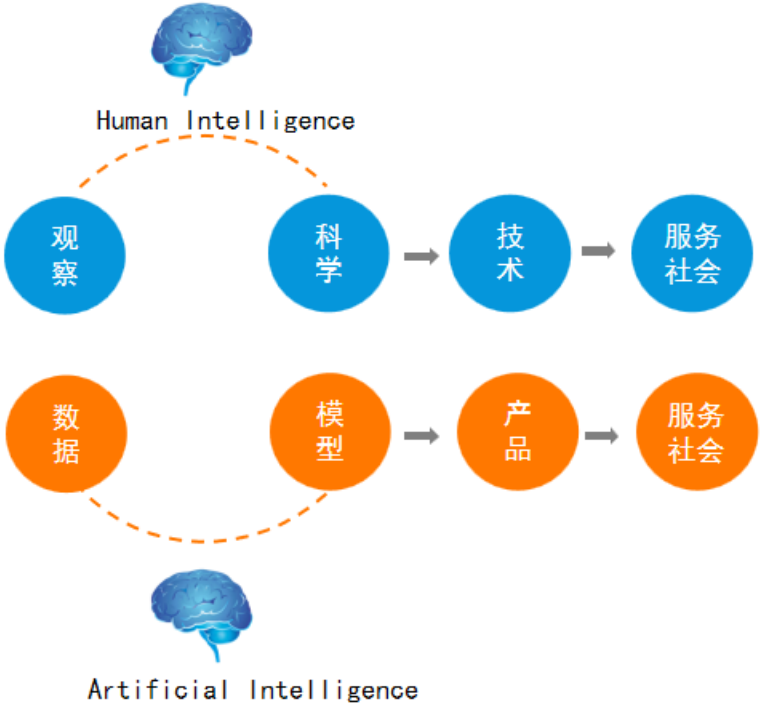
BaseBit.ai 翼方健数 | DataFun.

隐私计算，联邦学习，与数据原生时代的IT新基建

翼方健数 首席科学家 张霖涛

智能时代已来



人类有史以来第一次找到了第二个知识获取的平行途径：基于数据的机器认知

机器认知和人类认知的途径不完全重合，从而扩大了知识的绝对空间

机器认知具有强目的指向，从数据到服务社会的路径有可能更短，更高效

数据即知识
数据的流通就是知识的流通
数据的积累就是知识的积累

BaseBit.ai 翼方健数 | DataFun.

数据是智能时代最活跃的生产要素

生产率提高最快、对经济增长边际贡献最大，是社会资源配置围绕的中心、企业与国家竞争力的要害。

农业时代

最活跃的要素是**土地**

自然资源

工业时代

最活跃的要素是**资本**

自然资源过渡到人造虚拟资源

智能时代

最活跃的要素是**数据**

人造虚拟资源

数据-知识-智慧的价值生产链条迭代

数据伴生

数据伴随各类活动产生，具有“被动式、低价值、随机产生”的特点。特征通过对单一数据进行存储和分析，识别具体事件发展趋势并做出预测。

数据孪生

数据与物理空间相关联，通过数字化的方式，形成物理世界的映射，通过映射模拟预测趋势，辅助人类决策。该阶段物理空间“先感后知”，具有一定的滞后性。是将已有的知识应用于数字虚拟世界。

数据原生

物理空间本身就是信息空间，每个节点都是产生自我感知、自主运行的智能主体，物理空间“即感即知”。数据原生不同于过去，是生产人类认知之外的新知识。而数据价值的完整实现均可在终端实现。

数字排放，应用先行，人类决策

数据先行

数据作为一项新的生产要素的一些特点

- **信息时代的遗留物**
  - 质量参差、收集目的不同、非标准化的、非结构化的、相互隔离
- **独特的经济学特征**
  - 非竞争性的
  - 高昂的固定成本，低廉的可变成本
  - 外在性： 时间、环境、应用程序、网络效应
  - 可再用的
- **数据的非经济学维度**
  - 隐私、合规、机密、安全

中国产生数据的速度会在2025年超过美国

China's Datasphere on pace to becoming the largest in the world

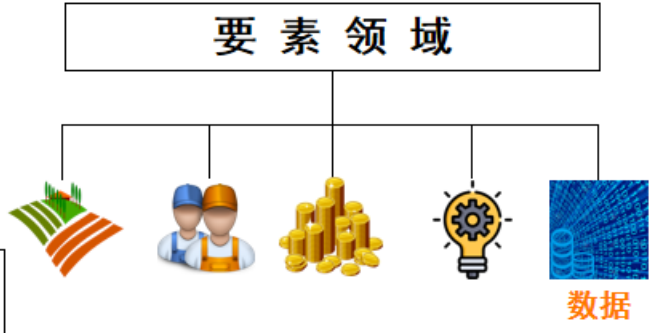
Every geographic region has its own Datasphere size and trajectories that are impacted by population, digital transformation progress, IT spend and maturity, and many other metrics. For example, China's Datasphere is expected to grow 26% on average over the next 5 years and will be the largest Datasphere of all regions by ... U.S., and R... population ... infrastructu... Asia-Pacific ... but not Chi...

