DataWorks（数据工场）

快速入门

快速入门

本指南将指引您快速完成一个完整的数据开发和运维操作。

> **注意：**
>
> 如果您是第一次使用 DataWorks（数据工场，原大数据开发套件），请确认已经根据 **准备工作** 模块的操作，准备好账号和项目角色、项目空间等内容，然后进入 **DataWorks 管理控制台** 页面，单击对应项目空间后的 **进入工作区**，便可进入 DataWorks 的 **数据开发** 页面开始数据开发工作。

通常情况下，通过 DataWorks 的项目空间实现数据开发和运维，包含以下操作：

步骤1：建表并上传数据

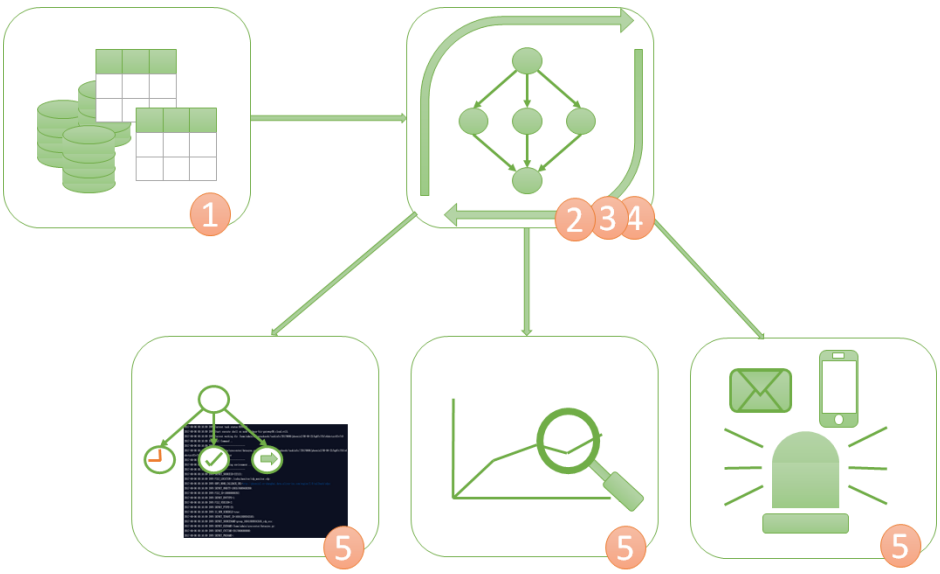
步骤2：创建工作流

步骤3：创建同步任务

步骤4：设置周期和依赖

步骤5：运维及日志排错

基本流程如下图所示：



本文将以创建表 bank_data 和 result_table 为例，说明如何创建表并上传数据。其中表 bank_data 用于存储业务数据，表 result_table 用于存储数据分析后产生的结果。

操作步骤

创建表 bank_data

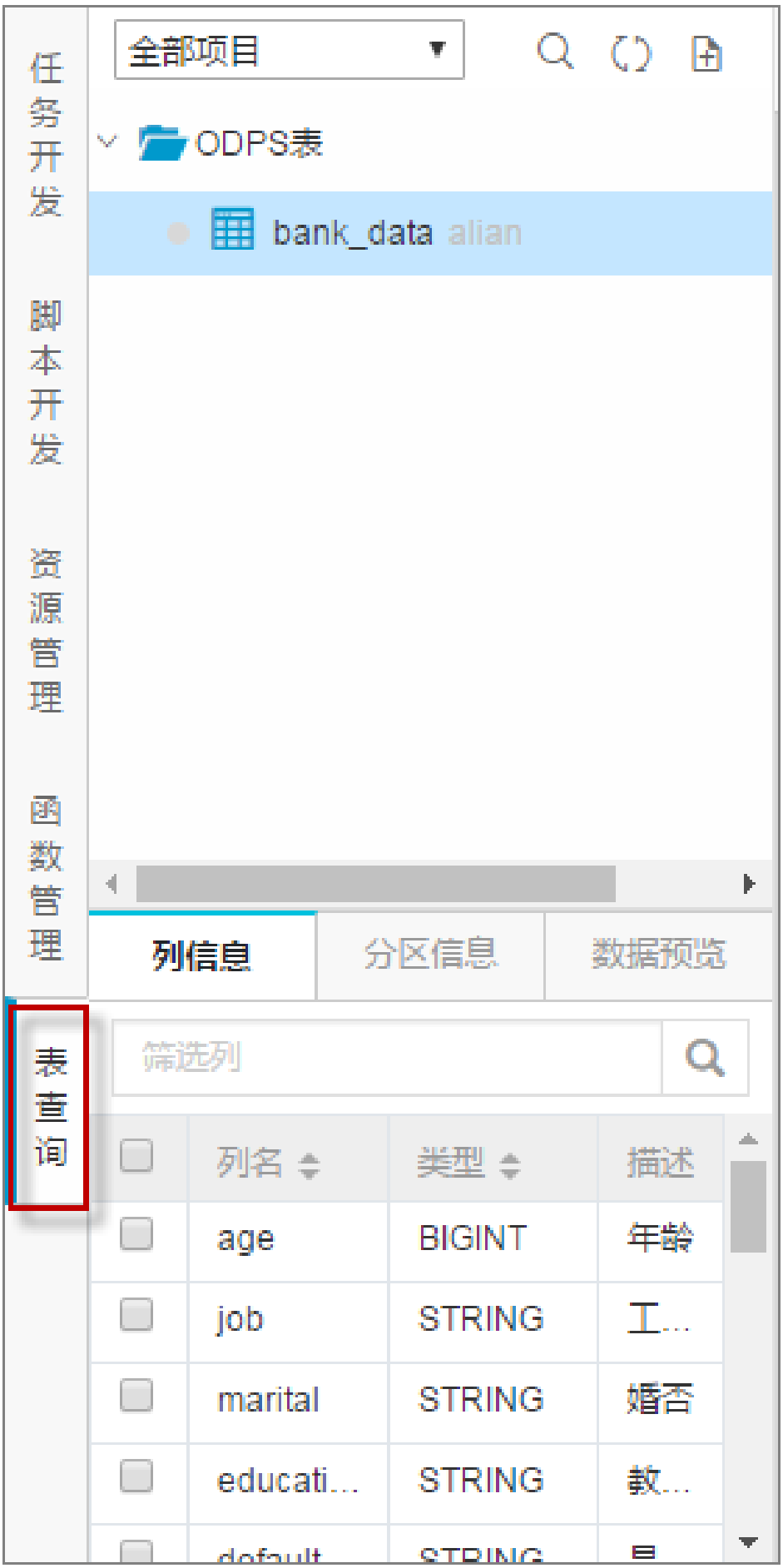
进入项目空间后，在 **数据开发** 页面单击 **新建**，选择 **新建表**。如下图所示：

