

区域医疗大数据白皮书 用友医疗大数据应用案例

医疗行业大数据

医疗大数据的应用，关键是整合所有的数据集，为机构和政策制定者来找到如何刺激经济并降低共享数据的技术门槛。

制药行业科研数据

数据拥有者 — 制药企业，学术界

数据集 — 临床试验，高通量药物筛选库 (HTS)

特点 — 数据的采集是数字化的；集合了来自个体或临床的基因或蛋白数据，它可以帮助识别药品不良反应和新功效

活动 (报销) 和成本数据

数据拥有者 — 支付方，医疗服务提供方

数据集 — 服务的使用，成本估算

特点 — 数据获取不标准，分散，老系统，不兼容格式

临床数据

数据拥有者 — 医疗服务提供方

数据集 — 电子病例，医疗影像

特点 — 数据分析的意义在于更加有力地推进电子病历的有意义使用 (MU) 以及质量的管理

病人行为和情绪数据

数据拥有者 — 院外消费者和利益相关者 (e.g., 零售渠道)

数据集 — 病人行为和爱好，零售历史，运动鞋采集的信息

特点 — 采集的是生活习惯等反映生活偏好的数据，用于健康管理的策略和计划的制定

以上信息来源: McKinsey Global Institute Analysis



中国医疗信息化的趋势

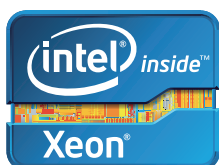
中国政府于 2009 年 4 月出台了《中共中央国务院关于深化医药卫生体制改革的意见》和《国务院深化医药卫生体制改革近期重点实施方案 (2009-2011)》，深化医院卫生体制改革工作从此进入实质实施阶段。

医改方案提出了医药卫生体制改革未来十年的总体目标，即建立健全覆盖城乡居民的基本医疗卫生制度。为群众提供安全，有效，方便，价廉的医疗卫生服务。实行人人享有基本医疗卫生服务，使人民群众健康水平进一步提高。

医药卫生信息化作为深化医药卫生体制改革“四梁八柱”的八柱之一，明确提出要建立实用共享的医药卫生信息系统。建立共享的医药卫生信息系统，区域卫生信息平台是基础，健康档案是纽带，基于健康档案的区域卫生信息平台是实现医药卫生信息共享的最佳实践。

在“健康中国 2020”的研究中也制定了未来的总体建设目标：改善城乡居民健康状况，提高国民健康生活质量，

UFIDA 用友



减少不同地区健康状况的差异。主要健康指标基本达到中等发达国家水平。

- 到 2015 年，基本医疗卫生制度初步建立，使全国国民人人拥有基本医疗保障，人人享有基本公共卫生服务。医疗卫生服务可及性明显增强，地区间人群健康状况和资源配置差异明显缩小。国民健康水平居于发展中国家前列。
- 到 2020 年，完全覆盖城乡居民的基本医疗卫生制度，实现人人享有基本医疗卫生服务，医疗保障水平不断提高。卫生服务利用明显改善，地区间人群健康差异进一步缩小，国民健康水平达到中等发达国家水平。

国际数据公司（IDC）发布的《中国医疗行业 IT 解决方案市场 2012-2016 预测与分析》报告显示，2011 年中国医疗行业 IT 花费是 146.3 亿元，比上年增长 28.9%。IDC 预测 2016 年中国医疗行业 IT 花费市场规模将达 339.9 亿元，2011-2016 年年复合增长率为 18.4%。

发展智能医疗是全球大趋势，对我国深化医药卫生体制改革、解决看病难及看病贵的问题有着很现实的意义。中国政府、医院近年来在医疗信息化领域投入较大，医疗信息化水平得到迅速提高。但总体来说，中国医疗信息化水平还处于初级阶段，存在医院之间信息共享机制不健全、全国各地发展不平衡、相关标准不统一等诸多问题。

