RMSCloud与科技文献云服务

吴广印
（中国科学技术信息研究所，北京 100038）

摘 要：在对云计算相关功能定义进行研究和分析的基础上，针对科技文献的云服务需求结合云计算的相关应用，介绍RMSCloud的相关核心技术及特点。最后基于RMSCloud对科技文献云服务应用的集成系统架构进行阐述。

关键词：RMSCloud；云计算；云服务；科技文献服务；云服务集成
中图分类号：G35 **文献标识码：**A DOI：10.3772/j.issn.1674-1544.2013.05.013

RMSCloud and S&T Document Cloud Service

Wu Guangyin
(Institute of Science and Technology Information Research of China, Beijing 100038)

Abstract: In this paper, the main definition of cloud computing related function has carried on the research and analysis, at the same time cloud service demand of the scientific documents, combined with the related application of cloud computing, this paper introduces the RMSCloud related core technologies and features. Finally introduced the RMSCloud based integration of scientific and technological documents cloud service application system architecture

Keywords: RMSCloud, cloud computing, cloud services, S&T document service, cloud service integration

1 引言

RMS是北京万方数据股份有限公司开发的一个统一的非结构化资源服务系统的简称。RMSCloud是资源服务系统的云计算服务缩写。RMSCloud在RMS架构基础上以云计算的技术架构为指导面向科技文献云服务的学术搜索引擎，在国家“863”课题“以科技文献为主的搜索引擎研制”资助下，历经2年多时间研制完成。

所谓“云计算”服务，就是直接为用户提供功能服务，而用户不必考虑平台、系统、应用软件甚至公共服务数据来源。用户在需要某种服务时，只需向服务提供商支付一定的服务费，即可获取这种直接的服务。显然“云计算”的服务模式是一种硬件、软件、系统资源的共享服务模式。云计算的最终目的是将计算、服务和应用作为一种公共设施提供给公众，从而大大提高资源的利用率。

在云计算环境下，用户的使用观念也会发生彻底的变化：从“购买系统”向“购买服务”转变，因为他们直接面对的将不再是复杂的硬件和软件，而是最终的服务。用户不需要拥有看得见、摸得着的硬件设施，也不需要为机房支付设备供电、空调制冷、专人维护等高昂费用，更不需要等待漫长的供货周期以及项目实施等冗长的时间，而只需要和云计算服务提供商签订服务合同，即可得到需要的直接服务。目前，由北京万方软件有限公司提供的“中国学术搜索网”云服务接口可为广大科技信息服务机构提供一体化的“科技文献搜索云服务”，从而最大限度地节约投资，提高服务效率。

作者简介：吴广印（1965- ），男，中国科学技术信息研究所研究员，北京万方软件有限公司董事长，研究方向：非结构数据库管理系统、中文信息检索。

基金项目：国家高科技发展计划（863计划）“云计算关键技术与系统（一期）”专项“以科技文献为主的搜索引擎研制”（2011AA01A206）。
收稿日期：2013年6月26日。

2 云计算及其核心技术

美国国家标准与技术研究所（NIST）提出的云计算的定义如下[1]：云计算是“一种无处不在且方便使用的计算模式，可按网络访问需求自动配置的计算资源共享池（例如网络、服务器、存储、应用程序和服务），可以最小的管理代价快速配置管理和发布资源，并且支持资源服务商和服务供应商的互动”。NIST提出云计算具有按需自助服务、宽带网络接入、资源池、快速弹性、量化服务等5个基本特征，软件即服务、平台即服务、基础设施即服务等3种服务模式，私有云、社区云、公有云、混合云等4类部署形式。

图1是国际上对3种不同云服务模式的用户控制权限的说明，其中打包软件为传统用户私有设施形式。

云计算系统运用了许多技术，其中以编程模型、数据管理与挖掘技术、数据存储技术、虚拟化技术、云计算平台管理技术最为关键。

（1）编程模型

Map/Reduce[2]是Google开发的java、Python、C++编程模型，它是一种简化易于理解的分布式编程模型和高效的任务调度模型，用于大规模数据集（大于1TB）的并行运算。该编程模型使云计算环境下的编程十分简单。Map/Reduce模式的思想是将要执行的问题分解成Map（映射）和Reduce（化简）的方式，先通过Map程序将数据切割成不相关的区块，分配（调度）给大量计算机处理，达到分布式运算的效果，再通过Reduce程序将结果汇整输出。Map/Reduce已经成为云计算领域分布式编程核心指导思想。

