阿里云开发者社区
ALIBABA CLOUD DEVELOPER COMMUNITY

阿里云数据中台产品矩阵 系列白皮书

Dataphin
智能数据构建及管理

2020云栖大会 特别版

数智俱乐部

钉钉扫描二维码 加入

阿里云数据中台 出品

钉钉扫码加入
阿里云数据中台交流群

阿里云开发者“藏经阁”
海量免费电子书下载

目录

1.数据构建与管理的问题与挑战

2009 年，阿里云诞生。自此，中国的各种云端服务开始涌现，传统的数据处理方式，特别是数据仓库领域，发生了剧烈变化。大数据平台和云端服务具有极高的性能、简单的部署、极强的可扩展性和轻松的可管理性，而成本较传统数据解决方案更低。因此，企业很快都在将其数据仓库从本地迁移到大数据平台或云服务中。尽管云服务可以降低存储与计算成本，但是大数据构建本身依旧面临着诸多挑战。

- 数据标准问题：烟囱式开发及局部业务服务支撑，导致同名指标不同口径的问题频发；历史不同业务系统逐步迭代上线，相同对象属性编码不一致等问题突出；
- 数据质量问题：重复建设导致任务链冗长、任务繁多，计算资源紧张，数据时效性不好；口径梳理定义的文档沉淀与开发代码实现脱节，数据准确性保障风险高；
- 需求响应问题：烟囱式开发的开发周期长、效率低，面向应用的服务化不足，导致业务响应速度慢，业务不满意的同时技术又觉得没有沉淀与成长；既懂业务又懂数据的人才不足，需求理解到开发实现涉及大量沟通，服务效率较差；
- 成本资源问题：烟囱式开发的重复建设浪费技术资源；上线难下线更难，源系统或业务变更不能及时反映到数据上，加之数据不标准，研发维护难上加难的同时，大量无用计算和存储造成资源浪费。

本白皮书致力于让企业数据开发人员和 IT 部门了解如何应用 Dataphin 构建企业数据中台，应对上述问题与挑战，构建和管理企业优质的数据资产中心，发挥数据价值。它深入探讨了数据中台构建的核心产品功能和推动数据产品开发的原则并讨论如何支持我们的客户和合作伙伴，并基于真实的客户案例来说明数据中台的实际能力。

2.解决之道：OneData 方法论（One Model+One ID+One Service）

企业的数据构建和管理可以分为三个发展阶段，上述问题和挑战会始终贯穿，但是关注侧重点应有所差异。

第一阶段：在线开发，这个阶段主要关心研发人员个人效率的提升；
第二阶段：数据平台构建与管理，即从在线开发平台升级到以 ETL 处理为中心的数据管理平台，这时对数据质量和资产管理的关注度开始提升；
第三阶段：数据综合治理，目标构建以数据资产为核心的综合数据治理平台。数据资产的核心是把数据按照业务含义进行逻辑化组织，形成规范的标准化管理体系，并能以便捷的方式被下游消费从而产生价值。

Dataphin 基于阿里巴巴集团自身多年数据构建与管理积累的经验，兼顾上述三个发展阶段中的关注点，形成了特有的 OneData 建设方法论，致力于帮助用户沉淀数据资产，构建自身的综合数据治理平台。该方法论体系由“3 个 One”组成，分别是 OneModel, OneID, OneService：

- OneModel：统一模型构建与管理。通过全域数据集成、数据分层架构、业务视角标准规范定义数据和处理数据，致力于统一数据口径、消除指标二义性；
- OneID：核心商业要素资产化。以业务和自然对象为基础，以标签数据为核心，能够实现全域实体识别与连接，数据价值深度萃取，助力企业构建标签体系、完成核心商业要素资产化；
- OneService：统一的主题式服务。以业务便捷消费数据为目标，建立主题式的数据服务单元，面向应用快速构建 API 以提供服务，建立起统一的数据服务中心。

