技术角度

大数据依托的NoSQL缺乏数据安全机制

从基础技术角度来看，大数据依托的基础技术是NoSQL（非关系型数据库）。

当前广泛应用的SQL（关系型数据库）技术，经过长期改进和完善，在维护数据安全方面已经设置严格的访问控制和隐私管理工具。

而在NoSQL技术中，并没有这样的要求。大数据数据来源和承载方式多种多样，如物联网、移动互联网、PC以及遍布地球各个角落的传感器，数据分散存在的状态，使企业很难定位和保护所有这些机密数据。

什么是隐私？

对于隐私这个词，科学研究上普遍接受的定义是“单个用户的某一些属性”，只要符合这一定义都可以被看做是隐私。我们在提“隐私”的时候，更加强调的是“单个用户”。那么，一群用户的某一些属性，可以认为不是隐私。我们拿刚才的例子来看，针对小明这个单个用户，“购买偏好”和“居住地址”就是隐私。如果公开的数据说住在五道口的小明爱买电子产品，那么这显然就是隐私泄漏了。但是如果数据中只包含一个区域的人的购买偏好，就没有泄露用户隐私。如果进一步讲，大家都知道小明住在海淀区五道口，那么是不是小明就爱买点此产品了呢？这种情况算不算事隐私泄漏呢？答案是不算，因为大家只是通过这个趋势推测，数据并不显示小明一定爱买电子产品。

所以，从隐私保护的角度来说，隐私是针对单个用户的概念，公开群体用户的信息不算是隐私泄漏，但是如果能从数据中能准确推测出个体的信息，那么就算是隐私泄漏。

历史案例:

1.美国在线（AOL）是一家美国互联网服务公司，也是美国最大的互联网提供商之一。在 2006年8月，为了学术研究，AOL 公开了匿名的搜索记录，其中包括 65 万个用户的数据，总共 20M 条查询记录。在这些数据中，用户的姓名被替换成了一个个匿名的 ID，但是纽约时报通过这些搜索纪录，找到了 ID 匿名为 4417749的用户在真实世界中对应的人。ID 4417749 的搜索记录里有关于“60岁的老年人”的问题、“ Lilburn地方的风景”、还有“Arnold” 的搜索字样。通过上面几条数据，纽约时报发现 Lilburn 只有14个人姓Arnold，最后经过直接联系这 14个人确认 ID 4417749 是一位62岁名字叫 Thelma Arnold的老奶奶。最后 AOL 紧急撤下数据，发表声明致歉，但是已经太晚了。因为隐私泄露事件，AOL遭到了起诉，最终赔偿受影响用户总额高达五百万美元。

2.同样是 2006年，美国最大的影视公司之一 Netflix，举办了一个预测算法的比赛（ Netflix Prize），比赛要求在公开数据上推测用户的电影评分。Netflix 把数据中唯一识别用户的信息抹去，认为这样就能保证用户的隐私。但是在 2007 年来自The University of Texas at Austin 的两位研究人员表示通过关联 Netflix 公开的数据和 IMDb（互联网电影数据库）网站上公开的纪录就能够识别出匿名后用户的身份。三年后，在2010年，Netflix 最后因为隐私原因宣布停止这项比赛，并因此受到高额罚款，赔偿金额总计九百万美元。

3.近期，Facebook上的用户数据被窃取，大数据分析公司Cambridge Analytica利用了从Facebook上不当收集的5000万用户的个人数据，“操纵民意”，在帮助特朗普击败希拉里过程中起到了重要作用。Facebook面临着前所未有的危机，不仅市值蒸发了超过500亿美元，而且“卸载Facebook”的口号也开始流行起来。

4.据报道，目前国内的滴滴等叫车软件利用大数据“杀熟”，新老用户起始地相同，路线相同，老客户的打车费用却更高。虽然滴滴回应绝无这种做法，但不信任的情绪在蔓延。

同时，携程也被曝光利用大数据，对老客户进行“杀熟”。有网友称，他经常通过某网站订某个特定酒店的房间，长年价格在380元-400元。偶然一次，他从前台得知酒店淡季的价格在300元上下。他用朋友的账号查询也是300元，但用自己的账号查，还是380元。网友也纷纷晒出了自己被宰的经历，大骂企业无良，不少媒体认为大数据杀熟是技术走了歪路，要有法律制裁才行。

无论是百度，还是嘀嘀，携程这类互联网企业，都拥有庞大的用户群体，掌握着众多的用户信息，他们才是信息的主导方，决定了我们接收什么样的信息。而对于个人来说，很多时候我们没办法保护自己的隐私，很大程度上因为我们根本没有选择权，我们无法选择信息公布还是隐藏，更无法决定屏蔽哪些信息，更谈不上我们能主动接收想了解的信息，说到底，我们就是被动的接收信息，信息完全不对等。

