**目录**

1 引言

大数据是什么？在写这篇文章之前，我发现身边很多IT人对于这些热门的新技术、新趋势往往趋之若鹜却又很难说的透彻，如果你问他大数据是什么？估计很少能说出一二三来。究其原因，一是因为大家对大数据这类新技术有着相同的原始渴求，至少知其然在聊天时不会显得很“土鳖”；二是在工作和生活环境中真正能参与实践大数据的案例实在太少了，所以大家没有必要花时间去知其所以然。

我希望有些不一样，所以对该如何去认识大数据进行了一番思索，包括查阅了资料，翻阅了最新的专业书籍，但我并不想把那些零散的资料碎片或不同理解论述简单规整并堆积起来形成毫无价值的转述或评论，我很真诚的希望进入事物探寻本质。

如果你说大数据就是数据大，或者侃侃而谈4个V，也许很有深度的谈到BI或预测的价值，又或者拿Google和Amazon举例，技术流可能会聊起hadoop和Cloud Computing，不管对错，只是无法勾勒对大数据的整体认识，不说是片面，但至少有些管窥蠡测、隔衣瘙痒了。……也许，“解构”是最好的方法。

2 大数据结构

首先，我认为大数据就是互联网发展到现今阶段的一种表象或特征而已，没有必要神话它或对它保持敬畏之心，在以云计算为代表的技术创新大幕的衬托下，这些原本很难收集和使用的数据开始容易被利用起来了，通过各行各业的不断创新，大数据会逐步为人类创造更多的价值。

其次，想要系统的认知大数据，必须要全面而细致的分解它，我着手从三个层面来展开：

第一层面是理论，理论是认知的必经途径，也是被广泛认同和传播的基线。我会从大数据的特征定义理解行业对大数据的整体描绘和定性；从对大数据价值的探讨来深入解析大数据的珍贵所在；从对大数据的现在和未来去洞悉大数据的发展趋势；从大数据隐私这个特别而重要的视角审视人和数据之间的长久博弈。

第二层面是技术，技术是大数据价值体现的手段和前进的基石。我将分别从云计算、分布式处理技术、存储技术和感知技术的发展来说明大数据从采集、处理、存储到形成结果的整个过程。

第三层面是实践，实践是大数据的最终价值体现。我将分别从互联网的大数据，政府的大数据，企业的大数据和个人的大数据四个方面来描绘大数据已经展现的美好景象及即将实现的蓝图。

3 大数据相关的理论

最早提出大数据时代到来的是麦肯锡：“数据，已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。”

业界（IBM 最早定义）将大数据的特征归纳为4个“V”（量Volume，多样Variety，价值Value，速度Velocity），或者说特点有四个层面：第一，数据体量巨大。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）；第二，数据类型繁多。比如，网络日志、视频、图片、地理位置信息等等。第三，价值密度低，商业价值高。第四，处理速度快。最后这一点也是和传统的数据挖掘技术有着本质的不同。

其实这些V并不能真正说清楚大数据的所有特征，下面这张图对大数据的一些相关特性做出了有效的说明。

古语云：三分技术，七分数据，得数据者得天下。先不论谁说的，但是这句话的正确性已经不用去论证了。维克托·迈尔-舍恩伯格在《大数据时代》一书中举了百般例证，都是为了说明一个道理：在大数据时代已经到来的时候要用大数据思维去发掘大数据的潜在价值。书中，作者提及最多的是Google如何利用人们的搜索记录挖掘数据二次利用价值，比如预测某地流感爆发的趋势；Amazon如何利用用户的购买和浏览历史数据进行有针对性的书籍购买推荐，以此有效提升销售量；Farecast如何利用过去十年所有的航线机票价格打折数据，来预测用户购买机票的时机是否合适。

那么，什么是大数据思维？维克托·迈尔-舍恩伯格认为：需要全部数据样本而不是抽样；关注效率而不是精确度；关注相关性而不是因果关系。

阿里巴巴的王坚对于大数据也有一些独特的见解，比如，

“今天的数据不是大，真正有意思的是数据变得在线了，这个恰恰是互联网的特点。”

“非互联网时期的产品，功能一定是它的价值，今天互联网的产品，数据一定是它的价值。”

“你千万不要想着拿数据去改进一个业务，这不是大数据。你一定是去做了一件以前做不了的事情。”

特别是最后一点，我是非常认同的，大数据的真正价值在于创造，在于填补无数个还未实现过的空白。

有人把数据比喻为蕴藏能量的煤矿。煤炭按照性质有焦煤、无烟煤、肥煤、贫煤等分类，而露天煤矿、深山煤矿的挖掘成本又不一样。与此类似，大数据并不在“大”，而在于“有用”。价值含量、挖掘成本比数量更为重要。

