ALIBABA CLOUD

阿里云

DataWorks

产品白皮书

文档版本：20201013

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.序言

根据2018年12月中国信通院发布的《数据资产管理实践白皮书3.0》：近年来，中国大数据和商业数据分析市场规模增速是世界平均增速的2倍，特别是银行业、离散制造业、流程制造业和政府等行业需求尤其旺盛。

然而，如此可观的增速背后却藏匿着较大隐患：

- 数据质量不过关：大部分企业数据质量较差，不良的数据质量会使企业额外花费15%到25%的成本。
- 数据难互通：绝大多数的企业存在数据孤岛问题，数据分析人员80%的精力都花在了获取数据的工作上，获取数据的效率非常低下。
- 数据安全难保证：自2013年以来全球数据泄露高达130亿条，其原因多半为管理制度不完善。
- 数据价值有待释放：大部分企业未建立有效管理和应用数据的模式，自身无法随着技术、市场、产业的变化不断迭代，更无法使数据资产管理成为持续和动态的过程，进而为数字化转型提供源源不断的动力。

因此，为了最大限度挖掘数据价值，企业的数据体系改革迫在眉睫：

- 管理方面，需要建立一套符合数据驱动的组织管理制度和流程。
- 技术方面，需要建设现代化数据平台、引入智能化技术，确保数据资产管理系统平台持续、健康地为数据资产管理体系服务。

2.发展背景

本文为您介绍DataWorks的获奖经历和发展历程。

DataWorks（数据工场）是新一代端到端的，具备数据集成与开发、生产运维调度、离线与实时分析、数据质量治理与资产管理、安全防护、数据共享与服务、机器学习和数据应用搭建等全栈数据研发能力的大数据平台。

获奖经历

2017年6月：DataWorks荣获中国国际软博会金奖。

2018年3月：Forrester Cloud Data Warehouse Q1 2018报告DataWorks与MaxCompute的组合为唯一入选的中国产品。

2018年5月：DataWorks荣获数博会十佳大数据案例。

2018年11月：Forrester Wave™ Q4 2018云数仓研究报告DataWorks成功入选竞争者（Contender）梯队，并领跑竞争者分组。

2018年12月：中国数字化转型与创新评选DataWorks为年度大数据创新产品。

发展历程

2009年：阿里云写下飞天第一行代码。

2009年8月：**DataWorks调度系统**正式上线（核心功能）。

2010年10月：统一淘宝内部计算任务调度。

2010年10月：核心功能 **DataX**上线（**数据集成**前身）。

2011年10月：在云端（DataWorks前身）发布，引领统一大数据平台的趋势。

2013年12月：**数据质量**DQC发布（核心功能）。

2014年12月：核心功能同步中心（**数据集成**前身）上线。

2015年1月：DataWorks对公共云用户输出。

2015年5月：DataWorks对专有云用户输出。

2017年4月：DataWorks开始国际化。

2018年10月：DataWorks V2.0版本发布，支持数据上云->数据研发->数据治理->数据分享的一站式数据研发体验。

2019年12月：DataWorks V3.0版本发布，支持跨地域任务依赖调度，并全新升级了多引擎架构，在MaxCompute的基础上，新增开源大数据引擎E-MapReduce、实时计算（Flink）、交互式分析和图计算服务（Graph Compute）等引擎服务。

3.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.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。

强大的基础能力

- 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/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。

大幅提升工作效率

非技术人员1~2小时即可掌握完整的数据研发流程，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。

直观的数据管理和权限管控方案

Web化的多租户模型确保用户数据被安全隔离，实现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。

数据准时产出有保障

仅需对最终产出任务配置单一基线，即可告别人工创建的成百上千的任务监控配置，实现对庞大任务流的全链路智能监控。

数据治理不再难

提供数据质量DQC、安全中心模块，将数据质量监控、数据审计等治理流程产品化，让您无需投入沟通成本、研发成本即可落地行业数据治理标准方案。帮助您的企业构建符合完整性、一致性、正确性、安全性的大数据体系。

零成本构建数据服务

告别接口开发、鉴权开发、API Server端运维的传统API构建方式，仅通过3步Web端配置即可快速构建数据HTTP API，助您以ServerLess的方式快速分享数据。

轻松构建数据应用

无需在本地搭建开发环境、长期维护程序运行Server，基于Web端的App Studio让您仅需几百行代码即可轻松构建属于您的数据智能应用。

一站式数据解决方案

基于DataWorks，您不仅可以体验到数据上云至应用开发的全生命周期开发流程，还可以快速构建企业级数仓和数据中台。DataWorks的一站式数据解决方案，助力企业实现数字化转型。

