CNGBdb

国家基因库生命大数据平台

区块链技术

如何保障基因大数据应用过程中的数据安全和隐私

华大区块链：李士森

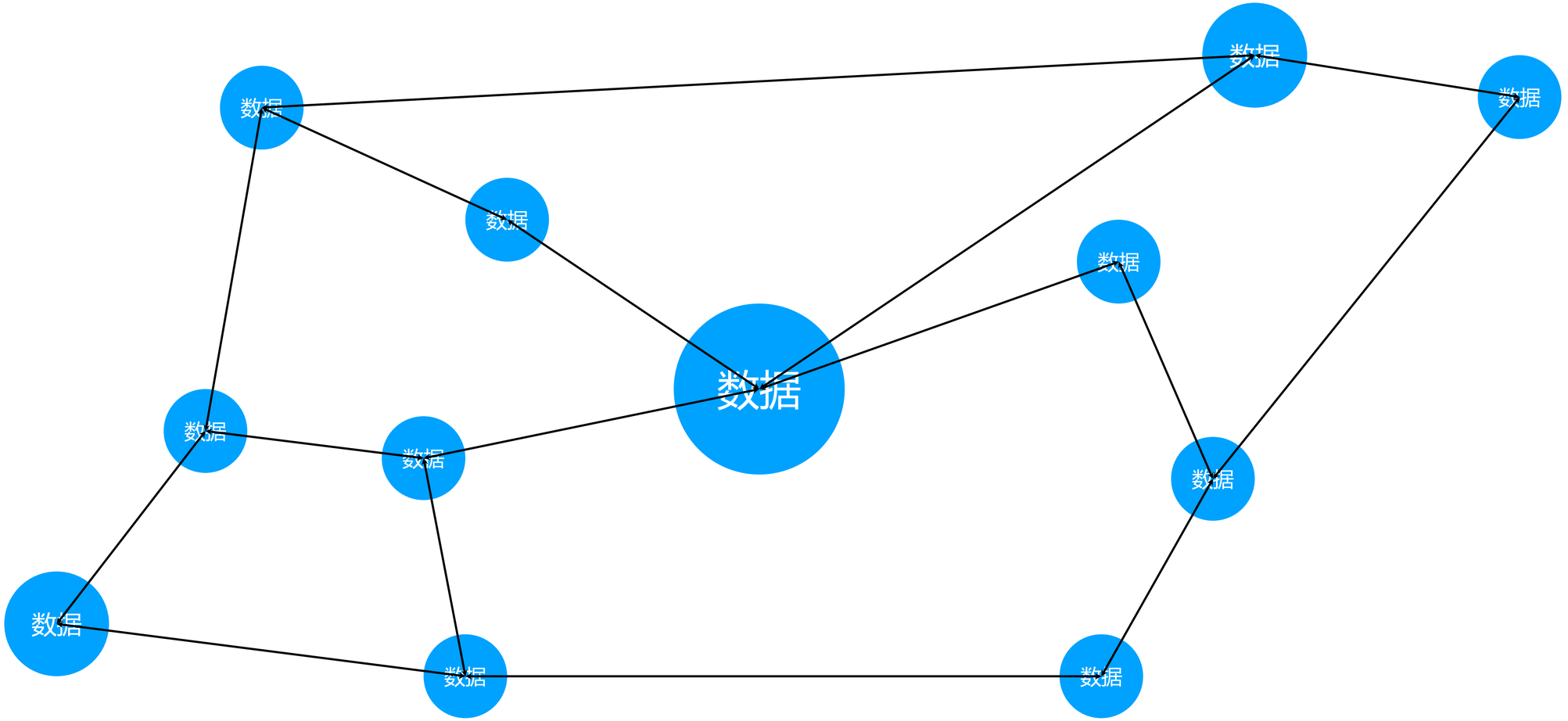
目录

01 数据安全的意义和关键问题

数据已成为基础性关键战略资源

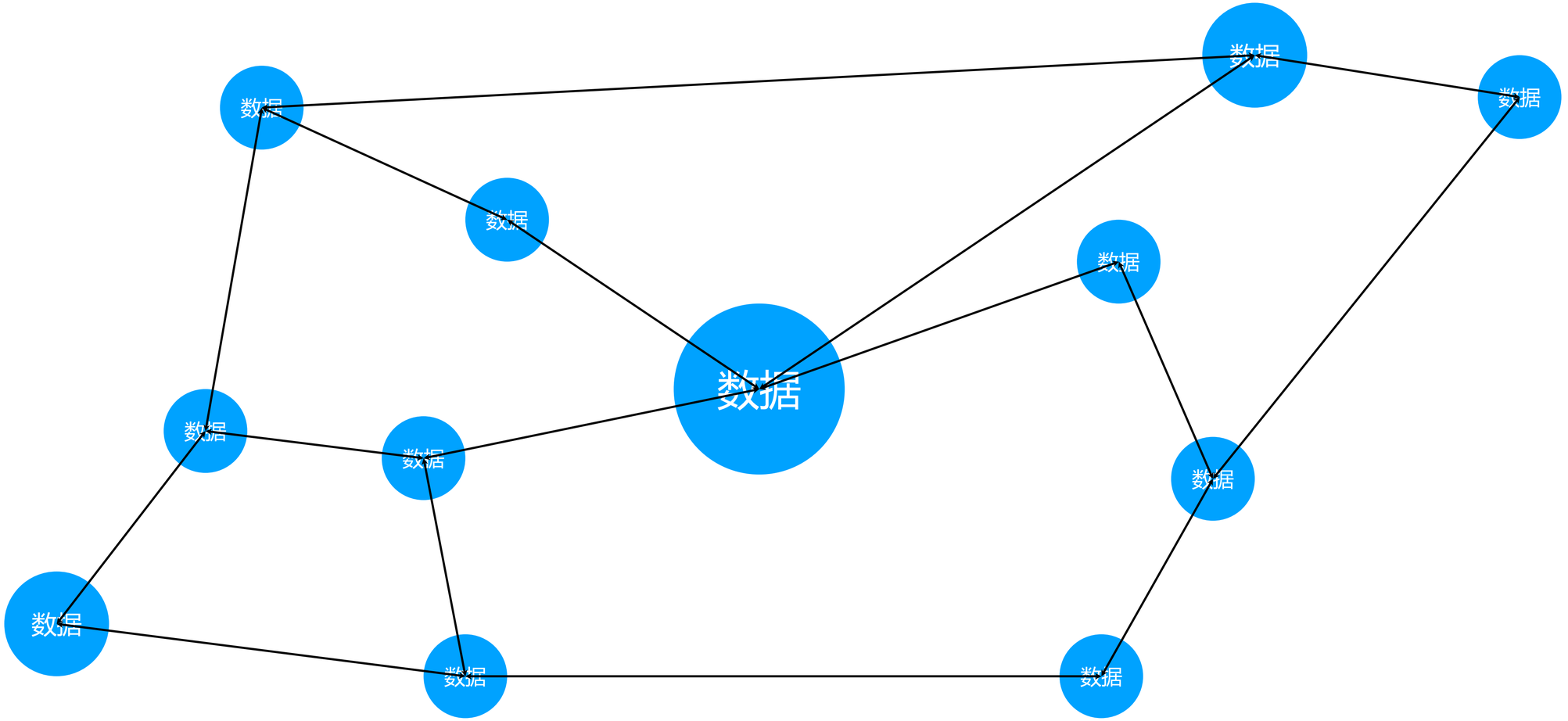
毫不夸张的说，当今社会，数据已经成为了比肩石油的基础性关键战略资源，正在颠覆各个行业的发展模式。

