Abstract

The CambridgeSoft Oracle Cartridge introduces a new way of dealing with chemical structure and reaction data in an Oracle database, improving portability and consistency in applications. Since the CS Cartridge is accessed through Oracle, this opens the door for programmers interested in manipulating chemical structure data directly from their code without learning a new language. Using the CS Cartridge is as easy as using a few simple SQL commands. The CambridgeSoft Oracle Cartridge supports CDX, CDXML, MolFile, RXN, and SMILES formats making it flexible enough to be included with both new and legacy data, without the need for conversion.

This document explains how the need for the CS Cartridge arose, and the benefits of integrating the technology with your own applications or CambridgeSoft’s Enterprise Solutions.
Overview

Chemical structure information is difficult to manipulate without utilizing special software. Oracle data cartridges define new recognized datatypes. CambridgeSoft’s Oracle Cartridge utilizes this technology making it possible to manipulate chemical structure and reaction data from within Oracle.

Chemical structure information is inserted into Oracle as a BLOB or a CLOB. The data are then indexed and are fully searchable. Since the cartridge is manipulated through SQL statements, after installation, the functionality can be used from any programming language.

Historically, chemical structure and reaction information was stored in raw files, separate from the application and the database. CambridgeSoft’s Oracle Cartridge eliminates the need for extra files, improving the efficiency of the application. The CS Cartridge allows the programmer to interact with Oracle directly, minimizing consistency and portability concerns.

Chemical Structure Data Types

There are several common file formats for structure data. For this reason, the Oracle Cartridge supports all widely used chemical structure data formats (CDX, CDXML, MolFile, RXN, SMILES).

Although the cartridge supports many chemical data formats, other formats do exist. The CS Cartridge also provides SQL statement extensions that allow users to index and search data in any chemical structure data format. Although the programming may be slightly more complicated, it is still possible to take advantage of the benefits of the cartridge.

How it Works

Oracle data cartridges extend the capabilities of the Oracle server. CambridgeSoft’s Oracle Cartridge defines for Oracle how to interpret, store, and manipulate chemical structure data.

Data Cartridges

Data cartridges allow users to define new data types and indicate how to interpret the data in regards to indexing and query optimization.

These cartridges are server based in that all components are found on the server and accessed through the server. Data are stored on the server and managed by the Oracle instance.

CambridgeSoft’s Oracle Cartridge

Oracle supports many data types, but no default data type currently handles chemical structure data. CambridgeSoft’s Oracle Cartridge extends the abilities of Oracle to include a chemical structure data type.

In an architecture utilizing the CS Cartridge, the client application communicates with an Oracle instance on the Server. This Oracle instance evokes the cartridge, when necessary, and returns results to the client.

In an application using this technology, the following steps are taken:

- A table to contain structures is created.
- That table is indexed using CambridgeSoft’s Oracle Cartridge index type.
- Structure data is added to the indexed table.
- Fast Substructure and Reaction Searches are now possible.
Indexing the Structure Table

Although not necessary, indexing a table containing structures to be searched will greatly increase the speed in which Oracle can search through data.

If the table containing structures is called `struc_table` and the field containing data is called `structures`, indexing the table is accomplished in one SQL statement:

```
CREATE INDEX ix ON struc_table(structures) INDEXTYPE IS CsCartridge.MoleculeIndexType;
```

With the execution of this statement, all structures stored in struc_table are now ready to be searched by characteristics such as substructure, molecular weight, and molecular formula. The MoleculeIndexType index type is a specially designed CS Cartridge index type containing cartridge operators and functions.

**NOTE:** The more structures a table contains, the longer it will take to index a table.

Inserting Data

Chemical structure data is inserted into Oracle as a CLOB (Character Large Object) or BLOB (Binary Large Object). After a table is indexed, subsequent additions to the table will be indexed automatically.

