

大数据操作系统产品

亚信科技DataOS产品白皮书

AISWare Dataos是亚信科技全力打造的数据操作系统产品，是以数据融合为基础、以智慧运营为衔接、以生产执行为服务为重点打造的数据操作系统产品，深化跨域数据整合并沉淀公共的数据能力实现数据的分层与水平解耦，同时整合数据处理、治理、安全、运营等多种能力实现精细数据服务满足数据开放和智慧运营要求。

目录

一. 摘要	3
二. 缩略语与术语解释.....	4
三. 产品概述.....	5
3.1 趋势与挑战	5
3.2 产品与定位.....	6
四. 技术介绍.....	7
4.1 亚信科技AISware BigData产品架构.....	7
4.2 AISWare BigData Dataos产品架构	8
4.2.1 数据治理中心.....	9
4.2.2 数据开发中心.....	12
4.2.3 数据交换中心.....	16
4.2.4 敏捷建模工具.....	19
4.2.5 数据开放中心.....	22
4.2.6 指标管理工具.....	25
4.2.7 数据安全中心.....	29
4.2.8 数据资产运营中心.....	33
4.3 关键技术	38
4.3.1 元数据驱动开发.....	38
4.3.2 集成开发环境IDE.....	38
4.3.3 技术平台能力的封装.....	38
4.3.4 基于Lambda架构的流批一体化开发.....	38
4.3.5 去中心化分布式调度架构.....	39
4.3.6 数据安全控制能力.....	39
4.3.7 数据查询安全控制.....	39
4.3.8 指标维度组存储设计.....	39
4.3.9 指标库文件存储设计.....	40
4.3.10 数据资产评估.....	40

五. 功能介绍	41
5.1 基础功能	41
5.2 特色功能	49
5.2.1 能力解耦和云化服务	49
5.2.2 开源及第三方快速集成	50
5.2.3 处理脚本自动生成与优化诊断	50
5.2.4 云端的数据开发环境	51
5.2.5 一体化DSL数据处理与访问	51
5.2.6 自动化和智能化的数据质量稽核	52
5.2.7 智能识别的主数据管理	52
5.2.8 数据虚拟化访问与安全控制	52
5.2.9 全方位管控的数据开放安全	53
5.2.10 服务的流控与信控	53
六. 带给客户的价值	54
七. 产品优势	55
八. 应用场景	56
8.1 设计即开发，敏捷建模协同交付	56
8.2 实时离线开发一体化，满足实时场景	56
8.3 体系化模型管理，缩短模型价值转化路径	57
8.4 基于目录按需数据分发共享，自主应用构建	57
8.5 构建企业关键指标体系，辅助科学管理、决策	58
8.6 构建模型一体化运营平台	58
九. 联系我们	59

[返回目录](#)

一. 摘要

当今社会已全面迎来数字化时代，物理世界的数字化属性不断挖掘、数字价值深度释放，以大数据、人工智能、区块链、云计算等为代表的新一代信息技术飞速迭代，加快向产业和行业下沉。数字化转型赋能产业、重塑业务模式，成为新旧动能转化，实现经济高质量发展的必由之路。

数据是数字化转型的基础也是关键，拥有数据的企业，都在考虑如何从拥有的数据中，挖掘新的市场机会，规避经营风险，提升企业竞争优势。但是随着万物互联，产生的数据越来越多，结构越来越复杂，如何管理数据，让数据更好的产生价值，服务于各行各业，对我们的系统支撑都带来了挑战。

为此，亚信打造了以数据为核心的DataOS产品，从数据治理、数据开发、数据开放、数据安全、数据运营五个方面进行全方位支撑。

本白皮书将从产品概述、技术架构，主要功能、客户价值、产品优势等几个方面阐述亚信AISWare DataOS产品。



[返回目录](#)

二. 缩略语与术语解释

缩略语或术语	英文全称	解释
元数据	Metadata	元数据是描述 数据 的数据，主要是描述数据 属性 的 信息 ，用来支持如指示存储位置、 历史 数据、 资源 查找、文件记录等功能。元数据主要有三类：业务元数据、技术元数据、管理元数据。
Data architecture	Data architecture	数据架构是实现数据规划的载体，是揭示业务本质、描述公司数据关系的全景视图，是统一数据语言、理顺数据关系、消除信息孤岛、建立数据互联的基础。
数据标准	Data standard	数据标准是公司层面共同遵守的数据含义和应用规则，是对数据项的统一解释。
主数据	Master Data	主数据 也称基准数据，是指系统间的 共享数据 。主数据的特点是变化缓慢，企业内部保持一致。 主数据 必须存在并加以正确维护，才能保证企业内部系统数据的 参照完整性
数据质量	Data quality	数据质量是指数据反映业务真实情况，且满足应用要求的程度。数据质量主要通过数据的准确性、完整性、及时性、一致性、规范性和关联性六个方面来衡量。
数据运维	Data operation	数据运维是为了满足数据共享与应用的需要，对数据进行的日常性运行维护操作
数据应用	Data Utilization	数据应用是指以数据价值发现为目标，对数据进行加工处理的过程。
数据资产目录	Data Assets Catalog	数据资产目录是指对数据中有价值、可用于分析和应用的数据进行提炼形成的目录体系
数据元	Data Element	通过定义、标识、表示以及允许值等一系列属性描述的数据单元。在特定的语义环境中被认为是不可再分的最小数据单元。

三. 产品概述

AISWare BigData Dataos是以数据融合为基础、以智慧运营为衔接、以生产执行为服务为重点打造的数据操作系统产品，深化跨域数据整合并沉淀公共的数据能力实现数据的分层与水平解耦，同时整合数据处理、治理、安全、运营等多种能力实现精细数据服务满足数据开放和智慧运营要求的产品。

3.1. 趋势与挑战

2019年，大数据正式进入业务智能化阶段，进入数据智能阶段后，整个行业呈现出两大趋势：第一，多技术融合。从客户需求来看，为了指导决策，需要基于复杂网络的推理和决策。数据智能正在向以数据挖掘、形成解决方案、辅助决策以及预警分析为特征“智能+”阶段迈进。第二，中台的形成。在企业数字化转型进程中，需要具备快速迭代升级的能力，基于数据驱动业务发展，这就需要建立一站式技术能力、统一的数据管理、快速配置开发业务的能力。

Forrester Research预测：2019将进入数字化转型的实用之年，即通过更加现实、可实践的步骤，为企业带来真正的价值。由于运营效率的原因或创新发展的需要，企业对数字化的需求会越来越高，为此也会对现有的数据管理带来挑战。

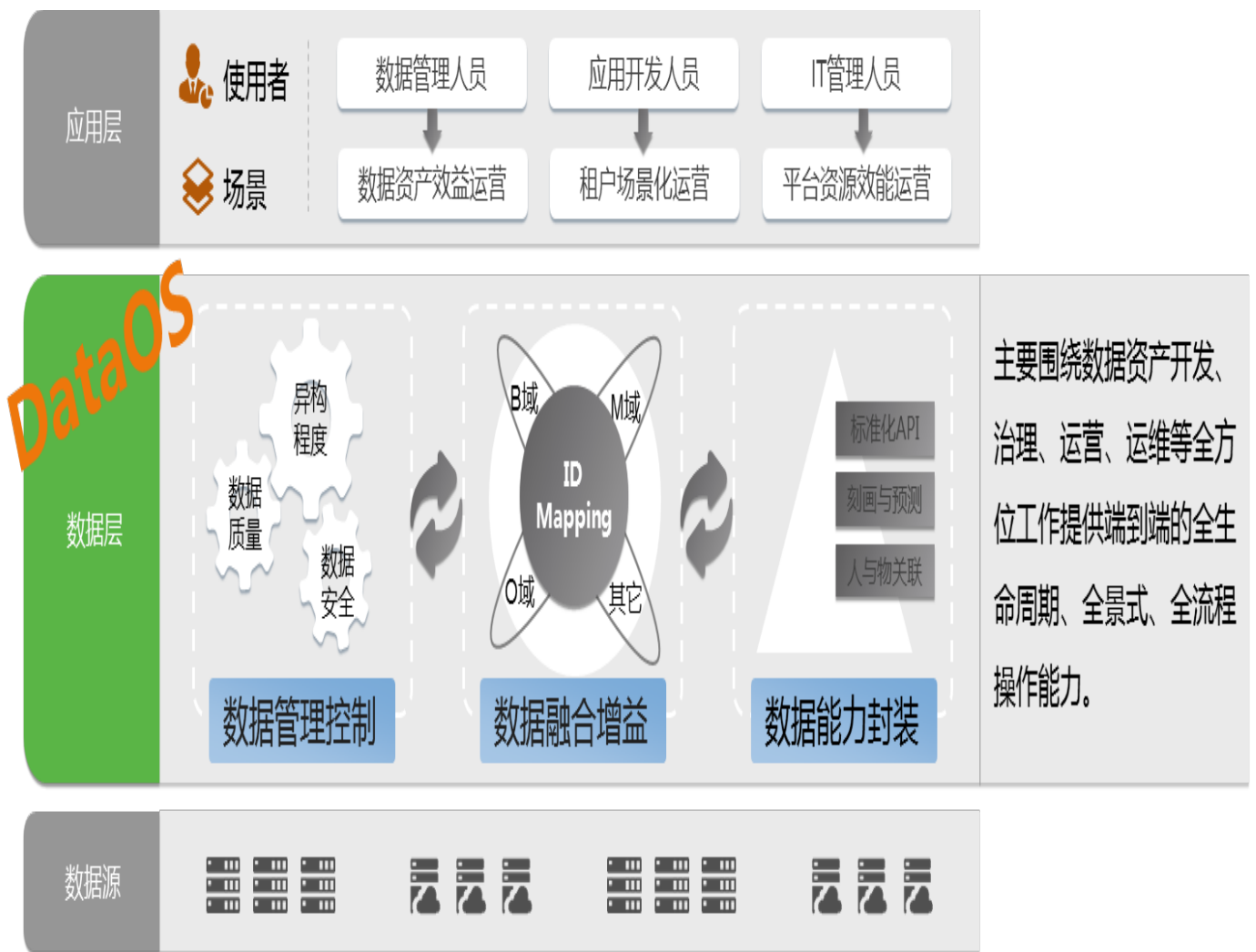
挑战1：数据服务实时性。对于一些实时性要求高的场景，比如实时营销、位置服务等。为此需要构建面向实际业务生产提供全流程打通的数据能力，将离线分析能力、实时处理能力进行融合，实现统一可视化编排及监控，并结合数据挖掘能力形成企业级端到端数据能力。

挑战2：数据资产管理全面性。随着5G，物联网技术的推进，越来越多领域的的数据汇聚在一起，数据量越来越大，结构越来越复杂，数据服务的对象越来越多，对数据的一致性、准确性有更高的要求，需要具备全域的全流程端到端的数据数据资产管理支撑能力。

挑战3；创新带来的变化性。在数字化创新过程中，上层应用的变化越来越快，如何快速的支撑上层对数据服务的需求已经迫在眉睫。为此需要以数据为中心，构建企业级数据中台来支撑前端对数据的需求，实现数据服务的沉淀与共享，快速满足前端业务的变化。

3.2. 产品定位

以建立最佳集成标准为着眼点，通过屏蔽技术多样性和数据多样性构建适用于业务领域的数据中台助力全域全体系数据中台支撑能力的打造，彰显在市场支撑、业务转型、产品运营及品牌提升方面的价值。



四. 技术介绍

4.1. 亚信科技AISware BigData产品架构

亚信大数据域产品集包括：

AISWare Integrated Operation Platform (集中运营平台)

AISWare DataApplicationPortal (数据应用门户)

AISWare Internet Insight Platform (互联网内容洞察)

AISWare DataDiscovery (数据探索分析平台)

AISWare DataOS (数据中台操作系统)

AISWare DataEngine (大数据计算引擎)

AISWare DataInfrastructure (数据基础设施)

AISWare AIOps (全域智能运维平台)