DATA AGE 2025

The Digitization of the World
From Edge to Core

IDC

数据是国家竞争力的 战略资源



BaseBit.ai 翼方健数 | DataFun.

数据流通的现实困境

1

数据的经济学特征

作为一种新型生产要素和虚拟资产的独特性，尤其是非竞争性，在传统共享和使用方式下，有数据资产流失或者转移的风险。

2

数据安全

当前的国家法律中，对于数据产生、使用和流通范围等各阶段涉及的相关方及其权利划分模糊不清且衡量标准、过程复杂。数据协作过程中的安全无法受到技术保障。

3

个人隐私保护

数据被使用过程中，对个人隐私数据的保护。同时哪些数据需要个人授权使用尚未明确。

4

数据资产保护

企业数据在发挥价值过程中数据没有保障，数据及算法均受到此困扰，价值发挥受限。

BaseBit.ai 翼方健数 | DataFun.

数据流通安全：所有权和使用权的分离

数据所有权的复杂性

以个人就医数据为例，是**患者个人数据**，但原始数据是在医疗机构里产生的。从某种意义上讲，**患者本人**和**医疗机构**都有一定的道理认为自身拥有**数据的所有权**。

数据所有权不明确

国家并**没有**给出一个明确的法律规定，说明数据的所有权是谁的。

所有权和使用权可以分离

以房屋出租为例，房屋的**所有权**是指对房屋全面支配的权利，房屋的**使用权**是对房屋的实际利用权力。在此情况下，就将房屋一定时期内的**占有、使用权**让渡给承租人来行使。通过一定法律契约，非房屋所有权人也可获得房屋的使用权。

数据所有权和使用权分离

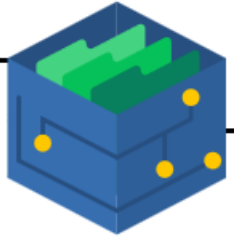
在数据的所有权未定的情况下，专注于数据的**使用权**（谁在什么场景下可以通过谁的授权使用数据）和**使用方法**（采用什么技术手段保证数据使用过程中的安全、隐私保护、资产保护）

数据安全管控及保障：通过安全技术实现

数据需要保护和隔离 vs 数据产生的价值在于联合计算和分析

数据本质上
不能安全流通

- 权责不清
- 可复制性
- 外部依赖性
- 隐私与机密需求



数据不流通
流通的是数据价值

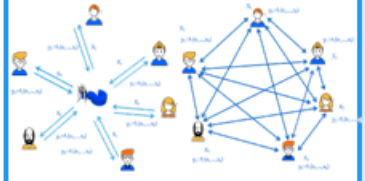
- 模型可应用
- 价值可流通
- 数据即知识
- 数据联合计算即新知识

BaseBit.ai 翼方健数 | DataFun.

