TOP1 100

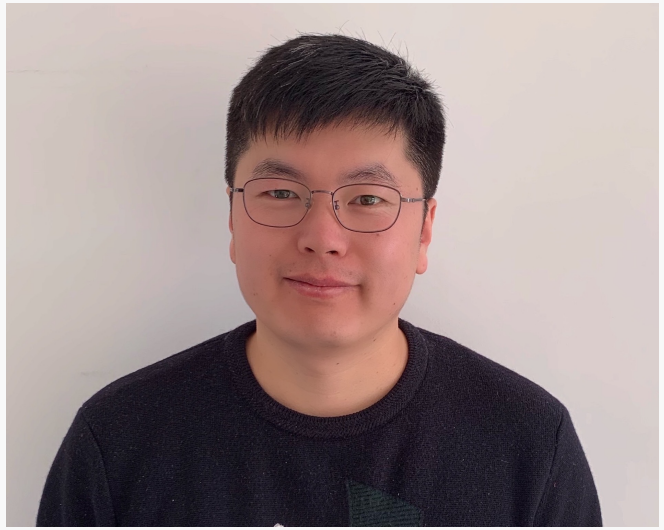
主办方 msup

《高效保障数仓一致性-美团数仓智能构建与治理平台的探索》

宋洪鑫 美团 高级技术专家

个人简介

个人简介



宋洪鑫
美团高级技术专家

- 2011~2013 曾就职于阿里，从事数据实时计算工作。
- 2014~至今 一直在就职于美团，近十年一直专注于大数据开发解决方案领域。
  - 负责离线和实时场景大数据开发平台工具建设，包括数仓建模工具、任务开发、测试与质量保障、以及元数据服务等。
  - 在大数据开发平台工具系统设计、数据安全、数据质量治理等方面有较多的经验。

目录

- 亮点介绍
- 数仓一致性问题
- 智能化解决方案
- 落地效果
- 总结与展望

亮点介绍

- **指标定义一致**：基于数据域、业务过程的统一定义规范，确保指标定义无重复、无歧义。
- **加工一致**：基于明细模型定义指标唯一的计算口径，所有下游ETL自动复用，同步修改。
- **消费一致**：基于模型内容与指标维度语义，提供统一的查询路由元数据服务。

- **应用层开发提效**：基于组件模型语义，复杂指标公式，预聚合配置，自动生成ETL任务等。
- **组件层配置提效**：根据词根快速定义模型和指标维度的计算口径，表之间的ER关系；双向门设计支持逆向建模等。
- **变更提效**：一处修改全局生效，可级联变更、测试、发布、回溯等。

- **数仓规划服务**：顶层设计数据域、业务过程、总线矩阵等指导和约束数仓规划和建设。
- **规范落地**：正向建模，强校验分层、主题、词根等规范100%落地

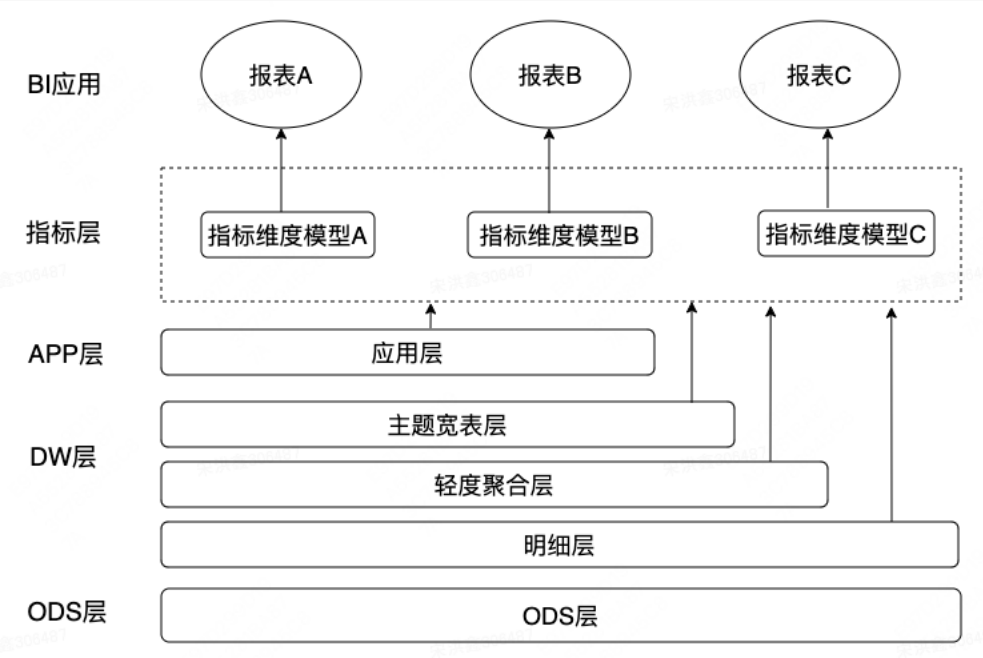
- **指标维度资产**：指标之间的层级关系、计算口径清晰可见，以及可被分析的维度支撑应用分析评估。
- **模型资产**：模型与指标、业务过程绑定、增强理解数，提升找数效率。

