SACC 2016 第八届中国系统架构师大会
SYSTEM ARCHITECT CONFERENCE CHINA 2016
架构 创新之路

爱奇艺大数据平台的构建之路

云平台技术总监
刘俊晖

目录

- 爱奇艺大数据平台的挑战
- 平台的构建之路
  - 1.0 专业化
  - 2.0 规模化
  - 3.0 生态化
- 案例
- 总结

爱奇艺大事记

2016

2月7日
独家直播2016猴年春晚
除夕当晚总播放量突破4500万
4月15日
《太阳的后裔》26亿播放量收官
微博话题阅读量达122亿
微指数峰值达到83万
5月6日
2016爱奇艺世界大会圆满收官
中国首个开放娱乐生态首次展现全貌
构建“爱奇艺世界观”
6月1日
爱奇艺有效VIP会员数已突破2000万

2015

2月18日
羊年春晚独家在线直播
7月6日
《盗墓笔记》全集上线
60小时总播放量破10亿
10月14日
爱奇艺VIP会员品牌全面升级
12月1日
爱奇艺VIP会员突破1000万

2014

7月17日
爱奇艺宣布成立影业公
提出“爱7.1电影大计划”

2013

5月7日
爱奇艺与PPS合并
提供更优质服务

2012

11月2日
爱奇艺
成为百度的
全资子公司

2011

6月23日
“奇艺出品”战略
11月26日
品牌战略升级为
“爱奇艺”

2010

4月22日
视频网站
“奇艺”
正式上线

Beyond the Vision

爱奇艺移动端 核心指标行业领先

移动端数据

- 爱奇艺移动端以**3.3亿**人的月度覆盖位列行业第一，总体占比高达**55%**，行业领先优势持续加大
- 爱奇艺移动端月度总使用次数（活跃度）达**212亿次**，成为视频用户首选

（数据来源：艾瑞MUT，2016年8月）

日均覆盖人数 NO.1

（数据来源：艾瑞MUT，2016年8月）

月度覆盖人数 NO.1

（数据来源：艾瑞MUT，2016年8月）

月度浏览时间 NO.1

（数据来源：艾瑞MUT，2016年8月）

挑战

30x

数据量

2PB+/day

日均处理量

10PB/人

人均运维量

爱奇艺大数据应用

大数据

- TA精算
- 大剧探针
- 爱奇艺指数

VIP服务

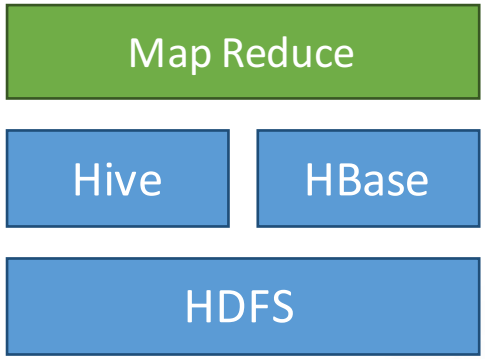
- 后羿（会员精准营销系统）
- 电影探针

广告投放

- 品牌分析
- 众里寻TA
- 剧场受众分析
- 一搜百映
- 追星族、接力赛、群英荟

1.0 专业化

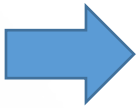
- 时间：2010 ~ 2013
- 规模
  - 集群：50 ~ 330台
  - 存储：1~6PB
  - 计算：
    - 日均作业：3万
    - 日均tasks数：220万
    - 日处理数据：150TB
- 开源服务
  - HDFS、MapReduce、Hive、HBase



1.0 专业化

痛点

- 业务自己维护集群
- 运维不规范
- 半监控状态
- 小文件多/存储压力大
- Jobtracker性能瓶颈



方案

- 集中到云平台管理
- 脚本化、流程
- Ganglia + Nagios
- Name/Space Quota

(1.3亿小文件，2PB冷数据）

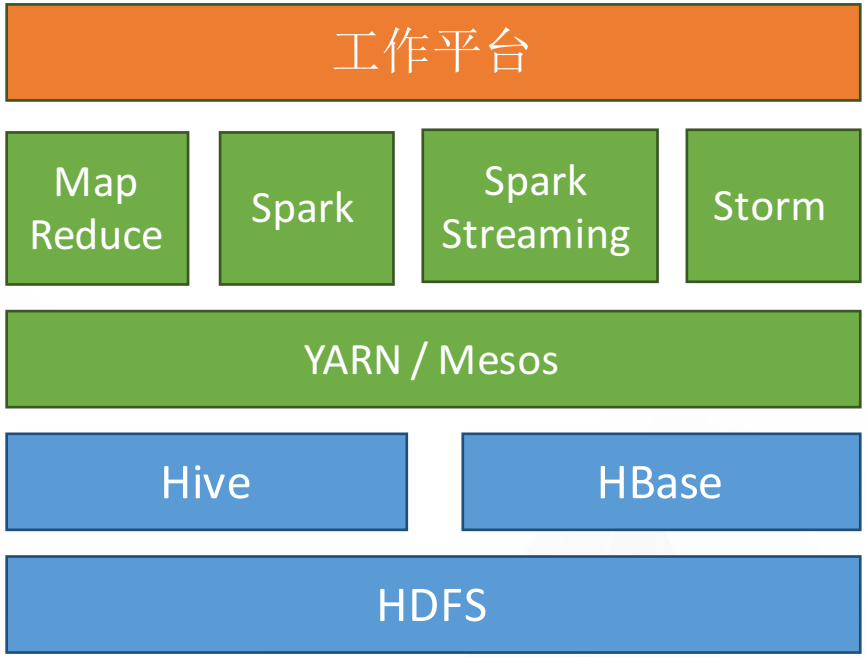
- 改源码，JT任务调度加快12倍

JobTracker调度性能差

- Hadoop 1.x + FairScheduler
- 同时运行的任务多 → 调度时间＞60ms → 心跳延迟大
- 解决方案：
  - 修改FairScheduler源代码，一次排序分配多个任务
  - 修改后调度时间＜5ms

