爱奇艺广告大数据实践

张超-Charles

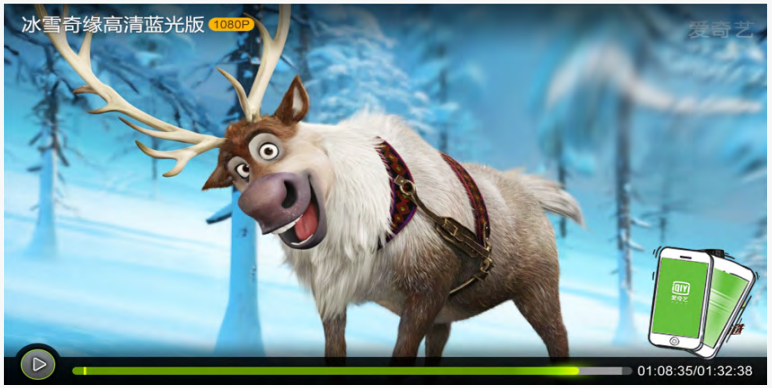
SACC2017

Outline

- 1. 爱奇艺广告简介
- 2. 广告数据的挑战和架构
- 3. 查询引擎
- 4. 数据质量保证
- 5. 实时计算
- 6. 总结

SACC2017

1. 爱奇艺广告简介



- 广告是爱奇艺商业变现的重要手段
- 全终端覆盖：
  - PC（含Mac），Mobile，TV
  - Flash，H5，iOS，Android，站外SDK
- 丰富的展现形式：
  - 30+ 创意模板
  - 贴片，暂停，创可贴，原创贴，TrueView，信息流，Banner，互动贴片
- 复杂的投放形态：
  - CPD，CPM，CPC，CPV，CPDownload
  - 品牌，效果，RTB实时竞价，PDB，内部推广

SACC2017

2. 爱奇艺广告数据应用场景

查询

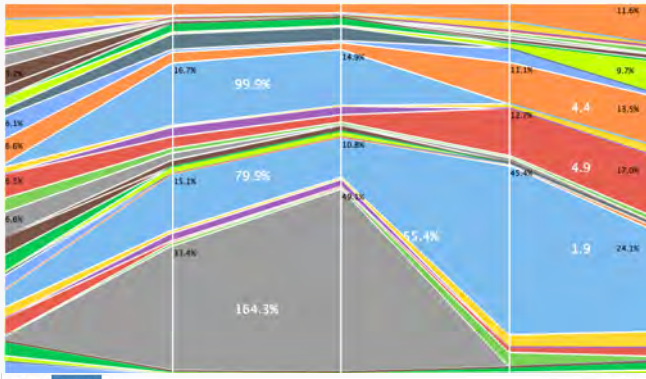
- **数据自助查询：**
  - 广告收入，分成
  - 订单投放效果，库存使用

