ALIBABA CLOUD

Alibaba Cloud

DataWorks

产品简介

文档版本：20201015

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

1.什么是DataWorks

本文为您介绍什么是DataWorks，以及DataWorks的功能和使用限制。

DataWorks（数据工场，原大数据开发套件）是阿里云重要的PaaS（Platform-as-a-Service）平台产品，为您提供数据集成、数据开发、数据地图、数据质量和数据服务等全方位的产品服务，一站式开发管理的界面，帮助企业专注于数据价值的挖掘和探索。

DataWorks支持多种计算和存储引擎服务，包括离线计算MaxCompute、开源大数据引擎E-MapReduce、实时计算（基于Flink）、机器学习PAI、图计算服务Graph Compute和交互式分析服务等，并且支持用户自定义接入计算和存储服务。DataWorks为您提供全链路智能大数据及AI开发和治理服务。

您可以使用DataWorks，对数据进行传输、转换和集成等操作，从不同的数据存储引入数据，并进行转化和开发，最后将处理好的数据同步至其它数据系统。

使用限制

仅支持Chrome浏览器54以上版本。

功能概述

- 全面托管的调度
  - DataWorks提供强大的调度功能，详情请参见调度配置。
    - 支持根据时间、依赖关系，进行任务触发的机制。详情请参见配置时间属性和依赖关系。
    - 支持每日千万级别的任务，根据DAG关系准确、准时地运行。
    - 支持分钟、小时、天、周和月多种调度周期配置。
  - 完全托管的服务，无需关心调度的服务器资源问题。
  - 提供隔离功能，确保不同租户之间的任务不会相互影响。
- DataWorks支持离线同步、Shell、ODPS SQL、ODPS MR等多种节点类型，通过节点之间的相互依赖，对复杂的数据进行分析处理。
  - 数据转化：依托MaxCompute强大的能力，保证了大数据的分析处理性能。
  - 数据同步：依托DataWorks中数据集成的强力支撑，支持超过20种数据源，为您提供稳定高效的数据传输功能。详情请参见数据集成和支持的数据源与读写插件。
- 可视化开发

  DataWorks提供可视化的代码开发、工作流设计器页面，无需搭配任何开发工具，简单拖拽和开发，即可完成复杂的数据分析任务。详情请参见界面功能点介绍。

  只要有浏览器有网络，您即可随时随地进行开发工作。
- 监控告警

  运维中心提供可视化的任务监控管理工具，支持以DAG图的形式展示任务运行时的全局情况，详情请参见运维中心。

  您可以方便地配置各类报警方式，任务发生错误可及时通知相关人员，保证业务正常运行。详情请参见智能监控。

2.基本概念

本文为您介绍DataWorks中，工作空间、业务流程、解决方案、组件、任务、实例、提交、脚本开发、资源、函数和输出名称等基本概念。

工作空间

工作空间是DataWorks管理任务、成员，分配角色和权限的基本单元。工作空间管理员可以加入成员至工作空间，并赋予工作空间管理员、开发、运维、部署、安全管理员或访客角色，以实现多角色协同工作。

> 说明 建议您根据部门或业务板块来划分工作空间。

一个工作空间支持绑定MaxCompute、E-MapReduce和实时计算等多种类型的计算引擎实例。绑定引擎实例后，即可在工作空间开发和调度引擎任务。

业务流程

针对业务实体，抽象出业务流程的概念，帮助您从业务视角组织代码的开发，提高任务管理效率。

> 说明 业务流程可以被多个解决方案复用。

业务流程帮助您从业务视角组织代码：

- 支持基于任务类型的代码组织方式。
- 支持多级子目录（建议不超过四级）。
- 支持从业务视角查看整体的业务流程，并进行优化。
- 支持根据业务流程组织发布和运维。
- 提供业务流程看板，帮助您更高效地进行开发。

解决方案

您可以自定义组合部分业务流程为一个解决方案。

解决方案的优势如下：

- 一个解决方案可以包括多个业务流程。
- 解决方案之间可以复用相同的业务流程。
- 组织完成的解决方案包含各类节点，可以让您进行沉浸式开发。

组件

您可以将SQL中的通用逻辑抽象为组件，提高代码的复用性。

SQL代码的处理过程通常是引入一到多个源数据表，通过过滤、连接和聚合等操作，加工出新的业务需要的目标表。组件是带有多个输入参数和输出参数的SQL代码过程模板。

任务（Task）

任务是对数据执行的操作的定义，示例如下：

- 通过数据同步节点任务，将数据从RDS同步至MaxCompute。
- 通过MaxCompute SQL节点任务，运行MaxCompute SQL来进行数据的转换。

