AdMaster Academy

BDTC
2014 中国大数据技术大会
BIG DATA TECHNOLOGY CONFERENCE
暨第二届CCF大数据学术会议

Hadoop在广告监测技术的实践

AdMaster(精硕科技)
卢亿雷

AdMaster
Data-Insight-Solution

目录

- 广告营销数据流程介绍
- 广告监测技术特点分析
- 广告监测数据差异分析
- 广告数据挖掘平台架构
- ADH在广告营销数据挖掘的特点
- AdMaster数据分析平台

广告营销数据流程介绍

全流程营销

- 展示广告
- Minisite
- 微博

实时竞价营销

- Ad Exchange（广告交易平台）
- DSP（Demand Side Platform，需求方平台）
- SSP（Sell-Side Platform，供应方平台）
- DMP（Data-Management Platform，数据管理平台）

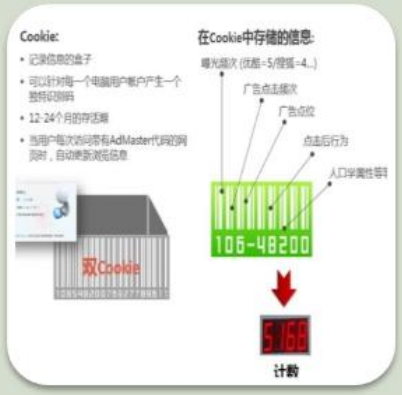
广告监测技术特点

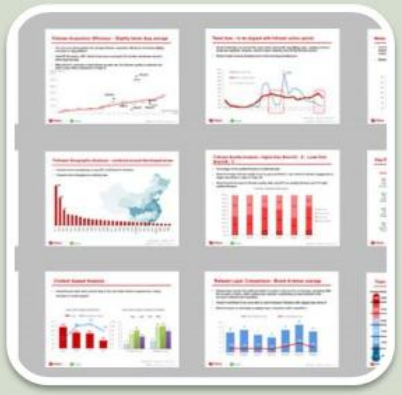
什么是Cookie:

- 指某些网站为了辨别用户身份而储存在用户本地终端（Client Side）上的数据（通常经过加密）--维基百科
- Cookie本身不能跨浏览器，更不能跨设备，且有过期时间限制

Cookie知识

广告监测工作流程





第一步

添加监测代码，在用户终端植入 **cookies**

第二步

广告浏览数据收集(包括流量、次数、人数等)

第三步

数据整理和计算(包括**人口属性推及**等)

第四步

深入的数据分析和洞察

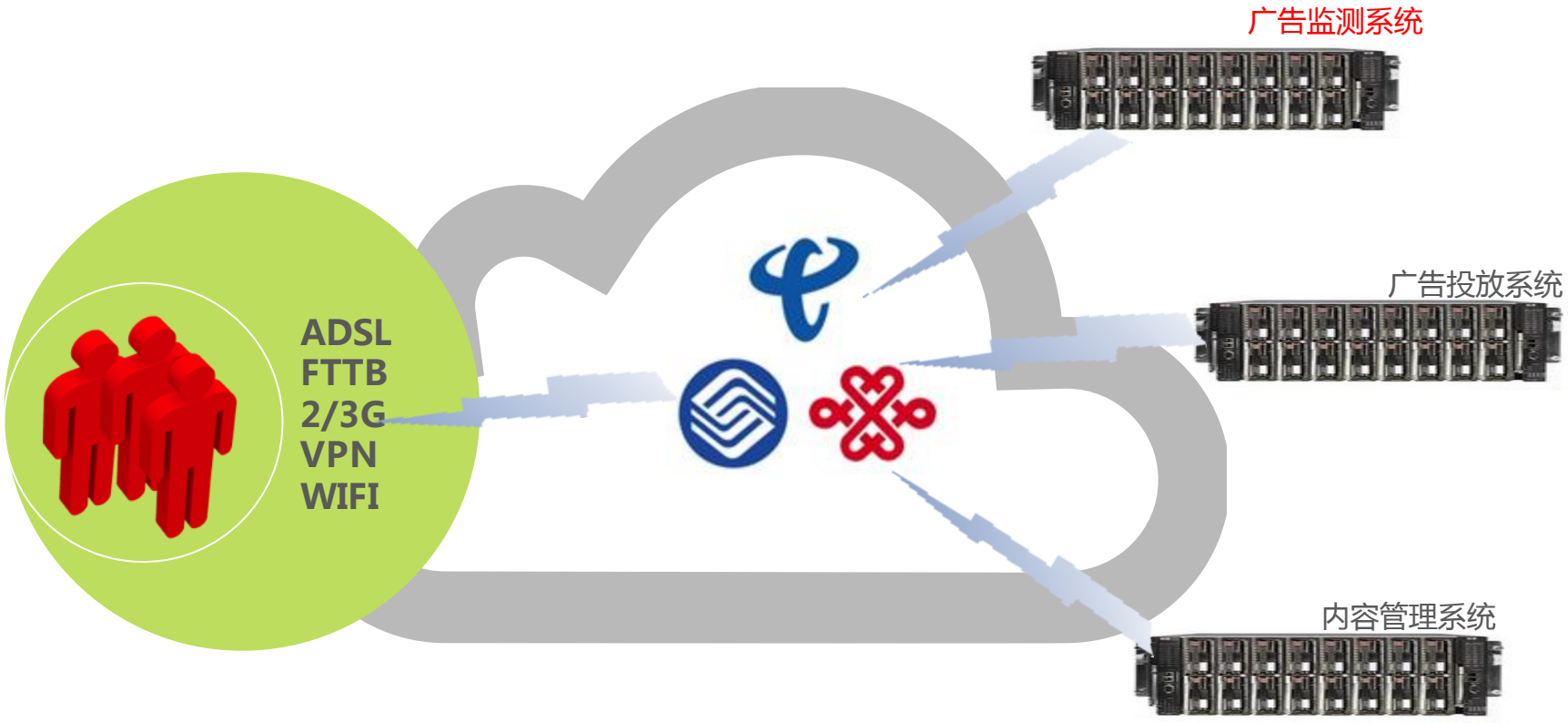
“数据的力量”:我们实时地告诉您“我们的网络广告与用户都发生了什么事情？为什么？”从而总结得到“**如何去改进或进一步规划**”