价值探讨

大数据是什么？投资者眼里是金光闪闪的两个字：资产。比如，Facebook上市时，评估机构评定的有效资产中大部分都是其社交网站上的数据。

如果把大数据比作一种产业，那么这种产业实现盈利的关键，在于提高对数据的“加工能力”，通过“加工”实现数据的“增值”。

Target 超市以20多种怀孕期间孕妇可能会购买的商品为基础，将所有用户的购买记录作为数据来源，通过构建模型分析购买者的行为相关性，能准确的推断出孕妇的具体临盆时间，这样Target的销售部门就可以有针对的在每个怀孕顾客的不同阶段寄送相应的产品优惠卷。

Target的例子是一个很典型的案例，这样印证了维克托·迈尔-舍恩伯格提过的一个很有指导意义的观点：通过找出一个关联物并监控它，就可以预测未来。Target通过监测购买者购买商品的时间和品种来准确预测顾客的孕期，这就是对数据的二次利用的典型案例。如果，我们通过采集驾驶员手机的GPS数据，就可以分析出当前哪些道路正在堵车，并可以及时发布道路交通提醒；通过采集汽车的GPS位置数据，就可以分析城市的哪些区域停车较多，这也代表该区域有着较为活跃的人群，这些分析数据适合卖给广告投放商。

不管大数据的核心价值是不是预测，但是基于大数据形成决策的模式已经为不少的企业带来了盈利和声誉。

从大数据的价值链条来分析，存在三种模式：

手握大数据，但是没有利用好；比较典型的是金融机构，电信行业，政府机构等。

没有数据，但是知道如何帮助有数据的人利用它；比较典型的是IT咨询和服务企业，比如，埃森哲，IBM，Oracle等。

既有数据，又有大数据思维；比较典型的是Google，Amazon，Mastercard等。

未来在大数据领域最具有价值的是两种事物：1-拥有大数据思维的人，这种人可以将大数据的潜在价值转化为实际利益；2-还未有被大数据触及过的业务领域。这些是还未被挖掘的油井，金矿，是所谓的蓝海。

Wal-Mart作为零售行业的巨头，他们的分析人员会对每个阶段的销售记录进行了全面的分析，有一次他们无意中发现虽不相关但很有价值的数据，在美国的飓风来临季节，超市的蛋挞和抵御飓风物品竟然销量都有大幅增加，于是他们做了一个明智决策，就是将蛋挞的销售位置移到了飓风物品销售区域旁边，看起来是为了方便用户挑选，但是没有想到蛋挞的销量因此又提高了很多。

还有一个有趣的例子，1948年辽沈战役期间，司令员林彪要求每天要进行例常的“每日军情汇报”，由值班参谋读出下属各个纵队、师、团用电台报告的当日战况和缴获情况。那几乎是重复着千篇一律枯燥无味的数据：每支部队歼敌多少、俘虏多少;缴获的火炮、车辆多少，枪支、物资多少……有一天，参谋照例汇报当日的战况，林彪突然打断他：“刚才念的在胡家窝棚那个战斗的缴获，你们听到了吗?”大家都很茫然，因为如此战斗每天都有几十起，不都是差不多一模一样的枯燥数字吗?林彪扫视一周，见无人回答，便接连问了三句：“为什么那里缴获的短枪与长枪的比例比其它战斗略高?”“为什么那里缴获和击毁的小车与大车的比例比其它战斗略高?”“为什么在那里俘虏和击毙的军官与士兵的比例比其它战斗略高?”林彪司令员大步走向挂满军用地图的墙壁，指着地图上的那个点说：“我猜想，不，我断定!敌人的指挥所就在这里!”果然，部队很快就抓住了敌方的指挥官廖耀湘，并取得这场重要战役的胜利。

这些例子真实的反映在各行各业，探求数据价值取决于把握数据的人，关键是人的数据思维；与其说是大数据创造了价值，不如说是大数据思维触发了新的价值增长。

现在和未来

我们先看看大数据在当下有怎样的杰出表现：

大数据帮助政府实现市场经济调控、公共卫生安全防范、灾难预警、社会舆论监督；

大数据帮助城市预防犯罪，实现智慧交通，提升紧急应急能力；

大数据帮助医疗机构建立患者的疾病风险跟踪机制，帮助医药企业提升药品的临床使用效果，帮助艾滋病研究机构为患者提供定制的药物；

大数据帮助航空公司节省运营成本，帮助电信企业实现售后服务质量提升，帮助保险企业识别欺诈骗保行为，帮助快递公司监测分析运输车辆的故障险情以提前预警维修，帮助电力公司有效识别预警即将发生故障的设备；

大数据帮助电商公司向用户推荐商品和服务，帮助旅游网站为旅游者提供心仪的旅游路线，帮助二手市场的买卖双方找到最合适的交易目标，帮助用户找到最合适的商品购买时期、商家和最优惠价格；

大数据帮助企业提升营销的针对性，降低物流和库存的成本，减少投资的风险，以及帮助企业提升广告投放精准度；

大数据帮助娱乐行业预测歌手，歌曲，电影，电视剧的受欢迎程度，并为投资者分析评估拍一部电影需要投入多少钱才最合适，否则就有可能收不回成本；

大数据帮助社交网站提供更准确的好友推荐，为用户提供更精准的企业招聘信息，向用户推荐可能喜欢的游戏以及适合购买的商品。

其实，这些还远远不够，未来大数据的身影应该无处不在，就算无法准确预测大数据终会将人类社会带往到哪种最终形态，但我相信只要发展脚步在继续，因大数据而产生的变革浪潮将很快淹没地球的每一个角落。

