



基于 Hadoop 的 OSS 域数据建模与采集方法研究

雷 蕾¹, 李景文¹, 宫大鹏¹, 王 睿¹, 苏 雷², 陈宁江³

(1. 中国移动通信集团广西有限公司 南宁 530022;

2. 亿阳信通股份有限公司 南宁 530022; 3. 广西大学计算机与电子信息学院 南宁 530004)

摘 要: 针对电信运营商在构建大数据中心中所面临的数据采集、建模问题, 分析了 OSS 域的数据现状和特点, 给出了一种基于层次分类梳理的建模方法论和数据模型管理思路, 可实现对 OSS 域数据的统一采集、集中建模和数据共享, 通过移动互联网端到端分析专题的应用案例进行验证, 可对电信行业建设大数据中心提供借鉴。

关键词: 企业数据中心; ETL; 运营支撑系统; 数据采集; Hadoop

doi: 10.11959/j.issn.1000-0801.2015021

Study on Data Modeling and Collection in OSS Based on Hadoop

Lei Lei¹, Li Jingwen¹, Gong Dapeng¹, Wang Rui¹, Su Lei², Chen Ningjiang³

(1. China Mobile Group Guangxi Co., Ltd., Nanning 530022, China;

2. BOCO Inter-Telecom InC., Nanning 530022, China;

3. College of Computer and Electronic Information, Guangxi University, Nanning 530004, China)

Abstract: According to the data acquisition and the modeling problem of telecom operators facing in the construction of EDC, the data status and data characteristics in the OSS was analyzed, a modeling method of hierarchical classification of combing theory and data model management idea were proposed, the unification of OSS data collection, sharing the modeling and data concentration were realized. Through the application of end to end analysis on mobile internet terminal as case subject, it provides reference for the construction on telecom industry EDC.

Key words: EDC, ETL, OSS, data acquisition, Hadoop

1 引言

随着“大数据时代”的来临, 电信运营商已意识到自己手中“数据金库”的价值, 正在积极推动传统的分析支撑体系向分布式大数据架构进行演进, 以构建集中化的企业级大数据中心, 实现整个企业的数据融合, 提供开放的数据能力, 逐步满足全网、全渠道、全业务一体化营销服务的要求。

在广西移动的企业级大数据中心建设及企业数据融

合工作中, 主要采用 Hadoop 架构搭建大数据中心, 数据来源主要包括电信行业通常所划分的 OSS (operation support system, 运营支撑系统)、BSS (business support system, 业务支撑系统)、MSS (management support system, 管理支撑系统) 3 个领域的的数据, 3 个域数据的采集处理复杂度不同, 服务于不同的场景。其中, OSS 域数据种类繁多, 数据量大, 采集过程复杂, 同时包含结构化及非结构化数据, 是三域数据中最复杂的一类, 也是建设大数据中心的重点及关键工作之一。本文结合业务实际需求, 介绍基于 Hadoop 平

台的 OSS 域数据采集和数据管理建模的解决方案和实现技术,目标是域数据有效地传递和组织,使得数据中心切实有效地对数据进行管理,对应用进行数据供给。

2 OSS 域数据背景与现状

OSS 域系统的主要数据类型包括网络的资源数据、告警数据、性能数据、网络测量数据、信令数据、工单数据、日志数据等,每天产生的数据量在 60 TB 左右,以烟囱的方式在网络优化系统、话务网管、数据网管、信令监测系统、综合网络资源管理系统等多套系统中采集和存储。

如图 1 所示,从数据量来看,来自于信令监测系统的信令数据占整个 OSS 域数据量的 97%, 信令数据记录了用户的通话记录和上网记录等信息,同时含有结构化数据和非结构化数据,是对用户行为进行分析、开展大数据分析营销的重要基础数据;从数据获取复杂度及采集频次来看,网络优化系统(简称网优)的数据获取复杂度最大,采集频次最高,网优数据全面记录了网络的各类性能指标,以结构化数据为主,是利用大数据技术开展 2G/3G/4G/WLAN 网络协同性能分析,提升用户网络使用感知的重要基础数据。

在上述的 OSS 域各系统中,都同时保留了各自采集到的原始数据和经过计算后的汇总数据,这种烟囱式的数据管理方式缺乏有效的管理机制,部分数据存在重复,如网优系统和话务网管系统同时都有话务量、掉话率数据,各系统间的数据也没有显性的联系,如信令系统中的信令数据与网优系统中的网络性能数据无法实现关联分析。因此,数据的孤岛现象严重,数据价值无法得到充分的挖掘和利用。国内许多互联网公司为了应对这样的数据自主研发了许多架构,例如腾讯为应对其自有业务的大数据处理构建了一套 TDW (Tencent distributed data warehouse, 腾讯分布式数据仓库)架构,小米对 HBase 进行不断改进和扩展。中国科学院计算技术研究所对行列混合式存储技术、HDFS 数据压缩技术的研究也已取得初步成效。

针对上述问题,为实现 OSS 域各类数据的统一采集、统一存储、集中建模和数据共享,笔者研发了一个基于 Hadoop 的 OSS 数据统一采集平台 (HD-OSS), 如图 2 所示。HD-OSS 平台基于 Hadoop 技术架构搭建,实现对各类数据的采集、清理、抽取和转换,并进行轻量级汇总计算。采用统一的云化 ETL 技术进行数据预处理,实现数据接入的统一管控、海量多样化数据处理,提供各个域数据融合

