML Schema support allows a standard data model to be defined for all data (structured and unstructured), and to use the database to enforce the data model. It also fully supports other standards such as XPath, Xquery, XSL Transformations, and DOM.

PL/SQL, Java, HTTP, FTP and WebDAV can be used to access the data in XML DB. The protocol support not only facilitates bulk upload and data migration but also allows XML documents to be viewed as files and folders. XML DB can also be used to manage other content, such as image files, along with the XML content. XML DB supports versioning and access control for content management and collaboration.

“As an Oracle partner, I am very excited about Oracle Database 10g Release 2. The new RDF capabilities in Oracle Spatial bring enterprise-class scalability and performance to graph databases. Cerebra will work with Oracle to apply these strengths for joint customers using semantic technologies to enable the real-time adaptive enterprise.”

—Jeff Pollock,
VP Technology, Cerebra

Oracle Spatial RDF Data Model

The life sciences industry is displaying a strong interest in Semantic Web technologies. As such, Oracle Database 10g Release 2 was enhanced to provide the industry’s first open, scalable, secure and reliable RDF management platform. Based on a graph data model, RDF triples are persisted, indexed and queried, in a similar manner to other object-relational data types. The RDF Data Model ensures that application developers benefit from the scalability of Oracle 10g, to deploy scalable and secure semantic applications.

LOB Data Type

LOB is an Oracle data type that enables storage and management of large blocks of unstructured data (such as text and images) inside or outside the database. There are three types of LOBs: Binary LOB (BLOB), Character LOB (CLOB), and Binary File (BFILE). BLOBs and CLOBs are stored within the Oracle databases. BFILES are large binary data files stored in operating system files outside the database. The BFILE column or attribute stores a file locator that points to an external file. You

can also specify the URIs (uniform resource identifiers) to external files from the database.

With LOBs, you can manage your unstructured data in the same table that contains the structured data. For example, you can access the images obtained from your experiment or the raw outputs from your instrument, no matter where these files are located, together with the name, date, and description of your experiment all from the same table.

BFILEs are read-only, however the data can be loaded into internal LOBs (BLOBs and CLOBs) to take advantage of full database functionality.

Oracle Text

Much life sciences content resides in text documents, for example scientific papers, and laboratory notebooks. Therefore, scientists need the ability to perform text searches in order to retrieve relevant information of interest. Oracle Text indexes the content of a document for fast and accurate retrieval of information and allows text searches to be combined with regular database searches in a single SQL statement. For Oracle database users, Oracle Text eliminates the need to buy an extra text-searching product. Oracle Text supports multiple languages and more than 150 document formats including popular ones like the Microsoft Office file formats, the Adobe PDF formats, as well as HTML and XML. The ability to find documents based on their textual content, metadata, or attributes can make the Oracle database the single point of integration for both text and relational data.

Terminology disparity is a common problem in life sciences. Researchers frequently call the same object by many different names, making it difficult to find information about a particular gene, protein, phenotype, or disease. Oracle Thesaurus, a feature within Oracle Text, supports the definition of multiple synonyms. Oracle Text also supports query expansion capabilities.

Search results provide documents with an often very broad subject base. Further effort is required to sub-select or prioritize the results. Oracle Text includes a number of text mining features, such as document clustering and classification, which provide both supervised and unsupervised document characterization.

Oracle Ultra Search, a feature of the Oracle Database, is a search application built on top of Oracle Text. Oracle Ultra Search enables searches across multiple repositories, whether they are databases, IMAP servers, or file systems.

“As an Oracle partner, I am very excited about Oracle Database 10g Release 2. The new RDF capabilities in Oracle Spatial bring enterprise-class scalability and performance to graph databases. Cerebra will work with Oracle to apply these strengths for joint customers using semantic technologies to enable the real-time adaptive enterprise.”

—Aram Adourian,
VP Computational Sciences, BG Medicine

Oracle Spatial Network Data Model

Oracle Spatial Network Data Model enables data to be stored in a relational architecture, but to be modeled as if it were in a graph representation, i.e. a series of nodes and links. Analytical functions available with the feature include shortest path, all paths, within cost, nearest neighbors, and minimum cost spanning tree. It is possible to add constraints to the analytics that include path cost, path length, and minimum bounding rectangle.

There a number of use cases for Network Data Model in the life sciences, including managing metabolic pathways, signal transduction pathways, protein-protein interaction maps, and text co-occurrence analysis.

Oracle *interMedia*

Much life sciences data from instruments is held in image format, for example sequence data, gene expression data, histology sections, and gels. Movies are also becoming more common, especially in areas such as monitoring cell growth or animal behavior, or in viewing simulations of protein folding. Researchers need to be able to manage this multimedia data and its metadata together with other relational data, such as annotations, in the database.

