浅谈大数据和云计算

Xxx

1. **什么是大数据**

大数据是指那些量大且复杂到无法用传统数据处理方法来处理的数据集合，它有四大特点。

1. 体量大

大数据的数据量，包括采集，存储和计算的量都非常大。大数据的起始计量单位至少是P（1000个T）、E（100万个T）或Z（10亿个T）。百度资料表明，其新首页导航每天需要提供的数据超过1.5PB，这些数据如果打印出来将超过5千亿张A4纸。而到目前为止，人类生产的所有印刷材料的数据量也不过为200PB

1. 类型和来源多样化

相对于以往便于存储的以文本为主的结构化数据（即有固定格式和有限长度的数据。例如填的表格就是结构化的数据，国籍：中华人民共和国，民族：汉，性别：男，这都叫结构化数据），非结构化数据（无固定格式的数据）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。

1. 数据价值密度相对较低

大数据虽然量大，但是可以说是浪里淘沙却又弥足珍贵。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。随着互联网以及物联网的广泛应用，信息感知无处不在，信息海量，但价值密度较低，如何结合业务逻辑并通过强大的机器算法来挖掘数据价值，是大数据时代最需要解决的问题。

1. 数据增长速度快，处理速度快

这是大数据区别于传统数据挖掘的显著特征。大数据数据处理遵循“1秒定律”，可从各种类型的数据中快速获得高价值的信息。比如搜索引擎要求几分钟前的新闻能够被用户查询到，个性化推荐算法尽可能要求实时完成推荐。

1. **大数据带来的好处**

大数据的数据来源有很多种，像是POS机数据、信用卡刷卡数据、电子商务数据这样的交易数据。移动设备上软件追踪到的移动通信数据。人们收发电子邮件、文档、图片、音频、视频，以及通过微信、博客等社交媒体产生人为数据。这些数据大多数为非结构性数据，需要用文本分析功能进行分析。大数据的无处不在，对各行各业都提供了极大便利。

1. 电子商务方面

大数据通过对客户的订单信息进行分类整理，根据客户的购买习惯、年龄、喜好、地域等区分进行推荐产品，进行个性化的页面展示。还可以根据以往数据，来决定库存数量和物流资源的动态调整。

1. 交通出游方面

通过WIFI+ibeacon或基站定位技术，收集到个体的出行数据。政府可以对每个车站、机场、道路交通各时段的流量规律信息，进行针对性的安防及其他部署。景区可以根据不同景点的客流密度、游客的停留时长以及游览路径，来决定景区的运营方向。通过分析公交车上GPS定位系统每天的位置和时间数据，结合时刻表预测出每一辆公交车的到站时间。通过记录每位用户在行驶过程中的实时数据例如行驶车速，所在位置等信息并进行数据汇总分析，而后计算出最佳线路，让用户能够避开拥堵。

1. 在医疗健康方面

手上的健身腕带可以收集有关我们走路或者慢跑的数据，例如行走步数、卡路里消耗、睡眠时长等数据与健康记录来改善我们的健康状况。对医院里众多病人的体质特征，病症信息，病史信息，分布情况进行分析处理，可以实现流行疾病预测，为各类疾病的治疗方案优化提供数据支撑。在加拿大多伦多的一家医院，针对早产婴儿，每秒钟有超过3000次的数据读取。通过这些数据分析，医院能够提前知道哪些早产儿出现问题并且有针对性地采取措施，避免早产婴儿夭折。

1. 在咨询娱乐方面视频分析方面

通过收集用户的以往阅读信息，分析用户的阅读喜好，从而进行个性化信息精准推送。通过收集用户的观影喜好，进行票房预测等。互联网电视能够追踪你正在看的内容，看了多长时间，甚至能够识别多少人坐在电视机前，来确定这个频道的流行度。美国国内规模最大的商业视频流供应商Netflix，收集的数据包括用户在看什么、喜欢在什么时段观看、在哪里观看以及使用哪些设备观看等。甚至记录用户在哪视频的哪个时间点后退、快进或者暂停，乃至看到哪里直接将视频关掉等信息。Netflix公司曾利用数据说服BBC重新翻拍了电视连结剧《纸牌屋》，而且成功的挖掘出演员Kevin Spacey和导演David Fincher的支持者与原剧集粉丝的关联性，确定新剧拍摄的最佳人选。

