基于 HDFS 的区域医学影像分布式存储架构设计

摘要：构建区域医学影像协作平台是均衡医疗资源、提高基层医院诊疗水平、降低医疗费用的有效途径，但是构建区域化影像平台在技术和成本上还存在巨大的挑战。本文详细分析了传统集中式存储和 HDFS（Hadoop Distributed File System）分布式存储系统的优缺点，设计了一种适合 HDFS 特点的 S-DICOM 文件格式，以及集中式存储（FC SAN）和分布式存储（HDFS 集群）结合的统一存储架构，开发了一套 SDFO（S-DICOM File Operator）中间件，为上层的 PACS 应用组件提供透明的存储访问接口。测试结果表明此架构可以满足海量医学影像资料的快速存取和处理需求。

随着 X 线机、CT、磁共振等大型影像设备在临床上的广泛应用，影像检查已成为临床诊断最重要的依据之一。但是昂贵的影像设备和重复的影像检查也成为医院和病人医疗支出的重要部分。同时，影像诊断难度大、操作复杂度高、专业性强，基层医院极其缺乏优秀的影像诊断人才。医疗设备和人才的不均衡，也是造成目前“看病难、看病贵”的重要原因。构建区域一体化的医疗协作平台，是均衡医疗资源、提高基层医院诊疗水平、实现“有序医疗”的重要途径。其中区域医学影像协作平台的构建，是区域医疗协作的重要组成部分，但是构建区域化的医学影像协作平台在技术和成本上还存在着巨大的挑战。

1 构建区域医学影像协作平台面临的挑战

数字医学影像技术目前已有成熟的国际标准，即 DICOM 3.0，遵照其标准建设的 PACS 系统也已从单机、科室逐步发展到全院、区域。目前国内许多大型三甲医院已开展全院 PACS 应用，实现了医院无胶片化。PACS 系统区域化将是下一阶段政府卫生部门和医疗机构的主要研究目标，但是构建大型区域医学影像中心和协作平台目前还面临巨大的挑战。

1.1 建设费用高

PACS 的数据量远远大于 HIS、LIS 等其它医疗系统，区域医学影像数据达到数百 TB 甚至 PB 级，采用传统存储架构（如 FC SAN/iSCSI 等）费用极高。

1.2 传输带宽存在瓶颈

即使是高性能的 FC SAN，其网络带宽和处理能力也难以达到 PB 级数据的快速处理和传输要求。

1.3 可用性受限

大型医院 PACS 系统常用“在线-近线-离线”的存储模式，离线数据大多存储在磁带库中，其可用性较差，数据不能实时获取。

1.4 缺乏一体化的应用平台

目前的医学影像协作，如远程影像会诊基本采用“点对点”的模式，缺乏一体化、跨平台、高可用的区域医学影像协同应用软件。随着云计算技术的飞速发展，为构建低成本、高可用、高性能的区域医学影像协作平台提供了一条有效的途径。云计算是 Google 率先提出来的一种新的技术和运营模式，从应用范围来看，云计算可分为公有云、私有云和混合云。从服务模式来看，云计算可分为 IaaS、SaaS 和 PaaS。区域医学影像云计算平台属于混合云的范畴，我们承担的课题就是研究通过医疗集团内部医院之间的高速城域网、医保网、电子政务外网、互联网等传输介质，为各类医疗机构提供 SaaS 模式的医学影像协作应用系统，包括 Web DICOM 终端、影像会诊、影像转诊、远程教育、数字胶片代存等服务。而高性能、高可靠的海量图像存储系统将是医学影像云计算平台的基础和关键，本文主要介绍一种基于 Hadoop 平台的分布式存储和传统集中式存储（FCSAN）相结合的存储架构的设计和实现。