<table>
<thead>
<tr>
<th>Field</th>
<th>Type of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOB</td>
<td>CDX</td>
</tr>
<tr>
<td>CLOB</td>
<td>CDX, CDXML, MolFile, RXN, SMILES</td>
</tr>
</tbody>
</table>

Searching

When a table is indexed, the CS Cartridge creates a cartridge-specific index table containing information about the structures found in the table.

The following command searches struc_table for structures containing a Benzene ring (`c1ccccc1` is the SMILES representation for Benzene) and a molecular weight between 200 and 500:

```
SELECT mol FROM struc_table WHERE CsCartridge.MoleculeContains(mol, 'c1ccccc1', '', 'MASSMIN=200, MASSMAX=500')=1
```

Structures of molecules matching the query are returned.

The CambridgeSoft Oracle Cartridge’s MoleculeContains operator is used, in the example above, to analyze the substructure and weight conditions entered.

Available Operators

<table>
<thead>
<tr>
<th>Operator</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoleculeIndexType</td>
<td>an index type definition for Oracle to use when indexing a table to hold structures</td>
</tr>
<tr>
<td>MoleculeContains</td>
<td>sets constraints on molecule characteristics such as substructure and molecular weight</td>
</tr>
<tr>
<td>FormulaContains</td>
<td>sets constraints on the molecular formula</td>
</tr>
</tbody>
</table>

The table below shows the available operators and their types of data:

- **BLOB**: CDX
- **CLOB**: CDX, CDXML, MolFile, RXN, SMILES
Available Functions

<table>
<thead>
<tr>
<th>Function</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConvertCDX.CDXToMolFile</td>
<td>converts CDX data to MolFile format</td>
</tr>
<tr>
<td>CDXTORxnFile</td>
<td>converts CDX data to RXN format</td>
</tr>
<tr>
<td>CDXTOSmiles</td>
<td>converts CDX data to SMILES format</td>
</tr>
<tr>
<td>CDXTOMolWeight</td>
<td>converts CDX data to a molecular weight</td>
</tr>
<tr>
<td>CDXTOMolFormula</td>
<td>converts CDX data to a molecular formula</td>
</tr>
</tbody>
</table>

Supported Platforms

CambridgeSoft’s Oracle Cartridge is supported on Windows, Linux/Intel, and Solaris 64bit. If needs for other platforms arise, this list will be expanded.

Maintenance

In general, data maintenance is performed entirely by Oracle. The CambridgeSoft Oracle Cartridge is transparent to the user and the programmer.

The database administrator may need to practice special care if the data table containing chemical structure data is exported. In this case the data table should be exported only and the index re-created in the new database (i.e. do not export the molecule index).

Uses

The most obvious use of the CS Cartridge is in taking advantage of the technology in new applications manipulating chemical structure data. It will simplify your code, and improve the overall portability of the application and database. The cartridge offers fast structure searching, more maintainable code, and chemical structure format flexibility.

In addition to writing your own applications utilizing the CS Cartridge, you can also take advantage of this technology in CambridgeSoft’s Enterprise Solutions. All ChemOffice WebServer applications are available in CambridgeSoft Oracle Cartridge versions.

Benefits

The CambridgeSoft Oracle Cartridge is not a plug-in, it is not a standalone application which can be integrated with existing code, it is an Oracle data cartridge, an extension to Oracle.

Data is stored on the server in an Oracle instance, never on the client, always keeping information consistent.

The CS Cartridge increases data format flexibility, maintainability of code, and may increase the speed of substructure searching in large databases compared to the same search in ChemFinder.

Simplifies Client Side Programming

CambridgeSoft’s Oracle Cartridge is accessed through Oracle when certain SQL commands are executed. This does not require any new private languages or interfaces, it adds to the list of operators and functions recognized by Oracle. It can be accessed from any language which supports database calls.

Programs running on the client machine will no longer have to keep track of or maintain data stored on the client. Applications will communicate only with Oracle. Oracle will in turn access the cartridge when necessary. With the CS Cartridge, chemical structure data can be dealt with as easily as any Oracle defined data types.