

# 阿里云 专有云企业版 DataWorks

## 产品简介

产品版本：V3.12.0

文档版本：20200706

# 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. 阿里云文档中所有内容，包括但不限于图片、架构设计、页面布局、文字描述，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
	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。	 <b>禁止：</b> 重置操作将丢失用户配置数据。
	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。	 <b>警告：</b> 重启操作将导致业务中断，恢复业务时间约十分钟。
	用于警示信息、补充说明等，是用户必须了解的内容。	 <b>注意：</b> 权重设置为0，该服务器不会再接受新请求。
	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	 <b>说明：</b> 您也可以通过按Ctrl + A选中全部文件。
>	多级菜单递进。	单击 <b>设置 &gt; 网络 &gt; 设置网络类型</b> 。
<b>粗体</b>	表示按键、菜单、页面名称等UI元素。	在 <b>结果确认</b> 页面，单击 <b>确定</b> 。
Courier字体	命令。	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

法律声明.....	I
通用约定.....	I
1 什么是DataWorks.....	1
2 产品优势.....	2
3 产品架构.....	4
3.1 功能架构.....	4
3.2 系统架构.....	5
3.3 安全架构.....	5
3.4 多租户模型.....	6
4 功能特性.....	7
4.1 数据集成.....	7
4.2 数据开发.....	8
4.2.1 数据开发概述.....	8
4.2.2 业务流程设计器.....	9
4.2.3 解决方案设计器.....	9
4.2.4 代码开发编辑器.....	10
4.2.5 代码管理与团队协作.....	11
4.3 监控运维.....	11
4.3.1 监控运维概述.....	11
4.3.2 运维概览.....	11
4.3.3 任务运维.....	11
4.3.4 智能监控.....	12
4.4 实时分析.....	12
4.5 数据服务.....	13
4.6 平台管理.....	13
4.7 数据资产管理.....	15
4.8 数据保护伞.....	15
4.8.1 数据保护伞概述.....	15
4.8.2 基本术语.....	16
4.8.3 规则配置.....	17
4.8.4 数据发现.....	17
4.8.5 数据访问.....	17
4.8.6 数据脱敏管理.....	18
4.8.7 分级信息管理.....	18
4.8.8 手动修正数据.....	18
4.8.9 数据风险.....	18
4.8.10 风险识别管理.....	18
4.8.11 数据审计.....	18
5 应用场景.....	19

5.1 云上数仓.....	19
5.2 BI应用.....	20
5.3 数据化运营.....	22
<b>6 使用限制.....</b>	<b>23</b>
<b>7 基本概念.....</b>	<b>24</b>



# 1 什么是DataWorks

DataWorks（数据工场，原大数据开发套件）是基于MaxCompute、EMR（E-MapReduce）等计算引擎、从工作室、车间到工具集都齐备的一站式大数据智能研发与治理平台，它能助力您快速完成数据集成、开发、治理、服务、质量、安全等全套数据研发治理工作。

DataWorks不仅具备海量数据的离线加工分析、数据挖掘的能力，也集成了数据集成、数据开发、生产运维、实时分析、资产管理、数据质量、数据安全、数据共享等核心数据工艺，同时还提供了数据服务、机器学习（PAI）在线研发平台，承上启下，让数据从采集到展现、从分析到驱动应用得以一站式解决。

DataWorks + MaxCompute在2018年获得著名分析评测机构Forrester的Cloud Data Warehouse云数据仓库世界排名第二的成绩，是唯一入选的中国产品。DataWorks V2.0在DataWorks V1.0的基础上新增业务流程、组件的概念，力求通过支持双项目开发、隔离数据开发和生产环境、保证数据研发规范、减少错误代码来完善数据研发体系。



## 2 产品优势

---

本文为您介绍DataWorks的产品优势。

- 超大规模计算处理能力

DataWorks与底层计算平台集成，能够轻松处理海量数据：

- 万亿级数据JOIN，百万级并发Job，作业I/O可达PB级/天。
- 离线调度支持百万级任务量，实时监控告警。
- 提供功能强大易用的SQL、MR引擎，兼容大部分标准SQL语法。
- 采用三重备份、读写请求鉴权、应用沙箱、系统沙箱等多层次数据存储和访问安全机制保护您的数据，确保不丢失、不泄露、不被窃取。

- 一站式的数据工场

DataWorks为您提供可视化的操作界面，并支持多人协同作业：

- 提供数据从集成、加工、管理、监控、输出服务的全流程所有功能。
- 提供可视化工作流程设计器功能。
- 多人协同作业机制，分角色进行任务开发、线上调度、运维、数据权限管理等功能，数据及任务无需落地即可完成复杂的操作流程。

- 海量异构数据源快速集成能力

DataWorks支持400对异构数据源的离线同步，支持分钟、小时、天、周和月多种调度周期配置。

- Web化的软件服务

DataWorks可在互联网/内部网络环境下直接使用，无需安装部署，拎包入住，开箱即用。

- 多租户权限模型

多租户模型确保您的数据被安全隔离，以租户为单位进行统一的权限管控、数据管理、调度资源管理和成员管理工作。

- 智能的监控报警

通过设置监控基线，您不仅可以从宏观把控整体任务链路的完成时间，也可以从微观对每一个节点任务状态进行全方位监控。

- 易用的智能SQL编辑器

通过智能代码提示功能、表Meta信息提示功能、代码格式化和折叠功能、预编译功能、炫酷皮肤切换功能来获得全新的SQL代码编辑体验。



- 完备的数据质量监控体系

支持多种异构数据源、离线数据、实时数据的质量校验、通知、管理。

- 便捷的数据服务开发接口

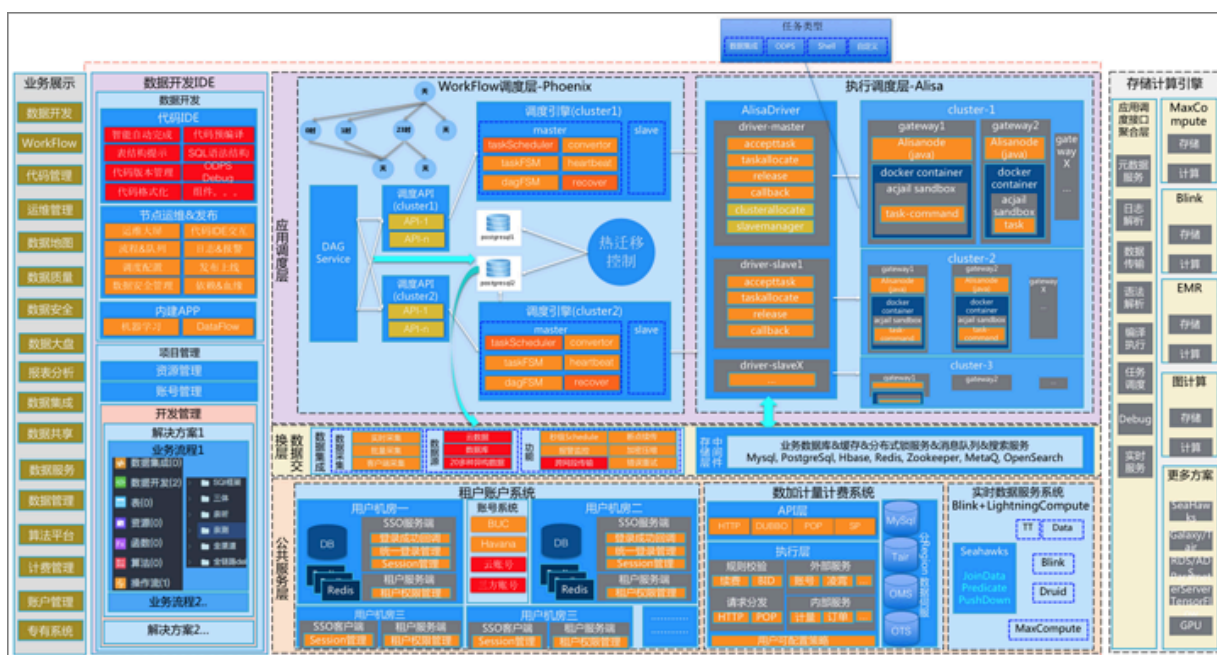
API网关服务、交互式数据服务引擎让您只需两步操作，即可将已有API和数据以服务的形式发布到数据共享与开放平台。