（2）海量数据分布存储技术

云计算系统由大量服务器组成，同时为大量用户服务，因此云计算系统采用分布式存储的方式存储数据，用冗余存储的方式保证数据的可靠性。云计算系统中广泛使用的数据存储系统是Google的GFS和Hadoop团队开发的GFS的开源实现HDFS[3]。GFS[4]即Google文件系统（Google File System），是一个可扩展的分布式文件系统，用于大型的、分布式的、对大量数据进行访问的应用。GFS的设计思想不同于传统的文件系统，是针对大规模数据处理和Google应用特性而设计的。它运行于廉价的普通硬件上，但可以提供容错功能，可以给大量的用户提供总体性能较高的服务。

一个GFS集群由一个主服务器和大量的块服务器构成，并被许多客户访问。主服务器存储文件系统所有的元数据，包括名字空间、访问控制信息、从文件到块的映射以及块的当前位置。它也控制系

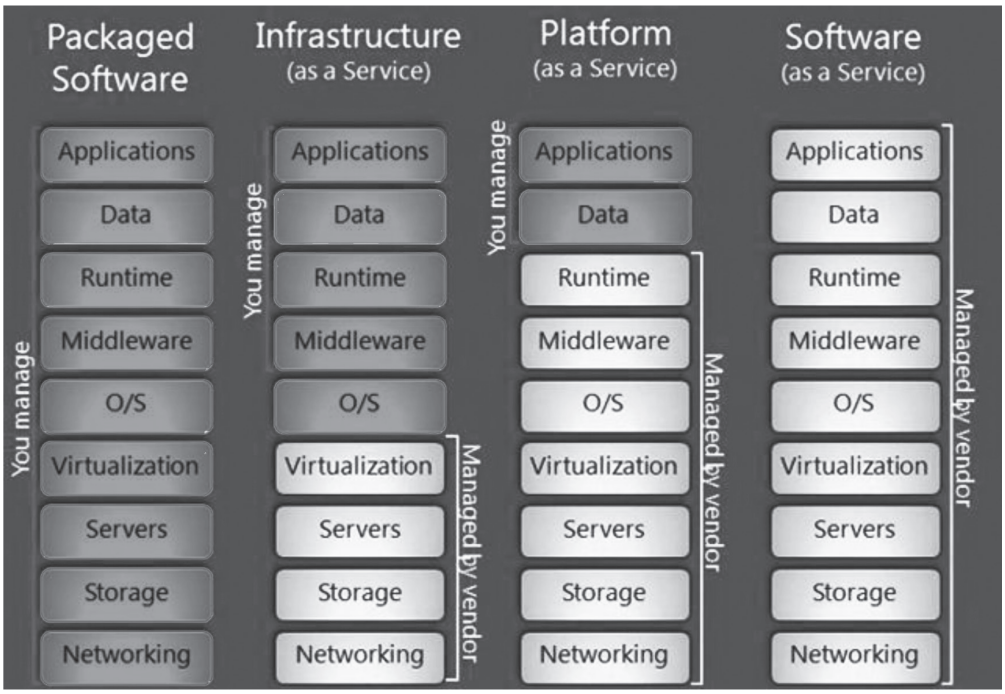

图1 云计算三种服务模式的控制权限差异（深色色块表示用户控制权限）

统范围的活动，如块租约管理、碎片数据块的整理与收集、块服务器间的块迁移。主服务器定期通过HeartBeat消息与每一个块服务器通信，给块服务器传递指令并收集它的状态。GFS中的文件被切分为64MB的块并以冗余存储，每份数据在系统中保存3个以上备份。

客户与主服务器的交换只限于对元数据的操作，所有数据方面的通信都直接和块服务器联系，从而提高了系统的效率，防止主服务器负载过重。

（3）海量数据管理技术

云计算需要对分布的、海量的数据进行处理、分析，因此，数据管理技术必需能够高效地管理大量的数据。云计算系统中的数据管理技术主要是Google的BT（BigTable）数据管理技术和Hadoop团队开发的开源数据管理模块HBase[5]。BT是建立在GFS、Scheduler、Lock Service和Map/Reduce之 上的一个大型的分布式数据库，与传统的关系数据库不同，它把所有数据都作为对象来处理，形成一个巨大的表格，用来分布存储大规模结构化数据。

