优酷土豆大数据架构和数据平台发展

--合一数据，唾手可得

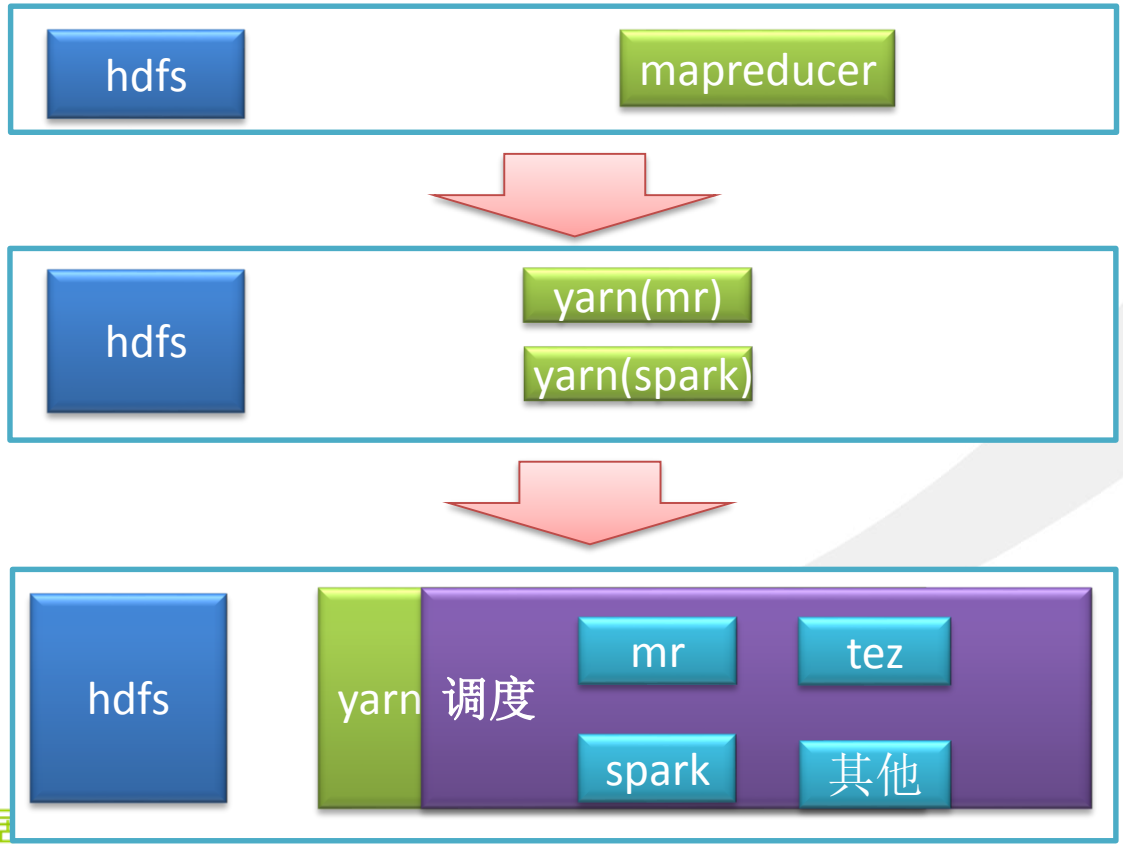
杨大海

2016年1月

合一数据平台现状

- Hadoop体系发展介绍
- 数据平台整体架构
- 实时计算流程,产品
- 数据产品架构介绍

