Optimización de la Fiabilidad de Datos en las Bases del DANE: Enfoque en Sensibilización, Recuento y Recolección

Wilson Baquero[1], Yesid Madera[2], Sebastián Rios[3]

[1−2]1Dpto.Posgrados FICB,
Universidad Uexternado
Pregrado Ciencia de Datos
Curso de Bases de Datos
Bogotá, Colombia
[1]wbaquerog@ucentral.edu.co, [2]ymaderam@ucentral.edu.co, [3]sriosv1@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 los procesos de recuento, sensibilización y recolección de la información para las investigaciones o encuestas sociales. Este proyecto implica la recopilación de información a nivel de viviendas, hogares y personas, desplegado en todos los departamentos 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 una base de datos centralizada. Esto no solo acelerará la manipulación de la información, sino que también permitirá una interacción más eficiente con herramientas analíticas como Power BI, Tableau o algunas otras herramientas para visualización. Este avance tecnológico promete optimizar de manera significativa la gestión de datos y facilitar la generación de ideas cruciales para el desarrollo y la toma de decisiones estratégicas en el ámbito estadístico del país.

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

Para el proyecto de base 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 del operativo en campo de las encuestas que hacen parte del proyecto transversal. El propósito es construir una base de datos que contenga toda la información de los operativos en campo de las investigaciones sociales de la entidad. El objetivo final de este proceso es tener un mejor control y comprensión del comportamiento de los diferentes indicadores que se obtienen a partir de esta recopilación de datos, esta iniciativa tiene como fundamento mejorar la gestión de los datos que surge debido al alto volumen de información que se obtiene de estos procesos.

2.1 Titulo del proyecto de investigación

Optimización de la Fiabilidad de Datos en las Bases del DANE: Enfoque en Sensibilización, Recuento y Recolección.

2.2 Objetivo general

Modernizar y optimizar la gestión de información, mediante la implementación de una base de datos centralizada y el uso de diversas herramientas de procesamiento. Esto permitirá una recopilación más eficiente de datos a nivel de viviendas, hogares y personas, con el propósito de comprender en profundidad el comportamiento de los indicadores estadísticos relevantes. El resultado final es la generación de insights cruciales que respalden el desarrollo y la toma de decisiones estratégicas en el ámbito estadístico del país.

2.2.1 Objetivos específicos

- Implementar una base de datos centralizada y eficiente en el DANE, que permita la recopilación, almacenamiento y gestión de datos.
- Facilitar la exploración y el análisis de las características de la población mediante el uso de herramientas avanzadas de procesamiento y análisis de datos.
- Optimizar la eficiencia de los procesos de recopilación, procesamiento y gestión de datos en el DANE.

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 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. Los resultado obtenidos en el procesamiento de la información, sirve para detectar cuales son los posibles errores que se están generando a la hora de cargar los datos, con ello poder idear mejores procesos de integración de las bases, con el fin de visualizar la información de forma mas concreta y eficiente.

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 Hipótesis

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 Estadística – 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, Tableau o una herramienta de visualización. 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 principalmente de las encuestas a higares , una de ellas es 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 Bases de Datos para nuestro proyecto proporciona 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 las bases de datos permiten gestionar grandes volúmenes de datos de manera óptima. Esto permitirá una toma de decisiones más ágil y precisa.

Además, esperamos que la normalización y estandarización de los datos facilite 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, esperamos en que 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 Bases 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)

4.1 Características del SMBD (Sistema Manejador de Bases de Datos) para el proyecto

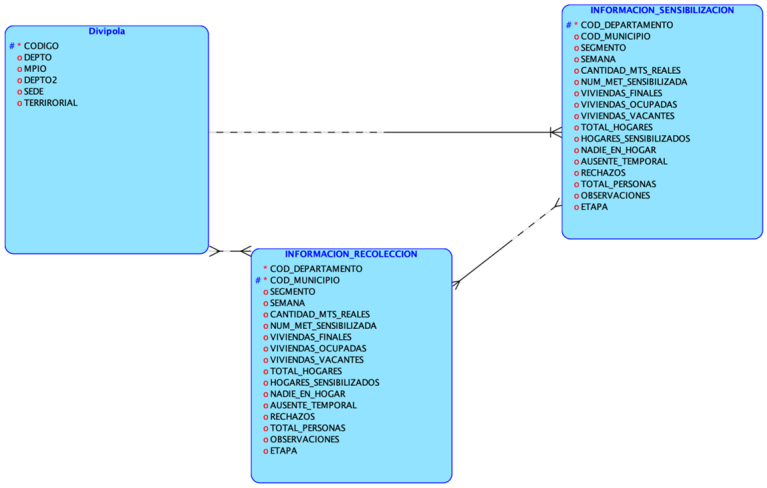
Para el presente proyecto se presentan las siguientes características en el sistema manejador de bases de datos (SMBD):

Modelo de Datos: Se utilizará un modelo de datos relacional donde los datos se organizan en tablas compuestas por filas y columnas. Cada fila representa un registro único, y cada columna representa un atributo o campo específico. Las relaciones entre las tablas se establecen mediante claves primarias y claves foráneas.

