HUAWEI

践行深度用云

# AI可信数据空间白皮书

# 编制委员会

PREPARATION COMMITTEE

**编制单位**　贵州省数据流通交易服务中心
贵州大数据集团
贵州贵旅数网科技有限公司
华为云计算技术有限公司

（排名不分先后）

**编委顾问**　朱宗尧　张　广　肖　霏

**编写成员**　潘伟杰　金　凯　艾晓松　胡琼元　黄籽渝
赵　将　吕　东　宋　雪　胡　鹤　袁　波
吴　越　邓龙江　刘志杰　刘泥君　代新敏
王似巍　冯俊峰　吴忠林　李　勇　龙　婕
杨文敬　张洪能　尹舒鹤　李　媛　杨　松
张　中　黄　涛　陈　媛　黄　松　杨舒宁
徐　强　徐　俊　悦　怡　张志刚　霍战鹏
邝逸鹤　周　岩　王　立　乔丽娜　陈　勇
邴　丽　孙梦龙　陈　玓　杨梦辉　唐如兵
雷鸿伟　郑　辉　张小军　柏　君　孙思东
张鑫洁　唐　文　张　溯　陈吉栋　叶　飞
白文武　何志强

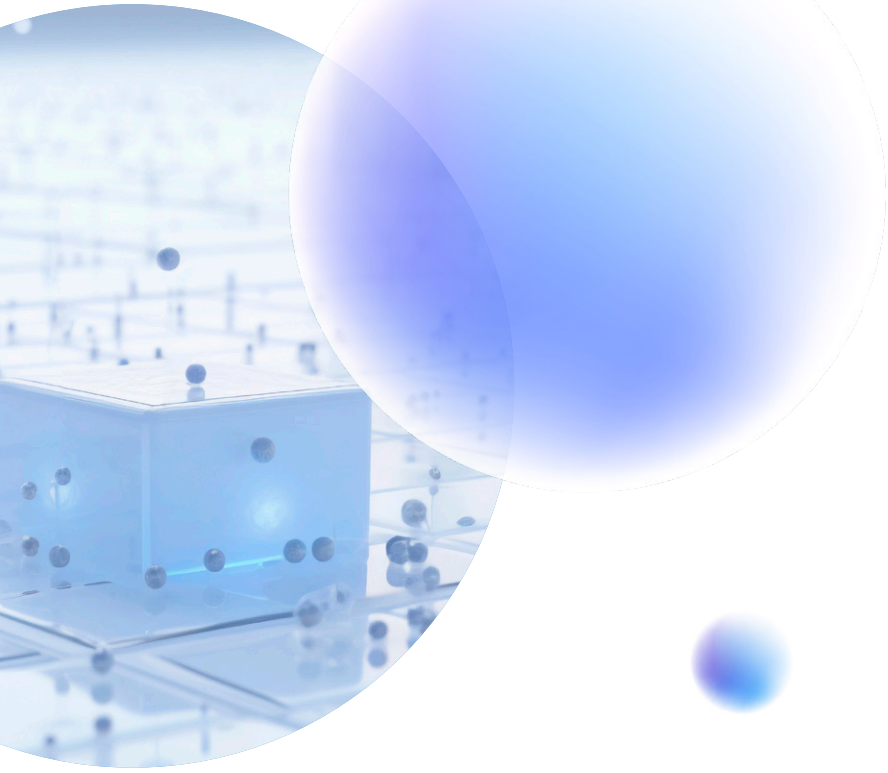
（排名不分先后）

# CONTENTS

目录

# 前言

# PREFACE



## 构建可信数据空间与人工智能协同创新范式，开启人工智能新时代。

在数字文明加速迭代进化的时代洪流中，数据与人工智能共生发展正重塑全球经济的格局。数据作为我国第五大生产要素，其价值释放的深度与广度，直接决定了人工智能产业从“感知智能”向“认知智能”跃迁的新高度。当前全球数据总量虽呈指数级增长（2023年突破175ZB），但高质量语料尤其是中文语料严重匮乏，实际流通率却不足5%。数据产业面临“不敢共享、不愿共享、不能共享”的困局如同无形枷锁禁锢着AI创新的步伐。比如医疗数据因隐私顾虑难以赋能疾病预测模型，工业数据因竞争壁垒阻碍供应链协同优化，金融数据因权属模糊制约风险管控精度，这些挑战的本质，是数据要素市场化配置过程中“信任基础”的不足，也将严重制约着人工智能产业的发展。

国家数据局重磅出台《可信数据空间发展行动计划（2024—2028年）》，以国家战略方式推动可信数据空间的加速建设，预计2028年国内将建成100个以上可信数据空间，包括城市、行业、企业、个人、跨境五类可信空间，打通数据要素流通的“最后一公里”。在AI时代，可信数据空间是战略级新型数据基础设施，它不是单一技术、工具的堆砌，而是制度规则、技术架构、生态系统三位一体协同的创新范式，将成为全域数字化转型的核心数字底座。

随着人工智能技术指数级增长，生产型AI场景爆发式应用，大模型对高质量数据的“饥渴需求”进一步放大了可信数据空间的价值。本白皮书结合新技术的理解与创新实践的探索，梳理可信数据空间与AI协同发展的技术路径、制度框架与实践蓝图，提出AI可信数据空间的顶层设计与创新架构，希望借此推动产业共识，联合各界探索数据要素市场化改革的灯塔，共同开启“数据可信即AI未来”的新纪元。

# 01

# 数据空间与人工智能协同发展挑战

## 1、数据空间发展与挑战

### 1.1 数据空间国内外发展趋势

国际数据空间 (IDS, International Data Space) 的概念最早起源于欧洲，旨在解决数据孤岛、隐私安全和权属不清等问题，推动数据的高效流通与价值释放，促进数字经济的快速发展。

2016 年，德国工业 4.0 战略率先提出工业数据空间的概念，随后欧盟推出《欧洲数据战略》，将数据空间建设提升至战略高度。美国、日本等国纷纷跟进，结合自身特点探索数据空间建设模式。截至 2024 年，全球已建成超过 200 个可信数据空间，覆盖工业、医疗、金融、能源、农业、交通等多个领域，实现了数据的安全可信流通与价值共创，成为数字经济高质量发展的关键基础设施。

欧盟是国际数据空间建设的先行者，其发展模式以联邦式去中心化为核心。欧盟通过《欧洲数据战略》、《数据治理法案》、《数据法案》等政策文件，强调数据主权、多方协同治理和标准化，GDPR (通用数据保护条例) 为数据保护奠定了法律基础。

在技术架构上，欧盟以 GAIA-X、IDSA 为代表，采用联邦式、去中心化架构。数据不集中存储，而是保留在数据源地，通过联邦学习、隐私计算等技术实现协同分析。此外，还利用区块链、分布式身份认证等技术，保障数据流通的安全与可控。通过连接器机制（Connector Mechanism），实现了不同系统间的数据互操作，提升跨域数据流通效率。

在生态与应用方面，欧盟已启动 14 个共同数据空间，覆盖工业制造、医疗健康、金融、能源、农业等领域。典型案例包括 Catena-X (汽车行业)、欧洲健康数据空间 (EHDS)、德国工业数据空间 (IDS) 等。这些数据空间推动了产业链上下游企业的数据共享与协同创新，为欧盟数字经济的发展注入了新动能。

全球可信数据空间建设发展日趋成熟，随着国家数据局的战略布局和重点工作推进，我国可信数据空间的建设已逐步从试点探索走向规模化建设。2024 年 11 月国家数据局发布《可信数据空间发展行动计划（2024—2028 年）》，首次在国家层面对这一新型数据基础设施系统布局，明确到 2028 年建成 100 个以上可信数据空间的目标。


图 1 可信数据空间建设发展历程

我国可信数据空间的建设可以分为四个主要阶段:

· **第一阶段: 建制度 (2020-2022 年) :** 国务院陆续发布《关于构建更加完善的要素市场化配置体制机制的意见》、《关于构建数据基础制度更好发挥数据要素作用的意见》等文件,明确数据列为新型生产要素,提出“三权分置”等制度, 形成基础的制度框架, 明确数据要素使用规则, 为数据要素市场化配置奠定基础。

· **第二阶段: 立顶设 (2023-2024 年) :** 数据局负责完成发布《数字中国建设整体布局规划》, 明确构建全国一体化数据资源体系, 提出让数据“供的出、流得动、用得好”的顶层设计方案。

· **第三阶段: 强行动 (2024-2025 年) :** 陆续发布《“数据要素 ×”三年行动计划 (2024—2026 年) 》、《公共数据资源授权运营实施规范 (试行) 》等管理要求, 推动数据要素在各行业的应用与价值释放, 明确从登记→授权→定价的完整链条。同时发布《可信数据空间发展行动计划 2024-2028》明确可信数据空间定义与定位, 提出 100+ 可信数据空间建设目标。

· **第四阶段: 促发展 (2025 年 -) :** 数据局发布《2025 年可信数据空间创新发展试点名单》、数标委发布《可信数据空间技术架构》, 进一步加强牵引各类数据空间的加快建设。

### 1.2 可信数据空间主要挑战

可信数据空间作为国家数据基础设施的重要组成部分, 作为全域数字化转型的新型底座通过构建安全可信的数据流通利用环境, 促进数据资源的共享共用, 进而释放数据要素价值。国家数据局发布行动计划明确推进企业、行业、城市、个人、跨境五类可信数据空间建设, 可信数据空间建设目前还处于发展初期面临诸多堵点问题与挑战:

· **挑战一、数据供给意愿不足**

①权属不清与价值分配难: 数据产权界定模糊, 数据供给方担心共享后失去控制权或收益被稀释。例如, 担忧数据泄露导致核心竞争力丧失;

②合规成本高: 数据分级分类标准不统一, 敏感数据 (如医疗、金融) 脱敏处理需专业团队, 数据供给方难以承担人力与资金成本。

· **挑战二、数据流通效率低**

①跨域系统数据协议不统一: 政府、企业、行业数据分散在异构应用系统中,数据结构、接口标准不统一, 跨域流通效率低;

②安全与实时性矛盾: 为满足“数据不出域、可用不可见”安全要求, 隐私计算 (如多方计算、同态加密、联合分析) 导致计算性能延迟增加 50% 以上, 难以支撑高价值场景 (金融、物流、交通) 等低时延响应要求。

**· 挑战三、高质量语料稀缺**

政府、金融、医疗、制造等行业大模型专业语料稀缺，海量多模态数据（文本、图像、传感器）待标注数据占比高，高质量语料转化率低（语义缺失、时效滞后等），无法满足行业大模型训练推理阶段对行业标注数据的诉求。

**· 挑战四、安全能力参差不齐**

数据流通涉及数据提供方、使用方、服务运营方等众多参与主体，不同主体的网络安全、传输安全、数据安全等防护水平差异较大，容易成为攻击者突破的薄弱环节，无法满足全链路数据安全防护要求。

## 2、人工智能大模型语料发展与挑战

### 2.1 人工智能大模型语料发展趋势

**1. 从大语言模型到多模态 / 具身智能大模型的语料演进**

当前大模型技术正经历从弱人工智能（机器学习、神经网络、大语言模型）向通用人工智能（Agent、多模态、具身智能）的范式跃迁。这不仅对于模型架构的设计理念进行了重构，同时也对大模型各阶段的训练语料提出全新要求。

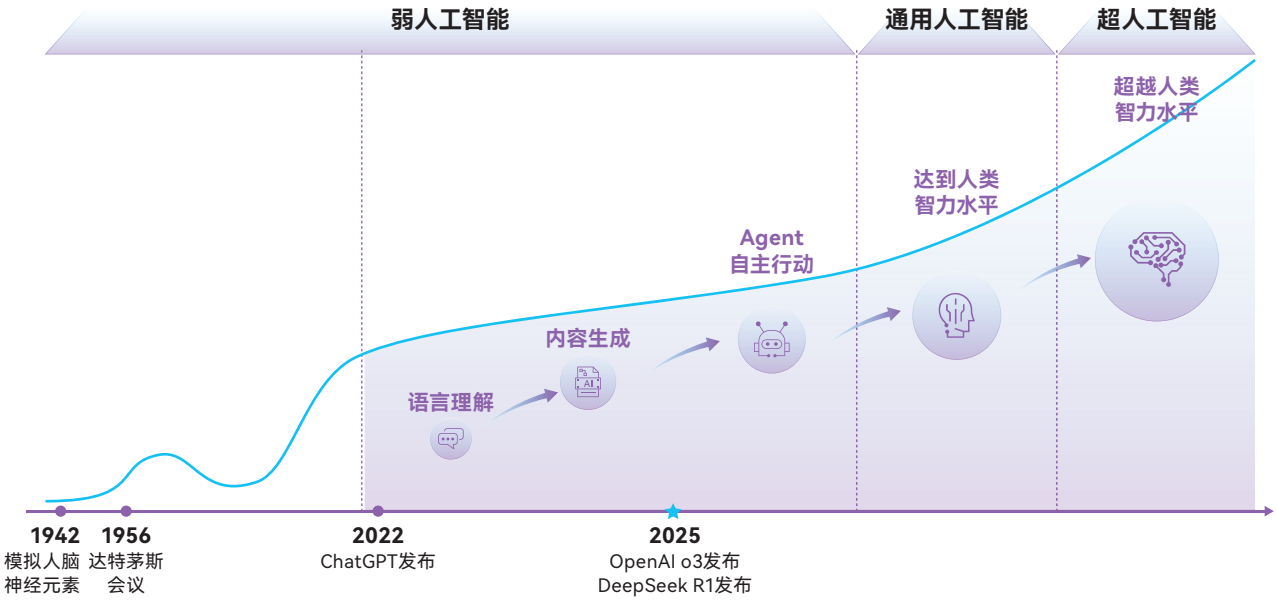



图 2 大模型技术发展演进阶段

首先，对多模态关联的复杂语料需求日益迫切。针对大模型语料质量，传统纯文本数据已经无法支撑多模态与具身大模型联合建模的需求。业界实践表明，在多模态场景下的语料供给，需要进行跨模态语料数据精准对齐，实践通过融合图文信息，运动轨迹，场景数据，使得复杂论文的解析准确率提升 37%。此类多模态场景在处理包含图表、公式的复杂文档时，需要在语料标准中体现语义逻辑关联能力，同比文本标注对于数据融合要求更高。

其次，语料规模需求同样呈现量级扩张。根据行业报告分析，多模态（视频）语料在智能监控领域的渗透率从 2023 年 18% 提升至 2025 年 35%，例如特斯拉自动驾驶系统当前需要超 20 万小时多模态场景数据进行模型能力提升。此外，具身智能模型更需要空间动态语料，比如物体运动轨迹、各类参数信息等。据求思咨询报告分析，全球 AI 语料市场规模预计在 2025 年突破 109 亿元。并且数据合成成为当前突破现在数据规模的主流路径，预计 2025 年合成数据在 AI 训练中占比将达 40%，例如工业数字孪生场景中，合成图像替代率从 30% 升至 65%。

同时语料数据的内容可信度保障面临更高要求。多模态语料的质检需建立跨模态的数据质量检验流程机制，避免图文内容不一致从而导致模型“越训越差”，造成模型认知偏差。再例如，具身智能模型将与物理世界进行高度交互，语料内容中需要包含安全边界参数（如机械臂操作力度阈值），确保与现实环境的适配。

**2. 从通用模型到行业大模型的语料深化**

随着通用模型发展从试点到落地，具有行业属性的行业大模型也逐步成为模型能力提升的重要转变，意味着 AI 技术从广度覆盖转向行业深度赋能，而行业大模型的性能则更是高度依赖领域的高质量语料。

