北京交通大学
BEIJING JIAOTONG UNIVERSITY

《大数据概论》
大数据安全与隐私保护

鲍鹏
软件学院

本章内容

- 大数据安全概述
- 大数据隐私问题
- 大数据安全技术

大数据安全的概念

- 大数据安全：指确保数据的保密性、完整性和可用性，不受到安全威胁影响。
  - 保密性：禁止用户在没有授权的情况下获取数据。
  - 完整性：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。
  - 可用性：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。

大数据安全问题形成原因

- 1. 大数据安全问题形成原因
  - 传统数据安全防护技术的缺陷
  - 大数据分布式存储的风险
  - 大数据平台安全机制的不足
  - 新型虚拟化网络技术的局限
  - 新型高级网络攻击的威胁

大数据安全问题形成原因

- (1) 传统数据安全防护技术的缺陷
  - 当前，网络攻击攻击效果已经从易察觉的系统宕机、信息泄露转向细小难以察觉的结果偏差。
  - 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。

大数据安全问题形成原因

- (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。
  - 如果云端服务器被攻陷，海量信息可在瞬间被集中窃取。

大数据安全问题形成原因

- (3) 大数据平台安全机制的不足：Hadoop为例
  - 用户的身份鉴别和授权访问等安全保障能力比较薄弱。
  - Hadoop中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。

大数据安全问题形成原因

- (4) 新型虚拟化网络技术的局限：
  - 接口的开放性会引发漏洞暴露和接口滥用的问题。
  - 网络功能虚拟化（NFV）部署时通常会外包给第三方虚拟化平台，易发生安全问题。

大数据安全问题形成原因

- (5) 新型高级网络攻击的威胁：
  - 例如高级可持续攻击（APT），攻击者将APT攻击代码长期隐蔽在大数据中。
  - APT攻击的发现难度更大，严重威胁着网络安全。

大数据安全问题的分类

- 2. 大数据安全问题的分类：
  - 大数据平台安全
  - 大数据自身安全
  - 大数据应用安全

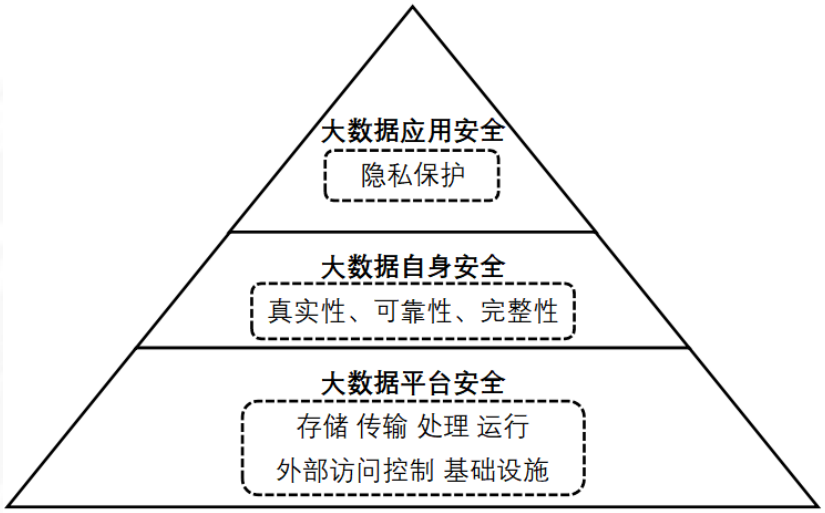


图 9.3 大数据安全的三个层面

大数据安全问题的分类

- (1) 大数据平台安全：
  - 大数据存储安全
  - 大数据传输安全
  - 大数据平台访问控制安全
  - 大数据运行计算安全
  - 大数据基础设施安全

大数据安全问题的分类

- (2) 大数据自身安全：
  - 保障数据源的真实可信性，防止源数据被伪造或刻意制造。
  - 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。

大数据安全问题的分类

- (3) 大数据应用安全：如用户隐私问题
  - 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。
  - 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。

本章内容

- 大数据安全概述
- 大数据隐私问题
- 大数据安全技术

大数据隐私问题

- 隐私权的定义：
  - 定义为“不受干涉”或“免于侵害”的独处权利。
  - 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。

