餐饮生态大数据架构实践

牛江浩@美团技术沙龙

自我介绍

1. 2011~2012 阿里巴巴后台研发工程师
2. 2013 加入原美团
3. 2017 餐饮生态数据应用方向技术负责人

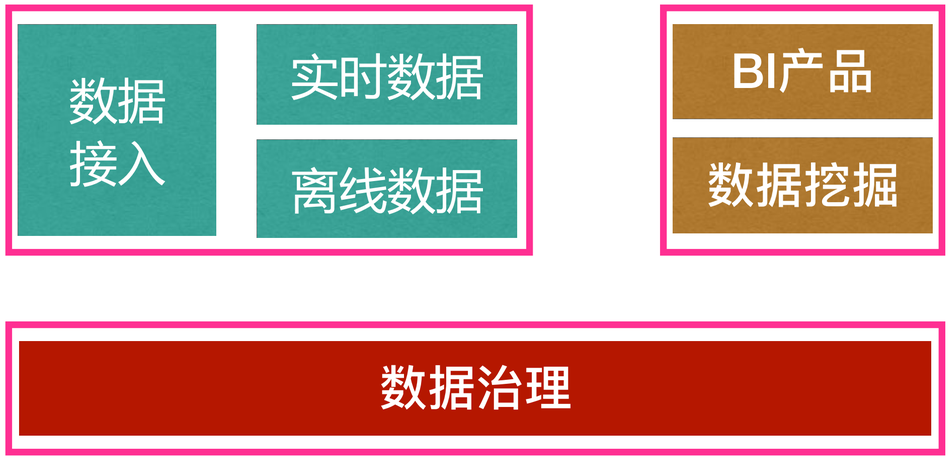
关注点：用数据技术解决业务问题、最大限度的发挥数据价值、支撑并驱动业务发展

团队组织架构

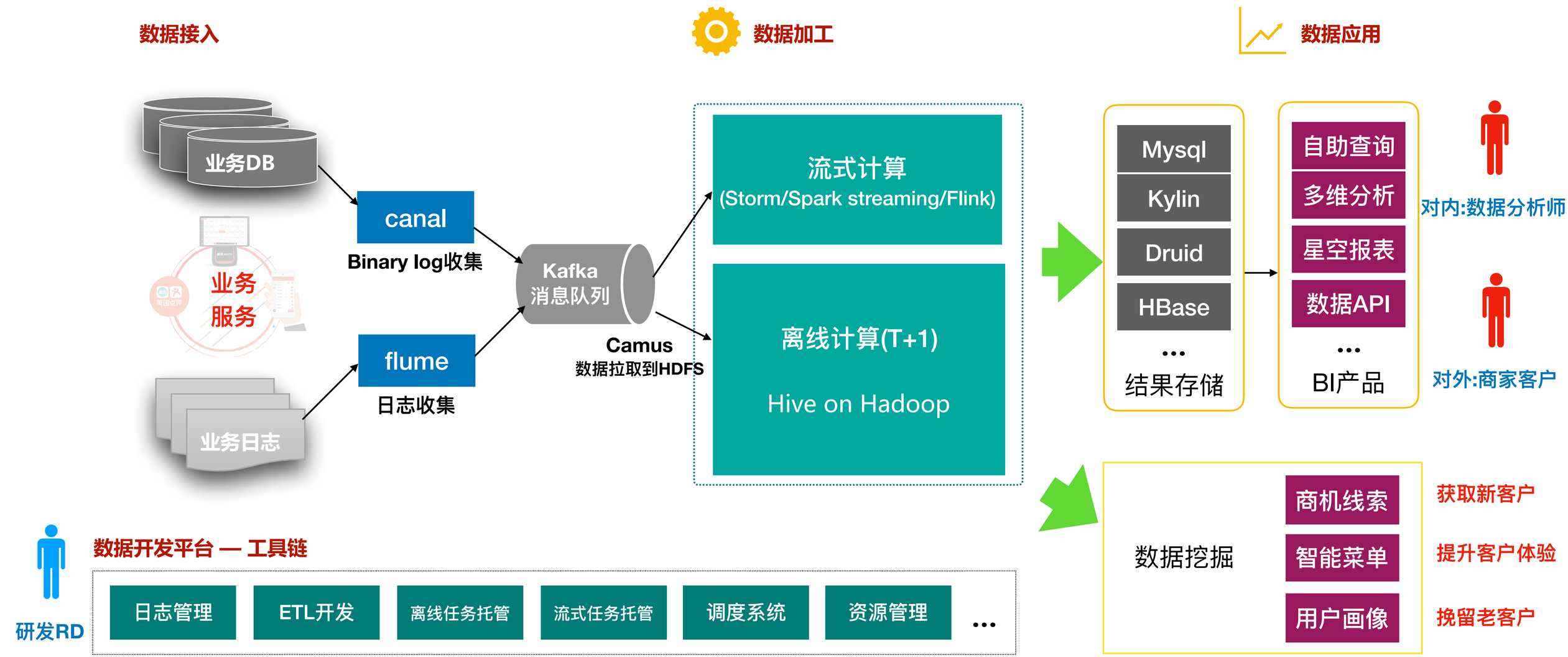

智 能
收 银
数据PM
数据RD
会员
营销
数据PM
数据RD
开放
平台
数据PM
数据RD
天网
系统
数据PM
数据RD
...
数据
接入
实时数据
离线数据
BI产品
数据挖掘
数据治理
餐饮生态数据组
公司统一基础数据研发平台

目录

- 离线数据
- 实时数据
- 数据应用
- 总结及思考