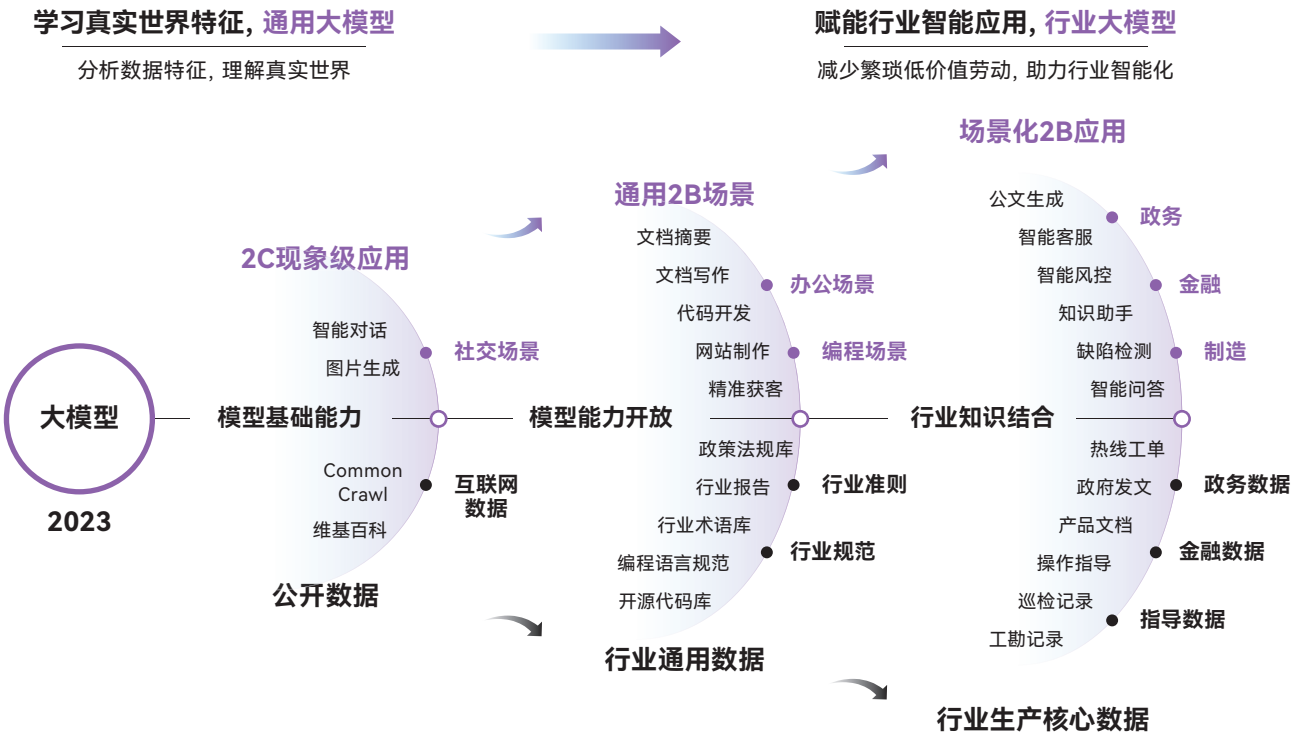



图 3 通用大模型到行业大模型的数据需求演进

语料质量: 随着通用语料的逐步耗尽，在高质量行业数据上的加工与使用成为各模型能力差距的主要原因。例如上海规划资源专项语料库，通过归集 1200 份技术标准、5.7 万份城建档案，建立“专家标注 +AI 校验”双重标注模式，可以实现地图信息的精准规划，从而代替通用语料仅能提供政策咨询渠道等简单能力，成为垂类模型的“高配置语料库”。

内容可信: 语料构建需要兼顾内容合规与机密性。尤其在医疗行业，领域语料常涉及大量个人隐私信息，医疗记录，企业信息等敏感内容，因此对于数据内容的隐私保护，催生出隐私计算技术的逐步成熟与落地。比如每日互动公司推出的 GAI Station 智能工作站，采用“本地小模型 + 云端大模型”架构，通过将企业内法务、财务等数据进行向量化处理，并结合联邦学习技术，使得跨机构数据协同合作的数据泄露风险降至 0.001% 以下。

### 2.2 人工智能大模型语料主要挑战

当前随着模型尺寸及模型场景的不断发展，语料的端到端建设与工程化能力也面临着挑战，具体展开为公开数据即将耗尽、领域数据流通困难、多模态对齐与合规性制约等三大维度，使得高质量数据资源成为模型能力提升的关键瓶颈。


图 4 人工智能大模型语料面临的关键挑战

公开数据即将耗尽: 随着智能化进程逐步推进，不同语种的语料资源质量与规模差异巨大，在全球数据训练集中，英语等主流语言拥有海量高质量数据，中文语料数据的占比仅为 1.3%。Epoch 研究表明，基于当前语料消耗速度，预计将在 2026 年，现有的公开高质量语言数据即将耗尽。

领域数据流通困难: 近年来，在数据流通、数据共享、数据开放已有很多先进探索，但依然存在 “主动找数、被动供数”，高质量领域数据的流通性问题仍然没有得到解决。虽然私域数据的专业性、可靠性、准确度可与行业场景更好适配，但专业领域知识积累门槛高，周期长，数据隐私要求高，所以领域知识的共享在实际落地上还存在难度。

多模数据对齐与合规制约: 不仅多模态数据的对齐与融合、数据合规流通等受制于技术的发展，在数据产权确权、数据资产入表等方面也需要通过政策法规制定数据加工的标准与规范。

## 3、数据空间与 AI 协同发展的关键挑战

在当前数智化转型加速的时代，数据与人工智能的协同创新成为推动各行业数字化、智能化升级的引擎。然而，这一融合过程中面临多重挑战，“数据壁垒、隐私合规、技术异构性、信任机制不透明”成为数据驱动 AI 创新的关键瓶颈，Data+AI 协同创新存在“三不可”的核心挑战。

### 3.1 挑战一、数据 AI 不可见

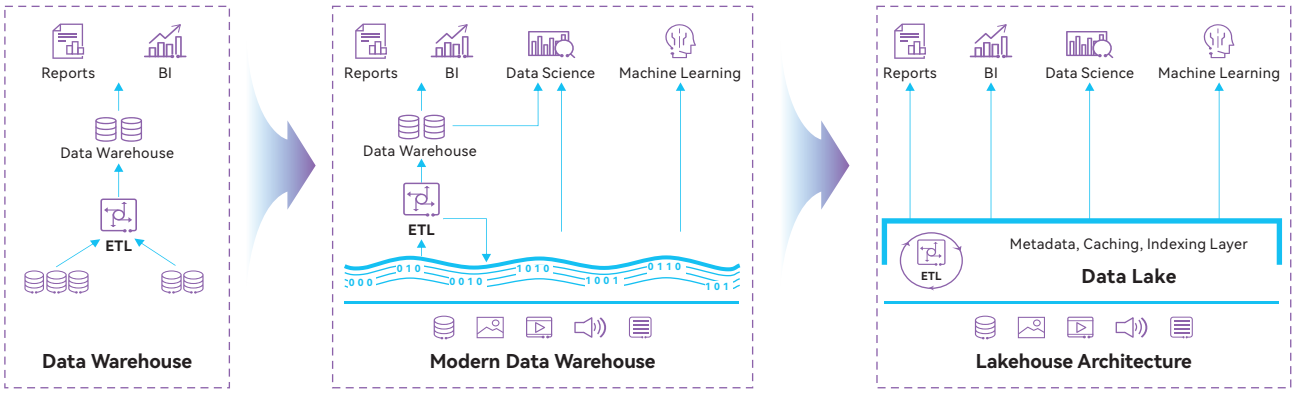



图 5 数据湖仓的架构演进发展历程

在过去的数十年时间内，各行业客户建设了大量的数据库、数据仓库、大数据等系统，形成了非常复杂的数据架构和庞大的数据资源体系。但由于上一阶段的数据建设通常是围绕交易、分析等特定目标建设的系统，从而逐步形成了大量的数据烟囱 (Data Silos) 架构，数据的协同共享面临诸多断点和壁垒，无法满足 AI 大模型时代对全量用数、高质量供数等新要求，因此 Data For AI 用数核心诉求是让全量数据“AI 看得见”。

**· 多形态异构技术体系，跨系统数据 AI 不可见**

当前模型训练 / 推理数据涉及多主体、多云、多业务系统间的数据集成与数据汇聚，由于多主体大数据平台建设采用异构技术平台、不同元数据管理、独立数据接口标准，导致跨主体跨域间数据无法高效流通、无法高效发现、汇聚、治理、加工、访问与检索，大量高价值数据难以被快速、高效集成至 AI 训练 / 推理生产流程中。

**①跨域多主体系统异构建设，跨系统横向集成难，无法为 AI 统一供数：**

• **缺少统一元数据管理，治理复杂度剧增：** Iceberg、Hudi、Delta Lake、语料集对象存储各表格式的元数据访问协议差异大，多个数据湖独立部署元数据管理，异构数据湖无法形成全局数据地图，跨系统数据发现成本高、血缘追踪割裂等。

• **缺少统一数据格式，跨格式数据计算性能差：** 多引擎 (Spark、Flink、模型训推等) 编排场景下兼容性差，跨格式联邦查询通过多 catalog 路由，查询计划复杂、性能差。

• **缺少统一权限管理，合规管控复杂度高：**异构元数据管理多重权限体系叠加、审计日志分散，满足统一 IAM 策略、统一审计的技术成本高、漏洞风险大。

**②缺少云边端一体化管理，中心训练 -> 边缘推理纵向数据供给不足：**

• **云边端数据孤岛导致语料碎片化：**大量高价值行业数据（工业设备、医疗检测、城市治理监测）滞留边缘端，导致垂直领域语料严重短缺；

• **云边端元数据描述不统一：**云边端数据缺乏统一元数据管理和存储标准，导致语义对齐困难，需要大量人工治理，高质量供给成本高、难度大；

• **云边端多级数据权限割裂：**大量高价值端侧生产数据，缺少统一的权限管理和脱敏管理，合规使用风险大。

### 3.2 挑战二、数据 AI 不好用

语料数据是大模型训练的重要“燃料”，但大模型语料因为“数据分散、质量参差不齐、多样性不足、过拟合风险”等问题存在数据 AI 不好用困境。基于私域行业数据加工高质量语料供给大模型训练与推理，需系统性解决清洗、标注及优化等高质量语料要求的核心问题。

**· 数据质量差，AI 不好用**

①数据噪声与错误：大量的语料存在拼写错误、语法错误、乱码、隐私非合规、内容非合规、重复内容、低质内容等，并且自动化和智能化去噪程度低、成本高、准确率低。

②标注质量、自动化程度低：行业标注标准不统一（例如医疗影像 - 磨玻璃影有多种定义）、专业人员参与人员不足（海量医疗影像，需要专业医学专家参与标注）、人工标注成本高，准确率低。

③数据时效性差：行业语料、特定领域知识库的词汇和术语未及时更新，无法覆盖新事件与新知识例如行业政策法规语料停留在“2024 年 9 月”。过时数据引发“假事实”、加剧模型幻觉。

**· 多模态数据碎片管理，跨模态语义难对齐，AI 不理解**

长期分散存储和管理大量结构化数据（OLTP/OLAP）、多模态数据（文本、图片、视频、语音等），导致多模态大模型训练与推理语料供给不足。

①数据特征异构性导致语义对齐失效：同一对象的跨模态描述难以对齐，例如“医疗报告的文本描述与 CT 图像的病症的映射关系不一致”；

②多模态元数据管理割裂，跨模态检索准确度低：不同数据系统的元数据标准不一致，无法跨数据系统的元数据多模态样本自动关联，例如“IT 系统身份证 ID 体系与人脸照片模式识别自动关联”。

**· 专业领域知识的理解瓶颈，高度专业化术语与上下文缺失**

①高价值结构化数据关联断裂：传统高价值结构化数据集缺少跨表跨字段的语义关联，结构化数据集的离散型导致大模型难以构建实体间的语义联系，例如大模型无法理解外键语义、无法理解多字段含义关联等；

**· 多时态数据采集不全，不满足 AI 对实时数据的关键诉求**

针对分析目的的数据系统建设，通常对历史数据、离线数据或者准实时数据进行了采集分析，但 AI 从训练进入到生产推理、智能体 / 智能应用投产阶段，对实时数据的采集和反馈尤为重要，尤其是针对制造等工业领域大模型来说，获得实时数据的反馈对提升模型精度至关重要，直接关系到 AI 在生产应用的实际效果和价值收益，因此 AI 进入生产阶段，数据平台需完成对实时、准实时、非实时全时态数据的全量采集和治理。

### 3.3 挑战三、数据 AI 不可信

由于多方协作对全链路、全过程的数据安全要求高，在数据语料供给方与行业大模型消费方会存在多方身份不可信、数据来源不可信、数据使用过程不可信、数据传输不可信、数据 AI 不可控等关键挑战。

• **参与方身份不可信：** 是阻碍数据要素高效流通的核心瓶颈，其本质在于传统身份验证机制无法适配数据跨域流通的复杂场景。一方面，不同机构 / 平台采用异构身份管理系统，缺乏统一的元数据标准和互操作协议，另一方面，动态身份管控难，数据离开持有方安全域后，缺乏对参与方行为的实施检测与权限回收机制，身份信任链断裂。

• **数据来源不可信，** 数据完整性受损及责任追溯困难。一方面，数据易篡改，数据生产链权责模糊，多主体参与导致源头数据被伪造或污染。另一方面，追溯机制缺失，缺乏全链路审计技术，数据泄露或滥用后难以定位责任主体。

• **数据使用过程不可信：** 也是根源性技术瓶颈之一。现有的 CA 体系仅能验证机构实体的身份，无法对数据应用实体（如虚机、容器）、计算环境进行可信认证，导致参与方对数据使用过程中执行环境真实性和安全性存疑。

• **数据传输安全不可信。** 一方面，传输协议存在安全缺陷，传统加密协议（如 SSL/TLS）存在中间人攻击风险，攻击者可以伪造证书截获明文数据。另一方面，跨域管控机制缺失，导致数据跨域流通时不能及时阻断数据截取行为，导致数据泄露事件。

• **数据权限策略不可控。** 多系统权限孤岛与多角色协同失效，导致多系统（如标注平台、训练集群、推理 API）权限策略不互通；权限粒度与效率的冲突，导致过细管控“按数据字段或操作步骤分权（如 ABAC 模型），导致审批流程冗长，拖累治理效率”；过粗管控“角色泛化（如“训练工程师”角色）可能赋予过量权限，增加误操作或恶意泄露风险”。审计链条断裂导致数据从采集到输出的全流程涉及多角色（标注员、算法工程师、运维人员），权限变更日志分散在各系统，无法关联分析。

02

# AI 可信数据空间
# 顶层设计

围绕数据空间与 AI 协同发展的“三不可”关键挑战，应对数据、人工智能产业需求、场景、技术快速变化的不确定性，迫切需要一揽子迭代升级的顶层架构和体系设计来应对未来的不确定性。全新升级的“三位一体”的顶层设计包括新体系、新架构、新生态三大部分：


图 6 AI 可信数据空间“三位一体”顶层设计架构

**① “数模协同”新体系**

从传统大数据时代的行业数据治理，到 AI 时代的数据工程、知识治理和 AI 智能用数的发展与变化，应对数据要素流通和大模型高质量可信供数，需要建设一套更为完善的“数模协同”体系，包括制度体系、技术体系、价值体系的深度协同。

**② “一湖一链一中枢”新架构**

通过 OneLake 技术架构升级，打造一个融合数据湖 + 一条端到端智能数据工具链 + 一个数据空间中枢的“一湖一链一中枢”新型数据底座平台，构建“全域入湖、数据可信、AI 好用”的 AI 可信数据空间方案，作为 AI 时代的新型数据基础设施，促进 Data+AI 的全面融合发展与创新。

**③ “数智共生”新生态**

Data+AI 时代数据与智能深度耦合在一起，你中有我，我中有你，难以剥离。尤其在生态体系方面更为密不可分，单纯的数据或 AI 生态体系都无法满足数据要素流通、大模型供数和 AI 智能业务等新场景新业态的诉求，因此未来必定是一个“数智共生”的全新生态体系，包括数据生态、模型生态、Agent 生态、应用生态、服务生态等不同维度的生态系统深度融合、相互促进，共同生长，形成繁荣的超级应用体系。

## 1、新体系：数模协同

在数智融合创新发展进入新时期，可信数据空间与大模型的协同治理，正在从技术耦合升级为数字生态的范式革命。制度协同锚定监管共识，通过跨主体权责契约与动态合规框架，破解数据主权与模型效能的两难；技术协同贯穿全链路防控体系，以隐私计算为盾、区块链为链、AI 治理为眼，实现从数据开发流通到模型价值释放的可信穿透；同时，价值协同激活要素功能，在安全可控的底座上，推动数据资产化、数据知识化与模型智能化的双向赋能，让流动的数据成为大模型创新的引擎，而非沉睡的资源 --- 以制度为纲、技术为脉、价值为魂，构建数模协同新体系。

