目录

[需要预先排序的转换器说明 1](#_Toc418499475)

[1. 什么是输入流预先排序 1](#_Toc418499476)

[2. 为什么需要输入流预先排序 1](#_Toc418499477)

[3. 哪些转换器需要输入流预先排序 1](#_Toc418499478)

[4. 需要输入流预先配置的转换器如何配置 2](#_Toc418499479)

[4.1. 合并记录 2](#_Toc418499480)

[4.2. 记录连接 3](#_Toc418499481)

[4.3. 记录去重 4](#_Toc418499482)

[4.4. 分组 5](#_Toc418499483)

[5. 如果输入流没有预先排序有什么影响 6](#_Toc418499484)

[6. 总结 6](#_Toc418499485)

需要预先排序的转换器说明

**BY 纪文龙**

# 什么是输入流预先排序

对于一个转换器来说，有输入、输出，输入的数据按照某一字段或多个字段进行排序，这样转换器获取的输入流是预先排好序的。

# 为什么需要输入流预先排序

数据转换流程是一个生产者消费者模型，每个步骤为一个节点，而节点之间是通过一个阻塞队列来传递数据，前一个节点一条一条往阻塞队列写入，后面节点一条一条从阻塞队列读取。所以在这种机制下，某些转换节点比如记录去重转换器，当前读取一条判断是否重复就是要与前一条进行比较，所以就要预先排好序比较才有意义。

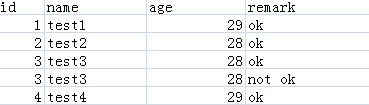
# 哪些转换器需要输入流预先排序

|  |  |
| --- | --- |
| 需要输入流预先排序的转换器列表 | |
| 转换器 | 描述 |
| 合并记录 | 将旧数据源数据与新数据源合并，得到最新版本的数据。相当于oracle中的merge操作。(判断合并依据的字段需要排序) |
| 记录连接 | 对具有来自两个不同输入步骤所得的数据集执行标准连接操作。连接选项包括 INNER、LEFT OUTER、RIGHT OUTER 和 FULL OUTER。（连接的字段需要排序） |
| 记录去重 | 从输入流中去除重复行。(判断重复依据的字段需要排序) |
| 分组 | 按照某些字段分组后，可以求聚合的值。(分组的字段需要排好序) |

