TDM TRUSTED INDUSTRIAL DATA MATRIX CONSORTIUM 可信工业数据空间生态链

工业互联网产业联盟 Alliance of Industrial Internet

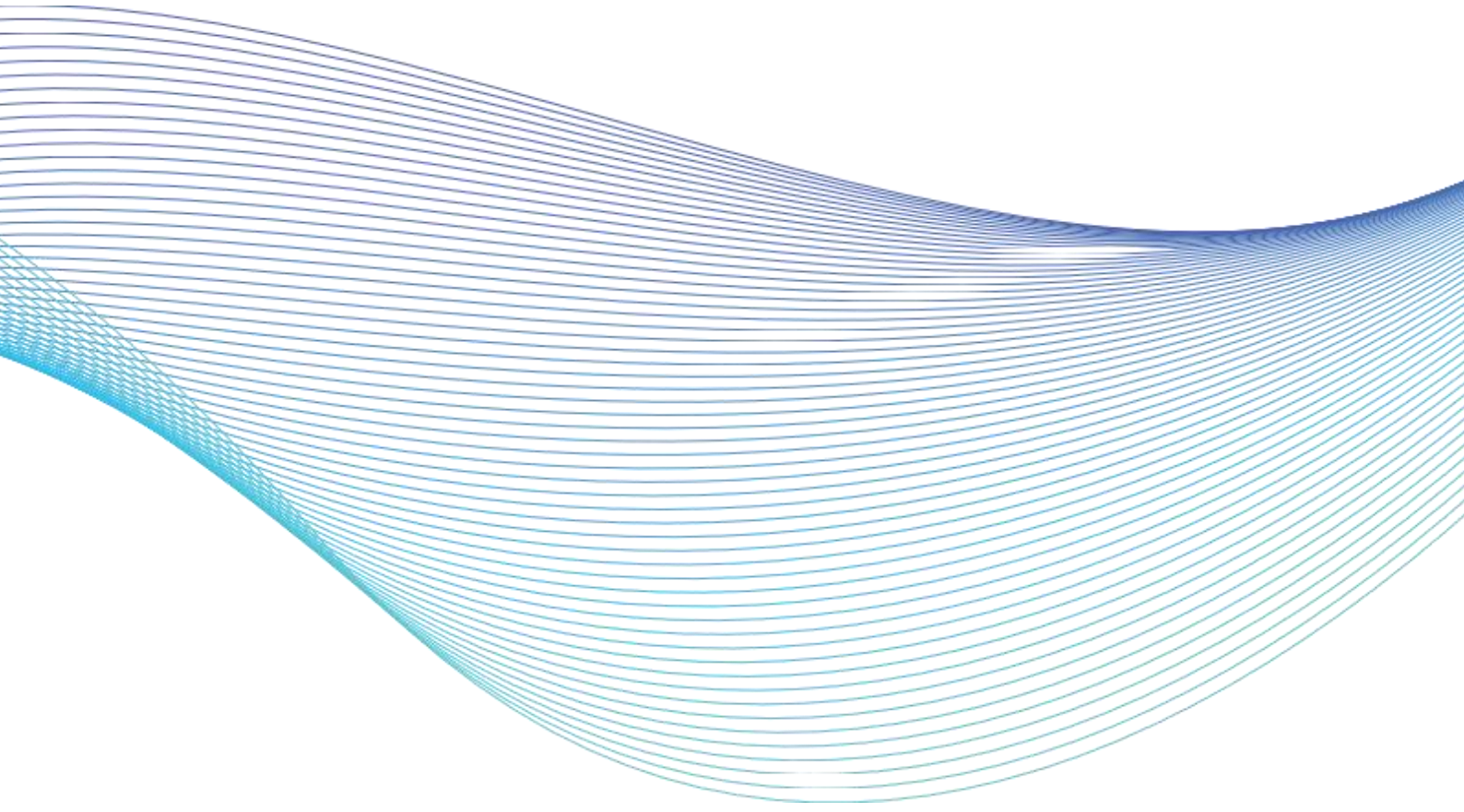
Trusted Industrial Data Matrix | 2022

可信工业数据流通关键技术研究报告

声明

Statement

本报告所载的材料和信息，包括但不限于文本、图片、数据、观点 、建议，不构成法律建议，也不应替代律师意见。本报告所有材料或内容的知识产权归工业互联网产业 联盟所有（注明是引自其他方的内容除外），并受法律保护。如需转载，需联系本联盟并获得授权许可。未经授权许可，任何人不得将报告的全部或部分内容以发布、转载、汇编、转让、出售等方式使用，不得将报告的全部或部分内容通过网络方式传播，不得在任何公开场合使用报告内相关描述及相关数据图表。违反上述声明者，本联盟将追究其相关法律责任。



牵头编写单位

可信工业数据空间生态链

中国信息通信研究院

工业互联网产业联盟

参与编写单位

中国科学院信息工程研究所

北京交通大学

南京理工大学

北京航空航天大学

天津大学

中国电信集团有限公司

华为技术有限公司

中控集团

信联科技(南京)有限公司

深圳数鑫科技有限公司

华控清交信息科技(北京)有限公司

广东一知安全科技有限公司

广域铭岛数字科技有限公司

中企云链(北京)金融信息服务有限公司

北京冲量在线科技有限公司

北京航天云路有限公司

前言

随着新一代信息技术与制造业的深度融合发展，全球工业数据应用已经进入纵深发展的新阶段，数据作为新型生产要素和重要战略资源，正在制造业数字化转型过程中发挥出更大的作用。在这一进程中，工业数据的流通共享受到广泛关注。顺应新发展形势，我国积极营造多方主体参与的数据共享流通生态，国务院先后发布《关于构建更加完善的要素市场化配置体制机制的意见》《要素市场化配置综合改革试点总体方案》《关于构建数据基础制度更好发挥数据要素作用的意见》，明确提出在确保数据安全的前提下，分级分类、分步有序推动部分领域数据流通以及合规使用。在此背景下，中国信息通信研究院提出可信数据空间的概念，并将其作为实现工业数据开放共享和可信流通的新型基础设施，发挥数据要素禀赋。本技术报告将会阐明满足可信工业数据空间功能要求的系列技术，为业界共同建设可信工业数据空间提供技术参考。

目 录

第一章
Chapter 1

背景介绍

（一）可信工业数据空间的概念

可信工业数据空间是在现有信息网络上搭建数据集聚、共享、流通和应用的分布式关键数据基础设施，通过体系化的技术安排确保数据流通协议的确认、履行和维护，解决数据要素提供方、使用方、服务方等主体间的安全与信任问题，进而实现数据驱动的数字化转型。

（二）工业数据流通场景

工业数据流通的场景，有众多的场景因子。从目前众多工业企业的实践来看，通常包括如下场景因子：数据的类型、数据的采集方式、数据的交换模式、数据的使用主体、数据处理的软件能力类型以及数据流通基础设施环境等。以上因子的组合，在企业内、企业间、生态系统内等场景形成数据价值闭环。

1. 复杂组织内部高密高价值数据流通

复杂组织内部通常分工侧重点明显，不同部门之间对应不同的业务流程和业务作业领域，如研发部门负责产品研究及设计、采购部门负责各个部门的部件采购、营销对口CRM等，不同业务部门所管理业务产生的核心数据构成了企业核心竞争力，业务域之间的有效协同通常也会涉及高密高价值数据，依赖数据可信交换流通。通常是基于结构化数据集，通过系统集成/人工上传方式，使用通用的数据加工软件进行简单的分析计算提取有价值信息。

2. 跨组织支持业务协同的数据流通

从企业内扩展到企业间，同样存在高密高价值的数据交换，通常是通过数据的价值再造(探索新的商业模式、业务创新、产品改进等)，由消费方主导，提供方协同，双方是协作关系，如采购与供应商之间的报价信息交换、企业间的技术合作材料交换、审计领域的审计原始材料管控等。跨组织高密数据传输，通常包含了一些专业数据处理软件或者AI提取技术以获取有价值信息。

3. 大型组织主导的生态圈内数据流通

大型企业或者产业领头者越来越多的将平台模式及其周围的生态系统构建作为战略重心。在此场景下，生态链链主与成员之间，既有生态圈内业务协作场景下数据流通，又有对各生态成员之间的数据泄漏担忧。因此，也亟需一套可控数据交换系统来支持生态的安全和持续发展。生态内的场景差异更多的体现在数据的交换模式和基础设施环境。

4. 基于可信服务方的数据增值服务

随着互联网、IOT等产业的兴起，大量企业拥有某产业独特的海量数据但自身缺乏强大的算力，也不具备数据价值挖掘和增值的解决方案设计能力。同时，也有一些深耕该产业的产品和服务提供商，虽拥有强大算力和解决方案，但缺乏真实的现网数据进行算法和解决方案验证。双方无法形成有效握手也是因为数据拥有者缺乏有效的数据安全流通方案，数据处理者通常也缺乏具有公信力的可信流通平台。产业数据流通中场景因子更多是通过传感器采集大量的物联网数据，使用强大的软件平台和专业设备等进行计算和提取。

（三）可信工业数据空间的作用

1. 数据共享架构类问题

(1)数据流通共享模型

在传统网络/信息访问控制模型基础上，增加新的控制要素，需要建立新的数据流通共享模型。该模型需要刻画数据流通要素、数据流通控制机制、数据流动控制策略，是可信工业数据流通和信息安全服务的根基。

(2)工业数据流通技术架构

在传统工业互联网单向中心化汇聚的基础上，增加分布式交换模式，需要提出新的工业数据受控流通技术架构。该架构需要支持设备动态扩展、多汇聚中心数据受控交换，是工业互联网扁平化数据安全交换的灵魂。

(3)信息互通协议与标准

在已有单厂商的封闭控制协议基础之上，需要制定新的不同设备类型间不同厂商间的工业设备信息互通协议与标准。该互通协议与标准包括统一数据交换设备接口、数据交换格式、统一工业通信规程，解决数据兼容存在问题，是实现工业数据流通的核心。

2. 数据安全可信类问题

(1)数据可控性保障能力弱

在法律法规等非技术方式的基础上，需要提出有效的数据流动交换控制机制，以此精准控制交换出管理域后数据按照数据拥有者意图流通和使用。该机制包括大规模数据自动标记、数据跨域流动监测、违规流动处置等技术。支撑法律法规等管理手段的落地，解决控制粒度粗、安全隐患大、安全风险高等问题。

(2)数据可信性验证能力弱

在现有基于单认证中心的端对端数据认证的基础上，需要提出有效的数据交叉认证方法，以此验证数据交换出管理域后数据的真实性。该方法包括数据源与传播路径交叉认证、高效可聚合签名与验签、信任动态度量与信任链构建等关键技术，是多源异构数据在流通过程中全生命周期可信的有效途径。

3. 数据隐私保护类问题

(1)隐私信息全生命周期保护

当前工业数据多模海量、泛在交换频繁、体系保护能力差，难以实现全流程可管可控。为了实现工业数据中隐私信息的全生命周期保护，需要从体系化保护角度出发，构建隐私保护技术体系。

(2)隐私信息跨系统迭代延伸控制

当前工业数据频繁跨终端、跨系统、跨生态圈交换流通已成为常态，隐私泄露事件层出不穷。为了解决交换流转过程中的隐私泄露问题，需要构建面向工业数据流通的隐私信息跨系统迭代延伸控制机制。

(3)面向信息流动的隐私保护系统稳定性

当前工业数据的产业应用多样、隐私需求各异，对隐私保护系统的稳定性要求高。为了提高工业数据隐私保护系统的普适性和代码稳定性，需要提出隐私保护的算法设计准则和通用算法框架，支撑隐私信息保护系统的代码稳定性和算法可扩展性。

（四）可信工业数据空间的作用

针对以上问题，可信工业数据空间实现了数据的开放共享和可信流动，数据所有者可对共享数据的使用对象、范围、方式进行控制，全程记录数据使用方对数据的使用、处理过程，实现数据的有效管理。可解决流通不畅、信息泄露、过度利用等风险，极大促进了数据要素资源的网络化共享、集约化整合、协作化开发、高效化利用。

空间为数据提供方提供数据使用对象、范围、方式的控制能力，高效融合工业跨域异构数据，统一工业数据质量管理标准，消除流通顾虑，释放数据供给。空间为数据使用方提供工业数据流通处理的日志存证，提供内外部合规记录，明确数据主权边界，实现工业数据资源有效管理。空间为数据供需双方提供数据要素流通中间服务，便利供需对接，加强隐私保护，促进应用场景创新和数据价值化配置。

（五）可信工业数据空间——系统架构

业务视角

可信工业数据流通系统共有三种不同利益相关方，分别为数据提供方、数据使用方和中间服务方，每个利益相关方在可信工业数据流通系统的两种不同模式中开展不同的活动，如图1、图2所示。

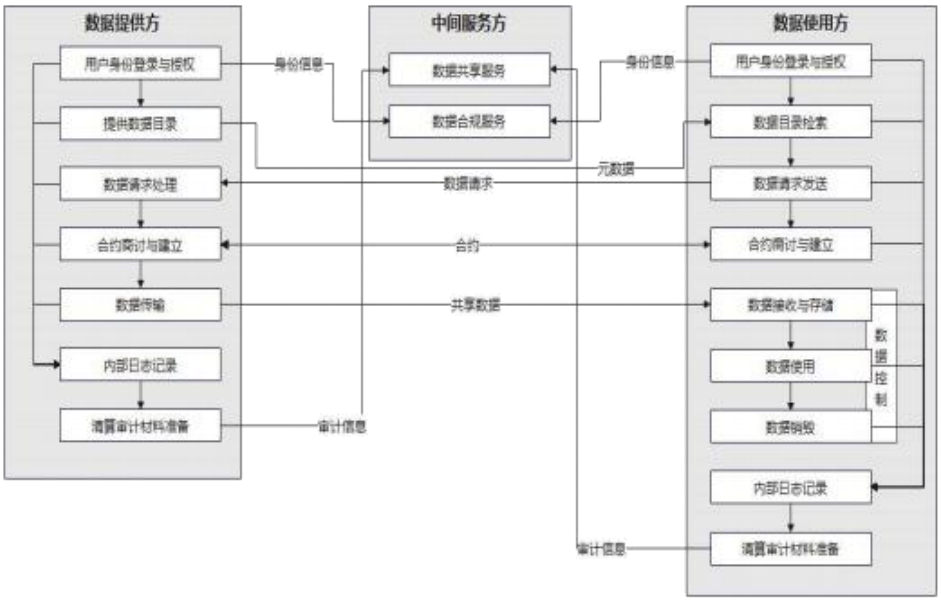


图1 分布式模式下业务视角

分布式模式下，数据提供方通过向中间服务方的数据合规服务方发送身份信息完成用户身份登录与授权，而后提供数据目录，处理数据使用方发送的数据请求，提供数据共享的合约或接受数据使用方发起的合约商讨请求并达成共识建立智能合约，数据提供方将共享数据传输至数据使用方。整个过程中每个活动发生时数据提供方进行内部日志记录，并周期性进行清算审计材料准备，将审计信息提供给中间服务方的共享服务方进行审计。

分布式模式下，数据使用方通过向中间服务方的数据合规服务方发送身份信息完成用户身份登录与授权，而后从数据提供方进行数据目录检索，向数据提供方发送共享数据使用请求。数据使用方可接收数据提供方预设好的共享数据使用合约或发起与数据提供方商讨共享数据使用合约的请求。数据使用方与数据提供方建立智能合约后，数据使用方接收数据提供方发送的共享数据并存储、使用、用后销毁，共享数据在数据使用方接收、存储、使用、销毁的过程中接受数据控制。整个过程中每个活动发生时数据提供方进行内部日志记录，并周期性进行清算审计材料准备，将审计信息提供给中间服务方的共享服务方进行审计。

分布式模式下，中间服务方仅提供基础的数据合规类服务以及数据交易中的审计服务。

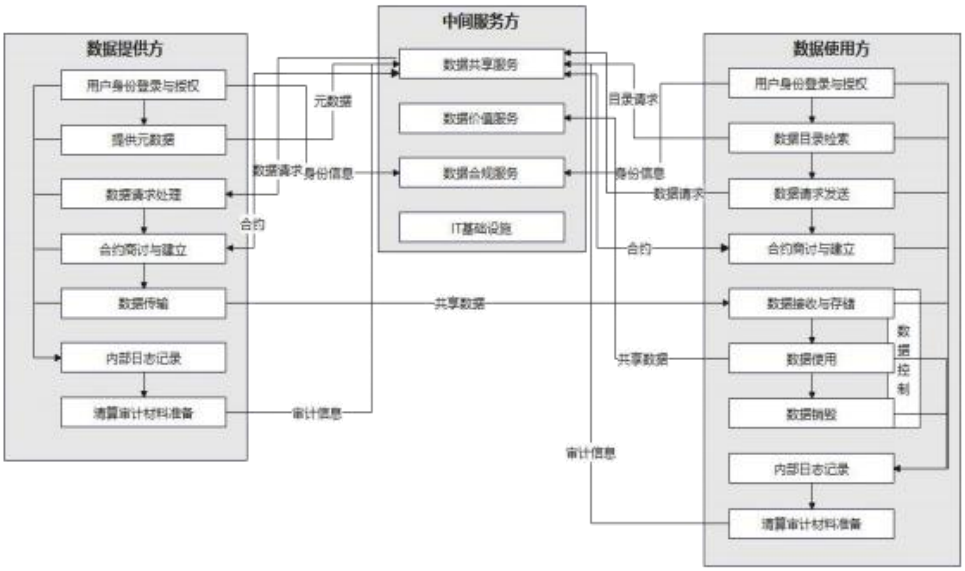


图2　中间服务模式下业务视角

中间服务模式下，数据提供方通过向中间服务方的数据合规服务方发送身份信息完成用户身份登录与授权，而后向中间服务方提供元数据并加入由中间服务方的数据共享服务方提供的数据目录服务。数据提供方从中间服务方获得数据共享请求，处理数据请求，通过中间服务方的数据共享服务提供数据共享合约或接受数据使用方发起的合约商讨请求并达成共识建立智能合约。数据提供方将共享

数据传输至数据使用方。整个过程中每个活动发生时，数据提供方进行内部日志记录，并周期性进行清算审计材料准备，将审计信息提供给中间服务方的共享服务方进行审计。

中间服务模式下，数据使用方通过向中间服务方的数据合规服务方发送身份信息完成用户身份登录与授权，而后从中间服务方的数据共享服务方进行数据目录检索，向中间服务方中的数据共享服务发送共享数据使用请求。数据使用方可通过中间服务方的数据共享服务方接受数据提供方预设好的共享数据使用合约或通过中间服务方的数据共享服务方发起与数据提供方商讨共享数据使用合约的请求。数据使用方与数据提供方建立智能合约后，数据使用方接收数据提供方发送的共享数据并存储、使用、销毁，共享数据在数据使用方接收、存储、使用、销毁的过程中接受数据控制，数据使用方也可通过中间服务方的IT基础设施服务方对数据进行存储，通过中间服务方的数据价值服务方对数据进行使用。整个过程中每个活动发生时数据提供方进行内部日志记录，并周期性进行清算审计材料准备，将审计信息提供给中间服务方的共享服务方进行审计。

中间服务模式下，中间服务方提供数据合规类服务、数据使用。

功能视角

业务视角规定了可信工业数据流通中的各利益相关方、各利益相关方的业务活动和各利益相关方之间的关系。功能视角规定了可信工业数据流通中各利益相关方所使用的系统功能，包括：数据资产控制、数据资产管理、可信环境、可信传输、供需对接、身份认证、日志存证、清算审计与数据增值类服务。

技术视角

技术视角规定了可信工业数据流通中实现各功能所需的关键技术。本报告将从功能视角展开，详细阐述每类功能所需的技术。

第二章
Chapter 2

关键技术

（一）数据资产控制相关技术

1. 数据控制

（1）定义

数据控制是指在数据的传输、存储、使用和销毁环节，通过机器可读的电子合约，实现对数据资产使用的时间、地点、主体、行为和客体等因素的控制。

数据控制是传统访问控制技术、使用控制技术在作用范围、作用周期和控制的细粒度等因素上丰富与革新。

（2）作用机理

数据控制技术的实现，可以拆解为以下三个步骤。

合约建立：数据提供方与数据使用方签订应能够被机器读取并执行的电子合约。一份电子合约由控制要求和控制策略构成。

控制要求描述了数据提供方对数据使用方提出的数据使用方式的限制，控制要求主要分为以下五类。

时间要求：如要求数据使用方10天后销毁数据等。

地点要求：如限定数据使用方访问数据时的IP等。

主体要求：如限定允许使用数据的用户，限定允许访问数据的应用程序等。

行为要求：如数据使用方不允许编辑、转发数据等。

客体要求：如要求数据在存储时保持加密状态。

控制策略是由控制要求所生成的，计算机可执行的代码。当数据提供方和数据使用方协定完成控制要求时，将自动生成计算机可执行的代码，控制策略从技术上保证了电子合约中控制要求可以被准确无误地执行。相较于接近自然语言的控制要求，控制策略更侧重于机器可读可执行，每一条控制策略都由对应控制要求一对一映射而来。

