



面向 DT 服务的数据分析系统架构

汪保友, 吴琮, 钱晶, 陶晓英

(中国联合网络通信集团有限公司上海市分公司, 上海 200050)

摘要: DT 时代, 数据分析系统肩负着运营商业创新和运营强支撑的更高要求。首先阐述了深化数据分析、增益数据价值所面临的需求与挑战; 接着, 提出面向开放服务的逻辑架构和技术架构以及数据治理与数据安全措施, 以实现能力开放、架构灵活的运营支撑; 最后示例介绍该平台的服务场景以及取得的良好效果和价值体现。

关键词: 大数据; 数据分析; 系统架构; 数据服务; 数据治理

中图分类号: TP311

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2016021

Architecture of service-oriented data analysis system

WANG Baoyou, WU Cong, QIAN Jing, TAO Xiaoying

Shanghai Branch of China United Network Telecommunication Co., Ltd., Shanghai 200050, China

Abstract: In the era of data technology, the higher requirements of telecom operators, such as business innovation and strong operation supporting, are being placed on data analysis system. Firstly, the requirement and challenge of the rising data analysis and the enhancing data value were expounded. Then, in order to realize open architecture and flexible operation support, the logical framework and technology architecture of the service-oriented data analysis system, as well as the data governance and the data security management were proposed. Finally, the service scenarios and the business value of the platform were introduced by an example.

Key words: big data, data analysis, system architecture, data service, data governance

1 引言

移动互联网时代, 变革日新月异, 在“互联网+”大潮下, 具有雄厚资金技术和用户规模的 BAT 公司频频涉足通信、智慧城市、云计算、物联网等信息服务领域, 各类 OTT 应用也动摇了运营商短信和语音业务收入的基础。虚拟运营商蓄势待发和互联网企业的跨界竞争对基础运营商而言, 呈现出空前的压力和挑战。大数据产业的发展, 给运营商带来了新的曙光。运营商处在大数据产业链的传输与交换中心地位, 具有丰富的高价值数据资

源, 具有开展大数据业务的独特优势。在这个数据为王的时代, 传统 IT 走向 DT。面对海量复杂数据和创新型应用, 面向 DT 服务的逻辑架构和技术架构以及数据治理与数据安全措施, 以实现能力开放、架构灵活的运营支撑; 最后示例介绍该平台的服务场景以及取得的良好效果和价值体现。

2 需求与挑战

运营商为消费者提供网络服务, 处在大数据产业链的传输与交换中心地位, 具有相对全面完整真实的高价值密度信息。包括用户的语音通话行为、上网痕迹信息、短/彩信使用信息、SP/CP 使用信息、地理位置移动信息、用户实



名信息、信令数据、网管日志数据、渠道偏好、终端喜好信息以及挖掘衍生的用户价值、用户消费习惯、交际圈信息、生活行为指纹、个人兴趣爱好等信息。

2.1 面对的需求

面向 DT 服务的数据分析系统,对深化数据分析和数据挖掘和增值数据价值提出了更高要求,需要面对企业内部支撑强能、企业外部创新增值的需求召唤。

(1) 精准化营销管理需求

电信行业客户群庞大、需求差异也大,运营商需要分析用户行为偏好,建立用户全息画像,通过标签化用户管理,细分市场、客户群,将最合适的业务及时、准确地推销给最需要的客户,实现业务和客户的最佳匹配,改善用户体验,提升用户消费和忠诚度。

(2) 流失预警与维系挽留需求

电信行业的竞争是大家有目共睹的。运营商间持续不断的“价格战”,使得客户变更业务及转网的选择余地和意愿越来越大,合约用户协议到期或促销活动优惠期结束后,大量低忠诚度用户通过转网或弃卡重入网方式,套取新的补贴优惠。有些老用户也会因各种原因,存在流失风险。如何提早识别预警流失倾向用户,对高价值、高风险的流失用户维系挽留,是运营商提升存量用户和收入保有率的需求。

