ALIBABA CLOUD

阿里云

DataWorks

迁移助手

文档版本：20200904

阿里云

法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。

2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。

3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。

4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的“现状”、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。

5. 阿里云网站上所有内容，包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含“阿里云”、“Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。

6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

通用约定

格式	说明	样例
危险	该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。	危险 重置操作将丢失用户配置数据。
警告	该类警示信息可能会导致系统重大变更甚至故障，或者导致人身伤害等结果。	警告 重启操作将导致业务中断，恢复业务时间约十分钟。
注意	用于警示信息、补充说明等，是用户必须了解的内容。	注意 权重设置为0，该服务器不会再接受新请求。
说明	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	说明 您也可以通过按Ctrl+A选中全部文件。
>	多级菜单递进。	**单击设置> 网络> 设置网络类型。**
粗体	表示按键、菜单、页面名称等UI元素。	**在结果确认页面，单击确定。**
`Courier字体`	命令或代码。	执行 `cd /d C:/window` 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	`bae log list --instanceid` *Instance_ID*
[] 或者 [a|b]	表示可选项，至多选择一个。	`ipconfig [-all|-t]`
{} 或者 {a|b}	表示必选项，至多选择一个。	`switch {active|stand}`

目录

1.概述

迁移助手可以帮您快速复制DataWorks上不同的版本、主账号、地域和工作空间中的开发成果。

迁移助手支持迁移**周期任务**、**手动任务**、**资源**、**函数**、**数据源**、**表元数据**、**临时查询**和**组件**等对象。您可以根据业务需求，选择**全量导出**、**增量导出**或**自选导出**等方式导出DataWorks中的开发成果。

> **注意**
>
> - 目前迁移助手功能处于公测阶段，支持的地域包括华东1（杭州）、华东2（上海）、华北2（北京）、华北3（张家口）、华南1（深圳）、西南1（成都）和亚太东南1（新加坡）。
> - 仅主账号和工作空间管理员能够进行导入和导出操作，其他角色成员仅支持查看导入、导出任务列表，无操作权限。

