一种基于 MongoDB 的大数据管理架构

钟 麟 员建厦
(中国电子科技集团公司第五十四研究所,河北 石家庄 050081)

[摘 要] 随着计算技术和互联网技术的不断发展,大数据处理技术也逐渐兴起,传统的关系数据库已经不能满足大数据管理的需求。结合增长速度越来越快的遥感影像数据,对“NoSQL 数据库”进行了介绍,对大数据的概念和特点进行深入分析,提出了通过 MongoDB 数据库实现对海量遥感影像数据进行管理的存储模型。

[关键词] 大数据 关系数据库 NoSQL MongoDB 副本集 分片
中图分类号:TP311 文献标识码:A 文章编号:1003-1739(2016)05-62-4

A Data Management Architecture Based on MongoDB

ZHONG Lin YUN Jian-sha
(The 54th Research Institute of CETC, Shijiazhuang Hebei 050081, China)

Abstract:With the development of computing and Internet technologies, the big data processing technology is gradually emerging. The traditional relational database has been unable to meet the needs of big data management. This paper describes NoSQL database combining with large quantity of remote sensing image data. The concept and characteristics of big data is analyzed in depth. The storage model of massive remote sensing image data management is put forward in this paper.

Key words:big data; relational database; NoSQL; MongoDB; replica set; shard

1 引言

伴随着计算机技术的发展,我国航天技术和卫星遥感技术也得到了空前的飞速发展,作为对地遥感信息的承载者,遥感影像数据的数据量也随着时间的推移而日益增加和不断积累,每天都以几 TB 甚至是几十 TB 的速度快速增长[1],而且随着遥感影像数据分辨率的不断提高,单个的遥感影像数据文件的数据量也在变大,导致整体的数据规模不断变大,这给当前的遥感影像数据管理系统带来了巨大的挑战。

在目前的遥感影像数据存储系统中,一般都是使用传统的关系型数据库保存遥感影像数据,随着需要管理的数据越来越多,这些系统在满足对空间的利用率、海量数据存储支持和数据高可靠性等方面的需求时已经显得越来越无能为力,在应对云存储技术和云计算技术等方面也表现出了很多难以克服的问题。遥感影像数据的数据量正在迅猛增长,其规模越来越庞大,对遥感影像数据的存储管理系统的要求也越来越高,应当具备支持海量数据、高性能、高可靠性、易扩展和资源的高重复利用率等特性,这些特性在以往传统的采用关系型数据库实现的存储管理系统是没有的,在这些需求越来越难满足的情况下,必须考虑采用新型的技术体制和更有效的手段进行数据的生产、管理和存储。

定稿日期:2016-02-12

NoSQL 数据库的出现,很好地解决了以往通过传统型的数据库在管理海量数据中的问题,使得在云计算环境下对海量数据进行管理不再存在性能瓶颈[2]。相对传统的关系型数据库来说,采用 NoSQL 数据库技术实现对遥感影像数据的存储管理是一个行之有效的方法。

2 NoSQL 数据库技术

2.1 NoSQL 介绍

NoSQL 数据库也被称为非关系型数据库,它是一系列与关系型数据库有较大差异的数据管理系统的统称,其中最主要的差异在于它不使用 SQL 作为基本的查询语言,由于它们不是用关系模型作为其主要的数据模型,并且不提供 SQL 接口,被统称为“NoSQL”[3]。

在 NoSQL 数据库中，主要有以下 3 种主流类型：键 - 值对（Key - value）型、文档型和面向列的存储型，分别如下：① 键 - 值对型：数据库中保存的都是 - 个键和相对应的值组成的键 - 值对，这样的数据结构非常简单，但是却能够提供远远高于关系型数据库的查询速度，正好满足大数据存储和高并发性的要求；② 文档型：文档模型的值是有语义的，可以在值上建立索引来为上层应用提供便利；③ 面向列存储型：使用表作为数据模型，访问数据时只返回在查询结果中的那些列的数据，这是由于数据库中的每一列都是数据库中的索引，面向列存储型的数据库不支持表之间的关联操作。

2.2 NoSQL 的特征

一般的 NoSQL 数据库通常具有如下的几个特征：

① 可实现水平扩展能力：在数据库运行过程中，可以通过简单的添加硬件或服务节点的方式来对系统进行扩展，以适应数据增长到一定规模时对性能的进一步需求，弥补了关系型数据库在扩展性上的不足[4]；

② 复制和分区（分片）的能力：大部分的 NoSQL 数据库可以运行在多台配置不高的 PC 机上，通过复制的功能来实现数据的冗余和系统的容灾机制，实现大规模集群的工作方式；

