# 大数据开发套件

## 快速开始

# 快速开始

如果您是第一次使用大数据开发套件，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 **大数据开发套件 — 管理控制台** 页面，点击对应项目空间后的 **进入工作区**，便可进入大数据开发套件的 **数据开发** 页面开始数据开发工作。

一般使用大数据开发套件的项目空间实现数据开发和运维遵循以下步骤：

步骤1：建表并传数据；

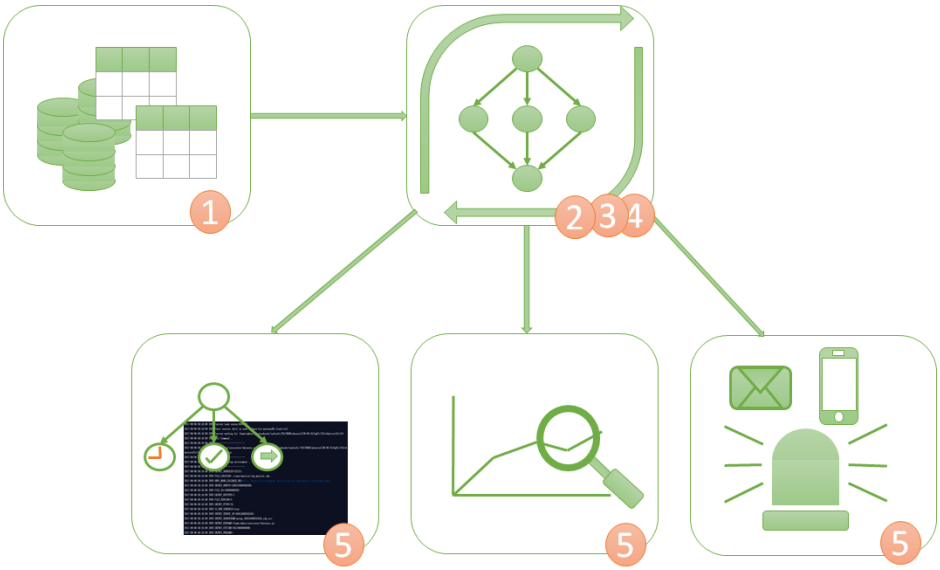
步骤2：创建工作流；

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。具体操作如下：

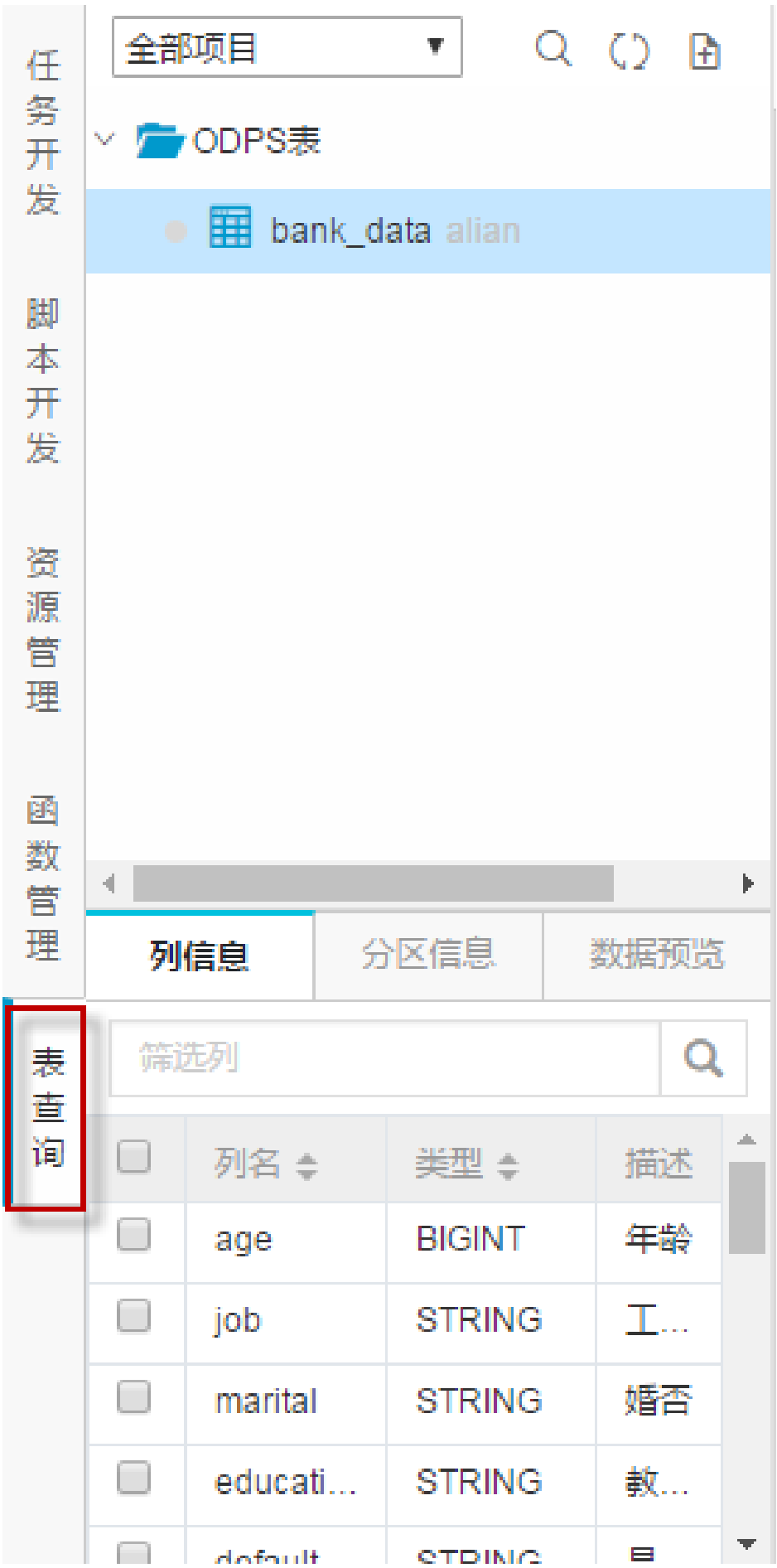
## 1. 创建表bank_data

a. 进入项目空间后，在**数据开发**页面点击**新建**，选择**新建表**；


b. 打开新建表后，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：