2.0 规模化

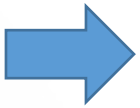
- 时间：2014 ~ 2015
- 规模
  - 集群：1000+台
  - 存储：~30PB
  - 计算：
    - 日均作业：~8万
    - 日均tasks数：~1800万
    - 日处理数据：~900TB
- 开源服务
  - HDFS、MapReduce、Hive、HBase
  - Spark、Storm



2.0 规模化

痛点

- 运维脚本较散乱
- 存储成本骤增
- 离线服务延迟大
- 资源利用率不够高
- 权限控制不够
- 故障处理慢



方案

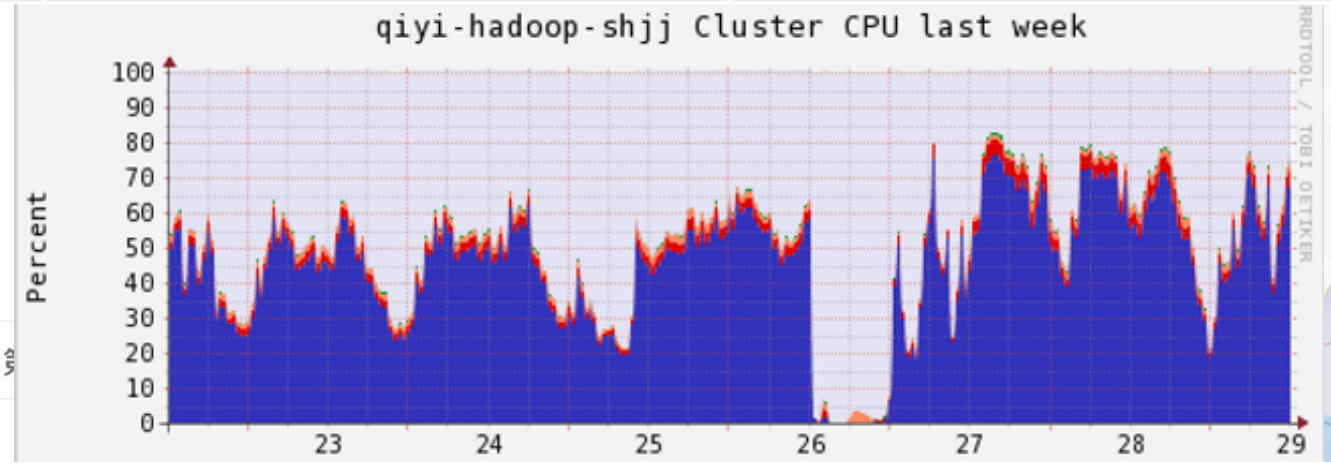
- Hadoop工作平台
- Parquet+gz组合（省20%）
- 推广Spark、Storm
- 升级到Yarn（提高21%）
- Kerberos、HDFS ACL
- 源码解决（贡献了45+ Patch）

Hadoop工作平台

- 后台管理（CMDB）
  - 集群、服务器、配置、用户等
- 运维管理
  - 运维操作Web化，配置与脚本分离
  - 脚本Ansible为主，Python为辅
- 数据管理
  - 数据注册与发现
  - Metadata API
- 公共库管理
  - Hive UDF

YARN升级

对比项	Hadoop 1.0	Hadoop 2.0 (YARN)
计算框架	只支持MapReduce	YARN成为一个通用的资源管理系统，支持MapReduce、Spark、Storm等目前比较流行的计算框架，甚至还允许用户自定义计算框架
调度	JobTracker负责所有的任务调度，负担较重	双层调度：应用调度、应用内tasks调度。 将应用内调度交给应用自己负责，减轻调度器负担
资源隔离	将资源简单地划分为slot，比如1 slot = (1CPU, 2GB Mem)； 将slot资源人为地划分为map、reduce，不适用于动态变化的生产环境	使用LXC进行隔离，用户可以自己申请需要多少资源，更加灵活、更充分利用； 2.6开始支持Docker
Availability	单点	HA
作业运行时间	475 s	201 s （降低57.7%）
资源利用率 (min/avg/max)	CPU: 15.4% / 40.7% / 61.2% Memory: 35.7% / 41.4% / 45%	CPU: 20.4% / 49% / 73.9% （高峰时提高21%） Memory: 24% / 33% / 40% （高峰时下降11%）



Spark in IQIYI

- 部署方式
  - Standalone（虚机为主）
  - Spark on Yarn （占Yarn20~30%资源）
- Spark on Yarn
  - 优点：资源共享、扩容方便
  - 缺点：对于实时任务，会受大集群波动影响
  - 优化
    - 使用yarn-cluster模式
    - 禁止YARN重启应用，用户通过平台（Europa）控制重试策略
    - 开发了Spark访问HBase的Kerberos验证，并定期刷新Token
    - 根据执行器的核数自动配置GC策略

