矩阵起源 Matrix origin

MatrixOne Intelligence

AI原生多模态数据智能解决方案白皮书

Your Data for Your AI

目录

前言

在当今时代，Gen 人工智能（Generative AI，简称 GenAI）正以前所未有的速度席卷全球，成为推动科技进步和产业变革的重要力量。从 ChatGPT 的横空出世到各类大模型的广泛应用，GenAI 不仅在技术层面取得了突破性进展，更在商业和社会层面引发了深远的影响。从文本生成、图像绘制到视频制作，GenAI 的应用场景日益丰富，为各行各业带来了前所未有的机遇与挑战。

据麦肯锡全球研究院（McKinsey Global Institute）的报告，到 2030 年，AI 技术有望为全球 GDP 贡献高达 13 万亿美元的增长。 Gartner 预计在 2026 年，超过 80% 的企业将使用 GenAI 应用程序编程接口（API）或模型，或者在相关生产环境中部署支持 GenAI 的应用程序。这一比例在 2023 年还不到 5%，这意味着在短短三年内，采用或创建 GenAI 模型的企业数量预计将会增长 16 倍。

在 GenAI 的架构中，数据处理的作用尤为关键。AI 技术与数据的紧密联系显而易见：庞大的数据集训练出强大的 AI 模型，而这些模型的功能又能促进数据处理的进一步优化。尽管如此，行业对 GenAI 技术栈中的算力层、模型层和应用层的各项能力及技术方案已有深入探索，但对数据处理层的重视程度仍显不足。在通用基座大模型越来越普及的趋势下，对企业自有数据的挖掘利用将变成 GenAI 落地企业级应用的最关键因素。

矩阵起源作为一家 Data+AI 领域的创业公司，在数据及 AI 领域已经有超过十年的行业经验沉淀。本白皮书将从矩阵起源的专业视角，深入剖析 Data+AI 领域的最新趋势和挑战，并给出如何对企业自有数据进行深度挖掘利用的详细蓝图，以实现更符合企业实际业务价值的 GenAI 应用落地。

GenAI 时代的数据挑战

类人脑计算能力的崛起

驱动 GenAI 技术发展的核心是大语言模型 LLM，其本质上是使用计算机构建巨大的神经网络结构模拟人脑神经元的构成，然后将海量的文本知识压缩到一个有庞大参数量的神经网络中。这样的架构可以给计算机赋予人类一样的交互能力，可以理解人类的语言和需求，再生成便于人类理解的数据。

GenAI 的类人脑计算能力与过去传统意义上计算机擅长的高速数学计算有根本性的区别：

1. 传统计算能力可以轻松完成人类在短时间内难以完成的复杂科学计算，而且工作准确度极高，相同的任务可能需要大量人力进行手动计算整合才能完成，且人类的工作经常出错，但是传统计算能力难以处理以人类自然语言构成的 NLP 任务，比如文档理解、对话理解、图片理解等，而对于人类而言即使是儿童也具备这些能力。
2. 而新型的 GenAI 计算能力是完全模仿人脑的结构所设计的，所展现的能力也跟人类行为极为相似，通过自然语言交互，同样能很轻松的胜任文档理解、对话理解、图片理解等任务，同时具备一定的创造力，可以生成现实中不存在的东西，但是并不擅长复杂的数学计算，准确度也是天生的缺陷。

因此，GenAI 真正带来的是一种全新的类人脑计算能力，它与传统的计算机精确数学计算能力共同构成了我们当今 IT 世界的新型计算基座。

非结构化数据价值开始被挖掘

数据作为 IT 世界的另外一个重要基石，在 GenAI 的新型计算能力加持下也发生了巨大的变化。

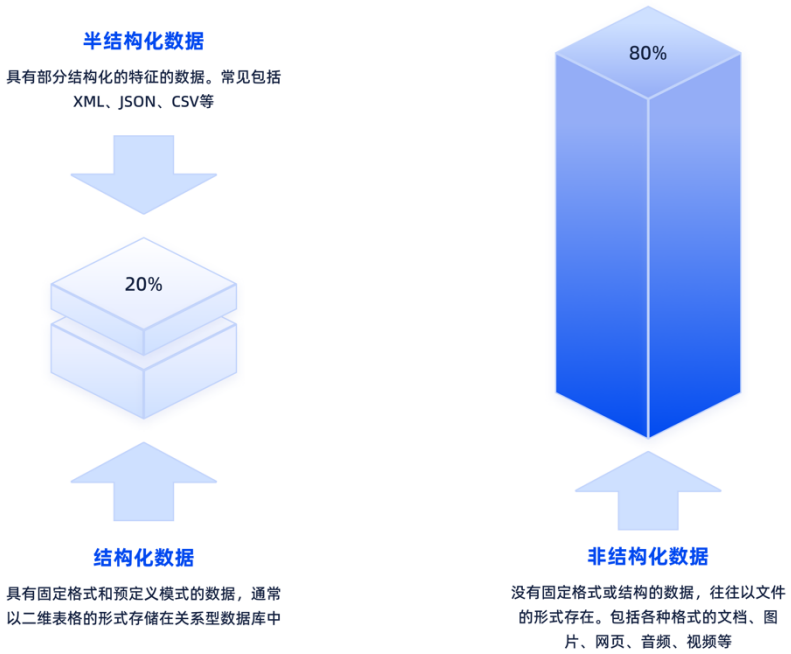
传统上在数据处理领域，我们会把数据分成三类，结构化数据、半结构化数据和非结构化数据：

- **结构化数据**是定量数据，由值和数字组成，是高度组织化的数据，易于访问和解释，它们往往以二维表格和数据库的形式存在。

- **非结构化数据**是定性数据，没有内部结构，由文本、视频和图像组成，包括各种格式的办公文档、图片、网页、音频/视频信息等，这些数据往往以文件的形式存在。
- **半结构化数据**则位于两者之间，它一般是自描述的，数据的结构和内容混在一起，没有明显的区分，如 JSON、XML 等格式的数据。

在过去数十年的 Data Infra 领域发展过程中，结构化数据和半结构化数据处理都是其中绝对的主角，结构化和半结构化数据由业务流程产生，与商业价值高度相关，这些数据与企业的流程业务及商业化息息相关，Data Infra 软件领域也逐渐演化出了非常成熟的产品及处理能力。

然而，根据 Gartner 的数据显示，结构化和半结构化数据仅仅占到全世界数据比例的不到 20%，其他 80%以上均是非结构化数据。在过去的技术能力下，非结构化数据难以处理，价值难以被挖掘和衡量，有研究显示大量办公文档类的数据在整个生命周期内最多只被使用过 2 次，相比较其被努力创造出来的投入相比产生的价值极为有限。因此，非结构化数据长期被当成企业负资产的存在。



本质上而言，非结构化数据实际上是为了方便人类与计算机交互，所创造出来的专为人类服务的各类格式，其与人类的理解能力及使用习惯息息相关，但是其对于传统的以数学计算为主要能力的计算机而言则难以被解析和处理。

而如今 GenAI 技术的出现则彻底打破了这个现状，一方面 AI 大模型本身即是由海量非结构化的文档及多模态数据训练而成，企业可以应用自身沉淀的大量非结构化数据进行模型训练及精调，另一方面在如 RAG 类型的技术框架的帮助下，非结构化数据可以通过 AI 解析及外挂向量数据库的方式得以实现解析及结构化，用户可以轻松实现如

ChatWithPdf 等类型的业务。

占全球数据 80%以上的非结构化数据得以实现价值解锁，其中蕴含的丰富业务洞察、客户需求和市场趋势，可以为企业创新、决策提供更进一步的数据价值。

企业落地 GenAI 的数据困境

在过去两年 GenAI 技术突飞猛进的背景下，企业普遍已经充分认识到了以 AI 大模型为基础的智能化升级的重要性，大量企业也都开始在开展与 GenAI 相关的技术预研及试验性的落地尝试。然而，由于通用 AI 大模型本质上是海量公开知识的压缩，在企业级场景落地中必然会碰到对企业相关语言和业务理解不准确的问题。

但是对于绝大部分企业而言，都对于更加准确的解决自身商业问题会提出较高的要求。而为了让通用大模型在行业中提高解决业务问题的精确度，不管是通用模型适应行业所进行精调方案，还是通过 RAG 架构进行知识外挂的方案，都离不开企业自有高质量数据的融入。同时面向行业的 GenAI 方案对于企业自有数据的要求往往是混合类型的，多模态的，既包括已经有相对较完善的 Data Infra 处理的结构化及半结构化数据，也包含了过去未经过处理的多模态非结构化数据，而这样的数据需求给企业落地 GenAI 提出了巨大的挑战。

在观察了大量企业实验落地 GenAI 的过程后，我们总结了以下问题：

● 严重的数据碎片化问题

在 GenAI 浪潮到来之前，企业的数据处理重点多集中于结构化数据的整合与优化，许多企业通过打破烟囱式业务系统构建了数据中台。然而，GenAI 应用场景对数据的要求远超以往，尤其是多模态数据的整合，其来源分散且管理复杂。非结构化数据通常分布在云盘、内部 IM 工具、对象存储、业务系统、服务器文件系统和个人设备中，创建与存储时缺乏统一的管理流程。而结构化数据在 GenAI 场景中也需与非结构化数据混用，不同数据类型间的关联进一步增加了碎片化程度。企业不仅需要高效整合这些异构数据源，还需确保权限与隐私的分级管理，以满足合规和安全要求。

● 异构多模态数据整合的复杂性

为了让 GenAI 在业务场景中真正创造价值，企业需要同时整合结构化、半结构化和非结构化数据进行融合使用。尤其是非结构化数据，因其多样的格式和模态 (如 Word、PPT、PDF、JPEG、WAV、MP4 等)，每种格式都涉及复杂的解析与治理流程。以 PDF 为例，

其处理链路包括版式检测与分割、内容识别（如文字、表格、图片）以及特征抽取。如果需要进一步与结构化和半结构化数据整合，整体链路的复杂性会成倍增加。对于缺乏深厚数据和 AI 工程能力的企业，这些技术门槛难以逾越。

