ORACLE DATABASE 12*c* GLOBALIZATION SUPPORT - NEW FEATURES

KEY FEATURES AND BENEFITS

NEW FEATURES

- Unicode 6.1 Support
- Collation Support with Unicode Collation Algorithm Conformance
- New Locales Support
- Database Migration Assistant for Unicode

BENEFITS

- Support the latest Unicode Standard 6.1 character definitions
- Support the industry standard multilingual collation with flexible capabilities
- Expand your application localization support with more locales coverage
- Significantly reduce the downtime, lower the costs, and simplify the tasks of migrating databases to the Unicode character set

Oracle Database 12c delivers enriched globalization support by introducing a set of new features that facilitate the deployment of databases in the Unicode® character set and the development of multilingual, standards-compliant enterprise applications. Oracle recommends a Unicode-based system architecture which enables the storage, processing, and retrieval of character data in any languages. The new built-in database capabilities offer enhanced usability and industry compatibility in building a complete Unicode solution that meets your business requirements.

Unicode 6.1 Support

Unicode Standard defines the universal character set for encoding characters used in most of the writing systems of the world. It provides a uniform representation of textual information independent of platform or programming language. Oracle has been supporting the Unicode character sets since Oracle 7. In Oracle Database 12*c*, this support has been updated to include version 6.1 of the Unicode Standard.

Version 6.1 of the Unicode Standard was released in January 2012 by the Unicode Consortium. It supersedes all previous versions and adds new characters to the Unicode Character Database for languages of China, other Asian countries, and Africa. As of version 6.1, the Unicode Standard encodes a total of more than 110,000 characters. Improvements have also been made in character properties and collation algorithms.

Oracle's Unicode database character set AL32UTF8 and national character set AL16UTF16 definitions have been updated to conform to Unicode 6.1. Using these character sets in Oracle Database 12*c* will equip you with the most comprehensive character repertoire support in multilingual databases for all your character data processing needs.

Unicode Collation Algorithm (UCA) Conformance

Different languages have different rules for performing comparison and sorting on strings of characters. The linguistic ordering of character strings usually deviates from the ordering of their binary representations. An important aspect of a genuine globalized application is its ability to compare and present information in a way that is consistent with end-users' linguistic conventions.

Unicode Collation Algorithm (UCA) is a Unicode standard for determining the linguistic order of Unicode strings. The UCA defines a Default Unicode Collation Element Table (DUCET) that supplies a reasonable default collation for all Unicode

characters. The DUCET is also customizable to accommodate the special ordering of specific languages. The UCA is fully compatible with the international collation standard ISO 14651 but offers extended features and flexibility in the collation behavior.

Oracle Database 12*c* introduces UCA support in addition to the existing database monolingual and multilingual linguistic collations. Oracle's implementation of UCA is compliant with Unicode Standard 6.1. The main features include:

- Full collation ordering based on Unicode 6.1 DUCET
- Multilevel comparison algorithm up to 4 collation levels
- Configurable options for sorting variable weighting characters (spaces, punctuations, symbols)
    - Blanked
    - Non-ignorable
    - Shifted
- 12 tailored language-specific UCA collations for Spanish, Traditional Spanish, Canadian French, Danish, Thai, Simplified Chinese (pinyin, stroke, radical), Traditional Chinese (stroke, radical), Japanese, and Korean

UCA is the recommended mechanism for sorting multilingual data. With the newly added UCA collations, customers can attain more fine-grained control on how the searching, sorting, and matching of multilingual data will be performed.

The linguistic operations involve transforming the character data into binary values called collation keys before evaluating the relative order. As the collation keys are represented in Oracle with the RAW data type, you can now sort longer text with higher precision in Oracle Database 12*c* since the maximum length limits of the VARCHAR2, NVARCHAR2, and RAW data types have been extended to 32767 bytes.