基于上述体系沉淀，Dataphin 以其为内核驱动，提供一站式数据构建与管理能力，帮助企业更好的构建数据中台，夯实数据基础，以实现数字化转型。

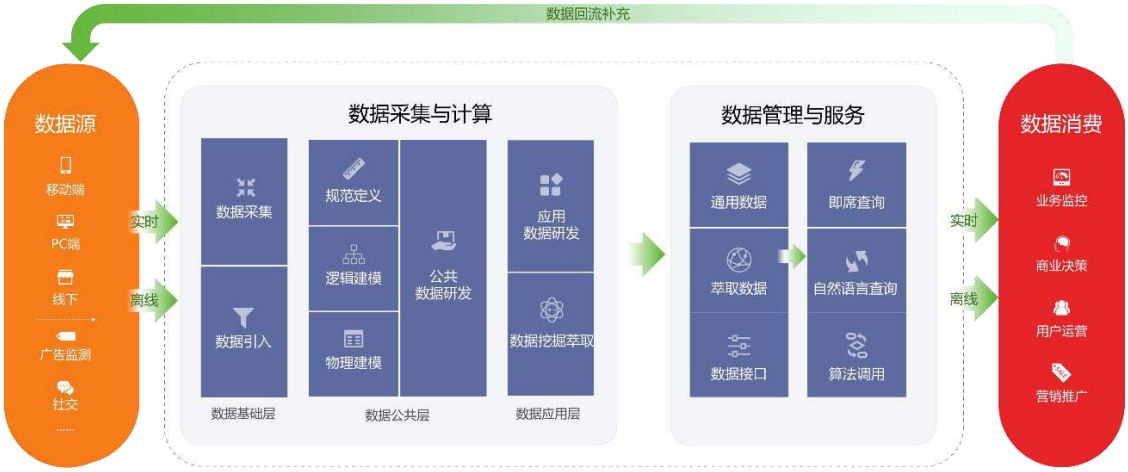
3.Dataphin 智能数据构建与管理

3.1 产品定位

Dataphin（智能数据构建与管理）是智能大数据建设平台，旨在面向各行各业大数据建设、管理及应用诉求，通过输出阿里巴巴集团数据中台长达十年实战沉淀的大数据建设 OneData 体系（OneModel + OneID + OneService）的产品、技术以及方法论，一站式提供集数据引入、规范定义、数据建模、数据研发、数据萃取、数据资产管理、数据服务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、融会贯通、资产化、服务化、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、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，同时支持数据表查询、智能语音查询等多种类型数据服务。

3.2 产品能力框架



3.3 产品使用流程

Dataphin 的产品使用大致分成几个部分：数仓规划、数据研发、资产管理和资产服务。

在 Dataphin 中配置好所需计算资源与存储资源后，通过数据集成功能，将散落各处的数据从源系统引入到平台内，形成基础数据中心，然后开始公共数据中心及萃取数据中心的建设。

数据建设有两种研发模式可选：第一种是自动化建模研发，第二种是编码研发。自动化建模研发标准、规范、高效，其实现过程为：通过可视化定义配置的方式完成数据逻辑建模、自动化物理建模及代码自动化的研发，核心流程包括业务视角的数据元素规范定义（维度、业务过程），并基于此进行的逻辑建模（维度逻辑表、事实明细逻辑表），并通过定义原子指标、业务限定等组装形成面向消费的派生指标（系统自动汇集为汇总逻辑表），系统将基于逻辑模型自动构建物理模型并自动化完成代码生成，提交发布生产后，系统可基于自动化形成的代码运行并处理数据。编码研发自由、灵活，通过手工编码的范式进行数据处理，可作为自动化建模研发模式的补充，扩展数据处理场景。研发完提交发布生产环境后，可对相关任务脚本进行运维，如调度执行情况查看、出错分析及重执行、监控告警配置与管理等。

所有处理的数据结果被视为资产进行统一的资产管理，包括：在资产全景与地图中对资产目录与详情进行查看（包括表结构、血缘关系、产出信息等）；在资产质量模块中可以配置规则对资产质量监控及分析；在资产治理板块可以对数据计算与存储资源进行分析并基于治理

