DTCC 2010-2022

第十三届中国数据库技术大会

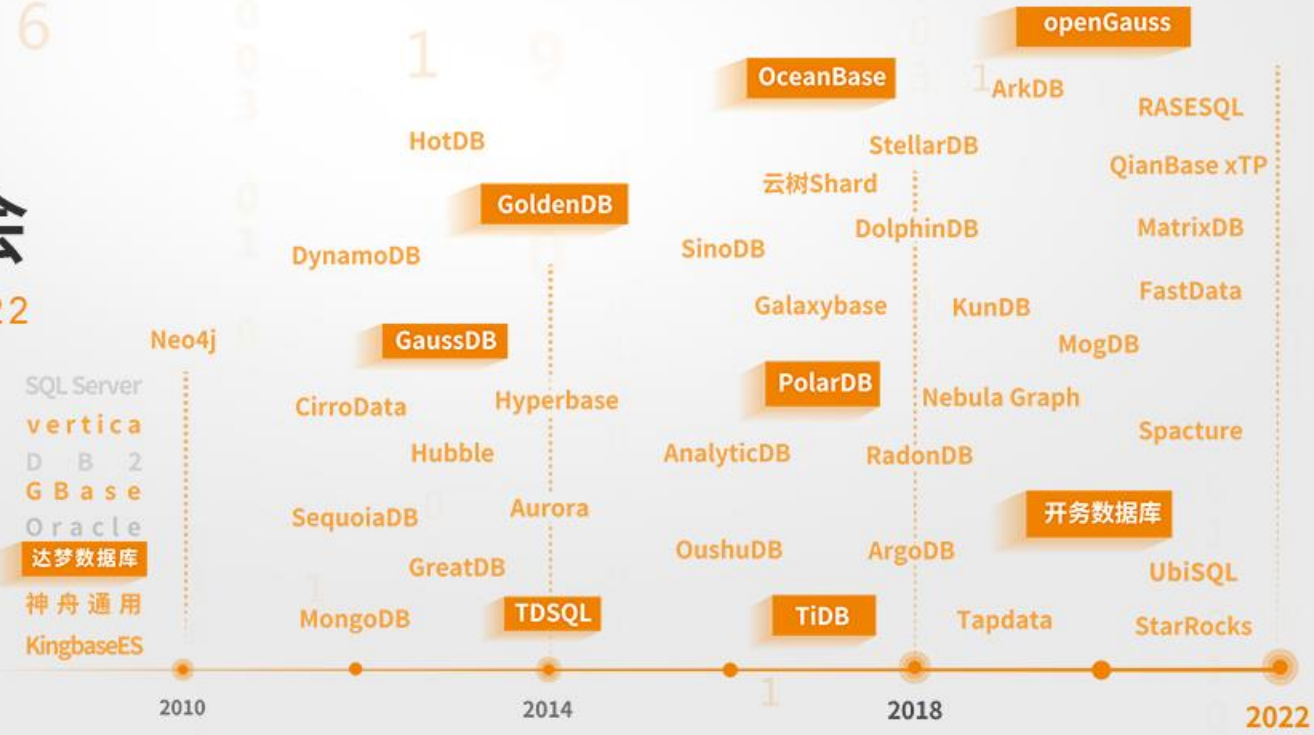
DATABASE TECHNOLOGY CONFERENCE CHINA 2022

数据智能 价值创新

IT168.com ChinaUnix.net ITPUB

线上直播 | 2022/12/14-16

数据来源：数据库产品上市商用时间



FastData DataFacts 建设数据智能平台的实践

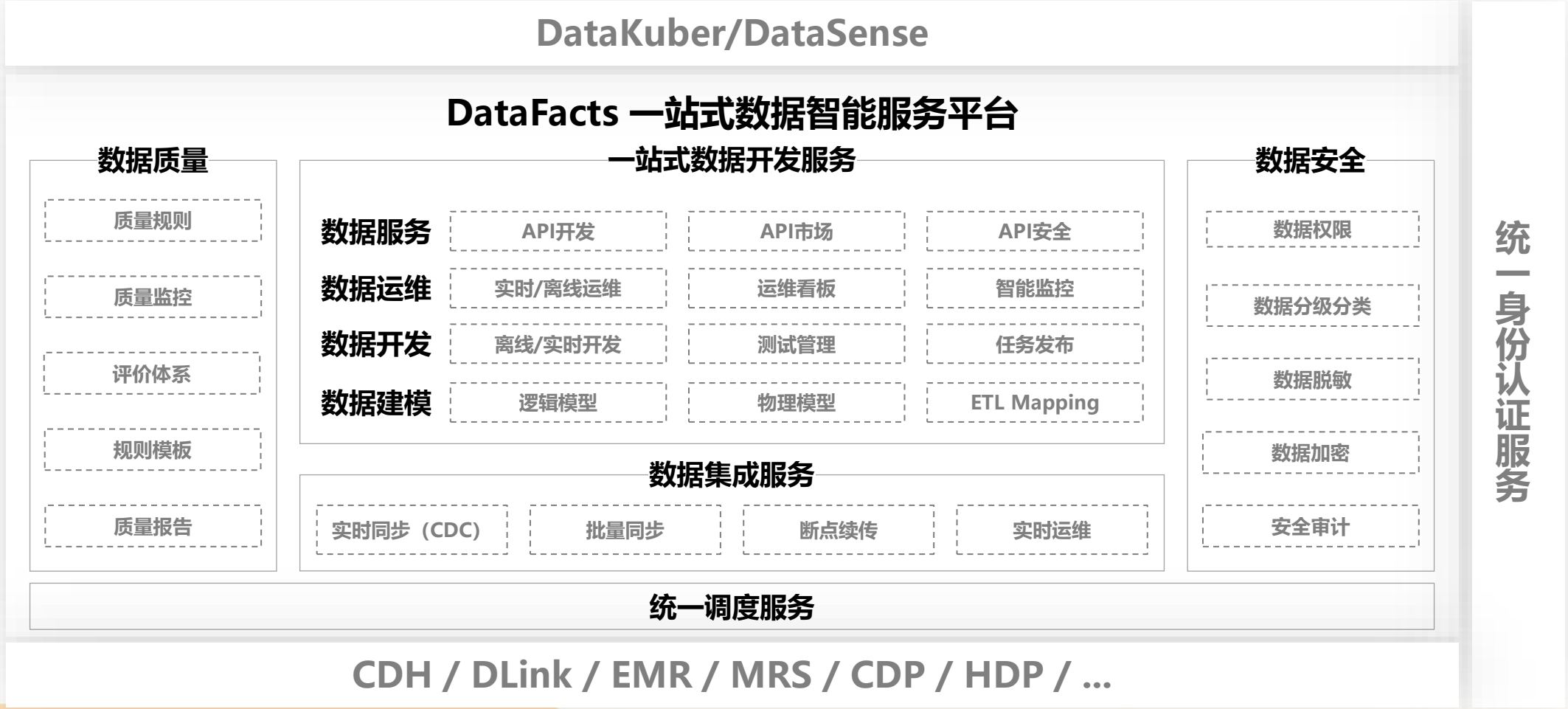
刘波 滴普科技 FastData产品线DataFacts产品总经理

PART 01

DataFacts产品概述

DataFacts产品定位: 一站式数据智能开发平台，实现数据价值可持续释放

DataFacts 是滴普提供的云中立、一站式的数据智能服务平台，为数据工程师提供**数据集成、数据建模、数据开发、数据服务、数据质量、数据安全**等开箱即用的服务能力，降低企业数字化转型实施门槛，提升数据智能平台构建效率，赋能企业数据资产持续沉淀，从而实现数据价值可持续释放。



