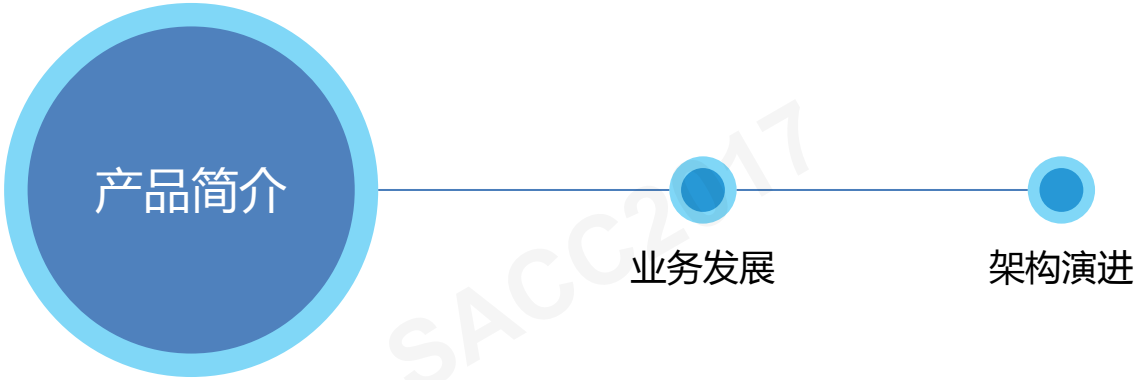
阿里云开放搜索
多租户实时计算架构的演进之路

阿里巴巴搜索事业部 邓万禧

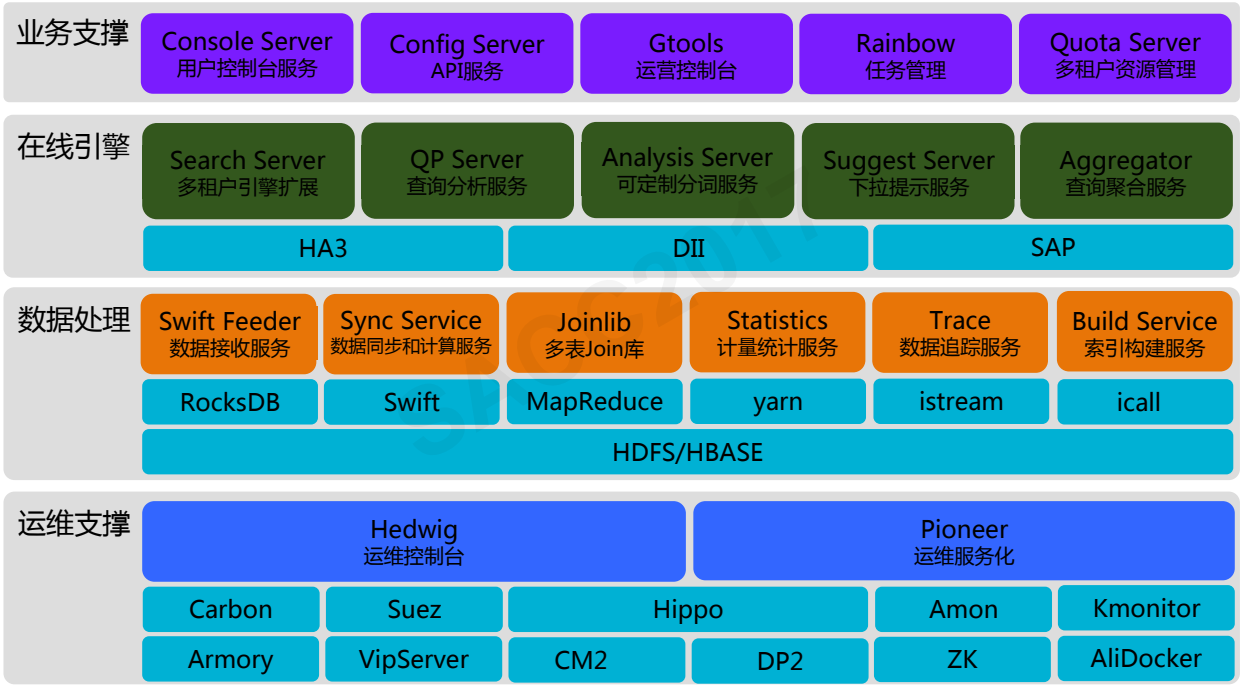
目录



Opensearch简介

- 是什么：完全自助式、可定制搜索托管服务
- 目标：低学习成本、灵活定制
  - 上线只需三步，可能产品经理就能搞定（降低门槛）
  - 数据结构定制、动态修改 (产品5分钟迭代)
  - 相关性定制、线上实时调试（解放算法同学生产力）

