技术篇

基于 NoSQL 和 NewSQL 新技术的大数据管理与分析

第 1 章　大数据及其特点

【本章要点】

- ✧ 大数据的定义与 5V 特点
- ✧ 大数据的应用场景
- ✧ 对大数据的误解
- ✧ CAP 理论与 BASE

1.1　大数据时代当前的状态

大数据（海量数据的全面升级版）已经成为风靡全世界的 IT（信息技术）新领域，毫不夸张地说，我们已经进入了一个大数据时代。Kantar Media CIC 每年都会通过一张信息图整理出中国互联网发展的数据。图 1.1 展示了 2017 年中国社交媒体、电子商务、共享经济等领域的大数据爆炸式增长。

类似地，图 1.2 展示了 2017 年短短 60 s 内，美国主要互联网公司产生的数据量。

在大数据时代，商务与科技人士对大数据及其发展的认知是十分积极的[1]。IBM（国际商业机器公司）调查了来自 70 个国家的 900 个商务和 IT 经理，这些商务领导者认为大数据带来的效应是[2]：

（1）他们基于数据进行绝大多数的决策的可能性增加到 166%。

（2）以数据分析为职业发展道路的可能性提升了 2.2 倍。

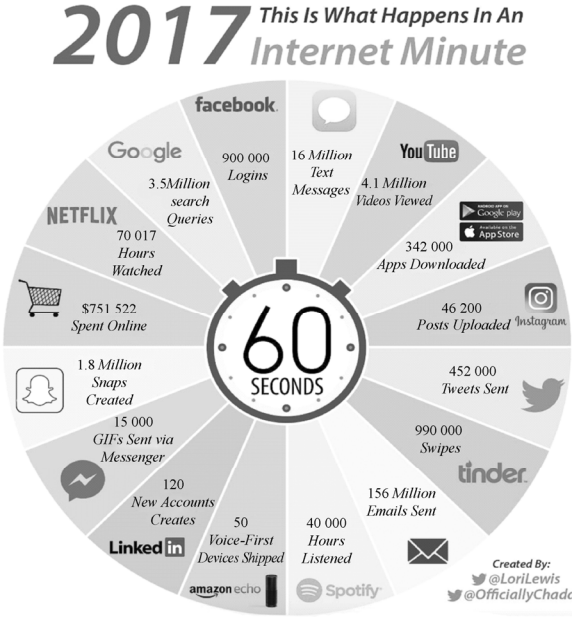
（3）他们在使用来自数据分析的关键价值资源方面增长了 75%。

（4）他们中 80%的人要衡量大数据对分析投资的影响力。

（5）他们中 85%的人拥有这样或那样的共享大数据分析资源。

中国 2017 年社会媒体
60 s 信息量
（扫码查看彩图）

图 1.1　Kantar Media CIC 发布的 2017 年中国社会媒体 60 s 信息量①



美国 2017 年互联网
60 s 信息量
（扫码查看彩图）

① Kantar Media CIC（中国社会化商业资讯提供商）2017 年发布。http：//www.ciccorporate.com/index.php?option=com_content&view= article& id=1379%3Akantar-media-cic-released-2017-every-60-seconds-in-china-infographic-big-data-for-understanding-chinese-social-media&catid=112%3AarcHives-2017&Itemid=223&lang=zh（2018-07-18 可访问）

图 1.2　美国 2017 年互联网 60 s 内的信息生成量

TEK 系统针对大数据调查了 2 000 多名 IT 专业人士和 1 500 多名 IT 领导，得到了以下的统计数据[2]：

（1）90%的 IT 领导者和 84%的 IT 专家相信在大数据上投入时间、金钱和资源是值得的。

（2）14%的 IT 领导者认为，在他们的组织中大数据的概念会经常应用。

（3）66%的 IT 领导者和 53%的 IT 专家报告，他们的数据存储在完全不同的系统中。

（4）60% 的 IT 领导者和 53%的 IT 专家报告，他们的组织缺少对数据质量的责任感。

（5）多于 50%的 IT 领导者质疑他们的数据的有效性。

（6）81%的 IT 领导者认为他们的组织缺少必需的专业人员，这些人应该能计划、建设和执行大数据行动。

1.2　大数据定义与特点

1.2.1　什么是“大数据”（Big Data）？

按照全球最具权威的 IT 研究与顾问咨询公司 Gartner 的定义[3]，数据是海量、高增长率和多样化的信息资产，它需要性价比高并具有创新性的处理模式，才能具有更强的决策力、洞察力和流程自动化。在 Gartner Group 的定义中首次定义了大数据 3V 的特点。

