运营商大数据AI能力层的构建方案探索

Exploration of the Construction Scheme of Operators´ Big Data AI Capability Layer

王　瑜，邓　程，蒋　涛，汪剑桥（中国联通江苏分公司，江苏 南京 210000）
Wang Yu，Deng Cheng，Jiang Tao，Wang Jianqiao（China Unicom Jiangsu Branch，Nanjing 210000，China）

摘　要：
电信运营商积累了大量的优质数据，介绍了构建大数据AI能力层的一种方案，大数据AI能力层能够帮助运营商更好地利用流量红利和数据红利。对外通过AI有效提升运营商客户服务水平与市场营销效果，同时拓宽运营商的服务类型和业务范围；对内使用AI推进网路虚拟化和云技术，提高自动化水平，降低资本和运营支出。

关键词：
人工智能；能力整合平台；大数据；数据挖掘
doi：10.12045/j.issn.1007-3043.2018.12.011
中图分类号：TP181
文献标识码：A
文章编号：1007-3043（2018）12-0051-06

Abstract：
Telecom operators have accumulated a large amount of high-quality data. It introduces a scheme of constructing big data AI capability layer. Big data AI capability layer can help operators make better use of traffic dividend and data dividend. Through AI, the service level and marketing effect of the operators can be effectively improved, and the service types and business scope of the operators can be broadened.

Keywords：
AI；Capability integration platform；Big data；Data mining

引用格式：王瑜，邓程，蒋涛，等. 运营商大数据AI能力层的构建方案探索[J]. 邮电设计技术，2018（12）：51-56.

0 前言

经过多年的高速发展，电信运营商目前已经积累了大量的数据，其中包括行业综合数据、用户使用交互信息、用户消费数据、设备日志记录等结构化数据，与文本、音视频、图片等非结构化数据。AI人工智能经过数十年的发展，很多算法已经非常成熟稳定，能够广泛应用到生产、生活的各个方面。2016年Alpha-Go事件以后，AI受到了全世界的瞩目，以谷歌、Face-Book、微软、阿里巴巴、百度等为代表的互联网企业在近几年也利用AI在各个行业积极布局。

目前我国电信产业已无法从人口红利模式中继续获取高速发展，转而逐渐重视流量红利和数据红利。运营商走在信息网络的最前沿，能获取用户最真实、最核心的数据，同时一直为用户提供全面的ICT服务。对外，AI的使用能够有效提升运营商客户服务水平与市场营销效果，同时能够拓宽运营商的服务类型和业务范围；对内，AI能够帮助运营商推进网路虚拟化和云技术，达到提高自动化水平，降低资本和运营支出的效果。

收稿日期：2018-11-23

1 AI能力层相关技术

1.1 人工智能

人工智能（AI），也称机器智能、智能模拟等，它是计算机科学、控制论、信息论、神经生理学、心理学、语言学等多种学科互相渗透而发展起来的一门综合性学科。人工智能是用来研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的科学。目前已在知识处理、模式识别、自然语言处理、博弈、自动定理证明、自动程序设计、专家系统、知识库、智能机器人

等多个领域取得举世瞩目的成果,并形成了多元化的发展方向。

1.2 Python

Python是一门解释性的、面向对象的、动态语义特征的高层语言。它的高层次的内建数据结构,以及动态类型和动态绑定,使得它非常适合于快速应用开发。Python的简单而易于阅读的语法强调了可读性,因此降低了程序维护的费用。Python支持模块和包,并鼓励程序模块化和代码重用。Python的解释器和标准扩展库的源码和二进制格式在各个主要平台上都可以免费得到、免费分发。

