网易大数据
Netease Big Data

Mammut

网易数据中台建设实践

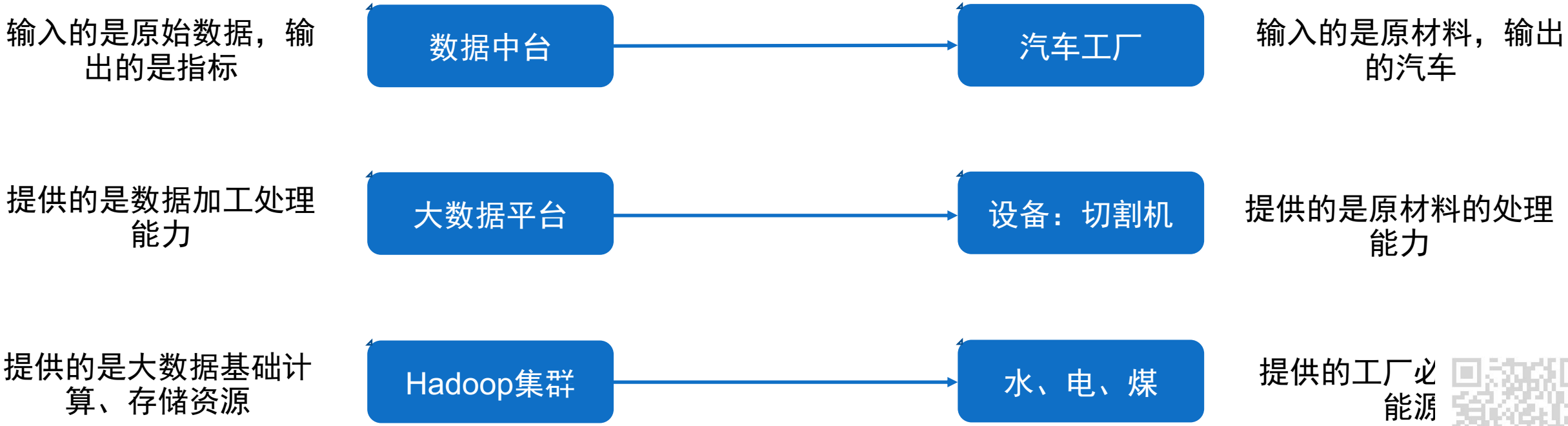
网易大数据平台负责人 郭忆

Agenda

1. 什么是数据中台？
2. 元数据中心：数据中台的基石
3. 数据治理：效率、质量、成本
4. 数据服务：数据中台的门户
5. 数据中台治理效果

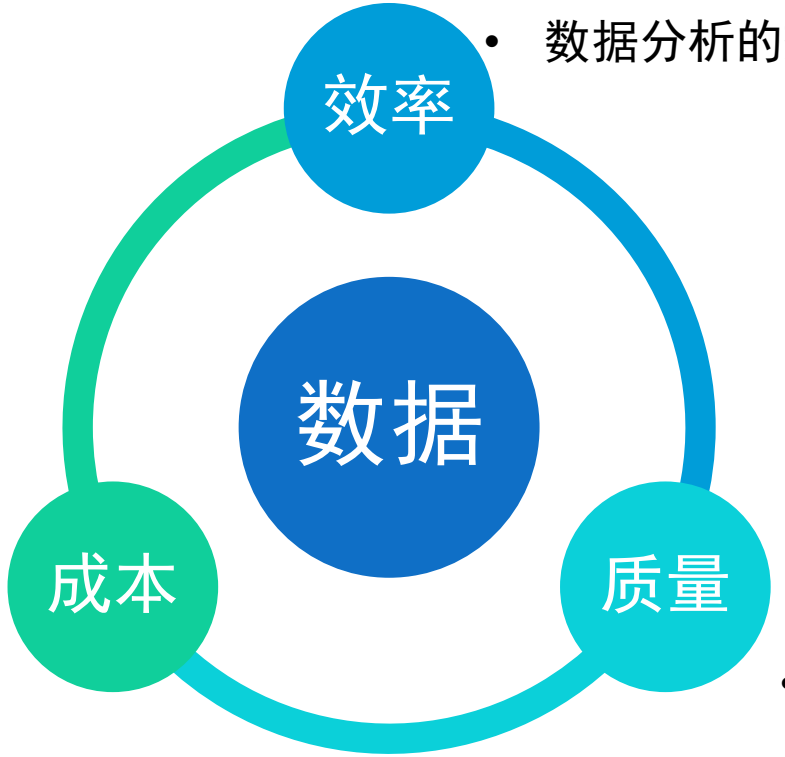
什么是数据中台

- 如果我们把数据中台比作一个汽车工厂



数据中台需要解决什么问题？

- 数据研发的效率
- 数据发现的效率
- 数据分析的效率



- 计算、存储资源成本
- 研发人力成本

- 指标一致性
- 数仓设计质量
- 数据质量

网易数据产品体系：以电商为例

业务场景	用户运营	商品运营	市场运营	供应链	管理层
数据产品	用户行为分析系统	商品运营系统	推广渠道管理系统	供应链决策协同系统	高层看板
	用户精准投放系统	商品舆情系统	活动实时直播		Vipapp

