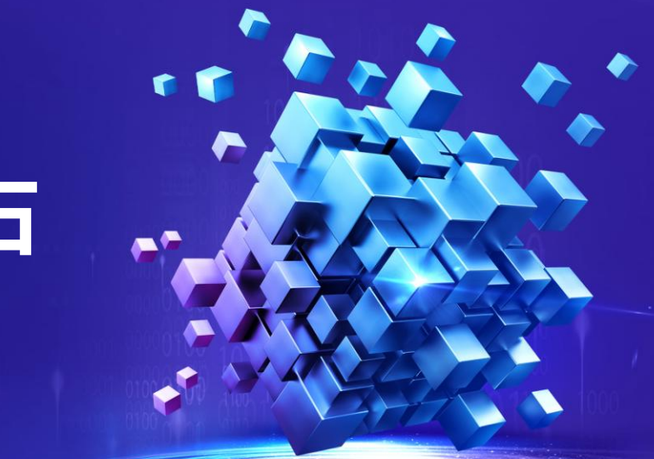
数据安全：保障数据高效合理开发利用的基石

冯登国



CONTENTS

数据与人们的生产生活已密不可分

- ✓ 从飞机、汽车的设计制造，到个人生活点滴的记录，数据已渗透到人类社会的各个方面
- ✓ 数据是资源、是钻石矿、是未来的新油田
- ✓ 数据意味着财富、意味着知识与信息、意味着企业甚至国家在科技浪潮中的核心竞争力

数据来源多样化

- ✓ 在互联网活动以及使用移动互联网过程中产生大量数据
- ✓ 由于物联网技术在智能工业、智能农业、智能交通、智能电网、安全监控等行业的广泛应用，各种类型的传感器被广泛部署，时时刻刻都在产生数据
- ✓ 交通、安防等领域部署的摄像设备产生的数字信号
- ✓ 人体本身就是一个无穷无尽的生物医学数据的重要来源，涉及临床医疗、公共卫生、医药研发等多个领域，类型非常广泛，包括电子病历、医学影像、临床实验数据、个人健康数据监测、基因组序列等
- ✓ 电信、金融、智慧城市、交通、科学研究等都会产生大量数据

数据安全形势严峻（1）

- ✓ 数据在带来巨大价值的同时，也引入了大量的安全风险与挑战
- ✓ 数据安全不仅关系到国家主权、安全和发展利益，而且关系到公民、组织的合法权益
- ✓ 数据安全是保障数据高效合理开发利用的基石

数据安全形势严峻（2）

✓ 数据泄露量逐年都在创新高，据报道2020年全球泄露量超过360亿条记录。数据泄露事件遍布于大小机构，以往认为相对安全的警察部门、电信运营商和互联网巨头也未能幸免

- 欧洲用户规模最大的德国电信子公司T-Mobile于2020年3月6日宣布其电子邮件供应商遭到黑客入侵，导致一些包括个人信息和财务信息的电子邮件账户泄露
- 自称“分布式揭密（DDoSecrets）”的激进团体于2020年6月22日在其网站上发布了一个296GB的数据包“警察揭密（BlueLeaks）”，并声称来自美国执法机构
- 全球规模第三的酒店集团万豪国际（Marriott International）于2020年4月1日宣布其发生数据泄露事件，导致520万客人的个人信息泄露；万豪国际的子公司曾于2018年发生过包括护照在内的数百万名旅客信息的数据泄露事件，被罚款9900万英镑
- 2021年3月10日有一则消息称，一群黑客自曝入侵了美国硅谷初创公司Verkada采集的大量安全摄像机数据，并盗取了15万个监控摄像头实时视频，偶然曝光美国电动汽车制造商特斯拉上海工厂实况

各国都高度重视数据安全

- ✓ 美国国防部于2020年10月8日发布《数据战略》，该战略强调通过数据融合实现军种联合，高度重视数据的安全性，强调数据全寿命周期的标准化能力
  - 从“网络中心战”向“数据中心战”转型
- ✓ 欧盟推出《通用数据保护条例》（GDPR）
- ✓ 日本于2017年部署实施了《个人信息保护法案》（APPI），日本内阁于2020年3月10日批准了《个人信息保护法案》修正案
- ✓ 2018 年美国加州为保障本州消费者的各项隐私权利，通过了《加利福尼亚州消费者隐私法案》（CCPA），该法案于 2020 年 1 月 1 日生效
- ✓ 2020年6月，十三届全国人大常委会审议通过了《中华人民共和国数据安全法（草案）》

