

An Oracle White Paper
February 2009

Managing Unstructured Data with Oracle Database 11g

Introduction

The vast majority of the information used by corporations, enterprises, and other organizations is referred to as Unstructured Data. Unstructured data are machine- or human-generated information where the data do not easily conform to standard data structures (such as rows and columns with well defined schema) and where the understanding of the data is not readily accessible without human or machine based interpretation. Common examples of are documents, multimedia content, maps and geographic information, satellite and medical imagery, and web content such as HTML.

The ways in which unstructured data are managed vary dramatically based on how the data are created and used.

- Huge volumes of data in desktop office systems (documents, spreadsheets and presentations) and specialized workstations and devices (geospatial analysis systems and medical capture and analysis systems).
- Multi-terabyte archives and digital libraries in government, academia and industry.
- Image data banks and libraries used in life sciences and pharmaceutical research.
- Public sector, telecommunications, utility and energy geospatial data warehouses.
- Integrated operational systems including business or health records, location and project data, and related audio, video and image information in retail, insurance, healthcare, government and public safety systems.
- Semantic data (triples) used in academic, pharmaceutical and intelligence research and discovery applications.

Advantages To Managing Unstructured Data in Oracle

Since the introduction of database management systems, database technology has been used to address the unique problems encountered when managing large volumes of unstructured data. Databases are often used to catalog and reference documents, images and media content stored in files through “pointer-based” implementations. To store unstructured data inside database tables, Binary Large Objects, or BLOBs have been available as containers for decades. Beyond simple BLOBs, for many years Oracle Database has incorporated intelligent data types and optimized data structures with operators to analyze and manipulate XML documents, multimedia content, text, and geospatial information. With Oracle Database 11g, Oracle is once again breaking new ground in the management of unstructured data through dramatic improvements in the performance, security, and types of unstructured data natively supported by a database management system.

There are many reasons organizations store unstructured data inside Oracle database management systems.

- **Robust Administration, Tuning and Management:** Content stored in the database can be directly linked with associated data. Metadata and content are maintained in sync; they are managed under transactional control. The database also offers robust services for backup, recovery, physical and logical tuning.
- **Simplicity of Application Development:** Oracle’s support for a specific type of content includes SQL language extensions, PL/SQL and JAVA APIs, Xpath and Xquery (in the case of XML) and, in many cases, JSP Tag Libraries, as well as algorithms that perform common or valuable operations through built in operators.
- **High Availability:** Oracle’s Maximum Availability Architecture makes “Zero data-loss” configurations possible for all data. Unlike common configurations where attribute information is stored in the database with pointers to unstructured data in files, only a single recovery procedure is required in the event of failure.
- **Scalable Architecture:** In many cases, the ability to index, partition, and perform operations through triggers, view processing, or table and database level parameters allows for dramatically larger datasets to be supported by applications that are built on the database rather than on file systems.
- **Security:** Oracle Database allows for fine-grained (row level and column level) security. The same security mechanisms are used for both structured and unstructured data. When using many file systems, directory services do not allow fine-grained levels of access control. It may not be possible to restrict access to individual users; in many systems enabling a user to access to any content in the directory gives access to all content in the directory.

BREAKING THE “PERFORMANCE BARRIER”

Prior to Oracle Database 11g, these benefits came at a cost. Database features like domain indexes, partitioning, and parallelism can make geospatial applications and query and update intensive XML applications perform better with content stored in the database than with content stored inside traditional file systems. However, in many other cases – multimedia applications, for example -- managing and retrieving unstructured data requires additional processing power and memory to achieve performance equivalent to file systems.

All that changes with Oracle Database 11g SecureFiles, a new, high-performance LOB that enables retrieval of unstructured data at speeds equal to and superior to that of equivalent file system configurations. SecureFiles is a major re-architecture featuring entirely new disk formats, space and memory management techniques; and delivers drastically improved LOB performance along with optimized storage.

