

# Proyecto de Bases de datos Clase Final

John Garcia<sup>1</sup>, Andres Velez<sup>2</sup>,

<sup>1-2</sup>Ingeniería y Ciencias Básicas, <sup>3</sup>Dpto. de Ingeniería,  
Universidad Central

Maestría en Analítica de Datos

Curso de Bases de Datos

Bogotá, Colombia

{<sup>1</sup>JohnDavidGarcia,<sup>2</sup>AndresVelez}jgarcia9@ucentral.edu.co,<sup>3</sup>avelezo@ucentral.edu.co

October 8, 2022

## Contents

<b>1</b>	<b>Introducción (Max 250 Palabras) - (<i>Primera entrega</i>)</b>	<b>3</b>
<b>2</b>	<b>Características del proyecto de investigación (Max 500 Palabras) - (<i>Primera entrega</i>)</b>	<b>3</b>
2.1	Titulo del proyecto de investigación () Primera entrega) . . . . .	3
2.2	Objetivo general (Max 100 Palabras) - ( <i>Primera entrega</i> ) . . . . .	3
2.2.1	Objetivos especificos (Max 100 Palabras) - ( <i>Primera entrega</i> )	3
2.3	Alcance (Max 200 Palabras) - ( <i>Primera entrega</i> ) . . . . .	4
2.4	Pregunta de investigación (Max 100 Palabras) - ( <i>Primera entrega</i> ) .	4
2.5	Hipotesis (Max 100 Palabras) - ( <i>Primera entrega</i> ) . . . . .	4
<b>3</b>	<b>Reflexiones sobre el origen de datos e información (Max 400 Palabras) - (<i>Primera entrega</i>)</b>	<b>5</b>
3.1	¿Cual es el origen de los datos e información ? (Max 100 Palabras) - ( <i>Primera entrega</i> ) . . . . .	5
3.2	¿Cuales son las consideraciones legales o eticas del uso de la información? (Max 100 Palabras) - ( <i>Primera entrega</i> ) . . . . .	6
3.3	¿Cuales son los retos de la información y los datos que utilizara en la base de datos en terminos de la calidad y la consolidación? (Max 100 Palabras) - ( <i>Primera entrega</i> ) . . . . .	6
3.4	¿Que espera de la utilización de un sistema de Bases de Datos para su proyecto? (Max 100 Palabras) - ( <i>Primera entrega</i> ) . . . . .	6

<b>4</b>	<b>Diseño del Modelo de Datos del SMBD (Sistema Manejador de Bases de Datos)(Primera entrega)</b>	<b>7</b>
4.1	Características del SMBD (Sistema Manejador de Bases de Datos) para el proyecto (Primera entrega) . . . . .	7
4.2	Diagrama modelo de datos (Primera entrega) . . . . .	8
4.3	Imágenes de la Base de Datos (Primera entrega) . . . . .	9
4.4	Código SQL - lenguaje de definición de datos (DDL) (Primera entrega) . . . . .	9
4.5	Código SQL - Manipulación de datos (DML) (Primera entrega) . .	9
4.6	Código SQL + Resultados: Vistas (Primera entrega) . . . . .	10
4.7	Código SQL + Resultados: Triggers (Primera entrega) . . . . .	11
4.8	Código SQL + Resultados: Funciones (Primera entrega) . . . . .	11
4.9	Código SQL + Resultados: procedimientos almacenados (Primera entrega) . . . . .	12
<b>5</b>	<b>Bases de Datos No-SQL (Segunda entrega)</b>	<b>13</b>
5.1	Diagrama Bases de Datos No-SQL (Segunda entrega) . . . . .	13
5.2	SMBD utilizado para la Base de Datos No-SQL (Segunda entrega)	13
<b>6</b>	<b>Aplicación de ETL (Extract, Transform, Load) y Bodega de Datos (Tercera entrega)</b>	<b>14</b>
6.1	Ejemplo de aplicación de ETL y Bodega de Datos (Tercera entrega)	14
<b>7</b>	<b>Lecciones aprendidas (Tercera entrega)</b>	<b>15</b>
<b>8</b>	<b>Bibliografía</b>	<b>16</b>

