大数据安全
Big Data Security

山东大学软件学院

第一章 绪论

1.1 大数据概述

1.2 大数据安全与隐私保护需求

1.3 大数据生命周期安全风险分析

1.4 大数据安全与隐私保护技术框架

1.5 大数据服务于信息安全

1.6 基本密码学工具

1.1 大数据概述

1.1.1 大数据来源

1.1.2 大数据应用

1.1.3 大数据技术框架

1.1.1 大数据来源

根据来源对象的不同，可以将其分为源自**人、机、物**等几类的大数据。若根据应用领域划分，则典型的大数据来源包括：

- 互联网大数据
- 物联网大数据
- 生物医疗大数据
- 电信大数据
- 金融大数据
- 智慧城市大数据
- 交通大数据
- 科学研究大数据
- ……

1.1.2 大数据应用

大数据被比喻为待开采的“金矿”，其用途是多样化的。目前大数据技术已经被广泛应用于**电子商务、金融、智能医疗、智能交通**等领域，各种新型应用模式层出不穷：

- **互联网大数据分析**

  电子商务平台通过对用户网络购物数据的分析，构建用户画像，可以更准确地掌握用户购物倾向，向其推荐可能感兴趣的产品，实现精准营销

- **交通大数据分析**

  交通管理部门可以对数据按时间切片分析，构建实时热点分布图，进行景区热力预警分析

- **医疗健康大数据分析**

  通过对大量电子病历的学习，医学研究机构可以更清晰地发现疾病演变规律，并作出更科学、准确的诊断

1.1.2 用户画像

完整还原用户全貌，有效捕捉活动场景

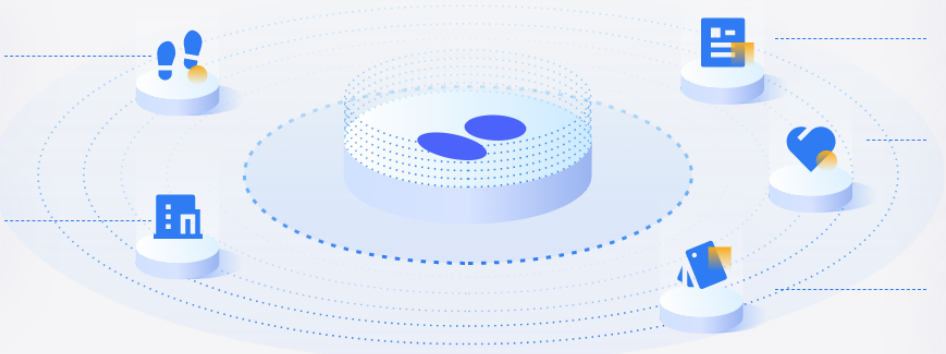
对用户线上和线下行为偏好深度洞察，构建全面、精准、多维的用户画像体系，为APP提供丰富的用户画像数据以及实时的场景识别能力，帮助APP全方位了解用户

