ALIBABA CLOUD

阿里云

智能数据构建与管理 Dataphin

产品简介

文档版本：20200922

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Dataphin

Dataphin（智能数据构建与管理）是一款用于大数据平台建设的智能引擎，旨在满足各行各业大数据建设、管理及应用需求。

视频介绍

功能特性

Dataphin遵循阿里巴巴集团多年实战沉淀的大数据建设OneData体系（OneModel、OneID、OneService），集产品、技术、方法论于一体，一站式地为您提供集数据引入、规范定义、数据建模研发、数据萃取、数据资产管理、数据服务等的全链路智能数据构建及管理服务。助您打造属于自己的标准统一、资产化、服务化和闭环自优化的智能数据体系，驱动创新。

功能	描述
平台管理	平台管理是Dataphin的基础功能，包含账号管理、系统设置和智能引擎。该功能帮助您系统地了解和熟悉整个产品，快速开始工作，并进行必要的系统管理与控制，保障各模块正常运转。
全局设计	基于业务全局，从顶层自下规划设计业务数据总线，包括划分命名空间、定义主题域及相关名词、划分管理单元（即项目）和定义数据源及计算引擎源。
数据引入	数据引入基于全局设计定义的项目空间与物理数据源，将各业务系统、各类型的数据抽取加载至目标数据库。这个过程可以实现各类业务数据的同步与集成，助您完成基础数据中心建设，为后续进一步加工数据奠定基础。
规范定义	基于全局设计定义的业务总线、数据引入构建的基础数据中心，根据业务数据需求，结构化地定义数据元素（例如维度、统计指标），保障数据无二义性地标准化、规范化生产。
建模研发	基于规范定义的数据元素，设计与构建可视化的数据模型。数据模型提交发布后，系统智能自动化地生成代码与调度任务，完成公共数据中心的全托管建设。
编码研发	基于通用的代码编辑页面，灵活地进行个性化的数据编码研发，完成任务发布。
资源及函数管理	• 支持管理各种资源包（例如JAR、文本文件），以满足部分数据处理需求。 • 支持查找与使用内置的系统函数。 • 支持用户自定义函数，以满足数据研发的特殊加工需求。
数据萃取	基于Dataphin数据建模研发沉淀的数据，萃取提供以目标对象为中心的数据打通和深度挖掘，并生成代码与调度任务，完成实体对象识别、连接及标签生产，可快速应用于各类业务。
调度运维	对建模研发、编码研发生成的代码任务进行基于策略的调度与运维，确保所有任务正常有序地运行。调度运维操作包括：部署数据生产任务、查看任务运行情况、管理及维护任务之间的依赖关系。
元数据中心	支持采集、解析和管理基础数据中心、公共数据中心、萃取数据中心的元数据。
资产分析	• 在元数据中心基础上，深度分析元数据，实现数据资产化管理。 • 为您可视化地呈现资产分布、元数据详情等，方便您快速查找、深度了解数据资产。

功能	描述
即席查询	支持用户通过自定义SQL等方式，查询数据资产中的数据。同时，通过查询分析引擎，快速获取物理表、逻辑表（即数据模型，或逻辑模型）的数据查询结果。
数据服务	数据服务为您提供高效便捷的主题式查询功能及有效的全链路企业内API生命周期托管，真正实现低门槛API开发，帮助您更好地进行数据资产应用以实现价值化。

更多功能模块详情，请参见Dataphin产品详情页。

功能详解：

- Dataphin快速入门
- 数据服务入门指导

为什么选择Dataphin

Dataphin可以屏蔽不同计算与存储环境的差异，助您快速引入数据并规范化地构建数据。您可以通过规范建模自动开发数据，萃取以实体对象为中心的标签数据体系，沉淀业务数据知识、数据资产，治理数据问题。同时，Dataphin还支持数据表查询、智能语音查询等多种类型的数据服务。

选择Dataphin，您可以轻松构建具有以下优势的数据体系：

- 数据规范统一：采用维度事实建模理论，对维度、维度属性、业务过程、指标字段等进行严格的标准化、规范化定义，保障数据质量，避免数据指标定义的二义性。
- 自动化编码：
  - 高效且自动化的编码：基于函数化理念，对通用数据计算逻辑进行组件化定义，并可自由组建统计指标，从而实现自助化建模研发，系统自动生成代码执行数据生产。
  - 智能计算优化：支持从业务视角进行逻辑建模。逻辑模型发布后，系统自动化进行物理建模、编码，从而降低对开发人员的技术能力依赖。
- 一站式研发体验：一站式地完成数据引入、建模、研发、运维、数据查找及探查等过程，研发链路统一且高效。
- 系统化构建数据目录：基于规范化建模、高效自动化的元数据抽取，以标准的技术框架系统地构建规范的业务化数据目录，形成数据资产地图，方便业务查找及应用。
- 高效的数据检索：基于元数据及业务数据构建数据图谱，实现快速、智能检索数据表及数据。
- 可视化的数据资产：系统化构建业务数据资产大图，从数据视角还原业务系统、提取业务数据，快速感知业务关键环节及数据。
- 数据使用简单可依赖：通过主题式数据查询服务，您可以快速查询和访问研发构建的数据逻辑表，简化约80%的查询代码。

