基于Seatunnel的生成式AI与非结构化数据迁移工具

VTS(Vector Transport Service) - 开源向量传输服务 By Zilliz

Nov.12/2024. Nian Liu

目录

- 背景介绍
- Motivation
- VTS(Vector Transport Service)
- Demo
- Roadmap
- Q/A

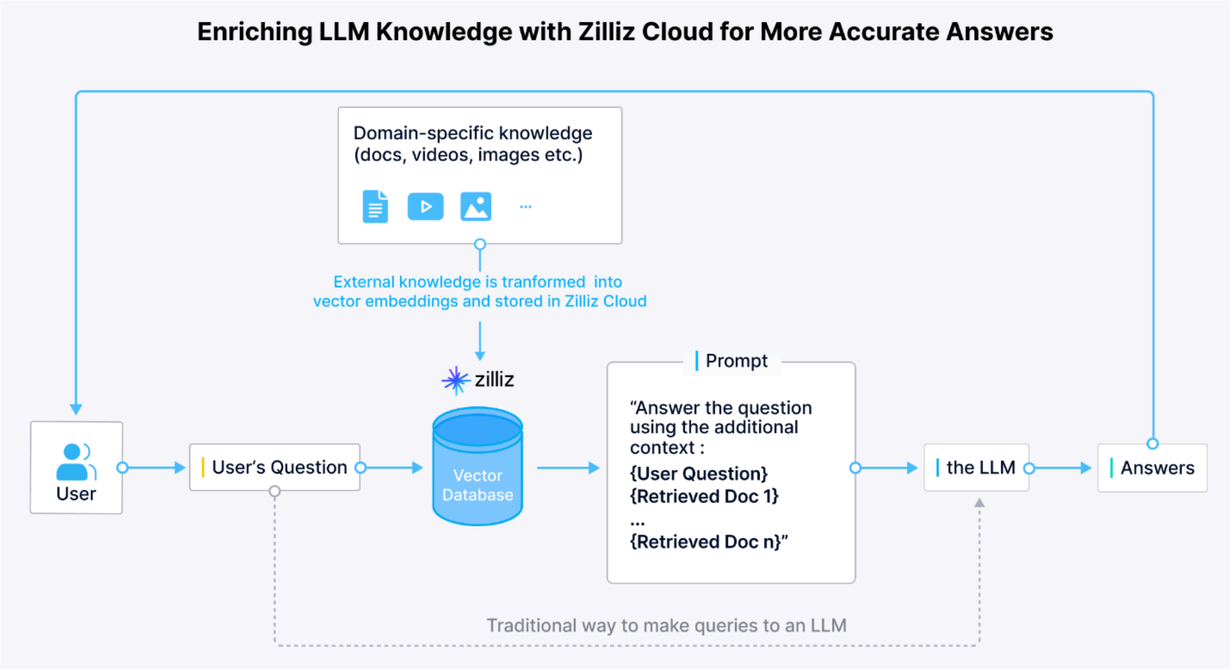
Mivus/Seatunnel

背景介绍

向量数据库

什么是向量数据库

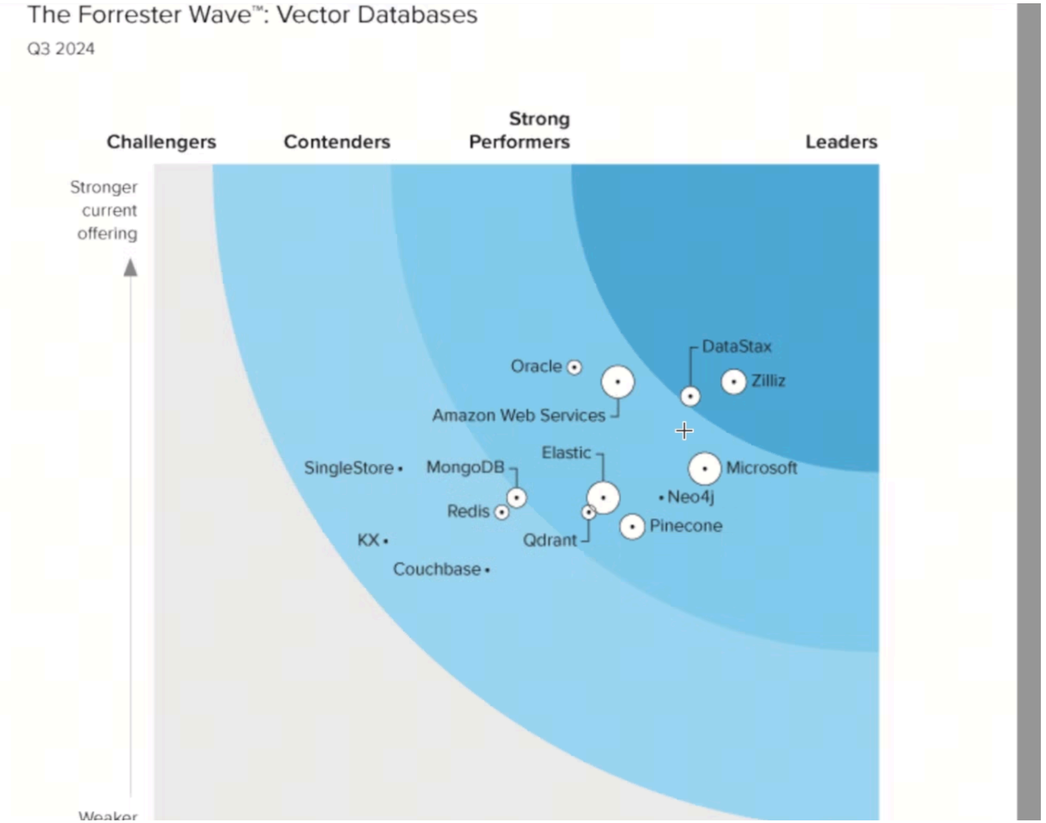
- 向量数据库是一种专门用于存储和检索向量数据的数据库系统
- 它能够高效处理高维向量数据，支持相似性搜索
  - 支持KNN(K-近邻)搜索
  - 计算向量间的距离(欧氏距离、余弦相似度等)