Oracle *interMedia* enables the management of audio, video, and image data by the Oracle Database, both inside and outside the database environment. This allows multimedia content and its metadata to be integrated with other data, and for it to be managed with the same level of reliability, availability, and data management functionality as relational data.

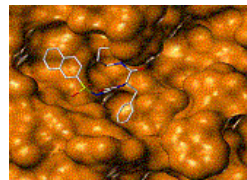
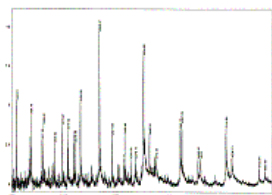


Figure 4. Oracle *interMedia* Manages Audio, Video and Image Data.

Oracle *interMedia* also provides image processing capabilities, such as metadata extraction, thumbnail generation, cropping, format transcoding, image compression/decompression, rotating and flipping an image, adjusting contrast and bit depth, brightening and darkening an image using gamma processing, and image matching using extracted features. Users can also supply their own algorithms for image analysis.

Oracle *interMedia* supports most of the popular image file formats. In Oracle Database 10g release 2, *interMedia* has been enhanced to provide support for the DICOM medical image format.

“With Oracle as the foundation, we were able to develop a solution that can secure a vast array of file-based data with vault like security.”

**—Bill Gargano,
President and COO, Taratec**

Oracle Files

Oracle Files, a feature of the Oracle Database, takes advantage of Oracle Text, Oracle *interMedia*, and LOBs, to provide a central, easy-to-use repository for content management of all data. Oracle Files combines the ease of use of a file system with the power of a database. It allows users to access files stored in the database through HTTP/WebDAV, FTP, SMB, NFS, and AFP protocols. For example, it is possible to map a network drive or a web folder from Windows Explorer to access data in Oracle as files and folders. Oracle Files allows researchers who are not software developers to access data stored in the database as easily as accessing files in a C drive. The protocol support facilitates bulk upload and data migration to Oracle. Applications on PCs or Macs can access the data stored in Oracle Files through SMB, AFP, or other protocols. Oracle Files includes a Java API that allows easy customization of the system and easy applications development with Java and XML. Oracle Files stores both structured and unstructured data, including text, multimedia, XML, and any other specialized data types as files and folders in the Oracle Database. For application specific data types, it allows you to build custom parsers and renderers to parse the information and metadata in any kind of data and render it out in any format. Metadata can be associated with files and relationships created in the data to facilitate searching. Files can be grouped using a folder hierarchy or a category to represent a project (A category is a group of metadata that can be associated with any file.) Groups of users can be created with different project access privileges. The functionality in Oracle Files makes it an ideal platform for a laboratory electronic notebook.

Cheminformatics companies have been able to develop chemical data cartridges due to the object-relational capabilities of the Oracle Database.

User-Defined Data Types and Extensible Indexing

As the Oracle Database is an object-relational database management system, users can define additional data types. Users can specify both the structure of the data and the ways of operating on it, and use these types within the relational model. User-defined data types make it easier to work with complex data such as chemical structures, spatial information and image data. With user-defined data types, for example, it is possible to do the following:

1. Create database abstractions for sequence, gene, and annotation.
2. Program the behavior for these abstractions such as `Size()` of a sequence.
3. Create collections of sequences to yield aggregations such as a chromosome.

The behavior for these types can be implemented in Java, C/C++, or PL/SQL.

User-Defined Operators

Typically, databases provide a set of pre-defined operators for the built-in data types. Operators can be related to arithmetic (+, -, *, /), comparison (=, >, <), Boolean logic (NOT, AND, OR), string comparison (LIKE), and so on. User-defined operators can be invoked anywhere built-in operators can be used, i.e., wherever expressions can occur. For example, if the user defines a new operator `ContainFragment()` that takes an Express Sequence Tag (EST) and a full sequence as input, and returns TRUE if the sequence contains the EST, then we can write a SQL query as `SELECT ID FROM DNATABLE WHERE ContainFragment(EST, sequence)`. The ability to increase the semantics of the query language by adding domain-specific operators is akin to extending the query service of the database.

Extensible Indexing

For complex scientific data types, special indexes are required to efficiently perform operations such as a three dimensional structural comparison. The Oracle Database provides extensible indexing to accommodate indexes on these complex data types. Extensible indexing enables the creation of application specific indexes as data cartridges on table columns or attributes of a user defined data type, providing efficient search and retrieval functions for these complex data. Many life sciences vendors have built cheminformatic cartridges using this technology.