2 Hadoop 平台简介

Hadoop 是目前应用最广泛的开源分布式存储和计算平台之一。它是根据 Google 的 GFS 分布式文件系统和 Map/Reduce 分布式计算技术而开发的开源平台，其设计目标是在普通的硬件平台上构建大容量、高性能、高可靠的分布式存储和分布式计算架构。Hadoop 目前已在 Yahoo、Facebook、亚马逊、百度、中移动等公司取得了广泛应用。其中 Yahoo、FaceBook 等公司已构建了数千至数万台普通服务器组成的大型 Hadoop 应用集群，FaceBook 上存储的图像数据量目前已超过 1 PB 即 1024 TB）。

2.1 Hadoop 集群的特点和适用性

Hadoop HDFS 分布式文件系统具有如下特点：（1）非常适合海量数据的存储和处理；（2）可扩展性高，只需简单添加服务器数量，即可实现存储容量和计算能力的线性增长；（3）数据冗余度高，缺省每份数据在 3 台服务器上保留备份；（4）适合“流式”访问，即一次写入，多次读取，数据写入后极少修改，这非常适合医学影像文件的特点；（5）除了数据存储能力外，Hadoop MapReduce 分布式计算框架还可充分利用各服务器 CPU 的计算资源，便于后期开展基于海量医学影像数据的图像融合、图像内容检索、三维重建等数据密集型计算。

2.2 存在的问题

Hadoop 在构建医学影像存储系统时还存在以下问题：1）Hadoop 的设计理念是针对大文件进行优化的，其默认的数据块大小为 64 MB，而医学影像资料中常见的 CT、MRI 的图像大小大多为 512 KB 左右，一次拍摄产生的图像数量大约为 100 ~200 幅，如果直接将这些大量的小文件存储在 HDFS 文件系统中，过多的小文件将导致 HDFS 的主节点 NameNode 内存消耗过大，降低整个集群的性能。2）HDFS 的设计理念不适合实时应用，在数据写入的过程中，每个数据块需复制 3 份，其写入性能大大低于读取性能，因此不太适合需要快速获取图像资料并撰写诊断报告的 PACS 实时应用。

3 系统设计

针对上述问题，我们设计了一种适合 Hadoop 平台的序列 DICOM 文件格式（S-DICOM），以及一套传统的集中存储和 HDFS 分布式文件系统相结合的 S-DICOM 文件存储架构。

3.1 S-DICOM 文件格式

CT、MRI 等 DICOM 文件大小虽然只有 512 KB 左右，但是病人的每个部位的检查通常都有 100~200 张图片，这样每个病人每次检查的数据量也将达到 50~100MB。而另一种常见的医学影像设备 X 线机（CR、DR），其单幅图像的数据量约为 8~20 MB，每次检查拍摄的图片一般为 2~4 幅，其总数据量也满足 HDFS 文件系统的要求。因此，将一个病人一次检查的所有图像合并成一个文件，再存储到 HDFS 中是比较合理的。我们采用了 Hadoop 的 SequenceFile 文件格式，将每个 DICOM 文件转化成健值对（key/value）的形式，然后合并成一个单独的 S-DICOM 文件，其中 key 为原 DICOM 文件名，value 为 DICOM 文件内容，文件格式（图 1）。

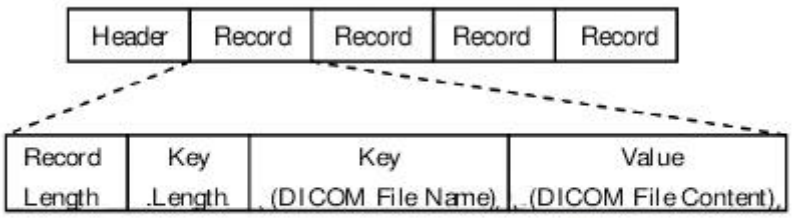

图 1 S-DICOM 文件格式

Fig.1 S-DICOM file format.

