个人简历

姓　　名：王　　亚

电　　邮：mr.wangya@qq.com

电　　话：13261749047

通讯地址：北京市昌平区朱辛庄

技能介绍

8 年工作经历，4 年数据分析经验，多次大数据架构实施经验；涉及 Hadoop 平台选型、搭建、维护、优化；以及数据采集(PC/APP/LOG)、存储分析、挖掘建模、搜索等架构设计、部署和优化等；熟练掌握网站分析、应用分析、游戏分析(略懂)常规衡量维度和指标，能从较为全面、专业的角度去找出适应该产品的衡量指标，从而为产品、运营、研发提供产品优化标准和改进方向；掌握基础机器学习(SparkML、numpy)建模方法，可以用 Python 编写简单的算法模型，比如朴素贝叶斯、线性回归、逻辑回归、决策树等，模型优化常用梯度下降算法，工程方面用 SparkML ALS 做过社交 APP 的用户推荐系统、垃圾文本过滤、色情广告图片分类；带着市场和产品思维，用专业技术解决问题。

主要工作：

1、搭建公司 40+节点 Hadoop 集群，日数据 900G，总容量 150TB，集群调优。

2、大数据数据仓库设计、ETL 开发、ODS 数据缓冲、ETL 清洗和自动校验、DW 增量构建、DM 数据集市构建、宽表构建。

3、数据分析和 APP 运营指标统计，如 ARRPU、DAU、LTV、留存、激活等。

4、反垃圾内容过滤，聊天/昵称/头像/相册的色情、广告、涉政违规检测；文本垃圾检测用 Word2Vec&BP 神经网络训练模型，NLP 用 Jcseg，词库用搜狗细胞词库+人工标注；图片检测用 OpenCV SIFT 实现。

5、推荐系统主要做首页推荐，特征抽取用 SparkSql+用户属性宽表+加权指标混合出用户特征矩阵，然后导入 SparkML ALS 协同过滤算法计算用户的偏好矩阵并推荐用户，上线后还要评估多个模型的推荐效果。

6、数据挖掘主要结合基础运营指标和用户行为做用户分群、用户画像以及各种付费相关性分析，给产品运营提供定向分析报告和产品改进建议。

个人信息

姓　　名：	王　　亚	性　　别：	男
年　　龄：	28	籍　　贯：	河　　南
婚　　姻：	已　　婚	户　　口：	非　　农
毕业时间：	2010 年 06 月	参工时间：	2010 年 07 月
最高学历：	专　　科	期望地点：	北　　京

联系电话：	13261749047	电子邮件：	mr.wangya@qq.com
离职状态：	离　　职	期望行业：	互 联 网
入职时间：	一 周 内	外包性质：	暂不考虑

教育经历

2008 年 09 月 至 2011 年 06 月　河南经贸职业学院　计算机应用技术专业　大专

主修课程：ASP.NET 网站开发与设计，数据库应用与开发，软件工程项目开发与管理，计算机网络技术，图形图像处理。

工作经历

上海连舟科技有限公司 2017-12 至 2018-05 大数据经理

工作描述：

- Hadoop 集群搭建与维护
- 数据采集、清洗以及数据仓库建设
- 紫石榴应用商店分析后台
- 创世加 APP 应用指标分析

北京炬鑫科技有限公司 2017-02 至 2017-12 大数据开发工程师

工作描述：

- 数据仓库建设、维度表、宽表、数据集市。
- 数据采集 APP&服务端、清洗、校验、归类、仓库增量更新。
- 净化环境 SparkML 文本转向量（TF-IDF、Word2Vec）分类（朴素贝叶斯、BP 多层感知神经网络），主要应用于用户画像(相同用户兴趣标签 KMeans 聚类)、聊天消息和直播间广告/色情/违禁内容的过滤。
- 大数据开放服务平台 GoLang 接口，提供了很多的基础接口服务，比如 GPS 反向地址解析的、IP 库的、NLP 分词的、BI 报表查询的、OpenCV 图片分类的、用户属性查询的、ES 聊天记录查询的等。
- 基于 SparkML 内置 ALS 算法实现的推荐系统，以促进付费相关性开展推荐，以及 AB 测试等。
- RTMP 直播流实时语音识别，同时支持 1000+路直播音频流分离和切割。

