美团大数据平台架构

演进过程与最新进展

xieyuchen@meituan.com

关于我

美团网 meituan.com

- 谢语宸

  11年 加入美团, 统计报表与数据仓库

  12年 数据仓库分布式化

  13年 数据开发开放平台

  14年 负责离线计算平台团队

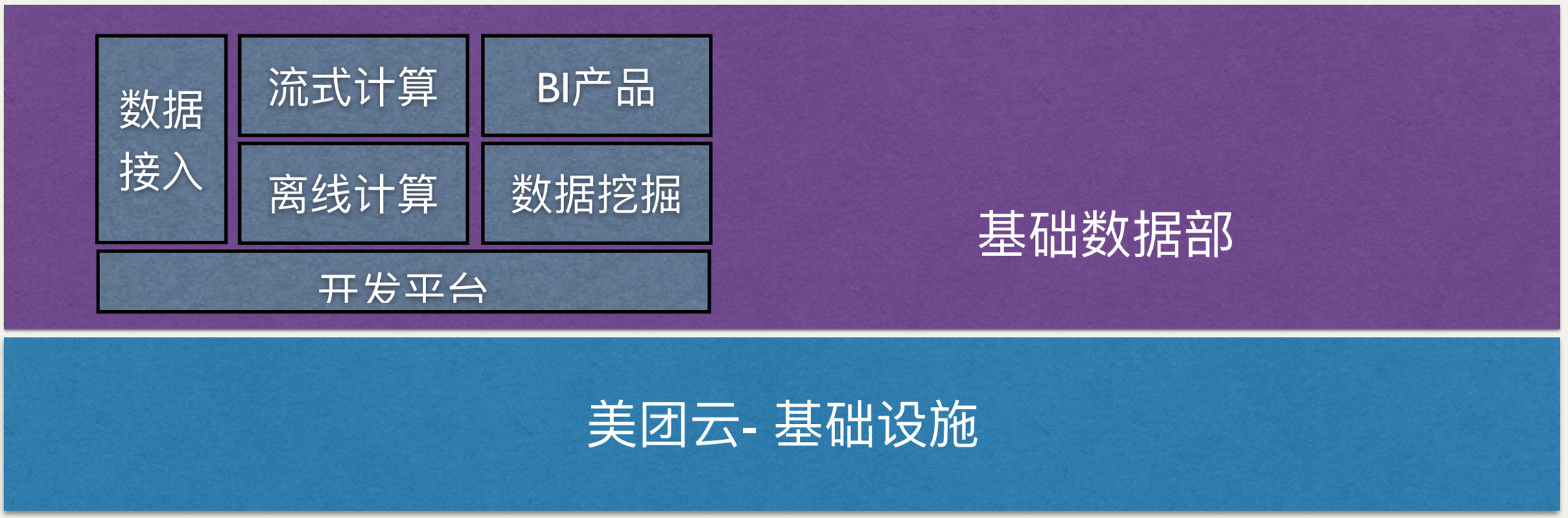
目录

- 美团大数据平台架构
- 平台演进时间线
- 近期挑战与应对
- 平台化思路总结

数据体系组织架构

美团网 meituan.com

到店餐饮	到店综合	酒店旅游	猫眼电影	外卖配送	广告平台	…
分析师	分析师	分析师	分析师	分析师	分析师	
数据	数据	数据	数据	数据	数据	



美团数据流架构图

美团网
meituan.com

数据接入

流式计算

离线计算

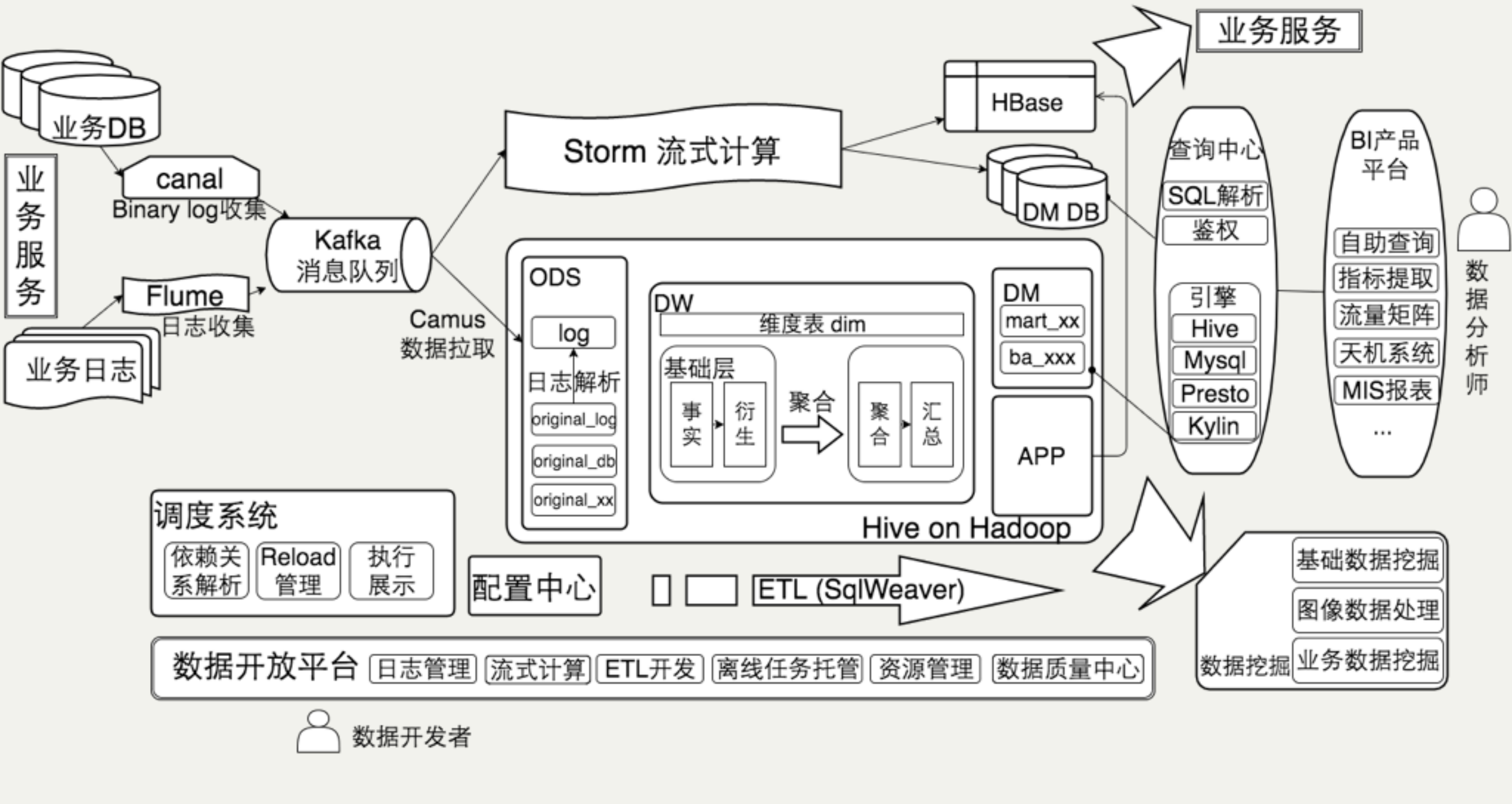
BI产品

数据挖掘

开发平台

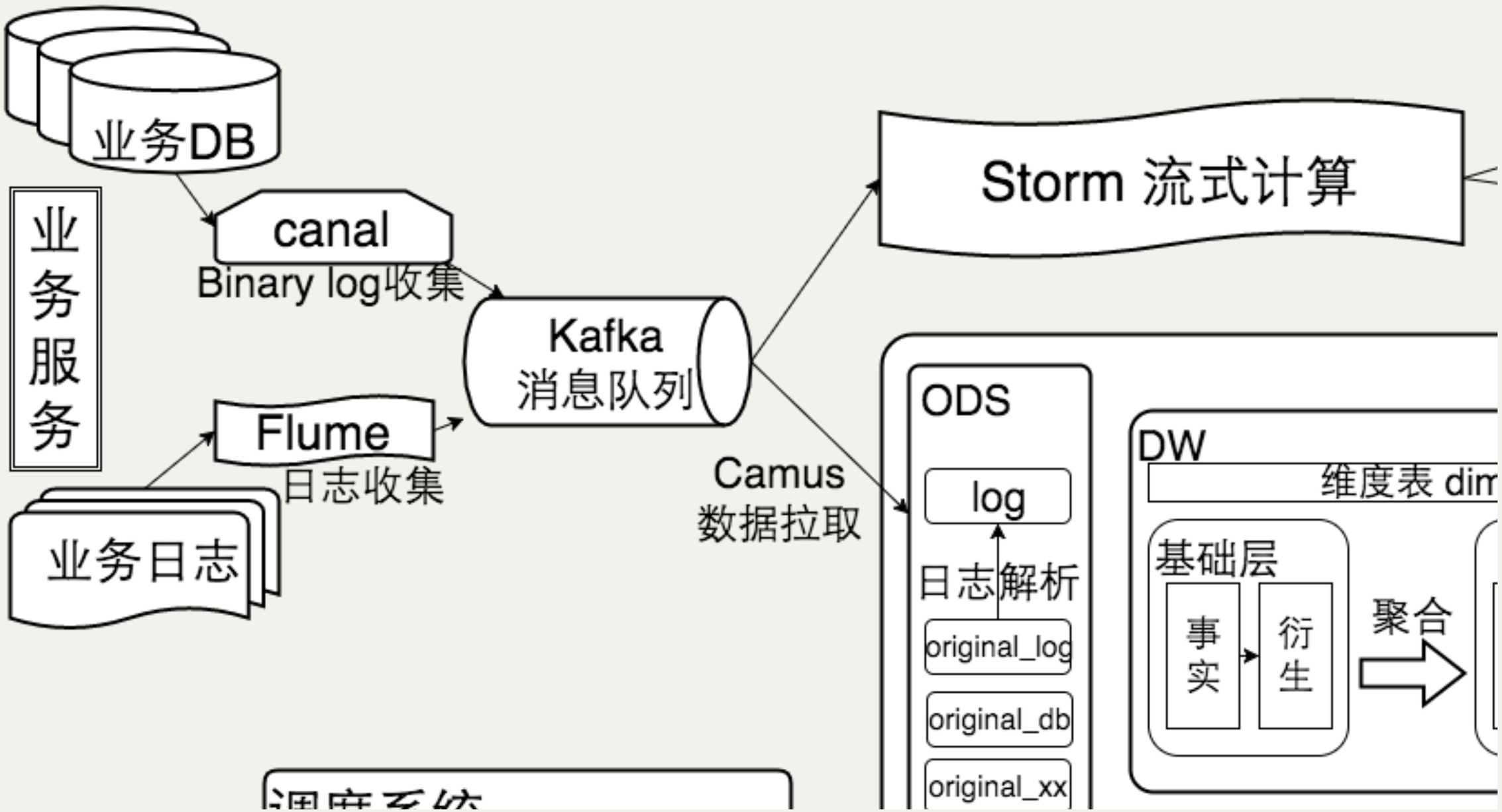
美团数据流架构图

美团网
meituan.com



数据接入与流式计算

美团网
meituan.com



数据收集特性

美团网 meituan.com

- 日志型数据多接口支持
- 关系型数据基于Binlog获取增量
- 消息队列集中化分发支持多下游

- 850+ 日志数
- 百万+ 峰值每秒消息接入

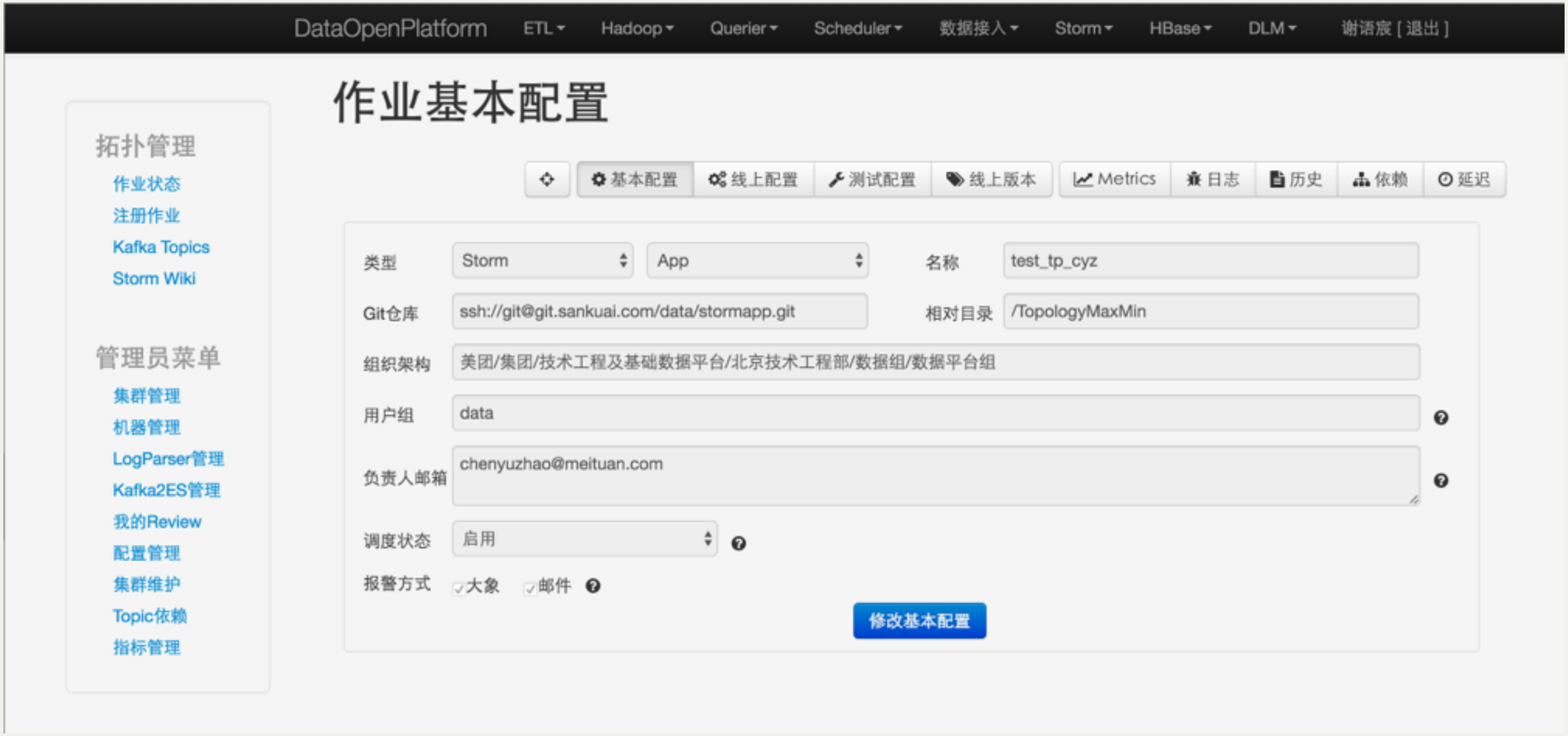
流式计算平台特性

- 测试开发平台化
- 拓扑开发框架
- 延迟统计与报警
- 拓扑间依赖关系解析

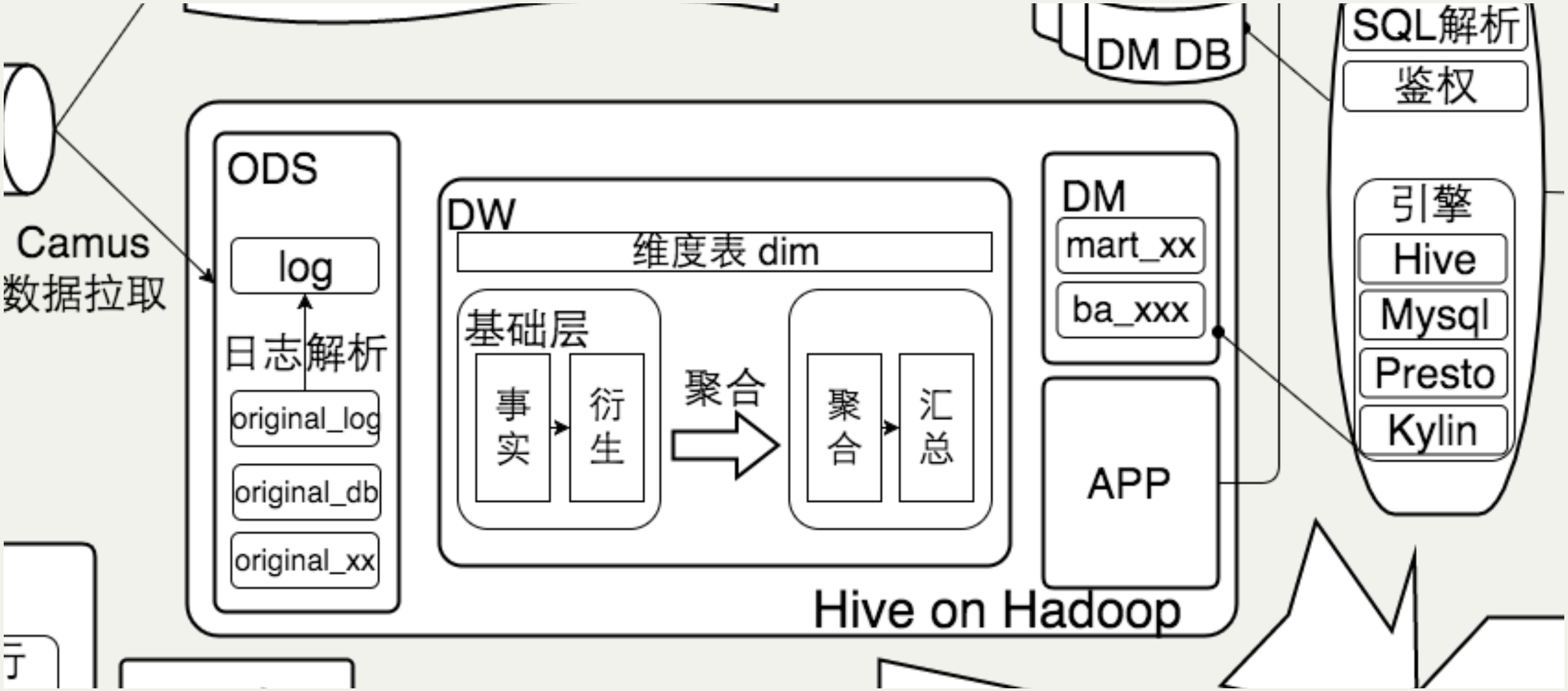
- 1100+ 实时拓扑