广告监测技术的出现，使我们真正认识了数字营销

从此，“盲人摸象”成为了历史……

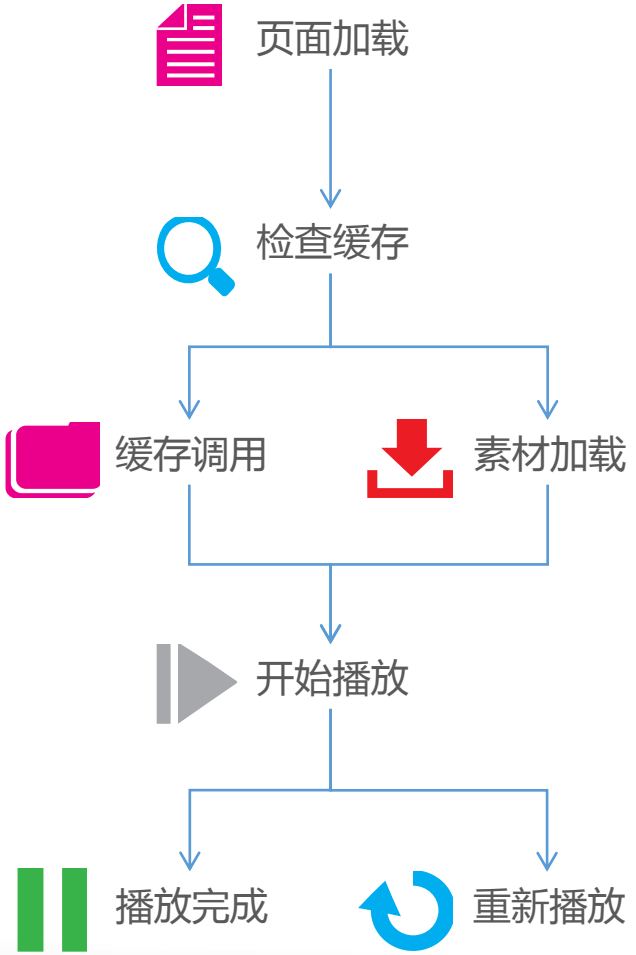
广告监测数据差异原因

- 对于同一个IP，采用不同IP库的系统可能会得出不同的地域结论
- 智能路由难题：
  中小宽带接入商智能地选择更快或成本更低的线路连接到服务器。
  两个独立系统（如监测和投放）同时采集同一个用户的IP，会取得不同的IP值。