网易在做数据中台前面临的挑战

- 指标口径不一致
- 数据重复建设
- 取数效率低
- 数据无法按时正确产出
- 成本指数增长

数据中台支撑产品：网易猛犸

网易有数　网易大屏　自助分析　行业数据产品　行业业务系统

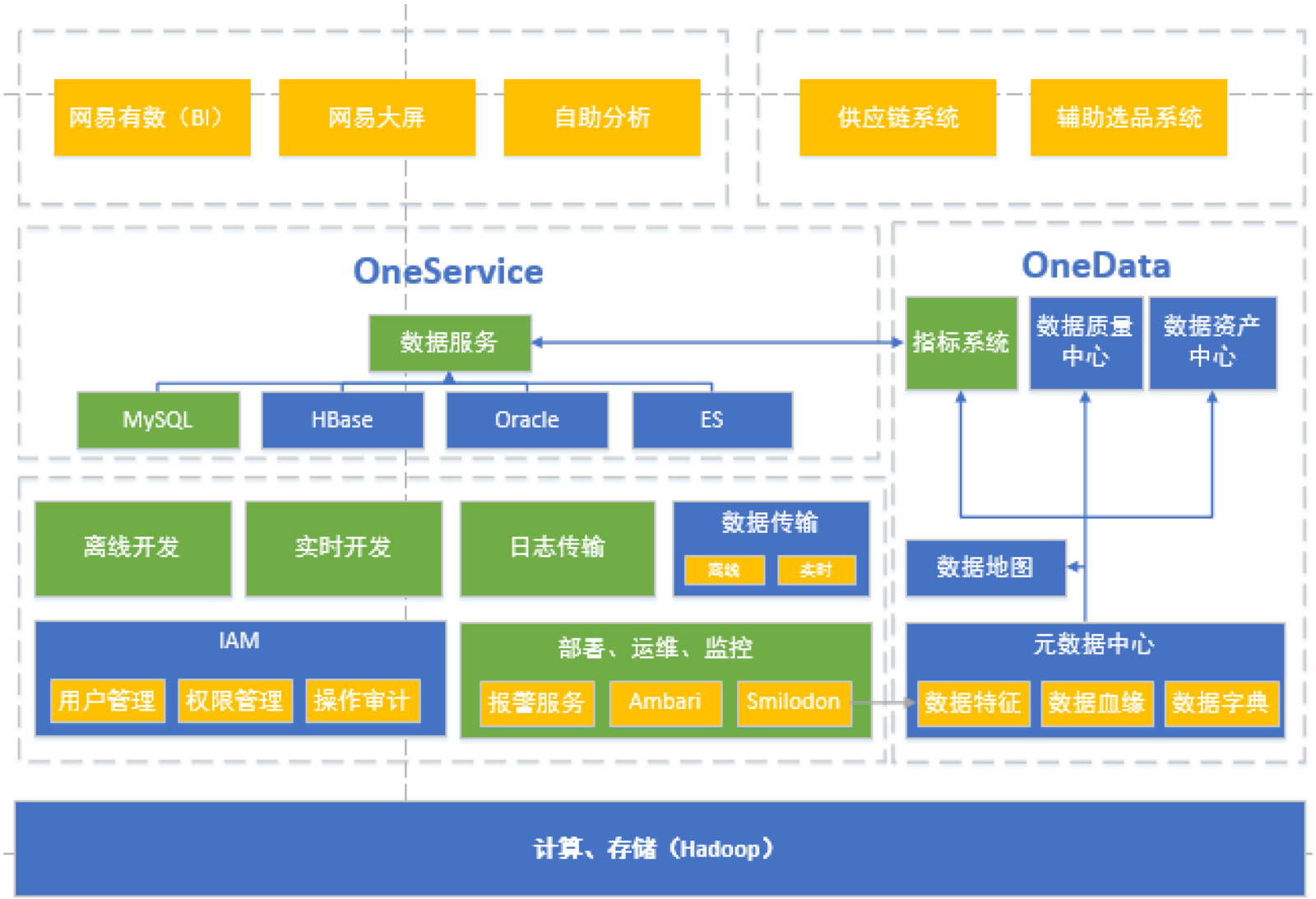
数据服务　数据服务

数据集成	数仓设计	数据开发	数据治理	运维安全
数据传输	指标系统	离线开发	成本治理	任务运维
日志采集	数仓设计	实时开发	质量治理	智能报警
数据填报		数据测试	数据地图	权限中心
埋点管理				

产品特色

- “组件式”产品架构，业务可以根据发展阶段选择性搭配
- “开放式”产品架构，聚焦核心通用产品，同时开放基础能力，允许业务集成新的产品
- “轻型易用”平台，通过“增强分析”降低用户使用的门槛
- 完美的支撑数据中台建设，减少重复建设，提高数据共享能力