控制要求通常为形式化的自然语言。一条明确的控制要求通常由一个判定条件和该判定条件触发后应执行的操作构成。例如，数据提供方要求数据不能在数据使用方处永久留存，只能留存14天。那么形式化后的控制要求为：该数据存放时间大于14天时，结束已开启的相关进程，销毁全部数据副本。

在用户可读的形式化控制要求(低代码形式可视化编辑)确定后，会自动生成机器可读的对应控制策略(代码)。部分控制要求如表1-1所示。

控制要求通常为形式化的自然语言。一条明确的控制要求通常由一个判定条件和该判定条件触发后应执行的操作构成。例如，数据提供方要求数据不能在数据使用方处永久留存，只能留存14天。那么形式化后的控制要求为：该数据存放时间大于14天时，结束已开启的相关进程，销毁全部数据副本。

在用户可读的形式化控制要求(低代码形式可视化编辑)确定后，会自动生成机器可读的对应控制策略(代码)。部分控制要求如表1-1所示。

动态监控：依据电子合约规定的控制要求，动态的监测和控制数据使用方对数据的使用过程。当数据使用过程与控制要求不符时，将按照合约的规定，执行对应的控制策略，对数据或进程进行相应的处理。

(3) 所解决的问题

数据控制技术主要解决了数据传输、存储、使用和销毁环节中的部分的安全可信问题。着重于实现数据和参与方的可控、可审计。

可控：解决了以往由于数据低成本可复制性，无形体性，难以控制数据资产不受控复制传播的问题。数据控制技术通过技术手段促使数据使用方履行电子合约中规定的内容，限制数据使用方二次转发数据资产给授权外第三方的无意或有意行为，确保数据提供方对外发送的数据资产及其全部副本在完成使用后被彻底销毁。

可审计：解决了以往对数据流通的过程难以监控的问题。电子合约中的控制策略在监测与执行时，会将用户的操作行为、数据的状态等信息作为副产物记录下来，从而形成可供审计的日志。

分类	控制要求	
	判定条件	触发后执行的操作
时间	该数据存放时间t>T时（t是当前存放时间，T是合约规定的存放时间）	结束已开启的相关进程，销毁全部数据副本
	该数据使用时间t>T时（t是当前使用时间，T是合约规定的使用时间）	
	该数据使用次数n>N时（n是当前使用次数，N是合约规定的使用次数）	
	该数据在规定的时间段外被访问时	拒绝该操作
地点	该数据被白名单外的IP/地区访问时	
	该数据被白名单外的设备/mac地址访问时	
	该数据在白名单外的空间内被访问时	
	该数据在安全等级较低的运行环境/网络环境下被访问时	
主体	该数据被白名单外的用户账户/法人实体访问时	
	该数据被白名单外的应用程序/进程访问时	
	该数据被白名单外API或服务调用时	
行为	该数据被复制时	
	该数据被删除时	
	该数据被转发/下载时	
	该数据被另存为新副本时	
	该数据的内容被编辑/加工/修改时	
	该数据被读取/写入/调用的频率f>F时（f是当前频率，F是合约规定的频率）	
	该数据的内容被截屏时	
	该数据的内容被拍照时	识别并留存日志
客体	该数据的新副本产生时（新副本包括由复制产生文件、另存为的新文件、输入算法/模型产生的新数据）	新副本文件继承相同的控制要求和策略
	该数据未被应用程序/进程/API调用时	保持加密状态
	该数据被应用程序/进程/API调用时	进行机密计算
	该数据被数据提供方申请撤回时	结束已开启的相关进程，销毁全部数据副本

表1-1 控制要求示例

(4) 适配性

本小节将从适用的数据类型、软硬件要求、云端兼容性、中心化/去中心化模式等方面分析该技术在可信工业数据空间中的适配性。

适用的数据类型

以结构化 / 半结构化 / 非结构化为分类标准：数据控制技术 a)支持结构化数据，如db等结构化数据库中的数据； b)支持半结构化数据，如json、xml、html等； c)支持非结构化数据，如doc / docx / txt等文档格式、pdf、ppt / pptx、xls / xlsx / csv等表格格式、dwg等图纸格式、jpg / bmp等图片格式、mp3/wav等声音格式以及avi/mp4等视频格式。以文件类 / 流数据分类：数据控制技术支持对文件类和流数据的控制。

软硬件要求

数据控制技术是相对轻量级的技术，其基本功能不需要硬件基础设施的支持，数据控制技术可以以应用程序 / 插件 / 控件的方式安装在数据提供方和数据使用方的设备系统。

部分进阶数据控制策略需搭配可信存储环境以及可信执行环境，如对处于计算中的数据进行控制。

云端兼容性

数据控制技术适用于云端部署的系统。相较于运行环境各异的用户终端，统一的云端环境更易管理与配置。云端在本质上是物理机或物理机的集合，作用于系统层的数据控制技术依然可以生效。

中心化 / 部分去中心化 / 去中心化模式

数据控制技术适用于 a)数据汇集的中心化模式,如数据湖、数据中台、大数据中心； b)数据分布在用户但服务集中的部分去中心化模式，如数据交易所，数据中介商； c)去中心化模式，即没有中间服务方的点对点模式。在各类模式之中，数据控制技术将会作用于数据资产所在位置，即中心化模式的数据中心 / 存储服务器一侧、部分去中心化和去中心化的用户一侧。

(5) 同类技术对比

本小节将把可用于数据控制的主流相关技术进行对比，包括基于角色的访问控制模型(Role-BasedAccessControl,RBAC)、基于属性的访问控制模型(Attribute-BasedAccessControl,ABAC)、基于使用控制的访问控制模型(UsageControl)、基于行为的访问控制模型、面向网络空间的访问控制模型(Cyberspace-OrientedAccessControl，CoAC)。

	基于角色的访问控制模型（RBAC）	基于属性的访问控制模型（ABAC）	基于使用控制的访问控制模型	基于行为的访问控制模型	面向网络空间的访问控制模型（CoAC）
概述	将用户映射到角色，角色拥有相对应的权限，用户通过“用户—角色”委派关系获得访问权限。目前 RBAC 商用广泛，但控制颗粒度较粗。	在模型定义、策略描述以及模型的实施中引入了属性概念，以主客体及其属性、环境属性为媒介，完成对数据访问的授权，及权限统一管理。	引入可变属性(mutable attribute)，将义务、条件和授权作为使用决策的核心，访问控制模型的延伸，额外生效于数据的使用和销毁环节。	梳理角色、时态和环境之间的相互关系，定义了“行为”的概念，将用户行为作为授权的媒介，可解决移动计算的信息系统中的访问控制问题，	综合考虑谁、何时、使用何设备、从何处接入网络、经由何网络、访问何资源、做何处理、数据保存多长时间等关键要素，将传统的信息系统访问控制和网络接入访问控制结合起来，提出了**网络与信息流转关联的访问控制机制**，构建资源/网络传播链来实施网络与信息系统联合的访问控制，并能支持跨系统跨流转管控，解决信息的脱离授权系统后的二次/多次转发的受控访问，支撑全生命周期访问控制
作用范围	用户自己的终端		作用范围延伸至数据使用方的终端（只影响电子合约规定的数据）		
作用周期	数据存储环节，用户访问时		数据存储、使用和销毁环节	数据存储、使用和销毁环节	数据产生、感知、发布、交易、传输、存储、使用、销毁环节
控制的细粒度	粗	中	中	中	细
配置难度	易	配置难度中等但管理维护复杂	中	中	中

表1-2 可用于数据控制技术的对比

2. 隐私计算

(1) 定义

隐私计算是面向隐私信息全生命周期保护的计算理论和方法，是隐私信息的所有权、管理权和使用权分离时隐私度量、隐私泄露代价、隐私保护与隐私分析复杂性的可计算模型与公理化系统。具体是指在处理视频、音频、图像、图形、文字、数值、泛在网络行为信息流等信息时，对所涉及的隐私信息进行描述、度量、评价和融合等操作，形成一套符号化、公式化且具有量化评价标准的隐私计算理论、算法及应用技术，支持多系统融合的隐私信息保护。隐私计算涵盖了信息搜集者、发布者和使用者在信息产生、感知、发布、传播、存储、处理、使用、销毁等全生命周期过程的所有计算操作，并包含支持海量用户、高并发、高效能隐私保护的系统设计理论与架构。隐私计算是泛在互联环境下隐私信息保护的重要理论基础。

(2) 作用机理

隐私计算是面向泛在共享(如图1-1所示)全生命周期保护(如图1-2所示)的理论与方法。隐私计算从“计算”的角度确立隐私信息产生、感知、发布、传播、存储、处理、使用、销毁等全生命周期的隐私计算架构(如图1-3所示)、延伸控制(包括按需脱敏、使用、删除等控制)、形式化描述方法、量化评估标准,以及脱敏算法的数学基础；基于延伸控制思想，抽象全生命周期各个环节对多模态隐私数据的操作，包括：隐私智能感知、分量保护要求量化、跨系统保护要求的量化映射、场景适配的隐私动态度量、按需迭代脱敏、多副本完备删除，以及根据保护效果自动迭代修正脱敏等；基于隐私计算语言支撑跨平台隐私保护的一致性;基于延伸控制和自存证实现泛在随遇的侵权判定，实时发现违规行为并取证溯源；基于隐私计算的算法设计准则和通用算法框架，支撑隐私信息保护系统的代码稳定性和算法可扩展性，并支撑高效能和高并发。

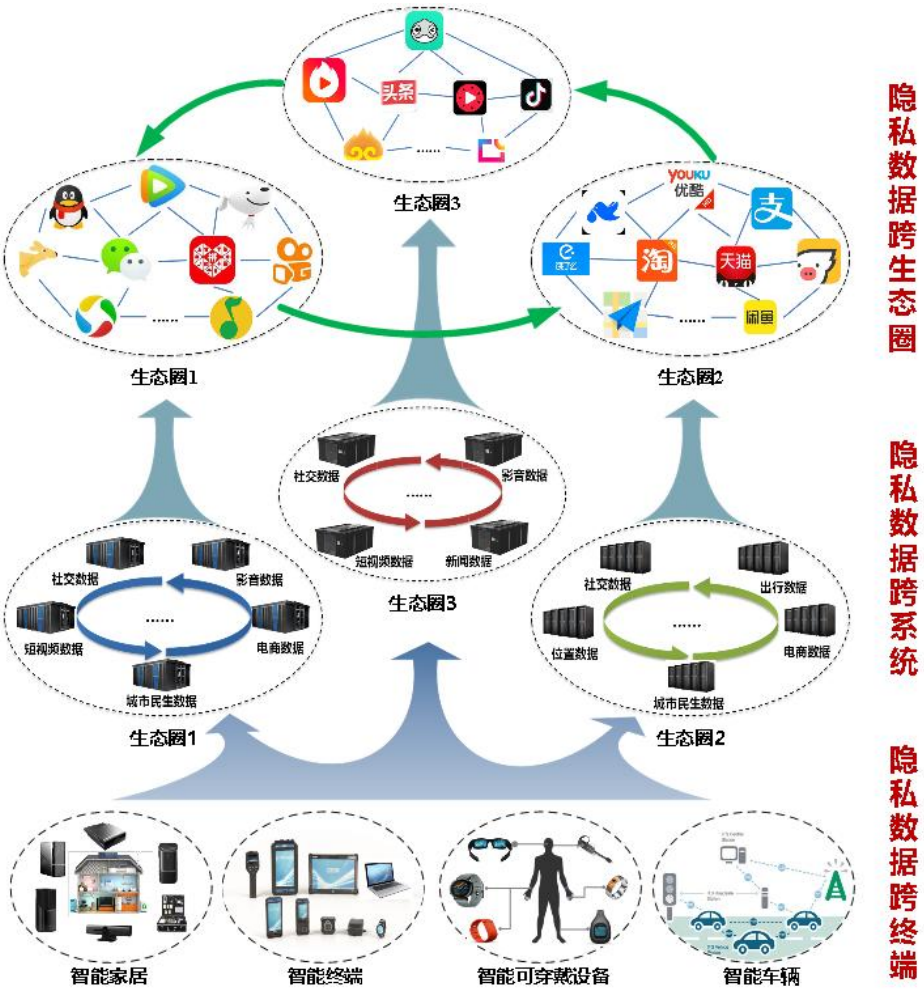


图1-1 隐私信息泛在共享的场景

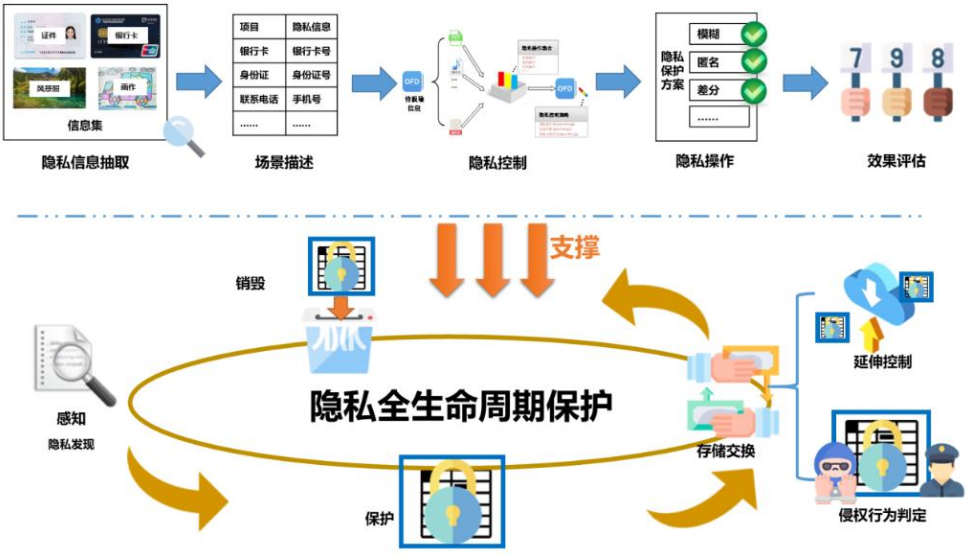


图1-2 隐私信息全生命周期保护示意图

隐私计算架构面向任意格式的明文信息M，具体包括以下5个步骤。

隐私信息抽取：根据明文信息M的格式、语义等，抽取隐私信息并得到隐私信息向量I。

场景抽象：根据I中各隐私信息分量的类型、语义等，对应用场景进行定义与抽象。

隐私控制：选取各隐私信息分量所支持的隐私操作，并生成隐私控制策略。

隐私操作：根据隐私信息的控制策略集合选择合适的隐私操作集合，构建隐私脱敏方案。

隐私保护效果评估：根据相关评价准则，使用基于熵或基于失真的隐私度量来评估所选择的隐私保护方案的隐私保护效果。

基于隐私计算的隐私信息保护系统应包括语义提取、场景抽象、隐私信息变换、隐私信息融合、隐私操作选取、隐私保护方案设计/选取、隐私保护效果评估等环节。

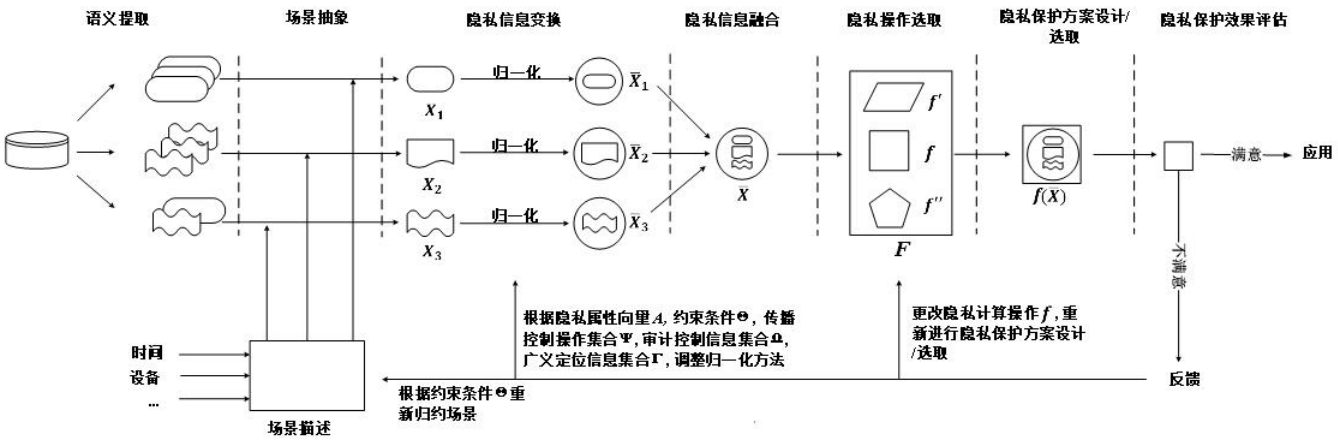


图1-3 基于隐私计算的隐私信息保护系统架构

(3) 在可信工业数据空间中所解决的问题

可信工业数据空间隐私信息的体系化保护：在隐私计算理论指导下，对工业数据空间中的隐私信息、约束条件、流转管控、审计监管、追踪溯源、确权鉴权等进行统一描述，支撑可信工业数据空间中隐私信息全生命周期过程的体系化保护。

隐私信息的迭代延伸控制：根据工业数据应用场景中的控制意图和接收者保护能力等因素生成隐私标签，并与隐私信息进行绑定，在流转过程中，后续节点持续根据下级节点的保护能力调整隐私标签，实现跨系统传播过程中场景适应的差异化访问权限控制和按需迭代脱敏，支撑工业数据泛在共享场景下隐私信息受控共享。

工业数据隐私保护系统的稳定性：根据隐私保护效果与可用性的应用需求，

对工业数据空间中不同场景下差异化算法进行归一化描述，抽象算法通用框架；根据用户的主观需求和客观环境，结合工业数据空间中涉及的用户终端状态，动态选择算法和确定算法参数，从而支撑隐私信息保护系统的代码稳定性和算法可扩展性，并支撑高效能和高并发。

隐私侵权行为溯源取证方法：对工业数据空间中的隐私信息、溯源记录信息和侵权行为判定结果，以及隐私信息全生命周期各环节的操作行为进行自存证，结合保护约束条件、传播控制，判定泛在共享环境下隐私操作行为的合规性，支撑工业数据全生命周期的隐私信息有序合规利用。

(4) 在可信工业数据空间中的适配性

隐私计算作为普适性理论与关键技术体系，适应各种软硬件要求、云/端部署、去中心化、数据类型等，对软硬件、网络架构、数据类型等无特殊要求。

(5) 与同类技术的关系

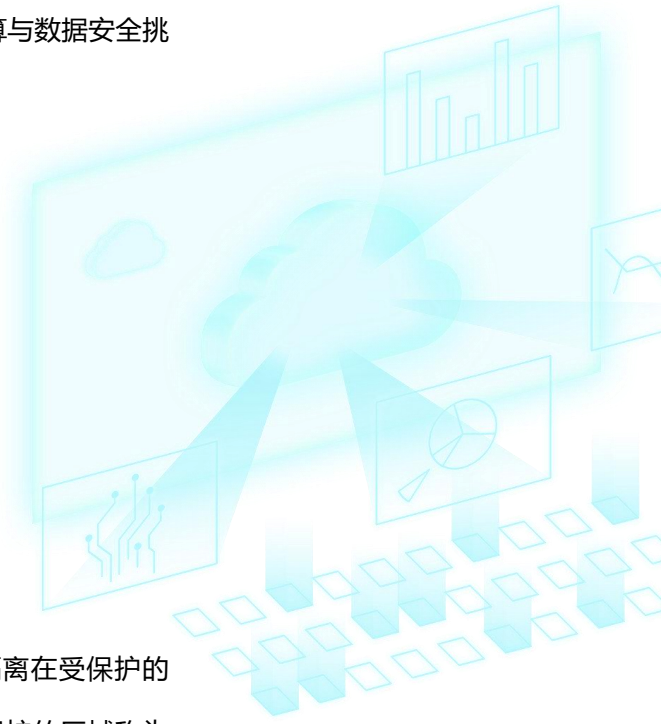
按照泛在共享环境下隐私全生命周期保护的需要，隐私计算与其他隐私保护技术进行综合比较如表1-3所示。

保护技术对比项	隐私计算	传统方法(k-匿名、差分等)
计算(使用)	√	√
计算开销	低	中
通信开销	低	低
需要专用硬件	×	×
交换	√	√
隐私防护	√	×
脱敏	√	√
迭代脱敏	√	×
有界系统(不出域)	√	√
无界系统(出域)	√	√
全生命周期	√	×
延伸控制	√	×
差异保护	√	×
量化与映射	√	×
评估	√	×
取证	√	×
用户数	海量	海量
计算粒度	对象级	单等级

表1-3 隐私计算与其他隐私保护方案的对比

(6) 成熟度分析

2015年，中国科学院信息工程研究所李凤华、西安电子科技大学李晖等学者在国际上率先提出并首次精准定义了隐私计算(PrivacyComputing)的概念、定义和研究范畴。2021年出版了首部隐私计算学术专著《隐私计算理论与技术》。相关学者组织了七届隐私计算国际学术研讨会(2015-2021)；中国中文信息学会在2018年成立了“大数据安全与隐私计算”专委会，举办了五届“大数据安全与隐私计算”学术会议(2018-2022)，并举办了首届“2022隐私计算与数据安全挑战赛”，持续推动隐私计算在国际和国内的学术研究与交流。