  - 快速检索最相似的向量
- 主要用于AI和机器学习应用场景
  - 图像检索系统
  - 推荐系统
  - 自然语言处理
  - 人脸识别
  - 相似商品搜索



Milvus

什么是 Milvus

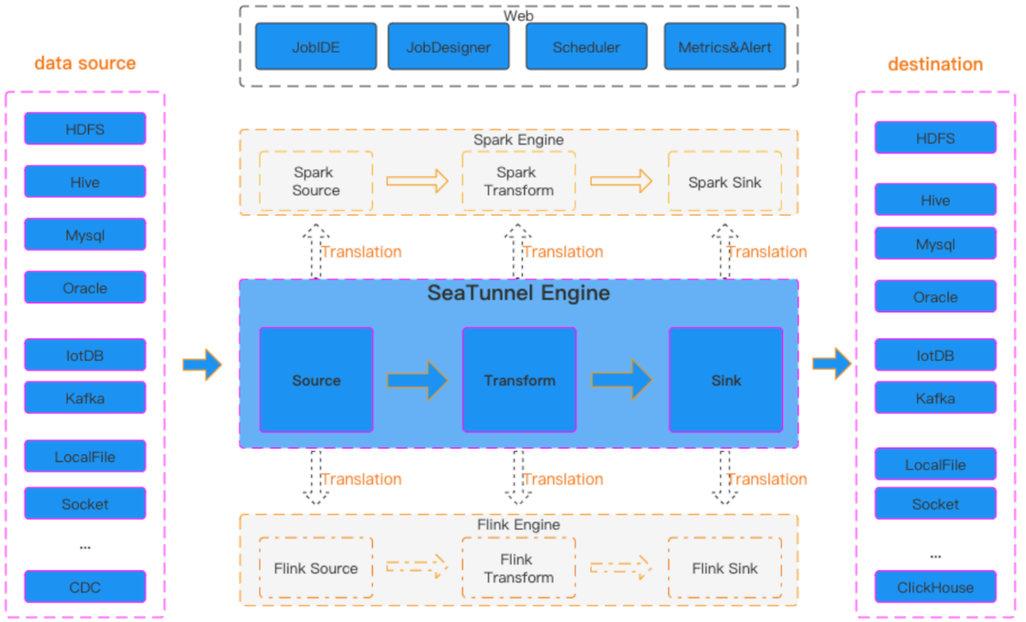
- 全球最受欢迎的开源向量数据库，Github 30000 Stars
- 向量数据库领导者 By Forrester Wave
- 云原生、高度可扩展，专为处理海量向量数据设计
- 支持多种部署环境
  - Milvus Lite
  - Milvus Standalong
  - Milvus K8s
- 提供开源版本和云服务版本(Zilliz Cloud)
  - https://zilliz.com/
  - https://milvus.io/