- 安全的数据共享机制

数据共享服务提供受保护空间，让详细数据以不可见、不落地的形式共享给其他租户，让数据真正安全地发挥大数据共享价值。

## 3 产品架构

DataWorks是阿里巴巴集团推出的大数据领域平台级产品，提供一站式大数据开发、管理、分析、挖掘、共享、交换等端到端的解决方案，利用MaxCompute（原ODPS）可处理海量数据，无需关心集群的搭建和运维。



由上图可以看出，DataWorks底层是基于MaxCompute的集成开发环境，包括数据开发、数据管理、数据分析、数据挖掘和管理控制台。

### 3.1 功能架构

本文为您介绍DataWorks的功能架构。



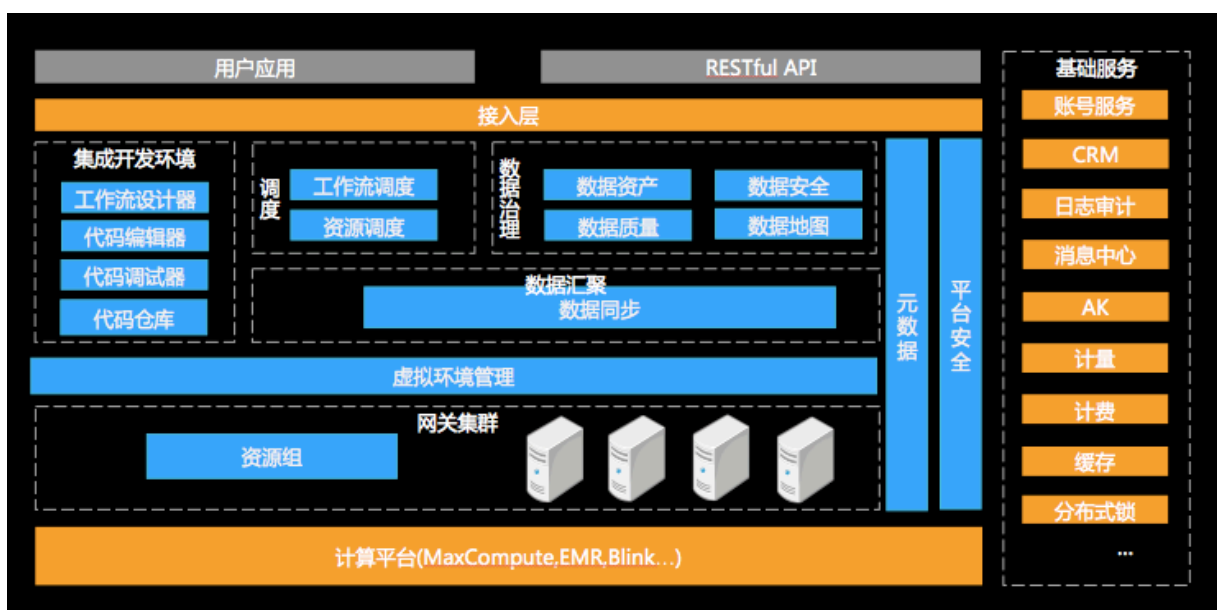
DataWorks提供以下八大主要功能模块：

- 数据集成：异构数据集成，将海量的数据从各种源系统汇集到大数据平台。
- 数据开发：数据仓库设计和ETL开发过程。
- 监控运维：ETL线上作业的运维监控，以及基于业务基线的大量任务实例监控。
- 数据分析：实时探查和分析数据。
- 数据资产管理：元数据管理、数据地图、数据血缘、数据资产大图等。
- 数据质量：数据质量探查、监控、校验和评分体系。
- 数据安全：数据权限管理，数据的分级打标、脱敏，以及数据审计。
- 数据服务：数据共享和数据交换，数据API服务。

## 3.2 系统架构

本文为您介绍DataWorks的系统架构。

DataWorks以数据为基础，以数据全链路加工流程为核心，提供数据汇聚、研发、治理、服务等多种功能，既能满足平台用户的数据需求，又能为上层应用提供各种行业解决方案。



## 3.3 安全架构

本文为您介绍DataWorks的安全架构。

DataWorks的安全架构，由平台自身的安全实现层、平台内置的安全服务层和租户可选的安全产品层构成：

- 平台自身的安全实现层：保障平台在代码实现和部署配置时的产品自身安全性。

- 平台内置的安全服务层：为租户和其用户提供平台基础性的安全服务能力，如租户资源隔离、身份认证、权限鉴别和日志合规审计等。
- 租户可选的安全产品层：为租户和其用户提供可选的、已集成的安全产品或工具，帮助租户根据其自行定义的安全策略对其拥有的系统、数据进行安全防护和运维管理。

## 3.4 多租户模型

DataWorks拥有自己的多租户权限模型。

- 弹性的存储和计算资源，租户可按需申请资源配额，独立管理自己的资源。
- 租户独立管理自有的数据、权限、用户、角色，彼此隔离，以确保数据安全。

## 4 功能特性

---

### 4.1 数据集成

本文为您介绍数据集成的概念和功能。

#### 概述

数据集成是阿里巴巴对外提供的稳定高效、弹性伸缩的数据同步平台。致力于提供复杂网络环境下、丰富的异构数据源之间数据高速稳定的数据移动及同步能力。

数据集成提供对业务方数据库进行抽取监控功能，能够对数据源头的的数据资源进行统一清点，并能够在复杂网络情况下对异构的数据源进行数据同步与集成，包括对关系型数据库、NoSQL数据库、大数据数据库、文本存储（FTP）等数据库类型支持，支持离线数据的批量、全量、增量同步，支持分钟、小时、天、周、月来自定义同步时间。

#### 支持多种数据通道

DataWorks数据集成支持多种数据通道：

- 元数据信息同步

元数据信息是整个平台数据的基础，数据集成系统可以从各个业务系统完成MySQL、SQLServer、Oracle、MaxCompute等20多种常见数据库的元数据信息的收集，避免对整体数据资产的情况不清楚，帮助数据管理者直接通过元数据进行后续资产的盘点及重点数据的同步。

- 关系型数据库同步服务

支持MySQL、SQLServer、Oracle、DRDS、PostgreSQL、DB2、RDS for PPAS等关系型数据库的读写操作。

- NoSQL数据库同步服务

支持HBase、MongoDB、Table Store等NoSQL数据库的读写操作。

- MPP数据库同步服务

支持HybridDB for MySQL、HybridDB for PostgreSQL等MPP数据库的读写操作。

- 大数据数据库同步服务

支持MaxCompute、HDFS的读写操作，并支持Analytic DB（ADS）的写操作。

- 非结构化存储同步服务

支持OSS、FTP的读写操作。

**说明：**

在数据集成中，数据源之间的数据传输是随机组合，目前已经支持400多对数据源的数据互传。

**支持数据流入管控**

支持各种数据类型的转换，精确识别脏数据，进行过滤、采集、展示，为您提供可靠的脏数据处理，让您准确把控数据流入内容，提供作业全链路的流量、数据量、脏数据探测和运行时汇报。