比如，Amazon的最终期望是：“最成功的书籍推荐应该只有一本书，就是用户要买的下一本书。”

Google也希望当用户在搜索时，最好的体验是搜索结果只包含用户所需要的内容，而这并不需要用户给予Google太多的提示。

而当物联网发展到达一定规模时，借助条形码、二维码、RFID等能够唯一标识产品，传感器、可穿戴设备、智能感知、视频采集、增强现实等技术可实现实时的信息采集和分析，这些数据能够支撑智慧城市，智慧交通，智慧能源，智慧医疗，智慧环保的理念需要，这些都所谓的智慧将是大数据的采集数据来源和服务范围。

未来的大数据除了将更好的解决社会问题，商业营销问题，科学技术问题，还有一个可预见的趋势是以人为本的大数据方针。人才是地球的主宰，大部分的数据都与人类有关，要通过大数据解决人的问题。

比如，建立个人的数据中心，将每个人的日常生活习惯，身体体征，社会网络，知识能力，爱好性情，疾病嗜好，情绪波动……换言之就是记录人从出生那一刻起的每一分每一秒，将除了思维外的一切都储存下来，这些数据可以被充分的利用：

医疗机构将实时的监测用户的身体健康状况；

教育机构更有针对的制定用户喜欢的教育培训计划；

服务行业为用户提供即时健康的符合用户生活习惯的食物和其它服务；

社交网络能为你提供合适的交友对象，并为志同道合的人群组织各种聚会活动；

政府能在用户的心理健康出现问题时有效的干预，防范自杀，刑事案件的发生；

金融机构能帮助用户进行有效的理财管理，为用户的资金提供更有效的使用建议和规划；

道路交通、汽车租赁及运输行业可以为用户提供更合适的出行线路和路途服务安排；

当然，上面的一切看起来都很美好，但是否是以牺牲了用户的自由为前提呢？只能说当新鲜事物带来了革新的同时也同样带来了“病菌”。比如，在手机未普及前，大家喜欢聚在一起聊天，自从手机普及后特别是有了互联网，大家不用聚在一起也可以随时随地的聊天，只是“病菌”滋生了另外一种情形，大家慢慢习惯了和手机共渡时光，人与人之间情感交流仿佛永远隔着一张“网”。

大数据隐私

你或许并不敏感，当你在不同的网站上注册了个人信息后，可能这些信息已经被扩散出去了，当你莫名其妙的接到各种邮件，电话，短信的滋扰时，你不会想到自己的电话号码，邮箱，生日，购买记录，收入水平，家庭住址，亲朋好友等私人信息早就被各种商业机构非法存储或贱卖给其它任何有需要的企业或个人了。

更可怕的是，这些信息你永远无法删除，它们永远存在于互联网的某些你不知道的角落。除非你更换掉自己的所有信息，但是这代价太大了。

用户隐私问题一直是大数据应用难以绕开的一个问题，如被央视曝光过的分众无线、罗维邓白氏以及网易邮箱都涉及侵犯用户隐私。目前，中国并没有专门的法律法规来界定用户隐私，处理相关问题时多采用其他相关法规条例来解释。但随着民众隐私意识的日益增强，合法合规地获取数据、分析数据和应用数据，是进行大数据分析时必须遵循的原则。

说到隐私被侵犯，爱德华?斯诺登应该占据一席之地，这位前美国中央情报局（CIA）雇员一手引爆了美国“棱镜计划”（PRISM）的内幕消息。“棱镜”项目是一项由美国国家安全局（NSA）自2007年起开始实施的绝密电子监听计划，年耗资近2000亿美元，用于监听全美电话通话记录，据称还可以使情报人员通过“后门”进入9家主要科技公司的服务器，包括微软、雅虎、谷歌、Facebook、PalTalk、美国在线、Skype、YouTube、苹果。这个事件引发了人们对政府使用大数据时对公民隐私侵犯的担心。

再看看我们身边，当微博，微信，QQ空间这些社交平台肆意的吞噬着数亿用户的各种信息时，你就不要指望你还有隐私权了，就算你在某个地方删除了，但也许这些信息已经被其他人转载或保存了，更有可能已经被百度或Google存为快照，早就提供给任意用户搜索了。

因此在大数据的背景下，很多人都在积极的抵制无底线的数字化，这种大数据和个体之间的博弈还会一直继续下去……

专家给予了我们一些如何有效保护大数据背景下隐私权的建议：减少信息的数字化；隐私权立法；数字隐私权基础设施（类似DRM数字版权管理）；人类改变认知（接受忽略过去）；创造良性的信息生态；语境化。

