专题:大数据技术与应用

电信运营商大数据发展策略探讨

黄勇军[1],冯　明[2],丁圣勇[1],樊勇兵[1]

(1.中国电信股份有限公司广东研究院 广州 510630;2. 中国电信集团公司 北京 100032)

摘　要:随着互联网业务和应用的迅猛发展以及移动互联网的爆炸式增长,电信运营商客户行为数据、网络运维数据、信令数据等海量数据的存储与分析日益成为电信运营商的重要挑战,大数据技术的出现与发展为电信运营商深挖数据提供了新的技术手段,同时也为其更好地服务客户提供了新的机遇。本文结合大数据的技术现状以及电信运营商的数据特点,分析了大数据技术在电信运营商的适用性,探讨了电信运营商应用大数据的策略,并提出了一种参考性的平台架构,以推动电信运营商对大数据技术的应用。

关键词:电信运营商;大数据;策略

doi: 10.3969/j.issn.1000-0801.2013.03.002

Big Data Development Strategy for Telecom Operators

Huang Yongjun[1], Feng Ming[2], Ding Shengyong[1], Fan Yongbing[1]

(1. Guangdong Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China;
2.China Telecom Corporation, Beijing 100032, China)

Abstract: With the rapid development of internet and dramatic growth of mobile services, the storage and process of massive data are becoming a big challenge for telecom operators. Big data technologies provide a new solution for the operators to mine massive data in depth as well as a new opportunity to provide better services for their customers. The applicability of big data technologies for operators were analyzed, based on the comprehensive analysis of big data technologies and telecom operators' data characteristics. The big data development strategy for telecom operators was discussed and a conceptual technical architecture was also proposed in order to push the application of big data technologies.

Key words: telecom operator, big data, strategy

1　引言

近年来,以海量数据处理为目标的大数据技术正成为新的研究热点。尽管没有严格的定义,但大数据的4V(volume——容量,value——价值,velocity——快速,variety——多样)特点以及主流的处理技术已经基本得到认可,使用传统技术在短时间内无法处理的任务或问题目前都可归为大数据问题,相应的处理技术被称为大数据技术。

大数据技术起源于互联网公司,最初主要用于解决海量非结构化网页数据的存储、分析以及检索等问题,在设计理念上采用经济的服务器构建超大规模的集群,以获得海量的数据存储和处理能力。与传统集群技术不同的是,在大数据解决方案下,尽管单台节点服务器的性能与可靠性不足以与高性能服务器媲美,但可以通过超大规模(上万台)集群以及冗余设计获得显著的成本和扩展优势。大数据技术的扩展性、先进性已被Google、微软、Yahoo、Facebook等顶级互联网公司所验证。

与此同时,随着电信运营商的全业务化运营以及3G推动下移动互联网业务的爆炸式增长,电信行业的数据类

型、数据规模、数据速度、数据价值在大数据的4个维度上得到显著体现。对于电信运营商来说,在不需要大量增加网络投资和运营成本的条件下,大数据体系极有可能成为未来企业新的价值增长点。但由于电信运营商在大数据的人才方面无明显优势,且内部系统复杂,大数据技术尚未在电信运营商中得到广泛应用并发挥价值,如何在电信行业中引入大数据技术并抓住大数据的机遇为客户提供更深入的服务,是当前一个急迫的问题,本文就电信行业的大数据应用策略展开探讨。

2 大数据技术框架

大数据技术的核心任务可分为两种:一种是基础的大数据存取,功能上类似于传统文件系统操作或数据库操作,但规模远超传统任务;另一种是数据挖掘分析,目的是从海量的数据中挖掘出有价值的信息。前者是较为简单的任务,是比较共性的需求,对应的大数据技术主要为分布式文件系统和分布式数据库;后者则需要通过在大数据计算平台上实现特定的算法才能完成相关任务,涉及的主要技术包括大数据计算平台和基于大数据计算平台的分布式数据挖掘技术。

(1)分布式文件系统技术

大数据时代的分布式文件系统利用大量普通服务器的存储能力,提供超大规模的文件存储能力,目前典型的技术是使用集中服务器维护数据分配信息。客户端对分布式文件系统进行存取操作时,首先通过集中服务器获得数据存取的节点以及相应分块位置,从而完成定位操作,一旦定位,数据存取就归结为普通的流读写。此外,为了提高可靠性,数据在写入时进行冗余复制,从而保证系统具有高度的可靠性。目前主流的分布式文件系统如Hadoop File System,多为基于Google开放的GFS技术实现。