- 秒级别实时数据流延迟

流式计算平台

美团网
meituan.com

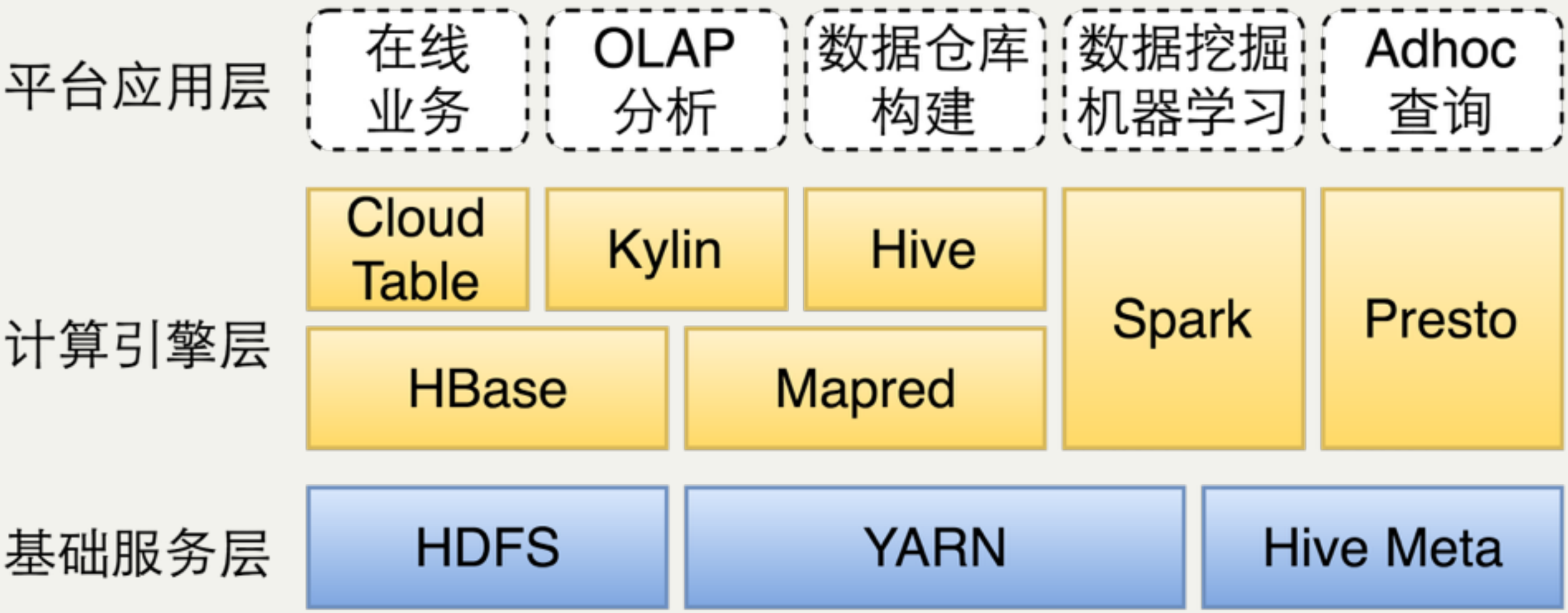


离线计算


DM DB
SQL解析
鉴权
ODS
log
日志解析
original_log
original_db
original_xx
DW
维度表 dim
基础层
事实
衍生
聚合
聚合
汇总
DM
mart_xx
ba_xxx
APP
引擎
Hive
Mysql
Presto
Kylin
Camus
数据拉取
Hive on Hadoop

离线计算部署架构

美团网 meituan.com



离线计算平台特性

- 高可用, 高可扩展
- 多计算框架支持
- 数据仓库开发模板

- 42P+ 总存储量
- 150K /天任务数
- 2500+ 节点, 3机房统一名字空间
- 16K 数据仓库数据表数

数据仓库开发模板

hmart_movie.detail_order_deal_info

版本列表 历史Review 依赖关系 调度配置 执行日志

```
##Description##
##-- 美团电影交易基础信息表（团购和选座）

##TaskInfo##
creator = 'lipeng10@meituan.com'

source = {
    'db': META['horigin_mobile'], ##-- 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':
'dt,order_id,deal_id,deal_time,use_time,come,market_city,business_id,business_name,main_poiid,main_poiname,cinema_id,cinema_name,cinema_city,bd_name,deal_price,closing_price,cost_price,cost_voucher,deal_ticket,user_id,is_new,is_real_new,activity_id,my_activity', ##-- 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['hmart_movie'], ##-- 单引号中填写目标表所在库
    'table': 'detail_order_deal_info', ##-- 单引号中填写目标表名
}

##Load##
#if $isRELOAD
set hive.exec.dynamic.partition.mode=nonstrict;
set hive.exec.dynamic.partition=true;

set hive.exec.max.dynamic.partitions=10000;
set hive.exec.max.dynamic.partitions.pernode=1000;
set hive.exec.reducers.max =1000;

set hive.exec.max.created.files=10000;
set hive.merge.mapfiles=true;
#end if

INSERT OVERWRITE TABLE `$target.table`
#if $isRELOAD
    PARTITION(dt)
#else
    PARTITION(dt = '$now.datekey')
#end if
select
    order_id,
    deal_id,
    buy_time deal_time,
    come,
```

