火山引擎

www.volcengine.com

数据中台DataLeap实现沉淀数据高效建设

——火山引擎DataLeap介绍

2022.12

数据建设痛点

01

企业数据建设痛点

需求响应慢

业务需求多，数据开发效率低，需求响应慢

资源成本高

数仓规划不合理，数据开发不规范，大量存储/计算资源浪费

数据质量差

数据质量参差不齐，业务使用数据有风险

资产管理难

数据来源多，数据资产缺少统一管理，难以发挥数据价值

安全无保障

用户数据保护，数据安全管控要求严格，出现问题难追溯

业务为先，数据中台工具的发展过程

- 缺啥做啥

- 数据研发驱动
- 开发效率提升

- 全生命周期
- 价值交付

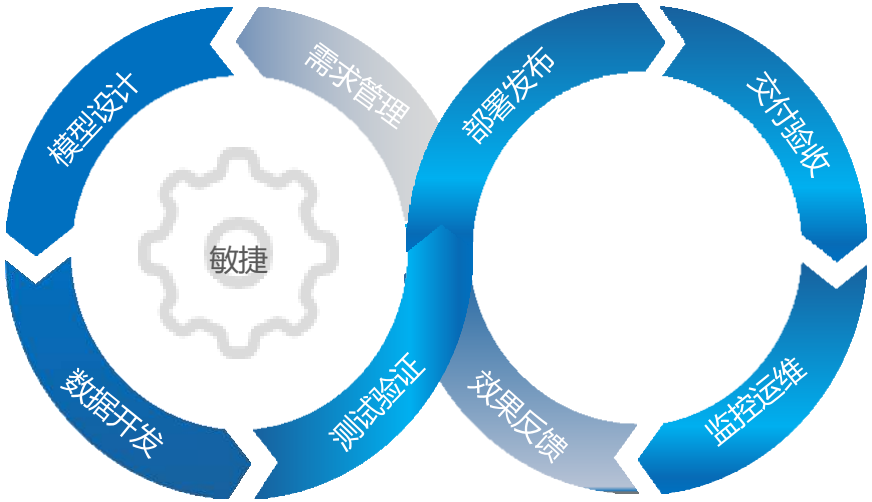
以业务为先为导向，沉淀数据建设最佳实践

服务评价体系

- 稳定性
- 需求满足度
- 数仓完善度
- 用户满意度

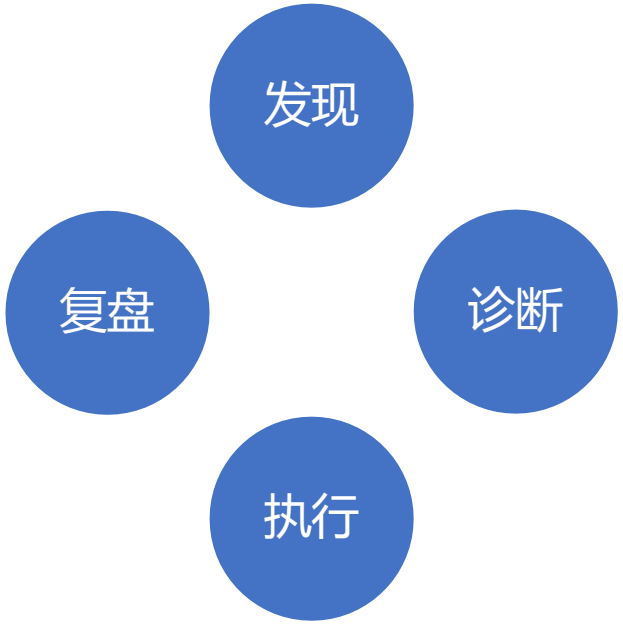
数据BP
0987机制

数据研发闭环



DataOps
敏捷开发

数据治理闭环



分布式治理
业务自治

“组织+方法论+平台”三位一体，实现高标准“0987”的业务价值目标

稳定性

SLA故障为0

需求满足度

业务需求满足率90%

数仓完善度

分析师查询覆盖率80%

用户满意度

NPS70%

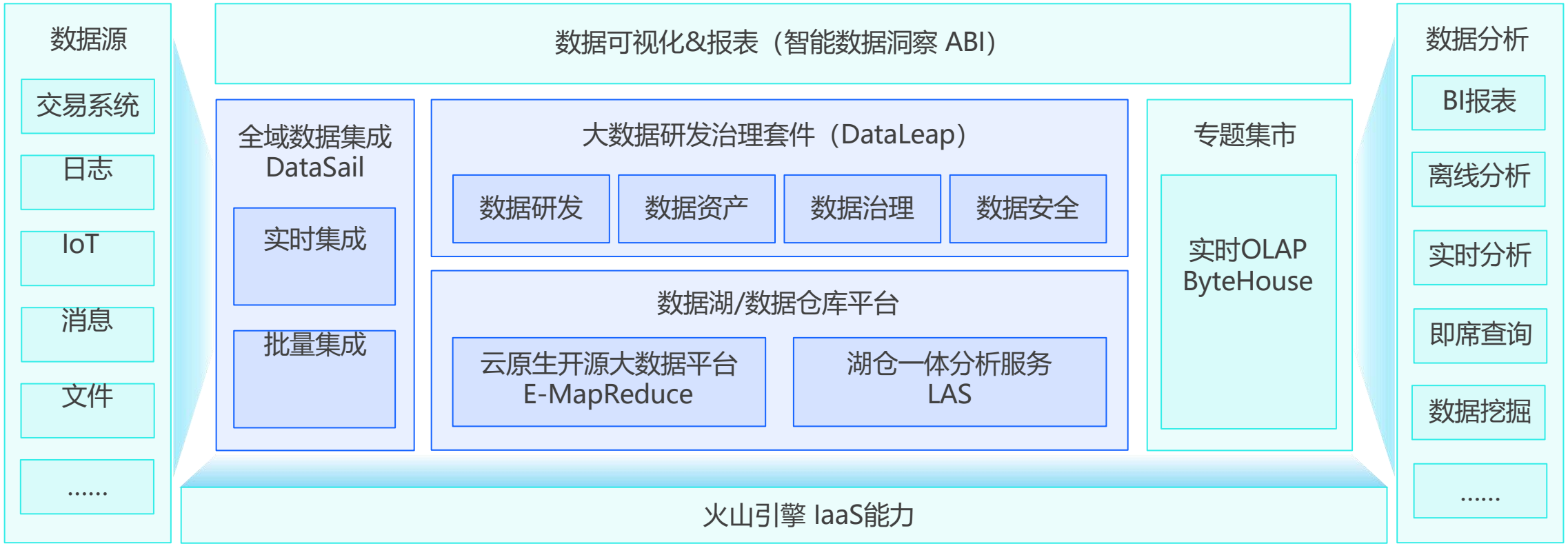
DataLeap解决方案

02

火山引擎数据中台方案

提供大数据存储、计算、分析、展示能力，支持TB~PB级数据的离线/实时/检索等场景数据处理分析

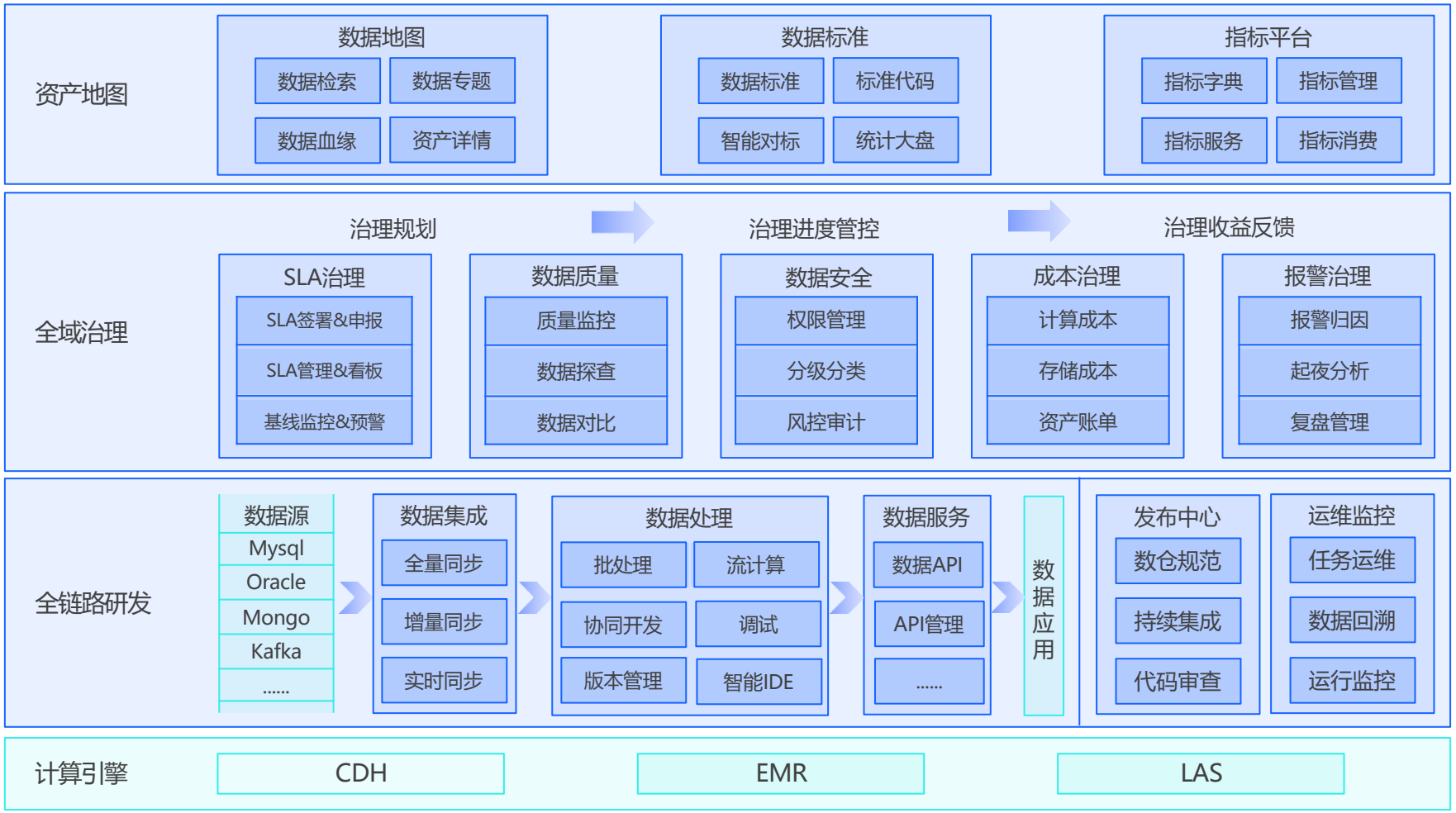
- **开源Hadoop生态大数据服务EMR：**开源Hadoop生态，企业级增强，集中管理运维，支持数据湖场景
