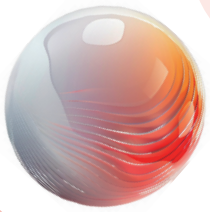
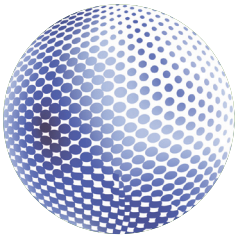
HUAWEI

迈向智能世界白皮书2024

数据存储

数据是数字化到数智化成功转型的关键要素





构建万物互联的智能世界

序言：

人类出现在地球上，已有数十万年的历史，但人类文明真正的高速发展时期也仅仅是最近几千年的时间。这里最关键的是纸张的出现，使得人类可以记录经验和知识，并借助纸张进行群体分享、学习、演进、发展，直接促进了人类社会文明的高速发展。值得一提的是，以前在中世纪欧洲采用羊皮进行重要文字的记录，当时一本书上千页，需要数百张羊皮来制作，是十分复杂和昂贵的，不利于知识的记录和传播。

在人们发明以数字化形式保存和传播信息后，人类进入数字时代，而数字化形式的信息则被称为数据。因为数据可以被高效处理，这促进了数据爆发式增长。而数据存储作为保存数据的载体，如同新时代的“纸张”，开始加速人类文明的跃迁。我们是新时代的数据存储缔造者、完善者、创新者，制造出面向数字化世界的“新纸张”。

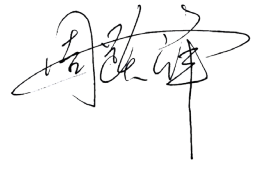
缺数据，不 AI。伴随着 AI 大模型持续成熟并走向多模态，数据逐渐成为 AI 的关键，因为 AI 是以电脑模拟人脑的思考方式、从海量数据中发现规律、总结知识，再将这些知识融入不同的业务场景，生成业务咨询师、编程机器人、智能客服等，让它们拥有自主学习的大脑并实现自我进化。人工智能角逐的决胜因素是数据的产生、保存和使用。

华为公司在数据存储产业上的大规模投入超过十五年，产品已进入全球超过 150 个国家和地区，广泛服务于金融、运营商、政务、制造、电力、能源、医疗、科研教育、交通等多个行业，在全球拥有超过 26,000 家客户。通过与业界专家、客户和伙伴深入交流，我们编写了《迈向智能世界白皮书 2024– 数据存储篇》白皮书报告，结合数字化、智能化，展望数据存储在各行业中的发展趋势和挑战，并提供行动建议。我相信这是一次有意义的探索，将凝聚更多的产业力量共同推进数据存储产业的发展。

过去三十年，新技术、新应用不断涌现，产生了海量数据。数据存储为这些数据提供了一个温馨的“家”，帮助这些新技术、新应用持续成长。华为数据存储产品线愿与社会各界更加紧密携手努力，汇聚产业力量，为更多新技术、新应用提供先进数据存力，共创数据存储美好未来。

周跃峰

华为数据存储产品线总裁



CONTENTS

目 录

CONTENTS

执行摘要

规模定律（Scaling Law）揭示了 AI 人工智能在当前深度学习算法框架下，算力和数据之间的关系：更强的算力加上更多的有效训练数据，可以得到更好的 AI 大模型。在规模定律的支持下，AI 大模型由单模态走向多模态，同时大模型能力和性能持续提升，这帮助了 AI 逐步走出中心训练、走向千行万业并得以应用，从办公辅助逐渐走向生产决策，从降低成本逐步走向增加效率，从管理当下逐渐走向预测未来，从高容错场景逐渐走向低容错场景，不断引发各行各业智能化转型和业务变革。在这个过程中，人们逐渐发现，进一步深化并加速业务数字化转型、以产生数量更多、类型更丰富的高价值数据，其重要性对于 AI 而言，不亚于唤醒历史沉睡数据。数字化和智能化以数据为纽带，相互促进、加速和融合，逐渐走向两者相结合的数智化，这对数据基础设施提出了新的更高要求，不断驱动着数据基础设施的演进。

数智化将持续高速发展，并将实现通用人工智能，帮助人类进入一个全新的智能世界。面向未来，我们对数智化必不可少的数据基础设施进行如下展望：

1. AI 大模型走向多模态，算力集群规模和数据规模持续增长，只有算力和存力协同演进、算存比可基于 AI 发展进行灵活调整，才能有效降低系统管理难度、助力 AI 在实际业务场景发挥不可替代的作用。

2. 在 AI 大模型训练阶段，伴随 AI 算力集群规模增长，相邻训练中断的间隔时间越来越短，这带来了更加频繁的 Checkpoint 存档，也带来了更加频繁的断点续训，亟需加速数据访问性能以快速完成 Checkpoint 的保存于加载。与此同时，智能化升级也在加速数字化转型，进而产生更多的业务数据，增加了数字化基础设施处理数据的复杂度和压力。

3. 智能化升级过程中，一方面加速了数字化转型，产生更多高价值业务数据，另一方面降低了黑客门槛，让勒索攻击更加频繁。

4. 伴随 AI 算力集群规模增长，对海量多源异构数据的高效管理逐渐成为 AI 赛道的关键竞争力。数据地图绘制、数据归集、数据预处理等工作，是 AI 大模型训练首当其冲的要务。

5. 千行万业在尝试将 AI 落地到行业应用的过程中，发现面临基础设施部署、大模型选择、二次训练和监督微调等方面的困难。复用基础设施厂商和 AI 大模型厂商的能力，成为千行万业快速落地 AI 的关键。

面向以 AI 大模型为代表的企业智能化新应用，新的数据基础设施架构也正在逐渐形成。为了构建 AI 大模型时代最佳的数据基础设施，我们建议：

1. 重视存算分离架构的灵活性和独立扩展，利用存算分离架构有效简化智算集群管理、让计算和存储分别按需扩展；关注横向扩展、性能线增、多协议互通等数智化时代数据基础设施基本能力。

2. 全闪存是数智化时代提升数据处理效率、满足业务需求的最优解，同时满足不断增长的数字化转型和日益深化的智能化变革；与此同时，配合向量 RAG、长上下文记忆存储等新兴数据范式，可以有效简化数据访问，实现以存强算，提升系统整体性能。

3. 不管是产生了更多数据的数字化，还是持续成长的智能化，均需要构建防治结合数据安全体系，从被动应对攻击走向主动全面防护。

4. 为 AI 算力集群建设 AI 数据湖底座，打破数据烟囱，实现数据的可视可管可用。

5. 针对 AI 大模型在行业场景的落地，用好训 / 推一体机，基于预集成了基础设施、工具软件等部件的一体化设备，并借助 AI 大模型供应商的系统集成能力，有效加速 AI 落地行业应用。

01

数字化快速走向数智化

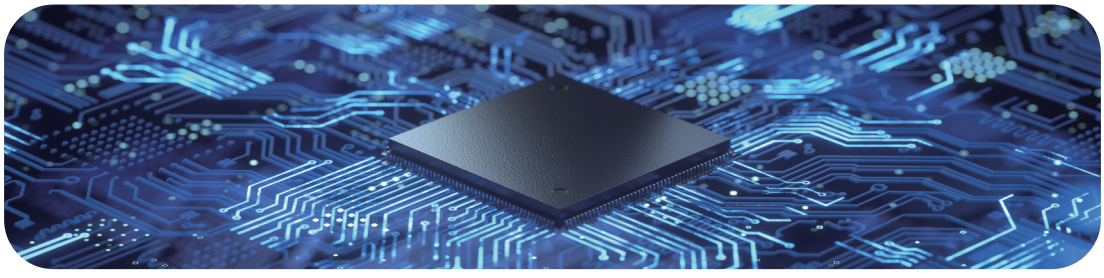
1956 年，人工智能（AI）被确立为一门学科。经过近 70 年的探索和发展，AI 进入了大模型时代，基于深度学习算法，通过大规模算力对海量数据进行学习和训练，从而得到较优的 AI 模型。今天，随着 AI 大模型能力和性能持续提升，AI 正逐渐由大模型中心训练走向行业应用，引发各行各业智能化转型和业务变革。

在 AI 大模型时代，算力、算法、数据构成了大模型训练三要素。规模定律（Scaling Law）揭示了在当前深度学习算法框架下，算力和数据之间的关系：更强的算力加上更多的有效训练数据，可以得到更好的 AI 大模型。

在规模定律的驱动下，人们一边不断构建更大规模的算力集群，一边竭尽所能获取更多的原始数据和训练数据，在实现了由单模态大模型向多模态大模型演进的同时，也在一些面向消费者的业务场景中获得商用。例如最新的智能办公本，除了进行日常笔记和阅读外，还可以进行图片文字识别、会议录音转文字记录、外语翻译、文案生成等多种智能操作，获得广泛接受。

相比大模型训练和 AI 应用于消费者，行业用户则更加关注 AI 大模型如何服务于业务、如何改善内部运营、如何增强竞争力。部分行业用户已在某些场景中找到 AI 的切入点，例如呼叫中心智能客服、医院诊疗助手、在线情景式教育、广告文案辅助生成、工业生产质检、复杂网络智能运维和自动驾驶等，并持续尝试在更多业务场景深入探索。在这个探索过程中，越来越多行业用户发现，AI 在行业的落地离不开高质量行业数据。一方面，行业和场景模型需要使用一定规模的行业数据对基础大模型进行二次训练和监督微调，进而得到一个面向特定行业的垂直模型；另一方面，在推理阶段用于消除幻觉的知识库，同样需要依赖高质量的、具有时效性的行业数据来生成。

可以看到，不管是基础大模型的训练，还是大模型在行业的应用落地，都离不开大规模高质量的数据。数据的规模和质量决定了 AI 智能的高度，也决定了 AI 在千行万业的应用成熟度。



1.1 金融

金融行业在数字化时代领航，开创了 FinTech。今天，AI 大模型与金融行业融合，在数字化所积累的海量数据资产基础上，金融行业具备在数智化时代继续领航的先发优势。以银行为例，正在从办公助手、智能填单等办公辅助逐步走向远程银行、信贷风控助手等生产场景。从对内办公辅助走向对外业务应用，意味着从高容错走向低容错。而正确的建议和选择，需要从海量数据中得出。针对海量数据的高效归集、快速处理、安全可靠，成为了新的挑战。

1.1.1 降本到增效：从办公辅助走向业务决策

金融机构一直是率先将新兴的 IT 技术应用于业务场景的行业。目前，领先金融机构已经纷纷投入人工智能 (AI) 技术，尤其是大模型技术的研发和布局，使能业务运营、产品营销、风险控制和客户服务等业务领域，从而提升金融服务的智能化。根据 IDC 相关报告，90% 的银行已经开始探索人工智能的应用，AI 技术成为银行技术创新的主要方向。

1. 在智能营销场景，通过 AI 技术分析大量的用户数据，并基于客户需求和偏好提供个性化的金融服务。这不仅提升了用户体验，同时增强了客户粘性。如，交通银行利用 AI 技术挖掘客户兴趣偏好，用大模型强化业务端留客能力，各类理财模型策略累计触客成交量近 4 千亿元，较传统方式成交率提升 16 倍。

2. 在智能理财场景，AI 技术通过机器学习和深度学习模型，能够帮助投资者更准确地做出投资决策。江苏农行和中国工商银行分别推出了类 ChatGPT 的大模型应用 ChatABC 和基于昇腾 AI 的金融行业通用模型，用于智能化地推荐理财产品。上海浦发银行则利用多模态人机交互、知识图谱等技术，推出了 AI“理财专家”，为消费者推荐合适的理财产品。

3. 在信贷审批的风控场景，AI 帮助简化和优化了从信贷决策到量化交易和金融风险管理的流程，亚太区域某头部银行通过 AI 技术实现了用户信贷申请过程从原来的数天缩短到只需一分钟完成申请，最快一秒钟获得审批。

4. 智能客服在金融服务中有着显著的应用。以招商银行信用卡公司为例，通过智能客服每天为客户提供超过 200 万以上的在线人机交互，并能够解决 99% 的用户问题。智能客服不仅能提升客户服务效率，相对于人工客服，还能够提供 24 小时不间断服务。

1.1.2 完善多源多元海量数据管理，加强数据安全合规建设

在人工智能应用逐步普及的过程中，金融机构在数据架构、数据安全和业务连续性等方面面临新的挑战。

1、首先，是庞大数据量的管理，金融行业在数据量方面已经达到了 EB（Exabyte，即艾字节）级别。以中国为例，根据北京金融信息化研究所（FITI）2023 年发布的最新报告，目前金融机构的数据量普遍达到 PB 级，其中大型金融机构的数据量超过 100PB，并且未来五年预计年均增幅将达到 24.33%。此外，国有大型银行的核心业务系统存储规模也已达到百 PB 级，票据影像等非核心系统存储规模更是达到了几十 PB 甚至百 PB 级。围绕金融行业海量业务数据，如何实现高可靠、高效率的访问，进一步实现数据价值最大化，是金融机构必须考虑的问题，例如，针对海量的数据量及不同的数据类型，采用高性能的存储设备以及优化存储架构，加快 AI 与金融行业的融合。

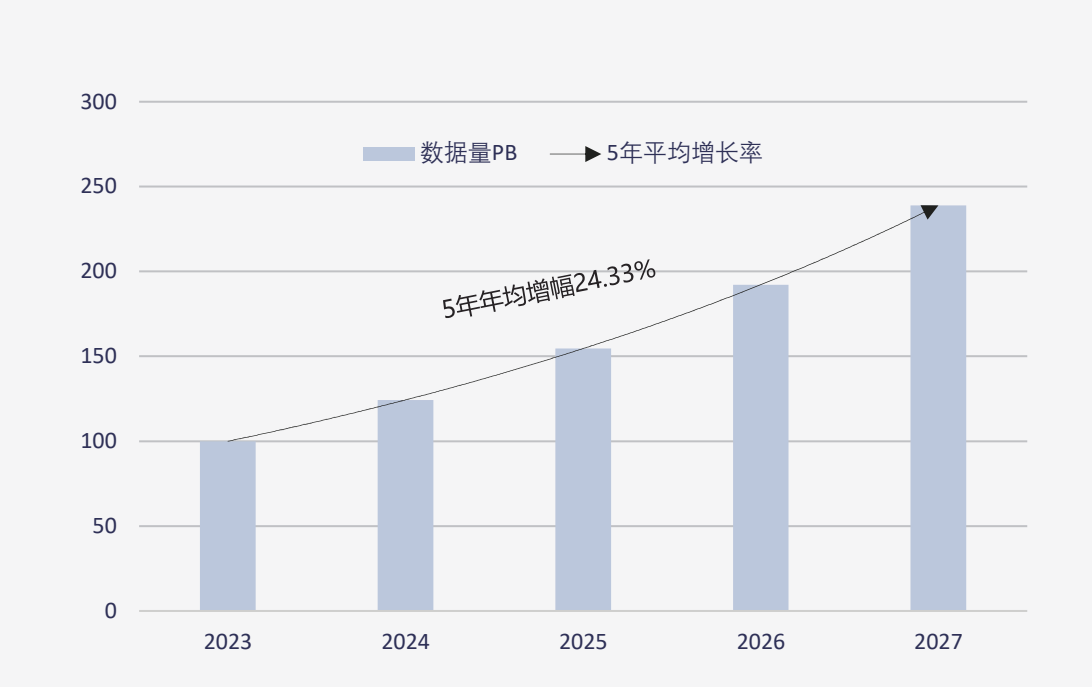


图 1：某大型金融机构数据量年均增长率

2、其次，金融业务需要处理种类多样的数据类型，经过多年的业务积累和沉淀下来的数据，比如：图片，视频，音频，以及互联网日志等各类金融数据，不但数据格式陈旧复杂，而且分散在不同的业务领域，甚至不同的地域。比如大小机核心系统的数据格式无法直接与开放平台的信用卡系统的数据格式进行数据交换；信贷业务，财富管理业务和互联网业务之间很难实现用户信息共享。将这些分散的数据整合并准备用于人工智能应用是一项艰巨的任务，急需建立一个完善的数据管理系统。如中国某头部银行一直将数据视为基础要素和战略资源，在建立大数据资源管理系统方面，面临有哪些数据，数据在哪里，如何有效利用这些数据的关键问题。

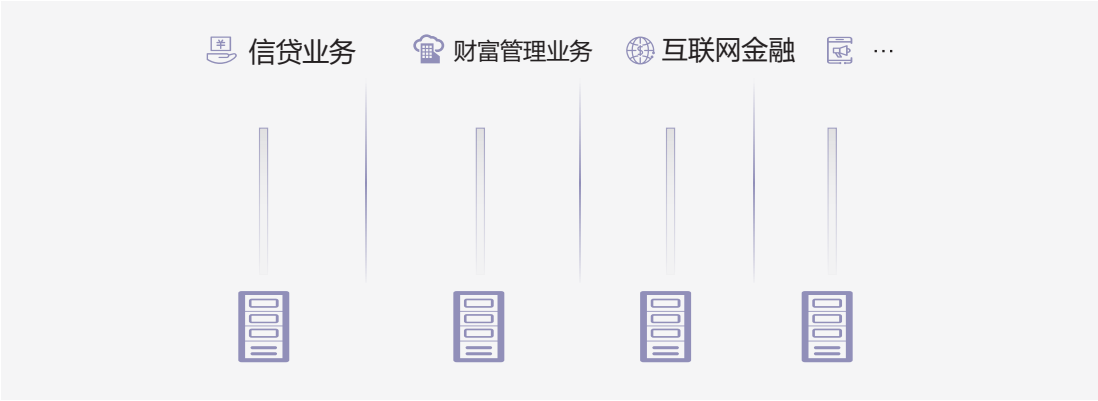


图 2：不同业务数据难共享

3、最后，金融行业数据处理，还必须满足行业监管和风险控制的合规要求。利用 AI 技术进行个性化推荐和精准广告投放的精准营销场景，对数据管理和隐私保护的挑战进一步增大，进而促进金融合规监管的要求提升。同时，人工智能应用增加了金融机构数据泄露的风险。2024 年 5 月，美国某知名银行遭 LockBit 勒索软件攻击，导致约上百万名客户数据被盗。2024 年 6 月，中国国家金融监督管理总局网站公布，中国某头部银行因数据安全管理不足、灾备管理不足，被罚数百万元。因此，以容灾为基本手段的数据物理安全，和以备份为基本手段的数据逻辑安全保障等多重手段在当前 AI 时代显得尤为重要。

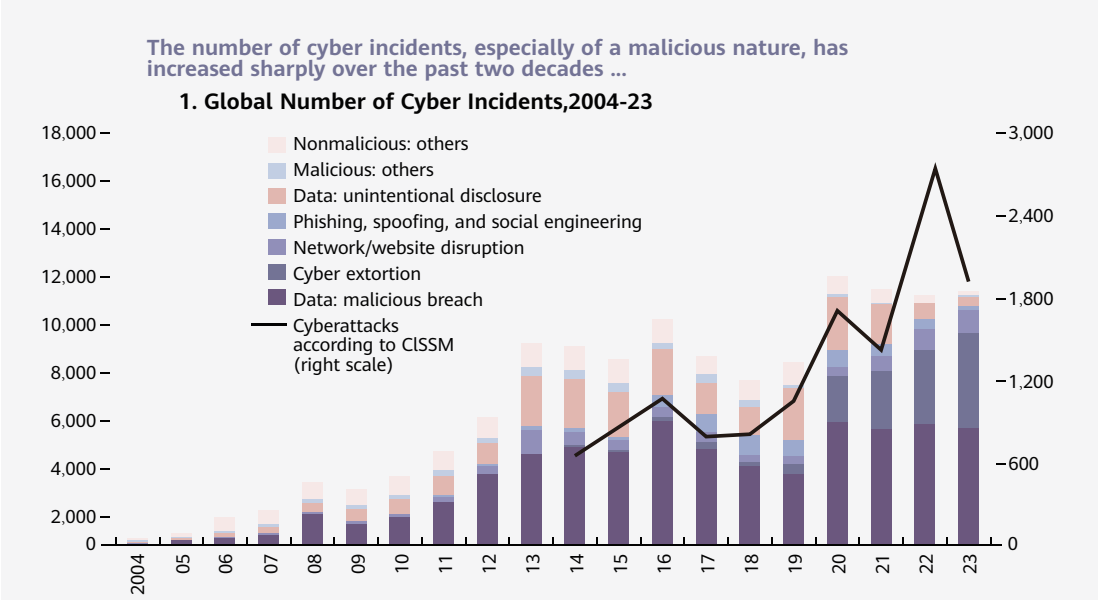


图 3：日益增长的人工智能和数字化应用，增加网络数据安全风险

如上图所示，国际货币基金组织 IMF 发布的《Global Financial Stability Report》指出，日益增长的人工智能和数字化应用，显著增加网络数据安全风险。