传统意义上的隐私计算技术

多方安全计算MPC/同态加密

如果没有一个可信的第三方，如何让多个数据所有者共同参与，安全地完成协同计算



HOW:
1）采用秘密分享、混淆电路、遗忘传输等方法将软件转化为逻辑阵列后进行安全计算
2）密文计算

联邦学习

如何让多个相互不信任的数据拥有方不必共享数据的基础上联合进行模型训练

HOW:
1）横向：每个参与者在本地训练计算自己样本，只分享模型训练的梯度
2）纵向：各参与者训练各自的embedding，共同训练上层模型

安全沙箱计算/TEE

如果有一个，或者可通过硬件建立一个可信任的第三方，让多个相互不信任的数据所有者共享数据进行计算

HOW:
1）利用可信任执行环境TEE防止操作系统恶意地查看应用执行环境的内容
2）利用安全沙箱防止恶意应用通过特殊调用控制操作系统

差分隐私

区块链

Privacy Computing

对抗神经网络

零知识证明

仅考虑 **计算进行时** 的安全，不考虑 **数据全周期的安全** 和 **隐私保护。**

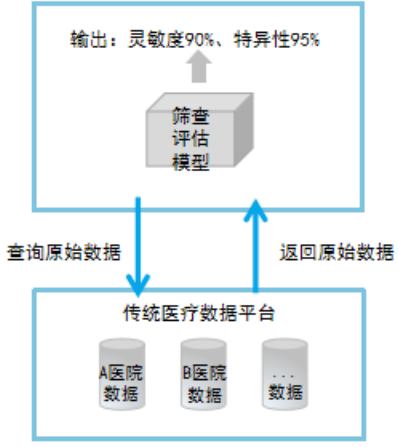
隐私安全计算

能够在**特定的信任假设**下
在**保护**数据所隐含的**隐私和机密**，
避免数据资产的**流失、转移和失控**的前提下，

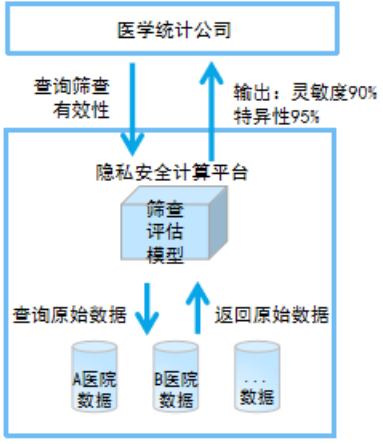
实现和分享数据价值的技术、产品和方法，
即为「**隐私安全计算**」。

隐私安全计算与数据流通的挑战

能够在**特定的信任假设**下在**保护**数据所隐含的**隐私和机密，**
避免数据资产的**流失、转移和失控**的前提下，
实现和分享数据价值的**技术、产品和方法，**即为「**隐私安全计算**」。



传统方式

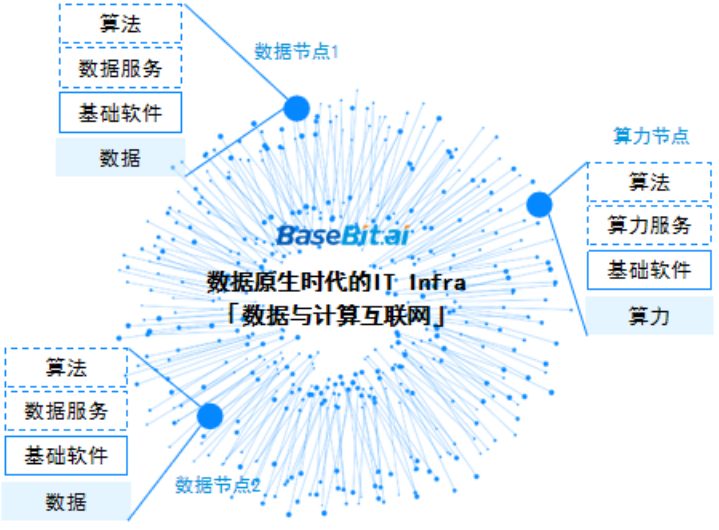


隐私计算方式

在隐私安全计算之外

- 应用方如何发现数据？
- 应用方如何获得数据的使用权？
- 应用方在获得数据授权之前，如何判断数据是否符合需求？
- 各个机构数据可能有不同的格式和字段，应用方如何有效使用这些数据？
- 应用方如何保证自己的模型IP不被平台管理员盗用？
- 如何使用“可用可见”的数据？
- 数据提供者如何信任数据在平台上存储安全？
- 数据提供者如何保证模型应用不将数据传出平台？
- 如何合理实现数据价值分配？
- 如何在训练好的模型不断对外输出服务的过程中保证没有数据隐私泄露？……

数字原生时代的新型基础设施



资产化的数据、算法、算力形成**基于隐私计算的「数据与计算互联网」**，正是**数据原生时代的新型IT Infra**。

翼方健数，是基于隐私计算的「数据与计算互联网」（“IoDC”）的建设者与运营商，致力于推动数据驱动的商业和产业，**实现数据价值，促成数据价值的流通**。

帮助数据源**实现数据价值**，是**节点广泛建立的前提**。而**数据要素资产化**和**隐私计算**是数据价值实现的必要条件。以商业价值促成数据拥有者、数据使用者和技术服务者良好协作。