AISWare BigData产品体系中产品间关系及AISWare BigData Dataos位置如图1所示。

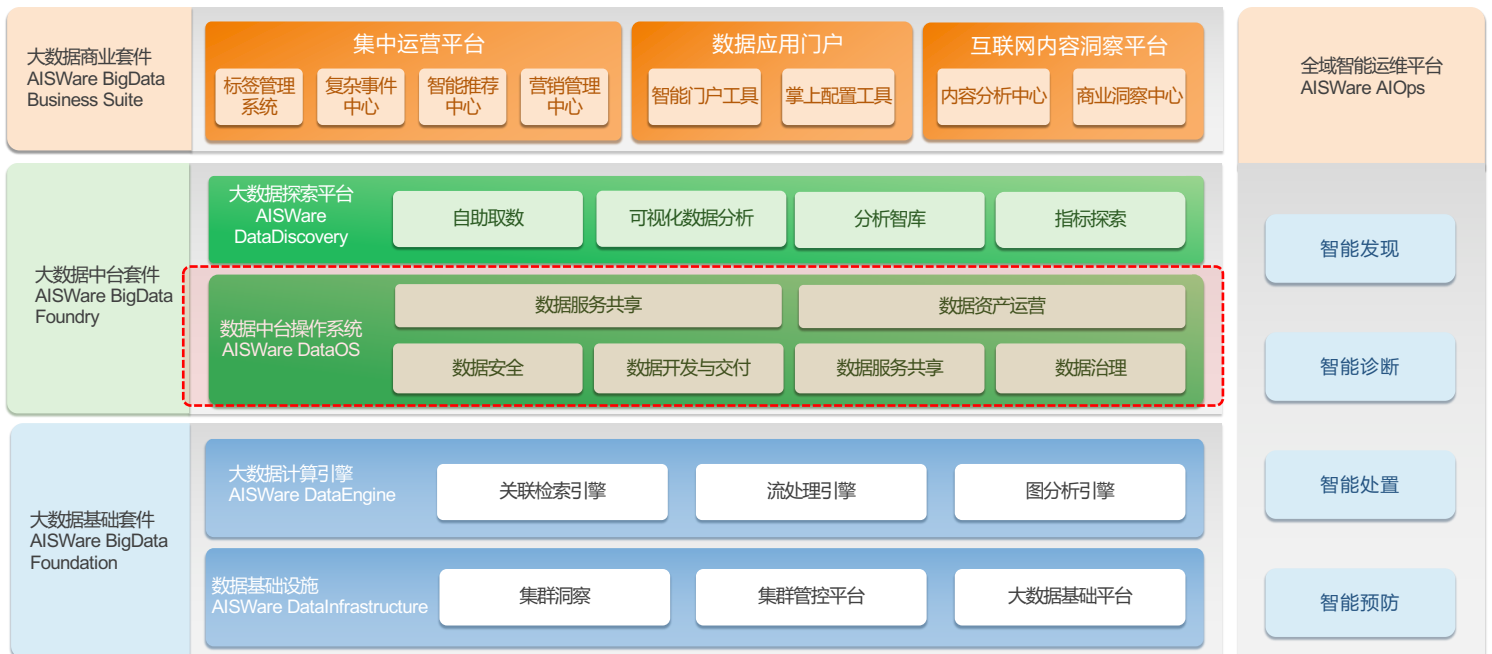


图1 亚信大数据域产品集总体架构

[返回目录](#)

4.2.亚信科技AISWare BigData Dataaos产品架构

亚信AISWare BigData Dataaos整体技术架构如图所示，其中包括数据治理中心、数据开发中心、数据交换中心、数据敏捷建模工具、数据开放中心、指标管理工具、数据安全中心、数据资产运营中心几部分。

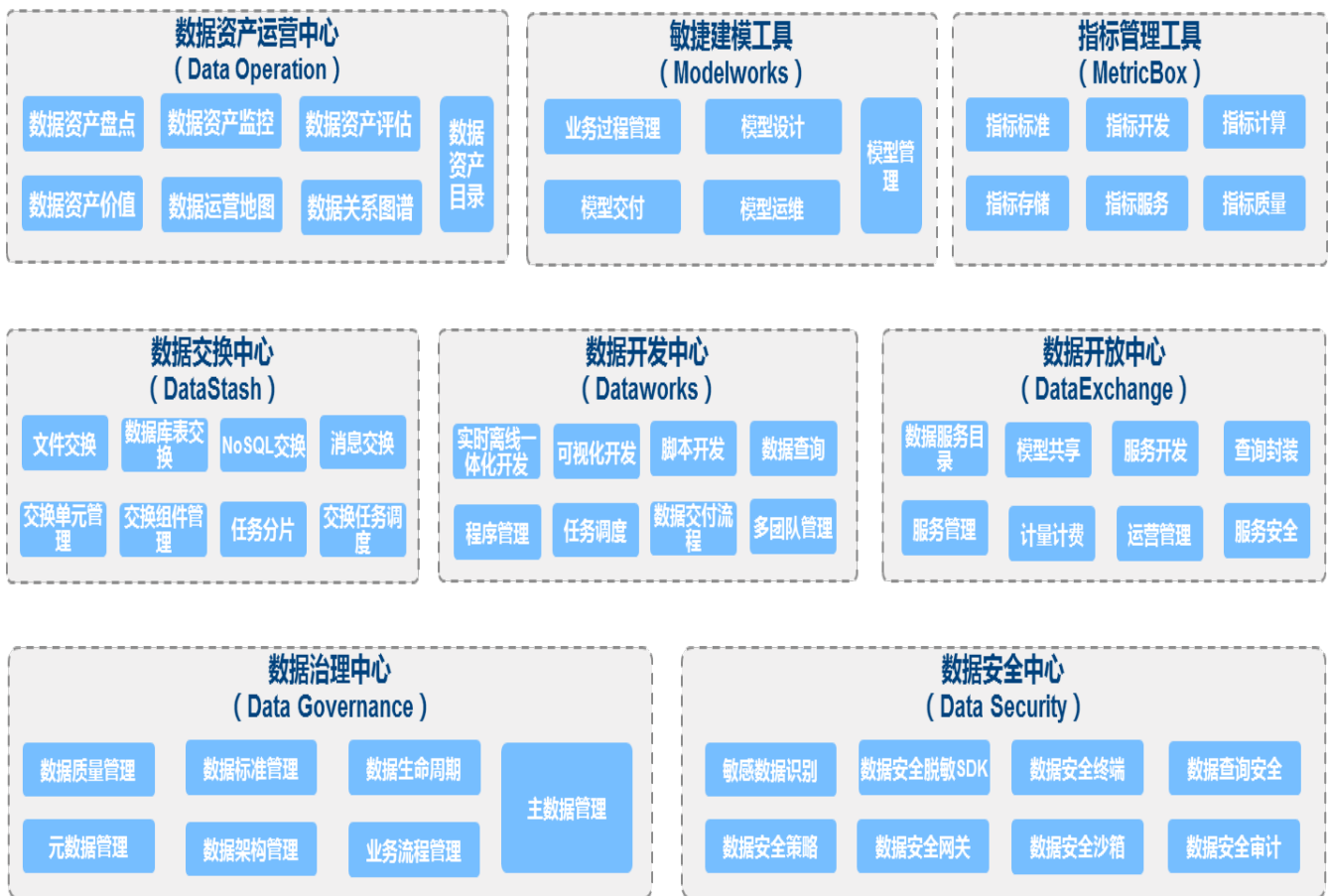


图2 AISWare BigData Dataaos产品架构

[返回目录](#)

4.2.1. 数据治理中心架构



[返回目录](#)

4.2.1.1. 数据的标准化

数据标准是大数据平台数据治理的基础性工作，是数据治理建设中的首要环节。数据标准为大数据平台提供统一的数据标准定义和平台逻辑模型，数据标准是大数据平台进行数据治理的依据和根本，数据标准是衡量企大数据平台数据资产运营和管理的评估依据，通过数据标准管理的实施，实现对大数据平台全网数据的统一运营管理。通过统一的数据标准制定和发布，结合制度约束、系统控制等手段，实现企业级大数据平台数据的完整性、有效性、一致性、规范性、开放性和共享性管理，提高企业级省大数据平台数据治理水平。

4.2.1.2. 元数据管理

元数据管理模块是数据资产管理系统的基础模块。以元模型驱动,通过元数据连接数据本身管理、IT过程管理和生产运行管理，支撑企业大数据中心数据处理全流程的结构化、非结构化数据的元数据管理，保障生产过程数据质量。将元数据管理贯穿数据的整个生命周期，将元数据管控融入到开发管控、数据质量管理、运维调度、数据开放、应用运营等数据管控的每个阶段。

4.2.1.3. 数据质量管理

数据质量管理是对提供业务需求的数据进行全面质量管理，通过数据质量相关技术手段及管理方法、组织、流程、评价考核规则的制定，及时发现并解决数据质量问题，提升数据的完整性、及时性、准确性及一致性，提升业务价值。

4.2.1.4. 开发管控

采用元数据驱动(MDA)设计理念，去规划元数据对象的创建、运行、评估、维护各环节，管控数据的整个生命周期。提供统一开发管理、可视化管理、过程管理、模型管理、开发质量管理、测试管理。

[返回目录](#)

4.2.1.5. 数据查询安全管控

SQL解析控制数据安全透明访问，即使用者按照正常的sql语句进行提交查询，同样的SQL语句，能够根据不同的使用者返回不同的数据结果。

4.2.1.6. 数据评估

数据评估是对数据管控成果的集中体现，并总体评估数据管控的成熟度。通过对数据的集成度、分布、冗余度、数据热度、重要性等多种角度实现数据评估，并基于评估结果对数据进行相应的处置。

4.2.1.7. 数据查询安全管控

SQL解析控制数据安全透明访问，即使用者按照正常的sql语句进行提交查询，同样的SQL语句，能够根据不同的使用者返回不同的数据结果。

[返回目录](#)

4.2.2. 数据开发中心架构



图4 AISWare DataInfrastructure-CI产品架构

4.2.2.1. 开发管理

从系统资源管理，数据权限管理，开发需求及流程管理等方面，实现对大数据开发平台的总体管控，为各个角色合理使用平台奠定基础。主要有以下功能：

开发资源管理：采用多租户的思路，将数据能力和数据平台数据处理能力按需、可控的进行开放，在保障数据安全性、数据可控性的前提下，通过规范化封装的数据操作，可视化开发工具开放给业务运营部门，由其自行进行数据操作开发；

开发团队管理：通过映射开发管理平台帐号及执行平台帐号，以租户的方式实现用户及用户组管理，以达到资源管控及数据权限控制的目的。

开发流程管理：支持ETL上线，下线，变更，数据上线，数据下线等流程管理；也可自定义业务流程。

开发架构管理：定义企业的数据架构和存储架构；包括数据的主题、层次，所涉及的业务部门、所用到的数据库类型等；存储架构设计考虑在异构数据库的环境下，不同的数据库有不同的参数设置，为减少开发人员在模型开发过程中的操作，同时也保障模型在落地到物理环境的正确性，通过在存储架构中设置默认的参数配置来保障开发的准备性。存储架构支持的数据类型有Gbase、Hive、RDB等各类不同的数据库

开发工具管理：平台提供了大量的开发工具组件，如数据交换，数据处理，流程控制，文件处理等，通过对团队的授权，团队成员能够有权使用相关开发工具。

4.2.2.2. 数据开发

基于Lambda架构，构建流、准实时、批处理统一的DAG编排能力，完成多种环境的全量与增量数据处理能力，简化开发过程，提高开发效率。开发管理以图形化开发界面、拖拽式开发方式、统一封装的函数库、灵活脚本编辑能力对开发过程进行支撑。包含以下功能：查看元数据、数据模型注册、数据模型导入、设计、开发、测试、上线、优化等。

4.2.2.3. 智能调度

打通任务侧到底层资源侧数据流向，提高任务运行分析效率，基于元数据的数据关系和资源感知，实现零配置的调度任务生产。

[返回目录](#)

4.2.2.4. 数据探索

作为开发者统一数据查询入口，无需不同数据库之间的切换查询，能够支持不同类型数据库（包括rdb、mpp、hive）查询能力，支持数据查询中断，查询日志记录、简化查询操作，快速查看数据字典，开发人员只能查看少量样例数据辅助开发，提高开发效率。

4.2.2.5. 调度能力

分布式调度主体能力：提供了从任务的生成，依赖判断，任务执行到状态维护的任务全生命周期管理。通过实时心跳检测，对任务的健康度进行监控。并提供了在异常状态下自我恢复的能力。提供模版调度配置，通过模版批量生成子任务，。支持批处理与流任务的统一调度。

插件：支持阻塞型与非阻塞插件类型。支持在任务执行超时/任务失败/任务启动超时/任务完成超时 状态下，运行指定的插件。系统自带的插件类型包含：sql插件，质量检查插件。

租户化：提供了按集群组进行调度资源的隔离。

4.2.2.6. 运维能力

运维能力包括智能分析与智能运维，其中智能分析包括：

阻断分析：通过任务的血缘，对执行路径上非根本原因的节点主动过滤，寻找最底层的阻断原因，提高运维问题查找的效率。

血缘分析：展示任务完整的血缘，提供完整的任务链路视图

影响分析：展示任务的影响节点，提供表格及图形两种查看模式。支持用户根据标签筛选/反选任务，进行选择性地重做。

智能运维包括：

任务运维包括：权限管控：可根据用户在团队内的角色，控制用户的可操作范围。过程审计：记录用户对任务的操作记录。自主监控：支持自定义监控的内容，根据用户习惯保存监控列表