(3) 创新增值需求

在不侵犯用户隐私的前提下,对数据进行深度加工,形成大数据能力输出和数据平台开放,服务于社会,是运营商实现“数据资产”运营和商业模式转型的业务创新需求。

2.2 面临的挑战

现有经营分析系统转型升级为面向 DT 的数据分析系统,需要面对如下挑战。

(1) 平台能力待提升

目前经营分析系统是服务企业内部,主要功能体现在:KPI 监控、统计报表、综合分析和客户群细分等方面。虽然取得了一定效果,但仍存在数据分散、分析内容不全面、展现界面不友好、同名指标不一致等方面的问题。经分使用者很难迅速查到自己想要的信息,特别是管理者很难一目了然地看到公司运营出现的关键问题。突出体现在数据量大与使用率低的矛盾、功能灵活与使用简便的矛盾、业务多变与指标稳定性的矛盾。经营分析人员在使用经分系统时,往往还需要进行大量的数据汇总整理、数据合并、深入分析和数据挖掘等工作,工作难度大、工作量大、效率

低;管理者和决策者在使用经分系统时,很难直接看到想要的信息和分析结果,难以迅速有效进行决策。

(2) 平台架构待扩展

目前经分系统采集的数据主要是 BSS 数据,处理的都是账单、详单、用户资料、产品服务订购、业务受理记录等结构化数据,包括 CRM、计费、账务、OCS、总部下发清单、结算、客服等,缺乏对 OSS 域、MSS 域的数据采集聚合。处理海量数据时系统性能瓶颈过大,存在各类时间窗口无法控制的风险。需要引入分布式并行处理技术架构,扩展数据采集范围,提高采集时效性、完备性。

(3) 数据治理需提高

运营商的数据分散,大数据 4V 特征明显,因各种原因存在少量“脏”数据,采集数据时需要排除数据噪音,分析数据时需要加强数据关联性,健全数据质量管控机制,提升数据质量。对涉密敏感信息需要脱敏加密,加强数据输出服务的安全防护和流程管理,规范数据资产运营机制。

3 面向 DT 服务的数据分析系统架构

基于以上的需求与挑战,需要构建“能力开放、灵活支撑、安全服务”的面向 DT 服务的数据分析系统,有效支撑各类创新应用,提升数据资源的价值转换能力。

3.1 系统逻辑架构

数据分析系统一般分为集团总部、省分两级架构,省分数据分析系统逻辑架构如图 1 所示。

省分数据分析系统主要包括 4 层结构,即数据获取层、数据存储层、数据应用层和数据服务层。采集的数据源包括 CRM、计费、账务管理、OCS、客服、VAC 平台、cBSS 等系统的基础数据以及集团总部下发文件明细数据、扩展采集整合 OSS 域、MSS 域的网管日志、网元信令、地理位置移动信息、上网痕迹、财务核算等数据。数据采集的方式可以是接口文件、DBlink 访问、ETL 加工处理、人机交互等方式。通过系列加工流程、深度分析和信息挖掘,在数据存储层形成企业数据仓库和数据集市。数据应用层表现形式包括运营总览、预定义报表、纵向上传、异动监控、电子书、即席查询、OLAP 分析等基础功能,精准营销、客户维系、网格管理等生产支撑体系以及外部输出服务功能。在数据服务层,可通过个性化定制、信息推送、用户搜索、能力开放等方式,实现对内对外服务。在整个数据加工处理、流转服务过程中,数据质量、数据标准、元数据、生命周期等数据管理措施贯穿始终。通过安全制度、安全技术、安全运营、安