数据流通–释放数据无限价值



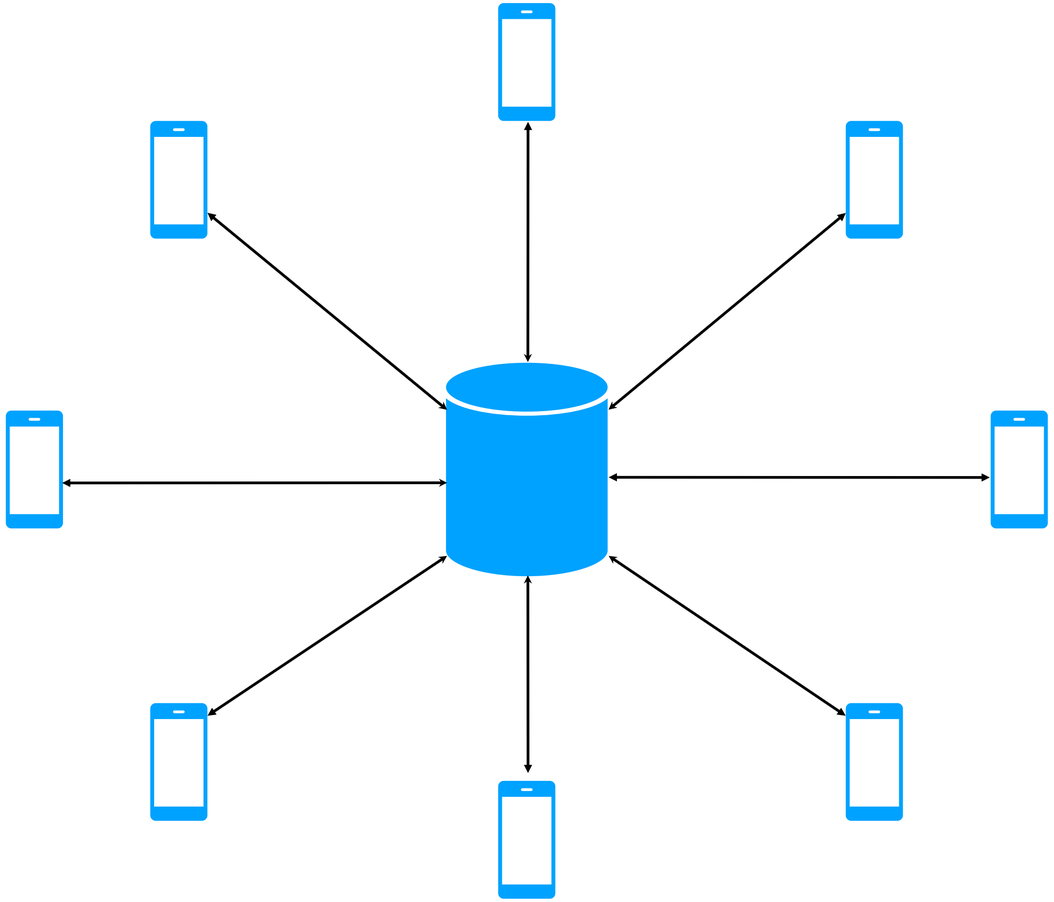
随着大数据产业市场规模的迅速扩大，大数据的开放共享、交换流通成为趋势。可以说，数据流通是释放数据价值的关键环节。

数据流通的基本概念



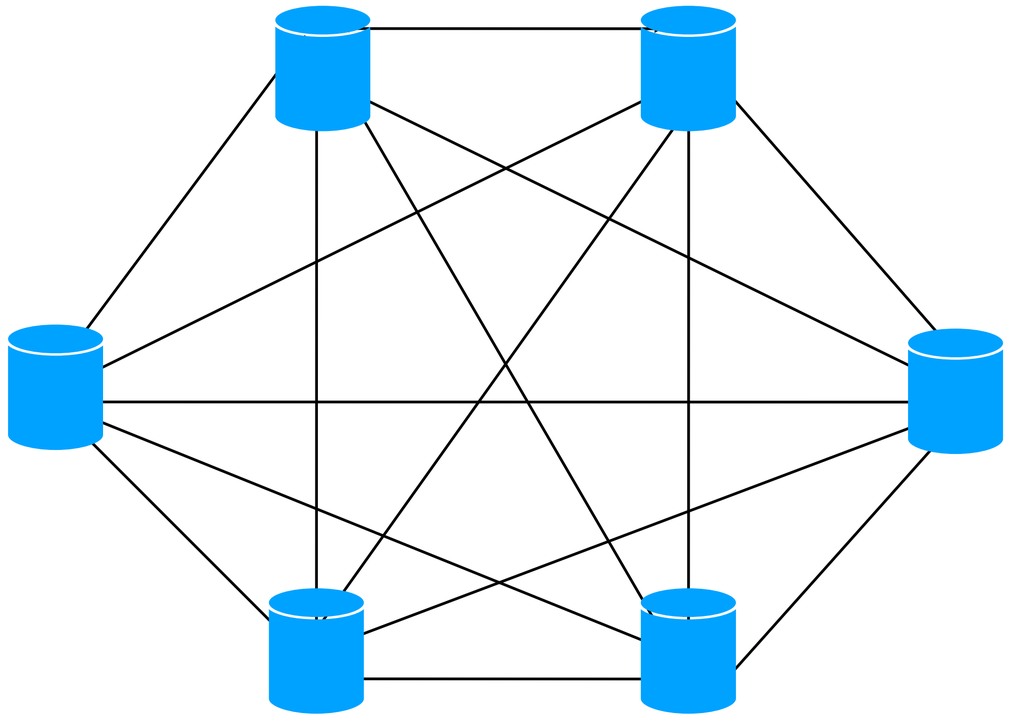
数据流通可以被定义为某些信息系统中存储的数据作为流通对象，按照一定规则的从供应方传递到需求方的过程。

点对点流通模式是数据的供需双方两个节点组成的数据流通系 统，数据供应方按照需求方的要求与约定规则，向需求方开放数据资 源库完成流通对象的转移。



星型结构流通模式

星型结构流通模式是由一方作为数据资源收集者，收集与整理所 有其他数据资源拥有者的数据资源，并对所有数据资源进行再次加工， 最后提供给所有数据资源需求方的模式。这种流通模式多存在于数据 平台以中间代理人身份为数据提供方和数据购买方提供数据交易撮合服务。数据提供方往往选择一种交易方式对数据自行定价出售，并按特定交易方式设定数据售卖期限及使用和转让条件。