(2)分布式数据库技术

分布式数据库技术用来实现海量数据的存取,这里的数据以记录形式存在,一般具有固定的属性,有别于流式的文件数据。分布式数据库的核心技术反映在CAP(consistency,availability,partition tolerance)定理中。CAP定理理论上证明了任何数据库都无法同时满足一致性、可用性、分区容忍性的要求。目前的分布式数据库放弃了传统数据库的一些特性,如事务操作,实现超大规模的数据读写能力,将分布式数据库的核心定位为<key,value>的快速存取问题。目前主流的分布式数据库平台有Hbase、MongoDB等。传统数据库强调严格的关系模型以及事务操作,对可用性、一致性要求很高,而分区容忍性较差,具体表现为实时性高、事务性强,但对非结构化数据的处理支持能力相对较弱,在容量的扩展性上也不如分布式数据库。

(3)分布式计算平台技术

分布式计算平台是为应用程序提供并行化的计算平台,能够将计算任务自动地加载在多台机器上并执行,将相应结果进行汇总。分布式计算平台不能支持为单机编写的普通程序,只能支持遵循其编程模式和规范的程序,即使用分布式计算平台的开发者必须根据分布式计算平台的特性自行设计任务分解方法,这也是使用分布式计算平台的主要困难所在。目前有两种典型的分布式计算平台,介绍如下。

- 实时流计算平台,支持实时流数据处理,开发者可定义每条数据的处理环节以及相应的处理方法,平台每接收到一条新数据就会自动调用不同的处理环节,以保证每条数据都被完整处理。这种平台通过将处理环节自动部署在不同的节点上,实现并行化的处理能力。典型的平台有Storm。
- 批量式计算平台,与实时流计算平台不同,其任务输入是已经存在的数据集合,执行任务时将这些数据集合分成若干块,每块启动一个任务进行处理并自动汇总结果。典型的平台有Hadoop MapReduce。

(4)分布式数据挖掘技术

分布式数据挖掘基于分布式计算平台实现数据挖掘算法,从而支持大规模的数据挖掘分析。将各种传统的数据挖掘算法(如聚类算法、分类算法)根据底层计算平台的要求进行并行化实现,必要时进行适当的简化以适应底层平台的要求。典型的分布式数据挖掘算法有迭代式*K*-means、基于Gibbs采样的LDA以及SVM等。

经过几年的发展,以上技术已被互联网巨头公司广泛使用,并且形成了一系列开源平台,如Hadoop、Storm、Hbase、MongoDB等。大数据技术的成本和扩展性优势已毋庸置疑,但由于大数据平台和产品最初多定位为满足互联网公司的自身需求,从可运营、可管理的角度看,尚不能完全满足运营级的产品要求,现有大数据平台往往需要深度的优化才能稳定可靠地运行。相比一些顶级互联网公司,电信运营商在这方面起步较晚,特别是在大数据平台技术

方面，还没有形成足够的积累。但另一方面，电信运营商在基础设施（如数据中心建设、用户网络行为分析、市场经营分析等）方面有较强的技术积累，这些积累为运营商快速应用大数据奠定了良好基础，包括提供大数据基础设施服务以及开展各种前向、后向合作运营大数据的机会。

3 电信运营商的大数据适用性

电信运营商的系统本质是为用户与用户、设备与设备、用户与设备之间提供通信信道，每天承载着海量信息，是互联网大数据的源头。电信运营商大体上掌握 3 类数据：第 1 类是支撑网络运营的设备状态及资源利用率数据，这类网络运维数据与用户无关，是纯粹的信道层面的数据，对网络优化扩容极其重要；第 2 类是与用户紧密相关的数据，具体又包括两部分，一是相对静态的体现用户身份的账号数据，伴随着用户业务的开通产生，另一种是实时的用户行为数据、用户通话的信令数据、用户网络访问日志等，是内容层面的数据，对经营分析极为重要；第 3类是增值服务类数据，如流媒体内容数据、视频监控数据、网页数据等。图 1 展示了前两类数据的来源、挖掘分析的服务对象及应用价值。

