DTCC

第十一届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2020

架构革新 高效可控

IT168.com ChinaUnix.net ITPUB

北京国际会议中心 | 2020/12/21-12/23

爱奇艺数据中台建设与应用

马金韬
爱奇艺数据中台 负责人

目录
CONTENT

01 数据生态现状和问题

02 爱奇艺数据中台介绍

03 爱奇艺数仓体系

04 数据中台场景应用

05 展望未来

01 数据生态现状和问题

随着数据的价值不断发挥，大数据生态也在形成和不断变迁，在这个过程中，数据基础设施和工具不断涌现，业务场景不断加深，遇到的问题也变得越来越突出

数据生态发展现状

随着大数据在各行各业发挥的价值越来越大，各公司对数据工作的投入也不断加大；在技术领域，为了满足不同场景的数据需求，大数据基础设施也在快速演进；数据的价值除了对数据使用能力的要求外，数据的厚度、宽度和广度也成为了数据价值的重要影响因素。

大数据
技术设施

数据应用
场景

数据治理
体系

数据组织
方式

数据生态遇到的问题

数据生产

数据生产不规范，跨业务和场景难以复用，同时，也难以保证投递的数据质量

数据采集

数据源多种多样，离线和实时采集差异大，数据接入成本高

数据开发

大数据技术门槛高，开发、运维和调优工作复杂，同时需要理解大数据架构、原理和代码

数据存储

资源管理复杂，跨集群存储需要额外成本，技术升级需业务适配

数据管理

数据资产缺乏有效的管理，所有数据一视同仁，导致资源浪费，数据使用效率低下

数据发现

数据相关信息缺乏有效的整合和输出会严重影响数据使用和交换效率

数据治理

数据质量和数据生产的稳定性对数据生态影响巨大，同时，口径不一致会导致错误决策

数据服务

数据会以多种形式输出到各种场景，缺少标准和通用的服务影响数据使用和复用效率

02 爱奇艺数据中台介绍

爱奇艺是一家以科技创新为驱动的伟大娱乐公司，新场景和新业务的探索成为公司持续发展的基础，同时成熟业务也会不断深耕业务价值，在这种背景下，需要有一套数据能力能够快速的支持好业务的各种诉求，数据中台应运而生。

爱奇艺业务矩阵


奇遇VR
VR终端设备
VR终端设备
爱奇艺VR
VR设备App
银河奇异果
智能电视
电视果
手机投屏App
电视果
手机投屏硬件
爱奇艺泡泡
明星粉丝社区
QiBubble 奇巴布
奇巴布
儿童
iQIYI 爱奇艺
叭嗒
二次元
爱奇艺体育
大众竞技娱乐
随刻
随刻创作
视频创作
电影票&演出票
爱奇艺票务
潮流电商
爱奇艺商城
斩颜
饭饭星球
主播直播
奇秀
网络游戏
爱奇艺游戏
天象互动
网络文学
爱奇艺文学
轻小说
漫画
爱奇艺漫画
知识付费
爱奇艺知识