卫生部于 2010 年编制了全国卫生信息化“十二五规划”，简称“3521-2”工程。具体来说，“3”代表国家、省、市三级卫生信息平台；“5”代表医疗服务、公共卫生、医疗保障、药品保障和综合管理五大卫生信息化业务领域；“2”代表健康档案和电子病历两大基础资源库；“1”代表卫生信息专网；后一个“2”代表信息安全体系和标准规范体系。其中，市级卫生信息平台，即区域卫生信息化，是卫生信息化的核心内容之一。

以“35212”工程为蓝图的信息化平台建设，极大地促进了区域内健康档案信息的采集和共享，国家在支持在上海、浙江、安徽、重庆、新疆等五省（区、市）开展基层医疗卫生信息系统建设试点工作的基础上，各地也积极开展基于居民电子健康档案区域医疗信息平台的建设。从各地的建设来看，积累了大量的健康档案数据。

中国医疗行业中的“大”数据

伴随着中国医疗卫生服务的信息化进程推进，必将产生大量的数据。这些数据主要来源于医疗业务活动、健康体检、公共卫生 9 项服务等医疗卫生服务。数据内容主要包括来自医院的大量电子病历、区域卫生信息平台采集的居民健康档案等。其中大量充斥着非结构化 / 半结构化的数据，包括图像，office 文档，以及 XML 结构文档等。

随着国家积极倡导的 3-5-2-1 区域医疗系统的建设，预计在全国会出现上百个医疗数据中心，每个数据中心都将承载近 1000 万人口的医疗卫生数据并提供各种类型的服务。根据估算，中国一个中等城市（一千万人口）50 年所积累的医疗数据量就会达到 10PB 级（来源：赛迪网）。而随着个人健康管理的推进，将产生越来越多的个人日常健康监测信息，这个数据的规模和增长速度将远超想象。

从目前各地的医疗信息系统的建设来看，尽管数据已经电子化，但绝大多数的医疗数据是处于归档状态。如果要想快速检索是十分复杂的，这些数据仍然分散的存储于不同的业务系统中。过去不是没有整合这些数据的需求，而是缺乏适合的技术手段。

单纯的使用传统的存储技术和分析处理方法，存在两个问题：一个是多样的和变化的数据格式，另一个是大数据量的数据获取速度。

大数据在医疗行业的应用

大数据解决方案在医疗行业的应用场景众多。在中国医疗行业，大数据技术目前最具实际应用的应用在于居民健康档案数据的管理和服务。“健康档案是个人全生命周期的医疗 / 健康数据的管理”。

- 从医生及卫生行政管理人员的角度来看，全生命周期的健康档案调阅有着现实意义。例如，对于慢性病患者，以往病程的变化，治疗的过程都对医生诊断和处置有着重要的辅助作用。过敏史，不良反应这些数据对避免出现医疗差错和事故也有着积极的作用。
- 对于海量的医疗及健康数据进行统计和分析，为管理决策、监管实施，提供了更为科学的依据。
- 传统的临床科研往往基于抽样调查进行，而随着健康档案数据的丰富可以大幅减轻工作量，同时提高科研数据的质量和数量以及数据处理的效率。

未来的大数据的利用前景是十分广阔的，不仅用于临床诊断，临床科研，而且为政府公共卫生决策及个人管理健康都会发挥积极的作用。

大数据应用中的挑战

大数据在医疗卫生行业的应用，也面临着来自不同方面的挑战。根据我们的观察，这些挑战远远超出了技术领域，主要包括以下几个方面：

首先是意识问题。我们是不是充分认识到数据服务的必要性，数据服务的复杂性和数据服务的迫切性。可能还有些企业用户在设计信息系统的时候还沿用过去的思路和方法设计，这可能就是还缺乏这样的意识。电子商务的发展，移动互联网的发展都让人感到是突如其来的迅猛，我们如果没有前瞻的思维，一旦医疗数据暴涨，数据服务需求增长，我们势必无法应对。

