NIST 大数据定义（草案）

NIST Big Data Public Working Group
Definitions and Taxonomies Subgroup

2015 年 3 月 2 日

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 研究报告“DRAFT NIST Big Data Interoperability Framework:Volume 1, Definitions”的大力支持（排名不分先后）：

山西天地科技有限公司

北京市闪联信息产业协会

清华大学

深圳市金蝶中间件有限公司

中科恒源信息科技有限公司

浪潮软件集团有限公司

北京世纪互联宽带数据中心有限公司

中国科学院计算机网络信息中心

华为技术有限公司

浪潮电子信息产业股份有限公司

北京数码大方科技股份有限公司

山东省标准化研究院

目录

1 介绍

1.1 背景

大数据在激发创新、刺激商业、推动社会进步方面的巨大潜能已得到商业界、学术界和政府部门的共同认可。大数据是用来描述网络化、数字化、传感器化、信息化社会数据泛滥的常用术语。很多以前无法解答的问题，如今通过对大数据资源的有效利用，已经可能解答，这些问题主要包括：

- 如何及早发现流行性疾病并做好预防？
- 如何在高性能新材料合成以前，就通过数据分析来发掘它们？
- 如何扭转网络攻击优于防守的趋势，防范网络安全威胁？

但同时，各方对于大数据相对传统方法究竟有多大优势，也有清醒认识——数据本身的容量、速度和复杂性的增长率，已超出现有科技在数据分析、数据管理、数据传输以及用户领域的发展进度。

如上，对于大数据的内在潜力和目前局限，各方都达成越来越广泛的共识，但对一些重要的、根本的问题缺乏定论，也持续困扰着大数据的潜在用户并阻碍进步。这些问题主要包括以下：

- 大数据解决方案属于什么性质？
- 大数据与传统的数据环境和相关应用有何不同？
- 大数据环境的本质特征是什么？
- 大数据环境如何与既有架构集成？
- 哪些关键科技和标准化方面的挑战亟需解决，从而加速大数据解决方案的部署实施？

在此背景下，2012 年 3 月 29 日，白宫宣布启动大数据研究和开发，该计划的目标包括加快科学和工程领域探索，巩固国家安全，并且通过提高从大量繁琐数据资源中分离提取精华的能力，来改革教学模式。

6 个联邦部门和它们的分支机构承诺至少出资 2 亿美元扶持至少 80 个项目，力图通过显著改善设备和技术，从海量数据中获取、整理并得出结论。同时，该计划鼓励产业界、研究型大学和非盈利组织一起联手联邦政府，最大限度地利用大数据带来的机遇。

受政府和民众双重意愿的推动，美国国家标准与技术研究院（NIST）接受挑战，联合各行业专业人士，确保大数据计划安全有效实施。2013 年 1 月 15 至 17 日，NIST 召开了“云和大数据论坛”，受大会启发，NIST 决定创建一个公共工作组，开发大数据互操作性框架。论坛与会者指出，该框架应当定义并区分大数据技术需要满足的需求，包括互操作性、可移植性、可重用性、可扩展性、数据使用、分析及技术架构。通过这些工作，该框架将促成最为安全有效的大数据方法和技术。

2013 年 6 月 19 日，NIST 大数据公共工作组（NBD-PWG）成立，全国各地工业界、学术界和各级政府纷纷加入。这个公共工作组将致力于打造工业、学术和政府利益共同体，旨在对大数据的定义、分类、安全参考架构、安全隐私需求和技术路线图形成共识，最终形成一个中立于供应商并在技术和基础设施方面独立的框架。基于此框架，大数据利益相关者能够运用最好的分析

工具，选择其最适合的计算平台和集群，处理问题或者解决可视化需求，大数据服务商也可从中挖掘增值机会。

《NIST 大数据互操作性框架草案》主要包括 7 卷内容：

- 第 1 卷：定义
- 第 2 卷：分类
- 第 3 卷：案例和总体需求
- 第 4 卷：安全和隐私
- 第 5 卷：架构白皮书调查
- 第 6 卷：参考架构
- 第 7 卷：标准路线图

1.2 定义与分类分组的范围和目标

本卷是由 NBD-PWG 定义与分类分组完成的，主要关注数据科学、参考架构和模式领域建立大数据的概念和定义相关的术语。

本卷的目的是为大数据相关的对象提供一个通用词库。对于管理者来说，本卷的术语可以帮助他们分清在大数据这一快速发展的领域需要理解的各种概念；对于采购人员来说，本文将提供用于讨论组织需求的框架，并帮助他们分清各种可行的方案的不同；对于市场人员来说，本文将提供推销解决方案和创新点的方法；对于技术社区来说，本卷将提供一种通用语言，以更好的区分不同的需求。

1.3 报告制作

大数据和数据科学已经成为包含很多概念的流行用语。为了更好地定义这些术语，NBD-PWG 定义与分类分组首先对这一混乱领域需要的各种概念进行梳理。然后对这两个最重要的术语（“大数据”和“数据科学”）以及他们包含的概念进行了澄清。

为了保持数据主题和数据系统的可管理性，分组尝试着将讨论限制在由于大数据的存在而带来的差异上，而那些扩展的主题，如数据类型或分类分析以及元数据，仅仅在出现了对大数据产生影响，或者带来问题的时候才进行讨论。当然，分组也确实引入了其他的需要用来理解新大数据方法论的主题。

术语的开发不依赖任何特定的工具或者执行方法以避免强调特定的执行方法，并且在本领域不可避免的发展的情况下保持术语的足够的通用性。

分组也注意到，一些领域，如法律领域，可能会使用与这里提供的定义不同的特定的语言。当前版本仅仅反映了本分组成员的知识范围。在评论期我们期望更多的参与者来支出在本文中提出的术语与各领域实际使用的不同。

1.4 报告结构

本卷期望能够澄清两个广泛应用的术语（大数据和数据科学）的意义，第 2 节中对这两个术语进行了讨论；第 3 节中对能够提供进一步信息的更多的基础概念和数据进行了深入的讨论；第

4 节则涉及到了几个更加详细的概念。《NIST 大数据互操作性框架：第 1 卷 定义》的第一版对在进行框架选择的时候能够确定分类或者功能性能力的一些基础的概念进行了描述。