数据流架构全景图

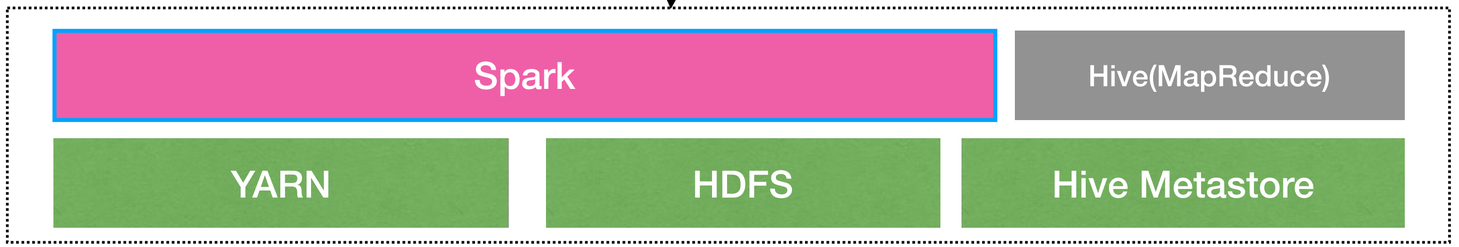

数据接入
数据加工
数据应用
业务DB
业务服务
业务日志
canal
Binary log收集
flume
日志收集
Kafka
消息队列
Camus
数据拉取到HDFS
流式计算
(Storm/Spark streaming/Flink)
离线计算(T+1)
Hive on Hadoop
Mysql
Kylin
Druid
HBase
...
结果存储
自助查询
多维分析
星空报表
数据API
...
BI产品
对内:数据分析师
对外:商家客户
数据挖掘
商机线索
智能菜单
用户画像
获取新客户
提升客户体验
挽留老客户
研发RD
数据开发平台 — 工具链
日志管理
ETL开发
离线任务托管
流式任务托管
调度系统
资源管理
...

