

# 面向多业务场景的数据仓库构建及 开源协同

分享人：赵德栋 (dantezhao)  
增值业务部

# 目录

---



- 微信游戏业务现状
- 整体思路
- 数据模型
- 数据规范
- 数据系统
- 开源协同

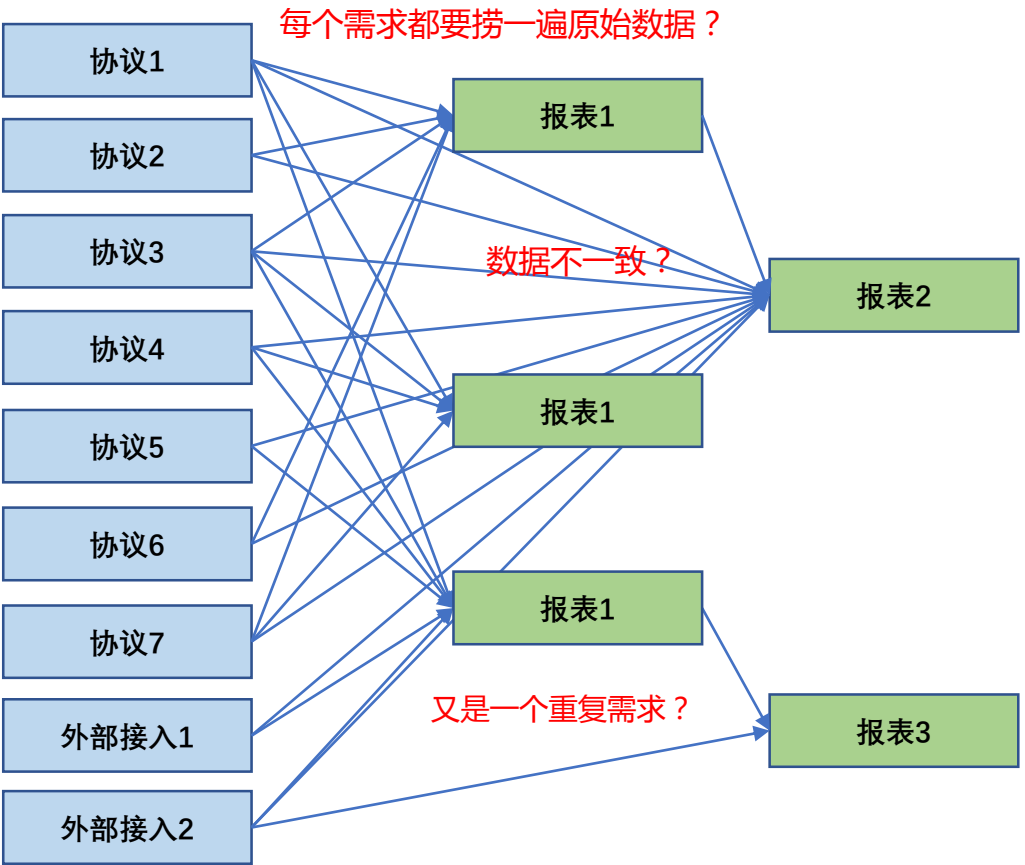
# 微信游戏业务现状

这些业务的痛点，你们也有遇到吗？

# 数据的痛，你也共鸣吗？



## 烟囱式数据模型



## 核心痛点问题



# 整体思路

在少人力的支撑下，如何高效支持多个业务场景的数据需求？

# 适合的，才是最好的！



找到业务场景中的难点，团队的特点！

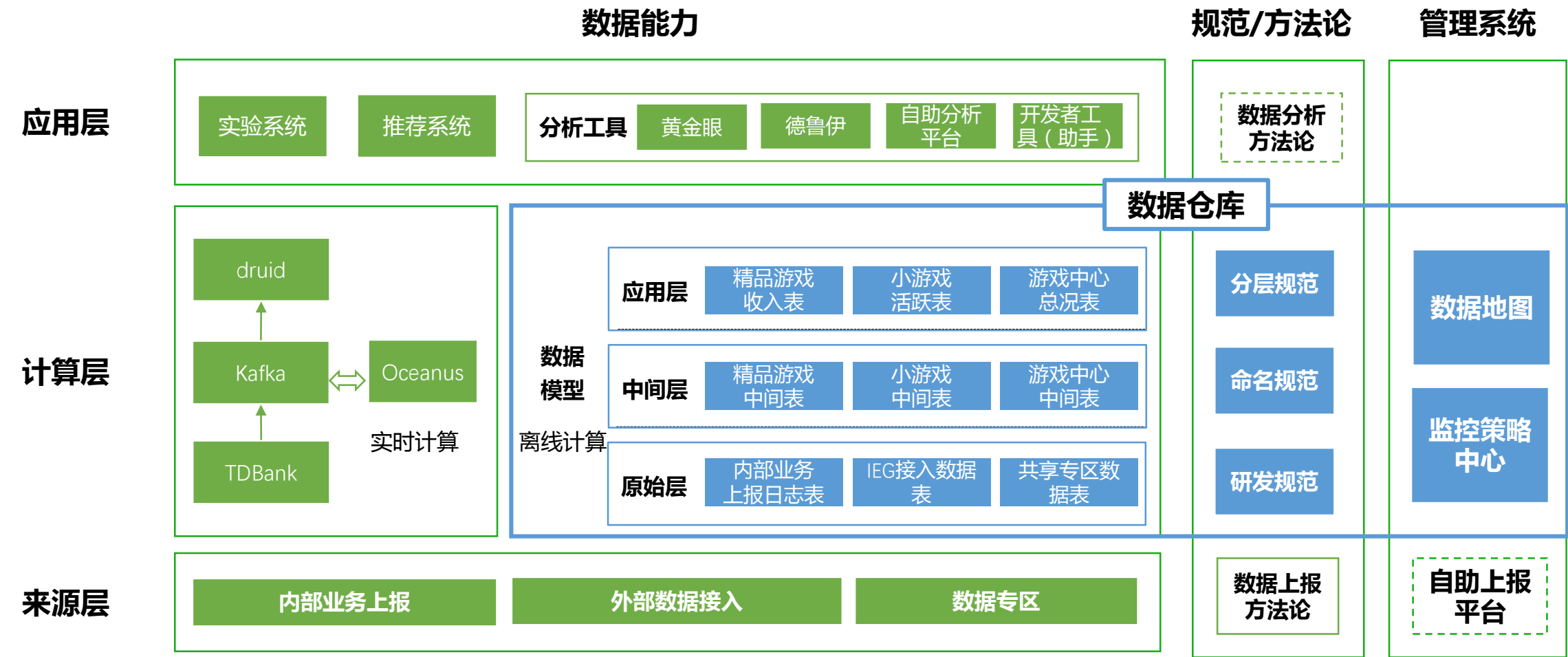
	阿里 - OneData	某信息安全业务场景	XX业务场景
业务背景	电信行业为主的数据分析场景	全业务场景的用户内容安全舆情数据分析	覆盖游戏、社区和内容业务的数据分析场景
数据背景	<div>1. 阿里系的电商行业的数据为主</div> <div>2. 业务形态比较接近</div> <div>3. 需求类型多，多种数据需求类型</div> <div>4. 有完善的公司级别的数据系统支持</div>	<div>1. 全业务场景的行为和审核数据</div> <div>2. 信息安全领域的个性化数据分析需求</div> <div>3. 支持的业务方和需求类型相对固定</div> <div>4. 缺少公司和部门完善的数据系统支持</div>	<div>1. 数据来源多：业务上报、外部接入、共享专区</div> <div>2. 业务形态各异：游戏、社区、内容等业务</div> <div>3. 需求类型多：报表类、分析类、画像类、推荐类</div> <div>4. 系统支持弱：缺少完善的数据系统支持</div>