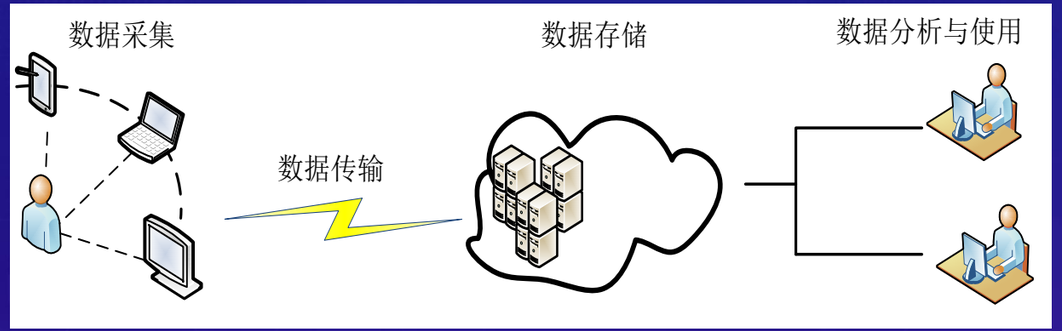
什么是数据安全

- ✓ 2020年6月，十三届全国人大常委会审议通过了《中华人民共和国数据安全法（草案）》，这份草案将数据和数据安全分别定义为
  - 数据：任何以电子或者非电子形式对信息的记录
  - 数据安全：通过采取必要措施，保障数据得到有效保护和合法利用，并持续处于安全状态的能力
- ✓ 这里主要关注电子化、数字化的数据

CONTENTS

数据全生命周期

- ✓ 数据的生命周期包括数据产生、采集、传输、交换、存储、分析、使用、分享、销毁等诸多环节，每个环节都面临着不同的安全威胁，需要进行全链条创新研究
- ✓ 其中，安全问题较为突出的环节是数据采集、数据传输、数据存储、数据分析与使用



数据主要安全威胁

- ✓ 数据泄露：数据被偷取、窃听、窃取或泄露而造成数据泄露，通过通信流量分析也可能导致信息泄露，也可能从公开的数据推理出敏感数据而造成数据泄露
- ✓ 数据破坏：数据被篡改（如更改、插入、删除、重放等）或假冒而造成数据破坏，系统或设备感染病毒、蠕虫等恶意代码而导致数据破坏，电磁干扰也可能造成数据破坏
- ✓ 隐私泄露：通过数据分析、处理或推理等手段都有可能导致个人隐私（如用户身份、社交关系、属性、轨迹等）泄露
- ✓ 数据失控：新的数据处理或应用模式（如云计算）而导致用户数据失控，攻击者利用攻击手段获得数据中心控制权而导致数据失控
- ✓ 数据滥用：数据被非法使用，或者被非授权使用或越权使用，也包括数据不可溯源、不可追踪
- ✓ 数据损坏或丢失：存储设备或硬盘驱动器损坏而造成数据损坏或丢失，人为操作失误可能会误删除系统的重要文件或修改影响系统运行的参数导致系统宕机而造成数据损坏或丢失，地震、火灾等自然灾害也可能造成数据损坏或丢失，电源供给系统故障也可能导致存储设备或硬盘上的数据损坏或丢失

数据安全研究目标与方向

- ✓ 围绕数据全生命周期，突破一批关键技术，全面提升数据安全防护能力，数据安全治理能力和数据安全威慑能力

数据安全关键技术

- ✓ 数据安全问题的解决离不开配套的法规和政策的支持以及严格的管理手段，但更需要有可信赖的技术手段支持
- ✓ 近年来，涌现出一大批数据安全新技术
  - 密文检索技术
  - 密文计算技术（如同态加密、函数加密）
  - 基于风险分析的访问控制技术
  - 差分隐私保护技术
  - 安全多方计算技术
  - 完整性验证技术
  - 零信任技术
  - ……