行为标签

近期活跃的应用，近期活动的场景

场景标签

机场、商圈、电影院景区、自定义场景等



属性标签

性别、年龄层次、消费水平、职业等

兴趣标签

购物、教育、影音、游戏、金融理财等

定制化标签

针对不同行业可定制化行业特色标签

广告追踪 >

私域流量池 >

营销自动化 >

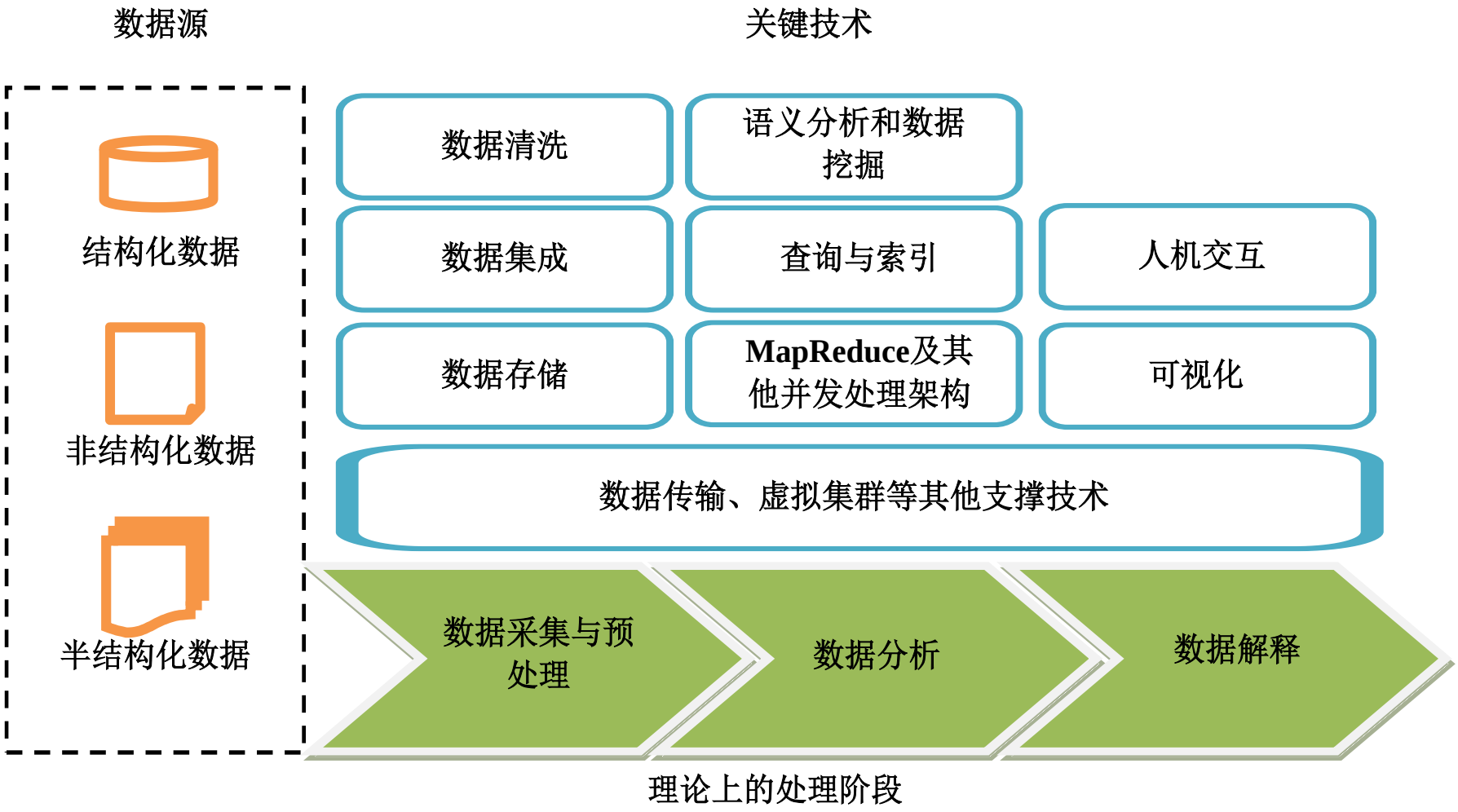
全域SCRM >

客户画像 >

数据报表 >

生命周期管理 >

1.1.3 大数据技术框架


数据源
关键技术
结构化数据
非结构化数据
半结构化数据
数据清洗
语义分析和数据挖掘
数据集成
查询与索引
人机交互
数据存储
MapReduce及其他并发处理架构
可视化
数据传输、虚拟集群等其他支撑技术
数据采集与预处理
数据分析
数据解释
理论上的处理阶段

