# CITTA算子设计文档

## 算子类型

### 1 关联算子

#### 1.1 算子设置

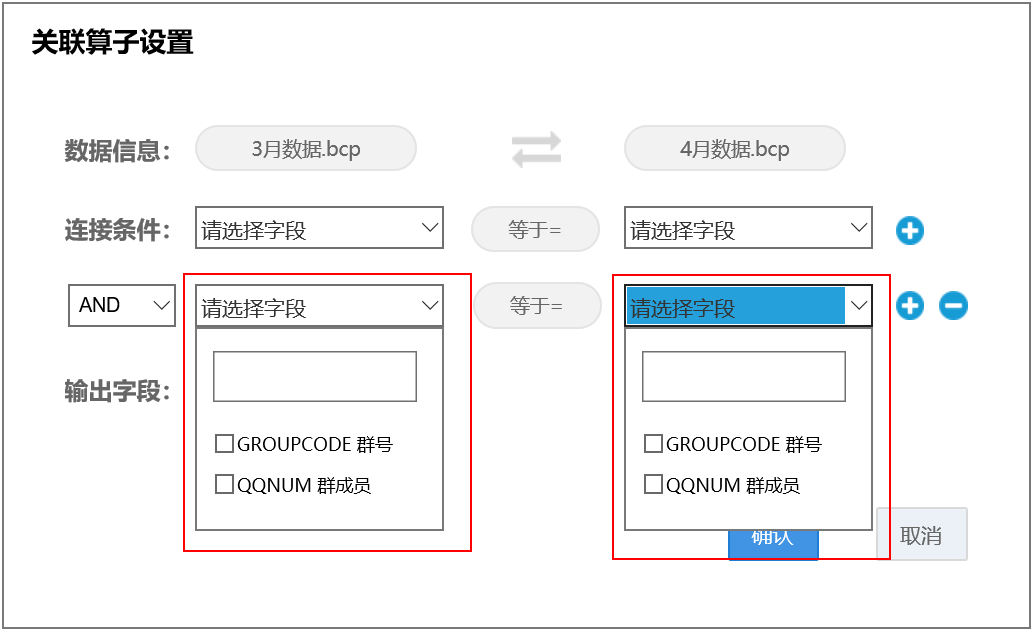


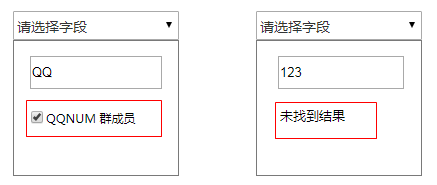
根据选择的关联条件将两个数据表进行连接，默认左连接。

用户根据数据表选择设置连接条件(等于=)，可设置多个连接条件（AND、OR）；

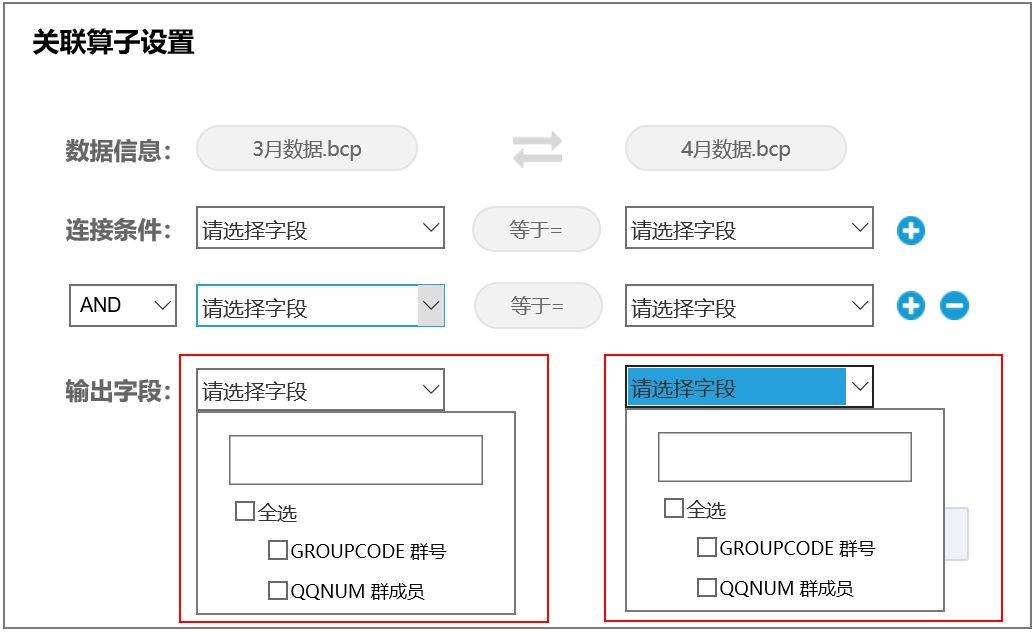


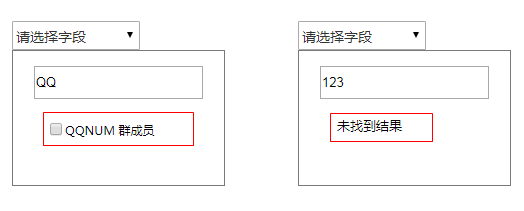
连接条件选择字段为单选框，支持用户输入查找字段，若查找到匹配项，则展示匹配结果进行勾选；若无匹配项，显示“未找到结果”。





用户分别选择两表中需要输出字段，获得关联结果；默认左连接，用户可在画布上调整画线的连接更改左右表的连接次序。输出字段选择为复选框，可选择多个字段进行结果输出。支持用户输入查找输出字段，若查找到匹配项，则展示匹配结果进行勾选（遇到进行多次查找的情况时，保留之前查找勾选结果）；若无匹配项，显示“未找到结果”。





#### 1.2 算子逻辑

1. **连接条件（等于=）**

左表3月数据：



右表4月数据：



连接条件：



连接结果：



输出字段：



输出结果：



1. **连接条件（小于等于≤）**

连接条件：



连接结果：



输出字段：



输出结果：



1. **连接条件（大于等于≥）**

左表4月数据：



