Proyecto2 - Bases de Datos II

Portada

Proyecto 2: API de restaurantes y Capa OLAP
Curso: Bases de Datos II
Integrantes:

- Daniel Alemán
- Luis Meza

Índice

Enlace de GitHub

Repositorio del Proyecto

Enlace sobre la Arquitectura del Proyecto

Arquitectura del Proyecto

Descripción del Proyecto

Este proyecto implementa un sistema completo para gestión de restaurantes con una arquitectura de microservicios. Permite administrar restaurantes, menús, productos, reservaciones y pedidos a través de un conjunto de APIs RESTful. El sistema está diseñado con alta disponibilidad, escalabilidad y rendimiento como prioridades, empleando tecnologías modernas como balanceo de carga, sharding de bases de datos, caché distribuida y búsquedas optimizadas. También implementa mediante Airflow el uso de DAGs para poder realizar ETL y reindexar los productos de manera automática en Elastic. Por otra parte, también incluye la funcionalidad de visualizar datos de analíticas significativas.

Arquitectura

Arquitectura Lógica

El proyecto sigue una arquitectura de microservicios con los siguientes componentes clave:

1. **Microservicios API**:
    - API principal para operaciones CRUD de restaurantes, menús, productos, reservas y pedidos
    - Servicio de autenticación para gestión de usuarios y tokens JWT
    - Servicio de búsqueda optimizado con Elasticsearch
    - Servicio de indexación y operaciones de ruteo dentro de un grafo mediante Neo4J
2. **Balanceo de Carga**:
    - Nginx como proxy inverso y balanceador para distribuir las peticiones entre instancias
3. **Persistencia**:
    - MongoDB: Almacenamiento principal con sharding y replicación para alta disponibilidad
    - PostgreSQL: Almacenamiento alternativo configurable
    - Elasticsearch: Índice de búsqueda para consultas optimizadas
    - HiveDB: Almacenamiento como Datawarehouse
4. **Caché**:
    - Redis como almacén de caché distribuida para mejorar el rendimiento
5. **CI/CD**:
    - Pipeline automatizado para pruebas, construcción y despliegue

Estructura del Proyecto

La estructura de directorios del proyecto está organizada por funcionalidad:

```
proyecto2-bases2/
├── airflow/                # Servicio de Airflow para realizar DAGs automatizados
├── analytics_service/      # Servicio de dashboards de analítica
├── api/                    # API principal
├── auth_service/           # Servicio de autenticación
├── drivers/                # Drivers necesarios para la metadata de Hive
├── etl_service/            # Servicio con propósito específico de ETL completo
├── graph_service/          # Servicio de indexación y ruteo con Neo4j
├── pruebas/                # Scripts de prueba y generación de datos
├── search_service/         # Servicio de búsqueda
├── spark_analytics/        # Script básico de consultas con Spark y Hive
├── docker-compose.yml      # Configuración de contenedores
├── hive_warehouse_init.sql # Inicialización del warehouse mediante HiveQL
├── init_cluster.sh         # Script para inicializar el cluster MongoDB
├── init_hive_warehouse.sh  # Script para inicializar Hive
├── init.sql                # Inicialización de las tablas de Postgres para los
servicios
├── set_config.sh           # Script para configurar el entorno
├── nginx.conf              # Configuración del balanceador de carga
└── README.md               # Documentación
```

Cada servicio sigue una estructura MC (Modelo-Controlador) con separación clara de responsabilidades:

```
servicio/
├── src/
│   ├── config/             # Configuración
│   ├── controllers/        # Controladores
│   ├── dao/                # Objetos de acceso a datos
│   ├── db/                 # Conexiones a bases de datos
│   ├── middlewares/        # Middlewares
│   ├── models/             # Modelos de datos
│   ├── routes/             # Definición de rutas
│   └── app.js              # Aplicación principal
├── swagger/                # Documentación de API
├── tests/                  # Pruebas
│   ├── integration/        # Pruebas de integración
│   ├── unit/               # Pruebas unitarias
│   └── utils/              # Utilidades para pruebas
└── server.js               # Punto de entrada
```

Componentes Principales

API Principal

La API principal maneja todas las operaciones CRUD relacionadas con:

- Restaurantes

- Menús
- Productos
- Reservaciones
- Pedidos
- Repartidores

Características clave:

- Implementación RESTful con Express.js
- Documentación completa con Swagger
- Escalabilidad horizontal con múltiples instancias (api1, api2)
- Interconexión con servicios de autenticación y búsqueda