爱奇艺数据中台

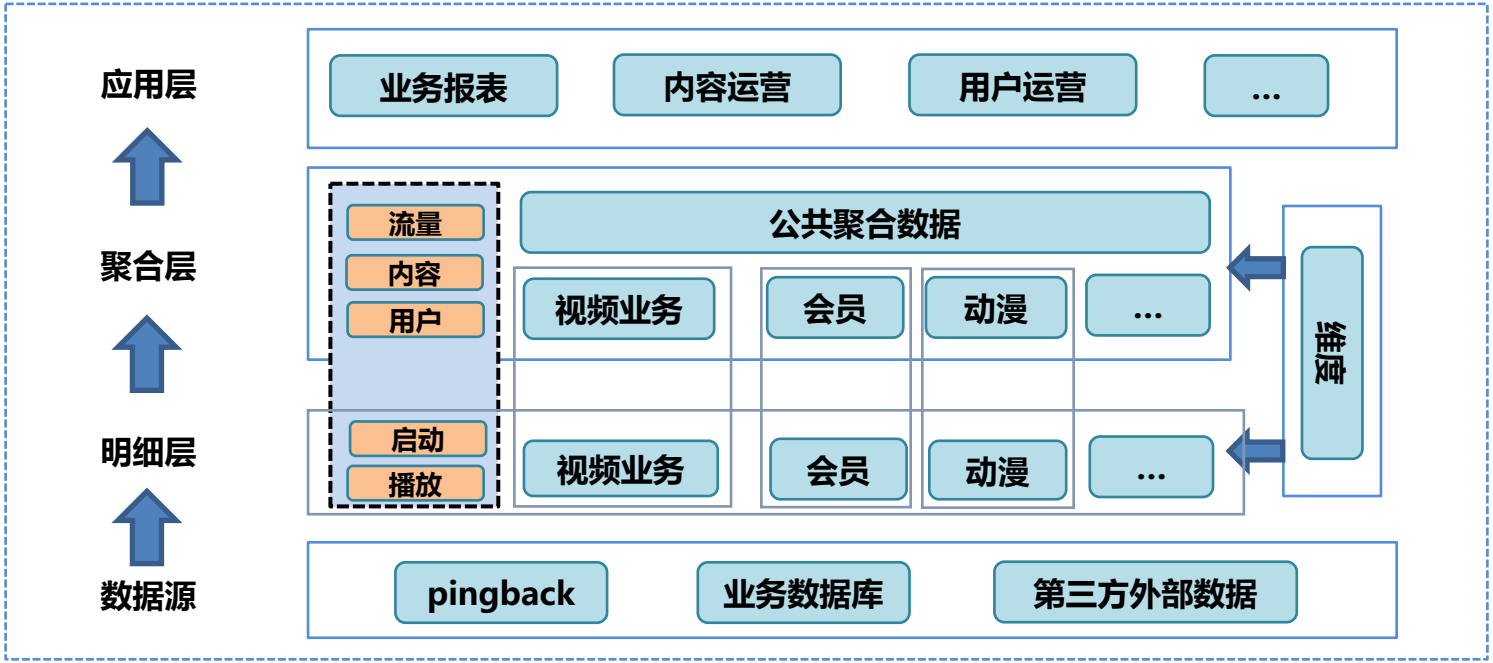

报表经营
数据科学
运营分析
数据应用
统一数据服务
可弹性 配置化 可监控 可组合
统一数据层
中心化 标准化 高质量 安全性
统一数据生产/接入
标准 简单 可配置
统一数据开发平台
一体化 智能化 灵活性
统一云服务
标准化 可弹性 高可用
统一数据治理
全覆盖
易分析
便理解
强推动