1.3 Docker技术

Docker是一个开源的容器引擎,可以方便地对容器进行管理。其对镜像的打包封装,以及引入的DockerRegistry对镜像的统一管理,构建了方便快捷的"Build,Ship and Run"流程,它可以统一整个开发、测试和部署的环境和流程,极大地减少运维成本。另外,得益于容器技术带来的轻量级虚拟化,以及Docker在分层镜像应用上的创新,Docker在磁盘占用、性能和效率方面相较于传统的虚拟化都有非常明显的提高。因为Docker是基于容器技术的轻量级虚拟化,相对于传统的虚拟化技术,省去了Hypervisor层的开销,而且其虚拟化技术是基于内核的Cgroup和Namespace技术,处理逻辑与内核深度融合,所以在很多方面,它的性能与物理机非常接近。

1.4 TensorFlow

TensorFlow是谷歌基于DistBelief进行研发的第二代人工智能学习系统,是用来制作AlphaGo的一个开源的深度学习系统,其命名来源于本身的运行原理。张量(Tensor)意味着N维数组,流(Flow)意味着基于数据流图的计算,TensorFlow为张量从流图的一端流动到另一端的计算过程。TensorFlow是将复杂的数据结构传输至人工智能神经网中进行分析和处理过程的系统。TensorFlow可被用于语音识别或图像识别等多项机器深度学习领域,对2011年开发的深度学习基础架构DistBelief进行了各方面的改进,它可在小到一部智能手机、大到数千台数据中心服务器的各种设备上运行。TensorFlow完全开源,任何人都可以用。

2 大数据AI能力层构建方案

2.1 运营商大数据现状

从数据规模来看,截至2018年6月,中国4G用户数已突破11亿,移动用户已近15亿,基于如此庞大的用户数,无论是用户信息、消费记录还是设备日志数据均体量庞大且保持快速增长。江苏联通天玑数据中心目前每日入库数据超过600亿行,每日集群各任务使用数据量超过100 TB。

从数据质量来看,运营商数据都是基于真实用户使用记录以及设备运行记录,具备真实性和完整性,是非常优质的数据。江苏联通天玑数据中心作为O域综合数据中心,目前接入了移动网核心网信令数据、移动网MR信令数据、移动网性能数据、客服类数据、固网信令数据、固网认证数据、设备告警数据、部分B域数据等多种数据源,全部为用户、设备日常产生的真实数据。

从数据使用来看,运营商数据使用者类型各异,需求类型众多。有些数据分析如位置营销数据需要实时输出分析结果,有些数据如核心网信令统计数据则需要非实时的大数据量运算;有些数据使用者需要使用集群的计算资源与数据资源,有些数据使用者需要使用自有计算框架,有些使用者需要使用专用AI计算设备。

2.2 大数据AI能力层构建

结合运营商大数据特点与数据使用现状,采用单一数据源、松耦合、高异构的原则构建江苏联通天玑数据中心AI能力层。天玑数据中心Hadoop集群提供唯一数据源,软件上搭建Docker集群、Python计算节点、TensorFlow节点等多种计算平台,硬件上部署了Hadoop集群服务器、单节点高性能计算服务器、GPU加速AI服务器等,各平台之间既相互独立又能够互通数据、互相调用计算资源。

图1示出的是天玑数据中心分层架构。

2.2.1 Docker集群

在运营商大数据使用场景中,经常会有临时的大数据量分析任务,如重大节假日保障、重要会议保障、自然灾害临时保障等数据分析场景,这些任务开启时通常需要临时调度大量计算资源,但场景结束后这些计算资源就完全闲置下来。大数据分析算法通常具有非常复杂的架构,很多数据使用者需要在多台节点上重复部署大量依赖环境,浪费了很多时间、精力。