Oracle SecureFiles

SecureFiles is designed with a completely new paradigm on how the database handles file data and delivers comparable file system-like performance for basic query and insert operations. The optimized algorithms with SecureFiles make it up to 10x faster than older LOBs. SecureFiles can take advantage of several advanced Oracle Database capabilities that are not possible with file systems. In an Oracle RAC environment, SecureFiles offer high levels of scalability that goes far beyond what is offered in file systems. SecureFiles allow for easy migration from older LOBs using Online Table Redefinition without affecting existing applications. Applications no longer have to deal with multiple interfaces for manipulating relational and associated file data. With SecureFiles, unstructured data can be part of a database transaction, thereby freeing the application from the complexity of guaranteeing atomicity, read consistency and other backup and recovery procedures.

SecureFiles extends the Transparent Data Encryption (TDE) capability to LOB data. The database supports automatic key management for all LOB columns within a table and transparently encrypts/decrypts data, backups and redo/undo log files. Applications require no changes and can transparently take advantage of TDE capabilities with SecureFiles LOBs. SecureFiles supports the following encryption algorithms:

- 3DES168: Triple Data Encryption Standard with a 168-bit key size.
- AES128: Advanced Encryption Standard with a 128 bit key size.
- AES192: Advanced Encryption Standard with a 192-bit key size. (Default)
- AES256: Advanced Encryption Standard with a 256-bit key size.

Storage Optimization in SecureFiles

Also available with SecureFiles are advanced file system features such as Deduplication and Compression. Deduplication eliminates multiple, redundant copies of SecureFiles data and is completely transparent to applications. Oracle automatically detects multiple, identical SecureFiles data and stores only one copy, thereby saving storage space. Deduplication not only simplifies storage management, but also results in significantly better performance, especially for copy operations.

LOB data will be compressed using industry standard compression algorithms resulting in significant savings in storage and improved performance. Oracle automatically determines if the SecureFile data is compressible or if compression savings are beneficial. SecureFiles use a server-wide default LOB compression algorithm and provide for varying levels of compression. Each compression level represents a tradeoff between compression factor and speed. Organizations can choose the compression level that best suit their needs based on storage and CPU usage constraints. SecureFiles are compressed and uncompressed automatically and are transparent to applications.

SPECIALIZED DATA TYPES AND DATA STRUCTURES

In the same way that database management systems include data types, storage and index structure, and operators to allow for meaningful query and analysis of structured data, they require these elements to add value when managing unstructured data. These features of Oracle Database 11g offer unique advantages specific to the management of XML, Text, Spatial, Semantic, and Multimedia and DICOM data.

Oracle XML DB

XML has been widely adopted in just about every conceivable industry. XML based standards can be found in the Health-care, Manufacturing Financial Services, Government and Publishing sectors. The introduction of XML-based standards, such as XBRL, has led to XML becoming the de-facto mechanism for exchanging information among application systems. This has led to a growth in the use of XML as a persistence model for mission critical data.

To meet this need, Oracle developed Oracle XML DB. Oracle XML DB is a high-performance, native XML storage and retrieval technology that is delivered with all versions of Oracle Database. It provides full support for all of the key XML standards, including XML, Namespaces, DOM, XQuery, SQL/XML and XSLT. Oracle XML DB is the first platform to deliver true hybrid relational / XML capabilities, making it possible to bring the full power of

the SQL language to bear on XML content and the full power of the XML paradigm on relational data.

With the release of Oracle Database 11g, Oracle extends its industry leading XML support ensuring that Oracle remains the best platform for storing, managing and querying all possible types of XML content. New features in Oracle Database 11g offer improved performance and scalability and enable complete support for the flexibility that makes the XML data model so attractive to so many different organizations.

Oracle Database 11g offers a number of improvements for users of Oracle XMLSchema-optimized XML storage.

