

数据抽取

1、功能概述

- 1、功能概述
 - 2、入口及界面
 - 3、设置说明
 - 4、数据抽取备份
 - 数据恢复
 - 5、数据抽取示例
 - 5.1 全量抽取
 - 示例效果
 - 设置步骤
 - 5.2 增量抽取
 - 示例效果
 - 设置步骤
 - 5.3 全量抽取和增量抽取区别
- 1、功能概述
 - 2、入口及界面
 - 3、设置说明
 - 4、数据抽取备份
 - 数据恢复
 - 5、数据抽取示例
 - 5.1 全量抽取
 - 示例效果
 - 设置步骤
 - 5.2 增量抽取
 - 示例效果
 - 设置步骤
 - 5.3 全量抽取和增量抽取区别

数据抽取是指从源数据库中抽取原始数据到高速缓存库，它可以保证秒级获取大级别量的数据结果，提高系统性能。

系统支持数据抽取功能的模块有：数据模型、自助数据集、可视化数据集、SQL数据集、原生SQL数据集、存储过程数据集、Java数据集、即席查询、透视分析、加载Excel数据。

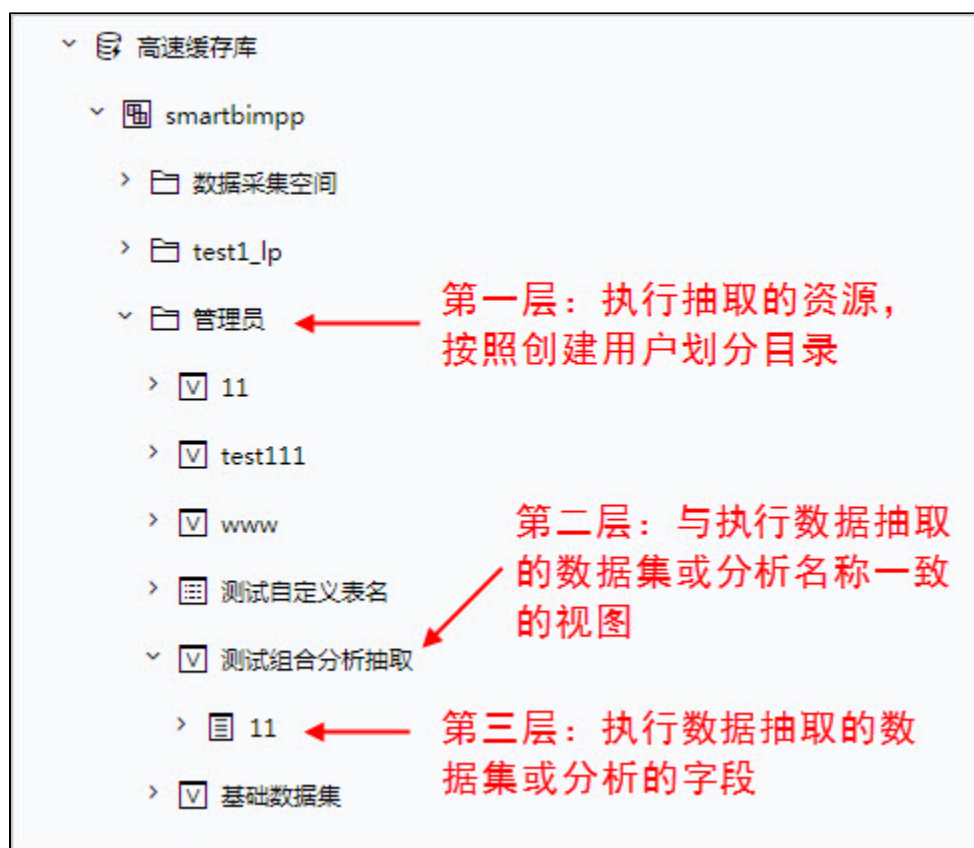


除数据模型以外的数据集的创建入口、资源显示和使用统一通过V10 license控制。

若是希望在V10版本中继续使用除数据模型外的数据集，则需更新授权包含旧数据集的license文件。

数据抽取功能的机制如下：

- 数据集或分析确定结果字段。
- 发起数据抽取指令后，从源数据库中将字段的所有数据抽取到高速缓存库，在高速缓存库的“DEFAULT”节点下生成对应的视图和字段：



- 再次查询当前数据集或分析的数据时，从高速缓存库获取数据。

补充说明

数据集有参数情况下，抽取之后基于该数据集创建的资源不会走mpp。

数据集没有参数的情况下，数据集抽取之后基于此创建的资源会走mpp，然后抽取之后的数据集预览的时候是实时查询不走mpp。



- 1、数据抽取功能必须在当前数据集已保存的前提下才能被激活使用。
- 2、系统支持“可视化数据集”、“即席查询”和“自助数据集”通过数据行权限控制数据抽取的结果。
- 3、数据集抽取时，如果包含参数，则只会抽取参数默认值相关的数据。

2、入口及界面

系统支持数据抽取功能的模块有：数据模型、自助数据集、可视化数据集、SQL数据集、原生SQL数据集、存储过程数据集、Java数据集、即席查询、透视分析、加载Excel数据。

- 数据模型：在“抽取”连接模式下，在表数据区单击 **抽取设置** 按钮；详情可参考文档[数据模型-抽取](#)。

抽取设置

查询列表


查询名称	抽取方式	
产品表	全量	↕
订单表	全量	↕
产品类别表	全量	↕
顾客表	全量	↕
订单明细表	全量	↕

抽取方式：
☒ 全量抽取 ☐ 按次抽取

抽取参数值： 无参数

抽取并建表

确定 取消

 数据模型抽取后的表在“高速缓存库”中不显示出来。

- 即席查询：在已保存的即席查询的编辑界面，单击工具栏上的 **抽取** 按钮（），打开“数据抽取设置”窗口。

数据抽取设置

☐ 实时

☒ 全量抽取 ☒ 清空数据

抽取表设置

抽取表名

高级设置：

异常处理

抽取出错时 ☒ 回滚 ☐ 继续

定时抽取

☒ 启用

设置


☐ 禁用

执行时间：未设置

立即抽取

确定(O)

取消(C)

- 非自助数据集：在已保存的非自助数据集的编辑界面，单击工具栏上的 **抽取** 按钮（），打开“数据抽取设置”窗口。

数据抽取设置

实时

全量抽取

清空数据

增量抽取按时间戳

抽取表设置

抽取表名

高级设置：

异常处理

抽取出错时

回滚

继续

定时抽取

启用

设置

禁用

立即抽取

确定(O)