数据团队情况	<div>&gt;300人</div> <div>集团统一的数据中台团队支持</div>	<div>&gt;10人</div> <div>专职的数据基础建设童鞋支持</div>	<div>&lt;6人</div> <div>没有专职的数据基础建设岗位，业务需求为主</div>
数据体系设计特点	<div>1. 以电商行业为主，构建全公司的数据中间层</div> <div>2. 使用公司层面提供的数据管理系统</div>	<div>1. 统一规范数据接入层的数据</div> <div>2. 主要提供宽表层的结果数据</div> <div>3. 自建数据管理系统</div>	<div>1. 根据各业务场景构建通用的数据中间层</div> <div>2. 构建基于核心维度的数据宽表层</div> <div>3. 建设数据管理系统：数据地图、数据质量监控</div> <div>4. 构建自助分析平台，解放研发人力</div>

# 数据模型为核心，数据规范为保障，数据系统为支撑！



实施部门的**数据仓库**（业务数据模型+研发规范+配套数据管理工具），对内支撑各业务线（中间表复用率70%，开发效率提升4倍），对外输出影响力

- ① **数据能力**：构建统一的数据中间层模型，为业务提供服务
- ② **管理系统**：从人工管理转变为系统管理
- ③ **规范/方法论**：统一规范标准，提升应用效率



# 数据模型

研发共建统一的中间层数据，应对基础数据的复杂性和应用灵活性

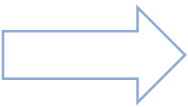
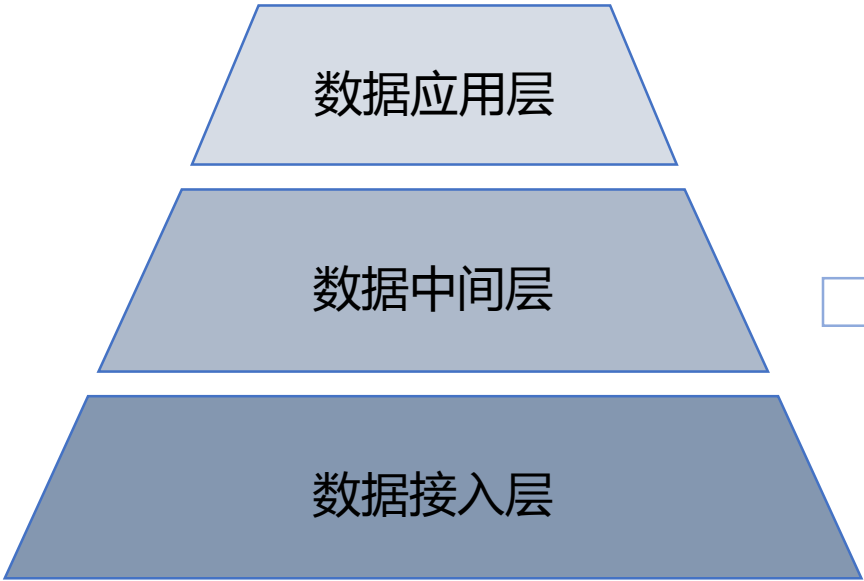


# 数据分层模型



## 数据分层设计原则：

- 1. 结合业界数据仓库设计，提出适合增值业务部的分层设计
- 2. 分层次对外提供服务，APP层直接对接业务，DW层提供通用数据中间层支持
- 3. 细化DWS层的数据设计，增加mid、rs、summary层