hadoop架构演变



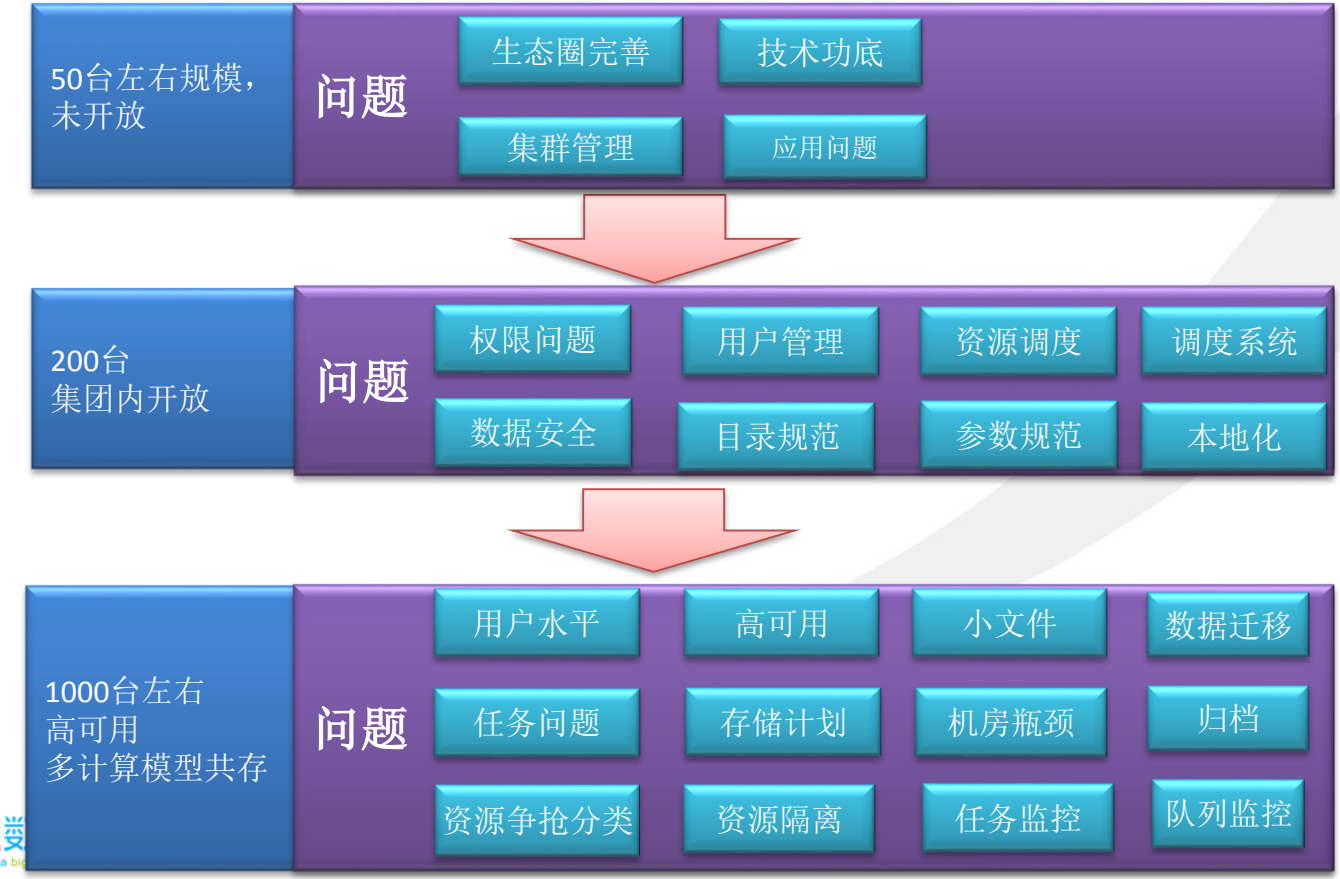
hadoop规模发展

2012.一个团队在用（50台左右规模），裸奔无任何控制

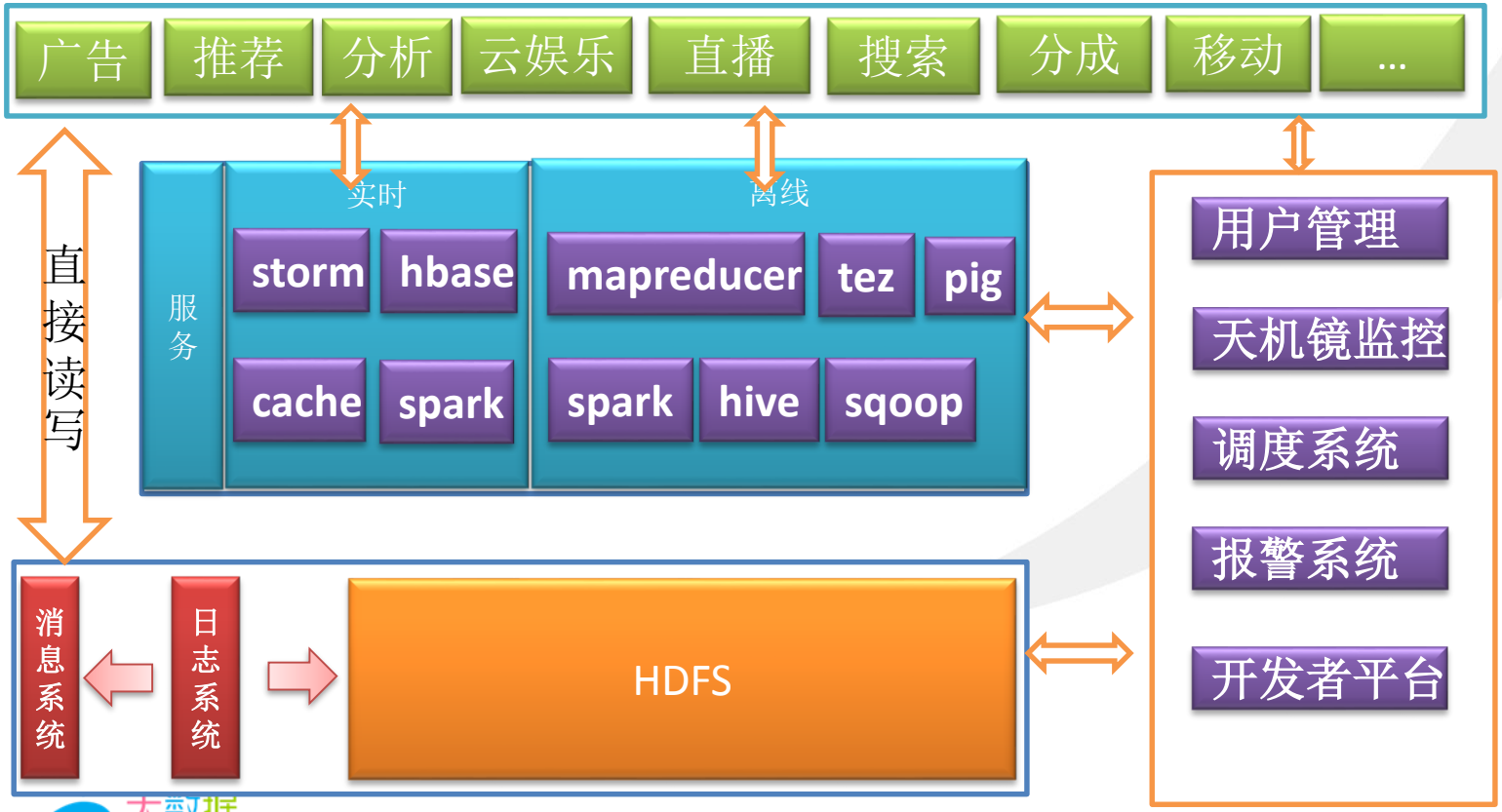
2013.平台集团内开放（200台左右），加入规范和权限控制