- ✓ 这里重点介绍一种数据安全新技术：同态加密技术

同态加密技术

✓ 背景

- 云计算环境下，数据的所有权和处理权相分离
- 为了安全，用户一般用密态存储数据

✓ 需求

- 如何对这些密文进行处理成为急需解决的问题
- 经典安全机制如AES难以满足这种需求，需要先解密才能处理数据，这个过程使数据暴露在众多漏洞及威胁之下，这就需要更多具有新型功能的安全机制，如同态加密

同态加密的基本思想

- ✓ 同态加密（Homomorphic Encryption，HE）的思想是由Rivest等人于1978年提出的，亦称“隐私同态”（Privacy Homomorphism）
- ✓ 基本思想
  - 在不使用私钥解密的前提下，能否对密文数据进行任意的计算，且计算结果的解密值等于对应的明文计算的结果
  - 形式化地讲，非对称性场景下的同态加密问题可以定义为：给定一组消息$(m_1,m_2,...,m_t)$在某个公开加密密钥PK下的密文为$(c_1,c_2,...,c_t)$，给定任意一个函数f，在不知道消息$(m_1,m_2,...,m_t)$以及私钥解密密钥SK的前提下，可否计算出$f(m_1,m_2,...,m_t)$ 在PK作用下的密文，而不泄漏关于$(m_1,m_2,...,m_t)$以及$f(m_1,m_2,...,m_t)$的任何信息?

同态加密的计算模式

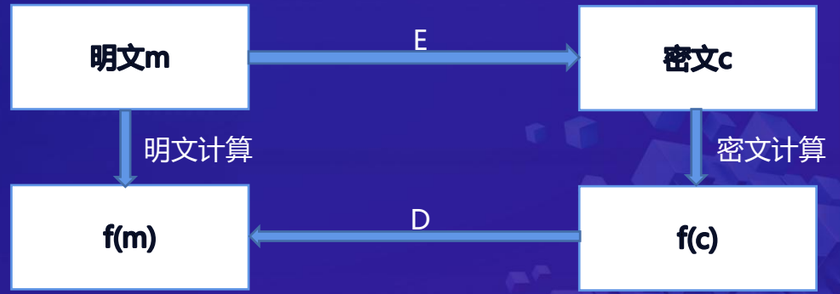
用户

$(c_1, c_2, \ldots, c_t) = (E(PK,m_1), E(PK,m_2), \ldots, E(PK,m_t))$

f

$E(PK, f(m_1, m_2, \ldots, m_t))$

云服务器



同态加密的发展历程

- ✓ 从单同态加密到类同态加密（Somewhat Homomorphic Encryption，SWHE）
- ✓ 再到全同态加密（Fully Homomorphic Encryption，FHE），经历了30多年的历程，最终于2009年由时为Stanford大学计算机科学系博士生的Craig Gentry基于理想格构造出第一个FHE方案，解决了这一重大问题
- ✓ 这一问题一直被密码学界视作一个“海市蜃楼”般的问题，一些密码学家甚至不吝誉之为“密码学圣杯”。“海市蜃楼”以及“圣杯”，足以说明这个问题的困难性以及此前整个密码学界对解决这个问题的不乐观态度！例如，在基督教中，圣杯一直被视作是一个“永远无法找回”的宗教象征！

同态加密方案

- ✓ 自从同态加密技术诞生以来，许多密码学研究者开始致力于同态加密方案的研究，并提出了大量的支持一定同态能力的加密方案
  - 支持任意次乘法同态操作的加密方案主要有RSA加密方案和Elgamal加密方案
  - 支持任意次加法同态操作的加密方案主要有GM加密方案、Benaloh加密方案、OU加密方案、NS加密方案、Paillier加密方案、DJ加密方案
  - 支持任意次加法同态操作和一次乘法同态操作的加密方案主要有BGN方案
  - 此外，Fellows等人于2006年提出的PC加密方案可支持任意电路但误差随密文规模呈指数级增长，Sanders等人使用隐私电路（Circuit-private）加法同态加密构造的隐私电路SYY加密方案可处理NC_1电路，Ishai等人使用线路图（Branching programs）同态处理NC_1电路