因此，金融机构拥抱 AI 新技术应用，重塑服务模式，唤醒数据价值的同时，要关注 AI 技术对数据管理所带来的挑战，才能有效提升金融服务的效率和品质。

1.2 运营商

“从电信企业向科技企业转型”已成为全球大部分运营商数字化转型的战略共识。随着生成式 AI 走深向实，电信运营商作为通信基础设施的建设者和运营者，拥有先天的资源优势、数据优势、行业使能经验优势，既为 AI 的发展提供基础设施支撑，又将会是 AI 应用落地的先行者。

1.2.1 开发到应用：蓄力大模型训推，对内运营增效，对外赋能千行万业

当前全球运营商形成三波 AI 阵营，第一波智能化先锋正在构建“终端设备、智算资源、模型应用”的全栈 AI 能力，如韩国 SKT、中国移动等；第二波运营商积极布局行业大模型，如新加坡电信 Singtel、德国电信、阿联酋 e& 等合资成立全球电信 AI 联盟（Global TelcoAIAlliance，GTAA），以专门开发及推出多语言的电信语言大模型服务话务中心和智慧运营；第三波务实型运营商关注 AI 带来的实际价值，尝试借助第三方合作伙伴的 AI 能力实现降本增效，如 Orange、Vodafone 等计划通过 AI 提升智能客服效率。

未来两到三年，运营商的大部分应用和业务都将被 AI 重塑。据 Valuates 预测，2027 年全球电信 AI 市场规模将增长到 150 亿美元，近三年年均复合增长率 42.6%。生成式 AI 主要通过两个方面助力运营商行业：

1、AI 应用与运营商现有业务结合，实现业务效率提升

利用人工智能的分析、策略优化与预测等能力来赋能网元、网络等业务系统，有助于提升电信网络的智能规建、运维、管控能力，并最终实现 L4/L5 级网络自动驾驶。如：韩国 KT 的 AI 语音机器人具备实时自动总结等功能，将客户请求的时间从 20 秒减少到了 5 秒。中国移动反诈骗系统月度拦截电话量超过 1400 万，准确率高达 98%。

2、对外赋能产学研用，推动智能升级

一方面，运营商可以直接为大模型企业或教育研究机构提供智算服务，做 AI 淘金时代“卖铲人”。另一方面，运营商可以将大模型能力外溢至行业客户，面向政务、教育、医疗等推出行业大模型新应用。如：中国移动九天政务大模型为甘肃打造智慧政务助手，构建 20 万实体和 1000 万业务关联的政务知识图谱以及 100 万级标准问答，为省内 2500 万的百姓提供便捷、高效的数智政务服务。

1.2.2 盘活海量数据，助力高效训练，使能大模型行业落地

运营商要抓住大模型的发展机遇，需要构建 AI-Ready 的基础设施，AI-Ready 的前提是 Data-Ready。与此同时，AI 集群规模不断扩大，现在已经已迈入万卡时代，大投入能否带来显著收益，将面临两大具体挑战：

挑战一：如何盘活运营商数据资产，更好地让大模型应用服务自身业务？

在数字化、智能化的趋势下，数据已经成为继土地、劳动力、资本、技术之后的“第五大生产要素”，是驱动数字经济深化发展的核心动力。特别是随着生成式 AI 大爆发，AI 大模型赋予了数据新的生命力，数据蕴含的价值进一步涌现，没有充足、优质的数据，大模型的学习能力将大打折扣。以中国移动计划在 2025 年实现全网 L4 高阶自治为例，需要汇聚 600 万 + 个 4G/5G 基站、9.9 亿用户和全国“4+N+31+X”数据中心等各类数据，当前核心数据规模已达 650PB，每日还会新产生至少 5PB 数据。只有将这些分散在不同省份、不同用户、不同应用的高价值数据有效组织起来，为大模型注入源源不断的数据“燃料”，才能实现 L4 级网络自动驾驶要求的智能基站节能、智能天线权值优化、投诉智能管理、网络费用稽核等能力，并给出科学的“规、建、维、优、营”的策略建议。

挑战二：如何降低 AI 开发和运营成本、拓展政企 AI 边缘应用，加速实现运营商 AI 商业闭环？

AI 集群是成本和能耗的吞金兽，如 GPT-3 单次训练的电力消耗相当于 500 吨二氧化碳排放当量，相当于 300 个家庭一年的用电量，而 Sora 的单次训练消耗是 GPT3 的 1000 倍。AI 集群可用度低造成了算力建设成本高、电力空耗等问题，推高了建设和运营成本。运营商需要考虑从“堆算力”到“挖潜力”，科学规划智算底座，比如：合理配置存储集群性能，选择高性能、高可靠的外置存储，提升 AI 集群可用度。

此外，生成式 AI 的商业正循环很重要的场景在边缘应用，尤其在 ToB 政企市场有大量 AI 应用市场前景，如医疗自助问诊、制造工业质检、金融智能客服、政务办事助手等，这些场景迫切需要“私域知识库 + 训练 / 推理 GPU+ 检索增强生成 RAG+ 场景化大模型”这样的一体化方案，运营商需要考虑采用一站式的训 / 推超融合一体机快速推出产品，实现大模型的商业兑现，打通大模型应用落地“最后一公里”。如中国移动九天超融合信创一体机，为行业用户提供了开箱即用的大模型服务，搭载 139 亿参数语言大模型以及 10 亿级参数视觉大模型，实现设备检查、皮带堆煤、皮带异物、煤量识别、人员违章等功能，助力某矿山客户井下安全管控和生产。

1.3 政务

在政务领域，人们正在探索通过人工智能应用于出入境管理、税收监管、政务问答等公共服务领域，提高公共服务组织的管理效能与风险分析、以及改善与服务对象的互动。

人工智能嵌入公共服务治理也面临着实时数据待共享、历史数据待激活、敏感数据待保护等风险挑战。

1.3.1 服务到治理：优化公共办事服务效率，增强公共业务治理能力

牛津智库 Oxford Insights 发布的 2023 年政府人工智慧完备指数（Government Artificial Intelligence Readiness Index 2023），报告对全球国家和地区政府对运用人工智能提供公共服务的准备程度做出评估，涵盖愿景、治理与道德、数字能力等 10 个维度 42 个指标。其中，数据是政务领域人工智能演进的关键推动因素，最常见的是语言类数据总量是图片类数据总量的 8 倍，而当前数据还主要用于客服系统、审批系统、分析决策辅助，并且高收入国家和低收入国家之间在数据收集、数据应用、数据安全方面的差距尤为明显，这反映了全球数字鸿沟的存在。美国在政务领域的人工智能应用的得分排名第一，其次为新加坡和英国，中国排名第十六。世界各国纷纷抢抓人工智能发展的重大机遇，并积极应对人工智能部署于公共服务中所遇到的政策、社会、经济、技术等问题，强调推行国家级的战略计划将会为社会各界带来变革的契机。

图 4：政府 AI 就绪指数的支柱

以出入境管理、税收监管、政务问答等公共服务领域为例，AI 正在深化这些场景，服务大众：

1、出入境管理

人工智能能够快速处理和分析大量出入境数据，实现自动化身份验证、智能风险评估和实时数据分析，预测移民趋势，优化资源配置，简化审查流程。例如，AI 可以通过生物识别技术快速核实旅客身份，减少人工审核的时间和错误率。同时，AI 还可以分析大量出入境数据，预测潜在的安全威胁，帮助管理部门提前采取措施。这种智能化的管理方式不仅提高了工作效率，还增强了安全性和用户体验。

2、税务系统

可以升税务管理的效率和准确性。通过 AI 技术，税务部门可以实现自动化数据处理、智能化数据分析和风险评估。例如，AI 可以通过自然语言处理技术自动解析税务文件，提取关键信息，减少人工审核的时间和错误率。实际上，AI 还可以分析历史税务数据，并结合其它类型数据，识别潜在的税务风险，帮助税务人员提前采取措施。例如，通过比较房地产公司的交易数据和实际税务申报数据，并结合建筑行业的标准成本数据（水泥、钢筋等基础材料），快速的评估出税收漏报的可能性。

3、政务问答

各业务部门对于政策的传播、规则的遵从，以及具体案例的咨询，都存在着大量的问询工作。通过自然语言处理（NLP）和机器学习技术，AI 可以快速理解并回答市民的各种政务问题，提供 24 小时不间断的服务。例如，AI 问答机器人可以在政府网站、微信公众号和 APP 等多个渠道上运行，随时为市民提供政策解读、办事指南和常见问题解答。这种智能化的问答系统不仅减少了人工客服的工作量，还提高了信息获取的便捷性和准确性。

1.3.2 共建跨部门数据流动，保护敏感数据，助力政通人和

人工智能在公共服务治理中的应用，虽然能够显著提升效率和服务质量，但也面临着诸多风险和挑战，例如：实时数据的共享需要确保数据的准确性和及时性，同时避免数据孤岛的形成。其次，历史数据的激活和利用需要克服数据格式不统一、数据量庞大等问题。最为关键的是，敏感数据的保护必须得到高度重视，采用加密技术和权限管理措施，确保数据在传输和存储过程中的安全性。

1、实时数据待共享

以中国社会信用体系建设为例，通过数据共享和信息交换，促进社会诚信建设以及对政府各部门、企业、个人等各种主体的信用评价和监管。针对企业，AI 系统可以实时查看企业的经营状况，税务记录，环保检测情况等信息，及时发现异常行为并发出预警；针对个人，个性化还款计划：AI

可以根据你的财务状况和还款能力，制定个性化的还款计划，帮助你更有效地管理债务，避免逾期还款，增加信用等级，这些 AI 应用不仅提高了社会信用体系的效率和准确性，还促进了信用体系的透明度和公正性。因此数据共享和信息交换对于 AI 非常重要，而数据共享和信息交换的基础是数据可视、可管、可用，这对数据存储提出了高要求。数据存储需要具备高效的数据管理能力：**可视**——数据资产的拥有者和管理者，需要对所有的数据有全貌概览，了解有哪些数据、数据的保存地点以及数据量、数据类型等，相当于维护了一份数据地图。**可管**——在确定了需要进行归集的数据后，需要有一个机制，来实现基于策略的数据流动。**可用**——这意味着原始数据需要被预处理、被转换为 AI 可识别和直接使用的数据。

2、历史数据待激活

全球多国政务机构持续探索基于历史数据提升服务能力，以税务为代表的部委正在积极激活税收历史数据并应用于 AI，以显著提升税务管理和决策的智能化水平。**辅助政策制定**——不同地区的经济发展水平和税收基础不同，AI 可以分析同一政策在不同地区的效果，帮助政府制定更具针对性的区域税收政策。**评估政策效果**——通过分析过去 5 年的税收数据，AI 可以评估某一税收政策实施前后的税收收入变化。某一减税政策是否真正促进了经济增长，增加了税收收入，还是导致了税收流失。**预测政策结果**——假设政府颁发一种税收优惠政策，通过对相关历史数据的分析，AI 可以预测未来几年内该税收优惠政策给特定行业带来的投资影响和发展变化。AI 对于历史数据需求，超乎了我们的想象，这对数据存储的读取速度提出了极高的要求。为了满足 AI 模型的快速训练和实时推理，存储系统必须具备超高的读取速度，以便迅速访问和处理海量数据。这不仅要求硬件层面的高性能存储设备，如 NVMe SSD，还需要优化的数据管理和缓存策略，以确保数据能够以最快的速度被读取和利用。

3、敏感数据待保护

公共服务领域涉及到大量的关键敏感数据，例如出入境管理涉及到敏感数据包括个人身份信息（如姓名、出生日期、护照号码）、生物特征数据（如指纹、虹膜数据）、旅行记录（如出入境时间、地点、航班信息）、签证信息等。AI 技术虽然提升了数据处理和分析的效率，但也带来了数据泄露和滥用的潜在风险，特别是在跨境数据传输过程中，敏感信息可能会被不法分子利用。因此建立公共数据管理制度和技术手段必不可少，而作为数据安全的最后一道防线，数据存储起着至关重要的作用。**数据加密**——所有存储的数据必须进行加密处理，以防止未经授权的访问和数据泄露。**访问控制**——严格控制对数据的访问权限，确保只有经过授权的人员才能访问敏感数据。**数据备份**——定期进行数据备份，确保在数据丢失或损坏时能够及时恢复。**日志记录**——记录所有对数据的访问和操作日志，以便在发生安全事件时进行追踪和审计。**数据隔离**——将敏感数据与其他数据隔离存储，减少数据泄露的风险。

1.4 制造

AI 在智能制造领域提升生产效率和产品质量，应用于 CAD 设计、需求预测、智能排产、预测性维护和决策支持。与此同时，数据收集与分析中的数据量激增、历史数据汇聚、数据清理和数据标签等挑战依然存在。

1.4.1 局部到全程：覆盖设计、生产、经营、售后，助力端到端增效

随着科技的飞速发展，人工智能技术在制造业中的应用已经从基础的售后机器人扩展到整个生产流程的各个环节，极大地提升了生产效率和产品质量。

1、AI 辅助进行 CAD 设计

对于大多数制造企业而言，计算机辅助设计（CAD）技术被大范围用于产品的设计阶段，包括外观设计、零部件设计、结构件设计、机械零件设计、模具设计等等。只有在设计阶段制图、建模、仿真越精准，后续在生产阶段才能更快速投产并快速出货。在 AI 时代以前，CAD 设计只能依赖有经验的员工进行产品设计，而后进行评审和检验，耗时耗力且有可能出现错误。AI 的到来带来了质的变化，在设计阶段可以通过 AI 来辅助自动生成 CAD 系统设计方案，也可以根据历史最佳实践快速形成新的设计，甚至支持多阶段并行设计，减少设计周期。

图 5：AI 辅助进行 CAD 设计

2、AI 支持需求预测与智能排产

对于大部分制造企业而言，在一年中有销售高峰和冷淡期，而销售的潮汐关联着采购、生产、仓储、供应等多个部门的工作。以往只能通过销售预测来进行排单，预测的准确性直接影响着整个产线。进入 AI 时代后，通过分析销售历史数据、供应链状态和市场价格等因素，AI 可以预测产品在一年中不同阶段的需求量，从而制定合理的生产计划，确保资源的最优配置，进而优化库存水平、降低生产和物流成本，减少生产延误和物料浪费。某大型半导体显示屏制造企业，分析了历史生产数据，并采集分析了整个制造过程中的设备数据、环境数据、产品数据，进而运用 AI 技术对整个制造过程进行智能化改造，实现了制造过程的自动化和智能化，其生产效率和产品质量得到了显著提升，同时还降低了生产成本，保持了半导体显示技术在业界的领先地位。

3、AI 在生产过程中做预测性维护

生产过程中的设备不可避免的会出现故障甚至停机，动辄小时级的维修严重影响产品的生产进度，尤其在交单高峰期的停机甚至会影响到公司信誉。以往只能通过有经验的老师傅多班 24 小时巡检保障，费人费时费力还没法完全避免设备故障。被动响应到主动维护一直是设备运维的进阶，AI 可以通过实时监测设备运行状态，预测潜在的故障并提前维护，这样可以减少设备停机时间，显著降低维修成本；同时也可以利用机器算法优化生产工艺，调整产品生产参数，并利用 AI 系统自动识别产品缺陷，并大幅提高产品检测速度和准确性。

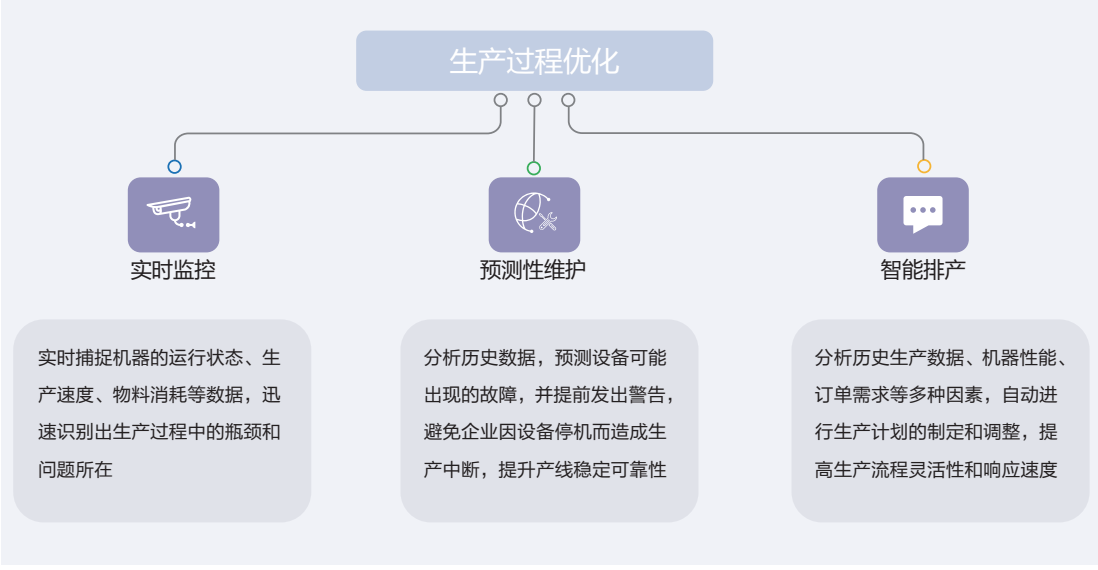


图 6：AI 在生产过程中的优化

某跨国大型生产 PLC（可编程逻辑控制器）的数字化工厂，通过整合和改造数据基础设施，与 PLM、MES、ERP 等数字化系统和平台无缝集成，并广泛应用 IoT 技术收集各类传感器数据 5000 万条 / 天，约 1TB/ 天并存储。而后使用多种 AI 技术，包括实时数据分析、机器视觉系统等对其中约数百 GB 数据进行分析，包括生产过程监控、产品质量检测、设备主动维护等，不仅提高

了生产效率和产品质量，还实现了生产过程的透明化和可追溯性，该工厂的产品上市时间缩短了近20%，生产效率提高了 13%，并且产品质量也得到了显著提升。

4、AI 在经营管理中支持决策

制造企业先一步发布有竞争力的产品，大概率可以快速赢得市场，甚至在一定程度上影响着市场的走向。怎样通过市场分析和经营管理进行精准的决策一直是企业高层思考的问题？以往只能通过大规模的用户访谈、多年的市场经验、多部门集体研讨来进行决策。虽然部分决策也赢得了市场，但是缺乏数据支持和详尽的决策流程，难以固化为标准决策机制。进入 AI 时代，通过大数据分析 ERP 系统，关联产品的设计、生产、测试、采购、仓储、供应、销售各个流程的数据，同时结合产线、人力、市场趋势、消费水平等多方面情况，为公司高层提供经营决策分析，有理有据，全过程数据分析链条完整，并且可以根据各个流程的变化，快速分析决策，缩短决策过程的时间，而且决策流程可以被固化。