# 业务行为数据矩阵：设计覆盖部门全业务场景的数据模型



## 业务行为数据矩阵设计：

- 1. 从全局角度设计，覆盖增值业务部六大业务场景
- 2. 设计全业务场景下的业务数据矩阵，覆盖行为、内容、社交和资料等数据

主题分类	业务主题 数据主题	精品游戏	小游戏	游戏中心	圈子	视频直播	个性化数据 (王者、吃鸡)
通用行为	预约	√		√			
	曝光	√	√	√			
	点击	√	√				
	下载	√					√
	注册	√	√	√	√		√
	登录	√	√	√	√		√
	分享	√	√	√			
	付费	√	√				
衍生行为	活跃	√	√	√	√	√	√
	留存	√	√	√	√	√	√
	流失	√	√	√	√		
	回流	√	√	√	√		
其它行为	其它	触达数据	双栖用户		点赞、Feeds、 发表、评论		
	宽表	√		√	√		

# 通用维度设计



## 通用维度设计特点：

1. 设计符合增值业务部特点的**通用数据维度**
2. 结合数据分层模型，不同层次数据使用不同的通用维度

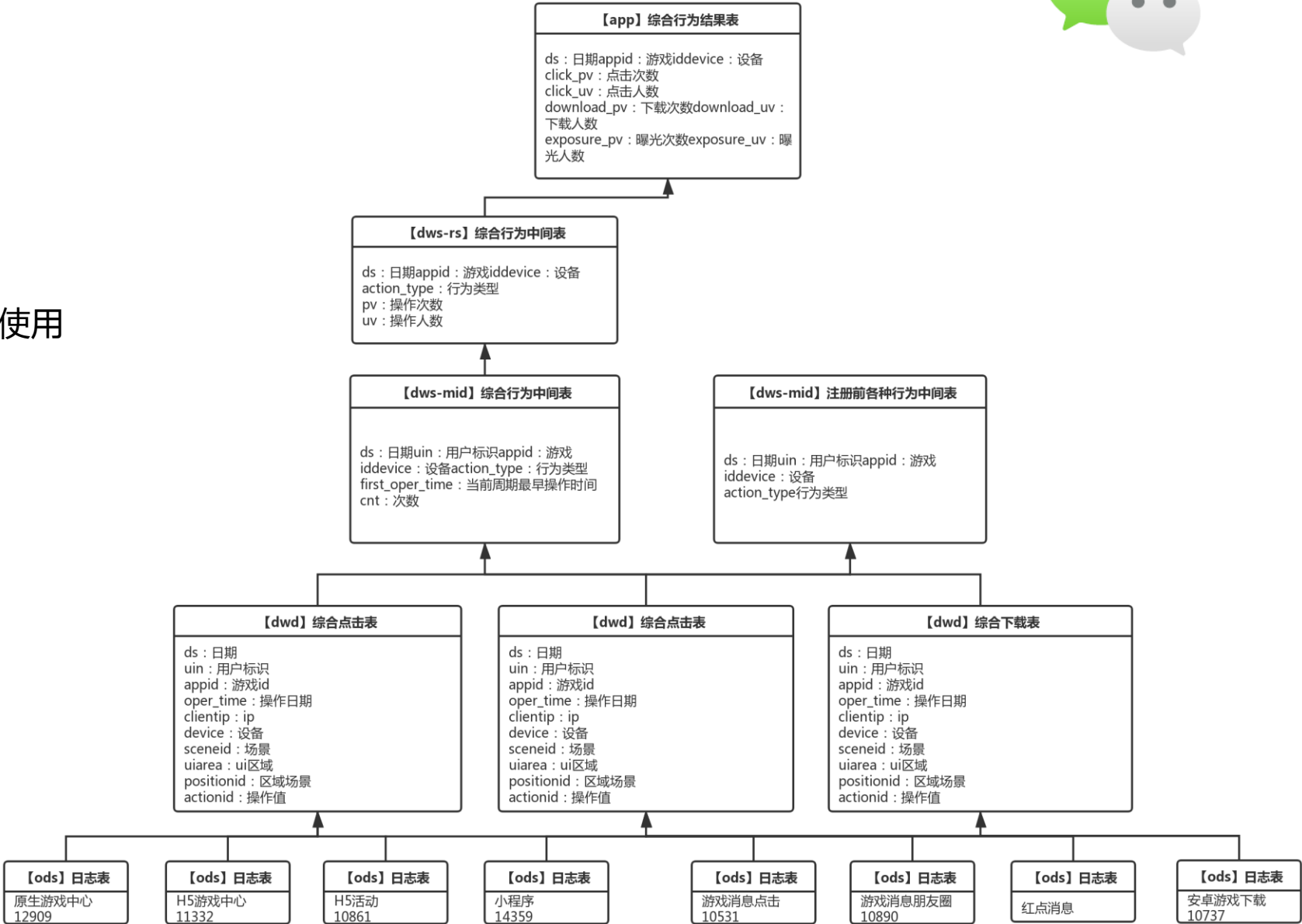
	字段	字段类型	字段名	备注
分区	logid	string	日志类型	10861, 11332, fcircle_ads等来源数据名字
	ds	int	数据日期	格式：20190101
用户标识	uin	string	微信唯一用户ID	
	commid	string	commonid	若无，则空
	openid	string	微信/qq的openid	若无，则空
	qq	bigint	QQ号	若无，则空
游戏	appid	string	游戏的appid	
时间	oper_time	string	操作时间	时间戳
地点	clientip	bigint	客户端IP	
设备环境	device	bigint	设备类型：	1：iphone；2：android 3：s60v3 4：s60v5；5：wp7；13：ipad；-1：其他
	platform	string	平台类型类型：	iOS, Android, 以后统计以大平台为主 为了兼容两种统计方式，故device和platform都保留
	networktype	string	网络类型	同原协议中的取值
场景	sceneid	bigint	场景id	同原协议中的取值
	uiarea	bigint	UI区域	同原协议中的取值
	positionid	bigint	UI位置ID	同原协议中的取值
	sourceid	bigint	源场景ID	同原协议中的取值
	destinationid	bigint	目标场景ID	同原协议中的取值
	noticeid	bigint	活动ID	同原协议中的取值
行为	actionid	bigint	操作行为值	统一所有log中的字段为actionid，只统一名字，具体取值同原协议

