GIAC

全球互联网架构大会

GLOBAL INTERNET ARCHITECTURE CONFERENCE

小米广告数据BI平台实践

贾菁辉　小米　广告数据技术主管


GIAC
全球互联网架构大会
GLOBAL INTERNET ARCHITECTURE CONFERENCE
GIAC 是中国互联网技术领域行业盛事，组委会从互联网架构最热门领域甄选前沿的有典型代表的技术创新及研发实践的架构案例，分享他们在本年度最值得总结、盘点的实践启示。
2018年11月 | 上海国际会议中心
msup
关注msup
公众号获得
更多案例实践
架構
高可用架构
改变互联网
的构建方式

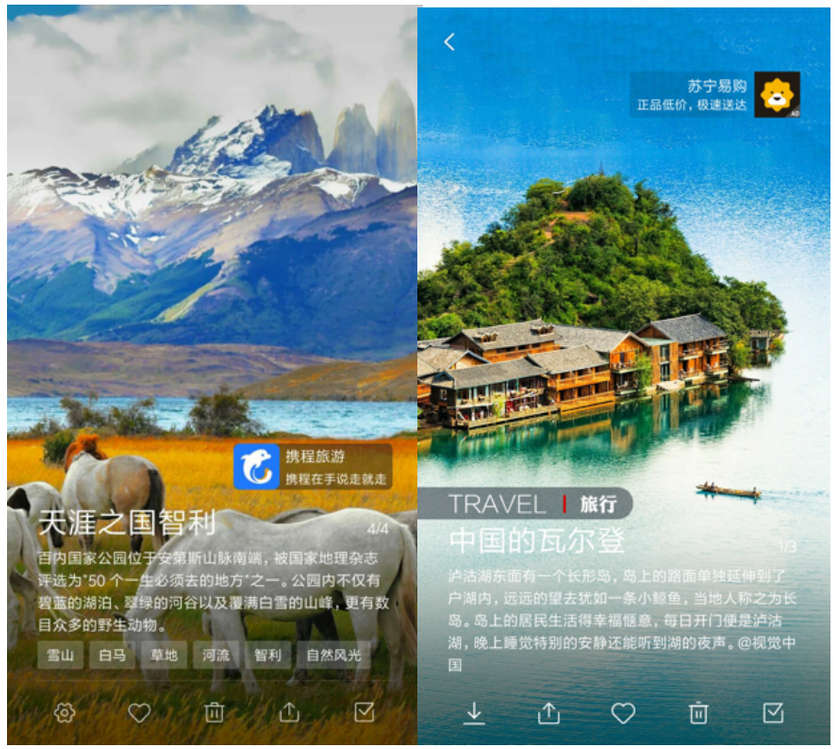
目录

目录

小米的广告业务


浏览器
视频
音乐
新闻资讯
28°
天气
16
日历
电视/盒子


13:54
电信
联通 3G
好听！TFBOYS献唱快本主题曲《同一秒快乐》
映客直播
性感女神
又|玩|出|了|新|花|样
立即下载
广告
热歌榜
歌手馆
领米币
最热MV
私人电台
推荐歌单
TOP50
NEW MUSIC
2017上半年最热新歌TOP50
流行
此生最幸运的是与你相遇
华语 / 流行 / 放松
每日新歌
华语 / 欧美 / 韩语
开始下载...
我的
精选
榜单
专栏


携程旅游
携程在手说走就走
天涯之国智利
4/4
百内国家公园位于安第斯山脉南端，被国家地理杂志评选为"50 个一生必须去的地方"之一。公园内不仅有碧蓝的湖泊、翠绿的河谷以及覆满白雪的山峰，更有数目众多的野生动物。
雪山
白马
草地
河流
智利
自然风光

苏宁易购
正品低价，极速送达
TRAVEL | 旅行
中国的瓦尔登
1/3
泸沽湖东面有一个长形岛，岛上的路面单独延伸到了户湖内，远远的望去犹如一条小鲸鱼，当地人称之为长岛。岛上的居民生活得幸福惬意，每日开门便是泸沽湖，晚上睡觉特别的安静还能听到湖的夜声。@视觉中国

