NIST Special Publication 1500-4

NIST 大数据通用框架草案

第四卷 安全与隐私

Final Version 1

NIST Big Data Public Working Group
Security and Privacy Reference Architectures

National Institute of
Standards and Technology
U.S. Department of Commerce

鸣　谢

特别感谢以下单位对翻译《NIST 大数据通用框架草案 第四卷 安全与隐私》的大力支持（排名不分先后）：

阿里云计算有限公司

中国电子技术标准化研究院

中国科学院信息工程研究所

国家信息中心

北京工业大学

华为技术有限公司

中国电子科技集团公司第三十研究所

国家信息技术安全研究中心

中国银联

目　录

执行摘要

《NIST 大数据互操作框架：卷 4，安全与隐私》由 NIST 大数据公开工作组（NIST Big Data Public Working Group，以下简称 NBD-PWG）安全与隐私小组编写，它确定了大数据特有的安全与隐私问题。

大数据的应用领域域包括医疗保健、药物发明、保险、金融、零售和其他多个来自私营部门和公共部门的领域，在这些应用领域域中的场景有健康变化、临床试验、兼并和收购、遥测设备、精准营销和国际反盗版。安全技术域包括身份识别、授权、审计、网络和设备安全、跨越信任边界的联盟。

显然，在理解和执行安全与隐私要求上，大数据的降临触发了需求模式的根本转变。这些变化显著且持续不断，特别是在扩展现有解决方案，以满足大数据的体量大、多样性、速度快和真实性，并重新定位技术基础架构的安全解决方案目标变化，例如，分布式计算系统和非关系型数据的存储。另外，多种多样的数据集变得更容易访问，并且越来越多的包含个人内容。一组新出现的问题必须得到解决，其中包括平衡隐私与实用性，对加密数据开展分析和治理，以及核查认证用户和匿名用户。

基于大数据的体量大、多样性、速度快、真实性等关键特征，小组从各个志愿者搜集用例，开发了一个达成共识的安全与隐私分类方法，将分类方法与 NIST 大数据参考架构（NIST Big Data Reference Architecture，以下简称 NBDRA）关联，并通过将用例映射到 NBDRA 的方式验证了 NBDRA。

NIST 大数据互操作性框架由七卷组成，每卷解决一个特定重点主题，随着 NBD-PWG 的工作产生。七卷具体如下：

- 卷 1:定义
- 卷 2:分类
- 卷 3:用例和通用要求
- 卷 4:安全与隐私
- 卷 5:架构白皮书调研
- 卷 6:参考架构

- 卷 7:标准路线图

NIST 的大数据互操作框架将会推出三个版本，分别对应 NBD-PWG 三个阶段的工作。这三个阶段实现如下目标：

- 阶段 1：确定高层次的大数据参考架构的关键组成部分，分别是技术、基础设施和中立供应商
- 阶段 2：定义 NBDRA 组件之间的通用接口
- 阶段 3：在通用接口上面构建大数据通用应用，对 NBDRA 进行验证

在阶段 2 工作小组的未来可能工作领域在本卷 1.5 节中突出显示，本卷记录的当前成果反映了在大数据高速发展领域形成的概念。

• 介绍

1.1 背景

在商业、学术和政府领导者之间已经就大数据的巨大潜力达成广泛共识，那就是大数据将引领创新，促进商业发展并推动进步。大数据是用来描述今天网络化、数字化、传感器承载和信息驱动世界的数据泛滥的通用术语。

海量数据资源的获得为解决以前遥不可及的问题带来了可能性，问题以下：

- 如何可靠检测、尽早干预潜在的流感？
- 是否能在材料合成之前，提前预测具有先进性能的新材料?
- 网络安全事件威胁中，如何逆转攻击者领先防御者的优势？

业界对大数据的能力超越了传统方法也达成广泛共识，数据量、速度和复杂性的增涨速率都超过了数据分析、管理、传输和用户应用领域科学和技术的进步。

尽管普遍认同了大数据固有的机会和目前的限制，对一些重要的根本性问题仍然缺乏共识，给潜在用户造成迷惑并阻碍进步。这些问题如下：

- 大数据解决方案要定义什么属性？
- 大数据与传统数据环境和相关应用程序有什么不同？
- 什么是大数据环境的基本特征？
- 如何在环境中集成目前部署的架构？
- 为了加快部署健壮的大数据解决方案，什么挑战需要核心系统、技术和标准

化来处理？

在此背景下，2012 年 3 月 29 日，白宫宣布了大数据研究和发展计划。该计划的目标包括帮助加快科学与工程的探索步伐，加强国家安全，以及转化教学和学习，提升从庞大而复杂的数据集合中提取知识和见解的能力。

联邦政府六大部门及其机构承诺，投入超过 2 亿美元涉及 80 多个项目，其目的是显著地提高访问、组织以及从海量数据得出结论所需要的工具和技术。该计划向行业、研究大学和非营利组织提出了挑战，要求他们与联邦政府一道创造大数据的机会。

在白宫计划和公众建议的推动下，美国国家标准与技术研究院（NIST）已接受挑战，促进行业专业人士的合作来进一步的安全和有效采用大数据。作为 NIST 云计算和大数据论坛 2013 年 1 月 15-17 日的一个结果，强烈鼓励 NIST 创建一个公开的工作组来开发大数据互操作性框架。论坛与会者指出，路线图应该定义大数据需求并区分优先次序，包括互操作性、可移植性、可重用性、可扩展性、数据使用、分析和技术设施。这样做，路线图将加快最安全和有效的大数据技术与科技的采纳。

2013 年 6 月 19 日，NIST 的大数据公开工作组（NBD-PWG）成立，得到国内行业，学术界和政府的广泛参与。NBD-PWG 的范围涉及到形成一个所有部门的利益共同体，包括工业界，学术界和政府，目标是就定义、分类、安全参考架构、安全与隐私达成共识，并将这些组成标准路线图。这项共识会创建一个厂商中立，技术和基础设施独立的框架，使大数据的利益相关者可以识别和使用最好的分析工具，在最适合的计算平台和集群上实现处理及可视化的需求，同时也让大数据服务提供商增值。

NIST 大数据互操作性框架由七卷组成，每卷解决一个特定重点主题，随着 NBD-PWG 的工作产生。七卷具体如下：

- 卷 1:定义
- 卷 2:分类
- 卷 3:用例和通用要求
- 卷 4:安全与隐私
- 卷 5:架构白皮书调研

- 卷 6:参考架构
- 卷 7:标准路线图

NIST 的大数据互操作框架将会推出三个版本，分别对应 NBD-PWG 三个阶段的工作。这三个阶段实现如下目标：

阶段 1：确定高层次的大数据参考架构的关键组成部分，分别是技术、基础设施和中立的供应商

阶段 2：定义 NBDRA 组件之间的通用接口

阶段 3：在通用接口上面构建大数据通用应用，对 NBDRA 进行验证

NBDRA 创建于阶段 1，并在阶段 2 和阶段 3 中进一步开发，它是一个高层次的概念模型，设计为一个便于公开讨论大数据固有的需求，结构和操作的服务工具，在《NIST 大数据互操作框架 卷 6: 参考架构》会进行详细讨论。工作小组阶段 2 未来可能的工作领域在本卷 1.5 节中强调，本卷记录的当前成果反映了在大数据高速发展领域内形成的概念。

1.2 安全与隐私小组的范围和目标

NBD-PWG 安全与隐私小组的关注点就是要形成一个工业，学术界和政府的利益共同体，目标是就参考架构达成共识，以处理所有利益相关者的安全与隐私问题。这包括了解什么标准是可用的或正在开发，并识别哪个关键组织正在研究这些标准。

小组的工作范围包括以下主题，其中有些将在本卷未来版本得到解决：

提供一个开始大数据特定的安全与隐私讨论的环境

从所有利益相关者收集与大数据处理、存储和服务相关的安全与隐私问题作为输入

分析安全与隐私要求面临的挑战列表并按照优先级排序，这些挑战可能会延迟或阻止大数据部署

- 开发一个安全与隐私参考架构来补充 NBDRA
- 制定大数据安全与隐私记录的工作草案
- 开发大数据安全与隐私分类
- 探索大数据安全与隐私分类与 NBDRA 之间的映射

- 探索用例和 NBDRA 之间的映射

虽然困扰大数据安全与隐私的问题有很多，小组的关注点是大数据安全与隐私的技术层面。

1.3 报告制定

NBD-PWG 安全与隐私小组通过探讨大数据安全与隐私的各个方面，制定这个文档。参与这项工作的主要步骤包括：

宣布 NBD-PWG 安全与隐私小组是向公众开放，以吸引和招揽政府、行业和学术界领域的专家和利益相关者

- 确定大数据安全与隐私的特定用例
- 开发一个详细的安全与隐私分类
- 展开 NBDRA 的安全性和保密性结构，并确定与 NBDRA 组件相关的特定主题
- 开始将识别的安全与隐私用例映射到 NBDRA

PWG 是本报告的编制贡献者，这是一个团体的成果，有一些主题覆盖到安全与隐私相关方面。尽量我们努力来契合主题，仍可会出现一些偏差，这些偏差在本文档的第二版来解决。

1.4 报告结构

- 第一节是介绍章节，本文其余部分的组织如下：
- 第二节讨论大数据特有的安全与隐私问题
- 第三节介绍安全与隐私相关的用例
- 第四节提供了一个安全与隐私初步的分类
- 第五节介绍了 NIST 大数据安全与隐私参考架构草案的细节，与整体 NBDRA 的关系
- 第六节将第三节的用例映射到 NBDRA
- 附录 A 讨论了特有的安全与隐私话题
- 附录 B 中包含云技术信息
- 附录 C 罗列了分类中出现的术语和定义

- 附录 D 包含本文档中使用的缩略语
- 附录 E 罗列了本文档使用的参考资料

1.5 未来工作

NBD-PWG 安全与隐私小组计划为后续版本（例如，第二版）进一步开发几个主题，这些主题包括以下内容：

仔细检查其他文献的现有模板：模板可以适用于大数据安全与隐私的结构，解决差距和搭建小组与其他工作组努力的桥梁。

- 进一步开发安全与隐私的分类
- 增强安全与隐私分类与 NBDRA 组件之间的联系
- 开发安全与隐私结构与 NBDRA 之间的连接
- 在本卷的范围内展开隐私讨论
- 探索大数据生态系统的治理、风险管理、数据所有权和价值估计，关注点是安全和隐私
- 将确定的安全和隐私用例映射到 NBDRA
- 从所处环境考虑 NBDRA 附录 B 的内容
- 探索 NBDRA 与隐私有关的可操作条款

我们保证未来会加入更多主题和方向，基于未来的输入以及对小组的作用，包括那些在公众评议期间收到的反馈。

2. 大数据安全与隐私

NBD-PWG 安全和隐私工作小组尝试通过识别大数据安全与隐私项目与传统实现的很多不同之处来研究大数据的安全与隐私问题。虽然这些概念不是一直都适用,但是下面的七个原则被认为是大数据具有代表性的一组较大的差异点:

- 大数据项目通常包括很多异构组件，并且从一开始就没有为这些异构组件设计单一的安全策略。
- 大部分安全和隐私的方法设计用于批处理或联机事务处理系统。大数据项目越来越多地涉及一个或多个流数据源,这些流数据源经常和静态数据结合，用来创建独特的安全和隐私场景。

- 使用原本不打算一起使用的多个大数据来源，可能会侵犯隐私或安全，甚至两者兼而有之。类似匿名化个人身份信息(PII)的技术在大数据时代可能不再满足需求了，因为这些技术并不是为大数据安全设计的。
- 大数据对传感器流数据的依赖越来越严重,比如物联网(物联网;如：智能医疗设备、智能城市、智能家居)中的传感器流，这种依赖会产生安全漏洞，这些隐含在流数据中的漏洞在积累成海量数据规模之前是比较容易管理的。
- 在大数据应用之前，某些类型数据被认为太大而无法分析,比如地理空间和视频成像，但在大数据时代，这些都成为数据源了。这些应用方式以前并没有想到，因此，可能没有相应的安全措施来保障这些数据的安全和隐私。
- 真实性的问题,溯源和管辖权问题在大数据方面被极大地放大了。很多组织、利益相关者、政府和越来越多的公民发现，关于他们自己的数据被包括在大数据分析中。
- 易失性是很重要的,因为大数据场景设想默认是永久性的。安全是一个快速发展的领域，可能拥有多个攻击向量和对策。数据被保留的时间可能超出被设计来保护它们的安全措施的生命周期。

2.1 概述

随着大数据的产生和利用，以及公共数据存储和可用度越来越大，安全和隐私措施变得越来越重要。

随着大数据的产生、访问和利用的不断发展，安全性与隐私措施变得越来越重要。数据生成将每两年翻一番，预计在 2020 年达到 40000 个艾字节。据估计,如果经过分析，2020 年将有超过三分之一的数据可以是有价值的。2010 年,只有不到三分之一的数据需要保护，而 2020 年将会有超过 40%的数据需要保护。

大数据的安全和隐私措施涉及到一种与传统系统不同的方法。大数据越来越多存储在公共云基础设施，这些云基础设施由各种硬件、操作系统和分析软件构成。传统的安全措施通常只解决小规模的系统，即持有静态防火墙和半封闭的数据网络。流媒体云技术的迅猛发展需要极其快速的应对能力，来应对安全问题和威胁。