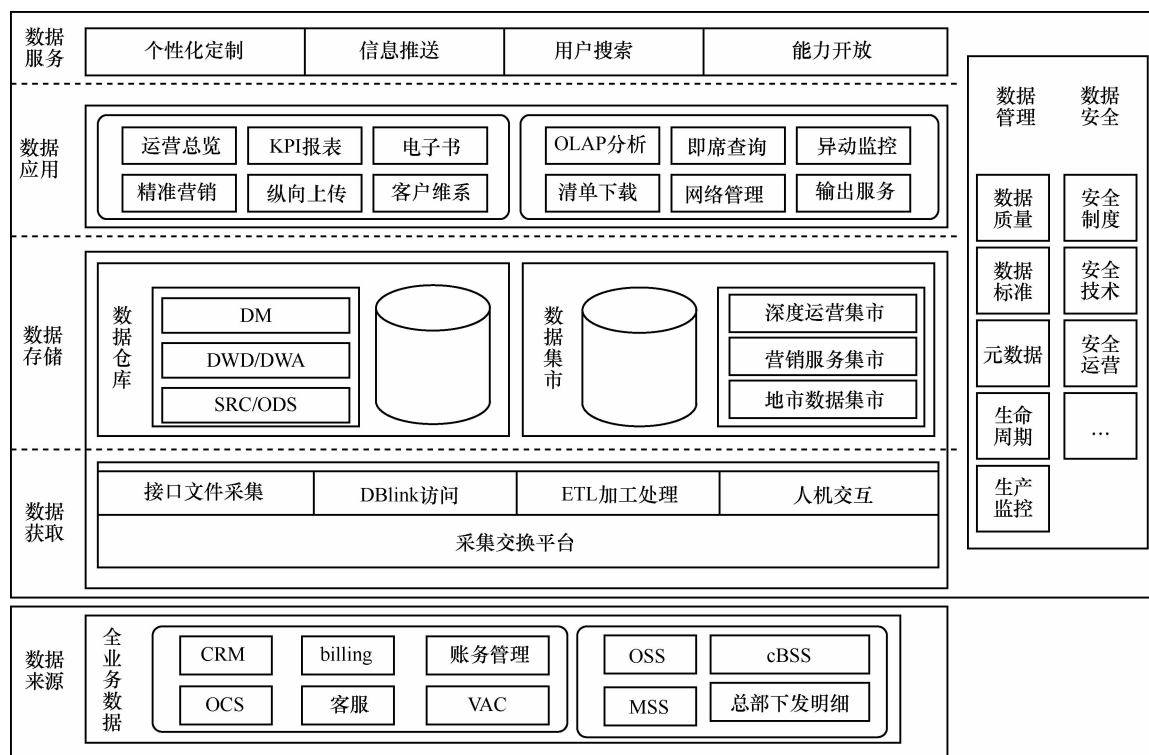


图1 省分数据分析系统逻辑架构

全教育等运营机制确保数据安全。

3.2 系统技术架构

技术架构上采用 Hadoop、MPP、传统关系型数据库混搭模式,扬长避短。Hadoop 对海量数据、非结构化数据的处理能力强,且易扩展;MPP 数据库对数据的处理性能高,可提升数据实时分析能力;传统成熟关系型数据库作为主数据仓库,可确保核心数据服务能力的稳定。数据分析系统技术架构如图 2 所示。

图 2 中,Hadoop 适用于大集群海量离线数据分析,可扩展至 1 000 多个节点,可处理 PB 级海量数据。利用 Hadoop 对海量非结构化数据的数据处理能力,可对网管日志、网元信令、上网痕迹等进行预处理整合。由于 Hadoop 的分布式并行计算和易扩展特性,且对硬件平台的要求相对较低(比如 x86 平台集群),可实现灵活配置,按需、快速分配资源,有效控制成本。

Hadoop 生态圈中,Impala 与 Hive 都是构建在 Hadoop 之上的数据查询工具,各有侧重,Hive 适用于静态批处理查询分析,而 Impala 适用于实时交互式 SQL 查询。可以先用 Hive 数据处理能力,进行数据转换聚合;再用 Impala 在 Hive 处理后的结果集上实施快速数据分析。HBase 是运行于 HDFS 顶层的 NoSQL,具备随即读写功能,是一种列存