小米的广告业务

售卖方式：

- CPC、CPD、CPM
- CPT、Schedule
- RTB

定向：

- 地域：国家、省(州)、市(县)
- 设备型号：手机、电视、盒子
- 时间：小时级
- 人群：性别、年龄 等
- 内容：上下文、剧集、CP
- 特殊：天气状况

频次控制：

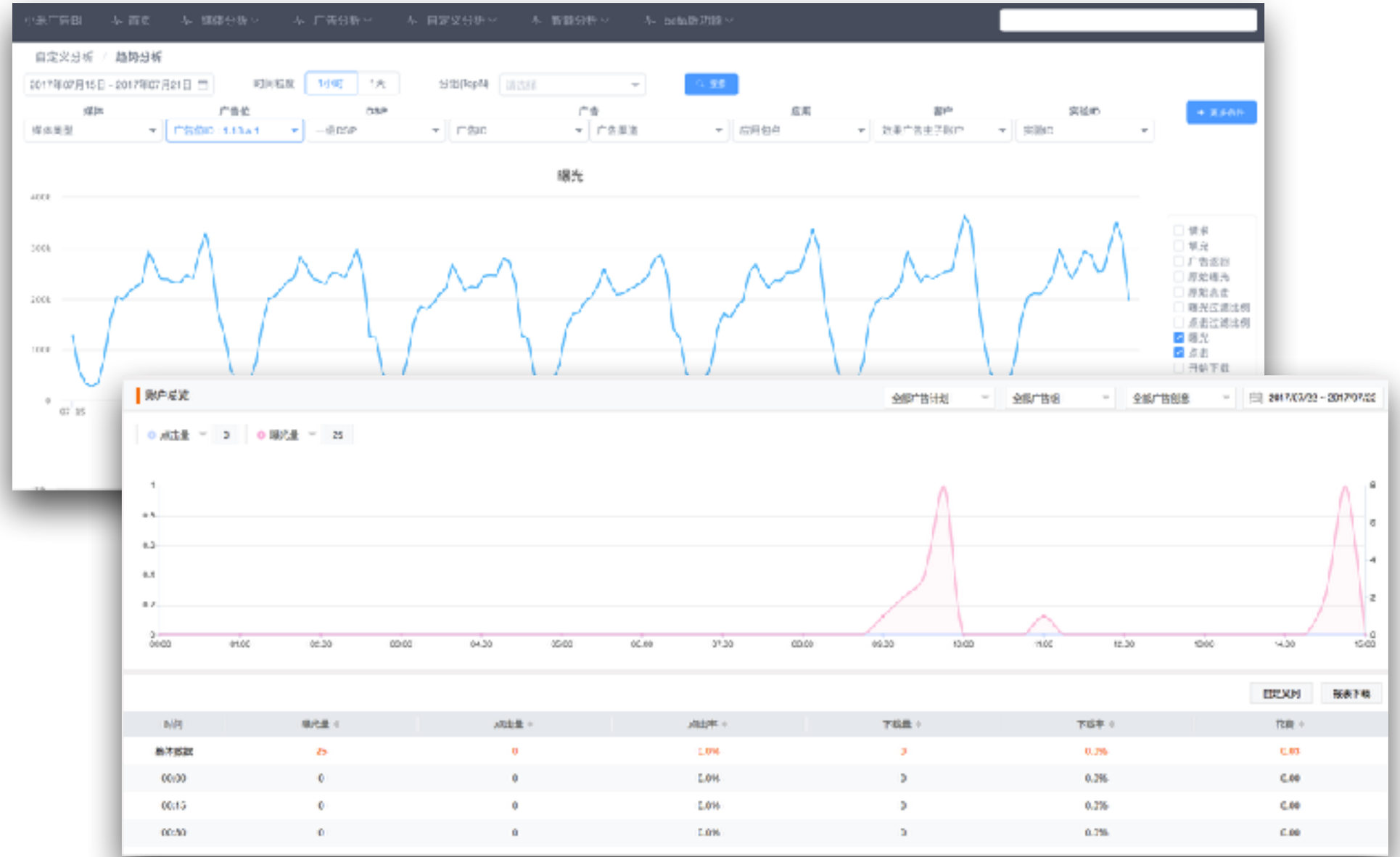
- 小时、日、周、月

BI需要解决的问题

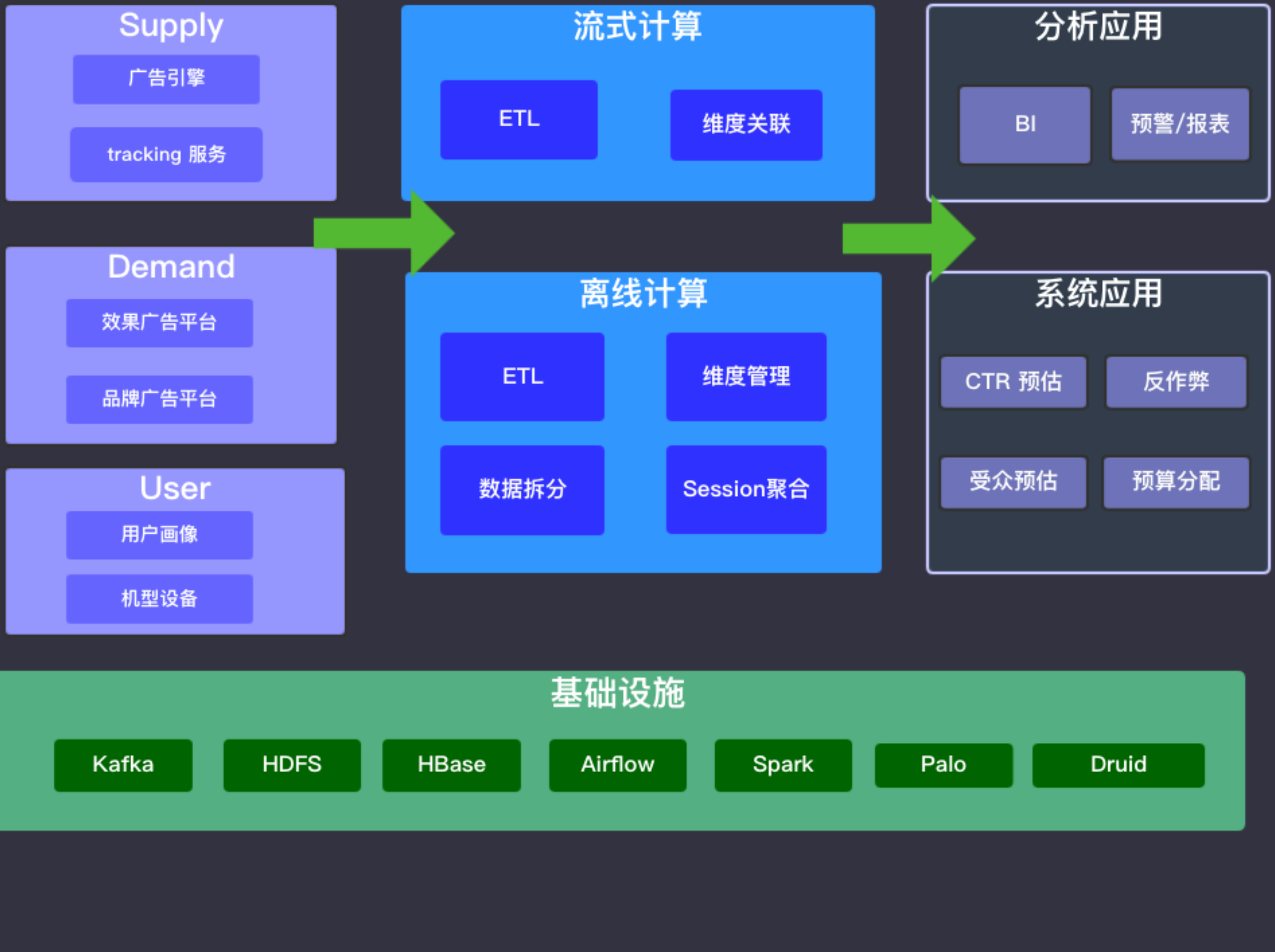
- 流量碎片化
- 业务国际化
- 客户需求复杂
- 内部数据获取效率问题

BI应用

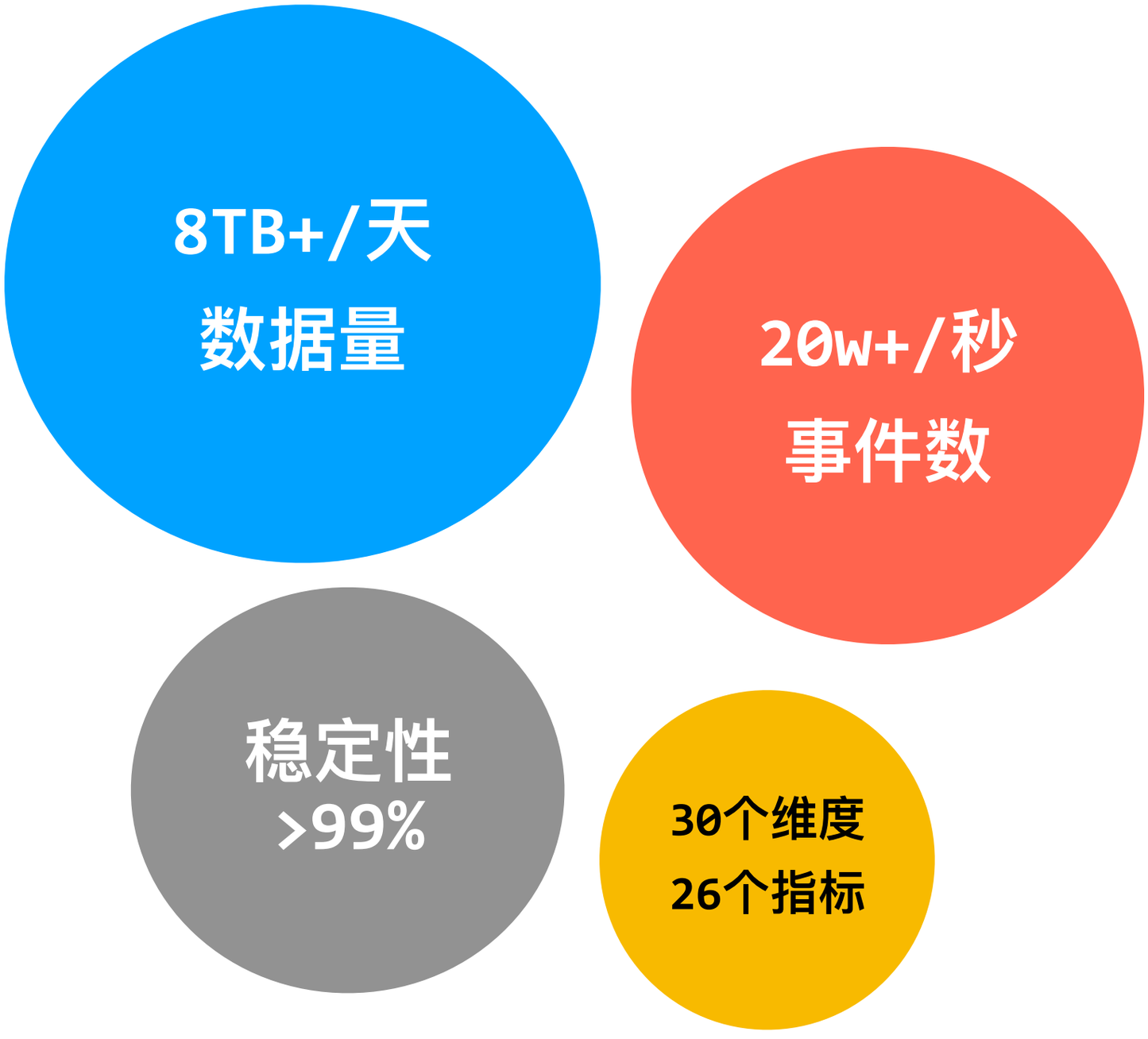
- **报表**
  - 业务邮件报表
  - 广告主报表
  - …
- **分析**
  - 多维分析
  - 归因分析
  - 账户诊断
  - A-B Test
  - …
- **监控&预警**