## **1 Introducción (Max 250 Palabras) - (Primera entrega)**

Para el año 2021 los estudiantes matriculados en Colombia fueron 9.797.677 alumnos, cifra inferior al año 2020 con 9.712.511 alumnos, muestra una disminución de 85.166 estudiantes (DANE,2021). El presupuesto asignado para educación en el país fue de 49,4 billones de pesos de los 350 billones de pesos aprobados del Presupuesto General de la Nación (DNP,2021).

De estos 49,4 billones el 89 por ciento se destinan para funcionamiento, es decir 43,9 billones deben ser usados para que 9.797.677 estudiantes e instituciones universitarias, brinden garantías y den resultados óptimos de calidad, algo injusto en algunos aspectos teniendo en cuenta los años de atraso en materia de infraestructura, en actualización de contenidos, de mejoras en los accesos a la educación, en garantías para docentes, entre tantas otras.

De igual forma, se complica aún más el acceso y las garantías cuando en un país como Colombia, se presenta casos como el de Centro Poblados en el cual se perdieron 70 mil millones de pesos (Revista Portafolio, 2021), destinados para para garantizar el acceso a internet en zonas alejadas del país, o el PAE (plan de alimentación Escolar), destinado a brindar una alimentación adecuada a los niños de bajos recursos.

Teniendo en cuenta este panorama, se hace un poco más complicado garantizar un adecuado acceso a una educación de calidad en zonas excluidas históricamente, con presencia de conflicto, con carencias generales y alejadas del bogo centrismo que gobierna Colombia, Departamentos ubicados en la región caribe (Atlántico, Bolívar, Cesar, Córdoba, Sucre, La Guajira, Magdalena, San Andrés y Providencia)

## **2 Características del proyecto de investigación (Max 500 Palabras) - (Primera entrega)**

### **2.1 Título del proyecto de investigación () Primera entrega)**

EL IMPACTO EN LOS REGISTROS EDUCATIVOS DE LA REGIÓN CARIBE CON BASE EN LA VIOLENCIA Y CORRUPCIÓN, EN COMPARACIÓN CON EL RESTO DEL PAÍS

### **2.2 Objetivo general (Max 100 Palabras) - (Primera entrega)**

Determinar el aporte realizado en Cobertura Neta en la Región Caribe para los años 2019 al 2021, en comparación con la Cobertura Neta del resto del País.

#### **2.2.1 Objetivos especificos (Max 100 Palabras) - (Primera entrega)**

- Determinar el departamento que menor Cobertura Neta aporta a cada nivel educativo de la Región Caribe.
- Definir el departamento que tiene la menor tasa de reprobación aporta a cada nivel educativo de la Región Caribe.

- Identificar qué departamento presenta la tasa más alta de deserción en cada nivel educativo de la Región Caribe.
- Identificar qué departamento presenta la menor tasa de repitencia en cada nivel educativo de la Región Caribe.
- Determinar qué departamento presenta la tasa más alta de aprobación en cada nivel educativo de la Región Caribe..

### **2.3 Alcance (Max 200 Palabras) - (*Primera entrega*)**

La educación colombiana se mantiene en vilo no solo por los coletazos que ha dejado la pandemia del año 2020 por un déficit histórico que tiene el país y por un giro a la izquierda en las políticas nacionales con el cambio de gobierno en el año 2022, porque, aunque el gobierno saliente indica que la cobertura a nivel general “aumento”, en contraposición un estudio brindado por el DANE indica que aun hacen falta 206.260 mil cupos que deberían ayudar a garantizar el acceso a la educación de los 793.311 niños hasta los 5 años. El presente escrito busca determinar cuál es el porcentaje de Cobertura Neta en los niveles de transición, básica primaria, básica secundaria y la educación media en la región Caribe, en comparación con el porcentaje de Cobertura Neta Nacional, para los años 2019 al 2021, estableciendo como influyen variables como la corrupción y la violencia en el desarrollo académico de los niños de esta región.