储数据库。而 Hive 支持 HSQL,是一种类 SQL 编程接口,可以将 SQL 语句转换为 MapReduce 任务运行,它本身不存储数据。

MPP 是一种海量数据实时分析架构。适合小集群(100 以内)、低并发(50 以内)的场景,可处理 TB 级海量数据。比如 EMC Greenplum、HP Vertica、IBM Netezza、Teradata Aster 等。MPP 数据库存储明细数据、轻度汇总数据等。

MPP 采用不共享架构,每个节点运行自己的操作系统和数据库等,节点之间信息交互通过网络传输实现。与 Hadoop 相比,扩展性、并发性比不上 Hadoop,这可从 CAP 理论中得到解释。根据 CAP 定律(consistency, availability, partition tolerance theorem):在分布式计算机系统中,一致性、可用性和分区容错性这 3 种保证无法同时得到满足,最多满足两个。因为 MPP 始终还是 DB 数据库,一定要考虑一致性,其次考虑可用性,最后才在可能的情况下尽量做好分区容错性。而 Hadoop 是专为并行处理和存储设计的,所有数据都以文件形式存储,它优先考虑的是分区容错性,然后是可用性,最后再考虑一致性。所以 Hadoop 的可扩展性当然优于 MPP。

传统 RDB 数据库(Oracle、DB2 等)用于存储用户标签库、用户 361°全息数据、粗粒度汇总数据、报表数据、多维

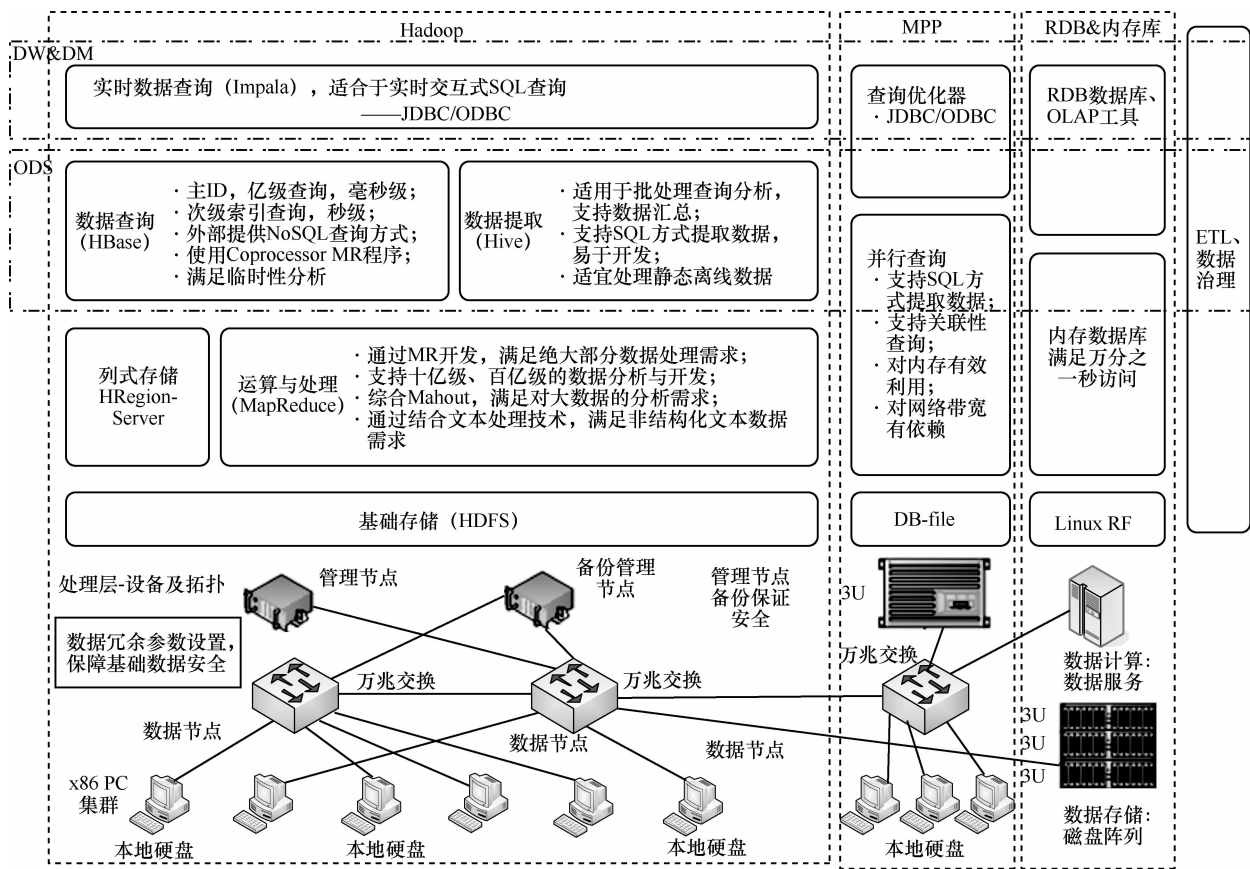


图2 数据分析系统技术架构

数据、指标库等结果数据,确保核心数据服务能力的稳定。

