借助 IBM PureData System for Analytics 生成数据的四个简单方法

2014 年 8 月 20 日上午 8:39；作者：Paul Froggatt

您是否需要快速生成数据集？或者需要扩展现有数据集？那么，下面将为您介绍通过 IBM PureData System for Analytics 实现这些需求的四个示例。

通过 PureData System 生成数据

1.基本构建模块

基于 Netezza 技术的 IBM PureData System for Analytics 包含一个名为 _v_vector_idx 的简单视图。该视图中包含了一个名为 idx 的列，该列可返回 1,024 行数据，其中 idx 的值是唯一的，范围在 0 到 1,023 之间。

不错，这没有什么特别的，之前也听说过。不过，我们还是来看一下它有哪些功能。

首先，您可能需要的行比最初更多，并且每行都具有唯一的值，那么就需要了解如何扩展从该视图中返回的结果。举例来说，以下查询将返回 32,000 个值：

```
SELECT ((mult.idx * 1024) + base.idx) idx

FROM _v_vector_idx base,

(SELECT idx FROM _v_vector_idx WHERE idx < 32) mult

ORDER BY idx;
```

当然，您可以在上面的示例中添加一个 WHERE 语句，来限制结果集，从而进一步扩展。既然您已经了解了如何控制唯一值的总体数量，那么可以使用这种基本方法来获取更加有意思的结果。

2.生成一系列日期值

在一个日期集内，如果您希望创建的表格中每个日期都对应一行数据，您可以运行类似于以下示例的脚本，通过该脚本可返回 1940 年到 2010 年之间的日期。

```
CREATE TABLE lots_of_dates AS

SELECT  '1900-01-01'::DATE + ((mult.idx * 1024) + base.idx) as the_date

FROM _v_vector_idx base,

(SELECT idx FROM _v_vector_idx) mult

WHERE the_date between  '1940-01-01' and  '2010-12-31';
```

3.生成一系列时间值

如果您想要创建一天内每一秒均包含一行数据的表格，您可以运行以下示例脚本：

```
CREATE TABLE all_seconds AS

SELECT DISTINCT(’ 00:00:01’::TIME * ((mult.idx * 1024) + base.idx)) as the_time

FROM _v_vector_idx base,

(SELECT idx FROM _v_vector_idx) mult

WHERE the_time BETWEEN ’ 00:00:00’and ’ 23:59:59’;
```

4.扩展数据集

您还可以在现有数据集的基础上对其进行扩展。举例来说，如果您拥有一个包含有一整天数据的表格，但您想要将其扩展到包含一个月（30 天）的数据，您可以运行以下示例脚本：

```
CREATE TABLE one_month_table AS

SELECT date_col + idx as date_col, col2, col3 /* and any more columns */

FROM one_day_table CROSS JOIN _v_vector_idx

WHERE idx BETWEEN  0 and 30;
```

正如上文所述，有很多种方法可用于生成我们所需的数据集。我希望大家能够找到这种简单、有用的方法，我非常愿意听听大家在各自环境中所采用的独特方法，欢迎大家在下方留言。

关于 Paul Froggatt

Paul 是英国 IBM Software Group 大数据技术销售团队中的一名客户技术专家，擅长于 IBM PureData System for Analytics（又称作 Netezza）技术，主要面向零售和电信行业。在 IBM 于 2010 年收购了 Netezza 之后，Paul 加入到了 IBM 的团队之中，之间他曾在英国的一家领先电信公司建立并管理着一支非常成功的信息管理团队，他主要负责根据客户的数据仓储需求，评估、选择及部署 Netezza 技术。2013 年，Paul 凭借其在践行 IBM 人的价值观及推动客户与 IBM 实现双赢方面的卓越表现，被评选为“最佳 IBM 人 (the Best of IBM)”。