# Proyecto de Integración y Automatización de Datos para IA para el proyecto Modernización y Automatización de la Gestión de Datos Estadísticos del DANE en Colombia

Yesid Madera[1], Wilson Baquero[2], Tatiana Pinzon[3], Jhuliana Bueno[4]

Facultad de Ingeniería y Ciencias Básicas
Universidad Central
Maestría Analítica de Datos
Automatización e Integración datos para IA
Bogotá, Colombia
{[1]*ymaderam*,[2]wbaquerog,[3]cpinzont2,[4]jbuenod}@ucentral.edu.co

November 25, 2023

## Contents

# 1 Introducción

Dentro del marco de la gestión de información del Departamento Administrativo Nacional de Estadística (DANE) en Colombia, se destaca un proyecto transversal que tiene la responsabilidad de llevar a cabo el recuento, sensibilización y recolección de datos. Este proyecto implica la recopilación de información a nivel de viviendas, hogares y personas, desplegado en diversas sedes a lo largo del país. Hasta la fecha, este proceso se ha realizado utilizando herramientas convencionales como Excel.

El propósito de nuestro proyecto consiste en modernizar esta operación mediante la implementación de un sistema de automatización e integración de datos centralizado. Esto no solo acelerará la manipulación de la información, sino que también permitirá una interacción más eficiente con herramientas analíticas como Power BI. Este avance tecnológico promete optimizar de manera significativa la gestión de datos al automatizar la recopilación y procesamiento de información, y facilitar la generación de ideas cruciales para el desarrollo y la toma de decisiones estratégicas en el ámbito estadístico del país. La automatización permitirá una mayor rapidez y precisión en la recolección de datos, además de la capacidad de integrar datos de diversas fuentes de manera más eficiente para obtener una visión más completa y actualizada de la situación estadística en Colombia.

# 2 Características del proyecto de investigación que hace uso de Integración y Automatización de Datos para IA

Para el proyecto de automatización e integración de datos, se busca explorar el uso de diversas herramientas de procesamiento con el objetivo de obtener los resultados deseados dentro del contexto que se manejará. En este caso, se analizarán tres tablas correspondientes a tres procesos diferentes que se llevan a cabo en el DANE, donde se recopila información sobre los hogares de Colombia. El propósito es construir una base de datos centralizada y automatizada que contenga diversas características que se pueden encontrar en una población.

El objetivo final de este proceso es lograr un mejor control y comprensión del comportamiento de los diferentes indicadores que se obtienen a partir de esta recopilación de datos, utilizando herramientas de automatización e integración de datos. La automatización de la recopilación, procesamiento y almacenamiento de datos permitirá gestionar de manera más eficiente el alto volumen de información que se obtiene de estos procesos. Además, facilitará el acceso a información actualizada en tiempo real y la generación de análisis más precisos y oportunos para tomar decisiones estratégicas basadas en datos en el ámbito estadístico del país.

## 2.1 Titulo del proyecto de investigación

Modernización y Automatización de la Gestión de Datos Estadísticos del DANE en Colombia.

## 2.2 Objetivo general

Implementar un sistema de modernización y automatización de la gestión de datos estadísticos en el Departamento Administrativo Nacional de Estadística (DANE) en Colombia, con el fin de optimizar la recopilación, procesamiento y análisis de información sobre viviendas, hogares y personas, permitiendo una toma de decisiones más informada y eficiente en el ámbito estadístico del país.

### 2.2.1 Objetivos especificos

- Desarrollar un sistema automatizado para la recopilación de datos en el DANE, que permita la captura eficiente y precisa de información sobre viviendas, hogares y personas, reduciendo errores humanos y acelerando el proceso de recolección.
- Implementar una base de datos centralizada que integre la información de los diferentes procesos estadísticos, asegurando la consistencia y disponibilidad de los datos en tiempo real, lo que facilitará el análisis y la generación de informes.
- Habilitar la capacidad de utilizar herramientas analíticas avanzadas, como Power BI, para analizar los datos recopilados de manera más eficiente y precisa, lo que permitirá a los analistas y tomadores de decisiones obtener insights más rápidos y relevantes para la toma de decisiones estratégicas en el ámbito estadístico del país.

## 2.3 Alcance

El alcance de este proyecto se enfoca en la implementación de una base de datos centralizada en el Departamento Administrativo Nacional de Estadística (DANE) para la dirección o área de recolección y acopio. Esto incluye la configuración, despliegue y optimización de la base de datos para la recopilación, almacenamiento y gestión eficiente de datos a nivel de viviendas, hogares y personas. Además, abarca la integración de herramientas avanzadas de procesamiento y análisis de datos para facilitar la exploración detallada. El proyecto se centra en mejorar la eficiencia y el seguimiento que se les realizan a los procesos de recuento, sensibilización y recolección para un mejor procesamiento y gestión de datos.

## 2.4 Pregunta de investigación

¿Cómo impactará la implementación de una base de datos y el uso de herramientas avanzadas de procesamiento de datos en la eficiencia y calidad de la gestión en

los procesos operativos de recuento, sensibilización y recolección de la Dirección de Recolección y Acopio del DANE en Colombia?

