隐私计算发展综述

闫树　吕艾临

(中国信息通信研究院云计算与大数据研究所,北京 100191)

摘要:数据流通是释放数据价值的关键环节,隐私计算技术为数据流通提供了解决方案。综述了隐私计算的技术原理,列举了隐私计算当前主要应用场景,分析了国内外隐私计算产业的发展情况。在此基础上,总结了隐私计算发展面临的问题,并从技术、产业、应用等视角分析了隐私计算未来的发展趋势。

关键词:隐私计算;发展情况;发展趋势

中图分类号:TP309.7　　**文献标识码**:A

引用格式:闫树,吕艾临. 隐私计算发展综述[J]. 信息通信技术与政策,2021,47(6):1-11.

doi: 10.12267/j.issn.2096-5931.2021.06.001

0　引言

当前,数据已成为比肩石油的基础性关键战略资源,正在颠覆全球社会的发展模式。2020 年 4 月,中共中央、国务院发布《关于构建更加完善的要素市场化配置体制机制的意见》,将数据同土地、劳动力、资本、技术等传统生产要素并列,作为一种新型生产要素参与分配。作为释放要素价值的关键环节,数据资源的开放共享、交换流通成为重要趋势,其需求日益强烈。

随着近年来数据安全事件频发,数据安全威胁日益严峻。既要应用数据,又要保护数据安全。如何兼顾发展和安全,平衡效率和风险,在保障安全的前提下发挥数据价值,是当前面临的重要课题。以多方安全计算(Secure Multi-Party Computation,MPC)、可信执行环境(Trusted Execution Environment,TEE)、联邦学习(Federated Learning,FL)等为代表的隐私计算技术为流通过程中数据的“可用不可见”提供了解决方案,有助于破解数据保护与利用之间的矛盾,已在金融、医疗、政务等领域开始推广应用。

权威机构 Gartner 发布的 2021 年前沿科技战略趋势中[1],将隐私计算(Privacy Preserving Computing)——其称为“隐私增强计算”列为未来几年内科技发展的九大趋势之一。随着各领域关注度的日益提升,隐私计算已成为发展火热的新兴技术,成为商业和资本竞争的热门赛道。本文将从技术原理、应用场景、产业发展现状和问题以及未来发展趋势等对隐私计算进行系统的梳理和展望。

1　隐私计算的技术原理

隐私计算是指在保护数据本身不对外泄露的前提下实现数据分析计算的一类信息技术,包含了数据科学、密码学、人工智能等众多技术体系的交叉融合。

从技术实现原理上看,隐私计算主要分为密码学和可信硬件两大领域。密码学技术目前以多方安全计算为代表;可信硬件领域则主要指可信执行环境;此外,还有基于以上两种技术路径衍生出的联邦学习等相关应用技术。

1.1　多方安全计算

多方安全计算技术的核心思想是设计特殊的加密算法和协议,基于密码学原理实现在无可信第三方的情况下,在多个参与方输入的加密数据之上直接进行计算。

伴随着公钥密码技术的出现,作为密码学领域中的一个重要研究方向,多方安全计算由姚期智[2]等人于 20 世纪 80 年代提出,以交互不可逆的密文数据的方式实现了对数据的安全保护,每个参与方不能得到

其他参与方的任何输入信息,只能得到计算结果。由于密码学原理的复杂性,多方安全计算的性能相对较低且技术开发难度较大,近几年产业界的高度关注使得其性能得以迅速提升,技术可用性得到很大的提升。图 1 为基于多方安全计算的数据流通产品技术架构。

多方安全计算的实现包含多个关键的底层密码学协议或框架,主要包括不经意传输(Oblivious Transfer) 、混淆电路(Garbled Circuit) 、秘密分享(Secret Sharing) 、同态加密(Homomorphic Encryption) 等。

(1) 不经意传输,也称茫然传输,提出了一种在数据传输与交互过程中保护隐私的思路。在不经意传输协议中,数据发送方同时发送多个消息,而接收方仅获取其中之一。发送方无法判断接收方获取了具体哪个消息,接收方也对其他消息的内容一无所知。

(2) 混淆电路,是一种将计算任务转化为布尔电路并对真值表进行加密打乱等混淆操作以保护输入隐私的思路。利用计算机编程将目标函数转化为布尔电路后,对每一个门输出的真值进行加密,参与方之间在互相不掌握对方私有数据的情况下共同完成计算。混淆电路是姚期智院士针对百万富翁问题提出的解决方案,因此又称姚氏电路。

(3) 秘密分享,也称秘密分割或秘密共享,给出了一种分而治之的秘密信息管理方案。秘密分享的原理是将秘密拆分成多个分片(Share) ,每个分片交由不同的参与方管理。只有超过一定门限数量的若干个参与方共同协作才能还原秘密信息,仅通过单一分片无法破解秘密。

(4) 同态加密,是一类实现在基础的加密操作之上直接完成密文数据间运算的加密算法。数据经过同态加密后进行计算得到的结果与用同一方法在明文计算下得到的结果保持一致,即先计算后解密等价于先解密后计算。

站在技术效果的角度,同态加密也是在无可信第三方的情况下,实现了“多个参与方共同完成一个约定函数的计算”,因此可以将同态加密归为多方安全计算的实现方案之一。但在经典的多方安全计算中,两方计算主要采用不经意传输与混淆电路结合的方案,三方及以上的计算则进一步结合了秘密分享,因此也有观点将同态加密视作一套基于密码学理论但独立于多方安全计算的隐私计算技术。

1.2 可信执行环境

可信执行环境的核心思想是构建一个独立于操作系统而存在的可信的、隔离的机密空间,数据计算仅在该安全环境内进行,通过依赖可信硬件来保障其安全。

可信执行环境的概念源于 Open Mobile Terminal Platform(OMTP) 于 2006 年提出的一种保护移动设备上敏感信息安全的双系统解决方案[3],在传统系统运行环境(Rich Execution Environment,REE) 之外,提供