右表3月数据：



连接条件：



连接结果：



输出字段：



输出结果：



### 2 取并集算子

#### 2.1 算子设置

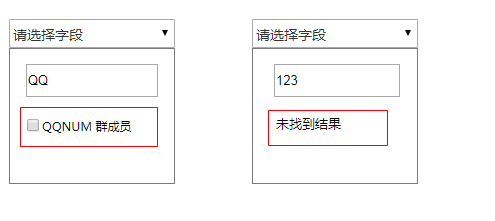


对两个数据表的选择字段进行合并。

用户通过数据表设置取并集条件，可设置多个取并集条件（默认AND）；



并集条件选择字段为单选框，支持用户输入查找字段，若查找到匹配项，则展示匹配结果进行勾选；若无匹配项，显示“未找到结果”。



考虑到两表中字段属性一致但字段名不同的情况，故可设置选择字段别名；别名输入框默认显示“别名”二字，鼠标点击“别名”消失，用户输入别名，输出结果显示为用户设置的别名；若用户未输入字段别名，则默认左表字段名输出。



用户选择是否去除重复项（单选项）后，获得取并集结果。



#### 2.2 算子逻辑

左表3月数据：



右表4月数据：



并集条件：



去重后输出结果：



### 3 碰撞算子

#### 3.1 算子设置



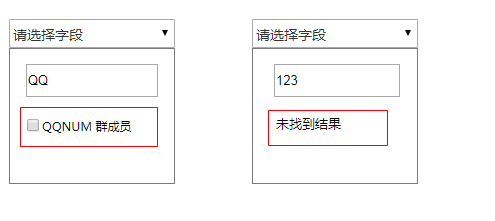
对两个数据表的选择字段进行取交集，输出两表都存在的数据。

碰撞算子即取交集算子，用户根据数据表设置两表中取交集的字段，可选择多个取交集字段（AND、OR）；



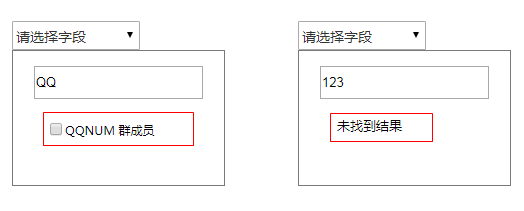
碰撞条件选择字段为单选框，支持用户输入查找字段，若查找到匹配项，则展示匹配结果进行勾选；若无匹配项，显示“未找到结果”。