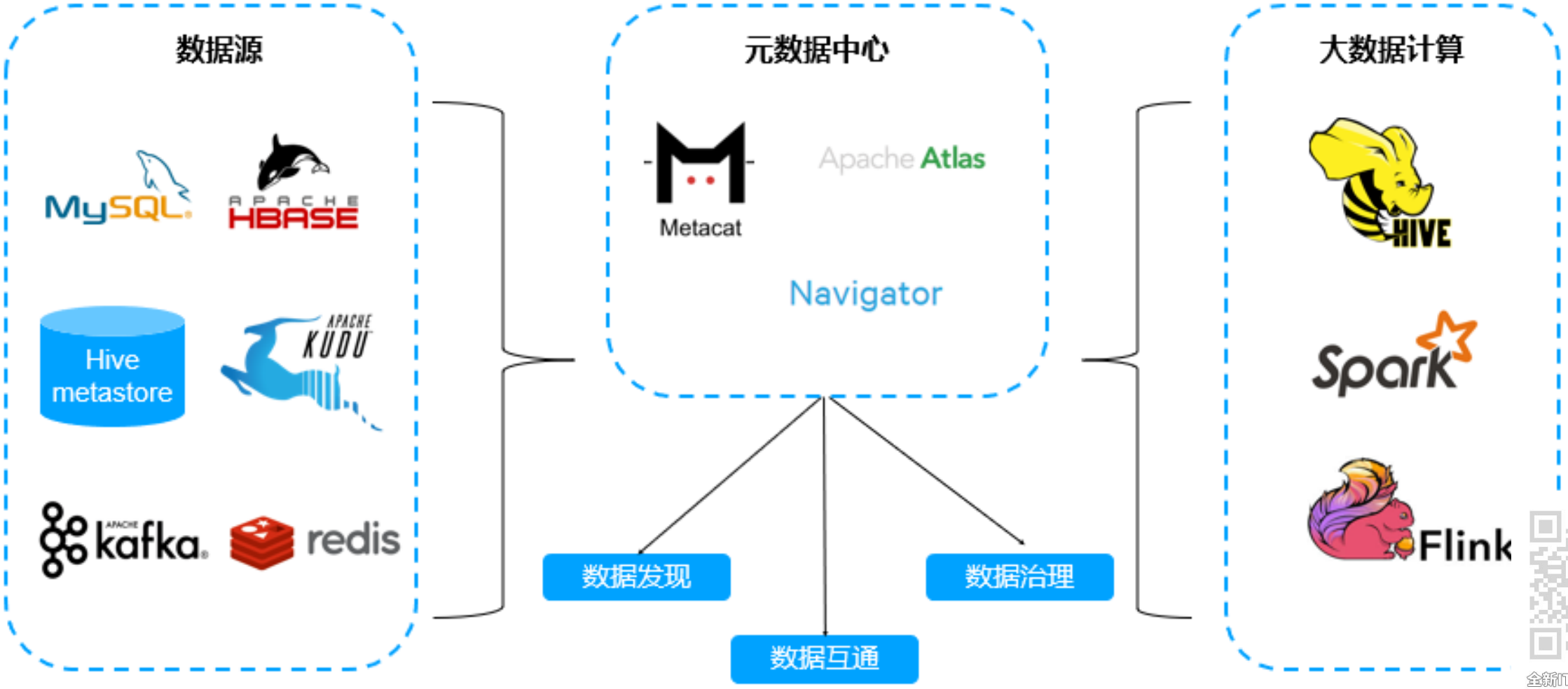
网易数据中台产品架构



Agenda

1. 什么是数据中台？
2. 元数据中心：数据中台的基石
3. 数据治理：效率、质量、成本
4. 数据服务：数据中台的门户
5. 数据中台治理效果

元数据中心：数据中台基石



元数据中心

数据标签

- 通过丰富的不同类型的标签，完善数据特征体系
- 指标标识、数仓的主题域、分层信息，是否是数仓维护的推荐表都以标签形式存在

多租户，多业务线

能够支持电商（考拉、严选）、互娱（音乐、游戏）、传媒、教育

多种数据源支持

能够覆盖网易所有的数据源，甚至包括Kafka，Redis，Hbase等Schema less KV系统

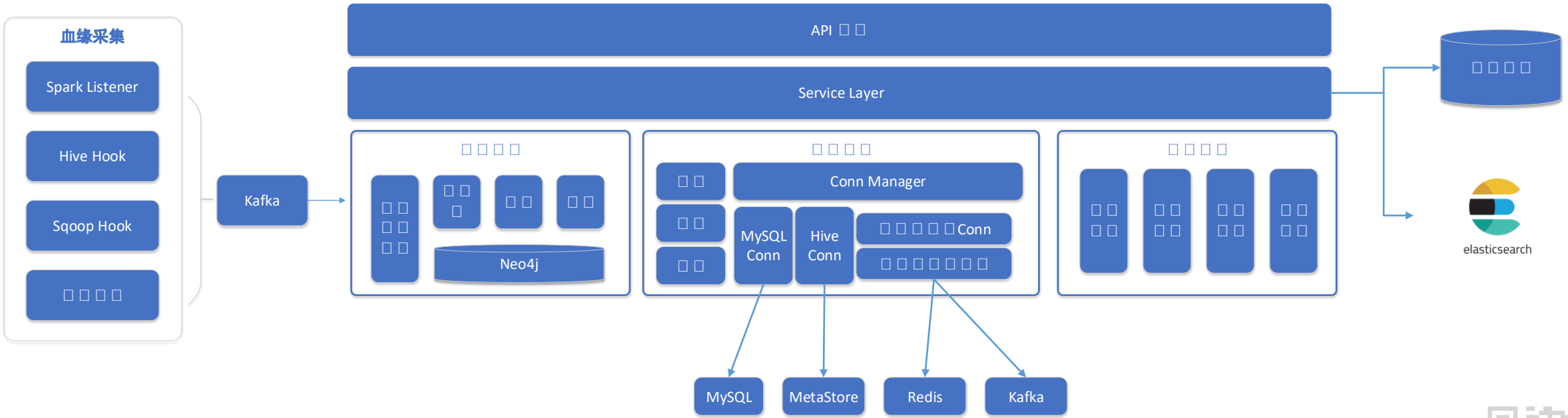
与大数据系统集成

- 与Ranger结合，允许通过自定义标签的方式对数据进行动态授权
- 数据传输、自助分析与元数据中心集成
- 基于元数据中心，构建数据质量中心、数据资产管理中心、数据地图

数据血缘

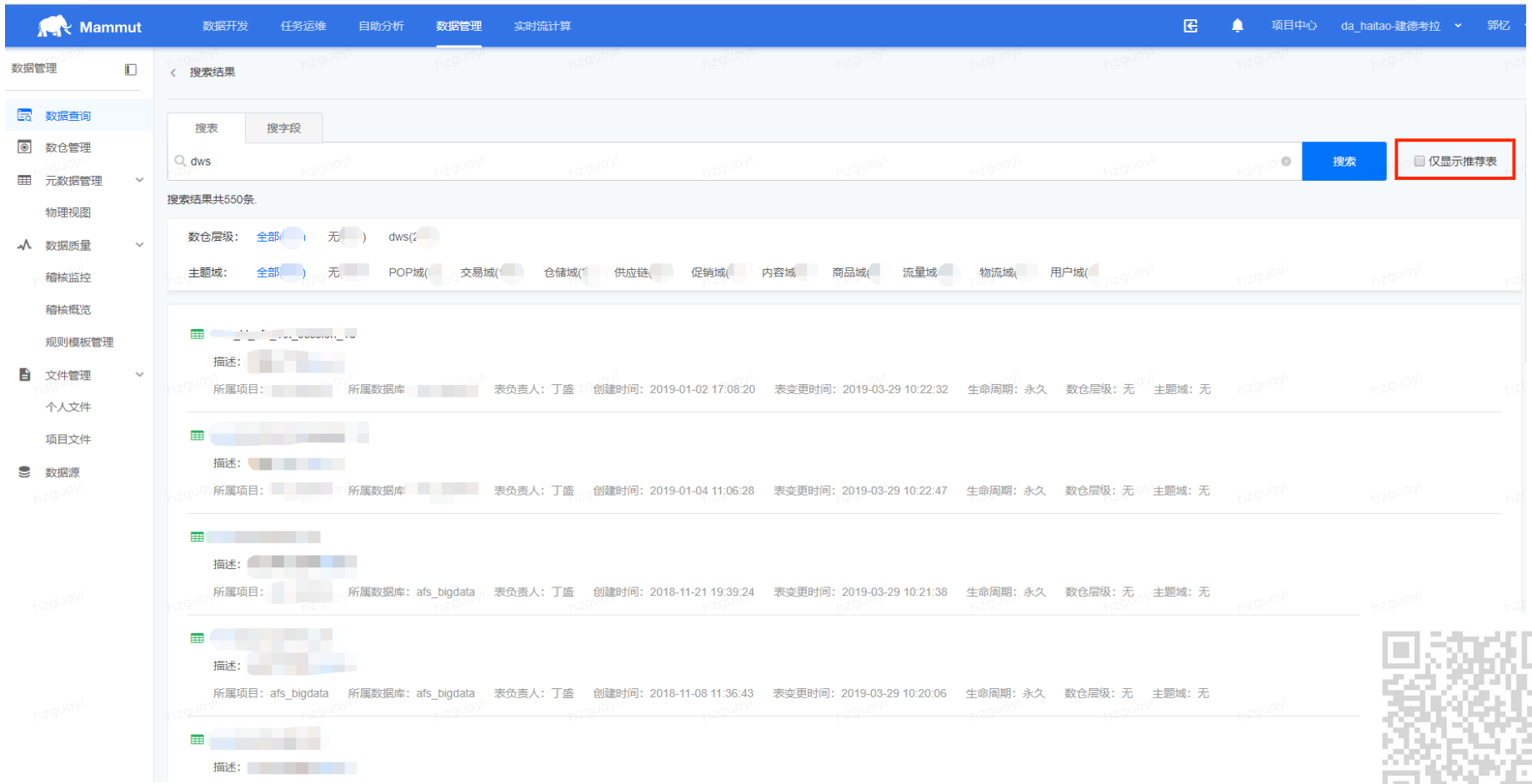
- 静态血缘 & 动态血缘
- 血缘支持时间戳，可以按照时间戳读取，过期
- 血缘覆盖率以及血缘采集性能

元数据中心



数据地图

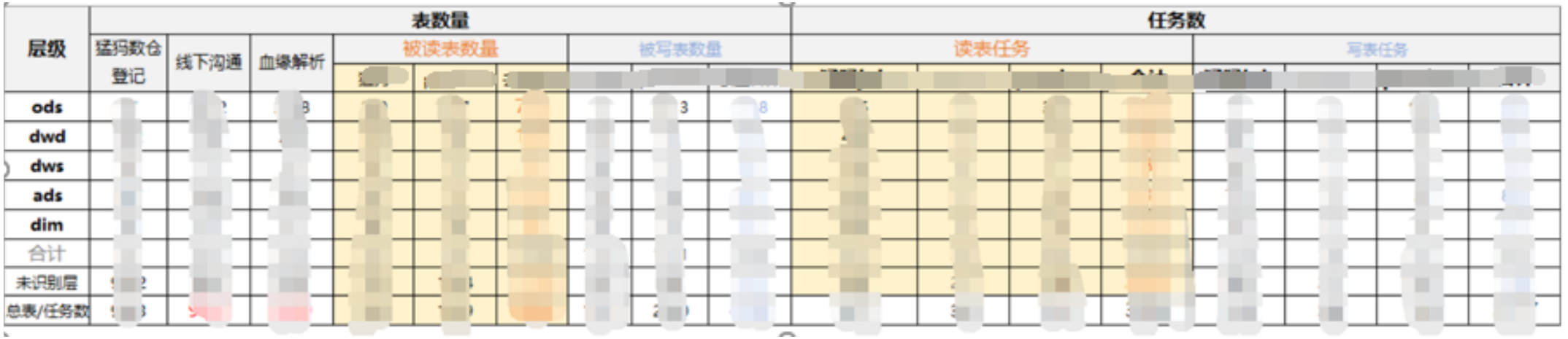
- 解决“有哪些数据可用？”，“到哪里找数据？”



