TRANSWARP 星 环 科 技

股票代码 688031

Transwarp Hippo

星环分布式向量数据库

星环信息科技（上海）股份有限公司

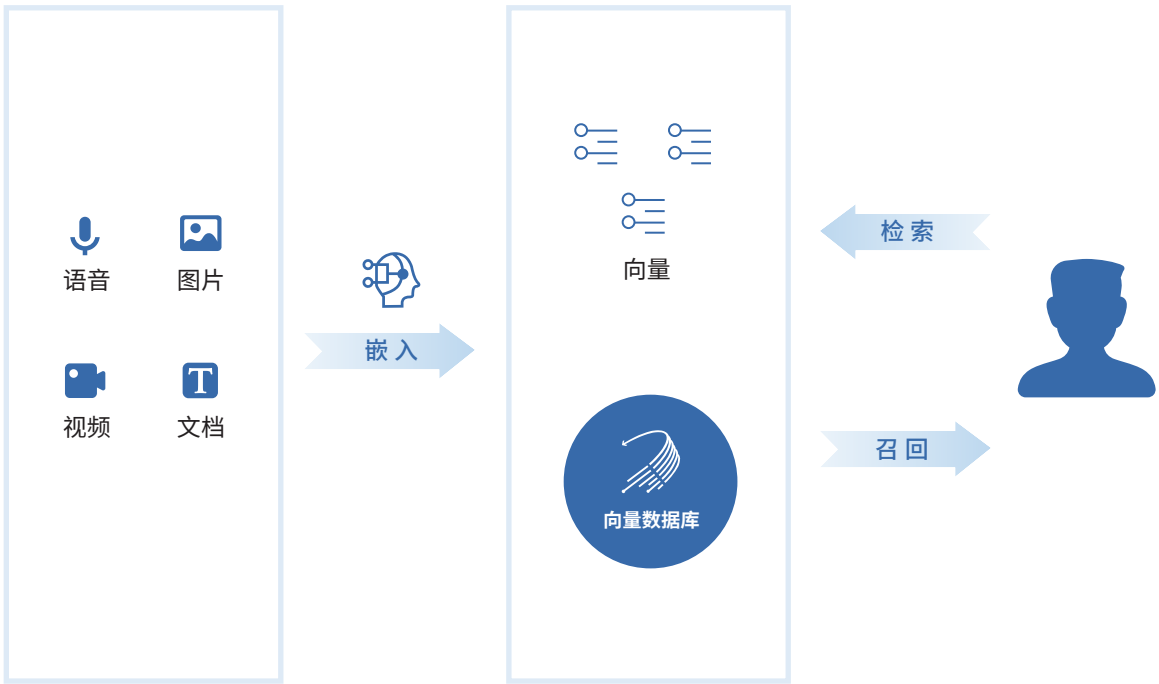
星环分布式向量数据库

产品介绍

伴随着企业对海量非结构化数据管理的需求的不断加深，以及深度学习在工业界的广泛落地，向量数据在实际应用场景下的数据量级开始直线增加。想要高效处理这些海量的向量数据，就需要更细分、更专业的数据基础设施，为向量构建专门的数据库处理系统。