3. 机密计算

本小节侧重于描述利用TEE等技术进行可信计算的过程。

(1) 定义

机密计算面向云计算应用，在数据处理过程中将敏感数据隔离在受保护的CPU区域中，该区域称为飞地(enclave),当前更为通常的是将受保护的区域称为可信执行环境TEE(TrustedExecutiveEnvironment)。在TEE中处理的数据以及处理方法只有授权的代码才能访问，包括云服务提供商在内的任何其他程序、设备或者人都无法知道。

(2)作用机理

随着企业越来越多的使用公共云和混合云服务，云中的数据安全成为最受关注的问题。机密计算的主要目标是向云计算的使用者提供更大的数据安全保证，确保数据所有者在云中的数据受到保护和保密。与存储加密和传输加密不同的是，机密计算通过保护正在处理或运行时的数据，消除了许多系统层面的数据安全漏洞。目前TEE的主要实现技术包括IntelSGX和ARM的TrustZone技术。

IntelSGX

Intel公司发布了基于其公司处理器架构的可信执行环境IntelSGX(如图2-11)，是一组增强应用程序代码和数据安全性的指令，为它们提供更强的保护以防泄漏或修改。SGX将应用程序分为了可信区域和非可信区域，其中可信区域被称为enclave。调用可信区域中的程序时，需要定义ecall借口，声明传递的数据的结构和大小。英特尔提供了包括本地证明、远程证明。数据密封等多个基础组件，并提供了丰富的软件开发包供开发者使用。

SGX允许用户态及内核态代码定义将特定内存区域，设置为私有区域，此区域也被称为飞地(Enclave)。其内容受到保护，不能被本身以外的任何进程存取，包括高权限级别运行的进程(例如操作系统内核进程)。

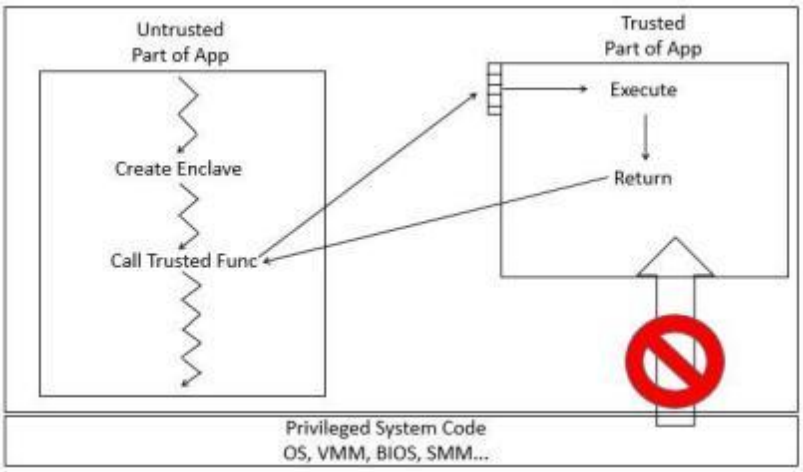


图1-4 IntelSGX系统架构

IntelSGX实现过程中应用程序分为安全部分和非安全部分：

应用程序启动enclave，它被放置在受保护的内存中。

当enclave函数被调用时，只有enclave内的代码可以看到它的数据，外部访问总是被拒绝；当它返回时，enclave数据保留在受保护的内存中。

ARMTrustZone

ARM公司提出的TrustZone技术实现硬件隔离机制，主要针对嵌入式移动终端处理器。TrustZone在概念上将SoC的硬件和软件资源划分为安全(SecureWorld)和非安全(NormalWorld)两个世界。所有需要保密的操作在安全世界执行(如指纹识别、密码处理、数据加解密、安全认证等)，其余操作在非安全世界执行　　(如用户操作系统、各种应用程序等)，安全世界和非安全世界通过一个名为MonitorMode的模式进行转换，如图2-5：

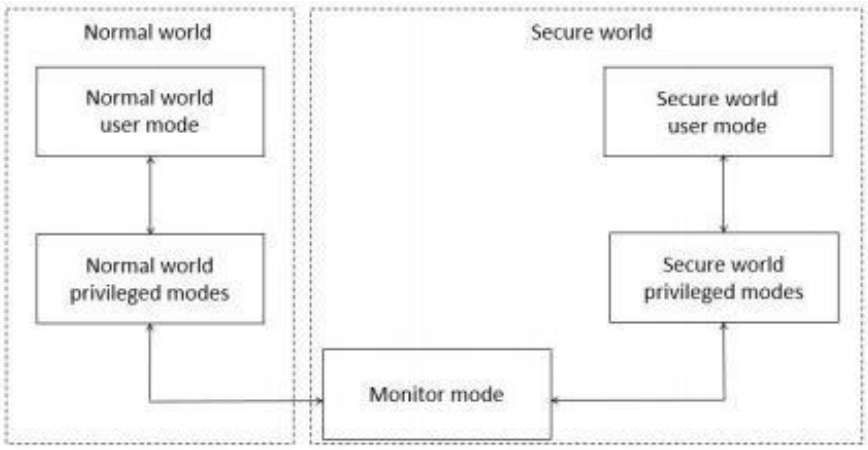


图2-5　Trustzone架构

处理器架构上，TrustZone将每个物理核虚拟为两个核，一个非安全核(Non-secureCore, NSCore)，运行非安全世界的代码；另一个安全核(SecureCore)，运行安全世界的代码。

两个虚拟核以基于时间片的方式运行，根据需要实时占用物理核，并通过MonitorMode在安全世界和非安全世界之间切换，类似同一CPU下的多应用程序环境，不同的是多应用程序环境下操作系统实现的是进程间切换，而Trustzone下的MonitorMode实现同一CPU上两个操作系统间的切换。

（二）数据资产管理相关技术

1．元数据管理

(1) 定义

元数据是对信息资源进行描述、解释、定位或使信息资源更易于被检索、利用及管理的结构化信息。元数据也被称作数据的数据(dataaboutdata)或信息的信息(informationaboutinformation)。元数据管理是对数据的组织、数据域以及其关系信息的管理，贯穿整个数据价值流，覆盖从数据产生、汇聚、加工到消费的全生命周期。通过元数据管理，可以提升共享、重新获取数据和理解企业信息资产的水平。

(2)作用机理

元数据管理实现了业务数据中实体和数据字段元素的定义、语义、业务规则和算法以及数据特征，便于完成业务数据的收集、组织和管理。元数据管理技术包括对元数据的采集、元数据的存储以及元数据的管理，如图2-1所示。

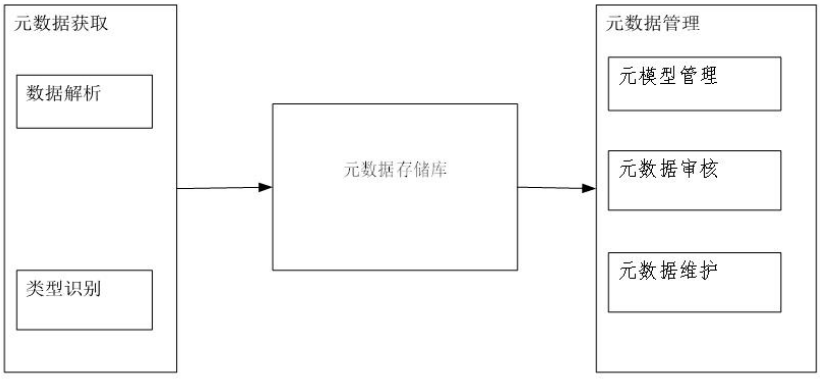


图2-1 元数据管理技术

元数据采集

在采集数据的时候，面对的大问题就是多种类数据源解析适配，以及数据调度任务的抽象，必须开发对应的工具来实现各种场景的元数据解析，对元数据采集的能力有两方面要求：

数据解析：适配解析各种数据源特点，文件格式，SQL脚本，抽象任务等，完成标准元数据的转换沉淀;

类型识别：十分复杂的一个节点，类型在描述数据的时候至关重要，结构化存储可以直接读取，文件类结构通常需要类型转换标识，任务流程会直接统一管理，依次保证数据在不同环境中的合理存储;

元数据的常见来源如下表：

序号	元数据来源类型	元数据来源
1	关系数据库	Oracle、MS SQLServer、DB2 等
2	建模工具	ERWin、PowerDesigner 等
3	数据集成工具	DataStage、PowerCenter 等
4	BI 报表工具	Cognos、SQL Server Reporting Services 等
5	调度工具	Automation
6	开发语言及脚本	Perl(日志方式)、SP(注释方式)
7	其他	元数据采集虚拟库等

表2-1 元数据来源

元数据存储

包括存储元数据以及元模型。元数据包含数据库信息、数据表信息、表的字段信息等元信息，元模型是描述元数据的元数据。这些信息梳理存储在关系型数据库中。

元数据通常分为业务、技术和操作三类：

业务元数据：描述业务领域的相关概念、关系和规则的数据，包括业务术语、信息分类、业务指标、业务规则、资产目录、Owner、数据密集等。

技术元数据：实施人员开发系统时使用的数据，包括模型的表与字段、ETL规则、集成关系等信息。

管理元数据：描述运营管理领域的相关概念，包括人员角色、岗位职责和管理流程、调度评率、访问日志等。

元数据管理

对收集到的元数据进行管理包含如下3个方面：

元模型管理：即基于元数据平台构建符合CWM规范的元数据仓库，实现元模型统一、集中化管理，提供元模型的查询、增加、修改、删除、元数据关系管理、权限设置等功能，支持概念模型、逻辑模型、物理模型的采集和管理，让用户直观地了解已有元模型的分类、统计、使用情况、变更追溯，以及每个元模型的生命周期管理。同时，支持应用开发的模型管理。

元数据审核：主要是审核采集到元数据仓库但还未正式发布到数据资源目录中的元数据。审核过程中支持对数据进行有效性验证并修复一些问题，例如缺乏语义描述、缺少字段、类型错误、编码缺失或不可识别的字符编码等。

元数据维护：对信息对象的基本信息、属性、被依赖关系、依赖关系、组合关系等元数据的新增、修改、删除、查询、发布等操作,支持根据元数据字典创建数据目录，打印目录结构，根据目录发现、查找元数据，查看元数据的内容。元数据维护是最基本的元数据管理功能之一，技术人员和业务人员都会使用这个功能查看元数据的基本信息。

(3) 所解决的问题

元数据管理作用于数据管理的全生命周期，能够有效促进数据资产的管理、交换、共享和开发利用，其在工业数据空间中可解决如下问题：

促进工业数据共享，元数据管理通过搭建标准统一的工业领域元数据体系，从业务、技术、操作3个方面，进行分类和定义，赋予数据意义，可帮助使业务人员快速获取可信数据。

提高工业数据的整合和溯源能力，元数据管理通过对所有的关键数据进行系统整合和处理，用户可以对数据处理流程追本溯源，了解业务处理规则，数据流通情况等。

激活工业数据活性：通过元数据管理，可以实现暗数据的透明化，增强数据活性，帮助解决工业数据资产盘点的问题。

支持工业数据增值变现：元数据管理通过分析表模型可敏捷响应海量增长的数据分析需求，支持工业数据增值和数据变现。

建立工业数据质量稽核体系，元数据管理通过建立筛选、核实、管理、报警、监控的机制，支撑业务管理规则有效落地，保障数据内容的合格、合规。

(4) 适配性

关系型数据库：能适用来自Oracle、DB2、SQLServer、MySQL等关系型数据库的库表结构、视图、存储过程等元数据。

非关系型数据库：支持来自MongoDB、Redis、Neo4j、HBase等非关系型数据库中的元数据。

数据仓库：对于主流的数据仓库，可以基于其内在的查询脚本，定制开发相应的适配器对其元数据进行管理。

其他元类型元数据：ER/Studio等建模工具、Kettle等ETL工具、PowerBI等前端工具、Excel格式文件的元数据。

(5) 同类技术对比

本小节将把元数据管理技术与主流相关技术进行对比，包括主数据管理(MDMMasterDataManagement)和产品数据管(ProductDataManagement)，主要从时效性、成本、对数据的控制、管理范围、管理方法等几个方面进行对比，如表2-2所示。

	主数据管理(MDM)	元数据管理	产品数据管理(PDM)
概述	自动、准确、及时地分发和分析企业中数据，并对数据进行验证	对描述数据的数据进行管理，即对数据及信息资源的描述性信息进行管理	管理所有与产品相关信息和所有与产品相关过程的技术
时效性	低	高	中
成本	需要管理主要的数据，实施成本低	需要管理数据的描述性信息，成本高	需要管理所有与产品相关的信息，成本高
对数据的控制	低	高	高
管理范围	业务对象的、具有持续性、非交易类的数据	描述数据的相关信息而存在的数据	产品生命周期内生产的所有数据
管理方法	使用MDM应用程序法和MDM平台法	建符合CWM规范的元数据仓库进行管理	协调组织整个产品生命周期内的过程事件

表2-2 元数据管理技术与主流相关技术对比

(6) 技术成熟度分析

元数据管理技术目前有着较高的技术成熟度，能够支持多种数据格式，包括结构化数据类型，半结构化数据类型，能够适应不同场景下元数据的采集。例如ApacheFlink是一个流式的数据流执行引擎，其针对数据流的分布式计算提供了数据分布、数据通信以及容错机制等功能，ApacheFlink提供了数据仓库的元数据管理工具，使用了一种能获取元数据资产的算法,通过该算法能够得到SQL中的表、字段之间的依赖关系。实现了字段级血缘依赖，全面、细粒度的元数据资产将会极大地减少后期数仓的维护成本，可以最大限度地减少数据本身造成的问题。在元数据的管理策略中，目前有集中式元数据管理策略、分布式元数据管理策略和无元数据管理策略，这三种不同的元数据管理策略能满足不同场景的需求，能在多种实际应用场景中发挥用途。

2. 数据安全态势感知

（1）定义

数据安全态势感知是对数据全生命周期各个环节的操作状态、合规状态、异常状态进行采集、融合分析，在此基础上对数据安全风险进行评估，并预测未来数据安全走势以及可能产生的影响范围和程度。

数据安全态势感知，一般分为三个层次的处理，要素感知、态势理解、态势预测。要素感知是感知数据存储和服务环境中涉及数据安全的相关要素，包括数据属性、使用状态等信息。态势理解是对一定范围内数据的使用状态进行综合判定。态势预测是在态势理解的基础上对数据安全未来演变趋势进行评估。

（2）作用机理

数据安全态势感知如图2-2所示，机理包括轻量级全网覆盖的数据使用状态采集、分层分级动态汇集、数据安全区域态势分析、数据安全全网态势分析、态势全景折叠可视化展示等部分。

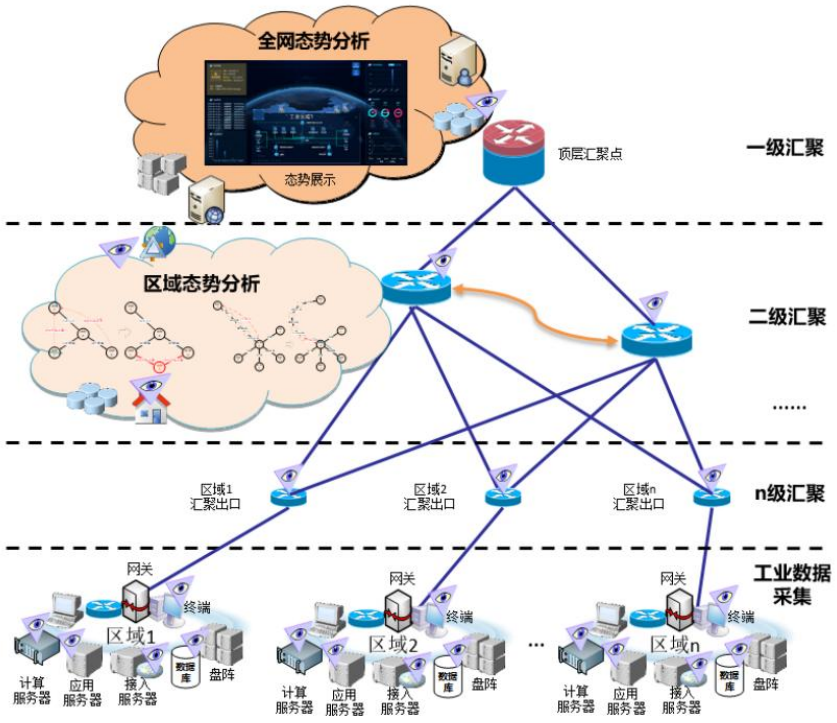


图2-2 数据安全态势感知体系

（3）成熟度分析

没有人精确定义数据安全态势感知，在本报告中给出了一种定义。

数据安全态势感知是新兴的安全技术，与发展多年的网络安全态势感知相比，数据安全态势感知存在几个特点：(1)在数据安全信息采集、汇聚、分析评判方面未形成标准建议，各个安全产品厂家根据自己的理解进行数据的采集、组织管理和算法研发，在不同数据场合形成不同的采集格式和数据清洗方法，整合不同厂家数据接口和服务的数据安全态势感知实现不易，且工作量大。(2)数据安全态势感知涉及的数据环节多，需要感知采集的数据量更大，数据分析的算力整合和分解调度面临巨大挑战。(3)数据安全态势感知需要理解数据的语义，与数据的业务流程、信息系统形成紧耦合，需要较多的投入完成所需的安全分析。(4)数据安全态势感需要在边界节点对数据的跨域进出、授权流通细微状态进行掌控，面临不同厂家多类设备上的数据存储和处理进行使用状态采集的需求，需要与多个厂家的多款设备、多种流程、多个数据保护方案对接，涉及厂家不愿配合、采集程序按需开发、采集接口难以规范化等问题。

由于存在海量、多源、流转关系复杂的数据使用场景，并且数据本身又具有多样性、敏感程度不一、关联关系复杂等特征，数据安全态势感知的实现存在投入大、设计复杂的挑战，促进其推广和发展需要解决以下两点问题：

明确数据安全的战略地位：需要明确数据安全对于可信数据数据流通的重要意义，从战略高度明确数据安全态势感知的重要价值，对数据使用和安全管理进行清晰规划与指导，在管理层面达成一致共识，推进数据语义、数据采集、数据授权等数据安全多个方面的标准化，为数据安全态势感知工作提供资源保障。

规范掌控全局数据使用状态：数据使用的复杂性和多样性给数据安全态势感知提出了较大挑战，掌握掌控全局数据使用状态是数据安全态势感知的重要前提，需要提出可信数据数据流通中各类数据的存储、使用、流通等环节的状态监管要求，并规范状态采集接口。

3. 标识解析

(1) 概念

标识解析体系通过赋予每一个实体物品(产品、零部件、机器设备等)和虚拟资产(模型、算法、工艺等)唯一的“身份证”，实现全网资源的灵活区分和信息管理，是实现企业数据流通、信息交互的关键枢纽。

(2) 作用机理

标识解析体系的核心包括：标识编码，标识解析系统，标识数据服务等三个部分。

标识编码，是能够唯一识别机器、产品等物理资源和模型、算法、工艺等虚拟资源的身份符号，类似于“身份证”。

标识解析系统，能够根据标识编码查询目标对象网络位置或相关信息的系统，对机器和物品进行唯一性的定位和信息查询。

标识数据服务，能够借助标识编码资源和标识解析系统开展标识数据管理和跨企业、跨行业、跨地区、跨国家的数据流通及基于数据的其他增值服务。

(3) 在可信工业数据空间中所解决的问题

通过标识实现数据资产管理及安全可信高效的数据共享流通，促进数据产业链共享、流通模式。具体可实现数据资产主动感知、自动化理解与灵活获取三个方面：

基于主动标识等技术，可以实现第三方数据资产服务的分布式发布，形成数据资产服务目录，从而支撑用户对工业数据空间中的数据与服务进行自动化快速感知。

基于标识赋予每一个物理实体和虚拟数据唯一身份信息，并对实体进行结构化、标准化表征，从而支撑用户对第三方数据资产服务的自动化准确理解。

基于解析技术，结合标识编码方案，实现第三方数据资产服务的高效寻址与灵活获取。

(4) 在可信工业数据空间中的适配性

本小节将从适用的软硬件要求、云端兼容性、中心化/去中心化模式等方面分析该技术在可信工业数据空间中的适配性。

软硬件要求

标识解析系统可以在通用设备上运行。此外，标识解析技术不依托DNS服务，因此，无需额外部署DNS服务。对比传统的基于DNS的改良路径标识方案，如EPC、OID等均需要对现有DNS架构进行扩充，通过将编码树映射为DNS树的一部分提供服务，即将编码解析服务覆盖在DNS服务之上，解析服务依赖DNS资源记录，安全防护依托于DNS安全保障措施。因此，传统的编码解析方案需要依托DNS系统进行建设。

云端兼容性

标识解析系统适用于云端部署。相较于运行环境各异的用户终端，统一的云端环境更易管理与配置。云端在本质上是物理机或物理机的集合，因此作用于系统层的标识解析技术依然可以生效。目前标识解析技术有企业自建与云托管两种模式，企业可以选择在本地部署企业节点，提供标识解析服务，也可以选择托管到二级节点。而用户则可以使用浏览器进行访问。