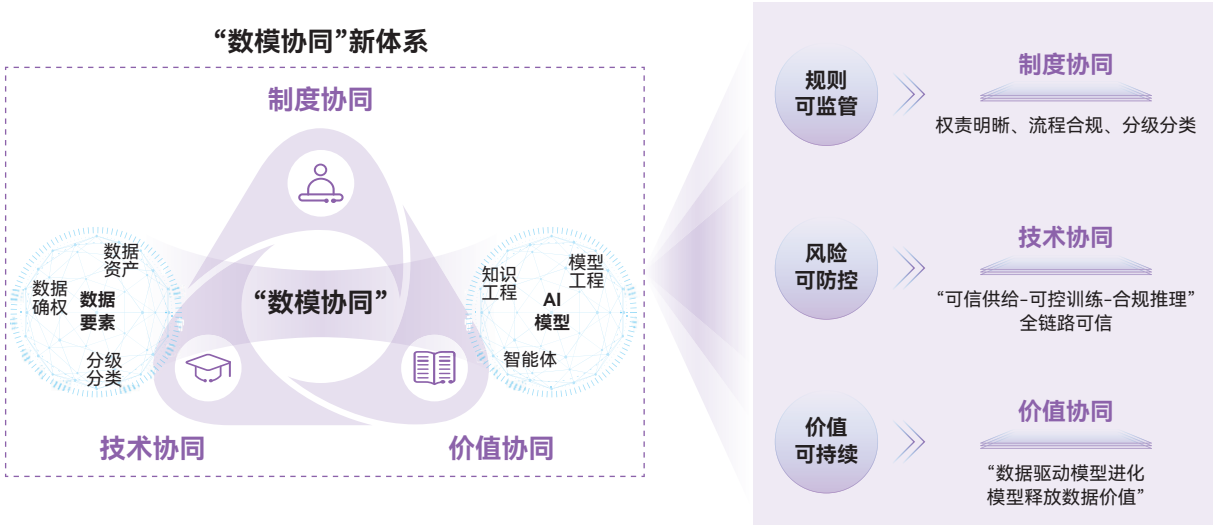



图 7 “数模协同”体系框架

### 一、制度协同：建立“权责明晰、动态确权与授权、健全授权运营”的治理机制

“数据二十条”以解决市场主体遇到的实际问题为导向，创新数据产权观念，淡化所有权、强调使用权，聚焦数据使用权流通，创造性提出建立数据资源持有权、数据加工使用权和数据产品经营权“三权分置”的数据产权制度框架，构建中国特色数据产权制度体系，旨在破解数据确权难题、激活要素市场。但在实际落地中，面临以下三大现实挑战：

| 挑战类型 | 核心问题 |
| --- | --- |
| 法律配套 | “三权分置”虽然淡化所有权，强调使用权流通，但数据原始所有归属权仍缺乏法律明确定义，导致权利边界，尤其在数据多主体加工后，权属链条断裂引发纠纷。 |
| 技术实现 | 动态确权技术不成熟，数据经过多层加工后，原始数据价值贡献者权益难以追溯。<br>数据匿名化合规标准不统一，无法判定匿名化。 |
| 流通机制 | 敏感数据未明确分级授权机制与标准。 |

表 1 “数据二十条”面临的关键挑战

**1、立法赋权**

建议通过逐步完善数据权责类法律法规，明确“三权”的法律属性：创设独立的数据财产权体系，界定“持有权”为事实管控权、“加工使用权”为受限用益权、“经营权”为产品化收益权；规定权利取得方式：原始取得（数据采集）、继受取得（授权许可），并明确“持有权”以合法控制为前提。

“数据资源持有权”，亦称数据持有权，指权利人有权对数据实施自主管控。未经权利人同意或出现法定事由，任何人不得侵害权利人对数据的持有状态。持有权的常见权利形式为消极的“防御”。

“数据加工使用权”或数据使用权，指的是权利人有权通过分析、加工、融合等各类方式，对其依法持有的数据进行自主利用。数据加工是行使数据使用权的重要方式,权利人通过行使该使用权,可发挥数据价值，以满足自身产业需求。

“数据产品经营权”或数据经营权，指权利人可通过转让、许可使用、合作开发及设立担保等方式，将数据上的部分或全部财产权益向外部进行转让。权利人行使该经营权以实现数据的流通与复用。

上述三种权利存在不同的组合状态。权利可因原始的数据生成行为等，同时获得数据持有权、数据使用权和数据经营权。在数据流通交易过程中，应允许不同市场参与主体通过合同交易获取数据三权中的部分或全部。

当然，数据权利的形式应当受到一定限制。这些前置条件包括遵循《个人信息保护法》《数据安全法》等法律规定，不得侵犯个人隐私及信息权益，不得损害数据安全。

**2、技术探索 - 确权与授权全链追溯**

数据要素确权追溯与安全可控技术探索，不仅仅促进数据要素市场化，降低交易成本，激活数据资产流通，同时，通过动态确权、动态授权、多方隐私计算等，构建“以技术赋能监管，以合规促进流通”的新型治理体系。

1) 动态确权：通过区块链的分布式、不可篡改、共识机制特性，将数据权属信息及流转记录上链存证，确保历史记录无法被篡改。同时将数据确权规则（使用权、收益权分配）编码为智能合约，自动执行权属转移和权限控制，减少人为干预风险，当满足预设条件时，智能合约自动触发数据所有权变更，并在链上更新权属记录。

2) 动态授权：在数据要素流通领域，通过结合区块链、隐私计算、智能合约等前沿技术，实现了数据权属的实时调整与精准管控，实现“智能合约驱动，一次确权、动态授权”，例如基于预设规则（时间、地域等）自动调整数据权限即限定数据在特定时段或者区域内使用，超范围自动失效。

3) 可用不可见：结合联邦学习、安全多方计算、密态训练与推理，在数据不离开本地的前提下进行协同计算或者可信计算环境下密态训练与推理，确保原始数据不被泄露的同时，权属和使用过程可审计、可追溯。从数据生成、数据计算、数据流通的每个环节均上链存证，形成不可篡改的审计轨迹，从而支持全生命周期的合规溯源。

**3、流通机制: 从被动开放走向主动赋能**

在数据要素流通领域, 公共数据授权机制与流程优化是释放数据价值、平衡安全与效率的关键。建设公共数据授权平台, 标志着公共数据资源迈入“价值化转型、集约化开发、规模化流通”新阶段。“公共数据授权运营空间”以区块链、隐私计算、使用控制等技术为基础支撑, 采用“开发工具库 + 业务管控台”模式, 按照数据资源、数据治理、数据开发、数据运营、数据流通五大层级进行架构。一是通过零信任、云桌面等技术确保终端接入安全; 二是开展数据分类分级制定对应安全管控策略; 三是通过数据脱敏、数据水印、隐私计算、API 接口管控等能力确保数据开发、测试及使用安全; 四是建设数据安全监管平台, 通过采集流量和日志实现全流程风险预警和及时处置; 五是通过区块链技术确保全流程操作可溯源存证。

## 二、技术协同: 建立“可信供给 - 可控训练 - 合规推理”的全生命周期的防控体系

数据要素与大模型的技术协同, 需要融合“可信数据空间”与“AI 安全防控”双轨能力, 构建覆盖数据加工与流通 -> 模型训练 -> 推理部署 -> 应用监控等系统性防护框架。

**1、可信供给: 数据可用性与可信流通**

通过可信数据空间构建安全合规的数据供给与流通技术体系, 确保语料来源合法、权属清晰、质量可控。基于区块链存证与动态确权, 记录语料来源及流转路径, 实现全链路可追溯, 结合动态脱敏、合规算子、多级过滤等关键技术, 系统性的抵御数据投毒、对抗样本攻击、价值偏见渗透等语料污染与危害。

**2、可控训练: 环境隔离与模型安全**

通过训练环境隔离、隐私保护机制、污染防御与模型鲁棒性实现模型训练分层防护“环境可信、数据可控、模型安全”。

1) 训练环境安全隔离: 构筑数据与模型参数“防火墙”, 将敏感数据与模型参数隔离在受保护的区域, 阻止外部攻击与窃取。例如 可信执行环境 TEE 硬件级隔离。

2) 污染防御与模型鲁棒性: 预训练阶段嵌入异常检测模型和知识图谱一致性, 识别被篡改数据, 抵御恶意数据侵蚀; 遵循“先攻击后防御”的对抗训练优化逻辑, 通过模拟攻击方法和对抗数据样本增强 (例如 FGSM、PGD) , 提高模型鲁棒性以及泛化能力, 从而覆盖训练集未见的边界, 减少过拟合。

**3、合规推理: 环境隔离与输入输出管控**

大模型在推理阶段面临的安全合规技术挑战涉及多个层面, 需要兼顾模型行为可控性、数据隐私保护、内容合规等。

1) 对抗性攻击与思维链劫持: 攻击者通过篡改或操纵模型的推理步骤, 诱导模型跳过安全审查, 输出有害内容。例如部署语义防火墙拦截越狱指令, 结合检索增强实时校验内容的准确性和合规性。

2) 数据隐私泄露: 用户与模型的交互内容可能被拦截或者非法存储。攻击者可通过 API 漏洞或者身份验证不足, 窃取隐私数据。例如通过密态推理和区块链存证实现密文推理和泄露溯源。

3) 输出内容合规: 训练数据的隐性偏见或者投毒攻

击，可能导致模型输出歧视性内容，带来合规偏差和意识形态风险。例如通过多层级审核引擎、敏感数据实时监测、价值观对齐指标等系统级防御。

## 三、价值协同：建立“数据驱动模型进化，模型释放数据价值”可持续双向闭环

数据要素与大模型的协同发展构成了双向赋能的闭环价值体系，数据作为核心生产资料驱动大模型的能力进化，大模型通过智能化释放数据的深层次价值。

**1、数据驱动模型进化：**大模型训练依赖大规模、高质量、多模态、实时更新的行业专属数据集，通过高质量数据集提升认知广度和语言泛化能力以及模型输出的准确性，剔除冗余和敏感信息，确保训练数据“可信合规”，避免大模型生成错误和偏见内容。同时，从原始数据资源到“模型燃料”的转化，成为模型可识别的标准化输入，支撑模型实时检索增强，减少幻觉问题。对于文本、图像、语音等跨模态数据的联合标注与配对，驱动模型理解跨模态语义对齐，实现物理世界交互场景的深度认知。充分利用人类和模型反馈数据（交互日志、错误修正样本）持续优化输出，动态生成高质量训练数据，推动模型自主进化。

**2、模型释放数据价值：**大模型突破传统数据分析的局限，实现数据价值的跃迁，从海量信息提取行业级的知识发现，例如在行业预测大模型，通过结合文本、图像、时序数据，构建实时分析能力，实现故障根因溯源、风险预测、故障率预测、决策精度、流程动态调优、自动化元数据管理与知识图谱构建等。

双向的价值协同不仅推动数据要素从“资源”向“资产”跃迁，更通过“数据驱动模型进化，模型重构数据价值”的飞轮效应，成为数字化转型的核心引擎。

## 2、新架构: AI 可信数据空间

### 2.1 “三可”架构原则

可信数据空间作为“人工智能 +”行动的数据基座，旨在构建“数据高质量供给 -> 模型深度赋能行业 -> 模型释放数据价值”的闭环。当前行业大模型语料质量差、资源结构单一、流通成本高“等制约人工智能 + 的深度发展。高质量数据集供给与流通作为系统性复杂的软件工程，涉及多模态数据工程、隐私计算、区块链、向量数据库、RAG 等创新技术组件的统一集成、统一管控、统一监控与运维。

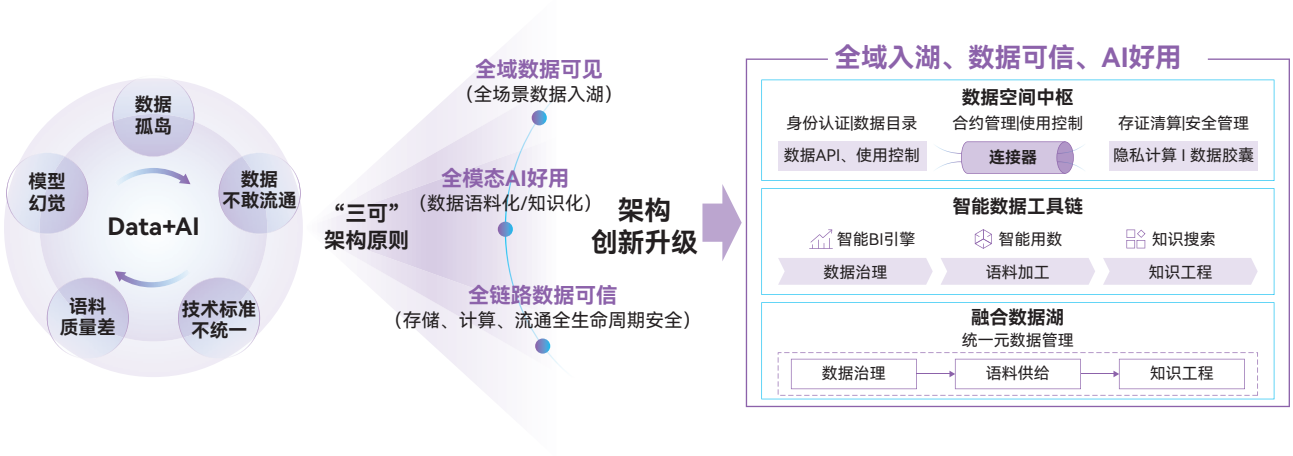



图 8 AI 可信数据空间“三可”架构原则示意图

**“三可架构”原则**

• **全域数据可见:** 通过统一数据架构平台，打破跨系统、跨技术堆栈、跨主体的数据壁垒，实现多源数据的全面汇聚、元数据管理、权限管理以及一站式端到端的汇聚 -> 清洗 -> 加工-> 标注 -> 合成 -> 评测等，其核心是构建横向贯通的数据网络，支撑传统数据分析、AI 语料加工治理、大模型训推数据集的高质量供给。

• **全链路数据可信:** 基于“数据可用不可见、可控可计量、训练语料可控、推理安全合规”技术与治理规范，确保数据从采集、加工到模型训练、推理的全生命周期满足准确性、内容可信、一致性、安全性等要求，建立贯穿 Data+AI 数据流的价值信任链。

• **全模态 AI 好用:** 通过数据 -> 语料 -> 知识一站式算子 +工具链体系，确保数据接入到 AI 服务输出的全链路高质量输出，整体架构不仅能处理多模态数据，而且系统保障数据质量、语料合规性与知识转换率，从而支撑大模型在动态数据环境中持续稳定、可信的推理能力。

华为以“全域数据可见、全链路数据可信、全模态 AI 好用”的架构原则为基石，构建 “全域入湖、数据可信、AI 好用”新一代 Data+AI 创新架构。

## 2.2 “一湖一链一中枢”架构蓝图

“一湖一链一中枢”架构通过对传统数据平台的系统性升级，不仅解决了“全域数据可见”的多源异构数据整合难题，更通过数据血缘追溯、统一权限控制、跨域流通策略控制确保“全链路数据可信”，同时依托智能数据工具链赋能供给复杂智能用数场景真正实现全模态 AI 好用。

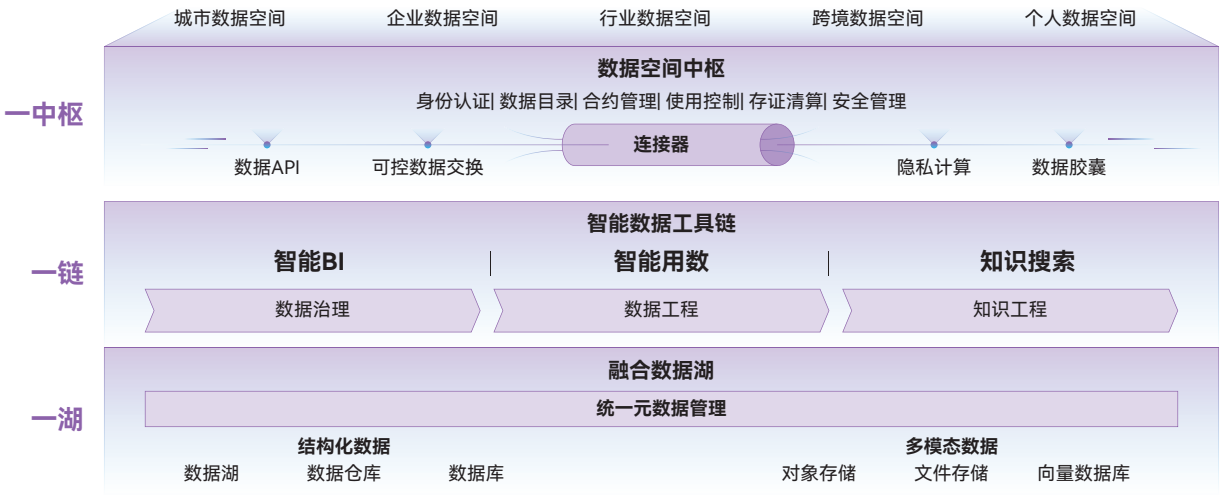



图 9 “一湖一链一中枢”体系架构

### 2.2.1 融合数据湖

大数据时代数据库、数据仓库、数据湖的建设主要围绕结构化数据、商业智能 (BI) 而建设，经过多年建设与发展，逐渐形成了多主体数据孤岛、多技术平台烟囱、多模态数据语义割裂无法融合等问题，无法满足 AI 时代对全模态、全时态、全形态数据集成与融合的需求，因此迫切需要通过 AI 数据湖升级，解决传统数据 AI 不可见的问题。

