应用实践

搭建基于云计算的开源海量数据挖掘平台

赵华茗
(中国科学院国家科学图书馆　北京 100190)

【摘要】通过分析亚马逊弹性 MapReduce(EMR)平台构架,针对信息情报机构内部数据处理的迫切需求,提出通过开源技术 Xen和 Hadoop平台构建基于云计算的动态可伸缩的海量数据处理平台并给出实施方案、海量文本数据处理案例和开源 EMR 平台的优势分析。实施方案主要分为三部分:搭建动态虚拟的云计算环境、安装制作 Hadoop虚拟服务器模板、配置运行 Cloudera和 Cloudera Desktop。通过开源 EMR 架构的应用,可以有效解决服务器蔓延问题,提高网络计算资源的利用效率和分布式数据挖掘服务的快速布署能力及灵活性。
【关键词】云计算　海量数据挖掘　虚拟技术　分布式计算　Xen　Cloudera　Hadoop
【分类号】TP393

Building the Open Source Mass Data Mining Platform Based on Cloud Computing

Zhao Huaming
(National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

【Abstract】Aiming to meet the internal data processing needs of information organizations, this paper, by analyzing the frameworks of Amazon Elastic Map/Reduce (EMR) platform, puts forward to build the dynamic and elastic open source mass data mining platform based on cloud computing, and provides a roadmap of successful implementation, an example of massive text data processing and the analysis of advantages of open source EMR platform. This implementation plan includes three parts: building dynamic virtual environment of cloud computing, creating the virtual server template of Hadoop, and deploying and running Cloudera and Cloudera Desktop. Through the application of open source EMR platform, the problem of server sprawl can be solved effectively, the utilization ratio of network computing resource is improved, and the rapid deployment capability and agility of distributed data processing services are enhanced.
【Keywords】Cloud computing　Mass data mining　Virtualization　Distributed computing　Xen　Cloudera　Hadoop

1 引　言

互联网促进了信息流通,也带来了信息的爆炸式增长,最新的 IDC研究报告指出 2010年全球信息量将进入 ZB时代,并且每年以 60%的速度在上升,这意味着每 18个月全球信息数据量将被翻倍[1]。面对不断拓展的惊人的数据规模,海量信息的存储与管理、实时处理、数据搜索、数据挖掘与智能应用等信息处理能力面临新的挑战,信息技术架构迫切需要以动态可伸缩为特点的支持海量数据处理的新的存储计算模式。

收稿日期: 2010-09-26
收修改稿日期: 2010-09-28
* 本文系“第二十四届全国计算机信息管理学术研讨会”论文。

云计算因为其弹性可伸缩的计算模式, 受到以IBM、亚马逊、谷歌等为代表的众多高科技公司的重视, 成为各公司应对海量信息处理的利器。近年来, 出现了众多各具特色的云计算应用产品, 包括应用在服务托管领域的亚马逊弹性云、著名的谷歌搜索、Zoho在线办公应用等。而在海量数据存储挖掘领域的典型云计算应用也出自于亚马逊公司, 即亚马逊的 Hadoop架构服务, 称为弹性 MR (Elastic MapReduce, EMR) [2,3], 其整体架构如图 1所示:

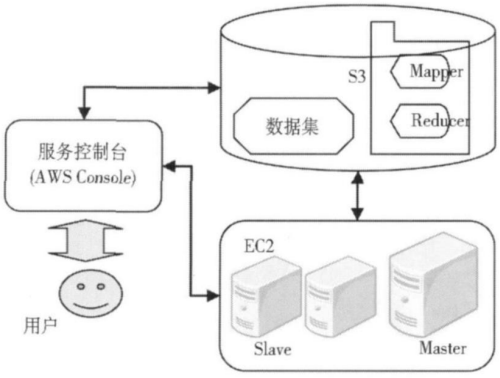

图 1　亚马逊 EMR 整体构架图[2]

