大数据的思考

3170103240 张佳瑶

一、 大数据概念

大数据是规模宏大、种类丰富、形式多变的数据集。从体积上看，大数据的规模是以往数据量的数倍，甚至是不能计算的大小。在不同时代，用来定义“大”的数据规模是不同的。在未来，现在的大数据可能已不再是大数据。但是利用有限的硬件资源处理远大于硬件承受能力的数据的任务一直在被执行。在不同行业，用来定义“大”的数据规模是不同的，定义数据是否有意义也是不同的。挖掘有价值的数据，凝炼数据中的价值是我们追求的目标。从速度上看，互联网产品每时每刻都在产生数据，是一个实时运作的数据源，数据量呈爆炸式指数级增长。相应的，把控数据的时间向价值，一个实时运作的大数据分析处理工具需要诞生。从种类上来看，文本、音频、视频和图片等结构化、半结构化、非结构化的数据一同存在。这些数据呈现低密度、非结构化的特征，处理这些数据耗时耗力。半结构化、非结构化的数据对数据预处理带来了挑战。

二、 大数据应用

大数据有存在的价值和意义。随着计算能力的不断提高，技术软件的不断进步，大数据帮助我们获取之前未知的信息，解决之前无能为力的问题。无论是作为企业，还是作为消费者，都能感受到大数据技术革新带来的高效便利。

大数据作为信息技术产业之一，国家制定了大数据战略和发展行动纲要推动其跨越发展。国务院印发的《“十三五”国家战略性新兴产业发展规划》中提到：“信息革命进程持续快速演进，物联网、云计算、大数据、人工智能等技术广泛渗透于经济社会各个领域，信息经济繁荣程度成为国家实力的重要标志。”在我们生活中，一个资源共享、金融、工商登记、社保缴费等信息集成的统一平台已经建成，“智慧城市”建设也正在如火如荼地开展。

如今，大数据已经成为一种重要的生产要素，一种必不可少的基础资源，隐藏着大量的经济利益。公共设施、国防安全、金融经济等领域都有基于大数据的应用。大数据公司层出不穷。Google、IBM、亚马逊等国际知名企业都是大数据行业的主要推动者，相继推出大数据产品。国内的大数据企业代表有百度、阿里巴巴、腾讯。阿里建设阿里云，提供 ODPS 在线服务。精准营销是大数据最早成熟的应用领域。互联网上有大量的用户，积累了海量的数据。商家通过搜集用户行为，分析内部数据，整合外部数据，构建用户画像，更好地了解用户需求，向用户投放精准广告，增强用户体验，最终提高自身的盈利。智能产品多为实时交互，大数据可以帮助人们完成快速高效、近乎实时的故障检测、问题分析。有时候打开手机上的淘宝 app，软件下方的商品推荐恰好是自己感兴趣的，因为淘宝会根据购买记录等信息制定个性化推荐。随着使用次数的增加，百度新闻推送越来越符合理想。科大讯飞将海量数据用于训练语音识别模型，准确率高达 97%，实现实时语音转写和翻译。

三、 大数据工具

这些海量的数据不能够用传统的数据处理软件来采集、储存、管理和分析。大数据工具正在更新换代。

数据采集技术对数据进行 ETL 操作：抽取、转换、加载。采集到的数据可以来源于系统日志。许多公司的业务平台会产生大量的日志记录，可以从中挖掘出有价值的信息。Scribe 是一个 Facebook 开发、汇总许多服务器实时流式传输日志数据的开源服务器。它从各种日志源上收集日志，存储到一个中央存储系统上，实现了“分布收集，统一处理”。网络上有大量用户留下来的数据，可以通过爬虫等方式获取。这些网络数据多为半结构化、非结构化的。Apache Nutch 是一个 Java 编写的分布式爬虫框架，具有高度可拓展、功能丰富的特点。传统的关系型数据库如 MySQL 等会被企业用来存放数据。数据库中的数据会由特定的系统进行分析处理。然而，SQL 数据库不适用于应对用户信息、地理位置、社交网络这些成倍增加

且不具备固定模式的数据。NoSQL 是一种全新的思维，提倡非关系型数据存储，为这种超大规模数据提供灵活的存储方式。

在大数据处理平台方面，Hadoop 是知名的大数据分析系统，是当下流行的大数据处理平台，具有可伸缩、高效的特点，能够处理 PB 级别的数据。它的思想之源来自于谷歌的三驾马车：GFS、MapReduce、BigTable。Hadoop 由两部分构成。Hadoop 的分布式文件系统用于底层文件管理。MapReduce 计算框架用于计算分析海量数据。基于 Hadoop，改进 Hadoop 是研究热点：提高 Hadoop 平台性能，建设更高效的查询处理，连接 Hadoop 和数据库系……

四、　　大数据思考

数据来源是否合法，数据是否涉及到用户的隐私，用户是否会敏感平台收集自己的数据……大数据技术的发展不可避免地伴随着一些问题，但是大数据技术不会停下革新的脚步。

云计算、物联网快速发展，数据洪流涌动，大数据时代已经到来。企业携手大数据，进行高密度分析，更深入地剖析目标用户，获取潜在的消费需求，制定精准的营销策略，降低风险，提高盈利。用户制造数据，使用大数据技术下的产品，体验愈加友好快捷的服务。政府利用大数据技术提高行政效率，建设更好的中国。机遇与挑战并存。硬件的存储能力、计算能力是有限的。能源价格上涨，大数据管理如何实现低能耗制约着其发展。应对大数据的挑战，我们要致力于挖掘新的数据，发现数据价值，创新数据应用，将大数据处理技术落于实地。