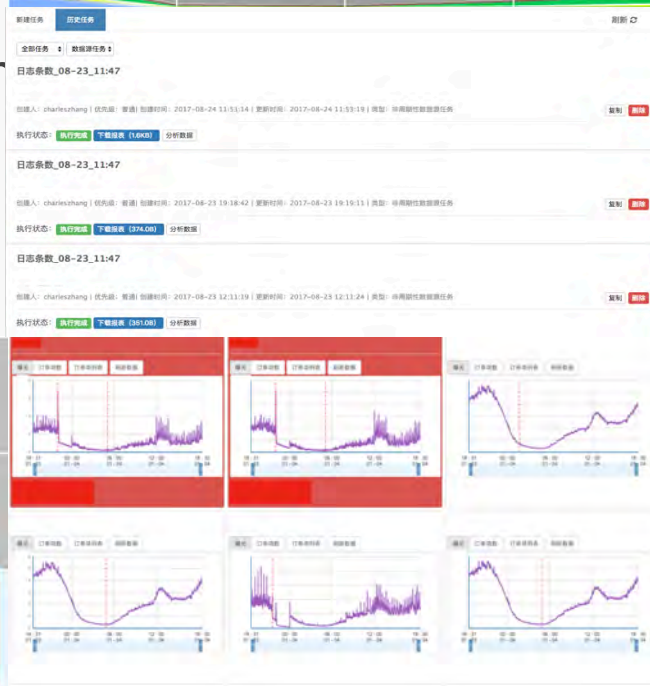
分析

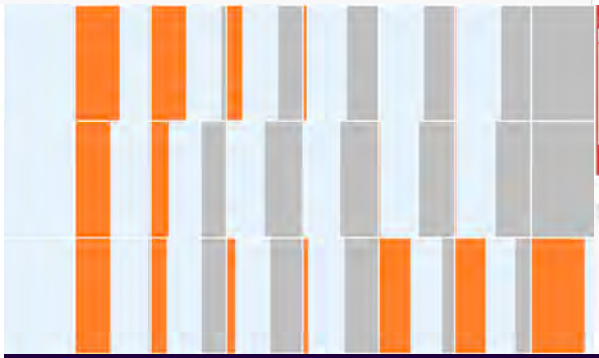
- **可视化分析：**
  - UV转化漏斗
  - Post-buy人群，N+Reach