4 数据治理与数据安全

在移动互联网、云计算和大数据的穹顶之下,未来的IT运维管理被赋予了更多的数据治理、数据挖掘和数据安全管理重任,传统IT走向DT。“Garbage in, Garbage out”表示输出质量是由输入质量决定的。面向DT服务的数据分析系统,数据治理管控机制如图3所示。

图3中列出数据生产中业务规范、数据采集加工过程中常规稽核规则配置以及和生产系统互动、与数据需求方互动运维协作机制。

面向DT服务的数据分析系统还必须确保“云、管、端”的数据安全,对涉密敏感信息需要脱敏加密,加强数据输出服务的安全防护和流程管理,从管理、技术和控制3个视角,建立数据资产安全运营规范机制。

- 从管理视角看,包括规章制度与作业标准的制定,岗位角色的设定,管理流程模型的制订等。
- 从控制角度看,涵盖网络层、系统层、应用层和数据

层的安全控制、安全审计等。

- 从技术视角看,需要采用立体交互的技术防护、监控与审计措施,做到安全预警、保护、检测、反应、恢复、反击,以保障系统安全稳定不间断应用。

总体来看,技术、人员、管理构成信息安全的三要素。其中,技术是核心,人员是关键,管理是保障。管理和技术并重,技术和人员结合,充分发挥人的作用,做到人、技术和管理和谐统一,在法律和安全规章的约束下,才能确保企业信息安全的真正有效落地。

5 数据服务的应用场景

5.1 数据服务的迭代优化流程

如何利用运营商拥有的丰富、多维、高价值的“数据资产”,解决自身营销和客户服务精准性,以改善用户体验,提升用户价值与忠诚度;如何对外数据输出和能力开放,服务于产业链各方合作伙伴,服务于社会;这是个渐进的过程,图4表示这种迭代优化流程。