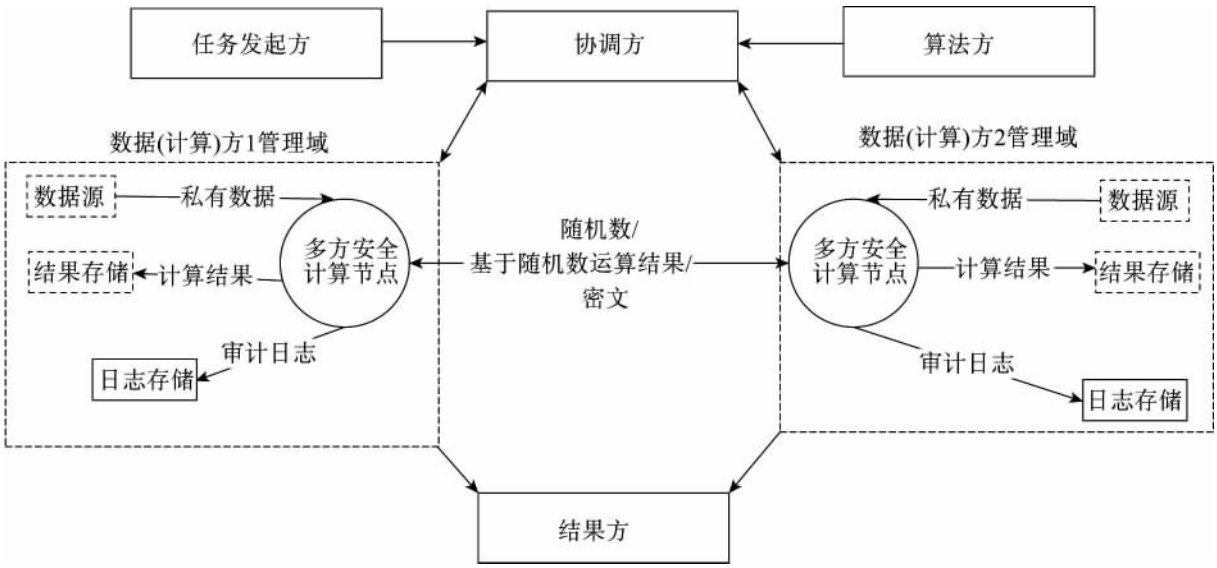

图 1 基于多方安全计算的数据流通产品技术架构

一个隔离的安全系统用于处理敏感数据。2010 年 7 月,Global Platform(致力于安全芯片的跨行业国际标准组织,简称 GP)起草制定了一整套可信执行环境系统的体系标准[4],成为当前许多商业或开源产品定义其各种功能接口的规范参考。

可信执行环境的最本质属性是隔离,通过芯片等硬件技术并与上层软件协同对数据进行保护,且同时保留与系统运行环境之间的算力共享。目前,可信执行环境的代表性硬件产品主要有 Intel 的 SGX、ARM 的 TrustZone 等,由此也诞生了很多基于以上产品的商业化实现方案,如百度 MesaTEE、华为 iTrustee 等。图 2 为基于可信执行环境的数据计算平台技术架构。

严格来讲,可信执行环境并不属于“数据可用不可见”,但其通用性高、开发难度低,在通用计算、复杂算法的实现上更为灵活,使得其在数据保护要求不是特别严苛的场景下仍有很多发挥价值的空间。

1.3 联邦学习

除了前两类隐私计算技术之外,国内外还衍生出了联邦学习、共享学习、知识联邦、联邦智能等一系列旨在解决多方数据联合机器学习的“联邦学习类”技术。

联邦学习的本质是分布式的机器学习,在保证数据隐私安全的基础上,实现共同建模,提升模型的效果。对于基于数据隐私保护的分布式机器学习,早在 2012 年即有学者发表了相关研究成果,直到 2016 年谷歌率先提出联邦学习的概念,才逐步受到更广泛的关注[5]。

联邦学习的目标是在不聚合参与方原始数据的前提下,实现保护终端数据隐私的联合建模。根据数据集的不同类型,联邦学习分为横向联邦学习、纵向联邦学习与联邦迁移学习。图 3 为基于联邦学习的数据流通产品技术架构。

1.3.1 横向联邦学习

横向联邦学习更适用于在特征重合较多,而样本重合较少的数据集间进行联合计算的场景。以样本维度(即横向)对数据集进行切分,以特征相同而样本不完全相同的数据部分为对象进行训练。谷歌在 2016 年提出的安卓手机模型更新数据联合建模方案就是利用单个用户使用安卓手机时,不断在本地更新模型参数并上传到安卓云上,从而使特征维度相同的各数据拥有方联合建模[5]。