目录

数仓一致性问题背景

1、**数据指标多，增长快**：随着业务的发展，对数据的需求更加全面、更精细，多数业务指标都膨胀到了千级万级，如某个业务就有4,000+等。

2、**应用层开发需求多**：ETL任务占比40%+，日常新增变更任务应用层也占大头（60%+），秒级响应要求的数据产品、报表的需求更多来自应用层数据支撑。



烟囱建设

痛点：

1、**口径乱、一致性问题多**：从不完统计的检测结果看，个别业务不一致占比在10%以上。

2、**事后治理，难收敛**：仅通过下线，换指标名的方式，治标不治本。

3、**应用层开发效率低**：复用性差，ETL任务开发时间长。

4、**指标口径变更协作难**：一个指标口径变更，要人工同步修改多个ETL保持一致。

数仓一致性问题背景

指标一致性(同名不同义、同义不同名、同名同义不同值)问题分析

定义：40%

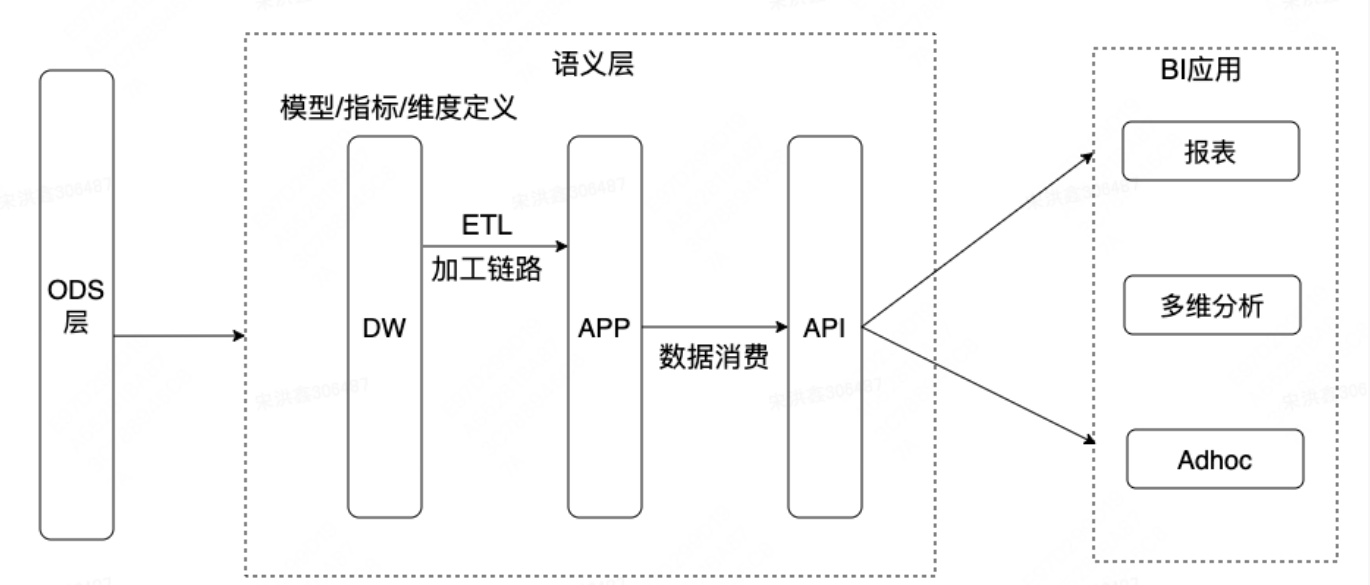
- 缺定义管理标准
- 制定标准未有效落地

加工链路：40%

- 数据源不一致
- 加工逻辑不一致
- 数据回溯周期不一致

应用（API）：20%

- 多消费口径不一致
- 指标维度二次计算不一致



一致性问题解决思路

指标维度定义一致

- 提供标准的顶层管理方法
- 系统化保障标准有效落地

结构化、唯一管理

加工链路一致

- 指标和表的技术口径可管理
- 上游加工逻辑与下游保持同步
- 数据回溯周期保障一致

自动化构建下游ETL
（复用上游逻辑）

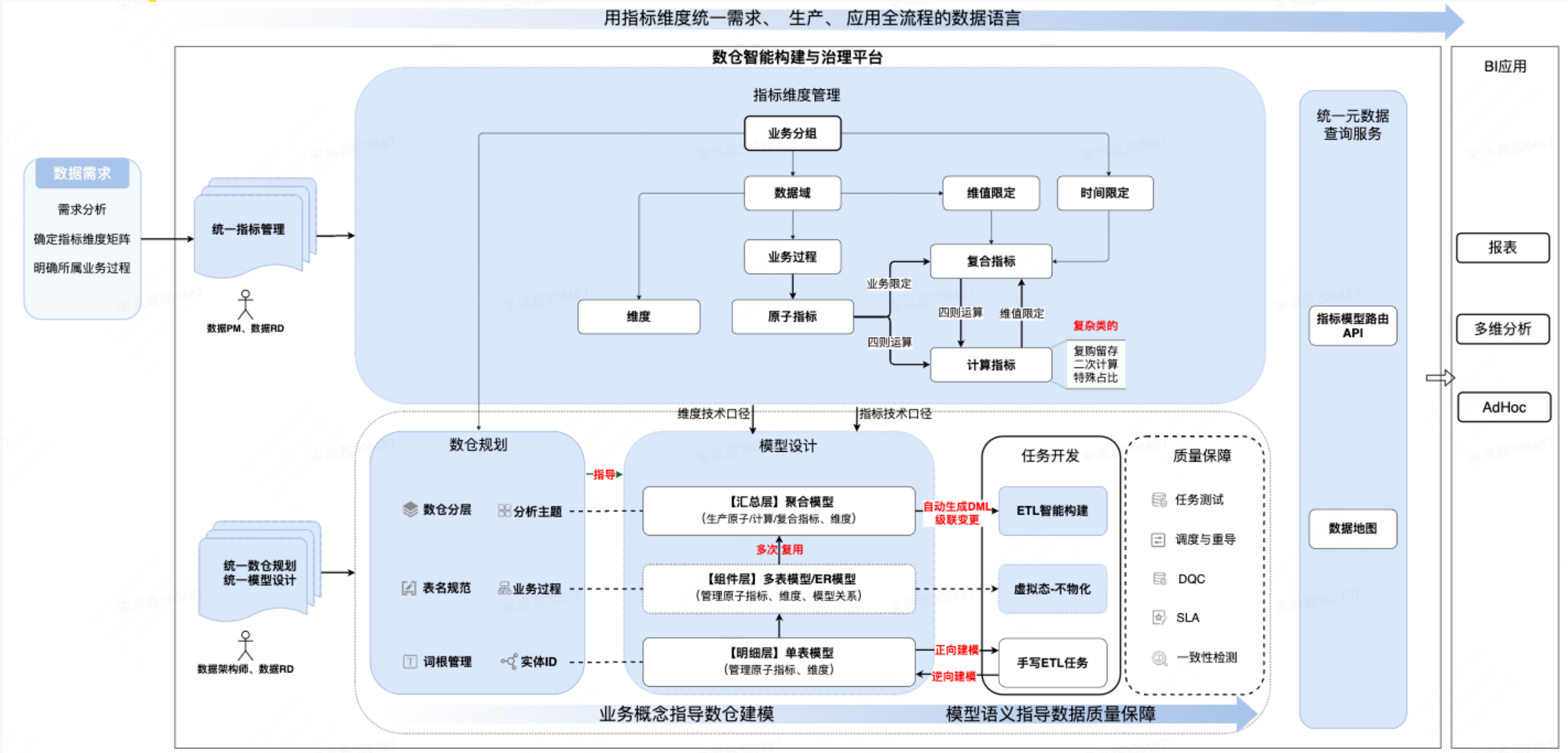
消费一致

- 统一的指标查询元数据路由

查询收口到统一API

目录

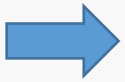
智能化解决方案


用指标维度统一需求、 生产、 应用全流程的数据语言
数仓智能构建与治理平台
数据需求
需求分析
确定指标维度矩阵
明确所属业务过程
统一指标管理
数据PM、数据RD
指标维度管理
业务分组
数据域
维值限定
时间限定
业务过程
复合指标
维度
原子指标
业务限定
四则运算
维值限定
复杂类的
四则运算
计算指标
复购留存
二次计算
特殊占比
统一元数据查询服务
指标模型路由API
数据地图
BI应用
报表
多维分析
AdHoc
维度技术口径
指标技术口径
统一数仓规划
统一模型设计
数据架构师、数据RD
数仓规划
指导
数仓分层
分析主题
表名规范
业务过程
词根管理
实体ID
模型设计
【汇总层】聚合模型
（生产原子/计算/复合指标、维度）
多次复用
【组件层】多表模型/ER模型
（管理原子指标、维度、模型关系）
【明细层】单表模型
（管理原子指标、维度）
自动生成DML
级联变更
正向建模
逆向建模
任务开发
ETL智能构建
虚拟态-不物化
手写ETL任务
质量保障
任务测试
调度与重导
DQC
SLA
一致性检测
业务概念指导数仓建模
模型语义指导数据质量保障

方案核心挑战

业务视角

- 一致性是否可持续?
- 且整体数据需求支持效率是否会降低?
- 成本是否可控?



平台挑战

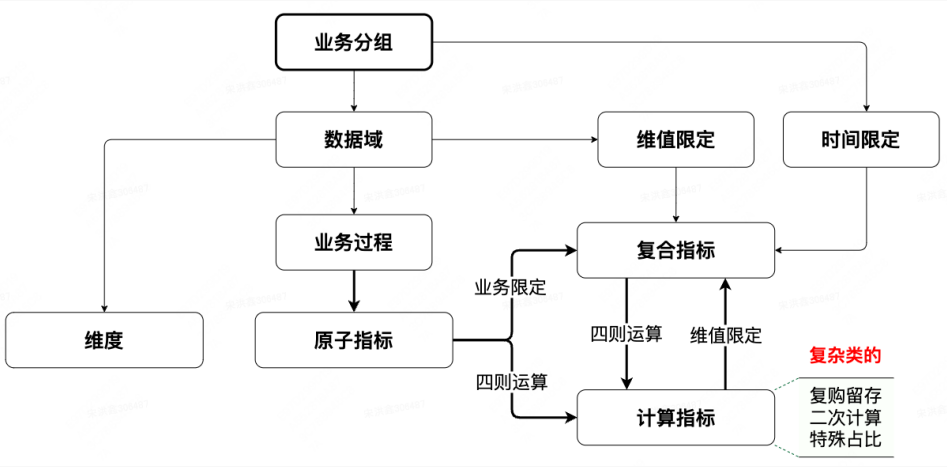
- **一致性保障**：方案复杂度、理解成本、落地可行性
- **开发效率（场景覆盖）支持**：指标维度定义、ETL构建语义支持等
- **任务执行性能保障**：可调优、可运维能力
- **兼顾业务高优需求**：开发和治理并行

