ALIBABA CLOUD

阿里云

专有云企业版

智能数据构建与管理 Dataphin

技术白皮书

产品版本：V3.14.0

文档版本：20210429

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 ---- 05
1.1. Dataphin简介 ---- 05
1.2. 功能概览 ---- 05
1.3. 产品优势 ---- 06
2.技术优势 ---- 07
3.产品架构 ---- 08
3.1. 系统架构 ---- 08
3.2. 技术架构 ---- 08
4.功能原理 ---- 10
4.1. 平台管理 ---- 10
4.2. 全局设计 ---- 10
4.3. 数据引入 ---- 11
4.4. 规范定义 ---- 11
4.5. 建模研发 ---- 12
4.6. 编码研发 ---- 13
4.7. 资源及函数管理 ---- 14
4.8. 调度运维 ---- 15
4.9. 元数据中心 ---- 16
4.10. 资产分析 ---- 16
4.11. 安全管理 ---- 17
4.12. 即席查询 ---- 18

1.什么是Dataphin

1.1. Dataphin简介

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。

通过输出阿里巴巴集团多年实战沉淀的大数据建设体系（OneData、OneEntity、OneService），集产品、技术、方法论于一体，一站式为您提供集数据引入、规范定义、数据建模研发、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。助力政府机构和企业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以驱动创新。

我们致力于屏蔽不同计算与存储环境差异，帮助用户快速引入数据、标准规范化构建数据、通过建模化方式自动开发数据、萃取以实体对象为中心的标签数据体系、沉淀业务数据知识与数据资产、治理数据问题。同时还支持数据表查询、智能语音查询等多种类型数据服务。

1.2. 功能概览

本文为您介绍Dataphin的模块及功能。

平台功能

支持对整个产品系统地了解熟悉、全局化功能设置。帮助您学习使用产品功能、快速开始工作，以及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

支持从业务全局出发，从顶层自下规划设计对应的业务数据总线，包括命名空间划分、主题域及相关名词定义、管理单元（即项目）划分、数据源及计算源定义。

数据引入

基于全局设计定义的项目空间与物理数据源，支持将各业务系统、各类型的数据抽取并加载至目标的大数据存储，完成数据的同步与集成，完成各垂直业务数据集成后的垂直数据中心的建设。同时也为后续数据的进一步加工处理准备了基础。

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，支持根据业务数据需求，结构化地定义、组件化地构建数据对象（维度、统计指标等），以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，支持可视化地设计与构建数据模型。提交发布后由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生产。

编码研发

基于通用的代码编辑界面，支持自由灵活地进行个性化的数据编码研发，并完成对应任务发布。

资源及函数管理

支持各种资源包（如JAR、文档文件）管理以满足部分数据处理需求，支持原生的系统函数查找与使用，支持自定义函数以满足数据研发特殊的函数加工需求。

调度运维

支持对建模研发、编码研发、数据萃取生成的代码任务进行基于策略的调度与运维，包括数据生产任务部署、任务运行及依赖情况查看并管理维护，以确保所有任务正常有序地运行。

元数据中心

支持采集、解析、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，支持元数据深度分析并实现资产化管理数据，以可视化地呈现资产分布、元数据详情等，以便捷查找及深度了解数据资产。

即席查询

支持自定义SQL等方式查询数据资产中的数据，并通过查询分析引擎快速实现物理表及面向主题的逻辑表（也即数据模型，或逻辑模型）数据查询及结果获取。

1.3. 产品优势

本文为您介绍Dataphin产品优势。

Dataphin包含以下优势：

- 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。
- 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组件化定义并可自由组建统计指标，从而自助地实现建模研发、系统自动生成代码执行生产数据。
- 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：数据引入、建模、研发、运维、数据查找及探查等过程一气呵成，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可读的业务化数据目录，形成数据资产地图，方便业务查找及应用。
- 高效数据检索：基于元数据及数据构建数据图谱，实现数据表及数据简单且快速的智能检索。
- 可视化数据资产：系统化构建业务数据资产大图，数据视角还原业务系统、提取业务数据知识，并可快速感知业务关键环节及数据。
- 数据使用简单可依赖：定义即服务，研发构建的业务主题式数据逻辑表可被直接、快速地查询和访问，可简化约80%的查询代码。
- 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自助ETL并快速满足数据需求。其中贯穿的OneData、OneEntity、OneService思想与方法论（已申请专利）可完成模型&指标抽象与自助定义、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、数据资产全链路分析追踪与优化，优化计算及存储资源分配，从而降低数据生产及消费成本。