数据原生IT Infra的挑战和机遇

数据治理与发现

如何高效完成适用于不同应用目的的数据治理，如何发现可利用的资产化的数据，如何让数据能够被发现，在数据不可见的前提下评估数据价值

价值实现

如何释放数据价值，将行业洞察应用和资产化的数据、算力、算法结合起来，降本增效、开源节流

隐私安全计算

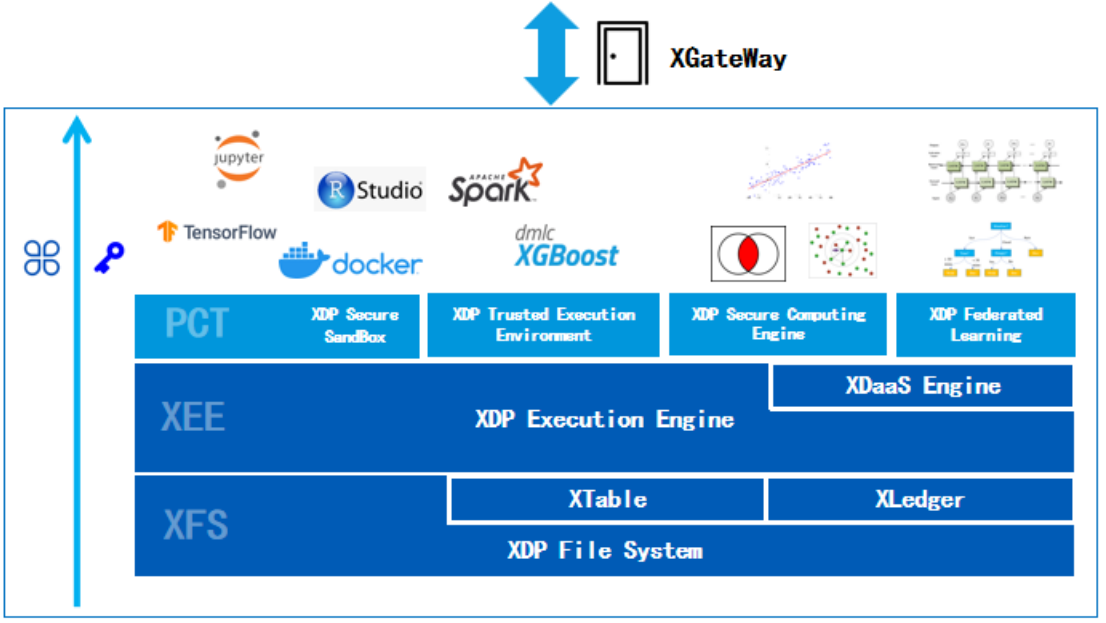
如何保证数据治理、发现、存储、计算和使用全周期的隐私和资产安全，如何保证数据的合法授权使用

异构资源配置调度

如何最有效地利用不同平台节点的异构存储和计算资源，如何支持不同的应用对不同类型资源需求的平衡

数字化转型的全栈IT技术矩阵

翼方健数的XDP技术架构



进出平台安控

数据价值实现

隐私计算技术

数据发现

计算资源调度

价值计量和分配

数据安全存储

XDP File System（XFS）：
为 IoDC 打造的分布式文件和数据编排系统

高规格的安全保护

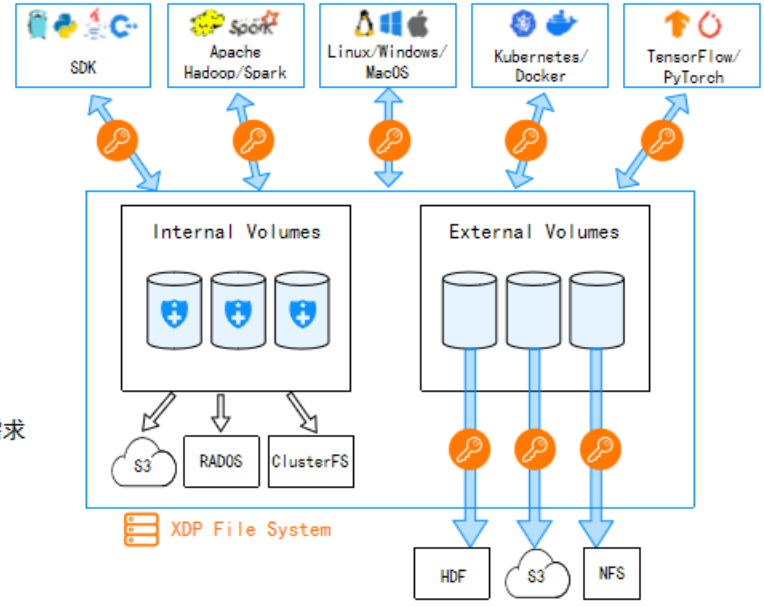
- 基于加密的数据保护和配套的**密钥管理系统**（KMS）
- 细粒度的访问控制保证数据**“最小可用原则”**