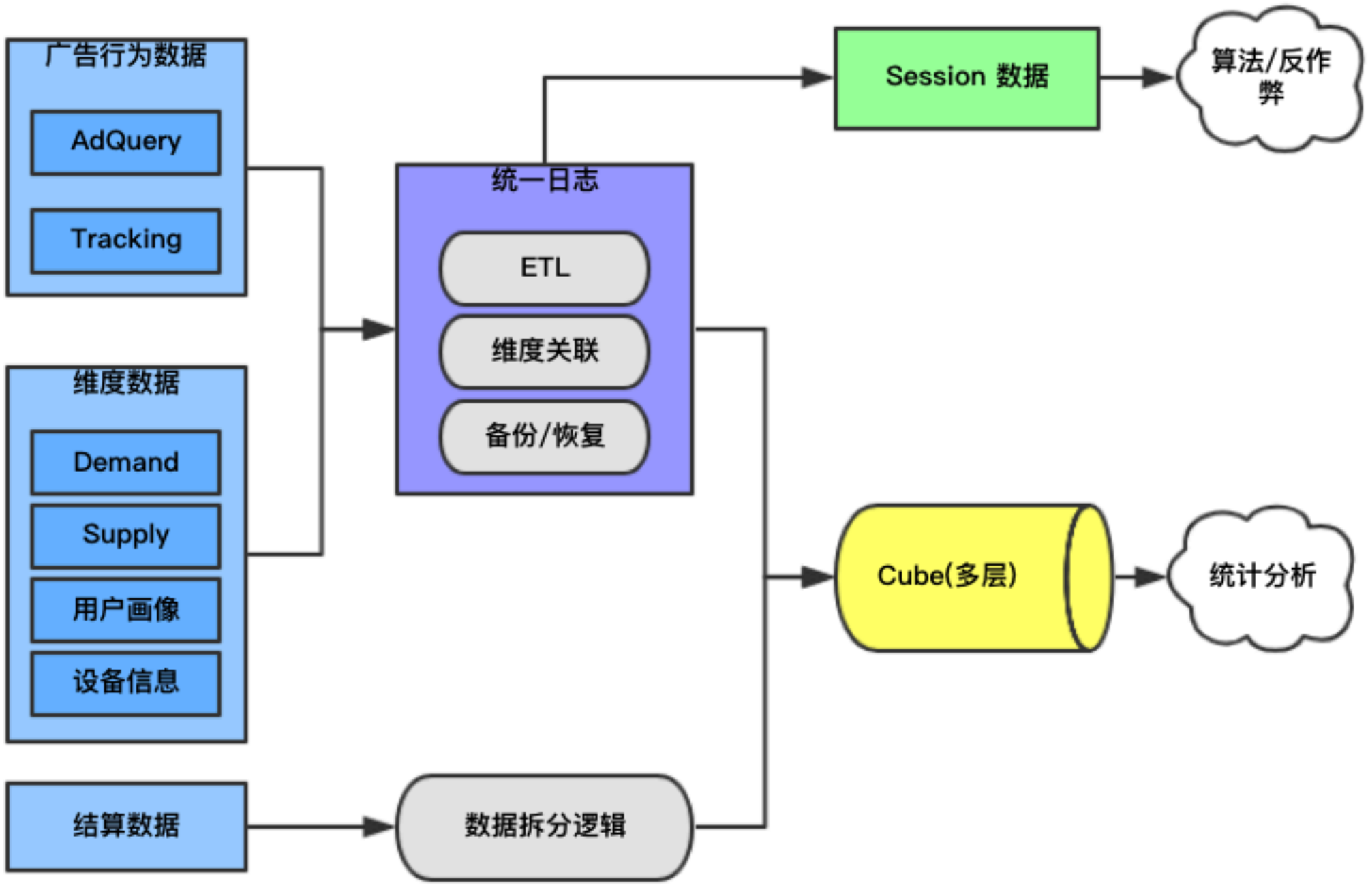
目录


Supply
广告引擎
tracking 服务
Demand
效果广告平台
品牌广告平台
User
用户画像
机型设备
流式计算
ETL
维度关联
离线计算
ETL
维度管理
数据拆分
Session聚合
分析应用
BI
预警/报表
系统应用
CTR 预估
反作弊
受众预估
预算分配
基础设施
Kafka
HDFS
HBase
Airflow
Spark
Palo
Druid