网状结构流通模式

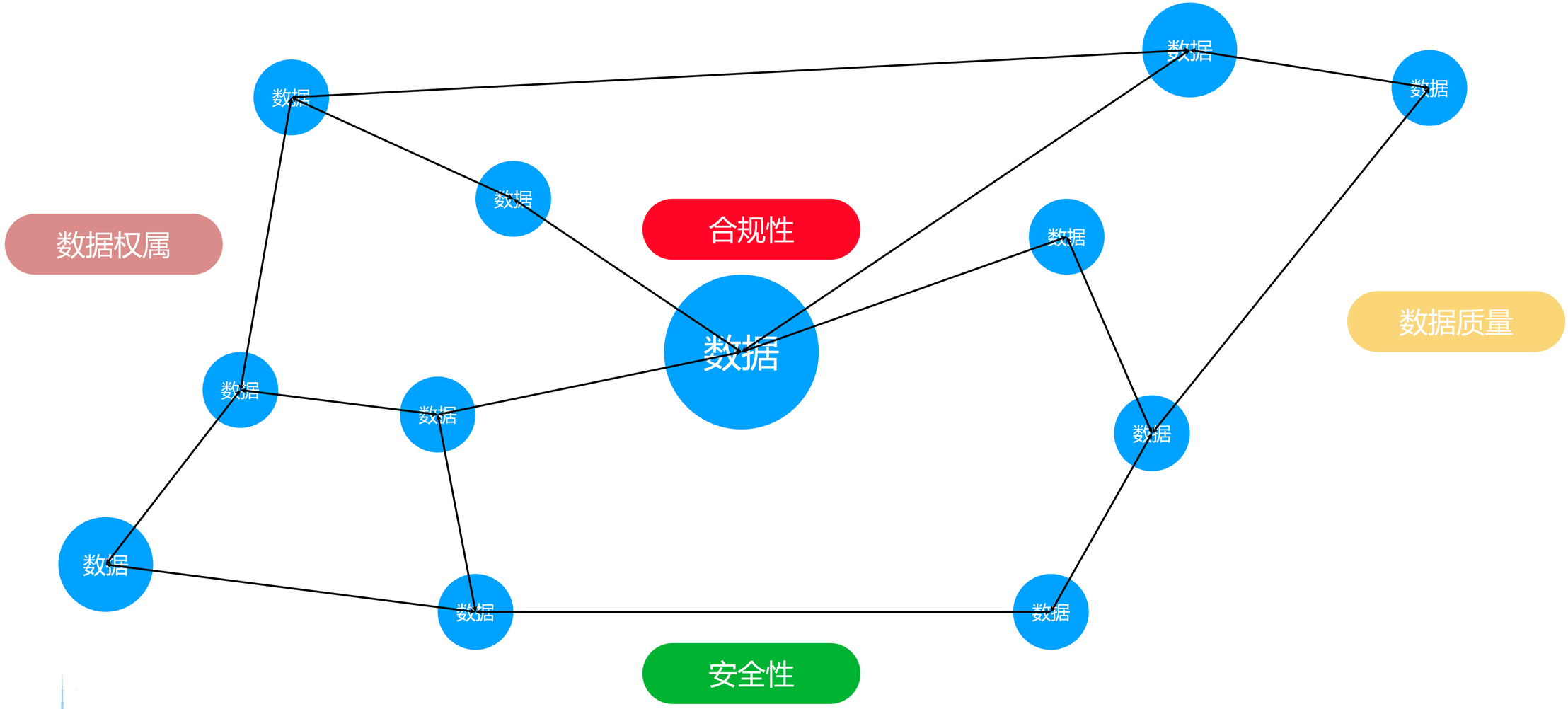
网状结构流通模式，是由一个或多个数据资源组织方，组织数据 资源的需求与供给，形成流通任务，并组织不同供需多方之间按需传 输数据资源的模式。按照通信、互联网等具有网络特性的行业发展规律，随着数据流通相关技术的逐渐成熟，社会整体数据资源将逐渐汇聚连接，最终形成网状结构的数据流通模式。

数据流通共享的意义

数据流通

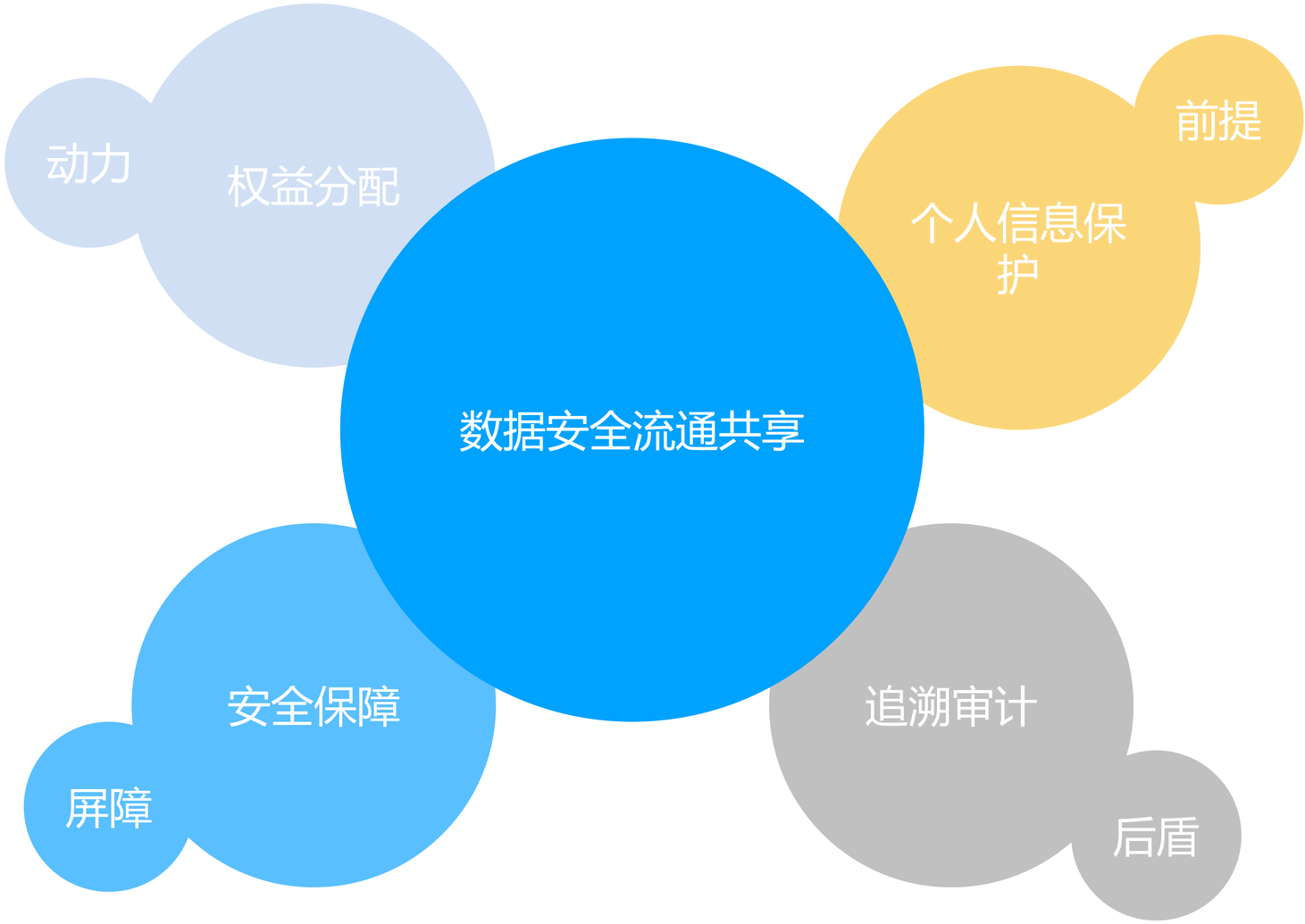
- 数据流通是促进数据资源融合的重要技术领域
- 数据流通是“互联网+”行业应用的新动力。
- 数据流通是数据资源成为数据资产必要条件。
- 数据流通是中国数字经济战略的基础保障。