- **湖仓一体分析服务LAS：**Serverless湖仓一体分析，多模引擎，兼容开源生态，支持数据仓库&数据湖场景
- **实时数据分析引擎ByteHouse:** 海量数据高性能写入/查询
- **一站式大数据研发套件 DataLeap：**兼容Hadoop开源生态，支持Flink，Spark等多种计算引擎，覆盖数据集成、开发、运维、治理及资产管理全链路。
- **智能数据洞察产品ABI：**以数据洞察为导向，从数据接入、数据整合、到查询、分析，最终以数据门户、大屏、管理驾驶舱的可视化形态呈现给业务用户，让数据发挥价值。



DataLeap大数据研发治理套件支撑字节集团数据中台建设

火山引擎 数据中台

DataLeap一站式数据研发治理平台



数据研发全链路管理

整合全域数据，支持20+多源异构数据集成，灵活对接各类业务系统。敏捷开发CI/CD，覆盖需 求、开发、测试、发布、运维等研发全链路管理。

数据全生命周期治理

结合基线监控、数据质量、 SLA治理等能力，提供事前预警、事中处理、事后复盘及推荐优化的全生命周期的数据治理能力

沉淀数据规范

统一数据标准及数据查询出口，沉淀数仓建设规范的最佳实践，提升数据开发效率，保证数据质量，快速精准为业务赋能

保障数据安全

更细粒度的行、列权限控制，表及字段级别的血缘管理，加上行为监控等功能，构成真正意义上的数据安全屏障

多云多引擎

提供公有云PaaS服务及灵活的私有化部署方案。同时，更可低成本、高效适配客户已有大数据平台，控制迁移成本，降低业务影响

基于DataOps的敏捷的数据研发全链路管理解决方案

全链路覆盖

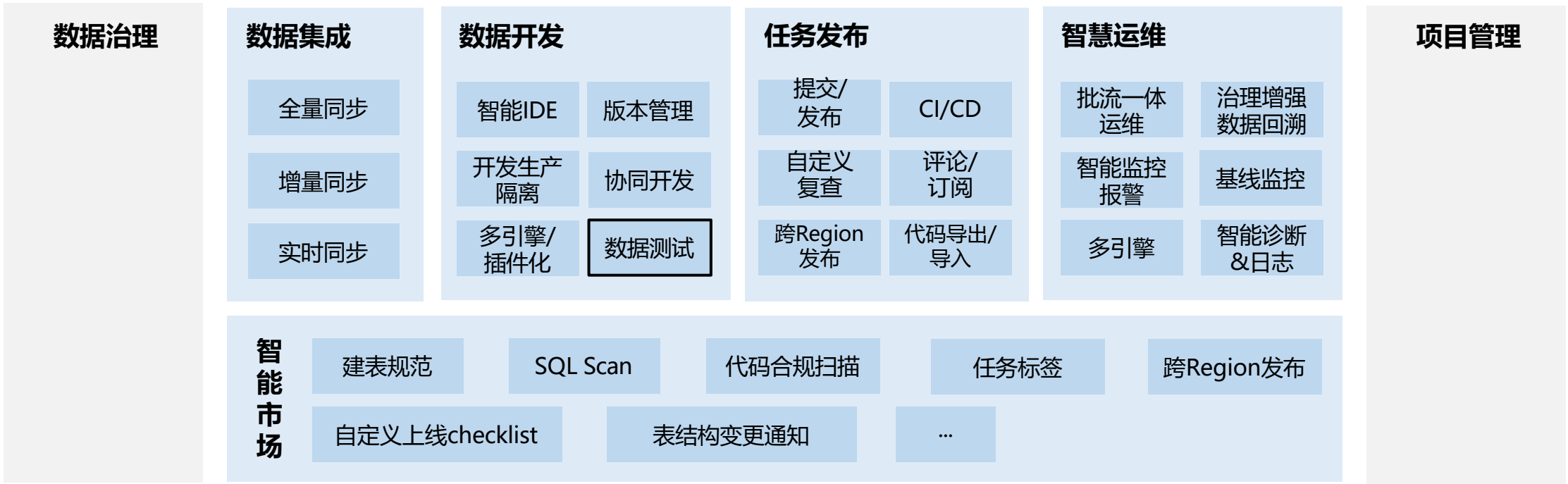
- 流程全覆盖，“集成/开发->测试->发布->验收->运维”
- 场景全覆盖，兼容Spark、Flink等多种计算引擎，提供HSQL、Spark、Python、Flink SQL等10+数据开发能力
- 管理全覆盖，丰富角色定义，开发规范配置