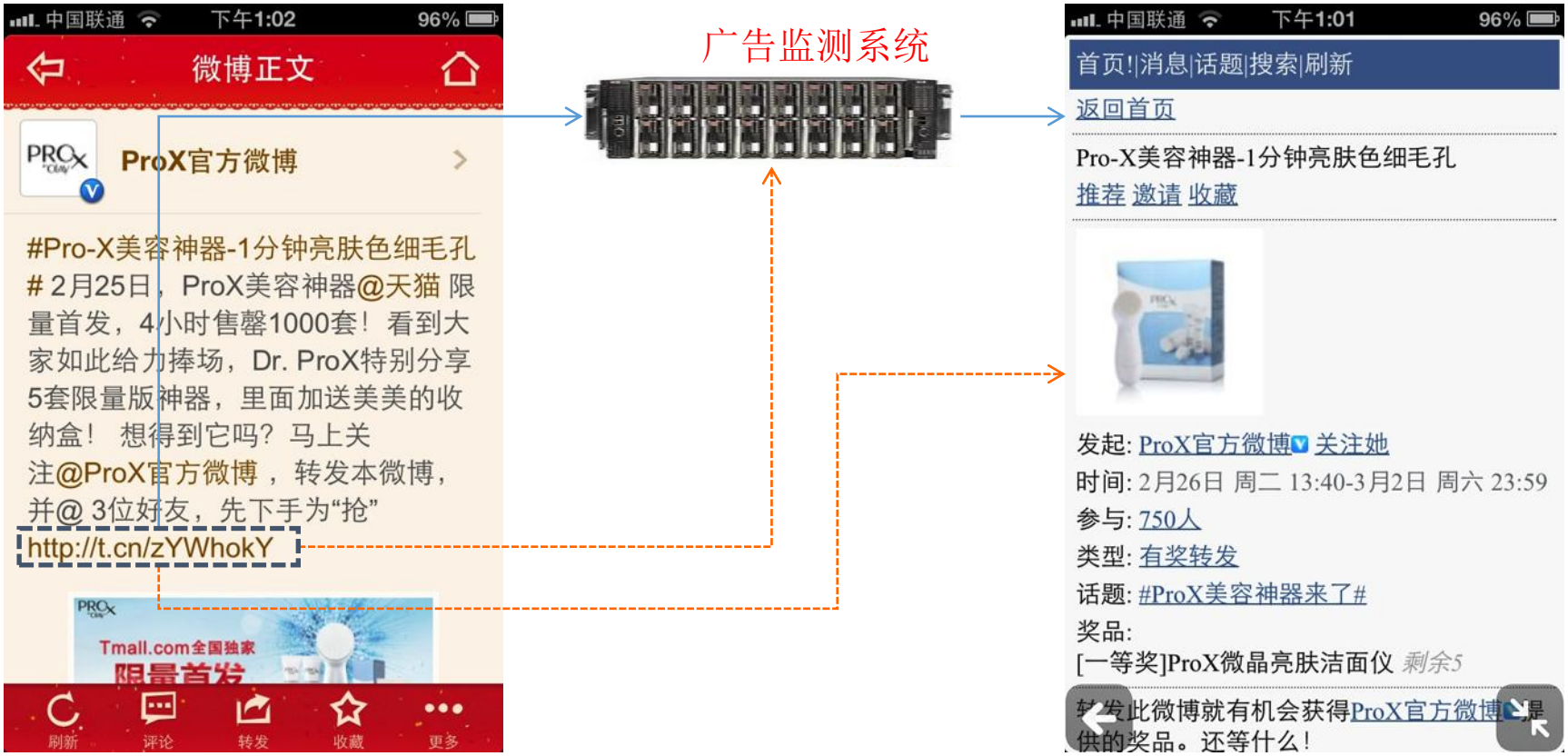
广告监测数据差异原因

- 监测代码的部署时点的不同
- 监测机制和指标定义的差异



广告监测数据差异原因

- 同步（串联）点击和异步（并联）点击监测；主要为了适应移动APP较不稳定的网络环境



同步（串联）点击监测　　异步（并联）点击监测

广告监测数据差异原因

- 浏览器 Cookie 和Dual Cookie

浏览器Cookie容易被清除，不能跨浏览器和PC客户端，因此需要使用Flash Cookie 进行校正

广告监测中存在的数据异常

无中生有

曝光造假
（曝光代码放在其他无广告页面）

点击造假
（嵌到其他点位骗点 / 刷点击代码）

频次造假
（控制机器人清除 Cookie 刷曝光）

重复调用监测造假
（一条广告刷多条曝光代码）

鱼目混珠

定向内容掺水
（利用非热门剧目和频道）

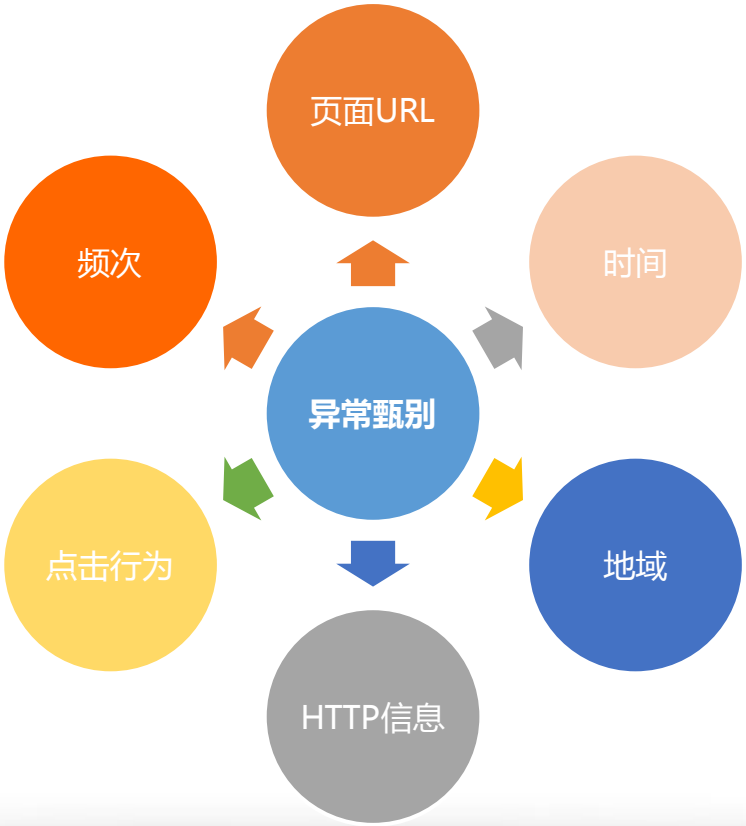
定向地域掺水
（利用三四线城市库存流量）

播放顺位掺水
（前帧贴后帧 / 轮播位置换序）

站外流量掺水
（够买廉价长尾流量）

广告异常甄别的六大维度

- 素材曝光、曝光后点击、点击后互动构成互相牵制的时间线
- Cookie ID / IP地址 / User Agent 组成受众的甄别信息
- 广告素材播放页面的URL，用于判定点位匹配度、定向质量度