数据安全流通面临的问题



目前数据流通在数据资源、数据质量、数据定价和流通合规性等方面存在问题与挑战。

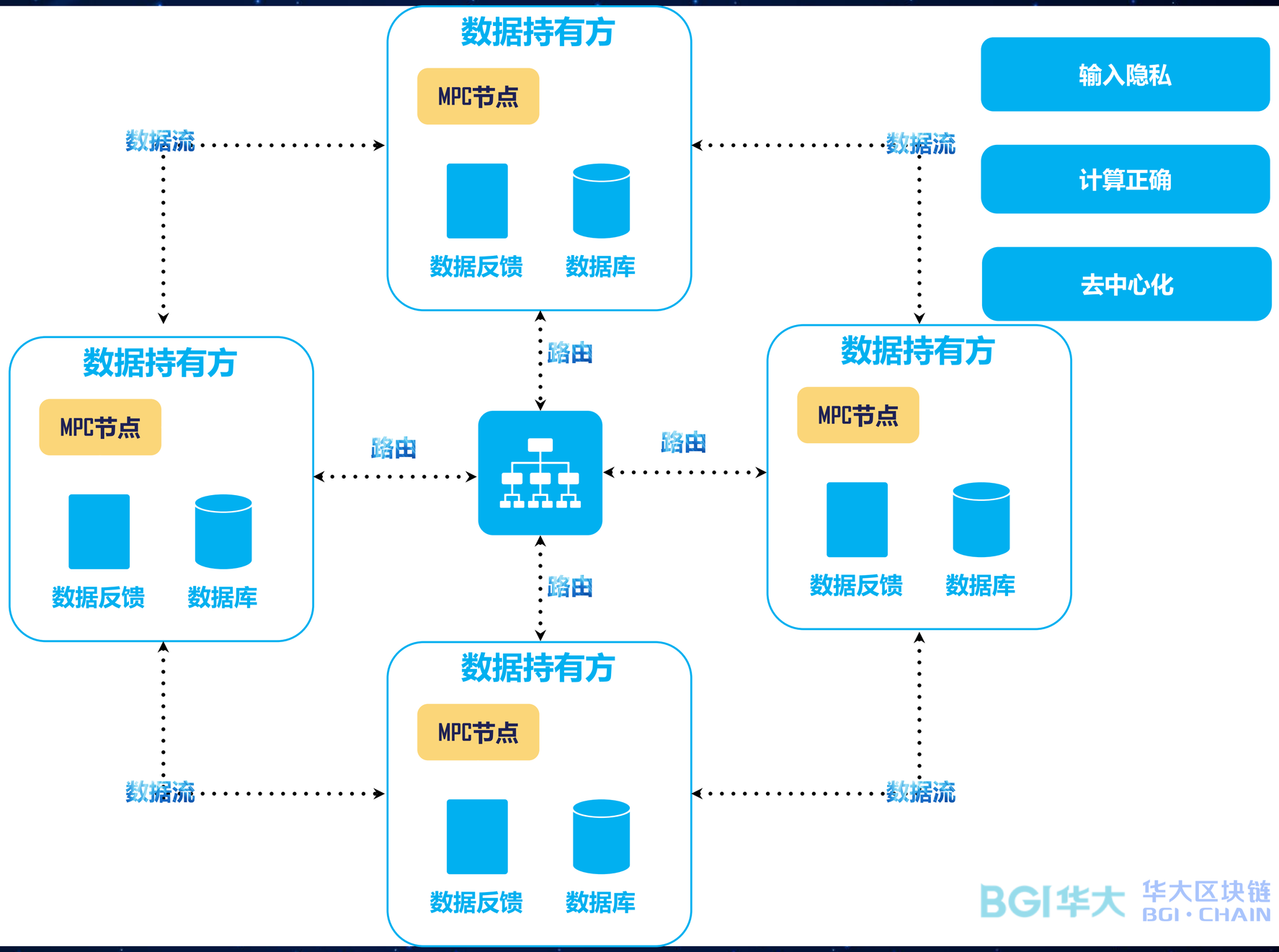
数据安全流通的技术需求



02 数据安全的主要技术框架

安全多方计算框架

安全多方计算(MPC)最初由图灵奖获得者、中国科学院院士姚 期智教授在1982 年通过百万富翁问题提出。安全多方计算的研究主 要针对无可信第三方情况下，安全地进行多方协同计算问题。即在一 个分布式网络中，多个参与实体各自持有秘密输入，各方希望共同完成对某函数的计算，而要求每个参与实体除计算结果外均不能得到其 他用户的任何输入信息。



安全多方计算的特点

输入隐私

安全多方计算研究的是各参与方在协作计算时如何对各方隐私数据进行保护，重点关注各参与方之间的隐私安全性问题，即在安全多方计算过程中必须保证各方私密输入独立，计算时不泄露本地任何 数据

计算正确

多方计算参与各方就某一约定计算任务，通过约定 MPC 协议进行 协同计算。计算结束后，各方得到正确的数据反馈。

去中心化

传统的分布式计算由中心节点协调各用户的计算进程，收集各用 户的输入信息。而在安全多方计算中，各参与方地位平等，不存在任 何有特权的参与方或第三方，提供一种去中心化的计算模式。

参与方个数区分

分为两方计算和多方计算，之间存在本质的区别。

计算场景区分

分为特定场景和通用场景。特定场景是指针对特定的计算逻辑，比如比较大小，确定双方交集等。

安全多方计算的适用场景

数据可信交互

安全多方计算理论为不同机构间提供了一套构建在协同计算网 络中的信息索引、查询、交换和数据跟踪的统一标准，可实现机构间 数据的可信互联互通，解决数据安全性、隐私性问题，大幅降低数据 信息交易摩擦和交易成本，为数据拥有方和需求方提供有效的对接渠道，形成互惠互利的交互服务网络。

数据安全查询

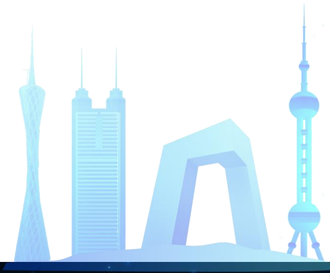
数据安全查询问题是安全多方计算的重要应用领域。使用安全多计算技术，能保证数据查询方仅得到查询结果，但对数据库其他记 录信息不可知。同时，拥有数据库的一方，不知道用户具体的查询请 求。