- CDH5.10 集群搭建与维护，部署应用服务包含：HDFS、Zookeeper、HBase、Spark1.X、Spark2.X、Flume、Kafka、Oozie、Hue、Hive、Yarn 等，以及集群升级、节点升降、CM 迁移、单机角色迁移、HA、NodeManage 编组、HBase JVM 性能调优。
- 数据挖掘分析主要使用 SparkSql 统计出来基础指标、结合分类算法看新上线产品数据的情况和付费的相关性指标。
- Scaka on Spark 2.2，是的，我们用的 Java、Scala、GoLang。

业务描述：

一款社交 APP，分为安卓和 IOS，数据采集分客户端和服务端。

日志缓冲落地采用 Flume Http Source、Kafka Channel、Hdfs Sink、Kafka

数据 ETL 采用 SparkSql、JsonSchema、Sqoop、Kettle、DataX、MySqlBinlog

数据分析方案采用 Hive、SparkSql、AkkaHttp、MySql

数据挖掘采用 SparkSql、分析报告

推荐系统采用 SparkSql、SparkML ALS、AkkaHttp、Redis

用户画像采用 HBase、SparkSql、Oozie、Hue

维度宽表采用 HBase、SparkSql、MongoDB、Redis、Hbase RegionObserver

亿级聊天记录、用户属性宽表查询采用 ElasticSearch

净化环境采用 SaprkSql、SparkStreaming、SparkML、AkkaHttp、GoLang

北京银创科技公司 2014-03 至 2016-08 大数据开发工程师

工作描述：

A、数据分析相关（大数据技术实现）

- 通金视频用户行为数据分析系统（video.17mf.com）
- 超级云脑股市舆情资讯挖掘中心（www.17mf.com）
- 娱乐秀场用户行为数据分析系统（www.cube168.com）
- 银创统计（tongji.wybi.net）
- 量化策略平台功能模块用户使用情况分析（www.spcs518.com）
- 消息系统聊天日志数据的采集/存储/分析(NodeJS+Kafka+HDFS+HIVE）

B、金融产品相关

- 华佗诊股（si.mfniu.com）
- 赤兔狙击（warning.88mf.com/t/）

- 视频点播（v.mfniu.com）

C、推广系统相关

- SEM、SEO、整合营销推广系统（dc.mfniu.com）
- URL 短链接管理、跳转分发管理系统

D、金融结算系统相关

- 渤海商品交易所 2 家会员单位佣金结算平台
- 东北亚贵金属交易所 1 家会员单位佣金结算平台
- 上海黄金交易所 1 家会员单位佣金结算平台

E、BI 系统相关

- 通金视频房主魔币结算业务(MYSQL)
- 通金视频、娱乐秀场推广用户、媒体、渠道、游客、注册用户量、听课时长、连通率、在线峰值、流失率等业务的数据分析
- 通金魔方客户软件每个功能使用量情况分析、用户二日、七日回流数据分析
- 娱乐秀场真实用户属性鉴定分析（分析黄牛用户，非机器人）

备注介绍：

1、除 B、D、E1 系列采用.NET 技术外，其他均采用大数据技术实现，数据展现层采用 ASP.NET+SqlServer+ECharts

2、日志采集系统相关的技术采用 Nginx+Lua+Redis 来满足后端业务的离线、实时分析需求

3、数据分析类业务均采用 HDFS+HVIE+HBASE+SQOOP 等技术满足分析需求，大量统计需求最常用的统计时间维度是：5 分、30 分、1 小时、1 天

4、准实时功能类数据业务均采用：KAFKA、ZOOKEEPER、Spark Streaming

5、大数据集群生产环境 CDH5.4.2；线下开发环境 Apache 原生；所有服务器均为自己动手安转、部署、开发、维护

北京安捷天盾科技发展有限公司 2013-04 至 2014-03 软件开发工程师

工作描述：

- 视频监控人脸识别系统（终端机、识别仪、服务器）
- 大规模海量人脸检索
- 快速身份验证仪 3.0
- 人脸识别开放平台（Restful API、平台门户网站、开发者中心）