与本文紧密相关的一些信息可以在《NIST 大数据互操作性框架》的其他卷中找到：《第 2 卷 分类》中提供了 NIST 大数据参考架构（NBDRA）的更详细的组件的描述；NIST 大数据参考架构（NBDRA）在《第 6 卷 参考架构》中描述；安全和隐私相关的概念在《第 4 卷 安全和隐私》中进行更详细的描述；为了理解这些系统是如何被组织起来满足用户的需求，读者可以参考《第 3 卷 案例和总体需求》；《第 7 卷 标准路线图》对第 1 卷到第 6 卷建立的框架进行了综述，并对 NBDRA 相关的标准工作进行了讨论。将本卷中的相关节进行比较，可以获得本卷与 NBD-PWG 的一致性的更深入的理解。

1.5 本卷后续工作

本卷仅仅体现了 NBD-PWG 起始阶段在为了制定规则和理清这一新兴的快速发展领域的工作成果。大数据包含了大范围的数据类型、研究领域、科技和技术，通过从不同的视角进行研究可以凝练出一个统一的、基础的定义集，然而，通过不同的视角的讨论，还可以得出对大数据的更广泛意义的理解。随着本领域的成熟，本文将需要引入本领域的更多创新的观点。为了确保本文的观点是正确的，未来 NBD-PWG 的任务还包括下面这些：

- 定义大数据源的不同通信模式，以更好的理清可采用的不同方式；
- 对第 1 卷进行更新，以包含其他工作组织，如国际标准化组织（ISO）第一联合技术委员会（JTC1）和事务处理执行委员会的工作成果；
- 改进对数据治理和拥有问题的讨论；
- 开发管理节；
- 开发安全和隐私节；
- 增加对数据价值的讨论。

2 大数据和数据科学定义

数据生成和存储的生长速度一直在呈现指数增长。在 1965 年的一篇论文中，Gordon Moore（戈登摩尔）估计在一个集成电路板上，晶体管的密度每两年增加一倍。这被称为“摩尔定律”，这样的增长速度已经被应用到从时钟速度到内存的计算的各个方面。数据量的增长速度为每十八个月增加一倍以上，比摩尔定律更快。这种数据爆炸现象为组合和使用数据来发现价值的新方法创造了机会，也因为被管理和分析数据的规模问题，带来了重大挑战。一个显著的变化在于非结构化数据的数量。历史上，结构化数据通常是大多数企业分析的重点，并且一直通过使用关系数据模型来处理。最近，诸如微文本、网页、关联数据、图像和视频等非结构化数据的总量激增，这种趋势表明会更多地结合非结构化数据去产生价值。大数据分析的主要好处在于具有处理大量不同类型信息的能力。大数据并不意味着现有数据量只是简单比原有数据量大，也不是比现有技术能够高效处理的数据量大。对更高性能或效率的需求在持续发展的基础上不断产生。然而，大数据代表了一个需要高效处理当前数据集的体系结构的根本性变化。

在数据系统的演化进程中，人们对于高经济效益以及高效率的数据的分析需求迫使现有技术变化多次。例如，可靠地处理数据的结构化转变的方法导致了向以关系代数为模型的数据存储范式的转变，与此同时数据模型开始向关系模型转变。这是一个在数据处理方面的根本性变化。因为关系数据模型不再能高效地处理当前所有对大量且通常是非结构化数据集的需求，所以当前被称为大数据的技术革命发生了。因为数据规模已经稳步增长了几十年，所以大数据革命并非只是数据量比原来更大的问题，而是体系结构上的一次性、根本性的转变，就像向关系模型的转变是一个一次性的变化那样。正如关系数据库在几十年里发展到了更高的效率，大数据技术也会持续不断地演进。其实，大数据的许多概念基础已经存在好多年了，只不过在最近十年里，它们快速发展成熟并且应用到大规模数据系统。

大数据这一术语已经被用于描述大量概念，在某种程度上这是因为几个不同的方面一直相互影响。为了理解这次革命，我们必须考虑如下四个方面的相互作用：数据集的特征、对数据集的分析、数据处理系统的性能以及对经济效益的商业考虑。

在本规范余下部分，大数据和数据科学这两个广义概念将会被分解成特定的个体术语及概念。

2.1 大数据定义

大数据指的是传统的数据体系结构无法高效处理新的数据集合的状况。推动新的体系结构的大数据的特征是：***数量***（即数据集的规模）、***多样性***（即来自多种数据仓库、领域或类型的数据）和数据的动态特征：***速度***（即：流量的速率）和***可变性***（即在其它特征里的变化）。这些特征——数量、多样性、速度和可变性——被俗称为大数据的几个“V”，我们将在第三部分对此进一步讨论。这些特征中的每一个都会影响到大数据系统的整体设计，并导致不同的数据系统体系结构或不同数据生命周期过程排序，来实现需要的高效性。

大数据*包含大量的数据集，这些数据集主要具有数量、多样性、速度和/或可变性等特征，并且需要一种能够高效存储、处理及分析数据的可扩展的体系结构。*

需要注意的是，上述定义包含了数据特征与对扩展来实现所需的性能和经济效益的系统体系结构的需求之间的相互作用。系统扩展有两种完全不同的方法，通常形象地描述为“垂直”和“水平”扩展。***垂直扩展***意味着通过增加处理速度、存储和内存等系统参数来获得更好的性能。

这种方法受到了物理性能的局限，摩尔定律中描述到实现物理性能的改进需要更多复杂的原件（硬件、软件等）来增加时间和成本。另一种方法是使用***水平扩展***，将集成的一组个体（通常是商品）资源作为一个单独的系统。水平扩展是大数据革命的关键。

大数据*范式由使用水平耦合的独立资源来实现对大量数据集的高效处理所需的可扩展性的数据系统的分布组成。*

上述新范式带来了大量的概念定义，这些定义表明当数据扩展导致数据管理成为系统体系结构设计中一个重要推动力的时候，便存在大数据问题。此定义没有明确地在大数据范式中提及水平扩展。

综上所述，大数据范式是一个从支持垂直扩展的单片系统（即为现有的机器增加更多的能力，例如更快的处理器或硬盘），到使用一组松散耦合的并行资源的水平扩展的并行系统（即向可用的集合里面添加更多的机器）的数据系统体系结构上的根本性转变。这种类型的并行化转换始于 20 年前的模拟社区，那时科学模拟开始使用大规模并行处理系统。

大规模并行处理*指的是多个处理器通过并行工作来实现一个特定*程序*的过程。*

计算机科学家们能够通过将代码和数据分开到单独的处理器的不同组合来极大地拓展模拟能力。然而，这也产生了这个领域里的很多并发问题，诸如在等待其他资源去完成计算任务时，信息传输、数据移动、使用不同资源的一致性延迟、负载平衡、系统低效等问题。