EMR框架分为三部分, 控制接口 (AWS Console)、存储服务 (Simple Storage Service, S3) 和弹性计算 (Elastic Compute Cloud, EC2)。通过 EMR平台, 企业、研究人员、数据分析师可以花费较少的费用轻松处理大数据集而不用担心计算设备问题。显而易见, 随着数据宇宙时代的到来, 这种新的动态可伸缩的数据处理模式必将在信息技术领域扮演越来越重要的角色。本文在云计算开源框架下给出如何搭建信息分析机构内部的弹性海量数据挖掘平台解决方案并给出了实施步骤。

2　云计算开源框架 Cloudera和 XenServer

在可信计算尚不成熟的时候, 通过开源技术实现随需弹性处理机构内部数据是机构信息人员搭建机构内部的 EMR框架的出发点。分析亚马逊 EMR服务框架, 不难看出动态可伸缩的大数据集计算模式主要解决两个关键问题: 动态布署虚拟 Hadoop节点服务器和在节点服务器间快速配置搭建 Hadoop分布式环境。考察目前已有较成熟的开源技术, 在开源框架下搭建这样动态灵活的大数据处理解决方案可以有几种组合方式, 如: 最接近亚马逊 EMR框架的 Eucalyptus+ Hadoop组合、基于虚拟操作系统上的分布式文件系统环境 VMware+ Hadoop组合和 XenServer+ Cloudera组合等。目前, 还没看到介绍以开源框架搭建 EMR平台的相关文献。本文使用最后一种平台组合构建本系统机构内部的开源 EMR平台, 实现类似于亚马逊 EMR框架中大数据处理所需的控制、存储、计算等相应的三部分关键功能。

2.1　开源 Hadoop发行版 Cloudera系统平台

Cloudera是一款开源标准的 Hadoop布署及调度平台[4], 可以有效提升 Hadoop环境的易管理性, 降低对使用人员的要求, 使 Hadoop初学者也可用该软件搭建谷歌式云计算平台, 处理海量数据。据统计, 目前大约有 75%的 Hadoop新用户使用 Cloudera[5]。

2.2　Hadoop交互管理平台 Cloudera Desktop

Cloudera Desktop[6]提供了一个图形化的基于 Web的针对 Hadoop框架的交互管理平台。用户可以以可视方式进行文件系统操作、权限管理、MapReduce任务管理、提交、浏览、监控计算任务状态并提供监控图表功能、浏览存储资料等。支持主流浏览器 (Firefox, Chrome, Safari和 IE8+)。

2.3　免费的虚拟机管理平台 XenServer

虚拟技术是云计算基础构架技术, 也是本文搭建开源 EMR平台的基础构架技术。目前, 典型的虚拟机技术实现有 Critrix XenServer[7,8]、VMware ESX Server[9]、Microsoft Hyper- V[7] 等。相较其他两种, Citrix XenServer的虚拟机的性能更接近真实硬件环境, 硬件支持广泛, 具有更好的安全性、兼容性和开源性, 也是本文选用的主要原因。

3　搭建开源 EMR环境

本文设计的开源 EMR平台包括虚拟云平台和分布式计算平台两部分, 其整体架构如图 2所示。

虚拟云平台是一个基于网络的动态可伸缩的虚拟设备环境, 弹性管理网络设备资源, 如: 动态调配物理设备资源、存储设备资源及动态创建的 Hadoop虚拟服务器, 有效支持对计算资源的规模化集约化管理。本文的虚拟云平台使用 XenServer和 XenCenter实现。而分布式计算环境是一个基于 Hadoop Map/Reduce框架的开源大数据并行计算环境, 选用 Cloudera和 Cloudera