使用场景

- 备份任务

  定期备份任务代码，减少误删项目导致数据丢失的情况。建议您选择**导出类型**为**全量导出**。

- 快速复制业务

  抽象出通用的业务流程，并快速复制业务。建议您选择**导出类型**为**自选导出**。

- 快速创建测试环境

  复制全部任务代码，快速搭建一个测试环境。您只需要修改生产数据库的数据输入为测试数据。建议您选择**导出类型**为**全量导出**或**自选导出**。

- 混合云异地开发

  导出公共云的代码至专有云中，实现同时在公共云和专有云进行数据开发。建议您选择**导出类型**为**自选导出**。

- 隔离开发环境和生产环境

  当开发环境和生产环境的网络完全隔离后，您可以通过迁移助手，导出开发环境中已完成开发的任务至生产环境。

2.任务上云

2.1. 导出开源引擎任务

DataWorks提供任务搬站功能，支持将开源调度引擎Oozie、Azkaban的任务快速迁移至DataWorks。本文为您介绍导出任务的文件要求等相关信息。

背景信息

您需要先导出开源调度引擎的任务至本地或OSS，再导入至DataWorks。导入的详情请参见导入开源引擎任务。

导出Oozie任务

导出文件的要求及结构如下：

- 导出文件的要求

  导出的文件需要包含XML和配置项等信息，导出后即为一个Zip包。

- 导出文件的结构

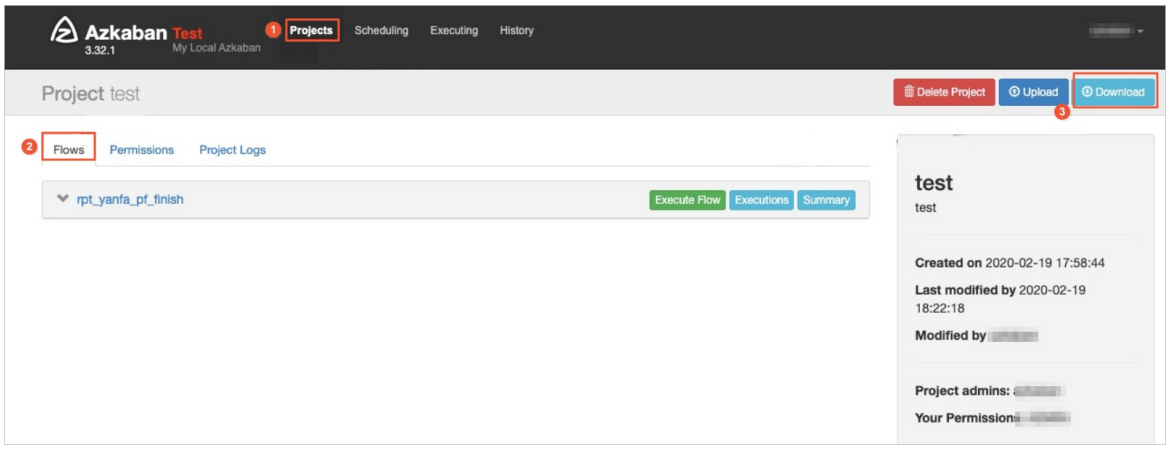
  Oozie的任务描述在HDFS的某个Path下。以Oozie官方的Examples为例，Examples包中的apps目录下，每个子目录都是一个Oozie的Workflow Job。该子目录包含Workflow的定义XML和配置项等信息。

```
$tree
.
├── aggregator
│   ├── coordinator-with-offset.xml
│   ├── coordinator.xml
│   ├── job.properties
│   ├── job-with-offset.properties
│   ├── lib
│   │   └── oozie-examples-4.2.0.jar
│   └── workflow.xml
├── sqoop
│   ├── db.hsqldb.properties
│   ├── db.hsqldb.script
│   ├── job.properties
│   └── workflow.xml
├── cron
│   ├── coordinator.xml
│   ├── job.properties
│   └── workflow.xml
├── cron-schedule
│   ├── coordinator.xml
│   ├── job.properties
│   └── workflow.xml
```

导出Azkaban任务

Azkaban有自己的Web控制台，支持在界面下载某个工作流（Flow）：

1. 登录Azkaban控制台的Projects页面。
2. 进入相应的Project页面，单击Flows，为您展示该Project下所有的工作流。
3. 单击页面右上方的Download，下载Project的导出文件。



Azkaban导出包的格式无特别限制，是原生Azkaban即可。导出包Zip文件内部为Azkaban的某个Project下，所有任务（Job）和关系的信息。

导出其它开源引擎任务

DataWorks为您提供标准模板便于导出除Oozie和Azkaban外的开源引擎任务。导出任务前，您需要下载**标准格式模板**并参考模板的文件结构修改内容。下载模板及目录结构的介绍请进入**开源引擎导出**页面进行查询：

1. 进入**数据开发**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。
2. 单击左上方的图标，选择**全部产品 > 其他 > 迁移助手**。
3. 在左侧导航栏，单击**任务上云 > 开源引擎导出**，进入**开源引擎导出方案选择**页面。
4. 单击**标准模板**。
5. 在**标准模板**页签下，单击**标准格式模板**进行下载。
6. 根据模板中的格式修改内容后，即可生成导出包。