选择输出字段，输出字段为复选框，可选择多个字段。支持用户输入查找字段，若查找到匹配项，则展示匹配结果进行勾选（遇到进行多次查找的情况时，保留之前查找勾选结果）；若无匹配项，显示“未找到结果”。





以左表为基表，根据用户选择的取交集条件将两表进行取交集计算；输出字段默认只能选择左表字段，根据用户选择的输出字段输出取交集结果。用户可通过在画布上交换数据表连接次序，选择输出字段，分别对两张表中取交集结果进行查看。

#### 3.2 算子逻辑

3月数据：



4月数据：



碰撞条件：



去重后输出结果：



### 4 取差集算子

#### 4.1 算子设置



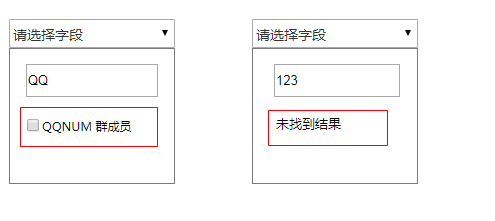
取差集算子比较查找左表存在而右表不存在的数据行。

以左表为基表，选择两表中取差集的字段，可选择多个取差集字段（AND、OR）；



差集条件选择字段为单选框，支持用户输入查找字段。若查找到匹配项，则展示匹配结果进行勾选；若无匹配项，显示“未找到结果”。

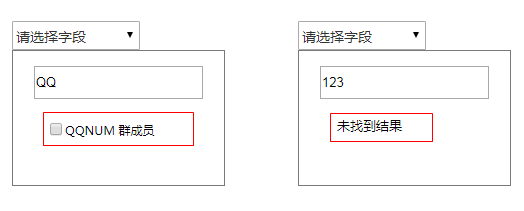




根据选择取差条件将两表进行比较计算；输出字段选择默认只能选择左表字段，若用户想要选择右表字段输出结果，可在画布上调整数据表的连接次序。

输出字段为复选框，可选择多个字段。支持用户输入查找字段，若查找到匹配项，则展示匹配结果进行勾选（遇到进行多次查找的情况时，保留之前查找勾选结果）；若无匹配项，显示“未找到结果”。