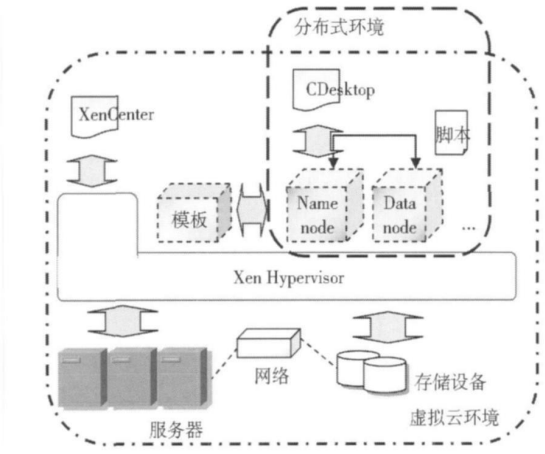

图 2 开源 EMR 整体构架图

Desktop实现。Cloudera Desktop是一个访问控制接口，提供基于 Firefox浏览器的简单界面。开源 EMR的实现过程可以分为三部分：搭建动态虚拟的云计算环境、安装制作 Hadoop虚拟服务器模板、配置运行 Cloudera和 Cloudera Desktop。第一部分是虚拟云环境，第二、三部分是分布式计算环境。

3.1 搭建动态虚拟的云计算环境

通过 XenServer和 XenCenter搭建动态虚拟的云计算环境的过程较复杂[10]。主要注意事项如下：

(1)XenServer要求安装在 64位 X86服务器上(32位服务器不支持 64位的虚拟机)，并且不支持多系统，不支持多系统引导，不能再安装运行其他应用程序。

为充分利用虚拟平台的动态资源调度特性，应以资源池方式管理虚拟云环境中的服务器设备资源和存储设备资源，因此所有安装 XenServer软件的主机应配置静态地址，其中一台为资源池的管理机，其他物理主机作为普通服务器加入，所有主机的管理员和密码最好相同。

(2)XenCenter安装要求有.NET框架 2.0及以上版本的支持，可以安装在普通 Windows管理机上。安装过程中应注意 XenCenter的安装路径，推荐修改为“C:\Citrix\XENCenter\”，以确保后期程序调用 XenServer API接口时不会出现路径指向问题。

(3)云计算环境中的数据安全很关键，因此存储设备和云计算环境分开是最理想的。XenCenter支持的存储设备主要有三种：基于 NFS VHD的存储池、基于 iSCSI的存储池、基于 Hardware HBA的存储池。配置存储池时要注意存储设备的接口配置提示信息，基于 NFS VHD方式的阵列设备的正确配置信息应该是“<nfs_ipaddress>:/<sr_share_dir>”(XenServer5.5版的提示信息模糊)。配置基于 HardwareHBA方式的阵列设备时，要在 XenServer安装或系统恢复时将光纤临时断开，系统运行后再物理连接上，系统会自动找到已连接上的阵列设备。

3.2 安装制作 Hadoop虚拟服务器模板

制作 Hadoop虚拟服务器模板是搭建开源 EMR框架的关键点之一，通过 Hadoop模板，可以在已搭建好的虚拟云环境中快速创建 Hadoop虚拟服务器节点，有效节省系统安装时间，并将 Hadoop分布环境的搭建重点放在 SSH安全数据传输连接和参数同步上，从而快速完成 Hadoop分布环境的部署和调度，与传统 Hadoop分布式环境相比，基于虚拟环境的 Hadoop分布式环境更灵活、更简捷、更有效地利用网络计算资源。主要制作过程和注意事项如下：

(1)安装 Linux基本操作系统(Base System)

通过 XenCenter创建新的服务器模板，在确定模板所需的操作系统类型、处理器、内存及硬盘大小后，即可根据提示完成模板的安装。根据 Hadoop分布环境安装要求，在测试过程中，选择安装 Linux操作系统 CentOS，虚拟处理器 Xeon E7420一个，虚拟内存 1GB，虚拟硬盘空间 60GB。整个安装过程与单机安装 CentOS操作系统的实际过程类似。

