基于 Hadoop 的医疗大数据分析系统的研究与设计①

廖 亮[1], 虞宏霄[2]

[1](南华大学 附属南华医院, 衡阳 421002)
[2](南华大学 计算机学院, 衡阳 421001)

摘 要: 针对目前部分医院对于庞大医疗数据处理能力匮乏问题, 设计了一个基于 Hadoop 的医疗大数据分析系统. 该系统可提供辅助诊断和医疗数据统计两大功能, 同时融合了多节点分布式计算技术, 可以根据患者的医检数据快速生成初诊结果, 并能够有效地改善传统医疗数据信息系统分析效率较低的现状.

关键词: Hadoop; 智能医疗; 大数据; HIS

Research and Design of Medical Mega Data Analysis System Based on Hadoop

LIAO Liang[1], YU Hong-Xiao[2]

[1](The Affiliated Nanhua Hospital, University of South China, Hengyang 421002, China)
[2](School of Computer Science and Technology, University of South China, Hengyang 421001, China)

Abstract: For solving the problem of lack of large medical data computing ability in some hospitals presently, a medical mega data analysis system based on Hadoop is designed. The system can provide two functions of auxiliary diagnosis and medical data statistics, combing with the technology of multi-node distributed computing. So, the preliminary diagnosis results can be concluded immediately according to patient's medical data. And at the same time, the proposed system also has more efficient analysis capability than the traditional hospital information system.

Key words: Hadoop; intelligent medical; mega data; HIS

1 引言

近年来, 计算机技术和互联网技术得到了前所未有的飞速发展, 人类社会迈入了大数据时代, 医疗产业信息化建设也随之不断加速. 据卫生部统计, 2014 年我国投入到医疗行业信息化建设的资金为 275.1 亿元, 2015 年总计投入规模超过 300 亿元[1]. 与此同时, 各类医疗信息数据量呈现出了爆炸式的增长趋势, 而传统的以数据仓库存储模式为主体医院信息系统(HIS)由于受到硬件成本的限制, 对于大量非结构化数据处理时容易遇到性能瓶颈, 很难做到存储能力和计算能力的双向扩展. 因此, 本文借鉴了当前大数据处理领域的最新科研成果, 设计了一个基于Hadoop的医疗大数据分析系统, 以更好地满足医院对于海量医疗数据的整合加工和定量分析的需求.

2 大数据及其处理技术概述

在 IT 系统和计算机网络的相关基础设施及应用中时刻都在产生大量的数据信息, 如何在合理时间内将各类纷繁复杂的数据进行有效地撷取管理, 并整合成为具备支持决策的实用数据源已成为现阶段的研究热点, 大数据的概念由此应运而生.

大数据的概念最早由全球知名资讯公司麦肯锡提出, 所谓大数据即 big data(或 mega data), 是指大小超出了常规数据库或工具软件的分析处理能力, 被迫采用非传统方式处理的数据集[2]. 大数据具备 4V 特征, 即 Volume(体量大)、Velocity(处理快)、Variety(类别多)、Veracity(可靠性高), 要求以高扩展存储和分布式处理方式完成数据查询及管理功能, 目前众多机构虽然已经拥有了数量较大的原始积累数据, 却普遍缺乏高效的数据挖掘分析手段, 同时数据仓库的日常维护成本也在逐年升高. 因此, 以 Hadoop 构架为代表的分布式文件系统得到了广泛的应用.

Hadoop 是由 Apache 基金会开发的适合大规模数据处理的分布式系统基础架构, 其核心部分包括 HDFS

① 收稿时间:2016-04-29;收到修改稿时间:2016-12-08 [doi:10.15888/j.cnki.csa.005845]

(Hadoop Distributed File System)和 Map/Reduce 编程模型[3]. HDFS 是一种采用主/从(master/slave)式架构, 同时具备高容错性特点, 可以通过大量部署在普通 PC 上实现多数据节点对超大数据集进行分块存储管理的分布式文件系统. 另外, HDFS 为文件访问提供“一次写入, 多次读取”的响应模型, 简化了数据一致性问题, 适合大数据流的高吞吐率操作应用. Map/Reduce 是由谷歌实验室提出的一种全新的分布式程序设计模型, 主要通过 Map(映射)和 Reduce(化简)两个步骤来并行处理大规模数据集. 首先, Map 函数在不改变原始文件列表的情况下, 对切割后的小块文件所形成的独立元素组进行逐一映射操作, 并创建多个新的列表用于保存 Map 的处理结果. 然后, 再由 Reduce 函数对映射后输出的中间文件依据 Key-Value 值进行适当的合并或缩减. 最后, 将大量结构不同甚至互不相关的原始数据经由特征抽取后产生的结果保存至磁盘[4].

