

ALIBABA CLOUD

阿里云

DataWorks
产品简介

文档版本：20220712

 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 <code>cd /d C:/window</code> 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	<code>bae log list --instanceid</code> <code>Instance_ID</code>
[] 或者 [a b]	表示可选项，至多选择一个。	<code>ipconfig [-all -t]</code>
{ } 或者 {a b}	表示必选项，至多选择一个。	<code>switch {active stand}</code>

目录

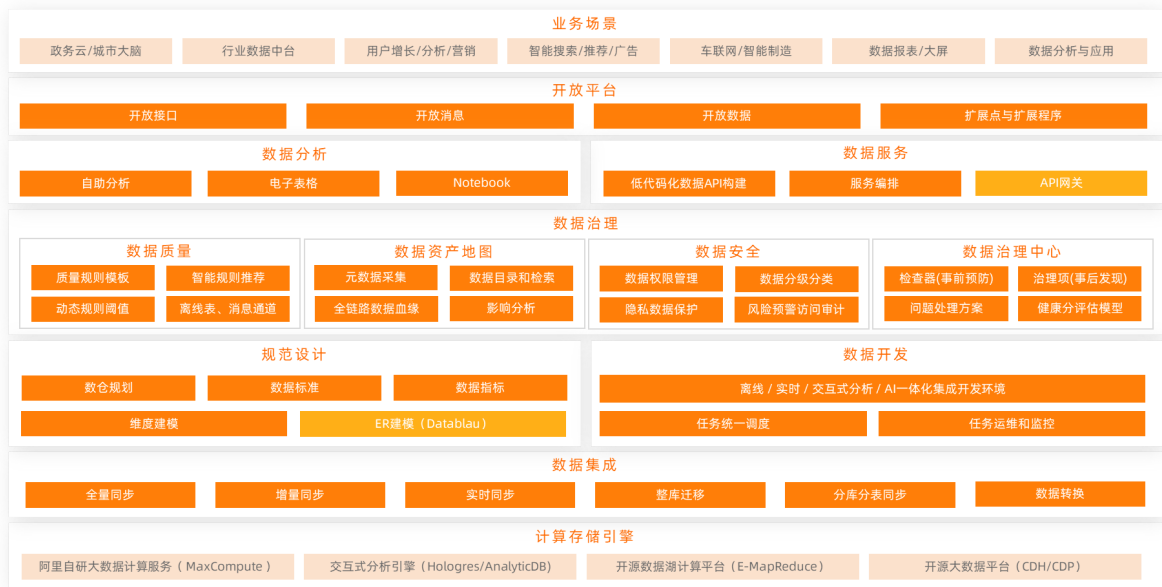
1.什么是DataWorks	05
2.功能特性	08
2.1. 数据集成：全领域数据汇聚	08
2.2. 数据开发与运维中心：数据加工	09
2.3. 数据建模：智能数据建模	11
2.4. 数据分析：即时快速分析	14
2.5. 数据质量：全流程的质量监控	15
2.6. 数据地图：统一管理，跟踪血缘	15
2.7. 数据服务：低成本快速发布API	15
2.8. 开放平台：能力全面开放	16
2.9. 迁移助手与迁云服务	17
2.10. 简单模式和标准模式的区别	18
3.产品优势	24
4.应用场景	25
4.1. 受众与核心能力	25
4.2. 构建数据仓库	25
4.3. 通用数据开发	27
5.客户案例	29
6.相关的云服务	35
7.基本概念	36

1.什么是DataWorks

DataWorks基于MaxCompute、Hologres、EMR、AnalyticDB、CDP等大数据引擎，为数据仓库、数据湖、湖仓一体等解决方案提供统一的全链路大数据开发治理平台。从2009年起，DataWorks不断沉淀阿里巴巴大数据建设方法论，支撑数据中台建设，同时与数万名政务、央国企、金融、零售、互联网、能源、制造、工业等行业的客户携手，不断提升数据应用效率，助力产业数字化升级。

产品架构

DataWorks十多年沉淀数百项核心能力，通过**智能数据建模**、**全域数据集成**、**高效数据生产**、**主动数据治理**、**全面数据安全**、**数据分析服务**六大全链路数据治理的能力，帮助企业治理内部不断上涨的“数据悬河”，释放企业的数据生产力。

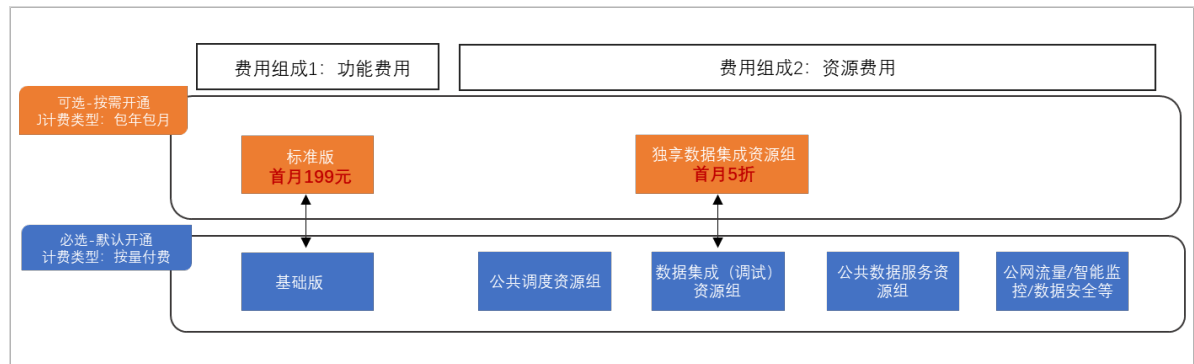


产品开通

说明 当前DataWorks仅支持在PC端Chrome浏览器69以上版本和新版Edge浏览器上使用。

新用户推荐参考如下配置开通使用产品，更多购买配置推荐请参见[购买指引](#)。

● 推荐配置



● 推荐原因

- 功能：推荐购买专业版，专业版满足企业专业的数据仓库构建需求，覆盖数据开发、任务运维、数据地图、数据质量等绝大部分功能。
- 资源：推荐购买独享数据集成资源组，以支持更好的同步体验，支持离线同步、实时同步、全增量同步等多种同步解决方案。

典型客户

- 国家电网大数据中心**：通过DataWorks实现总部+27家省（市）公司PB级数据的统一管理，通过全链路数据中台的治理与监测运营体系，加快电网整体数字化转型升级。
- 世界500强亿滋中国**：通过DataWorks智能数据建模进行全链路的数据模型治理，极大提升数据中台的自服务能力，让企业数据决策实现下放，释放新零售的数字化力量。
- 上市公司创梦天地**：基于开源的EMR引擎，用DataWorks替换自研调度系统，企业内部的技术人员可以更加专注业务，助力游戏行业的数据化运营。

更多客户案例请参见[客户案例](#)章节。

DataWorks发展之路

阿里巴巴集团内发展历程

从2009年产品立项开始，DataWorks与阿里巴巴业务共同发展，结合MaxCompute、Hologres等大数据计算引擎的能力，跨越多个技术阶段，支撑阿里巴巴数据中台与数据治理建设。目前阿里巴巴集团内DataWorks每天活跃用户数超过5万人，平均每3个人就有1个人使用DataWorks，支持300多个数据应用，**服务100多个阿里巴巴集团事业部**。

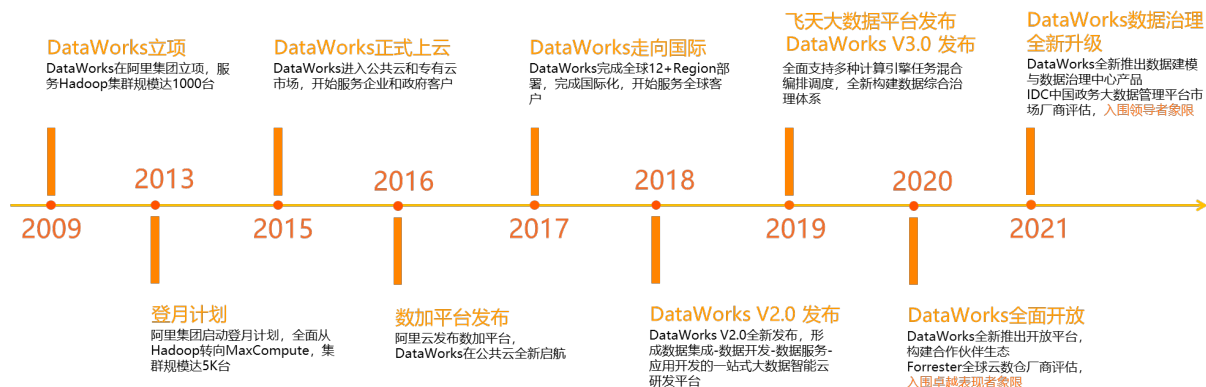
阿里巴巴数据平台发展的四个阶段

业务与数据平台十二年共同发展



阿里云上发展历程

2015年DataWorks正式上云，将多年沉淀的大数据建设方法论产品化输出，服务阿里云上客户，通过不断迭代的产品能力，DataWorks正在与各行各业的客户与合作伙伴一起，通过全链路数据治理，管得好数据、用得着数据，让数据从低质低效向高质高效流动。



获奖经历

- IDC：大数据平台公共云市场份额中国第一
- IDC：政务大数据产品能力、市场份额均位居中国第一
- Forrester：全球云数据仓库卓越表现者象限，国内唯一
- 中国信通院：首个通过577项技术要求的数据平台整体解决方案评测
- 中国电子学会科技进步特等奖
- 中国国际软件博览会金奖
- 浙江省科技进步一等奖
- 中国国际大数据产业博览会十佳大数据案例
- 中国信通院：数据集成工具、数据管理工具、数据开发平台、数据脱敏工具、数据分类分级等评测

学习路径

您可以通过DataWorks文档首页的学习路径，快速了解DataWorks的相关概念、基础操作及进阶操作等内容。



产品支持

您可以通过登录并进入[阿里云工单系统](#)提交售后问题，或[点击这里](#)加入DataWorks钉钉交流群进行售前售后咨询，咨询可直接@智能机器人，值班时间段内也可联系值班人员。

2. 功能特性

2.1. 数据集成：全领域数据汇聚

DataWorks的数据集成功能模块是稳定高效、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高速稳定的数据移动及同步能力。