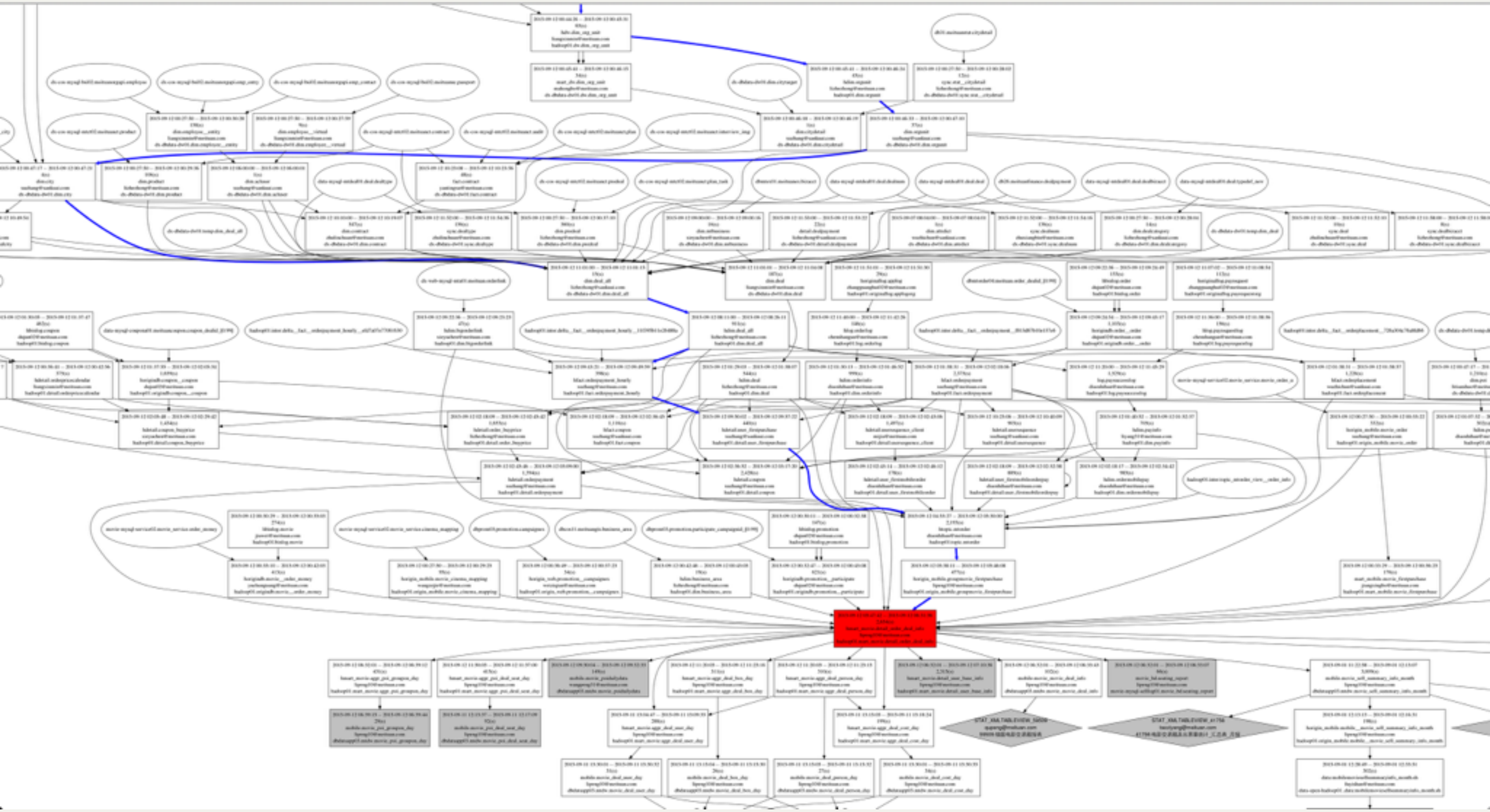
数据管理体系


original_xx
Hive on Hadoop
调度系统
依赖关系解析
Reload管理
执行展示
配置中心
ETL (SqlWeaver)
数据开放平台
日志管理
流式计算
ETL开发
离线任务托管
资源管理
数据质量中心
数据开发者

数据管理体系特性

- 自动依赖关系识别
- 业务线成本核算
- 任务SLA保障
- 数据质量监控

数据管理



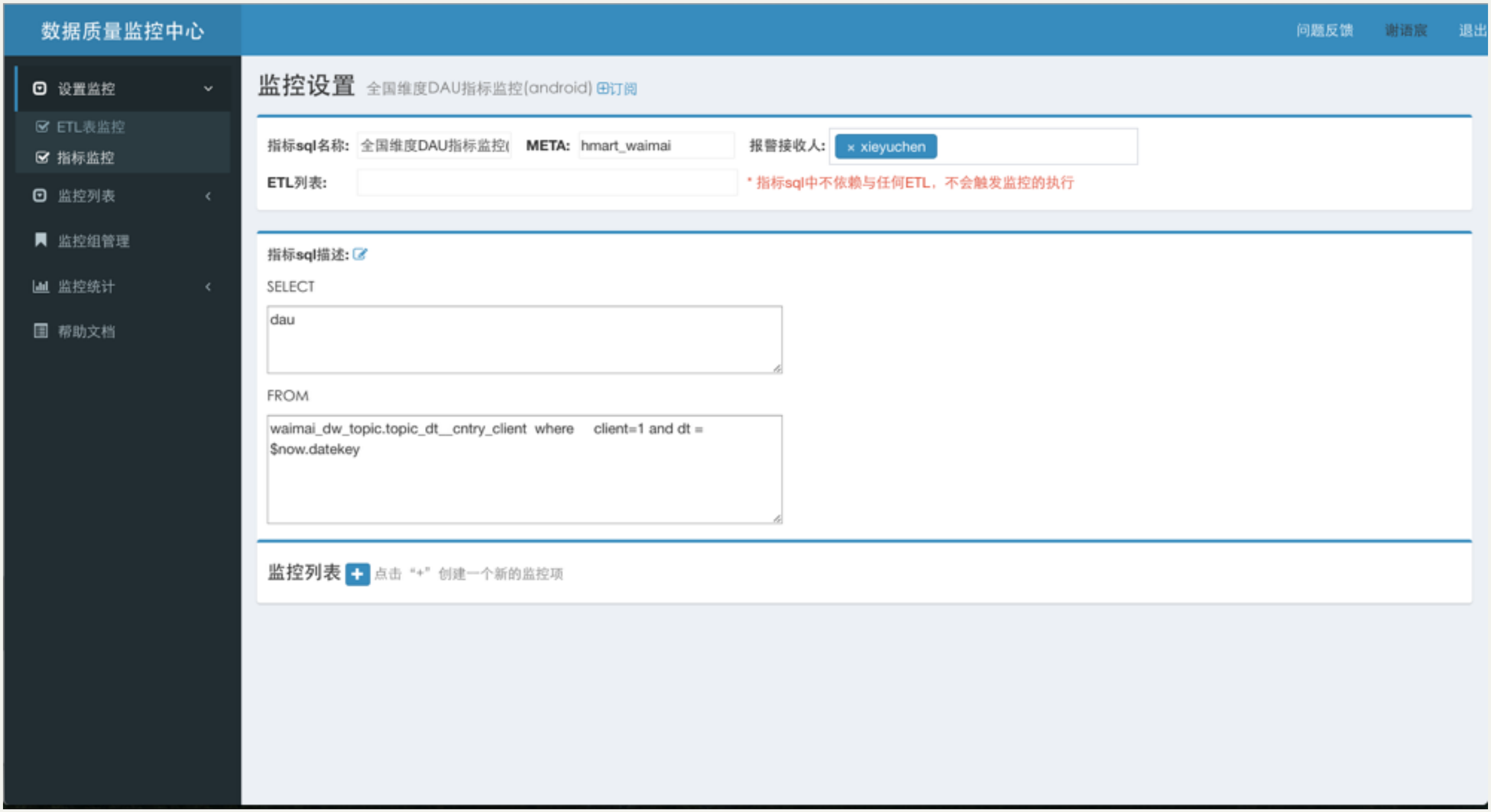
数据管理

美团网 meituan.com

事业群	内存配置量(G)	内存配置占比	CPU配置量	CPU配置占比	内存使用量(G)	CPU使用量
金融发展部	801	0.35%	356	0.35%	890	290
企业平台研发部	1109	0.49%	493	0.49%	202	89
支付平台部	2256	0.99%	1003	0.99%	4390	1428

数据管理

美团网 meituan.com



数据管理

美团网 meituan.com

搜索广告部数据组 数据仓库SLA日报 20160420

产出时间颜色说明
红色表示晚于期望产出时间(即打破SLA)
黄色表示有时效性风险(即实际产出时间距期望产出时间不足2小时,可能因为平台运维而打破SLA)

摘要:

sla名称	按时产出对标表数/总对标表数	期望产出时间	实际产出时间	环比	7日均值	与均值比
BI-北京广告-广告ABTEST数据	5/5	9:30:00	5:56:41	+5分24秒	6:04:06	-7分25秒
BI-北京广告-广告效果数据	1/1	7:30:00	6:44:01	+8分46秒	6:52:24	-8分23秒
BI-北京广告-广告点击用户行为数据	1/1	7:00:00	6:38:45	+9分1秒	6:35:41	+3分4秒
BI-北京广告-广告请求点击下单转化	3/3	8:00:00	5:25:29	+1分55秒	5:31:22	-5分53秒
BI-北京广告-搜索筛选数据	5/5	7:30:00	4:43:24	-8分7秒	4:53:25	-10分1秒
BI-北京广告-用户userid_uuid映射数据	1/1	7:00:00	6:15:24	+9分8秒	6:07:44	+7分40秒