1. 甚至奥巴马在总统竞选中都使用大数据分析来收集选民的数据，让他可以专注于对他最感兴趣的选民。
2. **大数据处理的几个步骤**
3. 数据的收集

数据的收集有爬取和推送两种，爬取即利用爬虫等技术把网上所有信息都下载到数据库中心。而推送是利用像是手环等这样的终端收集数据，再传到数据库里。

1. 数据的传输

一般会通过队列的方式进行，一个内存里面的队列肯定会被大量的数据挤爆掉，于是就产生了基于硬盘的分布式队列，这样队列可以多台机器同时传输，不管数据量多大，只要队列足够多，管道足够粗，就能够撑得住。

1. 数据的存储

随着结构化数据和非结构化数据量的持续增长，以及分析数据来源的多样化，此前存储系统的设计已经无法满足大数据应用的需要。存储厂商已经意识到这一点，他们开始修改基于块和文件的存储系统的架构设计以适应这些新的要求。一台机器的文件系统肯定是放不下的，所以需要一个很大的分布式文件系统来做这件事情，把多台机器的硬盘打成一块大的文件系统。

1. 数据的处理和分析

上面存储的数据是原始数据，原始数据多是杂乱无章的，有很多垃圾数据在里面，因而需要清洗和过滤，得到一些高质量的数据。对于高质量的数据，就可以进行分析，从而对数据进行分类，或者发现数据之间的相互关系，得到有效信息。对大量的数据做分解、统计、汇总，一台机器肯定搞不定，处理到猴年马月也分析不完。于是就有分布式计算的方法，将大量的数据分成小份，每台机器处理一小份，多台机器并行处理，很快就能算完。例如著名的Terasort对1个TB的数据排序，相当于1000G，如果单机处理，怎么也要几个小时，但并行处理209秒就完成了

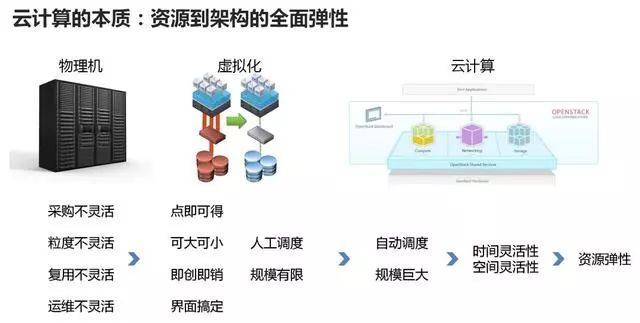
1. 数据的检索和挖掘

检索就是搜索，方便找寻想要的信息，就像是google和baidu这样的搜索引擎。另外就是挖掘，仅仅搜索出来已经不能满足人们的要求了，还需要从信息中挖掘出相互的关系。比如财经搜索，当搜索某个公司股票的时候，该公司的高管是不是也应该被挖掘出来呢？如果仅仅搜索出这个公司的股票发现涨的特别好，于是你就去买了，其实其高管发了一个声明，对股票十分不利，第二天就跌了。所以通过各种算法挖掘数据中的关系，形成知识库，十分重要

1. **云计算**

1.云计算与大数据

云计算最初的目标是对资源的管理，管理的主要是计算资源、网络资源、存储资源三个方面。一开始，为了解决数据处理的物理设备不灵活等问题，人们想出了一个方法——虚拟化。数据中心的物理设备都很强大，从物理的CPU、内存、硬盘中虚拟出一小块来给客户，同时也可以虚拟出一小块来给其他客户。每个客户只能看到自己的那一小块，但其实每个客户用的是整个大的设备上的一小块。这样空间灵活性和时间灵活性就基本解决了。但是随着集群规模的扩大，人工配置的过程越来越复杂，越来越耗时。另一方面也影响空间灵活性：当用户数量多时，这点集群规模，还远达不到想要多少要多少的程度，很可能这点资源很快就用完了，还得去采购。所以随着集群的规模越来越大，基本都是千台起步，动辄上万台、甚至几十上百万台。如果去查一下BAT，包括网易、谷歌、亚马逊，服务器数目都大的吓人。这么多机器要靠人去选一个位置放这台虚拟化的电脑并做相应的配置，几乎是不可能的事情，还是需要机器去做这个事情。于是人们发明了各种各样的算法来做这个事情，算法的名字叫做调度(Scheduler)。通俗一点说，就是有一个调度中心，几千台机器都在一个池子里面，无论用户需要多少CPU、内存、硬盘的虚拟电脑，调度中心会自动在大池子里面找一个能够满足用户需求的地方，把虚拟电脑启动起来做好配置，用户就直接能用了。这个阶段我们称为池化或者云化。到了这个阶段，才可以称为云计算，在这之前都只能叫虚拟化。



