

字节跳动基于 DataLeap的DataOps 实践

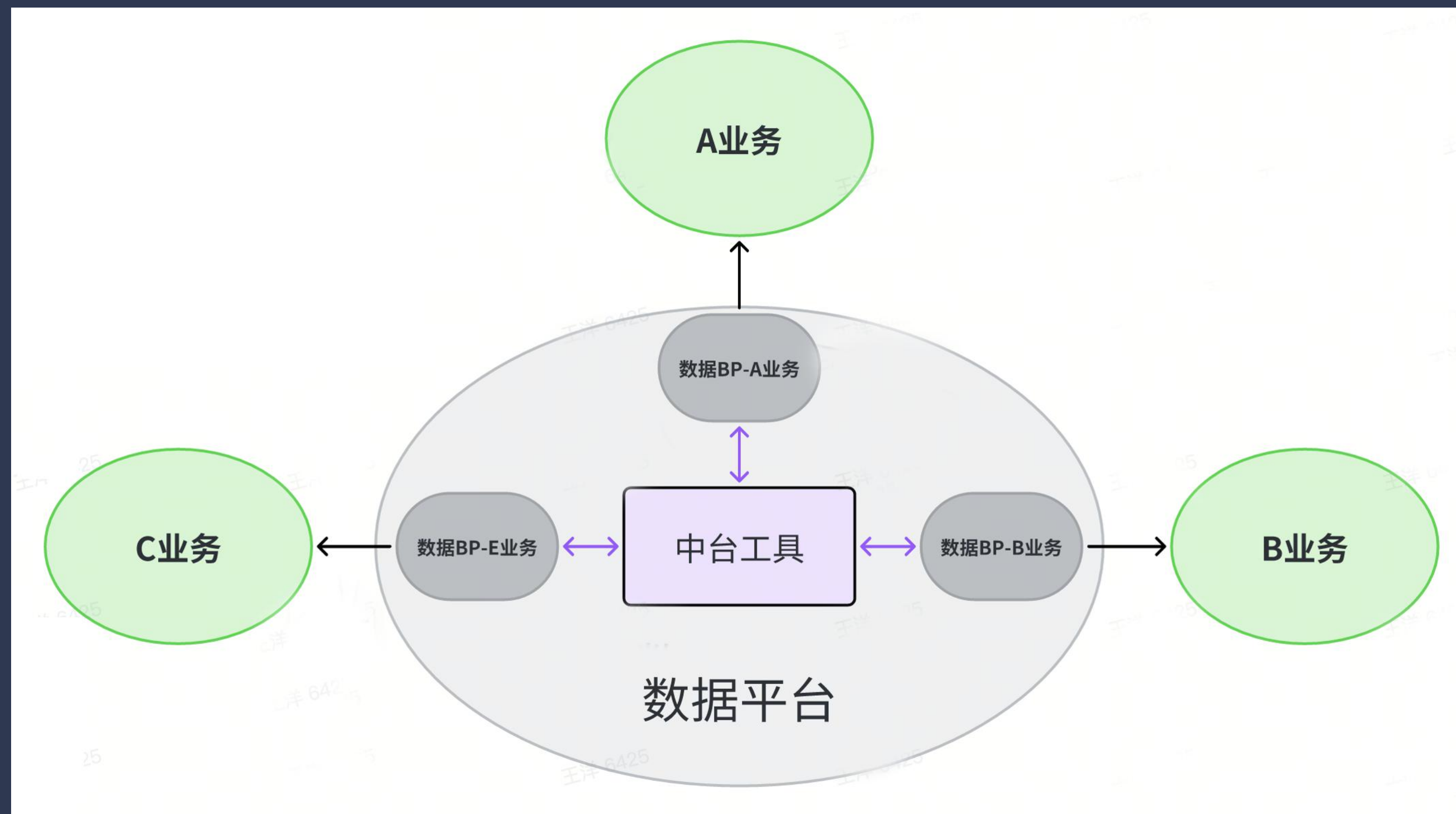
抖音直播数据研发负责人/王洋

目录

- 字节跳动数据研发的模式与挑战
- DataOps理念在字节的具象
- DataOps产品化及落地
- 最佳实践
- 未来展望

字节跳动数据研发的模式与挑战

➤ 中台+数据BP模式



字节跳动数据研发的模式与挑战

➤ 数据BP的业务支持

场景	需求类型	支持模式	优先级
决策	<ul style="list-style-type: none">• 战略决策• 产品决策	<ul style="list-style-type: none">• 战略指标体系的建设和产品化支持• 适配业务场景的数仓建设及分析支持• 内部数据产品如数据门户建设	高
运营	<ul style="list-style-type: none">• 客户运营• 内部运营	<ul style="list-style-type: none">• 对外运营平台中的数据模块建设，提供页面/API/数据表等不同方式的支持• 内部运营所需的BI、取数类支持	中