1.1.3 大数据技术框架

（1）数据采集与预处理：数据采集与预处理（Data Acquisition &Preparation）是大数据应用的基础，进行预处理、数据清洗等。

图　数据清理原理

（2）数据分析：数据分析（Data Analytics）是大数据应用的核心流程。根据不同层次大致可分为三类：**计算架构、查询与索引，以及数据分析和处理**。

（3）数据解释：数据解释（Data Interpretation）旨在更好地支持用户对数据分析结果的使用，涉及的主要技术有**可视化技术和人机交互技术**。

（4）数据传输、虚拟集群等其他支撑技术：为大数据处理提供技术支撑。

1.2 大数据安全与隐私保护需求

1.2.1 大数据安全

1.2.2 大数据隐私保护

1.2.3 区别与联系

大数据安全与隐私保护需求

1.2.1 大数据安全

大数据普遍存在巨大的数据安全需求。在大数据场景带来如下各项新技术挑战：

（1）**如何在满足可用性的前提下实现大数据机密性和完整性**

在大数据场景下，数据的高速流动特性以及操作多样性使得两者之间的矛盾更加突出。

（2）**如何实现大数据的安全共享**

在大数据访问控制中，用户难以信赖服务商正确实施**访问控制**策略，且在大数据应用中实现用户角色与权限划分更为困难。

（3）**如何实现大数据真实性验证与可信溯源**

当一定数量的虚假信息混杂在真实信息之中时，往往容易导致人们误判。最终影响数据分析结果的准确性。需要基于数据的来源真实性、传播途径、加工处理过程等，了解各项数据可信度，防止分析得出无意义或者错误的结果。

大数据安全与隐私保护需求

1.2.2 大数据隐私保护

大数据普遍还存在隐私保护需求。大量事实表明，未能妥善处理会对用户的隐私造成极大的侵害。

（1）由于去匿名化技术的发展，实现身份匿名越来越困难

仅数据发布时做简单的去标识处理已经无法保证用户隐私安全，通过链接不同数据源的信息，攻击者可能发起**身份重识别攻击（re-identification attack）**，逆向分析出匿名用户的真实身份，导致用户的身份隐私泄露。

（2）基于大数据对人们状态和行为的预测带来隐私泄露威胁

随着深度学习等人工智能技术快速发展，通过对用户行为建模与分析，个人行为规律可以被更为准确的预测与识别，刻意隐藏的敏感属性可以被推测出来。

总体而言，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。

大数据安全与隐私保护需求

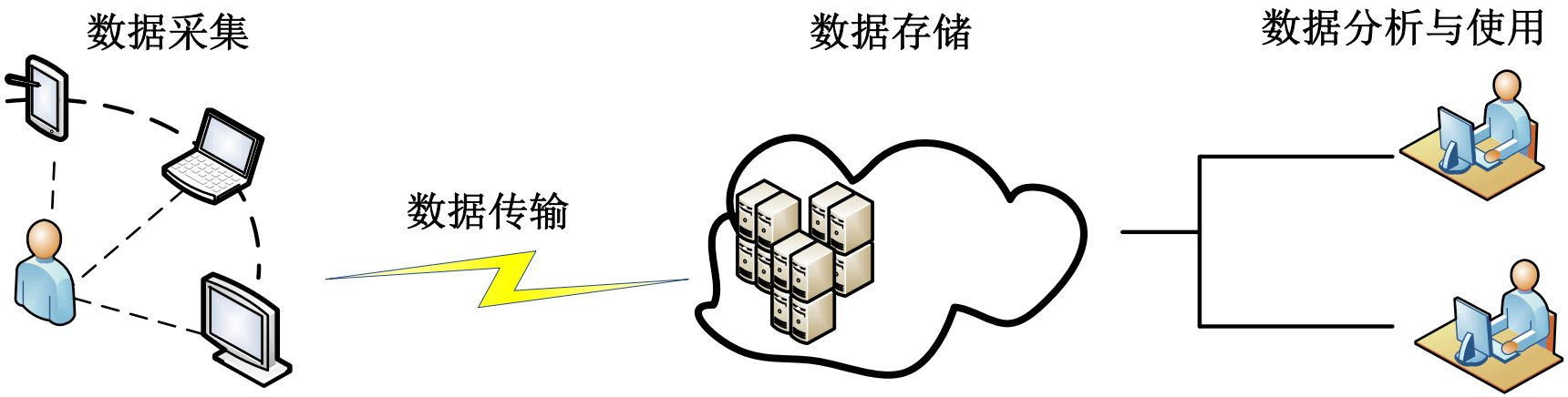
1.2.3 区别与联系

（1）大数据安全需求更为广泛，关注的目标不仅包括**数据机密性，还包括数据完整性、真实性、抗抵赖性/不可否认性**，以及平台安全、数据权属判定等。而**隐私保护需求一般仅聚焦于匿名性**。