平台能力

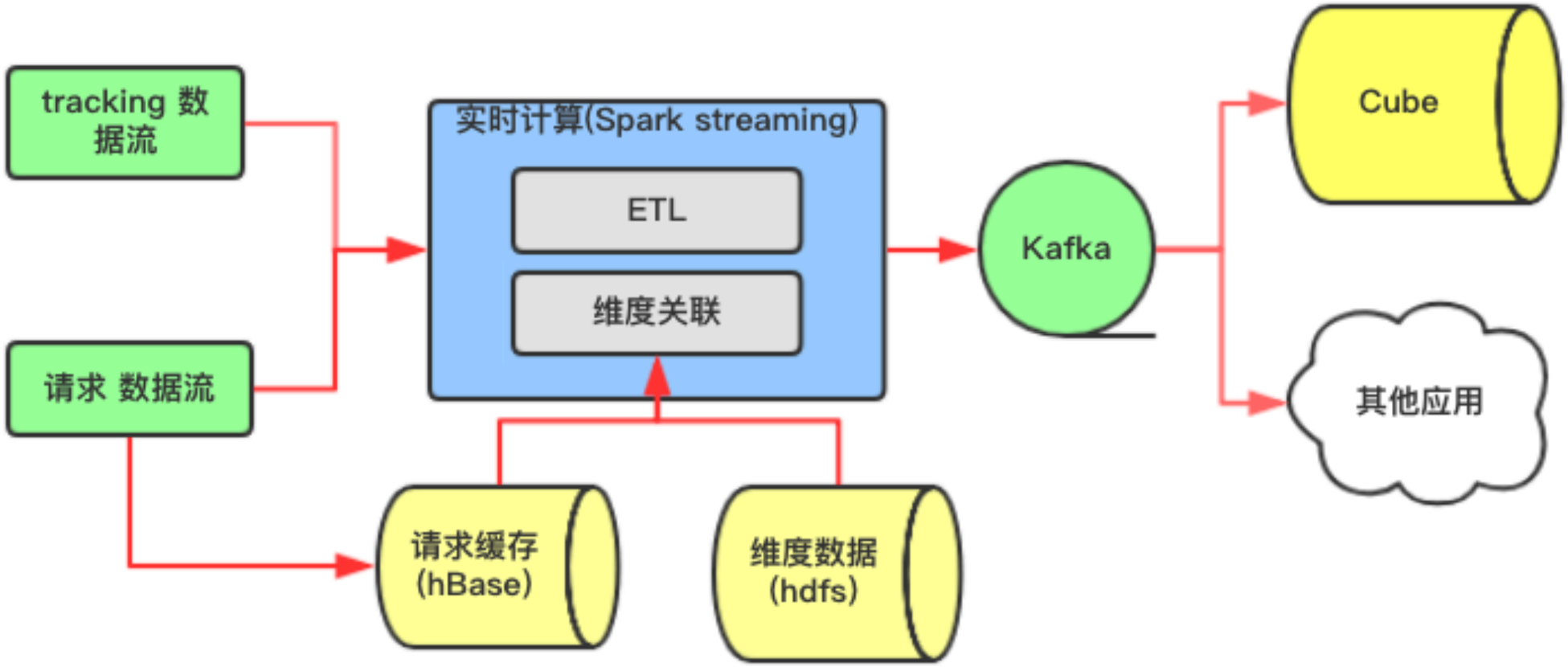


离线计算过程


广告行为数据
AdQuery
Tracking
维度数据
Demand
Supply
用户画像
设备信息
结算数据
统一日志
ETL
维度关联
备份/恢复
数据拆分逻辑
Session 数据
算法/反作弊
Cube(多层)
统计分析

