

oracle认证:详解Oracle数据库中文全文索引Oracle认证考试

PDF转换可能丢失图片或格式，建议阅读原文

https://www.100test.com/kao_ti2020/565/2021_2022_oracle_E8_A

E_A4_E8_c102_565773.htm 一、历史背景 Oracle数据库的全文

检索技术已经非常完美，Oracle Text使Oracle9i具备了强大的文本检索能力和智能化的文本管理能力。Oracle Text

是Oracle9i采用的新名称，在Oracle8/8i中它被称作Oracle interMedia Text，在Oracle8以前它的名称是Oracle ConText Cartridge。

二、Oracle Text 索引文档时所使用的主要逻辑步骤如下：(1) 数据存储逻辑搜索表的所有行，并读取列中的数据。

通常，这只是列数据，但有些数据存储使用列数据作为文档数据的指针。(2) 过滤器提取文档数据并将其转换为文本表示方式。

存储二进制文档(如 Word 或 Acrobat 文件)时需要这样做。过滤器的输出不必是纯文本格式 -- 它可以是 XML

或 HTML 之类的文本格式。(3) 分段器提取过滤器的输出信息，并将其转换为纯文本。包括 XML 和 HTML 在内的不同

文本格式有不同的分段器。转换为纯文本涉及检测重要文档段标记、移去不可见的信息和文本重新格式化。(4) 词法分析

器提取分段器中的纯文本，并将其拆分为不连续的标记。既存在空白字符分隔语言使用的词法分析器，也存在分段复杂的

亚洲语言使用的专门词法分析器。(5) 索引引擎提取词法分析器中的所有标记、文档段在分段器中的偏移量以及被称为

非索引字的低信息含量字列表，并构建反向索引。倒排索引存储标记和含有这些标记的文档。

三、需要的权限 要使用Oracle Text，必须具有CTXAPP角色或者是CTXSYS用户

。Oracle Text为系统管理员提供CTXSYS用户，为应用程序开

发人员提供CTXAPP角色。具有CTXAPP角色的用户可执行以下任务：创建索引，管理 Oracle Text 数据字典，包括创建和删除首选项，进行Oracle Text 查询，使用 Oracle Text PL/SQL 程序包。

四、具体的实现 文本装入文本列后，就可以创建Oracle Text索引。文档以许多不同方案、格式和语言存储。因此，每个 Oracle Text 索引有许多需要设置的选项，以针对特定情况配置索引。创建索引时，Oracle Text可使用若干个默认值，但在大多数情况下要求用户通过指定首选项来配置索引。每个索引的许多选项组成功能组，称为"类"，每个类集中体现配置的某一方面，可以认为这些类就是与文档数据库有关的一些问题。例如：数据存储、过滤器、词法分析器、相关词表、存储等。每个类具有许多预定义的行为，称之为对象。每个对象是类问题可能具有的答案，并且大多数对象都包含有属性。通过属性来定制对象，从而使对索引的配置更加多变以适应于不同的应用。

(1) 存储 (Storage) 类 存储类指定构成Oracle Text索引的数据库表和索引的表空间参数和创建参数。它仅有一个基本对象：BASIC_STORAGE，其属性包括：I_Index_Clause、I_Table_Clause、K_Table_Clause、N_Table_Clause、P_Table_Clause、R_Table_Clause。

(2) 数据存储 (Datastore) 类 数据存储：关于列中存储文本的位置和其他信息。默认情况下，文本直接存储到列中，表中的每行都表示一个单独的完整文档。其他数据存储位置包括存储在单独文件中或以其 URL 标识的 Web 页上。七个基本对象包括：Default_Datastore、Detail_Datastore、Direct_Datastore、File_Datastore、Multi_Column_Datastore、URL_Datastore、User_Datastore。

(3) 文档段组 (Section Group) 类 文档

段组是用于指定一组文档段的对象。必须先定义文档段，然后才能使用索引通过 WITHIN 运算符在文档段内进行查询。文档段定义为文档段组的一部分。包含七个基本对象

：AUTO_SECTION_GROUP、BASIC_SECTION_GROUP、HTML_SECTION_GROUP、NEWS_SECTION_GROUP、NULL_SECTION_GROUP、XML_SECTION_GROUP、PATH_SECTION_GROUP。（4）相关词表（Wordlist）类相关词表标识用于索引的词干和模糊匹配查询选项的语言，只有一个基本对象BASIC_WORDLIST，其属性有

：Fuzzy_Match、Fuzzy_Numresults、Fuzzy_Score、Stemmer、Substring_Index、Wildcard_Maxterms、Prefix_Index、Prefix_Max_Length、Prefix_Min_Length。（5）索引集

（Index Set）索引集是一个或多个Oracle 索引（不是Oracle Text索引）的集合，用于创建 CTXCAT类型的Oracle Text索引，只有一个基本对象BASIC_INDEX_SET。（6）词法分析器

（Lexer）类 词法分析器类标识文本使用的语言，还确定在文本中如何标识标记。默认的词法分析器是英语或其他西欧语言，用空格、标准标点和非字母数字字符标识标记，同时禁用大小写。包含8个基本对象：BASIC_LEXER

、CHINESE_LEXER、CHINESE_VGRAM_LEXER、JAPANESE_LEXER、JAPANESE_VGRAM_LEXER

、KOREAN_LEXER、KOREAN__MORPH_LEXER

、MULTI_LEXER。（7）过滤器（Filter）类 过滤器确定如何过滤文本以建立索引。可以使用过滤器对文字处理器处理的文档、格式化的文档、纯文本和 HTML 文档建立索引，包括5个基本对象：CHARSET_FILTER、INSO_FILTER INSO

、NULL_FILTER、PROCEDURE_FILTER、USER_FILTER。

(8) 非索引字表 (Stoplist) 类 非索引字表类是用以指定一组不编入索引的单词 (称为非索引字)。有两个基本对象

：BASIC_STOPLIST (一种语言中的所有非索引字)、MULTI_STOPLIST (包含多种语言中的非索引字的多语言非索引字表)。具体操作实践 1、创建词法分析器及相关词表

Begin-- 定义一个词法分析器 ctx_ddl.drop_preference(cnlex).
ctx_ddl.create_preference(cnlex,CHINESE_LEXER). --针对中文
-- 定义一个相关词表 ctx_ddl.create_preference(mywordlist,
BASIC_WORDLIST).

ctx_ddl.set_attribute(mywordlist,PREFIX_INDEX,TRUE).
ctx_ddl.set_attribute(mywordlist,PREFIX_MIN_LENGTH,1).
ctx_ddl.set_attribute(mywordlist,PREFIX_MAX_LENGTH, 5).
ctx_ddl.set_attribute(mywordlist,SUBSTRING_INDEX, YES). end.
// 上面的语句中，如果是针对英语语种则可以采用下面的语句来定义词法分析器

ctx_ddl.create_preference(mylex,BASIC_LEXER).
ctx_ddl.set_attribute(mylex,printjoins,-). 2、在需要创建全文索引的表中创建索引（索引类型必须是 ctxsys.context，即应用上下文索引） create index idx_person_desc on
personinfo(persondesc) indextype is ctxsys.context parameters (
DATASTORE CTXSYS.DIRECT_DATASTORE FILTER
CTXSYS.NULL_FILTER LEXER cnlex WORDLIST mywordlist).
-- 请注意此处采用的是NULL_FILTER过滤器，如果采用INSO_FILTER则不能对中文进行全文索引 3、进行全文索引的同步 exec ctx_ddl.sync_index(idx_user_info, 20M). 100Test 下

载频道开通，各类考试题目直接下载。详细请访问
www.100test.com