Lenguaje de Consulta: Para consultar y manipular los datos en la base de datos se utilizará el lenguaje SQL(Structured Query Language)

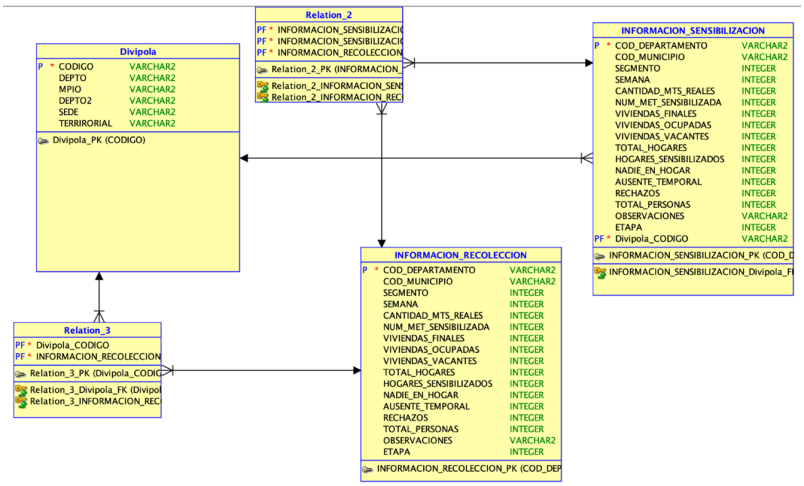
4.2 Diagrama modelo de datos

Modelo lógico :



(a) Diagrama 1

Modelo entidad relación:



(a) Diagrama 2

4.3 Imágenes de la Base de Datos

Base Divipola (vista preliminar):

	CODIGO	DEPTO	MPIO	DEPTO2	SUBSEDE	TERRITORIAL
1	15407	15	VILLA DE LEYVA	BOYACÁ	TUNJA	2 – CENTRO
2	15425	15	MACANAL	BOYACÁ	TUNJA	2 – CENTRO
3	15442	15	MARIPÍ	BOYACÁ	TUNJA	2 – CENTRO
4	15455	15	MIRAFLORES	BOYACÁ	TUNJA	2 – CENTRO
5	15464	15	MONGUA	BOYACÁ	TUNJA	2 – CENTRO
6	15466	15	MONGUÍ	BOYACÁ	TUNJA	2 – CENTRO
7	15469	15	MONIQUIRÁ	BOYACÁ	TUNJA	2 – CENTRO
8	15476	15	MOTAVITA	BOYACÁ	TUNJA	2 – CENTRO
9	15480	15	MUZO	BOYACÁ	TUNJA	2 – CENTRO
10	15491	15	NOBSA	BOYACÁ	TUNJA	2 – CENTRO

Figure 3: Tabla 1

Base Sensibilización (vista preliminar):

	ETAPA	SEDE	COD_DPTO	COD_MUNICIPIO	SEGMENTO	SEMANA	VIVI_FINALES	VIVIENDAS_OCUPADAS	VIVIENDAS_VACANTES	TOTAL_H
1	2202	BARRANQUILLA	8	001	101214	7	13	13	0	
2	2202	BARRANQUILLA	8	001	101226	8	10	10	0	
3	2202	BARRANQUILLA	8	001	101238	5	12	10	2	
4	2202	BARRANQUILLA	8	001	101250	6	10	10	0	
5	2202	BARRANQUILLA	8	001	101262	7	11	10	1	
6	2202	BARRANQUILLA	8	001	101274	8	11	10	1	
7	2202	BARRANQUILLA	8	001	101286	5	11	11	0	
8	2202	BARRANQUILLA	8	001	101298	6	11	9	2	
9	2202	BARRANQUILLA	8	001	101310	7	14	10	4	
10	2202	BARRANQUILLA	8	001	101322	8	12	10	2	

Figure 4: Tabla 2

Base Recolección (vista preliminar):

	WEEK	ID	COD_DPTO	COD_MCIPIO	CLASE	NOMBRE_DPTO	NOMBRE_MCIPIO	SEDE	ESTRATO_MUESTRAS	ESTRATO_CAMPO
1	4	(null)	63	190	2	QUINDÍO	CIRCASIA	ARMENIA	0	2
2	4	(null)	63	190	2	QUINDÍO	CIRCASIA	ARMENIA	0	2
3	4	(null)	63	190	2	QUINDÍO	CIRCASIA	ARMENIA	0	2
4	4	(null)	63	190	2	QUINDÍO	CIRCASIA	ARMENIA	0	2
5	4	(null)	63	190	2	QUINDÍO	CIRCASIA	ARMENIA	0	3
6	4	(null)	63	190	2	QUINDÍO	CIRCASIA	ARMENIA	0	2
7	4	(null)	63	190	2	QUINDÍO	CIRCASIA	ARMENIA	0	2
8	4	(null)	63	302	2	QUINDÍO	GÉNOVA	ARMENIA	0	0
9	4	(null)	63	302	2	QUINDÍO	GÉNOVA	ARMENIA	0	0
10	4	(null)	63	302	2	QUINDÍO	GÉNOVA	ARMENIA	0	2

Figure 5: tTabla 3

4.4 Código SQL - lenguaje de definición de datos (DDL)

Listing 1: Crear la tabla $\mathrm{WB}_{D}IVIPOLA$

```
CREATE TABLE WB_DIVIPOLA (
    COD             INT PRIMARY KEY,
    DEPTO           INT,
    DEPTO2          VARCHAR(100),
    SEDE            VARCHAR(100),
    TERRITORIAL     VARCHAR(100)
);
```