大数据隐私问题

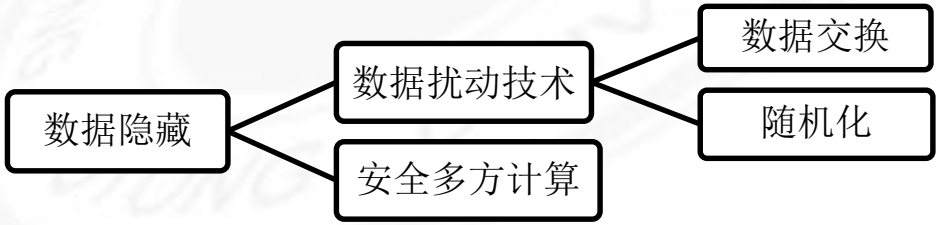
- 大数据时代的隐私问题的变化呈现如下特征：
  - 隐私范围扩大且难以界定
  - 隐私权利归属复杂
  - 隐私保护困难

大数据隐私问题

- 大数据隐私保护技术有：
  - 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

大数据隐私问题

- 1. 数据隐藏技术：
  - 针对数据挖掘的隐私保护技术。
  - 防范数据发掘方法所引发的隐私泄露。

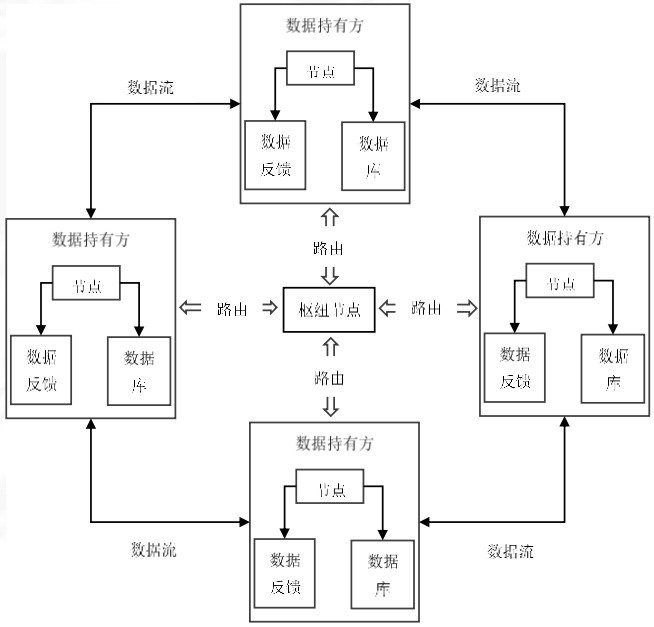


大数据隐私问题

- (1) 数据扰动技术：对数据进行变换，使其中敏感信息被隐藏。
  - 数据交换：在记录之间交换数据的值。
  - 随机化：在原始数据中添加一些噪声。

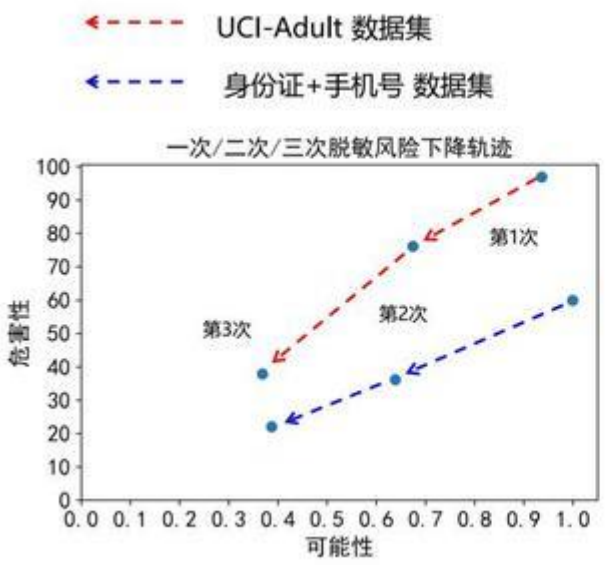
大数据隐私问题

- (2) 安全多方计算
  - 针对无可信第三方情况，允许多个数据拥有者进行协同计算。
  - 输出计算结果，确保各个参与者只能得到既定的输出结果。
  - 保证参与者的任何隐私信息不会被泄露。