Agenda

1. 什么是数据中台？
2. 元数据中心：数据中台的基石
3. 数据治理：效率、质量、成本
4. 数据服务：数据中台的门户
5. 数据中台治理效果

如何评价一个数仓设计好坏？



- 大量的表没有明确的主题域、业务过程，分层信息，数仓组织混乱
- 超过50%的任务直接引用ODS 层原始数据，30%的表存在跨层引用，DWD建设完善度较低
- DWS 层表复用性差，平均表引用系数低
- 依然有查询ODS 层原始数据的Query， DWS，ADS Query 覆盖率低，取数效率差
- 表、字段命名规范混乱，数据发现困难

规范化数仓设计

EasyDesign

数仓设计度量

规范化管理

团队协作

- 各层表的分布以及各层被下游表和任务引用情况，Query 查询覆盖率
- DWD： ODS 被跨层引用的表的数量
  DWD 平均被下游表引用系数
  ODS 被Query 查询情况
- DWS： DWS 平均被下游表引用系数
  DWS Query 覆盖率
- 度量管理
- 维度管理
- 基础字典管理
- 模型设计
- 审批流程

数仓升级的目标

覆盖度

- 消灭ADS/DWS 直接引用ODS 层原始数据
- 消灭Query直接查询ODS 层原始数据
- DWS/ADS Query 覆盖度上升

复用性

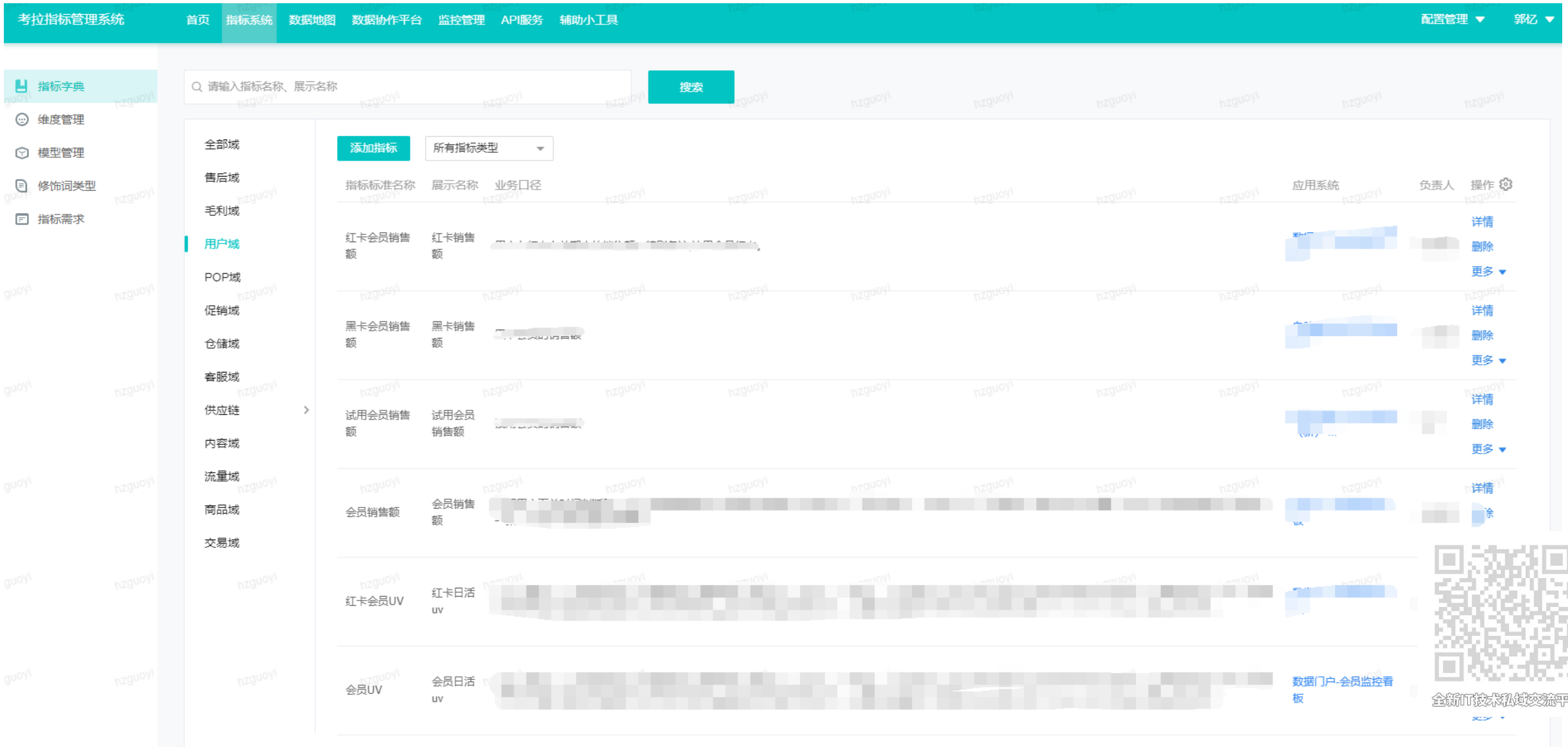
- DWS、DWD 平均每张表被下游表引用数量增加

规范性

