美团点评数据仓库开发模式演进

DTCC

2016中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2016

数 据 定 义 未 来

SequeMedia 盛拓传媒 IT168.com ChinaUnix.net ITPUB

关于我

宋洪鑫

- 2011-2013 阿里巴巴
  - 数据实时计算
- 2014-至今 美团点评
  - 数据仓库开发
  - 数据治理
  - 资源管理与权限控制
  - 任务例行执行解决方案

大纲

美团网
meituan.com

- 现状介绍
- 开发模式回顾
- 面临的挑战
- 解决方案
- 总结与展望

大纲

美团网
meituan.com

- 现状介绍
- 开发模式回顾
- 面临的挑战
- 解决方案
- 总结与展望

数据仓库现状介绍

美团网 meituan.com

- ETL任务数：17000+
- 数据RD：600+
- 集群规模：3000+节点
- 数据总量：40P+ ， 300T/天增量

数据仓库现状介绍

美团网 meituan.com

- 支撑的业务：40+业务场景
  - 餐饮
  - 旅游
  - 电影
  - 外卖
  - 配送
  - 广告
  - 风控等

数据仓库现状介绍

美团网 meituan.com

- 美团从2010年3月上线至今，数据仓库是如何发展成今天这样规模的？

- 开发模式：
  - 数据仓库开发过程中，数据需求方/分析师，业务方数据RD，以及数据平台RD是之间是如何协作的，角色定位是怎样的，以及这个过程中平台机制上是如何支持的

数据仓库现状介绍

- 开发模式演进时间线

大纲

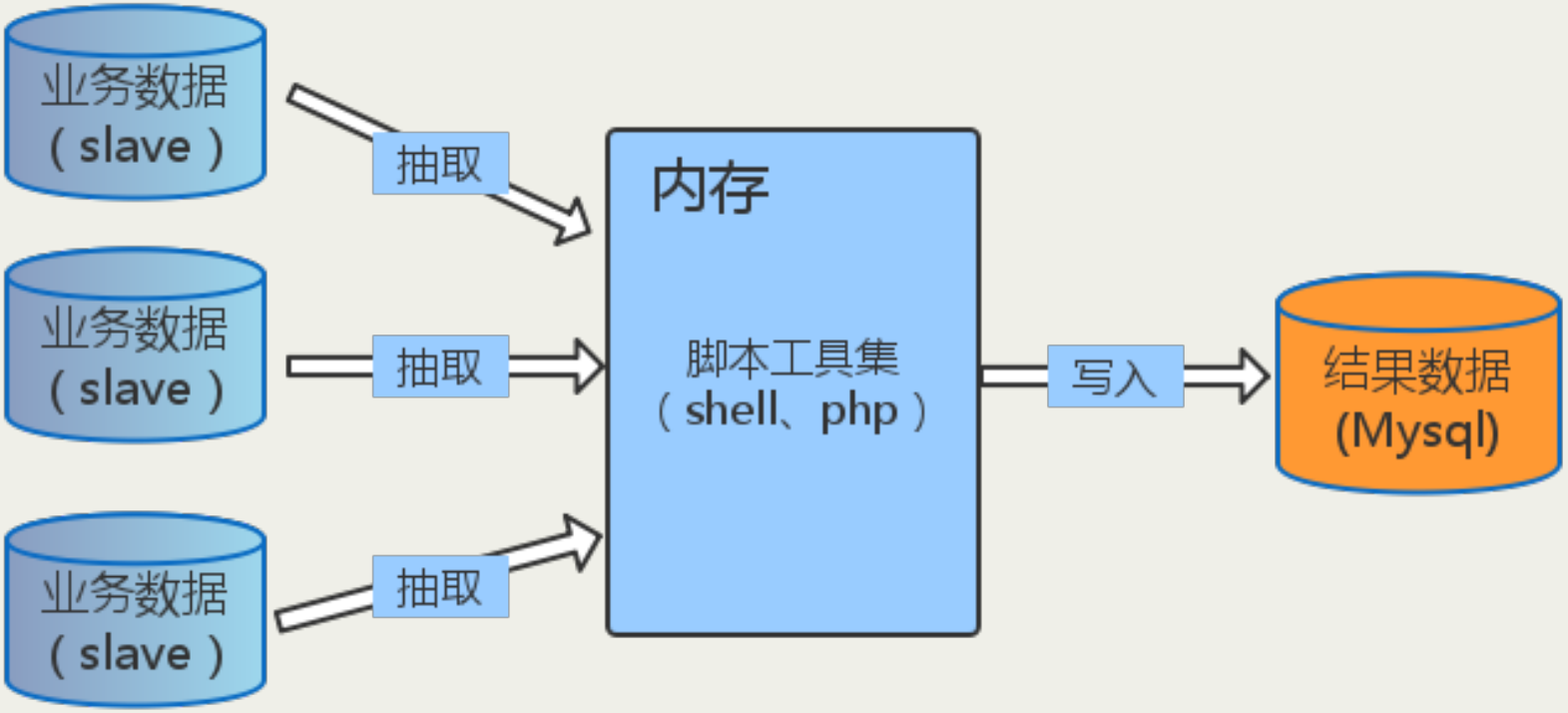
美团网
meituan.com

- 现状介绍
- 开发模式回顾
- 面临的挑战
- 解决方案
- 总结与展望

史前模式

美团网 meituan.com

- 业务上包办：数据组RD直接对接公司内部所有数据需求（主站团购）
- 技术上简陋：



史前模式问题

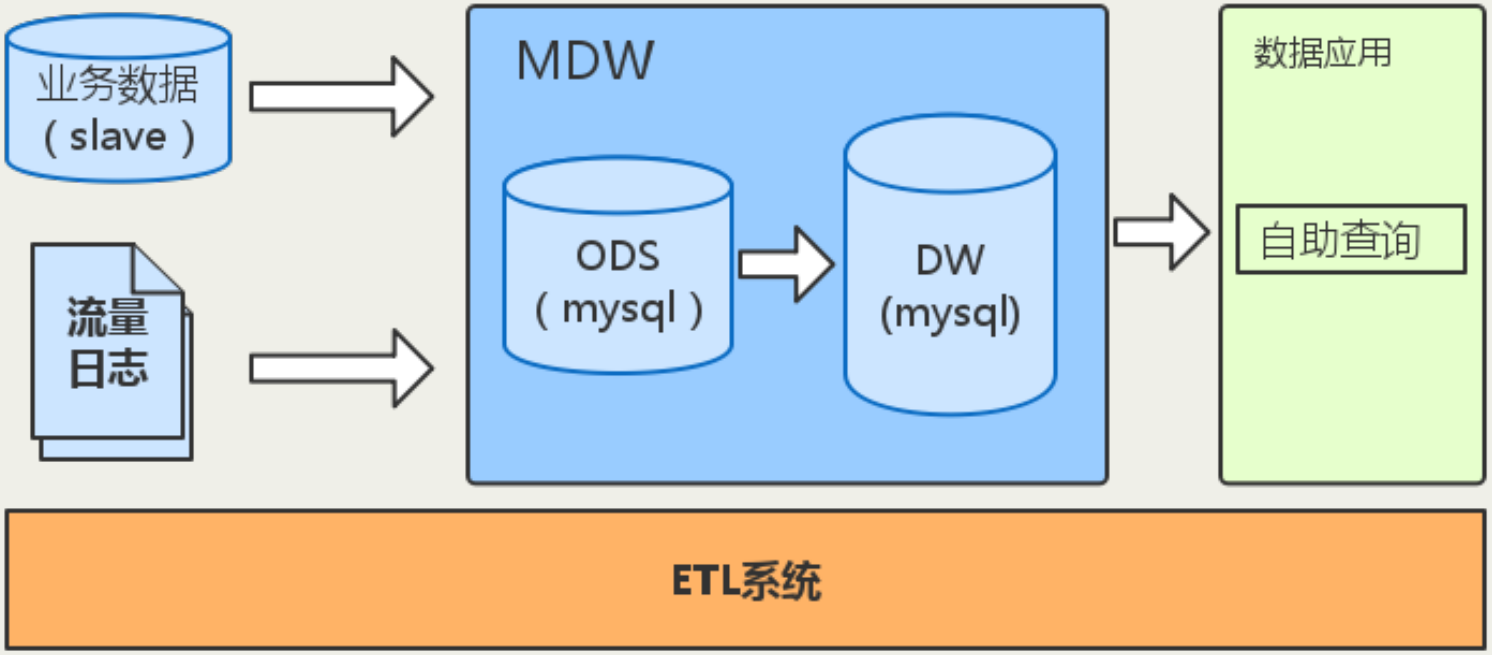
- 数据层面
  - 没有集成
  - 难以复用
- 工具层面
  - 重复开发
  - 管理混乱