大数据包含三类数据：无结构化数据、结构化数据、半结构化数据。无结构化数据是指数据没有预定义的结构、类型、模式或数据模型等，如 PDF、email、文本式数据。网页的 HTML 数据虽然有标签，但只是用于面向浏览器的文档显示样式渲染，并没有捕捉、存储和自动处理信息内容的功能，所以仍然是无结构化的。结构化数据是数据具有预先定义的符合规则的结构、类型和模式等，具有可处理、存储、使用的元数据信息，如传统的关系数据库数据。半结构化数据具有很有限的结构、数据类型或模式定义，如 XML。

1.2.2　大数据的 5V 特点

大数据的 5V 特点是 IBM 提出的，分别是数据量（Volume）、多样性（Variety）、高速（Velocity）、价值（Value）和真实性（Veracity），具体要点如图 1.3 所示。

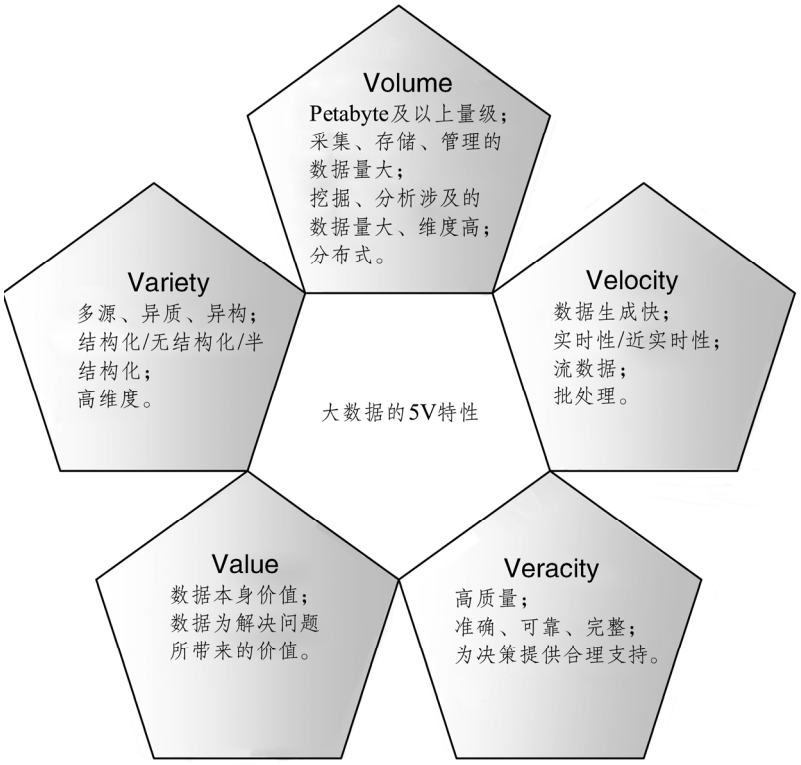


图 1.3　大数据的 5V 特性和具体要点

（1）Volume：表示数据量巨大，包括数据采集、存储和计算的量都非常大。大数据的起始计量单位至少是 PB（1 000 个 TB）、EB（100 万个 TB）或 ZB（10 亿个 TB）级别。

（2）Variety：数据的来源、类型、格式、语义等是多种多样的，具有多源异质异构性。80%的大数据是半结构化和非结构化的，例如：网络日志、社交网络平台数据、网页文件、电子商务交易数据、设备传感器数据、音频、视频、图片、地理位置信息等。多源异质异构数据对处理能力提出了更高的要求。