2.技术优势

本文为您介绍Dataphin的技术优势，帮助您更全面地了解Dataphin。

- **数据规范统一，高效且自动化的编码**
  - 标准建模：基于维度事实建模理论，结合函数化理念进一步拓展，实现对通用数据的计算逻辑进行组件化定义。定义出的标准化数据组件，可以进行逻辑上的组合，生成逻辑模型及其计算逻辑代码（即逻辑建模）。
  - 自动优化代码：基于计算和存储最优的规则，对逻辑模型的实际物理表进行水平拆分及物理拆分，优化用户输入的计算逻辑组合，生成实际执行的物理代码。
  - 自动调度执行：基于对物理代码的语义识别，结合用户手动配置的调度信息，系统会自动生成最优调度拓扑图，实现最佳执行时序，保障数据产出。
  - 关键指标：秒级建模，模型提交后分钟级代码优化生成。支持四种维度模型（普通、层级、枚举、虚拟）、两种事实模型（事务型、周期快照型）、两种指标（原生原子指标的统计指标、衍生原子指标的统计指标）。支持雪花模型及星型模型，支持模型、字段及指标并发运行。
- **系统化构建数据目录**
  - 提取全局元数据：自动化抽取建模及物理代码元数据。
  - 标准化解析元数据：基于数据资产模型，系统自动加工更新数据的元数据，提取数据资产。
  - 可视化呈现：基于全局、流动、结构三个视角，为用户可视化呈现数据资产信息，部分还原及解读业务系统的数据场景。
- **数据使用简单可依赖**
  - 逻辑模型查询：支持基于*[逻辑模型.关联维度.关联维度.….属性]*（例如，订单.买家.会员类型.类型值）查询所需数据，使常见分析类SQL的编写长度简化60%左右。
  - 优化查询SQL：逻辑模型查询转换为物理执行SQL时，基于计算最优规则，生成最终执行SQL。这种方式相较于单纯的物理表查询，提升了执行效率、降低了资源耗费。
- **一站式研发体验**
  - 数据引入、建模、研发、运维、数据查找及探查等开发过程一体化，研发链路统一而高效。
  - 支持编辑器智能提示函数、表等信息。
  - 支持100人在线协同工作，支持字段粒度、表粒度数据权限控制，并提供基于角色、审批等方式的权限管理。
- **高效调度**：支持百万级任务调度、小时调度。支持解析用户的资源设置，自动分配资源量。
- **集成异构数据源**：提供多种异构数据源的数据读取和写入能力，并提供脏数据过滤、流量控制等功能。
- **兼容多计算引擎**：支持MaxCompute、Hadoop CDH5.11.2（网络及元数据互通）两种计算引擎类型。
- **深度萃取数据价值**（暂未上线）：简单定义对象、设置计算参数，即可快速完成ID识别与连接、ID之上的标签体系构建，尤其是以“人”为中心的ID及标签数据连接，从而降低了营销DMP构建的门槛。

3.产品架构

3.1. 系统架构

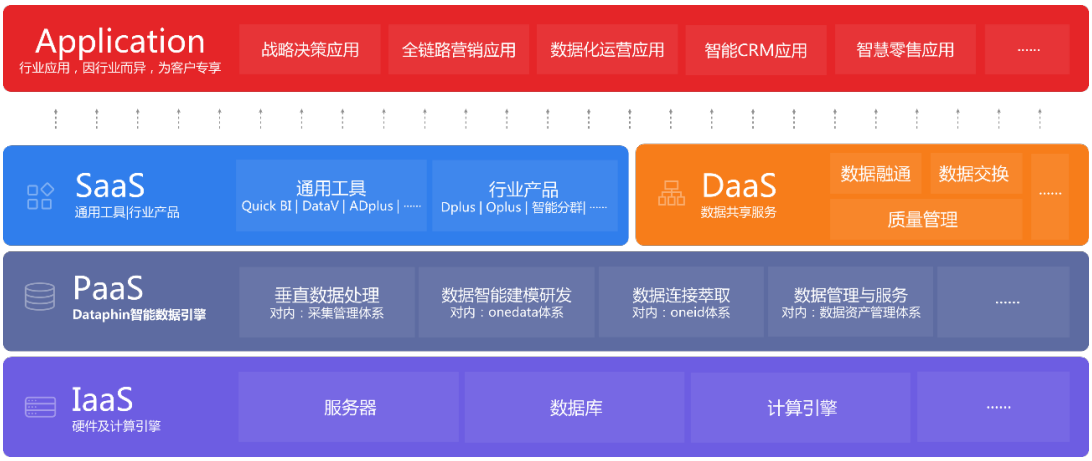
本文为您介绍Dataphin的系统架构和业务架构。

Dataphin在业务系统中的位置，请参见系统架构。

系统架构

对于业务系统而言，配套的业务平台成为数据平台的PaaS基础层，可以快速地输出易用的数据服务来支撑上层多样化的数据产品，从而让业务数据化运营。作为数据平台的基础层，向下可兼容不同的硬件设施，向上可对接各类应用产品，从而构建出从IaaS到SaaS的数据通路，输出面向业务的规范标准、连接融合和可管理查询的数据，请参见业务系统。