效率低！

近代模式

美团网
meituan.com

- 模式：由包办开始转向协作
- 技术：



近代模式问题

美团网 meituan.com

- 数据存储上：
  - 不易扩展
- 数据接入效率低
  - 专属数据RD对接
  - 新业务理解时间长
- 后期维护成本高
  - 人员流动
  - 需求变更

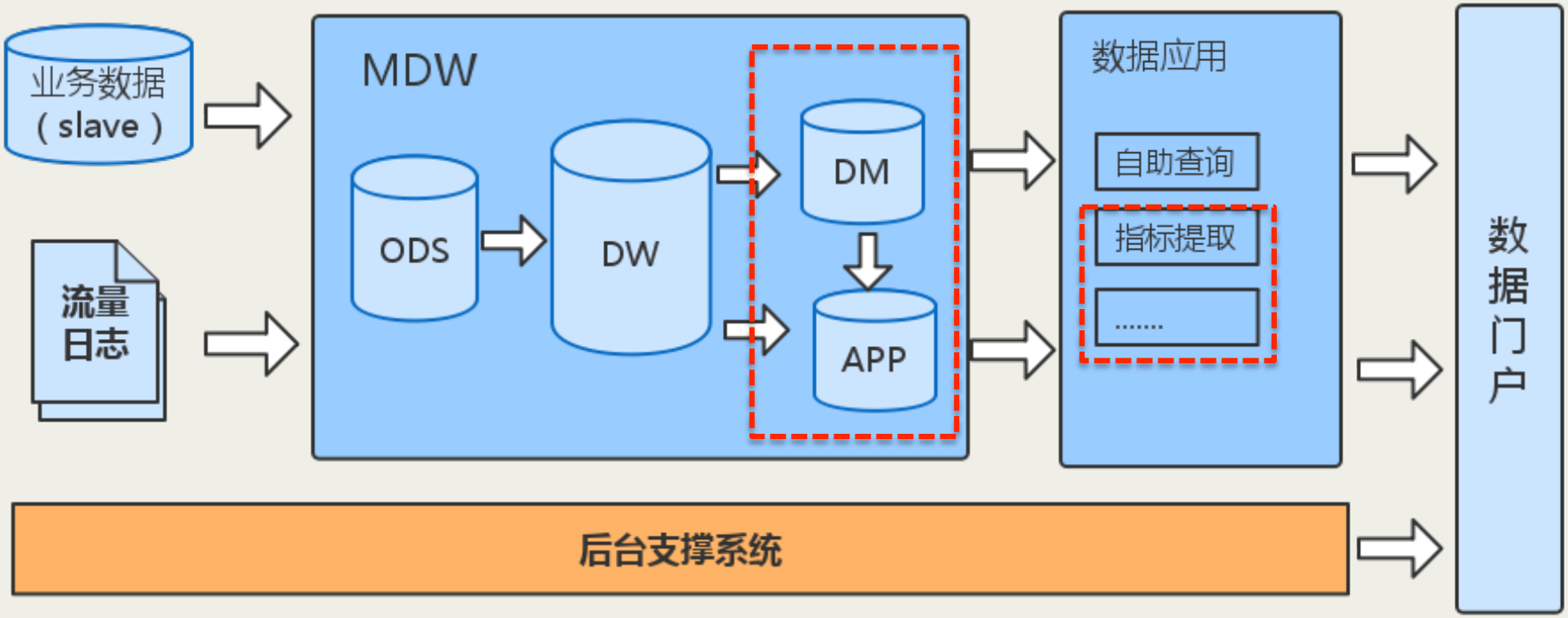
现代模式

- 业务上：数据组主导+业务方协作
  - 业务量：Etl数量由几百增加到1600+

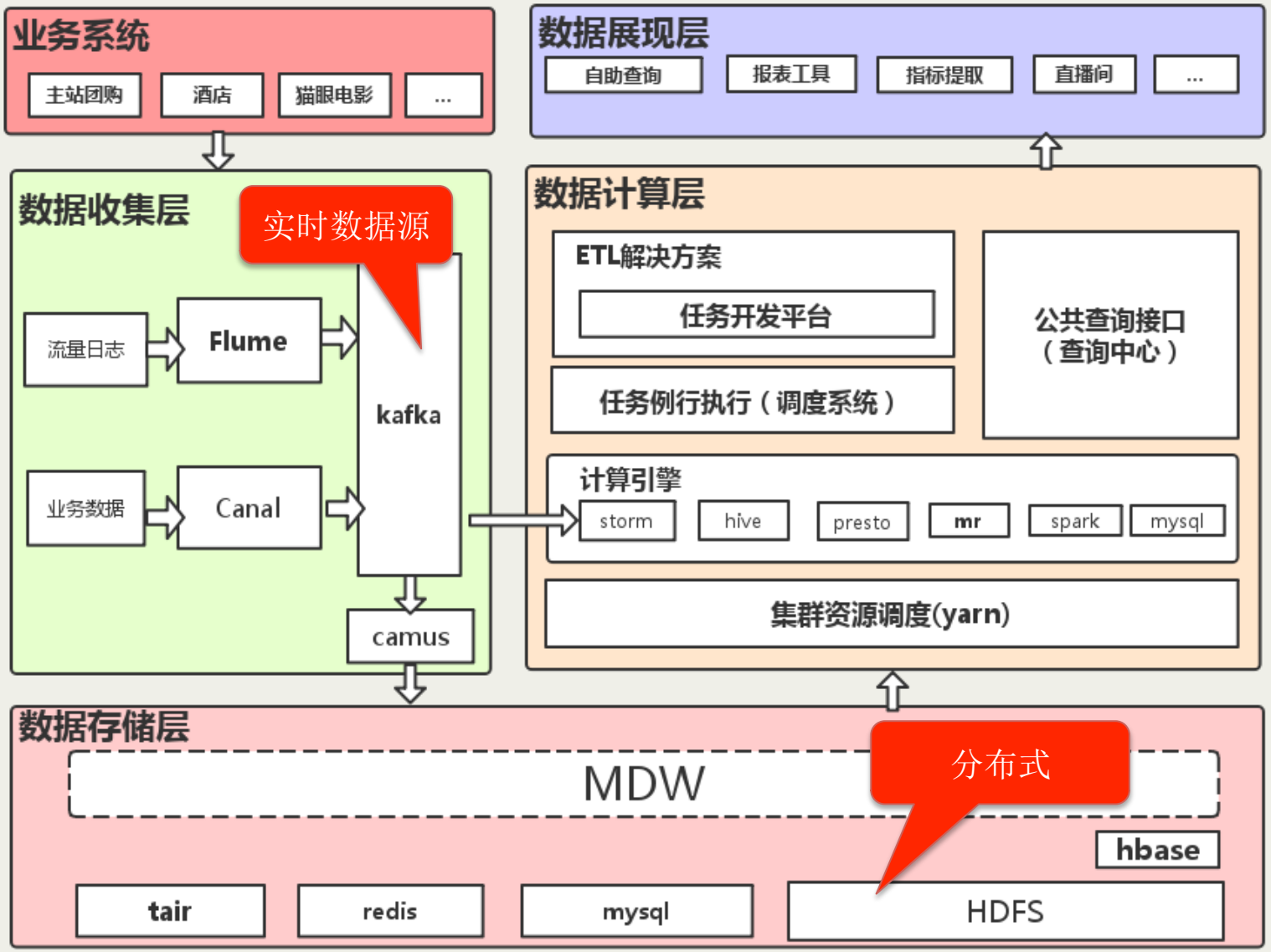
现代模式

美团网 meituan.com

- 数据仓库模型：自上而下



平台架构


业务系统
主站团购
酒店
猫眼电影
...
数据展现层
自助查询
报表工具
指标提取
直播间
...
数据收集层
实时数据源
流量日志
Flume
业务数据
Canal
kafka
camus
数据计算层
ETL解决方案
任务开发平台
任务例行执行（调度系统）
公共查询接口
（查询中心）
计算引擎
storm
hive
presto
mr
spark
mysql
集群资源调度(yarn)
数据存储层
MDW
分布式
hbase
tair
redis
mysql
HDFS