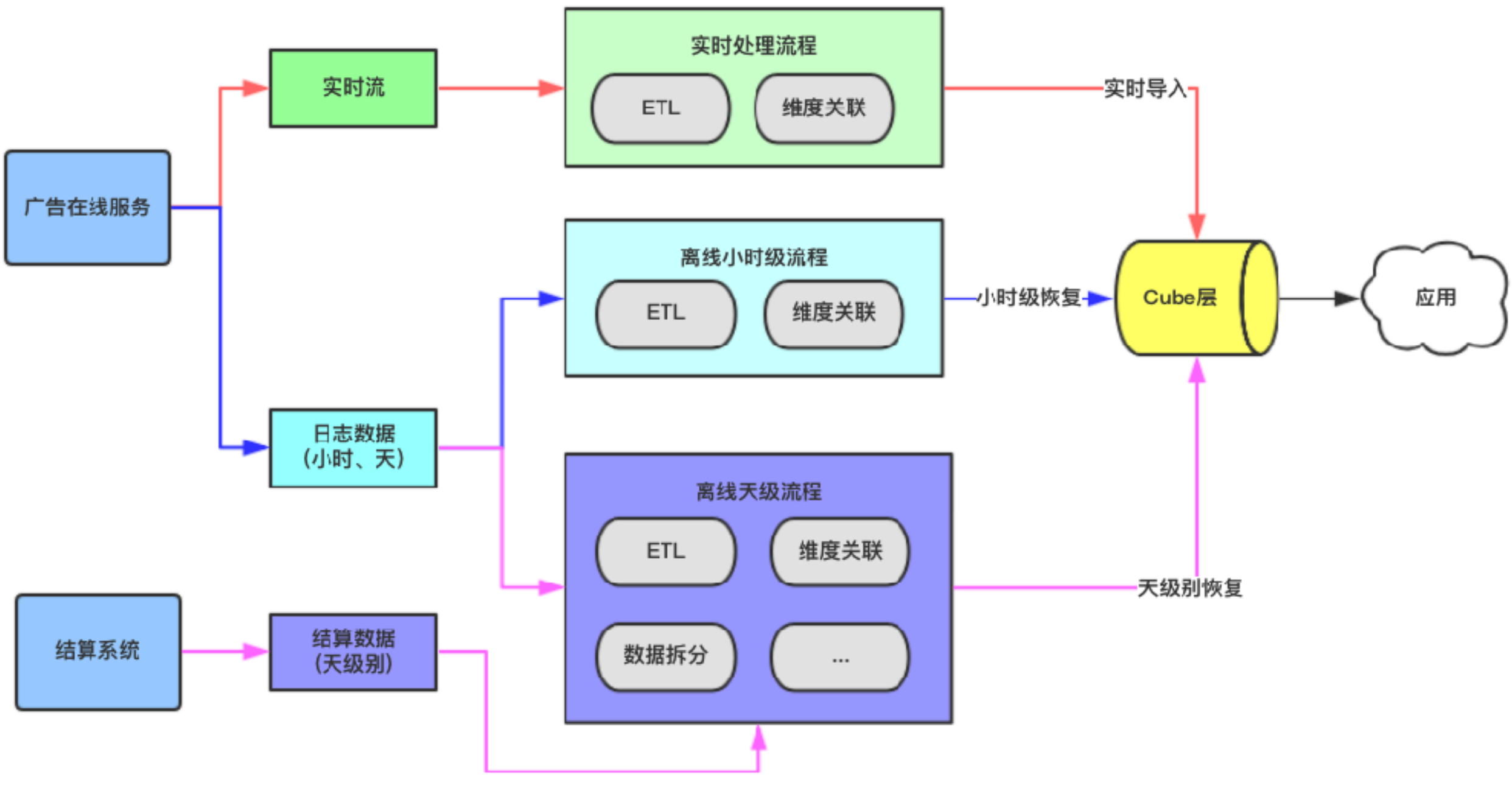
流式计算过程

- **基于spark-streaming**
- **维度关联：补全打点数据所需的维度信息**



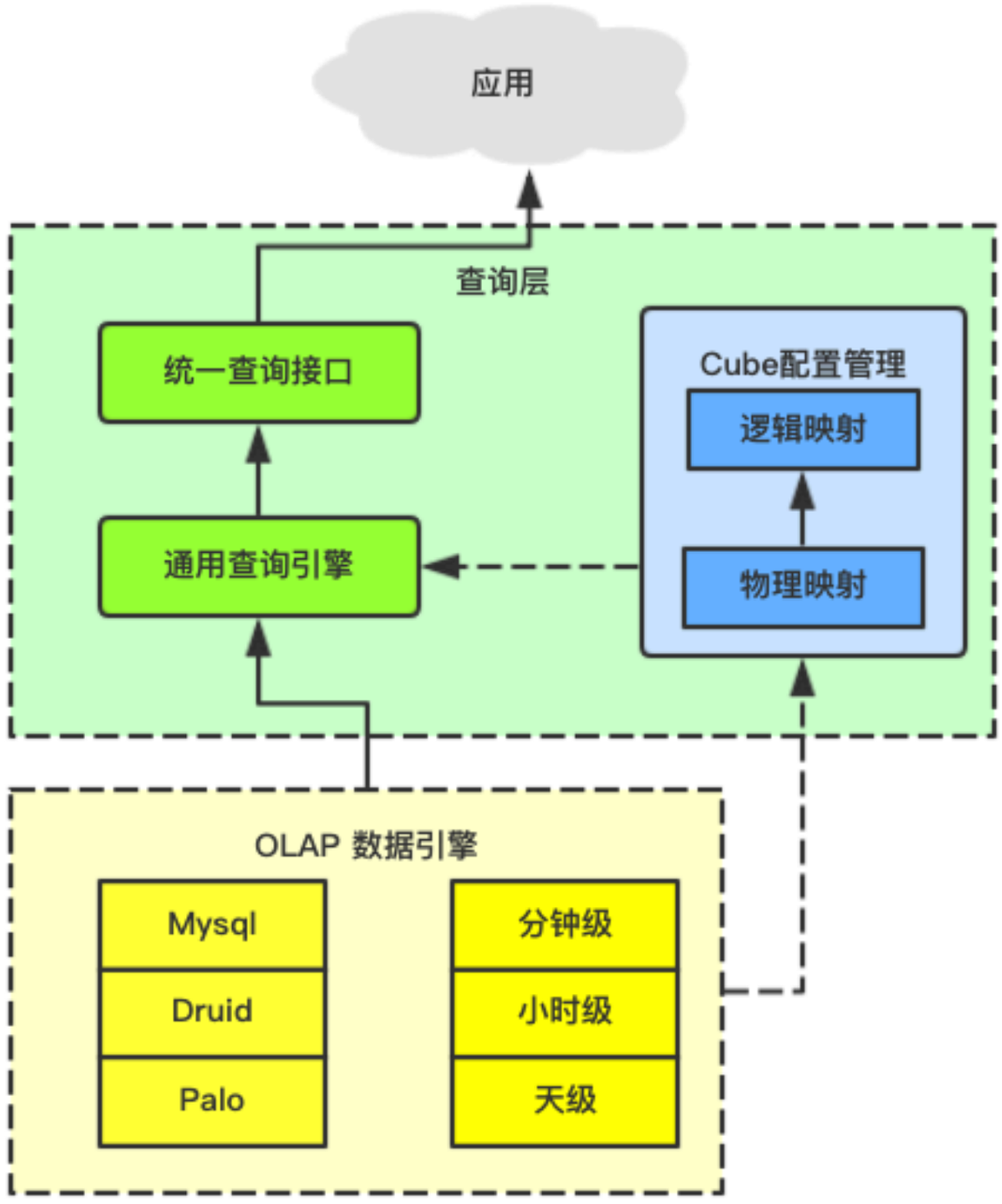
离线与实时数据流的整合

使用小时级和天级离线数据对实时流进行修正，解决数据准确度问题



Cube层结构

- 数据引擎层
  - 多引擎
  - 按时间粒度分层存储
- 查询层
  - 配置化管理
  - 通用查询引擎执行查询逻辑



目录

OLAP解决的问题

userId	time	type	accountId	campaignId	city	adId	fee	event
1	2017-11-1	cpc	12	14	beijing	1	0	view
2	2017-11-1	cpd	13	1	shanghai	2	0	view
1	2017-11-1	cpc	12	14	beijing	1	100	click
3	2017-11-1	cpm	12	15	guangzhou	3	10	view
3	2017-11-1	cpm	12	15	guangzhou	3	0	click
…								

明细数据

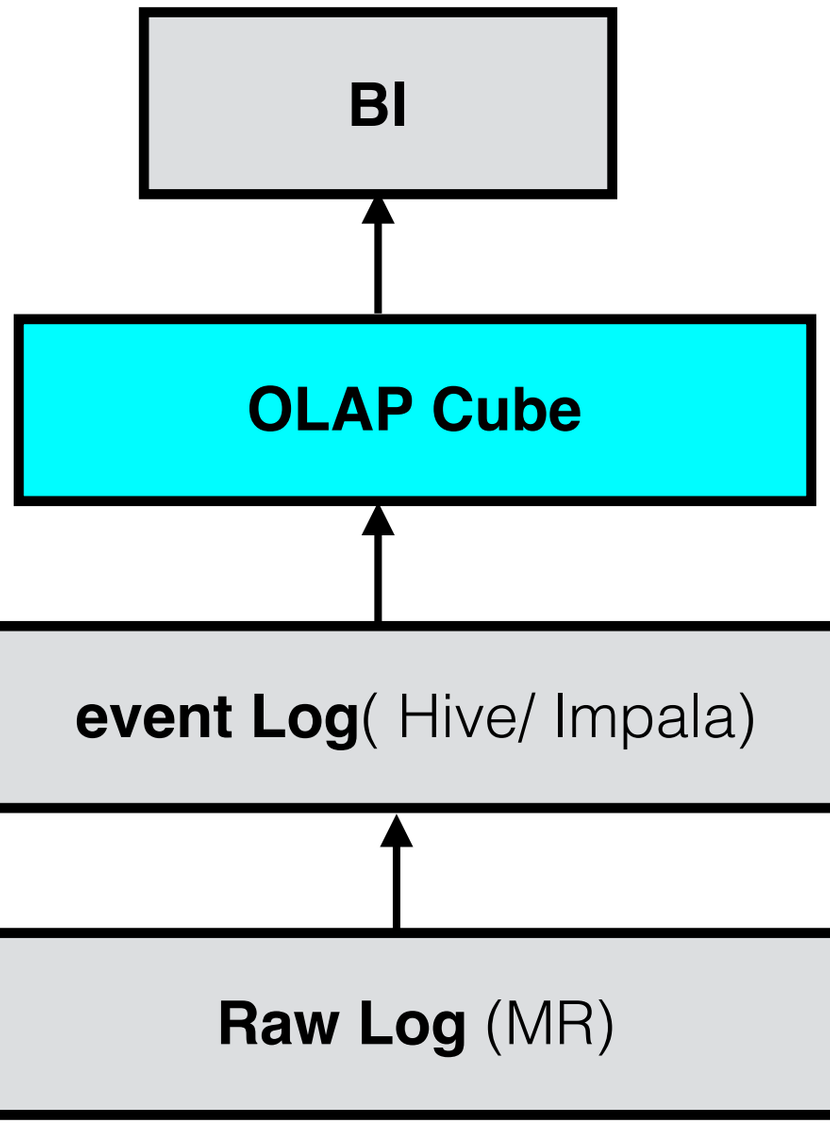
userId	time	type	accountId	campaignId	city	adId	fee	view	click	fee
1	2017-11-1	cpc	12	14	beijing	1	0	2	1	100
2	2017-11-1	cpd	13	1	shanghai	2	0	1	0	0
3	2017-11-1	cpm	12	15	guangzhou	3	10	1	1	10
…										

聚合数据

Q：2017-11-1 广告主A 收到来自北京的收入是多少？平均点击率怎么样?

OLAP所处位置

查询要求：



数据量：

秒级

GB级别

分钟级

TB 级别

10分钟级

TB 级别

技术需求

- 查询性能：常规查询在秒级返回
- 数据实时性：可查询到过去1分钟的数据
- 数据准确性：核心指标无误差，次级指标小于万分之一
- 系统稳定性：> 99.9%，1~2人可维护