广告异常甄别的六大维度

利用浏览器Referrer 信息，获取广告所在页面URL地址

页面URL

URL & Referrer

同一cookies显示、点击间隔
从广告显示到第一次点击的决策时间

时间

Time

广告显示和点击地域匹配度

地域

Geographic

点击前曝光频次，如果点击产生前是0曝光，该点击存在异常

频次

Frequency

点击率、到达率、跳失率、访问时间和访问深度的五维识别模型

点击行为

Conversion & Post-click Action

Session、浏览器版本、操作系统版本等信息；机器客户端模拟的行为
上述信息异常

HTTP信息

HTTP Header & User Agent

广告营销数据案例分析

页面浏览行为 | 网站浏览行为 | 社交网络行为 | 调研问卷结果 | 网上购买行为

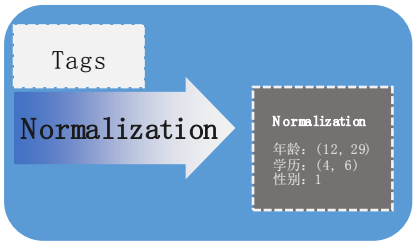
喜欢浏览品牌页面并喜欢看时政新闻

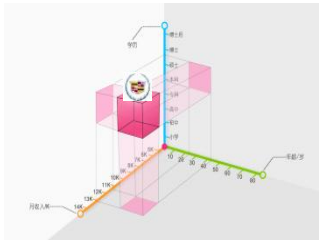
每日多次浏览品牌官方首页并参与活动

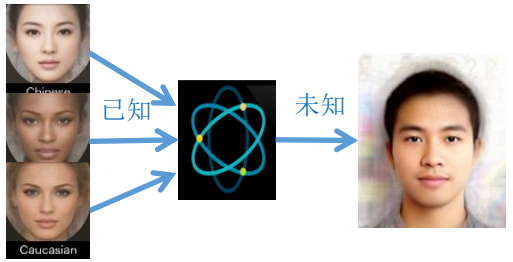
多次转发品牌官微；最近关注家电类，多关注家电类相关微博

填写品牌调研问卷

最近经常家电，多次关注家电竞拍活动

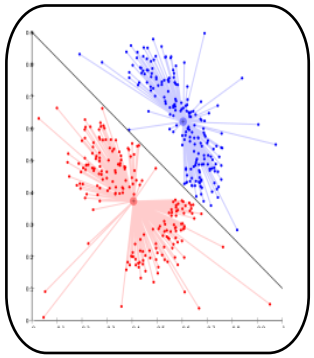






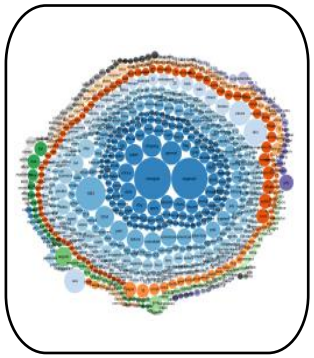
家电 | 新闻 | 冰箱 | 男 | 20-25岁

广告营销数据特点分析-算法



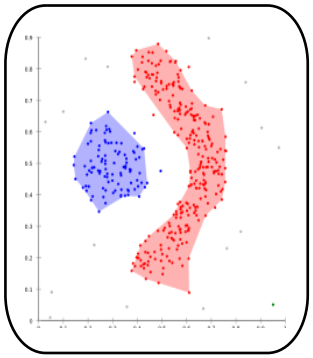
支持向量机（SVM）

- 判断用户男女性别
- 判断用户年龄分段
- 判断品牌投放是否安全？



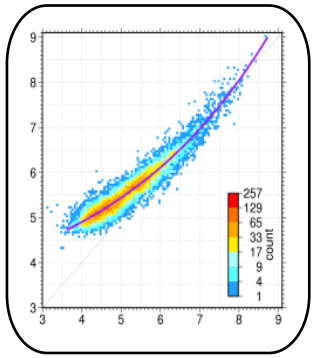
自然语言处理

- 判断页面内容的主题分类
- 判断用户分享内容的兴趣特征
- 判断用户评论的感情倾向



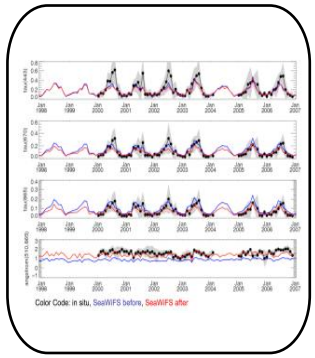
聚类分析

- 根据已有人群查找类似的潜在人群受众
- 根据人群历史数据特征推断人群的学历及收入等属性



回归分析

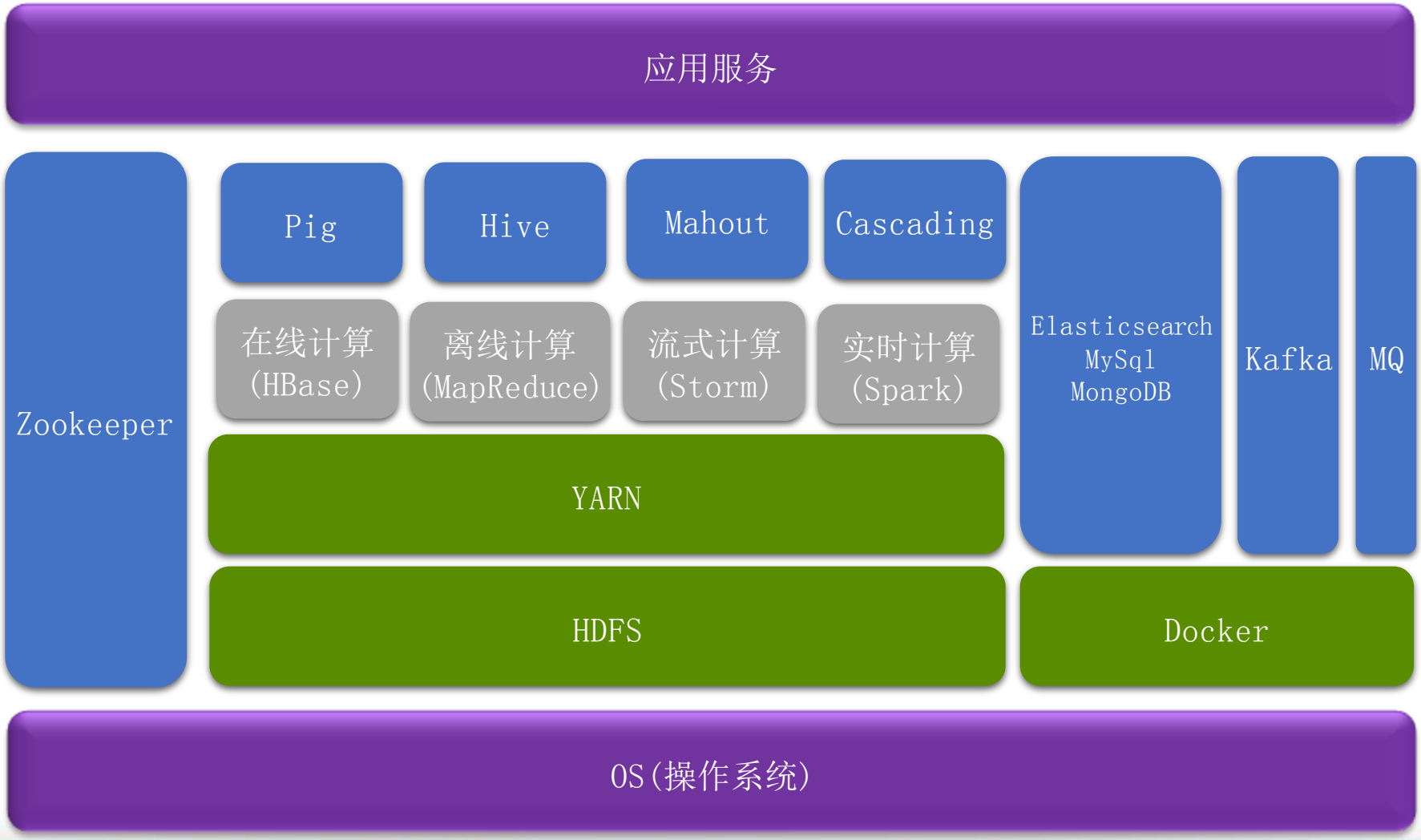
- 依据广告历史数据预测新广告投放的CTR
- 根据历史数据评估广告的综合投放效果