```
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'
);
```

c. 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...

# 2. 创建表result_table

a. 在**数据开发**页面点击**新建**，选择**新建表**；

b. 打开新建表后，输入建表语句，单击确认。建表语句如下：

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

c. 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。

# 3. 本地数据上传至bank_data

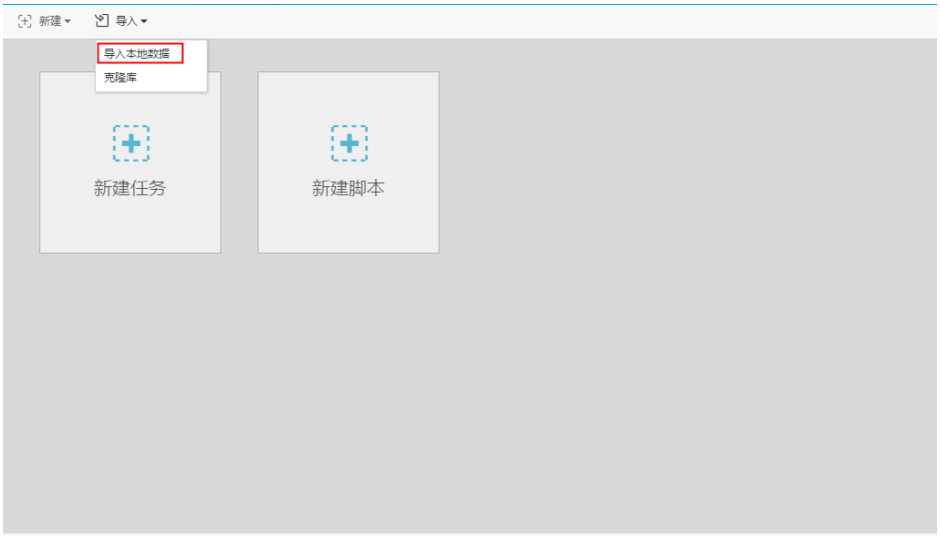
大数据开发套件支持用户将保存在本地的文本文件中的数据上传到工作空间的表，也支持通过数据集成模块将业务数据从多个不同的数据源导入到工作空间。本节将使用本地文件作为数据来源。