Servicio de Autenticación

Este microservicio gestiona todo lo relacionado con usuarios y seguridad:

- Registro de usuarios
- Inicio de sesión
- Verificación de tokens JWT
- Gestión de roles (administrador/cliente)

Características:

- Autenticación basada en JWT para seguridad
- Almacenamiento seguro de contraseñas con hash
- Alta disponibilidad con múltiples instancias (auth_service1, auth_service2)

Servicio de Búsqueda

Un microservicio independiente dedicado a proporcionar funcionalidad de búsqueda optimizada para el sistema:

- **Funcionalidades principales**:
    - Búsqueda de productos por texto libre
    - Búsqueda de productos por categoría
    - Indexación automática de productos nuevos
    - Reindexación manual de todo el catálogo
    - Sincronización con la base de datos principal
- **Características técnicas**:
    - Integración con Elasticsearch como motor de búsqueda de alto rendimiento
    - Índices optimizados para consultas rápidas y flexibles
    - Búsquedas tolerantes a errores tipográficos
    - Resultados relevantes con ponderación inteligente
    - Actualización en tiempo real del índice cuando cambian los productos
    - Alta disponibilidad mediante múltiples instancias (search_service1, search_service2)
    - Resiliencia ante fallos mediante manejo de errores robusto

El servicio está diseñado para funcionar de manera independiente, lo que permite que el sistema principal siga operando incluso si el servicio de búsqueda experimenta problemas temporales.

Servicio de Grafos

Este microservicio se encarga de resolver consultas complejas mediante estructuras de grafos, permitiendo obtener rutas óptimas, relaciones de co-ocurrencia, y recomendaciones inteligentes. Su propósito es brindar capacidades de análisis y búsqueda especializadas dentro del ecosistema del sistema principal.

- **Funcionalidades principales**:
    - Cálculo y actualización de relaciones de co-compras entre productos.
    - Identificación de usuarios influyentes a partir de patrones de recomendación.
    - Cálculo y optimización de la ruta de entrega más eficiente para repartidores.
    - Asignación automática del repartidor más adecuado para cada pedido.
- **Características técnicas**:
    - Utiliza Neo4j como motor de base de datos orientada a grafos, optimizada para consultas relacionales complejas.
    - Implementa índices especializados para búsquedas rápidas y flexibles.
    - Soporta búsquedas tolerantes a errores tipográficos, mejorando la experiencia del usuario.
    - Ofrece resultados relevantes mediante ponderación inteligente basada en popularidad, contexto y conexiones.
    - Realiza actualizaciones en tiempo real del grafo ante cambios en los productos o relaciones.
    - Garantiza alta disponibilidad mediante múltiples instancias paralelas (graph_service1, graph_service2).
    - Cuenta con mecanismos de resiliencia para manejar fallos sin afectar al sistema principal.

Este servicio opera de forma desacoplada del núcleo del sistema, asegurando que la funcionalidad general no se vea comprometida ante caídas temporales del servicio de grafos.

Servicio de ETL y Data Warehouse

Este microservicio se encarga de la extracción, transformación y carga de datos (ETL) desde las fuentes operacionales hacia el almacén de datos analítico, proporcionando una base sólida para el análisis OLAP y la generación de reportes estratégicos del negocio.

- **Funcionalidades principales**:
    - Extracción automatizada de datos desde MongoDB y PostgreSQL utilizando conectores Python especializados.
    - Transformación de datos mediante Apache Spark con SparkSQL para procesamiento distribuido de grandes volúmenes.
    - Carga optimizada de datos transformados hacia Apache Hive siguiendo esquemas estrella y copo de nieve.
    - Validación de integridad y calidad de datos durante todo el proceso ETL.
- **Características técnicas**:

- Utiliza Apache Spark como motor de procesamiento distribuido para transformaciones complejas y análisis de tendencias.
- Implementa Apache Hive como Data Warehouse principal, optimizado para consultas analíticas y almacenamiento columnar.
- El warehouse se inicializa automáticamente mediante scripts de shell que configuran esquemas, particiones y estructuras necesarias.
- Soporta procesamiento incremental para minimizar el impacto en recursos y tiempo de ejecución.
- Mantiene conexiones persistentes y pools de conexiones para optimizar el rendimiento de transferencia de datos.
- Implementa mecanismos de recuperación ante fallos y reintentos automáticos para garantizar la consistencia.

Servicio de Orquestación con Apache Airflow

Este componente centraliza y automatiza la ejecución de procesos de datos mediante flujos de trabajo programables y monitoreables, asegurando la actualización continua y confiable del ecosistema analítico.