**传输速度快**

数据集成充分利用单机网卡能力，并使用分布式模型架构，保障数据吞吐量水平扩展，能够提供GB级、TB级的数据流量。

**控制友好**

数据集成提供精准流控保证，支持通道、记录流、字节流三种流控模式，并提供完备的容错处理，支持线程级别、进程级别、作业级别多层次局部或全局的重跑。

**支持同步插件**

支持以插件的方式部署采集工具至数据源端的服务器，完成数据信息的同步采集工作。

**支持跨网络传输**

支持各种复杂网络环境下的数据传输。例如，本地跨私网环境、VPC环境等。

**说明：**

对长链路传输通过协议加速能更高效、稳定的传输大批量数据。

## 4.2 数据开发

### 4.2.1 数据开发概述

数据开发为数据使用者提供一站式的集成开发环境，可满足数据资源平台下，数据开发者进行ETL开发、数据挖掘算法开发、数据主题库建设等需求。

当底层数据进行聚合后，数据仍然处于零散的状态，数据是无法直接为上层智能算法和DI应用提供对应数据的，此时需要对数据进行汇聚加工。数据管理和开发人员需要在数据资源平台建立对应的数据中心，进行对应数据的加工。

## 4.2.2 业务流程设计器

本文为您介绍业务流程设计器的组成。

### 概述

业务流程设计器可以实现将不同类型业务节点组织在一起的开发方式，让用户能以业务流程为中心组织数据开发逻辑。该功能通过各种类型开发节点的容器看板，将相关的工具和管理操作围绕数据看板中的对象来组织，使得开发的管理更加方便和智能化。

目前业务流程设计器支持ODPS\_SQL、ODPS\_MR、Shell、机器学习、数据同步、虚拟节点、PyODPS、SQL组件节点之间的组合使用，不仅每个业务流程内的节点可以相互依赖，不同业务流程之间的节点也能跨流程依赖，同时所有节点均能根据业务需要设置个性化调度时间以便适时运行。

### 节点任务

节点任务包含ODPS\_SQL、ODPS\_MR、Shell、机器学习、数据同步、虚拟节点、PyODPS、SQL组件任务，可以被本业务流程内的节点及其他流程内的节点依赖，并能够被调度系统调度。

### 节点属性配置

双击某节点会展开节点任务设计器，您可以对节点进行基本业务配置（如对SQL节点编写SQL、对数据集成节点配置数据传输规则），并可以在页面右侧查看或配置节点的调度属性、血缘关系、版本和结构。

### 历史版本

您可以查看任意版本的节点代码（仅限ODPS\_SQL、ODPS\_MR、Shell等节点类型），回滚节点的历史版本。

### 任务发布

任务发布提供了简单易用的发布功能，您在标准模式的项目内，可以发布已测试通过的业务流程至生产环境。

## 4.2.3 解决方案设计器

全面升级数据开发模式，让您可以通过项目>解决方案>业务流程的模式进行数据开发工作。

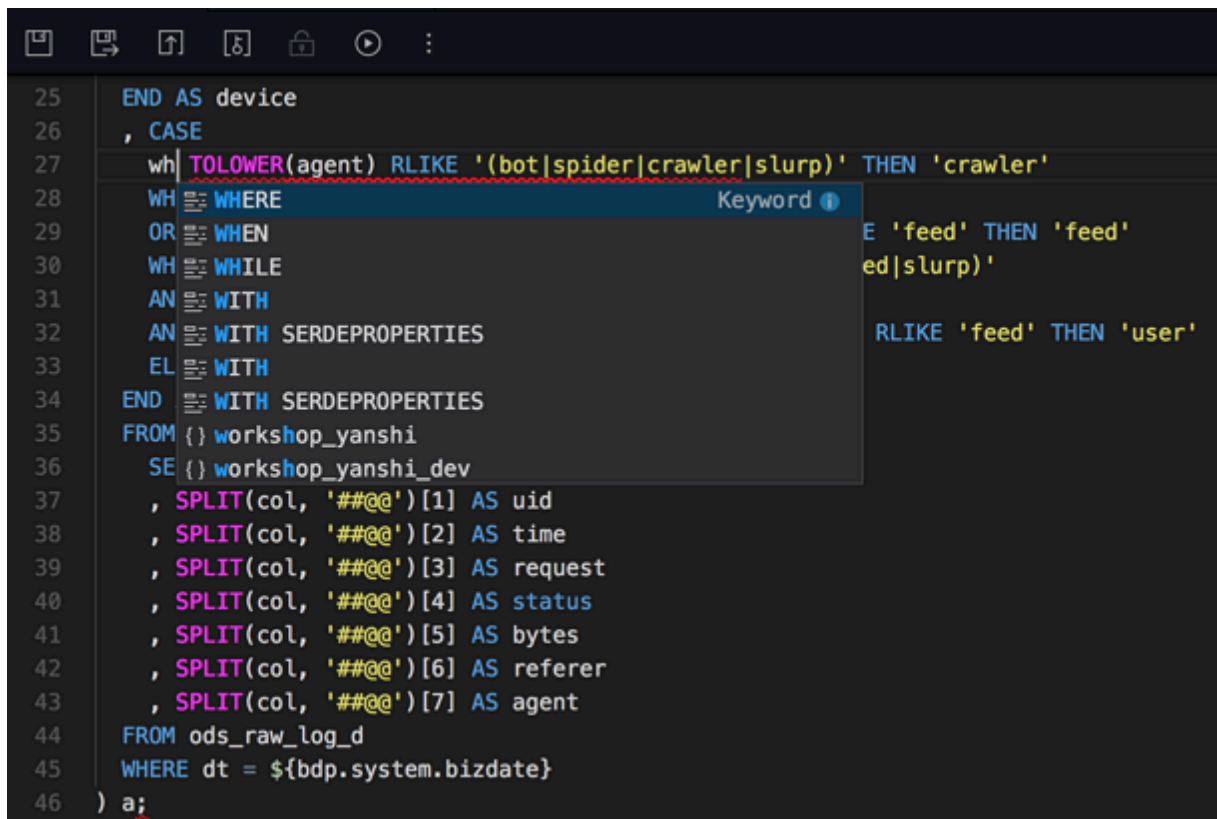
您可以通过解决方案设计器将不同类型的业务流程节点组合在一起，站在更高视角横跨多个业务流程做开发。同时，业务流程可以被多个解决方案复用，您只需要从解决方案视角来考量业务。

#### 4.2.4 代码开发编辑器

代码开发编辑器支持SQL编程、MR编程、Resource资源文件、注册UDF函数和Shell脚本编程。

## SQL编程

MaxCompute提供基于Web端的SQL编程，包括辅助编程（自动SQL提示、格式化、代码高亮等）、代码调试运行等编程功能。



## MR编程

MaxCompute支持将MR编译的JAR包以资源的形式上传并在ODPS MR节点中引用的形式来使用。

## Resource资源文件

支持上传如下类型的本地资源文件:

- JAR包：编辑好的Java Jar包，供UDF或ODPS\_MR程序调用。
- Python程序：供UDF程序调用。
- file文件：支持自定义参数的Shell脚本、xml配置文件、txt配置文件等资源文件。
- Archive类型：通过资源名称中的后缀识别压缩类型，支持的压缩文件类型包括.zip/.tgz/.tar.gz/.tar/jar。



## 注册UDF函数

UDF包含Java UDF和Python UDF两种，您需要先上传JAR包和Python程序后，再注册UDF函数，即可使用自定义函数进行数据开发。

## Shell脚本编程

提供在线的Shell脚本编程与调试环境。

## 4.2.5 代码管理与团队协作

代码管理与协作功能让数据开发可以多人同时进行编辑协作，提高开发效率。

代码管理提供工作流任务和代码的锁机制，保证同一个工作流任务或代码在同一时间内只能被一个用户编辑。您也可以通过获取锁的方式来得到编辑权限，并实时发送系统提示信息给对应用户。