在新建表页面，输入建表语句，单击 **确认**。创建表的更多 SQL 语法请参见 MaxCompute 创建/查看/删除表。

本示例的建表语句如下所示：

```
CREATE TABLE IF NOT EXISTS bank_data
(
age BIGINT COMMENT '年龄',
job STRING COMMENT '工作类型',
marital STRING COMMENT '婚否',
education STRING COMMENT '教育程度',
default STRING COMMENT '是否有信用卡',
housing STRING COMMENT '房贷',
loan STRING COMMENT '贷款',
contact STRING COMMENT '联系途径',
month STRING COMMENT '月份',
day_of_week STRING COMMENT '星期几',
duration STRING COMMENT '持续时间',
campaign BIGINT COMMENT '本次活动联系的次数',
pdays DOUBLE COMMENT '与上一次联系的时间间隔',
previous DOUBLE COMMENT '之前与客户联系的次数',
poutcome STRING COMMENT '之前市场活动的结果',
emp_var_rate DOUBLE COMMENT '就业变化速率',
cons_price_idx DOUBLE COMMENT '消费者物价指数',
cons_conf_idx DOUBLE COMMENT '消费者信心指数',
euribor3m DOUBLE COMMENT '欧元存款利率',
nr_employed DOUBLE COMMENT '职工人数',
y BIGINT COMMENT '是否有定期存款'
);
```

创建表后，可以在左侧导航栏 **表查询** 中输入表名进行搜索，查看表信息。如下图所示：


任务开发
脚本开发
资源管理
函数管理
表查询
全部项目
ODPS表
bank_data alian
列信息
分区信息
数据预览
筛选列
列名
类型
描述
age
BIGINT
年龄
job
STRING
工...
marital
STRING
婚否
educati...
STRING
教...
default
STRING

创建表 result_table

进入 **数据开发** 页面，单击 **新建**，选择 **新建表**。

在新建表页面，输入建表语句，单击 **确认**。建表语句如下所示：

```
CREATE TABLE IF NOT EXISTS result_table
(
education STRING COMMENT '教育程度',
num BIGINT COMMENT '人数'
);
```