同时，Dataphin可以为您提升构建数据体系的效率，降低成本：

- 提升效率：提供全链路、一站式、智能化的数据构建与管理工具，降低数据建设门槛。不同背景的开发人员可以自助ETL，快速满足业务需求。通过OneData（OneModel、OneID、OneService）方法论体系，可以完成模型和指标的抽象与自助定义、代码自动化生产、主题数据自动聚合并输出服务。
- 降低成本：以元数据为基础、算法智能为驱动，实现物理和逻辑分层的智能自动化生产。同时，分析数据资产全链路，优化计算及存储资源分配，从而降低数据生产及消费成本。

如果您想了解Dataphin是否适用于您的需求场景，请参见应用场景。

Dataphin定价

建议您先提供企业数据建设诉求及背景信息进行咨询，确认Dataphin功能及版本是否符合需求，再进行开通购买。

Dataphin支持按月购买的付费模式，关于Dataphin的计费标准请参见计费说明。

相关信息

您可以通过如下信息，进一步了解Dataphin：

- 了解阿里巴巴数据中台。
- Dataphin最佳实践：
  - 刚入职的数据分析师，如何使用1周的时间完成上千个数据指标的开发？
  - 如何通过Dataphin构建数据中台以实现新增100万用户？
  - Dataphin的代码自动化如何助力商业决策？
- Dataphin使用体验：
  - Dataphin体验系列-功能易用性及界面设计。
  - Dataphin体验系列-功能完备性。
- Dataphin功能详解：
  - Dataphin帮助企业构建数据中台系列-萃取数据中心。
  - Dataphin数据服务系列-API配置、管理和消费。
  - Dataphin支持哪些数据源？
  - Dataphin功能解读。

2.应用场景

本文将为您介绍Dataphin适用的典型应用场景。

智能构建云上数仓，提高战略决策效率

场景：某集团在全国经营多家连锁超市，线上线下零售渠道及形态众多。

痛点：因为业务系统多、数据来源多，经营所需的数据需求高频且多样化。但数据体系复杂、数据不统一，数据分析速度和数据准确一致性难保障，战略决策与数据化运营受阻。

解决方案：

- 数据融合：通过数据引入功能，将业务系统数据集成、融合一体，统一基础数据。
- 数据建模：通过规范建模功能，结合业务发展需求，自顶向下设计标准的数据模型，统一公共数据。
- 数据生产：基于建模后系统代码自动化托管生产功能，快速响应业务需求。模型设计输出后，自动化生成代码、周期性调度产出任务。

价值：

- 数据建设统一：数据标准规范定义。
- 数据研发提效：自动化代码生成。
- 战略决策高效：数据分析准确，数据需求响应及时。

推荐搭配组合：Dataphin + MaxCompute

MaxCompute详情请参见什么是MaxCompute。

输出主题式数据服务，提高数据化运营效率

场景：某公司是一家大型跨省直营餐饮品牌公司，具有线上线下多个客户触达渠道，以爆款思维策划公司品牌。

痛点：因业务扩张快，用户数据丰富，拉新留存效率、营销及转化效果急需提高。但各个获客渠道的用户数据分散，会员管理体系单一，推荐准确度不高，会员营销方式有限。

解决方案：

- 数据融合：通过数据引入功能，将各渠道数据沉淀至数据仓库内，丰富基础数据。
- 数据建模：通过数据建模及代码自动化生成功能，以会员为中心，构建完整的会员数据模型，集成会员属性、统计指标等数据。
- 主题服务：通过数仓即席查询功能，面向应用，自动输出会员主题的汇总数据模型，高效完成进一步的会员日报分析、会员门户搭建等。

价值：

- 数据建设统一：数据标准规范定义。
- 数据研发提效：自动化代码生成。
- 资产管理便利：数据丰富融通，主题化服务更智能。

推荐搭配组合：Dataphin + Quick BI + MaxCompute

- Quick BI详情请参见什么是Quick BI。
- MaxCompute详情请参见什么是MaxCompute。

3.基本概念

本文为您介绍Dataphin的基本概念。

业务板块

业务板块是逻辑空间的重要组成部分，是基于业务特征划分的命名空间。在同一个业务板块中可能包含多个不同的项目，业务板块与项目的关系为1:N。例如，某企业的业务涉及零售和文娱，且系统间相互独立，则零售和文娱就是两个业务板块。