- 人脸识别门禁、指纹锁

郑州新开普电子股份有限公司 2010-07 至 2013-03 软件开发工程师

工作描述：

- 公交城市一卡通资金清算平台
- 公交城市一卡通公交管理平台
- 公交城市一卡通综合缴费平台
- 网络舆情监测系统

项目经验

2017-12 至 2018-05 紫石榴应用商店分析后台

项目简介：

软件环境：CentOS 7.2、CDH5.11.2、PrestoDB 0.196-lzkj

数据采集清洗完毕之后，落入数据仓库模型中，统计任务主要用 SparkSQL 实现，调度分为按小时、按日调度、按周调度、按月调度，统计报表存储到 Oracle 数据库中，前端可视化采用 Struct2+MyBatis+ECharts+Bootstrap。用户画像属性宽表采用 HBase+MongoDB 二级索引实现。自定义个性化查询提数采用 PrestoDB(Hive、Oracle、MongoDB)实现。后台功能如下：

分类统计：应用数量、浏览量(人数)、展示量(次数)、下载量、查看量

用户统计：登录次数、查看次数、下载次数、安装次数

应用统计：展示次数、点击次数、下载次数、安装次数

时长统计：在线时长

活跃统计：时活跃、日活跃、周活跃、月活跃

留存统计：2 日留存、3 日留存、5 日留存、7 日留存、14 日留存、30 日留存

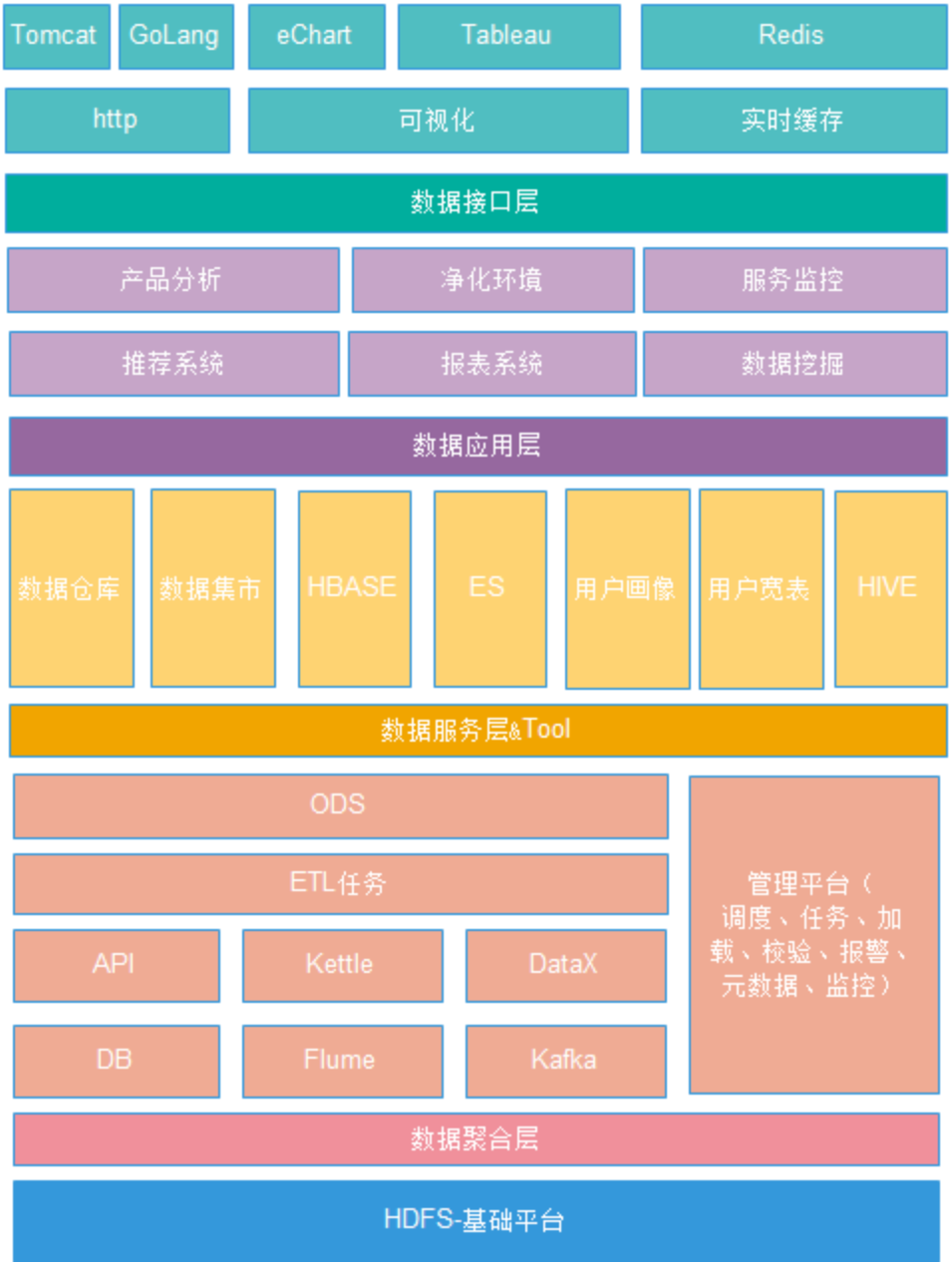
地域统计：活跃用户城市分布(国家、省份、城市、区县)