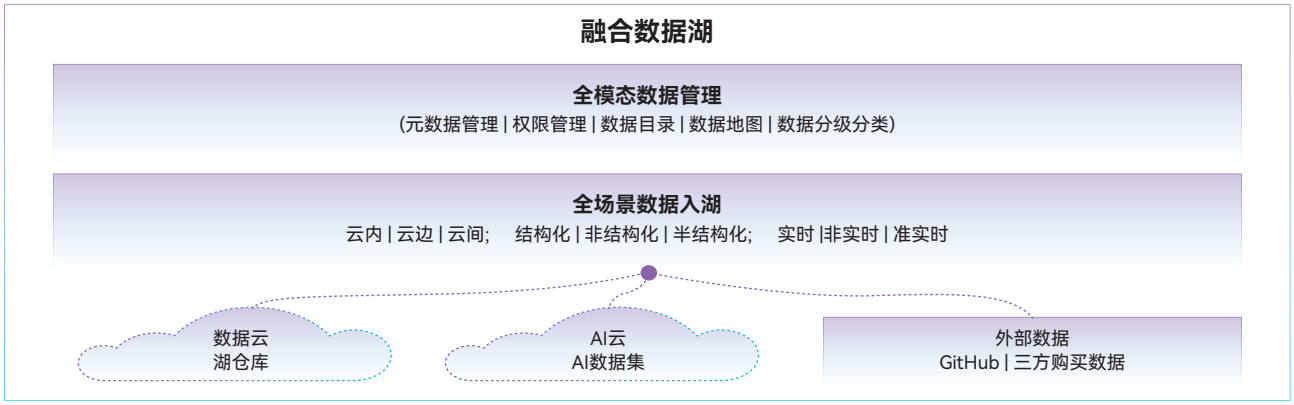



图 10 融合数据湖功能架构

为打破现有数据壁垒，满足企业全量数据让 AI 可见的关键需求，AI 数据湖架构升级必须实现“全场景数据全域入湖、全模态数据统一管理”的关键设计，实现对上层不同智能用数场景的高质量供数。

**关键设计一: 全场景数据, 全域入湖**

- **跨云全网数据集成:** 主要解决云内、云边、云间系统数据集成问题

| 集成场景 | 场景说明 | 举例 |
| --- | --- | --- |
| 云内集成 | 部署云上的OLTP、OLAP、OBS等数据 | 如云上核心系统产生数据 |
| 云间集成 | 部署不同云服务商之间的数据系统，<br>数据接入标准、协议、技术栈不完全一样 | 如不同云服务商，混合云与公有云上的系统产生数据 |
| 云边集成 | 部署在边缘设备上的系统产生的数据，包括IoT、视频等 | 如边端智能设备、传感器产生的实时数据 |

表 2 跨云全网数据集成面临的关键问题

- **跨域业务数据集成:** 主要解决 IT 与 OT 数据融合集成问题

IT 数据通常以结构化形式存储于数据库中, 主要用于企业的管理决策和业务流程; OT 数据则来源于工业生产、设备监控等领域, 多为实时性较强的时序数据。采用数据抽取、转换和加载 (ETL) 工具, 结合流式计算技术, 实现对不同类型跨域数据的统一集成。

**关键设计二: 全模态数据, 统一管理**

- **统一管理元数据、权限、标准与监控**

1) 统一元数据管理: 通过统一元数据管理技术, 实现对结构化、半结构化和非结构化数据元数据的采集、存储、查询和维护。例如, 对于非结构化的文本数据, 元数据可以包括数据的来源、创建时间、作者等信息; 对于图像数据, 元数据可以包含图像的分辨率、格式、拍摄设备等信息。

2) 统一权限管理: 基于角色的访问控制 (RBAC) 模型, 结合数据分级分类, 实现对不同用户和角色的数据访问权限管理, 有效防止数据泄露和滥用, 保障企业的数据安全。例如, 对于敏感的客户个人信息数据, 只有授权的客服人员和管理人员能够访问; 对于公开数据, 企业内的所有员工均可进行查询。

3) 统一标准管理: 通过建立统一的数据标准体系, 涵盖数据格式、编码规则、数据质量等方面的标准, 实现全模态数据的共享和交换。

4) 统一监控管理: 对数据的采集、存储、处理和使用过程进行实时监控。通过设置监控指标和预警机制, 及时发现数据异常情况, 如数据丢失、数据错误、数据访问异常等, 并采取相应的措施进行处理。

- **统一数据目录与数据地图**

1) 统一数据目录: 构建统一的数据目录, 将结构化、半结构化和非结构化数据按照业务主题、数据类型、数据来源等维度进行分类, 每个数据条目包含数据的名称、描述、格式、存储位置、更新时间等信息。

2) 统一数据地图: 数据地图以可视化的方式展示数据之间的关系和流向, 帮助数据使用者更好地理解数据的整体架构和数据资产分布。通过数据血缘分析和数据拓扑建模, 生成直观数据地图。

### 2.2.2 智能数据工具链

#### 一、数据工程

当前各行业面向结构化数据的治理方法和工具相对比较成熟，但是面向 AI 语料对非结构化、多模态数据的清洗、标注、自动合成、质量评价等一系列工具的相对匮乏现状，需围绕高质量语料自动化加工建立新的数据 / 语料 / 知识加工生产线，满足 AI 时代的高质量、高效率、自动化加工的诉求。


图 11 数据工程功能架构

数据工程平台聚焦构建高效、精准的数据加工链路，为大模型提供源源不断的高质量数据，全方位提升模型训练的效率与效果。它提供数据获取、数据加工、数据标注、数据发布、数据管理、数据安全六大全链路数据工程服务，通过一站式平台，支撑企业高质量用数。

①数据获取，支持多样化数据来源渠道，支持全模态数据类型，满足大模型训练所需各类数据采集诉求。

②数据加工，包括数据提取、数据过滤和数据转换 & 增强等，其中，数据提取，支持各种格式的文本、图像和音视频数据提取，自动化算子提升效率 10+ 倍；数据过滤，内置丰富数据过滤算子，支持用户自定义规则过滤，拒绝低质和不合规数据影响模型效果；数据转换 & 增强，提供丰富的数据转换和增强算子，构建高质量数据集支持模型效果更佳。

③数据标注，支持文本 / 视频 / 图像全类型标注，可以辅助预标注提效 10 倍，进一步借助人工审核提升标注准确率。

④数据发布，提供多种发布方式，一键发布到模型训练平台直通训练，同时可借助胶囊封装防止数据泄露。

⑤数据管理，提供数据地图实现资产全视角管理，全链路数据血缘支持正向和逆向数据追溯。

⑥数据安全，提供数据工程处理全流程安全管理，实现安全高效数据处理。

除了要构建全新面向 AI 语料加工的全流程工具链外，面向不同大模型场景的数据链路和工程实践也尤为重要，比如面向预测、NLP、CV、多模态不同类型语料建立完备的工具链与工程基线。

①预测数据加工链路，致力于处理时间序列等预测相关数据。通过多种数据源，如传感器数据、业务报表数据等，进行数据获取。利用异常值检测、数据平滑等清洗技术，去除噪声和错误数据，确保数据的可靠性。接着依据行业标准和业务规则，对数据进行标准化处理，使不同来源的数据具备统一格式，便于后续分析。运用特征工程方法进行数据转换，提取如滑动窗口特征、趋势特征等，增强数据的表达能力。在数据评估阶段，通过交叉验证等方式评估数据质量和预测模型的初步效果，根据评估结果优化数据处理流程。最后将符合要求的数据发布到模型训练平台，为预测类大模型提供坚实的数据支撑，助力提升预测的准确性和稳定性，例如在金融风险预测、销售趋势预测等场景发挥关键作用。

② NLP 数据加工链路，是围绕自然语言文本展开的。在数据获取环节，会涵盖网页文本、社交媒体内容、文档资料等多种来源，采用文本去重、停用词过滤等清洗手段，净化文本数据，依据自然语言处理的规范和模型需求，对文本进行词性标注、句法分析等标准化操作。同时，利用词嵌入、文本摘要等技术进行数据转换，将文本转化为适合模型处理的向量表示。除此之外，通过人工标注与自动评估相结合的方式，对 NLP 数据进行质量评估，确保数据的标注准确性和语义完整性。并通过发布高质量的 NLP 数据，为语言生成、文本分类、情感分析等大模型提供优质语料，显著提升大模型在语言处理任务中的表现。

③ CV 数据加工链路，专注于图像和视频数据处理。在数据获取环节，它从摄像头采集、图像库、视频平台等多渠道获取数据。运用图像去噪、图像增强等清洗技术，提升图像质量。按照图像识别、目标检测等任务要求，对数据进行标注和格式标准化。通过图像特征提取、目标检测框生成等方式进行数据转换，提取关键视觉特征。借助标注一致性检查、模型验证等评估手段，保证 CV 数据的质量。将处理后的 CV 数据发布到模型训练环境，为图像识别、目标跟踪、视频分析等大模型提供丰富的训练素材，有效提升大模型在视觉领域的识别精度和泛化能力。

④多模态数据加工链路，整合 NLP、CV 等多种类型数据。在数据获取时，同步采集文本、图像、音频等多模态数据。通过多模态数据对齐、融合等清洗和转换技术，打破模态间的隔阂，实现数据的有机结合。依据多模态任务的标准，对融合后的数据进行标准化处理。运用多模态特征融合算法进行数据转换，生成综合多模态信息的特征向量。采用多模态评估指标，从多个维度评估数据质量。发布多模态融合数据，为多模态大模型提供全面的数据支持，使模型能够综合理解和处理多种信息，在智能客服、智能安防、智能驾驶等复杂场景中发挥更强大的作用。

## 二、智能 BI

在可信数据空间中，需要借助丰富的数据开发工具，将原始数据加工成为具备流通价值的数据资产和数据产品。然而，由于数据来源、类型、规模以及流通和展现形式的多样性，要求提供多样化的数据开发工具进行支撑，包括统一的数据服务门户和智能化数据分析工具。

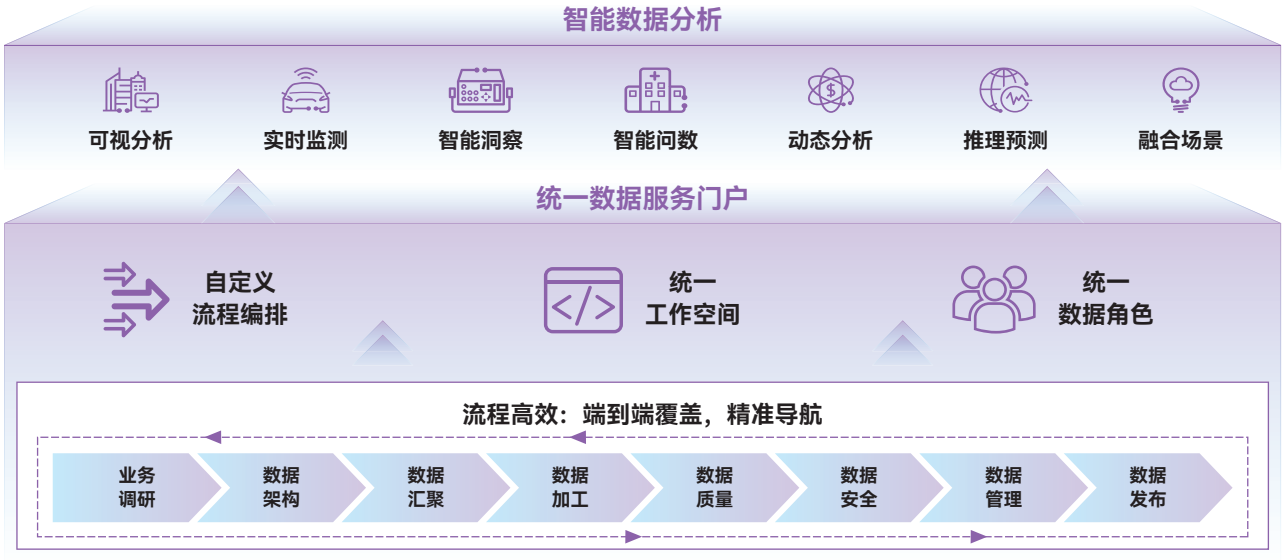



图 12 智能 BI 体系架构

・统一数据服务门户

传统工具在设计时通常只关注各自领域和阶段的数据管理能力，通过相互独立的权限体系进行角色授权，导致管理人员在环境准备、工具权限分配以及开发人员在工具使用过程中存在大量重复操作，难以实现一致的管理和使用体验，最终严重影响数据产品的开发利用效率。

・智能数据分析

在公共数据分析处理中，原始数据包括征信数据、企业信息、统计报表等结构化数据，以及政策文本、新闻报道、公众意见等非结构化数据。这些数据经过清洗与标注后，转化为可分析的结构化数据，并通过知识图谱构建非结构化知识的潜在关联。AI 可分析结构化数据与非结构化数据的关联，生成政策解读报告，提升决策的科学性。智能数据分析技术（如 NL2SQL）将自然语言转化为 SQL 查询，帮助工作人员快速获取所需数据，AI 可直接从结构化数据中获取结果，并结合非结构化数据生成全面分析报告。这种智能化方式提升了数据分析效率，助力政府更精准地制定政策，满足公众需求。


图 13 智能数据分析功能架构

开展智能数据分析时，将数据可视化与 AI 问答相结合，实现智能化的数据分析与洞察。在企业运营中，能够实时分析销售数据与市场趋势，生成动态的可视化报告，并通过自然语言交互提供业务建议。这种技术不仅提升了数据分析的效率，还为企业决策提供了更加直观的支持。

智能数据洞察提供可视、实时、易用、安全的企业智能分析数据服务，以最自然高效的方式获取业务见解，支撑业务实时高效决策。适配云上云下多种数据源，提供丰富多样的可视化组件，采用拖拽式自由布局，轻松实现数据分析和报表搭建，快速定制专属数据大屏。

## 三、智能用数

Data Agent 通过生成式 AI 将自然语言转换为可智能化执行的端到端数据自助分析与决策呈现，成为“智能用数”的核心载体，其本质是以 AI 为引擎，大幅降低数据使用门槛，使业务人员直接驱动数据从资源转化为行动依据。用户只需提出分析需求，agent 智能化完成数据到洞察报告的端到端流程，将传统“提需求 - 等开发”的天级周期压缩至分钟级，真正实现了“所思即所得”。

Data Agent 是指用户可以通过生成式 AI 技术，灵活构建自己的对话式数据助手，通过自然语言交互的方式获取对应数据，并生成各类可视化报表和洞察报告。实现从原始数据到生成洞察报告的端到端智能化处理，加速数据价值释放。

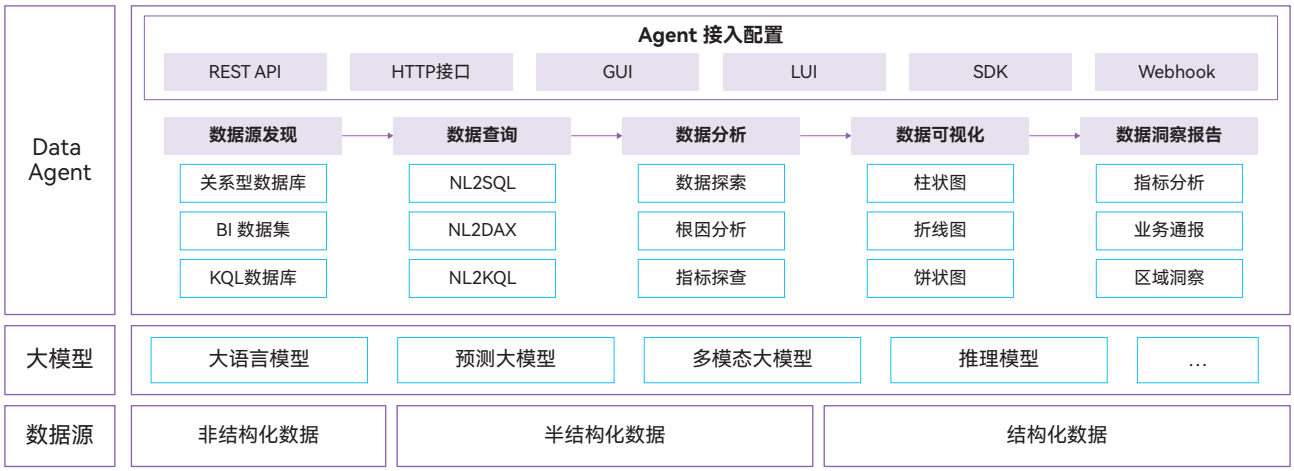



图 14 Data Agent 功能架构

**数据源发现:** 首先, 对用户问题进行理解, 明确所需数据源, 其次, 根据用户权限访问数据源, 确保系统可获取的数据及数据结构符合合规要求, 再根据可用的数据源及 agent 指令确定最相关数据源。

**数据查询:** 确定最相关数据源后, 通过大模型将用户问题进行改写, 使其更加清晰和结构化, 同时自动生成 SQL、DAX 公式、实时日志 / 时序数据查询语言等, 结合上下文进行数据查询。

**数据分析:** 基于查询的结果进行多维度的深度数据挖掘, 自动识别数据分布、异常值、相关性特征; 定位关键波动因素 (如销售额骤降) , 构建归因模型; 追踪核心指标趋势, 进行同环比、阈值预警、聚类归因等自动化诊断。

**数据可视化:** 根据已分析出的特征数据与分析目标, 智能选择图表类型, 例如: 柱状图 (对比分析) 、折线图 (趋势追踪) 、饼图 (构成占比) , 以及热力图 (相关性矩阵) 、散点图 (聚类分布) 、地理地图 (空间分析) 等高级图表。