Opensearch简介：整体架构

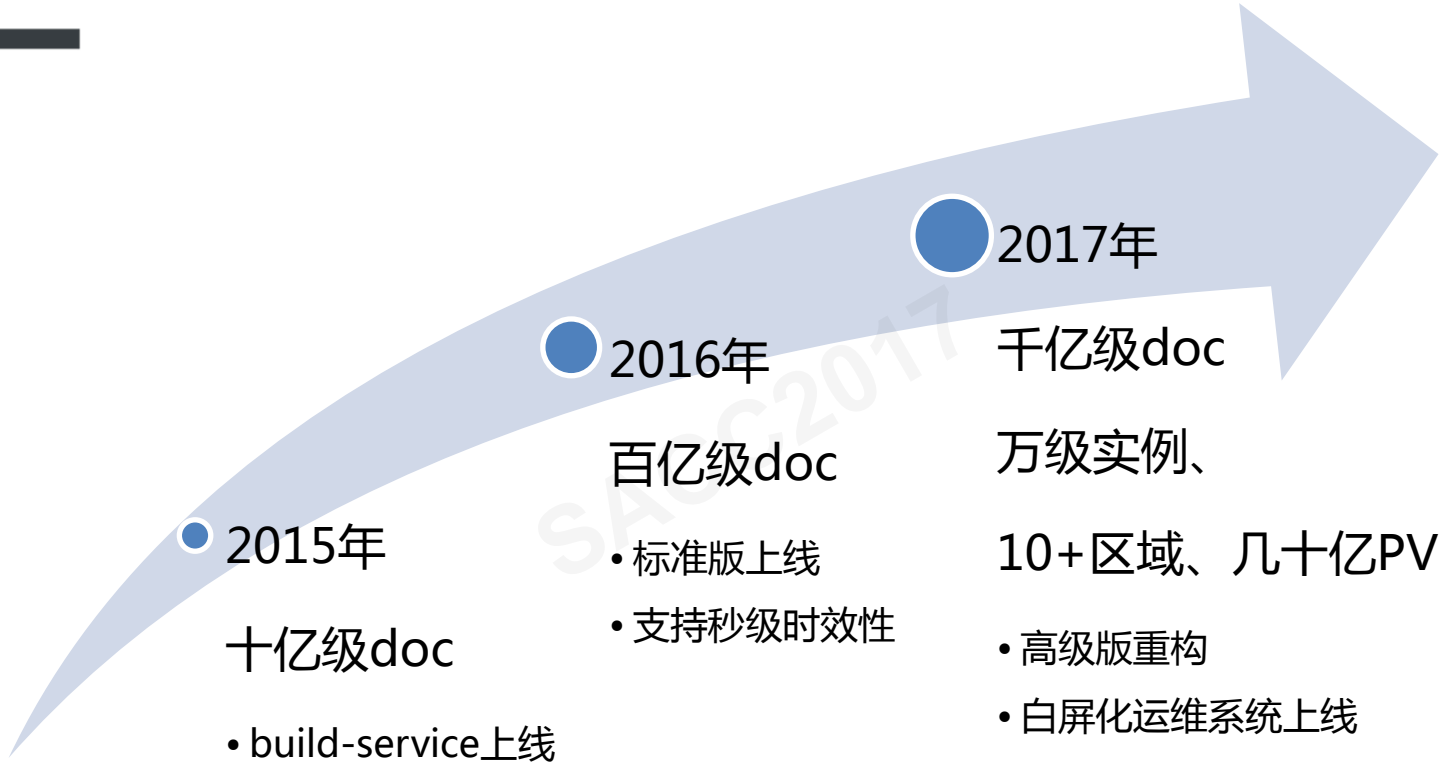


目录

业务发展

阿里巴巴集团内各BU	• 手淘；电商；O2O；APP；....
阿里云公有云业务	• 垂直门户；CRM/大数据；视频媒体；...