产品定位

核心目标： DataFacts 面向PB级实时/离线数据开发、运维场景，提供集数据建模、数据集成、数据开发、数据服务等能力于一体的一站式数据开发平台，屏蔽复杂的技术环节，在保证数据安全底线的基础上，让数据工程师专注于面向业务需求的数据开发，**实现易用、便捷、安全的高质量数据生产能力。**

关键痛点

数据平台建设慢

- 数据平台涉及组件多，构建难度大、数据开发人员不能专注于数据业务。

数据任务类型多

- 缺乏有效手段统一离线/实时数据处理；
- 数据类型多样化加剧数据流转及融合难度。

数据开发协同难

- 数据开发工具分散，开发任务难以统筹协同，开发效率低、不敏捷；
- 数据任务缺乏统一审核、发布、运维机制，易出错、易返工、难维护。

数据安全风险高

- 对数据的访问、流通缺乏统一的安全策略。

核心能力

数据集成

- 基于CDC技术，提供不侵入业务系统的企业级实时数据同步服务，保障数据时效性、可用性；
- 基于WAL架构下的CKP异常自动保存技术，实现断点续传，面对再复杂的网络状况，也能保证数据传输的稳定性；
- 插件式能力扩展，快速迭代数据集成能力和数据源适配范围。

数据建模

- 支持逻辑模型、物理模型设计，保障建模流程规范可控；
- 支持模型逆向工程，快速容纳管理企业存量数据模型；
- 融合多个行业最佳实践方法论，提升建模效率。

数据开发

- 支持WEB SQL的可视化离线/实时任务开发，降低用户学习成本；
- 多种任务DAG组织形式，实现跨流程、跨项目任务依赖，方便支持多种业务场景；
- 丰富的大数据组件，根据资源现状灵活实现多种任务，资源利用率更高；
- 支持数据的开发与生产环境隔离、多人协同开发，更安全、更高效。

数据运维

- 数据处理任务以DAG组织并监控，任务修复重跑、暂停、kill等操作更优雅；
- 完备的告警体系，支持自定义告警规则和丰富的日志信息，提高运维效率。

数据服务

- 拖拽式工作流编排，实现复杂api的场景；
- 统一的企业数据共享服务，严格管控数据使用权限；
- 多视角监控及分析服务的使用情况，高效评估数据资产的价值。

数据质量

- 覆盖数据资产化全流程进行质量监管和检验，保障数据完整性、有效性、及时性、一致性、准确性、唯一性；
- 内置质量检测规则模板并支持自定义规则，使质量检查场景更丰富；
- 支持与ETL任务关联执行质量检查，及时发现问题数据，减少数据污染。

数据安全

- 贯穿数据资产化全程，提供对隐私数据的脱敏、权限管理和安全审计等多种数据安全管理措施，全方位保障数据的安全运作。

产品价值

智能运维，及时高效

- 数据任务智能化监控，异常告警及时处理。

大规模数据同步及ETL

- PB级离线/实时数据规模；
- 百万级任务复杂调度。

敏捷、及时、轻松

- 实现数据团队DataOps；
- 数据开发更专注、更高效。

数据安全有保障

- 数据安全策略贯穿数据资源化全流程。

建设快、门槛低

- 平台组件统一部署；
- 开发工具开箱即用，可视化开发维护易使用。

产品优势


核心优势
实时高性能数据同步
基于CDC技术，实现复杂网络环境下的实时数据实时同步，轻松应对海量异构数据源，全面覆盖各类数据同步场景。
技术兼容开放
兼容开源生态，支持多种计算引擎并可随心切换；兼容多种存储标准，满足多模态数据存储需求。
简单易用多人协作
提供敏捷开发IDE，通过简单SQL编译即可让企业拥有大数据开发能力；多人协作设计、开发提高效率，缩短开发周期。
云中立
兼容主流公有云，支持混合云环境，打通云壁垒，保障各云之间的数据一致性管理，规避云计算厂家对企业的数据绑架。
生态开放
对外提供数据基础设施的各项能力，通过OpenAPI，助力实现多种大数据应用业务场景。
全流程数据质量监控
支持ETL全流程的事前、事中、事后数据质量检查和告警。
一站式数据开发
覆盖数据采集、建模、离线/实时开发、运维监控、服务开发等环节，数据开发者只需专注于业务开发。
支持复杂调度场景
支持ETL流程多层嵌套、跨项目、跨流程依赖调度。

PART 02

DataFacts功能介绍

产品功能架构图

数据服务
服务市场 | API监控 | API安全 | API开发 | 上架管理 | 订阅管理

数据开发

数据集成	数据开发	数据检索	测试	发布	运维监控
实时采集	离线/实时	SQL脚本管理	测试用例	发布申请	离线运维
离线采集	自动化DAG	查询结果下载	测试数据	发布审核	实时运维
组件管理	规范性检查	引擎配置	测试执行	发布记录	智能监控
运维监控	插拔式任务	SQL执行日志	测试报告	发布对比	故障诊断

项目空间：基础配置 | 成员管理 | 权限管理 | 多环境管理 | 导入导出

数据建模
数仓规划 | 逻辑模型 | 物理模型 | ETL Mapping

基础服务
计算资源：集群管理 | 引擎配置
数据源管理：RDBS | NoSQL | 消息 | 文件
统一调度服务

数据质量
质量规则 | 质量监控 | 规则模板 | 质量报告 | 工单管理

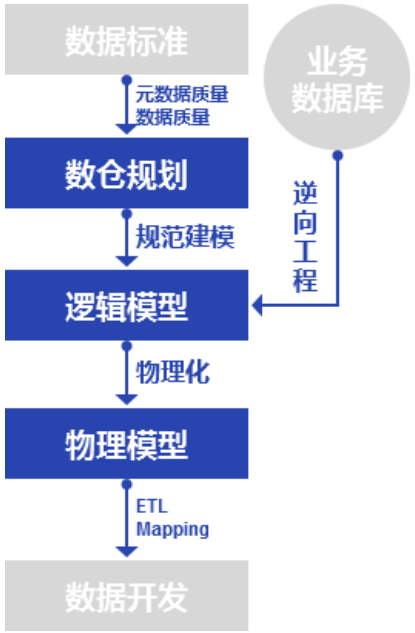
系统管理
数据安全 | 账户管理 | 审批中心 | 消息中心

DataKuber
元数据应用 | 数据标准管理 | 元数据管理 | 数据生命周期管理

CDH/DLink/EMR/MRS/CDP/HDP/ ...