（2）虽然隐私保护中的**数据匿名需求**与安全需求之一的**机密性需求**看上去比较类似，但后者**显然严格得多，要求非授权用户完全不可访问**。

（3）在**大数据安全**问题下，一般来说数据对象**是有明确定义**。而在涉及隐私保护需求时，所指的**用户“隐私”则较为笼统**，可能具有多种数据形态存在。

1.3 大数据全生命周期安全风险分析



- 大数据的生命周期包括**数据产生、采集、传输、存储、使用、分享、销毁等诸多环节**，每个环节都面临不同的安全威胁。
- 安全问题较为突出的是**数据采集、数据传输、数据存储、数据分析与使用四个阶段**。

1.3.1 数据采集阶段

- **数据采集：**是指采集方对于**用户终端、智能设备、传感器**等产生的数据**进行记录与预处理**的过程。在大多数应用中数据不需要预处理直接上传，而在某些特殊场景下，例如传输带宽存在限制、或采集数据精度存在约束时，数据采集方需要先**进行数据压缩、变换甚至加噪处理等步骤**，以降低数据大小或精度。一旦真实数据被采集，则用户隐私保护完全脱离用户自身控制。
- **数据采集是数据安全与隐私保护的第一道屏障**，可根据场景需求选择安全多方计算等密码学方法，或选择本地差分隐私（LDP）等隐私保护技术。

1.3.2 数据传输阶段

- **数据传输**：是指将采集到的大数据由用户端、智能设备、传感器等终端传送到大型集中式数据中心的过程。
- 数据传输阶段中的主要安全目标是**数据安全性**。为了保证数据在传输过程中内容不被恶意攻击者收集或破坏，有必要采取安全措施保证数据的**机密性和完整性**。
- 现有的密码技术已经能够提供成熟的解决方案，例如目前普遍使用的**SSL通讯加密协议、或采用专用加密机、VPN技术等**。

1.3.3 数据存储阶段

- 大数据被采集后常汇集**存储于大型数据中心**，而大量集中存储的有价值数据无疑容易成为高水平黑客团体的攻击目标。
- 大数据存储面临的安全风险是多方面的，不仅包括来自**外部黑客的攻击**、来自**内部人员的信息窃取**，还包括不同利益方**对数据的超权限使用**等。
- 因此，该阶段集中体现了**数据安全、平台安全、用户隐私保护**等多种安全需求。