其次是人才的问题。目前在大数据方面还是人才匮乏。真正掌握这个技术的公司，人员少之又少。期望有实力的公

司，企业积极推动这项技术在医疗领域的应用，带动大数据的发展和人才的培养。

再次是业务应用模式的探索。由于很多企业对医疗业务的不熟悉，无法真正发现医疗中运用大数据的关键业务，从而制约了大数据的发展。

区域医疗大数据整体框架

目的

在很多区域卫生信息化（RHIN）的项目的实施过程中，开发商多采用传统的数据库来处理健康档案的存储和服务。有些服务商也考虑采用了某些 No SQL 系统对非结构化的健康档案进行存储管理，例如使用 MongoDB。然而，使用单一的组件和产品，无法形成整体的解决方案。使用 Intel 大数据解决方案来进行健康档案的管理，很好的处理了数据服务和业务生产系统的关系。将业务系统和数据服务有机的分离开来，从性能和动态扩展方面都奠定了良好的基础。

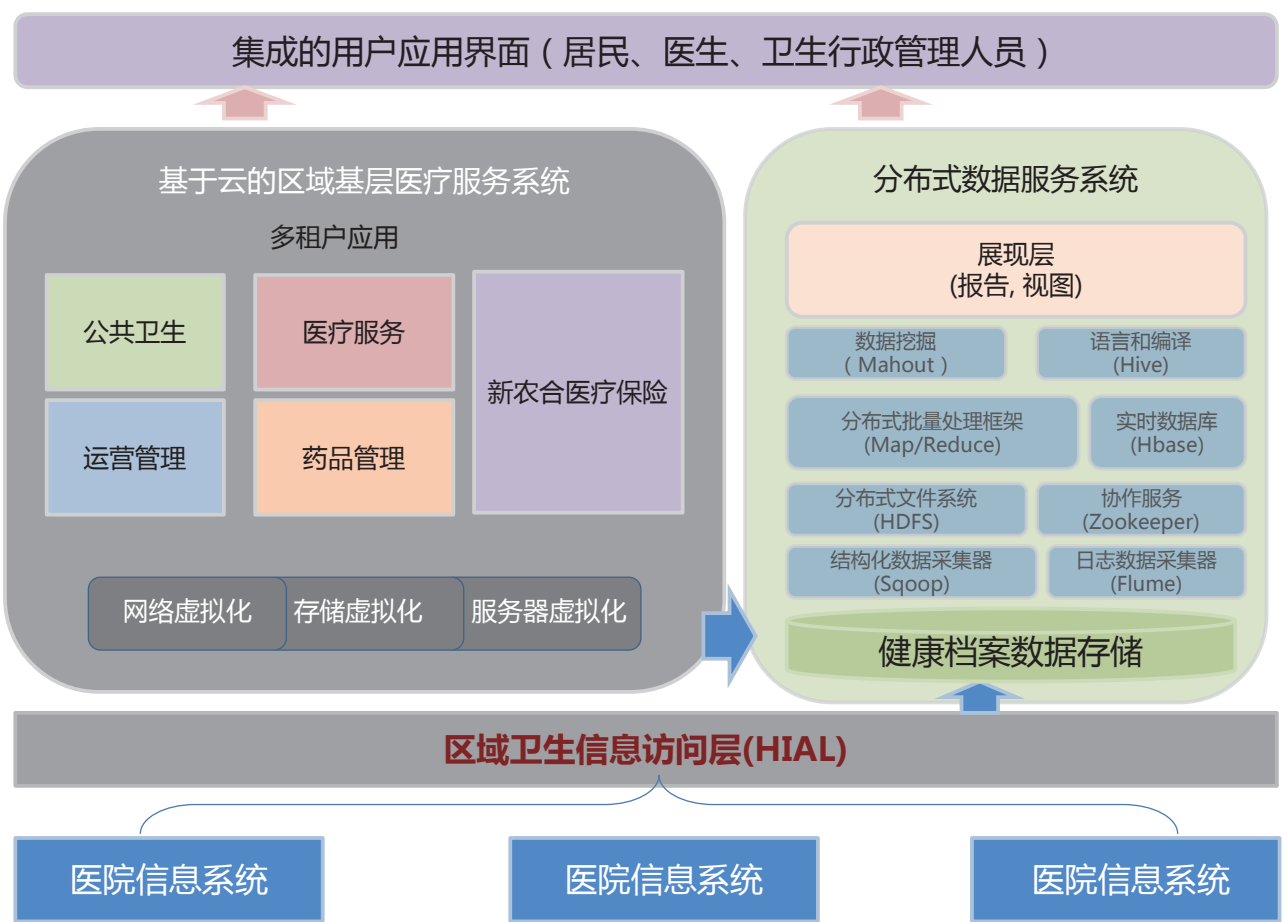


图 1 区域医疗（RHIN）及基层医疗机构信息系统大数据解决方案

定位

大数据解决方案在区域医疗平台中主要的定位在于对健康档案的管理和服务，它将伴随着区域医疗平台的建设而实施。从图中可以看出，大数据解决方案是架构在 HIAL 之上的数据服务系统。它将由 HIAL 采集来的医疗数据进行有效的存储和管理。面向医生，管理人员和病患提供综合的数据服务。它和基于虚拟化云技术的基层医疗机构信息系统进行整合，一方面采集来自基层的医疗相关数据，另一方面又为基层医疗机构提供全局的数据服务，这些服务的内容经过整合到医生，管理人员及病患的业务系统的界面中，使得用户在日常操作中平滑的获得数据服务。

功能