#### 4.2 算子逻辑

左表3月数据：



右表4月数据：



取差条件：



取差结果：



### 5 平均值算子

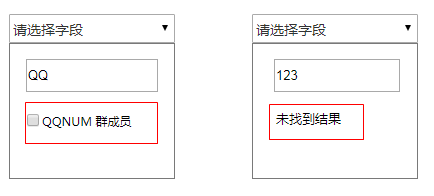
#### 5.1 算子设置



计算选择字段的平均值。

用户根据数据表选择需要计算平均值的字段，取平均值选择字段为单选框，支持用户输入查找字段，若查找到匹配项，则展示匹配结果进行勾选；若无匹配项，显示“未找到结果”。





根据选择字段计算平均值，以表格形式输出平均值结果。

#### 5.2 算子逻辑

3月数据：



取平均值字段：



结果：



### 6 取最大值算子

#### 6.1 算子设置

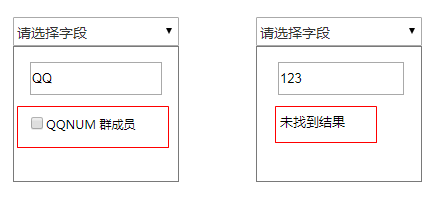


提取选择字段的最大值，查看相关数据信息。

用户通过数据表选择取最大值字段，取最大值选择字段为单选框，支持用户输入查找字段，若查找到匹配项，则展示匹配结果进行勾选；若无匹配项，显示“未找到结果”。

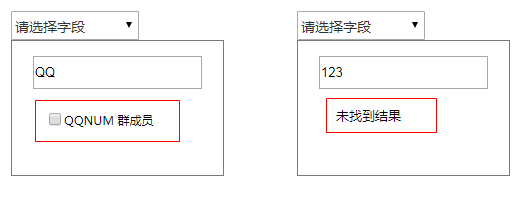
根据用户选择的取最大值字段，找出最大值所在行数据。





用户可选择输出字段，获得输出结果。输出字段为复选框，可选择多个字段进行结果输出；支持用户输入查找字段，若查找到匹配项，则展示匹配结果进行勾选（遇到进行多次查找的情况时，保留之前查找勾选结果）；若无匹配项，显示“未找到结果”。





#### 6.2 算子逻辑

3月数据：



取最大值字段：



输出字段：



输出结果：



### 7 取最小值算子

#### 7.1 算子设置



提取选择字段的最小值，查看相关数据信息。

用户通过数据表选择取最小值字段，取最小值选择字段为单选框，支持用户输入查找字段，若查找到匹配项，则展示匹配结果进行勾选；若无匹配项，显示“未找到结果”。

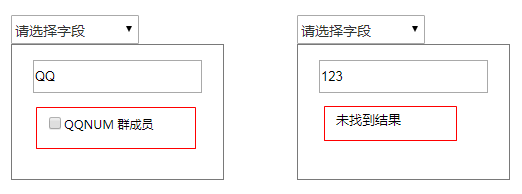
根据用户选择的取最小值字段，找出最小值所在行数据。





用户可选择输出字段，获得输出结果。输出字段为复选框，可选择多个字段进行结果输出；支持用户输入查找字段，若查找到匹配项，则展示匹配结果进行勾选（遇到进行多次查找的情况时，保留之前查找勾选结果）；若无匹配项，显示“未找到结果”。





#### 7.2 算子逻辑

3月数据：



取最大值字段：



输出字段：



输出结果：



### 8 随机采样算子

#### 8.1 算子设置



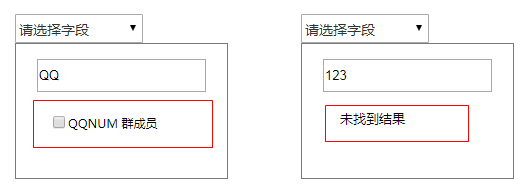
随机输出n条数据。

根据用户手动输入的需要查看的数据条数，随机选择数据行，若用户选择随机条数超过数据最大行数，则将所有数据全部输出，提示“随机条数超过最大行数，输出全部数据”；



并根据用户选择的输出字段，输出结果。输出字段为复选框，可选择多个字段进行结果输出；支持用户输入查找字段，若查找到匹配项，则展示匹配结果进行勾选（遇到进行多次查找的情况时，保留之前查找勾选结果）；若无匹配项，显示“未找到结果”。





