评语：精硕科技的混合异构营销数据平台在数据实时采集、交互分析等方面具有创新性，商业智能关键技术在数据分析领域具有突破性。

混合异构营销数据平台实践

卢亿雷
精硕科技（北京）股份有限公司，北京 100192

摘要
简要论述了混合异构营销数据平台在实施过程中解决的业务需求问题、用到的相应技术方案选型思考。从技术的角度描绘了构建整个平台过程中的发明创新、遇到的问题和解决方案以及获取的经验教训。

关键词
大数据；精准营销；人工智能；数据可视化

中图分类号：TP311 文献标识码：A doi: 10.11959/j.issn.2096-0271.2018034

Practices of hybrid heterogeneous marketing data platform

LU Yilei
Jing Shuo Technology Co., Ltd., Beijing 100192, China

Abstract
The problem of business requirements solved by the Hybrid Heterogeneous Marketing Data Platform in the process of implementation was elaborated, and the corresponding technical solutions were selected. The inventions and innovations in the construction of the entire platform, the problems and solutions encountered, as well as the lessons learned were depicted from a technical point of view.

Key words
big data, precision marketing, artificial intelligence, data visualization

1 引言

随着大数据技术的飞速发展，越来越多的企业采用了开放架构和开源软件。企业利用大数据技术及其各类应用组件，以自建团队的方式进行信息化建设非常普遍。笔者所在团队在整个数据平台实施过程中，尝试并使用了大量最新的业界开源技术，并自行研发了数据治理、数据生命周期管理、多租户使用和实时监控以及针对客户行业本身的自然语言处理算法及图像识别、深度学习算法，提供广告监控、舆情分析、人群画像等多方面的技术支持，每天处理的数据包含广告数据、舆情数据、第三方数据等多样化且数量较大的数据。为了保证公司多方数据源的对接和融合，满足公司多业务、多产品线的需求，如何利用先进的大数据技术把每天上百亿的数据进行整合和计算是现在面临的主要挑战。

2 存在的问题及解决方案

2.1 现有技术方案存在的问题

针对大数据整合和计算问题，国内外有很多同类型的产品平台，基本可以分为两大类：一类是购买的商业平台；一类是基于开源软件搭建并附加自己的产品逻辑构建的平台。这些产品平台存在以下几方面的问题。

（1）系统资源方面

购买的平台一般成本较高，与公司规模成正比。开源软件也并不是完全免费的，其可能涉及意料之外的实施、管理和支持成本，也可能由于缺乏人员维护而陷入停顿和死亡状态。另外，开源软件也要与自己平台上的其他软件兼容。

（2）功能特点方面

购买的平台一般功能较强，能满足当时的业务要求。但是由于是商业产品，如果接下来业务发展变化，原来的功能不满足需求，需要二次购买。开源软件一般也可以达到要求，但是需要组合使用多种软件。

（3）技术指标方面

购买的软件技术指标固定，系统一般较稳定，但也仅限于中规中矩，不适用于多样的数据营销业务。开源软件开源产品质量差别较大，需要在技术选型时就做好判断，一般需要有专业的技术人员支撑其稳定性。对于数据营销行业来说，没有现成的正好满足需求的产品。

（4）扩展和维护方面

商业购买的软件一般由厂商专门的维护人员负责维护，但是维护人员一般不负责扩展。而且，有的维护是另外收取费用的。虽然开源社区的支持能解决一些共通的问题，但周期相对较长，因此，开源软件基本上靠使用者自身维护。

（5）市场反馈方面

商业购买的软件增加了公司成本，因此相应地，公司的产品价格会较高，而且对市场需求的变化反馈会较慢。开源软件虽然成本低，可以随意切换技术架构，相对灵活，但是每一次技术的革新都是对整个系统的一次考验。

