Intel IT Center

大数据 101：非结构化数据分析

2012 年 6 月

大数据和新兴技术速成课

大数据和非结构化数据分析究竟意味着什么？您是否对它有所担心？本简报将为您提供有关大数据的速成课：它为什么重要、对 IT 部门的影响、非结构化数据分析的新兴技术以及英特尔如何提供帮助。

大数据为什么重要

数据正以惊人的速度激增。从出现文明到 2003 年，人类总共才创造 5 EB（5 ExaBytes, 10^{18} 字节）的数据，但是我们现在仅在两天内就创造出相同的数据量！[1] 到 2012 年，全球数字数据量将增长至 2.72 ZB（ZettaBytes, 10^{21} 字节），并以每两年翻番的速度增长，到 2015 年将达到 8 ZB。举例来说，这相当于 1800 万个国会图书馆。[2] 数十亿台连接的设备 — 从个人电脑和智能手机到 RFID 读取器和交通摄像头等传感设备，都在不断生成复杂的结构化数据和非结构化数据。

大数据是指庞大的数据集，它们有着比以往更大的容量（volume，按数量级）、更高的多样性（variety）和复杂性，以及更快的生成速度（velocity）。这三个关键特性有时被称为大数据的三个 V。

非结构化数据本质上是异构和可变的，同时具有许多格式，包括文本、文档、图形、视频等等。非结构化数据的增长速度比结构化数据的增长速度更快。根据 2011 年的 IDC 调查，[3] 它将占未来十年所创造数据的 90%。作为一个新的尚未开发的信息源，非结构化数据分析可揭露之前很难或无法确定的重要相互关系。

大数据分析是一项技术推动的战略，旨在获得更加丰富、深入和更加准确的客户、合作伙伴以及商业洞察，并最终获得竞争优势。通过处理稳定的实时数据流，与以往相比，组织可更加快速地做出时间敏感的决策、监控最新趋势、快速调整方向并抓住新的商机。

Sponsors of Tomorrow.™ intel

大数据对 IT 部门的影响

大数据极具爆发力，为 IT 组织同时带来了机会和挑战。为发掘它的全部潜能，大数据分析需要使用全新方法来捕获、存储和分析数据。

三个 V 概括了大数据的主要特点，同时定义了 IT 部门需要解决的主要问题：

- **容量。**非结构化数据的大规模和增长超过了传统存储和分析解决方案的发展速度。
- **多样性。**可从之前从未考虑过的来源收集大数据。传统的数据管理流程无法处理异构和可变的大数据，这些数据可能来自不同的格式，如电子邮件、社交媒体、视频、图像、博客和传感器数据，以及“阴影数据”，如访问日志和网络搜索历史记录。
- **速度。**数据实时生成，同时要求按需提供可用信息。

这三个 V 的组合还推动了第四个因素：**价值。**对于任何希望成功地从大数据获取价值的企业来说，必须平行解决容量、多样性和速度问题。不全面的解决方案绝不可取。

基础设施挑战

Hadoop* 和 MapReduce 等新兴技术设计用于应对大数据的三个 V。它们还对支持非结构化数据分析的分布式处理的基础设施提出了很高要求，这些要求包括以下：

- 为大规模分布式数据密集型作业而设计的基础设施，将问题分布到整个集群服务器节点
- 经济高效的存储，足以捕获和存储 TB 级别（如果不是 PB 级别）的数据，拥有智能能力来减少数据足迹，如数据压缩、自动数据分层和重复数据删除
- 可快速导入大型数据集然后复制到各节点进行处理的网络基础设施
- 保护高度分布式基础设施和数据的安全能力
- 使用统计数据、算法、数据挖掘和可视化技术识别机会所需的人力资源技能组合

数据科学家的兴起

寻找技能熟练的人才是与大数据分析相关的主要挑战之一。成功的大数据分析计划要求 IT 部门、业务用户和“数据科学家”之间的紧密协作，以识别和实施可解决正确商业问题的分析。数据科学是一个新兴领域，同时数据科学家是拥有特殊技能的全新专业人员。数据科学家负责为复杂的业务问题建模、发现业务洞察并识别机会。对于这种能够将流入组织的大量数字信息流变成有用信息的人员，市场需求很大。

支持大数据分析的新兴技术

新技术正在不断出现，使得非结构化数据分析变得可行和经济高效。通过充分利用计算资源的分布式网格的能力，新方法重新定义了管理和分析数据的方式。它使用了可轻松扩展的“无共享”（SNA）架构、分布式处理框架以及非关系和平行关系数据库。