功能	<ul style="list-style-type: none">• 产品功能• 业务流程	<ul style="list-style-type: none">• 用户产品中的实时/在线数据能力支持• 业务流程链路中的数据支持	低
策略	<ul style="list-style-type: none">• 算法策略• 经营策略	<ul style="list-style-type: none">• 线上算法策略的评估分析• 业务经营策略的支持	低

字节跳动数据研发的模式与挑战

➤ 数据BP的核心指标：0987

0

数据事故数为0

9

需求满足率>90%

8

分析覆盖率>80%

7

用户NPS>70%

字节跳动数据研发的模式与挑战

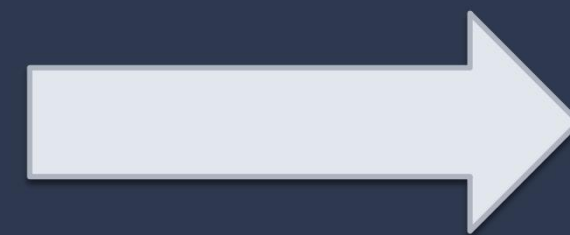
➤ 来自质量挑战

- 链路复杂：最长任务链路节点超X000，单任务1级下游最大超X000
- 变更频繁：每周线上任务变更次数超X000，其中风险场景超X00
- 事故易发：22年全年数据研发事故涉及到研发规范的占比56%

