北京交通大学
BEIJING JIAOTONG UNIVERSITY

大数据安全与隐私保护

2022.04.25

本章教学要点

- 本章主要介绍大数据安全、大数据隐私的相关理论概念。
- 其中需了解大数据安全问题的分类、大数据安全防护技术以及大数据隐私保护技术的研究内容；
- 了解大数据安全的概念、造成大数据安全问题的原因以及大数据安全防护方法相关知识点；
- 了解大数据隐私问题发展历程、大数据隐私保护政策的内容。

目录

01

大数据安全概述

PART ON

大数据安全的概念

大数据安全是指确保数据的保密性、完整性和可用性，不受到信息泄漏和非法篡改的安全威胁影响。

保密性

01

又称为机密性，是指禁止非法用户在没有授权的情况下访问、获取数据，避免数据遭受破坏或泄露而造成安全隐患，**数据加密、数据隐藏、访问控制**等是实现机密性要求的常用手段。

完整性

02

完整性是指在传输、存储数据的过程中，**确保数据不被未授权者篡改、损坏、销毁，或在篡改后能够被迅速发现**，常见的完整性的技术手段有数字签名。

可用性

03

可用性是指**保证合法用户在需要时可以使用所需的数据，并且数据在传输过程中没有失真**，使用数据的过程是可控的，常见手段如备份与恢复技术、防火墙技术等。

大数据安全问题的形成原因

大数据平台安全机制的不足

大数据分布式存储的风险

新型虚拟化网络技术的局限

传统数据安全防护技术的缺陷

新型高级网络攻击的威胁

大数据安全问题的形成原因

• 1.传统数据安全防护技术的缺陷

目前，针对大数据平台的网络攻击目的已经从单纯地窃取数据、瘫痪系统转向**干预、操纵分析结果**，攻击效果已经从直观易察觉的系统宕机、信息泄露转向**细小难以察觉的分析结果偏差**。同时，基于**大数据海量、多源、动态的特征以及大数据环境分布式、组件多、接口多的特点**，传统的基于监测、预警、响应的安全防护技术难以应对大数据安全问题的动态变化。

• 2.大数据分布式存储的风险

由于大数据在云端的分布式集中存储和处理，使得安全保密风险也向云端集中，一旦云端服务器受到攻击，**海量信息可在瞬间被集中窃取。**

大数据安全问题的形成原因

- **3.大数据平台安全机制的不足**

大数据时代，数据平台大多是基于Hadoop体系结构的，但是这种体系结构在自身安全机制方面存在局限性:

①在Hadoop体系结构中，**用户的身份鉴别和授权访问**等安全保障能力比较薄弱，它依赖于Linux的身份和权限管理机制，身份管理仅支持用户和用户组，权限管理仅有可读、可写和可执行3个，不能满足基于角色的身份管理和细粒度访问控制等新的安全需求。

②在安全审计方面，Hadoop只有分布在各组件中的日志记录，**没有原生安全审计功能**，需要使用外部附加工具进行日志分析。

③由于Hadoop是开源的，因此缺乏严格的测试管理和安全认证，**对组件漏洞和恶意后门的防范能力不足，存在漏洞和恶意代码**。

大数据安全问题的形成原因

- **4.新型虚拟化网络技术的局限**

为应对大数据环境下**网络架构的可扩展性需求**，以软件定义网络(Software Defined Network，SDN)和网络功能虚拟化(Network Function Virtualization，NFV)为代表的新型网络虚拟化技术近来发展迅速。

① 转发层集中管理所有网络设备，进行数据处理、转发和分配。

② 控制层以通用接口的方式与转发层进行数据传输通信，实现数据资源的管理编排以提供网络服务。

③ 应用层包括各种不同的业务应用。

大数据安全问题的形成原因

- **4.新型虚拟化网络技术的局限**

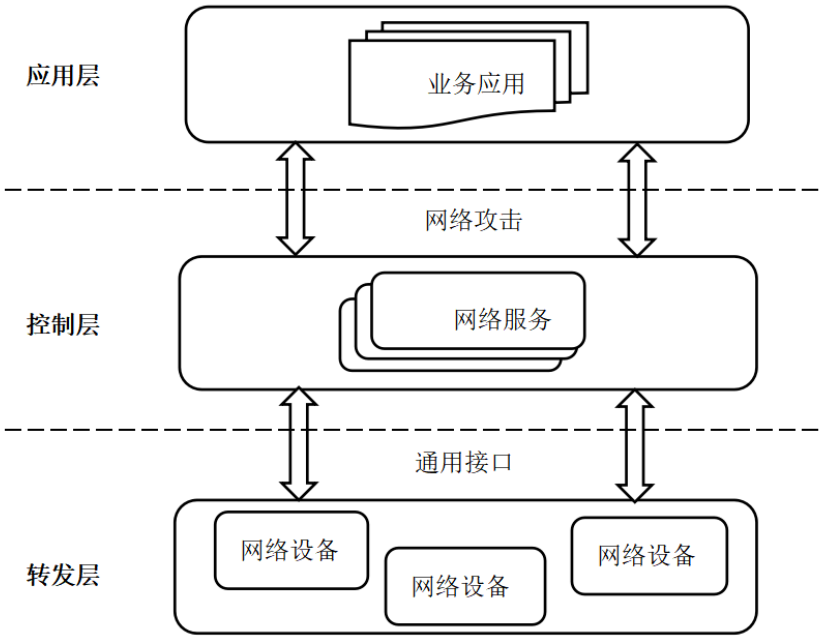


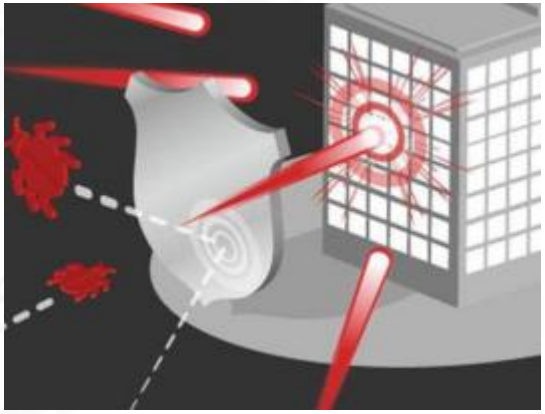
图 9.2 SDN 三层架构图

在这3层架构中，**通用接口的开放性会引发漏洞暴露和接口滥用**的问题，从而遭受更多的网络攻击。同样，NFV公开部署时通常会**外包给第三方虚拟化平台**，如云平台，而共享、非可信和虚拟化的环境使得NFV的管理编排和安全运行面临着很大的威胁。

大数据安全问题的形成原因

- **5. 新型高级网络攻击的威胁**

大数据具有的巨大潜在价值不仅使之在各行业领域的应用中被广泛重视，也使得它更容易成为攻击者的重点目标，从中挖掘出有利于攻击者实施破坏行为的信息，而且在大数据存储、计算、分析等技术快速发展的同时，也催生了很多新型高级的网络攻击手段。例如高级可持续攻击（Advanced Persistent Threat，APT），攻击者将APT攻击代码长期隐蔽在大数据中，大数据的价值低密度性，使得安全分析工具难以聚焦在价值点上，因此APT攻击的发现难度更大，攻击也更加精准，严重威胁着网络安全。



大数据安全问题的分类

• 1. 大数据平台安全

1 **大数据存储安全**

云存储平台并不是完全可信的，面临着非法入侵、泄露或篡改的风险;数据量的指数级增长，对各种类型和结构的数据进行数据存储，极易造成数据存储错位和数据管理混乱，为大数据后期的处理带来安全隐患。

2 **大数据传输安全**

大数据传输环节面临着数据失真、泄漏、篡改以及被数据流攻击者利用等风险。

3 **大数据平台访问控制安全**

大数据场景中，需要实施身份和权限管理，然而对未知的大量数据和用户进行**角色预设**十分困难。

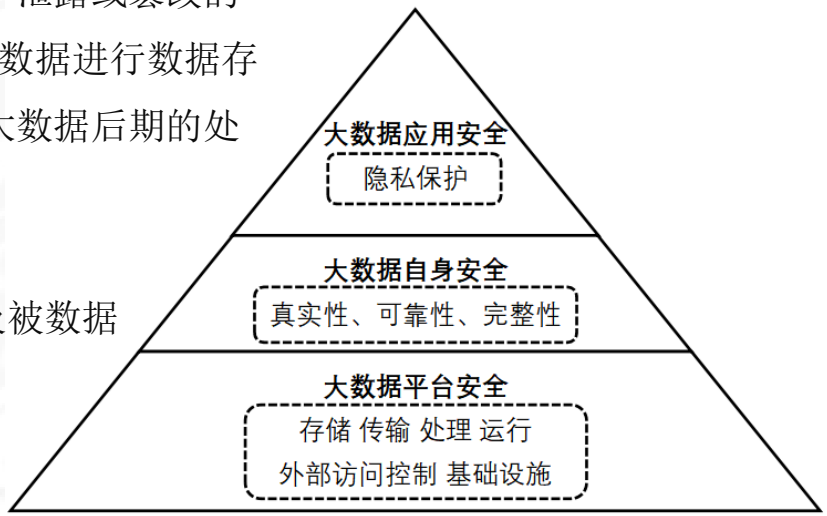


图 9.3 大数据安全的三个层面

大数据安全问题的分类

- ## 1. 大数据平台安全

4 **大数据运行计算安全**

各种数据应用背景不同，频繁的数据共享和交换促使**数据流动**路径变得交错复杂，再加上大数据的分布式、虚拟化处理模式，这些都使得数据在分析和处理过程中面临着被盗取的威胁。