**洞察报告生成:** 基于已有的分析结果及可视化报表, 生成完整的业务洞察。例如: 指标分析报告, 包含关键指标健康度诊断与归因摘要; 业务通报, 自动生成周报 / 月报 / 季报, 并能包含多图表整合与执行报告; 区域洞察, 包含地理维度业绩对比、市场渗透率热区识别等。

综上所述, Data Agent 致力于降低数据使用门槛, 利用生成式 AI 高效驱动业务分析与流程闭环, 辅助业务人员提升运作效率, 高效决策。

## 四、知识搜索

知识搜索引擎是实现数据与 AI 协同创新的核心工具。它通过结合大模型与知识库, 构建高效、智能的检索增强生成 (RAG) 能力, 帮助用户快速获取精准的知识内容, 提升决策效率与业务洞察力。

知识搜索引擎的核心目标是解决大模型在知识准确性、专业性、时效性等方面的不足, 通过引入结构化知识库和向量检索技术, 实现对海量数据的高效检索与智能生成。例如, 在政府政策解读场景中, 知识搜索引擎能够快速检索相关政策条款, 并结合历史案例生成个性化的解读报告, 显著提升 AI 的推理能力与响应效率。

但是，知识库的构建与搜索需要处理海量的多模态数据，包括文本、图片、视频等，数据清洗、标注与结构化处理的复杂性较高。其次，大模型在知识准确性、专业性、时效性等方面的不足，可能导致检索结果与用户需求不完全匹配。此外，非结构化数据的处理难度较大，AI 在理解与转化过程中容易出现偏差，难以生成高价值的内容。最后，数据安全与隐私保护问题也对知识搜索引擎提出了更高的要求，如何在提升检索效率的同时确保数据的合规性与安全性，仍是亟待解决的难题。

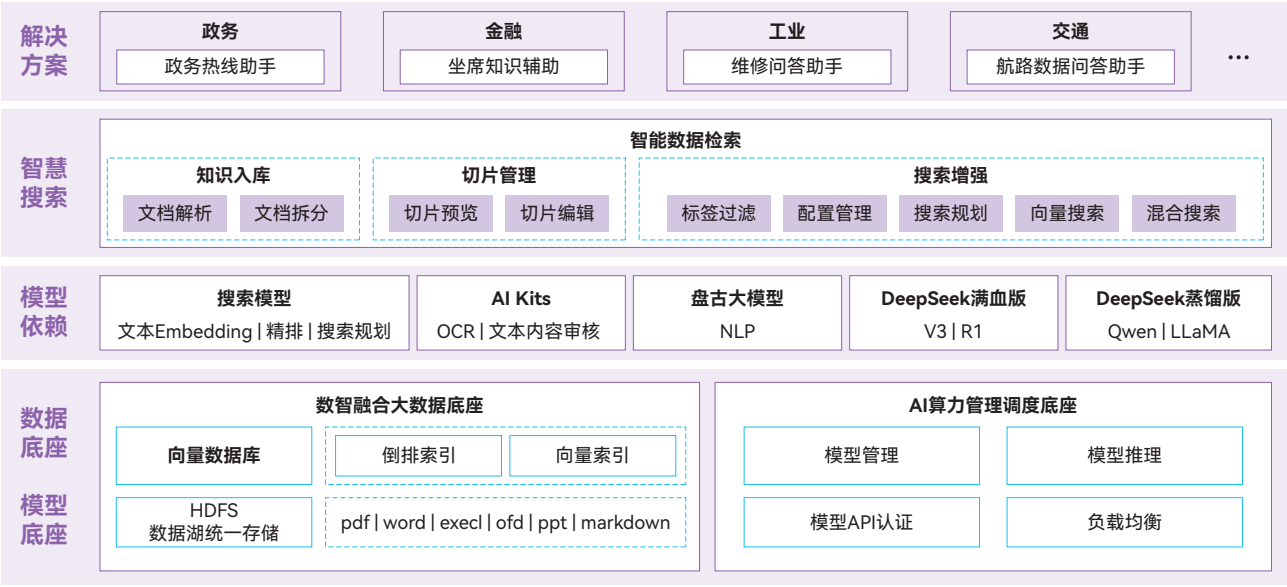



图 15 知识搜索引擎功能架构

基于上述知识库搜索面临的问题与挑战，在可信数据空间中的知识搜索引擎需要提供如下核心能力：

- 智能文档解析与知识构建

知识搜索引擎支持多种文档格式（如 PDF、Word、PPT、图片等）的解析与知识提取，将非结构化数据转化为 AI 可理解的知识图谱或向量数据。例如，政策文件中的条款会被转化为知识图谱，清晰展示政策间的关联关系；图片和表格则会被转录并标注，生成结构化的数据格式。这种能力使得 AI 能够快速理解复杂业务场景，为用户提供精准的知识服务。

- 智能检索与生成

通过向量数据库和知识图谱的结合，知识搜索引擎能够实现高效的知识检索与生成。例如，在企业市场分析中，AI 可以快速检索历史销售数据与市场趋势，生成动态的可视化报告，并通过自然语言交互提供业务建议。这种智能化方式不仅提升了数据分析的效率，还为企业决策提供了直观的支持。

- 灵活部署与安全管控

知识搜索引擎支持多种部署形态（如物理机、云上云下），满足不同行业对数据安全的要求。同时，平台提供统一的数据服务门户，实现工具的统一管控与数据的全生命周期管理，确保数据的可信、可用与安全。

### 2.2.3 数据空间中枢

随着数据价值增长和数据隐私保护意识的提高，用户越来越关注在 AI 时代下对于数据安全的新要求。对于数据加工的安全，应该在存储、加工、发布端到端业务流程中能保证数据不被泄露、不被盗取、不被恶意篡改或销毁；对于数据内容的安全，应该做到无敏感数据、无黄暴政数据，确保数据供给到模型的价值观正确。

**可信数据空间中枢**

为实现数据合规高效流通，推动数字经济高质量发展，华为依托自身可控数据交换、隐私计算、区块链等技术积累，针对数据要素产业发展面临的信任缺失、确权困难等痛点，打造数据空间中枢平台，构建“数据可用不可见”的可信数据基础设施。数据空间中枢作为可信数据空间关键内核，通过标准化数据要素加工、流通利用的全生命周期业务流程，实现对数据资源、参与主体、业务活动、工具资源、跨空间互联互通的统一管理调度。

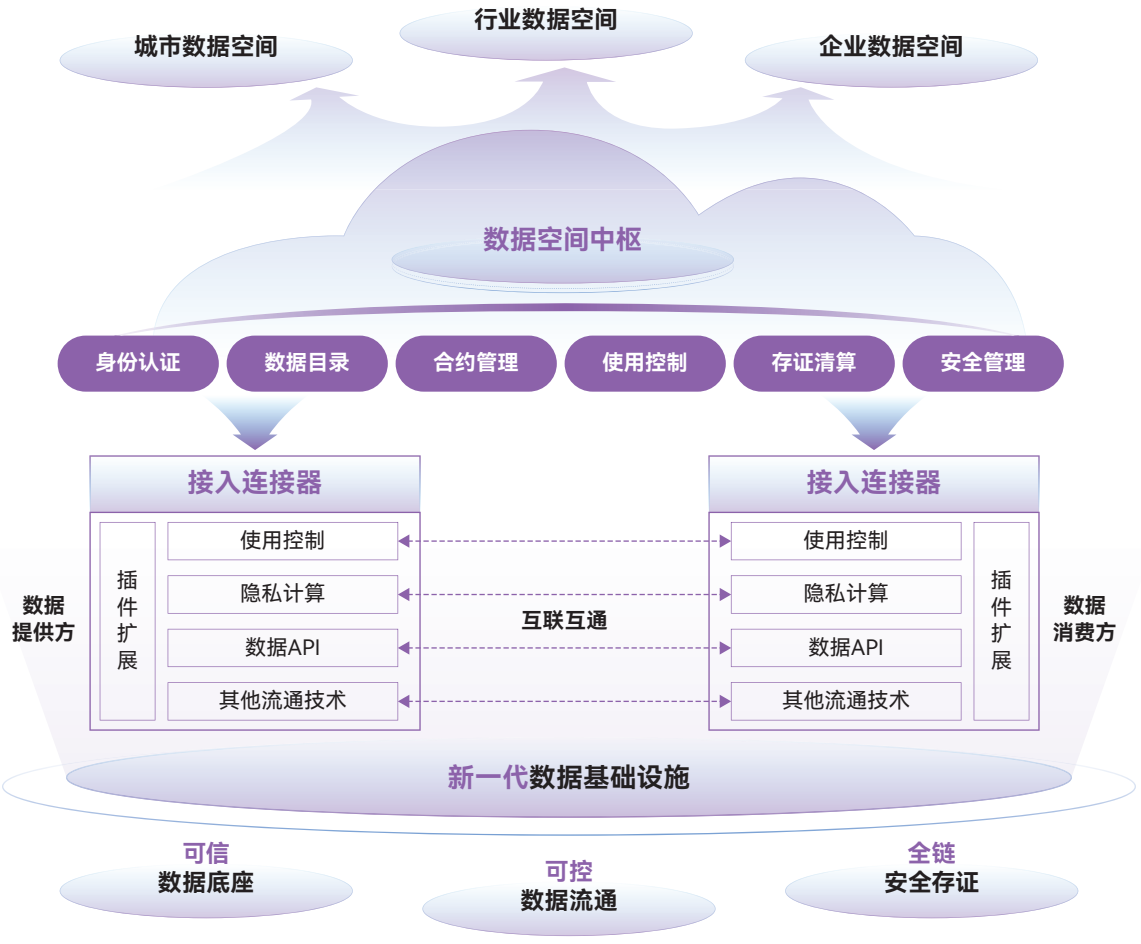



图 16 数据空间中枢体系架构

数据空间中枢由身份认证、数据目录、数字合约、使用控制、存证清算、安全管理六大模块组成。

身份认证模块落实全要素接入认证要求，对空间内主体身份、接入连接器及核心组件开展认证以建立互信，通过统一认证实现管理，借助连接器及开发工具认证保障技术组件和工具适配标准协议，确保安全可靠。

数据目录模块负责数据产品统一发布，包含数据资源目录与数据产品目录，接入数据和开发形成的产品分别按对应编目要求编目、审核后上架数据市场。数据市场提供数据语义发现和元数据智能识别能力，方便数据使用方快速找到数据资源。

数字合约和数据使用控制模块落实全过程动态管控，对数据共享共用全生命周期进行控制，中枢统一管理策略与合约，积累行业场景标准化策略模板构建智能推荐能力。数据提供方可分配权限和指定控制策略，智能控制中枢编排访问与使用策略并生成可机读策略，封装于连接器；使用方在连接器内按合约消费数据，通过合约管理优化履约，建立合作规范。

存证清算模块践行全链路存证溯源，构建行为和数据存证体系，依托审计日志对数据开发利用全过程溯源存证。同时提供详细的计量监测数据等使用消耗情况，支撑运营平台按定价策略计费，完成支付指令交换，在相关方之间清算费用。

安全管理模块实现全链路风险统一感知，构筑端到端安全防护体系，覆盖数据全流程安全以确保严进严出、过程可控；基于云原生保障数据资产可控，根据分级分类识别风险并实现策略自动化推荐、日志上报和策略下发；构建多层多级授权体系实现权限隔离，通过“租户 + 云服务 + 存算平台”三级授权拉通权限模型，提供细粒度权限隔离能力，满足最小化授权等原则。

**可信数据流通安全**

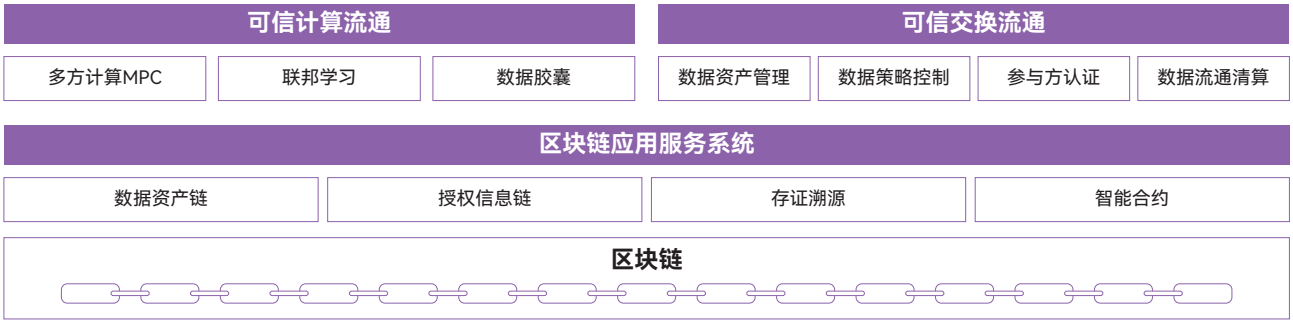



图 17 可信数据流通安全功能架构

可信计算流通通过多方计算、联邦学习和机密计算等技术，实现数据可用不可见，满足高密级数据流通使用的要求。

可信交换流通主要由策略控制中心、数据连接器、审计清算中心、认证中心等核心部分组成，定位为提供数据空间能力，通过提供合法性认证、数据使用安全策略控制、审计追溯等能力，实现数据在提供方与消费方数据空间交换过程中，端到端的安全保护。

区块链服务管理平台为可信存证管理系统提供区块链技术支撑，包括区块联盟管理、存证溯源、智能合约引擎、共识机制、共享账本、密码算法、运维监测等服务支撑。各相关参与方包括数据运营公司、大数据中心、管理机构等可通过区块链平台快速方便地创建和部署自身的区块链节点，并共同组成可信存证管理联盟链。

## 2.3 创新技术方向

### 2.3.1 数字护照

数字护照 (Digital Passport) 是基于数字身份 (Digital Identity, DID) 和相关技术构建的一种新型身份凭证，旨在以安全、隐私保护且可验证的方式，将个人或实体的身份信息数字化，并支持跨平台、跨国家的可信交互。

| 特性 | 传统电子护照（ePassport） | 数字护照（基于DID） |
| --- | --- | --- |
| 存储方式 | 嵌入式芯片（RFID） | 用户控制的数字钱包（如手机App） |
| 验证依赖 | 中心化数据库（如政府系统） | 分布式账本和密码学验证 |
| 隐私控制 | 信息全部暴露 | 选择性披露、最小化数据共享 |
| 国际互认 | 需双边协议（如ICAO标准） | 通过开放标准（DID/VC）实现 |

表 3 数字护照与传统电子护照对比

**数字护照的关键特性**

- **用户主权 (Self-Sovereign Identity, SSI)**

用户完全控制自己的数字护照，自主决定何时、向谁披露哪些信息，避免数据被第三方垄断。

- **跨域互操作性**

基于国际标准 (如 W3C DID/VC)，数字护照可在不同国家、平台间通用，例如用于国际旅行、跨境金融等场景。

- **防欺诈与安全**

加密签名和分布式验证机制防止伪造。生物识别 (如面部识别) 可绑定数字护照，进一步确保“人证合一”。

- **最小化披露**

支持选择性披露 (如仅展示护照有效期，而非全部信息)，减少隐私泄露风险。

**数字护照的价值及应用场景**

在日益数字化的世界中，数字护照正从概念走向现实，其核心价值在于为个体在虚拟空间构建一个安全、可信、便捷且用户自主可控的数字身份凭证。它的意义远不止于替代实体证件，而是重塑数字交互方式、赋能个体权利、提升社会效率的关键基础设施。

- **跨境旅行:** 机场安检时，旅客通过数字护照出示加密签名的签证和健康证明，海关通过扫描二维码即时验证真伪，无需纸质文件。

- **Web3.0& 稳定币:** 用户使用数字护照登录去中心化应用 (DApp)，证明真实身份的同时保护匿名性。

- **数字服务接入:** 银行或医疗平台要求用户提供“已成年”证明，用户仅需发送一个 ZKP 驱动的 VC，无需上传身份证照片。

数字护照是 DID 技术在身份领域的落地实践，通过密码学和分布式系统将现实身份转化为可编程、高隐私的数字凭证。其核心价值在于平衡便利性与安全

性，重塑个人与机构之间的信任关系，其以用户为中心，在保障安全隐私的前提下，通过便捷高效的身份管理，重塑信任建立的方式，释放个体潜能，提升社会整体运行效率，为构建一个更包容、更繁荣、更可信的数字未来奠定坚实基础。