发现

- **异常检测**
- **数据挖掘**







SACC2017

2. 广告数据的特点和挑战

数据量大

- 日均新增百亿级日志，10T+，存量达PB级
- 单表最高40+个维度，3000亿行数据
- 时间跨度长：需要保存至少2年以上

准确性要求非常高

- 收入结算
- 合作方广告分成
- 库存预估

业务复杂

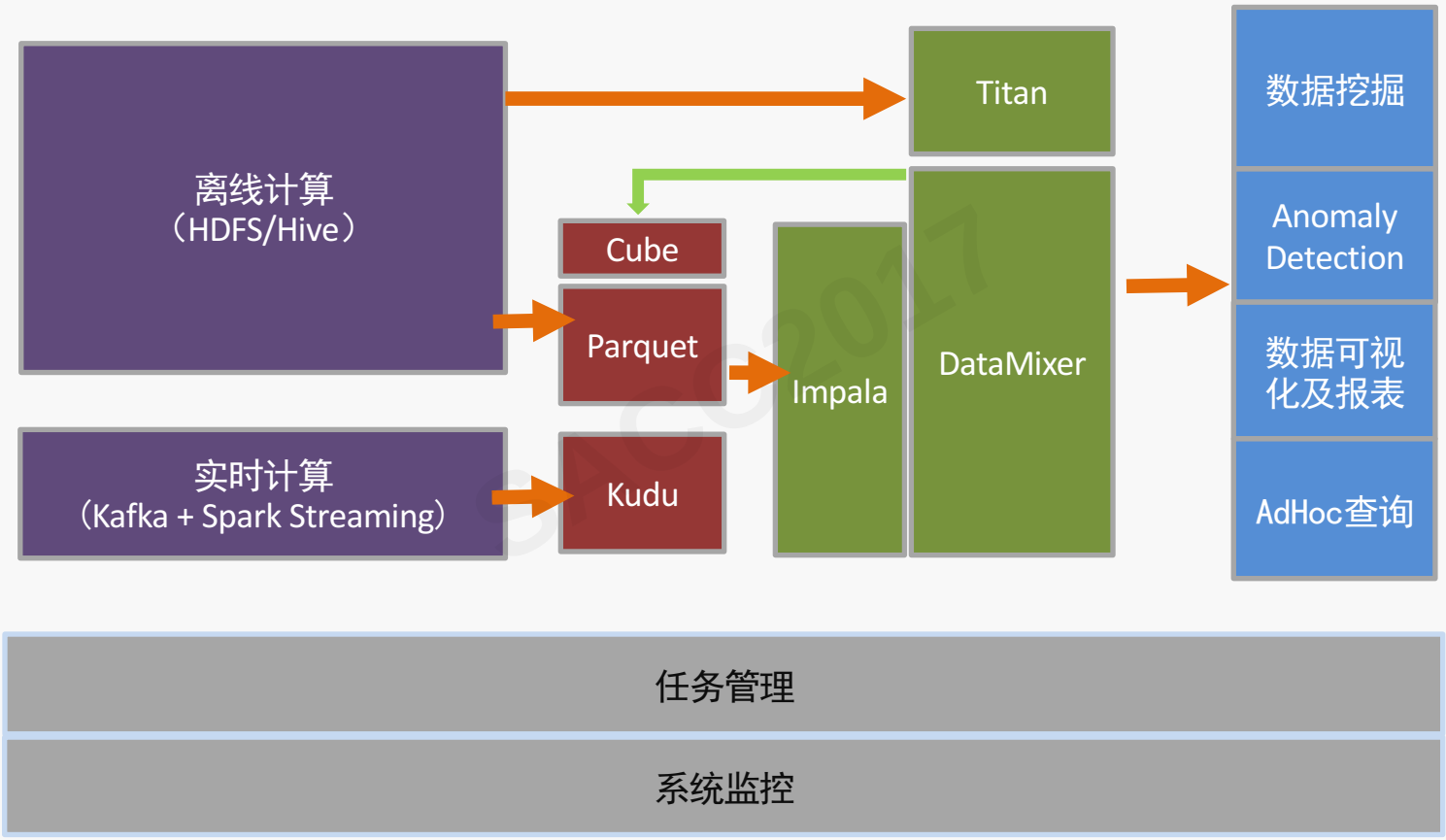
- 竞价 vs 定价
- 品牌 vs 效果
- 数据表300+

业务数据实时可调

- 实时数据需求强烈，广告主依赖数据实时决策
- 数据计算要保证业务数据调整后数据准确性