大数据隐私问题

- 2. 数据脱敏技术：对敏感信息根据脱敏规则，进行数据变形，实现隐私数据保护。



大数据隐私问题

- 3. 数据发布匿名技术：隐藏数据记录与特定个人之间的对应联系。

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

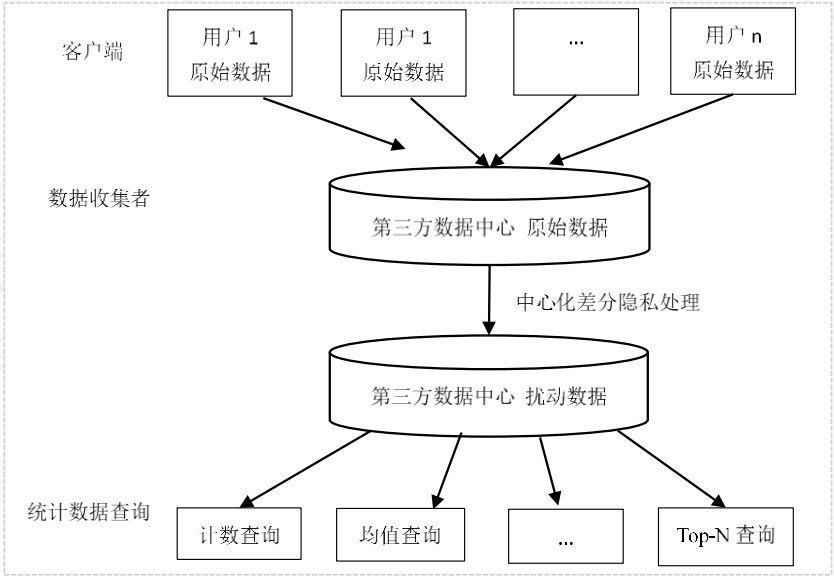
大数据隐私问题

- 4. 基于差分隐私的数据发布：保留统计学特征的前提下去除个体特征。

- 根据数据隐私化处理实施者的不同，分为：
  - 中心化差分隐私
  - 本地化差分隐私

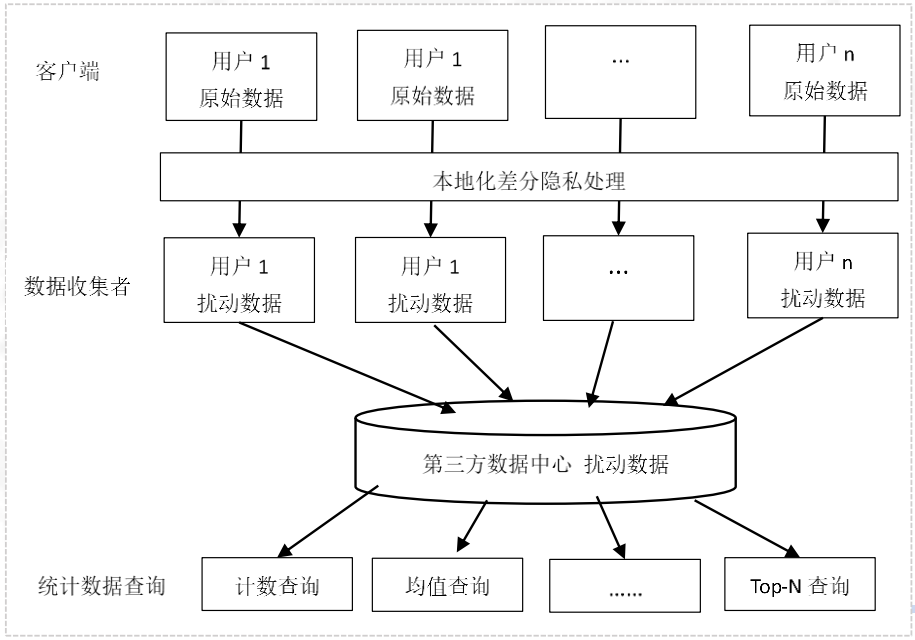
大数据隐私问题

- (1) 中心化差分隐私
  - 数据收集者将数据汇集到第三方数据中心。
  - 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。



