电信运营商大数据应用典型案例分析

余 飞
上海邮电设计咨询研究院有限公司 上海 200092

摘 要 移动互联网时代，云计算、物联网、智能终端等新技术新应用不断涌现，移动互联网的迅猛发展给电信运营商带来流量收益的同时，也带来了新的机遇和挑战。文章结合大数据的技术现状以及电信运营商的数据特点，分析电信运营商大数据发展遇到的问题，探讨电信运营商应用大数据的策略，最后提出一种适合电信运营商的大数据平台架构和方案。

关键词 大数据；电信运营商；Hadoop；移动互联网；数据流量

1 大数据技术简介

随着移动互联网、云计算、物联网技术和业务的发展，全球数据量正在呈爆炸性指数级增长。根据IDC发布的报告显示，2012年全球数据量约为2.8ZB，并以大约每两年翻一番的速度增长，预计到2020年，全球将产生35ZB的数据量。这意味着我们正进入大数据时代。

维基百科将大数据定义为：大数据是很多各种数据集汇合起来的数据集合，规模非常大并且复杂，以至于很难用常规的数据管理工具或传统的数据管理技术来处理这些数据。Facebook、Twiteer、微博等各类社交网络，各种智能终端，医疗影像、监控录像等各类视频以及遍布全球各个角落的各种传感器，无一不是数据来源。大量新数据源的出现导致非结构化的数据迅猛增长，占比超过80%，超越了传统关系型数据库的管理能力，使得大数据的存储、管理和处理很难利用传统的关系型数据库去完成，进而无法提取个中价值[1]。

以Hadoop为代表的大数据技术应运而生，它是一种非关系型数据库系统及分布式运算架构。近几年，Facebook、Google、Amazon、Yahoo、阿里巴巴和百度等开始了大数据化的进程，他们依托自己的数据优势，采取灵活深入的分析方法进行基于大数据的挖掘，从中摸索崭新的商业模式[2]。

2 大数据驱动电信运营商转型

当前，移动互联网OTT业务的快速成长，给电信运营商的基础语音业务和短信业务带来了不小冲击，运营商缓慢增长的网络流量收入和网络建设成本之间不断增加的剪刀差，正不断侵蚀着运营商的利润。面对互联网公司的激烈竞争，运营商要如何做才能扭转逐步被“管道化”的趋势？

在日常网络运营中，运营商积累了大量用户数据，这些数据相比较互联网公司的用户数据有着明显的优势：一是用户实名，真实详细的个人基本信息，比如年龄、性别、工作单位、职位等；二是位置信息，运营商通过技术手段，能轻易获得通话者的地理位置，且精确度非常高；三是通话信息，包括话费、对方信息等。这些数据正是最具战略性的资产，使得运营商在利用大数据方面具有天然优势。但是，没有管理的数据就像埋藏在地下的矿产，价值无法体现。运营商当前由于没有全局性大数据管理体系，现存数据呈现出碎片、割裂和孤岛状的特点，难以深入应用。

对于大数据的应用已经成为一种必然趋势，其发展势头非常强劲。大数据驱动不仅是电信运营商增强业务能力和网络能力的抓手，更重要的是，大数据驱动能促使电信运营商切实学习和领会互联网的思维，真正实现

以用户为中心，多维度了解用户，实现数据化运营，借助大数据中蕴含的价值和动力将转型发展落到实处。

3 电信运营商大数据发展遇到的问题

我国电信运营商由于技术、数据系统限制，用户隐私和商业模式不明确等问题，目前大数据应用只处在探索阶段，主要遇到以下问题。1)系统分散建设，难以实现资源共享。经营分析、信令监测、上网日志留存等众多数据系统分专业建设，其中部分系统还分省建设，造成资源无法共享。2)数据处理种类多，单一技术难以实现。各大数据系统数据模型不统一，只具备结构化数据处理能力，无法支持非结构化、半结构化数据处理，无法满足互联网业务发展要求。3)如何避免隐私泄露。人们对于隐私问题越来越重视，数据公司掌握大量数据和数据制造者要求隐私权之间的矛盾，使得大数据应用变得困难。4)尚未确立商业运营模式。运营商掌握的数据很多，但是这些数据应该怎样应用、给谁用、应用收益是否可以抵消数据开发分析的成本，这一系列问题也让运营商非常困扰[3]。