业务系统

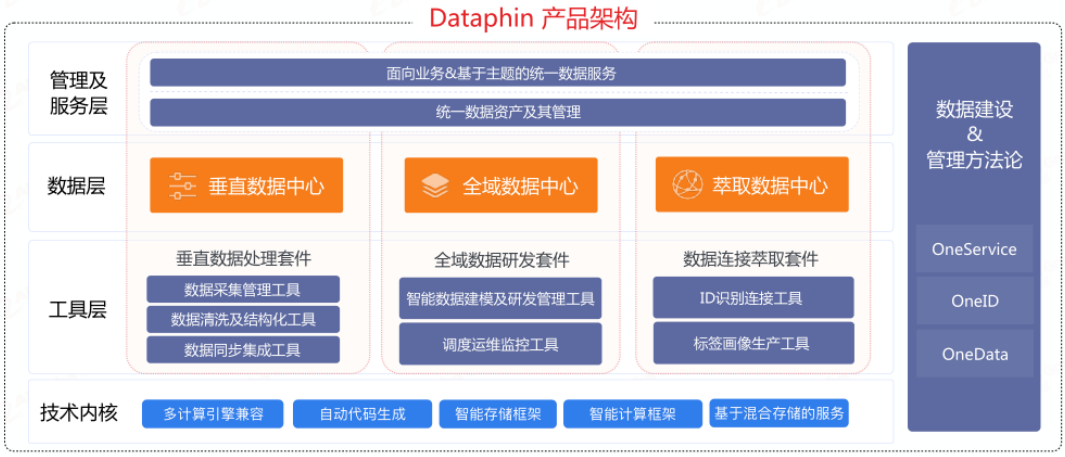


3.2. 技术架构

本文为您介绍Dataphin的产品架构和数据流。

Dataphin产品及架构，以及架构间的关系，详情请参见产品架构。

产品架构



基于OneData、OneID、OneService的数据建设与管理方法论，Dataphin系统由四部分组成，它们分别是**技术内核**、**工具层**、**数据层**和**管理服务层**。通过这四部分的协同合作，最终形成如下可控的数据流向，详情请参见数据流。

数据流

- **技术内核**：一套屏蔽底层计算、存储、软件系统差异的技术框架，保证数据研发可兼容多计算引擎与计算时效，代码自动化生成并实现智能存储与计算，数据服务支持混合存储等。
- **工具层**：面向开发者的数据构建与管理的工具，包括基础数据的标准规范及集成引入，公共数据的标准规范定义、智能建模研发、调度运维及机器学习，萃取数据的ID识别连接及标签生产。
- **数据层**：在技术内核基础上，通过工具加工生产，输出三种层次的结构化数据，构建出高保真、面向各业务的基础数据中心，模型化、面向主题的公共数据中心，深度加工、以实体为中心的萃取数据中心。
- **管理及服务层**：将数据及数据服务以资产化视角进行管理，以支持数据研发人员及业务人员都可以获取高质量且统一的数据资产；从业务视角将已有数据包装加工为主题式的数据服务，以保障业务可以统一地查询与调用数据。

4.功能原理

4.1. 平台管理

平台管理是Dataphin的基础功能，以保证系统中所有用户可控、有序、顺利地进行数据研发。本模块主要包括必要的全局功能设置，即账号管理、计算管理、首页引导及语言支持，以保证超级管理员用户能够把握平台全局，其他用户能够快速进入目标模块。

账号管理

从系统功能使用安全的角度进行用户账号控制。基于企业已有账号系统识别和配置产品的用户范围，由最高权限的用户管理其他用户账号及权限。

计算管理

- 作为PaaS层的平台产品，从系统计算稳定和统一的角度进行计算类型、底层硬件的设置与管理，以兼容IaaS层计算引擎，实现不同环境下的数据建设工作。
- 支持MaxCompute、Hadoop两类主流计算引擎，并支持对这两类引擎元数据的自动采集与解析。元数据采集部署初始化详情请参见元数据中心。

首页引导

- Dataphin首页为您提供数据构建与管理的入口、调度运维和项目空间的全局指标以及快速入口。
- Dataphin首页展示了数据生产、管理与服务的全流程，便于用户在正式开始工作前系统地了解学习产品功能，并根据需求快速进入相应的功能模块。

国际化-语言支持

Dataphin会根据系统的语言版本识别并选择默认语言（包括中文、English），方便不同国家地区的用户使用。

4.2. 全局设计

从业务全局出发进行数据架构的顶层设计，是数据建设工作中奠基的一步。保证数据的管理可控，数据研发、萃取、管理时定义及设计的数据体系满足中长期的业务需求，业务获取的数据与服务统一、面向主题、易用。