数据编排（Data Orchestration）能力

- 全面管理**IoDC**内的数据资源，高效对接外部数据源
  - 基于**POSIX**语义的分布式文件系统
  - 对接外部多种存储资源和I/O模式
- 不同计算方式的支持
  - 实现Spark、Hadoop、CSI等接口，满足多种计算方式的需求
- 实现**存储和计算的解耦**

高性能的存储效率

- 不同模式下顺序、随机读写高效率



BaseBit.ai 翼方健数 | DataFun.

XDP Execution Engine (XEE)：
为 IoDC 定制的计算资源适配与调度引擎

适配多种底层计算基础设施

底层算力资源的抽象和统一表达

云、私有化部署、混合云的支持

云原生的应用构建和部署方式

基于浏览器的多种交互方式

大幅提高平台开发/使用者效率

支持多种应用计算模式

如批处理、CS/BS、Big Data、云原生

兼容各类数据处理方式，零代码迁移

统筹管理调度IoDC下全网计算资源

结合XFS提供跨节点数据编排和计算能力

使“东数西算”成为可能

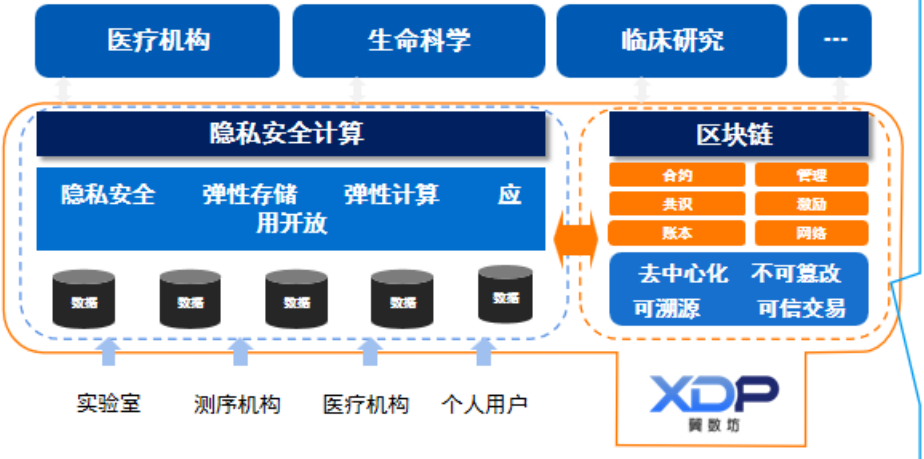
XDP DaaS Engine（XDaaS）：为 IoDC 打造的数据发现与整合

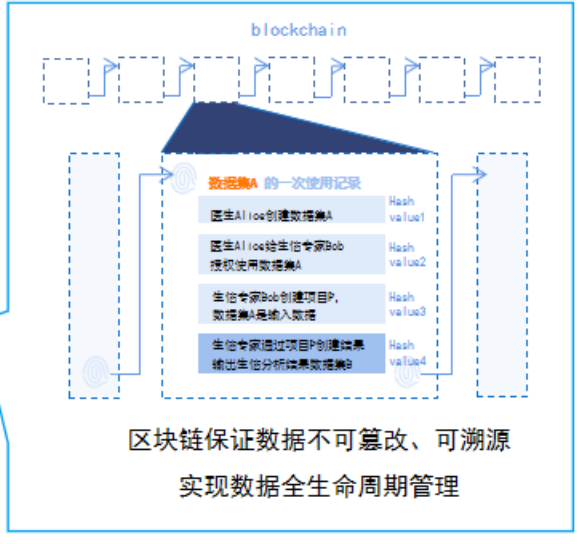
自主研发的XDaaS（Data as a Service）

- XDaaS提供可扩展的主数据和数据组织方式，将多源数据纳入统一的数据模型(common data model)之下，实现数据源间的有效融合
- 提供跨平台分布式高效的数据探查能力，为后续应用打造坚实的数据基础
- 采用差分隐私、加密查询等方式，保护原始数据安全，防止用户利用查询结果反推原始数据
- XDaaS在数据融合过程中实现cell级别的来源追踪，并提供细粒度的授权模式，进一步保护对敏感数据的使用