大数据系统的展现层依赖操作者和角色，它们只是代表着安全与隐私的不同

的很小的一个面。大数据系统应适应新兴大数据应用蓝图，这些都体现在许多商业和开放源代码的访问控制框架中。这些安全的方法可能会持续一段时间,也可能会发展的新兴大数据应用蓝图。附录 C 考虑了大数据安全和隐私方面的操作者和角色。

大数据正在产生得越来越多，大数据被应用到不同的行业,如医疗、药物开发、金融、保险、包装消费品的销售业。这些不同行业的有效沟通需要安全和隐私条款的标准化。NBD-PWG 安全和隐私工作小组旨在鼓励参与者参与全球大数据的讨论，识别对于大数据来说复杂和困难的安全和隐私需求。

在数十年的学术研究和商业解决方案中，形成了大量的在安全和隐私领域的工作成果。虽然大部分工作不是概念上不同于大数据,但它可能在使用不同的假设时形成。本文的主要目标之一是，在缺乏典型大数据特征的情况下，了解大数据安全和隐私需求是如何出现的,以及这些需求是如何不同于传统的安全和隐私要求的。

下面的列表是一个典型的区别清单（这个清单可能并不详尽），它列出了大数据中的新东西和传统需求之间的区别，这些传统需求是在大系统安全性和隐私问题出现之前被提出的。

- 大数据可能来自不同的端点。相比传统的数据提供商和消费者，现在的大数据使用者包括更多的类型—数据所有者,比如移动用户和社交网络用户是现在大数据的主要参与者。一些收集不同数据流的物理设备也可能是大数据的使用者。这本身并不是新事物,但是人类和设备类型的结合在规模上是史无前例的。威胁因素和减轻威胁的潜在保护机制相结合是新的。

- 数据的聚合和分发必须限制在一个正式的可理解的框架环境中。数据的可用性，以及确保数据消费者现在和先前关于数据使用的透明性是大数据的重要方面。但现实情况可能是，大数据系统是可运行的，但独立于那些外部正式和可理解的框架,这些框架由一个设计师团队设计，并带着一组明确定义的目标。但在一些环境中,当这样的框架缺失或没被系统性组成时,就会需要有公共场所或“围墙花园”门户网站，并需要监察员一样的角色对存储的数据进行监察。这些系统组合以及那些不可预见的组合需要不断更新的大数据框架。

- 数据搜索和选择可能导致隐私和安全政策问题。当前我们缺乏对数据提

供者所应该具有的能力的系统性理解。除了通过排除可预见的数据库或限制查询来防止用户身份标识的重构外，一个高素质用户、高素质架构师和系统保护组合是需要的。如果大数据的一个关键特性是“有从任何规模数据的高级分析中获得有差别性见解的能力”,一位分析师这么描述，那么搜索和选择分析的方面将会加重安全性和隐私问题。

- 大数据需要隐私保护机制,如用来保护个人身份信息(PII)的安全机制。因为可能会在数据所有者、提供者和消费者之间有不同的、可能出乎意料的处理步骤，来自端点的数据的隐私和完整性在每个阶段应该受到保护。大数据在端到端信息保障方面的实践并没有不同于其他系统,但必须被设计在一个更大的规模下。

- 大数据正在超越传统信息对于信任、开放、和责任的定义。数据治理正成为大数据系统一个日益重要的内在设计考虑, 在大数据以前，数据治理是托付给通常受雇于大组织的某些静态角色。

- 大数据系统的信息保障和灾难恢复可能需要独特的紧急措施。因为它的极端的可伸缩性,大数据提出对信息保障(IA)的挑战，而灾难恢复(DR)的实践,没有以系统方式来解决。传统的备份方法对大数据来说可能是不切实际的。此外, 大数据副本的测试、验证和溯源保障可能做不到按时完成来满足时序要求，而这些要求在小系统中容易满足。

- 大数据能够创造增加价值的潜在目标。攻击者完成系统攻击所需的努力可被扩展，以满足机会价值。大数据系统给攻击者提供集中的、高价值目标。随着大数据变得无处不在,这样的目标越来越多——这本身就是新信息技术场景。

- 未经同意的大数据可追踪能力增加了去匿名化和转移 PII 的风险。安全和隐私可以通过无意或恶意的数据完整性攻击而被侵犯。管理大数据相关的数据完整性带来了额外挑战，关系到大数据特点，特别是 PII。虽然有技术可用于开发匿名化的方法,一些专家警告说,同样强大的方法能将大数据与个人信息看成相同的东西。例如,预期外数据集的可用性能够使重构用户身份定成为可能。即使技术能保护隐私,恰当的许可和使用也可能遵循路径数据，而不通过托管人。

- 开放的数据和大科学下的新兴风险。数据标识、元数据标签、聚合和分割，即广泛预期数据科学与开放的数据组——如果不妥善管理,真实性可能被降级,因为它们是派生的,而不是主要信息源。因为不恰当的数据解释而导致的同行

评议研究被撤回，可能像研究人员利用第三方大数据一样越来越普遍。

2.2 大数据的特点对安全和隐私的影响

多样性、容量、速度和可变性是大数据的关键特征,通常被称为大数据的四个V。在适当的地方,这些特征在 NBD-PWG 的安全和隐私工作小组内形成了讨论。尽管四个 V 为大数据提供一个有用的简略描述,能够用于大数据的公众论述,但仍有其它重要的大数据特点影响着安全和隐私,如真实性、有效性和易失性。下面讨论这些因素以及有关它们对大数据安全和隐私的影响。

2.2.1 多样性

多样性描述数据的组织结构——不管数据是结构化的，半结构化的还是非结构化的。重新将传统关系数据库安全定位到非关系数据库一直是一个挑战。传统数据库系统设计并没有考虑安全和隐私，这些功能通常通过插件来实现。传统的加密技术也阻碍了基于语义的数据结构。标准加密的目的是提供语义安全，这意味着任意值的加密结果与其它任意值加密是无法区分的。因此，一旦应用了加密，任何依赖于数据值属性特点的任何数据组织本身都将变得无效，而元数据的结构仍然可能是有效的,因为它可能是未加密的。

一个由大数据多样性引起的紧急现象已经得到相当多的重视，那就是通过将匿名化数据集与表面上无伤大雅的公众数据库进行关联，能够推断出人的身份。当前，很多能够解决隐私保护数据披露的形式化模型被提出,在实践中,敏感数据通过匿名化和聚合过程充分移除明显的唯一身份标识后被共享。但这只是一个解决单点问题的过程,并且通常只是基于经验证据，因此，导致了很多通过匿名化与公开可用数据结合推断出用户身份的事例发生。

2.2.2 容量

大数据量描述了有多少数据进入。按照大数据的说法,这一般范围从千兆字节到艾字节。因此,海量大数据需要存储在分层存储媒体中。但数据在层与层之间的移动导致需要对威胁模型进行重新分类，以及对新型技术进行调研。面向基于

网络的、分布式、自动分层存储系统的威胁模型主要包括以下主要场景:机密性、完整性、溯源、可用性、一致性、共谋攻击、回退攻击和记录保留干扰。拥有大量数据的另一个好处是可以通过分析数据来帮助检测安全违规事件。这是一个使用大数据技术增强安全性的大数据应用实例。这个文档解决大数据安全的两方面。

2.2.3 速度

速度描述数据处理的速度。数据通常是成批到达或是以流式连续方式传输。像某些其它非关系数据库，分布式编程框架在开发过程中是不注重安全和隐私的。非正常计算节点可能会泄露机密数据。由于系统的高度连接性和依赖度，针对基础设施的局部攻击可能会危害到系统的相当大的一部分。如果系统不强制执行地理上分散节点间的认证，恶意节点就可以添加到系统中，从而可以偷听到机密数据。

2.2.4 真实性

大数据的真实性和有效性包括几点：

溯源——一些人称之为真实性是为了与大数据的五 V 特征保持一致，对数据质量和安全保护以及维护隐私政策而言都很重要。大数据频繁在个人、团体以及组织的利益之间移动，并与地区，国家和国际边界都有关系。溯源旨在解决了解数据原始出处的问题，如通过元数据，虽然这个问题超出了元数据维护。人们尝试了各种方法，例如糖蛋白质组学研究，但仍然没有明确的指南存在。

一个普遍的理解认为，溯源数据其实就是元数据，这些元数据可以用来建立数据谱系和数据保管链，包括校准、错误、遗失数据（例如，时间戳，位置，设备序列号，交易号码，以及授权）。

一些专家将定义和维护元数据的挑战视为最首要的原则，甚至高于溯源，尽管这两者是相互关联的。

真实性（在一些圈子里也被称为溯源，虽然这两者是不一样的）也包含对信息收集方法的信息保障措施。例如，当使用传感器时，就需要对数据的追踪、校准、版本、采样和设备配置。

信息综合处理是一个整体概念，它将数据真实性和溯源绑定到数据治理原则

上，同时要确保数据质量。信息综合处理可以通过修正错误，提升原始数据质量，填补空白，建模，校准值，以及对数据采集进行排序。

有效性是指数据的准确性和正确性。传统意义上，这被称为数据质量。在大数据安全场景中，有效性指的是一群关于数据的假设，在这些假设前提下可以对这些数据实施分析操作。例如，连续和离散的测量有不同的属性。“性别”可以被编码为 1 =男性，2 =女性，但 1.5 并不意味着一半之间的男性和女性。

如果失去这样的约束条件，则一个分析工具可以给出不恰当的结论。有效性有很多类型，其约束更为复杂。基于定义，大数据可以在不同的数据集上执行聚集和信息收集，并且这些数据聚合和收集方法可能并不能由系统设计师预先想到。

关于大数据的几个“无效”应用例子已经被引用。点击欺诈，可以利用大数据的规模效应实现。但同样可以使用大数据技术检测出来，这个点击欺诈事件曾被报道过，因为它被认为是导致厂商花费 11.6 亿美元做无用广告的主要原因。一个软件主管列出了七种不同类型的在线广告欺诈，包括非人为产生的观感和点击，隐藏的广告，歪曲的来源，广告网站，恶意广告注入，和违反政策的内容，例如色情或隐私侵犯。这些可以利用的大数据的规模效应来实现，并且可能需要大数据解决方案来对其进行检测和打击。

除了最初的积极性，一些趋势预测应用使用社交媒体来预测流感的发病率出现了一些问题。Lazer 的一项研究指出，在对 108 周数据中的 100 周数据进行大数据分析的结果可能夸大了流感的流行度。当试图用“喜欢”和“追随”等的不精确含义和意图来刻画或预测消费者行为时，导致对社会媒体的不认真解读是可能的。这些例子表明，大数据的所谓“有效性”可能无害地在翻译、解释或故意损坏恶意意图中丧失。

2.2.5 易失性

数据的易失性，即随着时间变化如何管理数据的变化，将直接影响到数据溯源。大数据是可以转换的，部分原因是系统可能产生无限的持久数据，这些数据比收集它们的工具持久，比设计数据采集、处理、汇总和存储软件的工程师活得还长，甚至比最初确定项目的数据消费者的赞助商活得还长。

角色本质上具有时间依赖性。安全和隐私要求可以相应地转移。随着负数据

治理责任的组织合并或消失，数据治理责任也可以转移。

虽然人们开展了对时序数据的管理研究（例如，在 e-Science 的卫星仪器的数据），但少有几个标准超出了简单时间戳，甚至很少有可作为指南使用的通用实践。为了长生命周期的大数据的安全和隐私，数据的时序性应该加以考虑。

2.3 与云的关系

许多大数据系统将使用云架构设计。在大数据云生态系统企业架构行业中，任何战略实现适当的访问控制和安全风险管理必须解决包括云计算引发的特定安全需求与云计算特性的相关复杂度，包含不限于以下：

- 宽泛网络接入
- 降低消费者的可视性和控制力度
- 提供者和消费者之间动态系统边界和混合角色以及责任。
- 多租户
- 数据残余
- 测量服务

数据规模按数量级增长（按需）、动态（弹性和成本优化），和复杂性（自动化和虚拟化）

这些云计算特点，相比传统信息技术解决方案，往往会为组织带来不同的安全风险，改变了组织的安全态势。

为了将数据迁移到云中并确保数据安全性，组织需要提前确定所有和云相关的、基于风险调整的安全控制或组件。在某些情况下，通过合同手段和服务级别协议（SLA）从云服务提供商获得请求可能是必须的，所有要求安全组件和控制得到了全面、准确地实施。进一步讨论云生态系统内的内部安全考虑，可以附录 B 中查询。未来这个文件的版本背景，将在 NBDRA 的附录 B 的内容中研究。

3 安全和隐私使用案例

大数据在科学和工程上有相当大的挑战。这方面的许多内容在 NIST 的《大数据互操作框架：卷 3 使用案例和通用需求》（NIST Big Data Interoperability Framework: Volume 3, Use Cases and General requirements）中进行了描述。但是，