5、AI 支持售后 7*24 咨询服务

智能聊天机器人在各个领域均有应用，在制造企业也不例外，通过智能聊天机器人提供 7*24 客户咨询服务在制造行业已经被成熟应用。AI 时代的来临，使得回应的内容更加准确、专业、及时，能够快速响应客户需求，解决客户问题，提高客户满意度。

1.4.2 唤醒历史沉睡数据，增强全流程生产效率

AI 在智能制造领域的应用不仅是技术的革命，也是推动制造业全面数字化转型的核心动力。在智能制造的端到端流程中，可以发现，无论是在经营管理阶段所做的决策分析、设计阶段的辅助设计、生产阶段需求预测和智能排产以及设备维护和产品检测、售后阶段的智能机器人服务响应，均需要大量的数据支持 AI 在对应阶段的作用。而 AI 在使用数据进行分析面临的挑战主要表现在以下几个方面：

1、数据收集与分析过程中的挑战

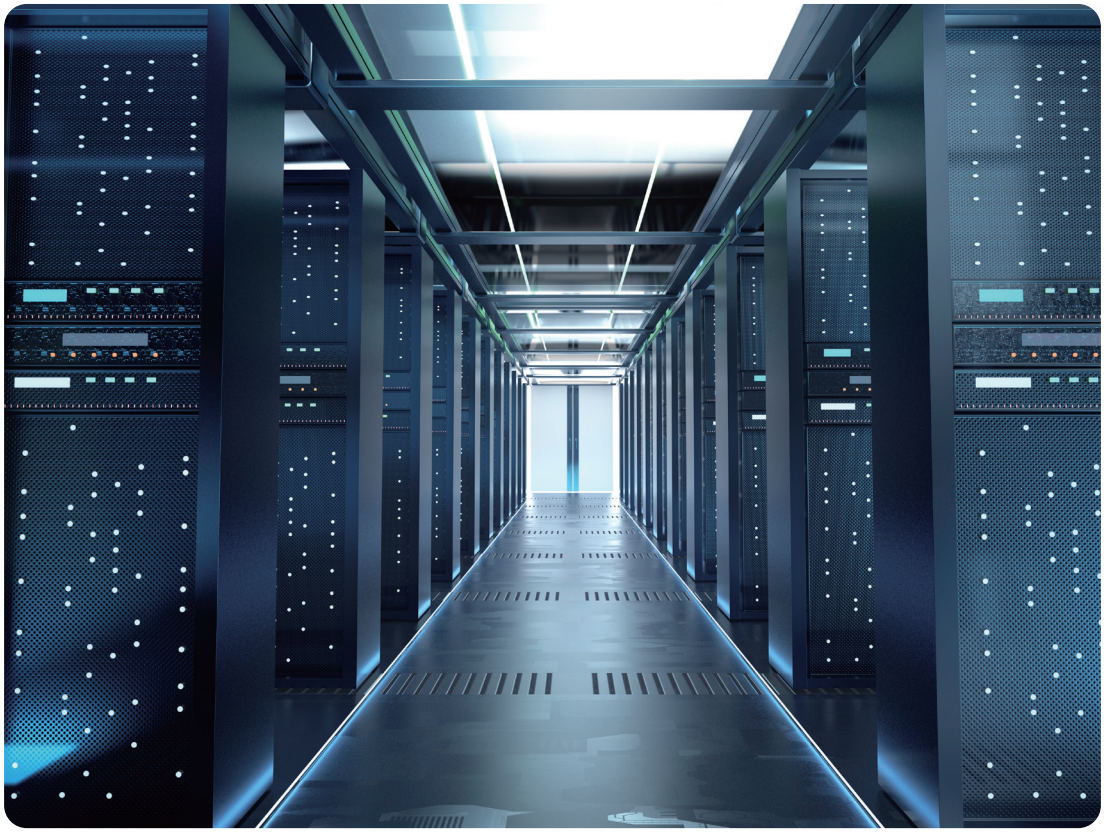
诸如在产品的生产测试过程中，实时数据收集与分析，是产品良率和设备长时间正常运行的保证。企业借助各类传感器和物联网设备，能够实时收集制造设备的运行数据，包括温度、速度、压力等关键参数。这些数据经过实时分析，可以即时反馈生产状态，确保生产过程的稳定性和可靠性。除了制造设备的传感器实时数据以外，还需要长时间、高频率的收集产品的质检数据，包括产品质检过程中的图片、音频等。如果要 AI 分析的更加精准，收集这些数据的周期也会从以往的小时级收集到如今分钟 / 秒级收集递进。伴随着收集的数据种类、数据类型、数据格式、收集时间的变化，收集的数据量呈指数级增长，每天采集到的数据从 MB 到 GB 级甚至 TB 级增长。某全球工程机械的领先企业通过 56 万 + 台的物联网设备一天收集的数据量就从原来的 GB 级到当前的 10TB 级。收集数据的增多不仅仅是因为 AI 可处理的数据量增多，从原来的 MB 到如今的 GB 级，那么收集的数据量也需要数量级的增长；同时该企业也认识到 AI 给制造行业带来的巨大机会，更多的数据意味着在今后的变局中拥有更多的“有效资产”，具备更牢靠的市场地位和强大的话语权。收集的数据量增加了 1000 倍以上，如何保障这些实时数据能够快速存的下、用的好，就成为制造企业不得不考虑的问题。

对于一些大型企业而言，海量的历史数据如何被激活价值，而不是机房里冷冰冰的机器，也是需要考虑的问题。如何进行降本增效一直是摆在制药行业面前的主要难题。一家大型制药企业也面临同样的挑战，怎样在不增加过多质检人力、不增加产线设备的情况下提升产品良率，进而增加盈利？多次探寻均无果的情况下他们盯上了已有的历史记录数据。通过对分布于多个地域的多种类型的历史数据进行多源汇聚的整合，并对生产过程中的生产工艺和设备运行状态进行 AI 分析，识别出了 9 个关键生产工艺参数，通过 AI 模拟实验对这些参数进行优化，最终将药品产率提升了 50%，良率也提升了 3%，公司因此每年在单个药物品种就增收 500−1000 万美元。诸如此类，通过深度学习算法，AI 可以从海量历史数据中识别出模式和异常，为生产、测试、仿真、决策提供科学依据。沉睡的历史数据被唤醒，并再次得以分析使用。怎样快速、简单、高效的汇聚多源的历史数据且不影响现有生产系统，汇聚后的数据高效地供 AI 系统调用就成为制造企业需要关注的问题。

2、数据归类与整理过程中的挑战

a) 数据清理：要确保收集的数据能够被 AI 使用和产生价值，务必要进行数据清理，包括数据补充缺失值、清洗数据集格式、纠正数据的物理和逻辑错误等。某电子制造集团有限公司在使用 AI 进行智能排产时发现，通过未经过清理的数据推理的结果总会出现偏差，有时候甚至不能出现结果。通过建立单独 AI 工业数据空间，接入多个工业软件系统，对数据进行汇聚、处理和交叉验证，保障数据和行为可信、可证，同时纠正数据的逻辑错误，输出正确的数据格式，进而才使用这些数据进行 AI 分析和使用，提升了排产效率。为了在数据清理时更加简单，清理的时候无需做太多的无用功，这就要求数据在采集到写入的过程中安全可靠，避免无意义的数据丢失和逻辑错误。怎样保障收集写入的数据安全可靠，不出现逻辑错误，或者在出现逻辑错误的时候能够自动修复，就成为制造企业必须考虑的问题。当然，除了收集中的问题，也需要考虑数据清理的时候可能会对原始数据造成的损坏和污染问题，这也是需要未雨绸缪的地方。

b) 数据标签：经过清理的数据只有打上数据标签，才能够帮助 AI 在训练时清晰理解数据的上下文，从而做出准确预测，并且相关的数据在使用过程中也会有不同的标签，尤其是生产数据、设备数据、经营数据、运维数据等，在决策、设计、生产、排产、售后过程中的作用不同。不同种类、不同类型、不同容量的数据如果仅依靠人力标注，成本高、耗时长且容易出现人为错误。怎样准确、高效、快速、低成本地去标注数据标签就成为了制造企业也需要考虑的问题。



1.5 电力

电力系统作为保障国计民生和支撑经济增长的关键基础设施，持续面临电网规模扩大、负荷增 长等挑战。利用 AI 辅助发电管理、输配电网负荷预测、安全巡检和隐患识别等，可有效助力电力供应安全。

1.5.1 预测到协同：精准的电力供需预测，使能高效的发输变配协同

在新型电力系统的建设中，电力供需预测的精准度和发输变配的高效协同至关重要。通过引入人工智能技术，电力企业可以实现对负荷动态和电价变化的精准预测，从而更好地匹配供需两侧的需求。这种协同不仅提高了电力系统的整体效率，还为实现清洁低碳、安全充裕、经济高效的电力供应奠定了坚实基础。

1、发电阶段：AI 建模优化发电管理，减少停机概率，识别潜藏问题

在世界 500 强的电力公司中，90% 已经使用智能电力分析系统，通过 AI 对包括火力 / 风力发电机、太阳能板等发电设备的健康状态进行实时诊断，从而优先更换高风险零件，减少计划外的停机时间。比如土耳其电力公司 ENERJISA，通过 AI 分析即时掌握发电机组与输配电路的运作状态，降低了 35%−45% 设备停机的时间，确保发电量处于可控标准之内。

同时，电力公司还会在发电机内装入 IoT 感测器，使用 AI 分析感测器所收集的信息，实时监控发电机的马达及零件状态，提前找出潜在问题。比如通过风速和发电量建立非监督学习 (一种 AI 算法) 的异常检测模型，描绘出正常状态曲线，当发动机的实时状态偏离正常状态曲线，就会及早安排检修，识别是否有潜藏问题。

2、供电阶段：AI 分析精准预测发电量和需求量，解决可再生能源的集成问题，平衡供需

过往使用燃煤、天然气等一次性能源的发电方式，发电量较易估算。但再生能源由于影响的变量太多，以光电、风电等为代表的再生能源发电量很难预估；且在预测用电需求中，也会因气候异常和生活型态改变等影响，无法通过历史用电资料精准预测用电需求。比如澳洲能源公司 Red Energy，出现过因为用电需求预测模型精准度较低，导致备转电力容量不足，必须临时向其他电力公司高价购买电力，增加公司营运成本。通过改进 AI 的预测模型后，Red Energy 的预测准确率达到 98%，并通过完善的事前规划，以较低价格购入电力，节省超百万美金的购电费用。

3、用电阶段：AI 用户分析找出异常数据，排查窃电、篡改电表等异常数据，减少损失，确保电网稳定度

以往电力公司在侦测窃电中，只有在专家检修或更换电表时才发现异常，或者有的电力公司会使用随机挑选的方式进行抽查。这两种方式都属于被动式的人工侦测，且投入成本高、排查效率低。电力公司可通过 AI 进行用户分析，在既有业务规则、用户有无篡改电表历史行为的基础上，结合窃电行为模式、用电量和用电目的之间的关联性等分析模型，精准地判断出各个电表的窃电风险，再交由相关人员做进一步的调查，提高侦测率并省下侦测成本。比如，巴西第二大电力公司，通过这种方式不只识别集团窃电的风险，更避免每个月数十万美元的窃电损失。

1.5.2 加强多维、高频数据采集和安全留存，促进更精准电力供需预测

通过多维度、高频率的数据采集，电力系统能够实时监测和分析各个环节的运行状态。这种数据采集不仅涵盖传统的电力负荷和电压数据，还包括气象、市场需求、设备健康状态等多方面的信息。利用这些丰富的数据，通过人工智能技术，能够帮助电力企业对发电、输电、变电和配电各环节进行精准控制和优化调度。

1、AI 预测用户用电量，增加数据采集量、提高数据采集频率，以得到更精准的预测

电力行业通过 AI 分析远程收集的用户用电数据来预测未来的使用电量，以提前准备供需电量。当采集的数据量不足、采集频率过少会导致预测结果产生偏差，这使得电力行业不断提升用户侧的监控器的采集频率以满足 AI 分析模型预测的要求。比如，在 IoT 抄表场景中，最初的设计可通过按周 / 月收一次用来计费来预测下月的使用电量，后续在 AI 预测模型的训练过程中，发现间隔更短的数据可以更加精准的预测用户使用量，从而提高到数分钟一次，使得能够更高效地预测，满足供需平衡。更大数据采集量、更频繁的数据采集周期，给数据存储设备带来了更大容量和更高性能的需求。

2、AI 分析电力设备，扩大数据采集维度、增加采集参数，以提前检修并减少停机概率

在发电管理场景，电力行业通过 AI 分析发电机内 IoT 感测器收集的电气元件信息，来提前找出潜在问题，及时安排检修。在最初的设计中，会收集发电机各元器件老化程度、故障零件数量和类型等来提前准备替换的零件库。随着感测器收集的数据量增多，在 AI 训练中发现一些非强关联的数据也可以加强 AI 模型的预测准确度，从而增加了 AI 分析仪表盘的维度，如发动机的运转状况、设备健康度、产出能源量等，提高了潜藏问题的发掘能力，降低了停机损失。

3、电力安全：勒索攻击不是会不会发生，而是什么时间发生

电力作为涉及国计民生的行业，一旦遭遇勒索攻击可造成大量的业务停摆，且随着电力行业的数字化建设深入，近年来已成为黑客的首要攻击目标之一。今年 8 月份，网络安全公司 Bitdefende 公

开了 Solarman 和 Deye 太阳能管理平台中的重大安全漏洞，可影响全球 20% 的光伏发电，涉及 190 多个国家和地区的 200 多万个光伏电站。

新型的勒索攻击不仅使用 AI 模型批量生成新的病毒样本，并且潜伏周期更长，隐蔽性更高，可轻松绕过普通的病毒检测库。比如，非洲某大型电力公司，近年曾经遭受勒索攻击，并被要求支付赎金数十万美元。

AI 在电力行业可通过收集核心业务生产系统的正常行为，在数据存储设备上建立 AI 侦测分析模型，判断数据存储侧的异常行为（加密、删除等）、以及短期内的异常存储容量变化，识别潜伏期的勒索攻击，降低被攻击风险。

1.6 科研教育

教育科研行业正经历 AI 带来的深刻变革，智能化展现出巨大潜力，深刻改变了教学、研究与管理方式，同时也给教育科研的 IT 系统建设带来了诸多机遇和挑战。

1.6.1 教学到探索：个性化教学，科研加速，AI 反向赋智人类

智能化在科研教育行业已经涌现出一批新兴场景的应用，通过以 AI 大模型为代表的智能化技术与科研教育场景应用深度结合，提升教学和研究的效率与质量。

1、个性化教学 / 智能教学辅助

AI 根据学生习惯、能力水平和兴趣点提供定制化学习计划和资源，提升学习评估精准度，提供智能辅助教学；典型的应用包括个性化教学方案、智能教育辅助、多元化教学资源整合、虚拟教室、实时学情监测等，通过智能化的实时反馈不断提升教学质量和效率。

图 7：人工智能赋能课程教学各环节

2、AI 辅助科研

AI 大模型帮助研究人员快速筛选和分析大量文献，通过语义分析确定研究领域的最新趋势和关键概念。同时 AI for Science（人工智能驱动的科学研究）这一新兴科学研究手段加速发展，它使用已知科学规律进行建模，同时挖掘海量数据的规律，在计算机的强大算力的加持下，进行科学问题研究。例如医疗领域用 AI 分析海量生物医学数据，以探索新治疗方法和药物。

教育科研智能化应用一方面提升了科研教育工作的效率，另一方面通过数据的汇聚、分析和萃取，进一步促进了知识的传承和共享。例如上海交大建设的 “交我算”与“教我算”两个平台，覆盖科研和教学服务，需要对接 AI、HPC 等不同算力平台，面临着数据访问协议多样、数据访问效率低等问题。因此，这套平台需要建设统一的存储底座、提供多协议互通等技术来满足多样化的应用需求，以便打造一套高效、高智能的科研平台。

1.6.2 围绕高性能、可靠安全的数据供应，构筑反向赋智的基石

科研教育在深入发展和应用 AI 的同时，也在数据处理上面临新的要求和挑战：

1、超大规模复杂数据集实时分析

教育科研智能化的特征主要体现在数据量的庞大、数据类型的多样性以及数据更新与分析的实时性。例如，在个性化教学场景，智能捕捉、收集学生在上课过程中的表情、动作、行为信息，将视频、图像、文本等多种类型的数据进行综合分析。这使得需要保存下来的数据量和复杂度都呈现出指数级的增长，容量扩展受限、机房空间受限、功耗受限成为让数据“存不下”的关键痛点；基于这些分析对学生的未来表现做精准预测，并通过该预测，智能推送个性化学习方案及教学调整建议，这对多类型混合负载的海量数据处理实时性也提出了新的挑战，对视频等大文件处理的高带宽要求和 AI 训练、文本等小文件的高 IOPS 要求难以同时满足，导致数据“用不好”。

2、数据安全性要求高

AI 时代数据安全和隐私保护成为重要议题，特别是在涉及敏感的教育信息时数据安全就显得尤为重要。例如，科研机构因为其财力雄厚，同时其科研项目往往拥有非常宝贵的数据、一些数据涉及尖端研究相关知识产权，更容易成为黑客攻击和勒索的对象。而教育科研机构面向共享及公开访问的网络设计，实验室、办公甚至移动设备等多设备、跨人群的广泛接入更是为数据安全保护带来了巨大的挑战。

3、数据的高效汇聚与流动

教育科研领域数据存在资源彼此联系、信息交织汇集、数据来源多样、要素关系分散的特征，需要建立更完善的数据采集和管理系统，实现全局的管理和高效的流动，以确保不同来源和类型的数据能够被有效利用。随着智能化应用的增加，数据跨组织、跨地域、跨时间、跨领域的共享和协同需求将大幅增加，当前 IT 系统的孤岛化建设将极大的制约数据价值的挖掘和发挥。

为了应对这些挑战，教育科研行业必须构建更加高效、稳定和可扩展的数据基础设施，包括高效的数据存储解决方案、先进的数据分析工具以及严格的数据管理政策。例如北京大学现代农业研究院小麦抗病遗传育种团队，通过大数据与人工智能应用对植物基因组进行持续研究，以大幅提升

主栽小麦品种的韧性。但对作物基因组的研究分析过程极其复杂，海量数据的处理和读写带来了巨大的挑战。首先，作物基因组研究中涉及到大量的基因组测序、表达谱测定、SNP 分析等数据产生，需要充足容量、巨大吞吐量的数据底座支撑；其次，由于基因测序的整个过程会有持续化的碎片文件读写，绝不允许被中断，这就要求支撑测序应用的存储系统具备极致的稳定性和可靠性，确保数据不会丢失或损坏；其三，在冷冻电镜和基因数据分析工作中，对存储系统的整体性能、小文件处理能力提出更高要求。因此，针对海量的作物基因簇数据的存放、无中断访问、高性能访问，成为摆在该团队面前首先要解决的问题。

1.7 医疗

作为知识密集型行业的代表，医疗行业相对更加容易获益于生成式 AI。人工智能正在为医疗行业注入新的活力：辅助诊断，药物研发，疾病预警等。与此同时，如何共享和汇集数据、并保护病患隐私和医疗数据安全，成为医疗行业拥抱 AI 所必须面临的挑战。