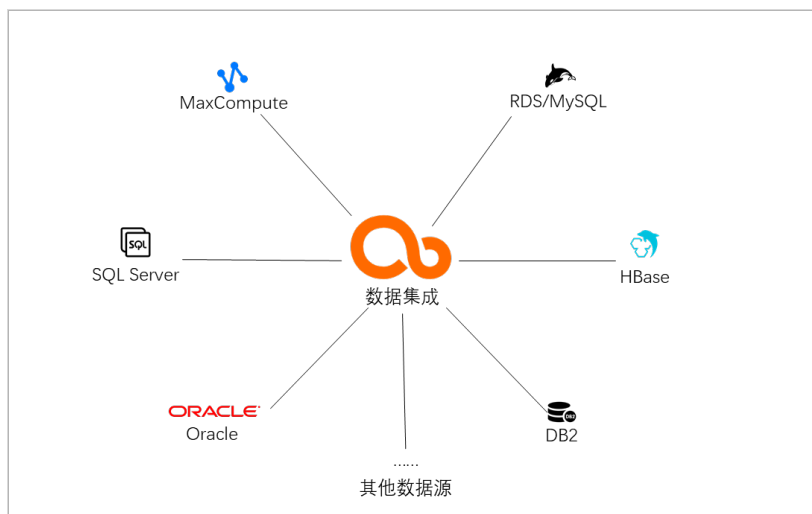
功能概述

DataWorks数据集成支持离线同步、实时同步，以及离线和实时一体化的全增量同步。其中：

- 离线同步场景下，支持设置离线同步任务的调度周期。
- 支持数据库、数仓、NoSQL数据库、文件存储、消息队列等近50多种不同异构数据源之间的数据同步。
- 支持在各类复杂网络环境下，连通数据源的网络解决方案，无论数据源在公网、IDC还是VPC内，均可使用DataWorks数据集成实现网络连通。
- 支持安全控制与运维监控，保障数据同步的安全、可控。

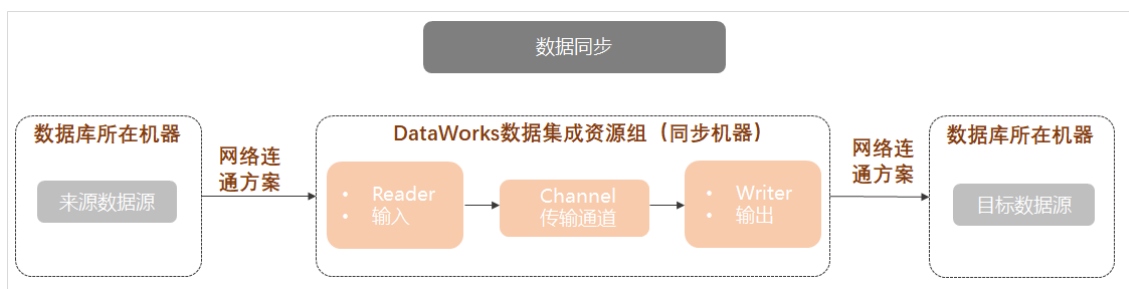
核心技术与架构

- 引擎架构



采用星形引擎架构，数据源接入数据集成后，即可与其他各类型数据源组成同步链路进行数据同步。当前支持的数据源请参见[支持的数据源与读写插件](#)、[实时同步支持的数据源](#)。

- 数据集成资源组与网络连通



如上图所示，进行数据同步前，需通过合适的网络连通方案将数据库和数据集成资源组间的网络打通。DataWorks当前支持使用独享数据集成资源组或自定义数据集成资源组，不同资源组的属性与适用场景不同，网络连通方案详情请参见[配置资源组与网络连通](#)。

适用场景

DataWorks数据集成适用于数据入湖入仓、分库分表、实时数据归档、云间数据流转等数据传输场景。

计费

运行数据集成任务可能产生的费用由以下几部分组成：

- 运行数据集成任务所需的数据集成资源组费用。

计费详情请参见[独享数据集成资源组计费说明：包年包月](#)、[公共数据集成（调试）资源组：按量计费](#)。

- 运行数据集成任务所需的调度资源组费用。

计费详情请参见[独享调度资源组计费说明：包年包月](#)、[公共调度资源组计费说明：按量付费](#)。

- （可选）运行数据集成任务所需的公网流量费用。

如果数据集成任务通过公网进行数据传输，会涉及公网流量费用，计费详情请参见[公网流量计费说明](#)。

 **说明** 运行数据集成任务时，可能会产生由数据同步任务相关配置引起的账单和费用，如同步上下游数据库、计算引擎系统计算和存储费用，所需网络服务费用（如高速通道、共享带宽、EIP）等等，此类收费不属于DataWorks相关收费范畴，账单也不会体现在DataWorks产品下。

开通使用

开通DataWorks版本后，您即可根据需要购买数据集成资源组，选择合适的同步方案开发数据集成任务。开通使用的操作详情请参见[数据集成](#)章节。

2.2. 数据开发与运维中心：数据加工

DataWorks的数据开发（DataStudio）是数据加工的开发平台，运维中心是智能运维平台，基于这两个功能模块，您可以在DataWorks上规范、高效地构建和运维数据开发 workflow。

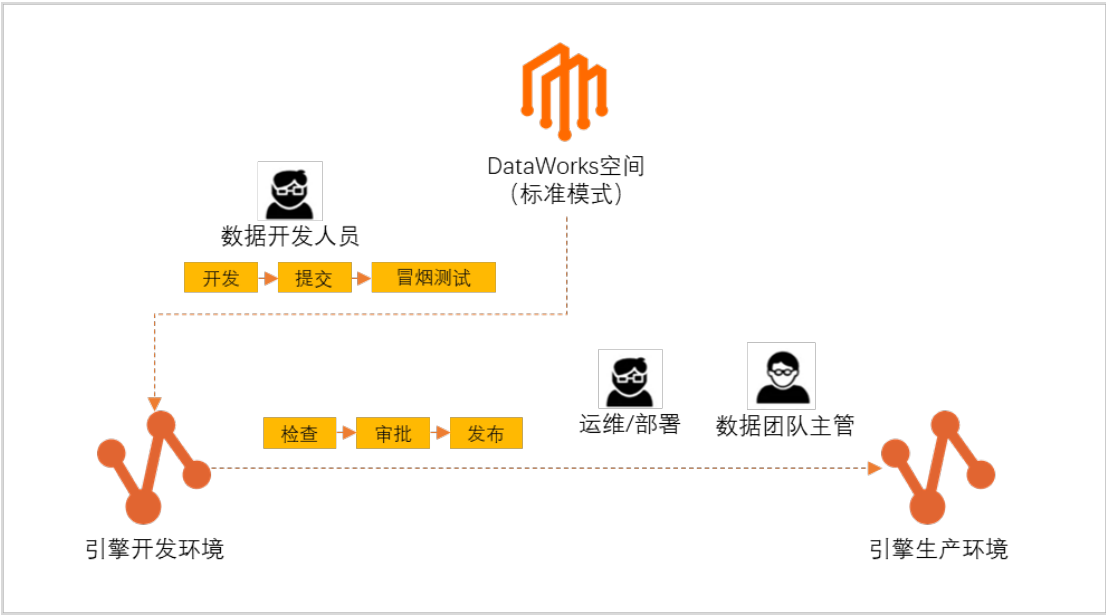
功能概述

DataWorks的数据开发的亮点功能如下。

- DataStudio支持MaxCompute、EMR、CDH、Hologres、AnalyticDB、Clickhouse等多种计算引擎，支持在统一的平台上进行各类引擎任务的开发、测试、发布和运维等操作。
- DataStudio支持智能编辑器、可视化依赖编排，调度能力经过阿里集团内调度任务、复杂业务依赖的反复验证。
- DataStudio提供隔离的开发和生产环境，结合版本管理、代码评审、冒烟测试、发布管控、操作审计等配套功能，帮助企业规范地完成数据开发。
- 运维中心支持数据时效性保障、任务诊断、影响分析、自动运维、移动运维等功能。