业务发展 — 计算架构大事记



2015年

十亿级doc

- build-service上线
- 去daily全量

2016年

百亿级doc

- 标准版上线
- 支持秒级时效性

2017年

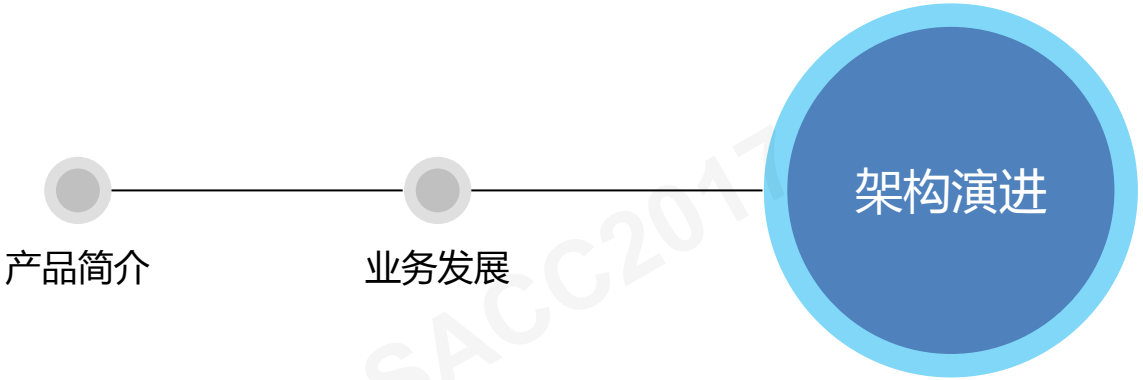
千亿级doc

万级实例、

10+区域、几十亿PV

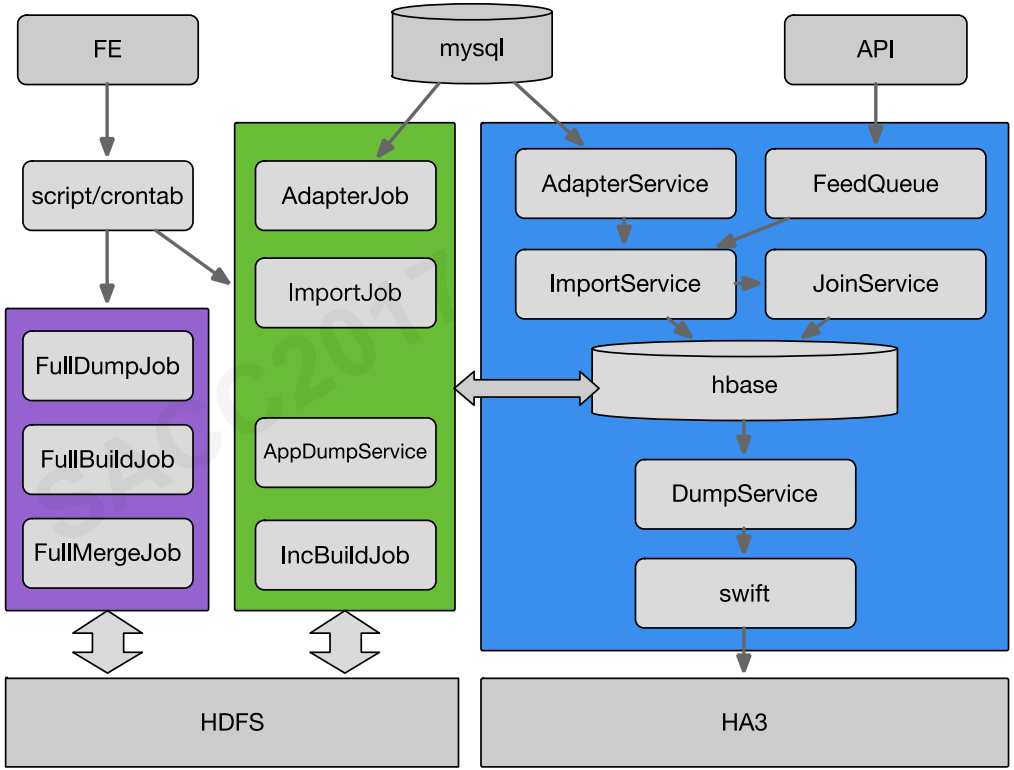
- 高级版重构
- 白屏化运维系统上线

目录



计算架构 — 2014

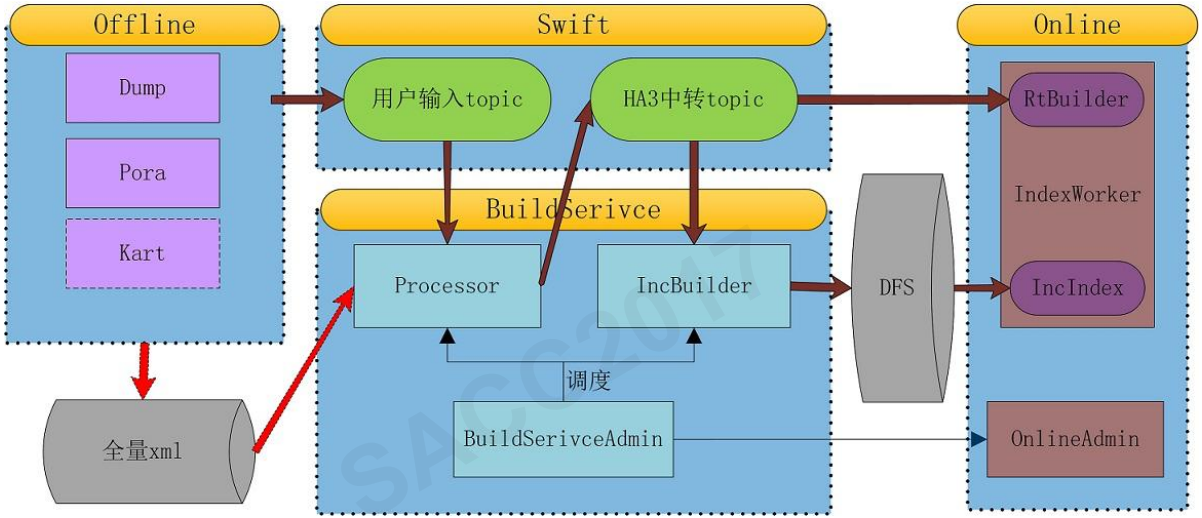
- 基于阿里淘系离线架构
- 分钟级时效性
- 三条数据流



计算架构 — 索引构建任务的问题

- 运维脚本开发维护复杂：
  - 调度：任务起停、全增量切换
  - 与引擎交互：zk信号
- 无法自定义数据源
- 成本高：多个在线引擎需要build多次

去全量 — Build Service