每个任务使用0或0个以上的数据表（数据集）作为输入，生成一个或多个数据表（数据集）作为输出。

任务主要分为节点任务（Node Task）、工作流任务（Flow Task）和内部节点（innerNode）。

任务类型	描述
节点任务（Node Task）	一个数据执行的操作。可以与其它节点任务、工作流任务配置依赖关系，组成DAG图。
工作流任务（Flow Task）	满足一个业务场景需求的一组内部节点，组成一个工作流任务，建议工作流任务小于10个。 工作流任务内部节点，无法被其它工作流任务、节点任务依赖。工作流任务可以与其它工作流任务、节点任务配置依赖关系，组成DAG图。 说明 从DataWorks V1.0升级的任务，仍保留工作流的概念。DataWorks V2.0及以上版本已无法创建工作流任务，您可以选择创建业务流程进行后续操作。
内部节点（innerNode）	工作流任务内部的节点，与节点任务的功能基本一致。您可以通过拖拽形成依赖关系，其调度周期会继承工作流任务的调度周期，无法进行单独配置。

实例（Instance）

实例是某个任务在某时某刻执行的一个快照。调度系统中的任务，经过调度系统、手动触发运行后，会生成一个实例。实例中会有任务的运行时间、运行状态和运行日志等信息。

例如设置每天2:00运行Task1实例，调度系统会在每天23:30根据周期节点定义好的时间，自动生成一个快照，即Task1第二天2:00运行的实例。到第二天2:00时，如果判断上游实例已经完成，Task1实例便会如期启动运行。

> 说明 您可以进入**运维中心 > 周期任务运维**页面，查询实例的相关信息。

提交（Submit）

提交是指开发的节点任务、业务流程，从DataWorks开发环境发布至调度系统的过程。完成提交后，相应的代码、调度配置全部合并至调度系统中，调度系统根据相关配置进行调度操作。

> 说明 未提交的节点任务、业务流程不会进入调度系统。

脚本开发（Script）

脚本开发是提供给数据分析使用的一个代码存储空间。脚本开发的代码无法发布到调度系统，无法进行调度参数配置，仅可以进行部分数据查询分析的工作。

资源、函数

资源、函数均为MaxCompute的概念，详情请参见资源和函数。

您可以在DataWorks中，通过界面管理资源和函数。如果通过MaxCompute的其它方式进行资源、函数管理，则无法在DataWorks中进行相关的查询。

输出名称

输出名称：每个任务（Task）输出点的名称。它是您在单个租户（阿里云账号）内设置依赖关系时，用于连接上下游两个任务（Task）的虚拟实体。

当您在设置某任务与其它任务形成上下游依赖关系时，必须根据输出名称（而不是节点名称或节点ID）来完成设置。设置完成后该任务的输出名也同时作为其下游节点的输入名称。

> **说明** 输出名称可以作为某个Task在同租户内，区别于其它Task的唯一概念对象，每个节点的输出名称默认为工作空间名称.系统生成9位数字.out。您可以对Task增加自定义输出名，但需要注意输出节点名称在租户内不允许重复。