4 电信运营商大数据策略

电信运营商大数据策略的核心在于从这些数据中挖掘价值，因关注点不同可区分为以下四种类型。第一，在市场层面，通过大数据分析用户行为，改进产品设计，并通过用户偏好分析，及时、准确且有针对性地开展营销与维系，不断改善用户体验，增加用户信息消费以及对运营商的黏度；第二，在网络层面，通过大数据分析网络流量、流向变化趋势，及时调整资源配置，同时还可以分析网络日志，进行全网优化，不断提升网络质量和网络利用率；第三，在企业经营层面，可以通过业务、资源、财务等各类数据的综合分析，快速准确地确定公司经营管理和市场竞争策略；第四，在业务创新层面，在保障用户隐私的前提下，可以对数据进行深度加工，对外提供数据分析服务，为企业创造新的价值。这样，大数据将助力运营商实现从网络服务提供商向信息服务提供商的转变[4]。

5 大数据分析处理应用平台

5.1 建设背景

以上海某电信运营商为例，2013年流量经营目标十分艰巨，要求月户均流量达到160M，流量经营收入达到23亿。面对如此艰巨任务，采用传统流量包营销模式已经不能满足市场经营分析和前端营销的需求。并且，面对每天以TB级速度增长的业务数据，该运营商在如何提升分析和管理能力方面遇到较大的瓶颈。另外，企业各类数据分散在各个系统中，缺乏集约化的数据管理和应用手段，导致了需求响应混乱无序，数据安全风险增大，数据无法有效进行关联从而形成数据资产。

针对以上问题，急需设计一套大数据分析处理应用平台，作为现有数据仓库系统的有益补充，形成企业大数据统一汇聚平台。

5.2 建设目标

大数据分析处理应用平台(以下简称大数据平台)的建设目标主要分为以下三点。1)通过对移动互联网上网行为数据、固网宽带的数据分析和快速分类，将有价值的用户行为信息进行再次整合后，推送到针对性营销平台和客户维系挽留平台，完成各个渠道的主动派单，实现快速营销。2)基于移动互联网流量营销功能，提升数据支撑能力，提供流量数据查询。3)提升EDA现有用户行为数据分析中时点统计分析能力和深化分析维度。

5.3 平台技术架构

5.3.1 平台技术方案要求

大数据平台面向企业内外部提供数据服务，要求系统必须具备高并发、实时动态数据获取和更新、海量数据的高效率存储和访问、高可扩展性和高可用性等特性，传统的数据处理技术已经无法很好应对新的挑战。本平台引入Hadoop等分布式解决方案以提升系统海量数据和高并发任务处理性能，提升系统可扩展性。平台支持离线批量处理、流式处理、在线处理、

交互式探索等多种计算框架，具备多租户模式支撑数据应用基础能力。

5.3.2 基础存储、计算平台技术方案分析

1) 基础平台技术架构分析及建议。针对海量数据的分析处理，目前业界主流解决方案有以下几种，如表1所示。①传统商业数据库方案：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＋传统关系型数据库的解决方案。②数据仓库一体机方案：基于一体机的BI集成化解决方案，一体机含大数据服务器、大数据存储、数据处理软件等。③基于X86开放平台的MPP海量数据方案：采用海量数据处理软件，基于X86服务器的大规模并行处理解决方案。④基于X86开放平台的Hadoop为代表的NoSQL分布式方案(通常具有如下特点：高性能、海量存储、高扩展性、高可用性)：采用Hadoop架构，基于X86服务器的大规模分布式解决方案。

表1 体系架构比较

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)	1.计算、查询效率一般 2.运行稳定性高 3.适合于OLTP系统 4.结构化数据处理	中	技术成熟、海量数据存在处理瓶颈
数据仓库一体机解决方案(一体机，如Teradata、Oracle Exadata等)	1.大数据量处理TB-PB级 2.并发查询、计算能力高 3.线性扩展能力强 4.适合OLAP系统 5.结构化数据处理	较高	技术成熟、要注意部分产品跨代兼容问题
MPP数据库(X86开放平台，EMC Greenplum、HP Vertica等)	1.大数据量处理PB级 2.并发查询、计算能力高 3.扩展能力非常强 4.适合OLAP系统 5.结构化数据处理 6.非结构化数据处理	较低	新技术，集成实施要求高
Hadoop为代表的NoSQL方案	1.大数据量处理PB级 2.并发查询、计算能力高 3.扩展能力非常强 4.适合查询型系统和数据关系简单的OLAP系统 5.结构化数据处理 6.非结构化数据处理	低	新技术，集成实施要求高