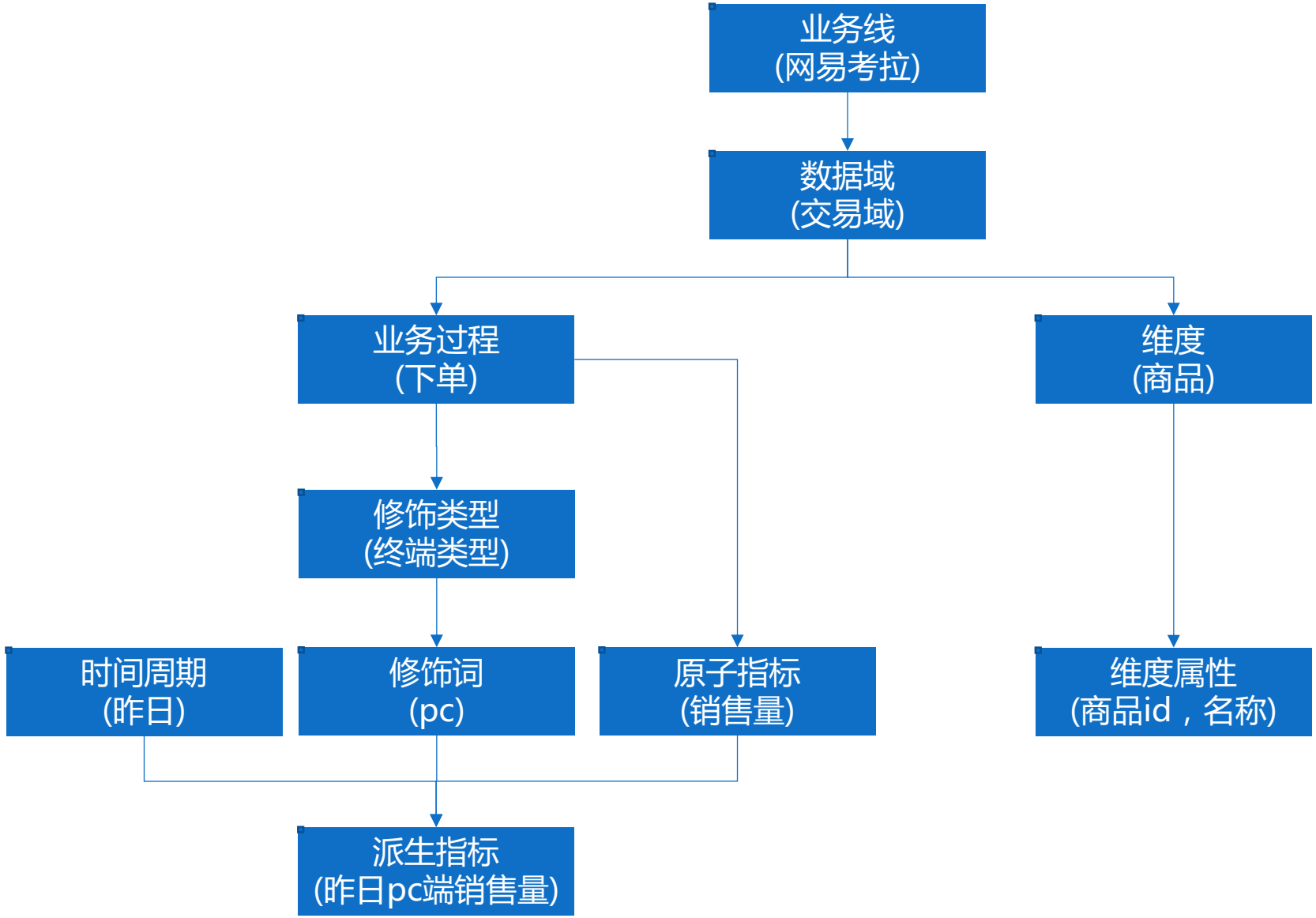
- 表、字段命名规范统一
- 建表流程审核

需求响应速度提升、查询速度提升、查询成本降低，数据使用者满意度提升！

指标系统

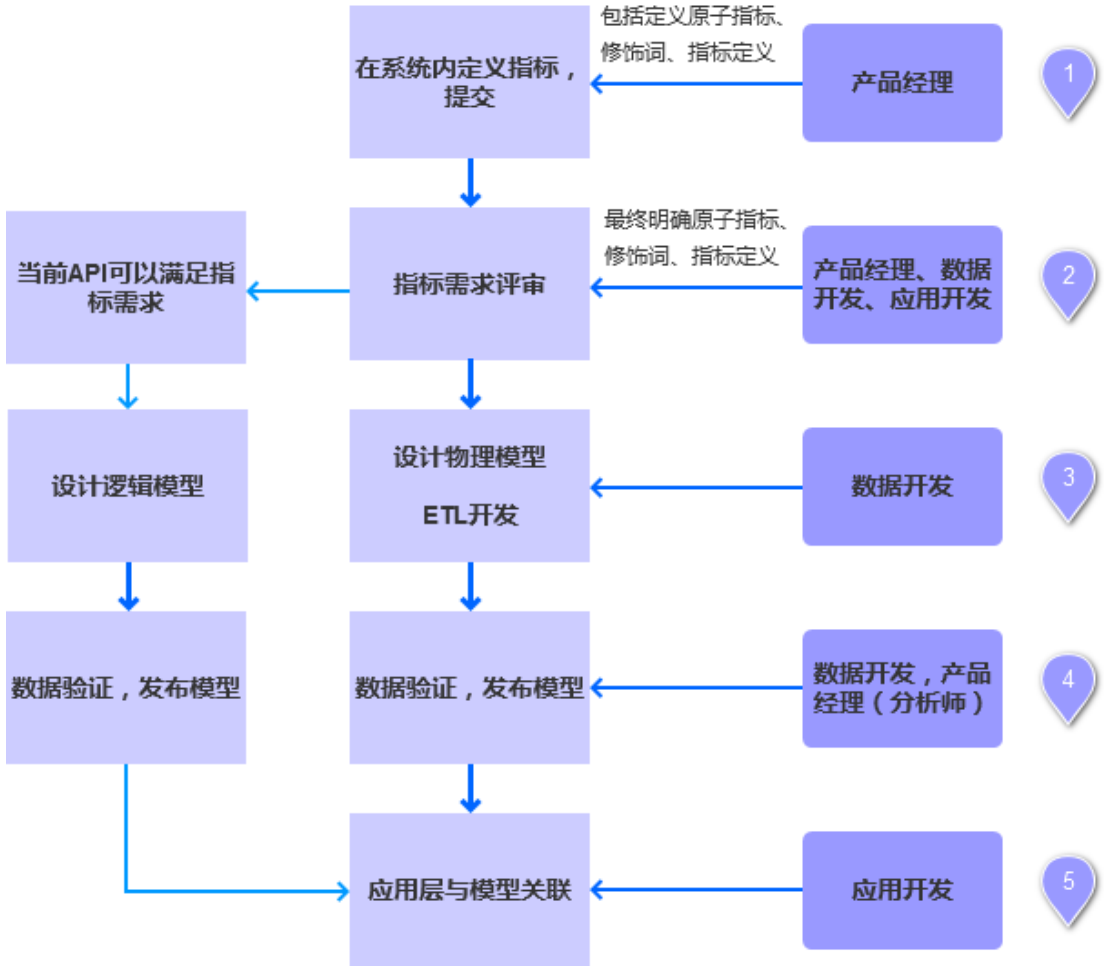


指标系统规范定义



指标开发实施规范

指标开发工作流程

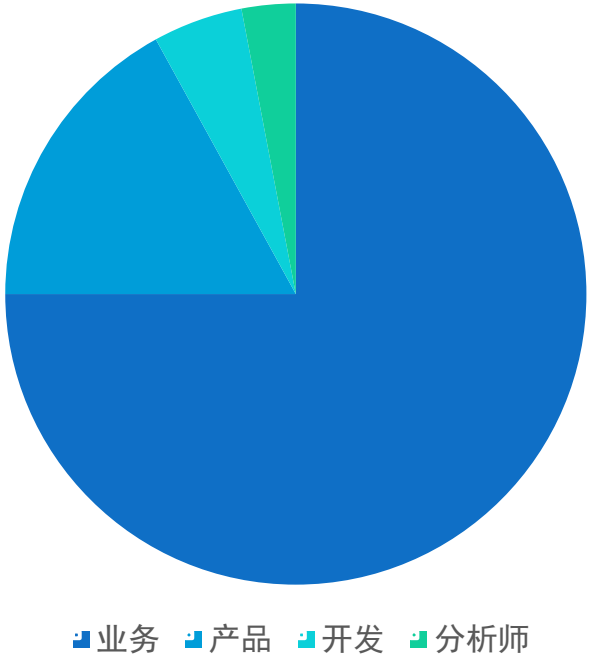


1. 创建任务，提交指标需求。
   - 现有指标：直接选取指标库中的指标，并选择维度和计算周期
   - 新指标：自己新建指标，定义指标，选择指标所需的维度及计算周期。
   - 新维度：新建维度
   - 新计算周期：新建计算周期
2. 评审指标需求，明确需要开发的指标任务。
   - 针对新指标：在需求评审明确后可发布到指标库中。数据产品定义该指标的原子指标、业务限定。
   - 针对新维度：在需求评审明确后可发布到指标库中。数据产品定义该新维度并维护
   - 针对新计算周期：在需求评审明确后可发布到指标库中。数据产品定义该新计算周期并维护
3. 开发需求流转到对应开发，开发设计应用模型并开发数据
4. 对模型数据进行审核验证并发布
   - 对照指标定义，数据架构师对指标技术口径进行验证。
   - 产品经理或者分析师针对数据对指标进行验证。
   - 模型发布有审核流程。
5. 应用开发发送请求通过统一API调用对应模型内的指标数据及指标信息，如指标定义等。