注意事项：免费版的 XenServer5.5对虚拟硬盘空间大小有限制，不能在创建服务器(基于已创建好的模板)时动态调整虚拟硬盘空间大小(但可以动态调整处理器和内存的大小)，因此在创建服务器模板时要考虑具体应用环境的需要，确定合适的虚拟硬盘空间。

(2)安装 Hadoop软件包

经过第一步，Linux操作系统的基本系统安装完成后，即可开始安装 Hadoop软件包，为后面布署分布式计算环境做准备，本文使用 Hadoop的企业优化发行版 Cloudera平台工具。Cloudera被简化优化后，和 Hadoop安装过程略有不同，注意默认的安装路径即可，目前 Cloudera的稳定发行版支持到 Hadoop-0.20。主要安装过程如下：

```
#./jdk-6u16-Linux-i586-rpm.bin //安装 Java环境
#curl http://archive.Cloudera.com/redhat/cdh/Cloudera-cdh2
repo > /etc/yum.repos.d/Cloudera-cdh2.repo //yum 更新配置文件
```

```
# yum - y install Hadoop- 0.20- conf- pseudo //安装 Cloudera
# yum - y install Hadoop- 0.20- conf- pseudo- desktop//安装 Cloudera Desktop
```

注意事项: Cloudera安装没有用户限制(非 root用户使用 sudo yum安装即可), 但 Hadoop分布式文件系统格式化必须由 Hadoop用户执行, 因此, Hadoop服务器节点模板创建时最好同时创建 Hadoop用户并授予 sudo超级权限。

3.3 配置运行 Cloudera和 Cloudera Desktop

Hadoop虚拟服务器模板制作完成后, 即可在 XenCenter监控窗口中看到该模板, 双击后根据提示创建基于该模板的 Hadoop虚拟服务器, 同理, 根据需要可以创建多个 Hadoop虚拟服务器节点。每个节点都已经安装好 Linux操作系统和 Hadoop软件包, 用户只需要配置运行 Hadoop分布式环境, 即可实现大数据处理所需的计算环境。Hadoop平台有三种运行方式: 单机模式、伪模式和完全分布式模式, 这里仅讨论完全分布模式的配置。本文参考 Hadoop分布式环境配置[1]和 Cloudera安装过程[2]并在虚拟环境中测试后, 整理主要配置过程和注意事项如下:

(1)配置网络安全协议 **SSH**

在每个虚拟节点服务器上利用 SSH生成密钥对, 并且将彼此公钥追加到自身和其他节点机的 authorized_keys文件中, 以保证各节点之间能够通过 SSH工具不输入密码直接登录, 自身也必须保证不输入密码直接登录。

注意事项: 因为 Hadoop分布式文件系统格式化必须由 Hadoop用户执行, 所以各个节点上的 SSH配置最好以 Hadoop用户身份进行。

(2)分布式环境参数配置

Cloudera的默认参数文件路径是"/etc/Hadoop/conf/", 主要配置文件为 Hadoop-env.sh、core-site.xml、hdfs-site.xml、mapred-site.xml、conf/masters和 conf/slaves, 与 Hadoop的分布式环境配置相似。Cloudera默认参数为伪模式配置, 完全分布模式配置主要是修改 Java运行环境路径、节点机名称或节点机 IP地址等相关信息。配置好主节点 Namenode的参数后, 将配置文件拷贝到其他节点机上, 同步整个 Hadoop环境参数。参数数据同步命令如下:

```
$ scp- r /etc/Hadoop/confHadoop-slaves /etc/Hadoop
```

注意事项: Cloudera除了上述的 6个配置文件外, 还有一个专门针对 Cloudera Desktop的参数配置文件"/usr/share/Cloudera-desktop/conf/Cloudera-desktop.ini", 参数修改如下:

```
namenode_host= localhost修改为 namenode_host= namenode
jobtracker_host= localhost修改为 jobtracker_host= namenode
```