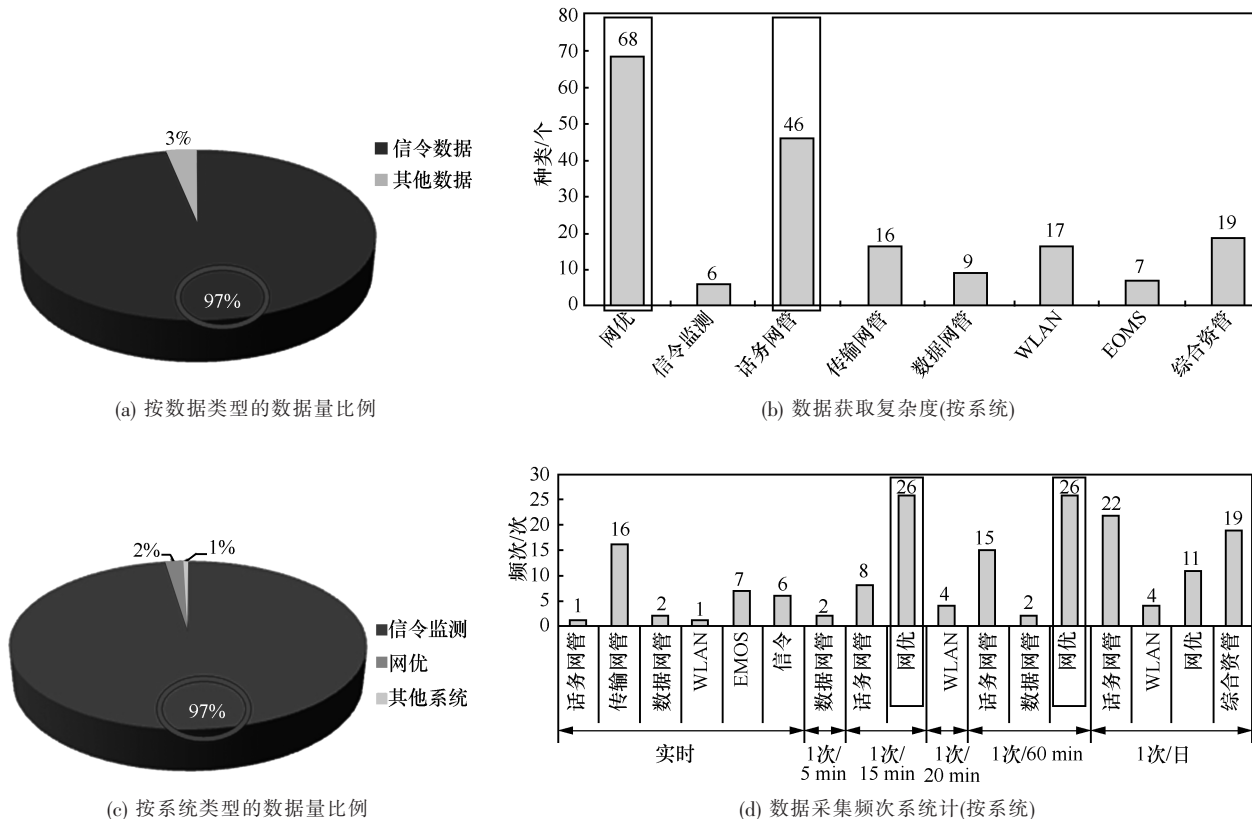


图 1 OSS 域数据分布

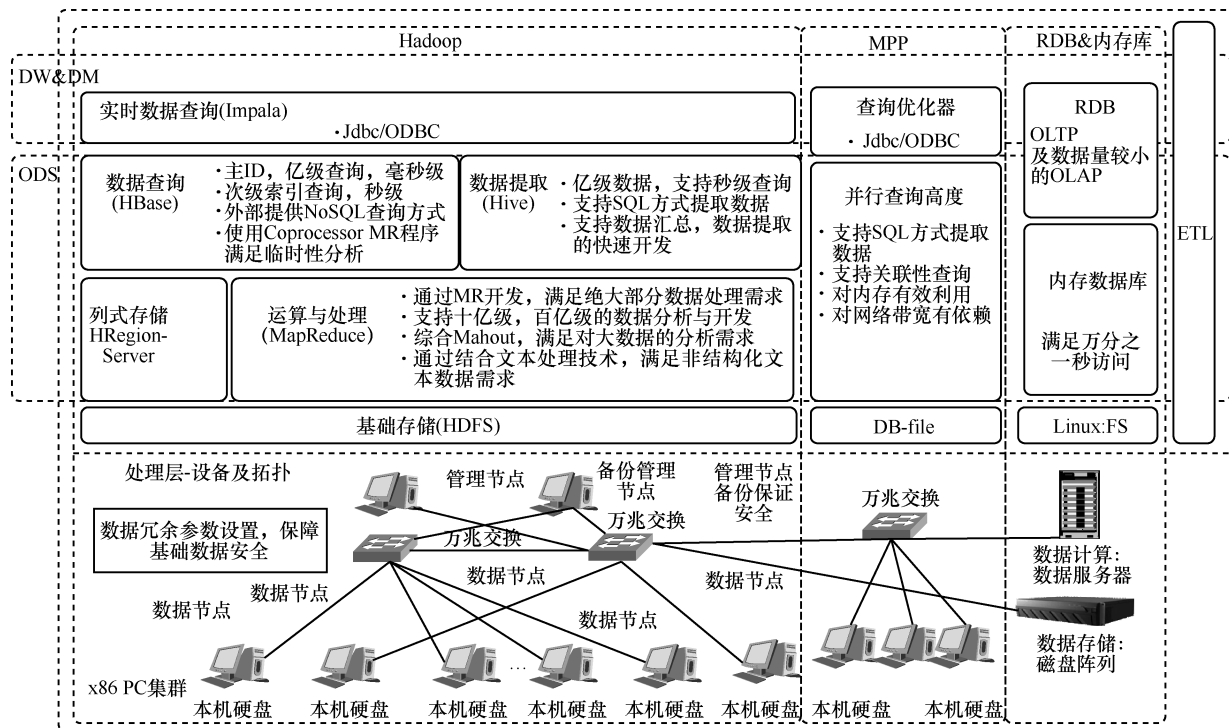


图2 混搭式数据中心架构

的基础,清洗后的数据在ETL平台完成统一的调度转换,最后进入数据中心。采用异构分层存储架构,引入大规模并行分布式(MPP)数据库和关系型数据仓库,按照数据在生命周期中不同阶段对存储的性能需求,数据分类分级后分布存储在数据仓库和MPP数据库。HD-OSS平台还需实现统一数据建模、统一编码管理和统一指标管理。在平台的建设中,数据采集管理和数据建模是确保大数据中心数据的完整性、准确性和一致性的关键问题。本文主要对它们进行研究并提出解决方案。OSS域数据面临的最大问题是业务的影响会直接扩散到网络设备采集层,一旦上层的应用需求改变,底层数据采集的时间粒度、空间粒度、指标算法、存储方式等可能会发生改变,OSS域数据与应用解耦是关键课题。