Google的很多项目使用BT来存储数据，包括网页查询，Google earth和Google金融。这些应用程序对BT的要求各不相同：数据大小（从URL到网页到卫星图象）不同，反应速度不同（从后端的大批处理到实时数据服务）。对于不同的要求，BT都成功地提供了灵活高效的服务。

（4）虚拟化技术

通过虚拟化技术可实现软件应用与底层硬件相隔离，它包括将单个资源划分成多个虚拟资源的裂分模式，也包括将多个资源整合成一个虚拟资源的聚合模式。虚拟化技术根据对象可分成存储虚拟化、计算虚拟化、网络虚拟化等。计算虚拟化又分为系统级虚拟化、应用级虚拟化和桌面虚拟化[6]。

（5）云计算平台管理技术

云计算资源规模庞大，服务器数量众多并分布在不同的地点，同时运行着数百种应用，如何有效地管理这些服务器，保证整个系统提供不间断的服务是巨大的挑战。云计算系统的平台管理技术能够使大量的服务器协同工作，方便地进行业务部署和开通，快速发现和恢复系统故障，通过自动化、智能化的手段实现大规模系统的可靠运营。

3　RMSCloud涉及的核心技术

RMSCloud是在云计算技术架构基础上提供科技文献云服务的学术搜索引擎。RMSCloud云搜索引擎是基于RMS系统的变长数据存储管理、多样化索引控制技术、中文智能分词技术，实现对于中文科技文献文本信息的快速准确分词，采用独特B*树文件索引算法，进行索引构建索引文件，利用多项检索优化算法实现了基于复杂布尔表达技术的全文检索。其先进索引技术，可以使检索词快速定位，检索速度几乎不受索引文件大小的限制，为海量科技文献信息的学术搜索提供了全文索引和检索技术支持，通过跨语言自动翻译和词表扩展技术，确保实现系统的查全、查准率。同时，RMSCloud云搜索引擎采用云计算架构和并行计算技术，通过索引分片，减少单索引数据量，提高索引检索速度；通过索引副本，实现全文索引在集群多节点之间的分布，实现多节点并行计算；通过无主从集群节点通信，实现节点数据同步，为集群节点可靠并行计算与云搜索服务提供保障。

RMSCLoud云搜索引擎核心搜索服务技术框架如图2所示。RMSCloud云搜索引擎分层结构及模块组成主要包括API接口、传输协议支撑、Java Netty框架、监控、RMS中文智能分词、第三方插件支持、云集群通信、脚本解析引擎、RMS全文索引、RMS全文检索、索引映射配置、数据源、分布式RMS索引目录支持、文件系统持久化网关等模块。RMSCloud云搜索引擎在研制和构建过程中，应用了大量的云计算技术，实现集群与并行计算支持，满足大数据量科技文献学术搜索与知识挖掘分析需求。

（1）集群与分布式并行计算

RMSCloud云搜索引擎支持分布式并行计算技术，主要依赖于以下途径实现。

集群技术：集群中有多个节点，其中有一个为主节点，这个主节点可以通过选举产生，主从节点是对于集群内部来说的。对于集群外部来说，就是去中心化，从外部来看集群，在逻辑上是个整体，与任何一个节点的通信和与整个集群通信是等价的。集群节点故障不影响整个集群的对外服务，从而保证集群的可靠性。

索引分片：可以把一个完整的全文索引分成多个分片，这样的好处是可以把一个大的全文索引进行拆分，分布到不同的节点上，在检索时，依托多个节点的计算能力进行并行计算和分布式检索。