2.2. 导入开源引擎任务

本文为您介绍如何导入从开源引擎导出的任务至DataWorks。

操作步骤

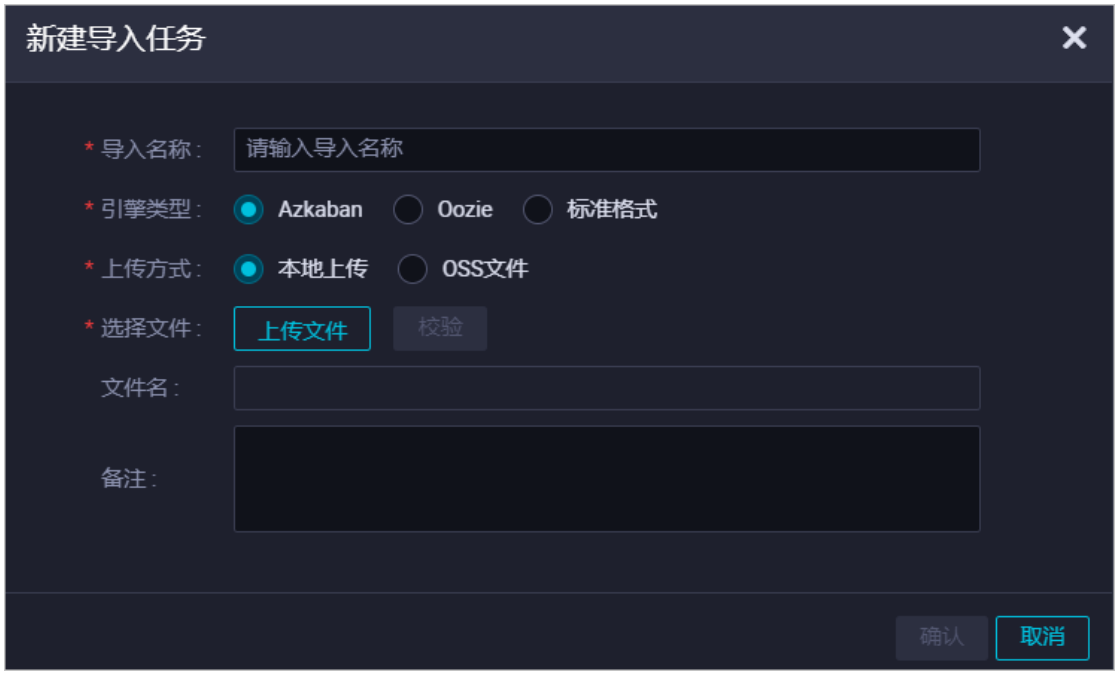
1. 进入**开源引擎导入**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。
   iv. 单击左上方的图标，选择**全部产品 > 其他 > 迁移助手**。

v. 在左侧导航栏，单击**任务上云 > 开源引擎导入**。

2. 创建导入任务。

i. 在**开源引擎导入**页面，单击右上方的**新建导入任务**。

ii. 在**新建导入任务**对话框中，配置各项参数。



参数	描述
导入名称	导入任务的名称，仅支持大小写字母、中文、数字、下划线（_）和英文句号（.）。
引擎类型	包括Azkaban、Oozie和标准格式。
上传方式	包括本地上传和OSS文件： ■ 当您选择上传方式为本地上传时，请进行以下操作： a. 单击上传文件。 b. 在本地选择需要上传的文件，单击打开。 c. 单击校检。 d. 待页面显示资源包校检成功，确认文件的格式和内容无误。 说明 本地文件最多支持上传30 MB。如果超过30 MB，请使用OSS文件上传。 ■ 当您选择上传方式为OSS文件时，请输入OSS链接，并进行校检和文件预览。
备注	对导入任务进行简单描述。

iii. 单击**确认**，进入**编辑导入任务**页面。

3. 编辑导入任务。

i. 在**编辑导入任务**页面，筛选**导入对象**。该页面默认显示**导入对象**为**周期任务**的对象。如果您导入的是其它类型的对象，请在**导入对象**下拉列表中进行筛选。

ii. （可选）单击**高级设置**，在**高级设置**对话框中，设置节点类型和计算引擎的映射关系。如果目标工作空间有多个计算引擎，请进行**高级设置**，选择任务的映射关系。目前支持设置映射的任务类型包括Shell、Hive和Sqoop。