# 举个栗子：精品游戏数据分析需求，有效提升数据应用效率



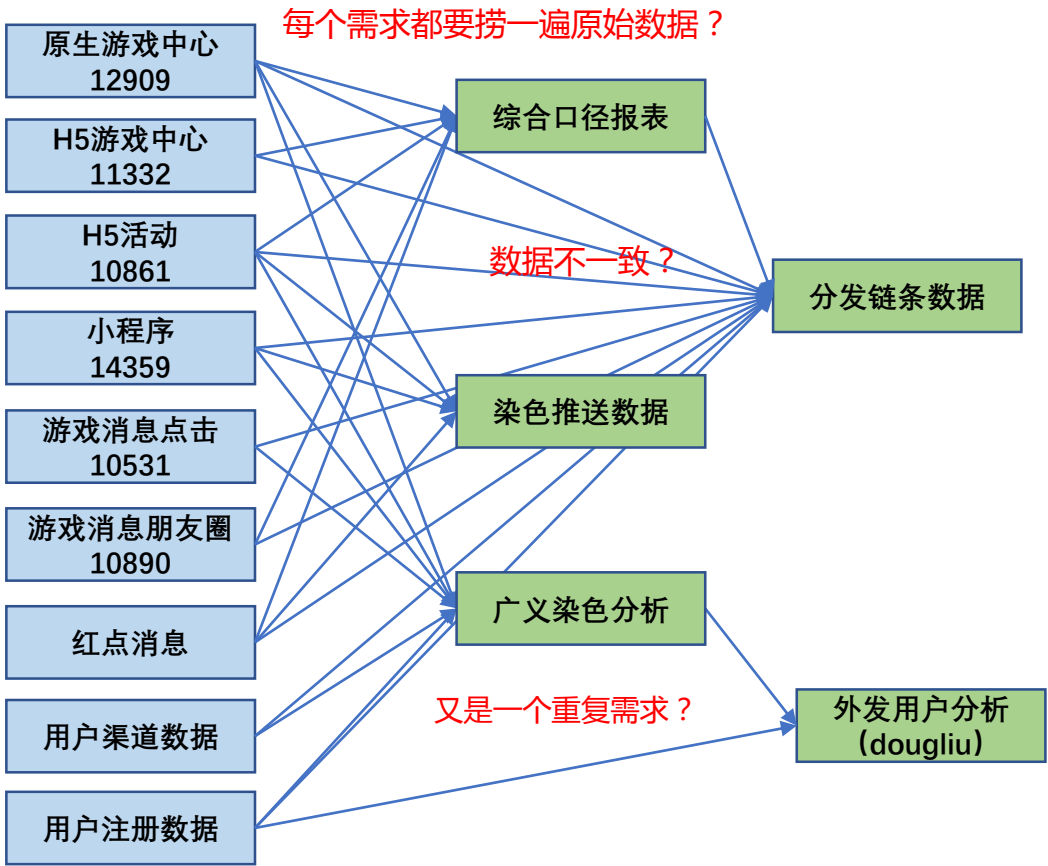
## 设计思路：

- 1. 越往上层，聚合度越高，数据量越少
- 2. 通过中间层屏蔽底层协议的影响
- 3. Dw层提供至自助分析平台，供自助分析使用
- 4. 不同层次的表，采用不同的维度