New Locale Coverage

As part of the continued effort to expand the scope of globalization support and address fast evolving customer requirements, Oracle Database 12*c* has introduced a set of 12 new languages and 32 new territories to supported database locales, covering additional regions of Asia, Africa, Americas, and Europe:

- New languages – Amharic, Armenian, Dari, Divehi, Khmer, Lao, Latin Bosnian, Maltese, Nepali, Persian, Sinhala, Swahili
- New Territories – Afghanistan, Armenia, Bahamas, Belize, Bermuda, Bolivia, Bosnia and Herzegovina, Cambodia, Cameroon, Congo Brazzaville, Congo Kinshasa, Ethiopia, Gabon, Honduras, Iran, Ivory Coast, Kenya, Laos, Maldives, Malta, Montenegro, Nepal, Nigeria, Pakistan, Paraguay, Senegal, Serbia, Sri Lanka, Tanzania, Uganda, Uruguay, Zambia

Moreover, it also includes the support for Ethiopian calendar, a calendar system

based on the Coptic calendar with a 13th month of either 5 or 6 days in length.

Database Migration Assistant for Unicode (DMU)

Migrating to the Unicode character set is an intricate process that involves many different operational aspects which can be both time-consuming and resource intensive. Any misstep along the way can lead to data loss and serious business consequences. Oracle Database Migration Assistant for Unicode (DMU) is a next-generation migration tool that streamlines the entire migration process with an intuitive GUI to minimize the DBA's manual workload and decision-making. It helps ensure all migration issues are addressed beforehand and the conversion of data is carried out correctly and efficiently. The DMU migration workflow covers:

- Enumeration - auto-identification of database objects containing textual data that requires conversion
- Scanning - comprehensive assessment of migration feasibility and discovery of potential data issues
- Cleansing - sophisticated toolsets for iterative data analysis and cleansing to ensure data safety
- Conversion - automated in-place data conversion to minimize time and space requirements

The DMU was first released in April 2011 on OTN as a free downloadable product. The latest DMU version 1.2 is bundled with Oracle Database 12*c* and is the officially supported method for migrating databases to the Unicode character set. The DMU also supports migrating selected prior database releases of 10.2, 11.1, and 11.2. The legacy command-line utilities CSSCAN and CSALTER have been desupported.

During the migration to the Unicode character set, data issues can arise due to various scenarios ranging from incorrect character set configuration of applications, to data expansion beyond column limits, storage of non-textual data in character columns, and more. The DMU provides tools and aids to diagnose the data issues and implement the corresponding cleansing solutions. For databases containing mixed encoding data, the DMU can help determine the actual data encodings and use them to migrate the data accordingly. The new extended VARCHAR2 type limit of 32767 bytes in Oracle Database 12*c* guarantees that any VARCHAR2 column with the pre-12.1 limit of 4000 bytes or less can just be lengthened to accommodate longer values resulting from data expanding in conversion to Unicode. Neither truncation nor migration to the CLOB data type is now necessary for such columns. You can choose to perform a cleansing action immediately or schedule it to be performed in the conversion downtime window if it is a schema-related change with application impact.

To effectively reduce the migration downtime window, the DMU employs a novel in-place data migration strategy that focuses on only the data that needs to be converted. Since the majority of character data in real-world production databases is 7-bit ASCII, which does not change binary representation in the Unicode UTF-8 encoding, this strategy offers huge performance advantages over the traditional

export/import approach, which is much more costly due to the processing of unnecessary data and the need to setup multiple instances. An innovative new architecture leverages dedicated database server-side migration functionality and powerful parallel features to produce the maximum possible data throughput and scalability. It also utilizes built-in intelligence to evenly assign the workload among multiple worker threads and recommend the optimal execution plan and conversion method based on the data distribution characteristics.

The DMU goes beyond helping migration to the Unicode character set by providing the capability to conduct ongoing health check of your post-migration database to maintain data compliance with the Unicode Standard. Even with a database that uses the Unicode character set, incorrectly configured applications may introduce invalid character codes into the database and cause data corruption. The DMU's Validation Mode feature can help expose the source of the issues and detect data problems before these issues are noted by end-users.

Contact Us

For more information about Oracle Database 12*c* Globalization Support, please visit oracle.com or call +1.800.ORACLE1 to speak to an Oracle representative.

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