中心化/去中心化模式

目前标识解析技术存在多种技术路径，包括中心化方式与去中心化方式、并且彼此之间的兼容互通方案也在建设之中。根据数字资产提供者与数字资产消费者的关系，标识解析技术可以提供多种服务模式。在中心化标识解析技术中，标识由管理节点集中分配，该种方案适用于支撑一个逻辑主体内部的多个节点间的工业数据共享，如一个企业内部的多个部门；在去中心化标识解析技术中，标识由参与方分布式自主生成，该种方案适用于支撑多个逻辑主体间的工业数据共享，如多个企业。而正在建设中的中心化/去中心化系统兼容互通方案可以支撑工业全产业、全链条的数据共享。

(5) 与同类技术的关系

标识解析技术是传统编码寻址方案在系统架构、标识对象和输入输出、解析方式等因素上丰富与革新。

本小节将把标识解析技术与主流相关技术进行对比，包括基产品电子代码(EPC, electronicproductcode)、对象标识符(OID, objectidentifier)，物联网统一标识(Ecode, entitycodeforIoT)、句柄(Handle)、泛在识别技术(UID, ubiquitousID)等。

技术名称	系统架构	标识对象	输入输出	解析方式	是否需要依托DNS系统建设
标识解析技术	混合结构	物理或数字资产	标识映射为资产信息	递归	否
DNS	树状结构	主机	域名映射为IP	递归，迭代	/
EPC	树状结构	物理对象	标识映射为URL	迭代	是
OID	树状结构	物理对象或逻辑对象	标识映射为URL或IP	递归	是
Handle	树状结构	数字对象	标识映射为一组值的集合	迭代	否
UID	树状结构	物理对象或逻辑对象，以及他们之间的关系	标识映射为背景描述	递归	否
GNS	图状结构	用户、物品和组织	标识映射为公钥或IP	迭代	否
BNS	树状结构	主机	标识映射为IP		否

表2-4 同类技术对比

（三）可信环境

1. 可信执行环境

(1) 定义

可信执行环境(TrustedExecutionEnvironment，简称为TEE)是计算平台上由软、硬件方法构建的一个安全区域，可保证在安全区域内部加载的代码和数据在机密性和完整性方面得到保护。在数据运算过程时通过该安全区域中执行加解密等保证安全。计算结束后，在安全区域内销毁原始数据，确保原始数据不被泄露。

(2) 作用机理

隔离性

X86架构的隔离机制从Intel80286处理器开始，Intel提出了CPU的两种运行模式，并且逐步衍生出后来的不同的特权界别，再后来提出了安全区域更小的SGX机制实现可信执行环境TEE。同样的，ARM架构通过Trustzone技术实现了相关软硬件的隔离性，实现安全世界与非安全世界的隔离。TEE通过隔离的执行环境，提供一个执行空间，该空间有更强的安全性，比安全芯片功能更丰富，提供其代码和数据的机密性和完整性保护。

软硬协同性

虽然标准定义可以通过软件方式或硬件方式实现TEE，但实际生产场景下，行业内更多通过软硬结合的方式进行安全性的保障与支持。

富表达性

TEE与单纯的安全芯片或纯软件的密码学数据保护方案相比支持的上层业务表达性更强，由于只需要定义好业务层面可信区域和非可信区域的逻辑划分，而不会对定义可信区域内的算法逻辑的语言有可计算性方面的限制(图灵完备的)。同时由于TEE已经提供了“安全黑盒”，安全区域内数据无需进行密态运算，所以其运算效率高。

(3) 所解决的问题

其目标主要是确保共享数据或使用共享数据的程序按照预期执行，在可信执行环境下保证初始状态的机密性、完整性，以及运行时状态的机密性、完整性。

计算时的数据保护

不同计算参与方将各自加密后的数据通过安全链路传至TEE计算环境中，在TEE中进行数据解密并进行联合计算，如联合用户画像、金融风控预测、社会风险识别等。计算结束后将计算结果通过安全链路返回给各个参与方，原始数据则在TEE环境中进行销毁。

密钥保护

密钥安全是数据安全、系统安全的关键要素。为了增强密钥的安全性，通常将密钥进行加密存储或引入专用的密钥设备。使用TEE计算环境技术，在硬隔离的TEE环境中进行密钥存储，可以简化传统密钥管理的复杂性，又保证密钥使用过程中的安全可靠。

密钥保护应用场景在TEE计算环境中实现密钥生成、管理等功能，同时支持通用加密算法。一方面收到用户密钥请求时使用密钥生成功能生成随机密钥、证书等安全凭证，并通过安全链路传递给用户；另一方面也支持在TEE中使用生成的密钥结合通用加密算法对原始数据、预测结果等进行加密计算。当密钥过期、失效时，在TEE环境中对密钥进行销毁，增强密钥生命周期整体安全性。

计算模型保护

多个企业利用在生产过程中积累的数据进行联合建模分析，为了得到更精确的结果，这些企业可以引入具有成熟算法模型的合作伙伴。在这种场景需求下，采用TEE计算环境保证算法模型安全。

数据加密存储

数据作为企业资产管理的一部分，关键性数据比如经营分析数据、财务数据、生产数据等更是具有极高的价值。政企客户通常极为重视这类关键数据的机密性保护，为防止数据泄露需要对数据进行加密存储。使用TEE计算技术可以为客户提供数据加密存储能力。该场景使用TEE计算环境作为数据加密模块，用户通过远程证明对TEE环境进行验证，并在TEE环境中生成密钥。客户关键数据在TEE环境内进行加密，加密后的数据可储存在TEE环境中，也可以存储在外部介质中。

隐私查询

在金融、电商、社区治理等领域需具备针对用户身份进行隐私查询能力，如通过指纹，人脸等信息对人员身份进行比对认证。在医疗领域同样存在对患者疾病病历、基因测序等数据的隐私查询。这些隐私数据往往来自多个政府部门或企业。基于TEE计算环境技术是实现这类具有隐私查询需求场景下数据可用不可见的有效方法之一。在该类隐私查询场景中，在TEE计算环境环境中构建数据汇交、统计、查询能力。数据提供方将各自的原始数据经加密安全链路传递至TEE计算环境环境中进行数据汇交与统计分析。数据查询方调用查询接口对其所需的内容发送隐私查询请求，TEE环境中的查询模块根据数据查询方的身份权限向数据查询方返回查询结果。同时也可以结合区块链等技术对数据查询方的查询操作进行存证。采用TEE计算环境进行隐私查询，数据提供方的原始数据与查询的整个过程置于硬件隔离的TEE计算环境环境中，可以实现多方数据的联合汇交，丰富数据库的同时有效降低敏感信息泄露的风险。

(4) 适配性

适用于 a)数据汇集的中心化模式,如数据湖、数据中台、大数据中心；b)数据分布在用户但服务集中的部分去中心化模式，如数据交易所，数据中介商；c)去中心化模式，即没有中间服务方的点对点模式。在各类模式之中，可信执行环境将会作用于数据资产所在位置，即中心化模式的数据中心/存储服务器一侧、部分去中心化和去中心化的用户一侧。

（5）同类技术对比

TEE需要基于预置集成了可信执行控制单元的CPU计算芯片来实现，这便需要确保芯片厂商的安全可信。虽然国外的芯片厂商相比中国厂商拥有更为成熟的产品和技术方案，但是国产芯片厂商拥有更强的自主可控性。在国产化自主可控的需求驱动下，国内芯片厂商将通过持续的研发投入来不断提升国产化可信硬件技术能力。目前国内的兆芯、海光、飞腾等芯片厂商相继推出了TEE技术方案。

技术	国外			国内		
	IntelSGX	TrustZone	AMDSEV	海光CSV	飞腾TrustZone	兆芯TCT
发布时间	2015	2005	2016	2020	2019	2017
指令集架构	X86_64	ARM	X86_64	X86_64	ARM	X86_64
是否支持任意代码运行	是	是	是	是	是	是
硬件安全密钥	有	无	有	有	无	有
完整性认证与封存	支持	不支持	支持	支持	不支持	支持
内存加密	是	否	是	是	否	否
内存完整性保证	支持	不支持	不支持	支持	不支持	支持
TEE安全I/O	不支持	支持	支持	支持	支持	支持
可用内存空间	≤1T	系统内存	系统内存	系统内存	系统内存	系统内存
TCB	硬件：CPUPackage 软件：Enclave内的 代码实现	硬件：安全虚拟核 软件：安全世界OS 和TA	硬件：AMDsecure processor 软件： 虚拟机镜像	硬件：海光SME 软件： 虚拟机镜像	硬件：安全虚拟核 软件：安全世界OS和TA	硬件：CPU&TPCM

表3-1 国内外的可信执行环境(TEE)技术对比

2. 可信软件与系统环境

(1) 定义

可信软件环境是指在软件层面，将App运行时隔离，防止其他App调取数据，缺点无法防范主系统攻击。

可信系统环境是指将受保护App放入虚拟系统隔离，主系统APP无法轻易访问隔离系统内App，但是hypervisor被攻破，所有隔离系统均会被沦陷。常见技术选型有虚拟化以及半虚拟化技术两种。

(2) 已有技术选型

容器化技术

容器化技术通常分为容器化运行技术以及容器化编排技术。这里重点研究容器化运行技术，最为常见的技术就是docker，docker是一种容器化运行技术或平台，基于Linux内核的cgroup和Namespace对进程进行了封装隔离，通过隔离apps的bin/lib,以容器的形式将应用程序及所有依赖软硬协同性打包在一起，做到App运行时隔离。

在对比虚拟化以及半虚拟化技术之前，这里先来补充下CPU如何通过Ring级别进行访问控制的。CPU通过Ring级别进行访问控制，Ring0是最高级别，Ring1，Ring2，Ring3依次降低。以LinuxX86为例，操作系统(内核)的代码运行在Ring0上，可以使用特权指令，可以使用特权指令，控制中断，修改页表，访问控制等。

应用程序的代码运行在最低级别Ring3上，不能做受控操作。如果需要访问磁盘、写文件，需要通过执行系统调用(函数)，执行系统调用时，CPU的运行级别从Ring3到Ring0依次切换，并跳转到系统调用的内核代码位置执行，由内核完成设备访问，之后再从Ring0返回Ring3，实现用户态和内核态的切换。

虚拟化技术

基于二进制的全虚拟化技术和容器化技术不同的是在HostOS主系统上加载了Hypervisor，这里我们常用的有vmware，virtualbox等，通过对内存运行的隔离，实现不同guestos间的隔离。这里区分“主机双系统形式”，虽然两者都是在内存上做了隔离，但是双系统是无论在运行时和非运行时都对主机内存进行了划分与占用，这本身是对计算机资源的一种浪费。而虚拟化技术在guestos非运行状态下，会对占用内存进行释放。

因为宿主机系统工作在Ring0，客户操作系统不能运行在Ring0，当客户操作系统执行特权指令时，就会发生错误。虚拟机管理管理程序(VMM)就是负责客户操作系统和内核交互的驱动程序，运行在Ring0上，以运动程序的形式体现(驱动程序工作在Ring0，否则不能驱动设备)。当客户操作系统执行特权指令时，会触发异常(cpu机制，没权限的指令，触发异常)，VMM捕获这个异常，在异常处做翻译、模拟，返回处理结构到客户操作系统内。客户操作系统认为自己的特权指令工作正常，继续运行。该处理过程复杂，性能损耗比较大。

半虚拟化技术

通过修改客户操作系统代码，将原来在物理机上执行的一些特权指令，修改成可以和VMM直接交互的方式，实现操作系统的定制化。半虚拟化技术XEN，就是通过为客户操作系统定制一个专门的内核版本，和X86、MIPS、ARM这些内核版本等价。这样，就不会有捕获异常、翻译和模拟的过程，性能损耗比较少。这也是XEN这种半虚拟化架构的优势，也是为什么XEN只支持Linux的虚拟化，不能虚拟化Windows的原因。

（四）可信传输

1. 信源和信道加密

（1）定义

信源加密，对信源采取保护措施及对信源发送的信息明文或代表明文的电信号进行加密，使消息不被非法截获或破译的保密方式。

信道加密是采用使窃密者不易截收到信息的通信信道，如采用专用的线路、瞬间通信和无线电扩频通信等。

（2）作用机理

因篇幅原因且技术较为成熟，本小节略。

（3）所解决的问题

信源加密是解决数据存储安全和流通时传输安全：信源加密可以在工业信息数据产生的同时完成对数据的加密，即对数据进行机密性保护，以便于数据信息在数据生成设备临时存储时以及发送给数据接收设备的传输过程中，始终保持数据的机密性，防止数据合法接收者以外的非法窃听、截获或破译。

信道加密是解决数据流通时的传输安全：信道加密是从通信传输信道的起点开始就对传输的各类指令和数据进行加密，直到通信传输信道的终点才解密使用，确保在数据在通信传输信道中流通过程中，始终得到机密性保护，防止通信传输信道路由上非法窃听者的监听、理解和侵入。

(4) 适配性

信源加密主要分为2种，软件加密和硬件加密。其中软件加密指的是通过软件加解密模块对数据进行机密性和完整性保护，数据加解密的整个过程都是软件来完成。软件加密，其基本功能不需要硬件基础设施的支持，可以以应用程序、动态库、软件包的方式安装在数据提供方和数据使用方的设备系统。硬件加密指的是通过专用的加密芯片或独立的处理芯片，如密码机、密码卡、智能密码钥匙等，实现密码算法运算。因此如果采用硬件加密方式，需要硬件基础设施提供支撑，还需要解决应用系统与硬件密码模块之间的硬件和软件接口适配性等问题。

信道加密，无论是有线信道加密还是无线信道加密，均需要硬件基础设施的特点进行调整，此外还需要在软件层面进行进一步地优化。

(5) 同类技术对比

信源加密即在信息传播的源头进行加密，可采用的方法有对称加密和非对称加密两大类。安全设计人员可以根据数据业务和场景需求，采用SM4、AES等对称分组加密算法，也可采用RC4、ZUC等对称流密码算法，或者采用SM2、SM9和RSA等非对称密码算法。

信道加密即在信道的两端采取加解密措施保护整个信道内传输数据的机密性，常见的方式有：

(1)根据网络链路的特点，在链路两端各部署加密机，在链路的发送端加密在链路的接收端解密，例如SDH加密机、ATM加密机和IP网络加密机等解决方案；

(2)采用信道加密的协议来实现，例如IPSec、SSL和TLS等。IPSec协议为IP层设计的通信保护协议，主要通过认证头(Authentication Header，AH)为IP数据报提供无连接数据完整性、消息认证以及防重放攻击保护，或者通过封装安全载荷(Encapsulating Security Payload，ESP)提供机密性、数据源认证、无连接完整性、防重放和机密性，使用IPSec协议可以保护基于TCP和UDP协议的数据通信。SSL是Netscape公司所提出的安全保密协议，在套接字客户端和服务器之间构造安全通道来进行数据传输，SSL运行在TCP层之上、应用层之下，为应用程序提供加密数据通道。SSL通过互相认证、使用数字签名确保完整性、

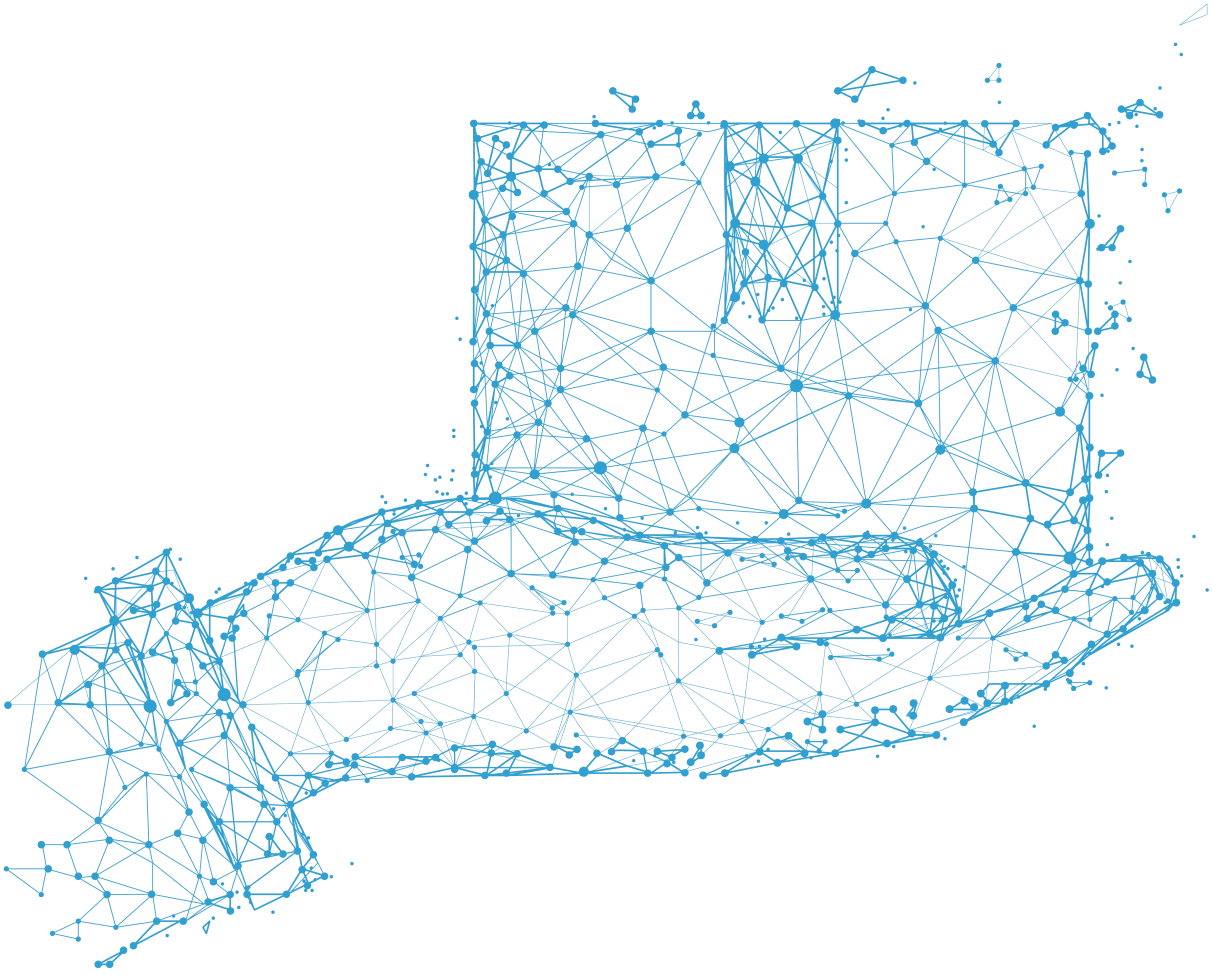
使用加密确保私密性，以实现套接字客户端和服务器之间的通信安全，该协议只能保护基于TCP协议的数据通信，不能提供针对UDP协议数据的保护。TLS是SSL的标准化后的产物，它建立在SSL3.0协议规范之上，是SSL3.0的后续版本。

(6) 技术成熟度分析

信源加密技术较为成熟，因篇幅原因，其成熟度分析略。

信道加密技术存在两大主流协议IPSec和TLS，这两类协议设计逐渐完善，目前正在得到广泛应用和推广。IPSec协议在1986年开始由美国国家安全局赞助下进行研发，在1995年形成标准RFC1825，后来在2011年曾经一度被IPv6标准RFC6434定义为强制要求支持的标准，在2019年在新的IPv6标准RFC8504改为非强制要求的标准。由于IPSec协议支持保护UDP协议数据，因此有不少支持IPSec的产品和通信网关。TLS以及其前身SSL协议，由于其不改变IP层的结构且有便利的开发工具，一直得到广泛应用支持，在1996年形成了SSL3.0规范，

在2008年TLS1.2成为互联网安全通信标准RFC5246，在2018年TLS1.3成为互联网安全通信标准RFC8446。



2. VPN

(1) 定义

虚拟专用网络(Virtual Private Network, VPN)是使用密码技术在公用网络上建立专用网络的技术。通过对网络数据的封包和加密传输，在一个公用网络建立一个临时的、安全的连接，从而实现在公网上传输私有数据，达到私有网络的安全级别。

(2) 作用机理

虚拟专用网络大致分两种：一种是企业内部虚拟网，另一种是客户端远程访问虚拟网。

企业内部虚拟网通过公用网络进行企业总部和各个分部之间互联，是传统的专线网或其他企业网的扩展或替代形式，其实质是在企业总部和各个分部配备VPN设备(或含VPN功能的设备如路由器、安全网关、服务器等)，通过公用网络在各个VPN设备之间建立VPN安全隧道来传输企业的私有网络数据，用于构建这种VPN连接的隧道技术有IPSec、GRE等。大致示意如下图所示。