2.2 解决方案

针对现有技术平台的不足，笔者团队提出了分布式混合异构大数据系统平台，对开源软件进行了改造和整合。平台利用先进的大数据技术帮助品牌指导营销策略、预先测算、量化营销效果，实现每月处理超过2 000亿笔数据、迭代超过100种数据模型的目标。该平台具有以下特点。

- 基本没有买入成本，虽然需要投入

研发成本，但这部分投入可以转化为自主知识产权。

● 高度整合了开源软件和公司的业务，甚至可以形成新的产品作为收益的来源。由于对开源软件做了相应的修改，相比简单使用软件来说更加合理和高效，目前国内像百度、腾讯、阿里巴巴这样的大型互联网公司都采用这样的方式。

● 由于整个系统平台是公司技术人员做出来的，因此公司了解系统的每个细节，可以按业务要求做出各种调整，可谓量身定做。

● 平台的建设和扩展维护是贯穿系统整个生命周期的。一直随着公司业务的变化而进化，并且可以选择性地吸收开源社区的各种改进。该平台的数据处理流程如图1所示。

由于针对业务做了系统构建，该平台可以充分利用所有资源，性价比较高，而且能够快速地反馈市场需求。

3 平台架构与关键技术

3.1 整体架构

平台整体架构如图2所示。该平台针对不同行业领域的信息进行爬取，可以支撑电商、微信、微博、新闻、论坛、问答等多种渠道每天数十亿的互联网页面信息的爬取，是大数据平台和业务需求中不可或缺的文本型半结构化数据。图2中，Blue Air是反作弊系统，Zing是为数据科学家提供的支持多种语言的计算平台。

图1 数据处理流程

在多维度、多结构的数据源和大数据平台强大的数据处理能力的支持下，系统中人工智能（artificial intelligence，AI）和商业智能（business intelligence，BI）的经验和模型发挥了极大作用，数据聚类算法、自然语言识别技术（neuro-linguistic programming，NLP）、数据关联分析、受众画像建模等技术得以快速发展，提升了公司的技术水准和服务效率，也帮助客户获得了更好的营销效果。

3.2 关键技术

3.2.1 大数据存储与管理技术

（1）分布式存储技术

攻关元数据管理、系统弹性扩展技术、存储层级内的优化技术、应用和负载的存储优化技术，针对营销活动的各环节，采用不同的方案，从最底层优化存储，并为后续的计算优化打下基础。