现有的大数据范式是相似的。数据系统需要一定程度的可扩展性，来与数据的规模相匹配。为了获得这种程度的可扩展性，我们需要使用不同机制去分配数据及使用松散耦合资源的数据检索处理过程。

实现对不同资源高效的可扩展性方法会持续不断地演进，但是范式的转变（类似先前模拟社区的转变）是一次性发生的。最终，在对一个处理过程或使用大量不同资源、并行工作的数据系统的分配的基础上，一个新的范式转变将可能产生。未来的革命需要用新的术语来描述。

大数据以如下的自引用观点为重点：因为数据需要可扩展的系统去处理，所以它是大规模的；相反地，因为有处理大数据的需求，所以产生了有更好扩展性的体系结构。我们很难去描述一个多大规模的数据集可以被称为大数据。如果可扩展的新体系结构的使用比传统垂直扩展体系结构的经济或性能更加高效（即如果传统单一平台的计算资源不能实现相似性能），数据通常被认为是大数据。这种数据特征与数据系统性能之间的循环作用关系在仅考虑其中一个方面的情况下，可以导致对大数据的不同定义。

大数据的一些定义集中于大数据特征所需的系统创新。

> ***大数据工程****包含一些先进技术，当数据集特征需要高效存储、处理和分析的新体系结构时，这些先进技术利用独立的资源来建立可扩展的数据系统。*

上述定义是耦合的，大数据工程会在数据特征需要的时候进行使用。不能被传统关系模型高效处理的数据集的作用日益显著，这种现象推动了在数据层面上新的工程技术的发展。对结构化数据可扩展的访问需求促进了以键值对范式为基础的软件的产生。文件分析重要性的提升促进了面向文档数据库范式的产生，关系数据日益提升的重要性促进了对面向图的数据存储的更加高效的使用。

新的非关系型数据库范式通常被称为 ***NoSQL*** （不是或不仅是结构化的查询语言[SQL]）系统，我们将会在第三部分对此进一步讨论。用 NoSQL 来定义大数据存储范式存在两个问题：首先，这种使用基于一种集合理论进行数据查询和检索的语言来描述数据存储方式的做法是不合适的；其次，相对于新的非关系型数据仓库，SQL 查询语言的应用能力一直在提高。尽管 NoSQL 这

一术语得到了广泛使用，并且它将继续指代在关系模型之上的新数据模型，但是我们仍然希望这一术语可以被一个更加合适的术语替代，因为将新的存储范式命名为和当前正在使用的查询语言相对立是不明智的。

> ***非关系模型****通常被称为 NoSQL ，指的是用于存储及处理数据的、不遵循关系代数的逻辑数据模型。*

另一个相关的工程技术是与大数据的多样性特征有关的联合数据库系统。

> *一个****联合数据库系统****是一种元数据库管理系统，它透明地将多个自治的数据库系统映射到一个单独的联合数据库。*

一个联合数据库是由一些基本数据库系统组成的数据库系统。大数据系统也可以从许多资源里获取多样的数据，但是底层仓库不需要都遵循关系模型。

需要注意的是，对于系统和分析处理过程来说，大数据范式的转变也导致了传统数据生命周期处理过程的一些变化。一个对端到端的数据生命周期的描述将处理过程分成收集、准备、分析、作用四个步骤。不同的大数据用例以数据集特征以及端到端数据生命周期的时间窗作为其特点。数据集特征通过不同方式改变着数据生命周期的处理过程，比如通过改变生命周期过程中数据被放置在永久存储器上的时刻。在一个传统的关系模型中，数据在准备阶段（例如提取-转换-负载和清理过程）之后被永久存储起来。在一个高速度数据的使用案例里，我们准备和分析数据来做预警，然后将数据（或数据集合）永久存储。在一个大规模数据的使用案例里，在被清理掉及组织起来之前，数据通常以一种未加工的状态被存储在生成它的地方（通常被称为提取-负载-变换）。用未加工的形式来持久化数据带来的结果是仅在准备和分析阶段检索数据时使用到数据模式或模型，这种大数据概念被称为读时模式。

> ***读时模式****是指从数据库读取数据时，数据模式在诸如转换、清理、整合的准备阶段的应用。*

大数据的另一个概念通常是指*将处理过程移动到数据，而不是将数据移动到处理过程。*

计算的可移植性是计算向数据位置移动的能力。

上述定义指出，数据分布太广泛，以至不能被查询及移动到另一个资源去做分析，因此相反地将分析程序分配给持有数据的资源，最终结果聚集在一个远程资源上。实际上，数据局部性的概念是并行数据结构的一个重要方面。其他的系统概念是互操作性（各种工具协同工作的能力）、可重用性（将工具从一个领域应用到另一个领域的能力）、可扩展性（为新领域添加或修改现有工具的能力）。这些系统概念并非专门针对大数据，但是我们可以在大数据参考体系结构的检测中理解它们在大数据中的存在。我们在 *NIST 大数据互操作性框架：卷 6，参考体系结构*中讨论了这一系列问题。

大数据的其他概念涉及到数据分析的变化，我们将在 2.2 部分对此进一步讨论。我们也使用了许多其他术语（尤其是以字母 V 开头的术语），其中有几个涉及到数据科学处理及其利益的方面，但它们不是大数据的新特征。这些术语包含***真实性***（即数据的准确性）、**价值**（即分析组织结构的价值）、**波动性**（即数据结构随时间变化的趋势）和**有效性**（即数据用于其预期用途的适当性）。虽然这些特征以及其他一些特征——包括质量控制、元数据和数据源——比大数据更早出现，但是他们对大数据系统的影响仍然是很重要的。我们将在 3.4 部分对几个关于大数据分析的术语进一步讨论。

实质上，大数据指的是数据仓库和使用不同资源并行工作的数据处理程序的可延展性，计算密集型仿真社区二十年前也同样包含了大规模并行处理。通过使用资源通信的方法，相同的可扩展性现在可以适用于数据密集型的应用程序。

2.2 数据科学定义

从字面的意思看，数据科学是在理论科学、实验科学和计算科学之后科学的第四范式。第四范式是由吉姆•格雷在 2007 年提出的一个术语。它将数据分析的实施作为一种经验科学，能够直接对数据本身进行学习获取分析结果。数据科学作为一种范式可包含假设的设立、用于说明该假设的数据集（新的数据或已经存在的数据）以及经过分析后对假设的确认或否定（或者确定是否还需要额外的信息或学习）。在很多数据科学项目中，未经处理的数据是首先被浏览的，它会触发一个假设，然后这个假设会被探讨。同其他的实验科学一样，最终的结果可能是原先的假设本身需要被重新定义。数据科学的关键概念是经验科学，它对数据直接实施科学化的分析流程。假设的设立可由业务需求驱动，也可来自于依据某个技术上的假设对业务的重新陈述。