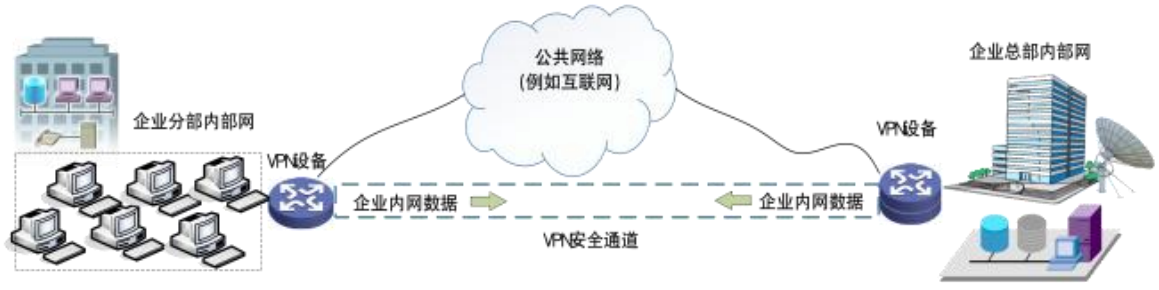


图4-1企业内部虚拟网

客户端远程访问虚拟网是指，企业在公共网络搭建VPN服务接入点，企业的

内部人员如出差人员使用网络隧道协议(一般的通用VPN客户端软件可提供)通过公共网络与企业VPN服务接入点建立一条的隧道连接从而访问企业内部网资源。大致示意如下图所示。

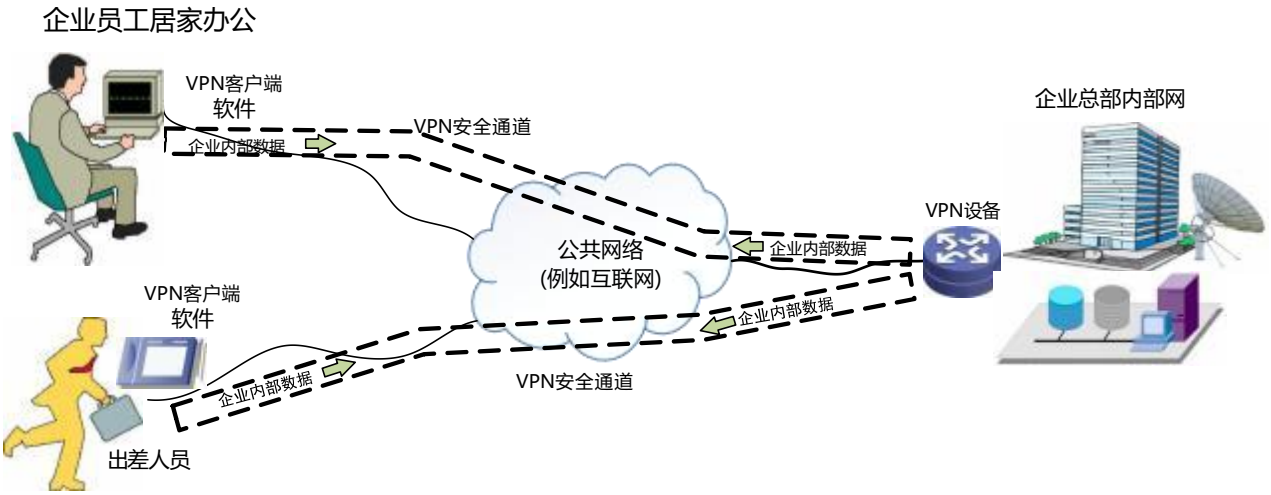


图4-2客户端远程访问虚拟网

(3)所解决的问题

VPN技术确保了数据流通时的传输安全：综合采用了非对称签名验签算法、非对称加解密算法和对称加解密算法，其中非对称签名验签算法确保了通信数据来源的认证性，防止非法人员接入并实施假冒通信；非对称加解密算法和对称加解密算法确保了通信数据的机密性，防止非法人员进行数据的窃听和理解。总之，VPN技术相当于给VPN通信的源点到终点之间建立了一条认证和机密的传输通道，确保了数据流通时的传输安全。

VPN加密隧道方式确保了内网地址保护：通过隧道协议对内网地址进行统一的接口地址转化，在加密隧道路由上的任何人只能看到隧道的起点接口地址和终点接口地址，无法知道内网中哪个地址进行了加密通信，防止非法窃听者对VPN通信中的内网地址进行分析。

(4) 适配性

①搭建企业内部虚拟网，需要在企业总部和各个分部配备VPN设备(或含VPN功能的设备如路由器、安全网关、服务器等)，其中至少一个VPN设备部署在公用网络上具有公网IP地址，然后在各个VPN设备配置VPN安全隧道，通过密钥协商等方式建立安全隧道来传输企业的私有网络数据。

②客户端远程访问虚拟网，需要企业在公共网络搭建VPN服务接入点(VPN设备或VPN服务器)，远程用户配备VPN客户端软件，通过公共网络与企业VPN服务接入点建立一条的隧道连接从而访问企业内部网资源。

（5）同类技术对比

协议名称	主要特点
L2F	由CISCO公司开发的点对点隧道协议。L2F协议本身并不提供加密，而是依赖于传输内容结构中的安全协议以提供数据机密性保护。
L2TP	第二层隧道协议，支持在IP、ATM、帧中继、X.25等多种网络中建立隧道。可以在运营商的链路层网络(例如ATM、帧中继等)单独使用，也可以在IP网络(例如互联网)中结合PPP协议使用，自身不提供加密与可靠性验证的功能，可以和传输内容结构中的安全协议搭配使用，从而实现数据的加密传输，常与IPSec搭配。
PPTP	点对点隧道协议，利用改进的通用路由封装协议(GRE)来实现点对点的数据封装和响应控制，以此形成一个点对点的数据传输隧道，但PPTP协议本身不提供加密或身份验证，依赖于PPP协议完成认证，完成认证后的隧道同样没有安全防护，因此需要结合其他安全协议来保护传输数据。
IPSec	IPSec是一个协议框架，通过对IP协议的分组进行加密和认证来保护基于IP协议的网络数据。
TLS/SSL	TLS/SSL协议主要由握手协议和记录协议组成，它们共同为应用访问连接提供认证、加密和防篡改功能。TLS/SSLVPN是解决远程用户访问企业内网最简单安全的解决方案。

表4-1 VPN技术协议对比

(6)技术成熟度分析

随着IP技术和因特网的快速发展，国内外在基于IP网络的VPN技术的探索也得到广泛的开展。基于IP网络的VPN技术研究，是在因特网事实的标准化组织-互联网工程任务组织(IETF)的架构下，由来自世界各地、各公司和组织的研究人员组成工作组联合完成的，并由IETF推动较成熟技术的标准化工作。

目前VPN框架已基本成熟，在VPN体系下发展了 IPSecVPN 、SSLVPN等技术，且均已成型为国际标准。国内，针对 IPSecVPN 和 SSLVPN 也出台了相关的密码标准，《 IPSecVPN 技术规范》《 IPSecVPN 网关产品规范》《 SSLVPN 技术规范》《 SSLVPN 网关产品规范》。今后VPN发展趋势主要是继续丰富和扩展VPN框架内允许的密码算法。

(五) 供需对接相关技术

(1)定义

数据确权是指确定工业数据在流通过程中的所有权和权益权，保护数据所有权人对数据财产直接控制和支配的权利，支撑数据要素市场的有效运行。

(2)作用机理

数据准确确权可引导数据资源被合理高效地利用，确保公平有序竞争环境。数据确权包括：数据内容确认与属性标识、权利登记与权属转移、权属仲裁等环节，其作用机理如5-1所示。

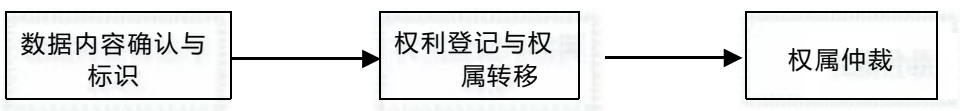


图5-1 数据确权核心环节

数据内容确认与属性标识：首先确认工业数据的数据内容，对数据进行唯一标识，并规划化描述数据属性，其中数据属性包括数据主体、数据类别、数据大小、数据产生时间等。

权利登记与权属转移：在确保数据合规、真实、准确的基础上，评估数据价值，定价数据资产，在权威第三方进行权利登记，必要时权属转移。

权属仲裁：数据在流通过程中争议发生时，采用证据交叉认证、审计等方式仲裁数据所有权、使用权和收益权。

2. 数据目录

(1)定义

数据目录(datacatalog)是企业中所有数据资产的详细目录，是元数据的集合，与数据管理和搜索工具相结合。同时数据目录作为可用数据的清单，提供数据资产的预期用途，帮助数据专业人员快速找到适合分析达到业务目的的最合适的数据。

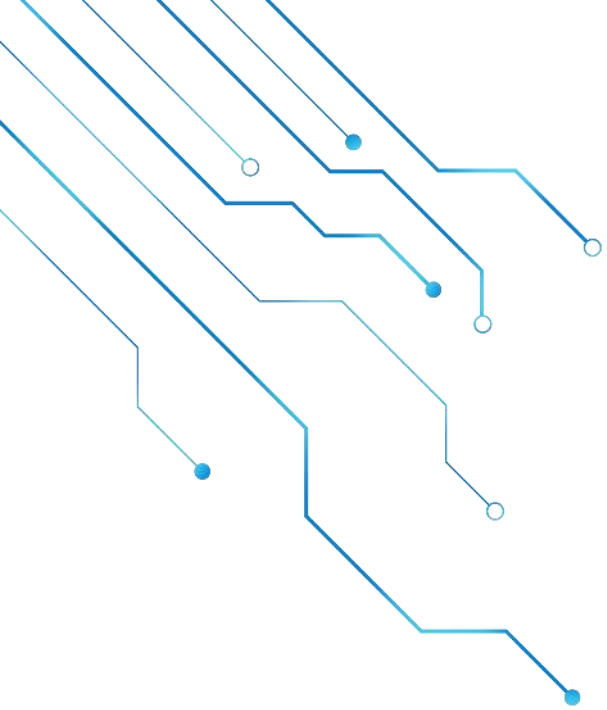
(2)作用机理

数据目录依赖于数据编目的核心能力——收集识别和描述可共享数据库的元数据。数据专业人员使用机器学习等人工智能方法进行元数据收集、语义推断和标记，能从自动化中获得最大的价值并最大限度地减少人工努力。

数据目录以元数据为核心，能够实现许多特征和功能，最基本的功能包括以下三个方面。

(a)数据集搜索：数据目录具备灵活的搜索和过滤选项，从而赋能用户快速找到相关数据集，以实施数据科学、分析或数据工程。数据目录的搜索能力，包括通过面、关键词和商业术语进行搜索。

(b)数据集评估：重要性的数据集评估包括预览数据集、查看所有相关元数据、查看用户评级、阅读用户评论和馆长注释以及查看数据质量信息。



(c)数据访问：从搜索到评估，再到数据访问的路径应该是一个完整的服务，数据访问功能包括对敏感数据的机密性、隐私性和合规性的控制。

除此之外，数据目录也提供许多其他功能，包括支持数据策划和协作数据管理、数据使用跟踪、智能数据集推荐和各种数据治理功能。

(3)在工业数据空间解决的问题

数据目录可以应用在数据治理的以下场景：

(a)提高数据理解能力：分析师可以通过数据目录找到数据的详细描述，包括数据用户的评论，帮助其更好地了解数据与业务的相关性。

(b)提高运营效率：数据目录对用户和IT数据提供不同类型的功能，用户可以更快地访问和分析数据，IT员工可以将更多时间集中在高优先级任务上。

(c)降低数据泄露风险：数据目录帮助根据行业和数据隐私法规使用授权用于特定目的的数据，还可以快速查看注释和元数据，以发现可能影响分析的空字段或错误值。

(d)高效制订数据管理计划：降低数据分析中发现、访问、准备和信任数据的难度，从而提高商业智能计划和大数据项目成功的可能性。

(e)提高数据分析的效果和速度：数据专业人员可以根据上下文数据，分析和回答快速应对问题、挑战和机遇。

(4)同类技术对比(数据目录与数据字典的区别)

(a)数据字典(DataDictionary)是对数据库、数据模型或数据源中数据结构的规范和描述。它由实体、表、数据集及其字段、列、数据元素的列表组成。数据字典可以包含各种范围的信息，具体取决于用例。其中一些是数据类型、描述、关系、别名、约束、源等。

数据字典常见形式如下图所示：

	Key	Name	Data type	Null	References	Description
	🔑	BusinessEntityID	int		Person.BusinessEntity	Primary key for Person records.
		PersonType	nchar(2)			Primary type of person: SC = Store Contact, IN = Individual (retail) customer, SP = Sales person, EM = Employee (non-sales), VC = Vendor contact, GC = General contact
		NameStyle	bit			0 = The data in FirstName and LastName are stored in western style (first name, last name) order. 1 = Eastern style (last name, first name) order.
		Title	nvarchar(8)	✓		A courtesy title. For example, Mr. or Ms.
		FirstName	nvarchar(50)			First name of the person.
		MiddleName	nvarchar(50)	✓		Middle name or middle initial of the person.
		LastName	nvarchar(50)			Last name of the person.
		Suffix	nvarchar(10)	✓		Surname suffix. For example, Sr. or Jr.
		EmailPromotion	int			0 = Contact does not wish to receive e-mail promotions, 1 = Contact does wish to receive e-mail promotions from AdventureWorks, 2 = Contact does wish to receive e-mail promotions from AdventureWorks and selected partners.
		AdditionalContactInfo	xml	✓		Additional contact information about the person stored in xml format.
		Demographics	xml	✓		Personal information such as hobbies, and income collected from online shoppers. Used for sales analysis.
	🔑	rowguid	uniqueidentifier			ROWGUIDCOL number uniquely identifying the record. Used to support a merge replication sample.
		ModifiedDate	datetime			Date and time the record was last updated.

图5-2 数据字典的常见形式

(b)二者的关系

数据目录通常包括数据资产的数据字典。因此，可以将数据字典视为数据目录的构建基块，两者都是元数据管理策略的重要组成部分。

(C)二者区别

数据字典记录特定数据库的技术元数据，而数据目录是跨域数据资产的所有元数据(技术、治理、操作、协作、质量和使用)统一访问、控制和协作层。数据字典有助于更好地理解和信任数据库中的数据，而数据目录有助于查找、理解、信任和协作处理数据。

数据字典和数据目录的区别如下表所示：

	数据字典	数据目录
定义	数据集和元素的定义	企业范围数据资产清单
类型	元数据(信息)	软件或带有实际数据库的软件
服务范围	数据源或数据模型	企业数据管理
元数据	数据集、字段、关系、定义等	数据资产、业务术语表、分类、数据沿袭
目的	描述数据库中的数据	编目企业数据以进行分析

表5-1数据字典和数据目录的区别

(5)数据目录的技术成熟度分析

数据目录的原始驱动力是“政务数据资源共享”，最早由政府提出，并在2007年正式发布国标：《GB/T21063-2007政务信息资源目录体系》。

Okera的研究人员揭示了2021年将出现的五个关键数据行业趋势和预测，其中明确表示“企业在数据目录和元数据管理上的投资将会带来回报”。到2023年，越来越多的企业将利用元数据创建通用而灵活的业务规则和请求处理。

当前，数据目录技术正处于高速发展，今后数据目录的发展将呈现以下趋势：1.随时随地接入高速网络，具备实现超级连接的能力；2.数据目录系统向云端迁移，工作负载与位置相匹配；3.通过机器学习增强元数据目录，提高数据的适应性。

3. 数据血缘追踪

(1) 定义

数据血缘(Datalineage)，又称数据血统、数据起源、数据谱系，是指数据的全生命周期中，数据从产生、处理、加工、融合、流转到最终消亡，数据之间自然形成一种关系，展现了数据产生的链路关系。数据血缘追踪也是元数据管理的重要应用之一，其梳理系统、表、视图、存储过程、ETL、程序代码、字段等之间的关系，并采用图数据库进行可视化展示。根据集成的数据库或视图，通过数据血缘，获得结果数据的来源信息;更新数据时能够反映原始数据库的变化，查看数据在数据流中变化过程。

(2) 作用机理

数据血缘追踪，即对各资源涉及的数据流经路径进行跟踪，类似于追踪数据的“血缘关系”。其可针对数据向下做影响分析或向上做溯源分析，有助于用户管理资源和排查问题。具体为：

影响分析：了解资源(如数据源、数据表、API、数据模型、SQL、模型等)被下游的使用情况，便于在更改资源时评估影响。

溯源分析：对资源(如图表、数据模型)的错误、疑问进行溯源，查明根因。主要特征数据血缘关系主要包含4个特征：

归属性：数据是被特定组织或个人拥有所有权的，拥有数据的组织或个人具备数据的使用权，实现营销、风险控制等目的。

多源性：同一个数据可以有多个来源，数据是由多个数据加工生成的，或者由多种加工方式或加工步骤生成的。

可追溯：数据的血缘关系体现了数据的全生命周期，从数据生成到废弃的整个过程，均可追溯。

层次性：数据的血缘关系是具备层级关系的，一个用户拥有多个数据库，一个数据库中存储着多张表，而一张表中有多个字段。他们有机结合在一起，形成完整的数据血缘关系。

(3)在工业数据空间解决的问题

数据血缘追踪可以应用在数据治理的以下场景：

数据溯源

用户分析处理的数据，可能来源很广泛，不同来源的数据，其数据质量参差不齐，对分析处理的结果影响也不尽相同。当数据发生异常，用户能追踪到异常发生的原因，把风险控制在适当的水平。依托于数据血缘的可塑性特点，根据血缘中的数据链路关系，可实现指定数据的来源、去向的追溯，可帮助用户理解数据含义、在全流程上定位数据问题、进行数据关联影响分析等，解决多层复杂逻辑处理后的数据难以理解、难以应用、出现问题难以定位的问题。

数据价值评估

传统的数据价值评估，往往完全依靠相关法规要求和业务经验，缺少在具体应用场景中的评估依据，数据价值评估脱离了数据的应用场景和真实的业务价值。而数据血缘则提供了一种基于数据实际应用的价值评估方法：使用者越多(需求方)、使用量级越大、更新越频繁的数据往往更有价值。

数据受众：在血缘关系图上，数据流出节点表示受众，亦即数据需求方，数据需求方越多表示数据价值越大；

数据更新量级：数据血缘关系图中，数据流转线路的线条越粗，表示数据更新的量级越大，从一定程度上反映了数据价值的大小；

数据更新频次：数据更新越频繁，表示数据越鲜活，价值越高。在血缘关系图上，数据流转线路的线段越短，更新越频繁。

数据质量评估

数据血缘清晰地记录了数据来源以及数据流转过程中的处理方式和处理规则，能实现对各个数据节点的分析和数据质量评估。

数据归档参考

数据血缘中记录了数据的去向，可清晰地掌握数据被消费的情况，一旦数据没有消费者，那也就意味着数据已经失去价值。此时，可以对数据进行进一步评估，考虑进行归档或销毁处理。

(4) 同类技术对比(数据血缘与影响分析的区别)

(a) 影响分析(ImpactAnalysis)的起点是当前分析对象，终点是受其影响的最末

端子代，按照影响关系逐层扩展。影响分析反映了当前对象在统一数据集成平台中，参与了哪些数据的形成。用户可以借助影响分析观察该对象的影响能力，即对于当前数据修改，会对哪些后代数据造成影响。

(b) 二者的关系

数据血缘追踪和影响分析两个关键特性为用户重建了整个数据管理的构建过程，刻画了家族成员彼此连接的脉络和途径。当数据出现错误或者异常时，用户可通过数据血缘向上分析锁定问题产生的源头，当对某些数据进行修改时，可通过影响关系向下分析，发现哪些数据实体中的数据会受到影响。

充分理解并运用这两种数据**分析**方式，将帮助用户在对海量数据进行分析时，降低排查错误的难度，预测并控制即将造成的影响，最终达到提升数据质量的效果。

(C) 二者区别

数据血缘追踪是分析数据的上游数据信息，用于追溯数据的来源和加工过程。影响分析是分析数据的下游数据信息，用于掌握数据变更可能造成的影响。

(5)技术成熟度分析

随着数据的爆发式增长，数据之间的关系也变得越发复杂。在这样的背景下，具备可塑性、归属性等特征的数据血缘最终将数据治理过程中发挥越来越大的作用。数据的血缘对于分析数据、跟踪数据的动态演化、衡量数据的可信度、保证数据的质量具有重要的意义。

2022年，目前的数据血缘大多是基于技术的梳理，一般服务技术人员的需求。2020年，随着数据服务走向前台，服务业务分析和CDO的业务数据血缘，目前已经有相关产品，通过数据的语义分析，将技术元数据映射到业务元数据上，将数据血缘以业务流程方式发布共享出来，辅助商务决策，这是未来的发展方向之一。

4. 区块链与智能合约

(1) 概念

区块链是一种按时间顺序将不断产生的信息区块以顺序相连方式组合而成的一种可追溯的链式数据结构，是一种以密码学方式保证数据不可篡改、不可伪造的分布式账本，是一个分布式存储数据库。区块链技术包含分布式数据存储、共识机制、P2P点对点传输及加密算法等，是计算机技术的新型应用模式。

智能合约：是一套以数字形式定义的承诺，是区块链上的所有节点共同接受的可编辑自动执行的通用协议，是由事件驱动的、具有状态的、运行在可复制的共享区块链数据账本上的一段计算机代码程序。区块链所有参与者接受编写好的智能合约，在符合执行条件的情况下自动执行条约，并更新数据库记录合约的执行情况。

