• 医学信息技术 •

基于 Hadoop 的电子健康档案云平台设计和实现 *

黄海平

（肇庆医学高等专科学校 肇庆 526040）

〔摘要〕 阐述基于 Hadoop 的电子健康档案云平台架构设计，包括服务对象及需求、逻辑架构、软件架构等方面，介绍基于 HBase 的电子健康档案云平台数据预处理模型，进行实验环境的搭建和配置，通过实验完成 Hadoop 集群的启动。

〔关键词〕 电子健康档案；Hadoop 云平台；HBase；数据预处理

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2016.01.004

Design and Implementation of Hadoop－based EHR Cloud Platform *HUANG Hai－ping, Zhaoqing City Medical College, Zhaoqing 526040, China*

〔**Abstract**〕 The paper elaborates the architecture design of Hadoop－based Electronic Health Records (EHR) cloud platform, including service objects and requirements, logical architecture and software architecture, etc. It introduces the data preprocessing model of HBase－based EHR cloud platform and how to establish and configure the experimental environment, as well as accomplish the startup of Hadoop cluster by experiments.

〔**Keywords**〕 Electronic Health Records (EHR); Hadoop cloud platform; HBase; Data preprocessing

1 引言

我国的医疗系统由大中小型医院、基层卫生机构、政府医疗卫生管理机构组成，数量庞大且类型各异。各个医疗机构的减少增加等变更，都会牵涉到其内部服务器上网络资源的重新调整[1]。而其中最重要的网络资源是电子健康档案（Electronic Health Relords，EHR）。《2014 年中国卫生统计年鉴》最新数据统计显示，广东省各地区医疗机构门诊年度新增病历数量达到了近 4 000TB，各医院入院人数新增病历数据量达到了近 600TB。以上数据还不包括以大文件形式存在的 EHR，如 B 超视频流、CT 视频流、医学影像图片等。

传统 EHR 的独立服务器，无论性能还是投入成本都无法负荷当下医疗数据对空间日益庞大的需求。从可扩展和动态化的 EHR 数据存储需求出发，建立一个基于 Hadoop 的 EHR 云平台是当前医疗信息技术条件下的首选。基于 Hadoop 的 EHR 云平台通过其分布式文件系统，能满足海量医疗数据的查询和存储需求；具有开放性，可实现信息共享；其结构化的数据库语言更可方便进一步

〔修回日期〕 2015－05－12

〔作者简介〕 黄海平，硕士，讲师，发表论文 4 篇。

〔基金项目〕 肇庆市科学技术局创新计划项目“云计算下临床数据交换的安全保护方案”（项目编号：肇科（2015）63 号）。

的数据挖掘。

2 EHR – Cloud 平台架构设计

2.1 EHR – Cloud 服务对象及其需求分析

电子健康档案云（EHR – Cloud）平台的组成部分，见图 1。

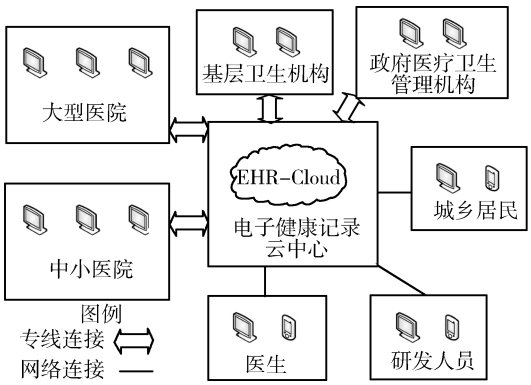

图 1 EHR – Cloud 平台服务对象组成

2.1.1 *各大中小型医院* 大医院可以为下面的中小型医院以及基层卫生机构，提供及时的诊断服务；并且也可以通过 EHR – Cloud 平台实时获取病人的诊断信息。

2.1.2 *基层卫生机构* 基层的卫生机构由于医疗设备及技术所限，需要通过该平台，联合大医院做出诊疗判断。基层机构可以把病人 EHR 上传到 EHR – Cloud 平台，由大医院抽取相关信息，做出诊疗判断，再把诊断结果反馈给平台传输返回给基层的卫生机构，以此作为依据，对病人做后续治疗。

2.1.3 *政府医疗卫生管理机构* 在遇到重大的社会性疾病的时候，EHR – Cloud 平台其自身的关系数据库能为政府医疗卫生管理机构第一时间提供所需的统计数据，以及医疗的具体情况记录，方便做出监督和决策。

2.1.4 *城乡居民* 城乡居民通过任意联网计算机就可以查询到自己的健康档案，享受个性化的私人健康管理服务。

2.1.5 *医生* 通过终端浏览器，不同医院的医生可以对同一个病人的 EHR 进行跨院实时访问，避免了重复诊断，对病患的病史也有一个全面的了解。

2.1.6 *研发人员* 在病人隐私得到充分尊重的情况下，相关的医学研究人员通过 EHR – Cloud 平台可以得到第一手的海量研究数据支持，给相关医学理论研究提供充分的资源。