同时数据开发也提供代码版本管理，您每一次提交的节点或工作流任务的版本都会被系统记录保存下来，您可以查看任意两个版本的对比。

## 4.3 监控运维

### 4.3.1 监控运维概述

监控运维为数据开发者和维护者提供一站式的数据运维管控能力，您可以自主管理作业的部署、作业优先级、以及生产监控运维。

DataWorks上数据量庞大、数据类型多样、数据业务复杂，数据处理任务也非常多，数据处理环节和流程周期长，需要支持高并发、多周期、支持多种数据处理环节的统一数据任务调度机制，按照策略进行数据任务调度。

DataWorks提供数据监控运维、任务运行情况监控、异常情况告警、日常运维数据统计等功能。

### 4.3.2 运维概览

运维概览主要用来展示调度任务的指标数据情况。

目前概览页面包含任务完成情况、任务运行情况、任务执行时长排行、调度任务数量趋势、近一月出错排行、任务类型分布和30天基线破线次数排行。

### 4.3.3 任务运维

可视化展示调度任务DAG图，极大地方便您对线上任务进行运维管理。

- 支持任务运行状态监控告警，支持单任务重跑、多任务重跑、Kill、置成功、暂停等操作。
- 支持两种模式选择：包括列表、DAG模式。
- 可以针对周期运行、测试运行、手动运行任务查看任务运行状态。

- 可以针对任务进行重跑、停止、查看运行日志、查看节点代码、查看节点属性。

### 4.3.4 智能监控

智能监控（Intelligent Monitor）是DataWorks任务运行的监控及分析系统。

根据监控规则和任务运行情况，智能监控决策是否报警、何时报警、如何报警以及给谁报警。智能监控会自动选择最合理的报警时间、报警方式以及报警对象。

智能监控旨在：

- 降低您的配置成本。
- 杜绝无效报警。
- 自动覆盖所有重要任务。

**智能监控**拥有一整套的监控报警逻辑，您只需要提供所关注业务的重要任务名称，即可监控整体任务的产出过程，并生成对应的标准统一的报警机制。智能监控还提供了轻量级的自助配置监控功能，让您可以根据自己的需求定义报警规则。

## 4.4 实时分析

实时数据分析主要提供临时查询和个人表两个核心功能，通过MaxCompute取数工具的准实时模式来加快分析速度。

### 原理介绍

service mode准实时模式与标准模式的区别

准实时模式的开关（系统默认打开）：`set ODPS.service.mode=[all|off|limited]`。

两个模式的区别，如下所示：

- 准实时模式采用预分配进程池的方式，不通过Fuxi Job调度，省去了Fuxi Job分配的几秒延迟。
- 准实时模式Map或Reduce的shuffle数据不再通过盘古来中转，直接通过网络传输。

### 技术要点

- mode=all会开启准实时模式，但是如果当前SQL需要的MaxCompute资源不足（例如当前空闲的worker数不足以新建需要的instance），会转而执行Fuxi Job。
- 准实时仍然有复杂的调度过程，但在fuxi job的基础上做了许多耗时方面的优化。
- 当设置了service mode，会优先尝试使用service mode方式运行。如果系统资源不够，或者service mode运行发生错误，会fall down到Fuxi Job方式运行。如果service mode发生未知异常，会以Fuxi Job方式再次进行尝试。

## 4.5 数据服务

DataWorks数据服务旨在为企业搭建统一的数据服务总线，帮助企业统一管理对内对外的API服务。

数据服务提供快速将数据表生成数据API的能力，同时支持您将现有的API快速注册到数据服务平台以统一管理和发布。同时，数据服务已与API网关（API Gateway）打通，支持将API服务一键发布至API网关。数据服务与API网关为您提供了安全稳定、低成本、易上手的数据共享与开放服务。数据服务采用Serverless架构，您只需关注API本身的查询逻辑，无需关心运行环境等基础设施，数据服务会为您准备好计算资源，并支持弹性扩展，零运维成本。

数据服务是数据交换的核心组件。数据交换包含数据共享服务和数据开放服务。数据服务为数据交换构建了安全、灵活、可靠的服务总线，可以支撑政务系统内部实现跨部门、跨层级、跨网络的数据共享，也可以支撑政务数据对社会公众开放的政务数据开放服务。

### API生成

数据服务支持将关系型数据库、NoSQL数据库（例如OTS）、分析型数据库（例如AnalyticDB）的表通过可视化配置的向导模式快速生成数据API，您无需具备编码能力，即可在几分钟之内生成好一个数据API，并且立即可以调用。同时为了满足高阶用户的个性化查询需求，数据服务也提供了自定义SQL的脚本模式，允许您自行编写API的查询SQL，并支持多表关联、复杂查询条件以及聚合函数等能力。

### API注册

数据服务也支持将您手中现成的API服务注册上来，与通过数据表生成的API统一管理。目前支持Restful风格的API注册，包含GET、POST、PUT和DELETE四类常见请求方式，支持表单、JSON和XML三种数据格式。

### API网关

API网关（API Gateway），提供API托管服务，涵盖API发布、管理、运维、售卖的全生命周期管理。辅助用户简单、快速、低成本、低风险的实现微服务聚合、前后端分离、系统集成，向合作伙伴、开发者开放功能和数据。数据服务已与API网关产品一键打通，在数据服务中配置生成以及注册的API都可以一键发布到API网关，并通过API网关来管理API的授权鉴权、流量控制、计量等服务。

## 4.6 平台管理

平台管理主要从系统层面，为管理者对参与数据资源平台使用的用户进行对应管控。

项目空间作为代码管理、成员管理、角色和权限分配的基本单元，每个团队都可具有独立的项目空间。您加入项目空间并被分配相关权限之后，才可查看或编辑代码。

**说明：**

一个用户可以同时加入多个项目空间，在不同的项目空间中被授予不同的角色。

平台管理包括组织管理、项目管理、成员管理和权限管理。

**组织管理**

显示组织详情信息以及组织Owner账号、AccessKey和AccessSecret信息，并可以对组织对应人员进行成员管控。

**项目管理**

DataWorks管理员可对您的项目空间进行列表展现，并提供创建、配置、激活、禁用项目空间的对应管理功能，方便数据资源层管理员对项目空间进行整体管控。

**成员管理**

**成员管理**页面以列表的形式显示本工作空间的成员名称、登录名称、成员角色等信息：

- 支持模糊搜索工作空间成员，并可以从本工作空间移除成员。
- 用户加入工作空间，仅项目管理员可以新增工作空间成员，支持模糊匹配查找的方式将用户添加至本工作空间。

**说明：**

在添加用户到工作空间时，必须为其指定至少一种角色。

项目管理员可以主动清退用户。

**说明：**

用户移出本工作空间后，会失去之前在本工作空间分配的所有权限。

**权限管理**

权限管理主要完成平台用户、角色、权限等管理由统一的管理提供。

下表为各角色对应的权限特征。

数据资源平台角色	平台权限特征
数据项目管理员	指数据项目空间的管理者，可对该项目空间的基本属性、数据源、当前项目空间计算引擎配置和项目成员等进行管理，并为项目成员赋予项目管理员、开发、运维、部署、访客角色。
数据开发	开发角色的用户能够创建工作流、脚本文件、资源和 UDF，新建或删除表，同时可以创建发布包，但不能执行发布操作。

数据资源平台角色	平台权限特征
数据运维	运维角色的用户由项目管理员分配运维权限；拥有发布及线上运维的操作权限，没有数据开发的操作权限。
数据部署	部署角色与运维角色相似，但是它没有线上运维的操作权限。
数据访客	访客角色的用户只具备查看权限，没有权限进行编辑工作流和代码等操作。