项对相关成本浪费点进行识别和治理。最后可以在数据服务板块配置生成主题式的服务单元及 API，对外系统提供便捷的数据访问服务。

3.4 产品功能列表

功能名称	功能描述
计算引擎设置	支持选择不同计算引擎进行数据处理，包括：MaxCompute、Hadoop Hive、AnalyticDB PostgreSQL、Flink。
数据集成	提供多种异构数据源的数据读写能力，支持传统 IDC 数据库、阿里云数据库、非结构化存储、大数据存储等多种数据存储作为数据同步来源或目标；支持可视化拖拽式数据同步管道配置开发，并提供轻度数据清洗转换、脏数据过滤、流量控制等能力。
数据建模	提供体系化、系统化建模及研发能力，将数据仓库理论以工具化、半自动化的方式实现：自顶向下快速构建业务维度、业务过程，并进一步细化构建维度表、事实表、汇总表、应用表，沉淀标准统一的数据资产，便于业务快速分层并进行智能数据应用，同时优化计算存储。 基于规范定义后的数据对象，面向业务视角，可视化构建数据逻辑模型。模型包括维度逻辑表、事实逻辑表及汇总逻辑表。模型定义完成发布生产后，系统将自动化生成物化代码并定时执行调度，完成公共数据中心的全托管生产。
数据研发	支持基于 MaxCompute、AnalyticDB PostgreSQL、Hive 的多种离线脚本类型，如 SQL、MapReduce、Python、Shell、Spark Jar；支持基于 Flink 大数据引擎的实时脚本类型如 SQL、DataStream（原生 API 模式）。支持编码任务的自定义调度配置，资源配置，任务参数配置。支持查看代码结构及执行计划，支持查看历史版本及回滚。在建模研发的基础上扩展支持编码研发，满足多样化、多时效的业务场景诉求，加持智能大数据构建。 支持用户自定义函数辅助任务研发过程，支持系统内置函数，支持 Jar、Python、Json 等多种资源包类型，支持自定义数据源进行数据输入输出源类型扩展，支持代码模板实现 SQL 片段复用能力，支持文件克隆、全局代码搜索等辅助研发工具。 支持基于计算引擎快速实现物理表与逻辑模型的数据查询与结果获取，实现轻量化数据查询服务。

数据萃取	在基础数据中心及公共数据中心基础上，支持以“目标对象”为中心，用参数选配的方式可视化识别与连接业务对象并提取对象行为与标签，智能完成数据打通、深度挖掘并生成定时调度任务，完成萃取数据中心的全托管生产。
任务运维	支持千万级任务的稳定调度与运维，新增支持各种不同周期（天、小时、跨周期）任务调度。基于数据建模研发、编码研发、数据萃取生成的任务进行调度与运维管控。包括生产任务运行、依赖管理与维护、资源消耗管控等。支持生产任务监控告警与自定义告警规则配置，确保任务正确生产与调度。
数据资产管理中心	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的基础上，支持元数据深度分析并实现资产化管理数据，支持可视化业务数据大图、资产分布及元数据详情等，帮助全局数据查找与分析。
资产质量	支持完备的数据质量管理与监控能力，保障数据的生命线。支持全局质量概况，可快速定位与识别数据质量问题；支持自定义质量校验规则配置，基于模板定义质量规则执行，汇总质量规则的校验结果并生成质量报告，为数据消费者呈现数据质量现状。
资产治理	支持全局资源消耗增速、资源消耗分布和资源治理概况总览，支持针对项目空间进行问题诊断，同时对治理问题点进行分析；分析治理后效果，可帮助评定治理情况并进一步推动治理优化。
数据服务	支持基于业务主题式的逻辑表查询访问服务，也支持物理表的查询访问服务。面向 API 提供者的开发中心，支持基于数据源的数据表创建生成主题式的服务单元及 API；面向 API 调用者的应用中心，支持基于业务应用的视角对 API 进行集中式的查看和调用，支持权限控制和流量控制。

4.产品的特色能力

Dataphin 产品核心特色是以元数据驱动的数据智能构建与管理能力，以业界领先性的元数据智能、数据虚拟化、端到端的全链路血缘为理论基础。

1）元数据智能方面：

Dataphin 自动化地采集多引擎、多存储的元数据，构建在线、实时和离线数据资产多时效的统一元数据资产，形成数据智能构建与管理的驱动力，在线的元数据服务于数据智能引擎的建模与自动化编码，实时元数据主要服务于数据资产全景与目录，离线算法模型则服务于数据质量、安全与成本高效治理能力。在 Dataphin 中不仅是基于元数据对已有的数据采建管用链路中的功能进行增强，而是以数据为驱动力地实现了比如自动化编码、模型自优化、管理自优化等能力。元数据智能构成企业数据资产构建与管理的数字化基础。

2）数据虚拟化方面：

Dataphin 通过数据虚拟化技术从业务视角结构化组织数据、标准化定义指标，通过技术视角的数据组织和生产由系统完成，最终以业务视角的逻辑形式表达数据提供消费。这一过程在确保数据高效高质生产的同时，大大降低数据理解成本，为企业内数据资产的快速流通与价值化提供基础。