数据科学范式*是经过发现、假设和假设测试的流程从数据中提取可用的知识的过程。*

数据科学可理解为在系统架构中的流程层发生的一系列行为，它可对存储在数据层中的数据进行操作，从而通过完整的数据生命周期从未经处理的数据中提取知识。

数据生命周期*是将未经处理的数据转化为可用的知识的一组流程。*

传统的数据生命周期包含收集、准备、分析和作用 4 个步骤，分析作为术语通常用于指数据生命周期中分析这一阶段。

分析*是从信息中综合推理出知识的过程。*

在新的大数据范式中，分析不再与数据模型以及分布于并行资源中的数据分割存在。当结构化数据通过关系模型被专门存储为组织好的信息时，我们可以设计一个面向这一结构的分析过程。然而，当数据科学范式的定义值得是直接从数据中学习，在大数据范式中学习的过程可隐式的与数据生命周期的各个阶段关联，而分析只是其中的一个子集。

数据科学*是依靠经验推理从未经处理的数据中经过完整的数据生命周期过程获取可用的知识的科学。*

数据科学横跨整个数据生命周期，交叉运用不同学科和领域的原理、技术和方法，包括数学分析领域、数据挖掘（如机器学习和模式识别）、统计学、运筹学和可视化，以及系统、软件和网络工程等领域。数据科学家和数据科学团队需要与一个或多个上述领域的资深专家共同解决复杂的数据问题，这些问题通常与业务策略的上下文关联，或者需要某个专业领域的知识。另外，个人在交流、表述和学习方面的能力对于处理大数据系统的复杂交互同样非常重要。

数据科学家*是数据科学方面的专业人员，具有足够的知识处理复杂的业务需求，且掌握领域的知识，具有分析的技能，且能够通过软件和系统工程在数据生命周期的各个阶段管理端对端的数据流程。*

当上述技能被数据科学家掌握时，团队中的成员的能力也会覆盖到图 1 中的各个方面。

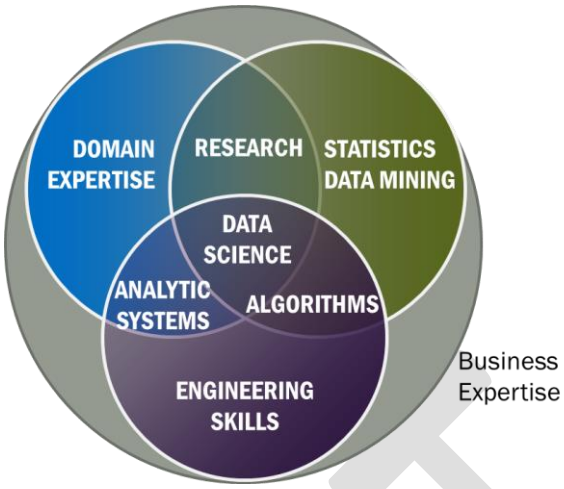

图1. 数据科学所需的能力

数据科学并非只涉及到分析，同样也涉及端对端的实验周期，对于数据科学来说，数据系统就是科学设备。数据科学家须了解数据的来源及出处，数据转换过程的合理性和准确性，转换算法和流程之间的相互关系，以及数据存储机制。数据科学家是总揽全局的角色，他应确保数据生命周期的每个阶段所执行的操作能够对证明假设产生意义。分析的概念将会在 3.4 节中做进一步讨论。

数据科学正越来越多的被用于对业务产生影响的决策中。在大数据系统中，证明某个相关性通常就足够用于业务的决策。例如，如果能够证明在网站中使用蓝色能够比使用绿色带来更大的销量，那么这一相关性就可以用于业务。而在这个例子中，不许要论证客户对于颜色偏好的理由，发现这一相关性就足矣。

正在数据科学社区中讨论的众多话题中，包含如下两个话题：数据采样，以及更多的数据优于更好的算法这一个观点。

数据采样是统计学的一个核心概念，它讨论的是从一个大规模数据集中选取一个子集的问题。这个子集能够在某种程度上代表整个数据集，可在点对方法进行试验的过程中作为分析流程的输入，或者用于描述问题。例如，数据采用可用于计算决定试验过程产出所需的数据（如在临床药物测试过程中）

当数据挖掘社区兴起时，讨论的重点通常在于如何将已有的数据用于其他目的（即，被用于训练模型的数据是从大规模的数据集中采样而来，而这些数据原本被采集用于其他目的）。在这方面，应确保分析不会趋向于过度拟合（即，分析的模式能够匹配数据采样，但在作用于整个数据集时却会失效），而这一过程却常常被忽视。新的大数据范式并不需要从全局数据中进行数据采样，因为大数据系统在理论上可以在没有性能损失的情况下处理整个数据集。然而，即便使用了全部的可用数据，它也只能代表一部分个体的行为，这些数据也是由这部分个体产生的，而这些数据并非全部个体的真实关注。例如，对 Twitter 的数据进行研究，分析人群的行为习惯，分析的结果并不能代表全部人群，因为并不是人人都用 Twitter。如果只有较少的采样用于数据科学流程中，我们在描述业务问题时需要考虑到隐式的数据采样。

更多的数据优于更好的算法这一断言暗示了更好的结果可以通过分析更大的数据采样来得到，而不需要对算法进行优化。这一断言的核心在于少量错误数据元素素可能不会影响到整个分

析的结果。如果分析所需要的仅仅是关联关系而非因果关系，那么这一断言较容易证明。如果分析是对行为趋势的判断，分析过程对数据的质量要求较高。

为了更好的对分析进行描述，分析行为可分解为不同的类别，包括发现、探索性分析、关联性分析、预测建模以及机器学习。这些分析类别并不限于大数据，一些类别的分析更多的出现在数据科学的应用中。

数据可续与大数据的关系紧密，它指的是对数据的管理以及整个数据处理的执行过程，包括数据系统的行为。因此，数据科学包含了分析，但分析并无法包含数据科学。

2.3 大数据的其它定义

为了帮助理解大数据这一新领域的内涵，已经有一些大数据的定义被建议使用。前面章节中已经讨论过一些大数据的概念，分别摘自一些博客文章。这些正式的和非正式的概念，在一定程度上已经被接受为大数据的术语。表 1 列举了大数据的概念和定义的一些例子。NIST 大数据公共工作组的定义非常类似高德纳公司给出的定义，同时强调横向规模化是提供成本效益的因素。表 1 列出的大数据的概念和定义并不是全面的，而是旨在把它们与大数据这样一个笼统术语的内在关系描绘出来。