字节跳动数据研发的模式与挑战

➤ 来自硬件成本的挑战

基于预算的
成本控制



基于需求的
精细化控制

字节跳动数据研发的模式与挑战

➤ 来自人效的挑战

- 如何证明团队当前的状态是高效的？
- 如何用更少的人员创造更大的业务价值？

DataOps理念在字节的具象

➤ DataOps的定义

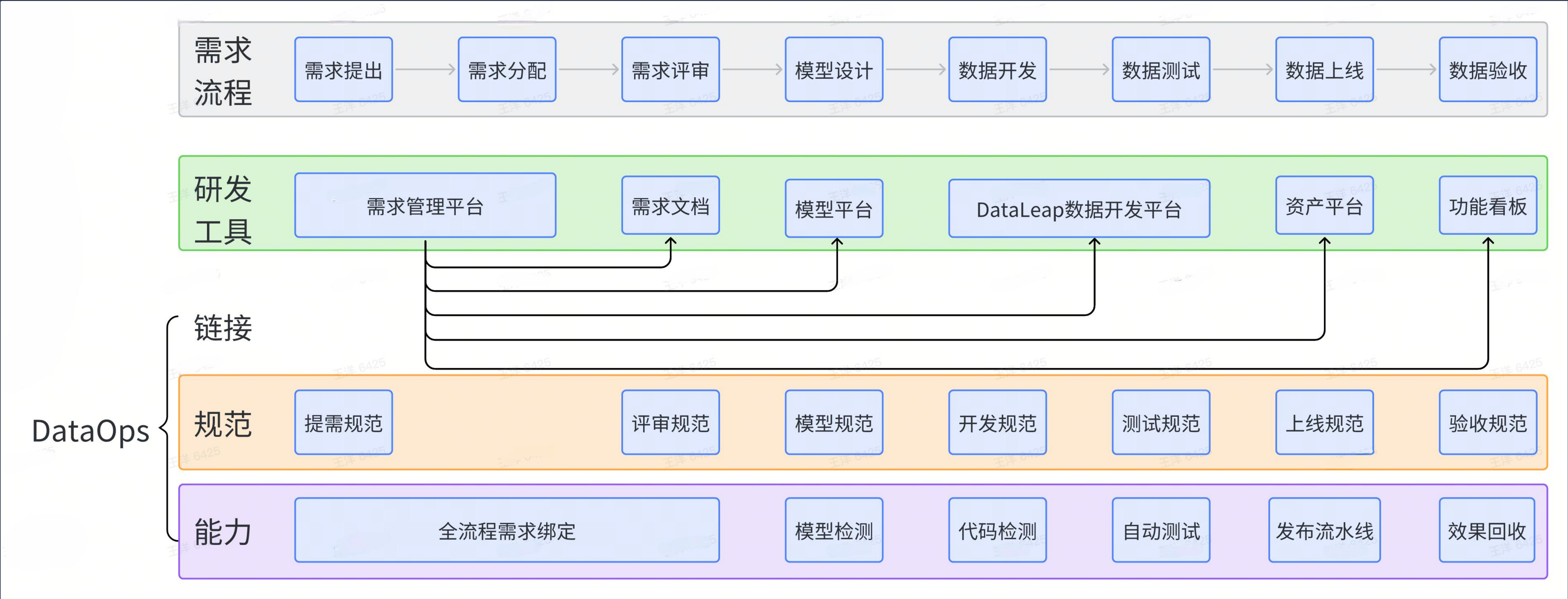
- 数据研发运营一体化（**DataOps**）：是数据开发的新范式，将敏捷、精益等理念融入数据开发过程，通过对数据相关人员、工具和流程的重新组织，打破协作壁垒，构建集开发、治理、运营于一体的自动化数据流水线，不断提高数据产品交付效率与质量，实现高质量数字化发展。

FROM [信通院](#)

