运营技术广角

基于大数据的移动用户行为分析系统与应用案例

谷红勋,杨珂
(中国电信股份有限公司河南分公司,河南 郑州 450016)

摘 要:本系统基于 Hadoop 架构采集运营商网络侧产生的大数据,并对数据进行深度加工,挖掘其中相关的用户行为属性,构建用户行为分析模型,并对具体用户行为应用进行有效探索。针对技术选型、系统实现、数据采集、模型设计与应用案例,完整展示了移动用户行为分析系统的设计思路与实现。

关键词:Hadoop;ETL;数据模型;用户行为分析
中图分类号: TN91 文献标识码:A
doi: 10.11959/j.issn.1000-0801.2016039

Mobile user behavior analysis system and applications based on big data

GU Hongxun, YANG Ke
Henan Branch of China Telecom Co., Ltd., Zhengzhou 450016, China

Abstract: Based on Hadoop's architecture, this system collects and analyzes the telecom operator network's data to build up user behavior model for effective exploration of big data applications. The whole process was discussed, including data collection, system design, implementation and application cases.

Key words: Hadoop, ETL, data model, user behavior analysis

1 引言

随着智能终端、云计算、物联网与 4G 网络的普及,电信网络系统产生了海量数据。与传统数据相比,电信运营商的数据具有数据量大、数据多样性、增长快速、价值密度低等特点。传统数据挖掘工具(如 Oracle、SPSS、SAS 等)并不具备大数据挖掘能力,所有的数据必须在单一的服务器上处理,硬件能力成为大数据应用的瓶颈。对电信运营商而言,必须寻找新一代的数据处理技术,以实现大数据的分析与挖掘。

同时,以往对大数据的探索主要集中在技术层面,实际应用案例较少,本文针对运营商网络数据进行深度加工,对原本只用于计费的通话详单进行深度加工,挖掘其中的用户行为属性,构建用户行为分析模型,并成功应用于养卡用户监控等具体业务。

目前在大数据应用方面的研究具体介绍如下。

- 参考文献[1]提出从数据挖掘的角度,分析大数据的数据建模与传统的数据建模之间的差异,并提出基于大数据设计数据模型的具体思路,包含数据来源、数据挖掘和分析、用户兴趣建模与安全隐私等。
- 参考文献[2]提出电信运营商如何采集、利用移动数据的相关议题。移动数据不只是用来理解用户的过去和现在,也可以预测用户未来的行为、活动和状态。

收稿日期:2015-07-28;修回日期:2015-12-15

· 参考文献[3]针对标准的移动电话记录，建立一套全新的用户行为分析指标，能够精确预测用户的个性与行为，为移动用户行为分析模型提供参考依据。

2 主流建设方案

传统数据挖掘工具（如 Oracle、SPSS、SAS 等）并不具备大数据挖掘能力，同时要求所有的数据必须在单一的服务器上处理，硬件能力成为大数据应用的瓶颈。随着数据量的大量增加，产生了新的数据存储和处理能力问题，传统数据仓库无法支撑线性扩容，造成管理难度加大、成本高、扩容压力大、效率下降等问题。电信运营商需要探索大数据系统的建设方案，解决上述问题。目前主流的大数据系统建设方案如下。

· 传统数据库升级解决方案：由高性能的主机与大容量存储组成，通常为“UNIX 服务器+存储磁盘阵列+数据仓库软件”的开放式解决方案。

· 一体机解决方案：基于一体机的 BI 集成化解决方案，一体机包含数据仓库服务器、数据仓库存储、数据仓库软件等。

· 基于 x86 开放平台的海量数据解决方案：在开源 Hadoop 技术的基础上开发的海量数据处理软件，基于 x86 服务器的大规模并行处理解决方案。

随着集约化运营管理思路的提出，系统处理的数据量会越来越大，传统的小机数据库模式难以支撑海量数据处理的要求，而一体化产品（硬件+数据库软件捆绑销售）投资高、性价比低。总结主要厂商解决方案，几种技术方案特点比较见表 1。