针对以上使用场景,部署了Docker集群。Docker集群所具备的快速部署、快速调度的特点非常适用于重保数据分析场景,任务开启时可以临时开启大批量

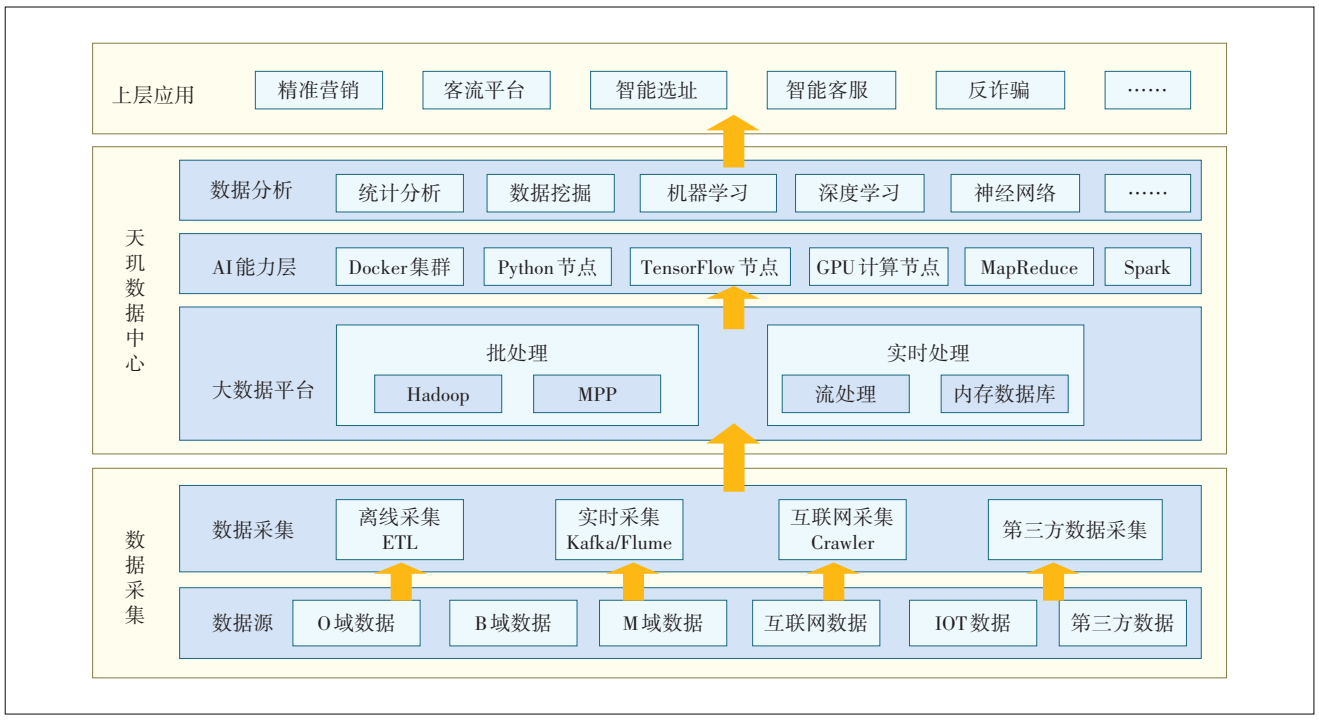

图1　天玑数据中心分层架构

Docker容器参与计算，任务结束后Docker容器即可关闭，无需冗杂的资源释放、清理工作。Docker容器能将完整的程序运行环境进行一次封装、多处调用，节省了大量计算环境部署的时间。

2.2.2　Python计算节点

Python用于机器学习开发环境，具有如下优点。

a）Python是解释语言，程序写起来非常方便。写程序方便对做机器学习的人很重要。因为经常需要对模型进行各种各样的修改，这在编译语言里很可能是牵一发而动全身的事情，Python里通常可以用很少的时间实现。

b）Python的开发生态成熟，有很多有用的库可以用。Python具备NumPy、SciPy、NLTK、os（自带）等丰富的API库，极大地方便了算法开发者，使其将精力专注于算法的设计上来。Python灵活的语法还使得包括文本操作、list/dict comprehension等非常实用的功能非常容易高效实现（编写和运行效率都高），配合lambda等使用更是方便。

c）Python的效率很高。解释语言的发展已经大大超过许多人的想象。很多比如list comprehension的语法都是贴近内核实现的。除了JIT之外，还有Cython可以大幅增加运行效率。最后，得益于Python对C的接口，很多像gnumpy、theano这样高效、Python接口友好的库可以加速程序的运行。