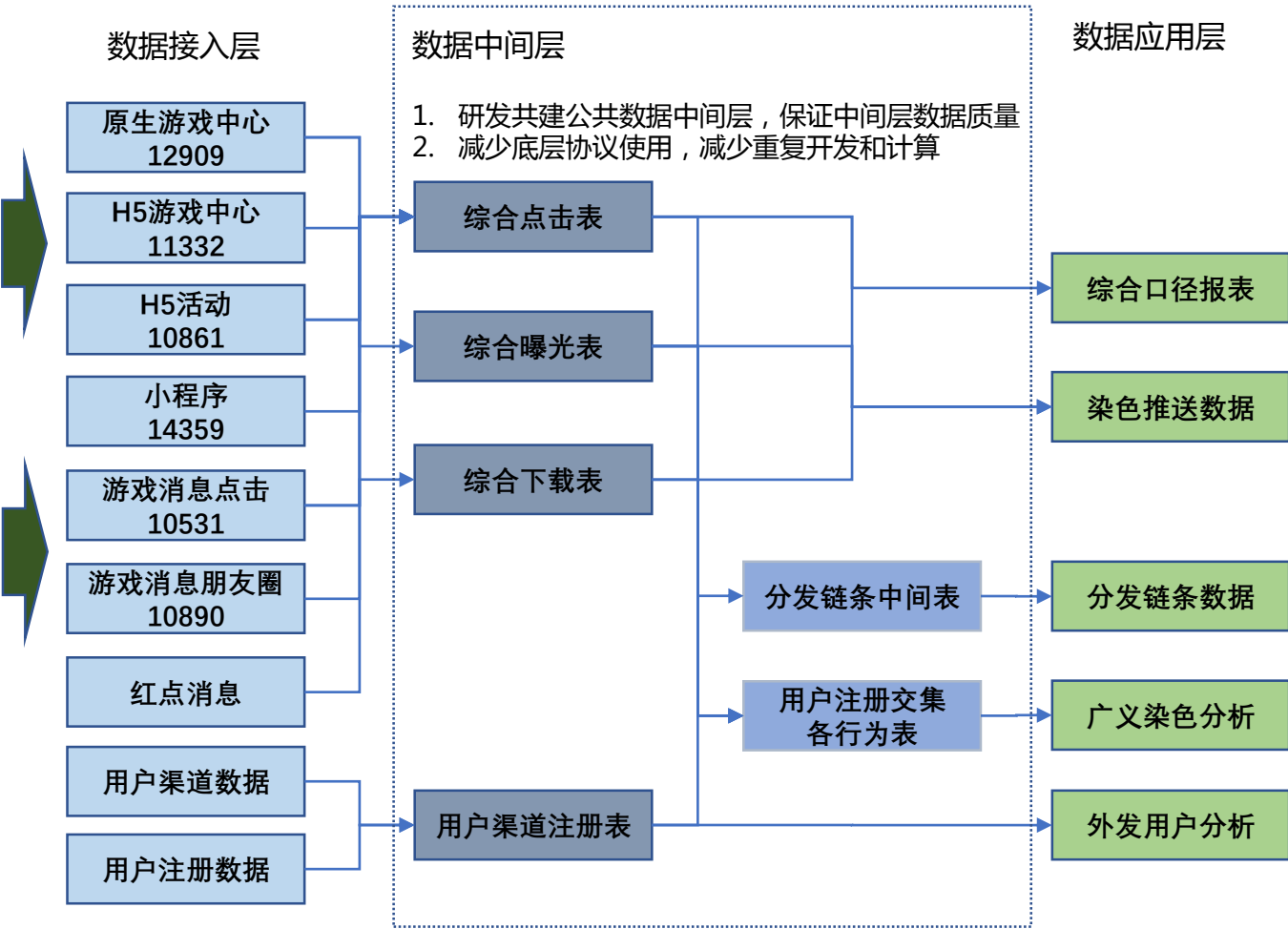




烟囱式数据模型



层次化数据模型



# 数据规范

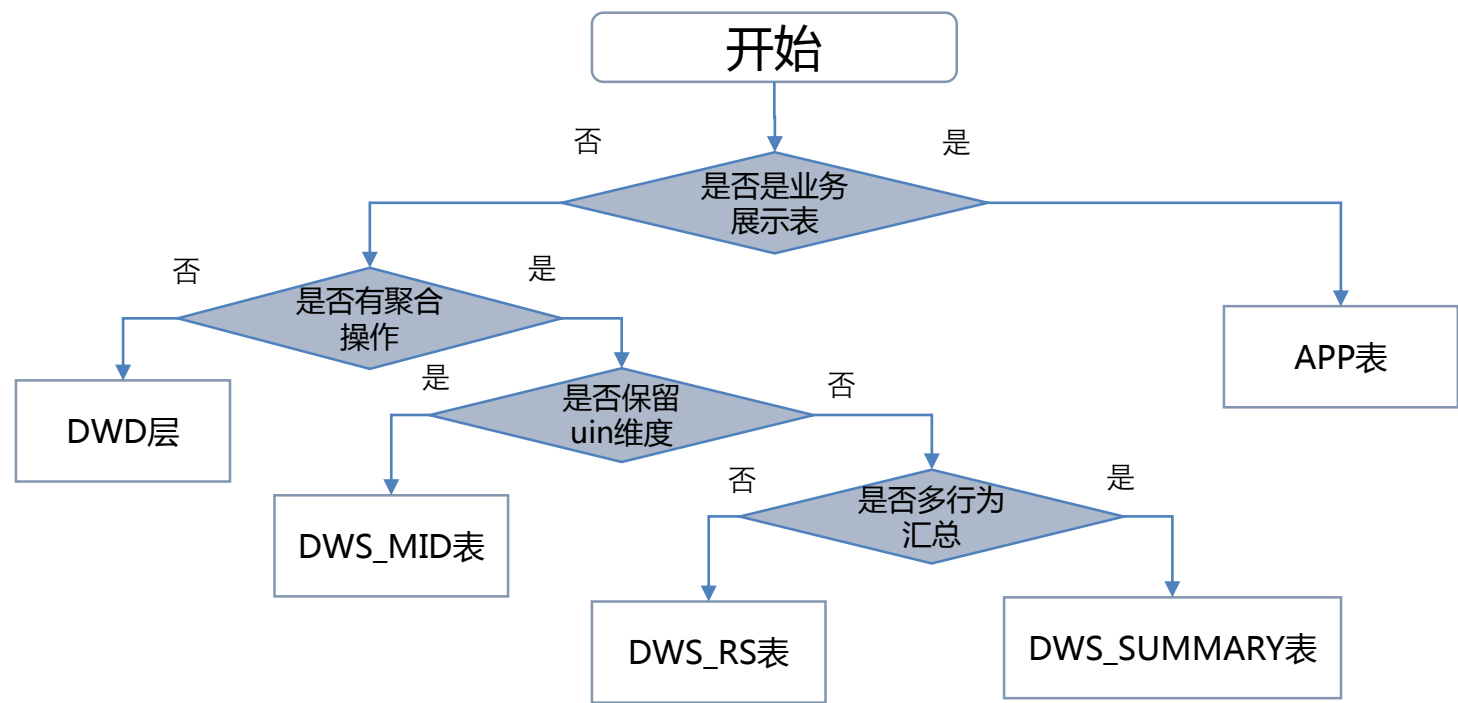
构建有明确的执行边界、可落地的数据规范

# 分层和命名规范：建立有明确标准、可落地执行的规范



## 规划设计特点：

- 1. 分层规范有**明确的划分标准**，保证**可落地执行**，没有边界模糊的地方
- 2. 命名规范结合分层规范，做到**见名知意**

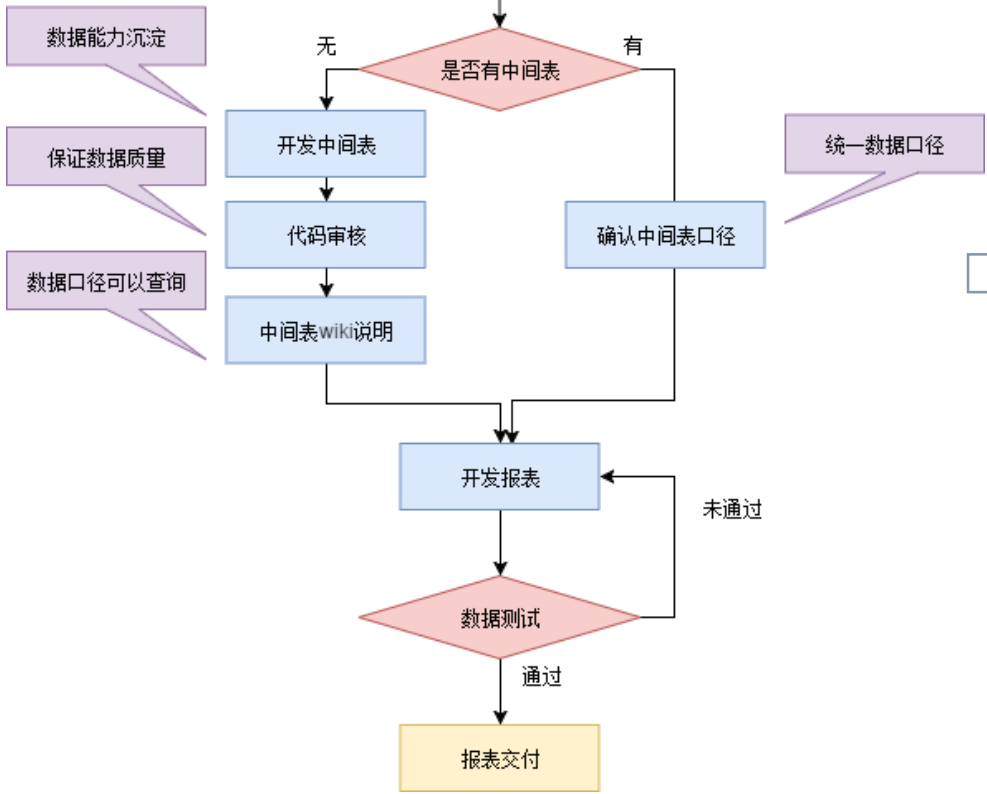
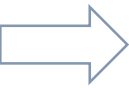
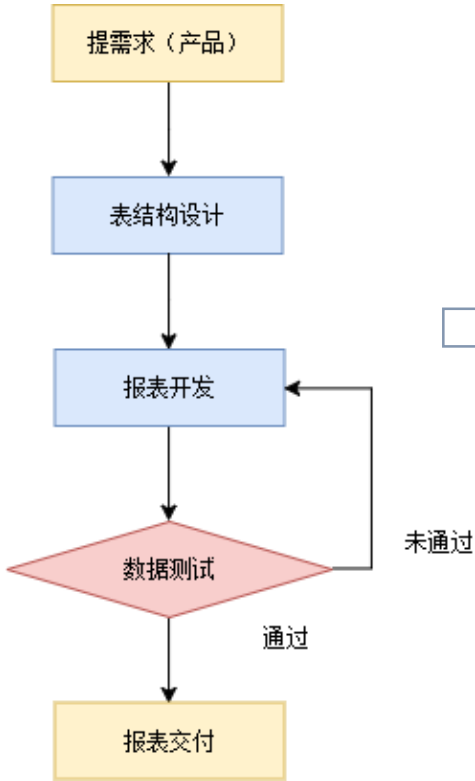


# 研发流程



## 规范化建设：

- 1. 建设中间表：沉淀通用数据能力
- 2. 使用中间表：统一数据口径
- 3. 代码审核：确保代码和数据质量



## 效果：

- 1. 代码统一Git管理
- 2. 代码上线均有同事Review

Commit statistics: master 2018-07-06 - 2019-08-22

commits	commits	contributed by
768	1.9	11
during the past 412 days	per day	authors

Chat history showing a discussion about data口径 (data口径) and code review. The conversation includes messages from @qfan and @redsealliao, discussing the need for a unified data口径 and the importance of code review.