OLAP引擎选型

- 早期方案： Mysql + Saiku
- 优势：搭建快速，简单易用
- 劣势：数据维度支持极少，查询性能差，不支持实时数据

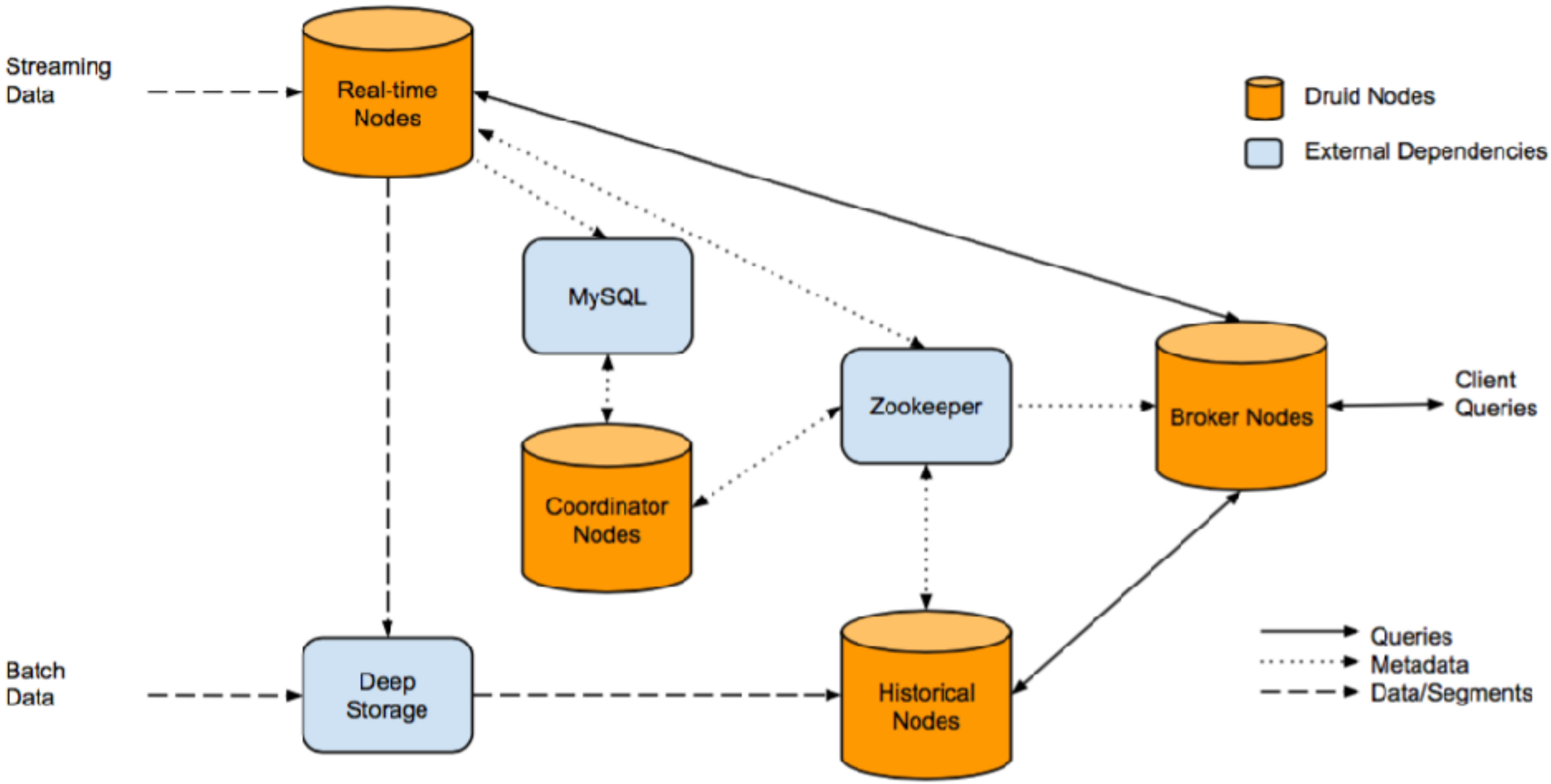
OLAP引擎选型

- 调研方案：Kylin
  - 特点：基于hBase的k-v存储，空间换时间
  - 优势：查询性能高
  - 劣势：需要手工建物化视图，当年不支持实时数据，数据膨胀

OLAP引擎选型

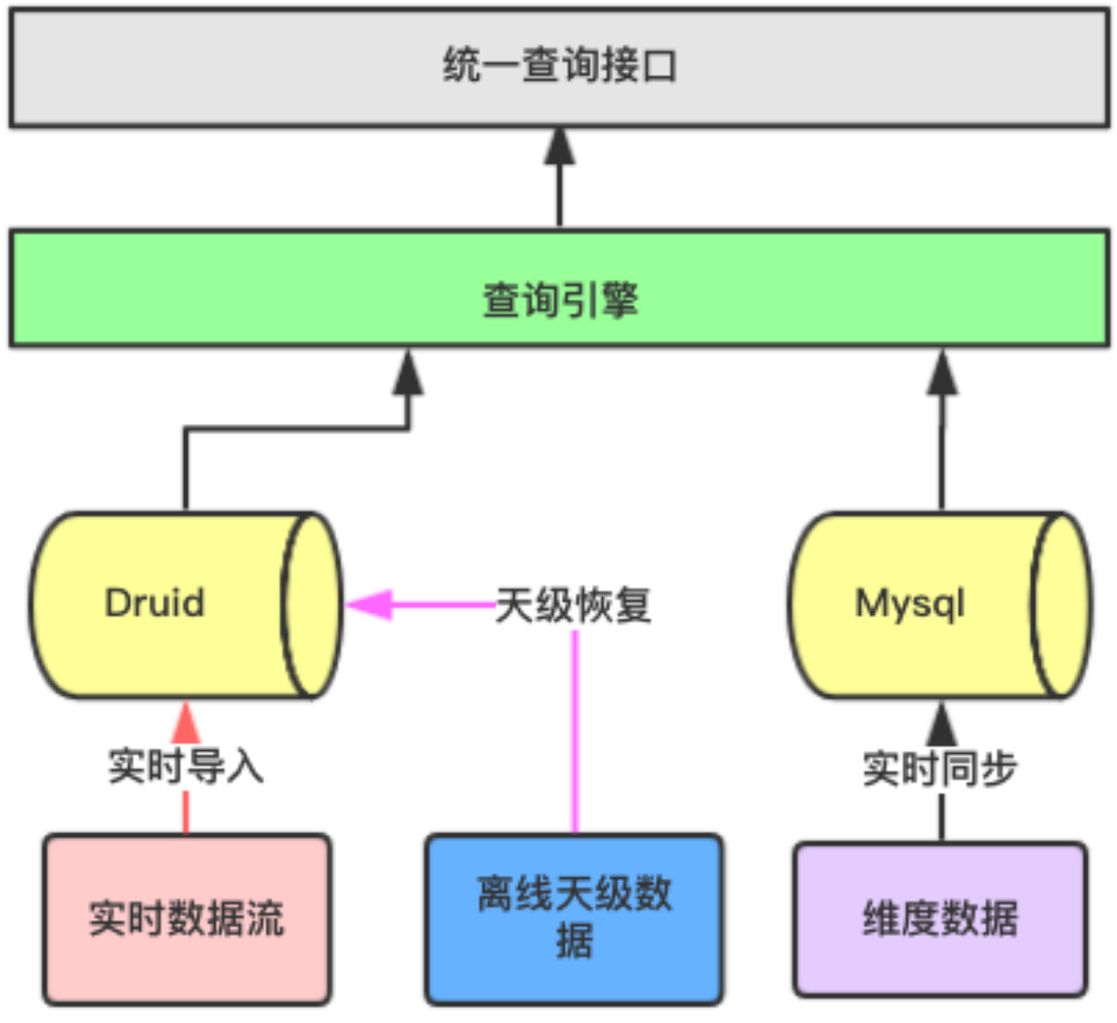
- 中期方案：Druid
- 特点：列存储 + bitmap索引
- 优势：原生支持实时数据；查询性能高；系统稳定
- 劣势：大型GroupBy查询失败率高

Druid系统稳定性


Streaming
Data
Real-time
Nodes
MySQL
Zookeeper
Broker Nodes
Client
Queries
Coordinator
Nodes
Batch
Data
Deep
Storage
Historical
Nodes
Druid Nodes
External Dependencies
Queries
Metadata
Data/Segments