- In-place evolution of XML Schemas.
- Oracle Partitioning of XML Schema optimized storage.
- Intelligent defaults for XMLSchema-optimized for an optimal storage model.
- XQuery operations on Schema-Optimized storage improvements
- Support for replication of text-based XMLType storage via Oracle Streams.

To address non-schema based XML in an optimal manner, Oracle Database 11g introduces a new Binary XML storage option and new XML Indexing capabilities that deliver high performance insert, update and query operations. Oracle's Binary XML format allows very efficient path based indexing of XML content. The format provides optimization of both XQuery execution and fragment extraction. The new XML Indexing capabilities of Oracle Database 11g take full advantage of this.

Oracle Database 11g simplifies the implementation of Light-Weight Service Oriented Applications by exposing Oracle PL/SQL packages, procedures and functions directly as web services. The Oracle XML DB repository now includes an event model to support workflow type applications where the simple task of creating or modifying a file automatically initiates the appropriate process.

Oracle Text

Oracle Text is the leading text searching, retrieval and management system to be integrated into a database environment. With Oracle Database 11g Release 1, Oracle Text introduces new features which aim to keep it in the leading position. These new features may be grouped into four target areas:

- Performance
- Minimization of application downtime
- Internationalization

- Ease of Maintenance

The performance of “mixed queries” – queries that have a text search part and a structured part – has been improved through the introduction of SDATA Sections and Compound Domain Indexes. The number of supported partitions has been dramatically increased; in Oracle Database 10g, the maximum number of partitions that could be used was 9999; in Oracle Database 11g the limit for text index partitions is now the same as the limit for table partitions – $2^{20} - 1$, or 1,048,575.

With Oracle Database 11g, Oracle Text supports Incremental Indexing and Online Index Recreation to improve system availability. Incremental Indexing allows you to create an index gradually at quiet times for your system. Online Index Recreation lets you create a “shadow” text index that can be built while the original index is still in use. When the index build is complete, the original index can be exchanged for the newly built shadow index. As soon as this is done, queries will automatically transition to the new index.

In addition, with Oracle Database 11g, additional internationalization support enables automatic language identification, stemming and segmentation operations for many more languages.

Oracle Spatial

In repeated studies by IDC, Oracle is the most widely used enterprise spatial database server with over 80% of the enterprise spatial database market. Every Oracle database includes Oracle Locator, built-in location features that enable any business application to directly incorporate location information and realize competitive advantages.

Oracle’s advancedgeospatial option, Oracle Spatial 11g is a comprehensive spatial database offering, including native support for vector and raster data, topology and network models, 3D data, geocoding, routing, and OGC-standard Web Services to meet the needs of business and government applications, including business intelligence and advancedgeospatial systems for land management, utilities, defense, and homeland security. Oracle’s open, native spatial support eliminates the cost of separate, proprietary systems, and is supported by all leading GIS vendors. Only Oracle delivers industry-leading security, performance, scalability, and manageability for mission critical spatial assets stored in Oracle’s native type.

With Oracle Spatial 11g, Oracle introduces:

- Integration with Oracle Business Intelligence Suite Enterprise Edition, Oracle Fusion Middleware WebCenter, and Oracle Business Activity Monitoring, through Oracle Fusion Middleware MapViewer.

- Support for storage and management of 3-dimensional data, point clouds, and terrain models.
- OpenGIS Web Services standards: Web Map Service (WMS), Web Feature Service – Transactions (WFS-T), Web Catalog Services (CS-W), and Open Location Services (OpenLS).
- GeoRaster data type and network data model now handles significantly larger data sets with greater simplicity.
- Routing engine, geocoding, topology data model, and network data model enhancements.

With Release 11g, Oracle Spatial provides significant new functionality that makes it the complete data management platform for any geospatial or enterprise location-enabled application. The spatial geometry data type has been enhanced with support for 3-dimensional data and new data types have been added with support for applications in domains such as urban planning, homeland security, or Lidar-based map production. These applications require storage and management of urban models, point clouds, and terrain models. Oracle Spatial now supports geospatial web services standards, to provide a secure, scalable service-oriented architecture platform. The GeoRaster data type and network data model have been enhanced to handle data sets larger by orders of magnitude with high performance and with greater ease of use.