5 **大数据基础设施安全**

大数据基础设施为大数据平台组件的运行提供所需的**存储、传输、网络等资源**，包括物理资源和虚拟化资源。攻击者往往会通过非授权访问、在网络传输过程中破坏数据完整性、传播网络病毒等方式对大数据基础设施造成安全威胁。

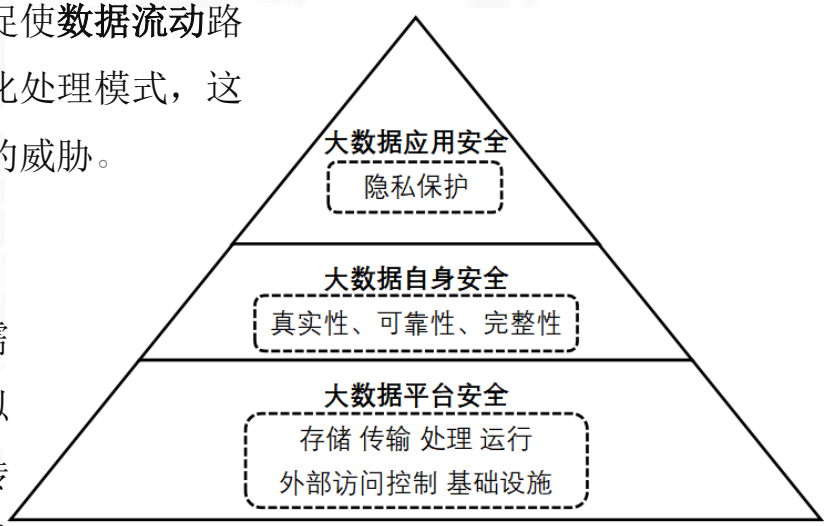


图 9.3 大数据安全的三个层面

大数据安全问题的分类

- **2. 大数据自身安全**

大数据具有来源广泛、类型多样、数据量增长速度快等特点，这给大数据自身的安全提出了更高的要求：

①需要保障数据源的**真实可信性**，防止源数据被伪造或刻意制造，并且考虑时间、数据版本变更等因素导致的数据失真。

②需要保障数据源的**可靠性和完整性**，尽可能减小数据采集过程中由于人工干预带来的误差，确保数据不被篡改、损坏，避免影响后期数据分析结果的准确性。

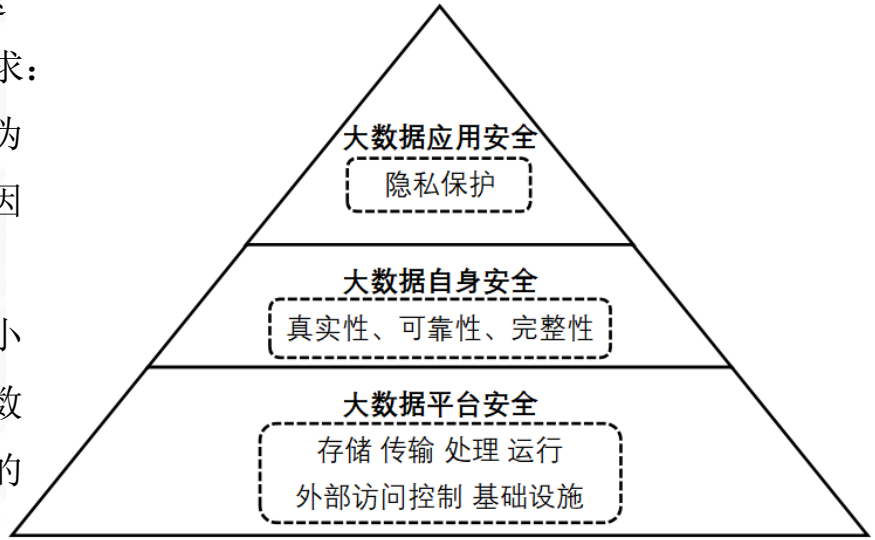


图 9.3 大数据安全的三个层面

大数据安全问题的分类

- **3. 大数据应用安全**

大数据在应用过程中尚存在一些安全问题，例如数**据共享、数据外包过程中的数据泄露和破坏，大数据市场资格认证和准入机制的缺乏以及受到持续关注的用户隐私保护问题**。

隐私保护是指利用去标识化、匿名化、密文计算等技术保障个人数据在平台上处理、流转过程中不被泄露。

大数据时代的隐私保护不仅仅是**保护个人隐私权，还包括在个人信息收集、使用过程中保障对个人信息的自主决定权利**，即用户应有权利决定自己的信息如何被使用，从而实现用户可控的隐私保护。

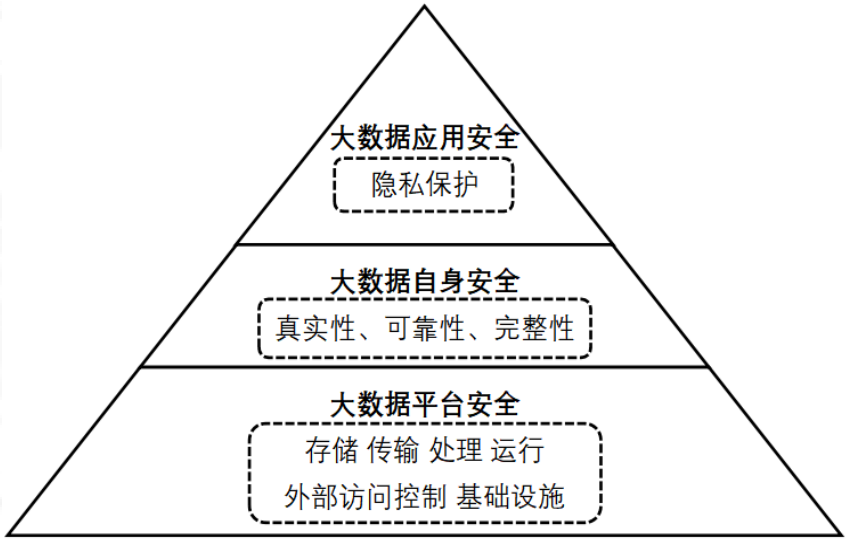


图 9.3 大数据安全的三个层面

02

大数据隐私问题

PART TW

大数据隐私问题的发展历程

1890年，美国著名学者沃伦（Warren）与布兰戴斯（Brandies）在《隐私权》一书中就清晰地提出了“隐私权”的概念。他们将隐私权定义为“不受干涉”或“免于侵害”的独处权利。这意味着隐私具有**个体自决性**，即个人具有决定隐私的对象、范围等的能动性，个人享有不被他人干涉和个人不愿被公开的信息被防护的权利。大数据时代的隐私问题的变化呈现如下特征：

1 **隐私范围扩大且难以界定**

隐私不再只是姓名、性别、联系方式等个人属性信息，它还扩展到了用户**在媒体信息行为中留下的行为痕迹信息、位置信息、消费信息、社会网络关系信息等个人行为信息。**

大数据隐私问题的发展历程

2 **隐私权利归属复杂**

① 在**个人**层面上，要求得到充分防护隐私的权利；

② **企业和组织**也同样具有信息产权，即在网络上通过合法方式搜集到的用户信息的权利；

③ **政府**则拥有以整个国家为主体而产生的所有数据。

3 **隐私保护困难**

隐私侵犯成为了更加普遍、更加有利可图的行为，侵犯个人隐私的形式也愈加复杂多样，对于界定是否构成侵权行为，根据目前的法律却无法准确判断。

大数据隐私保护政策

- **1. 国外针对隐私保护的主要相关政策**

年份	国家和地区	内容	说明
1970	德国	《联邦数据防护法》	对信息泄露的惩处
1974	美国	《隐私法》	基础法
1986	美国	《电子通信隐私法》	保证用户个人的通讯安全
1995	欧盟	《个人数据防护指令》	防护个人数据的最低标准
2002	欧盟	《关于在电子通信领域个人数据处理及保护隐私权的指令》	电子商务消费者隐私权
2003	日本	个人信息防护法	个人信息防护法律依据
2009	美国	《2009个人隐私与安全法案》、《数据泄漏事件通报法案》	数据泄漏通报标准
2012	美国	《消费者隐私权力法》	消费者隐私拥有权
2013	欧盟	《数据防护基本条例》	严格数据保障机制
2015	美国	《网络安全信息共享法案》	保障网络空间安全
2017	法国	《防护个人数据法案》	未成年人数据防护
2019	欧盟	《通用数据防护条例》	欧盟居民提供商品服务的境外数据处理