全同态加密方案（1）

- ✓ Gentry于2009年基于理想格构造的第一个FHE方案发表在ACM STOC2009国际会议上，国际ACM协会在其旗舰刊物Communications of ACM（2010年第3期）上以“一睹密码学圣杯芳容”为题并以“重大研究进展”的形式对这一成果进行了专题报道
- ✓ Gentry方案的基本构造思路：
  - 首先，构造一个类同态加密（SWHE）方案。SWHE方案不能做到全同态，只是一个“有点同态”的加密方案，只能对加密数据进行低次多项式计算的加密方案，也就是说只能同态计算“浅的电路”的加密方案
  - 其次，给出一种将SWHE方案修改为自举（Bootstrappble）同态加密方案的方法
  - 最后，通过递归式自嵌入，任何一个自枚举型同态加密方案都可以转化为一个全同态加密方案
- ✓ Gentry方案的安全性建立在理想格上的有界距离编码问题（BDDP）和稀疏子集和问题（SSSP）的困难性假设上，BDDP假设用于保证类同态加密方案的选择明文安全性（CPA），SSSP假设则是由于电路压缩（Squashing）引入的额外假设

全同态加密方案（2）

- ✓ 目前，全同态加密方案主要有两大类
  - 一类是无限层FHE方案，也称无界自举型FHE方案，这是真正意义上的FHE方案，其典型代表是Gentry方案。由于这类方案采用基于同态解密的Bootstrapping技术，所以无限层FHE方案理论上可以进行无限深度的同态操作，但付出的代价是同态运算的计算开销、密钥规模和密文尺寸都比较大
  - 另一类是层次型FHE方案，其典型代表是BGV方案。这类方案需要预先给定所需要的同态计算深度d，以致可以执行深度为d的多项式同态操作，从而可以满足绝大多数应用需求
- ✓ 关于FHE的研究进展可归纳为以下几个方面
  - 方案设计研究，如基于整数的设计、基于编码的设计
  - 效率改进和算法可用性研究，如基于误差学习（LWE）和ring-LWE问题的全同态加密方案、基于Gentry初始方案的改进
  - 实现和应用研究，各种FHE方案的软硬件实现、标准化和应用研究也受到高度重视，最具代表性的是开源的代码库HElib

同态加密的研究也受到了各国政府的高度关注

- ✓ 美国DARPA于2020年3月启动“虚拟环境中的数据保护（DPRIVE）”项目，其目的是设计一个用于计算全同态加密（FHE）的硬件加速器，以显著降低当前存在的计算负担，大幅加快FHE计算速度
- ✓ 美国DARPA于2010年7月推出了“密文可编程计算（Programming Computing on Encrypted Data，PROCEED）”项目，其目的是为“密文未经解密即可进行计算”而研发实用化的方法，以及为达到此目标所需要的新的编程计算语言
- ✓ 欧盟于2015年1月启动了“同态加密应用与技术（Homomorphic Encryption Applications and Technology，HEAT）”项目，其目的是开发同态加密的软硬件工具库以及安全性分析和参数推荐
- ✓ 美国情报高级研究计划局IARPA于2017年启动了“具有低开销的同态加密计算技术（HECTOR）”项目，其目的是聚焦同态密码编程语言与格式表达、密文计算协议、集成与标准化

同态加密的软件实现进展

类型	调研算法	消息空间（比特）	效率（毫秒，ms）
单同态加密	ElGamal算法（乘法同态）	40	0.001毫秒/次
	Paillier算法（加法同态）	1024	0.008毫秒/次
类同态加密 层次型同态加密（5层[1]）	BFV算法（整数运算，微软SEAL库，公开版[2]）	20	加法门：0.074毫秒/次 乘法门：17.279毫秒/次
	CKKS算法（浮点数运算，微软SEAL库，公开版[2]） 当前版本可能有安全隐患	浮点数	加法门：0.090毫秒/次 乘法门：0.818毫秒/次
单比特全同态加密（支持自举）	TFHE算法（TFHE库）	1	每个逻辑门:18毫秒/次
大比特明文空间全同态加密（支持自举）	BGV算法（IBM HELib库）	12	加法门：5.949毫秒/次 乘法门：210.843毫秒/次

注：

1. 在公开的库中没有找到100层类同态加密/层次型加密的数据，直觉上100层应该足够支持自举操作了。SEAL库最大支持64层，不过没有数据用于测试。此外，性能上猜测100层参数在不考虑批处理的情况下的时间单位应该至少是秒级。
2. 据说SEAL库的微软内部版可以支持自举，目前没有公开源码和数据。
3. 不同方案的选取参数的安全性有差别，支持的明文空间的差也较大，相互之间的性能很难公平比较。这里的软件实现没有考虑批处理。