2.2 EHR – Cloud 逻辑架构设计

2.2.1 *硬件虚拟化平台* EHR – Cloud 平台上多种应用程序协同运作，其用户群体以及机构的数目庞大，并且需要容纳实时变化的海量大文件，对服务器的性能以及负荷能力提出更高的要求。应用虚拟内存、硬盘、CPU 已经是大势所趋。业已成熟的虚拟硬件技术能通过负荷均衡、动态迁移、虚拟硬件空间，从而达到在不同服务器上运行不同应用程序的效果。这样一来，提高了系统的性能，所形成的基础设施即服务模式，大大降低了硬件的空间成本。

2.2.2 *EHR 数据中心* EHR 云数据中心，除了为区域内各大小医院、医生、病患以及相关的卫生机构提供 EHR 数据存储服务以外，还提供 EHR 数据灾难备份以及恢复服务。采用软件即服务应用平台的医院可以实时查询并上传病患的 EHR 资料，跨院进行病例诊断，从而实现转诊和远程医疗服务。

2.2.3 *软件即服务应用平台* 通过不断开发服务并且网络分享服务，从而达到降低系统软件维护及硬件安装成本的目的。中小型医院以及基层医疗卫生机构，对 EHR 实时性要求不太高的情况下，选择软件即服务的系统模式能有效降低软硬件维护成本。相对而言，大医院由于对于病患 EHR 的实时性要求高，普遍采用有固定单独服务器的传统 EHR 系统，该服务器能够存储海量的 EHR 数据，在本院内实现实时访问。软件即服务具有开放性和可扩展性，同时兼容多个 EHR 系统协同运作，方便中小型医院以及基层医院与大医院之间的转诊、会诊等在线的实时信息交流。

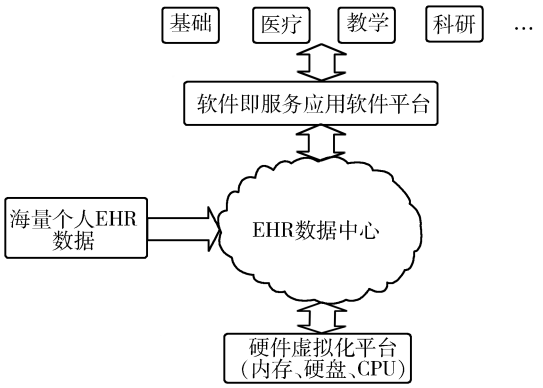

图 2 EHR – Cloud 逻辑架构组成

综上所述，图 2 中的 EHR – Cloud 平台通过硬件虚拟平台，扩充自己的容量和提升性能；软件即服务应用平台不断开发出不同的应用软件，方便各种类型的用户群体，在不同的终端设备上运行其软件，实现用户随时随地向 EHR 数据中心上传自己的 EHR 数据，可以读取个人的 EHR 数据。

2.3 EHR – Cloud 软件架构设计

2.3.1 设计原则 该电子健康档案云平台采用面对服务的架构，通过对异构软件松耦合的集成设计方法，有效实现了平台的横向扩展，兼顾功能上的灵活性和维护的安全性与便捷性。本文介绍的 EHR – Cloud 平台整体架构，见图 3。整个软件系统实现了模块化，按照自上而下的原则逐层分解。