数据建模

通过数据建模模块可以对数据分别按照业务和技术角度进行自定义的主题域分类和数据分层管理，并提供版本、订阅等能力，满足数据模型在根据业务需求动态调整的过程中，可以进行变更通知、版本追溯，保证基于数据标准下，构建有预见性的数据架构，提升后续数据开发、数据应用中各个环节的效率。



数仓规划

基于数仓规划能力，企业可以系统高效地完成业务调研及数据调研，依托数据标准所约束的各项质量规范，从纵向主题域、横向数据分层，立体统一地规划合理紧致的数据架构，形成包含维表、事实表的数仓矩阵，从根本上保障数据模型的可复用性。

逻辑模型

基于规范建模方法论，模型设计师可以在数仓矩阵中从全局视角高效有序地进行逻辑模型填充，通过配置化方式快速设计模型细节，拖拉拽方式构建关联表之间的关联关系，提升模型设计效率、合理性及预见性。

逆向工程

支持通过批量导入的方式，快速将业务数据库中的表逆向为逻辑模型，节省重复建设成本。

物理模型

通过配置化方式，依照数仓规划指引，快速将过审的逻辑模型物理化到数仓位置，降低技术门槛。

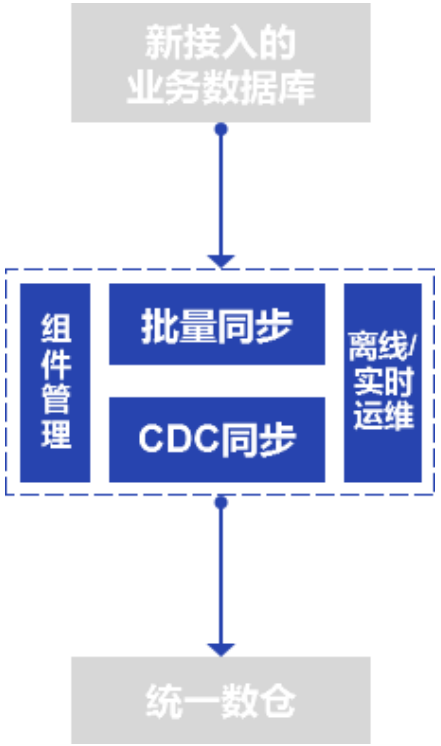
ETL Mapping

基于模型中所定义的Mapping规则，智能生成ETL代码模板，提升数据开发效率。

数据集成

“数据集成模块”提供不侵入业务系统的企业级实时数据同步服务，支持20多种主流数据源异构融合，基于CDC技术，依托断点续传能力，保障复杂网络情况下数据传输的稳定性、时效性、一致性。

覆盖mysql、sqlserver、oracle、postgresql、db2、hana等业务系统主流数据源，支持实现整库迁移、双数据中心、数据整合分发等海量数据高速同步场景，将所有数据同步到统一的存储计算引擎上。



组件管理

覆盖OLAP、RDBS、NoSQL、文件数据库、消息引擎、数据湖等20+主流存储引擎的离线/实时数据传输场景，内置丰富的读取、写入及数据转换组件，支持插件式扩展，快速迭代数据采集能力和数据源适配范围，降低数据集成任务开发门槛的同时，兼具灵活性。

CDC同步/批量同步

基于CDC技术，提供对不侵入业务系统的企业级实时数据同步服务，满足整库迁移、双数据中心、数据整合分发的各类同步场景需求，保障数据时效性、可用性。

离线/实时运维

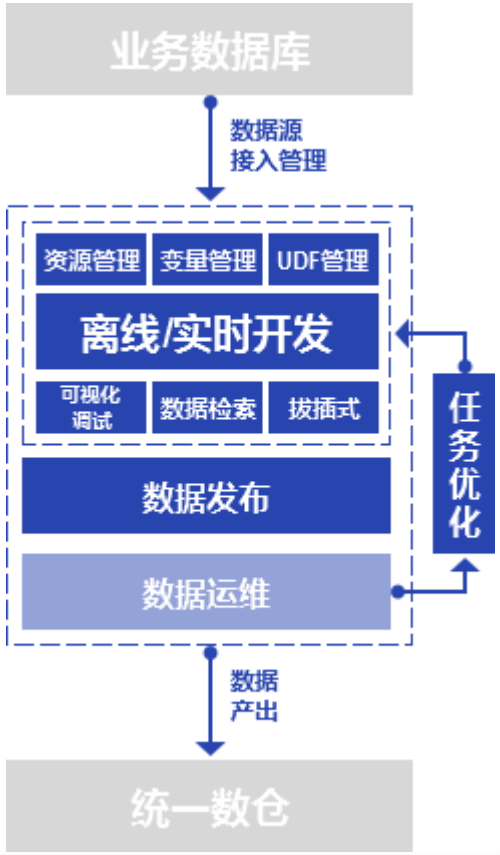
支持对离线/实时同步任务运行状态的统一监控及下探分析，基于对同步错误、任务错误、运行时间等指标的监控及异常处理策略设置，轻松实现大规模复杂同步场景的统一运维。

断点续传

基于WAL架构下的CKP异常自动保存技术，实现断点续传，面对再复杂的网络状况，也能保证数据传输的稳定性。

数据开发

数据开发的目的是将众多业务系统中零散的、不规整的业务数据，在一定的数据治理规范和业务要求下，整合加工为标准、可用、有重点的数据资源，是数据治理的重要底层支撑。DataFacts数据开发提供了界面化、智能高效的数据数据开发与测试体验，支持用户基于基于Web SQL进行可视化的离线/实时任务开发，实现针对批量数据或流式数据的清洗、加工、统计、归档等目标。



离线开发

支持HiveSQL、SparkSQL、ImpalaSQL、Python、Shell、Jar、Spark等任务。支持拖拉拽方式设计DAG流程。支持任务参数、跑数设置等调度配置。支持SQL数据血缘解析和任务版本管理。

实时开发

支持FlinkSQL任务和基于资源上传的Jar任务，FlinkSQL任务可以通过模板快速生成sink和source代码。支持任务参数、运行参数、检查点、TTL参数、任务并行数等调度配置。支持数据血缘配置和任务版本管理。

统一调度

支持跨项目流程、跨流程、跨周期等任务依赖统一调度。

数据发布（CI/CD）

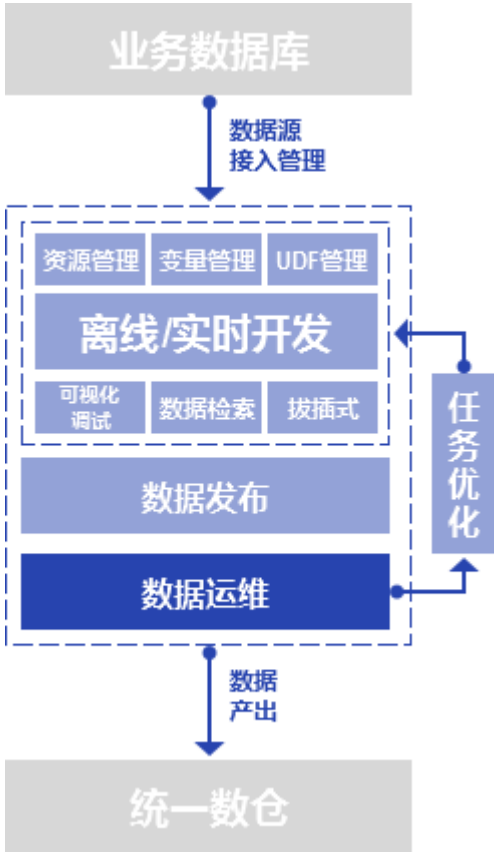
支持数据发布CI/CD，数据开发工程师申请发布后，开发Leader可通过查看任务、比对版本来审核数据任务。审核通过后，运维工程师方可进一步发布到生产环境。