Seatunnel

About Seatunnel

- 分布式数据集成平台
- 丰富且可扩展的连接器系统
  - 近百个connector,涵盖Event Data, Transactional Data, Saas Data, Cloud DB, etc
- 多引擎支持(Zeta、Flink、Spark)
  - High throughput, low latency
- 近百家公司生产环境使用

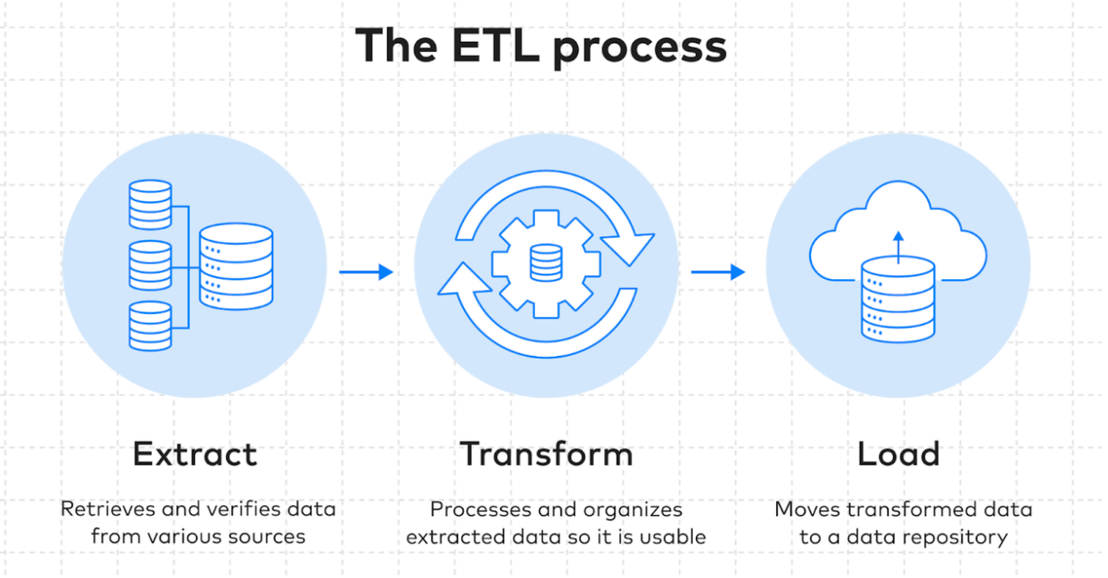


Motivation

为什么我们需要一个向量迁移工具

Current Situation
现状

- 传统ETL工具（AirByte/Fivetran/DataX, etc）
  - Close Source
  - 不支持向量数据处理
  - 缺乏向量特定的优化
  - 难以跟进技术演进
- 数据形态的挑战
  - 数据分布在多个平台
  - 数据格式多样性
  - 向量数据库能力差距



Motivation
动机

- 业务需求的压力
  - 用户迁移上云
- 迁移成本高
  - 已经生产应用
  - 不能停机
- 灵活性受限
  - 数据库切换

VTS(Vector Transport Service)