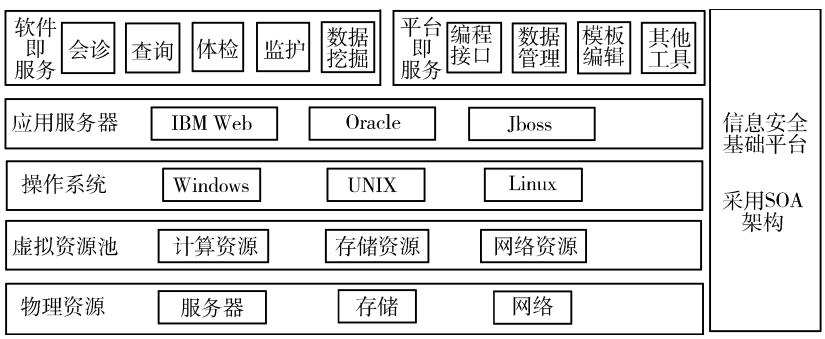

图 3 EHR – Cloud 软件架构

2.3.2 分层架构 整个 EHR – Cloud 软件架构[2]是在面向服务的总框架下分层搭建的，目的是开发出一个平台即服务模式以及软件即服务模式的 EHR 应用系统。该架构以物理层作为基础，物理层由服务器、存储介质、网络实体构成；再上面一层是通过硬件虚拟化技术构造的虚拟资源池，目的是把 CPU、内存以及硬盘抽象成为一个性能可靠、收缩自如的计算和存储平台，从而进一步实现该虚拟层的负载平衡以及动态迁移的功能。再上一层，在虚拟服务器上，不同操作系统并存，Linux、Windows、Unix、MacOS 自由切换；最上层的平台上开发软件即服务，实现了为软件即服务开发数据接口的功能。贯穿以上整个面向服务架构的是企业服务总线（Enterprise Service Bus，ESB），支撑该总线的是传统的中间件以及 XML、Web 服务等技术的合成体，目前主要的 ESB 有 IBM WebSphere ESB 和 Microsoft ESB 等。

3 基于 HBase 的 EHR – Cloud 数据预处理

3.1 概述

电子健康云存储平台为了实现海量的病人生理与动作数据，例如心电、呼吸、心跳、脉搏、血压等异构数据的存储，需要在这些数据进入 EHR – Cloud 的数据库之前，进行预处理。EHR – Cloud 采用了 Hadoop 的 HBase 数据库，是由于该数据库的数据存储类型均为字符型，兼容性比较好；另外该数据库数据的列族名均根据 EHR 标准建立，经过预处理后，一些生理与动作数据可以很好地实现存储，并且便于进一步的数据分析和数据挖掘。HBase 的 EHR – Cloud 数据预处理，严格遵循统一的标准，即《健康档案基本架构和标准（试行）》以及当前事件的《基本数据标准》[3]。

3.2 案例

3.2.1 数据预处理流程 以一个病例进行具体阐述。患者于 2015 年 1 月 1 日感觉肠道绞痛，于 1 月 3 日前往某医院某门诊（机构代码为 H122M500）就医，经过彩超检查诊断为急性肠胃炎。根据图 4 基于 HBase 的 EHR – Cloud 数据预处理流程如下：首先分析当前的医疗事件，提取主键信息，主要是身份证号（如李某，441202198010121040），提取事件的日期（如 2015 年 1 月 3 日），机构地点（如某医院某门诊），根据《健康档案相关卫生服务基本数据集标准目录》，可知应该采取门诊诊疗基本数据集，数据集标示符为 HRC00.01，然后根据《HRC00.01 门急诊诊疗基本数据集标准》查找到 ICD – 10 国际疾病分类标准编码，得出急性肠胃炎的 ICD – 10 编码为 K20，再根据《CV5199.01 检查/检验类别代码》，知道彩超检查编码为 9。最后按照列族、列名的方式完成对数据的创建。数据的预处理是系统后台进行的，前端界面的操作者，也就是医生只是通过 Web 客户端进行文档的输入，确认最后系统生成的信息匹配与否。

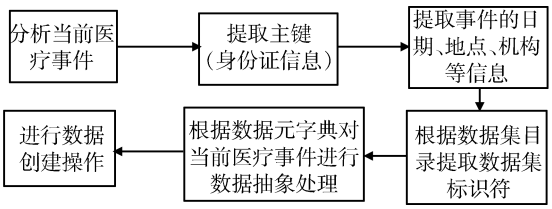

图 4 基于 HBase 的 EHR – Cloud 数据预处理模型设计

3.2.2 查找标准编码 根据表 1 可知应当采用门诊诊疗基本数据集，数据集标示符为 HRC00.01。根据表 2《HRC00.01 门急诊诊疗基本数据集标准》部分内容，查找到 ICD – 10 国际疾病分类标准编码，获取急性肠胃炎的 ICD – 10 编码为 K20。根据表 3《CV5199.01 检查/检验 – 类别代码》，得出彩超影像检查的编码值为 9。

表 1 健康档案相关卫生服务基本数据集标准目录

序号	数据集标准名称	数据集标识符
29	门诊诊疗基本数据集	HRC00.01
30	住院诊疗基本数据集	HRC00.02
31	住院病案首页基本数据集	HRC00.03
32	成人健康体检基本数据集	HRC00.014

表 2 《HRC00.01 门急诊诊疗基本数据集标准》部分内容

内部标示符	数据元标识符（DE）	数据元名称	定义	数据类型	表示格式	数据元允许值
HRC00.01.207	HR55.02.040	疾病诊断名称	略	S	A…50	—
HRC00.01.208	HR55.02.004	疾病诊断代码	略	S	AN7	ICD – 10 国际疾病分类标准编码
HRC00.01.209	HR42.02.112	发病日期时间	略	DT	DT15	—
HRC00.01.210	HR42.02.012	诊断日期	略	D	D8	—
HRC00.01.310	HR51.99.002.01	检查/检验 – 类别代码	略	S	N1	CV5199.01 检查/检验 – 类别代码

