大数据时代降临

钟婧

随着网络信息化时代的日益普遍，移动互联、社交网络、电子商务大大拓展了互联网的疆界和应用领域，我们正处在一个数据爆炸性增长的"大数据"时代，大数据在社会经济、政治、文化，人们生活等方面产生深远的影响，大数据时代对人类的数据驾驭能力提出了新的挑战与机遇。 “大数据”作为时下最火热的IT行业的词汇，随之而来的数据仓库、数据安全、数据分析、数据挖掘等等围绕大数据的商业价值的利用逐渐成为行业人士争相追捧的利润焦点。

大数据是指:无法用现有的软件工具提取、存储、搜索、共享、分析和处理的海量的、复杂的数据集合。大数据的主要特征为：  
  一是数据体量巨大。截至目前，人类生产的所有印刷材料的数据量是200PB(1PB=210TB)，而历史上全人类说过的所有的话的数据量大约是5EB。当前，典型个人计算机硬盘的容量为TB量级，而一些大企业的数据量已经接近EB量级。  
  二是数据类型繁多。这种类型的多样性也让数据被分为结构化数据和非结构化数据。相对于以往便于存储的以文本为主的结构化数据，非结构化数据越来越多，包括网络日志、音频、视频、图片、地理位置信息等，这些多类型的数据对数据的处理能力提出了更高要求。  
  三是价值密度低。价值密度的高低与数据总量的大小成反比。以视频为例，一部1小时的视频，在连续不间断的监控中，有用数据可能仅有一二秒。如何通过强大的机器算法更迅速地完成数据的价值“提纯”成为目前大数据背景下亟待解决的难题。  
 四是速度快。这是大数据区分于传统数据挖掘的最显著特征。根据IDC的“数字宇宙”的报告，预计到2020年，全球数据使用量将达到35.2ZB。在如此海量的数据面前，处理数据的效率就是企业的生命。

大数据主要可分为：数据采集，数据存取，基础架构，数据处理，统计分析，数据挖掘，模型预测，结果呈现等8种技术。主要形成了批处理、流处理和交互分析三种计算模式。大数据的处理方法有很多，普遍适用的大数据处理流程，可以概括为四步，分别是采集、导入和预处理、统计和分析，最后是数据挖掘。1、采集。大数据的采集是指利用多个数据库来接收发自客户端的数据，并且用户可以通过这些数据库来进行简单的查询和处理工作。2、导入/预处理。虽然采集端本身会有很多数据库，但是如果要对这些海量数据进行有效的分析，还是应该将这些来自前端的数据导入到一个集中的大型分布式数据库，或者分布式存储集群，并且可以在导入基础上做一些简单的清洗和预处理工作。导入与预处理过程的特点和挑战主要是导入的数据量大，每秒钟的导入量经常会达到百兆，甚至千兆级别。3、统计分析。统计与分析主要利用分布式数据库，或者分布式计算集群来对存储于其内的海量数据进行普通的分析和分类汇总等，以满足大多数常见的分析需求。统计与分析这部分的主要特点和挑战是分析涉及的数据量大，其对系统资源，特别是I/O会有极大的占用。4、挖掘。与前面统计和分析过程不同的是，数据挖掘一般没有什么预先设定好的主题，主要是在现有数据上面进行基于各种算法的计算，从而起到预测的效果，从而实现一些高级别数据分析的需求。

大数据的应用范围非常广。有机构预测，“大数据”的发展，将使零售业净利润增长60%以上，制造业的产品开发、组装成本将下降50%以上。制造行业，企业通过对网上数据分析了解客户需求和掌握市场动向，并对大数据进行分析后，就可以有效实现对采购和合理库存量的管理，大大减少因盲目进货而导致销售损失。在商业上，国外一些超市利用对手机的定位和购物推车获得商场内顾客在各处停留时间，利用视频监视图像软件分析顾客购物行为，优化商场布局和货架排列。在政府决策上，分析几十年来的天气数据，将各地降雨、气温、土壤状况和历年农作物产量做成精密图表，就可以预测农产品生产趋势等等。

大数据，是思维、技术与数据的三足鼎立。大数据不仅指规模庞大的数据，它首先是一种思维方式的变化，其次是对这些数据的处理和应用，是数据、处理技术与应用三者的统一的一列处理技术，最后，大数据的前提必然是充裕互通的数据本身。