## 4.7 数据资产管理

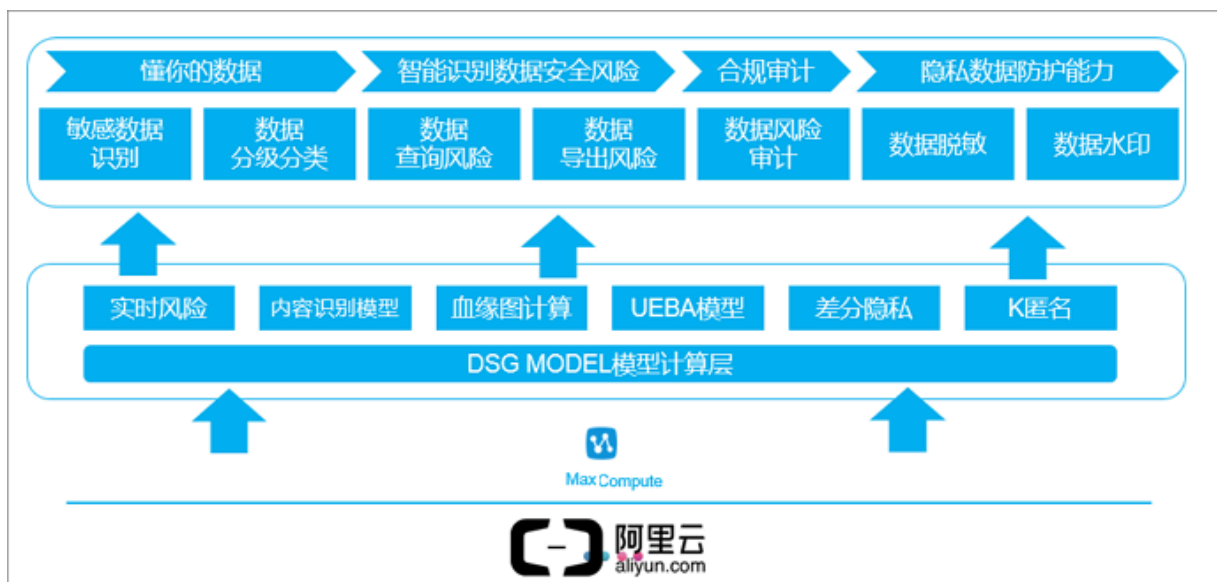
数据资产管理有独立的权限点控制，由于数据资产管理属于组织级别的模块，需要在组织管理中添加权限。

业务系统及DataWorks中有大量的数据表、API等各类数据资产，数据管理者通过数据集成工具同步数据、通过数据开发加工数据后，需要对整个平台数据进行统一管控，了解平台的核心数据资产，提供对应数据资产管理规范。

## 4.8 数据保护伞

### 4.8.1 数据保护伞概述

数据保护伞平台是一款数据安全产品，提供数据资产识别、敏感数据发现、数据分类分级、脱敏、访问监控、风险发现预警与审计能力。



数据保护伞平台底层是基于MaxCompute（原ODPS）的集成安全管理服务。

数据保护伞的特性如下：

- 智能地发现敏感数据

基于自学习的模型算法，自动识别企业拥有的敏感数据，并以直观的形式展示具体类型、分布、数量等信息。同时支持自定义类型的数据识别。

- 精准的分级分类

支持自定义分级信息功能，满足不同企业对数据等级管理需要。

- 灵活的数据脱敏

提供丰富多样、可配置的数据脱敏方式，实现使用环节的动态脱敏。

- 用户异常操作风险监控和审计

利用多维度关联分析及算法，主动发现异常风险操作，提供预警以及可视化一站式审计。

## 4.8.2 基本术语

本文为您介绍数据保护伞中的组织、项目空间、数据脱敏等基本术语。

### 组织 (Organization)

组织是指使用系统或计算资源的客户，广义上也包括了用户在云计算资源中构建的所有数据，包括账号、权限以及客制化的应用程序等，都属于组织的范畴。

### 项目空间 (Project)

项目空间是阿里云大数据开发平台最基本的组织对象，类似于传统数据库的DataBase。阿里云大数据开发平台的项目空间，是进行多组织隔离和访问控制的主要边界，也是您管理表 (Table)、资源 (Resource)、自定义函数 (UDF)、节点 (Node)、权限等的基本单元。

### 正则表达式 (Regular Expression)

正则表达式是对字符串操作的一种逻辑公式，就是用事先定义好的一些特定字符、及这些特定字符的组合，组成一个规则字符串。该规则字符串用于表达对字符串的一种过滤逻辑。您可以使用正则表达式规则实现对特殊数据的识别。



#### 说明：

字符串包括普通字符（例如a~z之间的字母）和特殊字符（称为元字符）。

### 数据分级分类 (Data Classification)

根据数据价值、敏感性、数据风险及法律法规要求，结合泄露后带来的影响，对数据进行分级分类。

### 数据扫描 (Data Scansion)

通过您对某类数据的识别规则定义（正则表达式、字段名），在用户端进行数据扫描。

## 数据脱敏 (Data Desensitization)

数据脱敏是指对某些敏感信息通过脱敏规则进行数据的变形，实现敏感隐私数据的可靠保护。

### MaxCompute (原ODPS)

MaxCompute是阿里巴巴自主研发的海量数据离线数据处理平台，主要服务于实时性要求相对不高的批量结构化数据的存储和计算，可以提供海量数据仓库的解决方案以及针对大数据的分析建模服务。

## 4.8.3 规则配置

数据安全管理员可以在配置监控规则，定义敏感数据。

如果已定义好敏感数据，请直接跳转至[识别数据分布](#)、[数据访问行为](#)以及[数据导出](#)模块进行相关操作。

## 4.8.4 数据发现

数据安全管理员在完成敏感数据规则配置T+1之后，即可在识别数据分布中查看数据分布情况，分整体、按等级分布以及字段明细。

根据您的查看需求，提供按project、规则名、规则类型、风险等级（即分级）进行过滤选择。

## 4.8.5 数据访问

本文为您介绍访问和导出数据。

数据访问包括访问行为和导出行为：

- 访问行为包括create、insert和select操作，但不包括访问失败的行为。
- 导出行为是指数据从MaxCompute导出的行为。

### 访问行为

数据安全管理员在完成敏感数据规则配置T+1之后，即可在数据访问行为中查看数据使用情况，包括整体、按等级分布以及字段明细。根据您的查看需求，提供按project、规则名、规则类型、风险等级（即分级）、访问人员进行过滤选择。

### 导出行为

数据安全管理员在完成敏感数据规则配置T+1之后，即可在数据导出中查看用户从MaxCompute中把数据导出至外部的情况，包括整体、TOP导出用户以及导出明细。根据您的查看需求，提供按规则名、规则类型、导出大于选项进行过滤选择。

### 4.8.6 数据脱敏管理

数据脱敏配置页面为您提供新建、修改、删除和测试脱敏规则的功能。

您可以对每一条数据识别规则，自定义配置相应的脱敏方式，并可以对相应的数据脱敏规则配置不需脱敏的白名单。

### 4.8.7 分级信息管理

如果规则配置中的分级选择无法满足需求时，您可以在分级页面管理中进行设置。

您可以在**分级信息管理**页面新建、删除分级，并可以调整分级优先级和规则优先级。

### 4.8.8 手动修正数据

在规则识别的敏感数据不准确的情况下，您可以在修正数据页面手动修正，包括删除识别错误数据、更改识别数据类型以及批量处理。

### 4.8.9 数据风险

数据风险页面为您提供通过手工进行风险数据识别、风险识别管理（风险规则配置识别、AI识别）产生的风险数据清单，同事可以对风险数据进行审计备注。

### 4.8.10 风险识别管理

风险识别管理页面为您提供风险数据规则配置。

您可以通过配置风险数据的规则，识别日常访问中的风险以及启动AI识别自动识别数据风险。识别后的风险数据统一在数据风险页面进行展示和审计操作，同时也会在数据访问页面的相应数据后打上识别标志。