(2) 作用机理

共识机制、智能合约和Token机制是区块链技术体系的核心特征。

(a)共识机制

区块链中的每一个新增区块要经过记账者确认，其他参与者对新确认区块进行共识验证，确定新区块合法性。典型解决方案通过工作量证明(PoW)、权益证明(PoS)和实用拜占庭容错(PBFT)来实现。其中PoW中参与者依赖运算能力来获取记账权，PoS中持有权益越高的参与者获得记账权的难度越低，PBFT则是在区块链上不同参与者两两进行信息交换和形成共识。

(b)智能合约

区块链上的所有节点共同接受的可编辑自动执行通用协议，典型方案为区块链所有参与者接受编写好的智能合约，在符合执行条件的情况下自动执行条约，并更新数据库记录合约的执行情况。

(c)Token机制

加密算法和分布式账本使Token成为可流通的加密数字权益证明，通常以一定预设机制生成并分发给部分或全部区块链参与者，通过设置数量上限控制Token的贬值速度。

(3) 在可信工业数据空间中所解决的问题

区块链技术主要解决了工业数据空间中的身份认证、自动化交易与可信存证问题。

(a)身份认证

区块链结合证书技术可支撑构建可信工业数据空间中的身份认证商，并替换传统集中式身份认证方案。通过链上存储证书验证方式与本地持有证书相结合的方式，解决大量身份数据需存储在集中式节点的挑战；并且用户不必依赖可信第三方身份认证商来访问不同域的服务，从而解决传统方案中，身份提供商可以看到用户和服务提供商之间的所有交易从而带来的信息泄露问题，保证用户身份属于用户；同时，区块链因其不可篡改特性，保证身份验证的可信性。

(b)自动化交易与可信存证

区块链技术可以实现工业数据空间中的高效交易清算与可信存证，在数据使用之前，数据的使用者与数据的提供者通过协商，形成智能合约；在数据使用阶段，当供需关系、用户意图与使用要求等符合执行条件时自动执行智能合约，并更新数据库记录合约的执行情况，完成工业数据空间中自动交易，实现多个节点分布式互通与高效协同，保证交易存证与可追溯。

(4) 在可信工业数据空间中的适配性

(a) 软硬件要求

在硬件方面，传统的区块链可以在通用设备上运行。以到目前为止运行时间最长的比特币为例。在硬盘需求方面，比特币从08年开始运行，区块链内容大小达210G，因此要求硬盘至少500G的空间；在内存方面，一般要求4G，但如果需要查询历史记录，则需要加载完整的交易索引表-tindex，这导致需要8G+的内

存，所以一般16G的内存较为适宜；在CPU方面，4核即可；在带宽方面，要求至少8M/s的下载带宽，推荐配置25+M/s的下载带宽。以太坊和超级账本由于运行时间尚短，对硬盘的要求甚至比比特币的要求更低，其余配置要求相似。由此可见当前区块链技术对通用计算机的适配良好。目前也有多个区块链平台已经投入对嵌入式平台和轻量化设备市场，如嵌入式区块链平台tendermint和轻量级区块链Mina。其中Tendermint实现了拜占庭容错，任何正常工作的机器都会收到相同的交易日志，并分别推导出相同的状态，可以在不超过1/3的机器失效时依然正常工作。Mina是一个轻量级的区块链，区块链大小可维持在22KB左右，这允许节点以低门槛的硬件条件参与，哪怕是运算能力相对较弱的移动端，类似手机、平板电脑等，也可以同步验证Mina网络。

在软件方面，现有区块链技术对软件的要求比较容易实现。以比特币为例，部署所必需的依赖库只有libssl、libboost、libevent三项，其中libssl用于加密及随机数生成，椭圆曲线加密算法。libboost为工具库、线程库，用于数据结构等。libevent为网络库，用于独立于OS的异步网络。其余区块链技术的部署要求也类似。

(5) 与同类技术的关系/对比分析

(a) 与中心化系统对比

中心化系统是指中央控制器通过直接命令或使用权力等级制度来对系统的底层构成进行控制的一种系统。此类系统往往采用单根树状结构，该种方案部署简单，但带来服务节点权限不对等问题，可能导致服务被非法控制，并且面临海量数据时存在单点负载过重、服务拥塞等问题。

(b) 与分布式哈希技术对比

分布式哈希技术(DHT)是一种不需要中心服务器的分布式存储方法，通过某种协议将数据分散地存储在多个节点上，可有效解决集中式架构单一故障带来的服务瘫痪，同时通过散列运算进行存储查询。DHT技术拥有对等、无中心等特点，但不具备区块链防篡改、安全等优势，因此，目前该技术常与区块链技术相结合使用，通过DHT技术构建区块链的链下存储系统，从而提升系统的整体运行效率。

（六）工业数据流通面临的问题

1. 非对称密码算法SM2/SM9/RSA

（1）定义

非对称密码算法或公钥密码算法是指运用陷门单向函数原理编制的加密密钥公开、解密密钥保密的密码算法。RSA是一种国际商用公钥密码算法，被广泛用于安全数据的传输，算法安全性由大整数分解的困难性来保证。SM2是国密局发布的椭圆曲线公钥密码算法，算法安全性由离散对数困难问题来保证。其在我国商用密码体系中被用来替换RSA算法。SM9标识密码算法是国密局发布的一种基于身份标识的密码算法(Identity-BasedCryptography，IBC)，IBC算法以用户的身份标识作为公钥。

（2）作用机理

因篇幅原因且技术较为成熟，本小节略。

（3）所解决的问题

公钥加解密算法确保了存储和交换时数据的机密性：采用RSA、SM2、SM9等公钥加解密算法，可以加密数据，防止数据在存储时以及传输过程中被非法窃听和理解，解决数据的机密性问题。公钥加解密算法加密对称密钥，用对称加解密算法对数据进行加密，同样能防止数据在存储时以及传输过程中被非法窃听和理解，解决数据的机密性问题，运用得当时加解密效率更高。

公钥签名验签算法确保了存储和交换时数据的完整性：采用RSA、SM2、SM9等公钥签名算法，可以对数据进行签名，由数据接收者或其他人对数据及签名进行公钥验签测试，解决存储时及交换过程中数据的完整性问题，解决了攻击者对消息进行篡改的问题，确保收到的消息与发出的消息保持一致，保证消息未被复制、插入、修改、更改顺序或重放。采用公钥签名验签算法与哈

希算法相结合的方式，可以给存储和传输过程中的不定长数据生成数据摘要，由公钥签名算法对数据摘要进行签名，使得其他人可以使用公钥验证算法进行数据、数据摘要和签名的验签，数据完整性验证的效率更高。

公钥签名验签算法确保了存储和交换时数据的不可否认性：RSA、SM2、SM9等公钥签名验签算法，使用数据发送者的私钥对数据进行签名，其他人可以用其公钥进行验签，确认是否来自预期的发送者以及确认数据的来源是否可信。在可能涉及法律责任认定的应用中，公钥签名技术提供数据原发证据，实现数据原发行为的不可否认性。消息进行篡改的问题，确保收到的消息与发出的消息保持一致，保证消息未被复制、插入、修改、更改顺序或重放。采用公钥签名验签名验签算法与哈希算法相结合的方式，可以给存储和传输过程中的不定长数据生成数据摘要，由公钥签名算法对数据摘要进行签名，使得其他人可以使用公钥验证算法进行数据、数据摘要和签名的验签，数据完整性验证的效率更高。

公钥签名验签算法确保了存储和交换时数据的不可否认性：RSA、SM2、SM9等公钥签名验签算法，使用数据发送者的私钥对数据进行签名，其他人可以用其公钥进行验签，确认是否来自于预期的发送者以及确认数据的来源是否可信。在可能涉及法律责任认定的应用中，公钥签名技术提供数据原发证据，实现数据原发行为的不可否认性。

(4) 同类技术对比

RSA是目前最有影响力和最常用的公钥加密算法，它能够抵抗到目前为止已知的绝大多数密码攻击，已被ISO推荐为公钥数据加密标准。随着密码技术和计算机技术的发展，目前1024位RSA算法已经被证实存在被攻击的风险，美国国家标准技术研究院在2010年要求全面禁用1024位RSA算法，升级到2048位RSA算法。

SM2算法，由国家密码管理局于2010年12月发布，是我国自主设计的公钥密码算法。SM2基于更加安全先进的椭圆曲线密码机制，在国际标准的ECC椭圆曲线密码理论基础上进行改进而来，其加密强度更高，在安全性能、速度性能等方面都优于RSA算法，在我国商用密码体系中被用来替换RSA算法。

为了降低公开密钥系统中密钥和证书管理的复杂性，1984年提出了标识密码(Identity-BasedCryptography)的理念。标识密码将用户的标识(如邮件地址、手机号码、QQ号码等)作为公钥，省略了交换数字证书和公钥过程，使得安全系统变得易于部署和管理，非常适合端对端离线安全通讯、云端数据加密、基于属性加密、基于策略加密的各种场合。2008年标识密码算法正式获得国家密码管理局颁发的商密算法型号：SM9(商密九号算法)，为我国标识密码技术的应用奠定了坚实的基础。

本小节将三种公钥密码技术与另一主流相关技术ElGamal算法进行对比。对比细节如下表所示。

算法名称	RSA	ElGamal	SM2	SM9
算法类别	非对称加密算法(基于大整数分解困难问题)	非对称加密算法(基于离散对数问题)	非对称加密算法(基于椭圆曲线 ECC)	标识加密算法(IBE)
应用领域	数据加密、数字签名	数据加密、数字签名	数据加密、数字签名	端对端离线安全通讯
安全性	截至 2020 年，最低建议至少 2048 位	1024bit 为默认	256 位强度已超过 RSA-2048	强度等同于 RSA-3072
算法结构	基于特殊的可逆模幂运算	基于离散对数困难问题	基于椭圆曲线(ECC)	基于τ-DHI 问题和 τ-Gap-BDHI 问题
计算复杂度	亚指数级	亚指数级	完全指数级	完全指数级
存储空间(密钥长度)	2048-4096bit	1024bit	192-256bit	256bit
密钥生成速度	慢	较慢	较 RSA 算法快百倍以上	较 RSA 算法快百倍以上

表6-1 公钥密码技术对比分析

2. 对称密码技术 SM4 与 AES

(1) 定义

对称密码技术是指，信息发送方将明文在一个密钥的作用下经过特殊对称密码编码技术处理后，使其变成随机复杂的加密密文，而信息接收方收到密文后在同一个密钥的作用下经过特殊对称密码解密技术处理，得到明文的过程。对称密码技术的特点是加密和解密使用相同的密钥。

- **SM4**：SM4算法是2006年我国国家密码管理局公布的国内第一个商用密码算法。SM4算法是一种分组密码算法，其数据分组长度为128位(即16字节，4字)，密钥长度也为128位(即16字节，4字)。其加解密过程采用了32轮迭代机制，每一轮需要一个轮密钥以字节(8位)和字(32位)位单位进行数据处理。

- **AES**：高级加密标准(Advanced Encryption Standard，AES),又称Rijndael加密法，是美国联邦政府在2000年批准采用的一种对称密码算法。它是一种迭代型分组密码算法，分组长度和密钥长度都可变，各自可以独立地使用128、192和256位的密钥来加密和解密分组中的数据。

(2) 作用机理

因篇幅原因且技术较为成熟，本小节略。

(3) 所解决的问题

SM4和AES算法等对称密码算法主要解决了数据的安全传输和安全存储问题，主要用于实现数据信息的机密性。无论是SM4、AES算法还是其他的分组对称密码算法，他们从以下几个方面解决了数据的机密性问题：

明确了数据分组问题。明文的长度是不固定的，需要设计如何分组适合当前软硬件的移位、置换、异或操作。当明文长度不是分组长度的整数倍时，是否需要填充且如何填充等问题。这些都是确保明文加密后能解密出一模一样的明文，需要解决设计的问题。

设计了数据分组的衔接问题。分组之后的各个块以何种形式组织起来实现整体的加解密，这里蕴含着安全性和加解密效率的设计。

数据机密性保护问题。这些算法设计了如何通过移位、置换、异或等操作实现基于对称密钥针对每个块的加解密处理，这个是不同的分组加密算法的核心部分。

(4) 同类技术对比

SM4和AES是分组算法，都属于对称密码体制，其加密与解密密钥相同，主要用于大量数据的保密传输。

1977年，DES成为美国政府的商用加密标准，并授权在非密级政府通信中使用，随后该算法在国际上得到广泛使用。但该算法的56bit密钥太短，已不适合用于当今分布式开放网络对数据加密安全性的要求。AES算法在此阶段应运而生，并最终成为取代DES的新一代数据加密标准。在DES向AES过渡的过程中，NIST将3DES指定为过渡的加密标准，3DES是DES的一个安全变形，通过执行3次DES达到增加密钥长度和安全性的目的。2012年3月21日，我国国家密码管理局发布了SM4算法。SM4、AES与3DES算法整体特性如下表所示。

SM4的安全强度和计算效率介于AES与3DES之间。考虑3DES算法安全性较低，且现有的应用系统正逐步用AES替代3DES。

在计算量方面，SM4与AES-128算法的计算量差别较大。在安全性方面，SM4的安全强度等同于AES-128，但是近年来一些密码分析表明，SM4的安全性略弱于AES-128。由于SM4的密钥长度固定为128bit，没有提供更长的可选密钥长度，在安全等级要求越来越高的情况下，SM4可能面临应用范围受限的问题。

	SM4	AES	3DES
分组长度	128bit	128bit	64bit
密钥长度	128bit	128/182/256bit	112/168bit
迭代论数	32	10/12/14	8
算法结构	非平衡 Feistel 网络	SP（代换-置换）网络	平衡 Feistel 网络
算法特性	基本轮函数加迭代，含非线性变换	排列、置换加迭代，含非线性变换	使用标准的算术和逻辑运算，先替代后置换，不含非线性变换

表6-2 SM4、AES、3DES间的比较

3. 密码散列函数SM3和SHA算法

（1）定义

密码散列函数，是用于将一个长度不定的数字消息映射对应到固定长度的字符串(又称消息摘要)的算法，应具备的特点是，如果输入的消息不同，那么应该映射对应到不同字符串。

SM3是中华人民共和国政府采用的一种密码散列函数标准，能计算出一个数字消息所对应到的，长度固定的字符串(又称消息摘要)的算法。由国家密码管理局于2010年12月17日发布。相关标准为“GM / T0004 - 2012《SM3密码杂凑算法》”。

SHA即安全散列算法(SecureHashAlgorithm，SHA)，是一个密码散列函数家族，是由美国国家安全局(NSA)所设计，并由美国国家标准与技术研究院(NIST)发布的安全散列算法。SHA算法包括多个不同长度的算法，分别是SHA-1、SHA-224、SHA-256、SHA-384和SHA-512。

（2）作用机理

因篇幅原因且技术较为成熟，本小节略。

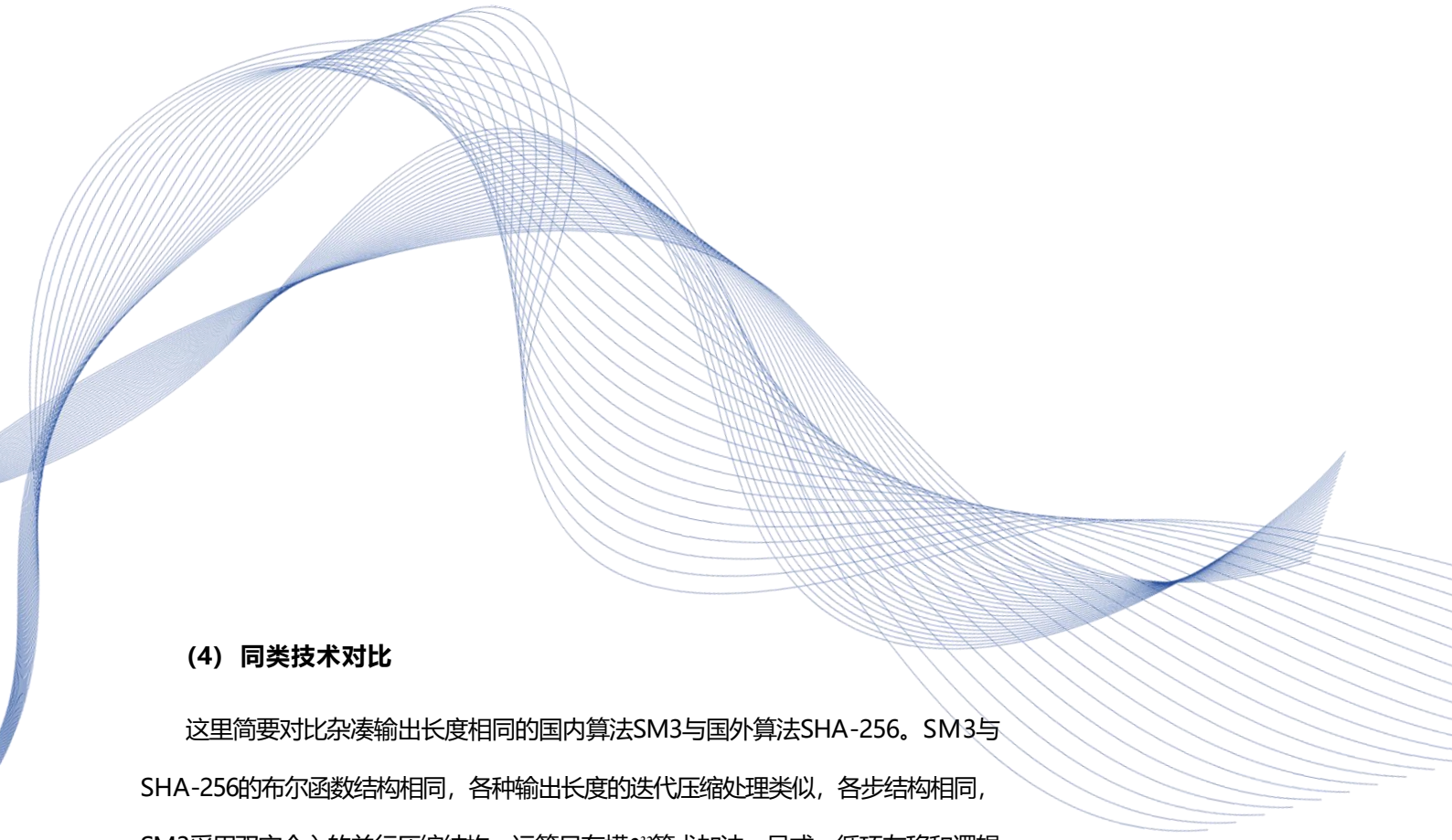
（3）所解决的问题

密码散列函数通过将不定长的数据映射为定长的摘要，能解决和支撑解决以下问题：

密码散列函数可实现数据的完整性校验。通过结合对称密钥、或者结合非对称签名算法，对数据进行密码散列函数计算，形成完整性验证码或完整性签名，可检测数据中的细微的篡改，接收者可以确认自己所收到的数据与发送者所提供的数据是否一致。

密码散列函数可实现口令的快速验证。将用户名口令等数据进行散列函数计算后，与系统预留的散列值对比，可完成用户名口令的快速验证。将口令和伪随机数生成器产生的随机值混合后计算其散列值，然后将这个散列值用作口令的校验，通过此方法能够从某种程度抵御针对口令的字典攻击。

单向散列函数可以构造伪随机数生成器。由于输入数据的细微变化就能引起摘要的大幅度变化，让单向散列函数的输出具有不可预测性，可以利用单向散列函数生成伪随机数。



（4）同类技术对比

这里简要对比杂凑输出长度相同的国内算法SM3与国外算法SHA-256。SM3与SHA-256的布尔函数结构相同，各种输出长度的迭代压缩处理类似，各步结构相同，SM3采用双字介入的并行压缩结构，运算只有模2^{32}算术加法、异或、循环左移和逻辑与、或、非等基本运算，适于ASIC实现，构成的四级加法链流水线有利于硬件的优化实现。据测试，在智能卡中实现SM3算法，算法程序占用RAM共112Bytes，占用ROM共828Bytes。执行27567个机器周期。同等条件下资源占用远远少于SHA-256，运算速度大大高于SHA-256。

4. 双/多因素认证

（1）定义

身份认证是指在信息交互过程中，参与者的其中一方对另一方或多方的身份进行判定及确认的过程。证明或验证实体身份的数据元素称为认证因素，为增强认证安全强度，同时采用两种/多种认证因素的身份认证方案则称为双/多因素认证。

（2）作用机理

在身份认证中，需证实自己身份的一方称为示证者(Prover)，另一方为验证者(Verifier)。验证身份的一般方法为：在接收到来自示证者的含认证因素的认证消息后，验证者将经公式和算法运算所得到的结果与从存储中读取的信息经公式和算法运算所得结果进行比较，根据比较结果得出身份判定结论。身份证明主要依靠以下三类认证因素之一或组合实现：