除了成本因素外，本系统需要处理结构化、文件型和非结构化数据，还需要考虑数据结构问题，具体因素如下：

· 对于海量的结构化数据处理，如何保障系统的稳定性和高性能；

· 对于文件型和非结构化数据处理，先以分布式集群系统平台进行预处理，形成结构化数据后交由 MPP 或关系型数据库进行处理。

综合考虑技术成熟度、性价比和数据处理需求，采用基于 Hadoop 的分布式集群系统的平台架构。该技术架构具备下列优点。

· 高性能：采用分布式存储、并行计算技术，充分利用设备性能，提升数据处理速度。

· 高可靠性：多任务并行计算、数据冗余存储，有效避免设备单点故障，提供高可靠服务。

· 高扩展性：x86 架构可以通过增加节点，完美支持计算和存储能力的线性扩容。

· 高性价比：利用低成本的基于 x86 的主机设备，有效降低一次性投入成本，更能支持小成本的平滑升级与扩容。

3 关键技术

3.1 Hadoop 开源软件

Hadoop 是对大量数据进行分布式处理的软件框架。Hadoop 系统以可靠、高效、可伸缩的方式进行大数据处理，以并行的方式工作，通过并行处理加快处理速度，具有以下几个优点。

· 高可靠性：假设计算元素和存储可以出错，可维护多个工作数据副本，确保能够针对失败的节点重新进行分布处理。

· 高扩展性：能够在可用的计算机集簇间分配数据并完成计算任务，这些集簇可以方便地扩展到数以千计的节点中。

· 高效性：能够在节点之间动态地移动数据，并保证各个节点的动态平衡。

· 高容错性：能够自动保存数据的多个副本，并能够自动重新分配失败的任务。

3.2 多租户管理技术

由于本系统面向多个用户提供多种服务，各类型用户通过访问本系统获取自己的数据，必须保障这些数据不被

表 1 技术方案特点比较

方案	特性	大规模部署成本	备注
传统商业数据库方案（如 Oracle+小型机、DB2+小型机等）	计算、查询效率一般；运行稳定性高，适合于 OLTP 系统；结构化数据处理	中	技术成熟、海量数据存在处理瓶颈
MPP 数据库（一体机，如 Teradata、Oracle Exadata 等）	大数据量处理 TB、PB 级；并发查询、计算能力高；线性扩展能力强；适合 OLAP 系统；结构化数据处理	较高	技术成熟、要注意产品跨代兼容问题
分布式数据库（如 x86 开放平台、Hadoop 等）	大数据量处理 PB 级；并发查询、计算能力高；扩展能力非常强；适合 OLAP 系统；支持结构化与非结构化数据处理	较低	新技术，集成实施要求高，需要在相关技术上深度封装

其他用户随意访问或篡改。因此如何实现多租户安全，保证多用户间隔离、数据安全和防止有害代码的威胁，是本系统必须解决的问题。

本系统采用多租户管理技术，对数据库结构进行特殊的设计，在安全和隔离性方面也要有所保障，实现如下功能。

- 资源隔离：控制高资源消耗任务，通过容量/公平调度器，控制资源分配以保证重要工作的资源。
- 数据隔离：用户数据保存在用户专有的目录中，其他未被授权的用户不能访问。
- 安全隔离：保证不同用户和组的安全，保证对集群的所有操作都是经过授权认证的。

3.3 基于 shared hardware 的架构设计

为了实现对“多租户”的支持，需要配套相应的多租户架构（multi-tenancy architecture），本系统基于硬件共享（shared hardware）架构，为多租户提供一个应用容器集群环境，应用运行在应用容器中，实现资源与数据的安全隔离。