Extensible Optimizer

For user-defined operators and domain indexes, the extensible optimizer gives users control over the three main inputs used by the optimizer: statistics, selectivity, and cost. The extensibility of the optimizer lies in the user's ability to collect domain-specific statistics, and based on such statistics, predict the selectivity and cost of each domain-specific operation. The optimizer estimates the cost of various access paths while choosing an optimal plan. The user-defined costs can be in the form of default costs that the optimizer simply looks up, or can be full cost functions based on user-collected statistics.

Life sciences research is generating vast amounts of data due to the increase in automation and high throughput technologies. It is anticipated that data generated from systems biology and medical images will place still greater demands on organizations.

Manage Vast Quantities of Data

Life sciences data is growing rapidly, and often scales to many terabytes. In many life sciences applications, such as high volume genomic sequencing operations, uptime availability and the ability to scale as data volume grows are very important.

Oracle Database 10g has been designed to scale to exabytes of data, offers unbreakable reliability, and enhanced manageability features. Many of these enhancements are made available as part of Oracle's enterprise grid architecture. Support for grid concepts are provided via a range of features for both data and computational provisioning.

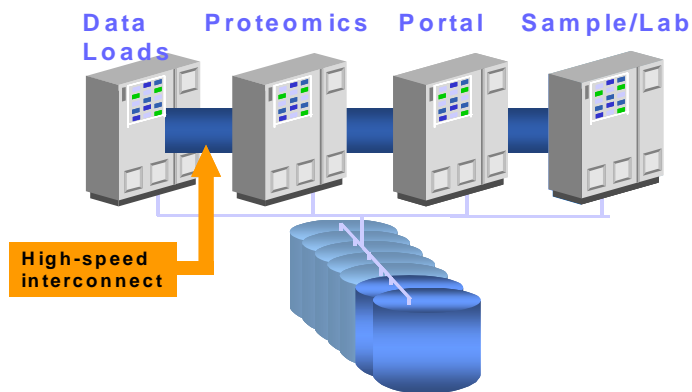
Real Application Clusters

Real Application Clusters (RAC) extends the Oracle Database to clustered systems using a Cache Fusion architecture that utilizes a shared cache for cache coherency and greatly reduces disk IO.

RAC is a technology that enables applications to scale by the seamless addition of compute nodes as required by changing computational loads or storage requirements. For example, if Oracle Data Mining (ODM) BLAST were running in a RAC database, the instance can grow to incorporate additional compute nodes if the computational load is exceptionally high due to the nature of the job or the high number of users.

Applications also do not need to be modified in order to run in a RAC environment. RAC is a high availability feature, because if there is a node failure, the database will just continue running on the remaining nodes.

RAC architecture is easy to manage as a single system image is preserved across the clusters. The DBA only needs to manage one virtual server. In addition, in Oracle Database 10g RAC introduces a new service framework that allows administrators to configure, manage, and monitor application workloads as a service, deployed across a number of nodes, in a large scale cluster deployment. This workload management service framework is integrated with the existing availability extensions provided in an earlier release of RAC, allowing administrators to not only monitor and manage performance levels for a given service, but also



provide these services continuously.

Figure 5. Real Application Cluster Architecture

“We trust Oracle in its ability to run terabyte-class databases in clustered environments with high availability. And we’re please to say that Oracle has not disappointed us.”

**—Toru Suzuki,
Project Manager, Takara Bio Inc.**

Grid Computing

Grid Computing extends RAC and is a new IT architecture that produces more resilient and lower cost enterprise information systems. With grid computing, groups of independent, modular hardware and software components can be connected and rejoined on demand to meet the changing needs of businesses. Higher quality of service results from having no single point of failure, a powerful security infrastructure, and centralized, policy-driven management. Lower costs derive from increasing the utilization of resources and dramatically reducing management and maintenance costs. Rather than dedicating a stack of software and hardware to a specific task, all resources are pooled and allocated on demand, which eliminates underutilized capacity and redundant capabilities. Grid computing also enables the use of smaller individual hardware components, which reduces the cost of each individual component and providing more flexibility to devote resources in accordance with changing needs.

Automated Storage Management

Oracle Database 10g can be configured to take advantage of an Oracle provided storage virtualization layer that automates and simplifies the optimal layout of all Oracle database managed disk storage, including data files, control files and log files. Automated storage management will configure disk groups, providing data redundancy and optimal layout of all data. As storage requirements grow, administrators can simply add inexpensive disks to the disk groups to meet the overall capacity requirement, and the automated storage management capabilities will automatically re-balance and redistribute the Oracle Database files to ensure optimal performance across the changed configuration.

Partitioning

Partitioning addresses key issues in supporting very large tables and indexes by letting users decompose them into smaller and more manageable pieces called partitions. Partitioning offers these advantages:

- It enables data management and system maintenance at the partition level. This results in reduced downtime and increased availability.
- It improves query performance because in many cases, you just need to query a subset of partitions, instead of the entire table.
- It can be implemented without any modification to applications.