在做数据质量前业务面临的现状

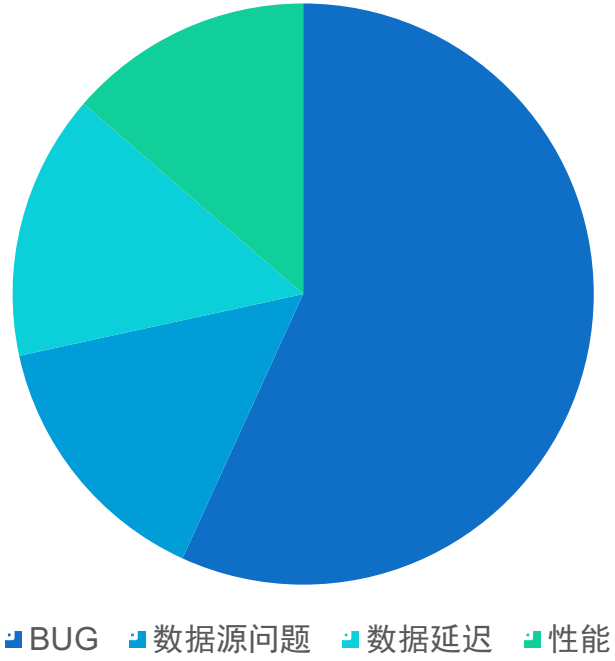
- 超过90%的问题是由业务和产品发现

报障人群统计



- 收集的问题中存在研发bug的占比超过50%

故障问题分类



数据质量方法论

完整性

是指需要的数据已完整记录，可以分为记录数完整性和字段值完整性。

时效性

是指数据的时效对应用的满足程度，可以分为数据任务结束时间和运行时长。

准确性

是指数据在数仓中的值和实际值是否相同，可以分为口径实现结果和数据逻辑合理性。

一致性

是指系统内外部数据源之间的数据一致程度，可以分为指标一致性和信息一致性。

数据稽核监控规则

完整性

- 表数据量波动监控和绝对值监控
- 主键唯一监控
- 字段为空，为0的监控
- 数据完整性监控，订单24小时，终端覆盖

准确性

- 数值计算逻辑监控：商品只能归属一个BU
- 数据格式的监控，比如IP，URL
- 码表的监控
- 数据异常监控，比如日期还没发生

一致性

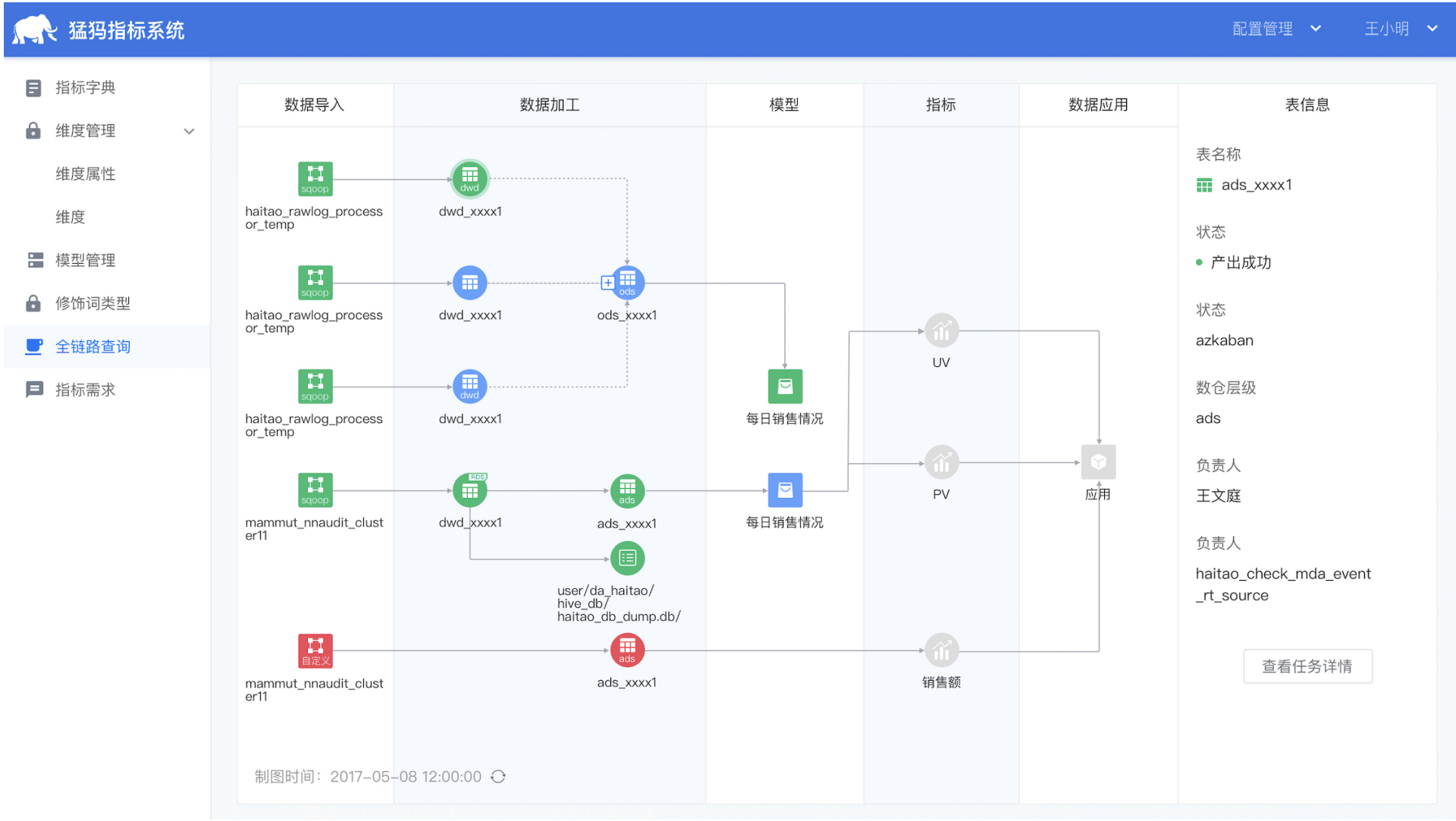
- 同一个指标在不同模型不一致监控
- 相关指标趋势监控，比如uv和pv走势一致
- 聚合逻辑一致性监控
- 不同数据源对同一个实体的值一致