无共享架构是无状态的，没有节点共享内存或磁盘存储，因硬件、数据管理和分析应用技术发展的融合而成为可能。

- **硬件架构。**商用服务器的集群（如基于英特尔® 至强® 处理器的服务器）为在整个分布式网格的大量并行处理提供了计算能力和速度。
- **分析应用架构。**新的数据处理系统通过管理和推送数据到单个节点、发送指示给联网服务器以并行运行、收集单个结果，然后重组数据以生成有意义的结果，从而确保计算网格正常运行。在驻留地点处理数据比首先传输数据到集中系统进行处理更加快速高效。
- **数据架构。**为处理非结构化数据的多样性和复杂性，数据库从关系型转为非关系型。与结构化、规范化和密集填充的关系数据库不同，非关系数据库可扩展、以网络为主导、半结构化并松散填充。NoSQL 数据库解决方案无需固定表格模式，避免连接操作并可水平扩展。

分布式框架：Apache* Hadoop* 的出现

Apache* Hadoop 正在演进为非结构化数据分析的最佳新兴方法。Hadoop 是一个开源架构，使用简单的编程模型以允许在计算机集群中分布式处理大数据集。完整的技术堆栈包括常用设施、分布式文件系统、分析和数据存储平台，以及管理分布式处理、并行计算、工作流程和配置管理的应用层。除了提供高可用性之外，较传统方法相比，Hadoop 是一种更加经济高效的大型非结构化数据集处理方法，同时提供极大的可扩展性和速度。

随着越来越多的企业意识到与大数据相关的价值和优势，Hadoop 的采用正不断增长。Apache 在 2012 年 1 月推出了 Apache Hadoop 1.0 的首个完整生产版本。有关 Hadoop 部署的详细信息请参见*《在英特尔® 平台上进行云设计与部署的英特尔® 云构建计划指南：Apache* Hadoop》**。

Hadoop 生态系统

Hadoop 的商用版本也呈增长趋势。Hadoop 生态系统是一个复杂的厂商和解决方案的联合，包括老牌厂商和若干新厂商。众多厂商都提供了他们自己的 Hadoop 分发，并集合了其他 Hadoop 项目的基本堆栈，如 Hive*、Pig* 和 Chukwa*。其中一些分发可与数据仓库、数据库和其他数据管理产品集成，允许分析引擎访问和查询多个来源的数据。

Hadoop 基础设施：大数据存储和网络

Hadoop 集群通过主流计算和存储资源的极大改进而成为可能，并补充了万兆位以太网（10 GbE）解决方案。10 GbE 带来的带宽增长是导入和复制（在多台服务器之间）大型数据集的关键。英特尔® Ethernet 10 Gigabit 融合网络适配器提供了高吞吐量连接，同时英特尔 SATA 固态硬盘为原始存储提供了高性能、高吞吐量存储选择。为提高效率，存储需要支持其它高级能力，如压缩、加密、自动数据分层、重复数据删除、纠删码和自动精简配置 — 现有的英特尔至强® 处理器 E5 系统都支持这些功能。

大数据和云的情况如何?

随着云计算的出现，组织现在可访问他们自己的联网服务器数据中心和 Amazon* 网络服务等公共云基础设施服务中的大型社区计算机网格。在大数据时代，云为数据分析提供了潜在的自助计算模型。云计算和大数据分析都是虚拟化技术和网格计算模型的延伸，使得云成为可以远低于传统数据平台的成本提供业务支持的灵活数据平台。Hadoop 正快速演进为云中大数据的实际框架。

英特尔如何提供帮助

英特尔是一家提供数据中心基础设施 — 服务器、网络、存储、数据库和数据仓库 — 以及相关技术支持的公司，可通过以下方面助您实施大数据分析：

- 提供专为面向大数据分析项目扩展而设计的优化技术
- 帮助您更加快速地推进您的全新大数据分析项目
- 使用先进的分布式分析应对明天的挑战

可供了解更多信息的英特尔资源

英特尔 IT 中心提供了简单、无误、公正的信息，介绍了英特尔可帮助 IT 专业人员部署大数据分析等战略项目的方法。有关大数据分析的规划指南、同行研究、真实客户参考、厂商亮点和现场活动的信息，请访问 intel.com/bigdata

1 “Google Chief Eric Schmidt on the Data Explosion”。*首席信息官的 I-Global 智能（2010 年 8 月 4 日）*。 www.i-cio.com/features/august-2010/eric-schmidt-exabytes-of-data

2 “Big Data Infographic and Gartner 2012 Top 10 Strategic Tech Trends”。*Business Analytics 3.0*（博客）（2011 年 11 月 11 日） practicalanalytics.wordpress.com/2011/11/11/big-data-infographic-and-gartner-2012-top-10-strategic-tech-trends/

3 “Extracting Value from Chaos”。*IDC IView*，EMC 公司（2011 年 6 月）。www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

与同事分享

0612/RF/ME/PDF-CN 327439-001

Sponsors of Tomorrow.™