#### 8.2 算子逻辑

3月数据：



随机条数设置：2

输出字段：



输出结果：



### 9 条件筛选算子

#### 9.1 算子设置

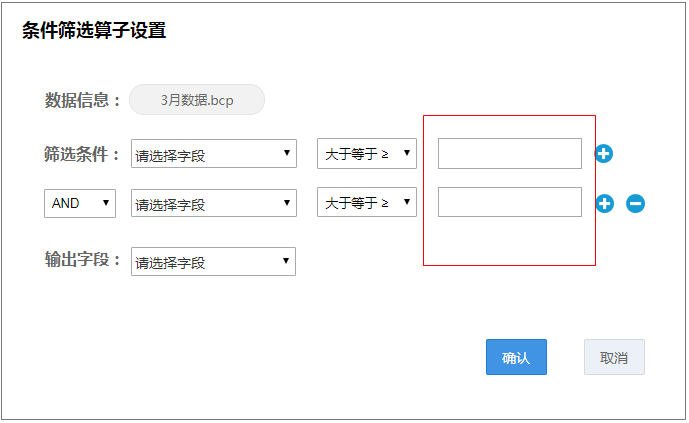


根据过滤条件设置查看符合条件的所在行数据。

用户根据数据表选择过滤字段，可选择多个过滤字段（AND、OR）；



用户手动输入过滤条件；



过滤字段选择字段为单选框，支持用户输入查找字段，若查找到匹配项，则展示匹配结果进行勾选；若无匹配项，显示“未找到结果”。

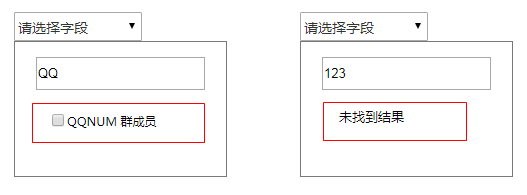




过滤条件有六种类型：大于>、小于<、等于=、不等于≠、大于等于≥，小于等于≤；根据用户输入的过滤条件，对相应字段进行数据过滤，生成过滤结果。

根据用户选择输出字段，输出过滤结果。用户选择单个或多个输出字段；输出字段为复选框，可选择多个字段进行结果输出；支持用户输入查找字段，若查找到匹配项，则展示匹配结果进行勾选（遇到进行多次查找的情况时，保留之前查找勾选结果）；若无匹配项，显示“未找到结果”。





#### 9.2 算子逻辑

3月数据：



过滤条件：



输出字段：



输出结果：



### 10 频率统计算子

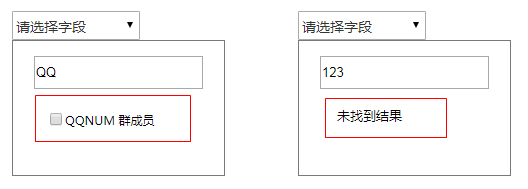
#### 10.1 算子设置



统计选择字段出现次数。

统计字段为复选框，可选择多个字段进行频率统计；支持用户输入查找字段，若查找到匹配项，则展示匹配结果进行勾选（遇到进行多次查找的情况时，保留之前查找勾选结果）；若无匹配项，显示“未找到结果”。