全局设计包括以下内容：

- 基于业务特征划分业务总线——业务板块维护及权限控制、数据域定义维护及权限控制、公共定义的全局性统计周期设置与管理。
- 基于数据独立管理及开发协作需求划分项目空间——项目基本信息及计算资源配置管理、成员管理。
- 基于项目计算资源及业务数据需求定义数据来源——数据源的配置管理。

业务总线

业务总线是基于业务特征以划分定义逻辑意义上的命名空间、主题分类、名词术语，以实现从管理顶层设计、控制建设过程中的数据定义标准化。

项目空间

项目空间是基于数据研发与管理团队对数据研发项目独立管理、对数据资源质量等高效管理的需求，为实现资源隔离、用户成员分组、数据建设约束条件配置等，而定义的物理命名空间。

物理数据源

支持数据源创建、修改等操作以注册及注销所需数据库，支持的数据源类型包括MaxCompute、MySQL、SQL Server、PostgreSQL等。数据源一方面作为数据同步传输的来源或者目标，另一方面一些特殊数据源类型（如MaxCompute）可作为对应计算引擎类型设置后项目计算存储基座。

4.3. 数据引入

本文为您介绍如何将数据引入Dataphin平台。

数据引入是基于企业数据全局架构中基础数据层设计，选定所需的业务数据存储，并根据存储及数据时效、数据质量等需求，制定的数据同步、清洗、结构化策略。

作为数据建设的初始化环节，数据同步套件基于阿里巴巴在业务数据、日志数据等多种类型数据交换同步方面多年实践的沉淀，可实现业务原始数据高效地录入。并通过管道对传输元数据的采集与统计能力，对数据传输量和数据内容方面，可支持简单规则校验及数据容错自定义机制等灵活管控，实现数据高质的同步。

数据源配置

数据源配置管理模块，支持多数据源的接入与管理，清单列表的展示方式可一目了然管控已接入的数据源，同时支持多种类型的数据源新增。数据同步中心目前支持数据源包括MaxCompute、MySQL、SQLServer、PostgreSQL、Hive等。

数据同步

数据同步模块，支持选择来源数据与目标数据，配置相关参数实现增量或全量同步，确定源数据与目标数据字段的映射关系，同时支持传输流量、并发数等的配置，并生成相关任务节点发布后进行调度。

4.4. 规范定义

本文为您介绍如何规范定义维度、业务过程、原子指标、业务限定和派生指标。

概述

在传统的数据研发过程中，数据建模以及指标定义等具体而重要的数据建设与研发工作，很多情况下依赖于研发人员的专业能力，命名方式没有统一的标准。仅基于个性且变动的文档，进行研发工作规范及设计的传递说明，非常容易造成不同指标命名冲突或者重复计算等一系列问题。

Dataphin基于OneData方法论，标准规范化维度、业务过程、指标定义等重要数据元素的定义，保证口径、算法、命名等的唯一性，从数据设计顶层杜绝指标二义性的产生。Dataphin可以帮助您基于表单式操作，便捷、批量地创建指标，降低数据研发门槛，还可以快速赋能有基本数据分析能力的业务人员，大大提升了研发效率。

规范定义包括维度、业务过程、原子指标、业务限定、派生指标共五个模块。

维度

- 维度在业务板块内唯一，并唯一归属于一个数据域，实现命名与主题分类的归一与规范化。
- Dataphin支持主子维度关系定义，以统一维度对象、归一维度特征。
- Dataphin支持不同类型的维度定义，包括枚举维度、虚拟维度、普通（层级）维度、普通维度。
- Dataphin支持查看和管理业务板块、项目两个范围内已有的维度清单，也支持对单个维度的快速查看与编辑。

业务过程

业务过程是指在业务中发生的最小单元的行为或事务，例如创建订单，浏览网页。业务过程产生的行为明细（例如支付了一笔订单，浏览了某个网页），最终都会汇集到事实表中。大部分情况下，事实表都会聚焦于某个特定的业务过程。

与维度的意义相同，业务过程是OneData方法论的顶层设计概念之一，与维度共同明确数据建设的架构。Dataphin支持业务过程的规范定义，不仅可以看到组织的业务全貌，还可以通过业务过程方便地对事实表进行分类管理。

为保障后续事实模型统一、标准、规范地构建，业务过程在业务板块内唯一，并唯一归属于一个数据域，实现命名与主题分类的归一与规范化。

Dataphin支持查看和管理业务板块、项目两个范围内已有的业务过程清单，也支持对单个业务过程的快速查看与编辑。

原子指标

原子指标是对指标统计口径、具体算法的一个抽象。为了从根源上解决定义、研发不一致的问题，Dataphin创新性地提出了**设计即开发**的理念。指标定义同时即明确设计统计口径（即计算逻辑），不需要ETL二次或者重复研发，提升了研发效率，也保证了统计结果的一致性。