### 4.8.11 数据审计

数据审计页面多维度展示您的风险处理结果和风险分布情况。

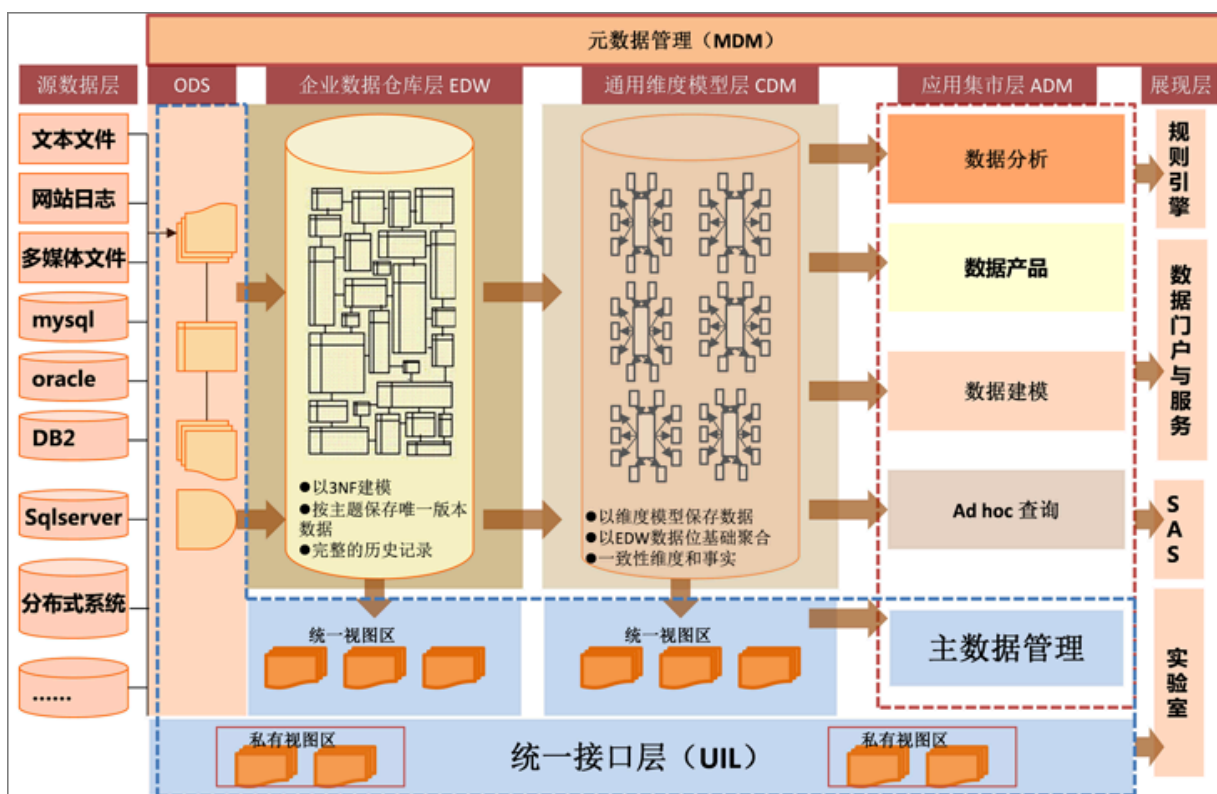
您可以在**数据审计**页面查看**风险总量**、**已处理风险数**、**未处理风险数**、**风险量趋势**和**风险事件维度分析**。



## 5 应用场景

### 5.1 云上数仓

大型企业可以在专有云环境下使用DataWorks来构建超大型的数据仓库。

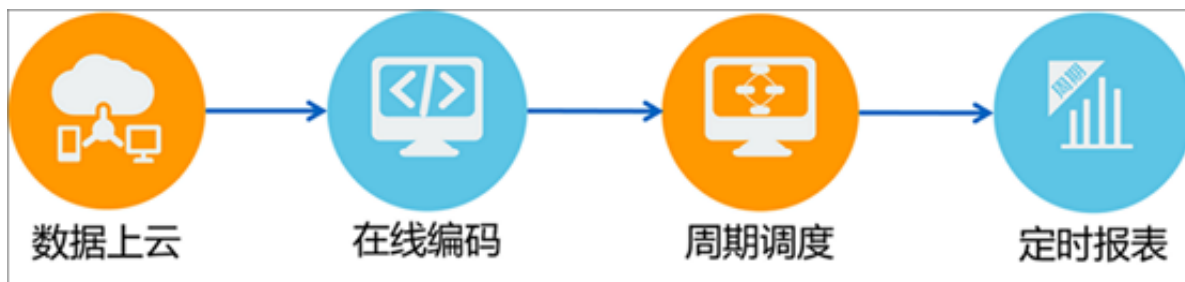


DataWorks提供卓越的海量数据集成能力：

- 海量存储：可支持PB、EB级别的数据仓库，存储规模可线性扩展。
- 数据集成：支持多种异构数据源的数据同步和整合，消除数据孤岛。
- 数据开发：基于MaxCompute（原ODPS）的大数据开发，支持SQL、MR等编程框架，以及贴近业务场景的白屏化 workflow 设计器。
- 数据管理：基于统一的元数据服务来提供数据资源管理视图，以及数据权限审批流程。
- 离线调度：可以提供多时间维度的周期性调度能力，支持每天百万级的调度并发，并对任务调度实时监控，对错误及时告警。

## 5.2 BI应用

本节为您介绍如何使用DataWorks制作报表。



基于网络日志，完成如下分析需求：

- 统计网站的PV（浏览次数）、UV（独立访客），按用户的终端类型（如Android、iPad、iPhone、PC等）分别统计，并给出这一天的统计报表。
- 网站的访问来源，了解网站的流量从哪里来。

截取一条真实的数据如下所示。

```
xx.xxx.xx.xxx - - [12/Feb/2014:03:15:52 +0800] "GET /articles/4914.html HTTP/1.1" 200 37666  
"http://xxx.cn/articles/6043.html" "Mozilla/5.0 (Windows NT 6.2; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/xx.x.xxxx.xxx Safari/537.36" -
```

1. 新建表：在导入数据之前，需要先创建一张MaxCompute目标表，把表名命名为ods\_log\_tracker。

**方案一：【业务流程】-【右键-表】-【新建表】**

字段英文名	字段中文名	字段类型	描述	主键	操作
ip		string	client ip address	是	
user		string		是	
time		datetime		是	
method		string	HTTP request type, such as GET POST...	是	