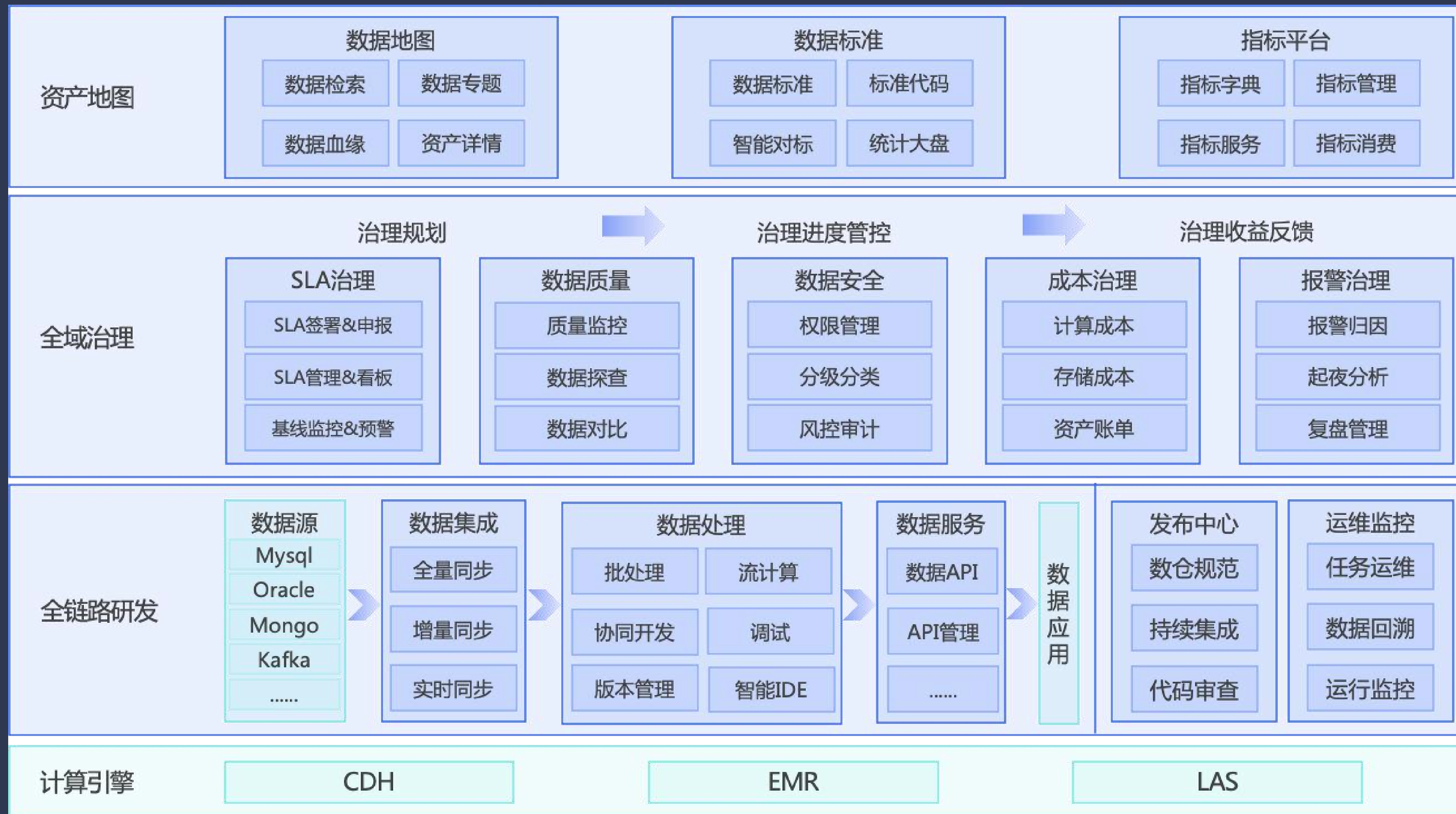
DataOps理念在字节的具象

- 什么是，什么不是？
- DataOps是作用于人+流程+工具的一套方法论，目标是提高数据质量和开发效率，主要通过敏捷协作、自动化/智能化、以及清晰的度量监测，让数据流水线达到持续集成、部署、交付（CI/CD），在DataLeap体系内，DataOps主要以规范研发流程为目的，涵盖对规范研发流程的“已有能力集成”，形成一站式研发体验，同时也包括规范研发流程所需关键的“新能力建设+集成”，除此以外的数据开发基础能力迭代不作为DataOps的一部分