根据用户选择的单个或多个频率统计字段，以及是否去重后进行频率统计，对单个或多个字段相同的数据行进行累加统计，最后根据用户选择的排序方式将所有字段结果输出。

#### 10.2 算子逻辑

3月数据：



统计字段：归属地

经去重、从大到小排序后：



### 11 排序算子

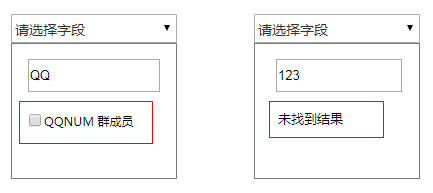
#### 11.1 算子设置



根据选择字段进行排序，支持数据去重。

排序字段为单选框，支持用户输入查找字段，若查找到匹配项，则展示匹配结果进行勾选；若无匹配项，显示“未找到结果”。





根据用户选择的排序字段以及排序方式（升序或降序，选择按钮为单选按钮）对数据表进行排序处理；



用户可选择是否去重，对数据进行去重处理后，根据用户手动输入的输出条数进行结果输出。用户若没有填写输出条数，默认输出所有排序结果；



手动输入从第n行开始，若输入行数<1,则默认从第一行开始。



手动输入输出条数若超过数据最大行数，则显示“输出行数超过最大行数”。



#### 11.2 算子逻辑

3月数据：



排序字段：年龄

从小到大、经去重排序后，输出第2行到第4行，输出结果：



### 12 分组算子

#### 12.1 算子设置

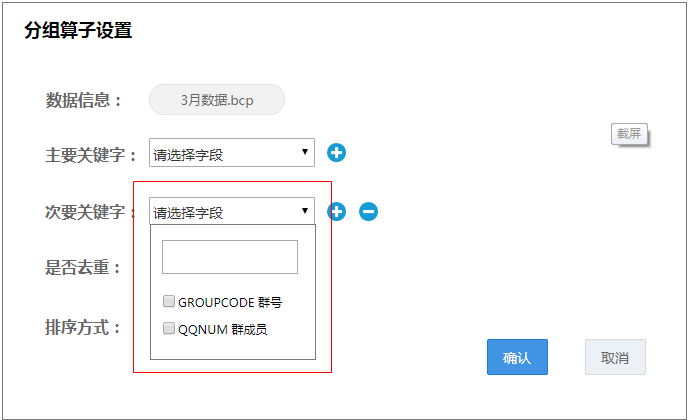


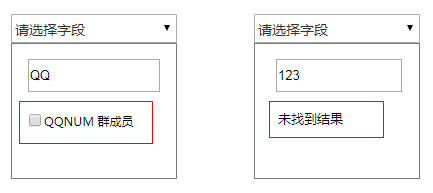
根据选择字段对文本进行分组展示。

用户根据数据表选择分组关键字，可选择多个分组关键字（AND）。默认选择第一个分组关键字为主要关键字，后续添加的关键字为次要关键字；依次按照主要关键字、次要关键字对文本进行排序分组。



分组关键字为单选框，支持用户输入查找字段，若查找到匹配项，则展示匹配结果进行勾选；若无匹配项，显示“未找到结果”。





按照主要关键字、次要关键字对文本进行排序分组后，根据用户选择的排序方式、去重方式计算后，进行结果输出。（将分组选择字段按主要关键字、次要关键字放在列表最前）。



#### 12.2 算子逻辑

3月数据：



主要关键字：归属地

次要关键字：年龄

经去重操作、排序方式选择从小到大排序后，输出结果（将分组选择字段按主要关键字、次要关键字放在列表最前）：



### 13 AI实践算子

#### 13.1 算子设置

AI实践算子主要是完成算子运算展示功能。用户可选择输入文件、输出文件，可设置执行时间和输出字段。AI实践算子为一元自定义算子。

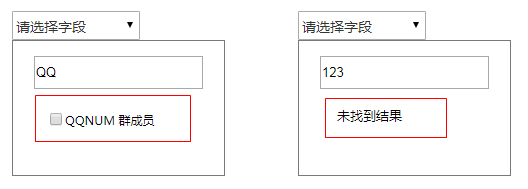


两种执行时间方式可供用户选择，若选择固定秒数，则默认执行30秒（文本框默认显示30秒，光标点击数字消失），用户也可手动输入固定执行秒数；若选择随机秒数，用户要手动输入随机区间，后台根据用户设置的秒数区间先选择随机数，再执行运算。



输出字段为复选框，可选择多个字段进行结果输出；支持用户输入查找字段，若查找到匹配项，则展示匹配结果进行勾选（遇到进行多次查找的情况时，保留之前查找勾选结果）；若无匹配项，显示“未找到结果”。





对于结果文件，用户可点击“浏览”按钮从本机中直接选择输出文件，并在文本框中显示选择文件名。



### 14 多源算子

#### 14.1 算子设置

多源算子主要是完成算子运算展示功能。用户可选择输入文件、输出文件，可设置执行时间和输出字段。多源算子二元自定义算子。

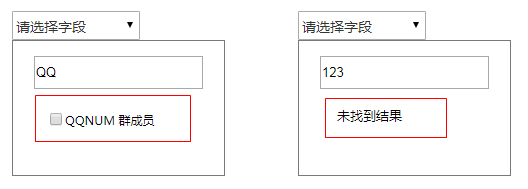


两种执行时间方式可供用户选择，若选择固定秒数，则默认执行30秒（文本框默认显示30秒，光标点击数字消失），用户也可手动输入固定执行秒数；若选择随机秒数，用户要手动输入随机区间，后台根据用户设置的秒数区间先选择随机数，再执行运算。