4 大数据用户行为分析系统设计

4.1 需采集处理的数据类型

依据业务需求，必须支持 TB 级数据采集，主要采集的数据类型如下。

- 企业经营数据：BSS 中的计费详单、用户、客户、套餐、服务、渠道等数据。
- 企业运营数据：OSS 中的资源、服务开通等数据。
- 企业管理数据：MSS 中的人力、财务等数据。
- 移动 DPI 数据：访问移动互联网的行为数据，包括用户手机号、访问 URL、应用、访问时间等信息。
- 移动 AAA 系统数据：用户信息及行为信息，包括用户手机号、IP 地址、认证时间、基站位置等信息。
- 固网 DPI 数据：访问互联网的行为数据，包括用户 IP 地址、访问 URL、访问时间、用户 UA、cookie 等信息。
- 固网 AAA 系统数据：用户互联网访问的 IP 地址和 AD 账号的对应关系。
- 位置信令数据：用户的地理位置信息。
- 业务平台数据：能力类、产品类、支撑类平台的用户增值业务、基地业务、行业应用等数据。

具体数据采集类型与数量见表 2。

4.2 系统架构设计

系统架构分为存储层、服务层、处理层和管理层，主要功能如图 1 所示。

（1）存储层

支持异构的存储设备，通过存储虚拟化技术，将存储设备统一到资源池中，通过部署分布式文件系统，对上层提供统一的存储服务。系统同时支持低成本的本地磁盘方案。

（2）服务层

服务层为 ETL 平台提供必需的底层服务。其中流程

表 2　具体数据采集类型与数量

数据源信息	每天数据量	采集平台
客户信息，业务订购，产品变更，产品、销售品、营销活动的发布、查询、变更等数据	10 GB	CRM
套餐、余额、账单、清单等数据	200 GB	计费系统
各种报表信息	4 GB	ODS
经分指标、市场分析、产品分析、报表	2 GB	EDW
合作伙伴信息	3.3 GB	电子渠道
使用易信、QQ、语音等方式的交互的客户报障、投诉等信息	300 GB	客服系统
增值业务订购、计费等数据	667 MB	ISMP
点到点短信、SP 到点和点到 SP 短信的详单，包含主被叫方号码、类型、计费方、信息量等	10 GB	短信中心
点到点彩信、SP 到点和点到 SP 彩信的详单，包含主被叫方号码、类型、计费方、信息量等	6 GB	彩信平台
全网用户（包括移动、宽带、固话等）登录的认证方式、登录时间、登录 IP 地址等信息	50 GB	UDB 平台
告警信息、各专业网管数据信息	60 GB	智能网管
用户访问移动互联网的行为数据，包括用户手机号、访问 URL、应用等信息	1.4 TB	移动 DPI
宽带用户访问互联网的行为数据，包括用户 IP 地址、访问 URL、访问时间、用户 UA 等信息	10 Gbit/s	固网 DPI
用户信息及行为信息，包括用户手机号、IP 地址、认证时间、基站位置等信息	10 GB	移动 AAA 系统
实时获得用户互联网访问的 IP 和宽带账号的对应关系	30 GB	固网 AAA 系统
用户的位置信息数据	100 GB	信令监测平台
用户上网日志信息	2 TB	移动上网日志留存系统
基地产品业务数据	1.7 GB	集团基地业务平台