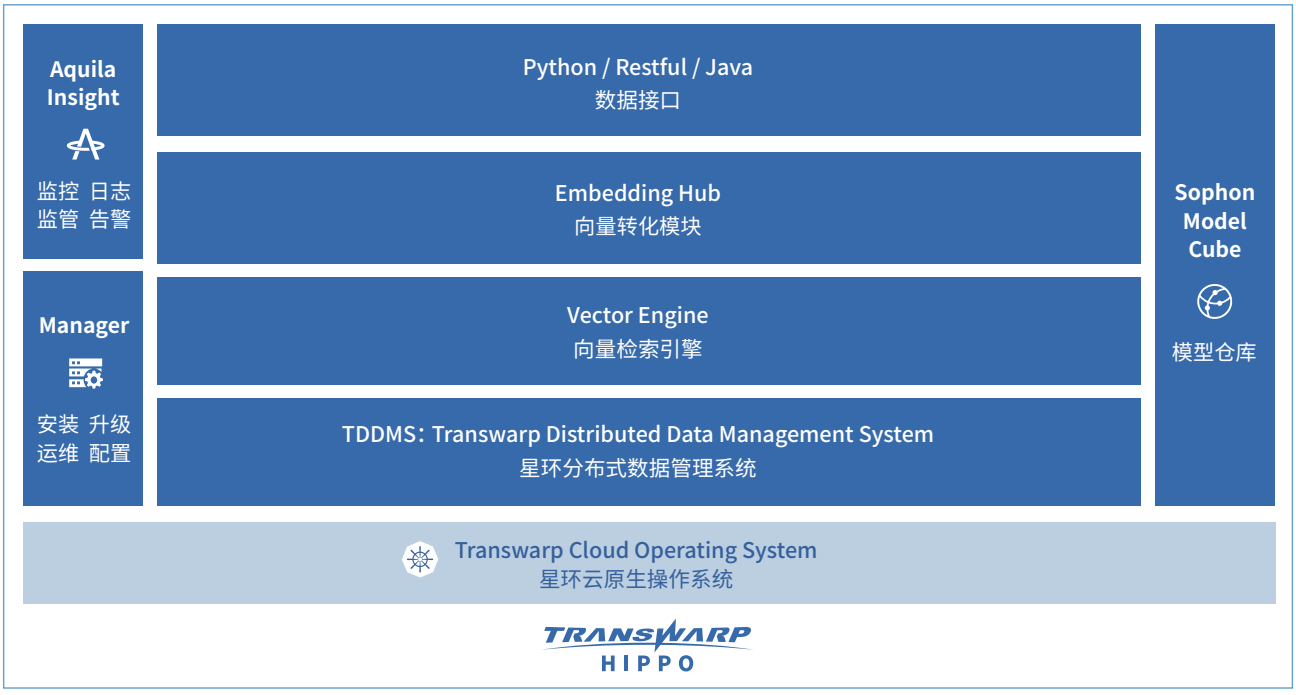
Transwarp Hippo 是星环科技自主可控的一款企业级云原生分布式向量数据库，支持存储，索引以及管理来自深度神经网络或者各类机器学习模型所生成的海量向量数据，能够高效的解决向量相似度检索以及高密度向量聚类等问题。Transwarp Hippo 具备高可用，高性能，易拓展等特点，支持多种向量搜索索引，支持数据分区分片，数据持久化，增量数据摄取，向量标量字段过滤混合查询等功能，能够很好的满足企业针对海量向量数据的高实时性查询、检索、召回等场景。

同时，Transwarp Hippo 也可以高效的服务于大模型，有效地解决大模型在知识时效性低、输入能力有限、准确度低等问题，通过将最新资料、专业知识、个人习惯等海量信息向量存储在 Hippo 中，可以极大地拓展大模型的应用边界，让大模型保持信息实时性，并能够动态调整，使大模型拥有“长期记忆”，在一定程度上解决“AI 幻觉”的问题。



核心组件

产品架构图



组件介绍

TDDMS

自主研发的分布式数据管理系统，采用 shared-nothing 架构，通过多副本机制实现数据服务高可用，使用 raft 协议保证副本之间的数据一致性；支持弹性扩缩容、自动故障恢复、权限控制、多租户与冷热数据分层存储等功能。

Vector Engine

自主研发的向量搜索引擎，支持海量向量数据的检索，同时在计算框架上进行了大幅度优化，具备高准确性与高性能的相似检索能力。

Embedding Hub

Hippo 内置的向量转化工具，提供标准化接口连通各类大模型并实现数据的向量嵌入。

大模型有哪些？

向量嵌入？

Model Cube

模型仓库，一站式完成模型生命周期中的模型上架、模型评估和模型部署，大幅提高模型的可维护性和可操作性；支持管理多源异构模型，支持镜像模型、文件模型和组合模型，并提供模型评估和模型体验功能，最大化模型价值。

产品特性与优势

产品特性

云原生系统

采用全面容器化部署，支持服务的弹性扩缩容；同时具备多租户和强大的资源管控能力。

分布式部署

具备分布式部署能力，有丰富的大规模集群部署经验；通过Raft算法确保数据的强一致性；提供故障迁移，数据修复等数据保障能力。

数据多模

基于多模型统一架构，支持与结构化、非结构化数据统一存储管理，实现数据跨模型联合分析，提高数据分析效率，同时避免了部署多套系统带来的架构复杂、开发运维成本高等问题。