基于Apache Seatunnel开发的

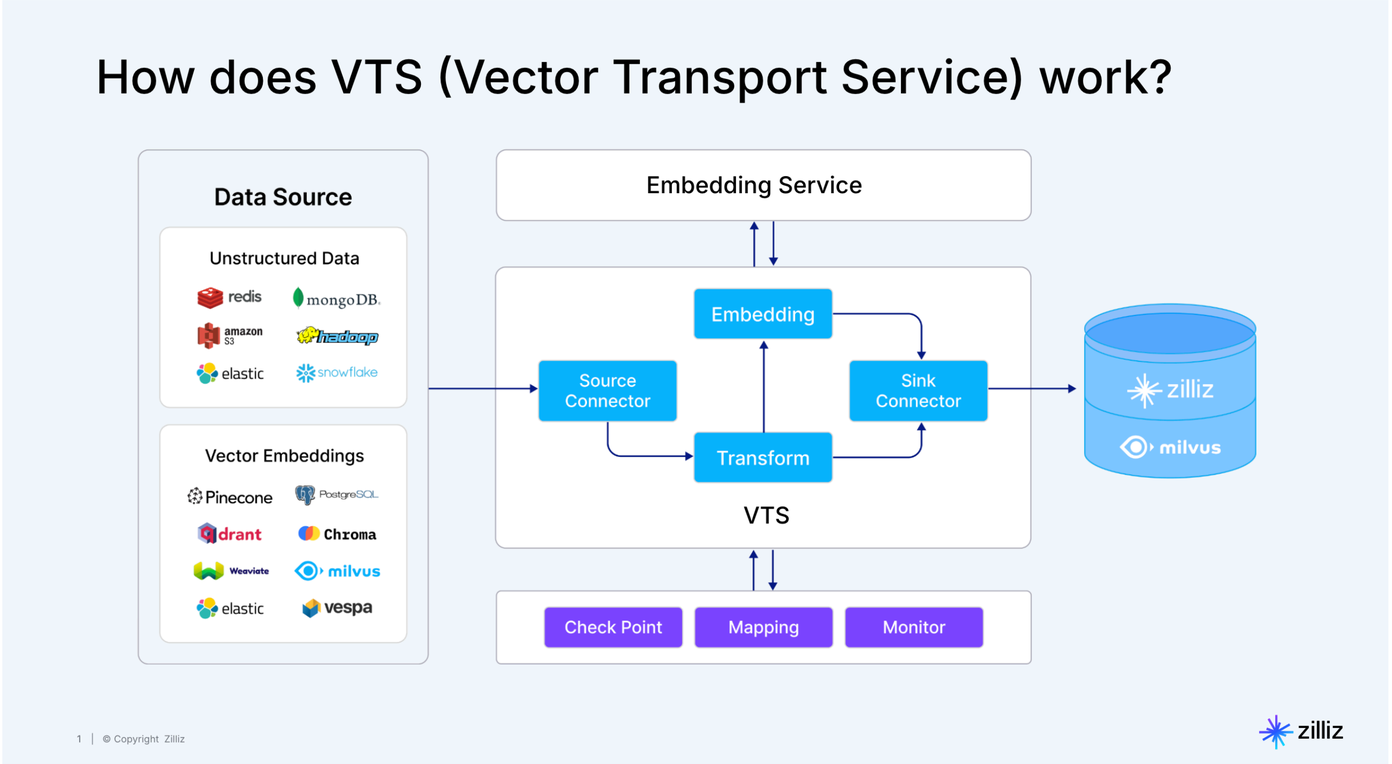
数据传输服务

什么是 VTS

- 全称Vector Transport Service(向量传输服务)，由Zilliz基于Apache Seatunnel开发
- 专注向量和非结构化数据迁移的开源工具
- 打通向量数据库与传统数据系统的桥梁，结合Milvus和Seatunnel的优势，解决向量数据迁移特有挑战
- https://github.com/zilliztech/vts
- 验证测试后会merge到seatunnel官方分支