3 系统构架设计

针对目前各大医院内部医疗数据信息化建设的实际运行情况, 本文所提出的医疗大数据分析系统的框架包括: 数据层、访问控制层和应用层三个部分, 系统体系结构如图 1 所示.

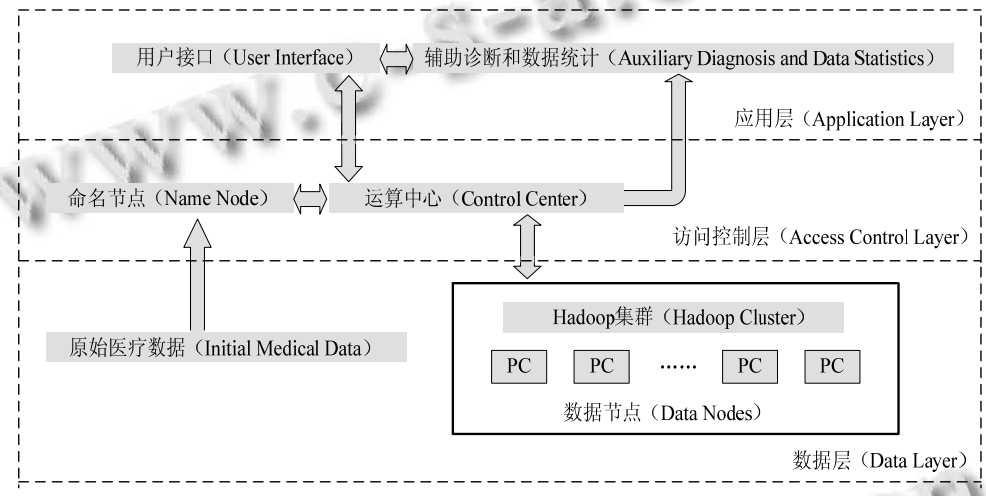

图 1　医疗大数据分析系统体系结构

该系统采用层次化结构设计原则, 最底层是数据层, 负责将现有的医院信息系统提交的各类医疗数据文件通过切割分块的形式保存至 Hadoop 集群数据节点, 实现文件的分片管理和负载均衡控制. 访问控制层是中间层, 由命名节点管理命名空间镜像以及各文件块和数据节点的对应关系, 运算中心通过调用命名节点提供的元数据信息, 对原始数据集进行 Map/Reduce 处理, 指导文件的读写流程, 并将处理结果上交至应用层. 应用层是系统的最高层, 为用户提供了操作界面接口, 用户可以通过该接口向访问控制层下达操作指令以及接收系统的辅助诊断报告和数据统计分析结果.

4 系统功能的设计与实现

该系统通过与传统的医院信息系统进行协同工作, 可以对现有单节点医疗数据库中存放各类医疗数据转为分布式存储管理; 并通过调用运算中心设计的 Map/Reduce 算法, 实现对海量数据的高效统计分析和医疗辅助诊断.

4.1 数据存储功能的设计与实现

数据层是由一系列安装了 Linux 操作系统的普通 PC 和现有医院信息系统的医疗数据库构成, Hadoop 分布式文件系统(HDFS)运行在众多 PC 构成的数据节点集群中, 主要负责对原始医疗数据进行导入和分布式存储管理, 其工作原理如图 2 所示[5].

目前现有的医院信息系统(HIS)主要由电子病历子系统(EMR)和影像归档通信子系统(PACS)构成, 其中 EMR 用于存放病人的基本信息、医检结果以及诊断报告等结构化数据, PACS 存放的是各类数字化医学影像、声音等非结构化数据. 在 Hadoop 项目中, 除 HDFS 和 Map/Reduce 编程模型外, 还包括了结构化数据仓库基础

构架Hive, 非关系型数据库Hbase, 以及传统数据仓库与HDFS之间的数据导入工具Sqoop等第三方模块.