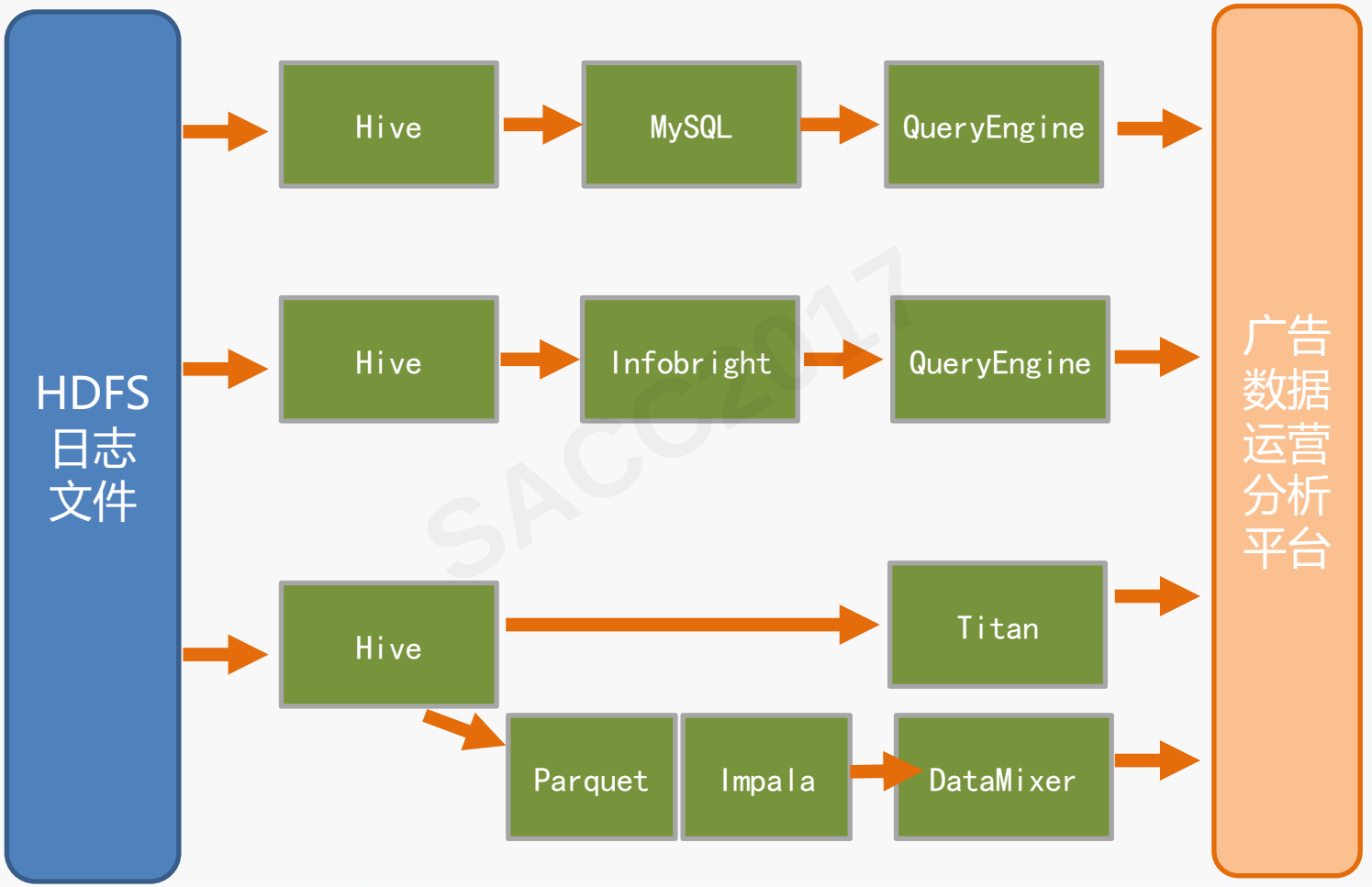
SACC2017

2. 爱奇艺广告数据架构



SACC2017

3. 查询引擎-架构演进

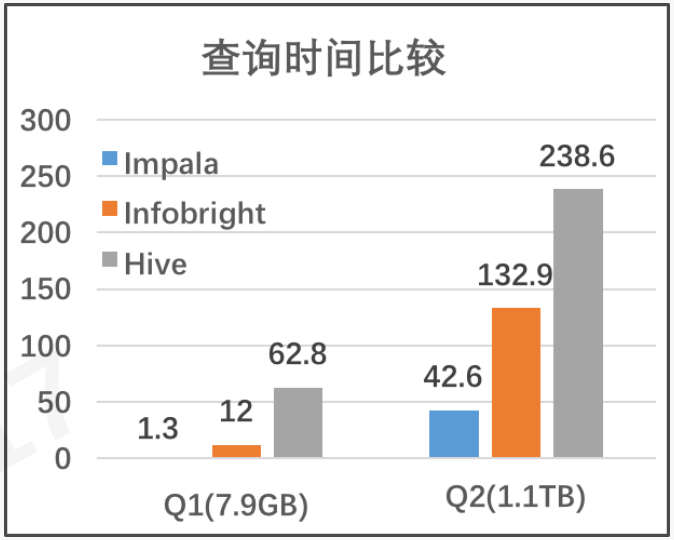


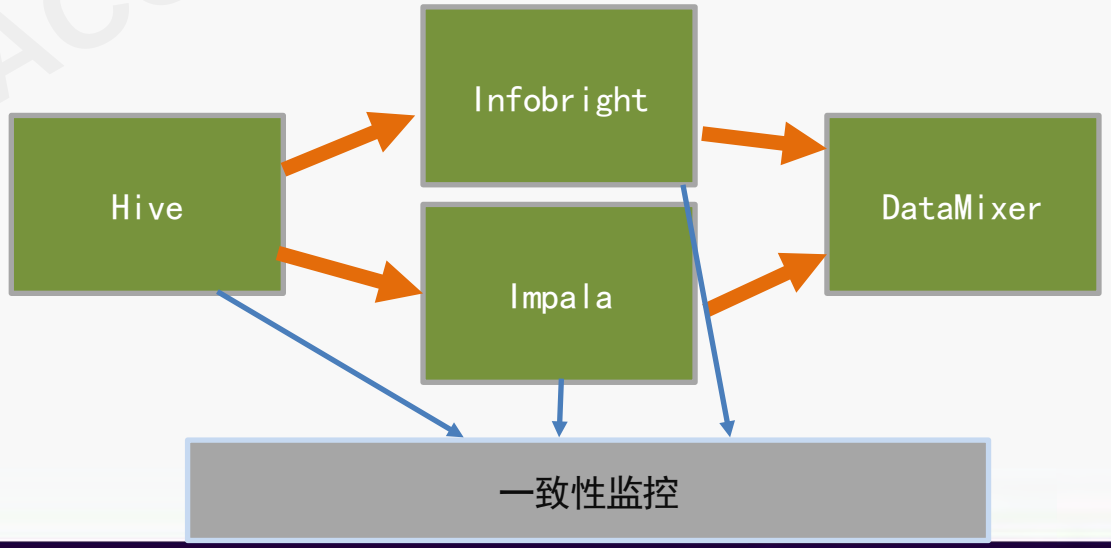
SACC2017

3. 查询引擎 - Impala

- 选型原因：
  - 性能
  - 支持SQL，支持Join
  - 实时与离线统一：无缝支持kudu
  - 水平扩展
  - 与hadoop生态体系兼容

- Infobright-> Impala切换:
  - 双引擎并存
