第1章 绪论
Chapter 1

1.1 大数据的特点

大数据（Big Data）通常被认为是一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合。随着大数据研究的不断深入，我们逐步意识到大数据不仅指数据本身的规模，而且包括数据采集工具、数据存储平台、数据分析系统和数据衍生价值等要素。IBM 提出大数据的 5V 特点：Volume、Velocity、Variety、Value、Veracity。这里，我们归纳大数据主要具有以下特点。

1）数据量大

现有的各种传感器、移动设备、智能终端和网络等都无时无刻不在产生数据，数量级别已经突破 TB 级，发展至 PB 乃至 ZB 级，统计数据量呈千倍级别上升。伙伴产业研究院（PAISI）研究统计，2017 年全年数据总量超过 15.2ZB，同比增长 35.7%；到 2018 年，全球数据总量达 19.4ZB；未来几年，全球数据的增长速度为每年 25%以上；以此推算，到 2020 年，全球的数据总量将达到 30ZB。

2）类型多样

当前大数据包含的数据类型呈现多样化发展趋势。以往数据大多以二维结构呈现，随着互联网、多媒体等技术的快速发展和普及，视频、音频、图片、邮件、HTML、RFID、GPS 和传感器等产生的非结构化数据每年都以 60%的速度增长。预计非结构化数据将占数据总量的 80%以上[1]。

3）运算高效

由 Apache 基金会所开发的 Hadoop 利用集群的高速运算和存储，实现了一个分布式运

行系统，以流的形式提供高传输率来访问数据，适应了大数据的应用程序。而且，数据挖掘、语义引擎、可视化分析等技术的发展，使得可以从海量数据中深度解析和提取信息，实现数据增值。

4）产生价值

数据中的价值是大数据的终极目标，企业可以通过大数据的融合获得有价值的信息。特别是在竞争激烈的商业领域，数据正成为企业的新型资产，企业追求数据最大价值化。同时，大数据价值也存在密度低的特性，需要对海量的数据进行挖掘分析才能得到真正有用的信息，形成用户价值。

1.2 大数据平台

大数据有三种常用的处理框架：Hadoop、Spark 和 Storm。

Hadoop 是一种专用于批处理的处理框架，是首个在开源社区获得极大关注的大数据框架。Hadoop 基于谷歌发表的海量数据处理相关的多篇论文，重新实现了相关算法和组件堆栈，使大规模批处理技术变得更容易使用。新版 Hadoop 包含多个组件，通过配合使用可处理批数据。

（1）HDFS（Hadoop Distributed File System）是一种分布式文件系统层，可对集群节点间的存储和复制进行协调。HDFS 确保了无法避免的节点故障发生后数据依然可用，可将其用作数据来源，用于存储中间态的处理结果，并可存储计算的最终结果。

（2）YARN（Yet Another Resource Negotiator）可充当 Hadoop 堆栈的集群协调组件。该组件负责协调并管理底层资源和调度作业的运行。通过充当集群资源的接口，YARN 使得用户能在 Hadoop 集群中使用比以往的迭代方式运行更多类型的工作负载。

（3）MapReduce 是 Hadoop 的原生批处理引擎。Spark 是一种具有流处理能力的下一代批处理框架。与 Hadoop 的 MapReduce 引擎基于相同原则开发而来的 Spark 主要侧重于通过完善的内存计算和处理优化机制加快批处理工作负载的运行速度。

Spark 可作为独立集群部署（需要相应存储层的配合），也可与 Hadoop 集成并取代 MapReduce 引擎。与 MapReduce 不同，Spark 的数据处理工作全部在内存中进行，只在一开始将数据读入内存，以及将最终结果持久存储时需要与存储层交互。所有中间态的处理结果均存储在内存中。使用 Spark 而非 MapReduce 的主要原因是速度。在内存计算策略和先进的 DAG（Directed Acyclic Graph，有向无环图）调度等机制的帮助下，Spark 可以用更快的速度处理相同的数据集。Spark 的另一个重要优势在于多样性，可作为独立集群部署，或与现有 Hadoop 集群集成。Spark 可运行批处理和流处理，运行一个集群即可处理不同类型的任务。除了引擎自身的能力外，围绕 Spark 还建立了包含各种库的生态系统，可为机器学习、交互式查询等任务提供更好的支持。相比 MapReduce，Spark 任务更易于编写，因此可大幅提高生产力。