1.7.1 诊疗到预防：辅助提升诊疗效率，加速康复减少疾病

随着 AI 技术的不断发展，AI 与医疗行业的结合越来越深入，其在医疗领域的应用也越来越广泛，AI 给医疗行业带来的变化更加显著，从辅助诊疗、药物研发到疾病预警等多个应用场景，AI 都发挥着重要作用。未来，AI 在医疗领域的发展趋势将深刻影响医疗行业的格局和患者的就医体验。

图 8：“医疗 +AI”应用场景

1、辅助诊疗

AI 技术在基层卫生健康服务中的应用试点启动实施，形成了可复制使用的医学人工智能基层辅助诊疗应用系统。这些系统通过智能分诊、AI 辅助诊疗等方式，帮助医生提升诊疗水平，赋能基层诊疗。例如，一款 AI 智能分割及规划算法的设备适用于脑出血抽吸引流、颅内活检等临床场景，通过 AI 找到斑块位置，精准定位到脑出血点，协助医生完成手术，提高了手术安全性和精准性。

2、药物研发

传统的药物创新研发遵循“倒摩尔定律”，AI 技术通过数据和算法模型建立的优势，正在为药

物研发带来革命性的变革。通过深度学习模型，可以更快速地分析分子结构，从而加速新药的发现并减少昂贵的实验需求。例如，有研究利用 AI 成功地识别出了一种能够对抗抗生素耐药细菌的药物，在短短 21 天内被发现，并在 46 天内完成了实验验证，这比传统的药物研发过程快了数年，大大缩短了药物研发的时间和成本。

3、疾病预警

AI 与大数据模型的应用使得疾病预警有了“工具”。通过分析国际卫生部门各种疾病方面的资讯和信息，帮助疾控等部门更准确地预测疾病的发展趋势和高发期，从而提前采取相应的防控措施。例如，AI 可以“收集”眼科医生无法识别的细微信息，通过大数据模型分析某种疾病患者视网膜变化，最终完成具有明确标记的疾病检测任务。

1.7.2 打通诊疗数据共享，保护数据安全，维护病患隐私

随着 AI 技术在医疗领域的广泛应用，在医疗行业的数据上也面临着数据收集难、数据隐私和安全、被勒索病毒攻击等诸多挑战。

1、数据收集难

AI 以数据为食，获得的数据越多、质量越好，其越能在任务中表现出色。收集的数据必须来自可靠的来源，从不可靠的来源收集数据可能会对 AI 训练的输出产生不利影响。因此，为了获得准确的输出，医院必须从可靠的来源收集训练数据，如从患者的历史和当前病历中找到可靠的数据。

2、数据隐私和安全

医学领域涉及大量敏感数据，如患者的身份信息、健康状况、疾病诊疗情况、生物基因信息等，不仅涉及患者隐私，还具有特殊的敏感性和重要价值，一旦泄露，可能给患者带来身心困扰和财产损失，甚至对社会稳定和国家安全造成负面影响，因此医疗领域的数据安全非常重要。

然而，医疗 AI 的研发与应用，必须依赖大量的医疗数据用于算法训练，数据量越大、越多样，其分析和预测的结果将越精准。但数据收集、分析处理、云端存储和信息共享等大数据技术的应用，加大了数据泄露的风险。

3、被勒索病毒攻击

AI 技术的发展使得勒索软件可以更精准地选择目标、定制攻击，并且更具欺骗性。通过分析目标的数据和行为模式，勒索软件可以更有效地选择目标，并制定更有针对性的攻击策略。此外，AI 可以使得勒索软件在攻击过程中更具自适应性，能够根据受害者的反应来调整攻击方式，增加攻击成功的几率。《2023 年中国企业勒索病毒攻击态势分析报告》显示，医疗行业已经成为勒索病毒攻击的重灾区。自 2018 年以来，全球已发生 500 次公开确认的针对医疗组织的勒索软件攻击，导致近 1.3 万个独

立设施瘫痪，并影响到近 4900 万份病患记录，这些攻击仅由停机造成的经济损失就已超过 920 亿美元。据第三方统计数据显示，医疗行业连续 12 年成为数据泄露成本最高的行业，2022 年医疗机构该数据高达 1010 万美元，与 2020 年相比激增 42%。

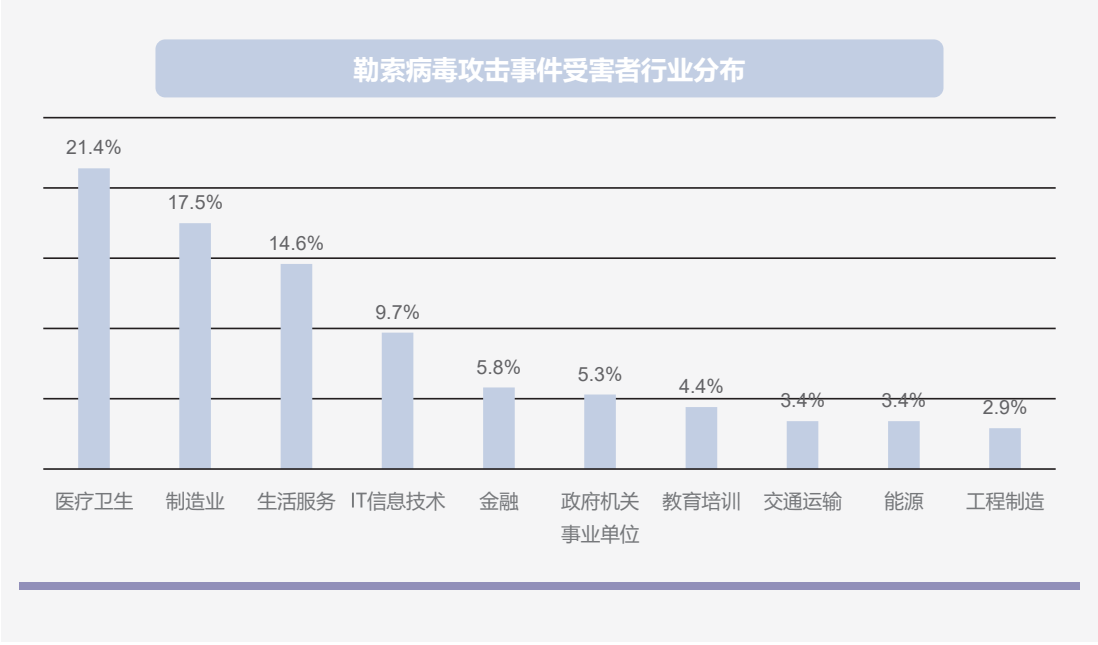


图 9：2023 年中国企业勒索病毒攻击态势分析报告

为了有效解决面临的诸多数据层面的挑战，医疗行业亟需采用专业数据存储产品，通过专业的存储内生安全、容灾备份、安全可信数据流动、防勒索保护技术等，让数据存的下、存的放心、用的安心，助力 AI 加速医疗领域迈向智能世界。

1.8 行业数智化：数据是关键

今天，包括金融、运营商、政务、制造、电力等在内的多个行业，数字化和智能化都在不断改变这些行业的面貌。

数字化将人类社会生产和日常生活中所产生的信息转变为数字格式的数据，极大地提高了信息记录、处理和传播效率。智能化，通过 AI 算力基于数字化所产生的数据进行训练和推理，最大程度地释放数据价值。

数字化为智能化提供必需的数据；智能化通过释放数据价值以牵引更多业务场景积极拥抱和扩大数字化；更多的数字化，又产生了更多的数据。可以看到，数字化和智能化相互依赖，又相互促进，逐渐融合成为数智化。数智化是数字化被赋智后的自然延伸，它通过学习数据以产生智能，并将智能应用于数字化，进而推动各行各业数字化向更高效、更智能的方向发展。随着技术的不断进步，数智化将继续深化其在各个领域中的应用，推动社会向智能世界迈进。

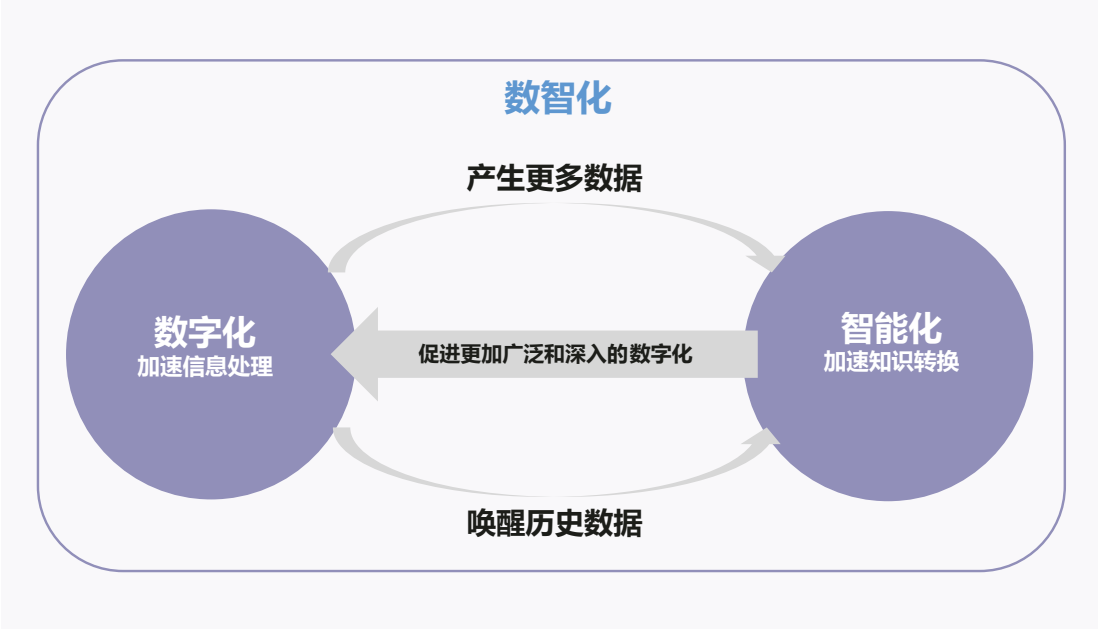


图 10：以数据为纽带，加速数字化和智能化融合成为数智化

数字化和智能化分别使用通用算力和智能算力对数据进行处理、分析、价值创造。数据则是连接数字化和智能化的纽带，是两者融合成为数智化的基石，是数字化到数智化成功转型的关键。

02

数据为纲：行业数智化呼唤高质量数据和高效数据处理

数据的规模和质量决定了 AI 智能的高度。 以 GPT 为例：GPT−1 采用了 4.8GB 原始数据进行训练；GPT−2 采用了 40GB 经过人类过滤后的数据进行训练；GPT−3 采用了 570GB 数据进行训练，而这 570GB 数据来自对 45TB 原始数据的过滤；ChatGPT/GPT−4 在 GPT−3 训练数据基础上，加入了高质量的标注。从 GPT−1 到 GPT−4，模型架构相似，而模型参数规模、数据集规模和质量不同，产生了不同的 AI 大模型训练结果。GPT 的演进，用事实证明了许多 AI 学者的观点：AI 以数据为中心。

千行万业在数智化的过程中，不管是对基础大模型的二次训练和监督微调，还是在应用推理阶段，均离不开大规模高质量数据。在实践中，大多数企业通过唤醒历史数据、采集并保存更多生产数据、人工合成数据这几种方式相互配合，为 AI 算力提供数据。

在数据规模和质量满足企业数智化所需的同时，数据效率同样不可被忽略。数据效率从数据保存、访问、能耗和安全等维度，帮助企业用户对数据进行更多维度的审视，让数据不仅供得上，还要供的快、供的稳。

2.1 数据觉醒：充分发挥历史数据价值

缺数据，不 AI。数据短缺成为制约大模型发展的瓶颈。

当前，大模型正在赋能千行百业，但也面临着海量、优质的行业数据集严重匮乏的挑战。行业数据包含领域特有的知识、术语、规则、流程和逻辑，这使得其往往难以在通用数据集中充分覆盖。与此同时，行业数据具有稀缺性的特点，据智源研究院统计，当前已知的所有开源行业文本类数据集仅有约 1.2TB，远远无法满足千行万业的模型需求。

数据在人工智能（AI）领域中扮演着至关重要的角色，在训练模型阶段，AI 模型需要大量的数据来进行训练。这些数据用于学习模式、预测结果和优化性能。没有足够的数据，模型的准确性和效果将受到限制。

1、数据驱动决策：AI 系统的决策基于数据。从金融预测到医疗诊断，数据支持着 AI 系统的智能决策。

2、迭代改进：数据允许 AI 系统不断迭代和改进。通过分析用户反馈、监控性能指标和更新数据，AI 可以不断优化自身。

3、个性化体验：数据使得 AI 能够为每个用户提供个性化的体验。例如，推荐算法根据用户的历史行为和偏好来推送内容。

数据觉醒是从数字化时代向智能化时代演进的必经之路：

1、激活业务闲置数据

业务运转过程中，产生大量的数据。一部分数据是热数据，被频繁访问，随时可能被修改。另外一部分数据，则随着时间的推理，热度逐渐降低，虽然依然保存在主存储中，但是几乎不太可能被再次访问，例如大量的医疗影像数据，在病人痊愈后，相关影像数据可能就不再被访问，进入闲置状态，直到主存储容量被占满后，再被转移到其他存储上。

随着人工智能大模型规模不断扩大，对训练数据的需求呈指数级增长。将业务闲置的数据纳入训练数据资源，可有效帮助大模型训练。

2、唤醒历史归档数据

如企业档案、历史记录、文献资料等，正逐渐被挖掘和利用。这些数据蕴含着宝贵的历史信息，且数据量巨大，可以用于模型训练、趋势洞察、业务预测。

a) 更丰富的训练数据：历史数据包含了过去的经验、事件和知识。通过激活这些数据，我们可以获得更丰富、更多样化的训练样本，用于训练机器学习模型。这有助于提高模型的准确性和泛化能力。

b) 趋势洞察分析：历史数据可以用于预测未来趋势。通过分析过去的数据，我们可以发现模式、周期性和趋势，从而预测未来可能发生的事件。这对于业务决策和规划至关重要。

c) 异常检测和故障预测：历史数据中的异常情况和故障信息可以帮助我们构建异常检测模型。这些模型可以用于实时监测，及早发现潜在问题，从而避免损失。

为了确保人工智能的持续进化，必须投资于高质量训练数据的收集和管理：

维基百科当前的内容规模约为 4.2 亿个单词。根据 ARK Invest 的“Big Ideas 2023”报告，到 2030 年，模型训练需要具有惊人的 162 万亿个单词。人工智能模型规模和复杂性的增加无疑将导致对高质量训练数据的更大需求。

在计算规模不断下降的世界中，数据将成为人工智能发展的主要制约因素。随着人工智能模型变得更加复杂，对多样化、准确和庞大数据集的需求将继续增长。在管理各种历史数据、唤醒历史数据的过程中，需要关注：

1、数据来源多样化

从各种来源收集数据有助于确保人工智能模型在多样化且具有代表性的样本上进行训练，从而减少偏差并提高其整体性能。对数据基础设施的要求需要能够存储大规模的数据集，包括多样化的来源；快速读写和检索数据，以满足训练模型的需求；保护数据免受未经授权的访问和损坏以及确保数据的持久性和可靠性。

2、确保数据质量

训练数据的质量对于人工智能模型的准确性和有效性至关重要。应优先考虑数据清理、注释和验证，以确保最高质量的数据集。此外，加载了数据标注，数据清洗等技术的数据基础设施，可以帮助最大化可用训练数据的价值。同样对于数据基础设施的要求是，具备大容量以存储高质量的数据集，而最重要的是可以实现近存计算，构建在存储侧构建数据清洗、标注、验证的能力。

3、解决数据隐私问题

随着对训练数据的需求不断增长，解决隐私问题并确保数据收集和处理遵循道德准则并遵守数据保护法规至关重要。实施隐私计算等技术可以帮助保护个人隐私，同时仍然能为人工智能训练提供有用的数据。

2.2 数据生成与合成：让数据为数智化而生

将海量历史数据唤醒，利用这些历史数据进行 AI 大模型训练和推理，有效帮助了 AI 大模型高速发展。在 AI 发展过程中，人们逐渐意识这些海量历史数据虽然对 AI 起到了不可替代的作用，但并非为 AI 而生，例如在数据采集频度、数据格式、数据多样性、数据留存等维度，均存在可以改善的空间。

举例，某工厂的汽油储罐需要进行泄露监控和检测，历史上通过摄像头 + 人工检查的方式开展；伴随着机器视觉大模型的成熟，人们可以利用 AI 对储罐上的油斑进行实时分析，以提前发现泄露隐患；但是，原有监控系统仅仅保留最近 30 到 90 天的监控数据，缺少历史上泄露隐患暴露前的油斑视频，也就让 AI 训练缺少了相应的数据；另外，除了缺少历史视频数据外，老摄像头的清晰度可能仅满足人类肉眼预判隐患，但是 AI 可以分析清晰度更高的视频，进而更加精确地预判隐患。

可见，在 AI 的驱动下，人们不仅要思考如何利用好已有的历史数据，还应该思考如何在既有数字化业务中改进数据生成，通过提升数据规模和质量来加速数字化到数智化的转型。

另外，除了在现实业务中生成更多高质量数据外，对于某些难以通过在实践中获取的数据，也可以考虑数据合成的方式。

2.2.1 数据生成

一般来说，可以参考 5F 方法来思考如何生成并留存更多的高质量数据供 AI 使用。

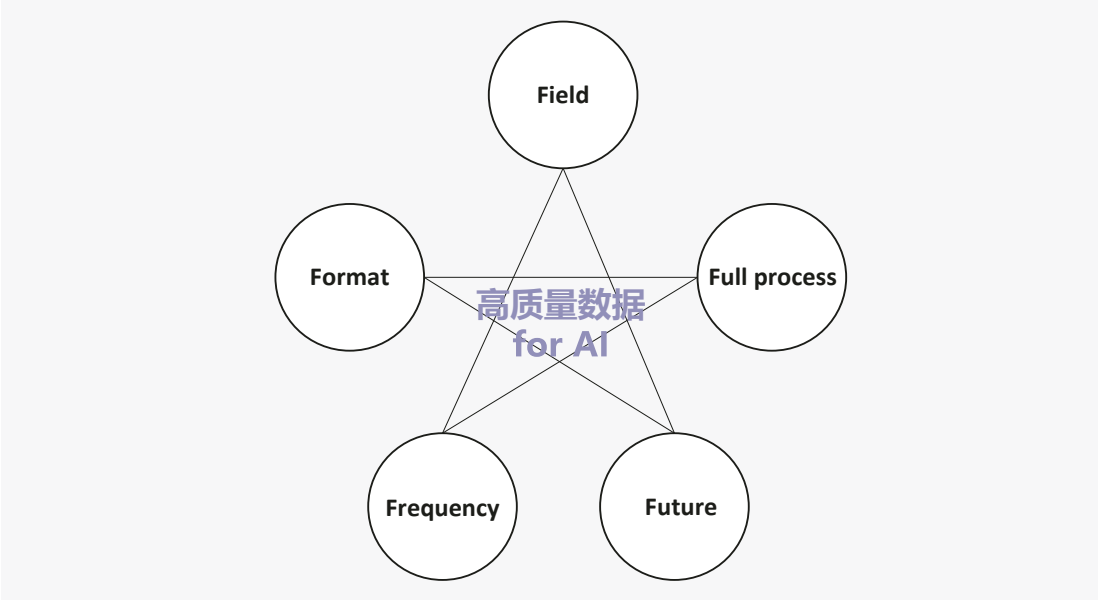


图 11：5F 维度思考高质量数据生成和采集

5F 方法是一个思考框架，帮助行业用户从五个维度去思考如何生成、采集、留存更多高质量数据供 AI 使用。这五个维度分别是： Field（数据生成 / 采集的现场）、Format（数据生成 / 采集的格式）、Full process（业务全流程数据）、Frequency（数据生成 / 采集的频率）、Future（面向未来的数据留存周期）。