Partitioning is a key tool for building systems such as genomic sequencing operations or large-scale biological image databases that store terabytes of data.

In Oracle Database 10g partitioning can increase the throughput of applications that perform a very high number of concurrent inserts via support for hash partitioning of global indexes. In addition, partitioning capabilities have been expanded to include list partitioning support for index-organized tables (IOTs), partitioning of IOTs containing large object binaries (LOBs), and automatic global index management.

“In the beginning, we considered using MySQL, Oracle, and another database. But when we evaluated our project needs over the next ten years and realized that our databases could grow to terabytes, we decided we needed a scalable database and one that was reliable. We didn’t want to be forced to change databases in the middle of the project.”

**—Joshua Li,
Sr. Computational Scientist, UCSD**

Oracle 10g Application Server

All of Oracle’s core middle-tier services have been integrated into one product, Oracle 10g Application Server, enabling customers to build and deploy portals, transactional applications, and business intelligence applications with a single product. Oracle AS Web Cache stores frequently accessed pages in memory and allows the database queries to be processed faster. Also, by reducing the load on the backend database, this service allows the database to support more users.

Additionally, Oracle 10g AS makes intensive use of connection pooling and load balancing mechanisms. Such mechanisms enable applications to run with fewer connections to the database than the number of actual users, thus increasing the total number of users that can be supported by the database.

The integration of these features in Oracle 10g AS results in dramatic increases in scalability and performance without additional computer resources, making Oracle 10g AS the most scalable middle-tier infrastructure for all types of applications.

Oracle Data Guard

In discovery departments, 24 x 7 availability of the database is crucial. Downtime often means lost revenues and lost productivity. Additionally, data integrity is essential for organizations to meet government regulatory requirements. Oracle Data Guard protects data from user errors, disasters, storage failures, and planned outages. It provides an out-of-the box rapid deployment and management interface for a standby database, a real time copy of the production system. When disasters occur, DBAs can switch over instantly to a standby database with no loss of data. Oracle Data Guard also gives users the option to set a delay in applying changes to the standby database. This feature enables the possibility to correct any human errors. Flashback capabilities can also be used to reduce downtime caused by human errors.

With Logical Standby Database, Data Guard applies SQL statements, instead of redo logs, to the standby system. Production and logical standby system can therefore be physically different. This allows the standby system to be optimized for reporting purpose and enables production work to be undertaken on the standby system.

In Oracle Database 10g, support for rolling upgrades of hardware, operating system, or database version reduces the downtime associated with database and application upgrades. Hot log mining capabilities are supported, where the Log Miner can automatically determine the relevant logs to mine for required information. Performance and security of data transmission between production and standby databases has also been improved, with both databases authenticating themselves before shipping or accepting encrypted redo data. To improve transmission speed the transmitted data can be compressed, and to improve reliability the transmitted data is check-summed. Management of a Data Guard environment has also been improved in this release, with support for more fine-

grained supplemental logging at the database, schema and table level, as well as improved monitoring capabilities.

Oracle Scheduler

The Oracle Database scheduler provides enterprise-scheduling functionality. The scheduler gives administrators the ability to schedule a job to run at a particular data and time. It also provides the ability to create libraries of the scheduler objects thus allowing existing objects to be shared by other users. It also enables scarce computing resources to be allocated appropriately among competing jobs, thus aligning job processing with the business's service level needs. Jobs that share common characteristics and behavior can be grouped into larger entities called Job Classes that can be prioritized by controlling the system resources allocated to each. For finer control, the prioritization among the Job Classes can also be based on a schedule.

Oracle BPEL

BPEL is emerging as the standard for assembling a set of discrete services into an end-to-end process flow, radically reducing the cost and complexity of process integration initiatives. Oracle BPEL Process Manager offers a comprehensive and easy-to-use infrastructure for creating, deploying and managing BPEL business processes.

Collaboration efforts among biotech, pharmaceutical, and research institutions often place complex requirements for data access and security.

ENABLE SECURE COLLABORATION

In life sciences communities, researchers often need to collaborate and share information with other researchers inside and outside an organization. . This not only creates challenges in terms of physically sharing data, but also presents many security issues. Life sciences companies also strive to achieve regulatory compliance and secure their intellectual property.

Oracle Database, Oracle Application Server and Oracle Collaboration Suite provide rich collaboration and security features that greatly facilitate regulatory compliance and collaborations among researchers and heterogeneous applications.

Virtual Private Database and Oracle Label Security

Oracle provides strong security features that ensure the safety of intellectual property. Oracle sets a new standard in database security with Virtual Private Database (VPD). The VPD enables, within a single database, per-user or per customer data access with the assurance of physical data separation. It ensures that, no matter how a user gets to the data, the same strong access control policy is enforced.