03 爱奇艺数仓体系

在整个数据中台建设过程中，统一数据层和元数据将是对用户最为直接的表现形式，本部分将对从数仓和数据图谱角度对这两部分进行介绍。

爱奇艺数仓1.0

缺陷　生产/使用效率低下、口径不统一、烟囱式建设、无工具支持



爱奇艺数仓+

统一数仓

提供底层全面且通用的数据，并作为规范的制定和约束者，
为上层提供数据和底层模型，是数据仓库建设的基础

业务集市

基于统一数仓的数据和模型，
结合业务数据分析目的，建
设满足各业务的数据集

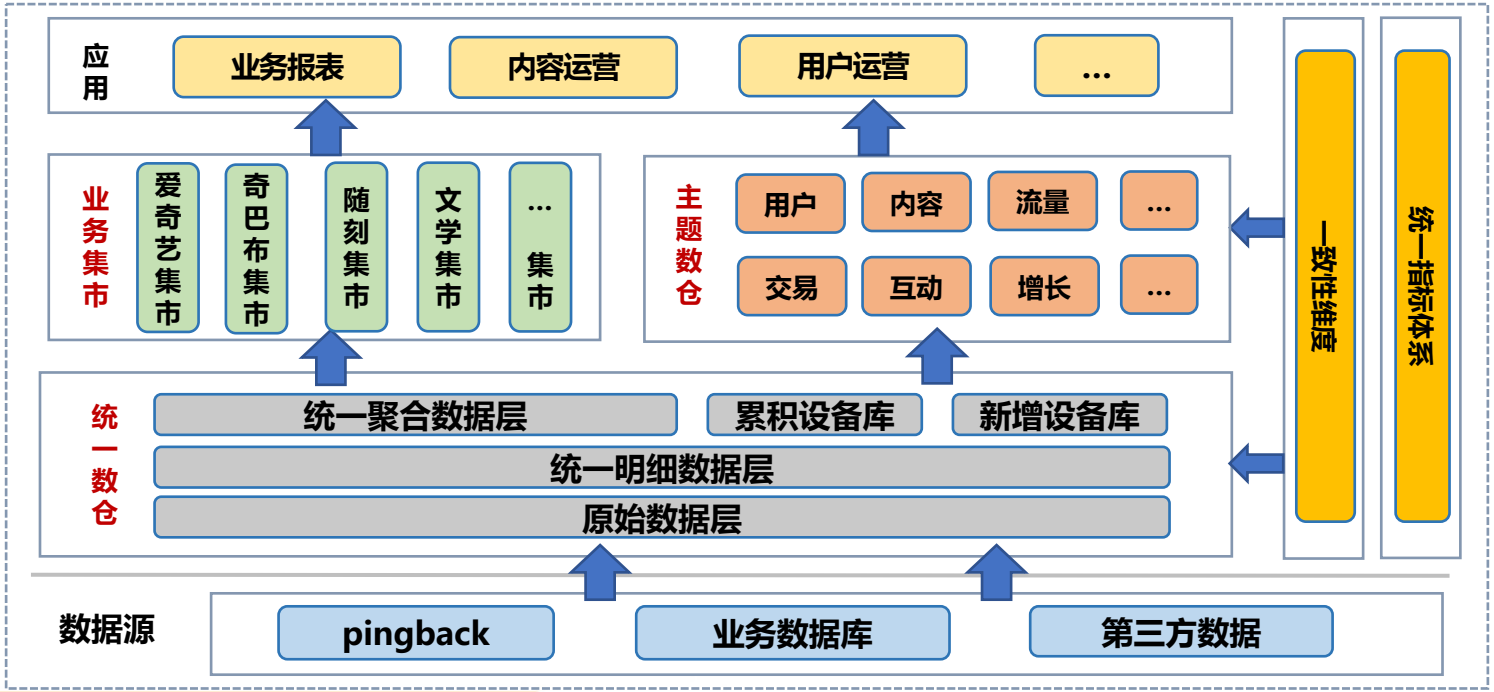
主题数仓

基于统一数仓的数据和模型，
面向不同实体的分析，建设用
户、内容、增长等领域的数据
集合

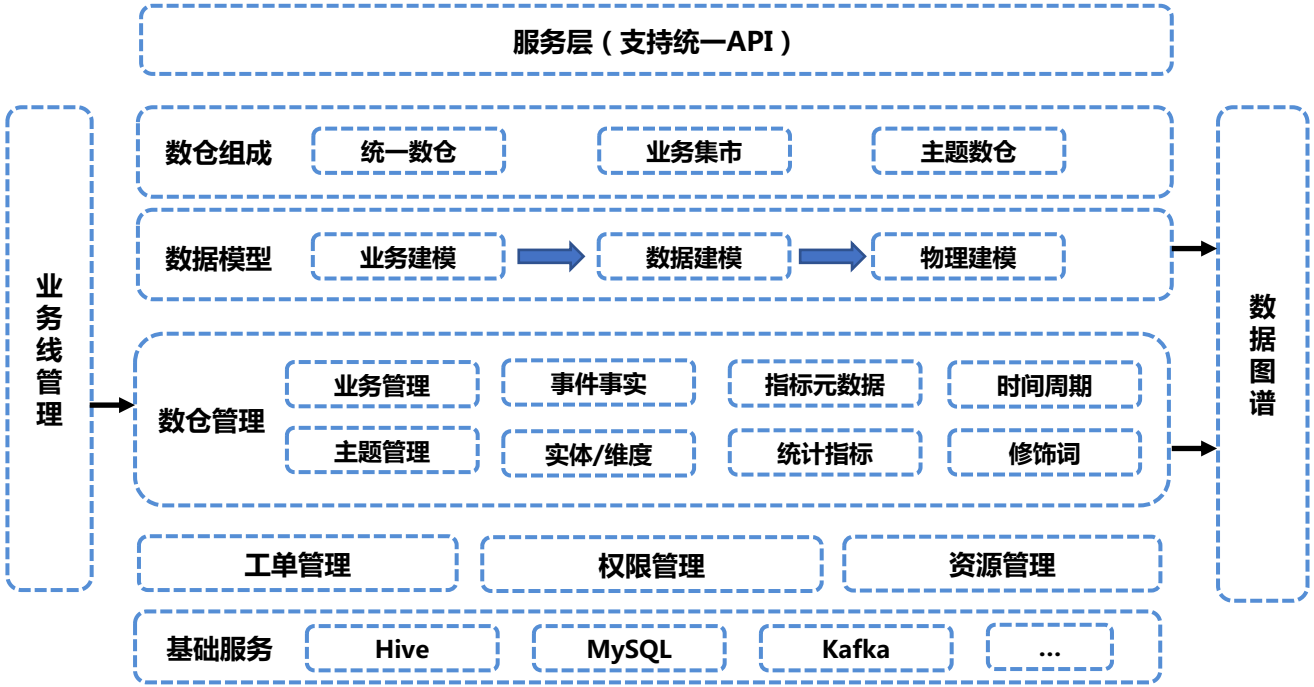
高内聚、低耦合

爱奇艺数仓+

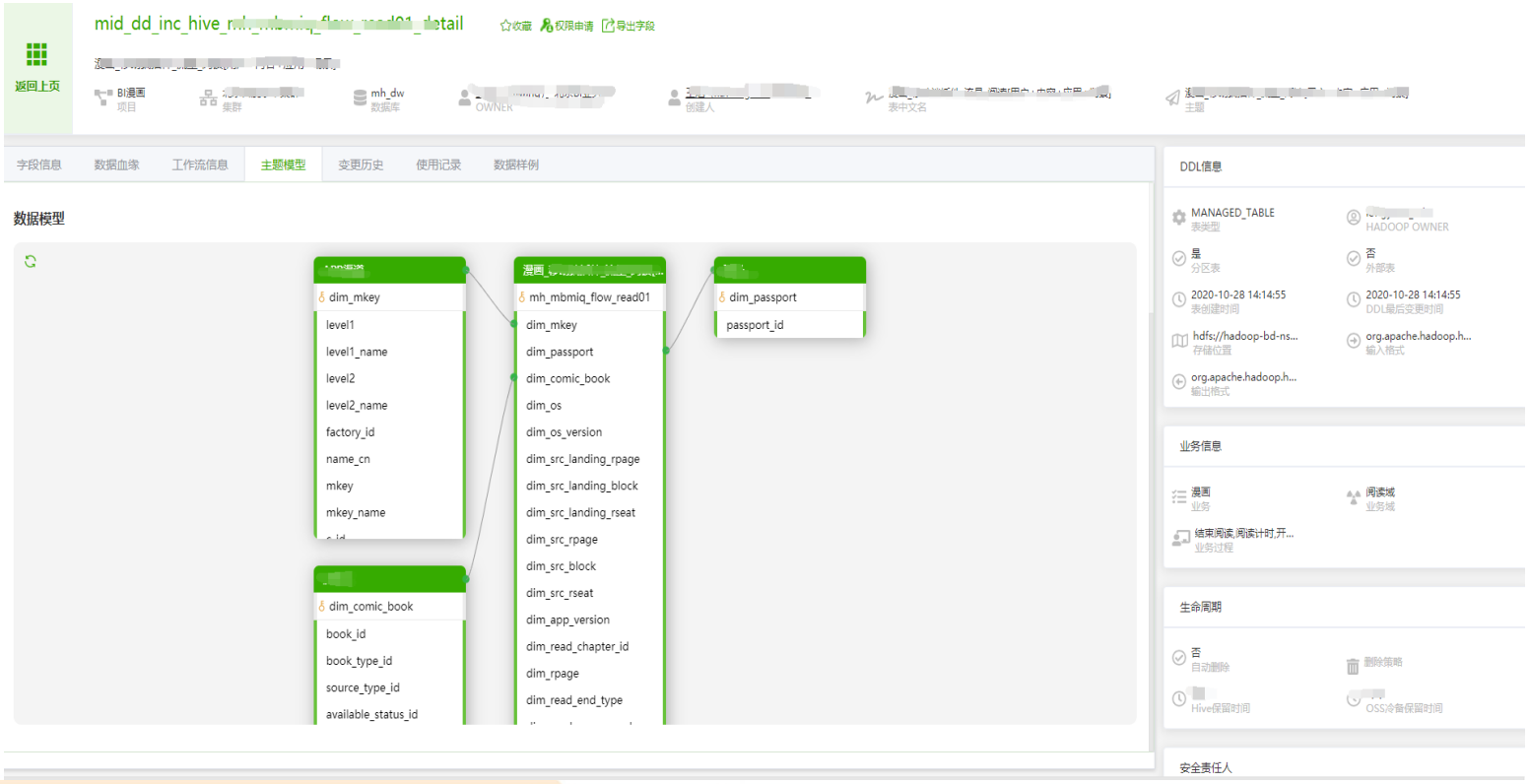
目标　明确分层/组成、规范化/标准化数据建模、口径统一、提高效率



爱奇艺数仓平台架构


服务层（支持统一API）
业务线管理
数仓组成
统一数仓
业务集市
主题数仓
数据模型
业务建模
数据建模
物理建模
数仓管理
业务管理
事件事实
指标元数据
时间周期
主题管理
实体/维度
统计指标
修饰词
数据图谱
工单管理
权限管理
资源管理
基础服务
Hive
MySQL
Kafka
...

爱奇艺数仓平台


返回上页
mid_dd_inc_hive_rnl_...detail
收藏
权限申请
导出字段
BI漫画
项目
mh_dw
数据库
OWNER
创建人
表中文名
主题
字段信息
数据血缘
工作流信息
主题模型
变更历史
使用记录
数据样例
数据模型
dim_mkey
level1
level1_name
level2
level2_name
factory_id
name_cn
mkey
mkey_name
mh_mbmiq_flow_read01
dim_mkey
dim_passport
dim_comic_book
dim_os
dim_os_version
dim_src_landing_rpage
dim_src_landing_block
dim_src_landing_rseat
dim_src_rpage
dim_src_block
dim_src_rseat
dim_app_version
dim_read_chapter_id
dim_rpage
dim_read_end_type
dim_passport
passport_id
dim_comic_book
book_id
book_type_id
source_type_id
available_status_id
DDL信息
MANAGED_TABLE
表类型
HADOOP OWNER
是
分区表
否
外部表
2020-10-28 14:14:55
表创建时间
2020-10-28 14:14:55
DDL最后变更时间
hdfs://hadoop-bd-ns...
存储位置
org.apache.hadoop.h...
输入格式
org.apache.hadoop.h...
输出格式
业务信息
漫画
业务
阅读域
业务域
结束阅读,阅读计时,开...
业务过程
生命周期
否
自动删除
删除策略
Hive保留时间
OSS冷备保留时间
安全责任人