③ 接口简单：NoSQL 数据库一般不提供自己的开发语言包，只提供会话级别的接口或协议，简化了接口方式；

④ 较弱的事务模型：只支持较弱的事务，只需满足最终一致性的要求；

⑤ 使用分布式索引：在分布式集群中，可以将文档保存在不同的服务器上，通过分布式的索引在其中快速定位数据，而且可以通过内存存储数据的方式提高数据的读写速度，提高访问性能[5]；

⑥ 灵活的数据模型：不需要预先定义表结构，随时可以为数据记录动态添加或修改属性，不影响已有数据，在软件实现过程中仍然可以存储自定义的数据格式，而不需要想关系型数据库那样事先为要存储的数据建立字段。

2.3 NoSQL 数据库与关系型数据库的比较

常见的三类 NoSQL 数据库与关系型数据库的比较如表 1 所示。

表 1 NoSQL 数据库与关系型数据库的比较

数据模型	列存储	键-值存储	文档存储	关系模型
架构	主+从服务器模式	无中心分布式	自动分片+副本集	主备
查询	不支持复合条件	只支持主键查询	支持丰富的查询	能进行复杂查询，支持多表连接
读写特性	版本满足一直	总可写，读复杂	写复杂，所有鞋版本满足一致性	严格的事务处理
扩展性	添加从服务器	添加节点	添加分片	很难
负载均衡策略	通过数据迁移	一致性哈希算法	自动分片	数据划分
数据版本	时间戳	向量时钟	即时完成写操作	单一版本

3 大数据及遥感影像数据

大数据指快速增长并难以被普通数据管理软件在可容忍的时间范围内进行捕捉、存储、查询、共享、分析和显示的数据。一般认为大数据具有数据量大、结构复杂、实时性高和价值密度低 4 个特征，并将之归纳为“四 V 特征”。

大数据技术横跨多个技术领域，从数据存储、虚拟化和云计算，到数据库管理、并行计算和数据挖掘。大数据来源的多样性导致了其结构的复杂性，网络日志、视频、图片及地理位置均可作为大数据的数据源，其中大部分是半结构化或非结构化数据。

遥感影像数据是一种典型的大数据，并且包括半结构化的元数据和非结构化的实体数据，具备大数据的基本特征。从其应用方面来看，遥感影像数据已在国家安全、社会持续发展、国土资源的合理开发与整治、资源的综合利用和有效管理、优化生产力的整体布局以及外交等领域发挥着越来越大的作用。由于遥感影像数据具有超高容量、可靠性强和方便及时等特点，使其在交通管理、土地规划、军事、资源、环境和防灾等很多领域都起着不可替代的用途。通过遥感影像数据还可以对农业、灾害、资源环境和公共安全等重大问题进行宏观决策，是保障国家安全的基础性和战略性资源。利用卫星平台，人们可以迅速得到几天前甚至几小时前的拍摄的高分辨率的遥感影像，使获取的信息更加及时准确。

遥感技术和遥感影像数据给人们的生产生活带来很多方便和进步，但是遥感影像数据的数据量很大，例如一幅分辨率为 2 m 的多光谱影像文件就可以达到 500 MB 左右，而随着技术的逐渐发展，遥感影像数据的分辨率还在不断提高，数据量便会不断增加。如何才能有效组织、大量存储、快速检索、快速浏览和方便使用这些海量的数据便成为摆在我们面前的一个迫切需要解决的问题，而 NoSQL 数据库是实现大数据管理的一个有效途径。

4 基于 NoSQL 的存储模型

4.1 MongoDB 介绍

由 lOgen 公司开发并维护的 MongoDB 数据库是一种面向文档存储的 NoSQL 数据库，可以实现对海量数据的存储管理，它的特点是高性能、易部署和易使用，非常适合存储大数据。MongoDB 是一种高性能、开源和无模式的文档型数据库，它能够支持与关系型数据库中类似的很多操作，也能够提供类似关系数据库中很多的功能，其执行语句的语法特点和关系型数据库的 SQL 语法语法也很接近。使用该数据库不需要

预先定义模式和数据库结构，它采用 BSON 语法格式存储数据，支持的数据结构非常松散，支持嵌入文档等复杂结构，可以创建多级索引，也可以在每一列单独建立索引，使用非常灵活方便。

在 MongoDB 中，文档(document)是其核心概念，类似于关系型数据库中的元组，由多个键及其相应的值有序地存放在一起组成。若干个文档便组成集合(collection)，集合类似于关系型数据库中的表，可包括多个文档。多个集合可以组成一个数据库，一个 MongoDB 的实例可以承载多个数据库，每个数据库之间可以是完全独立的，文档、集合和数据库的层次结构关系如图 1 所示。