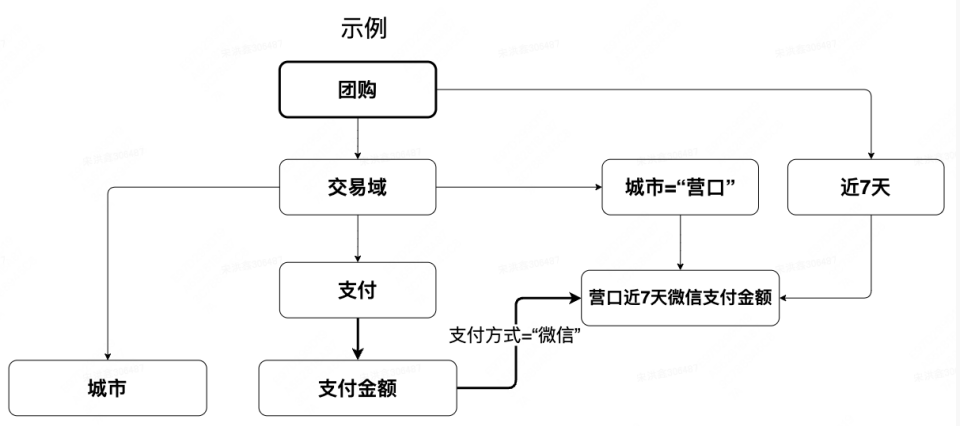
一致性保障-指标口径管理

参考：Kimball维度建模、阿里onedata理论

维度、指标结构化分层管理

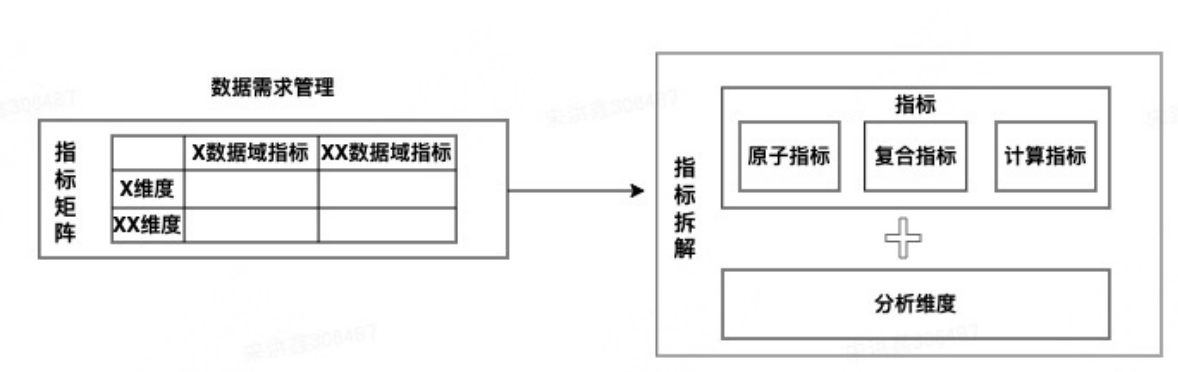
解决“同名不同义”问题





一致性保障方法：

- 原子指标通过归属业务过程区分
- 跨数据域指标通过结构化拆解区分



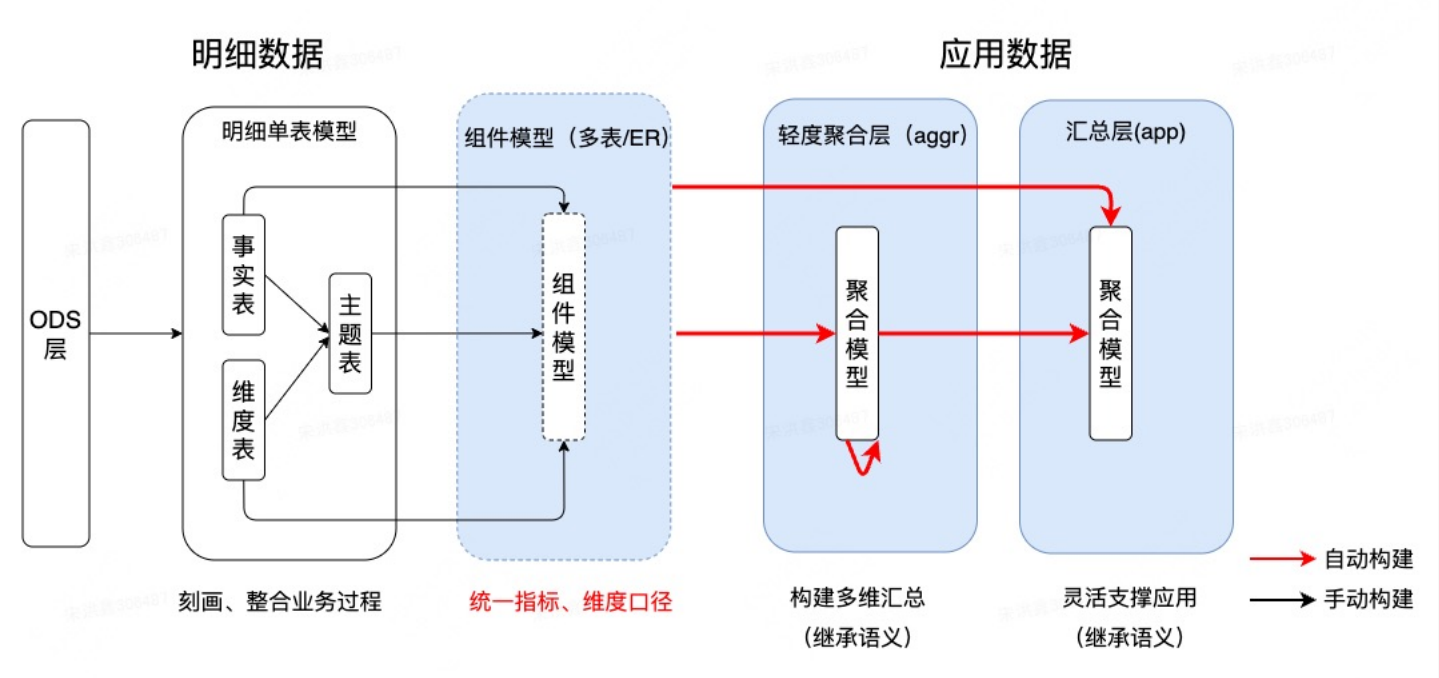
一致性保障-数仓建模

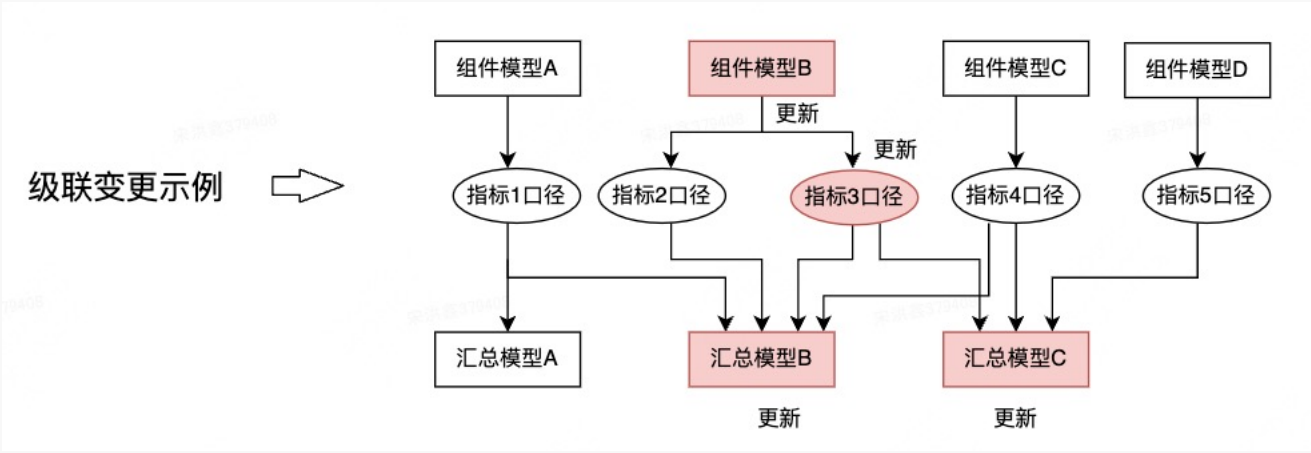
数仓分层变化：

- 新增组件模型
- 聚合层以上变薄

