自助ETL-界面介绍

ETL工作流示例

ETL工作流必须以数据源为起点，以目标源为终点：从数据源中抽取需要进行处理的数据，通过拖拽内置的预处理方法，之后将处理过的数据存储到目标源中。

数据源中支持的关系数据源有：

oracle、mysql、星环、DB2、gbase 8a、gbase 8S 、PostgreSQL、SQL Server 、SelectDB、vertica、greenplum、gauss100、gauss200、hive、达梦6、达梦7、sybase、aliyun MaxCompute、aliyun AnalyticDB、clickhouse、infobright、huawei FusionInsight HD、presto、MariaDB、KingBase、KingBase_V8、KingBaseANALYTICS、TERADATA、SHENTONG、OBASE、INFORMIX、IMPALA、KYLIN、SAP HANA、SelectDB 数据库。关于数据源的更多信息请参见 数据源 章节。

系统支持的数据预处理方法包含：采样、拆分、过滤与映射、列选择、空值处理、合并列、合并行、元数据编辑、JOIN、行选择、去除重复值、排序、增加序列号、聚合、分列、派生列等。这些预处理方法的使用详情请参见 数据预处理 章节。

目标源中支持的关系目标源有：

oracle、mysql、星环、DB2、gbase 8a、gbase 8S 、PostgreSQL、SQL Server 、vertica、greenplum、gauss100、gauss200、达梦6、达梦7、sybase、clickhouse、infobright、MariaDB 、SelectDB关于目标源的更多信息请参见 目标源 章节。