索引副本：可针对索引及分片设置多个索引

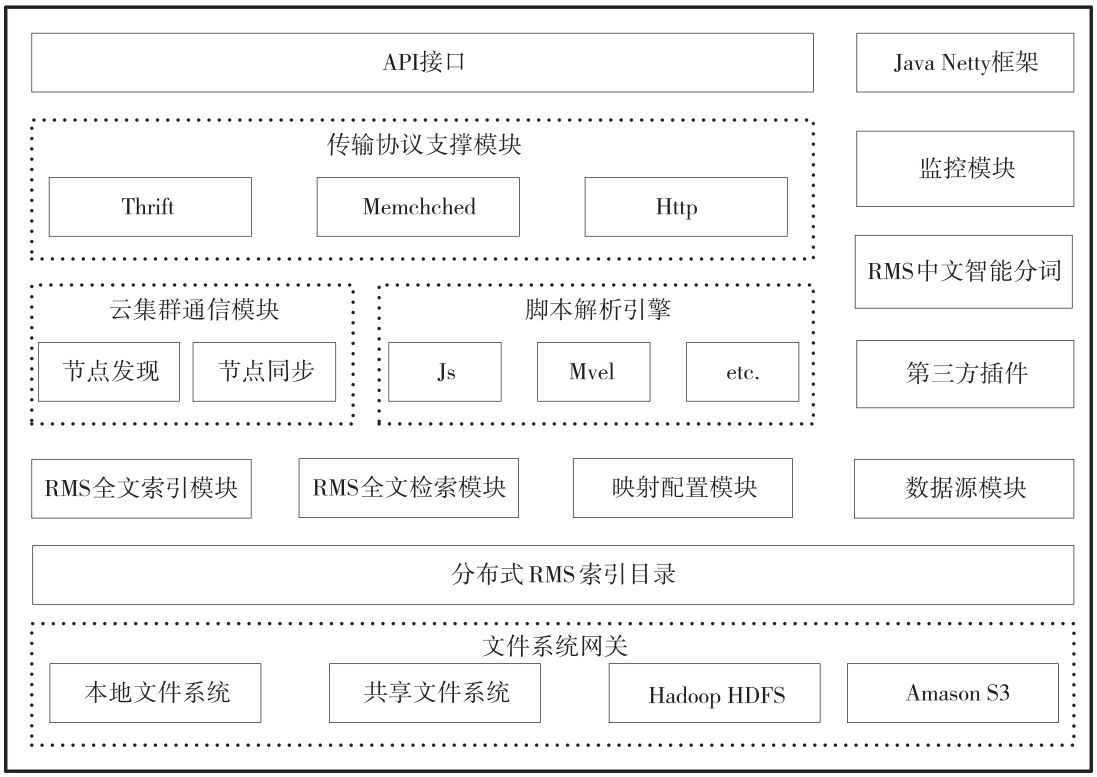

图2 RMSCLoud云搜索引擎核心搜索服务技术框架

的副本。副本的作用一是提高系统的容错性，当某个节点某个分片损坏或丢失时可以从副本中自动恢复；二是提高检索效率，可以自动对搜索请求进行负载均衡，调度到其他节点副本上进行分布式检索。

（2）自动化维护与管理

RMSCloud云搜索引擎基于分布式计算模式，支持节点自动发现、节点扩展，数据自动重新分布、索引自动持久化存储等能力，可以实现对于云计算集群的自动维护管理功能。

自动节点发现：类似一个p2p的系统，它先通过广播寻找存在的节点，再通过多播协议来进行节点之间的通信，同时也支持点对点的交互。

数据重新分布：在有节点加入或退出时会根据机器的负载对索引分片进行重新分配，挂掉的节点重新启动时也会自动进行数据恢复。

数据源自动索引更新：可支持从数据源中自动读取数据并同步索引到RMSClouded云搜索服务集群中。

索引持久化存储网关：RMSClouded默认是先把索引存放到内存中，当分配内存满了时再持久化到硬盘等存储网关。当这个集群关闭再重新启动时就会从存储网关中读取索引数据。RMSClouded支持多种类型的持久化存储网关，有本地文件系统，共享文件系统，Hadoop的HDFS和Amazon的s3云存储服务。

4 RMSCloud科技文献服务相关核心技术

在RMSCloud的研制过程中，除借鉴传统搜索引擎在云计算应用方面的思路，同时也考虑了学术搜索引擎的专有特性。

（1）词表与中文智能分词技术的大量应用

RMSCloud在数据索引、用户检索需求处理等方面大量应用了词表和中文智能分词技术，中文分词技术的好坏直接影响系统的“查全/查准率”。

万方科技文献主题词库：用于文献分词与索引构建，检索语句的分词与扩展检索、相关检索词提示等。这些主题词来源于万方数据期刊、学位论文等数据库中的作者形成的主题词项，通过二次规范加工建立。

汉语叙词表：用于对检索关键词基于词间关系，包括上位词、下位词、相关词、代用词等主题词本体扩展与相关检索、相似词推荐。该词表以中国科学技术信息研究所建立的“工程词表”为