离线计算

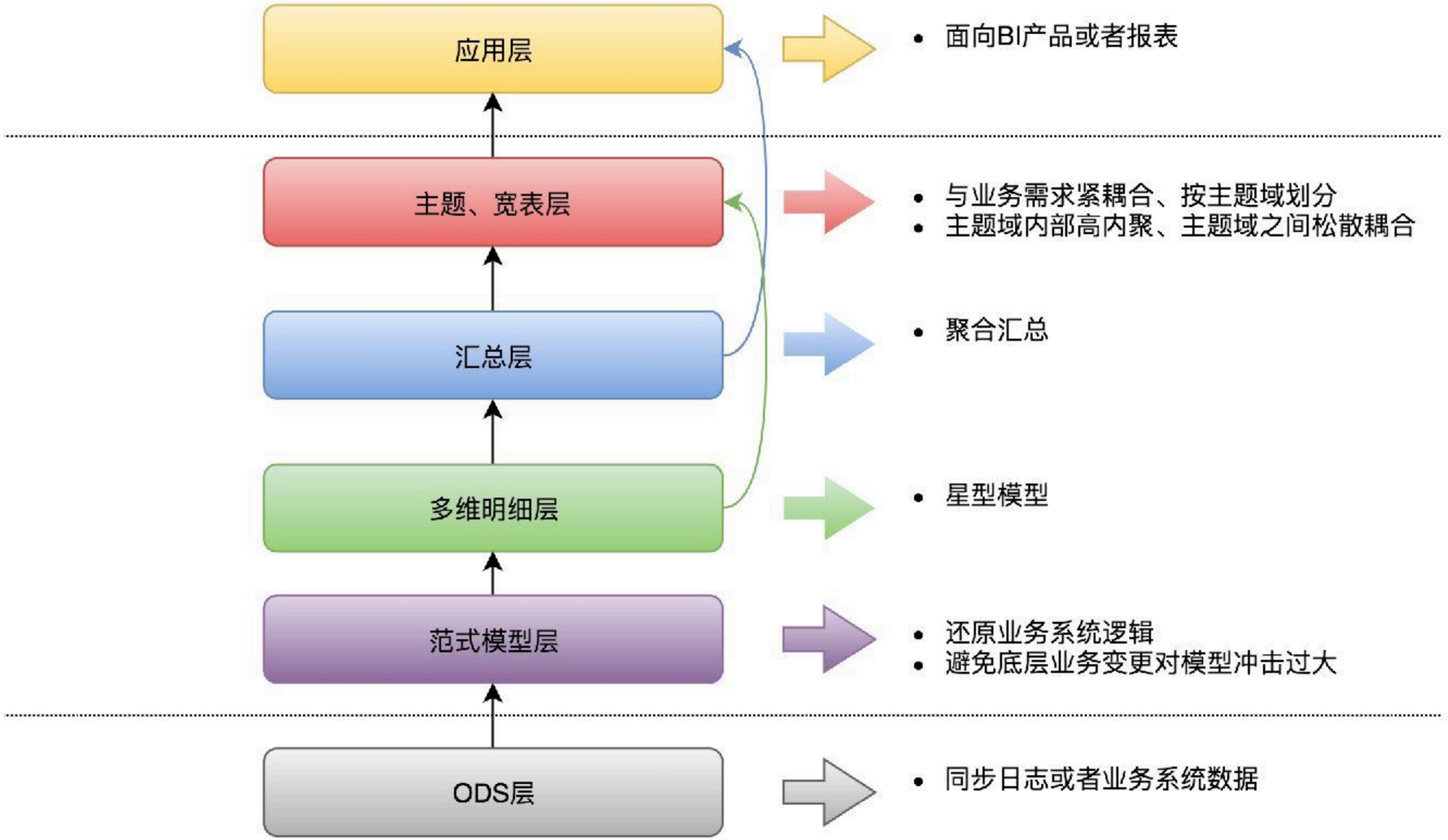
前台：开发平台

调度系统

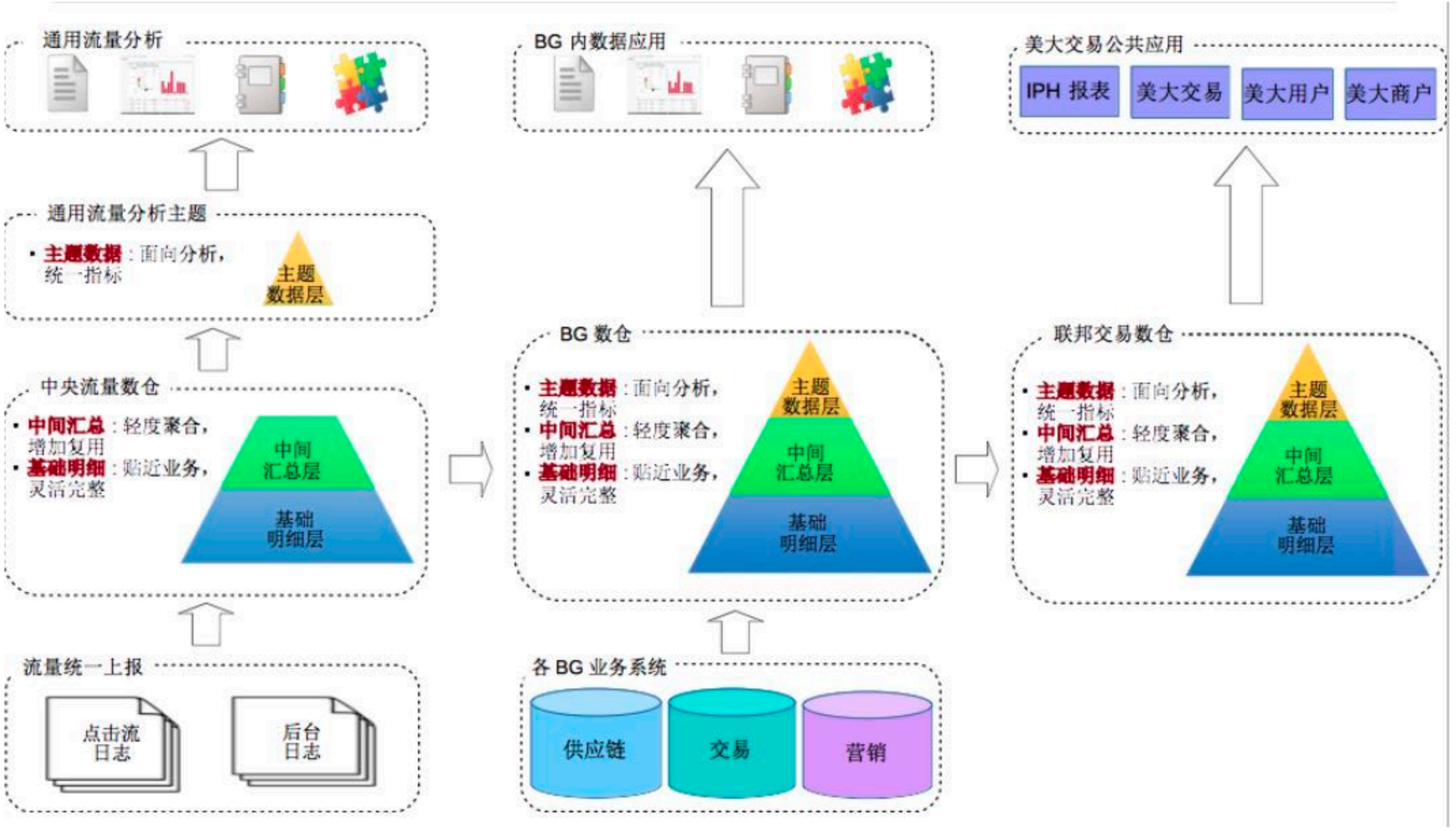
后台：计算、存储引擎



离线数据仓库建设


应用层
主题、宽表层
汇总层
多维明细层
范式模型层
ODS层
面向BI产品或者报表
与业务需求紧耦合、按主题域划分
主题域内部高内聚、主题域之间松散耦合
聚合汇总
星型模型
还原业务系统逻辑
避免底层业务变更对模型冲击过大
同步日志或者业务系统数据

全域数据联动


通用流量分析
BG 内数据应用
美大交易公共应用
IPH 报表
美大交易
美大用户
美大商户
通用流量分析主题
主题数据：面向分析，统一指标
主题数据层
BG 数仓
主题数据：面向分析，统一指标
中间汇总：轻度聚合，增加复用
基础明细：贴近业务，灵活完整
主题数据层
中间汇总层
基础明细层
联邦交易数仓
主题数据：面向分析，统一指标
中间汇总：轻度聚合，增加复用
基础明细：贴近业务，灵活完整
主题数据层
中间汇总层
基础明细层
中央流量数仓
中间汇总：轻度聚合，增加复用