大数据解决方案在区域医疗中的功能包括基本服务，数据分析及依从性管理三个层次。基本服务内容包括提供基本的存储，查询，浏览。分析服务将针对主题，对临床数据，公共卫生管理数据，绩效考核数据以及农合付费管理等不同领域进行分析，在海量的不同结构的数据中找到可能的趋势和风险。依从性管理应该是大数据服务的高级阶段，它利用业务系统中的业务活动，触发数据服务请求，数据服务经过对大数据集的调阅，分析给出特定性结果。从而控制和导引业务操作的路径。这种类型的服务只有在大数据技术的支撑下，才能获得实时的，准确的，可用的应用。

价值

针对医疗业务数据的多样性，多变性及大规模的特点。大数据技术在医疗中发挥着它特有的价值。Solution Stack 目的旨在为目前各地建设的区域医疗平台，针对数据管理及服务系统建设提供参考。其价值在于能够为未来的区域医疗健康档案服务奠定一个良好的基础，并且从成本的角度，大大节省投资，同时满足未来的扩展性要求。

英特尔大数据解决方案

英特尔提供拥有强大计算能力的硬件平台，并且提供基于 Hadoop 的大数据软件平台，形成了为企业级客户服务的完整的大数据解决方案。

英特尔用“芯”支持大数据，至强 E5 处理器可用于大数据应用中的分布式处理和高性能计算，E7 则可以服务关系型数据库和商业智能技术及应用。

基于英特尔®至强®处理器的双路 / 多路平台所具备的开放

式、普及性、易优化、灵活等特点，是实现大数据的最理想平台。英特尔®至强®处理器计算能力，在计算能力密切相关的数据中心和超级计算等方面已经得到了充分的体现。此外，在大数据处理所需要的高可扩展性上，英特尔®至强®处理器在服务器和存储系统上也具有无可比拟的优势。对于商业智能来说，基于英特尔®至强®处理器的多路平台还具备高性能、高能效、灵活扩展以及高性价比等优势。