创建表后，可以在左侧导航栏 **表查询** 中输入表名进行搜索，查看表信息。

本地数据上传至 bank_data

DataWorks（数据工场，原大数据开发套件）支持以下操作：

将保存在本地的文本文件中的数据上传到工作空间的表中。

通过数据集成模块将业务数据从多个不同的数据源导入到工作空间。

注意：

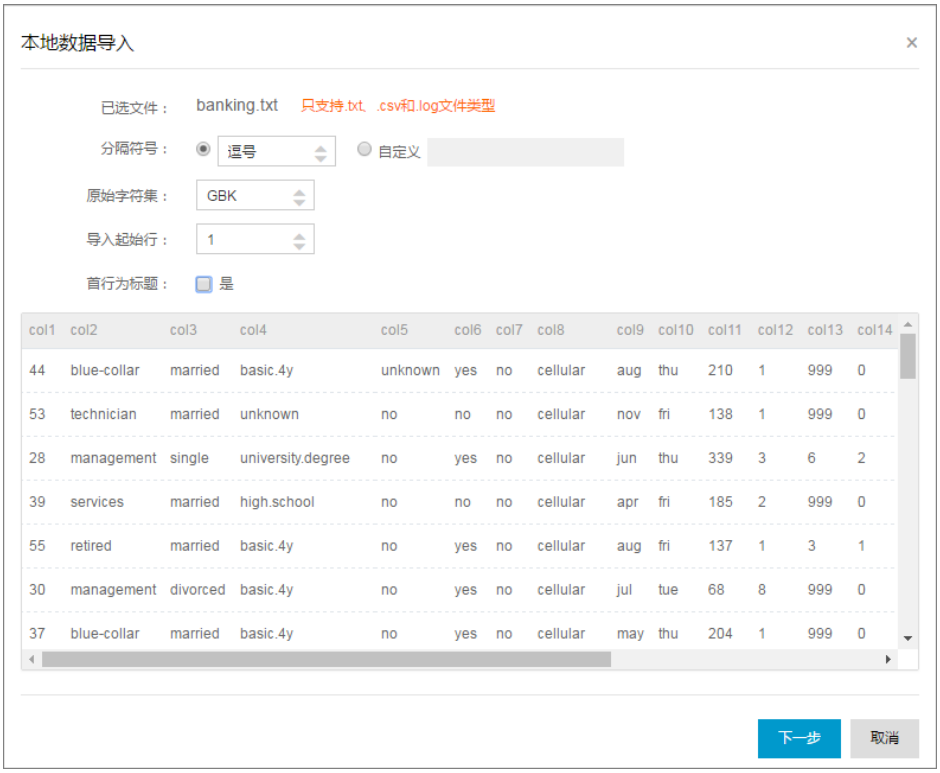
本文将使用本地文件作为数据来源。本地文本文件上传有以下限制：

- 文件类型：仅支持 .txt 和 .csv 格式。
- 文件大小：不超过 10 M。
- 操作对象：支持分区表导入和非分区表导入，但不支持分区值为中文。

以导入本地文件 banking.txt 到 DataWorks 为例，操作如下：

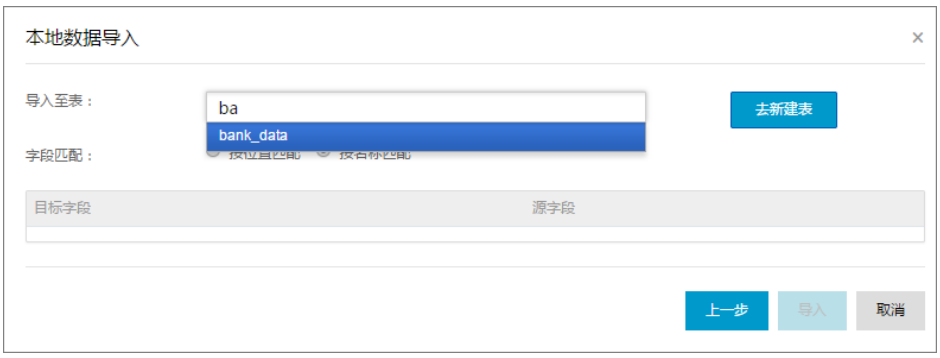
1 . 单击 **导入**，选择 **导入本地数据**。如下图所示：

2 . 选择本地数据文件，配置导入信息，单击 **下一步**。如下图所示：

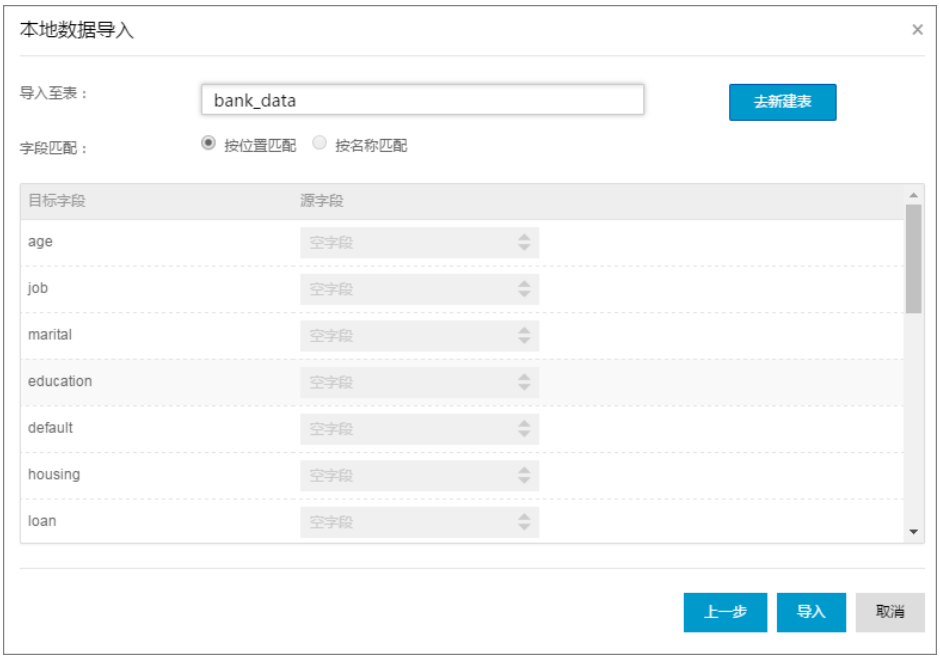


col1	col2	col3	col4	col5	col6	col7	col8	col9	col10	col11	col12	col13	col14
44	blue-collar	married	basic.4y	unknown	yes	no	cellular	aug	thu	210	1	999	0
53	technician	married	unknown	no	no	no	cellular	nov	fri	138	1	999	0
28	management	single	university.degree	no	yes	no	cellular	jun	thu	339	3	6	2
39	services	married	high.school	no	no	no	cellular	apr	fri	185	2	999	0
55	retired	married	basic.4y	no	yes	no	cellular	aug	fri	137	1	3	1
30	management	divorced	basic.4y	no	yes	no	cellular	jul	tue	68	8	999	0
37	blue-collar	married	basic.4y	no	yes	no	cellular	may	thu	204	1	999	0

3 . 至少输入2个字母搜索表名，选择需导入数据的表，如：bank_data。若需新建，可单击 **去新建表**，如下图所示：



4 . 选择字段匹配方式（本示例选择按位置匹配），单击 **导入**。如下图所示：



5 . 文件导入后，系统将提示您数据导入成功或失败。

其他数据导入方式

创建数据同步任务

适用范围：

保存在 RDS、MySQL、SQL Server、PostgreSQL、MaxCompute、OCS、DRDS、OSS、Oracle、FTP、dm、Hdfs、MongoDB 等多种数据源中的各种数据。

通过 DataWorks 创建数据同步任务的具体操作请参见 创建数据同步任务。

本地文件上传

适用范围：

文件大小不超过 10M，支持 .txt 和 .csv 文件类型，目标支持分区表和非分区表，但不支持中文作为分区。

通过 DataWorks 进行本地文件上传，具体操作如上文 **本地数据上传至 bank_data** 所示。

使用 Tunnel 命令上传文件

适用范围：

大小超过 10M 的本地文件和其他资源文件等。

通过 MaxCompute 客户端 提供的 Tunnel 命令来进行数据的上传及下载，当本地数据文件需要上传到分区表时，可以通过客户端 tunnel 命令方式进行上传。

详情请参见 Tunnel 命令操作。

使用 dataX 开源工具

适用范围：

大批量的本地数据导入，二维表结构的数据等，上述3种方式无法支持的其他场景。详情请参见 DataX。

更多 DataX 开源介绍，请参见 DataX 开源地址。

后续步骤

现在，您已经学习了如何创建表并上传数据，您可以继续学习下一个教程。在该教程中您将学习如何创建工作流来对项目空间的数据进行进一步的计算与分析。详情请参见 创建工作流分析数据。

DataWorks（数据工场，原大数据开发套件）的数据开发功能支持图形化设计数据分析工作流，以工作流任务和内部节点的方式实现对数据的处理和相互依赖。目前支持包括 ODPS_SQL、数据同步、OPEN_MR、SHELL、机器学习、虚节点等多种任务类型，每种任务类型的具体使用方法请参见 任务类型介绍。

本文将以创建工作流 work 为例，说明如何在工作流中创建节点并配置依赖关系，以方便地设计和展现数据分析的步骤和顺序，并简要说明如何利用数据开发功能对工作空间的数据做进一步的分析和计算。