本平台采用以Hadoop为代表的分布式架构解决方案，原因如下。

①Hadoop等分布式架构通常用于非结构化/半结构化数据处理，已被广泛应用于多种大数据应用场景，成为业界大数据处理的最主流解决方案之一，具有可靠、高效、可伸缩的特点。另外，Hadoop等分布式架构可以解决系统的I/O问题，通过各服务器的列式数据的关联，并发生成数据，可解决海量数据的关联、入库、查询、共享等需要。

②Hadoop等分布式解决方案，已有成熟的组件适应各应用场景。如：Hadoop中可采用HDFS存储层存储、Hive方式关联入库、HBASE方式查询；具备可扩展性高的特点，并支持数据节点在线调整，扩展更多应用。

③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，避免传统方案数据库海量数据处理瓶颈。

④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。

⑤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。

⑥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。

⑦数据源采用非结构化/半结构化数据处理，有利于未来进行各种业务的扩展，有效提高数据的可用性。

2) 开源二次开发版本/商业版本对比分析及建议。基于Hadoop的架构特性及平台需求的多样化，业界有开源二次开发版本及商业版本的使用情况，对比如表2所示。综合考虑投资额、业界使用案例等因素，本平台采用开源二次开发版本。在实际建设中，需重点评估、考核支撑厂家的开发、支撑、服务能力，以保障平台未来的运营。

3) Hadoop版本对比分析及建议。目前主流的开源Hadoop版本分为Hadoop1.0、Hadoop2.0。Hadoop源代码可分为Apache版本和CDH版本，比较如表3所示。

①Hadoop1.0仍存在单点故障的问题；Hadoop2.0已消除单点故障(目前为2.2版本，Hadoop2.4版本即将开放)。②相比CDH版本，Apache版本多部门版本并行开发，更新速度较快，及时发布补丁和更新，因更新速度较快对运维能力要求较高。③开源amban管理模块采用开源模式，可以基于此进行二次开发。

表2 二次开发版本/商业版本对比

类别		开源二次开发版本	商用版本
支撑能力		针对开源版本的二次开发，需厂家具备较高的支撑能力，对核心组件具备研发能力	商用版本具备较强的支撑能力
技术		可针对较高版本Hadoop进行二次开发，更新速度较快	商用Hadoop普遍版本较低，更新速度较慢，Hadoop新生功能难以转化为生产力
性能		基于2.0以上版本进行二次开发，使用新的资源调度模型和算法，极大提高硬件资源利用率；且持续改进完善	对部分算法和代码进行改进，性能比未二次开发开源版本有小幅度提升
后期运维		可通过二次开发，实现与业务统一的监控、统计报表、告警管理；并可开发更多的运维管理功能	独立的监控、统计报表、告警管理，但cdh-manager管理模块为闭源，不支持二次开发
安全性		提供完善的安全解决方案，具备完善的权限管理；配置用户管理和防火墙安全策略，能满足企业对安全性的要求	加强安全性管理，增加文件加密等
稳定性		2.0以上版本，消除单点故障，满足企业级应用对高可靠性的要求	消除了单点故障，提高了可靠性
使用案例		互联网公司普遍采用开源版本。越来越多的运营商逐步走向互联网技术路线，接受开源项目	国内有使用案例，但普遍存在厂家绑定情况
投资	分析	免费，因涉及二次开发，需增加应用软件费用	按license计费，无需二次开发Hadoop，因此应用软件开发量较小
	投资额	主要为Hadoop基础平台二次开发费用	按节点收取license费用，部分厂家费用高昂
	扩容	功能已开发，无需扩容费用	系统扩容需新增license费用
发展		开放和标准化是未来软件发展的目标，开源软件模式在行业内占据着越来越重要的位置	完全受厂家的控制，系统升级和架构持续演进受限制