### **2.4 Pregunta de investigación (Max 100 Palabras) - (*Primera entrega*)**

¿La Cobertura Neta en los niveles de estudio para transición, básica primaria, básica secundaria y educación media en la región Caribe es mayor en su conjunto que la Cobertura Neta del resto de departamentos de Colombia para los años 2019 a 2021?

### **2.5 Hipotesis (Max 100 Palabras) - (*Primera entrega*)**

Por Antonomasia se considera que los departamentos, las regiones, los lugares en los cuales se presentan mayores niveles de corrupción, de violencia, de carencias, son esos lugares donde la niñez tiene una menor proyección académica, pero que sucedería si los números, las estadísticas rompieran la tendencia o esta falsa creencia. Por medio de este escrito se procurará comprobar con el uso de bases de datos académicas, de violencia y de pobreza. Sí los estudiantes de la Región Caribe presentan números más bajos en calidad de aprobación en sus niveles educativos de educación básica en comparación con el resto del País.

### 3 Reflexiones sobre el origen de datos e información

(Max 400 Palabras) - (*Primera entrega*)

La fuente de datos de la operación es el registro realizado, normalmente, por los rectores de las sedes y establecimientos educativos de las Entidades Territoriales Certificadas (ETC) en el Sistema Integrado de Matrícula (SIMAT), es decir, que la información plasmada en la base de datos que presenta el Ministerio de Educación Nacional (MEN), debe estar actualizada y debe ser lo más fehaciente posible.

Esta base de datos se comenzó a publicar en el año 2016, es de carácter anualizado y se va actualizando con base en la información que se va agregando al SIMAT por medio de las Entidades Territoriales Certificadas, el rector de establecimiento educativo o el personal administrativo responsable de la entidad territorial.

De igual forma, se usan 3 bases de datos adicionales para darle continuidad a los objetivos, Pobreza, Conflicto armado y seguridad ciudadana, Educación, las cuales son brindadas por la página de la Dirección Nacional de Planeación (DNP) y su aplicativo TERRIDATA, en la cual reposan los datos brindados y los cuales son de carácter público.

Dentro del proceso de recolección y publicación de la información se debe realizar, el proceso de planeación de las entidades, sus capacidades institucionales, así como las proyecciones de cupos estudiantiles para el siguiente año, de igual forma, se deben tener en cuenta la cantidad de solicitudes, la asignación de los cupos educativos, la cantidad real de matriculados, lo cual deriva en una auditoria directa sobre la información brindada por cada ETC, como lo establece la Resolución 7797 de 2015 del Ministerio de Educación.

Realizado el proceso de recolección de la información se procede a desagregar de forma regional, luego departamental, municipal, zonal (sea urbano o rural), posteriormente por sedes y por establecimiento. De igual forma, se desagrega por sexo, grado, nivel educativo (transición, básica primaria, básica secundaria y media), el sector (público o privado), la zona, la jornada (mañana, tarde, nocturna) y por modelo educativo.

Esta forma tan detallada de recolectar y desagregar la información debería mostrarle al país los datos más precisos y exactos posibles, pero como no es del todo posible acceder a ciertas zonas de la geografía nacional o en algunas ocasiones la información suministrada en el SIMAT es manipulada por los mismos encargados de brindarla, resulta muy complicado y aunque es la más acercada a la realidad no implica que no se presenten fallos en la información consignada o que haya fallos de juicio al momento de desagregarla.

#### 3.1 ¿Cual es el origen de los datos e información ? (Max 100 Palabras) - (*Primera entrega*)

Se toma como base la información de DATOS ABIERTOS acerca de las estadísticas de educación en los niveles de preescolar, primaria y media, brindadas por el ministerio de educación de Colombia, para los años 2019 y 2021, de igual forma

se usan los datos para violencia y pobreza brindados por TERRIDATA de la Dirección Nacional de planeación (DNP).