Listing 2: Crear la tabla WB$_R ECOLECCION$

```
CREATE TABLE WB_RECOLECCION(
    cod_departamento            VARCHAR2,
    cod_municipio               VARCHAR2,
    segmento                    INTEGER,
    semana                      INTEGER,
    cantidad_mts_reales         INTEGER,
    num_met_sensibilizada       INTEGER,
    viviendas_finales           INTEGER,
    viviendas_ocupadas          INTEGER,
    viviendas_vacantes          INTEGER,
    total_hogares               INTEGER,
    hogares_sensibilizados      INTEGER,
    nadie_en_hogar              INTEGER,
    ausente_temporal            INTEGER,
    rechazos                    INTEGER,
    total_personas              INTEGER,
    observaciones               VARCHAR2,
    etapa                       INTEGER
    );
```

Listing 3: Crear la tabla WB$_S ENSIBILIZACION$

```
CREATE TABLE WB_SENSIBILIZACION  (
    cod_departamento            VARCHAR2,
    cod_municipio               VARCHAR2,
    segmento                    INTEGER,
    semana                      INTEGER,
    cantidad_mts_reales         INTEGER,
    num_met_sensibilizada       INTEGER,
    viviendas_finales           INTEGER,
    viviendas_ocupadas          INTEGER,
    viviendas_vacantes          INTEGER,
    total_hogares               INTEGER,
    hogares_sensibilizados      INTEGER,
    nadie_en_hogar              INTEGER,
    ausente_temporal            INTEGER,
    rechazos                    INTEGER,
    total_personas              INTEGER,
    observaciones               VARCHAR2,
    etapa                       INTEGER,
    divipola_codigo             VARCHAR2
);
```

4.5 Código SQL - Manipulación de datos (DML)

- INSERTAR DATOS EN LAS TABLAS:

```
-- Actualizar datos en la tabla WB_DIVIPOLA
UPDATE WB_DIVIPOLA
SET TERRITORIAL = 'Territorial Actualizado'
WHERE COD = 1;

-- Actualizar datos en la tabla WB_RECOLECCION
UPDATE WB_RECOLECCION
SET observaciones = 'Nueva observaci n'
WHERE cod_departamento = '101' AND cod_municipio = '001';

-- Actualizar datos en la tabla WB_SENSIBILIZACION
UPDATE WB_SENSIBILIZACION
SET etapa = 5
WHERE cod_departamento = '101' AND cod_municipio = '002';
```

- ELIMINACIÓN DE REGISTROS DE LAS TABLAS :

```
-- Eliminar un registro de la tabla WB_DIVIPOLA
DELETE FROM WB_DIVIPOLA
WHERE COD = 1;

-- Eliminar un registro de la tabla WB_RECOLECCION
DELETE FROM WB_RECOLECCION
WHERE cod_departamento = '101' AND cod_municipio = '001';

-- Eliminar un registro de la tabla WB_SENSIBILIZACION
DELETE FROM WB_SENSIBILIZACION
WHERE cod_departamento = '101' AND cod_municipio = '002';
```

4.6 Código SQL + Resultados: Vistas

La vista "Vista_Combinada" combina datos de las tres tablas utilizando una consulta SQL que realiza una unión ("JOIN") entre las tablas "WB_DIVIPOLA", "WB_RECOLECCION" y "WB_SENSIBILIZACION" utilizando el campo "COD" como clave de unión entre "WB_DIVIPOLA" y las otras dos tablas.

Listing 4: Crear una vista que combina datos de $\mathrm{WB}_{D}IVIPOLA$

```
CREATE VIEW Vista_Combinada AS
SELECT
    DIV.COD AS Divipola_COD,
    DIV.DEPTO AS Divipola_Depto,
    DIV.DEPTO2 AS Divipola_Depto2,
    DIV.SEDE AS Divipola_SEDE,
    DIV.TERRITORIAL AS Divipola_TERRITORIAL,
    REC.cod_departamento AS Recoleccion_CodDepartamento,
    REC.cod_municipio AS Recoleccion_CodMunicipio,
    REC.segmento AS Recoleccion_Segmento,
    REC.semana AS Recoleccion_Semana,
    REC.cantidad_mts_reales AS Recoleccion_CantidadMtsReales,
```

```
    REC.num_met_sensibilizada AS Recoleccion_NumMetSensibilizada,
    REC.viviendas_finales AS Recoleccion_ViviendasFinales,
    REC.viviendas_ocupadas AS Recoleccion_ViviendasOcupadas,
    REC.viviendas_vacantes AS Recoleccion_ViviendasVacantes,
    REC.total_hogares AS Recoleccion_TotalHogares,
    REC.hogares_sensibilizados AS Recoleccion_HogaresSensibilizados,
    REC.nadie_en_hogar AS Recoleccion_NadieEnHogar,
    REC.ausente_temporal AS Recoleccion_AusenteTemporal,
    REC.rechazos AS Recoleccion_Rechazos,
    REC.total_personas AS Recoleccion_TotalPersonas,
    REC.observaciones AS Recoleccion_Observaciones,
    REC.etapa AS Recoleccion_Etapa,
    SEN.etapa AS Sensibilizacion_Etapa,
    SEN.divipola_codigo AS Sensibilizacion_DivipolaCodigo
FROM
    WB_DIVIPOLA DIV
JOIN
    WB_RECOLECCION REC
ON
    DIV.COD = REC.cod_departamento
JOIN
    WB_SENSIBILIZACION SEN
ON
    DIV.COD = SEN.cod_departamento;
```