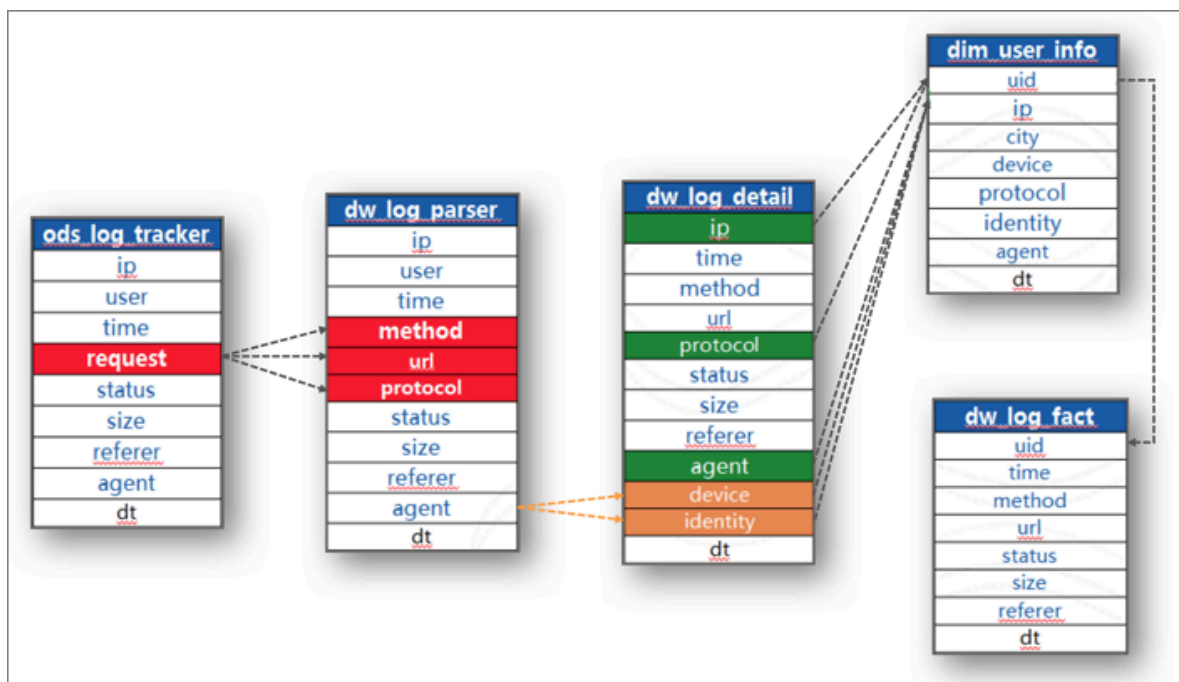
**方案二：【业务流程】->【右键-数据开发】->【ODPS SQL】**

```
--odps sql  
--author: dataworks_demo2  
--create time: 2018-08-02 15:57:55  
CREATE TABLE IF NOT EXISTS dw_log_parser(  
  ip STRING COMMENT 'client ip address',  
  user STRING,  
  time DATETIME,  
  method STRING COMMENT 'HTTP request type, such as GET POST...',  
  url STRING,  
  protocol STRING,  
  status BIGINT COMMENT 'HTTP response code from server',  
  size BIGINT,  
  referer STRING,  
  agent STRING)  
PARTITIONED BY (dt STRING);
```

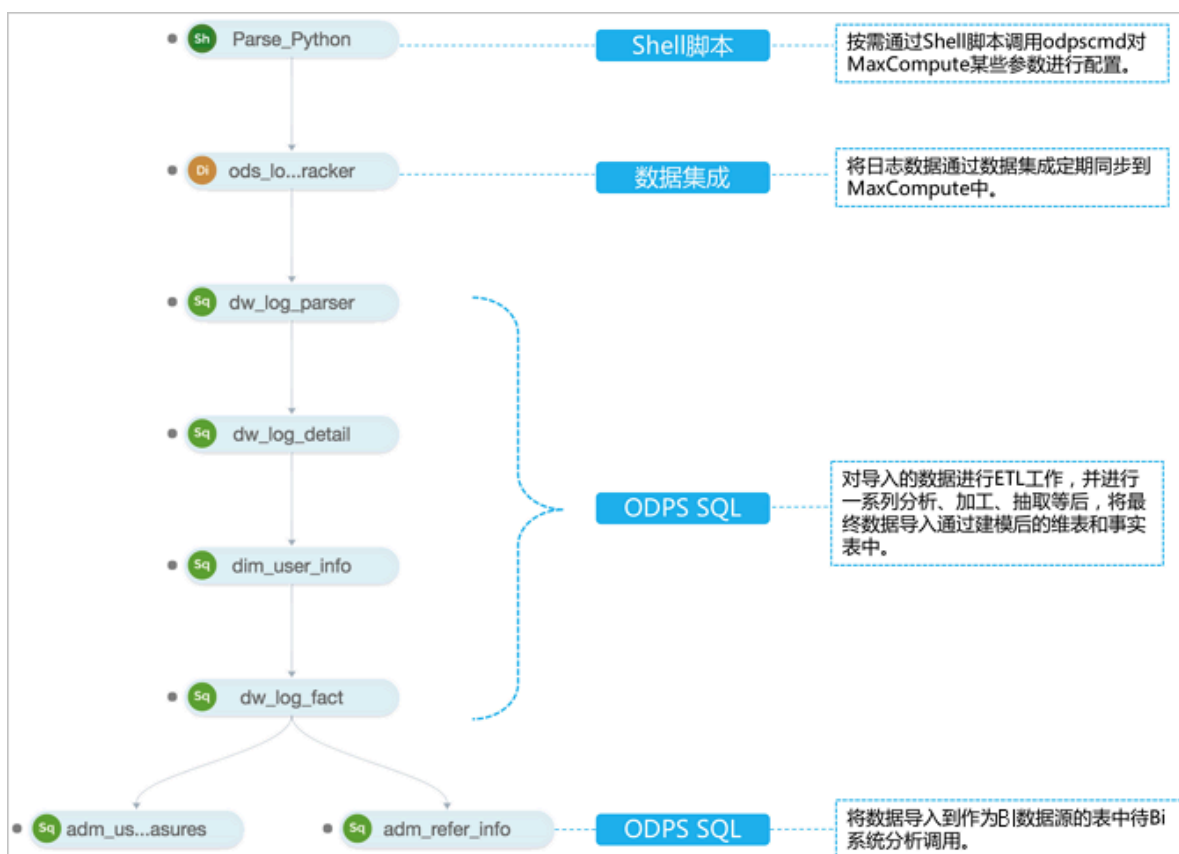
**ods\_log\_tracker**

ip
user
time
request
status
size
referer
agent
dt

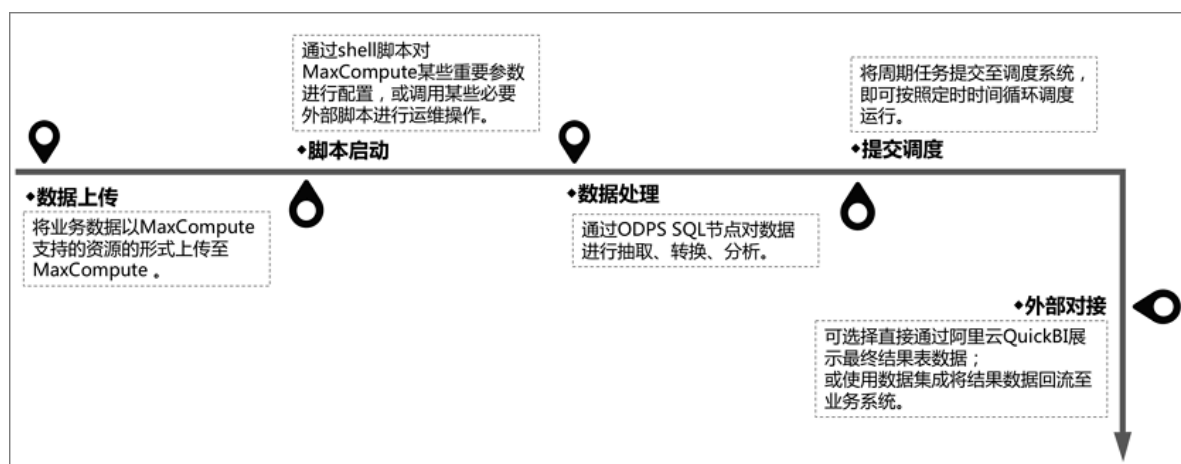
## 2. 设置各表之间的依赖关系。



## 3. 分别新建Shell、数据同步和ODPS SQL节点任务。



#### 4. 进行任务开发。



### 5.3 数据化运营

- 创新业务：通过数据挖掘建模和实时决策系统，将大数据加工结果直接应用于业务系统。
- 中小企业：基于DataWorks可快速使用和分析数据，助力企业的经营决策。