5.产品架构

5.1. 概述

DataWorks提供九个核心功能模块：以数据为基础，以全链路加工为核心，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，整体功能架构如下图所示。

5.2. 数据集成：全领域数据汇聚

数据集成（Data Integration）是提供了可跨异构数据存储系统能力、可靠、安全、低成本、可弹性扩展的数据同步平台。数据集成组合400多对数据源，提供不同网络环境下的全量/增量数据同步通道，具备可视化向导模式和脚本模式两种任务配置方式。

目前数据集成主要支持数据批量（离线）、增量两种同步方式，并提供整库迁移和批量上云等解决方案。数据集成提供的批量创建同步任务的快捷工具，能让您快速将单个或多个数据库内所有表批量上传到MaxCompute中，节省大量时间与人力成本。

数据集成与大数据开发套件深度融合，完全复用开发套件的调度能力和同步任务的监控、报警等运维能力。

5.3. 数据开发：数据融合加工

DataWorks的数据开发提供SQL格式化、智能补齐、关键字高亮、错误提示、SQL内部结构等人性化功能，从细节上带来更顺滑的SQL开发体验。

数据开发支持用户自行组合SQL、MR、Shell、实时计算和PAI机器学习各类任务的混编工作流，并实现分钟级调度、逻辑控制和上下游传参。此外，您可以从业务视角整体管理工作流，将同类业务组织为解决方案，实现沉浸式开发。数据开发提供开发、生产环境隔离的**标准工作空间模式**，从宏观上实现更稳定、更可靠的生产环境。

5.4. 数据分析：即时快速分析

数据分析功能为DataWorks本身的查询结果、本地Excel文件、MySQL等常见数据源提供快速灵活的透视分析能力。基于OLAP引擎Hologres，数据分析可以为您提供庞大数据量分析的加速能力。

数据分析功能支持维表编辑和转存、简单图表和结果分享，满足分析师对海量数据的日常分析需求。详情请参见概述。

在DataWorks中查询结果后，单击**数据分析**，即可跳转至**数据分析**页面。

5.5. 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

5.6. 智能监控：保证数据供给稳定

智能监控（Intelligent Monitor）是DataWorks运维监控中负责任务运行监控及分析的系统。

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。智能监控可以筛选关键依赖、对比实时关键路径，最终自动选择最合理的报警时间、报警方式以及报警对象，辅助您的决策优化。

> 说明 DataWorks基础版不支持智能监控中的基线管理和事件管理。

5.7. 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。

5.8. 数据资产管理（仅专有云）

业务系统及数据资源平台中存在大量的数据表、API等各类数据资产，数据管理者通过数据集成同步数据、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，提供对应的数据资产管理规范。

不同角色的用户可以在数据资产管理模块进行以下操作：

- 资产使用者可以快速查找、申请和使用资产。
- 管理者可以直观查看文件的详情，并上传和编辑文件。
- 资产使用者和管理者可以构建数据资产类目、归集数据表的业务信息。

有关DataWorks专有云的更多详情，请参见专有云文档。

5.9. 数据安全：为数据保驾护航

DataWorks的数据安全功能针对数据资产管理，提供数据识别、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。

数据安全针对用户权限管理，提供可视化申请审批流程，并可进行权限的审计和管理，提高您的数据安全等级，方便您进行数据权限管控。

5.10. 数据服务：低成本快速发布API

数据服务为企业搭建统一的数据服务总线，助您统一管理对内和对外的API服务。

数据服务支持您将现有的API快速注册到数据服务平台，统一管理和发布。通过与阿里云API网关（API Gateway）打通，数据服务还支持您将API服务一键发布至API网关。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，实现零运维成本。

在您通过图形化界面完成API发布后，数据服务还能自动为您生成API文档，并且实时更新文档。

5.11. 数据应用开发：轻松构建数据应用

DataWorks通过App Studio实现数据应用开发功能。

App Studio是一款数据开发工具。您无需下载、安装本地IDE和配置环境变量，只需一个浏览器即可编写、运行和调试应用程序，体验和本地IDE一样的编程效果，在线发布应用。App Studio提供了丰富的前端组件，通过自由拖拽，即可简单快速搭建前端应用。