CI/CD自动化

- 研发流水线，插件化、模板化抽象数据测试、代码发布流程，提高开发效率
- 高扩展性，智能市场承载多种自定义插件，与研发流程无缝融合

治理深度融合

- 研发过程与数据治理深度融合，通过事前、事中治理联动，降低事后治理成本
- 挖掘元数据价值，丰富开发规范，优化研发链路，提升管理水平



关键能力：一站式数据研发全链路管理

稳定、安全、高效的数据集成

- 20+多源异构数据集成
- 覆盖常见的业务存储系统
- 提供全量、增量、实时的数据同步能力，整合全域数据

一站式、全栈数据研发

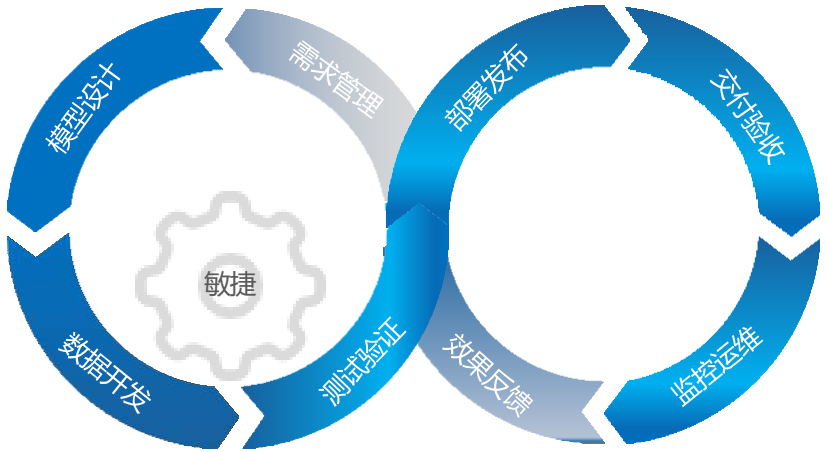
- 沉淀数仓建设规范，支持建模规范自动化检查
- 兼容Spark、Flink等多种计算引擎，提供HSQL、Spark、Python、Flink SQL、Notebook等10+数据开发能力
- 协同开发、智能IDE提高开发效率，在线调试、数据测试加快代码验证流程
- 敏捷开发CI/CD，支持开发生产隔离，跨域/项目代码同步，实现代码持续集成与部署

全面的运维能力

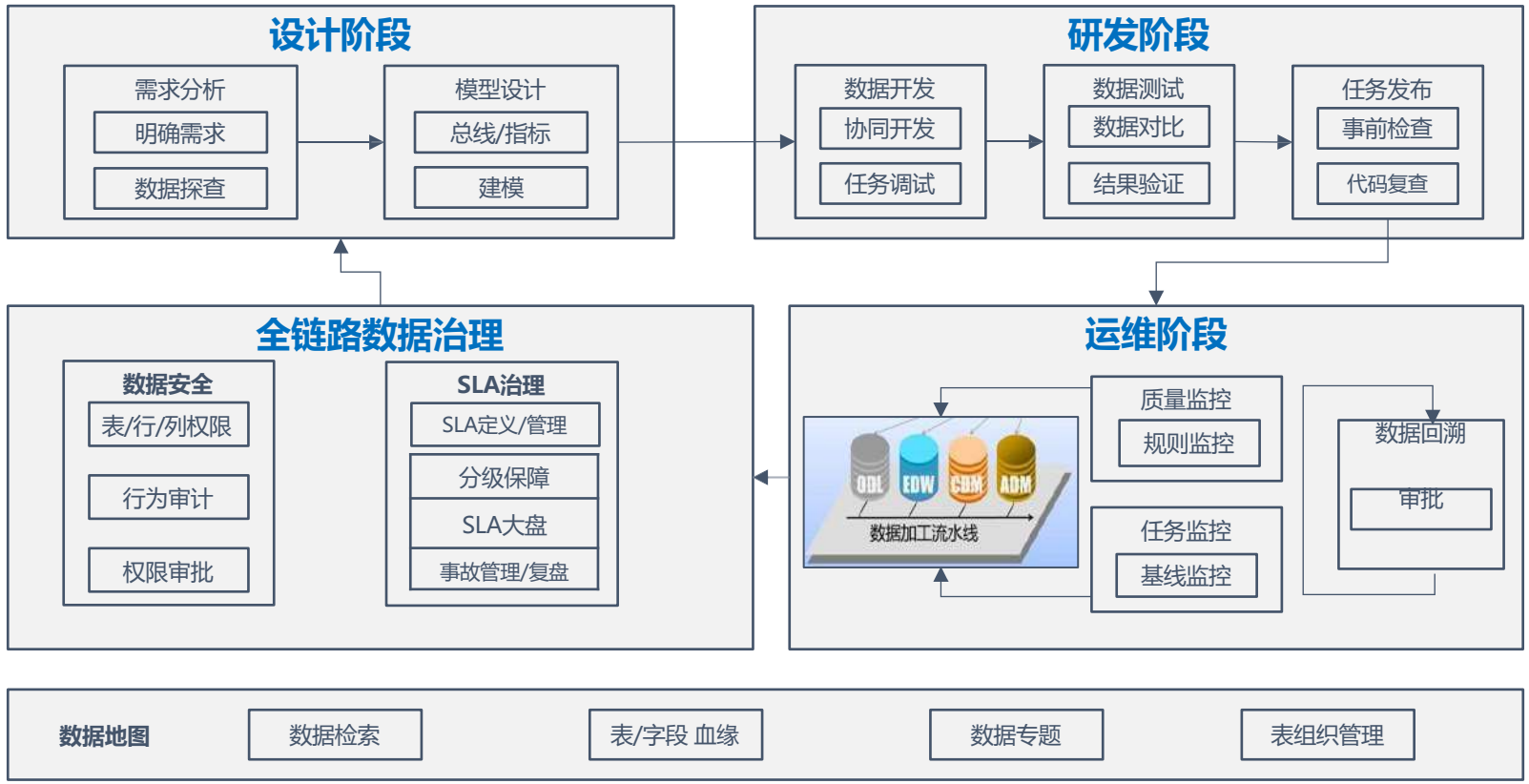
- 丰富的批、流任务监控规则，归类业务运维管理，监控全链路任务运行
- 支持天级/小时级基线告警，降低监控运维成本，报警治理最佳实践
- 复杂业务场景验证的数据回溯能力，解决补数据、重刷历史数据等问题
- 深入底层引擎逻辑，沉淀智能诊断服务，赋能常见运维问题的自检测能力