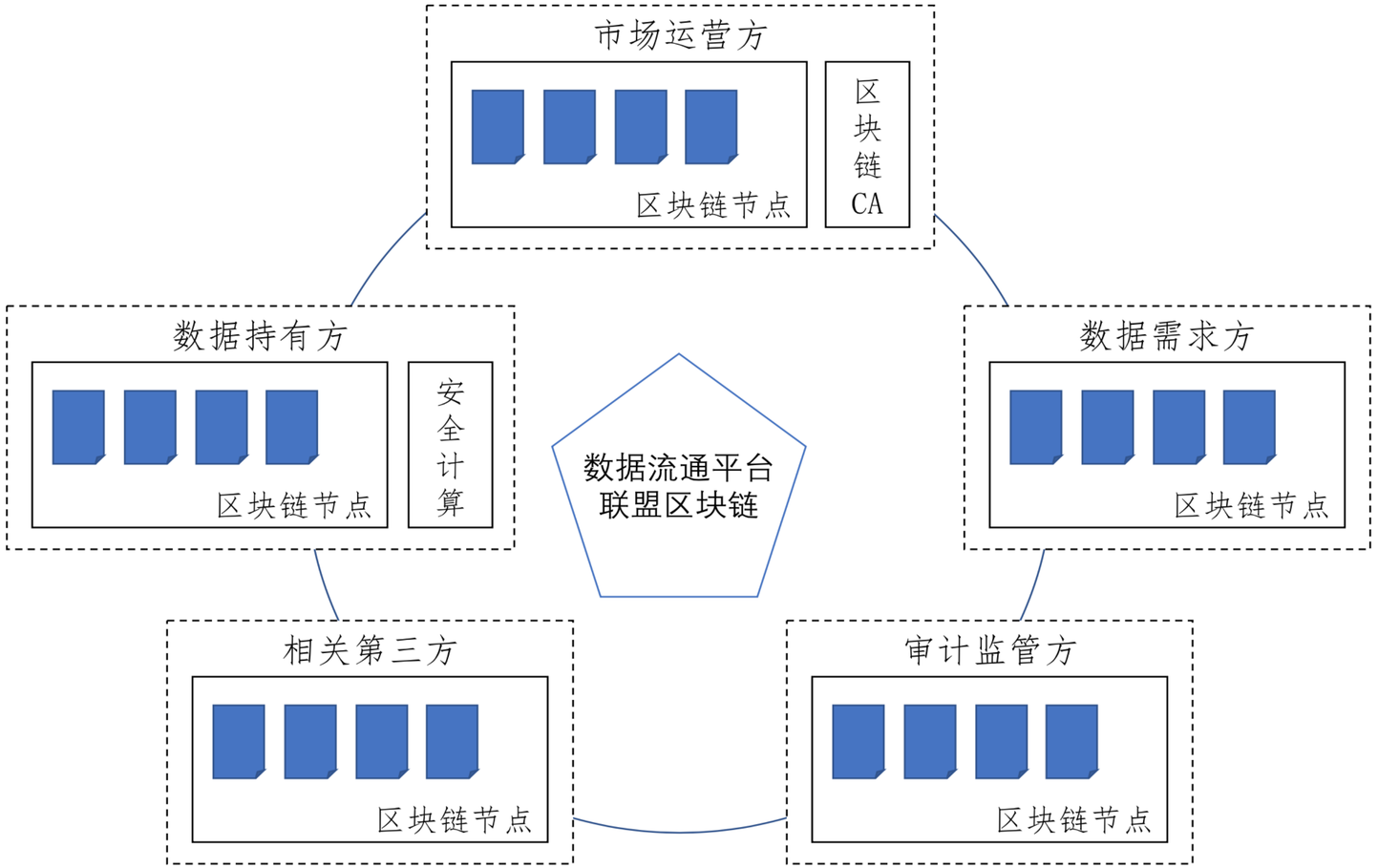
数据联合分析

随着大数据技术的发展，社会活动中产生和搜集的数据和信息量 急剧增加。敏感信息数据的收集、跨机构的合作以及跨国公司的经营 运作等给传统数据分析算法提出新的挑战，已有的数据分析算法可能 会导致隐私暴露，数据分析中的隐私和安全性问题得到极大的关注。 将安全多方计算技术引入传统的数据分析领域，能够在一定程度上解决此问题，其主要目的是改进已有的数据分析算法，通过多方数据源协同分析计算，使得敏感数据不被泄露。

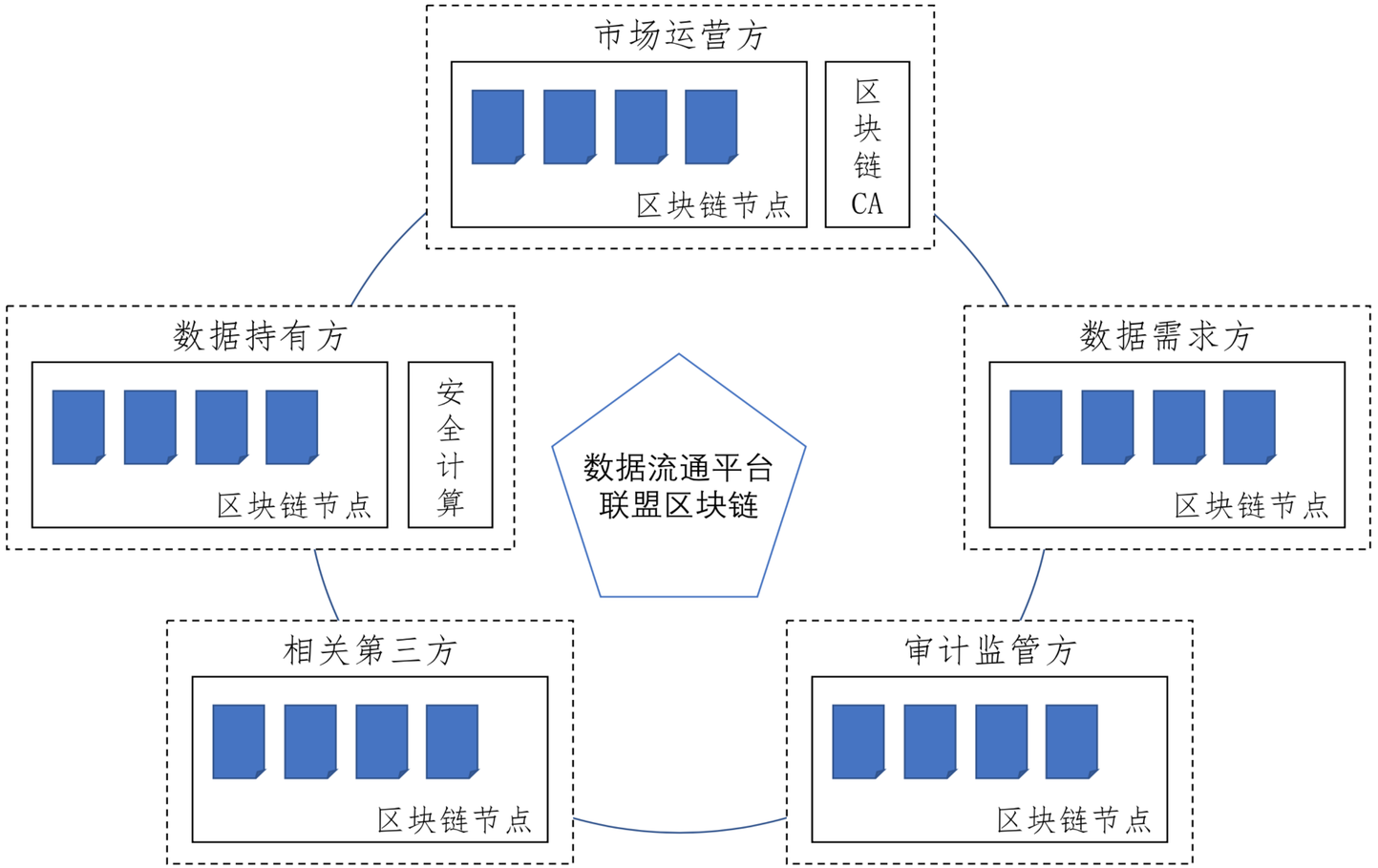
什么是区块链?

使用密码学技术将共识确认的区块链按照顺序追加形成的分布式账本





目前区块链应用主要有三种模式:公有链、联盟链、私有链。公 有链是运行在互联网的完全分布式区块链;联盟链则是由多个关联机 构共同发起和运营，带有准入机制;私有链是公有链的私有化部署， 往往由单个机构主持运行。联盟链模式提供的准入控制，更易于参与 者的身份筛选和责任认定，有助于数据安全和质量保障，因此更适合数据流通领域。



合约生效前需要各方用户达成一致的内容包括但不限于:数据内容、数据量、数据使用限制、交易期限、计费方式等。数据授受双方 通过合约执行数据核验、数据计算、结果提交和放行等，并由市场运 营方、审计监管方规范合约履行过程。数据计算由链下的安全计算环 节来完成，在链上完成结果汇总。这一安全计算在数据持有方认可的 保密执行环境中完成，避免计算过程数据外流或失窃。

区块链技术技术的特点

去中心化

在区块记录生成过程中，区块链参与方的权利和 义务平等。去中心化同时也是多中心化，即在部分节点失效，甚至恶 意化的情况下，仍能保证区块链的正常运行。

自信任

区块链所有节点之间无需信任也可以进行交互。因 为区块链账本的存储是多副本的，合约执行和记录添加是基于公开的 机器代码和共识机制，所以节点的任何行为都是可预期的。

防篡改

已生成的区块链记录由全体成员共同保存。而且任何节点的本地账本都自动与共识版本对齐。在一定的规则和时间范围 内，区块记录的更改行为都是不可实现的。

加密技术

区块链的加密技术应用一方面可保证信息传递的隐秘安全、不可 篡改，另一方面是用户身份的唯一性认定。

共识机制

在区块链中共识机制就是在信道可靠的前提下，保证分布式系统 认定合法的区块按照一定的顺序组织成链条。

共识机制包含以下三种属性:
(1)一致性，即不存在两个正确的 决策节点决定出不同的区块;
(2)合法性，即决定的区块必须是由一个节点提出的提案;
(3)可终止性，即一致性的区块必须在有限时间内完成。