表一：大数据概念的举例

概念	作者	定义
4V 特征 (数量大, 种类多，高速，易变) 和工程学	高德纳	“大数据是数量巨大、高速产生、种类众多的信息集合，需要通过高效、创新的信息处理方法，来达到提高洞察力和决策制定能力的目的。”
数量大	Techtarget	“虽然大数据并没有指明具体的数量，但是当谈到 PB 级和 EB 级的数据时这个术语经常被使用。”
	牛津英语词典（OED）	“大数据，名词。超大规模的计算数据，通常在某种程度上对它的操作和管理会对逻辑带来重要挑战；（也指）涉及这些数据的计算分支。”
更大的数据	安妮特 格雷纳	“大数据是数据，因为其各种不同的数据规模，所以需要足够的观察来进行不同寻常的处理，虽然从一个规范到另一个规范以及时间上的变化并不是经常发生。”
不仅仅是数据量	昆廷 哈代	“大数据中的大不仅仅指数据库的规模，而是指我们拥有大量的数据来源，例如数字传感器和在世界范围内迁移的行为追踪器。”
	克里斯 诺伊曼	“...我们最初的定义是一个能够存储 10TB 的或者更多数据的系统...随着时间的推移，数据的多样性开始在这些系统中变的越来越常见（尤其是混合结构化和非结构化的数据的需求），这些造成了把“3V 特征”(数据量大，高速和种类多)作为大数据定义的广泛采用。
大数据工程	IDC[11] [16]	“大数据技术描述了一种新一代的技术和体系结构，被设计用来通过快速获取、发现和分析，从而达到经济地从非常大量的各种数据中获取价值的目的。”

概念	作者	定义
	哈尔·范里安	“大数据意味着这种数据不能够简单地被放进标准的关系型数据库里。”
	麦肯锡	“大数据是一种数据集合，其规模超出了传统数据库软件工具的捕获、存储、管理和分析的能力。”
更少的取样	约翰·福尔曼	“大数据是当你的企业想使用数据来解决一个问题、回答一个疑问、生产一个产品等情况下，…，不用简单抽样或数据变换就能充分利用数据而制定出一个解决方案。”
	皮特·斯科莫洛奇	“大数据最初描述的是消费互联网行业中把算法运用到不断变大的异构数据上来解决那些在小型数据集上有次优解的一种做法。”
新的数据类型	汤姆·迪文波特	“在过去的十年左右的时间内，大量的、新的数据类型已经出现了。”
	马克·冯·里吉门纳姆	“大数据不仅仅在于量大，更在于把不同的数据集合整合在一起来，进行实时分析，进而达到为组织获得洞察力的目的。因而，大数据的正确定义应该是：混合的数据。”
分析	瑞安·斯万斯特洛姆	“以前大数据意味着单个机器不能够处理的数据。现在任何与数据分析或可视化相关的东西都在使用大数据这个流行词。”
数据科学	乔尔·古林	“大数据描述了这样一些数据集合，它们是如此的巨大、复杂、或者快速变化，以至于突破了我们的分析能力的极限。”
	约什·弗格森	“大数据是一个用来描述由于生活方方面面的数据变得可用而带给我们的机遇和挑战的广义词。但它不仅仅是数据，还包括把数据变得有意义的人、过程和分析。”
价值	哈伦·哈里斯	“对我来说，‘大数据’描述了这样一种情形：一个组织能够（可论证地）使用到需要重构、理解和模拟的他们所关注的某一部分世界。”
	杰西卡·柯克帕特里克	“大数据指的是用复杂的数据集合来推动一个公司或机构的焦点、方向和决策制定。”
	希拉里·梅森	“大数据就是指一种收集信息和查询信息的能力，从而使得我们能够学到以前无法获得的关于这个世界的一些东西。”
	格雷格瑞·皮亚特斯基·夏皮罗	“我见过的最好的定义是‘当数据的规模成为问题的一部分时，数据就是大数据’。但是，这个定义仅仅谈到了规模。现在流行词‘大数据’指的是商业、科学和技术领域的数据驱动新范式，在这种范式下，巨大的数据规模和范围能够带来更好的服务，产品和平台。”
文化上的改变	德鲁·康威	“大数据，起源于分布式计算领域里的一种技术创新，现如今是一种文化运动，通过它我们能持续探索人类如何与这个世界进行大规模交互，以及人类之间、世界之间的大规模互动。”
	丹尼尔·吉利克	“‘大数据’代表了一种文化变迁，越来越多的决策是运用算法对记录的不变的证据进行透明逻辑和运算得到的。我认为‘大’更多的指的是这种变化的普遍性而不是数据的特殊数量。”

概念	作者	定义
	凯茜·奥尼尔	“‘大数据’有很多含义，一个重要的方面是它作为修辞手法的用途，它能够被用来欺骗、误导或者炒作。

3 大数据特征

第 2 章讨论的大数据概念差异在第 3 章大数据特性的讨论中也有相似的反映。第 3 章将会讨论一些大数据的术语和概念，以便于理解在已有的数据架构和分析上下文中大数据范例所带来的新观点。

3.1 数据元素素和元数据

个人数据元素素并没有随大数据发生变化，本文并不会对其进行详细地讨论。如果想了解与数据类型相关的其他信息，读者可以查看 ISO 标准 ISO/IEC 11404:2007 通用目的的数据类型，作为示例，ISO 21090:2011 健康信息将该标准扩展成了保健信息数据类型。

大数据非常重要的一个概念就是元数据，元数据通常也被描述为“数据的数据”。元数据描述了与数据有关的其他信息，例如如何以及何时收集数据、怎样处理数据。元数据本身应该被看作是包含跟踪、变更管理和安全需要的所有信息的数据。人们正在制定很多与元数据相关的标准，这些标准可用于元数据、通用元数据（例如 ISO/IEC 11179-x）、特定学科元数据（例如针对地理空间数据的 ISO 19115-x）。

描述一个数据集历史的元数据称为它的起源，该部分将会在 3.6 中讨论。随着开放数据（数据对其他人可用）和链接数据（与其他数据连接在一起的数据）成为标准，有信息描述数据是如何被收集、转换和处理的变得越来越重要。当改变原始收集过程中数据的用途以便于从中提取额外价值的时候，起源类型的元数据可以指导用户纠正数据的使用。

