



运营商大数据技术应用研究

陈涛¹, 鲁萌², 陈彦名¹

(1. 中国移动通信集团设计院有限公司, 北京 100080;

2. 中国信息通信研究院, 北京 100191)

摘要: 大数据技术的快速发展为运营商的传统业务(如网络规划和优化)提供了快速的海量数据处理方案, 也为运营商开拓新业务(如征信、个性化推荐、互联网业务)奠定了数据基础, 对行业的发展具有重要意义。研究了大数据处理的相关技术, 介绍了运营商在业务中应用大数据技术的工程实践经验和待解决的关键问题, 为相关研究人员提供参考。

关键词: 大数据处理; 网络规划; 网络优化

中图分类号: TP311.13

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2017002

Research on operators' big data technologies and applications

CHEN Tao¹, LU Meng², CHEN Yanming¹

1. China Mobile Group Design Institute Co., Ltd., Beijing 100080, China

2. China Academy of Information and Communication Technology, Beijing 100191, China

Abstract: The rapid development of big data technology provides fast mass data processing method for operators' traditional businesses, such as network planning and optimization, but also laid the data foundation for new businesses, such as credit, personalized recommendations, internet services. It is important to the development of the industry. The big data processing cutting-edge technology was studied, relative engineering experience and key issues were introduced, reference for other researcher was provided.

Key words: big data processing, network planning, network optimization

1 引言

中国移动 2016 年上半年业绩报告显示, 流量收入超过传统业务, 成为最主要的收入来源。在实施运营成本总量控制时, 传统语音和短信业务市场逐渐饱和或萎缩, 而手机上网流量迅猛增长。如何利用大数据技术进行精细化经营、提升网络质量、开拓新业务成为运营商面临的

主要挑战。

大数据指数据的大小和复杂性无法通过现有常用的软件工具, 以合理的成本在可接受的时间内进行捕获、管理和处理的数据集。如搜索引擎、电子商务、社交网络等互联网公司或天文学、高能物理、生物学等科学研究领域都会产生 TB(10^{12} byte)到 EB(10^{18} byte)级的数据, 并且要求在短时间内完成处理, 相关业务通常要求 IT 系统在秒级

或分钟级别给予计算反馈,长时间的批处理计算变得越来越难以容忍。这样大规模的数据采集、存储、基于业务模型的统计计算要求已经超出了基于传统关系型数据库 IT 系统的处理能力,从而催生了以流式计算(Storm)、Google 文件系统(GFS)^[1]、内存计算(Spark)为代表的大数据传输、存储和处理技术。

运营商在支撑电信业务的同时,也产生了大量的经营分析数据、网络性能和资源数据、测量报告和信令数据等。在大数据处理技术出现以前,这些数据大多数都没有得到充分的积累和利用。网络规划和优化工作通常需要手工搜集经营数据,或依靠路测系统、客户投诉系统获得最初的故障信息,完成日常的网络建设和维护工作。随着网络流量的增长和大数据处理技术的成熟,运营商迫切需要对所拥有的数据进行积累和分析,提高网络规划和优化的能力。同时挖掘网络数据分析的应用潜力,发挥运营商的资源整合能力,在基于内容的服务领域取得竞争优势。本文对大数据处理技术和在运营商数据处理领域的应用进行了研究,为相关研究人员提供参考。

2 大数据处理关键技术

大数据处理任务要解决大数据的采集、存储、统计和模型计算带来的挑战。比如,在典型网络优化大数据应用中,单个数据表大小平均在 130 GB 左右(1:5 数据压缩配置),最大单表超过 500 GB,数据的增加、更新和查找都需要利用分布式存储系统来执行。每天原始数据的采集量在 PB 量级左右,入库数据在几亿量级,而集群 I/O 峰值在每秒几万量级,必须采用流式计算和内存数据库进行缓冲和优化。复杂统计和模型的计算往往要求对原始数据进行多阶段汇总计算,如果将中间计算结果存回磁盘很难满足计算时间的要求,也需要借助内存计算技术完成。因此,运营商大数据处理系统是在上述计算技术快速发展的基础上实现的。

2.1 分布式存储系统对业务系统的支撑

运营商大数据通常存储在 MPP (massive parallel processing) 数据库(如 Greenplum、GBase 数据库)或基于 GFS 和 Big Table 原理^[2]实现的开源组件 Hadoop 文件系统和 HBase 数据库中。这两种分布式存储系统都是将运算分布到集群各个节点中独立运算后进行结果合并,但是在运营商大数据处理系统中使用的情景有所不同。MPP 数据库在数据并发写入速度(特别是单行字段较多时,中小 MPP