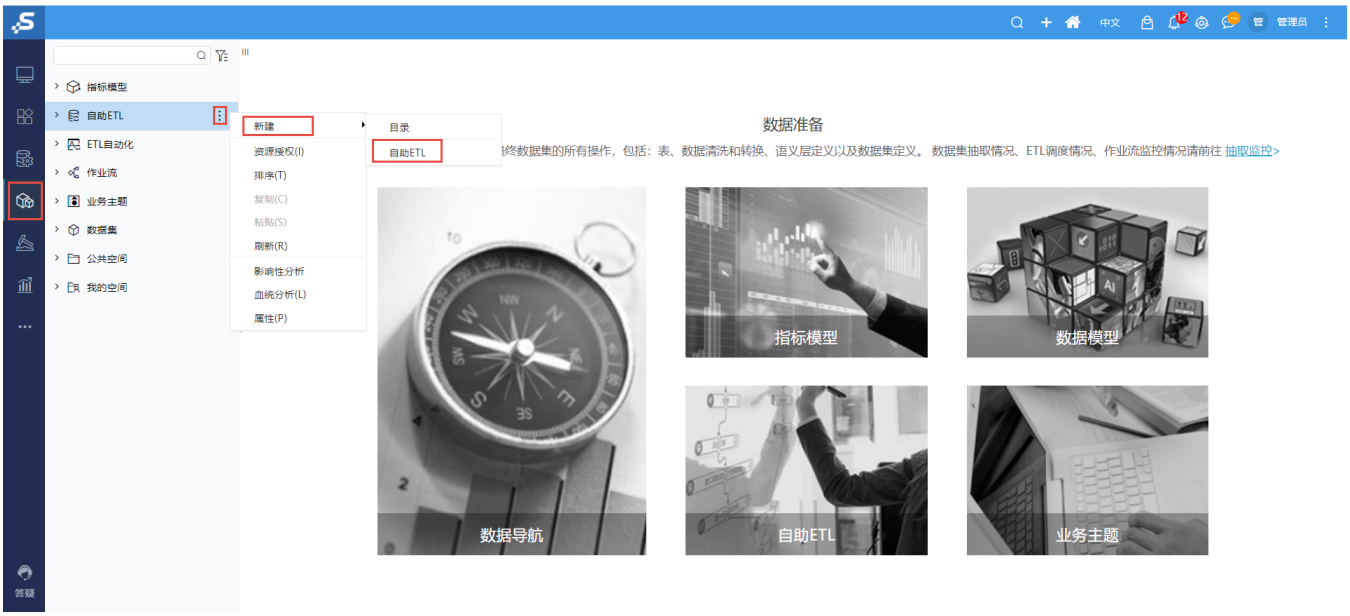
功能入口

ETL工作流定制界面的操作入口有如下三个：

入口1：在系统主界面的快捷菜单中选择 **数据准备 > 自助ETL**，进入“新建自助ETL”界面：

入口2：在系统导航栏中选择 **数据准备**，进入“数据准备”界面并单击快捷菜单 **自助ETL**，进入“新建自助ETL”界面：

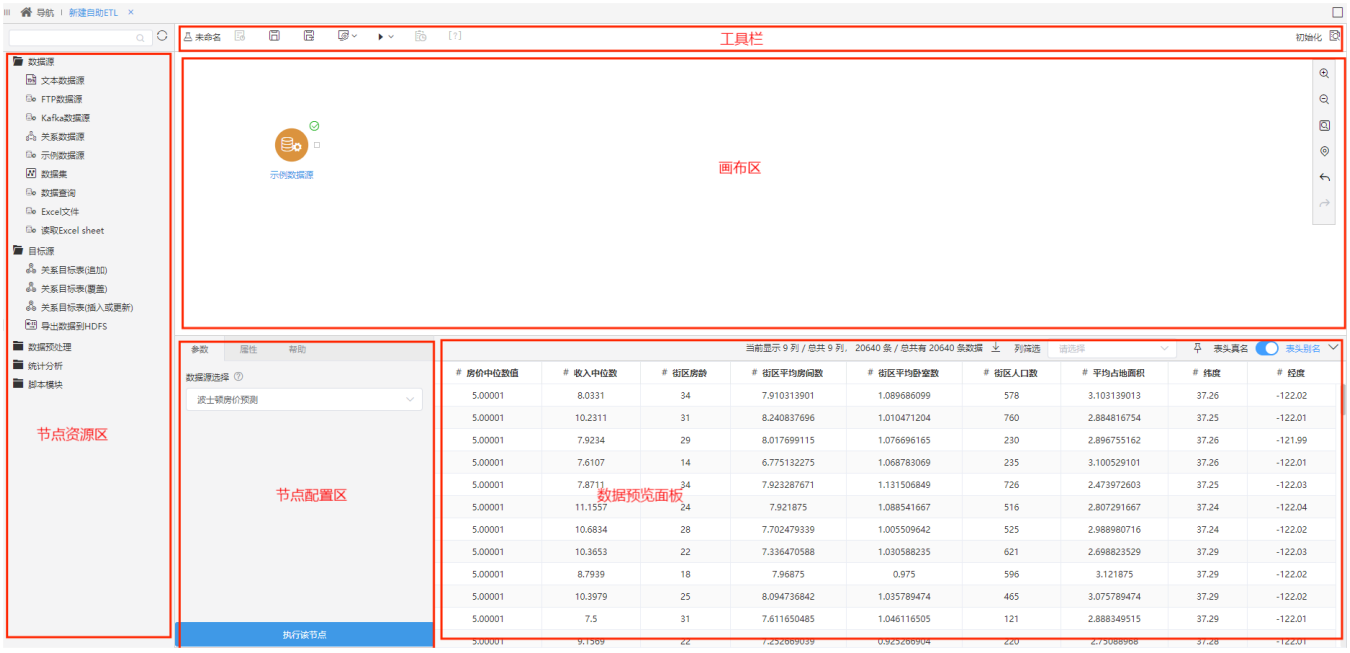
入口3：在系统导航栏中选择 **数据准备**，在左侧资源树自助ETL的更多中选择“自助ETL”，进入“新建自助ETL”界面：