### **3.2 ¿Cuales son las consideraciones legales o eticas del uso de la información? (Max 100 Palabras) - (Primera entrega)**

Al tratarse de un tema tan delicado como la educación en menores de edad, se presenta el dilema legal y ético por el uso de la información, pero al tratarse de un conjunto de datos, en los cuales no se reflejen datos personales, sino el manejo de cifras bases de datos públicas y al no contener información sensible, confidencial o que vulnere los derechos de los menores de edad, no se estaría incurriendo en una agresión a los estatutos de manejo de información ni a la ética

### **3.3 ¿Cuales son los retos de la información y los datos que utilizara en la base de datos en terminos de la calidad y la consolidación? (Max 100 Palabras) - (Primera entrega)**

Al realizar el proceso de combinar la base de datos principal de educación, con las bases de datos de Pobreza y Violencia se pueda consolidar la información requerida para poder observar en que medida afectan estos dos detonadores influyen directamente en el rendimiento y los estándares de Educación de la Región Caribe en comparación con otras zonas que hayan sufrido en mayor, menor o igual medida el impacto de estos dos detonantes.

### **3.4 ¿Que espera de la utilización de un sistema de Bases de Datos para su proyecto? (Max 100 Palabras) - (Primera entrega)**

Se espera que el uso de MySQL facilite no solo la consolidación de la información, sino la forma en que se pueda estructurar, dando un margen de mejora de esta, sin tener que dar inicio desde ceros a todo el proceso. De igual forma, se espera que al ser un sistema de gestión permita una mejor lectura e interpretación de las de datos los datos usados para el estudio.

## 4 Diseño del Modelo de Datos del SMBD (Sistema Manejador de Bases de Datos) *(Primera entrega)*

### 4.1 Características del SMBD (Sistema Manejador de Bases de Datos) para el proyecto *(Primera entrega)*

MySQL es un sistema de gestión de bases de datos relacionales (RDBMS) de código abierto respaldado por Oracle y basado en el lenguaje de consulta estructurado (SQL) más extendido en la actualidad para almacenar y administrar datos.

- Admite muchos tipos de datos, como enteros con o sin signo, float, texto, data time, entre otros.
- MySQL permite almacenar y acceder a los datos a través de múltiples motores de almacenamiento, incluyendo InnoDB, CSV y NDB.
- Es capaz de replicar datos y particionar tablas para mejorar el rendimiento y la durabilidad
- Los usuarios de MySQL no tienen que aprender nuevos comandos; pueden acceder a sus datos utilizando comandos SQL estándar
- MySQL está escrito en C y C++ y es accesible y está disponible en más de 20 plataformas, como Mac, Windows, Linux y Unix
- Soporta grandes bases de datos con millones de registros y admite muchos tipos de datos
- Para la seguridad se utiliza un sistema de privilegios de acceso y contraseñas encriptadas que permite la verificación basada en el host.
- MySQL es posible conectarse a MySQL Server utilizando varios protocolos, incluyendo sockets TCP/IP en cualquier plataforma
- Es un componente de LAMP, plataforma de desarrollo web que utiliza Linux como sistema operativo, Apache como servidor web.

Sentencias básicas: SELECT es usada para consultar datos. DISTINCT Sirve para eliminar los duplicados de las consultas de datos. WHERE Es usada para incluir las condiciones de los datos que queremos consultar. AND y OR es usada para incluir 2 o más condiciones a una consulta. ORDER BY Es usada para ordenar los resultados de una consulta. INSERT Es usada para insertar datos. UPDATE Es usada para actualizar o modificar datos ya existentes. DELETE Es usada para borrar datos.

