Unidad 1. Evidencia de Aprendizaje 1 - Creación de una base de datos analítica

Andres Felipe Callejas

Yonier Alexis Quiceno Rodriguez

Big Data

IU Digital (Institución Universitaria Digital de Antioquia)

Colombia  
2025

Contenido

[Introducción 2](#_Toc197896041)

[Descripción del problema. 3](#_Toc197896042)

[Objetivos. 4](#_Toc197896043)

[General 4](#_Toc197896044)

[Especifico 4](#_Toc197896045)

[Descripción de los datos disponibles. 4](#_Toc197896046)

[Solución propuesta (Elección del SGBD y esquema diseñado). 5](#_Toc197896047)

[Metodología empleada. 5](#_Toc197896048)

[Resultados y Conclusiones. 6](#_Toc197896049)

[Bibliografía (Normas APA). 7](#_Toc197896050)

# Introducción

En el contexto actual del año 2025, el café colombiano continúa consolidándose como uno de los productos agrícolas más representativos y prestigiosos a nivel mundial. Su sabor, aroma y calidad han sido objeto de estudio por parte de múltiples entidades, tanto nacionales como extranjeras. Sin embargo, entender la percepción internacional sobre su consumo y preparación sigue siendo un reto, especialmente ante la creciente competencia de cafés de otras latitudes.

Este proyecto nace con el objetivo de crear un sistema de gestión de base de datos que permita almacenar, consultar y analizar los resultados de una encuesta internacional sobre hábitos de consumo de café. A través de este sistema, se pretende facilitar la toma de decisiones en estrategias de comercialización y posicionamiento del café colombiano en los mercados de exportación.

# Descripción del problema.

La necesidad de visibilizar y posicionar el café colombiano de exportación requiere información clara, estructurada y útil sobre el comportamiento del consumidor internacional. ¿Dónde lo toman? ¿Cómo lo prefieren? ¿Cuánto están dispuestos a pagar? ¿Qué tanto valoran la calidad frente al precio?

Actualmente, se cuenta con una base de datos dispersa, en inglés y poco sistematizada, lo que dificulta su análisis profundo y su uso estratégico. Por ello, se hace necesario construir una base estructurada y accesible, que facilite responder preguntas clave para el sector cafetero nacional.

# Objetivos.

## General

Desarrollar un sistema de gestión de base de datos para almacenar y analizar los resultados de una encuesta internacional sobre consumo de café colombiano tipo exportación.

## Especifico

* Traducir, estructurar y almacenar la información de la encuesta en un modelo de base de datos relacional.
* Permitir consultas rápidas sobre variables demográficas, hábitos de consumo y disposición de pago.
* Generar visualizaciones que faciliten el análisis de resultados por región, grupo etario o preferencias.

# Descripción de los datos disponibles.

Se cuenta con dos fuentes principales:

**QUESTION\_KEY.csv:** Contiene la codificación de las preguntas aplicadas en inglés, con su tipo, sección temática y posibles respuestas.

**GACTT\_RESULTS\_ANONYMIZED\_v2.csv:** Conjunto anonimizado de respuestas obtenidas en la encuesta internacional, estructurado por columnas según cada pregunta.

Estas fuentes serán transformadas a español para facilitar su uso en el contexto colombiano. La base incluye más de 100 variables relacionadas con edad, frecuencia de consumo, lugares habituales, métodos de preparación, gastos, percepción del valor y más.

# Solución propuesta (Elección del SGBD y esquema diseñado).

**Elección del SGBD:**  
Se propone el uso de **MySQL** por ser gratuito, ampliamente documentado, compatible con múltiples plataformas y soportado por herramientas de análisis como Power BI y Excel.

**Esquema de base de datos diseñado:**  
El modelo relacional consta de las siguientes tablas principales:

preguntas (id, orden, texto, tipo, sección, opciones)

respuestas (id, id\_encuesta, id\_pregunta, respuesta)

encuestas (id, fecha, pais, edad, genero, educación, estado\_laboral, hijos, afinidad\_política, etc.)

Este diseño permite almacenar tanto preguntas de selección única como múltiple, además de campos abiertos. Se incluye la metadata demográfica de cada encuestado.

# Metodología empleada.

La metodología se estructuró en cuatro fases:

**Revisión y limpieza de datos:** Identificación de preguntas relevantes, eliminación de columnas vacías o irrelevantes.

**Traducción y normalización:** Traducción técnica de los textos al español conservando el sentido original.

**Diseño del modelo relacional:** Normalización hasta 3FN para asegurar eficiencia y evitar redundancias.

**Implementación del sistema:** Creación de las tablas en MySQL, carga de datos con scripts en Python, generación de consultas clave.

Esta metodología responde directamente a los objetivos del proyecto, priorizando la usabilidad del sistema y su adaptabilidad.

# Resultados y Conclusiones.

El sistema desarrollado permite realizar análisis como:

¿Qué métodos de preparación son más usados por los jóvenes de 18 a 24 años?

¿Cuánto estaría dispuesto a pagar el consumidor promedio europeo por una taza?

¿Dónde se consume más café: en casa, en cafés o en la oficina?

Gracias a la estandarización de los datos y su integración en una base relacional, se facilita la toma de decisiones para exportadores, cafeterías de especialidad y agencias de promoción del café colombiano.

Este proyecto constituye un primer paso hacia un sistema de inteligencia de mercados centrado en el consumidor internacional.

# Bibliografía (Normas APA).

Callejas Jaramillo, A. F. (2025). Documentación académica para el desarrollo de sistemas de análisis de datos con Python y SQLite. Institución Universitaria Digital de Antioquia. Asignatura: Big Data (PREICA2501B020109).

GitHub, Inc. (2024). GitHub: Código, control de versiones y colaboración. Recuperado de https://github.com

Maven Analytics. (s.f.). GACTT Coffee Dataset. Recuperado el 11 de mayo de 2025 de https://mavenanalytics.io/data-playground?page=4&pageSize=5

Python Software Foundation. (2024). Python (Versión 3.9) [Lenguaje de programación]. Recuperado de https://www.python.org