表3 Hadoop版本选型对比

性能	Hadoop1.0	Hadoop2.0
可靠性	存在单点故障	满足企业级应用对高可靠性的要求
安全性	无集中的用户管理系统，不支持安全控制	提供完善的安全解决方案，具备完善的权限管理；配置用户管理和防火墙安全策略，能满足企业对安全性的要求
易管理	ambari开源项目，图形化安装、配置、监控	更多开源和免费的Hadoop管理项目；如Cloudera Manager系统提供更多企业级管理功能
算法	粗粒度的资源调度算法	新的资源调度模型和算法，极大提高硬件资源的利用率；且持续改进和完善

综上所述，建议本平台采用基于Hadoop2.0以上开源版本进行二次开发的Hadoop版本。

4) Spark框架及支持建议。

①Spark与Hadoop的对比。Spark是UC Berkeley AMP lab所开源的类Hadoop MapReduce的通用并行计算框架，拥有Hadoop MapReduce所具有的优点；但不同于MapReduce的是，Spark的中间数据可以保存在内存中，对于迭代运算效率更高，更适合于迭代运算比较多的ML和DM运算。

Hadoop只提供了Map和Reduce两种操作类型，而Spark提供了很多种数据集操作类型，比如map、filter、flatMap、sample、groupByKey、reduceByKey、union、join、cogroup、mapValues、sort和partionBy等，同时还提供Count、collect、reduce、lookup和save等多种actions操作。这些多种多样的数据集操作类型，给上层应用的开发者提供了方便。各个处理节点之间的通信模型不再像Hadoop那样是唯一的Data Shuffle模式，用户可以命名、物化、控制中间结果的存储和分区等；因此，Spark编程模型比Hadoop更灵活[5]。

②Spark与Hadoop的结合。Spark可以直接对HDFS进行数据的读写，同样支持Spark on Yarn；Spark可以与MapReduce运行于同集群中，共享存储资源与计算；数据仓库Shark实现上借用Hive，几乎与Hive完全兼容。让Spark运行于Yarn上与Hadoop共用集群资源可以提高资源利用率。

③Spark支持建议。考虑到Spark生态系统的成熟现状及发展前景，建议本平台以Hadoop生态系统为主体，同时支持Spark计算框架，可采用Spark on Yarn架构，提升基础平台的统一性和灵活性。

5) 分布式K-V内存数据库。内存跟传统磁盘相比，具有更高的读写速度。在海量的、高并发的简单关系查询中，应用内存数据库将有效提升系统性能。目前各种类型的内存数据库已经得到了广泛的应用，如关系型内存数据库TimesTen、fastdb，Key-Value型的数据库等，往往作为前端数据库的角色出现，处理和存储短时间段内的实时数据，根本原因在于内存数据库容量有限，大量的业务应用不得不由后台的磁盘数据库负责。为解决传统集中式内存数据库的容量、并发、扩展性、

持久化等问题，目前可持久化的分布式内存数据库使用越来越多。

Key-Value内存数据库的主要特点就是具有极高的并发读写性能。以Redis为例，它是一个高性能的Key-Value数据库，同时支持磁盘数据的持久化。Redis使用内存提供主存储支持，而仅使用硬盘做持久性的存储。

Redis支持存储的Value类型，包括string(字符串)、list(链表)、set(集合)和zset(有序集合)等。这些数据类型都支持push/pop、add/remove及取交集、并集、差集和更丰富的操作，而且这些操作都是原子性的。在此基础上，Redis支持各种不同的方式排序。与memcached一样，为了保证效率，数据都是缓存在内存中。区别的是Redis会周期性把更新的数据写入磁盘或者把修改操作写入追加的记录文件，并且在此基础上实现了M/S同步。

综上所述，本平台建议引入K-V内存技术。

6) 分布式消息中间件。消息中间件是指支持与保障分布式应用程序之间同步/异步收发消息的中间件。消息中间件利用高效可靠的消息传递机制进行平台无关的数据交流，并基于数据通信来进行分布式系统的集成。通过提供消息传递和消息排队模型，它可以在分布式环境下扩展进程间的通信。