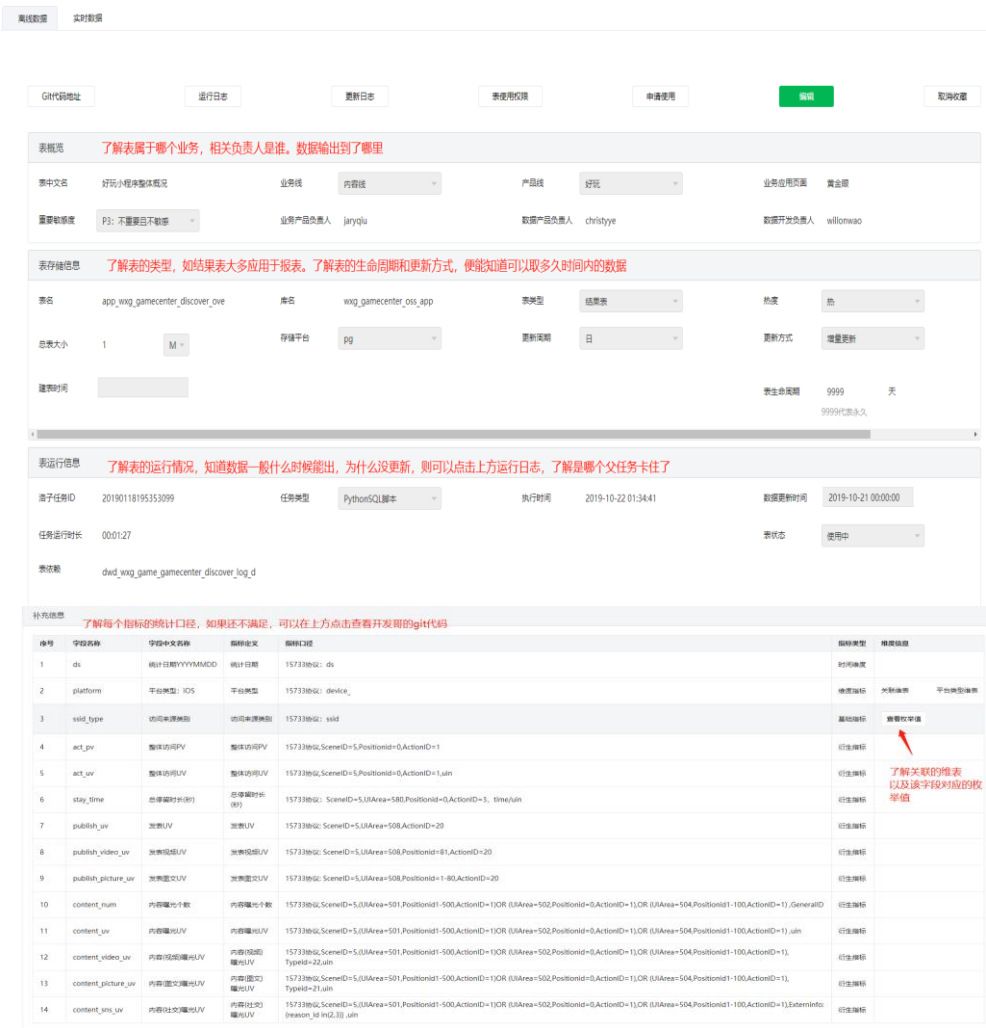
# 数据系统

数据地图提供统一的数据查询入口， 数据监控策略中心提供数据质量的保证！

# 数据地图：整合数据仓库内容，提供全局数据视图

## 数据地图：

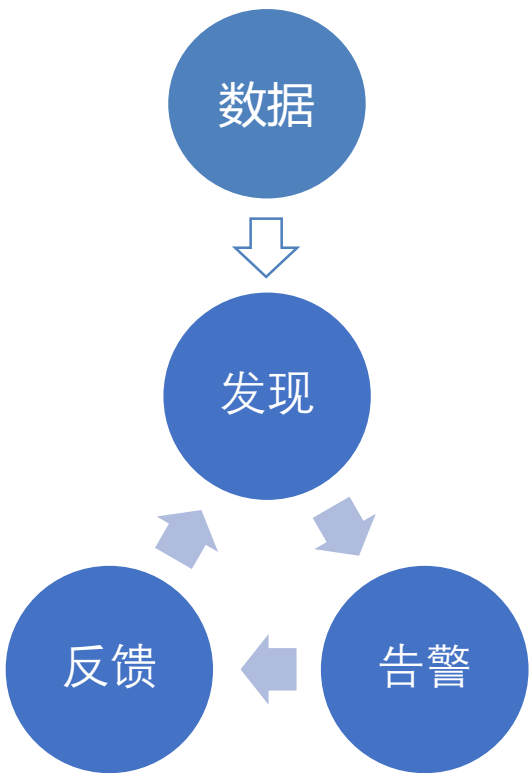
- 1. 提供统一查询数据入口
- 2. 解决“不知道有什么数据”、“数据在哪？”的困境
- 3. 打通实时和离线数据从统一接入，管理和应用的链条



# 监控策略中心：数据质量监控的平台化解决方案

## 数据质量建设：

- 1. 异常发现->异常告警->异常反馈的闭环处理能力
- 2. 增加异常告警工单处理，提高**业务感知度**



表名	字段名	维度内容	操作
游戏中心基础指标结果表	platform	platform(Android,iOS)	删除
游戏中心基础指标结果表	appid	sappid(QQ飞车,爱消除)	删除
游戏中心基础指标结果表	version	(H5游戏中心,原生游戏中心)	删除

逻辑关系 AND

下载UV 新增条件 新增条件组

下载UV 阈值 >= 1

执行周期 每天 15点

告警处理 下发工单 调用接口

数据周期 7 单位

告警列表

cloudyang

cloudyang

cloudyang

cloudyang

触发告警

处理中

处理中

已结单

2019-08-23 18:47:22

2019-07-29 17:54:16

2019-07-29 19:29:09

2019-07-31 20:12:34

cloud测试离线告警

告警监控

游戏中心下载UV监控

展示的时间范围: 08-16 00:00:00~08-22 00:00:00

指标: 下载UV

平台类型: iOS

游戏中心版本: H5游戏中心

游戏的appid: 爱消除

异常告警

1.8w

1.5w

1.3w

1.1w

8.5k

08.16

08.18

08.20

08.22

告警指标: 下载UV

告警区间: [1, ∞)

告警值: 9320

详细数据

添加评论

请输入评论

vConsole

提交

# 数据应用

七大基础能力，支持常规数据分析需求自主化

# 自助分析平台：七大功能助力” 常规 “需求自助化，解放研发人力



数据提取

模板列表

任务列表

任务实例

分析提取

数据分析

画像分析

跟踪分析

跟踪分析列表

定时任务列表

数据管理

数据包管理

数据包审批

元数据管理

我的任务（通用的任务可以设置为任务实例）

ID	任务名称	任务描述
68205	test20191010	分析提取任务描述
68204	游戏中心8.1-8.31活跃	分析提取任务描述
68203	游戏中心3.1-7.31活跃	分析提取任务描述
68202	游戏中心3.1-8.31活跃	分析提取任务描述
68201	游戏中心9.1-9.30活跃	分析提取任务描述
68200	电竞来源sept	分析提取任务描述