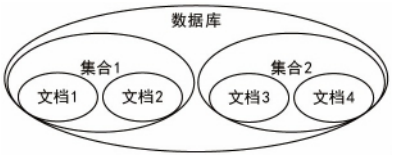


图 1 MongoDB 的层次结构

传统的关系型数据库首先由数据库组成，在数据库中可以创建若干个数据表，每个表中保存具体的记录。对应于这 3 个层次概念，MongoDB 数据库也是由、集合和文档对象 3 个层次组成，分别与关系型数据库中进行对应。另外，MongoDB 中的文档对应于关系型数据库里的表，但是由于 MongoDB 有模式自由的特点，集合中没有列、行和关系概念。

4.2 总体存储结构

针对海量遥感影像数据的管理，系统总体的存储架构自下而上可以简单分为数据层、数据服务层和数据处理层 3 层，如图 2 所示。

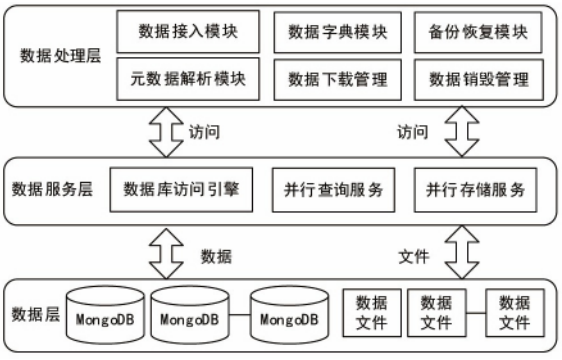


图 2 总体存储架构

在图 2 中，①数据层：采用 MongoDB 的集群数据库，为上层提供数据库服务支持，既包括数据库数据，也包括数据文件的存储管理，成为该管理系统中的数据基础；② 存储架构的中间层为数据服务层：通过该层实现对数据库的访问引擎和并行查询服务。通过数据服务层可以实现对各种数据库提供的不同数据源进行屏蔽，这样，系统在存取数据时不只局限对某一个数据库的操作，可以适应处理不同数据的要求，具有较好的可扩展性和完备性，方便管理和部署；③ 向应用提供的数据处理层：通过该层可直接面向应用提供数据接入、元数据解析、数据字典管理、数据下载、备份恢复以及数据销毁管理等功能。

4.3 MongoDB 的架构

在系统总体的存储架构中，MongoDB 的集群数据库的架构如图 3 所示：① 集群中的不同服务器可以作为不同的分片，每个分片可以是单个服务器，也可以使用副本集来提供分片的主副本和备份副本[6]；② 副本集由至少 3 台服务器实现，其中一台服务器为主节点，负责数据的写入，其它为副节点，只能读出主节点写入的数据；③ 配置服务器存储了集群中的元数据，能够记录每个分片的基本信息和块的信息；④ 查询路由器（mongos）可以当作一个路由和协调进程，通过它对 MongoDB 访问可以使集群中的多个组件看起来像是一个单一的系统，使客户端无需纠结需要访问集群中的哪个服务器。

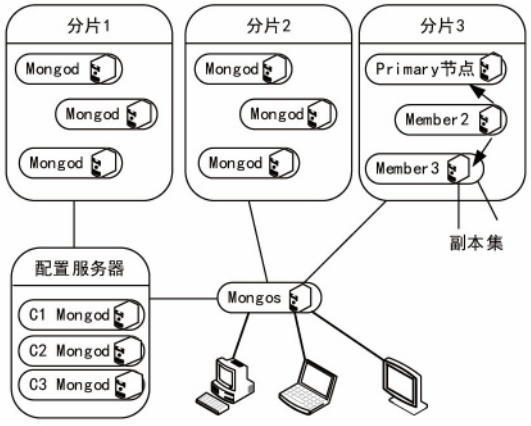
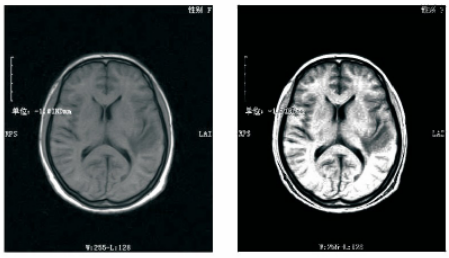


图 3 总体存储架构

5 结束语

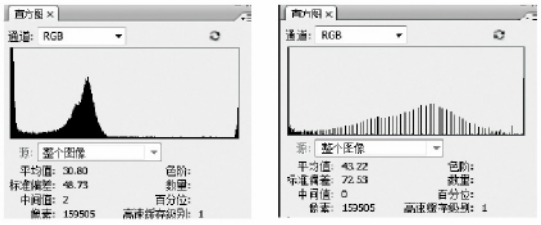
在本系统中，使用了基于 NoSQL 存储技术的 MongoDB 数据库系统，通过该数据库系统来存储需要共享的遥感影像数据，并且基于该数据库系统设计了一种具有高并发、高可靠性和高效存储等特点的存储架构，在这种大数据存储架构中，对数据的访问速度优于传统的关系型数据库，由于数据库能够提供丰富的查询和检索方案，使之满足了遥感影像数据的存储访问需求。该存储架构中的低成本的横向扩展模式提供了基于自动分片机制，不同的分片之间可以负载均衡，使系统具有了更好的可靠性和数据处理性能。另外，MongoDB 数据底层采用了列存储机制，如果某条数据记录中含有空列并不占用实际空间，大大节省了空间资源。 （下转第 74 页）