- **Funcionalidades principales**:
  - Orquestación automatizada del pipeline ETL completo desde extracción hasta carga final.
  - Coordinación de dependencias entre tareas y servicios del sistema de datos.
  - Monitoreo y alertas automáticas ante fallos o anomalías en los procesos.
  - Gestión de reindexación de catálogos de productos en ElasticSearch.
- **Características técnicas**:
  - Implementa dos DAGs principales especializados para diferentes necesidades operacionales:
    - **DAG de ETL**: Ejecuta el proceso completo de extracción, transformación y carga cada 6 horas, manteniendo actualizado el Data Warehouse con datos frescos para análisis.
    - **DAG de Reindexación**: Gestiona la actualización automática de índices de ElasticSearch cuando se detectan cambios en el catálogo de productos, garantizando búsquedas actualizadas.
  - Utiliza sensores y operadores especializados para integración con Spark, Hive y ElasticSearch.
  - Proporciona interfaz web para monitoreo en tiempo real del estado de ejecución y logs detallados.
  - Implementa políticas de reintentos configurables y manejo de excepciones para alta confiabilidad.
  - Mantiene historial completo de ejecuciones para auditoría y análisis de rendimiento.

Este servicio opera como el cerebro coordinador del ecosistema de datos, asegurando que todas las transformaciones y actualizaciones se ejecuten de manera ordenada, puntual y confiable, sin intervención manual.

Balanceador de Carga (Nginx)

El sistema utiliza Nginx como balanceador de carga y proxy inverso para distribuir el tráfico entre múltiples instancias de cada microservicio:

- **Configuración implementada**:
    - Balanceo de carga para API principal entre instancias api1 y api2
    - Balanceo de carga para servicio de autenticación entre auth_service1 y auth_service2
    - Balanceo de carga para servicio de búsqueda entre search_service1 y search_service2
        - Balanceo de carga para servicio de grafos entre grap_service1 y graph_service2
    - Enrutamiento basado en prefijos de URL (/api/, /auth/, /search/, /graph/)
    - Terminación SSL centralizada
- **Características técnicas**:
    - Algoritmo de balanceo round-robin para distribución uniforme de carga
    - Compresión de respuestas para optimizar el ancho de banda
    - Buffer y timeouts configurados para operaciones de larga duración
    - Redirección inteligente basada en path de URL
    - Health checks periódicos para detectar instancias no disponibles
- **Beneficios para el sistema**:
    - Alta disponibilidad mediante múltiples instancias de cada servicio
    - Escalabilidad horizontal sencilla (añadir más instancias sin cambios en la aplicación)
    - Resistencia ante fallos de servicios individuales
    - Punto único de entrada para los clientes con enrutamiento transparente
    - Capacidad de actualizar servicios individuales sin interrumpir el sistema completo

Nginx corre en su propio contenedor Docker, y su configuración se monta desde el archivo nginx.conf en el sistema host.

Sistema de Caché (Redis)

El proyecto implementa Redis como un sistema de caché distribuida para optimizar el rendimiento y reducir la carga en las bases de datos:

- **Casos de uso implementados**:
    - Caché de productos individuales y listados completos
    - Almacenamiento temporal de resultados de consultas frecuentes
    - Caché de datos de autenticación y sesiones
    - Invalidación automática de caché al modificar recursos
- **Estrategia de caché**:
    - Implementación de patrón Cache-Aside para recursos frecuentemente accedidos
    - TTL (Time-To-Live) configurado según el tipo de datos
    - Invalidación selectiva al modificar recursos relacionados
    - Manejo de versiones de datos en caché
- **Beneficios medibles**:
    - Reducción de tiempos de respuesta de API en hasta un 80% para recursos en caché
    - Alivio significativo de carga en MongoDB y PostgreSQL

    - Mayor resistencia del sistema ante picos de tráfico
    - Acceso ultrarrápido en memoria para datos de alta demanda

Las pruebas de integración incluyen verificaciones específicas del comportamiento del caché, asegurando que el sistema maneja correctamente la obtención, almacenamiento e invalidación de datos en caché.