（2）存储方案

根据不同行业的特性，针对数据的管理、查询以及分析的性能需求，数据规模和吞吐量的增长需求，主要采用关系型数据库、非关系型数据库、实时数据库、列式数据库等存储方案。

3.2.2 数据可视化技术

传统的数据可视化工具仅将数据进行组合，通过不同的展现方式提供给用户，用于发现数据之间的关联信息。采用可视

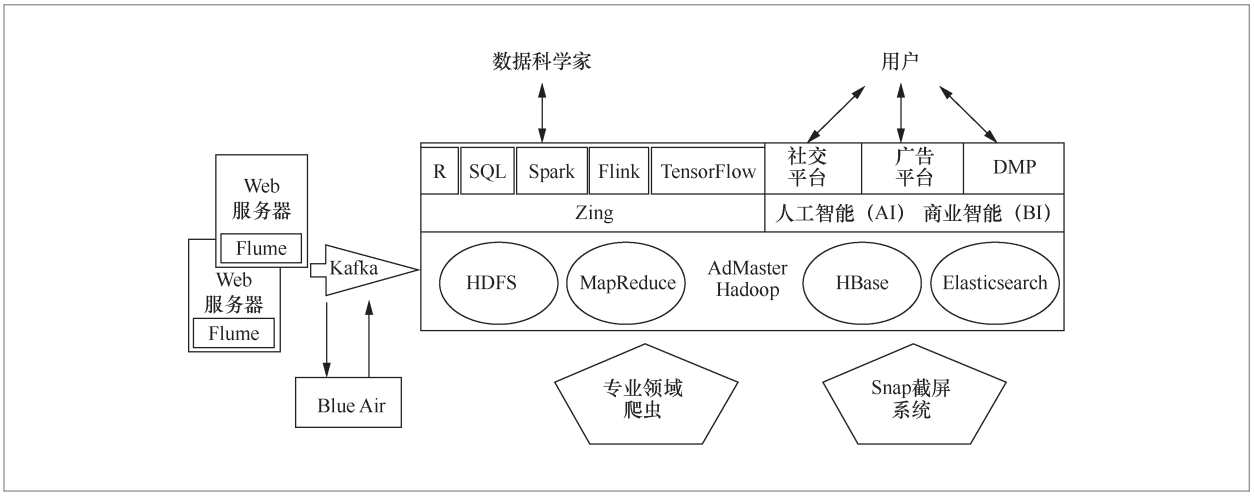

图2 平台整体架构

化开发工具Spark、GraphX、Zeppelin、Matplotlib、recharts和moojnn等解决方案或技术，保证营销数据可视化应用效果，能够快速地收集分析数据并对数据信息进行实时更新。数据可视化展现包括表格、图表、图形、示意图、地图和标签云图，从而使用户能够创建简单的仪表板或监管信息图表和可视化效果。

通过最新的HTML5、React、Redux等前端技术，最大化呈现丰富的数据内容，使用响应式设计，可同时兼容平板电脑、智能手机和个人计算机，实现随时随地的移动监管。

3.2.3 大数据分析与挖掘技术

采用统计分析技术，把计算推向数据并且并行处理，使Hadoop获得了强大的深度分析能力。经过算法的并行化，突破原有的可处理数据的技术瓶颈，利用并行计算模式大幅度提高了工具的计算性能，同时赋予了系统技术深度分析的能力。

3.2.4 人工智能和商业智能技术

在人工智能方面，在NLP、图像模式识别、跨设备多ID用户识别等方向取得了可喜成果，并运用在了社交情感分析、数据管理平台（data management platform，DMP）和受众分析等商业产品。可以在上百亿条的数据上分析客户关注的关键词的社会反响，涵盖新闻、论坛、微信和微博等公众平台。经过历史验证，情感分析的准确度已经可以达到70%以上，热点事件发现时延在1 h以内，让客户可以第一时间知道自己产品的社会影响，同时也可以知道竞争产品的相关情况。

在图像模式识别方面，采用了深度学习算法，笔者所在单位应用于Snap监播实录的特定广告识别技术，使用自研的特征识别和匹配算法，部署在40多台服务器的计算集群上，全天候不停歇地对已经投放的广告进行识别，每天监控的广告数在5亿次以上。不仅可以识别简单的横幅广告，还可以对视频广告进行有效的识别。这项技术的应用，代替了以往占用了大量人力的“人肉广告作弊识别”，为广告代理公司和顾问团大大降低了成本，而且监控的范围也比以前有了100倍以上的提升。随着手机和各种移动电子设备的普及，拥有两部以上的手机或电脑和其他

智能设备已经变得相当普遍，如果可以实现跨设备多ID用户识别，无疑可以更好地帮助广告主识别受众。笔者团队充分利用多年数字营销的经验，以超过200亿条的广告数据为基础，基于每秒400万次并发查询的高效key-value查询系统，研发了多ID识别算法，很好地实现了一人多机的识别。

在商业智能方面，主要致力于多维数据组合查询、用户自定义数据分析、基于“人—事件”的受众人群分析等方向。具体来说，多维数据组合查询主要应用于广告数据分析。笔者团队的大数据分析平台，每天分析100亿条以上的广告数据，从50多个维度的各种组合来透视这些广告数据的价值，每天早上5:00以前为所有广告主提供准确的报表。使用HTTP 2.0及WebSocket技术，动态平滑升级，可以做到用户无感知。数据查询速度也在秒级完成。作为固定多维度查询的补充，广告系统还支持用户自定义数据分析：通过提供一整套的可视化数据处理接口，用户只需要在页面上用鼠标点几下，就可以按自己熟悉的语言来定制特殊的数据处理逻辑。笔者团队提供近100个营销数据维度作为原材料，用户可以按时间、地域、广告媒体、广告位置、广告创意等多个细分维度计算数据的相关性，计算数据分布和其他多种指标，1 TB数据的处理时延在10 s以内。基于“人—事件”模型的受众人群分析系统，则按“以人为本”的使用方法提供全方位的支持。