数据库集群只能完成每秒几千次写入),集群最大可扩展节点、数据存储总量上都比 Hadoop/HBase 小。但是 MPP 数据库支持 SQL 语言进行数据查询且查询速度很快,构建数据分析业务系统比较方便。所以,运营商大数据处理系统可以使用 Hadoop/HBase 来存储原始日志数据文件或原始测量数值。而使用 MPP 数据库存储加工汇总后的经营分析数据或重要网络指标汇总数据。

Google 的 Spanner/F1^[3,4]系统示范了一个分布式关系型数据库。它改进了 Big Table 存储架构,并且支持事务性操作和全球分布的 OLTP 和 OLAP 能力,实际上融合了上述 MPP 和 Hadoop/HBase 的优势,开源社区正努力提供稳定的开源实现供运营商大数据工程选用,是取代 MPP/NoSQL 数据库混搭的一个重要方向。

2.2 内存计算技术的发展

大数据数据处理时间通常面临如下挑战:当前数据处理必须在下一阶段采集的数据到来之前完成或同步完成;必须充分利用集群的 I/O 资源完成大数据读写,又不能超过 I/O 负荷,造成数据的丢失;满足实时业务的计算反馈。

内存计算^[5]利用分布式内存通过改进存储模型和并行编程模型,将大数据尽量保留在内存中进行缓存和处理,从而尽量避免 I/O 操作的一种新型的以数据为中心的并行计算模式。内存计算技术对大数据处理提供了以下助力。

- 大数据处理系统可以利用分布式内存数据库,如 Redis 集群缓存数据处理中间结果或常用配置参数,只在必要时才写回硬盘,从而减少磁盘 I/O 操作,极大地缩短大数据的汇总和模型计算时间;也可以借助 memcached 缓存热点内容,加快对用户的反馈。
- MapReduce 计算模型可以利用廉价的集群构建高性能的数据处理系统。但是在对时延和吞吐量要求较高的应用中,HaLoop^[6]、Spark 等在内存迭代处理数据的通用系统更有优势。比如利用 Spark 处理数据时,只需要一次从文件系统中读取数据,Spark 将中间结果缓存在内存中,直接用于下一阶段的算子操作,可以快速完成计算任务。
- 运营商日志和测量数据的产生处理过程呈现明显的流特征,且并发采集量较大。如果使用通常的并发采集技术,20 个并发进程向集群写入数据就可能造成中小数据库集群 I/O 拥塞而丢失



原始数据,实际数据处理系统数据量通常需要 200 个采集进程同时工作。利用 Storm 或 Spark Streaming 技术配合 Kafka 等分布式消息系统可以缓冲采集数据,实时汇总数据指标的任务。

2.3 机器学习算法的应用

机器学习理论在互联网推荐系统、社交网络的挖掘、机器翻译领域取得了重要成果。事实上,运营商在网络规划、建设和优化的过程中积累了大量的运维经验。在小数据集上利用聚类、分类、回归分析等方法十分常见,但是如何在大数据上利用机器学习方法进行数据规律的挖掘并指导工程实践还是一个挑战。主要困难是电信领域专家经验的提取、算法并行化、工程实现等问题。通过在运营商大数据上使用机器学习算法可以为运营商充分挖掘网络数据应用潜力、开拓互联网业务提供基础的数据支撑。其应用前景十分广阔。

2.4 大数据技术对运营商数据处理系统架构的改进

表 1 对支撑运营商规划和优化等传统业务所需要分析的主要数据量进行了估计,利用上述大数据处理关键技

表 1 运营商主要数据类型与数据量估计

数据类型	数据量	处理时间要求与主要处理技术
经营分析数据	TB 级	天或月度汇总处理,传统 ETL 和分析技术
网络性能与资源数据	TB 级	分钟级别汇总处理,流式处理技术
MR 测量数据	PB 级	分钟级别汇总处理,流式处理技术
日志类数据	TB-PB 级	分钟级别汇总处理,MapReduce 批处理技术
信令数据	ZB 级	分钟级别汇总处理,MapReduce 批处理技术