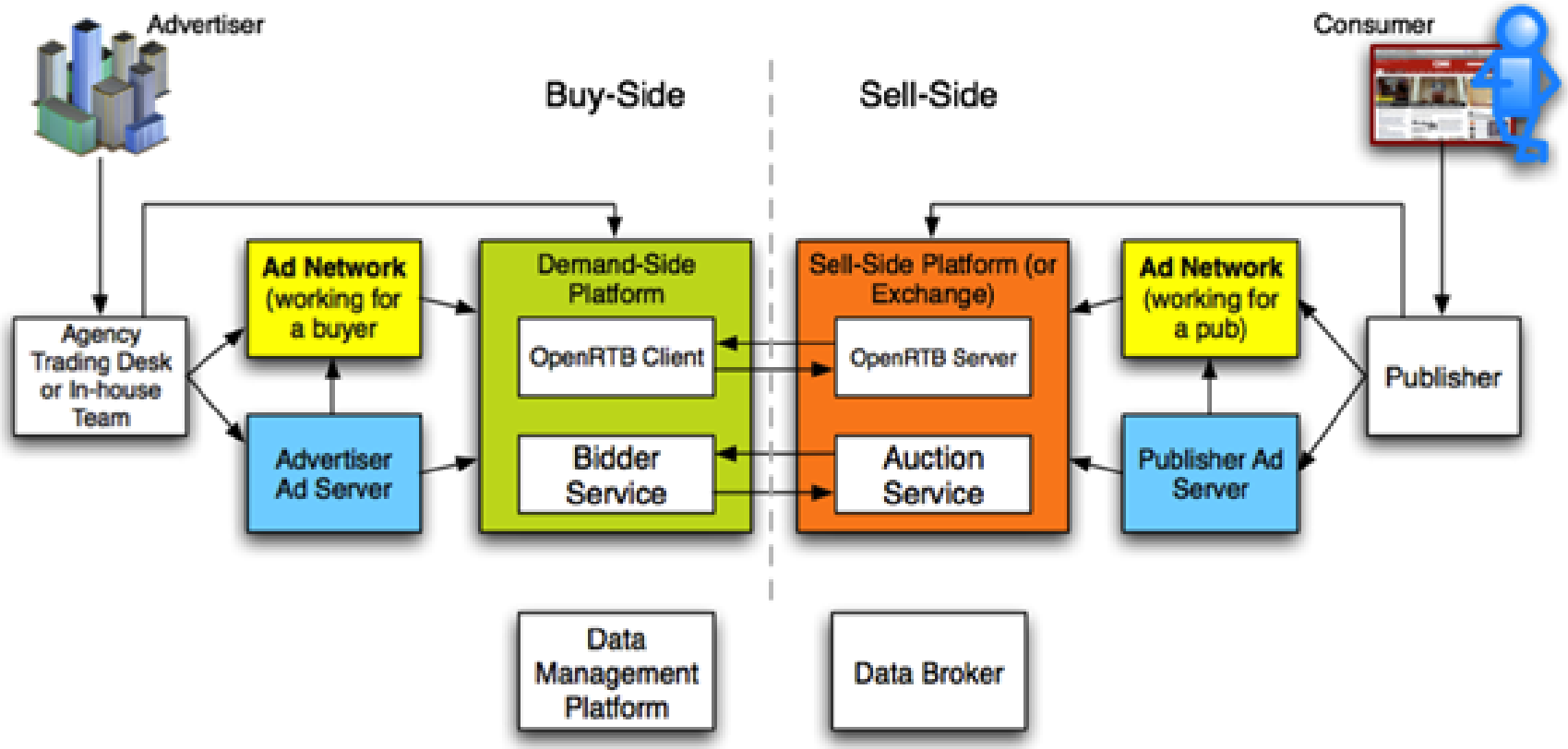
时间序列分析

- 预测用户在特定时期的兴趣强度
- 预测用户在特定时期的购买意愿强度

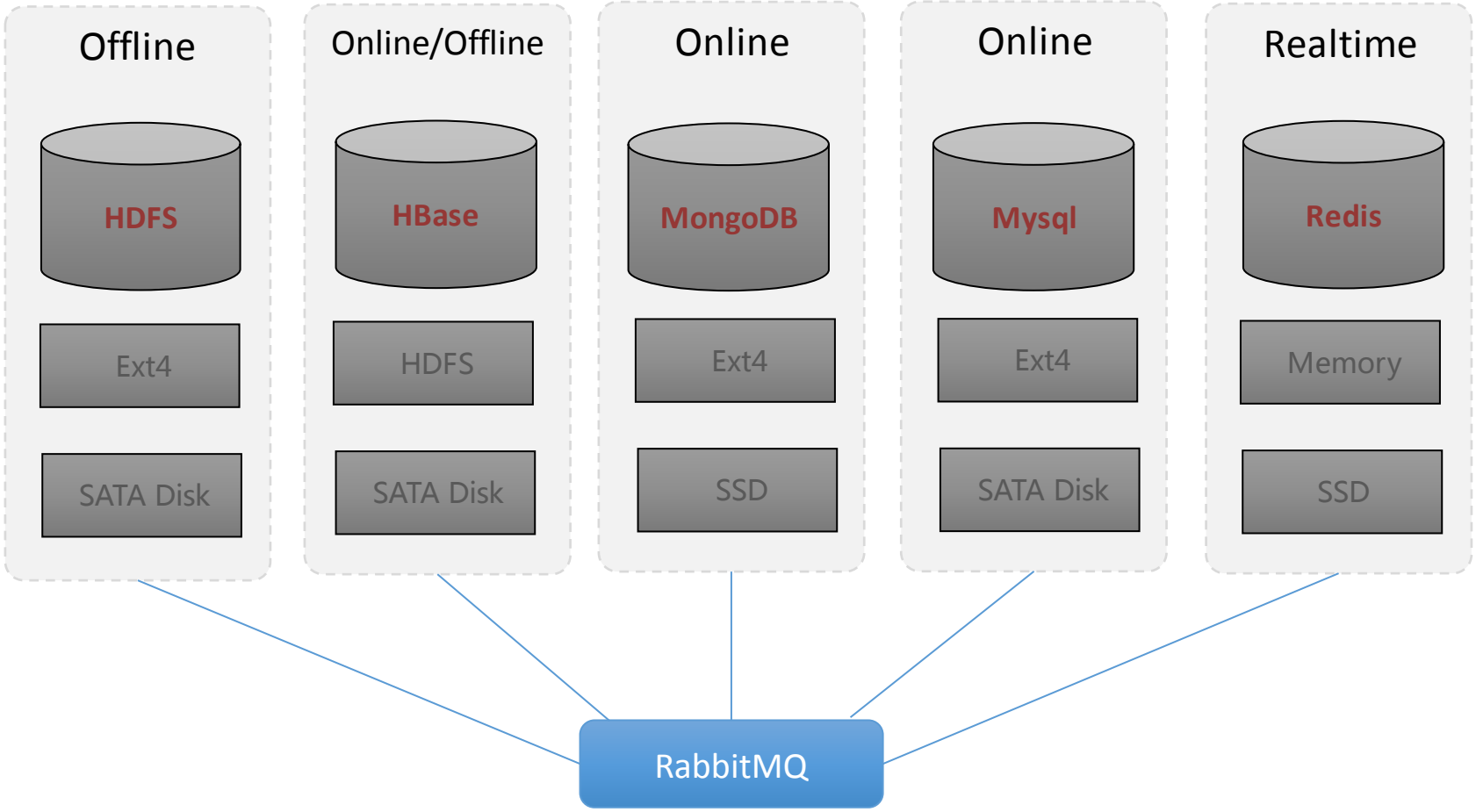
广告数据挖掘平台架构


应用服务
Zookeeper
Pig
Hive
Mahout
Cascading
在线计算 (HBase)
离线计算 (MapReduce)
流式计算 (Storm)
实时计算 (Spark)
YARN
HDFS
Elasticsearch
MySql
MongoDB
Kafka
MQ
Docker
OS(操作系统)