基于Druid的查询架构

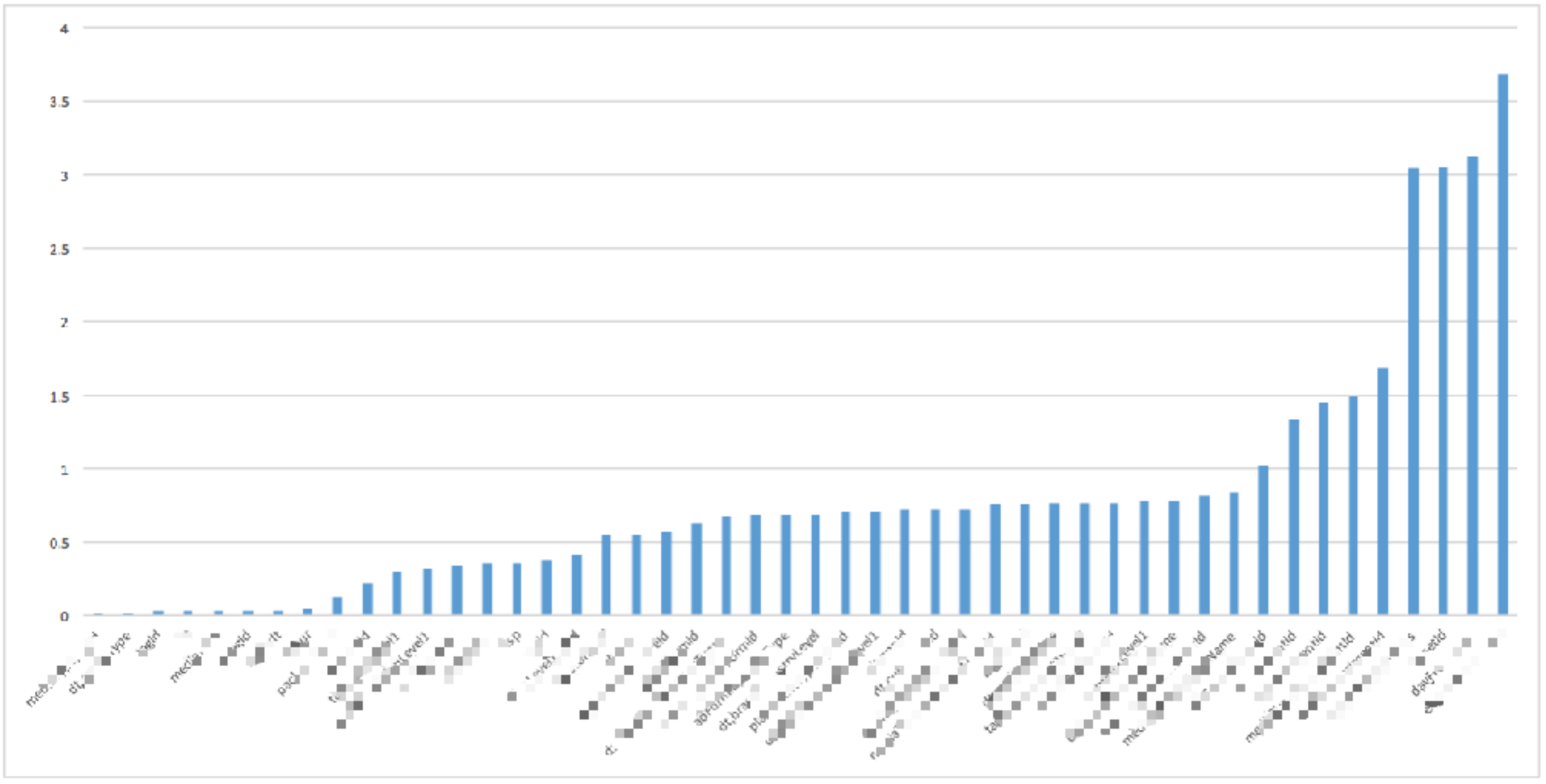
- 优势：
  - 原生支持实时数据流；
  - 查询性能高；
  - 系统稳定
- 不足：
  - 对维度表支持不友好
  - 大型GroupBy查询失败率高
  - 依赖内存



引入Palo

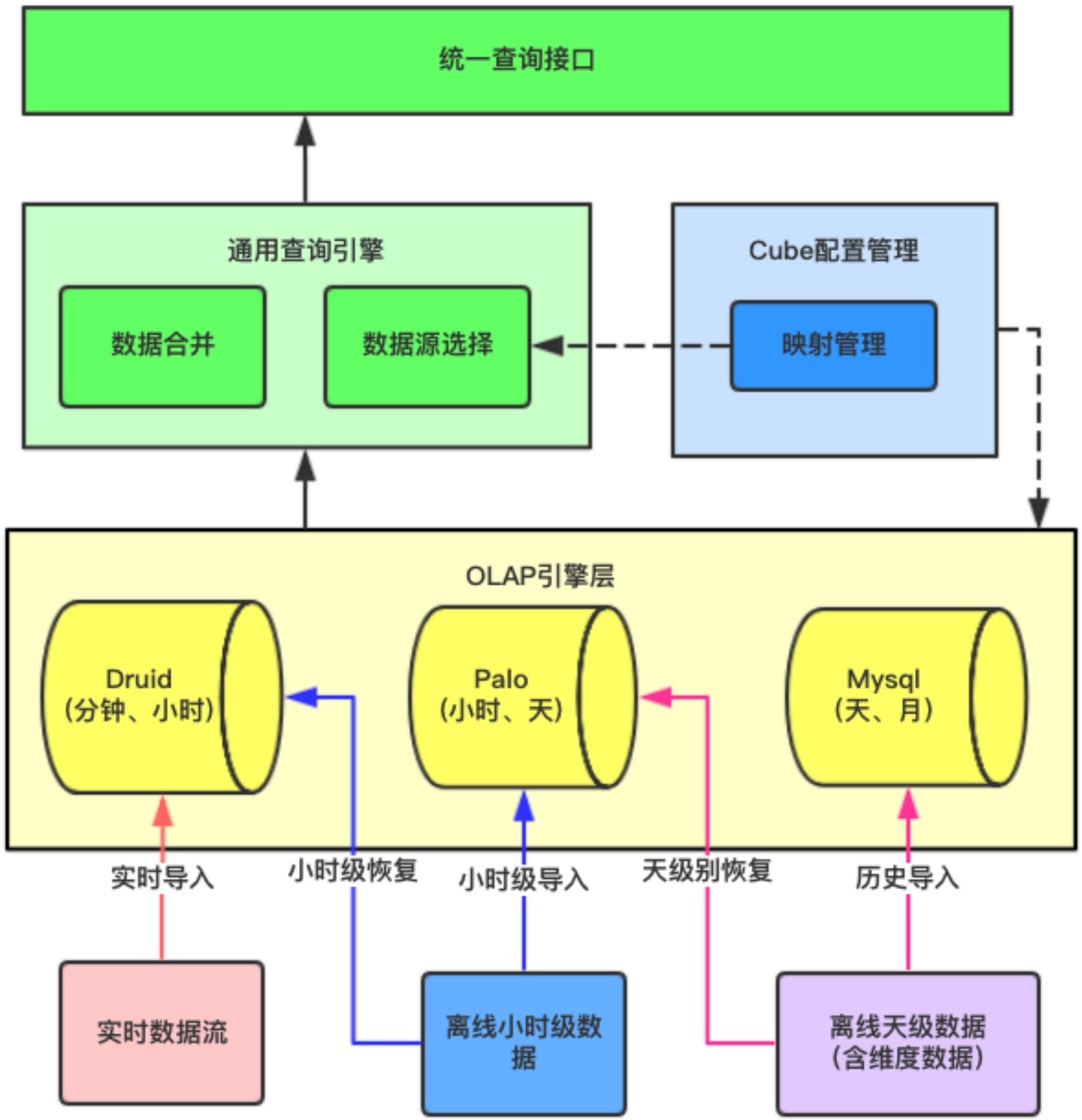
- 支持星形模型
- Group By 查询稳定性高
- 部署简单，只需要上传预编译包和设置be，默认情况下都不需要修改config文件
- mysql接口与操作，上手快，易集成
- 对一个事实表可以根据需要建立多个rollup，平衡了效率和存储空间