但是这些都很难立即见效或者有实质性的改善。

比如，现在有一种职业叫删帖人，专门负责帮人到各大网站删帖，删除评论。其实这些人就是通过黑客技术侵入各大网站，破获管理员的密码然后进行手工定向删除。只不过他们保护的不是客户的隐私，而大多是丑闻。还有一种职业叫人肉专家，他们负责从互联网上找到一个与他们根本就无关系用户的任意信息。这是很可怕的事情，也就是说，如果有人想找到你，只需要两个条件：1-你上过网，留下过痕迹；2-你的亲朋好友或仅仅是认识你的人上过网，留下过你的痕迹。这两个条件满足其一，人肉专家就可以很轻松的找到你，可能还知道你现在正在某个餐厅和谁一起共进晚餐。

当很多互联网企业意识到隐私对于用户的重要性时，为了继续得到用户的信任，他们采取了很多办法，比如google承诺仅保留用户的搜索记录9个月，浏览器厂商提供了无痕冲浪模式，社交网站拒绝公共搜索引擎的爬虫进入，并将提供出去的数据全部采取匿名方式处理等。

在这种复杂的环境里面，很多人依然没有建立对于信息隐私的保护意识，让自己一直处于被滋扰，被精心设计，被利用，被监视的处境中。可是，我们能做的几乎微乎其微，因为个人隐私数据已经无法由我们自己掌控了，就像一首诗里说到的：“如果你现在继续麻木，那就别指望这麻木能抵挡得住被”扒光”那一刻的惊恐和绝望……”

和大数据相关的技术

云技术

大数据常和云计算联系到一起，因为实时的大型数据集分析需要分布式处理框架来向数十、数百或甚至数万的电脑分配工作。可以说，云计算充当了工业革命时期的发动机的角色，而大数据则是电。

云计算思想的起源是麦卡锡在上世纪60年代提出的：把计算能力作为一种像水和电一样的公用事业提供给用户。

如今，在Google、Amazon、Facebook等一批互联网企业引领下，一种行之有效的模式出现了：云计算提供基础架构平台，大数据应用运行在这个平台上。

业内是这么形容两者的关系：没有大数据的信息积淀，则云计算的计算能力再强大，也难以找到用武之地；没有云计算的处理能力，则大数据的信息积淀再丰富，也终究只是镜花水月。

4 大数据需要的云计算技术

这里暂且列举一些，比如虚拟化技术，分布式处理技术，海量数据的存储和管理技术，NoSQL、实时流数据处理、智能分析技术（类似模式识别以及自然语言理解）等。

云计算和大数据之间的关系可以用下面的一张图来说明，两者之间结合后会产生如下效应：可以提供更多基于海量业务数据的创新型服务；通过云计算技术的不断发展降低大数据业务的创新成本。

如果将云计算与大数据进行一些比较，最明显的区分在两个方面：

第一，在概念上两者有所不同，云计算改变了IT，而大数据则改变了业务。然而大数据必须有云作为基础架构，才能得以顺畅运营。

第二，大数据和云计算的目标受众不同，云计算是CIO等关心的技术层，是一个进阶的IT解决方案。而大数据是CEO关注的、是业务层的产品，而大数据的决策者是业务层。

*4.1 分布式处理技术*

分布式处理系统可以将不同地点的或具有不同功能的或拥有不同数据的多台计算机用通信网络连接起来，在控制系统的统一管理控制下，协调地完成信息处理任务—这就是分布式处理系统的定义。

以Hadoop（Yahoo）为例进行说明，Hadoop是一个实现了MapReduce模式的能够对大量数据进行分布式处理的软件框架，是以一种可靠、高效、可伸缩的方式进行处理的。

而MapReduce是Google提出的一种云计算的核心计算模式，是一种分布式运算技术，也是简化的分布式编程模式，MapReduce模式的主要思想是将自动分割要执行的问题（例如程序）拆解成map（映射）和reduce（化简）的方式， 在数据被分割后通过Map 函数的程序将数据映射成不同的区块，分配给计算机机群处理达到分布式运算的效果，在通过Reduce 函数的程序将结果汇整，从而输出开发者需要的结果。

再来看看Hadoop的特性，第一，它是可靠的，因为它假设计算元素和存储会失败，因此它维护多个工作数据副本，确保能够针对失败的节点重新分布处理。其次，Hadoop 是高效的，因为它以并行的方式工作，通过并行处理加快处理速度。Hadoop 还是可伸缩的，能够处理 PB 级数据。此外，Hadoop 依赖于社区服务器，因此它的成本比较低，任何人都可以使用。

你也可以这么理解Hadoop的构成，Hadoop=HDFS（文件系统，数据存储技术相关）+HBase（数据库）+MapReduce（数据处理）+……Others

Hadoop用到的一些技术有：

HDFS: Hadoop分布式文件系统(Distributed File System) － HDFS HadoopDistributed File System)

MapReduce：并行计算框架

HBase: 类似Google BigTable的分布式NoSQL列数据库。

Hive：数据仓库工具，由Facebook贡献。

Zookeeper：分布式锁设施，提供类似Google Chubby的功能，由Facebook贡献。

Avro：新的数据序列化格式与传输工具，将逐步取代Hadoop原有的IPC机制。

Pig:大数据分析平台，为用户提供多种接口。

Ambari：Hadoop管理工具，可以快捷的监控、部署、管理集群。

Sqoop：用于在Hadoop与传统的数据库间进行数据的传递。

说了这么多，举个实际的例子，虽然这个例子有些陈旧，但是淘宝的海量数据技术架构还是有助于我们理解对于大数据的运作处理机制:

淘宝大数据

如上图所示，淘宝的海量数据产品技术架构分为五个层次，从上至下来看它们分别是：数据源，计算层，存储层，查询层和产品层。

数据来源层。存放着淘宝各店的交易数据。在数据源层产生的数据，通过DataX，DbSync和Timetunel准实时的传输到下面第2点所述的“云梯”。

计算层。在这个计算层内，淘宝采用的是Hadoop集群，这个集群，我们暂且称之为云梯，是计算层的主要组成部分。在云梯上，系统每天会对数据产品进行不同的MapReduce计算。

存储层。在这一层，淘宝采用了两个东西，一个使MyFox，一个是Prom。MyFox是基于MySQL的分布式关系型数据库的集群，Prom是基于Hadoop Hbase技术的一个NoSQL的存储集群。

查询层。在这一层中，Glider是以HTTP协议对外提供restful方式的接口。数据产品通过一个唯一的URL来获取到它想要的数据。同时，数据查询即是通过MyFox来查询的。

最后一层是产品层，这个就不用解释了。

*4.2 存储技术*

大数据可以抽象的分为大数据存储和大数据分析，这两者的关系是：大数据存储的目的是支撑大数据分析。到目前为止，还是两种截然不同的计算机技术领域：大数据存储致力于研发可以扩展至PB甚至EB级别的数据存储平台；大数据分析关注在最短时间内处理大量不同类型的数据集。

提到存储，有一个著名的摩尔定律相信大家都听过：18个月集成电路的复杂性就增加一倍。所以，存储器的成本大约每18-24个月就下降一半。成本的不断下降也造就了大数据的可存储性。

比如，Google大约管理着超过50万台服务器和100万块硬盘，而且Google还在不断的扩大计算能力和存储能力，其中很多的扩展都是基于在廉价服务器和普通存储硬盘的基础上进行的，这大大降低了其服务成本，因此可以将更多的资金投入到技术的研发当中。

以Amazon举例，Amazon S3 是一种面向 Internet 的存储服务。该服务旨在让开发人员能更轻松的进行网络规模计算。Amazon S3 提供一个简明的 Web 服务界面，用户可通过它随时在 Web 上的任何位置存储和检索的任意大小的数据。 此服务让所有开发人员都能访问同一个具备高扩展性、可靠性、安全性和快速价廉的基础设施，Amazon 用它来运行其全球的网站网络。再看看S3的设计指标：在特定年度内为数据源提供 99.999999999% 的耐久性和 99.99% 的可用性，并能够承受两个设施中的数据同时丢失。

S3很成功也确实卓有成效，S3云的存储对象已达到万亿级别，而且性能表现相当良好。S3云已经拥万亿跨地域存储对象，同时AWS的对象执行请求也达到百万的峰值数量。目前全球范围内已经有数以十万计的企业在通过AWS运行自己的全部或者部分日常业务。这些企业用户遍布190多个国家，几乎世界上的每个角落都有Amazon用户的身影。

此外，云创大数据的cStor云存储系统采用了先进的云计算技术、网络通信技术以及分布式文件系统技术，将硬件存储节点组织管理起来，以提供高性能、高可靠的存储。基于此，cStor A8000云存储系统一体机集中供电、集中散热，每个机架最大可搭载总存储容量高达3.8PB，但整体功耗却比传统方式节省10倍，全面展现了新一代高密度云存储产品的高容量、高性能以及节能环保的绿色魅力，已经广泛用于电信、平安城市等多个领域的海量数据存储与处理。

*4.3 感知技术*

大数据的采集和感知技术的发展是紧密联系的。以传感器技术，指纹识别技术，RFID技术，坐标定位技术等为基础的感知能力提升同样是物联网发展的基石。全世界的工业设备、汽车、电表上有着无数的数码传感器，随时测量和传递着有关位置、运动、震动、温度、湿度乃至空气中化学物质的变化，都会产生海量的数据信息。

而随着智能手机的普及，感知技术可谓迎来了发展的高峰期，除了地理位置信息被广泛的应用外，一些新的感知手段也开始登上舞台，比如，最新的”iPhone 5S”在home键内嵌指纹传感器，新型手机可通过呼气直接检测燃烧脂肪量，用于手机的嗅觉传感器面世可以监测从空气污染到危险的化学药品，微软正在研发可感知用户当前心情智能手机技术，谷歌眼镜InSight新技术可通过衣着进行人物识别。

除此之外，还有很多与感知相关的技术革新让我们耳目一新：比如，牙齿传感器实时监控口腔活动及饮食状况，婴儿穿戴设备可用大数据去养育宝宝，Intel正研发3D笔记本摄像头可追踪眼球读懂情绪，日本公司开发新型可监控用户心率的纺织材料，业界正在尝试将生物测定技术引入支付领域等。

