缺失值节点

作者：张青松

目录

[1. DataStudio中的缺失值 2](#_Toc433907499)

[2. 使用缺失值节点处理缺失值 3](#_Toc433907500)

[2.1. 节点设置界面 3](#_Toc433907501)

[2.2. 处理方式 5](#_Toc433907502)

# DataStudio中的缺失值

缺失值是指在输入数据二维表中，某些单元格的数据不存在。在读取这样的数据时，DataStudio使用一个缺失值（MissingCell）占位符表示该单元格中的数据。



如下表格所示的数据，下表包含了常见的几种类型并且都有缺失值：

|  |  |  |  |
| --- | --- | --- | --- |
| integer | double | string | unknown |
| 1 | 1.2 | a | 2015/7/7 |
|  | 0.33 | b | 2015/7/7 |
| 3 |  | + | 2015/7/7 |
| 4 | 1.6 |  | 2015/7/7 |
| 10 | 5.5 | j |  |

表1 缺失值数据表

在表格中的每一列依次为整形（Integer）、双精度（Double）、字符串（String）、和其他类型（Unknown）类型。并且存在各种类型的缺失值。

然后，使用“文件读取”节点读取该数据得到的数据如下表



其中“文件读取”节点暂时自动将String类型的缺失值处理成空字符串（“”）。其他标示为“<null>”的单元格即为缺失值。

# 使用缺失值节点处理缺失值

## 节点设置界面

考虑到数据表中存在的缺失值会影响对数据的计算分析过程。在必要的情况下，可以使用“缺失值”节点对缺失值进行处理。

缺失值处理分为两种方式：

1. 按照字段类型处理。
2. 指定具体字段处理。

注意：当同一个字段同时符合按字段处理，并且又指定了具体字段时，以指定字段处理方式为准。

“缺失值”节点设置面板如下所示：



图2 设置面板“默认”页签



图3 设置面板“自定义”页签

如上图2和图3所示，分别展示了按字段类型处理和指定字段处理。

## 处理方式

对于缺失值提供了如下几种处理方式：

1. 不处理：指定的缺失值不做任何处理。
2. 移出该行：将指定缺失值的行从数据表中移出。
3. 最大值：针对数值类型的缺失值，用所在列的数据中的最大值填充缺失值位置。
4. 最小值：同上，用最小值填充。
5. 平均值：同上，用平均值填充。
6. 词频最高：对于字符串类型这样的可枚举的类型，使用出现频率最高的填充。
7. 指定值：用输入的值填充单元格。

并不是各种数据类型都能用上述所有处理方式来处理缺失值，针对不同类型的字段，提供了不同的处理方式。

1. 数值类型，包括整形（Integer）和双精度（Double）类型，提供的处理方式有：不处理、移出该行、最大值、最小值、平均值和指定值。
2. 字符串（String）类型：提供处理方式为不处理、移出该行、词频最高以及指定值。
3. 未知类型：对于其他类型，仅仅提供通用的不处理和移出该行。