BASES DE DATOS
COM-12101
PROYECTO FINAL

Introducción

Con el presente proyecto, el alumno se enfrentará a un conjunto de datos desnormalizado que le permitirá poner a prueba sus habilidades de diseño de bases de datos, definición de esquemas, manipulación y consulta de datos.

El proyecto se evaluará por equipos (los mismos equipos que en las tareas) y debido a la naturaleza del curso, **deberá efectuarse en su totalidad usando SQL como herramienta de estudio, limpieza y normalización de datos**.

Conjunto de datos

El equipo debe seleccionar una fuente de datos de su preferencia para la ejecución del proyecto. La fuente de datos deberá ser real, pública y de interés para el resto del grupo. Algunas **sugerencias** de portales de sets de datos son:

- Chicago Data Portal (https://data.cityofchicago.org)
- United States Government's Open Data (https://data.gov)
- Portal de Datos Abiertos de la CDMX (https://datos.cdmx.gob.mx)

Las siguientes restricciones aplican para la selección de conjunto de datos:

- Debe poder descomponerse en al menos 3 entidades
- La carga inicial debe constar de al menos 5,000 registros
- La carga inicial debe constar de al menos 10 atributos distintos
- No puede ser una base de datos normalizada que se desnormalice como parte del proyecto

Contenidos del proyecto

La entrega final del proyecto se realizará en un repositorio alojado en GitHub. Por cada rubro, se debe tener por lo menos un script independiente en el repositorio y una sección dedicada en el README donde se documenten los scripts y se realicen las explicaciones y demostraciones pertinentes.

Las actividades a considerar son:

Entrega	Actividad	Fecha	Entregas Evaluadas
1	A	29-enero-2025	1
2	B, C	2-abril-2025	1, 2
3	D, E	30-abril-2025	2, 3
FINAL	**TODAS**	**14-mayo-2025**	**TODAS**

A) Introducción al conjunto de datos y al problema a estudiar considerando aspectos éticos del conjunto de datos empleado

El equipo deberá seleccionar un set de datos y detallar los siguientes aspectos en el README del repositorio:

- Descripción general de los datos
- ¿Quién los recolecta?
- ¿Cuál es el propósito de su recolección?
- ¿Dónde se pueden obtener?
- ¿Con qué frecuencia se actualizan?
- ¿Cuántas tuplas y cuántos atributos tiene el set de datos?
- ¿Qué significa cada atributo del set?
- ¿Qué atributos son numéricos?
- ¿Qué atributos son categóricos?
- ¿Qué atributos son de tipo texto?
- ¿Qué atributos son de tipo temporal y/o fecha?
- ¿Cuál es el objetivo buscado con el set de datos? ¿Para qué se usará por el equipo?
- ¿Qué consideraciones éticas conlleva el análisis y explotación de dichos datos?

B) Carga inicial y análisis preliminar

Se debe documentar en el repositorio cómo realizar la carga inicial del set de datos a una base de datos de tipo Postgres. Así mismo, se deben agregar los scripts pertinentes para la creación del esquema inicial de la carga.

También, mediante el uso de consultas SQL, que deben ser incluidas en un script en el repositorio, se deberá realizar un análisis exploratorio de los datos. Algunas sugerencias son:

- ¿Existen columnas con valores únicos?
- Mínimos y máximos de fechas
- Mínimos, máximos y promedios de valores numéricos
- Duplicados en atributos categóricos
- Columnas redundantes
- Conteo de tuplas por cada categoría
- Conteo de valores nulos
- ¿Existen inconsistencias en el set de datos?

C) Limpieza de datos

El equipo debe detallar qué actividades de limpieza se deben efectuar al set de datos para su uso, siempre teniendo en mente el objetivo planteado para el proyecto. En el README se debe incluir una sección con las actividades realizadas, explicar cualquier operación no trivial usada y explicar por qué fue necesaria dicha actividad de limpieza. Además se debe tener en el repositorio al menos un script para efectuar la limpieza con los datos en bruto.

D) Normalización de datos hasta cuarta formal normal

El equipo debe proponer y justificar una descomposición intuitiva de los datos en diversas entidades e incluir el diagrama de entidad-relación resultante.

A partir del diseño intuitivo, el equipo debe enlistar todas las dependencias funcionales y multivaluadas no triviales presentes explicando su naturaleza. A partir de estas, se deben normalizar los datos hasta cuarta forma normal justificando en cada momento las proyecciones realizadas así como el diseño final. De esta etapa también se debe presentar el diagrama de entidad-relación resultante.

Con base en el diseño en cuarta forma normal, se debe incluir los scripts pertinentes para efectuar las descomposiciones necesarias para alcanzar dicho diseño con los datos limpios. Se debe tener cuidado de no generar tuplas idénticas durante el proceso de descomposición y añadir identificadores artificiales como llave primaria a cada entidad resultante.

E) Análisis de datos a través de consultas SQL y creación de atributos analíticos

Empleado los datos normalizados, se deben crear consultas SQL (incluidas y documentadas en el repositorio) para atacar el objetivo planteado para el proyecto. Estas consultas deben utilizar los atributos naturales para crear atributos enriquecidos que sean útiles para análisis superiores a partir de filtros, agrupaciones, composiciones y funciones de ventana. En el README se deben detallar los resultados principales en forma tabular y/o gráfica explicando los hallazgos y su importancia para el proyecto.

F) Creación de APIs

Usando FastAPI se deben crear las APIs para realizar todas la operaciones CRUD en las tablas resultantes del diseño normalizado. El README debe especificar cómo se debe configurar el ambiente de Python, ejecutar el servicio e interactuar con el mismo.

Evaluación

La evaluación del proyecto se efectuará de acuerdo a los siguientes porcentajes:

Repositorio de replicación del proyecto 30%
README 40%
Presentación final 30%