爱奇艺数据图谱

搜索服务
- 多数据源
- 表名、维度、指标

业务图谱
- 结构化
- 向导式

数据血缘
- 寻根溯源
- 影响性分析

数据标准
- 业务信息、指标、维度、主题模型

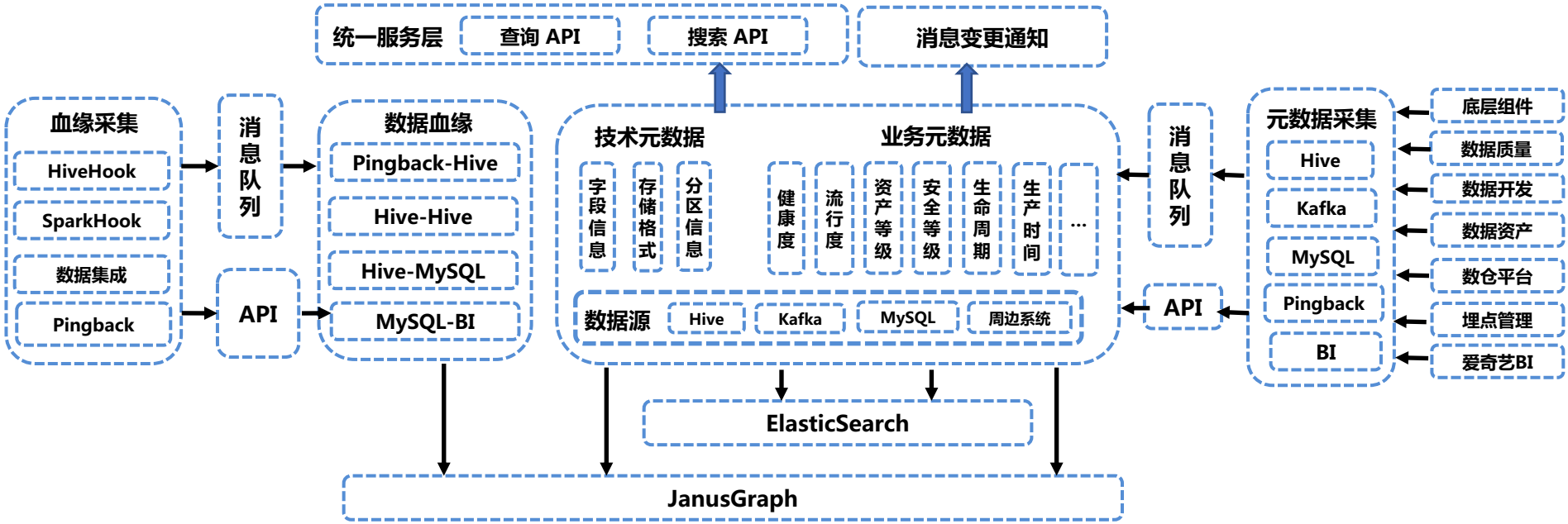
元数据中心

基础信息
- Schema
- 生产者/安全责任人
- 数据样例
- 变更历史

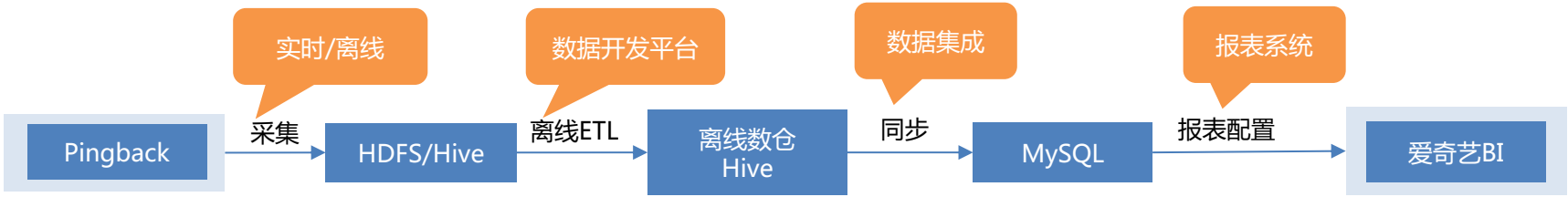
数据资产
- 资产等级
- 健康度/质量分数
- 数据流行度

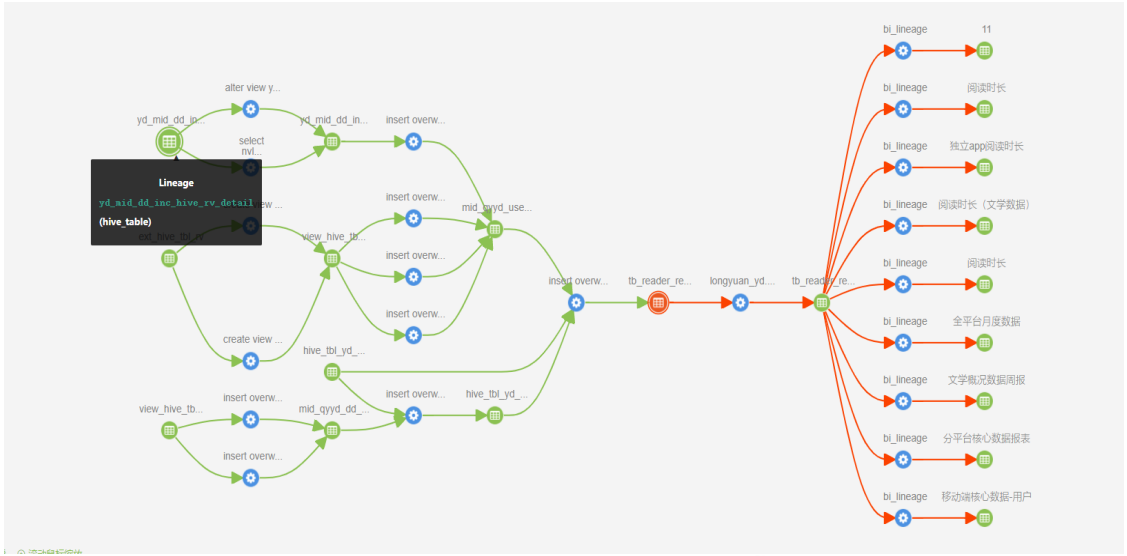
创建高效的环境，快速的“找数据”，直观的理解和使用数据，实现指导数据开发、提高开发效率等“用数据”需求

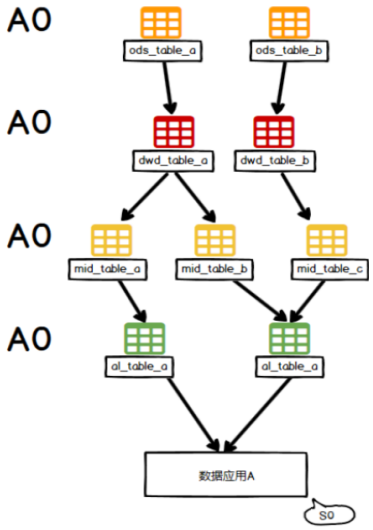
爱奇艺数据图谱


统一服务层
查询 API
搜索 API
消息变更通知
血缘采集
HiveHook
SparkHook
数据集成
Pingback
消息队列
API
数据血缘
Pingback-Hive
Hive-Hive
Hive-MySQL
MySQL-BI
技术元数据
字段信息
存储格式
分区信息
业务元数据
健康度
流行度
资产等级
安全等级
生命周期
生产时间
...
数据源
Hive
Kafka
MySQL
周边系统
消息队列
API
元数据采集
Hive
Kafka
MySQL
Pingback
BI
底层组件
数据质量
数据开发
数据资产
数仓平台
埋点管理
爱奇艺BI
ElasticSearch
JanusGraph

爱奇艺数据血缘


实时/离线
数据开发平台
数据集成
报表系统
Pingback
采集
HDFS/Hive
离线ETL
离线数仓
Hive
同步
MySQL
报表配置
爱奇艺BI


Lineage
yd_mid_dd_inc_hive_rv_detail
(hive_table)
bi_lineage
11
bi_lineage
阅读时长
bi_lineage
独立app阅读时长
bi_lineage
阅读时长（文学数据）
bi_lineage
阅读时长
bi_lineage
全平台月度数据
bi_lineage
文学概况数据周报
bi_lineage
分平台核心数据报表
bi_lineage
移动端核心数据-用户


A0
A0
A0
A0
ods_table_a
ods_table_b
dwd_table_a
dwd_table_b
mid_table_a
mid_table_b
mid_table_c
al_table_a
al_table_a
数据应用A
S0