Core Capability
核心能力

- 向量数据库迁移
- AI应用数据Pipeline构建
- 向量数据实时同步
- 非结构化数据转换与加载
- 跨平台数据集成



zilliz

VTS

支持的 Connector

- Milvus
- Pinecone
- Qdrant
- Postgres SQL
- ElasticSearch
- Tencent Vector DB

VTS

支持的 Transform

- TablePathMapper
  - 更改表名
- FieldMapper
  - 增删列
- Embedding
  - 文本向量化

VTS

支持的数据类型

- Float Vector
- Sparse Float Vector
- 多向量列
- 动态列
- 数据插入
  - Upsert
  - Bulk Insert（离线，大批量）

Demo

Pinecone -> Milvus

Performance

速率为2961/s，同步1亿向量需要约9个半小时(4core/8GB memory)

VTS

非结构化数据支持

- Shopify
- PDF
- Google Doc
- Slack
- Image/Text

应用

商品推荐场景-Shopify

1. 从shopify同步product, Inventory
2. Call embedding service -> 存入Milvus
3. 相似度搜索
4. 返回最相似的商品
5. Supported Already in VTS

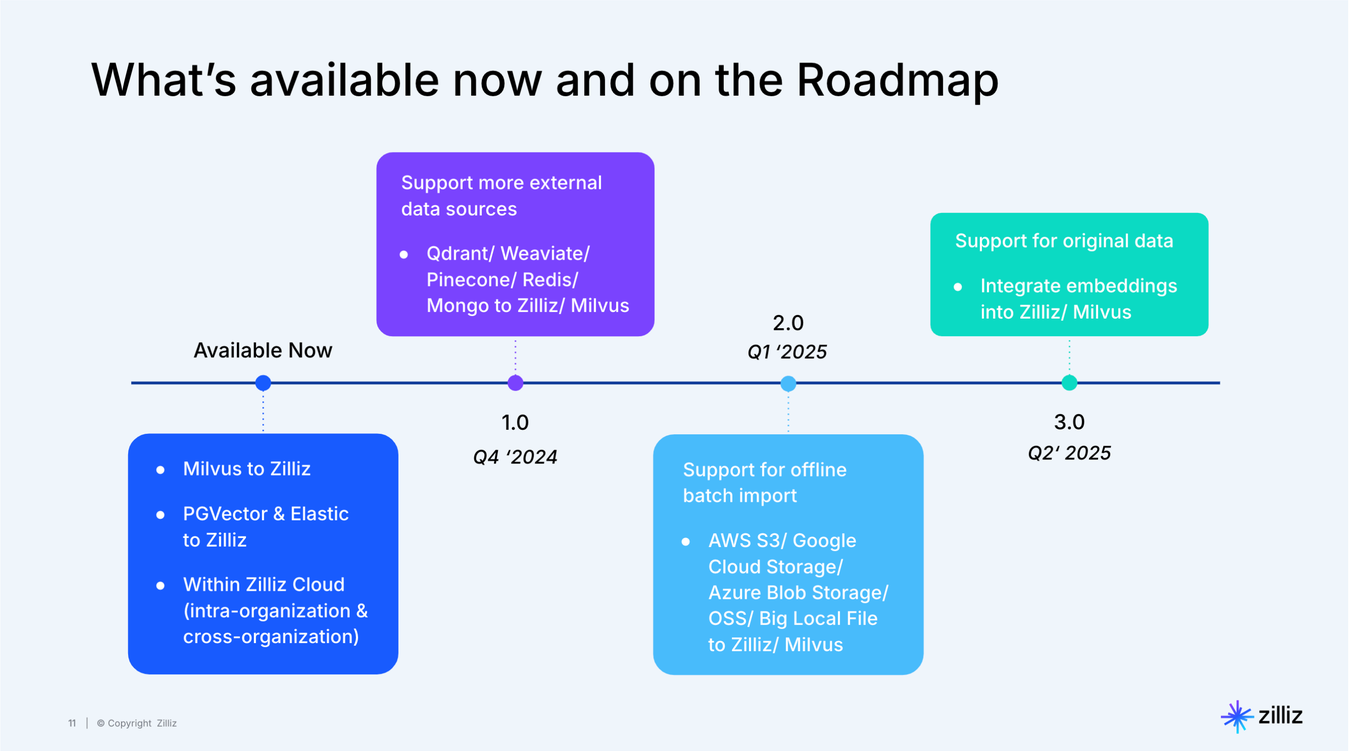
路径规划

Road Map

Roadmap

更多的数据源支持

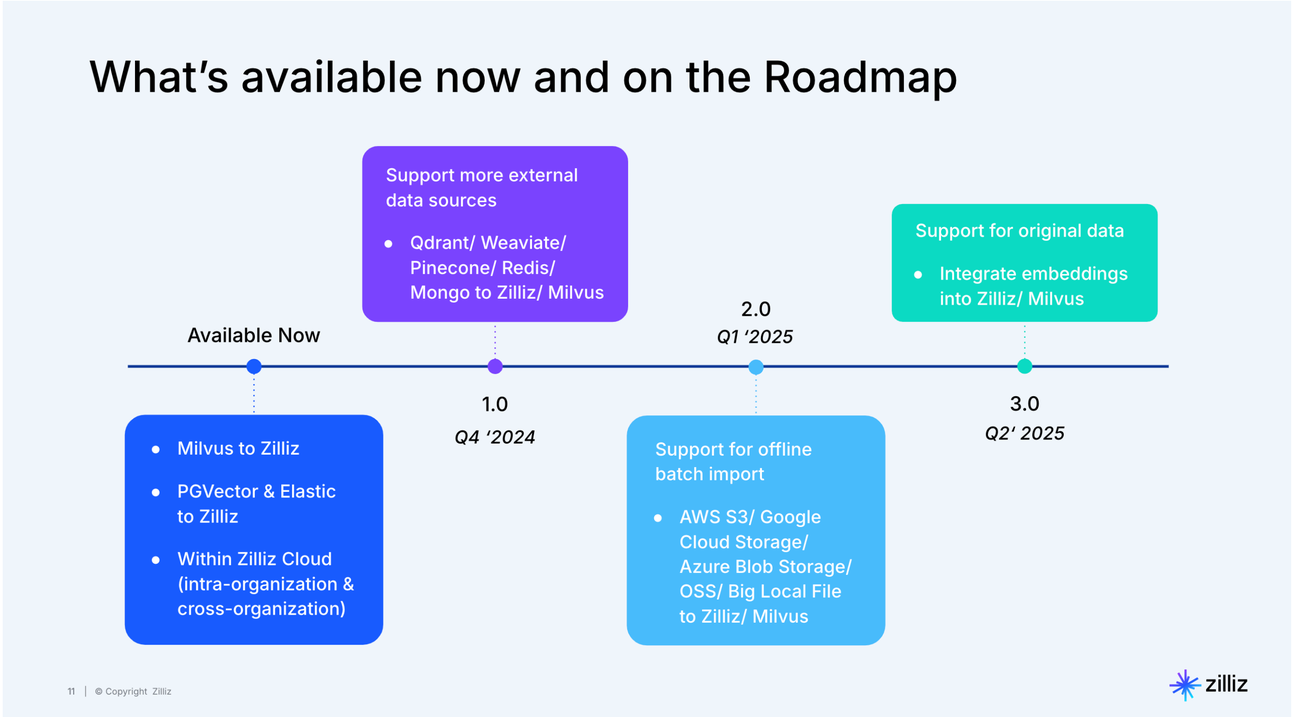
- Chroma DB
- DataStax(Astra DB)
- DataLake
- Mongo DB
- Kafka(real time AI)
- Object Storage Import
- ...



Roadmap

Milvus Data in/out

- Insert raw data directly
- Search with raw data
- Expected in Milvus 2.5



Roadmap

ETL pipeline for GenAI?

- 任务流编排
- Embedding service
- 外部API
- DolphinScheduler

Thank You!

Q/A

关注我们

- https://github.com/zilliztech/vts
- https://github.com/milvus-io/milvus
- https://milvus.io/
- https://zilliz.com
- Zilliz官方公众号