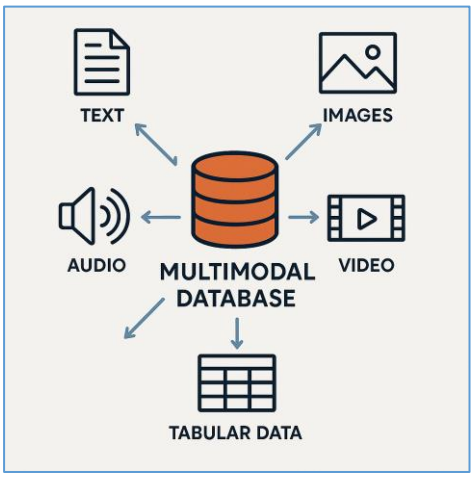
Proyecto Integrador
Sistema de Base de Datos Multimodal con Indexación Avanzada



1. Objetivo General

Diseñar e implementar un sistema de base de datos multimodal capaz de indexar y consultar datos estructurados y no estructurados, integrando técnicas de indexación avanzada. El proyecto final incluirá tres componentes: una API backend que incluye los principales componentes de procesamiento de consultas, indexación y persistencia, el cual estará conectado con aplicaciones FrontEnd orientadas a tareas prácticas de gestión de datos y aplicaciones para escenarios específicos. La integración multimodal permitirá gestionar y combinar diversos tipos de datos —texto, imágenes, audio, video y estructuras tabulares— ofreciendo una solución robusta y versátil para entornos de datos heterogéneos.

2. Ventajas de la Integración Multimodal

- **Contexto más rico y significativo:** La combinación de múltiples tipos de datos (texto, imágenes, audio, video, datos estructurados) proporciona un contexto más amplio y profundo, lo que permite una comprensión más completa de la información.
- **Aplicaciones más versátiles y adaptables:** Facilita el desarrollo de soluciones en dominios donde los datos provienen de fuentes heterogéneas, como sistemas de recomendación, análisis de contenido multimedia, salud digital, vigilancia inteligente, entre otros.

Jiaheng Lu and Irena Holubová. 2019. Multi-model Databases: A New Journey to Handle the Variety of Data. ACM Comput. Surv. 52, 3, Article 55 (May 2020)

3. Arquitectura General del Proyecto

Se usará una arquitectura orientada a microservicios, cada aplicación es un **microservicio:**

3.1. Backend (API de minigestor multimodal de BD)

Expone endpoints REST para consultas y administración.
Componentes internos:

a) **ParserSQL Personalizado**
   - Traduce las consultas SQL-like en un Plan de Ejecución Interno.
   - Extensiones para **consultas vectoriales**.

b) **Query Engine (Motor de Ejecución)**
   - Coordina las operaciones sobre los distintos módulos de almacenamiento.
   - Decide qué índice usar según la consulta (Optimizer básico).

c) **Módulo de Almacenamiento Tabular**
   - **Gestor de Archivos Tabulares** (CSV, o bloques binarios).

- Métodos CRUD básicos.
- Implementación de manejadores de índices clásicos:
  - B+Tree Index
  - Hash Index
  - Sequential File

d) **Módulo de Almacenamiento Vectorial**
- Gestión de embeddings (imágenes, audio, texto).
- Indexación con:
  - **IVF Flat / PQ**
  - **Índice Invertido para Descriptores Locales**
- Consultas k-NN y por rango.

e) **Extractor de Embeddings (Pipeline IA externo o integrado)**
- Texto → Bag of Words, TF-IDF.
- Imagen → CNN / CLIP.
- Audio → MFCC.
- Devuelve un vector que se indexa en el **Módulo Vectorial**.

f) **Persistencia**
- Archivos tabulares en disco (formato propietario, simula RDBMS).
- Índices en disco (B+Tree, Hash).
- Vector Store en disco (binarios optimizados para búsqueda KNN).
- Los modelos IA entrenados en un **repositorio de modelos**.
- Metadatos.

3.2. Frontend (UI cliente)

- WebApp ligera (React o Flask/Django con templates).
- Funciones:
  - Enviar consultas SQL personalizadas al backend.
  - Visualizar resultados tabulares.
  - Subir archivos tabulares (CSV).
  - Subir imágenes/audio/texto → para vectorización e indexación.
  - Panel de exploración de índices (B+Tree, Hash, Sequential).

3.3. Capa de Aplicaciones

Módulos que **se enchufan al backend** y usan los datos/indexaciones ya disponibles:

- **Sistema de gestión de inventarios:** Gestión de productos en almacenes, con búsquedas por código, nombre, categoría o ubicación.
  - Soporte para productos con dimensiones físicas (peso, tamaño).
  - Se puede indexar por ubicación en un almacén 3D (R-Tree).
  - https://zlatanova.xyz/PhDthesis/pdf/ch7.pdf
- **Sistema de gestión geoespacial:** Gestión de ubicaciones, rutas o zonas geográficas:
  - Estaciones meteorológicas
  - Puntos de interés turístico
  - Rastreo de vehículos o envíos
  - https://www.veraset.com/insights/uses-of-geospatial-data
- **Aplicaciones de IA con datos multidimensionales:** aplicaciones que permitan la búsqueda por similitud sobre vectores numéricos:
  - **Reconocedor de rostros**: Búsqueda de imágenes similares en base a descriptores faciales.
  - **Detector de copias de audios**: Detección de similitudes entre clips de audio usando descriptores.
  - **Sistema de recomendación de noticias**: Basado en similitudes entre textos o embeddings.

- **Recomendación musical**: Usando similitudes en lyrics (texto) y características de audio.
- **E-Commerce**: Recomendación de productos basados en descriptores de imágenes y características textuales.

4. Estructura del Proyecto

El proyecto se realizará en dos fases:

1) Organización e Indexación Eficiente de Archivos con Datos Tabulares y Espaciales
2) Mapeando el Caos: Indexación y Organización de Datos No Estructurados para Datos Multimedia (texto, imágenes, audio, video).