本系统综合了广告产品线、社交产品线、调研产品线和数据交换产品线等多条产品线的所有数据，以受众为中心，把所有信息串联起来，建立了一套基于广告受众数据的类Siri智能问答系统，除了可以回答“这个ID在什么时间什么地方上过哪些网站”这类简单问题，还能回答诸如“这个ID是什么性别”“这个ID大概什么年纪”“这个ID喜欢足球吗”“这个ID家里有婴幼儿吗”等问题，结合上文提到的多ID识别，它还可以回答“这两个ID是同一个人吗”这样的问题。这一系统已经作为笔者团队的一项独立对外业务开放。

3.2.5 数据质量控制技术

从业务系统中获取的数据首先进入操作数据存储（operational data store，ODS），抽取—转换—加载（extract-transform-load，ETL）引擎根据元数据规范和业务规则对ODS数据进行整合处理，正常数据进入主题数据库完成整合，异常数据进入数据质量管理系统进行人工干预。数据质量管理系统对异常数据进行人工干预，通过人工干预可处理的批量数据进入临时库，由ETL根据新的规则进行处理。通用的干预措施形成元数据规范，进入元数据规范标准；无法处理的数据暂存到临时库，通知业务单位处理。

4 应用场景

4.1 广告效果分析

笔者团队的广告效果分析业务每天有百亿以上的数据增量，需要同时提供“天”级别的准确报表和秒级别的实时报表，并支持用户自定义的各种维度的组合查询。应用了以上系统后，完美地支撑了广告分析的全线业务，真正做到了准确又实时、灵活又高效。

4.2 受众标签计算

在混合异构大数据平台上，以多种格

式海量数据为支撑，运用机器学习的方法，受众标签计算得以高效运行。其中综合了用户的行为数据、第三方的标签数据、多ID的打通、社交数据分析等在百亿级数据集上的运算，为各产品线提供稳定可靠的服务。

4.3 定向广告投放

依托于混合异构大数据平台的广告投放系统，可以对复杂多样的投放要求提供100 ms以内的快速响应，每一个需求的满足都整合了最新的数据变化，而这个新数据的量级在每秒3万次以上，正所谓“线上一秒钟，线下十年功”，正是这个平台的支撑，才使SmartServing系统运行如此轻快。

5 应用效果

该应用平台部署之后，达到了非常理想的效果。

- 高并发数据采集服务端单个采集节点可以支持每秒12 000次的数据采集请求，并且支持分布式部署，从而支持全国范围的营销数据采集服务。满足全国多机房实时数据收集汇总的管理平台需求，数据汇总时延不超过3 min，兼顾备灾容错，保证数据不错不漏。自动识别服务器和数据结构的差异，并自动优化配置。
- 针对数据结构复杂、数据源多的业务特点，采用消息队列架构，为多种渠道的数据源对接提供了实时和稳定的消息收发支撑平台，实现了系统间协作的零时延。
- 混合异构大数据平台还提供了交互式的数据分析平台，支持数据建模团队使用R、SQL、Spark、Python等多种方式，进行实时的交互式数据分析和图表展示。数据科学家和数据工程师可以同时使用同一平台处理上千亿条的数据。
- 信息爬取系统可以支撑电商、微信、微博、新闻、论坛、问答等多种渠道每天数十亿的互联网页面信息的爬取。
- 采集大量地理经纬度和地理信息点（point of information，POI）信息。对于这些数据，混合异构大数据平台会自动进行曲面几何转换，再进行存储利用。
- 利用多数据源的整合和复杂数据结构的兼容，全方位地分析和整理多维度的数据、自动化调度优先级、智能调配计算资源，为业务决策提供有效支持。
- 计算任务瓶颈的辅助分析工具，为研发、测试、建模等团队提供代码级的优化建议。
- 利用页面仿真录屏系统，可以对用户浏览页面乃至广告的过程进行模拟并录屏，通过图像识别及深度学习，真实地还原页面浏览、广告播放的过程，也是重要的网页采样和信息留档数据。
- 构建了基于大数据的自动分析和建模体系，并在数据入库前的清洗阶段引入了虚假数据过滤模型（反作弊技术），模型可以动态自动学习和更新识别模式库，从而在第一时间发现异常数据，并可以自动向各相关业务系统发出异常预警。