（3）Velocity：数据增长速度快，I/O 速度快，从而要求数据处理速度要快，时效性高。例如：天气大数据的动态性，提出了快速处理要求；“ 双

十一”电商平台对高速增长的交易数据管理和处理；以毫秒级速率产生的各种传感器数据；等等。

（4）Value：数据应被用于解决特定的问题或完成特定的任务，因此对大数据本身及其带来的价值要求高。大数据具有巨量性，但原始未加工的数据因其量大而杂乱稀释了本身的价值；大数据的巨量性也造成了数据挖掘算法的低效，甚至失效，挖掘出的结果并未带来期望的价值，不能满足各类应用对大数据大价值的要求，这也是大数据的特点和挑战。

（5）Veracity：大数据质量不高，体现在准确性、可靠性、完整性等方面的挑战。因为期望从大数据中获得决策，所以需要高质量的大数据。

现在大数据的 5V 特性又被扩展为 6V、7V 等。其中可视化（Visualization）作为大数据所呈现出的特点比较勉强，但作为理解大数据的基本处理要求是迫在眉睫的，因为可视化技术是大数据分析最直观、易懂，也是最理想的方法。这是对信息的一种新的阅读和理解方式。

1.3　沃尔玛应用大数据的案例

沃尔玛（Walmart Inc）开始使用大数据的时间甚至早于这个词汇享誉整个业界。2012 年沃尔玛已将 10 个节点的 Hadoop 平台集群升级到了 250 个节点，同时将存放在 Oracle，Netezza 和 Greenplum 硬件上的数据迁移到其自己的系统上，目标是将 10 个不同网站集成为一个网站，以便在新的 Hadoop 集群上存放所有新产生和流入的数据。

沃尔玛的实验室研发了许多大数据工具，如 Social Genome，ShoppyCat 和 Get on the Shelf。

按照资料[4]的介绍，Social Genome 等大数据分析工具，能够分析百万至十几亿的 Facebook 信息、推文、YouTube 视频、博文等，使得沃尔玛可以与在网络上提到某类商品的顾客或顾客的朋友们联系，告知他们特定的商品信息，包括降价等。这个软件工具组合了来源于万维网的公开数据、社交数据和个人专有数据，如顾客购买数据和合同信息。这样便构建了一个巨大的、持续变化和更新的知识库，库中管理了上亿的实体和关系，从而使得分析师和决策者能更好地理解顾客在线表达的上下文。例如，一位女士定期在推特上讨论电影，当她某次发推说“我喜欢盐”时，沃尔玛能够理解她正在谈论著名的好莱坞电影“盐”（Salt，中文翻译为“特工绍特”），

而不是调味料“盐”。

沃尔玛在研发 Social Genome 时遇到了几个技术挑战。首当其冲的是如大水倾泻般流进集群的数据量，以及输入进来的数据速度。由于通常的 MapReduce 框架无法应对这样的数据量和速度，他们开发了名为 Muppet 的工具，可以实时处理所有集群节点上的数据并同时完成多个分析，Muppet 现在已经开源。

1.4 其他应用实例

1.4.1 教育领域[5]

教育领域充满了海量数据，这些数据与学生、院系、课程、培养成果等密切相关。正确地研究分析这些数据，可以洞察教育规律，提高教书育人质量和教育工作的有效性。例如，大数据分析在以下几个方面可以发挥巨大作用：

（1）个性化和动态学习。基于学生的学习历史可以为每个学生创建个性化的学习程序和模式，从而提高学生的学习成果。

（2）教学材料再组织。通过分析大数据再构造或组织教学材料。这些数据来源于学生喜欢学习什么内容，实时监测到的课程哪些部分更容易理解等。

（3）成绩系统。通过分析学生数据改进成绩分析等系统。

（4）事业预测。通过正确地分析研究每个学生的记录，可以帮助理解学生的进步、强项与弱项、兴趣等，从而帮助确定哪些事业对学生的未来是最合适的。

真实的应用：阿拉巴马大学有 38 000 名学生和海量数据。大学管理者能够使用大数据分析和可视化技术，获取学生学业相关的模式，从而改革大学的运作、招生、毕业或留级工作。

大数据的应用已为解决教育系统中一个最大的陷阱提供了解决策略，即解决“一种模式适配所有风格”的教育设置问题。

1.4.2 交通领域[5]