Spark 算法优化

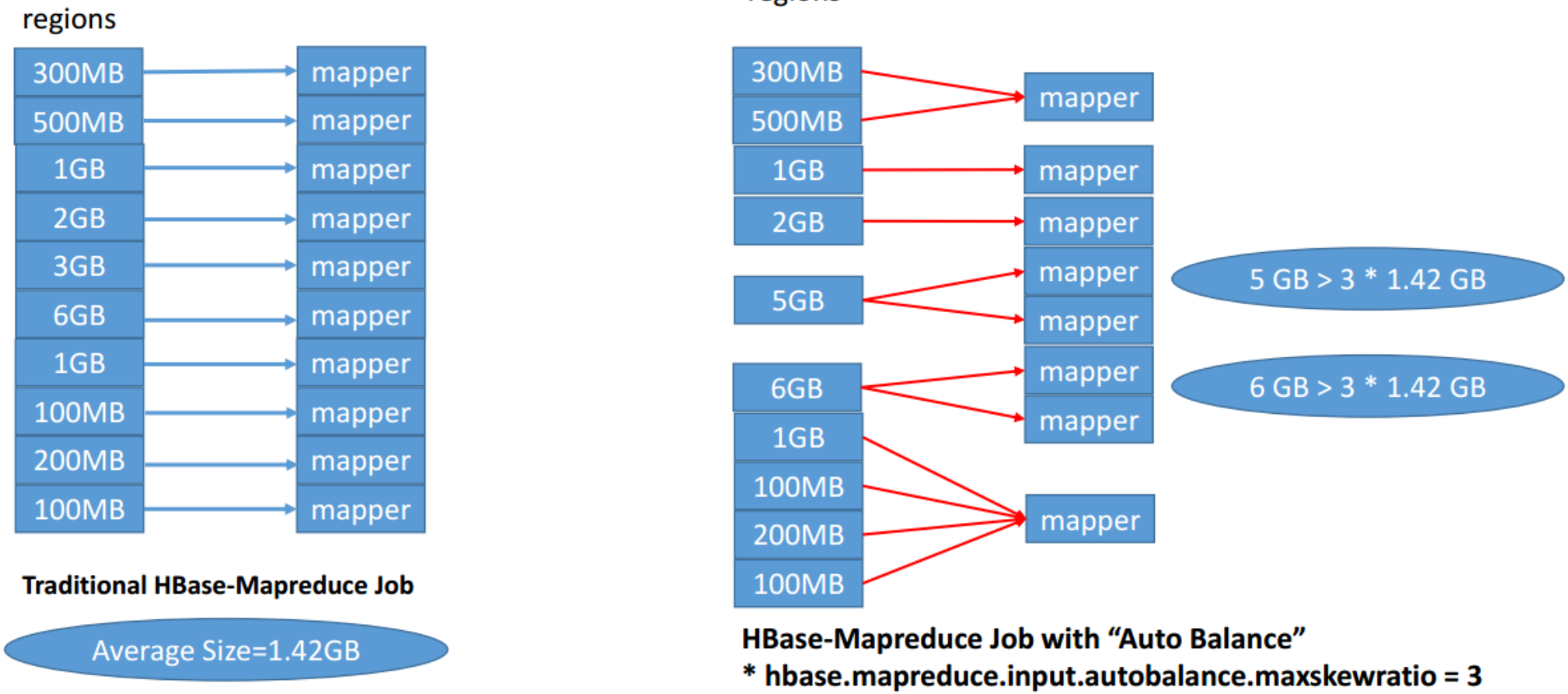
- Hadoop Mahout -> Spark Mllib
  - Logistic Regression / Decision Tree / LDA / ALS
- 算法实现优化
  - LR / ALS / FP-growth ： 实现调优和BUG修复
  - All Pair Similarity Search(APSS)
    - 计算出每个item最相似的TOP K个item，并返回它们之间的相似度
    - 实测性能大约对600万个item的集合计算两两相似度，并返回每个item的TOP 16
    - 2小时 -> 20分钟

社区贡献

- 爱奇艺向Apache Hadoop社区贡献45+ Patches，如：
  - HDFS-7798：解决Checkpoint失败问题
  - HDFS-8113：解决block report失败问题
  - YARN-3024：提高localization效率
  - YARN-3266：解决NodeManager识别问题
  - HIVE-11149：解决PerfLogger引起的Hive任务卡住
  - HBASE-12590：降低了数据倾斜的HBase表对计算资源的浪费

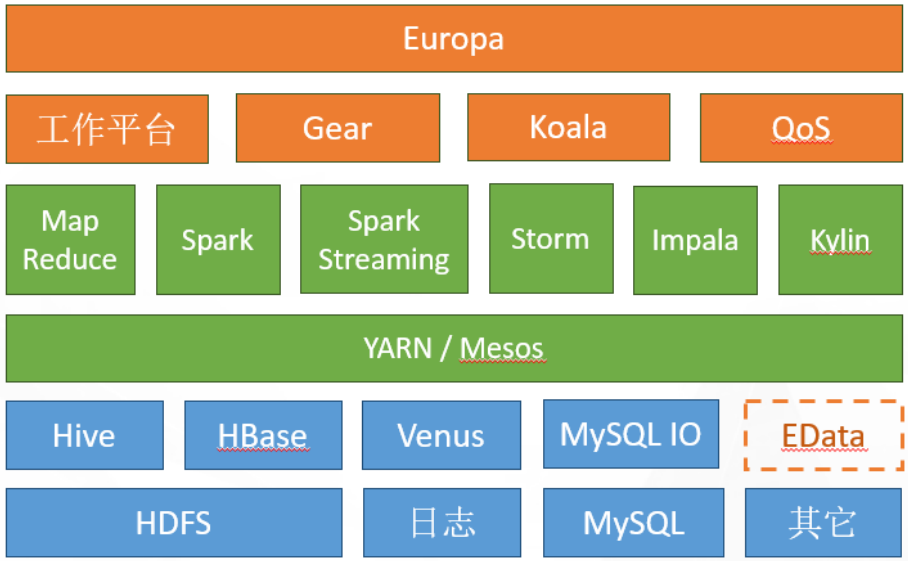
HBASE-12590

- For example, if we have a table with 10 regions, the average region size is 1.42 GB.



3.0 生态化

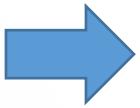
- 时间：2015 ~ 2016
- 规模
  - 集群：2000+台
  - 存储：~60PB
  - 计算
    - 日均作业：~15万
    - 日均tasks数：~4000万
    - 日处理数据：2PB+
- 开源服务
  - HDFS / MapReduce / Hive / HBase
  - Spark / Storm
  - OLAP : Impala / Kylin
- 自研系统
  - QoS / Koala / Gear / Venus / Europa



3.0 生态化

痛点

- OLAP查询慢
- 监控报警不够精准
- 运维成本高
- Crontab不易管理
- 日志不易管理、排障难
- 易用性不足、门槛高



方案

- Impala、Kylin
- Hadoop QoS
- Koala自动化运维系统
- Gear工作流管理系统
- Venus 日志收集计算平台
- Europa 大数据开发平台