VPD security policies can be defined to trigger on relevant column access, providing both better accountability and more fine-grained data security. VPD also introduces static security policies for security rules that are always enforced (i.e. not based on changing criteria such as time of access), which provides a performance

advantage in large-scale hosted environments. In addition, VPD allows parallel execution of application contexts, and improved performance and scalability in data warehouses where parallel queries are routinely used.

Oracle Label Security extends VPD to enforce label-based access control in the Oracle Database. Label-based access control provided by Oracle Label Security allows organizations to assign sensitivity labels to information, control access based on those labels, and ensure that data is marked with the appropriate sensitivity label. For example, you might mark your data as “Company Confidential” or “Public.” Further, there may be some information that can be shared with partners. And some that is only accessible by certain groups within the company. The ability to natively manage labeled data is a tremendous advantage in being able to provide the right information to the right people at the right level of secure data access, as FDA 21 CFR Part 11 requires that “ System access limited to authorized individuals. [11.10(d)]”

Oracle also offers many other security features such as support for selective data encryption, SSL for data transport, Public Key Infrastructure for authentication, as well as single sign-on and Oracle Internet Directory, or other LDAP-compliant directory. In Oracle Database 10g, support for both Kerberos based user authentication, and database to database communications based on Kerberos credentials has been added.

IT infrastructure needs to support collaboration and yet still enforce secure data access, authentication and integrity schemes.

Auditing

A critical aspect of life sciences IT is maintaining a record of system activity to ensure that users are held accountable for their actions. The FDA requires companies submitting new drug applications to have a “secure, computer-generated, time-based audit trails to independently record the date and time of entries and actions that create, modify, or delete electronic records. [11.10(e)]”

Oracle’s Auditing feature enables organizations to define audit policies which specify the data access conditions, such as querying the target tables, misuse of legitimate access, or database intrusion, that trigger the audit event. If the audit condition is met, an audit event entry, including username, SQL text, bind variable, policy name, session id, time stamp, and other attributes, is inserted into the audit trail.

Auditing also allows companies to use a flexible event handler to notify administrators that the triggering event has occurred or respond with other appropriate actions. Oracle allows companies to audit both successful statements and unsuccessful attempts, as FDA requires that “use of safeguards to prevent unauthorized use of passwords and identification codes, and to immediately report to the system security unit any attempts at their unauthorized use [11.300(d)]” specified in 21 CFR Part 11. Fine-grained auditing applies to query operations and DML operations, to provide strong accountability of all user operations.

Intellectual property protection and regulatory compliance have become critical factor for any company getting a drug to market.

Advanced Queuing

To manage business communications within its internal departments and with its customers and business partners and track these data, a messaging system must have the capability to track, manage, queue, and deliver data securely over the Internet. With Advanced Queuing, message queuing operations get all the benefits of a database such as security and reliability, yet in addition the message queuing operations are transactional. Once committed, messages are guaranteed to be delivered. The database also offers disaster protection for these messages and all the Advanced Queuing operations are automatically audited. All the messaging information can also be looked up using SQL views. Message queuing can take advantages of the type system of the Oracle database.

Oracle Workflow

Research organizations often need to automate their laboratory according to certain laboratory and business processes. Oracle Workflow takes advantage of Advanced Queuing to provide these capabilities through an extensible process driven architecture. Oracle Workflow automatically processes and routes information of any type, according to business rules specified, to any person or system inside or outside an enterprise. Oracle Workflow allows people to receive, analyze, and respond to notifications through any standard e-mail system, or any Web browser. Laboratory workers, scientists, IT personnel, customers, research partners, and external parties can easily be included in a workflow processes.

Oracle Workflow also lets users model and maintain Laboratory Information Management System (LIMS) or your bioinformatics processes and pipelines using a graphical workflow builder. Sophisticated processes can be defined that loop, branch into parallel flows and rendezvous, decompose into sub-flows, time out, and more. The business processes for expression laboratories, data analysis, and other systems can be defined. Oracle provides the Workflow Engine and the Business Event System in the Oracle Database.

Oracle 10g AS Web Services

Scientific research in life sciences involves an interconnected world of sharing and exchanging information, tapping into domain specific databases, and running specific applications to analyze the data. Life sciences companies often need to streamline their business processes and integrate their systems with partners' over the Internet, because in life sciences, no point solution is enough, nor can a single company cover the entire space effectively. A substantial part of the value in any solution lies in its ability to integrate with information produced elsewhere in the drug development process.

Web Services consists of a set of messaging protocols, programming standards, and network registration and discovery facilities that expose business functions to authorized parties over the Internet.