2017-02 至 2017-12 大数据平台与数据采集

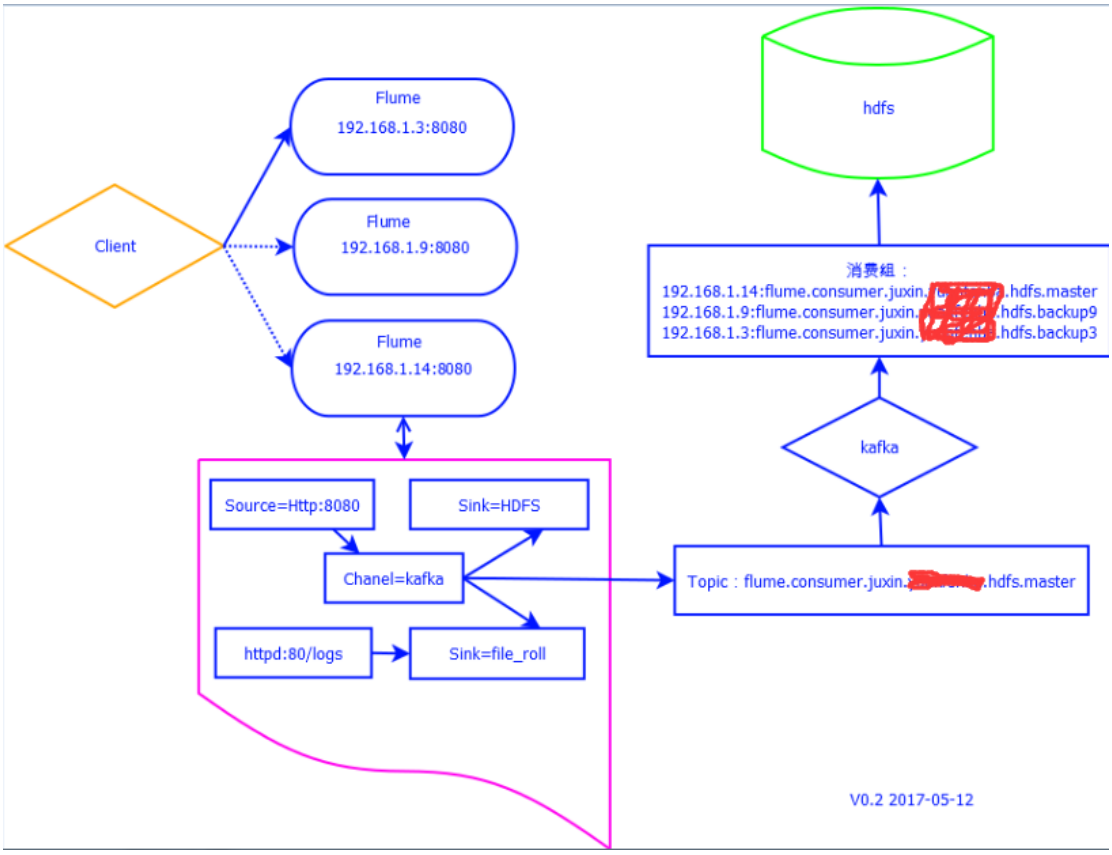
项目简介：

软件环境：CentOS 6.8/6.9、CDH5.10.1

201702 月份之前，旧集群采用的是 HDP Ambari，HDP 内核的优化配置有问题，高负载情况下会导致机器关机且无法重启，后来就升级到 CDH 了，安装方式采用本地自建 YUM 源离线安装，Spark2.X 采用 CSD 服务独立安装，当前的系统架构演变如下图所示：



整体数据架构以数据聚合层为先锋，数据经过逐层加工为上层提供各种个性化数据服务。数据接入层采用经过精心设计的 Flume 组合架构，来确保数据采集的稳定性和可靠性，数据发送方采用 GoLang Channel，Flume 架构如下：



日志数据采集协议我们进行了统一的规范，遵循文档先行，自动校验，自动清洗，自动文档，一个标准 Topic 生成的 Markdown 文档如下：

```
# <span id="OnlineInfo">OnlineInfo</span>
- 用户在线日志
- 正式环境,[返回顶部](#top)

```json
{
  "topic"   : "OnlineInfo",
  "group"   : "com.juxin.[redacted]",                    //组名称,默认值com.juxin.[redacted]
  "id"      : "guid(fa7a3071-90c4-4868-b9ea-0ac1e833545c)",    //消息永不重复的唯一标识,由发送者生成
