



集团化企业开放数据 平台构建之路

傅杰
2013年4月



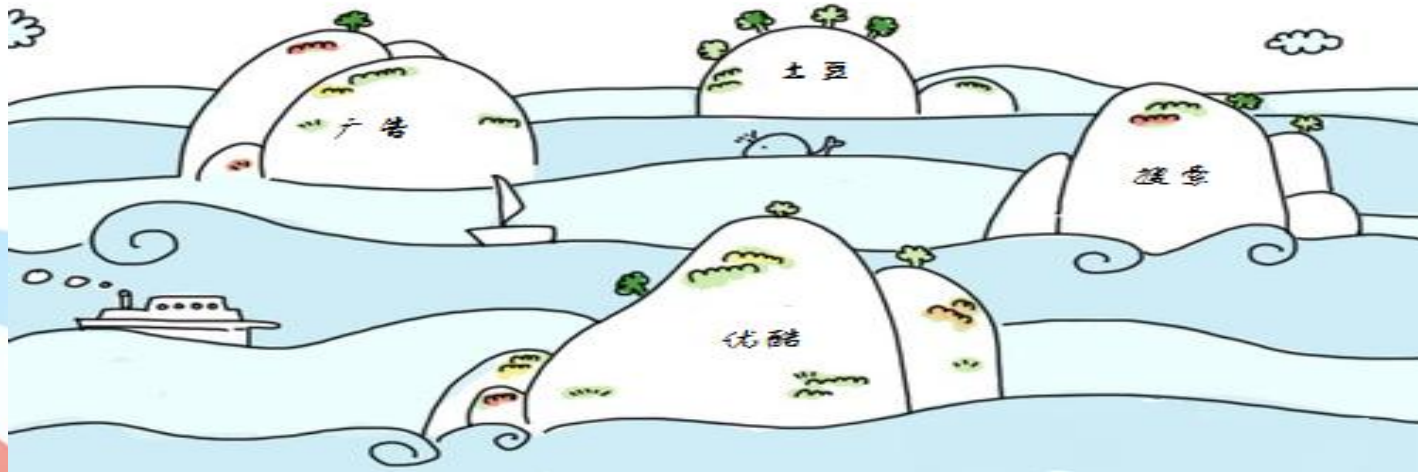
大纲

- 企业集团化带来的数据挑战
- 开放数据平台的构建之路
- 优酷土豆集团开放数据平台



数据孤岛