界面介绍

ETL工作流定制界面

ETL流程定制界面用于设计定制ETL工作流。如下图所示：



该界面主要分为如下几个区：

- 节点资源区：显示当前流程可拖拽使用的资源，最顶端的文本框支持输入资源名称关键字模糊匹配搜索结果。详细介绍请参考自助ETL-节点资源区介绍。
- 画布区：用于定制ETL工作流。
- 节点配置区：用于对"画布区"所选资源的参数和属性进行配置。该区默认显示当前流程的别名、描述及创建更新时间信息。

- 工具栏：用于对当前流程进行的操作，详情请参见 工具栏。
- 数据预览面板：用于查看选择的节点输出的数据。

工作流操作

工具栏

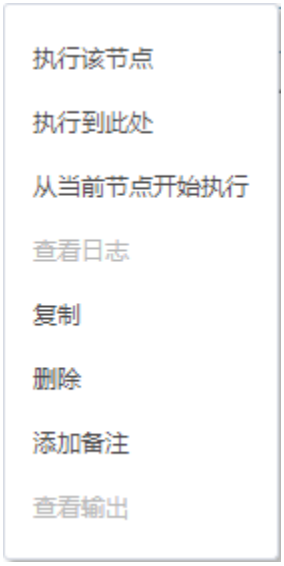
工具栏中有如下工具按钮支持工作流的相关操作。

这些工具按钮从左至右的说明如下：

按钮	说明
查看历史	用于查看定制的ETL工作流执行历史的相关信息。
保存	用于保存当前ETL流程。
另存	用于将当前ETL流程保存到其它路径下。
缓存	缓存策略： 1. 第一次执行，每个节点执行后，结果都会缓存起来，下游节点执行的时候，直接从缓存中获取上游节点执行结果 2. 第二次在界面上手工点执行，上次执行成功的节点，不会再重新执行，只会执行未执行或执行失败的节点。 3. 如果想执行成功的节点也能重新执行，比如关系数据源节点想重新执行获取最新数据，那就先点击清除缓存，然后再执行 4. 通过计划任务调起etl或者作业流调起etl，每个节点都会重新执行（为了保证能获取最新数据） 缓存作用： 1. 在界面上手工点执行，上次执行成功的节点，不会再重新执行，减少用户等待时间 2. 启用缓存，系统会更加稳定可靠。因为启用缓存，每个节点执行只需要从缓存中获取上游节点执行结果，不需要重新计算上游节点逻辑。不启用缓存，每个节点执行需要把上游节点逻辑重算一次，越到后面的节点，累积的计算逻辑越多，执行的时间越长，越容易出现卡死现象。 清除缓存：点击按钮清除缓存的节点数据。
运行	执行策略：用于运行当前ETL流程。 • 全量（默认）：运行数据源节点包含的全部数据； • 小批量：运行节点前1000条数据，当数据量较大时选择小批量运行，可减少用户等待时间。 ⚠ 1. 小批量功能目前只支持关系数据源与数据查询节点； 2. 需要配置缓存才能使用小批量功能，请参考 缓存 ； 3. 设置为小批量试运行不影响计划任务，计划任务还是按全量执行。
定时任务	用于修改当前ETL流程的调度设置。ETL调度详情请参见 抽取监控 章节相关内容。
参数设置	用于配置当前ETL高级查询的参数，详情请参考 数据挖掘-参数设置 。
日志	用于记录自助ETL运行状态信息。

右键菜单

各节点资源的右键菜单支持相关操作。
节点资源的右键菜单如下：



该右键菜单各项的说明如下：

右键菜单	说明
执行到此处	表示运行工作流到当前节点资源结束。
执行该节点	表示运行工作流时到当前节点资源结束。
从当前节点开始执行	表示运行工作流时从当前节点资源开始执行。
查看日志	用于查看当前节点资源的运行日志。
复制	用于复制选择的节点，与节点右键菜单的“粘粘”结合使用。
删除	表示删除当前节点资源。
添加备注	对实验或节点添加备注信息进行记录，详情请参考 文本组件 。
查看输出	用于查看当前节点资源的输出列表。
粘粘	在空白的画布区任意位置上粘贴复制的节点。
清空备注	清空当前实验上的所有备注