### 2.5 Hipotesis

La implementación de una base de datos para la dirección de recolección y acopio del DANE mejorará significativamente la eficiencia en el seguimiento, sensibilización y recolección de datos, permitiendo una mayor agilidad en los procesos.

La incorporación de herramientas avanzadas de procesamiento y análisis de datos optimizará la exploración y el análisis de los datos proporcionados por cada uno de los procesos que hacen parte de la dirección de recolección y acopio del DANE, proporcionando información más precisa y valiosa.

La centralización de la información y la optimización del seguimiento de los procesos en el DANE a través de esta base de datos contribuirá a una toma de decisiones más informada y estratégica.

## 3 Reflexiones sobre el origen de datos e información

Desde el Departamento Administrativo Nacional de Estadistica – DANE, en la Dirección de Recolección y Acopio se observa que el acceso a la información de los procesos que hacen parte de la dirección, como son el recuento, sensibilización y recolección es un poco difícil ya que esta se encuentra en Excel, esto ha dificultado el acceso y visualización de los datos por el gran volumen que se maneja.

La necesidad y los beneficios de migrar de herramientas convencionales como Excel a una base de datos centralizada, representa un salto significativo en la eficiencia y capacidad de manipulación de la información. La importancia de esta modernización para mantenerse a la vanguardia en términos de tecnología y para proporcionar una plataforma robusta para la gestión de datos.

La implementación de una base de datos centralizada puede influir en la calidad y fiabilidad de los datos recopilados; este cambio puede reducir posibles errores de entrada, duplicaciones o inconsistencias en la información. Las bases de datos pueden facilitar la validación y verificación de la información, lo que conduce a resultados más precisos y confiables.

La adopción de una nueva infraestructura de datos permitirá una interacción más eficiente con herramientas analíticas como Power BI. Esta permite la visualización de la información de forma más entendible, lo que ayuda a la generación de ideas y la toma de decisiones estratégicas en el ámbito estadístico.

## 3.1 ¿Cual es el origen de los datos e información?

El origen de los datos e información proviene de los procesos de recuento, sensibilización y recolección de la información de campo para la Gran Encuesta Integrada de Hogares – GEIH, precisamente controlada por la dirección de recolección y acopio del Departamento Administrativo Nacional de Estadística (DANE) en Colombia. Sin embargo, actualmente, esta información se encuentra almacenada en formatos convencionales como Excel. Esta limitación dificulta la visualización y manipulación de los datos debido al volumen que se maneja. La modernización hacia una base de datos centralizada es esencial para optimizar la eficiencia y fiabilidad de la información, permitiendo una interacción más efectiva con herramientas analíticas y facilitando la generación de ideas cruciales para la toma de decisiones estratégicas en el ámbito estadístico del país.

## 3.2 ¿Cuales son las consideraciones legales o éticas del uso de la información?

El uso de la información del Departamento Administrativo Nacional de Estadística (DANE) en Colombia conlleva consideraciones legales y éticas fundamentales. Desde una perspectiva legal, es crucial cumplir con las regulaciones de protección de datos y privacidad, garantizando la confidencialidad de la información de individuos y entidades. Además, se debe respetar la propiedad intelectual y los derechos de autor de los datos recopilados. Desde una perspectiva ética, se requiere transparencia en la recopilación y uso de datos, así como la obtención de consentimiento cuando sea necesario. Es esencial garantizar su uso para fines lícitos y en beneficio de la sociedad en su conjunto.

## 3.3 ¿Cuales son los retos de la información y los datos que utilizara en Integración y Automatización de Datos para IA?

Al implementar una base de datos en al Dirección de Recolección y Acopio - DRA del Departamento Administrativo Nacional de Estadística (DANE), se presentan varios desafíos en términos de calidad y consolidación de la información.

Algunos de los retos para garantizar la calidad de los datos, implica asegurar su precisión, integridad y consistencia mediante procedimientos rigurosos de entrada y validación. Además, se requiere la normalización y estandarización de la información proveniente de diversas fuentes y formatos para facilitar su integración y análisis en la base de datos. Asimismo, es esencial llevar a cabo una limpieza y enriquecimiento de los datos para corregir posibles errores, duplicaciones o datos incompletos, garantizando la coherencia y fiabilidad de la información. La consolidación de datos provenientes de distintas fuentes también representa un desafío, requiriendo la implementación de protocolos para asegurar su uniformidad. La seguridad y privacidad de los datos son

fundamentales y deben ser protegidos contra accesos no autorizados o vulnerabilidades mediante medidas de seguridad sólidas.

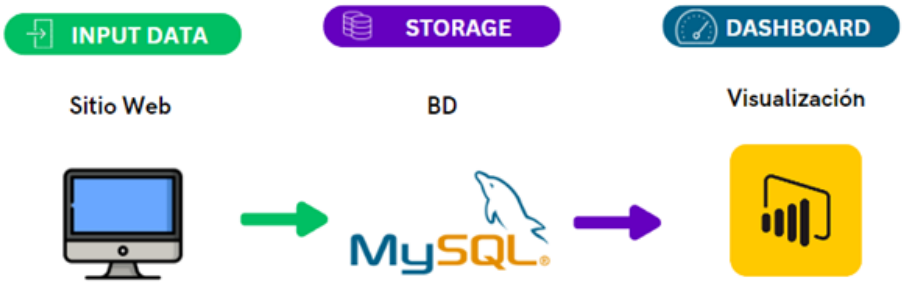
### 3.4 ¿Que espera de la utilización de Integración y Automatización de Datos para IA para su proyecto?