语义元数据是另一种类型的元数据，它指的是为了合理解释一个数据元素素而提供的定义性描述。一个本体（ontology）可以被概念化成一个图模型，表示实体之间的一种语义关系。本体是语义模型，它必须服从不同级别的逻辑模型。本体和语义模型早于大数据，本文档并不会对此进行详细地讨论。本体在本质上可以非常通用，也可以特定于某个领域。已经有很多机制可以用来实现这些唯一的定义性描述，读者可以参考万维网联盟（W3C）在语义网络方面的努力。在新的大数据范例中语义数据是很重要的，因为语义网络表示一个试图为术语提供切面含义的大数据。需要再次重申的是，语义元数据对于链接数据的努力非常重要。

分类（Taxonomies）表示与数据元素素关系相关的一些元数据。分类是实体间的层级关系，在这种关系中数据元素素会被分解成更小的组件部分。虽然这些概念非常重要，但是它们早于大数据范式转变。

3.2 数据记录与非关系模型

数据元素素被以某种特殊的形式记录了下来，如报告、事件或传输。以前，商务系统中的大部分数据都是结构化数据，此类数据所存储的记录都是结构化的且可以高效的匹配于关系型结构模型中。这些记录可以理解为是表中的每一行，而数据元素素就是其中的一个单元。非结构化数

据，如文本、图片、视频及关系数据，一直以来在数量上都保持着尤为明显的增长势头。虽然现代关系型数据库逐渐可以支持此类数据元素素，如分析、索引和处理，但此类方式的限制与使用在非标准的 SQL 扩展中是矛盾的。对于非结构化及半结构化数据的分析需求实际已经非常迫切，然而，大数据（框架）模式的提出其实更强调非结构化或关系型数据，当然不同的工程算法，处理方式也会更高效便捷。

其次，大数据工程中涉及到新的数据存储记录方式被称作语义元数据。在某些情况下，（数据）记录仍沿用（数据）表格结构的概念。一种存储范例即键值结构，记录包含了一个关键字及一串数据共同存放于一个值中，该条数据通过关键字来检索，同时非关系数据库软件用于控制对值中数据的访问。这可被看作是一种关系型数据库表的子集或简化，它通过单一索引域或列实现。这是一种文件存储的变形，文件具有不同的值空间（/域），每一个值空间（/域）都可以当作一个索引或关键字。与关系型表格的不同便是这一系列的文件不必具备相同的值空间（/域）。

另一种大数据记录存储类型是图形化模式。图形化模式代表了不同数据元素素之间的关系。数据元素素即节点，其间关系就是各节点之间的连接。图形存储模式将每一个数据元素素都呈现出一组主语、谓语、对象结构。通常情况下，对象与关系就是通过前文中提及的本体所展现。

再一种数据元素素关系是大数据模型中由来已久的概念之一：数据元素素间的“复杂性”。例如，在某些系统中，数据元素素一旦脱离其他相关数据元素素就无法被解析。这不难理解，好比人类基因序列，每个元素、它所处的位置、及其他相邻元素的关系无疑都是至关重要的。这种复杂性往往被归咎于大数据，但是它其实是数据元素素之间的或是跨数据记录的，不受数据集是否具有大数据特点的影响。

3.3 数据集特征与存储

特征抽取基本要求的总体要求如下：数据记录被分组组织成具有大数据 4V 特征的数据集。数据集的特征可以参照数据本身特性，或静态数据特性；在网络中传输的数据，或暂驻计算机内存可被读取或更新的数据，其特征可参照动态数据。动态数据将在 3.4 部分讨论。

静态数据：大数据时代，静态数据具有与传统显著不同的典型特征：体量大、类型多（volume、variety）。体量大是大数据背景下静态数据的重要特征。估计显示，全球数据量每两年翻一番。如果这种趋势得以延续，到 2020 年全球数据总量将是 2011 年全球数据总量的 500 倍，如此可见，数据无疑将是海量的。如 2.1 部分所介绍，数据量的指数级增长，催生了伸缩存储的新模式，实现了水平扩展资源的无缝集成。

静态数据的第二个特征是复杂数据资源使用需求的快速增长，复杂数据涉及领域众多、类型多样。传统上，复杂数据通过转换或预分析等处理，提取到与其他数据集成的数据特征。数据格式、逻辑模型、时间尺度和语义等分析中必定要处理的特征的多样化使该类数据集成变得更加复杂。如，从社会网络、图像或传感源原生数据等复杂数据对象中提取数据集成所需的文本。为处理繁多的数据格式，联邦数据库模型作为新型的数据库系统，实现了跨越底层数据库系统的整体控制和协同操作。对于迁移集成十分困难的大体量数据，以及那些创建数据系统的组织机构所不能控制的数据，复杂大数据强烈需求一系列新的大数据引擎方案，高效、自动实现多格式、多逻辑模型的跨库集成。

大数据引擎催生了新的数据存储模型，其处理非结构化数据比传统关系模型更加高效，也产生了数据集成机制方面的一些衍生问题。新的（水平）扩展技术通过将大数据散布到大量的便宜资源中提升管理和操控能力，而不是存储在传统的昂贵、垂直扩展性能良好的系统中。例如，针对分析需要，专门开发了主要存储和索引公共存储中异构数据的文档存储。下面讨论面向数据记录的新型非关系存储。

共用磁盘文件系统：SAN 网络存储（Storage Area Networks）和 NAS 网络附属存储（Network Attached Storage）等技术采用支持多计算资源访问的单存储池方案，解决了多节点同步访问大体量数据集的许多关键技术，但仍存在一些问题，如数据锁定、数据更新，以及从单个输入/输出操作访问公共存储池时产生的性能瓶颈，限制了面向大数据应用程序的扩展能力。这些问题在完全分布式文件系统实施中全部被解决。

分布式文件系统：分布式文件存储系统中，多结构化(对象)数据集分布于服务器集群的各个计算节点，以文件/数据集或者更常见的块粒度存储，支持集群中的多个节点同时操作一个大文件/数据集的不同内容。设计时，大数据框架常常充分利用节点数据的本地化优势，避免分布处理时跨节点的数据移动。此外，许多分布式文件系统还支持文件/块粒度的复制，所以为了可靠性与恢复(保证数据不会因一个节点的故障而丢失)，以及增强数据本地化，每个数据对象均在不同的机器上多次备份。任何类型的数据，任何体量的文件，均可采用一些高性能的大文件关键技术完成处理，而不需要借助于正式提取、转化或加载转换等操作。