输出字段为复选框，可选择多个字段进行结果输出；支持用户输入查找字段，若查找到匹配项，则展示匹配结果进行勾选（遇到进行多次查找的情况时，保留之前查找勾选结果）；若无匹配项，显示“未找到结果”。





对于结果文件，用户可点击“浏览”按钮从本机中直接选择输出文件，并在文本框中显示选择文件名。



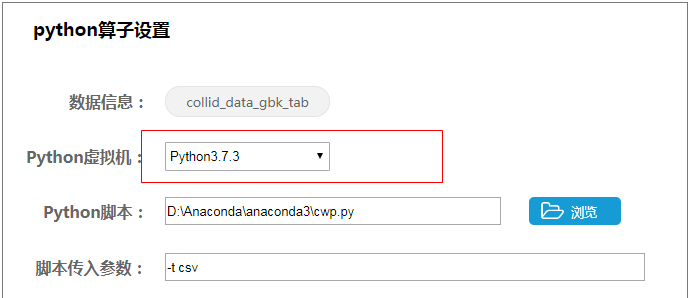
### 15 Python算子

#### 15.1 算子设置

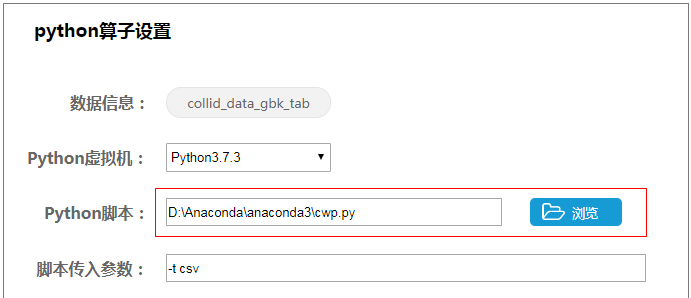
Python算子主要是用来处理原始数据较复杂，需要执行Python脚本生成结果文件的数据。



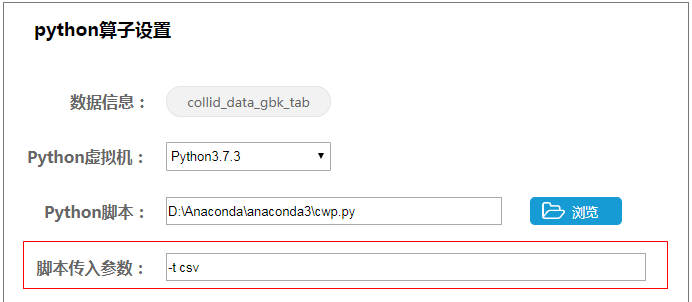
Python虚拟机选择为下拉单选框，用户在python引擎中将本机存在的Python虚拟机导入后，所有Python虚拟机可在下拉列表框中全部显示出来供用户选择。



点击“浏览”按钮可选择导入需要执行的Python脚本。



脚本传入参数中用户可添加设置其它参数，如＂-b＂、＂-o＂、＂-t＂等。



用户需要分别对输入文件和输出结果文件进行设置。

对于输入文件设置，用户有以下两种设置选择：

1. 无输入文件：脚本中已设置好输入文件的路径，可与算子连接的数据源路径不一致；



1. 传参指定输入文件：输入文件为对应连接的数据源路径信息，＂-f1＂默认指代输入文件设置参数。



对于结果文件设置，用户有以下四种设置选择：

1. 无结果文件：Python脚本中已设置好结果文件的路径，算子连接的生成结果为空；



1. 将脚本stdout重定向到结果文件：Python脚本中未设置结果文件的路径，脚本执行结果会全部打印到屏幕上，可将屏幕打印结果重定向到算子连接的结果文件中；



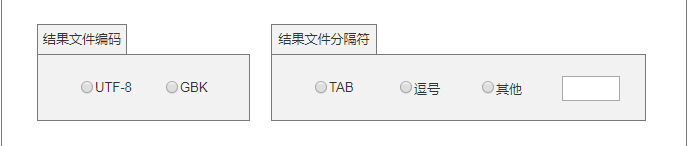
1. 传参指定结果文件：Python脚本中未设置结果文件的路径，算子连接的结果文件路径会自动显示出来，＂-f2＂默认指代输出文件设置参数，脚本执行结果也会直接存入算子连接的结果文件中。



