数据科学导论课文献阅读

1. Big Data Quality: A Survey

ABSTRACT：（摘要）

随着通信技术的进步以及生成、收集和存储的大量数据，以高效且经济高效的方式管理这些数据洪流的质量变得至关重要。存储、处理、隐私和分析是大数据面临挑战的主要关键因素，需要进行质量评估和监控。大数据社区已将质量视为其成熟度的一个重要方面。然而，这是一项至关重要的做法，应该在其生命周期的早期阶段实施，并逐步应用于其他关键流程。我们越早纳入质量，我们就能从洞察中获得全部好处。在本文中，我们首先确定了需要质量评估的关键挑战。然后，我们调查、分类和讨论大数据管理的最新工作。因此，我们提出了一个全面的质量管理框架，描述了通过不同的大数据阶段进行的关键质量评估实践。该框架可用于利用质量管理，并为数据科学家提供路线图，以更好地了解质量实践并强调管理质量的重要性。最后，我们总结了这篇论文，并指出了大数据质量的一些未来研究方向。

（1）INTRODUCTION（引言）

大数据 （BD） 已成为学术界和工业界非常有吸引力的研发领域。随着宽带互联网的普及和最近出现的大量服务（VOD、云存储和服务、数据集群），每天都会产生大量数据，强化了大数据时代。许多 IT 专业人员、研究人员、科学家和公司都在努力定义、描述和分析新的挑战以及可能用于应对这些挑战的技术和方法。通过探索现有技术和平台，数据科学家正在处理和分析这些海量数据，以产生可能对社会和人类福祉产生重大影响的相关见解。例如，预测市场增长、跟踪和隔离感染疾病、管理道路交通以及预测气象。然而，用于传统数据集的传统工具、技术和算法不再适用，因为大数据是动态的、本质上是连续的、采用各种格式、非结构化且规模大。因此，从头开始适应、重写、重新设计这些工具和算法以应对新的 Data 特征和相关挑战非常重要。

在大数据中，数据最初来自不同的方面，来自必须经过清理、过滤、处理、集成、合并、分区、运输、草图和存储的多个来源。所有这些步骤都是实时、批量或并行执行的，最好是在云端执行。虽然众所周知，从理论上讲，更高质量的数据会带来更好的预测能力和整体洞察力，但这些原始数据必须在预处理阶段通过质量评估来引导，在该阶段，我们会进行数据清理、重复数据删除、压缩、过滤和格式转换等活动。这个强制性步骤对于提炼、评估数据和确保其质量至关重要。

为了跟踪数据价值和相关性，以及上述预处理和处理转换的影响的严重性，数据质量的概念至关重要。此外，目标数据的性质，例如从社交网络生成的数据，以及没有质量参考的非结构化数据，表明必须在初始阶段对数据进行剖析并提供一定的质量信息。这也意味着必须在数据属性的整个生命周期中评估、改进和控制数据属性质量，因为它直接影响分析阶段的结果。

数据质量是数据库社区中一个众所周知的概念，多年来一直是一个活跃的研究领域。但是，将这些质量概念直接应用于大数据在数据预处理的时间和成本方面面临严峻的挑战。由于这些技术是为结构良好的数据开发的，这一事实加剧了这个问题。大数据揭示了新的特征，使其质量评估极具挑战性。大数据的多样性带来了复杂的数据结构，这增加了其质量评估的难度。此外，大数据量大涉及处理时间和资源，几乎不会影响其质量评估过程。此外，鉴于大数据的生成和波动速度，可变性、速度和波动性特征在管理和评估大数据质量方面带来了新的挑战。据我们所知，目前还没有出现针对大数据的标准质量管理框架。现有的大数据质量管理工作大多仍在调查中，尚未达到良好的成熟度。数据库社区过去的工作无法完全采用，因为上面提到的大数据新挑战。但是，一些质量评估实践可以重新调整以应对这些新问题。