消息中间件适用于需要可靠数据传送的分布式环境。采用消息中间件机制的系统中，不同的对象之间通过传消息来激活对方的事件，完成相应的操作。发送者将消息发送给消息服务器，消息服务器将消息存放在若干队列中，在合适的时候再将消息转发给接收者。消息中间件能在不同平台之间通信，它常被用来屏蔽掉各种平台及协议之间的特性，实现应用程序之间的协同，其优点在于能够在客户和服务器之间提供同步和异步的连接，并且在任何时刻都可以将消息进行传送或者存储转发。

互联网等大型分布式解决方案中，往往采用分布式的消息中间件。大数据分析系统分布式解决方案中，broker、producer、consumer都为集群，消息路由对顺序和可靠性有极高要求。

综上所述，本平台建议引入分布式消息中间件。

5.3.3 平台技术架构

大数据分析处理应用平台采用总分架构，通过统一的数据中心来汇聚各类数据源的数据，并进行关联和整合。整个平台包括四个部分：数据采集网关、数据存储处理平台、数据应用平台、数据管控平台，如图1所示。

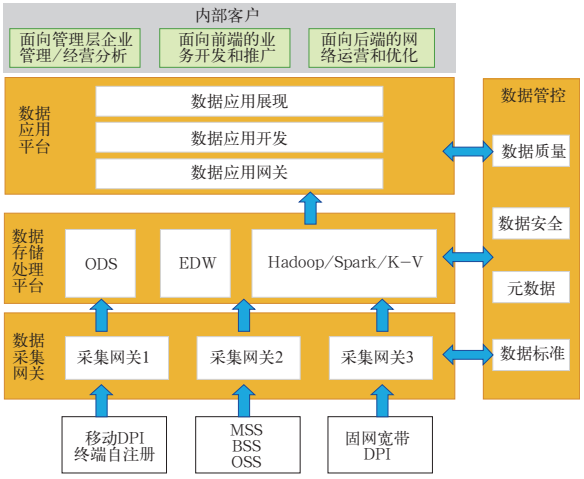

图1 大数据分析处理应用平台技术架构

1) 数据采集网关。负责数据的采集、清洗和安全传输等，采集范围包括移动DPI、固网宽带DPI数据、MSS/BSS/OSS等。其部署方式采用分布式前置部署，部署在数据采集节点。采集模式采用按专业建立采集通道，避免统一数据源由多方重复采集，进一步实现数据采集的标准化，提升数据时效性和传输效率。以移动DPI数据为例：利用现有DPI解析和清洗设备作为数据采集网关，当前不做安全控制，清洗策略按需固定，同时整合终端自注册数据和PCMD数据(Per Call Measurement Data)。

2) 数据存储处理平台。负责整合汇聚所有数据并进行关联，负责对外提供数据，如三户信息、套餐等IT相关数据、移动DPI和固网宽带DPI等；利用运营商原有ODS、EDW系统承担结构化数据存储和处理，建设新的Hadoop分布式平台负责海量话单、移动DPI和固网宽带DPI数据等非结构化和半结构化数据的存储和处理。

3) 数据应用平台。数据应用网关统一封装数据，

提供统一的数据共享接口，数据应用通过共享接口获取数据。在数据应用开发平台中，可以通过建立开发流程中的业务和技术组件，实现数据应用敏捷化和标准化开发，提高开发效率。