1. 浏览指定结果文件：Python脚本中已设置好结果文件的路径，执行Python脚本时在指定路径中可能会生成多个结果文件，点击“约定”按钮可从生成的多个结果文件中选择一个作为算子的生成结果，便可进行后续连接操作。



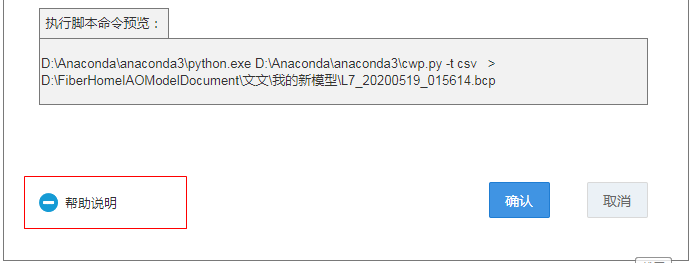
用户还可选择结果文件在数据预览中的文件编码格式和文件分隔符，避免数据预览过程中数据出现乱码的情况，并对数据展示格式进行规整。



根据用户以上对Python脚本的选择设置，自动生成脚本执行命令在执行脚本命令预览框中。



点击帮助说明按钮可显示算子设置的具体说明。



### 16 关键词过滤算子

#### 16.1 算子设置

关键词过滤算子主要用文本关键词对数据字段进行过滤。



数据源输入为双输入，分别为需过滤数据信息和包含文本关键词字段的文件。

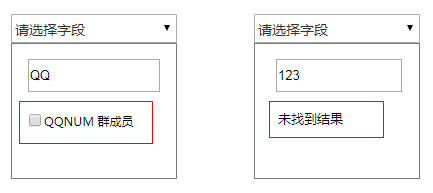


用户设置过滤条件，选择文本过滤字段以及关键词字段，条件可选择“包含”或“不包含”。



文本过滤条件选择字段以及关键词选择字段均为单选框，支持用户输入查找字段，若查找到匹配项，则展示匹配结果进行勾选；若无匹配项，显示“未找到结果”。



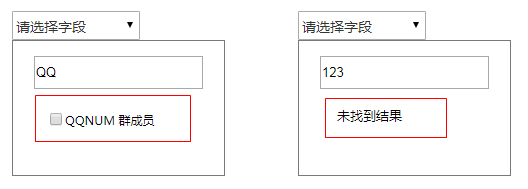


用户可对关键词条件进行预览，关键词行之间默认OR连接，每行空格之间默认AND连接。用括号包含每行AND连接的关键词，默认执行次序为先AND后OR。



输出字段为复选框，可选择多个字段进行结果输出；支持用户输入查找字段，若查找到匹配项，则展示匹配结果进行勾选（遇到进行多次查找的情况时，保留之前查找勾选结果）；若无匹配项，显示“未找到结果”。





### 17 数据标准化算子

#### 17.1 算子设置

当存在数据表中数据字段较多的情况时，用户可选择需要查看的字段，并对选择字段按照选择顺序进行查看。



用户可选择多个字段进行查看。

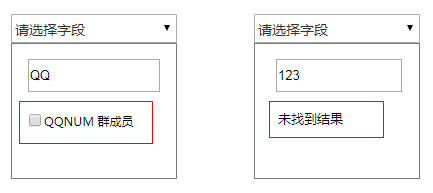


选择字段按照选择顺序进行自动编号。



添加条件选择字段均为单选框，支持用户输入查找字段，若查找到匹配项，则展示匹配结果进行勾选；若无匹配项，显示“未找到结果”。





用户选择查看字段可设置别名，光标移入别名框“别名”二字消失，输入别名信息。