2016.多元化计算模型，几百个用户和1000台左右节点
（1000台左右）总存储：28P使用21P，集群cpu使用80%以上

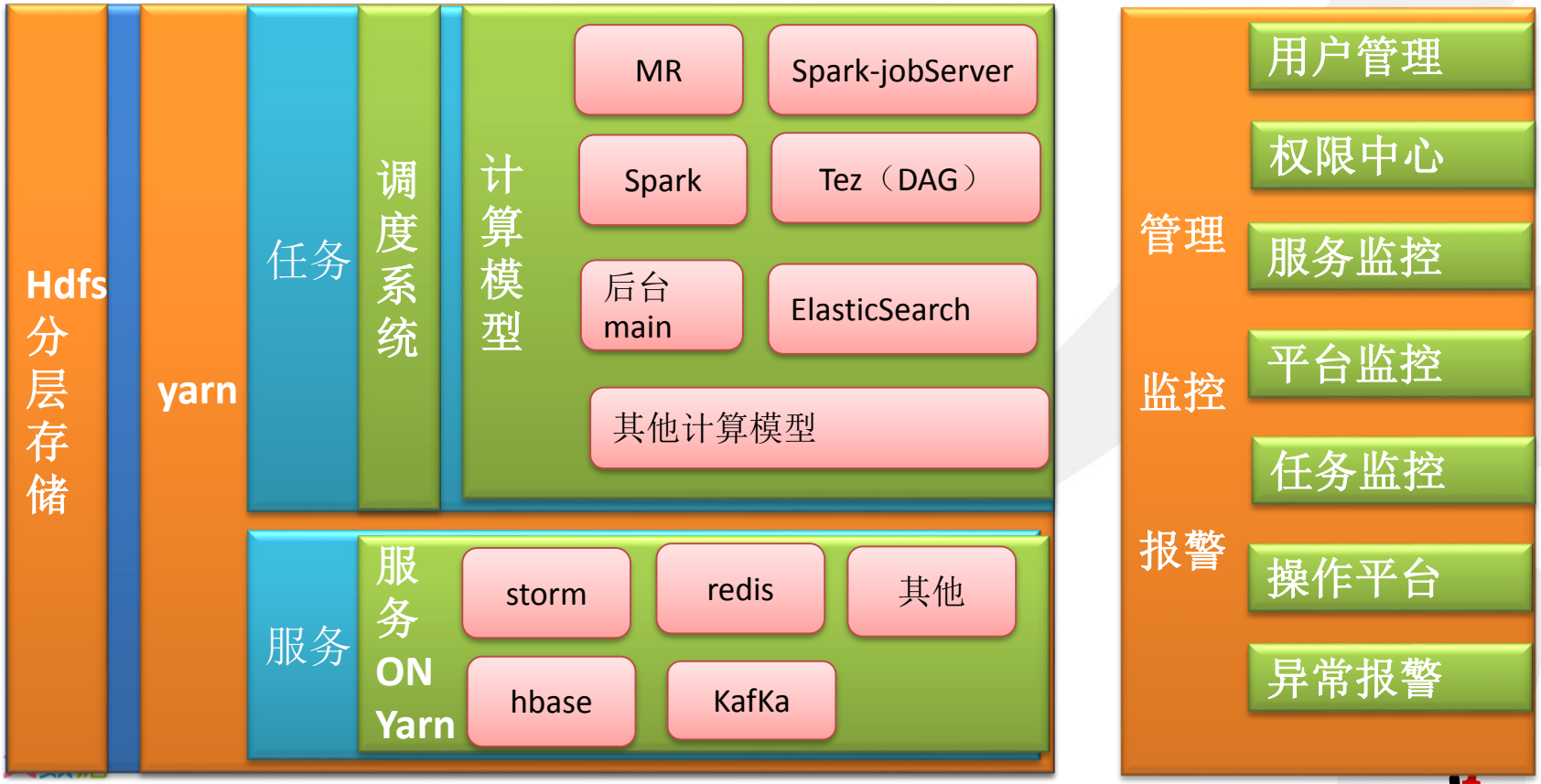
hadoop问题演变


50台左右规模，
未开放
问题
生态圈完善
技术功底
集群管理
应用问题
200台
集团内开放
问题
权限问题
用户管理
资源调度
调度系统
数据安全
目录规范
参数规范
本地化
1000台左右
高可用
多计算模型共存
问题
用户水平
高可用
小文件
数据迁移
任务问题
存储计划
机房瓶颈
归档
资源争抢分类
资源隔离
任务监控
队列监控

