

Trabajo Final – Integración de Python y PostgreSQL en el Análisis de Datos

Profesor: Yoseph Ayala Valencia

Objetivo del Trabajo Final

El propósito del trabajo final es que los alumnos demuestren su capacidad para:

- Seleccionar y justificar un dataset de Kaggle
 - Plantear una pregunta de investigación o definir una problemática de negocio que se abordará con el dataset.
 - Diseñar y justificar el esquema lógico de la base de datos a implementar en PostgreSQL.
 - Integrar Python con PostgreSQL para la carga, manipulación y consulta de datos.
 - Implementar consultas SQL avanzadas empaquetadas en funciones en Python que abarquen:
 - **Consultas Básicas y JOINS:** Ejemplos que integren datos de varias tablas.
 - **Funciones de Agregación y Agrupamiento:** Uso de SUM, COUNT, AVG, MIN, MAX, GROUP BY y HAVING.
 - **Agregación Condicional:** Uso de CASE WHEN en consultas de agregación.
 - **Funciones de Ventana y CTE:** Ejemplos que utilicen la cláusula OVER, CTE, ROW_NUMBER, RANK, LAG, etc.
 - Documentar todo el proceso en un Notebook de Jupyter, explicando desde la pregunta de investigación, el modelo lógico de la base de datos y el análisis de los resultados.
 - Realizar una presentación de impacto con los resultados encontrados.
-

1. Estructura del Proyecto

El proyecto final debe estar organizado de la siguiente manera:

```
final_project/
├── inputs/
│   └── (Archivos CSV)
├── modulos/ queries/
│   ├── __init__.py      # Archivo de inicialización (puede estar vacío)
│   └── script_queries.py #Script(s) con los queries
└── analysis/
```

| └─ Final_Analysis.ipynb # Notebook de Jupyter que documenta el proyecto, explica la problemática, modelo lógico y análisis, y llama a las funciones de `modulos/queries`

| └─ scripts/
| └─ load_data.ipynb # Notebook de Jupyter, bien documentado, para descargar y subir la información a PostgreSQL

2. Requerimientos del Trabajo

2.1 Selección y Planteamiento del Problema (Debe estar documentado en el `Final_Analysis.ipynb`)

- **Elección del Dataset:**
 - Describir brevemente el origen, la temática y la relevancia del dataset para el análisis.
- **Pregunta de Investigación / Problemática:**
 - Formular una pregunta clara y precisa. Ejemplo:
“¿Cómo influyen las variables de ventas y el comportamiento de los clientes en el desempeño general de Northwind Traders?”
o “¿Cuáles son los segmentos de clientes y productos con mayor potencial de crecimiento en función de sus ventas?”

2.2 Diseño del Esquema de la Base de Datos (Debe estar documentado en el `load_data.ipynb`)

- **Modelo Lógico:**
 - Elaborar un diagrama (por ejemplo, usando dbdiagram.io) que incluya las tablas principales y relaciones entre tablas.
 - **Justificación:**
Explicar brevemente las decisiones tomadas en el diseño del esquema y cómo ayudan a responder la problemática planteada.

2.3 Implementación en PostgreSQL y Carga de Datos (Debe estar documentado en el `load_data.ipynb`)

- **Creación de la Base de Datos y Tablas:**
 - Crear la base de datos en PostgreSQL y definir las tablas siguiendo el modelo lógico.
- **Carga de Datos:**
 - Utilizar el script `load_data.py` para automatizar el proceso de descarga y carga de los datos en las tablas de PostgreSQL.
 - El script debe estar bien documentado, explicando paso a paso la conexión, lectura y carga de datos.

