

基于 Hadoop 高校教育资源云存储平台构建研究

王凤领

(黑龙江财经学院 财经信息工程系,黑龙江 哈尔滨 150025)

摘要:云计算在整合高校教育资源方面的关键是海量的数据存储,把云存储的技术和方法运用到高校教育资源整合的建设中,节约高校教育资源,有效地提高高校教育资源的利用率,从而提高高校的教学水平。文中阐述了云计算和云存储的概念和特点,介绍了高校教育资源中存在的问题和优势,简单介绍了主流的云计算平台,重点介绍了 Hadoop 的 HDFS 存储系统的特点和结构;结合基于云体系结构,设计出高校教育云体系结构,实现基于 Hadoop 平台的高校教育资源网架构,使用 Hadoop 平台进行数据的存储和传输。实验结果表明,利用 Hadoop 云计算平台,为大规模高校数据处理提供良好的解决方法,有效提高海量数据的处理速度。

关键词:云计算;Hadoop;云存储;分布式文件系统;教育资源

中图分类号:G64

文献标识码:A

文章编号:1673-629X(2016)03-0176-05

doi:10.3969/j.issn.1673-629X.2016.03.041

Study on Construction of Cloud Storage Platform for College Education Resources Based on Hadoop

WANG Feng-ling

(Department of Financial Information Engineering, Heilongjiang University of Finance and Economics,
Harbin 150025, China)

Abstract: The key for cloud computing in the integration of university education resources is the mass data storage. The technology and method of cloud storage can be used in the construction of university resources integration. It can improve the utilization rate of resources, save resources and improve the teaching level. In this paper, it expounds concept and characteristics of cloud computing and cloud storage, presenting the existing problems and advantages in the higher education resources, then simply introducing the mainstream cloud computing platform, and focusing on the characteristics and structure of HDFS Hadoop storage system. With the combination of cloud system structure based on the system structure, the cloud system structure of education in colleges and universities is designed to realize the higher education resource network framework based on Hadoop platform. The data transmission and storage is conducted by Hadoop platform. Experimental results show that the use of Hadoop cloud computing platform can effectively improve the speed of mass data processing, which provides a good solution for large-scale and efficient data processing.

Key words: cloud computing; Hadoop; cloud storage; HDFS; educational resources

0 引言

整合教育资源,有效合理地使用教育资源,有利于高等教育的发展。云计算时代的来临,对于建立一个统一、开放和灵活的教育资源平台,实现资源共享,缩小教育信息化差距具有深远的影响。

把云存储技术和方法,运用到高校教育资源的整合建设中,有效地节约高校教育资源,并且提高高校资源的利用率。

1 云计算和云存储的概述

1.1 云计算的概念和特点

云计算作为一种新型的服务交付和使用模式,是基于互联网的计算方式,通过这种方式能够向用户提供所需的软硬件平台和网络资源^[1]。

(1)可扩展性:为满足用户的增长和需要,云计算的规模可以动态扩展和伸缩。

(2)虚拟化:用户通过终端设备登录到 Internet 中

的云服务平台,即可得到所需要的服务,不需要知道资源的具体所在位置。云计算将虚拟化的异构分布式的数字和物理资源,通过虚拟化技术统一存储在数字中心。

(3)稳定性:用户可将自己的数字资源放心存储在云服务平台,云计算通过虚拟化技术将数据资源存储在虚拟资源池中,有效地保证了数据的安全性。

(4)通用性:在云计算的支持下,可以支撑不同应用的运行,构造出各种功能的应用。

(5)运行成本低:用其廉价的节点构成云体系,通过云计算的特殊容错措施。云计算的通用性和公用性有效地提升了物理资源的利用率,由于云计算的自动化管理,也使其管理中心数据的成本大幅降低。

1.2 云存储的概念和特点

云存储把存储设备看作一个资源池,通过虚拟化技术,根据需要分配这些资源。云存储将庞大的数据自动拆分成若干个小数据块,并进行分布式存储,最后,交由存储资源组成的计算机集群系统^[2]。

云存储是将网络不同类型的存储设备,通过应用软件集合来协同工作,通过分布式文件系统、网格技术或集群应用等功能,来共同对外提供的业务访问和数据存储功能的系统^[3]。