任务优化

语句执行效率优化建议。

数据运维

数据运维通过对任务运行状态及性能、任务耗费资源进行全方位洞察，保障数据开发任务的正常运转，帮助数据开发工程师及时发现问题，解决问题，达到充分利用资源，快速优化任务的目的。



离线运维

支持运维工程师对生产环境上的流程进行运行管理，支持重跑、错误重试、停止、置成功、置失败、补数据、配置质量规则、查看日志等操作。支持为任务关联质量规则并实现告警阻塞。

实时运维

支持对任务运行性能的监控，包括FailOverRate、各Source 的数据输入量、各Sink 的数据输出量等指标。

智能监控

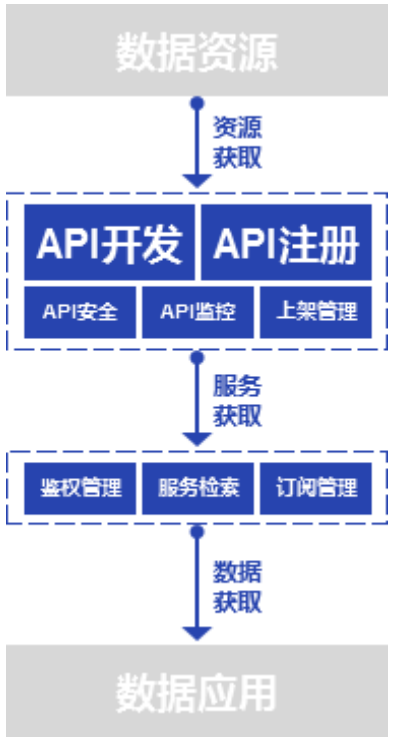
支持智能基线告警，保障任务准时完成。支持针对流程及任务的多种状态的监控策略及告警通知。

智能诊断

ETL流程故障信息收集、故障隔离、故障处理建议。

数据服务

数据服务是数据资源化后进行价值变现的主要方式，通过对数据进行计算逻辑的封装(过滤查询、数据分析和算法推理等)，生成数据服务API接口后，供上层数据应用调用，从而实现数据在业务场景的价值释放。“数据服务管理模块”提供了统一的数据服务开发管理、使用权限管理及服务订阅管理，帮助企业实现统一的数据服务目录，实现跨部门的数据服务共享。



API开发

支持以“向导配置”的方式，无需SQL，即可快速配置API实现对表的查询。配置好的API可自动生成SQL语句，用户可以方便的进行API测试后，发布上架。针对稍复杂的查询分析场景，可通过自定义SQL的方式生成API。针对更加复杂的查询分析场景，可以通过服务编排的方式将多个API节点和Python函数进行串联处理的方式提供数据服务，及针对输出的数据进行脱敏处理。

API注册

支持将企业已有的数据API注册到统一的网关中进行统一管理。

API监控

支持出错告警、耗时告警、限流告警等配置及查看。

API安全

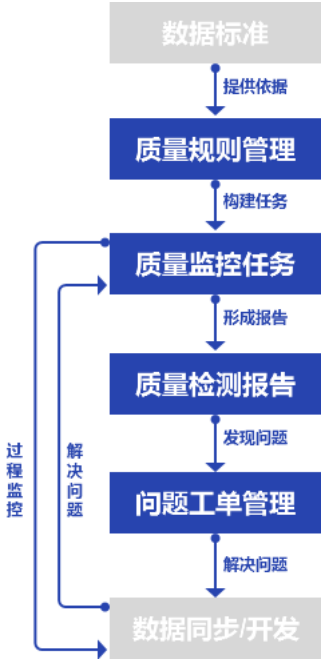
支持黑白名单、QPS限制、单位时间调用次数限制。

API订阅

支持通过API市场检索服务、订阅服务及设置使用期限。支持数据应用鉴权管理及成员管理，实现数据安全共享。

数据质量

数据质量模块通过质量监控、报告，从产前、产中、产后等不同环节，对原始数据、中间数据及应用数据进行质量稽核，挖出不合规的数据项，指导用户指定数据质量优化策略，保证数据的完整性、一致性、准确性等指标。



质量规则

内置多种质量规则，支持从合理性、完整性、唯一性、及时性、准确性、规范性、一致性、及时性等质量评分指标对数据质量进行监控；同时，支持用户通过SQL或正则表达式的方式自定义质量规则；

支持用户自定义质量评分指标，为不同的质量规则关联指标并分配计分权重，从而根据企业的数据质量评估体系，有层次有重点地对数据质量进行评分。

质量监控

支持用户创建单表稽核、多表比对、关联数据元数据集稽核等质量监控任务，并以配置化的方式设置多个质量规则及任务调度方式。

支持周期性调度、手动触发、关联任务等多种调度方式。其中，关联任务调度支持数据质量任务关联多个数据开发任务，对产出数据进行过程质量稽核，如触发强规则不通过，则自动阻塞下游节点。

检测报告&工单管理

用户可以查看数据质量评分、总体规则数量及错误告警数量、任务时长详情等情况总览，查看触发错误告警的未通过规则列表、趋势、详情等信息，并对相应的数据质量问题以工单形式进行流转和处理。