统计周期

统计的时间范围。例如最近1天、最近30天等。

项目管理

项目是一种物理空间上的划分，便于用户在数据中台建设过程中对物理资源及开发人员进行隔离化管理。一个业务板块可以包含多个项目，每个系统成员可以加入多个不同的项目。

维度

人们观察事物的角度，是指一种视角，是确定事物的多方位、多角度、多层次的条件和概念。

业务过程

业务过程即业务活动中的所有事件。

维度逻辑表

丰富维度的属性信息形成维度逻辑表。通过维度逻辑表，设计及加工处理公共对象明细数据，便于提取业务中对象的明细数据。

事实逻辑表

用于描述业务过程的详细信息。通过创建事实逻辑表，设计及加工处理公共事务明细数据，便于提取业务中事务的明细数据。

业务限定

圈定统计的业务范围。

原子指标

对指标统计口径、具体算法的抽象。Dataphin创新性地提出了设计即开发的理念，指标定义同时也明确了设计统计口径（即计算逻辑），提升了研发效率，并保证了统计结果的一致性。例如支付金额。

派生指标

即基于原子指标、时间周期和维度，圈定业务统计范围并分析获取业务统计指标的数值。`派生指标=原子指标+业务限定+统计周期+维度（维度的组合）（统计粒度）`。

统计粒度

定义数据汇总的程度。例如，如果维度为时间，则统计粒度为年、季、月、周或日等。

4.使用限制

本文为您介绍Dataphin的使用限制。

为保障软件系统稳定，Dataphin有部分使用上的限制或建议，详情如下表所示。

功能项	操作项	使用限制/建议

管理中心	成员管理	• 超级管理员（即超管）是您购买Dataphin后，由系统初始化自动生成的角色。
		  ○ 您的阿里云主账号即为超级管理员账号，通过**访问控制**功能可以创建RAM用户，详情请参见创建RAM用户。
		  ○ 一个Dataphin系统只有一个超级管理员账号，拥有系统内所有权限。
		• 阿里云RAM账号体系下，如果要更新用户列表、用户信息，则需要进行如下操作：
		  i. 使用超级管理员账号登录管理控制台介绍，配置Access Key做授权。
		  ii. 使用超级管理员账号执行**账号系统同步**，即可获取阿里云主账号下的RAM子账号，并添加为Dataphin的成员。

管理中心	配置计算引擎	• 全局配置只支持在系统内计算引擎源为空的情况下，由超管更新。 • MaxCompute计算引擎类型下，Endpoint地址配置详情请参见配置Endpoint。
计算引擎类型	选择设置	计算引擎设置需要提前采购计算引擎MaxCompute资源，系统以此来支持相关数据的建设工作。 • 需要选择计算引擎类型（目前仅开放MaxCompute计算类型），配置计算引擎所在的集群，例如Endpoint等信息。系统以此来支持该计算引擎类型下、该集群上，相关数据的建设工作。请根据您的计算引擎的集群情况选择设置。 • 如果您需要用到Spark on MaxCompute做计算，建议提交工单咨询，确认您的MaxCompute计算引擎所在的地域（Region）是否开通Spark服务。如果该地域未开通Spark服务，您的Spark任务将无法成功执行。 • 请您以Dataphin为唯一入口进行数据构建与管理，以免出现元数据错误、权限异常等问题。 **说明** 创建MaxCompute项目的具体方法请参见创建项目空间。
数据源管理	新增数据源	• 建议配置的数据源Access Key为管理级权限。可以通过配置主账号Access Key，或者给子账号Access Key授予MaxCompute所有权限来实现。 • 不建议将同一个物理数据库（配置完全相同）配为两个数据源。

功能项	操作项	使用限制/建议
项目管理	项目名称	• 建议当配置数据源为MaxCompute类型时，项目英文名必须与MaxCompute的Project英文名一致。 • 项目名不可以 `LD_` / `ld_` 开始，以免与业务板块名冲突，导致查询功能不可用。
项目管理	配置计算引擎源	• 对于已配置为项目数据源的物理数据库，不建议再从其他非Dataphin控制台进行数据的增、删、改操作。 • 不建议您为项目配置跨集群的计算引擎源。
研发工作台	规范建模	• 建议您谨慎命名规范定义和逻辑表对象的英文名，且推荐使用小写英文字母命名，以免因下游依赖约束导致英文名不可改且不易读。 • 请尽可能使用英文缩写，以免字段名称超出数据库限制，导致数据生产出错。
研发工作台	数据处理	• 不支持项目所属的计算引擎源在跨集群的情况下读取数据。 • 非Dataphin创建的表，Dataphin中元数据可能无法获取或者更新相关信息。
研发工作台	即席查询	逻辑表查询时，必须使用业务板块的英文名作为前缀。跨项目物理表使用时，必须使用项目的英文名作为前缀。 如果您需要查询开发环境数据，请在生产名称后加上_dev，系统会自动将生产业务板块、生产项目生成对应的变量。例如，您拥有业务板块LD_Trade，则系统自动生成业务板块变量${LD_Trade}。该变量在开发环境执行时默认被替换为LD_Trade_dev，在生产环境执行时默认被替换为LD_Trade。您也可以在执行时设置固定的值，提高代码在不同环境执行时的灵活性。