### 2.3.2 轻量 AI 机密计算

随着大语言模型的广泛应用，一系列新旧交织的安全威胁接踵而至，给 AI 产业生态的健康发展带来严峻的挑战。

当代的大型语言模型是海量算力、宝贵的训练数据与创新算法的汇聚的精华，是数字时代的高价值资产，自然成为攻击者觊觎的目标。用户提交给模型进行训练和推理处理的数据中，常常包含个人隐私、商业机密等敏感信息，一旦泄露，后果不堪设想。在缺乏有效防护的应用环境里，无论是心怀不轨的系统运维人员，还是利用漏洞、恶意软件非法获取访问权限的黑客，都可能对模型安全构成威胁。

正因如此，随着大模型应用规模的不断扩大，对其高价值资产的安全保护已刻不容缓。

在可用于大模型及推理数据安全防护的技术中，同态密码、多方计算等密码学技术虽凭借其严密的密码学根基具有很高的安全性，但在实际应用中却面临严重的性能瓶颈。

相比之下，基于硬件隔离的机密计算技术更贴合大模型应用的现实需求。机密计算通过在硬件层面构建可信执行环境（TEE），将大模型与推理数据置于受硬件保护的安全区域内，即使操作系统、虚拟机管理程序被恶意攻击，该区域内的数据与计算过程依然能保持机密性与完整性。在实际应用中，机密计算既能够保障模型与数据的安全，又能在性能损耗相对可控的前提下，满足大模型实时推理的效率要求，为大模型的安全应用提供了一条兼具安全性与性能的现实路径。

因此，华为鲲鹏推出了 virtCCA 方案，创新性地突破了传统 TEE 仅能运行小型可信应用与需要应用改造的难题。

同时，为了确保 virtCCA 机密虚机与昇腾 NPU 的安全高效交互，华为昇腾提出了基于 NPU 硬件可信执行环境 TEE 的 PMCC（Privacy and Model Confidential Computing）轻量级 AI 机密计算方案。

借助于 virtCCA+PMCC 方案，华为实现了从数据到模型的端到端硬件级安全保护，为 AI 机密推理与微调等高价值场景提供了坚实的可信根基。

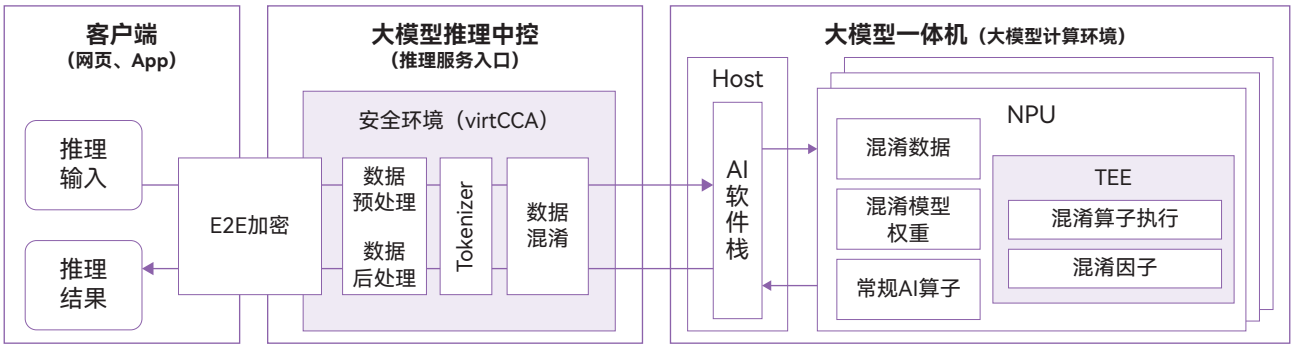



图 18 virtCCA+PMCC 轻量级 AI 机密计算方案示意图

由于 PMCC 的混淆变换是在昇腾 NPU TEE 的高安环境中执行的，因此可以实现变换因子与硬件的绑定，避免传统软件安全方案可被拷贝到其他硬件上执行的缺陷。同时，由于在 PMCC 方案中，昇腾 NPU 上运行的模型与数据都经过了混淆变换，所以即便通过未受保护的 H2D 信道和 D2D 信道传输，也不会影响方案的机密性保护效果。

当前，ARM V9 推出了 CCA (Confidential Computing Architecture) 方案，通过硬件级内存加密与指令集扩展，提供更高效的安全执行空间 Realm。其核心在于构建跨处理器、缓存及总线的全链路机密计算体系。未来华为将通过鲲鹏 CPU + 总线 + 昇腾 NPU 的协同设计，进一步实现 CCA 方案在华为计算的落地，借助于 ARM V9 CCA 的安全特性，通过软硬协同进一步降低 AI 机密推理（及微调）的性能损耗，为大模型机密推理与微调提供兼顾安全与效率的机密计算解决方案。

## 3、新生态：数智共生

### 3.1 培育多元数据生态主体

数据生态是指数据空间参与各方依据既定规则，围绕数据资源的流通、共享、开发、利用开展价值共创的生态系统，包括数据提供方、数据使用方、数据服务方、可信数据空间运营者等生态主体。针对本地优势产业和典型场景，制定生态主体培育政策，形成“政府引导、企业参与、科研支撑、行业协同”的生态体系。通过数据互通、资源共享、协同创新，推动数据空间可持续发展。各个生态主体定义和培育措施如下：

数据提供方：在可信数据空间中提供数据资源的主体，有权决定其他参与方对其数据的访问、共享和使用权限，并有权在数据创造价值后，根据约定分享相应权益。支持企业依法依规对其合法获取的数据进行开发利用，培育一批贴近业务需求的行业性数据资源企业。鼓励企业间按照市场化方式授权使用数据、共同分享收益，推动企业跨行业发展。

数据使用方：在可信数据空间中使用数据资源的主体，依据与可信数据空间运营者、数据提供方等签订的协议，按约加工使用数据资源、数据产品和服务。支持企业面向数据要素 X 典型应用、AI 场景、新兴产业和全域数字化转型需要，创新应用模式，更好发挥数据要素价值，赋能产业发展，培育一批深刻理解行业特征、高度匹配产业需求的数据应用企业。

数据服务方：在可信数据空间中提供各类服务的主体，包括数据开发、数据中介、数据托管等类型，提供数据开发应用、供需撮合、托管运营等服务。支持企业面向数据流通交易提供专业化服务，重点围绕数据业务咨询、数据供需对接，交易撮合、合规服务、数据资产服务等方面，培育一批数据服务企业，发展数据流通交易新模式新业态。

可信数据空间运营方：在可信数据空间中负责日常运营和管理的主体，制定并执行空间运营规则与管理规范，促进参与各方共建、共享、共用可信数据空间，保障可信数据空间的稳定运行与安全合规。可信数据空间运营方可以是独立的第三方，也可以由数据提供方、数据服务方等主体承担。支持企业面向数据接入、数据加工、数据流通、数据运营和数据安全，聚焦数据流通利用基础设施和多元生态主体协同机制，重点培育一批具有公信力、竞争力的数据运营商。

可信数据空间监管方: 指履行可信数据空间监管责任的政府主管部门或授权监管的第三方主体, 负责对可信数据空间的各项活动进行指导、监督和规范, 确保可信数据空间运营的合规性。

### 3.2 制定多元生态主体协同标准和机制

当前各个生态主体间存在信任顾虑, 担心数据安全等问题, 导致跨主体数据 “不敢共享、不会共享、不愿共享”。通过制定统一的数据资源管理标准、认证与信任机制、数据共享规则、技术标准体系、利益分配机制等, 以破除当前数据流通难题。

数据资源管理标准: 建立统一的数据目录和数据标识, 确保数据资源可被高效查询和跨主体互认。

认证与信任机制: 建立统一身份管理, 通过第三方认证机构, 对参与可信数据空间的实体进行安全评估和认证, 确保其合法合规, 并符合标准。建立审计与追溯机制, 确保数据操作的可追溯性, 记录数据的访问、修改、传输等行为, 以便审计和取证。

数据共享规则: 明确数据共享的范围、方式和权限, 以及数字合约协商机制、清算审计、纠纷解决等业务流程进行标准化, 确保数据在合法合规的前提下, 按照合同约定进行流通使用。

技术标准体系: 保障不同主体之间的系统和数据能够互联互通。

利益分配机制: 明确数据共享带来的经济利益或其他收益的分配方式, 激励各方参与合作。根据各主体的贡献程度合理分配收益, 可按照数据提供量、技术研发投入、市场推广效果等因素进行分配。

依托全国数据标准化技术委员会技术文件和可信数据空间发展联盟团体标准, 加强标准引领, 促进规范化管理。地方政府或行业组织, 应鼓励数据空间建设单位、运营机构、科研院校等积极参与可信数据空间相关国家标准、行业标准和地方标准制定。

### 3.3 搭建数据生态服务中心

建立可信数据空间的数据市场, 应用市场, 需求大厅等, 以需求导向进行场景挖掘, 供需对接, 数据供给; 同时通过“揭榜挂帅”机制吸引多方参与场景开发, 定期发布行业白皮书及优秀场景案例; 定期举办数据生态研讨会、项目对接会等活动, 加强各主体之间的面对面交流与合作。每季度举办一次数据生态研讨会, 邀请各主体代表分享经验、探讨合作机会。激发多方主体参与, 增强数据空间生态活力。

### 3.4 探索数据生态运营模式

构建可信数据空间推广策略, 第一探索商业模式, 包括免费试用、先用后付、应用分成、会员制、供需撮合佣金等多元商业模式, 提升各个生态主体的参与积极性,形成可持续发展的路径。第二分层培育行业主体, 引导龙头企业牵头,要求链主企业开发核心数据接口, 带动上下游中小企业接入; 扶持中小企业, 提供普惠性工具和服务补贴, 降低参与门槛和成本, 激发创新活力; 设立专项基金孵化第三方服务机构, 包括数据开发、数据经纪、数据托管、审计清算、合规审查等机构接入数据空间, 形成数据全生命周期的服务体系。第三定期举办产业沙龙活动, 行业数据空间峰会, 创新场景大赛等活动, 牵引产学研协同发展。

# 03
# 最佳实践案例

## 1、贵州大数据集团公共数据授权运营空间实践

### 1.1 项目背景

在全球数字文明演进与数据成为关键生产要素的时代背景下，贵州作为全国首个国家级大数据综合试验区，承担着为国家探索数据要素市场化改革路径的战略使命。以建设国家数据要素综合试验区为抓手，贵州着力构建可信数据空间体系，破解数据“不敢流通、不愿流通、不会流通”的难题，为全国推动数据资源转化为现实生产力提供实践样本。

作为“东数西算”核心枢纽，贵州不仅是全国一体化算力网的关键节点，更在探索数据跨区域流通与优化配置方面发挥引领作用。可信数据空间作为关键基础设施，有力支撑“东数西存、东数西算、东数西用”国家格局的形成。贵州数字经济占 GDP 比重达 42%、连续 9 年增速全国领先，正是数据要素驱动发展的生动体现。

依托十年数博会积淀与大数据交易实践，贵州构建了“立法保障 + 算力支撑 + 场景应用”的生态闭环，在全国率先建立数据要素法规制度体系，形成可信流通的“规则高地”。通过推动 24 个行业领域建设高质量数据集与大模型应用，贵州积极响应“数据要素 ×”战略，探索数据赋能实体经济新路径，为全国产业数字化转型提供可复制的“融合范式”，成为中国在全球数字竞争中的重要试验田。

### 1.2 解决方案

作为贵州省首个公共数据授权运营空间，是一级授权运营机构承接公共数据授权运营工作的重要载体，以“1+7+2+N”为总体框架，打造授权运营体系，该体系以《贵州省公共数据授权运营实施方案》为核心指导，围绕七大体系机制与两项规范制度展开建设，最终落地 N 个主题应用。“7”具体包括数据归集治理体系、数据产品体系、AI 数据集体系、数据流通体系、数据运营体系、数据安全体系以及数据工具体系；“2”是指运营生态制度和技术标准规范。通过多层级、全流程的机制设计与技术保障，该体系构建了安全可信、高效协同的数据运营环境，全面促进公共数据要素的有序流通、深度融合与创新应用，为贵州省数字经济发展提供坚实支撑。

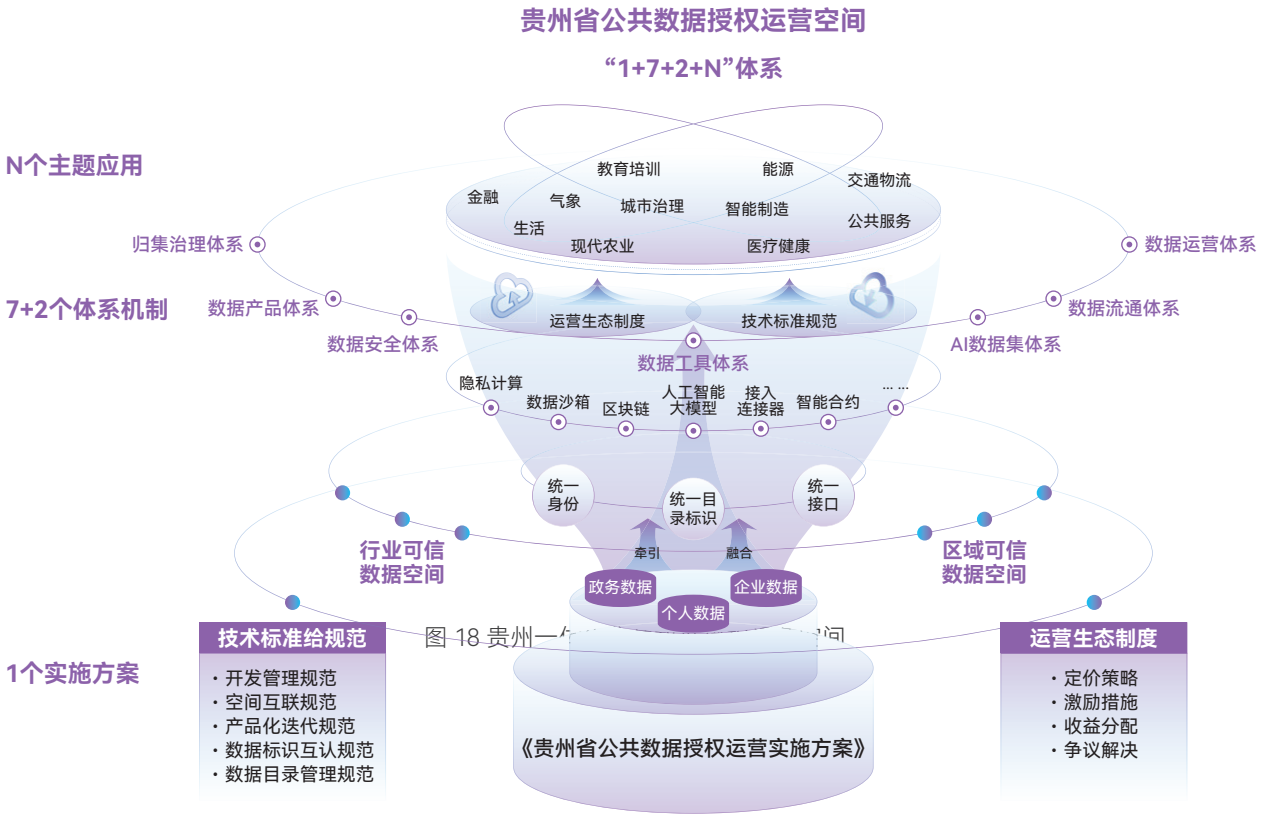



图 19 贵州省公共数据授权运营空间“1+7+2+N”体系

· 技术侧围绕可信空间技术系统要求，空间具备以下四大重点能力。一是一体化协同能力，实现与省公共数据服务平台的数据目录同步、资源接入和业务协同，支持场景申请、产品出域等审批流程，并提供智能化的审批辅助；二是全链路开发能力：提供数据清洗加工、产品建模、出域审核等功能，支持多源数据接入、低代码工作流编排、自定义算法开发与模型上架；三是多租户运营能力，支持多级机构按角色权限认证使用空间资源、工具与数据，可提供授权运营域内跨区域跨层级的业务拓展；四是价值运营能力，具备产品审核、计量计费、运营指标分析等功能，实现数据价值的共创共享。