  - DataMixer自动路由
  - 一致性检测

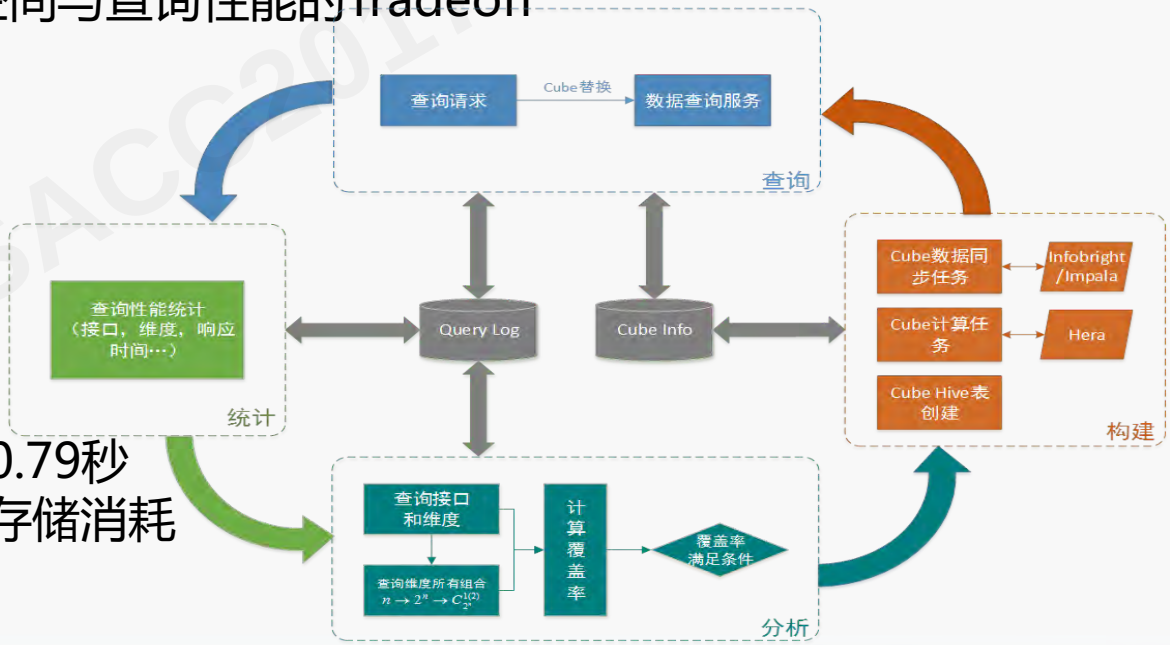




SACC2017

3. 查询引擎 – Cube智能构建

- 预聚合：
  - 把最常用的查询pattern预聚合成聚合表集合（Cube）
  - 空间换时间
- Full Cube vs Partial Cube：
  - Cube构建延迟，空间与查询性能的Tradeoff
  - 2/8原则
- Cube智能构建算法
  - 固定覆盖率
  - 维度组合聚类
- 效果：
  - 整体平均查询性能0.79秒
  - Cube构建以1%的存储消耗换来40%+性能提升