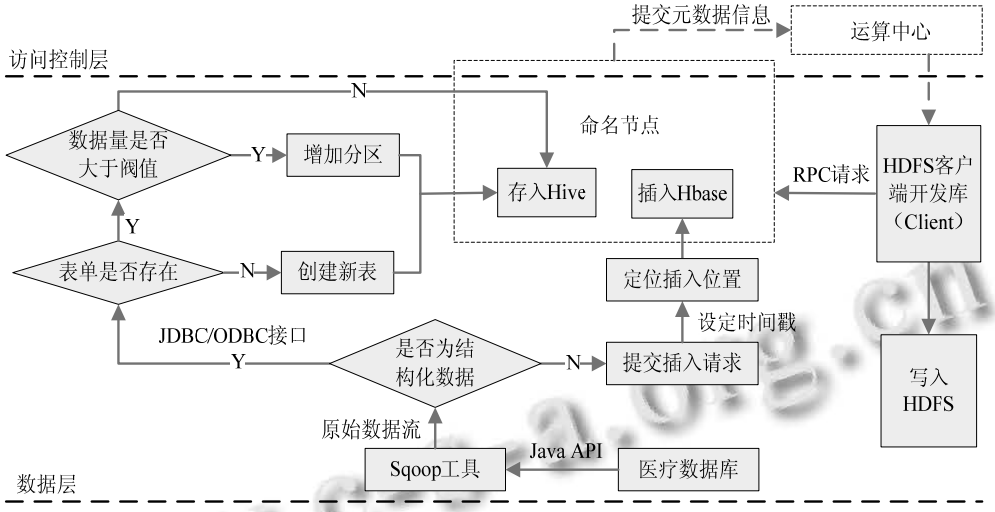

图 2 数据分布式存储实现原理

在将原始医疗数据进行分布式处理之前, 先在命名节点中安装 Hive 和 Hbase, 再利用 Sqoop 工具提供的 Java API 与传统医疗数据库连接. 需要导入的各类数据, 首先判断其是否为结构化数据, 如果是结构化数据, Sqoop 工具将通过 JDBC/ODBC 接口连接 Hive, 然后查询与该数据对应的存储表单是否已经存在, 如果不存在则创建新表后存入 Hive; 如果已经存在, 再判断数据量是否超过设定阀值(Threshold), 如果没有超过, 直接存入 Hive; 如果超过, 则需要增加分区后再存入 Hive. 当数据为非结构化数据时, Sqoop 工具将通过 Hbase 接口连接 Hbase, 并提交插入请求; 请求得到响应后, 再对 Hbase 表进行扫描并定位插入位置, 同时设定时间戳, 将数据插入 Hbase 数据库. 数据写入 HDFS 的执行流程如下所示:

(1) 客户端开发库(Client)启动数据节点, 并向上层的命名节点发起 RPC 请求.

(2) 命名节点会检查需要创建的文件是否已经存在以及创建者的操作权限, 若检查成功, 则为文件创建一个记录; 检查失败, 向客户端抛出异常.

(3) 当 RPC 写入请求得到响应后, 客户端开发库(Client)会将需要写入的文件切分成多个 Packets, 然后向命名节点申请新的 Blocks, 并将本地文件与 HDFS 数据块的映射列表, 以“块报告”的形式提交给命名节点.

(4) 命名节点向客户端返回所管理的数据节点的配置信息, 客户端将根据数据节点的 IP 地址, 以管道(Pipeline)的形式, 按顺序写入到每一个数据块节点中.

当原始医疗数据全部写入 HDFS 后, 命名节点将所有文件的元数据信息(如文件的属性; 文件的块列表; 文件块与数据节点的对应关系等)提交给运算中心, 运算中心会根据设计好的 Map/Reduce 算法对分布式文件集进行特定的读写操作和分析处理.

3.2 辅助诊断和数据统计功能的设计与实现

在患者实际就医过程中, 通常需要进行大量的医疗检查, 由于患者的体质差异, 同一类型疾病的医检项目可能会在不同患者的检查过程中呈现出不同的数据结果. 因此, 部分患者在医检过后, 还需要经过一段时间的入院观察治疗才能最终定性所患疾病的具体类型. 而在医院现存的电子病历中, 包含了众多已确诊病症的医检数据及患者的个人信息, 基于Hadoop的医疗大数据分析系统, 可以通过对 HDFS 中存放的所有电子病历文件进行Map/Reduce处理, 将不同病症的各类医检项目数据值进行区间归纳, 并生成辅助检测模板以提高医院的工作效率, 同时还可以对各年龄段患者的主要易发病进行高速数据统计. 算法实现如下[6]:

Mapper 算法:

(1) 打开电子病历文件, 当文件非空且文件未结束则循环读取字符串到变量 str 中;

(2) 如果 str=”年龄” then key1=年龄值(整数类型);

当 str=”诊断结果”时, value1=病症名称, 将(key1,

value1)写入中间文件;

(3) 如果 str=”诊断结果” then key2=病症名称(字符串类型).

当 str=”医检数据”时, 修改 key2=病症名称 & 该病对应的某种医检项目名称(字符串类型);

value2=与该病对应的某种医检项目的医检结果数据值,

将与该病对应的每一种医检项目分别生成(key2, value2)写入中间文件.

Reducer 算法:

(1) 创建 Hash 表 ht;

(2) 当 key 值为整型时, key=与 key1 对应的年龄段 & value1, value=value+1, 将(key, value)写入 ht;

(3) 当 key 值字符串类型时, 如果 value > max 则 max = value;

key= key2, value= max, 将(key, value)写入 ht;

如果 value < min 则 min = value;

key= key2, value= min, 将(key, value)写入 ht;

(4) 将 ht 中的每一组(key, value)写入最终结果文件;

由于 Mapper 算法所提供的(key, value)中的 key 与 value 可能为不同的数值类型, 而 HashTable 可以支持任何类型的 key-value 键值对, 因此需要创建一个 Hash 表用于保存 Reduce 处理的临时结果.

