1.2大数据的4V特征、关键技术

那么究竟什么是大数据呢，对大数据不同的人有着不同的理解，下面给出几种有关大数据的权威定义：维基百科认为：大数据是指利用常用软件工具捕获、管理和处理数据所耗时间超过可容忍时间的数据集。百度搜索的定义为：“大数据”是一个特别大的数据集，并且这样的数据集无法用传统数据库工具对其内容进行抓取、管理和处理。互联网周刊的定义为："大数据"的概念远不止大量的数据（TB）和处理大量数据的技术，或者所谓的"4个V"之类的简单概念，而是涵盖了人们在大规模数据的基础上可以做的事情，而这些事情在小规模数据的基础上是无法实现的。换句话说，大数据让我们以一种前所未有的方式，通过对海量数据进行分析，获得有巨大价值的产品和服务，或深刻的洞见，最终形成变革之力。

IBM给出了大数据的3V定义：规模性、多样性、高速性。而谷歌更是提出了大数据的4V特征：

第一个特征是数据量大，据IDC估测，数据以每年50%速度增长，每两年增长一倍（大数据摩尔定律），人类在最近两年产生的数据量≈之前产生全部数据量总和。

预计到2020年，全球将总共拥有35ZB的数据量。

相较于2010年，数据量将增长近30倍。

存储单位从过去的GB到TB，直至PB、EB。随着信息技术的高速发展，数据开始爆发性增长。社交网络（微博、推特、脸书）、移动网络、各种智能终端等，都成为数据的来源。淘宝网近4亿的会员每天产生的商品交易数据约20TB；脸书约10亿的用户每天产生的日志数据超过300TB。迫切需要智能的算法、强大的数据处理平台和新的数据处理技术，来统计、分析、预测和实时处理如此大规模的数据。

第二个V是指数据类型繁多：大数据与人类信息密切相关，数据被分为结构化数据和非结构化数据。相对于以往便储存的以数据库/文本为主的结构变化数据，人类信息90%都是非结构化数据，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。

典型的人为生成的非结构化数据包括：

文本文件：文字处理、电子表格、演示文稿、电子邮件、日志。

电子邮件：电子邮件由于其元数据而具有一些内部结构，我们有时将其称为半结构化。但是，消息字段是非结构化的，传统的分析工具无法解析它。

社交媒体：来自新浪微博、微信、QQ、Facebook，Twitter，LinkedIn等平台的数据。

网站： YouTube，Instagram，照片共享网站。

移动数据：短信、位置等。

通讯：聊天、即时消息、电话录音、协作软件等。

媒体：MP3、数码照片、音频文件、视频文件。

业务应用程序：MS Office文档、生产力应用程序。

典型的机器生成的非结构化数据包括：

卫星图像：天气数据、地形、军事活动。

科学数据：石油和天然气勘探、空间勘探、地震图像、大气数据。

数字监控：监控照片和视频。

传感器数据：交通、天气、海洋传感器。

在一分钟之内可以产生多少数据：新浪可以发送 2 万条微博，人人网可以发生 30 万次访问，苹果可以下载 4.7 万次应用，百度可以产生 90 万次搜索查询，淘宝可以卖出 6 万件商品。对于快速产生的数据，我们同样也需要快速的处理和分析

大数据的第三个V：处理速度快 (Velocity）：处理速度到底有多快呢？一般从数据的生成到消耗，时间窗口非常小，可用于生成决策的时间非常少。

1秒定律”或者秒级定律,就是说对处理速度有要求,一般要在秒级时间范围内给出分析结果,时间太长就失去价值了.这个速度要求是大数据处理技术和传统的数据挖掘技术最大的区别.谷歌(微博)已经开发出更新的技术Dremel，这是一种用来分析信息的方法，它可以在数以千计的服务器上运行，能以极快的速度处理网络规模的海量数据，从而让“大数据”看起来变小。Dremel可在大约3秒钟时间里处理1PB的数据查询请求。

大数据的第4个V是价值密度低，商业价值高。这个概念有点抽象，怎么去理解呢，大数据就是一个海量的数据，在大海中捞

金子，这金子就是我们的宝藏。但我们把这块金子经过一系列的分析处理过程之后，我们就能确定是在某一平方米的水域，那么这个密度就会高很多了，这块金子就分布在这一平方米中，在这一块区域去捞金子那么就容易得多了。以视频为例，连续不间断监控过程中，可能有用的数据仅仅有一两秒，但是具有很高的商业价值。

大数据技术，就是从各种类型的数据中快速获得有价值信息的技术。大数据领域已经涌现出了大量新的技术，它们成为大数据采集、存储、处理和展现的有力武器。