数据研发全链路管理

需求->设计->**开发->测试->发布->验收->运维->**效果



关键能力：贯穿全生命周期的治理体系



高效的数据探查

- 一键探查目标数据表的数据分布，快速了解数据特征
- 提高需求分析及建模的效率，快速了解数据表概况

丰富数据质量监控

- 自定义强、弱质量规则，与生产数据的任务关联，第一时间阻止数据污染，保障数据可靠
- 丰富的质量规则校验，可复用模板，也可灵活自定义
- 多表关联、多行多列等复杂质量规则配置，满足复查的业务场景需求

赋能业务自治，实现SLA全链路保障

- 按需申报，根据实际业务当前发展状况与实际需求，自发自驱进行申报
- 高效对齐，对生产链路进行精细分析，长链路任务也能快速对齐治理目标
- 全链路保障，SLA签署与复盘均闭环在产品中，签署完成后进行系统级保障

关键能力：数据治理闭环

按需申报

- 按实际业务发展阶段和要求，按需申报

根据实际业务当前发展状况与实际需求，自发自驱进行申报

高效对齐

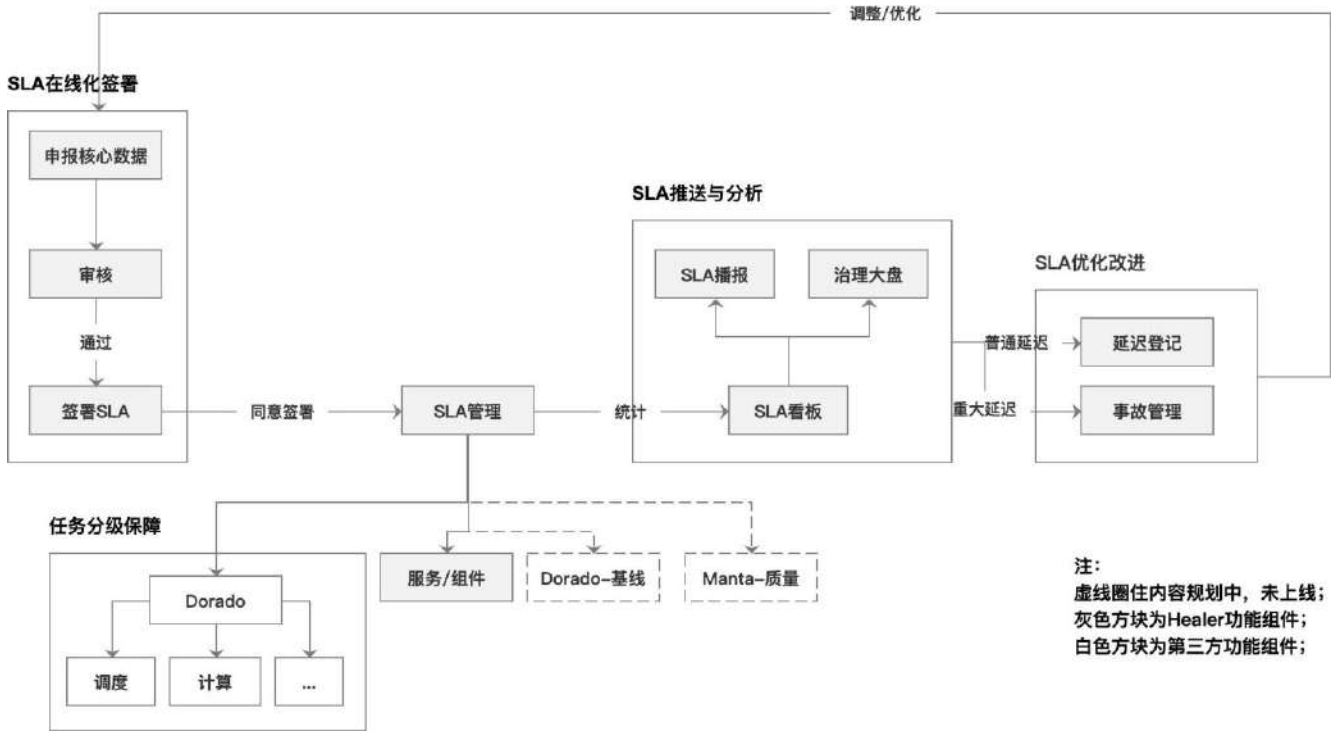
- 链路分析与算法机制，高效对齐

对生产链路进行精细分析，长链路任务也能快速对齐治理目标

全链路保障

- SLA治理线上化，全链路保障

SLA签署与复盘均闭环在产品中，签署完成后进行系统级保障



关键能力：端到端的数据质量保障

自主探查

- 事前-数据探查