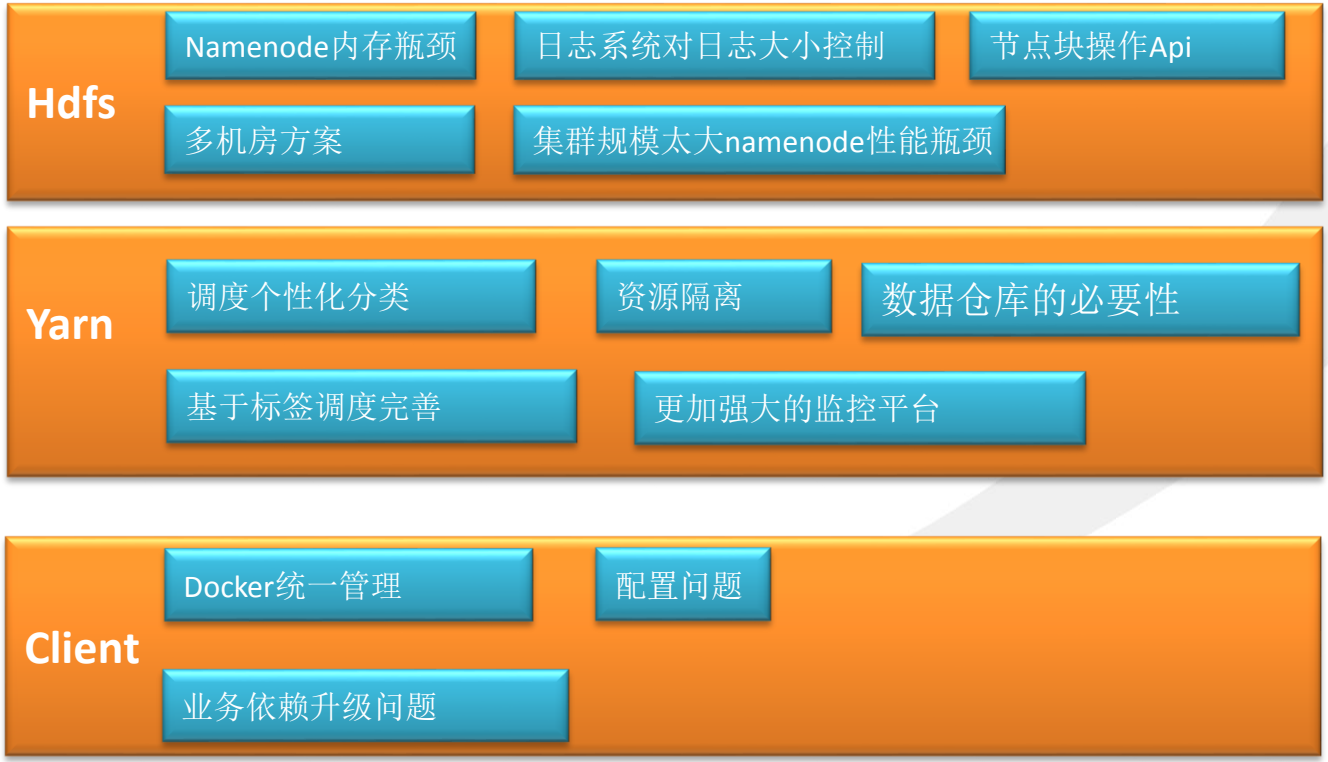
数据平台现状



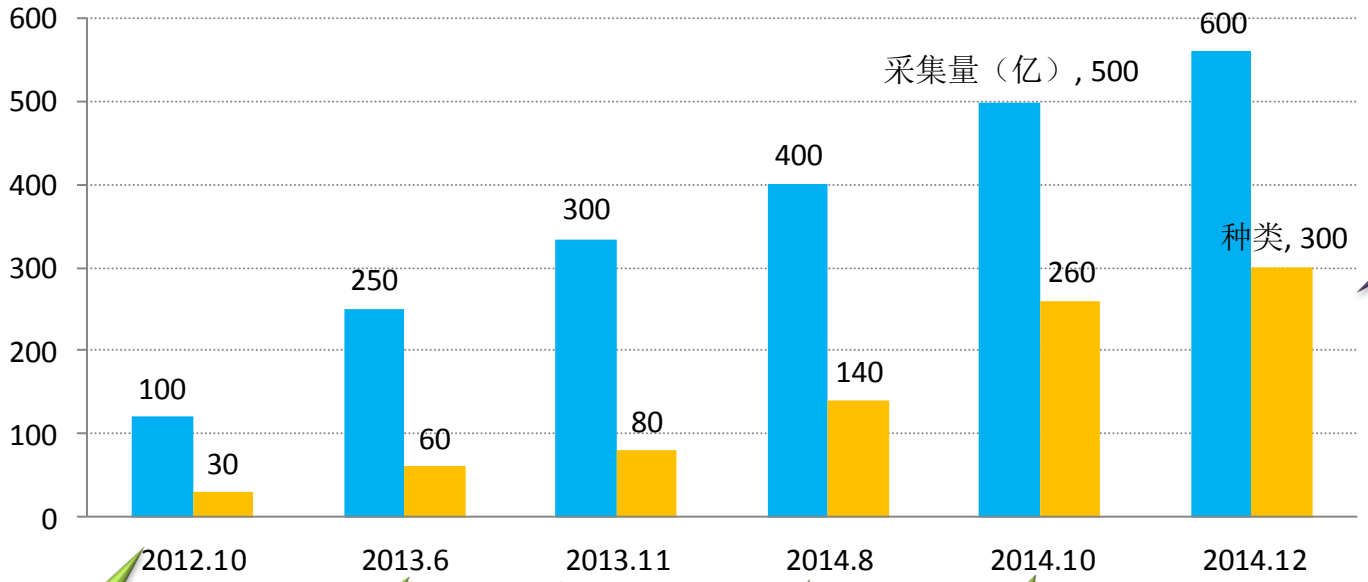
hadoop平台未来



hadoop平台挑战



日志系统：采集量两年6番



300种600亿
日增20T;
周增3种;