1、Field，数据生成 / 采集的现场

数据产生于不同的区域和位置。有的来自偏远户外，例如油气勘探、远洋科考；有的来自户内，例如智能电表、智能家居；有的来自于终端设备，例如手机网购、办公便携等。

在 AI 大模型爆发之前，人们大多在这些不同的地点，仅采集和记录当前可以被处理的数据。以智能电表为例，最开始仅用于取代人力抄表，实现 AMR（Automatic Meter Reading），后来追加 AMI（Advanced Metering Infrastructure），以对用电情况进行实时分析，支撑输电、配电的高效运作。实际上，现在已经有部分电力公司在思考，利用智能电表搜集更多的环境数据，如温度、湿度、气压、噪音等现场相关信息，借助 AI 进行更加精确的分析预测并提升能源使用效率。例如，某个区域 A 的平均气温在 30 度左右，但是湿度高达 90%，而另外一个区域 B 的平均气温在 35 度左右，而湿度低于 5%。虽然这些数据对于供电不直接相关，但是电力公司可以基于这些信息做出预判：区域 A 住户开启空调的概率高于区域 B，进而对不同区域的供电做出提前部署。

2、Format，数据生成 / 采集的格式

因为需求、预算等不同的原因，人们在数字化过程中，会选择适合当时业务的解决方案，而 AI 在过去大概率没有被作为一个考量因素。今天，伴随着 AI 逐渐走进千行万业，AI 在数字化建设时必须要被作为一个考量因素，而数据格式是 AI 考量要素的一个重要维度。

数据格式，泛指信息以什么样的方式被数字化。例如，一段音频，WAV、FLAC、MP3 等就是不同的格式；一张图片，JPG、GIF、PNG 等就是不同的格式。除了这里提到的编解码格式外，清晰度、分辨率等也属于数据格式的范畴。

针对数据格式的考量，需要考虑是否可以满足当前及可见未来的 AI 所需。

3、Full process，业务全流程数据

现在的 AI 训练，主要还是在学习结果，尤其是正确的结果。而人类实际上的学习过程，不仅仅是通过学习正确的结果来获取知识，同样也会通过学习错误的结果、学习计算 / 推导过程来获取知识，例如代码编程、图纸设计等。

AI 能力在持续提升，也在探索对错误结果数据、计算 / 推导过程数据的学习。而实际上，我们现在对于错误结果数据、计算 / 推导过程数据的保存，是不完善的，并没有将这些数据纳入到我们的数字化进程中。在 AI 时代，伴随着上下文窗口的持续增大，相信针对这种类型数据的学习，将会是帮助 AI 实现进一步跃升的关键。

4、Frequency，数据生成 / 采集的频率

在数字化时代，人们在生产过程中，对数据的采集和留存，整体上有两种方式：

a. 产生多少数据，就记录多少数据，并匹配相应的计算和网络资源，对这些数据进行处理。典型的场景是金融行业，如在线交易。

b. 对产生的数据进行周期性采样，并对采样数据进行保存。采样的周期，一般取决于业务系统当前的处理能力和精度。典型的场景是科学研究，如气象监测站、科学育种等。

伴随着 AI 带来的超大规模算力，针对上述第二种场景，现在的情况已经逐渐转变为“数据饥饿”，即 AI 算力等待更多的高质量数据输入。人们应该思考如何适度超前地提高数据收集频率并将这些现实世界中产生的高价值数据有效保存下来，为 AI 提供更多的数据燃料。

5、Future，面向未来的数据留存周期

在 AI 大模型时代之前，数据被长期留存，主要目的是作为存档以备后续查阅。现在，除了满足法规遵从要求的最短留存时间外，数据需要被留存的时长，需要充分考虑 AI 的发展，提前对数据的留存周期进行考察。即便现在用不到这么多的数据，也需要为未来做提前思考，做到适度提前。

例如某国海关对人员出入境记录要求保存 5 年时间，可以供人们按需查询近年的出入境记录等。但伴随着 AI 大模型持续成熟，5 年的出入境记录数据可能会逐渐无法满足 AI 大模型训练所需，该国海关正在考虑将这些数据保留更长的时间，延长到 10 年甚至 20 年。

2.2.2 数据合成

数据合成是一种通过计算机算法或模拟生成人工数据的方式，它模仿真实世界数据的统计特性和特征，但并不包含、或仅包含一部分现实世界的真实数据。通过数据合成得到的数据，被称为合成数据，可以用于多种目的，包括数据增强、数据隐私保护、以及在数据稀缺的情况下进行模型训练和测试。

合成数据的生成方法主要有四种：

1、基于统计分布数据合成

该数据合成方法首先分析真实数据集的统计分布，找出其统计学上的分布规律，例如正态分布等；然后，数据专家基于识别出来的分布规律，从零开始或者基于特定的初始数据集进行数据合成，进而创建出在统计学上与原始数据集相似的数据集。

2、基于机器学习数据合成

这种方式与第一种方式本质相同，唯一的不同是依靠训练机器学习模型以理解真实数据集的统计分布，并基于模型来生成数据，获得与真实数据具有相同统计分布的合成数据。

3、基于生成式 AI 数据合成

通过对生成式 AI 进行提问或者提示，引导生成式 AI 产生文字、图片、音频、视频等数据。基于生成式 AI 的数据合成，与基于机器学习的数据合成有一些相似，一般来说前者用于生成较为通用的文字、图片等数据，而后者多用于科学科研场景所需要的数值。

4、基于随机算法数据合成

在消除原始数据的敏感数据、保护个人隐私方面，随机算法可以很好的生成姓名、家庭住址等敏感信息，帮助对原始数据集进行脱敏。

合成数据的优势包括无限量生成数据的能力、隐私保护、减少偏差以及提高数据质量。它允许组织在不违反隐私法规的情况下使用数据，同时提供了一种经济高效的方式来获取更多数据。

然而，合成数据也有其局限性，例如合成数据可能无法完全捕捉真实数据的复杂性和多样性，且其生成过程可能需要高水平的专业知识和技术。

只要正确认识合成数据、合理利用合成数据，那么合成数据是对在现实世界中获取的真实原始数据的有益补充，可以解决数据稀缺和隐私保护等关键挑战，从而在 AI 研究和应用开发中发挥巨大的作用。

2.3 数据效率：以高效数据访问使能高效数据处理，加速行业数智化

数智化时代，AI 新质生产力成为社会发展的核心驱动力，数据成为最重要的生产资料，企业的数据量将呈现指数级的增加和留存，同时 AI 生产需要完成海量的数据归集、千亿级文件数据的加载，以及断点续训所需要的 Checkpoint 保存和加载，AI 生产力能否充分释放和发挥价值，数据效率将是关键。过去，主要是通过数据存储的性能、容量及可靠性三个维度的优化来不断提升数据效率；面向未来，数智时代的还需增加数据范式、绿色节能与数据编织三个新的维度提升，才能充分提升数据效率、释放数据生产力。

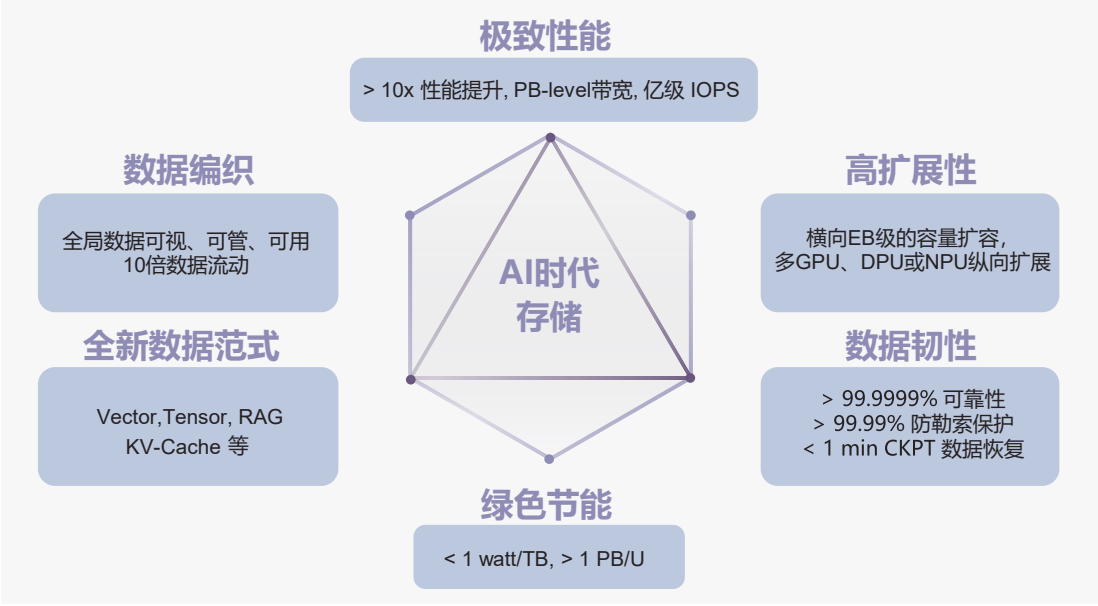


图 12：AI 时代存储

1、极致性能

数据源分散、归集困难，数据频繁搬迁，PB 级数据预处理往往需要数天；在训练中，大量的小文件加载慢、使得 GPU 等待时间长，同时 Checkpoint 恢复时间长导致 GPU 利用率低。因此为了提升 AI 训练集群的利用率和训练效率，减少算力等待时间，需要存储具有更高的性能，未来需要超越传统存储 10 倍的性能，支持 PB 级带宽以及亿级 IOPS，能够让海量数据的加载、Checkpoint 的写入更加迅速，同时还需要支持多数据协议的以减少数据的拷贝，从而才能极大提升生成式 AI 的全流程效率。

2、高扩展性

AI 时代，数据成为企业日趋重要的价值资产，数据的留存率越来越高。当前企业数据平均留存已从之前的数月上升到数十年，同时数据增速将会呈现指数级的提升，数据的爆发式增长对存储容

量提出更高的要求，未来存储集群需要能够支持 EB 级容量的横向扩展，同时每个引擎需要支持多 GPU、DPU 或 NPU 纵向扩展，以支持近存计算。

3、数据韧性

随着数据价值的不断提升，数据韧性及安全变得越来越重要，需要系列措施来保护数据的完整性和可用性。一方面体现在生产过程中的可靠性，数据不丢失、业务零中断，通过架构和技术创新，专业的存储设备在架构、节点冗余设计等方面的可靠性，能够构筑多级的安全可靠机制，实现 99.9999% 高可用性，不用再担心因数据的问题影响 AI 业务运行；同时针对日益增加的数据勒索风险，需要结合全面的动态检测、主动防御、联动恢复机制，变静态管理为动态检测，变被动响应为联防联动，打造立体化的防勒索解决方案，形成一个完整的防御体系，确保数据的安全。

4、数据编织

数据资产能够被高效利用的前提，是能否实现数据资产的可视、可管、可用。数据编织，就是能够实现可以随时随地使用任何数据。具体来讲，是通过数据资产的统一视图，实现跨域、跨站点、跨厂家等复杂数据的全局可视、实时更新；其次是能够实现数据目录的智能化，通过 AI 和自动化技术实现数据的自动标签、聚合、检索、呈现，推进数据按内容、合规、热度等维度的全自动化分类分级，并能够根据数据的热温冷分析实现数据的自动流动，实现数据高效经济的储存；同时能够面对海量的文件，实现千亿级文件秒级检索的能力，实现数据的高效查找。

5、全新数据范式

通过近存计算实现近数据预处理，让数据在存储完成部分过滤、归一、转码与增强的数据准备任务，减少数据搬移，从而提升 GPU 利用率。同时，把企业最新垂直化的数据进行向量化存储和检索，大幅度降低企业接入和使用 AI 大模型的难度。使能多维“张量”格式的数据，通过智能检索引擎，具备快速的张量数据检索能力；通过内嵌知识库，利用 RAG 技术消除 AI 大模型幻觉。另外，基于多层 KV-Cache 实现长记忆内存型存储，通过以查代算、推理过程数据在推理集群内共享等方式，减少推理算力的压力，进而有效支持千万 Token 长度的无损超长序列，全面提升推理效率和精度。

6、绿色节能

节能减排是社会持续发展的基础。到 2026 年，全球数据中心的耗电量预计将达到 2022 年的 2.3 倍，相当于日本一个国家全年的耗电量，其中数据中心一半以上的电力消耗都将被 AI 占据。存储作为 AI 数据的载体，随着数据量的增加，需要更加绿色节能的数据存储方案，单位数据存储的能耗优化改进是产业发展的必然需求。通过存储介质应用创新和整机硬件创新，实现小于 1Watt/TB 的存储能效和 1PB/U 的存储密度；同时结合介质创新的大容量 SSD 盘，在节约空间、节省能耗的同时，可以实现相同空间 10 倍容量的提升，从而助力数据中心能耗的降低。

03

数智化时代数据基础设施展望

3.1 基于存算分离架构的AI-Ready 数据基础设施

以存算分离架构部署 AI-Ready 数据基础设施，加速智能涌现

AI 大模型走向多模态，算力集群规模和数据规模持续扩大、系统管理复杂度日渐增长的同时，数据存力逐渐成为 AI 持续高速增长的关键。

存算分离架构的灵活性和独立扩展，可有效简化智算集群管理、方便计算和存储分别按需扩展；在此架构下，灵活横向扩展、性能线性增长、多协议互通等能力成为数据基础设施基本要求。

3.1.1 趋势

1、AI 训练走向多模态，数据规模持续增长、类型日趋复杂

伴随着 AI 大模型从 NLP 走向多模态，数据快速膨胀，带来了数据量的爆炸和数据处理复杂度的大幅提升。比如过去在 NLP 处理时，参数量规模通常在千亿级左右，训练数据都是简单的数字、文本、图像、音频等；而到了多模态大模型时代，参数量规模已经达到了万亿到十万亿级左右，训练数据追加视频、3D、4D 等等，每条训练数据有几十 GB。数据访问方式，数据归集方式，数据组织形式都发生的根本性的变化。

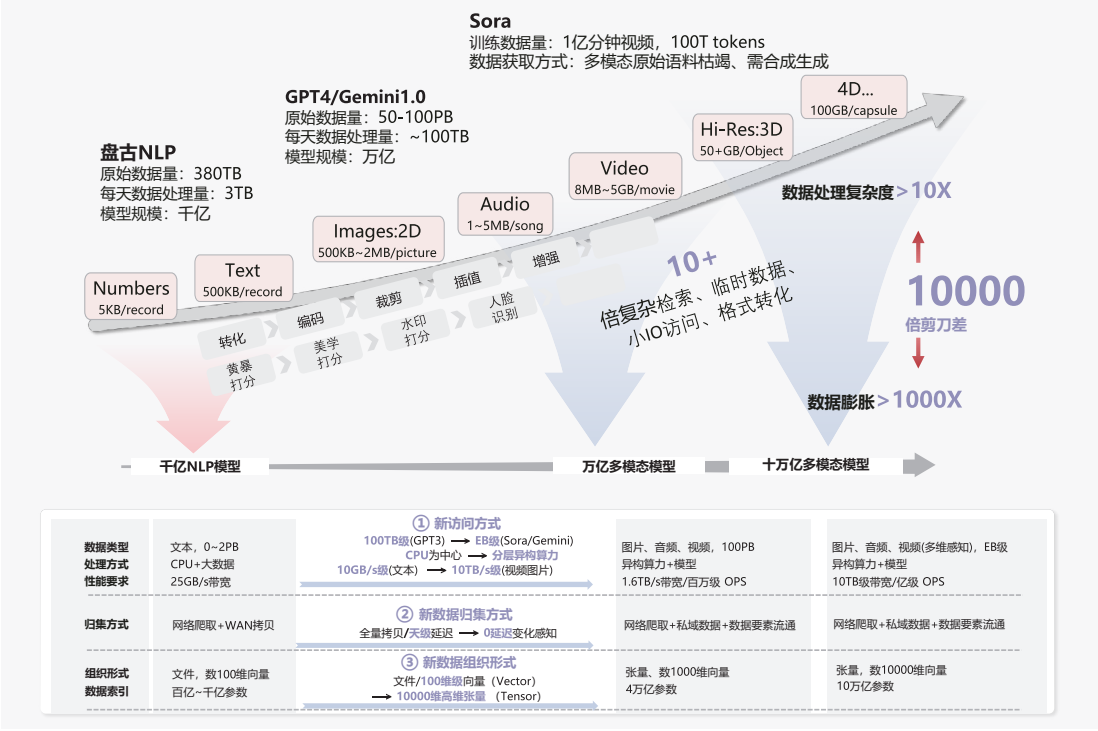


图 13：训练走向多模态，数据量越来越大、类型越来越复杂

2、伴随 AI 算力集群规模越来越大，算力利用率持续降低

AI 大模型的训练和推理过程，主要分为四个阶段：数据获取，数据预处理，模型训练，模型推理。

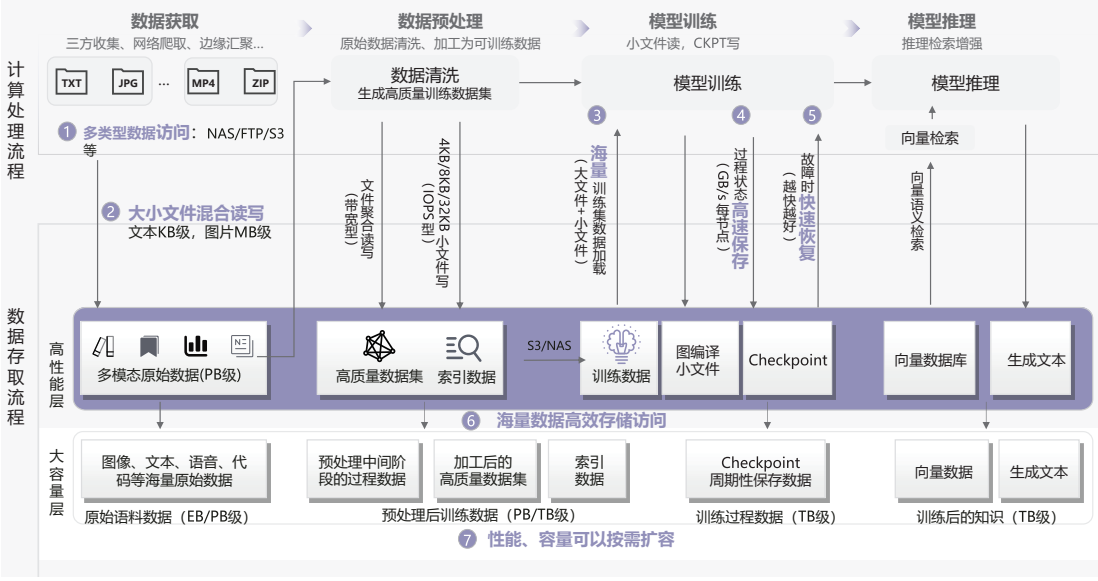


图 14：AI 全流程与计算处理、数据存取的关系

阶段一：数据获取，将不同数据源的数据导入到存储中（通常采用数据湖），通过 Spark 等分析软件进行数据收集、过滤、聚类和索引，用于以后的分析和处理。通常，这个阶段需要 EB/PB 级的原始语料数据，通过 NAS、S3 等不同的协议进行访问，涉及到 KB 级大小的文件、MB 级大小的图片等，是一个混合 IO 读写模型。