iii. 单击页面右上方的**开始导入**。

4. 查看导入报告。

i. 在**导入进度**对话框中，确认任务的导入进度。

ii. 待任务导入成功后，单击**返回导入任务列表**。

iii. 在**导入任务列表**页面，单击相应任务后的**查看导入报告**，查看导入任务的**基本信息**、**导入结果**、**明细**和**导入设置**。

3.DataWorks迁移

3.1. 创建和查看DataWorks导出任务

迁移助手支持导出周期任务、手动任务、资源、函数、表元数据、数据源、组件和临时查询等对象，本文为您介绍如何创建和查看导出任务。

前提条件

- 目前迁移助手功能处于公测阶段，支持的地域包括华东1（杭州）、华东2（上海）、华北2（北京）、华北3（张家口）、华南1（深圳）、西南1（成都）和亚太东南1（新加坡）。
- 仅主账号和工作空间管理员能够进行导入和导出操作，其他角色成员仅支持查看导入、导出任务列表，无操作权限。

背景信息

迁移助手支持通过**全量导出**、**增量导出**和**自选导出**等方式导出任务。不同导出类型的使用场景如下：

- **全量导出**适用于全量备份工作空间中的任务，主要用于备份代码、快速复制一个测试环境等场景。全量导出的版本为开发过程中最新的版本。

  全量导出仅支持导出保存成功的对象。当同一个任务有开发和生产等多个版本时，以开发侧保存的版本为主进行全量导出。

- **增量导出**基于对象的最后修改时间，筛选最近修改过的对象并导出。

  > **说明** 增量导出不支持选择导出黑名单。

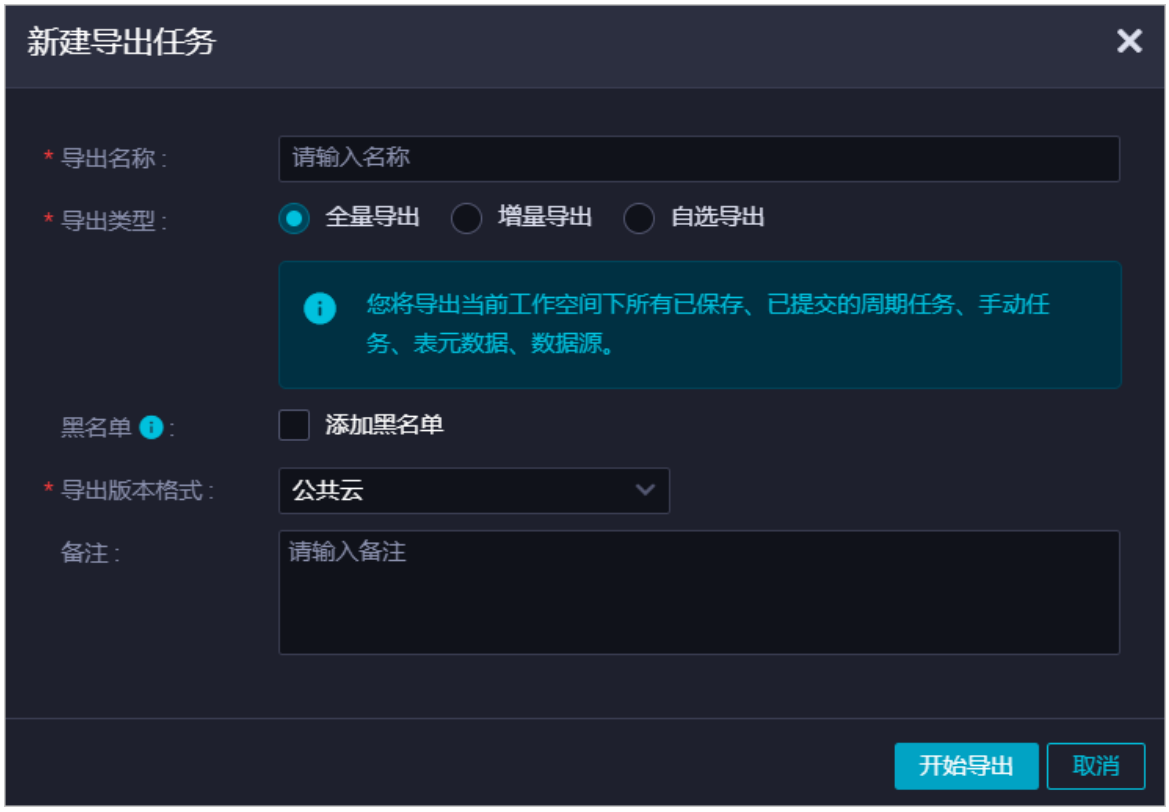
- **自选导出**适用于抽象出通用的业务流程，以便其它业务快速复制。在开发和生产集群完全隔离的状态下，您可以通过自选导出功能，完成类似于发布任务的操作。