· 在业务侧，该空间以区块链、隐私计算及使用控制等技术为底层支撑，采用“开发工具库 + 业务管控台”的双轮驱动模式，并结合“采、治、开、运、流、安”六项流程开展数据业务工作。其中，数据治理工具库涵盖多模态语料标注、数据清洗等功能；数据建模工具库则集成了多方安全计算、联邦学习、原子模型开发等先进工具，可支持数据产品与应用的快速构建；安全工具库配备数据沙箱、数据保险箱等工具，确保数据的安全可信；业务管控台方面，其提供租户空间管理、工具整合及运营支撑服务，同时搭载授权管理、计量计费等功能，能够有效破解数据定价与计量的难题。

· 运营侧，创新“政府 + 授权运营机构 + 市场主体”协同模式：政府负责政策引导与合规监管，授权运营机构承担数据治理与产品开发，企业用户通过空间获取标准化数据服务，形成“数据供给 - 开发 - 流通 - 应用”的闭环生态。


图 20 贵州省公共数据授权运营空间安全能力

• 安全侧在安全防护方面，平台侧构建了“双闭环”数据安全体系，通过技术实现数据“出域不离域”，管理上做到审批、预警与溯源闭环。平台安全采用“三位一体”架构，符合国家三级等保，部署零信任与密码服务，强化基础安全；在数据流程中实行分级管控，借助安全网关、云桌面、API 监测等手段，确保各级数据合规使用；并通过安全监管平台和行为分析，实现全流程可追溯。线下场地实行分区管理，按安全等级设置权限，并配备视频监控与防偷拍设备，保障物理环境安全。

### 1.3 项目成果

基于贵州省公共数据服务平台已经归集的 14.7 亿条公共数据资源，通过授权运营空间已形成 80 个公共数据产品。贵州将持续探索“政府 + 授权运营机构 + 开发利用机构”的公共数据授权运营协同机制，加大力度开展大模型与空间融合等技术攻关，培育带动一批数据技术和产品服务商，将平台打造成支撑构建全国一体化数据市场的重要载体。

## 2、贵州省旅游可信数据空间及大模型实践

### 2.1 项目背景

2024 年, 贵州省提出打造世界级旅游目的地, 大力推动智慧旅游发展。2025 政府工作报告指出, 持续提升旅游业发展质效, 大力发展智慧旅游, 加快“省旅游数智互联平台”等项目建设。

同时,《贵州省推动人工智能高质量发展行动方案（2025—2027 年）》提出: 1) 实施数据资源建设行动, 保障高质量数据供给。2) 实施行业大模型发展行动, 促进产业转型升级。3) 实施数字生态优化行动, 激发改革创新活力。

在文旅数据极度分散的情况下, 围绕游客的核心体验作为大模型构建的切入点, 打造游客端服务智能体。同时聚焦智能行程规划等复杂场景, 解决游客在旅游攻略耗时耗力的痛点, 通过构建智能体去集成相关信息、旅游的资源和服务, 实现贵州特色资源与游客需求的匹配, 构建以用户意图为核心的智能交互, 推动传统依赖于平台为中心的模式从“人找服务”的模式向“服务找人”模式进行转变。

### 2.2 解决方案

贵州旅游可信数据空间建设实践, 可总结为以下六大关键步骤:

(1) 从用户的视角设计主要数据维护: 基于多年实践, 和对旅游产业生态、群众服务的深度研究, 结合国家文旅行业数据分类方法, 以用户视角设计了一套新型数据标准。从四项核心维度完成数据汇聚, 以景区为例包括: 位置与设施、门票、主要景点与项目, 综合评价等, 同时将数据进行精细化标签管理。

(2) 建立“四维一体”的数据采集体系: 通过融合政府公共数据确保数据可信性、互联网公开数据确保场景丰富性、经营系统的运营数据确保动态准确性、人工填报数据确保细节完备性,形成一套全面、可信、精准可用的贵州文旅数据体系, 加快推动构建以数据为关键要素的数字经济, 发挥数据的基础资源作用和创新引擎作用。

(3) 形成一批文旅行业数字化转型工具、标准化接口, 让数据“供得出”: 通过搭建酒店智能体、景区智能体等数智化工具, 解决文旅行业中很多中小企业数据供不出的难题。同时, 开发多类标准化的接口, 以支撑供应商的系统连接。

(4) 构建省市两级联动的互联平台, 建立数据高效供给体系: 由市州运营平台负责联通本地各类旅游服务企业和商品, 组织做好本地交易结算和落地服务; 省级枢纽平台负责聚合全省旅游服务资源数据, 建成充分开放的省级本地网, 运用旅游服务智能体, 拓展全网和市场渠道, 为实体企业降本增效, 驱动交易、资金、数据的本地归集。

(5) 打造数据可信的流通体系: 依托省旅游数智互联平台, 逐步构建数据要素平台与可信数据空间管理平台, 推动数据要素价值释放。以场景需求为牵引, 基于各方共识规则, 完善“采 - 治 - 算 - 流 - 用”全生命周期的数据安全与合规管理, 赋能行业大模型提供更精准的游客画像和更好的产业服务。

(6) 建立开放共享的数智化生态，推动旅游产业转型升级。加强和各市州文旅部门以及各类涉旅主体以及行业协会的沟通对接，加强横向沟通、纵向贯通，横向上积极探索与携程、同程、美团、高德、一码游贵州、贵客荟等本地旅游平台和 OTA 平台合作，纵向上加强与市（州）、县（区）以及各涉旅市场主体的沟通，创新合作模式，共建旅游生态，共同推动平台的建设和发展，共享平台建设成果。


图 21 贵州省旅游数智互联平台暨旅游行业大模型总体架构

## 2.3 项目成果

依托旅游大模型，打造了“黄小西”旅游智能体生态，形成一套完整的开发和销售体系，通过“自主创新与开放合作并重”的技术路线，围绕“游客服务个性化、企业运营智能化、行业治理精准化”领域形成智能体产品矩阵。一是 C 端智能体产品，为游客提供行程规划、快捷订购、AI 伴游、游记生成等应用场景，实现行程规划到一键下单的完整闭环体验；并支持轻量化嵌入各类旅游服务平台，链接各类旅游企业并提供线下服务，随时随地为游客的贵州之旅提供便利。二是 B 端智能体产品体系，为酒店提供智能客房服务、智能问答、周边地图、旅行安排等功能，帮助酒店实现“经营自己”与“经营周边”的双向提升。同时景区智能体正在开发中，打造以 AI 为核心的游客体验系统，覆盖智能导览、智能讲解、景区购票、景区活动等领域。

## 3、上海数据集团城市数据空间实践

### 3.1 项目背景

作为全球数据汇聚流通的重要节点，上海在探索数据要素市场化的过程中，提出了自己的“上海方案”。为此，上海率先于 2022 年成立数据集团，打造数据要素市场化配置的核心载体，加速公共数据和国企数据的要素化。上海数据集团是以数据为核心业务的具有功能保障属性的市场竞争类市属一级国企。作为上海市公共数据授权运营主体和城市一体化大数据资源基础治理的支撑主体，以推进数据要素市场建设、激发数据要素潜能、保障数据安全为战略使命，以促进公共数据、社会数据、个人数据融合开发利用为主责主业，聚焦数字产业化、产业数字化和推动数据产业生态发展，践行“数据治理体系共建者、数据资源体系开拓者、数字经济发展引领者、数字政府建设推动者、国际数据合作先行者”的责任担当，致力于成为世界一流的数据要素型企业。通过整合公共数据空间、企业数据空间和个人数据空间，利用创新的技术寻找数据要素的价值场景，释放数据要素的生产力，帮助上海各政府机构、本地企业、民众挖掘和赋能数据要素的价值，为此联合华为云 Stack 打造“城市数据空间”新范式。

### 3.2 解决方案

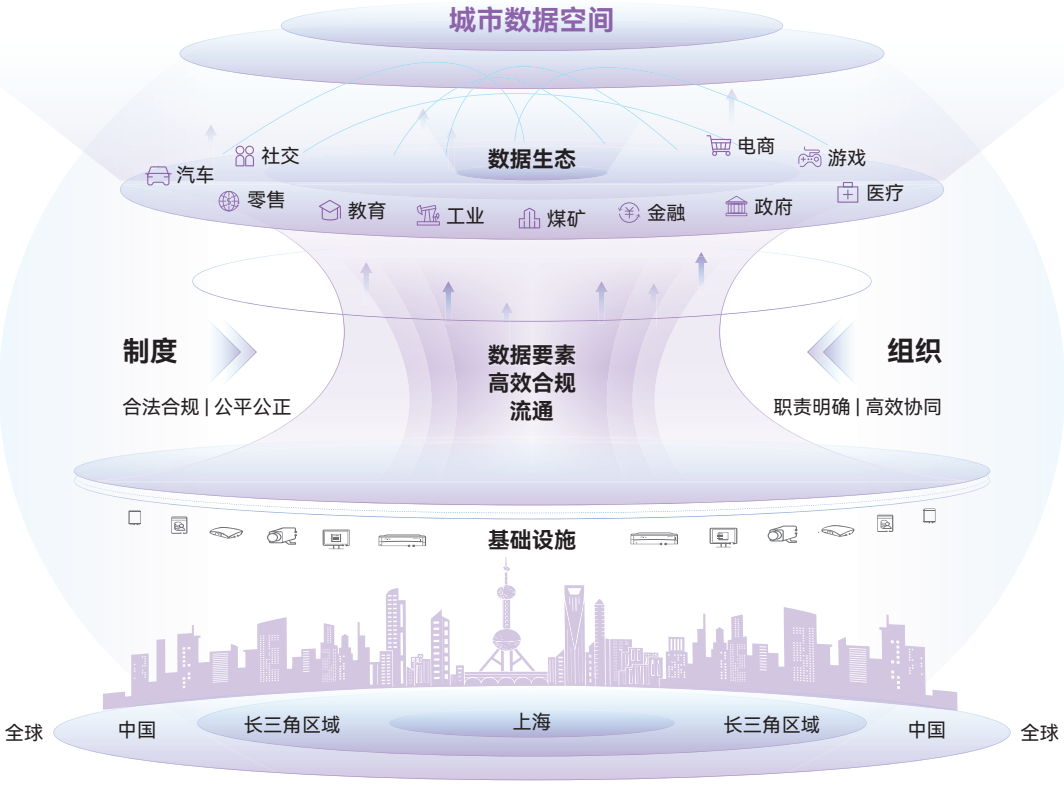



图 22 上海城市数据空间解决方案体系架构

上海城市数据空间体系架构由四个部分组成(2+1+1)首先是 2 个保障体系 (制度体系和组织体系) , 其中, 制度体系从立法和法规角度, 发布城市数据空间相关政策制度, 构建从地方数据条例、管理办法、实施细则到地方标准的四位一体式法规体系, 提供全方位保障。组织体系, 从地方数据局到协同各级责任主体及标准委员会, 形成清晰的组织架构保障。其次是 1 个基础设施, 包括基础硬件和基础软件, 是支撑上层数据生态进行数据共享、开发、上市和流通的大平台。最后是 1 个数据生态, 包括参与到数据要素全生命周期流通里面的所有参与方, 以及为这些参与方构建起来的生态培育环境。城市数据空间体系架构的四个组成部分, 相互依赖又有机协同。制度和组织体系是空气和水分, 基础设施是肥沃的黑土地, 数据生态是繁荣的枝叶, 数据要素高效合规流通是经脉主干道。

2023 年, 上海数据集团以公共数据为牵引, 构建城市数据空间的关键基础设施——“天机 • 智信”平台。采用技术领先的湖仓一体、存算分离架构, 满足以公共数据为牵引, 融合企业数据、行业数据等多源数据汇聚、治理和开发利用, 提供面向数据治理、数据产品、数据服务、数据应用的开发工具。围绕数据全生命周期, 提供信任安全和授权运营的管理能力, 以促进数据的社会化利用。“天机 • 智信”平台深度融合区块链、隐私计算等关键技术, 依托“浦江数链”、“数字信任”体系提供身份可认证、访问可控制、授权可管理、安全可审计、过程可追溯的关键技术能力, 打造城市级数据空间基础设施的标杆和示范。

“天机 • 智信”平台打造“1+2+4+X”整体架构, 如下图所示:

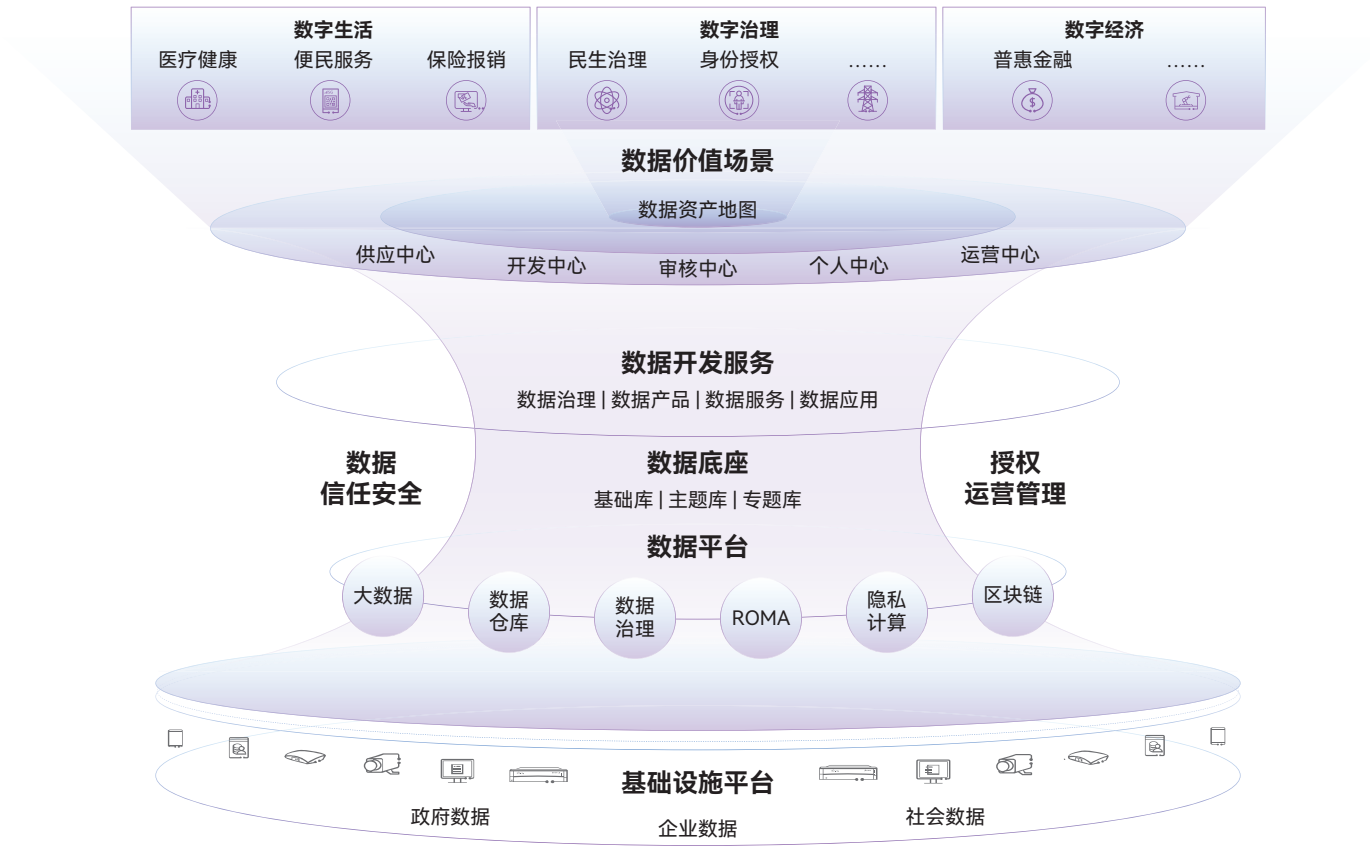



图 23 “天机 • 智信”平台“1+2+4+X”整体架构

• 1 个数据底座: 采用自主创新、安全可信的技术路线, 构建统一的数据汇聚、存储、治理加工、运维管理能力的数据底座;

• 2 套生命周期管理: 建设数字信任和数据安全体系, 实现数据全流程汇聚、采集、存储、加工、服务、使用的安全保障能力, 并在对外数据要素流通中通过可信计算系统实现数据不动、算法和模型可动, 避免数据外泄, 充分保障了数据隐私安全。通过可信存系统实现数据授权、数据使用、数据目录、数据服务等全流程的存证留痕不可抵赖第三方审计, 为数据社会化利用提供安全和运营支撑;

• 4 类数据开发服务: 提供数据治理、数据产品、数据服务、数据应用 4 类工具, 为数据标的提供开发支持; 提供对主数据、元数据、数据模型、数据标准、数据质量进行管理能力, 形成整体多类数据的元数据管理、模型管理、数据质量管理等数据管控体系, 打造数据资产管理能力;