  "version" : "0.1",                                     //消息版本,默认值0.1
  "message" : {                                          //消息正文
    "uid"              : 0 ,           //用户UID
    "time"             : 1493963414,   //发送时间戳
    "status"           : 0 ,           //状态|1下线|3上线
  }
}
```
```

2017-02 至 2017-12 大数据数据仓库

项目简介：

数据经过数据聚合层校验清洗之后，会增量更新到数据仓库，每小时一个更新周期频率，平均每小时数据量约 40G 左右，清洗耗时约 3 分钟。数据仓库采用雪花型模型构建，以产品业务划分 DM 层，公共维度共享的结构。

数据仓库目前拥有事实主题数据 134 个，维度属性 167，基本涵盖了每个产品线的各个功能模块，数据查询采用 ROLAP ON SparkSql&Hive，主要原因是产品更新迭代很快，采用 MOLAP 模型 CUBE 很容易就分裂了，并导致重新计算。

数据集市划分：直播集市、1V1 集市、捕鱼集市、抓娃娃集市、推荐系统集市、净化环境集市、BI 统计分析集市、数据挖掘集市。

数据仓库主要处理离线业务分析，有 1 小时数据延迟；实时数据处理直接接入 Kafka+Spark Streaming，以 Kafka 多个消费组来完成不同模块业务处理。