数据安全

在数据处理、流转和使用的过程中，数据安全模块覆盖所有的环节，提供权限管理、数据脱敏、操作审计等能力，进行事前事后风险控制，满足诸如隐私计算等场景需要。

数据分级

基于数据安全策略，为数据获取权限、敏感数据扫描、数据脱敏加密提供不同程度的管控密级。

数据权限

为数据开发、服务开发、数据共享提供库级、表级、字段级等颗粒度的权限管控。

数据脱敏

支持构建自定义敏感数据识别规则，并构建自动化的敏感数据发现任务及分布报表，并进行遮掩、截断、Hash等脱敏操作。

数据加密

支持各种数据加密算法、管理加密策略。

日志审计

提供针对各类系统操作及数据访问操作的日志审计。

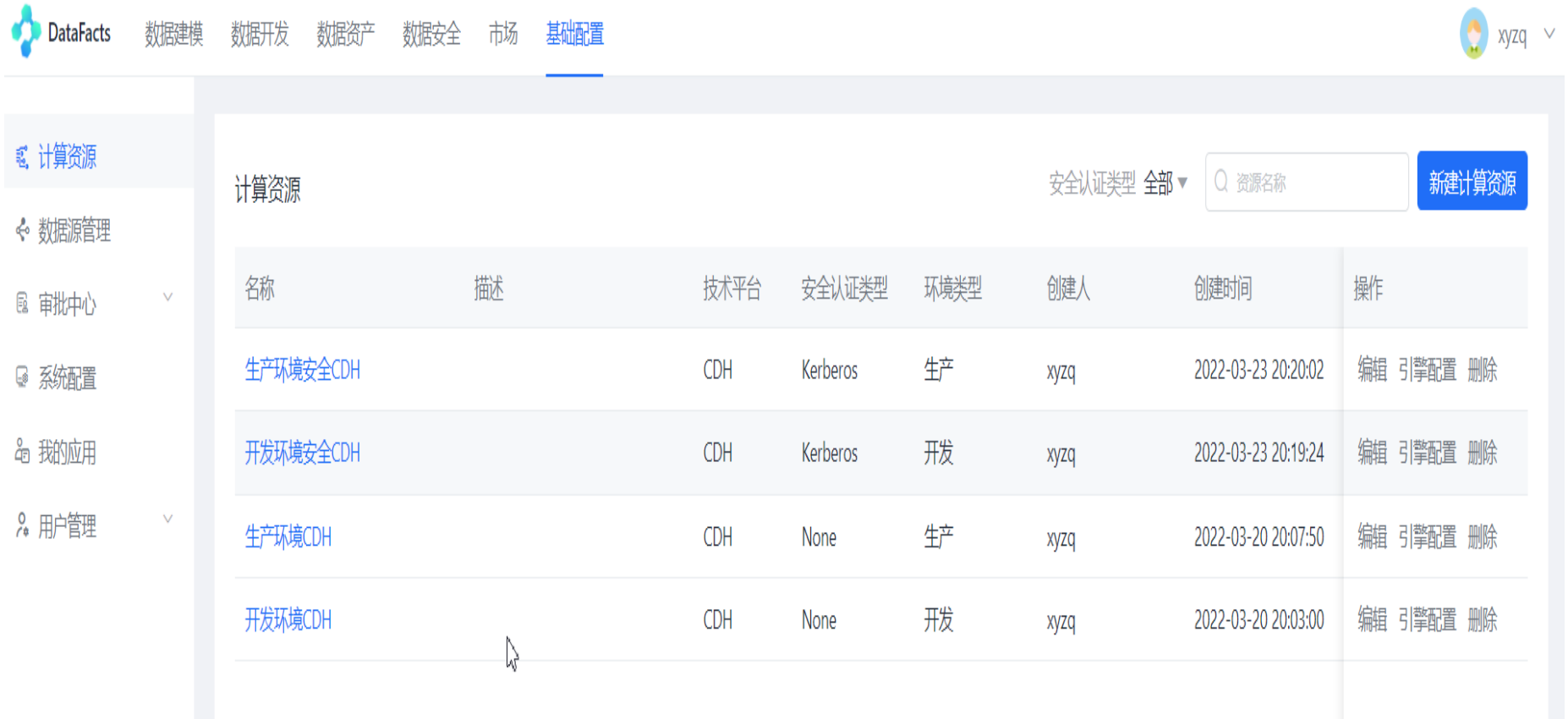
PART 03

DataFacts开发流程

一站式数据开发

- 基础配置
- 数据模型设计
- 项目配置
- ETL工程开发
- ETL发布
- ETL运维
- 数据服务开发

进行数据开发之前，需先完成计算、存储引擎配置及连通性测试，以及相关账号及权限配置工作。



一站式数据开发

基础配置

数据模型设计

项目配置