SACC2017

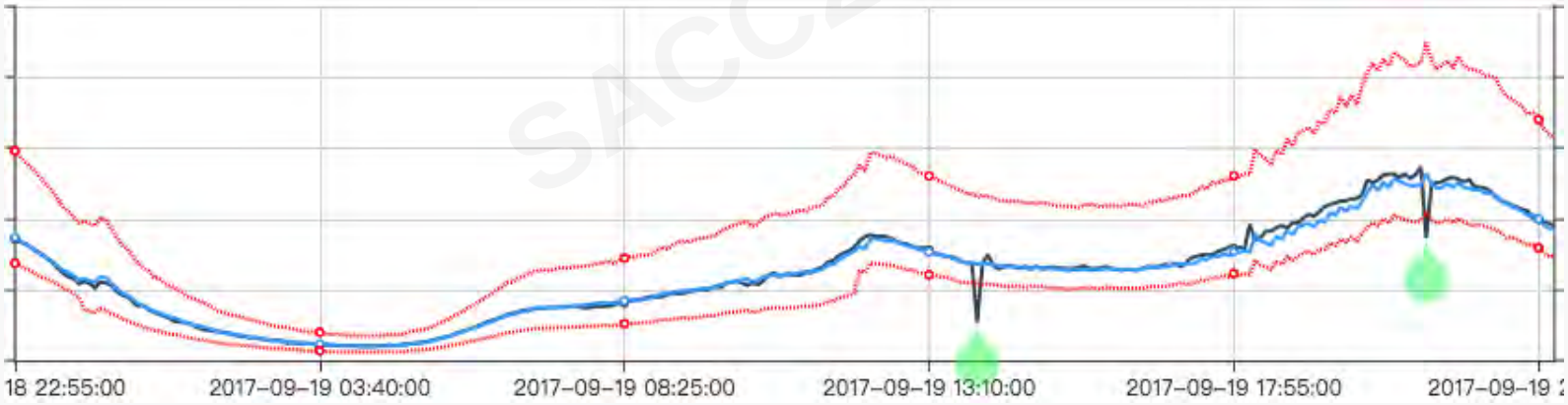
4. 数据准确性保证

- 广告数据直接影响收入及核心决策，准确性是根本
- 端到端数据流，数据出错的风险非常高
  - 数据源端多且复杂
  - 核心业务数据依赖于人工录入，且可变
  - 累积效应：
    - 多个环节的低风险事件累积成高风险

SACC2017

4. 数据准确性保证-准确性判断

- 如何判断准确？
  - 一致性：
    - 不同环节之间：base-》middle-》report
    - 不同计算管道：实时 vs 离线
  - 数据本身趋势合理性：时间序列异常检测

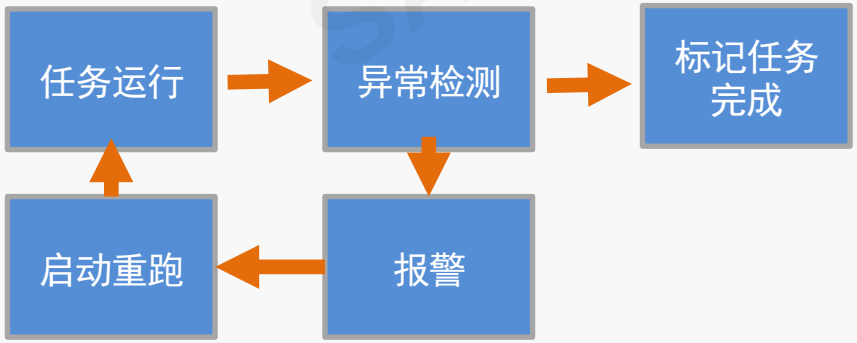


SACC2017

4. 数据准确性保证 – 异常检测

- 检测方法：
  - 恒定阈值：适合较为稳定的指标，如故障率
  - 波动率：适合发现突升突降，但无法区别是否是流量本身上涨引起
  - 动态阈值：基于预测算法，可以有效规避周末效应
    - 3Sigma，Holt-Winters，ARIMA，FB Prophet