广告数据挖掘平台架构


Advertiser
Consumer
Buy-Side
Sell-Side
Ad Network (working for a buyer
Demand-Side Platform
OpenRTB Client
Bidder Service
Sell-Side Platform (or Exchange)
OpenRTB Server
Auction Service
Ad Network (working for a pub)
Agency Trading Desk or In-house Team
Advertiser Ad Server
Publisher Ad Server
Publisher
Data Management Platform
Data Broker

广告数据挖掘平台架构


Offline
HDFS
Ext4
SATA Disk
Online/Offline
HBase
HDFS
SATA Disk
Online
MongoDB
Ext4
SSD
Online
Mysql
Ext4
SATA Disk
Realtime
Redis
Memory
SSD
RabbitMQ

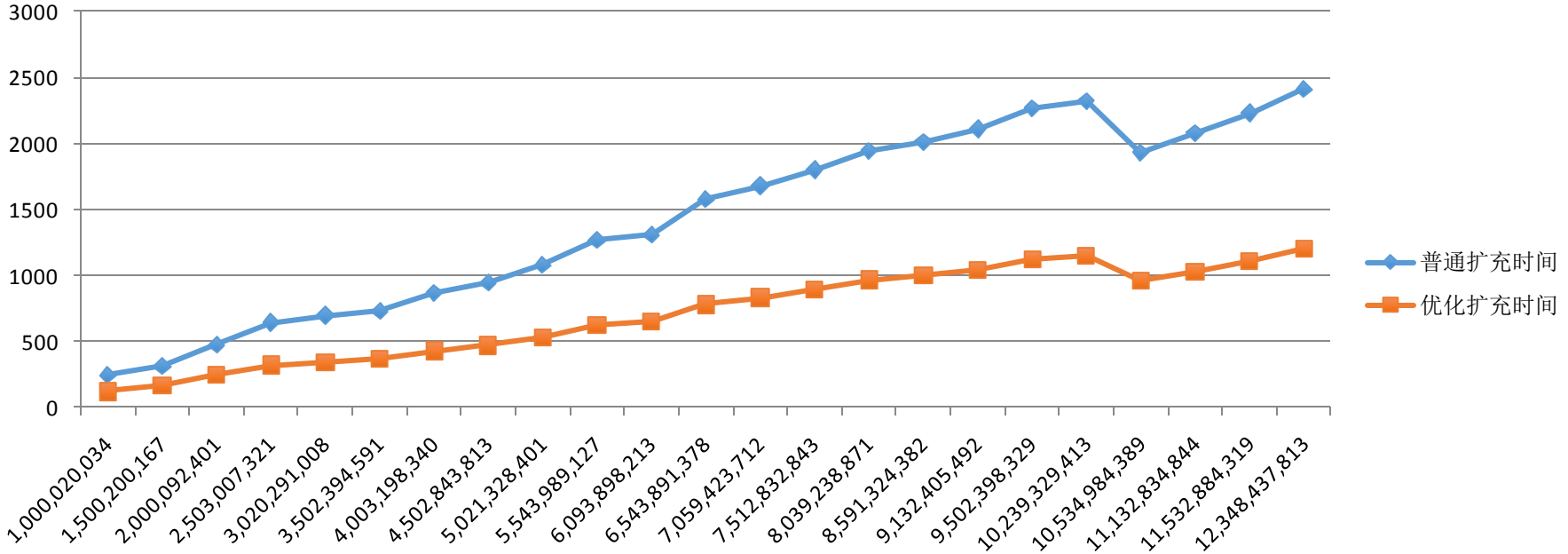
Advertising Distribution Hadoop(ADH)在广告数据挖掘的特点

ADH

- 优化合并过程，使采集数据直接生成客户所需格式，提高处理速度
- 内置广告行业算法，不需要编写MR就可以计算PV、UV等各种维度数据
- 优化HBase查询，专为社会化数据定制，提高处理性能
- 集成数据任务调度系统，可以根据业务需求自动调整计算资源
- 集成Storm，优化Storm传输，减小数据延迟，实时提供数据计算
- 集成Spark，优化迭代工作负载，优化RDD序列化，提高性能和存储效率

ADH在营销数据挖掘的特点-MR

- 减少数据扩充，优化合并过程，使采集数据直接生成客户所需格式，提高处理速度
  - 修改Reduce生成文件格式
  - 提高了近1倍的速度



ADH在广告营销数据挖掘的特点-算法

- 内置广告行业算法，不需要编写MapReduce就可以计算PV、UV等各种维度数据
  - 基础数据
  - 频次数据
  - 增量频次数据
  - 重合数据
  - 独占数据
  - 人口属性数据
  - 来源数据
  - IGRP数据

ADH在广告营销数据挖掘的特点-HBase

- 优化HBase查询，专为社会化数据定制，提高处理性能
- 内置多SCAN实现：
  - MapReduce单表多SCAN场景，优化Map初始化，把速度从O(N)降为O(1)
- 回收策略修改：
  - MinorCompaction
  - MajorCompaction