大纲

美团网
meituan.com

- 现状介绍
- 开发模式回顾
- 面临的挑战
- 解决方案
- 思考与总结

业务增长

美团网
meituan.com

酒店

电影

主站
团购

外卖

广告

配送

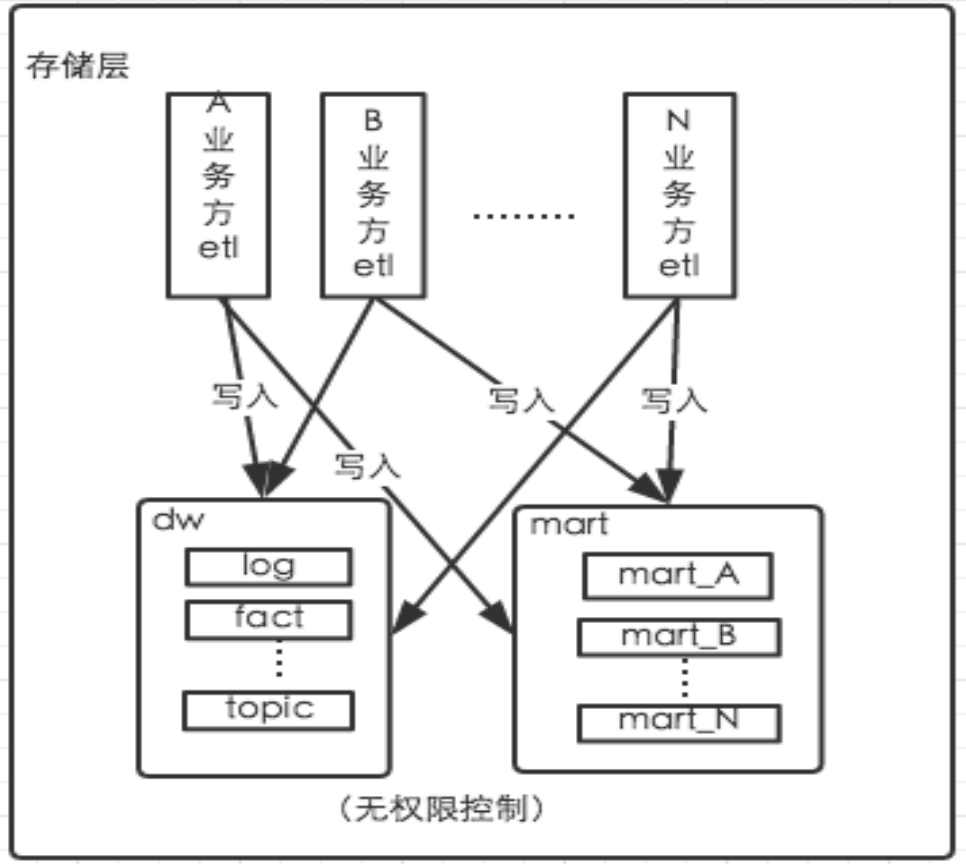
业务增长

美团网 meituan.com

- 带来的影响
  - 仓库要支撑业务场景变多：20+个业务
  - 平台服务的业务RD人数猛增：200+ 半年翻倍
  - ETL任务数：3800+ 成倍增长

面临的挑战

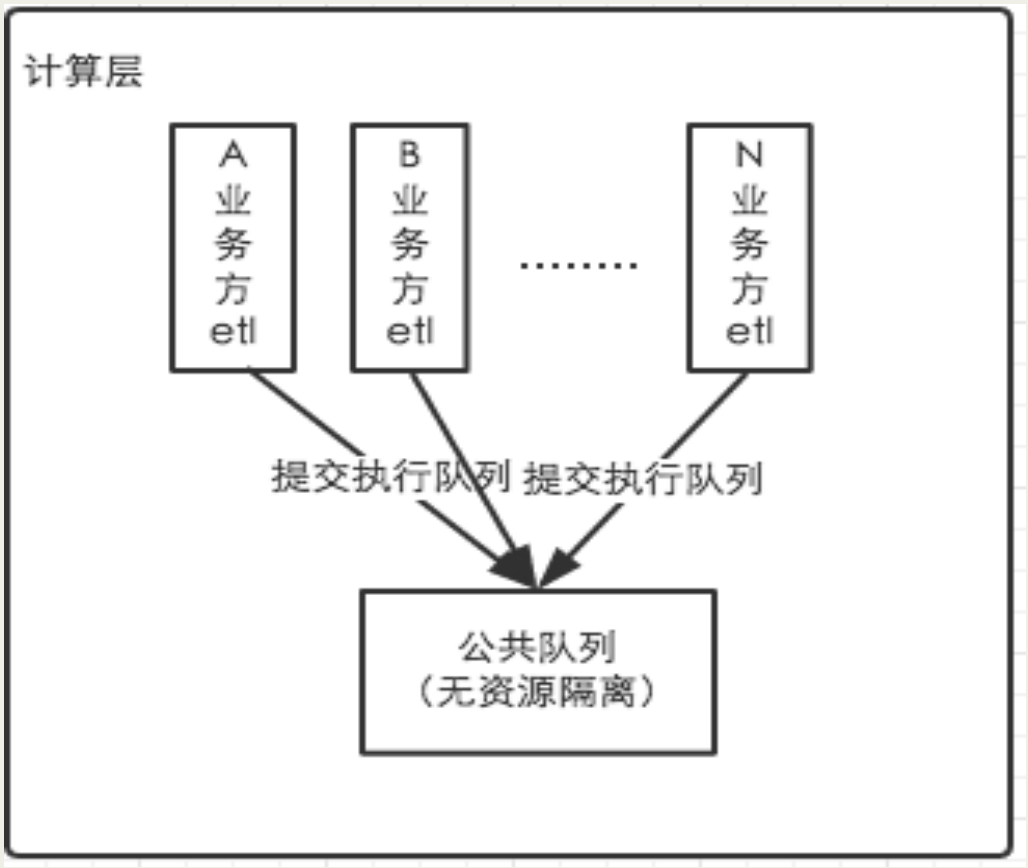
- ETL开发和数据存储管理出现混乱



面临的挑战

美团网
meituan.com

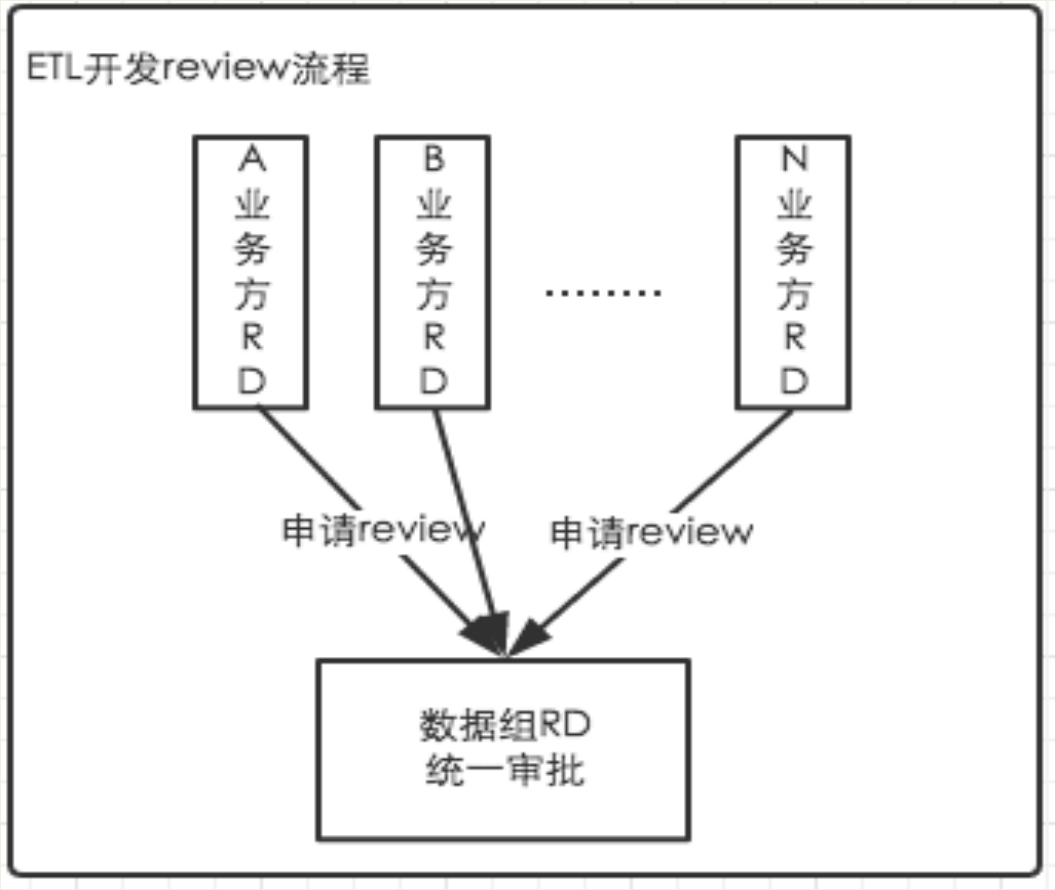
- ETL任务计算相互影响



面临的挑战

美团网
meituan.com

- ETL开发效率不能满足业务增长



面临的挑战

美团网 meituan.com

- 囧境：
  - 平台人力资源消耗严重
    - 任务Review
    - 任务管理
    - 各种对接
  - 各业务方吐槽平台
    - 任务开发效率低
    - 任务执行慢