本地文本文件上传限制：

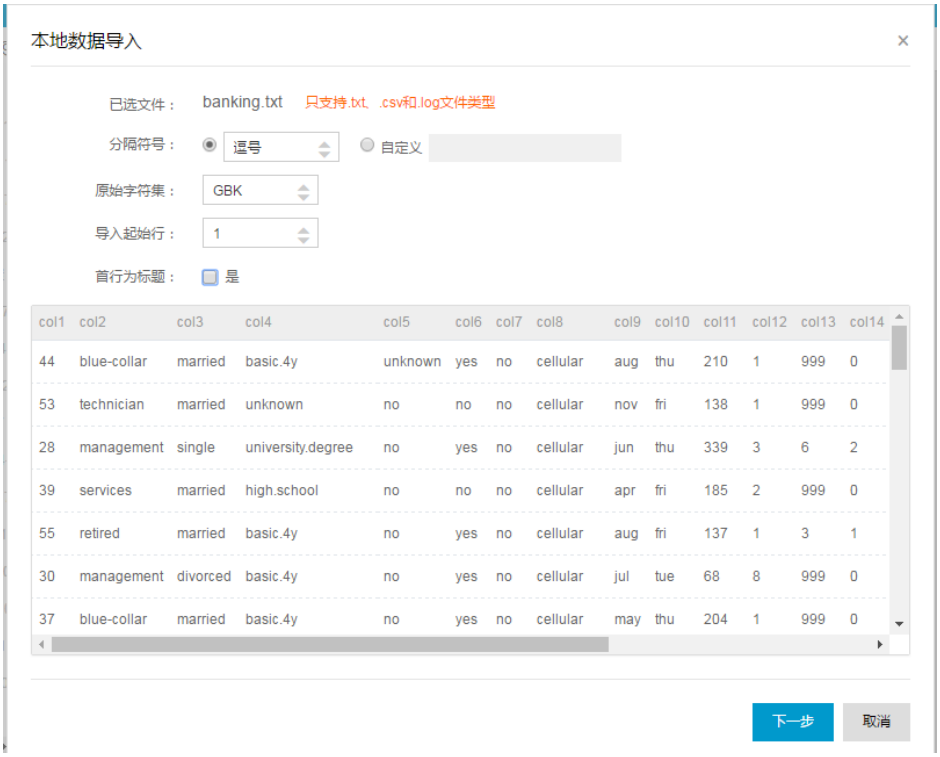
- 文件类型：仅支持 .txt 和 .csv 格式；
- 文件大小: 不超过 10 M；
- 操作对象：仅支持导入数据到非分区表，不支持MaxCompute分区表。

以导入 banking.txt 到大数据开发套件为例，操作如下：

a.点击**导入**，选择**导入本地数据**；

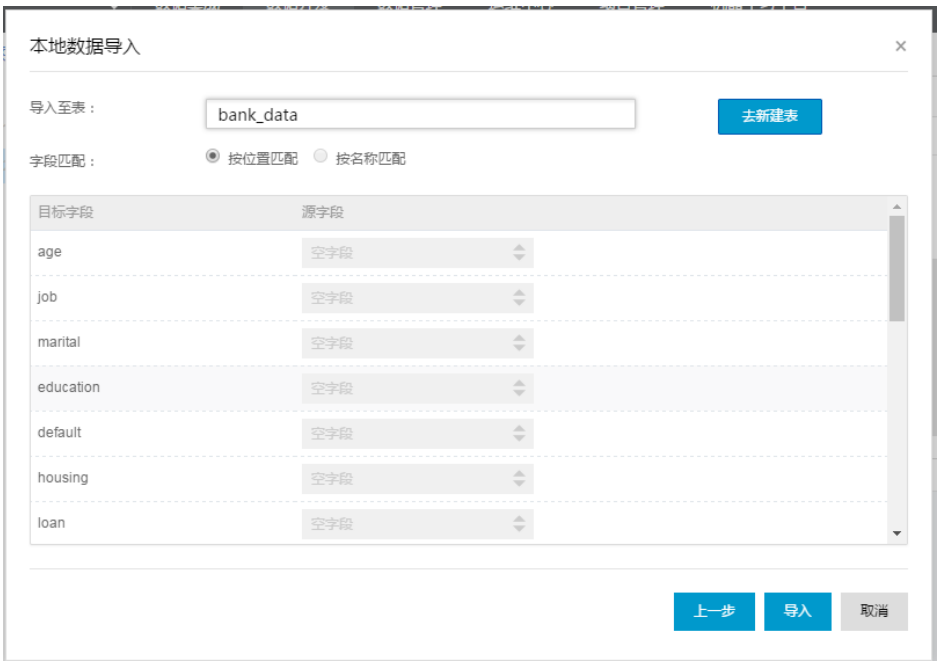



b.选择本地数据文件，配置导入信息，点击**下一步**；




c. 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，可点击**去新建表**，如下图所示：


d. 选择字段匹配方式（本示例选择按位置匹配），点击**导入**；




e. 文件导入后，系统将提示您数据导入成功或失败。

# 4. 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 DataIDE 创建数据同步任务的操作步骤请参见：**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，目标仅支持非分区表。

通过 DataIDE 进行本地文件上传，操作详情如“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**。

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，以方便的设计和展现数据分析的步骤和先后顺序，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。创建工作流分析数据的具体操作如下：

