



大数据处理对电子商务的影响研究^{*}

陈云海¹, 黄兰秋²

(1. 中国电信股份有限公司广东研究院 广州 510630; 2. 华南师范大学经济与管理学院 广州 510006)

摘要: 鉴于传统的 IT 基础设施难以对电子商务系统中日益增长的数据进行有效管理及利用, 本文在云计算的时代背景下, 对大数据处理技术进行分析阐述, 并在此基础上, 探讨了大数据处理对电子商务发展的影响。

关键词: 大数据; 电子商务; 云计算

doi: 10.3969/j.issn.1000-0801.2013.03.004

Study on the Impact of Big Data Processing in Electronic Commerce

Chen Yunhai¹, Huang Lanqiu²

(1. Guangdong Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China;

2. School of Economic and Management, South China Normal University, Guangzhou 510006, China)

Abstract: It is difficult for traditional IT infrastructure to manage and use the growing massive data in electronic commerce effectively. In view of it, the impact of big data processing on the development of electronic commerce was discussed based on discussing the big data.

Key words: big data, electronic commerce, cloud computing

1 引言

随着“巨无霸电子商务(以下简称电商)”的出现,并访问量级大大增加,数据呈爆炸性增长,数据类型多样化,淘宝和天猫在 2012 年 11 月 11 日“双十一”的总交易额高达 191 亿元,共吸引了 2.13 亿名独立用户访问。春运时期,铁道部售票网站日点击量在 5 亿次以上,伴随着业务量的剧增,电商平台每秒钟都在生成海量的非结构化、半结构化数据(如文本、应用、位置信息、图片、音乐、视频等)。高并发和大数据量访问成为当前电商系统发展的瓶颈,根据调查,46%的人认为处理大量数据是其最大的挑战^[1]。这些大数据量往往以 PB(1 000 TB)、EB(106 TB)或 ZB(10⁹ TB)为计量单位,称为大数据。传统的分布式计算、

并行计算和数据库集群等技术由于技术上的局限性,长期存在扩展性低、成本高、硬件复杂等问题,电商的大数据处理客观上要求引入具备高容错性、高扩展性、低成本、易用性的新的架构和大数据处理技术。

2 大数据时代电商 IT 基础设施的变革

后互联网时代,PC 服务器成为电商企业最广泛使用的 IT 基础设施,随着业务量的发展,电商企业大量使用了 PC 服务器集群技术和小型机技术,用于满足日益增长的在线业务交易量。同时,企业必须投入大量人/财/物等资源对 IT 基础设施进行常规性的维护、升级、扩容、更新。

2.1 电子商务中数据仓库系统的发展

当前,电子商务的数据大规模存在且持续增长,表现

^{*} 中国民航信息技术科研基地开放课题基金资助项目(No.CAAC-ITRB-201206))



为大规模、分布式、异构性,使得电商数据仓库系统发生了显著变化:数据量由 TB 级升至 PB 级,并仍在持续爆炸式增长。有关调查显示,2015 年最大数据仓库中的数据量将逼近 100 PB,其增长速度远超摩尔定律;分析需求由常规分析转向深度分析;硬件平台由高端服务器转向由中低端硬件构成的大规模集群平台,并行数据库的规模增大,成本急剧上升^[2]。电商数据分析的两大趋势和挑战是:数据量的膨胀;数据深度分析需求的增长^[3]。

目前仅具备传统数据处理技术和信息分析能力的数据库和 BI 工具难以完成 PB 级大数据的数据管理和分析工作。

2.2 大数据的 IT 基础云计算架构

云计算是新型分布式网络计算架构,特别适合向各种网络应用提供计算、存储、网络、软件等在线服务,NIST 认为这种架构具有 5 个关键功能、3 种服务模式和 4 种部署方式^[4],如图 1^[4]所示。

云计算的特征主要有:按需自助服务(on-demand self-service)、泛在网络访问(broad network access)、虚拟池化的资源(resource pooling)、快速可伸缩性(rapid elasticity)、可度量的服务(measured service)^[5],这些能力对当前电商的基础 IT 设施来说,都是必须要满足的要求。

3 大数据处理模式

大数据处理模式从传统的数据库集群演进到云计算 MapReduce 大规模并行处理架构,实现任务的分解处理和结果合并,从而实现对可处理数据规模的无限扩展,大数

据处理技术已被认为是继云计算、物联网之后 IT 产业又一次颠覆性的技术变革。

3.1 数据库集群模式

集群是指通过协同工作方式运行同一套应用程序,针对客户端及应用程序提供单一系统映像,使用特定的连接方式,将硬件设备结合起来,构成的松散耦合的计算节点集合,具备以下优势:性能提升、扩展性提升、可靠性提升。数据库集群是将集群技术引入数据库^[6]。