- 使得离线全量、增量、实时数据统一入口。
- 复杂处理操作在processor完成，多备份的情况下能节约机器资源。
- 支持全量阶段并发，加速全量流程。
- 支持多种输入方式（file/swift）

去全量 — 日常全量

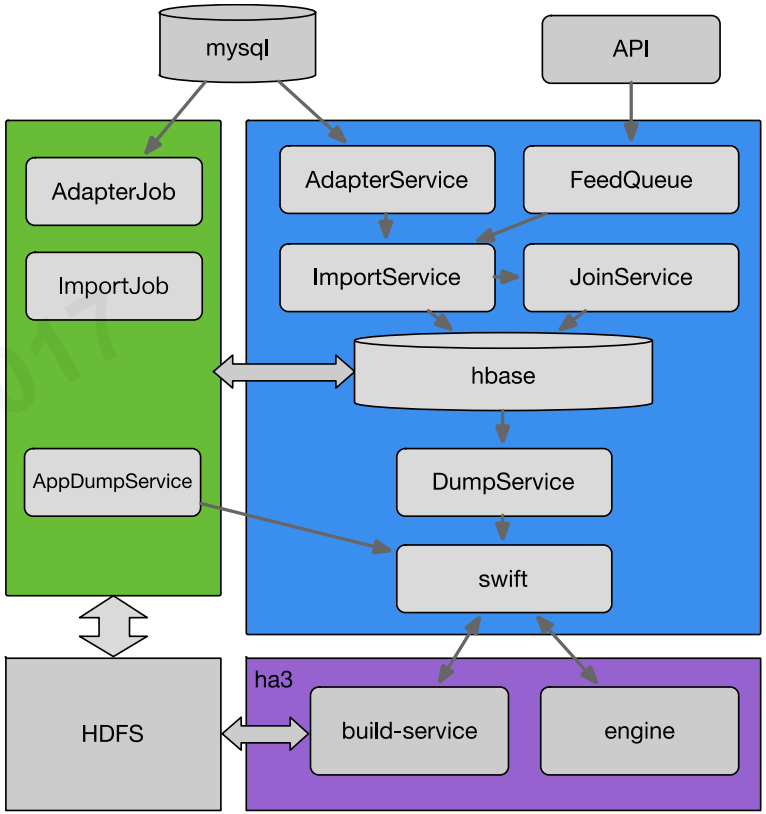
- 与数据回收耦合，按天运行
- 全量运行不稳定
- 对产出时间有严格要求
  - 不及时会影响引擎索引切换
  - 成本高

去全量 — Build Service

- 支持Restore机制：
  - 从引擎索引中恢复数据
- 离线merge回收过期数据
  - merge策略改进：deleted percent + balance tree；
  - 拆分全量segement，减小索引切换时对查询的影响

去全量 — 效果

- 稳定性增加
  - 解决了全量延迟导致时效性延迟问题
  - 全量构建时导致的时效性抖动
- 成本降低
  - 提高了资源利用率，无需为全量预留资源
  - 减小了索引构建的成本