进入迁移助手

1. 登录DataWorks控制台。
2. 在左侧导航栏，单击**工作空间列表**。
3. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。
4. 单击左上方的图标，选择**全部产品 > 其他 > 迁移助手**，默认进入**DataWorks迁移 > DataWorks导出**页面。

创建全量导出任务

1. 在**DataWorks导出**页面，单击右上方的**新建导出任务**。
2. 在**新建导出任务**对话框中，配置各项参数。



参数	描述
导出名称	导出名称仅支持大小写字母、中文、数字、下划线和小数点。
导出类型	选择全量导出，您将导出当前工作空间下所有已保存、已提交的周期任务、手动任务、表元数据和数据源。
黑名单	您可以根据业务需求决定是否选中添加黑名单，以筛选全量导出过程中无需导出的任务和资源。
导出版本格式	包括公共云和专有云（V3.6.1-V3.11）。DataWorks上不同版本的数据格式不一致，请先确认待导入环境中DataWorks的版本。
备注	对导出任务进行简单描述。

3. （可选）添加黑名单并导出任务。如果您选中添加黑名单，请执行下述操作：
   i. 在新建导出任务对话框中，单击添加黑名单。
   ii. 在选择黑名单页面，选择无需导出的对象。
   iii. 单击添加到黑名单。
   iv. 单击页面右上方的开始导出。
   v. 在导出确认对话框中，单击确认。
4. （可选）如果您未选中添加黑名单，请直接单击开始导出。
5. 在导出进度对话框中，查看任务的导出进度。待导出成功后，单击返回导出任务列表。

创建增量导出任务

1. 在DataWorks导出页面，单击右上方的新建导出任务。
2. 在新建导出任务对话框中，配置各项参数。

参数	描述
导出名称	导出名称仅支持大小写字母、中文、数字、下划线和小数点。
导出类型	选择增量导出，您将导出指定日期后修改的文件，包括已保存、已提交的周期任务、手动任务、表元数据和数据源。
增量开始日期	按照最后修改的时间进行增量。
导出版本格式	包括公共云和专有云（V3.6.1-V3.11）。
备注	对导出任务进行简单描述。

