大数据面临的数据安全

数据在当前学术界和产业界扮演至关重要的角色，它被认为是对我们生活、工作和思维方式的重大变革。然而，大数据时代在安全和个人隐私的方面存在许多风险，由此所引起的隐私泄露为个人带来困扰，同时伴随而来的虚假信息也将导致错误的分析结果。因此人们迫切需要技术以确保数据安全。本文将总结并分析大数据时代所带来的安全挑战和机遇，并提供相对应的关键对策。

一、 引言

在信息技术中，“ 大数据 ”是指一些使用目前现有数据库管理工具或传统数据处理应用很难处理的大型而复杂的数据集。其挑战包括采集、管理、存储、搜索、共享、分析和可视化。更大的数据集的趋势是由于从相关数据的单一大数据集推导而来的额外信息，与分离的较小的具有相同数据总量的数据集相比，能够发现相关性来“ 识别商业趋势（ spot business trends ）、确定研究的质量、预防疾病、法律引用链接、打击犯罪以及实时确定道路交通状态 ”。近几年大数据一词的持续升温也带来了大数据泡沫的疑虑，大数据的前景与目前云计算、物联网、移动互联网等是分不开的。目前大数据的发展仍然面临着许多问题，数据安全与隐私问题是人们公认的关键问题之一。当前，人们在互联网上的一言一行都掌握在互联网商家手中，包括购物习惯、好友联络情况、阅读习惯、检索习惯等等．多项实际案例说明，即使无害的数据被大量收集后，也会暴露个人隐私．事实上，大数据安全含义更为广泛，人们面临的威胁并不仅限于个人隐私泄漏．与其它信息一样，大数据在存储、处理、传输等过程中面临诸多安全风险，具有数据安全与隐私保护需求。本文介绍大数据时代的到来，重点分析了当前大数据所带来的安全挑战，详细阐述了当前大数据安全与隐私保护的关键技术．同时必须承认，大数据在引人新的安全问题和挑战的同时也为信息安全领域带来了新的发展机遇，即基于大数据的信息安全相关技术可以反过来用于大数据的安全和隐私保护。

二、 大数据时代的到来

大数据时代的到来，是全球知名咨询公司麦肯锡最早提出的。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。”当今，社会信息化和网络化的发展导致数据爆炸式增长．据统计，平均每秒有 200 万用户在使用谷歌搜索， Facebook用户每天共享的东西超过 40 亿， Twitter 每天处理的推特数量超过 3．4 亿．同时，科学计算、医疗卫生、金融、零售业等各行业也有大量数据在不断产生． 2012 年全球信息总量已经达到 2．7ZB，而到 2015 年这一数值预计会达到 8 ZB．这一现象引发了人们的广泛关注。下面先从大数据与目前云计算、物联网、移动互联网的联系讲起。

(一) 大数据市场格局

从严格意义上来说，早在 20 世纪 90 年代“数据仓库之父”的 Bill Inmon 便提出了“大数据”的概念。大数据之所以在最近走红，主要归结于互联网、移动设备、物联网和云计算等快速崛起，全球数据量大大提升。可以说，移动互联网、物联网以及云计算等热点崛起在很大程度上是大数据产生的原因。物联网，移动互联网再加上传统互联网，每天都在产生海量数据，而大数据又通过云计算的形式，将这些数据筛选处理分析，提前出有用的信息，这就是大数据分析。

(二) 大数据具有四个典型特征

大数据 (Big Data) 是指“无法用现有的软件工具提取、存储、搜索、共享、分析和处理的海量的、复杂的数据集合。”业界通常用四个 V（即 Volume、Variety 、Value、Velocity ）来概括大数据的特征。

第一，数据体量巨大（ Volume）。到目前为止，人类生产的所有印刷材料的数据量是 200PB（1PB=1000TB），而历史上全人类说过的所有的话的数据量大约是 5EB（1EB=1000PB）。当前，典型个人计算机硬盘的容量为 TB 量级，而一些大企业的数据量已经接近 EB量级。

第二，数据类型繁多 (Variety ）。这种类型的多样性也让数据被分为结构化数据和非结构化数据。相对于以往便于存储的以文本为主的结构化数据，非结构化数据越来越多，包括网络日志、音频、视频、图片、地理位置信息等等多类型的数据对数据的处理能力提出了更高的要求。