(1)可靠性:云存储把数据备份到多个硬件存储设备中,保证数据的安全可靠性。

(2)服务性:云当中的数据是进行分布式存储的,能够有效分担存储和访问过程中的压力,提高系统性能。

(3)存储效率:云存储为避免个别服务器由于工作量过大造成的瓶颈问题,能够将工作量很好地均匀分配在不同的存储服务器上,使存储系统发挥最大的效能。

(4)硬件冗余:云存储将文件复制并保存到不同的服务器,来解决硬件损坏所带来的难题^[4]。在硬件发生损坏时,云存储知道文件具体存放的位置,而保持服务的继续。

(5)管理性:将高校教育资源统一存储在云中,方便教育资源的管理,节省了人力和物力^[5]。

2 高校教育资源存在的主要问题

(1)教育资源分布不均匀。对有充足资金提供支持的著名大学,就有更好的教育资源。地方院校存在设备老化,资金有限,资源不足等问题。

(2)基础设施的闲置。由于多方面原因,对多数高校在基础设施购入后,设备在多数情况下一直处于闲置和浪费状态,对设备本身也是一种浪费^[6]。同时,也提高了设备的运行成本,浪费电力资源、资金等。

(3)软硬件投入的不均衡。这是多数高校教育资源建设过程中,所普遍存在面临的共同问题^[7]。

(4)基础设施重复购置。因高校自身教育资源建设的需要等原因,重新建立或者更换陈旧的基础设施,都需要花费大量资金购入新设施。

(5)教育资源共享程度低。各院校、职能部门之间的资源共享手段较单一,效率也比较低,跨部门、院系的教学和科研合作受制于教育资源管理水平。

3 云计算在资源平台建设的优势

3.1 云计算技术

云计算构建的目的在于同时为许多用户服务。云计算系统运用了很多技术,其中以虚拟化、数据存储、云计算平台管理最为关键。云计算系统用冗余存储方式保证数据的可靠性,采用分布式存储方式能有效存储数据^[8-9]。

(1)数据存储技术。为满足用户需要,云计算系统并行为用户提供服务。云计算的数据存储技术需具有较高的传输率、吞吐率和分布式的特点。

(2)虚拟化技术。虚拟化技术能够实现底层硬件与软件应用之间的隔离,包括聚合模式(是把多个资源整合成一个虚拟资源)和裂分模式(是将单个资源划分成多个虚拟资源)^[10]。

(3)云计算平台管理技术。云计算系统平台管理能使多个服务器协同工作,快速发现系统故障和恢复系统,方便进行工作部署和开通,系统通过智能化、自动化的方法达到可靠运营。

3.2 可行性分析

基于云计算的开放性教育资源平台的建设,率先迈出了重要的一步,云计算为高校教育资源的建设提供了重要机遇。其可行性分析如下:

(1)扩大教育资源的共享度,避免资源重复建设,有效提高基础设施利用率。由于云计算对终端设备支持跨平台,可以随时随地使用网络共享资源^[11]。

(2)提高教育资源管理。通过分布式和并行计算,云计算对高校教育资源进行分析管理、存储和计算,高可靠性能够保证开放教育资源平台的高效运行。

4 云计算平台

目前主要的云计算平台有:微软 Azure 平台、Hadoop(Apache 基金会)、阿里巴巴的阿里云、AbiCloud(Abiqio 公司)、谷歌 Google AppEngine、Amazon 的 EC2 和中国移动的 BigCloud 等^[6]。

文中采用 Hadoop 平台,实现分布式文件系统,简称 HDFS。HDFS 具有高容错性,使其可在廉价的硬件设备上部署,不仅提供了适合那些拥有大数据集的应

用平台,而且还提供了高吞吐率的数据读写。

1)HDFS 文件系统特点。

(1)HDFS 源代码是全球开发者共同研究探讨,是开放的,不收取任何费用。

(2)支持超大规模数据集。HDFS 经过 Yahoo 公司的验证,能够支持大规模数据集^[11]。HDFS 上的文件由几 GB 到几 TB,甚至到 PB 级别大小不等。

(3)HDFS 认为硬件错误不是异常,是常态。对每个部分都有出现错误的可能性,HDFS 的一个重要特点是它的故障检测和快速自动恢复^[5]。

(4)具有很强的扩展性。HDFS 中的节点增加、退出时,都不会影响用户的正常使用。

(5)HDFS 容错能力很强。HDFS 以块序列形式存储每一个文件,HDFS 存储众多的文件,设计在集群中的大量机器之间。

(6)HDFS 平台移植性。HDFS 运行在 Windows 和 Linux 等操作平台,是由 Java 语言实现的,具有很好的可移植性^[12]。

(7)提高数据访问性能。HDFS 设计不是用户交互处理,从而提高了吞吐量,节约收发数据的时间,数据就不会在网络上造成堵塞。

(8)HDFS 简化文件访问模型。文件一旦被创建,就不能再进行修改,以提高吞吐量,增加文件的一致性^[13]。

2)HDFS 系统架构。

HDFS 集群是一种典型的主从式架构,是由多个 DataNode 的数据节点和一个 NameNode 的控制节点组成。HDFS 群中还有一个 SecondaryNameNode,它配合 NameNode 进行检查点(checkpoint)操作,不与 DataNode 或 Client 交互。NameNode 将文件系统的元数据存储在内存之中,维护整个文件系统的命名空间,以控制和管理文件的相关操作,见图 1。