Impala

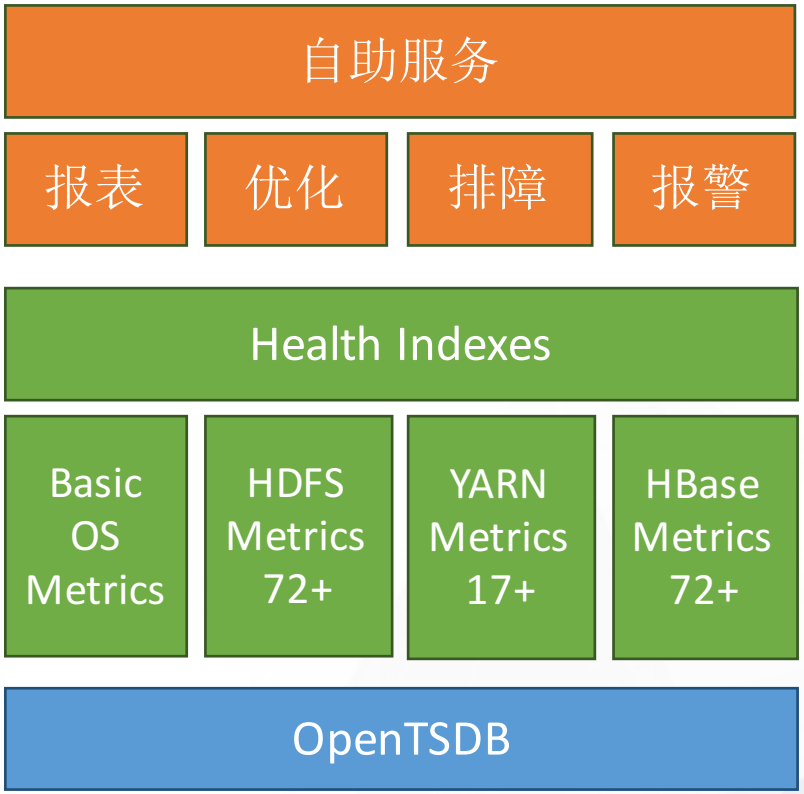
- 产品优势
  - 快速响应：通常为秒级到几分钟
  - Hadoop生态：方便集成、交互数据
  - SQL接口：使用简单
- 实际场景
  - 报表系统：Hive查询几十分钟
  - 广告：MySQL超过TB后很慢
- 测试结果
  - Impala比Hive快10倍；
  - 支持TB数据规模；

Kylin

- 产品特点
  - 优势：查询时间在秒级
  - 原理：空间换时间，预先计算每种维度组合的测量值
  - 限制：查询维度不宜过多，维度和测量指标需预定义
- 实际场景
  - 报表团队：行为表，每天产生3.5TB数据，Hive预处理存到MySQL，时间超过1天
- 测试结果
  - Kylin查询时间在1s以下
  - Kylin构建时间2.5小时（10倍）
  - KylinCube大小9GB（0.76%）

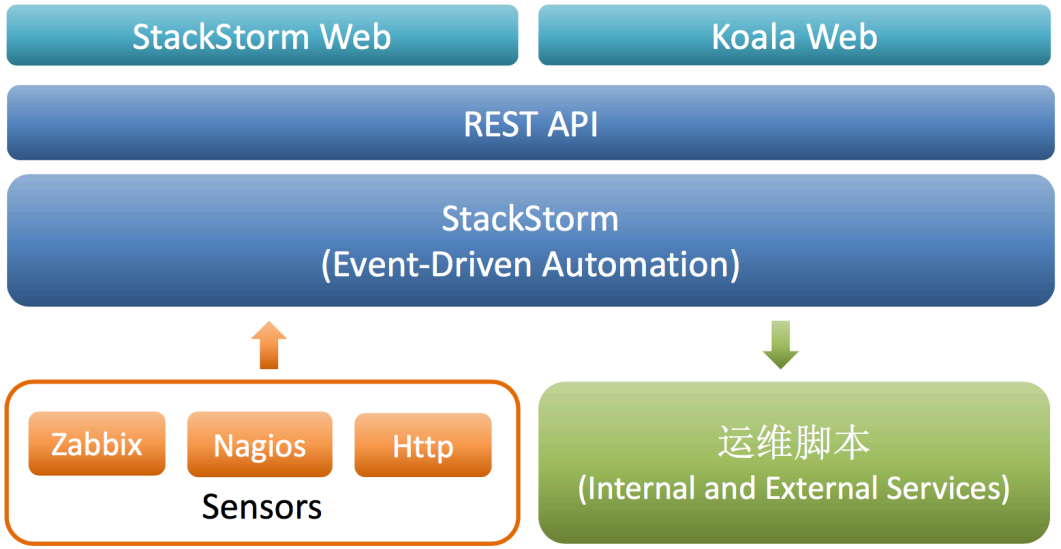
Hadoop QoS

- Ganglia缺点
  - 指标过多而且分散
  - 不能排序和聚合
  - 非功能导向
- 主要功能
  - 报表
    - HDFS、HBase、YARN健康度
  - 优化
    - HBase Region的热点分析
  - 排障
    - HBase的callQueue排查
  - 报警
    - 当前Inode数 / 预估最大Inodes数 > 85%