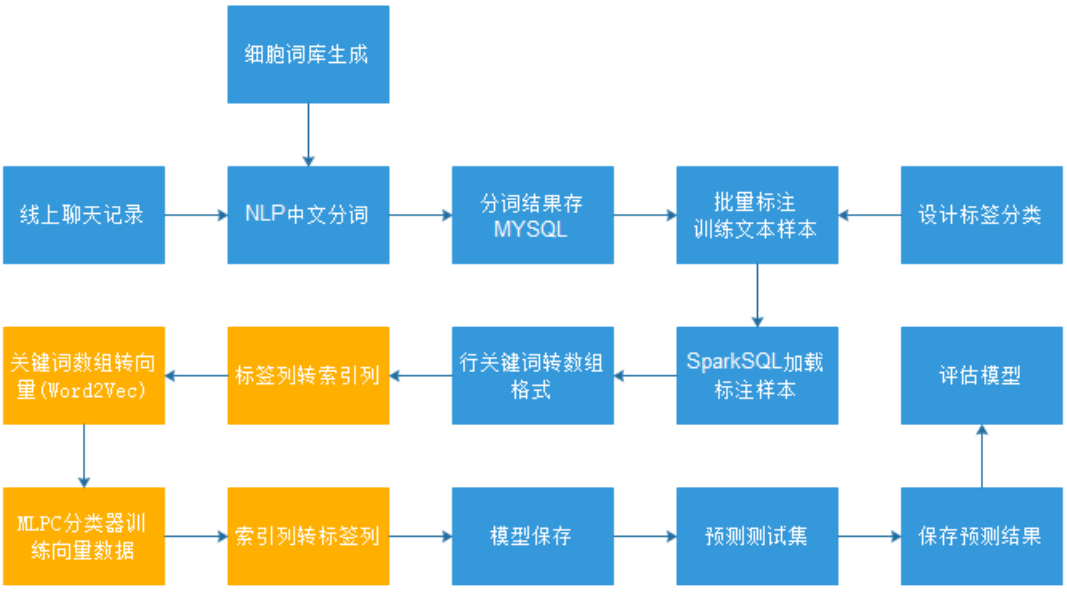
2017-02 至 2017-12 净化环境(内容反垃圾)

项目简介：

软件环境：SparkSQL、SparkML、SparkStreaming、AkkaHttp

分类定义：20 色情|21 广告|22 脏话|22 欺诈|23 违规

多文本分类模型主要检测聊天消息、直播聊天、昵称、标签；文本分类模型经过 2 个版本，第一版采用 TF-IDF、Naive Bayes，第二版采用 Word2Vec、MLPC(多层感知神经网络)，最新迭代的文本分类模型训练流程如下图所示：



图片监测当前采用 Spark Streaming+图谱实时过滤；Opcv SIFT 算法已封装好，正在标注训练集样本数据，还没有上线，图床采用 OSS 存储。

2017-02 至 2017-12 推荐系统

项目简介：

软件环境：SparkSql、SparkML、AkkaHttp、Redis

推荐系统主要做了 APP 的首页推荐、一键打招呼推荐，主要分为 3 个步骤：偏好数据量化和特征向量抽取，协同过滤生成候选列表，候选列表混合排序策略。

偏好数据选取和特征抽取，常用的有视频评价、被查看资料次数、聊天人数、被关注人数、好评率、颜值、同城、好友数、平均视频时长、收礼物数、认证时长、平均在线时长等指标，经过归一化、加权平均最终形成偏好数据评分，形成 A 对 B 偏好数据的整体过程就是：UserA 对 UserB 偏好=(UserA 对 UserB 评价指标*加权 A+UserB 自身指标*加权 B)/(加权 A+加权 B)。

协同过滤算法采用 SparkML ALS 算法训练模型，并产生一个推荐的候选列表，默认 1000 个，候选列表存入 Redis 中，每天凌晨 6:00 推荐 1 批候选用户列表。

候选列表混合排序第一版采用简单随机重复抽样方法，第二版本正采用决策树回归 GBTRegressor 算法优化。

2017-02 至 2017-12 BI 统计分析

项目简介：

软件环境：SparkSql、AkkaHttp

主要提供常用的运营统计分析指标，分为小时、日周期维度，各种常见的运营指标，比如：新增、注册、激活、留存、转化，DAU、LTV、ARRPU、版本、客户端类型、渠道、时长、频率、地域分布、机型、分享、错误分析、功能使用情况等。