前提条件

在开始本操作前请确保您已根据 创建表并上传数据 的操作，在工作空间中准备好业务数据表 bank_data 和其中的数据，以及结果表 result_table。

操作步骤

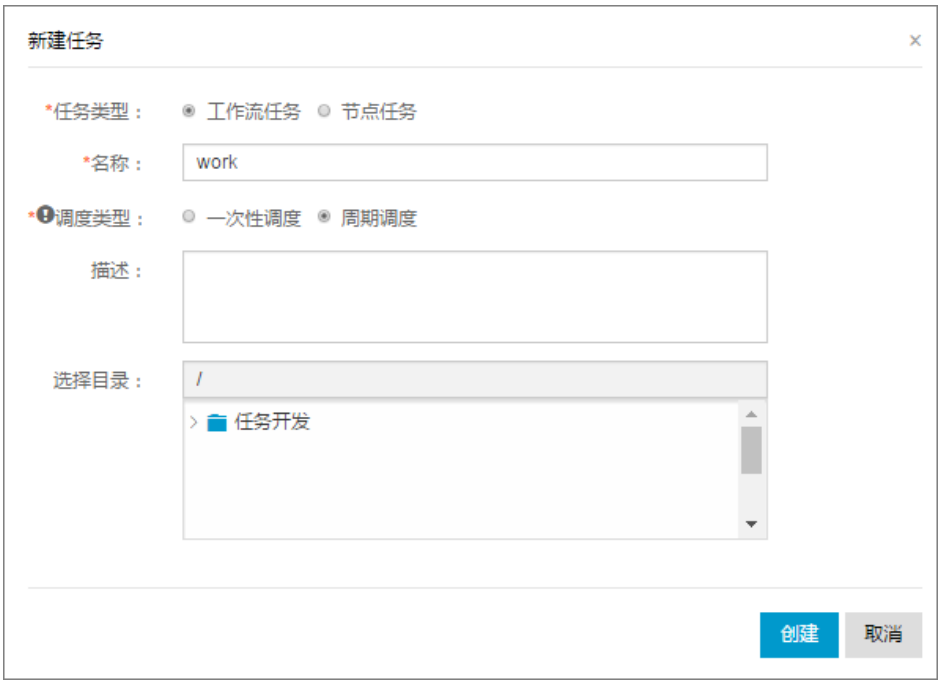
创建工作流

进入项目空间后，单击 **数据开发** 页面中的 **新建**，选择 **新建任务**。如下图所示：

选择弹出框中的相关内容，指定任务类型为 **工作流任务**。如下图所示：

注意：

下图中的调度属性一旦选定，不可以更改。



在工作流画布中创建节点和关系

本节将在工作流中创建一个虚节点 start 和一个 odps_sql 节点 insert_data，并配置为 insert_data 依赖于 start。

注意：

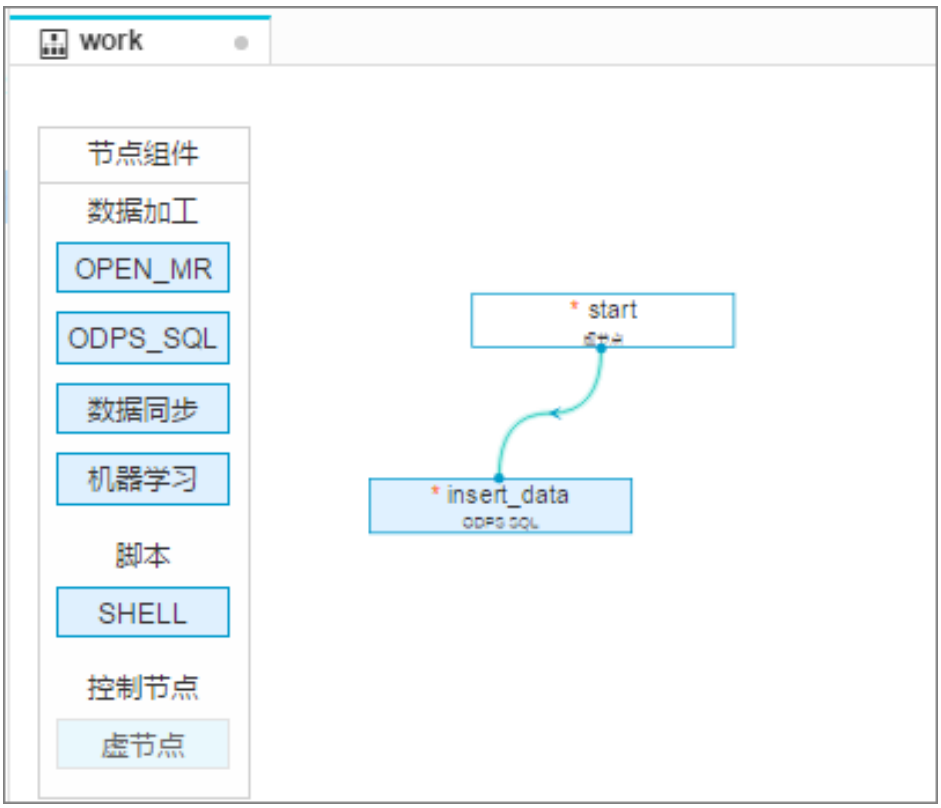
虚拟节点属于控制类型节点，在工作流运行过程中不对数据产生任何影响，仅用于实现对下游节点的运维控制。

虚节点在被其他节点依赖的情况下，如果被运维人员手动设置为运行失败，则下游未运行的节点将因此无法被触发运行，在运维过程中可以防止上游错误数据进一步蔓延。详情请参见 任务类型介绍 中的虚节点类型。综上所述，一般建议设计工作流时，默认创建一个虚节点作为根节点来控制整个工作流。

双击虚节点，输入节点名 start。

双击 **ODPS_SQL**，输入节点名 insert_data，如下图所示：

单击 start 节点并拖动连线到 insert_data 节点，使 insert_data 节点依赖于 start 节点，如下图所示：



在 ODPS_SQL 节点中编辑代码

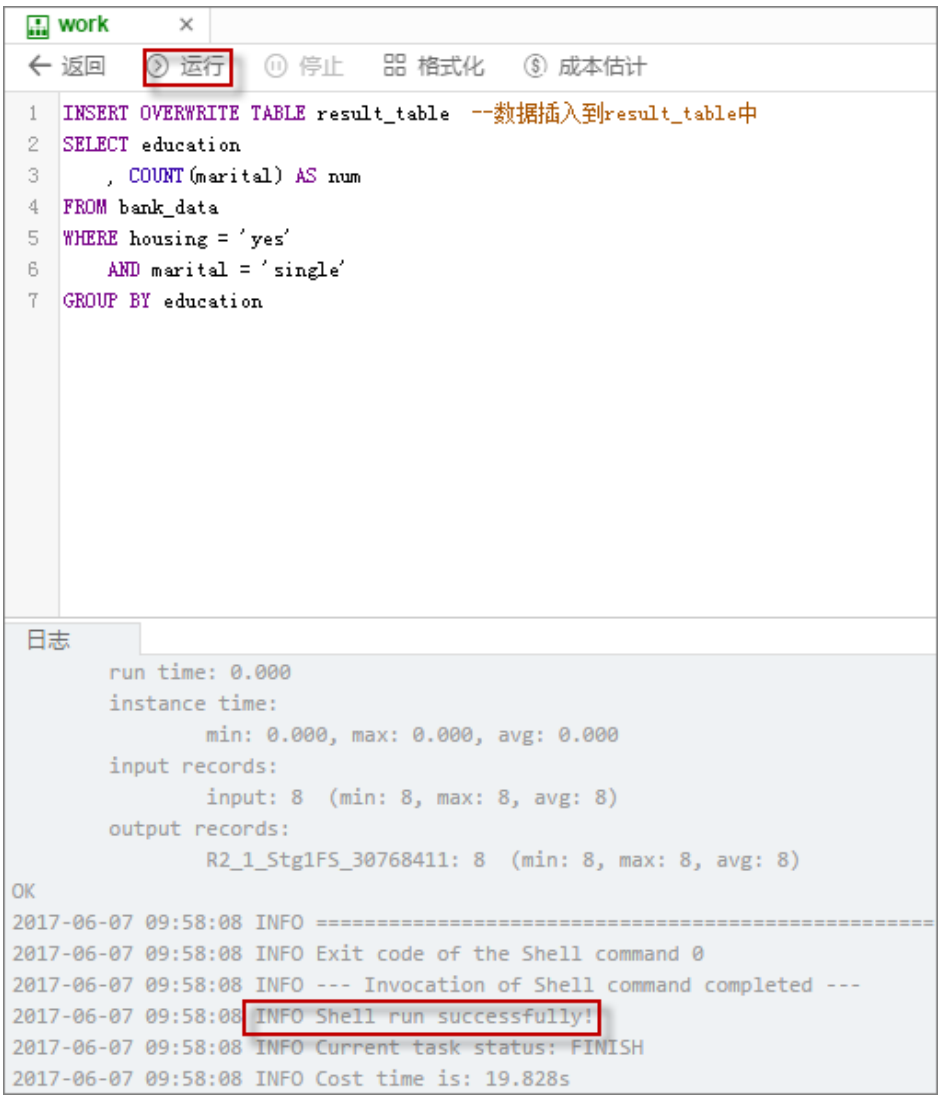
本节将在 ODPS_SQL 节点 **insert_data** 中用 SQL 代码查询不同学历的单身人士贷款买房的数量，并将结果保存下来以备后续节点继续分析或展现。SQL 语句如下所示，具体语法说明请参见 MaxCompute 文档。

```
INSERT OVERWRITE TABLE result_table  --数据插入到result_table中
SELECT education
, COUNT(marital) AS num
FROM bank_data
WHERE housing = 'yes'
AND marital = 'single'
GROUP BY education
```