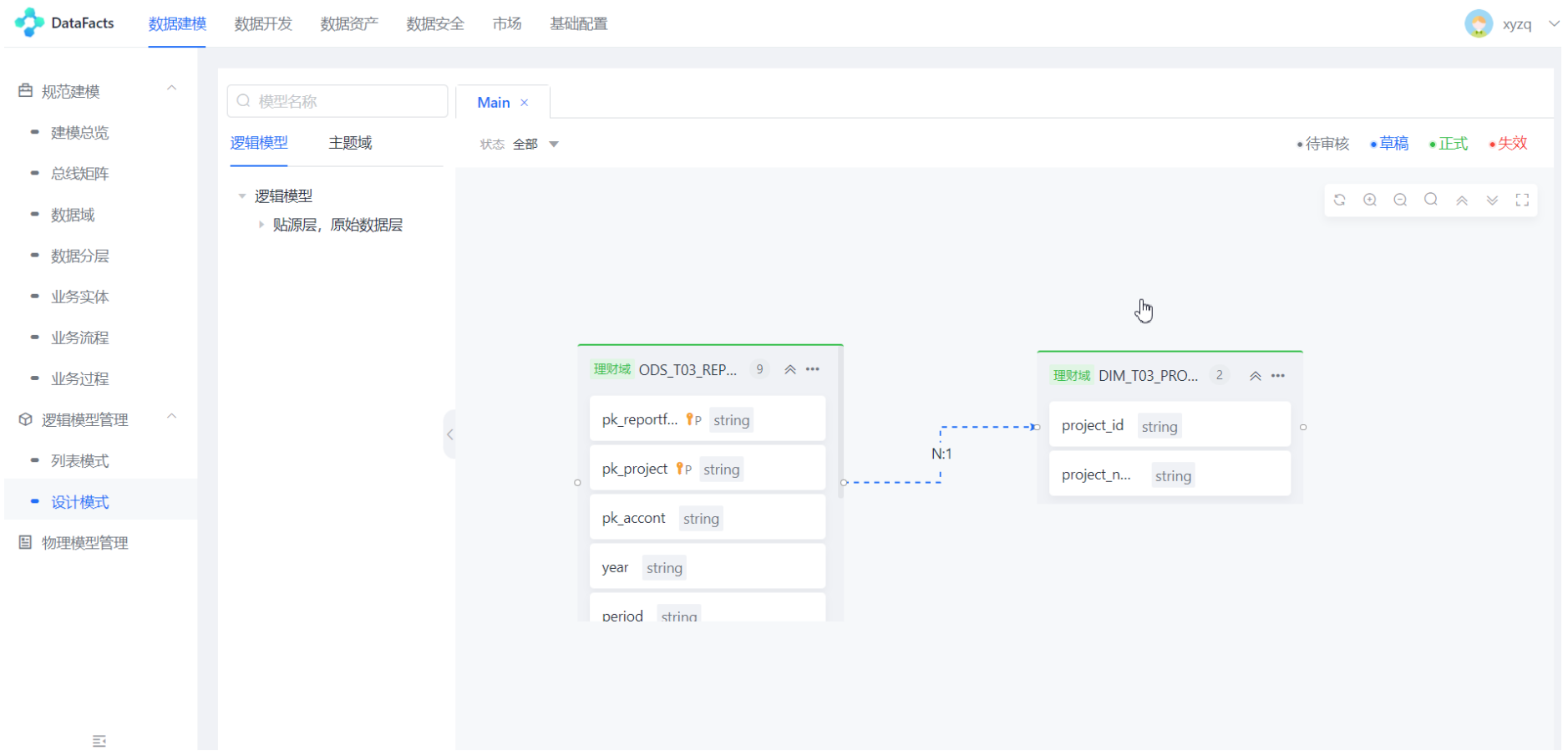
ETL工程开发

ETL发布

ETL运维

数据服务开发

1、逻辑模型设计，包括模型分层、所属主题、模型数据、模型关系等信息；
2、物理模型设计，包括数据库类型、表物理参数，创建的数据源等。



一站式数据开发

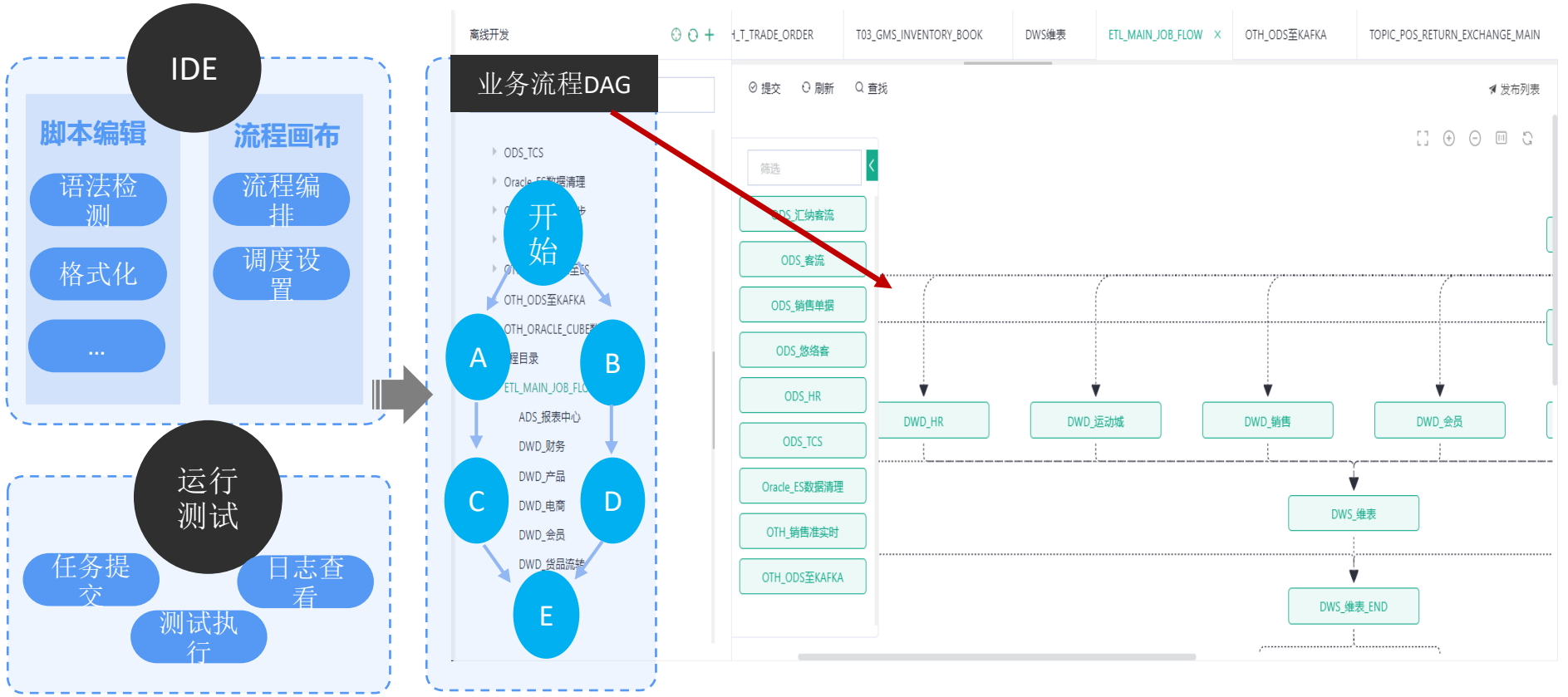
- 基础配置
- 数据模型设计
- **项目配置**
- ETL工程开发
- ETL发布
- ETL运维
- 数据服务开发

创建项目空间实现权限隔离，项目空间中可配置项目基础信息、生产&开发环境、成员管理、权限管理、数据源信息等。

一站式数据开发

基础配置
数据模型设计
项目配置
ETL工程开发
ETL发布
ETL运维
数据服务开发

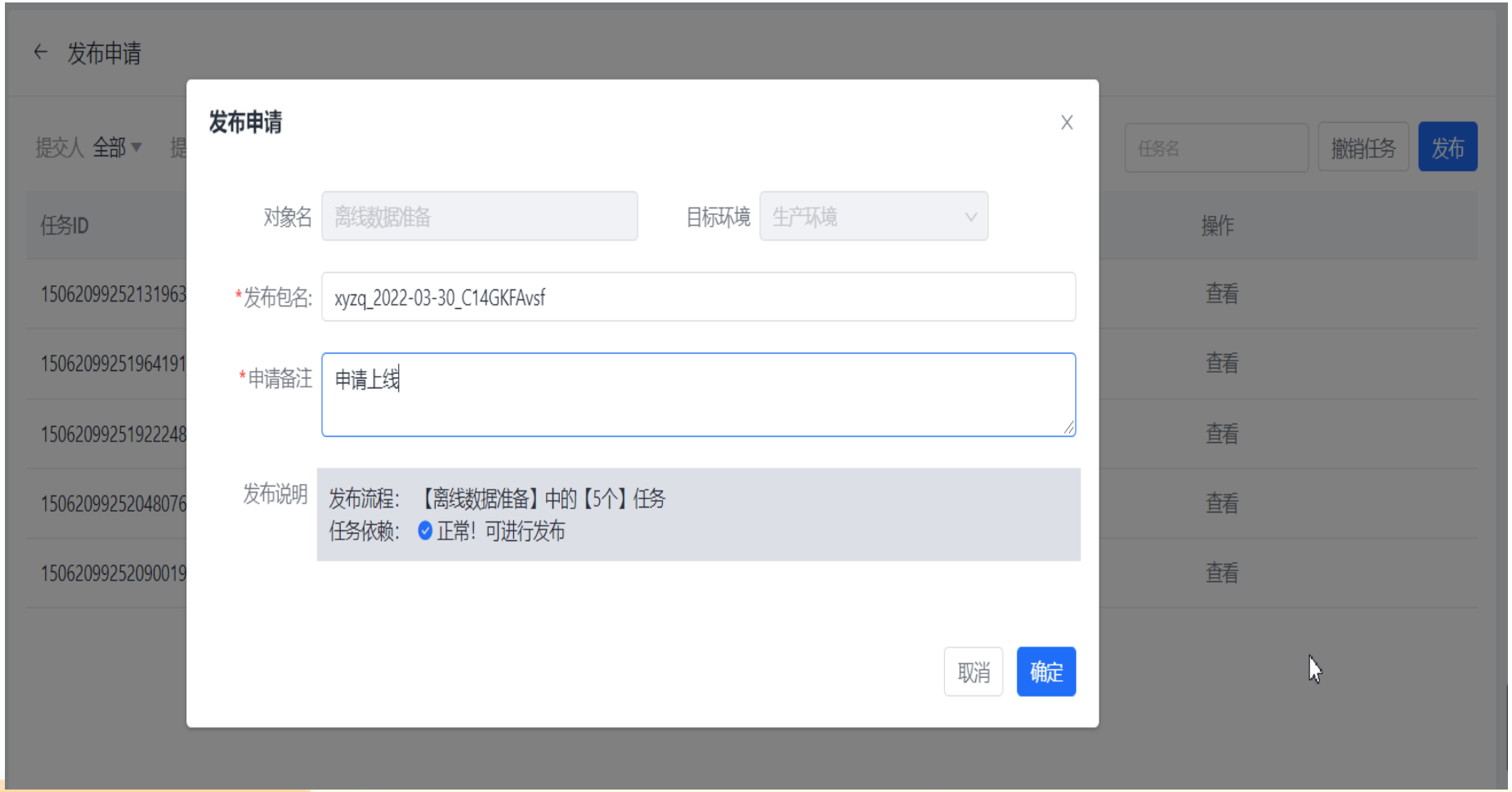
进入项目空间进行ETL工程开发，产品具备离线\实时开发能力，适配多种任务类型（Spark SQL、Hive SQL、Impala SQL、Flink SQL、Python、Jar等），支持资源管理、函数管理、变量管理等功能。