企业级安全

可提供基于SASL的用户认证能力，以及基于SSL/TLS的数据加密传输。

产品优势

一站式

提供向量转化工具和Embedding模型，一站式完成模型上架、模型评估和模型部署，降低用户使用成本，提高数据入库效率。

高精度

多类索引支持，一库搞定向量+全文联合检索，提高大模型召回准确率；结合自研图数据库，可进一步提高大模型精度。

高性能

支持多进程架构与GPU加速，充分发挥并行检索能力，结合软硬件深度优化，充分发挥CPU多核、高内存带宽等优势，为海量、多维向量提供强劲算力。

易对接

提供标准的Python、Restful、Java API等接口，可轻松对接各类应用和模型，提高应用开发和调用的效率。

应用场景

文本检索

传统搜索引擎更偏向于词 / 句的精确查询，Hippo 通过向量引擎提供自然语言处理能力，可以更好的支持基于语义的查询分析，让查询更满足人性化的需求。同时 Hippo 支持全文和向量混合检索，大幅提高召回精度。

语音/图像/视频检索

通过机器学习分析，各类数据可以被抽象成高维向量特征，Hippo 则可以将所有特征构建成高效的向量索引，用户可以基于向量索引实现数据的相似性检索，可以覆盖各类 AI 场景，如人脸识别、语音识别、视频指纹等。

个性化推荐

Hippo 支持与各类深度学习平台搭建的模型进行耦合，分析、挖掘用户行为与喜好等多方面相关数据向量化存储，通过向量相似度检索，将用户可能感兴趣的信息推送给客户，做到千人千面的个性化推荐效果。

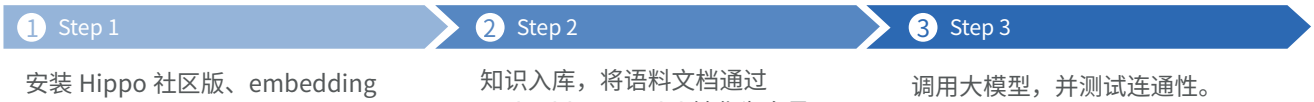
大模型应用

Hippo 在大模型的大规模应用中，同样可以发挥重要的作用。Hippo 可以作为 LLM 的中间载体承载 LLM 生成的各类内容，有效扩展 LLM 的时间与空间边界，使大模型拥有“长期记忆”，并协助解决目前企业最担忧的大模型泄露隐私问题。