集团化企业的数据孤岛现象更为明显！



企业集团化带来的数据挑战

物理孤岛

重复造轮

资源浪费

逻辑孤岛

数据关联

数据标准

挑战物理孤岛



- 集中存储

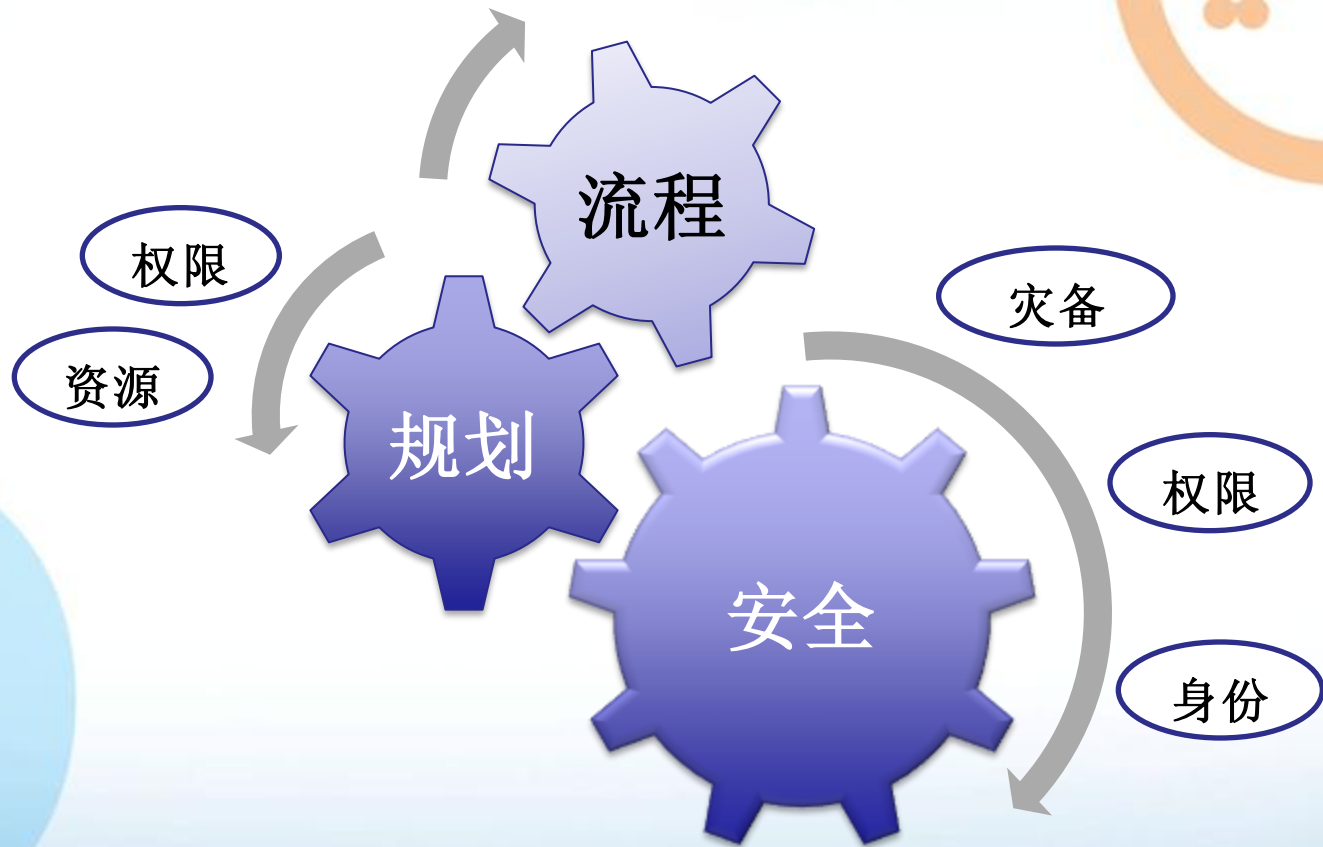
- 统一采集

- 开放计算