一站式数据开发

- 基础配置
- 数据模型设计
- 项目配置
- ETL工程开发
- ETL发布
- ETL运维
- 数据服务开发

ETL工程在开发环境完成开发及测试后，可通过“发布”功能经过审核后提交到生产环境。



一站式数据开发

基础配置

数据模型设计

项目配置

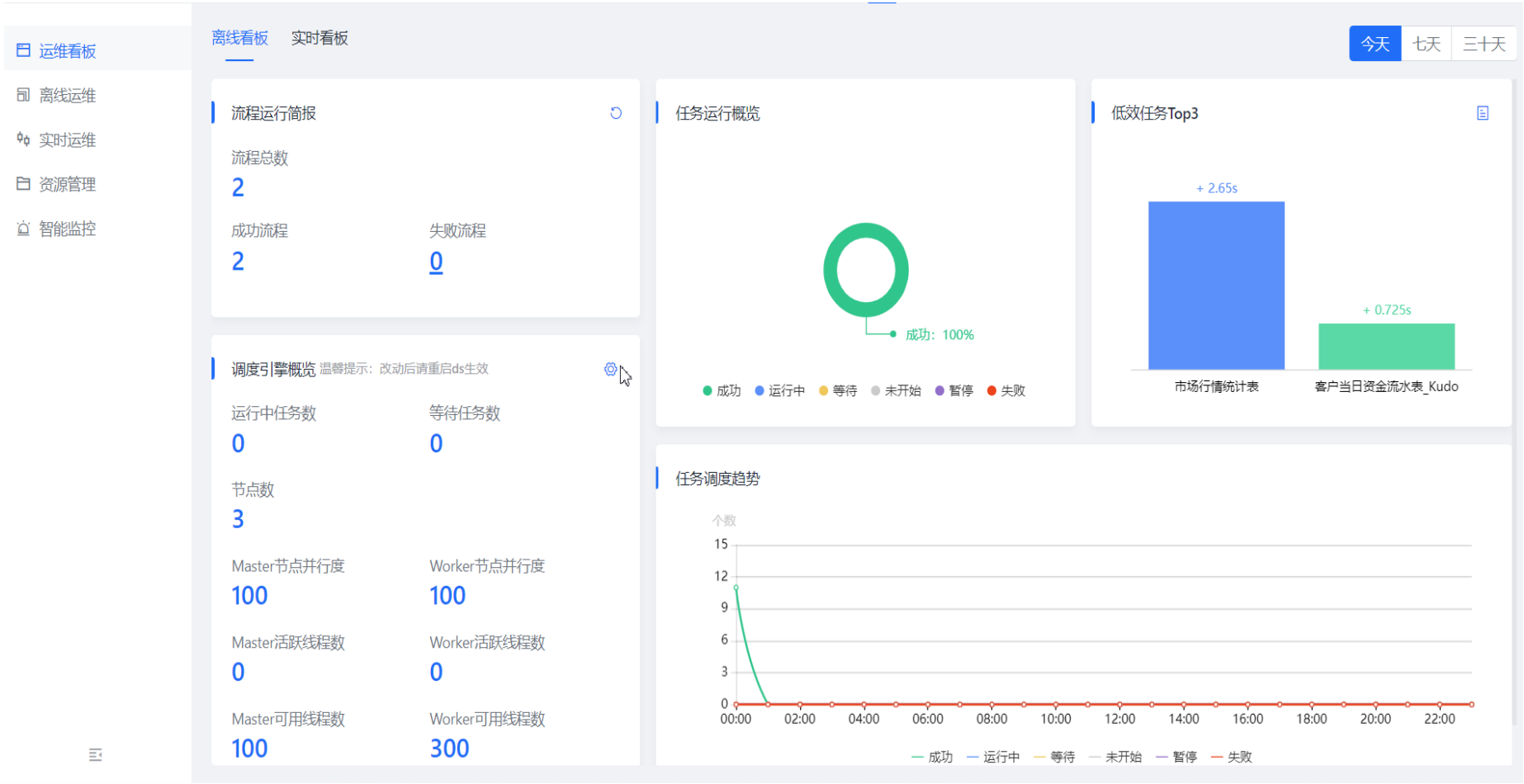
ETL工程开发

ETL发布

ETL运维

数据服务开发

在运维模块对生产环境任务做运维、监控、错误重试、补数据、流程管理、告警配置等自动化运维动作。



一站式数据开发

基础配置

数据模型设计

项目配置

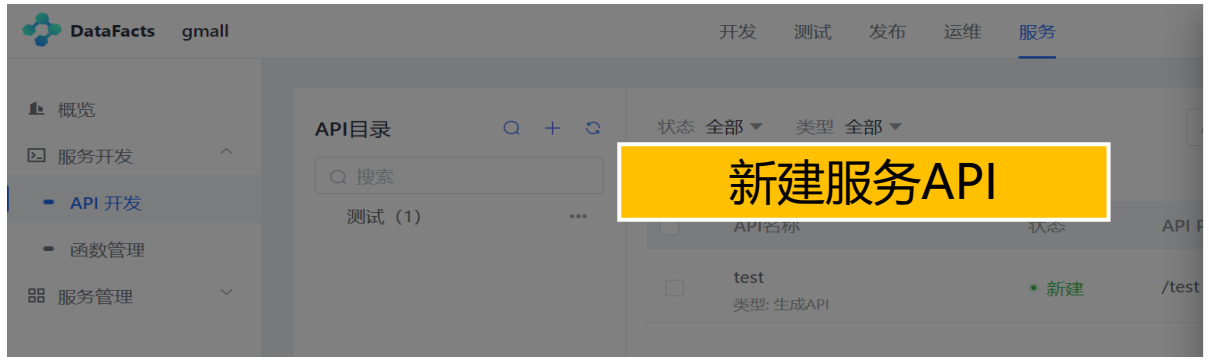
ETL工程开发

ETL发布

ETL运维

数据服务开发

数据服务，为企业上层应用提供统一的、标准的数据共享接口，提高数据利用效率和质量。



PART 04

数据平台案例实践

滔搏运动：数据平台建设，算力和服务响应效率提升，数据管理能力提升

案例背景

- 作为中国最具竞争力的体育用品零售商，2002年成立以来零售网络覆盖中国九大区域，300多个城市，拥有8000+门店。拥有多个国际运动及户外品牌经销权，业绩长期保持跨越式增长。公司于2019年10月在香港联合交易所主板上市，并于2020年被纳入MSCI中国指数，并获入选港股通。