这些使用案例主要针对科学和工程应用，其中安全和隐私是次要考虑的—如果后者对系统架构根本没有影响的话。

因此，在这份文档的准备过程中开发了一组不同的使用案例来专门探索安全和隐私的问题。其中部分使用案例描绘了不活跃的或传统的应用，但是因为它们展示了典型的安全/隐私设计模式因此也被选入。

针对安全和隐私的使用案例在下面的子章节进行介绍。包含的使用案例被进行如下分组：零售/市场，医疗保健，网络安全，政府，工业，航空和运输。但是，这个分组并不代表受大数据安全和隐私影响的整个产业范围。

这些使用案例是在参考架构还不成熟时选择的。它们被收集来以确定典型的安全和隐私场景，这些场景被认为是适合特别分类到大数据的。我们尝试把这些使用案例与 NBDRA 进行映射。在版本 2 中，将研究额外的使用案例到 NBDRA 和分类法的映射。这份文档的一部分是并行进行开发的，在版本 2 中将加强这部分内容之间的关联。

3.1 零售/市场

3.1.1 消费者数字媒体使用

场景描述：在决定一个交易前，有智能设备帮助的消费者已经变得对价格、方便性和接入非常在意。内容所有者将供消费者使用的数据特许给门户网站进行展示，例如 Netflix、iTunes 等。

来自于不同零售商、存储位置和/或交付选择、众包率（crowd-sourced rate）的价格比较已经成为选择的常见因素。为了竞争，零售商紧盯消费者的位置、兴趣和消费模式以动态地创建市场战略并销售给消费者并无预期想要的产品。

安全和隐私现状：个人数据通过几种方式被收集，包括智能手机 GPS（全球定位系统）或位置、浏览器使用、社交媒体和智能设备上的应用（apps）。

- 隐私

上面描述的大部分数据收集方法仅提供较弱的隐私控制。此外，消费者的无意识和疏忽允许第三方合法地捕获信息。在这种场景下，消费者仅能受限于期望没有隐私。

● 安全

控制不一致和/或未适当建立以达成以下：

- 隔离，容器化（containerization），和数据加密
- 监控和威胁检测
- 用于数据供应（data feed）的用户和设备识别
- 与其他数据源的接口连接
- 用户匿名化：一些数据收集和整合使用匿名化技术，个人用户能通过使用 其他公共大数据池被再次识别。
- 原件的数字版权管理（DRM）技术并不能满足对数据可预见的使用需求。“DRM 指的是一大类针对在各种不同设备上限制使用和拷贝数字内容的接入控制技术。”DRM 能被攻破，转为预期之外的目的，或者不能在大数据特征环境下运行—特别是高速和大量特征。

研究现状：对使能保护个人数据（无论匿名的或不匿名的）的隐私和安全控制的研究还很有限。

3.1.2 Nielsen 家庭调查：Apollo 项目

场景描述：Nielsen 家庭调查（Nielsen Homescan）是 Nielsen 收集家庭级零售交易的一个子公司。Apollo 项目是设计来在 Nielsen 成员中联合起来以更好地公告交易行为内容曝光的项目。Apollo 项目并没有超出一个有限的试验的范畴，但是反映了大数据意图。该项目中包含的来自于 Nielsen、Arbitron 或者各种合约商的描述是一个尽全力通用的描述，而并不是官方的观点。其提供的信息应该被视作说明性，而不是一个历史记录。

一个常规的零售交易有一张结账收据，包含了购买的所有 SKUs（库存量单位）、时间、日期、商店位置等。Nielsen 家庭调查使用一个随机化的国家取样统计来收集采购交易数据。自 2005 年起，该数据仓库已经是一个数 TB 的数据集。该仓库的建立基于结构化技术，以适应多 TB 的需求。Homescan 自己维护数据，并共享给消费者，这些消费者通过一个私有门户网站使用一个柱状数据库被允许部分接入。可以使用第三方软件进行额外的分析。其他消费者仅能接收到包含聚合数据的报告，更多粒度的需要付费购买。

安全和隐私现状（2005-2006）：

● 隐私：有大量的 PII 数据。调查参与者放弃市场细分数据、人口统计资料和其它信息的所有权作为交换从而得到补偿。

● 安全：传统接入安全涉及到使用数据库引擎在域级（field level）实现的组策略、组件级应用安全和物理接入控制。

● 适当的审计方法，但仅适用于内部员工。选择性退出数据清理是最低限度。

3.1.3 网站流量分析

场景描述：访问级网站服务器的记录是大粒度和数量巨大的。为了有用，日志数据必须与其它（潜在的大数据）数据源关联，包括页面内容（按钮、文本、导航事件），市场级事件例如活动、媒体分类等。对使用复杂的事件处理（CEP）实时对业务流进行分析的计划----如果还未部署----有讨论。一个不小的问题是区分业务流类型，包括内部用户社区，对其的收集策略和安全是不同的。

安全和隐私现状：

● 非欧盟：双向确认（Opt-in）默认依赖于获得访问者同意以进行跟踪、记录 IP 地址用于某些分析，直到在城市街区级确定访问者。

● 媒体接入控制（MAC）地址跟踪，使得分析师能确定 IP 设备，该 IP 设备是 PII 的一种形式。

● 一些公司允许按需清除数据，但是大部分不可能擦洗以前收集的网页服务器业务流。

● 欧盟对收集这些被当做 PII 的数据有更严格的规定。这样的网站流量仅在聚合甚至为了在欧盟内基于美国跨国操作时被擦洗（匿名化）或者上报。

3.2 医疗保健

3.2.1 健康信息交互

场景描述：健康信息交互（Health Information Exchanges，HIEs）促进医疗保健信息的共享，包括电子健康记录（EHRs），以使该信息能被相关覆盖实体

接入，当然前提是获得病人的同意。

HIEs 趋于联邦的，其中各自覆盖的实体保持对数据的保管。这对许多场景造成了问题，例如紧急事件，由于各种各样的原因包括技术（例如交互性）、商业和安全考虑。

通过强壮的密码学和秘钥管理来满足 HIPAA（Health Insurance Portability and Accountability Act，健康保险可移植性和问责制法案）对 PHI（protected health information，受保护的健康信息）的要求----理想情况下不需要云业务运营商签署 BAA（business associate agreement，商业关联协议）---- HIEs 的云可实施性可以提供数个好处，包括病患安全、降低的医疗保健花费、紧急事件例如击碎玻璃（Break-the-Glass）和 CDC（Centers for Disease Control and Prevention，疾病控制和预防中心）场景时的规定访问。

以下是 NBD PWG 提出的几个初步的场景：

Break-the-Glass：可能某些情况下由于医疗形势或者无法联系到监护人，病人不能提供同意书，但是一个授权方需要立即访问相关的病患记录。密码学增强秘钥生命周期管理能提供足够的可视性和不可抵赖性级别，这能在事后对违规行为进行跟踪。

知情同意：当 EHRs 在覆盖的实体和商业合作人之间转移时，病人能传达他们的同意权，以及制定他们 HER 中的哪些组件是可以转移的（例如，他们的牙科医生不需要看到他们的精神病学记录）是合适和必要的。通过密码学技术，可以对将传输的文本指定细粒度的密文策略。关于同意书相关的标准，参见 NIST 800-53（附录 J，IP-1 章节）、美国 DHS 健康 IT 策略委员会（US DHS Health IT Policy Committee）隐私和安全工作组（Privacy and Security Workgroup），以及针对数据接入同意、同意指令的 7 级健康水平（Health Level Seven，HL7）国际版本 3。

流行病援助：公共健康实体，例如需要该信息以促进公共安全的 CDC 和可能其它非政府组织需要受控地接入该信息，有可能某些情况下无法接入服务和基础设施。有合适密码控制的云 HIE 能够通过授权和审计以能促进场景需求的方式向授权实体发布关键信息。

项目现状和/或提出的安全和隐私：

- 安全
- 轻量级但是安全的云下（off-cloud）加密：需要有能力执行轻量级但是安全的 EHR 云下加密，该 HER 能存在于从浏览器到企业服务器的任意容器内，并利用强壮的对称密码学。
- 同态加密
  - 应用密码学：彻底减少，现实的威胁模型和有效的技术。
- 隐私
  - 差别隐私：能确保抵御 PII 不恰当泄露的技术
- HIPAA

3.2.2 遗传隐私

场景描述：一个策略制定者、倡导组织、个人、学术中心和产业界的联盟已经形成了一个倡议：数据免费！(Free the Data)，以填补由于缺乏可用的 BRCA1 和 BRCA2 基因信息而造成的公共信息空白。该联盟也计划在开放、可搜索的数据库中扩展提供其它类型的基因信息，包括国家生物技术信息数据库中心 ClinVar。该项目的主要创办人包括基因联盟（Genetic Alliance)、加利福利亚旧金山大学（the University of California San Francisco)，InVitae 公司，以及病患拥护者。

该倡议邀请个人在一个公共数据库中以适当的隐私设定共享他们自己的基因变异信息，这样他们的家庭、朋友和临床医生能更好地了解突变意味着什么。协同工作以创建该资源意味着朝更好地了解疾病、更高质量的病患护理和提升的人类健康的方向而努力。

安全和隐私现状：

- 安全
  - 基于 SSL（Secure Sockets Layer，安全套接层）的认证和接入控制。低证明级别的基本的用户注册
  - 对用户死亡后的数据所有权和保管的关注
  - 站点管理人员可能能接入数据----推荐强壮的加密和秘钥托管
- 隐私

- 对遗传信息进行透明的、记录的、策略监管的控制
- 全生命周期的数据所有权和保管权控制

3.2.3 制药临床试验数据共享

场景描述：公司例行公布他们的临床研究，与学术研究者合作，在公共网站上共享临床试验信息，临床上分为 3 个阶段：病患征募的时间，新药获批后，当调查研究项目终止。即使对研究人员和政府，接入临床试验数据也是受限的，且没有统一的标准存在。

美国制药研究和厂商（The Pharmaceutical Research and Manufacturers of America,即 PhRMA)代表了国家领先的生物制药学研究人员和生物技术公司。2013 年 7 月，PhRMA 加入了欧盟制药产业和联盟（European Federation of Pharmaceutical Industries and Associations，即 EFPIA）采用联合的原则以负责临床试验数据共享。通过该协定，公司将在资源的基础上应用这些原则作为公共基线，PhRMA 鼓励所有的医学研究人员，包括那些在学术界和政府的人员，通过采用时实施以下承诺，来提升医疗和科学进步：

- 增强研究人员间的数据共享
- 增强对临床研究信息的公共接入
- 共享试验信息的证明流程
- 重申对发布临床试验结果恪守承诺

安全和隐私现状：

PhRMA 没有直接提出安全和隐私，但是这些问题被 PhRMA 和该提案的评审者确定了。

- 安全：
  - 超过试验处置期后纵向的保管不清楚，特别是公司合并或解散后
  - 数据共享的标准不清楚
- 需要使用审计和安全
  - 出版限制：需要额外的安全来保护出版商的权利；例如，Elsevier 或者 Wiley
  - 隐私

- 病患级别的数据披露----选择性的针对每一个公司

- PhRMA 提到匿名化（身份再识别），但是提到了小样本的问题
- 研究级的数据披露----选择性的针对每一个公司

3.3 网络安全

3.3.1 网络保护

场景描述：网络保护包括各种数据采集和监控。现有的网络安全产品的监测高容量数据集，如事件日志，跨越数千工作站和服务器，但他们还没有能够扩展到大数据。改进的安全软件将会包括物理数据的相互关联（如设备以及建筑入口/出口使用的访问卡）和可能会更紧密地与应用集成，这将产生以前未有类型或大小的日志和审计记录。大数据分析系统将需要处理和分析这些数据，以提供有意义的结果。这些系统还可以是多租户，迎合更多不同种类的公司。

这个场景突出了两个子场景：

- 大数据安全
- 安全大数据

当前的安全和隐私：

- 安全在这方面是成熟的;隐私概念也少一些。
- 传统政策性安全盛行，虽然时间维度和监测条款修改事件往往是非标准或未经审计的
- 网络安全的应用程序运行在较高级别的安全性，因此需要单独的审计和安全措施
- 非跨行业标准为操作系统的收集数据方法以外的聚集数据而存在
- 实现大数据的网络安全应该包括数据管理，加密/密钥管理和租户数据的隔离/集装箱化。
- 在为大数据的网络安全设计备份和灾难恢复时应考虑波动性。日志的使用寿命可以延伸超过创造了他们的器件的寿命。
- 隐私：

- 企业授权数据发布给国家/国家组织

- PII 数据的保护

目前厂商都在采用大数据分析大批量规模的日志相关性和事故响应，如安全信息和事件管理（SIEM）。

3.4 政府

3.4.1 军事：无人驾驶车辆传感器数据

情景描述：无人驾驶车辆（或无人机）及其机载传感器（例如，视频流）可以产生应存储在非标准格式的 PB 级数据。这些流往往不被处理实时，但美国国防部门（DOD）在购买技术，使这个成为可能。因为相关性是关键，全球定位系统，时间和其它数据流必须被共同收集。布拉德利曼宁泄露的情况是安全漏洞的实例。