1.3.2 纵向联邦学习

纵向联邦学习更适用于样本重合较多,而特征重

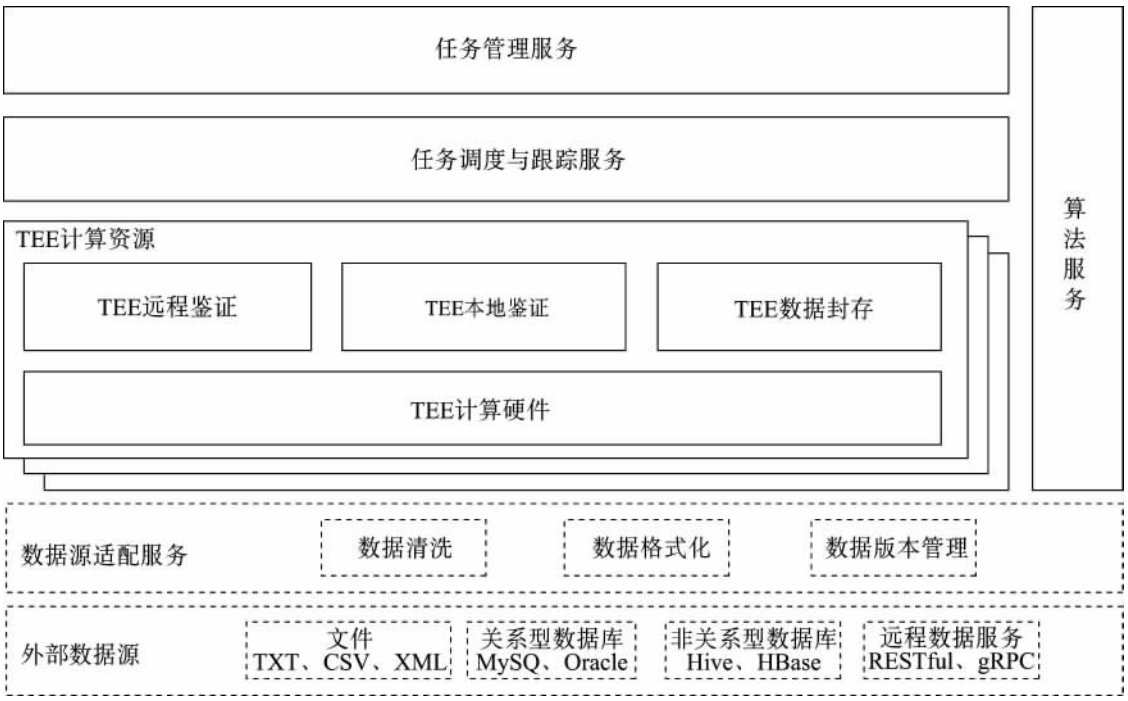

图 2 基于可信执行环境的数据计算平台技术架构

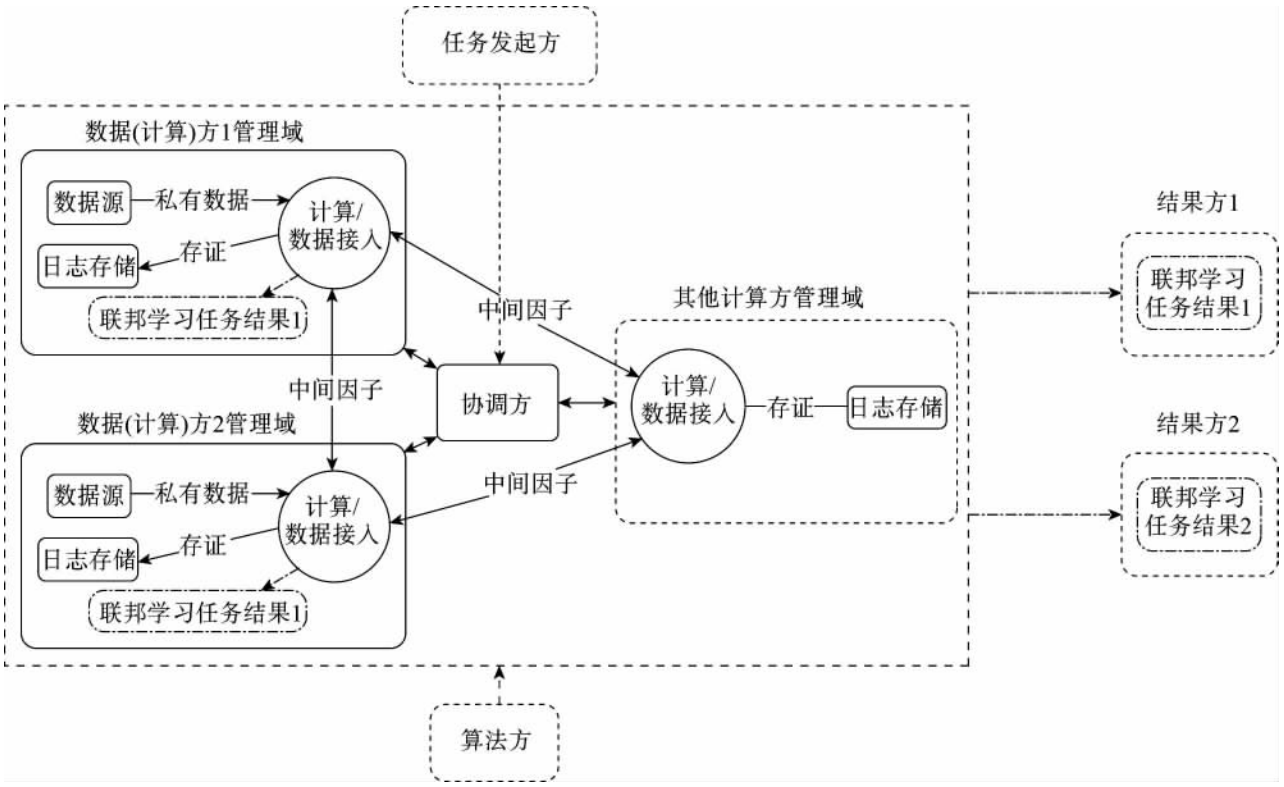

图 3　基于联邦学习的数据流通产品技术架构

合较少的数据集间联合计算的场景。以特征维度(即纵向) 对数据集进行切分,以样本相同而特征不完全相同的数据部分为对象进行训练。以同一地区的银行和电商为例,由于两机构在特定地区的用户群体交集较大,因此可以对两机构的不同维度的用户特征进行聚合以增强模型能力。