爱奇艺数仓地图

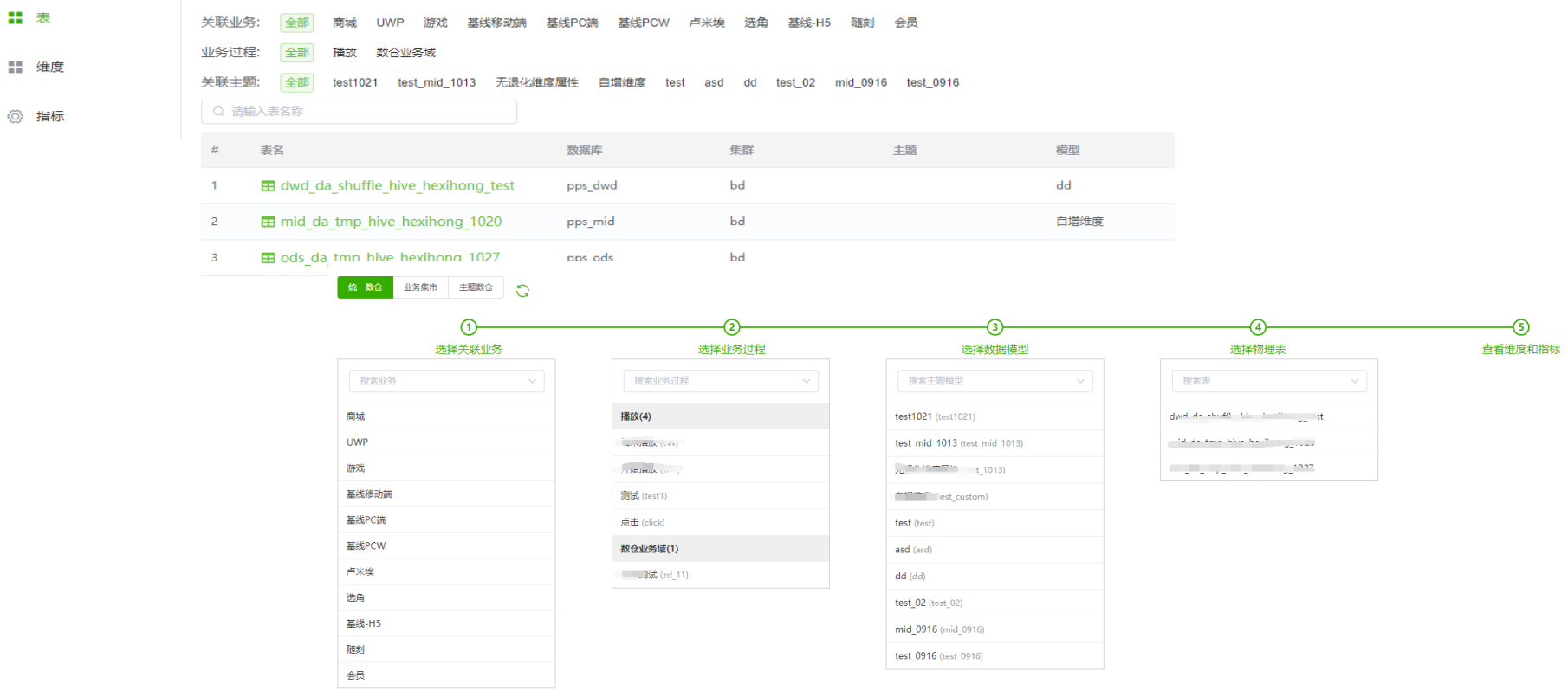


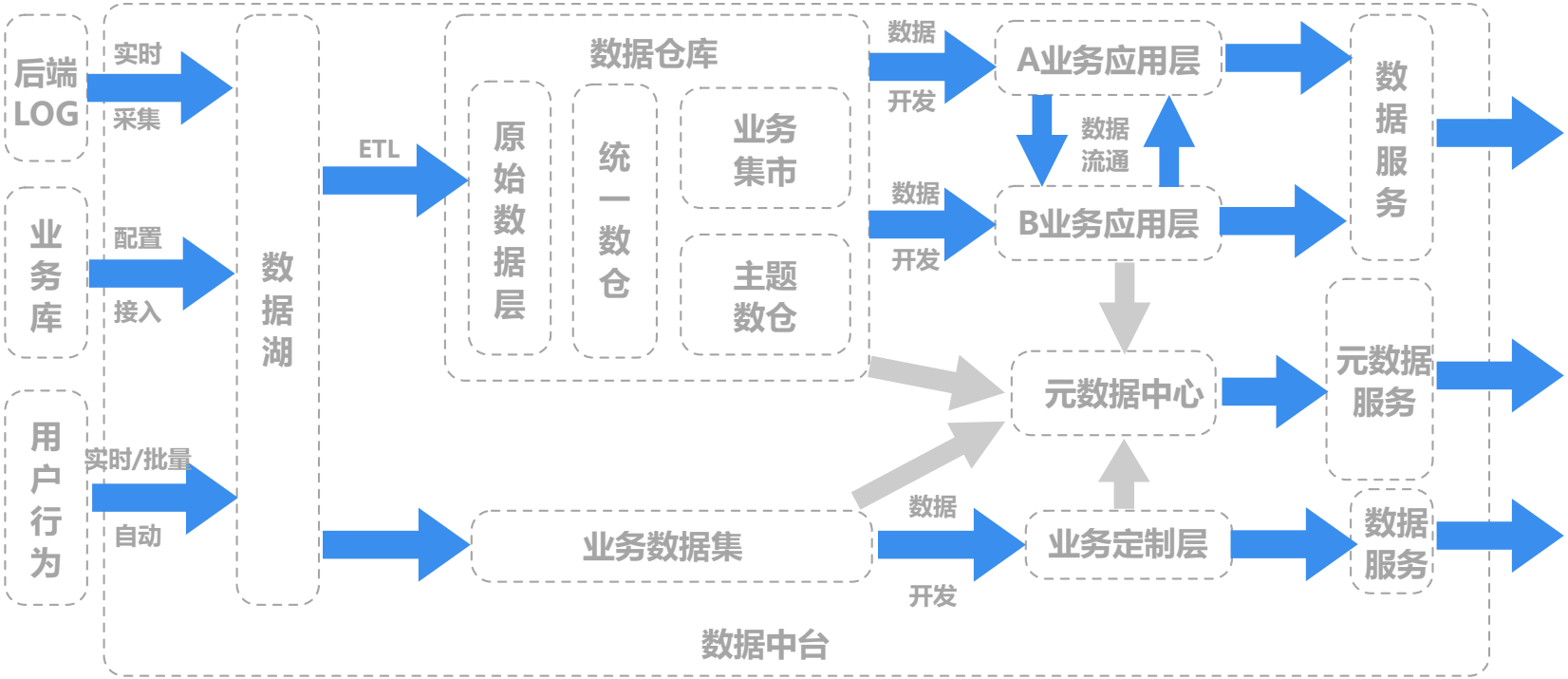
表
维度
指标
关联业务: 全部 商城 UWP 游戏 基线移动端 基线PC端 基线PCW 卢米埃 选角 基线-H5 随刻 会员
业务过程: 全部 播放 数仓业务域
关联主题: 全部 test1021 test_mid_1013 无退化维度属性 自增维度 test asd dd test_02 mid_0916 test_0916
请输入表名称
表名 数据库 集群 主题 模型
1 dwd_da_shuffle_hive_hexihong_test pps_dwd bd dd
2 mid_da_tmp_hive_hexihong_1020 pps_mid bd 自增维度
3 ods_da_tmp_hive_hexihong_1027 pps_ods bd
统一数仓 业务集市 主题数仓
1 选择关联业务
2 选择业务过程
3 选择数据模型
4 选择物理表
5 查看维度和指标
搜索业务
商城
UWP
游戏
基线移动端
基线PC端
基线PCW
卢米埃
选角
基线-H5
随刻
会员
搜索业务过程
播放(4)
测试 (test1)
点击 (click)
数仓业务域(1)
搜索主题模型
test1021 (test1021)
test_mid_1013 (test_mid_1013)
test (test)
asd (asd)
dd (dd)
test_02 (test_02)
mid_0916 (mid_0916)
test_0916 (test_0916)
搜索表