3 基于Hadoop的数据采集管理

OSS数据统一采集平台主要解决网络中网元的采集适配问题,统一管理对网元、信令、路测等设备的采集进程,监控整个采集过程和数据质量,通过ETL完成数据清洗和部分轻量级数据的汇总,将数据输送进入数据中心,这是整个OSS采集的核心。统一采集平台重点对告警、性能(含网优)、资源和信令四大类数据采用不同的方式进行处理,如图3所示。

- 对于资源数据来说,数据量较小,数据来源主要为工程配置和资源入网割接时录入的手工数据,对资源状态的实时性要求不高,每天仅需更新一次,因此每天由统一采集平台采集后,分别提供给大数据中心和综合资源系统。
- 对于告警数据来说,数据量大,具备流式处理的特征,而且实时性要求非常高,所以统一采集平台采集到数据后将数据分为两份,一份送到综合告警管理系统,一份送入企业大数据中心。
- 对于性能数据来说,数据量一般,但数据来源多、采集频次高、数据结构复杂,统一采集平台完成数据采集及解析后,再进行轻量级计算及汇总,将原始数据及汇总数据根据需要送到OSS域的各系统及企业大数据中心。
- 对于信令数据来说,数据量特别大,但数据来源单一,仅需要全量接入信令监测系统提供的数据或文件即可,但由于存在大量的非结构化数据,因此需要统一采集平台采用爬虫技术对非结构化数据进行分析,所有的数据全部送入到企业级大数据中心保存。

下面以最复杂的性能数据为例,说明统一采集平台基于Hadoop平台进行ETL数据处理的过程,如图4所示。

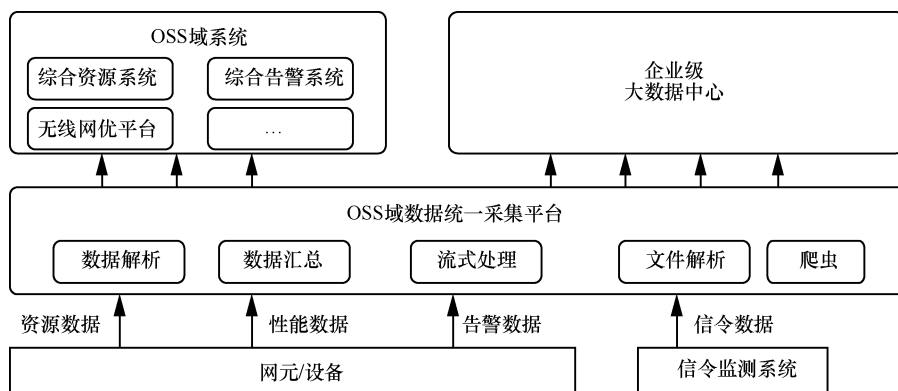


图3 OSS域数据统一采集示意

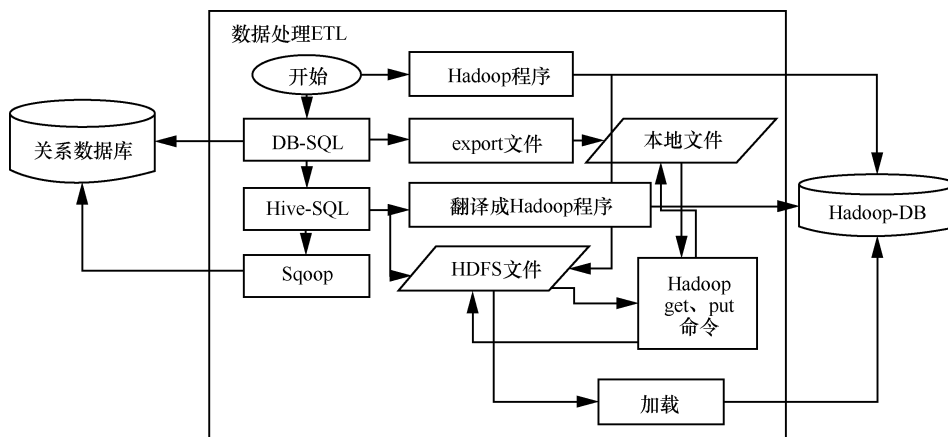


图4 基于Hadoop的ETL数据处理流程

基于Hadoop平台的ETL与普通的ETL过程没有本质区别。统一采集平台完成底层数据采集后即进入ETL过程，主要通过Oozie或其他调度工具实现周期调度，另外HDFS和FS系统可以通过流程打通，Hive与Hbase也可以通过Sqoop与其他数据库打通。大数据的处理过程也由SQL方式提供，可以引入一些数据挖掘的处理过程。

4 OSS域数据建模管理设计

4.1 数据建模方法论

通过总结实际经验，笔者提出了一种层次型分类梳理数据建模方法，如图5所示。数据建模和管理遵从自上而下的分类梳理对数据进行重新规划，主要根据数据源归属、特点、规模等情况对数据进行模型分层和数据粒度分层，按照ODS(operational data store,操作型数据存储)、DW(data warehouse,数据仓库)和DM(data market,数据集市)3层进行数据建模。ODS用于存放从数据源直接抽取出来的数据，这些数据在数据结构、数据之间的逻辑关系上与