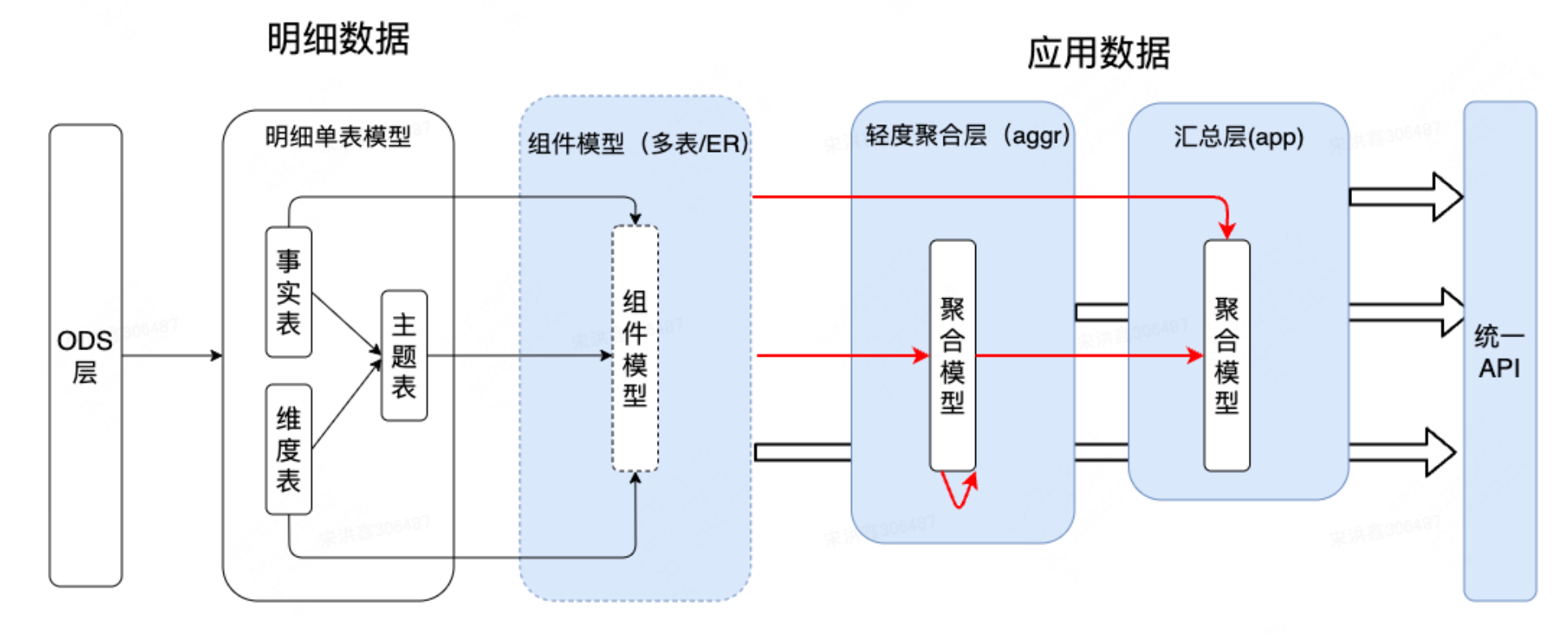
一致性保障方法：

- 是否指标有相同的表字段计算达式（解决同义不同名问题）
- 手动构建指标口径，统一到组件模型，确保唯一，下游模型通过自动构建复用逻辑，继承语义。（解决多口径下同名不同义问题）
- 级联变更、同步重导：组件层口径修改，级联变更下游构建逻辑。（解决链路变更、回刷周期不一致问题）





一致性保障-查询路由



- 一致性保障方法：通过统一的模型路由服务，保证明细组件模型、聚合模型均可查且语义保持一致（解决消费不一致问题）
- 路由策略：根据查询的信息（指标、分析维度、限定条件等）筛选模型，并权衡因素（性能、就绪时间等）进行最优选择。 因此模型语义非常关键，决定了查询效率和数据质量。

