

基于云计算的 OA 资源数据交换模式探讨

毛海波

(宁波大学 图书馆与信息中心, 浙江 宁波 315211)

摘要: 通过分析现有 OA 资源系统数据交换模式, 发现在进行海量 OA 资源数据交换时存在性能差、速度慢和效率低等问题. 文中提出云计算环境下的 OA 资源数据交换平台功能、结构和交换模式, 针对性地解决海量 OA 资源数据交换问题. 首先分析 OA 资源交换云应用系统体系, 设计 OA 资源交换云平台的子系统和功能. 其次结合云计算和大数据技术, 将 OA 资源交换云平台子系统部署到计算与存储、大数据管理、数据服务和 OA 资源交换四层结构中, 形成虚拟的数据交换平台. 最后设计 OA 资源交换云平台的数据交换模式, 以交换域的方式实现 OA 资源的云交换.

关键词: 开放存取; 云计算; 数据交换; 云交换

中图分类号: G250.76

文献标志码: A

文章编号: 1001-5132 (2016) 01-0028-05

开放存取(Open Access, OA)作为一种公开、高效的数字资源传播新模式^[1], 受到科技工作者、图书馆员、学术团体等积极参与和支持. OA 资源作为免费、共享的数字资源能更快、更广泛、更自由的传播, 更多地被他人引用, 更好地推动知识创新和学术交流^[2]. OA 资源系统作为 OA 资源获取、存储和访问的服务平台, 为众多用户提供全文检索、浏览、下载和打印等服务^[3].

目前, 互联网已经进入大数据时代^[4], OA 资源呈现出高速增长的趋势. 根据 DOAJ 统计, 近几年 OA 资源的年增长率保持在 15%以上, 每天平均新增 OA 资源高达 9 804 个^[5]. 为此, OA 资源的收割、整理、存储、发布、检索及下载等变得更加频繁, 对 OA 资源系统的数据交换性能、速度和效率提出更高的要求. 云计算技术的出现, 为提高 OA 资源数据交换的性能、速度和效率提供了新的技术方案. 本文就对云环境下的 OA 资源数据交换平台的结构、功能以及交换模式等方面进行阐述.

1 现有 OA 资源数据交换模式分析

1.1 OA 资源系统及分类

当前, 国内外的高校图书馆和学术机构已经

建立了功能各异的 OA 资源系统. 按照 OA 资源系统的应用类型, 可以大致分为 4 种 OA 资源类型: OA 资源导航型、OA 资源仓储型、OA 资源整合型和 OA 资源收割型.

(1) OA 资源导航型. 主要是国内高校图书馆网站中的 OA 资源导航网页, 如西南交通大学图书馆、安徽大学图书馆、天津高等教育文献信息中心、宁波大学图书馆等 OA 资源导航网页.

(2) OA 资源仓储型. 遵循 DC 元数据规范, 对论文、报告等文献以电子文档形式进行统一存储和管理, 并以机构知识库形式提供给用户检索、下载等服务. 其代表性系统有中国科技论文在线^[6]、奇迹文库、中国预印本服务系统、香港科技大学科研成果全文仓储、中国学术会议在线、中国开放式教育资源共享组织、NTIC 国外预印本门户等.

(3) OA 资源整合型. 对 OA 资源系统进行整合和集成, 提供统一检索、下载等服务, 并实现各类 OA 资源的动态链接. 其代表性系统有 CNKI 网络资源共享平台、上海交通大学的学术信息资源检索系统(AIRA)、佛罗里达州立大学的 D-Scholarship 系统、麻省理工学院的 D-Space 系统^[7]、Endeavor Information Systems 公司的 Encompass、ExLibris

公司的 Metalib 系统、Innovative 的 MAP 系统等。

(4)OA 资源收割型。遵循 OAI-PMH 协议, 收割 OA 资源仓储元数据, 加工后提供检索、下载等服务。其代表系统有 OAIster 检索平台、奥尔特多米尼大学的 ARC 和 DP9、CitebaseSearch、Scirus、Google Scholar 系统^[8]等。