4.7 Código SQL + Resultados: Triggers

Se crea un trigger con la finalidad de realizar un seguimiento automático de los nuevos datos qué se inserten en la tabla "WB_RECOLECCION" y copiar los datos relevantes en la tabla "WB_SENSIBILIZACION" cuando se inserte un nuevo registro en "WB_RECOLECCION, Este trigger se activará automáticamente después de cada inserción en la tabla "WB_RECOLECCION" (AFTER INSERT ON WB_RECOLECCION). Copiará los datos relevantes de la inserción a la tabla "WB_SENSIBILIZACION" y registrará la inserción en una tabla de registro llamada "RegistroInserciones."

```
-- Crear una tabla de registro para seguimiento de inserciones
CREATE TABLE RegistroInserciones (
    Id INT PRIMARY KEY,
    FechaInsercion TIMESTAMP,
    Detalles VARCHAR(255)
);


-- Crear un trigger que copia datos de WB_RECOLECCION a WB_SENSIBILIZACION
CREATE OR REPLACE TRIGGER Trigger_CopiaDatos
AFTER INSERT ON WB_RECOLECCION
FOR EACH ROW
BEGIN
    INSERT INTO WB_SENSIBILIZACION (
        cod_departamento,
        cod_municipio,
        segmento,
        semana,
```

```
        cantidad_mts_reales,
        num_met_sensibilizada,
        viviendas_finales,
        viviendas_ocupadas,
        viviendas_vacantes,
        total_hogares,
        hogares_sensibilizados,
        nadie_en_hogar,
        ausente_temporal,
        rechazos,
        total_personas,
        observaciones,
        etapa,
        divipola_codigo
    ) VALUES (
        :NEW.cod_departamento,
        :NEW.cod_municipio,
        :NEW.segmento,
        :NEW.semana,
        :NEW.cantidad_mts_reales,
        :NEW.num_met_sensibilizada,
        :NEW.viviendas_finales,
        :NEW.viviendas_ocupadas,
        :NEW.viviendas_vacantes,
        :NEW.total_hogares,
        :NEW.hogares_sensibilizados,
        :NEW.nadie_en_hogar,
        :NEW.ausente_temporal,
        :NEW.rechazos,
        :NEW.total_personas,
        :NEW.observaciones,
        :NEW.etapa,
        :NEW.divipola_codigo
    );

    Registrar la inserci n en el registro de inserciones
    INSERT INTO RegistroInserciones (Id, FechaInsercion, Detalles)
    VALUES (:NEW.Id, SYSDATE, 'Nuevo registro insertado en WB_RECOLECCION');
END;
```

4.8 Código SQL + Resultados: Funciones

Se crea una función que calcula el promedio de la cantidad de viviendas finales en la tabla "WB_RECOLECCION" para un departamento y municipio específico.

```
Crear una funci n que calcule el promedio de
viviendas finales
CREATE OR REPLACE FUNCTION CalcularPromedioViviendasFinales(
    departamento VARCHAR2,
    municipio VARCHAR2
)
```

```
RETURN DECIMAL
IS
    total_viviendas DECIMAL;
    total_registros INTEGER;
    promedio DECIMAL;
BEGIN
    SELECT SUM(viviendas_finales), COUNT(*) INTO total_viviendas, total_registros
    FROM WB_RECOLECCION
    WHERE cod_departamento = departamento AND cod_municipio = municipio;

    IF total_registros = 0 THEN
        RETURN 0;
        Evitar la divisi n por cero si no hay registros
    END IF;

    promedio := total_viviendas / total_registros;
    RETURN promedio;
END CalcularPromedioViviendasFinales;

Utilizar la funci n para calcular el promedio de
viviendas finales para un departamento y municipio
SELECT CalcularPromedioViviendasFinales('Bogot ', 'Bogot -Municipio')
FROM dual;
```

4.9 Código SQL + Resultados: procedimientos almacenados

Se crea un procedimiento almacenado que utiliza las tablas "WB_RECOLECCION" y "WB_SENSIBILIZACION" el procedimiento calcula y muestra el total de viviendas finales para un departamento y municipio específicos:

```
-- Crear un procedimiento para calcular y mostrar el total de viviendas finales
CREATE OR REPLACE PROCEDURE CalcularTotalViviendasFinales(
    departamento_in VARCHAR2,
    municipio_in VARCHAR2
)
AS
    total_viviendas NUMBER;
BEGIN
    SELECT SUM(viviendas_finales) INTO total_viviendas
    FROM WB_RECOLECCION
    WHERE cod_departamento = departamento_in AND cod_municipio = municipio_in;

    DBMS_OUTPUT.PUT_LINE('El -total -de -viviendas -finales -para -'
    || departamento_in
    || ',-'
    || municipio_in
    || '-es:-'
    || total_viviendas);
END;
```

5 Bases de Datos No-SQL

5.1 Diagrama Bases de Datos No-SQL

Meta-Modelo Conceptual:

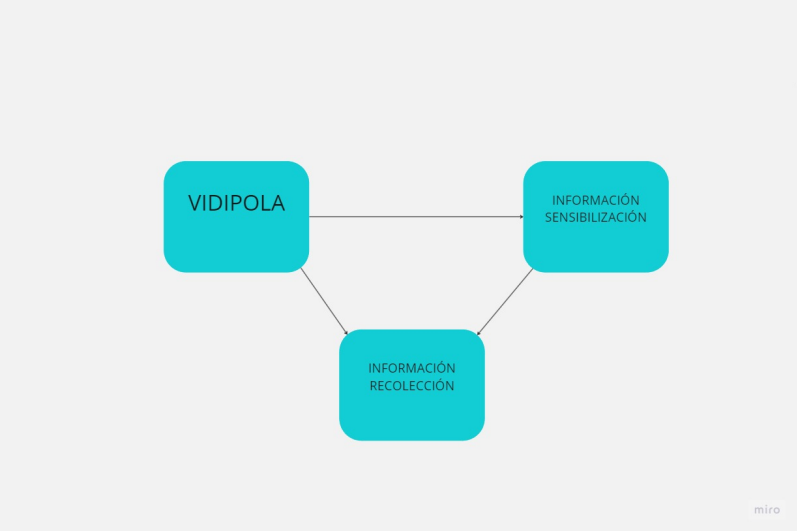


Figure 6: Diagrama 1

Meta-Modelo Lógico:

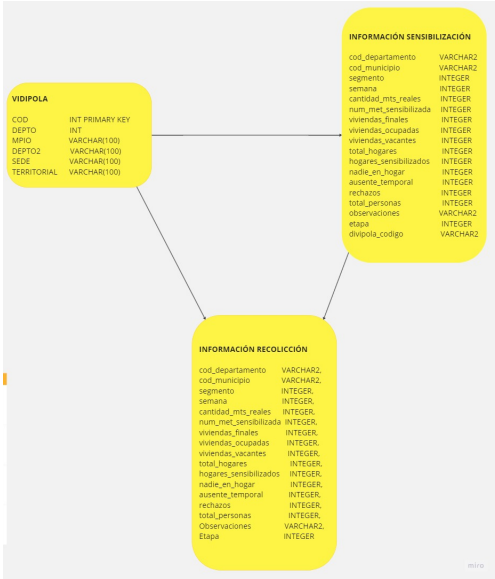


Figure 7: Diagrama 2

Meta-Modelo Físico:

Figure 8: Diagrama 3

5.2 SMBD utilizado para la Base de Datos No-SQL

Se utilizo el sistema de datos NoSql orientado a documentos MongoDB Se realizo la creación de una base de datos PROYECTO BASES DATOS en la cual se crea una colección para cada una de las tres tablas. **Meta-Modelo Físico:**

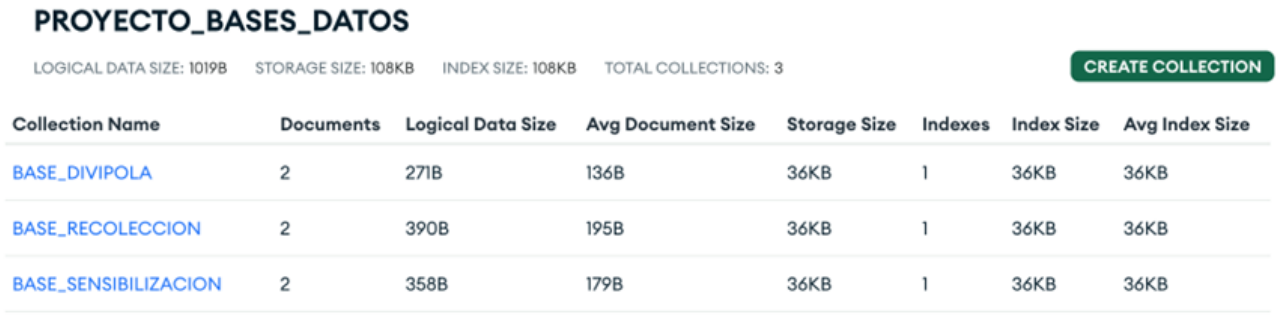


Figure 9: Diagrama 3

En cada colección se realizó la creación de 2 documentos con la información de las respectivas tablas

BASE DIVIPOLA:

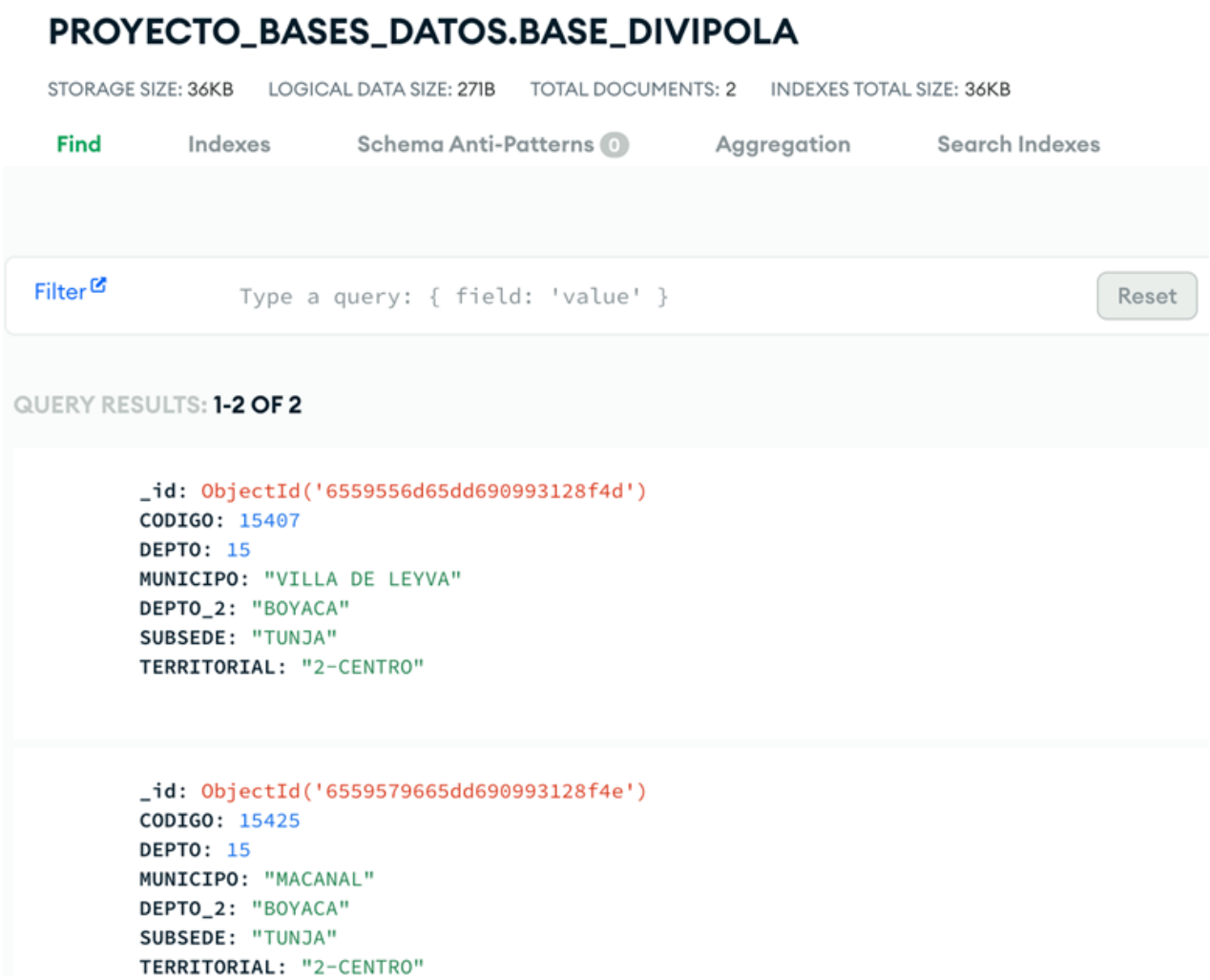


Figure 10: Tabla mongo

BASE RECOLECCIÓN:

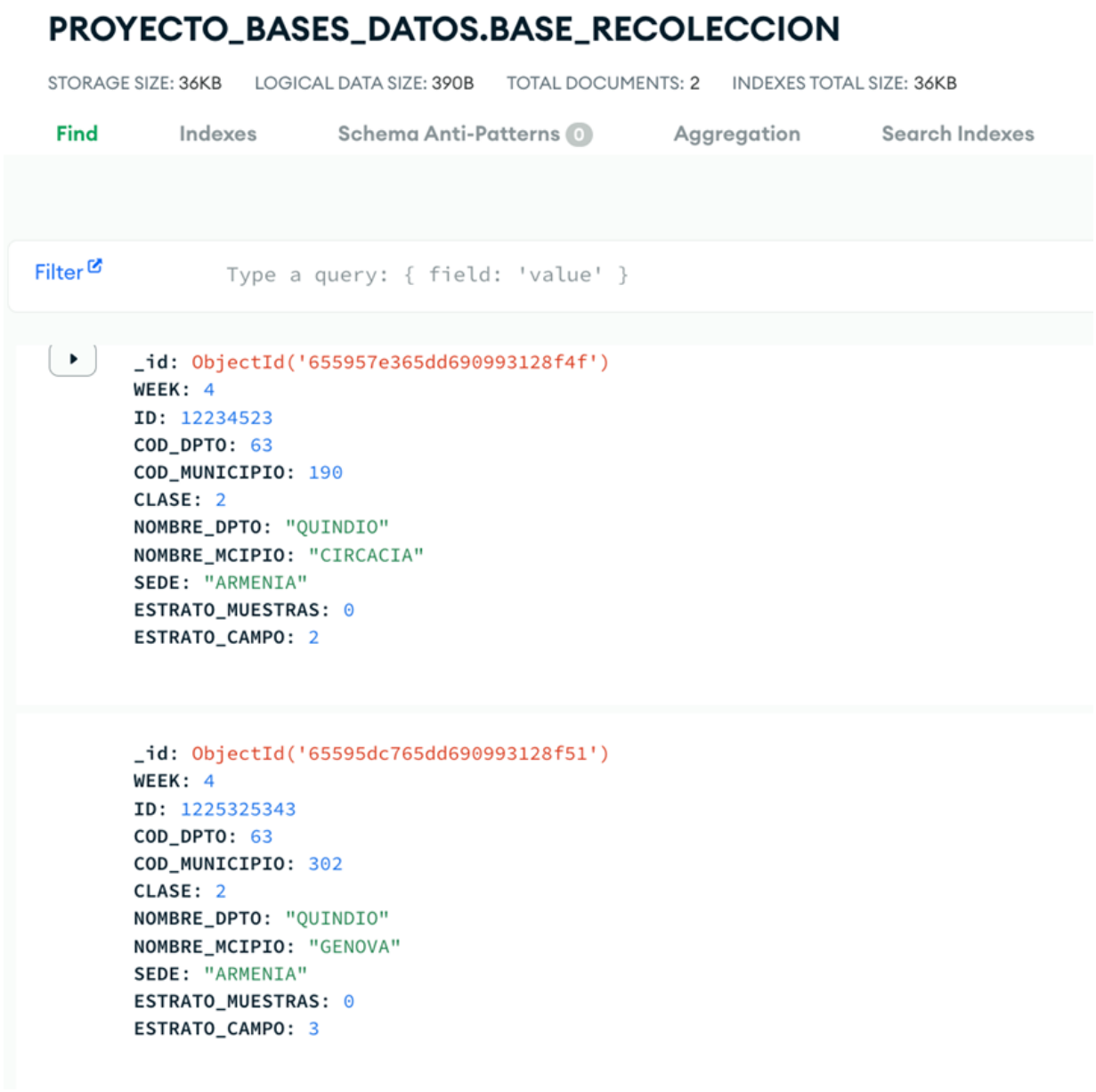


Figure 11: Tabla Recolección

BASE SENSIBILIZACION:

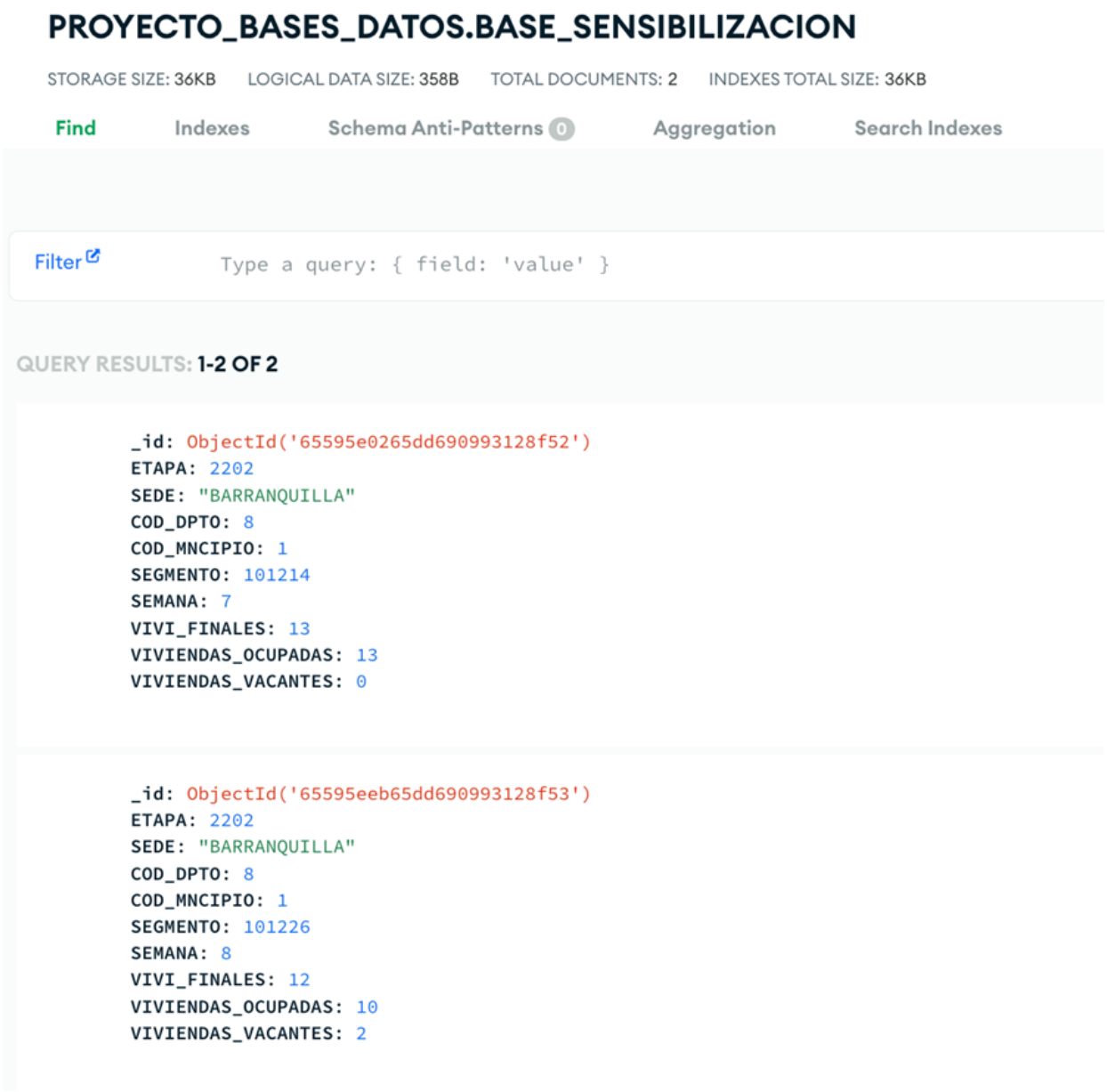


Figure 12: Tabla sensibilización

6 Aplicación de ETL (Extract, Transform, Load) y Bodega de Datos

6.1 Ejemplo de aplicación de ETL y Bodega de Datos

Proceso de extracción :

