

文章编号:1007-6735(2013)03-0289-05

引入资源即服务的云计算架构及其应用

朱小栋, 高春昌, 王恒山

(上海理工大学 管理学院, 上海 200093)

摘要: 提出引入资源的四层云计算构架, 阐述资源在云环境中的必要性. 同时从资源的角度, 给出云和云计算的概念. 设计云计算环境下的数据采集模型, 并从数据资源的角度, 设计应用广义搜索树的面向云计算的数据预处理机制. 与已有的云计算构架对比, 该构架更能体现出云计算技术的特性, 为云计算提供了一个思考方向.

关键词: 云; 云计算; 数据采集; 数据预处理; 架构; 资源

中图分类号: TP 315; TP 274 **文献标志码:** A

Cloud Computing Architecture with Resource and Its Application

ZHU Xiao-dong, GAO Chun-chang, WANG Heng-shan

(Business School, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: A cloud computing architecture with four levels was presented. The notion of resource as a service RAAS was proposed and interpreted. Compared with existing cloud computing architectures, it reflects the core characteristics of cloud computing technology more exactly. The definitions of cloud and cloud computing were given from the standing point of data resource. The data acquisition model was designed. The way of selecting the processing equipment in cloud environment in term of data resource was described, which provides a new direction for the application of cloud computing.

Key words: cloud; cloud computing; data acquisition; data preprocessing; architecture; resource

云计算将计算任务分布在大量计算机构成的资源池上, 使各种应用系统能够根据需要获取计算力、存储空间和各种软件服务^[1]. 人们在生活中逐渐感受到云计算的魅力. 云计算技术自 2006 年提出以

来, Google、Amazon、Yahoo 等互联网服务商, IBM、Microsoft 等 IT 厂商都纷纷提出了自己的云计算战略, 各电信运营商也对云计算投入了极大的关注, 云计算平台极低的成本、良好的扩充性已成为学术研

收稿日期: 2012-09-14

基金项目: 上海市教委科研创新基金资助项目(12YZ103); 教育部高等学校博士点基金资助项目(20123120120004); 教育部人文社会科学青年基金资助项目(12YJC870037); 国家自然科学基金资助项目(71071098); 上海理工大学人文社会攀登计划基金资助项目(5812114277)

第一作者: 朱小栋(1981-), 男, 讲师. 研究方向: 云计算、软件工程、电子商务、知识发现. E-mail: zhuxd@usst.edu.cn

究和 IT 产业的关注热点. Google 宣称由于使用了云计算技术,其计算成本仅为竞争对手的 1/100,存储成本仅为竞争对手的 1/30^[2]. 当今,数据采集已经在互联网及分布式领域得到广泛使用. 由于云环境下数据的海量、实时、分布式的特点为数据采集带来了极大的挑战,数据采集领域已经发生了重要的变化. 因此,及时准确地采集到所需的数据尤为必要. 在云平台的应用中,对海量数据的处理需求,给计算设备的性能带来了严峻的考验,同时也为数据处理带来了新的问题,因此数据的预处理工作已成必然. 云平台下的数据往往由于其数据的海量性、实时性等因素,不能够很好地被采集,引入 RaaS 的云平台构架将云计算的内容直指其核心——数据资源,为云平台下的数据采集等工作起到了很大的帮助.

针对现有的云计算构架中的不足,提出了引入 RaaS 的云计算构架,从资源的角度诠释云和云计算的概念,从数据资源的角度对云计算下的数据采集和预处理问题进行描述和设计.

1 云计算架构

1.1 相关研究

伴随着云计算及其技术的广泛应用,对于什么是云计算及其应用有着激烈的讨论,尹红风在文献[3]中回顾了钱学森先生晚年的开放复杂巨系统和思维科学,并分析这些理论对计算机科学的新发展、云计算技术的发展所做出的理论贡献. Armbrust 等^[4]对什么是云计算、和以前的模型如 SaaS (software as a service) 有什么不同、为什么现在是云计算发展的最佳时机、云计算将创造什么样新的机遇、有哪些挑战以及如何应对等一系列问题进行了详细的解答. Yau 等^[5]从软件工程的角度对云计算的应用系统开发所遇到的挑战问题进行了理论分析. 张兴旺等^[6]创建了基于云计算的大规模数据处理框架模型,说明了云计算在大规模数据处理中的可行性. 刘真等^[7]提出并实现了基于云计算的铁路数据模型,通过大规模铁路货票数据处理实例,验证了其可扩展性和高效性.

云计算描述了一种基于互联网的新的 IT 服务增加、使用和交付模式,常涉及通过互联网来提供动态、易扩展而且经常是虚拟化的资源. 中国云计算中心认为云计算是一种基于因特网的超级计算模式. 现有的云计算大多是将云计算架构简单地根据其应用平台、设施或软件将其分为几个层次,不能够体现

云计算的核心问题. 本文通过云计算对数据的使用,指明云计算的核心问题就是对资源的应用,明确地提出了引入 RaaS(resource as a service)的云计算架构,并对数据的采集和预处理进行了描述.