(3)运行分布式环境

初始化:

```
$ . /usr/lib/Hadoop- 0.20/bin/Hadoop namenode - format
```

启动 Cloudera

```
$ forx in /etc/init.d/Hadoop- 0.20- *; do $ x start done
```

启动 Cloudera Desktop:

```
$ sudo /etc/init.d/Cloudera- desktop start
```

成功启动 Cloudera和 Cloudera Desktop之后, 数据分析人员就可以通过浏览器方式简单实现对基于虚拟技术的分布式环境(EMR)的控制和管理, 并可以开始进行大数据处理。入口地址: "http://myserverip:8088/", 如图 3所示:

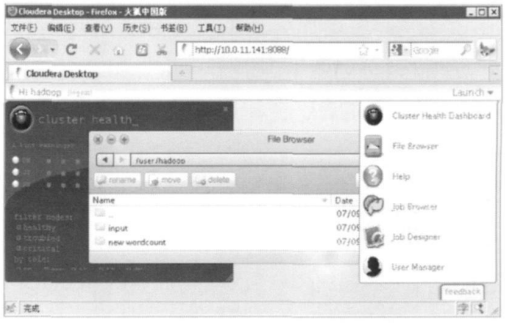

图 3 Cloudera Desktop服务入口示意图

3.4 Cloudera常见问题及解决方案

(1)大数据计算时常报"Name node is in safe mode"错误。原因分析: 在分布式文件系统启动的时候, 开始的时候会有安全模式, 当分布式文件系统处于安全模式的情况下, 文件系统中的内容不允许修改也不允许删除, 直到安全模式结束。

解决的命令:

```
$ bin/Hadoop dfsadmin - safemode leave /关闭 safemode
```

(2)大数据计算时报"Permission denied exception in job designer"错误。原因分析: 系统默认的临时目录"Hadoop tmp dir"的读写权限不够。

解决的命令:

```
$ chmod 777 /var/lib/Hadoop- 0.20/cache/Hadoop
```

3.5 增加 Hadoop虚拟服务器节点(Slavenode)

虚拟云计算是基于虚拟技术动态分配网络计算资

源,随需提供入门级应用的计算模式。而结合虚拟技术和分布式技术则在理论上提供了可无限动态扩展网络计算资源的随需提供企业级海量数据挖掘的计算模式。在开源海量数据挖掘平台(EMR)中,动态扩展存储和计算资源的方法是在增加 Hadoop节点服务器方法[13]的基础上,结合虚拟技术实现,关键步骤如下:

(1)在 XenCenter监控窗口中,基于 Hadoop虚拟服务器模板创建新的 Hadoop节点服务器 Slavenode;实际上,结合 XenAPI接口,该步骤也可以在浏览器的方式下在线完成[14];

(2)在 Slavenode上配置好 Hadoop节点服务器所需的运行环境,包括 SSH、相关 Cloudera配置文件的拷贝;

(3)将新的 Slavenode的 Host信息加到集群 Namenode及其他 Datanode的 /etc/hosts配置文件中去;将新的 Slavenode的 IP加到 Master的 conf/slaves中;

(4)重启 Hadoop Cluster,在 Cluster中看到新的 Slavenode节点。

注意事项:新添加一个节点到集群当中时,HDFS不会自动地移动文件块到新节点中去平衡磁盘空间,而新创建的文件将只会使用新节点的磁盘空间。解决方案如下:

(1)将文件复制一份,然后删掉源文件。

(2)将磁盘块满的节点关掉,等待直到文件块自动复制完成,再把节点加回去。这样,冗余数量变得过多,系统将会随机删去多余的冗余。

(3)运行 bin/start- balancer sh命令,这会很耗时间。

4 海量数据挖掘实例

在信息爆炸的今天,海量数据挖掘几乎是任何一个信息分析机构要面对的课题。而用户每天面临的80%信息都是文本信息。这里通过基本的 MapReduce WordCount算法测试说明基于机构内部自建的开源 EMR平台如何进行海量文本数据挖掘处理,同时介绍使用 Cloudera Desktop进行简单的大数据处理的过程。试验目的是将大数据处理的难点 ETL过程交给开源 EMR平台处理,数据分析过程使用传统方式,实现数据处理过程中的优势互补。测试数据为非结构化的文本文件。测试过程是将文本文件上传到开源 EMR平台中,然后调用 WordCount算法处理文本文件,最后将处理结果下载到本地做进一步分析。整个过程在图形化界面中通过简单的系统交互完成。