数据库集群在技术上具备一定的局限性,具体介绍如下:

- 采用 PC 服务器作为功能节点,系统线缆众多,硬件复杂度过高,实施架设难度较大,可扩展性受限;
- 高速互联设备必须通过主机的 PCI 插槽与主机相连,而 PCI 的传输速率无法满足并行数据库集群节点间的数据通信要求;
- 数据库安全性和数据集可扩展性提升空间极小,全面提升速度、数据同步、安全保证、可扩展性 4 个技术指标是一大难题;
- 随着设备量的增加和应用的复杂化,需要迅速追加投入以解决所引起的兼容性和可靠性等各类问题;
- 并行数据库主要采用 shared-nothing 结构,在扩展性、容错性、成本、对异构环境的支持能力等方面有所欠缺并相互影响,因此扩展性非常有限,目前尚未有数千节点规模的应用案例^[2,7]。

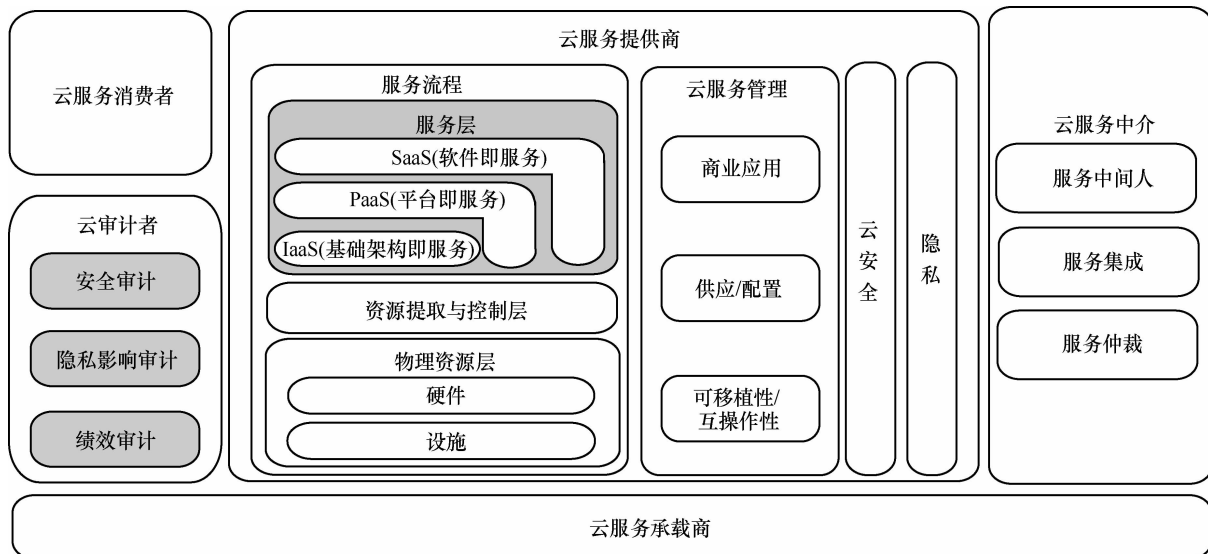


图 1 NIST 云计算参考模型

3.2 MapReduce 框架

云计算架构由大规模低端服务器组成服务器集群,提供海量存储空间和大规模数据的处理能力,具备可靠性、扩展性以及高可用性,因此中国科学院计算所、中国移动、百度和淘宝、网易等电子商务平台都使用主流云计算平台 Hadoop 架构进行 Web 搜索、大数据分析等^[2,8,9]。MapReduce 框架包括:分布式文件系统 (HDFS)、并行编程模型 MapReduce、并行执行引擎。

从 HDFS 角度来看,Hadoop 的节点由存储并提供定位块服务的数据节点(data node)和管理分布式文件系统命名空间的命名节点(name node)组成,HDFS 主从结构的体系架构设计大大简化了分布式系统架构。其文件系统的设计特点是:元数据集中管理、数据块(64 MB)分散存储以保证数据的安全性,数据复制(每份数据至少 3 个备份)实现高度容错^[3]。传统的数据库系统实时响应能力较高,但对于 TB 级或 PB 级别的大数据集,数据挖掘的检索速度则急剧下降,但引入 HDFS 与 RDBMS 相结合的机制可以充分利用两者的优势,实现高效率的数据挖掘与决策支持^[9]。