基础明细：贴近业务，灵活完整
中间汇总层
基础明细层
流量统一上报
点击流日志
后台日志
各 BG 业务系统
供应链
交易
营销

数据管理

自身业务实际面临的**核心问题**：质量差、效率低、价值少

数据质量管理

生命周期管理

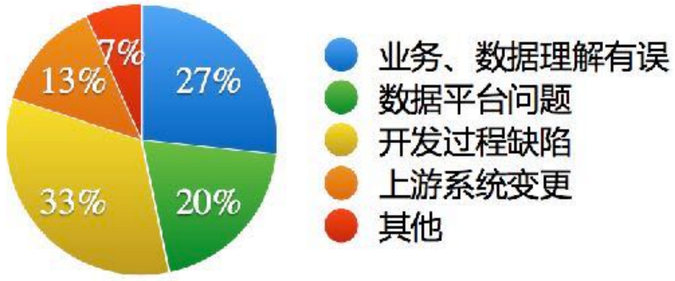
辅助开发工具

数据价值运营

数据管理

数据质量管理-事前识别、预防&规避风险

- **技能培训、加强review、**充分测试等 —>保证开发质量
- **数据设计评审** —>业务系统、数据、需求方**认知保持一致**
- **确立问责机制** —> 和上游业务系统做好约定、周知义务并**关联KPI**
- 其他诸如和数据平台做好约定、参数变更要周知……



数据管理

数据质量管理-事中监控、容错、降级

- 离线数据质量监控 —> **及时**发现问题
- **实时**数据质量监控 —> **提前**发现上游问题、提前介入、**避免**问题发生
- **数据质量问题截断** —> 挂起全部、更细粒度的控制 (血缘、影响指标)
- 关键数据进行 **降级、容错**
- ……