相比之下，Spark 是一个专门用来对分布式存储的大数据进行处理的工具，它不进行分布式数据的存储。由于 Spark 处理数据的方式不同，其数据处理速度比 MapReduce 快很多。MapReduce 分步对数据进行处理：从集群中读取数据，进行一次处理，将结果写到集群；再从集群中读取更新后的数据，进行下一次的处理，将结果写到集群等。Spark

在内存中以接近“实时”的时间完成所有的数据分析：从集群中读取数据，完成所有必需的分析处理，将结果写回集群。因此，Spark 的批处理速度比 MapReduce 快近 10 倍，内存中的数据分析速度则快近 100 倍。如果需要处理的数据和结果需求大部分情况下是静态的，且时间允许等待批处理完成，则 MapReduce 的处理方式也是完全可以接受的。如果需要对流数据进行分析，如来自现场的传感器收集的数据或者需要多重数据处理的应用，那么应该使用 Spark 进行处理。目前，大部分机器学习算法都需要多重数据处理。通常，Spark 可以应用在实时的市场活动、在线产品推荐、网络安全分析、机器日志监控等场景。

Storm 是一种侧重于极低延迟的流处理框架，是要求近实时处理的工作负载的最佳选择。该技术可处理非常大量的数据，通过比其他解决方案采用更低的延迟提供结果。Storm 的流处理可对框架中拓扑的 DAG 进行编排。这些拓扑描述了当数据片段进入系统后，需要对每个传入的片段执行的不同转换或步骤。拓扑包含：

（1）Stream：普通的数据流，这是一种会持续抵达系统的无边界数据。

（2）Spout：位于拓扑边缘的数据流来源，可以是 API 或查询等，从这里可以产生待处理的数据。

（3）Bolt：代表需要消耗流数据，对其应用操作，并将结果以流的形式进行输出的处理步骤。Bolt 需要与每个 Spout 建立连接，随后相互连接以组成所有必要的处理。在拓扑的尾部，可以使用最终的 Bolt 输出作为相互连接的其他系统的输入。

1.3 医疗健康大数据的应用需求

医疗健康大数据平台是以健康和医疗两大类数据的异构整合、统一存储、高效处理为基础，以深度分析和挖掘为核心，通过能力开放实现数据共享和产业链资源整合的一体化平台。作为基础支撑平台，它能够提供健康和医疗大数据应用的基础环境；针对医疗行业大数据应用特点，对来自异构业务系统，包括专业机构、公共卫生系统、院内系统、区域卫生平台的结构化与非结构化数据进行统一规划来满足医疗健康行业大数据应用平台的需求；保证系统具有高性能、高可靠、易扩展、易使用等特点，同时提供图形化的统一管理系统，简化用户的管理和维护工作。作为大数据应用平台，它在基础支撑平台的基础上，经过分布式并行数据处理、大规模数据分析和挖掘后，应用于卫生数据统计、决策分析、数据挖掘、疾病预警、健康预测、报表展现等场景。

面向大数据分析的医疗健康大数据平台架构包括异构医疗健康大数据整合、海量数据统一存储、分布式并行数据处理、医疗健康大数据分析和挖掘 4 个重要组成部分。

1）异构医疗健康大数据整合

有效的数据整合是大数据分析的前提。健康大数据主要是由各类可穿戴设备产生的多模态体征数据；医疗大数据则包括分散存储在 EMR、EHR、HIS、LIS、PACS 等医疗信息化系统中的医疗数据和其他类型的公共卫生数据。这些数据来源多样，存储在大量关系型数据库和文件中的数据需要经过数据的采集、清洗和转换过程，并经过抽取获得元数据信息，实现异构的多类型健康和医疗两大类数据的统一整合。HDFS 具有强大的可扩展性，能够支持 PB 级别的数据存储，基于 HDFS 可实现对大规模数据

的统一整合。

2）海量数据统一存储

海量数据统一存储是大数据分析的基础。异构数据完成整合后，利用针对医疗健康大数据专门设计的海量数据统一存储模型，以用户为中心进行设计，围绕着用户健康档案，按照统一的格式和规范，实现对体征、体检、病历、住院、妇幼、疾控和社保共 7 类数据的统一存储，从而真正地提高数据孤岛之间的数据共享能力，终结医疗健康数据的碎片化。海量数据统一存储模型支持对以上 7 种类型数据的增加、修改、查询和删除的操作，同时保证上述操作的可靠性和一致性。此外，由于 EMR 和 PACS 这类系统中的用户量和数据量的飞速增长，在数据的存储规模达到一定程度时，如何实现系统的存储容量自动增长和负载平衡也是一个非常关键的问题。对于上述各类数据中的重要数据，实现数据的安全、可靠存储，甚至是 7×24h 的数据存储和访问能力也是一个较大的技术挑战。