此外，Python还具备数据存储方便、数据获取方便、数据运算方便、输出结果方便、和其他语言交互方便、调用GPU加速方便、云系统支持方便等种种优点。

2.2.3　TensorFlow环境

TensorFlow可用于语音识别或图像识别等多项机器深度学习领域，并且TensorFlow完全开源，任何人都可以用。

TensorFlow表达了高层次的机器学习计算，大幅简化了第一代系统，并且具备更好的灵活性和可延展性。TensorFlow一大亮点是支持异构设备分布式计算，它能够在各个平台上自动运行模型，从手机、单个CPU/GPU到成百上千GPU卡组成的分布式系统。从目前的文档看，TensorFlow支持CNN、RNN和LSTM算法，这都是目前在Image、Speech和NLP最流行的深度神经网络模型。

在江苏联通MR共享层项目中，针对MR定位分析算法，TensorFlow框架的加入提升了30%的运算速度与10%的数据准确性。

2.3　各AI计算平台协作

天玑数据中心AI能力层采用松耦合架构，各AI计算平台皆可独立运行，但各平台如果全部独立运行无法发挥大数据优势，平台之间需要相互协作才能将各

自优势最大化。

图2示出的是各AI能力层相互协作结构。

天玑数据中心AI能力层建设原则为:所有数据存储在Hadoop平台中,各AI计算平台都需要通过接口机连接到Hadoop平台,进行身份验证和数据读取,然后才能进行数据分析,数据分析结果可输出也可写入Hadoop集群。

Docker平台本身是一个容器调度框架,可以在容器中封装Python、TensorFlow等计算环境,Docker集群中有一台Hadoop接口机,作为权限认证、数据读取的中间环节,Docker集群可以通过接口机从Hadoop集群中读取数据,然后进行数据处理、分析。

Python节点安装Hadoop代理,本身可以作为集群权限认证、读取数据的节点,可以将数据读取后进行分析。但是Python语言本身性能有限,并且无法进行并行计算,数据量增加到一定程度就无法很好地完成任务。可使用PySpark框架来通过Python调用Hadoop的计算资源。为了不破坏Spark已有的运行时架构,Spark在外围包装一层Python API,借助Py4j实现Python和Java的交互,进而实现通过Python编写Spark应用程序,其运行时架构如图3所示。