社区版介绍

Hippo 社区版，低成本、快速构建大模型应用

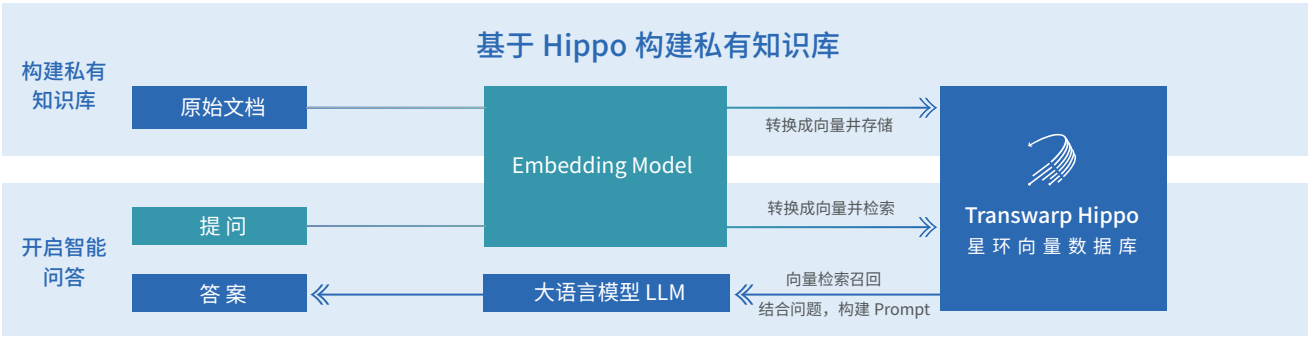
基于 Hippo 社区版快速搭建私有知识库，开始智能问答



安装 Hippo 社区版、embedding model，并验证；

知识入库，将语料文档通过 embedding model 转化为向量，存储到 Hippo 中；

调用大模型，并测试连通性。



多种版本，满足不同场景需求

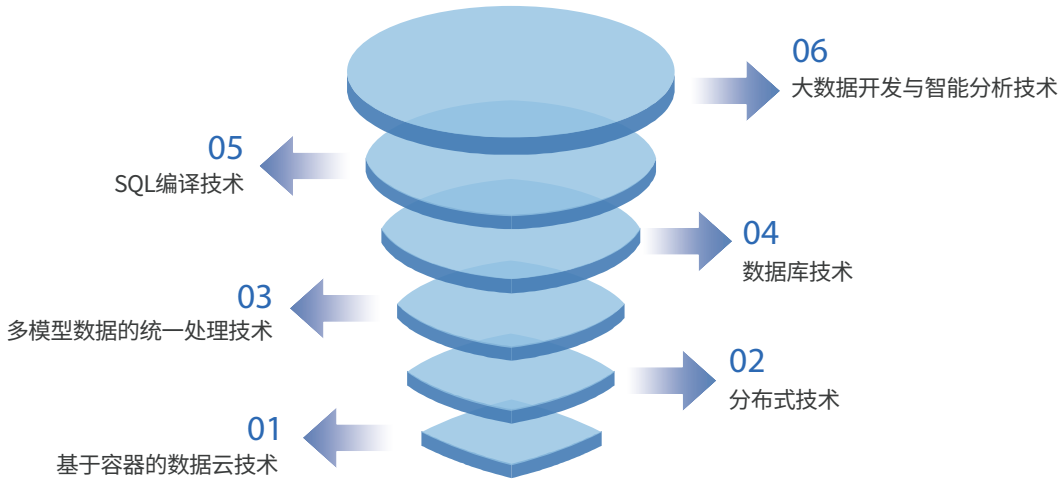
	Hippo Community 社区版	Hippo 存储密集版	Hippo Pro 高性能版
拓展性	单节点	可扩展至 500 节点	可扩展至 1500 节点
身份认证	–	SSO	
访问控制	✗	RBAC 细粒度	
容量（百万向量 / 服务器）	10M	20M	10M
性能（qps/ 服务器）	100	100	1000
容灾能力	✗	跨机房、跨城市多活部署	
跨集群热备	✗	表级跨集群实时热备	
数据备份	✗	全量备份、按时间点恢复	增量、全量备份、按时间点恢复
行级权限管控	✗	✗	✓
加密	✗	传输加密	数据加密 / 传输加密
指标监控	✓	✓	细粒度 Prometheus 指标监控
Vector similarity filter	✗	✗	✓
数据生命周期	✗	✓	✓
GPU 索引	✗	✗	✓
hybird search	✓	✓	✓
只写副本	✗	✗	✓（仅做高可靠，降低内存使用率）

星环分布式向量数据库

关于我们

星环科技（股票代码：688031）致力于打造企业级大数据基础软件，围绕数据的集成、存储、治理、建模、分析、挖掘和流通等数据全生命周期提供基础软件与服务，构建明日数据世界。公司以上海为总部，以北京、南京、广州、新加坡为区域总部，在郑州、成都、重庆、济南设有支持中心，同时在深圳、西安等地设有办事机构，并在加拿大设有海外分支机构。经过多年自主研发，星环科技建立了多个产品系列：一站式大数据基础平台TDH、分布式分析型数据库ArgoDB及交易型数据库KunDB、基于容器的智能数据云平台TDC、大数据开发工具TDS、智能分析工具Sophon和超融合大数据一体机TxData Appliance 等，并拥有多项专利技术。目前公司产品已经在十几个行业应用落地，拥有超过一千家终端用户。2016年公司成为中国首个进入Gartner数据仓库及数据管理解决方案魔力象限的厂商，且被评为最具前瞻性的远见者；2017年被IDC评为中国大数据市场领导者；2018年星环科技成为12年来全球首个完成TPC-DS测试并通过官方审计的数据库厂商；2020年再次被IDC评为中国大数据管理平台领导者。自2021年起，星环科技蝉联Gartner增强分析技术中国推荐供应商；2022年，公司入选Gartner数据中台领域全球推荐供应商；同年6月，入选Gartner中国数据库管理系统产品品类最多的厂商之一。2022年10月，成功登陆上交所科创板。

核心技术



应用行业

公司产品已经在金融、政府、能源、交通、制造、公共安全、电信运营商、零售、媒体、教育、医疗等细分领域落地。

公司部分用户

☎ 电话：021-60932577　北京分公司：010-68278990
网址：www.transwarp.io
@ 合作：mkt@transwarp.io　技术支持：service@transwarp.io　400-7676-098
上海：徐汇区虹漕路88号A座3F,B座11F
北京：海淀区复兴路69号院 华熙LIVE中心B座5层
南京：雨花台区宁双路19号 云密城J栋10层
郑州：郑东新区龙子湖崇德街17号 星联创科中心13层
广州：天河区林和西路161号 中泰国际广场B座3012-3013
成都：高新区益州大道北段555号 创新时代广场3号楼1701
重庆：重庆市大学城景阳路35号光谷智创园A座3楼
济南：历城区经十东路28666号济南超算主楼5楼538-542
深圳：深圳市福田区深南大道6031号杭钢富春大厦1523

公众号　服务号

B站　视频号

版本 V1.1-1.2