1.2 现有 OA 资源数据交换模式

通过对上述 OA 资源系统的数据交换模式分析, 归纳起来可以分为 3 种数据交换方式: 点对点数据交换、基于消息中间件的数据交换和基于 Web 服务的数据交换^[9]。

(1)点对点数据交换。它需要对不同的 OA 资源系统开发特定的接口程序, 接口程序的多少与需要连接的系统数有关。如果某个 OA 资源系统发生变化, 相应的接口都要进行调整, 这使得整个交换平台的结构越来越复杂, 维护成本不断增加, 难度越来越大。其代表性系统包括奇迹文库、中国预印本服务系统、中国学术会议在线、D-Space 系统等。

(2)基于消息中间件的数据交换。它是利用消息传送机制和消息队列模式组成的中间件技术, 以消息形式实现 OA 资源系统间的数据交换。它具有松耦合、异步可靠和持久化等特点, 但因消息中间件体系结构扩展性差, 难以满足大规模网络环境下 OA 资源系统间的数据交换需求。其代表性系统包括 CNKI 网络资源共享平台、OAIster、Google Scholar 等。

(3)基于 Web 服务的数据交换。利用 Web 服务技术, 对 OA 资源系统的接口进行定义、描述和发现, 通过服务的请求和调用来实现 OA 资源系统间的数据交换和共享。它的特点是基于开放的标准协议, 优点是实现异构平台间的数据交换。但是它面对大数据量和高速存取时, OA 资源系统间的数据交换性能和效率大大降低。其最具代表性系统是 Metalib 系统。

通过对上述数据交换技术分析和比较, 发现这 3 种数据交换模式都存在特定的问题, 导致数据交换平台都是专用的, 在系统的开放性、资源的共享性及业务的协同性方面也存在问题。面对大规模的数据数据交换时, 数据交换性能、速度和效率不高。因此引入云计算技术, 构建专用于 OA 资源交换的数据交换平台, 能有效解决现有 OA 资源数

据交换模式所存在的问题, 大幅度提升数据交换效率和性能。

2 OA 资源数据云交换模式

2.1 OA 资源数据云交换应用体系

OA 资源数据云交换应用体系框架包括外部 OA 资源系统、OA 资源应用服务、OA 数据交换平台和 OA 资源库。OA 资源数据交换云体系应用框架如图 1 所示。

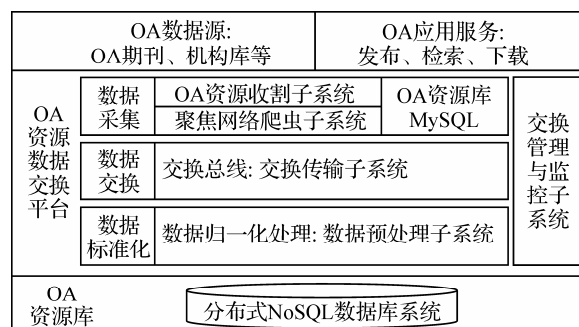


图1 OA 资源数据云交换体系应用框架

外部 OA 资源系统为数据交换提供数据源, 它包括 OA 期刊、机构知识库、学科知识库以及网络资源等 OA 资源。

OA 资源应用服务为用户提供结构化、标准化的 OA 资源信息, 通过 OA 资源门户为用户提供全文检索、浏览、下载和打印等服务。

OA 资源数据交换平台是数据交换的核心, 它包括负责 OA 资源采集的 OA 资源收割子系统和聚焦网络爬虫子系统, 用于数据交换的交换传输子系统, 实现数据标准化的数据预处理子系统, 以及全程监控平台运作的交换管理与监控子系统。

OA 资源库为原始数据、预处理数据和 OA 元数据提供存储好管理服务, 它包括 MySQL 数据库系统和分布式 NoSQL 数据库系统。

2.2 OA 资源数据云交换平台功能

OA 资源数据交换平台通过 OA 资源收割子系统和聚焦网络爬虫子系统获取 OA 资源, 并存储到前置 OA 资源库中, 在交换管理子系统的流程控制下, 通过数据预处理子系统、交换传输子系统, 将标准化、格式化的 OA 资源信息定向传输到 OA 资源库中。同时, 交换传输子系统还为 OA 资源应用服务提供数据访问接口。