图 5-1: 阿里小贷的数据业务运营模式图



## 6 使用限制

---

无

## 7 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。

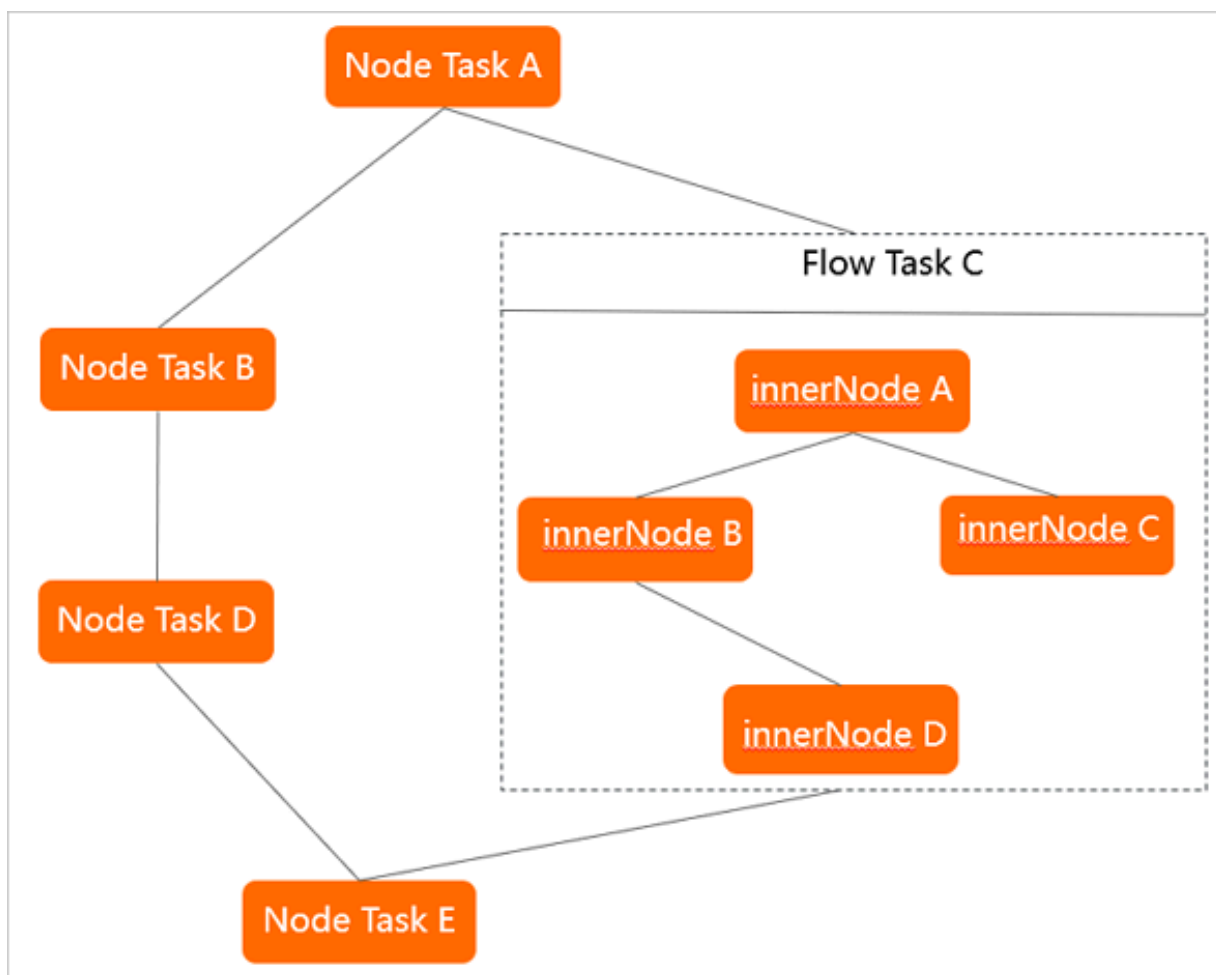
## 任务 (Task)

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。

任务类型	描述
工作流任务（Flow Task）	<p>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。</p> <p>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。</p> <div>  <b>说明：</b>            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。         </div>
内部节点（innerNode）	<p>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。</p>

### 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。



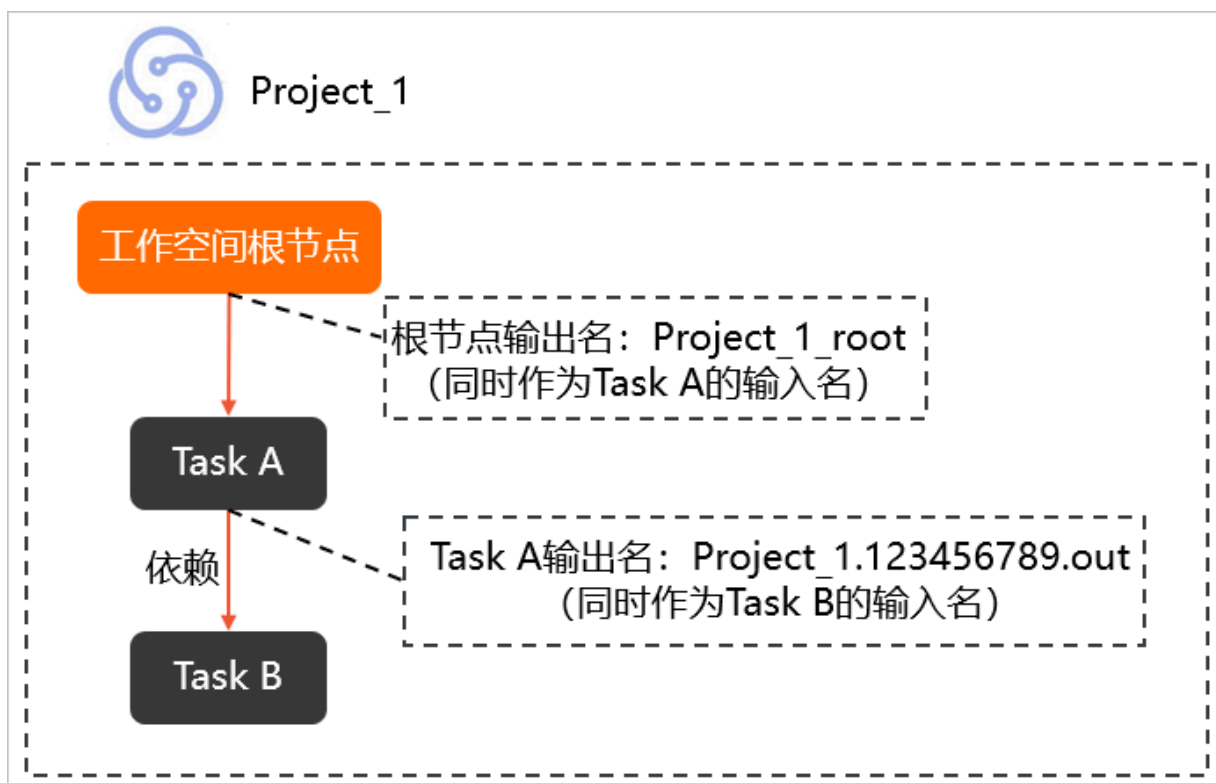
## 资源、函数

资源、函数均为MaxCompute的概念，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。