(a) 原图　　(b)均衡化后

图 7 图像均衡化处理图

直方图则为表示图像的每个颜色亮度级别的像素数目，展示像素在图像中的分布情况，提供图像色调范围或图像基本色调类型的快速浏览图，以便更好的校正色彩[2]，处理前后的直方图如图 8 所示。



(a) 原图　　(b) 均衡化后

图 8 均衡化后的直方图

“曲线”命令是精确调整颜色色调的命令，顶部点调整图像亮部，曲线中部点调整中级灰度，曲线底部点则调整图像暗部。通过曲线命令也可以调整图像曝光过少或过度，达到均衡化和曝光正常化[6]。照片的曝光正常和清晰程度有利于医生诊断鉴定,更加准确地掌握病人的情况 ;对于学生来说学习这种方法以后在工作中可以进行间接使用；同时在医学高职院校这种方法也提供了教师制作医学影像教学资料的途径，有利于读片实验教学的开展，有利于学生的学习，有利于教学资料的储存。

方法与步骤 :① 在 Photoshop 中选择“打开”命令选择图 7(a)的原始图片 ;② 选择“窗口”→“直方图”命令，打开直方图面板，观察直方图情况发现左端暗调像素分布密集，右端像素分布少，表明曝光不足，如图 8 (a) 所示 ;③ 选择“图像”→“调整”→“曲线”命令[5]，调整直方图的均衡化如图 9 所示，最终达到图 7(b)的效果。

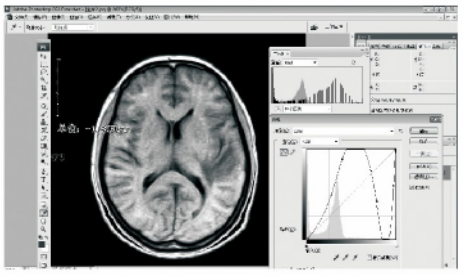


图 9 “曲线”命令调整均衡化

4 结束语

文章主要探讨如何使用 Photoshop 技术进行医学领域相关的图像处理的方法和步骤，为医学院校制作教学辅助课件以及学生通过 Photoshop 软件间接学习临床后期图像处理方法提供可借鉴途径之一。可以说 Photoshop 当前已经融入到医学相关领域中，正在为医疗卫生领域服务，我们应该重视和发展它。

参考文献

[1] 赵 岩.3ds Max 动画与后期制作完美风暴[M].北京:人民邮电出版社,2011:24 - 75.
[2] 汪 可.Adobe PhotoshopCS 认证考试指南[M].北京:人民邮电出版社,2006.
[3] 王 研,张志常.Photoshop 软件在临床影像诊断及医学形态学方面的应用[J].电脑与信息技术,2012(3):27 - 51.
[4] 蓝 鹏,云素珍,张 凌.腮腺良恶性肿瘤 CT 影像色度学定量研究[J].内蒙古医科大学学报,2013(6):429 - 433.
[5] 李 晶,杜润家,南晓东.Photoshop CS 软件在骨与软组织病变影像诊断中的应用[J].西北民族大学学报,2012(4):66 - 69.
[6] 穆天虹.基于 Photoshop 的图像处理[J].信息与电脑,2011(6):241 - 243.

(上接第 64 页)

参考文献

[1] 胡文波,徐造林.分布式存储方案的设计与研究[J].计算机技术与发展,2010,20(4):66 - 68.
[2] 张 恩,张广弟,兰 磊.基于 MongoDB 的海量空间数据存储和并行[J].地理空间信息,2014,12(1) :46 - 48.
[3] 杜卫华.浅析基于 MongoDB 的云数据管理技术的研究与应用[J].网络安全技术与应用,2014(8):89 - 90.
[4] 倪睿熙.一种基于 JSON 的异构数据查询方法[J].无线电通信技术,2013,39(1):73 - 76.
[5] 吴 淼,倪力舜.一种针对 MongoDB 数据库的证据获取方法[J].中国司法鉴定,2011(3):54 - 55.
[6] 梁于胜,王洪琳.无人机数据链自动测试系统设计[J].无线电工程,2014,44(4):20 - 22.