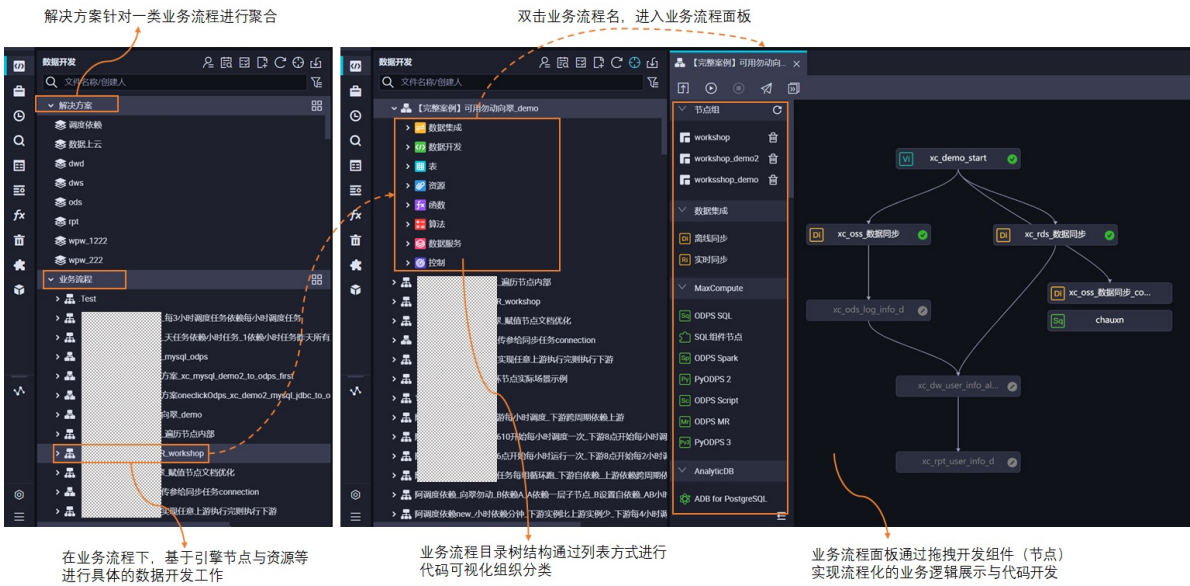
核心技术与架构

- 高效、规范的开发流程



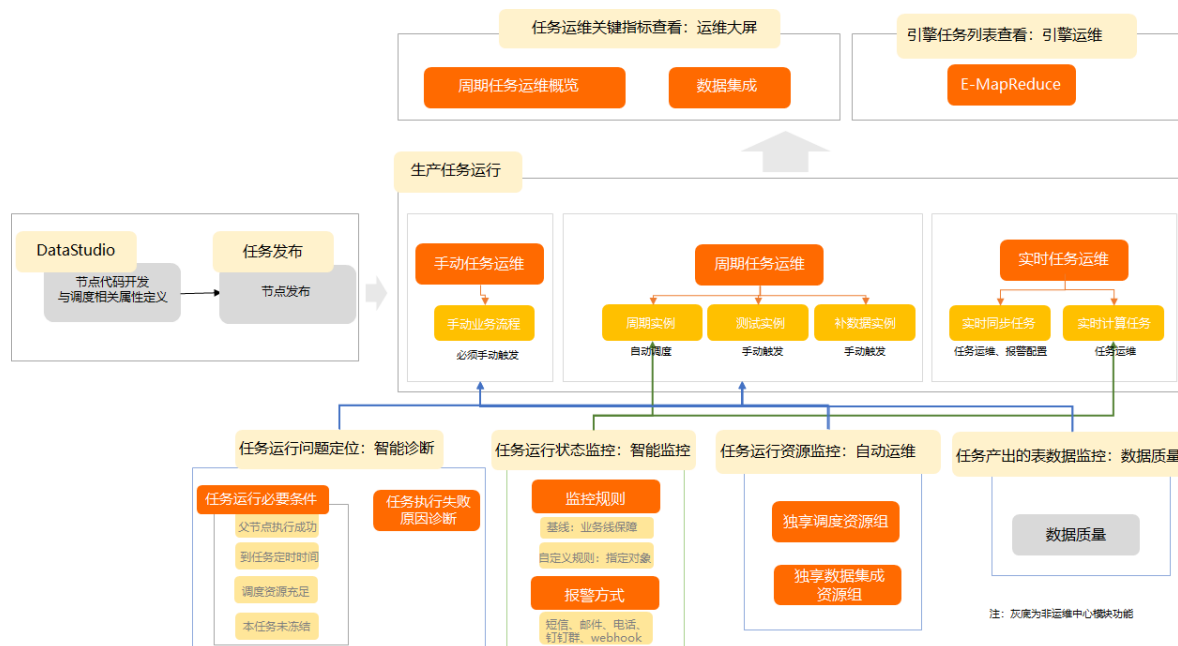
说明 DataWorks提供标准模式的工作空间，可隔离开发环境与生产环境，标准模式的详细介绍可参见[简单模式和标准模式的区别](#)。

● 可视化的开发界面



支持通过拖拉拽的方式构建任务流程，在统一的界面进行数据开发和调度配置。

● 任务监控与定位处理



计费

进行数据开发与运维可能产生的费用包括：

- 运行任务所需的调度资源组费用。
计费详情请参见[独享调度资源组计费说明：包年包月](#)、[公共调度资源组计费说明：按量付费](#)。
- 进行运维监控时，监控规则计算扫描所需的实例资源，和触发监控报警所需的通信费用。
计费详情请参见[基线实例计费说明](#)、[报警短信与报警电话计费](#)。

说明 进行数据开发时，可能会产生的计算引擎系统计算和存储费用，此类费用不属于 DataWorks 相关收费范畴，账单也不会体现在 DataWorks 产品下。

开通使用

开通 DataWorks 版本后，您即可根据需要购买调度资源组，绑定计算引擎，开始数据开发。开通使用的操作详情请参见[数据开发](#)章节和[运维中心](#)章节。

2.3. 数据建模：智能数据建模

智能数据建模是阿里云 DataWorks 自主研发的智能数据建模产品，沉淀了阿里巴巴十多年来数仓建模方法论的最佳实践，包含数仓规划、数据标准、维度建模及数据指标四大模块，帮助企业在搭建数据中台、数据集市建设过程中提升建模及逆向建模的能力，并通过数据建模快速构建企业数据资产。

功能概述

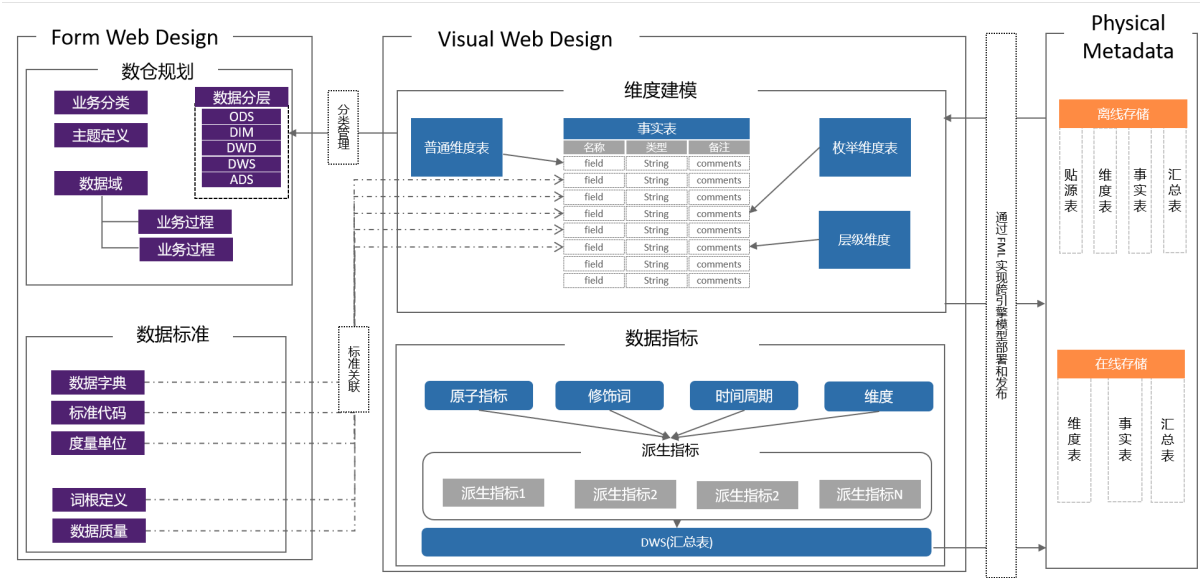
智能数据建模产品包含数仓规划、数据标准、维度建模、数据指标四大产品模块。



- **数仓规划：**数仓规划支持数仓分层、数据域、数据集市等的规划，支持设置模型设计空间，不同部门可共享一套数据标准和数据模型。
- **数据标准：**数据标准字段标准、标准代码、度量单位、命名词典的定义，支持标准代码自动生成质量规则，落标检查不再难。
- **维度建模：**维度建模支持逆向建模，解决现有数仓的建模冷启动难题，支持可视化数仓维度建模，支持通过Excel文件导入模型和通过FML（一种类SQL的DSL）快速构建模型，支持与数据开发DataStudio无缝打通，自动生成ETL代码。
- **数据指标：**数据指标支持原子指标、派生指标的定义与构建，与维度建模无缝打通，可根据原子指标和不同维度批量创建派生指标。

核心技术与架构

DataWorks智能建模包含的各模块架构图如下。



适用场景

DataWorks智能建模可助力企业构建自身建模能力，挖掘企业的数据资产价值，例如：

- 海量数据的标准化管理
企业业务越庞大数据结构就越复杂，企业数据量会随着企业业务的快速发展而迅速增长，如何结构化有序地管理和存储数据是每个企业都将面临的一个挑战。
- 业务数据互联互通，打破信息壁垒
公司内部各业务、各部门之间数据独立自主形成了数据孤岛，导致决策层无法清晰、快速地了解公司各类数据情况。如何打破部门或业务领域之间的信息孤岛是企业数据管理的严峻考验。
- 数据标准整合，统一灵活对接
同一数据不同描述，企业数据管理难、内容重复、结果不准确。如何制定统一的数据标准又不打破原有的系统架构，实现灵活对接上下游业务，是标准化管理的核心重点之一。
- 数据价值最大化，企业利润最大化
在最大程度上用好企业各类数据，使企业数据价值最大化，为企业提供更高效的数据服务。

计费

- ② 说明 DataWorks智能数据建模已于2022年4月25日24点结束免费公测，公测期间创建的所有数据均会为您提供长期保存服务。2022年4月26日0点开始正式收费，需要您购买后才能使用智能数据建模产品。
- ② 说明 DataWorks智能数据建模计划于2022年4月25日24点结束免费公测，公测期间创建的所有数据均会为您提供长期保存服务。2022年4月26日0点开始正式收费，需要您购买后才能使用智能数据建模产品。

使用智能建模产品时，可能涉及到的收费项和注意事项如下。

- 购买建模产品

购买智能数据建模产品，需要您保有或新购买DataWorks增值版本。购买DataWorks增值版本，详情请参见[版本服务计费说明](#)。

● 使用建模OpenAPI

DataWorks OpenAPI仅企业版或旗舰版支持使用。如需使用智能数据建模相关的OpenAPI，需要您同时保有DataWorks企业版或旗舰版，以及DataWorks智能数据建模产品的任一规格。

 **说明** 智能数据建模产品到期后，DataWorks企业版或旗舰版仍然保留，但智能数据建模相关的OpenAPI无法使用。

● 建模产品到期后使用

智能数据建模产品到期后，您可以继续在智能数据建模产品内浏览已创建的模型，如需新增模型或编辑已有模型，只需为智能数据建模产品续费即可。

计费详情请参见[智能数据建模计费标准](#)。

开通使用

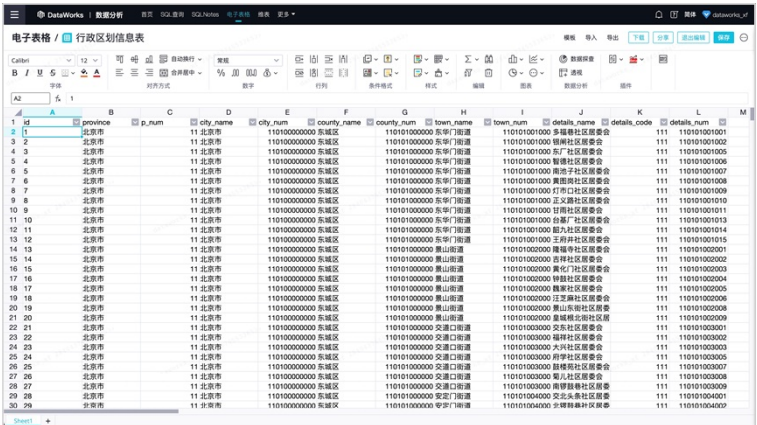
购买增值版本DataWorks后，您可开通智能建模产品，并使用智能建模的数仓规划、数据标准、维度建模、数据指标功能，操作详情请参见[智能建模](#)章节。

2.4. 数据分析：即时快速分析

数据分析基于“人人都是数据分析师”的产品目标，旨在为更多非专业数据开发人员，如数据分析、产品、运营等工作人员提供更加简洁高效的取数、用数工具，提升大家日常取数分析效率。