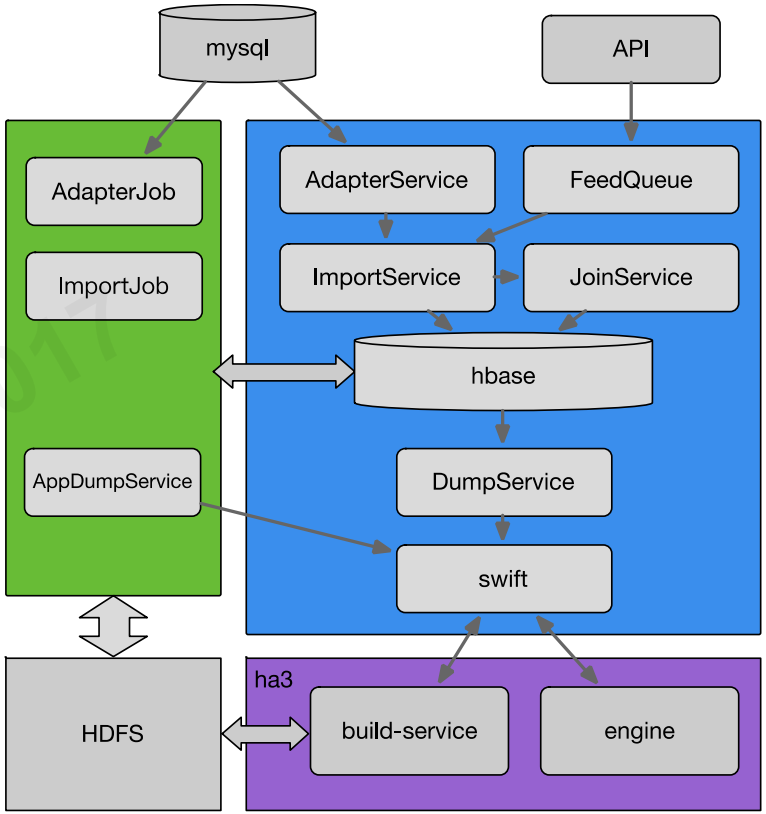


计算架构 — 新的用户需求

- 时效性：秒级
- SLA：99.9%
- 写入流量：双十一每秒百万次

计算架构 — 老架构的问题

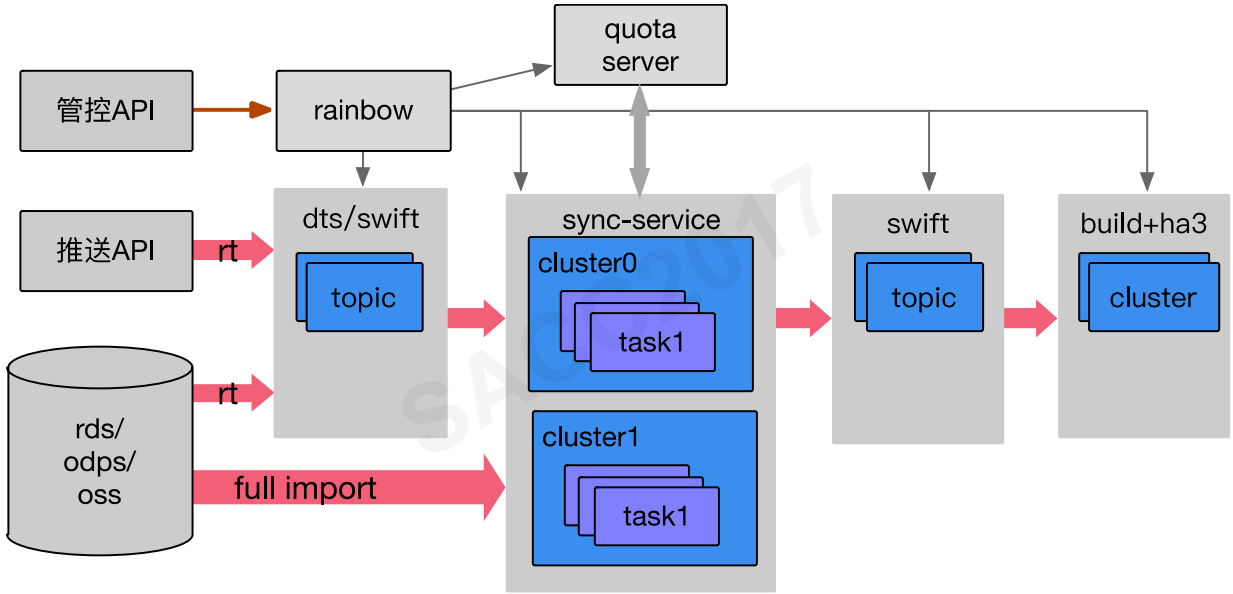
- 多租户互相影响引起的时效性抖动
  - 共享实例流控不完善
  - 无物理隔离的能力
  - 批量和实时间互相影响
- 组件多，无法灵活扩容