图4中,数据服务的基础是面向DT服务的数据分析

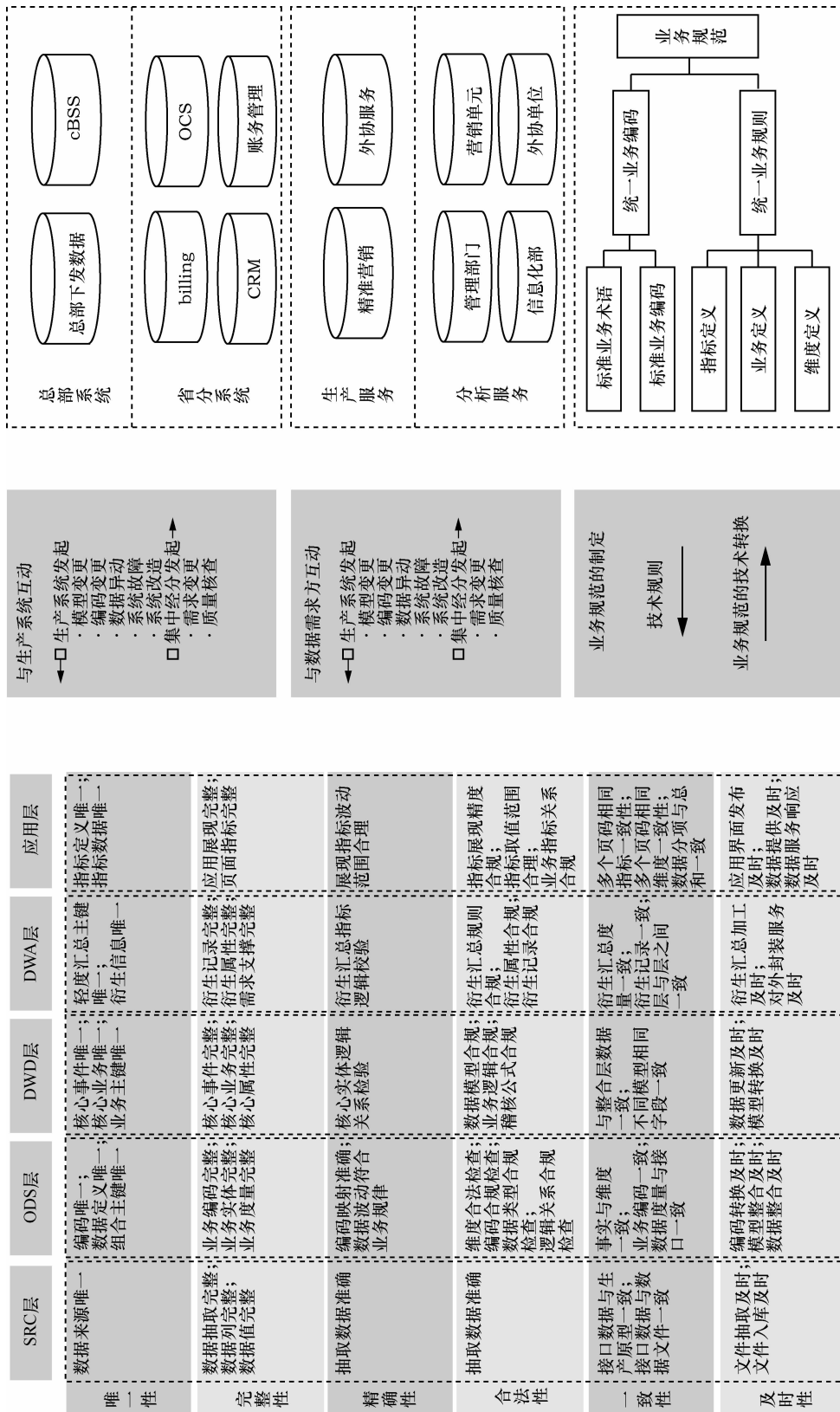


图 3 数据治理管控机制

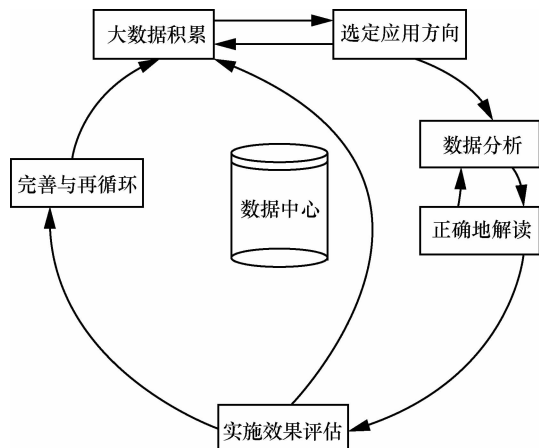


图4 数据服务的迭代优化流程

系统(数据中心),通过大数据积累、选定应用方向、数据分析、正确的解读、实施效果评估、完善与再循环闭环流程,

逐步提升服务效果和能力。其中比较关键的环节包括数据清洗、数据集成、构建模型、算法参数调优等。

5.2 精准营销服务支撑

在数据中心中构建客户 361°全息视图,如图 5 所示。

图 5 中,一些用户衍生属性,是通过 CHAID(卡方自动交叉检验)分类算法、随机森林模型等算法挖掘分析得出,比如忠诚度、流失倾向、兴趣爱好等。有些属性对衍生属性的动态变化起着关键作用,比如交际圈的收缩、主叫异常波动会显著影响忠诚度、流失倾向等评估,图 6 显示了流失预警的关键特征。