分布式计算：分布式计算的主流框架包括存储层，以及实现了多级与算法编程模型的处理层组合体。低成本服务器基于分布式文件系统存储数据，可以显著降低大规模数据(如：网页索引)计算的存储成本。对于分布式数据，MapReduce 已成为其计算模式的默认处理组件，处理结果常常被加载到分析环境。

廉价服务器主要适用于低速、批量处理的大数据应用，但不具有低延迟的高性能。计算时，MapReduce 基本处理应用限制了数据的更新或迭代访问。对于重复更新的需求，可以使用批量同步并行系统或新版 MapReduce。MapReduce 的改进和泛化已经实现了旧版本所缺乏内容的开发，包括容错、弹性迭代，中间层消歧和高效查询。

资源协商： 常见的分布式计算系统几乎没有内置的数据管理功能。当前，已经发展了一些提供必要管理功能的技术，以支持业务管理、工作流集成、安全和治理等。在资源管理发展中出现了一些特别重要的成果，如支持附加的处理模型(除了 MapReduce)等的新特征，以及多租户环境、高可用性、低延迟应用程序的约束等。

典型实现中，资源管理器是若干节点管理器的核心（hub）。资源管理器收到客户机或用户的访问后，依次向一个或者几个节点管理器中的应用程序宿主（master）发出请求。第二个客户端也可发出自己的请求，提交给在同一个节点或者其他节点管理器的其他应用程序管理宿主。基于可用的 CPU 和内存，确定任务优先级，并在节点内调配适当的处理资源。

数据移动通常是由传输和应用程序编程接口(API)技术实现，而不是资源管理器。少数情况下，点对点(P2P)通信协议可以实现网络上大规模的文件传播或迁移，技术上讲 P2P 网络也是分布式文件系统。最大的社交网络，也是大数据的主要用户，采用 P2P 技术实现了大量内部机器之间二进制大对象(BLOBs)的移动，单个对象文件超过了 1GB。内部应用案例还显示，该技术已经应用到私人文件同步， 两个终端通过该系统关联后即可实现本地文件夹的自动更新。

外部使用时， P2P 系统的每个终端为数据移动提供带宽，因此该技术成为支持最大并发用户共同利用文档的最快方式。例如，NASA(美国国家航空航天局)使用该技术向公众开放 3GB 的图像。基于该技术，任何大体量的数据(如，视频、科学数据)都可以快速分发，而只需较低的带宽成本。

大数据还有许多其他正在快速发展变化的内容，在此不展开讨论，如集群管理、管理非关系模型数据的集群资源的其它通信机制。其他工业出版物中出现了新型软件定义存储（software defined storage）的多层存储 (例如内存，缓存，固态硬盘，硬盘，网络驱动器)使用的讨论内容。软件定义存储使用软件来确定存储层的动态分配，降低存储成本，同时保持所需的数据检索性能。

3.4 动态数据

文本特征抽取基本要求如下：大数据的一个重要特性是可以进行分析的时间窗口。动态数据采用实时，或接近实时地处理和分析，并且必须采用同静态数据（即持久化数据）非常不同的方式进行处理。动态数据往往类似于事件处理架构，并专注于实时或者可操作的智能应用。

在大数据时代，动态数据的典型特征是速度（velocity）和可变性（variability）。速度是指流中的数据被创建，存储，分析和可视化的速率。大数据速度意味着大量数据被在很短的时间量被处理。在大数据时代，数据被创建并实时地或接近实时地传递。增加数据流率创建了新的挑战，以使用实时或近实时数据。传统上这种概念已经被描述为流数据。虽然对某些行业来说这些方面是新概念，其他行业（例如电信业）已经处理了多年高容量和短间隔数据。然而新的并行换算方法确实为有效处理这些数据增加了新的大数据工程方案。

动态数据的第二个特征是可变性，它指的是在数据中任何随时间的变化，包括流量，格式，或组合物。鉴于在给定的时间里许多数据处理产生的数据量激增，需要新的技术来有效地处理这些数据。数据处理往往跟云环境中额外虚拟资源的自动配置进行捆绑。在其他关注于运营云架构的行业出版物中可以找到处理数据的技术的详细讨论。早期的大数据系统由互联网搜索提供商建立，其他系统经常部署在裸机上来实现在集群和多个存储设备分配 I/O 的最佳效率。尽管云（即虚拟化）的基础设施经常被用于大数据部署的测试和原型，最近的趋势是，为了提升产品解决方案的 I/O 虚拟化基础设施效率，部署在云或者 IaaS 平台。

具有高可变性的高速系统可以部署在云基础设施上，因为其具有能够添加或删除节点处理峰值性能的成本和性能效率。当不再需要提供显著成本节约用于操作这种类型的大数据系统时，能够释放这些资源。在非常大的实现方案和某些情况下，云供应商正在物理硬件之上实施该相同类型的弹性基础设施。这对于已经需要大量的基础设施，而只是需要通过可以改变应用负载来平衡资源的组织尤其如此。

3.5 大数据的数据科学生命周期模型

根据章节 2.1 的介绍，数据生命周期由下面 4 个阶段组成：

1. ***收集：***在这个阶段数据以原始形式（即行数据）收集和存储；
2. ***准备：***这个阶段包括将原始数据转换成清洁，有组织的信息的收集流程；
3. ***分析：***这个阶段涉及将有组织的信息生成综合知识的技术；
4. ***作用：***这个阶段涉及使用综合知识为企业创造价值的过程。

在传统的数据仓库，数据处理过程遵从以上的顺序（即收集，准备，存储和分析）。关系型模型以优化预期分析的方式来设计。因为大数据的不同特性影响了数据处理过程的顺序。这些变化的例子如下：

- ***数据仓库：***持久性存储发生在数据准备阶段之后
- ***大数据数量系统：***数据在准备阶段前立即以原始形式存储；准备阶段发生在读出中，并且被称为“在读之上的模式”
- ***大数据速度应用程序：***收集，准备和分析（预警）在动态改变中发生，并可能包括在存储之前进行概括或聚合

正如仿真器跨集群处理器拆分了分析处理，数据处理流程被重新设计以跨数据接点拆分数据转换。因为数据可能太大而移动不了，转换代码可以跨数据持久节点被并行发送，而不是数据被提取并送至转换服务器。