1.3.3　联邦迁移学习

联邦迁移学习则适用于数据集间样本和特征重合均较少的场景。在这样的场景中,不再对数据进行切分,而是利用迁移学习来弥补数据或标签的不足。以不同地区、不同行业机构之间进行联合建模为例,用户群体和特征维度的交集都很小,联邦迁移学习即用来针对性解决单边数据规模小、标签样本少的问题。

此外,许多企业推出的共享学习、知识联邦、联邦智能等一系列技术大多以联邦学习为基础进行改进,目的仍然是实现多方数据的机器学习。

1.4　各类技术对比

由于技术路径的不同,各类隐私计算技术均有其更加适用的场景: 多方安全计算技术不依赖硬件且具备较高的安全性,但是仅支持一些相对简单的运算逻辑; 可信执行环境技术具备更好的性能和算法适用性,但是对硬件有一定依赖; 联邦学习技术则可以解决复杂的算法建模问题,但是性能存在一定瓶颈。相关技术的主要对比如表 1 所示。

整体描述: 开发难度大、关注度高,使得性能提升迅速; 易开发、性能佳,但需信任芯片厂商(Intel、ARM 等) 综合运用各类密码学方法(主要针对机器学习) 。

2　隐私计算的应用场景

近年来,隐私计算技术和应用快速成熟。回顾隐私计算的发展历程,以 MPC 为代表的相关密码学理论早在 20 世纪 80 年代就已经被提出[1],但长期处在实验室研究阶段,并未得到实践应用,而以 TEE 为代表的可信硬件也主要应用于个人(C 端) 移动用户的数据安全保护中。随着机构间数据流通需求的日益旺盛,隐私计算的相关理论得到了产业界的更多关注,在算法协议不断优化、硬件性能逐步增强之下,隐私计算的可用性大大提升,以 MPC 为例,相比明文计算的耗时已经从数十万倍降至百倍之内[8]。自 2018 年开始,隐私计算的技术和产品成熟度迅速提升,在我国加快培

表 1 隐私计算相关技术主要对比

技术	多方安全计算	可信执行环境	联邦学习
安全机制	基于密码学原理 对数据加密	引入可信硬件	数据不动,模型动
性能	低~中	高	高
通用性	高	中	低
高效性	中	中	低
准确性	高	高	中~高
可控性	高	中	高
保密性	高	中~高	中
可信方	不需要	需要	不需要
整体描述	开发难度大、关注度高使得性能提升迅速	易开发、性能佳,但需信任芯片厂商(Intel、ARM 等)	综合运用各类密码学方法,主要针对机器学习

育发展数据要素市场、数据安全流通需求快速迸发的推动下,隐私计算技术的应用场景越来越多。

2.1 金融行业的应用场景

在金融科技深刻赋能业务的进程中,外部数据的共享应用成为金融机构的强烈需求,基于隐私计算的金融风控和获客成为目前国内最主要的隐私计算落地场景。

金融机构与外部数据源的合作过程中存在的风险主要来源于两个方面:一是涉及大量个人用户信息,受到的监管要求严格;二是机构自身业务积累的数据资产和商业秘密容易泄露。而利用隐私计算,金融机构之间、金融机构同运营商、互联网、电商平台之间等可以在不泄露原始信息的前提下对客户进行联合的精准画像,在信贷评估、产品推荐等场景下有效控制违约风险,提高业务效率。

以银行个人信贷业务为例,首先需要对客户进行三要素(姓名、身份证号和手机号)核验查询等以确认用户身份。为避免在面向外部机构进行查询调用过程中客户的敏感信息被缓存,银行可以利用隐私计算实现匿踪查询以保护数据安全。其次,银行还需要引入客户的行为数据和场景数据进行联合建模以准确判断其偿付能力和违约风险。利用隐私计算,参与建模的金融机构和外部数据合作方可以在不直接交互原始数据的前提下,实现多方数据的虚拟融合和样本对齐,各自在本地进行算法训练,仅对任务的中间因子进行安全交互,在敏感数据不出门的同时完成对用户的画像。

2.2 医疗行业的应用场景

医疗行业的数据共享与流通需求成为隐私计算的另一个关注重点。医学研究、临床诊断、医疗服务等对基于大数据的统计分析与应用挖掘有着强烈的需求,但其依赖的是众多病患的个人健康数据,这些数据规模大、价值含量高,但共享流通却十分困难。

一方面是跨机构的数据采集与整合难。相似疾病的不同病例、同一病患的不同病案等大量的诊疗数据往往分布在不同的医疗机构,各机构数据开放共享的意愿本就有限,再加上各医疗机构间的数据标准、编码方式各不相同,跨机构间的联合研究与诊断就更加困难。另一方面是跨机构的数据联合应用难。病患的个人医疗数据十分敏感,且复杂的诊疗数据在其使用过程中难以管控,面对个人隐私保护和数据安全要求,很多机构难免望而却步。

隐私计算为以上难点提供了解决思路。利用隐私计算,在建立分散存储的标准化数据库的基础上,可以实现分布式的联合统计分析,从而获得临床科研的研究成果。在抗击新冠肺炎疫情的过程中,隐私计算助力实现了全球范围内的疫情数据共享,基于多方安全计算等技术实现了允许用户在不公布己方数据的前提下,联合其他科研人员协同进行病例样本基因组的联

合分析并共享结果,实现了对病毒流行病学情况的实时追踪和对未来毒株演化的预测,成为助力抗疫情的一把利剑。

2.3 政务行业的应用场景

作为跨机构间数据流通的重要参与主体,政务行业有望成为隐私计算技术落地的下一个重要场景。政务数据的规模大、种类多、蕴含价值高,涉及公安、交通、税务、环境等各类人民生产生活和社会运行的数据,政务数据的流通与应用将释放巨大能量。