取消(C)

- 自助数据集：在已保存的自助数据集的编辑界面，先点击工具栏上的 **抽取** 按钮

Test-1

Test-1

筛选器：0 | +

抽取

点击抽取按钮

关于自助数据集的创建入口、资源显示和使用，统一通过V10 license控制。如希望在V10版本中继续使用自助数据集，则需更新授权包含自助数据集的license文件。

再点击弹出的 **设置**按钮，打开“数据抽取设置”窗口。

Test-1

Test-1

状态: 初始

设置

筛选器：0 | +

点击设置

抽取设置

全量抽取

增量抽取按时间戳

清空数据

抽取表设置

抽取表名

高级设置：

异常处理

抽取出错时

回滚

继续

定时抽取

启用

设置

禁用

执行时间：未设置

立即抽取

确定

取消

3、设置说明

“数据抽取”窗口中的设置项说明如下：

界面介绍	分类		功能说明
抽取方式	实时		表示不抽取。其中，自助数据集的不抽取设置通过 实时 按钮（   设置 ）实现。
	全量抽取	清空数据	<div><div><div></div><div>勾选 清空数据时，用户需要在定义数据集时，添加标识符字段用于区分抽取数据的历史版本。详情请参考 数据抽取示例。</div></div></div>
	增量抽取	增量抽数据按时间戳	<div>指与上次抽取结果中最大时间对比，将大于这个时间的数据进行集中抽取。</div> <div><div><div></div><div>数据集中含有时间信息的字段才支持增量抽取。</div></div></div>
		增量字段	表示与上次抽取结果的最大时间进行对比的字段，必须将记录了时间信息的字段做为增量字段。
		时间格式	时间格式用于将非DATE或非DATETIME类型的增量字段进行格式转化。例如：若增量字段为“订单日期”，“订单日期”是“string”类型，数值是“20150101”，则需要设置其时间格式为“YYYYMMDD”。
		增量参数	<div>表示按照选择的参数，抽取该参数相关数据。（需要把需要抽取得表中得最新“日期”值传给参数作为筛选值），抽取筛选后的结果。</div> <div><div><div></div><div>非自助数据集才有“增量参数”设置项，增量参数须为时间信息的字段。</div></div></div>

		忽略抽取当天数据	表示不包含当天的增量数据。
		覆盖最后抽取的N天数据	表示根据时间戳，重新抽取并覆盖高速缓存库中当前自助数据集的最后N天数据。目前只支持Vertica类型的高速缓存库允许“覆盖最后抽取的N天数据”设置项。
抽取表设置	抽取表名		<div>支持自定义表名。</div> <div><ul style="list-style-type: none">• 设置抽取表名：以“输入的表名”作为表名称，“数据集名称”作为表别名。• 不设置抽取表名：默认以“数据集ID”作为表名称，“数据集名称”作为表别名。</div> <div><div></div><div><div>1、输入的表名不支持“@#\$\$%^&*{}[]/”等特殊字符。</div><div>2、抽取到“星环”和“Presto+Hive”高速缓存库，输入的表名不支持“中文”。</div></div></div>
	高级设置	分区字段 分区类型	<div>“分区字段”和“分区类型”设置项用于设置集群的数据抽取，若未选择时间字段，则不使用Smartbiamp集群建表。</div> <div>“分区字段”要求为日期类型的。</div> <div><div></div><div><div>1、在不清空数据的情况下，集群多次抽取以首次选择的分区类型为准。如首次抽取分区类型选择“年”，再次抽取且不清空数据时，分区字段选择“月/日”，抽取后表数据追加，仍按年分区。</div><div>2、如果mpp是单机部署的，则无论是否有设置分区字段，都默认使用Log引擎创建表；如果mpp是集群部署的，则选择分区字段后，将使用MergeTree引擎创建表。</div></div></div>
异常处理	回滚		表示返回到数据抽取前的状态。
	继续		表示继续抽取下一条数据，并将这条错误数据写入异常日志，供用户下载查看异常原因。
定时抽取	启用		表示启用定时抽取，选择 启用 同时出现“设置”设置项，点击 设置 可定制计划任务，根据时间计划将数据定时抽取到高速缓存，详情请参见 计划 章节。
	禁用		表示不启用定时抽取。
	执行时间		简述定时抽取的任务内容，当定制完定时抽取的任务之后，会自动生成。
执行用户	资源创建者		表示当前自助数据集的创建用户，将只抽取该创建用户拥有的数据行权限内的数据。数据行权限详情请参考 数据权限 。
	特定用户		表示指定抽取的用户，通过用户名和密码指定，将抽取该指定用户拥有的数据行权限内的数据。数据行权限详情请参考 数据权限 。
排序字段			用于多线程抽取时对数据进行排序，避免抽取的数据重复。
立即抽取			表示立即抽取数据到高速缓存库。
上表中的“执行用户”设置项用于保证：只允许抽取资源创建者数据行权限内的数据。目前只有“可视化数据集”、“即席查询”和“自助数据集”的数据抽取受数据行权限控制。			



即席查询的数据抽取功能不支持“增量抽取”。

4、数据抽取备份

1) 在系统运维的 **系统选项>高级设置** 中，设置项 **“BACKUP_TAB_RETAIN_NUM”**可设置在数据库中保留的抽取表的个数，默认为5个，详情请参考 [系统选项-高级设置](#) 。

2) 在数据抽取中只要清空抽取表的数据，系统就会自动备份。

数据恢复

如果发生数据丢失，可通过手工恢复，即编写SQL语句或使用数据库工具，将某个备份表的数据还原替换到原始表中。SQL语句如下：

```
INSERT INTO table2
SELECT * FROM table1;
```