术改进后的典型的运营商数据分析平台,如图 1 所示。

传统数据分析平台,如经营分析平台或网管支撑平台主要以关系型数据库为主构建,随着所需要的数据量增加,在数据存取速度和处理速度上均面临巨大压力,主要通过周期性的扩容来应对计算量的增长,通常集群数量很难超过 100 台,不能满足需求。引入新计算组件后(如图 1 阴影部分),实时处理要求较高的采集数据可以将解码后

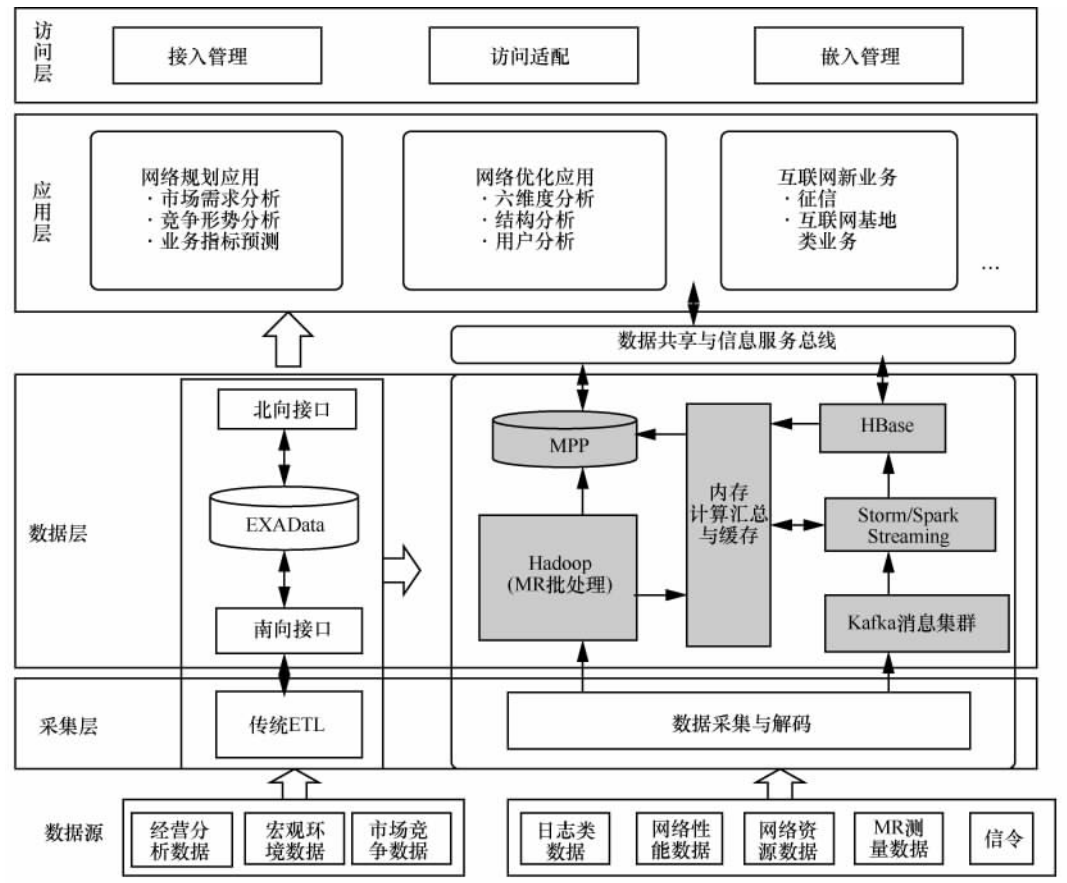


图 1 运营商数据处理系统的架构演进

的数据先投递到 Kafka 分布式消息集群中,利用流式计算或内存计算进行小时或天粒度的基础数据汇总操作,可以将以前需要三四个小时才能计算结束的计算任务缩短到 15 min 以内,处理时间大大缩短。利用 Hadoop 集群进行 MapReduce 批处理计算,可以将非结构化或半结构化的日志数据,如重要网站访问日志、市场一线人员的工作日志、爬取的互联网数据等纳入基础分析数据中,为传统业务带来了新的支撑手段。

3 运营商大数据技术应用

3.1 大数据技术对运营商传统业务的支撑

3.1.1 大数据技术在网络规划业务中的应用

传统上,运营商规划业务每年进行一次,本年度结束前完成下一年度的业务发展规划和网络建设规划。传统的规划业务流程如图 2 所示。业务发展指标确定后,分解到地市公司。各地网络建设根据业务指标要求进行工程实施。但是随着市场竞争的加剧和经营成本的约束,需要进行精细化的规划工作。比如一个地市的市场潜力有多大,如何合理地制定年度业务目标;如何对集团客户潜在价值进行评估,在市场推广成本一定的情况下,指导市场人员优先发展最有价值的客户;投资预算确定的情况下,如何将有限的建设资金用于新建和扩建最能带来价值的通信基站;如何根据竞争对手的情况,按月调整投资规划适应市场竞争等。