集群运维包括：一键启停：页面上一键启停。生命周期：自带生命周期管理，清理过期的日志文件及日志数据

[返回目录](#)

4.2.2.7. 运营能力

运营能力包括任务运营与集群运营。

任务运营：提供调度总览，直观查看平台总体任务情况。主动分析任务执行路径，提供路径优化建议。

集群运营：提供调度效能实时分析报告，为系统负载提供数据支撑

[返回目录](#)

4.2.3. 数据交换中心架构



图4 AISWare DataInfrastructure-CI产品架构

[返回目录](#)

4.2.3.1. 数据交换

使用多样的数据采集与交换场景，支持多种复杂的数据源与数据类型。数据采集与交换分为离线采集（不落地，落地），实时采集（流数据、CDC），采集与交换是IO与网络密集型操作，平台设计的目标是跑满网卡与磁盘；支持库到库、文件到库、文件到文件、库到文件交换能力；采用去中心化架构，提高并发性及高可用性。提供按租户方式进行数据交换权限分配，从开发、上线、监控整套流程，每个环节加强租户隔离控制；解决租户不同异构数据之间交换及访问问题。

4.2.3.2. 数据申请

支持数据申请功能，包括临时性数据交换申请和日常的批量数据交换申请。

4.2.3.3. 数据审批

使用者在进行数据申请、数据导出操作等均需管理人员审批，该功能需配置相应的人员管理权限。

4.2.3.4. 数据赋权

对于数据审批通过的数据，其权限需要对相应的角色或帐号进行赋权，以保障数据安全合理的使用。

[返回目录](#)

4.2.3.5. 数据任务配置

源数据读取配置，包括TxtFile文件读入、(s)ftp文件、HDFS文件读取、JDBC读取、Hive文件导出、Hbase读取。

目标数据写入配置，包括DB2导入、Gbase导入、Hive导入、JDBC导入功能。

4.2.3.6. 数据执行

源数据读取执行：包括TxtFile文件任务处理、FTP/SFTP文件任务处理、HDFS文件任务处理、JDBC任务处理、Hive文件导出

目标数据写入执行：包括TxtFile文件写入、FTP/SFTP文件写入、HDFS文件写入、JDBC数据写入、HBASE数据写入、DB2文件导入、Gbase导入。

4.2.3.7. 任务监控

运行任务监控：能够对在运行的临时与批量任务是否成功、是否挂死等异常情况进行监控；根据交换的数据特点制定质量监控点，包括但不限于数据的条数、文件大小等；

运行日志查看：对分发任务情况进行监控，收集运行开始时间、结束时间、运行时长、状态（成功、失败）等日志信息；

数据完整性稽核：通过记录数检查，比对源表、目标表数据量，检查分发的数据是否有缺失；

数据一致性稽核：数据内容检查，比对源表、目标表数据内容上是否有差别；

数据及时性稽核：需对分发完成时间进行检查，检查分发程序是否在规定时间内完成；

[返回目录](#)

4.2.4. 敏捷开发工具架构



图4 AISWare DataInfrastructure-CI产品架构

[返回目录](#)

4.2.4.1. 需求管理

需求详情：明确需求内容、成员、工作目标与计划，并提供文档、模型、业务口径等参考内容。

任务管理：可将需求拆分为多个任务，便于追踪需求进度，持续交付。并量化团队和个人工作量与交付质量。

需求进度：从整体上查看需求完成情况与工作量情况，并通过流程将需求结束。

4.2.4.2. 模型设计

流程图设计：提供多种流程图设计功能，包括数据流向图、业务流程图、实体关系图、脑图、UML图等。

概念模型设计：概念模型是从现实世界到信息世界的第一层抽象，确定领域实体属性关系等内容，提供模型、程序、文件等不同类型的概念实体。

物理模型设计：物理模型就是根据逻辑模型对应到具体的数据模型的机器实现。数据库中的一些对象如下：表，视图，字段，数据类型、长度、主键、外键、索引、是否可为空，默认值。提供物理模型设计功能。

程序设计：提供自动生成基础sql代码功能，辅助快速开发。并能解析sql代码，识别来源、输出、临时模型实体。

多人协作建模：多人协作迭代式进行开发与交付，提高开发效率，提高设计开发一致性。管理、设计、开发、测试、维护高效协作。

敏捷建模版本管理：对设计图、模型、程序进行管理，保存每次变更过程。

4.2.4.3. 模型测试

测试用例设计：依据模型分区设计、逻辑约束、数据标准等，生成相关测试用例。

测试用例执行：通过手工，或在程序中增加测试步骤，对模型进行临时或周期性单元测试，并展示测试结果。

[返回目录](#)

4.2.4.4. 模型管理

模型打包上线：将设计图中待上线模型、程序打包，通过上线流程一次性提交。

生命周期管理：基于模型设计中的生命周期配置，对数据进行管理。

模型知识发布：基于模型设计中的生命周期配置，对数据进行管理。

[返回目录](#)

4.2.5. 数据开放中心架构

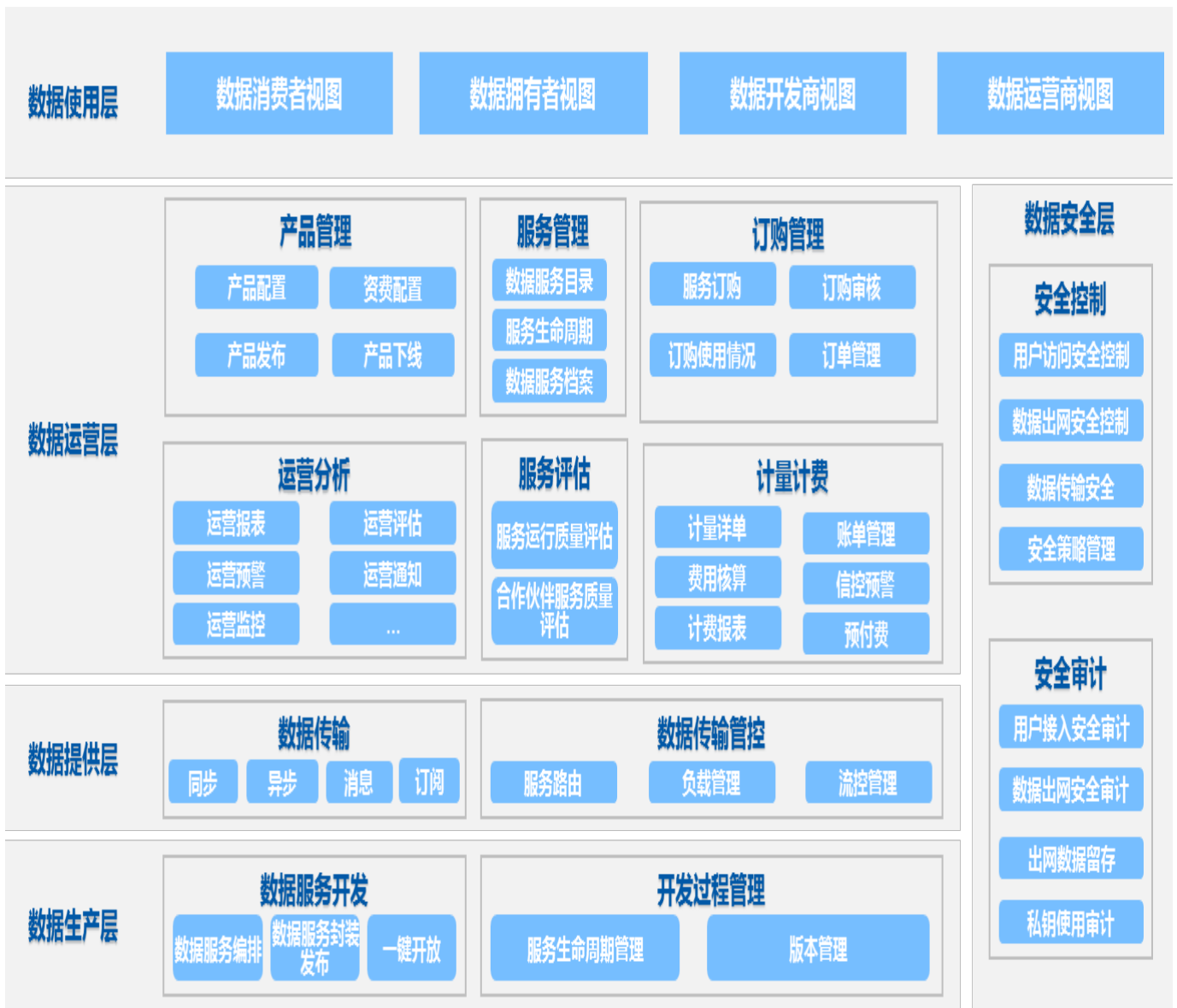


图4 AISWare DataInfrastructure-CI产品架构

[返回目录](#)

4.2.5.1. 服务生产层

服务生产层提供对接底层数据平台，提供可视化编排方式进行服务开发，并对服务生产全过程进行管理。

服务生产能力：服务生产层对接RDB、Redis、HBase、Hive等数据平台进行服务封装。提供服务代理、SQL查询、文件封装、Redis查询、数据脱敏、参数转码等多种封装组件进行服务开发；提供可视化流程编排的方式进行服务开发生产；提供便捷的数据一键开放，大幅提高服务开发效率；提供多个API进行组合开发，实现不同API之间的联动。

服务开发过程管理：提供对服务开发、测试、上线、暂停、下线、版本管理等服务开发过程的管理。

4.2.5.2. 服务提供层

服务提供层面向于数据消费者，采取多种方式提供数据服务，并保障服务调用的高效、稳定和安全。

数据服务提供方式：支持同步、异步、消息、订阅等多种服务提供方式。

数据服务传输管控：数据提供的过程中实现对服务的鉴权、路由、负载、流控、配额等的统一管理。

数据服务安全管控：提供数据开放服务的同时可对数据进行数据脱敏、访问控制、内容级出网校验等以保证出网数据的安全性。

4.2.5.3. 数据运营层

数据运营层提供对用户、服务的统一管理、授权与审计，针对有变现需求的客户提供产品管理与计费运营等功能。

服务运行生命周期：提供的服务的注册、授权、监控以及服务生命周期管理功能。

产品管理与计费运营：提供数据服务产品运营管理功能，包括产品管理、资费管理、订购管理、服务计费、运营分析等功能。

用户管理：提供用户注册、开通、信息维护、应用分配、用户审计等用户管理功能。

[返回目录](#)

4.2.5.4. 服务视图层

用户登录后可查看服务视图层功能。包括可共享的共享数据资源目录、数据服务目录以及在售的数据产品。

共享资源目录：提供数据共享资源的注册，共享资源的详情展现等功能。用户登录后可查看平台提供哪些可共享的数据资源。

数据服务目录：提供当前可以使用的数据服务的查看，包括服务的基础信息、调用方式、调用参数、返回示例和授权情况等。

数据服务市场：针对数据服务对外销售提供数据产品的详情展示，以及产品包含的服务、计费方式、订购方式等。

[返回目录](#)