阶段二：数据预处理，经过清洗之后的数据，通过数据预处理软件，进行特征提取，特性建模，并进行向量化，我们称之为“特征库”。

阶段三：模型训练，通过 AI 训练集群进行轮训（Epoch），并在每个 epoch 期间调整权重和偏置以优化模型质量，最终输出能够解决某类问题的“模型数据库”。这个阶段，每次训练前，需要将海量的训练数据集加载到 GPU 内存中，过程中需要周期性地将 TB 大小的 Checkpoint 文件保存到存储中，故障时又需要从存储中快速地加载 Checkpoint 进行恢复。特别强调的是，这个过程对存储的性能要求极高，而且是越快越好。Meta 的 Llama 3 大模型进行训练的过程中，Meta 动用了 1.6 万块 GPU 集群，该训练过程中遭遇了 419 次意外组件故障导致的训练中断，平均每 3 小时发生一次，频繁的故障严重影响了 AI 模型的训练效率和稳定性。集群的业务中断时间为：

$$=(1-\left(1-\frac{\frac{\text{备份间隔CKPT}}{2}+\text{故障恢复时间MTTR}}{\text{平均无故障时间MTBF}}\right))*365*24$$

那么年均集群中断时间为：

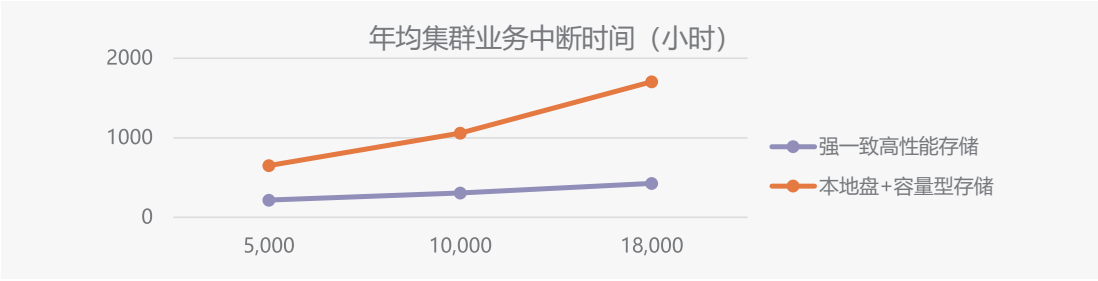


图 15：年均集群业务中断时间

那么在每次故障后如何快速读取数据、尽快重新恢复训练就显得尤为重要。

以 Checkpoint 的读写为例：每个 GPU 训练过程中会同步写一个 Checkpoint 分片，所有 GPU 产生的 Checkpoint 最终拼装成一个完整的 Checkpoint。任何一个分片错误都将造成这个周期的训练无效。

如下图所示，每个训练节点在 T0 时刻产生 N 个分片，组合成一个 T0 时刻完整的 Checkpoint 0。如果这些 Checkpoint 分片保存在服务器本地盘中，那么所有节点会通过异步的方式同步至外置存储中。

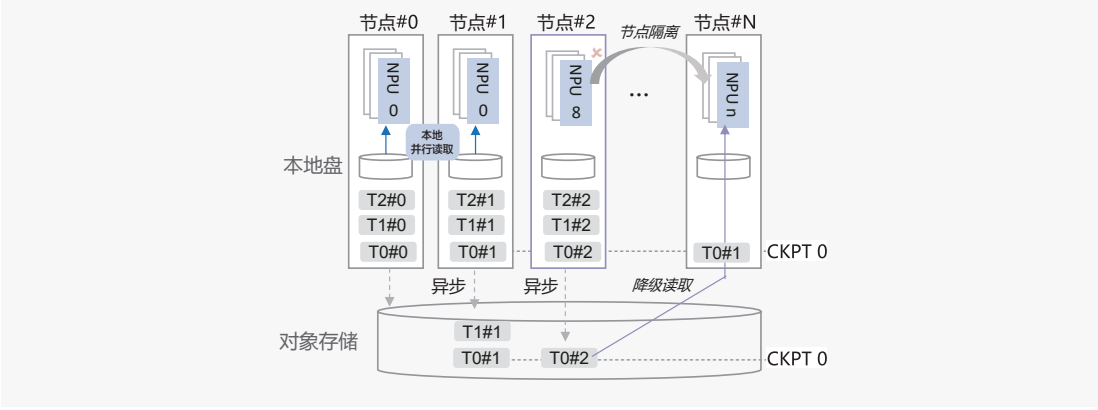


图 16：节点故障导致训练无效

如果节点 2 发生故障，此时训练任务首先会删除该故障节点，切换至新的节点 N，但由于服务器件本地盘无法共享数据，所以只能从外置存储进行加载。

由于是异步同步机制，只能加载到数个周期以前的 Checkpoint 分片，造成这几个周期的训练任务无效。另外，外置的对象存储往往性能很差，加载时间很长，在这个加载过程中，整个训练任务处于等待状态，1 个节点拖慢整个集群的恢复效率。

阶段四：模型推理，用户输入查询问题时，为了提升大模型推理的准确性，避免其出现幻觉，企业一般都会利用私域的知识对大模型进行微调，并通过检索增强生成（RAG）技术提升回答问题的准备性。

3、幻觉普遍存在于 AI 推理过程中

导致 AI 幻觉的原因是多方面的：

a）通用大模型的数据质量不高，规模不够大。如果使用不准确或者错误的数据进行训练，大模型就会产生 AI 幻觉。大模型训练所使用的数据可能包含错误信息，这些信息可能来源于数据收集过程中的错误、数据处理阶段的问题，或者是历史数据遗留问题。不准确的数据会直接影响模型的判断和预测能力，导致模型输出不可靠的结果。如果训练数据在不同群体、类别或场景中存在偏见，那么这种不公平会在模型的推理结果中被放大，进一步影响模型的公正性和普适性。例如，如果一个用于对象识别的模型主要是用浅色对象的数据训练的，它在深色对象上的识别效果可能会显著下降。随着时间推移，某些数据可能会失去现实意义，如果继续使用这些过时数据训练模型，会导致模型无法适应最新的应用场景和需求变化。当模型训练的数据规模不够大时，模型的泛化能力会受到限制，即模型对未见过的数据和新场景的适应性会较差。这通常表现为模型在训练集上表现优异，但在实际应用或测试集上性能明显下降。大规模数据集合应涵盖丰富的场景和多样性，以确保模型具备广泛的知识理解和处理能力。若数据规模虽大但多样性不足，同样会限制模型的应用范围和性能表现。

b）通用大模型运用于行业中进行二次训练和微调时，行业数据不够多，数据质量不高，规模也不够大。当行业数据量有限时，通用大模型在进行二次训练时，模型容易在少量的训练数据上过度拟合，导致其在新的、未见过的数据上表现不佳，这种情况在机器学习和深度学习中十分常见，特别是在复杂的模型结构中。另外，小规模的数据集可能不足以涵盖行业中所有重要的场景和情况，这会导致模型的训练不具备足够的代表性，从而在实际应用中出现预测偏差。特定行业的数据分布可能存在明显的长尾效应，即大部分数据集中在少数类别，而其他类别数据稀少，这会造成模型在常见类别上表现良好，而在少数类别上表现较差。如果行业数据质量不高，包含噪声或者错误信息，也可能是标注不一致，甚至是关键信息缺失，这都将影响模型的判断准确度，进而影响最终的应用效果。

c）推理缺少行业共识或者基础知识，缺少行业实时信息，时效性不够。大模型如果缺乏对行业共识和基础知识的理解，其推理过程可能无法深入到行业实际问题的核心，导致分析结果停留在表面。在行业决策过程中，模型由于缺乏必要的行业背景知识，可能无法提供有效的决策支持，影响决策的准确性和可靠性。行业特有的模式和规律需要大量专业知识支撑才能识别和学习，缺少这些知识的模型难以准确把握行业特性。另一方面，行业实时信息是模型预测未来趋势的重要依据，对于模型的时效性至关重要，如金融市场的价格变动、供应链管理的库存动态等，缺乏实时信息将导致模型输出过时，无法及时响应行业变化。

3.1.2 建议

1、采用存算分离架构，分别部署智能算力和存力，各自按需演进

在 AI 大模型的部署中，将算力和存力分开部署的存算分离架构显得尤为重要。这种架构不仅能够有效地提升资源利用效率，还能为模型训练和推理提供强大的支持。存算分离使得计算和存储资源可以独立进行横向或纵向扩展，根据实际需求增减资源，避免过度投资和资源浪费。同时，在现阶段 AI 大模型发展中，改变粗放式堆算力模式，选择高性能、高可靠的专业外置存储，合理配置存储集群性能，从 AI 训练的全流程角度优化，降低训练任务中断，提升算力可用度。为了保障整个集群的负载均衡性，在需求高峰期，可以增加计算资源以处理更大的数据量，而无需担心存储瓶颈；反之，在数据密集型任务中，可以单独增强存储性能，提升整体处理速度。用户可以根据不同资源价格走势和自身业务特点，选择性价比最优资源组合，有效控制成本。

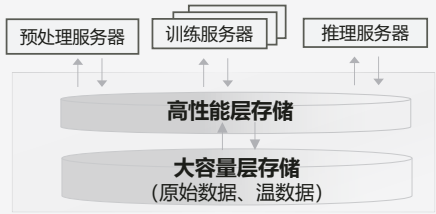


图 17：存算分离架构帮助各自按需演进

另一方面，AI 的发展也会伴随着算力、算法和数据的不断向前演进。存算分离架构允许计算资源和存储资源独立进行技术更新和升级。这意味着可以在不影响到另一方的情况下，采用最新的处理器或优化算法提升计算性能，或者采用新的存储技术提高数据读取速度。在 AI 领域，模型和算法的迭代速度非常快。存算分离架构可以快速适应这些变化。例如，当一个新的 AI 模型需要更多的计算资源时，可以迅速增加 GPU 或 TPU 节点，而无需担心存储瓶颈。由于计算和存储资源是独立的，因此更容易集成最新的技术进展，如新型神经网络架构或优化算法，只需在相应的计算或存储层面进行升级即可。存算分离架构还支持多租户环境，不同的用户可以共享计算和存储资源，同时又能保证资源之间的隔离和实验安全。数据存储独立于计算资源，可以更专注于数据的安全和备份，减少数据错误和丢失的风险。采用高性能（带宽、IOPS）、灵活扩展、可靠的专业 AI 存储，提升集群的可用度。

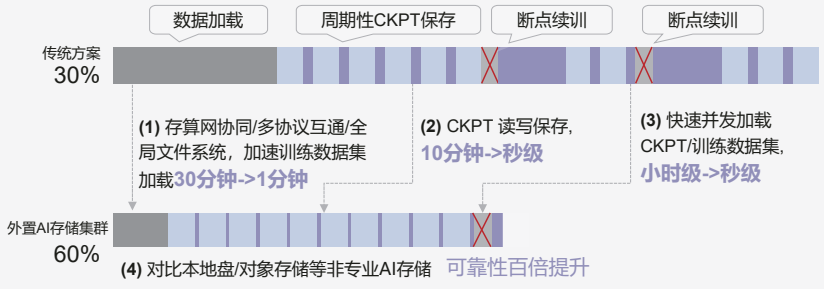


图 18：可靠的专业 AI 存储，提升集群可用度

2、数据基础设施具备横向扩展能力，性能随容量线性增长

当前的 AI 大模型已经从处理单一类型的数据（如文本）发展到处理多种类型的数据（如文本、图像、音视频等）。这种多模态甚至全模态的发展路径将使得训练数据集的规模从 TB 级别上升至 PB 乃至 EB 级别。AI 大模型的参数量也从千亿级别向万亿甚至十万亿规模迈进。这意味着所需要的计算资源和存储资源将同步增加，存储系统必须能够适应这一变化，提供足够的容量以及与之匹配的性能。存储需要支持 EB 级的容量扩展，并且在容量扩展的同时性能也要随容量线性增长。随着模型复杂性的增加，数据存取和预处理的复杂度也在上升。存储系统不仅要应对大规模数据的高速存取需求，还要支持复杂的数据处理流程，因此还需要支持 GPU、DPU、NPU 等横向扩展能力，用于 IO 处理的加速。AI 存储系统应该被设计为同时具备高性能层和大容量层，且对外呈现统一的命名空间。这种设计允许数据首次写入时根据策略放置于不同的层级，并可根据访问频度和时间等策略自动进行数据分级迁移，从而优化整体性能与容量利用率。为了应对 AI 全流程中的数据存储和访问需求，AI 存储系统需覆盖从数据获取、预处理、模型训练到模型部署的各个阶段。这不仅简化了数据流转过程，还减少了因数据迁移带来的时间和资源消耗。理想的存储架构应具备全对称式架构设计，无独立的元数据服务节点。随着存储节点数的增加，系统的总带宽和元数据访问能力能够实现线性增长，满足 AI 训练过程对高性能的需求。

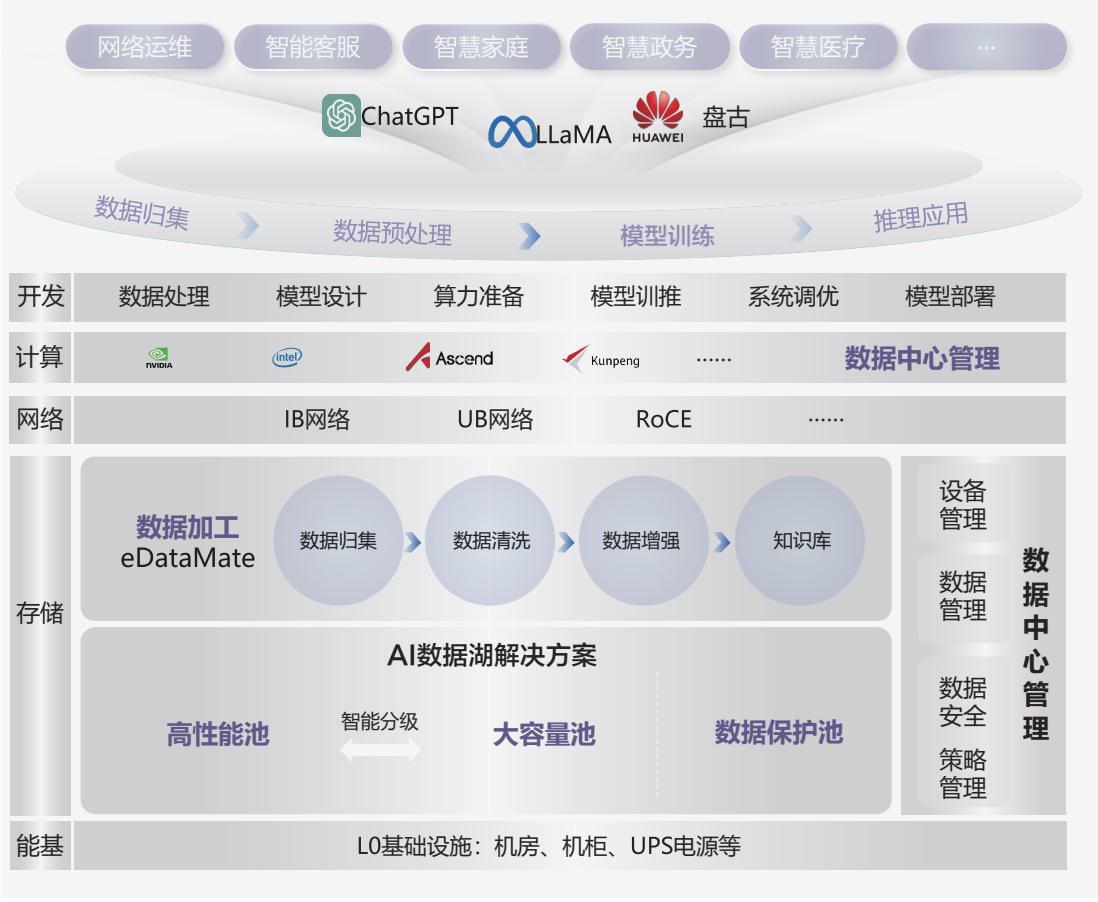


图 19：华为 AI 数据湖解决方案

3、数据基础设施支持多协议，且协议之间互通

在 AI 的数据预处理阶段，数据清洗、数据集成、数据转换和数据消减是四个关键步骤。然而，这些步骤往往需要耗费大量时间和资源。数据准备的过程不仅需要处理大量的数据，还需要确保数据的准确性和一致性。由于数据源的多样性和复杂性，处理过程中可能会遇到各种问题，如数据缺失、不一致和冗余等，这些都需要仔细处理和验证。因此，数据准备阶段通常是整个 AI 项目中最耗时的部分之一，例如 PB 级数据，预处理就会历时数月。

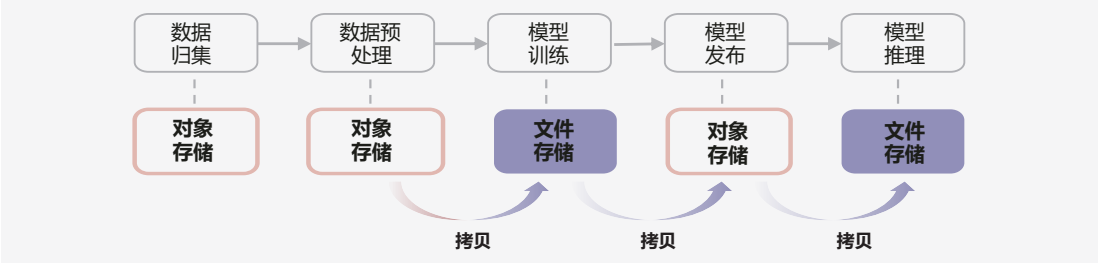


图 20：数据协议不同，数据在存储间需要多次拷贝

如图 20 所示，由于数据协议不同，数据在存储间需要多次拷贝。训练准备时涉及亿级文件拷贝，以天级到周级为单位，训练准备耗时长。

比如华为的盘古小艺语音模型训练，原始数据 2PB，根据上游业务需求，数据清洗过程膨胀为 30+PB，耗时长达几个月。AI 全流程涉及的工具链可能使用不同的协议。优秀的 AI 存储应该支持 NAS、大数据、对象等多种协议，且各协议语义无损，确保与原生协议相同的生态兼容性。另外，在 AI 的各个阶段中，数据应当能够实现 0 拷贝和 0 格式转换。通过全局文件系统和多协议互通来提升数据准备的效率，避免数据在数据中心间、设备间的拷贝。并且数据处理和 AI 训练与推理各个阶段之间无需数据拷贝，加速大数据和 AI 平台的部署与并行处理，减少等待时间和性能损失。存储系统还要支持高性能动态混合负载，需要在数据导入、预处理、模型训练等阶段同时处理大小文件的读写操作，并在这些操作中保持高性能，特别是生成 Checkpoint 时的大量写入操作，如图 21 所示。