根据计算逻辑复杂性，Dataphin将原子指标分为两种：

- 原生的原子指标，例如支付金额。
- 衍生原子指标-基于原子指标组合构建，例如客单价为支付金额除以买家数。

为保障所有统计指标统一、标准、规范地构建，原子指标在业务板块内唯一，并唯一归属于一个来源逻辑表，计算逻辑也以该来源逻辑表模型的字段为准，进行确定性定义。每个原子指标取来源逻辑表模型相关的所有逻辑表追溯所属分类，所以原子指标可能归属于多个数据域，实现命名、逻辑归一，主题分类规范化、标准化、系统化。

业务限定

原子指标是计算逻辑的标准化定义，业务限定则是条件限制的标准化定义。同原子指标，为保障所有统计指标统一、标准、规范地构建，业务限定在业务板块内唯一，并唯一归属于一个来源逻辑表，计算逻辑也以该来源逻辑表模型的字段为准，进行确定性定义。每个业务限定取来源逻辑表模型相关的所有逻辑表追溯所属分类，所以业务限定可能归属于多个数据域，从而实现命名、逻辑归一，主题分类规范化、标准化、系统化。

派生指标

派生指标即常见的统计指标。为保证统计指标标准、规范、无二义性地生成，OneData方法论抽象为如下四部分：

- 原子指标：明确统计口径，即计算逻辑。
- 业务限定：统计的业务范围，筛选出符合业务规则的记录。
- 统计周期：统计的时间范围，例如最近一天，最近30天等。
- 统计粒度：统计分析的对象或视角，定义数据需要汇总的程度，可以理解为聚合运算时的分组条件（即SQL中的group by）。粒度是维度的一个组合。例如，某个指标是某个卖家在某个省份的成交额，那么粒度就是卖家、地区这两个维度的组合。

基于上述定义组合机制，即可完成派生指标批量、无重复地快速创建，且保证概念定义、计算逻辑明确而不重复，业务人员也可自主定义生产。派生指标作为与字段同级的概念，在同一个统计粒度内唯一，保证对象组合的统计数据定义唯一与确定。

4.5. 建模研发

Dataphin提供体系化、系统化建模及研发的功能，将数据仓库理论以工具化方式深度应用实现：自顶向下快速构建业务维度、业务过程，并进一步细化开发维度表、事实表、汇总表、应用层，沉淀出标准统一的数据资产，便于业务分层数据应用，同时优化计算存储。

维度逻辑表

维度逻辑表是对维度逻辑模型的细化，包含维度的详细信息。支持对已有维度逻辑表的清单查看和管理，对单张维度逻辑表的可视化查看及编辑。

事实逻辑表

Dataphin支持用事实逻辑表来描述特定过程（即业务过程，如下单、支付等）或者状态度量（如账户余额、库存量等）属性的数据仓库模型。事实逻辑表的构建是基于优化的类雪花模型完成的，即：允许将除度量和关联维度外的属性信息退化至事实表，作为事实属性，并可以进行归类，以降低模型设计复杂度、提升用户使用友好度。

汇总逻辑表

汇总逻辑表是重要的数据仓库模型之一，模型内涉及两类元素：一类是描述某种统计粒度（由N个维度组成，N>=0，如：省份+产品线）的各种统计值（即派生指标，如最近7天销售额），另一类是组成统计粒度的各个维度（如：省份、产品线）的属性信息（如：省份名称、产品线名称、产品线等级等）。

代码自动化

维度逻辑表、事实逻辑表、汇总逻辑表提交发布后将自动完成物理模型设计、自动完成相关代码、自动生成对应调度任务（一个逻辑表一般会涉及多个任务）以生产所需数据，可在调度运维中查看了解其执行逻辑。

4.6. 编码研发

编码研发是Dataphin数据处理研发过程与建模研发并行的重要数据研发方式之一，提供用户基于计算引擎的代码编写方式编辑脚本文件，并可提交至调度系统进行任务生产，同时支持追溯查看节点版本，从而完成通用数据开发。

脚本文件会有多种类型（包括SQL、Shell、MapReduce等），对应不同的编写配置要求（如代码的语法特征、调度配置要求），提交发布成功会生成对应任务进行运行、生产数据，在调度运维的DAG图中又被称为节点。编码研发的核心功能点包括：代码文件管理（增删改查）、代码编辑、任务调度配置及发布、节点版本管理。

代码编辑器

代码编辑器提供在线代码编辑界面以完成数据开发任务。支持SQL编程、MR编程、Spark编程、Shell编程。

任务调度配置及发布

- **调度配置**

  支持手动和周期性任务的调度配置， 可对已经完成配置的任务进行发布。系统支持自动检测任务调度配置的完整性，对配置完整的任务才会放行然后进行发布，所有发布的任务会生成周期性任务在**调度运维 > 周期性任务**列表中展示。