4.2.6. 指标管理工具架构

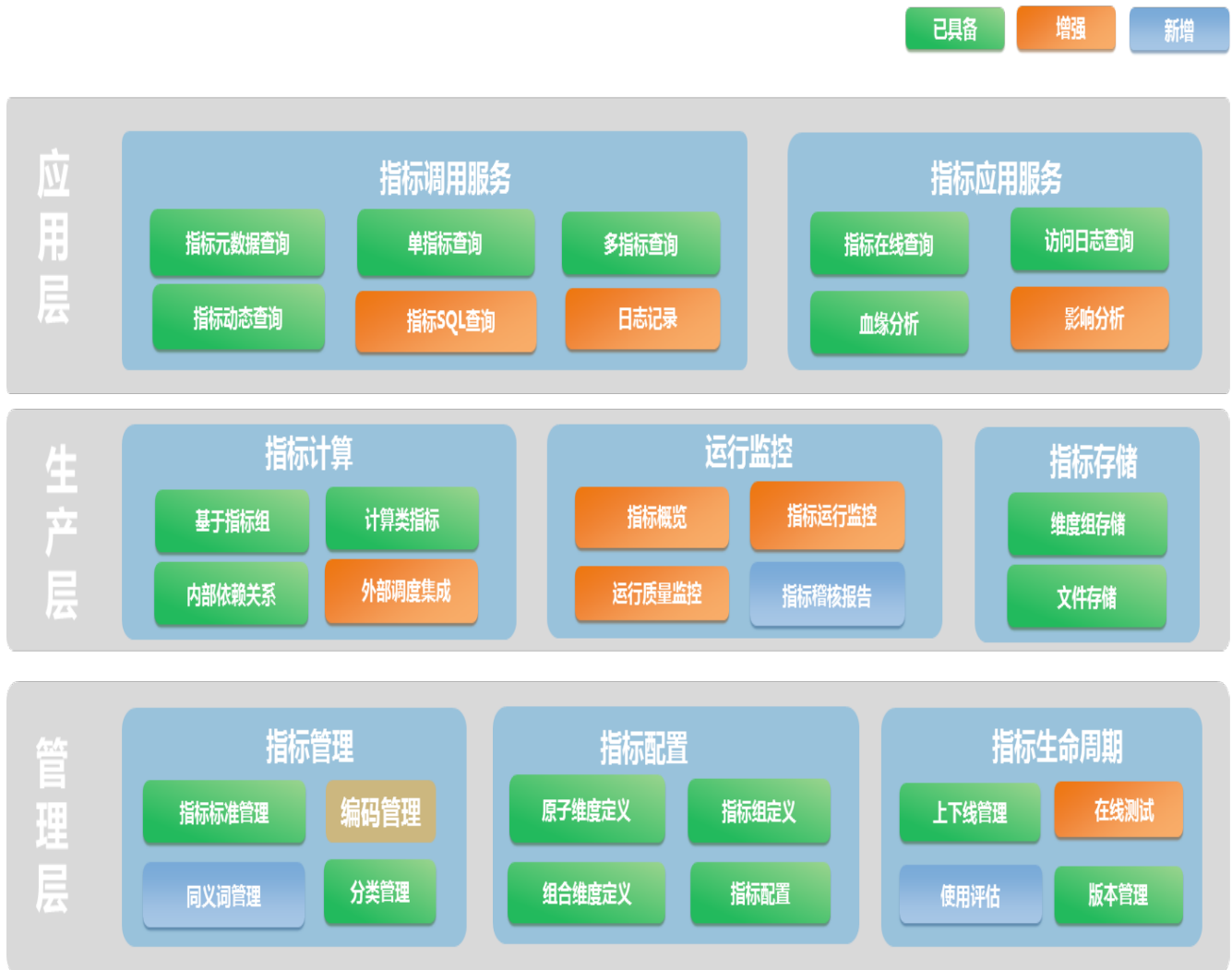


图4 AISWare DataInfrastructure-CI产品架构

[返回目录](#)

4.2.6.1. 指标管理

指标标准管理（逻辑指标）：指标标准主要是对指标的业务口径进行定义。

指标标准维护的内容包括：逻辑指标编码、逻辑指标名称、业务口径、指标单位、备注等。标准新增时，会基于同义词、近义词等对新增指标进行提示，避免出现重复指标的增加。

指标分类管理：支持对指标的分类进行管理，指标可以归属于自定义的指标大类，同时可对指标进行定义其标签。

基础信息维护包括编码管理、同义词/近义词管理。

编码管理：提供指标组编码规则、指标编码规则、指标周期编码规则、指标类型编码规则管理。可对编码的顺序，默认的编码字母进行调整。

同义词/近义词管理：提供同义词维护、近义词维护的页面，可将一对词组设定为近义词或同义词。同义词/近义词主要用于在指标标准建立、指标搜索时展示相关指标。

4.2.6.2. 指标配置

指标配置包括维度定义、指标组定义与指标定义。

维度定义：维度定义分为原子维度定义、组合维度定义两项。

原子维度即单维度，组合维度即用原子维度组合而成的多维度。

所有的指标的维度均基于组合维度选择，可以用一个原子维度构建组合维度。

指标组定义：指标组是一组指标生成的集合，其主体为SQL查询，该SQL用于计算一组可以从一个SQL计算出来的指标。

指标定义：指标定义基于组合维度、指标标准以及指标周期来进行组合。

指标的基础信息包括：组合维度、指标标准、指标周期、指标编码（自动生成）、指标名称、指标计算类型（基于指标组计算、基于其它指标计算、基于其它指标在使用时计算）、优先级、数据源、指标SQL(或表达式)、技术口径、备注。

4.2.5.4. 指标生命周期管理

上下线管理：上线流程包括指标的新建、开发、测试、申请、审批、发布等过程，只有上线的指标才会进入运行状态。下线管理包括下线申请、下线审批、下线影响分析等流程，处于下线状态的指标会停止运行。

在线测试：可对开发过程中的指标进行在线测试，对测试运行的结果进行在线反馈，只有在线测试通过的指标才能进行上线。

使用评估：通过指标API调用情况等信息进行汇总，展示指标被调用的次数。

版本管理：记录指标的版本信息，可对制定版本进行对比展示其变化内容。

4.2.5.4. 指标生产

指标生产包括指标概览、指标计算、指标监控、指标存储、血缘分析。

1、指标概览：统一展示指标的定义情况、指标运行情况、指标质量监控、指标服务调用情况等。

2、指标计算：分别提供指标组计算、指标计算的脚本执行命令，用于嵌入到调度系统中进行调度执行或者手工执行。基于指标组计算：需选择用于指标计算的指标组、指标SQL。此类指标的生成依赖于其对应的指标组是否成功生成。基于指标静态计算：需填写指标生成的表达式。此类指标的生成依赖于其相关的其他指标计算完成，并将基于表达的计算结果写入数据库。该类型适用于计算结果不会丢失维度、且能直接累加的指标。基于指标动态计算：需填写指标生成的表达式。此类指标在指标调用时进行计算，计算结果不写入数据库。特殊计算说明：在传统的四则运算的基础上，内置了部分特殊的计算场景。调度系统集成：调度系统直接调用指标组、指标执行命令。指标组依赖于表或者程序生产完成，指标则依赖于指标组生产或其它指标生产完成。

3、指标监控包括指标运行监控、指标质量监控规则、指标稽核、指标稽核告警。

指标运行监控：指标运行纳入调度系统管理，指标运行监控即调度系统的监控功能。

指标质量监控规则：指标质量监控规则设定，提供对单个指标的质量规则进行设定，主要的设定规则包括：波动检查（增长、降低）、值域检查（上限或下限）、平衡性检查、及时性检查。发现问题后，及时展示监控到的当前告警信息，并将错误信息发送至告警集中管理。

[返回目录](#)

指标稽核：支持在页面展示一定时间段的指标质量稽核汇总情况、变化趋势等，形成指标质量稽核报告。

指标稽核告警：运行检查指标稽核规则，对不通过的规则进行集中展示监控到的告警信息，提供标准回调接口方便与现场告警平台进行开发对接。

4、指标存储:包括存储方式与存储位置。其中存储方式为每个维度组指标放在一个表中，存储位置为指标计算在主仓库进行，指标存储到Hadoop文件

5、血缘分析:指标库内的血缘关系为计算指标与指标间的血缘关系，原子指标与指标组间的血缘关系，指标组依赖的表的血缘关系。指标库的影响分析是指，某一指标影响的其他计算指标，以及该指标影响的应用，如报表信息等。

4.2.5.5. 指标应用

1、指标搜索：基于分词、词性、同义词、近义词、模糊匹配搜索指标标准等。指标搜索页展示相关的指标标准信息，并展示最匹配的指标的数值、波动、近期变化值等信息。匹配优先级按模糊匹配、同义、近义、词性递减。

2、指标展示:支持单个指标查询、指标近期波动、指标排名等。

3、指标推送:支持用户订阅指标服务API，指标数据生成后会将指标数据定期推送到用户指定的地址（包括FTP、数据库等）。可将指标数据，全量或者部分分发到指定的数据库，方便etl或应用开发人员进行数据核查和使用。

4、多租户指标分享:指标服务权限管控：支持租户级的服务调用权限管理，包括服务授权、授权申请、授权回收等。支持多租户环境下的指标定义、开发与计算。

4.2.5.6. OPENAPI服务

在指标层提供基于指标的OpenAPI封装能力，通过API接口对外提供指标查询、数据供给等服务。

[返回目录](#)

4.2.7. 数据安全中心架构



图4 AISWare DataInfrastructure-CI产品架构

[返回目录](#)

4.2.7.1. 安全策略

整体定义企业的数据安全策略，根据企业的数据使用情况和业务构成，建立符合企业当前和未来一段时间其发展所需的数据安全体系和标准规范，例如分析企业数据使用场景、开放程度、数据资产安全级别、个性化安全组件等等。具体包括数据保护策略、规则库、流程控制、管理制度、账号体系、安全元数据等等。

4.2.7.2. 敏感数据定义

敏感字段规范定义：定义企业内部敏感数据的业务定义、安全级别、存储长度、业务范围。

静态密钥管理策略：对定义的敏感数据设置统一的加密策略。同时基于定制的伪随机数生成器辅助生成加密密钥。基于性能考虑目前只支持两种加密算法：可逆算法：AES；不可逆算法：SHA1

4.2.7.3 .敏感数据识别

敏感数据识别规则定义：敏感数据平台提供多种数据识别方式，识别方式也可以调整优先级。

正则表达式：基于正则表达式的数据识别，用于识别特定形态数据格式。例如手机号、imei等。

类SQL表达式：便于熟悉SQL语法的用户，产品同时提供类SQL like语法用于实现数据识别。

样本定义：对于企业内部特有数据进行样本提取，使用不可逆加密方式构建在内存数据库中。用户可以使用样本数据对数据中心内的数据进行识别。例如：家庭住址、用户姓名等，样本抽取任务可以通过配置实现，支持单次、每日、每月。

4.2.7.4. 敏感数据加固

对于关键信息的管控推荐使用静态加密的方式，数据中心可以设置独立的安全租户实现数据的加密工作。推荐关键数据有手机号、证件号、用户姓名、家庭住址、imei。同类型数据需要使用相同的加密密钥，保证后序的关联计算不受影响。

4.2.7.5. 敏感评估

数据使用评估，评估敏感数据各项指标，包括：

- 敏感数据分布：评估敏感数据在数据中心和租户之间的分布情况。
- 敏感数据使用：评估敏感数据在各个租户、用户、程序等使用情况。
- 敏感数据健康评估：评估数据中心内加密与未加密数据比例，加密数据表使用频率。

敏感数据监控，监控敏感数据使用：

- SQL日志采集：基于jobhistoryserver和yarn的SQL采集，动态密钥使用日志。
- 异常监控：通过动态密钥,监控用户敏感数据是否调用异常。
- 动态密钥使用跟踪：跟踪密钥使用情况,使用场景、租户、敏感字段。
- 敏感数据使用跟踪：跟踪用户密钥加解密。

[返回目录](#)

4.2.7.6. 密钥管理

动态密钥:动态密钥是针对用户数据开发时需要加解密数据，不同用户使用时，分配不同私钥解密，且能还原出相同数据；密钥被他人盗用，无法解密。针对具体场景，可一人一密、一时一密、一数一密。

密钥配送：根据密钥接收者的公钥采用非对称加密技术对静态密钥进行加密。

密钥完整性签名认证：为方式传输过程中，信息被第三方截取篡改，需要对CEK信息进行数据签名。

密钥分配日志：记录用户对密钥的申请信息，及密钥管理对用户的密钥分配信息

密钥分配跟踪：基于日志采集匹配跟踪静态密钥使用情况。

4.2.7.7. 统一接入认证

基于RestApi的开放接口由于是无session状态，为保证接口调用安全，需要对调用URL进行签名验证机制。

[返回目录](#)

4.2.8. 数据资产运营中心架构

数据资产运营地图

数据资产全景视图

数据资产分布视图

数据流向视图

租户运营视图

平台资源运行视图

数据资产评估

数据资产健康度评估

数据资产应用效益评估

数据生产效能评估

数据冷热度评估

数据资产价值评估

数据管理成熟度评估

数据资产监控

数据生命周期监控

数据生产运行监控

数据使用过程监控

数据资产目录

数据资源管理

数据资产目录

数据资产档案

数据资产图谱

数据资产应用集成

数据资产盘点

数据发现

数据摘要

数据自动识别

数据资产分类

数据资产登记

[返回目录](#)

4.2.8.1. 数据监控分析

数据生产质量分析:主要包括数据资产生产的完整性、一致性、有效性方面的质量分析,而且与具体业务相结合,针对具体业务、专题开展数据质量的相关分析。

数据生产效能分析:主要包括数据资产生产的资源投入产出分析,投入包括平台资源投入、开发成本投入、管理分摊投入;产出包括在外部市场拓展、内部能力支撑方面的产出等。

数据运维效益分析:主要包括数据运维的全程全域分析,含数据问题定界定位、作业队列运行状态、故障排除时效等。

敏感数据引用评估:主要包括敏感数据(模型级、字段级)在引用过程中情况分析,包括引用痕迹、引用量级、引用归口部门、引用效果等。

4.2.8.2. 数据评估

整体评估视图:提供功能数据资产整体效能评估功能,聚合需求响应效率评估、敏感数据安全监控,数据流向与交换合理性分析、应用使用效能情况、作业效率监控等评估信息在统一视图进行完整呈现。基于构建公允的数据资产评估体系,通过现状评估与持续周期性反馈相结合的思路,对数据平台能力与能力的综合价值进行跟踪和评价,为应用周期性的投资决策提供参考依据。