Bases de Datos

MongoDB (Principal)

- Configurado con sharding y replicación para alta disponibilidad y escalabilidad
- Estructura:
    - 2 shards con 3 réplicas cada uno
    - 3 servidores de configuración
    - 3 routers (mongos)
- Colecciones fragmentadas mediante hash de identificadores para distribución uniforme

PostgreSQL

- Base de datos relacional como alternativa configurable
- Ideal para consultas complejas y relaciones estructuradas

Elasticsearch

- Motor especializado para búsquedas de texto completo
- Indexación optimizada de productos para consultas rápidas

Instalación

Requisitos Previos

Para ejecutar este proyecto necesita:

- Docker y Docker Compose
- Git
- Node.js y npm (para desarrollo local)

Instalación del Proyecto

Si desea clonar el repositorio y hacer uso de él, basta con utilizar la siguiente serie de comandos:

```
git clone https://github.com/DanielAR27/proyecto2-bases2.git
cd proyecto2-bases2
```

Instalación de Módulos

Para cada servicio que requiere instalación de módulos:

- ./api

- ./auth_service
- ./search_service
- ./graph_service

Para ubicarse dentro de ellos, debe estar en la raíz del proyecto y utilizar el siguiente comando:

```
cd <nombre_del_servicio>
```

Una vez dentro, puede instalar o actualizar los módulos necesarios:

```
npm install
```

Construcción y Levantamiento de los Servicios

Para construir los contenedores e iniciar toda la aplicación, primero debe dar permisos de ejecución a los scripts:

```
chmod +x set_config.sh
chmod +x init_cluster.sh
chmod +x init_hive_warehouse.sh
```

Luego ejecute el script principal:

```
./set_config.sh
```

Este script realiza las siguientes acciones:

1. Levanta todos los servicios base con `docker-compose up -d`
2. Inicializa el cluster MongoDB con sharding y replicación mediante `./init_cluster.sh`
3. Inicializa el warehouse de Hive mediante `./init_hive_warehouse.sh`
4. Levanta los servicios backend (API, Auth, Search) con `docker-compose --profile backend up --build -d`

El script `init_cluster.sh` configura MongoDB con:

- Replica Set de configuración (3 nodos)
- Dos Shards Replica Set (3 nodos cada uno)
- Habilitación de sharding en la base de datos `apidb`
- Configuración de colecciones particionadas
- Preparación de metadatos en los routers

Agregación de datos sintéticos

Posteriormente, si desea generar datos sintéticos puede seguir los siguientes pasos:

1. Dirigase a la carpeta de pruebas mediante el siguiente comando, debe estar ubicado en la raíz del proyecto

```
cd pruebas/
```

2. Instale las dependencias necesarias para proceder con la generación de los datos utilizando el siguiente comando

```
npm install
```

3. Una vez instaladas las dependencias, puede hacer uso del siguiente comando para generar datos sintéticos

```
node generarDatosMasivos.js
```

Levantamiento de servicios ETL, Airflow y capa OLAP

Una vez insertados datos dentro de la base de datos, puede hacer uso del siguiente comando para levantar el servicio encargado de ETL

```
docker-compose --profile etl up --build -d
```

Debe esperar un tiempo a que se ejecute la primera vez para luego poder levantar el servicio de Airflow encargado de ejecutar el proceso ETL y la reindexación cada cierto tiempo, está configurado por defecto en 6 horas.

```
docker-compose --profile airflow up --build -d
```

Finalmente, para poder visualizar dashboards de analítica puede hacer uso del siguiente comando

```
docker-compose --profile analytics up --build -d
```

Para visualizar los dashboards puede hacer uso de la siguiente dirección: `http://localhost:8501/`

Si desea ver algunas métricas básicas, puede esperar a que cargue el contenedor y luego utilizar el siguiente comando

```
docker logs -f spark_analytics
```

Pruebas

Pruebas Unitarias y de Integración

El proyecto cuenta con pruebas automatizadas para cada servicio:

Servicio de Autenticación

```
# Construir contenedor de prueba
docker-compose --profile test build auth_test

# Ejecutar pruebas con cobertura
docker-compose --profile test run --rm auth_test
```

Servicio de Búsqueda

```
# Construir contenedor de prueba
docker-compose --profile test build search_test

# Ejecutar pruebas con cobertura
docker-compose --profile test run --rm search_test
```

API Principal

```
# Construir contenedor de prueba
docker-compose --profile test build api_test

# Ejecutar pruebas con cobertura
docker-compose --profile test run --rm api_test
```

Cobertura de Pruebas

El sistema utiliza Jest como framework de pruebas y se centra en dos tipos principales de pruebas:

- **Pruebas unitarias**: Verifican el comportamiento correcto de componentes individuales como DAOs, controladores y modelos.
- **Pruebas de integración**: Evalúan la interacción entre diferentes partes del sistema, incluyendo:
  - Flujos CRUD completos
  - Validación de datos de entrada
  - Manejo de permisos y autenticación
  - Integración entre servicios

    - Funcionamiento del sistema de caché

Cada prueba genera informes detallados de cobertura que muestran qué porcentaje del código está siendo evaluado.