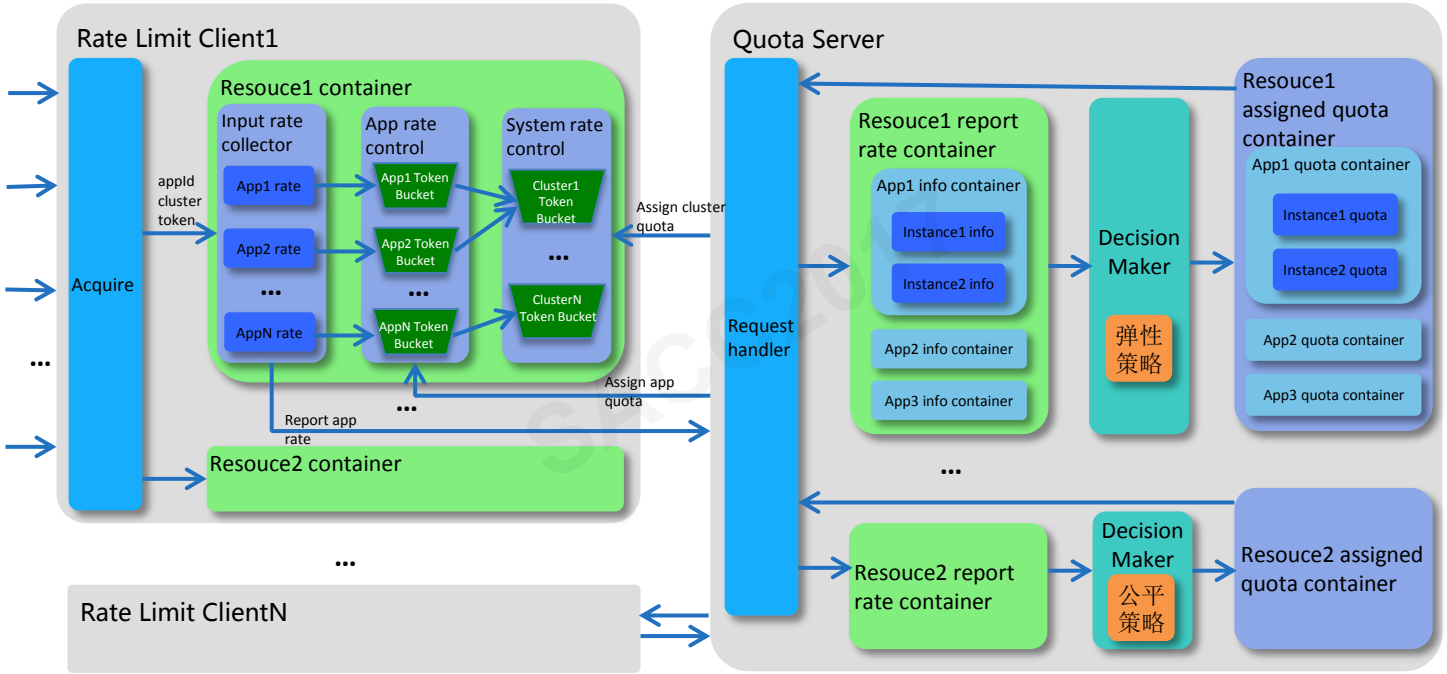
计算架构 — 解决问题的思路

- 多租户资源隔离
  - 模型：两维 cluster + task
  - 独享实例：物理隔离
  - 共享实例：精确流控
- 管控系统重构
  - 统一管理后端组件：包括流计算、存储和引擎
  - 提供Restfull API，支持实例功能的差异化
  - 支持两维调度