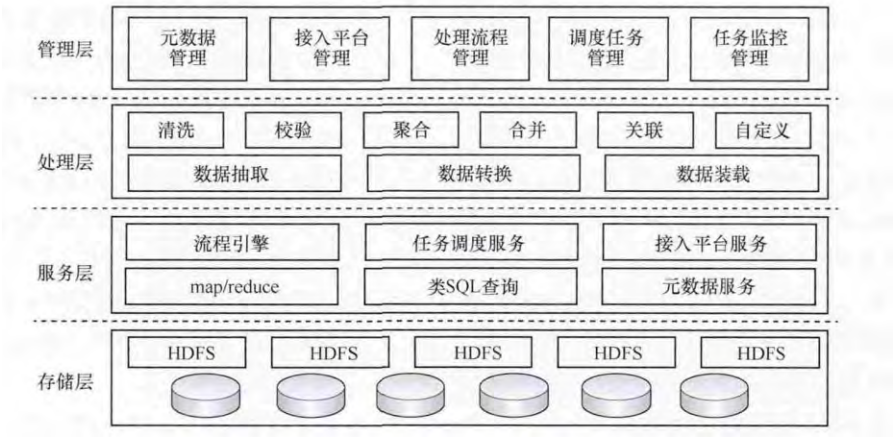

图 1 功能架构

引擎与任务调度服务,以定时或者触发方式执行预先配置的 ETL 任务,支持复杂流程的串并联;元数据与接入平台服务,提供不同种类、异构数据源的数据抽取能力;map/reduce 与类 SQL 查询提供并行计算与简便的分析功能。

(3)处理层

处理层是数据分析平台的核心功能,分为数据抽取、数据转换与数据装载 3 个过程,常见的 ETL 动作包括数据清洗、数据校验、聚合、关联等,支持自定义的数据处理动作。

(4)管理层

平台提供可视化、流程化的管理操作界面,便于业务人员使用。管理功能包括元数据管理、营销活动管理、目标客户管理管理与系统监控管理等。

4.3 ETL 功能设计

由于数据大,本系统对 ETL 处理能力提出了更高的要求:需要集中支撑大量的数据采集任务调度;需要集中支撑大量的数据处理任务调度。本系统采用分布式 ETL 调度框架进行任务调度,可以解决如下问题:

- 支持部署多个调度节点,解决调度节点单点故障问题,在任意一个调度节点挂死后都不会影响调度任务的调度与执行;
- 调度节点可扩展,可以根据具体需求动态扩展调度节点数,提高处理性能;
- 调度节点均衡负载,可以在多个调度节点中实现均衡负载,避免资源压力集中在某个节点上;
- 调度先进先出原则,需要保证工单执行的时序性。

4.4 系统服务器需求评估

本系统每月采集数据大约为 59 TB。系统需要的服务器计算过程见表 3,计算结果共需要 18 台服务器。

4.5 系统拓扑结构

本系统采用吉比特网络接入 Hadoop 平台,各个节点均配置 4 端口吉比特,分别接入两台相互冗余的接入交换机,并采用网卡聚合方式接入,以保障网络接入的安全稳定性。对于多台应用服务器的负载均衡访问,均由 DCN 接入层部署的负载均衡器提供。系统拓扑结构如图2所示。

5 用户行为分析模型设计与应用

5.1 用户行为分析模型设计思路

本系统对原本只用于计费使用的通信、上网数据进行深度加工,挖掘其中的用户行为属性,如规律性(regularity)、多元性(diversity)、空间行为(spatial behavior)、活动行为(active behavior)、使用行为(basic phone use)、关联性(correlation)6 类,并与这些关键指标构建用户行为模式。

(1)规律性

- 平均通话间隔(average inter-call time):计算用户通话(包括主被叫)间隔的平均值,单位为 s。从上一通电话开始,到下一通电话开始记为一次间隔。
- 平均短信间隔(average inter-text time):计算用户收发短信间隔的平均值,单位为 s,取样为每两条短信之间的时间间隔。
- 平均上网间隔(average inter-internet time):计算用户上网间隔的平均值,单位为 s,取样为每两次上网之间的时间间隔,上网行为包括通过 2G、3G、Wi-Fi 上网。