英特尔针对大数据的分发和管理需求，提供针对英特尔平台优化的 Hadoop 产品和服务。

Hadoop

- 最重要的大数据平台之一，来自开源 Apache 项目。
- 使用一个简单的编程模型，以支持在计算机集群上对大数据集进行分布式处理。
- 其完整的技术堆栈包含了通用实用程序、一个分布式文件系统、分析和数据存储平台以及一个用于管理分布式处理、并行计算、工作流程与配置管理的应用层。
- 除了提供高可用性外，Hadoop 框架还能比传统方法更为经济高效地处理大型的、复杂的或非结构化的数据集，可带来出色的可扩展性与速度。

英特尔 Hadoop 发行版 (Intel Distribution of Hadoop)，它包含了 Apache Hadoop 开源项目的完整发行版以及一个增值的监测和管理控制台 Intel Manager for Hadoop* 软件。此产品由英特尔提供充分支持，英特尔将承担起发行版密集的更新和维护流程。

- Intel 的 Hadoop 发行版是经过测试和验证的稳定版本，在客户生产环境成功部署运营，可以确保客户生产环境 7x24 小时不间断运行。
- Intel 的 Hadoop 发行版包括了 Intel 针对现有客户在实际使用中出现问题的解决方法以及大量改进和优化。这些改进弥补了开源 Hadoop 在实际使用中的缺陷和不足，并且包含了大量的性能优化。

- Intel 的集群管理工具和安装工具简化了 Hadoop 的安装和配置。可以根据用户的硬件环境自动生成最优化的集群配置，充分发挥集群的计算能力。
- 基于 Intel 在云计算研发上的经验积累，提供从项目规划到实施各阶段专业的咨询服务，帮助客户构建高可扩展高性能的分布式系统。
- 结合 Intel 的硬件平台，提供全面的软硬件解决方案设计。