其实，这些感知被逐渐捕获的过程就是就世界被数据化的过程，一旦世界被完全数据化了，那么世界的本质也就是信息了。

就像一句名言所说，“人类以前延续的是文明，现在传承的是信息。”

大数据的实践

互联网的大数据

互联网上的数据每年增长50%，每两年便将翻一番，而目前世界上90%以上的数据是最近几年才产生的。据IDC预测，到2020年全球将总共拥有35ZB的数据量。互联网是大数据发展的前哨阵地，随着WEB2.0时代的发展，人们似乎都习惯了将自己的生活通过网络进行数据化，方便分享以及记录并回忆。

互联网上的大数据很难清晰的界定分类界限，我们先看看BAT的大数据：

百度拥有两种类型的大数据：用户搜索表征的需求数据；爬虫和阿拉丁获取的公共web数据。搜索巨头百度围绕数据而生。它对网页数据的爬取、网页内容的组织和解析，通过语义分析对搜索需求的精准理解进而从海量数据中找准结果，以及精准的搜索引擎关键字广告，实质上就是一个数据的获取、组织、分析和挖掘的过程。搜索引擎在大数据时代面临的挑战有：更多的暗网数据；更多的WEB化但是没有结构化的数据；更多的WEB化、结构化但是封闭的数据。

阿里巴巴拥有交易数据和信用数据。这两种数据更容易变现，挖掘出商业价值。除此之外阿里巴巴还通过投资等方式掌握了部分社交数据、移动数据。如微博和高德。

腾讯拥有用户关系数据和基于此产生的社交数据。这些数据可以分析人们的生活和行为,从里面挖掘出政治、社会、文化、商业、健康等领域的信息，甚至预测未来。

在信息技术更为发达的美国，除了行业知名的类似Google，Facebook外，已经涌现了很多大数据类型的公司，它们专门经营数据产品，比如：

Metamarkets：这家公司对Twitter、支付、签到和一些与互联网相关的问题进行了分析，为客户提供了很好的数据分析支持。

Tableau：他们的精力主要集中于将海量数据以可视化的方式展现出来。Tableau为数字媒体提供了一个新的展示数据的方式。他们提供了一个免费工具，任何人在没有编程知识背景的情况下都能制造出数据专用图表。这个软件还能对数据进行分析，并提供有价值的建议。

ParAccel：他们向美国执法机构提供了数据分析，比如对15000个有犯罪前科的人进行跟踪，从而向执法机构提供了参考性较高的犯罪预测。他们是犯罪的预言者。

QlikTech：QlikTech旗下的Qlikview是一个商业智能领域的自主服务工具，能够应用于科学研究和艺术等领域。为了帮助开发者对这些数据进行分析，QlikTech提供了对原始数据进行可视化处理等功能的工具。

GoodData：GoodData希望帮助客户从数据中挖掘财富。这家创业公司主要面向商业用户和IT企业高管，提供数据存储、性能报告、数据分析等工具。

TellApart：TellApart和电商公司进行合作，他们会根据用户的浏览行为等数据进行分析，通过锁定潜在买家方式提高电商企业的收入。

DataSift：DataSift主要收集并分析社交网络媒体上的数据，并帮助品牌公司掌握突发新闻的舆论点，并制定有针对性的营销方案。这家公司还和Twitter有合作协议，使得自己变成了行业中为数不多可以分析早期tweet的创业公司。

Datahero：公司的目标是将复杂的数据变得更加简单明了，方便普通人去理解和想象。

举了很多例子，这里简要归纳一下，在互联网大数据的典型代表性包括：

用户行为数据（精准广告投放、内容推荐、行为习惯和喜好分析、产品优化等）

用户消费数据（精准营销、信用记录分析、活动促销、理财等）

用户地理位置数据（O2O推广，商家推荐，交友推荐等）

互联网金融数据（P2P，小额贷款，支付，信用，供应链金融等）

用户社交等UGC数据（趋势分析、流行元素分析、受欢迎程度分析、舆论监控分析、社会问题分析等）

同时，提供数据托管服务的大数据平台也应运而生，比如万物云与环境云。其中，作为智能硬件大数据免费托管平台，万物云可无限承载海量的物联网和智能设备数据。通过使用多种协议，各种智能设备将安全地向万物云提交产生的设备数据，在服务平台上进行存储和处理，并通过数据应用编程接口向各种物联网应用提供可靠的跨平台的数据查询和调用服务。万物云在大幅度降低物联网数据应用的技术门槛及运营成本的同时，也满足了物联网产品原型开发、商业运营和规模发展各阶段需求。目前，万物云的注册用户达到1605，入库数据超过55亿条。