4.1 WordCount算法

WordCount[15]是一个经典的 Map/Reduce的应用示例,它可以计算出指定数据集中每一个单词出现的次数,是文本数据挖掘处理中的基础部分和算法之一。在测试过程中,WordCount算法将上传的非机构化文本处理为词和词频数据对。

4.2 大数据挖掘的关键步骤

以 WordCount算法为例,处理结果如图 4所示:

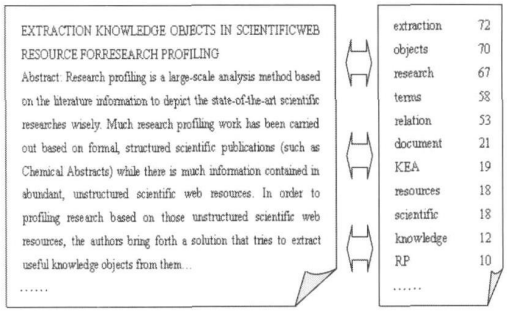

图 4 海量数据 WordCount算法处理结果示意图

(1)进入 Cloudera Desktop主界面,默认登录用户是系统创建的 Hadoop用户。成功登入后,用户可以看到图 3所示的界面:Cloudera Desktop有 4个主要功能模块,在界面右上角的 Launch菜单中,包括 Cluster Health Dashboard、File Brower、Job Browser以及 Job Designer;

(2)点击打开 File Brower,在 Hadoop用户主目录中,创建数据上传目标文件夹 INPUT,并上传需要处理的文本数据,如图 4中左边的文本就是测试中上传的文本文件内容,也可以是多个文本文件。

(3)点击打开 Job Designer,激活系统默认的基础算法,复制并修改 WordCount算法的数据输入(INPUT)和输出(随机命名输出文件夹名称,不能与已有文件夹重名,如使用 OP+Time方式命名)参数,完成新 WordCount Job任务的创建。

(4)运行该 Job任务,系统将自动创建 OPTime文件夹,显示任务执行的进度和结果,并将最终数据处理结果保存到该文件夹中。用户还可以通过 File Browser查看结果数据或将结果集下载到本地系统进一步深入分析。如图 4中右边的文件就是将处理结果下载到本地并排序后呈现的文本。

5 开源 EMR架构优势分析

数据挖掘是适应信息社会从海量数据中提取有用信息的需要而产生的。现在,政府、企业都把数据看成宝贵的财富,纷纷利用数据挖掘技术发现其中隐藏的信息。亚马逊在 2010年的 Hadoop峰会上表明其目前的数据挖掘业务比重非常大,并为提升其 EMR服务将有显著的持续投入[16]。总体来说,除了数据保密性外,信息机构内部的开源 EMR架构的应用很好地解决了如下几方面的问题:

(1)高效的网络存储和计算资源的控制利用,有效防止服务器蔓延,推动机构内部数据中心的绿色节能建设。通过虚拟技术将具有相类似的应用服务器整合到相对集中的资源池中,提高应用的稳定性和可用性,同时通过可视化监控界面动态配置、调整调度服务器及存储设备,提高计算资源的利用效率。

(2)加速分布式数据挖掘服务部署能力。通过分布式服务应用映像模板,用户可以根据数据挖掘的任务和数据规模,简单、灵活地创建和增减 Hadoop服务器节点,形成规模合适的容错性强的 Hadoop集群,低成本快速完成数据挖掘任务。任务完成后,还可以快速收回计算资源给其他应用使用,深层次挖掘计算资源的可利用空间。