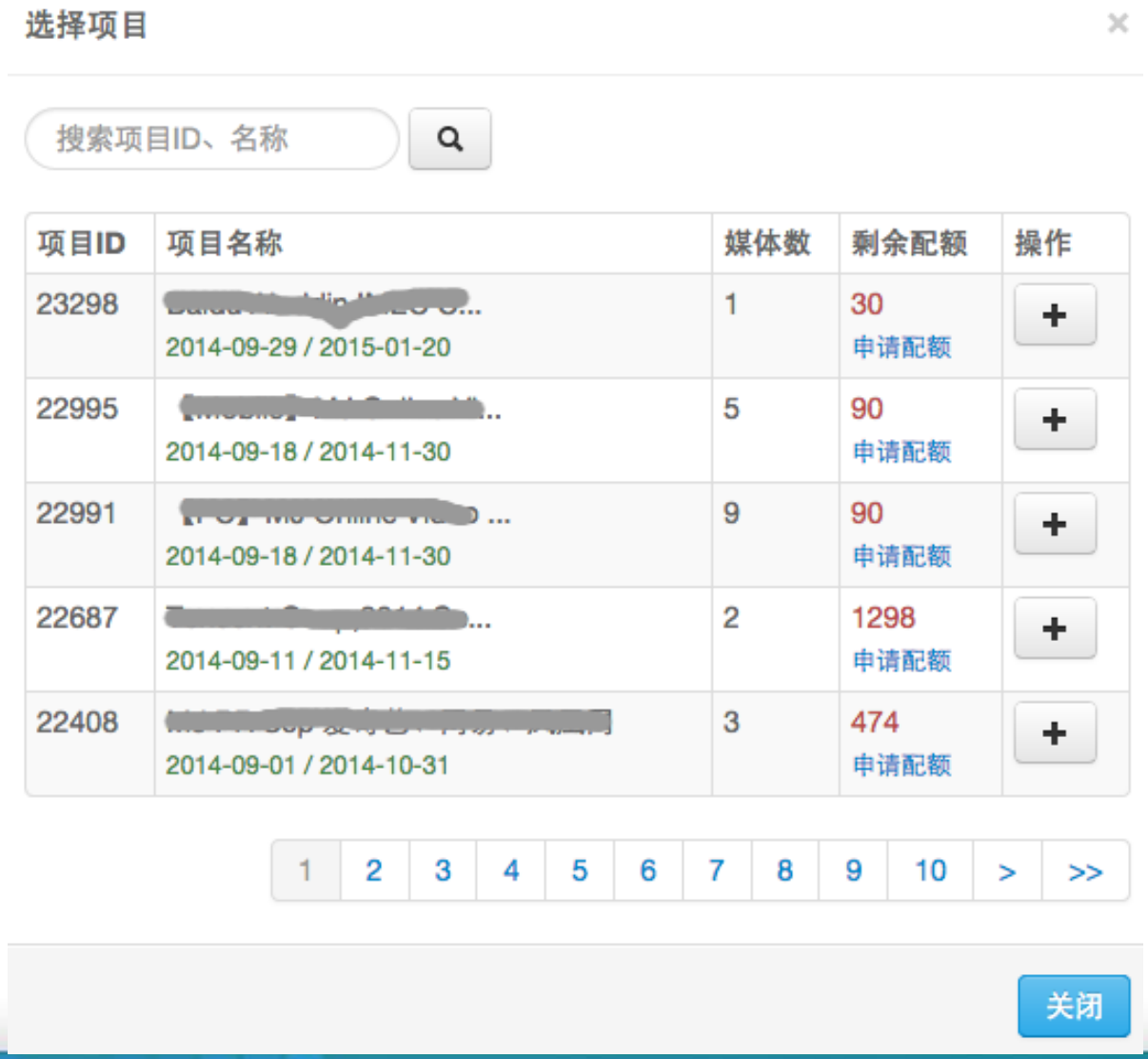
ADH在营销数据挖掘的特点-HBase

```
101    * number of splits matches the number of regions in a table.
102    *
103    * @param context The current job context.
104    * @return The list of input splits.
105    * @throws IOException When creating the list of splits fails.
106    * @see org.apache.hadoop.mapreduce.InputFormat#getSplits(org.apache.hadoop.mapreduce.JobContext)
107    */
108   @Override
109   public List<InputSplit> getSplits(JobContext context) throws IOException {
110     if (scans.isEmpty()) {
111       throw new IOException("No scans were provided.");
112     }
113     List<InputSplit> splits = new ArrayList<InputSplit>();
114
115     for (Scan scan : scans) {
116       byte[] tableName = scan.getAttribute(Scan.SCAN_ATTRIBUTES_TABLE_NAME);
117       if (tableName == null)
118         throw new IOException("A scan object did not have a table name");
119       HTable table = new HTable(context.getConfiguration(), tableName);
120       Pair<byte[][], byte[][]> keys = table.getStartEndKeys();
121       if (keys == null || keys.getFirst() == null ||
122           keys.getFirst().length == 0) {
123         throw new IOException("Expecting at least one region for table : "
124             + Bytes.toString(tableName));
125       }
126       int count = 0;
```

ADH在广告营销数据挖掘的特点-调度

- 集成数据任务调度系统，可以根据业务需求自动调整计算资源
  - Job 配额计算、配额查询
  - 项目的配额分配和更新
  - 查询任务优先级管理

ADH在广告营销数据挖掘的特点-调度



ADH在广告营销数据挖掘的特点-Storm

- 集成Storm，优化Storm传输，减小数据延迟，实时提供数据计算
  - 修改Storm底层传输协议
- 应用场景
  - 实时监控
  - 多机房数据同步

ADH在广告营销数据挖掘的特点-Storm

1. vim storm-core/src/jvm/backtype/storm/messaging/IContext.java (vim)

~ (zsh) | ~ (zsh) | ..social_ma... | ~/tmp (zsh) | ~ (zsh) | ~/tmp (zsh) | yilei@bj-s... | vim (vim) | ~ (zsh) | ..ase/mapr... | ..ckend/col... | ..social_ma... | ..ocail/sina... | ..4.15-cdh...

```
18 package backtype.storm.messaging;
19
20 import java.util.Map;
21
22 /**
23  * This interface needs to be implemented for messaging plugin.
24  *
25  * Messaging plugin is specified via Storm config parameter, storm.messaging.transport.
26  *
27  * A messaging plugin should have a default constructor and implements IContext interface.
28  * Upon construction, we will invoke IContext::prepare(storm_conf) to enable context to be configured
29  * according to storm configuration.
30  */
31 public interface IContext {
32     /**
33      * This method is invoked at the startup of messaging plugin
34      * @param storm_conf storm configuration
35      */
36     public void prepare(Map storm_conf);
37
38     /**
39      * This method is invoked when a worker is unload a messaging plugin
40      */
41     public void term();
42
43     /**
44      * This method establishes a server side connection
45      * @param storm_id topology ID
46      * @param port port #
47      * @return server side connection
48      */
49     public IConnection bind(String storm_id, int port);
50
51     /**
```

ADH在广告营销数据挖掘的特点-Spark

- 集成Spark，优化迭代工作负载，提高性能和存储效率
  - 优化迭代算法
  - 修改RDD序列化方式
- 应用场景
  - 增量频次计算
  - 人群计算

ADH在广告营销数据挖掘的特点-Spark

1. vim storage/StorageLevel.scala (vim)