基础，主要用于科技文献检索的后空扩展检索，在“中国学术搜索网”中得到应用。

中英文主题词对照表：用于中英文词的对照翻译与中英文混合检索扩展。

专家库：通过对万方的科技文献仓储进行数据挖掘，形成了600多万的科技专家数据库，通过人工辅助规范形成，用于对专家的同名识别和专家知识仓储库管理。

多层级机构库：通过对万方的科技文献仓储进行数据提取，然后利用万方软件自行研发的机构名称规范辅助工具进行处理，人工校对生成。主要用于对于机构名称的标引规范，和机构名称的归一化检索，提高机构名称的“查全/查准率”。在机构创新能力评价中意义更为重要。

（2）深度数据加工标引与多维度的聚类和知识挖掘分析支持

除RMSCloud相关核心技术研发之外，同时对科技文献的加工处理提出了较高的要求。对于中外文科技文献仓储知识库建设，制定了元数据加工标引、质量检查等一系列标准规范，提升数据加工标引的质量。同时，对于科技文献元数据，严格按照学科、主题、人物、机构、基金等“知识获取五要素”进行深度标引，为围绕五要素的检索、导航、多维度聚类和知识挖掘分析提供了基础。

（3）相关度计算排序与相似结果推荐

RMSCloud可以根据用户检索关键词进行自动识别，判断用户检索人物、机构、期刊、主题等检索意图，同时可提供按照检索词的相关度排序和相似结果推荐。基于云计算架构的学术搜索引擎通过对于数据库、字段及索引定义权重分值，以支持多字段过滤与相关度排序及相似结果计算。

关于RMSCloud的详细技术及研究内容介绍，参见《数字图书馆论坛》2013年第6期云计算专刊。

5　RMSClod应用示范

多年来，北京万方软件股份有限公司一直从事科技信息服务系统相关的技术研究开发工作，开发出了系列相关产品，包括非结构数据资源管理系统RMS、万方数据资源整合服务平台、科技文献自动分类与摘要服务系统、万方学术搜索、科技创新文献共享支撑平台等系列产品和服务系统。经过近10年的研究与开发实践，结合目前承担的国家“863”计划重大专项“以科技文献为主的搜索引擎研制”部分成果，尤其是结合最新的RMSCloud系统的开发成果，我们提出了基于“云服务”的国际科技文献服务系统总体架构，并通过“中国学术搜索网”和部分省市示范系统建设得到实施验证。

图3是我们在多年研究开发基础上设计提出的基于“云服务”的科技文献服务系统总体架构图。下面将对这一系统架构做详细功能解释说明。

万方科技文献仓储云服务中心：该中心是本系统架构的核心，它包括规范化的元数据仓储中心、相关知识库中心和管理这些数据的基于Web Service架构的资源管理与服务系统[7]RMS，RMSCloud为底层云学术搜索引擎。其中，科技文献仓储云服务中心包括中外文期刊、会议、学位论文、专利、标准、法律法规、科技成果、科技人物、机构等以事实数据为基础的元数据仓储，该仓储中心的数据规范原则，以本人提出的“知识获取五要素”为指导思想。该数据仓储数据规范的主要工作目标是解决科技信息服务中的人物重名和机构名称变迁、机构合并等引起的“查全/查准”问题。目前，该仓储的元数据记录数达6亿规模，几乎涵盖所有科技文献所涉及的中外文元数据记录。

知识库中心：包括知识获取五要素中涉及的学科、人物、主题、机构、基金等相关知识库，其中包括420多万的主题知识库和1200万作者相关的知识库，其中作者的科研合作网络和学术网络知识库是通过数据挖掘及其相关技术由计算机自动生成的，对外提供服务接口。

云学术搜索引擎RMSCloud：在元数据服务中心里，RMSCloud负责元数据的接收、存储、索引，并提供标准的云搜索服务。以RMSCloud为基础的云服务示范系统“中国学术搜索网”已经正式投入服务（http://www.sciinfo.cn）。