Combined with the performance, scalability, and security of Oracle Database, Oracle Spatial 11g is the most advanced spatial database platform available for enterprise class deployments.

RDF, OWL and Semantic Database Management

New software and data models are emerging to help in sharing of knowledge among multiple applications in areas such as data/content integration and enterprise application integration. This software will be based on semantic data modeling standards, such as RDF and OWL from the W3C.

Oracle Database 11g incorporates native RDF/RDFS/OWL support, enabling application developers to benefit from a scalable, secure, integrated, efficient platform for semantic data management. Application developers can add meaning to data and metadata by defining a set of terms and the relationships between them. These sets of terms (“ontologies”) enable query, analysis and actions based on semantic content, rather than simply data values. Ontologies are increasingly used to build applications that utilize domain-specific knowledge. Ontological data sets, often containing 100s of millions of data items and relationships, can be stored in

groups of three, or "triples" using the new RDF data model. Oracle enables scaling to billions of triples to meet the needs of the most demanding applications.

Oracle Multimedia

Oracle Multimedia (formerly Oracle *interMedia*) is a feature that enables Oracle Database to store, manage, and retrieve images, audio, video, or other media data in an integrated fashion with other enterprise information. Oracle Multimedia extends Oracle Database reliability, availability, and data management to multimedia content in traditional, Internet, electronic commerce, and media-rich applications.

With Oracle Database 11g, Oracle Multimedia includes significant performance and scalability improvements. Oracle Multimedia supports Oracle SecureFiles, to dramatically improve performance and significantly strengthen the native content management capabilities of Oracle Database. In addition, the size limit for individual media objects that can be stored and retrieved within database storage structures (BLOB) is increased to the BLOB size limit, which is between 8 terabytes and 128 terabytes.

In addition to storing and retrieving large images, Oracle Multimedia can also extract image attributes including height, width, and compressionFormat for images that contain up to two billion pixels, or with a resolution of up to 46000x46000.

Oracle DICOM Medical Content Management

With Oracle Database 11g, Oracle Multimedia includes features and delivers the performance necessary to build large-scale repositories and archives of DICOM format medical images. By extending Multimedia to store image, audio and video using SecureFiles in Oracle databases, all the security, performance and management tools that have made Oracle the standard for enterprise class databases are now available for huge archives of media objects as well.

Specifically for medical image applications, Oracle provides methods to:

- Convert images to formats useful in web applications to simplify development of visually oriented applications.
- Extract both standard and private metadata for indexing.
- Validate that the metadata conforms to the DICOM standard or local standards.
- Remove all patient private data to create anonymous images for research, or training.
- Create new images with corrected metadata.
- Create DICOM format images from non-DICOM images.

All of these features are built for easy customization to support local requirements using a powerful model driven programming methodology. A secure Data Model Repository is used to support the frequent changes in the DICOM standard and local requirements.

Conclusion

The dramatic performance and functional improvements in Oracle Database 11g make the two essential elements for better management of unstructured data possible. First, the ability to manage, secure, query, and administer information with the highest levels of performance, and second, the ability to derive understanding and knowledge in an open, standard manner from data which had previously been dependent upon proprietary application or device logic. Over a decade of development, research, and close collaboration with customers and application providers have resulted in unique capabilities for managing unstructured data only found in Oracle Database 11g.



Managing Unstructured Data with Oracle

Database 11g

February 2009

Author: James Steiner

Contributing Authors: Mark Drake, Roger Ford,

Melliyal Annamalai, Jean Ihm, Xavier Lopez

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



Oracle is committed to developing practices and products that help protect the environment

Copyright © 2009, Oracle and/or its affiliates. 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 is a registered trademark of Oracle Corporation and/or its affiliates. Other names may be trademarks of their respective owners.

0109