开发效率支持-构建语义

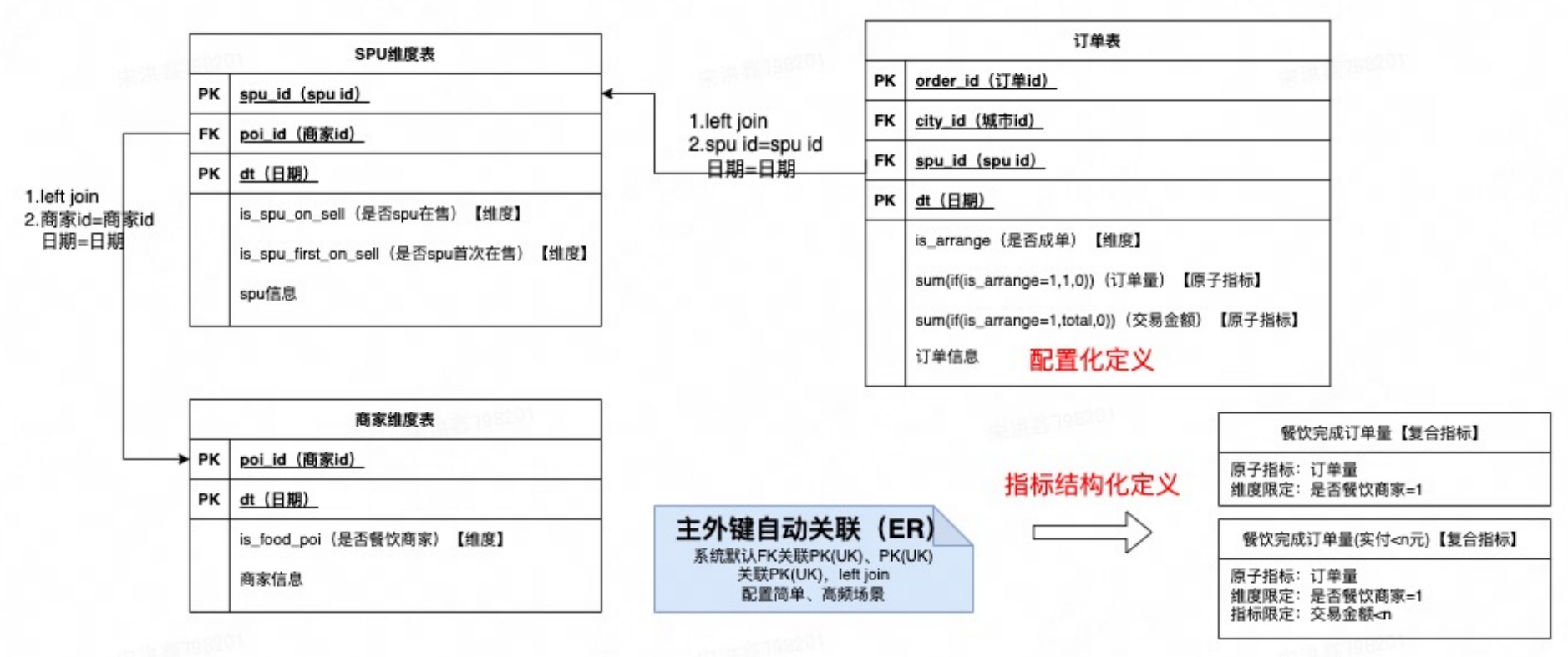
参考：模型配置方式上 参考MetricFlow

组件模型的构建方式一：

- ER模型：通过打主外键，自动联想构建关联。

模型配置语义：

- 字段绑定
- 虚拟表达式
- 自定义UDF



操作更加简单，一次配置多次引用，但关联、配置语义有限

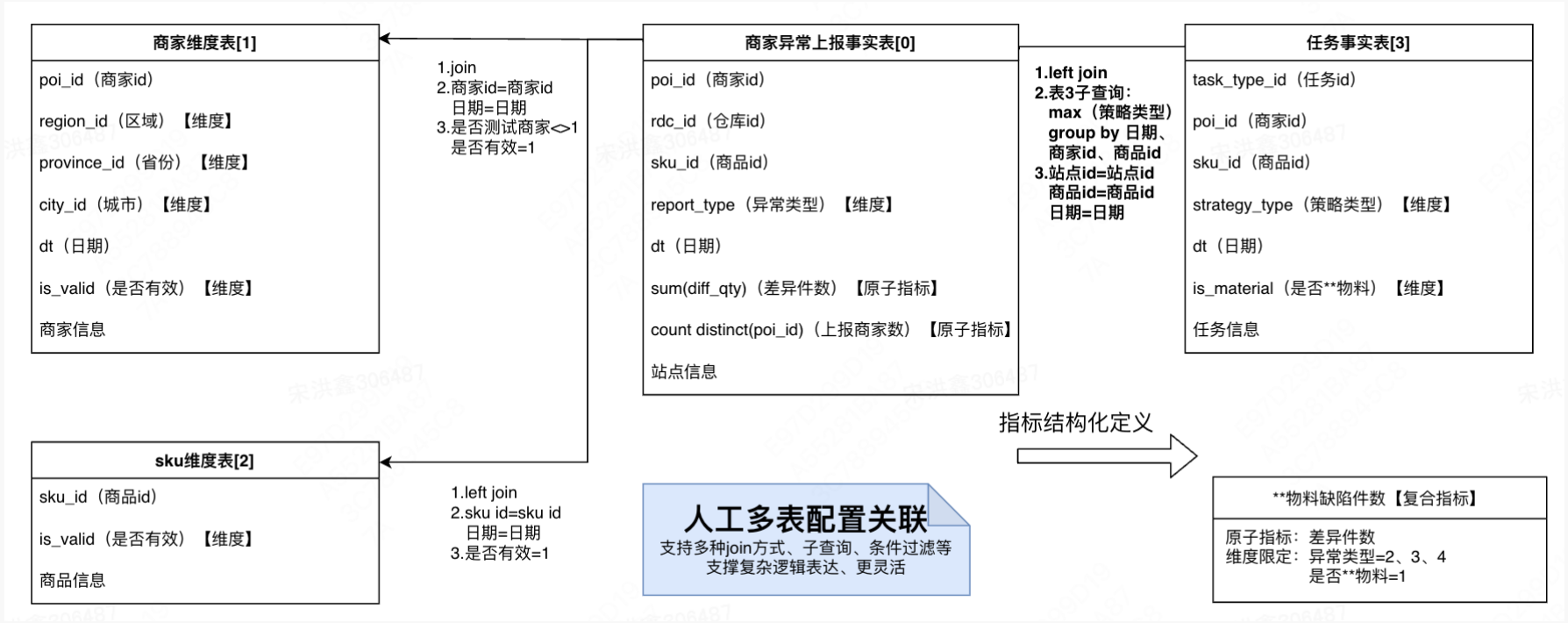
开发效率支持-构建语义

组件模型的构建方式上二：

- 多表模型：人工配置关联语义。

组件模型配置语义：

- 字段绑定
- 虚拟表达式
- 自定义UDF
- 开窗、侧视图等



支持场景更加灵活、可兼顾更多个性化需求的**表意能力**

开发效率支持-构建语义

聚合模型语义：

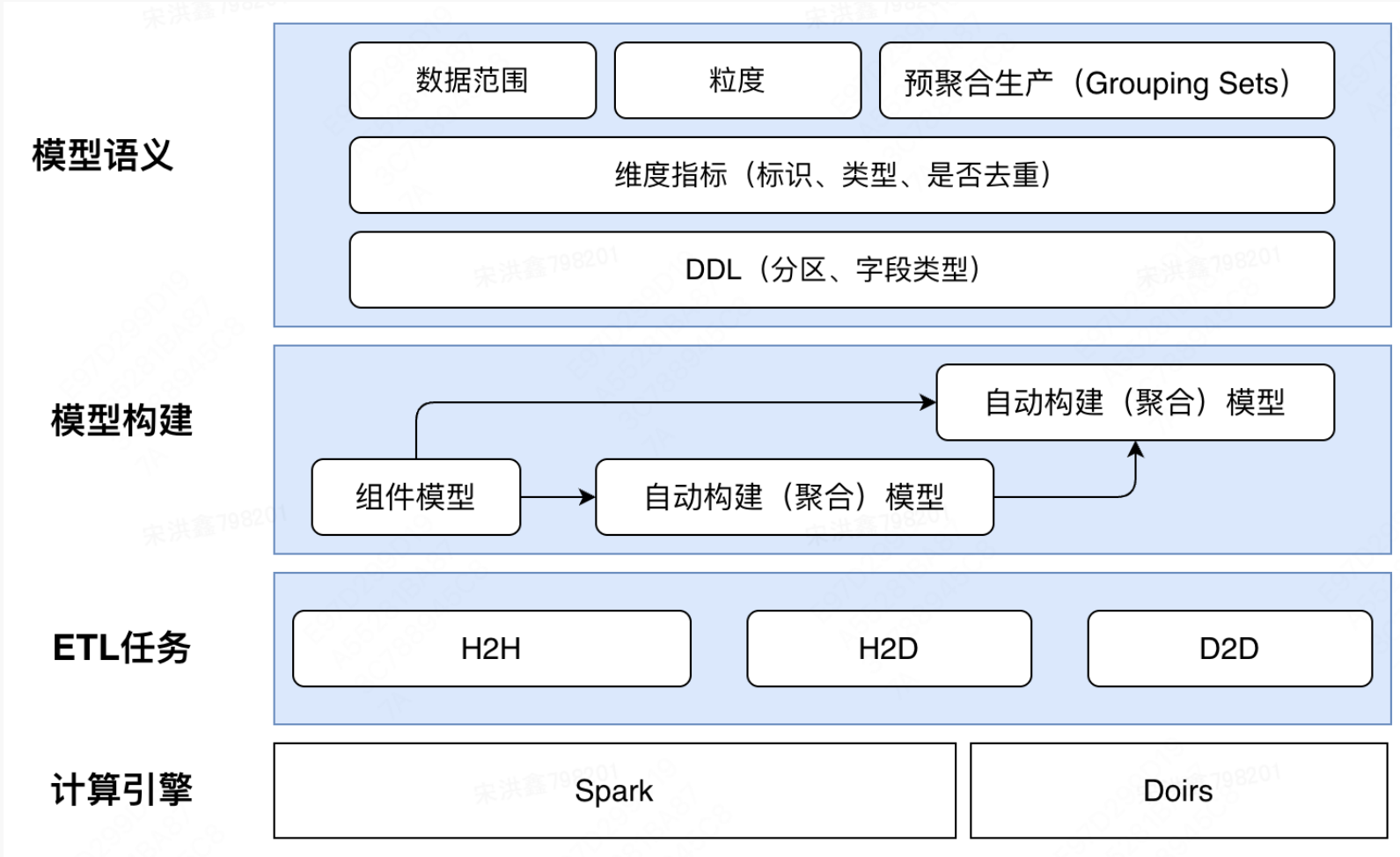