大量事实表明：大数据未被妥善处理会对用户的隐私造成极大的侵害。根据需要保护的内容不同，隐私保护又可以进一步细分为位置隐私保护、标识符匿名保护、连接关系匿名保护等。人们面临的威胁并不仅限于个人隐私泄漏，还在于基于大数据对人们状态和行为的预测。一个典型的例子是

5.某零售商通过历史记录分析，比家长更早知道其女儿已经怀孕的事实，并向其邮寄相关广告信息。而社交网络分析研究也表明，可以通过其中的群组特性发现用户的属性。例如通过分析用户的Twitter信息，可以发现用户的政治倾向、消费习惯以及喜好的球队等。当前企业常常认为经过匿名处理后，信息不包含用户的标识符，就可以公开发布了。但事实上，仅通过匿名保护并不能很好地达到隐私保护目标。

目前用户数据的收集、存储、管理 与 使 用 等 均 缺 乏 规 范 ，更 缺 乏 监管，主要依靠企业的自律。用户无法确定自己隐私信息的用途。而在商业化场景中，用户应有权决定自己的信息如何被利用，实现用户可控的隐私保护。包括：数据采集时的隐私保护，如数据精度处理；数据共享、发布时的隐私保护，如数据的匿名处理、人 工 加 扰 等 ；数 据 分 析 时 的 隐 私 保护；数据生命周期的隐私保护；隐私数据可信销毁等。

大数据隐私威胁的来源



Figure 1 大数据价值链示意图

1.产生环节 威胁来源

1）大数据时代，新浪、腾讯、淘宝等商家既是数据的生产者，又是数据的存储、管理者和使用者，单纯通过技术手段限制商家对用户信息的使用，实现用户隐私保护将极其困难。

2）目前用户数据的收集、存储、管理与使用等均缺乏规范，更缺乏监管，主要依靠企业的自律。用户无法确定自己隐私信息的用途。

3）服务商内部人员偷盗售卖、网络服务安全漏洞被利用、木马、钓鱼网站恶意盗取等也可能导致该环节的数据安全及隐私问题。

2.数据获取及传输环节

数据获取及传输根据传输渠道分为两类：线上传输和线下传输。线上主要指利用互联网等虚拟媒介而实现的一系列没有发生面对面交互的数据传输；线下可理解为有真实发生的、当面的、人与人有通过肢体动态的一系列活动，即通过移动硬盘等移动存储媒介传输数据。

对于线下传输方式，实体媒介的物理安全是数据安全的保障，实体媒介的遗失、被窃、损害等情况，都存在数据保密性、完整性被破坏的风险，可能导致用户隐私信息的泄露。

对于线上传输方式，比如Internet或Intranet上传输，可以使用 Internet 协议安全性 (IPSec)保证其安全。但用户的数据以明文的形式传输，被黑客截获，很可能存在数据被盗用、滥用情况，导致用户隐私信息泄露。因此，需要对获取和传输阶段的安全&隐私进行保障。

3.数据存储环节

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，以往的存储系统已经无法满足大数据应用的需要。对于占数据总量80%以上的非结构化数据，通常采用非关系型数据库（NoSQL）存储技术完成对大数据的抓取、管理和处理。虽然NoSQL 数据存储易扩展、高可用、性能好，但是仍存在一些问题。例如，访问控制和隐私管理模式问题、技术漏洞和成熟度问题、授权与验证的安全问题、数据管理与保密问题等。而结构化数据的安全防护也存在漏洞，例如物理故障、人为误操作、软件问题、病毒、木马和黑客攻击等因素都可能严重威胁数据的安全性。

4.在数据分析及应用阶段，存在安全、隐私问题，其原因包括：

1）关联属性：在大数据分析及应用阶段，可以通过用户零散数据之间的关联属性，将某个人的很多行为数据聚集在一起时，他的隐私就很可能会暴露，因为有关他的信息已经足够多，这种隐性的数据暴露往往是个人无法预知和控制的。(即通过大数据的关联分析)

2）基于大数据的个人状态及行为预测：在大数据时代，人们面临的威胁并不仅限于个人隐私泄露，还在于基于大数据对人们状态和行为的预测。例如零售商可以通过历史纪录分析，得到顾客在衣食住行等方面的爱好、倾向等；社交网络分析研究也表明，可以通过其中的群组特性发现用户的属性，例如通过分析用户的微博等信息，可以发现用户的政治倾向、消费习惯以及其它爱好等。