表 3　系统需要的服务器计算过程

序号	参数名称	参数值	单位	备注
1	移动数据中间层	9	TB	
2	CCG 数据中间层	18	TB	
3	ODS 数据中间层	1.8	TB	
4	其他数据中间层	0	TB	
5	中间层数据小计	28.8	TB	各项数据保存周期内，中间层数据合计
6	Hadoop 汇总数据保存周期	1 095	天	汇总数据，永久保存，暂时按 3 年评估
7	Hadoop 汇总数据比例	5‰		按照日原始数据量的 5‰评估
8	Hadoop 汇总数据小计	1.149 75	TB	
10	合计	31.42	TB	原始数据+中间层数据+汇总数据
		Hadoop 所需存储合计		
11	副本策略	2		2 副本
12	数据压缩比	2∶1		不同数据存储采用不同的压缩比，2∶1 为综合压缩比（经验值）
13	冗余系数	80%		系统负荷不能超过总容量的 80%
14	Hadoop 平台所需存储容量	58.91	TB	数据量合计×（1+副本策略）/压缩比/冗余系数
		Hadoop 数据处理存储估算		
15	每节点实际存储容量(Hadoop)	10	TB	2CPU，64 GB 内存，12×1 TB，2 块盘做 RAID1，安装操作系统，10 块盘不做 RAID，可用 10 块盘
16	满足存储估算的 data node 数量	6	台	向上取整
		所需 x86 服务器数量计算结果		
17	所需 name node 数量	2	台	同 data node，内存不小于 128 GB
18	所需 data node 数量	16	台	取性能估算或者存储估算的最大值
19	合计	18	台	

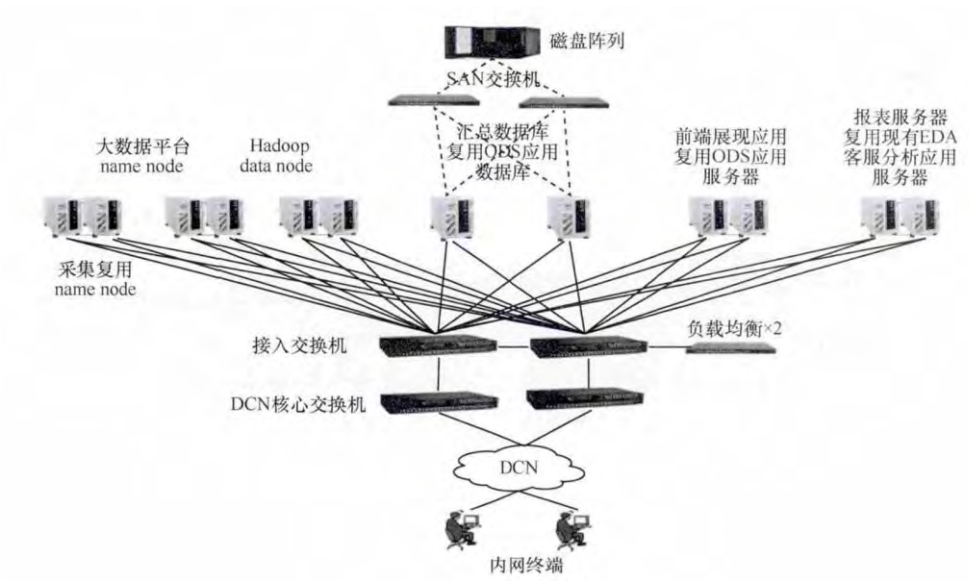

图 2　系统拓扑结构

· 通话间隔方差（variance of inter-call time）：用户两次通话之间间隔时间的方差，单位为 s^2，表示用户每通电话间隔同平均通话间隔的偏离程度。

· 短信间隔方差（variance of inter-text time）：用户两次短信之间间隔时间的方差，单位为 s^2，表示用户每个短信间隔同平均短信间隔的偏离程度。

· 上网间隔方差(variance of inter-internet time):用户两次上网之间间隔时间的方差,单位为 s^2,表示用户每个上网间隔同平均上网间隔的偏离程度。