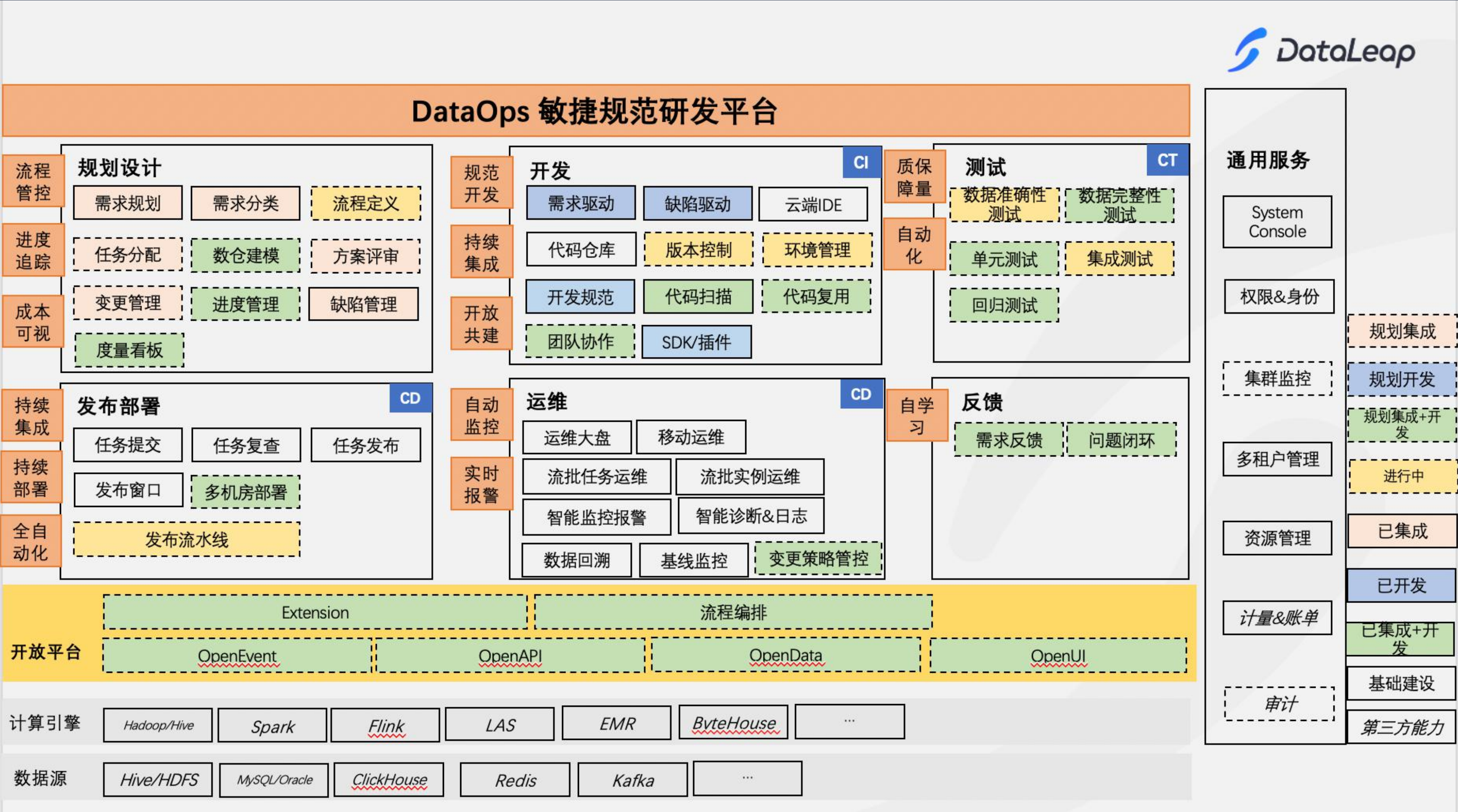
DataOps理念在字节的具象



DataOps产品化及落地-DataLeap

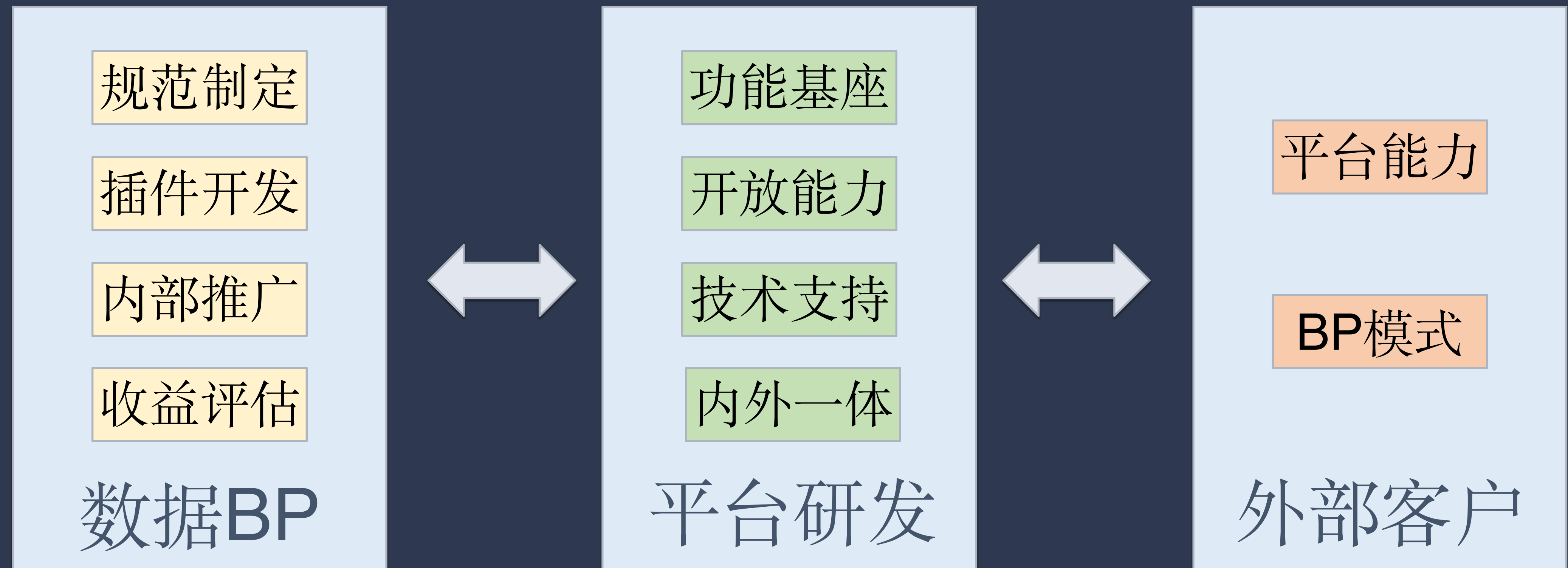


DataOps产品化及落地



DataOps产品化及落地

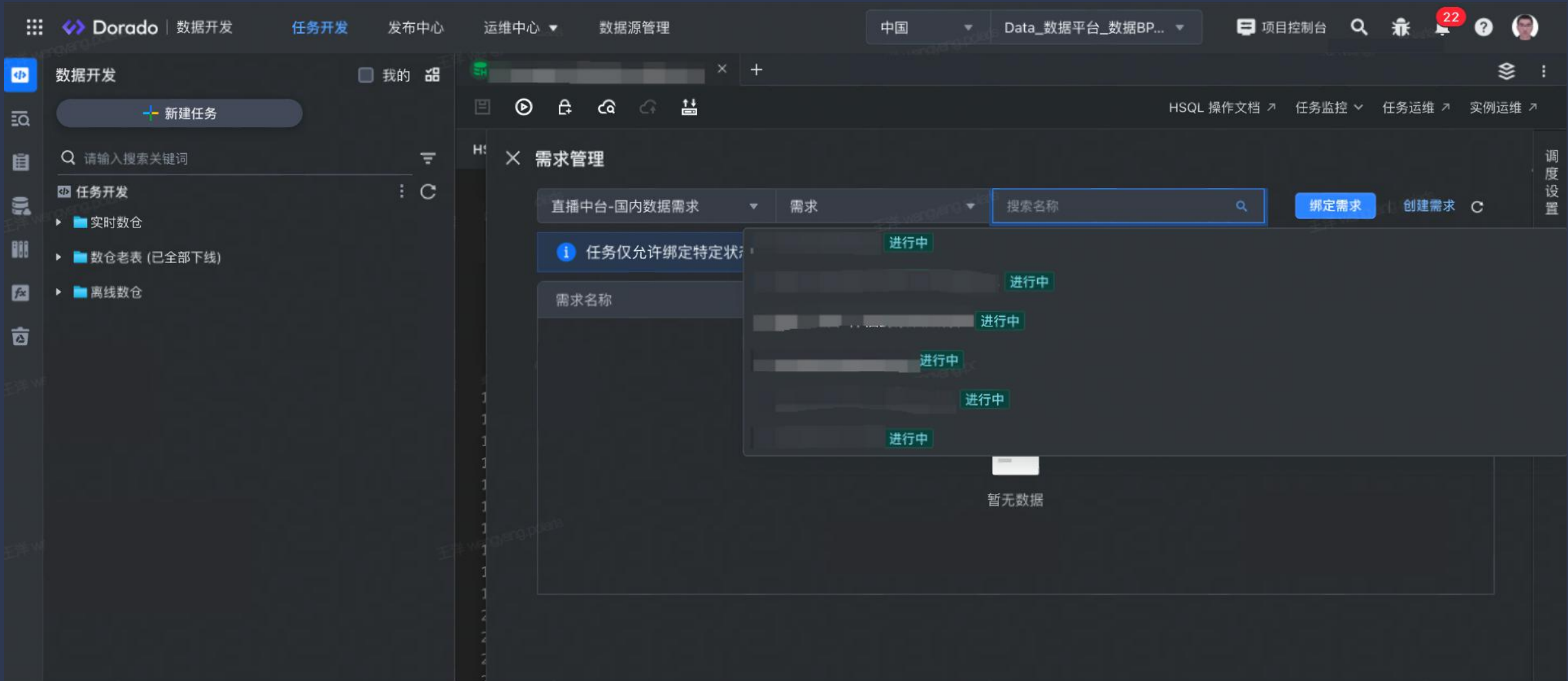
➤ 中台+BP模式



DataOps产品化及落地

➤ 需求管理

- 需求的准入要求
- 需求与开发过程及交付物绑定
- 需求的进度追踪
- 需求的价值评估



DataOps产品化及落地

➤ 流水线管理

- 测试流水线
- 发布流水线
- 离线&实时任务管理
- 任务优先级管理

扩展程序配置

扩展程序: CodeCt

显示名称: CodeCt

ID: d56563dc4f

自定义参数

运维配置

* 生效规则:

* 检测规则:

* 规则强度:

参数:

白名单:

生效时间:

上线时禁止写入测试分区

关键字select *禁用

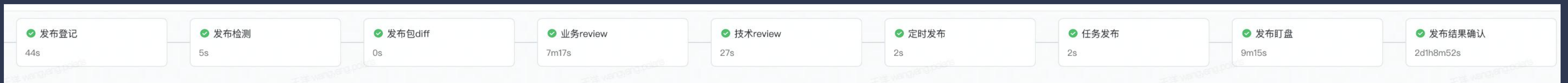
参数spark.executor.memory阈值

参数spark.dynamicAllocation.maxExecutors禁用

参数spark.driver.memory阈值

任务名和表名一致性校验

上线代码版本检查



最佳实践

- 推广运营：如何在公司范围内大规模落地DataOps?

鲶鱼效应

让某些团队先跑起来

拆箱即用

提供低成本的切换路径

自顶向下

先让leader认清价值

最佳实践

➤ 指标牵引



最佳实践

- 管理者视角：围绕数据开发团队的价值和未来，通过开放让数据团队有可输出的专业价值

业务
价值

专业
价值

最佳实践

- 开发者视角：如何获得工作中的成就感？
- 认可&执行：规范本身是反人性的，在团队内落地DataOps需要充分沟通，结合团队调整与个人发展，讲清为什么，避免粗暴落地
- 参与&贡献：构建人人可参与的开发环境，让数据开发可以深度的参与流程制定与落地的过程中来，促进个人影响力的提升

最佳实践

➤ 收益度量

- **规范**：在不同方向上规范制定与复用，保障流程100%落地
- **质量**：系统性的解决风险场景上的研发流程问题，因研发流程导致的数据质量事故数归0
- **效率**：预计可提升研发在业务需求满足中的开发效率10%+

未来展望

➤ 业务价值

- 数据需求价值度量标准
- 基于需求价值最大化的调度策略

未来展望

➤ 质量与效率

- 基于大模型的需求对接能力
- 基于大模型辅助开发的能力
- 低成本的数据测试及验证能力

未来展望

➤ 对外开放

- DataOps理念在字节落地的成果未来也会通过火山引擎DataLea对外输出，敬请期待
- **火山引擎DataLeap：一站式数据中台套件**，帮助用户快速完成数据集成、开发、运维、治理、资产、安全等全套数据中台建设，帮助数据团队有效的降低工作成本和数据维护成本、挖掘数据价值、为企业决策提供数据支撑。