## 4.2 Diagrama modelo de datos (*Primera entrega*)

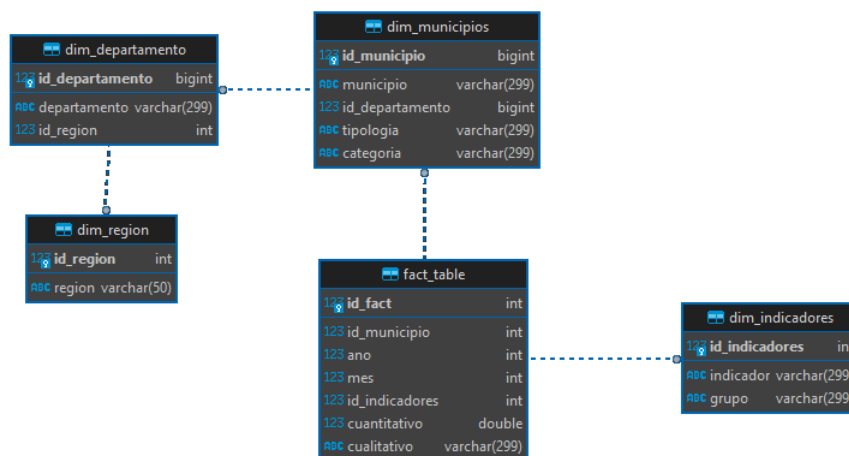


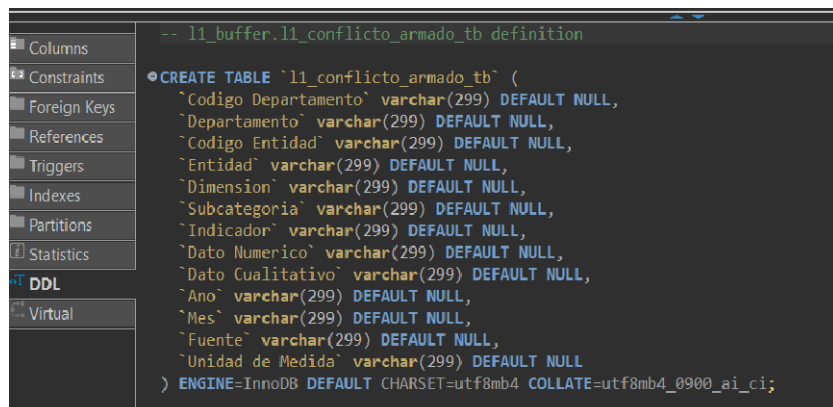
Figure 1: Modelo entidad de relación

EL modelo construido consta de 5 tablas, de las cuales 4 de ellas son dimensiones y una única tabla de hechos donde se plasman todos los indicadores medidos para los departamentos y municipios. En las dimensiones geográficas tenemos las segmentaciones propias de la DIVIPOLA a nivel nacional obteniendo, en la tabla de municipio, una tipología asociada al de nivel de ruralidad/urbanidad de este y una categoría en terminos de factores económicos.



### 4.3 Imágenes de la Base de Datos (*Primera entrega*)

### 4.4 Código SQL - lenguaje de definición de datos (DDL) (*Primera entrega*)



```
-- l1_buffer.l1_conflicto_armado_tb definition
CREATE TABLE `l1_conflicto_armado_tb` (
  `Codigo Departamento` varchar(299) DEFAULT NULL,
  `Departamento` varchar(299) DEFAULT NULL,
  `Codigo Entidad` varchar(299) DEFAULT NULL,
  `Entidad` varchar(299) DEFAULT NULL,
  `Dimension` varchar(299) DEFAULT NULL,
  `Subcategoria` varchar(299) DEFAULT NULL,
  `Indicador` varchar(299) DEFAULT NULL,
  `Dato Numerico` varchar(299) DEFAULT NULL,
  `Dato Cualitativo` varchar(299) DEFAULT NULL,
  `Año` varchar(299) DEFAULT NULL,
  `Mes` varchar(299) DEFAULT NULL,
  `Fuente` varchar(299) DEFAULT NULL,
  `Unidad de Medida` varchar(299) DEFAULT NULL
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci;
```

Figure 2: DDL

De las 4 bases de datos usadas para el desarrollo del proyecto se usó un operando CREATE para crear la tabla principal l3.edw, en la cual se ven consignados todos los cambios realizados a las bases iniciales, cambios como eliminación de datos nulos, organización de las palabras, en pocas palabras limpieza y organización de la data.