数据发布场景

1.数据提供方生成数据身份:数据提供方根据所要发布的数据文 件属性，生成数据身份，包含数据的各类特殊属性。
2.数据提供方提交数据发布合约:数据提供放根据数据发布合约 的模版，将数据身份填入合约，向区块链发送创建智能合约请求。
3.市场运营方核实数据发布合约:市场运营方对合约中各类特殊 属性完成核实。核实通过后，对该合约进行签名，完成数据发布合约。

数据发布场景

1.数据需求方提交数据授权合约:数据需求方可通过区块链检索、 查阅意向数据，并根据合约模板，向区块链提交合约。
2.合约签署:合约中规定的参与者在本地区块链收到智能合约请 求，并进行签署。市场运营方在签署前，需确认各个参与者资质。
3.数据使用:数据持有方根据合约的授权签名和算法要求，执行 数据计算。数据实体在使用过程不能离开该安全计算环境，对于多个 数据持有方，则引入安全多方计算。
4.数据结果提交及提取:数据持有方将结果提交到区块链合约中。
对于多个数据持有方，由合约完成结果的汇总。之后数据需求方从区 块链获取计算结果。

对于数据持有方

(1)数据隐私保护 区块链平台上公开的数据身份仅包含数据指纹、数据持有方 ID 等关键的概要信息。数据实体的核心记录仅存储在数据提供者认定的 安全计算环境中，即实体本身并未在区块链公开，以保证持有者的隐 私不泄漏。
(2)操作透明可监督
数据流通的发起，即区块链智能合约的发起，都必须经过数据持 有方的应答和审批。流通过程中的操作都根据合约中约定的策略依次 自动执行，并进行链内共识、记录。数据持有方可从区块链中读取批准合约的操作记录，以实现监控、核验和审计。

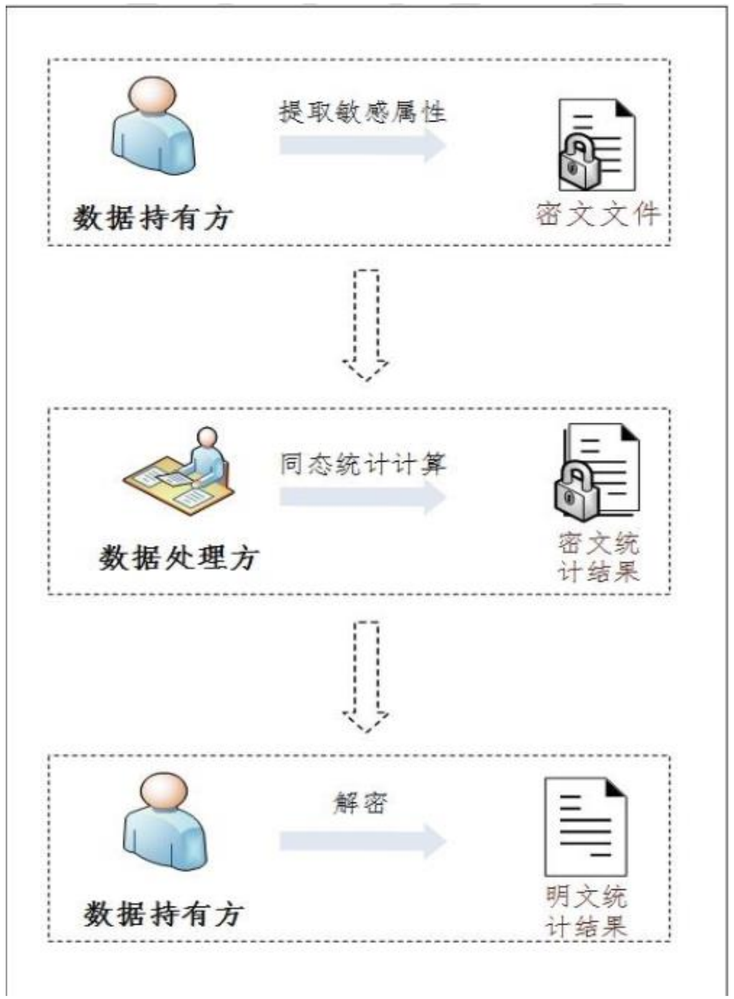
对数据需求方

(1)数据信息可获取 数据通过区块链发布并广播到每个参与方节点后，数据需求方可以在本地节点完成区块链检索、获取数据信息，简化了信息沟通流程。
(2)数据表达一致
通过对数据进行概念化、规范化的精确描述，链下的数据实体在 区块链上表达为数据身份对象。因数据身份对象具有唯一性和可辨伪 性，保证了市场上流通数据的可靠性。
(3)可引入第三方
区块链的联盟属性及智能合约能力，可支持消费者在数据流通过 程中自由加入第三方支持，如算法能力支持、结果投放能力支持等。 这为数据流通场景增强了扩展性，提供了更便捷的使用体验，同时更 益于构建稳定的数据生态。

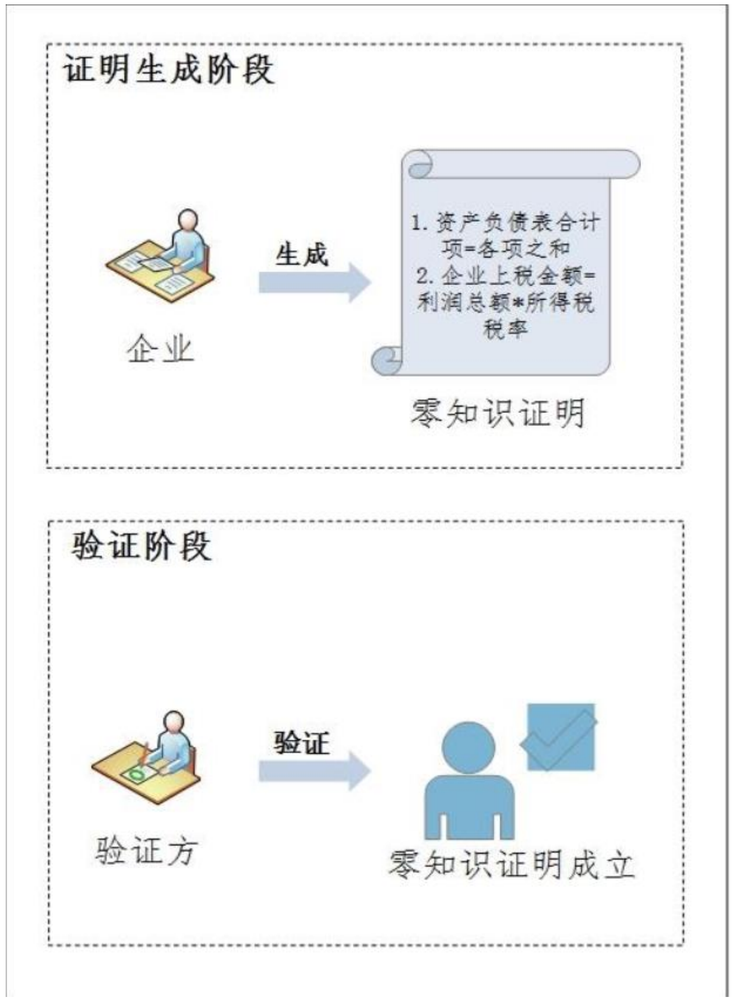
对数据需求方

(1)降低流通成本 区块链平台在线进行广播并达成合约，实现了最大程度的自动化，降低了数据流通市场运营成本。 (2)避免数据垄断
区块链可为数据流通引入广泛的市场参与者，实现有效的竞争生 态。区块链智能合约的灵活性还可实现数据的交叉融合，为不同数据 的自由组合及运用创造了条件。这两点都会有利于削弱市场中大而全的数据提供者的强势地位，从而促进数据的合理流通和高效使用。