当前的安全和隐私：

- 适用机构职责的单独法规
- 对于国内监控：美国联邦调查局（FBI）
- 对于海外的监控：多个机构，包括美国中央情报局（CIA）和国防部各机构
- 并非所有的用途将是军工;例如，美国国家海洋和大气管理局
- 军事安全分类是比较复杂的和在需要知道的基础上确定
- 信息保证措施严格遵循，不像在一些商业环境

目前的研究：

- 其中有审计手段的用途被审计，软件不被安装/部署，直到使用进行审计“认证”，并且开发周期具有相当的监督;例如，美国军队的军规 25-219
- 内部威胁（例如，爱德华·斯诺登，布拉德利·曼宁和间谍）正在处理中，如美国国防高级研究计划局（DARPA）的网络入侵者威胁（CINDER）项目。本研究和一些由行业提供的未备基金的建议或许被考虑。

3.4.2 教育：共同核心的学生成绩汇报

情景描述：145 个国家已决定为 K-12 学生成绩的测量统一标准。结果被用

于多种用途，并且该程序是初期的，但它会得到纵向大数据的状态。设想的数据集包括跨学生的整个学校的历史和整个学校和国家，以及考虑到测试激励的变化的学生级表现。

当前的安全和隐私：

- 数据由私人公司进行评分和转发给国家机构被收集。课堂上，学校和区标识符留在得分结果上。学生 PII 的状态是未知的;然而，众所周知，教师接收课堂教学级表现的反馈。学生/家长获得测试结果的程度尚不清楚。
- 国家不愿参加 InBloom initiative20 说明了围绕教育大数据的隐私有关的纠纷。
- 据一些报道，家长可以选择让学生不参加国家的测试，所以选择退出的记录也必须被收集并用于清除没有入选的学生记录。

目前的研究：

- 纵向表现数据对项目评估者来说将有价值，如果数据可扩大规模
- 数据驱动学习 22 将涉及获取学生的表现数据，可能比在测试时更频繁，并且在更高的粒度，因此需要更多的数据。其中一个企业例子是斯维塔斯学习的 23 学生的决策预测分析。

3.5 产业：航空

3.5.1 传感器数据存储和分析

情景描述：大多数商业航空公司都配备了上百个传感器不断地在一次飞行中捕获引擎和/或飞机的健康信息。在一个单一的飞行中，传感器可收集多个千兆字节的数据和传输这些数据流到大数据分析系统。几个企业管理着这些大数据分析系统，如零件/发动机制造商，航空公司，以及飞机制造商和数据可能会在这些公司间共享。汇总数据是分析维修调度，飞行路径等。航空公司等一个共同的要求，是为了安全和从竞争对手隔离其数据，即使当数据被传输到同一分析系统。

航空公司也更希望控制如何，何时以及和谁的数据是共享的，即使对于分析的目的。大多数这些分析系统现在被转移到基础设施云服务提供商。

当前和建议的安全和隐私：

- 静态加密：大数据系统应该加密存储在基础架构层，使云存储管理员无法访问数据
- 密钥管理：密钥管理应架构化，使终端用户（例如，飞机）对发布解密数据的密钥有唯一/共享控制
- 动态加密：大数据系统应验证在云提供商中传输的数据也被加密
- 加密使用：大数据系统当在内存中处理数据（尤其是在云提供商中）将希望完整的混淆/加密
- 传感器的标定和唯一标识（如设备身份管理）

研究人员目前正在研究以下安全增强功能：

- 在云提供商中映射虚拟化基础架构层
- Quorum 机制为基础的加密
- 多方计算能力
- 设备公钥基础设施（PKI）

3.6 运输

3.6.1 货运

以下的例子概述了航运业（如联邦快递，UPS，DHL），是如何经常使用大数据的。大数据被用在物品标识，运输和处理在供应链中的物品。对发送者，接收者，以及所有那些在之间有必要知道在运输途中物品的位置和到达时间的各方来说，该物品的标识是很重要的。目前，运输物品的状态不是通过整个信息链进行中继的。这将由传感器的信息，GPS 坐标，和一个基于 ISO 技术委员会 ISO JTC1 SC31 WG2 内新的国际标准化组织（ISO）29161 标准发展而来的独特的识别模式等提供。当卡车到达贮库或当物品被传送至接收方时数据被近实时地更新。中间条件在当前不被知道，位置不是实时更新，物品在一个仓库丢了或在装运里表现了家园安全的潜在问题。该记录被保持在一个档案里，并可以为系统确定的天数里进行访问。

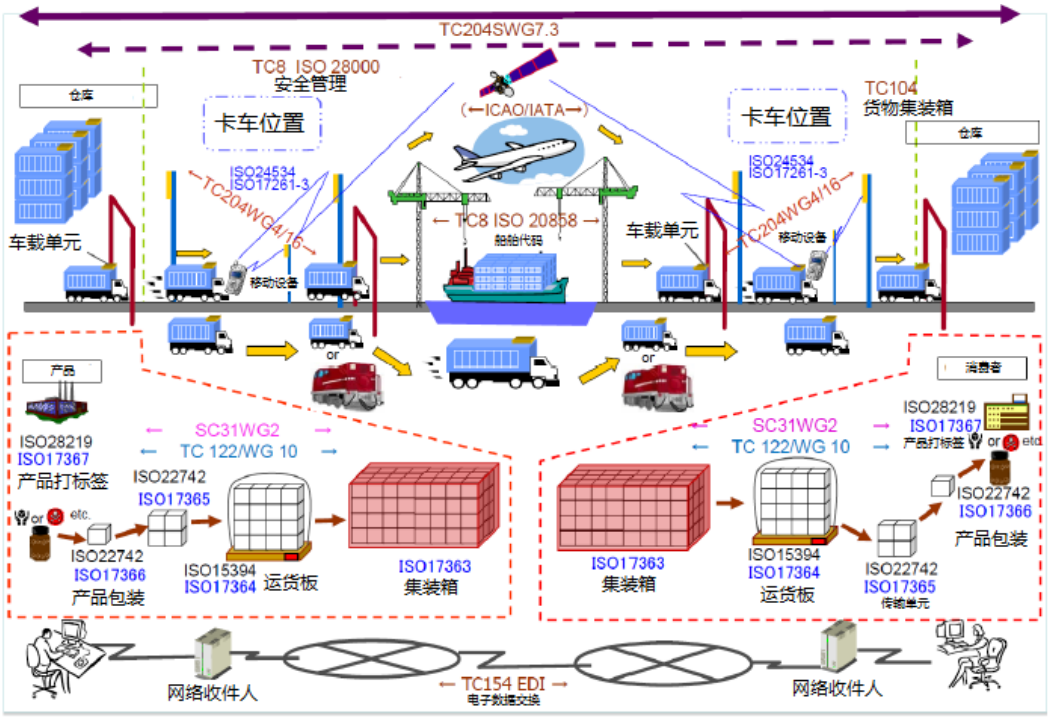

Figure 1: 货物运输场景

图 1：货物运输场景

4. 安全和隐私的分类

存在一个候选主题集被应用于开发这些安全和隐私的分类，这个候选集来源于大数据云安全联盟工作组的一篇文章:“大数据安全与隐私面临的前十项挑战”，候选主题和本章节中相关的资料详见参考附录 A。

一个大数据安全和隐私的分类应该涵盖现有的目标，有用的分类。尽管存在许多有关安全和隐私的概念定义，但是这里涉及的分类的目的是强调并且细化精确到大数据的新的原则。

下面各章节分别对各种安全和隐私类别进行了概述，同时并给出每种分类元素里所包含的主题列表。这些列表是分组初步讨论的结果并且可能会在版本 2 中进一步深入讨论。

4.1 安全和隐私的概念主题

如图表 2 所示，安全和隐私的概念分类包括四个主要方面：数据保密性；数据出处；系统健康和公共政策、社会和跨组织主题。

前三个方面大致与传统的保密性，完整性和可用性分类相对应，并向着并行大数据的重新定位思考。

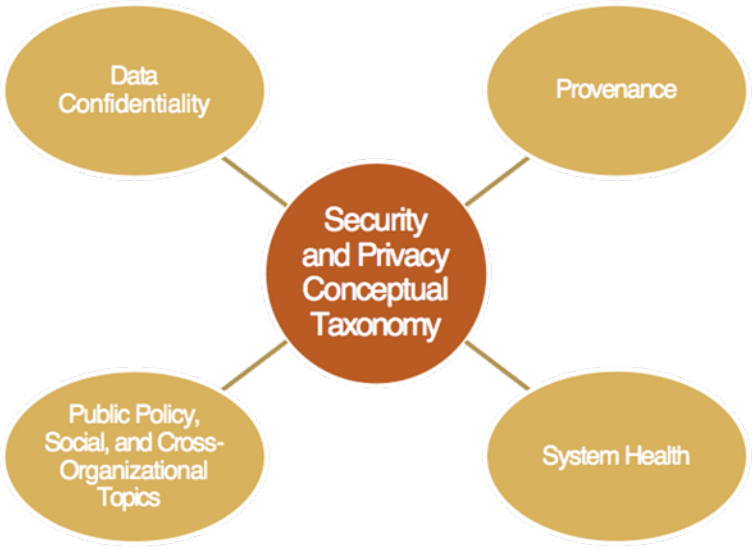

图 2 安全和隐私的概念分类

4.1.1 数据保密性

- 传输数据保密性：例如，通过使用传输层安全协议执行保密性
- 静态数据保密性
- 基于凭证访问数据的策略
  - 系统：通过使用诸如访问控制列表和虚拟机等系统构造来执行策略
  - 密码执行：通过使用诸如 PKI 和基于身份/属性的加密机制来执行策略
- 对加密数据进行计算
- 搜索报告：支持对加密数据搜索报告的加密协议--不能从搜索标准推断出来的任何纯文本信息是保证被隐藏的
- 同态加密：支持对加密的基本纯文字操作的加密协议--关于纯文本的任何信息是保证被隐藏的
- 安全数据聚合：聚合数据而不影响隐私
  - 数据匿名化
  - 识别保护隐私的记录
  - 密钥管理

● 就像 Chandramouli 和 Iorga 提及的，加密密钥的云安全，作为安全和隐私的一个基本组成模块，呈现出“额外的复杂性”，从而能够为大数据而被改写：（1）由于更多的云消费者-供应商的关系造成的更大的多样性，（2）更大的需求和基础设施的多样性，密钥管理系统和保护资源都将定位于此。

● 大数据系统不纯粹是云系统，但正如在本文档的其他地方所提到的，两者是密切相关的。一种可能性是将 Chandramouli 和 Iorga 为云服务模型开发的密钥管理框架重定向为 NIST 大数据参考框架安全和隐私结构。云模型将与 NIST 大数据参考框架和云安全的概念提出的结构相对应。NIST 提供云计算概念的定义，包括设施即服务（IaaS），平台即服务（PaaS），和软件即服务（SaaS）的云服务模型。

● 大数据密钥管理系统（KMS）的挑战，反映出大数据特征（即体积，速度，多样，变化）所施加的要求。例如，空闲密钥的生成以及遗留下来又难以更新的工作流，而且通常又是比较挑剔的，数据库密钥生成对于需要快速部署和缩放使用大量资源的大数据系统来说是不够的。一个大数据密钥管理系统的生命周期可能将比大数据系统架构师的设计工作时期还长。对大数据密钥管理的位置、规模、所有权、保管、出处和审计的设计是安全和隐私结构的一个方面。

4.1.2 溯源（可追溯性 / 完整性）

- 终点输入验证：验证输入数据是否是来自一个已认证源的机制，例如数字签名
  - 句法：句法层面的验证
- 语义：语义验证是一个重要的问题。一般而言，语义验证是验证典型的业务规则，如到期日期。有意或无意违反语义规则会锁住一个应用程序。当使用不能够识别特定变体的数据转换器时也会是锁住应用的情况发生。协议和数据格式通过供应商使用可能会被更改，例如，一个预留的数据空间，将允许他们的产品有区别于其他产品的能力。这个问题也可能出现在不同版本的消费者设备系统，包括移动设备。一个消息的语义和被传输的数据应该被验证已确认，在最低限度地符合任何应用标准。对于来自于一个传感器或数据提供商的数据已经通过使用验证器或数据检查验证是有效的，数字签名的使用将是重要的保证，因此也是有

效的。这种能力是重要的，特别是如果数据被转化或参与数据的管理。如果数据未能满足这些要求，则可能会被丢弃，并且如果数据继续出现问题，则源提交数据的能力可能会限制。这些错误类型将被记录并被阻止传播给消费者。

- 在大数据系统中，数字签名是非常重要的
- 通信完整性：数据在传输，执行过程的完整性，例如通过使用传输层协议（TLS）
  - 数据认证计算：确保在关键数据片段上的计算确实是预期的计算
  - 可信平台：通过可信平台的应用实施，如可信平台模块（TPMs）
- 加密实施：通过加密机制的使用实现
- 粒度审计：在高粒度下进行审计
- 价值资产的控制
  - 生命周期管理
- 保留与处置
  - 数字版权管理

4.1.3 系统健康

- 拒绝服务的安全（DoS）
- 主动抵抗 DOS 的加密协议的构造
- 安全大数据
- 安全情报分析
- 数据驱动滥用检测
- 大数据日志分析，网络物理事件，智能代理
  - 安全漏洞事件检测
  - 法医学
  - 支持恢复力的大数据