3）端到端的全链路血缘方面：

Dataphin 通过 OneLog 技术正逐步打通数据全生命周期、数据研发生产全链路的元数据，实现端到端的全链路血缘能力，基于此，企业可以真正意义上精准地追踪数据的来源和业务应用，并回收数据应用的价值表现，数据在业务中的价值评估成为可能；端端全链路血缘构成企业“可评估”的数据价值化的基础。

5.典型的场景与案例

阿里巴巴数据中台解决方案成型于零售行业，但绝不止步于零售行业。在过去的三年中，Dataphin 经过了不同行业不同场景的实践和检验，沉淀了丰富的客户案例和行业经验。

行业	客户案例
零售	某大型乳业企业，基于 Dataphin 构建数据中台。通过 Dataphin 与各个业务系统进行对接，完成这些系统的数据的沉淀与汇集，并根据各种业务场景对数据进行智能分析，然后用分析结果实时呈现业务状况，同时指导各个业务运营，构成一个闭环。 数据中台的建设该该企业带来诸多价值：完成数据的汇集与统一构建，让公司上下对业务能够有统一的认知；提供实时的数据呈现，支持企业业务决策的快速进行；能够帮助业务部门实现精细化运营与服务，如个性化的导购服务、恶意积分兑换识别与控制等。
电信运营商	某电信运营商，基于 Dataphin 完成阶段性的数据中台建设。在 Hadoop 大数据平台之上，采用 Dataphin 和 Quick BI 产品，汇聚用户 O 域（网络管理域 OSS），B 域（业务支撑域 BSS）和 M 域（信息管理域 MSS）全域数据，完成了 IPTV 用户分析和 LBS 服务分析，有效的支撑省及地市分公司数据建设和消费的同时，在大数据中心数据加工效率和资产管理能力方面也提升明显。
传媒	某大型电视台互联网业务，它面临着缺乏指标数据体系支持决策、数据标准缺乏统一、用户服务内容无个性化能力导致竞争力差及粘性不足等挑战，基于阿里云专有云平台及 Dataphin 等产品，遵循 OneData 的理念，构建日志、会员、内容等七大数据主题域，完成源数据层、中间层、应用层数据处理加工体系，统一了数据资产，通过实时直播大屏等大数据应用实现数据服务能力的大大提升，获得业务团队的一致认可。
地产	某地产集团，基于 Dataphin 的数据智能构建与管理能力，集成了其地产、物业、IoT、文旅、智慧案场和外部合作数据等近 60 个数据源的数据，体系化完成了数据资产化，构建数据中台。一阶段完成后，一方面是效率提升，能以 5 人的数据团队，服务集团 17 个区域及总部职能人员共计几百人的数据需求；另一方面是数据价值呈现，基于融合的数据，在业务决策分析、供应商管理、智慧客服等业务上发挥了举足轻重的作用。
航空	某大型航空公司，存在数据不完整、不统一、不开放的问题。为解决这些问题，其引入了阿里云 MaxCompute 大数据计算平台，同时基于 Dataphin 产品，完成几十个业务系统数据的完整集成、基于标准规范的数据模型实现数据统一，同时提供统一的数据服务，应用于收益分析、运行品质、地服品质、客户画像等。

6.售卖及服务方式

Dataphin 提供公共云在线服务和线下独立部署两种服务模式。

Dataphin 支持多种售卖版本和计算引擎，每种计算引擎对应不同的售卖版本。在每个版本最小功能合集的基础上，可根据实际需求场景灵活叠加购买增值功能包，以夯实数据构建与管理基础，更好地对接上层应用服务。

1）公共云在线服务

公共云环境下，Dataphin 支持按月订购的预付费模式，开通即可使用。企业根据自身数据处理规模、企业/数据所在地理位置、数据处理计算引擎选择、数据构建与管理的阶段差异诉求，可在华东、华南、华北地域选择基于 Maxcompute 等计算引擎的 Dataphin 服务，同时可叠加数据萃取、数据质量、资产治理、数据服务等增值功能，实现按需获取服务。

2）线下独立部署

独立部署环境下，Dataphin 提供一次购买软件并每年订购维保的买断式服务。与公共云类似的，可根据需要选购不同的产品功能规格。同时，除了产品服务，独立部署环境下 Dataphin 还支持按需订阅专家服务。

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的，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文档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，阿里云以大数据集成服务平台的“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”、“AliCloud”、“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。

钉钉扫码加入
阿里云数据中台交流群

阿里云开发者“藏经阁”
海量免费电子书下载