上线前的数据进行测试，查看内容的分布和数据特征，保证数据符合业务预期，避免下游用户因为数据错误导致决策失误

强规则熔断

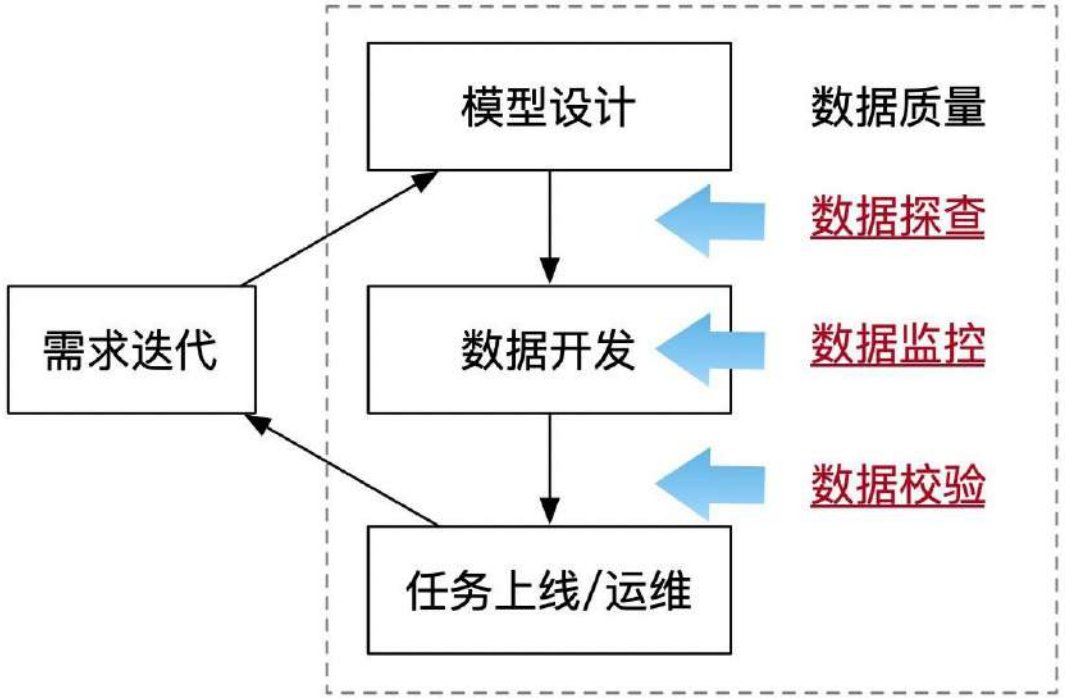
- 事中-数据监控

六要素监控模板，支持多表关联、多行多列等复杂质量规则配置，满足复杂的业务场景需求，强弱规则机制，阻断下游持续污染

数据校验

- 事后-数据对比

丰富的自定义对比能力，验证开发代码逻辑的准确性与数据结果，保障数据按预期产出



关键能力：数据资产发现与管理

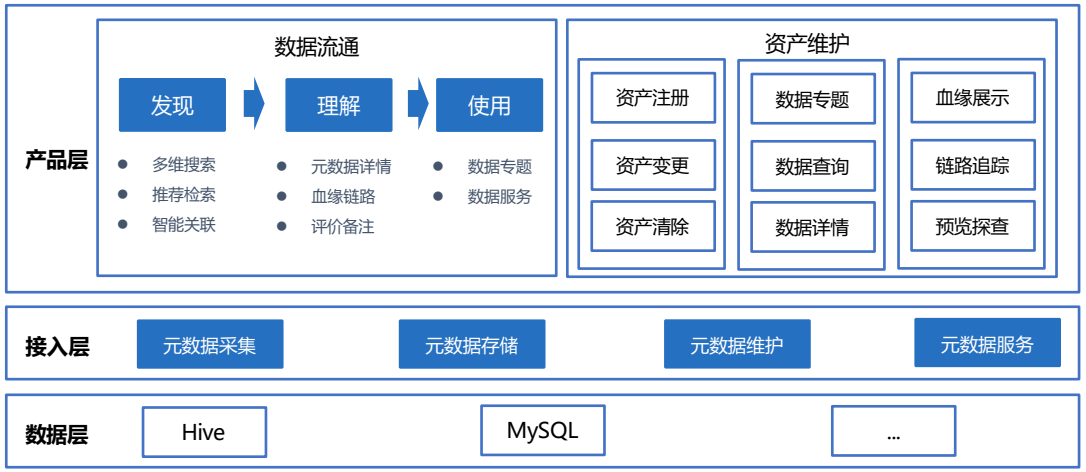
全域数据采集，基于血缘关系，挖掘数据价值

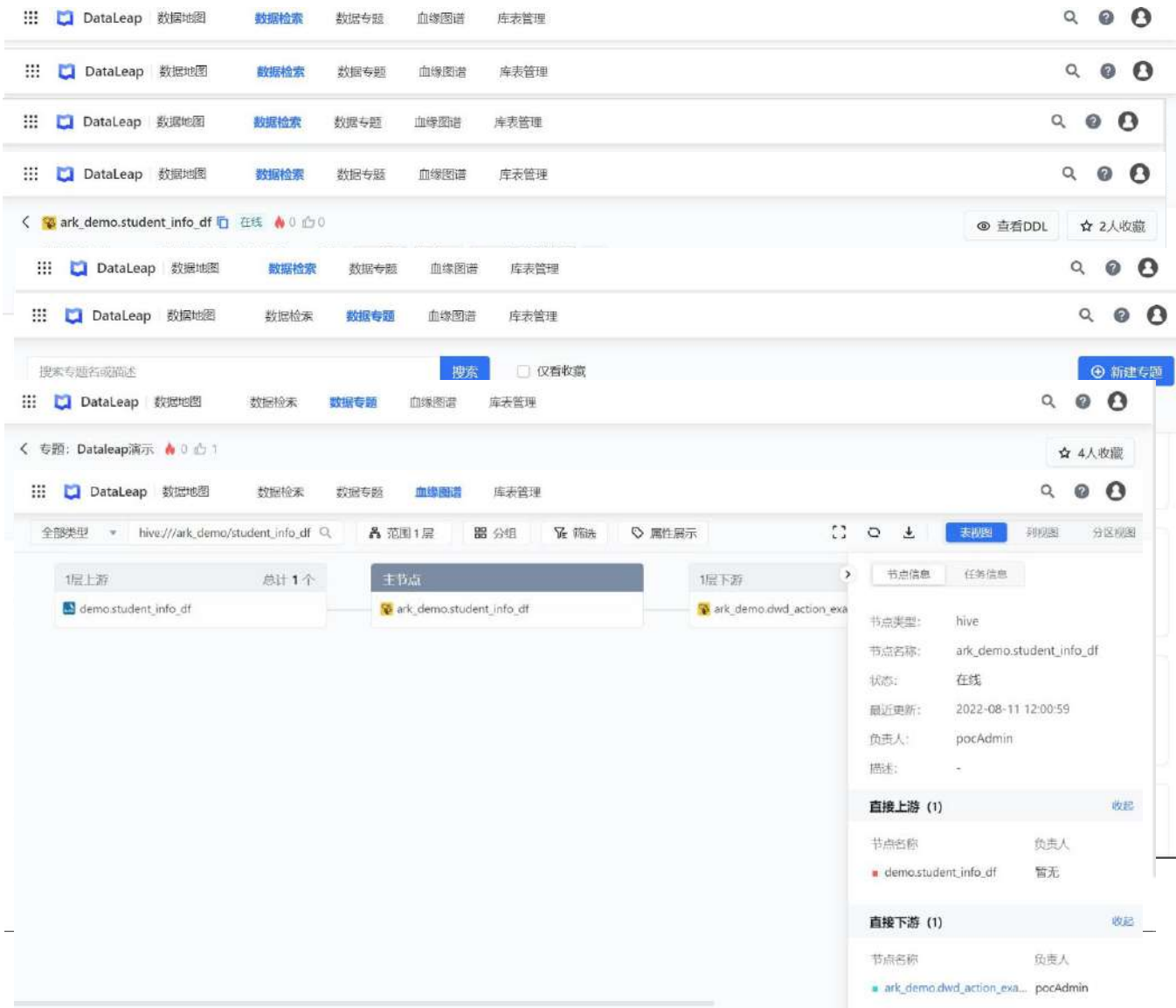
数据检索

- 强大的数据检索、丰富的元数据详细信息，结合数据血缘，帮助用户全面的探索与理解多种类型的数据内容，充分发掘数据的价值

资产管理

- 基于标签的数据资产管理模式，包括业务域、主题、产品线、专题等，提供灵活、全面的管理能力





火山引擎 数据中台

关键能力：内部实践举例-基于DataLeap构建电商实时数仓的资产管理

业务介绍

自2020年4月开始电商业务运营陆续有实时数据诉求以来，业务侧对数据的实时性要求越来越高，原本批处理的数据仓库建设体系无法满足现阶段需求，电商数据BP-实时数仓是服务于电商数据产品、分析师、运营等业务方的数据研发团队，**产出订单、流量等实时数据供下游进行实时看数、实时分析、实时决策。**