The screenshot shows the DataLeap product page with a blue and white color scheme. The main heading is '大数据研发治理套件 DataLeap'. Below it is a descriptive paragraph. The page features a navigation bar with '立即购买' (Buy Now), '控制台' (Control Panel), and '产品咨询' (Product Consultation). Below the navigation bar are four main sections: '帮助文档' (Help Docs), '快速入门' (Quick Start), '产品定价' (Product Pricing), and '数智平台VeDI' (Digital Platform VeDI).

大数据研发治理套件 DataLeap

一站式数据中台套件，帮助用户快速完成数据集成、开发、运维、治理、资产、安全等全套数据中台建设，提升数据研发效率、降低管理成本。搭配EMR/LAS大数据存储计算引擎，加速企业数据中台及湖仓一体平台建设，为企业数字化转型提供数据支撑

[立即购买](#) [控制台 >](#) [产品咨询 >](#)

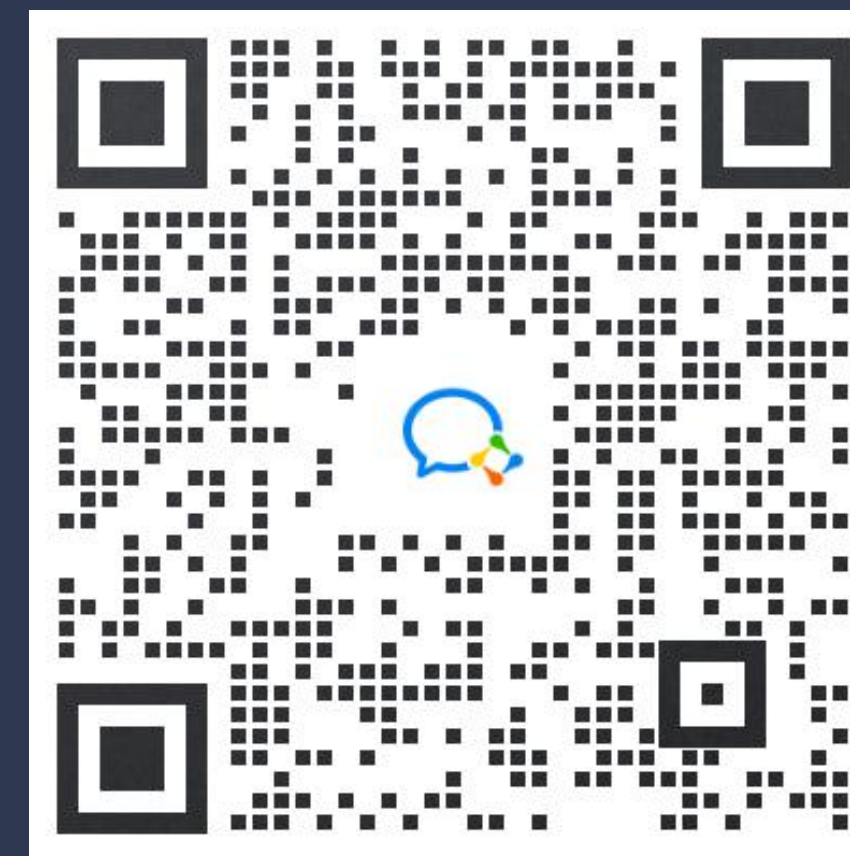
帮助文档 学习完整产品使用方法	快速入门 新手体验&简单上手	产品定价 限时特惠版来袭	数智平台VeDI 了解数智平台产品家族
---------------------------	--------------------------	------------------------	-------------------------------

关于我们



进入火山引擎DataLeap官网

了解更多产品信息



进入官方交流群

获取更多技术干货、活动信息

想一想，我该如何把这些
技术应用在工作实践中？

THANKS