在这种情况下，应开发数据质量模型以遵循一些大数据关键概念，如其应用的数据源、域、性质、格式和类型。在处理大型数据集时，正确管理这些质量方案至关重要。此外，现有的大数据架构不支持质量管理流程。然而，一些计划仍然局限于特定的应用程序域和范围。此外，质量的评估和估计必须在从数据开始到分析的所有生命周期阶段进行处理。这种评估对于提供增值服务和实现大数据愿景至关重要。质量测量、评估、实施、监控和适应是关键的质量流程，将说明大数据质量管理的含义。

本文的其余部分组织如下：下一部分介绍大数据和数据质量基础、定义、特征和生命周期。第 3 节介绍了大数据价值链的整体质量管理模型。第 4 节对大数据质量评估和管理方面最重要的研究工作进行了调查和分类。第 5 节确定了大数据质量管理的主要挑战和开放的研究方向，以及质量管理子流程。最后，最后一部分总结了本文的持续和具有挑战性的方向。

（2）BIG DATA AND DATA QUALITY FOUNDATIONS（大数据和数据质量基础）

根据 IBM 、Gartner 、McKinsey的数据，每天都会产生大量的数据;此数据表示 2.5 万亿字节（EB = 1018 字节）。在 2000 年，存储了 800,000 PB（1 PB= 10 字节）的数据。20 年后，即 2020 年，这个数字将达到 35 ZB（1 ZB= 10 字节）。数据存储的指数级增长源于 Google、Yahoo 等 Web 搜索公司;他们必须查询结构松散的数据源的非常大的分布式聚合。此外，包括 Facebook、Amazon、Twitter、YouTube、物联网传感器、移动智能手机在内的应用领域是主要参与者和数据生成器。他们每天生成的数据量为 5 到 10 TB（1 TB= 10 字节）。

1. Big Data(大数据)

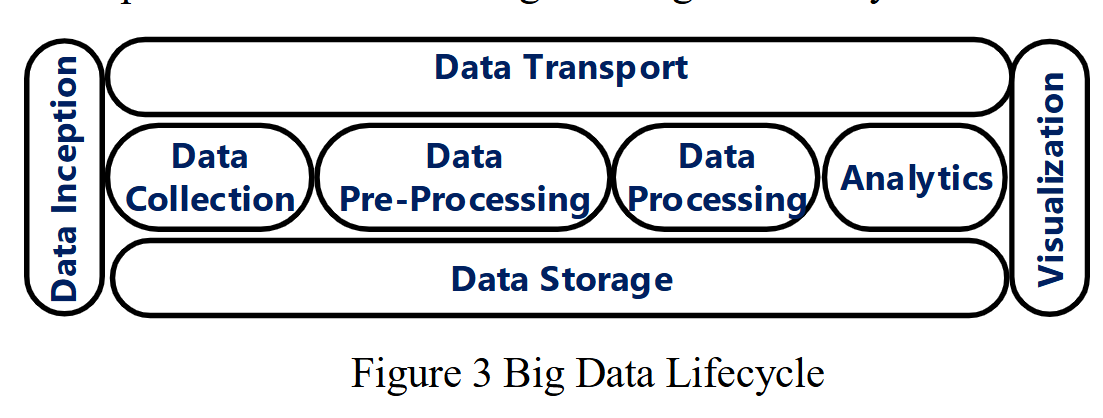
如果我们需要定义大数据，我们必须介绍它多年来的演变，同时将其与它的特性联系起来。顾名思义，这在某种程度上是关于传统数据库无法处理的大量数据文件。然后扩展到涵盖使用传统软件算法分析这些数据的困难。大数据是指整个价值链，包括几个阶段：数据生成、收集、采集、运输、存储、预处理以及处理、分析和可视化。我们可以从这个链条中提取的见解来自使用新技术和新架构的持续数据增长。