环境云则是一个全面而便捷的综合环境大数据开放平台，收录权威数据源(中央气象台、国家环保部数据中心、美国全球地震信息中心等)所发布的各类环境数据，接收云创自主布建的全国各类环境监控传感器网络(包括空气质量指标，土壤环境质量指标检测网络)所采集的数据，并结合相关数据预测模型生成的预报数据，依托数据托管服务平台万物云所提供的数据存储服务，推出了一系列功能丰富、便捷易用的综合环境数据REST API，配合详尽的接口使用帮助，为环境应用开发者提供丰富可靠的气象、环境、灾害以及地理数据服务。此外，环境云还为环境研究人员提供了自定义数据报表生成和下载功能，并向公众展示环境实况。目前，环境云的入库数据已经超过6亿条。

政府的大数据

近期，奥巴马政府宣布投资2亿美元拉动大数据相关产业发展，将“大数据战略”上升为国家意志。奥巴马政府将数据定义为“未来的新石油”，并表示一个国家拥有数据的规模、活性及解释运用的能力将成为综合国力的重要组成部分，未来，对数据的占有和控制甚至将成为陆权、海权、空权之外的另一种国家核心资产。

在国内，政府各个部门都握有构成社会基础的原始数据，比如，气象数据，金融数据，信用数据，电力数据，煤气数据，自来水数据，道路交通数据，客运数据，安全刑事案件数据，住房数据，海关数据，出入境数据，旅游数据，医疗数据，教育数据，环保数据等等。这些数据在每个政府部门里面看起来是单一的，静态的。但是，如果政府可以将这些数据关联起来，并对这些数据进行有效的关联分析和统一管理，这些数据必定将获得新生，其价值是无法估量的。

具体来说，现在城市都在走向智能和智慧，比如，智能电网、智慧交通、智慧医疗、智慧环保、智慧城市，这些都依托于大数据，可以说大数据是智慧的核心能源。从国内整体投资规模来看，到2012年底全国开建智慧城市的城市数超过180个，通信网络和数据平台等基础设施建设投资规模接近5000亿元。“十二五”期间智慧城市建设拉动的设备投资规模将达1万亿元人民币。大数据为智慧城市的各个领域提供决策支持。在城市规划方面，通过对城市地理、气象等自然信息和经济、社会、文化、人口等人文社会信息的挖掘，可以为城市规划提供决策，强化城市管理服务的科学性和前瞻性。在交通管理方面，通过对道路交通信息的实时挖掘，能有效缓解交通拥堵，并快速响应突发状况，为城市交通的良性运转提供科学的决策依据。在舆情监控方面，通过网络关键词搜索及语义智能分析，能提高舆情分析的及时性、全面性，全面掌握社情民意，提高公共服务能力，应对网络突发的公共事件，打击违法犯罪。在安防与防灾领域，通过大数据的挖掘，可以及时发现人为或自然灾害、恐怖事件，提高应急处理能力和安全防范能力。

另外，作为国家的管理者，政府应该有勇气将手中的数据逐步开放，供给更多有能力的机构组织或个人来分析并加以利用，以加速造福人类。比如，美国政府就筹建了一个data.gov网站，这是奥巴马任期内的一个重要举措：要求政府公开透明，而核心就是实现政府机构的数据公开。截止目前，已经开放了有91054 个datasets；349citizen-developed apps；137 mobile apps；175 agencies and subagencies；87 galleries；295 Government APIs。

企业的大数据

企业的CXO们最关注的还是报表曲线的背后能有怎样的信息，他该做怎样的决策，其实这一切都需要通过数据来传递和支撑。在理想的世界中，大数据是巨大的杠杆，可以改变公司的影响力，带来竞争差异、节省金钱、增加利润、愉悦买家、奖赏忠诚用户、将潜在客户转化为客户、增加吸引力、打败竞争对手、开拓用户群并创造市场。

那么，哪些传统企业最需要大数据服务呢？抛砖引玉，先举几个例子：对大量消费者提供产品或服务的企业（精准营销）；做小而美模式的中长尾企业（服务转型）；面临互联网压力之下必须转型的传统企业（生死存亡）。

对于企业的大数据，还有一种预测：随着数据逐渐成为企业的一种资产，数据产业会向传统企业的供应链模式发展，最终形成“数据供应链”。这里尤其有两个明显的现象：外部数据的重要性日益超过内部数据。在互联互通的互联网时代，单一企业的内部数据与整个互联网数据比较起来只是沧海一粟；能提供包括数据供应、数据整合与加工、数据应用等多环节服务的公司会有明显的综合竞争优势。

对于提供大数据服务的企业来说，他们等待的是合作机会，就像微软史密斯说的：“给我提供一些数据，我就能做一些改变。如果给我提供所有数据，我就能拯救世界。”

然而，一直做企业服务的巨头将优势不在，不得不眼看新兴互联网企业加入战局，开启残酷竞争模式。为何会出现这种局面？从 IT 产业的发展来看，第一代 IT 巨头大多是 ToB 的，比如 IBM、Microsoft、Oracle、SAP、HP这类传统 IT 企业；第二代 IT 巨头大多是ToC 的，比如 Yahoo、Google、Amazon、Facebook 这类互联网企业。大数据到来前，这两类公司彼此之间基本是井水不犯河水；但在当前这个大数据时代，这两类公司已经开始直接竞争。比如 Amazon 已经开始提供云模式的数据仓库服务，直接抢占 IBM、Oracle 的市场。这个现象出现的本质原因是：在互联网巨头的带动下，传统 IT 巨头的客户普遍开始从事电子商务业务，正是由于客户进入了互联网，所以传统 IT 巨头们不情愿地被拖入了互联网领域。如果他们不进入互联网，他们业务必将萎缩。在进入互联网后，他们又必须将云技术，大数据等互联网最具有优势的技术通过封装打造成自己的产品再提供给企业。