AR 系数(AR coefficient)为每个用户建立 AR 模型,如时间序列 X_t 包括用户周一早上 6 点到晚上 12 点,周二早上 6 点到晚上 12 点,…的通话数,模型如下:

$$X_t=c+\sum_{i=1}^{p}\varphi X_{t-i}+\varepsilon_t \tag{1}$$

AR 系数 φ 表示知道前 6 h 打了多少电话,预测未来 6 h 的通话次数。

(2)多元性

通话熵(entropy of call):表示用户同其他用户通话的信息量,用户通话联系人越多,通话熵越大。用户 A 同用户B 间通话熵的计算式为:

$$H_{1,A\to B}=-\sum_{B}f_{1,B}\ln f_{1,B} \tag{2}$$

其中,$f_{1,B}$ 为 A 同 B 通话的频率。

短信熵(entropy of text):表示用户同其他用户发短信的信息量,用户短信联系人越多,通话熵越大。用户 A 同 B 间短信熵的计算式为:

$$H_{2,A\to B}=-\sum_{B}f_{2,B}\ln f_{2,B} \tag{3}$$

其中,$f_{2,B}$ 为 A 同 B 发短信的频率。

上网熵 (entropy of internet):表示用户上网的信息量,用户上网次数越多,上网熵越大。用户 A 上网熵的计算式为:

$$H_{3,A}=-\sum f_3\ln f_3 \tag{4}$$

其中,f_3 为 A 上网的频率。

联系人通话比(contact to call ratio):表示用户联系人中有多少通过通话联系。联系人通话比为联系人数同通话联系人之比。

联系人短信比(contact to text ratio):表示用户联系人中有多少通过短信联系。联系人短信比为联系人数同短信联系人之比。

通话联系人数(number of call contact):通过通话的联系人数。

短信联系人数(number of text contact):通过短信的联系人数。

(3)空间行为

· 旋回半径(radius of gyration):包括用户所有位置的圆的最小半径,位置为用户停留大于 15 min 的基站。
· 旅行距离(distance traveled):为用户在一段时间内到访位置的连续距离。
· 地点数(number of place):用户停留地点总数。
· 地点熵(entropy of place):表示用户在某地点通话、发短信、上网的信息量,用户停留的地点越多,地点熵越大。用户 A 的地点熵计算式为:

$$H_{4,A@Z}=-\sum_{Z}f_{4,Z}\ln f_{4,Z} \tag{5}$$

其中,$f_{4,Z}$ 为 A 在 Z 地使用手机的频率。

(4)活动行为

· 通话回复率(call response rate):表示用户回复通话的比率,回复通话为用户甲同用户乙通话后 1 h 内用户乙回复用户甲的通话。通话回复率为回复通话次数占通话总次数的百分数。
· 短信回复率(text response rate):表示用户回复短信的比率,回复短信为用户甲发给用户乙短信后 1 h 内用户乙回复用户甲的短信。短信回复率为回复短信次数占总短信的百分数。
· 发起通话率(percent of call initiated):表示某用户同其他用户通话时有多少次为该用户主叫。发起通话率等于用户主叫通话的次数与通话总次数的比率。

(5)使用行为

· 通话次数(number of call):用户通话的次数。
· 短信数(number of text):用户发短信的次数。
· 上网次数(number of internet):用户上网的次数。
· 上网流量(flow of internet):用户上网的总流量,包括 Wi-Fi、2G、3G 上网。
· 互动次数(number of interaction):用户间互动行为的次数。互动行为包括通话和短信,1 h 内互动行为的往复记为互动。

(6)关联性

· 机卡比值(cellphone-card ratio):表示同一手机号对应终端数量的比值,比值越大,说明该号码曾被多个手机终端使用。计算时使用终端串码(IMEI)关联手机号。
· 卡机比值(card-cellphone ratio):表示某一终端使用过手机号的数量,比值越大,说明该终端使用过的手机卡越多。

· 销售员贝叶斯因子(retailer Bayesian factor):表示销售员拥有养卡前科的先验概率。贝叶斯推断中,在事件 1 发生的条件下事件 2 发生的概率,即后验概率,可由先验概率与调整因子得到。如销售员拥有养卡前科,则再产生养卡行为的概率会大。

5.2　用户行为模型应用案例

移动用户行为分析系统 2014 年 12 月开发完成,2015年开始测试使用,已针对养卡用户监控等开展具体应用。养卡用户是指渠道商为了获取号码卡销售后得到的酬金,私自激活并伪装号码正在使用的状态,以期获取运营商酬金。养卡用户属于无效用户,造成电信运营商大量营销资源与佣金的浪费。