Oracle AS provides a comprehensive platform to develop, deploy, and manage Web Services. It enables heterogeneous applications from different companies, written in different languages and running on different platforms, to interact programmatically in real-time over the Web. Oracle AS Web Services fully supports open Internet standards including UDDI, WSDL, SOAP, ebXML, RosettaNet, and Java/J2EE. It can also inter-operate seamlessly with those developed by Microsoft's proprietary .NET architecture.

Oracle AS Web Services offers significant benefits to life sciences companies:

- Increase research productivity – Oracle AS Web Services allows life sciences companies to integrate information and application functionality from a variety of internal sources (e.g., databases, annotations and documents) and external sources (e.g., collaborators, external data and services) using industry standard, XML-based communications protocols.
- Reduce costs – Automating the interaction between applications or laboratory processes reduces costs and minimizes human error. In addition, Web Services are re-usable components that allow developers to easily leverage existing content and functionality, reducing development costs.

Oracle has developed a new Web services framework that allows middle tier and desktop applications to search, retrieve, and extract business data from the database through standard Web services mechanisms. In addition, new capabilities can make the output generated from an external Web service call appear as a regular SQL table, where it can be used in the FROM clause of a SQL query.

Oracle 10g AS Portal

Oracle AS Portal is a Web-based application for building and deploying portals. It provides a secure, manageable environment for accessing and interacting with software services and information resources. Oracle AS Portal allows companies to build portals to integrate both external and internal web-based resources with standardized, reusable information components called portlets. It is possible to build portlets for a range of information sources, including GenBank, MEDLINE, calendar applications, newsletters, or electronic notebook applications. The deployment of portlets within a portal can be personalized for each user and managed by Oracle AS Portal. Oracle AS Portal provides services including single sign-on, content classification, enterprise search, directory integration, and access control. Portal administrators can selectively grant access to applications and information by making portlets available only to specific users or groups.

Oracle AS Portal's self-service publishing features allow users to post and share documents or web content with other users anywhere in the world. Researchers use intuitive controls for document/file upload, version control, page formatting/display, and access control to publish and manage their content; no technical expertise or knowledge of HTML is required. Users with minimal

development experience can build a variety of application components (Web forms, charts, reports, etc.) that display and interact with data in the database.

In summary, Oracle AS Portal provides an infrastructure that gives researchers fast, personalized access from a central place to all their commonly used applications and information resources. It also facilitates collaboration and information sharing among life sciences communities.

Oracle HTML DB

Oracle HTML DB is a declarative development tool and a framework for the development and deployment of database-centric Web applications. Oracle HTML DB accelerates application development through built in features such as design themes, navigational controls, form handlers and flexible reports. Using only a Web browser, it is possible to quickly build a sophisticated database driven Web application. This feature improves data access by deploying data that is currently locked in spreadsheets and personal databases to the Web, allowing concurrent updates by multiple users. Oracle HTML DB is comprised of an Application Builder that helps developers assemble an HTML interface on top of database objects, a SQL Workshop that lets developers through the Web view database objects, create database objects, run SQL commands and query by example, and a Data workshop that enables developers to import plain text and spreadsheet data into database tables and export data from database tables.

Analytical capabilities embedded within the database eliminate the data management overhead of needing to take data out of the database in order to analyze it

Find Patterns and Insights

There are vast amounts of data available in life sciences. However, only when researchers analyze the data to find hidden insights and patterns is knowledge gained from this data. Oracle Database 10g provides a wide range of analysis capabilities, including query and reporting, statistical functions, regular expression searches, data mining, text mining, OLAP, and the ability to create custom analytical functions. A key advantage of Oracle's analytics is easy data management as a result of the analytics being entirely embedded in the Oracle Database. This eliminates data movement, preserves security schemes, and retains a "single source of truth".

Oracle Data Mining

Mining life sciences data can help scientists to discover relationships. For example, mining genomic data can identify targets that may be associated with certain diseases, and leads that may be associated with cures. One challenge in finding patterns in life sciences data is the sheer size of the data—usually thousands of data fields, such as the gene expression data in a microarray experiment. Frequently significant new breakthroughs or insights are only obtainable by combining data of different types and identifying hidden information. Therefore, it is critical to consolidate data and make it available for data mining.

Oracle Data Mining (ODM) provides a number of data mining algorithms that are embedded natively within the database. By supporting data mining inside the database, data stays securely in the database and it eliminates the performance overhead of needing to move data out of the database.

ODM provides an Attribute Importance (AI) algorithm for feature selection to identify which variables are most influential in predicting a target field's value(s). The ODM AI algorithm can also be used, for example, to identify the genes that are most likely to be associated with a disease or to find the compounds that are most likely to modify gene or protein behavior.

