

数据抽取

功能概述

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：自助数据集、可视化查询、SQL查询、原生SQL查询、存储过程查询、Java查询、组合分析、透视分析、加载Excel数据。

数据抽取功能的机制如下：

- 1）数据集或分析确定结果字段。
- 2）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：



- 3）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。
- 注：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“可视化查询”、“组合分析”和“自助数据集”通过数据行权限控制数据抽取的结果。
 - 3、除“自助数据集”外，其它数据集如果包含参数，则只会抽取参数默认值相关的数据，如果参数没有默认值，将无法正常完成抽取。
 - 4、V8.0及以下版本只支持自助数据集的数据抽取。

入口及界面

- 1、非自助数据集及组合分析：在已保存的非自助数据集或组合分析的编辑界面，单击工具栏上的 **抽取** 按钮（），打开“数据抽取设置”窗口。

文档目录：

- [功能概述](#)
- [入口及界面](#)
- [设置说明](#)
 - [数据抽取示例](#)

数据抽取设置

实时

全量抽取

清空数据

异常处理

抽取出错时

回滚

继续

执行用户

资源创建者

管理员

特定用户

立即抽取

设置定时抽取

确定(O)

取消(C)

2、自助数据数据集：在已保存的自助数据集的编辑界面，先单击工具栏上的 **抽取** 按钮（  **设置**），再单击旁边的 **设置**（  **设置**），打开“数据抽取设置”窗口。

抽取设置

全量抽取

清空数据

增量抽数按时间戳

异常处理

抽取出错时

回滚

继续

执行用户

资源创建者

管理员

特定用户

立即抽取

设置定时抽取

取消

确定

设置说明

非自助数据集和组合分析的数据抽取功能不支持“增量抽取”。

“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	- 勾选清空数据：清空缓存数据并重新抽取。 - 勾除清空数据：保留每次抽取的数据记录，并再次抽取所有数据。 注：勾除清空数据时，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。
	增量抽取	增量抽数据按时间戳	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。 目前只有自助数据集支持增量抽取，且只有自助数据集中含有时间信息的字段才支持增量抽取。
		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”。
		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。
	异常处理		
执行用户	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。
	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 。
抽取时间	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 。
	立即抽取		表示立即抽取数据到高速缓存库。
	定时抽取		表示根据时间计划将数据定时抽取到高速缓存，其中定时抽取通过定制计划任务实现，详情请参见 计划 章节。

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“可视化查询”、“组合分析”和“自助数据集”的数据抽取受数据行权限控制。

数据抽取示例

当选择“全量抽取”并勾除“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。

如下示例中添加了“日期标识”字段，用日期来区分不同时间抽取的数据。

示例效果

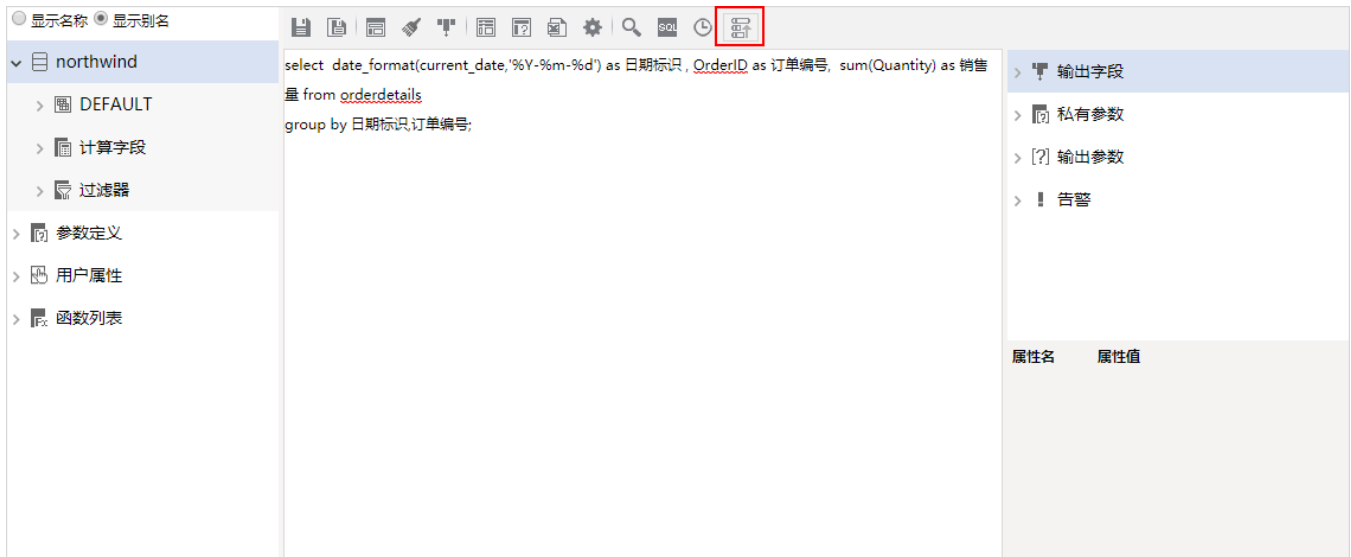
在高速缓存库中浏览该资源的数据，结果如图，包括了2018-12-19和2018-12-20抽取的数据：

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

设置方法

1、抽取2018-12-19的数据。

1) 点击SQL查询工具栏的 **数据抽取** 按钮，如图：



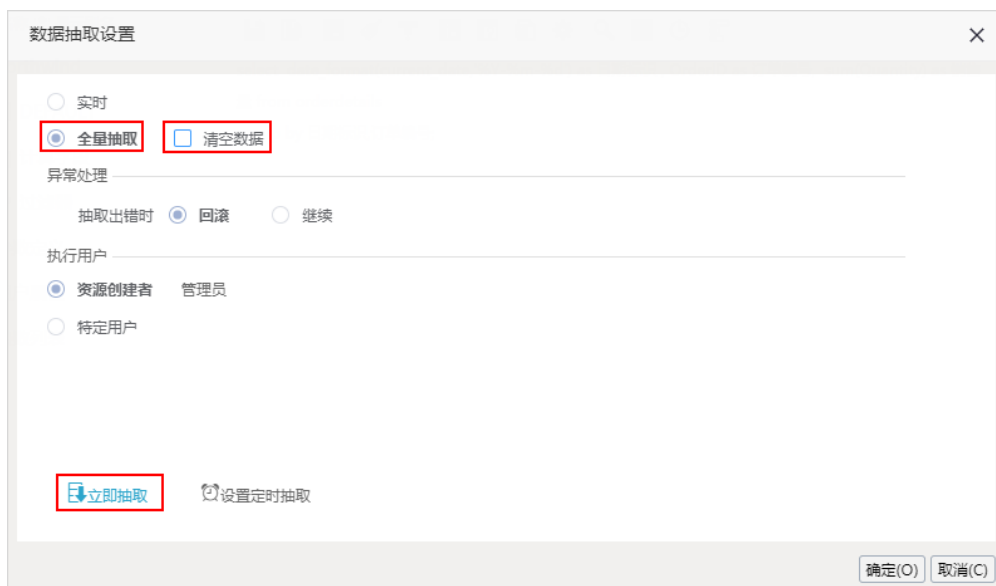
2) 弹出“数据抽取设置”界面，选择“全量抽取”后，点击 **立即抽取** ；



2、抽取2018-12-20的数据。

1) 点击SQL查询工具栏的 **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** ；



3、进行数据预览。

1) 在高速缓存库找到该资源，选中该资源，右键 > **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