近年来,各地政府积极推进政务数据的开放共享,但不同部门之间的数据孤岛难以快速消除,且政务数据涉及社会民生,数据合规和安全管控要求更加严格。因此,政务部门之间、政府与企业之间的数据共享应用十分困难。隐私计算为此提供了解决方案,在跨机构之间的个人身份确认、企业经营监管、智慧城市建设等众多场景中均有广阔的应用前景。在部分地方政府的相关规划里,已经有所涉及。

2.4 隐私计算应用现状

信任机制是隐私计算广泛应用的关键,隐私计算技术自诞生以来的重要使命便是保证隐私数据在被利用的过程中不被泄露。作为一项新技术,如何自证安全、持续强化安全、建立市场信任是其能否被广泛应用的关键。

(1) 自证安全是隐私计算应用的瓶颈问题,当前隐私计算应用主要通过深入介绍产品保密算法机制、签订严格保密协议和引入第三方评测机构评测产品来实现。

(2) 持续强化安全是隐私计算应用的长效保障,目前主要通过不断优化算法来防范恶意攻击,以更加严格地控制计算流程来封堵漏洞等方式实现。

(3) 建立市场信任是隐私计算应用的关键问题,在隐私计算的过程中,通过严格的数据授权、身份验证、状态监控预警等方式让数据提供方始终清楚己方数据的用量、用法、用途均不超出事先约定,可以充分建立用户信任乃至市场信任,但当前的应用在这一部分工作仍需加强。

3 隐私计算的产业现状

3.1 产业政策环境不断完善

近年来,我国数据立法进程不断加快,尤其强调数据应用过程中的数据安全[5]。早在 2016 年 11 月的《中华人民共和国网络安全法》中,即强调了数据隐私的重要性。2020 年 7 月,《中华人民共和国数据安全法(草案)》公布,明确了开展数据活动的组织、个人的数据安全保护义务,落实了数据安全保护责任规定支持促进的具体措施。2020 年 10 月,《中华人民共和国个人信息保护法(草案)》公布,进一步强调了个人信息在数据流通过程中的安全合规。

隐私计算是平衡数据利用与安全的重要路径,工业和信息化部、中国人民银行等各部委先后出台相应的政策支持技术发展。工业和信息化部先后在 2016 年发布的《大数据产业发展规划(2016—2020 年)》[12] 和 2019 年发布的《工业大数据发展指导意见(征求意见稿)》[13] 中提出支持企业加强多方安全计算等数据流通关键技术攻关和测试验证,并在工业领域积极推广应用,促进工业数据安全流通。中国人民银行在 2019 年 9 月颁布《金融科技(FinTech)发展规划(2019—2021 年)》[14] 也提出要利用多方安全计算技术提升金融服务安全性。政策的提前布局为我国抢占隐私计算关键领域的竞争优势奠定了重要基础。

3.2 技术产品市场发展迅速

2019 年,Gartner 首次将隐私计算列为处于启动期的关键技术[15]。2020 年,Gartner 又将隐私计算列为 2021 年企业机构九大重要战略科技之一,并预测隐私计算将迅速得到落地应用,预计到 2025 年应用范围将覆盖全球一半的大型企业机构[2]。近两年来,伴随着技术的不断成熟,国内外隐私计算产业化的步伐明显加快。可以预见,未来几年将是技术产品加速迭代,应用场景快速升级,产业生态逐步成熟的重要阶段。

3.2.1 国外隐私计算技术研究创新活跃,但商业化进展稍缓

从技术发展的历程来看,谷歌、Intel 等国际领军企业开创了隐私计算产业的时代潮流[6-8]。但从整体发展路径来看,相比国内企业,国际科技企业在学术研究和开源生态的建设上更为活跃;相比之下,商业化的产品形态较为局限,产业生态也尚未形成火热竞争或垄断格局。

微软研究院自 2011 年开始大规模推进多方安全计算的研究,从两方安全计算入手,逐渐拓展至三方计算和不存在交互行为的多方计算。但微软前期的

MPC 研究存在两个瓶颈，一是加密协议只针对一些简单的分析功能有效，如聚类分析、线性回归等；二是计算的执行必须运行在低水平的与或门电路中，执行过程麻烦而低效。2018 年，微软印度研究院推出了 EzPC 项目，希望克服上述两个问题。作为一个高效、可扩展的 MPC 协议，EzPC 是一个加密成本感知编译器，使用算术和布尔电路的组合，通过高级语言执行计算，支持神经网络训练和预测等复杂的算法。

谷歌是联邦学习的引路人，自 2017 年 4 月，谷歌便提出了联邦学习的概念，并于 2019 年发布论文给出了可扩展大规模移动端联邦系统的描述，用于改进谷歌输入法的自动关联与推荐。但与此同时，2019 年 8 月，谷歌又开源了名为 Private Join and Compute 的新型多方安全计算开源库，结合了隐私求交和同态加密两种基本的加密技术，帮助各组织和隐私数据集协同工作，针对个别项目还使用随机密钥进行高度加密，提高隐私性。

Intel 的 SGX 和 ARM 的 TrustZone 处于 TEE 硬件的垄断地位[9,11]。基于 ARM TrustZone 实现的可信执行环境是一种硬件隔离安全机制，以物理方式将系统划分为安全和非安全组件，确保在正常操作下的软件无法直接访问安全区域的数据；而基于 Intel SGX 实现的可信执行环境是一种算力和内存隔离的安全机制，使用特殊指令和软件将应用程序代码放入一个 Enclave 中执行，Enclave 可在虚拟机监控器、主操作系统和驱动程序均被恶意代码攻陷的情况下，仍然对其内的代码和内存数据提供安全保护。TrustZone 在 2008 年推出，而 SGX 最早在 2013 年推出，二者都是随着移动手机的大发展而繁荣起来，目前市场上可信执行环境的商业化落地都是基于 TrustZone 或 SGX 的解决方案。