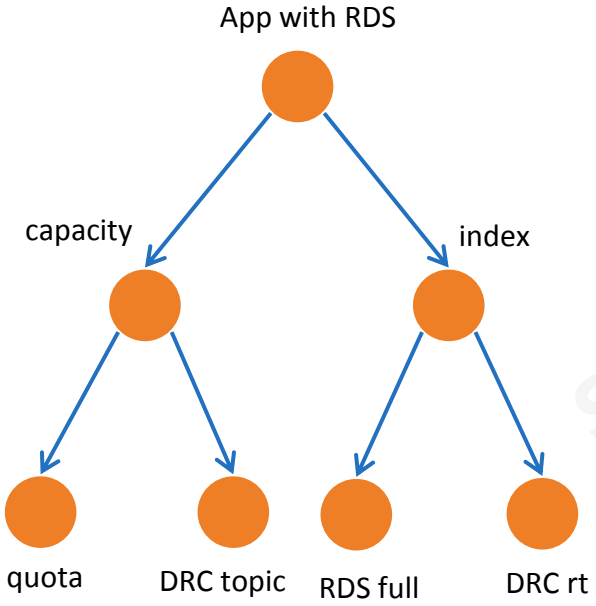
计算架构 — 2016

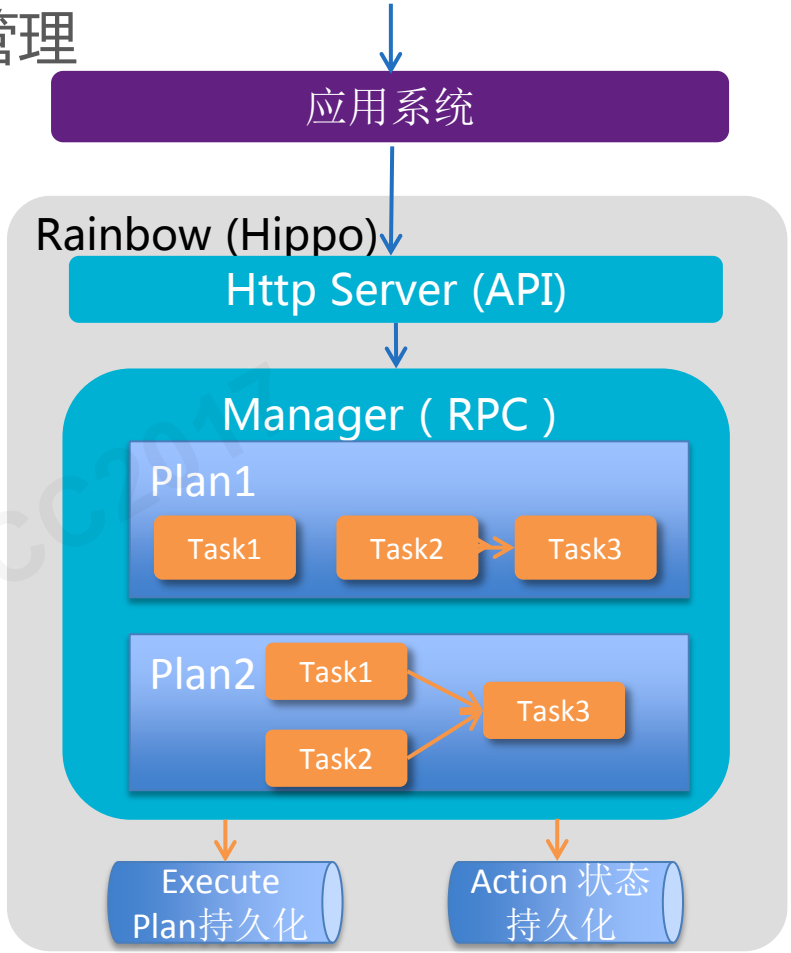

quota
server
管控API
rainbow
推送API
rt
rt
rds/
odps/
oss
full import
dts/swift
topic
sync-service
cluster0
task1
cluster1
task1
swift
topic
build+ha3
cluster