3 AI能力层应用案例

公司提出了推动2G用户迁转4G的要求,需要通过对天玑数据中心相关数据进行挖掘,智能推荐2G迁转4G目标用户群。本系统共计分为2个模型。

3.1 2G转4G潜在客户识别模型

在用户画像的基础上,分别筛选历史完成2G迁转

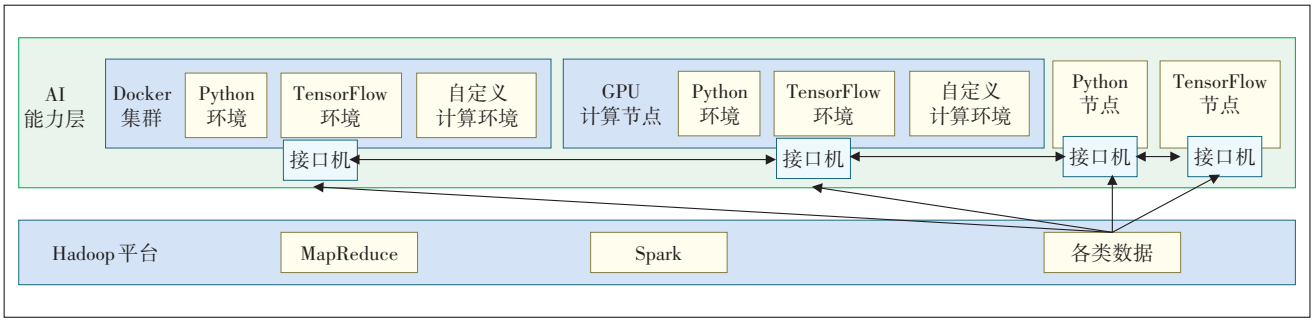

图2　各AI能力层相互协作结构

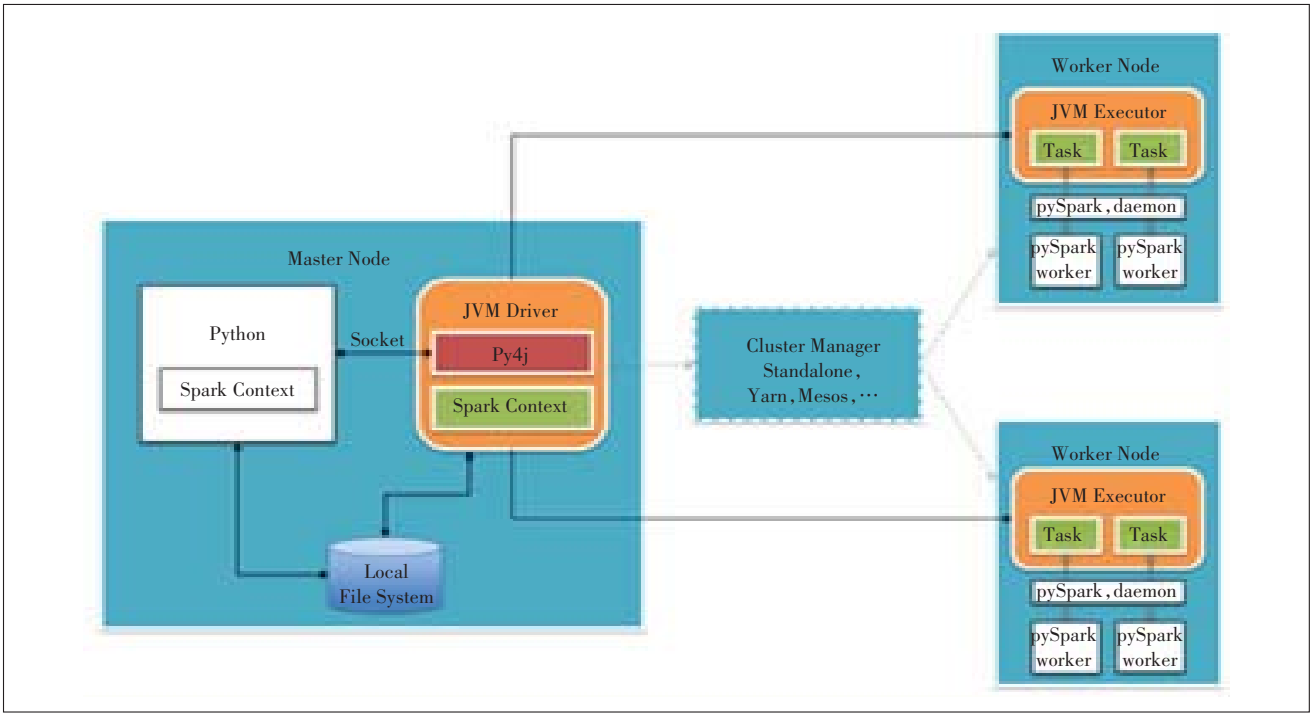

图3　PySpark架构

4G的用户历史消费数据，通过机器学习算法，提取出这些用户的特征，并以此来识别当前登网2G用户中，具备这些特征的消费者，将之识别为高推荐值的用户。

3.1.1 模型input参数

通过用户画像，主要输入无量纲字段带入模型中计算。引入无量纲字段具有如下优点。

a）避免了归一化的问题。

b）后续若要引入集成模型，进一步提高模型准确度，不需要额外的数据处理工作。

c）适用于BP神经网络模型。

导入模型的数据实例见表1。

研究2018年1—6月用户的情况，若1月份为2G用户，6月份为4G用户，则选择该用户1月份的数据，并设置为正例。若1月份为2G用户，6月份仍为2G用户，则选择该用户1月份的数据，并设置为反例。