大数据隐私问题

- (2) 本地化差分隐私
  - 由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。



本章内容

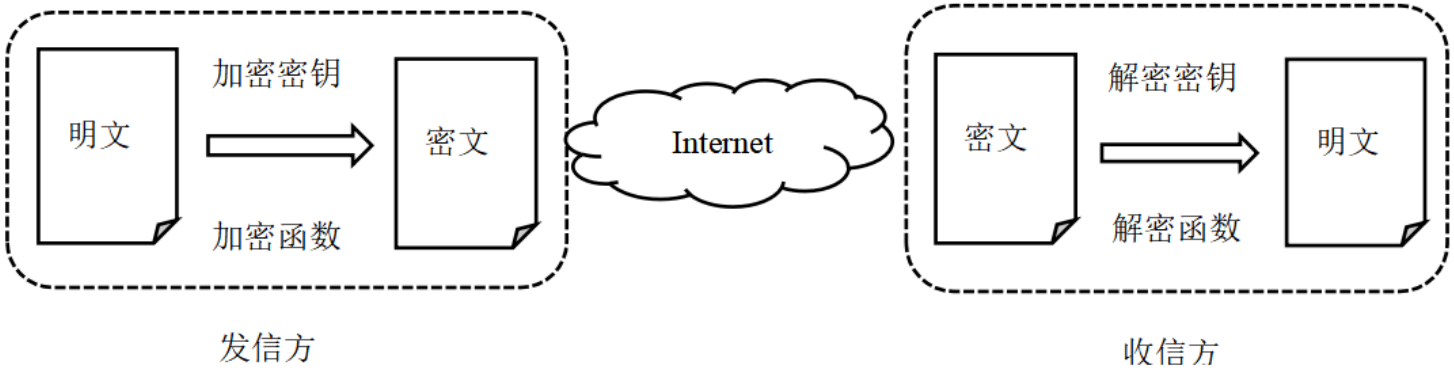
- 大数据安全概述
- 大数据隐私问题
- 大数据安全技术

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

大数据安全相关技术

- 1. 数据加密技术：
  - 加密阶段：将原始信息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。
  - 解密阶段：接收方则将此密文经过解密函数、解密密钥还原成明文。



大数据安全相关技术

- 2. 大数据真实性分析认证技术：保证大数据的真实可信性，对大数据的发布者进行认证检测。
  - 数字签名
  - 数字水印
  - 基于数据挖掘的认证技术

大数据安全相关技术

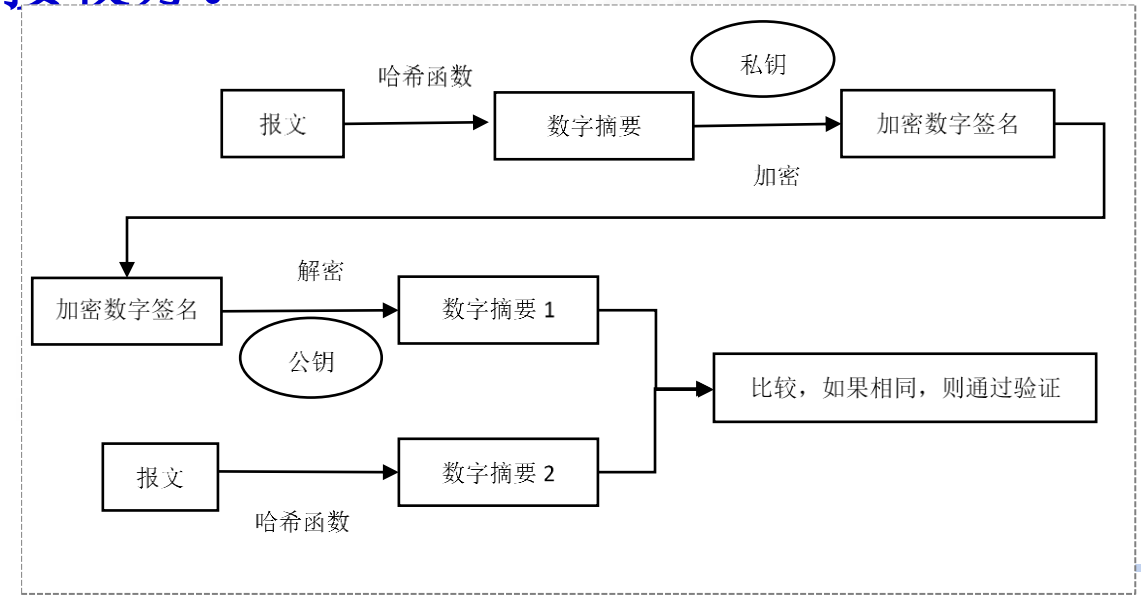
- (1) 数字签名（Digital Signature）技术：
  - 通过密码技术对电子文档形成签名。
  - 结合了哈希算法等公钥加密技术。
  - 目的是保证发送信息的真实性和完整性。