- **提交发布**

  有权限控制地允许项目成员操作，仅参数配置全面、依赖关系内容有效、非循环依赖的调度可以完成提交发布，生成调度任务，保证数据可以如期、稳定、有序生产。

代码管理

支持代码文件的新增、删除、更新、重命名、文件夹分类管理、内容编辑与查看等操作，方便代码文件整理、使用。

- **文件管理**

  支持单文件编辑、删除、下线、重命名操作管理之余，可查看单代码文件的发布状态、创建人和创建时间。这样，可实现整个大数据编码研发的文件的便捷创建、清晰查看与系统化管理。

- **文件夹管理**

  代码文件过多时需要进行分类管理，保证文件有序分布与展示，支持新建、重命名、删除文件夹，历史文件和新建的代码文件可被操作移动至对应文件夹目录内进行管理，支持多级文件夹管理。

团队协作编程

- **节点版本管理**

  支持对任务节点版本的历史追溯，可查看版本号、提交人、提交时间、提交备注，并可以查看各版本的代码详情，以比较差异性。目前面向Maxcompute_SQL、ODPS MR、Shell等节点类型。

- **协作开发**

  为了支持多人同时编辑协作，提高开发效率，提供脚本文件的锁机制，以有效保障协同编辑与冲突解决，即保证同一个代码在同一时间内只能被一个用户所编辑，用户通过获取锁的方式来得到编辑权限，被偷锁者也可获取偷锁进行，进行相应处理操作。

4.7. 资源及函数管理

资源及函数管理是编码研发中重要的辅助功能，数据开发者可以上传本地的资源，在节点运行时调用以实现特殊的加工诉求，或者直接使用计算引擎支持的脚本语言所对应系统自带函数，实现常见的数据加工处理。特别地，对于较高频率处理某类数据逻辑且无法用系统内建函数解决（如：按照某种业务逻辑进行数据转换）时，可基于上传的资源注册自定义函数实现。

资源管理

支持当前项目内数据开发者进行资源的新增、修改等操作。通过新建、编辑，可完成资源文件的命名、上传，然后复制引用至代码中，也可以手工删除不需要的资源文件。

- 资源新建与上传

  支持上传的本地资源文件类型包括如下，新增文件类型在标准接口下支持3日快速扩展，资源名称在项目空间中唯一、提交成功后文件名及上传的资源包不可改、上传仅支持单个文件并提示文件与所选类型保持一致。

- 资源引用

  通过copy引用即可复制并粘贴本资源至所需应用的代码编辑框的对应代码位置，编写相关执行语句调用该资源。

- 资源更新

  支持对已管理的资源进行描述信息的编辑更新，也支持删除已有资源，达到空间合理利用的目的。

函数管理

函数管理可以满足函数的查找、使用、管理。函数分为两类：一类是系统默认的内建函数，另一类是基于已上传的jar或python等资源的自定义函数。自定义函数支持在引用标准化的函数基础上进行扩展。

- 新建自定义函数

  自定义函数要求其名称在项目空间中唯一，且注册后不可修改名称。

- 函数引用

  系统内建函数或自定义函数，支持copy引用即可复制并粘贴本函数名至所需应用的代码编辑框的对应代码位置，以示例的命令格式编写相关执行语句进行加工处理。

- 函数更新

  支持自定义函数的更新，即：可以编辑变更函数的相关信息（名称除外），也可以删除注销不再需要的自定义函数。

4.8. 调度运维

本文为您介绍调度运维模块的任务列表和任务运维。

调度运维子产品作为数据研发工作后期环节的日常化维护控制部分，提供所有数据处理任务（周期性任务及手动任务）的清单、任务依赖关系DAG图、实际任务（周期性任务、手动任务、补数据任务）运行的实例清单、运行实例的依赖关系及状态DAG图，可实现任务执行时序安排、执行进程拆分、执行机器资源最佳分配、异常任务发现，保证所有任务能如期、稳定、可靠地执行并生产出数据，如有运行问题可及时管控。目前调度运维的功能模块包括任务列表和任务运维2大部分。

任务列表

任务列表提供不同项目下的周期性任务及手动任务的任务清单列表及任务依赖关系DAG图。

周期性任务

周期性任务，支持任务清单、任务查找、单任务的依赖关系查看。支持切换项目查看和查找任务，支持以任务节点名称和节点ID的模糊匹配搜索查找任务，支持对指定人员的任务节点和今天发布的任务节点的二次筛选，可对任务进行缩小范围或精确定位进行运维。

手动任务

支持任务清单、任务查找、单任务的详细情况查看。支持切换项目查看和查找任务，支持以任务节点名称和节点ID的模糊匹配搜索查找任务，支持对指定人员的任务节点和今天发布的任务节点的二次筛选，可对任务进行缩小范围或精确定位进行运维。

