Foundations of Big Data

在这个数字世界中，每个人都留下了痕迹。

从我们的旅行习惯到我们的锻炼和娱乐，越来越多的互联网

我们每天与之交互的连接设备记录了大量数据

关于我们。

甚至还有一个名字：大数据。

安永会计师事务所提供以下定义：“大数据是指动态，大数据和

由人员，工具和机器创建的不同数量的数据。

它需要新的，创新的和可扩展的技术来收集，托管和分析

处理收集的大量数据以获取实时业务洞察

与消费者，风险，利润，绩效，生产力管理和增强的股东有关

值。”

大数据没有一个定义，但是有一些共同的元素

涵盖不同的定义，例如速度，数量，种类，准确性和价值。

这些是大数据的V。

速度是数据累积的速度。

数据以极快的速度生成，并且永不停止。

近或实时流，本地和基于云的技术可以非常快速地处理信息

很快。

卷是数据的规模或存储的数据量的增加。

数量驱动因素是数据源，高分辨率传感器和可扩展性的增加

基础设施。

多样性是数据的多样性。

在非结构化数据中，结构化数据可以很好地适合关系数据库中的行和列

数据不是以预定义的方式组织的，例如推文，博客文章，图片，数字，

和视频。

多样性还反映出数据来自不同的来源，机器，人员和流程，

组织内部和外部。

驱动因素包括移动技术，社交媒体，可穿戴技术，地理技术，视频，

还有很多很多

准确性是数据的质量和来源，以及它与事实和准确性的一致性。

属性包括一致性，完整性，完整性和歧义性。

驱动因素包括成本和可追溯性的需求。

有了大量可用数据，有关数据准确性的争论越来越激烈

在数字时代。

信息是真实的还是虚假的？

价值是我们的能力，需要将数据转化为价值。

价值不只是利润。

它可能具有医疗或社会福利，以及客户，员工或个人满意度。

人们花时间去理解大数据的主要原因是从中获取价值

它。

让我们看一些V起作用的例子。

速度：每60秒钟，数小时的素材会上传到YouTube，

数据。

考虑一下数小时，数天甚至数年内数据累积的速度。

数量：世界人口约为70亿，其中绝大多数

现在正在使用数字设备； 手机，台式机和笔记本电脑，可穿戴设备，

等等。

这些设备都生成，捕获和存储数据-大约2.5兆字节

每天。

这相当于一千万个蓝光DVD。

多样性：让我们考虑一下不同类型的数据。 文字，图片，电影，声音，健康

来自可穿戴设备的数据，以及来自连接到移动设备的设备的许多不同类型的数据

物联网。

准确性：80％的数据被认为是非结构化的，我们必须设计出产生数据的方法

可靠而准确的见解。

必须对数据进行分类，分析和可视化。

如今，数据科学家从大数据中获取见解并应对挑战

这些庞大的数据集存在。

所收集数据的规模意味着使用常规方法是不可行的

数据分析工具。

但是，利用分布式计算能力的替代工具可以解决此问题。

Apache Spark，Hadoop及其生态系统等工具提供了提取，加载，分析，

并跨分布式计算资源处理数据，从而提供新的见解和知识。

这为组织提供了更多与客户联系并丰富服务的方法

他们提供。

因此，下次您系上智能手表，解锁智能手机或跟踪锻炼时，

请记住，您的数据正在开始一段可能遍及全球的旅程，

通过大数据分析，然后返回给您。