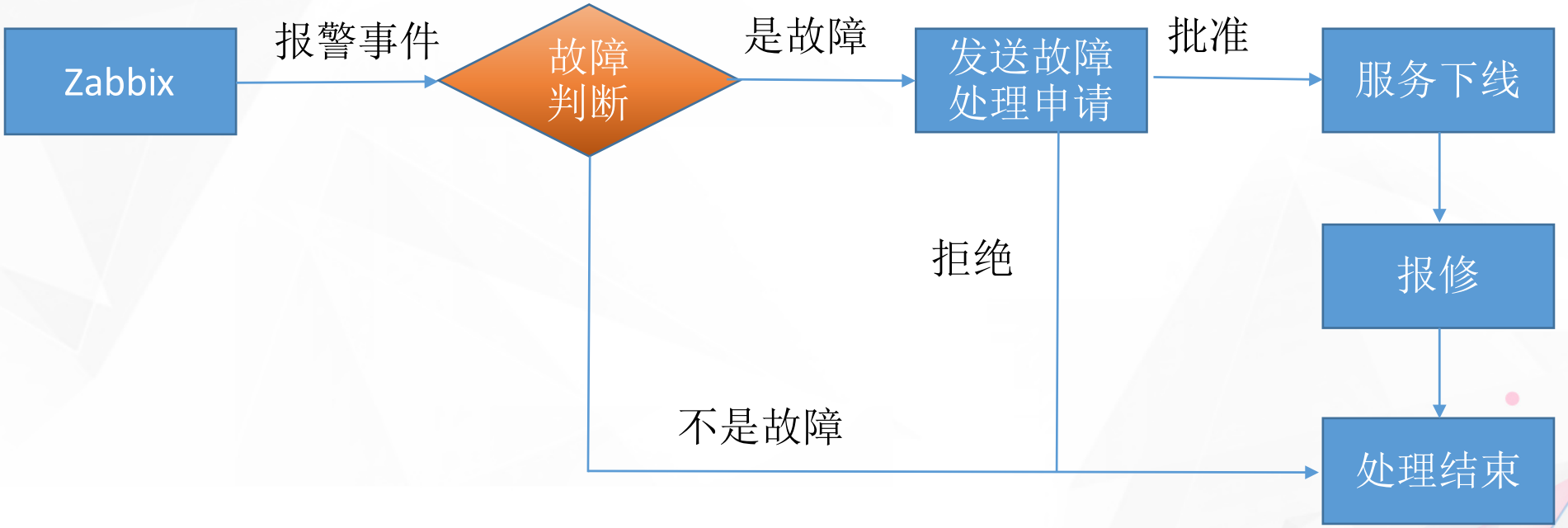
Koala 自动化运维系统

- 报警事件的自动处理
  - 快速响应
  - 无人值守
  - 失败后通知
- 运维脚本的统一管理
  - Gerrit / Gitlab
  - Web / RestAPI
- 审计
  - 操作记录的存储、查询、统计