开放Hadoop平台



开放三要素



Hadoop安全开放



- 集成Kerberos
- 自定义用户组
- 监控报表
- 开放注册用户120、团队16



Hadoop集群规模



- 日增2T+原始日志
- 日均6000道作业
- 日扫描上百T数据
- CPU利用率50-60%、峰值95%

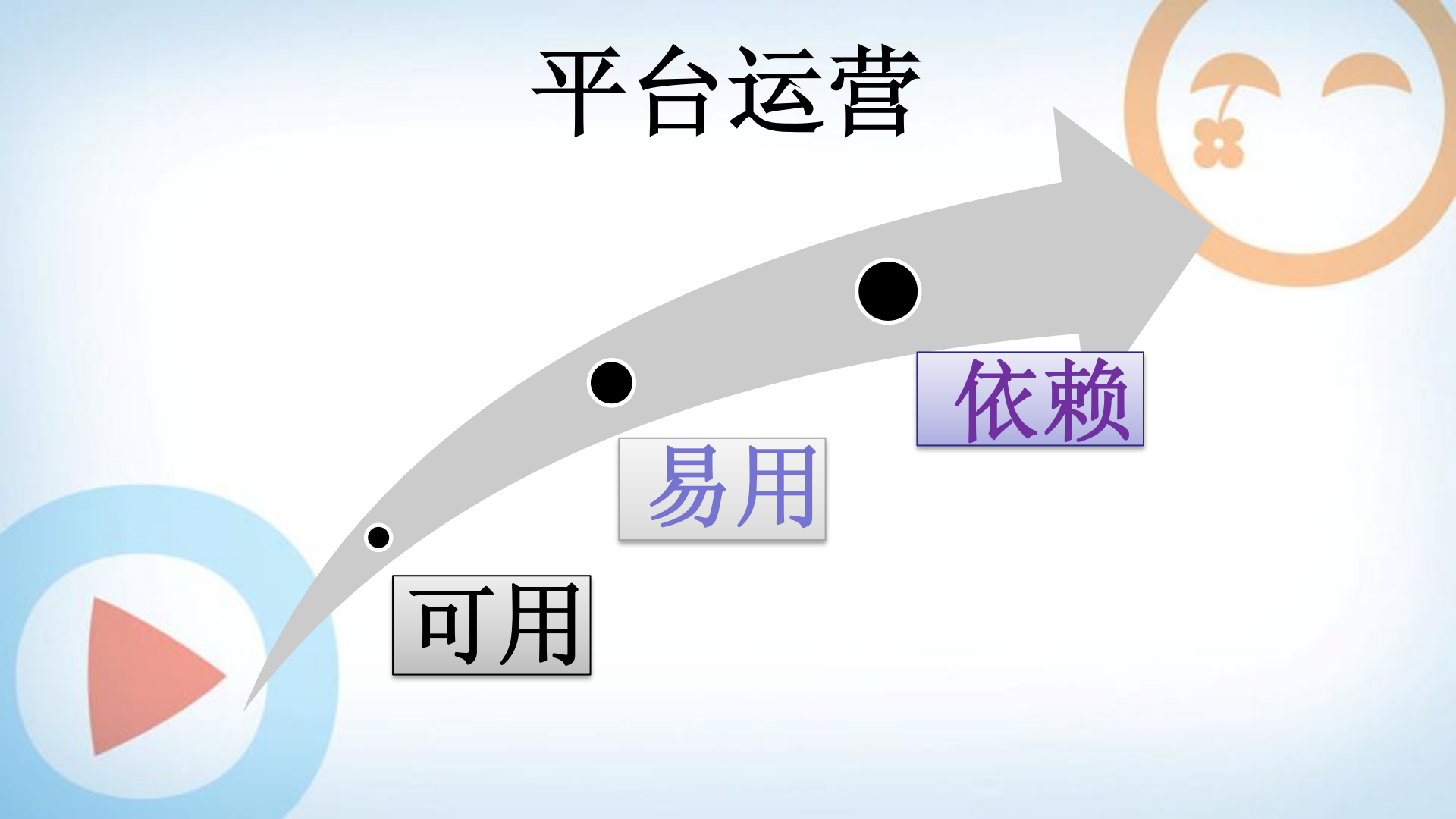


荆棘密布

- System.out
- 机房迁移(jaas bug)
- 公平调度器故障
- JobTracker 堆栈满
- IPV6



平台运营



挑战逻辑孤岛