1. 定义：根据许多参考资料，大数据没有明确和最终的定义。高容量、高速度和高种类的信息资产需要经济高效、创新的信息处理形式，以增强洞察力和决策能力。“大数据”用于描述大量的结构化和非结构化数据;因此，使用传统的数据库和软件技术很难对其进行处理。它还指组织处理和管理来自多个来源的大量数据所需的技术和存储设施。
2. 来源：数据来自四面八方：用于收集气候信息的传感器、社交媒体网站上的帖子、上传到媒体门户的数字图片和视频、购买交易记录和手机 GPS 信号等等。巨大的数据量并不意味着它是唯一需要考虑的特征。
3. Charachteristics（特征）

2011 年，大数据的早期阶段，McKinney Global Institute 的报告确定了大数据与任何其他数据概念不同的三个主要原始维度。Volume， Velocity 和 Variety，也称为 3 V，如图 2 所示。这些特性不受数量限制是至关重要的。因为大数据不仅仅是一个大小问题;这是一个在新的和新兴的数据和内容类型中发现见解以做出有用决策的机会。最近，维度数增加到 4、7 甚至 10 V 。

1. Big Data Lifecycle（大数据生命周期）

如前所述，如图 3 所示，大数据生态系统被组织为从数据开始到可视化的价值链生命周期。下面简要介绍了大数据生命周期的所有主要阶段。



1. 数据生成/开始：是创建数据的阶段，许多数据源负责这些数据：电生理信号、用于收集气候信息的传感器、监控设备、社交媒体网站的帖子、视频和静止图像、交易记录等等。
2. 数据采集：包括数据收集、数据传输和数据预处理。数据收集：数据以特定数据格式从不同来源收集：使用传感器和 RFID 进行真实世界数据测量，或使用特定设计脚本从任何来源获取数据来抓取网络。数据传输：使用互连网络将收集的数据传输到存储数据中心。数据预处理：它包括典型的预处理数据活动如数据集成、扩充、转换、缩减和清理等。
3. 数据存储：数据存储并分布在多个集群中的基础设施数据中心，数据中心在地理上分布。存储系统可确保多个容错级别，以实现可靠性和效率。
4. 数据处理与分析：应用数据挖掘算法、机器学习、人工智能和深度学习来处理数据并提取有用的洞察以更好地做出决策。数据科学家是此阶段最受期待的用户，因为他们拥有将所需内容应用于必须分析的内容的专业知识。
5. 数据可视化：评估已处理数据价值的最佳方法是直观地检查数据并做出相应的决策。可视化方法在大数据中的应用非常重要，因为它可以闭合价值链的闭环。
6. Data Quality(数据质量)

数据质量 （DQ） 领域的大多数研究来自数据库管理研究社区。数据质量不容易定义，其定义是数据域感知的。一般来说，人们一致认为数据质量始终取决于数据源的质量。

1. 定义：众所周知，DQ 有许多与上下文、域、区域或使用它的字段相关的定义。学术界对 DQ 的理解与工业界的理解不同。在 [21] 中，作者从 ISO 25012 标准中最已知和最常用的定义中总结了数据质量。在文献中，数据质量是 “适合使用”。在 [19] 中，数据质量被定义为使用或满足用户需求的适当性。
2. 数据质量维度 （DQD）：根据 [19]、[22]、[23] 中，DQD 提供了一种衡量和管理数据质量的方法。有多个质量维度，每个维度都与特定量度相关联。DQD 通常分为四类，如图 4 所示：内在、上下文、表征和可访问性 [23]–[27]。例如，上下文维度与信息相关，而 intrinsic 是指目标和原生数据属性。内部 DQD 的示例包括：准确性：衡量数据是否被正确记录并反映实际值。及时性：衡量数据是否作为数据货币和波动性是最新的 [28]。一致性：衡量数据是否与其格式和结构一致。主要是对数据约束的尊重。完整性：描述是否记录了所有相关数据。它测量属性的缺失值。
3. 数据质量指标：每个 DQD 都需要量化和衡量。DQD 指标表示评估这些维度的步骤。从简单的公式到更复杂的多变量表达式，这些指标给出了可测量性属性分配给 DQD。例如，属性的缺失值计算被视为评估 DQD 完整性的度量 [29]–[31]。
4. 数据质量评估：遵循数据驱动的策略需要对已生成的数据进行质量评估。因此，必须测量和量化 DQD。对于结构化或半结构化数据，数据可作为一组属性（以列或行表示），并分别记录它们的值。任何数据质量指标都应指定数据值是否遵循质量属性（维度）。作者在 [32] 中引用了数据质量测量指标倾向于评估二进制结果的正确或不正确（分别为 0 和 100%），并使用通用公式来计算应用于质量维度的这些属性（例如准确性）。测量使用其相关的已定义量度生成 DQD 的分数（例如，准确率分数是正确实例值的数量除以测量属性的实例总数）。
5. Big Data Quality Evaluation（大数据质量评估）