以IBM举例，上一个十年，他们抛弃了PC，成功转向了软件和服务，而这次将远离服务与咨询，更多地专注于因大数据分析软件而带来的全新业务增长点。IBM执行总裁罗睿兰认为，“数据将成为一切行业当中决定胜负的根本因素，最终数据将成为人类至关重要的自然资源。”IBM积极的提出了“大数据平台”架构。该平台的四大核心能力包括Hadoop系统、流计算（StreamComputing）、数据仓库（Data Warehouse）和信息整合与治理（Information Integration and Governance）

IBM大数据

另外一家亟待通过云和大数据战略而复苏的巨头公司HP也推出了自己的产品：HAVEn，一个可以自由扩展伸缩的大数据解决方案。这个解决方案由HP Autonomy、HP Vertica、HP ArcSight 和惠普运营管理(HP OperationsManagement)四大技术组成。还支持Hadoop这样通用的技术。HAVEn不是一个软件平台，而是一个生态环境。四大组成部分满足不同的应用场景需要，Autonomy解决音视频识别的重要解决方案；Vertica解决数据处理的速度和效率的方案；ArcSight解决机器的记录信息处理，帮助企业获得更高安全级别的管理；运营管理解决的不仅仅是外部数据的处理，而是包括了IT基础设施产生的数据。

个人的大数据

个人的大数据这个概念很少有人提及，简单来说，就是与个人相关联的各种有价值数据信息被有效采集后，可由本人授权提供第三方进行处理和使用，并获得第三方提供的数据服务。

举个例子来说明会更清晰一些：

未来，每个用户可以在互联网上注册个人的数据中心，以存储个人的大数据信息。用户可确定哪些个人数据可被采集，并通过可穿戴设备或植入芯片等感知技术来采集捕获个人的大数据，比如，牙齿监控数据，心率数据，体温数据，视力数据，记忆能力，地理位置信息，社会关系数据，运动数据，饮食数据，购物数据等等。用户可以将其中的牙齿监测数据授权给XX牙科诊所使用，由他们监控和使用这些数据，进而为用户制定有效的牙齿防治和维护计划；也可以将个人的运动数据授权提供给某运动健身机构，由他们监测自己的身体运动机能，并有针对的制定和调整个人的运动计划；还可以将个人的消费数据授权给金融理财机构，由他们帮你制定合理的理财计划并对收益进行预测。当然，其中有一部分个人数据是无需个人授权即可提供给国家相关部门进行实时监控的，比如罪案预防监控中心可以实时的监控本地区每个人的情绪和心理状态，以预防自杀和犯罪的发生。

以个人为中心的大数据有这么一些特性：

数据仅留存在个人中心，其它第三方机构只被授权使用（数据有一定的使用期限），且必须接受用后即焚的监管。

采集个人数据应该明确分类，除了国家立法明确要求接受监控的数据外，其它类型数据都由用户自己决定是否被采集。

数据的使用将只能由用户进行授权，数据中心可帮助监控个人数据的整个生命周期。

展望过于美好，也许实现个人数据中心将遥遥无期，也许这还不是解决个人数据隐私的最好方法，也许业界对大数据的无限渴求会阻止数据个人中心的实现，但是随着数据越来越多，在缺乏监管之后，必然会有一场激烈的博弈：到底是数据重要还是隐私重要；是以商业为中心还是以个人为中心。

5 大数据应用及未来展望

大数据应用及未来展望，紧跟互联网+，便捷未来人们的生活，方便用户。畅想未来人与人之间的联系，从社交网络、社区文化，到“六度空间”;展望教育对于国家的希望，大数据和教育整合、合理的早教，对个人的帮助，对社会和国家的贡献;帮助病人的控制病情，并发病等，结合医疗平台，预测依照现有的生活方式，对个人的影响度，精确医疗救助，帮助老人送诊就医;减少自然灾害对人类，对生态环境的影响，“蝴蝶效应”预测自然灾害的发生;从开发者角度，整合用户数据，适应市场变化，用户需求，猜“你”喜欢，开发出满足用户需求的应用;大数据和人脸识别的结合应用，人脸分析，动态推送广告，全“自动”，非“手动”应用，强调人和人之间的猜“你”喜欢的全新的社交方式。大数据在未来的应用必将更加广泛，如何获取大数据，掌握大数据，提取大数据，整合大数据，关系到人们未来生活的方方面面，谁掌握了大数据，就掌握了未来!

中英文对照

Volume 量

Variety 多样

Value 价值

Velocity 速度

Data Warehouse 数据仓库