### 4.5 Código SQL - Manipulación de datos (DML) (*Primera entrega*)

Se usó la sentencia SELECT para la selección adecuada de la información, en especial las columnas que se repiten en las 4 bases de datos, al igual que JOIN e INNER JOIN, para juntar las bases de datos si generar una repetencia de las columnas generadas ni una saturación de la información contenida en la tabla final l3.edw.

```

create table if not exists l3_edw.dim_municipios (primary key (id_municipio)) as
select
    cast(id_municipio as signed) id_municipio,
    municipio,
    cast(id_departamento as signed) id_departamento,
    b.id_region,
    tipologia as id_tipologia,
    categoria as id_categoria
from (
    select *,
    row_number() over(partition by id_municipio order by categoria desc) rn
    from l2_stage.distinct_municipios
) a
join l3_edw.dim_region b on a.region = b.region
where a.rn = 1;

```

Figure 3: DML

#### 4.6 Código SQL + Resultados: Vistas (*Primera entrega*)

se procede a agregar la vista y el código usado, con enfoque en la sentencia SELECT distinct

```

create table if not exists l3_edw.dim_departamento (primary key (id_departamento)) as
select distinct
    cast(id_departamento as signed) id_departamento,
    departamento
from l2_stage.l2_territorios_tb;

create temporary table if not exists l2_stage.distinct_municipios as
select distinct *
from l2_stage.l2_territorios_tb ltt;

```

Figure 4: Vistas

#### 4.7 Código SQL + Resultados: Triggers (*Primera entrega*)

En el código no se generaron triggers, se generaron índices de las tablas creadas con las bases de datos, en los cuales se pueden observar las llaves primarias de las tablas.

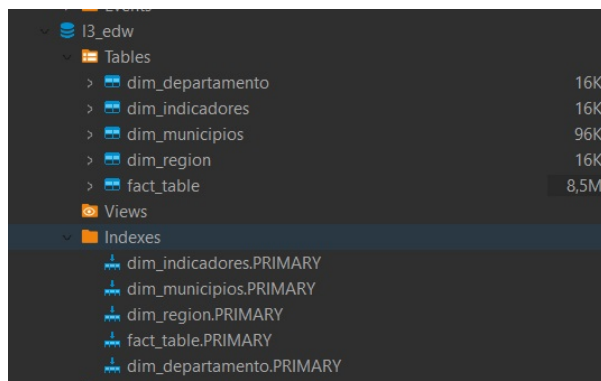


Figure 5: Indices

#### 4.8 Código SQL + Resultados: Funciones (*Primera entrega*)

se procedio a usar la función distinct para eliminar los datos repetidos en las tablas existentes en las tablas indicadores y municipios de las bases de datos.

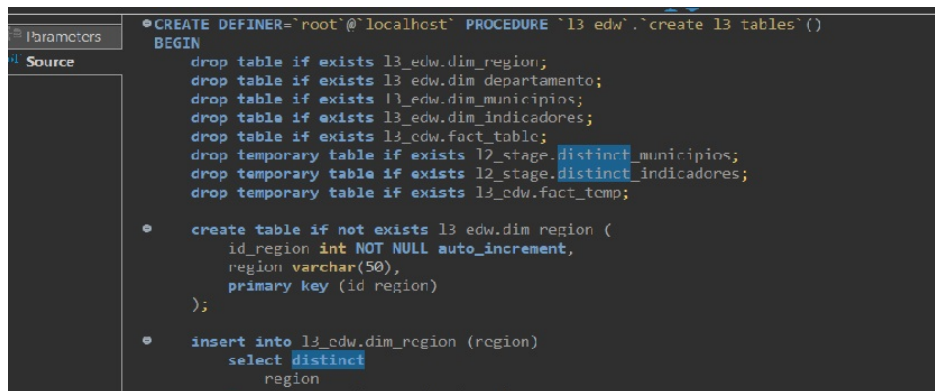
```
)
select
  cast(a.id_municipio as signed) id_municipio,
  cast(a.ano as signed) ano,
  cast(a.mes as signed) mes,
  b.id_indicadores,
  a.cuantitativo,
  a.cualitativo
from l3_edw.fact_temp a
inner join l3_edw.dim_indicadores b on a.indicador = b.indicador;

drop temporary table if exists l2_stage.distinct_indicadores;
drop temporary table if exists l2_stage.distinct_municipios;
drop temporary table if exists l3_edw.fact_temp;
END
```