数据质量在大数据生命周期中的重要性重新定义了处理数据监督的方式。管理数据质量涉及在每个阶段添加更多功能，并进行持续的质量控制和监控，以避免在生命周期的所有阶段出现质量故障。大数据质量评估关注性能、价值和成本等属性。在下一节中，我们将详细说明如何在大数据生命周期中管理和评估质量。

1. Big Data Quality Issues(大数据质量问题)

当数据值不满足质量要求时，就会出现数据质量问题 [33]。这些问题是由于发生在不同级别的几个因素或过程造成的：1） 数据源：不可靠性、信任、数据复制、不一致、多源和数据域，2） 生成级别：人类数据输入、传感器设备读数、社交媒体、非结构化数据和缺失值，以及 3） 过程和/或应用程序级别（采集：收集、传输）。数据预处理通过执行许多任务和活动（如数据转换、集成、融合和规范化）来提高数据质量。[21]、[34] 中的作者列举了影响数据质量的不良数据的许多原因，并提出了影响数据质量及其相关维度的因素列表。在 [19]、[34] 中，作者解决了按 DQD、粒度级别和数据源类型分类的不良数据原因的汇编，同时强调了这些之间的因果关系映射。

1. A HOLISTIC QUALITY MANAGEMENT MODEL FOR BIG DATA（大数据的整体质量管理模型）

在大数据中，数据管理对其质量管理是公正的。因此，我们需要识别生命周期每个阶段的质量问题和要求。为了确保高质量的价值链，改进程序是不可避免的，并且应该内置于生命周期的每个流程中。目标是在进行质量管理活动时，应在不增加不同大数据生态系统层的额外通信、处理和成本开销的情况下进行。

要启动任何大数据质量项目，必须列举一组参数和概念来识别流程，并描述用作其管理关键输入的数据和工作流类型。在数据质量方面正在实施两种策略：论文前面提到的数据驱动和过程驱动策略。数据驱动的方法作用于数据本身，评估其质量以提高其质量。流程驱动是一种预测方法，侧重于用于生成或操作数据的流程的质量。

在图 5 中，我们提出了一个整体质量管理模型，该模型捕获了重要的质量方面，并探索了如何在大数据的整个生命周期中处理大数据质量。我们确定了必须处理和解决数据质量问题的流程，并提供了质量评估方案以确保其有效管理。此顺序中最重要的阶段是：（1） 数据创建，（2） 数据源 （3） 数据收集，（4） 数据传输，（5） 数据存储，（6） 数据预处理，（7） 处理和分析，以及 （8） 可视化。此外，必须解决大数据生命周期阶段及其相关的质量信息和流程，以实现端到端的大数据质量管理驱动的生命周期。

在下文中，我们强调了质量要求考虑、质量实施和执行在大数据价值链的不同流程中的重要性。我们还描述了质量在价值链不同流程中的传播，以及这些流程之间的协调水平，以支持质量管理。最后，我们说明了通过环回和进程间交互来持续改进质量。

1. 确保质量的数据采集

数据采集分两个阶段处理：数据开始和数据收集。但是，数据源（图 5 中的 2）与通常以多种格式存储的现有数据相关.在从根本上解决大数据质量问题之前，首先，我们需要检查过去积累的现有数据，并根据这些数据构建知识库，以更好地预测为大数据重新设计其创建。在这种情况下，数据设计必须是一个迭代过程，它会影响数据检查和数据质量评估（在使用现有数据时通常遵循数据驱动的方法），以增强创建可供大数据生命周期流程使用的高质量数据的流程。