- 分组聚合
- 二次去重指标
- 生成模型数据范围

多种聚合模型构建路径：

- 组件到聚合
- 聚合到聚合。

支持多种ETL任务类型：

- Hive2hive
- Hive2doris
- doris2doirs 运行在spark、doris等引擎上。



覆盖更多聚合模型ETL生产场景

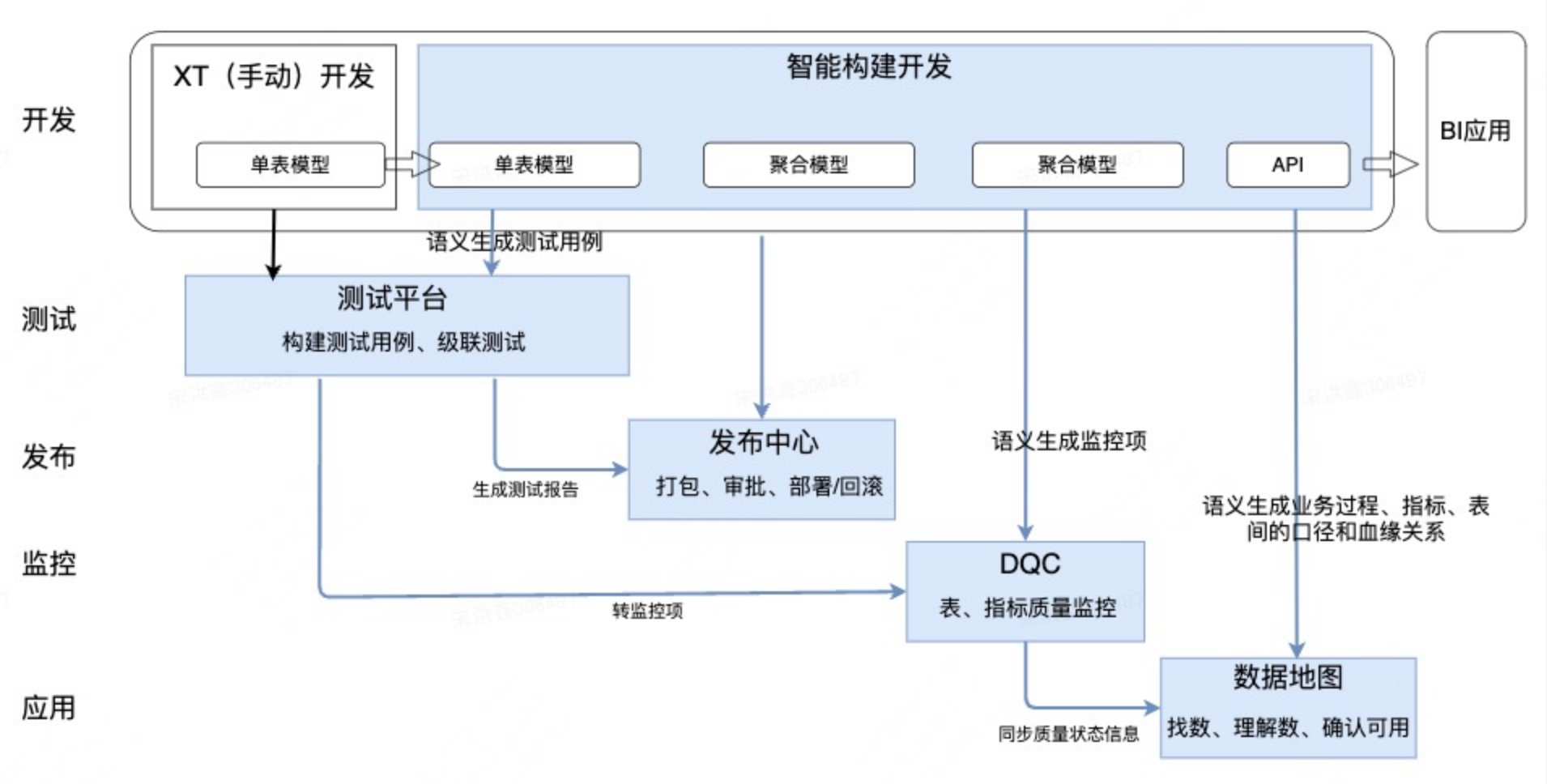
开发效率支持-工具体系

端到端的测试：
- 级联测试
- 测试用例生成
- 指标结果对数

语义指导质量监控：
- 主键重复
- 值域约束
- 漏斗关系

数据地图：
- 业务概念导航
- 找指标、找明细
- 资产沉淀与传承

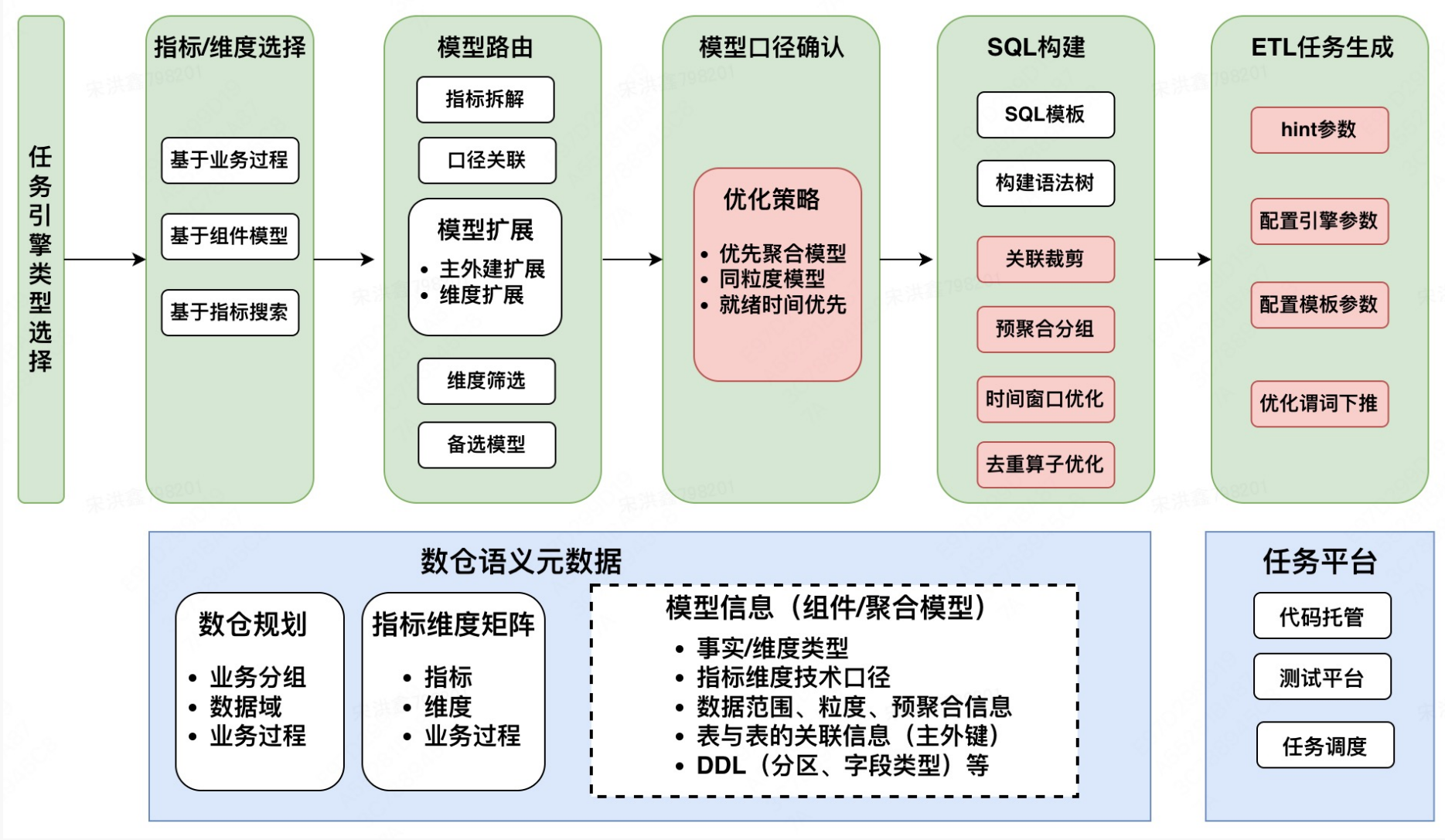
打通各个环节、形成正向循环



任务性能保障

性能保障:

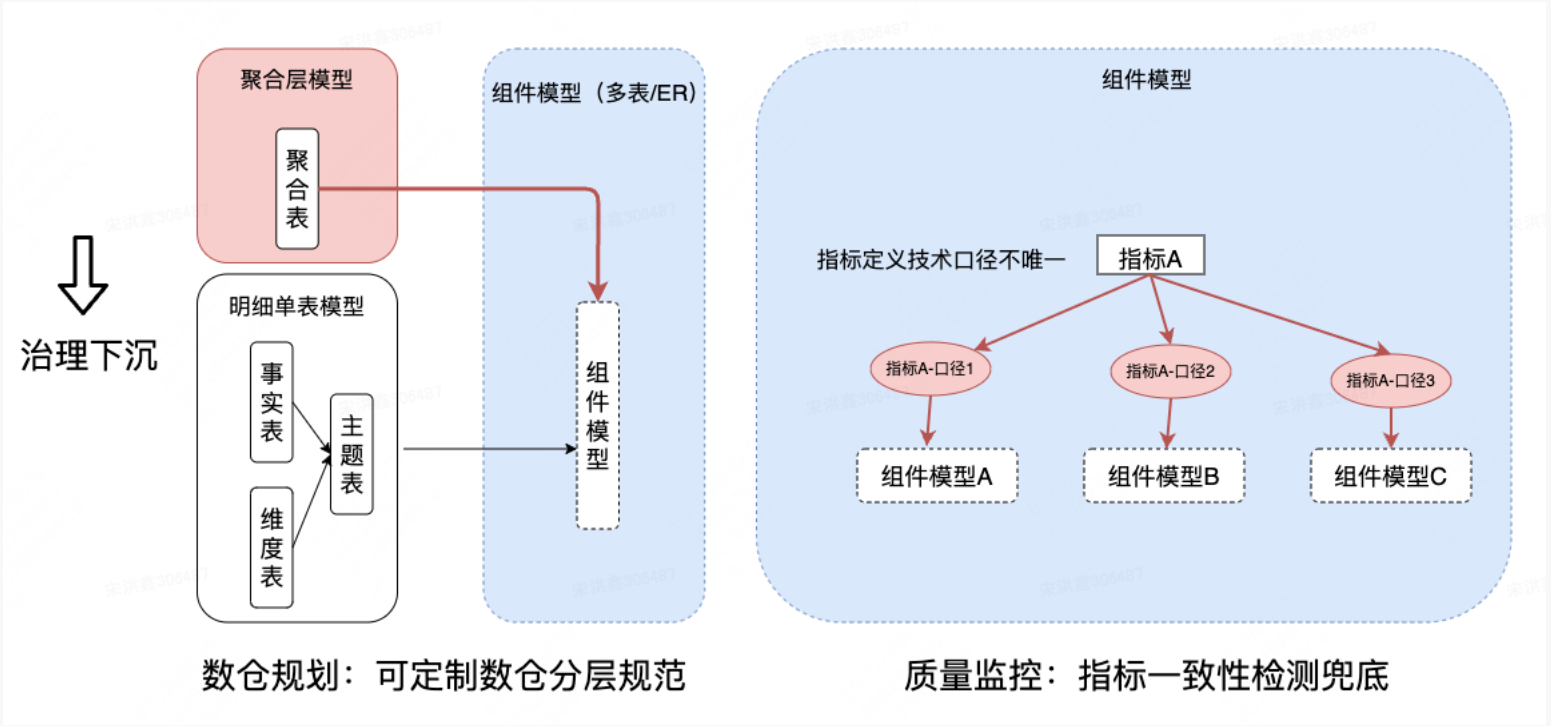
- 自动裁剪
- 引擎参数
- 运维参数
- 谓词下推



兼顾业务高优需求

双向门的设计：

- **数仓分层规范**：可以允许从聚合层进行组件建模定义
- **指标技术口径定义规范**：可以允许复合指标直接绑定组件模型中的物理字段
- **指标口径管理规范**：可以允许指标临时有多个计算口径。