大数据隐私保护政策

- **2. 我国针对隐私保护的主要相关政策**

年份	内容	说明
2009	《中华人民共和国侵权责任法》	隐私权
2013	《电信和互联网用户个人信息防护规定》	收集、使用个人信息的规则和信息安全保障措施要求
2015	《促进大数据发展行动纲要》	健全大数据安全保障体系
2015	《中华人民共和国刑法修正案（九）》	网络个人隐私信息刑法防护
2016	《中华人民共和国国民经济和社会发展第十三个五年规划纲要》	大数据安全管理制度
2016	《中华人民共和国网络安全法》	全面规范网络空间安全管理、基础性法律
2017	《中华人民共和国民法总则》	个人信息自主权
2018	《信息安全技术个人信息安全规范》	个人信息安全
2019	《数据安全管理办法（征求意见稿）》	网络运营者行为规定

03

大数据安全技术

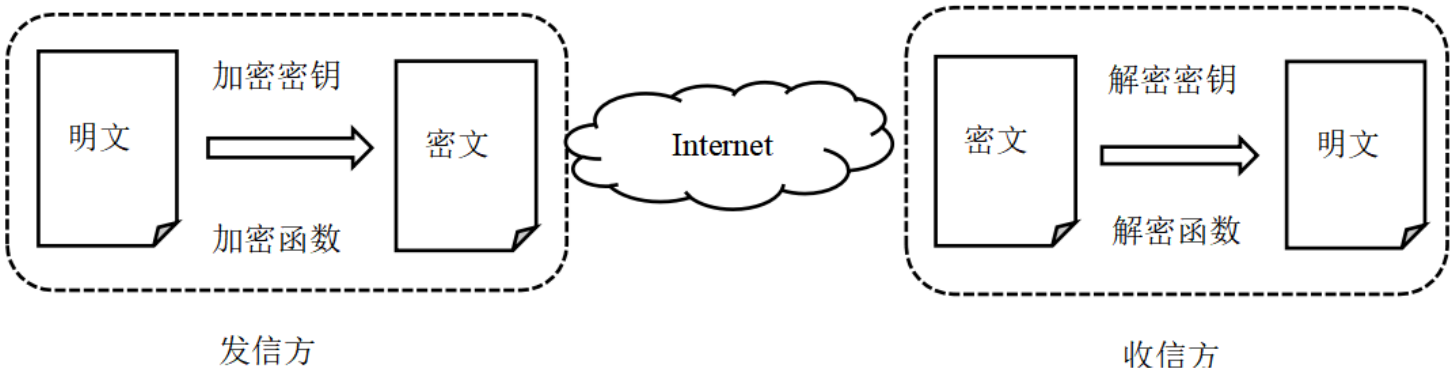
PART TH

大数据安全相关技术

大数据的安全防护技术的关键技术有**数据加密技术、数据真实性分析和认证技术、访问控制技术、安全审计技术、数据溯源技术、APT攻击检测技术**等。

• 1. 数据加密技术

数据加密技术的基本思路是将原始信息(或称明文)经过加密密钥及加密函数转换，变成无意义的密文，实现信息隐蔽，而接收方则将此密文经过解密函数、解密密钥原成明文。

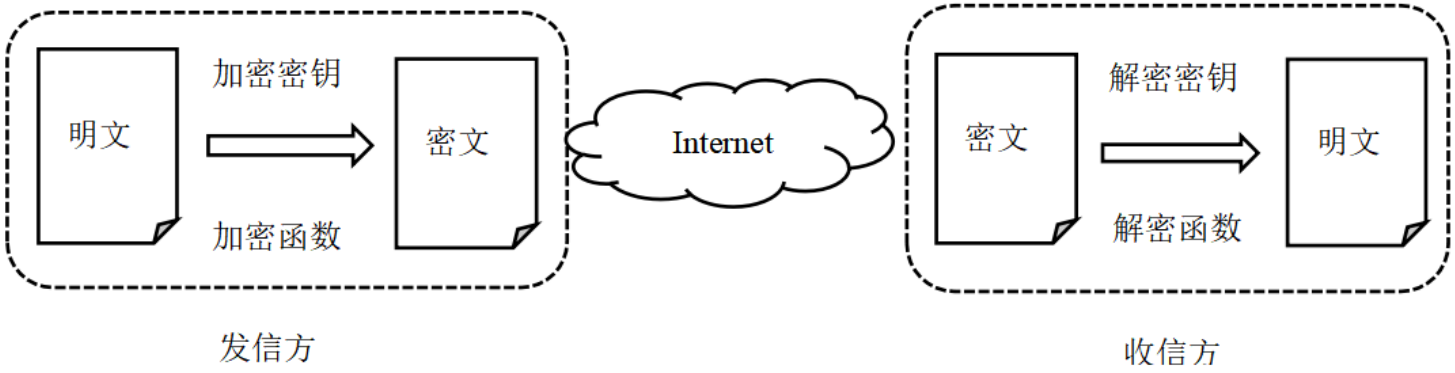


大数据安全相关技术

- **1. 数据加密技术**

（1）同态加密（Homomorphic Encryption）

明文（Plain Text）：没有加密的原始数据。
密文（Cypher Text）：加密以后的数据。
加密（Encryption）：把明文变换成密文的过程。
解密（Decryption）：把密文还原成明文的过程。
密钥（Key）：一般是单词、短语或一串数字，用于加密和解密的钥匙



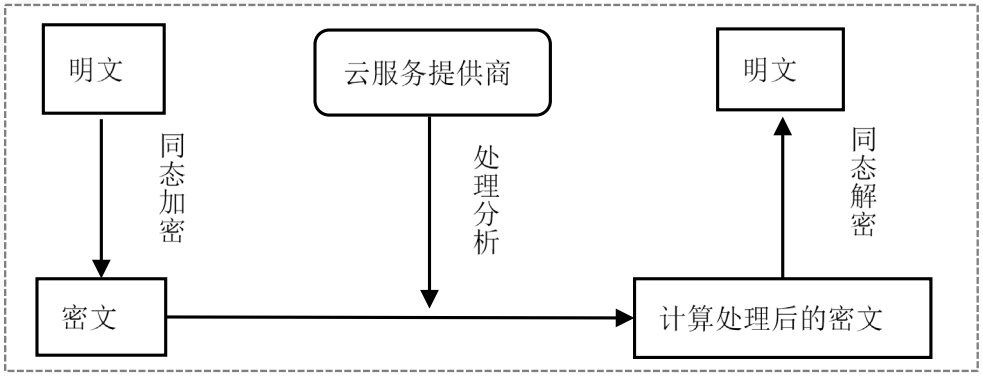
大数据安全相关技术

- **1. 数据加密技术**

（1）同态加密（Homomorphic Encryption）

一般的数据加密技术，用户是不能对密文做任何操作的，只能进行存储、传输，否则会导致错误的解密，甚至解密失败，因此**不能满足对除明文外的密文进行处理的需求**，而同态加密和可搜索加密可以在一定程度上解决以上难题。

同态加密除了能实现基本的加密操作之外，**还能实现密文间的多种计算功能**，即**先计算后解密可等价于先解密后计算**。



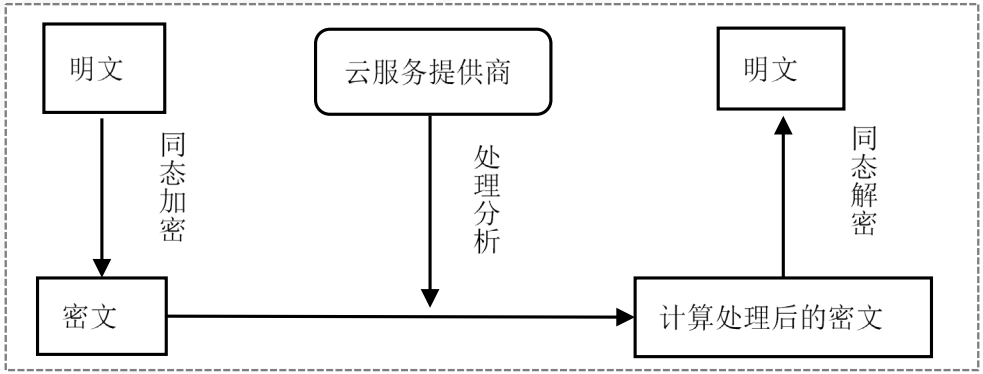
大数据安全相关技术

- **1. 数据加密技术**

（1）同态加密（Homomorphic Encryption）

用户采用同态加密对明文进行加密，并将加密后的数据发送至第三方云，充分利用云服务器的计算能力，实现数据处理分析；待第三方处理后将结果返回给用户，这个结果只有用户自身可以进行解密。

整个过程第三方云平台无法获知任何有效的数据信息，各种数据分析过程也不会泄露用户隐私。因此，同态技术可以使得在云环境下，实现对明文和密文信息的运算，而不会有损隐私数据。