3.1.2 模型的设置

模型采用xgboost，在构建模型阶段，要特别注意数据不平衡的问题。从历史数据看，反例样本数据量为正例样本数据量的5~7倍，远远高于正例样本的数据量。为了消除样本不平衡的问题，可采用如下方法。

a）在调参中设置样本不平衡参数，进行纠偏。

b）选取部分反例样本，舍弃多余样本，达到正反例平衡。

c）增加正例样本的数目，可以通过复制部分正例，或者通过算法增加虚拟的正例样本数目。

d）切割反例样本n份，每份反例样本均与正例进行模型计算，计算结果进行投票统计或者求取平均值。

本模型采用第4种方法。

模型示意图如图4所示。

3.1.3 模型训练环境

通过PySpark组件，直接调用Hadoop集群Spark组件，使用Hadoop集群的计算资源和数据资源进行模型训练。

3.2 潜在换机预测模型

选取入网时长超过9个月的现网存量用户，与前8个月关联，保留存量用户。根据用户历史数据，预测下个月是否会换手机。

3.2.1 输入数据清洗

选取入网时长超过9个月的现网存量用户，与前8个月关联，保留存量用户，并筛选出手机用户，手机IMEI号为正常字符的用户。选取字段表中相应字段，再与终端表中的mzie_type、is_dual、mz_type字段进行关联。最终得到700万条数据。

数据字段表见表2。

3.2.2 数据建模

针对换机预测课题，构建基于xgboost算法模型，将筛选的700万条数据分批导入模型进行训练，给用户打上换机可能性的标签。市场操作时，可以调整可能性阈值，来筛选换机可能性较大的用户，精确营销。

模型示意图如图5所示。

表1　导入模型数据实例

msisdn	age_index	user_sex	hprovince	hcity	vcity	innet_years	type_vip
156***	0.0555	0	130	13002	13002	2.0833	0.2
nds_score	online_ratio	ratio_4g	dinner_index	card_index	phone_index	roam_rate	price_index
0.5	0.2	0.5	0.971998	0.097222	0.8	0.9	0.5
flow	talklen	bill	wireless_ratio	secondary_card	user_loyalty	potential	label
0	0.25	0	0.8076	0.3679	0.0945	0.110	1

图4　2G转4G潜在客户识别模型示意图

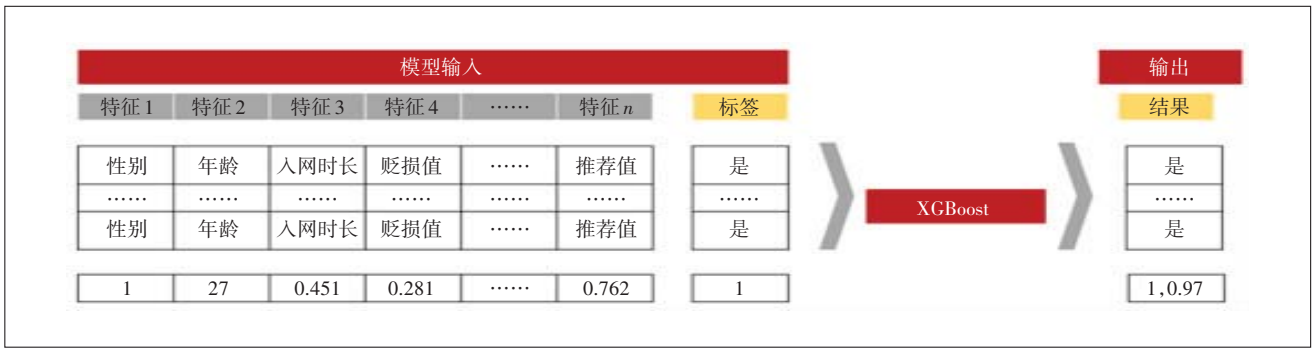

图5　潜在换机预测模型示意图

表2　输入数据字段表

基本字段	指标解释	指标说明	信息来源
talken	通话时间	信令层捞取变量	xdr.cs_calllog_acess
phone	本端号码		
phone_vcity	本端号码拜访地		
rphone_hcity	对端号码归属地		
rphone	对端号码		
eventtype	事件类型		
subevttype	事件子类型		
day	日期	分区变量	自定义
vcity	拜访地市（本段号码）		
sum_talklen	累计通话时长	中间变量	聚合计算
sum_counts	累计通话频次		