统计结果存在 MYSQL，结合 AkkaHttp 提供 Restful 接口。

2017-02 至 2017-12 用户行为数据挖掘

项目简介：

软件环境：SparkSql、SparkML

数据挖掘主要提供定向分析报告，即除常规 BI 统计分析任务之外的需求。主要围绕 2 大块开展：留存相关性、付费相关性。

即用户的哪些行为指标可以导致留存率提升？用户哪些行为指标可以导致付费率提升？挖掘出相关的因素之后，以报告形式提供给产品和负责人，作为产品改进优化的依据。

发布一个版本之后，增加了一些功能，这些功能效果好不好，指标有木有上升或下降，都需要跟踪这些新增功能产生的数据或指标，是否对产品进行了提升。

比如预测流失率属于留存相关性的一部分，APP 内有很多的功能，这些功能都以 Event 的形式进行了埋点，所以用户产生的任何行为全部都会记录下来，将收集上来的 Event 汇总计数，采用 SparkSql 的 Poive 函数保持 UID 维度不变将 Event 旋转，进一步归一化之后得到用户使用 APP 的行为特征矩阵，

再关联上用户属性宽表（比如 N 日留存字段）从而得到 Label 列，经过以上步骤已经把一个用户的属性和行为特征完全填充满了，也是一个训练和测试集了。

挑选一个分类算法，比如朴素贝叶斯，预测用户在 N 日后的留存概率，把预测结果留存概率高的用户和留存概率低的用户归一化之前的 Event 统计结果提取出来，会能明显的看到哪些功能可以显著的提高用户的留存。

作为产品的改进方案，计算出留存概率高的用户功能使用频率的平均指标，将那些留存率概率低的用户筛选出来，结合运营方案和推送系统，向这个高留存率指标的用户靠拢。

更深入的进一步做法，还可以将用户分群，将不同用户群体的功能使用频率指标进一步细化。

关于用户行为数据挖掘，需要结合具体的产品和业务做一些深度的调整，方法也会有所不同，付费相关性暂不介绍。

2015-10 至 2016-02 超级云脑舆情中心

项目简介：

软件环境：Solr、HBase、Kafka、Tomcat、MySql、Jcseg、爬虫技术

作为项目的负责人全程参与项目设计、开发、上线流程，主要职责如下：

1、舆情数据分析挖掘技术方案的设计。

2、网络爬虫采集系统架构的设计与开发。

3、协调产品人员一起完成界面 UI 的设计与数据接口对接方案。

4、功能模块的开发以及项目进步的跟踪。

5、跟进产品功能的测试与上线。

项目用到的技术：HBASE、Kafka、Solr、Jcseg、ActiveMQ、MapReduce、HDFS、Redis，以及少量的 Spark ML 尝试工作内容，SPark ML 主要做文章的自动分类。项目网址：http://yun.17mf.com

2015-06 至 2015-10 阿尔法网站运营统计

项目简介：

软件环境：CDH5.4.2、HIVE、MapReduce、Kafka、HBase、Nginx、LUA、Redis、ASP.NET MVC、Signalr、Linq、EF、RestFul、ECharts

作为项目的负责人全程参与项目的设计、开发、上线流程。

1、需求分析与功能模块业务逻辑文档编写。

2、项目技术架构方案的设计、日志采集系统设计、离线、实时分析关键技术方案的设计。

3、功能模块的开发以及项目进步的跟踪。

流量分析	来源分析	访客分析	受访分析
今日统计(小时)	来源分类	地域分析	热力图
昨日统计(按日)	搜索引擎	系统环境	鼠标轨迹
实时访客	搜索词	新老访客	受访域名
流量趋势	来源域名	忠诚度	受访页面
	来源页面	活跃度	用户轨迹