# 1. 创建工作流work

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。

a. 进入项目空间后，在**数据开发**页面点击**新建**，选择**新建任务**；


b. 在弹出框选择创建**工作流任务**；

**注意**：下图中的调度属性一旦选定，不可以再更改。



## 2. 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。

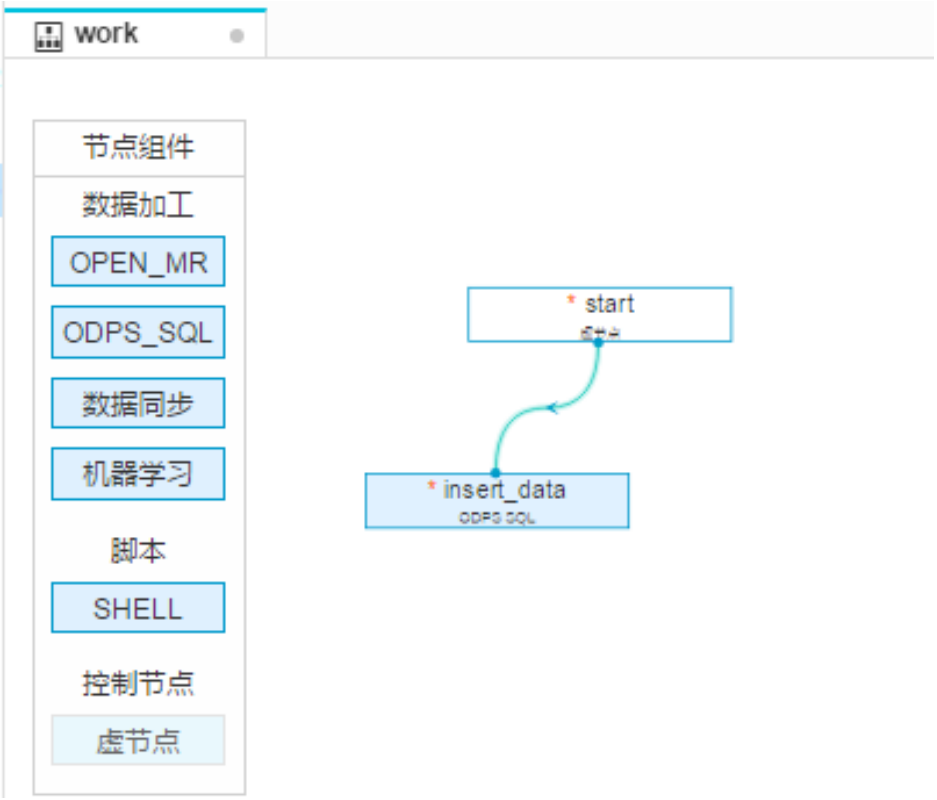
**注意：**

- 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。具体介绍请参见：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。具体操作如下：