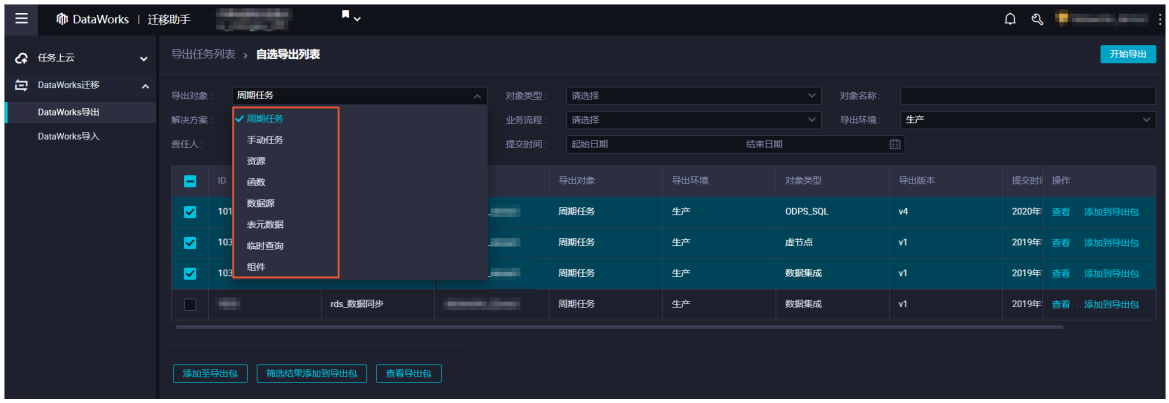
3. 单击开始导出。

创建自选导出任务

1. 在DataWorks导出页面，单击右上方的新建导出任务。
2. 在新建导出任务对话框中，配置各项参数。

参数	描述
导出名称	导出名称仅支持大小写字母、中文、数字、下划线和小数点。
导出类型	选择自选导出，您可以自由选择需要导出的文件，包括已保存、已提交的周期任务、手动任务、表元数据和数据源。
导出版本格式	包括公共云和专有云（V3.6.1-V3.11）。
备注	对导出任务进行简单描述。

3. 单击选择导出内容。
4. 在自选导出列表页面，根据导出对象的类型筛选需要导出的对象。迁移助手支持的导出对象包括周期任务、手动任务、资源、函数、数据源、表元数据、临时查询和组件。



5. 选中需要导出的对象后，单击添加至导出包。您也可以根据导出对象、对象类型和导出环境等条件进行筛选，单击筛选结果添加到导出包，添加筛选的所有结果。
6. 单击页面右上方的开始导出。

任务状态

在导出任务列表页面，您可以查看导出任务的导出任务名称、导出类型、任务创建人、状态、更新时间和备注等信息，不同状态的任务可以进行不同的操作：

- 当导出任务的状态为导出成功时，您可以：
  - 单击查看导出报告，查看导出任务的基本信息、概览和明细。

  - 单击下载导出包，下载导出任务至本地。
  - 克隆不同导出类型的任务：
    - 单击导出类型为全量导出，且未添加黑名单任务后的克隆。在克隆对话框中，输入导出名称，单击开始导出。
    - 单击导出类型为全量导出，且已添加黑名单任务后的克隆。在克隆对话框中，输入导出名称，单击添加黑名单。

      在选择黑名单页面，选择无需导出的对象，单击添加到黑名单后，单击页面右上方的开始导出。
    - 单击导出类型为自选导出任务后的克隆。在克隆对话框中，输入导出名称，单击选择导出内容。

      在自选导出列表页面，选择需要导出的对象，单击添加至导出包后，单击页面右上方的开始导出。
- 当导出任务的状态为导出失败时，除查看导出报告和下载导出包外，您还可以单击重试，重新导出任务。
- 当导出任务的导出类型为自选导出、状态为编辑中时，您可以：
  - 单击编辑，在任务的编辑页面根据导出任务类型进行不同的操作。
  - 单击查看导出包，在导出包详情页面查看导出任务的基本信息、概览和明细。
  - 单击删除，在删除对话框中，单击确认。
- 当导出任务的导出类型为全量导出、状态为编辑中时，除可以编辑和删除导出任务外，您还可以单击查看黑名单。在黑名单预览对话框中，确认添加的黑名单，单击开始导出或关闭。