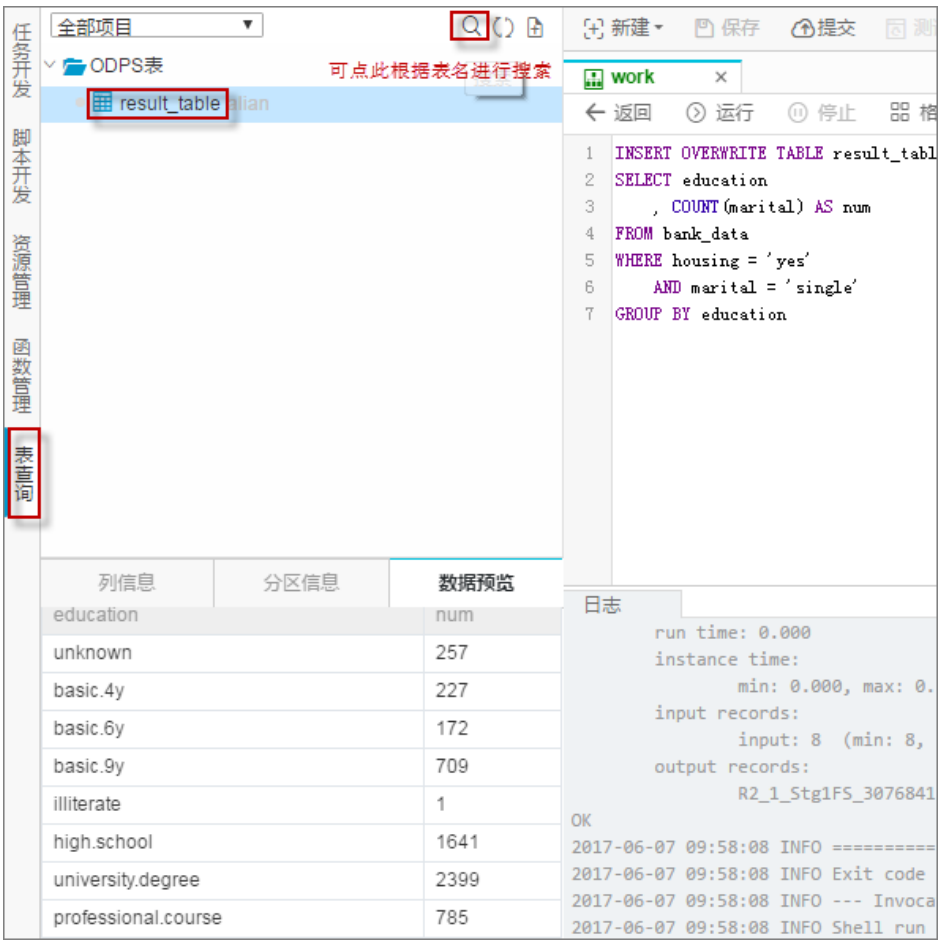
运行并调试 ODPS_SQL 节点

在 insert_data 节点中编辑好 SQL 语句后，单击 **保存**，防止代码丢失。

单击 **运行**，查看运行日志和结果。如下图所示：

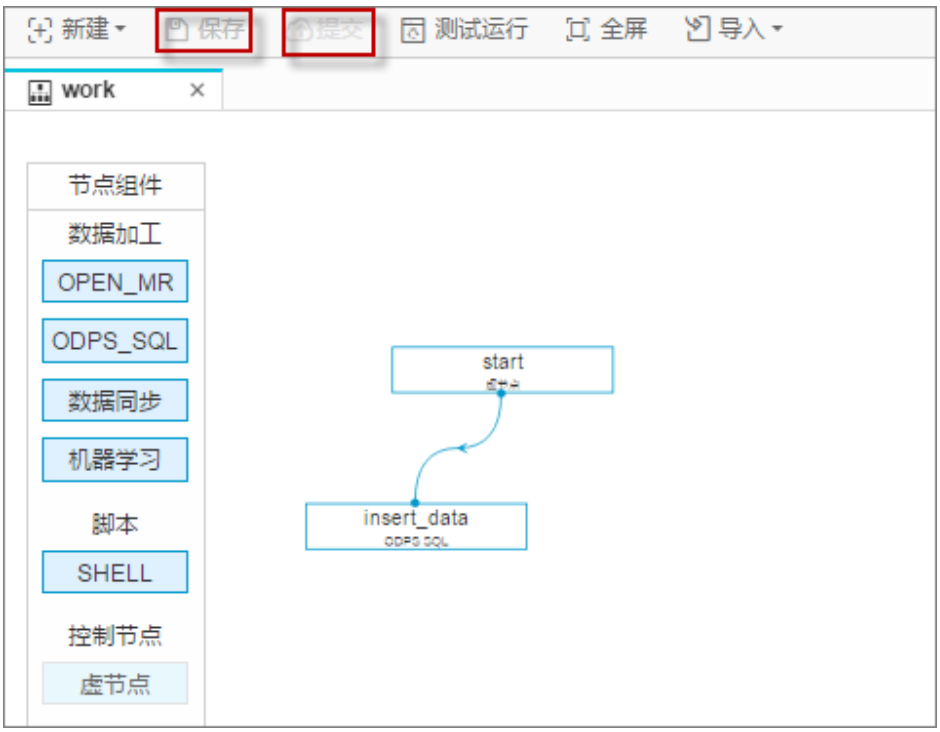


完成以上操作后，您可以在左侧的 **表查询** 中查询这张表中的数据。



保存并提交工作流

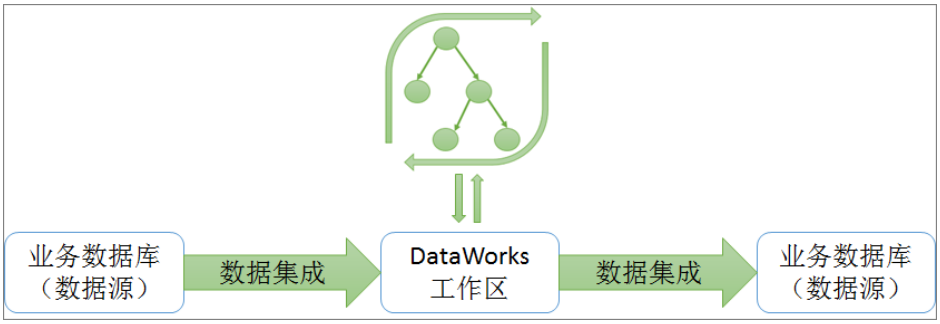
运行并调试好 ODPS_SQL 节点 insert_data 后，返回工作流页面，保存并提交整个工作流。如下图所示：



后续步骤

现在，您已经学习了如何创建工作流，并对其进行保存和提交，您可以继续学习下一个教程。在该教程中您将学习如何通过创建同步任务来把数据导出到不同类型的数据源中。详情请参见 **创建同步任务导出结果**。

在 DataWorks（数据工场，原大数据开发套件）中，通常使用数据集成功能，将您的系统中产生的业务数据定期导入到工作区，通过工作流任务的计算后，再将计算结果定期导出到您指定的数据源中，供进一步展示或运行使用。



目前数据集成功能支持从以下数据源中将数据导入工作空间或将数据从工作空间导出：RDS、MySQL、SQL Server、PostgreSQL、MaxCompute、OCS、DRDS、OSS、Oracle、FTP、 dm、Hdfs、MongoDB 等，详细的数据源类型列表请参见 **支持数据源类型**。

本文将以 MySQL 数据源为例，说明如何利用数据集成功能将 DataWorks 中的数据导出到 MySQL 数据源中。

前提条件