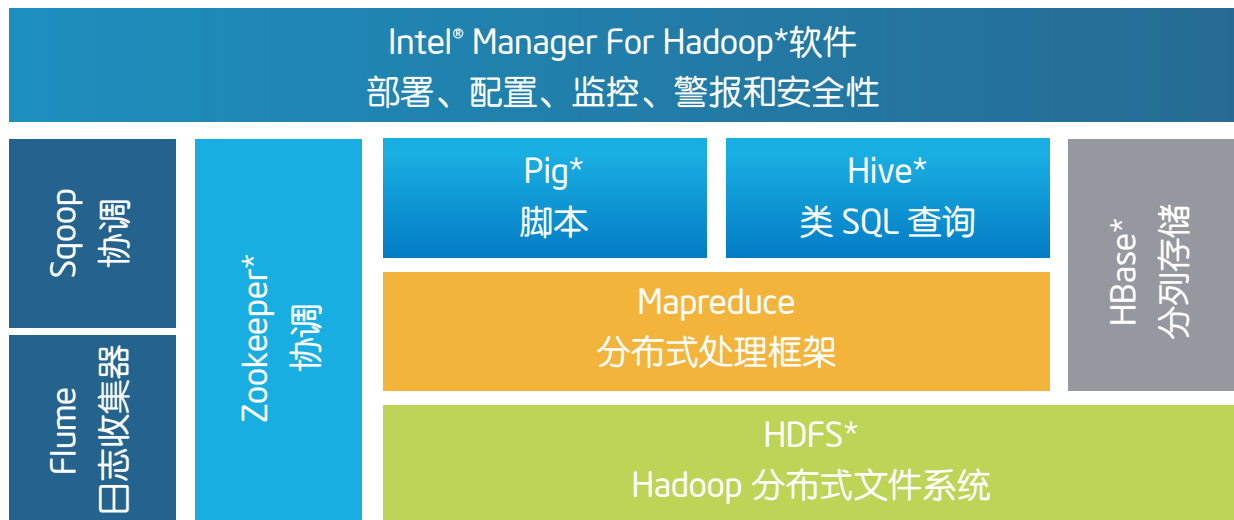


图 2 英特尔 Hadoop 发行版

用友区域医疗平台大数据应用案例

锦州，户籍人口规模约 310 万，辽宁西部城市群区域性中心城市和重要港口城市。2012 年 3 月，锦州市政府正式启动“健康城市”战略。建立以城市为单位建设的区域卫生数据中心，以居民健康卡为纽带，覆盖卫生信息化的主要业务领域，存储的数据将包括居民健康档案、电子病历、公共卫生及综合管理等相关数据，20 年左右的数据量预计将达到 PB 级。随着“健康城市”理念的提出，个人健康数据将伴随着居民对于个人健康的关注，以及健康管理相关服务的推广而大幅增加。

挑战

用友医疗卫生信息系统有限公司（简称：用友医疗），作为锦州“健康城市”战略区域卫生数据中心的主要建设方，拥有大量省市级以及区县级智慧健康城市区域卫生解决方案项目建设成功经验。用友医疗在采用传统关系型数据库在锦州的区域卫生数据中心建设中，逐步发现一些潜在的建设问题：

数据量庞大

- 区域卫生数据中心以城市为单位，将存储健康档案、电子病历、卫生管理三大数据库，锦州作为 300 万人口

的中等规模城市，其卫生数据中心的规模在 20 年后预计会达到 PB 级：

- 传统关系型数据在大数据存储实现中存在局限性，在一张表中存储 500GB 之上的数据存在性能方面的问题。

数据类型复杂多变

锦州的区域卫生数据中心，将会存储大量的非结构化数据和半结构化数据。如果采用传统关系型数据库，则面临众多不易解决的问题：

- PACS 影像，B 超、病理分析等业务产生的非结构化数据大小不一，从数百 KB 到数百 MB 都有，病人的一次诊断活动中需要存储、调阅数百张影像数据；
- 临床电子病历数据一般采用符合 HL7 CDA 标准的 XML 文件格式，文件格式随着时间变化，会不断的演变；
- 由于医疗卫生行业的业务复杂性，很难制定统一的数据标准，这就给数据访问和交换带来新的挑战。

另外，未来的数据处理也将面临巨大挑战，譬如未来海量数据的存储，备份，扩容等，快速的检索特定数据，以及高效的数据交换等等，这也是锦州区域卫生数据中心未来必会产生的新的需求。

基于 Hadoop 实现海量数据的处理和分析

为了解决上述诸多问题，英特尔公司协助用友医疗进行了合理的架构分析和指导，对于基于大数据分析的解决方案进行了深入的探索和研究。并且制定了基于英特尔大数据解决方案的区域卫生数据中心建设目标：

- 在海量数据情况下，数据中心必须具有文档快速检索的性能；

- 随着未来医疗系统的升级，医疗标准的版本升级，数据中心必须能够解决存储模式如何满足数据模式的更新的问题；
- 数据中心必须具备水平扩展能力和对应用程序透明的能力，要求能做到的底层扩展对上层业务的隔离，通过更多的服务器成比例的透明化扩展容量和性能。