业务需求与痛点

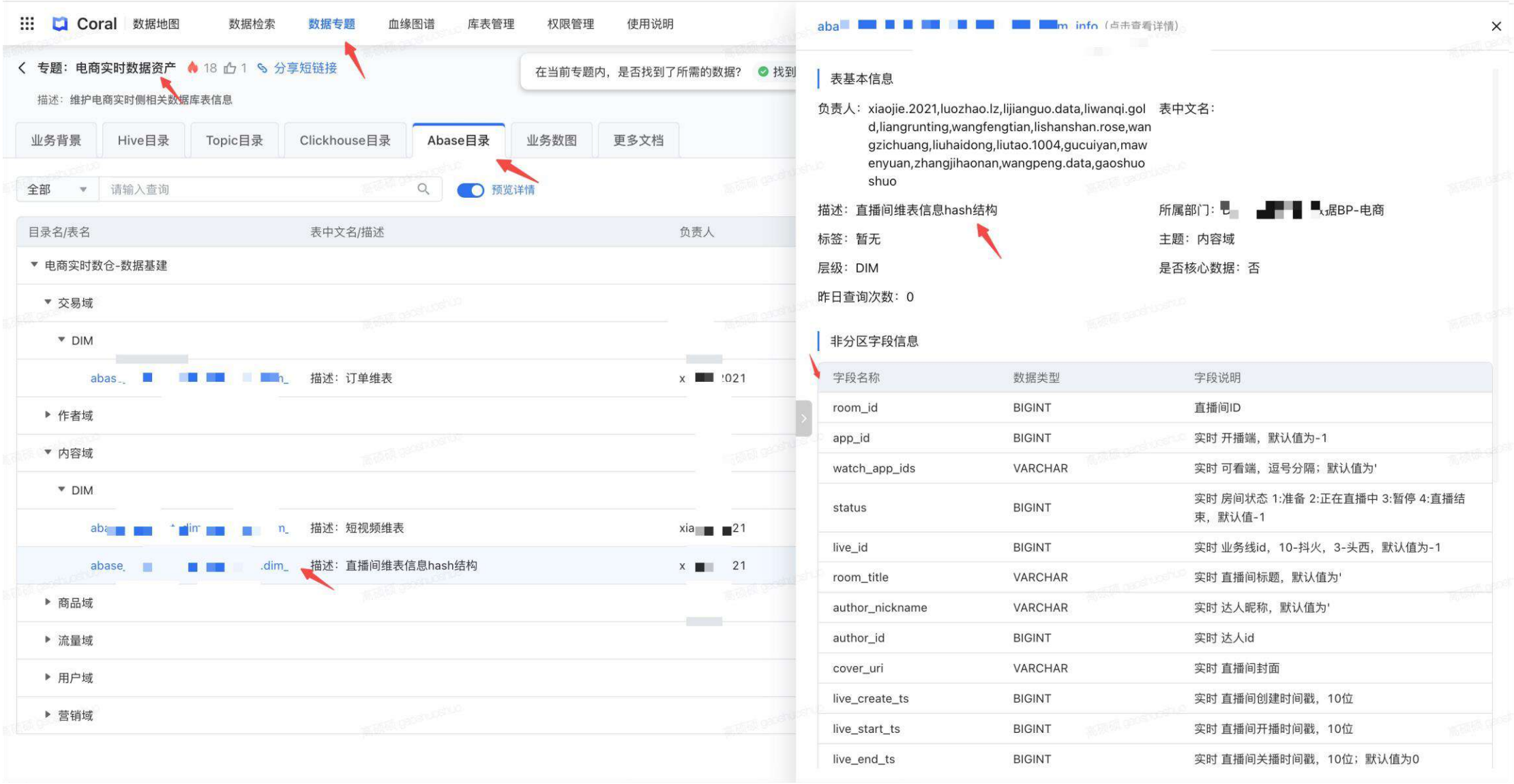
电商实时数仓一共有几千个任务在Leap上进行开发和运维，构建数仓数据体系以及任务维护痛点如下：

1. 实时数仓的都是在任务中进行定义，每次开发都需要手动使用 DDL 语句定义表结构后方能使用，**工作重复繁琐，有时还会引发线上错误。**
2. Flink 读写数据源的表结构等**元信息没有统一的地方维护，**每次定义及使用需要一定的上下游沟通成本或者需要查询手动维护的文档，元信息变动维护成本高且易出错；
3. 实时数仓中表和元数据运行进行数据域和主题的划分，**开发时很难找到可以使用的表信息**，也不明确目前数据域下有哪些表可以使用。
4. **缺少统一的元数据管理平台构建任务和字段级的血缘关系**，一旦出现下游数据某个字段有问题的情况，排查生产链路成本极高

解决方法

目前电商实时数仓已经完成BMQ和Abase元数据表的设计，BMQ完成200+表的录入，Abase表完成20+维表的录入，并且在Leap通过电商实时数据资产专题来维护电商实时数仓的资产内容，通过Leap和元数据+数据专题的功能解决了实时数仓数据表的管理，维护和查询的问题。