1.3.4 数据分析与使用阶段

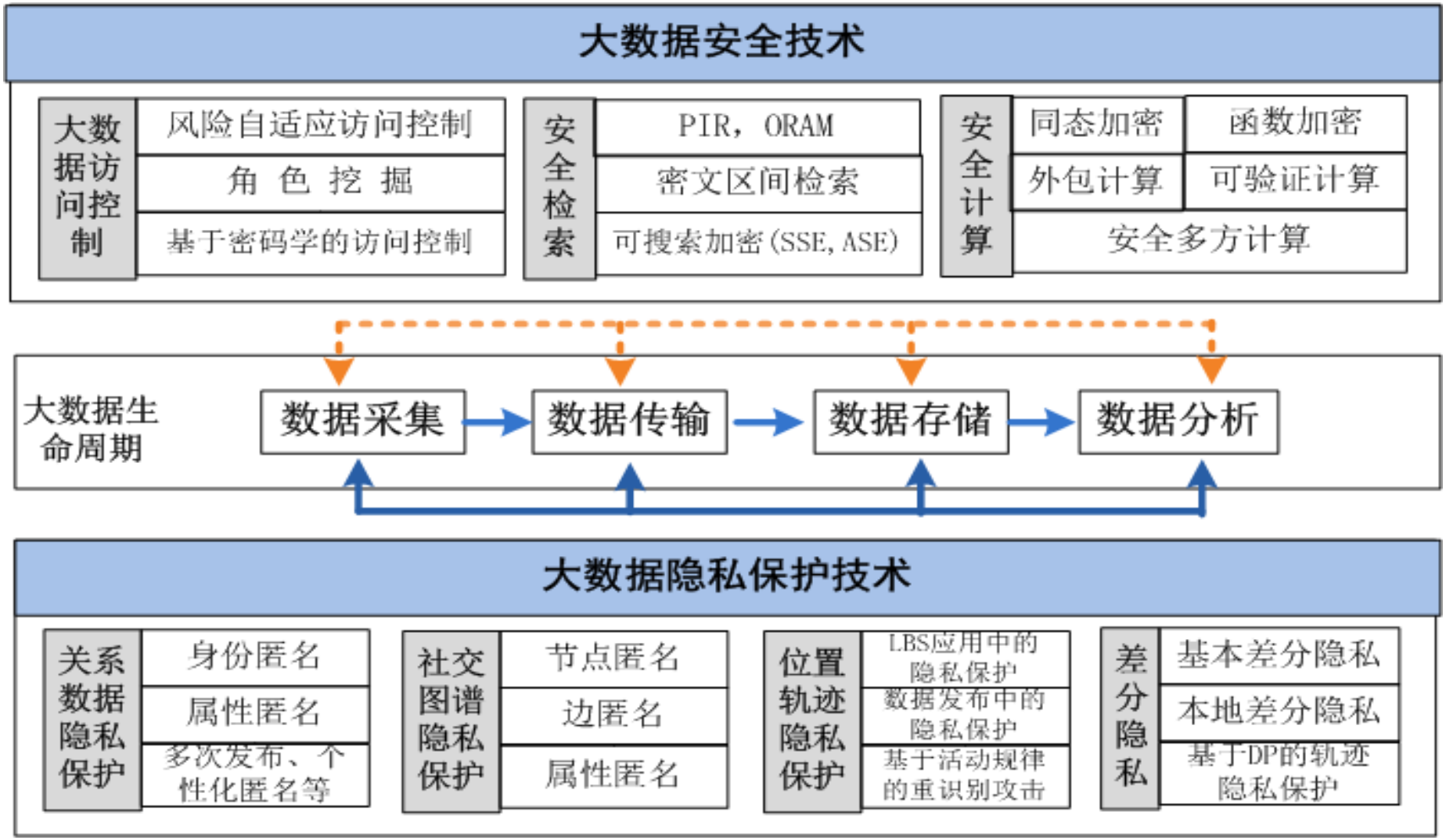
- 大数据采集、传输、存储的**主要目的是为了分析与使用**，通过**数据挖掘、机器学习等算法处理**，从而**提取出所需的知识**。
- 本阶段焦点在于**如何实现数据挖掘中的隐私保护**，降低多源异构数据集成中的隐私泄露。防止数据使用者对用户数据挖掘，得出用户刻意隐藏的知识；防止分析者在进行统计分析时，得到具体用户的隐私信息。

1.4 大数据安全与隐私保护技术框架

1.4.1 大数据安全技术

1.4.2 大数据隐私保护技术

1.4 大数据安全与隐私保护技术框架



大数据安全与隐私保护技术框架

1.4.1 大数据安全技术

(1) 大数据访问控制

- 角色挖掘
- 风险自适应访问控制
- 基于密码学的访问控制

(2) 安全检索

- PIR系列与Oblivious RAM
- 对称可搜索加密
- 非对称可搜索加密
- 密文区间检索

(3) 安全计算

- 同态加密
- 可验证计算
- 安全多方计算
- 函数加密
- 外包计算

大数据安全与隐私保护技术框架

1.4.2 大数据隐私保护技术

(1) 关系型数据隐私保护

- 身份匿名
- 属性匿名
- 多次发布模型与个性化匿名

(2) 社交图谱数据隐私保护

- 节点匿名
- 边匿名
- 属性匿名

(3) 位置与轨迹数据隐私保护

- 面向LBS应用的隐私保护
- 面向数据发布的隐私保护
- 基于用户活动规律的攻击分析

(4) 差分隐私

- 基本差分隐私
- 本地差分隐私（Local Differential Privacy，LDP）
- 基于差分隐私的轨迹隐私保护

1.5 大数据服务于信息安全

1.5.1 基于大数据的威胁发现技术

1.5.2 基于大数据的认证技术

1.5.3 基于大数据的数据真实性分析

1.5.4 大数据与“安全即服务”

