Oracle Text – the journey of a token

Eugen Iacob

Masstech - Romania

Keywords:

Oracle; SQL; PL/SQL; Development

Content:

Sometime programmers start the search in an application by using LIKE or REGEXP_LIKE operators and expand it to the entire application level. This approach is not well suited at application level and it will not perform as we expect.

A keyword index like Oracle Text will help in this case. Other similar products exist, but we first need to understand how Oracle Text works and how we can get most of its performance before assessing other products. Many times it can fulfill our needs and we can have good results with Oracle Text. This index is available even in the free edition Oracle XE and it evolved along with the database starting with Oracle 8i.

Data source can rest inside the database (table column / multi columns in the same table or not / nested tables / parent-child relationship tables) or outside (files in OS / FTP / HTTP). Beside plain text content, many binary formats are recognized by the filtering engine (PDF, Word, Excel, archived files…). If needed, we can create our custom filtering engine as external operating system file or internal as a PL/SQL procedure.

A simple command as:

```
CREATE INDEX … ON table_name (column_name) INDEXTYPE IS CTXSYS.CONTEXT;
```

will be enough to have a functional index that is used with CONTAINS operator:

```
SELECT … FROM table_name WHERE CONTAINS(column_name, 'keyword')>0;
```

The index is highly customizable and we can set a large variety of parameters. Under the hood, there are few tables that are created automatically by CREATE INDEX command and they are maintained by database engine without our intervention. In these tables is kept in a BLOB encoded format all information about every token, in which document is located, at which position. Structured data can be also indexed inside the text index for better performance. It is important to see how the information in these internal tables is kept during the lifetime of a table record and to understand how terms as fragmentation, on commit, transactional affects the performance and what should we do to improve it.

Performance of a text index over 4 TB of content will be presented.

I hope you will enjoy the session as much as I enjoyed working with this keyword index technology.

Contact:

Phone: +40742001323

Email: eiacob@gmail.com