大纲

美团网
meituan.com

- 现状介绍
- 开发模式回顾
- 面临的挑战
- 解决方案
- 总结与展望

解决方案

美团网 meituan.com

- 思路：
  - 提升效率
  - 安全可控
  - 管理透明
- 权衡：
  - 开放的后果
  - 满足业务需求第一

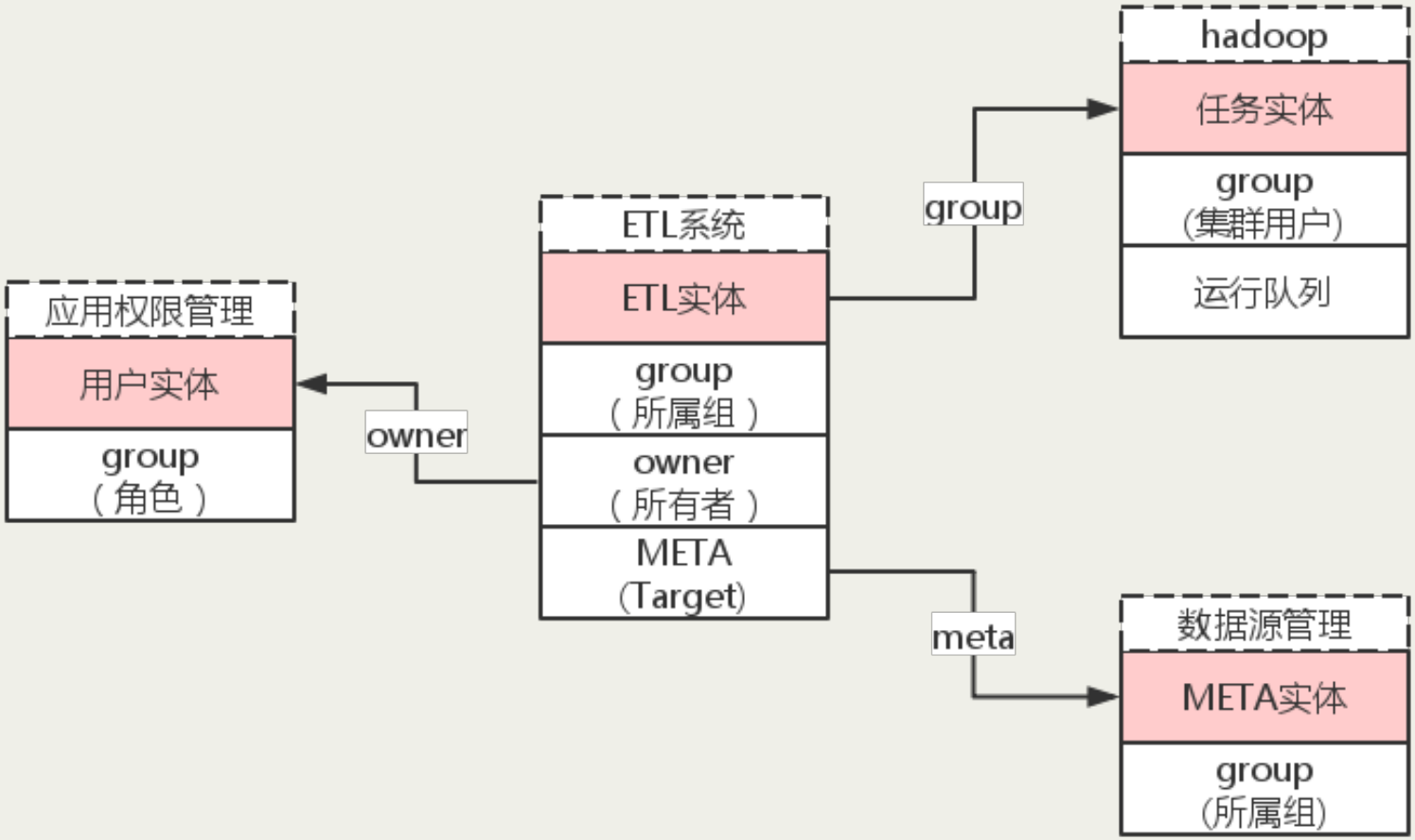
解决方案

美团网 meituan.com

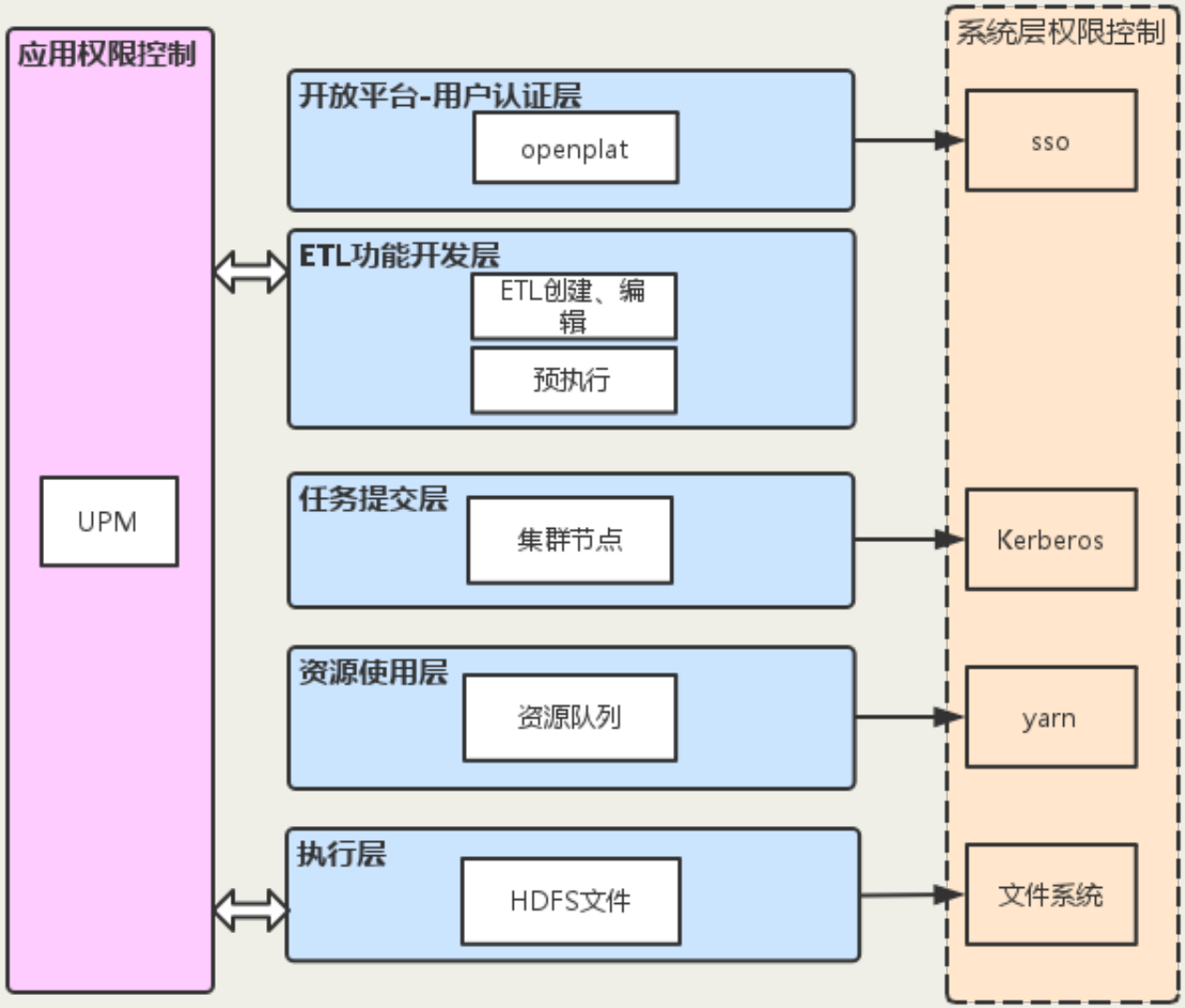
- ETL任务分组