大数据安全相关技术

- 数字签名技术的相关定义：
  - 公钥：公钥用于加密信息和解密数字签名。
  - 私钥：私钥用于解密信息和加密消息摘要。
  - 消息摘要：对消息使用哈希算法获取的固定长度的字符串。
  - 数字签名：使用私钥加密的消息摘要。

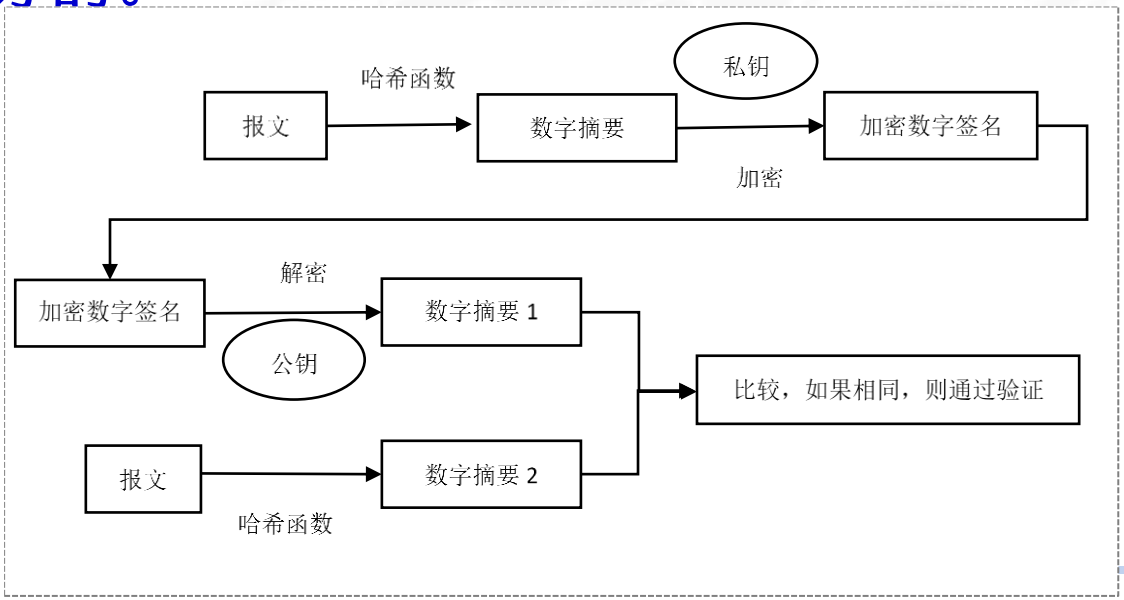
大数据安全相关技术

- 数字签名技术的加密流程：
  - 发送方用哈希函数从报文文本中生成数字摘要。
  - 用发送方的私钥对这个摘要进行加密。
  - 加密后的摘要作为报文的数字签名和报文一起发送给接收方。