4.1.4 公共政策，社会和跨组织主题

下面的主题集是从计算机协会（ACM）团队得到的。每一个主题都有大数据安全性和隐私性，这可能会影响结构覆盖如何实现一个特定的大数据目标。例如，

一个医疗设备项目可能需要解决人的安全风险，而一个银行项目可能会涉及到适用于大数据跨越边界的不同准则。研发这些大数据的概念的深入工作有望能够在各团队中实现。

- 滥用和设计计算机的犯罪
- 电脑相关的公共/私人健康系统
- 行为准则（在数据科学内，但也跨专业）
- 人类安全
- 知识产权及相关信息管理
- 管制
- 跨境数据流
- 使用/滥用权力
- 残疾人士的辅助技术（例如，为人群中某些小群体增加或设置不同的安全/隐私措施可能是需要的）
- 就业（例如，适用于工作场所法的规定可以控制被员工生产或者管理的大数据的恰当使用）
- 电子商务的社交方面
- 法律：审查、税收、合同执行、执法取证

4.2 安全和隐私的运行层面术语

当前关于保障大数据系统安全的实践呈现多样性，通过运用不属于统一概念框架的各种不同方法来实现。图 3 所示的运行分类的各要素代表了实践性方法论的组合，这些要素被归类于运行层面主要是由于它们针对了大数据系统运行过程中的特定缺陷或风险管理面临的挑战。此时在标准的发展过程中,这些方法论还未合并成一个有凝聚力的安全结构的一部分，它们是具有潜在价值的清单模式(checklist-style)要素，可以满足特定安全或隐私需求。未来的工作必须将这些方法论与危机管理指导意见（如：NIST 关于将风险管理框架运用到联邦信息系统和 COBIT 风险 IT 框架的特别出版指南）更好地整合起来。

在已提出的运行层面术语中，概念分类法的广泛考虑表现出循环使用的特征。例如，通信的保密性可以运用于静态数据管理和存取管理，同时还作为安全元数

据模型的一部分。

运行分类在关注大数据的特定议题时，会与小数据分类相重叠。

4.2.1 设备和应用注册

- 设备，用户，资产，服务以及应用程序注册：包括机对机(M2M)的设备注册和 IoT 网络，DRM 管理(DRM-managed)资产，服务，应用程序和以及角色。
- 安全元数据模型
  - 元数据模型与一个安全系统的所有元素都有关联，同时也所有的基础存储库保持联系。大数据通常需要这种附加的复杂性，因为它具有更长的生命周期，更广泛的用户群体或其他方面的原因。
  - 一个大数据模型必须处理数据测速方面以及有关数据和安全模型组件生命周期的时间方面。
  - 政策执行
  - 环境构建
  - 部署策略实施
- 治理模型
  - 细微政策的审核稽查
- 特定角色的行为分析

4.2.2 身份和访问管理

- 虚拟化层的身份（如：云主机，平台即服务[PaaS]）
  - 可信赖平台
- 应用程序层身份
  - 终端用户层身份管理
  - 功能作用
  - 身份提供者 (IdP)
  - 身份提供者由安全断言标记语言(SAML)定义。在一个有着数据提供者、协调器、资源提供者、框架提供商和数据消费者的大数据生态系统中，像安全断言标记语言(SAML)，安全令牌服务(STS)或者可扩展访问控制标记语言(XACML)这

样的方案对于分解安全分类中的各要素而言是有益而不排斥的方法。

- 大数据有大量身份提供者，一个身份供应者通过发布身份和角色来访问资源提供商的数据，在 SAML 框架中，信任是通过注册阶段的 SAML/web 服务机制来共享的。

- 在大数据中，由于数据密度，用户需要“漫游”数据（然而在传统的虚拟专用网络(VPN)模式下的情形中，用户漫游可以跨过信任边界）。因此，传统的认证/授权模型(authn/authz)需要得到进一步扩展，因为依赖方保管着其他人的数据，他们不再完全被信任。数据汇总可能来源于多个不同的资源提供者。
- 一个途径是通过拓展 SAML 中基于声明的方法来增加安全与隐私性保障。
- 附加可扩展访问控制标记语言概念
- 可扩展访问控制标记语言(XACML)中介绍了对大数据安全有帮助的附加性内容。在大数据中，一方不仅仅分享主张，同时还分享授权策略。在每一个策略所有权和编写的位置都有一个策略访问点，在每个数据访问处都有一个策略执行点。一个策略执行点调用指定的策略决定点来进行可审计的决策，通过这种方式，不可否认性以及可信任的第三方的通常意义在可扩展访问控制标记语言中得到了扩展。大数据假定了大量的策略“点”，身份发布者以及数据。
- 策略编写点
- 策略决定点
- 策略执行点
- 策略访问点

4.2.3 数据治理

然而由于数据体积、速度、多样性、可变性等特点，大数据表现得十分大而复杂，大数据管理在一些重要概念和实际规模下将会更为广泛。缺少了大数据治理的大数据对于利益相关者而言会变得没那么有用。为了刺激积极的变化，大数据管理将需要持续覆盖数据生命周期中的静止状态、运动状态、不完善阶段，并且处理针对新一代、老一代，个体即组织以及组织即组织的安全与隐私服务。这将需要培养经济利益与创新，同时要推动行动自由以及促进个体与公共福利。在集成完善人类要素时，它将需要不被完全理解的标准管理技术和实践。大数据管

理将需要一些能接受现有技术缓慢性与无效性的新视角。一些数据管理的注意事项如下所列:

大数据应用程序来支持管理: 新应用的发展运用了大数据原则，旨在加强管理，这可能是出现在即的大数据应用中最有效的一种方法。

- 加密与密钥管理
  - 静止
  - 内存中
  - 传输中
  - 分散/集装箱化
  - 存储安全
- 数据丢失的预防与检测
- 网络服务网关
- 数据转换
  - 汇总数据管理
- 验证计算
- 加密数据计算
- 数据生命周期管理
- 配置、迁移、保留策略
- 有“风险”的 PII 微数据
  - 标志转换与匿名化
  - 风险管理的再识别
  - 终端验证
  - 数字权限管理(DRM)
  - 信任
- 公开性
  - 公正与信息伦理

4.2.4 基础设施管理

基础设施管理涉及到有关硬件操作与维护方面的安全与隐私的注意事项。一

些与基础设施管理有关的主题如下所列：

- 威胁和漏洞管理
  - 抵御 DoS 攻击的加密协议
- 监视与警报
  - 正如关键基础设施网络安全框架(CIICF)指出的，大数据为大规模安全情报、复杂事件融合、分析与监测提供了新的机遇。
- 缓解
  - 大数据漏洞缓解计划可能会有定性或定量的不同
  - 配置管理
  - 配置管理是保护系统和数据整体性的一个方面，包括以下内容：
- 补丁管理
  - 升级
  - 记录
- 大数据必须生产和管理具备更大多样性和速度的记录。例如，分析和统计抽样可能需要在一个不断变化的基础上。
- 这是一个容易理解的领域，但大数据可以跨过传统系统所有权边界。回顾 NIST 的“识别、保护、检测、响应和恢复”框架可能会发现大数据所特有的计划。
- 网络边界管理
  - 为安全信道建立一个不可知数据的连接
  - 共享服务网络结构，如欧洲电信标准协会(ETSI)TS 102 484 智能卡规范中，指定为“安全渠道用例和需求”。
  - 区域/云网络设计（包括连通性）
- 弹性，冗余与恢复
  - 弹性

大数据系统的安全设备相对于其他系统而言可能比较脆弱。一个给定的安全或隐私构造可能需要考虑这一点。弹性需求是特定领域的，但可能需要大数据系统规模的几何增长。

  - 冗余

大数据系统的冗余表明了不同层次的挑战。在大数据系统中复制来保持有意冗余一般在一个软件层面产生。在另一层面上，用来支撑故障转移，弹性或者减少数据中心潜在因素的完全冗余的系统可能会更加困难，由于大数据的速度、体积或其他方面因素。

- 恢复

恢复大数据安全故障可能需要大量的提前配置，超出了所需的小数据，应急计划和与用户的沟通可能需要类似大的规模。

4.2.5 风险与责任

风险与责任包括以下话题：

- 责任
- 信息，过程以及角色行为问责可以通过各种途径得以实现，包括：
  - 透明度门户和监测点
  - 正反向出处检查
  - 依从性
- 大数据依从性跨越了安全与隐私分类的多个方面，包括隐私，报道以及本国法律
  - 取证

通过大数据启用的取证技术

取证运用于大数据时的失败场景

- 企业风险水平

大数据评估应该被映射到此分类法的各要素中去，企业风险模型可以引入隐私方面的考量。

4.3 大数据安全隐私角色

大数据安全隐私的讨论应面向广泛大众，其中应包括密码学、安全学、合规性、信息技术等方面专家，以及对安全隐私成本与影响有所研究的领域专家和企业决策者。此外，应在大数据安全隐私文档前加入分类信息，以便研究者找到相关内容，并方便研究者针对相关内容提供反馈意见。

大数据生态系统中，参与的组织机构往往具有多样化的角色和工作流程。本文不仅提出了用于识别重要个人角色及其责任的模型，而且用同样的方法对安全控制点进行了分类，使角色更容易与相关安全控制点对应。

4.3.1 基础设施管理者

重要个人角色通常指某些个人或团体，这些个人或团体在大数据生态系统实现前进行技术决策，在大数据生态系统实现后解决系统缺陷和安全问题。

大数据生态系统有时通过多个组织机构合作实现，这时个人的技术工作将涉及多种技术、多种基础设施、多种工作流程或这三者的整合。对于大数据安全领域，还将涉 及身份识别，认证，访问控制和日志聚合。

他们的背景、实践以及所使用的术语，趋向于一致化。组织内部持续对他们施加以较少成本完成更多功能的压力。节省成本是一个潜在的压力，当新的问题产生时，基础设施技术会面临更多的压力。

4.3.2 管理者、风险管理者、合规工作者

数据管理是数据和数据系统管理的基础，数据管理是指数据的管理、形式化或规范化（例如行为模式）。风险管理包括基于大数据处理的正风险评估和负风险评估。合规性包括大数据操作过程涉及的法律、规范、协议或其他规则。一般而言，管理、风险管理和合规性（GRC）由组织机构中各个部门合作完成，包括法律、人力资源、IT 部门、合规性。某些产业和机构十分重视合规性，将合规性与纪律规则分离。

合规性专家常具备相同的背景，使用通用术语，并且采用相同工作步骤和工作流程，对相关市场和领域的其他组织机构具有深远影响。

在同一组织机构中，合规性专家旨在保护组织机构，免受人员流失、组织内个人行为和重要市场合规性风险等造成的负面影响。

大型企业和政府机构，合规性专家常任命到法律部门、市场部门、会计部门或与首席信息官相关的职位。内部审计人员和外部审计人员经常包含其中。

由于新型大数据策略的使用、资源缺乏和其他特殊因素，小型组织机构可能在没有合规性系统和合规性操作的情况下生成、拥有和处理大数据。尽管合规性

在大数据中十分重要，但在小型组织机构中并未受重视。

新合规性策略更够很好的应对上述问题，即使由一个人构成的公司中也能便捷的构建大数据应用，内置大量合规性操作。

安全隐私的框架需要为合规性增加相关数据和工作流程，这些数据和工作流程的功能类似于控制 NIST 大数据参考框架的系统控制器，详细内容可参见第五章。

4.3.3 信息工作者

信息工作者是指工作于信息生成、传输或处理的个人或团体。由于信息技术自身的特点和他们工作相关的业务特点，他们更倾向于使用通用的专业化术语。不过，他们的角色、责任和相关工作流程具有一定局限性，例如一个数据学家对数据内容和传输有深入研究，但往往只有在评估特定领域数据或分析工具而需要额外的工作、成本、风险或合规性责任时，才会关注到数据的安全性和隐私性。

信息工作者通常是数据管理者，有些可能是研究管员，他们通常具有多重管理角色或多重信息管理角色，他们的职责可能包括内容修订，检索或部分法律程序的法律职责。

信息工作者与服务和产品紧密相关。他们在组织机构命令下进行大数据分析，或实现有用数据商业化，或以服务提供者的身份进行商业数据传输，或通过分析第三方数据进行业务优化和业务增强。

4.4 角色与大数据安全隐私分类的关系

大数据安全隐私分为以下四类：数据机密性大数据安全隐私，数据可靠性大数据安全隐私，数据可用性大数据安全隐私，政策法规相关大数据安全隐私、社会学相关大数据安全隐私和跨组织产生的大数据安全隐私。通过以下三个角色进行举例说明，股东可以被定义为直接影响大数据解决方案选择和实施的个人或团体。决策者可以被定义为在选择或实施前，进行解决备选方案评估的个人或团体。例如，一个第三方安全咨询师可能作为决策者被组织雇用。而一个内部 IT 部门的安全专业人员在任务的安全监控阶段、安全维护阶段和安全审计阶段可能同时作为决策者和股东。

后续章节内容将探讨股东的利益，以及不同决策者分别处于哪三类角色。

4.4.1 数据机密性

信息加密 IT 专业人员应清楚相关定义、威胁模型、假设、安全措施、核心算法和协议。这些人员更像决策者，而非股东。点到点安全 IT 专业人员应具备密码学知识基础，且理解、掌握密码学迁移到现有安全基础设施和控制点的方式。