在数据创建阶段采用流程驱动策略，其中设置质量约束以在创建不良数据之前消除不良数据。此外，在创建数据后，将启动收集过程，以将数据收集为更结构化和更有条理的格式，从而成为有用的数据。用于重新组织数据本身的一组技术可确保并保证数据的质量。数据收集过程被认为是数据管理过程，如果使用适当的数据科学方法和技术，它可以对数据质量产生积极影响 [35]。

用于设计数据的框架和工具各不相同，包括传统数据库系统到更多与大数据相关的系统，如 NoSQL 数据库（MongoDb、Cassandra、Hadoop HBase 和其他相关的 Hadoop 生态系统工具）。基于 NoSQL 的数据库设计将阻止复杂的表格结构查询，并导致使用连接器工具。

1. 质量感知型数据传输和存储

对于数据传输阶段，考虑到服务配置期间应满足的要求，支持 QoS 非常重要。确保数据质量更具体地与用于传输数据的底层网络以及保证在多个点之间传输数据而不会丢失或损坏数据的安全措施有关。数据网络的质量取决于客户的 QoS 要求，提供商 QoS 产品和实时 QoS 测量。QoS 的主要目标是提供优先级，包括专用带宽、受控的抖动和延迟（例如大数据实时流处理）以及改进的丢失特性。此外，提供商应考虑 SLA、主动和被动配置策略，以提供所需的 QoS。

对于数据存储，通过数据分发和复制来处理大数据质量。例如，使用 Hadoop 生态系统的存储依赖于多个节点来复制数据，以避免任何灾难性的数据丢失，并确保故障发生时的连续性。此外，数据存储质量取决于多种类型的大数据的存储介质 I/O 性能。读取和写入比率必须遵循每种数据类型（例如高清视频数据和流处理）的 I/O 要求。

1. 质量驱动的数据预处理

预处理是大数据中质量管理的最后补救措施。它是在被处理阶段消耗之前清理、集成、规范化、转换、集成数据以提高其质量的过程。它遵循严重依赖数据值的数据驱动方法。例如，数据清理依赖于完整性、一致性。质量设置（例如需求、基线/模型和质量报告）对于实现数据质量改进的流程至关重要。它确定数据、维度和分数的预期质量。模型或基线表示基本质量要求边界，可以根据质量评估过程以及质量与其要求的匹配进行调整。最后，质量报告与数据来源类似，记录了从开始到最终分析阶段的数据路径。质量报告内容会迭代更新，并使用新参数进行扩充，以用作大数据质量存储库。许多数据预处理工具已经出现，此外，数据库领域中使用的其他现有工具也已更新以处理大数据。大多数情况下，基于 Hadoop 和 spark 的预处理工具和框架被广泛使用（例如，用于大数据的 Talend Open Studio [36] 和 Open Refine [37]）。

1. 支持质量的数据处理和分析

处理和分析利用了之前质量评估活动中进行的质量评估。在处理阶段，质量管理包括验证预处理后的数据是否符合处理质量要求，此外还包括对处理算法（例如 基于 Hadoop 的处理）。但是，分析包括评估不同的分析方案和算法（例如深度学习），以确保在可视化从分析中检索的见解之前质量评估的有效性。

至于大数据预处理，处理框架与 Hadoop 和 spark 相关。在 [38] 中，已经详细介绍了许多用于预处理和处理的框架和方法。

1. 支持质量的数据可视化

它包括聚合可视化工具的质量要求、质量报告和各种可视化工具的质量。它还评估可视化数据，从而做出准确的决策。这个阶段非常重要，因为它是关闭大数据价值链中质量评估循环的最后一个过程，并触发一些持续改进的建议，以实现持续的质量监控。使用了许多大数据可视化工具，例如 Google Chart、Tableau、DataWrapper。