1.2 基于资源的云计算概念

定义 1 云是基于互联网的复杂网络系统.

云平台不仅仅是一个平台、系统或者设备,它将 PC 以及其它设备的大量信息和处理设备集中在一起协同工作,使用平台中空闲的处理和存储设备最大限度地使用平台内的设备,对信息进行并行计算,从而实现高效的数据处理.

云涉及多个系统,如软件提供系统、物联网系统等,从而形成一个涉及多个系统的复杂网络.

定义 2 云计算是利用系统资源池内的闲置资源,根据系统需求进行计算的集中计算模式.

云计算技术实现了不同资源的快速组建进而形成新的解决方案,可以实现异地资源本地化. 云计算通过用户终端付费或免费的方式实现使用本身不具有的事物,也无需耗费精力对其进行维护,而且还可以根据用户的需要而增减服务. 云计算能够确通过平台将资源进行整合,从而形成新的问题解决方案,为用户节省成本. 云计算通过并行计算实现了数据的快速处理,形成多种备选的方案,具有良好的可扩展性,通过免费或购买的方式实现这些服务产品的使用,具有快速的实施速度,并且在这一系列活动过程当中没有产生任何的环境负担,是绿色无污染的产业活动.

1.3 引入 RaaS 的云计算架构

传统的云计算服务模式包括 3 层:软件即服务 SaaS,平台即服务 PaaS(platform as a service),基础设施即服务 IaaS(infrastructure as a service). 本研究在此基础上提出新型的服务模式资源即服务 RaaS,提出了引入 RaaS 的云计算架构,如图 1 所示.

物质、能源、信息被称为 21 世纪的三大支柱,信息发挥越来越重要的作用. 因此,本文将信息资源作为云计算构架的一个层次,清晰地体现资源在云平台中的重要作用. RaaS 这一构架层次的引入,将资源作为云服务的内容.

这一构架层次实现以数据作为云计算的服务内容,利用云平台进行数据采集和数据处理,明确数据的搜索范围,提高了效率. 另一方面,节约不必要的数据处理设备,节省了成本,达到实现云计算的目的.

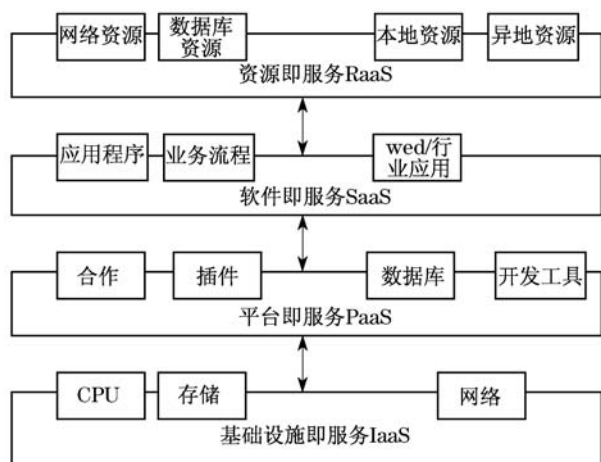


图1 引入 RaaS 云计算架构

Fig.1 Cloud computing architecture with RaaS

2 云计算下的数据采集模型

数据是云计算的对象,从 RaaS 的角度,进一步研究基于该架构的数据采集模型和数据预处理模型。

在云平台中数据往往具有很强的动态性,动态数据的特征为数据采集工作带来了极大的挑战,高质量的网络数据采集结果,对于云计算中的数据应用具有重要意义。

目前,网络数据采集大致包括利用网络爬行器 WebCrawler、网络抓包软件 wireshark、商业搜索引擎、Web 日志文件抓取,以及其它一些网络信息的数据采集方法等. Bharat^[8] 在 1998 年设计的所谓“连接服务器”,是以 AltaVista 一个拥有 1 亿 URL 地址的大型爬行器作为基础. 2000 年, Broder 等^[9] 进一步提出“连接服务器”的改进版,可以提供新近收集的数据,并利用图形描述 Web 的结构关系. 数据的采集结果直接关系到所要验证结果的准确性,这一环节至关重要。