表 3 CV5199.01 检查/检验 – 类别代码

值	值含义
1	问询
2	物理
3	实验室
9	影像

3.2.3 结果数据 上述得出的信息作为结果数据，结合我国 EHR 的统一规定，构建出该病例的 HBase 数据库的数据集合，见表 4。

表 4 病例的 HBase 数据库部分内容

行键	时间戳	…	HRC 00.01 .207	HRC 00.01 .208	HRC 00.01 .209	HRC 00.01 .210	HRC 00.01 .310	…	HRC 00.02 .101	HRC 00.02 .102	…
–	…	–	–	–	–	–	–	–	–	–	–
–	T4	–	–	–	–	–	9	–	–	–	–
–	T3	–	–	K20	–	–	–	–	–	–	–
–	T2	–	–	–	–	–	–	–	–	–	–
–	T1	–	–	–	–	–	–	–	–	–	–

4 EHR – Cloud 平台实现

4.1 实验环境部署

两台实体 PC 机通过运行 VMWare Workstation 各虚拟出 3 台 linux 电脑，形成总计 6 台 Linux 电脑的 Linux 集群。集群采取主从模式，主服务器即 Master 服务器命名为 usky2001，分配 IP 地址为：196.168.1.11，其余 5 台服务器作为从服务器，即 Slave 服务器，分别命名为 usky2002、usky2006，分配 IP 地址分别为 196.168.1.12，196.168.1.13，196.168.1.14，196.168.1.15，196.168.1.16，见图 5。通过 ping，Linux 集群搭建成功。在 Linux 集群上部署 Hadoop 系统，进行相关配置。

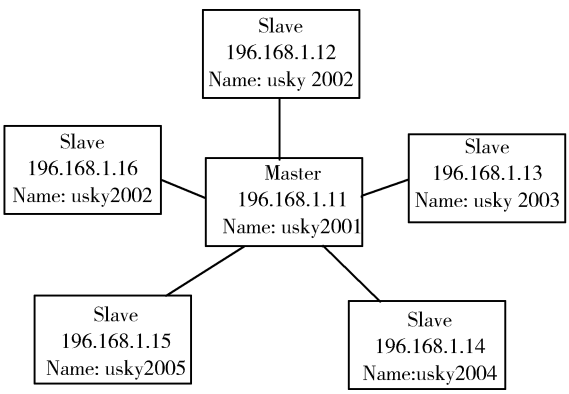

图 5 实验平台搭建

4.2 Hadoop 集群搭建

在搭建成功的 Linux 集群当中，解压并安装 Hadoop – 0.20.2 云平台软件，接着以集群的 6 台虚拟机中的主服务器作为 NameNode，其余 5 台从服务器作为 DataNode，修改 Hadoop 配置文件，分别完成启动主服务器和 5 台从服务器。通过所设置的局域网里任何的 PC 终端，在浏览器上输入 IP 地址 196.168.1.11，可以成功启动 Hadoop 集群，运行 Hadoop 集群后，分布式文件系统成功运行[4]。

5 结语

基于 Hadoop 的 EHR – Cloud 云平台能够满足海量异构大文件的个人健康档案的预处理以及存储的空间要求。将个人健康档案上传到云平台的数据中心，通过该平台的信息交换和协调运作，可以同时满足各级医疗机构、医生、病患以及政府卫生管理部门和医学研究者等多方面的需求[5]。而且云平台的无限扩展性以及资源整合能力，可以实现病患个人健康档案的信息共享，进而实现跨院的医疗诊断等信息化服务。当然目前而言，EHR 云平台的关键性技术，如虚拟化技术、分布式存储、分布式运算、分布式数据库目前还处于摸索阶段。全面利用 EHR 云平台推动医疗系统的信息化发展，具有巨大的社会和经济效益，但在技术层面的实现还有比较长远的路要走。

参考文献

1 胡铁军，李丹亚，王兆令，等. 中国医学信息网络建设的回顾与展望［J］. 医学信息学杂志，1995，(4)：1 – 4.
2 盛芳菲. 云存储服务在智慧医疗上的应用研究［D］. 北京：中国科学院大学，2013.
3 徐婷，鲍勇. 基于云计算远程平台的社区健康管理服务运行新模式的思路与建议［J］. 中国全科医学，2014，17 (1)：81 – 83，90.
4 李军生. 基于 Hadoop 的医疗信息共享平台的设计与实现［D］. 长沙：湖南大学，2013.
5 牟磊. 一种面向电子健康档案的 Hadoop 云计算平台实现［D］. 成都：电子科技大学，2014.