客户 361°全息视图,在企业内部精准营销服务实际运用中,取得较好效果,据业务部门营销后评估报告,流量类业务的针对性营销较先前普销型推荐的成功率提升了 7.5 个百分点,客户对营销的满意度提升近 10 个百分点,存量公

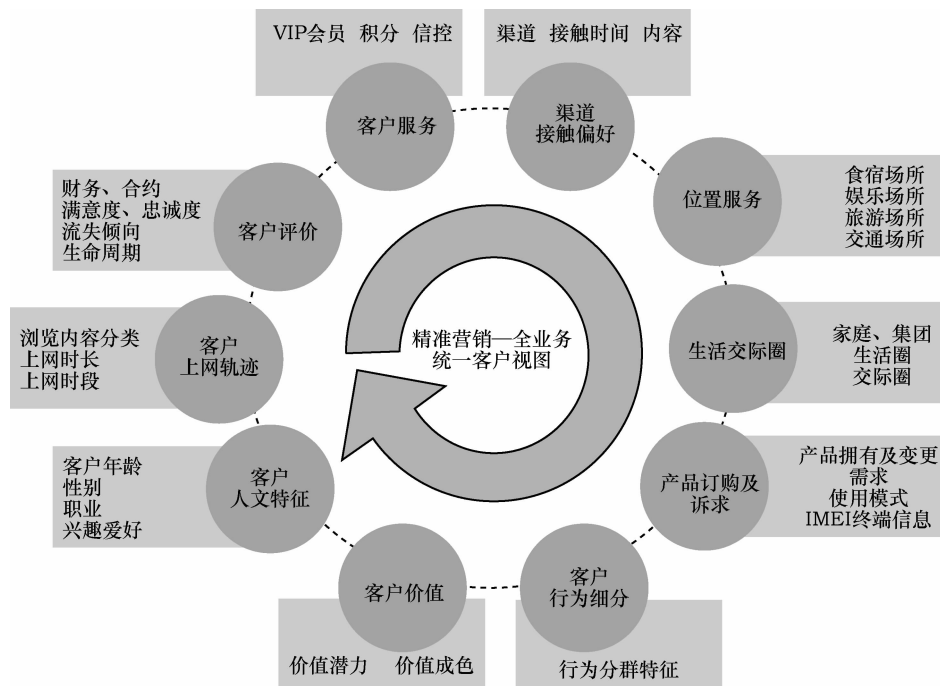


图5 客户 361°全息

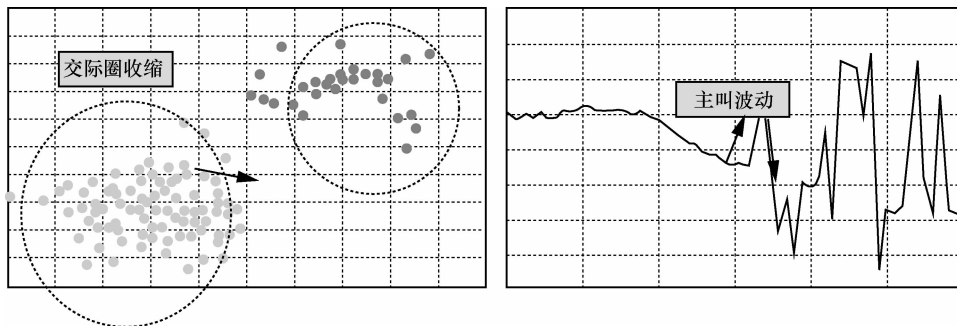


图6 流失预警关键特征示例

众用户的保有率与2014年同比提升1.4个百分点,续约用户的ARPU值也有明显提升。此外,通过用户4G终端选购因素的挖掘分类训练,得出实验结果:训练分区查准率71%,测试分区查准率68%,查全率平均65%,预计经过迭代优化,对精准营销也会发生积极作用。