ODM provides four supervised learning techniques for classification and prediction: Naive Bayes, Adaptive Bayes Networks, Decision Trees and Support Vector Machines; and four unsupervised learning algorithms: Association Rules, Non-Negative Matrix Factorization, Hierarchical k-means Clustering and Orthogonal partitioning Clustering (O-Cluster). ODM provides an Attribute Importance (A priori) algorithm to identify the attributes most influential on a specified target attribute and a Nonnegative Matrix Factorization (NMF) algorithm for data reduction and feature creation. Additionally, ODM 10gR2 adds support for anomaly detection using a one class support vector machine algorithm.

In supervised learning, a target field or dependent variable is identified. The supervised-learning technique then sifts through data trying to find patterns and relationships between the independent variables and the dependent variable. For example, supervised learning can be used to predict which patients may respond to a drug, and can provide some transparency or explanation of the relationships found by ODM that explain why.

In unsupervised learning, an objective is not indicated to the data mining algorithm. Associations and clustering algorithms make no assumptions about the target field. Instead, they allow the data mining algorithm to find associations and clusters in the data. For example, ODM's unsupervised learning algorithms can be used to identify new sub-disease classes by clustering the gene expression profiles of patients with the disease. Understanding the specific nature of various sub-classes of a disease is key to developing efficacious drugs.

ODM functionality is accessible via the Oracle Data Miner Graphical User Interface (GUI) and Java and PL/SQL APIs. Oracle Data Miner provides mining activity guides and wizards to assist end users through the data mining process. Application developers can either develop their own data mining applications or leverage the models and analyses created using Oracle Data Miner.

"InforSense's technology is tightly coupled with Oracle technology. We welcome the new analytics in Oracle Database 10g Release 2, such as decision trees, providing our joint customers with a comprehensive choice of in-database processing capabilities for enterprise-wide integrative analytics."

**—Yike Guo,
CEO, InforSense**

“Oracle 10g’s new BLAST feature will enable us to easily integrate multiple types of genomic and proteomic data for complicated queries used in the mining of our proprietary protein-protein interaction and cDNA sequence datasets.”

**—Jake Chen,
Principal Bioinformatics Scientist,
Prolexys Pharmaceuticals**

The Basic Local Alignment Search Tool (BLAST) family of algorithms (BLASTN, BLASTP, BLASTX, TBLASTN and TBLASTX) are available as part of the ODM option to Oracle Database 10g Enterprise Edition. BLAST searches have been implemented using the Oracle Table Functions feature and can now be performed with a single SQL statement.

BLAST_MATCH can be invoked to retrieve the sequence identifier and similarity results; and BLAST_ALIGN can be invoked to retrieve the sequence identifier, similarity results and full alignment information.

ODM BLAST relieves the burden of moving data out of the database, eliminates the need to parse data files, and allows BLAST results to be integrated with existing relational data. For example, it now becomes possible to write a query that retrieves sequences where a similarity threshold is met, and where perhaps the data was entered after July 2002, and the DNA sample was taken from John.

Text Mining

Oracle provides powerful means of mining text-based data utilizing a combination of applications including Oracle Text, Oracle Data Mining and Network Data Model. Information retrieval, extraction and mining operations can be assembled into powerful text management, processing and analysis workflows. Standard methods include document searching, theme identification, gist summarization, relationship extraction, clustering and classification. ODM’s Non-negative Matrix Factorization introduces informative feature extraction from a large collection of unstructured text documents. These features can be used independently or in combination with additional structured data in ODM’s advanced data mining algorithms. Networks and graphs of extracted entity relationships can be analyzed and visualized as a network data model with the help of the Oracle plugin to the Cytoscape viewer.

Oracle OLAP

Oracle OLAP offers integrated online analytical processing (OLAP) for fast, interactive drill-down and analysis of data. It also provides predictive analytical functions such as modeling, forecasts, and scenario management using multidimensional data. Because Oracle OLAP is fully integrated into the relational database, all data and metadata is stored and managed from within Oracle, minimizing the need for data replication and providing the extremely fast query response time.

New OLAP capabilities are available in Oracle Database 10g through built in analytical workspaces. New PL/SQL and XML interfaces are provided for the creation of workspaces based on the cubes and dimensions defined in the OLAP catalog in the database. These new interfaces can be used directly or through Oracle Enterprise Manager. New cross-tabular analysis capabilities support the aggregation of attributes within a dimension, and new parallel capabilities are provided for

AGGREGATE and SQL IMPORT operations making it faster to load and materialize the analytical workspaces from relational information.

Oracle Discoverer

Oracle AS Discoverer provides a simple, intuitive graphical user interface for interactive query, reporting and visualization of data. Oracle AS Discoverer makes it easy for scientists and managers to query and drill-down on a workbook of data and to publish graphical and tabular results to Oracle Portal. Discoverer Workbooks provide a pre-defined environment of data, calculations, and reports that enables users with little training to view and explore their data.