● 规模化部署和管理难度高

GenAI 的应用和多模态数据处理高度依赖强大的 IT 基础设施。构建一个使用数十份文档的 RAG Demo 相对简单，但在真实的大型生产环境中，企业通常需要处理 PB 级别的数据，并进行复杂的模型精调和训练。这不仅需要大量高性能 GPU 和 CPU 的算力支持，还需依托大容量存储和高带宽、低延迟的网络架构。同时，底层资源管理平台也必须具备资源调度与自动化扩展的能力，支持多模态数据的预处理与存储，并以云原生架构为基础，确保跨环境的灵活部署和高效管理。

● 数据召回与输出准确率的局限

作为 GenAI 的核心技术，大模型本质上是基于概率分布生成输出内容，这一特性决定了其难以在高要求的企业业务场景中做到完全精确。准确率不足直接影响了商业价值，因此结合企业自身数据进行模型精调或采用 RAG（检索增强生成）框架成为必然选择。然而，这些优化技术本身存在较高的技术门槛。例如，在 RAG 框架中，基础的向量语义搜索对于短词短句的提问效果不佳，难以满足高精度的业务需求，需要引入多种搜索方式（如关键词匹配和全文检索）进行补充。而在企业普遍期待的 Chat2BI 应用场景中，直接使用大模型生成的 SQL 往往准确率较低，需通过工程化手段如语法校验与结果优化，来提高其实际可用性。

典型行业场景的落地难题

以下列举了三个典型企业场景的真实案例，展示 GenAI 企业级落地中数据层面的主要障碍：

● 报业传媒集团公司

该集团企业已经成立近 30 年，其看到了 GenAI 在内容生产上的强大能力，希望在内容生产领域能对自身业务进行赋能，在内容生产的工作流程中嵌入 AI 的能力。但是通用大模型的效果不够理想，而该集团企业拥有海量的媒体素材，包括历史报刊的数字化文件、大量的图片、音视频资料等，因此其希望将自有海量素材与大模型能力结合起来，再嵌入到自身的工作流。然而经过梳理和盘点后发现，这些素材数据散放在各种业务系统、硬盘、

云网盘等空间里，碎片化极为严重；同时缺乏手段可以从中找到与希望生产的内容主题相关的素材，一线编辑仅能凭记忆和少量筛选的模式来从中获取极少量素材；另外如何将这些素材与大模型结合起来落地，不管是做精调还是 RAG 方案，对于该企业而言也存在巨大的技术和资源门槛。

● 大型电子制造公司

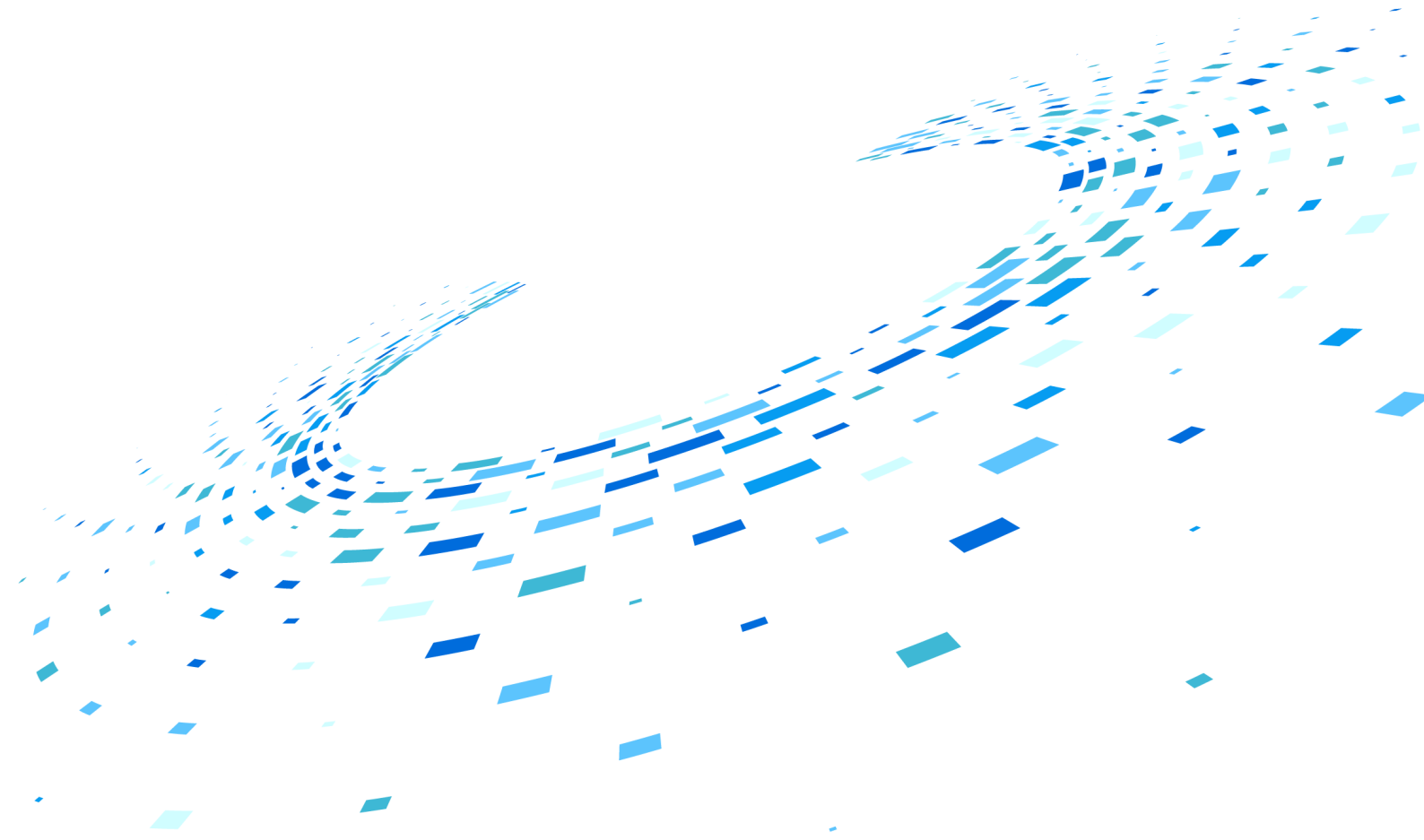
该公司是一家年产值上百亿元的电子生产企业，拥有多家工厂和数十条电子产品的生产线。该企业长期在产线上采集大量各类型数据，包括生产设备产生的结构化、文档、图片数据，还有员工操作的音视频数据，结构化数据相对已经比较好地被 MES 系统数据库所承载，然而其他类型的数据还一直无法得到有效处理。举例来说，针对工人在某道工序上的操作规范，每个工位有摄像头采集了员工操作视频，该企业目前只能通过人工抽查视频的形式判断员工是否戴手套，是否有磕碰等行为，但是其覆盖率仅有不到 5%，同时很难再与其他系统数据进行关联分析。GenAI 的出现对该企业而言是一个新的契机，然而对于多模态数据的统一治理及与大模型的联动，同样超出了该公司的技术能力。

● 市级政府规划部门

该部门为某市级政府的发展规划部门，该市经济主要以传统工业为主，因此招商引资的政策倾向也比较看重新能源相关的高附加值工业。该部门的招商人员长期以来都需要关注多个细分产业的市场动向及各地政策情况，以对自身发展决策提供依据。然而长期以来都只能依靠人工方式去网络上搜索获取相关信息，再经过人工的整合归纳后，结合自身产业数据，按月发布相关报告，以供决策层领导使用。GenAI 技术出现以后，招商人员开始通过使用公开的大模型工具来进行更进一步的搜索和整合，提高了一定的工作效率。然而面对如行业咨询报告，上市公司财报，工商注册信息等更加复杂及多元的数据，实际上通用 GenAI 大模型工具输出效果并不理想，特别是涉及到当地产业的各类文件、政策、统计数据等，处理复杂度将进一步提升。

总结

GenAI 的价值已被企业广泛认可，并在实际应用中初步落地。然而，要真正解决业务问题并发挥其商业潜力，GenAI 的实施必须依赖高质量的企业自有数据。长期以来，非结构化数据的潜在价值未被充分挖掘，而 GenAI 的出现为这些数据的激活带来了全新的可能性，同时也提升了其在企业应用中的商业价值。然而，大多数企业的数据仍未达到 AI-Ready 的高质量标准，面临工程复杂性和资源高成本的挑战。当前，行业亟需一套高效且全面的解决方案，来应对混合多模态数据的整合与利用。



MatrixOne Intelligence AI 原生多模态数据智能解决方案

MatrixOne Intelligence 概述

矩阵起源自成立以来，一直以为数字世界提供简捷强大的数据智能操作系统作为使命，致力于让企业和用户简单、敏捷、高效地拥抱数据价值。

MatrixOne Intelligence 是一套面向多模态数据的 AI 数据智能解决方案，旨在帮助企业应对数据碎片化、多模态数据整合复杂、GenAI 应用落地困难等挑战。通过集成数据治理、智能解析、多模态搜索和超融合数据底座等功能，MatrixOne Intelligence 为企业提供了一站式的端到端平台解决方案。该平台基于创新的云原生架构和存算分离设计，支持结构化、半结构化和非结构化数据的统一管理和高效处理，具备高度灵活的部署能力，可适配公有云、私有云及本地数据中心的多种环境。MatrixOne Intelligence 解决方案的目的是将企业内部的自有数据变成可以服务于 GenAI 落地应用的 AI-Ready 数据，并且对业务产生价值。而这个目标本质上就是提高大模型在企业应用场景下的准确度。

MatrixOne Intelligence 致力于赋能企业，帮助企业充分挖掘和释放自身数据的潜能，让企业自有数据在 AI 时代得到充分利用，成为其独特竞争力的关键来源。

解决方案架构介绍

在前文中，我们探讨了 GenAI 在企业级应用落地中面临的数据挑战，包括数据碎片化、异构多模态数据整合复杂、以及自有数据的价值难以充分释放。这些问题严重限制了企业在数据智能时代的竞争力和效率。而 MatrixOne Intelligence 作为一套面向多模态数据的 AI 数据智能解决方案，正是针对这些关键痛点设计，为企业提供了一条从数据到智能的全新路径。