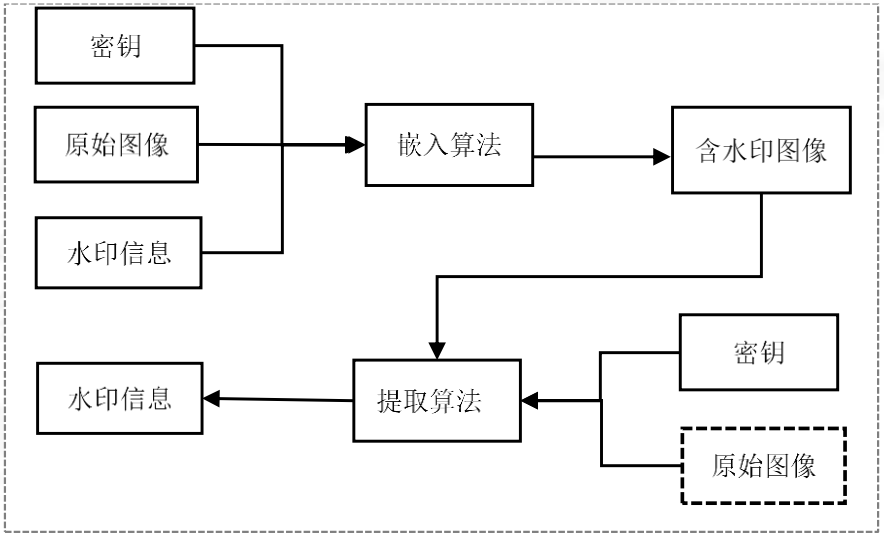
大数据安全相关技术

- 数字签名技术的解密流程：
  - 接收方用公钥来对报文附加的数字签名进行解密。
  - 再用相同的哈希函数从原始报文中计算报文摘要。
  - 两个摘要相同，则接收方就能确认该数字签名是发送方的。



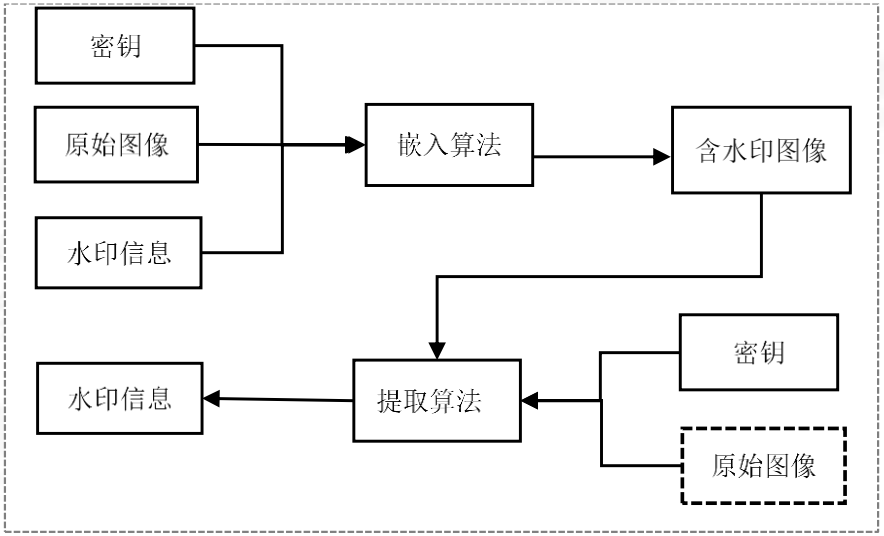
大数据安全相关技术

- (2) 数字水印（Digital Watermark）技术：
  - 应用计算机算法嵌入载体文件的防护信息
  - 以难以察觉的方式直接嵌入数据载体内部
  - 主要分为嵌入过程和提取过程



大数据安全相关技术

- 数字水印技术的过程：
  - 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，输出为含水印的图像。
  - 提取过程：根据水印的完整性，判断原始数据的完整性。