时效性

- 任务延迟监控
- 表产出时间的监控
- 源数据延迟监控

全链路数据质量监控

- 覆盖数据产出的完整生命周期
- 全链路数据血缘的实时监控
- 快速了解哪些数据产品的哪些指标异常
- 故障恢复时间的请准预估



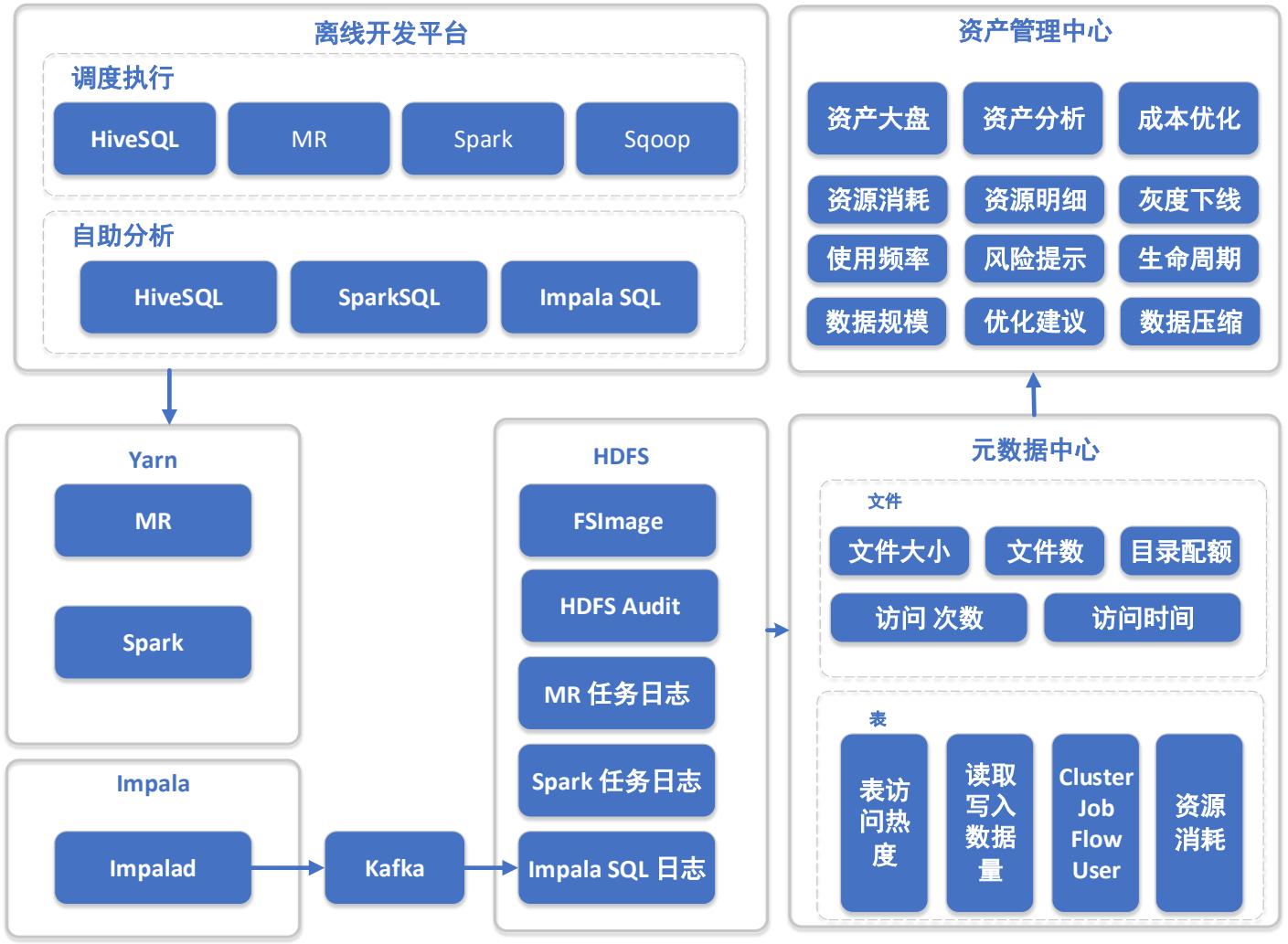
成本治理

- 每个项目花了多少钱？
- 每张报表花了多少钱？
- 每个数据产品花了多少钱？
- 每个人花了多少钱？
- 各个业务线预算符合度？
- 根据表的热度，存储空间、加工表消耗的资源，确认表是否可以优化？给出优化建议？
- 表的一键下线

成本治理


成本治理方法论
队列
空闲队列
指标： 队列的资源利用率
队列提交权限
清理队列权限
任务
任务优化
指标： 任务消耗的计算资源
指标： 任务的耗时（实际执行时间，排除任务依赖等待时间）
SQL 优化
任务责任人落实
任务提交人早已变更
无用任务的清理
非本项目的任务迁移其他的队列
消峰填谷
调整任务的调度时间到空闲时间
表
无用表的清理
指标: 表的访问频率
指标：最近访问时间
表生命周期管理
外部表生命周期文件清理
文件
无用文件的清理
指标： 文件的访问频率
指标: 文件的最近访问时间
应用
应用的上游任务的资源消耗

系统架构


离线开发平台
调度执行
HiveSQL
MR
Spark
Sqoop
自助分析
HiveSQL
SparkSQL
Impala SQL
资产管理中心
资产大盘
资产分析
成本优化
资源消耗
资源明细
灰度下线
使用频率
风险提示
生命周期
数据规模
优化建议
数据压缩
Yarn
MR
Spark
Impala
Impalad
Kafka
HDFS
FSImage
HDFS Audit
MR 任务日志
Spark 任务日志
Impala SQL 日志
元数据中心
文件
文件大小
文件数
目录配额
访问 次数
访问时间
表
表访问热度
读取写入数据量
Cluster Job Flow User
资源消耗