• X 个数据价值场景: 面向上海城市数字化转型, 重点面向数字治理、数字经济及数字生活, 支持 X 个数据价值场景的对外服务和发布。

### 3.3 项目成果

当前, 以普惠金融场景为例, 上海数据集团已经成功开放超过 3000 项公共数据, 向 33 家金融机构提供数据标准化服务, 帮助金融机构优化信贷评估模型, 提升评估效率, 为中小微企业完成了超过 6800 亿元的信贷评估发放, 缓解中小微企业融资难, 融资慢的问题。对于政府来说, 也改善了区域营商环境, 为社会经济长效发展注入动力。

面向未来, 上海数据集团将持续推进城市数据空间创新, 以驱动城市数字经济高速增长。以全上海的城市数据授权运营为目标, 实现公共数据、企业数据及其他数据的汇聚、供给、授权、运营及市场化开发利用, 服务更多城市应用场景。

## 4、深圳南山数据可信流通服务探索实践

### 4.1 项目背景

南山区积极推进数据要素可信流通服务平台建设，通过建设数据云底座、公共数据授权运营平台、数据可信加工域、应用场景、运营机制等，有效提升了公共数据开发效率，促进了数据流通利用，探索形成了一批数智融合开发应用实践成果，包括医疗健康、科技创新等场景，有效赋能区域数据产业的发展。在案例可研和实施的过程中,需破解以下关键难题与瓶颈问题。

一是公共数据授权运营制度流程待完善。开展公共数据授权运营需要解决以下难题，一是确定区域授权运营模式，依据南山区实际情况，建立可发挥数商主观能动性、激发数据开发活力的运营模式；二是建立场景需求审批的跨部门协同机制，高效开展供需对接，三是建立合规审核机制，完善对开发主体资质审核、场景的合规性审核、数据产品的出域审核机制等；四是建立全流程的安全防护机制，包括开发方管理、安全要求、监督机制、应急预案等。

二是数据应用场景待拓展。公共数据应用主要集中于政务、金融信贷等少数领域，跨行业数据融合场景匮乏，行业间的数据壁垒仍存在；场景开发在广度与深度上均显不足，缺乏创新性、可复制的成熟应用场景，数据价值释放不充分。

三是数据可信流通技术待突破。在公共数据和行业数据开发利用及流通过程中，数据泄露、二次散播、本地存留等自身权益不能保障的问题,使数据提供方不愿、不敢或不能提供数据，阻碍了数据要素流通产业发展。需加强数据安全技术能力建设，灵活运用于数据流通应用场景，保障数据主权，使跨域、跨体系、跨行业数据“可用不可见，可见不可得”，促进数据要素市场繁荣。

### 4.2 解决方案

南山区通过建立数据要素可信流通服务平台，完善公共数据授权运营体系，统筹规划建设数据基础设施，探索试点行业数据要素创新应用，最终实现数据资源的优化配置和高效利用。


图 24 南山区数据要素可信流通服务平台总体架构

南山区政务服务和数据管理局牵头制定了《关于数据要素赋能南山区高质量发展的实施意见》，系统提出南山区数据要素生态发展的整体规划，确定了南山区数据基础设施建设思路。包括对数据要素业务运营环境的构建，依托数据湖、数据仓库，一体化数据开发平台等数据底座，应用区块链、隐私计算、数据空间等关键技术，持续构筑数据开发利用和可信流通能力。以上共同形成南山区坚实的数据基础设施底座。具体措施主要有以下几点：

1) 建设南山区数据要素可信流通服务平台：

南山区数据要素可信流通服务平台作为南山区公共数据开发利用和行业数据空间的战略载体，通过“**1+4+N**”的架构规划，推动南山区数据要素产业的发展。

“**1**”是指一套区公共数据授权运营系统，基于区数据要素可信流通服务中心，开启数据要素价值释放的新篇章。主要建设方案内容包括：

• 夯实统一数据底座：进一步完善数据底座能力，通过补充高性能数据湖存算能力、完善分布式数据仓库环境、引入通用关系型数据库、区块链等，为数据的存储和管理提供坚实基础。

• 构筑数据开发平台：引入一站式数据开发利用平台，集数据采集、汇聚、加工、流通、安全等能力于一体，实现端到端数据开发处理，高效支撑数据产品的开发和流通利用。远期规划语料加工、数据标注、AI 算法开发与场景化建模等 AI 相关开发能力。

• 建设流通利用能力：补充数据 API 服务网关、多方安全计算及联邦学习模型开发等隐私计算能力，以及可信数据空间连接器能力等，满足多种场景下高价值数据产品的跨域安全流通需求，实现“数据可用不可见”，“数据可见不可得”，“数据不动价值动”。

“**4**”是指对 4 类数据服务的信息检索，分别是“找数据”“找算力”“找服务”和“找政策”；4 个找满足了对南

山数据要素相关资源或服务一站式便捷查找和供需撮合，推动数据产业信息共享，促进各主体围绕流通服务平台高效开展作业，有利于数据生态汇聚和产业培育及发展。

“**N**”是指 N 个行业数据空间的特色专区。专区充分结合南山区的发展规划和产业分析，以成熟一个规划一个的方式，逐步推动南山区特色产业专区的建设，现阶段规划的是人工智能、低空经济、医疗健康、具身智能等专区。

2) 探索可信数据流通创新场景:

南山区切实推动数据要素 × 千行百业行动。协同发改、工信、科创、卫健等业务主管部门，打造第一批 40+“数据要素 ×”应用场景，推动数据要素深度融入重点领域。

结合南山区丰富的医疗科研资源、医疗产业基础，南山区先行先试规划医疗健康领域数据空间专区，构建医疗数据合规、安全的统一供给渠道，通过隐私计算、可控数据交换等数据流通技术，帮助科研团队等机构在数据“可用不可见、可见不可得”的前提下，使用医疗数据完成科研分析，提升医疗科研创新能力。

3) 完善一套合规高效的运营机制:

出台实施《深圳市南山区公共数据授权运营管理暂行办法》，明确由区数据主管部门建设公共数据授权运营统一通道，实现授权运营工作集约化管理；全面落实“首席数据官”，明确专人负责本机构公共数据资源管理；编制包括公共数据授权运营平台用户管理、公共数据资源处理、开发利用、授权运营安全管理、应急预案等共 18 个运营配套制度文件，保障公共数据合规、高效流通。

### 4.3 项目成果

南山区数据要素可信流通服务平台利用大数据、数据仓库、隐私计算、区块链等技术，提升平台能力。采用先进湖仓技术支持对海量公共数据进行高效挖掘分析，加工成高价值的数据产品；规划数据工程能力，为大模型训练提供高质量语料集；利用区块链技术保障数据的安全性和可追溯性，确保数据在授权运营过程中的可信度，并结合隐私计算、可控交换等数据流通技术，为数据“供得出”“流得动”“用得好”“保安全”提供技术保障。

在“数据要素 × 医疗”场景中，打造“数据 + 平台 + 场景”一体化的医疗数据安全流通解决方案，当前可信流通服务平台已连接南山医院等各方主体，实现了医疗健康领域的数据可信互联互通。通过医疗行业数据高质量供给，支撑重点应用场景创新，如:

• 在医疗科研场景，医疗影像数据经过数据开发方的标注之后，形成高质量数据集，并基于可信数据空间实现数据集在医疗机构和科研机构间可控交换，辅助诊断模型训练效率大幅提升。

• 在医疗健康场景，基于医院的诊疗记录、住院结算费用等数据，在病人离院结算时，实现商业保险公司的直接赔付，借助合规开放就医数据，实现商保快赔平均 30 分钟内、商保直赔最快 3 秒内免材料赔付，降低保险公司理赔调查成本 50% 以上。

医疗数据空间专区激活了医疗健康行业数据，形成可开发利用、可流通的高价值数据资产，逐步赋能健康数字经济产业的应用落地，支撑建立健康数字经济生态圈。未来南山区将继续规划人工智能、具身智能、低空经济等产业专区，促进各新兴产业的数据流通利用，促进产业的茁壮发展。

## 5、华为企业数据空间探索实践

### 5.1 项目背景

随着数字化转型的推进，" 以数据为核心 " 的理念正在重塑企业的能力，其效果正逐步显现。在这一过程中，企业内部的数据流通和消费模式也在发生变化。过去，数据主要沿着企业内部的一条或多条业务流程进行集成和共享。而现在，这种模式正在向跨主体、跨边界的数据流通和交换转变。

数据在企业中的角色也经历了转变，从仅仅是 IT 系统的一部分，逐渐演变为一种战略资产，最终成为企业生产的关键要素。然而，数据的非排他性和无损复制性等固有特性，也为企业的数据流通带来了新的挑战。这些特性要求企业在数据管理上采取更加灵活和创新的方法，以确保数据的安全、合规和有效利用。

为应对数据作为生产要素流通的挑战，华为提出了创新的数据空间解决方案。该方案的关键在于通过技术手段确保数据交换协议得到有效执行，以解决数据的非排他性和可复制性问题，并保障数据流通过程中合约的遵守。数据空间整合了数据连接器、注册认证服务、使用控制服务和清算服务等核心功能模块，并建立了一套完善的“可信、可控、可证”的技术体系。

华为的鲲鹏 / 昇腾生态链覆盖了从底层硬件到基础软件，再到上层行业应用的全产业链。这一生态系统具有参与方众多、业务覆盖面广、数据商业价值高和协同需求强等特点。其中，数据在生态链中的自由流通是关键需求，它有助于促进数据价值的释放和支撑技术创新。基于数据空间的理念，项目针对生态业务中的特定问题和需求，构建了鲲鹏 / 昇腾数据空间，并在联合研发、供应链协同、质量管理追溯等场景中实现了有效应用，充分发挥了数据要素的价值。

### 5.2 解决方案

针对鲲鹏 / 昇腾生态业务中的问题和诉求，华为构建了 EDS (Enterprise Data Space，企业数据交换空间)，实现了数据流通的“可信、可控、可证”。

1) 构建跨主体数据流通的信任机制：

• 生态业务各参与方的身份认证。CA (Certificate Authority，认证中心) 经过评估认证，向参与方颁发数字证书，该数字证书是唯一的身份标识。

• 访问鉴权与可信通信。数据提供方与数据使用方技术部件之间的通信，基于 PKI (public key infrastructure，公共密钥基础设施) 完成鉴权、数字签名，从而保证数据传输的安全性、完整性和实名性。

• 安全、平等、可信的执行环境。这个环境基于容器管理技术和基础安全能力打造而成。

2) 构建数据使用控制机制，实现全流程的精细管控：

• 数据使用控制策略：基于“4W2H” (Who、When、Where、Do What、How To、How Many) 设计原子策略，灵活组合，实现精准的数据使用控制。

• 全流程精细管控：全流程操作通过日志记录 + 区块链存证 + 全链路血缘，实现数据提供方可查证追溯、数据使用方可自证清白、数据监管方可监管审计。

3) 建跨主体的数据交换合约，并通过技术手段保证合约的有效履行：

• 跨主体的数据交换生成智能合约，基于契约化的合约，实现双方要求与承诺的 IT 化，并通过区块链进行存证，防止篡改。

• 合约中包含的数据使用控制策略，通过技术手段被 IT 强制遵守，避免人工执行带来的不可控风险。

### 5.3 项目成果

鲲鹏 / 昇腾产业生态数据空间于 2021 年 9 月上线使用，目前已包含 15 个参与方（华为 2 个，生态伙伴 13 个），上架 25+ 数据交换资源，应用 21+ 使用控制策略，累计数据交换量 14000+ 次。

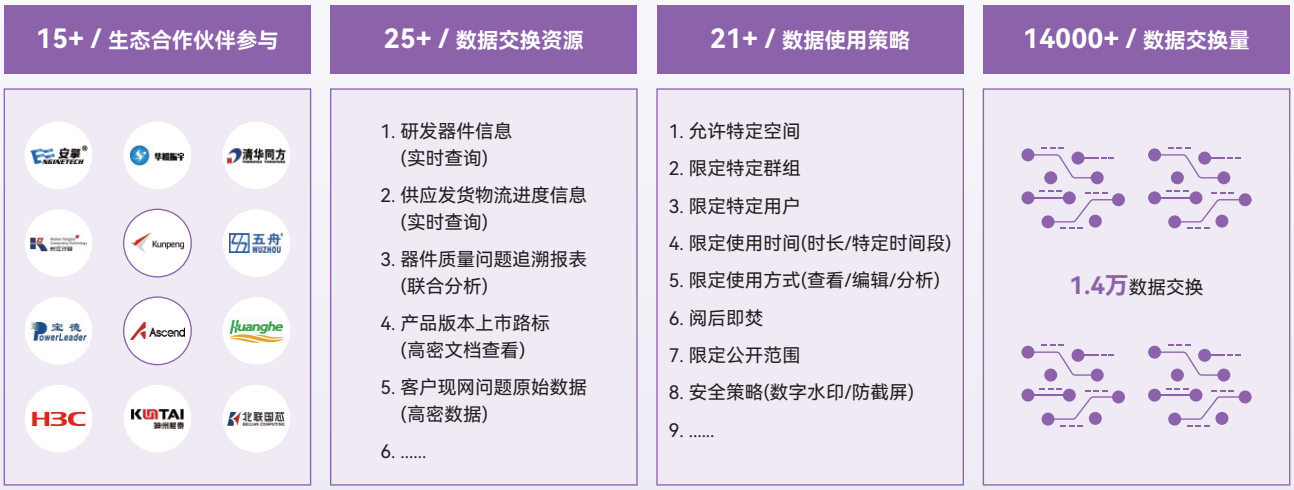



图 25 鲲鹏 / 昇腾生态数据空间业务概况

鲲鹏 / 昇腾产业生态数据空间的使用，有效支撑了华为与生态伙伴在研发、生产制造、产品服务等领域的业务深度协同，并积累了 4 类典型应用场景：

**（1）高密的资料文档交换。**主要应用于研发、销售等领域，如涉及芯片 / 部件相关的技术文件、研发路标规划等，为非结构化数据文件，如 Word、PPT、Excel、PDF 等文件，从华为向生态伙伴单向传递。属于华为高密数据，需要严格控制使用范围（仅具体合作项目成员），只允许查看。

**（2）敏感业务信息共享查询。**主要应用于制造、物流、服务等领域，如华为向整机厂商供货的物流计划与状态数据，为结构化数据，从华为向生态伙伴单向传递。属于华为受控数据，需要控制数据使用范围（伙伴采购相关岗位人员），按需即席查询。

**（3）多方数据联合分析。**主要应用于质量追溯、运营分析等场景，如汇聚各环节产品质量问题数据，进行质量联合追溯，为结构化、非结构化数据，需要华为与生态伙伴互相交换，并支撑端到端的质量追溯及运营分析。属于机密数据，需要严格控制数据使用范围（双方质量运营团队），允许数据整合分析。

**（4）重要项目的互动协同。**主要应用于能力评估协同、伙伴能力提升等场景。如华为帮助生态伙伴进行能力评估，华为向生态伙伴提供评估项，生态伙伴反馈举证数据，基于评估结果制定提升计划，为半结构化、非结构化数据，双向多次交互。属于高度机密数据，需要严格控制数据使用范围（评估项目专家团队），允许查看和编辑。

# 04
# 参考引用

[1] MoonFox & 中欧 AI 与管理创新研究中心,《AI 产业全景洞察报 2025》, 2025

[2] Stanford HAI ,《2025 年人工智能指数报告》, 2025

[3] 深圳市工业和信息化局,《深圳市加快推进人工智能终端产业发展行动计划 (2025—2026 年) 》, 2025

[4] 华为公司,《AI 大模型技术和趋势洞察 2.0》, 2024

[5] 上海数据集团有限公司, 华为云计算技术有限公司,《城市数据空间 CDS 白皮书》, 2023

[6] 中国国际问题研究院,《国际问题研究》, 2024

[7] 华为云开发者联盟,《分布式身份: 重新定义你的“身份”管理》, 2021

[8] https://www.ibm.com/cn-zh/think/topics/data-engineering, IBM Think 主页, 2024.5.31

[9] 数据要素白皮书, 中国信息通信研究院, 2023.9

[10] “用基础制度破解基础难题, 开启数据要素价值释放新时代”, 政策解读, 中华人民共和国国家发展和改革委员会, 2022.12.20

[11] 当贝 AI 生态扩容: 通义 Qwen3-235B 多模态能力开放: https://tech.china.com/articles/20250723/20250723231703758.html

[12] 2025 年 AI 语料行业现状及未来发展趋势预测: https://www.chinairn.com/hyzx/20250725/160006166.shtml

[13] 上海量子城市突破: 专项语料库助力 AI 治理, 精准“锁”住幻觉: http://www.itbear.com.cn/html/2025-07/894602.html

[14] 每日互动发布 GAI Station 智能工作站, 破解企业 AI 应用私有化部署难题: http://www.itbear.com.cn/html/2025-07/895468.html

[15]《Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics》, Michael Armbrust, Ali Ghodsi, Reynold Xin, Matei Zaharia