SQL语句详情请参考 [INSERT INTO SELECT](#) 。

5、数据抽取示例

5.1 全量抽取

全量抽取表示把数据全部抽取，抽取时，先将数据抽取到临时表；抽取成功后，drop原始表，rename到目标表。

- 示例效果

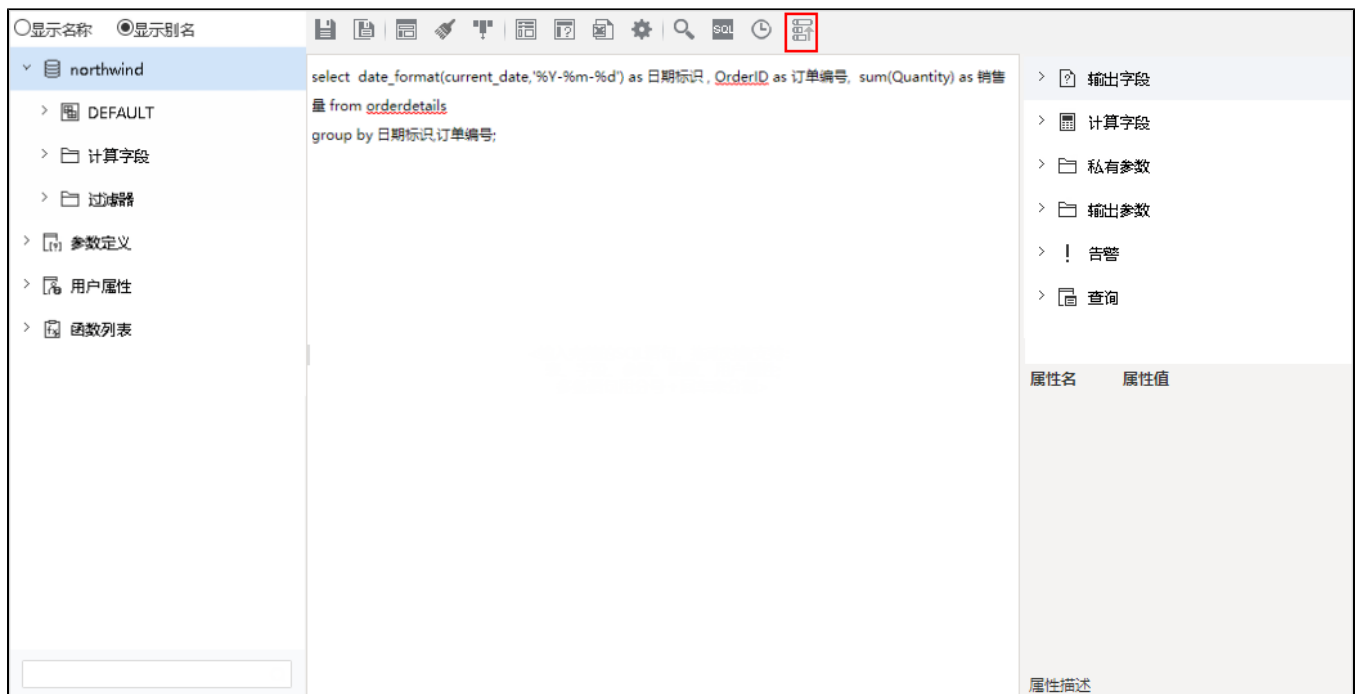
在高速缓存库中浏览该资源的数据，添加了“日期标识”字段，用日期来区分不同时间抽取的数据。结果如图，包括了2018-12-19和2018-12-20抽取的数据：

日期标识	订单编号	销售量
2018-12-19	10249	49
2018-12-19	10250	60
2018-12-19	10251	41
2018-12-19	10252	105
2018-12-19	10253	102
2018-12-19	10254	57
2018-12-20	10249	49
2018-12-20	10250	60
2018-12-20	10251	41
2018-12-20	10252	105
2018-12-20	10253	102
2018-12-20	10254	57
2018-12-20	10255	110
2018-12-20	10256	27
2018-12-20	10257	46

- 设置步骤

第一步：抽取2018-12-19的数据。

1) 点击SQL数据集工具栏的 **数据抽取** 按钮，如图：



2) 弹出“数据抽取设置”界面，选择“全量抽取”后，点击 **立即抽取** ；



第二步：抽取2018-12-20的数据。

1) 点击SQL数据集工具栏的 **数据抽取** 按钮进行抽取。

2) 弹出“数据抽取设置”界面，选择 **全量抽取** ，勾除 **清空数据** 后，点击 **立即抽取** ；

数据抽取设置

实时

全量抽取

清空数据

增量抽取按时间戳

抽取表设置

抽取表名

高级设置：

异常处理

抽取出错时

回滚

继续

定时抽取

启用

设置


禁用

立即抽取

确定(O)

取消(C)

第三步：进行数据预览。

1) 在高速缓存库找到该资源，选中该资源的更多操作  中，选择 **数据集监控管理 > 浏览数据**，如图：

2) 浏览数据效果如图：

日期标识	订单编号	销售量
2018-12-19	10249	49
2018-12-19	10250	60
2018-12-19	10251	41
2018-12-19	10252	105
2018-12-19	10253	102
2018-12-19	10254	57
2018-12-20	10249	49
2018-12-20	10250	60
2018-12-20	10251	41
2018-12-20	10252	105
2018-12-20	10253	102
2018-12-20	10254	57
2018-12-20	10255	110
2018-12-20	10256	27
2018-12-20	10257	46



当选择“全量抽取”并勾选“清空数据”时，用户需要在定义数据集时，添加标识符字段用于区分抽取数据的历史版本。

5.2 增量抽取

• 示例效果

浏览增量抽取前高速缓存库内数据，如下图所示：

表名称:	Test_1	表别名:	Test_1
data	id	quantity	
20201109	1.00	50.00	
20201109	2.00	51.00	
20201109	3.00	52.00	

增量抽取完成后，在高速缓存库中浏览该资源的数据，结果如下图所示：

表名称:	18a8a90130175aa97aa97d7e80175ac096afe04f2	表别名:	Test-1
data	id	quantity	
20201109	1.00	50.00	
20201109	2.00	51.00	
20201109	3.00	52.00	
20201110	NaN	NaN	
20201110	4.00	53.00	