大数据安全相关技术

• 1. 数据加密技术

（1）同态加密（Homomorphic Encryption）

同态加密技术具有以下一些优势：

1. **减少计算代价**
   计算复杂性上，可以先对多个密文进行计算之后再解密，不必对每一个密文解密而花费高昂的计算代价。

2. **减少通信代价**
   通信复杂性上，可以实现无密钥方对密文的计算，即密文计算无须经过密钥方，这样通过转移计算任务，平衡了各方的计算代价。

3. **安全性**
   可以实现让解密方只能获知最后的结果，而无法获得每一个密文的消息，从而提高信息的安全性。

大数据安全相关技术

• 1. 数据加密技术

（2）可搜索加密(Searchable Encryption)

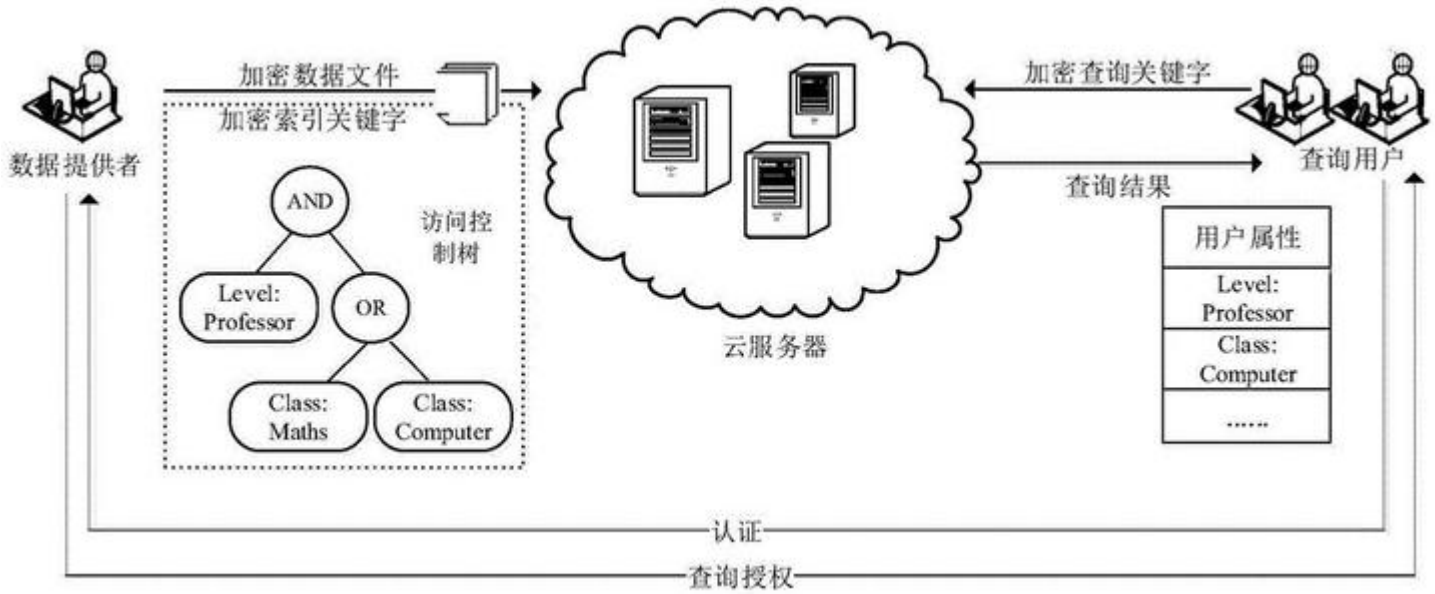
传统的数据加密技术，用户需要寻找包含某个关键字的相关信息时，只能按明文搜索，这造成了很严重的信息泄露，因为任意的恶意服务器都可以获取查询用户的查询关键字、查询结果等信息，严重危害个人的安全和隐私。可搜索加密技术解决了以上难题，它是一种**基于密文进行关键字搜索查询**的方案，在这种模式下，通过密码学的基本技术来保证用户的隐私信息和人身安全。

在应用上，可搜索加密技术非常适用于**云端隐私数据的防护**，不会降低对云端数据的提取和使用效率。在云端隐私数据的**高效共享**方面，可搜索加密也能发挥巨大作用，可以有效地支持最基本形式的隐私数据共享，即文件的发送和接收。

大数据安全相关技术

- 1. 数据加密技术

（2）可搜索加密(Searchable Encryption)



大数据安全相关技术

• 1. 数据加密技术

（2）可搜索加密(Searchable Encryption)

可搜索加密技术具有以下一些优势：

1 **安全性**

① 可证明安全，即不可信服务器仅仅通过密文不能获得有关明文的任何信息；

② 控制搜索安全，即不可信服务器不能在没有合法用户的认证下进行搜索；

③ 隐藏查询安全和查询独立安全，即用户向服务器发起有关一个关键字的查询，不可信服务器在整个搜索的过程中除了查询结果之外，不会获得搜索关键字内容以及任何明文信息。

2 **访问效率**

用户不需要为了没有包含关键字的文件浪费网络开销和存储空间；对关键字进行搜索的操作交由云端来执行，充分利用了云端强大的计算能力。

3 **资源节约**

用户不必对不符合条件的文件进行解密操作，节省了本地的计算资源。

大数据安全相关技术

• 2. 大数据真实性分析认证技术

为保证大数据的真实可信性，需要对大数据的发布者做认证检测，如利用**数字签名、数字水印、口令等认证技术**，近年来，**指纹、人脸等生物识别**等方式也在各个领域投入使用。另外，随着数据挖掘技术的发展，一种**基于数据挖掘的认证技术**也应运而生。

（1）数字签名（Digital Signature）

数字签名是一种通过密码技术对电子文档形成的签名，结合了哈希算法等公钥加密技术，是加密后得到的一段数字串，如十六进制形式的一串字符“A00117EFF3132……3CB2”。目的是保证发送信息的真实性和完整性，防止欺骗和抵赖的发生。

大数据安全相关技术

- **2. 大数据真实性分析认证技术**

（1）数字签名（Digital Signature）

基本原理：每个人都有一对数字身份，其中一个只有本人知道，称为**私钥**，另一个公开的，称为**公钥**。**一般公钥用于加密，私钥用于解密，或用私钥实现数字签名，而用公钥来验证签名。**

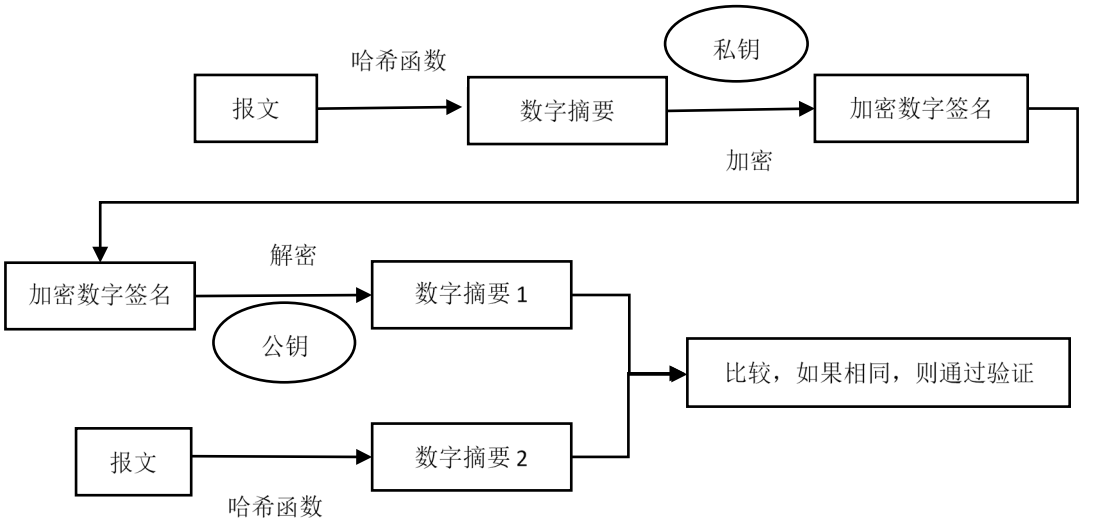
哈希函数的输入为任意长度的消息M，输出为一个固定长度的散列值，称为消息摘要(Message Digest）。**哈希函数是消息M的所有位的函数并提供错误检测能力，即消息中的任何一位或多位的变化都将导致该散列值的变化。**