当系统做数据统计时, 首先判断接收到的 key 值是否为整型, 是整型则按照数值大小归入对应的年龄段, 然后将该年龄段与所患病症组合成新的 key 值, 并判断该 key 是否已经存在于 ht 中, 如果尚未存在, 则在 ht 中加入该 key; 如果已经存在, 则将该 key 对应的 value 值(即该年龄段患该种疾病的人数)加 1.

当接收到的 key 值为字符串类型时, 如果判断该 key 对应的 value 值大于现存的最大值 max, 则将 max 替换成该 value; 如果判断该 key 对应的 value 值小于现存的最小值 min, 则将 min 替换成该 value. 如此反复比对, 即可实现某种疾病不同患者的各项医检项目数据值的区间归并, 最终将所有疾病的医检项目数据值区间进行分类提取, 形成辅助检测模板.

5 系统性能测试

为了测试系统的实际运行效果, 作者为本系统配置了 20 个数据节点, 随机抽取了各年龄段共计 50535 份电子病历进行了数据分析, 生成的易发病统计表如表 1 所示.

表 1 各年龄段易发病统计

0-12 岁	易发病名称	病毒感染	呼吸道疾病	消化道疾病	外伤	手足口病	发热	阑尾炎
(6196 人次)	所占比例(%)	11.9	28.24	11.43	19.97	5.26	13.87	7.13
……	……	……	……	……	……	……	……	……
60 岁以上	易发病名称	心血管病	呼吸道疾病	消化道疾病	肿瘤	脑血管病	结石	糖尿病
(18754 人次)	所占比例(%)	18.24	13.01	13.45	17.64	18.52	7.63	9.48

最后, 将本系统与现有的医疗数据库利用程序控制台的 Begin()和 End()函数中所记录的时间进行了工作效率比较, 两者在数据处理过程中的时间消耗如表 2(不包含数据写入磁盘所消耗的时间)所示. 工作效率对比折线图如图 3 所示.

表 2 大数据分析系统与传统数据库的耗时对比

	所消耗的时间(ms)	
电子病历数量(份)	大数据分析系统 (Hadoop)	现有医疗数据库 (SQL Server)
10000	103	2179
20000	148	4127
30000	169	6396
40000	187	8783
50000	206	11025

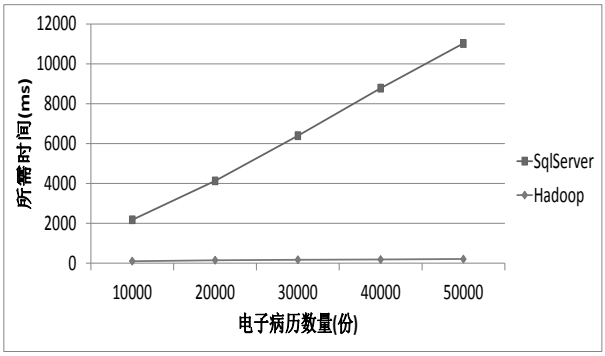

图 3 工作效率对比

通过实验结果可以看出, 随着被处理的电子病历的数量增加, 传统的单节点数据库耗时呈线性增长趋

势; 而基于 Hadoop 的医疗大数据分析系统, 由于在数据统计过程中采用了分布式的处理方式, 时间消耗并未显著增长.

6 结语

本文的主要创新点有两个: (1)提出了一个基于 Hadoop 的大数据分析系统的体系结构, 并对该系统所提供的功能进行了详细的分析与设计; (2)为医疗辅助诊断和数据统计设计了一个切实可行的 Map/Reduce 算法, 优化了医疗诊断流程并实现了海量数据的高速统计. 最后, 通过具体实验验证了基于 Hadoop 的医疗大数据分析系统比传统的单一节点数据库具备更高的工作效率.

参考文献

1 中国产业信息网.2015 年中国医疗信息化市场运营报告, http://www.chyxx.com, 2015.

2 马建光,姜巍.大数据的概念特征及应用,国防科技,2013, 34(2):10–17.

3 White T. Hadoop: The Definitive Guide. 3rd Ed. O'Reilly Media, 2012, 5.

4 Gillick D, Faria A, DeNero J. Mapreduce: Distributed computing for machine learning. Berkley, 2006, 12.

5 Shvachko K, Kuang H, Radia S, et al. The hadoop distributed file system. 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST). IEEE. 2010.

6 Boyd S, Parikh N, Chu E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning, 2011, 3(1).