## 业务成果

- 共建设500+张中间表，覆盖增值业务部3大业务线。
- 数据仓库中间表复用率达到70%，单任务计算资源使用减少85.7%。
- 日常数据分析需求开发效率提升4倍，节省人力60%。

## 方法论沉淀

- 1. 一种通用的数据仓库分层方法：<http://km.oa.com/articles/view/392305>
- 2. 数据仓库实践之业务数据矩阵的设计：<http://km.oa.com/articles/view/393290>
- 3. 数据质量监控的一些思考：<http://km.oa.com/articles/view/403505>
- 4. 元数据系统的设计：<http://km.oa.com/articles/view/343125>
- 5. 数据仓库之维度建模：<http://km.oa.com/articles/view/341829>

# 开源协同

以增值业务的实践为基础， 向公司开源事业贡献一份力量

# 开源协同：数据资产管理（河图）



**Oteam介绍：**通过“数据治理+数据资产化”的管理过程对传统数据管理方式进行扩展与升级，从而有效实施数据管控、保护、交付和提高数据资产价值。

**获得荣誉：**

- 1. 腾讯“河图”代表互联网行业参与制定《数据标准管理实践白皮书》，推动行业标准化建设
- 2. 信通院公布国内行业数据管理平台图谱，腾讯数据资产管理（“河图”）作为互联网行业代表上榜
- 3. 腾讯公司级开源协同奖

**个人职责：**PMC成员之一，参与数据仓库规范编写。目前已经完成《[腾讯数据仓库规范体系](#)》第一版。

## 腾讯数据仓库规范体系

版本号	版本描述	编辑	审阅
V0.1	创建文档	chaoboxiong(熊超波); dantezhao(赵德栋); winstonliu(刘玮)	christyye(叶萍);














### 一、数据仓库模型设计





国内数据管理平台企业、产品和应用行业

2019 数据资产管理大会  
Data Asset Management Conference 2019

企业	产品	主要应用行业	企业	产品	主要应用行业
 Tencent 腾讯	河图—腾讯数据资产管理	互联网	 inspur 浪潮	浪潮数据湖平台(iDataLake)	政府、医疗
 阿里云 aliyun.com	阿里云大数据平台 ( DataWorks&DataQ )	政务、零售	 Dt Dream 数梦工场	DTSphere 数据管理平台	政府、公安
 Bai du 百度	百度数智平台-百度数据治理 Dayu	零售、金融、政务、 互联网	 YEEXUN	逸迅企业数据治理工具软件	政府、物流、金 融、制造、零售
 HUAWEI	智能数据湖运营平台DAYU	工业、通信	 浩鲸科技	鲸灵数据治理平台 ( ZSmartDGP )	政府、交通
 中国南方电网	数据资产管理平台	能源	 datablau	Datablau	金融、交通、能 源
 GRIDSUM 国双	大数据治理软件	政府、传媒、司法	 亿信华辰 ESEN SOFT	数据治理管理平台—睿治	金融、健康、能 源
 CS&S 中软国际	DAM数据资产管理平台	工业、通信、金融、 政府	 Sefonsoft 四方伟业	SDC Govern数据治理	政府、金融、电 信
 中国移动 China Mobile	大云数据管理套件	通信	 ENJOYOR 银江股份	数据分析与决策辅助平台	交通、健康
 TRANSWARP 星环科技	Transwarp Data Hub企业级一 站式大数据综合平台	金融、能源、交通、 政府	 NEUSOFT 东软	SaCa Data Integration 东软大数据 治理平台、UniEAP report报表平台	政府
 东方金信 Eastern Jin Technology	海盒数据资产管理系统	政府、金融、工业	 Hudaque — 华傲数据 —	数据标准治理系统 ( auDSG )	司法
 H3C	DataEngine平台产品 DataEngine BI 商业智能产品	通信、金融、政府、 公安、教育	 ZTE中兴	中兴数据管理平台	通信、政府
 百分点 BAIFENDIAN.COM	大数据操作系统( BD -OS )	政府、公安	 PRIMETON · 普元	元数据平台、数据质量平台	金融、政府、电 信

# 总结与展望

数据服务平台整体架构及演进历程回顾

# 数据服务平台演进历程



DataMore

## 进阶阶段

## 优化阶段

## 初始阶段

### ● 数据工具

- ✓ 三方共建Face画像平台
- ✓ 黄金眼报表
- ✓ Idea提取模板

### ● 方法论

- ✓ 《数据上报流程》
- ✓ 《需求评审流程》

### ● 工具平台

- ✓ 自助分析平台
- ✓ 实时上报监控

### ● 方法论

- ✓ 《数据上报优化流程》

### ● 模型能力

- ✓ 推荐模型
- ✓ 挖掘模型

### ● 工具平台

- ✓ 实时分析平台- 德鲁伊
- ✓ 数据管理平台- 数据地图
- 监控策略中心- 业务监控

### ● 方法论

- ✓ 《数据仓库模型方法》
- ✓ 《数据仓库研发规范》

### ● 模型能力

- ✓ 统一推荐平台
- ✓ 全业务画像标签

### ● 工具平台

- 数据管理平台- 血缘分析
- 监控策略中心- 数据治理
- 自助上报平台
- 智能业务分析

### ● 方法论

- 《数据上报规范和自助化流程》
- 《数据分析方法论》

### ● 模型能力

- 全局业务评估
- 在线学习

人工借力

探索自研

夯实中台

开放赋能

**THANKS**

CODING IN  
TENCENT

腾讯开放技术