实例运维

提供不同项目下的周期性任务实例、手动任务实例、补数据实例的清单列表及任务实例运行详情。

周期性实例

支持实例清单查看、实例查找、单实例的详情查看。可查看所有常规实例的运行状态、对应任务的唯一性编号NodeID、节点名称、任务的负责人、任务运行开始时间和结束时间及时长等信息。支持切换项目查看和查找任务实例，支持以任务节点名称和节点ID的模糊匹配搜索查找任务，支持对我的实例、出错实例、未完成节点的二次筛选，同时也支持对运行日期进行二次筛选，可实现对实例缩小范围或精确定位进行运维。

手动实例

支持实例清单查看、实例查找、单实例的详情查看。可查看所有手动实例的运行状态、对应任务的唯一性编号NodeID、节点名称、任务的负责人、任务运行开始时间和结束时间及时长等信息。支持切换项目查看和查找任务实例，支持以任务节点名称和节点ID的模糊匹配搜索查找任务，支持对我的实例、今天发布实例的二次筛选，可实现对实例缩小范围或精确定位进行运维。

补数据实例

提供补数据实例的清单，可查看补数据的节点名称，补数据的时间分区范围信息及状态，被补数据的任务节点ID、名称、负责人，补数据的运行时长信息。也支持对补数据实例的搜索、筛选等查找，便捷定位所需查看的具体实例。

逻辑表

支持搜索、查看逻辑表及其包含的物理节点清单，查看单逻辑表详情。可切换查看逻辑表任务和逻辑表实例。逻辑表任务右侧DAG图默认展示当前逻辑表内包含的所有节点和内部节点间的依赖关系（包含“非直接依赖”样式）。逻辑表实例右侧DAG图默认展示当前逻辑表对应的所有节点实例及其状态（如运行中、运行成功、运行失败）。

4.9. 元数据中心

Dataphin包含强大的元数据管理能力，支持MaxCompute、Hadoop、Hive、MySQL、PostgreSQL、Oracle等元数据的采集与抽取，支持对上述计算/存储引擎中元数据的实时追踪，并通过对不同类型存储元数据的抽象，构建统一元数据模型。

Dataphin支持多类型元数据快速扩展，通过元数据中心建设，实现元数据丰富多样、标准统一，为数据地图、数据治理提供强大、稳定的元数据保障。

元数据中心是数据资产管理的核心基础，元数据中心建设内容包括：

- 元数据采集规范：维护统一标准的数据建设规范，确保模型建设、数据表创建、血缘依赖关系记录都能够一致，提高元数据在检索及服务中的可用性。
- 元数据时效与质量：保障元数据产出时间、数据质量，提高资产管理应用数据时效性及研发用户的精准检索。
- 元数据模型体系：通过构建统一的元数据公共模型，兼容多种数据类型，保障数据地图的一站式服务能力。

4.10. 资产分析

数据通过采集、集成、加工等流程构建完成后，通过数据资产模块进行系统化管理。

基于OneData和数据资产的方法论设计应用原则以及元数据全面采集提取、解析管理及加工的技术内核，实现所有数据如资产般分类整理、质量监测、资源优化，保证所有数据的成本最小化消耗、价值最大化呈现并用于业务。

驱动数据资产管理功能的是一系列技术内核。实时事件/订阅服务支撑表、任务等元数据的实时更新，规则引擎支撑治理项规则的高效准确判定与健康模型的构建，日志动态解析支撑每日海量生产任务执行与机器运维日志的分析，图计算则支撑数据血缘关系的分析与构建，Onelog全链路数据追踪技术则支撑数据生产、服务与消费过程中元数据的互通，插件式的元数据接入与加工架构则支撑了对多计算/存储引擎的兼容数据资产管理，是阿里巴巴基于多年海量数据管理经验沉淀出的一整套集分析、治理、应用、运营为一体的方法论与产品，覆盖数据的建、管、用、销全生命周期。

资产分析的主题包含两个关键词：“全域”和“融通”。“全域”是对全域数据的盘点，是通过OneData体系中维度、业务过程、关联关系等构建起数据资产大图，即以模型的语言来盘点描述数据资产；“融通”则是对数据资产在生产过程中的成本与价值进行分析，通过连接度、贡献度两大模型，描述出数据资产中不同数据集在资产大图中发挥的不同作用。

Dataphin数据地图模块，以对企业数据资产分析构建起的数据资产目录为基础，综合用户使用行为，通过元数据画像+搜索引擎，实现对企业数据资产的高效检索。

资产全景

基于OneData构建的企业数据资产，可以结构化展示，以不同形状的物料组件表示业务实体，以不同样式的连线表示实体间不同的业务过程关系， 如此将同一业务板块下数据全景清晰描绘。