传统的规划工作主要是提取过往公司经营业绩指标等数据进行小数据集的业务分析来完成。大数据处理技术为改进规划工作提供的可能。政策经济要素(如 GDP、

CPI)、社会环境要素(如人口规模消费、家庭户数)、自然地理要素(如城区面积、人口密度)、消费行为要素(如人均收入、人均电信消费支出)和一线片区经理的工作日志、竞争性要素(如竞争公司的基站位置、广告推广的趋势分析)、基站价值分析数据等多维数据都可以参与到规划运算中。通过文本大数据处理技术和 Hadoop/MR 处理的非结构化数据,如一线市场经理工作日志或从互联网抓取的竞争对手广告信息,可以及时了解竞争公司的竞争策略,迅速调整规划方案。

这种依赖于大数据处理技术对传统电信规划业务的改进主要体现在更多的非结构化数据和公司外经营环境数据可以被采集和运算,规划业务可以精细化到高价值的楼宇和高价值的基站粒度,为节约投资,低成本发展用户提供了数据支撑。如何在大数据上构建多维数据模型,筛选出影响电信业务指标的关键环境因素,如何通过降维方法,加快数据处理仍然是行业探索的重点。

3.1.2 大数据技术在网络优化业务中的应用

传统的网络优化工作主要依靠网络性能监控系统、路测系统和用户投诉系统来发现网络问题,优化成本较高,且随着网络数据规模日益增加,数据处理速度不能满足网络问题快速处理的要求。当用户网络感知出现恶化时,运营商很难及时发现。

基于大数据的网络优化平台主要通过流处理技术实时采集分析性能、资源、MR 测量数据和信令数据,利用 Spark 等内存计算模式进行关键指标的汇总,通过固化成熟的网络优化模型和开发自动化参数优化工具完成传统

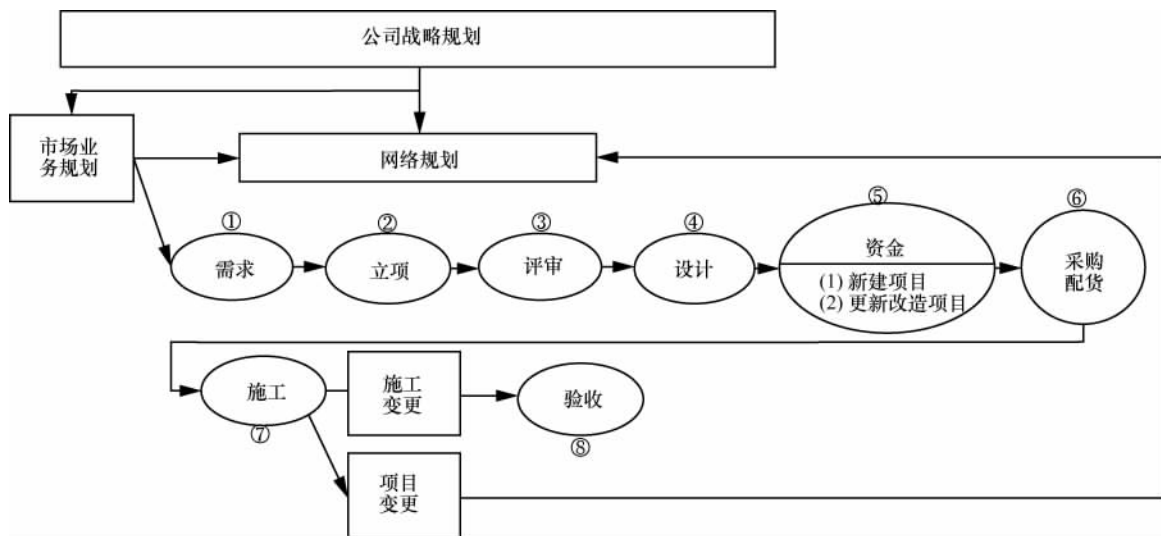


图 2 运营商传统规划业务流程



的网络优化目标。指标汇总和网络问题发现的速度可以由小时级提高到分钟级,可以实现实时用户网络感知指标的监控与优化。目前大数据处理主要尝试用于支持重大节日和事件的通信保障,实施虚拟业务拨测,进行用户网络感知管理和故障快速定界,自动发现并选点改善 4G 室内弱覆盖等问题,取得了良好的效果。