功能概述

数据分析支持基于个人视角的数据上传、公共数据集、表搜索与收藏、在线SQL取数、SQL文件共享、SQL查询结果下载及用电子表格进行大屏幕数据查看等产品功能。



电子表格 / 行政区划信息表

	A	B	C	D	E	F	G	H	I	J	K	L
	id	province	p_num	city_name	city_num	county_name	county_num	town_name	town_num	details_name	details_code	details_num
1	1	北京市	11	北京市	1101000000	东城区	1101010000	东单门牌	1101010000	多福源社区居委会	111	110101001001
2	2	北京市	11	北京市	1101000000	东城区	1101010000	东单门牌	1101010000	锦麟社区居委会	111	110101001002
3	3	北京市	11	北京市	1101000000	东城区	1101010000	东单门牌	1101010000	五广社区居委会	111	110101001005
4	4	北京市	11	北京市	1101000000	东城区	1101010000	东单门牌	1101010000	智德社区居委会	111	110101001006
5	5	北京市	11	北京市	1101000000	东城区	1101010000	东单门牌	1101010000	南池子社区居委会	111	110101001007
6	6	北京市	11	北京市	1101000000	东城区	1101010000	东单门牌	1101010000	黄面园社区居委会	111	110101001008
7	7	北京市	11	北京市	1101000000	东城区	1101010000	东单门牌	1101010000	灯市口社区居委会	111	110101001009
8	8	北京市	11	北京市	1101000000	东城区	1101010000	东单门牌	1101010000	正阳门社区居委会	111	110101001010
9	9	北京市	11	北京市	1101000000	东城区	1101010000	东单门牌	1101010000	甘雨社区居委会	111	110101001011
10	10	北京市	11	北京市	1101000000	东城区	1101010000	东单门牌	1101010000	合顺广社区居委会	111	110101001013
11	11	北京市	11	北京市	1101000000	东城区	1101010000	东单门牌	1101010000	福九社区居委会	111	110101001014
12	12	北京市	11	北京市	1101000000	东城区	1101010000	东单门牌	1101010000	王村社区居委会	111	110101001015
13	13	北京市	11	北京市	1101000000	东城区	1101010000	泰山街道	1101010000	魏家社区居委会	111	110101002001
14	14	北京市	11	北京市	1101000000	东城区	1101010000	泰山街道	1101010000	魏家社区居委会	111	110101002002
15	15	北京市	11	北京市	1101000000	东城区	1101010000	泰山街道	1101010000	魏家社区居委会	111	110101002003
16	16	北京市	11	北京市	1101000000	东城区	1101010000	泰山街道	1101010000	魏家社区居委会	111	110101002004
17	17	北京市	11	北京市	1101000000	东城区	1101010000	泰山街道	1101010000	魏家社区居委会	111	110101002005
18	18	北京市	11	北京市	1101000000	东城区	1101010000	泰山街道	1101010000	魏家社区居委会	111	110101002006
19	19	北京市	11	北京市	1101000000	东城区	1101010000	泰山街道	1101010000	魏家社区居委会	111	110101002008
20	20	北京市	11	北京市	1101000000	东城区	1101010000	泰山街道	1101010000	魏家社区居委会	111	110101002009
21	21	北京市	11	北京市	1101000000	东城区	1101010000	泰山街道	1101010000	魏家社区居委会	111	110101002010
22	22	北京市	11	北京市	1101000000	东城区	1101010000	泰山街道	1101010000	魏家社区居委会	111	110101002011
23	23	北京市	11	北京市	1101000000	东城区	1101010000	泰山街道	1101010000	魏家社区居委会	111	110101002012
24	24	北京市	11	北京市	1101000000	东城区	1101010000	泰山街道	1101010000	魏家社区居委会	111	110101002013
25	25	北京市	11	北京市	1101000000	东城区	1101010000	泰山街道	1101010000	魏家社区居委会	111	110101002014
26	26	北京市	11	北京市	1101000000	东城区	1101010000	泰山街道	1101010000	魏家社区居委会	111	110101002015
27	27	北京市	11	北京市	1101000000	东城区	1101010000	泰山街道	1101010000	魏家社区居委会	111	110101002016
28	28	北京市	11	北京市	1101000000	东城区	1101010000	泰山街道	1101010000	魏家社区居委会	111	110101002017
29	29	北京市	11	北京市	1101000000	东城区	1101010000	泰山街道	1101010000	魏家社区居委会	111	110101002018
30	30	北京市	11	北京市	1101000000	东城区	1101010000	泰山街道	1101010000	魏家社区居委会	111	110101002019
31	31	北京市	11	北京市	1101000000	东城区	1101010000	泰山街道	1101010000	魏家社区居委会	111	110101002020

电子表格
支持本地数据上传与在线数据源，兼容大总分Excel函数计算，支持透视分析和数据探查。

SQL查询
支持多种数据源，更便捷更轻量的SQL查询取数工具

公共数据集
提供公共数据集

适用场景

适合更多非专业数据开发人员，如数据分析、产品、运营等工作人员进行高效、海量、流动、安全地数据分析工作。

- 海量：借助计算引擎的能力，可以高效分析全量、海量的数据。
- 流动：在线数据分析可以从不同业务系统的数据库获取数据进行分析。DataWorks数据分析支持导出数据为MaxCompute表，或分享结果数据至指定成员，并赋予其权限。数据可以在不同的系统和人员之间流动。

- 安全：所有SQL查询、SQL结果下载等功能的操作都可以接入安全审计。

计费

DataWorks的数据分析功能不收取功能费用，您开通DataWorks后即可使用数据分析功能，但不同版本的数据分析功能细节不一致，详情请参见[DataWorks各版本详解](#)。

开通使用

购买DataWorks版本后，您即可登录DataWorks控制台使用数据分析功能，操作详情请参见[数据分析](#)章节。

2.5. 数据质量：全流程的质量监控

DataWorks的全流程数据质量监控功能为您提供35种预设表级别、字段级别和自定义的监控模板。

数据质量帮助您第一时间感知到源端数据的变更与ETL（Extract Transformation Load）中产生的脏数据，自动拦截问题任务，有效阻断脏数据向下游蔓延。

 说明 ETL是抽取、转换和加载源端数据至目的端的过程。

数据质量以数据集（DataSet）为监控对象，支持监控MaxCompute数据表和DataHub实时数据流。当离线MaxCompute数据发生变化时，数据质量会对数据进行校验，并阻塞生产链路，以避免问题数据污染扩散。同时，数据质量提供历史校验结果的管理，以便您对数据质量进行分析和定级。详情请参见[数据质量](#)。

数据质量为您解决以下问题：

- 数据库频繁变更问题
- 业务频繁变化问题
- 数据定义问题
- 业务系统的脏数据问题
- 系统交互导致质量问题
- 数据订正引发的问题
- 数据仓库自身导致的质量问题

2.6. 数据地图：统一管理，跟踪血缘

DataWorks的数据地图功能可以帮助您实现对数据的统一管理和血缘的跟踪。

[数据地图](#)以数据搜索为基础，提供表使用说明、数据类目、数据血缘、字段血缘等工具，帮助数据表的使用者和拥有者更好地管理数据、协作开发。



2.7. 数据服务：低成本快速发布API

DataWorks的数据服务功能模块是灵活轻量、安全稳定的数据API构建平台，旨在为企业提供全面的数据共享能力，帮助用户从发布审批、授权管控、调用计量、资源隔离等方面实现数据价值输出及共享开放。

功能概述

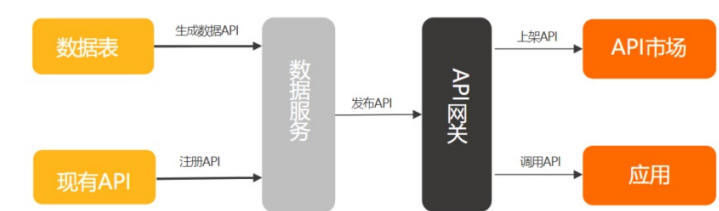
作为数据仓库与上层应用系统间的“桥梁”，DataWorks数据服务为企业搭建统一的服务总线，帮助企业统一创建及管理对内、对外的API服务，解决数仓、数据库与数据应用间的“最后一公里”，加速数据的流动和共享。



- 数据服务支持通过零代码或自助SQL的双模式，将各类数据源下的数据表生成数据API，同时支持函数计算来辅助加工API的请求参数及返回结果。
- 数据服务采用Serverless架构，用户无需关心运行环境等基础设施，即可将API服务一键发布至API网关。

核心技术与架构

数据服务采用Serverless架构，您只需要关注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。



计费

使用DataWorks的数据服务功能创建数据API后，API调用请求需使用数据服务资源组，您可以根据业务需要选用，对应资源组的计费详情请参见[独享数据服务资源组计费说明：包年包月](#)、[数据服务公共数据服务资源组计费说明：按量付费](#)。

开通使用

购买DataWorks版本后，您即可登录DataWorks控制台使用数据服务功能，操作详情请参见[数据服务](#)章节。

2.8. 开放平台：能力全面开放

DataWorks开放平台是DataWorks对外提供数据和能力的开放通道。DataWorks开放平台提供开放API（OpenAPI）、开放事件（OpenEvent）、扩展程序（Extensions）的能力，可以帮助您快速实现各类应用系统对接DataWorks、方便快捷的进行数据流程管控、数据治理和运维，及时响应应用系统对接DataWorks的业务状态变化。

功能概述

DataWorks开放平台提供开发API（OpenAPI）、开放事件（OpenEvent）、扩展程序（Extensions）等能力。

- 开发API（OpenAPI）

通过OpenAPI可以实现您的自有应用与DataWorks的深度集成，例如实现批量创建任务、发布任务、运维任务等，提升您的大数据处理效率，减少人工操作成本。

关于OpenAPI功能，具体请参见[开放API（OpenAPI）](#)。

- 开放事件（OpenEvent）

通过OpenEvent可以允许您订阅DataWorks中的系统事件，实时获取并响应事件变化，例如订阅表变更事件实现对核心表的实时监控，订阅任务变更事件实现自定义实时任务监控大屏。

关于开放事件功能，具体请参见[开放事件](#)。

- 扩展程序（Extensions）

Extensions则是将OpenAPI和OpenEvent有机结合起来的服务级插件，通过Extensions允许您对DataWorks中的流程控制进行自定义，例如您可以自定义任务发布管控插件，从而对不符合规范和要求的任务进行拦截。

关于扩展程序功能，具体请参见[扩展程序](#)。

适用场景

DataWorks开放平台提供了全面的开放能力，可以实现深度的系统集成、自动化操作、流程定义、业务监控等，欢迎广大用户及合作伙伴，基于DataWorks的开放平台来实现行业化、场景化的数据应用和插件。

计费

- OpenAPI当前仅DataWorks企业版和旗舰版支持使用，不同版本用户可免费调用OpenAPI的次数不一致，超出免费额度后会按调用次数按量收费，计费详情请参见[OpenAPI计费说明](#)。
- OpenEvent和Extensions当前邀测中，仅DataWorks企业版用户可参加邀测，申请邀测后，在邀测期间不收取其他费用。

开通使用

购买DataWorks所需版本后（企业版或旗舰版），您即可使用OpenAPI功能，操作详情请参见[API参考](#)章节；OpenEvent和Extensions功能需以[提交工单](#)参与邀测并试用，操作详情请参见[OpenEvent](#)章节、[Extensions](#)章节。