第 1 类数据以结构化为主，处理逻辑相对简单，局限于某个区域网络，数据量也相对较小。但在全网范围看，由于网络节点多，设备数量大，传统技术已很难实现长时段、全网级的统计分析。由于这类数据的结构化属性较强，统计方法相对简单，使用分布式文件系统和分布式数据库技术能够基本满足数据存储和基础分析的需求。

第 2 类数据具有典型的大数据 4V 特点，即规模大、变化速度快、价值高、类型复杂。在规模方面，国内主流运营商的用户数达到数亿规模，用户每天的网络行为日志无疑构成海量数据，并不亚于顶级互联网公司。在速度方面，时刻都在变化，以记录海量用户的实时行为。在类型方面，具有典型的多样性，首先体现在数据来源方面，数据可能来自宽带网络，也可能来自无线网络或 3G 网络；其次体现在结构方面，既包含结构化的用户账号数据，也包含半结构化的用户访问日志。在价值方面，蕴含着用户兴趣、位置、身份信息，无论是对电信企业自身还是对外部互联网企业，都具有无穷的价值。但由于这类数据涉及电信运营商的核心业务，很多数据处理任务对事务性、实时性、可靠性都有极高的要求，目前的大数据解决方案尚不能完全满足这些要求。因此对于这类数据，大数据初期适合定位于补充性的分析处理，如提供用户话单查询、基于用户网络访问日志的用户兴趣画像等。

第 3 类数据具有明显的非结构化特性，特别适合利用大数据技术处理。如流媒体和视频监控数据，可以使用分布式文件系统代替传统的存储系统（如 SAN），可以使用实时流计算平台进行编解码处理；再如互联网增值应用中的网页数据抓取、分析、索引，完全可以借鉴互联网公司广泛使用的大数据处理技术。

4 电信运营商的大数据应用策略

大数据技术在数据挖掘的广度、深度方面都带来了新的机遇，电信运营商应当把握大数据时代的契机，加强数据挖掘与分析工作，将特有的数据资源转化为资产与核心竞争力。但与此同时，电信运营商应当认识到，大数据技术和产品具有互联网化的特点，目前大数据技术没有成熟的

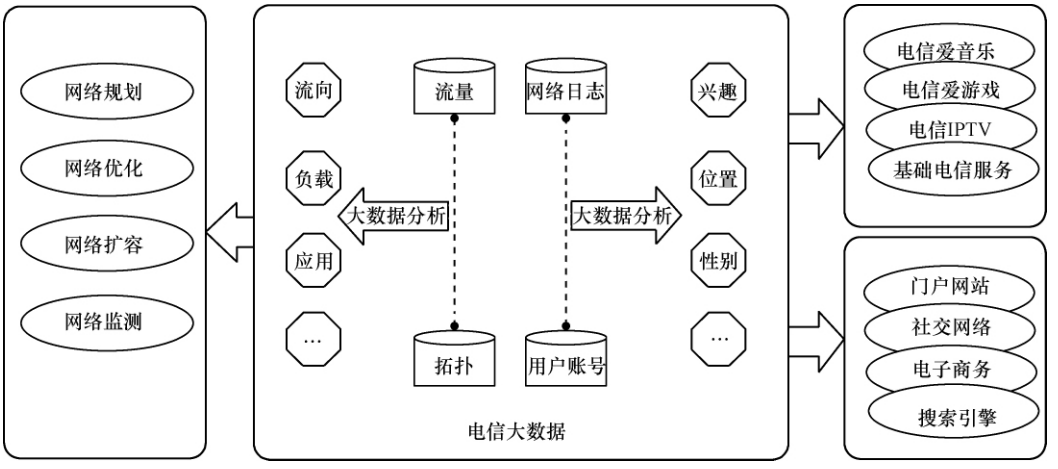

图 1 电信核心数据来源示意

可直接运营的产品,大数据的上下游产业链也远不如传统网络设备完善,这意味着大数据的应用是个长期渐进的过程,也是需要自主研发、运营、优化的过程。依据目前大数据的技术及产业链现状,从“人才培养、技术研发、平台建设、应用切入、业务运营”的角度,分析探讨电信行业应用大数据的策略。