3.2 混合式存储架构

单纯的 HDFS 分布式文件系统不适合实时应用，但是具备低成本、高可扩展、高性能、高可靠的特点，传统的集中存储（FC SAN）则非常适合小文

件的快速写。因此，结合两者的优点我们设计了一套混合式存储模式，其核心是 SDFO（S-DICOM File
Operator）中间件，主要用于屏蔽底层操作细节，为上层的 SaaS 模式医学影像应用系统和 DICOM 应用组件提供统一的图像查询、读取和写入接口。SDFO 的核心主要由 SDFO Lo-cator、SDFO Reader、SDFO Writer、SDFO Converter、SDFO Client 五个部分组成。SDFO Locator 用于检索 DICOM 文件的存储位置，SDFO Reader 用于读取 DI-COM 文件，SDFO Writer 负责将从影像设备获取的图像写入集中存储（FC SAN），SDFO Converter 负责定时将 FC SAN 中的 DICOM 图像转换为 S-DICOM 格式，合并后存储到 HDFS 中。其系统框架（图 2）。

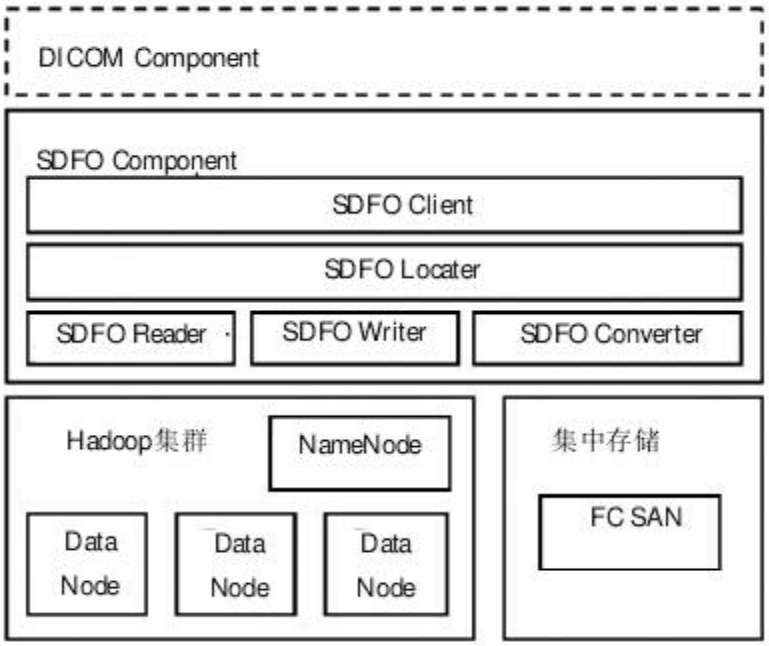

图 2 系统框架

Fig.2 The system architecture.

医院 PACS 系统中存储的图像，超过 3 个月后，其访问量将大大下降，因此我们将 3 个月内的 DICOM 文件以其原始文件格式存储在 FC SAN 中，超过 3 个月的图像则定时转换成 S-DICOM 格式，存储到 HDFS 中（也可根据需要设置存储超期时间）。利用 Hadoop HDFS 的线性扩展能力，我们可以将传统 PACS 的“在线-近线-离线”模式简化为“在线-近线”模式，解决离线数据可用性差的问题。

3.2.1 图像读取流程

SDFO 从 Hadoop HDFS 集群和 FC SAN 中检索和读取图像的流程（图 3）。

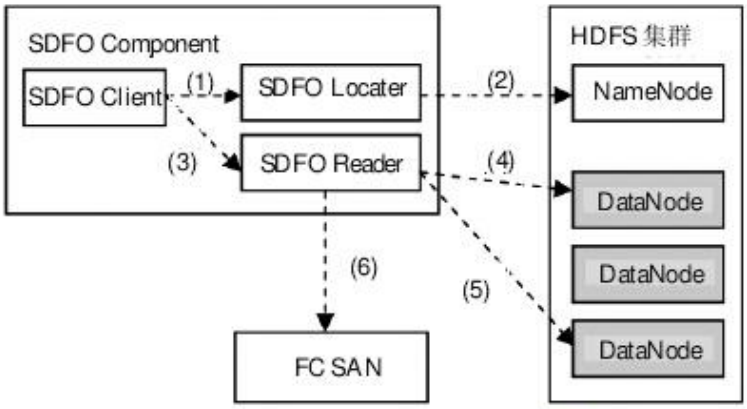

图 3 图像读取流程

Fig.3 The image reading process.