4) 数据管控平台。利用运营商现有数据运营管控各子系统，如数据质量稽核、元数据管理等功能，保证数据安全可用，数据运营稳定高效。平台的数据处理流程如图2所示。

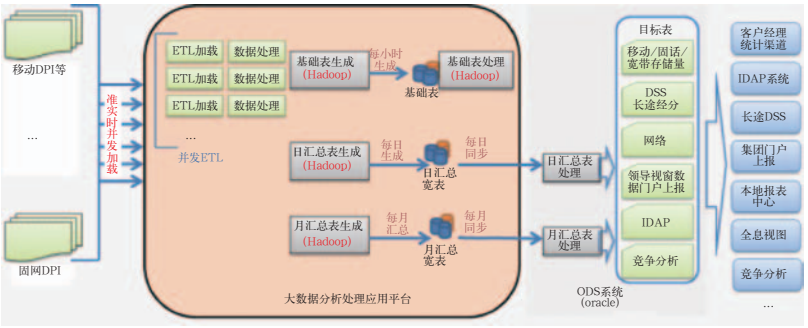

图2 大数据分析处理应用平台数据处理流程

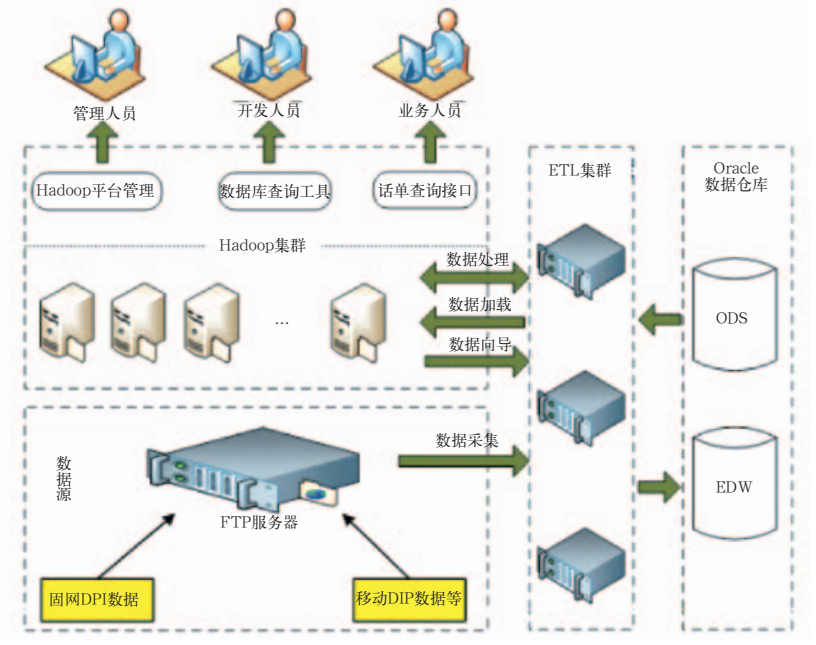

图3 大数据分析处理应用平台数据处理流程

①移动DPI、固网DPI等数据每5分钟内准实时加载。②由于运营商目前基于oracle的ODS分析报表展现体系比较完善，日汇总数据量平均为详单的1/20，因此将日汇总数据的分析结果倒回ODS系统，进行相关的多维分析和渠道展现。③大数据分析处理平台的准实时分析(小时级)直接在大数据平台上查询展现。平台的逻辑架构如图3所示。

5.4 主要功能

1) 固网宽带DPI数据分析功能和应用。包括上网行为总体分析、Top100网站访问排名、Top1000关键字排名、分类网站分析、用户偏好总体分析和竞争对手网站总体分析和竞争对手网站轨迹分析。

2) 移动互联网行为数据分析功能和应用。

①移动互联网流量分析。包括：移动用户上网流量分析、使用终端客户端上网分析、重点增值业务流量分析、区局3G流量跟踪分析、存量用户3G/4G推荐、低流量用户增值应用推荐、超量用户升档/加装包推荐、写字楼白领圈及高校圈升档/加装包推荐等。

②移动互联网流量营销。包括：流量使用情况及通话情况实时查询、套餐流量情况实时查询及统计、移动流量用户行为数据分析、优酷等视频客户端搜索指定关键词的用户清单、安卓市场等应用市场类软件搜索指定关键词的用户清单等。

③互联网分析应用营销派单。包括：重点增值业务营销派单、重点推荐手机软件营销派单、基于用户位置的营销派单等。

5.5 性能指标

1) 数据采集性能指标。数据采集处理频率为5分钟一次，高峰时段每频次需要采集的数据量约为1.5亿，故数据采集的性能指标为15 000/5/60=50万条/秒。

2) 数据处理性能指标。数据采

集有两个数据源：移动互联网上网行为数据(每天80亿条)和固网宽带DPI数据(每天100亿条)。

①移动互联网上网行为数据的处理性能指标。移动DPI一天的数据量约为80亿条，每天有4个汇总处理需求，每个汇总的时间要求为1小时(共需4小时)，性能指标为222万条/秒。