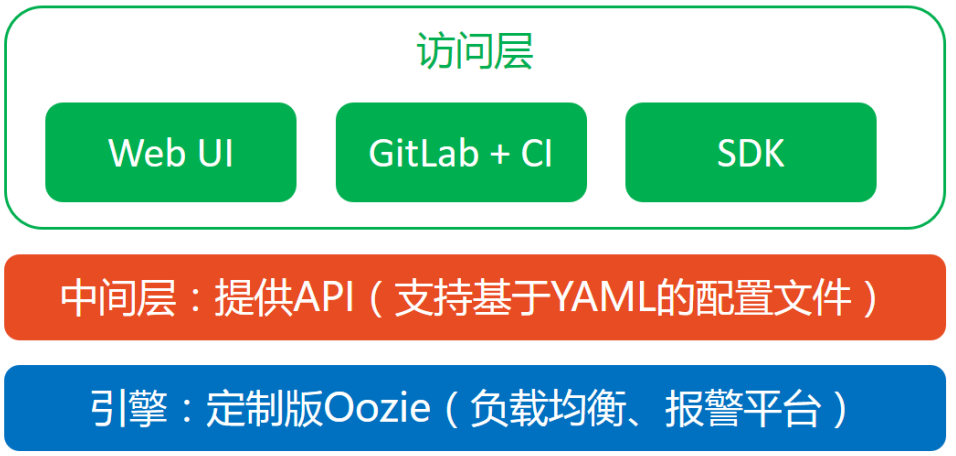
Koala 自动化运维系统

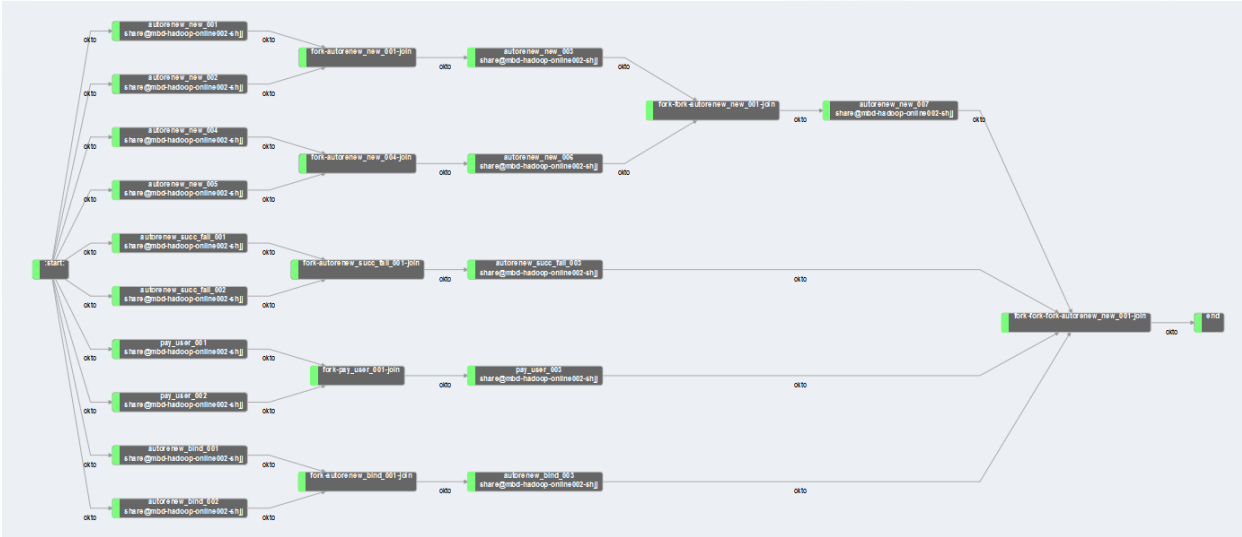
- 硬件故障自动报修



Gear工作流管理系统

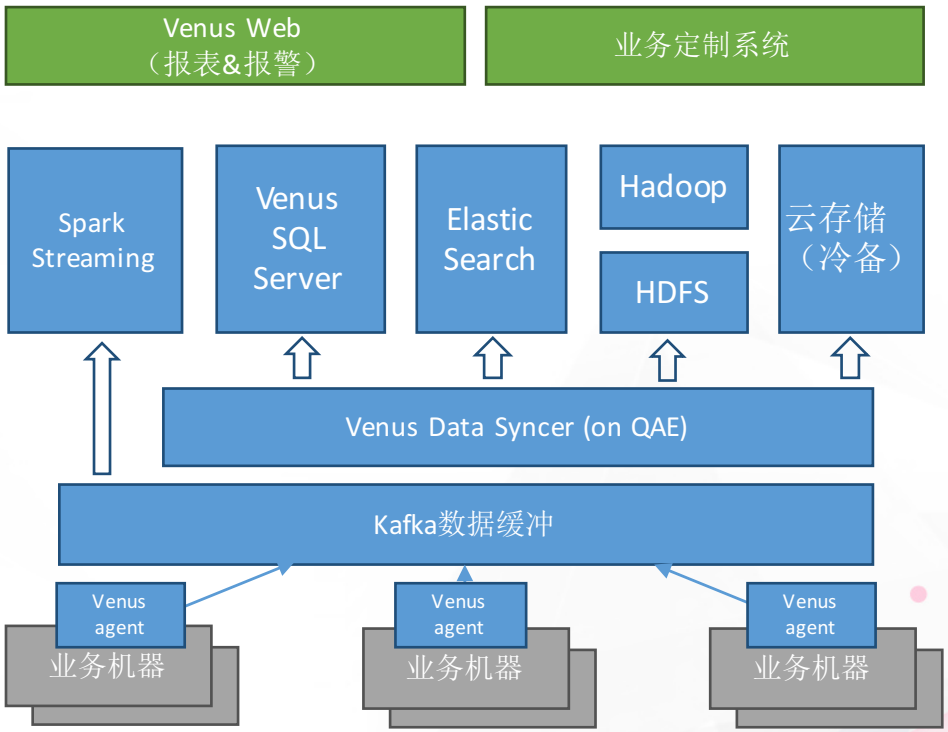
- 基本功能
  - 作业管理、定时启动
  - 依赖管理、重试机制
- 特色功能
  - 基于YAML的配置文件
  - 使用GitLab管理，自动提交
  - 报警订阅
  - 自定义报警接口
  - 任务机负载均衡
- 业务应用
  - 已上线80个项目，360个工作流，1500+个任务





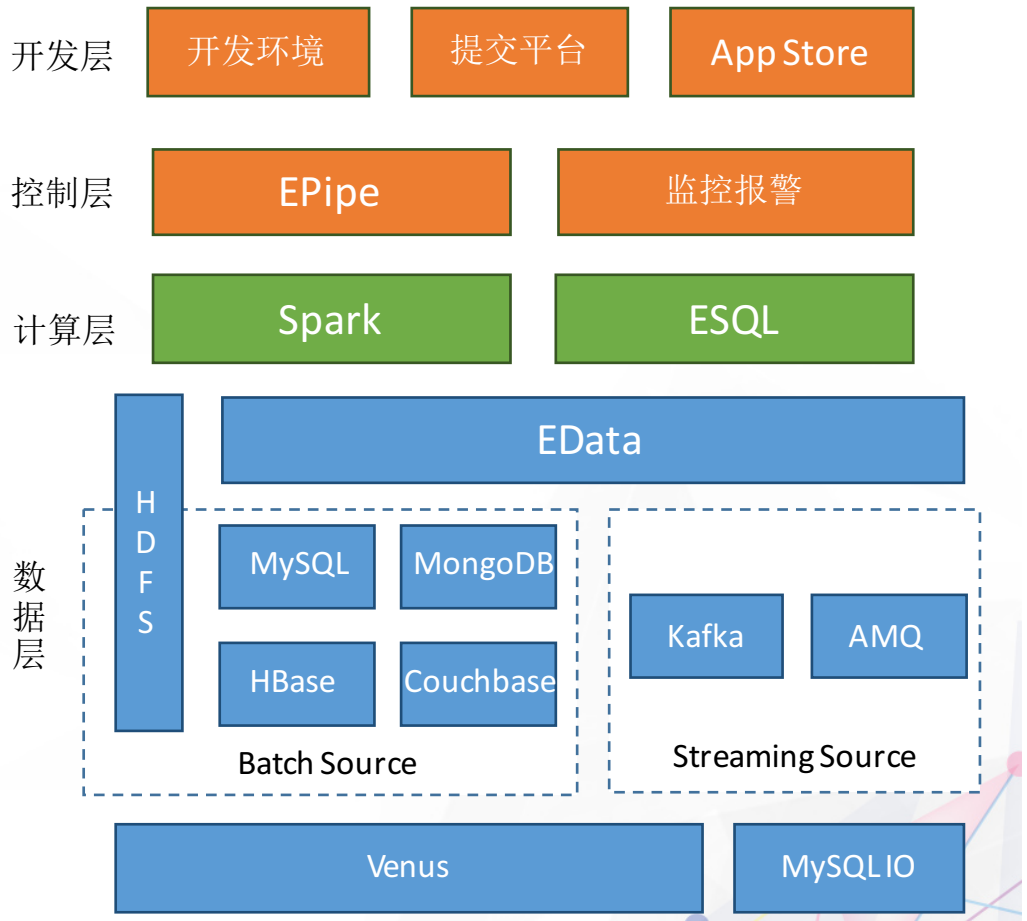
Venus 实时数据收集计算平台

- 主要功能
  - 公司统一的日志入口（Pingback+机器日志）
  - 一站式的收集、分析、报表展示方案
  - 自助服务
  - 全链路分析
  - Venus SQL Server
- 目前状况
  - 核心业务全部接入
  - 接入机器数3000+
  - 每日数据吞吐180TB+
  - 峰值330万条日志/秒
- 典型应用
  - 线上排障（播放、会员）
  - 大数据分析（安全）
  - Docker日志
  - 重要日志备份



Europa – 大数据开发平台

- 目标：提高大数据分析的效率
- 开发层
  - 提交方式：网页、命令行、SDK、Maven插件
  - App Store：共享和构建
- 控制层
  - Epipe：基于Gear工作流管理系统
  - 监控报警：错误检测、报警订阅、资源审计
- 计算层
  - 高级用户使用Spark
  - 入门级用户使用ESQL
- 数据层
  - Edata方便操作HDFS以外的数据源（官方库有BUG且接口不统一）