2.9. 迁移助手与迁云服务

DataWorks迁移助手支持将开源调度引擎的作业迁移至DataWorks，支持作业跨云、跨Region、跨账号迁移，实现DataWorks作业快速克隆部署，同时DataWorks团队联合大数据专家服务团队，上线迁云服务，帮助您快速实现数据与任务的上云。

功能概述

迁移助手与迁云服务主要功能包括：

- 任务上云：实现将开源调度引擎的作业搬迁至DataWorks上。
- DataWorks迁移：实现DataWorks体系内的开发成果互相迁移。

适用场景

主要适用于：

- 任务上云：开源调度引擎的作业搬迁至DataWorks上。
- 备份任务：用户可通过迁移助手定期备份任务代码，减少误删项目带来的损失。
- 业务快速复制：用户可抽象出通用业务，通过迁移助手导出导入的方式快速复制。
- 快速创建测试环境：通过迁移助手全量复制业务代码，只需修改生产数据库的数据输入为测试数据，可快速搭建测试环境。
- 跨云开发：支持公共云的DataWorks和私有云的DataWorks之间进行互导，达到协同开发。

计费

您开通DataWorks后即可使用迁移助手功能，但不同版本的DataWorks支持的功能细节不一致，详情请参见[DataWorks各版本详解](#)。

开通使用

购买DataWorks版本后，您即可登录DataWorks控制台使用迁移助手功能，操作详情请参见[迁移助手](#)章节。

2.10. 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。



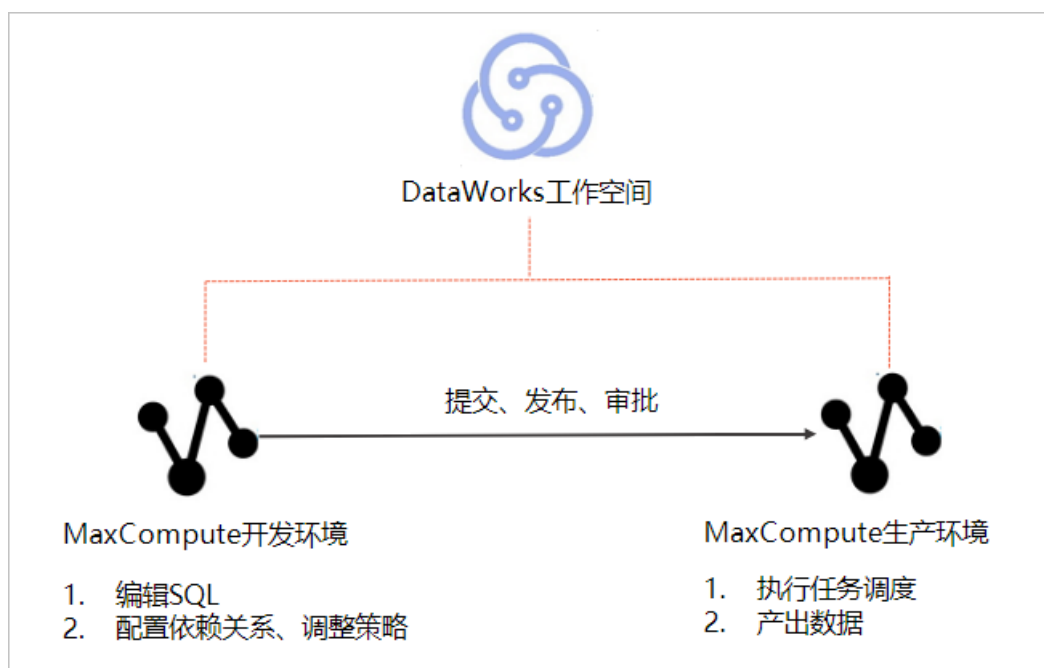
标准模式的工作空间

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[计算引擎信息](#)。

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RAM用户 - 阿里云RAM角色
	EMR（E-MapReduce）	开发环境	- 快捷模式下访问身份：统一使用集群内的Hadoop用户。 - 安全模式下访问身份：任务执行者
		生产环境	- 快捷模式下访问身份：统一使用集群内的Hadoop用户。 - 安全模式下访问身份（可选）： - 任务责任人 - 阿里云主账号 - 阿里云RAM用户
	Hologres	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RAM用户
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）： - 阿里云主账号 - 阿里云RAM角色 - 任务负责人：任务Owner账号的身份
	EMR	开发环境即生产环境	- 快捷模式下访问身份：统一使用集群内的Hadoop用户。 - 安全模式下访问身份（可选）： - 任务责任人 - 阿里云主账号 - 阿里云RAM用户

工作空间模式	计算引擎类型	环境	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RAM用户

AnalyticDB for MySQL、AnalyticDB for PostgreSQL等引擎与您新增引擎实例时不同环境绑定的账号有关，权限与账号在数据库中的权限一致。

不同模式工作空间的权限管理特征

DataWorks采取RBAC权限模型供用户管理DataWorks所有页面可见功能以及API的使用权限，同时这套权限体系与MaxCompute的RBAC角色体系存在天然的映射关系，详情可参见[角色及成员管理：空间级](#)。不同工作空间类型的权限管理特征与优缺点不一致，以下表格为您对比介绍两种空间类型的权限细分特点。

细分特点	简单模式	标准模式
权限概述	在简单模式空间下，DataWorks的“开发”角色因为与所绑定MaxCompute项目的“Role_Project_Dev” Role进行了映射，因此DataWorks开发角色天然能够读取MaxCompute项目内的所有数据。	在标准模式空间下，DataWorks的“开发”角色因为与所绑定MaxCompute项目（dev环境）的“Role_Project_Dev” Role进行了映射，因此： • DataWorks开发角色天然能够读取MaxCompute项目（dev环境）内的所有数据。 • 由于没有和MaxCompute项目（PROD环境）的role映射，因此默认情况下DataWorks开发角色无MaxCompute（PROD环境）的数据权限。
优点	简单、方便、易用。 仅需要授权数据开发人员“DataWorks开发角色”即可完成所有数据仓库开发工作。	安全、规范。 • 具备安全、规范的代码发布管控流程（包含代码评审、代码DIFF查看等功能），保障生产环境稳定性，避免不必要的因代码逻辑引起的脏数据蔓延或任务报错等非预期情况。 • 数据访问得到有效管控，数据安全得以保障。
缺点	存在不稳定、不安全的风险。 • 开发角色可以不经任何人审批，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。 • 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，可随意对表进行增加、删除和修改等操作，存在数据安全风险。	流程相对复杂，一般情况下无法一人完成所有数据开发、生产流程。

MaxCompute引擎在不同模式工作空间访问的相关问题

场景	工作空间模式	数据开发界面	生产运维中心
执行账号	标准模式	当前操作人	调度引擎指定账号
	简单模式	调度引擎指定账号  说明 简单模式工作空间下，如果调度引擎指定账号为阿里云主账号，无论是否是主账号进行执行任务的操作，其都将使用主账号身份执行任务。	
资源所在环境	标准模式	projectname_dev.tablename/function/resource	projectname.tablename/function/resource
	简单模式	projectname.tablename/function/resource	
访问资源	标准模式	场景一： <code>select col1 from tablename</code> - 用个人账号访问开发环境下该表。 - 即用个人账号访问：<code>projectname_dev.tablename</code> 场景二： <code>select col1 from projectname.tablename</code> - 用个人账号访问生产环境下该表。 - 即用个人账号访问：<code>projectname.tablename</code>  说明 阿里云RAM用户（非调度引擎指定账号）默认无权限访问生产环境，生产表权限需要在安全中心申请。	
	简单模式	<code>select col1 from tablename.</code> - 用调度引擎指定账号访问生产表。 - 即用调度引擎指定账号访问生产表：<code>projectname.tablename</code>  说明 简单模式工作空间下，如果调度引擎指定账号为阿里云主账号，无论是否是主账号进行执行任务的操作，其都将使用主账号身份访问该资源。	
	标准模式	个人所拥有的资源权限。	调度引擎指定账号所拥有的权限。
	简单模式		

场景 访问资源权限问题	工作空间模式	数据开发界面	生产运维中心
	简单模式	调度引擎指定账号所拥有的权限。  说明 如果当前简单模式工作空间使用的调度引擎指定账号为阿里云主账号，权限过大，不建议您直接使用阿里云主账号作为简单模式工作空间调度引擎执行账号。	

MaxCompute引擎在不同模式下数据库表命名规范

简单模式下不区分开发环境和生产环境，开发库即生产库。标准模式下，支持开发环境和生产环境隔离，开发环境和生产环境的数据库表命名有所区别，如果需要在开发环境访问生产环境的数据库表，请根据以下命名规范严格区分数据库表名，避免误操作生产环境。

环境类型	标准模式	示例
开发环境	项目名_dev.表名	在projectA项目下创建一个开发库表user_info，则数据库表名为：projectA_dev.user_info。
生产环境	项目名.表名	在projectA项目下创建一个生产库表user_info，则数据库表名为：projectA.user_info。

3. 产品优势

DataWorks具有强大的基础能力，可以为您大幅提升工作效率，保障数据准时产出，助力数据治理，让您零成本构建数据服务。

学习成本低

非技术人员1~2小时即可掌握完整的数据开发、治理流程，告别传统命令行，节省巨大的学习成本。

让您可以在同一DAG图中，构建异构计算引擎形成混编任务流（数据同步+SQL+MR+MaxCompute Spark+实时计算+ML），无需分别维护各技术栈，助您高效组合混编任务流。

人效提升快

一键开通即可完成开箱即用的数仓技术架构搭建，告别繁重的自研、部署、维护工作，为企业免去数十人运维开发团队。

产品功能全

功能涵盖数据传输、开发、生产、治理、安全全领域，每个领域深度覆盖大数据全生命周期，轻松帮助企业应对在搭建数仓、搭建数据中台、数字化转型项目中遇到的难题。

- 支持复杂网络环境、常见数据源的数据同步上云以及实时、历史数据的批量与增量同步。
- 在数据源性能不受限情况下，让您的同步速度达到万兆。
- 支持单用户千万级别的复杂任务调度，让数据加工更流畅。

4. 应用场景

4.1. 受众与核心能力

本文为您介绍DataWorks的产品定位、产品受众和产品的核心能力。

产品定位

DataWorks致力于为数据开发者、数据分析师、数据资产管理者，打造一个具备开放自主开发与全栈数据研发能力的一站式、标准化、可视化、透明化的智能大数据全生命周期云研发平台。DataWorks赋予用户仅通过单一平台，即可实现数据传输、数据计算、数据治理、数据分享的各类复杂组合场景的能力。

同时，DataWorks持续打造符合企业级数仓、数据中台构建要求的功能模块，为企业业务的数字化转型提供支持。

产品受众

- 从事数据开发、算法开发等岗位的技术人员
- 从事销售运营、商业智能分析等岗位的业务人员
- 从事数据安全与合规工作的管理人员
- 从事数据应用开发的开发人员
- 把控公司核心数据资产的管理人员