项目介绍：https://github.com/ljja/SmartAnalytics

2014-08 至 2015-10 阿尔法推广分析

项目简介：

软件环境：HDFS、HIVE、HBASE、MapReduce、ActiveMQ、Kafka

全程负责该项目的开发与维护工作，该项目历史有 2 个版本：

1、基于.NET+SqlServer+MSMQ 方式的数据仓库版。

2、技术 CDH 大数据集群+ActiveMQ 的大数据版本。

该项目全部基于用户行为日志实现，半结构化文本文件格式数据，主要核心业务如下（个性化定制业务）：

1、新增游客、注册用户数量与明细。

2、新增游客、注册用户的来源属性、推广渠道、媒体统计。

3、计费游客、整合营销、SEO 三大渠道的游客、注册用户明细数量的分类。

4、钱眼移动端游客、注册用户数量分类明细统计。

5、视频直播用户的进出日志听课市场统计。

6、关键词与搜索词停留市场分析。

7、通金视频功能标签统计，参见用户行为日志采集系统_trackClsTag 事件（仿京东用户画像原理）。

备注说明：

1、以上统计方式全部统一维度为按小时统计分析。

2、数据统计主要采用 HIVE+MapReduce 实现，可视化展示采用 ASP.NET+SqlServer+ECcharts。

3、HBase 作为每个游客和注册用户的来源的历史属性大表。

4、Kafka 作为实时访客与产品连通日志的实时消息管道，供后端作为连通率统计分析的数据来源。

5、其他用户的行为轨迹、热力图等通用分析已归并到《银创统计》项目中。

6、无特殊说明部分均采用 MR 清洗数据，Hive 统计，HBase 作为历史大表查询。

2013-04 至 2013-07 视频监控人脸识别系统

项目简介：

基于人脸识别、比对技术的网络视频监控系统，系统架构分为三个子系统：终端机、识别机、服务器。系统预热成功后自动从服务器同步黑名单到识别仪；实时通信框架采用 Restful API 确保报警信息实时准确的推送到各台终端机；视频实时人脸比对采用多线程、24 帧频率从黑名单库中比对。

2013-07 至 2013-09 大规模海量人脸检索

项目简介：

系统分为批量导入人像库、检索海量人脸库，以人脸比对技术为核心实现 30 秒内从一千万人像库中检索出相似的目标人。数据库采用 Oracle10GB，MongoDB 作为高速内存缓存。采用多线程、异步并发的方式快速检索一千万数据。

2013-09 至 2013-10 快速身份验证仪 3.0

项目简介：

采用 WPF 技术，在定制主板的 X86 架构上运行的触屏版快速身份验证软件。设备应用场景于金融机构业务办理前置身份验证。采用一比一技术（人、身份证）确保人证正确无误。

2013-10 至 2014-02 人脸识别开放平台

项目简介：

Facecore.cn 服务于互联网广大的开发者，开发者基于简单、易用 API 接口快速开发出基于人脸识别的应用。平台分为：开发者门户网站、API 服务器、运营监控中心。上线第一天 14 小时内独立 IP 访客数量突破 1027 人。

2012-02 至 2012-05 网络舆情监测系统

项目简介：

该项目分为三个子系统：网络信息采集子系统、舆情信息分析监控子系统、舆情信息综合管理平台。系统主要的功能为如下几点：

1、热点话题、敏感话题的识别，可以根据新闻出处权威度、评论数量、发言时间密集程度等参数，识别出给定时间段内的热门话题。利用关键字布控和语义分析，识别该敏感话题。

2、倾向性分析，对于每个话题，对每个发言人发表的文章的观点，倾向行进行分析和统计。

3、主题跟踪，分析新发表文章、帖子的话题是否与已有的主题相同。

4、自动摘要，对各类主题、各类倾向能够形成自动摘要。

5、趋势分析，分析某个主题在不同的时间段，人们关注的程度。

6、突发事件分析，对突发事件进行跨时间、跨空间综合分析，获知事件发生的全貌并预测事件发展的趋势。

7、报警系统，对突发事件，涉及内容安全的敏感话题及时发现并报警。