3）分布式并行数据处理

高效的分布式并行数据处理是大数据分析的关键。为了支撑各类复杂多样的大数据应用场景，需要频繁地对这些繁杂、大规模、结构复杂的结构化与非结构化数据进行处理，因此实现对这些数据的高效分布式并行处理非常关键。常见的数据处理需求有 ETL 操作、大规模数据的实时排名、数据校验、异常分析、数据统计和数据迁移。这些大数据处理通常包含 3 种模式：离线批处理、流式实时处理和内存计算。基于 MapReduce 编程模型和 Hive、Pig 等大数据处理工具，可以有效地进行离线批处理操作：Storm 能提供高性能的流式实时处理支持；Spark 内存计算新技术能够满足小数据集上处理复杂迭代的数据处理场景下的计算需求。

4）医疗健康大数据分析和挖掘

医疗健康大数据分析和挖掘是大数据价值变现的关键环节。基于健康和医疗专业知识，构建复杂的算法和模型。针对特定的分析场景，利用可插拔的方式为每种分析场景实现相应的大数据分析服务引擎，如健康预测与疾病控制引擎、慢性病趋势分析引擎、统计学分析引擎、协同推荐引擎和可视化处理引擎。同时，结合综合管理、公共卫生、交换共享和其他卫生主题等业务需求，通过采集不同医疗机构业务系统数据，对各项医疗业务进行汇总统计、构成分析、对比分析、因素分析、增量函数分析等，并通过各种图表形象、直观地表达出来，能够有效地反映医疗管理机构或服务机构的整体运营、管理等情况。另外，还有利于管理层正确分析并做出有效决策，强化医卫管理，优化资源配置，控制不合理因素，最终实现基于大数据分析和挖掘技术的业务支撑、决策支持、科研辅助和管理支持等数据应用。

国家卫生健康委员会 2018 年 12 月 22 日发布《关于加快推进电子健康卡的普及应用工作的意见》，提出加快推进电子健康卡的普及和应用，推动居民电子健康档案在线查询和规范使用，到 2020 年，实现电子健康档案数据库与电子病历数据库互联对接，全方位记录、管理居民健康信息。同时，结合区域全民健康信息平台，实现现有公共卫生信息系统与居民电子健康档案的联通整合，健全高血压、糖尿病等老年慢性病及食源性疾病管理网络，推进母子健康手册信息化，加强对严重精神障碍患者发病报告的审核、数据分析、质量控制等信息管理。

1.4 国外研究现状及趋势

随着信息网络技术的飞速发展，全球进入大数据时代，大数据已经成为一个国家的重要战略资源。2012 年 3 月 29 日，美国政府颁布了《大数据研究和发展计划》，将大数据从商业行为上升到国家意志和国家战略，提出了三大战略目标：①开发大数据技术来收集、存储、保护、管理、分析、共享海量数据；②利用大数据技术加速科学与工程发展步伐，加强国家安全，实现教、学转型；③增加开发与使用大数据技术所需的人员数量。目前，在工业界，围绕大数据已经形成了非常庞大和复杂的大数据平台软件体系，其中比较核心的产品线包括 Hadoop 和 Spark 两大家族。Hadoop 是一个由 Apache 基金会开发的分布式系统基础架构。Hadoop 用户可以在不了解分布式底层细节的情况下，开发分布式程序。Hadoop 家族软件包括 HBase、Hive、Pig、Sqoop、Cassandra 等，其产品线更加适合离线计算。Spark 是 UC Berkeley AMPLab 开发的类 Hadoop MapReduce 的通用并行框架。与 Hadoop 的开源集群计算环境相似，但与 MapReduce 的不同在于 Job 中间输出结果可以保存在内存中，从而不再需要读/写 HDFS，因此 Spark 能够更好地适用于数据挖掘与机器学习等需要迭代的 MapReduce 算法。

与此同时，大数据的安全问题也日益凸显，成为大数据应用发展的一大瓶颈。针对严峻的大数据安全形势，美国、英国、法国、德国、日本、澳大利亚和欧盟等世界主要国家和地区采取了颁布大数据安全发展战略、制定大数据安全法规、成立大数据管理机构、加强大数据安全监管、研发大数据安全技术、培养大数据安全人才等一系列举措，为大数据安全发展提供强力支撑。目前，包括 Cloudera、Intel 在内的多个 Hadoop 发行版厂商，都在实行或制订安全方面的计划。