MapReduce 由 Google 设计,用于对集群上的大数据集进行并行计算处理,是非关系型数据管理和分析技术的典型代表。MapReduce 将数据处理任务抽象为一系列的 map (映射)和 reduce(化简)操作对,分别完成数据的过滤和聚集操作,并通过简单的界面进行管理。其计算流程和基本原理简单地说,就是将大数据集分解为成百上千个小数据集,每个(或若干个)数据集分别由集群中的一个节点进行

处理并生成中间结果,这些中间结果又由大量的节点进行合并,形成最终结果,如图 2 所示。

基于 MapReduce 计算模型编写分布式并程序的主要编码工作就是实现 map 和 reduce 函数,其他的分布式存储、工作调度、负载平衡、容错处理、网络通信等复杂问题均由 MapReduce 框架负责处理。MapReduce 原理简单、技术简洁、数据处理效率高,在系统层面解决了数据库集群难以解决的扩展性、容错性等问题,MapReduce 免费开源,基于异构廉价服务器搭建可弹性伸缩的大规模集群,并行、分布式地处理和分析大规模数据,其构建成本远低于数据库集群所采用的并行数据库^[9]。几种大数据处理模式的比较分析见表 1,可见,MapReduce 大数据处理模式具有相对显著的优势。

4 大数据处理对电子商务业务的影响

在 IT 基础设施上,淘宝率先引入云计算技术,阿里云成为我国第一家专门从事云计算服务的公司,淘宝的所有交易系统都为自建,通过电商平台集成海量数据,以下主要以淘宝为例,分析引入 MapReduce 大数据处理模式给电商业务带来的革命性影响。

4.1 强大的信息检索服务

商品的丰富性直接影响电商的竞争力,而海量的商品数目、繁杂的分类体系以及复杂的非结构化的商品属性数据等都需要 IT 基础设施具备足够的灵活性和强大的检索能力。

云平台架构提供的超大规模计算能力和大数据处理

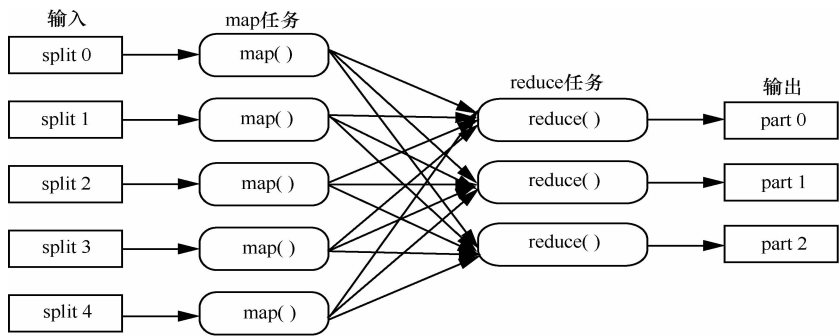


图 2 MapReduce 并行计算流程^[3]

表 1 大规模数据处理模式的比较分析

特征	扩展性	完整性与安全性	灵活性	数据处理能力	处理效率	异构存储	成本
数据库集群	非常有限,上百节点	事务处理、灾备、容错性差(节点失效影响整体)	低	主要支持结构化数据	决定于优化的分布程序设计	欠缺	硬件复杂度高,追加成本急剧上升
MapReduce	跨平台可扩展,上千节点	高容错性、强健性(节点失效不影响整体)	高	支持复杂的数据处理,包括结构化和非结构化数据	高	高支持	硬件需求低,扩展成本很低



能力能够提供强大的个性化信息检索功能,即根据用户的个体差异、个人兴趣和需求特征进行智能海量检索,并高效率返回高查全率和查准率检索结果^[9]。另外,还能实现信息推送服务、热点信息推送、信息推荐等新型信息检索服务^[9]。

云计算的技术优势使得信息检索和服务可以很好地解决长期存在的人类自然语言理解、知识推理等问题,充分发挥深度数据挖掘和知识发现的功能,以迅速准确地分析处理用户信息行为、理解用户自然语言表达并进行相应智能检索,得出符合用户需求的信息和产品,提高用户服务的速度和精准度^[10],最大限度地提升客户满意度。

淘宝商品具备数 10 种属性,还包括视频、图像等多媒体数据,淘宝用户可以通过颜色、价格、品牌等几十种属性在海量数据中进行智能检索和过滤,检索的实时性和准确率很高。2010 年淘宝注册用户数达到 3.7 亿户,在线商品数达到 8 亿件,最多每天有 6 000 万人访问淘宝网,平均每分钟出售 4.8 万件商品^[14],实时呈现订单结果。