画布工具

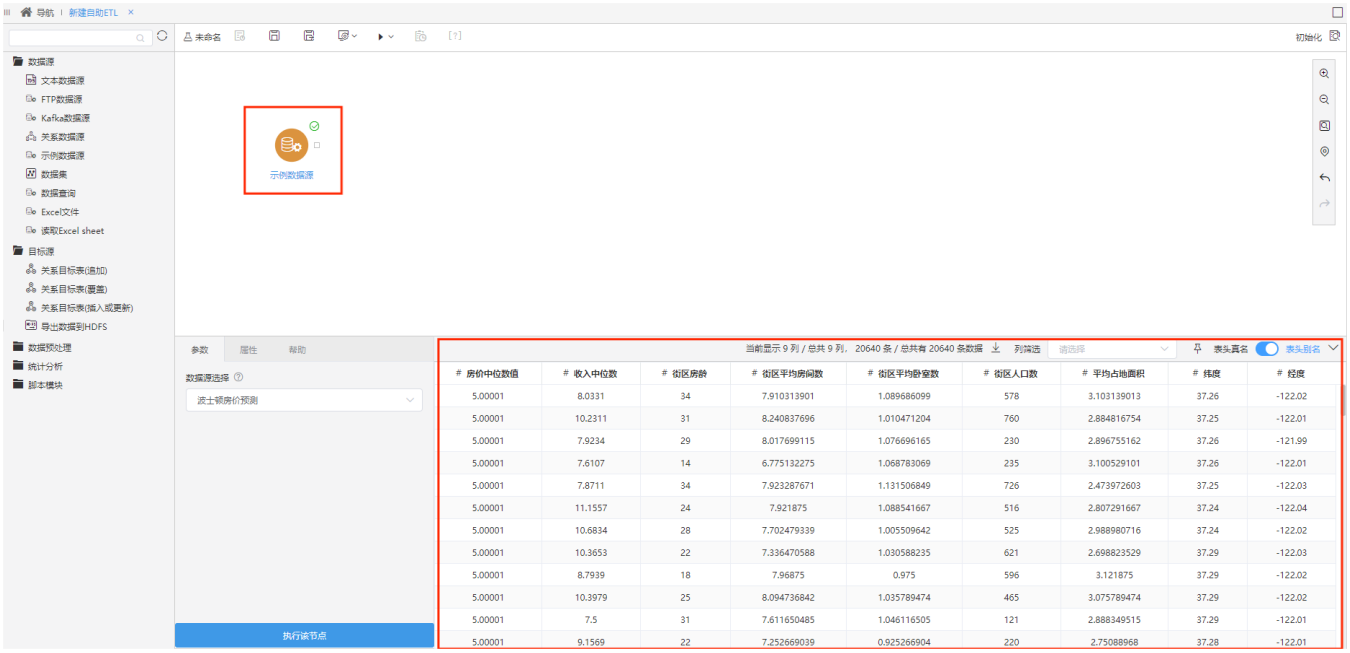
画布内含缩放工具，用于对工作进行放大、缩小操作：

该工具箱中从上到下依次是：放大、缩小、原始大小、定位到节点、撤销、还原。

数据预览面板

用于查看选择的节点输出的数据。

打开数据预览面板，点击节点可查看节点输出后的数据。



各项说明如下：

设置项	说明
节点状态	• 节点执行成功，则数据预览面板显示对应节点的数据预览； • 节点未执行或执行报错，则数据预览面板提示“暂无数据”。
列筛选	选择一个或多个列的方式来查看数据。
表头真名/表头别名	选择显示表头真名或别名。
下载	下载预览的数据到本地。 此处会把预览的数据以csv文件的方式下载到本地。为了保证数据安全，默认不会下载全量数据，数据量为100条。如果需要下载更多数据，可以到 系统运维——数据挖掘配置——执行引擎——节点数据存储行数 中配置，然后重新执行该节点即可。