Europa – ESQL

- 痛点：Spark学习成本高、调试困难、维护难
- 方案：ESQL提供了配置文件+SQL的方式编写Spark程序
- 案例：Venus标准化日志 -> 匹配 -> 存到HDFS
- 对比：数百行代码 V.S. 几行配置 + SQL

```
"test": {
  "desc": "从Kafka读取json格式数据，通过正则表达式抓取字符串，并以parquet格式保存到HDFS",
  "strategy": "SparkStreamingStrategy",
  "compositor": [
    {
      "name": "KafkaStreamingCompositor",
      "params": {"metadata.broker.list": "xxx.xxx.xxx.xxx:9092,...", "topics": "vis-nginx-cache-access"}
    },
    {
      "name": "JSONTableCompositor",
      "params": {"tableName": "raw_table", "schema": "dc:string,server:string,raw:string"}
    },
    {
      "name": "SQLCompositor",
      "params": {"sql": "select dc, server, regexp_extract(raw, '((?<= (tvid=)).+?(?= (&|,|\s)))') as tvid from raw_table"}
    },
    {
      "name": "SQLParquetOutputCompositor",
      "params": {"path": "/data/.../vis-nginx-cache-access.parquet/date=${yyyyMMdd}/hour=${HH}", "mode": "Append"}
    }
  ]
}
```

MySQL IO (Inside Out)

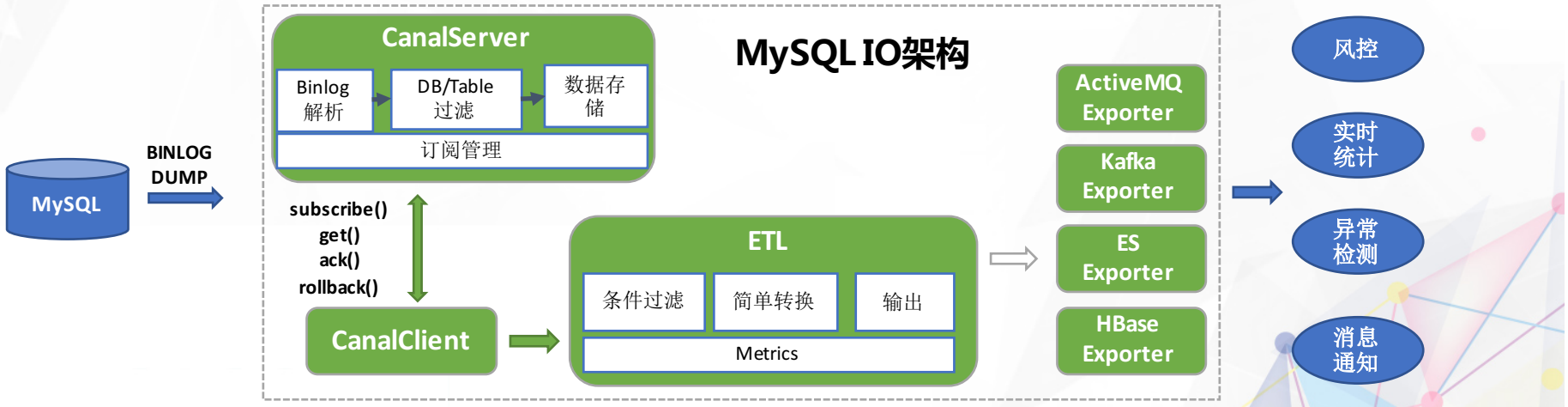
- MySQL数据变更订阅服务
- 基于binlog实时解析的MySQL数据变更增量订阅与同步工具，基于阿里开源的 canal 实现

主要功能

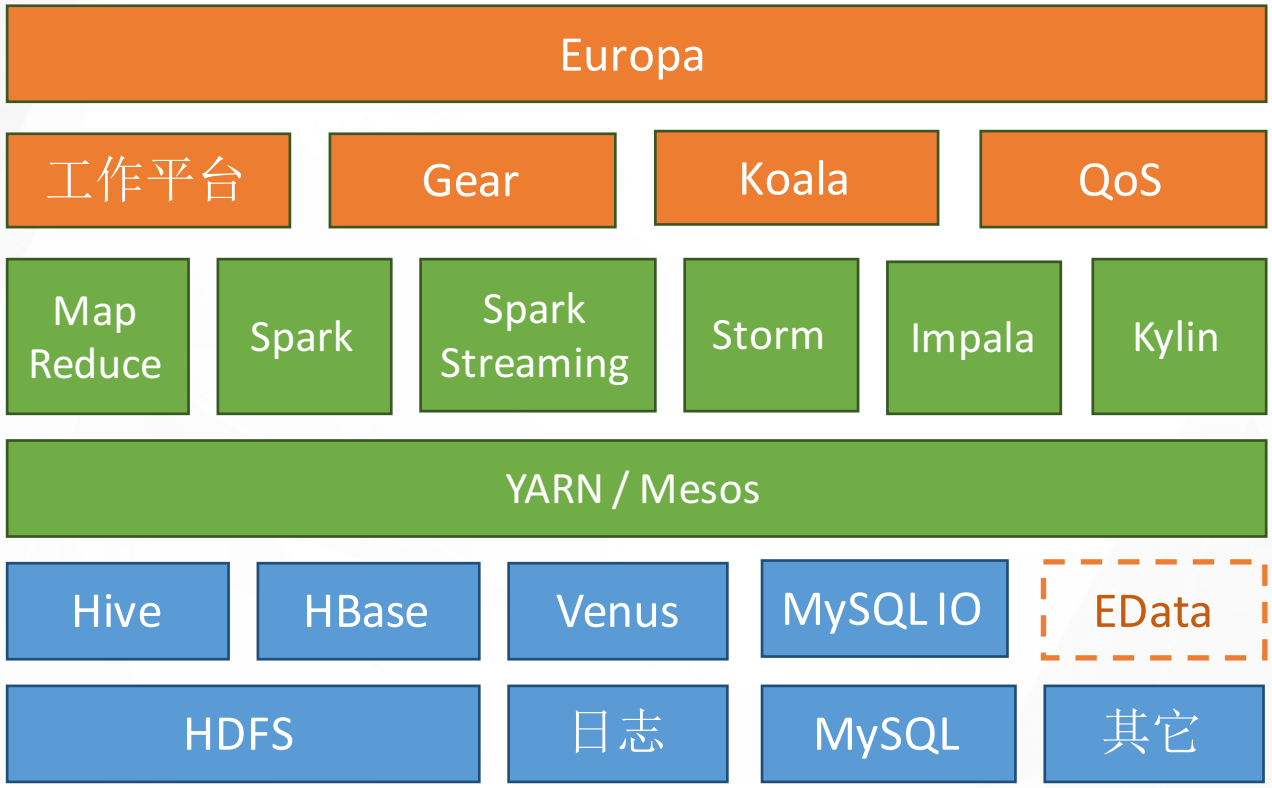
- 实时解析binlog，数据变更信息写入MQ
- 支持读/写binlog数据的事务性，数据不丢失
- 支持数据过滤，包括基于操作类型、列名、列值的过滤
- 支持数据转换，包括库/表/列的映射、空值处理等
- 支持数据写入到ActiveMQ、Kafka、Hbase和ES