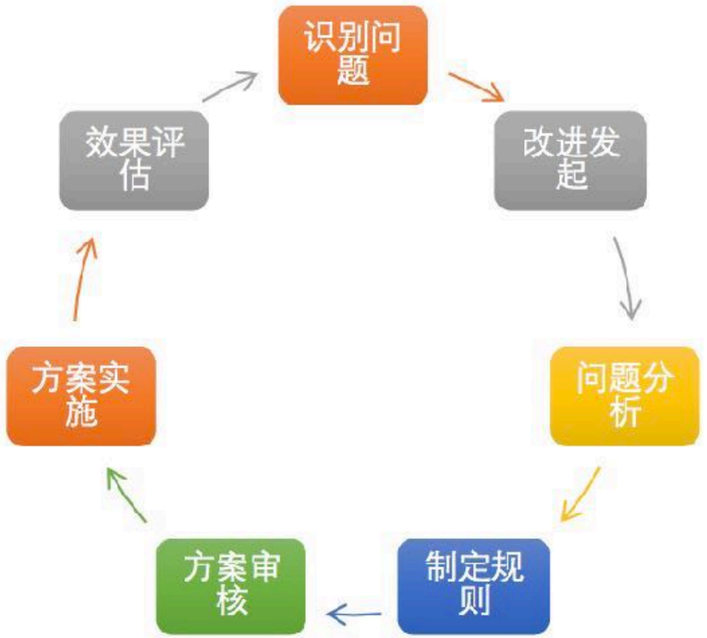
数据内容&时效性监控

- 注册SLA保障
- 数据内容监控

数据管理

数据质量管理-事后复盘

- 分析问题**本质-5why**
- 建立反馈**闭环-PDCA**
- 归纳总结**沉淀-规范**



识别、定义、解决问题、**系统化解决问题、**验收复盘、持续提升数据质量

数据管理

工具化辅助、自动化提升效率

痛点：规范 执行成本高、效果差

- 建模工具化
- 审核工具化
- 测试工具化
- ...