1) 所知(Knowledge)：该身份所掌握的知识或信息，如口令字、私钥等。

2) 所有(Possesses)：该身份所具有的实体，如身份证、Smartcard等。

3) 个人特征(Characteristics)：该身份所具有的特性，如人的生物特征或行为特征、设备指纹等。

根据系统的安全需求、安全水平、用户可接受性、成本等因素，可以选择适当的组合设计实现双/多因素认证方案。以FIDO(FastIdentityOnline)联盟提出的通用身份认证框架(UniversalAuthenticationFramework,UAF)为例(如图2-37所示)。FIDO在注册时为用户产生密钥对，私钥通过用户PIN码或生物特征数据加密存储在FIDO认证器中(如手机、手提电脑等)，对应的公钥传输到认证服务器保存。认证过程如图6-1所示，用户端通过输入生物特征或PIN码验证后解密获得私钥，用私钥对认证服务器发来的随机挑战进行签名并返回给服务器，服务器通过验证签名的正确性完成对用户的认证。在该认证协议中，结合使用了用户口令字(PIN码)/生物特征与私钥两种认证因素，且两种因素在使用上是具有关联性的，并非简单地依次叠加。一般情况下，先后完成多种独立的单因素认证方案不被认为是多因素认证，必须要多种认证因素信息有所关联地验证才能有效地提升认证方案的安全性。

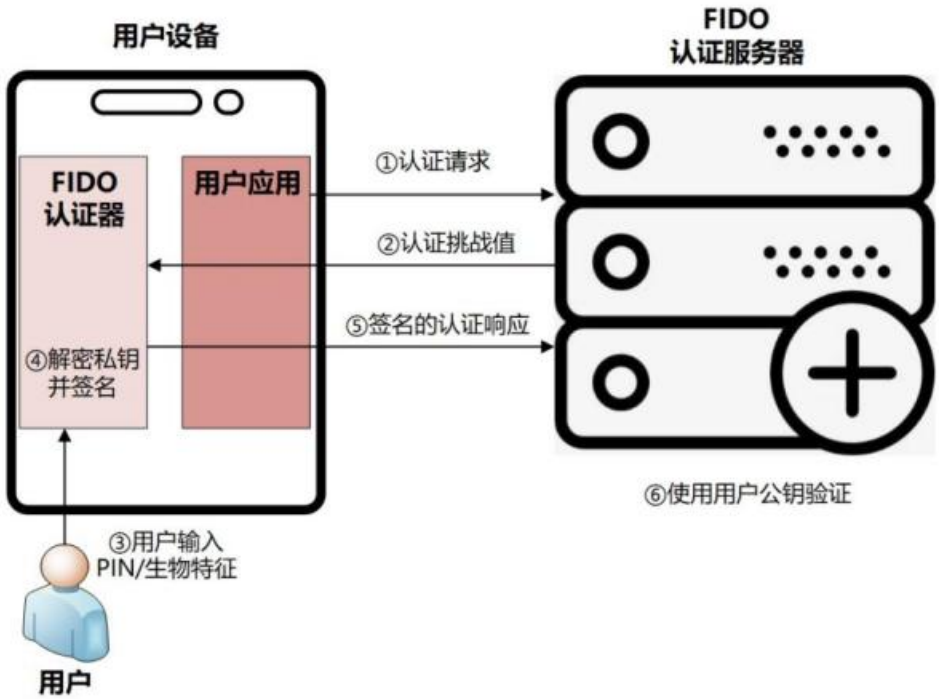


图6-1FIDOUAF双因素认证流程

（3）所解决的问题

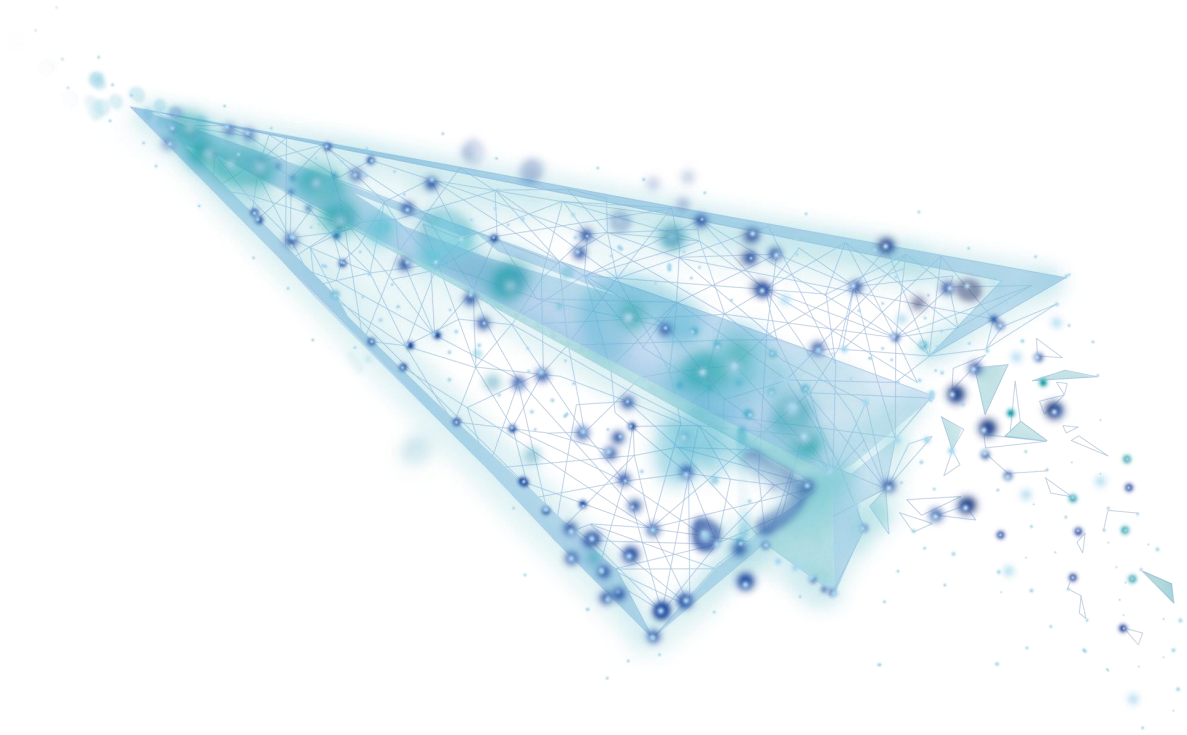
从用户安全性角度分析，非法用户通常采用以下手段进行攻击：

①窃取口令：非法用户获得合法用户身份的口令，可能在未授权下直接访问系统数据或资源；②流量分析：非法用户对公开信道上传输的消息进行分析，试图获取或还原有用数据；③重传/修改/伪造合法用户消息；④阻断服务：破坏系统资源的合法管理和使用。

双/多因素认证可针对上述问题为工业信息系统提供以下安全功能：

①数据/消息源的可信性，即信息的来源是可信的，系统所获得的信息是不是由冒充者发出。②不可抵赖性，即可以确认信息来源是系统中已注册的某唯一用户，该用户不可否认他在系统内的操作，同样，系统也不能否认已收到消息。③访问控制，即非法用户不能访问系统资源，合法用户只能访问系统授权的指定资源。④抗协议攻击，在多因子认证方案中，即使用户某认证因素泄露，如口令字被窃取，但因认证需要多重因素才能成功，极大地提高了安全强度。此外，通过随机挑战、一次性口令字、短信验证码认证等认证方法的加入，可以有效防止消息的重传、篡改等情况。

通过对用户的认证及身份管理，可对用户进行监管，以减少恶意用户行为。单一因素认证方案在细粒度管理的工业数据系统中在安全性及灵活性上远低于双/多因素认证方案，因此在实际应用中，多对用户采用多因素认证方案。



（4） 同类技术对比

主流的单因素认证主要为静态口令字、智能卡、UKey、生物(行为)特征、动态口令认证等，各方案认证原理如下：

1) 静态口令字。口令字(一般称作密码)是由用户自己设定的一串静态数据，系统通过哈希等算法比对口令字的运算值判定是否为合法用户。口令字虽然使用及部署简单，但容易遭受容易字典攻击、窃取、监听、重放攻击、木马攻击等。

2) 智能卡。智能卡芯片具有很高的集成度，其私钥不可读且在卡内进行签名及验证操作，可以有效防止硬件克隆，而且能使解密者对软件端代码的跟踪、调试、侦听数据的手段失效，其支持基于数字签名的认证方案，但只有已通过国际安全机构检测和认证(EAL4+)的专业安全芯片制造商才能提供智能卡芯片。

3) UKey。UKey是集智能卡与读卡器于一体的USB设备，支持热插热拔和即插即用，作为密钥存储器，自身硬件结构决定了用户只能通过厂商编程接口访问数据，这就保证了保存在UKey中的数字证书无法被复制，其基于挑战-应答的认证并且每一个USBKey都带有PIN码保护，因此其为一种双因素认证方案。

4) 生物特征。生物特征认证是指通过自动化技术利用人体的生理特征和(或)行为特征进行身份鉴定。目前利用生理特征进行生物识别的主要方法有：指纹识别、虹膜识别、手掌识别、视网膜识别和脸相识别；利用行为特征进行识别的主要方法有：声音识别、笔迹识别和击键识别等。一个典型的生物特征识别系统包括生物特征识别传感器、特征提取、匹配其和系统数据库四个模块，以及采集生物特征样本、预处理、特征提取和特征匹配四个处理过程。

5)动态口令认证。又称一次性口令认证，主要针对口令窃取/窃听攻击。这类方案的主要设计思路为在口令字中添加不确定因素，通过某种运算(通常Hash函数等单向函数)使每次登录时口令字都不相同，以此增强系统认证的安全性。根据不确定因素，即动态因素的不同，主要有三种动态口令认证机制：①挑战/响应机制；②时间同步机制；③事件同步机制。动态因素由两部分构成，一部分是用户信息、终端信息及共享密钥等固定信息，另一部分为时间、计数器及Hash链等动态信息。各项技术对比分析如表6-3所示。

方案	易用性	安全性	成本	备注
静态口令字	高	低	低	安全性低，维护密码困难
智能卡	中	中	中	需要智能卡及配套读卡器
Ukey	中	高	高	1、采用通用的 USB 接口 2、需要提供独立的 UKey 硬件，增加了少量应用成本
生物特征	低	中	高	1、需要额外生物特征提取设备 2、生物识别技术是单向认证技术
动态口令	高	中	中	1、需要额外部署动态口令生成系统，一般需要联网使用或专有硬件 2、缺少数字签名技术，无法保证不可抵赖性
双/多因子	中	高	依据因素	依据因素，可动态调整认证强度、功能及可用性

表6-3单因素认证对比

5. 数字证书认证技术

（1）定义

数字证书是网络中标识实体身份信息的电子化文件，由权威公正的第三方机构，即CA(CertificateAuthority)中心签发，基于数字证书的认证方案称为数字证书认证技术。

（2）作用机理

数字证书的结构在Satyam标准中定义。国际电信联盟(ITU)于1999年推出X.509标准，是国际数字证书的管理标准，其对应的ISO标准是ISO/IEC9594-8。X.509标准的公钥证书共有V1、V2和V3三个版本。V3格式在V2的基础上添加了一些扩展字段，特殊的扩展字段类型可以由任何组织或者社区定义和注册。X.509V3是目前最广泛采用的证书格式。

X.509描述了两个级别的认证：简单认证和强认证。

简单认证也称为弱认证，是指基于使用用户名和口令的方式来验证用户身份，目前很多应用中仍使用这种简单鉴别。所谓强认证就是利用公钥密码体制实现的认证，它是基于PKI/CA对其用户签发证书证明用户的身份，用户在验证过程中使用私钥对特定信息签名，任何人可通过证书获取用户公钥，利用公钥验证用户签名，从而达到确认身份的目的。

X.509又将强认证分为单向认证、双向认证和三向认证三种认证形式，以适应不同的应用环境。这三种认证过程都使用公钥签名技术，并假定参与各方都可从目录服务器获取对方的公钥证书，或对方最初发来的消息中包括公钥证书，即假定参与方都知道对方的公钥。三种认证过程如6-2所示。

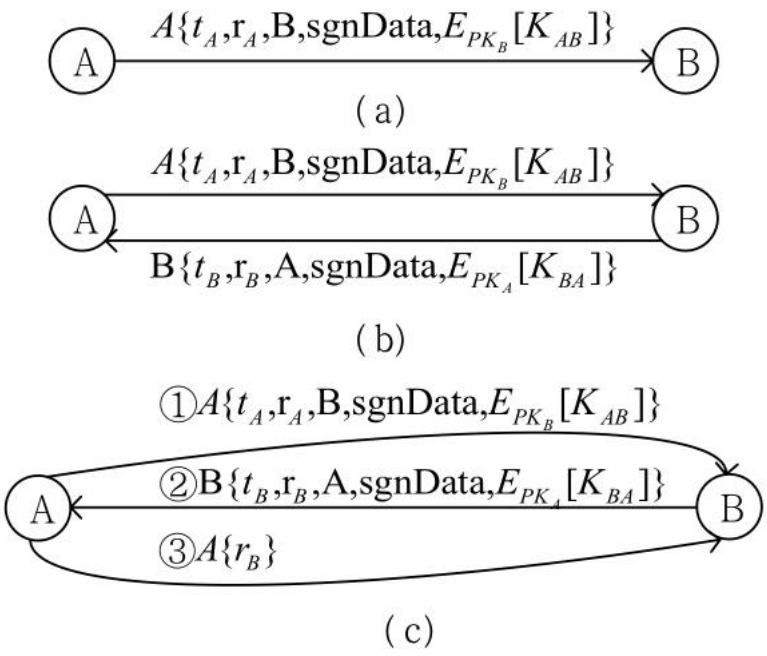


图6-2 X.509的认证过程
(a)单向认证；(b)双向认证；(c)三向认证

1) 单向认证

单向认证指用户A将消息发往用户B，以向用户B表明A的身份，消息是由A产生的。这个鉴别过程需要验证信息的发送方A的身份。消息的接收者是B，B的身份不需要进行验证，同时，必须要保证消息的完整性。

为实现单向认证，A发往B的消息应是由A的秘密密钥签署的若干数据项组成。数据项中应至少包括时间戳t_A、一次性随机数r_A、B的身份，其中时间戳又有消息的产生时间(可选项)和截止时间，以处理消息传送过程中可能出现的延迟，一次性随机数用以防止重放攻击。r_A在该消息到截止时间以前应该是这一消息唯一所有的，因此B可在这一消息的截止时间以前，一直存有r_A，以拒绝具有相同r_A的其他消息。

如果仅单纯为了认证，则A发往B的上述消息就可作为A提交给B的凭证。如果不只是为了认证，则A用自己的秘密密钥签署的数据项还可包括其他附加信息sgnData，对信息进行签名时也会把该信息包含在内，以保证该信息的真实性和完整性。此外，数据项中还可包括一个双方意欲建立的会话密钥K_{AB}(这个会话密钥需

要使用B的公开密钥加密保护）。图2-3中A{}表示以A的私钥对{}中的数据计算数字签名。

2）双向认证

双向认证是指通信双方A、B需要相互鉴别对方身份。为了完成双向认证，在上述单向认证的基础上，B需要对A发送的消息作出应答，以证明B的身份。应答消息是由B产生的，应答的接收者是A，应答消息必须保证完整性。应答消息中包括由A发来的一次性随机r_A(以使应答消息有效)、由B产生的时间戳t_B和一次性随机数r_B,与单向认证类似，应答消息中也可包括其他附加信息和由A的公开密钥加密的会话密钥。B{}表示以B的私钥对{}中的数据计算数字签名。

3）三向认证

在完成上述的双向认证之后，A再对B发来的一次性随机数签名后发往B，即构成第三向认证。三向认证的目的是双方将收到的对方发来的一次性随机数又都返回给对方，因此双方不需检查时间戳，只需检查对方的一次性随机数即可检查出是否有重放攻击。在通信双方无法建立时钟同步时，就需使用这种方法。

（3）所解决的问题

数字证书为公钥基础设施（PublicKeyInfrastructure，PKI）提供可信的数字身份，是PKI的主要部分。公钥基础设施的目的是从技术上解决网上身份认证、电子信息的完整性和不可抵赖性等安全问题，为信息应用提供可靠的安全服务。PKI是标准化的密钥管理平台，能为所有网络应用透明地提供采用加密和数字签名等密码服务所需要的密钥和证书管理，因此在实际应用中，使用基于PKI的数字证书所产生的可信身份包含身份标识及结合密码机制（如签名、加密算法等），可以提供认证、授权或数字签名验证等服务，实现安全、易用、灵活的统一认证及资源可控管理，PKI体系所包含的证书机构、注册机构、策略管理、密钥与证书管理、密码备份与恢复、撤销系统等功能模块的有机结合，可以提高系统的互操作性和可扩展性。此外，安全应用程序的开发者不必再关心复杂的数学模型和运算，只需直接按照标准使用API接口即可实现相应的安全服务。

通过对用户的认证及身份管理，可对用户进行监管，以减少恶意用户行为。单一因素认证方案在细粒度管理的工业数据系统中在安全性及灵活性上远低于双/多因素认证方案，因此在实际应用中，多对用户采用多因素认证方案。

（4）适配性

目前，我国PKI产品被广泛应用于电子政务、电子商务、电子银行等相关领域，通过实现身份认证与访问控制等功能，工业控制系统在发展中，各管理域之间的数据互联互通、身份互认需求的增多、至上而下的安全建设策略，都对以PKI为核心的基础设施建设及应用的普及提供支撑。工业4.0带来了基于物联网的工业模式，其中对于智能设备的远程控制攻击极大地增高了工控系统的风险，在这些环境中构建的产品在制造过程以及使用和使用过程中均要求具有强身份，PKI的核心是根密钥，它充当设备和服务的安全锚：①使用根密钥，可以验证证书链，以检查谁可以访问系统数据。②根密钥对代表设备身份的证书链进行签名——它控制对网络的访问，防止伪造并确保客户数据的安全性。由PKI生成和保护的身份可以较好的适用于该场景，基于数字证书的认证是工控系统中常用的认证方式。

（5）同类技术对比

为了简化系统中的证书管理问题，提高公钥密码系统的效率，Shamir1984年在美洲的密码学会上首次提出了基于身份的密码技术(Identity-basedCryptography，IBC)。在IBC中，用户的公钥可以根据某个公开的算法由用户的身份如姓名、身份证号码、电话号码、Email地址等直接计算出来，用户与其身份相匹配的私钥d_{ID}由私钥生成器(PrivateKeyGenerator，PKG)按照某种公开的算法产生。IBC使得任意两个用户可以直接通信，不需要交换公钥证书，不必保存公钥证书列表，也不必使用在线的第三方，只需要PKG为每个首次加入系统的用户发行一个与其身份相匹配的私钥即可。

二者的不同之处主要表现在以下几个方面：

(1) 用户密钥生成过程不同。在IBC系统中，用户的公钥就是其被公开的身份信息，或者由身份信息演化得到，用户的私钥可以在用户需要时由PKG使用主密钥计算生成。

(2) 私钥传输方式不同。在IBC系统中，用户的私钥由PKG产生，为保证私钥的秘密性，PKG必须通过安全的通信信道把生成的私钥传送给对应的用户；PKI中的私钥由用户个人生成且个人独有，不在任何场所中传输。

(3) 用户公钥确认方式不同。在基于证书的PKI系统中，用户和其对应的公钥由CA颁发的证书绑定，公钥的表现形式是看似随机的字符串，需要在验证了CA的签名才能被其他用户接受；在IBC系统中，用户公钥就是公开的身份信息，或者由身份演化得到，无需一个权威机构签名。

(4) 公钥撤销方式不同。在PKI系统中，公钥的撤销通常由CA维持的CRL实现；在IBC系统中，目前采用的方式一般是在用户身份字符串后串联一个表示公钥生命周期的字符串，该方式给PKG带来了负担。

(5) 公钥存储不同。在基于证书PKI系统中，需要一个公开目录来存放用户的证书(公钥)。在IBC中，由于公钥由用户的身份得到，无需证书支持，可节省资源。

(6) 密钥对的生命周期不同。在基于证书的PKI系统中，密钥对的生命周期长；在IBC系统中，密钥对的生命周期较短。

(7) 系统参数发布形式不同。在基于证书PKI系统中，所有参数作为公钥的一部分；在IBC系统中，公开参数由公开参数服务方发布。

(8) 可信第三方工作状态不同。在基于证书的PKI系统中，CA必须时刻在线以便能够响应第三方的查询；在IBC系统中，每个PKG只是在系统的建立阶段提供服务。

IBC相应于PKI存在的一个主要问题是当用户私钥泄露后，作为公钥的用户身份无法撤销。因此简单地以用户身份作为公钥在实际应用中存在问题，一般会在身份之后再增加有效期、序号等数据，以支持公钥的更新。这一方式使得IBC本身方便获取公钥的优点被弱化了。

（6）技术成熟度分析

美国是最早提出PKI概念的国家，与PKI相关的绝大部分标准都由美国制定，其PKI技术在世界上处于领先地位，已建成的政府PKI体系为联邦政府、工作机构与商业机构等景下电子数据交换提供信息安全保障。欧洲在PKI建设中强调技术中立、隐私保护等原则。较有影响力的PKI产品Entrust/PKI5.0已能较好的曼珠商业、企业的实际需求。VeriSign提供的PKI服务为Internet的很多软极提供签名认证。