Palo 性能测试



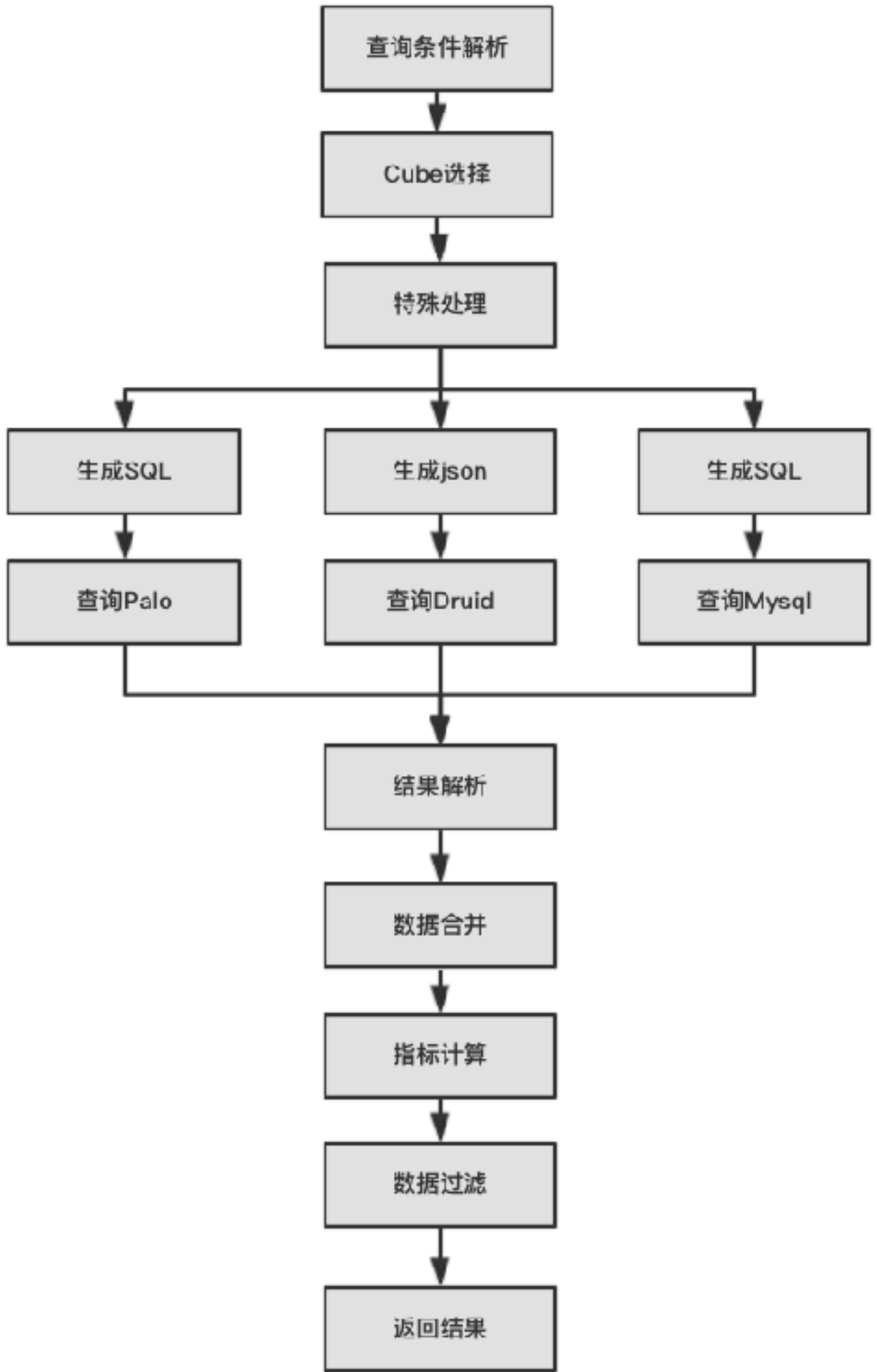
多引擎的查询架构

- 多引擎
  - Druid 实时数据
  - Palo 小时、天级别数据
  - Mysql 历史精确财务数据



通用查询引擎逻辑

- 查询条件解析
- Cube自动选择：
  - 根据维度、指标、粒度、实时性选择支持的cube
- 特殊处理：
  - 特殊指标计算，如：填充率
- 生成不同引擎的查询语句
- 结果解析
- 数据合并
  - 实时与离线数据合并
- 指标计算
- 数据过滤
  - 根据权限进行过滤



BI应用： 统计分析


请选择筛选条件
时间
媒体类型
媒体名称
时间
媒体类型
媒体名称
请求
填充
点击
20171104
20171104
20171104
20171104
20171104
请选择筛选条件
请求
800M
600M
400M
200M
0
11-04
11-05
11-06

BI应用： 数据监控


阈值设置
2016年01月25日 - 2016年02月00日
请求
请求
请求置橙色阈值
请求置红色阈值
测试阈值
5
%
10
%
近29天, 红色报警3天, 橙色报警10天
日期
请求
日环比 (当天比前一天)
2018/01/06
0.06%
2018/01/07
0.50%
2018/01/08
-1.26%
2018/01/09
-10.70%
2018/01/10
-5.89%
2018/01/11
2.25%
2018/01/12
-0.78%
2018/01/13
-1.79%
2018/01/14
12.33%


规则1
规则名称
筛选条件
广告组ID：1.1...
媒体
广告组测试
一级DSP
数据为所有同时满足条件交集的合计结果
监控项
请求
对比项
日环比(当天比前一天)
阈值设置
橙色阈值 ≥5%
红色阈值 ≥10%
+添加下一规则
下一步


我的报警
新建报警
报警名称
规则
收件人
抄送
发送时间
状态
操作
核心开屏广告位报警
共 1 条规则
每日 10:00
暂停中
编辑
测试
启动
删除
测试报警3
共 1 条规则
每周周三 21:00
暂停中
测试报警2
共 2 条规则
每日 17:00
暂停中
测试报警
共 1 条规则
每周周二 00:00
暂停中

BI应用： 波动归因

媒体分析 / **收入波动分析**

2018年05月25日 [广告位]- 搜索

广告位ID	广告位名称	总收入	昨日收入	日环比变化	上周收入	周环比变化
	测试			↓4.55% (		↓6.45% (↓77112.94)

共 1 条 1 50 条/页 跳至 1 页

指标	当前值	基准值	变化率	收入波动贡献值
dsp填充率			-4.43%	-510
效果填充率			-1.47%	-8
DSP原始ecpm			-24.65%	-59.60
人均请求数			-2.41%	-6.16
效果曝光占比			-0.49%	-1.1
效果ECPC			-0.44%	-1.7
dsp曝光占比			-8.62%	-0.
品牌原始ecpm			↑21.80%	↑1.6
效果Ctr			↑1.65%	
品牌曝光占比			↑55.03%	
曝光留存率			-3.80%	
请求DAU			-4.14%	-10
品牌填充率			-59.78%	-5

应用：数据服务

- **CTR 预估实时模型**
  - 最新1小时内，任一在投广告在不同feature组合下的点击率
- **预算平滑**
  - 最近1小时内，某个广告账户的预算消耗趋势(精确到分钟)
- **受众预估**
  - 过去一周，指定人群定向条件下(性别、年龄、地域...) 广告曝光量和覆盖人数

目录

总结&思考

- 开源or自研：将有限的研发资源投入到最能产生**业务价值**的地方
- 不但要建立数据系统，更要建立**数据标准**
- 架构是随着业务不断**演化**，没有最好的架构，只有最适合的

谢谢！