②固网宽带DPI数据的处理性能指标。固网DPI一天的数据量约为100亿条，总共1.9TB，每天1个汇总处理需求，汇总时间要求为1小时，性能指标为278万条/秒。

5.6 组网方案及软硬件配置

本大数据分析处理应用平台的网络拓扑如图4所示。

由图4可知，本平台硬件配置方面主要有：1台日志采集服务器、2台AAA Radius采集服务器、4台固网HTTP Get前置采集服务器、4台固网双向DPI前置采集服务器、11台固网双向DPI清洗服务器、11台固网HTTP Get清洗服务器、93台Hadoop数据节点服务器和2台Hadoop控制节点服务器(需要安装Hadoop、hive、hbase、Zookeeper、pig、ganglia、Spark、K-V)、16台Hadoop ETL节点服务器(需要安装sqoop、flume)、2台门户服务器、10台应用服务器、4台一级汇聚分流设备、1台二级汇聚分流设备、2台前置采集交换机及若干核心交换机、防火墙等。

本平台软件配置方面主要包括1套大数据分析处理应用软件、1套Hadoop分布式平台基础软件。

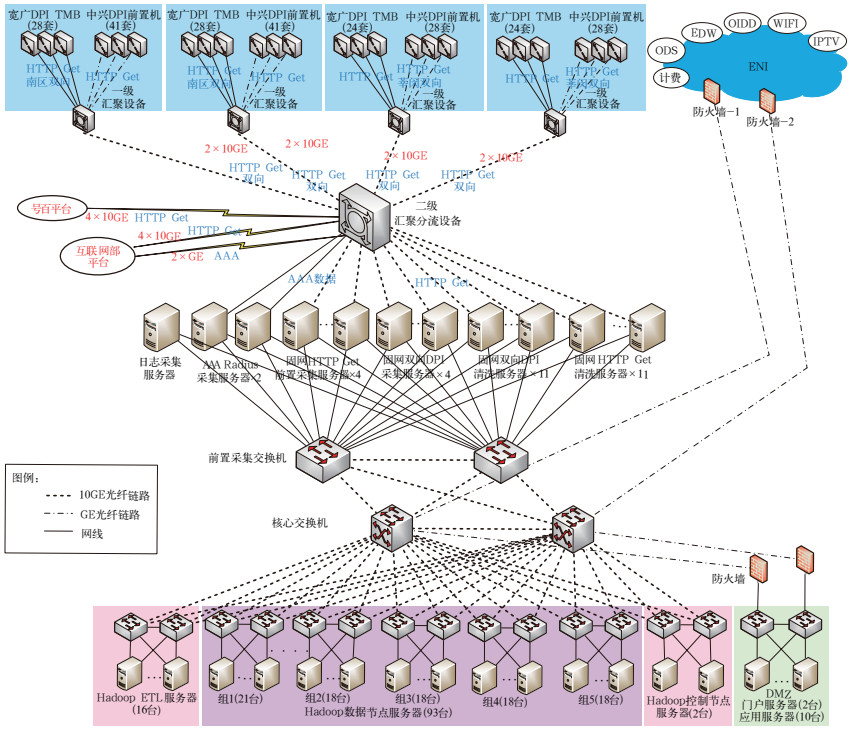

图4 大数据分析处理应用平台网络拓扑

6 总结

近年来伴随云计算和大数据的发展热潮，数据作为一种无形资产的价值正在日益得到社会广泛认可。面向大数据时代，运营商的及时转型成为必然，否则将有被互联网企业超越的可能性。电信运营商需要重视并建立大数据体系，掌握大数据技能，发掘大数据价值，对内可实现智慧运营，为用户提供精细化营销服务，对外可提供增值化业务，将数据提供给零售行业、金融业和保险业等，实现数据的二次营销，从而为自身的转型发展提供强劲的动力。

参考文献

[1] 冯明丽,陈志彬.基于电信运营商的大数据解决方案分析[J].通信与信息技术,2013(05):12-13

[2] 于艳华,宋美娜.大数据[J].中兴通讯技术,2013(03):57-58

[3] 顾基发.大数据要注意的一些问题[J].科技促进发展,2014(01):25-26

[4] 陈勇.大数据及其商业价值[J].通信与信息技术,2013(01):10-11

[5] 夏俊鸾,邵赛赛.Spark Streaming:大规模流式数据处理的新贵[J].程序员,2014(02):21-22

(下转83页)