混合异构大数据平台项目深度整合了多个开源大数据平台，并融入了自己的技术基因，真正把开源技术灵活运用，一方面解决了公司在技术研发过程中遇到的问题，另一方面回馈开源，为大数据的综合应用方案提供了可行的案例。

6 结束语

本文介绍了混合异构大数据平台的关键技术与应用。在平台开发过程中，笔者总结了一些技术经验和教训，具体如下。

● 多个开源软件整合时，需注意不同软件依赖的版本，从而提高开发效率，尽量少走弯路。

● 操作系统级别要尽量做到环境一致，比如Openfiles、HugePage等的配置，需要由运维部门统一提供最优的配置。

● 选择开源系统时，尽量根据最简单、用户最多、自身团队较熟悉所使用语言的原则选择合适的开源系统。

● 开源软件也可能存在错误或不支持某些需求，必要时需根据自身需求更改源代码。

● 需经过多次迭代，方可搭建稳定高效的数据平台。

参考文献:

[1] DEAN J, GHEMAWAT S. MapReduce: simplified data processing on large clusters[C]//The 6th conference on Symposium on Opearting Systems Design & Implementation,December 6-8, 2004, San Francisco, USA. New York: ACM Press, 2004, 51(1): 10.

[2] CHANGF, DEAN J, GHEMAWAT S, et al. Bigtable: a distributed storage system for structured data[J]. ACM Transactions on Computer Systems, 2008, 26(2): 1-26.

[3] POLYZOTIS N, ROY S, WHANG S, et al. Data management challenges in production machine learning[C]//ACM International Conference on Management of Data, May 14-19, 2017, Chicago, USA. New York: ACM Press, 2017: 1723-1726.

[4] VENKATARAMANS, YANG Z H, LIU D. SparkR: scaling R programs with Spark[C]//International Conference on Management of Data, June 26-July 1, 2016, San Francisco, USA. New York: ACM Press, 2016: 1099-1104.

[5] LI H Y, GHODSI A,ZAHARIA M, et al. Tachyon: reliable, memory speed storage for cluster computing frameworks[C]// ACM Symposium on Cloud Computing, November 3-5, 2014, Seattle, USA. New York: ACM Press. 2014: 1-15.

[6] BAILISP, DAVIDSON A, FEKETE A, et al. Highly available transactions: virtues and limitations[J]. Proceedings of the VLDB Endowment, 2013, 7(3): 181-192.

[7] SHUTE J, OANCEA M, ELLNER S, et al. F1: the fault-tolerant distributed RDBMS supporting Google's Ad business[C]// AcmSigmod International Conference on Management, May 20-24, 2012, Scottsdale, USA. New York: ACM Press, 2012: 777-778.

作者简介

卢亿雷（1980-），男，精硕科技（北京）股份有限公司技术副总裁，主要负责数据的采集、清洗、存储、挖掘等整个数据流过程，确保提供高可靠、高可用、高扩展、高性能系统服务，提供Hadoop、HBase、Storm、Spark等离线、流式及实时分布式计算服务。

收稿日期： 2018-02-09