

数据挖掘-派生列

概述

派生列节点是用于在数据集中生成可行的新特征字段。它可对现有数据的某个特征操作，允许用户自定义生成特征名称，并将生成的新特征字段添加到原数据集中。

- [概述](#)
- [输入/输出](#)
- [参数设置](#)
- [示例](#)



输入/输出

输入	一个输入端口，用于接收数据集。
输出	一个输出端口，用于输出派生列的结果。

参数设置

设置派生列的参数：



设置说明如下：

参数	说明
添加表达式	派生列配置 输入关键字搜索列 添加/编辑表达式 1 2 3 STRING T_O LONG Time TAZ_O TAZ_D 函数 累积 字符串拼接(拼接符) 字符串拼接 字符串截取 小数点保留 1 2 3 1 2 3 1. 添加/编辑表达式：选择对应的函数和字段，拖入添加/编辑表达式区。 2. 新增表达式：输入派生列名，单击 增加 按钮添加表达式到列表。 3. 保存表达式：点击 确定 按钮进行保存。

修改已定义的派生列列名

在列表区选择表达式，单击 **修改** 按钮后，修改列名再单击 **增加** 按钮。

添加/编辑表达式

1

派生列名:

请输入或左侧拖入派生列名

操作:

增加

修改

Year

year(Date)

确定

取消

目前派生列支持的函数有：

函数名称	函数及示例	说明
累积	sum(<column>) over(partition by <columns> order by <columns>)	对数字型字段进行累计计算。 - Sum(): 用于设置求和计算字段; - Partition by(): 用于设置分组的字段; - Order by(): 用于设置排序的字段;
字符拼接（拼接符）	concat_ws("sep", str1, str2, ..., strN)	通过sep连接符将字符串拼接。 - sep由用户自定义。 - str: 表示连接的字符。
字符拼接	concat(str1, str2, ..., strN)	不需要连接符，直接将字符串连接。 str: 表示连接的字符。
字符截取	substring(<column>, pos[, len])	从字段中截取指定位置的len个字符。 <column>: 表示截取的字段; - pos: 表示从第几个位置开始截取，设置为正整数; - len: 表示一共取多少个字符，设置为正整数;
小数点保留	format_number(<column>, num)	保留浮点型字段小数点位数。 <column>: 表示设置小数点位数的字段; - num: 表示小数点后保留的位数，设置为0和正整数;
内容提取（年）	year(<column>)	截取时间字段中的年信息。 <column>: 表示截取的字段;
内容提取（月）	month(<column>)	截取时间字段中的月信息。 <column>: 表示截取的字段;
内容提取（日）	day(<column>)	截取时间字段中的日信息。 <column>: 表示截取的字段;
内容提取（时）	hour(<column>)	截取时间字段中的时信息。 <column>: 表示截取的字段;

内容提取（分）	minute(<column>)	截取时间字段中的分信息。 <column>：表示截取的字段；
内容提取（秒）	minute(<column>)	截取时间字段中的秒信息。 <column>：表示截取的字段；
向上取整	ceiling(<column>)	获取大于或等于取整字段的最小整数。 <column>：表示取整的字段。
向下取整	floor(<column>)	获取小于或等于取整字段的最大整数。 <column>：表示取整的字段。
四舍五入	round(<column>)	获取四舍五入后的整数。 <column>：表示四舍五入的字段。

示例

1、原先文本数据源有1000条数据，想要获取Date字段中关于年份的信息，选择派生列节点，表达式为“year(Date)”。

The screenshot displays a data processing workflow in a software interface. On the left, a sidebar lists various data processing steps, with 'Derive Column' (派生列) highlighted. The main workspace shows a flow starting from a 'Text Data Source' (文本数据源) node, which connects to a 'Derive Column' (派生列) node. A red arrow points from the 'Derive Column' node in the sidebar to the 'Derive Column' node in the workspace. Another red arrow points from the 'Derive Column' node in the workspace to the configuration panel. The configuration panel, titled 'Derive Column Configuration' (派生列配置), shows the expression 'year(Date)' in the 'Expression' (表达式) field. Below the expression field, the 'Derived Column Name' (派生列名) is set to 'Year'. The panel also includes a 'Confirm' (确定) button and a 'Cancel' (取消) button.

2、输出结果如图：

当前显示 100 条 / 总共有 1000 条数据 提示:点击单元格可查看超出的内容 获取字段Date中关于年的信息，生成新的一列Year

As Date	# Time	As T_O	# TAZ_O	# TAZ_D	# NUMKY	# Year
2017-09-27 00:00:00.0	17	2017-09-27 17	320	1177	11	2017
2017-10-01 00:00:00.0	15	2017-10-01 15	320	1177	11	2017
2017-10-07 00:00:00.0	23	2017-10-07 23	320	1177	11	2017
2017-10-08 00:00:00.0	16	2017-10-08 16	320	1177	23	2017
2017-09-19 00:00:00.0	16	2017-09-19 16	320	1192	11	2017
2017-10-08 00:00:00.0	6	2017-10-08 06	320	1192	11	2017
2017-09-22 00:00:00.0	7	2017-09-22 07	320	1223	11	2017
2017-09-22 00:00:00.0	11	2017-09-22 11	320	1223	11	2017
2017-10-01 00:00:00.0	12	2017-10-01 12	320	1223	11	2017
2017-10-07 00:00:00.0	17	2017-10-07 17	320	1223	11	2017
2017-10-08 00:00:00.0	8	2017-10-08 08	320	1223	11	2017

注意：表头中◆表示特征列，*表示标签列

表头真名 表头别名