目录

[应用分区实现大数据量抽取 1](#_Toc422910189)

[1. 基本概念 1](#_Toc422910190)

[2. DI分区功能说明 1](#_Toc422910191)

[4. 案例-数据分区 2](#_Toc422910192)

[4.1. 步骤一：数据库连接设置分区 2](#_Toc422910193)

[4.2. 步骤二：项目属性设置分区 3](#_Toc422910194)

[4.3. 步骤三：转换项设置分区 3](#_Toc422910195)

[4.4. 测试验证 4](#_Toc422910196)

[5. 案例-求余分区 5](#_Toc422910197)

[5.1. 步骤一：项目属性设置分区 5](#_Toc422910198)

[5.2. 步骤二：转换项设置分区 5](#_Toc422910199)

# 应用分区实现大数据量抽取

By 段丽霞

# 基本概念

数据库分区主要有两种形式：水平分区和垂直分区。如下表所示，说明了分区的概念：

|  |  |  |
| --- | --- | --- |
| **对比项** | **水平分区** | **垂直分区** |
| **定义** | 按记录进行分区，不同的记录可以分开保存，每个子表的列数相同。  例如，表T1，可以把id为单数的放到数据文件P1，双数的放到数据文件P2 | 按列进行分区，即把一条记录分开多个地方保存，每个子表的行数相同。  例如，表T1，可以把id和name放到数据文件p1，把qty放到数据文件p2。 |
| **特点** | 结构相同，数据不同。 | 结构不同，数据可能不同，可能相同 |

分区是为了增加数据的并行度，根据系统资源如存储，CPU等增加横向扩展能力。

# DI分区功能说明

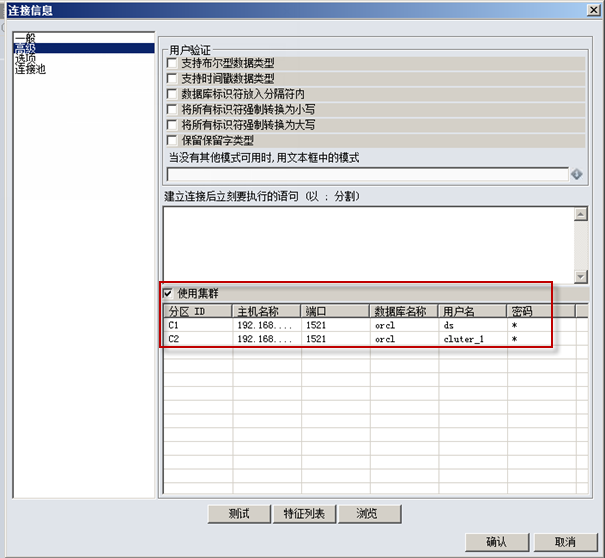
DI产品中支持两种分区方案：数据库分区和求余分区。如下表所示，介绍了DI产品中的两种分区方案：

|  |  |  |
| --- | --- | --- |
| **对比项** | **数据库分区** | **求余分区** |
| **定义** | DI中数据库分区支持oracle数据库水平分区， | 对某字段值通过求余将表数据划分为N个。 |
| **应用场景** | 数据量大，可以应用数据库物理分区进行并行操作，提升性能。 | 数据量大，可以应用求余法将数据分割为逻辑上的几个分区并行操作。 |
| **操作步骤** | 1. 数据库连接中定义。目前支持Oracle集群分区。 2. 项目属性中导入分区设置。 3. 转换中设置应用定义好的分区。 | 直接在转换中应用分区。 |
| **注意事项** | 1. 目前只针对Oralce数据库分区。 2. 所有分区镜像：只针对表输入。 3. 分区设置要注意设计逻辑，否则不仅不会提升性能，还会降低性能。 | 一般来说，求余之后划分的表分区比较均匀。如果差异很大，则需要重新设计。选择字段和选择求余的数字是否合适。 |