在本架构中，万方科技文献仓储云服务中心属于公共云服务中心范畴，它除了管理万方软件自己的仓储数据外，还可以为用户提供数据共享服务。目前，该中心支持15种标准格式的元数据交换，涵盖期刊、会议、图书、方志、报告、视频等科技文献数据。同时，该服务中心属于本架构方案中的最底层，除了网络和系统上的安全措施外，对存储在中心的所有数据均采用了高强度的128位加密算法进行磁盘级保护。目前，该服务中心已正式对图书、情报等信息服务部门提供服务，用户通过接口直接调用本中心（中间经过云调度中心的认证

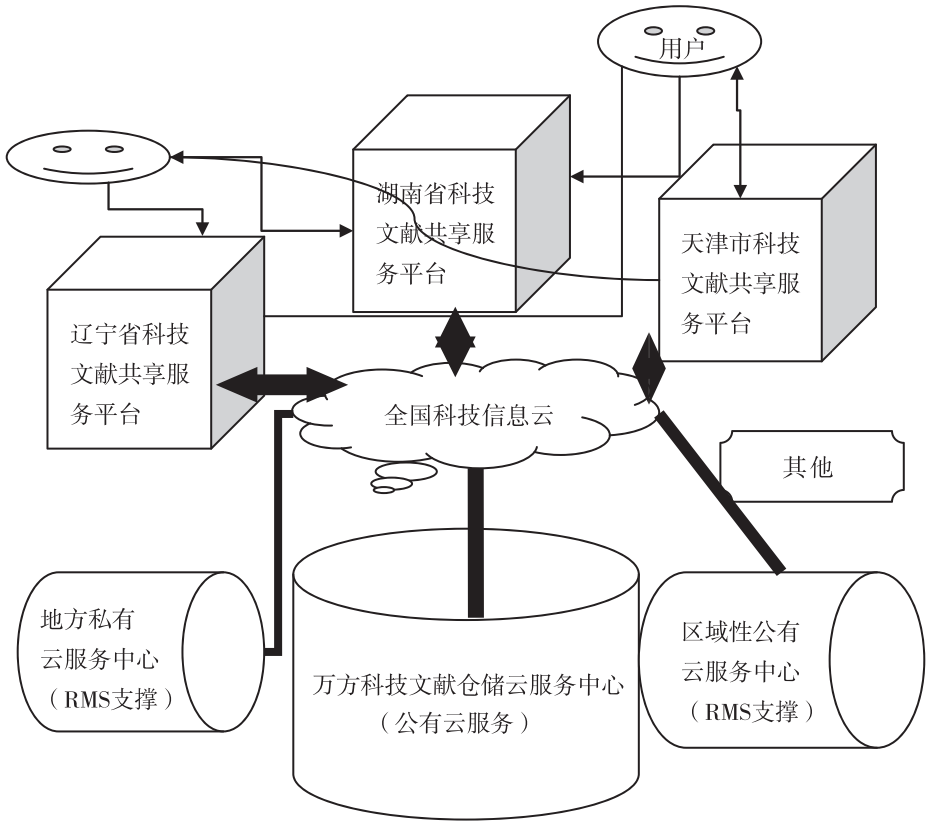

图3　基于“云服务”的国家科技文献服务平台总体架构图

和管理）提供的80多种服务。该中心提供的服务包括数据库管理、检索、数据交换、自动标引与分类、基于“知识获取五要素”的导航服务、聚类统计分析等。另外，云服务中心的硬件设备也可根据用户规模、资源规模进行快速扩展部署。

地方私有云服务中心：公共云服务模式客观存在潜在的用户关键数据的安全风险（起码从技术上是这样的），用户的关键数据放在公共云服务中心，虽然节省投资，提高了效率，但毕竟放在别人那里。因此，在总体架构里面提供了对私有云服务中心的支持，私有云服务中心可提供和公共云服务中心一样的功能。不同之处在于，该中心还支持对其他关系数据库的管理，这样也可以方便将原有老架构的系统纳入新的云服务管理架构继续使用。私有云服务中心主要用来管理用户的本地关键数据，规避云服务潜在的安全问题。另外，由于私有云服务中心提供的各类服务相对公有云服务要简单的多，可采用集中式搜索引擎RMS系统为搜索引擎。