关键能力：内部实践举例-基于DataLeap构建电商实时数仓的资产管理



关键能力：数据资产细粒度权限安全管控

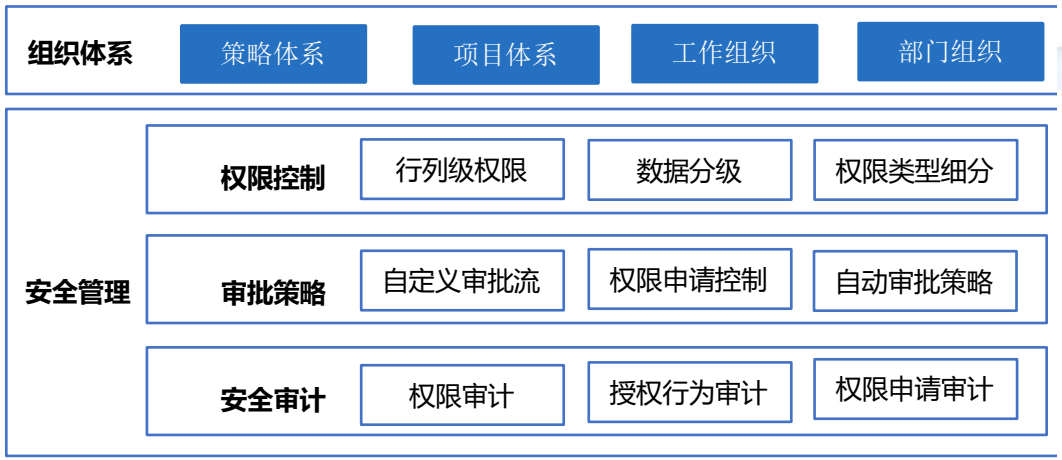
事前、事中与事后全方位保障生产数据安全

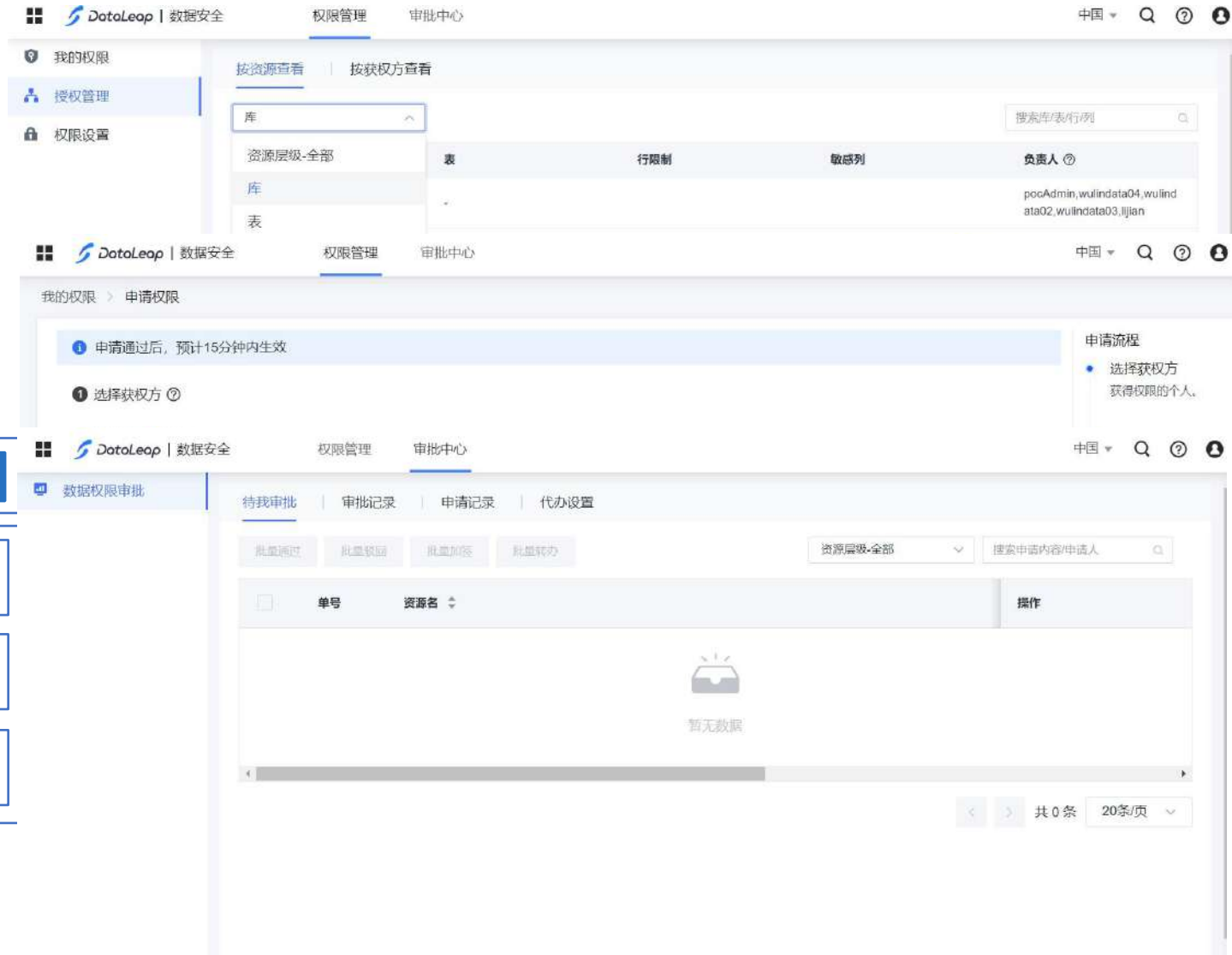
行列权限控制和数据分级管理

- 最小授权原则，支持行列最细粒度的权限管控，精确控制数据权限
- 支持数据分类分级规则

强大的数据审计

- 提供完善的安全审计功能，包括权限审计、授权行为审计、权限申请审计





客户案例

03

案例：得到

客户需求

得到是国内在线教育/音视频内容的头部企业，旗下有得到课程、得到大学、听书、电子书、直播等多个业务线。内容业态覆盖PCG、UGC等多种方向，随着业务的快速发展，在精细化运营、数据治理等方向遇到了不少挑战：

1、知识城邦是UGC社区，需要产出大量的优质内容对专业知识频道进行补充，除了产出优质内容，还要能面向得到用户进行内容精准分发，从而提高用户粘性。

2、随着用户体量的激增，如何有效挖掘用户价值？提高增值服务的渗透和付费转化率是运营迫切需要优化的环节。运营优化高度依赖数据体系建设，得到的数据基建存在埋点管理困难、口径不一致、监测指标无法多方对齐等问题

案例：得到

客户Re景

国内在线教育/音视频内容的头部企业，旗下有课程、大学、听书、电子书、直播等多个业务线。内容覆盖PCG、UGC等多个方向。

业务痛点

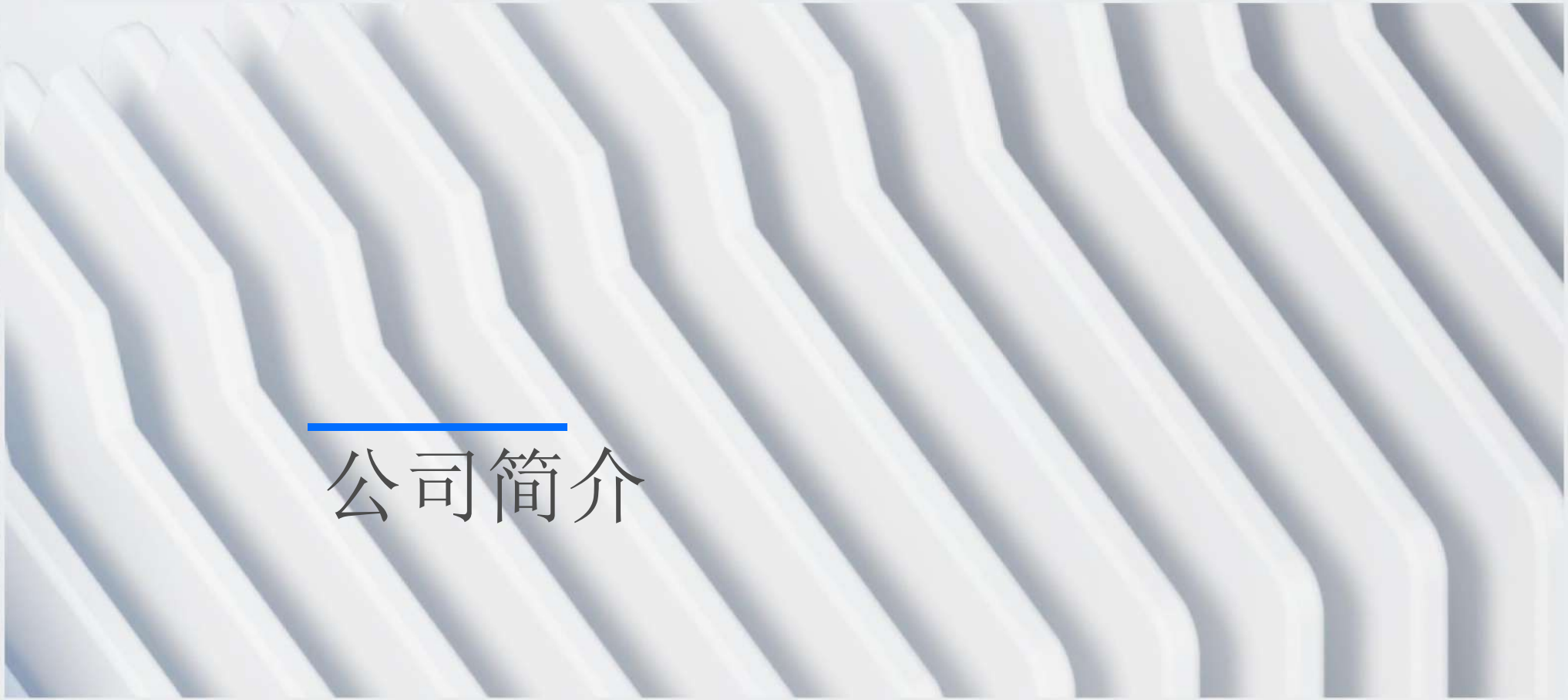
数据治理问题突出。
数据质量差，业务用数无保障。
数据建设缺少规范，数仓层级混乱、数据口径不统一，找数用数困难。
缺少全链路保障，数据产出延迟情况频繁发生，严重影响业务发展。

解决方案

“组织+方法论+平台“
组织上引入字节“数据BP”机制，配合专家咨询服务，搭建可持续治理体系。
结合字节数仓建设白皮书，在建模/数仓分层/研发/管理等方面，建立规范。
通过DataLeap提高数据质量和SLA达成率，解决数据产出延迟和脏数据问题；结合数据地图提高元数据完整度，提高找数用数效率。