（1）从 DICOM Locator 获取图像存放的路径，如果图像存放在 FC SAN 中，则跳至第 6 步；

（2）从 HDFS 的 NameNode 获取文件数据块所在的 DataNodes 位置；

（3）调用 SDFO 的 read 方法，开始获取图像；

（4）从 HDFS 的 DataNode 1 获取第一个数据块，以此类推至其它的数据块，此步骤可以并行操作；

（5）从 HDFS 的 DataNode n 中获取最后一个数据块，将所有的数据块合并成完整的文件，关闭 HDFS 数据流，并将其转换成标准的 DICOM 图像；

（6）存放在 FC SAN 中的 DICOM 文件直接通过 JAVA 的本地文件系统接口读取。

3.2.2 图像写入流程

SDFO 中间件中 DICOM 文件的写入方式与传统的文件写入方式相同，直接通过 JAVA 本地文件系统接口写入 FC SAN。

3.2.3 图像转换流程

图像转换流程定时将 FC SAN 中的 DICOM 文件合并成 S-DICOM 文件，存入 HDFS 中。其转换流程（图 4）。

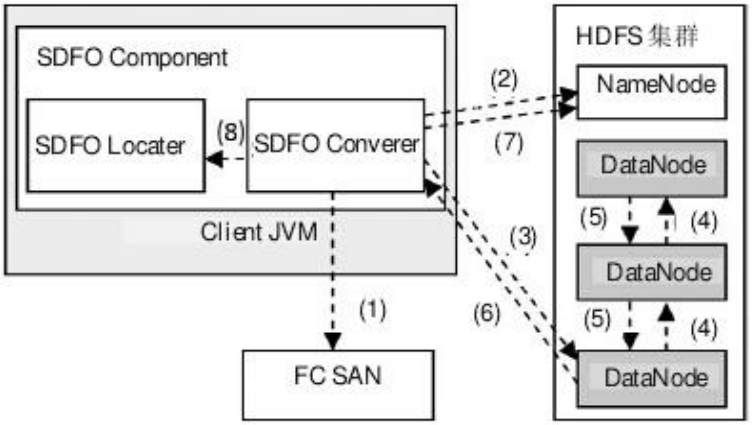

图 4 图像转换流程

Fig.4 The image converting process

（1）调用 JAVA 的本地文件系统接口，循环获取 FCSAN 中某个文件夹下的文件列表（每个病人每次检查的所有图像存放在一个单独的文件夹中），将每个 DI-COM 文件转化成一个健值对（key/value），将 key/vlaue 健值对顺序写入一个单独的 S-DICOM 文件数据流；

（2）调用 DistributeFileSystem 的 create 方法，通过 NameNode 的 RPC 接口创建文件，并获取用于存放数据块的 DataNodes 列表；