尊重业务现状、结合工具能力迭代速度：治理策略上先松后严

目录

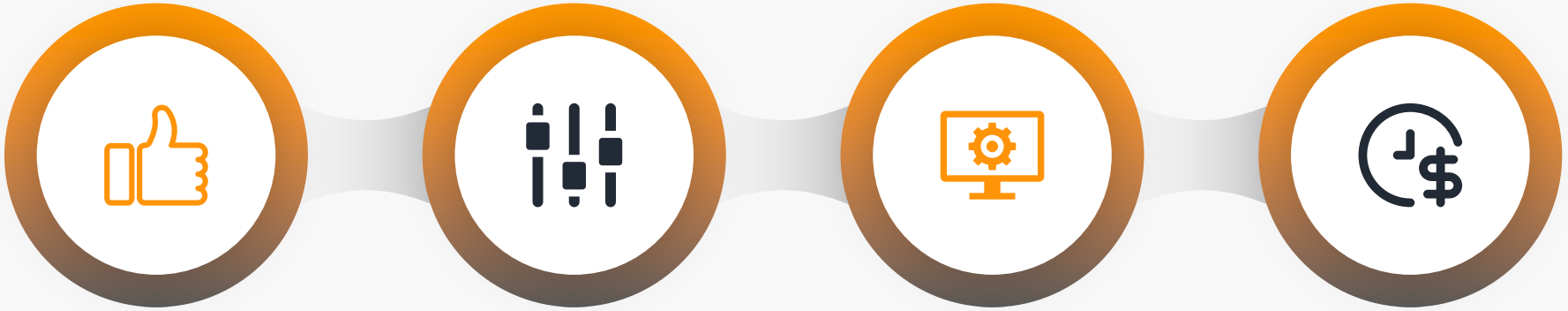
落地效果

开发效率提升60%

应用层ETL开发时长，由人均3PD，降到1PD左右

原子指标复用率5.8

较之前提升了4倍左右



指标不一致率<1%

无加工链路侧导致的一致性case

数仓链路覆盖50%+

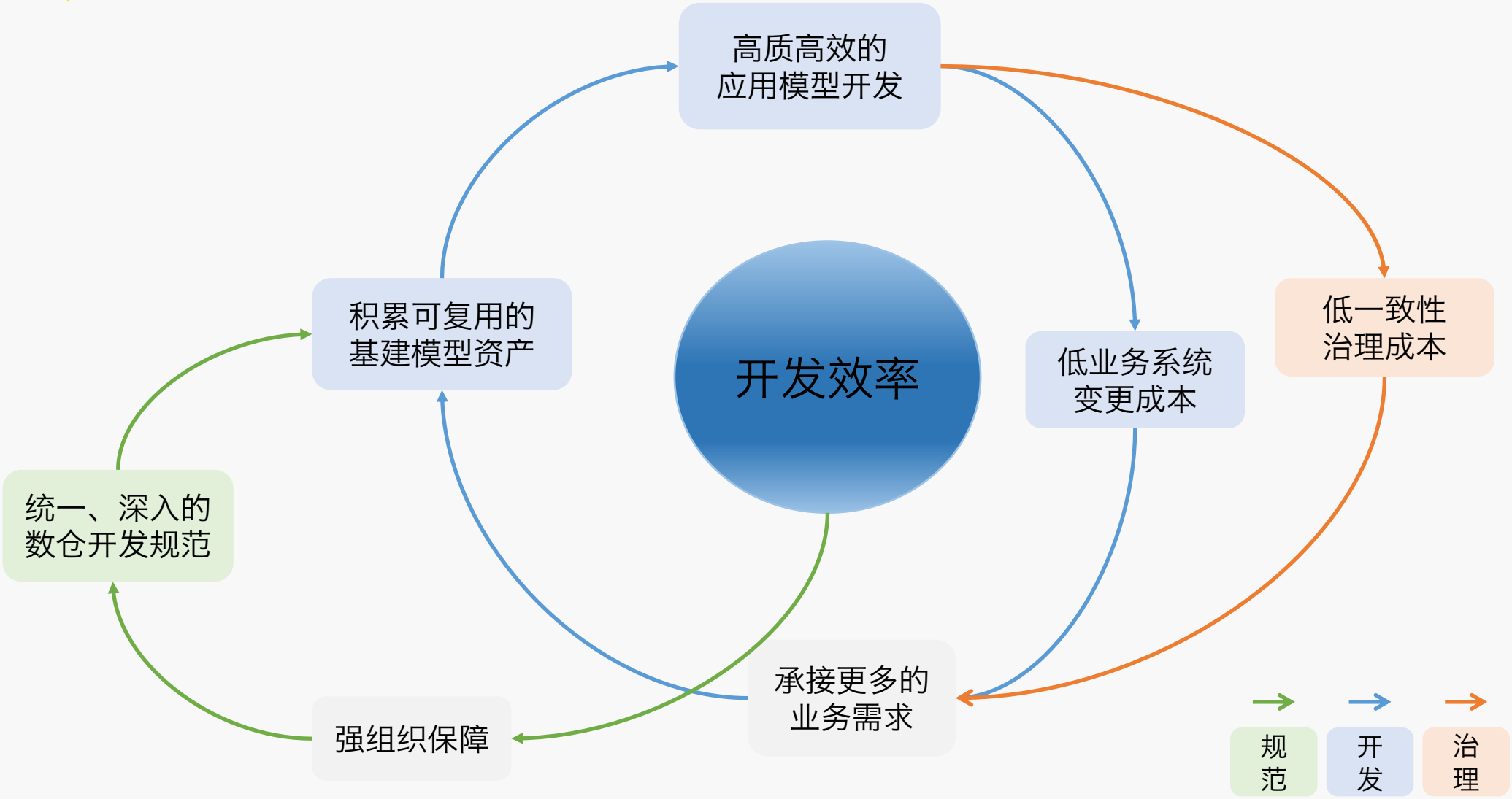
智能构建模型1000+以上，新增占比60%+

目录

经验总结

- **组织保障**:
  - 数仓架构师共识建模方法认知
  - 推动使用新生产模式的决心
  - 产研支持不同优先级需求的协同
- **工具能力**:
  - 构建能力支撑场景覆盖度
  - 关注工具使用效率
  - 双向门的设计
- **流程协作**:
  - 让用户（数据RD）形成黏性
  - 深度BP机制，快速突破卡点

治理飞轮

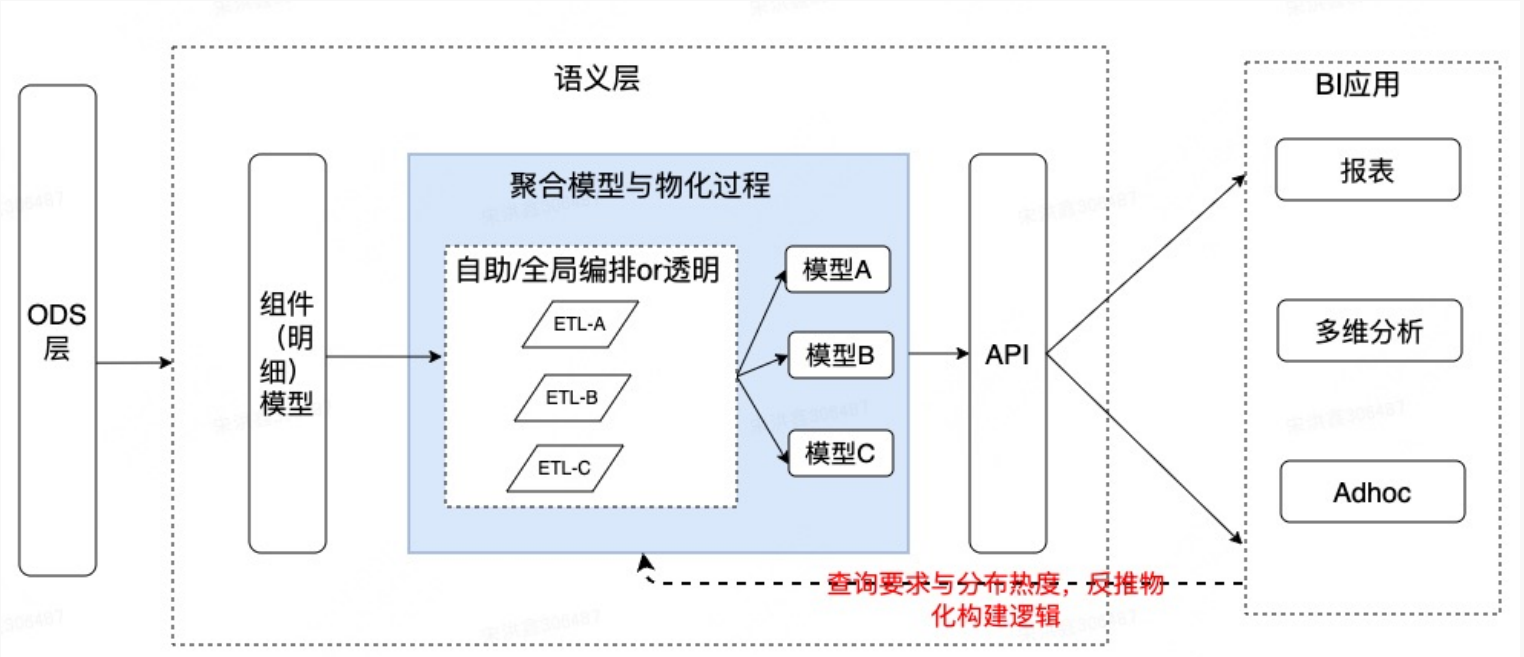

高质高效的
应用模型开发
低业务系统
变更成本
低一致性
治理成本
承接更多的
业务需求
强组织保障
统一、深入的
数仓开发规范
积累可复用的
基建模型资产
开发效率
规范
开发
治理

下一步

- **智能构建能力的持续迭代：**
  - 更多复杂指标、维度的构建能力支持：条件判断指标、衍生维度等
  - 支持基于指标维度结构特征、标签生产能力
  - 一套配置化语义，在多引擎执行的能力（spark, doris, flink）
- **关于智能构建（物化）分工的探索：**
  - 自助物化 VS 全局物化 VS 透明物化

满足性能要求下： 优化资源效率，开发效率

挑战：变更、运维



谢谢聆听！

TOP100

主办方 msup

微信官方公众号：壹佰案例
关注查看更多年度实践案例

更多技术干货，欢迎关注“美团技术团队”