3.6 大数据分析

分析过程经常定义为对初始假设描述的探索，对建立一个分析特定假设过程的开发，和把分析封装到可操作系统中的应用。虽然大数据已经涉及所有三种类型的分析过程，但开发和应用分析大部分的变化来自于开发和应用分析。新的大数据工程技术改变了可能的分析类型，但不会导致完全新型的分析。不过，如果数据检索的速度足够快，分析师便可以以过去不可行的方式与数据交互。传统的统计分析技术在分析之前会缩减、采样或者汇总数据。这么做是为了使得分析硬件无法承载的大规模数据集成为可能。大数据分析经常侧重在数据总体做运算的价值，这样分析才能有更多的机会来判断因果关系，而不仅仅是相关性。但是，当我们知晓某些趋势或者方向到足够以采取行动时，相关性仍然是有用的。今天，大多数的统计分析和数据挖掘注重因果关系，能够描述为什么事情正在发生。发现原因有助于角色改变趋势或结果。角色，这在系统开发中可以代表个人、组织、软件或硬件等，讨论参见《NIST 大数据互操作性框架：第 2 卷，分类》。大数据解决方案，使得对大型、复杂、异构数据实施因果关系类型的复杂分析成为可能。

除了数据量(volume)，速度(velocity)，数据类型(variety)，以及数据可变性(variability)外，有许多以 V 开头的单词已被用于描述系统架构方面的大数据的需求。这些术语和数据分析非常相关。准确性(Veracity)和数据来源（provenance）就是这样的两个术语，将在下面讨论。

准确性(Veracity)是指数据的完整性和精度，并用来描述由来已久的数据质量问题，用白话讲就是“垃圾进，垃圾出”。如果是因果关系分析，那么每个数据元素素的质量是极为重要的。如果是基于大规模数据集的相关性分析或者趋势分析，那么个别坏的数据元素素会在总体的统计中丢失，趋势仍旧是准确的。正如在 2.2 节说的那样，很多人讨论“是不是更全面的数据优于更好的算法”，但是这个话题在其他地方讨论更好。

在 3.1 节已经讨论过，数据来源，或者说数据的历史，正逐渐成为大数据分析中的必要要素，因为越来越多的数据会被改变原来的用途，用以和当初创建数据时完全不同机制的新类型分析。由于数据的使用远远超出了数据产生者的控制，可访问和数据相关的整个创建和处理历史的元数据变得越来越重要。此外，至关重要的是要知道什么分析会产生数据，因为置信度范围、误差范围以及精度/召回率的限制都是和分析的输出相关的。

另外分析中需考虑的是分析过程和承担做出可操作洞察的人或者进程之间的交互速度。分析数据处理速度会由于数据处理的连贯程度由批处理转向流处理而下降。虽然数据处理的连贯程度问题在大数据时代来临之前就已经出现了，预期的连贯程度是选择体系结构和工具组件的重要因素。基于由集群扩展导致的大数据内更大的查询和分析速度，互动（即实时）处理变得越来越重要。快速分析周期允许分析师在数据上做探索性发现，在任意时间框架内浏览相比其它可能更多的数据空间。数据处理的连贯程度的更深入讨论参见《NIST 大数据互操作性框架：第 6 卷，参考架构》。

3.7 大数据指标和基准

大数据工程使用的初始条件包括确定具体情况下的数据量级门槛，当数据量超过该门槛值时数据才被认定为大数据。在确定数据量级门槛值时，必须考虑到多方面因素，每个应用可能都会

对应一个不同的结果值。正如在本文 2.1 节中所述，大数据特征导致大数据工程技术的应用，让数据系统可以高效率、低成本的运行。对于一个具体的应用来说，能否获得一定的性能和成本效率需要进行设计分析，这已超出本报告的范围。

对大数据指标和基准的一个重要需求是为大数据系统提供性能标准。事务处理性能委员会 TCP-xHD 大数据委员会正在着手解决这一问题，他们的成果和相关信息将会被收录到本报告的后续版本中。

3.8 大数据安全和隐私

《NIST 大数据互操作性框架：第 4 卷，安全和隐私》，讨论了大数据的安全和隐私。

虽然关于大数据公开的讨论都聚焦于数据量大（海量性）、处理速度快（高速性）、数据类型多（多样性）等三个特点，然而大数据还有一些其他的重要特征同样会影响到安全性和隐私，例如大数据的准确性和波动性。

《NIST 大数据互操作性框架：第 4 卷，安全和隐私》对大数据各项特征的影响进行了详细描述，一些关键点概括如下：

- 多样性：把传统关系型数据库的安全性迁移到非关系型数据库一直是一个挑战。大数据多样性所引起的一个突发问题已得到重视，即通过关联无关的公共数据库，就能够从匿名数据集中推断出身份。
- 海量性：大数据的海量性使存储必然要处于多层存储介质中。数据在各层之间移动，已经产生了对威胁模型作系统分析的需求，以及研究和开发新技术的需求。
- 高速性：和非关系型数据库一样，诸如Hadoop等分布式编程框架在发展过程中并未把安全性作为一个主要目标。
- 准确性（真实性）：当数据跨越个人边界到群组、兴趣社区、州际、国家甚至全球范围时，在保护数据的完整性和维护隐私策略方面已经遇到复杂的挑战。
- 波动性：根据负责大数据收集、处理、汇总、存储角色的时效性，安全和隐私的需求要能够随之改变。在负责的组织合并甚至消失时，数据治理也能随之改变。

3.9 数据治理

数据治理是数据和数据系统管理的基本要素。

数据治理是指围绕数据管理，对数据进行的形式化、条理化（例如，行为模式）工作。

数据治理的定义包括数据全生命周期管理，包含数据处于静态、动态、未完成状态或交易状态下的管理。为了最大化数据治理产生的效益，数据治理应考虑各年龄段人群、个体经营者以及企业的隐私和安全问题。在全球互联网大数据经济背景下，数据治理被用来解决很多重要问题。例如，许多企业提供数据托管平台来管理系统用户产生的数据。虽然从数据托管公司角度看，治理的策略和流程很常规，但是治理和数据的控制权对于数据提供者来说却是需要面对的新问题。这些问题包括：用户是否还能自己掌控数据，还是被数据托管公司管理？数据生产者是否可以删除数据？他们是否能够控制别人获得查看数据的权限？

治理问题与双方利益相关，其中一方（例如，数据托管公司）希望通过数据获取利益，而数据提供方却希望保留获取自身利益的权利。在大数据模式下产生的新的数据治理问题，将在本标准的下一版本中进行深入讨论。