除上述提到的科技巨头外，国外互联网、AI、区块链领域的相关企业和机构也快速布局了隐私计算，Facebook 将基于 PyToych 的隐私计算机器学习框架 CrypTen 进行开源；AI 公司 Zama 开源了基于全同态加密的软件库 Concrete；麻省理工学院创办的区块链公司 Enigma 推出了基于多方安全计算的新加密系统，并与 Intel 合作研发基于 SGX 的可信执行环境；创业公司 Sharemind、Privitar 致力于搭建自研的多方安全计算平台。但从应用场景来看，目前的主要应用大多局限于将 MPC 技术应用于分布式密钥管理领域，如美国的 Unbound Tech 和丹麦的 Sepior。

从技术路径上看，各国际企业相对更关注基于可信执行环境的隐私计算。2019 年成立的 Linus 基金会旗下的机密计算联盟（Confidential Computing Consortium）便聚焦于此，关注基于可信硬件和云服务生态的数据安全。该联盟的创始会员包括阿里、腾讯、ARM、谷歌、Intel、微软、百度、华为等世界级企业，2020 年 AMD、英伟达、埃森哲、R3 等新一批知名企业也陆续加入。

3.2.2 国内隐私计算技术产品蓬勃发展，形成一定优势

我国的隐私计算技术产业化在 2018 年后开始进入快速启动阶段，形成了互联网大厂、大数据公司、运营商、金融机构和金融科技企业、隐私计算初创企业为代表的五大类市场主要参与者。

阿里巴巴、百度、腾讯、京东、蚂蚁等各互联网巨头凭借自己在技术领域的积累，自 2019 年开始纷纷推出了各自的隐私计算产品，形成了跨业务、多团队、强支撑的发展态势，集团内部不同业务根据自身的业务特点和需求，选择一种或多种技术方案融合的方式进行开发；作为大规模数据资源拥有者的电信运营商为拓展业务形态，不仅三家运营商均在集团层面开始了隐私计算技术的选型与应用，天翼支付、电信云等子公司还自建平台服务于内部或其他机构的数据流通业务；金融机构是数据流通与安全应用最主要的需求者，国有银行的研究院或是事业部也均开始了隐私计算技术的研究工作，新心数科、神谱科技、平安科技、百融云创、度小满等金融科技类企业也将传统的数据建模、数据分析等业务拓展到基于联邦学习平台等的隐私计算服务中；同盾科技、星环科技、Talking Data、京信数科等代表性的大数据技术厂商也快速布局基于隐私计算的数据流通产品或平台。

以上企业的商业化路径大多是既要服务于企业自身运营需求，也可作为服务方为政务、金融等领域提供技术支撑。区别于此，如富数科技、华控清交、矩阵元、翼方健数、数牍科技、锘崴科技、光之树科技、零知识科技等一批专注于隐私计算产品化的初创企业也不断涌现。作为促进数据流通的关键技术，在国内大数据产业稳步发展、数据要素市场化配置加快推进的背景下，我国隐私计算技术产品日渐成熟，各领域应用场景加

速落地,产业快速发展。面对隐私计算技术领域的国际竞争,我国已初具竞争优势,有望占据有利地位。

从技术路线上来看,多方安全计算的复杂度高、开发难度大,以华控清交、富数科技、矩阵元等为代表的隐私计算初创企业多致力于此,专注于打造以底层多方安全计算技术为基础的数据流通基础设施;可信执行环境对于硬件的局限及国外芯片的强依赖,使得其在国内的产品选型相对较少,较集中于百度、阿里巴巴等互联网大厂和冲量在线、隔镜科技等初创企业,但目前已出现冲量在线与兆芯在国产化硬件研发上的合作探索;对于联邦学习,由于机器学习类应用需求的突出,且有较成熟的开源社区为基础,开发难度相对轻松,因而运营商、金融科技公司等业务需求方大多专注在基于联邦学习的隐私计算产品化中。

2020 年,为提升行业认知,推进隐私计算技术与应用的融合,在工业和信息化部相关司局的指导和支持下,中国信息通信研究院云计算与大数据研究所牵头成立公益性合作平台“隐私计算联盟”(Privacy Preserving Computing Alliance)[16],目前联盟已有包含技术厂商、政府单位、运营商和金融机构等在内的 50 余家成员单位。

3.2.3 隐私计算的开源生态逐渐显现

开源社区的知识共享和多方协同有利于加快技术升级迭代和商业化项目落地的效率。对比传统的大数据技术工具,开源已成为生态中的绝对主流。作为保障数据合作与安全的重要基础,隐私计算有望进一步拥抱开源。近两年,很多大厂不断提供开源资源,目前国内外科技巨头在隐私计算领域的开源项目情况如表 2 所示[16]。

从目前国内外影响力较强的隐私计算开源项目来看,联邦学习主要有 PySyft、TF-Federated 和 FATE。

PySyft 和 TF-Federated 目前仅支持试验环境。PySyft 是开源社区 OpenMined 开源的隐私计算框架,主要针对实现基于隐私计算的深度学习。PySyft 将联邦学习、多方安全计算以及差分隐私、远程执行等技术结合在一个编程模型中并集成到不同的深度学习框架中,如 PyTorch、Keras 或 TensorFlow;谷歌基于 TensorFlow 开源的 TF-Federated,则主要针对类似谷歌输入法案例的横向联邦学习。

FATE 是国内联邦学习商业化产品的主要贡献力量,由微众银行于 2019 年 2 月开源。FATE 提供了一种基于数据隐私保护的分布式安全计算框架,为机器学习、深度学习和迁移学习算法提供高性能的安全计算支持,支持同态加密、SecretShare 等多种多方安全计算协议,简单易用。目前,社区内已有超 370 家企业、164 所高校合作[16]。