03 数据安全技术工具

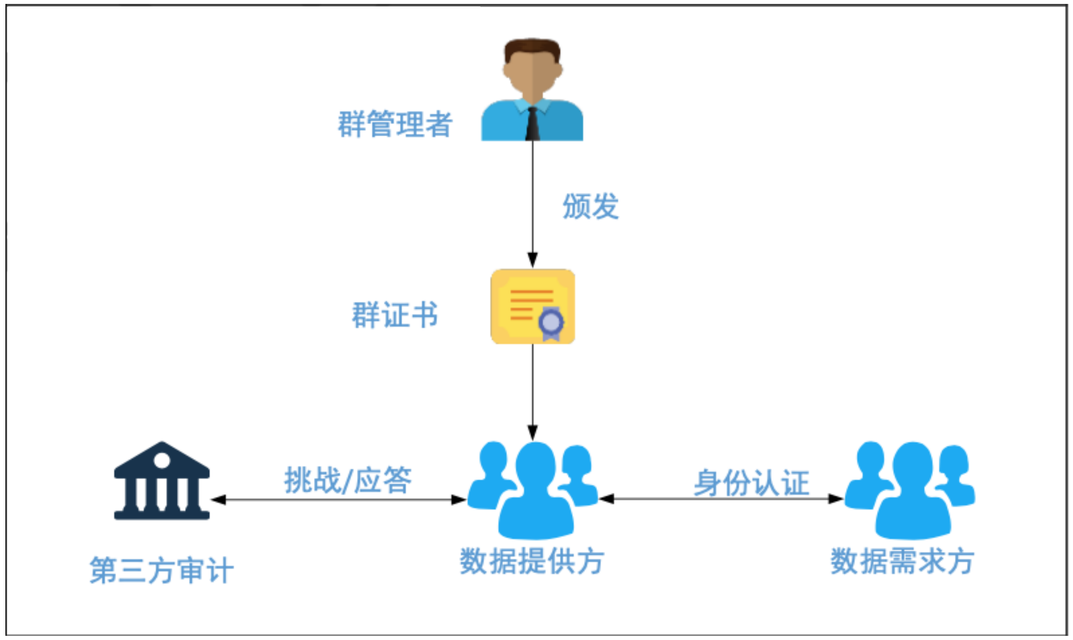


同态加密是指对其加密数据进行处理得到一个输出，将此输出进 行解密，其结果与用同一方法处理未加密原始数据得到的结果一致。

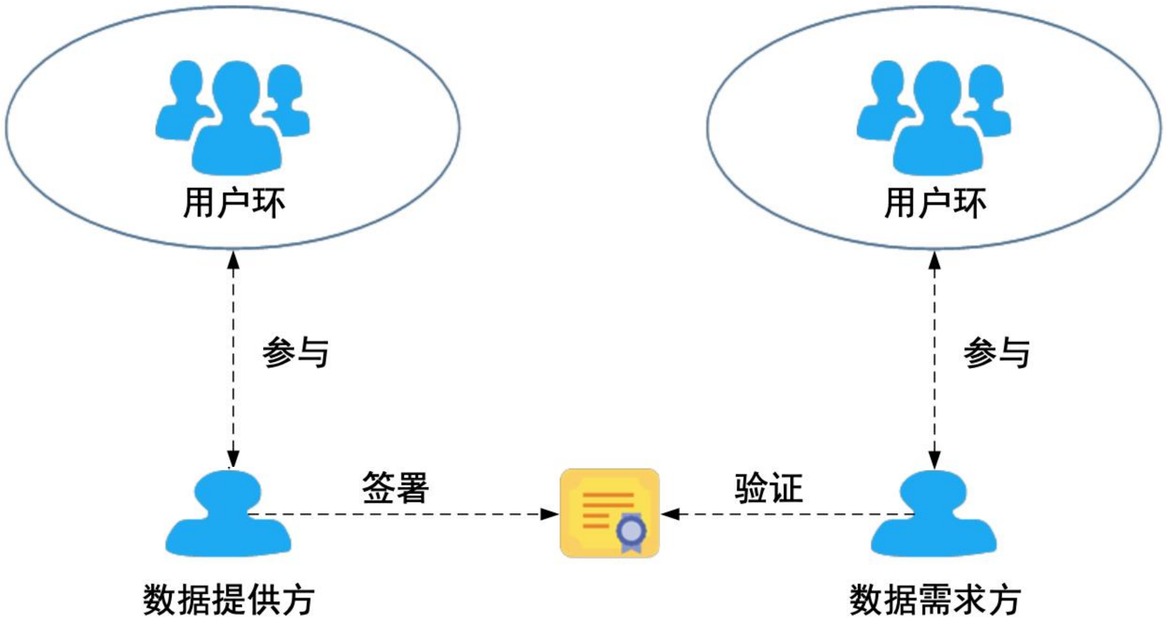


在一个零知识证明协议中，证明者向验证者证明一个声明的有效 性，而不会泄露除了有效性之外任何信息。使用零知识证明，证明者无需任何事件相关数据向验证者证明事件的真实性。

数据安全计算工具–群签名

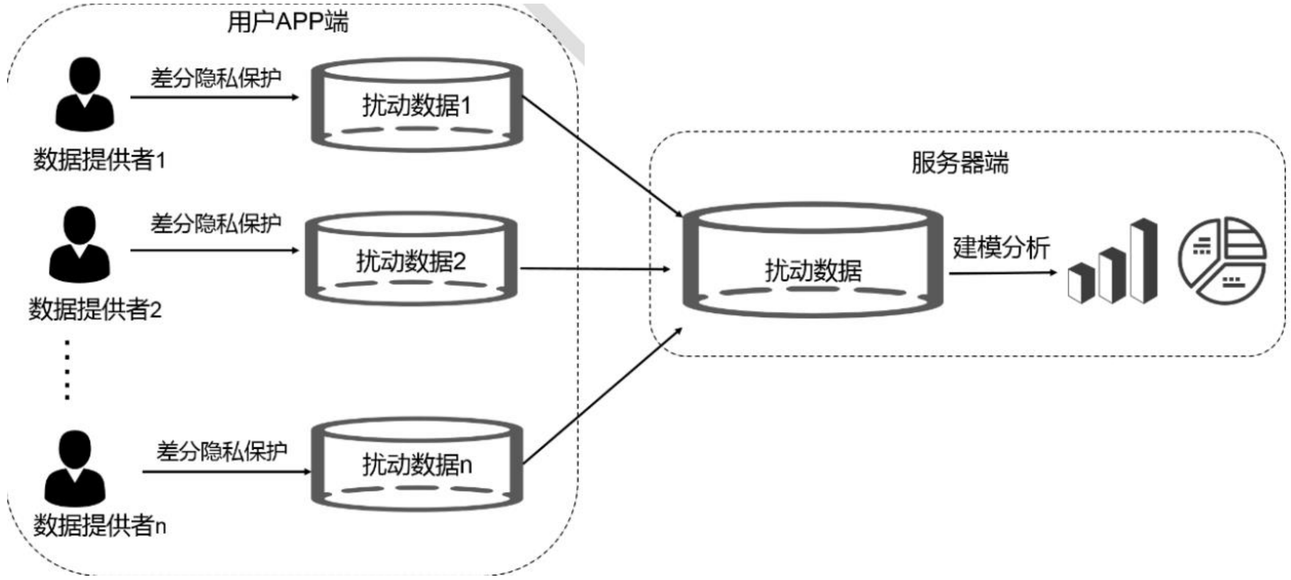


群签名技术是一种允许一个群体中的任意成员以匿名方式代表 整个群体对消息进行签名，并可公开验证的机制。
由于群签名能为签署者提供较好的匿名性，同时在必要时又通过 可信管理方追溯签署者身份，使得群签名技术在诸如共享数据认证，身份认证及金融合同签署等事务中，发挥重要作用。