我国的PKI技术从1998年开始起步，2001年PKI技术被列为“十五”863计划信息安全主题重大项目，目前，我国已全面推动PKI技术的研究与应用以及将CA证书应用到真武系统及个人安全邮件等多个方面，目前PKI基于证书的认证标准完善，认证技术较为成熟。

6. 数字签名认证技术

（1）定义

使用公钥密码算法对认证因素数据进行签名以实现实体身份认证的方案称为数字签名认证技术。

（2）作用机理

基于非对称密码技术的挑战-应答机制是一种典型的数字签名认证技术,上节中所述证书认证也为数字签名认证。ISO/IEC将以下协议挑战-应答机制标准化为数字签名认证机制的基本结构，签名算法可以根据实际安全需求进行替换,如替换为聚合签名算法，则可实现实体身份的批量认证。

使用公钥的ISO两次传输单方认证协议：

① $B \rightarrow A$： $R_B \| Text1$；

② $A \rightarrow B$: $Cert_A \| Token_{AB}$。

其中 $Token_{AB} = R_A \| R_B \| ID_B \| Text3 \| 꾌ign_A R\_A \| R\_B \| ID\_B \| Text2$，Cert是被认证者的公钥证书。在收到该消息后，B验证签名，若签名正确，则B认可A的身份。

一种典型的证书认证流程如下：

①客户端发送登录请求至服务器。

②服务器端查询用户的有效性，若该用户有效则生成随机挑战**R**发送至客户端。

③客户端一般要先在本机通过口令字(PIN码)认证获取证书中的私钥文件，并用私钥对**R**的Hash值签名，因为通过Hash函数可以将**R**规范为固定长度。随后用户将签名值发送至服务器端。

④服务器端接收到签名文件后，在数据库中查询该用户的公钥，检验签名的合法性。若签名合法，则将认证成功消息返回至客户端。整体认证流程如图6-3所示。

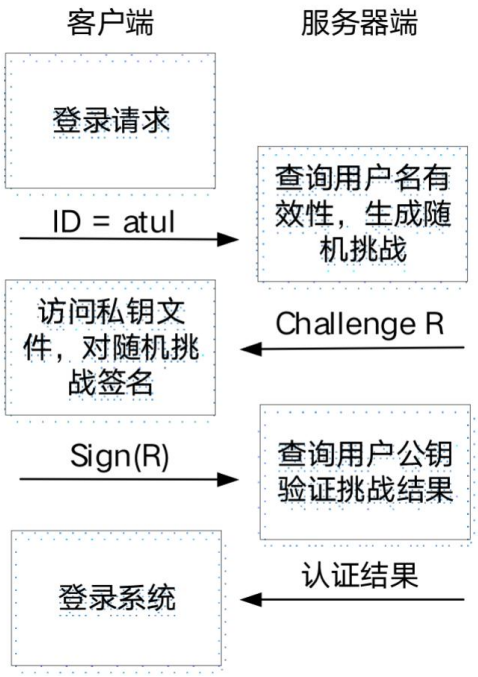


图6-3 证书认证流程

（3） 所解决的问题

数字签名机制作为保障网络信息安全手段之一，可以解决伪造(冒充)、抵赖和篡改问题，其能够抵御的网络攻击如下：

1) 防冒充（伪造）：其他人不能伪造对消息的签名，因为私钥仅由签名者保有，能够找出正确的签名结果数据。进而，接受方利用签名验证可鉴别发送发宣称的身份，接受方使用返送方的公开密钥对签名进行验签运算，若签名有效，则证明对方身份是真实的。

2) 防篡改（保护信息的完整性）：若所签名或被签名的消息在传输过程中遭到篡改（或破坏），则验签无法得到正确结果，从而在公钥正确地前提下可以确认签名或消息在传输中已经被篡改。

3) 防重放：在数字签名中，若在签名报文中添加流水号、时戳等技术，可以防止重放攻击。

4) 防抵赖：被通过验签的签名报文可以确认来自于签名者，可作为追溯证据防止抵赖。

此外，数字签名有大量的不同种类的方案，如聚合签名、盲签名、环签名、组签名等等，基于不同的数字签名方案可以实现安全功能多样的认证协议，如聚合签

名可以实现群组认证，组签名可实现组认证但成员匿名等功能。

（4）适配性

公钥密码体制可以很好的解决1对多的问题，在工控系统中可以对大量设备进行统一的身份管理，使用代理签名等技术还可实现一对多的工业数据安全共享工业控制系统与政务信息系统相同，也是通过网络技术来实现与外界的互联互通，在通信建立前或者某外部设备接入到内部网络时，可根据工业控制系统的安全等级要求，考虑采用基于消息鉴别码机制，或基于公钥密码算法的数字签名机制进行单向或双向身份鉴别，并将证书信息存储于安全介质中，对证书的申请、发放、使用、吊销等过程通过技术手段严格控制，并建立相关制度保障。智能卡也是一种典型的数字签名认证方案，将用户密钥对、CA公钥、数字证书等存储在智能卡上能为用户提供更高级别的安全保障。在认证的同时，复用数字签名机制等密码技术对访问控制信息进行完整性保护，可保证访问控制信息不被恶意篡改，因此在系统部署方面，基于数字签名的认证方案是较为方便的。

（5）同类技术对比

在上节中所述基于PKI及IBC的认证机制均为有证书的数字签名认证协议。经对比分析，PKI证书的管理过程复杂，对离线应用的支持也不好，而IBC具有特点2的无需分发公钥的优势，在用户的标识符管理上就简便很多；IBC的应用场景需要对中心无条件的信任，也就是无条件地相信中心不会滥用私钥，比如组织内部的电子邮件加密系统。PKI在应用中私钥可以由用户端产生，中心也无法获知，因此，互联网上的电子邮件系统更适合选用PKI；IBC无法在保留原身份标识符的前提下实现用户密钥对的撤销或更新，也就是说换密钥就需要换用户标识符。当用户是物品时，一般只需要知道真伪即可，很少有更换密钥的需求；但当用户是人员时，更换人员的身份标识符(比如手机号)是非常不方便的，有时(比如身份证号)甚至是不可行的。因此，IBC更适合物联网场景下对物品身份真伪的鉴别。

同时，为了克服私钥托管问题，密码学家2003年首次提出了“无证书(Certficateless)”公钥密钥的概念，同样不需要证书，但也克服了私钥托管的弊端。无证书公钥中，实体的公钥由两部分组成，除了实体标识符之外，还需要密管中心分

发的一个可公开的个性化参数。其私钥也是两部分组成，一部分必须由密管中心产生，另一部分可以由实体自己产生并保存，密管中心无法获得实体的完整私钥，可以在密钥管理中心是半信任时使用。但其基于双线性对等密码算法，带来了额外的计算开销，在签密方案中，基于PKI及IBC的签名方案仍然为主流。但是随着物联网技术引入工业控制系统，无证书公钥密码的应用日益扩大，成为物联网密码应用的潜在解决方案。

（6）技术成熟度分析

目前基于数字签名的认证技术拥有标准的协议设计模式，但仍需针对数字签名算法的功能进行创新设计，有许多具有特殊用途的数字签名算法被相继提出，如盲签名、群签名、不可否认签名、代理签名、门限签名、失败-停止签名及环签名等，但目前公钥密码体制普遍存在计算消耗大等弱点，在椭圆曲线密码体制提出后有所改善，但提高公钥密码算法的计算效率仍是重点研究方向。替换数字签名算法可以增强认证协议的安全强度及功能性，也赋予了基于数字签名算法的认证协议更大的发展空间。

(七) 日志存证和清算审计相关技术

1. 数据安全审计

(1) 定义:

数据安全审计是基于用户在平台上签署的合约，对用户在可信空间管理终端上操作的存证记录进行审计和核验，以备用户违反合约约定的事件发生后，有效地追查责任。

(2) 作用机理:

数据安全审计技术的实现，可以拆解为以下四个步骤。

A.信息收集

信息收集功能实现对指定用户相关业务的合约信息和日志存证数据的采集汇聚，为数据安全审计提供审计依据。为了更好地支持审计工作，要求数据具有良好的数据结构，数据库管理系统具有高可靠性和高完整性，数据库管理系统要为审计的需要设置相应的特性。

B.数据审计

数据审计支持人工审计和半自动审计和全自动审计等多种审计方式。数据安全审计人员会定期审计用户数据操作，确保用户数据操作都是合法合理合规的。

C.审计报告

数据安全审计的结果填报到审计管理系统中，并支持输出审计报告，审计报告根据审计类型的不同分为手动编写报告、自动生成报告、自动生成+人工改写等多种方式。

审计报告中应明确指出审计中发现的问题，包括审计异常事件类型、安全级别、引发事件的原因、操作的用户、操作时间、操作类型、操作结果、违反的合同条款等。

D.异常处置

数据安全的审计结果需要进行及时的处置，以确保签订合约的有效履行。平台应根据平台管理制度和合约签订的违约条款在规定时间内进行有效处置。处置方式包括在线自动处置和线下人工处置。

线上自动处置：当平台审计有违规行为发生时，按照合约的规定可以自动执行如：自动关闭进程、自动删除文件、弹出告警信息、发送告警短信、通知数据提供方、通知平台管理方等处置措施。

线下人工处置：当平台审计到或接到线上自动发送的告警消息时，需人工介入进行事件调查、责任划分、违约处罚等处置。

（3）所解决的问题：

数据安全审计技术主要解决了数据使用过程中异常行为的发现、调查、分析及事后追查处置的全流程管控，为平台的稳定、合规、可持续运行提供技术保障。

2. 存证与证据交叉认证

存证至少发生在类似这种环节：数据发送方发送数据时，以及数据接收方接收数据时。其目的显而易见：为了证明数据传送行为是真实有效的，防止后期抵赖。因此，存证应具有不可伪造、不可篡改、可回溯等特征，配合审计等手段可为后期纠纷提供有力的法律依据。具体实现时需要结合其他技术(如密码技术、可信执行环境、区块链等)。具体要求如下：

(1) 数据提供方在数据加密后发送的时候，存证数据的生成和数据的加密是原子操作，保证不可能用真实数据生成存证数据而将假数据加密后发送出去；

(2) 数据接收方解密数据的时候，存证数据的生成和结果的解密同样需要是原子操作，保证不可能解密得到真实数据而用假数据存证；

(3) 存证信息需要是公开的，但是存证对应的数据是加密的，而且无法利用存证信息反推原始结果；

(4) 存证信息的校验需要通过公开可用的工具进行，使得存证校验的过程具有透明性。

可以采用本地存证和外部存证两种方式。本地存证由可信计算环境加密后保存在本地，外部存证(输出数据和计算结果的哈希值)可以存储在区块链。外部存证可被用于快速解决争议，最终通过本地存证可完成数据溯源。

任务开始时，部署在数据提供方上的可信计算环境与接收方之间建立连接。可信执行环境处理原始数据从而得到一份混淆后的密文信息，用于后续的验证。接收方收到的只有来自于可信执行环境的信息，因此无法伪造出与发送方不同的输入数据。最后得到的存证信息又可分两部分：经过混淆后的数据信息，以及原始数据的密文。混淆后的数据信息是公开的，而且支持公开验证，而原始数据的密文只有当存在争议时才需要解密。使用可信计算以及区块链技术（同时借助了对称密钥和哈希算法）可以实现前述的四点要求。

存证贯穿数据流通合约的执行全过程：

(1) 在合约生成阶段，合约一旦签署则意味着正式生效，因此合约生效时一般需要进行存证，起到后期防抵赖、可追溯的作用；

(2) 在合约计算阶段，由于涉及大量数据传输，存在多个存证流程节点：数据提供方发送的数据，中间第三方对接收到的数据，中间第三方对传送的结果数据，以及数据使用方接收到结果数据；

(3) 在合约结算阶段，需要对结算结果进行存证。

基于存证信息能够回溯整个合约执行过程，因此在发生纠纷时可对区存证信息进行审计，为法律判决提供依据。

（八）数据增值类服务

1. 多方安全计算

（1）定义

多方安全计算技术是一种密码技术，其源于我国姚期智院士在1982年提出的“百万富翁问题”，并在1986年被姚院士首次证明，其后由Goldreich、Micali和Widgerson等扩展至多方。多方安全计算技术通常采用秘密分享(SecretSharing)、混淆电路(GarbledCircuit)、同态加密(HomomorphicEncryption)等密码算法实现，支撑联合统计、联合建模、隐私集合求交和隐匿查询等功能的实现。

（2）作用机理

数据安全审计技术的实现，可以拆解为以下四个步骤。

多方安全计算的安全性和准确性可进行严格的密码学证明，在多个互不信任的参与方之间进行“密文”形式的协同计算，求通用函数$F(x_1,x_2,\ldots,x_n)$的结果，其中x_i为参与方i私有的输入数据，F为公开函数。执行时可确保：

- 对输入数据的机密性：多方安全计算协议执行过程中，攻击者无法推断出任何有关私有输入数据的信息。

- 对输出计算结果的正确性：诚实参与方不会得到错误的计算结果。

（3）所解决的问题

可信工业数据空间的建设，其主要目的是保证来自多个数据源的数据，在空间中安全的共享及融合应用。多方安全计算可解决其中安全和应用的问题：一是多方安全计算具备信息论上的安全性以及完备的理论依据，常被用于个人信息等敏感数据的保护，可满足空间对数据安全的要求；二是与传统加密技术相比，多方安全计算技术可对计算过程中的“动态”数据进行保护，有利于空间内数据安全流通；三是多方安全计算可支持多方的信息交换和联合计算，但是当参与方数量增加时，计算量呈

爆炸式增长；四是多方安全计算技术可执行各种通用运算，从基本的加法、乘法和比较，到复杂的机器学习算法，可支撑各种工业数据应用。

（4）适配性

一般认为，多方安全计算第一次的成功实践是2009年的“丹麦甜菜拍卖”，已有十余年的工程化落地基础，在金融、医疗等领域，国内外均已有丰富的应用。各技术厂商在私有化部署或云部署，中心化及去中心化部署，专用密码芯片或一体机等，已在项目落地和产品化上开展了多样的尝试，可适配不同的工业数据应用场景。但由于该技术在工业领域的应用还处于初步探索，还需要结合具体场景拓展，并且在算力资源的需求较高。

（5）与同类技术的关系/对比分析

与原有的明文大数据处理相比，由于多方安全计算是基于密码学技术，虽然保障了安全性，但其计算效率相较于明文有较大差距。近年来围绕数据安全共享需求，围绕密码学基础理论、底层协议、分布式计算、系统、算法、编译、芯片以及软硬件结合等方面，科研与技术开发人员进行大量的研究，使得多方安全计算技术的性能耗费已低到明文的10到100倍。

2. 联邦学习

（1）定义

联邦学习是一种训练数据去中心化的机器学习解决方案，2016年由谷歌公司提出，目的在于通过对分布式数据开展训练学习形成一个高质量的机器学习模型，解决集中算力的需求和数据不出域的问题。联邦学习的流程为：在符合条件的计算节点集合中挑选出部分节点，分别从服务器端下载当前的模型；被选择的节点用各自的数据训练模型；各个节点将训练好的模型传输给服务器；服务器将接收到的各个节点的模型聚合成一个新的模型。不断循环以上步骤，直至训练出最终模型。

联邦学习的实质是多节点利用自身拥有数据完成机器学习模型训练的一种分布式架构，合作节点之间交换训练中间结果和模型参数，而不交换数据本身，自然而然减少了数据泄露，联邦学习的中间结果也会泄露数据的部分信息。因此，联邦学习是人工智能领域模型训练的一种计算模式，在可信工业数据流通中仅解决模型训练需要数据汇集的问题，不能解决其他任何数据安全和隐私保护问题。

（2）作用机理

联邦学习按照联合建模的场景需求可分为三类：横向联邦主要用于业态相同或相似的双方间进行样本联合；纵向联邦主要用于业态不同但用户相同或相似的双方进行特征联合；联邦迁移是迁移学习(TransferLearning)在联邦学习框架中的应用，主要用于业态和用户均交集较少的双方间的迁移学习。

由于采用分布式建模的方式，联邦学习训练得到的模型与传统数据集中建模得到的模型相比可能会有一定的性能损失。考虑到实际应用的需求，联邦学习应该保证与传统建模相比模型性能的损失足够小。

（3）所解决的问题

随着人工智能技术的发展，机器学习建模和预测等越来越普遍，在建设可信工业数据空间时，可通过联邦学习相关技术的应用，在保证原始数据不出工业数据供给方本地的前提下，实现共同的机器学习模型训练，支撑逻辑回归、决策树、神经网络等机器学习算法。

（4）适配性

Google采用的横向联邦学习一般基于客户端-服务器(CS)模式，在服务器端融合多个客户端上传的梯度数据。而国内通常采用纵向联邦学习，以直连模式建立两个机构间的点对点连接，共同进行模型训练。并且由于联邦学习主要基于本地的机器学习，因此要求参与联邦学习的机构应该在其本地具备机器学习相关的数据、算法、计算资源，以及部署相应的联邦学习平台。

（5）同类技术对比

联邦学习的安全性建立在几方交换的梯度系信息(或中间计算结果)中，不会暴露原始数据的信任基础之上。相比于传统机器学习算法，联邦学习中各合作节点之间仅交换训练中间结果和模型参数，而不交换数据本身，自然而然减少了数据泄露。然而已有研究证明，可通过联邦学习交换的中间计算结果反推或恢复参与方的原始数据，因此在实际应用中，一般会结合如差分隐私、多方安全计算、可信执行环境等技术，对中间计算结果的融合进行保护。

3. 其他

(1)数据模型

数据的加工处理方式体现了数据的组织方式以及数据上的相关操作。抽象的说，每一种算法逻辑(或者一段计算机程序)都定义了一个明确的数据加工处理方式。数据模型是基于一些行业或领域经验并经过一段时间积淀下来的相对成型的数据组织方式或处理方式，而APP就更加成熟和固化，主要针对特定应用。

在可信数据空间中，这些算法、数据模型、APP通常涉及开发、认证、上架等活动。在开发环节，基于用户需求，通过用户提供的试用数据甚至集成开发环境(IDE)进行开发，保证成品的有效性。开发后一般需要经过专家或检测认证机构的评测、认证，通过后会有证书。最后是申请上架(如在应用商店APPStore上架)，供用户选择使用。

(2)数据计算

无论是哪一种数据加工处理方式，都应该在数据提供方充分授权并与数据使用方达成合约的情况下才能被用于数据计算。按照数据加工处理活动在哪一个参与方上发生，存在以下几种数据计算模式：

(a) 数据提供方进行数据计算。计算后将结果发送给数据使用方。

(b) 数据使用方进行数据计算。数据提供方将数据发送给数据使用方，数据使用方按照合约规定的数据加工处理方式进行数据计算并获得结果。计算结束后删除原始数据。

(c) 数据提供方和数据使用方同时参与计算。

(d) 委托第三方进行代理计算。一般发生在数据提供方、数据使用方没有算力基础、计算效率过低，或者多方数据融合计算的场景(如云计算场景)。

在整个数据空间中每时每刻都可能存在大量数据流通，数据计算存在并行现象。因此需要对所有计算任务进行统一调度管理。每次计算完成代表数据流通通过了关键环节，相应合约中的数据可用次数等内容应被更新。

（3） 数据标识服务

数据标识服务主要包括标识分配和标识解析两大服务类型，一般涉及标识服务方、数据提供方、最终用户三类角色。

标识分配是为数据提供方的数据资产分配唯一标记的过程。一般存在两种方式:

(a)数据提供方向标识服务方提供数据资产描述信息，标识服务方为此数据资产分配一个唯一ID，确定绑定关系并返回给数据提供方；

(b)标识服务方为数据提供方预先生成一系列标识集合，数据提供方为自己的批量数据资产逐一分配标识号，并报标识服务方审核确认。

标识解析过程由用户端发起。用户通过某终端设备(一般配置APP)获得现场某产品标识(如通过扫码)，并将解析需求发送给标识服务方。标识服务方直接返回标识对应的数据资产信息，或者向数据提供方请求详细信息后再返回给最终用户。

结语

虽然鉴于安全性、必要性、回报率和技术成熟度等各方面的因素，目前工业数据流通的场景和实际案例仍然不多，但“数据已经成为继土地、劳动力、资本、技术之外的第五大生产要素”，数据要素将成为数字经济时代的核心生产要素，数据要素市场化配置已上升为国家战略，推动工业数据流通是一种可以预测的历史趋势。

我国未来数据流通模式将为多主体、多种流通模式共存的模式。首先，数据交易所、数据经纪商、大数据平台等不同服务主体可灵活接入空间，获得技术架构一致、面向数据全生命周期管理和控制的标准化服务；其次，不同行业和地区可以根据业务和数据敏感性需求，在空间内构建行业数据空间，从而形成地方枢纽、行业枢纽、龙头企业枢纽。可信数据空间或将成为未来数据要素市场的核心组成和数字经济的关键数据基础设施。

欢迎扫码关注
加入可信工业数据空间生态链
共同助力数据要素流通

工业互联网产业联盟
Alliance of Industrial Internet
北京市海淀区花园北路52号
010-62305887
http://www.aii-alliance.org