(1)人才培养

无论在哪个行业,应用大数据都需要同时熟悉行业知识以及大数据分析或者大数据技术的综合人才。从业界观点来看,大数据改变的是从“样本分析”到“全量分析”的模式,分析方法是一个巨大的挑战,而企业需要的数据人才也大致包括产品和市场分析、安全和风险分析以及商业智能3大类。电信运营商的数据人才,一方面应是数据分析和研发人才,能够建立适应电信运营商的数据架构,提供有效的机器学习和数据挖掘分析模式的能力;另一方面应熟悉电信自身的业务,即电信行业的数据科学家。综合考虑大数据上下游产业链的不完善现状以及大数据技术对未来企业发展的重要作用,电信运营商应当加强大数据人才的储备,引入高层次大数据人才,并通过内部大数据应用快速培养人才。

(2)技术研发

大数据研发包括平台型研发和基于平台的应用型研发。从现状出发,电信运营商应该基础设施与应用并重,但首先以应用型研发为主,即能够首先用好大数据,与此同时适度进行平台型研发,以支撑大数据应用。在积累到一定经验后,加大平台型研发的投入,以逐步从对内服务转向对外运营。大数据的应用十分广泛,但完整部署需要较长的时间,应用切入的方式有利于兼顾近期运营和长远规划,而大数据基础设施也是一个逐步完善的过程,建议以自有研发力量为主建立核心研发团队,打造未来成为企业价值核心的大数据系统。

(3)平台建设

电信企业的各套系统基本上都需要大数据支持。每套系统独立建设大数据平台不仅浪费,且不具备相应人才。大数据在基础设施层面尽量实现共享,以发挥大数据规模集群的优势。结合电信运营商的系统及管理现状,可建设省级大数据中心及全国级大数据中心,省级大数据中心定位于满足省内各种应用和系统对大数据能力的需求,全国级大数据中心定位于满足全国性系统对大数据能力的需求。

(4)应用切入

任何一个新技术都很难一次性替换原有系统的技术。大数据技术的初期切入策略可定位为补充式切入,即重点实现传统技术难以实现的问题,如全网流量分析、用户行为画像、用户话单查询、“号百”餐饮搜索等。

(5)业务运营

在基础设施层面,考虑到行业对大数据需求的普及以及大数据技术对基础设施平台的规模和弹性要求,运营商可结合云数据中心提供云化的大数据基础设施服务,为大数据服务提供商或用户提供高质量的、专业的弹性基础设施平台,并在平台上部署基础的大数据平台软件和分析系统,同时嵌入特有的电信能力,发挥运营商的基础设施服务优势。在数据挖掘分析层面,大数据技术初期以优先服务内部系统为主,从解决内部系统的实际需求出发,积累大数据的开发、运营经验,在充分掌握大数据技术的基础上逐步对外提供大数据分析服务,积极寻求与行业或者企业开展大数据运营的合作。同时要注意的是,电信行业大数据的运营应该充分发挥运营商已有的企业数据仓库(EDW)系统体系,用好电信运营商传统的数据体系,结合新的海量用户行为数据,创造更大的数据应用价值。

在具体技术架构方面,电信运营商大数据平台可以依托开源项目,采用分层、模块化思想对主要平台元素进行设计,各层相对独立,通过标准接口向外部应用系统开放。参考技术架构如图2所示。

其中,最底层为硬件平台层,将PC服务器集群、存储、网络等基础设施资源组合在一起,形成大规模的计算机集群,供上层应用系统使用。

第2层是分布式平台层,在物理平台的基础上部署分布式文件系统、分布式数据库、缓存服务、任务分解、资源调度等一系列分布式软件,把多台独立的PC服务器组合成具有超大规模计算和存储能力的系统。分布式平台屏蔽了分布式系统任务分解、资源调配等复杂的底层工作,简化了上层分布式应用的开发流程。

第3层是基础能力层,将基于数据分析的一系列公共基础服务抽象成功能模块,开放给上层系统和应用开发者。对外提供包括数据仓库查询分析、数据挖掘、统计分析等在内的基础数据分析功能。