资产地图

通过业务板块-数据域-维度&业务过程的关系汇总展示了企业数据构成，并与企业资产全景相对应；同时，结合用户搜索、访问及收藏等主动行为，为用户提供高效、快捷、准确的数据查找与探索的入口。

4.11. 安全管理

Dataphin聚焦智能数据的构建与管理，对于数据安全管理非常重视，面向数据从产生到销毁全生命周期，提供数据访问控制与隔离、数据安全分类分级、个人信息合规管理、数据脱敏、数据使用安全审计等功能，全面保障数据安全。

随着对大数据技术的应用深化，数据安全也成为重要课题。国内《中华人民共和国网络完全法》于2017年6月1日开始实施，鼓励开发网络数据安全保护和应用技术。国际上《欧盟数据保护条例》(General Data Protection Regulation) 于2018年5月25日生效，旨在加强个人信息等数据保护。

数据安全管理最首要的工作是数据访问控制与隔离。Dataphin提供了完善的数据使用权限申请、审批及其生命周期管理，支持多租户访问隔离措施，支持字段粒度的权限管控，并提供基于ACL的数据访问授权模型。

Dataphin基于数据生命周期构建全面的数据安全保障体系，从数据行为、数据内容、数据环境等角度提供技术和管理措施。在大数据构建与管理过程中，Dataphin结合阿里云数据安全管控体系，提供可用不可见的大数据交换共享安全环境、字段粒度的权限访问管控、严格的权限申请审批流程管控、健全的数据使用行为追踪及审计能力等，保障大数据在存储、流通、使用全过程中的安全管控。

目前Dataphin提供权限分层分级及权限的申请、审批、赋权、交还及鉴权的全链路管理流程。

权限类型

为了保证您能安全、有控制地使用产品及访问数据，Dataphin提供角色和资源两个类型的权限机制。

- 角色权限

  为了统一管理用户在平台上的操作，Dataphin除了提供账号管理机制获取超级管理员及系统成员，从平台层面控制用户准入方式。还提供基于项目管理的组织粒度的数据资源获取与操作的权限管理（称之为角色管理），以批量、可管控地赋予系统内用户一组数据资源及操作权限。

- 资源权限

  为了统一管理用户对项目中数据资源的操作，Dataphin提供数据访问控制机制。在项目空间独立管理、成员与资源逻辑隔离的情况下，控制跨项目的存储资源的访问，实现数据在不做迁移时也可在不同项目空间内使用，达到数据共享的目的。

权限管理

- 权限申请

  数据研发者在使用数据地图找到所需数据表并查看该表详细元数据信息后，使用该表需要预先申请权限。

  在具体的权限申请过程中，Dataphin支持根据来源数据表信息，自动显示该表类型和业务板块归属等信息，还支持显示该表字段元数据。

  权限申请流程支持对遵从最小化按需申请原则的响应，即：

  - 支持字段级别的权限申请。
  - 支持多种权限时效的选择，可以自定义选择日期区间或者快速选择30天、90天、180天及1年。
  - 支持用户输入权限申请使用场景及目的说明，以便审批人可以据此判断是否可授权，从而实现根据场景按需申请。

- 申请记录管理

  支持通过权限列表查看申请记录及当前审批状态，并可通过单击**详情**查看申请信息，单击**撤回**撤销工单。对于审批通过的权限，可以查看相关列表清单及具体的字段级信息。

- 权限审批

  权限提交后，系统支持将权限审批的工单随机分发给数据表所在项目的管理员进行审批。在审批人视角，可以通过**我的审批**查看申请人提交的申请信息，并进行授权或驳回。

- 权限交还

  当用户的角色有转岗或者离职变动时，需要提前交还权限，以防止数据及对应生产任务无人管理的情况发生。Dataphin的安全管理支持在**我的权限**页面，单击**交还**按钮，将权限转交给项目管理员，实现权限的有效回收。

4.12. 即席查询

即席查询基于Dataphin强大的OneService引擎，提供高性能的数据临时查询与探索。既支持传统的简单查询方式，也支持主题式查询，在代码简洁度、查询执行速度等方面表现优异。

语法特点

- 支持基于所有建模的逻辑表的离线查询，智能查询引擎会基于产出时间、查询性能等因素选择最优的物理表。
- 支持基于雪花模型的关联查询，使SQL更加简洁和智能。
- 支持物理表、逻辑表以及物理表和逻辑表的混合查询。
- 支持 MaxCompute SQL、Hive SQL等多计算引擎的语法。
- 支持SQL的智能提示、预编译、 格式化等功能。
- 支持逻辑表与物理表字段级别的权限管理与鉴权。

查询执行

在查询脚本文件中，可以根据需要自由输入查询语句，脚本编辑器会根据输入智能给出提示，快速定位所需要的数据表或者字段，同时也会帮助用户校验脚本语法的有效性。