Un ejemplo de los aspectos verificados en las pruebas de integración:

- Creación, lectura, actualización y eliminación de recursos
- Comportamiento correcto frente a entradas inválidas
- Verificación de permisos según roles de usuario
- Comportamiento del caché Redis
- Resiliencia ante fallas en servicios externos

CI/CD Pipeline

El proyecto implementa un pipeline completo de Integración Continua y Despliegue Continuo utilizando GitHub Actions, lo que garantiza la calidad del código y facilita el proceso de entrega:

Workflow de CI/CD

El pipeline se activa automáticamente en los siguientes casos:

- Con cada push a las ramas main, master o develop
- Al crear Pull Requests hacia main o master

Etapas del Pipeline

1. **Etapa de Pruebas (Test)**:
    - Clona el repositorio y configura el entorno usando template.env
    - Levanta la infraestructura completa con Docker Compose
    - Inicializa el cluster MongoDB con la configuración de sharding
    - Construye y arranca los servicios backend
    - Ejecuta las pruebas unitarias y de integración para cada servicio:
        - Pruebas del servicio de autenticación
        - Pruebas de la API principal
        - Pruebas del servicio de búsqueda
    - Genera informes de cobertura de código
2. **Etapa de Construcción y Publicación (Build-and-Push)**:
    - Se ejecuta solo después de que las pruebas sean exitosas en ramas main o master
    - Configura Docker Buildx para construcción multiplataforma
    - Autentica con GitHub Container Registry
    - Construye imágenes Docker optimizadas para cada servicio:
        - Servicio de autenticación
        - API principal
        - Servicio de búsqueda
    - Publica las imágenes en GitHub Container Registry con el tag "latest"
3. **Etapa de Despliegue (Deploy)**:

    - Crea un artefacto de despliegue que incluye:
        - Docker Compose configurado
        - Variables de entorno (.env)
        - Configuración de Nginx
        - Scripts de inicialización
    - Genera documentación detallada con instrucciones paso a paso para el despliegue
    - Sube los artefactos al sistema de almacenamiento de GitHub Actions
    - Prepara instrucciones para actualizar un despliegue existente

Este pipeline garantiza que:

- Todo el código pase las pruebas automatizadas
- Solo el código verificado se construya y publique
- Se generen artefactos consistentes para cada versión
- El proceso de despliegue sea reproducible y documentado

El workflow completo está definido en el archivo ci-cd.yml en la raíz del repositorio.

Acceso a Interfaces

Documentación de la API Rest

Para visualizar la documentación interactiva generada con Swagger:

- API Principal: `http://localhost/api/api-docs/`
- Servicio de Autenticación: `http://localhost/auth/api-docs/`
- Servicio de Búsqueda: `http://localhost/search/api-docs/`
- Servicio de Grafos: `http://localhost/graph/api-docs/`

Visualización en Tiempo Real de Postgres DB

Use PgAdmin para gestionar la base de datos PostgreSQL:

```
http://localhost:5050
```

Pasos para configurar PgAdmin:

1. Click derecho en "Servers" → Register → Server
2. En General: Nombre = "PG Docker"
3. En Connection:
    - Host: `postgres_container`
    - Port: `5432`
    - Database: `apidb`
    - Username: `postgres`
    - Password: `postgres`

Visualización en Tiempo Real de Mongo DB

Use Mongo Express para gestionar MongoDB:

```
http://localhost:8081
```

Pasos para acceder:

1. Introduzca las credenciales configuradas en el .env
2. Acceda a la base de datos "apidb" para ver las colecciones
3. Explore las colecciones: counters, menus, pedidos, productos, reservas, restaurantes, usuarios

Elasticsearch y Kibana

Para gestionar y monitorear Elasticsearch:

```
http://localhost:5601
```

Kibana ofrece una interfaz intuitiva para:

- Explorar índices
- Crear y probar consultas
- Visualizar datos
- Monitorear rendimiento

Reinicio Completo del Entorno

Si desea eliminar todos los contenedores, redes y volúmenes:

```
docker-compose down -v
```

Esto restablecerá completamente el entorno y eliminará todos los datos almacenados.

Autores: Daniel Alemán, Luis Meza
Última actualización: *11/5/2025*