（3）调用 FSDataOutputStream，将 S-DICOM 文件转换成内部的数据队列，将数据写入第一个 DataNode；

（4）数据块写入成功后，第一个 DataNode 将写入的数据块复制到第二个 DataNode，依次类推至第三个 DataNode。

（5）按相反的顺序，第三个 DataNode 写入成功后，依次向第二个和第一个 DataNode 返回 ack packet，确认数据写入成功；

（6）循环写入所有的数据块后，调用 close 方法关闭 FSDataOutputStream；

（7）向 NameNode 发送 complete 指令，确认文件写入完成，更新 NameNode 的元数据；

（8）向 DICOM Locator 写入 DICOM 文件的存储路径。

4 应用测试效果

4.1 软硬件配置

我们目前已搭建了 20 台服务器组成的 Hadoop 集群。CPU：Intel Xeon E5504；内存：8 GB DDR3；网卡：两块 1000 Mbps 以太网卡；硬盘：4 块 1 TB SATA。存储空间共计 80 TB，按照 Hadoop 缺省配置，每个数据块在 3 台不同的服务器上保存副本，因此实际存储容量约为 27 TB。每台服务器均接入千兆汇聚层交换机，汇聚层交换机万兆上联。操作系统：64 位 CentOS 5.4；Java 环境：JDK 1.6.0-b09；Hadoop 平台：Hadoop 0.20.2。

4.2 测试结果

DICOM 图像的写入以及 3 个月内图像的读取均是直接通过 FC SAN 完成的，其性能与普通的 PACS 环境区别不大，因此我们主要测试读取 3 个月以前的 S-DI-COM 图像以及将 DICOM 图像合并转换成 S-DICOM 图像的性能。Hadoop 支持分布式读写，我们分别测试了 1~5 个 SDFO Client 的情况下，S-DICOM 读取和转换的性能如下表所示（单位：MB/s）：

表 1 测试结果 (MB/s)

	Clients数量				
模式	1	2	3	4	5
SDFO Reader	83.24	159.62	225.29	275.30	324.21
SDFO Converter	10.27	16.94	22.86	27.95	32.88

从测试结果可以看出 SDFO 的读性能基本是与 Client 数量线性相关的，这是由于 Hadoop 中的数据块是均匀分布在各 DataNode 中的，读取文件时可以聚合各 DataNode 的网络带宽，随着 DataNode 数量的增大，其聚合的总带宽将远远超过传统的 FC SAN 传输速率。根据测试情况来看，客户端同时读取和转换一个病人一次检查的 S-DICOM 文件时间约为 1~2 s 左右，这样的延时对 PACS 系统的操作是可以忽略的。

从测试结果也可看出 Hadoop 的写入性能不佳，单个 Client 写入 HDFS 的速率只能达到 10 MB/s 左右，这是由于 HDFS 写入文件时需要同时写入 3 个副本相关。

但随着 SDFO Client 数量的增加，写入速率也相应增大，当 SDFO Client 数量为 5 时，总写入速率约为 33MB/s。一个大型三甲医院 PACS 系统每天产生的图像数据量约为 20 GB 左右，全部转换成 S-DICOM 文件耗时约 10 min，对于拥有较多医院的区域，可以通过增加 SDFO 客户端数量的方式，近似线性地提高转换和存储性能，在每天的夜间空闲时段进行数据转换任务也是可以接受的。

5 总结与展望

Hadoop 平台是构建超大规模数据集群，实现存储聚合和数据密集型分布式计算的优秀平台，它可以有效解决构建区域医学影像数据中心的成本高、可扩展性差、传输带宽不足、离线数据可用性差的问题。但是 Hadoop HDFS 也存在不适合 CT、MRI 等小文件的存储及实时应用的问题。为此我们设计了一种 S-DICOM 文件格式，使其适应 HDFS 的特点，同时通过传统的集中式存储（FC SAN）和分布式存储（HDFS 集群）组合的存储架构，开发了一套 SDFO 中间件，为上层的 PACS 应用组件提供透明的 DICOM 文件访问接口。该系统在测试平台上取得了比较满意的效果，能满足大型区域医学影像中心的功能和性能要求。今后我们将在此基础上开展进一步的研究工作：1）进一步提高系统的安全性，完善应用系统、存储架构和网络拓扑等方面的加密和授权机制，确保病人的隐私和数据安全；2）充分利用 Hadoop 集群的分布式计算能力，开发基于 MapReduce 算法的图像融合、图像内容检索、三维重建等应用。

作者：李彭军，陈光杰，郭文明