大数据安全相关技术

• 2. 大数据真实性分析认证技术

（1）数字签名（Digital Signature）

数字签名的过程：发送方用一个哈希函数从报文文本中生成数字摘要，然后用发送方的私钥对这个摘要进行加密，这个加密后的摘要将作为报文的数字签名和报文一起发送给接收方。



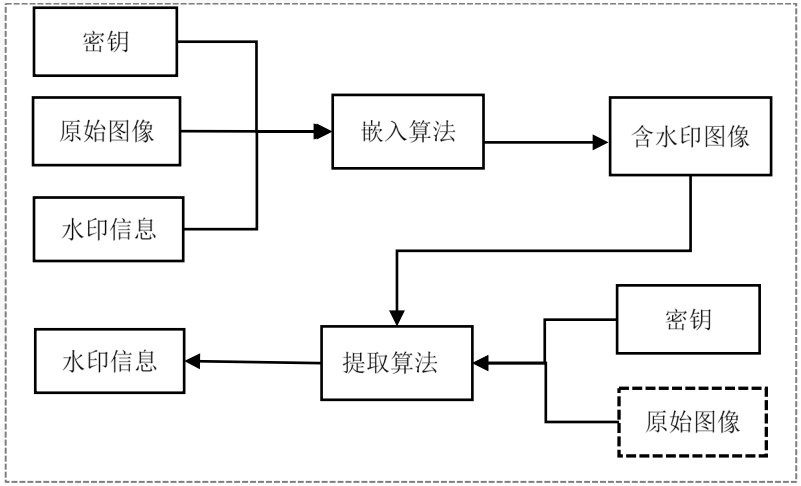
数字签名的验证过程：接收方首先用自己的公钥来对报文附加的数字签名进行解密，再用与发送方一样的哈希函数从接收到的原始报文中计算出报文摘要，如果两个摘要相同，那么接收方就能确认该数字签名是发送方的

大数据安全相关技术

- **2. 大数据真实性分析认证技术**

（2）数字水印（Digital Watermark）

数字水印是一种应用计算机算法**嵌入载体文件的防护信息**。数字水印技术是一种基于内容的、非密码机制的**计算机信息隐藏技术**，它将标识信息（即数字水印）以难以察觉的方式直接嵌入数据载体内部且原载体的使用价值，也不容易被探知和再次修改，但是可以被生产方识别和辨认，用以确定数字产品的所有权或检验数字内容的原始性。

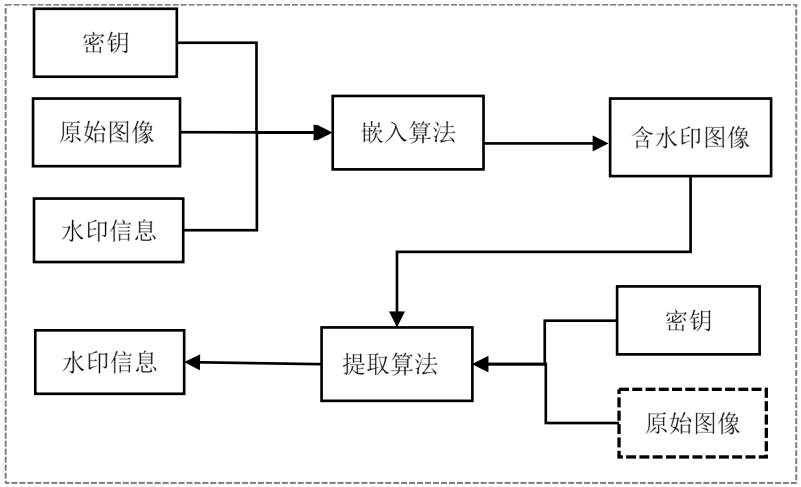


大数据安全相关技术

- **2. 大数据真实性分析认证技术**

（2）数字水印（Digital Watermark）

数字水印的嵌入过程：将密钥、原始图像和水印信息作为嵌入算法的输入，输出含水印的图像。在某些水印系统中，水印可以被精确地提取出来，这一过程被称作**水印提取**。通过提取出的水印的完整性，可判断原始数据的完整性，如果提取出的水印发生了部分的变化，说明原始数据被篡改，而且还能通过变化的水印的位置来确定原始数据被篡改的位置。



大数据安全相关技术

• 2. 大数据真实性分析认证技术

（3）基于数据挖掘的认证技术

指的是收集用户行为和设备数据，并对这些数据进行分析，通过鉴别操作者行为及其设备使用信息来确定其身份。相比数字签名和数字水印技术，该技术具有以下优点：

1. **安全性**
利用大数据技术所能收集的用户行为和设备特征数据是多样的，如用户使用系统的时间、设备、位置信息、操作习惯、消费数据等。通过这些数据能够建立用户行为特征轮廓，而攻击者很难在方方面面都模仿到用户行为，两者之间必然存在一个较大偏差，因此，攻击者模仿的用户信息很难被认证通过。

2. **减轻了用户负担**
用户行为和设备特征数据的采集、存储和分析都由认证系统完成，避免了由于用户所持有凭证不同而带来的种种不便。

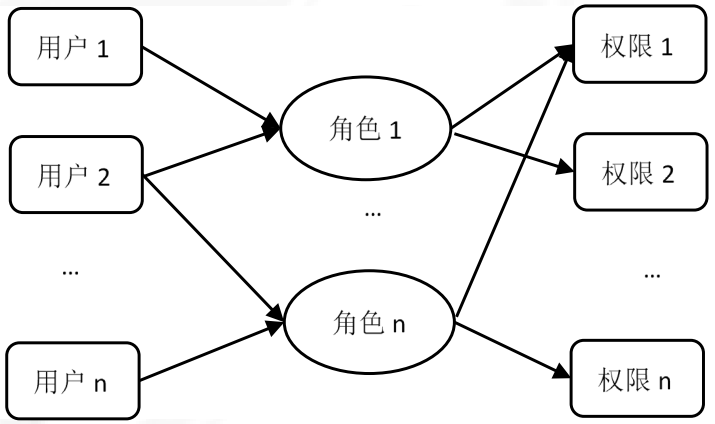
3. **更好地支持各系统认证机制的统一**
可让用户在整个网络空间采用相同的行为特征进行身份认证，而避免不同系统采用不同认证方式。

大数据安全相关技术

- **3. 访问控制技术**

（1）基于角色的访问控制（Role-Based Access Control，RBAC）

基本思想是对系统操作的各种权限不是直接授予具体的用户，而是在**用户集合与权限集合之间建立一个角色集合**，即一个用户拥有若干角色，每一个角色拥有若干权限，**这样就构造成用户-角色-权限的授权模型**。在这种模型中，用户与角色之间，角色与权限之间，一般者是多对多的关系



大数据安全相关技术

- **3. 访问控制技术**

（1）基于角色的访问控制（Role-Based Access Control，RBAC）

优点：

不必在每次创建用户时都进行分配权限的操作，只要分配用户相应的角色即可，而且角色的权限变更比用户的权限变更要少得多，这样将简化用户的权限管理，减少系统的开销，提高企业安全策略的灵活性。

缺点：

在分布式环境下，存在严重的管理规模和控制粒度问题，并不能发挥系统所满意的效果；它不能抵抗合谋攻击，即多个成员联合起来解密资源。

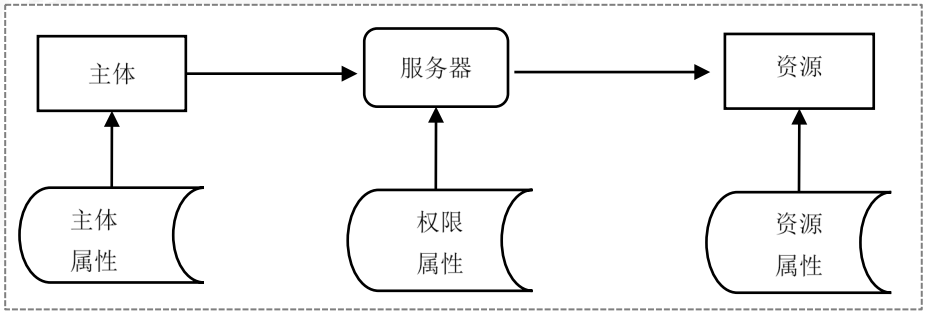
大数据安全相关技术

- **3. 访问控制技术**

（2）基于属性加密的访问控制（Attribute-based Encryption Access Control）

基本思想是用一系列属性集来描述用户的**身份信息（称为主体属性）和资源信息（称为资源属性）**，加密者在加密时设定**访问规则**，并以密文的形式存储在服务器上（**称为权限属性**）。

当接收者向服务器进行身份认证时，需要出示与自身属性相关的信任证书，**当接收者拥有的属性超过加密者所描述的预设门槛**时，用户便可对资源进行解密的，服务器对应的资源发送给接收者。



大数据安全相关技术

- **3. 访问控制技术**

（3）基于风险的访问控制（Risk-based Access Control）

由于大数据应用系统的复杂性，通常会存在一些特定的访问需求在设计策略时没有考虑，或者访问需求的变化引起访问控制策略不再适合等。如果严格按照预先定义的策略执行访问控制，将产生授权不足无法完成业务的情况。而基于风险的访问控制不再严格地按照预先分配的权限进行访问控制，而是开始**衡量访问行为所带来的风险是否为系统可接受的。**因此，当发生一些未预料到的访问行为时，若其风险是可接受的，则仍然可以允许该访问。

大数据安全相关技术

• 4. 数据溯源技术

（1）标记法

是指**用标注的方式来记录原始数据的一些重要信息**，如原始信息的背景、作者、时间、出处等，并**让标注和数据一起传播**，最后通过查看目标数据的标注来获得数据的溯源。