3.2. 创建和查看DataWorks导入任务

您在DataWorks导出任务后，可以将其导入至相应的工作空间中，完成任务的迁移。

前提条件

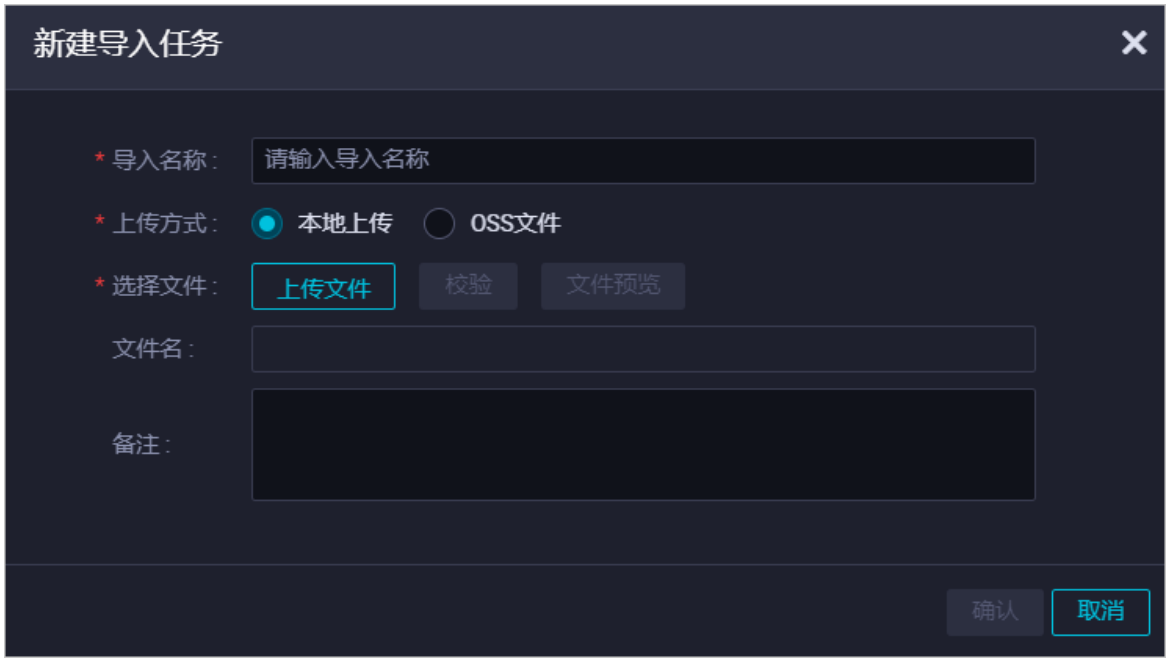
- 目前迁移助手功能处于公测阶段，支持的地域包括华东1（杭州）、华东2（上海）、华北2（北京）、华北3（张家口）、华南1（深圳）、西南1（成都）和亚太东南1（新加坡）。
- 仅主账号和工作空间管理员能够进行导入和导出操作，其他角色成员仅支持查看导入、导出任务列表，无操作权限。

进入迁移助手

1. 登录DataWorks控制台。
2. 在左侧导航栏，单击**工作空间列表**。
3. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。
4. 单击左上方的图标，选择**全部产品 > 其他 > 迁移助手**，默认进入**DataWorks迁移 > DataWorks导出**页面。

创建导入任务

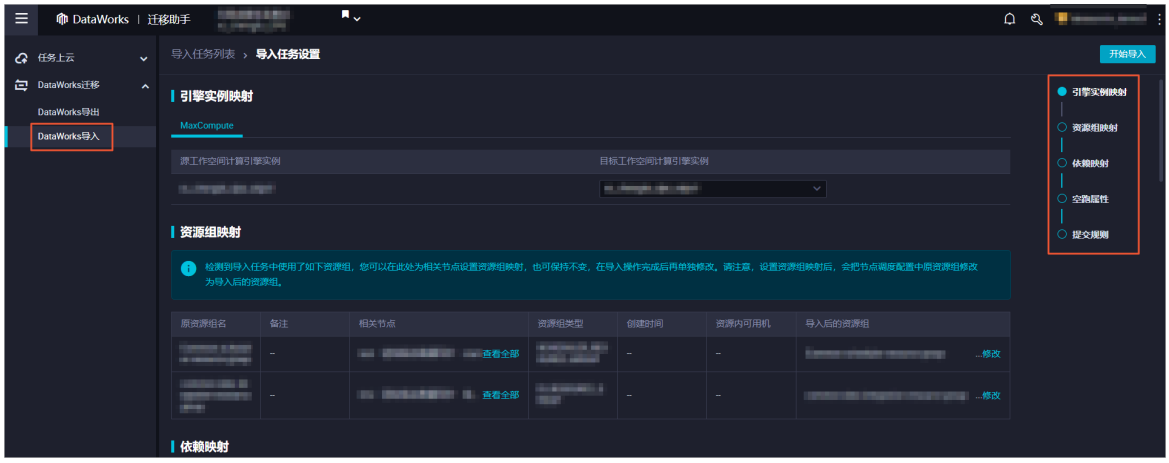
1. 在迁移助手的左侧导航栏，单击**DataWorks迁移 > DataWorks导入**。
2. 在**DataWorks导入**页面，单击右上方的**新建导入任务**。
3. 在**新建导入任务**对话框中，配置各项参数。