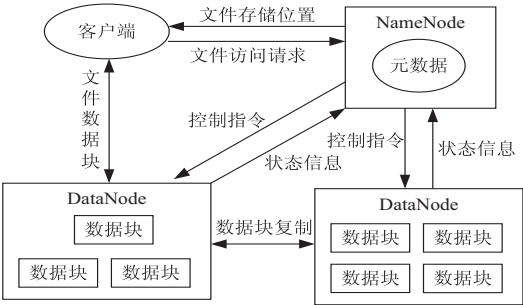


图 1 HDFS 系统架构示意图

HDFS 的系统架构中,单一控制节点 NameNode 提供一个统一文件命名空间,负责对文件系统树、目录和文件等元数据信息的维护和管理;众多数量的数据节点 DataNode 具有计算能力,管理自身携带的存储资源,也具有存储能力。在 HDFS 中,一个文件将被分成

若干个数据块,数据块能够通过 DataNode 相互之间进行复制,这些数据块被分散在不同的 DataNode 上存储。DataNode 与 NameNode 定期保持联系,以便 NameNode 掌握自身的工作状态^[14]。由 NameNode 向 DataNode 发送控制指令,例如,对文件数据块的创建、复制和删除等操作。

5 云存储在整合高校资源的体系结构

5.1 云存储体系结构

云存储的体系结构由应用层、平台层、资源层组成,如图 2 所示。

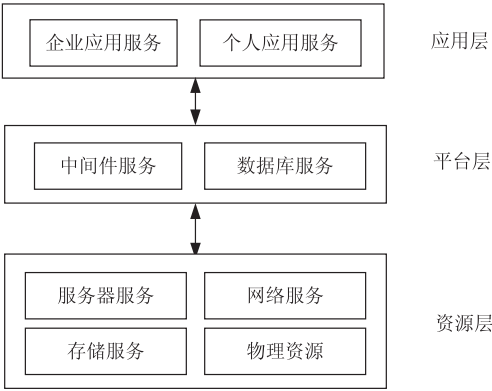


图 2 云计算体系结构

(1)应用层。主要为用户提供相关软件服务。

(2)资源层。主要是指基础架构的云计算服务,用户可以基于这些服务来搭建自己的应用,它把基础架构的各种功能都提供给用户。这种服务隐藏物理资源的复杂性,可提供虚拟化的资源。

(3)平台层。中间件服务为用户提供可扩展的事务处理中间件等服务。数据库服务为用户提供可扩展的数据库处理能力。

5.2 高校教育资源云结构

首先建立云计算数据中心,整合高校教育资源,通过虚拟化技术来提高虚拟化资源池和硬件设备的利用率。文中采用 Hadoop 平台实现数据传输^[15],具体如图 3 所示。