核心能力

基于DataWorks，您可以获得如下能力：

- **数据集成**：复杂网络环境、丰富数据源之间的数据传输与上云。
- **数据开发**：在线批处理、流处理和机器学习等多引擎任务开发，构建复杂的调度依赖，提供开发、生产环境隔离的研发模式。
- **实时分析（仅公共云）**：提供基于电子表格的快速、灵活的即时查询。
- **数据服务**：零代码快速生成Serverless化的API。
- **数据质量**：通过表级别、字段级别监控规则定义，第一时间感知脏数据。
- **智能监控**：一键实现复杂工作流的全链路监控报警配置。
- **数据地图（公共云）/数据管理（专有云）**：提供强大的数据搜索、数据类目、数据血缘等能力。
- **数据资产管理（仅专有云）**：统一管理整个平台的数据表、API等各类数据资产。
- **数据安全**：数据审计、数据脱敏、权限控制等能力。
- **应用开发（仅公共云）**：基于Web端的组件拖拉拽轻松构建数据应用。
- **工作空间管理（公共云）/平台管理（专有云）**：从系统层面，为管理者提供对使用DataWorks的用户（成员）权限、DataWorks底层计算引擎配置的管理能力。

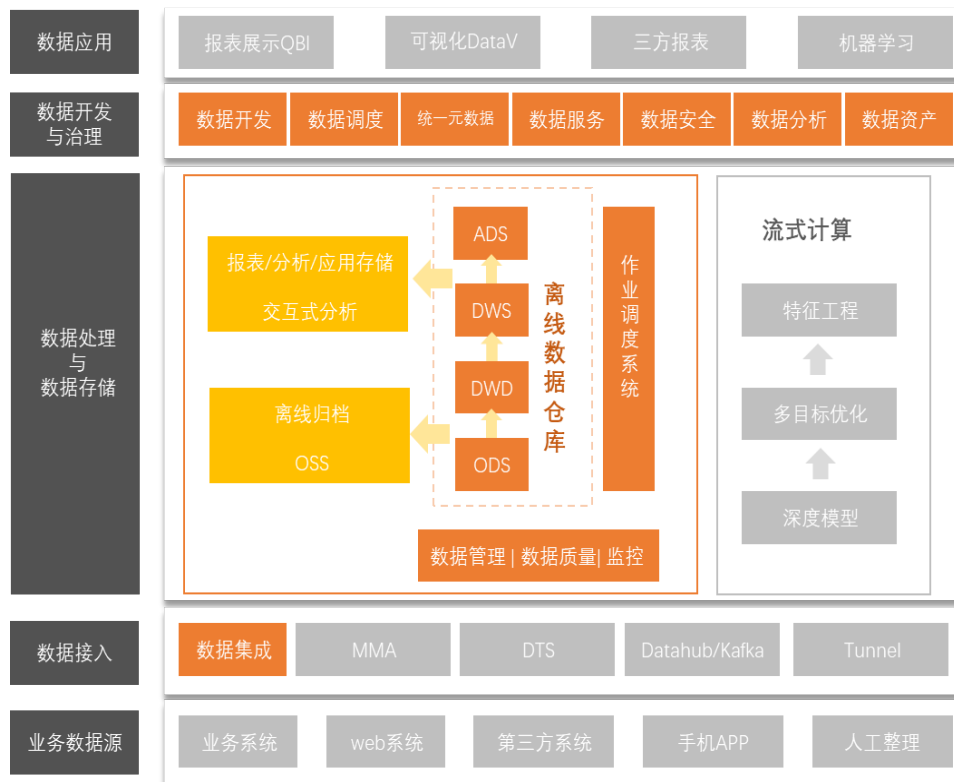
总之，使用DataWorks，您不仅可以进行海量数据的离线加工分析，还能完成数据的汇聚集成、开发、生产调度与运维、离线与实时分析、数据质量治理与资产管理、安全审计、数据共享与服务、机器学习、应用搭建等覆盖大数据全生命周期的最佳实践。让数据从采集到展现、从分析到驱动应用得以一站式解决，真正实现数据业务化、业务数据化。

4.2. 构建数据仓库

DataWorks具有通过可视化方式实现数据开发、治理全流程相关的核心能力，本文将为您介绍DataWorks在构建云上大数据仓库和构建智能实时数据仓库两个典型应用场景下的应用示例。

构建云上大数据仓库

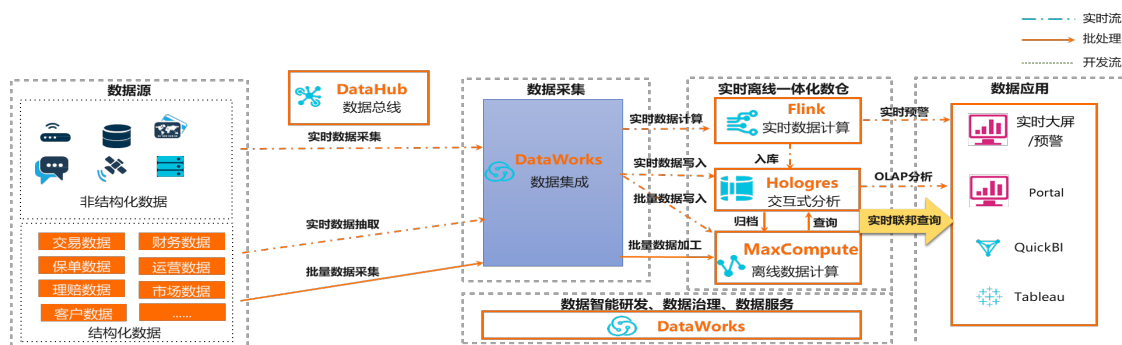
本场景推荐的架构如下。



- **适用行业**：全行业适用。
- **方案优势**：阿里巴巴大数据最佳实践，高性能、低成本、Severless服务，免运维、全托管模式，让企业的大数据研发人员更聚焦在业务数据的开发、生产、治理。
- **产品组合**：MaxCompute + Flink + DataWorks。
- **场景说明**
 - 用户数据来源丰富，包括来自云端的数据、外部数据源，数据统一沉淀，完成数据清洗、建模。
 - 用户的应用场景复杂，对非结构化的语音、自然语言文本进行语音识别、语义分析、情感分析等，同时融合结构化数据搭建企业级的数据管理平台，并且计算和存储成本最低。
 - 平台支撑多种形式的的应用，包括使用机器学习算法进行复杂数据分析、使用BI报表进行图表展现、使用可视化产品进行大屏展示、使用其他自定义的方式消费数据。

构建智能实时数据仓库

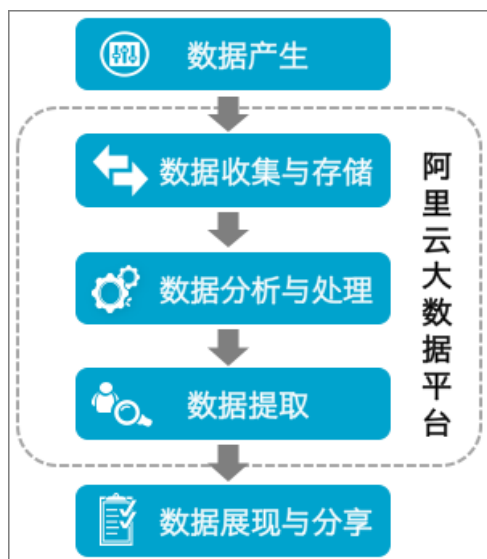
本场景推荐的架构如下。



- **适用行业：**适用于电商、游戏、社交等互联网行业大规模数据实时查询场景。
- **方案优势：**
 - 阿里云实时数仓全套链路 with 离线数仓无缝打通。
 - 满足一套存储，两种计算（实时计算和离线计算）的高性价比组合。
- **产品组合：**DataHub+实时计算Flink+交互式分析+MaxCompute+DataWorks+Quick BI / DataV
- **场景说明：**
 - **数据采集：**通过DataWorks（批量）、DataHub（实时）进行统一数据采集接入。
 - **数据开发：**基于DataWorks进行数据全链路研发，包括数据集成、数据开发和ETL、转换及计算等开发，以及数据作业的调度、监控、告警等。DataWorks提供数据开发链路的安全管控的能力，以及基于DataWorks数据服务模块提供统一数据服务API能力。
 - **实时数据：**按实际业务需求使用Flink进行实时ETL（可选）、结果入库，使用交互式分析产品构建实时数据仓库、应用集市，并提供海量数据的实时交互查询和分析。
 - **交互式分析：**提供实时离线联邦查询。历史离线数据存放于MaxCompute，实时分析数据存放于交互式分析。基于阿里云Quick BI或第三方数据分析工具（如Tableau）行数据可视化，以及构建各业务板块数据服务门户应用。

4.3. 通用数据开发

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。



② 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：

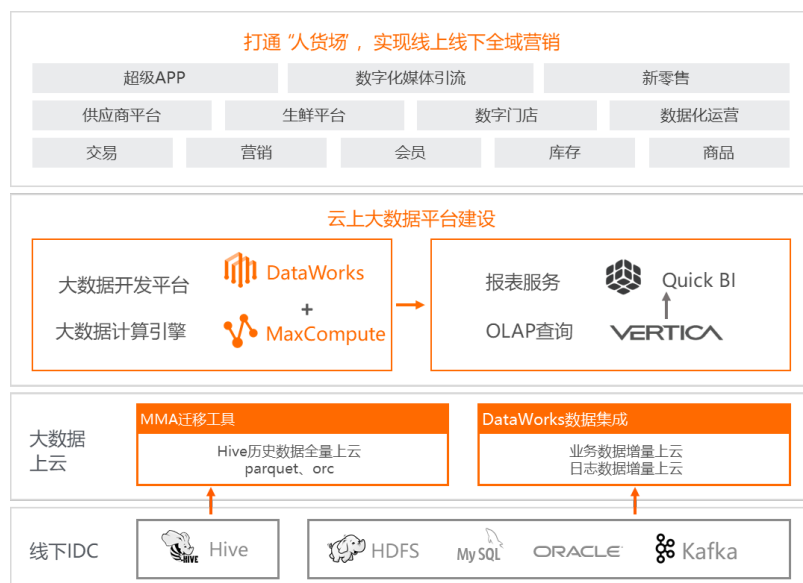
1. **数据产生**：业务系统每天会产生大量结构化的数据，存储在业务系统所对应的数据库中，包括MySQL、Oracle和RDS等类型。
2. **数据收集与存储**：您需要同步不同业务系统的数据至MaxCompute中，方可通过MaxCompute的海量数据存储与处理能力分析已有的数据。
DataWorks提供数据集成服务，可以支持多种数据源类型，根据预设的调度周期同步业务系统的数据至MaxCompute。
3. **数据分析与处理**：完成数据的同步后，可以对MaxCompute中的数据进行加工（MaxCompute SQL、MaxCompute MR）、分析与挖掘（数据分析、数据挖掘）等处理，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. 客户案例