增量抽取的数据



增量抽取时对比上次抽取结果的最大时间，仅抽取了大于上次抽取时间的20201110的数据，忽略了20201109的数据。

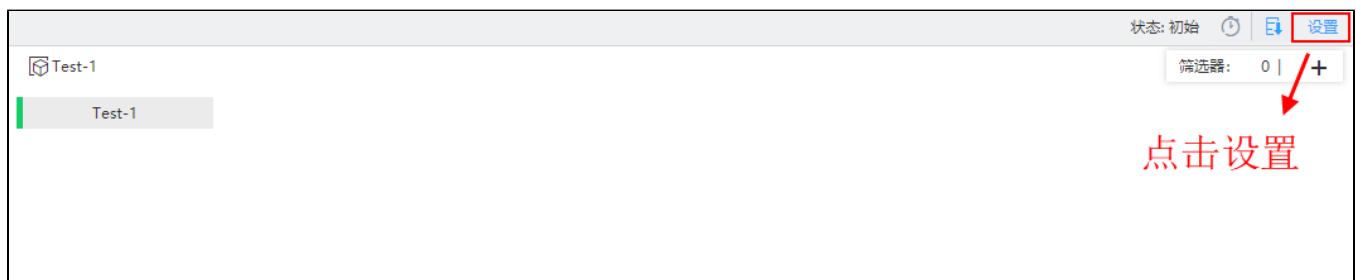
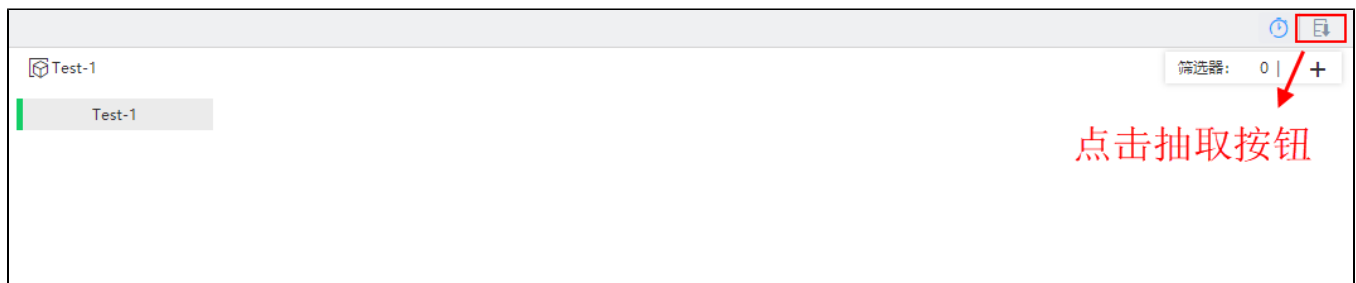
• 设置步骤

第一步：首次增量抽取

1) 自助数据集中浏览初始数据

   <input type="checkbox"/> 立即刷新			<input type="checkbox"/> 显示隐藏字段 <input checked="" type="checkbox"/> 显示别名 100 行
As Test-1 data	# Test-1 id	# Test-1 quantity	
20201109	1.00	50.00	
20201109	2.00	51.00	
20201109	3.00	52.00	

2) 点击抽取按钮再点击弹出的设置按钮



3) 进行相关增量抽取的设置

抽取设置

全量抽取

增量抽取按时间戳

增量字段

Test-1 / data

时间格式

YYYYMMDD

☐ 忽略抽取当天数据

☐ 覆盖最后抽取的

0

天数据

抽取表设置

抽取表名

Test_1

异常处理

抽取出错时

回滚

继续

定时抽取

立即抽取

确定

取消

⚠ 当选择“增量抽取按时间戳”时，数据集中需含有时间信息的字段作为增量字段。

4) 抽取成功后前往高速缓存库相应路径下预览数据

The screenshot shows the 'High-Speed Cache Library' (高速缓存库) interface. The left sidebar displays a tree structure under the 'DEFAULT' folder, including '数据采集空间', '管理员', and '1'. The 'Test_1' folder is selected, and a context menu is open, showing options: '打开(O)', '预览数据(A)', '设置数据权限', '调度管理', and '删除(D)'. The '预览数据(A)' option is highlighted.

表名: Test_1			表别名: Test_1		
data		id	quantity		
20201109		1.00	50.00		
20201109		2.00	51.00		
20201109		3.00	52.00		



首次增量抽取时等同于全量抽取。

第二步：第二次增量抽取

1) 自助数据集中浏览新增的数据

立即刷新			显示隐藏字段 <input type="checkbox"/> 显示别名 <input checked="" type="checkbox"/> 100 行		
Test-1 data		# Test-1 id	# Test-1 quantity		
20201109		1.00	50.00		
20201109		2.00	51.00		
20201109		3.00	52.00		
20201110					
20201110		4.00	53.00		

2) 进行增量抽取设置

抽取设置

☐ 全量抽取

☒ 增量抽取按时间戳

当前已抽取最大时间为: 20201109

增量字段 Test-1 / data 时间格式 YYYYMMDD

☐ 忽略抽取当天数据 ☐ 覆盖最后抽取的 0 天数据

抽取表设置

抽取表名 Test_1

异常处理

抽取出错时 ☒ 回滚 ☐ 继续

定时抽取

立即抽取

确定 取消

3) 查看抽取日志



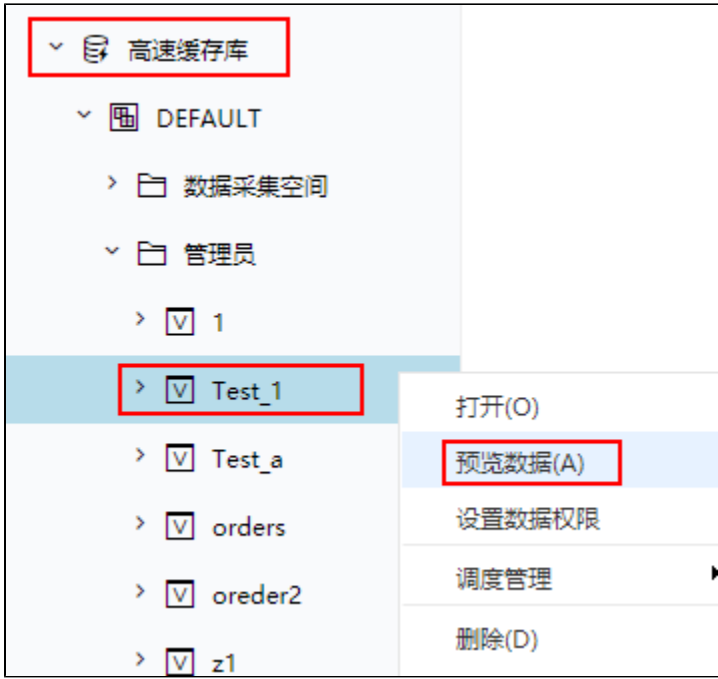
日志[数据集名称: Test_1]

开始时间	结束时间	日志	抽取时间	类型	状态	抽取条数	抽取失败信息
2020-11-09 16:21:24	2020-11-09 16:21:26	数据抽取成功	2秒	增量按时间戳	同步完成	2	---
2020-11-09 16:06:34	2020-11-09 16:06:35	数据抽取成功,由于表不存在,本次为全量抽取	1秒	增量按时间戳	同步完成	3	---

确定(O) 取消(C)

第三步：进行数据预览

1) 抽取成功后对应路径下预览数据



2) 浏览数据效果如图:

表名称:	8a8a90130175aa97d7e80175ac096afe04f2	表别名:	Test-1
data	id	quantity	
20201109	1.00	50.00	
20201109	2.00	51.00	
20201109	3.00	52.00	
20201110	NaN	NaN	
20201110	4.00	53.00	

增量抽取的数据

5.3 全量抽取和增量抽取区别

全量抽取	抽取所有数据
增量抽取	指与上次抽取结果中最大时间对比，将大于这个时间的数据进行集中抽取。

1、功能概述

- 1、功能概述
 - 2、入口及界面
 - 3、设置说明
 - 4、数据抽取备份
 - 数据恢复
 - 5、数据抽取示例
 - 5.1 全量抽取
 - 示例效果
 - 设置步骤
 - 5.2 增量抽取
 - 示例效果
 - 设置步骤
 - 5.3 全量抽取和增量抽取区别
- 1、功能概述
 - 2、入口及界面
 - 3、设置说明
 - 4、数据抽取备份
 - 数据恢复
 - 5、数据抽取示例
 - 5.1 全量抽取
 - 示例效果
 - 设置步骤
 - 5.2 增量抽取
 - 示例效果
 - 设置步骤

数据抽取是指从源数据库中抽取原始数据到高速缓存库，它可以保证秒级获取大级别量的数据结果，提高系统性能。

系统支持数据抽取功能的模块有：数据模型、自助数据集、可视化数据集、SQL数据集、原生SQL数据集、存储过程数据集、Java数据集、即席查询、透视分析、加载Excel数据。

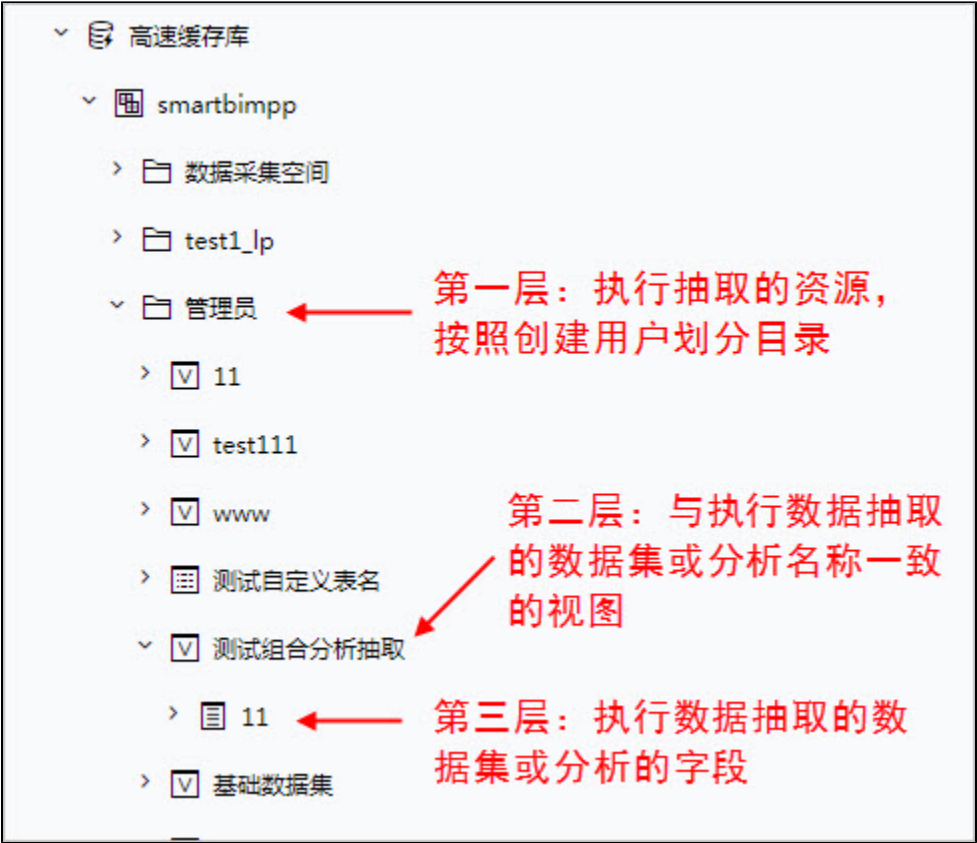


除数据模型以外的数据集的创建入口、资源显示和使用统一通过V10 license控制。

若是希望在V10版本中继续使用除数据模型外的数据集，则需更新授权包含旧数据集的license文件。

数据抽取功能的机制如下：

- 数据集或分析确定结果字段。
- 发起数据抽取指令后，从源数据库中将字段的所有数据抽取到高速缓存库，在高速缓存库的“DEFAULT”节点下生成对应的视图和字段：



- 再次查询当前数据集或分析的数据时，从高速缓存库获取数据。



补充说明

数据集有参数情况下，抽取之后基于该数据集创建的资源不会走mpp。

数据集没有参数的情况下，数据集抽取之后基于此创建的资源会走mpp，然后抽取之后的数据集预览的时候是实时查询不走mpp。



- 1、数据抽取功能必须在当前数据集已保存的前提下才能被激活使用。
- 2、系统支持“可视化数据集”、“即席查询”和“自助数据集”通过数据行权限控制数据抽取的结果。
- 3、数据集抽取时，如果包含参数，则只会抽取参数默认值相关的数据。

2、入口及界面

系统支持数据抽取功能的模块有：数据模型、自助数据集、可视化数据集、SQL数据集、原生SQL数据集、存储过程数据集、Java数据集、即席查询、透视分析、加载Excel数据。

- 数据模型：在“抽取”连接模式下，在表数据区单击 **抽取设置** 按钮；详情可参考文档[数据模型-抽取](#)。

抽取设置

查询列表

查询名称	抽取方式	
产品表	全量	
订单表	全量	
产品类别表	全量	
顾客表	全量	
订单明细表	全量	

抽取方式：☒ 全量抽取 ☐ 按次抽取

抽取参数值：无参数

抽取并建宽表

确定取消



数据模型抽取后的表不会出现的“高速缓存库”。

- 即席查询：在已保存的即席查询的编辑界面，单击工具栏上的 **抽取** 按钮（），打开“数据抽取设置”窗口。

数据抽取设置

☐ 实时

☒ 全量抽取 ☒ 清空数据

抽取表设置

抽取表名

高级设置：

异常处理

抽取出错时

☒ 回滚 ☐ 继续

定时抽取

☒ 启用

设置

☐ 禁用

执行时间：未设置

立即抽取

确定(O)