3.应用场景

本文将为您介绍DataWorks的应用场景示例。

日志大数据分析

- 提高工作效率。

  将日志数据同步至MaxCompute，通过SQL语句进行分析与处理，提高工作效率。

- 提高存储利用率。

  降低整体存储和计算的费用的同时，并提高性能和稳定性。

- 降低大数据使用门槛。

  MaxCompute提供多种开源软件的插件，可以轻松完成数据上云。

推荐搭配使用：

DataWorks + 数据集成 + AnalyticDB for MySQL + Quick BI + MaxCompute

精细化运营

- 提升业务洞察能力。

  通过MaxCompute计算能力，可以实现针对百万用户的精细化运营。

- 业务数据化。

  可以提升对业务数据的分析能力并进行有效监控，更好地业务赋能。

- 快速响应业务需求。

  可以根据新业务的数据分析需求，快速灵活地进行响应与满足。

推荐搭配使用：

DataWorks + 数据集成 + Quick BI + MaxCompute

数据安全管理

- 敏感数据识别。

  通过用户自定义规则，自动识别敏感数据，并标记对应的级别。

- 敏感数据展示脱敏。

  提供设置脱敏规则功能，实现敏感数据查询展示脱敏。

- 敏感数据操作风险监控。

  可视化监控数据分布、数据使用和数据导出，提供自定义风险识别和审计功能。

推荐使用DataWorks的数据保护伞功能。

4.数据开发流程

通常数据开发的总体流程包括数据产生、数据收集与存储、数据分析与处理、数据提取和数据展现与分享。

> **说明** 上图中，虚线框内的开发流程均可基于阿里云大数据平台完成。

数据开发的流程如下所示：

1. **数据产生**：业务系统每天会产生大量结构化的数据，存储在业务系统所对应的数据库中，包括MySQL、Oracle和RDS等类型。
2. **数据收集与存储**：您需要同步不同业务系统的数据至MaxCompute中，方可通过MaxCompute的海量数据存储与处理能力分析已有的数据。

   DataWorks提供数据集成服务，可以支持多种数据源类型，根据预设的调度周期同步业务系统的数据至MaxCompute。
3. **数据分析与处理**：完成数据的同步后，可以对MaxCompute中的数据进行加工（MaxCompute SQL、MaxCompute MR）、分析与挖掘（数据分析、数据挖掘）等处理，从而发现其价值。
4. **数据提取**：分析与处理后的结果数据，需要同步导出至业务系统，以供业务人员使用其分析的价值。
5. **数据展现与分享**：数据提取成功后，可以通过报表、地理信息系统等多种展现方式，展示与分享大数据分析、处理后的成果。

5.简单模式和标准模式的区别

为方便不同安全管控要求的用户生产数据，DataWorks为您提供简单模式和标准模式两种工作空间模式。本文为您介绍两种模式工作空间的区别和数据访问模式。

简单模式的工作空间

简单模式下，一个Dataworks工作空间对应一个计算引擎（项目、实例或数据库），无法设置开发环境和生产环境，只能进行简单的数据开发。简单模式的工作空间无法对数据开发流程和表权限进行强控制。

您使用简单模式工作空间的优势和风险如下：

- 优势：使用方便。提交代码后，您无需发布，代码即可进入调度系统周期性执行，产出结果数据。
- 风险：开发角色可以不经过任何人审批，随时新增、修改代码并提交至调度系统，给生产环境带来不稳定因素。同时，当面向MaxCompute计算引擎时，开发角色默认拥有当前MaxCompute项目所有表的读写权限。开发角色的用户可以随意对表进行增加、删除和修改等操作，存在数据安全风险。

以MaxCompute为例，简单模式工作空间的流程如下。

标准模式的工作空间

标准模式的工作空间下，一个DataWorks工作空间对应两个计算引擎（项目、实例或数据库）。与简单模式的工作空间相比，标准模式的工作空间有如下不同：

- 所有代码仅支持在开发环境编辑，您无法修改生产环境的代码。
- 提交任务后，任务会进入开发环境调度系统。但实际不会自动调度，仅作为冒烟测试使用。如果您需要自动调度运行任务，请发布任务至生产环境。

  发布任务前，需要项目管理员或运维角色的成员审批通过，才能发布成功。

以MaxCompute为例，标准模式工作空间的流程如下。

不同模式工作空间的数据访问模式

您可以在**工作空间配置 > 计算引擎信息**区域，设置不同模式下，工作空间的数据访问模式。详情请参见配置工作空间。

工作空间模式	计算引擎类型	环境	访问身份
标准模式	MaxCompute	开发环境	页面运行任务（不可选）：默认为执行任务者（当前登录者）
		生产环境	调度访问身份（可选）： • 阿里云主账号 • 阿里云子账号
	EMR（E-MapReduce）	开发环境	页面运行任务和调度访问身份均统一设置，即**新增EMR集群**对话框中输入的**Access ID**和**Access Key**对应的访问身份。
		生产环境	
		开发环境	页面运行任务（不可选）：默认为执行任务者（当前登录者）

工作空间模式	计算引擎类型 Hologres	环境	访问身份
		生产环境	调度访问身份（可选）： • 阿里云主账号 • 阿里云子账号
简单模式	MaxCompute	开发环境即生产环境	页面运行任务（不可选）：默认为执行任务者（当前登录者） 调度访问身份（可选）： • 任务责任人：任务Owner账号的身份 • 阿里云主账号
	EMR	开发环境即生产环境	页面运行任务和调度访问身份均统一设置，即**新增EMR集群**对话框中输入的**Access ID**和**Access Key**对应的访问身份。
	Hologres	开发环境即生产环境	页面运行任务（不可选）：默认为执行任务者（当前登录者） 调度访问身份（可选）： • 阿里云主账号 • 阿里云子账号