优势

- 实时共享数据变更，写入方无需做额外工作
- 对MySQL没有侵入性，不影响线上系统稳定性
- 支持HA，各组件均有热备节点
- 支持Metrics监控和告警

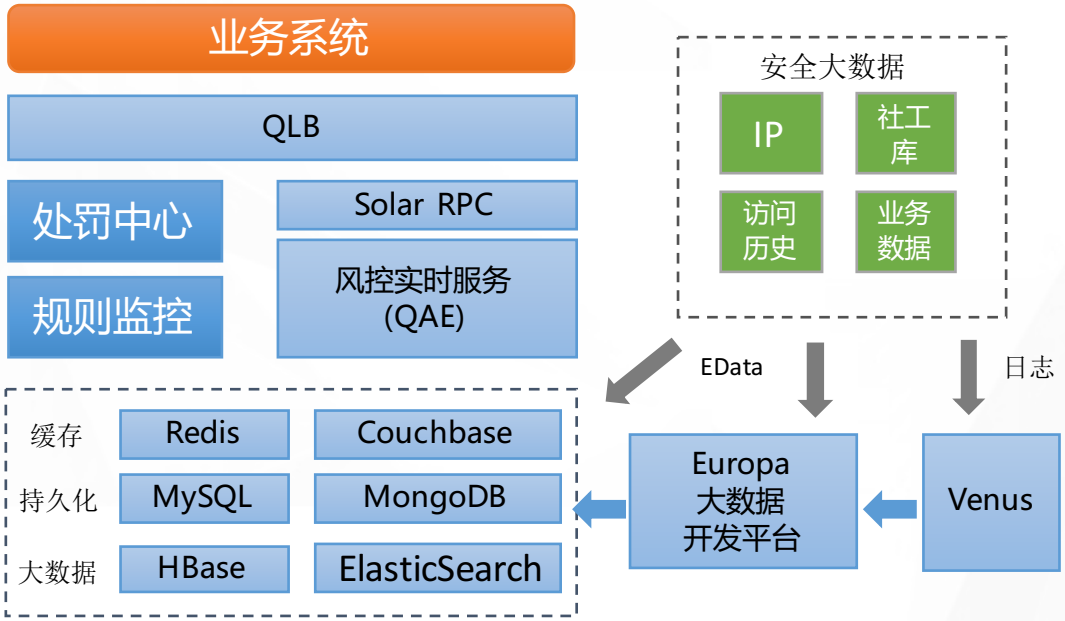


爱奇艺大数据平台架构


Europa
工作平台
Gear
Koala
QoS
Map Reduce
Spark
Spark Streaming
Storm
Impala
Kylin
YARN / Mesos
Hive
HBase
Venus
MySQL IO
EData
HDFS
日志
MySQL
其它

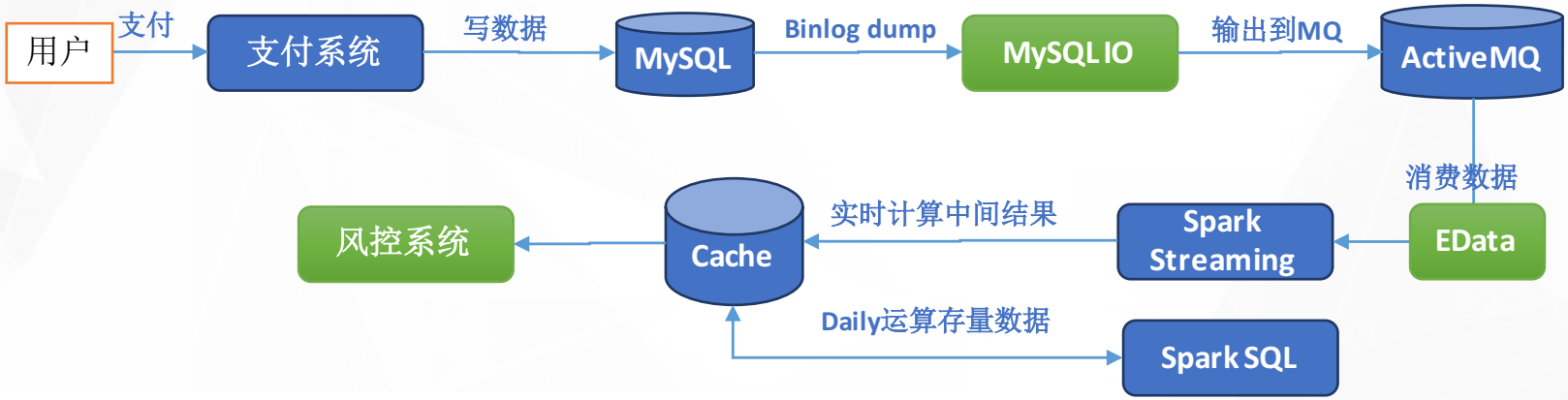
案例：风控系统

- 利用云服务高效构建：四个月，2.5个人力
- 数据：跨业务、异构的数据系统
- 计算：各种复杂的批处理和流处理



案例：风控系统

- 举例：支付表监听
- 特点
  - 需要实时获取业务方数据变化
  - 支付系统存在已久,不便于进行相关数据的实时投递开发
  - 结合实时与存量数据，进行事前/中/后风险决策



以支付表监听举例,风控需要实时获取用户的当天支付数据,提现数据等,用于风险决策

总结

- 阶段感受
  - 专业化：专人做专事、规范化
  - 规模化：技术深入、突破规模瓶颈
  - 生态化：平台、工具链、易用性
- Rome was not built in a day
- 每个阶段需要根据人员和ROI调整优先级

THANKS

SequeMedia 盛拓传媒 | IT168.com 专注导购16年 | ChinaUnix.net | ITPUB www.itpub.net