a. 双击虚节点，输入节点名 start；


b. 双击“ODPS_SQL”，输入节点名insert_data，如下图所示：


c. 点击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：



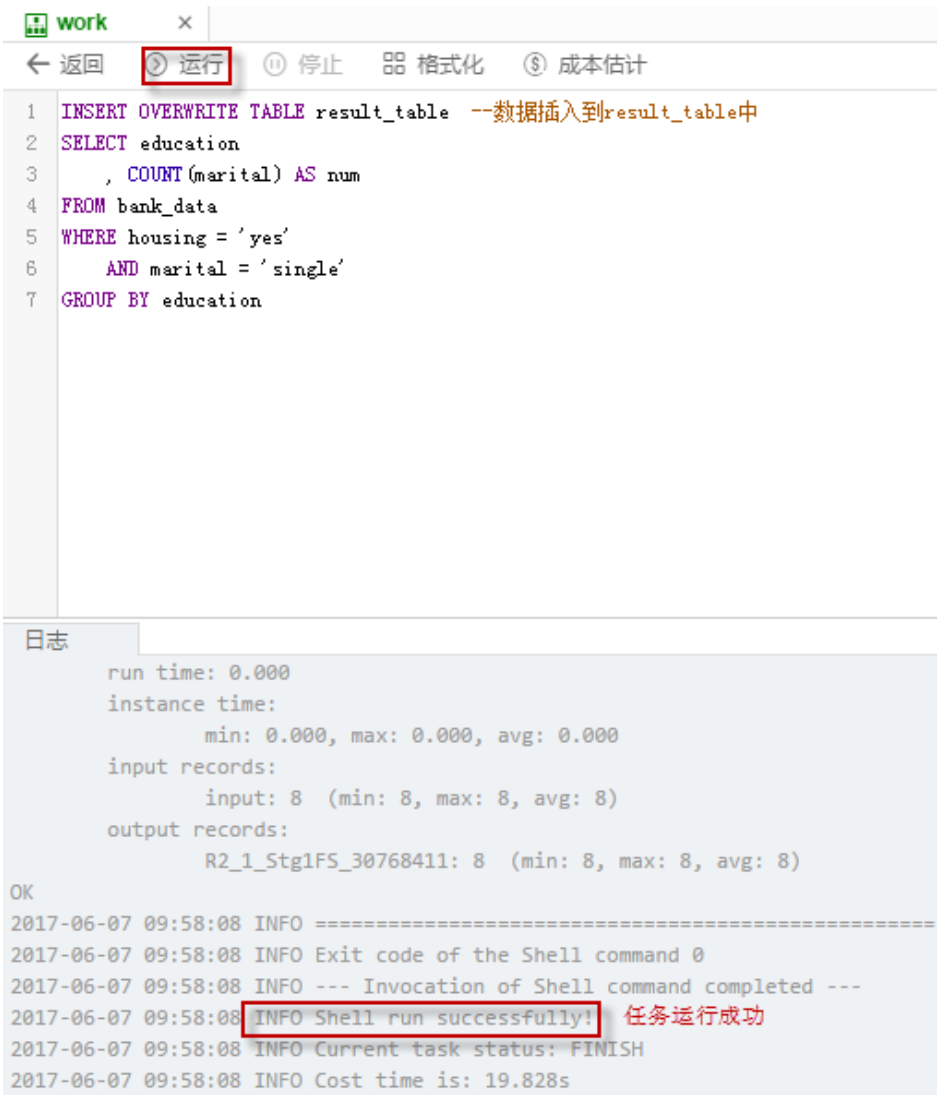

# 3. 在ODPS_SQL节点中编辑代码

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。

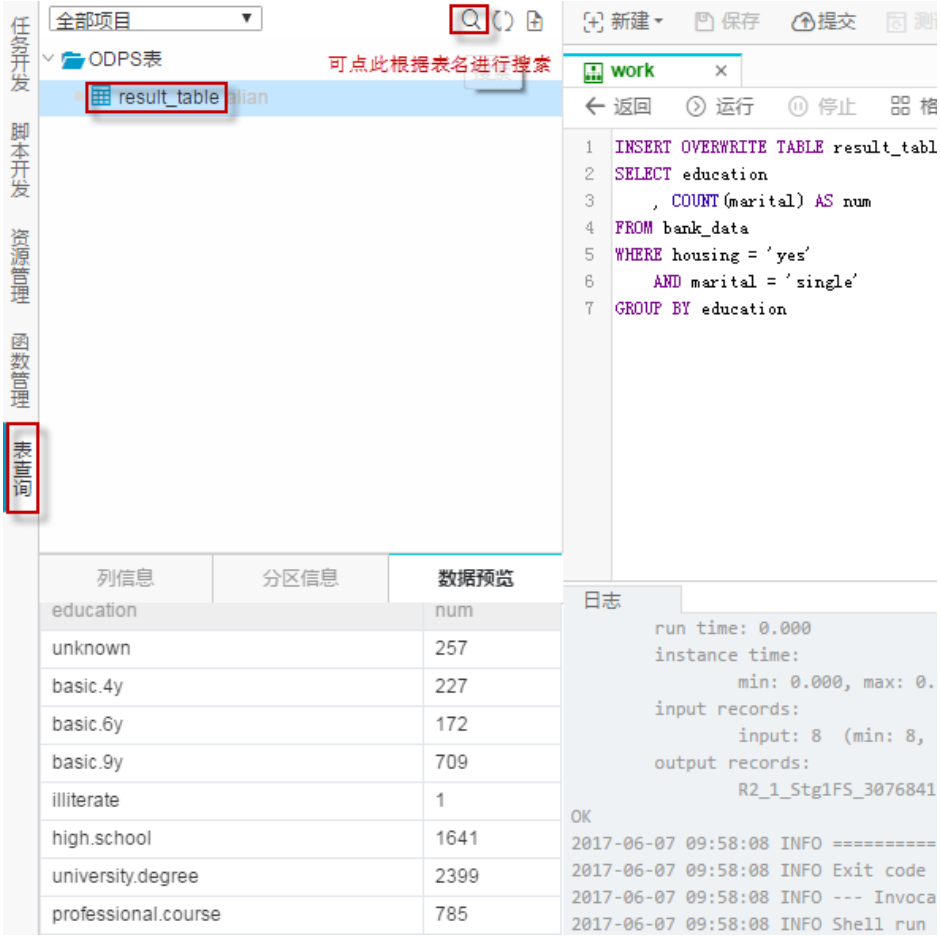
```
INSERT OVERWRITE TABLE result_table  --数据插入到result_table中
SELECT education
, COUNT(marital) AS num
FROM bank_data
WHERE housing = 'yes'
AND marital = 'single'
GROUP BY education
```