合规性工作者应当权衡关键需求（如电子健康档案相关 HIPAA 要求）和决策者对密码学及安全给出的评估。合规性工作的经理会先从有沟通需求的决策者做起。由于参加内部审计、外部审计和工作流程制定，这些角色的人员中也有股东。

4.4.2 数据可靠性

数据可靠性（准确性）与数据隐私相关，由于业务需要而保护知识产权，防止直接泄漏或者大数据分析导致的间接泄漏。它将信息工作者划归为决策者。信息工作者需要与密码学和安全的决策者合作来传达业务需求和实施控制

同样的，当一个组织时获取和消费数据时，信息工作者需要确认数据是可用的，并且需要在拷贝给其他组织前确认数据的完整性、非法地址、构造。

如果数据供应商没有进行适当等级的过滤数据标记化处理，那么组织机构可能会面临额外的风险。值得注意的是，美国卫生和人类服务部（HHS）在新闻发布会上宣布最终 HIPAA 的总括规则：

“今天宣布的变化，扩展了很多接收到健康信息的承包商和分包商等的业务需求。有些报告至 HHS 的大型漏洞包含商务联系。对于不合规的处罚金有所上涨，最高为金额 1500000.“

组织机构在生态系统模式使用或分享健康数据，这些生态系统模式包括手机应用和 SaaS 提供商，需要确定合适的法律规定来保障信息的纯确性和可靠性。。

4.4.3 数据可用性

系统健康是一个传统 IT 领域，其 IT 管理者是技术、协议和产品的股东和决策者。IT 管理者同样负责系统健康维护的职责划分，这个系统将提供数据、分析

或服务，属于电话行业的 OSS 领域，在联合服务方面经验丰富。

安全和密码学专家应该仔细检查系统健康店的运营架构和潜在缺陷。当系统基础设施包含多种学术和产品时，这种缺陷可能会扩大化。

系统健康类似于伞的概念，处在信息工作者和基础设施管理的交叉点。与人体的健康一样，大数据系统监控会快速产生大量大数据，其中大量和快速是大数据的特点。与人类健康诊断相似，有些潜在的迹象能够反映白细胞的防御措施。别的可能反映缺乏免疫力的健康状况，如高血压。同理，大数据系统可采用安全信息和事件管理，或大数据分析来监控系统健康。

大数据系统健康的体量、速度、种类和变化，明显不同于小型系统。现有系统的健康工具和设计模式可能不足以处理包括大数据安全和大数据隐私的大数据。有商业网络服务供应商报告说，相比单一应用程序，内部会计和系统管理工具使用的资源更多。系统事件的数量和事件交互的复杂性是一个巨大的挑战，需要大数据解决方案，来维护大数据系统。管理系统的健康，包括安全需要的角色定义和组织化管理的工具集。从另一个角度看，大数据正在改变计算机安全工作人员的角色。

例如，被 DevOps 运动所吸引（例如，移向应用程序开发或系统运维团队开展的混合任务），这项任务快速、方便配置、大数据系统部署与分配。跟踪、恶意或意外配置变更在大数据领域有所增加。

4.4.4 政策法规、社会和跨组织

安全隐私相关公共策略由联邦贸易委员会、美国食品和药物管理局、DHHS 办公室和联邦协调委员会等联邦机构发布。其中，美国国土安全部安全警备小组负责美国本土计算机安全。社会话角色包括非政府组织、利益集团、专业机构和标准制定组织。跨组织角色包括某些产业或产业间通用的设计模式，如医药、物流、制造、分销（便于数据共享）、综合处理和综合管理的设计模式或通用设计模式。大数据框架受跨组织化影响，也将影响大数据跨组织的发展。

4.5 其他类别

虽然其他大数据安全隐私类别已经提出，但是却未详细定义。并且这些类别

是应归入已有类别，还是应定义和研究成为新的类别，这个问题尚未定论。其中部分类别简述如下。

4.5.1 供应、计量和计费

供应、计量和计费是金融系统用于资产管理，资产使用计量和资产使用计费的基本要素。如果系统能提供灵活的服务，匹配合适的计费方式，并且将计费系统整合其中，大数据金融化将更加敏捷。虽然流程目前仅适用于少数参与者的情况，但是对存在许多供应商、消费者和服务供应商这种情况，这一流程可以迅速复杂化从而适应这种新的情况。

参与现有业务流程的信息工作者和 IT 专业人员，可能会成为决策者候选人或股东，他们将决定供应、计量的安全和隐私是否成为系统设计的一部分。由于计量数据和计费数据呈爆炸式增长，而潜在应用和风险信息极易被淹没，所以潜在应用和风险并没有被充分挖掘。

此外，还有一些关注金融系统准确性和有效性的类别。而合规性类别，例如审计和恢复，可能与供应和计量有所重叠。

4.5.2 数据出售

数据是可以进行购买和销售的资产是大数据系统的特点之一。谷歌搜索依赖用户的废弃搜索信息，也就是用户进行大数据搜索后保存到表单上的搜索信息。谷歌和脸书可以选择对数据进行重新包装然后为其他人提供有偿服务。

类似于服务出售，如果参与者拥有大数据供应者、传输者、消费者中多种角色，那么此数据生态系统将具有更高的价值。此外，如何选择数据出售模型是是需要进一步研究的领域。为适应 PII、可靠性和管理，信息工作者和 IT 专业人员也就是决策候选人和股东需要研究更复杂的数据出售模型。而数据出售还包括安全隐私的风险和责任转移。

5 安全和隐私组织

对 NIST 大数据参考架构而言，安全和隐私考虑是其不可或缺的一方面。根

据收集的材料和 NBD-PWG 安全和隐私子组成员及其他人之间开展大规模的头脑风暴，形成了以下对安全和隐私组织的建议。

安全和隐私组织：安全和隐私方面的考虑是 NIST 大数据参考架构的一个必要方面。这一点在图 4 中通过安全和隐私组织围绕着五大主要组件被几何地呈现出来，因为所有组件都受安全和隐私考虑的影响。于是安全和隐私的作用通过与组件之间的关系被正确地描述出来，却没有详述细节，虽然这些细节可能更精确，但最好在安全和隐私参考架构中进行详述。数据提供者和数据消费者都包含在安全和隐私组织中，因为至少他们应该同意在合适的位置部署安全协议和安全机制。安全和隐私组织是一种类比，用来暗示错综复杂互相关联的本质以及安全和隐私在 NIST 大数据参考架构中的无处不在。

无处不在这一方面体现在图 4 中，安全和隐私组织围绕着所有的功能组件。NBD-PGW 决定将数据提供者，数据消费者以及大数据应用和架构提供者都囊括到安全和隐私组织中，因为这些实体应该同意在合适的位置部署安全协议和安全机制。NIST 大数据互操作框架：卷 6，参考架构文稿中详细讨论了 NIST 大数据参考架构中的其他组件。

此时，对提出的组织概念如何在 NIST 大数据参考架构的每个组件中实现的说明是粗略的，更多是建议性而不是规定的。但是，相信迟早模板将进化并形成可靠的根据用于更加精细的 迭代。

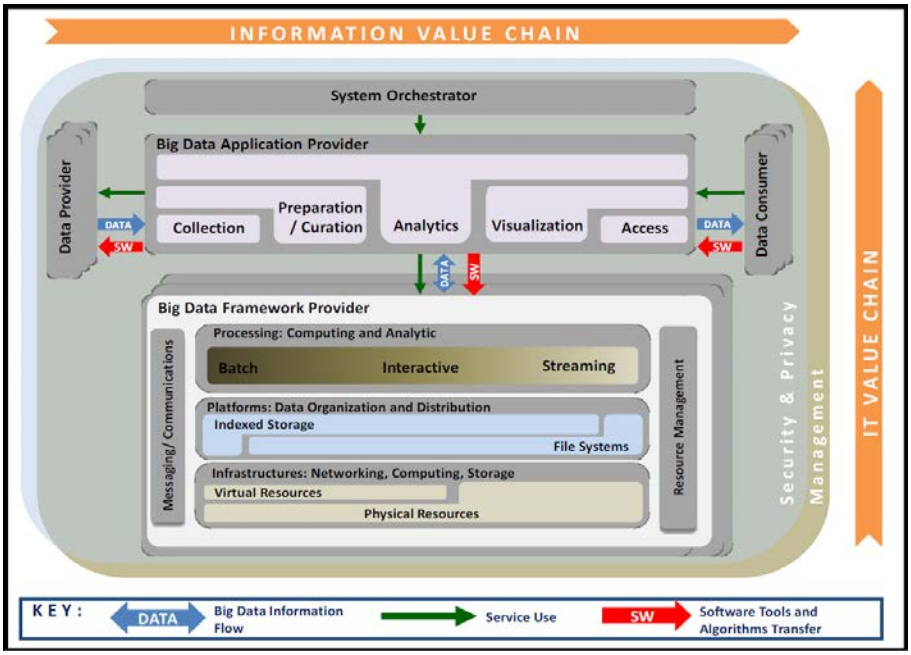

图 4：NIST 大数据参考架构

图 4 引入了两个新的概念：信息价值链和 IT 价值链，这对安全和隐私考虑来说是非常重要的。

信息价值链：虽然信息价值链不适用于所有域，但通过信息价值的增加，减少，细化，定义或其他变化，存在一系列隐含的处理操作。在每一阶段的物源保存和其他安全机制的应用，可能受限于国家特定的对信息价值的贡献。

IT 价值链：平台特定的考虑适用于大数据系统扩张期。在缩放过程中，特定的安全，隐私，或 GRC 机制或实践都需要被调用。

5.1 NIST 大数据参考架构中的安全和隐私组织

图 5 提供了一些安全和隐私话题的概述，这些话题与 NIST 大数据参考架构中某些关键组件和接口相关。本图代表了对安全和隐私组织与 NIST 大数据参考架构相互交织特性的最初描述。

图 5 是否会在本文档第二版本中进一步开发还尚未可知。然而，安全和隐私组织和 NIST 大数据参考架构的关联以及安全隐私分类和 NIST 大数据参考架构的关联将会在本文档第二版本中进一步研究。

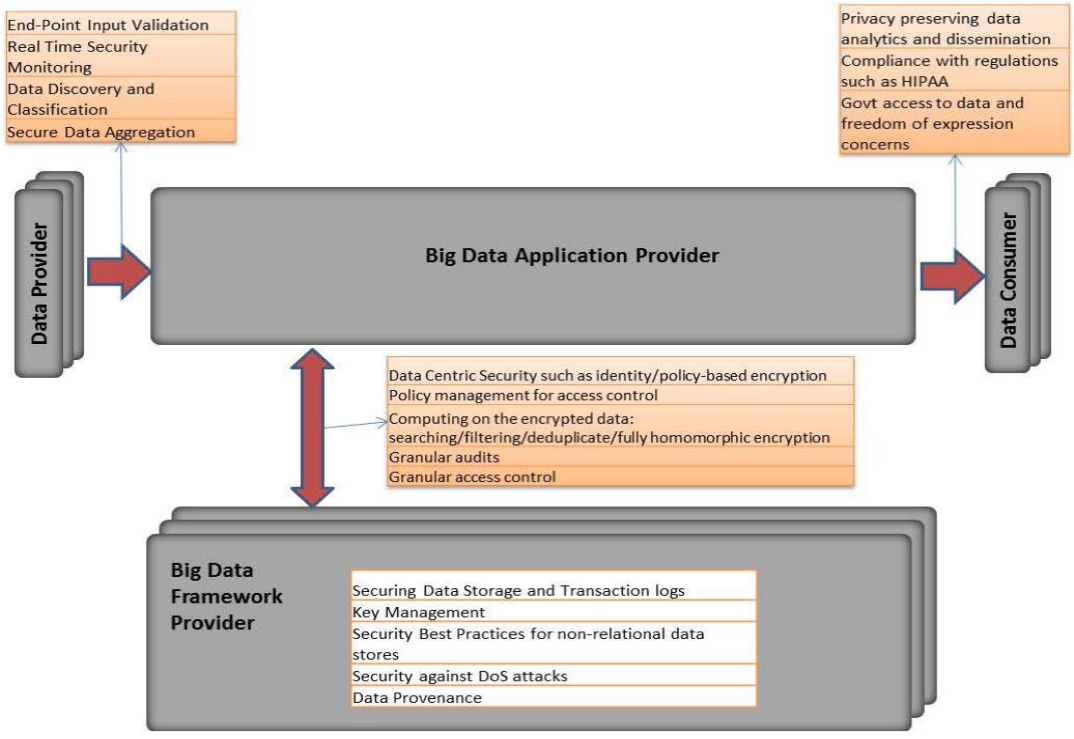

图 5 理论上安全和隐私组织与 NIST 大数据参考架构的重叠

图 5 中组和接口描述如下。

- 数据提供者→大数据应用提供者之间的接口

来自数据提供者的数据可能要被验证完整性和真实性。传入流量可能被恶意用于发动 DoS 攻击或者利用软件漏洞进行攻击。因此，实时安全监控是有效的。数据发现和分类应该在尊重隐私的前提下执行。