- 数据易用
- 封装数据服务
- 打造数据产品

开放数据平台



探索逻辑孤岛



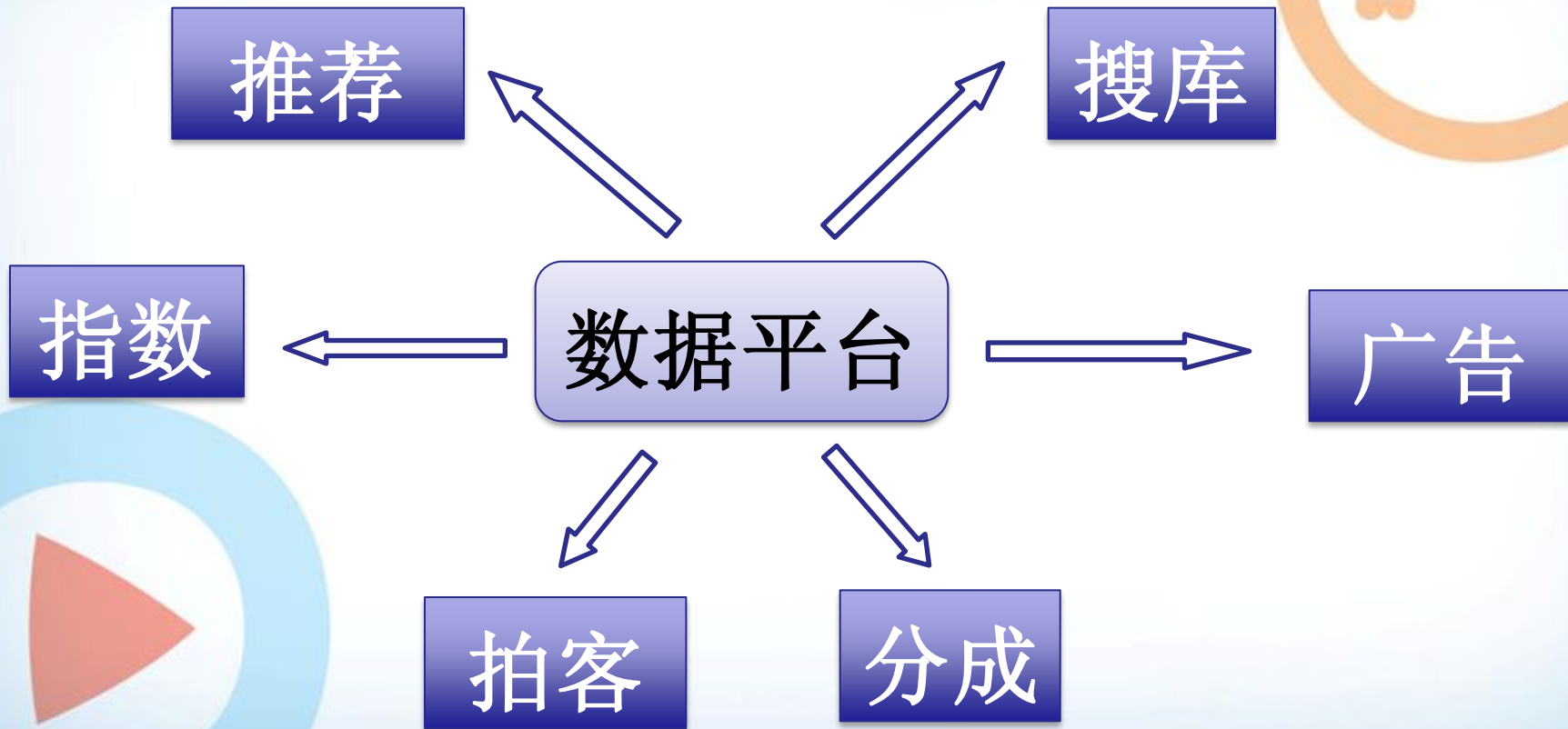
- 制定业务数据标准
- 元数据定义
- HIVE库互通
- 数据产品