如果您是 **ECS 上自建的数据库** 或者是 **RDS/MongoDB 等数据源**，需要在自己的 **ECS 添加安全组** 或 **RDS/MongoDB 等管控台添加白名单**，详情请参见 添加白名单和安全组。

注意：

若使用自定义资源组调度 RDS 的数据同步任务，必须把自定义资源组的机器 IP 也加到 RDS 的白名单中。

操作步骤

新增数据源

注意：

只有项目管理员角色才能够新建数据源，其他角色的成员仅能查看数据源。

以项目管理员身份进入 DataWorks **管理控制台**，单击 **项目列表** 下对应项目操作栏中的 **进入工作区**。

进入顶部菜单栏中的 **数据集成** 页面，单击左侧导航栏中的 **数据源**。

单击右上角的 **新增数据源**，如下图所示：

填写新增数据源弹出框中的各配置项，如下图所示：

配置说明如下：

数据源类型：有公网 IP。

数据源名称：字母、数字、下划线组合，且不能以数字和下划线开头。比如：abc_123。

数据源描述：不超过 80 个字符。

JDBC URL：JDBC 连接信息，格式为：jdbc:mysql://host:port/database。

用户名/密码：数据库对应的用户名和密码。

不同数据源类型对应的配置说明，请参见 **数据源配置**。

单击 **测试连通性**。

若测试连通性成功，单击 **保存** 即可。

若测试连通性失败，请根据自身情况参见 ECS 上自建的数据库测试连通性失败 或 RDS 数据源测试连通性不通。

确认作为目标的 MySQL 数据库中有表

在 MySQL 数据库中创建表 odps_result ，建表语句如下所示：

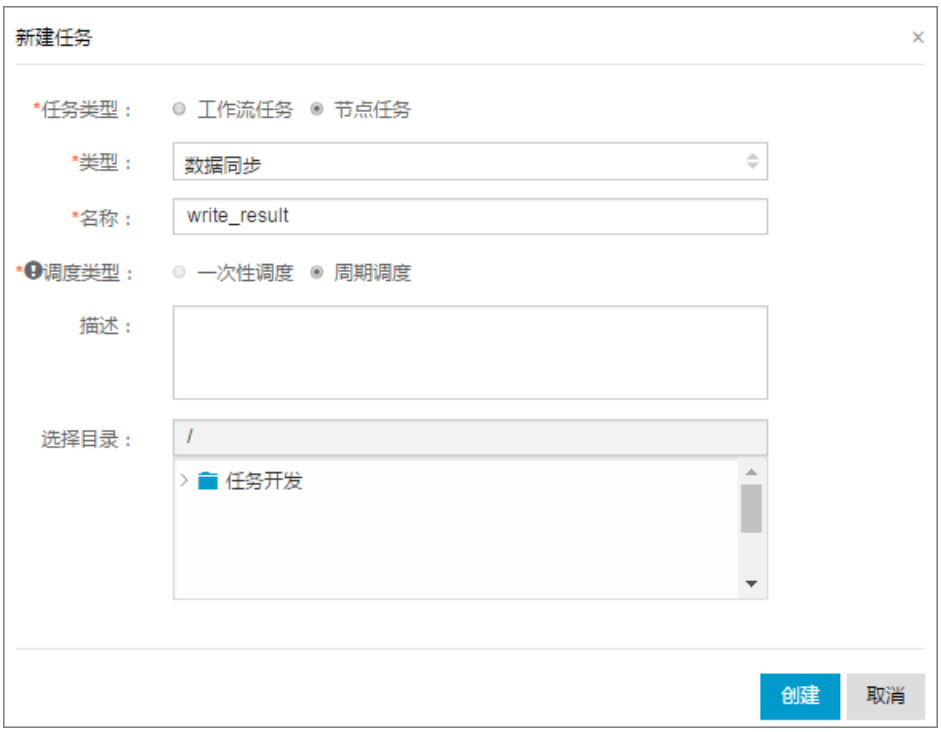
```
CREATE TABLE `ODPS_RESULT` (
`education` varchar(255) NULL ,
`num` int(10) NULL
)
```