表 2 代表性隐私计算开源项目

序号	项目名	开源时间	机构	技术路径
1	PySyft	2017 年 7 月	OpenMined	多方安全计算、联邦学习
2	TF-Encrypted	2018 年 3 月	DropoutLabs、Openmined、阿里巴巴	多方安全计算
3	Asylo	2018 年 5 月	谷歌	可信执行环境
4	MesaTEE	2018 年 9 月	百度	可信执行环境
5	FATE	2019 年 2 月	微众银行	联邦学习
6	TF-Federated	2019 年 8 月	谷歌	联邦学习
7	Private Join & Compute	2019 年 8 月	谷歌	多方安全计算
8	PaddleFL	2019 年 9 月	百度	联邦学习
9	CrypTen	2019 年 10 月	Facebook	多方安全计算
10	Fedlearner	2020 年 1 月	字节跳动	联邦学习
11	Rosetta	2020 年 8 月	矩阵元	多方安全计算
12	KubeTEE	2020 年 9 月	蚂蚁集团	可信执行环境

3.3 配套标准体系日渐完善

技术体系的发展壮大需要配套标准指引的支撑。《多方安全计算技术框架》和《基于 TEE 的安全计算》两项国际标准分别于 2019 年 4 月和 2020 年 9 月在电气电子工程师学会(IEEE) 立项,但现有标准的内容主要是给出了通用性的技术框架,尚没有深入到应用中的细节。

中国信息通信研究院依托中国通信标准化协会大数据技术标准推进委员会于 2018—2020 年分别牵头制定了《基于多方安全计算的数据流通产品》《基于联邦学习的数据流通产品》《基于可信执行环境的数据计算平台》《区块链辅助的隐私计算技术工具》4 项隐私计算技术产品功能上的系列标准[16]。随着技术的火热发展,这些标准正在快速迭代、不断完善,针对不同产品的性能和安全性标准也正在加速制定中。

与此同时,中国信息通信研究院云计算与大数据研究所还依据已有标准积极开展技术产品的标准化评测,帮助市场建立起对于市场产品的客观评价体系,助力行业行稳致远。自 2019 年下半年开始启动的隐私计算技术产品评测已完成 3 批共 40 余次的产品评测[16]。透过每一批评测中产品数量的快速增长,也可见证国内隐私计算产业的火热发展。

4 隐私计算发展面临的问题

我国隐私计算发展具备一定优势、存在广阔应用空间,但由于技术发展仍不完善,因此也面临着一些问题。

4.1 市场对于隐私计算的认知度、认可度仍然不足

隐私计算技术仍然在蓬勃发展的上升期,技术发展尚未成熟,市场培育也尚未完成。一方面由于隐私计算技术复杂且常常呈现“黑盒化”现象,大部分用户对隐私技术难以理解和信任;另一方面,对技术理解的不够全面,会导致用户对技术应用的效果产生过度预期。

4.2 技术推广所需的成熟商业模式尚未形成

隐私计算的应用必将重构过去数据流通、融合等相关的业务流程,形成新的商业模式。但是,当前市场正处于快速发展的早期阶段,明确的激励机制、利益分配机制和通用的平台收费机制等商业化落地模式尚未形成,难以支撑技术的大规模推广。特别是在政务、医疗等系统相对固化、开放动力不足的行业,推广隐私计算技术的难度很大,在近几年内难以看到大规模应用的情况。

4.3 现有法律法规未对隐私计算地位进行明确定位

《中华人民共和国网络安全法》中规定“未经被收集者同意,网络运营者不得向他人提供个人信息”,同时设置了“经过处理无法识别特定个人且不能复原”的例外条款。隐私计算仅仅避免了原始数据转移的过程,但完成了基于多方数据的计算,所以仍然涉及到个人信息的提供和使用。目前,将个人信息用于隐私计算,以及如何在符合个人信息保护的要求下使用隐私计算技术,现有法律法规及相关标准等并无明确界定。这正在成为制约隐私计算发展的无法回避的问题。

4.4 在满足通用性功能的基础之上,产品性能提升任重道远

目前来看,大多数隐私计算产品已能满足用户的通用性需求,但实现技术大规模商用的瓶颈不在于理论体系的创新,而在于计算成本的控制与计算效率的支撑。无论是工程化实现的创新突破还是软硬件之间的优化适配,隐私计算的性能提升都任重道远。

5 隐私计算未来的发展趋势

5.1 从技术角度来看,隐私计算仍在加速成熟

(1) 软硬件协同优化性能的提升、技术的可用性。硬件加速在隐私计算性能提升方面正在发挥越来越关键的作用,在算法不断优化的基础上,一些专用芯片和控件的使用将进一步提升隐私计算的性能。

(2) 逐步向大规模分布式计算迈进。2020 年以来,隐私计算逐渐成熟的一个表现就是分布式隐私计算的逐渐应用,为解决隐私计算在计算量方面的瓶颈提供了优秀实践。

(3) 提供工具化、模块化的服务能力。如何满足用户的个性化与定制化需求、提升用户使用效率将成为产品形态趋同之下,技术提供者提升竞争力的关键。低代码甚至零代码开发、图形化拉拖拽替代编码和多版本轻量化部署等将成为产品升级优化的关键之一。

5.2 从应用角度来看,隐私计算将加速与其他技术的协同以推进大规模落地应用

(1) 增强与区块链等其他技术的不断协同。区块链与隐私计算的功能是天然互补的,借助区块链去中

心化、不可篡改、公开透明的特性,将增强隐私计算任务的可验证性、可审计性,目前已成为诸多厂商的技术融合方向。此外,隐私计算与云计算的协同,将在支持云端数据存储、处理的同时加强任务过程中的安全与隐私控制;而隐私计算与人工智能的协同,将有力推进数据智能的应用和发展。