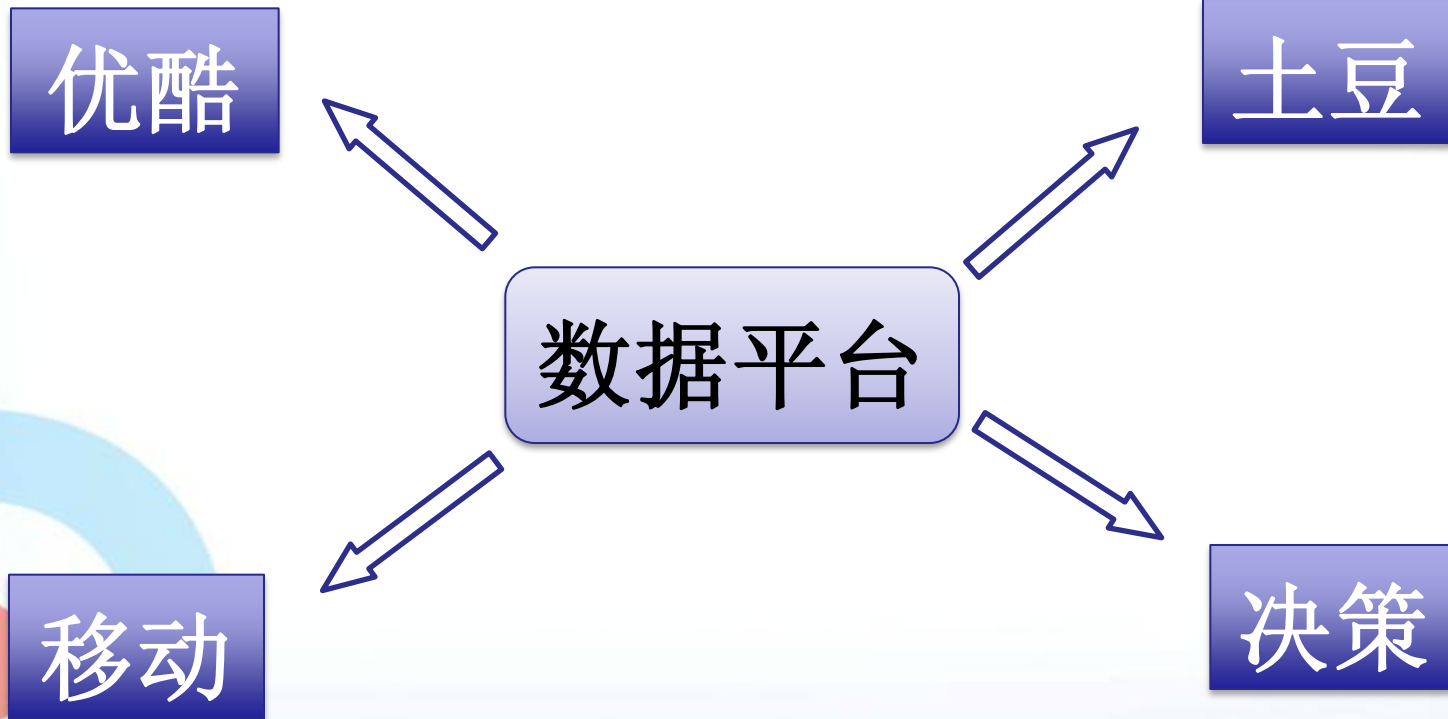
优酷土豆集团开放数据平台

DATA, SO EASY

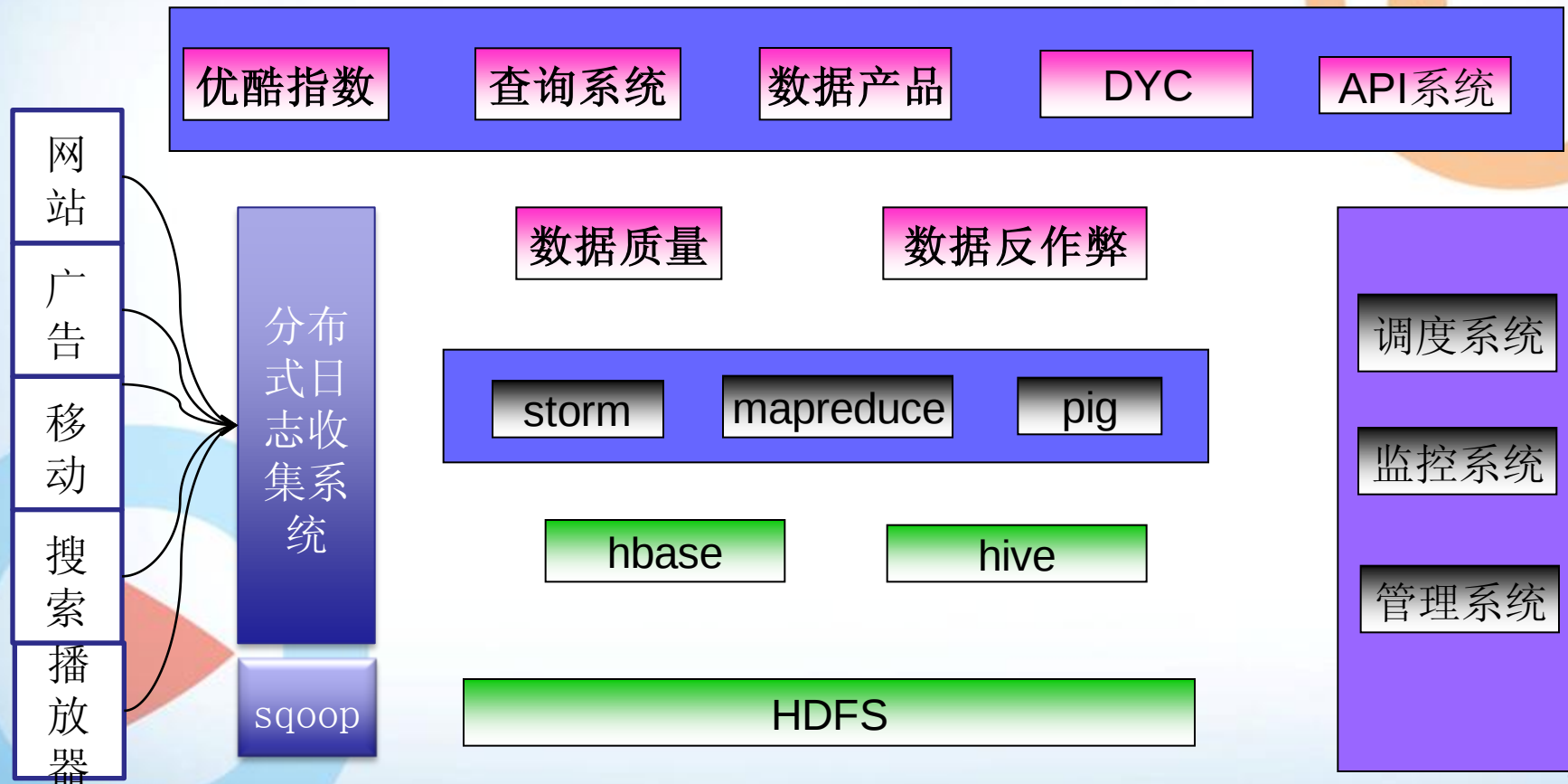
数据平台辐射范围



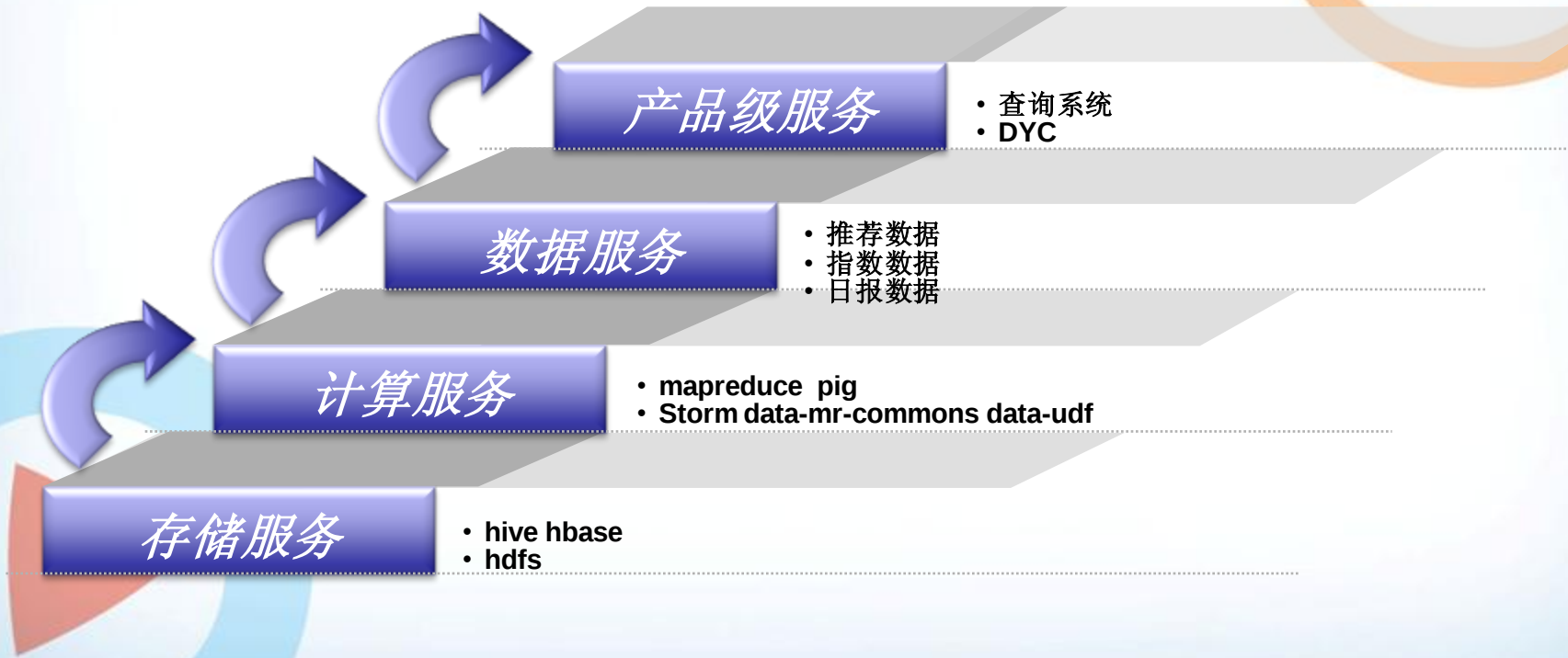