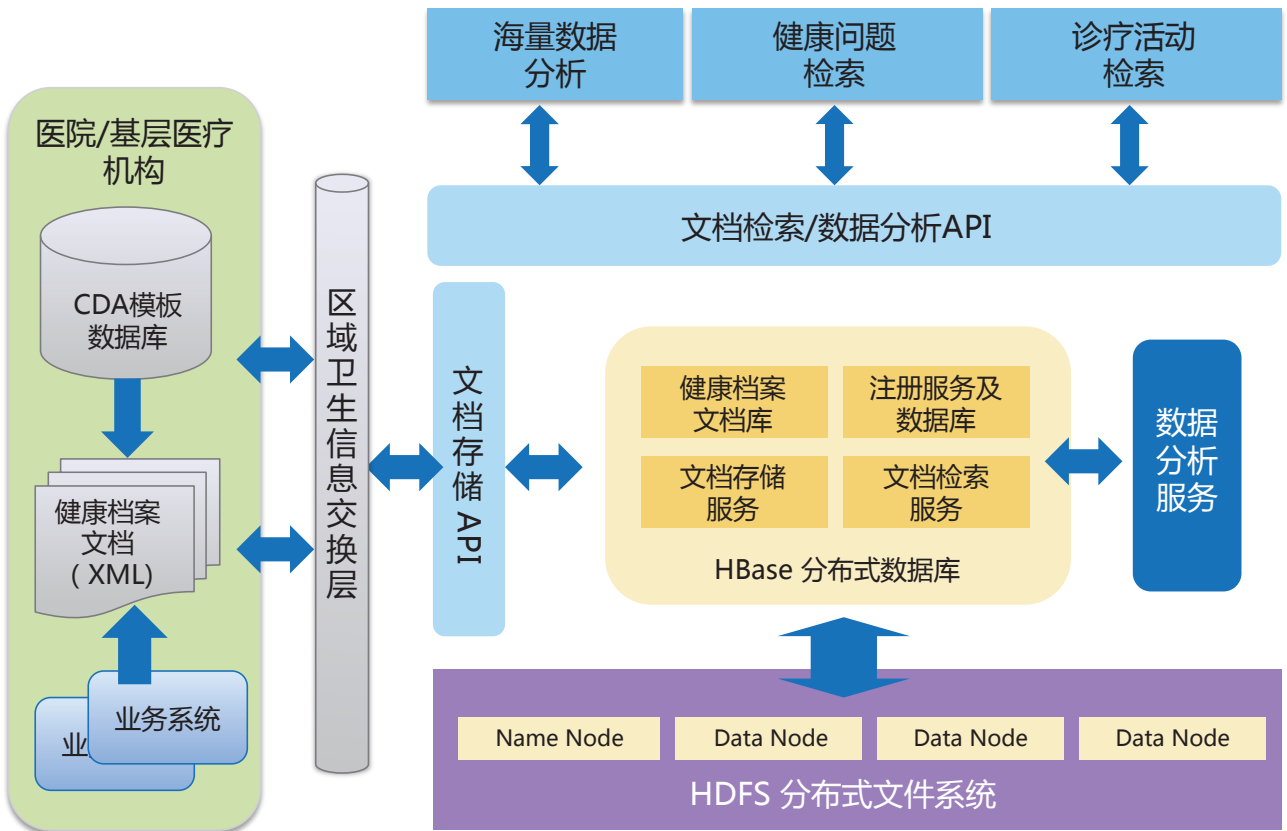


图 3 基于 Hadoop 的区域卫生信息平台数据处理流程

- 个人基本健康信息登记：1 千万条
- 门诊摘要：3 千万条
- 处方记录：1 亿条
- 糖尿病专档：2 百万条
- 糖尿病随访：2 千 4 百万

区域内的各医疗卫生服务机构，调用区域卫生信息平台提供的服务，将依从 HL7 CDA 模板设计的健康档案文档（XML 格式），上传至区域卫生信息平台。采集的 CDA 文档类型包括个人基本健康记录登记、门诊摘要、处方记录、糖尿病专访和糖尿病随访等不同类型的文档。文档上传后，通过 XML 解析，提取关键数据元素（META-DATA），调用文档存储服务将关键数据元素和 XML 文档存储在 hbase 数据库中，形成健康档案文档库。

基于大数据的区域卫生信息平台提供文档检索服务和数据分析服务，通过开放的文档检索 / 数据分析接口，针对个人（居民、患者）、医生和卫生管理人员提供不同类型的数据和业务服务，满足不同用户类型对于海量数据访问、处理和分析的需求。