全同态加密研究的重大意义

✓ 对称密码

- 从有人类历史以来发展到今天
- 1949年成为一门科学
- 解决了加密保护问题

✓ 公钥密码

- 1976年诞生
- 解决了安全认证问题
- 遇到量子计算等新型计算技术的严峻挑战

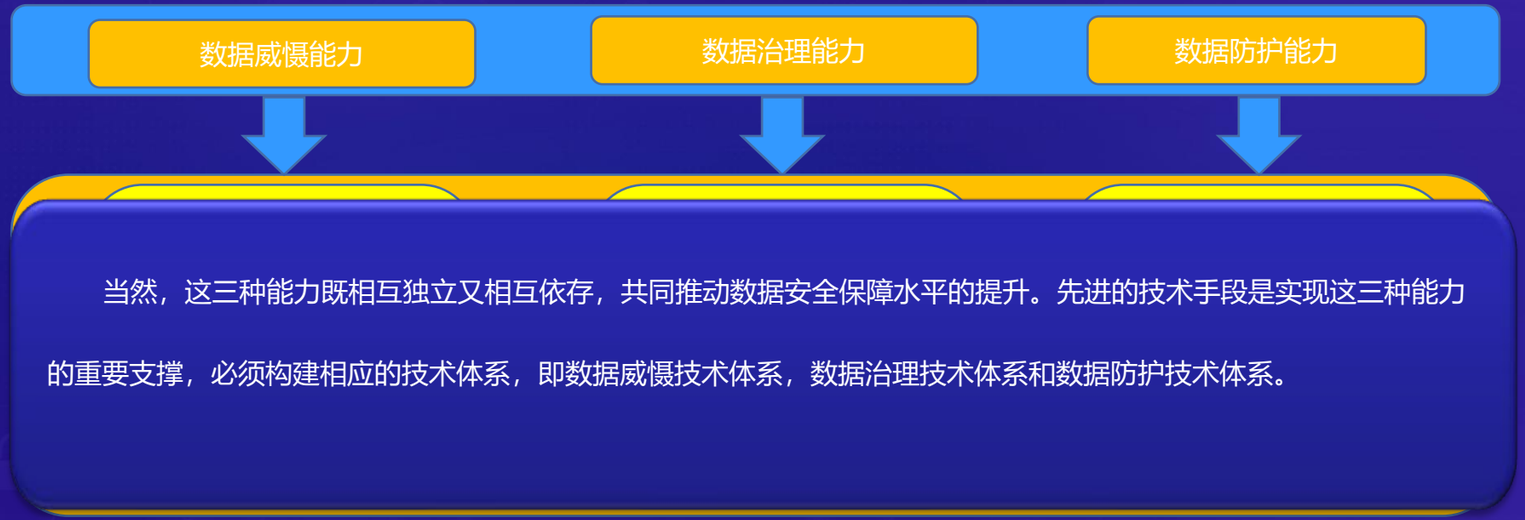
✓ 全同态密码

- 1978年提出，2009年真正解决了存在性问题
- 同态加密技术的重要特性是在保障数据机密性情况下执行数据计算
- 从理论上解决了**机密/隐私计算**问题
- 当前只是可实现，还远未达到高性能实现，离实际应用还有很大距离
- 一旦其高性能实现问题得到解决，必将对安全计算产生巨大的推动作用，可大大提升数据安全性