基于用户行为分析系统针对养卡用户的行为进行深度挖掘,养卡用户行为与正常用户行为比对特征见表 4。

(1)模型分析期

指用户入网行为、通信行为产生时间段,即模型输入变量的时间窗口(分析期为 2014 年 11 月、2014 年 12 月)。

(2)模型应用期

指异常用户名单输出时间,即应用模型异常名单,开展管控工作的时间窗口(管控期为 2015 年 1–8 月)。

(3)用户行为特征

用户满足低活跃度,则认为用户满足必备条件。

· 低活跃度:月主叫时长小于或等于 3 min、月被叫次数小于或等于 3 次、月短信小于或等于 3 条、月流量小于或等于 3 MB,满足其中任意 3 项则认为其低活跃度。

· 串码集中:5 个以上号码使用同一终端注册(取最后一次使用终端)。

· 用户信息:10 个以上用户使用相同身份证开户。

· 联系号码集中:月(主动+被动)联系号码数小于或等于 3 个,某网点当月发展用户中 10 个及以上用户拨打同一个号码(不含 10000、11888 等客服号码)超过 3 次,上述两项中满足任意一项即判定符合此条件。

该系统上线后,实现对养卡网用户的精确判别,同时在市场部门的配合下,开展“养卡专项清理”活动,实时监控入网渠道,建立追溯机制,对发展渠道进行追责。模型应用后,养卡用户数量得到有效控制,由 2015 年 1月的 259 312 人,下降到 8 月的 145 219 人,模型效果显著,如图 3 所示。

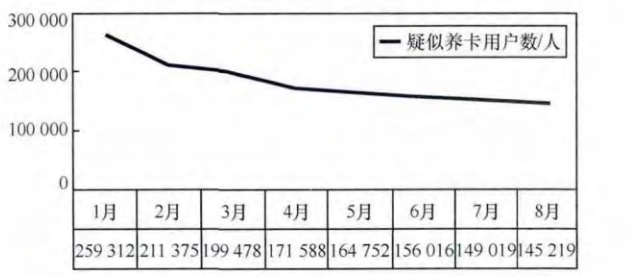

图 3　模型效果

6　结束语

本系统采用 Hadoop 架构采集运营商网络侧数据,对大数据技术选型、ETL 过程、数据吞吐量、平台实现方案等进行探索。对原先只用于计费的详单数据,进行深度加工,分析其中的户行为属性,并对养卡用户监控场景进行有效的实践。

对于运营商而言,大数据包括 3 个层面的含义:第一个层面是“大数据”资产,囊括高形态复杂度的超大规模数据;第二个层面是“大数据”平台,实现全新的、强大的数据处理机制;第三个层面是“大数据”运营,带来创新的业务机会与商业模式。在成功实现第一、二层面的业务探索与系统建设后,本系统已初步具备第三层面的大数据运营能力,并已成功应用在养卡用户识别等营销活动中。未来将结合用户上网数据、用户位置数据等,进一步扩大指标体系的范围与有效性,争取在 4G 发展、终端升级、流量经营、存量经营、流失预警等方面,建立大数据驱动的经营新模

表 4　养卡用户行为与正常用户行为比对特征

特征	实际养卡用户	暂停使用手机(如出国)	流失用户
规律性	几乎无通话/短信/上网行为。平均通话/短信/上网间隔长	有平均通话/短信/上网间隔,通话/短信/上网间隔方差较大,用户停用手机呈跳跃性	平均通话/短信/上网间隔较小,用户手机使用呈平滑曲线减少
多元性	通话多元性单一	通话多元性较单一	通话多元性正常
空间行为	空间范围小	空间范围极大	空间范围正常
活动行为	活动行为单一,互动联系人少	活动行为几乎没有	活动行为较少
使用行为	用户手机使用少,各指标近乎零	用户手机使用少,各指标近乎零	用户手机使用减少,各指标低于正常水平
关联性	关联性复杂	关联性唯一	关联性唯一

式,将数据变为生产力。

参考文献:

[1] WU X D, ZHU X Q, WU G Q, et al. Data mining with big data[J]. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(1): 97-102.