(1)OA 资源收割子系统。遵循 OAI-PMH 协议

的 OA 期刊、机构知识库和学科知识库由 OA 资源收割子系统统一处理,对收割过来的 OA 资源存储到前置 OA 资源库 MySQL 中,等待数据清洗、转换处理等数据预处理。

(2) 聚焦网络爬虫子系统. 对分布在网络中的 OA 期刊和其他网络资源,按照 OA 资源元数据定义格式,通过聚焦网络爬虫子系统进行收割,统一存储到前置 OA 资源库 MySQL 中,等待数据清洗、转换处理等数据预处理。

(3) 数据预处理子系统. 按照期刊论文、学位论文、会议论文、电子图书等 OA 资源元数据定义,对前置 OA 资源库中的数据进行数据清理、变换、集成以及归一化处理,将标准化、格式化的 OA 资源信息存储到 OA 资源库 NoSQL 中。

(4) 交换传输子系统. 采用 SOA 和 ESB 服务总线技术构建数据交换传输子系统,即数据总线服务系统,遵循 XML 数据交换格式,为 OA 资源收割、数据预处理及应用服务提供数据交换通道,实现交换数据的打包、转换、传递、路由、解包等功能。

(5) 交换管理与监控子系统. 作为交换平台的指挥和管理中心,监控和协调 OA 资源收割和数据预处理子系统的有效运行,管理 OA 资源收割原则、数据交换规则、OA 资源数据标准等信息。

(6) OA 资源库. MySQL 数据库系统作为前置 OA 资源库,用来存储和管理收割过来的 OA 资源数据. 分布式 NoSQL 数据库系统作为 OA 资源库,用来存储和管理标准化后的 OA 资源数据。

2.3 云环境下的 OA 资源数据交换平台

OA 资源数据交换云平台引入云计算和大数据相关技术^[10-11],构建成低成本、高可靠性、高扩展性和高性能以及集成数据采集、传输、存储等功能的数据交换平台. 利用云计算服务模式,将 OA 资源交换云平台设计为计算与存储管理平台(IaaS)、大数据管理平台(PaaS)、数据服务平台(DaaS)和 OA 资源交换平台(EaaS)四层结构. 整体的交换云平台技术框架如图 2 所示。

(1) 计算与存储管理平台(IaaS). 通过 Hadoop 的分布式文件管理系统,对基于 Linux 操作系统的 X86 服务器集进行集群管理,形成庞大的计算池和存储池,为 OA 资源收割子系统、聚焦网络爬虫子



图2 OA资源交换云平台技术框架

系统、交换传输子系统、数据预处理子系统、交换管理与监控子系统以及 OA 资源库等提供弹性的计算单元和存储单元。

(2) 大数据管理平台(PaaS). 利用 Hadoop 提供的 Hbase、MapReduce、ZooKeeper 等功能,对海量的 OA 资源数据进行分布式的存储管理、协同管理和分布式计算,统一管理 MySQL 数据库系统和分布式 NoSQL 数据库系统的计算和存储。

(3) 数据服务平台(DaaS). 从 PaaS 平台中分离出来,专注于数据转化、数据清洗、数据规格化、数据挖掘、数据服务等操作. 通过数据预处理子系统对海量 OA 资源进行过滤、清洗、映射、转换的数据处理,并将标准化、格式化的 OA 资源数据存储到 OA 资源库中. 同时,为应用服务提供数据访问的 API 和 Web 服务数据接口。

(4) OA 资源交换平台(EaaS). 从 DaaS 平台中分离出来,专注于数据采集、数据交换、数据监控等操作. 利用大数据和分布式技术,建立包括分布式消息、并行适配器、分布式 ESB 总线功能的交换传输子系统,为 OA 资源收割子系统、聚焦网络爬虫子系统、数据预处理子系统提供数据交换服务. 并通过交换管理与监控子系统统一管理 OA 资源收割子系统、聚焦网络爬虫子系统、数据预处理子系统以及 OA 资源库的数据交换规则,同时监控各子系统的性能。