# 4. 运行并调试ODPS_SQL节点insert_data

在insert_data节点中编辑好SQL语句后，点击**保存**，防止代码丢失。然后点击**运行**，查看运行日志和结果。如下图所示：

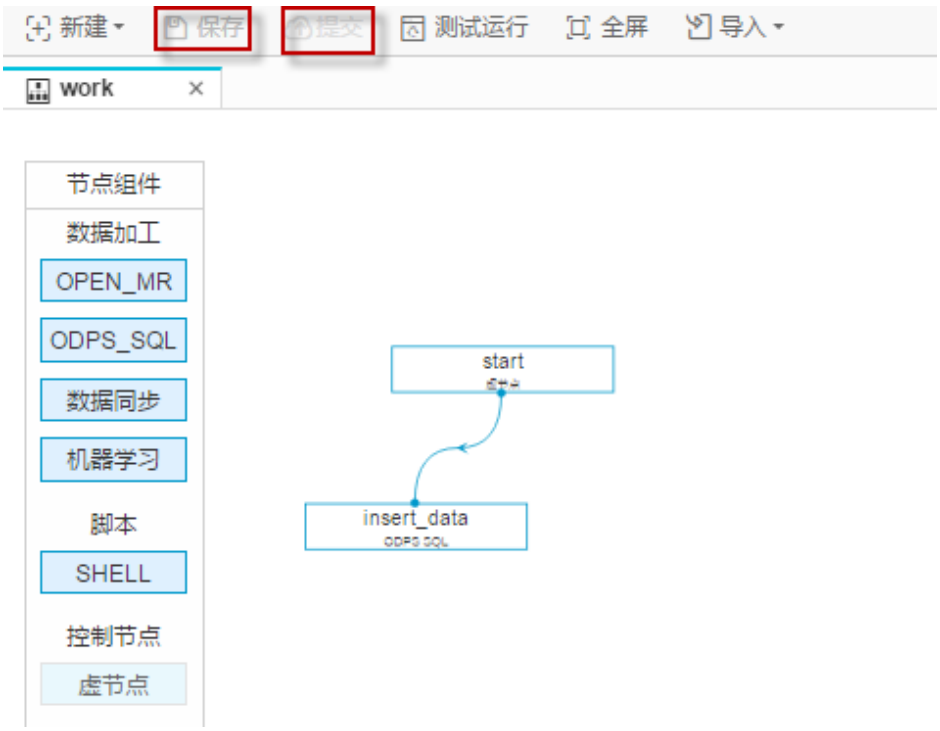



完成以上操作后，您可以在左侧的**表查询**中找到这张表里的数据：




# 5. 保存并提交工作流work

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：

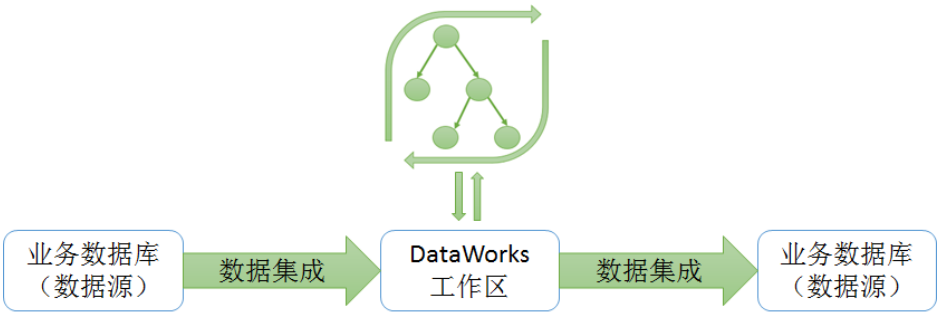



# 后续步骤

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**。

在大数据开发套件中，通常使用数据集成功能，将用户自己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用户指定的数据源中，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大数据开发套件中的数据导出到MySQL数据源中。详细操作如下：

# 1. 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。

a. 以项目管理员身份进入 **阿里云数加平台>大数据开发套件>管理控制台**，点**项目列表**下对应项目操作栏中的**进入工作区**；

b. 进入顶部菜单栏中的**数据集成**页面，点击左侧导航栏中的 **数据源**；

c. 点击右上角的 **新增数据源**，如下图所示：


d. 在新增数据源弹出框中填写相关配置项，如下图所示：


- 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。
- 数据源类型：根据自身需求进行选择，请确认选择的数据源内有表。
- 网络类型：根据自身需求进行选择。
- JDBC URL：< jdbc:mysql://host:port/database>。
- 用户名/密码：数据库对应的用户名和密码。

名词解释：

经典网络：统一部署在阿里云的公共基础网络内，网络的规划和管理由阿里云负责，更适合对网络易用性要求比较高的客户。

专有网络：基于阿里云构建出一个隔离的网络环境。您可以完全掌控自己的虚拟网络，包括选择自有的IP地址范围、划分网段、配置路由表和网关。支持公网连接，网络类型选择经典网络即可。需要注意公网