2.云计算与大数据

随着数据越来越大，很多小公司没有这么大的数据处理能力，这是该怎么办呢。 云计算就是为了来解决这些问题。云计算可以为大数据的运算提供资源层的灵活 性。而云计算也会部署大数据放到它的PaaS平台上，作为一个非常非常重要的通 用应用。现在公有云上基本上都会有大数据的解决方案了，一个小公司需要大数 据平台的时候，不需要采购一千台机器，只要到公有云上一点，这一千台机器都出来了，并且上面已经部署好了的大数据平台，只要把数据放进去算就可以了。云计算需要大数据，大数据需要云计算，二者就这样结合了。

1. **大数据带来的隐患**

不同行业的人对于大数据的理解是不同的，对于投资人和创业者而言，大数据是个热门的融资标签。对于大多数互联网公司或者工程师而言，大数据实际是各种运算和处理加上几个数据科学家搞点数据挖掘或者机器学习，但是对于消费者或者互联网用户来说大数据却是另外一个意思——尽可能地搜集跟终端消费者相关的隐私，然后进行营销。从理论上说，大数据公司通过搜集用户行为，可以更好地了解消费者的需求，增强用户体验。但是在实践上，这些所谓的智能推荐还停留在很初级的阶段，因此当我在淘宝上搜索大数据入门书籍结果在微博上不停地看到跟大数据培训班相关的广告。

大数据安全管理困难、安全意识薄弱等因素增加了大数据产业的发展风险。一是数据的海量存储增加了数据防护的难度。随着大数据域智慧城市、云计算、工业互联网等领域深度融合，大量数据一旦遭到损坏或丢失将造成巨大损失。二是政府、企业及个人都存在安全意识不足的问题。数据无序滥用将侵害个人隐私及企业商业秘密，甚至危害国家安全。网络空间数据安全问题的严重性、紧迫性，在很大程度上已超越其他传统的安全问题

1. **自身感受**

大数据带来的方便是能切身感受到的，打开一款音乐软件，它的每日推荐中每一首都是自己喜欢的歌，上网购物，总能给我推荐我最近一直想买的东西，而且它还会给我找到同类商品中性价比最高的店。但他的缺陷也摆在眼前，信息太多了，隐私太少了。感觉自己时刻处于被人监视的状态，在互联网的另一端有无数双眼睛注视着你的一举一动。我们似乎对垃圾邮件，骚扰电话已经习以为常，以至于忽视了信息泄露带来的危机和严重性。网上说技术总是一把双刃剑，风险与机遇并存。我们都知道在这个信息为王的时代，数据就是一座金矿，但很少有人意识到有金矿其实是不够的，提炼金子才是个技术活。国内很多公司不缺数据，却的是提炼金子的能力。他们守着金子干着急，赚不到钱，情急之下能做的恐怕也只有卖用户的隐私为生了吧。正如李彦宏所言：中国人愿意用隐私换便利，当然我们也要遵循一些原则，如果这个数据能让用户受益，他们又愿意给我们用，我们才会去使用它。而这个原则带入到用户的角度，那就是要在用户知情、允许的前提下，变向地用隐私换取服务。“懂你，但不认识你。”这便是大数据发展与个人隐私之间的一个平衡点，是数据生命周期中的最好状态

掌握着大量数据的企业，首先应该做的便是利用技术把握好这个平衡点，而非一心利用数据追求商业价值。只有突破了数据隐私保护的瓶颈，大数据企业才能迎来真正的春天。