计算架构 — 多租户实时流控的实现



计算架构 — 管控系统之任务管理



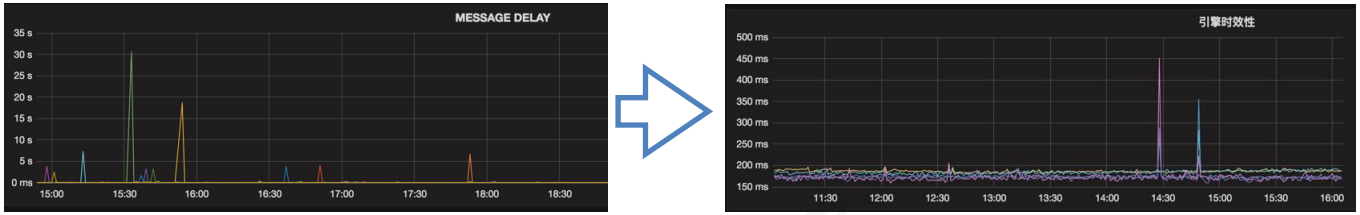


计算架构 — 其他特性

- 实例支持多版本
  - 完善运维支持：平滑升级/回滚
  - 应用场景：

- 更完善的数据导入机制
- 更完善的运维支撑系统

计算架构 — 2016



- 流量平稳：
  - 多个应用间严格隔离
  - SLA：时效性99.9% 1s
- 双十一：
  - 支持了百万级别的写入
  - 整体延迟在200ms左右

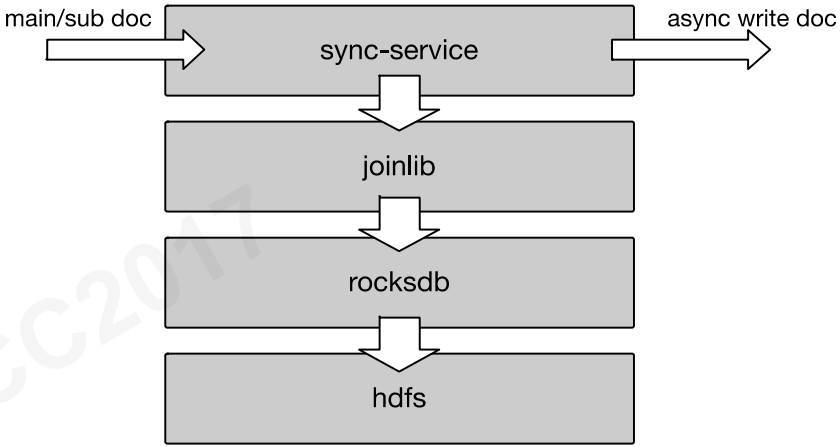
计算架构 — 2017多表join

计算架构 — 2017多表join

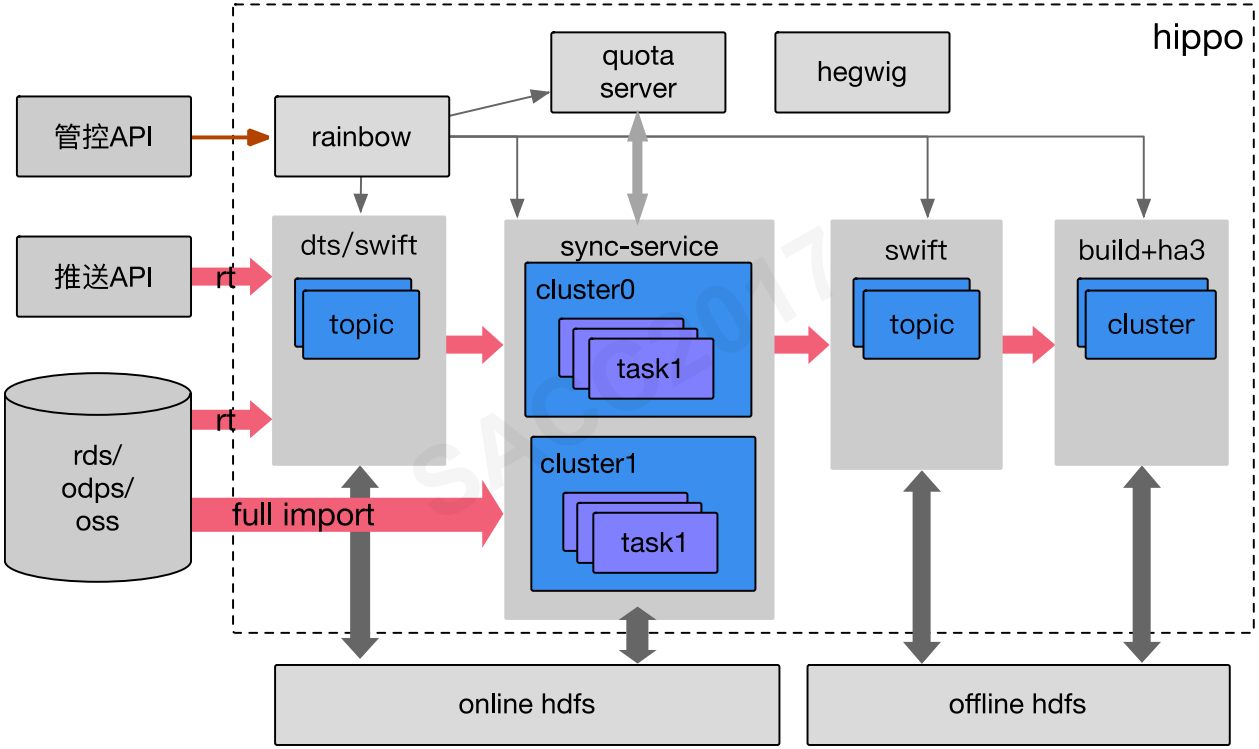
- 基于hbase实现多表join遇到的问题：
  - 时效性抖动：用户数据有热点；rs failover和major compact等等
  - 运维成本高：几十个hbase集群
  - 隔离性：实例间互相影响；功能间互相影响；
- 思路：
  - 租户间需要有存储物理隔离的能力
  - 低成本
  - 运维简单

计算架构 — joinlib

- 可以嵌入到多种流计算框架
  - sync-service or blink
- 存储可选：rocksdb/hbase
- default：rocksdb on hdfs
- 其他特性：
  - 附表更新流控
  - 支持protobuf
  - 基于消息队列和checkpoint实现recover，关闭WAL



计算架构 — 2017


hippo
quota server
hegwig
管控API
rainbow
推送API
rt
dts/swift
topic
sync-service
cluster0
task1
swift
topic
build+ha3
cluster
rds/
odps/
oss
rt
full import
cluster1
task1
online hdfs
offline hdfs

计算架构 — 效果和适用场景

- Sync-Service + Joinlib
  - 效果：
    - 时效性极大提升，接近秒级
    - 多实例间严格隔离，附表、主表更新准确流控
    - 性能极大提升：200台机器->50台
  - 应用场景：适用于中小数据量实例(十几亿级别)

计算架构 — 后续规划

- 计算框架升级：blink
- 存储自定义
- 白屏化运维：
- 聚焦搜索场景：提升相关性定制能力，分词定制/查询分析/相关性定制
- 支持数据分析场景的新产品：

Opensearch：https://www.aliyun.com/product/opensearch
Elasticsearch：https://data.aliyun.com/product/elasticsearch

THANKS