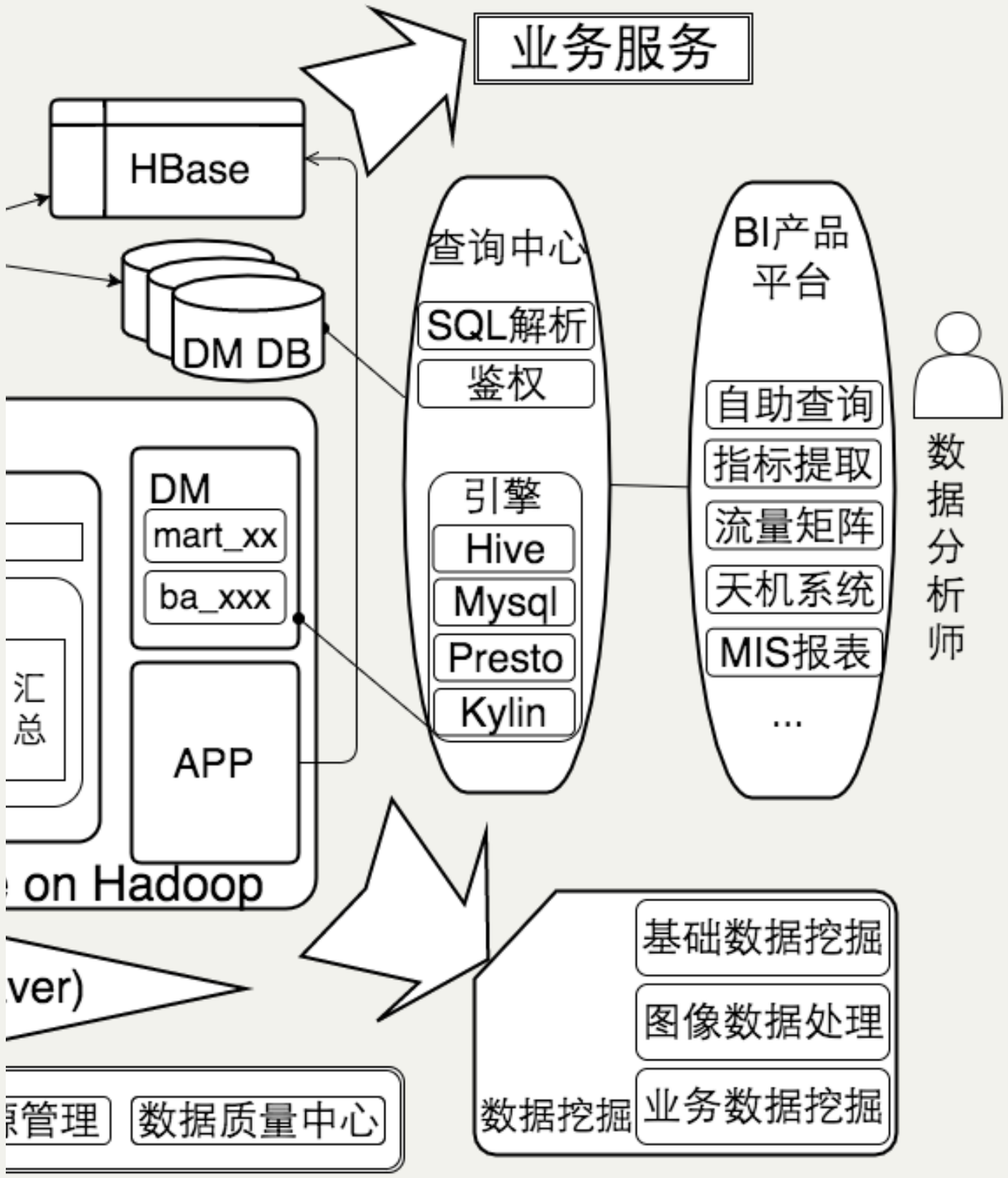
异常表:

sla名称	表名	期望产出时间	调度产出时间	任务成功

存在时效性风险(期望时间-产出时间<2h)的表:

sla名称	表名	期望产出时间	调度产出时间	任务成功
BI-北京广告-广告效果数据	hmart_ads.mtdm_cpm_cpc_daily_effect_detail [火星]	7:30:00	6:44:01	是
BI-北京广告-广告点击用户行为数据	hmart_ads.mtdm_cpm_cpc_daily_traffic_detail [火星]	7:00:00	6:38:45	是
BI-北京广告-用户userid_uuid映射数据	staging.userid_uuid_mapping [火星]	7:00:00	6:15:24	是

BI产品


业务服务
HBase
DM DB
DM
mart_xx
ba_xxx
APP
on Hadoop
ver)
查询中心
SQL解析
鉴权
引擎
Hive
Mysql
Presto
Kylin
BI产品
平台
自助查询
指标提取
流量矩阵
天机系统
MIS报表
...
数据分析师
基础数据挖掘
图像数据处理
业务数据挖掘
数据挖掘
数据质量中心

BI产品 - 指标提取

美团网 meituan.com

指标提取工具 | 查询列表 | 新建查询 | 编辑查询 | BUG/建议/反馈 | 功能/使用说明 | 指标字典

指标 维度

分类

美团

交易

全部

在结果中搜索

关键词

只显示可选

美团订单数

美团购买用户数

美团交易额

美团毛利率

美团毛收入

美团券数

美团首次购买用户数

美团税前消费收入

美团下单交易额

查询 保存

查询状态 已完成(运行时间 3 秒 [2015-08-11 19:20:57 ~ 2015-08-11 19:21:00])

行 平台(二) ✖ 指标 ✖

列 日 ✖ 城市 ✖

指标 美团订单数 ✖ 美团交易额 ✖ 美团毛利率 ✖ 美团毛收入 ✖

日期 (今天是: 8月11日)