1. 质量传播和持续的质量改进

图 5 中提出的质量管理模型说明了质量管理活动如何在大数据价值链的不同流程中传播。在每个阶段，都会考虑质量要求，反映质量测量结果，并生成质量报告。此质量报告将转发到价值链的下一阶段，以跟踪、验证、丰富和提高数据和流程的质量。通过流程之间的环回，可以在需要时监控、修改和提高质量，从而实现持续的质量改进。这是一个非常重要的功能，可确保整个大数据价值链中质量评估的准确性和改进。

（4）BIG DATA QUALITY: RESEARCH CLASSIFICATION（大数据质量：研究分类）

如图 6 所示，我们选择了大量主要解决并讨论了大数据中的数据质量的论文。他们中的一些深入探讨了特定的质量属性，然而，另一些则解决了大数据质量问题，并提出了一些基于质量评估的解决方案，在应用各种技术的同时进行改进。指定的文献揭示了数据质量已经从不同的认知（例如数据、过程、应用程序及其管理）中得到解决，重点是质量属性评估、质量评估过程和与大数据相关的质量模型。该分类的目标是提取与大数据质量相关的主要研究趋势，确定迄今为止在其质量管理中已解决的问题，以及需要进一步探索以充分发挥其潜力的内容。此外，我们预测这些发现将为未来的研究方向指明道路。我们提出了 10 个课程（从 I 到 XI），用于确定大数据质量的主要研究趋势。这些类别还分为 5 个主要集群（A 到 E），用于定义与数据质量和大数据相关的最重要关注领域。

1. 大数据价值链

在 [35]、[39]–[49] 中，作者强调了许多数据质量问题及其在大数据生命周期阶段的影响;作为预处理、处理和存储。他们主要提出了一个管理流程作为分类法，以控制预处理的质量，以及在分析之前验证数据质量、数据源和数据格式的处理。其他人提出了一个组合质量模型来检测预处理缺陷并采取行动纠正任务缺陷。

对于数据处理，一些研究利用分析技术、机器学习和分类实践，并评估它们对大数据的适用性。然后，他们分析了数据大小的影响，以评估这些方法应用的准确性。其他人建议调整这些技术来处理大数据特征并解决不同的处理质量问题。

由于存储对于大数据至关重要，因此作者讨论了多存储提供商及其对预处理性能和效率的影响，当采用多个云提供商上的高数据分发时进行处理。以前的大多数工作都以临时的方式解决了大数据中的质量问题，而没有遵循考虑质量特征、流程和底层基础设施的综合模型。这种模式将确保价值链中的端到端质量管理。大多数作品单独讨论了大数据阶段，而没有在其整个生命周期中利用它们的影响。

1. 大数据管理和特性

在这一类别中，大数据的两个重要方面被针对[31]，[35]，[39]–[41]，[44]，[46]–[59]，即其特征（V）和管理。大多数作者都同意 V 代表了 Big Dat 的一个重要方面，如果数据质量没有得到有效的管理，就会对数据质量产生很大影响。他们通常将 V 与 DQD 联系起来，以便发现相关性和相互影响。其他人强调，可扩展性、完整性和资源优化与大数据 V 高度成正比，因为它们代表了其 Management 解决方案的关键要素。

在大数据管理中，作者调查并提出了处理存储、预处理和处理的管理模型。此外，还仔细研究了管理过程中涉及的每个过程的技术和方法的最新审查。管理大数据需要将其价值链阶段与相关流程和子流程进行映射。质量在大数据管理中的重要性通常没有得到解决。这样一个具有端到端质量管理和执行的框架非常具有挑战性。

1. 大数据问题和数据质量问题

将大数据问题与数据质量问题聚集在一起是合理的，因为这两个概念之间存在密切的关系。任何数据质量问题都将反映在分析中。精选文献的作者[25]、[31]、[41]、[43]、[45]、[46]、[48]、[49]、[51]–[56]、[58]–[60]强调，尽早发现质量问题并将其与生命周期中的大数据问题进行映射非常重要。这将有助于隔离和调整必须处理这两个问题的流程。大多数数据质量问题已在研究界得到大量解决，但仍然没有在大数据中得到适应。与 DQD 和 V 相关的进一步讨论被认为是成比例的。