需求响应效率评估:管理平台租户的需求提出-响应-开发-关闭等全周期效能评估过程,按需求类别展示需求支撑完成率、需求平均滞留时长(区分曾经挂起需求与普通需求);根据需求频率、处理流程、状态、实施计划等维度来组织后评估模块,关注需求(按业务类别/按需求部门归口)的投入产出效益,凸显需求支撑效率对业务发展的促进作用;根据集中化多租户平台需求支撑的日常汇报形式,将需求后评估各项能力集成到需求后评估指数魔方界面,实现全面、有效、及时的呈现中心需求支撑工作情况。

数据流转与交换评估:展现数据中心采集层,加载层,建模层的数据资产的流转、交换情况。根据数据资产流转与交换的流通效率(时间效率、成功失败占比、队列等候时长等)、流通内容(包括数据规模、模型数量、核心字段、有效标签数量等)、流通效果(支撑应用、专题、服务的情况)三方面的情况来整体评估数据资产的流通效能,为数据交换与开放的关键平台能力提供改善与提升建议。

[返回目录](#)

应用效果评估:展示平台各子应用系统总数、总体使用人数、总体点击次数、总体登录次数,以及应用系统整体预算投入与市场支撑价值等信息。通过共性指标(应用系统基础使用和投资情况)和个性指标(业务承载情况、业务价值评估)两个维度的评估指标体系来呈现数据资产的应用效果。

作业效率评估:根据作业的运行日志分析,监控并评估作业的执行与运行效率(时间与过程),开展作业失败的根因分析、回溯作业的执行过程,具体包括对作业执行状态监控、作业异常定位定界分析、作业前置分析、作业调度影响分析等功能。

4.2.8.3. 数据地图

资产目录定义:1.管理员角色查看全局2.租户角色看自身3.租户累计等于全局数量4.按照数据、平台、应用、管理四方面组织,下级节点可按需配置。

资产目录盘点:1.基于数据目录(系统、主题、层次、租户),再增加行业、业务、应用等属性,构建资产汇集初级视图,实现简单归集2.为数据资产配置提供基础能力(配置可根据栏目点选)。

资产目录配置:1.指标定义(数据口径、单位)2.动态开发场景下调用DACP开发调度能力3.从业务价值、数据质量、数据规模、时间价值等维度定义数据资产价值4.根据资产价值配置增值保值策略⑤考虑数据资产的生命周期5.业务与资产的关系要呈现。

总体链路查看:按照数据架构层次、主题、数据生产流程、数据血缘关系等分层分级的链路结构来查看数据全貌,从不同颗粒度的数据关系来呈现数据规模、质量、使用冷热度、模型及字段关联等关键信息,并通过可拖拽伸缩功能实现数据地图逐层的钻取,针对局部关注数据信息进行查看和分析。支持数据地图逐层钻取和缩放功能,定位数据模型的所处层次、主题、生产流程位置、并能呈现跨层次、主题的数据血缘关系。

选择具体数据模型,可以展示数据链路详情,包括模型使用冷热频率、模型跨域跨主题的血缘关系(表和字段级)等信息。

[返回目录](#)

数据流向地图:1.整体增长：分层统计分析一段时间内的数据增长情况，包括接口数量、数据量、字段总数、调度任务数、调度任务完成情况等数据资产周期变化统计情况。2.数据引入：从省经、一级系统引入数据，基于时间展示数据占比、数据增量趋势等情况；主要包括分层的模型数量、记录总数、数据总量、任务调度完成情况、趋势分析曲线情况；分主题域的子域数量、接口数量、新增接口数量、字段数量等数据引入情况。3.数据处理：对汇总层、中间层数据规模、增长趋势等展示，包括应用、指标、标签、算法分类数量、数据总量、任务调度完成情况、趋势分析曲线。4.应用分析：分类统计展示经分数据支撑报表应用的数量，还包括应用分类与数量、指标分类与数量、标签分类与数量、算法分类与数量等数据资产统计情况。

数据分布地图:1.整体分布分析：统计分析总体数据冷热度，包括模型、标签、视图等数据资产的使用热度、存储利用率等分布和使用情况。2.分层分布分析：展示分层存量数据冷热分布情况，包括各层各主题域中模型、标签、视图、OpenAPI等数据资产的分布和使用情况。3.存量数据分析：统计存量数据情况，包括存储、模型数、实体表数等；支持对存量数据热、温、冷数据的总体统计分布分析。

资产价值地图:展示资产价值总体情况（包括开放能力/接口分类数量、开放能力/接口/标签调用累计、合作伙伴数量/内部客户数量等）、对内支撑效果（包括营销成果排名情况，营销成功率、维系挽留率，收入提升情况；热点模型/标签排名情况）、对外服务效果（包括热点模型/标签排名情况、服务调用排名情况（合作伙伴）、标签调用趋势（合作伙伴））等。

4.2.8.4. 数据知识图谱

数据资产图谱实体模型:1.模型实体：是指数据中心对数据表的抽象表示。模型是数据中心中最重要的实体对象之一。模型实体包括模型、实体表、字段等子实体。2.程序实体：主要是指程序主要是指数据中心运行的ETL程序。包括程序元数据、业务口径、程序名、参考口径等。3.接口实体：包括数据文件接口、分发接口等。模型内容包括接口元数据、接口类型、接口数据量等。4.数据源：是指为数据中心提供数据来源的源系统。包括数据源名称、数据源类型等。5.指标实体：是指指标是数据中心通过汇总，形成的低维度的数据指标。内容包括指标编码、指标类型、指标口径、指标标准等。6.实体关系图谱：资产关系是指将利用关系来描述各类资产实体的之间的对应的关联关系。7.实体事件图谱：实体事件是用于描述实体的动态发展变化信息，体现事件与实体间的关联，并利用时序来描述事件的发展线条。

[返回目录](#)

数据资产图谱构建:1.采集装载：自动采集是指通过约定的采集规则，对元数据管理系统中的数据实体信息进行定时采集以获取最新的数据信息。2.数据转换：目前系统元数据一般采用关系数据库方式进行存储，数据从关系数据库到图数据库需要提供数据转换与映射工具，以保证关系数据结果到图数据结果的快速、准确转换。3.接口实体：包括数据文件接口、分发接口等。模型内容包括接口元数据、接口类型、接口数据量等模型构建：数据资产图谱模型的构建即对整个知识图谱的结构进行定义，并将相应的数据采集、转换、装载到模型中。

数据资产图谱应用:1.资产实体查询：通过数据资产实体的关键属性来查询资产的相关信息，并同时展示该资产较受关注的关系。用户可以通过相关的实体继续点击展开其资产关系。2.资产关系查询：通过资产查询其相关的关系，可对关系的类型进行裁剪，可设置关系的层次。资产关系采取关系图的方式进行展现。3.事件查询：通过资产实体或属性查询其相关事件，及其关联实体的相关事件。资产实体的事件以时间轴的方式进行展现。

4.3关键技术

[返回目录](#)

4.3.1 元数据驱动开发

元数据（Metadata）是描述数据的数据，可将其按用途的不同分为两类：技术元数据（Technical Metadata）和业务元数据（Business Metadata）。数据的标签、说明，数据库字段、字段说明，数据的业务理解、业务定义，数据加工过程的业务说明和技术参数等都是元数据。在大数据环境下中，元数据管理的复杂度和重要性远高于传统的数据仓库。

元数据驱动的开发模式，采用以数据模型为主体，把元数据、数据处理、数据质量管理等以模型的方式融合在元数据血缘调度里统一处理，实现业务元数据和技术元数据的强制性前向获取，建设保证业务元数据/技术元数据/技术实现这三者一致性的体系。

4.3.2集成开发环境IDE

集成开发环境（Integrated Development Environment，IDE）是一种辅助开发人员开发软件的应用软件。开发服务平台为开发人员提供统一的集成开发环境，让开发人员可以通过图形化界面完成数据采集、存储、应用逻辑等的开发。集成开发环境集成了组件调用、代码编写、编译、调试功能等一体化的开发应用服务套件和图形用户界面工具，实现各租户的应用设计、开发、测试、部署、运行、监控和管理的全流程支持。通过集成开发环境可进行数据建模、数据处理、数据流及调度作业配置，完成开发调试、测试上线、运行监控及预警等开发流程。

4.3.3 技术平台能力的封装

数据开发将通用的能力下沉，以方便调用。从数据处理和数据应用两个方面，对底层技术能力的封装，就可以支持基于IDE的快速开发，也可以支持采用脚本进行更为复杂的类型开发。

4.3.4基于Lambda架构的流批一体化开发

Lambda整合离线计算和实时计算，集成了Hadoop，Kafka，flink，redis，Hbase等各类大数据组件。实现了真正的流、批协同处理，同时统一了编程接口，使得整体处理架构简单化。

[返回目录](#)

4.3.5 去中心化分布式调度架构

平台采用去中心化架构，实现集群高可用和负载均衡能力，避免因节点单点瓶颈引发故障，通过节点机争抢方式，减缓了运行队列任务挤压，提高了任务运行速率。

采用去中心化架构：屏蔽Master-slave引起的单点故障，集群支持scale-out横向扩展。

容错处理FailOver：整个集群中所有的节点的角色都是相同的，地位是平等的。集群中任一节点down机，它的工作会被其他节点接管，对整个集群几乎是没有任何影响。

心跳检测机制：采用zookeeper临时节点技术，检测broker节点健康状态，自动接管异常节点任务；同时任务运行定时汇报心跳状态，超时自动处理。

4.3.6 数据安全管控能力

传统安全视角，更多关注平台级安全（网络、设备、防入侵等），而较少涉足业务级安全（如经营风险、业务保障、过程管控等）但在数据安全问题上，更为值得关注的反而是业务领域（数据可靠性/有效性、脱敏、去隐私等），需打破传统藩篱，实现跨域协作发展。在数据加工链上涉及多个环境和技术形态，需要首先梳理需要具备哪些能力用于安全管控和审计。

4.3.7 数据查询安全管控

SQL解析控制数据安全透明访问，即使用者按照正常的sql语句进行提交查询，同样的SQL语句，能够根据不同的使用者返回不同的数据结果。

输入权限控制：根据提交的SQL语句，解析其所用到的表、字段。

权限管控：根据数据拥有者的表和字段授权使用情况，检查当前使用者是否有权限使用

输出安全控制检查：数据拥有者允许数据某些属性可以做关联查询，但不愿意被输出。需要对输出字段的安全进行权限检查

SQL改写：因为查询的脚本是明文的，但查询的数据是加密的，因此需要对明文的sql中的关联条件、比较条件进行同态加密提交，才能得到正确的结果。

4.3.8 指标维度组存储设计

将同一维度组合的指标数据存放在一个实体表中，不采用以往指标库通常将指标存储在一张表的设计方式，通过该方式实现了指标维度的灵活组合。

[返回目录](#)

4.3.9 指标维度组存储设计

将同一维度组合的指标数据存放在一个实体表中，不采用以往指标库通常将指标存储在一张表的设计方式，通过该方式实现了指标维度的灵活组合。

指标数据存储在文件中，每个文件存储一个维度组、指标、时间周期的数据，文件存放在HDFS文件系统中以保证文件的快速读取，同时降低对同一文件/表的同时读写而导致的锁死。传统的指标结果一般存储在关系数据库或者HBase中。存储在关系数据库中，由于指标大量的读取和存储，有较大的概率导致锁表，从而影响指标运行及时性。由于HBase的表设计限制，更适合支持固定维度指标的设计，从而导致可支持的应用大量减少。

4.3.10 数据安全管控能力

数据评估是对数据管控成果的集中体现，并总体评估数据管控的成熟度。通过对数据的集成度、分布、冗余度、数据热度、重要性等多种角度实现数据评估，并基于评估结果对数据进行相应的处置。