- 资源隔离
- 权限控制
- 开放ETL-Review

ETL分组

- ETL按业务进行分组

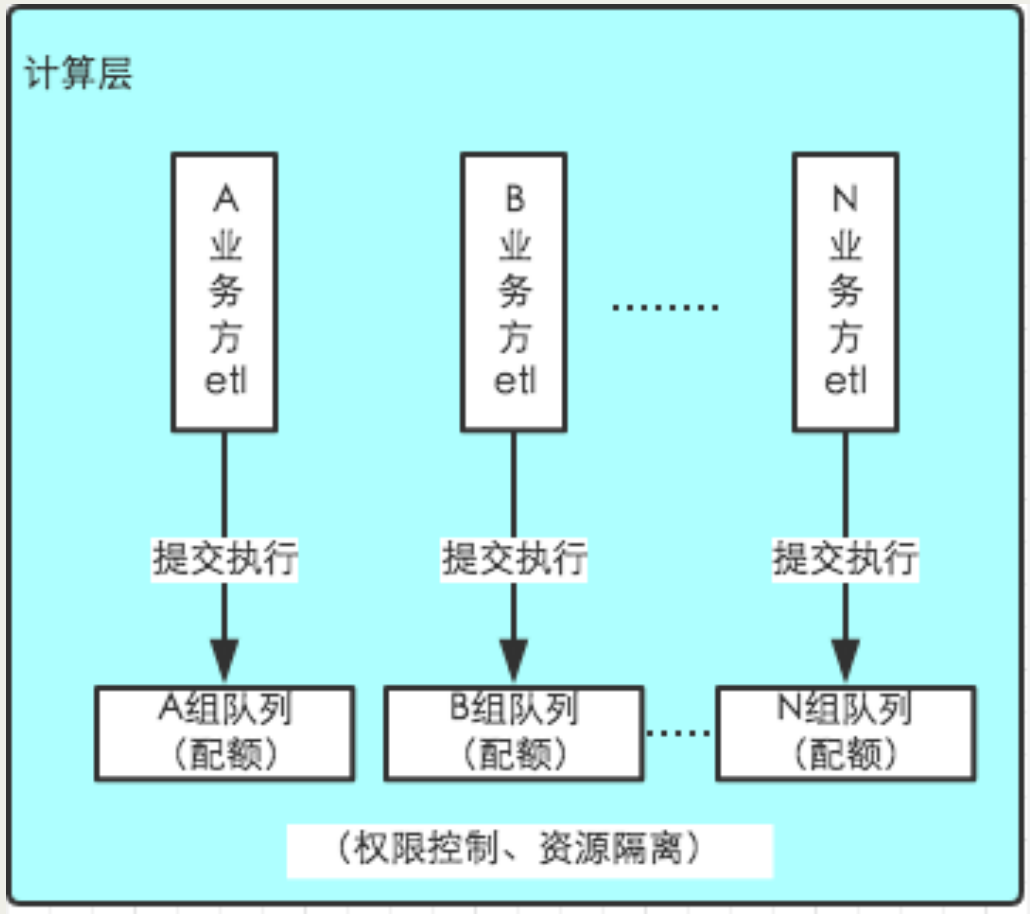


权限模型

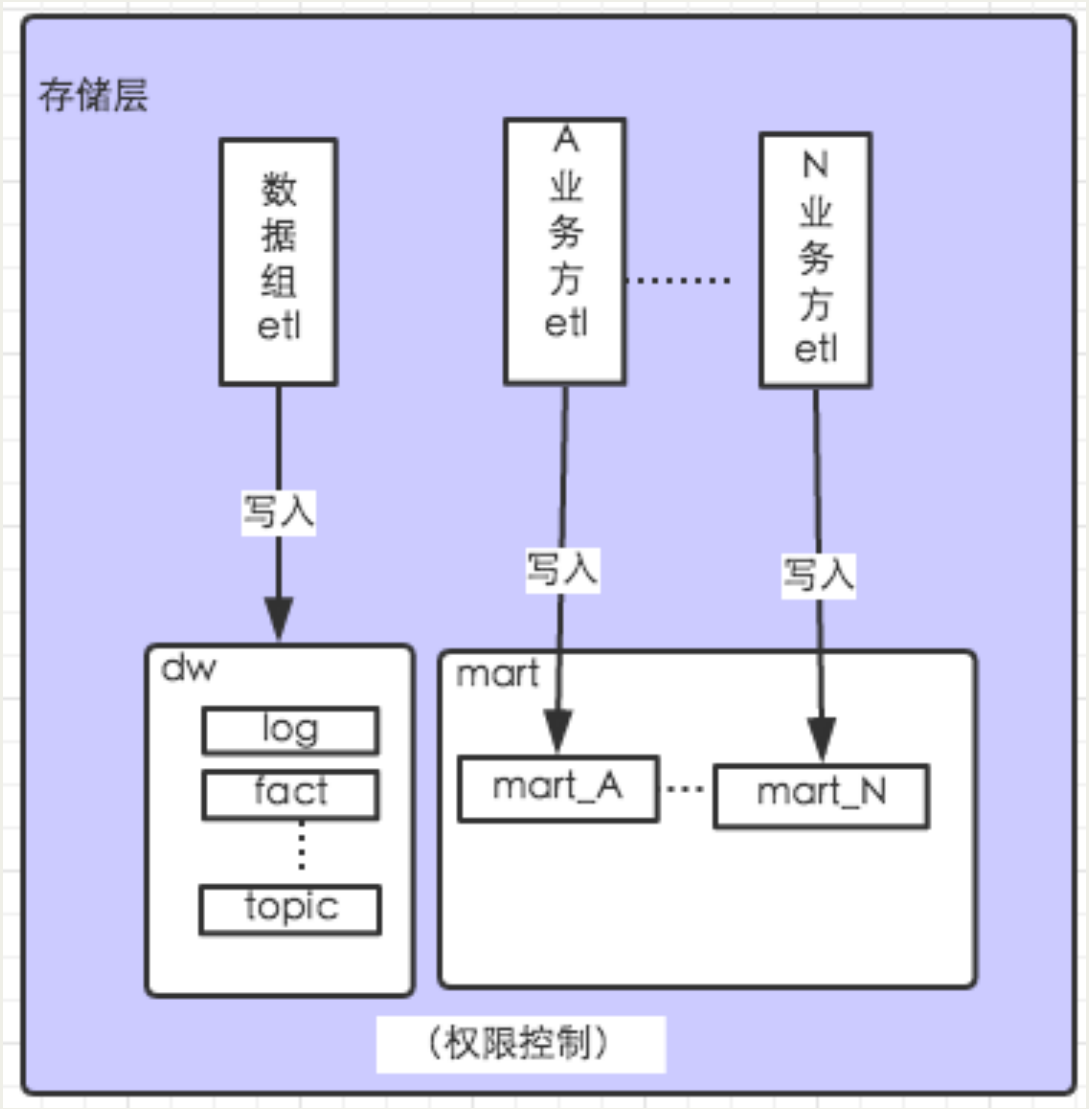

应用权限控制
UPM
开放平台-用户认证层
openplat
ETL功能开发层
ETL创建、编辑
预执行
任务提交层
集群节点
资源使用层
资源队列
执行层
HDFS文件
系统层权限控制
sso
Kerberos
yarn
文件系统