效果反馈

客户整体**数据治理能力实现3年的跃进。**
数据治理显著提高，完成**0-1体系搭建**；数仓易用性提升，分析效能**提升60%**；数据研发**提效50%**，4人团队管理超3000任务。



公司简介

04

关于字节跳动

覆盖150+个国家和地区

支持75+语种

抖音日活:6亿
字节月活:19亿

海外

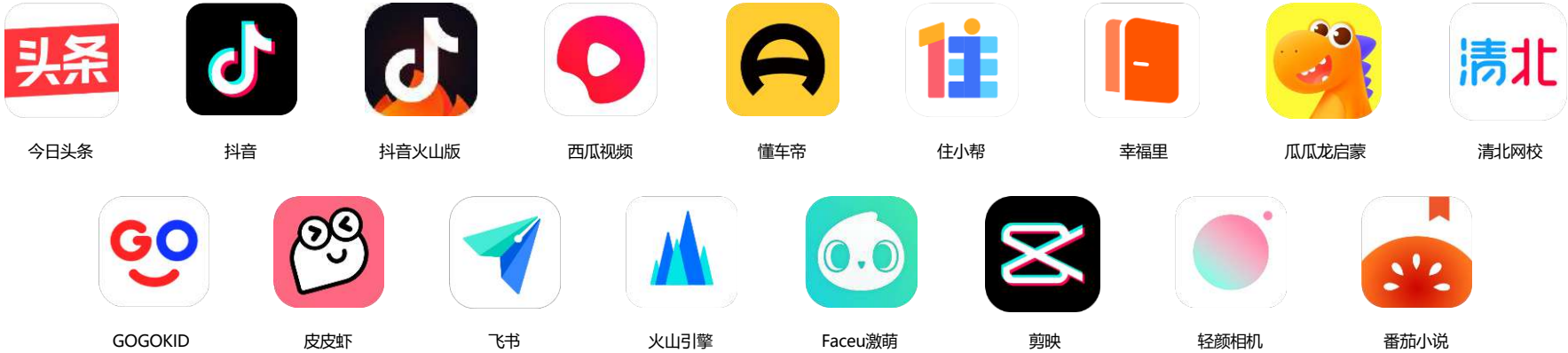
Babe　Helo　TikTok　Lark　Ulike　CapCut　Resso

中国

今日头条　抖音　抖音火山版　西瓜视频　懂车帝　FaceU　图虫

火山引擎 数据中台

字节增长背后的技术与组织能力



内容中台

内容生产 | 内容定制 | 内容审核

生产高质量内容
准确分发到用户
多维度分析效果

内容运营

AI用户体验

智能创作 | 语音 | 视觉 | 推荐算法

AI工程开发与服务

数据中台

数据开发 | 数据资产 | 数据洞察 | 数据运营

用户运营

UG中台

增长链路 | 厂商预装 | 应用商店 | 信息流投放

增长策略 | 冷启承接 | 激励任务 | 频次提升

广告投放
效果分析
投放优化

敏捷研发平台 基于多云的云原生基础设施

火山引擎的使命和愿景

激发创造
释放潜能

火山引擎将字节跳动经过验证的技术积累和成功实践，孵化成能帮助企业实现业务创新、创造业务价值的技术红利，释放企业持续增长的潜能。

火山引擎智能增长技术整体架构

解决方案

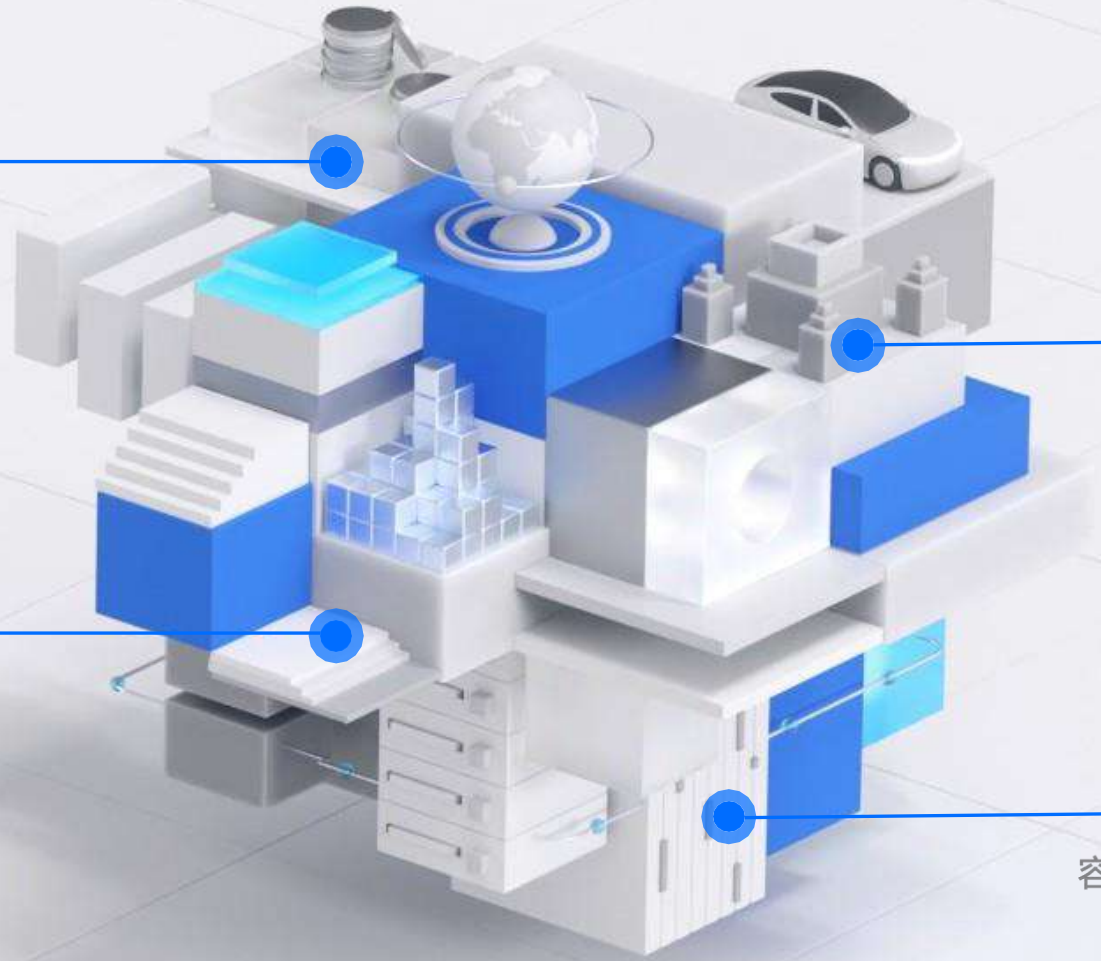
全方位覆盖各行业的业务场景

新零售、汽车、金融、
文旅、泛互联网

技术中台

全链路的用户转化触点

AI中台、多媒体中台
研发中台、数据中台



智能应用

全面的智能体验套件

智能内容、内容创作、智能体验
个性化推荐、智能营销
智能运营、业务安全

统一基础服务

快捷的基础服务搭建能力

容器服务、存储、公有服务、私有云

www.volcengine.com

火山引擎「IGT 智能增长技术」产品架构

火山引擎高效稳定支撑21年春晚活动

搭载了火山引擎的部分企业伙伴

Thanks.

联系人 喻沛枫
电话 18362005736
微信 18362005736
Email 1102090552@gmail.com

火山引擎数据中台

www.volcengine.com