If a collection of life sciences data shares a common attribute “Disease A”, Oracle AS Discoverer allows scientists to interrogate other attributes of the records such as Age and Symptoms to find their relationships, and report and publish the analysis results through Oracle Portal for internal or external usage.

Statistics

Oracle’s support for statistical functions has been expanded in Oracle 10g to include summary and descriptive statistics (e.g. DBMS_STAT_FUNCS: summarizes numerical columns of a table and returns count, min, max, range, mean, stats_mode, variance, standard deviation, median, quantile values, +/- 3 sigma values, top/bottom 5 values), parametric and non-parametric hypothesis testing (e.g. Student t-test, F-test, Binomial test, Wilcoxon Signed Ranks test, Chi-square, Mann Whitney test, Kolmogorov-Smirnov test, One-way ANOVA), , distribution fit tests, correlations and cross tab statistics. Oracle Database supports linear regression, moving and cumulative aggregates, ranking functions, and other basic statistical functions. Oracle provides the 20% of statistical functions that are used 60-70% of the time—and they are all available as standard functionality in the Oracle Database 10g. All statistics functions are accessed via SQL and provide a platform for developing analytical applications. With Oracle, the data analysis occurs in the database without having to extract the data to a separate statistical software package to perform even the most basic statistical functions.

Oracle Database 10g introduces inter-row calculation clauses (via the SQL language), which add support for symbolic cell addressing, and automatic and custom formulas. With the SQL MODEL clause, you can define a multidimensional array on query results and then apply rules on the array to calculate new values. The rules can be sophisticated interdependent calculations. By integrating advanced calculations into the database, performance, scalability and manageability are enhanced significantly compared to external solutions. Rather than copying data into separate applications or PC spreadsheets, users can keep their data within the Oracle environment.

Support is also provided for looping constructs, reference datasets (lookup tables), and recursive model solving. These new capabilities make it easy to build models and perform complex calculations without needing to code multiple joins and

“Oracle 10g’s new BLAST feature will enable us to easily integrate multiple types of genomic and proteomic data for complicated queries used in the mining of our proprietary protein-protein interaction and cDNA sequence datasets.”

**—Marcel Davidson,
Head of Database Administration,
Prolexys Pharmaceuticals**

clauses. More aggregates are supported per query; performance of single row DML operations involving bit-mapped indexes has been improved, queries using bind variables are eligible for rewrite to use materialized views, and multiple materialized views can be used in a query rewrite.

Regular Expression Searches

Regular Expression Searches is supported in SQL and PL/SQL in Oracle Database 10g. Regular Expressions provide a powerful search and replace capability used by many in Unix and Java environments. Support in the database will allow developers to write one line queries that previously would have taken multiple lines of SQL code. This POSIX compliant implementation also supports multilingual queries and is locale sensitive.

An example use case of Regular Expression Searches in the life sciences is for the extraction of protein motifs in sequence data.

Table Functions

Many computational applications require extensive data processing with complex algorithms. Table Functions allows researchers to implement their own compute intensive algorithms in PL/SQL in the database or Java, C or C++ outside the database. Oracle Table functions accept a set of rows as input, and provide a set of rows as output, and can be used seamlessly in application.

Benefits of Table functions are that they allow users to integrate additional functionality with the database and make that new functionality accessible via SQL. In addition, they take advantage of much of the databases functionality, for example procedural logic, parallelism and pipelining, which gives them very powerful functionality and performance.

IEEE Support

Oracle Database 10g provides IEEE support for the industry standard for treatment and precision of numbers. Its addition to the database also results in performance enhancements.

CONCLUSION

Oracle Database 10g, Oracle Application Server 10g and Oracle Collaboration Suite 10g provide a complete IT infrastructure called Oracle's Platform for Life Sciences. Oracle's Platform for Life Sciences has a range of features that enable distributed data access, integration of a wide variety of data types, management of vast quantities of data, finding of patterns and insights, and collaborating securely. By leveraging and adopting the new features in the Oracle Life Sciences Platform, organizations can greatly increase productivity.

Oracle's Platform for Life Sciences has a range of features that enable distributed data access, integration of a wide variety of data types, management of vast quantities of data, collaborating securely and finding patterns and insights.



Oracle's Platform for Life Sciences
[September] 2005
Author: Susie Stephens
Contributing Authors: Charlie Berger

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
oracle.com

Copyright © 2005, Oracle. All rights reserved.

This document is provided for information purposes only and the contents hereof are subject to change without notice.

This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission. Oracle, JD Edwards, PeopleSoft, and Retek are registered trademarks of Oracle Corporation and/or its affiliates. Other names may be trademarks of their respective owners.