集成度评估：计算公式为接口单元数/DWD层模型实体数。值越高，说明对来自不同系统的接口的进行了越高的集成。

数据分布评估：获取经分域所有系统数据，按关系数据库、hadoop，数据所属层次统计数据存储空间。

冗余数据评估：对冗余数据进行清理，提高数据质量。

热数据发现：根据数据的使用频率来判断数据热度。

重要性评估：根据访问频次和业务需求角度出发进行数据重要性评估。

五. 功能介绍

5.1. 基础功能

一级功能点	二级功能点	功能点描述
数据治理中心	元数据管理	通过元数据采集汇聚，形成“物理分散、逻辑统一”的统一元数据管理，保障元数据日常运营，并提供元数据查询、血缘及影响分析等对外服务能力：包括元数据注册、元数据采集、元数据解析、元数据稽核、元数据分析、元数据维护、元数据查询、元数据服务、业务元数据自动补充
	数据架构管理	通过数据架构管理使庞大的数据更为清晰和有条理，简化异构数据环境下的存储，高效的管理海量数据。包括数据架构、存储架构、业务流程
	数据标准管理	通过统一的数据标准制定和发布，结合制度约束、系统控制等手段，实现企业级大数据平台数据的完整性、有效性、一致性、规范性、开放性和共享性管理，提高企业级省大数据平台数据治理水平。包括标准制定、标准发布、标准绑定、标准稽核、标准评估、标准服务、标准废止、标准库自动构建
	数据质量管理	数据质量管理支撑对数据时效性、完整性、准确性、一致性的保障要求。包括质量分类管理、质量对象管理、质量规则管理、质量稽核、质量监控、质量总览、质量规则自动设置、质量阈值自动调整
	数据生命周期管理	通过不同的数据存储策略，实现数据生命周期管理。包括存储规则配置、存储信息监控、数据存储清理、存储分析评估、存储周期自动识别
	主数据管理	基于ID识别技术驱动数据连接，以OneID的形式对内外部系统提供统一、高质量、高价值的主数据服务。主要包括：主数据设计、主数据映射、主数据集成、主数据服务

[返回目录](#)

一级功能点	二级功能点	功能点描述
数据开发中心	数据开发	提供统一开发平台，保障开发过程规范化、流程化。包括数据开发IDE、取数操作IDE、多租户管理、资源管理、服务管理、程序管理、流程管理
	敏捷模型交付	通过可视化设计图串接DevOps能力，构建贯穿模型设计、程序开发、流程测试的可视化数据建模工具，实现多人协作迭代完成设计与交付。包括需求管理、模型设计、模型测试、模型发布
	任务调度	支撑跨平台、跨系统、租户隔离的统一调度管理。包括任务配置、运维监控、执行引擎、集群管理、对外服务
	采集与交换	支持丰富多样的数据采集与交换方式。包括单元管理、运维监控、执行引擎、交换组件、集群管理、对外服务
	批流一体处理	支撑实时开发能力，支撑应用实时性需求。包括实时采集、实时编排、执行引擎、实时调度

[返回目录](#)

一级功能点	二级功能点	功能点描述
数据交换中心	单元管理	主要功能包括单元配置、单元生命周期管理、模版管理、全局变量管理
	运维监控	主要功能包括单元监控、单元运维、资源监控、告警管理
	执行引擎	主要功能包括单元生成、单元编译、分布式执行框架、资源规划、交换执行、健康检测、异常处理
	交换组件	主要功能包括cli组件、pipeline组件、控制组件
	集群管理	主要功能包括集群配置、集群监控运维、多租户集群支持
	对外服务	主要功能包括单元监控运维、单元创建管理

[返回目录](#)

一级功能点	二级功能点	功能点描述
敏捷建模工具	需求管理	主要功能包括待办流程、发起流程、流程查询、流程处理、需求详情、任务管理、需求进度
	模型设计	主要功能包括流程图设计、概念模型设计、物理模型设计、程序设计、多人协作建模、敏捷模型版本管理
	模型测试	主要功能包括测试用例设计、测试用例执行
	模型发布	主要功能包括模型打包上线、生命周期管理、模型知识发布

[返回目录](#)

一级功能点	二级功能点	功能点描述
数据开放中心	服务开发	提供高效、便捷的数据服务开发环境，支撑将数据以服务接口方式对外。提供包括服务编排、服务测试、服务组合、服务生命周期
	服务提供	支撑数据开放所需服务。包括服务组件、服务接入、服务负载、服务计量、服务日志
	服务控制	提供服务过程的管控。包括流量控制、安全控制、返回内容控制
	服务监控	实现服务的监控，包括调用日志、服务分析
	服务运营	根据服务使用情况进行服务的运营。包括服务上架、服务目录、服务订购、个人中心
	计费帐务	支撑计量计费功能，包括资费配置、计费处理、账务处理、账务报表

[返回目录](#)

一级功能点	二级功能点	功能点描述
指标管理工具	指标标准	主要功能包括指标规范管理、指标分类管理
	指标管理	主要功能包括原子维度管理、组合维度管理、指标组定义、指标配置、指标生命周期
	指标计算	主要功能包括指标处理、指标存储架构、指标血缘分析
	指标质量	主要功能包括指标质量监控、指标稽核告警、指标质量自动发现、指标质量报告
	指标访问	主要功能包括指标查询、指标应用API服务、指标订阅推送、查询日志记录

[返回目录](#)

一级功能点	二级功能点	功能点描述
数据安全中心	敏感数据管理	支撑敏感数据的全过程管理，包括敏感数据定义、敏感数据识别、敏感数据加固、敏感数据开发管理、敏感数据评估、密钥管理
	数据安全网关	包括安全JDBC、服务网关、终端网关、数据库连接客户端
	安全网关审计	包括日志处理、日志审计
	安全基础组件	包括安全基础管理、密钥组件、安全组件

[返回目录](#)

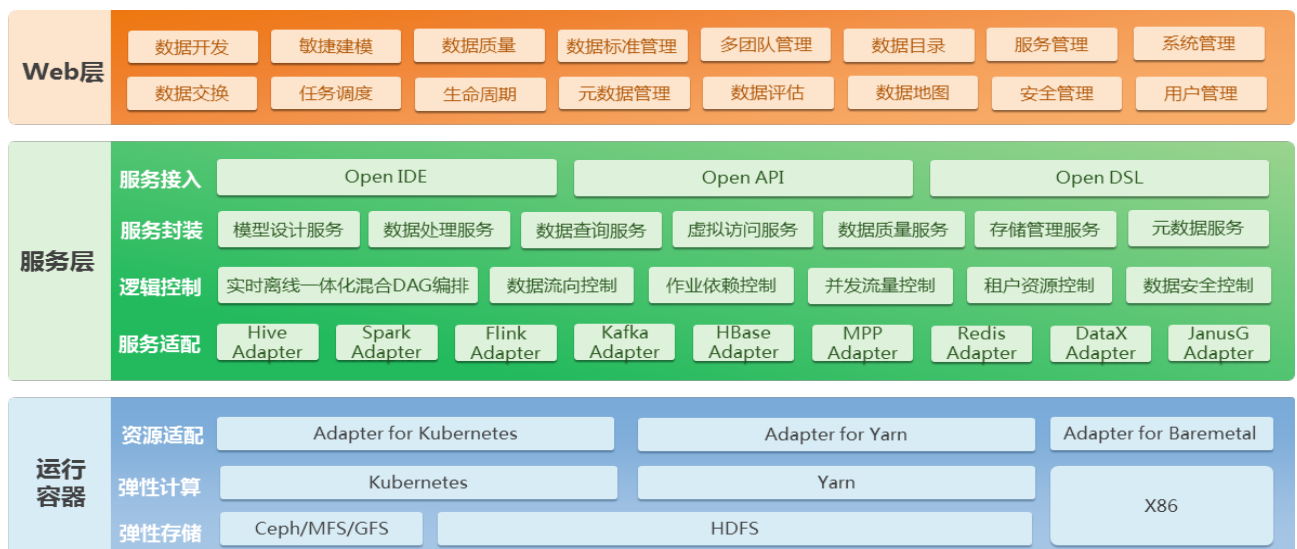
一级功能点	二级功能点	功能点描述
数据资产运营中心	数据目录	包括数据资源目录、数据资源管理、私有数据目录、数据资产目录、数据资产盘点
	数据评估	包括数据资产健康评估、数据资产运营评估、数据管理能力成熟度评估、数据资产价值评估
	数据地图	包括数据链路地图、数据流向地图、数据分布地图
	模型超市	包括模型分享、模型超市、模型移植
	数据图谱	包括数据资产图谱模型、数据资产图谱构建、资产图谱查询、资产图谱应用

5.2. 特色功能

5.2.1. 大数据多租户能力模型

产品技术架构向能力解耦、微服务的技术架构进行演进。在运行容器提供对弹性计算与存储不同资源的适配；服务层中提供统一的服务接入，并封装原子服务向外提供能力支撑，逻辑控制层提供混合DAG、数据流向依赖、并发资源与安全等多类控制；Web层也容器化由K8S提供部署服务。

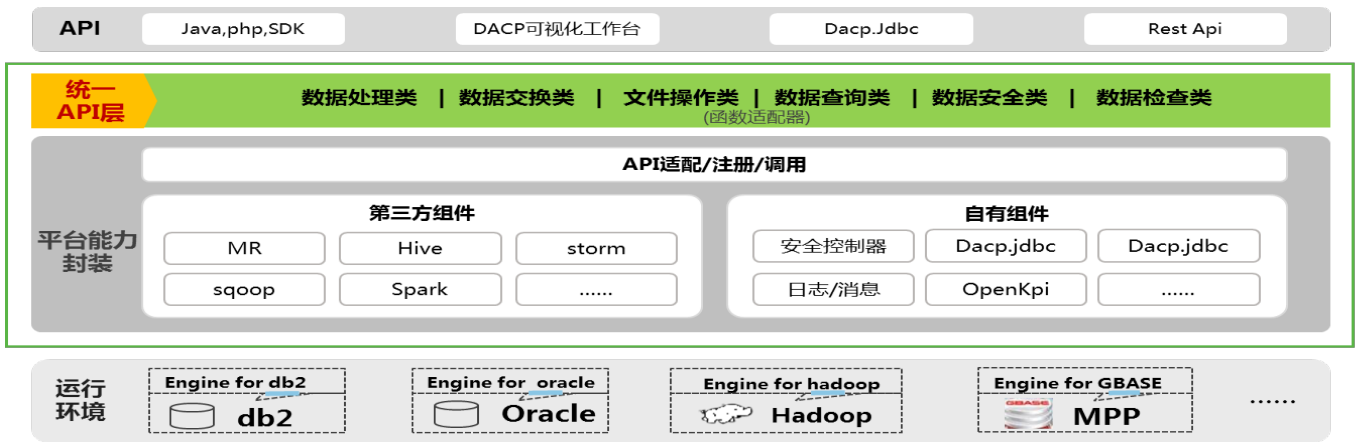
- 全容器Web服务：全容器化部署WebServer。
- 前后台分离：代码前后台完全分离。
- 服务接入：统一服务接入，支持不同类型。
- 服务封装：能力解耦封装原子服务。
- 逻辑控制：抽象逻辑控制层实现不同类型控制。
- 服务适配：插件式方式集成不同大数据能力。
- 资源适配：解耦并适配不同的弹性计算与存储资源。
- 储算分离：计算与存储完全分离架构。



[返回目录](#)

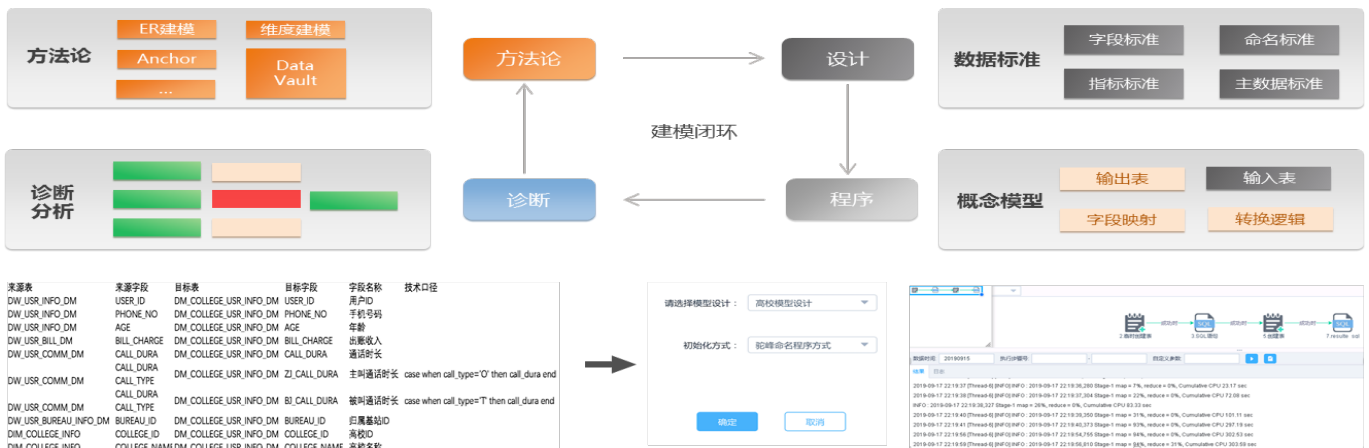
5.2.2. 开源及第三方快速集成

拥有统一封装函数库，能屏蔽底层差异性，具备很强的跨平台统一开发能力；强有力的API管理能力，数器坊可快速集成开源软件及其他第三方软件，并对其进行统一管理和运行监控。