(2) 促进跨技术平台间的互联互通。隐私计算的目标在于促进多方数据之间的互联互通,但从应用现状看,不同技术路径之间的差异明显,而同一路径下不同产品的实现方案也相互独立,数据资源的互联互通只能基于不同的技术平台分块实现,无疑增加了应用侧的使用成本。从长期发展来看,跨技术路径、跨系统平台之间的隐私计算技术工具的互联互通将成为广泛需求。

5.3 从产业角度来看,隐私计算还需与数据治理相互配合,共同促进数据要素的共享利用和价值释放

(1) 隐私计算有望成为数据流通的关键基础设施。随着近两年国内隐私计算的技术产品快速落地,越来越多的行业客户开始在数据流通活动中实施部署相应的技术解决方案,隐私计算逐步推动着传统数据流通模式和流程的变革,当技术能力和应用模式越发成熟之时,隐私计算才有望成为全社会数据流通网络的支撑型基础设施。

(2) 技术发展将随着数据合规要求的不断变化而动态演变。在《数据安全法(二审草案)》《个人信息保护法(二审草案)》等法案逐步出台后,金融、电信、互联网等各个行业将陆续出台数据流通相关的监管规定,虽然监管层面逐步认可并鼓励应用隐私计算以促进数据流通与权益保护之间的平衡,但法律法规并不会明确给出技术应用的具体路径。随着攻击手段和破解技术的不断发展,数据合规要求将不断更新,技术的发展也将随之动态演变。

6 结束语

随着各领域关注度的日益提升,隐私计算已成为发展火热的新兴技术,成为商业和资本竞争的热门赛道。数据流通是释放数据价值的关键环节,隐私计算技术为数据流通提供了解决方案。本文对隐私计算的技术原理进行了阐述,列举了隐私计算当前的主要应用场景,分析了国内外隐私计算产业的发展情况。在此基础上,总结了隐私计算发展面临的问题,并从技术、产业、应用等视角分析了隐私计算未来的发展趋势。我国隐私计算发展具备一定优势、存在广阔应用空间,但由于技术发展仍不完善,因此也面临着一些问题。无论是工程化实现的创新突破还是软硬件之间的优化适配,隐私计算的性能提升都任重道远。

参考文献

[1] Gartner. Gartner 发布 2021 年重要战略科技趋势[EB/OL]. (2020-10-20)[2021-04-01]. https://www.gartner.com/cn/newsroom/press-releases/2021-top-strategic-technologies-cn.

[2] Yao A C. Protocols for secure computations[C]//23rd Annual Symposium on Foundations of Computer Science (sfcs 1982). IEEE, 1982: 160-164.

[3] OMTP. OMTP advanced trusted environment OMTP TR1 v1.1[EB/OL]. (2009-05-28)[2021-04-10]. http://www.omtp.org/OMTP_Advanced_Trusted_Environment_OMTP_TR1_v1_1.pdf.

[4] GlobalPlatform. TEE system architecture v1.0. [EB/OL]. [2021-04-10]. https://globalplatform.org/specs-library/tee-system-architecture-v1-2.

[5] 李凤华, 李晖, 贾焰, 等. 隐私计算研究范畴及发展趋势[J]. 通信学报, 2016, 37(4): 1-11.

[6] 韩嵩, 顾绵雪, 李凤华, 等. 一种基于区块链的数据链接隐私计算方法:, CN109635584B[P], 2019.

[7] 李晖. 隐私计算—面向隐私保护的新型计算[J]. 信息通信技术, 2018, 12(6): 6-8.

[8] 杨强. 联邦学习: 人工智能的最后一公里[J]. 智能系统学报, 2020(1).

[9] Yan H. Efficient secure multi-party computing and its applications[J], 2010.

[10] 王童, 马文平, 罗维. 基于区块链的信息共享及安全多方计算模型[J]. 计算机科学, 2019, 46(9): 162-168.

[11] Rehak D R, Dodds P, Lannom L. A model and infrastructure for federated learning content repositories[Z], 2005.

[12] 工业和信息化部. 大数据产业发展规划(2016—2020年)[R], 2016.

[13] 工业和信息化部. 工业大数据发展指导意见(征求意见稿)[R], 2019.

[14] 中国人民银行. 金融科技(FinTech)发展规划(2019—2021 年)[R], 2019.
[15] Gartner. Gartner 2019 年十大技术趋势[EB/OL]. (2018-10-18)[2021-04-01]. https://wenku.baidu.com/view/55b57b7a0875f46527d3240c844769eae009a3e6.html.
[16] 中国信息通信研究院. 隐私保护计算技术研究报告[R], 2020.

作者简介:

闫树 中国信息通信研究院云计算与大数据研究所大数据与区块链部副主任,高级工程师,博士,主要从事大数据、数据流通等方面的研究工作

吕艾临 中国信息通信研究院云计算与大数据研究所大数据与区块链部工程师,主要从事数据流通技术与流通合规等方面的研究工作

Overview of the development of privacy preserving computing

YAN Shu, LYU Ailin

(Cloud Computing & Big Data Research Institute, China Academy of Information and Communications Technology, Beijing 100191, China)

Abstract: Data circulation is the key to release the value of data. Privacy preserving computing technology provides a solution for data circulation. This paper summarizes the technical principles of privacy preserving computing, lists the main application scenarios of privacy preserving computing, and analyzes the development of privacy preserving computing industry all around the world. On this basis, this paper summarizes the problems faced by the development, and analyzes the future development trend of privacy preserving computing from the perspective of technology, industry and application.

Keywords: privacy preserving computing; development status; development trend

(收稿日期: 2021-04-22)