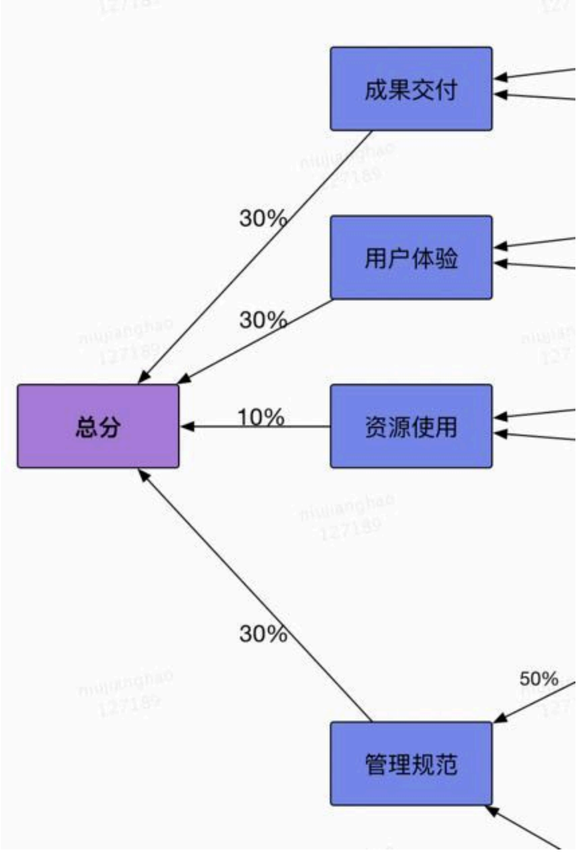
数据管理

数据生命周期管理

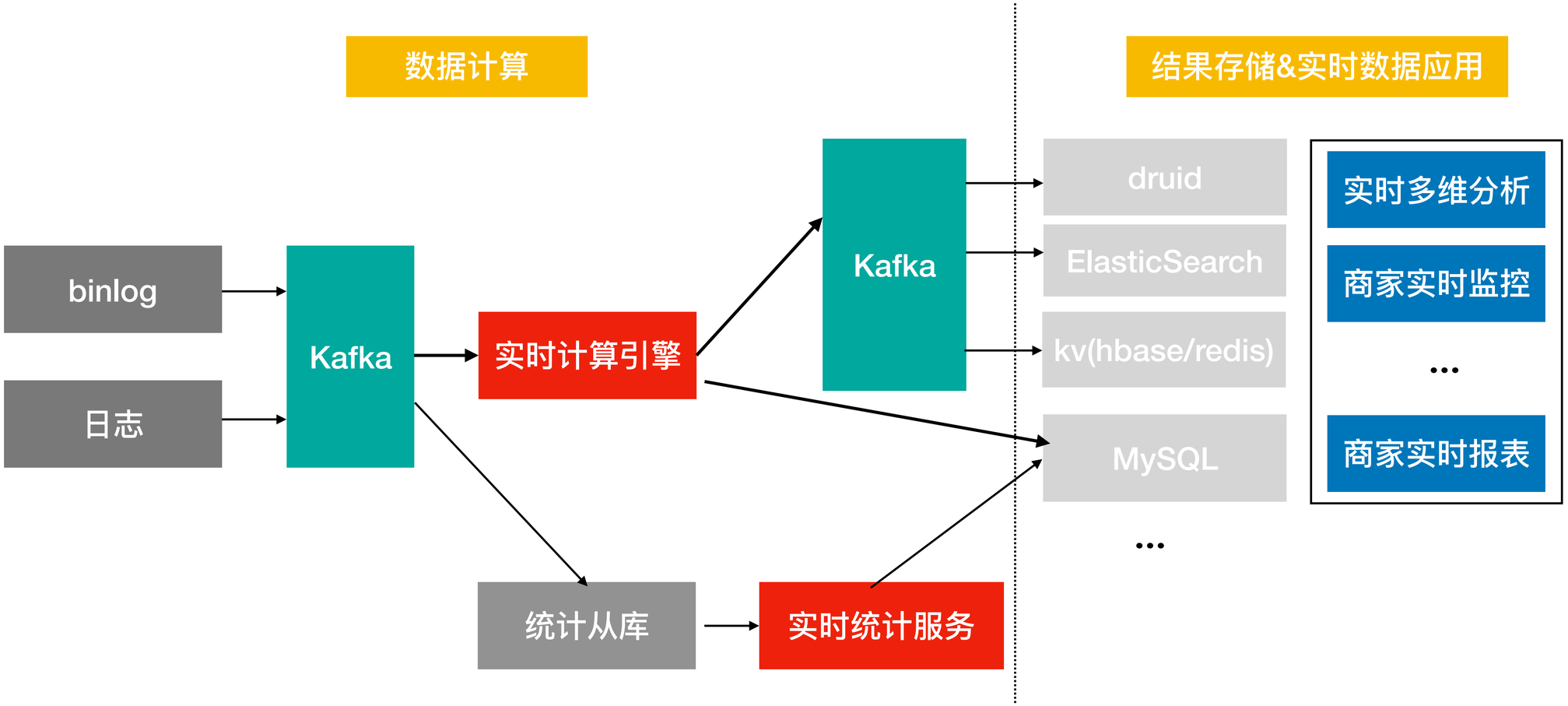
应用元数据 + 数据血缘，及时下线老化、无效数据，保证数据仓库数据整体干净有效

量化数据仓库健康度

通过**过程指标**为数据仓库建设提供**管理和优化**的**手段**