区域性公有云服务中心：目前，部分省市信息服务机构已经开始建设区域性重点行业科技创新服务系统，这类系统不同于现有的文献服务系统，主要表现出区域性、内容涵盖面广、交叉性等特征。区域性：是为本地区重点产业的关键业务提供支撑。内容涵盖面广：不仅仅是科技文献服务，还包括基于互联网信息产业动态、研究报告、政策法规、专家互动、竞争情报、成果转化与服务等一系列产业信息服务。交叉性：虽然产业服务是某个省市根据自身区域业务需求提出的，但在全国范围内和部分区域仍然存在一定的交叉性。鉴于这种情况，万方软件提出的区域性云服务的架构思想，主要是为了避免不同省市间产业信息的重复建设。比如，辽宁省已经建设完成了车床产业服务平台，吉林等其他省市也需要这样的产业服务，我们建议以辽宁为主，其他省市参与共建共享。

区域性云服务中心的管理架构和万方元数据云服务中心一样，只是内容的归属有所区别，也可体现多个信息服务机构的共建、共享宗旨，其搜索引擎视规模来选择RMS或RMSCloud。

云服务调度中心：云服务调度中心是本架构中“云服务”最为核心的部分，是RMSCloud云服务的基础，所有基于“云服务”的管理、调度模式都在这里得到体现。它主要包括整个云的安全防护与

认证、用户管理、服务机构管理、云数据服务中心配置、管理与调度、服务缓存、服务负载均衡等功能。无论公有云、私有云，还是区域性云服务都通过本调度中心进行管理与调度。实际上该调度中心是“国家科技文献服务”的资源调度中心。

省市科技文献共享服务平台：省市科技文献共享服务平台是各省科技信息（情报）研究所（院）根据自身业务特点提出的面向本省市的科技文献共享保障平台，具有明显区域特征和个性化服务模式。目前，大部分服务平台在公共性文献信息服务方面基本上都是利用万方科技文献仓储云服务中心所提供的数据和相关接口服务，将自己拥有的特色数据存放在自己的私有云服务系统中。采用万方软件提供的科技创新文献共享支撑平台，可直接调用万方云服务平台的示范系统“中国学术搜索网”提供的所有服务，同时可以调用私有云的所有服务。

到本文截稿时，除“中国学术搜索网”已经正式对外提供服务外，辽宁、吉林、黑龙江、山东、山西、湖南、河南、云南等省市科技文献服务平台的搜索引擎服务都已经正式接入到万方科技文献云服务中心。其中，最具代表性的是“甘肃省科技文献服务平台”，平台门户及业务平台均由他们自己开发完成，其中的数据搜索、数据挖掘与分析、主题趋势分析、原文定位等均是调用RMSCloud的云服务接口完成。由于本文主要目的是阐述RMSCloud的科技文献云服务功能，涉及科技文献服务的相关核心技术没有做更多的介绍。详细参见参考文献[8]和文献[9]。

6　结语

RMSCloud是针对科技文献服务的需求特点采用云计算核心技术架构，开发完成的专用科技文献云搜索服务平台。万方软件利用RMSCloud系统对原来基于RMS资源服务系统构建的省市科技文献服务系统进行了全面升级，使用的科技文献元数据全部来源于仓储中心，节省了大量服务器和搜索引擎部署，大大提高了系统的功能和性能。同时基于“云服务”的分布式搜索引擎RMSCloud的研制成功，可为未来科技文献系统的“大数据”应用提供自主知识产权技术支撑和保障。

参考文献

[1] NIST. Final Version of NIST Cloud Computing Definition Published [M/OL]. [2013−04−18]. http://www.nist.gov/itl/csd/cloud−102511.cfm.

[2] Map Reduce: Simplied Data Processing on Large Clusters [M/OL]. [2013−01−09]. http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/zh−CN//archive/mapreduce−osdi04.pdf.

[3] Hadoop [EB/OL]. [2012−03−19]. http://hadoop.apache.org.

[4] 田嵩,晏伯武,杨慧等.基于GFS的分布式云存储应用技术的设计[J].福建电脑,2012(10):23−25.

[5] 刘星.Hbase性能深度分析[J].程序员,2011(7):102−104.

[6] 朱学迅.虚拟化技术研究[J].电信技术研究,2008(5):28−31.

[7] 吴广印.基于Web Service构架的资源共享技术研究与实现[J].情报学报,2007(6):851−857.

[8] 吴广印.RMS系统架构与情报检索系统的功能需求研究[J].数字图书馆论坛,2013(6):31−38.

[9] 吴广印.分布式学术搜索引擎研制及其大数据应用[J].数字图书馆论坛,2013(6):10−18.