带宽的速度和相关网络费用消耗，无特殊情况不建议使用。

不同数据源类型对应的配置说明，请参见：**数据源配置**。

e. 点击**测试连通性**；

f. 若测试连通性成功，点击**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**。

# 2. 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：

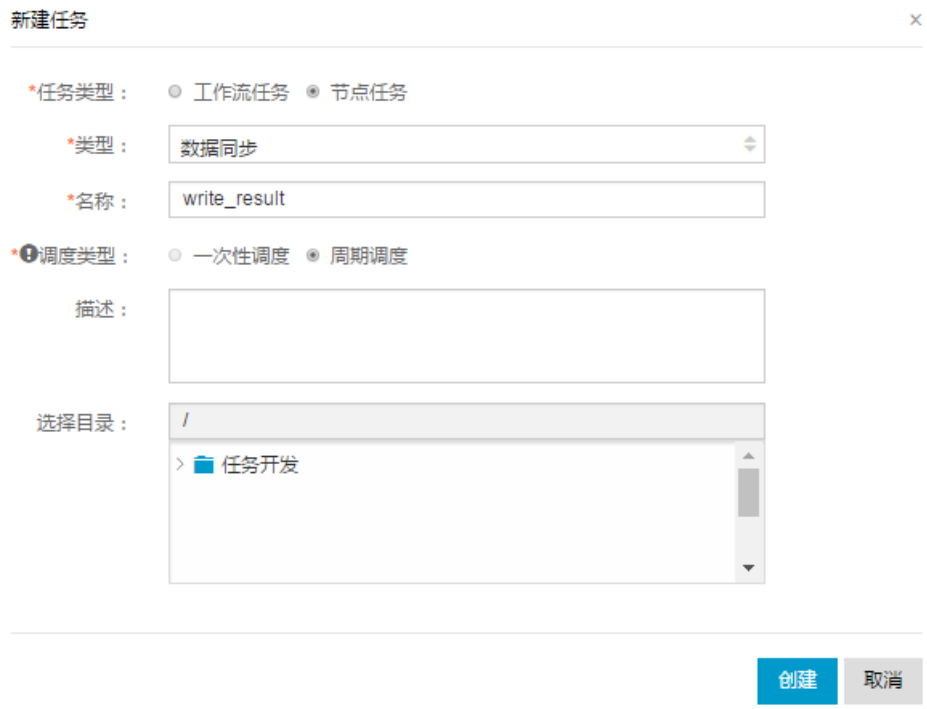
```
CREATE TABLE `ODPS_RESULT` (
`education` varchar(255) NULL ,
`num` int(10) NULL
)
```

建表完成后，可通过desc odps_result；语句查看表详情。

# 3. 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：

a. 新建同步节点write_result，如下图所示：




b. 选择来源；

选择ODPS数据源及源头表result_table，选择后点击**下一步**，如下图所示：


c. 选择目标；

选择mysql数据源及目标表odps_result，选择后点击**下一步**，如下图所示：


d. 映射字段

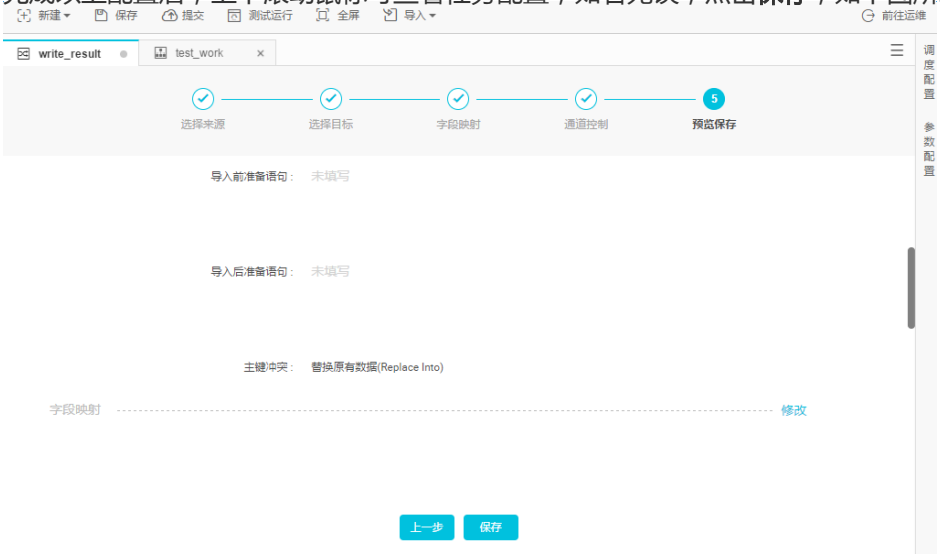
点击下一步，选择字段的映射关系。需对字段映射关系进行配置，左侧“源头表字段”和右侧“目标表字段


”为一一对应的关系 。

e. 通道控制点击下一步，配置作业速率上限和脏数据检查规则，如下图所示：


f. 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，点击**保存**，如下图所示：




# 4. 提交数据同步任务

同步任务保存后，点击右边的**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。

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，说明大数据开发套件的调度配置和任务运维功能。具体操作如下：

## 1. 为同步任务配置调度属性

a. 在数据开发—任务开发下找到需要配置的同步任务（write_result），点击右侧的**调度配置**，即可为任务配置**调度属性**，如下图所示：


- 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。

## 2. 为同步任务配置依赖属性

配置完同步任务的调度属性后，继续配置依赖属性，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达定时时间，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。