第三，价值密度低（ Value）。价值密度的高低与数据总量的大小成反比。以视频为例，一部一小时的视频，在连续不间断监控过程中，可能有用的数据仅仅只有一两秒。如何通过强大的机器算法更迅速地完成数据的价值“ 提纯 ”是目前大数据汹涌背景下亟待解决的难题。

第四，处理速度快（ Velocity ）。这是大数据区分于传统数据挖掘最显著的特征。根据 IDC 的“ 数字宇宙 ”的报告，预计到 2020 年全球数据使用量将会达到 35.2ZB。在如此海量的数据面前，处理数据的效率就是企业的生命。

(三) 大数据对现实生活的影响

大数据能带来什么变化呢？通过对卫星以及全球数亿传感器、 RFID 标签、带 GPS 的相机和智能手机实时收集的数据做可视化处理，人类就可以感知、测量、理解和影响人类的生存方式，实现先辈们遥不可及的梦想。

2012 年 9 月 25 日到 10 月 2 日，邀请全球各地参与者通过“ 大数据人类面孔 ”这一应用来“ 测量我们的世界 ”。这一应用可以让人们用手机作为传感器参与一系列活动，他们同时可以比较全球其它参与者对一些值得深思的问题给出了什么答案。参与者可以绘制出自己每天的路径，分享那些带给他们好运的物品和仪式，了解其他人想要在一生中经历的特别体验，发现自己身边以前没有意识到的秘密。参与者还能够得出自己的“ 数字身影 ”。

三、 大数据给信息安全带来的机遇和挑战

大数据在带来了新安全风险的同时也为信息安全的发展提供了新机遇。大数据正在为安全分析提供新的可能性，对于海量数据的分析有助于信息安全服务提供商更好的刻画网络异常行为，从而找出数据中的风险点。对实时安全和商务数据结

合在一起的数据进行预防性的分析，以便识别钓鱼攻击，防止诈骗和阻止黑客入侵。网络攻击行为总会留下蛛丝马迹，这些痕迹都以数据的形势隐藏在大数据中，利用大数据技术整合计算和处理资源有助于更有针对性的应对信息安全威胁，使得网络攻击行为无所遁形，有助于找到发起攻击的源头。

四、 数据安全策略

数据的安全有两层含义：一个是逻辑上的安全，比如防止病毒的破坏、黑客入侵等等。另一个就是物理上的安全，比如人为的错误或不可抗拒的灾难。前者需要系统的安全防护，后者需要数据存储备份的保护。接下来主要针对目前在数据存储备份及数据恢复方面采用的技术进行介绍。

(一) 数据备份

1. 磁带存储

目前使用磁带存储设备解决企业数据备份保存问题依然是行之有效的方法。目前磁带存储主要应用的技术有三种： LT0(开放线性磁带) 、DLT(数码线性磁带) 和AIT(先进智能磁带)。对于用户的不同需求，这三种技术都有各自的优缺点，了解它们各自的性能特点，将会帮助用户选择到底自己适合使用哪种技术。

2. 网络存储技术

网络存储技术是基 f 数据存储的一种通用网络术语。网络存储结构大致分为三种：直连式存储 (DAS)、网络存储设备 (NAS)和存储网络 (sAN)。

3. 第三方数据灾难备份

我们上面所说的这些准备工作都是针对自己实施数据容灾而言的，我们还可以通过第三方的服务来完成数据灾备工作。特别是针对数据灾备而言，选择第三方服务已经成为数据容灾领域的一个非常莺要的发展趋势。外包数据灾备业务为国内企业广泛实施提供了一条切实可行的道路。目前 C A、EMC IBM、惠普、赛门铁克等知名 IT 厂商已经开始在国内提供数据灾备服务，这些厂商所拥有的技术能力和

服务经验可以极大的降低实施数据灾备的技术风险和成本投入。一些银行也能够为企业提供存储数据的服务，相对于自己实施数据灾备来说，第三方的数据灾备服务可以以更少的投入获得更高的效率，这对很多确实有数据容灾需求甚至正在观望的用户都是极具吸引力的。