数据平台内部支撑



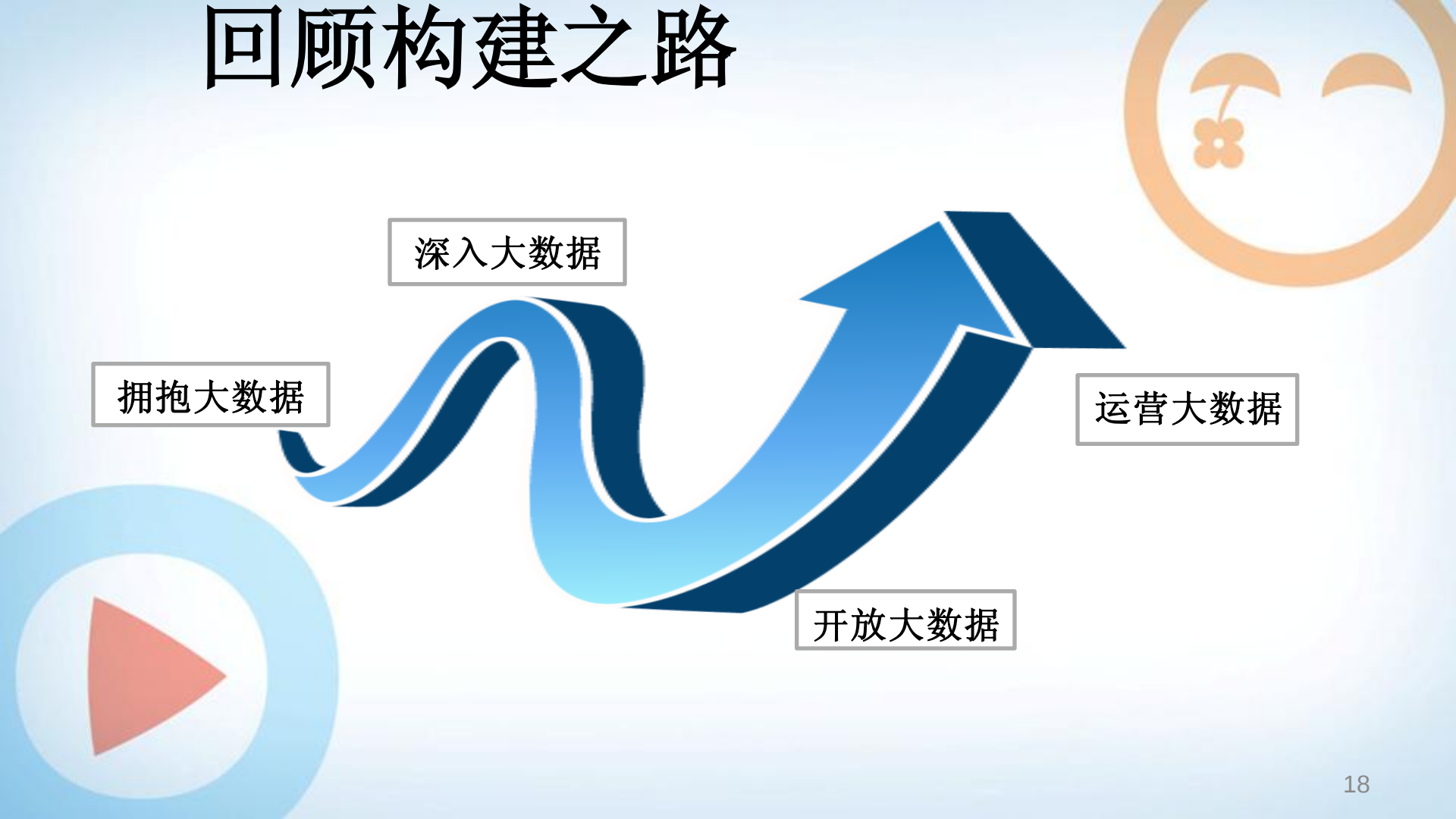
数据平台架构



平台服务



回顾构建之路



近期工作

- 监控运营报表
- 实时计算应用
- 日志采集系统优化
- API层架构规划



Q&A