结论

经过多次反复单业务负载测试、大数量测试和调优等多种技术相结合的手段，英特尔公司和用友医疗在锦州区域卫生数据中心，利用基于 Intel 英特尔 E5 至强服务器平台，以及英特尔 Hadoop 发行版 (Intel Distribution)，成功的搭建了的完整的区域医疗大数据计算架构。该架构可以满足海量数据（一亿以上记录数）的高并发检索和实时数据分析的性能要求。证明了基于英特尔至强 E5 服务器平台以及英特尔 Hadoop 发行版的大数据整体解决方案，可以作为基于大数据分析的智慧健康云服务平台提供数据处理、检索和分析等服务的基础架构。

价值

业务价值

英特尔与用友医疗合作的基于大数据量优化的整体解决方案：

- 使得医生可以快速检索就诊患者的个人基本信息、既有病史、就诊及处方等重要信息。通过辅助信息，帮助医生做出更准确的诊断，并且有效避免重复用药、药物不良反应等；
- 支持高并发的大数据检索，使得个人（居民、患者）通过便捷快速的方式访问自己的健康档案成为可能。与居民健康教育相结合，可以帮助居民建立起个人健康管理的理念，从治疗转向预防；
- 支持灵活的数据分析建模和数据分析挖掘，为区域卫生管理者提供了接近实时的数据分析结果，可以广泛应用在疾病控制、医疗行为监管、医疗质量管理等领域。

IT 价值

英特尔与用友医疗合作的基于大数据量优化的软硬件解决方案，实现了：

- 海量数据存储：应对超过一亿条记录(文档)的存储需求；
- 数据格式动态扩展：Hbase 提供的列式存储方式，可以轻松应对数据格式的灵活调整；
- 海量数据快速检索：高并发的并行海量数据检索，满足居民和医生实时检索健康档案的需求；
- 统计分析：开放的统计分析框架，接近实时的统计分析能力；
- 平滑扩容：横向扩展性能卓越，应对未来业务和数据量的不断增长。

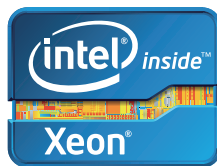
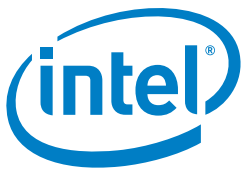
经济价值

- 降低主机存储建设成本：与传统数据库和小型机的方案相比，英特尔大数据解决方案具有更好的开放性和更好的经济性；
- 高可扩展性，集成的图形化界面提供了便捷的节点管理和横向扩展功能，大幅降低了管理成本；
- 充分利用现有硬件资源，降低升级成本。



参考资料

- 1) McKinsey Global Institute Analysis
- 2) 基于健康档案的区域卫生信息平台建设指南 (试行), 卫生部
- 3) 基于健康档案的区域卫生信息平台建设技术解决方案 (试行), 卫生部
- 4) 基于电子病历的医院信息平台建设技术解决方案 (1.0 版), 卫生部
- 5) 区域卫生信息平台白皮书, 英特尔
- 6) “健康中国 2020” 战略研究报告, “健康中国 2020” 战略研究报告编委会
- 7) 中国医疗行业 IT 解决方案市场 2012-2016 预测与分析, IDC



UFIDA 用友

法律声明:

本文档提供有关英特尔®产品的信息。本文档并未以明示或暗示, 或以禁止反言或其他方式授予任何知识产权许可。除英特尔在其产品的销售条款和条件中声明的责任之外, 英特尔概不承担任何其他责任。并且, 英特尔对英特尔®产品的销售和 / 或使用不作任何明示或暗示的担保, 包括对产品的特定用户适用性、适销性或对任何其他任 何专利权、版权或其他知识产权的侵权责任等, 均不作担保。除非已获得英特尔的书面认可, 否则英特尔®产品并非设计为或用于任何英特尔®产品故障可能导致人身 伤害或生命危险的应用。

版权所有©2013 英特尔公司。保留所有权利。英特尔、Intel 标识、Xeon、至强、Xeon inside 是英特尔公司在美国和其他国家 (地区) 的商标。

* 文中涉及的其它名称及商标属于各自所有者资产。