4.2 快速的弹性处理能力

电商系统必须具备无以伦比的快速弹性处理能力,能够处理突发的访问量、海量订单和客户浏览请求,而且需要根据需求和业务量的上涨不断扩容服务器和增加数据存储设备。

基于云计算技术的云存储平台拥有理论上无限的海量存储和超大规模计算等资源,能够存储和处理 TB 级乃至 PB 级的海量数据^[11],企业不用安装硬件,就可以廉价、快速地部署应用系统并实现弹性伸缩,以提高资源的管控能力和促进优化利用。这种丰富的 IT 基础设施和弹性处理能力能够廉价、快捷地输出给中小企业。淘宝和天猫 2012 年中国互联网时尚消费数据显示,淘宝在线销售的原创商品每日更新超过 100 万件,由于淘宝的成功营销,2010 年、2011 年、2012 年的 11 月 11 日,淘宝单日交易额分别为 9.36 亿元、52 亿元、191 亿元^[12],销售额的激增和“双十一”营销的成功,验证了淘宝云计算平台的运转效率及快速的弹性处理能力。

4.3 精准的海量数据分析

实时性的海量数据分析越来越成为电商的核心竞争力,大数据的价值关键在于信息分析和利用。云计算可在极短时间内对海量数据和大数据进行收集、存储、分析和处理,极大地提高了企业的信息分析能力,使得电商需要的实时精准的海量数据挖掘和大数据深度分析等成为可

能。淘宝每天数以千万计的交易产生大量的交易时间、商品价格、购买数量等交易数据和利益相关方的年龄、职业、地址等个人特征信息,从这些海量数据中,淘宝实时准确地进行各类店铺排名和个性化智能推荐;进行用户行为数据分析,得到电商用户所需的个性化信息与产品,便于开展精准营销;商家根据历史信息和“淘宝指数”进行生产、进销存计划;买家得以获得更符合个性化需求的商品信息,提高客户满意度。

4.4 云化的信息安全服务

信息安全是电商企业业务可持续的最关键保障。在大数据时代,大数据成为国家和企业的核心资产,大数据蓝海成为未来竞争的制高点。但是,大数据往往更复杂、更敏感,更易成为网络攻击的显著目标,加大了隐私泄露风险,大数据深度分析技术让黑客的攻击更精准。电商系统不可能防止外部数据商挖掘个人信息,各社交网站均不同程度地开放用户所产生的实时数据,外部数据提供商能够通过收集、监测、分析这些数据得出用户的信息体系,常规的安全方案与措施无法满足大数据时代数据非线性增长的需求,用户隐私安全问题将更为显著。例如,能够通过智能手机定位分析精确锁定个人位置。

另一方面,大数据处理技术能够全面、及时、精确地监测并获取各类网络异常行为或网络攻击行为的结构化和非结构化数据,实时进行安全分析和预防性分析,以便度量企业安全级别和安全风险,更有针对性地设计、实施信息安全方案,应对安全风险,寻找攻击源,识别钓鱼攻击,防止诈骗和阻止黑客入侵等。

云计算技术能够把专业可靠的信息安全方案封装为云服务,为用户提供优质、廉价、全面的安全和备份服务。所有信息资源都托管在云端,由云计算强大的服务器集群和虚拟化技术提供冗余、灾备、数据备份和自动故障恢复等功能,专业的 IT 管理团队负责维护电商企业数据,提供专业化的信息安全与保密方案^[9,13]。

淘宝网构建了系统的安全体系,包括支付安全、信用评价、店铺评分、物流保障、网络安全、风险控制、消费者保障等,并且不断加强自身系统安全、打击各类网络诈骗行为、开展用户安全教育,2012 年淘宝安全中心共拦截 592 万个钓鱼网站,拦截木马 1 477 万次,引导超过 2 400 万名淘宝用户绑定二次验证、淘宝安全中心等安全产品^[14],这些安全举措较为成功地净化了网购环境,保障了电商的规模发展。目前淘宝网的消费投诉率不到

1%, 低于实体零售商。

5 结束语

云计算、大数据已被公认为信息技术的热点和趋势, 本文主要探讨如何借助云计算和大数据处理技术改善电商系统的 IT 基础设施、转变服务理念、创新商业模式、发展电子商务。笔者认为, 大数据处理在电子商务上的未来发展趋势主要有以下几方面。