大数据安全相关技术

- (3) 基于数据挖掘的认证技术：
  - 收集用户行为和设备数据。
  - 对这些数据进行分析。
  - 鉴别操作者行为及其设备使用信息来确定身份。

大数据安全相关技术

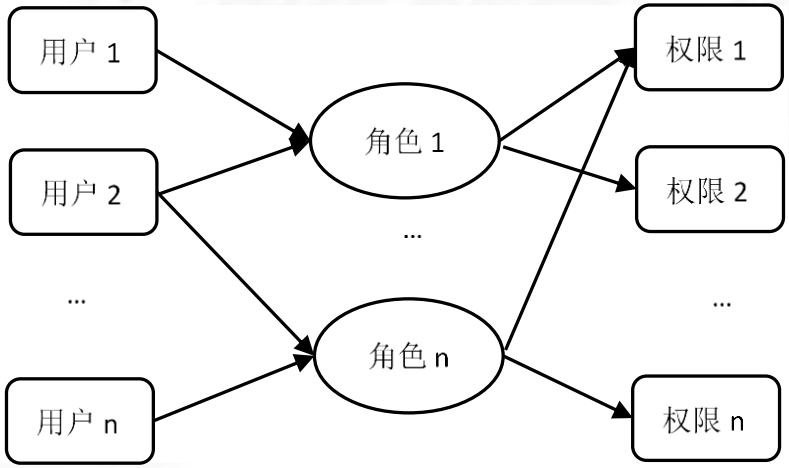
- 基于数据挖掘技术的优势：
  - 安全性：攻击者很难再模仿到用户行为。
  - 减轻用户负担：避免了由于用户所持有凭证不同而带来的种种不便。
  - 认证机制统一：避免不同系统采用不同认证方式。

大数据安全相关技术

- 3. 访问控制技术。
  - 基于角色的访问控制
  - 基于属性加密的访问控制
  - 基于风险的访问控制

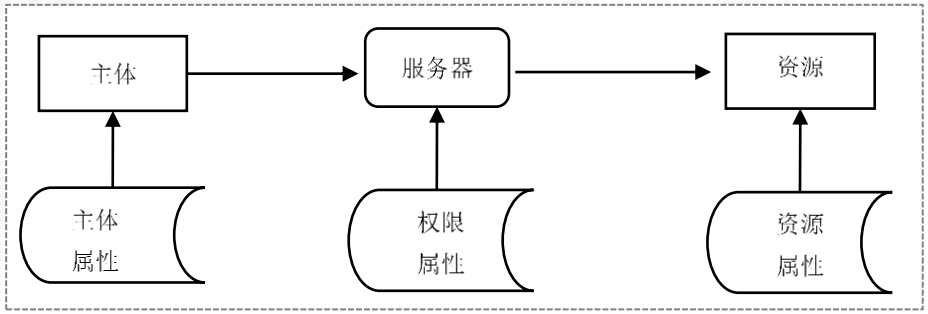
大数据安全相关技术

- (1) 基于角色的访问控制：
  - 不是直接授予具体权限给用户。
  - 在用户集合与权限集合之间建立一个角色集合。
  - 构造成用户-角色-权限的授权模型。



大数据安全相关技术

- (2) 基于属性加密的访问控制：
  - 用一系列属性集来描述用户的身份信息。
  - 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) 基于风险的访问控制：
  - 大数据应用系统的复杂性，会存在一些特定的访问需求在设计策略时没有考虑。
  - 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。
  - 基于风险的访问控制开始衡量访问行为所带来的风险是否为系统可接受的。

大数据安全相关技术

- **4. 数据溯源技术**
  - 标记法
  - 反向查询法

大数据安全相关技术

- (1) 标记法：
  - 用标注的方式来记录原始数据的一些重要信息。
  - 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。

大数据安全相关技术

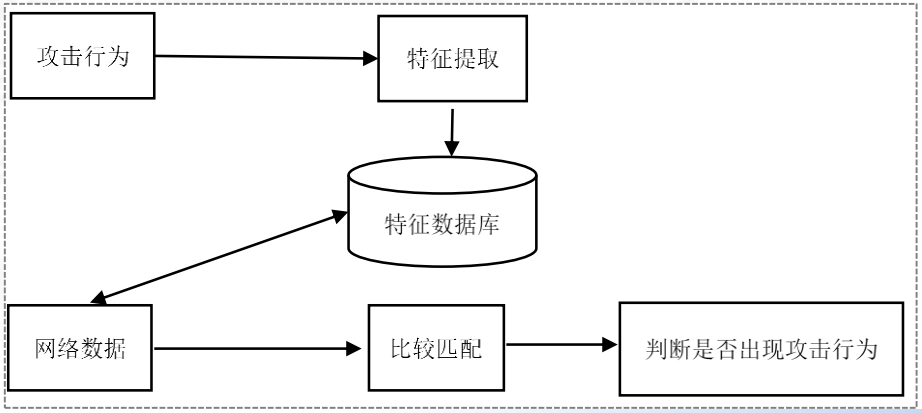
- (2) 反向查询法：
  - 反向查询法通过构造原函数的反函数对查询求逆。
  - 由结果追溯到原数据，更适合于细粒度数据。