大数据服务于信息安全

1.5.1 基于大数据的威胁发现技术

由于大数据分析技术的出现，企业可以超越**以往的“保护-检测-响应-恢复”（PDRR）模式**，更**主动地发现潜在的安全威胁**。相比于传统技术方案，基于大数据的威胁发现技术具有如下优点：

□ 分析内容的范围更大

在威胁检测方面引入大数据分析技术，可更全面地发现针对数据资产、软件资产、实物资产、人员资产、服务资产和其他为业务提供支持的无形资产等信息资产的攻击

□ 分析内容的时间跨度更长

引入大数据分析技术后，威胁分析窗口可以横跨若干年的数据，因此，威胁发现能力更强，可有效应对APT类攻击

□ 攻击威胁的预测性

基于大数据的威胁分析，可进行超前的预判。它能够寻找潜在的安全威胁，对未发生的攻击行为进行预防

□ 未知威胁检测

大数据分析的特点是侧重于普通的关联分析，而不侧重因果分析，因此，通过采用恰当的分析模型，可发现未知威胁

1.5.2 基于大数据的认证技术

认证技术中引入大数据分析具有如下优点：

- **攻击者很难模拟用户行为特征来通过认证，安全性更高**

  攻击者很难在方方面面都模仿到用户行为

- **减小了用户负担**

  用户行为和设备行为特征数据的采集、存储和分析都由认证系统完成

- **可更好地支持各系统认证机制的统一**

  基于大数据的认证技术可以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

虽然基于大数据的认证技术具有上述优点，但同时也存在一些问题和挑战亟待解决，例如：1）初始阶段的认证问题。以及2）用户隐私问题。

1.5.3 基于大数据的数据真实性分析

- 目前，基于大数据的**数据真实性分析**被广泛认为是最为有效的方法。许多企业将其应用于**过滤垃圾邮件**、**识别虚假评论**等。
- 基于大数据的数据真实性分析技术能够**提高垃圾信息的鉴别能力**。表现在以下两个方面：
  - 引入大数据分析可获得更高的识别准确率
  - 在进行大数据分析时，通过机器学习技术，可发现更多具有新特征的垃圾信息

1.5.4 大数据与“安全-即-服务”

- 未来必将涌现出更多、更丰富的安全应用和安全服务，大数据也必将充分展现“安全-即-服务（Security-as-a-Service)”的理念。
- 对于大多数信息安全企业来说，可以通过某种方式获得大数据服务，再结合自己的技术特色领域，对外提供专业化安全服务。
- 一种未来的发展前景是，以底层大数据服务为基础，各个企业之间组成相互依赖、互相支撑的信息安全服务体系，总体上形成信息安全产业界的良好生态环境。