为解决这些挑战，MatrixOne Intelligence 通过统一的底层资源管理、全链路数据治理、多模态数据融合存储、建模及搜索能力，搭建了一套端到端的数据智能架构。如下图所示，该解决方案自下而上分为四个层次，分别是基础设施层、数据库及 AI 服务层、数据集成与治理层、以及应用交互层。这四个层次环环相扣，共同构建出一个强大的数据智能解决方案。

● 基础设施层

基础设施层是整个解决方案的 IT 资源底座，它整合了 CPU 和 GPU 计算能力，支持大规模的并行处理，确保 AI 模型和数据处理的高效运行。容器编排与管理功能提升了系统的可扩展性和灵活性，为企业提供高效的资源调度和负载均衡。

● 数据库及 AI 服务层

数据库及 AI 服务层提供了完善的数据库及 AI 模型能力底座。其支持结构化、半结构化及非结构化数据的融合存储与建模，同时提供 LLM 模型、Embedding 模型和自定义模型训练功能，也提供了快速构建智能体的工作流工具能力。

● 数据集成与治理层

数据集成与治理层负责从左侧各类数据源中采集、清洗和转换数据，进行统一的预处理和特征工程。它支持结构化、半结构化及非结构化数据的整合与处理，确保数据的质量和一致性，为后续的分析和 AI 模型训练提供可靠的数据基础。

● 应用交互层

应用交互层是用户与整体方案的界面接口，用户既可以直接使用我们提供的多模态搜索及 Chat2BI 等终端应用，也可以通过 API 及工作流工具的形式自行构建相关应用。

核心产品概述

如前文架构图所示，MatrixOne Intelligence 解决方案包含五款核心软件产品，它们分别对应解决方案架构中的不同层次，构成了完整的技术体系。这些产品通过协同工作，将基础设施、数据集成、治理、存储、分析以及交互能力无缝连接起来，提供了一套一站式、端到端的多模态数据智能解决方案。

接下来，我们将逐一介绍这五款核心产品，详细阐述它们在不同层次中的功能定位和独特价值，展示它们如何协作以应对企业在 GenAI 落地中的数据和智能挑战。

● MatrixDC 高性能算网调度平台

MatrixDC 是一套高性能算网调度平台，它作为资源底座，通过 K8s 容器、RDMA 高速网络、对象存储等基础能力打造了一套将 CPU 及 GPU 服务器统一纳管、组网、调度及运营的平台。MatrixDC 集成了全面的容器编排与管理能力，通过 Kubernetes 等云原生技术实现算力网络及存储资源的弹性扩展与高效利用。MatrixDC 支持多种计算资源的整合，包括 CPU、GPU 以及存储和网络资源，能够满足从小规模实验到大规模生产环境的多样化需求。通过容器化技术和分布式部署架构，MatrixDC 为企业提供了灵活的资源分配方式，支持 Serverless 化服务调用，帮助用户在应对复杂计算任务的同时大幅降低运维成本。此外，MatrixDC 具备低延迟、高吞吐的网络优化能力，能够保障多节点间高效通信，是多模态 AI 任务运行和大模型训练的强大技术基石。

● MatrixOne 超融合云原生数据库

MatrixOne 是 MatrixOne Intelligence 平台的核心数据管理底座，旨在为企业提供一套全面的超融合数据库解决方案，以支持面向 GenAI 的多模态数据的高效处理。其采用存算分离与云原生架构设计，支持结构化、半结构化和非结构化数据的统一存储与查询。MatrixOne 具备多模态数据融合处理能力，可同时支持事务型（OLTP）、分析型（OLAP）、向量检索、全文搜索和时序数据查询，极大地简化了企业复杂数据负载的管理需求。此外，MatrixOne 具备强大的快照功能，为 GenAI 中快速动态变化的训练集、验证集和评估集的数据版本化提供了可靠支持。通过与 MatrixGenesis 及 MatrixPipeline 的深度集成，MatrixOne 能够快速完成数据解析、向量化和特征工程，并支持高性能的多维度检索与召回。

● MatrixGenesis AI 智能体应用开发平台

MatrixGenesis 是 MatrixOne Intelligence 平台中的 AI 服务模块，专注于为企业提供

大模型支持与智能应用开发能力。作为企业 AI 转型的核心工具，MatrixGenesis 涵盖从模型训练与精调到推理部署的全生命周期管理，帮助企业将 GenAI 快速应用于实际业务场景。通过整合先进的大模型服务（如 LLM 和多模态模型）和 MaaS（模型即服务）平台，MatrixGenesis 支持灵活配置和扩展，适应多样化的行业需求。此外，MatrixGenesis 具备强大的 Agent 工作流设计与开发功能，使企业能够快速构建面向特定业务场景的智能体应用。凭借高效的工作流管理工具和便捷的模型集成能力，MatrixGenesis 大幅降低了企业在 AI 应用开发中的技术门槛，为 GenAI 的规模化落地提供了坚实支撑。

● MatrixPipeline 多模态数据工程平台

MatrixPipeline 是 MatrixOne Intelligence 平台中的数据处理与治理模块，专为企业提供多模态数据的高效接入、转换和管理能力。作为数据流的核心引擎，MatrixPipeline 支持从结构化、半结构化到非结构化数据的统一接入，通过灵活的连接器与自动化 ETL 流程，帮助企业轻松整合多源数据。其内置的预处理与解析功能能够针对不同数据格式（如 PDF、Word、JPEG、视频、音频等）进行智能解析、内容抽取和特征工程，为后续模型训练和推理提供高质量的数据支持。此外，MatrixPipeline 还具备数据清洗、增强和标注能力，结合大模型提供的嵌入式标注与自动化特征生成功能，大幅提升数据治理的效率与准确性。通过与 MatrixOne 数据库的深度集成，MatrixPipeline 可以实现无缝的数据流管理，支持高效的数据版本管理和全生命周期追踪。作为企业数据智能化的基础模块，MatrixPipeline 简化了复杂的数据管道构建流程，显著降低了多模态数据治理的技术门槛。

● MatrixSearch 多模态智能搜索引擎

MatrixSearch 是 MatrixOne Intelligence 平台的多模态智能搜索引擎，专为企业提供强大的跨模态检索与语义查询能力。通过集成向量检索、全文检索和结构化查询，MatrixSearch 支持文本、图像、音频和视频等多种数据类型的高效检索，帮助企业从多模态数据中快速获取关键信息。其创新的混合搜索机制结合了语义理解与自然语言查询，能够深入解析用户意图，无论是结构化问题（如 SQL 查询）还是非结构化场景（如语音指令或文档问答），均可提供精准的检索结果。此外，MatrixSearch 内置多路召回与智能排序算法，将向量匹配与关键词检索结果进行综合优化，确保结果的相关性与准确性。凭借对多模态数据的全面支持以及与业务场景深度结合的灵活性，MatrixSearch 为企业实现数据驱动决策和 GenAI 的智能化应用提供了重要支撑。

解决方案技术特点及优势

MatrixOne Intelligence 采用现代 IT 架构设计的核心原则，构建了一个 模块化、高可扩展性和高可靠性的技术体系，充分适应企业多样化的数据和 AI 应用需求。整个平台基于云原生架构，利用容器化、微服务和分布式计算技术，实现了系统的灵活部署和弹性扩展。模块化设计使各功能组件（如数据集成、治理、存储、AI 模型服务、搜索引擎等）独立运行并可自由组合，方便企业根据需求快速调整和扩展业务能力。MatrixOne Intelligence 解决方案在以下六个方面展现了显著的优势。

● 一站式端到端平台能力

MatrixOne Intelligence 提供了一个高度集成的多模态数据智能平台，涵盖从数据接入、治理、分析到应用的全流程。企业无需在多个独立系统间迁移数据或自定义开发，大幅降低实施复杂度和开发成本，实现业务快速上线。

● 弹性高效的资源调度

平台采用云原生架构和 Serverless 计算模式，支持 CPU、GPU 及存储资源的按需扩展和动态调度。存算分离的设计进一步增强了灵活性和经济性，使企业无需复杂的资源规划即可轻松应对业务波动，优化了整体成本结构。

● 超融合数据处理能力

MatrixOne 以单一引擎支持结构化、半结构化和非结构化数据的统一存储与计算，同时兼容 OLTP、OLAP、向量、全文、时序等多种混合负载。相比传统的多系统架构，这种超融合方式简化了数据管理流程，显著减少企业在架构设计和运维上的投入，快速释放数据潜能。

● 动态数据版本管理

MatrixOne 内置强大的快照能力，可以对数据版本进行灵活管理，支持对多版本数据的记录、比较和回溯，确保数据处理的可追溯性和一致性。这不仅满足了企业在数据审计和法规合规方面的需求，还能加速 AI 模型的迭代优化，增强数据管理的灵活性。