CONTENTS

数据安全发展思路

- ✓ 应从国家安全、社会组织安全、公民个人安全三个层面关注数据安全问题，全面提升数据安全的三种能力



关于数据安全的几点建议

深刻认识数据安全的内涵

大数据时代的数据安全不仅包括传统的机密性、完整性、可用性等，也包括隐私保护；不仅包括防止数据泄露的隐私保护，也包括数据分析意义下的隐私保护。

紧跟国际数据安全技术发展趋势

大数据时代的数据安全技术成为国际学术界关注的热点，聚集了大量的科研投入，新技术、新突破层出不穷，在进行原始创新和集成创新的同时，应充分借鉴国际先进成果。

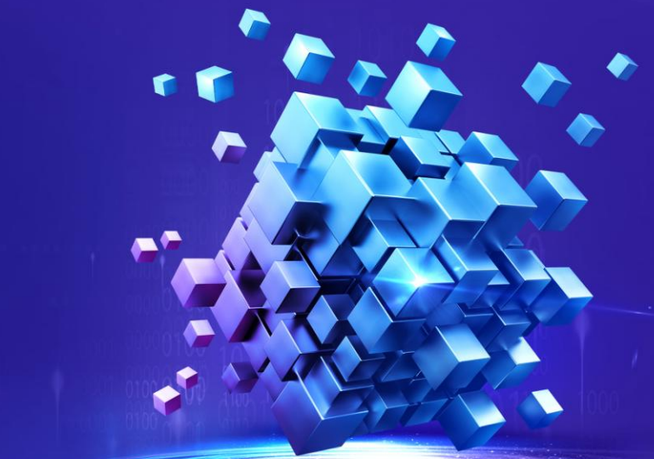
加强数据安全法律法规研究与制定

通过法律手段规范市场、强化监管，合理平衡数据管制与自由流动，营造良好的法治环境。

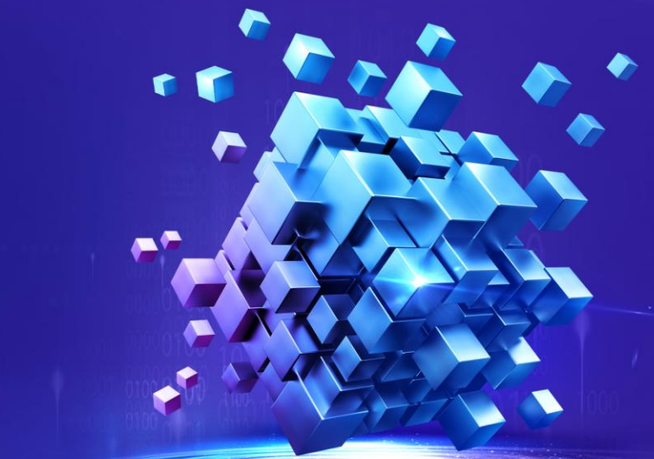
紧密结合产业和应用实际

自主掌控一批核心关键技术和产品，推出切实可行的安全解决方案和标准规范，为保障数据产业健康稳定发展保驾护航。

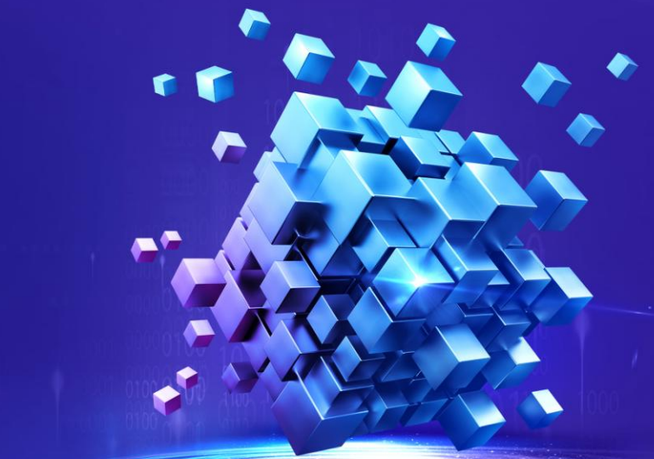
谢谢!



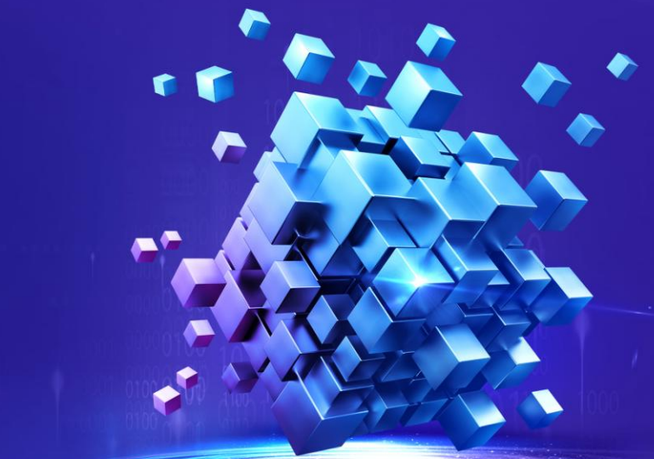
谢谢!



谢谢!



谢谢!



谢谢!