Figure 6: Funciones código

## 4.9 Código SQL + Resultados: procedimientos almacenados (Primera entrega)

Se procede a mostrar el concepto general del conjunto de datos que fueron manipulados y los cuales generaron el desarrollo general del concepto de la base l3.edw En la segunda imagen se puede ver el resultado del procedimiento aplicado y como este compila en l3.edw la creación general de la tabla con las modificaciones y limpieza anteriormente aplicadas.



```
CREATE DEFINER='root'@'localhost' PROCEDURE `l3_edw`.`create l3 tables`()
BEGIN
    drop table if exists l3_edw.dim_region;
    drop table if exists l3_edw.dim_departamento;
    drop table if exists l3_edw.dim_municipios;
    drop table if exists l3_edw.dim_indicadores;
    drop table if exists l3_edw.fact_table;
    drop temporary table if exists l2_stage.distinct_municipios;
    drop temporary table if exists l2_stage.distinct_indicadores;
    drop temporary table if exists l3_edw.fact_temp;

    create table if not exists l3_edw.dim_region (
        id_region int NOT NULL auto_increment,
        region varchar(50),
        primary key (id_region)
    );

    insert into l3_edw.dim_region (region)
    select distinct
        region
```

Figure 7: Procedimiento

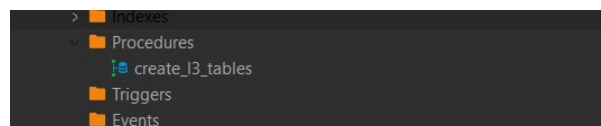


Figure 8: Procedimiento Resultados

## 5 Bases de Datos No-SQL (*Segunda entrega*)

### 5.1 Diagrama Bases de Datos No-SQL (*Segunda entrega*)

### 5.2 SMBD utilizado para la Base de Datos No-SQL (*Segunda entrega*)

## 6 Aplicación de ETL (Extract, Transform, Load) y Bodega de Datos (*Tercera entrega*)

### 6.1 Ejemplo de aplicación de ETL y Bodega de Datos (*Tercera entrega*)

## 7 Lecciones aprendidas (*Tercera entrega*)

## 8 Bibliografía

<https://www.mineducacion.gov.co/portal/normativa/Resoluciones/351282:Resolucion-No-07797-de-2015>

<https://www.dane.gov.co/files/investigaciones/educacion/educacion-formal/2021/bol-EDUC-21-pdf>

[https://www.dane.gov.co/files/investigaciones/boletines/educacion/bol\\_EDUC20.pdf](https://www.dane.gov.co/files/investigaciones/boletines/educacion/bol_EDUC20.pdf)

[https://www.dane.gov.co/files/investigaciones/boletines/educacion/bol\\_EDUC19.pdf](https://www.dane.gov.co/files/investigaciones/boletines/educacion/bol_EDUC19.pdf)

<https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/pobreza-multidimensional>

[https://www.datos.gov.co/Educaci-n/MEN\\_ESTADISTICAS\\_EN\\_EDUCACION\\_EN\\_PREESCOLAR-B-SICA/nudc-7mev](https://www.datos.gov.co/Educaci-n/MEN_ESTADISTICAS_EN_EDUCACION_EN_PREESCOLAR-B-SICA/nudc-7mev)