客户痛点

业务动因：

- **系统差异问题：**虚拟店、库存口径、营业目标、分类汇总、算法逻辑、流程规范等问题待解决；
- **业财差异问题：**自营批发划分、库存分类差异、管理架构、O2O销售、内购券算法、多品外招柜、业态分析标准、批发加盟客户统计口径问题；
- **电商差异问题：**不可共享库存、产品主数据差异、销售差异、库存口径差异等。

技术动因：

- **数据口径不一：**业务财务管理视角不一致，各方对数据理解有冲突
- **部分数据维度缺失：**数据分析有的从业务出发，有的财务视角的数据，数据缺乏统一存储，管理，利用，难以业务财务综合分析；
- **原有技术架构性能不足：**数据治理能力难以支撑快速迭代的业务需求

滴普解决方案

实现场景

数据资产标准化

对现有业务逻辑、系统逻辑的梳理，构筑滔搏数据中台，实现统一的数据视图

数据赋能业务

基于统一数据中台，构建数据模型和数据分析，统一可靠地数据来源赋能业务

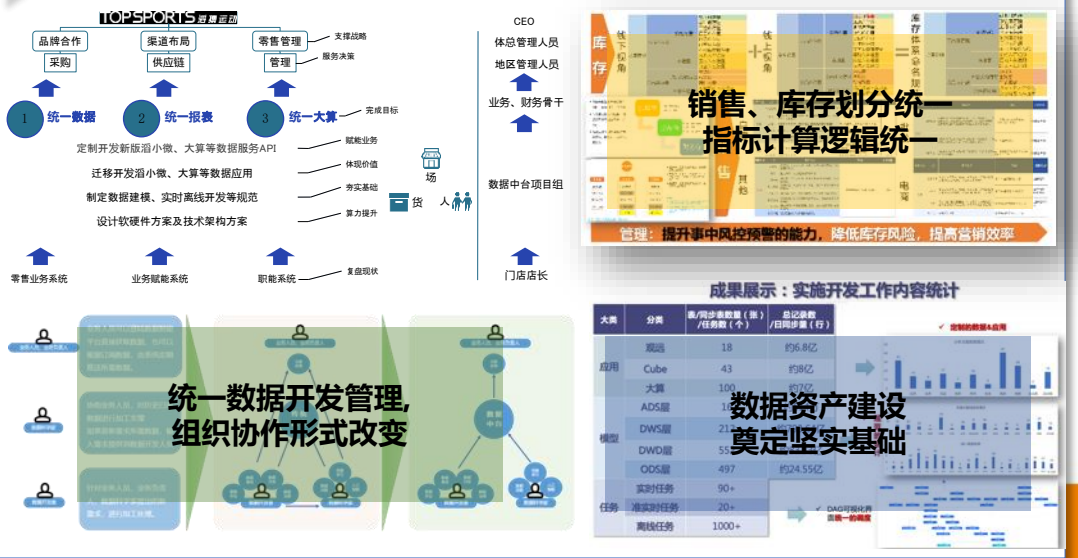
数据可视化管理

通过使用DataFacts平台作为数据管理抓手进行数据安全管理、数据服务管理等管理动作

使用产品

DataFacts

产品运营效果图示



给客户的价值

数据平台算力提升：

大数据集群以分布式+高可用部署直接提高了底座的健壮性；实时数据计算采用Flink 等最新技术保证了算力进一步提升。

服务响应提升：

通过使用DataFacts平台，用户可以直接了解当前数据现状，快速定位需求，这种以终为始的响应模式保证响应效率进一步提升。

业务标准规范化提升：

项目过程中制定的数据开发规范、数据建模规范、数据维护规范等标准规范，有利于保证后期数据运维可以持续性的提升。

数据管理能力成熟度提升：

通过使用DataFacts平台作为数据管理抓手进行数据安全管理、数据服务管理等管理动作，有利于保质保量的达成统一数据、统一报表、统一大算的目标。

百丽时尚：基于FastData湖仓一体架构，优化成本、性能和效率

数据解决方案

全域数据库

内部数据

ERP CRM 业务库

系统数据

门店视频/IoT

外部数据

电商平台/舆情

天气/高德等

DIC咨询设计

商品八大业务场景：经营计划 | 商品企划 | OTB | 采购管理 | 货品首铺 | 补货迭代 | 调拨调价 | 下沉降解

FastData 数据资产管理运营平台

识别服务 | 统计服务 | 分析服务 | 预警服务 | 决策服务

模型：单品销售预测模型 | 商品相似度模型 | 商品畅滞销模型 | 库存分析模型 | 尺码拆分模型 | 商品等级模型 | 门店等级模型 | 店货匹配模型

标签体系：商品体系 | 渠道体系 | 门店体系

指标体系：商品域 | 零售域 | 交易域 | 财务域

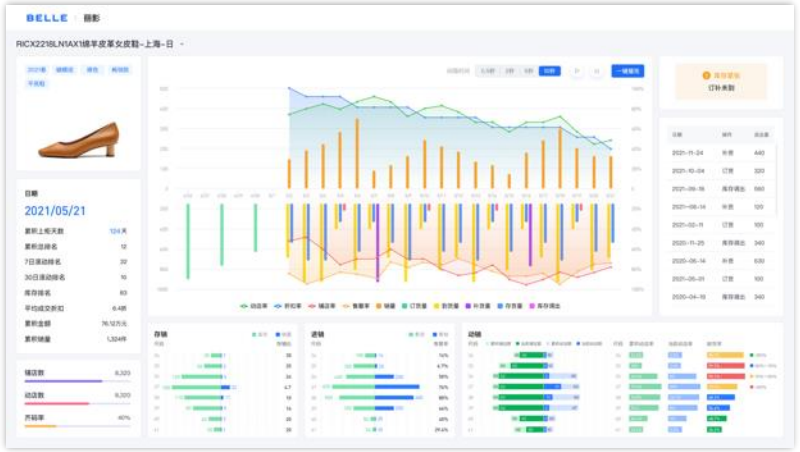
FastData 数据平台基础底座

一站式数据治理和数据服务平台

流批一体数据分析引擎

私有云 + 公有云 | HDFS | PB级数据存储 | 云原生架构

业务价值场景



✓ 预测工具

✓ 指标和标签赋能百丽补货平台升级


openGauss
OceanBase
ArkDB
RASESQL
HotDB
StellarDB
QianBase xTP
云树Shard
GoldenDB
DolphinDB
MatrixDB
DynamoDB
SinoDB
FastData
Galaxybase
KunDB
SQL Server
vertica
D B 2
GBase
Oracle
达梦数据库
Spacture
开务数据库
UbiSQL
神舟通用
TDSQL
TiDB
Tapdata
StarRocks
KingbaseES
2010
2014
2018

THANKS