资源隔离

- 对ETL的计算资源按组进行分队列

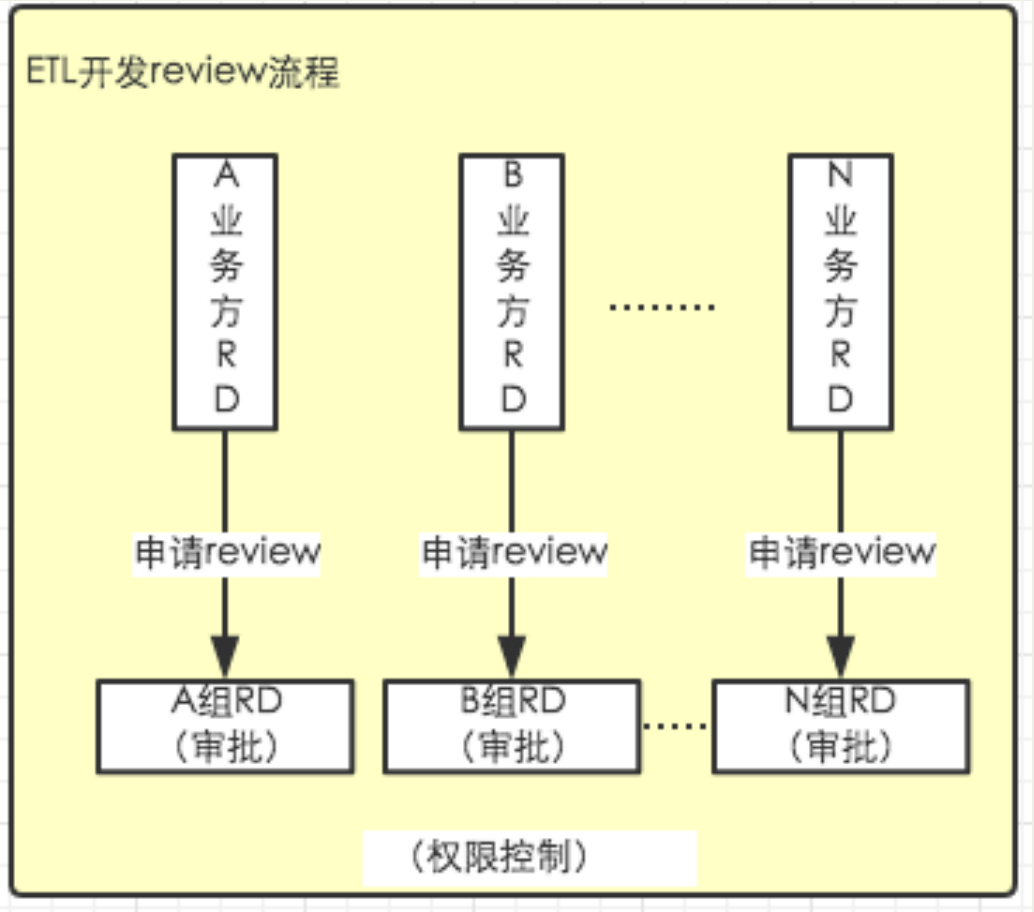


权限控制

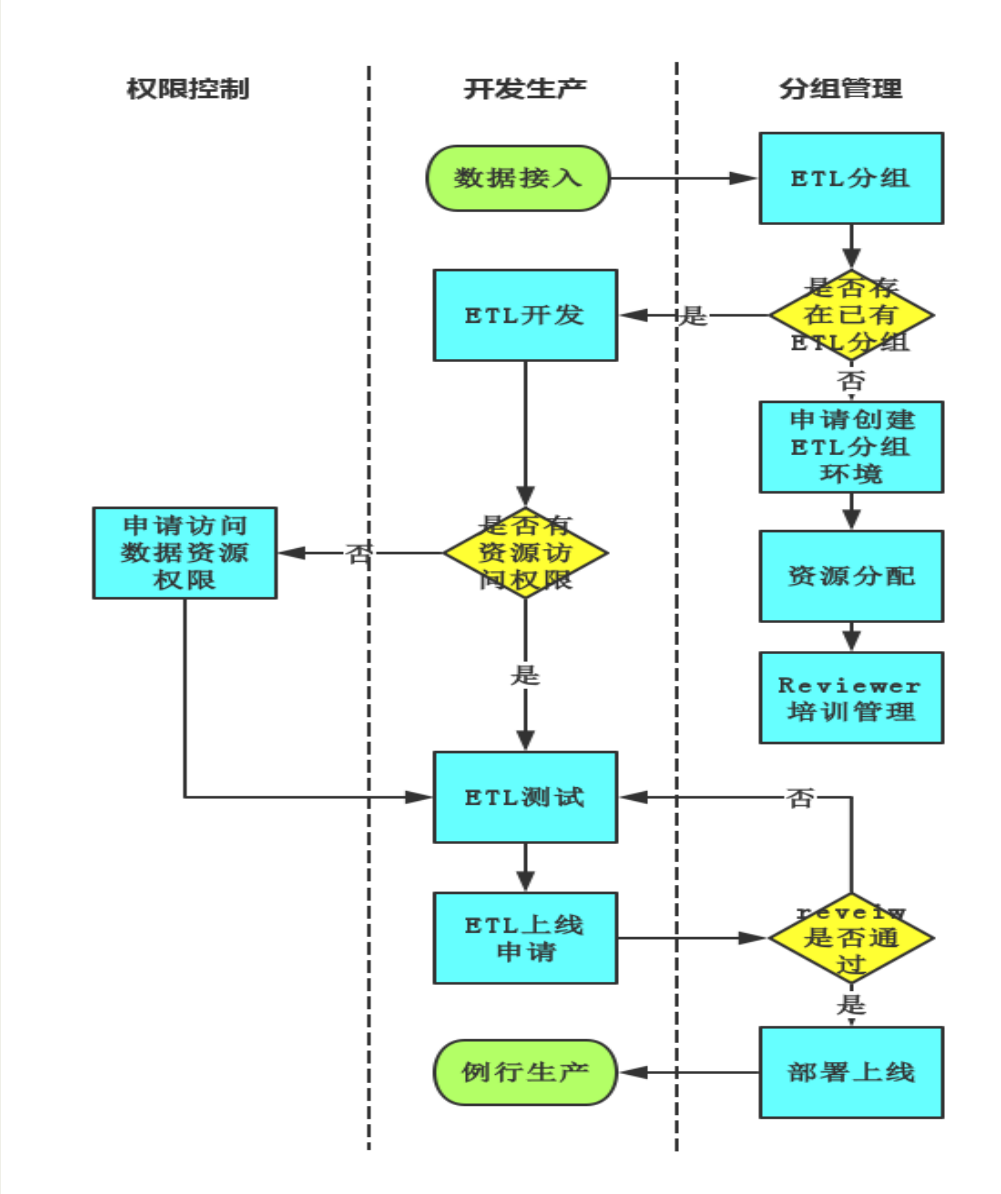

存储层
数据组etl
A业务方etl
N业务方etl
写入
写入
写入
dw
log
fact
topic
mart
mart_A
mart_N
（权限控制）

开放ETL-Reivew

- 对各组Reviewer培训