新版上线，统一优土采集

开发完成长连接日志采集，时长采集更精确

发布v2版本，性能提高，多机房部署

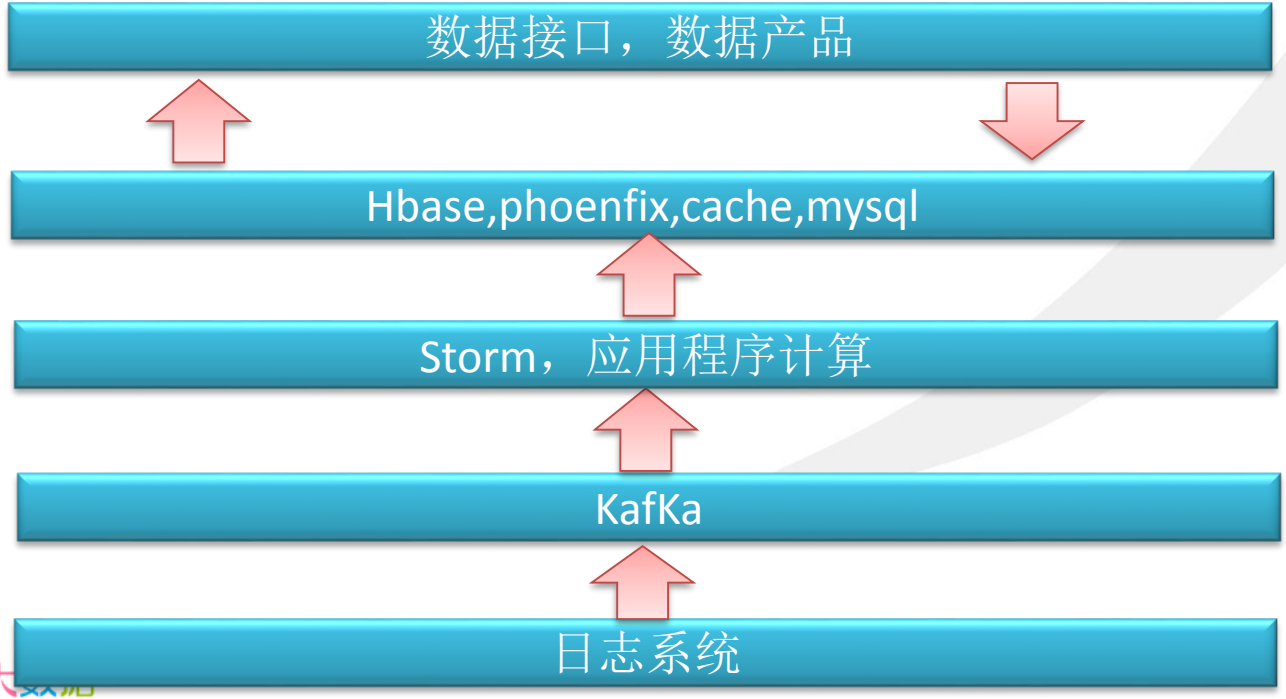
来疯、移动部分日志上线

线上广告日志全部切到日志系统

移动API相关日志正在切

合一集团

实时计算架构流程



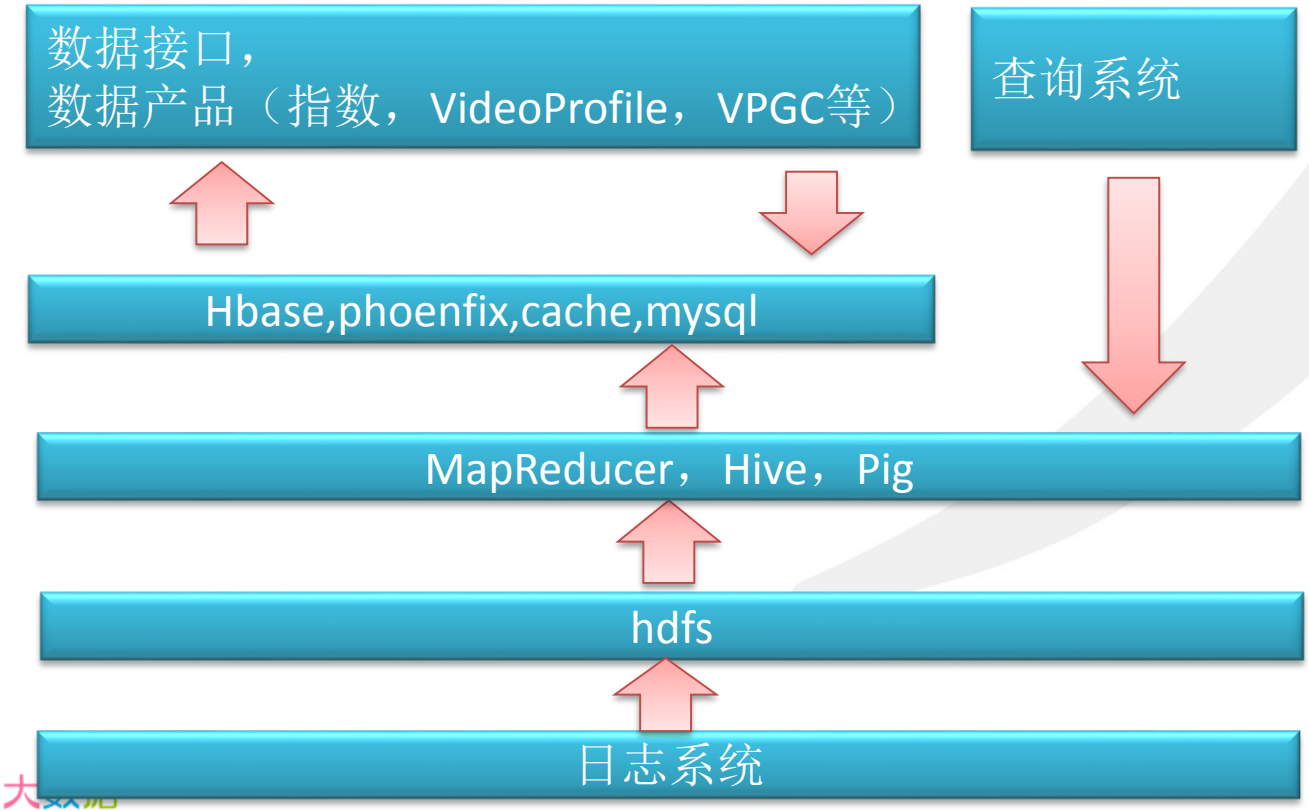
数据分析产品背景

- 指数:index.youku.com全网视频榜单
- VIdeoProfile 全站视频播放行为分析
- Vpgc用户行为分析
- 查询系统，个性化用户需求处理

大数据
big data big thoughts

合一集团

数据分析产品架构



数据分析架构-问题

1.缺少数据仓库：缺少数据仓库导致数据重复扫描集群占用资源后期无法控制的情况

2.数据接口*平台*应统一包装：数据接口应有数据平台统一包装开放，业务团队只用申请使用，做到数据接口和数据产品分离（好处是数据接口共享，数据接口可以作统一的流量请求监控等）

2016 大型企业信息运维高峰会

Big Enterprise Information System Maintance Summit

企业信息运维的互联网+

主办单位：黑龙江嘉和安泰商务服务有限公司

协办单位：国网黑龙江省电力有限公司信息通信公司　黑龙江省电力调度实业公司