大数据作为一种新兴的网络应用模式，数据存储和使用与云计算紧密相关。大数据的安全考虑需要遵循传统信息系统的安全技术标准和云计算安全技术标准。相关的国际组织已经制定了一系列云计算安全技术标准，在大数据安全平台的设计中应当遵循及参考。相关标准主要包括：①ISO/IEC 27017《云计算服务信息安全管理指南》；②ISO/IEC 27018《云端系统的数据保护》；③NIST SP 800-125《完全虚拟化技术安全指南》；④NIST SP 800-144《公有云中的安全和隐私指南》；⑤NIST SP 800-145《云计算定义》；⑥NIST SP 800-146《云计算概要及建议》；⑦CSA《云安全指南》；⑧CSA《云控制矩阵》。

从基础技术角度看，大数据平台对数据的聚合增加了数据泄露的风险。Hadoop 是一个具有代表性的分布式系统架构，可以用来应对 PB 级甚至 ZB 级的海量数据存储。作为一个云化的平台，Hadoop 自身也存在着云计算面临的安全风险，平台需要实施基于身份验证的安全访问机制，Hadoop 派生的新数据集也同样面临着数据加密的问题。同样，大数据依托的基础技术——NoSQL（Not Only SQL，非关系型数据库）与当前广泛应用的 SQL（关系型数据库）不同，没有经过长期改进和完善，在维护数据安全方面也未设置严格的访问控制和隐私管理。并且由于大数据中数据来源和承载方式的多样性，NoSQL 内在安全机制缺乏保密性和完整性，导致将很难定位和保护敏感信息。

由于医疗健康大数据平台的极端重要性和广泛应用前景，欧美政府和研究机构对此展开了广泛的研究。医疗健康领域的大数据平台初期主要以电子健康网络、物联网的形式存

在，ACM 自 2006 年以来发起了专门的国际会议 Pervasive Health[2]，以推动这一领域的研究和交流。医疗健康大数据平台的网络安全与隐私保护面临特殊的挑战[3,4]。这些挑战主要包括以下两个方面。

（1）网络安全与隐私的极端重要性。与传统安全失败仅仅导致数据和经济损失不同，医疗健康大数据平台的安全失败还可能导致生命的逝去，大面积的安全失败甚至可能引发严重的社会问题，因此必须对医疗健康大数据平台的安全给予高度重视，不容闪失。

（2）网络中信息资源高度分散、动态和开放共享。出于提高医疗健康质量的考虑，医疗健康大数据平台提供多种手段支持信息资源的开放共享，这与医疗健康大数据平台安全与隐私保护需求之间形成了尖锐的冲突。海量数据的外包存储、持续动态更新和大规模用户环境更是加剧了这一冲突。

1.5 国内研究现状及趋势

国家政策对大数据的支持力度正在不断提升，大数据已上升至国家战略。自 2014 年 3 月“大数据”首次出现在《政府工作报告》中以来，国务院常务会议一年内 6 次提及大数据运用，李克强总理多次强调大数据运用的重要性。2015 年 7 月 1 日，国务院办公厅印发了《关于运用大数据加强对市场主体服务和监管的若干意见》。为加快推进云计算标准化工作，提升标准对构建云计算生态系统的整体支撑作用，工业和信息化部组织相关单位、标准化机构和标准化技术组织于 2015 年 10 月发布了《云计算综合标准化体系建设指南》。阿里巴巴公司的阿里云“飞天”云计算平台已在金融服务、政府管理、医疗健康、气象、电子商务等多个领域应用。曙光公司掌握了包括云基础设施、云管理平台、云安全、云存储、云服务等一系列云计算核心技术与产品，可以为用户提供“端到端”云计算自主可控的整体解决方案。在大数据分析处理设备方面，华为、浪潮、曙光等公司推出了大数据一体化解决方案。可见，我国大数据领域的自主软硬件产品发展势头良好，已经能满足一定范围的业务应用需求，为构建我国自主可控的大数据安全奠定了一定的基础。