1. 数据质量

在接下来的工作中，数据质量已经得到了充分的研究[25]、[31]、[39]、[41]-[46]、[48]-[59]、[61]，这证实了它对大数据的重要性。大多数作者认为 DQD 是一个重要的模型。他们基于与 V 映射的 DQD 开发了大数据质量模型，以解决可扩展性和可靠性问题。其他解决方案解决了为图像、二进制数据和非结构化数据等异常数据选择 DQD 和指标的问题。他们提取了特征并组合了许多 DQD 来衡量质量维度分数。要提出大数据质量评估方案，需要解决许多挑战。作者列出了许多必须考虑的功能：例如：准确性、一致性、出处、不确定性。

因此，对于大数据中的数据质量及其管理，没有完整的参考模型。DQM 必须确保数据一致性，并遵循自成立以来的所有步骤以进行质量评估。换句话说，新的技术和决策挑战使 DQM 应用程序变得更加复杂。

1. 大数据应用和质量改进

它是一个基于价值链的应用程序，遵循从数据创建到可视化的各个阶段。一些作者专注于如何通过评估资源管理指标（如存储和处理）来管理这些应用程序中的质量。其他人则提出了提高数据质量的解决方案，同时应用作为预处理一部分的清理任务和活动（例如 BigDed 和 Nadeef）[51]– [53]、[56]、[68]。数据质量包括评估质量维度，以及将上述所有内容结合起来，以汇总一个反映大数据生命周期应用程序质量的质量分数。

1. DISCUSSION AND FUTURE DIRECTIONS（讨论和未来方向）

确保质量被认为是大数据时代最具挑战性的问题之一。学术界和工业界目前提出的解决质量问题的方法和解决方案尚未达到令人信服的成熟度。评估大数据质量的重要性与它为用户（例如政府、企业）产生的价值至关重要。此外，遵循经过充分研究的数据质量管理计划，使用正确的评估方案，采用适当的质量测量方法，并利用适当的工具和平台共同开展不同的质量评估活动，将有助于实现高质量的评估结果。此外，解决整个大数据价值链的质量问题可以实施端到端的质量评估和管理，并带来更好的质量改进。最后，评估质量评估的开销保证了具有成本效益的质量管理流程。

最后，值得一提的是，大数据质量的研究工作和解决方案仍处于初级阶段，在大数据质量的发展和标准化方面还有很多工作要做。这是一个多学科、复杂和多变体的领域，新的评估、处理技术和分析算法、存储技术和处理平台将在这一活跃研究领域的发展和成熟中发挥重要作用。我们预计学术界的研究人员将为开发新的大数据质量方法、算法和优化技术做出贡献，这些方法、算法和优化技术超越了数据库和数据仓库中使用的传统方法、算法和优化技术。但是，各行各业将引领新平台、解决方案和技术的开发计划，以支持大数据生命周期内的端到端质量管理。

1. CONCLUSION（结论）

大数据已成为处理大量、连续、变化和复杂数据的新范式。它的质量是其接受和有用的关键。数据质量差可能会导致严重后果。这将失去以适当的方式分析和探索大规模数据集的好处。使用传统技术来管理大数据再合适不过了。因此，高度要求设计和应用有效的方法来管理质量。在本文中，我们确定了大数据质量方面的关键研究挑战，并强调了它们的重要性。然后，我们调查、分类和讨论了最全面的研究计划。之后，我们提出了大数据质量管理模型的整体视图，该模型强调在整个价值链中要进行的关键质量评估活动。Finlay，我们讨论了大数据质量评估的主要趋势，并指出了一些未来的研究方向。我们计划进一步扩展这项工作的范围，并深入描述如何在真实的大数据项目背景下实施质量评估活动，以及质量至关重要的地方。