数据源基本保持一致，DW主要根据企业模型而来，而从业务专题出发建立的模型则会固化在DM层。在ODS层和DW层，一般会以运营数据模型和企业模型为出发点，指导数据在采集计算层面大致分出哪些是业务支撑驱动的数据，哪些是运维支撑驱动的数据，例如借助eTOM模型定义出企业规划、运营支撑、故障、计费、开通、保障等层面数据覆盖范围和数据归属，并能结合现状调研初步估算出数据的使用频度和粒度大小。而DM层的业务模型则主要根据业务应用方向分专题建模，如流量经营业务模型、客户感知业务模型等。

4.2 数据模型的分层原则

OSS域数据将按照如下原则划分到数据模型的ODS、DW、DM 3个层次。

(1) ODS层模型

- 原则一：按照10亿~1 000亿的亿级数据进行分层。
- 原则二：按照数据之间的推演规则进行分层，如图6所示。

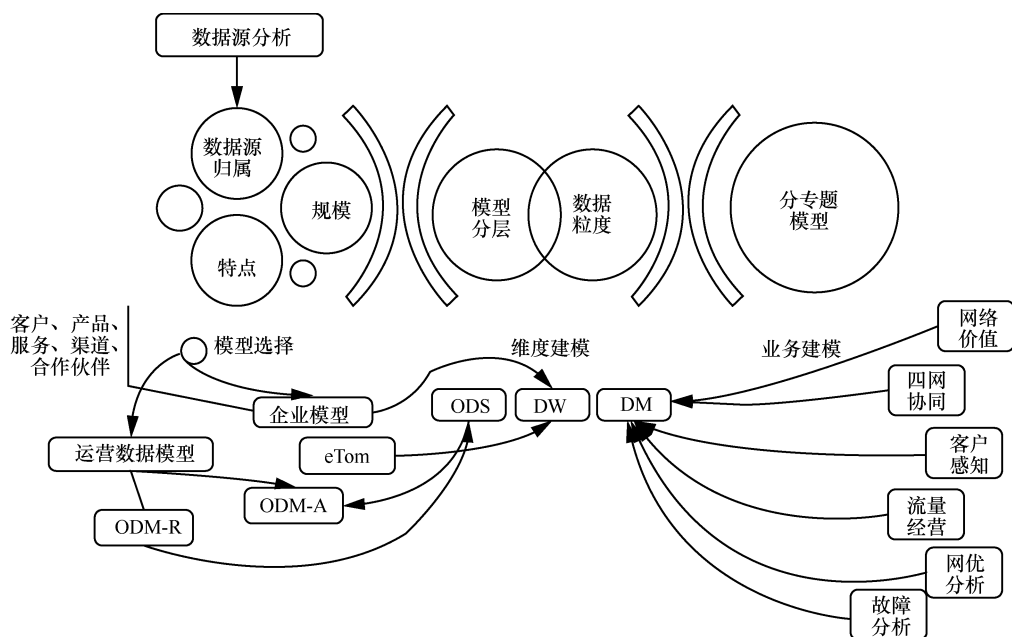


图5 建模方法论

	网管	监控	网优	信令	EOMS	短彩中心	B域
用户级				IMEI-MSISDN	投诉工单		客户信息 客户账单
细粒度网元	动环 局数据 厂商计数器 BTS 交换 设备	告警	细粒度网元性能 小区性能		故障工单		
会话级				Gn Mc PDP			
话单级						短信 彩信	客户话单
测量级			MR				
会话状态				Gb			

逻辑在ODS层, 物理在DW层

该数据保存在ODS层

图6 ODS层模型

(2)ODS层

ODS层也被称作数据缓冲区,存储包括O域专业网性能、告警、资源、DPI等系统数据。建立清单级、会话级数据(PI、KPI级数据)模型。例如网管系统的性能数据会按专业组织最细粒度的数据模型表示。DW层模型如图7所示,在ODS层保存清单级数据后,通过进一步关联汇总形成DW层数据,主要为了支撑上层应用分析和钻取分析,根据业务需求对事实数据和维度数据进行有效的组织和规范,提高数据的访问效率。按企业模型和不同细分粒度组织建模,并设计出模型间的关联关系。一般

DW层按4类模型进行汇总,分别汇总为维度表、事实表、聚集表、临时表。

- 维度表:根据系统各个主题逻辑数据模型的维度设计的物理数据库表,记录模型维度信息。
- 事实表:记录各专题最细粒度的事实信息,物理数据库表的设计是依据逻辑数据模型设计的。
- 聚集表:记录各专题汇总粒度的事实信息,物理数据库表的设计是依据逻辑数据模型设计的。
- 临时表:根据数据ETL设计定义,即为中间表,无模型设计意义。