  - 传承仓库建设经验
- 开发平台保障机制
  - 保证仓库建设基本规范



开发流程图

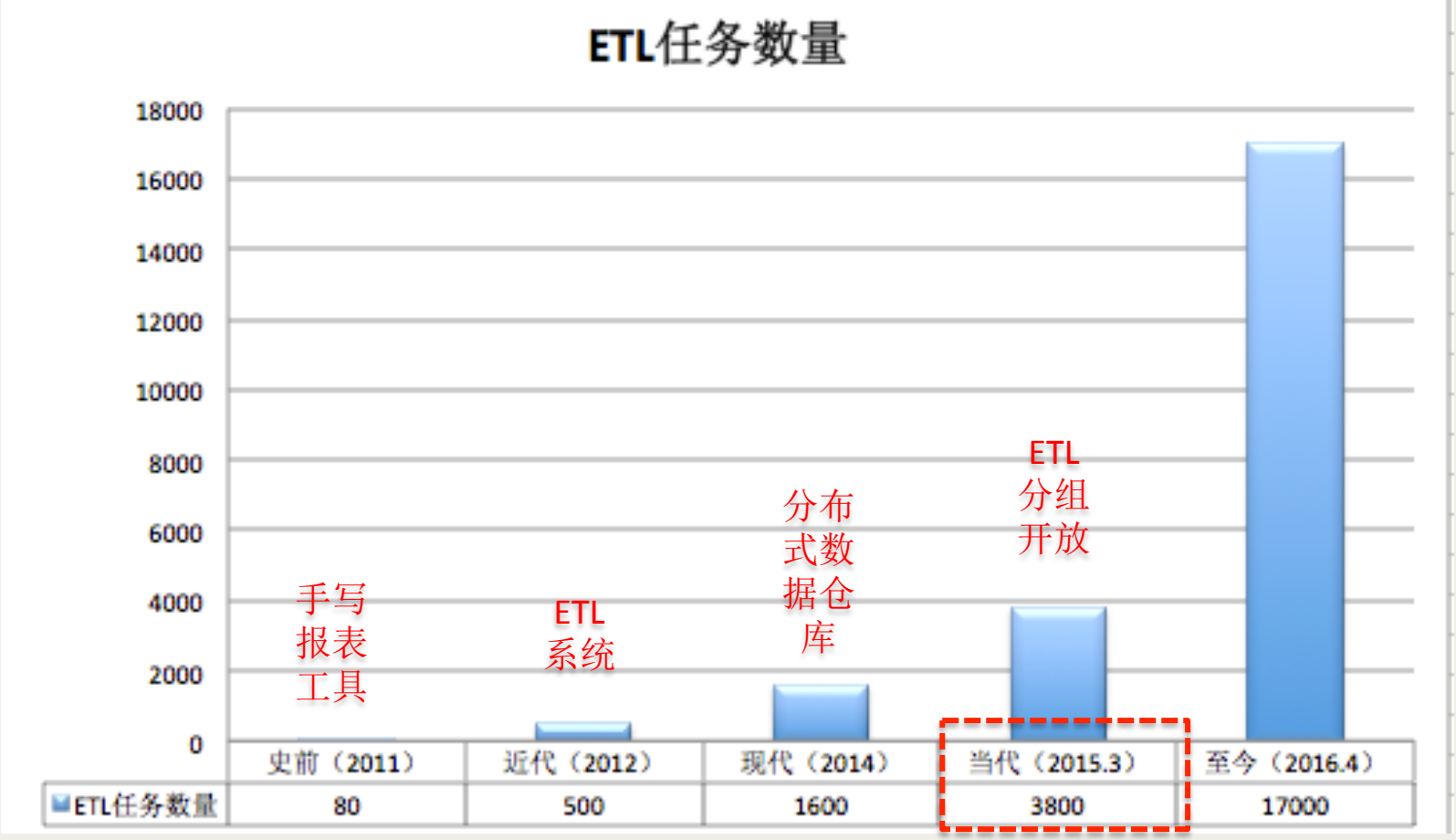

权限控制
开发生产
分组管理
数据接入
ETL分组
是否存在已有ETL分组
是
ETL开发
否
申请创建ETL分组环境
资源分配
Reviewer培训管理
是否有资源访问权限
否
申请访问数据资源权限
是
ETL测试
否
ETL上线申请
是否通过
是
例行生产
部署上线