建表完成后，可通过 desc odps_result； 语句查看表详情。

新建并配置同步节点

本节将新建一个同步节点 write_result 并进行配置，以把表 result_table 中的数据写入到自己的 MySQL 数据库中。具体操作如下：

新建同步节点 write_result，如下图所示：



选择来源。

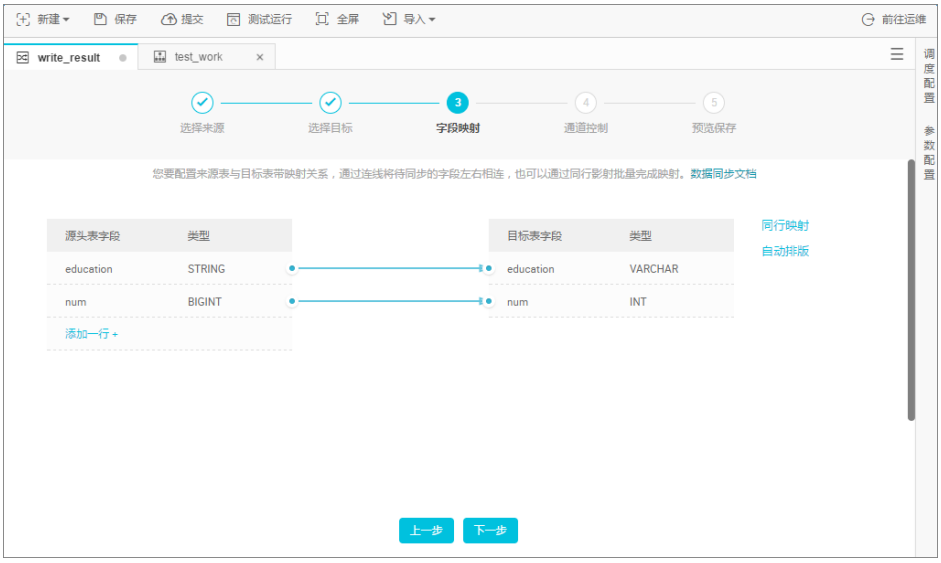
选择 MaxCompute 数据源及源头表 result_table，然后单击 **下一步**，如下图所示：

选择目标。

选择 Mysql 数据源及目标表 odps_result，然后单击 **下一步**，如下图所示：

映射字段。

选择字段的映射关系。需对字段映射关系进行配置，左侧 **源头表字段** 和右侧 **目标表字段** 为一一对应的关系。

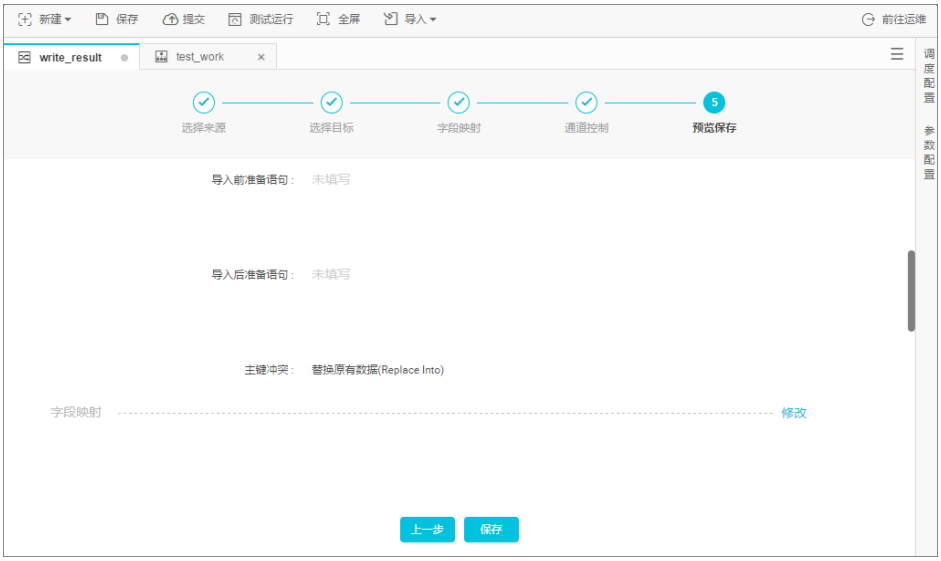


通道控制。

单击 **下一步**，配置作业速率上限和脏数据检查规则，如下图所示：

预览保存。

完成以上配置后，上下滚动鼠标可查看任务配置，如若无误，单击 **保存**，如下图所示：



提交数据同步任务

同步任务保存后，单击右边的 **提交**，将同步任务提交到调度系统中，调度系统会按照配置属性在从第二天开始自动定时执行。

后续步骤

现在，您已经学习了如何创建同步任务把数据导出到不同类型的数据源中，您可以继续学习下一个教程。在该教程中您将学习如何设置同步任务的调度属性和依赖关系。详情请参见 设置任务的调度属性和依赖关系。

DataWorks（数据工场，原大数据开发套件）提供了强大的调度能力，支持按照时间、依赖关系的任务触发机制，支持每日千万级别的任务按照 DAG 关系准确、准时运行。支持分钟、小时、天、周和月多种调度周期配置，详情请参见 调度配置介绍。

本文将以 创建同步任务 中创建的 write_result 为例，将其调度周期配置为周调度，说明 DataWorks 的调度配置和任务运维功能。

操作步骤

配置同步任务的调度属性

进入 **数据开发 > 任务开发** 页面，双击打开需要配置的同步任务（write_result），单击右侧的 **调度配置**，即可为任务配置 **调度属性**，如下图所示：

配置参数说明：

调度状态：勾选后即为暂停状态。

出错重试：勾选后即开启。

生效日期：任务的有效日期，根据自身需求进行设置。

调度周期：任务的运行周期（月/周/天/小时/分钟），比如以周为调度周期进行调度。

具体时间：任务运行的具体时间，比如将任务配置为在每周二的凌晨2点开始运行。

配置同步任务的依赖属性

配置完同步任务的调度属性后，展开依赖属性继续配置，如下图所示：

依赖属性中可以配置任务的上游依赖，表示即使当前任务的实例已经到定时时间，也必须等待上游任务的实例运行完毕才会触发运行。

如上图所示的配置表明：当前任务的实例将在上游 write_result 任务的实例运行完毕后才会触发执行，您在上游任务中输入 work，即可给 write_result 配置上游任务。

如果没有配置上游任务，则当前任务默认由项目本身触发运行，故在调度系统中，该任务的上游默认为 project_start 任务。每一个项目中默认会创建一个 project_start 任务作为根任务。