标注法具有实现简单、容易管理等优点，但缺点是只适合小型系统，对于大型系统而言，很难为细粒度的数据提供详细的数据溯源信息；此外还需要额外的存储空间，对存储造成很大的压力。

大数据安全相关技术

• 4. 数据溯源技术

（2）反向查询法

反向查询法通过**构造原函数的反函数对查询求逆**，由结果追溯到原数据，更适合于细粒度数据。

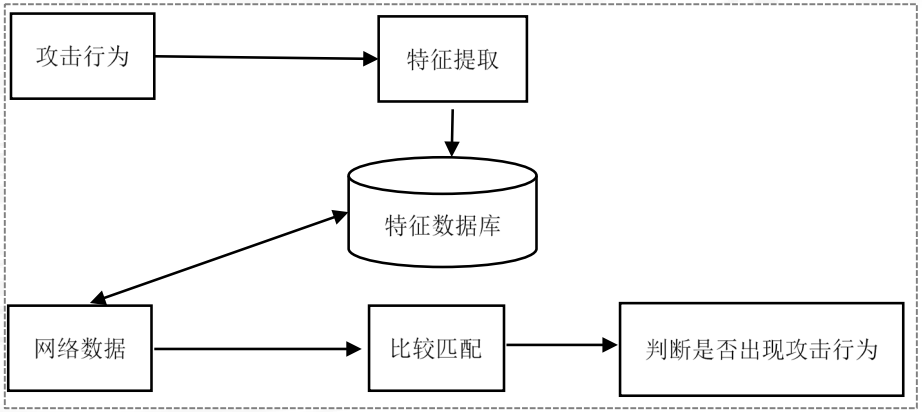
与标注法相比，它需要的存储空间更小，并且追踪比较简单，只需存储少量的元数据就可实现对数据的溯源追踪。其缺点是用户需要提供反函数和相对应的验证函数，但并不是所有函数都具有反函数，因此具有一定局限性，实现相对比较复杂。

大数据安全相关技术

- **5. 大数据安全审计技术**

（1）基于规则的安全审计

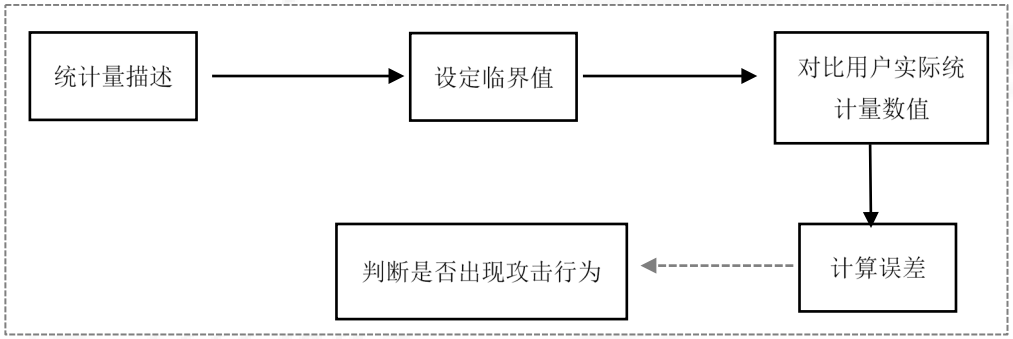
基本思想是：将已知的攻击行为进行**特征提取**，之后放入**特征数据库**中，当进行安全审计分析时，将收集到的网络数据与特征数据库中的特征进行**比较匹配**，判断是否出现网络攻击行为，对此采取相应的响应机制。



- **5. 大数据安全审计技术**

（2）基于统计的安全审计

基本思路是：统计正常情况下对象的**统计量描述**，比如一个网络流量的平均值、方差等，审计人员根据经验**设定临界值**，**即正常数值和非正常数值的分界点**，然后与用户实际生活中使用的统计量数值进行对比，根据与临界值的误差大小判断是否收到网络攻击，采取相应的响应机制。



大数据安全相关技术

- **5. 大数据安全审计技术**

（3）基于机器自学习的安全审计

对于已知的入侵模式，基于规则和统计的安全审计方法能较好地应对，但不适用于未知的入侵模式。而基于机器自学习的安全审计能通过**数据挖掘分析和关联分析**，**对未知的入侵模式提供更快的异常活动的检测**，更有针对性地观察事件行为趋势，从而对可疑行为进行预警。

大数据安全相关技术

• 6. APT攻击检测技术

APT攻击，也称高级可持续威胁攻击，是指某组织对特定对象展开的持续有效的攻击活动，主要特点是有组织、目标明确、持续性、破坏力大、隐蔽性。

APT攻击的检测难度主要表现在以下3个方面：

1 **先进的攻击方法**

攻击者能适应防御者的入侵检测能力，**不断更换和改进入侵方法**，具有较强的隐藏能力，攻击入口、途径、时间都是不确定和不可预见的，使得基于特征匹配的传统检测防御技术很难有效检测出攻击。

2 **持续性攻击与隐藏**

APT通过长时间攻击成功进入目标系统后，通常采取**隐藏策略进入休眠状态**，待时机成熟时，才利用时间间隙与外部服务器交流，在系统中并无明显异常，这使得基于单点时间或短时间窗口的实时检测技术和会话频繁检测技术也难以成功检测出异常攻击。

大数据安全相关技术

• 6. APT攻击检测技术

3 长期驻留目标系统

①攻击者一旦侵入目标系统便会积极争取目标系统或网络的最高权限，**实现程序的自启功能**。

②攻击者会在目标网络中基于已控制的网络主机**实现横向转移和信息收集**，规避安全检测，扩大被入侵网络的覆盖面，寻找新的攻击目标。

③一旦其找到了想要攻击的最终目标和适当传送信息的机会，攻击者便会通过事先准备好的隐藏通道获取信息、窃取数据或执行破坏活动，且不留任何被入侵的痕迹。

常用的APT攻击**检测技术**主要有网络流量异常检测、主机恶意代码异常检测和社交网络安全事件挖掘等。

大数据安全相关技术

• 6. APT攻击检测技术

1 **网络流量异常检测**

使用数据流抓取工具采集网络数据流信息，以此作为输入，并提取和选择用于检测异常的数据属性，然后通过统计分析、数据挖掘和机器学习等方法，发现异信息。

2 **恶意代码异常检测**

通过数据挖掘技术**建立恶意代码特征数据库**，然后对海量样本程序的特征进行关联分析，从而识别代码是否具有恶意行为，它可以有效检测数量快速增长的未知恶意程序。

3 **社交网络安全事件挖掘**

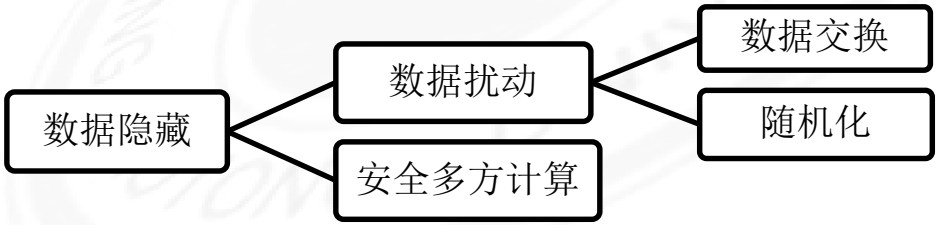
从社交网络海量数据中**挖掘分析用户正常行为模式、社交关系网、用户间的信任关系等社会属性**，通过在线监控**将违背行为模式和信任关系的异常行为归纳为威胁事件**，并据此快速定位攻击者的不轨行为和社会属性，进而为攻击检测、计算机取证和信息安全防护提供指导和依据。

大数据隐私保护技术

目前应用最广泛的隐私保护技术有**数据隐藏、数据脱敏、数据发布匿名技术、基于差分隐私的数据发布技术**等。

• 1. 数据隐藏技术

即使是经过匿名处理后的数据，通过关联分析、聚类、分类等数据挖掘方法后，依然可以分析出用户的隐私。数据隐藏技术是一种针对数据挖掘的隐私保护技术，目的是在保证大数据可用性的前提下，防范数据发掘方法所引发的隐私泄露。



大数据隐私保护技术

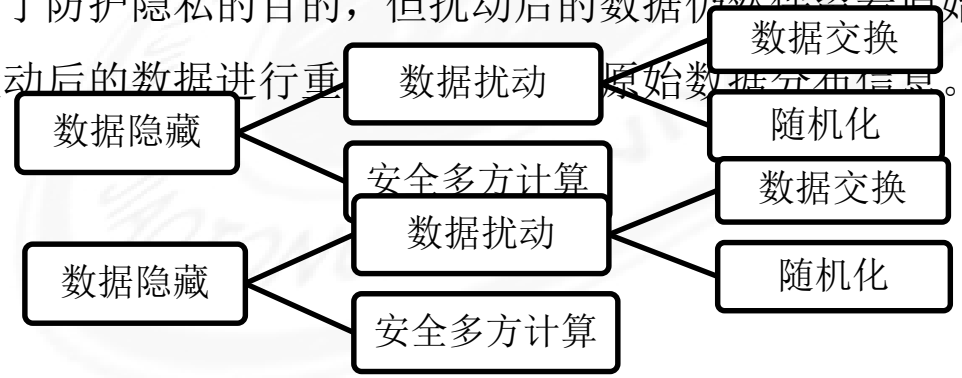
• 1. 数据隐藏技术

（1）数据扰动技术（Data Perturbation）

数据扰动技术即对数据进行变换，使其中敏感信息被隐藏，只呈现出数据的统计学特征。

① 数据交换即在记录之间交换数据的值，保留某些统计学特征而不保留真实数值。

② 随机化是指在原始数据中添加一些噪声，然后发布扰动后的数据，从而隐藏真实数值，达到了防护隐私的目的，但扰动后的数据仍然保留着原始数据的分布信息，通过对扰动后的数据进行重[illegible]原始数据分布信息。