ETL开发示例


ETL 名称
hmart_waimai.order
TARGET_META名称
hmart_waimai
ETL 模板
一般ETL模板
##-- 这个是sqlweaver(美团自主研发的ETL工具)的编辑模板
##-- 本模板内容均以 ##-- 开始,完成编辑后请删除
##-- ##xxxx## 型的是ETL专属文档节点标志, 每个节点标志到下一个节点标志为本节点内容
##-- 流程应该命名成: 目标表meta名(库名).表名
##Description##
##-- 这个节点填写本ETL的描述信息, 包括目标表定义, 建立时的需求jira编号等
##TaskInfo##
creator = 'AnonymousUser@meituan.com'
source = {
'db': META[''], ##-- 这里的单引号内填写在哪个数据库链接执行 Extract阶段, 具体有哪些链接请点击"查看META"按钮查看
}
stream = {
'format': '', ##-- 这里的单引号中填写目标表的列名, 以逗号分割, 按照Extract节点的结果顺序做对应, 特殊情况Extract的列数可以小于目标表列数
}
target = {
'db': META[''], ##-- 单引号中填写目标表所在库
'table': '', ##-- 单引号中填写目标表名
}
##Extract##
##-- Extract节点, 这里填写一个能在source.db上执行的sql
##Preload##
##-- Preload节点, 这里填写一个在load到目标表之前target.db上执行的sql(可以留空)
##Load##
##-- Load节点, (可以留空)
##TargetDDL##
##-- 目标表表结构
CREATE TABLE IF NOT EXISTS `$target.table`
(
) ENGINE=MyISAM DEFAULT CHARSET=utf8 COMMENT=''
ETL负责人
songhongxin
ETL分组
dep-waimai
保存
以
作为参数测试
提交review
查看META

收益

- 保证了ETL随业务的同比增长趋势



收益

美团网 meituan.com

- 在Review请求数量增加1.5倍的情况下，响应时间降低了40%
- ETL-Review开放节约了数据组RD30%左右的时间,能投入到平台建设工作中
- 完成按业务分组(40+),确定了各组对接人和资源配额，规范了数据仓库建设

大纲

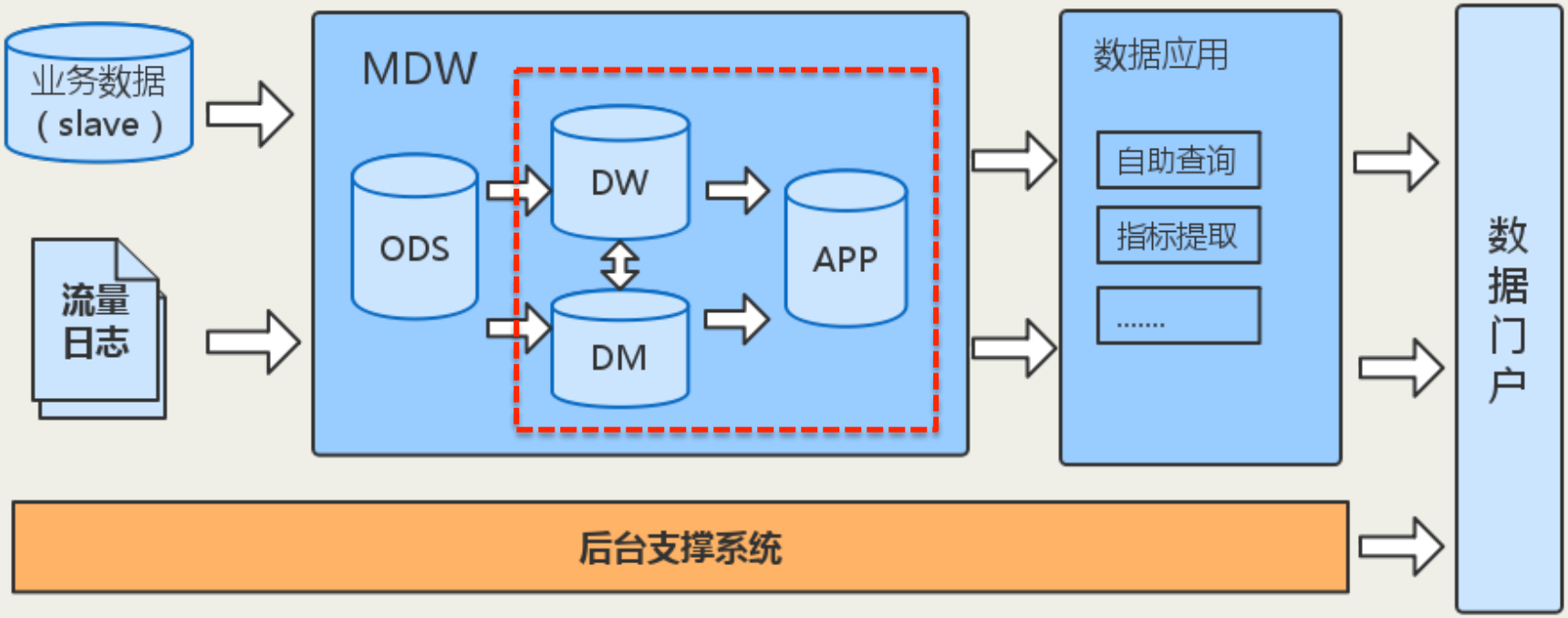
美团网
meituan.com

- 现状介绍
- 开发模式回顾
- 面临的挑战
- 解决方案
- 总结与展望

当代模式

美团网 meituan.com

- 业务上：由主导+协作变成业务自治
- 仓库建设：由自上而下到自下而上结合



经验总结

美团网 meituan.com

- 注重效率：提高数据接入、ETL开发，上线、生产效率
- 平台开放：仓库规范和数据安全可控下、足够开放
- 降低成本：保证资源合理分配，降低数据管理成本
- 技术与时俱进：拥抱开源，关注最新技术，选择适合业务场景
- 勇敢决策：果断迭代仓库开发模式，适应业务发展需求，并及时回顾、调整

展望

美团网 meituan.com

- 加强数据治理：
  - 自下而上带来的数据整合问题
  - 存储和计算资源浪费
  - 数据质量问题不受重视
- 建设好协同开发平台：
  - 业务变更频繁导致分组管理问题
    - 任务、资源、权限、负责人
  - 工具链环节过多，使用成本高
    - 数据源接入、任务开发、生产、运维

欢迎加入

美团网
meituan.com

美团点评技术团队

宋洪鑫

songhongxin@meituan.com

THANKS

SequeMedia 盛拓传媒 IT168.com ChinaUnix ITPUB