最后值得一提的是，无论是自己实施数据灾备还是委托第三方进行数据灾备服务，数据灾备计划的执行和后续工作都是非常重要的，就像进行灾备之前都要进行准备工作一样。即使将数据灾备业务外包出去也不代表企业人员就摆脱了所有的责任，其实在外包方式下食、№需要进行的沟通协调工作并不少于自己建设数据容灾设施。

(二) 数据恢复技术

当数据丢失而又未作好备份或备份丢失的情况下，只能进行数据恢复。实现数据恢复技术主要靠软件技术、硬件技术及二者的结合。软件技术主要有杀毒软件内嵌的恢复数据功能，其恢复数据功能比较有限，而且对于其它类型造成的数据丢失效果不理想。其次是专业的数据恢复软件，其除了具有备份数据、存储数据功能外，还采用了许多先进的理念，如数据直接读取、分析处理数据、修补数据技术，其数据灾难恢复能力较强。常用的恢复软件包括 Scandisk， PC一 3000 亦是磁盘检测工具功能更强大；现在新兴的修复数据误删除、误格式化的数据的软件较多，各有优势，需要在大量的实践中找到适合自己应用技巧的软件。

用软件方法恢复速度较快，但硬件方面造成的数据丢失一般无法恢复。数据恢复中的硬件恢复能在不破坏原有存储介质系统的条件下，对多种存储介质、多种硬件平台、软件平台下的大多数原因造成的数据丢失进行数据恢复，但费用不菲，同时需要一些高精尖设备的支持。

数据的恢复操作一般应由具在资质的专业人员进行。当用户计算机系统、存储介质、软硬件发生意外，导致重要数据无法显示或读取，应立即切断电源，停止使用，保护磁盘，保护数据恢复的现场。避免进行任何非专业的操作，这样才能最大限度地保证为数据的恢复提供最大的可能。非专业的数据恢复尝试，有些时候不但不能挽救数据，而是对数据二次破坏甚至导致数据无法恢复。据统计在日常应用

中大部分数据丢失与逻辑数据损坏有关，所以需要使用者和维修专业人员对本部分知识有更加深入的学习。

数据恢复工作是在数据遭受损失，采用常规方法无法再次获取数据文件时采取的补救性措施，这就要求我们养成良好的数据备份的习惯，多学习数据备份的方法及技巧，特别要学会用云盘保存重要数据。当然，也应研究新的加密技术使数据不被非法者所读取。

五、 基于各类新平台的数据安全技术

云计算使用灵活高效的加密机制保护用户数据的机密性，针对加密数据难以检索的问题部署了密文检索方法．为保证用户数据的完整性，用户根据自己的秘密信息，以挑战一应答模式发起连续的验证请求，根据返回有限结果可以判断远程海量信息的存在性及正确性．经实验表明，该平台能够较好地保证数据的机密性与完整性，并具有良好的可用性和可扩展性。

设置开机密码、数据转移或删除、计算机定位等方面加强个人计算机的数据安全， 可以提高数据安全的可靠性，通过基于手机的方式能够提高功能实现的方便性和高效性。

移动云服务相比传统云具有移动互联、灵活终端应用和便捷数据存取等特点。然而，丰富的移动云服务应用也带来了更多的安全与隐私泄露问题。在阐述移动云服务的基本概念、应用与安全问题的基础给出了其安全与隐私保护体系结构，主要围绕安全协议与认证、访问控制、完整性验证、移动可信计算和基于加密、匿名、混淆的隐私保护等关键技术。

智能电网方案充分利用了 HBase 高性能优势和现代密码技术，将密钥与密文的管理分离，具有安全性好、密钥管理方便及效率高等特点。开发了基 Hadoop 的原型系统，对方案的时间开销进行了分析，并通过相关实验证明了方案的有效性和可行性。

结论

尽管大数据带来了新的安全问题，但它自身也是解决问题的重要手段，既是挑战又是机遇。从大数据的来源，数据安全等角度出发，可见当前大数据安全与隐私保护息息相关．但总体上来说，当前针对大数据安全与隐私保护的相关研究刚刚起步，尚未达到统一的标准，新技术和新思维仍在不断被提出中。我认为，通过技术手段与相关政策法规等相结合，才能更好地解决大数据时代下数据安全与个人隐私保护

问题。相信在技术人员和相关立法机构和监管部门的共同努力下，数据安全和数据共享最终必能兼得。