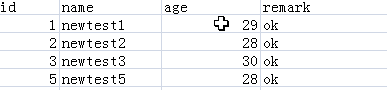
# 需要输入流预先配置的转换器如何配置

通过这些转换器的配置更好的来理解为什么需要输入流预先排序。数据准备：

Input1的数据如下

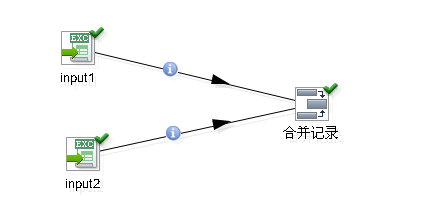


Input2的数据如下

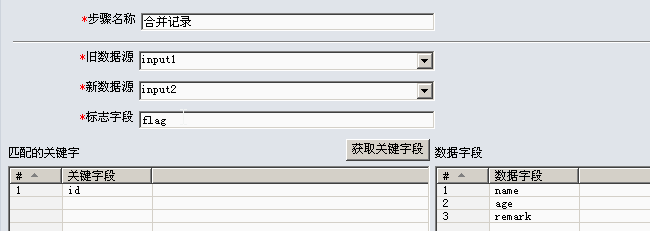


## 合并记录

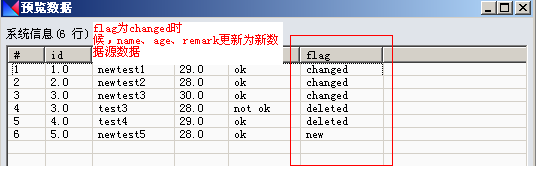
转换配置如下



* 1. Input1作为旧数据源，input2作为新数据源，判断记录合并的标准是id字段，前提就是id得排好序，配置如下：

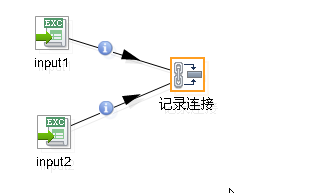


结果预览

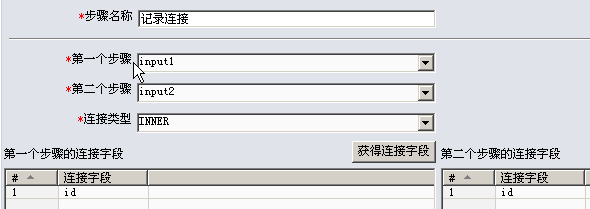


## 记录连接

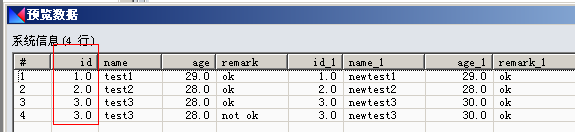
转换配置如下



记录连接将input1和input2通过id进行连接，前提就是id得排好序，配置如下：

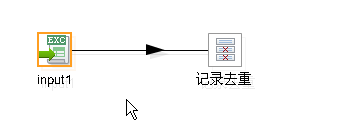


结果预览

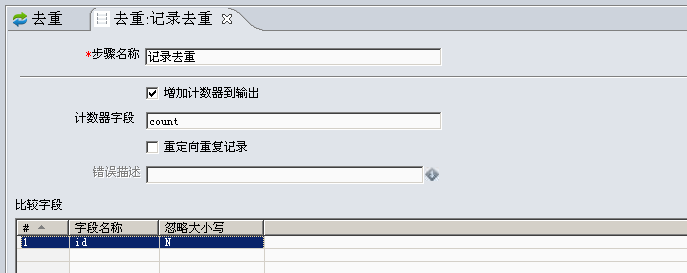


## 记录去重

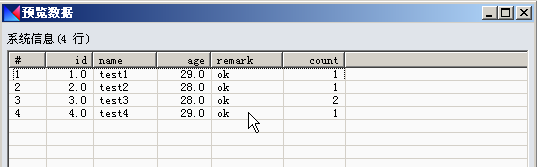
转换配置如下



认为id重复就是重复数据，前提条件就是id得实现排好序，去重配置如下



去重结果查看

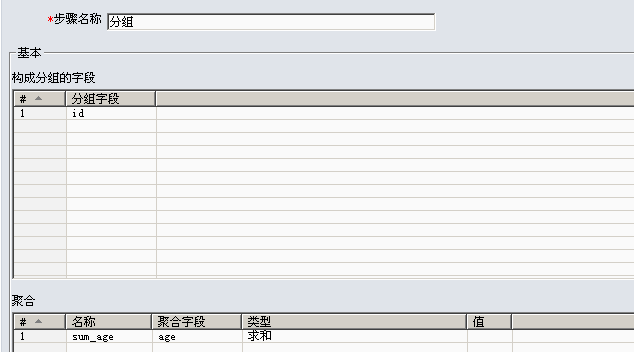


## 分组

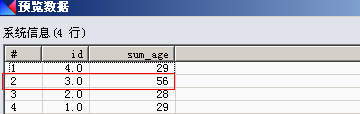
转换的配置如下



前三个转换器我的input1和input2是排好序的，如果没有排好序，可以增加个排序记录进行排序，这里对input1的id进行降序排列。然后再加分组求和



结果预览



# 如果输入流没有预先排序有什么影响

如果输入流没有预先排序，那么会导致输出结果不正确，比如记录去重，如果输入流没有排序，那么只有相邻重复的行才会检查是否重复。

# 总结

排序包括升序和降序，都可以，如果原始的输入文件没有排序，可以增加排序记录节点进行排序，如果原始输入是表输入，可以在表输入的sql查询语句处添加order by语句，也可以添加排序记录节点进行排序。