云环境下的 OA 资源数据交换平台,实现物理交换平台向虚拟交换平台的跨越. 通过云计算技术,在 1 个物理平台上建立多个相互独立 OA 资源虚拟交换平台,每个 OA 资源虚拟交换平台独立完成一组 OA 资源收割、交换、转换、存储等数据交

换任务。

2.4 OA 资源数据云交换模式

具有虚拟化、超计算能力、高可靠性、高扩展性等特点的 OA 资源虚拟交换平台与现有的 OA 资源系统数据交换模式有很大的差别, 在数据交换性能、速度和效率上有显著的提高。云环境中的 OA 资源虚拟交换平台我们定义为交换域, 即 1 个虚拟交换平台就是 1 个交换域。OA 资源数据交换平台可由 1 个或多个交换域组成, OA 资源云交换平台的数据交换模式如图 3 所示。

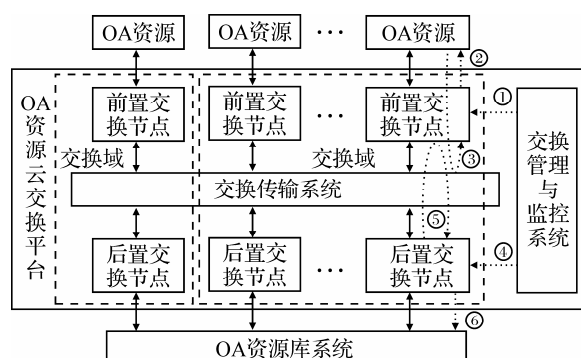


图3 OA 资源交换云平台数据交换

交换域由交换节点、交换传输系统和交换管理与监控系统组成。交换节点分为前置交换节点和后置交换节点, 其中, 前置交换节点包括 OA 资源收割子系统、聚焦网络爬虫子系统和前置 OA 资源库, 后置交换节点包括数据预处理子系统和 OA 资源库, 在 1 个交换域内可以设计 1 个或多个交换节点, 交换节点实现点对点的数据交换。交换传输系统是基于 SOA 架构的 ESB 数据传输总线, 实现 OA 资源的打包、转换、传递、路由、解包等功能, 交换传输系统是基于 Web 服务的数据交换。交换管理与监控系统是对数据交换过程的流程设置、部署和执行以及交换云的运行情况监控和分析, 交换管理与监控系统主要采用消息中间件的数据交换。

因此, 交换域的数据交换模式既有点对点模式, 又有 Web 服务模式, 还有消息中间件模式。云环境下 OA 资源虚拟交换平台的数据交换模式是根据 OA 资源数据交换平台子系统的自身特性, 充分发挥各子系统数据交换的性能和效率, 实现数据协同交换。其具体的数据交换过程如图 3 所示, 具体的数据交换步骤如下:

(1) 交换管理与监控系统向 OA 资源收割系统 (OA 资源收割子系统或聚焦网络爬虫子系统) 发出收割 OA 资源指令。

(2) OA 资源收割系统根据 OA 资源收割指令和地址开始采集 OA 资源数据。

(3) OA 资源收割系统采集到 OA 资源数据后, 通过交换传输系统提供的 OA 资源写入接口, 将数据保存到前置 OA 资源库 MySQL 数据库中, 等待数据预处理系统处理。

(4) 交换管理与监控系统向数据预处理系统发出清理、变换、归一化、标准化等 OA 资源数据处理指令。

(5) 数据预处理系统通过交换传输系统提供的 OA 资源读取接口, 将数据保存在前置 OA 资源库中数据取出, 再根据交换管理与监控系统提供数据处理规则和要求对数据进行处理。

(6) 数据预处理系统将标准化、格式化的 OA 资源信息存储到 OA 资源库 NoSQL 数据库中。

2.5 OA 资源数据云交换实现

利用 3 台服务器、Hadoop 系统、分布式爬虫和分布式数据处理系统组建小型的 OA 资源数据云交换平台, 具体的平台构成见表 1。

OA 资源数据云交换平台采用 Master/Slave 架构, 即它包含 1 个 Master 节点, 2 个 Slaves 节点, 由 Master 对 Slaves 进行统一管理。其中, 1 个 Slaves 节点运行分布式爬虫系统, 对中国科技论文在线