- 大数据应用提供者→数据消费者之间的接口

数据，包括交付给数据消费者的汇总结果，必须尊重隐私。第三方或其他实体访问数据需要遵循法律法规，例如 HIPAA。关注点包括政府对敏感数据的访问。

- 大数据应用提供者→大数据框架提供者的接口

数据可以在加密状态下被存储和检索。应该存在合适的接入控制策略，以确保只有使用正确的凭证以需要的粒度才能访问数据。先进的加密技术可以允许应用程序可以对数据进行丰富的基于策略的访问，对加密数据进行搜索，过滤，以及对明文的计算。

- 大数据框架提供者内部

应该保持静止数据和交易日志数据是安全的。密钥管理是访问控制和跟踪密钥的关键。非关系型数据库应该有一个安全措施层。数据起源对持有合适的上下文来保证每一阶段数据安全和数据功能来说是非常重要的。应该减少拒绝服务攻击以保证数据的可用性。

- 系统控制器

系统控制器在识别，管理，审计和测序大数据各组件进程中发挥着关键作用。例如，一个将数据从收集阶段转移到进一步准备阶段的工作流可以执行安全或隐私方面的工作。

系统控制器为对手提供了一个额外的，诱人的攻击面。系统控制器往往需要永久性或暂时性提高的权限。系统控制器为贯彻安全机制，监控源，系统接入管理工具，审计要点提供以及无意征服隐私或其他信息的保证措施提供了机会。

5.2 隐私工程原则

大数据安全和隐私应该充分利用现有的标准和实践。在隐私领域，一个在整个过程中考虑隐私的系统方法，对在大数据场景下考虑采用安全和隐私实践是非

常有用的指导。结构化信息标准促进组织（OASIS）隐私管理参考模型（PMRM），包含七个基本原则，为大系统架构师提供了恰当的基础指导。当涉及到任何个人数据时，隐私应该是大数据系统设计中一个必要元素。

其他隐私工程框架也在考虑范围内。

相关原则包括身份管理框架，例如在网络空间可信身份的国家策略（NSTIC）提出的和在 NIST 云计算安全参考架构中考虑的身份管理框架。有助于安全和隐私组织的身份管理方面将在本文档后续版本中解决。

大数据框架也可以用于加强安全性。通过安全情报，事件检测和取证，大数据分析可以用来检测隐私泄露。

5.3 大数据安全操作分类和 NIST 大数据参考架构之间的关联

表 1 描述了操作分类和 NBDRA 组件之间的初步的映射关系。每个操作分类单元（4.2 章节）的话题和活动被分配到每个 NBDRA 组件中，如表 1 活动栏所示。描述栏中提供了每个 NBDRA 组件安全和隐私方面的额外信息。

表 1 安全操作分类与 NBDRA 组件的映射

活动	描述
系统控制器	
• 策略执行 • 安全元数据模型 • 数据损失预防，检测 • 数据生命周期管理 • 威胁和漏洞管理 • 缓解 • 配置管理 • 监测，报警 • 恶意软件监控和修复 • 跳回，冗余和恢复 • 问责制 • 遵从性 • 取证 • 业务风险模型	几个安全功能已经映射到系统控制器块内，因为他们需要架构级别的决策和意识。这些功能方面都是和安全组织强相关的，这样可以以不同形式的操作细节通过不同的节点来触摸整个架构。这样的安全功能包括国家特有的合规要求，取证广泛扩大的要求以及特定领域具有隐私意识的业务风险模型。
数据提供者	

• 设备，用户，资产，服务，应用程序注册 • 应用层身份 • 最终用户层身份管理 • 终点输入验证 • 数字版权管理 • 监测，报警	数据提供者要保证数据的真实性，相应地，敏感的，受版权保护的或有价值的数据应该得到充分的保护。这将导致实体注册和身份生态系统的可操作性方面。
数据消费者	
• 应用层的身份 • 最终用户层身份管理 • Web 服务网关 • 数字版权管理 • 监测，报警	数据消费者在义务与要求方面展现出与数据提供者的二元性，只有他们面对应用提供者的访问/可视化方面。
应用提供者	
• 应用层的身份 • Web 服务网关 • 数据传输 • 数字版权管理 • 监测，报警	应用提供者接口存在于数据提供者和数据消费者之间。它参与了与这些块之间所有安全接口协议，以及维持了与框架提供者之间的安全交互。
框架提供者	
• 虚拟层身份 • 身份提供者 • 加密和密钥管理 • 隔离/集装箱 • 存储安全 • 网络边界控制 • 监测，报警	框架提供者负责数据/计算的安全性，占据了数据生命周期的大部分份额。这包括通过加密和访问控制确保静止数据的安全；通过隔离/虚拟化的计算安全；以及与应用提供者的通信安全。

6. 映射用例至 NBDRA

在本小节，第 3 节中展示的安全与隐私相关的用例将被映射到图 5 中的 NBDRA 组件和接口，图 5 是虚拟化安全和私有化构造重叠到 NBDRA。

6.1 用户数字化媒体用例

通过演示门户，内容所有者为消费者使用授权数据。消费者使用数字媒体会

产生大数据，包括用户层级的人口统计资料以及类似播放顺序、推荐和内容导航这样的使用模式。

表 2 映射消费者数字媒体使用至参考架构

NBDRA 组件和接口	安全和隐私主题	用例映射
数据提供者→ 应用提供者	结点输入验证 实时安全监测 数据发现和分类 安全数据融合	数据异同且依赖于供应商。存在电子欺骗的可能性。例如，由安全的微软权限管理服务提供保护。安全/多用途互联网邮件扩展（S/ MIME） 内容创作安全 通过媒体、*种群*和频道，发现/分类是有可能的 供应商提供的服务－安全融合措施是不透明
应用提供者→ 数据消费者	隐私保护的数据分析 遵守规定 政府访问数据和言论自由的问题	向内容所有者汇总报告 PII 披露问题层出不穷 各种问题；例如，播放恐怖播客以及非法回放
数据提供者→ 框架提供者	数据中心安全，如身份/基于政策的加密访问控制涉及的政策管理 计算处理加密数据：搜索/过滤/去重/全同态加密 审计	未知 用户，管理员回放，库维护以及审计员 未知 审计使用 DRM 的费用
框架提供者	安全数据存储和事务日志 秘钥管理 非关系型数据存储的最佳安全做法 安全应对 DoS 攻击 数据源	未知 未知 未知 N/A 数据所有者，生产者，消费者的可溯性是被保存的
Fabric	安全情报分析 事件检测 取证	未经批准的使用/访问的机器智能 定义“回放”粒度 播放记录的法律纠纷传票

6.2 Nielsen 家庭调查：Apollo 项目

Nielsen Homescan 涉及家庭层级的零售交易和相关媒体曝光，使用了统计学上有效的国家样本。一般性描述由供应商提供。该项基于 2006 年阿波罗计划架构。（阿波罗计划并没有出现其原型状态。）

表 3 映射 Nielsen Homescan 至参考架构

NBDRA 组件和接口	安全和隐私主题	用例映射
数据提供者→ 应用提供者	结点输入验证 实时安全监测 数据发现和分类 安全数据融合	数字源的特定设备键； 经内部扫描的接受源并调解到家庭 ID（角色问题） 无 基于数据源的分类（例如，零售商店，装置和纸源） 汇总成人口交叉表。内部分析师此前访问 PII
应用提供者→ 数据消费者	隐私保护的数据分析 遵守规定 政府访问数据和言论自由的问题	（有时）汇总到特定产品，统计学上有效的独立变量预先进行面板数据权利保，并通过组织实施控制 N／A
数据提供者→ 框架提供者	数据中心安全，如身份/基于政策的加密 访问控制涉及的政策管理 计算处理加密数据：搜索/过滤/去重/全同态加密 审计	加密不到位；仅用于数据中心到数据中心传输。XML（可扩展标记语言）多维数据集安全映射到 Sybase IQ 和报告工具 广泛的基于角色控制 N/A Schematron 的流程和审核的步骤
框架提供者	安全数据存储和事务日志 秘钥管理 非关系型数据存储的最佳安全做法	通过基础设施团队确保具体项目审计 由项目首席安全官（CSO）进行管理。发行为客户和内部用户单独的密钥对 通过 XML 模式验证常规数

	安全应对 DoS 攻击 数据源	据完整性检查 为查询子系统提供业界标准的虚拟主机提供商保护
Fabric	安全情报分析 事件检测 取证	没有具体项目计划 N/A 用法，*cube*－*creation* 和设备合并审计记录被保留作取证和计费

6.3 网络流量分析

访问层的网络服务器的日志信息具有高粒度和数据量大的特点，网络日志和其他信息源是息息相关的，这些信息源包括页面内容（页面按钮、文本、导航项）和市场营销活动，如广告宣传和媒体分类。

表 4 映射网站流量分析到参考架构

NBDRA 组件与接口	安全和隐私话题	案例使用的映射
数据提供→应用提供	终端输入校验	依赖设备。入侵容易。
	实时安全监控	网络服务器监控
	数据发现和分类安全数据的聚合	聚合数据到设备，访客，按钮，网页事件，以及其它
应用提供→数据消费者	隐私保护数据分析	IP 的匿名和时间戳的降解
		用户指定的内容隐藏
	遵循规则	匿名服务需要欧盟制定的规则。自愿选择匿名与否
	政府监管数据并且关注自由	是
数据提供→框架提供	数据中心的安全，如 认证/基于策略的加密	很依赖于数据管理员
	访问策略管理控制	系统和应用级的访问控制

	加密数据的计算：搜索/过滤/重复/全同态加密	未知
	审核	支持客户审核的准确性和完整性
框架提供	保护数据存储和事务日志	存储归档是个大问题
	密钥管理	首席安全官和应用
	非关系型数据的存储的最佳实践	未知
	针对拒绝服务的攻击	标准
	数据源	服务器,应用,IP 身份认证,页面内任意时间点的文档对象模型(DOM),和任意时间点的市场营销事件
结构	安全情报分析	访问 Web 日志往往需要较高的权限
	事件监测	可以推断出如下信息：巨大的销售量，市场营销以及整个网络的健康状态
	取证	参考 SIME

6.4 健康信息交换

健康信息交换（HIE）数据是从各种数据提供者聚集得到的，其中可能包括的涵盖实体，如医院和指定参加临床试验的合同研究组织（CROs）。数据消费者将包括急诊室人员，疾病预防控制中心，以及其他授权卫生（或其他方式）的组织。因为任何一个城市或地区有可能实现自己的 HIE，这些交换也可能服务于数据消费者和数据提供者彼此之间。

表格 5 HIE 映射参考体系架构

NBDRA 组件和接口	安全和隐私主题	用例映射
	端点输入验证	强大的身份验证，可能通

数据提供者-->应用程序提供者		过的 X.509v3 证书，SAFE(所有的签名和身份验证)的潜在影响力，代替一般 PKI 的桥梁
	实时安全监控	来电记录的验证，以通过签名验证保证完整性,通过确定的加密 PHI 来保证 HIPAA 的隐私。可能需要检查知情同意书证据。
	数据发现和分类	利用健康等级 7（HL7）和其他的标准格式投机，但要避免在模式规范化方面的尝试。一些列将被高度加密的，而其他将被特殊加密的（或与之相关联的加密元数据）启用发现和分类。可能需要根据数据源的策略或 HIE 服务提供商执行列过滤。
	安全数据融合	清除文本的列可以进行重复数据删除，可能是确定性加密的列。其他列可以具有用于促进聚集和重复数据删除的加密的元数据。保留规则可以假设，但性格规则在合规的相关领域是不能被假设的。
应用程序提供者-->数据消费者	隐私保护数据分析	搜索加密数据和数据占有的证明。由于临床试验的参与，导致潜在不良识别的经验。识别潜在的职业病人。趋势和流行病，以及这些共关系环境等效果。药物是否被执行的决定将会产生不良反应，不破坏双盲。患者将需要被提供访问和使用他们的电子病历数据的详细记帐。
	规则的一致性	HIPAA 安全和隐私将需要访问电子病历(EHR)数据的明细核算。促进这一点，

		记录和提醒，将需要数据消费者的联合身份集成。
	政府访问数据和言论自由问题	CDC，执法，传票及认股权证。访问可能根据流感的出现（例如，CDC）或许可证据（例如，执法）会被切换。
数据提供者--> 框架提供者	数据中心的安全，如身份/基于策略的加密策略管理访问控制 计算上的加密数据：搜索/过滤/重复数据删除/全同步加密 审计	行级和列级访问控制 基于角色和基于声明为主。定义 PHI 单元 隐私保护访问相关的事件，异常，以及 CDC 的趋势和其他有关卫生机构 促进 HIPAA 准备和 HHS 审计
框架提供者	安全数据存储和事务日志 密钥管理 非关系数据存储的安全最佳实践 针对分布式拒绝服务（DDoS）攻击安全性 数据源	完整性和保密性需要被保护，而且以强调可用性的方式建立完整性 联和的整个涵盖实体，有必要通过数据源的多个涵盖实体来管理钥匙的生命周期 终端到终端的加密，以最小熵相关的特殊情景方案来提供更丰富的查询操作而不影响可以容忍的隐私 强制性要求：系统必须幸免于 DDoS 攻击 完整性和所有访问和修改的数据的完整性。此信息可以是和数据一样的敏感，可以作为受相称访问策略
构造	安全情报分析	知情患者同意的监控，授