5.2.3. 处理脚本自动生成与优化诊断

通过在线设计模块，填写来源表、目标表字段设计及技术口径，自动生成程序样例脚本，简化开发。



[返回目录](#)

5.2.4.云端的数据开发环境

提供云端开发环境，开发者受限于难以在本地模拟生产环境。通过Web直接访问WebIDE，实现随时随地开发。基于容器的服务化Web应用开发，保障快速搭建一致的开发环境。



5.2.5.一体化DSL数据处理与访问

提供云端开发环境，开发者受限于难以在本地模拟生产环境。通过Web直接访问WebIDE，实现随时随地开发。基于容器的服务化Web应用开发，保障快速搭建一致的开发环境。

基于Flink引擎打造全域类SQL语法的DSL，统一库内和库外数据处理形态，使得非结构化数据和非数据库数据达到统一的数据处理体验。

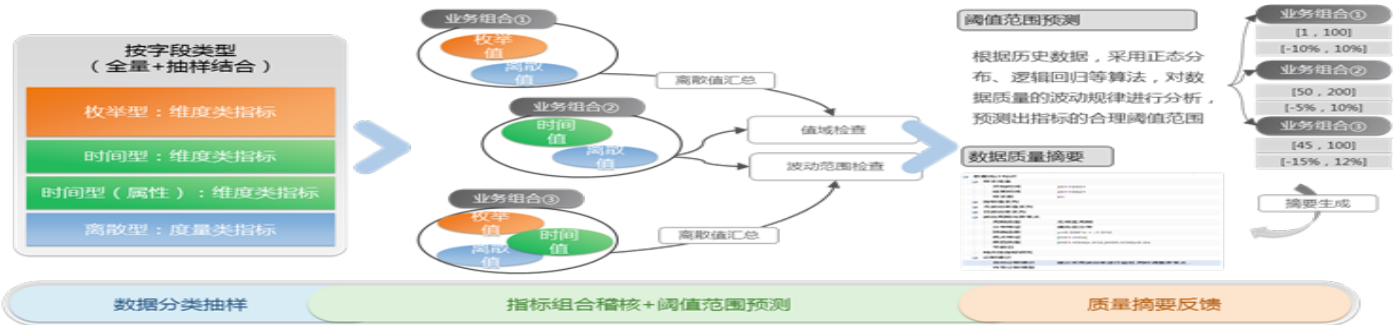
- Catalog：替换hive数据库的catalog为DataOS的数据目录。
- Yaml Script：自研基于yaml定制脚本语法，实现数据处理过程中的AST执行树，并对接flink的原始核心api
- DSL：在原有flinkSQL的语法上进行了扩展，实现了（where insert流复制、case insert流路由、merge into流更新、create table可扩展的数据对象集成）
- 表类型：在原有flinksql表定义的基础上增加pipe内存表的概念，实现复杂业务流数据的中间状态转换，从而不依赖外部存储。



[返回目录](#)

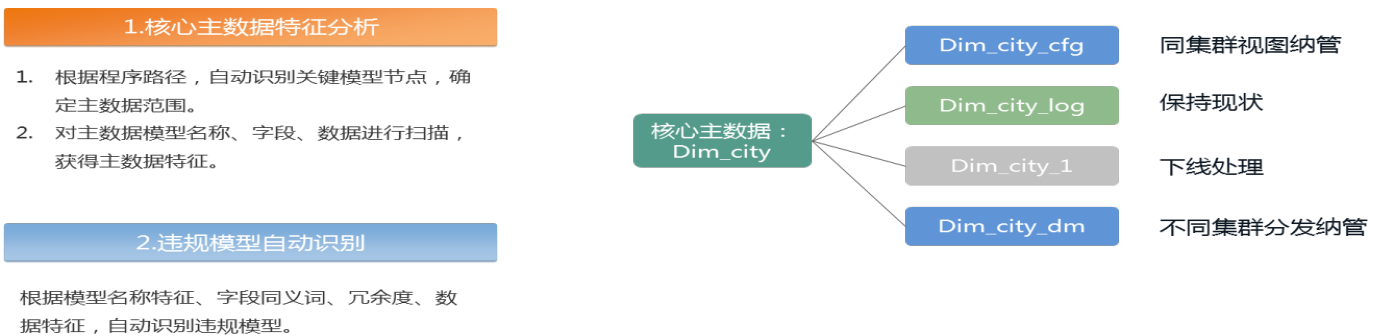
5.2.6. 自动化和智能化的数据质量稽核

通过分析数据字段内容摘要（抽样结合全量的方式），将枚举维度与时间维度自由组合，预测目标阈值是否在合理波动范围内，以达到质量问题的自动扫描，减少人工配置，实现高效感知-反馈-改善。



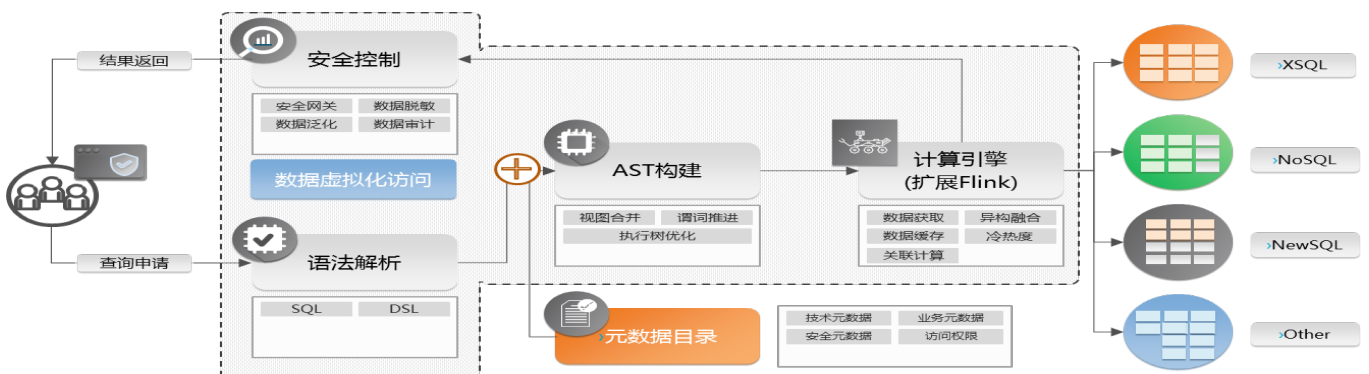
5.2.7. 智能识别的主数据管理

基于主数据智能识别，分析主数据模型特性与数据摘要，自动识别违规模型。对违规模型逐步采取纳管措施，最终下线、或合并，提高模型效率。



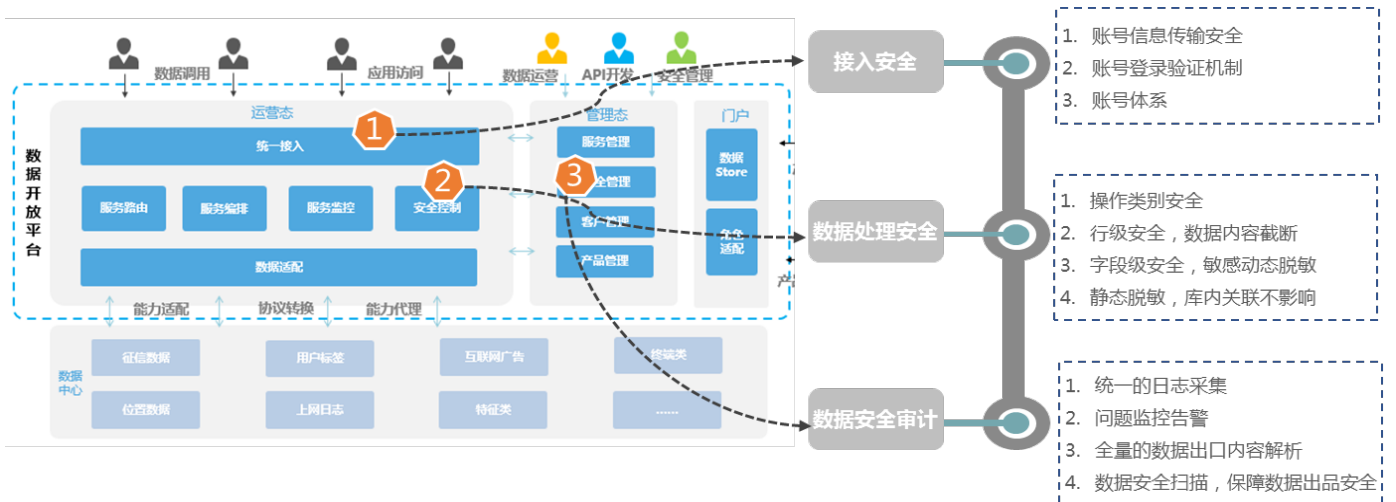
5.2.8. 数据虚拟化访问与安全控制

数据虚拟化访问与元数据目录（MetaCatalog）服务协同，实现各类异构外部数据的注册，帮助客户通过传统工具访问异构跨域目标数据。



5.2.9. 全方位管控的数据开放安全

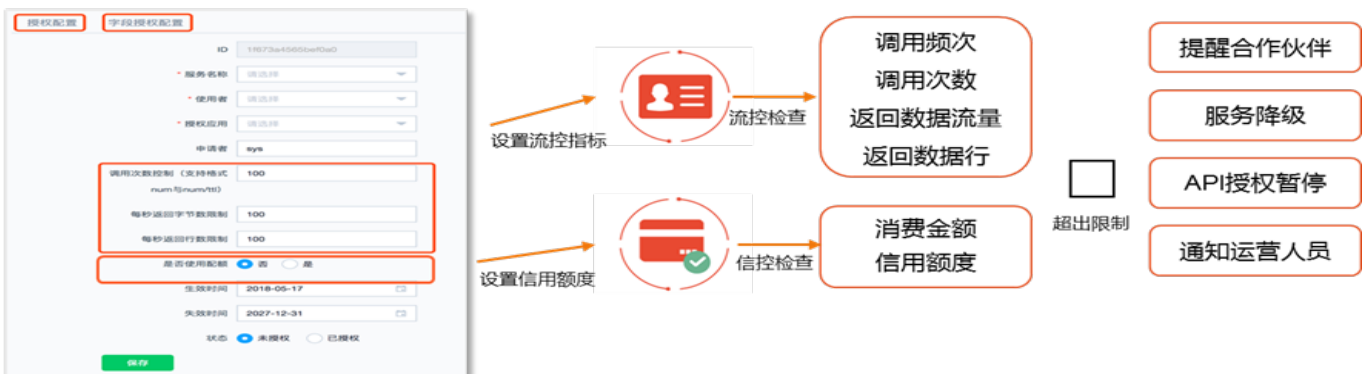
从接入安全、数据处理安全、数据安全审计三个方面，覆盖事前、事中、事后的全过程管控。



5.2.10. 服务的流控与信控

API流量控制能力：API在对不同合作伙伴、应用进行服务授权时，可设置其访问次数、频次、返回数据流量以及行数，降低服务器的峰值压力，避免安全隐患。

消费信控能力：合作伙伴在消费超出信用额度时，可临时暂停其服务授权，并通知合作伙伴及运营人员，以便及时进行跟踪处理。



[返回目录](#)

六. 带给客户的价值

- **轻松获得实现“平民享用”**：将企业内部所有数据统一处理形成标准化数据，挖掘出对企业最有价值的信息，构建企业数据资产库，对内对外提供一致的、高可用大数据服务。
- **重塑数据化运营和决策流程**：提供数据连接共享、数据服务共享、数据模型共享等多种数据共享形式辅以高效的共享流程，构建以数据化、算法、产品等形成反馈闭环，建立新的反馈机制。
- **着眼数据管理，提升价值效益**：以数据资产为核心，通过数据资源盘点与规范管理，形成企业的“数据资产管理”体系与标准，保障运营过程的合规、风险可控和价值实现。



七. 产品优势

[返回目录](#)

锻造全链路集成开发模式：结合前向数据治理能力实现数据操作过程标准化、流程化，通过集成组件调用、代码编写、调试等一体化的面向数据开发的套件工具，闭环涵盖各类数据开发场景。

灵活普适的对外服务模式：以业务为导向沉淀可被BU复用的能力，通过Data API、Open IDE、Embedded DSL三类面向数据域的标准方式支撑数据能力、工具能力和模型能力的场景化开放。

