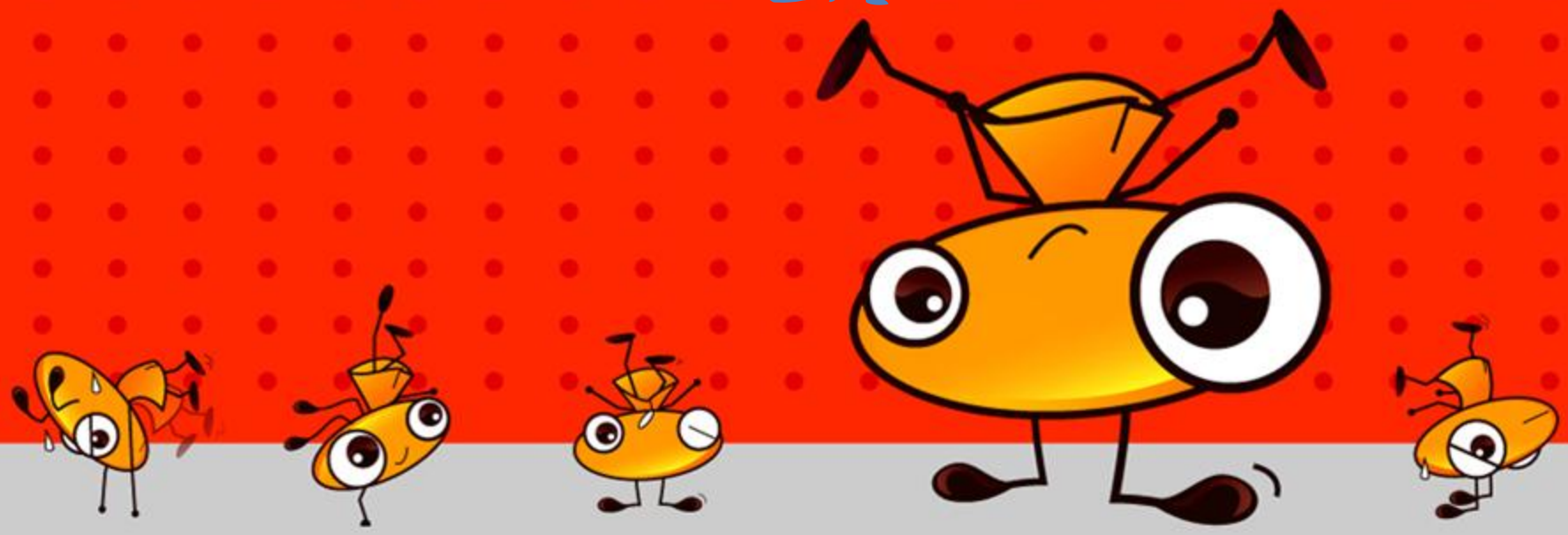
淘宝网
Taobao.com

淘宝数据分析挖掘实践及变革

毛波 2013-04



淘宝网
Taobao.com

目录

- 淘宝数据四阶段
- 系统变迁及平台架构
- 数据应用格局
- 新的探索
- 一些观点

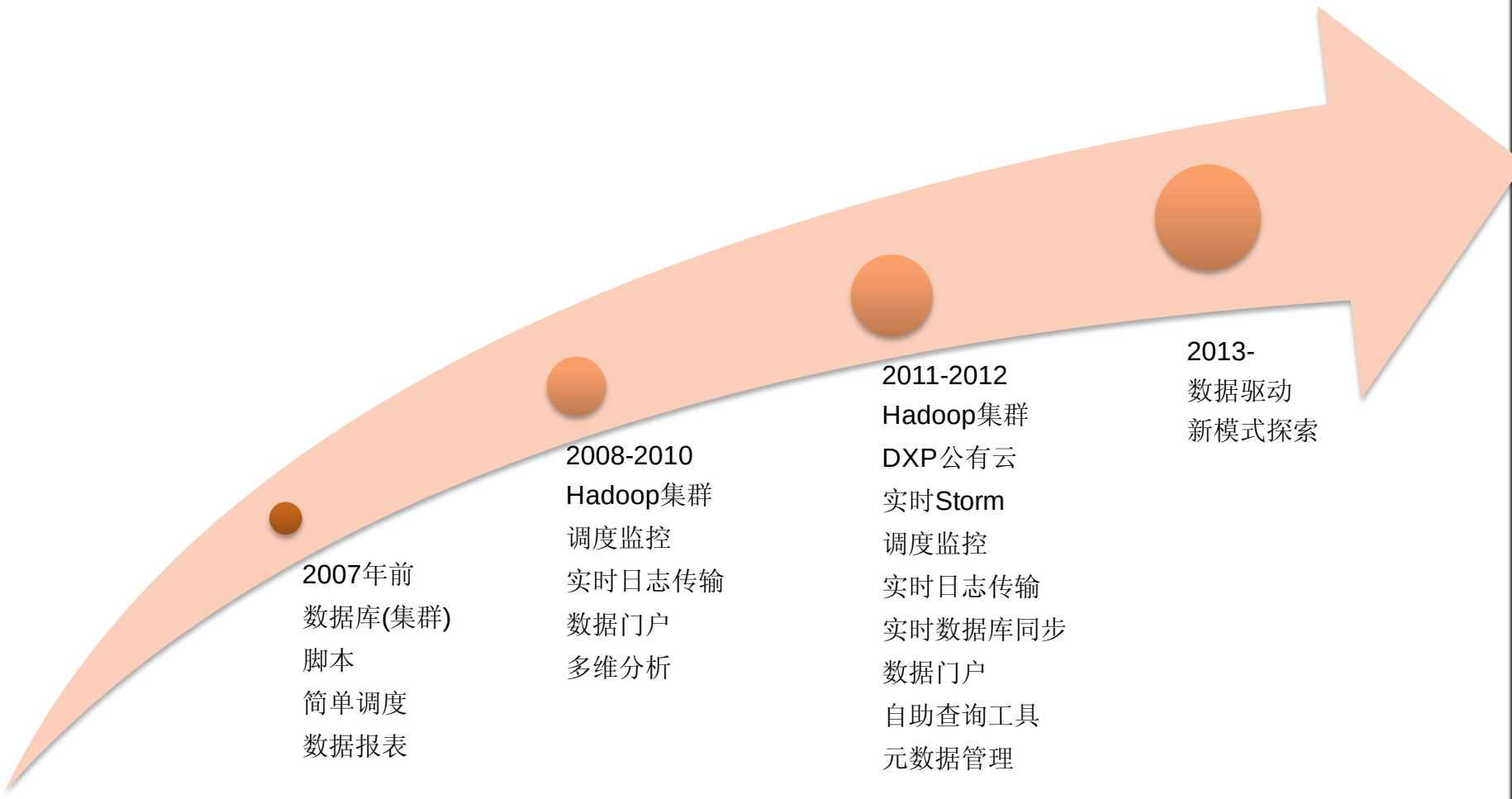
淘宝数据四阶段

- 被动响应
  - 2007年前
- 主动变革
  - 2008-2010
- 优化完善
  - 2011-2012
- 引领驱动
  - 2013-

淘宝网
Taobao.com

系统变迁及平台架构

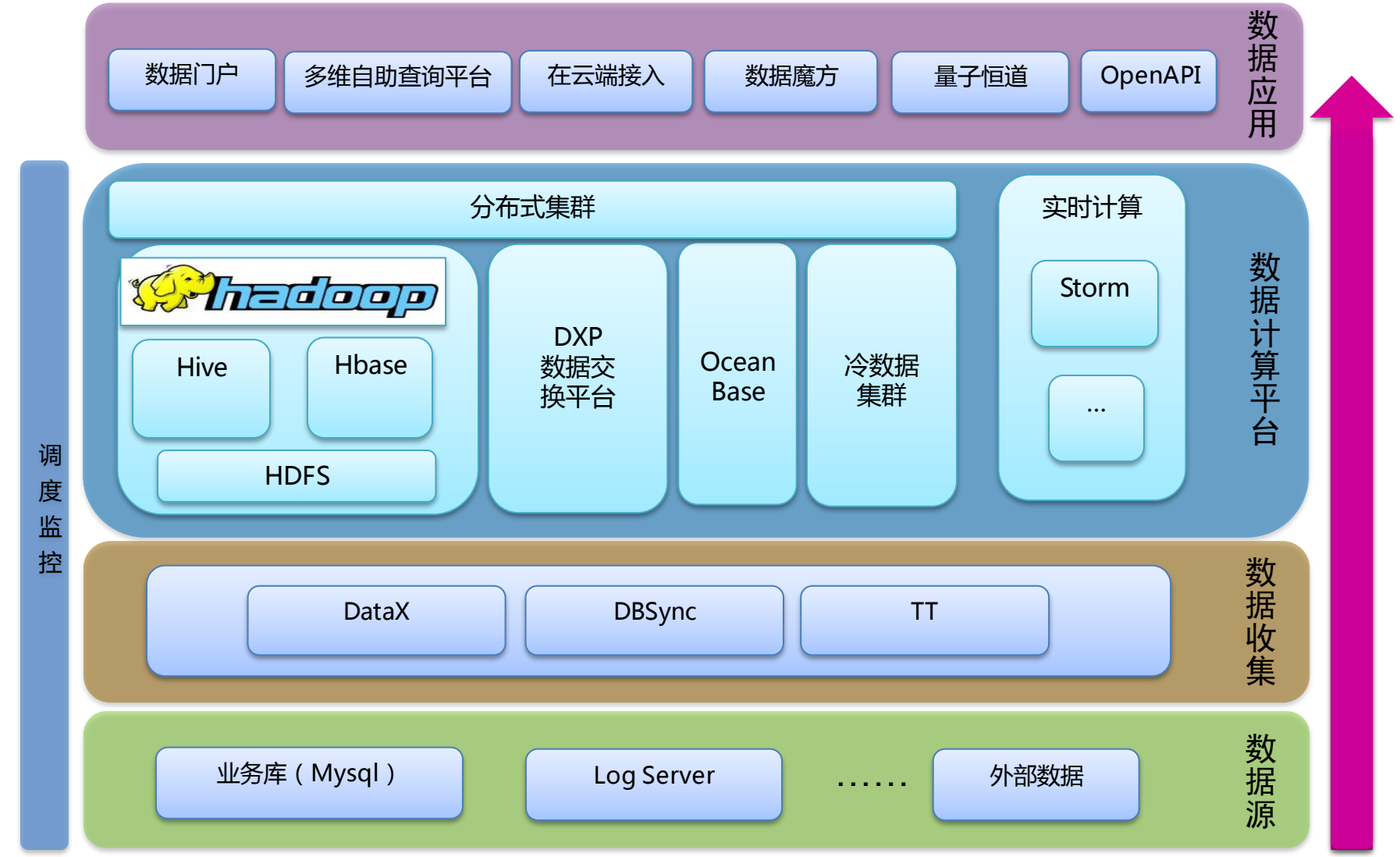
数据系统变迁


2007年前
数据库(集群)
脚本
简单调度
数据报表
2008-2010
Hadoop集群
调度监控
实时日志传输
数据门户
多维分析
2011-2012
Hadoop集群
DXP公有云
实时Storm
调度监控
实时日志传输
实时数据库同步
数据门户
自助查询工具
元数据管理
2013-
数据驱动
新模式探索

数据系统变迁

- 性能
- 扩展性
- 运维

数据平台架构


数据门户
多维自助查询平台
在云端接入
数据魔方
量子恒道
OpenAPI
数据应用
调度监控
分布式集群
hadoop
Hive
Hbase
HDFS
DXP
数据交
换平台
Ocean
Base
冷数据
集群
实时计算