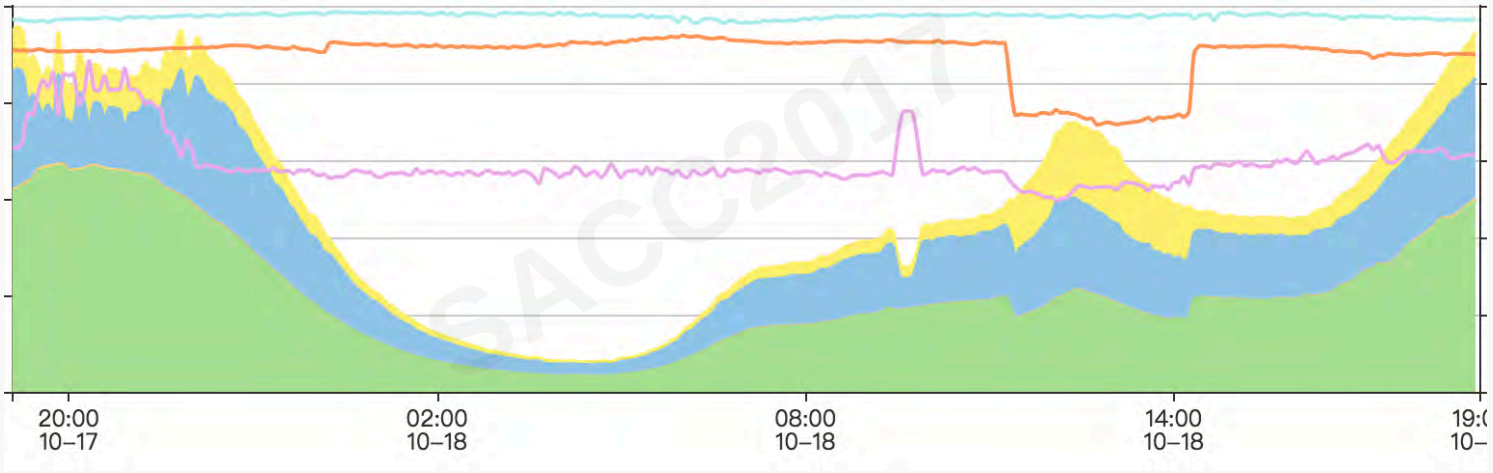
- 应用落地：
  - 自动暂停，提示人工确认
  - 一键重跑



SACC2017

5. 实时计算

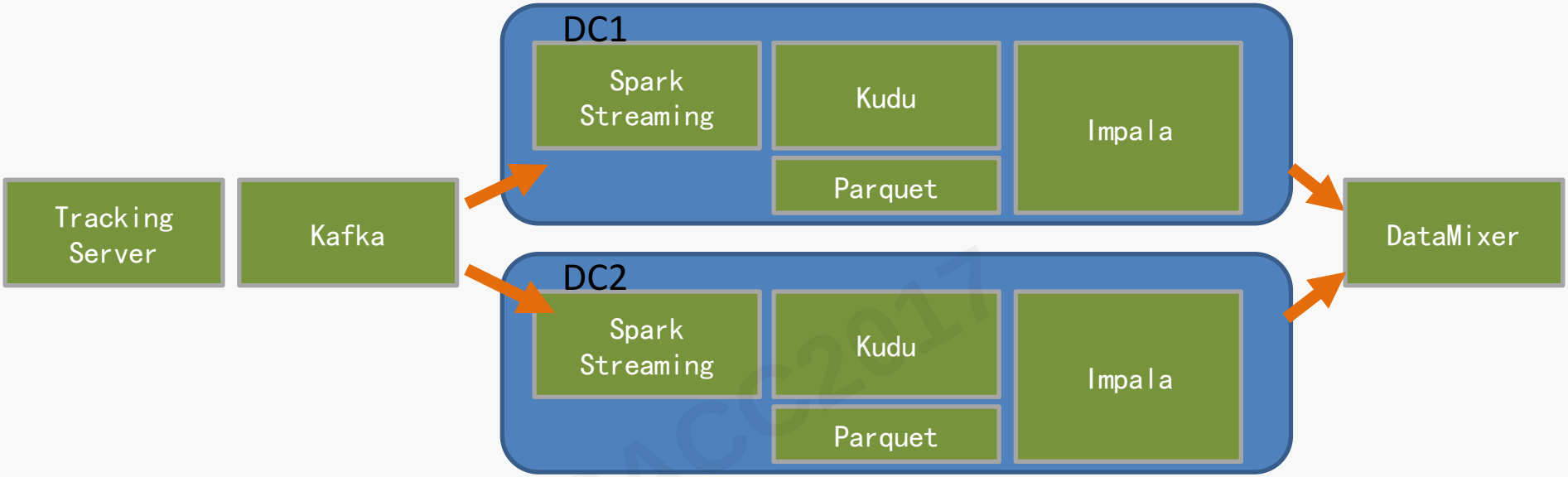
- 实时数据在广告的应用场景：
  - 广告投放效果分析，支持广告主实时决策
  - 运营/运维监控，实时异常检测
  - 实时数据挖掘: 广告主预算预警，CTR预估等



- 挑战：
  - 吞吐量大，尖峰QPS 100W+
  - Exactly Once
  - 数据一致性
  - 高可用

SACC2017

5. 实时计算 – 架构



- Kudu：
  - 优点：写入即可查，聚合查询性能优于ES
- Lambda架构：最终一致性
- 异地双活：
  - 高可用：根据延迟，可用性状态自动切换集群
  - TODO：脏数据自动判断
    - 根据数据突变及一致性自动判断脏数据，切换到健康集群

6. 总结

- 爱奇艺广告数据架构的特点：
  - Impala+Parquet/Kudu
  - Query serving 层： DataMixer+Titan
  - 异地双活保证高可用性
  - 通过异常检测来保证数据正确性

- 展望：
  - 实时和离线在计算层统一
  - DataMixer与Titan整合
  - 新技术框架的探索：Druid，Flink

SACC2017


THANKS