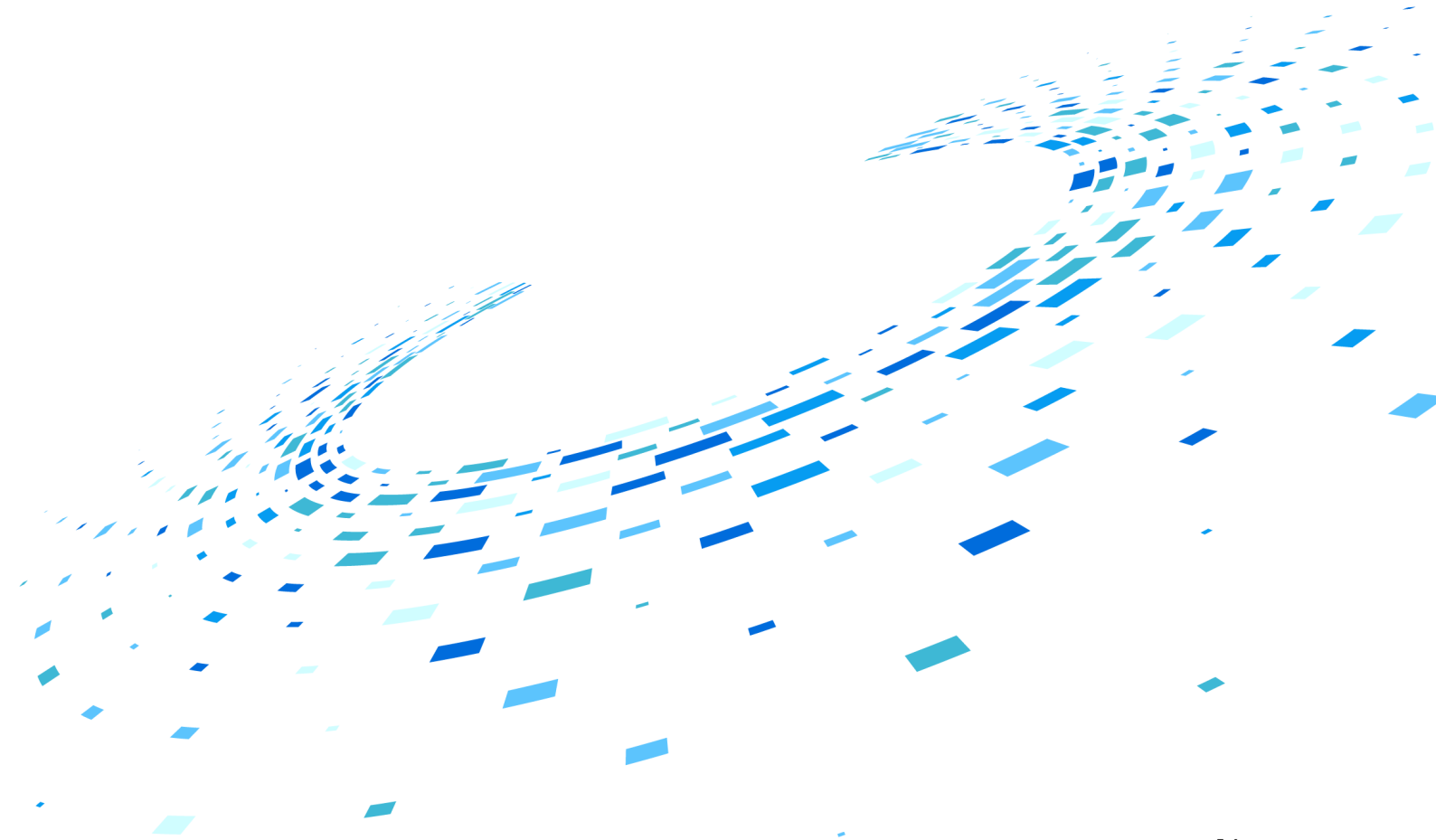
● AI 驱动的高效数据治理

平台内置 AI 能力，可自动处理文本、图像、音频和视频等多模态数据，完成数据的提取、标注、分类和特征工程，全面提升数据治理效率。企业能够快速构建高质量的 AI-ready

数据资产，为 GenAI 的落地提供可靠支撑。

● 混合多模态搜索引擎

MatrixSearch 集成语义搜索、全文检索和结构化查询能力，支持跨数据库、文档、音视频等多模态数据的检索。其多路召回与混合重排算法确保结果的相关性和准确性，为企业用户提供高效的知识获取体验，并显著提升数据服务能力。



解决方案技术流程详解

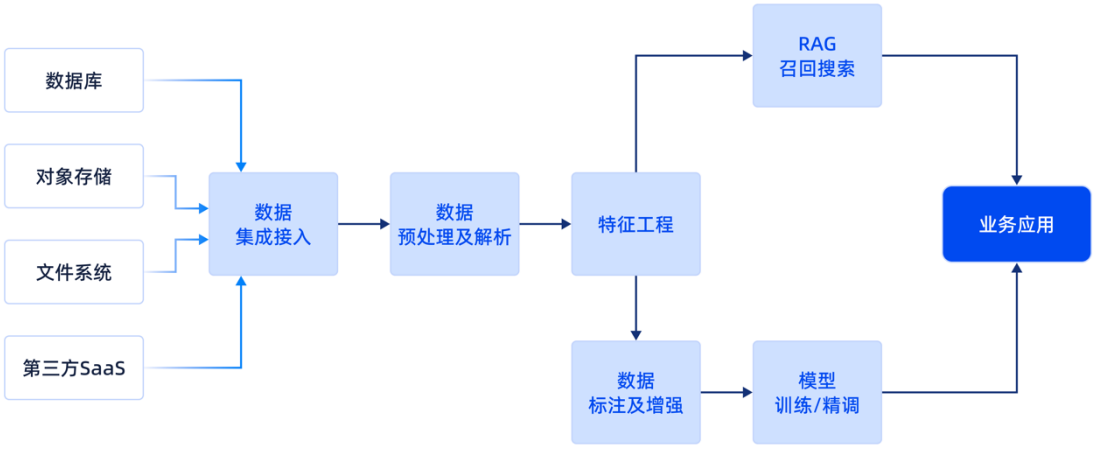
在明确了解决方案的整体架构及核心能力之后，本章将从数据流转链路 Data Pipeline 的角度，详细拆解 MatrixOne Intelligence 解决方案的技术实施流程，展示从数据接入到智能应用的完整闭环。作为一个面向多模态数据的 AI 数据智能平台，MatrixOne Intelligence 的实施流程涵盖了数据接入与整合、预处理与治理、标注与特征工程、存储与管理、模型训练与评估，以及最终的数据召回与搜索。这些环节共同构成了一个高效的技术体系，帮助企业将分散的多模态数据转化为可驱动业务的智能资产。

整体数据流程

MatrixOne Intelligence 解决方案的目的是将企业内部的自有数据变成可以服务于 GenAI 落地应用的 AI-Ready 数据，并且对业务产生价值。而这个目标本质上就是提高大模型在企业应用场景下的准确度。

当前行业中有四种较为常见的做法来实现这个目标：提示词工程，RAG，模型精调，预训练。其中提示词工程与 RAG 都需要基于对企业数据的挖掘，对大模型输入进行更符合场景背景的提示用语，我们可以将其归纳为面向推理的 GenAI 数据工程。而模型精调及预训练则是将自有数据用于训练更行业化的模型，我们可以将其归纳为面向训练的 GenAI 数据工程。这两条链路构成了企业 GenAI Data Pipeline 的基本框架，这两者也可以同时存在并相互配合。而在数据被加工到服务于模型训练或者推理之前，会有公用的数据链路，如数据接入、数据清洗及预处理、数据解析和特征工程的工作。

整体 Data Pipeline 可以总结到如下流程图中：



接下来我们会逐个分析其中每个关键环节的场景，数据加工的技术要求，以及 MatrixOne Intelligence 解决方案中的产品能力如何匹配该环节的需求。

数据接入与整合

● 环节概述

前文已经详细描述过企业客户在面向 GenAI 应用场景时，普遍面临新一轮的数据孤岛问题。各类数据源可能分布于不同的数据库（如关系型数据库、NoSQL 数据库）、文件系统（本地或云存储）、第三方 SaaS 应用（如网盘、IM 工具）以及边缘设备等环境中。这些数据不仅物理位置分散，格式上也高度异构，涵盖结构化数据（如数据库表）、半结构化数据（如 JSON、XML）以及非结构化数据（如 PDF 文档、图像、视频、音频等）。

这种分散和多样化的数据形态带来了以下关键问题和需求：

1. 数据获取与整合复杂：数据分布在多个系统和位置，缺乏统一的接入和管理方式，导致数据整合工作量大且效率低下。
2. 非结构化数据处理压力：非结构化数据体量巨大（如视频和音频文件），完全采用中心化的接入方式会带来带宽瓶颈、高延迟和高成本问题。
3. 多模态数据标准化：数据格式不一致，解析和标准化过程繁琐，难以直接为 AI 建模

和应用提供支持。

4. 安全性与权限管理：跨部门或跨系统的数据访问需要精细化的权限控制，确保数据在接入和管理过程中的安全性和合规性。

因此，本环节的核心目标是解决数据的分散和异构性问题，构建一个支持多数据源统一接入、云边协同处理和分布式管理的架构。通过高效整合结构化、半结构化和非结构化数据，并提供灵活的权限控制和标准化处理能力，为后续的 AI 建模和智能化应用奠定坚实的数据基础。

● 技术流程

在数据接入和整合环节，MatrixOne Intelligence 通过 MatrixPipeline 产品提供了一整套强大的功能模块，以高效、安全地实现多数据源的整合和管理，具体包括以下过程：

多源异构数据的统一接入

- **广泛的数据源支持**：支持连接结构化数据（如 MySQL、PostgreSQL）、半结构化数据（如 JSON、XML）以及非结构化数据（如 PDF、图像、音视频），并通过标准化接口与主流第三方 SaaS 应用（如百度网盘、飞书）无缝集成。
- **虚拟化接入**：通过 Data Fabric 架构，支持对分布式数据的逻辑统一访问，无需数据物理迁移即可完成接入与治理。

云边协同数据处理

- **边缘侧数据初步处理**：针对非结构化数据量大的场景（d 如视频、图片数据），支持在边缘设备完成数据采集、过滤和压缩，将精简后的数据上传至云端，减少带宽占用和传输延迟。
- **云端集中管理**：云端处理复杂的多模态数据解析和深度检索任务，实现云边协作，提升处理效率和数据响应速度。

实时与批量同步能力

- **实时数据接入**：通过实时流处理工具，支持实时数据流的接入，确保企业能够及时响应动态业务需求。
- **批量历史数据加载**：支持从传统存储系统或数据仓库中高效导入历史数据，为全量分析构建完整的数据视图。

分布式元数据管理

- **全局元数据目录**：建立分布式元数据索引，对多位置、多格式的数据进行统一管理和定位，快速检索数据而无需直接访问源文件。
- **智能化数据调度**：根据数据访问频率和业务需求，动态优化数据的存取路径，在边缘和云端间实现资源最优分配。

安全性与权限控制

- **精细化权限管理**：基于角色的访问控制（RBAC）机制，为不同用户和部门提供多层级权限配置，确保数据的使用和共享安全合规。
- **加密与脱敏**：支持数据传输中的 SSL 加密，以及敏感数据的脱敏存储，全面保障数据安全。