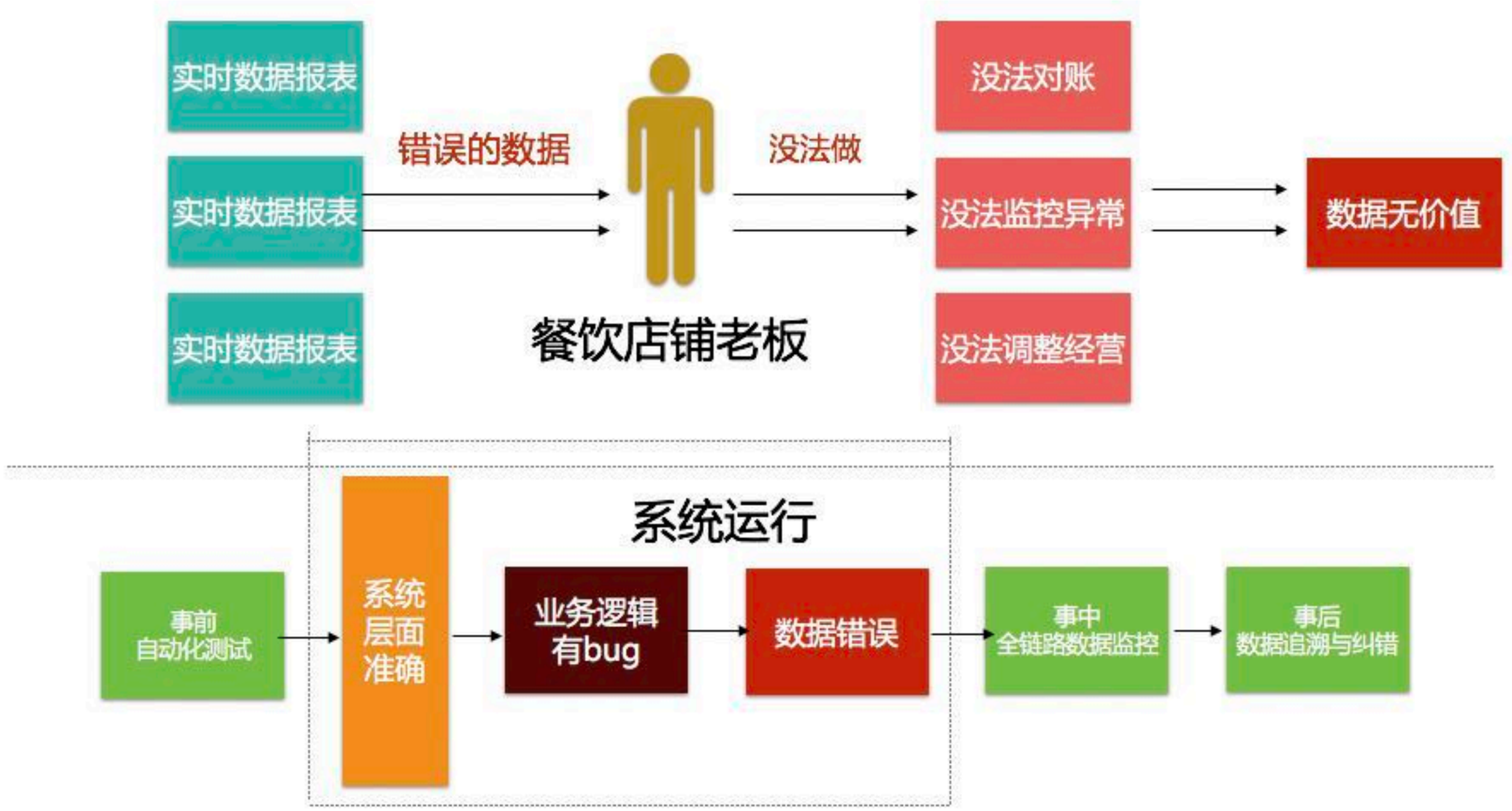
实时数据清洗



- 时效性
- 多样化窗口计算
- 开发效率
- 精确性

实时数据清洗

极高的**准确性、可用性**要求



提前识别&规避风险+缩短故障发现时间+缩短故障处理时间

数据应用

满足内部分析师查看和分析数据的需求

- 自助查询
- 多维分析
- 星空报表
- 数据API
- ...