提交同步任务

保存同步任务 write_result，单击 **提交**，将其提交到调度系统中，如下图所示：

任务只有提交到调度系统中，才会从第二天开始自动按照调度属性配置的周期在各时间点生成实例，然后定时运行。

> **注意：**
>
> 如果是 23:30 以后提交的任务，则调度系统从第三天开始才会自动周期生成实例并定时运行。

后续步骤

现在，您已经学习了如何设置同步任务的调度属性和依赖关系，您可以继续学习下一个教程。在该教程中您将学习如何对提交的任务进行周期运维并查看日志排错。详情请参见 周期运维并查看日志排错。

在之前的操作中，您配置了每周二凌晨 2 点执行同步任务，将任务提交后需要到第二天才能看到调度系统自动执行的结果，那么如何确认实例运行的定时时间和相互依赖关系符合预期呢？DataWorks（数据开发，原大数据开发套件）提供了测试运行、补数据和周期运行 3 种触发方式，详情如下：

测试运行：手动触发方式。如果您仅需确认单个任务的定时情况和运行，建议使用测试运行。

补数据运行：手动触发方式。如果您需要确认多个任务的定时情况和相互依赖关系，或者需要从某个根任务开始重新执行数据分析计算，可以考虑使用此方式。

周期运行：系统自动触发方式。提交成功的任务，调度系统在第二天 0 点起会自动生成当天不同时间点的运行实例，并在定时时间达到时检查各实例的上游实例是否运行成功，如果定时时间已到并且上游实例全部运行成功，则当前实例会自动触发运行，无需人工干预。

注意：

手动触发和自动调度的调度系统根据周期生成实例的规则一致：

无论周期选择天/小时/分钟/月/周，任务在每一个日期都会有对应实例生成。

仅在指定日期的对应实例会定时运行并生成运行日志。

非指定日期的对应实例不会实际运行，而是在满足运行条件时将状态直接转换为成功，因此不会有运行日志生成。

本文将为您说明如何实现以上三种触发方式，具体操作见下文。关于任务运维的更多操作和功能说明，请参见任务运维。

测试运行

手动触发测试运行

单击工作流页面中的 **测试运行** 按钮，如下图所示：

根据跳转页面的提示，单击 **确认** 和 **运行**，如下图所示：

单击 **前往运维中心** 查看任务运行状态，如下图所示：

查看测试实例的信息及运行日志

单击任务名，即可看到实例 DAG 图。在实例 DAG 视图中，右键单击实例，可以查看该实例的依赖关系和详细信息并进行终止运行、重跑等具体操作。在实例 DAG 视图中，双击实例，即可弹出任务属性、运行日志、操作日志、代码等，如下图：

注意：

测试运行是手动触发任务，只要定时时间到了，立即运行，无视实例的上游依赖关系。

根据前文所述的实例生成规则，配置为每周二凌晨 2 点运行的任务 write_result，测试运行时选择的业务日期是周一（业务日期=运行日期-1），实例会在 2 点真正运行；如果不是周一，则实例在 2 点转换为成功状态，且没有日志生成。

补数据运行

手动触发补数据运行

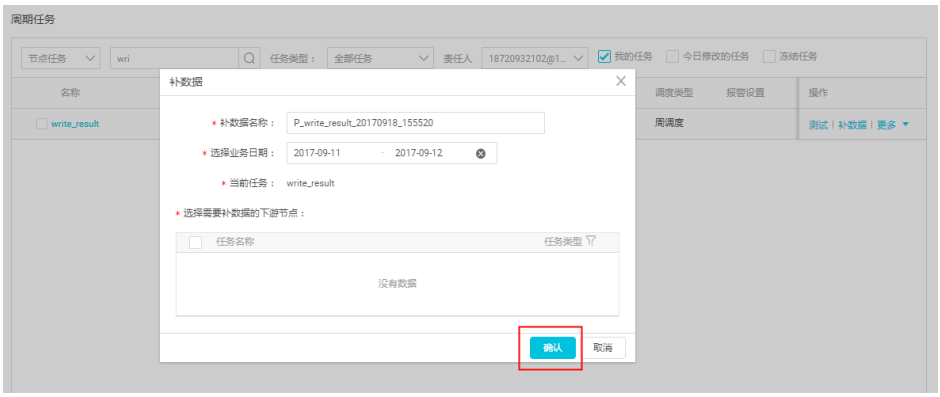
若需要确认 **多个任务** 的定时情况和相互依赖关系，或者需要从某个根任务开始重新执行数据分析计算，可进入 **运维中心 > 任务列表 > 周期任务** 页面，选择任务，单击 **补数据**，来补跑某段时间的多个任务。

操作步骤

进入 **运维中心 > 周期任务页面**，输入任务名称。

选中任务查询结果，单击 **补数据** 按钮。如下图所示：

设置补数据的业务日期为 2017-09-11 到 2017-09-12 ，选择 write_result 节点任务，单击 **确认**。如下图所示：



单击 **前往查看补数据结果。**

查看补数据实例的信息及运行日志

在 **补数据实例** 页面下找到任务实例：单击任务名，即可看到实例 DAG 图。在实例 DAG 视图中，右键单击实例，可以查看该实例的依赖关系和详细信息并进行终止运行、重跑等具体操作。在实例 DAG 视图中，双击实例，即可弹出任务属性、运行日志、操作日志、代码等，如下图所示：

注意：

上图中的任务失败的原因是：同步的表中，源头没有这个分区值，所以读取失败。

补数据任务的实例是一天依赖一天的，比如：补了2017-09-15 到 2017-09-18 这段时间的任务，如果 15 号的实例运行失败了，则 16 号的实例也不会运行。

根据前文所述的实例生成规则，配置为每周二凌晨 2 点运行的任务 write_result，补数据运行时选择的业务日期是周一（业务日期=运行日期-1），实例会在 2 点真正运行；如果不是周一，则实例在 2 点转换为成功状态，且没有日志生成。

周期自动运行

周期自动运行，由系统根据所有任务的调度配置自动触发，故页面没有操作入口。查看实例信息和运行日志有

以下两种：

进入 **运维中心 > 周期实例** 页面，选择业务日期或运行日期等参数，搜索 write_result 任务对应的实例，然后右键查看实例信息和运行日志。如下图所示：

单击任务名，即可看到实例 DAG 图。在实例 DAG 视图中，右键单击实例，可以查看该实例的依赖关系和详细信息并进行终止运行、重跑等具体操作。在实例DAG视图中，双击实例，即可弹出任务属性、运行日志、操作日志、代码等，如下图所示：

注意：

该任务未运行的原因是：上游任务未运行。

若任务的实例初始状态为未运行，当定时时间到达时，调度系统会检查这个实例的全部上游实例是否运行成功。

只有上游实例全部运行成功并且定时时间到达的实例，才会被触发运行。

处于未运行状态的实例，请确认上游实例已经全部成功且已到定时时间。