取消(C)

- 非自助数据集：在已保存的非自助数据集的编辑界面，单击工具栏上的 **抽取** 按钮（），打开“数据抽取设置”窗口。

数据抽取设置

实时

全量抽取

清空数据

增量抽取按时间戳

抽取表设置

抽取表名

高级设置：

异常处理

抽取出错时

回滚

继续

定时抽取

启用

设置

禁用

立即抽取

确定(O)

取消(C)

- 自助数据集：在已保存的自助数据集的编辑界面，先点击工具栏上的 **抽取** 按钮

Test-1

Test-1

筛选器：0 | +

抽取

点击抽取按钮

关于自助数据集的创建入口、资源显示和使用，统一通过V10 license控制。如希望在V10版本中继续使用自助数据集，则需更新授权包含自助数据集的license文件。

再点击弹出的 **设置**按钮，打开“数据抽取设置”窗口。

Test-1

Test-1

状态: 初始

设置

筛选器：0 | +

点击设置

		覆盖最后抽取的N天数据	表示根据时间戳，重新抽取并覆盖高速缓存库中当前自助数据集的最后N天数据。目前只支持Vertica类型的高速缓存库允许“覆盖最后抽取的N天数据”设置项。 <div><div> 只有自助数据集才有该设置项。</div></div>
抽取表设置	抽取表名		支持自定义表名。 <ul style="list-style-type: none">设置抽取表名：以“输入的表名”作为表名称，“数据集名称”作为表别名。不设置抽取表名：默认以“数据集ID”作为表名称，“数据集名称”作为表别名。 <div><div> 1、输入的表名不支持“@#%`&*{}[]/”等特殊字符。</div><div> 2、抽取到“星环”和“Presto+Hive”高速缓存库，输入的表名不支持“中文”。</div></div>
	高级设置	分区字段 分区类型	“分区字段”和“分区类型”设置项用于设置集群的数据抽取，若未选择时间字段，则不使用Smartbiamp集群建表。 “分区字段”要求为日期类型的。 <div><div> 1、在不清空数据的情况下，集群多次抽取以首次选择的分区类型为准。如首次抽取分区类型选择“年”，再次抽取且不清空数据时，分区字段选择“月/日”，抽取后表数据追加，仍按年分区。</div><div> 2、如果mpp是单机部署的，则无论是否有设置分区字段，都默认使用Log引擎创建表；如果mpp是集群部署的，则选择分区字段后，将使用MergeTree引擎创建表。</div></div>
异常处理	回滚		表示返回到数据抽取前的状态。
	继续		表示继续抽取下一条数据，并将这条错误数据写入异常日志，供用户下载查看异常原因。
定时抽取	启用		表示启用定时抽取，选择 启用 同时出现“设置”设置项，点击 设置 可定制计划任务，根据时间计划将数据定时抽取到高速缓存，详情请参见 计划 章节。
	禁用		表示不启用定时抽取。
	执行时间		简述定时抽取的任务内容，当定制完定时抽取的任务之后，会自动生成。
执行用户	资源创建者		表示当前自助数据集的创建用户，将只抽取该创建用户拥有的数据行权限内的数据。数据行权限详情请参考 数据权限 。
	特定用户		表示指定抽取的用户，通过用户名和密码指定，将抽取该指定用户拥有的数据行权限内的数据。数据行权限详情请参考 数据权限 。
排序字段			用于多线程抽取时对数据进行排序，避免抽取的数据重复。
立即抽取			表示立即抽取数据到高速缓存库。
上表中的“执行用户”设置项用于保证：只允许抽取资源创建者数据行权限内的数据。目前只有“可视化数据集”、“即席查询”和“自助数据集”的数据抽取受数据行权限控制。			
<div><div> 即席查询的数据抽取功能不支持“增量抽取”。</div></div>			

4、数据抽取备份

- 在系统运维的 **系统选项>高级设置** 中，设置项 **“BACKUP_TAB_RETAIN_NUM”** 可设置在数据库中保留的抽取表的个数，默认为5个，详情请参考 [系统选项-高级设置](#) 。
- 在数据抽取中只要清空抽取表的数据，系统就会自动备份。

抽取设置

全量抽取

清空数据

增量抽取按时间戳

抽取表设置

抽取表名

高级设置：

异常处理

抽取出错时

回滚

继续

定时抽取

启用

设置

禁用

执行时间：从 2021-03-18 起，每 1 天，01:00

立即抽取

确定

取消

3) 每抽取一次在数据库中就会备份一张表（第一次除外），如图：

数据库导航

项目

输入表格名称的一部分

新连接 3

aaa

aaab

abc

default

haid_fact

haid_source

haidmpp

smartbimpp

表

011dab600ccd43e5a78dd55d917abcea_etl

01ec14d3fcd64c66b85762225f1586d3_etl

0318

0318_bak1

0318_bak2

0318_bak3

0318_bak4

0318_bak5

111

111111

Project - General

Name

DataSource

<none> Script

0318_bak1

属性 数据 ER 图

0318_bak1 | 输入一个 SQL 表达式来过滤结果 (使用 Ctrl+Space)

网格	ABC CategoryName	ABC ShipRegion	123 Freight
1	点心	华南	12.76
2	饮料	华南	12.76
3	海鲜	华南	12.76
4	日用品	华南	0.12
5	点心	华南	0.12
6	日用品	华南	0.12
7	饮料	华南	5.74
8	日用品	华南	5.74
9	调味品	华南	17.68
10	点心	华南	17.68
11	日用品	华南	1.61
12	饮料	华南	1.61
13	肉/家禽	华南	64.5
14	特制品	华南	64.5
15	饮料	华南	64.5
16	调味品	华南	64.86
17	海鲜	华南	64.86

Record

Save Cancel Script

200+ Rows: 1

备份规则：在抽取数据之前，默认备份抽取表（第一次除外）到数据库中，每抽取一次进行一次滚动备份，如果备份表的个数到设置的上限，下次备份则先移除最旧的备份表再进行备份。

数据恢复

如果发生数据丢失，可通过手工恢复，即编写SQL语句或使用数据库工具，将某个备份表的数据还原替换到原始表中。SQL语句如下：

```
INSERT INTO table2
SELECT * FROM table1;
```