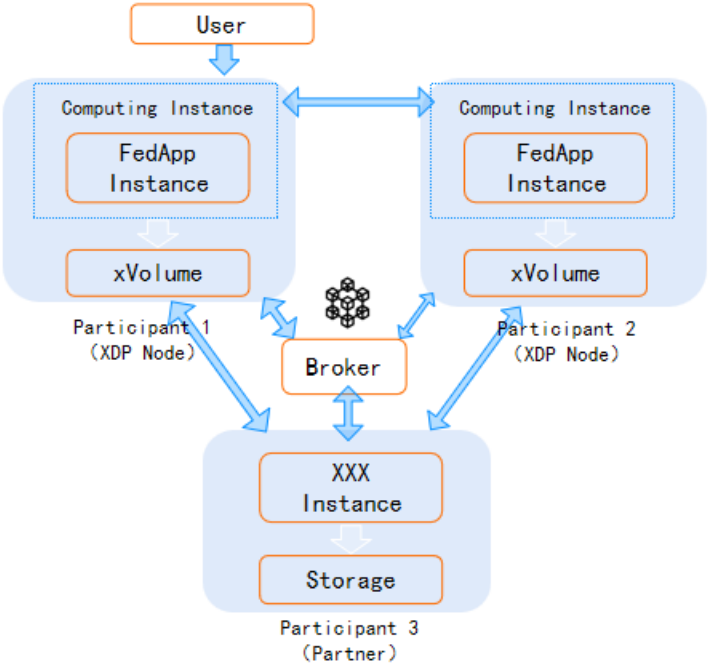
XLedger：隐私安全计算与区块链技术融合

为XDP联盟和IoDC提供不可篡改的数据存证与智能合约





XFederation：为IoDC订制的网络协作协议



节点互联

-节点认证

- 通过中心节点（XDP Fed Broker）或区块链认证、管理联盟内数据资源和计算资源
- 参照互联网层级设计，通过中心节点间的互联，在保证局域网络的管理下实现IoDC
- 各节点自主，中心节点仅限于证书签发、远程验证等基础认证服务

-节点发现

- 全网发现已经注册认证后的节点
- 针对节点能力定向搜索和连接
- 相互验证及多层证书为节点间通信提供信任基础和安全保障

数据互联

-联盟数据探查

- 构建联盟数据资源目录，提供全网数据搜索能力
- 分布式的数据查询实现全网数据探查，便于后续计算协作

-联盟数据授权

- 针对网络化信任假设提供多种数据授权模式
- 去中心化身份（DID）实现P2P场景下数据的授权使用

计算互联

- 通过XEE实现计算资源本地与远端的统一抽象并提供全网的计算编排能力
- 通过FedApp的封装抽象实现对联盟协议无感知的分布式计算
- 节点间点对点的任务分发管理实现自发的计算协作
- 支持其他框架，如FATE

BaseBit.ai 翼方健数 | DataFun.

XDP Privacy Computing Tech(PCT):为 IoDC 定制的隐私计算引擎

XSSBox

自主研发的安全沙箱

XDP Secure SandBox

应用之间，应用和系统之间的强隔离

XTEE

自主研发的可信执行环境

XDP Trusted Execution Environment

为单体平台提供“零信任”的本地计算环境

XSCE

自主研发的密文计算框架

XDP Secure Compute Engine

IoDC原始数据不出域，安全高效联合密文计算

XFL

自主研发的联邦学习框架

XDP Federated Learning

IoDC原始数据不出域，安全高效联合建模

多方安全计算

联邦学习

可信任执行环境

CAICT 中国信息通信研究院

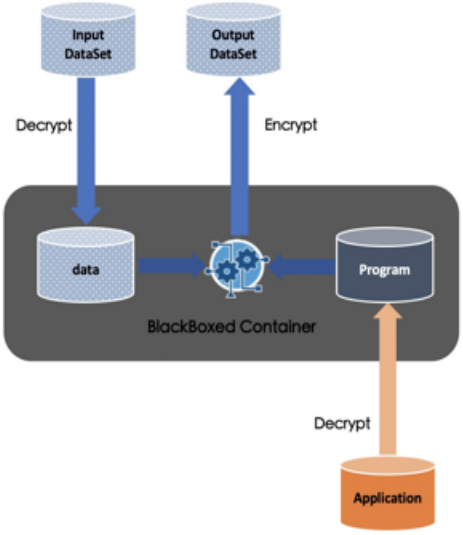
三大主流安全计算方式

都通过信通院认证

翼数安全沙箱(XDP Secure Sandbox)