BI产品

数据应用

商家数据可视化：满足商家经营的需求

- 自助查询
- 多维分析
- 星空报表
- 数据API
- ...
- BI产品

数据API开发平台

数据API面临的挑战

可用性

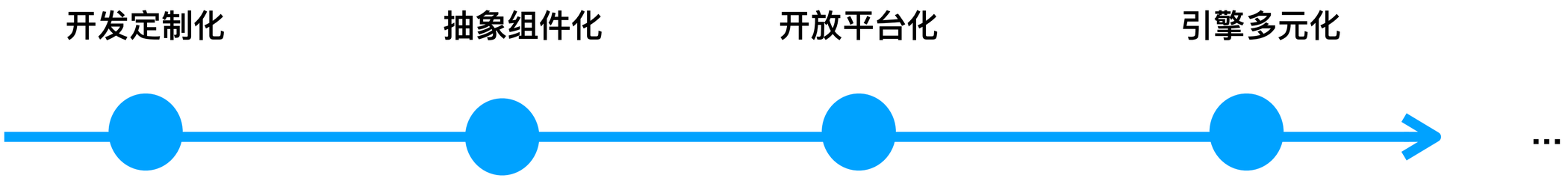
性能

效率

成本

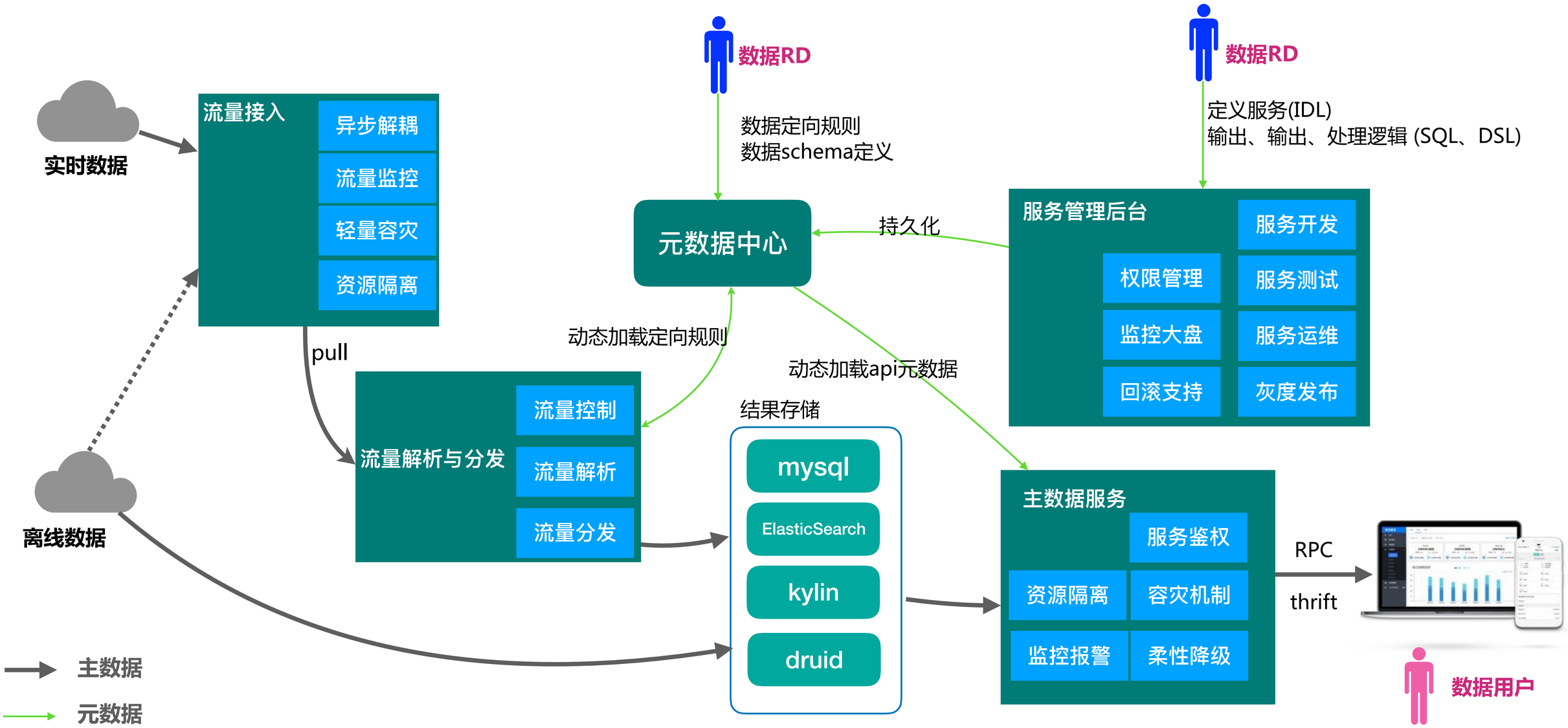
...

数据API开发模式的演进



随着业务的发展，主动识别&**预见** 技术体系中的痛点，下沉抽象**解决主要矛盾**

系统逻辑架构


数据RD
数据RD
数据定向规则
数据schema定义
定义服务(IDL)
输出、输出、处理逻辑 (SQL、DSL)
实时数据
流量接入
异步解耦
流量监控
轻量容灾
资源隔离
元数据中心
持久化
服务管理后台
服务开发
权限管理
服务测试
监控大盘
服务运维
回滚支持
灰度发布
pull
动态加载定向规则
动态加载api元数据
结果存储
流量解析与分发
流量控制
流量解析
流量分发
mysql
ElasticSearch
kylin
druid
主数据服务
服务鉴权
资源隔离
容灾机制
监控报警
柔性降级
RPC
thrift
离线数据
主数据
元数据
数据用户

取得效果

- 让不具备后台开发能力的数仓同学具备后台服务输出能力
- 2人力、98%+数据API、累积600+个
- 过去1个Q 接口累积迭代1000余次

总结与思考

- 站在业务&用户的角度看问题
- 面对问题、解决问题
- 升维&降维
- 视野是技术的全局思考力

谢谢大家

Q&A