Storm
...
数据计算平台
DataX
DBSync
TT
数据收集
业务库（Mysql）
Log Server
.....
外部数据
数据源

数据应用格局

对外数据产品

- 数据魔方/淘宝指数
  - 行业趋势
  - 人群特征
  - 成交排行
  - 市场细分
- 量子恒道
  - 销售分析
  - 营销效果
  - 来源分析
- 搜索排行榜

对外数据产品

- 淘宝时光机
  - http://me.taobao.com/
  - 回忆的感动
- 排行榜

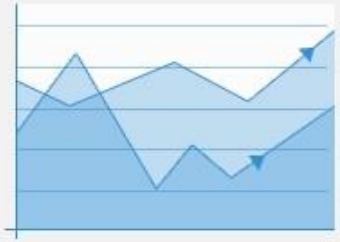
对外数据产品–淘宝指数

淘宝指数能告诉你...

长周期走势

淘宝上连衣裙的搜索趋势是怎样的？

任一关键词（如商品、行业、事件等）的搜索和成交走势。



人群特性

淘宝上搜索、购买iPhone4的都是什么样的人？

用淘宝指数查看不同商品的**消费人群特征。**

成交排行

最近7天淘宝最火的搜索词、行业和品牌是？

基于淘宝搜索和成交的排行榜，宏观数据清晰呈现。

市场细分

北京女白领和20岁大学生都买过什么面膜？

淘宝指数告诉你**不同标签的人买过什么商品。**

对外数据产品-量子恒道


汇总数据分析
店铺诊断
健康日报
新
经营概况
PC端店铺分析
流量分析
实时客户访问
流量概况
按小时流量分析
按天流量分析
宝贝被访排行
分类页被访排行
店内搜索关键词
首页被访数据


销售分析
销售总览
销售详解
宝贝销售排行
买家购买分析