标准化模型管理支撑能力：基于多年行业理解构建形成全数据资产管理框架体系，以“所见即所得”的设计理念完成模型设计到模型实现流程的标准规范性落地，实现模型管理过程的可度量。

数据服务目录的多样化：基于CWM标准的企业元数据管理的技术架构，依托元数据和图谱关系服务构建面向企业全景的数据目录，实现“多组织、分层级、自定义”的数据字典能力提供。

数据管理的效能评估化：结合国内数据发展的现实情况，通过有型的产品支撑和方法论落实数据成熟度评价体系，构建资产效能评估体系用以指导数据生命周期管理策略落地，为企业数据管理改进提供指导。

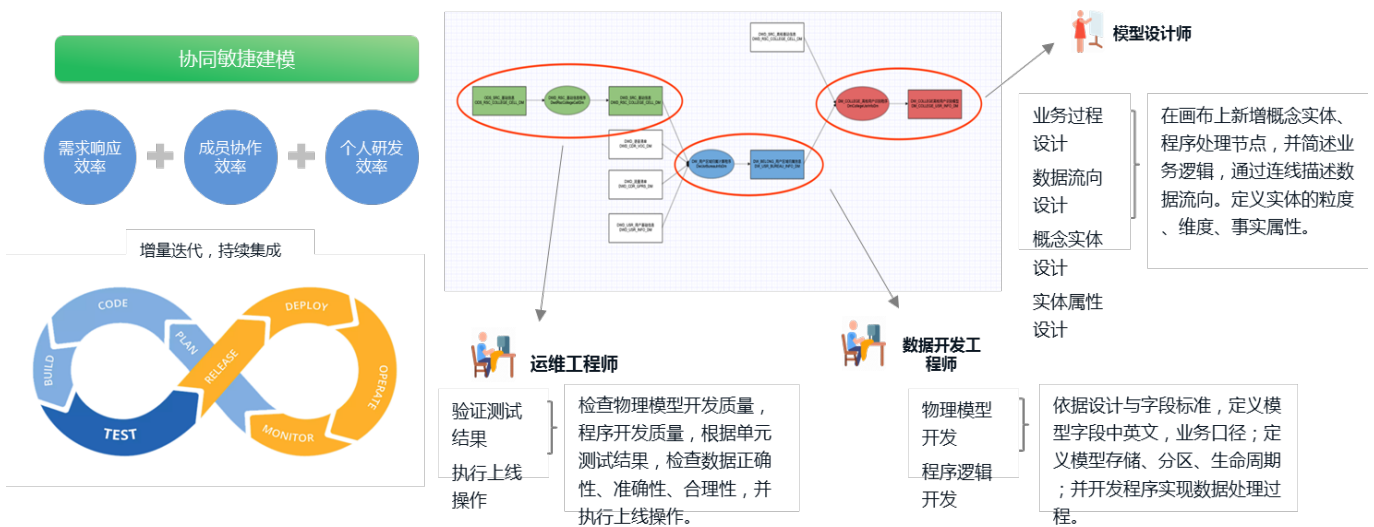
数据治理环节的智能化工具：结合机器学习等技术具体应用到数据治理等环节，构建从基于规则的人工操作到自动操作的处理流程，支撑数据质量、数据安全与作业调度在策略管理与应对处理方面的能力提升。



八. 应用场景

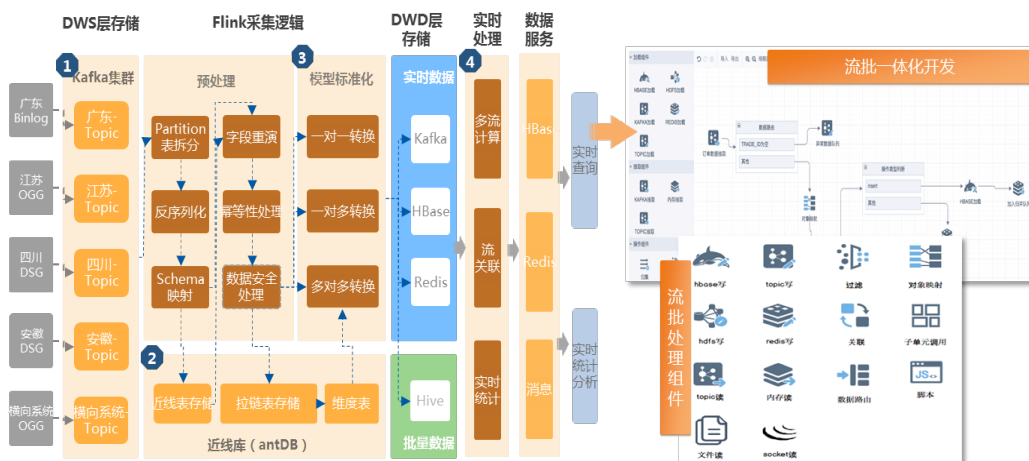
8.1. 设计即开发，敏捷建模协同交付

使用可视化设计图串接DevOps能力，构建贯穿模型设计、程序开发、流程测试的可视化数据建模工具，实现多人协作迭代完成设计与交付。保证从设计、开发到上线、整个链条过程中设计思维、规范文档的全局一致性，所见即所得。



8.2. 实时离线开发一体化，满足实时场景

源端业务系统通过CDC采集与小文件的方式把实时数据生产数据送到集团kafka，在整合层完成31省模型的标准化工作后的实时数据与批量数据的存储存储。在统一开发界面，进行流批数据的协同处理。



8.3. 体系化模型管理，缩短模型价值转化路径

建立整体模型管理体系，实现对全部模型的管控，提升模型效率形成模型知识沉淀

贯申模型全生命周期的智能分析管理能力，从模型的创建、发布、评估、变更、共享到模型的下线、还原，实现模型全生命周期的可视化管控，构建了全面的分类管理体系，实现企业级范围内模型全生命周期管理：

- 基础模型
- 融合模型
- 自建模型
- 挖掘模型

重新定位营销管理人员、建模师和一线营销人员三者的生产关系，简化大数据营销流程，实现大数据建模人员与一线营销人员的直达连接，缩短模型价值的转化路径。

重塑精确营销生产关系

- 传统方式下营销管理人员是所有营销要素的集成者，花费了大量时间在协调沟通等事务型工作上；建模师被动支撑，一线人员等任务；
- 基于获客宝平台，营销管理人员发布营销活动后，建模师成为营销数据的生产端，一线营销人员成为数据的消费端，扁平化双向选择。

缩短模型价值转化路径

· 实现建模人员与一线人员的点对点连接，实现大数据模型直连一线。当前已两个城市试点，累计发布模型200个，模型流转环节从6个缩减到2个，流转时间从原来1-2天减少到1个小时，其中流量营销模型实现了一笔2笔的迭代速度，营销成功率从5%提升到30%。

传统营销生产关系

营销人员发布营销活动

营销人员与建模师沟通

建模师建模

建模师与营销人员沟通

营销人员发布营销活动

➔

新型营销生产关系

营销人员发布营销活动

营销人员与建模师直连

建模师建模

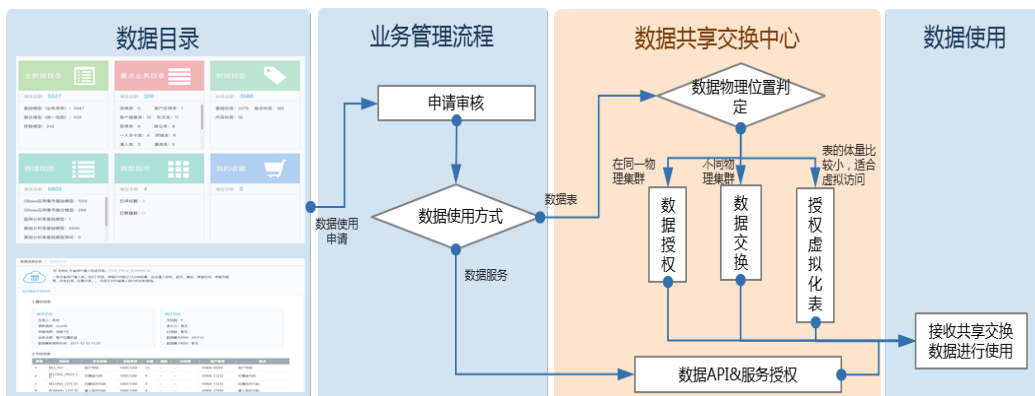
建模师与营销人员直连

营销人员发布营销活动

借助互联网思维助力某产品，将流程环节从6个下降到2个，平均审批时长从2天下降到1小时，效率得到极大提升。

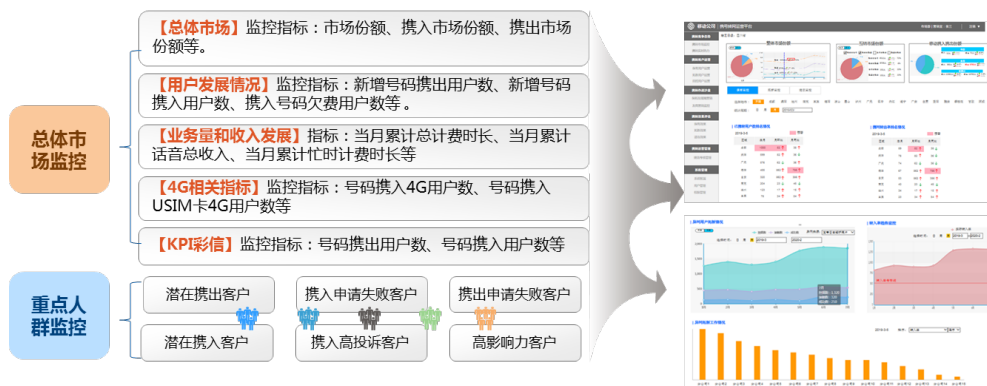
8.4. 基于数据目录的按需数据分发共享，自主应用构建

随着租户的入驻，租户与数据中心之间、租户间存在大量的数据共享需求。通过数据目录提供完善数据服务和共享交换能力，方便租户高效地获取数据信息，打造企业统一的数据共享交换中心。



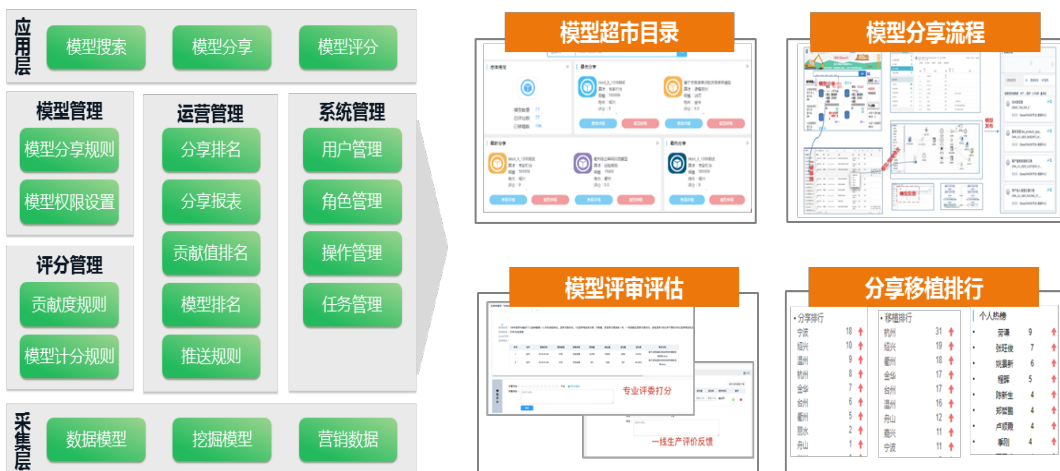
8.5. 构建企业关键指标体系，辅助科学管理、决策

通过对公司战略目标的分解和业务产出结果的提取，制定企业关键的KPI指标，对关键业务指标进行监控，维护，生成经营战报，手机经分，快速锁定识别管理漏洞和经营风险。



8.6. 基于数据目录的按需数据分发共享，自主应用构建

大数据模型超市，打造模型的一体化运营平台，包括模型发布、稽核、分享、审核、移植、反馈等系列功能，沉淀模型经验，开创省市、地市间模型共建共享通路新模式，实现智慧运营最佳实践的快速落地。



[返回目录](#)

九. 联系我们

亚信科技（中国）有限公司

地址：北京市海淀区中关村软件园二期西北旺东路10号院东区亚信大厦

邮编：100193

传真：010-82166699

电话：010-82166688

Email： DataOs@asiainfo.com

网址： www.asiainfo.com





Thank you



亚信科技依托产品、服务、运营、集成能力助力企业数字化，持续创造新价值。