## 3. 提交同步任务

保存同步任务write_result，点击**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。

**特别说明**：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？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。

说明：

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，但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

## 1. 手动触发测试运行

a. 在工作流页面点击**测试运行**按钮，如下图所示：


b. 根据跳转页面的提示点击**确认**和**运行**，如下图所示：



c. 点击前往**运维中心**查看任务运行状态，如下图所示：


## 2. 查看测试实例的信息及运行日志

在**测试**页面下找到任务实例，右键可以查看定时时间 / 配置属性 / 代码等信息，也可以查看运行日志，如下图所示:

说明：

- 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), 则实例会在 2 点真正运行；如果不是周一，则实例在 2 点转换为成功状态并且没有日志生成。

# 补数据运行

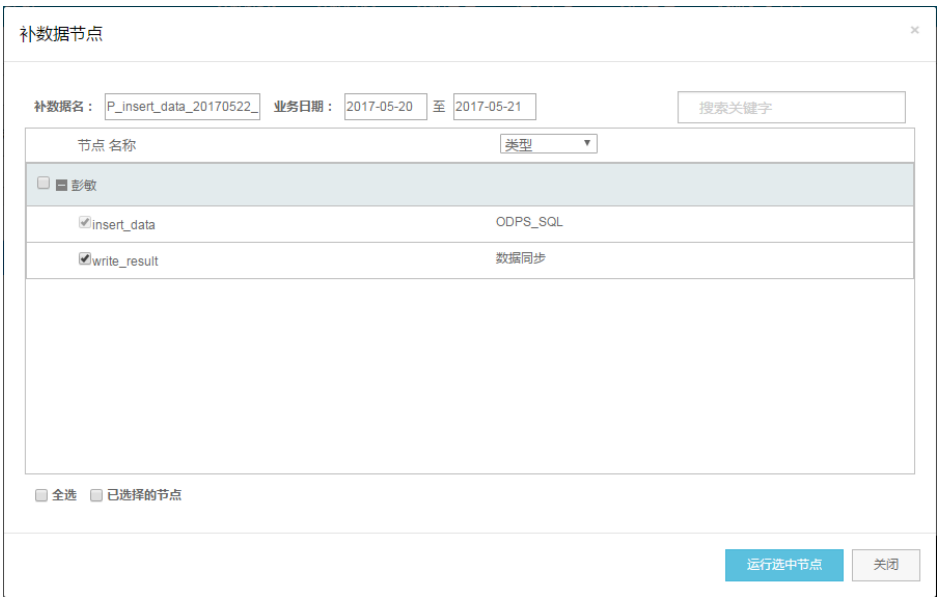
## 1. 手动触发补数据运行

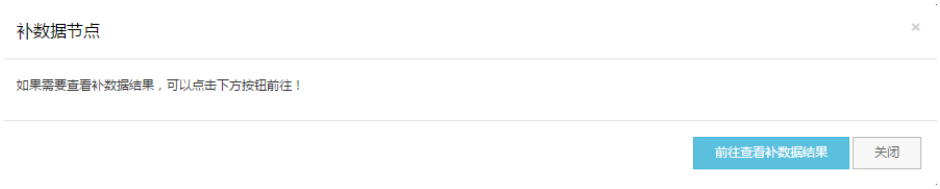
若需要确认**多个任务**的定时情况和相互依赖关系，或者需要从某个根任务开始重新执行数据分析计算，可以在**运维中心**下的**任务管理**中，选择**补数据任务**，来补跑某段时间的多个任务。操作如下：

a. 在运维中心下的任务管理中，输入任务名称；

b. 选中任务查询结果，查看右侧的任务关系图，右键单击任务，选择**补数据**节点。如下图所示：


c. 设置补数据的业务日期为 2017-05-20 到 2017-05-21 ，选择 insert_data 和 write_result 节点任务，点击**运行选中节点**。如下图所示：




d. 点击**前往查看补数据结果**。


## 2. 查看补数据实例的信息及运行日志

在**补数据**页面下找到任务实例，右键可以查看定时时间 / 配置属性 / 代码等，也可以查看运行日志，如下图所示：


说明：

- 补数据运行是手动触发方式，但生成的实例会与周期自动运行的实例存在上游依赖关系；若该任务有上游任务没有运行成功，定时时间到了不会触发运行。故补数据时建议从上游开始触发，或者当任务为“未运行”状态时，检查周期运行的下游任务是否已经运行成功不会单独运行。
- 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), 则实例会在 2 点真正运行；如果不是周一，则实例在 2 点转换为成功状态并且没有日志生成。

# 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：

点击**运维中心**下的**任务运维**中的**运维**，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：


选择任务，右键单击**查看节点运行日志**。如下图所示：


说明：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。故未运行状态的实例，请确认上游实例已经全部成功且已到定时时间。