为单体平台提供“零信任”的本地计算环境

- 不同于传统沙箱计算环境，翼数安全沙箱满足单体平台上“软件可信，用户不可信”的安全假设
- 安全信任根构建于系统管理员之外，充分防范平台运维的违规操作
- 通过计算存储分离、安全容器环境、软件定义网络等手段，实现平台用户间强隔离
- 确保对数据在运行时的保护，完成单体平台数据保护的闭环



翼数可信执行环境(XTEE)

提供基于硬件的安全、高效、通用的端到端可信执行环境

远程证明

- TrustZone/TPM等硬件构建信任根
- 定义并实现硬件无关的远程证明协议
- 为XDP平台内和平台间计算任务提供统一、安全的远程证明服务

TEE运行时

- 基于OCI容器镜像的统一计算抽象
- 实现或兼容加密VM、LibOS等多种云原生容器安全运行时
- 为用户提供有或无TEE硬件辅助等多种场景下的通用安全Enclave
- 全链路数据安全

加密文件系统

- 实现完全POSIX语义的本地和分布式加密文件系统
- 为运行在安全Enclave下的用户应用提供通用、可靠、高效的容器安全存储
- 最大程度地满足各类应用计算范式。

翼数密文计算框架 XSC

高效完备、灵活部署、集成开放的跨平台密文计算

高效、完备的算法库

算法全面：PSI/PIR/联合统计/特征工程/逻辑回归/线性回归/Softmax/CNN推断等；

性能优化：1）PSI算法亿级数据10分钟级；2）逻辑回归/线性回归100万样本/1千特征训练时长分钟级。

应用部署灵活

应用模式：支持应用端SDK接入计算节点；

计算节点：支持两方/三方计算节点部署；

安装部署：一键式部署/迁移。

集成开放

横向集成：集成MPC/FHE/ DP/ZKP 多种隐私计算技术；