参数	描述
导入名称	导入名称仅支持大小写字母、中文、数字、下划线（_）和英文句号（.）。

参数	描述
上传方式	包括本地上传和OSS文件： ○ 当您选择上传方式为本地上传时，请进行以下操作： a. 单击上传文件。 b. 在本地选择需要上传的文件，单击打开。 c. 单击校检。 d. 待页面显示资源包校检成功后，您可以单击文件预览，查看待导入的文件。 说明 本地文件最多支持上传30 MB。 ○ 当您选择上传方式为OSS文件时，请输入OSS链接，并进行校检和文件预览。
备注	对导入任务进行简单描述。

4. 单击确认，进入导入任务设置页面。导入任务前，您需要校检导入文件的格式和内容。通过校检后，才可以单击确认。
5. 配置导入任务。配置导入任务时，必须配置引擎实例映射。其它配置为可选操作，您可以根据业务需求设置。



说明 如果是同租户、同地域下不同工作空间的互导，您只需要设置引擎实例映射。

（可选）

i. 在引擎实例映射区域，设置源工作空间和目标工作空间的计算引擎实例映射关系。如果源工作空间绑定多种类型的计算引擎，目标工作空间仅绑定一种类型的计算引擎，则目标工作空间无其它类型节点的创建权限，导致导入任务失败。

ii. （可选）在资源组映射区域，修改源工作空间和目标工作空间的资源组映射关系，避免出现运行任务时无法找到资源组的情况。

iii. （可选）在依赖映射区域，为相关节点设置项目映射。在导入任务时，有任务代码使用源工作空间名称。您可以修改新项目名，修改范围为任务代码、本节点输入名称和输出名称。待导入完成后，会快速替换为新的工作空间名称。

iv. （可选）在**空跑属性**区域，单击相应节点后的**设置空跑**。您也可以选中多个需要空跑的节点，单击**批量设置空跑**。

该配置项用于为周期任务设置调度参数中的时间属性。设置空跑后，节点会直接运行成功，不会生成数据。

v. （可选）在**提交规则**区域，您可以设置**资源**、**函数**和**表**的提交规则，并可以**修改责任人**。

> **说明**
> - 如果目标工作空间已存在同名的对象，会出现提交失败的情况。
> - 如果您选择不修改责任人，且源任务无责任人，则会设置提交人为任务的责任人。

6. 单击右上方的**开始导入**。
7. 在**请确认**对话框中，单击**确认**。

查看导入任务

在**导入任务列表**页面，不同状态的任务会显示不同的操作：

- 待任务导入成功后，您可以在**导入任务列表**页面，单击相应任务后的**查看导入报告**，查看导入任务的**基本信息**、**导入结果**、**明细**和**导入设置**。

- 如果是**编辑中**的任务，您可以进行以下操作：
  - 单击相应任务后的**继续编辑**，在**导入任务设置**页面修改任务的配置。
  - 单击相应任务后的**预览**，查看导入文件的**基本信息**、**概览**和**明细**。
  - 单击相应任务后的**删除**，在提示框中单击**确认**，删除该导入任务。
- 如果是**导入失败**的任务，您可以单击相应任务后的**重新导入**。在**导入进度**对话框中，确认导入完成后，请单击**返回导入任务列表**。