6 结束语

大数据时代给运营商业创新和涅槃重生提供了新的曙光。数据资产运营是一个迭代优化的过程,构建易扩展、性能高效的灵活系统架构是需求的推动,也是产生价值的源泉。目前建设的统一数据分析系统平台,已取得初步成效和商业价值。平台架构的设计方案、数据治理的协同管理对业界有一定参考价值。从未来的发展看,大数据资产对运营商业贡献会越来越多。与第三方合作,构建大数据生态体系和能力开放已在开展并会持续增强,数据租赁、数据分析、数据分享服务等将是运营商提供通信网络服务外的一大亮点。

参考文献:

- [1] 王珊,王会举,覃雄派,等. 架构大数据:挑战、现状与展望[J]. 计算机学报, 2011, 34(10):1741-1752.
WANG S, WANG H J, TAN X P, et al. Architecting big data: challenges, studies and forecasts [J]. Chinese Journal of Computers, 2011, 34(10):1741-1752.
- [2] “基于大数据的互联网化存量经营”项目组. 运营商存量经营大数据平台及其关键技术研究 [J]. 电信科学, 2014, 30(6): 118-125.
Project Team of Big Data for Inventory Management. Research on big data platform and key technologies for inventory management of carriers [J]. Telecommunications Science, 2014, 30(6):118-125.
- [3] 吴军. 大数据和机器智能对未来社会的影响 [J]. 电信科学, 2015, 31(2):1-10.
WU Jun. Big data, machine intelligence and their impacts to the future world[J]. Telecommunications Science, 2015, 31(2): 1-10.
- [4] 雷蕾,李景文,宫大鹏,等. 基于Hadoop的OSS域数据建模与采集方法研究[J]. 电信科学, 2015, 31(1):128-138.
LEI L, LI J W, GONG D P, et al. Study on data modeling and collection in OSS based on Hadoop [J]. Telecommunications Science, 2015, 31(1):128-138.
- [5] 孟小峰,慈祥. 大数据管理:概念、技术与挑战[J]. 计算机研究与发展, 2013, 50(1):146-169.
- MENG X F, CI X. Big data management: concepts, techniques and challenges [J]. Journal of Computer Research and Development, 2013, 50(1):146-169.
- [6] 张海峰,董昭. 面向大数据的经营分析系统建设思路分析[J]. 互联网天地, 2015(2).
ZHANG H F, DONG Z. Analysis of construction ideas of business analysis system for big data[J]. China Internet, 2015(2).
- [7] 陈可寒,韩盼盼,吴健. 基于用户聚类的异构社交网络推荐算法[J]. 计算机学报, 2013, 36(2):349-359.
CHEN K H, HAN P P, WU J. User clustering based social network recommendation [J]. Chinese Journal of Computers, 2013, 36(2):349-359.

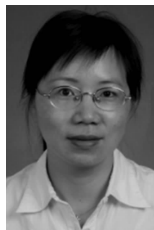
[作者简介]



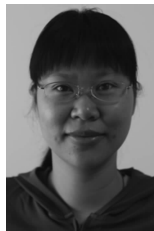
汪保友(1968-),男,博士,中国联合网络通信集团有限公司上海市分公司高级工程师,主要研究方向为数据分析与数据挖掘、数据科学、数字签名、信息安全。



吴琮(1975-),女,中国联合网络通信集团有限公司上海市分公司信息化事业部总经理,主要研究方向为电信业务支撑系统的规划设计、移动互联网、云计算。



钱晶(1970-),女,中国联合网络通信集团有限公司上海市分公司工程师,主要研究方向为数据科学、移动互联网、通信网络规划。



陶晓英(1978-),女,中国联合网络通信集团有限公司上海市分公司工程师,主要研究方向为数据科学、电信业务支撑系统的规划设计、移动互联网。