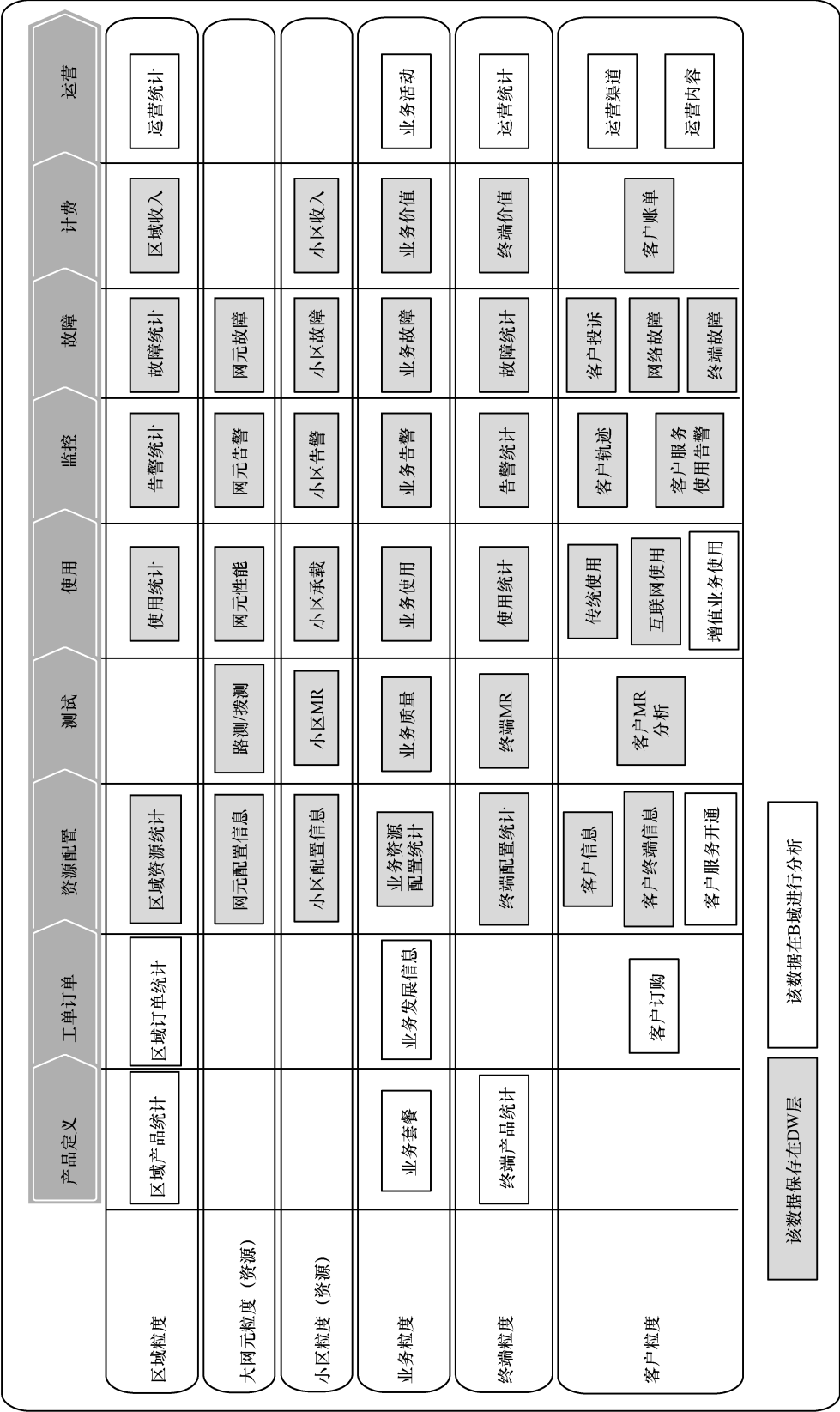


图 7 DW 层模型



(3)DM 层模型

DM 层模型面向应用组织建模,以业务需求应用为单位对 DW 层的数据进一步组织和存储,是面向需求以及未来需求变更、扩展的模型。一般以业务作为驱动,在 DW 模型基础上按业务专题所需要的业务再次进行数据组织,通过按上层业务组合、业务钻取等形成最终面向应用的 DM 层模型;各专题域模型引用 ODS 层企业流程模型,构成逻辑专题域数据。

以几个分析专题为例,将所需要的业务进行模型组

织,按上层业务组合、业务钻取等建立模型,具体见表 1。

通过以上过程,即完成了数据中心架构支撑下的分析应用建模支撑。图 8 给出了一个整体业务数据建模例子,在数据中心中组织了从基础数据、数据仓库、数据集市 3 层数据模型,通过数据抽取进行逐层汇总,汇聚到 DM 层以后,即可形成高价值业务分析、高流量用户分析等面向分析专题的模型数据。当上层业务发生变更,开发者可根据情况在 DW 层再次快速组织数据汇总,省去重新采集建模的开发时间。按这样层次组织的数据模型即可实现对应用

表 1 分析专题 DM 模型示意

专题名称	表说明	表名	数据粒度	存储周期
四网协同	GN 口日统计	TM_SIGN_GN_D	天	30 天
	GN 口月统计	TM_SIGN_GN_M	月	24 月
	天粒度高流量用户	TM_BSS_HIGH_FLOW_CUST_D	天	30 天
	月度高流量用户	TM_BSS_HIGH_FLOW_CUST_M	月	24 月
	...	...	...	...
价值小区分析	移动自有业务月统计	TM_BSS_SERVICE_M	月	24 月
	价值小区收益月粒度表	TM_NIOS_VALUE_CELL_INCOME_M	月	24 月
	BSC 收益月粒度表	TM_NIOS_VALUE_BSC_INCOME_M	月	24 月
	GPRS 收益月粒度表	TM_DNMS_VALUE_GPRS_INCOME_M	月	24 月
	...	...	...	...