我国的大数据发展很快，但是从数据应用的角度看，我国仍然处于大数据发展的初期阶段，绝大部分都是基于开源生态圈展开的应用开发，大数据分析处理技术不高，国内各行业、企业对大数据的安全防护与安全应用仍处于研究与摸索阶段。在国内大数据的分析领域，具有通用大数据分析能力的厂商基本被国外机构所垄断，表 1.1 列出了中国市场大数据分析能力排名前 20 的厂商。从表 1.1 中可以看出，具有较强大数据分析能力的企业大多为国外企业。国内企业只有阿里、百度、腾讯三家互联网公司较有实力，而互联网公司的数据分析能力是内向型的，即主要服务于自身业务。目前，国内行业、企业的大数据分析技术与平台还存在信息容易泄露、安全技术落后、防控能力不足等问题，亟待加强自主可控技术产品使用，同时从信息安全体系建设、大数据安全技术应用等方面加以解决。加快研发大数据中心安全防护技术，确保基于云服务的数据中心安全。针对大数据分散存储、分头管理、共享应用等特点，着力研发大数据管理系统、海量数据挖掘与预测分析、海量数据融合与集成等关键技术，加快构建自主可控信息系统，力求在高速组网、集群计算机编程、扩展云计算能力、广泛应用部署、数据安全和隐私保护等方面取得突破，全面提高大数据安全技术水平。针对医疗健康大数据这一重要领域的特殊需求，国内目前还没有相

关的机构能够提供专业的数据分析与安全服务。

表 1.1　中国市场大数据分析厂商

排　名	厂　　商	综合评分（10 分）	分项得分（10 分）			
			创新能力（35%）	服务能力（20%）	解决方案（30%）	市场影响力（15%）
1	IBM	9.1	10	8.5	8.5	9
2	Oracle	8.7	9	8	8.5	9
3	Google	8.6	9	8	8.5	8.5
4	Amazon	8.5	9	8	8.5	8
5	HP	8.3	8.5	8	8.5	8
6	SAP	8.2	9	8	7.5	8
7	Intel	8.1	9	8	7.5	7.5
8	Teradata	8.0	8.5	8	7.5	8
9	Microsoft	7.9	8	7.5	8	8
10	阿里	7.7	8.5	7	7	8
11	EMC	7.4	8	7	7.5	6
12	百度	7.0	8	5	7.5	6
13	Cloudera	7.4	7.5	8	7.5	6
14	雅虎	7.0	8	6.5	6	7
15	Splunk	7.1	8.5	7.5	6	5.5
16	腾讯	7.0	7	6	7	8
17	Dell	6.6	7	6.5	7	5
18	Opera Solutions	6.3	7	5.5	6.5	5
19	Mu Sigma	6.0	6.5	5	6	6
20	Fusion-io	6.1	7	5.5	5.5	6

全国信息安全标准化技术委员会积极推动产学研用单位参与大数据安全标准化工作，开展大数据安全标准的研制，为大数据产业安全有序发展提供标准化支撑。2017 年 4 月 8 日，全国信息安全标准化技术委员会 2017 年第一次工作组“会议周”在武汉召开，《大数据安全标准化白皮书》正式发布。《大数据安全标准化白皮书》由中国电子技术标准化研究院、清华大学、四川大学、阿里云计算有限公司等 25 家企事业单位共同编制，重点介绍了国内外的大数据安全法规政策、标准化现状，重点分析了大数据安全所面临的安全风险和挑战，给出了大数据安全标准化体系框架，规划了大数据安全标准工作重点，提出了开展大数据安全标准化工作的建议。2018 年 4 月 16 日，发布了《大数据安全标准化白皮书（2018 版）》。2019 年 5 月 13 日，国家标准新闻发布会在市场监管总局马甸办公区新闻发布厅召开，网络安全等级保护制度 2.0 标准正式发布，将于 2019 年 12 月 1 日开始实施。网络安全等级保护制度 2.0 标准在 1.0 标准的基础上，注重全方位主动防御、安全可信、动态感知和全面审计，实现了对传统信息系统、基础信息网络、云计算、大数据、物联网、移动互联和工业控制信息系统等保护对象的全覆盖。针对大数据的扩展要求包括管理流量与业

务流量分离、大数据授权与分类分级管理、大数据层面入侵防范与告警、大数据应用安全管理。

为了有效地整合多源异构的医疗健康资源，开展医疗健康大数据平台研究也是极其迫切的，我国高校和研究机构在相关领域开展了研究工作，取得了一些初步成果，如传感器技术[5]、实时数据处理形式化方法[6]、系统仿真[7]及可用于保护电子健康网络的信息安全技术[8]，包括认证码、数字签名、数据保密、秘密分享、安全多方计算和零知识证明等理论与技术。然而，现有的工作尚缺少对医疗健康大数据安全进行系统性、针对性的研究。

依托科技部的国家重点研发计划专项，我们项目承担团队的目标是研发出一整套生殖健康大数据平台数据挖掘计算与安全软件，产生一批符合健康大数据市场应用需要的数据融合、数据管理、趋势预测等模型工具。相关研究成果对提高我国机构在相关行业的竞争能力，填补相关领域的技术空白，具有非常重要的战略意义。