(3)大数据处理的简单化,开发方便。通过图形化 Hadoop平台管理界面,海量数据处理对专业数据分析人员来说不再是复杂的服务器集群软、硬件和数据挖掘算法的整合过程,系统屏蔽掉底层,数据分析师可以将主要精力放到数据挖掘算法上。这种大数据处理过程的简单化趋势将推动知识挖掘、发现的快速发展。

6 结　语

本文通过分析亚马逊 EMR海量数据处理平台构架,针对信息分析机构内部数据处理的迫切需求,提出通过开源技术 XenServer和 Cloudera版 Hadoop平台构建信息机构自己的动态可伸缩的海量数据处理平台并给出实施方案和文本数据处理案例。目前,开源 EMR平台在算法、多节点计算速度、实时数据处理、中文支持等方面还有很多限制和不足,这也是笔者下一步改进的方向。

参考文献:

[1] 2010 Digital Universe Study [EB/OL]. [2010-09-27]. http://gigaom.files.wordpress.com/2010/05/2010-digital-universe-iview_5-4-10.pdf

[2] Amazon Introduces Elastic MapReduce (Hadoop Framework) Service[EB/OL]. [2010-09-27]. http://www.byteonic.com/2009/amazon-introduces-elastic-mapreduce-Hadoop-framework-service/.

[3] Amazon Elastic MapReduce [EB/OL]. [2010-09-26]. http://aws.amazon.com/elasticmapreduce/.

[4] Cloudera Enterprise[EB/OL]. [2010-09-27]. http://www.Cloudera.com/products-services/enterprise/.

[5] Hadoop中国 2009云计算大会[EB/OL]. [2010-09-27]. http://Linux.chinaunix.net/news/2009/11/15/1144192.shtml

[6] Developing Applications for HUE[EB/OL]. [2010-09-27]. http://www.Cloudera.com/blog/2010/07/developing-applications-for-hue/.

[7] Pratt I, Fraser K, Hand S, et al. Xen 3.0 and the Art of Virtualization [EB/OL]. [2010-09-27]. http://www.Linuxsymposium.org/2005/Linuxsymposium_procv2.pdf

[8] Technical and Commercial Comparison of Citrix XenServer and VMware[EB/OL]. [2010-09-27]. http://www.citrix.com/site/resources/dynamic/salesdocs/XS_vs_VMware_comparison.pdf

[9] VMware vSphere[EB/OL]. [2010-09-27]. http://www.vmware.com/products/esx/.

[10] XenServer Installation Guide[EB/OL]. [2010-09-26]. http://support.citrix.com/servlet/KbServlet/download/18052-102-19049/installation.pdf

[11] Hadoop Cluster Setup[EB/OL]. [2010-09-26]. http://hadoop.apache.org/common/docs/r0.20.0/cluster_setup.html

[12] Hadoop 5_minute Quick Start[EB/OL]. [2010-09-26]. http://nightly.cloudera.com/docs-backup/hadoop_5_minute_quick_start.html

[13] Hadoop添加节点的方法[EB/OL]. [2010-09-26]. http://wenku.baidu.com/view/e57ffe3e0912a2161479291e.html

[14] 赵华茗,李春旺,周强.基于 XenServer的数字图书馆云服务平台实现研究[J].电信科学,2010,26(8A):33-38.

[15] Hadoop Map/Reduce Tutorial [EB/OL]. [2010-09-27]. http://Hadoop.apache.org/common/docs/r0.18.2/mapred_tutorial.html

[16] Amazon Elastic MapReduce Updates from Hadoop Summit 2010 [EB/OL]. [2010-09-27]. http://www.infoq.com/news/2010/07/amazon-elastic-mapreduce-updates

(作者 E-mail:zhaohm@mail.las.ac.cn)