1.6 基本密码学工具

（1）**加密技术**

- 对称加密技术
- 公钥加密技术

（2）**数字签名技术**

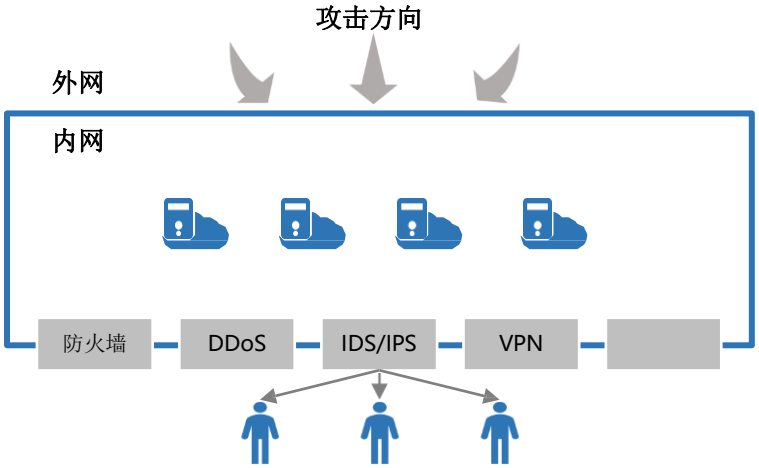
（3）**杂凑函数/散列函数Hash和消息鉴别码MAC技术**

（4）**密钥交换技术、身份鉴别和认证**

零信任安全

传统安全

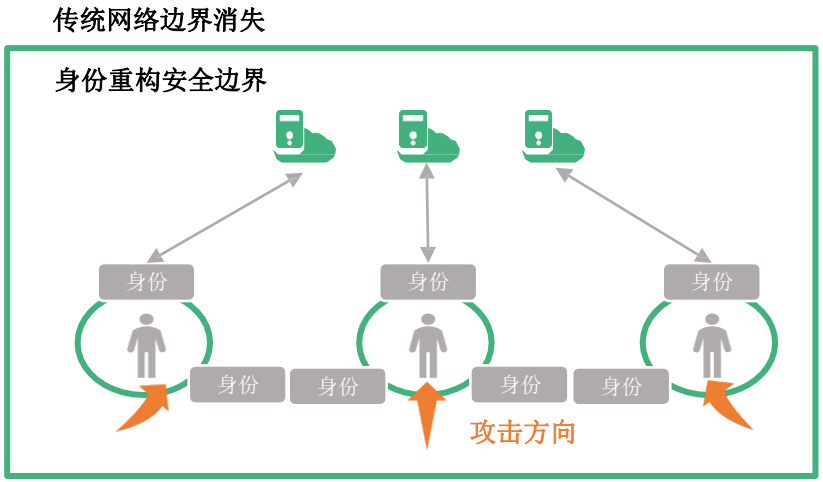
- 背景：数据流向简单，保护核心网络
- 方式：通过防火墙/DDoS/IDS/IPS边界保护



传统安全模型认为内网环境是可信任的，一旦攻破网络边界，处于高度威胁之中

业务安全

- 背景：数据多维流动，“人”是网络核心价值点
- 方式：传统边界消失，“零信任模型”



任何用户都不被信任，在“零信任模型”下， 任何时刻任何环节，设备、身份及权限都有必要被验证。

智能行为认证(IPA)

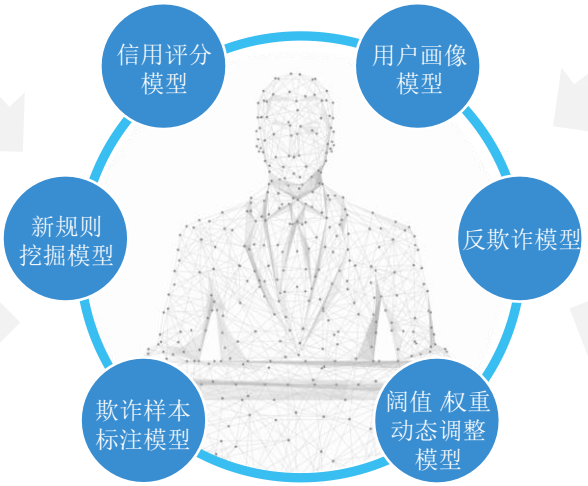
- 基千海量多涌异构行为数据和和亿万量级精准欺诈数据，结合机器学习算法，形成了一整套由上干组规则因素、无监督学习引擎组成的智能实时身份反欺诈系统，在零信任模型下，不预设身份认证方式，根据实时安全环境确定认证方式，实现正常用户的无感知身份认证和欺诈用户精准识别

反欺诈平台：灵活、高效的引擎及管理

- 高性能规则引擎
- 分布式实时引擎
- 可视化管理能力
- 欺诈图谱检测

多算法支持：与数据特点结合的机器学习算法

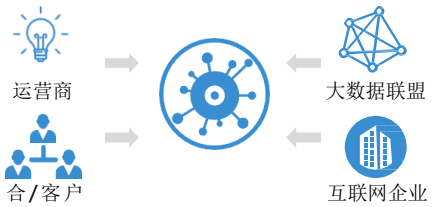
- 有监督学习
- 半监督学习
- 无监督学习
- 迁移学习



多模态引擎：多引擎应对不同阶段、目标

- 机器学习引擎
- 数据探索引擎
- 图计算引擎
- 流式特征引擎
- 离线计算引擎
- 可视化建模引擎
- 深度学习引擎

海量数据：结合内外部数据进行辅助决策



针对不同情景采用不同规则、算法，实现无感知认证和反欺诈