El proceso de extracción inicia con la obtención de datos e información mediante los procedimientos de recuento, sensibilización y recolección de datos de campo para la Gran Encuesta Integrada de Hogares – GEIH, supervisada por la dirección de recolección y acopio del Departamento Administrativo Nacional de Estadística (DANE) en Colombia. Desde este proceso, se derivan tres tablas fundamentales: Divipola, que contiene información sobre la ubicación geográfica; Sensibilización, que registra datos relevantes acerca de la concientización en el proceso; y Recolección, que almacena información específica recopilada durante el proceso de recolección de datos

Proceso de transformación:

Para las tres fuentes de datos se realizan los siguientes procesos de transformación

- Limpieza de Datos: Se eliminan o corrigen datos incorrectos, incompletos o inconsistentes. Esto puede incluir la corrección de errores tipográficos, la estandarización de formatos y la identificación y manejo de valores nulos.
- Transformación de Datos: Los datos pueden se transforman para adaptarse al formato requerido.
- Integración de Datos: Los datos provienen de múltiples fuentes por lo que se deba integrar la información.
- Deduplicación: Se eliminan duplicados de los datos para garantizar la consistencia y la integridad.
- Validación de Datos: Se verifica la integridad y la coherencia de los datos transformados para asegurarse de que cumplan con los estándares y las reglas de negocio establecidas.

6.2 Automatización de Datos

La optimización del proceso de gestión de datos fue significativa al migrar de un enfoque manual basado en Excel a una solución automatizada implementada en Oracle Developer. Este cambio podría representar mejoras sustanciales en varios aspectos clave del flujo de trabajo, destacando los siguientes puntos:

- Eliminación de Tareas Manuales: La transición a Oracle Developer permitió eliminar las tareas manuales que antes implicaban la manipulación de datos en hojas de cálculo de Excel. Las acciones repetitivas, propensas a errores y que consumían tiempo, como la entrada manual y la actualización de datos, fueron reemplazadas por procesos automáticos.
- Integración de Datos Simplificada: La automatización permitió una integración más sencilla de datos desde diversas fuentes. Oracle Developer facilita la conexión y la importación de datos de manera eficiente, contribuyendo a una gestión más efectiva de la información.
- Mayor Escalabilidad: La solución implementada en Oracle Developer es altamente escalable, lo que significa que puede adaptarse fácilmente a cambios en la escala de operaciones sin requerir un aumento significativo en los recursos humanos. Esto es fundamental para mantener la eficiencia a medida que los volúmenes de datos evolucionan.

la migración a Oracle Developer ha llevado a una transformación significativa en la eficiencia y la confiabilidad de la gestión de datos. Al eliminar la dependencia de procesos manuales y al aprovechar la automatización, se han logrado mejoras notables en la velocidad, la precisión y la escalabilidad de las operaciones relacionadas con los datos.

6.3 Integración de Datos

La integración de datos en este caso de uso específico tiene como objetivo unificar la información dispersa en las tablas Divipola, Sensibilización y Recolección, facilitando así un análisis más completo y una toma de decisiones informada en el ámbito de la Gran Encuesta Integrada de Hogares en Colombia. La integración de datos para este caso se llevo de la siguiente manera:

- Consolidación de Datos: Las tablas Divipola, Sensibilización y Recolección contienen información valiosa sobre la ubicación geográfica, el proceso de sensibilización y los datos recopilados durante la encuesta, respectivamente. La integración implica consolidar estos conjuntos de datos en una única fuente centralizada.
- Transformación y Normalización: Dado que los datos pueden provenir de diferentes formatos y estructuras, es crucial normalizarlos durante la integración. Esto puede implicar la estandarización de esquemas, la conversión de formatos de fechas o la unificación de unidades de medida para garantizar coherencia.
- Resolución de Duplicados: La integración aborda la posibilidad de duplicados o inconsistencias en los datos. Se implementan mecanismos para identificar y resolver duplicados, asegurando la integridad y calidad de la información resultante.

7 Proximos pasos

Los siguientes pasos se centraron en poner en marcha el proceso implementado en el DANE y en agregar un componente adicional dirigido hacia la data visualización. A través de esta iniciativa, se busca integrar el modelo de datos establecido en Oracle con una herramienta de visualización, tal como Power BI. Esto permite obtener perspectivas valiosas sobre el proceso, posibilitando la toma de decisiones fundamentadas mediante la interpretación visual de los datos. Este enfoque no solo contribuirá a la eficacia operativa, sino que también facilitará la identificación de tendencias y patrones, promoviendo así una toma de decisiones más informada y estratégica.

8 Lecciones aprendidas

Una gestión efectiva y un modelo de datos sólido son esenciales para el éxito de una empresa o proceso. Se ha reconocido que contar con un modelo de datos bien estructurado no solo facilita la toma de decisiones informada, sino que también contribuye significativamente a la eficiencia operativa, la calidad de los datos, el cumplimiento normativo y la capacidad de adaptación frente a cambios. Esta experiencia destaca la importancia de invertir tiempo y recursos en la gestión adecuada de datos y en el diseño de un modelo que respalde las necesidades presentes y futuras de la organización.

9 Bibliografía

- *(S/f). Oracle.com. Recuperado el 24 de noviembre de 2023, de* `https://education.oracle.com/es/oracle-database-introduccin-a-sqlplsql/courP_3110`