2.4 Desarrollo de Consultas y Análisis en Python

- **Implementación de Querys en Python (Carpeta `modulos/queries`):**

- Funciones que contengan consultas básicas y JOINS, integrando datos de varias tablas.
- Funciones que implementen consultas con funciones de agregación (SUM, COUNT, AVG, MIN, MAX), agrupamiento con GROUP BY y HAVING, y uso de CASE WHEN para agregación condicional.
- Funciones que implementen consultas con funciones de ventana (usando OVER), CTE, ROW_NUMBER, RANK, LAG, etc.
- **Análisis en el Notebook:**
 - **En el notebook Final_Analysis.ipynb**, se debe documentar desde la **pregunta de investigación** y el **modelo lógico** hasta el **análisis** de los resultados.
 - Incluir gráficos llamativos (barras, pastel, líneas, etc.) y explicaciones detalladas de cada sección.
 - El notebook debe estar bien documentado, explicando la lógica y el propósito de cada función y resultado obtenido.

2.5 Análisis y Resultados Esperados

- **Exploración Inicial:**
 - Resumen del dataset, descripción de las tablas y análisis exploratorio inicial.
- **Interpretación de Resultados:**
 - Los alumnos deben interpretar los gráficos y resultados, explicando cómo estos insights pueden ayudar a responder la pregunta de investigación o a resolver la problemática de negocio.

3. Rubrica de Evaluación Final

Trabajo Final (60% de la nota)

Criterio	Peso (%)	Descripción
Selección y Justificación del Dataset	5	- Elección adecuada del dataset. - Descripción clara del origen, temática y relevancia para el análisis de negocio/investigación.
Pregunta de Investigación / Problemática	5	- Formulación clara y precisa de la pregunta de investigación o problemática. - Relación directa entre el problema planteado y el dataset seleccionado.
Diseño del Esquema de la Base de Datos	5	- Presentación de un modelo lógico coherente (diagrama y descripción). - Justificación de las decisiones de diseño (tablas, relaciones, etc.).
Implementación en PostgreSQL y Carga de Datos	5	- Creación correcta de la base de datos y tablas. - Evidencia de carga de datos desde Python usando load_data.ipynb con buena documentación.

Criterio	Peso (%)	Descripción
Consultas y Funcionalidades SQL en Python	15	- Implementación de consultas avanzadas empaquetadas en funciones (agregación, agrupamiento, CTE, funciones de ventana, etc.). - Calidad y claridad del código.
Notebook Final de Análisis	25	Notebook: Final_Analysis.ipynb que integre y documente todo el análisis, con explicaciones detalladas, visualizaciones llamativas y conclusiones interpretativas.

Exposición (40% de la nota)

Criterio	Peso (%)	Descripción
Claridad y Organización	20	- La exposición sigue el Principio de la Pirámide : comenzar con la idea principal, luego detallar el soporte y finalmente los detalles. - Explicación clara y convincente de los resultados obtenidos.
Interpretación y Resultados	20	- Uso de gráficos llamativos y ejemplos prácticos que respalden los insights. - Capacidad para responder a la pregunta de investigación y sugerir recomendaciones basadas en los resultados.

4. Instrucciones de Entrega

- **Formato de Entrega:**
 - **Entregar un archivo comprimido (ZIP)** que incluya la estructura completa del proyecto:
 - Carpetas: inputs/, queries/, analysis/, scripts/.
 - El Notebook Final_Analysis.ipynb con la documentación y análisis completo.
 - El script load_data.ipynb debidamente documentado.
- **Fecha Límite:**
La entrega final del archivo comprimido (ZIP) se realizará hasta el **11/03**
- **Exposición:**
 - Se hará en la última clase del curso **12/03**
 - Cada exposición podrá durar como máximo 15 minutos.
 - No es necesario que todos los integrantes expongan. Puede ser solo 1 si así desean.
 - Cada equipo realizará una exposición basada en el **Principio de la Pirámide**:
 - **Idea Principal:** Resumen de los hallazgos clave.
 - **Soporte:** Detalle de los análisis, métodos y resultados.
 - **Detalles:** Explicación de la metodología y ejemplos de código.
 - La presentación debe incluir diapositivas (powerpoint, canvas, etc).
- **Originalidad y Buenas Prácticas:**

- El código debe estar bien documentado, modularizado y seguir buenas prácticas de programación.
- La presentación deben ser claros, coherentes y bien estructurados.