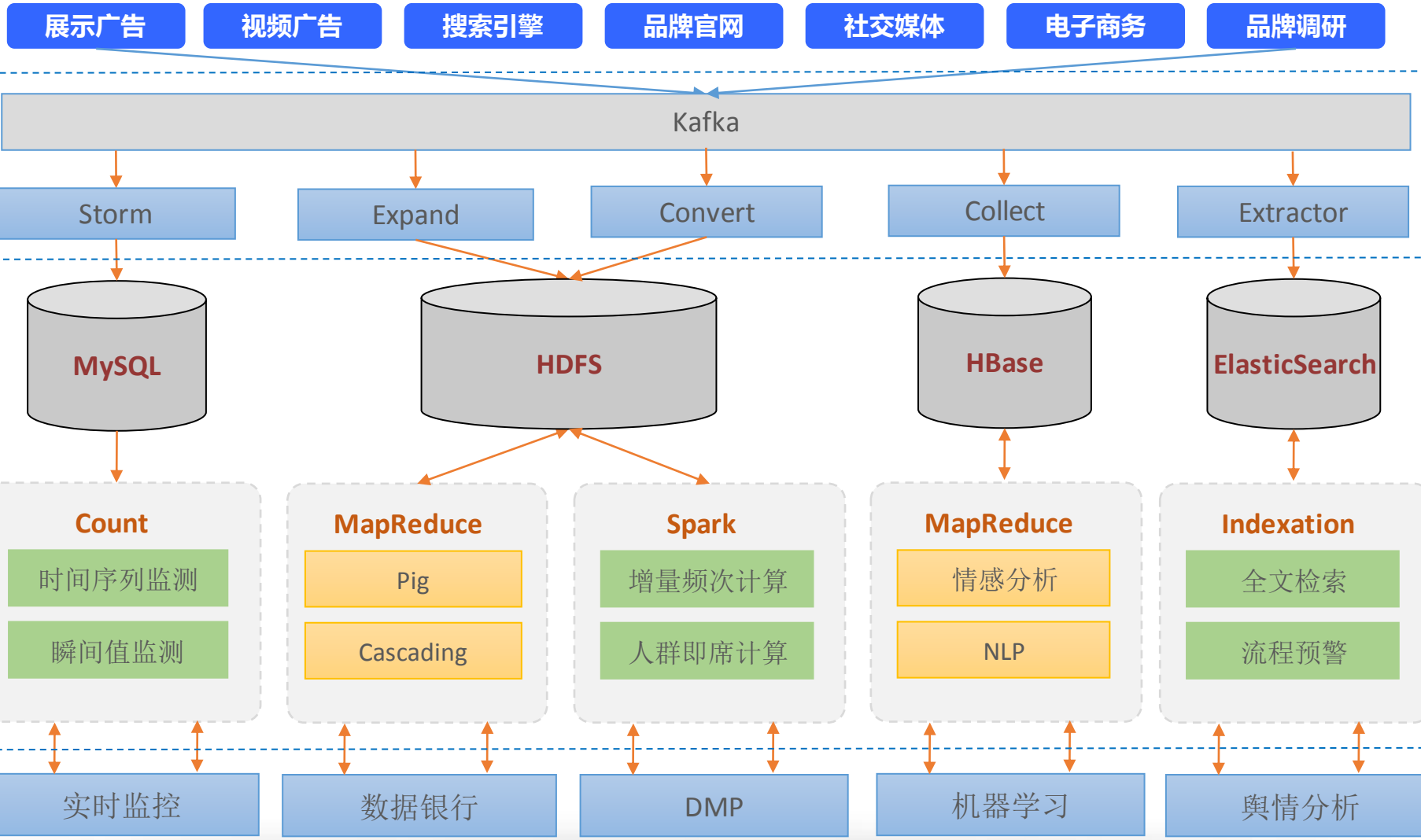
~ (zsh) | ~ (zsh) | ..social_ma... | ~/tmp (zsh) | ~ (zsh) | ~/tmp (zsh) | yilei@bj-s... | vim (vim) | ~ (zsh) | ..ase/mapr... | ..ckend/col... | ..social_ma... | ..ocail/sina... | ..4.15-cdh...

```
137  * new storage levels.
138  */
139 object StorageLevel {
140   val NONE = new StorageLevel(false, false, false, false)
141   val DISK_ONLY = new StorageLevel(true, false, false, false)
142   val DISK_ONLY_2 = new StorageLevel(true, false, false, false, 2)
143   val MEMORY_ONLY = new StorageLevel(false, true, false, true)
144   val MEMORY_ONLY_2 = new StorageLevel(false, true, false, true, 2)
145   val MEMORY_ONLY_SER = new StorageLevel(false, true, false, false)
146   val MEMORY_ONLY_SER_2 = new StorageLevel(false, true, false, false, 2)
147   val MEMORY_AND_DISK = new StorageLevel(true, true, false, true)
148   val MEMORY_AND_DISK_2 = new StorageLevel(true, true, false, true, 2)
149   val MEMORY_AND_DISK_SER = new StorageLevel(true, true, false, false)
150   val MEMORY_AND_DISK_SER_2 = new StorageLevel(true, true, false, false, 2)
151   val OFF_HEAP = new StorageLevel(false, false, true, false)
152
153   /**
154    * :: DeveloperApi ::
155    * Return the StorageLevel object with the specified name.
156    */
157   @DeveloperApi
158   def fromString(s: String): StorageLevel = s match {
159     case "NONE" => NONE
160     case "DISK_ONLY" => DISK_ONLY
161     case "DISK_ONLY_2" => DISK_ONLY_2
162     case "MEMORY_ONLY" => MEMORY_ONLY
163     case "MEMORY_ONLY_2" => MEMORY_ONLY_2
164     case "MEMORY_ONLY_SER" => MEMORY_ONLY_SER
165     case "MEMORY_ONLY_SER_2" => MEMORY_ONLY_SER_2
166     case "MEMORY_AND_DISK" => MEMORY_AND_DISK
167     case "MEMORY_AND_DISK_2" => MEMORY_AND_DISK_2
168     case "MEMORY_AND_DISK_SER" => MEMORY_AND_DISK_SER
169     case "MEMORY_AND_DISK_SER_2" => MEMORY_AND_DISK_SER_2
170     case "OFF_HEAP" => OFF_HEAP
```

AdMaster数据分析平台

- 每天请求数约100亿左右
- 每天增长几TB级数据
- 每天对几千亿条记录进行几百种维度的计算

AdMaster数据分析平台



BDTC
2014 中国大数据技术大会
BIG DATA TECHNOLOGY CONFERENCE
暨第二届CCF大数据学术会议

谢谢

@卢亿雷

johnlya@163.com