3.2.3　模型训练环境

通过Hadoop集群Hive数据仓库读取相关数据，读取到Python计算节点，使用相关硬件加速接口调用GPU进行计算加速。

3.3　2G迁转4G用户数据挖掘模型效果

以IMEI号为标识，挖掘用户近9个月的终端更换情况，并统计这9个月用户更换手机次数。预测月前3个月账单、流量、通话的增长、减少情况。用户年龄、性别、入网时长、NPS评分等信息。将数据送入xgboost算法进行训练，得到换机预测算法模型。再将需要预测月份的数据送入模型，得到该月用户换手机的可能性指标。筛选可能性大于某一阈值的用户名单，系统直接推送给相关市场部门，进行精准营销。本次以沭阳的4个营业厅作为潜在换机客户营销试点。一个星期的时间，共拨通106户，其中成功办理14户，拨通办理成功率为13.2%。相较于非精准营销的普通营销方式，每月到营业厅办理中国联通合约机，占每月中国联通用户换手机总数的2%这个比例，有了500%以上的提升，达到了精准营销的目的。

参考文献：

[1] 王浩，马艳. 关于人工智能的应用与发展[J]. 考试周刊，2009(8)：148-150.
[2] 黄勇军，冯明，丁圣勇，等. 电信运营商大数据发展策略探讨[J]. 电信科学，2017，29(3)：7-11.
[3] 张云帆. 电信运营商大数据发展策略与价值挖掘[J]. 移动通信，2016，40(5)：20-23.
[4] 童晓渝，张云勇，房秉毅，等. 大数据时代电信运营商的机遇[J]. 信息通信技术，2013(1)：5-9.
[5] 梁柏青，陆钢，李慧云，等. 运营商能力开放架构研究及发展思路探讨[J]. 电信科学，2011，27(4)：7-11.
[6] 程学旗，靳小龙，王元卓，等. 大数据系统和分析技术综述[J]. 软件学报，2014(9)：1889-1908.
[7] 李文栋. 基于Spark的大数据挖掘技术的研究与实现[D]. 济南：山东大学，2015.
[8] NANDI A. Spark for Python Developers[M]. Packt Publishing Ltd，2015.
[9] SAMOSIR J，INDRAWAN-SANTIAGO M，HAGHIGHI P D. An evaluation of data stream processing systems for data driven applications[J]. Procedia Computer Science，2016，80：439-449.
[10] ABADI M，BARHAM P，CHEN J，et al. Tensorflow: a system for large-scale machine learning[C]//OSDI. 2016，16：265-283.
[11] 章敏敏，徐和平，王晓洁，等. 谷歌TensorFlow机器学习框架及应用[J]. 微型机与应用，2017，36(10)：58-60.
[12] 华为Docker实践小组. Docker进阶与实战[M]. 北京：机械工业出版社，2016.
[13] 董西成. HADOOP技术内幕——深入解析YARN架构设计与实现原理[M]. 北京：机械工业出版社，2013.

作者简介：
王瑜，毕业于南京邮电大学，业务质量管理技术总监，硕士，主要从事网络大数据平台规划、技术架构制定、课题预研等工作；邓程，毕业于南京师范大学，工程师，硕士，主要从事网络大数据平台管理与维护、大数据平台数据采集管理等工作；蒋涛，毕业于大连海事大学，工程师，本科，主要从事大数据分析、大数据建模、机器学习算法研究等工作；汪剑桥，毕业于南京工业大学，工程师，本科，主要从事大数据平台数据采集实施及日常维护等工作。