Agenda

1. 什么是数据中台？
2. 元数据中心：数据中台的基石
3. 数据治理：效率、质量、成本
4. 数据服务：数据中台的门户
5. 数据中台治理效果

数据服务

指标口径

- 相同指标在多个表中存在
- 指标口径不统一
- 指标复用率低，指标重复加工

安全隐患

- 底层表直接暴漏
- 所有对数仓的访问权限控制

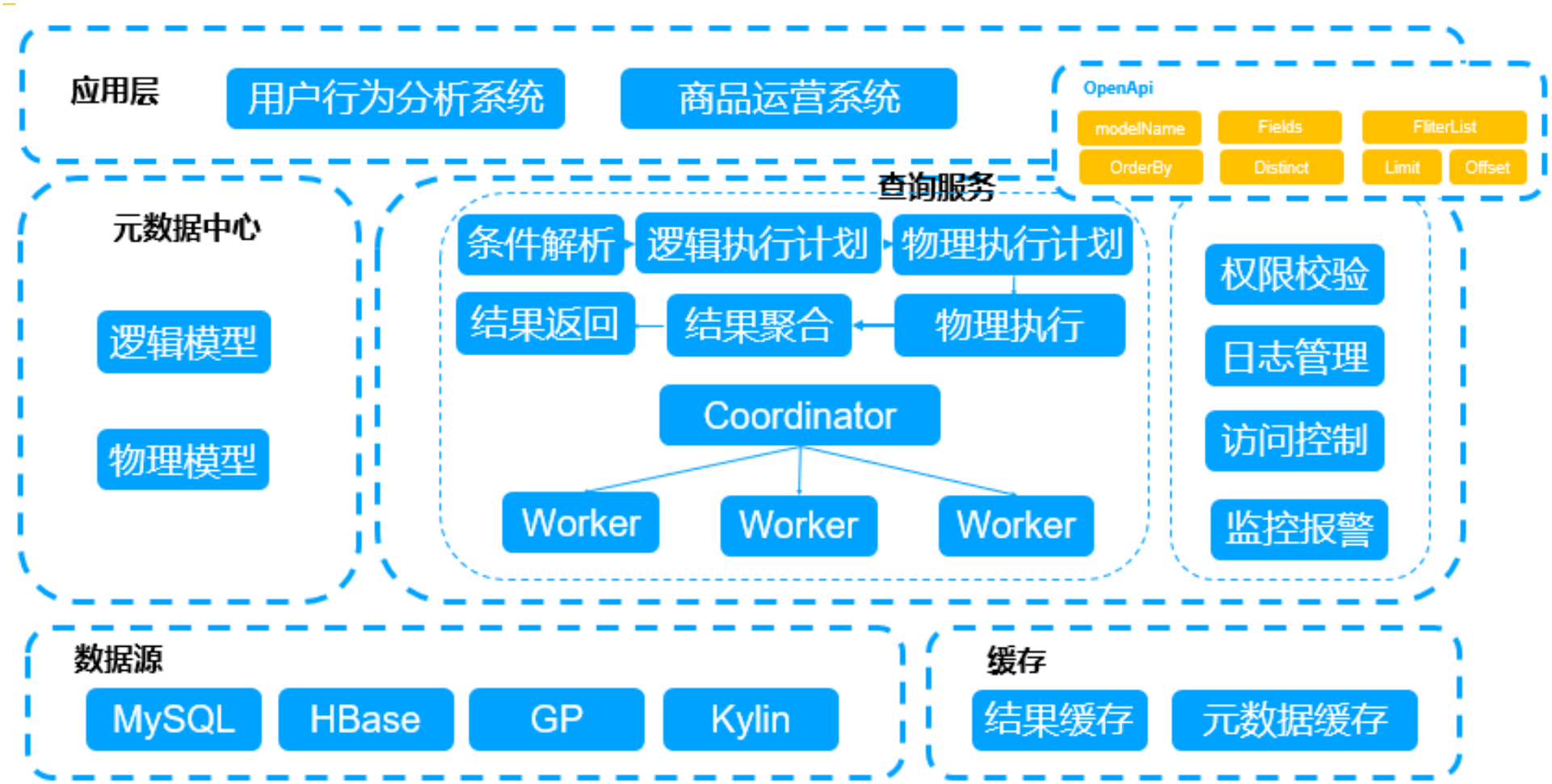
效率

- 所有需求一个接口
- SDK 接入效率高

维护性

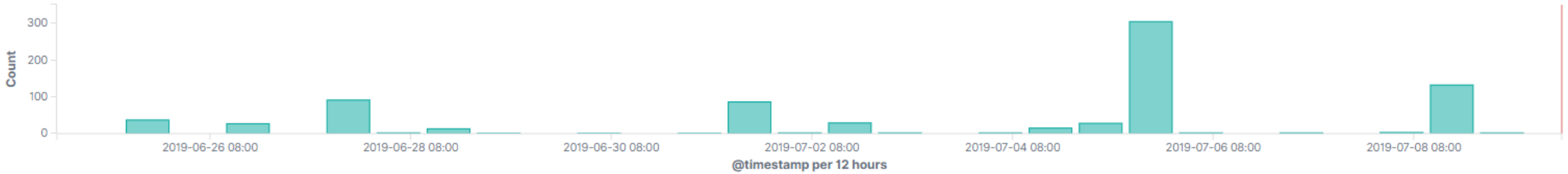
- 打通了数据应用-指标-数仓表的全链路监控
- 底层表变更，需要应用层修改

数据服务


应用层
用户行为分析系统
商品运营系统
OpenApi
modelName
Fields
FliterList
OrderBy
Distinct
Limit
Offset
查询服务
元数据中心
逻辑模型
物理模型
条件解析
逻辑执行计划
物理执行计划
结果返回
结果聚合
物理执行
Coordinator
Worker
Worker
Worker
权限校验
日志管理
访问控制
监控报警
数据源
MySQL
HBase
GP
Kylin
缓存
结果缓存
元数据缓存

数据服务

Jun 24, 2019 @ 19:24:17.752 - Jul 9, 2019 @ 19:24:17.752 — Auto



Time	appName	requestId	modelName	costTime	cache	statusCode
Jul 9, 2019 @ 07:00:00.259	严选应用	bce9842f-52b1-4abe-a29f-4547420b0b61	adi_kl_yanxuan_order_di	75	false	200
Jul 9, 2019 @ 07:00:00.118	严选应用	07d9970a-4766-42df-b243-c8309af83567	adi_kl_yanxuan_order_di	101	false	200
Jul 8, 2019 @ 21:46:20.568	选品系统	db12b0c1-3d0b-4bf8-9d78-a3db0a61dbe8	adi_kl_xuanpin_goods_stat_df	8	true	200
Jul 8, 2019 @ 21:46:20.548	选品系统	3e7e759f-0432-4b03-bd94-5de4559a8a42	adi_kl_xuanpin_goods_stat_df	9	true	200
Jul 8, 2019 @ 21:46:20.535	选品系统	1a5f3ce1-4d21-436e-ab56-a97ed1f109a5	adi_kl_xuanpin_goods_stat_df	4	true	200
Jul 8, 2019 @ 21:46:20.518	选品系统	ad8348dc-7d07-4b20-bb06-593b2f447fa2	adi_kl_xuanpin_cat_stat_df	7	true	200
Jul 8, 2019 @ 21:46:20.499	选品系统	c7c3abae-5819-4bc8-8c63-6786214fc49c	adi_kl_xuanpin_cat_stat_df	8	true	200
Jul 8, 2019 @ 21:46:04.852	选品系统	7b060dd1-d987-4fe3-aa7b-3c69b6499536	adi_kl_xuanpin_goods_stat_df	8	true	200
Jul 8, 2019 @ 21:46:04.828	选品系统	8080eafa-7d5f-47e5-9875-74bec7621bbd	adi_kl_xuanpin_goods_stat_df	13	true	200
Jul 8, 2019 @ 21:46:04.811	选品系统	858f003b-8c33-46fd-b0f7-a2499e88d84f	adi_kl_xuanpin_goods_stat_df	9	true	200
Jul 8, 2019 @ 21:46:04.790	选品系统	1bdc7959-a3ab-40e4-837a-d9d8b1ea7320	adi_kl_xuanpin_cat_stat_df	10	true	200
Jul 8, 2019 @ 21:46:04.774	选品系统	a22e5119-6463-43b0-b422-378eebf6251a	adi_kl_xuanpin_cat_stat_df	5	true	200

Agenda

1. 什么是数据中台？
2. 元数据中心：数据中台的基石
3. 数据治理：效率、质量、成本
4. 数据服务：数据中台的门户
5. 数据中台治理效果

数据中台的实施价值

100% 数据产品的指标通过系统化管理，有明确的业务口径、计算逻辑、数据来源，完全消除指标二义性

基于数据地图，实现100%自助取数，取数效率提升300%！分析师满意度提升！

所有数仓维护的表都有明确的分层、主题域，业务过程，ODS 层引用任务降低，100%消灭ODS 层Query，DW层引用任务增长，DWS/ADS Query 99.9%覆盖，表的引用系数提升

全链路数据跟踪，回答“数据准不准？”“哪些数据故障？”“什么时候恢复？”，加速数据故障的排查定位，助力99.8%SLA达成

通过全链路资产分析，消灭低价值的资产，为业务节省20%的资源成本

总结

数据标准规范统一

元数据互通

成本治理

服务化

数据质量治理