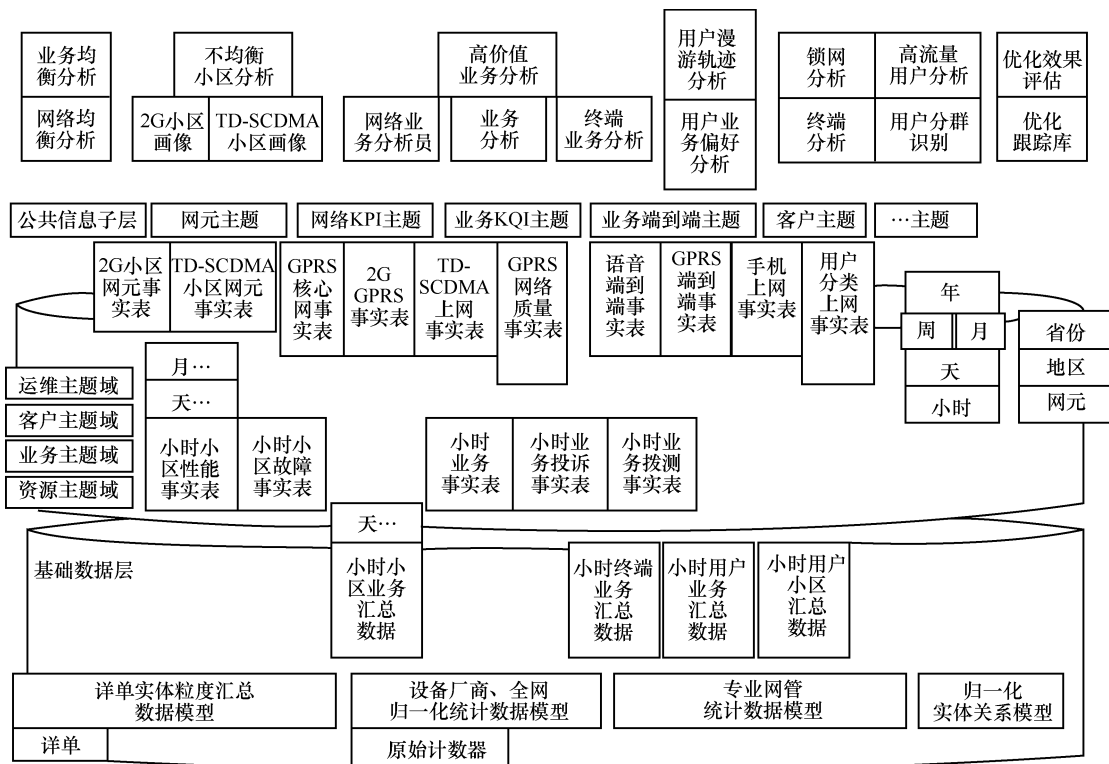


图 8 按业务分层的数据模型例子

开发的快速模型支撑,响应性能分析需求的高时效性。

5 应用场景

图 9 为 OSS 域性能分析示意。

基于 OSS 域数据的分析应用主要集中在客户感知、业务质量、网络性能 3 个方面,这些分析专题的数据来源复杂,结构化程度不统一,以下重点以移动互联网端到端分

析专题为例,进行数据采集及模型管理的介绍。

如图 10 所示,移动互联网端到端分析专题重点围绕 LTE 网络移动互联网业务,开展端到端感知分析。从业务感知指标入口,总结自上而下的问题溯源关联规则,提供各类型业务(浏览、下载、视频和即时通信四大类业务类型)的业务质量端到端分析,实现精准的问题定界和定位。