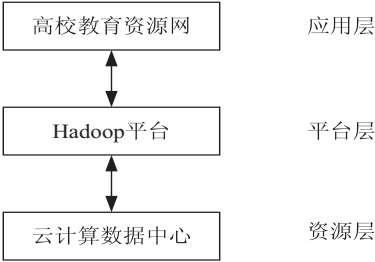


图 3 高校教育资源云结构

5.3 基于 Hadoop 平台的高校教育资源网架构

用户想从高校教育资源网上得到自己要的资源,首先要向云计算服务的提供商发送请求,从云存储里

通过 NameNode 找到相应 Block 的 ID,并根据 Block 的 ID 获得自己所要的数据资源,具体如图 4 所示。

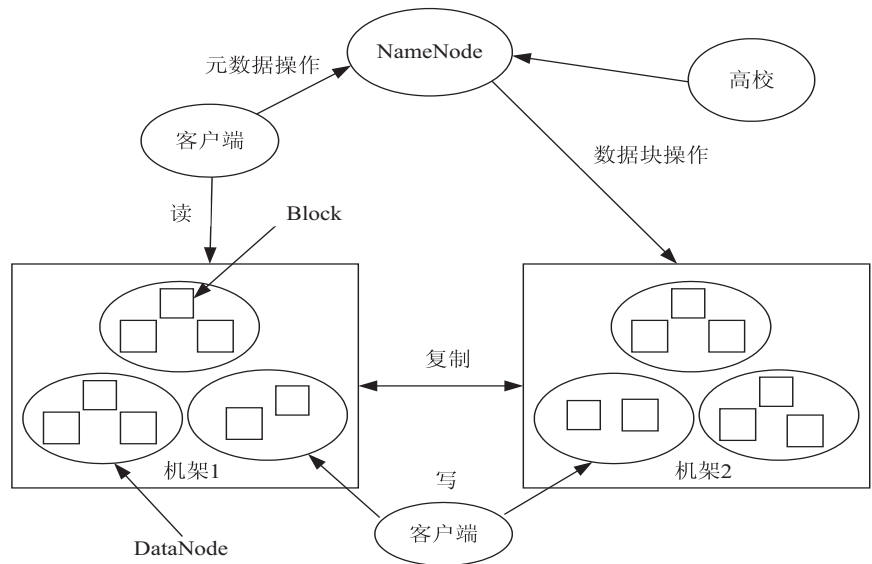


图 4 基于 Hadoop 平台高校教育资源的架构

由图 4 可知,高等学校相当于 DataNode,由所在高等学校管理员向云存储服务系统发送数据,为节省高校服务器的存储空间,由管理员决定是否删除已发送的数据。NameNode 把数据发送给云服务提供商的 DataNode,同时由 NameNode 进行记录,来保存到映像文件里,由 Secondary NameNode 来对 NameNode 中的元数据进行相应备份,并把数据备份文件更新在 NameNode 中,保存最新的映射文件和事务日志。当用户需要数据时,要向云服务提供商发送数据请求,NameNode 根据映射文件找到 Block 编号,发送给 Client,DataNode 数据根据用户发送的对应编号找到 Block,最后,再把数据发送给客户端。

可以通过 Hadoop 平台模拟实现由高校向 NameNode 发送数据的传输过程,NameNode 从 HDFS 拿到数据,该数据由 DataNode 上传到 HDFS。

在 Hadoop API 里涉及到如下接口:

```
public void mkdir(String src)
throws IOException{
//Create the given dir
//Throws:
//IOException
}
```

后台代码的编写是实现云存储平台将数据文件上传到 HDFS 集群上的关键。使用 HDFS 中的类 DistributedFileSystem 的 copyFromLocalFile 方法,使用小文件处理模块,来调用 HDFS API,将本地文件上传到 HDFS 上,根据所给的文件路径,实现文件由本地用户上传到 HDFS 集群。通过调用该类的相关方法来实现对元数据的相关操作。客户端要想访问 HDFS 集群,必须使用该类来实现相关文件操作^[16]。

在云存储平台中,用户根据需要下载或上传文件。在 Java 中,通过下面的代码,实现文件的上传功能。

```
public static void uploadLocalFile2HDFS( String s, String d)
throws IOException
{
Configuration config = new Configuration();
FileSystem hdfs = FileSystem.get( config );
Path src = new Path(s);
Path dst = new Path(d);
hdfs. copyFromLocalFile( src, dst);
hdfs. close();
}

public static void uploadLocalFile2HDFS( String s, String d)
throws IOException
{
Configuration config = new Configuration();
FileSystem hdfs = FileSystem.get( config );
Path src = new Path(s);
Path dst = new Path(d);
hdfs. copyFromLocalFile( src, dst);
hdfs. close();
}
```

在 Java 中,通过如下代码读取文件:

```
public static byte[] readHDFSFile( String dst) throws Exception
{
Configuration conf = new Configuration();
FileSystem fs = FileSystem.get( conf );
Path path = new Path( dst );
if ( fs. exists( path ) )
{
FSDataInputStream is = fs. open( path );
FileStatus stat = fs. getFileStatus( path );
```



```
byte[] buffer = new byte[Integer.parseInt(String.valueOf(
stat.getLen()))];
is.readFully(0, buffer);
is.close();
fs.close();
return buffer;
}
else
{
throw new Exception(" the file is not found. ");
}
}
```

以上代码,描写了一个 DataNode 向 NameNode 发送数据和接收数据的过程,与高校向云服务提供商发送数据的基本原理相同。

6 结束语

云存储作为一个新兴的概念,运用云存储的技术和方法整合高校资源,能够有效地提高高校资源的利用率,节约高校教育资源。构建高校教育资源共享云平台,结合云计算技术,对其体系架构和实现进行深入剖析,在 Hadoop 平台下,实现数据的传输与存储。该平台的实现借助于现有较成熟的云平台应用实例,需要多租户、分布式计算和虚拟化等多种技术之间的协同,构建高校教育资源共享云服务平台。此方案能够为高校资源的共建和共享提供一个新思路。

参考文献:

- [1] 朱晓蒙. 基于云存储的高校教育资源平台[D]. 西安: 西安科技大学, 2012.
- [2] 沈丹丹, 江正. 对高校教育信息资源共享策略的思考[J]. 软件导刊·教育技术, 2010(11): 80-82.