(1) 专业化数据处理服务与产品日益丰富

超大规模电商企业(集团、联盟)、政府、公益机构、盈利机构等在云环境下得以协同工作, 免费或收费地提供专业化数据处理服务或产品, 并可输出给没有足够 IT 能力的中小电商企业。

(2) 大数据定制化系统

大数据定制化系统成为超大规模电商的商机和趋势, 并可作为 IT 能力输出给中小电商企业, 目前, 浪潮集团等企业已发布了定制化、专业化的大数据系统或产品。

(3) 构建新型电商的商业模式

云计算带来的规模经济效应、个性化服务、长尾效应、蓝海、众包等, 不仅变革了 IT 基础设施, 更重要的是重塑了经济学概念, 促进了企业业务模式的改变, 从而使得可以快速迈进服务经济时代^[4]。传统商业模式和业务模式得以彻底变革和创新, 随着技术与营销思想的良性互动, 将涌现出越来越多的新模式、新思路、新方法。

(4) IT 基础设施成为核心竞争力

云计算架构和大数据处理的巨大可扩展性, 为海量计算资源、存储资源的整合管控能力和 7×24 h 的高可靠性, 提供更加卓越的计算处理性能、简单方便的部署使用方案和全方位的监控管理措施, 可以显著提高数据中心和高性能计算中心的生产率和工作效率, 最大限度地降低 IT 运营和管理维护成本, 给传统 IT 基础设施带来革命性变革。这种技术优势突破了电商交易规模的发展瓶颈, 在技术的基础上提升了电商行业的发展上限, 使得 IT 基础设施逐渐成为电商企业的核心竞争力。

参考文献

- 1 亿邦动力网. 零售商认为大数据有利于电子商务. <http://www.ebrun.com/20121018/58764.shtml>, 2013
- 2 王珊, 王会举, 覃雄派等. 架构大数据: 挑战、现状与展望. 计算机学报, 2011(10): 1741~1752
- 3 覃雄派, 王会举, 杜小勇等. 大数据分析——RDBMS 与 MapReduce 的竞争与共生. 软件学报, 2012(1): 32~45
- 4 Liu F, Tong J, Mao J, *et al.* NIST cloud computing reference architecture. http://www.nist.gov/customcf/get_pdf.cfm?pub_id=909505.pdf, 2011
- 5 Mell P, Grance T. The NIST definition of cloud computing. <http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf>, 2012
- 6 董旭, 叶剑, 刘煜. 数据库集群技术浅析. 邮电设计技术, 2012(10): 23~25
- 7 数据库集群技术. <http://blog.csdn.net/sillybanny/article/details/532583>, 2013
- 8 赵春雷, 乔治·纳汉. “大数据”时代的计算机信息处理技术. 世界科学, 2012(2): 30~31
- 9 黄兰秋. 基于云计算的企业竞争情报服务模式研究. 南开大学博士学位论文, 2012
- 10 于晓燕. 基于网格的数字图书馆模式构建研究. 南开大学博士学位论文, 2010
- 11 赵海峰. 档案馆应用云计算技术的可行性分析. 兰台世界, 2012(20): 23~24
- 12 中国电子商务研究中心. 2010 年淘宝双 11 总销售额 9.36 亿元. <http://www.100ec.cn/detail-6071538.html>, 2012
- 13 孔楠. 基于云计算平台的商业服务模式研究. 上海外国语大学硕士学位论文, 2010
- 14 中国电子商务研究中心. 淘宝网发布 2013 春节网购安全宝典. <http://www.100ec.cn/detail-6082154.html>, 2012
- 15 CIO 时代网. 从经济学的角度来看云计算. <http://www.ciotimes.com/cloud/cjs/59434.html>, 2012
- 16 李满意. 大数据安全. 保密科学技术, 2012(9): 71~72
- 17 新浪科技. 淘宝宣布注册用户达到 3.7 亿. <http://tech.sina.com.cn/i/2011-01-06/20285067308.shtml>, 2012

[作者简介]



陈云海, 男, 中国电信股份有限公司广东研究院信息业务研发室主任、工程师, 主要研究方向为视频编解码算法、互联网应用、虚拟化、云计算、自然语言处理、知识管理及应用, 发表期刊论文近 10 篇。



黄兰秋, 女, 博士, 华南师范大学经济与管理学院讲师, 主要研究方向为云计算、数据挖掘、情报学, 发表期刊论文近 15 篇。

(收稿日期: 2013-03-04)