制约大数据处理在传统网络优化领域的应用的主要问题是数据的完整性和规范性需要进一步提高,用户定位算法与多数据源的关联分析技术准确性有待提高。此外,固化网络优化专家的优化经验,通过优化模型,自动发现网络优化问题,完成溯源、定界、派单也是网络优化自动化的研究方向。

3.2 大数据技术对运营商新业务开展的支撑

运营商拥有自营的互联网基地业务和宽带业务,从原始数据中分析得到的用户业务偏好有助于帮助业务的营销推广。如通过信令、经分、网络和终端、位置数据,结合用户的基本性别、年龄等信息,可以将用户阅读和消遣偏好进行分类,通过协同推荐等方法给用户推送图书、音乐、视频和宽带产品,从而提高产品的交易成功率。实践表明,利用位置数据支撑精准宽带业务营销通常可以将新增交易提升 10%~20%。

在进行必要的用户隐私处理后,用户的位置信息、通话通信记录等也可以包装为征信^[7]、旅游、广告等行业信息产品对外提供服务。如为互联网金融公司提供个人征信信息,或为旅游管理部门提供实时客流监控、客户归宿地分析、景区客户轨迹分析等。运营商自有的网络信息有可能不能完全满足各个行业的数据需求,还需要与电商、搜索和社交网络的用户数据结合。在运营商难以整合各方数据资源的情况下,比较好的业务开展模式是借鉴气象机构的数据服务方式,为各行业提供应用统计排名、用户偏好等“特供”分析数据。

4 结束语

大数据处理技术为运营商传统规划和优化业务、互联网新业务的开展提供了新的手段,有利于降低成本、提升用户满意度,并可以对外输出信息服务,带来新的收入增长。但是,运营商大数据的可靠性、规范性尚需完善,传统数据分析模型在大数据上应用也需要进一步验证。此外,

大数据商业模式选择也需要进一步探索,通过运营商的连接优势整合行业数据独立提供数据服务或者对外提供“特供”数据服务都是值得探索的新方向。

参考文献:

- [1] GARCIA H, LUDU A. The Google file system [J]. ACM Sigops Operating Systems Review, 2003, 37(5): 29-43.
- [2] CHANG F, DEAN J, GHEMAWAT S, et al. Bigtable: a distributed storage system for structured data [C]//Unix Symposium on Operating Systems Design and Implementation, November 6-8, 2006, Seattle, USA. New Jersey: IEEE Press, 2006: 15.
- [3] CORBETT J C, DEAN J, EPSTEIN N, et al. Spanner: Google's globally-distributed database[J]. ACM Transactions on Computer Systems, 2013, 31(3): 251-264.
- [4] SHUTE J, VINGRALEK R, SAMWEL B, et al. F1: a distributed SQL database that scales [J]. VLDB Endowment, 2013, 6(11): 1068-1079.
- [5] 罗乐, 刘轶, 钱德沛. 内存计算技术研究综述 [J]. 软件学报, 2016, 27(8): 2147-2167.
- LUO L, LIU Y, QIAN D P. Survey on in-memory computing technology[J]. Journal of Software, 2016, 27(8): 2147-2167.
- [6] BU Y, HOWE B, BALAZINSKA M, et al. HaLoop: efficient iterative data processing on large clusters[J]. VLDB Endowment, 2010, 3(1-2): 285-296.
- [7] 陈一听, 谷静宜, 王蕾, 等. 基于中国联通大数据的个人征信评估方法[J]. 信息通信技术, 2015(6): 36-41.
- CHEN Y X, GU J Y, WANG L, et al. Personal credit rating based on big data of China Unicom [J]. Information and Communications Technologies, 2015(6): 36-41.

[作者简介]



陈涛(1976-),男,博士,中国移动通信集团设计院有限公司高级工程师,CCF 会员,主要研究方向为大数据处理技术、信息安全技术。

鲁萌,现就职于中国信息通信研究院,主要研究方向为大数据处理技术。

陈彦名(1981-),女,中国移动通信集团设计院有限公司高级工程师,主要研究方向为平台级云软件架构、大数据处理技术、数据网管接口测试标准制订。