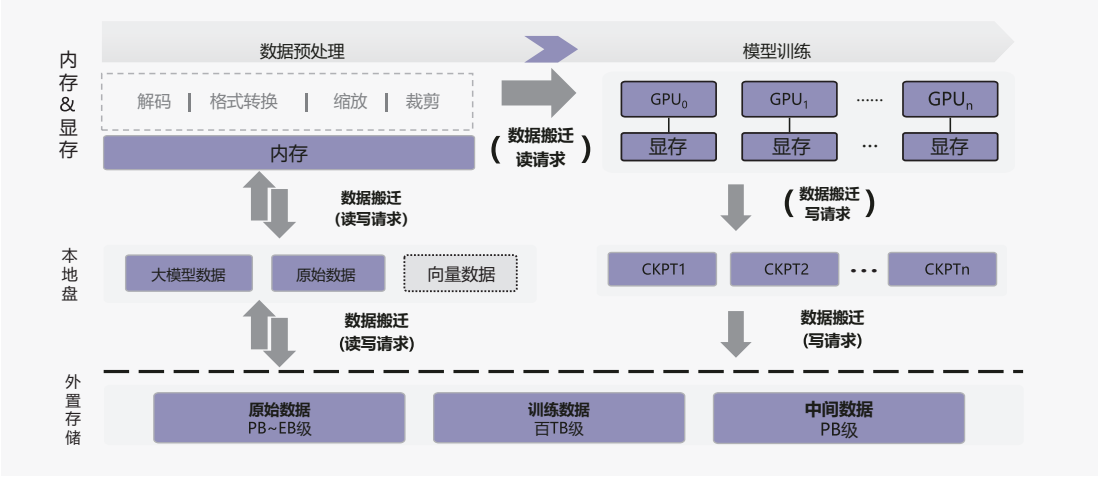


图 21：AI 的数据预处理阶段，PB 级数据搬迁影响全流程效率

3.2 全闪存助力高效数据处理

以全闪存提升数据处理效率，加速数据价值释放

伴随 AI 大模型算力集群规模不断增长，算力等待数据所产生的算力空载问题日渐突出，亟需加速数据访问效率以提升算力利用率。与此同时，智能化升级也在加速数字化转型，进而产生更多的业务数据，增加了数字化基础设施处理数据的复杂度和压力。

全闪存是数智化时代提升数据处理效率、满足业务需求的最优解，同时满足不断增长的数字化转型和日益深化的智能化变革；与此同时，配合向量 RAG、长上下文记忆存储等新兴数据范式，可以有效简化数据访问，实现以存强算，提升系统整体性能。

3.2.1 趋势

1、多源异构海量数据预处理日趋复杂，传统数据管理走向综合数据治理

当今的数字环境中，社交媒体、物联网设备、在线交易、传感器网络等丰富的数据来源持续产生数据。在数据冗余与复杂度的累积之下，为了从海量无序的数据中加速汲取“营养”，企业需要剔除无效数据和噪音数据以寻找有效特征和价值信息，这需要更强大、更智能的数据处理技术来进行数据的存储、治理与分析。

譬如，自动驾驶训练需要汇集各种稀有路况、极端天气下的设备运行状态，以覆盖未来不同场景下的行为预测与决策动作。仅 Waymo 一家公司的公开数据集就包括约 1000 辆测试车、共计 10 万英里驾驶时长的采集数据，单辆测试车每天就会生成 20TB 以上的原始数据。在日益激烈的市场竞争之下，从测试验证到规模商用的周期被不断压缩，迫使车企更加快速地从所采集的海量数据中提炼出优质算法，如迁移学习、少样本学习和自监督学习，以快速提升模型的适应性，这就需要更高的数据读取效率。

在医疗影像分析（如 CT、MRI、X-ray 等）中也是如此。一次全身 CT 扫描产生数千张图像，数据量可达 GB 级，但真正有诊断价值的决定性信息通常只占很小一部分，一个微小的肿瘤或异常组织可能只占据几张图像的极小区域，其余部分则是正常组织或无关区域，这对关键病灶识别与筛选的效率提出了更高的要求。

另外，传统的数据存储主要关注数据的存取、管理和备份恢复，依赖于关系型数据库、文件系统等，以求确保数据的持久性和可访问性。然而今天，企业对数据的态度已经逐步走向数据的综合

治理，一方面强调对数据的全生命周期管理，如整合、清洗、标注、保护、合规处理与价值挖掘。这通常需要将来自不同系统、不同平台、不同设备的数据集成到统一的数据环境中，提供全面统一的数据视图和分析能力，支持数据的跨部门协作与跨地域共享。例如，零售商通过集成来自线上渠道和线下门店的数据（如销售数据、客户反馈、库存信息），提供全渠道的综合客户视图，优化库存管理和营销策略。

2、更大规模的算力要求数据存储提供更高性能的数据访问

深度学习模型中的神经网络层数与参数量越来越多，催生了越来越高的数据维度和量级。为了训练这些模型，传统的数据处理方法已难以满足需求，传统的关系型数据库与存储主要以索引和关系模型为基础，在处理高维度数据（如嵌入向量）和复杂查询时效率显著降低，比如面对 100 万条记录的响应时间高达 1~5 秒，而专为高维数据而设计的向量数据库仅需几十毫秒。

向量数据允许从数百万个数据点中快速进行相似度计算和最近邻搜索（k-Nearest Neighbors, k-NN），这对于处理大量数据的任务（如图像检索、文本匹配）非常重要，能够大幅提升模型优化、数据处理的效率。例如，在电商营销推荐系统中，用户和商品以特征向量的格式来计算出相似度与关系，从而进行个性化推荐。

3、数据实时处理逐渐成为多种业务的基础需求

AI 技术逐步融入金融交易、自动驾驶、智能制造等行业，不仅需要传统数据分析能力来定期处理历史静态数据（如季度报表等），更需要实时处理动态的数据流，这需要系统必须能在毫秒级的时间内处理和分析数据，从而做出准确的决策，以帮助企业获得差异化优势。比如，纳斯达克股票交易所需要处理来自全球各地的市场数据，包括股票价格、交易量、订单信息等，每秒需处理数百万个订单和数据包并实时执行交易决策。流式数据处理框架的兴起，如 Apache Flink 和 Kafka Streams，要求数据存储能够更全面地融合各种实时数据格式、更快速地响应数据读写请求，让分析和训练更实时，支持 AI 系统的动态决策能力。

3.2.2 建议

1、构建以数据综合治理为目的的数据基础设施

数据存储从传统的数据管理走向数据综合治理，一方面实现多源异构海量数据的快速归集和汇聚，另一方面通过专业的数据预处理工具链，从海量数据中高效提取所需的训练数据。

综合治理一般分为三个层级。首先是设备管理层，在数据中心维度将所有数据存储设备管理起来，做到统一管理、统一运维。其次是数据管理层，借助全局文件系统，将企业分散在所有数据中心的数据都纳入到同一张数据地图，实现可视化管理和调度。最后是数据过滤层，只有将原始数据

过滤处理（也被称为预处理）后，所形成的高质量数据集才能被包括 AI 在内的多种分析平台所高效处理。

2、通过全闪存存储和语义创新为算力高效提供数据

全闪存存储可极大地缩短数据读取和写入的时间，能提供更高的 IOPS 和更低的响应时延，提升现代数据中心的性能，从而满足企业对实时数据处理和分析的极致要求，显著提高数据处理的效率。

不论是面向关系型数据库的集中式架构，还是面向海量非结构化数据的分布式架构，都可以利用闪存的高性能、大容量、低功耗，在有限空间内提供惊人的性能密度和容量密度，从而满足大规模算力对数据的高速访问，支撑大规模算力发挥出其应有的作用。

同时，创新的数据访问语义（内存语义、向量语义等）可以缩短算力和数据之间的路径，加速算力对数据的访问。

3、统一数据基础设施平台，实现数据高效流转

提供数据全生命周期的管理，从数据的生成、存储、处理到最终的归档和销毁，均能高效而可靠地进行。实现多协议融合互通，使得数据可以在不同的存储和计算环境中高效流转，无需进行繁琐的数据迁移操作。这种免迁移的数据流转方式，不仅节省了大量时间和资源，还确保了数据在传输过程中的安全性和完整性，进一步提升了数据处理的效率。

3.3 存储内生安全成为基本需求

数据存储是数据安全的起跑线，数据安全不能输在起跑线上

智能化升级过程中，一方面加速了数字化转型，产生更多高价值业务数据，另一方面降低了黑客门槛，让勒索攻击更加频繁。

不管是产生了更多数据的数字化，还是持续成长的智能化，均需要在数据基础设施层面构建防治结合的数据安全体系，基于存储内生安全，从被动应对攻击走向主动全面防护。

3.3.1 趋势

1、数据量增长而备份窗口有限，呼唤更强备份能力

ChatGPT、盘古等 AI 大模型的蓬勃发展驱动了数字化领域对于数据价值挖掘能力的需求。各行各业利用 AI 技术挖掘大量结构化和非结构化数据中的隐藏模式和知识，揭示其中的关联、趋势和规律，为大模型提供丰富的训练材料，以产生正确的决策结果。这些数据价值挖掘诉求驱动了用户收集更多维度、更高频次的数据，使得数据量呈指数级增长，数据价值也比以往更高。

面对数据的爆发式增长，数据备份迎来新的挑战。在数据短期留存场景中，在原有相同大小的备份时间窗口内，备份存储需要完成更多的高价值数据备份任务，这要求更先进的备份介质和架构，比如采用全闪化的备份介质、利用重删压缩算法备份更多数据、使用数据直通的备份一体机等。对于数据长期留存的场景，很多 AI 模型会调取历史的经验数据来进行二次训练，且由于场景不同，时常出现同一份数据多份数据拷贝的情况。这使得备份归档介质在解决数据留存期问题的基础上，不仅需要具有温冷数据自动分级的能力，还需要具备备份归档数据快速切换的能力。对此，业界厂商已经尝试使用备份归档融合的架构同时保存短长留存周期的数据，通过架构内部的自动分级，实现长期留存数据的快速恢复。

2、AI 降低勒索攻击门槛，全面数据保护势在必行

生成式 AI 出现以来，传统的安全自动化大大提升，但随之带来的是：勒索软件的变体迭代也更加频繁，网络攻击的门槛被大幅降低。有研究表明 WormGPT、FraudGPT 等工具的出现，生成式 AI 导致网络钓鱼邮件攻击增长 135%。据最新市场调研报告数据，生成式 AI 和云的广泛应用使得恶意机器人（Bad bots）暴涨，占互联网总流量的 73%。日本一位没有任何专业 IT 知识的男子，仅使用生成式 AI 的问答功能，制造出能对电脑资料加密、索要赎金的勒索病毒。

同时，生成式 AI 还可优化勒索攻击的攻击方式，使攻击内容更加难以被辨别，如借助 Bot 自动化攻击手段，让攻击者可以更快速、准确地扫描漏洞或对网络发起攻击，大幅增加网络攻击的波及面和有效性。2023 年 11 月份，中国某黑客组织借助 ChatGPT 进行病毒程序优化、漏洞扫描、渗透获取许可权、植入勒索病毒等一系列攻击手段，造成某公司服务器全部挂死，并以此对受害公司进行勒索。

3.3.2 建议

1、建设全闪存备份存储，提升备份效率

通过全闪介质实现同时间窗内更快的备份与恢复效率，利用重删压缩算法在同容量备份更多副本，恢复更多数据；通过数据直通的三合一架构（备份软件、备份服务器、备份存储）提升可靠性，避免传统服务器堆叠架构的链路闪断风险。针对长期留存和短期留存数据共存的场景，采用备份归档一体架构，将备份归档数据融合，实现数据的无损分级，备份归档数据的无缝切换。

2、建设多层防勒索，从被动走向主动，防治结合

通过存储、网络等基础设施的结合，采用多层次、端到端的有效防护，可提供抵御勒索软件的最佳防御。网络与存储多层检测及联动的数据保护，通过有效的攻击前预防、攻击时的精准检测及响应和攻击后快速恢复，使勒索攻击防护从被动响应向主动防御转变，帮助用户及时发现并拦截勒索攻击，保护数据不被非法加密和窃取，利用存储快速安全恢复数据，全方位构建防勒索安全防护体系。

3.4 AI数据湖使能数据可视可管可用

建设 AI 数据湖底座，打破数据烟囱，实现数据可视可管可用

伴随 AI 算力集群规模增长，海量多源异构数据的管理已经成为主要挑战之一。数据地图绘制、数据归集、数据预处理、海量数据分级管理和安全保护等工作，是 AI 大模型训练首当其冲的要务。

为数智化转型建设 AI 数据湖底座，基于数据编织能力打破数据烟囱，才能实现海量多源异构数据存得下、流的动、用得好。

3.4.1 趋势

1、数据逐渐成为 AI 的差异化竞争力

“缺数据，不 AI”已经成为业界共识，数据的规模和质量决定了 AI 的高度。根据《2023 Global Trends in AI Report》调研统计，构建 AI 基础设施的主要挑战中，数据资产的有序有效管理超越数据安全与计算性能，成为 TOP 1 的挑战。未来 AI 大模型的好坏，20% 由算法决定，80% 由数据决定。在 DataLearner 大模型综合排行榜中，Meta 公司的 LLaMA3 大模型依靠 70B 参数 +15 万亿 Token 数量获得 82 分，远超 LLaMA2 大模型使用 70B 参数 +2 万亿 Token 数量获得的 68.9 分。企业尤其需要关注行业数据、日常运营数据等核心数据资产的原始积累，充足的数据、高质量的数据将帮助企业显著提升 AI 训练和推理的效果。

图 22：大模型在各领域的业务场景示意图

2、数据资产管理成为企业开展 AI 实践的关键准备

数据质量是数据资产管理的核心问题之一，在整个 AI 的作业流程中，准备好高质量的数据所耗费的时间占整个 AI 作业的 80%。多数企业面临数据来源众多，数据质量参差不齐，导致很难快速准备好训练 AI 模型所需的大量数据。关键数据资产入库、进行清单化管理是企业开展 AI 实践的关键准备。

在大模型训练环节，高质量的 QA 问答对，可以显著改善 AI 大模型的模型精调效果。依赖人工生成问答对存在效率低、输出质量不稳定的问题，业界采用 Self-QA 和 Self-Instruct 技术，通过工具自动生成高质量的 QA 问答对语料。

在大模型推理环节，检索增强生成技术（RAG，Retrieval-Augmented Generation）是提升大模型推理精度的关键措施。企业需要将数据资产进行向量化后，在向量数据库中进行保存，以便在 RAG 系统中进行高效的信息检索和生成。

3、从训练走向推理，让 AI 进入千行万业已经成为业界共识

随着大模型参数规模、上下文长度等技术演进，向量检索库容从千万级走向十亿级，检索时延和精度随之恶化，索引重建需要数周时间，影响大模型推理的商业使用。同时，上下文长度决定大模型的记忆推理能力，长序列推理能够使语义更丰富，生成内容更连贯、准确，超长序列成为大模型推理的主流技术选择。但长序列也面临诸多挑战，例如推理算力成为瓶颈，推理响应缓慢等。因此，无损成为人们在实现长序列过程中的焦点。为实现无损长序列，人们一方面注意到单服务器推理模式已经很难满足业务诉求，推理走向集群化成为必然选择，另一方面模拟人脑的快慢思考方式，基于强一致性的外置独立存储，构建多层 KV-Cache 等技术，帮助推理集群具备长记忆能力，在推理集群内以查代算、过程结果共享，减少推理算力压力。大模型推理的效率和成本，成为商业正循环的核心竞争力。

4、善于应用 AI 的企业将从竞争中胜出

AI 大模型应用从知识问答、文生图、文生视频等通用应用，演变为大模型 + Copilot 辅助 +Agent 自主决策的综合应用。能够熟练评估大模型能力，掌握大模型使用和优化方法，将极大提升企业的综合竞争力。

比如金融行业，采用 AI 大模型技术可以帮助银行实现精准客户画像，提供更好的个性化推荐和定制化服务；通过人机交互打通智能客服、智能网点等流程，大幅提升终端用户体验。

图 23：PB 级非结构化数据将被激活

比如医疗行业，采用 AI 大模型可以通过预约就诊、智能分诊等改善患者院前就医体验；在就诊过程中影像辅助诊疗、辅助病理诊断、精准医疗等，减少医生工作量，提升诊断效率和诊断质；诊后，AI 通过健康管理、知识问答等功能，协助患者进行健康管理，从被动治疗转向主动预防。

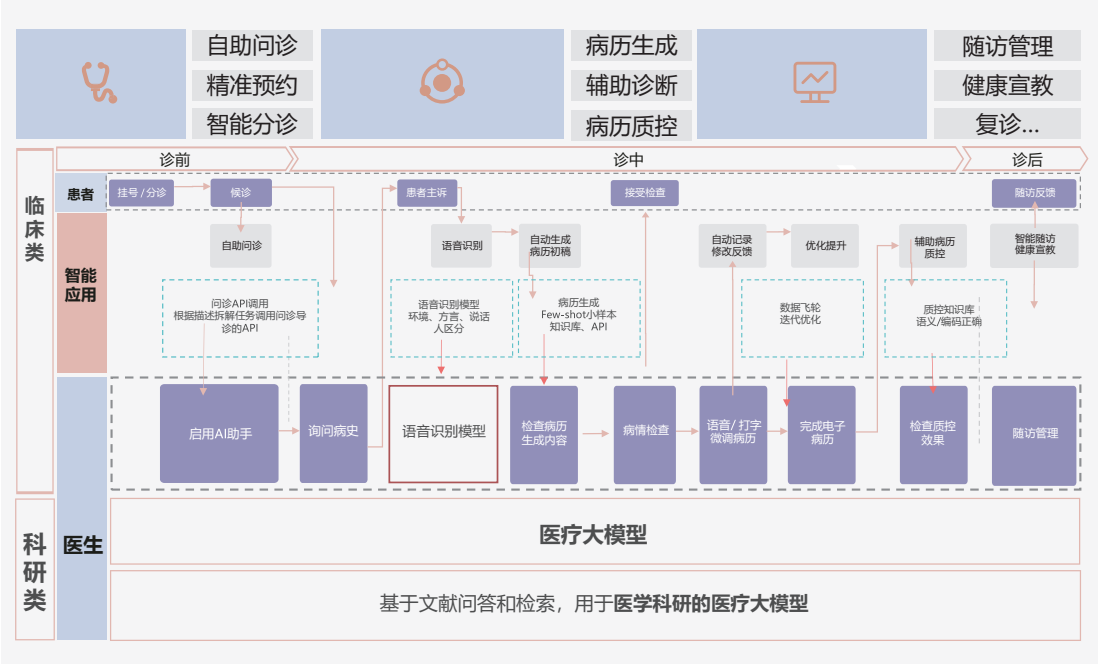


图 24：医疗大模型工作流程

3.4.2建议

1、建立统一 AI 数据湖，实现数据资产可视、可管、可用

更多的行业知识、企业知识的积累，是 AI 大模型迭代升级的前提。当前，企业大量的数据资产分散在分支机构、生产现场，这些数据种类繁多且可能来自不同地域的业务系统、不同合作单位或生态伙伴、甚至是不同厂商的公有云或私有云，形成一个个数据烟囱，制约着 AI 大模型应用的健康发展。

企业需要建立统一的数据湖底座，实现全域数据资产的可视、可管、可用。首先是数据资产一张图，实现跨域、跨站点、跨厂家等复杂数据的全局可视、实时更新；其次是数据目录智能化，满足数据自动标签、聚合、检索、呈现，推进数据按内容、合规、热度等维度的全自动化分类分级；最后再结合算存网协同配合，让归集后的数据可以被高效访问和处理，让数据做到真正可用。只有解决跨组织、跨地域、跨应用的数据统一调度问题，为大模型注入源源不断的数据“燃料”，才能让企业的大模型更好地服务自身业务。