2015-08-01 ~ 2015-08-31

日

搜索

全选

20150801

20150802

20150803

20150804

20150805

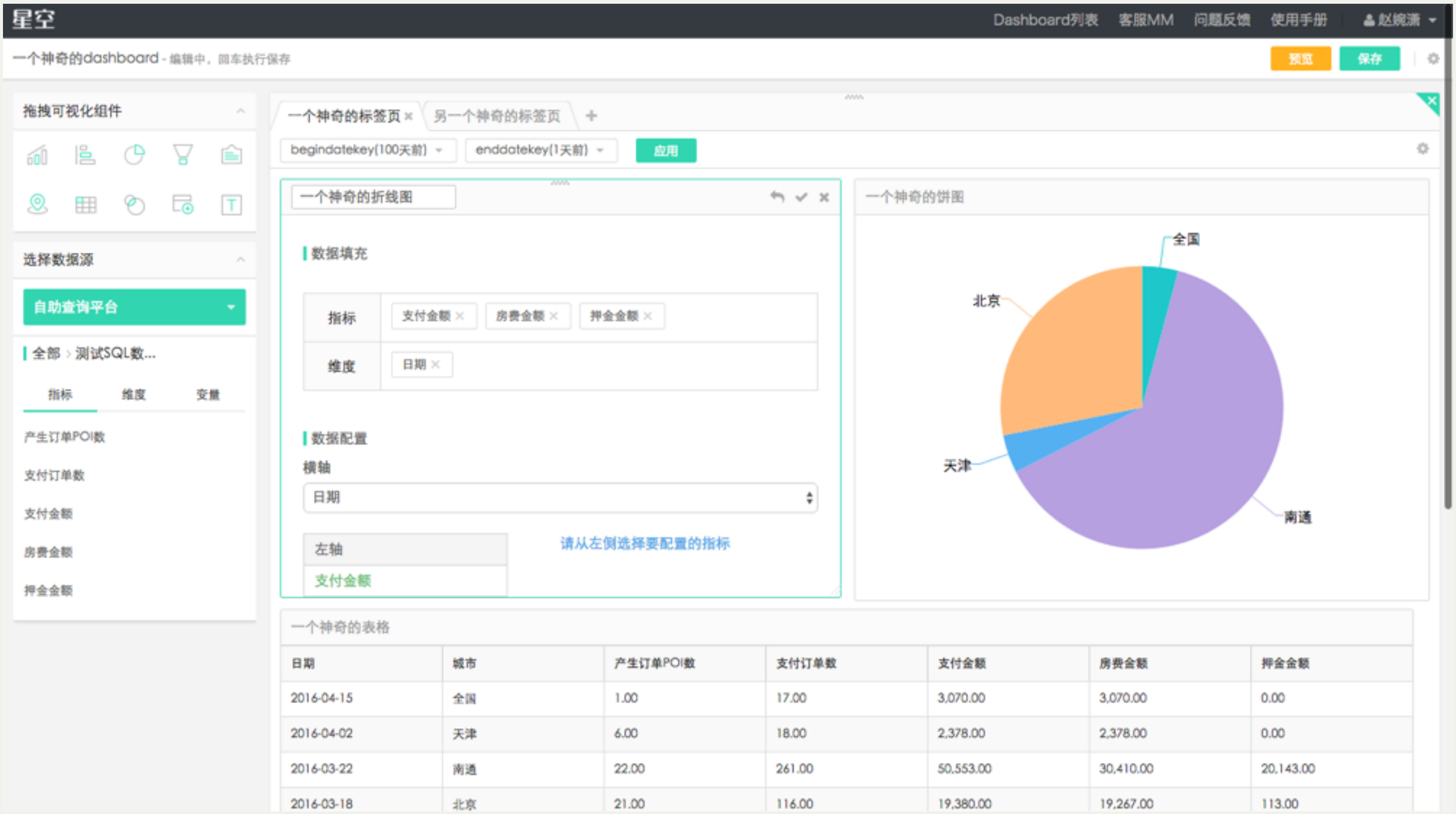
20150806

20150807

		MOB				PC			
		美团交易额	美团毛利率	美团毛收入	美团订单数	美团交易额	美团毛利率	美团毛收入	美团订单数
20150801	上海	[redacted]							
	北京								
	南京								
	天津								
	广州								
	杭州								
	深圳								
	西安								
	重庆								
	上海								

BI产品 - 星空图表

美团网 meituan.com



数据平台时间线

美团网 meituan.com

“手写”
PHP
报表

SQL
报表
工具

自研
调度
系统

ETL
开发
平台

Binlog
收集

引入
Spark

Hadoop
多机房

OLAP
引擎

2011 2012 2013 2014 2015 2016

SQL
ETL
工具

数据
仓库
概念

上Hadoop

实时
计算
平台

Hadoop2.0
YARN

离线
任务
平台

数据
质量
&
SLA
保障

交互式
开发

数据平台时间线

“手写” PHP 报表

SQL 报表 工具

自研 调度 系统

ETL 开发 平台

Binlog 收集

引入 Spark

Hadoop 多机房

OLAP 引擎

2011 2012 2013 2014 2015 2016

SQL ETL 工具

数据 仓库 概念

上Hadoop

实时 计算 平台

Hadoop2.0 YARN

离线 任务 平台

数据 质量 & SLA 保障

交互式 开发

数据平台时间线

“手写”PHP报表

SQL报表工具

自研调度系统

ETL开发平台

Binlog收集

引入Spark

Hadoop多机房

OLAP引擎

2011

2012

2013

2014

2015

2016

SQL ETL 工具

数据仓库概念

上Hadoop

实时计算平台

Hadoop2.0 YARN

离线任务平台

数据质量&SLA保障

交互式开发

数据平台时间线

美团网 meituan.com

2011
- “手写” PHP 报表
- SQL 报表工具
- SQL ETL 工具

2012
- 自研调度系统
- 数据仓库概念

2013
- ETL 开发平台
- 上Hadoop
- 实时计算平台

2014
- Binlog 收集
- 引入 Spark
- Hadoop2.0 YARN

2015
- Hadoop 多机房
- 离线任务平台
- 数据质量 & SLA 保障

2016
- OLAP 引擎
- 交互式开发

数据平台时间线

2011
- “手写” PHP 报表
- SQL 报表 工具
- 自研 调度 系统
- ETL 开发 平台

2012
- SQL ETL 工具
- 数据 仓库 概念

2013
- 上Hadoop

2014
- Binlog 收集
- 引入 Spark
- 实时 计算 平台
- Hadoop2.0 YARN

2015
- Hadoop 多机房
- 离线 任务 平台
- 数据 质量 & SLA 保障

2016
- OLAP 引擎
- 交互式 开发

数据平台时间线

美团网 meituan.com

最新与进展

美团网 meituan.com

- Hadoop 单 NameSpace 多机房
- 任务托管 与 交互式开发
- OLAP引擎探索

Hadoop多机房架构

美团网 meituan.com

- 背景
  - 15年初, 被告知机房总机架位500节点
  - 新离线机房最早9月交付
  - 15年6月预估1000节点, 15年12月预估1500节点
  - 业务任务耦合重, 难以快速拆分

Hadoop多机房架构

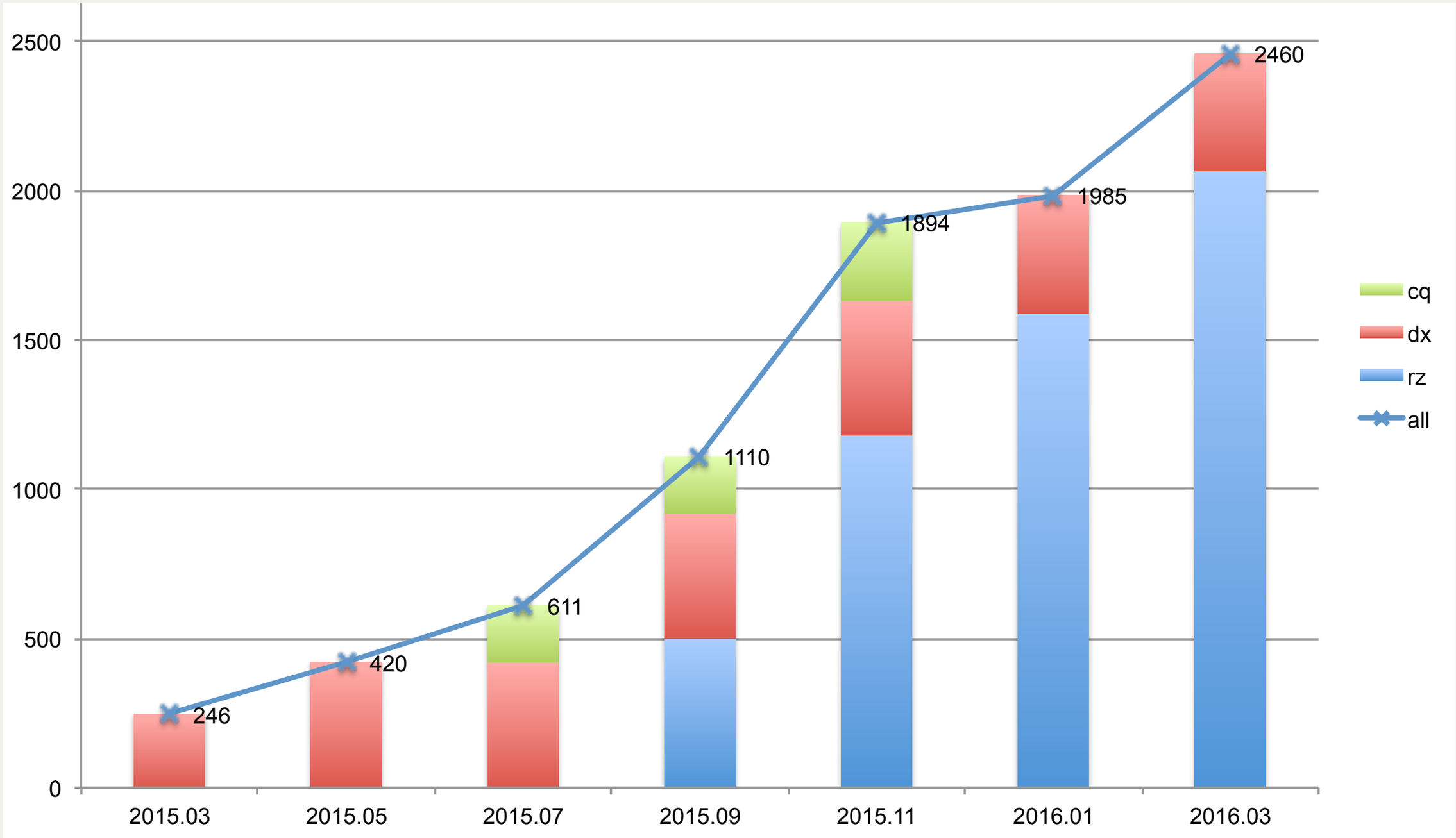
美团网 meituan.com

- Hadoop多机房核心问题
  - 跨机房带宽小, 延迟高
  - Hadoop分布式系统, 跨节点就可能跨机房
- Hadoop跨节点数据流场景
  - App内部container间网络交换
  - 非DataNode本地读取
  - HDFS写入pipeline