第4层是服务能力层,在基础能力层上形成搜索引擎、位置信息、内容分发等功能更为完善的服务。这些服务独立

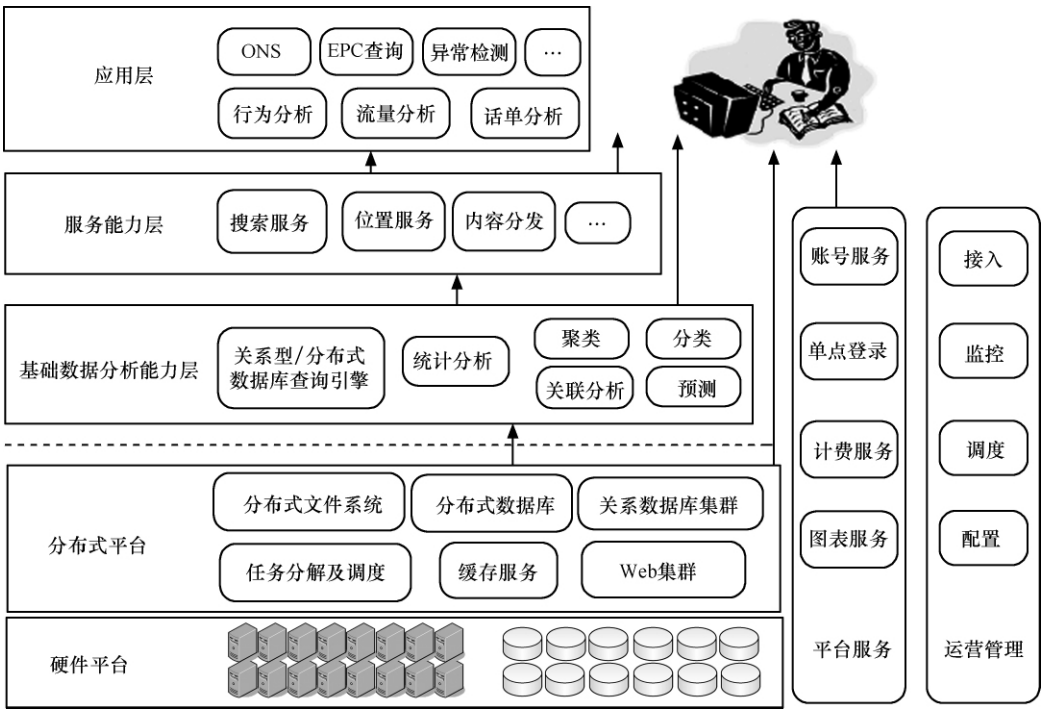

图 2 电信大数据平台参考架构

于应用系统的业务逻辑,可作为应用程序的能力补充。

第 5 层是应用层,即需要使用大数据服务的一系列电信应用系统,如大规模用户行为分析、全网流量分析等,通过调用平台的大数据服务接口快速实现大数据能力。

此外,为满足大数据平台的可运营、可管理要求,平台需实现运营管理能力,为应用系统提供访问控制、资源分配等一系列管理服务。

5 结束语

结合大数据的技术框架和电信运营商的数据特点,对大数据适用性进行了分析,并在此基础上给出电信运营商大数据发展的策略建议,旨在为电信运营商应用大数据技术提供参考。云计算服务模式的出现,拉近了网络运营商、企业、用户之间的距离,大数据的普及将使得三者之间形成更为紧密的依赖关系,为电信运营商更好地服务客户提供新的机遇。

参考文献

1 White Tom.Hadoop: the Definitive Guide. O'Reilly Media, p.3. ISBN978-1-4493-3877-0, 2012

2 Nancy Lynch, Seth Gilbert. Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services. ACM SIGACT News, 2002, 33(2):51~59

3 Ghemawat, Sanjay, Howard Gobioff, *et al*. The Google file system. ACM SIGOPS Operating Systems Review, 2003, 37(5)

[作者简介]

黄勇军,男,硕士,中国电信股份有限公司广东研究院高级工程师、副院长兼云计算研究所所长,长期从事通信网络发展规划、新技术应用研究和投资管理工作,曾从事电信市场经营管理工作,主要研究方向为宽带网络技术、云计算。

冯明,男,中国电信集团公司教授级高级工程师,主要研究方向为数据网络、云计算及物联网等。

丁圣勇,男,中国电信股份有限公司广东研究院工程师,长期从事互联网技术的应用研发工作,主要研究方向为网络大数据挖掘分析。

樊勇兵,男,博士,中国数据中心产业发展联盟云计算服务专家委员会第一批专家委员,现就职于中国电信股份有限公司广东研究院,主要研究方向为云计算、IDC 等。

(收稿日期:2013-03-05)