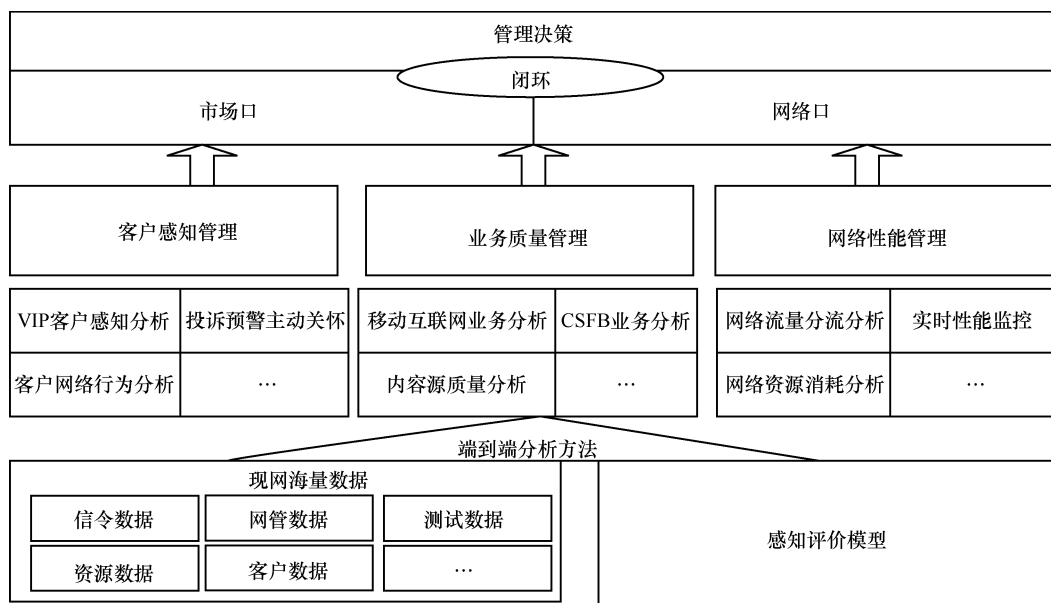


图 9 OSS 域性能分析示意

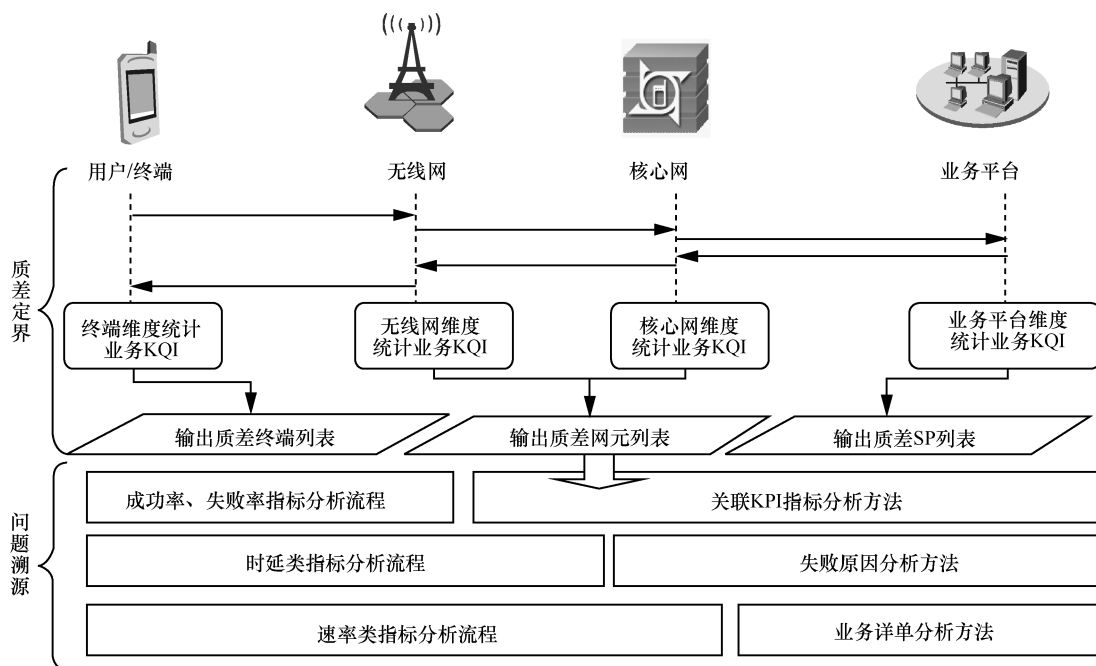


图 10 移动互联网端到端业务质量指标定位



针对业务端到端质量问题,进行问题溯源分为两大步骤:第一,质差定界;第二,问题溯源。两大步骤均需要通过梳理指标体系实现。根据业务梳理得到的指标集结合模型分层原则可设计出采集及模型管理的要求,见表2。

目前通过业务质量管理平台,能够获取移动互联网的4类业务,分别为网页浏览、视频业务、即时通信、应用下载业务数据。通过DPI平台可对用户面S1-U口、Uu口与SGi口信令解析,获取KQI/KPI/PI指标。表3为过程指标分解。

浏览类业务主要包括附着、承载激活、DNS解析、TCP链接、HTTP请求的5个阶段,当成功率类指标发生波动时,首先关联KPI指标确定业务失败在哪个过程,然后根据这个过程中出现的错误码分析失败原因,最后通过FM/PM/CM排查KPI,横向对比,定界具体异常点。基于业务过程的KPI指标能够定位业务失败原因,对业务质量进行管理和优化。通过对两类指标进行业务关联,即可完成故障溯源的过程。

采集方面,通过OSS统一采集平台实现初步ETL过程,更好地解决算法不透明、缺乏管控手段、统计口径不一致、接口杂乱等问题,平台统一管理所有的采集通道和时间轴。在ETL完成后,数据按专业维度存放在数据中心的ODS层。

一般在ODS层保存两年的清单级数据(PI、KPI级数

据),通过进一步关联汇总为DW层模型。在DW模型基础上,按互联网端到端分析专题所需要的业务再次进行数据组织,按上层业务组合、业务钻取等建立模型,形成最终面向应用的DM层模型,即完成了整个数据中心架构支撑下的分析应用建模支撑。

关于OSS数据对大数据业务的其他支撑,如图11所示,还可以通过对信令数据进行深度挖掘,研究客户的上网行为、基本特征、位置轨迹、消费偏好等信息,对数据进行“脱敏”后,可完成用户画像,为自有业务营销、客户服务提供数据支撑。

6 结束语

电信运营商数据中心建设必须重视OSS域数据建设,这是保证整个数据中心质量、支撑应用分析业务开展的重要基础。目前中国移动正在全国推动数据中心的建设,在建设过程中怎样处理好数据采集与数据存储的关系,受到业界公司的重点关注。本文介绍了对OSS域数据进行大数据中心架构管理的采集流程和数据建模的关键方法,为其他省建设企业数据中心提供可实施的建设经验。目前广西移动正在借助已有的平台对数据中心进行重新规划,对于OSS域的数据供给,建设完成后可以对架设在数据中心之上60%~70%的应用群提供OSS侧的数据支撑。下一步的工作主要包括:适应业务发展的多样化,需要研究将更多的数据采

表2 互联网端到端指标体系

业务类型	信息交互的可用性	信息交互的及时性	信息交互的稳定性
浏览类	首页响应成功率 页面显示成功率	首页响应时长 页面显示时长 页面下载速率	
视频类	流媒体播放成功率	流媒体播放等待时长 流媒体业务下载速率	流媒体播放中断率
即时通信类	业务登录成功率	业务登录时延 大数据(视频/图片)传输上下行速率	异常掉线率
应用下载类	应用下载成功率	应用下载速率	

表3 过程指标分解

附着	承载建立	DNS解析	TCP链接	HTTP过程
EPS附着请求次数	专用承载激活请求次数	DNS解析成功率	TCP建立成功率	get响应成功率
EPS附着成功次数	专用承载激活成功次数	DNS解析时长	TCP建立时长	get响应时长
...	...	...	...	...