环签名是一种简化的群签名，环签名中只有环成员没有管理者， 不需要环成员间的合作。在环签名中不需要创建环，改变或者删除环， 也不需要分配指定的密钥，无法撤销签名者的匿名性，除非签名者自 己想暴露身份。环签名在强调匿名性的同时，增加了审计监管的难度。

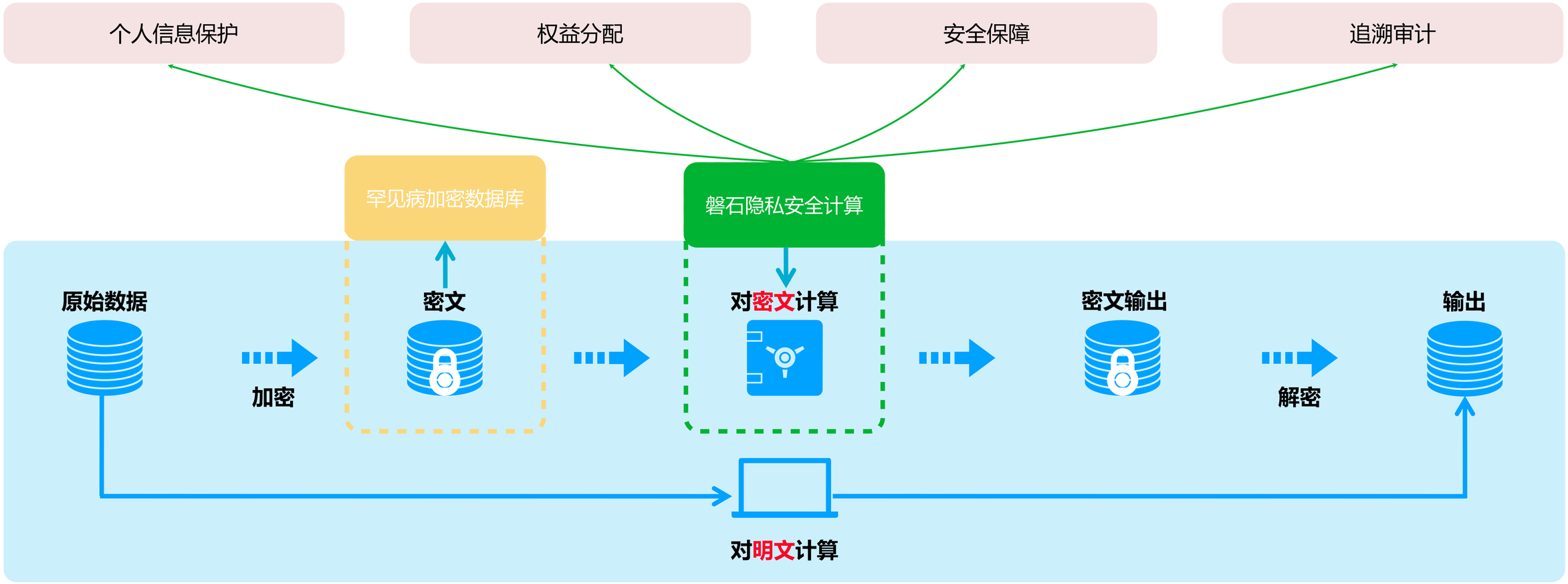
数据安全计算工具–差分隐私



差分隐私的基本思想是对原始数据的转换或者是对统计结果添 加噪音来达到隐私保护效果，相对于传统的隐私保护模型，差分隐私 具有以下两个优点:不关心攻击者所具有的背景知识;具有严谨的统 计学模型，能够提供可量化的隐私保证。

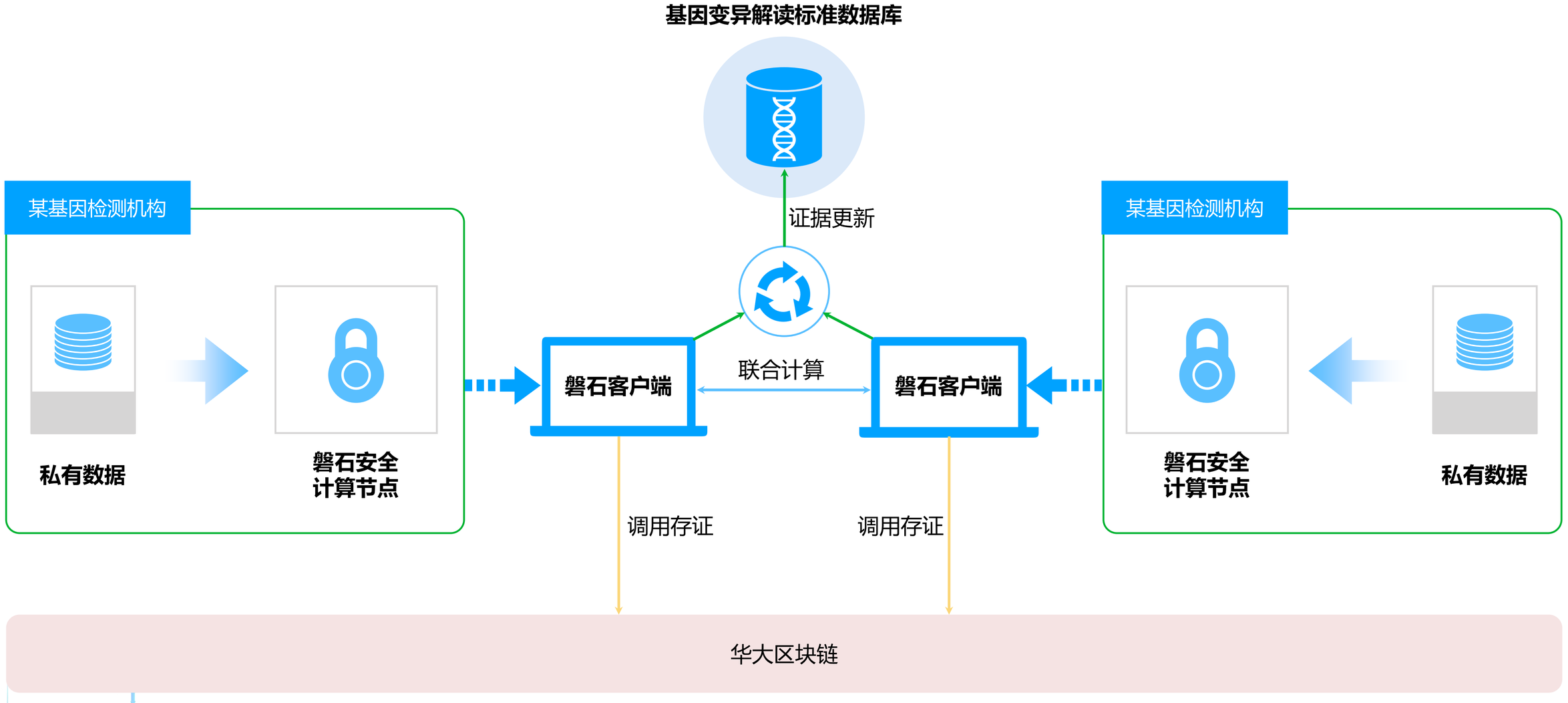
04 数据安全实践案例

基于区块链技术的罕见病加密数据查询



数据经过算法加密后，进行计算得到计算的输出结果，对这个输出结果进行解密，其结果与用同一方法计算明文数据得到的结果是一致的

区块链+MPC的联合数据分析



05 总结

要指出的是，这些技术并不是孤立存在的。每个技术本身可能 有不同的算法实现，技术的方案也会以使用场景的变化而变化。
如何在保护个人信息、商业秘密和国家安全的前提下，构建统一协调、多方协同、需求驱动、合规使用的数据流通生态体系，一直是 深化互联网和数字经济发展、释放数据红利的重要课题。

感谢聆听