通过上述能力，可以有效的解决企业在多模态数据接入与整合环节的难题，为后续的 AI 建模和智能化应用奠定了坚实的数据基础。

● 产品能力

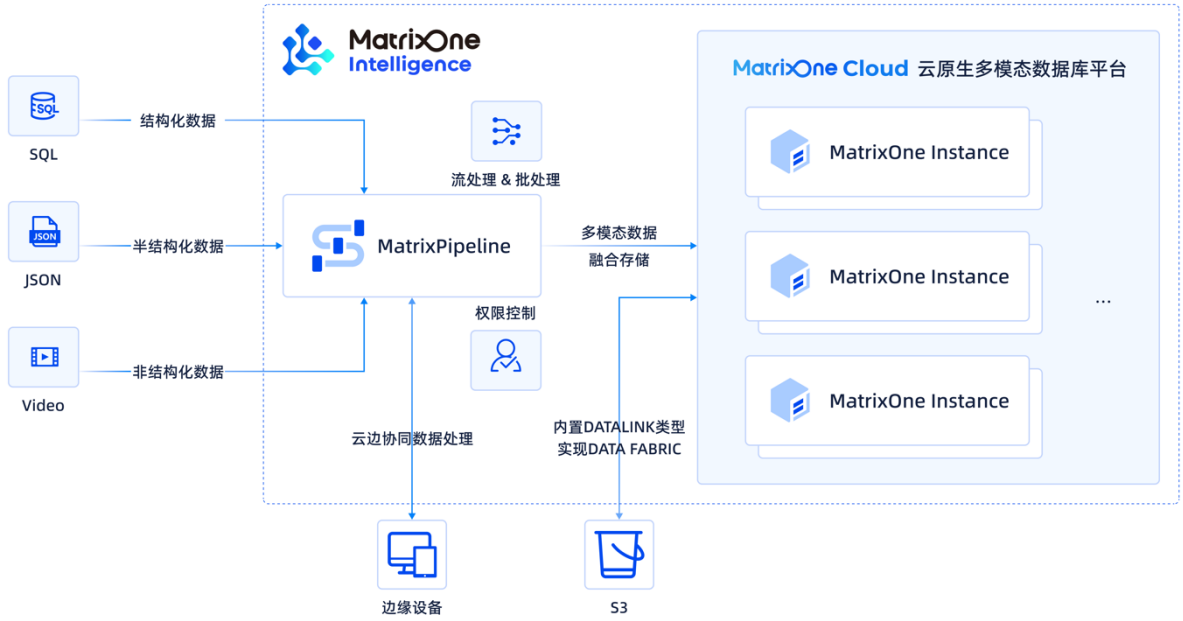
在数据接入与整合环节，MatrixOne Intelligence 通过以下核心产品提供支撑能力：

MatrixPipeline 数据连接器

- 提供灵活的各类数据连接器，支持多种异构数据源的快速接入。
- 虚内置流处理和批量数据同步能力，支持实时与历史数据的高效导入。
- 提供数据标准化工具，包括格式转换、元数据生成和权限控制功能。
- 支持边缘节点与云端协同工作，通过边缘设备完成数据的初步解析和压缩。
- 在云端集中处理复杂任务，并通过智能调度优化资源使用。

MatrixOne 多模态数据管理

- MatrixOne 作为统一的云原生多模态数据库平台，支持结构化、半结构化和非结构化数据的融合存储。
- MatrixOne 通过 Datalink 及 Stage 能力直接链接外部存储中的数据，实现 Data Fabric 架构。
- MatrixOne 具备 ACID 能力，可以保证在数据导入和传输过程中 exactly-once 的能力。
- MatrixOne 提供分布式元数据管理和全局索引服务，支持跨节点快速检索。



多源异构数据统一接入：从 MatrixPipeline 到 MatrixOne 融合存储

数据预处理与解析

● 环节概述

在整体方案中，预处理和解析是数据从原始状态转化为高质量 AI-ready 数据的关键环节。从上一个环节中，我们提取了大量以各种格式存在的非结构化数据，例如文档类的 DOCX、PPT、PDF、Markdown 等，图片类的 JPG、BMP、SVG 等，音频类的 WAV、MP3、WMA 等，视频类的 MP4、MOV 等，网页类的 HTML。然而，这些数据由于格式多样化、内容复杂性高、数据质量不一，无法直接输入到 AI 训练或推理流程中。预处理和解析的目标是统一处理这些多模态数据，将其转化为结构化或半结构化的形式，同时提升数据的质量和一致性。这包括清洗冗余数据、修复缺失值、提取内容的核心特征，例如从文档中提取文本信息，从图片中识别对象和场景，从音频中转录语音文本，以及从视频中提取关键帧与标签。通过标准化的方式统一格式并消除噪声，为后续的建模、训练和推理打下坚实基础。同时，该环节还需要支持自动化处理流程，以便应对大规模、多格式、多模态数据的高效转换和解析需求，从而最大限度降低手动操作成本并提升处理效率。

● 技术流程

在数据预处理及解析环节，MatrixOne Intelligence 结合 MatrixPipeline 自动化管道和 MatrixGenesis 的智能解析能力，提供了一整套高效、灵活的解决方案，覆盖从数据清

洗到数据解析的完整链路。

数据预处理

针对所有的非结构化数据文件，都会经过以下三个基本流程：

1. 首先会经过格式的校验，检查文件名中标记的文件类型与实际类型是否匹配。
2. 其次会进行数据去重，针对文件进行 MD5 的校验，以去除相应的重复数据。
3. 然后，再将数据进行格式的归一化，文档类及网页类数据统一转换成 pdf 格式，图片类数据统一转换成 jpg 格式，音频转换成 wav 格式，视频转换成 mp4 格式，以对后续的流程进行统一管理。

这里的数据预处理工作都可以通过 MatrixPipeline 中预制的数据预处理模块而完成，同时也支持用户自己编写代码，将自定义的数据预处理脚本打包成服务注册到 MatrixPipeline 中进行执行，以输出相应的结果。

文档数据解析

在各个类型的数据都被统一成相应的格式后，针对每一类型数据都有相应的解析模块。针对被统一成 pdf 的文档数据，其将经历以下的解析流程，以尽量多的从中提取出有效数据及元信息：

1. **PDF 版式与信息块识别**：对输入 PDF 文档进行布局解析，识别并分块提取图片、表格、图表和文本等信息块。
2. **文本数据解析**：提取文本的元数据和原始内容，并将文本按照用户设定的逻辑或系统指定的逻辑进行切片，后续可针对切片进行向量化 Embedding。
3. **图片数据解析**：提取 PDF 中的图片内容，对其同时进行视觉模型反推、OCR 文本提取，后续再将图片本身进行向量化 Embedding。
4. **表格数据解析**：利用表格识别算法提取表格中的结构化数据，并通过元数据描述其版式，支持复杂嵌套表格的递归解析。
5. **手动调整与优化**：同时支持用户对自动化解析结果进行手动调整，优化分块、标注和元数据内容，提升解析质量。

多媒体数据解析

而面对图片、音频、视频等多媒体类型，其将经历额外的数据预处理后，沿用或增加部分数据解析流程，以形成更有效的结构化数据：

1. 面对 JPG 图片数据，将直接复用文档类数据中提取到图片后的相关解析流程。
2. 面对 WAV 音频数据，则先采用 ASR 使其变成文本，再同时将音频数据及文本数据同时进行 Embedding 向量化。
3. 面对 MP4 视频数据，将先其拆成语音及视频数据，语音数据复用上一步流程，而视频将采用差分抽帧后再走图片的解析流程。

● 产品能力

在数据预处理与解析环节，以下产品能力对上述技术流程可以形成强有力的支撑：

MatrixPipeline 数据管道能力

- 提供自动化的数据管道能力，支持数据预处理模块的配置与执行。
- 内置丰富的预处理模板，如数据格式校验、去重、归一化，用户可扩展自定义功能。
- 通过可视化操作界面简化管道设计，支持大规模数据的并行处理与调度。

MatrixOne 多模态数据统一建模

- 支持多模态数据的统一存储及建模，包括元数据、解析数据及 Embedding 数据。
- 提供动态分区和分布式存储能力，保障数据存取的高效性和一致性。
- 内置强大的查询能力，可快速定位解析结果，为后续分析和建模提供支持。

MatrixGenesis 模型服务及 AI 数据解析

- 提供智能解析模块，支持 PDF 布局分析、OCR、ASR 等多模态数据解析功能。
- 集成大模型能力，用于图像反推、文本语义提取及多模态特征生成。
- 通过分布式计算及 GPU 加速并行计算支持大规模解析任务的高效执行。

特征工程

● 环节概述

特征工程是将从数据转化为模型可用的特征表示的关键环节，在 AI 模型训练与推理中起着核心作用。一个高效的特征工程流程不仅需要支持特征的生成与管理，还需要解决训练和推理特征一致性的问题，确保模型在生产环境中的稳定性和准确性。上一环节已经从多模态数据中解析出详细的各种格式的内容，而本环节将进一步根据需要训练及推理模型的特点从中提取相关数据特征，形成特征库 Feature Store。

MatrixOne Intelligence 通过提供强大的 Feature Store 能力，构建统一的特征管理平台，实现特征的生成、存储、共享和复用。Feature Store 在训练和推理流程中扮演了桥梁的角色，通过统一特征存储与访问机制，确保训练和推理使用的数据一致性，大

幅提升 AI 应用的开发与运营效率。

● 技术流程

特征生成

- **特征提取**：从结构化和非结构化数据中提取关键特征，例如文本向量、图片的视觉向量、音频的频谱特征等。

特征加工与派生

- **上下文增强**：对于语义 Embedding，结合上下文窗口（sliding window）策略生成片段级、文档级语义特征。使用链式提示（Chain-of-Thought Prompting）生成逻辑增强特征。
- **对齐与正则化**：在多模态场景中，采用对齐损失（Contrastive Loss）优化不同模态之间的嵌入表示。标准化特征范围以适配不同模型的输入需求。
- **特征分层**：针对多任务场景，生成任务专用特征（如分类任务的标签增强特征，生成任务的提示优化特征）。