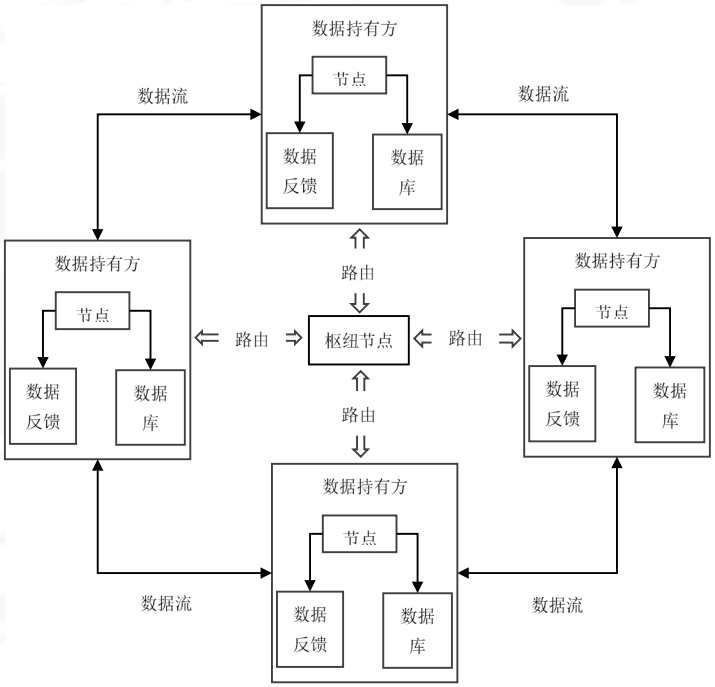


大数据隐私保护技术

- **1. 数据隐藏技术**

（2）安全多方计算（Secure Muti-Party Computation，SMC）

安全多方计算是指针对无可信第三方的情况下，**允许多个数据拥有者进行协同计算**，输出计算结果，该计算方式确保**各个参与者只能得到既定的输出结果**，参与者的任何隐私信息不会被泄露。



大数据隐私保护技术

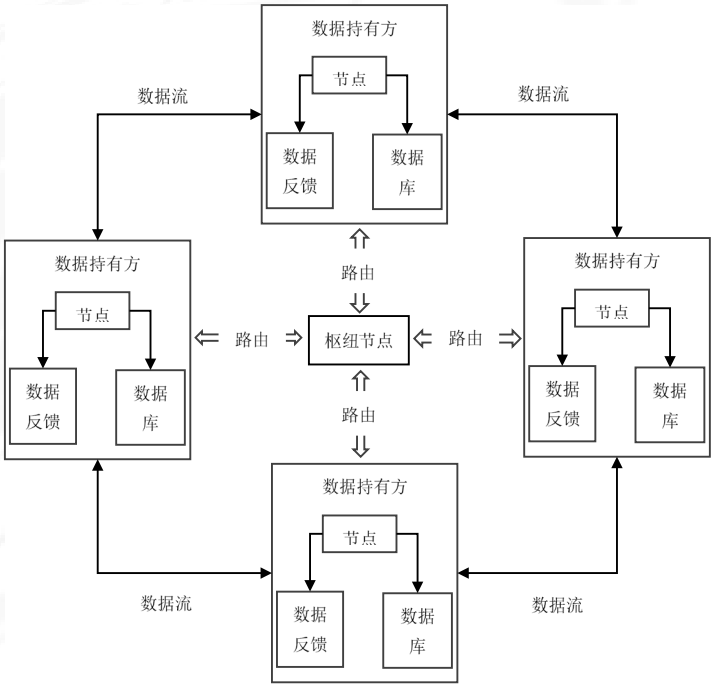
• 1. 数据隐藏技术

（2）安全多方计算（Secure Muti-Party Computation，SMC）

当一个安全多方计算任务发起时，枢纽节点传输网络及信令控制。每个数据持有方可发起协同计算任务：

- 通过**枢纽节点**进行路由寻址，选择其余数据持有方进行安全的协同计算；
- 从**本地数据库**查询所需数据，共同就安全多方计算任务在数据流间进行协同计算。

在保证输入隐私性的前提下，各方得到正确的**数据反馈**，整个过程中本地数据没有泄露给其他任何参与方。



大数据隐私保护技术

- **1. 数据隐藏技术**

（2）安全多方计算（Secure Muti-Party Computation，SMC）

特点：

① **输入隐私性：**安全多方计算过程中必须保证各方私密输入独立，计算时不泄露任何本地数据。

② **计算正确性：**各参与方通过安全多方计算协议进行协同计算，计算结束后，各方得到正确的数据反馈。

③ **去中心化：**传统的分布式计算由中心节点协调各用户的计算进程，收集各用户的输入信息，而安全多方计算中，各参与方地位平等，不存在任何有特权的参与方或第三方，提供一种去中心化的计算模式。

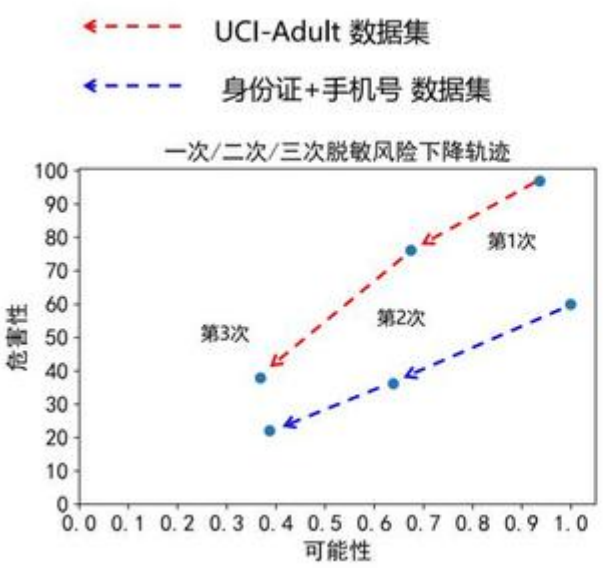
大数据隐私保护技术

• 2. 数据脱敏 (Data Masking)

数据脱敏是指对某些识别到的敏感信息**通过脱敏规则进行数据的变形**，实现敏感隐私数据的可靠保护。

在涉及客户安全数据或者一些商业性敏感数据的情况下，在不违反系统规则条件下，对真实数据进行改造并提供测试使用，如身份证号、手机号、卡号、客户号等个人信息都需要进行数据脱敏。

识别数据对象中的敏感信息时，通常采用**自动化敏感信息识别技术和机器学习方法**，构建已知敏感信息知识库，而后对疑似敏感信息进行匹配。



大数据隐私保护技术

• 3. 数据发布匿名技术

数据发布匿名是匿名技术在数据发布中的应用，在确保所发布的数据在公开可用的前提下，**隐藏数据记录与特定个人之间的对应联系**，从而防护个人隐私。典型的数据发布匿名技术有k-匿名、l-diversity匿名、m-invariance匿名等。

引入4个概念：

（1）**标识符**：能直接确定一个个体的属性，如用户ID，姓名等。

（2）**准标识符集**：通过和外部表连接来间接确定一个个体的最小属性集，如｛省份，出生时间，性别，邮编｝。

（3）**链式攻击**：攻击者通过对发布的数据和从其他渠道获取的外部数据进行链接操作，以推理出隐私数据。

（4）**数据泛化**：**用较高层次的概念替换较低层次的概念**，从而**汇总**数据，例如把年龄的具体数值范围替换为青年、中年和老年层。

大数据隐私保护技术

- **3. 数据发布匿名技术**

***k*-匿名通过对数据进行泛化，发布精度较低的数据，使得同一个准标识符集至少有*k*条记录，观察者便无法通过准标识符链接记录。**

表 9.5 原始信息表

用户 ID	邮编	年纪	病种
1	47677	29	Heart Disease
2	47602	22	Flu
3	47679	27	Cancer
4	47905	43	Flu
5	47909	52	Heart Disease
6	47906	47	Cancer
7	47605	30	Heart Disease
9	47673	36	Cancer
9	47607	32	Cancer