垂直集成：集成数据治理/文件存储等上下层组件；

第三方集成：集成第三方硬件加速部件；

翼数联邦学习框架 XFL

- 海量隐私第三方数据
- 产生数据价值
- 海量模型库
- 海量训练插件
- 快速自定义训练流程

主流算法覆盖

- 横向算法 – 几乎全覆盖，可解决实际工程中绝大多数场景
- 纵向算法 – 在速度性能与模型损失上优于市面上已有框架

海量插件覆盖

- 提供大量模型训练插件，即插即用

高扩展性

- 联邦算法支持两方及任意多方，适应不同的联邦规模

高安全性

- 用户原始数据不出域
- 交换数据不可解密或不可反推原始数据
- 关键协议支持端到端加密

丰富的自定义接口

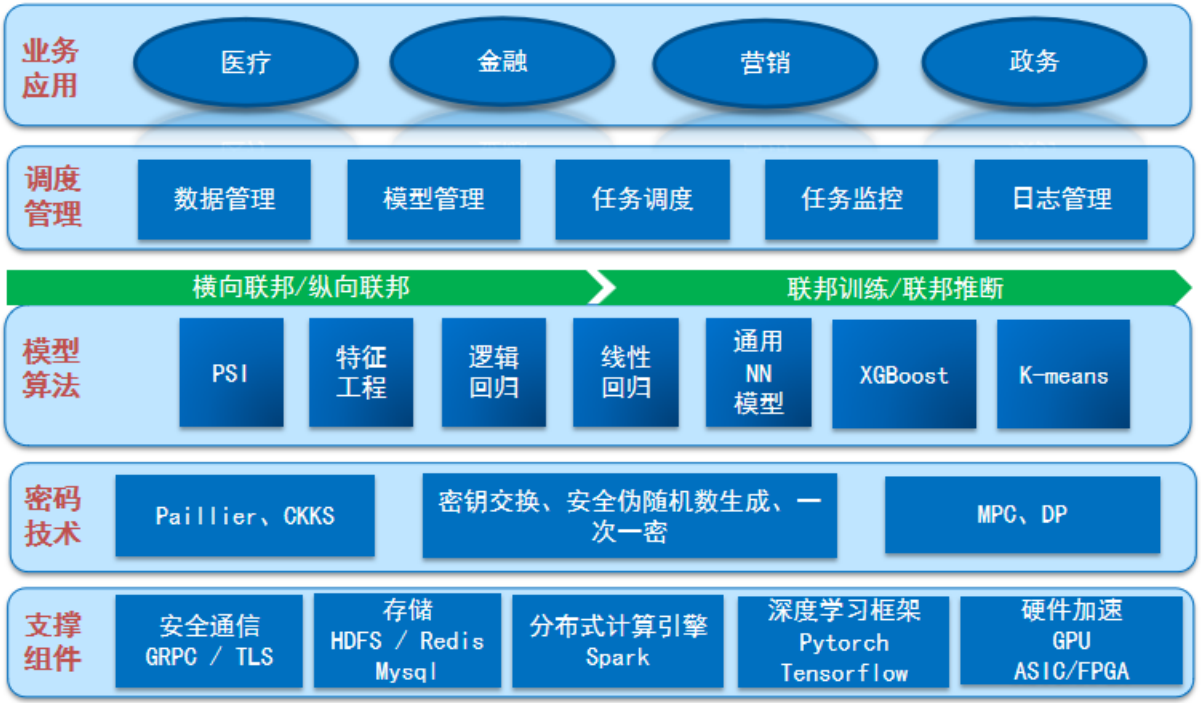
- 接口定义清晰，可依据任务需求进行快速开发

自主研发的联邦学习框架

在IoDC环境下安全联合建模

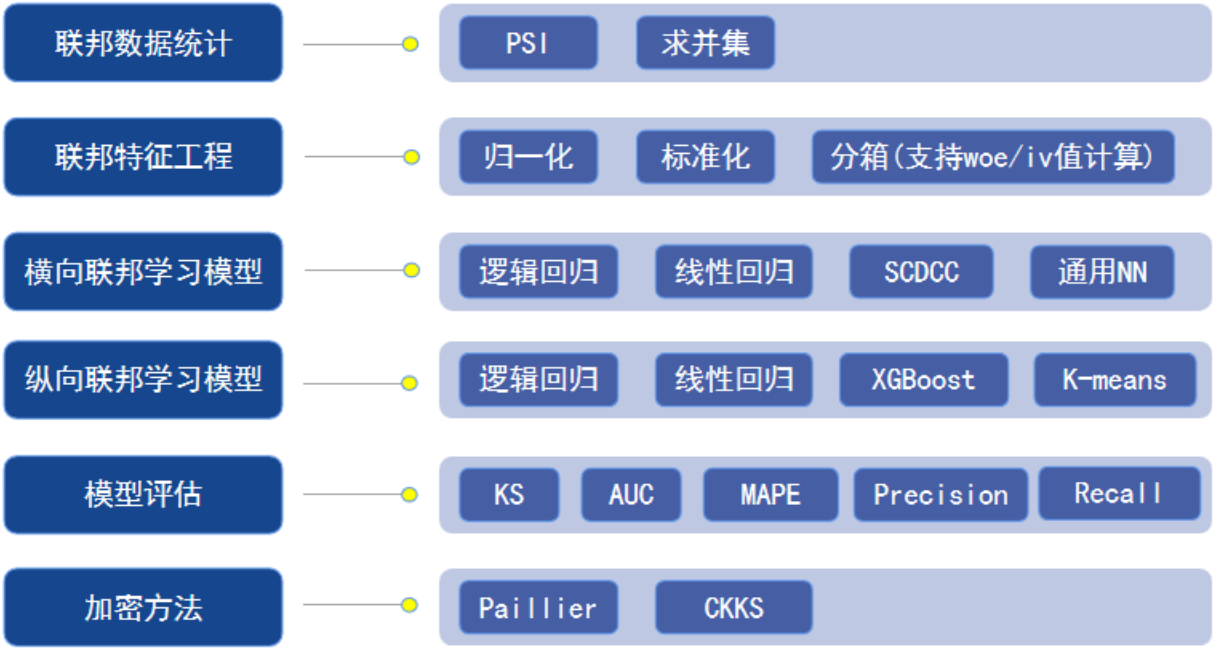
BaseBit.ai 翼方健数 | DataFun.

XFL架构图



BaseBit.ai 翼方健数 | DataFun.

XFL算法模块



BaseBit.ai 翼方健数 | DataFun.

BaseBit.ai 翼方健数®

成立于2016年1月，是**“数据和计算互联网”**的先行者，是一家专注大数据、人工智能和隐私安全计算的高科技公司。以隐私安全计算为核心，为**医疗、政务、金融、营销**等行业，建设在数据安全和个人隐私保护基础上的数据开放生态和数据共享协作环境，并在此基础上发展人工智能的能力，为行业赋能。

广州 · 南京 · 杭州 · 济南 · 西安

非常感谢您的观看

BaseBit.ai 翼方健数 | DataFun.