DataWorks在多个行业中均有典型的案例落地，帮助多个行业的企业解决数据痛点，挖掘数据价值，本文为您介绍典型行业中已落地的客户案例。

新零售行业：大润发云上数据中台建设

客户架构如下。



● 客户简介

为了快速数智化转型，拥抱新零售，大润发计划两年内将IT系统全面迁移到阿里云上，不再自建IDC。同时与阿里云合作启动数据中台项目，可以帮助大润发降低TCO的同时，更好的依托云上生态，实现数据资产业务化闭环。

● 客户需求

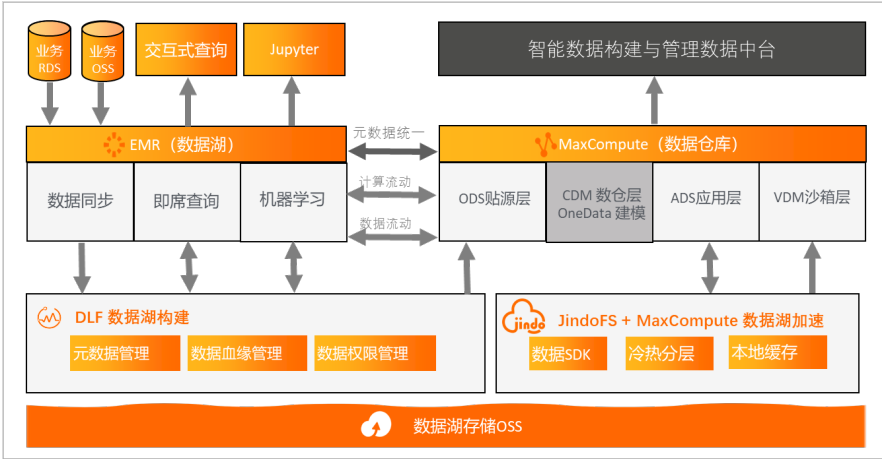
- 基于Hadoop开源生态打造，软硬件维护成本高昂，稳定性问题不断，严重影响业务经营分析。
- 线上业务爆发，需求积压严重，期望有整体解决方案，能够快速灵活支持业务发展所需的技术扩展。

● 价值体现

通过MMA工具，15天完成400TB+历史数据迁移，同时保证了迁移的准确性，让客户拥有平滑高效的上云体验。基于飞天大数据平台产品DataWorks+MaxCompute大大提高了数据业务的开发效率，构建大润发独有的数据中台体系。

新金融行业：某互联网金融公司湖仓一体案例

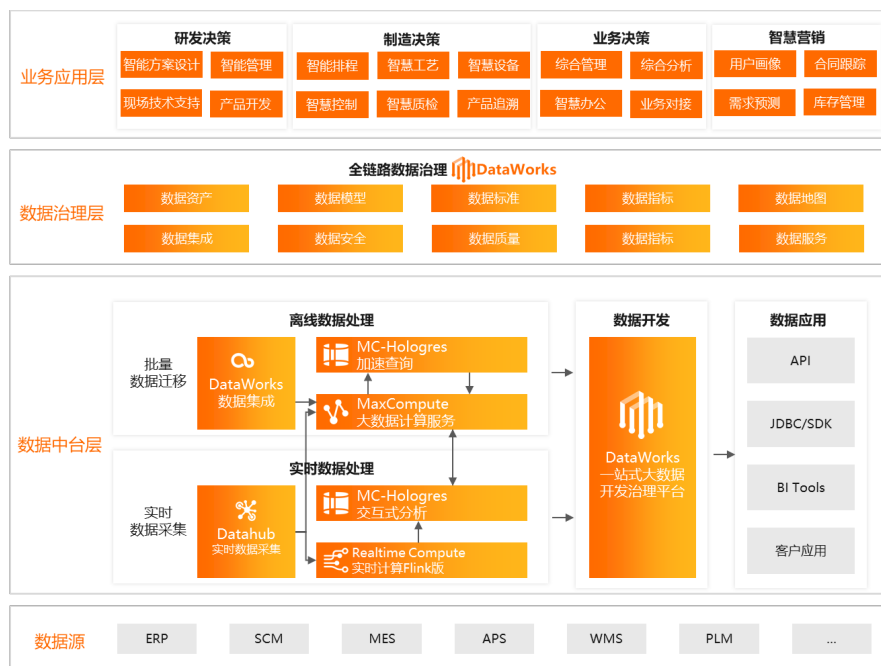
客户架构如下。



- 客户简介
公司的第一代数据湖是基于Hadoop + OSS搭建的，同时引入的数据中台的执行引擎和存储是MaxCompute，两套异构的执行引擎带来存储冗余、元数据不统一、权限不统一、湖仓计算不能自由流动。
- 客户需求
如架构图所示，MaxCompute和EMR不同引擎用于不同的业务场景，使用阿里云数据湖构建DLF统一做元数据管理和统一用户权限管理。通过DataWorks进行全链路数据治理，提升数据质量与应用能力。
- 价值体现
 - 将EMR的元数据统一到DLF，底层使用OSS作统一存储，并通过湖仓一体打通EMR数据湖和MaxCompute数仓两套体系，让数据和计算在湖和仓之间自由流动。
 - 实现湖仓数据分层存储。数据中台对数据湖数据进行维度建模的中间表存储在MaxCompute上，EMR或其他引擎消费ADS层。

新能源：某能源客户基于DataWorks全链路数据治理案例

客户架构如下。



● 客户简介

- 多家子公司经过多年建设，系统数量多，技术路线复杂多样。
- 数据分散，数据标准定义混乱，各类数据出现断层，无法有效用于分析。
- 数据管理权责不明，缺乏数据治理，没有有效的数据共享机制。

● 客户需求

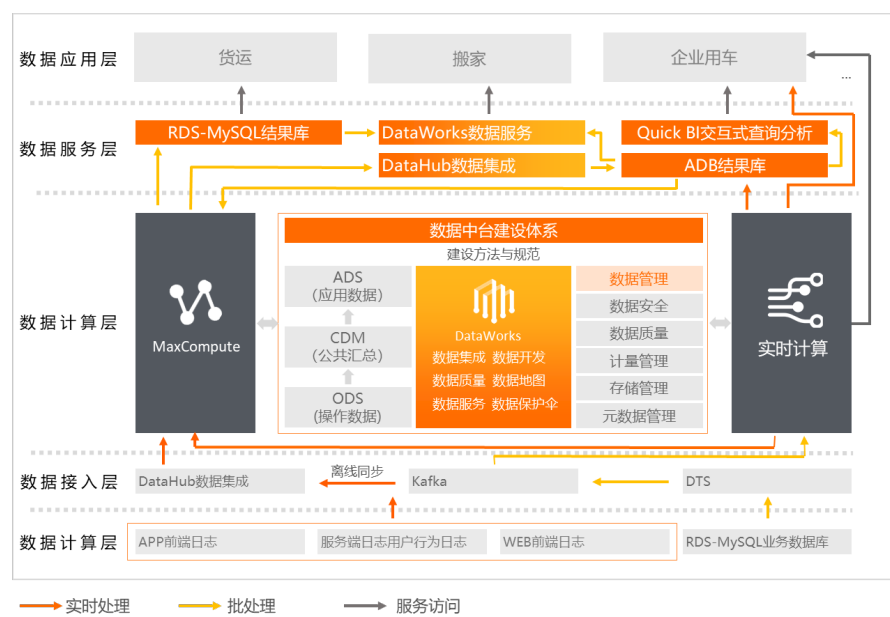
- 通过DataWorks+MaxCompute搭建数据中台，打破数据孤岛。
- 通过Realtime Compute+MaxCompute交互式分析（Hologres）提升数据中台实时性。
- 通过DataWorks进行全链路数据治理，提升数据质量与应用能力。

● 价值体现

- 打造B2B智慧营销系统，实现智能制造+互联网模式落地。
- 打造离线实时一体化数据中台，构建统一、完整的大数据应用链路，服务内部几大核心业务。
- 全链路数据治理提高数据可用性，让数据在中台进行自由流动，保证数据准确、准时、一致，成本削减1亿元。
- 提高业务迭代效率，数据更新频率由1天变成10分钟，新需求上线由1周变成1天。

互联网行业：快狗打车云上大数据仓库

客户架构如下。



● 客户简介

快狗打车则一直坚持通过“连接网络化”、“运力共享化”、“过程数据化”、“匹配智能化”等数字信息化解决方案，将闲散运力统一整合到平台上，通过大数据将运力精准匹配市场需求，实现运力的节能减排，降低空驶率，有效提升行业运行效率，积极推动绿色物流发展。

● 客户需求

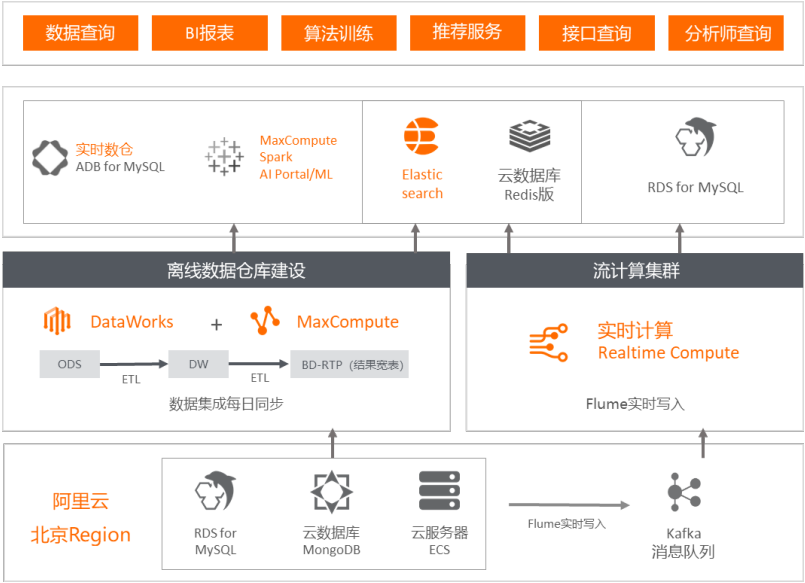
- 海量数据处理效率下降，离线数据计算时长不稳定。
- 实时计算不开发维护代价大，希望对数仓进行综合治理。

● 价值体现

基于飞天大数据平台产品，快狗打车不仅机器成本节约30%以上，数据开发效率提升100%。从Java Storm 迁移到Flink SQL使实时计算开发周期大大减少，维护更加容易，数据一致性得到更好地保障，提升了业务监控大屏的准确性和实时性，用户可以更专注于业务，加速了业务的实时化。同时，阿里云的24小时运维服务保证了集群稳定，实现了零故障。