特征存储与版本管理

- **嵌入向量存储**：将文本、图像和多模态生成的 Embedding 向量统一存储到向量数据库，便于高效检索和相似度计算。
- **元数据存储**：保存特征生成的上下文信息（如模型版本、生成时间、输入特性），便于追溯和分析。
- **版本控制**：为生成的特征分配版本号，确保模型训练与推理使用一致的特征版本。

特征优化与选择

- **语义优化**：对 Embedding 特征进行去噪处理，例如通过降维技术（如 PCA）或稀疏化处理减少无效信息。
- **对抗性增强**：利用对抗性样本（adversarial samples）生成更加鲁棒的特征，以增强模型对异常输入的适应能力。
- **任务相关性评估**：通过特征重要性分析（如 SHAP 值）评估特征对具体任务的贡献，优化特征集合。

特征验证与服务化

- **验证一致性**：检查训练阶段与推理阶段的特征一致性，确保生产环境的稳定性。
- **实时服务化**：将生成的 Embedding 特征提供为实时服务，支持在线推理和相似度检索需求。
- **跨场景复用**：支持特征跨任务、跨场景复用（如通用语义 Embedding 在搜索、对话和推荐场景中的共享）。

● 产品能力

在特征工程环节，MatrixOne Intelligence 提供了强大的产品能力支持，涵盖多模态存储、版本管理、在线服务和 Embedding 特征生成，具体包括：

MatrixOne 多模态存储及版本管理

- **多模态支持**：MatrixOne 数据库能够统一存储来自文本、图像、音频和视频的 Embedding 向量及相关元数据，支持多模态特征的高效管理。
- **高并发与低延迟**：MatrixOne 支持 OLTP 的负载，同时分布式架构支持大规模特征存取和高并发在线服务，满足实时推理和相似度检索的需求。
- **动态版本控制**：MatrixOne 提供快照机制，自动记录特征生成的版本状态，确保训练与推理使用一致的特征数据版本。
- **回滚与追溯能力**：支持对特定版本特征的回滚操作，方便模型问题排查和历史重现。

MatrixGenesis 的 Embedding 支持

- **预训练模型支持**：MatrixGenesis 内置强大的预训练模型（如 BERT、CLIP、Wav2Vec2.0 等），支持文本、图像和音频的语义 Embedding 生成。
- **模型可扩展性**：支持用户加载自定义的 Embedding 模型，以满足不同业务场景的特定需求。

数据标注与增强

● 环节概述

数据标注与增强环节是在原始数据解析的基础上，针对特定训练任务和模型需求进行数据的进一步加工、治理和生成，以构建高质量的训练集、验证集和测试集。这一环节旨

在满足多样化模型（如大语言模型 LLM、文生图模型、视频理解模型等）的精调需求，同时确保数据集的格式与内容符合训练要求，并具备灵活的分类、管理和更新能力。

数据标注与增强的核心输入是经过解析和初步特征化的数据，输出为特定任务定制的标注数据集，并支持进一步的数据增强与分类操作。通过这一环节，企业能够快速构建满足精调需求的数据集，为高质量模型训练和评估提供保障。

● 技术流程

数据标注与训练集生成

- **面向大语言模型（LLM）精调**：从数据库中召回相关语料和知识数据。利用预训练大模型生成初步的 input-output 数据对。对生成结果进行人工审核和优化，确保数据质量和格式一致性，最终生成适合 SFT、LoRA 等方法的 input-output 格式数据集。
- **面向文生图模型精调（如 Stable Diffusion）**： 提取原始图片数据并生成关键词形式的描述。通过反推模型生成初步描述，或结合人工标注优化文本内容。将图文对组合为 input 文-output 图格式，符合精调要求。
- **面向图像反推模型精调**：提取图片内容并生成描述文本，强调语义和细节关联。优化描述内容以满足反推模型的文本生成要求，以生成以图文对为核心的 input 图-output 文数据集。
- **视频理解模型精调**：根据语义对原始视频进行切片和整合，提取关键帧作为图片。为每段视频生成文本描述（如场景描述、行为描述），形成 input 图-output 文格式。

数据增强

- **文本数据增强**：调用大模型生成同义替换版本的句子或段落，保持语义一致性。补充语境信息，例如在对话式训练集中生成更复杂的上下文链条。
- **图文数据增强**：为原始图片生成多版本文本描述（关键词、短语、长段文字）。调用扩展模型生成与原始图像风格相似的新图片，配合对应描述文本，扩展训练集规模。
- **视频数据增强**：在现有视频切片的基础上，调用大模型生成新的语义描述。增加视频帧的多种切片组合，扩展小段视频样本的数量和多样性。

数据分类与版本管理

- **数据集划分**：将增强后的数据集按照任务需求划分为训练集、验证集和测试集（如 70%-20%-10%）。采用随机分组（如 ID 哈希、时间戳分组）或特征驱动分组，确保数据分布均衡。
- **版本管理**：记录每次数据集更新的版本，确保数据的可追溯性和一致性。在多次模型训

练和评估后，选择最佳版本的训练集用于大规模精调。

● 产品能力

MatrixOne 快照与分组

- 提供快照和版本管理功能，确保数据集的更新可追溯性和一致性。
- 高效分组能力支持大规模随机数据划分操作。

MatrixGenesis 大模型服务

- 提供各类大模型托管服务能力，用于生成图文对、文本描述和语义增强内容。
- 支持结合人工审核优化生成结果，提升数据标注效率与质量。

MatrixPipeline 数据 pipeline 任务

- 提供自动化数据增强与分类工具，通过可配置流程完成大规模数据的治理和分组。
- 支持用户自定义数据处理逻辑，灵活应对不同任务需求。

模型训练与评估

● 环节概述

模型训练与评估是 GenAI 落地过程中的核心环节，旨在通过对高质量数据集的训练构建符合业务需求的 AI 模型，并通过科学的评估方法验证模型性能，确保其在实际场景中的可用性和稳定性。在 GenAI（如大语言模型 LLM、文生图模型、视频理解模型）应用中，训练往往需要处理大规模数据，涉及深度模型参数优化、分布式计算以及高效资源管理。

本环节的核心输入是数据标注与增强阶段生成的训练集、验证集和测试集，以及预训练模型或基础模型（如 Qwen、Stable Diffusion）。核心输出是经过精调的任务专用模型和全面的评估指标结果，用以指导模型的上线与优化。

● 技术流程

训练准备

- **数据加载与预处理**：从数据库中加载训练集、验证集及测试集，按批次组织输入数据。结合训练任务的需求 (如多模态对齐、序列预测等) 对数据进行实时预处理 (如归一化、分词、补全)。
- **资源调度**：分配计算资源（如 GPU 集群）并配置分布式计算环境。动态调度存储、计算和通信资源，确保资源利用率最大化。

模型精调

- **精调方法**：
  - 全参数精调：针对高优先级任务，通过全参数训练调整模型。
  - 增量训练：针对小规模数据集，采用 LoRA 或参数高效微调方法提升效率。
  - Prompt 调优：基于任务设计 Prompt 模板，通过优化输入模式提升生成效果。
- **优化过程**：使用分布式梯度下降（如 AdamW、LAMB）优化模型参数。
  - 结合混合精度训练（FP16/FP32）提升训练效率，减少显存占用。

模型评估

- **验证过程**：在每轮训练后，使用验证集评估模型性能 (如损失值、准确率、F1 分数等)。
  - 针对生成任务，采用 BLEU、ROUGE 等语言生成指标，或 FID、CLIP Score 等视觉生成指标。
- **测试与分析**：在测试集上评估模型的通用性能和泛化能力。
  - 对比不同模型版本，选取性能最优的版本进行部署。

迭代优化

- **模型调参**：分析训练过程中的超参数对性能的影响，优化学习率、批次大小、正则化参数等。
- **数据回馈**：结合评估结果，分析错误案例，对训练数据集进行优化与增强。

模型保存与版本管理

- **模型存储**：将精调后的模型及相关元数据 (如超参数配置、评估结果) 存储到数据库中。
- **版本控制**：对每次训练生成的模型进行版本化管理，支持回滚、比较和复用历史模型版本。

● 产品能力

MatrixOne 融合存储及快照

- **高效数据加载**：支持从多模态数据存储中快速加载训练数据，提供高性能分布式查询和批量数据预处理能力。
- **模型与元数据存储**：统一存储精调后的模型及其相关元数据，支持版本化管理和快速检索。
- **快照功能**：记录训练数据及模型的状态，确保训练与评估过程的可追溯性和复现性。

MatrixGenesis 训练工具箱

- **预训练模型支持**：内置丰富的预训练模型（如 Qwen、Stable Diffusion），支持多模态和任务特定的模型精调。
- **高效优化框架**：提供分布式训练框架，支持大规模模型的高效训练与微调。
- **多模态评估**：内置针对文本、图像、视频等多模态任务的多样化评估工具。

RAG 召回与搜索

● 环节概述

RAG（Retrieval-Augmented Generation）是 GenAI 的一种关键技术，通过将知识检索与生成模型相结合，使大模型能够在推理过程中动态调用外部知识库，提升生成内容的准确性和可控性。

在 RAG 召回与搜索环节，系统的核心目标是从海量数据中快速检索与用户查询相关的高质量内容，并将其作为上下文提供给生成模型，以增强模型的生成效果。这一过程既包括传统的全文检索（如基于关键字的 BM25 算法），也包括语义级别的向量检索（如基于 Embedding 的语义匹配）。最终，结合多模态数据的搜索与检索优化，RAG 系统能够满足从文本、图片到音视频的多样化应用需求。

该环节与应用侧的交互息息相关，用户将使用自然语言及多模态数据进行输入查询，而系统将负责从用户自有数据中召回最相关的数据，并返回给用户。

● 技术流程

多模态数据索引构建

- **数据预处理**：对存储库中的多模态数据（文本、图片、视频等）进行预处理和标准化。
- **索引类型**：通过 BM25 等传统方法为结构化和文本数据构建关键字索引。利用 Embedding 模型生成向量化表示，并通过 FAISS、ScaNN 等工具构建高效向量索引。
- **多模态支持**：针对图像、视频等非文本数据，生成语义嵌入向量，支持跨模态检索。