Hadoop多机房架构

- 架构决策
  - 先多机房, 再拆NameSpace
  - 每个节点计算所属机房
  - YARN队列增加机房属性, 修改调度策略单任务只调度到单机房
  - HDFS修改addBlock策略, 只返回client所在机房DataNode构成的pipeline. 读数据时优先读client所在机房.
  - 修改HDFS Balancer策略只在单机房内部迁移
  - 走Balancer挪块的接口, 构造文件Block分布/迁移控制工具

Hadoop多机房架构 - 效果



Hadoop多机房架构

美团网 meituan.com

- 特点
  - 代码改动小, 范围可控
  - 快速开发, 顶住资源问题
  - 业务透明迁移

任务托管与交互式开发

美团网 meituan.com

- 背景
  - 业务基于Hadoop/Spark客户端开发
  - 编译, 测试, 开发效率低
  - 环境部署效率低, 管理成本高
  - 代码编译环境/执行环境不确定, 问题追查周期长
  - Spark开发效率更高, 但是学习/尝试成本高

任务托管与交互式开发

美团网 meituan.com

- 架构决策
  - 任务托管平台
    - 任务代码, 打包, 执行统一平台化管理
  - 交互式开发工具
    - 调研了ipython notebook+spark和zeppelin
    - 基于后者开发, 修复一系列bug / 补充登陆&认证
    - 解决了开发者尝试代码逻辑, 互相参考代码实现的需求

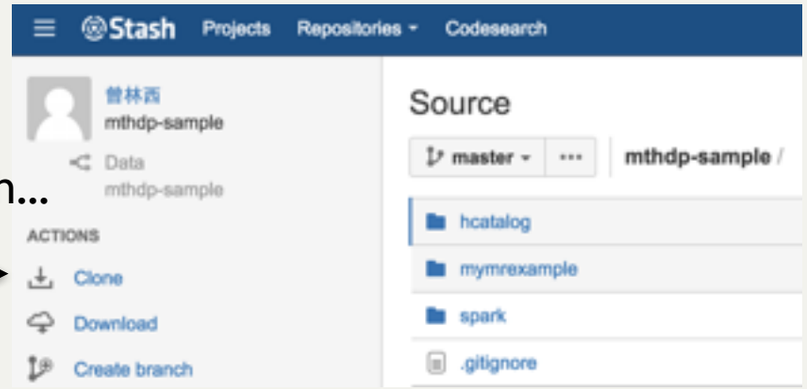
任务托管平台

美团网 meituan.com

user code...

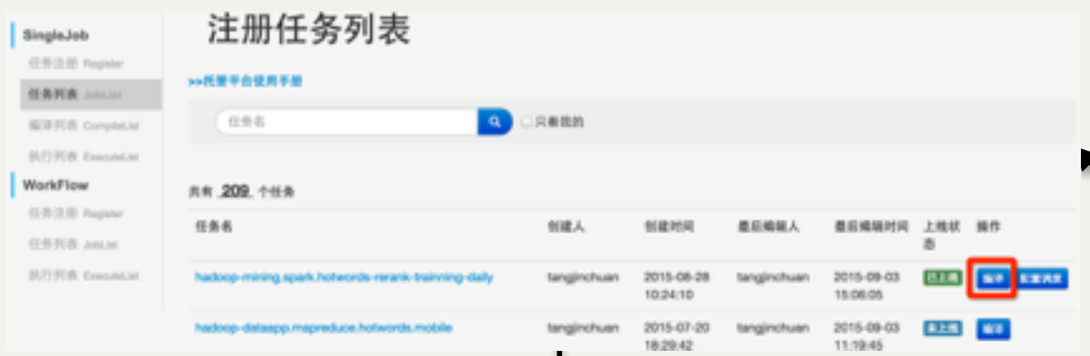
```scala
import java.awt.Polygon
val aor_list = sqlContext.sql("select id, point_string from mart_waimai.fact_wm_aor_monfst_snapshot where dt='20150601'")
val aor_list_p = aor_list.map(aor => {
  val p = new Polygon()
  aor.getString(1).split(";").map(loc => {
    val lat_lng = loc.split(",").map(s => (s.toDouble * 1000000).toInt)
    p.addPoint(lat_lng(1), lat_lng(0))
  })
  (aor.getInt(0), p)
})
val local_aor = aor_list_p.collect()
val aors = sc.broadcast(local_aor)

val fact = sqlContext.sql("select id, longitude, latitude from mart_waimai.fact_ord_submitted where dt='20150601'")
val matched = fact.map(l => {
  (l.getInt(0), aors.value.find(a => a._2.contains(l.getInt(1), l.getInt(2))).map(_._1))
}).filter(! _._2.isEmpty)
val res = matched.count
matched.saveAsTextFile("/tmp/hcatalog-example/multiinput")
```

git push...

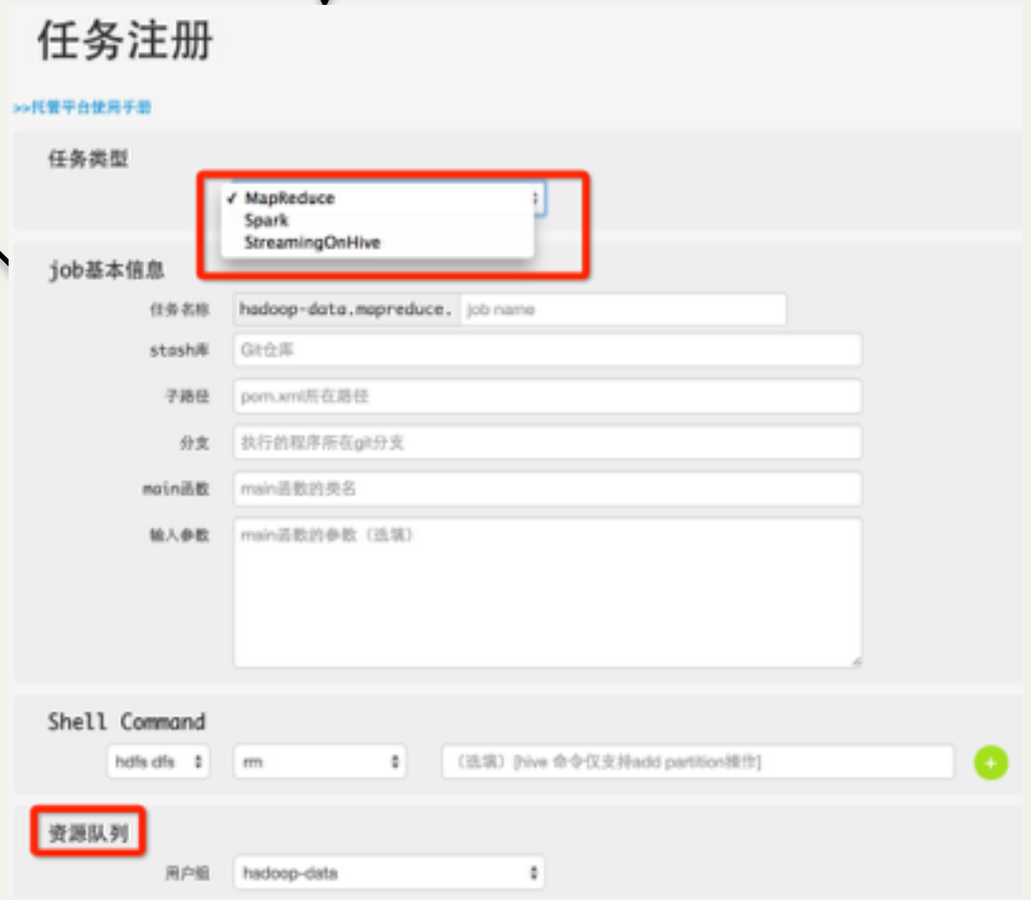


Hadoop作业管理

托管平台注册…



Hadoop作业编译、执行…



交互式开发

美团网 meituan.com

Zeppelin Notebook Interpreter

adp#zhengdong#搜索补余数据统计

```
val hotel = """hotelsearch""".r
val poi_supp_imp_with_hotel = sqlContext.sql(
    s"""
    select
        `_mt_servername`, recommendstids, uuid
    from
        log.dataapp_recapi_search
    where
        dt='20160306' and length(recommendstids)>7 and length(searchct_pois)<3 and length(searchstids)<3
    """).map(
            r => (r.getString(0), r.getString(1), r.getString(2))
        ).map{
            case(servername, stids, uuid) => (hotel.findFirstIn(servername), servername, stids, uuid)
        }
poi_supp_imp_with_hotel.count()
```

```
hotel: scala.util.matching.Regex = hotelsearch
poi_supp_imp_with_hotel: org.apache.spark.rdd.RDD[(Option[String], String, String, String)] = MapPartitionsRDD[2420] at map at <console>:48
res141: Long = 1016319
```