(上接第 175 页)

- preservation of interactive multimedia performances[C]//Proceedings of the international computer music conference. Belfast; [s. n.], 2008.
- [11] Mikroyannidis A, Theodoulidis B. Ontology management and evolution for business intelligence[J]. International Journal of Information Management, 2010, 30(6): 559-566.
 - [12] 李伟超. 数字资源长期协同保存问题研究[J]. 创新科技, 2013(7): 58-59.
 - [13] Bao J, Hu Z, Caragea D, et al. A tool for collaborative construction of large biological ontologies[C]//Proc of 17th international conference on database and expert systems applications. Krakow, Poland; [s. n.], 2006.
 - [14] Mikroyannidis A. A semantic framework for cloud learning en-

- [3] 褚军亮, 宋立荣. 云计算用于网络科技信息资源整合服务初探[J]. 中国科技资源导刊, 2010(6): 42-47.
- [4] 王文清, 陈凌. CALIS 数字图书馆云服务平台模型[J]. 大学图书馆学报, 2009, 27(4): 13-18.
- [5] 高宏卿, 汪浩. 基于云存储的教学资源整合研究与实现[J]. 现代教育技术, 2010, 20(3): 97-101.
- [6] 王胜川. 基于云计算的存储技术研究[J]. 石油工业计算机应用, 2011(2): 53-55.
- [7] 张建勋, 古志民, 郑超. 云计算研究进展综述[J]. 计算机应用研究, 2010, 27(2): 429-430.
- [8] 张兴. 基于 Hadoop 的云储存平台的研究与实现[D]. 成都: 电子科技大学, 2013.
- [9] Likhterov B, Kopeika N S. Motion-Blurred image restoration using modified inverse all-pole filters[C]//Proceedings of SPIE. [s. l.]: [s. n.], 2002: 56-62.
- [10] Chen Kang, Yin Yubing, Zheng Weimin. Teaching large scale data processing: the five-week course and two years' experiences[C]//Proc of the first ACM summit on computing education. Beijing, China: ACM, 2008.
- [11] White T. Hadoop The definitive guide[M]. [s. l.]: O'Reilly Media, Inc., 2009.
- [12] Sultan N. Cloud computing for education: a new dawn? [J]. International Journal of Information Management, 2010, 30(2): 109-116.
- [13] Holger Scroll Stefan Wind. Requirements engineering for cloud computing[J]. Journal of Communication and Computer, 2011, 8(9): 707-715.
- [14] Bordas M. Application of cloud computing in the public sector in hungary[J]. Journal of US-China Public Administration, 2012, 9(5): 534-551.
- [15] 李伟林, 陈成. 广播电视大学云计算教学平台的构建探索[J]. 湖南广播电视大学学报, 2012(1): 5-8.
- [16] 喻昕, 王敬一. 基于云计算技术的数字图书馆云服务平台架构研究[J]. 情报科学, 2011, 29(7): 1049-1053.


vironments[M]//Cloud computing for teaching and learning: strategies for design and implementatio. [s. l.]: IGI Global, 2012: 17-31.

- [15] 王兴, 何婷婷, 庄超. 本体演化及本体的版本管理机制研究[J]. 计算机与数字工程, 2006, 34(7): 7-10.
- [16] 徐静, 宋文. 本体演化管理的方法和关键技术研究进展[J]. 图书情报工作, 2011, 55(14): 122-124.
- [17] 薛建武, 赵娜, 高俊萍, 等. 本体演化过程中的实例完备性研究[J]. 情报杂志, 2013, 32(9): 175-179.
- [18] Mikroyannidis A. Evolving e-learning ontologies for personal and cloud learning environments[C]//Proc of 2011 seventh international conference on signal image technology & internet-based systems. [s. l.]: [s. n.], 2011.

基于Hadoop高校教育资源云存储平台构建研究

作者：[王凤领](#)，[WANG Feng-ling](#)

作者单位：[黑龙江财经学院 财经信息工程系, 黑龙江 哈尔滨, 150025](#)

刊名：[计算机技术与发展](#)

英文刊名：

年，卷(期)：2016, 26 (3)

引用本文格式：[王凤领](#), [WANG Feng-ling](#) [基于Hadoop高校教育资源云存储平台构建研究](#)[期刊论文]-[计算机技术与发展](#) 2016 (3)