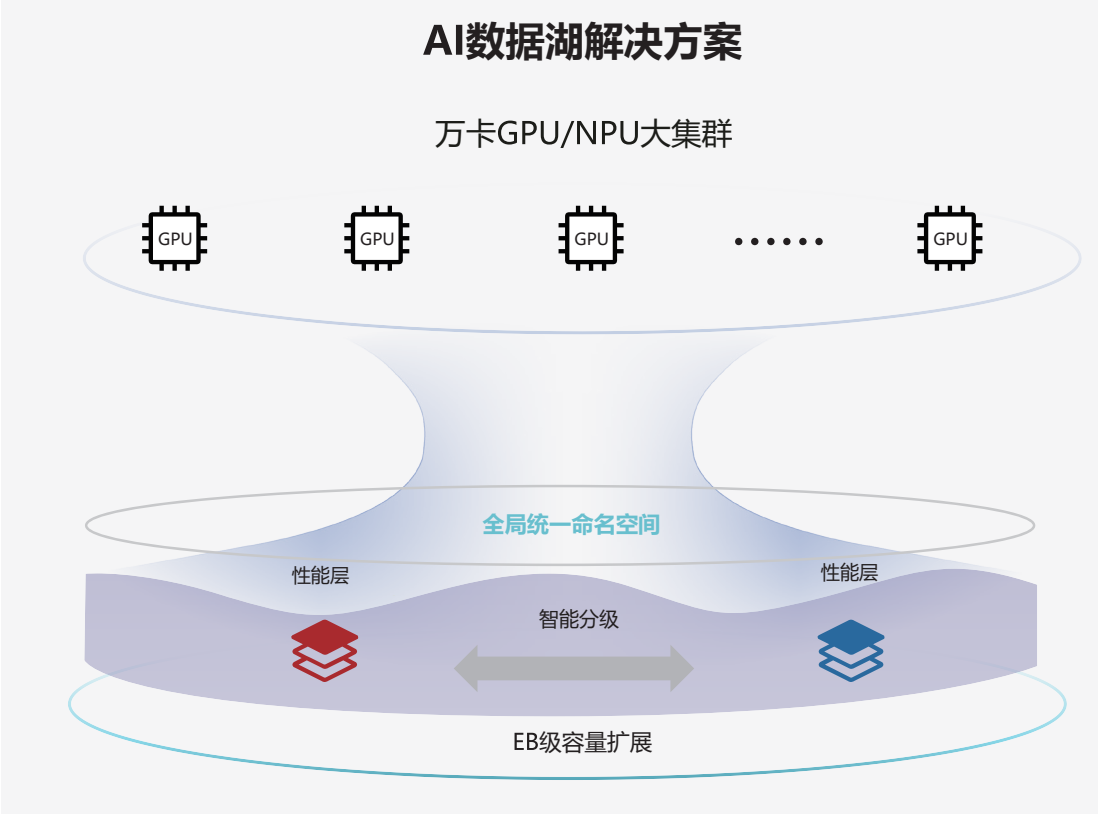


图 25：华为 AI 数据湖解决方案

2、面向训练，选择专业 AI 存储，提升算力利用率，最大化 AI 投资效率

大模型的 Scaling Law 法则持续有效，其技术复杂度正变得越来越高，模型参数量从千亿级到万亿级，集群规模从千卡级到万卡级，训练数据集从 TB 级到 EB 级。这意味着更多的数据要处理、更大参数的大模型、更频繁的再训练和调优。不符合要求的 AI 基础设施将会无形中为企业的智能化升级之路带来额外成本。在业界，NVIDIA 与专业存储厂商合作，基于标准文件系统 +Share Everything 存储架构，共同打造高性能 AI 训练集群。橡树岭国家实验室也在其下一代智算中心技术建议书中提出，只有 AI-Optimized Storage 才能满足大模型在处理 EB 级数据量时对性能、可靠性的要求。

企业需要科学规划智算底座，选择面向 AI 负载优化的专用 AI 存储，从粗放式“堆算力”到“挖潜力”提升集群效率。合理配置存储集群性能，选择高性能、高可靠的外置 AI 存储，可提升集群可用度 10% 以上，减少算力等数据造成投资浪费。

3、面向推理，采用 RAG、长序列等技术，提升大模型推理性能和准确度

企业知识数据日新月异，大模型的周期性训练很难保证时效性、以及在专业知识领域的准确性。从建设成本和应用效果考虑，企业应用 AI 改造方案已逐渐收敛到增强型检索（RAG 技术），通过大模型在生成结果时从数据库中检索出相关知识，生成有参考信息的回答，从而提高推理结果的可信度。在推理阶段，多轮对话、长序列上下文依赖大模型的记忆能力，通过智算处理器 xPU、内存 DRAM、外置存储 SSD 的三层缓存机制，可以将大模型的记忆周期从小时级延展至数年，提升推理的准确度，同时在类似问题的推理需求中通过查询历史结果替代推理来节省算力开销。

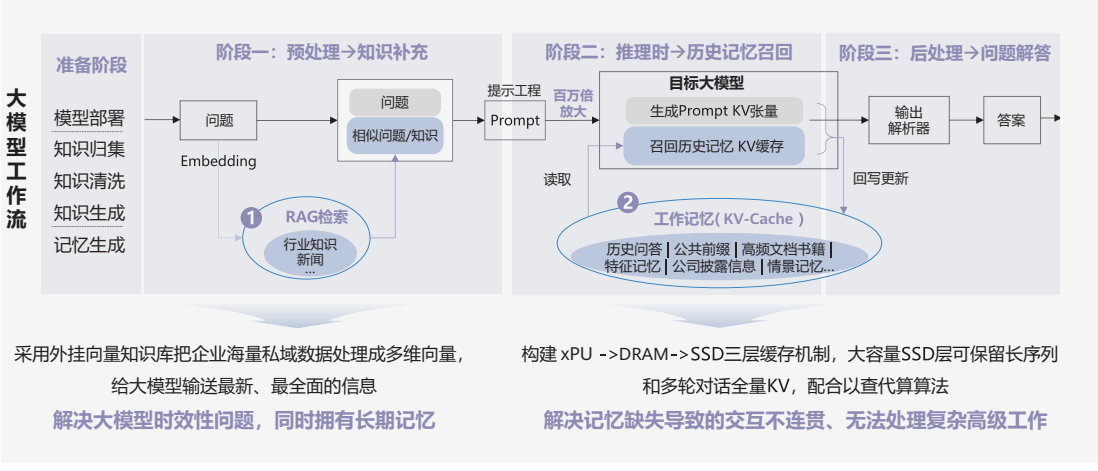


图 26：检索增强生成 RAG+KV-Cache 技术架构图

4、利用容灾、备份、防勒索等手段，加强数据分类分级保护

大模型诞生于海量数据，这些数据囊括用户的个人信息、企业的私域生产数据等敏感信息。伴随着大模型技术的迅猛发展，一系列数据安全风险也开始浮现。样本数据投毒攻击可能使得模型产

生误导性结果，严重影响决策的准确性。模型文件被窃取将导致数亿元投资的成果化为泡影。训练数据被勒索病毒加密则可能导致大模型被迫中断训练，影响企业生产安全。

企业需要重视数据资产的分类分级管理，确定数据的拥有者和使用者，确保数据的合规隐私，从管理、应用、网络到存储，构建全方位的安全解决方案。其中，作为数据的最终载体，存储可提供包括存储软硬件系统安全、数据容灾与备份、防勒索保护以及安全管理在内的一整套内生安全解决方案，为数据构筑最后一道安全防线。

5、增加 AI 人才培养机制，积极开展 AI 大模型实践

AI 大模型应用正在从知识问答、文生图、文生视频等应用，走向以大模型 +Copilot 辅助 +Agent 自主决策的复杂应用，从企业辅助生产走向核心生产，成为企业提升运营效率的关键抓手。

企业应该从顶层设计、组织架构、人才和团队建设等，全面评估生成式 AI 应用的能力预备水平。例如，在顶层设计上，企业是否建立了评估和跟踪开源 AI 大模型、数据和培训模型使用的指导方法，是否研究了业界 AI 基础设施最佳实践案例。在组织架构上，是否设立了相关的数据安全、隐私及伦理的专属团队等。在人才和团队建设上，企业应该培养更多具备对 AI 大模型、尤其是 AI 大模型存储方面拥有深入理解、实战经验的专业人员，构建 AI 大模型的人才培养体系。

3.5 训/推一体机加速AI大模型落地行业应用

助力 AI 大模型快速落地行业应用，训 / 推一体机使能千行万业数智化

AI 发展如火如荼，各行业均在尝试将 AI 落地到行业应用中，却面临基础设施部署、大模型选择、二次训练和监督微调等方面的困难。

训 / 推一体机通过将基础设施、工具软件等进行预集成，并与 AI 大模型供应商协同，可有效助力 AI 快速落地行业应用，使能千行万业数智化。

3.5.1 趋势

1、数据质量参差不齐，数据准备时间长

企业大量的原始数据，清洗成可用的数据集，耗时又复杂。首先，收集大量有代表性和高质量的数据并非易事，可能需要从多个来源获取并整合。其次，清洗数据以去除噪声、错误和重复信息需要耗费大量时间和精力。再者，对数据进行准确的标注以满足模型训练的需求，通常需要专业人员参与，这一过程既耗时又要求高度的准确性。另外，在数据准备过程中，由于各部门的参与度不一，数据质量难以统一，进而影响到大模型的使用效果。

2、硬件选型难、交付周期长，运维成本高

大模型应用需要选择适合的计算、存储、网络等硬件设施。然而硬件种类繁多，性能参数复杂，导致硬件选型难；同时硬件组装、调试、测试、上线等环节复杂，部署上线后的监控、维护和升级等环节繁琐且困难。

3、大模型幻觉严重，推理准确度无法满足业务需求

大模型在面对复杂场景时，输出结果失真，出现大模型幻觉，不仅降低了模型的准确性，在重要的决策场景中，基于错误的信息可能导致严重的后果。在学术研究和知识传播领域，不准确的内容可能误导读者和研究者，甚至可能引发道德和法律风险。

4、数据安全无保证，模型等核心数据资产易泄露

行业高价值数据是企业的核心资产，数据安全性要求高；对于模型厂商来说，行业模型是使能企业模型应用的核心组件，也同样需要保证模型的安全可靠，要避免模型泄露风险。AI 训练数据和模型的安全挑战包括以下几个方面：

a) 数据隐私：训练数据可能包含敏感信息，如个人身份信息、财务数据等。

b) 模型安全：攻击者可能会通过篡改模型参数、注入恶意代码等方式来攻击模型，从而影响模型的输出结果。

c) 对抗攻击：攻击者可能会通过对抗样本来欺骗模型，使其产生错误的输出结果。

d) 模型解释性：AI 模型的黑盒特性使得其输出结果难以解释，这可能会导致模型的不可信度和不可靠性。

e) 模型共享：在模型共享过程中，可能会泄露模型的敏感信息，如模型参数、训练数据等。

f) 模型部署：在模型部署过程中，可能会面临网络攻击、恶意软件注入等安全威胁，从而影响模型的安全性和可靠性。

5、投资回报挑战大

AI 大模型目前仍处在探索期，意味着在软硬件的投资不一定可以按时获得预期的回报，可能导致运营成本超预算执行。大规模的数据处理、图形渲染、深度学习训练等任务中，如果 GPU 利用率过低，会显著降低工作效率，延长任务完成时间。对于企业或研究机构而言，会阻碍创新和发展的速度，影响产品的推出或科研成果的产出。

3.5.2建议

1、通过预集成数据预处理工具链，快速生成高质量训练数据集

高质量的数据是 AI 实现精准推理的基石。AI 专业存储供应商一般会提供数据准备工具链组件，通常提供数十种高性能 AI 算子，能够对多种格式的数据进行自动化清洗（包括解析、过滤、去重、替换等），从而帮助企业用户快速将原始数据转化成高价值的数据集。

2、部署全栈预集成训 / 推一体机用于大模型行业落地

训 / 推超融合一体机通过将计算、存储、网络等硬件预集成、预调优，开箱即用，省去了企业繁琐的选型、组装和调试过程，大大节省了时间和人力成本（包括华为在内的诸多厂商推出训 / 推一体机，预集成成 GPU/NPU 服务器，网络，以及专业存储设备）。同时，通过预置的全栈设备管理软件，对计算、存储、网络、和容器平台等基础硬件和软件平台进行管理运维，大幅降低 IT 人员的日常运维负担，使其可以专注于 AI 业务的实现，而无需为基础设施的搭建和运维感到担忧。

很多训 / 推一体机可以提供高性能机密执行环境，以及数据和模型的机密防护措施。配合数据保护和防勒索，可以对企业用户的关键数据进行充分保护，避免数据资产泄露或者受损。

另外，大多数训 / 推一体机还支持横向扩展，即将多个训 / 推一体机组合在一起，形成一个更大的训 / 推一体化平台。这种能力可以帮助众多企业客户按需部署大模型应用，分散投资周期，减小投资回报风险和压力。

3、利用 RAG 知识库，消除幻觉实现精准推理

通过将高质量数据集嵌入到训 / 推一体机提供知识库存储中，每当用户提出问题，内嵌的 RAG 工程将快速从知识库中检索出预置的知识，帮助推理过程聚焦在正确的上下文环境中，有效解决幻觉问题。另外，通过实时更新知识库，大模型的回答也将具备时效性。内置的模型评估组件可对模型推理的准确度进行评估和追溯，最终实现精准推理。

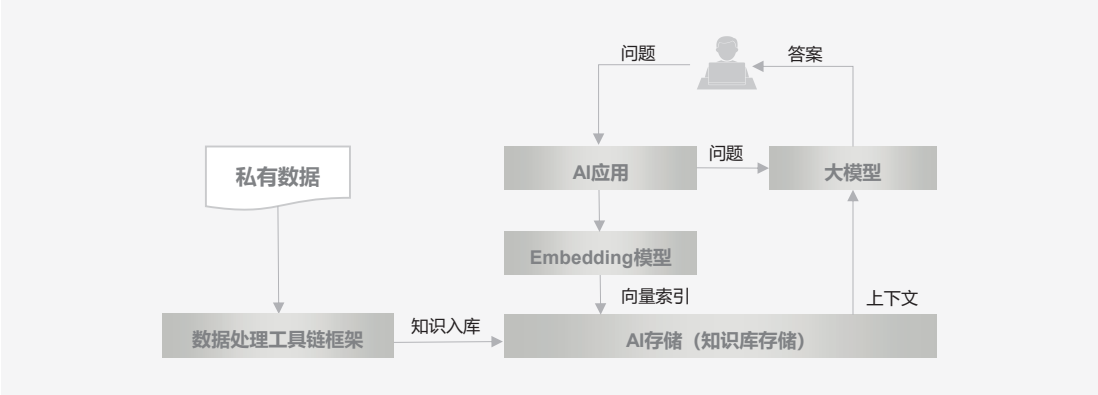


图 27：RAG 知识库工作流程

参考文献

1	《Government AI Readiness Index 2023》 https://oxfordinsights.com/ai-readiness/ai-readiness-index/
2	《人工智能嵌入公共服务治理的风险挑战》 https://www.secrss.com/articles/50686
3	《在税收风险分析中引入人工智能技术的思考》 https://mt.ucass.edu.cn/info/1035/1904.htm
4	《2023 年中国企业勒索病毒攻击态势分析报告》 https://www.qianxin.com/news/detail?news_id=10812
5	《2023 电信 AI 产业发展白皮书》 https://www.iotku.com/news/889569632678051840.html
6	《新一代人工智能基础设施白皮书》 https://www.sensecore.cn/whitepaper.pdf
7	《人工智能技术在智能制造中的典型应用场景与标准体系研究》 https://www.engineering.org.cn/ch/10.15302/J-SSCAE-2018.04.018
8	《36 氪研究院	2024 年中国 AI+ 制造产业研究报告》 https://36kr.com/p/2853458940627588
9	《生成式 AI“进军”制造业：应用范式、趋势与问题》 https://www.tisi.org/27034
10	《如何利用生成式 AI 优化制造企业的生产流程？》 https://m.huxiu.com/article/3082157.html
11	《AI 赋能制造业专题报告：从 9 个细分赛道谈起》 https://www.sohu.com/a/673958291_121124366
12	《什么是合成数据？》 https://aws.amazon.com/cn/what-is/synthetic-data/
13	《合成数据技术的现状、前景和挑战》 https://www.secrss.com/articles/55976
14	Data-centric Artificial Intelligence: A Survey https://arxiv.org/pdf/2303.10158
15	《AI 大模型成新赛点！银行业金融科技竞争加剧》 https://xhrbcj.com/newsDetail?id=d89bc8524614dbf5aa38dd0e92b16757&type=2
16	《同业首家！农业银行发布自主金融 AI 大模型应用 ChatABC》 https://www.cebnet.com.cn/20230331/102868614.html
17	《人工智能嵌入公共服务治理的风险挑战》 https://www.secrss.com/articles/50686
18	《工商银行发布首个金融行业通用模型：基于昇腾 AI，已实现广泛应用》 https://ai.qianjia.com/html/2023-04/04_400384.html

19	《大模型在银行业应用实践：以浦发银行为例》 https://new.qq.com/rain/a/20240131A09C0400#:~:text
20	《金融存力基础设施发展研究报告》 https://www.yunduijie.com/businessreport/view/3956.html
21	《未来已来 – 全球 AI 创新融合应用城市排名及展望》 https://www.baogaozhan.com/57705.html
22	《Global Financial Stability Report》 https://www.imf.org/en/Publications/GFSR/Issues/2024/04/16/global-financial-stability-report-april-2024?cid=bl-com-SM2024-GFSREA2024001
23	《2023 Global Trends in AI Report》 https://www.weka.io/wp-content/uploads/files/resources/2023/08/2023-Global-Trends-AI-Report.pdf
24	《大模型综合能力评测对比表》 https://www.datalearner.com/ai-models/llm-evaluation
25	《OLCF-6 Request for Proposals》 https://www.olcf.ornl.gov/draft-olcf-6-technical-requirements/
26	《上海交大超算平台用户手册》 https://docs.hpc.sjtu.edu.cn/
27	《北京大学现代农业研究院基因分析平台：一粒种子如何撬动世界？》 https://e.huawei.com/cn/case-studies/solutions/storage/institute-of-advanced-agricultural-sciences-peking-university
28	《How AI is Creating Explosive Demand for Training Data》 https://www.unite.ai/how-ai-is-creating-explosive-demand-for-training-data/
29	《全球数据中心用电量到 2026 年或翻番》 https://www.stdaily.com/index/kejixinwen/202401/a3bb743e34134d159c5e7f1e50069ce7.shtml
30	《再生能源发展 借力 AI 少走冤枉路》 https://www.sas.com/zh_tw/insights/articles/analytics/ai-renewable-energy.html
31	《60 Hurts per Second – How We Got Access to Enough Solar Power to Run the United States》 https://www.bitdefender.com/blog/labs/60-hurts-per-second-how-we-got-access-to-enough-solar-power-to-run-the-united-states/
32	《Scam email cyber attacks increase after rise of ChatGPT》 https://technologymagazine.com/articles/scam-email-cyber-attacks-increase-after-rise-of-chatgpt
33	《Bad Bots Account for 73% of Internet Traffic: Analysis》 https://www.securityweek.com/bad-bots-account-for-73-of-internet-traffic-analysis/
34	《黑客攻击破坏、贩卖公民个人信息类案件发案更为频繁》 http://www.chinapeace.gov.cn/chinapeace/c100047/2024-01/05/content_12705164.shtml

华为技术有限公司
深圳龙岗区坂田华为基地
电话：+86 755 28780808
邮编：518129
www.huawei.com

顾问：周跃峰
指导委员会：肖德刚、庞鑫、杨柏梁、方卫峰、樊杰、王振、申珅、孙睿、薛寒、黄亭、罗燚、仇东华
主编：龚涛、裘方佳
编委：华娴、韩茂、庞良硕、陈洋、高覃、苗永刚、蒋华虎、陈振华、冯真、李文秀、刘玮琦、陈辉、蓝国平、陈琳、余乐清、曾帆、袁燕龙、曹晓辉、范珏
（以上排名不分先后）