表1 小型的 OA 资源数据云交换平台

序号	名称	功能描述	数量
1	服务器	CPU 1 个, RAM 2 GB, 硬盘 146 GB	3
2	操作系统	64 位 ubuntu 14.04.1 的 Linux 操作系统	3
3	Hadoop 系统	Hadoop-2.6.0, jdk1.8.0-40, hbase-1.0.0, hive-0.14.0, sqoop-1.4.4	1
4	分布式爬虫系统	Ebot 分布式开源爬虫系统	1
5	分布式数据预处理系统	根据 OA 资源元数据定义, 自主研发。	1
6	开发工具	MyEclipse8.6	1

(<http://www.paper.edu.cn>)上的期刊论文数据进行收割;另1个Slaves节点运行分布式数据预处理系统,用于OA数据的清理、变换、归一化、标准化等处理。通过本数据云交换平台的实验,证明云环境下的数据交换平台比传统的数据交换平台的综合性能提高2倍以上。

3 结语

采用云计算架构的OA资源数据交换平台,有效解决大型OA资源系统访问、存储以及与其他OA资源系统数据交换的性能、速度和效率瓶颈问题,提高OA资源系统运行效率。同时,降低硬件资源需求和运维成本。利用云计算的SaaS服务模式,根据用户OA资源交换申请要求,在云计算的虚拟环境中分配一组独立的OA资源数据交换服务,按照预先设定的数据交换规则,自动实现OA资源数据交换。最终,实现用户个性化的OA资源访问和数据交换服务。

参考文献:

- [1] OSF. Open Access Initiative[EB/OL]. [2014-11-15]. <http://www.opensocietyfoundations.org/grants>.
- [2] MAVODZA J. A review of the open access concept in the UAE[J]. New Library World, 2013, 114:259-266.
- [3] OJSTERSEK M, BREZOVNIK J, KOTAR M, et al. Establishing of a Slovenian openaccess infrastructure: a technical point of view[J]. Electronic Library and Information Systems, 2014, 48:394-412.
- [4] 维克托·尔耶·舍恩伯格. 大数据时代[M]. 杭州: 浙江人民出版社, 2012:75-126.
- [5] CHADWELL F, SUTTON S C. The future of open access and library publishing[J]. New Library World, 2014, 115:225-236.
- [6] 教育部. 中国科技论文在线[EB/OL]. [2014-10-23]. <http://www.paper.edu.cn/>.
- [7] DURASPACE. DSpace[EB/OL]. [2014-09-10]. <http://www.dspace.org/>.
- [8] GOOGLE. Google 学术搜索[EB/OL]. [2014-12-01]. <http://scholar.google.com/>.
- [9] 董蕙亚, 薄剑勇, 贾学敏. 地方政务数据交换云架构研究与设计[J]. 宁波大学学报(理工版), 2011, 24(3):25-28.
- [10] MILLER M. 云计算[M]. 北京: 机械工业出版社, 2009: 55.
- [11] 孟小峰, 慈祥. 大数据管理: 概念、技术与挑战[J]. 计算机研究与发展, 2013, 50(1):146-170.

Probe: OA Resource Data Exchange Mode Based on Cloud Computing

MAO Hai-bo

(Library and Information Center, Ningbo University, Ningbo, 315211, China)

Abstract: By analyzing the existing data exchange mode in OA resource system, the poor performances characterized with low efficiency are identified in the process of the massive OA resource data exchange. In this paper, a platform design for OA resource data exchange in cloud computing environment is presented, in which OA resources are used to solve the massive data exchange problems. First, the OA resource cloud exchange of application system architecture is analyzed, the OA resource cloud switching subsystem and features of the platform are designed. Second, through the combination of cloud computing and big data technology, the OA resource exchange platform subsystems is deployed for purposes including cloud computing and storage, big data management, OA cloud resource exchange and data services in the four-layer structure. Third, the OA resource cloud exchange platform's data exchange mode is designed for cloud exchange of OA resources by switching domain.

Key words: open access; cloud computing; data exchange; cloud exchange

(责任编辑 章践立)