1-4-cn
1
大家好，我是北京理工大学计算机学院数据科学与知识工程研究所的车海莺，本节我们讨论大数据的特征
2
原来大数据的 Characters 是 3 Vs，即 volume、velocity、variety 随着信息技术的发展，系统或设备记录的数据急剧增加。
如图所示。一开始，Originally the Characters of Big data is 3 Vs, which are volume, velocERP（企业资源计划）系统记录采购明细、采购记录和付款记录，主要是交易记录，数据规模为 MB。
然后他们扩展了客户的内容，在 CRM 中将细分、报价详情、客户接触、支持联系人信息记录在系统中，这是与组织边界的交易和交互，数据规模为 GB。
进一步在 Web 的帮助下，Web 日志、报价历史记录、A/B 测试、动态定价、附属网络、搜索营销、行为定位、动态漏斗也可以重新编码和存储有用的信息。其中是与组织内外的交易和交互，数据规模为 TB。
在大数据范围内，传感器/RFID/设备、用户点击流、移动网络、情感分析、用户生成的内容、社交互动和提要、空间和 GPS 坐标、外部人口统计、业务数据供给、高清视频、音频、图像、语音转文本、产品/服务日志、短信/彩信等。数据范围非常全面，有交易、交互和观察，数据规模为 PB，所以从这个发展中，我们可以看到，数据量和种类，速度也在增加。让我们一一介绍这些特征。
3
1. 容量
容量是指大型组织每秒收集和生成的大量数据。
数据规模从 2009 年到 2020 年增长了 44 倍，从 0.8 zettabytes 到 35 zettabytes， 数据量呈指数级增长，这些数据来自不同的来源，
例如物联网设备、社交媒体、视频、金融交易和客户日志。数据量可能会有所不同。
例如，文本文件为几千字节，而视频文件为几兆字节。
IDC《数据时代 2025》报告，到 2025 年，全球数据将达到 175 ZB（2009 年的 218.75 倍）
4
地球范围是世界上最大的科学项目。
该天文台旨在追踪北美的地质演化，记录超过 380 万平方英里的数据，积累 67 TB 的数据。
它分析了圣安地列斯断层的地震滑动，当然还有黄石公园下方的岩浆羽流等等。
5
另一个最重要的大数据特征是它的多样性。
它指的是不同的数据来源及其性质。多年来，数据来源发生了变化。 早些时候，它仅在电子表格和数据库中可用。如今，数据存在于照片、音频文件、视频、文本文件和 PDF 中。
包括关系数据（表/事务/遗留数据）
文本数据（网络）
半结构化数据（XML）
图表数据 社交网络、语义网（RDF）、……
流数据您只能扫描一次数据
单个应用程序可以生成/收集多种类型的数据
公共大数据（在线、天气、金融等）
数据的多样性对其存储和分析至关重要。为了提取知识，所有这些类型的数据都需要链接在一起
6
例如，如果您想分析您的客户，您想了解他的所有不同方面。
包括他的购买偏好、我们在交易系统、银行、金融历史、社交媒体、游戏和娱乐等方面的已知历史，
通过所有这些数据，您将了解有关您客户的全面信息并提供更好的服务。

7
该术语是指创建或生成数据的速度。
数据生成速度快，需要快速处理这种数据产生的速度也与处理这些数据的速度有关。
这是因为只有经过分析和处理，数据才能满足客户/用户的需求。
传感器、社交媒体网站和应用程序日志产生了大量数据，而且所有这些数据都是连续的。
如果数据流不连续，那么在其上投入时间或精力是没有意义的。
迟到的决定会导致错失良机
例如，E-Promotions：根据您当前的位置，您的购买历史来分析您想要什么和 立即为您旁边的商店发送促销信息，如果您在距离商店很远的时候发送了促销信息，那是没有用的。
以及医疗保健监测，传感器监测您的活动和身体，当发生任何异常测量时，需要立即做出反应
8
许多数据是实时的。在社交媒体和网络中，我们所有人都在生成数据。
科学仪器正在收集各种数据。移动设备一直在跟踪所有对象。传感器技术和网络正在测量各种数据。
人和设备都在生成数据。
收集数据的能力不再阻碍进步和创新但是，通过及时和可扩展的方式管理、分析、总结、可视化和从收集的数据中发现知识的能力。
9
实时分析很重要，它可以使实时做出更好的决策，在发生时采取行动。
例如：相关且引人注目的产品推荐。
了解客户为何转向竞争对手及其报价及时反击在促销活动仍在进行时提高促销活动的营销效果。
预防欺诈，因为它正在发生并更主动地预防。
10
让我们看看 5Vs，1 规模，静态数据，TB 到 EB 的现有数据要处理，
2 速度，运动中的数据，流数据，毫秒到秒的响应。
3 种类繁多，数据形式多样，结构化、非结构化、文本、多媒体等另外两个是真实性和价值。
4 真实性：可疑的数据，由于数据不一致和不完整、歧义、延迟、欺骗、模型近似而导致的不确定性。大数据的这一特性与前一个特性相连。它定义了数据的可信度。由于您遇到的大多数数据都是非结构化的因此过滤掉不必要的信息并将其余信息用于处理非常重要。
5 价值：在大数据的特征中，价值也许是最重要的。无论数据的生成速度或数量有多快，它都必须可靠且有用。否则，数据不足以进行处理或分析。
研究表明，质量差的数据可能导致公司收入损失近 20%。数据科学家首先将原始数据转换为信息。然后清理这个数据集以检索最有用的数据。
分析和模式识别是在这个数据集上完成的。如果该过程成功，则可以认为数据是有价值的。即使是有价值的数据，它的价值密度也很低，就像在沙滩上淘金一样。几个小时的视频中的几秒钟可能是有用的。
11
大数据是商业、营销、销售、分析和研究等主要领域背后的驱动力。 它改变了全球以客户和产品为基础的公司的业务战略。 因此，在分析和决策制定时，必须对所有大数据特征给予同等重视。
在本节中，我们了解了大数据的特征、数量、速度、多样性、准确性和价值，这些是您在进行大数据分析时需要考虑的关键因素。