SQL语句详情请参考 [INSERT INTO SELECT](#) 。

5、数据抽取示例

5.1 全量抽取

全量抽取表示把数据全部抽取，抽取时，先将数据抽取到临时表；抽取成功后，drop原始表，rename到目标表。

- 示例效果

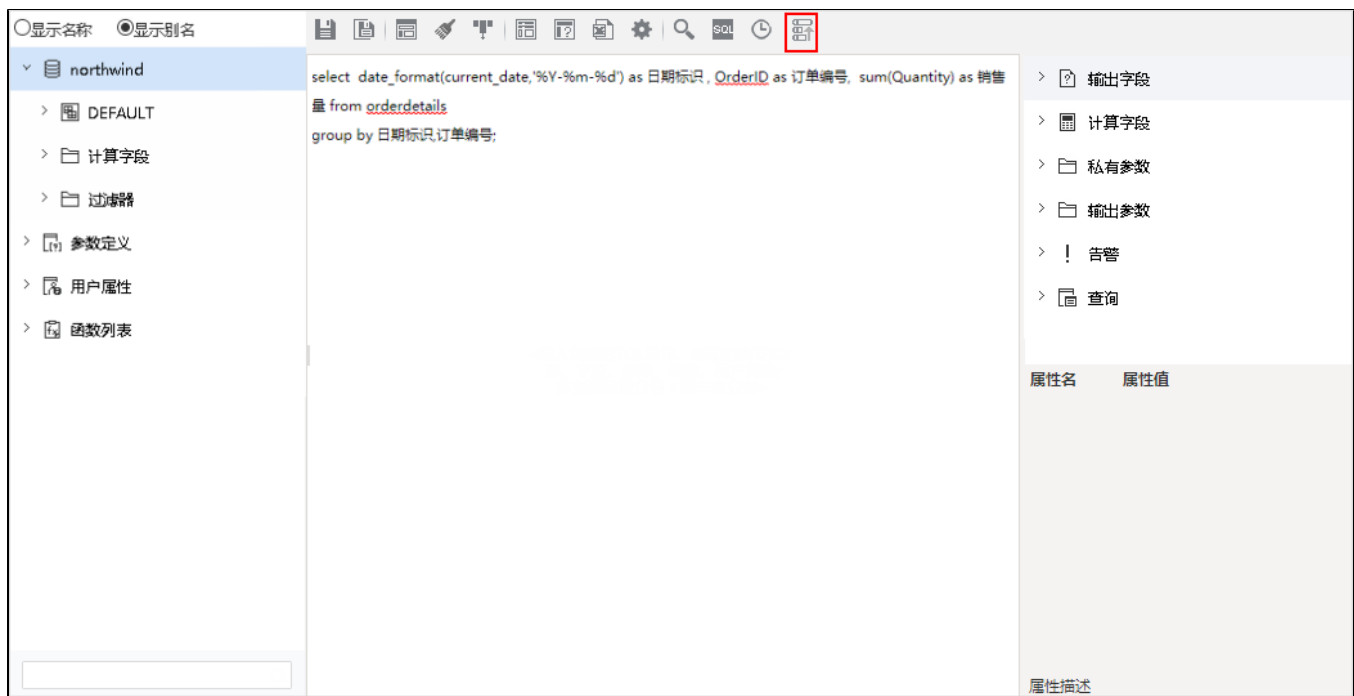
在高速缓存库中浏览该资源的数据，添加了“日期标识”字段，用日期来区分不同时间抽取的数据。结果如图，包括了2018-12-19和2018-12-20抽取的数据：

日期标识	订单编号	销售量
2018-12-19	10249	49
2018-12-19	10250	60
2018-12-19	10251	41
2018-12-19	10252	105
2018-12-19	10253	102
2018-12-19	10254	57
2018-12-20	10249	49
2018-12-20	10250	60
2018-12-20	10251	41
2018-12-20	10252	105
2018-12-20	10253	102
2018-12-20	10254	57
2018-12-20	10255	110
2018-12-20	10256	27
2018-12-20	10257	46

- 设置步骤

第一步：抽取2018-12-19的数据。

1) 点击SQL数据集工具栏的 **数据抽取** 按钮，如图：



2) 弹出“数据抽取设置”界面，选择“全量抽取”后，点击 **立即抽取** ；



第二步：抽取2018-12-20的数据。

1) 点击SQL数据集工具栏的 **数据抽取** 按钮进行抽取。

2) 弹出“数据抽取设置”界面，选择 **全量抽取** ，勾除 **清空数据** 后，点击 **立即抽取** ；

数据抽取设置

实时

全量抽取

清空数据

增量抽取按时间戳

抽取表设置

抽取表名

高级设置：

异常处理

抽取出错时

回滚

继续

定时抽取

启用

设置


禁用

立即抽取

确定(O)

取消(C)

第三步：进行数据预览。

1) 在高速缓存库找到该资源，选中该资源的更多操作  中，选择 **数据集监控管理 > 浏览数据**，如图：

2) 浏览数据效果如图：

日期标识	订单编号	销售量
2018-12-19	10249	49
2018-12-19	10250	60
2018-12-19	10251	41
2018-12-19	10252	105
2018-12-19	10253	102
2018-12-19	10254	57
2018-12-20	10249	49
2018-12-20	10250	60
2018-12-20	10251	41
2018-12-20	10252	105
2018-12-20	10253	102
2018-12-20	10254	57
2018-12-20	10255	110
2018-12-20	10256	27
2018-12-20	10257	46



当选择“全量抽取”并勾选“清空数据”时，用户需要在定义数据集时，添加标识符字段用于区分抽取数据的历史版本。

5.2 增量抽取

• 示例效果

浏览增量抽取前高速缓存库内数据，如下图所示：

导航

预览数据[Test_1]

x

表名称:

Test_1

表别名:

Test_1

data	id	quantity
20201109	1.00	50.00
20201109	2.00	51.00
20201109	3.00	52.00

增量抽取完成后，在高速缓存库中浏览该资源的数据，结果如下图所示：

🏠 导航 | 预览数据[Test-1] x

表名称: l8a8a90130175aa97aa97d7e80175ac096afe04f2

表别名: Test-1

data	id	quantity
20201109	1.00	50.00
20201109	2.00	51.00
20201109	3.00	52.00
20201110	NaN	NaN
20201110	4.00	53.00