Took 54 seconds

```
case class NoresAll(hotel:Option[String], servername: String, stids: String, uuid: String)
val poi_supp_table = poi_supp_imp_with_hotel.map(r => NoresAll(r._1, r._2, r._3, r._4)).toDF()
poi_supp_table.show()
```

```
defined class NoresAll
poi_supp_table: org.apache.spark.sql.DataFrame = [hotel: string, servername: string, stids: string, uuid: string]
+-----+------------------+------------------+------------------+
|hotel|          servername|             stids|              uuid|
+-----+------------------+------------------+------------------+
| null|dx-dataapp-recapi...|[{"stid":"1441282...|CF8826291EB19AC90...|
| null|yf-dataapp-recapi...|[{"stid":"8214696...|E438CE6E9FADF1C59...|
| null|yf-dataapp-recapi...|[{"stid":"7350005...|CBB68923F11C80C2A...|
| null|dx-dataapp-recapi...|[{"stid":"1873628...|41846607936416148...|
| null|yf-dataapp-recapi...|[{"stid":"4324761...|67F3CC18D9F5E1EA2...|
| null|dx-dataapp-recapi...|[{"stid":"7218742...|6B81FE3C1BBC9471E...|
```

OLAP引擎

美团网 meituan.com

- 需求特点
  - 亿级别事实, 50以内指标
  - 千万级别维度, 20个以内类别
  - TP99 < 3S
  - 多种维度组合聚合查询
  - 去重指标要求精确

OLAP引擎

- 可能的方案
  - Presto / Hive / Spark on ORC File宽表
  - Hive grouping set 导入HBase + 二级索引
  - Druid
  - ElasticSearch
  - Kylin

OLAP引擎

- 探索思路
  - 考虑稳定性, 成熟度, 掌控力, 社区活跃度, 先大力尝试Kylin, 并在业务线尝试落地
  - 基于Star Schema Benchmark, 构造OLAP场景测试数据, 对社区方案对比测试
  - 分享进展并收集业务特殊需求后, 迭代测试用例

Kylin - OLAP分析引擎

美团网 meituan.com

Kylin | Query | Cubes | Jobs | Tables | Admin | Help | Welcome, ADMIN

Project: waimai_dolphin **Cube Name:** Filter ... + Cube

Cubes

Name	Status	Cube Size	Source Records	Last Build Time	Owner	Create Time	Actions	Admins
app_dt__poi_audit	READY	22.54 GB	35,704,918	2016-04-15 17:14:28 GMT+8	ADMIN	2016-04-15 14:19:09 GMT+8	Action	Action

Grid | Visualization | SQL | JSON(Cube) | JSON(Model) | Access | Notification | HBase

Cube Designer

Cube Info (1) — Data Model (2) — Dimensions (3) — Measures (4) — Filter (5) — Refresh Setting (6) — Advanced Setting (7) — Overview (8)

Filter ...

ID	Name	Expression	Param Type	Param Value	Return Type
1	_COUNT_	COUNT	constant	1	bigint
2	AUDIT_ACT_COST	SUM	column	AUDIT_ACT_COST	decimal
3	AUDIT_ACT_ORD_AMT	SUM	column	AUDIT_ACT_ORD_AMT	decimal
4	OLP_ORD_NUM	SUM	column	OLP_ORD_NUM	bigint

Apache Kylin | Apache Kylin Community

StarSchemaBenchmark

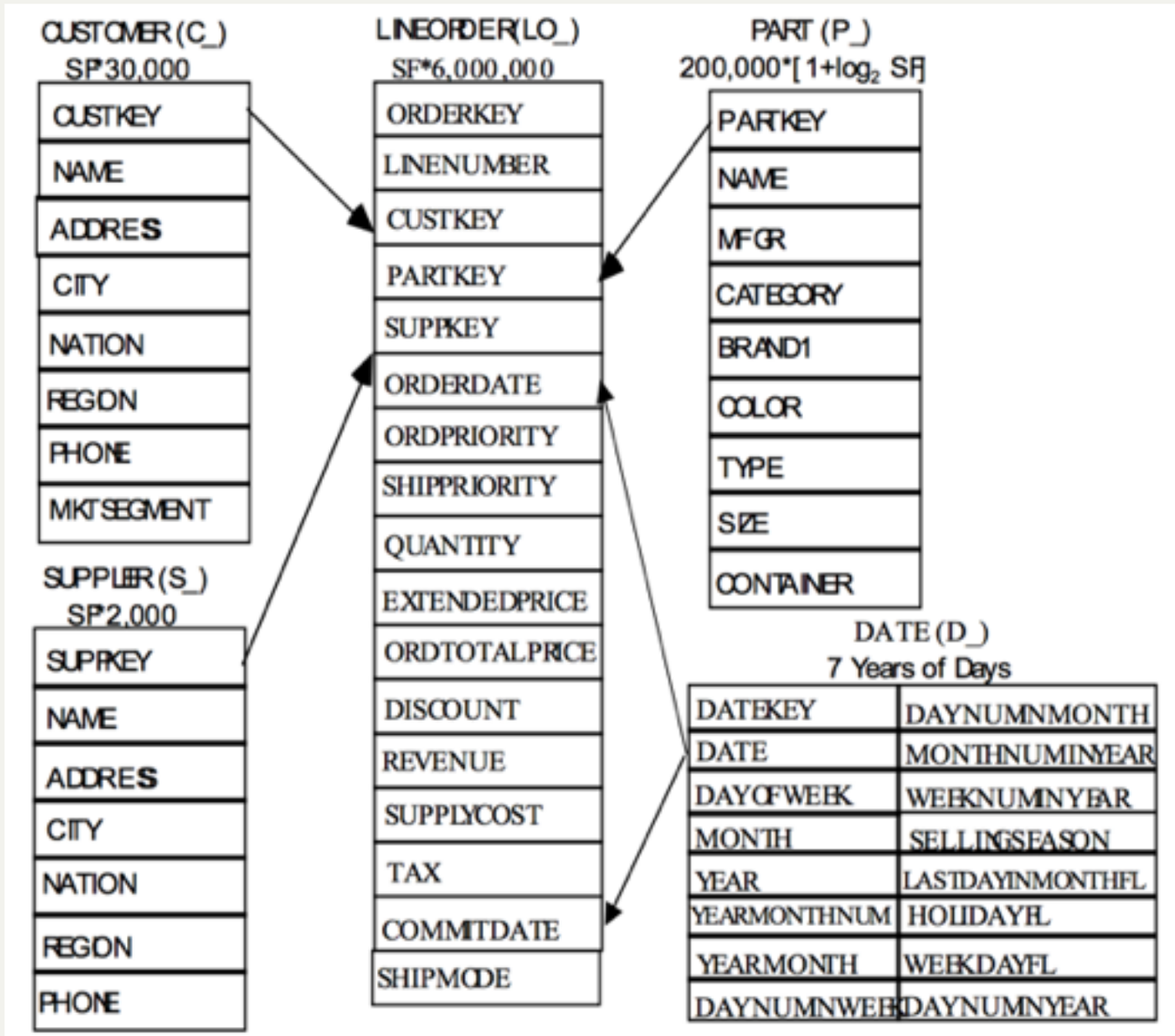


Figure 1.2 SSB Schema

OLAP引擎

- 目前进展
  - 完成Presto、Kylin1.3、Kylin1.5、Druid测试
  - 支持某BI项目7个数据立方体
  - 业务开发周期7天 -> 1~2天
  - 3亿行数据, TP95%查询响应时间在1s内, 日查询量2万

平台化思考

- 平台的价值
- 重复的事情做一次, 做得精
- 统一化, 减少业务间对接成本
- 为业务效率负责, 第一时间考虑业务成本

平台化思考

- 平台的发展
- 支持业务是第一位的
- 与先进业务同行，辅助并沉淀技术
- 设立规范，用积累的技术支撑后发业务

平台化思考

美团网 meituan.com

- 关于开源
  - 持续关注 & 前瞻性调研
  - 共性patch进行贡献
  - 选择性重构
  - 理智权衡 & 选型

美团网

meituan.com

谢谢！