云计算可以实现数据的规则更改,对不符合采集规则的任务进行二次设定,从而得出规范的数据,为下一步的数据预处理工作带来极大的方便. 云环境下数据采集模型如图 2 所示. 数据采集的各个阶段划分很重要,能够帮助用户获得合适的数据。

a. 确定采集对象 从研究问题的角度出发,选择正确的数据采集方向。

b. 设定采集规则 由于数据的海量性,往往一批数据中包含众多的冗余信息,选取需要的数据进行采集,舍弃冗余度不必要的数据. 同时检查采集规则是否合适,如不合适,则重新设定采集规则. 之后进行数据采集。

c. 导出数据 当数据采集结束后导出数据,并以所需的格式对数据进行存储。

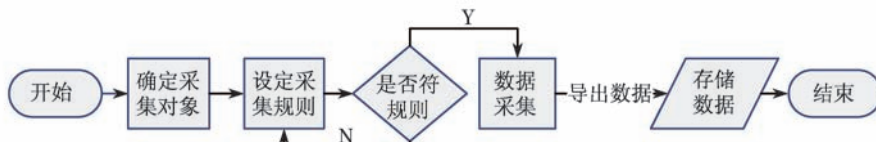


图2 云环境下数据采集模型

Fig.2 Data gathering model in cloud environment

3 面向云计算的数据预处理

数据采集所得到的大多数原始数据都是脏数据,严重影响了数据分析的准确性. 伴随着云计算的兴起,对数据处理的速度和精确程度有了更高的要求. 因此,对数据进行预处理使其规范化和可操作化,对数据的后期挖掘与分析极为有利. 网络数据与日俱增,使得传统的数据预处理方式在云计算环境下已经不再适用. 如何对网络数据进行及时、高效预处理引起了越来越多学者的关注。

Dean 等^[10]介绍了一个大型集群上简化数据处理的 MapReduce 方法,其具有很强的容错能力,不会因

为部分数据问题而影响数据处理能力. Elteir 等^[11]研究了通过异步数据处理增强 MapReduce 的方法。

云计算是利用资源池中的处理设备对数据进行集中处理,因此,要找到平台当中空闲适合的处理设备. 采用广度优先树的方法能较好地搜寻平台当中的空闲设备。

图 3(见下页)给出了云计算下数据预处理模型示意图,从数据采集设备得到数据经过数据清洗与变换,云计算采用的是 MapReduce 进行数据处理,在 Map 环节中同时将空值点、噪声点和不一致点清除,并对需要变换属性的数据进行属性的变换. 在 Reduce 环节当中去除冗余和数据集成,从而得出精简的数据,如图 3 所示。

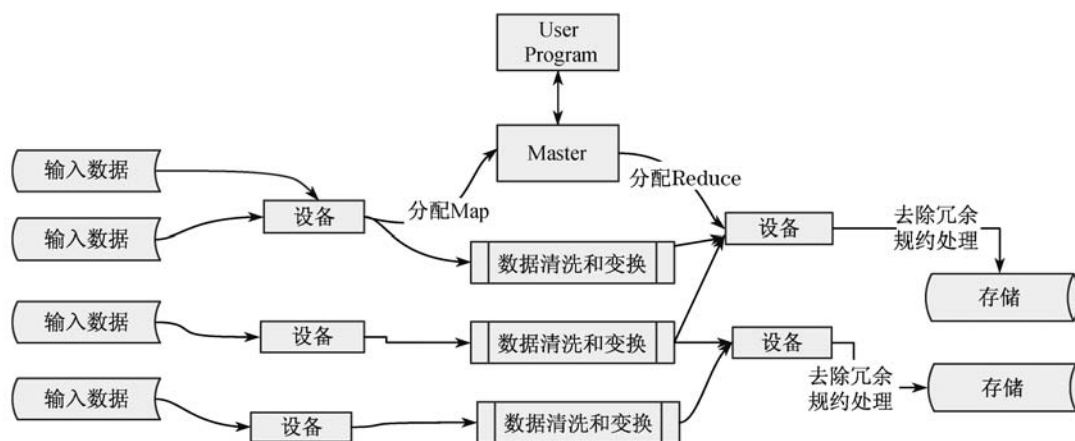


图3 云计算下数据预处理模型

Fig.3 Data preprocessing model in cloud environment

4 示例分析验证

微博作为一个复杂网络的典型实例,也是一个云,个人PC终端和云平台的交互形成了一个复杂网络,从而证实云是基于互联网的复杂网络.微博从无到仅仅用了几年的时间,如今中国四大门户都已开通微博.根据腾讯新闻,2011年上半年中国微博用户数量从6 311万增长到1.95亿,半年增幅达208.9%,手机微博的应用也成为亮点,手机网民使用微博比例也从2010年末的15.5%上升至34%.现有的研究都集中在微博信息的舆论以及信息情报机制等方面.微博作为云平台下流式数据的一个代表,对社会舆论导向具有很大的影响.样本数据3 360条,研究其数据演变情况,对于流式数据处理和引导社会舆论都具有重要意义.本文同样也采用微博数据作为研究对象.