大数据安全相关技术

- 5. 大数据安全审计技术
  - 基于规则的安全审计
  - 基于统计的安全审计
  - 基于机器学习的安全审计

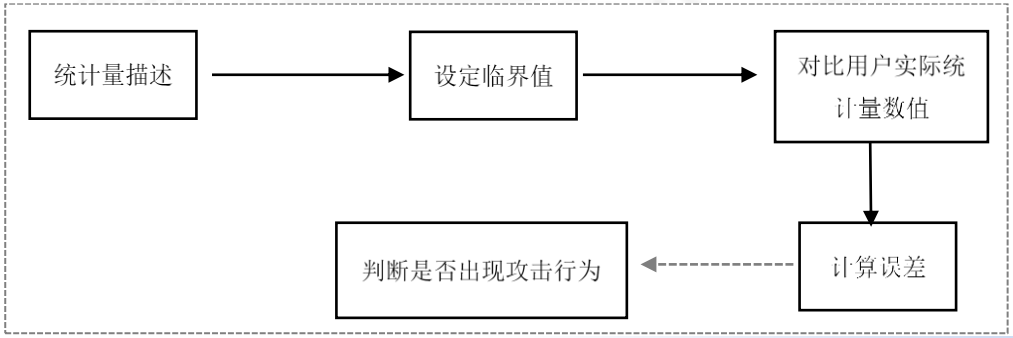
大数据安全相关技术

- (1) 基于规则的安全审计：
  - 提取已知的攻击行为特征，放入特征数据库。
  - 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。
  - 根据匹配状况判断是否出现网络攻击行为，对此采取相应的响应机制。



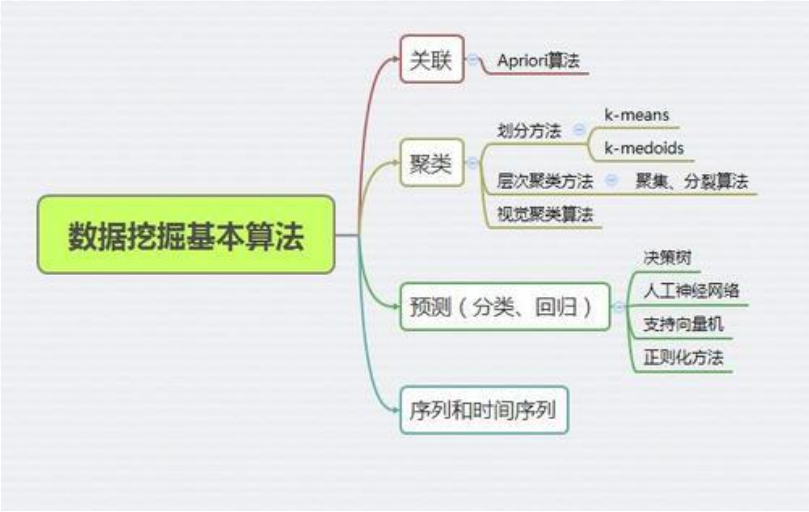
大数据安全相关技术

- (2) 基于统计的安全审计：
  - 统计正常情况下统计量描述，如方差、平均值。
  - 审计人员根据经验设定临界值。
  - 根据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

- (3) 基于机器学习安全审计：
  - 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。
  - 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：指某组织对特定对象展开的持续有效的攻击活动。

- APT攻击的检测难度表现在以下方面：
  - 先进的攻击方法
  - 持续性攻击与隐藏
  - 长期驻留目标系统

大数据安全相关技术

- 6. APT攻击检测技术
  - 网络流量异常检测
  - 恶意代码异常检测
  - 社交网络安全事件挖掘

小作业

简答题:

简析大数据生命周期中存在哪些安全风险，我们可以使用哪些技术手段实现安全目标?

提交方式:

课程平台，**word**格式

提交时间:

本周日（**6.19**）