表 9.6 经过 *K*-匿名处理后的信息表

用户 ID	邮编	年纪	病种
1	476**	2*	Heart Disease
2	476**	2*	Flu
3	476**	2*	Cancer
4	479**	>40	Flu
5	479**	>40	Heart Disease
6	479**	>40	Cancer
7	476**	3*	Heart Disease
9	476**	3*	Cancer
9	476**	3*	Cancer

• 4. 基于差分隐私的数据发布

差分隐私(Differential Privacy)是密码学中的一种手段，当从统计数据库查询数据时，能在**保留统计学特征的前提下去除个体特征**，最大限度减少识别用户隐私记录的机会，同时保证个人隐私的泄露风险不超过预先设定的风险阈值，常用的差分隐私的方法是**对数据加入噪音进行扰动**。

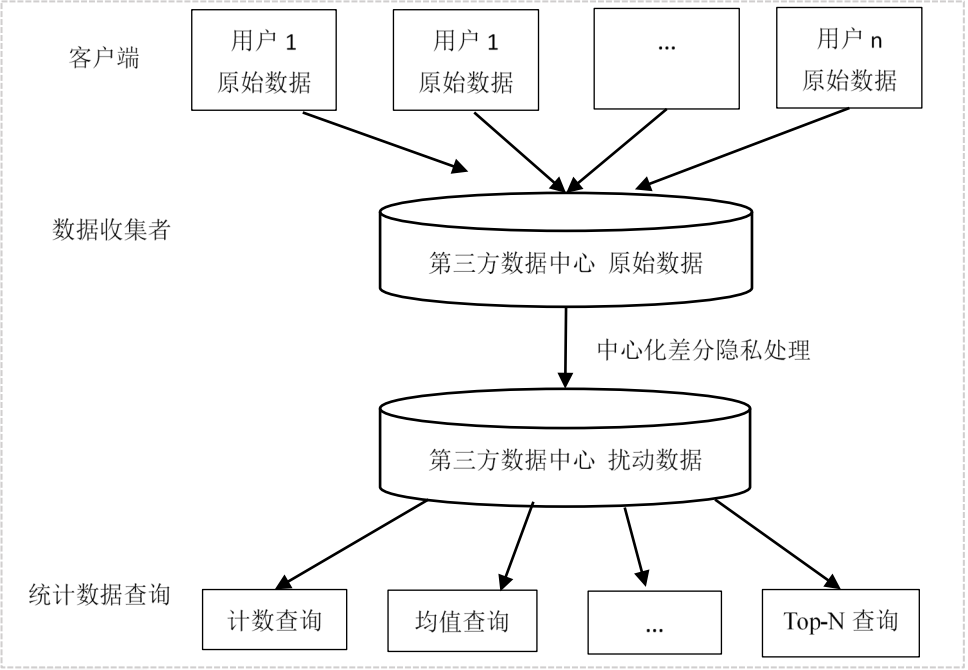
根据数据隐私化处理实施者的不同，差分隐私可分为中心化差分隐私(Centralized Differential Privacy，CDP)和本地化差分隐私(Local Differential Privacy，LDP)。

大数据隐私保护技术

- **4. 基于差分隐私的数据发布**

（1）中心化差分隐私

处理流程：数据收集者将多源客户端原始数据**汇集到第三方数据中心**，并由数据中心进行满足差分隐私的数据扰动，对外发布扰动数据后即可用于统计数据查询。

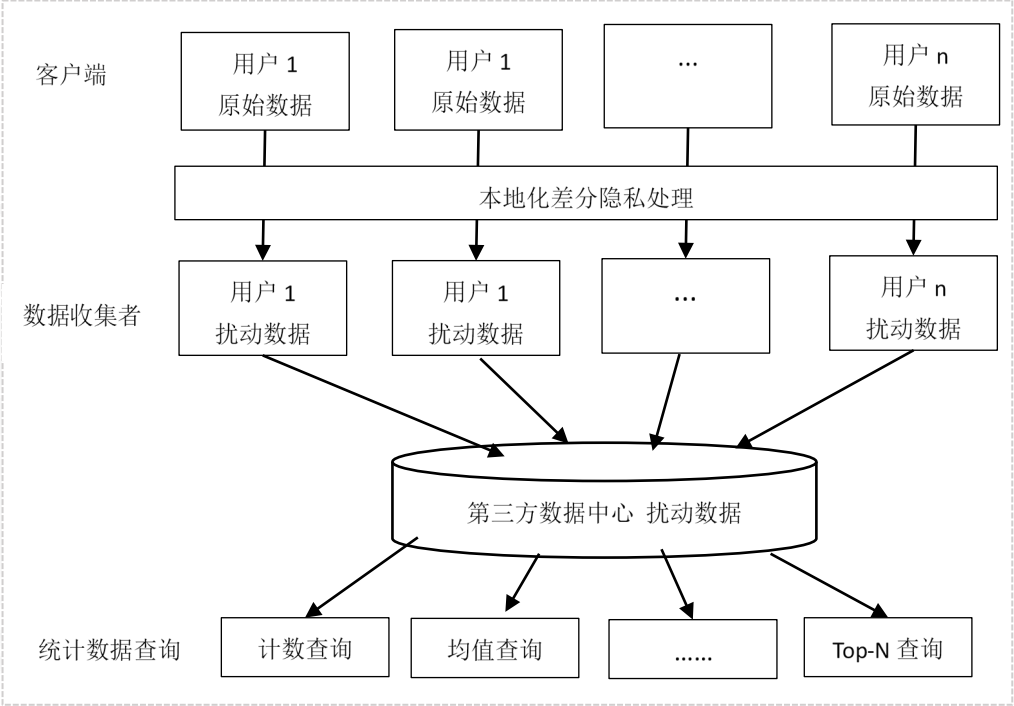


大数据隐私保护技术

- **4. 基于差分隐私的数据发布**

（2）本地化差分隐私

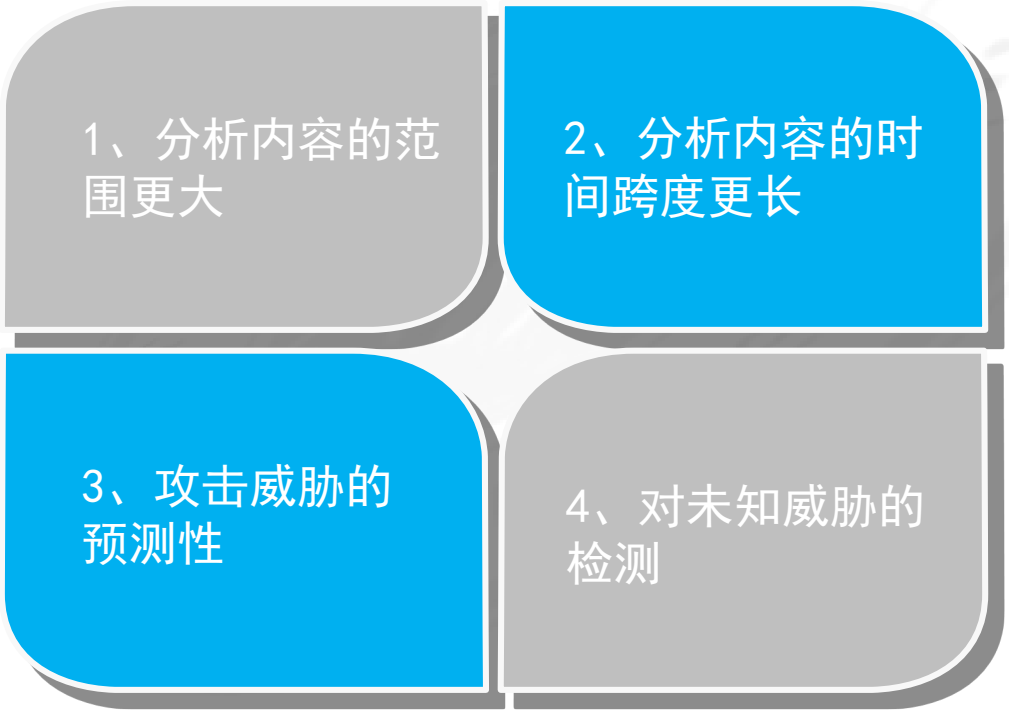
本地化差分隐私是针对第三方数据收集者的隐私处理操作的非可信性提出的，首先由客户端的用户在**本地进行满足差分隐私的数据扰动**，再将扰动数据发送给收集者，汇集在第三方数据中心。



大数据的应用

- **基于大数据的威胁发现技术**

基于大数据，企业可以更主动的发现潜在的安全威胁
相较于传统技术方案，大数据威胁发现技术有以下优点：



大数据的应用

- **基于大数据的认证技术**

身份认证：信息系统或网络中确认操作者身份的过程，传统认证技术只要通过用户所知的口令或者持有凭证来鉴别用户

传统技术面临的问题：

1、攻击者总能找到方法来骗取用户所知的秘密，或窃取用户凭证

2、传统认证技术中认证方式越安全往往意味着用户负担越重

大数据的应用

- 基于大数据的认证技术

基于大数据的认证技术：收集用户行为和设备行为数据，对这些数据分析，获得用户行为和设备行为的特征，进而确定其身份。