促销手段分析
推广效果
客户分析
百宝箱
实验室功能
手机端店铺分析
流量分析
销售分析
销售总览
宝贝销售排行
推广效果

数据嵌入产品中

- 搜索匹配、排序
- 广告匹配、排序
- 推荐
- 商家后台数据
- 营销效果
  - 直通车、展示广告、淘宝客

内部数据服务

- 淘数据门户
  - 用户分析
  - 商家云图
  - 活动效果分析
  - 例行数据报表
- 在云端
  - 低门槛接入分布式集群
  - 周活跃用户1000+

内部数据服务

- 多维数据自助查询平台
  - 数据仓库和索引技术结合
  - 随意组合维度
  - 秒级返回
- 日常数据需求管理
  - 数据接口人

淘宝网 Taobao.com

数据工具

- 天网调度
- 元数据管理
- 数据地图-定位、血缘分析
- DataX异源数据传输
- TimeTunnel实时日志传输
- 监控报警
- 生命周期管理

新的探索

- 金融服务
  - 小微企业贷款
  - 个人消费贷款
- 全网精准营销
  - DMP、DSP、AD Exchange、RTB
- 无线与PC数据打通
- 数据交换

淘宝网
Taobao.com

一些观点

- 数据处理是手段，数据应用是根本
- 云系统运维能力是核心竞争力
- 整合关联让数据价值指数级增长
- 数据可视化很重要
- 想大做小，迭代优化
- 关于隐私
  - 隐私和服务的权衡-GPS
  - 控制使用比控制收集更有效
  - 不针对具体个体

新浪微博：eNeolithic

QA