检索与召回

- **检索策略**：单路检索使用关键字匹配或语义匹配完成单一通道检索。多路召回结合全文检索和语义检索的结果，通过混合排序优化召回性能。
- **多模态检索**：支持跨模态查询，例如文本查询图片，图片查询视频内容，或者结合多种输入类型完成混合查询。
- **动态更新**：对新增数据或实时变化数据动态更新索引，确保召回结果的时效性。

候选上下文生成与排序

- **初筛阶段**：快速召回候选内容，依据检索算法生成初步相关性排名。
- **精排阶段**：结合多模态特征及上下文一致性对候选结果重新排序，确保与用户查询的语义和任务目标高度一致。
- **多模态融合**：根据检索结果类型（文本、图片、视频等）融合不同模态内容生成最终上下文。

上下文交付与模型增强

- **上下文拼接**：将召回的内容整合为输入格式（如文本段落或嵌入向量），提供给生成模型。
- **反馈优化**：结合用户反馈数据优化召回与排序策略，提升模型推理的精准性和相关性。

● 产品能力

MatrixOne 数据库的检索能力

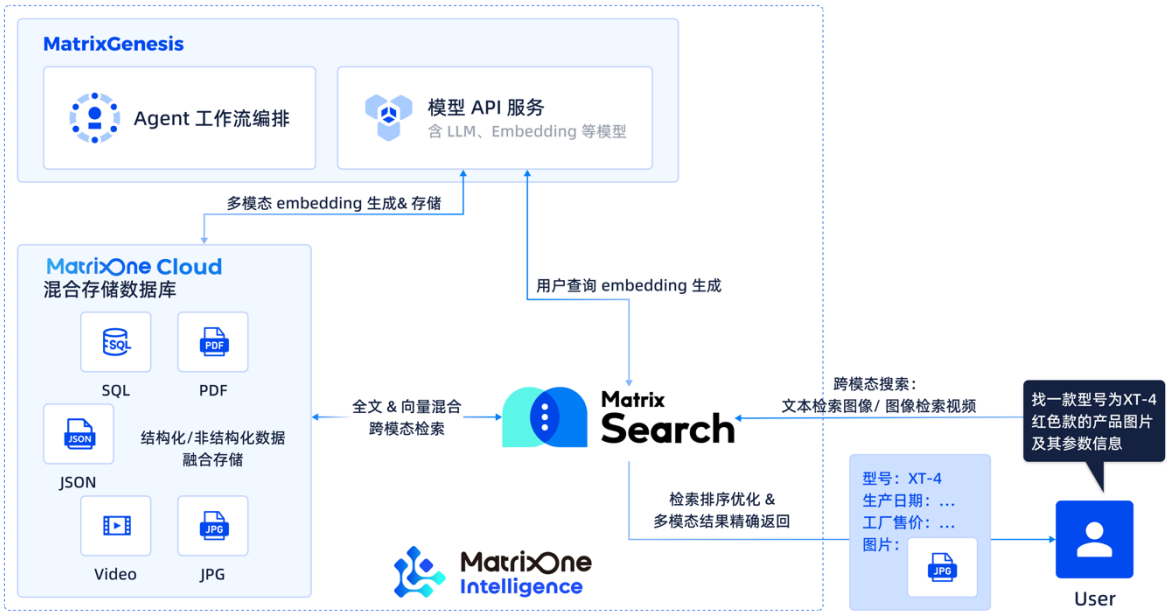
- **统一数据存储**：支持结构化和非结构化数据的融合存储，方便多模态索引与检索。
- **全文与向量检索结合**：内置全文检索和向量检索能力，支持混合查询与动态召回。
- **高效索引管理**：提供分布式索引构建与更新机制，确保大规模数据的高效检索性能。

MatrixGenesis 的模型支持能力

- **Embedding 模型支持**：内置语义嵌入模型（如 BERT、CLIP），针对文本、图像、音视频生成高质量向量化表示。
- **多模态支持**：支持跨文本、图像、视频的嵌入生成与模态对齐，为语义检索提供底层支持。

MatrixSearch 的多模态检索能力

- **多模态语义检索**：结合语义和全文检索，支持文本、图像、音频、视频等多模态数据的统一查询。
- **跨模态查询支持**：实现文本检索图像、图像检索视频等复杂查询，满足多样化业务场景需求。
- **分布式扩展与高并发性能**：支持大规模检索场景，确保高并发和低延迟的查询响应。



MatrixOne + MatrixGenesis + MatrixSearch 实现精准多模态搜索

总结

通过对技术流程的逐步拆解，MatrixOne Intelligence 全面展示了从数据接入到智能应用的完整闭环。针对企业在 GenAI 落地过程中面临的数据分散、异构复杂、规模化处理及智能化应用等挑战，方案提供了统一的数据接入与整合、高效的预处理与解析、多模态特征工程、精准的数据标注与增强，以及强大的模型训练与评估能力。通过 RAG 召回与搜索环节的优化，进一步提升了大模型推理阶段的准确性和业务适配性。借助 MatrixOne 数据库、MatrixPipeline、MatrixGenesis、MatrixSearch 等核心产品，方案实现了数据治理、存储、计算与智能模型能力的无缝协作。整体流程以模块化、自动化、高性能为设计原则，为企业构建了一套面向 GenAI 应用的高效数据智能平台，加速 AI 应用的开发与落地，为企业充分释放多模态数据价值提供了有力保障。

行业案例

极视角多模态数据与特征平台

● 客户背景

极视角是一家专注于计算机视觉算法研发的企业，其业务场景覆盖工业检测、智慧零售、智慧城市等多个领域。在企业的发展过程中，极视角面临着 AI 算法开发效率低下的挑战，尤其是在多模态数据的管理与使用上存在严重的痛点，包括数据分散、管理混乱、特征开发效率低等问题。为了提升 AI 算法的开发效率，极视角希望构建一套完整的多模态数据与特征平台，以支持大规模数据的高效管理、加工和复用。

● 解决方案

通过引入 MatrixOne Intelligence，极视角搭建了一套覆盖数据接入、解析、特征工程、存储与建模的端到端多模态数据与特征管理平台。具体实施流程包括以下几方面：

1. **数据接入与整合**：极视角将分散在不同存储系统（如本地文件系统、云存储等）的多模态数据统一接入到 MatrixOne 数据库中，涵盖图像、视频等核心数据类型。通过 MatrixPipeline 的自动化管道能力，实现对数据的批量归档、去重及格式规范化处理，确保数据的一致性与可管理性。
2. **数据解析与特征化**：针对海量图像和视频数据，利用 MatrixGenesis 的智能解析能力，从数据中提取语义标签、对象特征及嵌入向量，并将这些解析结果存储到 MatrixOne 数据库中。多模态特征被统一管理并版本化，极大提升了特征的可追溯性和复用性。
3. **特征工程与共享**：借助 MatrixOne Intelligence 的 Feature Store 能力，极视角实现了特征的集中化管理和分布式存储。通过统一的特征生成、优化与复用机制，不同团队可以快速调用已有特征，避免重复开发，显著提升了 AI 算法迭代速度。
4. **存储与建模支持**：MatrixOne 数据库支持高并发和低延迟的分布式存储，确保算法开发过程中对多模态数据与特征的高效访问。同时，特征的版本化管理能力为模型训练提供了稳定的数据基础，确保训练和推理阶段的数据一致性。

● 客户收益

通过构建基于 MatrixOne Intelligence 的多模态数据与特征平台，极视角实现了 AI 算法开发效率的大幅提升：数据接入效率提高了 60%，多模态数据的整合与管理更加规范

化。特征复用率提升了 70%，避免了因重复开发特征而浪费的资源。算法迭代周期从平均两周缩短到一周以内，产品开发效率显著提高。基于平台的稳定支持，极视角能够更高效地响应客户需求，加速算法落地场景的拓展。

深智城集团

● 客户背景

深智城集团是深圳智慧城市科技发展领域的重要参与者，其智慧交通系统需要处理来自人、车、道路、环境等多源异构数据的实时分析和决策需求。然而，传统的数据库系统在面对海量数据高频写入、实时分析、数据一致性及多模态数据管理等方面表现出明显的性能瓶颈，无法满足智慧交通场景的高效运行需求。此外，系统组件复杂、管理成本高，与云原生技术的兼容性不足，也进一步限制了智慧交通系统的扩展性和灵活性。

● 解决方案

深智城集团通过引入 MatrixOne Intelligence，基于超融合数据库 MatrixOne 的能力，对其交通大数据平台进行全面升级，打造了高性能、高效能的智慧交通数据基础设施。

1. 数据接入与整合

深智城通过 MatrixPipeline 将交通系统中多源数据（如传感器数据、视频监控数据、车辆轨迹数据等）接入至 MatrixOne 数据库中。通过标准化接入与预处理，实现了结构化与非结构化数据的统一管理，为后续数据分析和实时处理奠定了基础。

2. 实时分析与存储优化

a) **超融合架构**：MatrixOne 数据库将事务与分析能力融合在一个平台中，无需分离的 OLTP 与 OLAP 系统，大幅提升了实时查询与分析效率。

b) **实时性支持**：支持每小时处理 TB 级数据的高频写入与实时分析需求，实现了秒级响应时间。

c) **表结构动态变更**：通过在线表结构变更功能，为交通场景的多变业务需求提供了灵活支持，避免了因表结构调整而造成的系统中断。

3. 云原生兼容与弹性扩展

a) MatrixOne 深度兼容 Kubernetes 技术，通过容器编排实现了数据层与基础设施层的无缝整合：动态调度与弹性扩缩容能力，有效优化了资源利用率，降低了硬件开销。

b) 简化了数据库的部署和管理流程，提升了智慧交通系统的可扩展性和运维效率。

4. **架构优化与组件整合**
   a) **组件简化**：将原有的 5 个独立数据组件整合至 MatrixOne，减少了 80%的组件数量，大幅降低了架构复杂性。
   b) **一致性管理**：通过分布式事务和数据一致性支持，确保了交通大数据平台在多节点高并发场景下的稳定性。