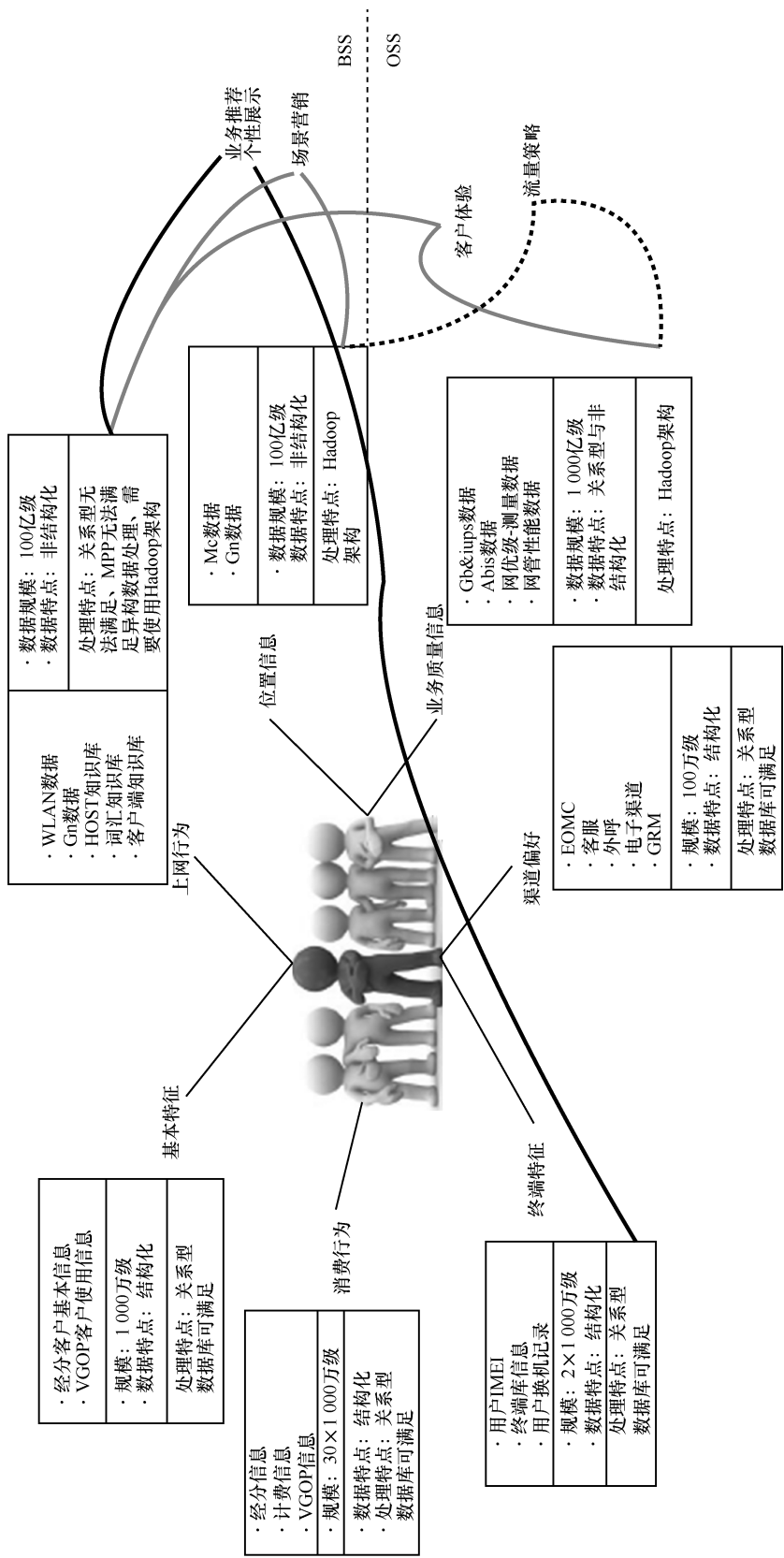


图 11 OSS 域数据的分析支撑



集、计算存储模型融入到数据中心的整体架构中;深入研究 OSS 在大数据支撑的应用,例如通过对信令数据的深度挖掘,研究客户上网行为、位置轨迹、消费偏好等信息,为自有业务营销、客户产品服务提供数据支撑。

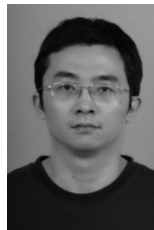
参考文献

- 1 Barroso L A, Dean J, Holze U. Websearch for a planet: The Google cluster architecture. IEEE Micro, 2003, 23(2):22~28
- 2 Xu Y, Kostamaa P, Qi Y. A Hadoop based distributed loading approach to parallel data warehouses. Proceedings of SIGMOD 2011, Athens, Greece, 2011
- 3 詹志强,孟洛明,邱雪松. 多专业网综合网管系统体系结构的研究. 北京邮电大学学报, 2003 (1)
Zhan Z Q, Meng L M, Qiu X S. Architecture of network management system for multi-technology network. Journal of Beijing University of Posts and Telecommunications, 2003(1)
- 4 Liu X F, Thomsen C, Pedersen T B. Cloud ETL: Scalable Dimensional ETL for Hadoop and Hive. DB Technical Report, 2012
- 5 Mohammed M, Mohd S A. A Framework for Interoperable Distributed ETL Components Based on SOA. Proceedings of ICSTE 2010, San Juan, PR, USA, 2010
- 6 陈桂汉. 综合电信管理解决方案. 北京: 电子工业出版社, 2002
Chen G H. Integrated Telecommunications Management Solutions. Beijing: Publishing House of Electronics Industry, 2002
- 7 Ghemawat S, Gobiuff H, Le-ung S T. The Google file system. Proceedings of 19th Symposium on Operating Systems Principles, Lake George, New York, USA, 2003
- 8 Corbett J C, Dean J, Epstein M. Google's globally distributed database. Proceedings of the 10th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2012, Hollywood, CA, USA, 2012
- 9 Samuel P, Mall R, Kanth P. Automatic test case generation from UML communication diagrams. Information and Software Technology, 2007(49): 158~171
- 10 Ling F, Chang E, Dillon T. A semantic network-based design methodology for XML documents. ACM Transactions on Information System(TOIS), 2002, 20(4): 1~6
- 11 Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. Proceedings of 6th Symposium on Operating Systems Design and Implementation, OSDI '04, San Francisco, USA, 2004

[作者简介]



雷蕾,女,中国移动通信集团广西有限公司信息系统部工程师,主要研究方向为云计算、大数据、网管支撑系统。



李景文,男,中国移动通信集团广西有限公司信息系统部工程师,主要研究方向为云计算、大数据、网管支撑系统、IT 架构。



王睿,男,中国移动通信集团广西有限公司信息系统部工程师,主要研究方向为大数据、IT 架构、云计算、网络安全管控。



宫大鹏,男,中国移动通信集团广西有限公司信息系统部项目经理,主要研究方向为网管支撑系统、系统规划。



苏雷,男,亿阳信通股份有限公司规划咨询工程师,主要研究方向为行业大数据、运营商 OSS 系统规划。



陈宁江,男,博士,广西大学计算机与电子信息学院教授,主要方向为网络分布计算、云计算、软件工程。