以Twitter网一周内的热点话题为例,样本数据3 360条,在这类微博信息中,要求的数据是热点话题,因此在数据采集时只要侧重这几个方面的采集规则即可.采集中的规则包括时间、标题、查询关键字(关键字)、事件等多个方面.鉴于在数据分析中(假设)不会用到所有的数据属性值,在预处理中对数据进行清洗、变换、去冗等处理,包括将文本数据转化为数据值型数据,合并同类项最终得到简单明确的1 176条数据.

针对Twitter网站中一周内的热点话题数据的采集和预处理之后,能够快速得出图4所示的曲线.

由图4可见,对于一般性的话题其持续时间 t 在20 h以下,显示了网络微博数据的短寿命,而对于凸显节日性的话题则持续时间比较长,且这类的数

据也是舆情监督的重点,对于非常规的热点话题进行监督并对可能出现的突发事件进行调控.

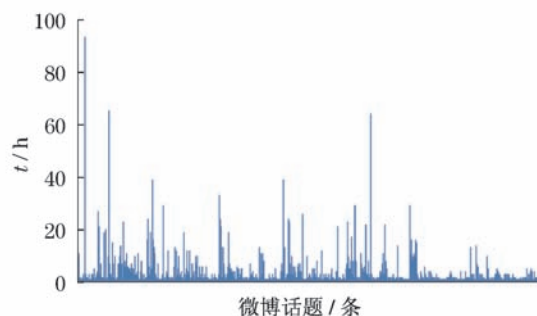


图4 热点话题持续时间

Fig.4 Duration of hot topics

5 结束语

云计算广泛应用于IT产业、电子商务、军事、物流等多个领域或者是多个领域相结合的复杂网络当中.如何提高云计算的服务效率,并为终端用户提供准确的信息服务等已经成为学术研究和各个云服务商产业的新焦点,本文对当前的云计算架构中的不足提出了引入RaaS的云计算架构,反映了云计算的信息资源服务层次,并对云环境中的数据采集和预处理进行了研究.云计算中的用户隐私和数据安全性等方面的研究,将是笔者下一步工作的重点.

参考文献:

- [1] 刘鹏. 云计算[M]. 北京: 电子工业出版社, 2010.
- [2] 孙健, 贾晓菁. Google云计算平台的技术架构及其成本的影响研究[J]. 电信科学, 2010(1): 38-44.
- [3] 尹红风, 戴汝为. 思维与智慧科学及工程[J]. 上海理

- 工大学学报,2011,33(1),18-23.
- [4] Armbrust M, Fox A, Griffith R, et al. Above the clouds: a Berkeley view of cloud computing[R]. Tech Rep UCB/EECS-2009-28. Berkeley: University of California, 2009: 1-23.
- [5] Yau S, An H. Software engineering meets services and cloud computing[J]. Computer Society, 2011, 44(10): 47-53.
- [6] 张兴旺,李晨晖,秦晓珠. 云计算环境下大规模数据处理的研究与初步实现[J]. 现代图书情报技术, 2011, 204(4): 17-23.
- [7] 刘真,刘峰,张宝鹏,等. 云计算模型在铁路大规模数据处理中的应用[J]. 北京交通大学学报, 2010, 10, 34(5): 14-19.
- [8] Bharat K, Broder A, Henzinger M, et al. The connectivity server: fast access to linkage information on the Web [J]. Computer Networks and ISDN System, 1998, 30(1): 469-477.
- [9] Broder A, Kumar R, Moghoul F. Graph structure in the web[J]. Computer Networks and ISDN System, 2000, 33(1): 309-320.
- [10] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters[J]. Communications of the ACM-50th anniversary issue: 1958-2008, 2008, 51(1): 107-113.
- [11] Elteir M, Lin H, Feng W. Enhancing MapReduce via asynchronous data processing [C] // IEEE 16th International Conference on Parallel and Distributed Systems, New York: IEEE Press, 2010, 397-405.
- (编辑:金虹)

(上接第 288 页)

参考文献:

- [1] 马德锋,朱孔敏,邱明. 基于 LabVIEW 的圆柱度误差测量软件设计[J]. 轴承, 2007(11): 43-47.
- [2] 郑鹏,张琳娜,陈明仪. 现行圆柱度测量方法分析及采样方案新策略[J]. 机械强度, 2011, 33(1): 152-158.
- [3] 张小梅,李春梅. 高精度圆柱度测量方法[J]. 计测技术, 2007, 27: 51-54.
- [4] 蔡俊,王宏光. 转子临界转速实验与计算的对比分析[J]. 上海理工大学学报, 2007, 29(5): 471-475.
- [5] 赵茜,王东霞,刘兰英. 圆柱度误差及其评定方法综述[J]. 计量与测试技术, 2006, 33(12): 1-2.
- [6] 荆学东,何凯,杨亮. 圆度和圆柱度测量虚拟仪器的开发[J]. 自动化仪表, 2011, 32(10): 52-54.