[2] MUSOLESI M. Big mobile data mining: good or evil[J]. IEEE Internet Computing, 2014, 18(1): 7-10.

[3] MONTJOYE Y A D, QUOIDBACH J, ROBIC F, et al. Social computing, behavioral-cultural modeling and prediction [M]. Berlin: Springer Heidelberg, 2013.

[4] MONTJOYE Y A D, HIDALGO C A, VERLEYSEN M, et al. Unique in the crowd: the privacy bounds of human mobility[J]. Open Access Publications from Université Catholique De Louvain, 2013, 3(6): 776.

[5] OLIVEIRA R D, KARATZOGLOU A, CONCEJERO C P, et al. Towards a psychographic user model from mobile phone usage [C]//CHI'11 Extended Abstracts on Human Factors in Computing Systems, May 7-12, 2011, Vancouver, BC.[S.l.:s.n.], c2011.

[6] 李文莲, 夏健明. 基于"大数据"的商业模式创新[J]. 中国工业经济, 2013(5): 83-95.
LI W L, XIA J M. Business model innovation based on "big data"[J]. China Industrial Economy, 2013(5): 83–95.

[7] 赵春雷. "大数据"时代的计算机信息处理技术[J]. 世界科学, 2012(2): 30-31.
ZHAO C L. Computer information processing technology in the era of big data[J]. World Science, 2012(2): 30-31.

[8] AGRAWAL D, BERNSTEIN P, BERTINO E, et al. Challenges and opportunities with big data[EB/OL]. (2011-10-29)[2015-07-28]. http://www.docin.com/p-633891531.html .

[9] 王秀丽. 数据挖掘功能特性及其应用流程分析[J]. 科技资讯, 2005(5): 151-152.
WANG X L. Functional characteristics and application of data mining [J]. Pioneering Withence & Technology Monthly, 2005 (5): 151-152.

[10] 王永生. 大数据时代的商业模式创新研究[J]. 南京财经大学学报, 2013(6): 47-51.
WANG Y S. Research on business model innovation in the era of big data [J]. Journal of Nanjing University of Finance and Economics, 2013(6): 47-51.

[11] 李璐. 实时分析迎战大数据[J]. 通信世界, 2012(29).
LI L. The challenge of the real-time analysis for large data[J]. Communications World, 2012(29).

[12] 陈晓霞, 徐国虎. 大数据业务的商业模式探讨 [J]. 电子商务, 2013(6): 16-17.
CHEN X X, XU G H. The study of the big data's business model[J]. E-commerce, 2013(6): 16-17.

[13] 汪维佳. 数量型数据关联规则挖掘及其在通信行业用户分析中的应用[J]. 统计科学与实践, 2005(3): 28-30.
WANG W J. Association rule of quantitative data and its application for communication industry[J]. Statistical Theory and Practice, 2005(3): 28-30.

[14] 徐光宪, 刘建辉, 黄素芬. 电信行业中数据挖掘的应用研究[J]. 现代管理科学, 2004(12): 8-9.
XU G X, LIU J H, HUANG S F. The application of data mining in telecom industry[J]. Modern Management Science, 2004 (12): 8-9.

[15] 郭明, 郑惠莉. 用数据挖掘法分析电信客户流失 [J]. 现代通信, 2005(3): 7-9.
ZHENG H L, GUO M. Analysis of telecom customer churn by data mining[J]. Communication Today, 2005(3): 7-9

[作者简介]

谷红勋(1972–),男,中国电信股份有限公司河南分公司副总经理、高级工程师,主要从事市场营销、企业信息化、互联网增值等工作。

杨珂(1972–),女,中国电信股份有限公司河南分公司移动互联网业务部主任、高级工程师,主要从事互联网增值业务工作。