互联网行业：宝宝树云上大数据仓库

客户架构如下。



● 客户简介

宝宝树成立于2007年，中国最大、最活跃的母婴类社区平台。作为最早做互联网2C的社区平台之一，宝宝树很早就建立了自己的IDC集群，而且规模越来越大。

● 客户需求

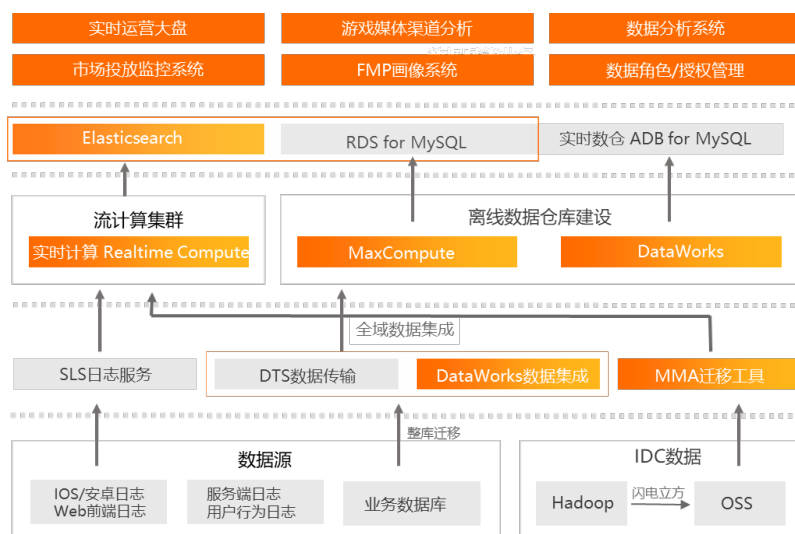
- 集群水位高，性能差，亟待大数据综合治理。
- IDC大数据每年投入成本高，希望降本提效。

● 价值体现

从大数据平台上云整体“降本增效”的方案快速切入，迁移到大数据MaxCompute、实时计算、DataWorks后，部分任务有10倍以上的性能提升，存储从自建Hadoop 3PB降到900T，利用Flink实时数据处理能力，将宝宝树现有的场景实时化（“基于用户ID维度和内容类型的实时行为”、“获取用户的实时群聊ID”及“获取文章的实时发布信息”），并且基于Flink进行实时推荐增加转化率。大数据平台整体成本节省30%以上。

游戏行业：DeNA中国游戏全链路运营

客户架构如下。



● 客户简介

DeNA是优秀的网络服务公司，随着游戏项目的生命周期越来越短，项目的各个阶段走向实时、精准的把控，需要构建经济、高效的精细化数据运营体系。

● 客户需求

- 存在Hadoop1.0、2.0两个集群，技术架构复杂，平台的稳定性和安全性、资源的弹性伸缩能力都遇到了瓶颈。
- 日志来源渠道多，实时性要求高，基于fluentd的文件采集服务，随着日志量的增多，性能、稳定性存在明显瓶颈。
- “人肉脚本”的数据开发方式，业务支持效率低，且hive计算性能无法满足需求。

● 价值体现

DeNA中国是游戏行业首家应用闪电立方+MMA工具项目，在无专线等环境下，1个多月就完成10年增量RDS库300TB+历史数据50TB的数据迁移，具备较高的技术复杂度。相较之前客户之前使用的基于python开源的airflow任务管理系统来说，DataWorks拥有以下优势：

- 任务管理一目了然，任务出错定位以及即时跳转到相关任务代码修复。
- 数据源一次性管理，不需要重复劳动，可被多种数据服务需求使用游戏业务拥有上百个数据源。
- 整体技术下沉，使得资源调度等都不需要自己耗费精力和“额外”的coding，而实现专注于管理开发。

迁移完成后，飞天大数据平台覆盖数据采集>存储&计算>实时/离线分析等游戏数据运营全链路。

6.相关的云服务

DataWorks作为阿里云一站式大数据开发与治理平台，通常会与计算引擎产品联合使用，此外使用DataWorks进行数据集成时通常联合进行数据传输的数据源产品一起使用。本文为您介绍典型场景下，使用DataWorks时通常会使用到的其他云产品。

计算引擎类云产品

DataWorks支持将计算引擎绑定至DataWorks的工作空间，绑定计算引擎后，您即可在DataWorks上创建对应引擎的计算任务，并进行周期调度。当前支持绑定的计算引擎类型如下：

- MaxCompute
- E-MapReduce
- Hologres
- ADB for Postgre
- ADB for MySQL
- CDH
- ClickHouse

绑定计算引擎的操作可参见[配置工作空间](#)。

数据源类云产品

使用DataWorks进行数据集成时，支持将数据从不同数据源间进行离线或实时同步，各类阿里云或自建关系型数据库、非结构化存储、大数据存储、消息队列等产品均支持添加为DataWorks的数据源，添加完成后即可使用DataWorks进行数据集成。

- 离线同步支持的数据源情况请参见[支持的数据源与读写插件](#)。
- 离线同步支持的数据源情况请参见[实时同步支持的数据源](#)。

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，提高您的使用体验。

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。

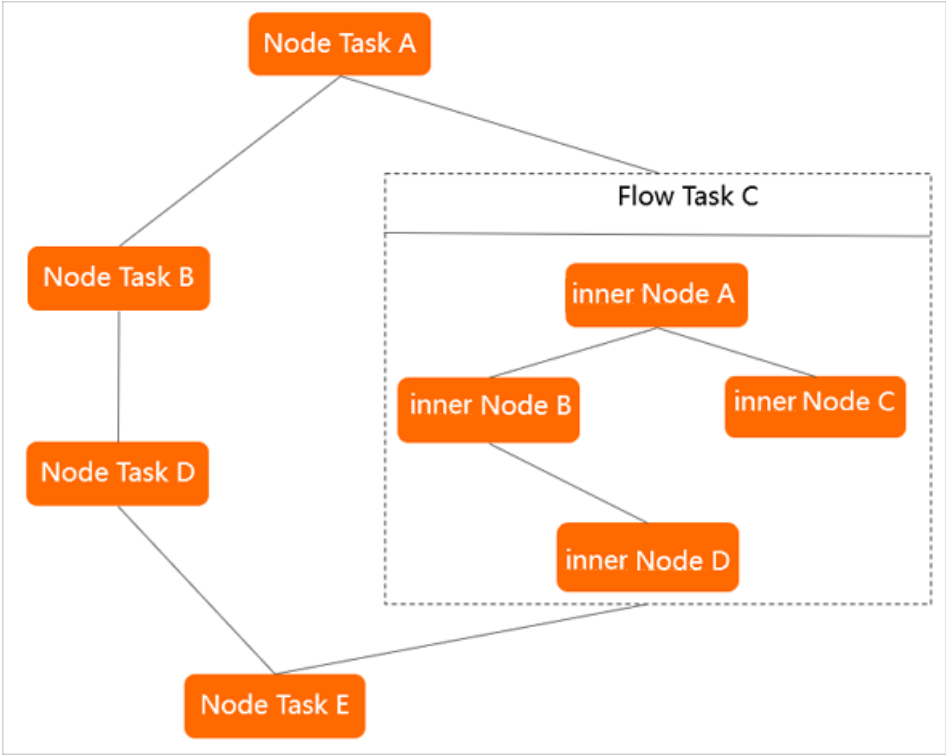
任务（Task）

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（inner Node）。



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。

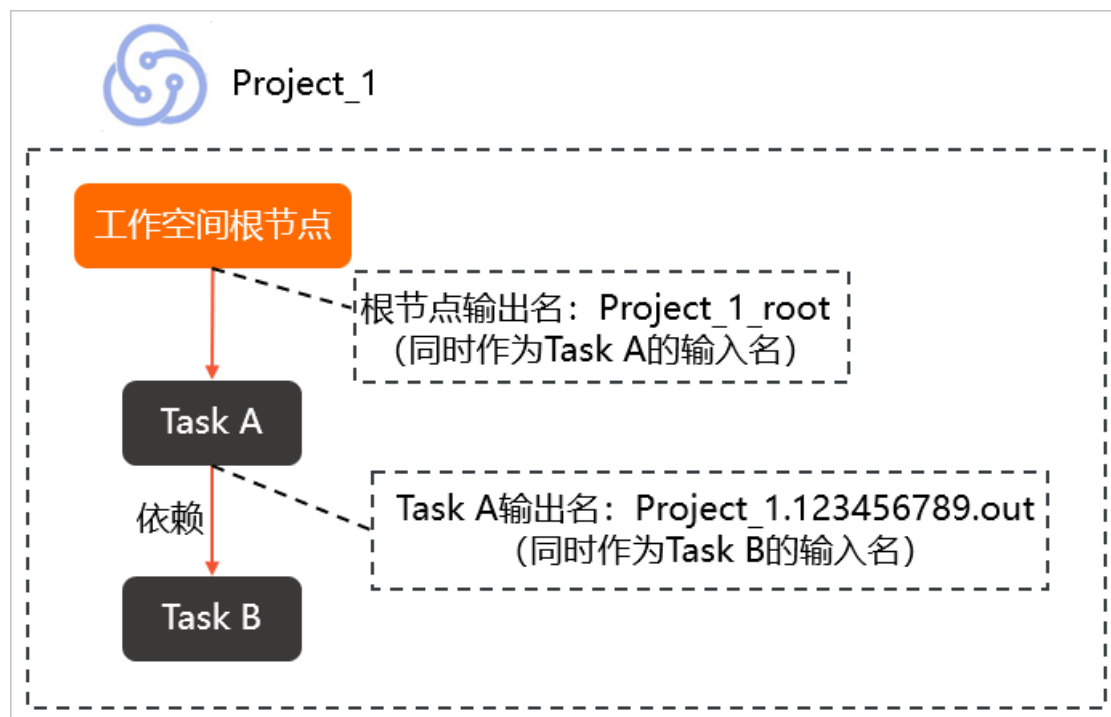
资源、函数

MaxCompute引擎使用的资源和函数，详情请参见[资源](#)和[函数](#)。

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。

元数据

元数据是数据的描述数据，可以为数据说明其属性（名称、大小、数据类型等），或结构（字段、类型、长度等），或其相关数据（位于何处、拥有者、产出任务、访问权限等）。DataWorks中元数据主要指库、表相关的信息，元数据管理对应的主要应用是[数据地图](#)。

补数据

完成周期任务的开发，将任务提交发布之后，任务会按照调度配置定时运行。如果您希望对历史时间段内的数据进行计算，您可以使用补数据功能。补数据操作生成的补数据实例将按照指定的业务日期运行。