增量抽取的数据




增量抽取的数据

增量抽取时对比上次抽取结果的最大时间，仅抽取了大于上次抽取时间的20201110的数据，忽略了20201109的数据。

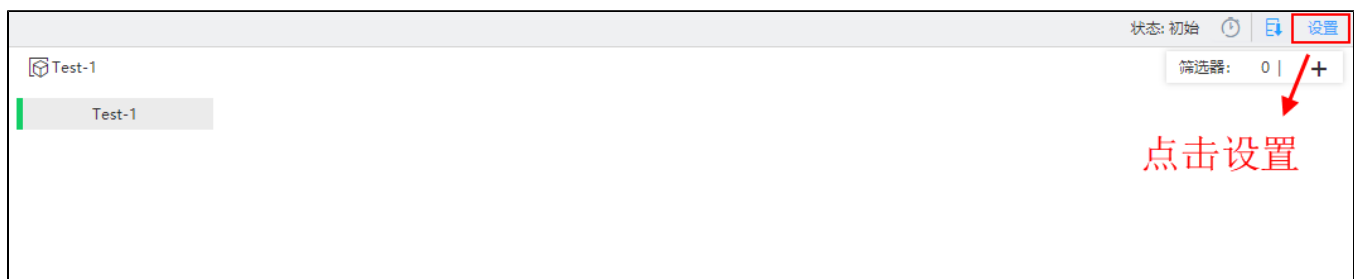
• 设置步骤

第一步：首次增量抽取

1) 自助数据集中浏览初始数据

   <input type="checkbox"/> 立即刷新			<input type="checkbox"/> 显示隐藏字段 <input checked="" type="checkbox"/> 显示别名 100 行
As Test-1 data	# Test-1 id	# Test-1 quantity	
20201109	1.00	50.00	
20201109	2.00	51.00	
20201109	3.00	52.00	

2) 点击抽取按钮再点击弹出的设置按钮



3) 进行相关增量抽取的设置

抽取设置

全量抽取

增量抽取按时间戳

增量字段

Test-1 / data

时间格式

YYYYMMDD

☐ 忽略抽取当天数据

☐ 覆盖最后抽取的

0

天数据

抽取表设置

抽取表名

Test_1

异常处理

抽取出错时

回滚

继续

定时抽取

立即抽取

确定

取消

⚠ 当选择“增量抽取按时间戳”时，数据集中需含有时间信息的字段作为增量字段。

4) 抽取成功后前往高速缓存库相应路径下预览数据

The screenshot shows a file explorer interface for the '高速缓存库' (High-Speed Cache Library). The tree view is expanded to show the following structure:

- 高速缓存库
 - DEFAULT
 - 数据采集空间
 - 管理员
 - 1
 - Test_1 (selected)
 - Test_a
 - orders
 - oreder2
 - z1

A context menu is open for the selected 'Test_1' item, with the following options:

- 打开(O)
- 预览数据(A) (highlighted)
- 设置数据权限
- 调度管理
- 删除(D)

🏠 导航 | 预览数据[Test_1] x

表名称: Test_1 表别名: Test_1

data	id	quantity
20201109	1.00	50.00
20201109	2.00	51.00
20201109	3.00	52.00

⚠️ 首次增量抽取时等同于全量抽取。

第二步：第二次增量抽取

1) 自助数据集中浏览新增的数据

🔍 📷 ↻ ☐ 立即刷新 ☐ 显示隐藏字段 ☒ 显示别名 100 行

Ab Test-1 data	# Test-1 id	# Test-1 quantity
20201109	1.00	50.00
20201109	2.00	51.00
20201109	3.00	52.00
20201110		
20201110	4.00	53.00

2) 进行增量抽取设置

⚙️ 抽取设置

☐ 全量抽取

☒ 增量抽取按时间戳

当前已抽取最大时间为: 20201109

增量字段 Test-1 / data 时间格式 YYYYMMDD

☐ 忽略抽取当天数据 ☐ 覆盖最后抽取的 0 天数据

抽取表设置

抽取表名 Test_1

异常处理

抽取出错时 ☒ 回滚 ☐ 继续

定时抽取

3) 查看抽取日志



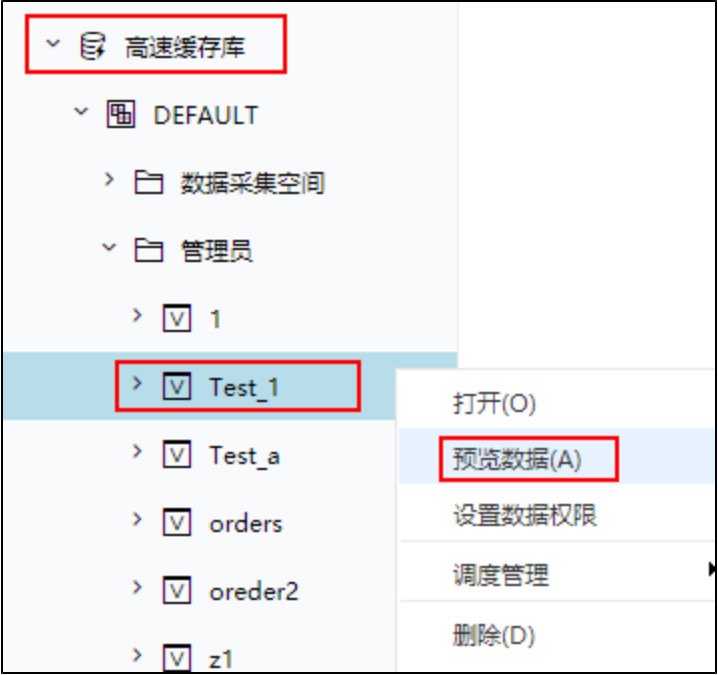
日志(数据集名称: Test_1)

开始时间	结束时间	日志	抽取时间	类型	状态	抽取条数	抽取失败信息
2020-11-09 16:21:24	2020-11-09 16:21:26	数据抽取成功	2秒	增量按时间戳	同步完成	2	---
2020-11-09 16:06:34	2020-11-09 16:06:35	数据抽取成功,由于表不存在,本次为全量抽取	1秒	增量按时间戳	同步完成	3	---

确定(O) 取消(C)

第三步：进行数据预览

1) 抽取成功后对应路径下预览数据



2) 浏览数据效果如图:

表名称: <input type="text" value="l8a8a90130175aa97aa97d7e80175ac096afe04f2"/>			表别名: <input type="text" value="Test-1"/>
data	id	quantity	
20201109	1.00	50.00	
20201109	2.00	51.00	
20201109	3.00	52.00	
20201110	NaN	NaN	
20201110	4.00	53.00	

增量抽取的数据

5.3 全量抽取和增量抽取区别

全量抽取	抽取所有数据
增量抽取	指与上次抽取结果中最大时间对比，将大于这个时间的数据进行集中抽取。