交通领域是大数据应用的活跃区：

（1）路线规划。大数据被用来理解和估计客户对不同路线和多种交通方式的需求，然后通过路线规划减少客户的等待时间。

（2）拥堵管理和交通控制。基于大数据，实时估计路途和交通模式已成为可能。例如，人们使用高德或谷歌地图确定最快速、少拥堵的路线。

（3）交通安全级别：通过大数据实时处理和预测分析，标识交通事故多发区域，可进行预警，增加交通安全性。

真实的应用：以 Uber（优步）为例，它产生和使用了大量数据，这些数据与司机、车辆、地点，以及每辆车的每次行程等相关。分析这些数据，可预测司机的供需和地点，预测每次行程的费用。

1.4.3　在音乐领域的应用[6]

如今，汽车已成为大多数人的代步工具，在车内听的歌曲很可能反映了车主的真实喜好。音乐元数据公司 Gracenote 采用智能手机和平板电脑内置的麦克风，基于大数据挖掘技术，识别用户车载音响中播放的歌曲。他们的技术可检测掌声或嘘声等反应，甚至还能检测用户是否调高了音量。这样，Gracenote 就可以发现用户真正喜欢的歌曲，以及听歌的习惯（听歌的时间和地点）。Gracenote 拥有数百万首歌曲的音频和元数据，因而可以快速识别歌曲信息，并按音乐的风格、歌手、地理位置等分类。

1.5　对大数据的误解

1. 大数据仅仅意味着数据巨大？

通常 PB 级的数据被视为大数据的门槛，但数据量仅仅是大数据 5V 特性之一，多样性和数据增长与处理速度也至关重要。大数据量与后两者相结合，才能符合 1.2.1 节中 Gartner Group 的大数据定义的中心思想。

当前的数据已经从传统的结构化演变为多种模式和多种类型，数据多样性对数据采集、管理、融合、处理和深入分析都带来了挑战。例如，传统的关系数据库在管理和分析声音、图像、网络日志、地理定位数据等方面是十分无力的。

1.2.2 节已经讨论了速度特性的意义。传统的技术和方法不适用于大数据引出的高速采集、管理和处理需求。因此，采用新的方法是必要的。

由此可见，大数据带来的挑战是全方位的，这才是“大”的真实内涵。

2. 大数据技术就是 Hadoop?

要解决巨量和快速的大数据带来的挑战，技术创新是核心。一个解决方案是采用分布式处理和存储，提高处理速度。

Hadoop 是 Apache 开源的分布式计算系统，采用了 MapReduce 计算模式和 HDFS 分布式文件系统的存储方式，但有一个较大问题是 MapReduce 读写效率低。Spark 是一个类 Hadoop MapReduce 的通用并行框架，同样是 Apache 开源系统，与 Hadoop 最大区别在于可以将中间结果存放在内存，不需要读写 HDFS，因此 Spark 可以提供超过 Hadoop 100 倍的运算速度，也更好地适用于数据挖掘与机器学习等需要迭代的 MapReduce 的算法。而 Storm 是一个开源的数据流处理系统，能够完成 Hadoop 不擅长的实时计算。

因此 Hadoop，Spark 和 Storm 是目前最重要的三大分布式计算系统，Hadoop 常用于离线的、复杂的大数据处理；Spark 常用于离线的、快速的大数据处理；而 Storm 常用于在线的、实时的大数据处理。

目前大数据技术层出不穷，覆盖了从大数据存储、计算，到分析和挖掘的大部分领域。但传统数据仓库技术依然有用武之地，现有的 IT 基础设施不仅会保留，还会继续发展。

所以，大数据技术十分丰富，Hadoop 仅仅是其中之一。

3. 大数据意味着非结构化数据？

“非结构化”数据在大数据类型中的比例很高，但“结构化”和“半结构化”数据同样存在。印度著名的塔塔咨询服务有限公司（TCS）就发现 51%的数据是结构化的，27%的数据是非结构化的，另有 22%的数据是半结构化的。这也是为什么 Hadoop 框架同时支持面向非结构化数据的 NoSQL 数据库 HBase 和面向结构化和半结构化数据的数据仓库组件 Hive，如图 1.4 所示。

4. 大数据只是社会媒体内容和情感分析?

大数据包含了各行各业所产生出的各类数据，包括网络流量、IT 系统日志、客户的情绪，或任何其他类型的数据。所以，社会媒体内容和用户情感仅仅是大数据中很小的一个方面，银行、保险业、航空、汽车、信用