# 案例-数据分区

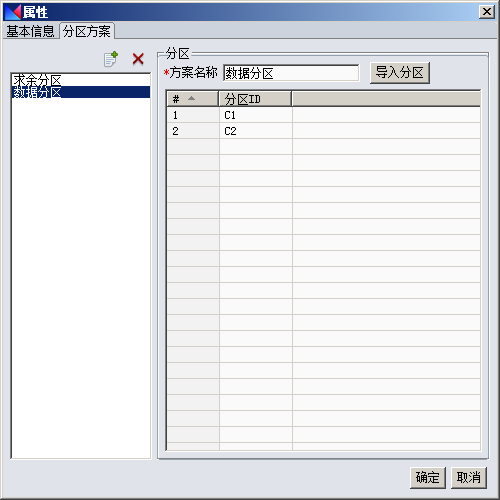
## 步骤一：数据库连接设置分区

如下图所示，编辑数据库连接，点击高级-使用集群，新建数据库水平分区方案，如下图所示：



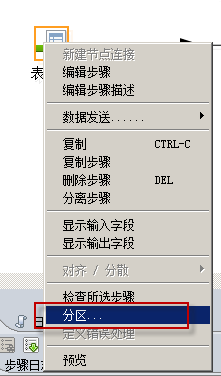
## 步骤二：项目属性设置分区

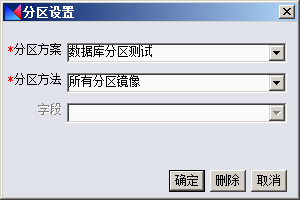
选中项目，右键点击属性，选择页签【分区方案】，如下图所示。点击按钮【新增】，输入分区方案名称，这里，选择【导入分区】（也可以手动输入）。

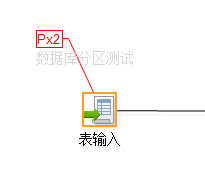


## 步骤三：转换项设置分区

选中需要设置分区的转换项，点击右键，弹出右键功能选项中的【分区】，如下图所示。







## 测试验证

最后，有必要验证一下是否有提升性能，达到预期效果。如果未达到效果，查看设计逻辑是否有问题，是否需要改进等。

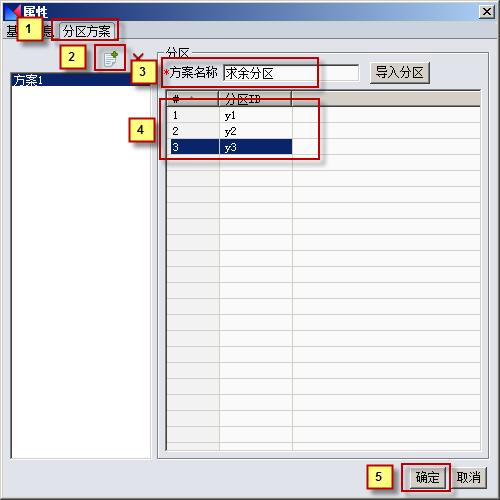
# 案例-求余分区

## 步骤一：项目属性设置分区

选中项目，右键点击属性，选择页签【分区方案】，如下图所示，点击按钮【新增】，输入分区方案名称，在字段【分区ID】中输入求余分区需要设置的分区ID。

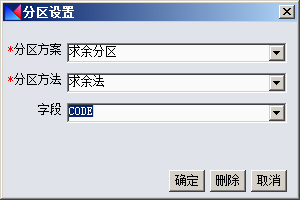
一般来说，分区ID的设置方法：

* 实施分区的列如果类型为整数型，那么分区方案为，取【列的值/分区ID的个数】的余数进行分区。
* 实施分区的列如果类型为字符型，那么分区方案为，取该列的哈希值进行分区。哈希算法略。



## 步骤二：转换项设置分区

选中需要设置分区的转换项，点击右键，弹出右键功能选项中的【分区】，如下图所示。



其中，【字段】为需要进行分区的列。

设置之后，工作区图标显示如下：