5.12. 多租户模型：更直观的权限管控

DataWorks拥有自己的多租户权限模型。

- 针对计算和存储资源，租户可以按需申请资源配额，独立管理自己的资源。
- 针对数据，租户可以独立管理自有的数据、权限、用户、角色，实现与其他租户间彼此隔离，确保数据安全。

5.13. 工作空间管理（公共云）与平台管理（专有云）

工作空间作为代码管理、成员管理、角色和权限分配的基本单元，每个团队都可以具有独立的工作空间。

您只有在加入工作空间并被分配权限后，才具备查看代码、编辑代码、代码发布申请、代码审批执行等权限。

在工作空间管理（公共云）与平台管理（专有云）中，管理员可将相关用户加入到工作空间，并赋予DataWorks预设的项目管理员、开发、运维、部署、安全管理员和访客等角色，实现多角色协同开发。

6.应用场景

6.1. 低成本、高效率构建企业级数仓

本文将为您介绍通过DataWorks实现低成本、高效率构建企业级数据仓库的应用场景。

DataWorks所具备的可视化配置业务流程、周期自动调度、表分层管理等功能，符合构建企业级数仓时所面临的ETL任务分层管理、数据分层管理、分层定时产出数据的基础强需求。

客户案例—天弘基金

- 背景简介

  天弘基金是中国用户数最多的公募基金公司。由于传统数据架构无法适应交易量的快速增长、规模的飞速扩大以及任务之间多重依赖，所以选择使用DataWorks高效构建企业级数据仓库。

- 解决方案

  使用DataWorks作为数仓解决方案后，运维成本显著降低。同时，DataWorks的数据质量和数据安全能力符合金融行业的数据治理要求。

6.2. 轻松构建数据中台

基于DataWorks数据集成、开发、数据治理、数据分享等产品化的大数据全生命周期能力，您可以轻松构建数据中台，实现数据标准范围可控、数据可连接及可萃取。

DataWorks可以为您解决如何衡量数据资产价值、如何驱动业务、如何降低成本并提高效率、如何挖掘数据价值等企业最关心的问题。让数据植入业务并赋能业务数据化，构建可持续发展的商业化数据产品平台，达到人人都能进行数据化运营的目的。

联华华商

- 背景介绍

  联华华商集团是一家从事零售连锁业经营的企业。近年来由于业务扩张，面对复杂的业务形态与繁多的商品大类，如果仅依靠传统ERP积累的数据则无法找到改善用户体验、支持未来决策的依据，面对新零售浪潮的冲击，联华对数据化转型的需求十分迫切。

- 解决方案

  2018年，联华以DataWorks+MaxCompute为基础，借鉴阿里巴巴集团的大数据之路，构建了基于数据中台的新零售体系。通过构建数据中台，切实提升了业务系统效能，匹配业务特性，让业务脱胎换骨。同时，有效地提高了数据效率，助力数据化运营。

数据城市大脑

- 背景介绍

  2018年，某省的政府部门想要在交通、政务、亚运、平安等公共事业领域深入突破，便需要强大的数据资源、计算资源、人工智能作为统一支撑。在业务发展过程中，基于各类数据资源的共享和需求获取，需明确城市数据资源，并提供数据申请、获取、开发的完整机制和工具。

- 解决方案

基于DataWorks的数据共享、数据开放能力构建的数据城市大脑门户实现了城市数据资源的开放，对产业发展起到催生、带动作用，促进传统产业转型升级。

智慧城市

- 背景介绍

  某地区经过持续十多年的快速发展，人口规模、土地空间、社会基础设施、人力资源、商业模式，尤其是交通及公共服务等均受到巨大挑战，面临严峻的资源环境及人力资源压力。

  因此构建一个统一共享的数据资源平台、确立具有战略观和全局观的智慧城市整体规划和顶层设计、因地制宜推进智慧城市的整体布局和分段发展，具有重大而积极的指导意义。

- 解决方案

  数据是城市发展的新的基础资源，通过汇集各部门的数据，建立一个统一规范、安全可控、充分共享的城市数据资源平台，是实现数据资源跨区域、跨层级、跨部门的互联互通、融合共享的基础。

  以智慧交通、智慧医疗、智慧旅游、智慧政务为基础，归集相关政府部门的数据，逐渐构建成一个统一、共享的数据资源平台。基于DataWorks数据服务、数据资产管理的数据共享交换能力以及二次开发数据资源平台，可以实现数据在政府部门内的共享及面向市民的数据开放。