● 客户收益

- 通过基于 MatrixOne Intelligence 的交通大数据平台改造，深智城集团实现了显著的技术和管理收益：数据组件数量减少 80%，系统架构更加简化，管理效率显著提升。
- 实现了实时性业务支持，每小时处理 TB 级数据，秒级响应时间满足智慧交通场景需求。
- 运维成本降低约 50%，基础设施资源利用率显著提升。
- 云原生兼容性增强，系统可扩展性和弹性部署能力大幅提高，为未来业务发展提供了良好基础。

江西铜业

● 客户背景

江西铜业是全球领先的铜生产企业，转炉作业是其核心生产环节之一。然而，转炉作业涉及复杂的工业流程，产生了大量的 IoT 数据（如传感器采集的温度、压力、气体浓度等）和多模态数据（如现场视频、设备运行日志等）。这些数据分散在不同的系统中，缺乏统一的管理和处理能力，导致难以有效利用数据进行智能化决策。江西铜业亟需构建一套整合 IoT 和多模态数据的智慧作业平台，以实现精准监控和高效作业优化。

● 解决方案

通过引入 MatrixOne Intelligence，江西铜业成功搭建了一套覆盖数据接入、解析、分析和智能推理的端到端智慧作业平台，为转炉生产的智能化提供了有力支持。

1. **数据接入与整合**：借助 MatrixPipeline，江西铜业将来自 IoT 设备的实时数据（如传感器数据）和多模态数据（如转炉运行视频）统一接入平台。通过边缘计算节点对高频 IoT 数据进行预处理（如压缩、清洗），并将处理后的数据与视频流数据上传到 MatrixOne 数据库，实现了多源数据的实时整合。

2. **数据解析与特征提取**：利用 MatrixGenesis 的智能解析能力，从多模态数据中提取关键特征。从 IoT 数据中提取时间序列特征，如温度波动趋势、压力异常点等，为生产监控提供关键指标。从视频数据中通过视频分析技术提取转炉作业过程中的关键帧，并结合算法识别设备显示屏上的参数信息，为作业优化提供数据支撑。
3. **实时监控与建模分析**：通过 MatrixOne 数据库实现多模态数据的统一存储与高效检索，支持转炉状态的实时监控和异常预警。基于历史数据和实时特征，构建机器学习模型预测转炉操作参数（如最佳切换时间），优化生产效率并降低能耗。
4. **智能推理与决策支持**：基于 RAG 技术，整合历史数据与实时数据，为作业员提供动态决策支持（如炉内状态推荐操作）。多模态搜索功能帮助生产团队快速定位异常视频片段及相关 IoT 参数，为问题排查和优化策略提供依据。

● 客户收益

通过构建基于 MatrixOne Intelligence 的智慧作业平台，江西铜业在转炉生产过程中实现了显著的智能化提升：IoT 数据与多模态数据的整合效率提升了 80%，实现了生产数据的全链路可视化。通过智能模型优化，转炉作业效率提升了 30%，能耗降低了 15%。异常检测和问题定位时间缩短了 70%，大幅提升了生产问题的响应速度。智能化决策支持帮助一线作业员显著减少了操作失误，提升了产品质量的稳定性。

金意陶

● 客户背景

金意陶是一家专注于瓷砖产品研发与销售的企业，其产品种类丰富，用户在选购过程中需要快速找到符合需求的产品。然而，传统的产品检索方式（如通过关键词或型号搜索）难以满足销售与客户沟通时候的快速选型需求。金意陶希望构建一个基于图像搜索的智能平台，支持销售通过拍照、上传图片或输入文字来精准找到相关产品，同时查询库存信息，从而提升客户体验并优化销售效率。

● 解决方案

基于 MatrixOne Intelligence，金意陶搭建了以 MatrixSearch 为核心的智能搜索平台，实现了从图片管理到搜索图片结果的完整闭环，具体包括以下能力：

1. **数据接入与整合**
   a) MatrixSearch 从后台管理系统接入产品图片和库存数据，并通过自动更新和 API 接口实现数据的同步和实时更新。
2. **搜索索引构建与优化**
   a) MatrixSearch 利用 EfficientNet 模型对上传的图片进行特征提取，生成高精度图像嵌入向量，构建图像检索索引。
   b) 支持混合检索，通过结合语义检索（以文搜图）和向量检索（以图搜图），提升搜索准确性和结果的相关性。
3. **智能搜索功能实现**
   a) 用户通过小程序入口上传图片或输入文字进行产品搜索。
   b) 系统调用 MatrixSearch 的搜索 API 快速返回匹配的产品结果，同时支持按分类筛选和系列查询，帮助用户快速找到目标产品。
4. **库存查询与展示**
   a) 对搜索结果进行后台过滤分类后，小程序展示匹配的产品信息，包括名称、规格、图片、库存等。
   b) 用户可直接查询产品库存情况，并进一步查看同系列其他产品，优化搜索体验。

● 客户收益

- 通过基于 MatrixSearch 的智能搜索平台，金意陶在产品检索和客户体验方面实现了显著提升：搜索效率提升 90%，销售能够快速找到符合需求的瓷砖产品。
- 系统化的库存查询功能帮助销售团队优化库存管理，减少人工操作时间。
- 基于图片特征的智能搜索功能显著提升了用户对产品选择的满意度，增强了品牌黏性。
- 灵活的小程序入口简化了用户交互流程，为客户提供了随时随地的高效服务。

素问 TechAgent

● 客户背景

素问 TechAgent 是一家专注于产业链舆情数据服务的企业，主要为制造业龙头企业及政府机构提供基于企业基本信息、研报、财报、专利及新闻等多模态数据的舆情分析。

然而，随着业务规模的增长，TechAgent 原有的数据架构面临以下挑战：

●多种数据存储和处理工具（如 MySQL、MongoDB、ElasticSearch、Faiss、ClickHouse）的使用增加了架构复杂度和运维难度。

●数据从采集、存储到处理需跨多个系统完成，效率低下，且需要大量人工干预。

●私有化交付时，数据库部署和调试周期冗长，影响整体交付效率。

TechAgent 迫切需要一套简化数据架构、提升处理效率并支持 GenAI 应用的智能数据平台。

● 解决方案

基于 MatrixOne Intelligence，TechAgent 构建了一套支持多模态数据存储、智能检索和实时分析的 AIGC 平台，大幅简化了数据架构并提升了平台效能。

1. 数据接入与整合

a) 通过 MatrixPipeline 自动化数据管道，将来自网络爬虫、API 接口及文件提取的多模态数据统一接入系统。

b) 数据在接入时自动完成格式规范化和冗余去除，确保一致性和高效性。

c) 支持多模态数据源（如文本、JSON、图片等）的批量处理和动态更新，显著减少人工干预。

2. 智能解析与特征提取

a) 借助 MatrixGenesis 的智能解析能力，从复杂数据中提取语义特征和结构化信息：

i. **文本数据**：利用预训练语言模型生成嵌入向量和语义标签。

ii. **JSON 数据**：解析嵌套数据结构并提取关键字段，简化后续检索和分析。

iii. **图片与文档**：通过 OCR 和视觉模型提取内容特征，实现跨模态关联分析。

b) 解析后的数据存储至 MatrixOne 数据库中，形成统一的知识存储库，支持后续检索和推理。

3. 检索与搜索优化

a) 借助 MatrixSearch 的能力，TechAgent 实现了全文检索与语义向量检索的混合模式，满足复杂场景需求。

b) 通过倒排索引与向量检索相结合，支持基于语义和关键字的混合搜索，提升了检索的精度与相关性。

4. 云原生部署与扩展

a) MatrixOne 基于 Kubernetes 实现全云原生化设计，支持容器化部署和动态扩缩容。

b) 支持负载隔离，为特定任务分配专属资源，保障高优先级任务的性能和安全性。

5. **实时处理与分析**

a) MatrixOne 的 HTAP（混合事务与分析处理）架构同时支持 OLTP 和 OLAP 负载，无需在 MySQL 与 ClickHouse 间进行 ETL 操作，大幅提升了实时分析效率。

b) 在生成报告或进行复杂数据分析时，数据处理从小时级缩短至分钟级。

● 客户收益

通过基于 MatrixOne Intelligence 的智能平台，TechAgent 在数据管理和业务交付方面实现了显著的提升：

- **数据接入效率提升**：借助 MatrixPipeline 自动化流程，接入和规范化数据的效率提升了 60%。
- **解析效率提高**：MatrixGenesis 智能解析减少了人工标注工作量，使数据预处理速度提升了 2 倍。
- **数据架构简化**：将原本多工具组合整合为单一数据库系统，减少 80%的运维复杂度。
- **数据处理效率显著提升**：数据处理效率从小时级缩短到分钟级。
- **私有化交付周期缩短**：私有化交付周期从 2 个月缩短至 1 周，显著提升交付效率。
- **精准检索与智能推理**：统一的检索平台支持多模态数据的语义搜索和快速查询，为客户提供更精准的舆情分析和智能化服务。

MatrixOne Intelligence 帮助 TechAgent 实现了从数据整合到智能解析的全流程优化，不仅解决了复杂数据架构和低效运维的难题，还为 GenAI 应用场景提供了坚实的技术基础，加速了业务的规模化拓展和落地。

总结

在 GenAI 快速发展的浪潮中，多模态数据已成为推动企业智能化升级的核心动力。本白皮书详细探讨了 MatrixOne Intelligence AI 原生多模态数据智能解决方案，通过对数据全生命周期的整合与优化，从数据接入与治理、预处理与解析，到特征工程、模型训练与评估，再到召回与搜索，为企业构建了一套全面、统一、高效的数据智能平台。通过这种以数据为中心的设计理念，MatrixOne Intelligence 帮助客户实现“Your Data for Your AI”的承诺，让企业自有数据成为 GenAI 应用的坚实基础和独特竞争力的来源。MatrixOne Intelligence 期待与您共同迈向数据智能与 GenAI 深度融合的未来！

矩阵起源 Matrix origin

Store Anywhere
Compute Anywhere
Innovate Anywhere

Web.
www.matrixorigin.cn

E-mail.
contact@matrixorigin.cn