04 数据中台场景应用

数据中台的能力往往以平台形式呈现，需要通过业务场景进行能力的输出，在这个过程中会将中台和平台的区别体现淋漓尽致

定制化应用场景

适用于成熟业务，对数据有更为定制化的诉求，同时自身具备一定的开发能力



通用化应用场景

适用于新场景，满足新业务数据的快速接入和数据分析，并能提供常规的报表输出

用户视角

Step 1 申请业务ID

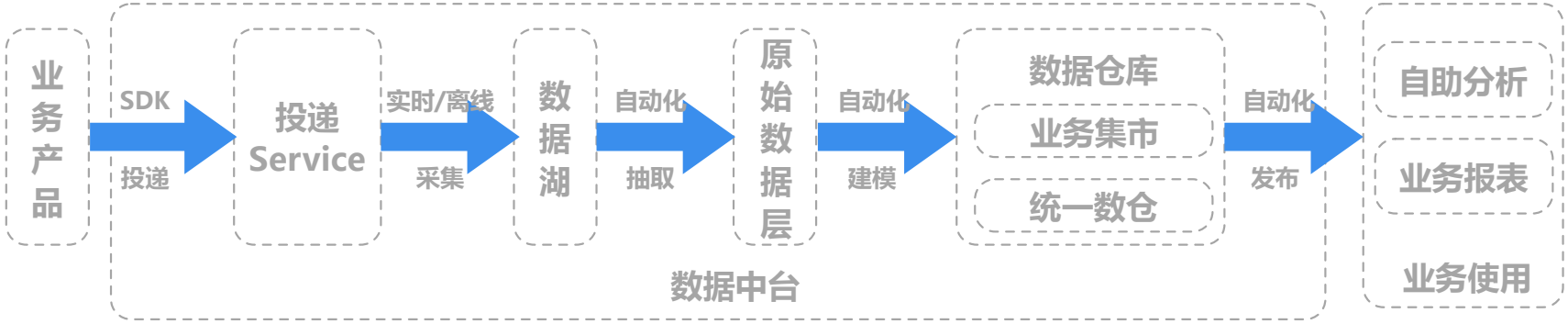
Step 2 Pingback 埋点/投递

Step 3 产品开发

Step 4 自助分析

Step 5 业务报表

底层逻辑



05 未来展望

数据的价值在不断被挖掘，基础设施快速发展，试图满足业务和用户对数据的各种应用场景，数据中台也需要持续迭代，融合更多的能力。

未来展望

1. 流批一体
2. 数据AI化
3. 数据湖
4. 智能解决方案

爱奇艺技术产品团队
公众号

THANKS