	事件检测 取证	权和未授权的转让，访问和修改 记录保管转移，记录的增加/修改（或单元），授权查询，未授权查询，并尝试修改 防篡改日志，证据篡改事件。识别监管的记录级转让和单元级访问或修改的能力

6.5 基因隐私

基因隐私的映射有待开发并且在未来的版本中将会出现。

6.6 临床试验数据共享

在某一工业贸易组织的提议下，新药物临床试验的数据可在企业内存储之外共享。

表 6：药物临床试验数据共享到参考架构的映射关系

NBDRA 组件和接口	安全和隐私话题	用例映射
数据提供方 → 应用提供方	终端输入验证； 实时安全监控； 数据发现和分类； 安全数据汇总；	不透明（Opaque）—公司特有的； 无； 不透明（Opaque）—公司特有的； 第三方汇总
应用供应方 → 数据消费者	隐私保护数据分析； 符合规定； 政府对数据的访问和表达观点的自由；	汇总数据不出现潜在细节统计特性； 负责的开发商和第三方托管； 有限应用于研究领域，但未来可能存在公共健康数据的担忧。 仅有临床研究报告，但要适当在研究和病人两个层面进行选择
数据提供方 ↔ 框架提供方	数据中心的安全，如基于身份/策略加密； 访问控制的管理策略；	TBD； 内置角色;第三方托管角色;研究者角色;参与的患者医生；

	加密数据的计算：搜索/过滤/重复数据删除/全同态加密； 审核；	TBD； 第三方审核
框架提供方	安全数据存储与业务日志； 密钥管理； 对非关系数据存储的最佳安全实践； 针对 DoS 攻击的安全措施； 数据源；	TBD； 公司内部各不同；外部 TBD； TBD； 不可能成为公开； TBD-关键问题
结构(fabric)	安全情报分析； 事件检测； 取证；	TBD； TBD；

6.7 网络保护

SIEM 是一个用来保卫和维护网络的工具组。

表 7：网络保护到参考架构的映射关系

NBDRA 组件和接口	安全和隐私话题	用例映射
数据提供方 → 应用提供方	终端输入验证； 实时安全监控； 数据发现和分类； 安全数据汇总；	软件供应商特定的;参考市售终端验证 52 --- 随工具不同而变化，但需在安全语义和来源的基础上进行分类 通过子网，工作站和服务器聚集
应用供应方 → 数据消费者	隐私保护数据分析； 符合规定； 政府对数据的访问和表达观点的自由；	平台定制； 适用，但监管活动对分析师不易可见; NSA 和 FBI 可以按需访问
数据提供方 ↔ 框架提供方	数据中心的安全，如基于身份/策略加密； 访问控制的管理策略； 加密数据的计算：搜索/过滤/重复数据删除/全同态加密； 审核；	操作系统的一般特征； 例如，针对事件日志的一组策略； 供应商和平台定制； 复杂--审核可能贯穿始终；
框架提供方	安全数据存储与业务日志；	供应商和平台定制；

	密钥管理； 对非关系数据存储的最佳安全实践； 针对 DoS 攻击的安全措施； 数据源；	首席安全官和 SIEM 产品密钥； TBD ； 大数据应用层 DDoS 攻击可以通过交通及其相关性分析的结合得到缓解； 例如，如何知道一个入侵记录实际上与特定工作站关联
结构(fabric)	安全情报分析； 事件检测； 取证；	当前 SIEMs 特征； 当前 SIEMs 特征； 当前 SIEMs 特征；

6.8 军事：无人驾驶车辆传感器数据

无人驾驶车辆（无人机）及其车载传感器（例如：流视频）能产生应该储存在非标准格式中的 PB 级数据。美国政府正在推行扩展用于大数据，如流视频的存储能力的功能。欲了解更多信息，请参阅国防信息系统局（DISA）在国防部私有云中艾字节的大数据对象协议 53。

表 8：军用无人机传感器数据到参考架构的映射

NBDRA 组件和接口	安全和隐私话题	用例映射
数据提供方 → 应用提供方	终端输入验证； 实时安全监控； 数据发现和分类； 安全数据汇总；	需要保护传感器（例如，相机），以防止欺骗/被盗的传感器流。在国防部管道中存在新的收发器和协议。传感器流将包括智能手机和平板电脑等来源； 板载（onboard）和控制站次级传感器的安全监控； 变化范围从媒体特定编码到复杂情景感知增强融合方案； 融合挑战从简单到复杂。视频流可能会遭到不安全或不集成地使用 54。
应用供应方 → 数据消费者	隐私保护数据分析； 符合规定； 政府对数据的访问和表达观点的自由；	地理空间的限制：超越通用横轴墨卡托投影（UTM）无法监视；军事秘密：起源隐私的目标和点 大量的，也有标准问题； 例如，谷歌街景诉讼

数据提供方 ↔ 框架提供方	数据中心的安全，如身份/基于策略的加密； 访问控制的管理策略； 加密数据的计算：搜索/过滤/重复数据删除/全同态加密； 审核；	基于策略的加密，往往是由传统的信道容量/类型决定； 转换往往是在 DOD/承包商制定的系统计划内做出的； 有时在供应商提供的架构内进行，或者通过图像处理并行； CSO 和监察长（OIG）审核
框架提供方	安全数据存储与业务日志； 密钥管理； 对非关系数据存储的最佳安全措施； 针对 DoS 攻击的安全措施； 数据源；	通常，加上数据中心安全级别严格管理（例如，营 VS 团 VS 司令部） CSO--指挥链； 目前不用不同方式处理;这由国防部改变； 国防部抗干扰电子措施； 必须跟踪到传感器的时间点配置和元数据
结构(fabric)	安全情报分析； 事件检测； 取证；	国防部开发战斗安全软件智能的特定领域--事件驱动和监测--往往是远程的 例如，在视频流中目标识别，从阴影推断目标的高度。融合来自卫星红外线与独立传感器流的数据； 用于行动后回顾（AAR）--全面支持传感器数据流的播放

6.9 教育：共同核心学生表现报告

为每位同学进行从“摇篮到坟墓”的表现评价现在成为可能 – 至少在基础教育中（K-12），而且很超越基础教育范畴。这可能包括曾经给予的每个测试结果。

表 9：共同核心 K-12 学生报告到参考架构的映射

NBDRA 组件和接口	安全和隐私话题	用例映射
数据提供方	终端输入验证；	依赖于应用的，欺骗成为可能；

→ 应用提供方	实时安全监控； 数据发现和分类； 安全数据汇总；	测试，测试者，管理者和数据的特定供应商监控； 未知； 典型：课堂级的
应用供应方 → 数据消费者	隐私保护数据分析； 符合规定； 政府对数据的访问和表达观点的自由；	不同的：比如，在所有同年级教室的教师水平分析； 家长，学生和纳税人的信息披露和隐私应用； 是。可能因为资金资助而需要披露，可能因为教师、管理员和地区的质量评价而需要披露；
数据提供方 ↔ 框架提供方	数据中心的安全，如身份/基于策略的加密； 访问控制的管理策略； 加密数据的计算：搜索/过滤/重复数据删除/全同态加密； 审核；	同时支持单个接入（学生）和分区汇总； 供应商（例如，皮尔森）控制，州级政策，联邦级政策;大概 20-50 种不同的角色目前都有明确规定 55； 已提议； 通过工会，国家机关，响应传票来支持内部和第三方审核
框架提供方	安全数据存储与业务日志； 密钥管理； 对非关系数据存储的最佳安全措施； 针对 DoS 攻击的安全措施； 数据源；	大型企业的安全，传输等级控制—从教室到联邦政府； 从教室水平到国家级别的 CSO； --- 标准； 溯源到测量事件需要在某个时间点捕获测试，这可能本身就需要一个大数据平台
结构(fabric)	安全情报分析； 事件检测； 取证；	各种商业安全应用； 各种商业安全应用； 各种商业安全应用；

6.10 传感器数据存储和分析

传感器数据存储和分析的映射正在开发之中，将包含在本文件未来的版本中。

6.11 货物托运

该用例提供了一个大数据应用相关的航运业的概述，其标准可能出现在不久的将来。

表 10：货物运输到参考架构的映射

NBDRA 组件和接口	安全和隐私话题	用例映射
数据提供方 → 应用提供方	终端输入验证； 实时安全监控； 数据发现和分类； 安全数据汇总；	确保传感器手机的数据的完整性； 传感器可以检测异常 ，为有特殊要求的包装检测温度/环境等条件，还可以检测泄漏/辐射； --- 安全汇聚来自传感器的数据
应用供应方 → 数据消费者	隐私保护数据分析； 符合规定； 政府对数据的访问和表达观点的自由；	传感器采集的数据可能是私人的，可以揭示这些封装和地理信息.对揭示信息需要进行隐私保护； --- 美国国土安全部可以监视可疑包裹进/出境动态
数据提供方 ↔ 框架提供方	数据中心的安全，如身份/基于策略的加密； 访问控制的管理策略； 加密数据的计算：搜索/过滤/重复数据删除/全同态加密； 审核；	--- 私人的，敏感的传感器数据和包数据只应提供给授权个体。第三方商用产品可以实现对数据的低级访问； 见上面“Transformation”章节； ---

框架提供方	安全数据存储与业务日志； 密钥管理； 对非关系数据存储的最佳安全措施； 针对 DoS 攻击的安全措施； 数据源；	传感器数据的日志记录对于跟踪包是必不可少的。处在休息状态的传感器数据应该保存在安全的数据存储中； 用于加密数据； 传感器类型和数据类型的多样性可能需要使用到非关系数据存储； --- 元数据应该被加密连接到采集到的数据上去，以便确保源头和进程的完整性。源头的完整保存，有时会要求一个独立的大数据应用
结构(fabric)	安全情报分析； 事件检测； 取证；	传感器数据的异常对数据流量的篡改/欺诈插入具有指示作用； 可以检测到异常事件，如货物搬离预定路线或者莫名静止； 记录数据的分析可以在事故发生后揭示它们的详细情况

附录 C:大数据参与者和角色：适用于大数据业务场景

早在 20 世纪初期，面向服务的体系结构（SOA）广为流传。虽然利用参与者和角色概念的相对较少，但 SOA 已经影响了系统分析过程，在一定程度上也影响了系统设计。Patig and Lopez-Sanz 等人指出，参与者和角色纳入统一建模语言后，使这些概念以服务的形式很好的展现出来。大数据将会更好的适用于这些概念。然而参与者和角色的概念还没有被安全组织完全采纳，小组觉得如何将这些概念从传统方式改变到适用于 SOA，对于大数据设计者是非常重要的。

业务流程执行语言（BPEL）和业务流程模型及规则（BPM&N）框架的相似编排，为大数据安全和隐私类标准提供了可借鉴模式。 Ardagna 等人建议如何调整可能会继续 SOA，但大数据系统提出不同的挑战。

大数据系统既包括简单的 M2M 通信用户关系，也包括人、机器的系统化、体系化的复杂用户关系组合. 用户意味着指定系统中的角色到自然人。从公民角度

来看，一个自然人可以与大数据系统中的应用和信息资源有着牵丝万缕的关系。

以下列表描述了大量用户以及用户随时间的变化。对于某些系统，角色仅能在某一时间点上有效。对于大数据系统用户安全是有挑战性的，例如传统架构没有考虑明确的归档或删除政策，具体如下：

- 零售组织涉及消费者或者有够买预期的自然人;然后,消费者成为一个客户;
- 与银行服务的金融机构的客户关系的自然人;
- 可以在不同组织或相近金融机构进行汽车消费贷款的自然人;
- 从不同银行或同一银行进行住房贷款的自然人;
- 对健康、生命、汽车、房屋或租房进行保险的自然人;
- 通过私立部门代发工资或通过公共部门就业战略受益的或被保险的自然人;
- 关心一家或多家公共或私立学校的自然人;
- 一个或多个国有、私人企业的雇员、临时工、承包方或第三方雇员的自然人;
- 未成年的，被法律或其他保护的自然人;
- 一个自然人可能同时承担一个或多个角色。
- 对于每一类角色，系统拥有者应该明确是否为用户做了如下方面：
- 对进入系统的 PII 进行了标识;
- 如何去身份标识，何时采取何种方法;
- 数据的完整性校验，错误、漏报和不确定性的纠正;
- 信息删除，采取自动化机制报告和验证删除效果;
- 大数据系统与其他系统发生关联时，可以加入多级退出体系;
- 验证数据没有超过监管（例如与年龄有关）、政府（例如一个地区或国家）或期满（例如不在是一个客户）的界限。

退出访问：

一个标准组织不断调整框架,直到利用这个框架可以保护到每个人的安全和隐私，同时，一些观察员认为两个简单的协议可以用于大数据中个人信息保护的治理，具体如下：

协议一：一个人仅能决定选择加入自身的个人数据资料，并且可以决定随时删除;

协议二：个人隐私和安全退出过程中，应使每一个人可以随时修改他们访问和审阅日志文件和报告的时间，创建自毁时间（类似欧盟的“被遗忘权”。）