Esperamos que la implementación de un sistema de automatización e integración de datos para nuestro proyecto proporcione una serie de beneficios significativos. En primer lugar, anticipamos una notable mejora en la eficiencia y rapidez en la manipulación y acceso a la información, ya que la automatización permitirá gestionar grandes volúmenes de datos de manera óptima. Esto facilitará una toma de decisiones más ágil y precisa.

Además, esperamos que la normalización y estandarización de los datos faciliten su integración y comparación, proporcionando una visión más completa y detallada de los indicadores estadísticos que se manejan en la Dirección de Recolección y Acopio del DANE. Asimismo, confiamos en que la automatización de la limpieza y enriquecimiento de los datos resulte en una mayor fiabilidad y precisión en los análisis que realicemos.

La seguridad y privacidad de la información también son prioridades, y confiamos en que el sistema de automatización e integración de datos proporcionará las herramientas necesarias para garantizar la protección contra accesos no autorizados o vulnerabilidades.

## 4 Diseño de integración y Automatización de Datos para IA (Diagrama)




## 5 Integración de Datos

La captura de datos se automatiza a través de un sitio web diseñado para recoger datos con precisión, los cuales se ingresan directamente en una base de datos MySQL. Este proceso de integración de datos está automatizado para asegurar que, tras la captura, los datos se transfieran de manera fluida y sin errores a la base de datos. La relevancia de Power BI en esta etapa es que se configura para extraer directamente los datos de MySQL, usando su capacidad de conectividad nativa, lo que permite una sincronización en tiempo real de los datos para su análisis y visualización.

## 6 Automatización de Datos

La automatización se extiende al mantenimiento de la base de datos y a la generación de informes. Se configuran tareas programadas que no solo actualizan la base de datos con nueva información del sitio web, sino que también sincronizan estos datos con Power BI. Esto se logra a través de la programación de refreshers automáticos dentro de Power BI, que se activan en los horarios establecidos para reflejar los últimos cambios de datos en los dashboards de manera autónoma, facilitando un seguimiento constante y actualizado del dato.

## 7 IA

Aunque la IA no se integró en las fases iniciales del proyecto, su incorporación futura es vital para mantener la vanguardia en el análisis de datos estadísticos.

La IA también tendrá un rol en la automatización de la calidad de los datos, implementando algoritmos de detección de anomalías que pueden indicar errores de captura de datos o posibles inexactitudes. Esto permitirá realizar ajustes proactivos y mantener la alta calidad de los datos en la base de datos MySQL.

# 8 Proximos pasos

Los próximos pasos incluyen la optimización continua de estos procesos de automatización y la exploración de integraciones adicionales, como la incorporación de fuentes de datos externas y la mejora de los algoritmos de IA para un análisis más avanzado.

La implementación piloto de la base de datos centralizada y los sistemas de automatización en una sede seleccionada. Se realizarán pruebas para asegurar la eficiencia y efectividad de los procesos implementados

# 9 Lecciones aprendidas

Una de las lecciones clave ha sido la importancia de una integración y automatización efectiva entre todas las plataformas (web, MySQL, Power BI) para mantener la integridad y actualidad de los datos. El proyecto ha demostrado que el uso coordinado de estas herramientas puede resultar en una gestión de datos más eficiente y una mejor toma de decisiones basada en datos.

Otra lección clave aprendida es la importancia de una planificación detallada y la involucración de todos los stakeholders en las fases tempranas del proyecto. La adaptación a nuevas tecnologías requiere un enfoque gradual y una formación adecuada para asegurar una transición exitosa. También se destacó la relevancia de una continua evaluación y adaptación de las herramientas implementadas para satisfacer las cambiantes necesidades de gestión de datos.

# 10 Bibliografía

Russell, S. J., Norvig, P. (2016). Artificial Intelligence: A Modern Approach (3rd ed.). Pearson.

Frye, C. (2021). Microsoft Power BI Cookbook: Creating Business Intelligence Solutions of Analytical Data Models, Reports, and Dashboards. Packt Publishing.

Silberschatz, A., Korth, H. F., Sudarshan, S. (2020). Database System Concepts (7th ed.). McGraw-Hill Education.

Marz, N., Warren, J. (2015). Big Data: Principles and best practices of scalable realtime data systems. Manning Publications.

Provost, F., Fawcett, T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media.

Tjaden, B. (2017). Fundamental of Business Intelligence. Springer.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2021). An Introduction to Statistical Learning: with Applications in R (2nd ed.). Springer.

Rob, P., Coronel, C. (2018). Database Systems: Design, Implementation, Management (12th ed.). Cengage Learning.