

**Máster universitario de Ciencia de Datos**

**Práctica 2**

**Diseño y uso de bases de datos analíticas – identificación, diseño y desarrollo de los procesos ETL.**

Autor:

Mario Ubierna San Mamés

|  |
| --- |
| Índice de Contenido |

[Índice de Contenido 3](#_Toc71554266)

[Índice de tablas 4](#_Toc71554267)

[Índice de ilustraciones 5](#_Toc71554268)

[1. Introducción 6](#_Toc71554269)

[2. Bibliografía 7](#_Toc71554270)

|  |
| --- |
| Índice de tablas |

**No se encuentran elementos de tabla de ilustraciones.**

|  |
| --- |
| Índice de ilustraciones |

**No se encuentran elementos de tabla de ilustraciones.**

|  |
| --- |
| Introducción |

## Presentación

A partir de la solución oficial de la primera práctica (PRA1), el estudiante debe diseñar, implementar y ejecutar los procesos de extracción, transformación y carga de los datos de las fuentes de datos proporcionadas.

Así pues, esta actividad tiene como objetivo identificar y desarrollar los procesos de carga del almacén de datos y que esta sea efectiva.

## Descripción

Si nos centramos en los subobjetivos, esta segunda parte del caso práctico consiste en lo siguiente:

* Identificar los procesos de extracción, transformación y carga de datos (ETL) hacia el almacén de datos.
* Diseñar y desarrollar los procesos ETL mediante las herramientas de diseño proporcionadas.
* Implementar con los trabajos (*jobs*) los procesos ETL para que su carga planificada sea efectiva.

Además del documento con la solución de la PRA2 que se debe entregar, también se tendrá en consideración la implementación sobre la máquina virtual proporcionada en el curso.

En resumen, el documento de la solución de la PRA2 debe incluir los siguientes aspectos:

* Descripción de todas las acciones que se han realizado.
* Capturas de pantalla que muestren todas las partes significativas del ETL, sus características y su correspondiente explicación.
* Capturas de pantalla que demuestren la correcta ejecución de la ETL y el tiempo de ejecución.
* Capturas de pantalla que demuestren las correcta carga de los datos (cargados en la base de datos).

|  |
| --- |
| Identificación de los procesos ETL |

A la hora de diseñar los procesos de carga de una base de datos analítica no hay una única estrategia. Es habitual estructurar los procesos ETL sobre la base de las entidades de datos que se deben actualizar, ya que existen diferencias conceptuales en la actualización de una dimensión con respecto a la de una tabla de hechos. La división del proceso de carga inicial en diferentes bloques de actualización facilitará el diseño de un orden de ejecución y la gestión de las dependencias. Cada uno de estos bloques de actualización se dividirá en las correspondientes etapas de extracción, transformación y carga.

Se identifican los dos bloques siguientes:

* **Bloque IN:** procesos de carga de los datos desde las fuentes a las tablas intermedias en el área de maniobras (*staging area*). Estos procesos se distinguen por el prefijo «IN\_» en el nombre.
* **Bloque TR:** procesos de transformación para cargar los datos desde las tablas intermedias hasta nuestro almacén, según el modelo multidimensional diseñado. Así pues, son diferentes los procesos ETL de transformación para cargar las dimensiones de aquellos que se realizan para cargar las tablas de hechos. Estos procesos se distinguen con el prefijo «TR\_» en el nombre.

## Bloque IN

Respecto al bloque In, el cual nos va a permitir almacenar la información en el staging area, tenemos los siguientes procesos:

|  |  |  |  |
| --- | --- | --- | --- |
| **Nombre ETL** | **Descripción** | **Orígenes de los dastos** | **Tabla de destino (stage)** |
| IN\_ DENUNCIAS\_ INFRACCIONES | Carga de los datos correspondientes a las estadísticas sobre los expedientes incoados por el artículo 36.6 LOPSC de desobediencia durante el estado de emergencia sanitaria COVID-19 en la comunidad de Euskadi. | ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx | STG\_Denuncias\_Infracciones |
| IN\_POBLACION | Carga los datos respectivos a las cifras de la población española. | población\_9687bsc.csv | STG\_Poblacion |
| IN\_MOVILIDAD | Movilidad de la población durante el estado de alarma. | 35167bsc.csv | STG\_Movilidad |
| IN\_AGLOMERACION | Porcentaje de la población que evitaba las aglomeraciones con motivo del coronavirus, por grupo de edad y provincia. | statistic\_id1104235\_covid19\_-poblacion-que-evitabalas-aglomeraciones-segunedad-en-espana-2020.xlsx | STG\_Evitar\_Aglomeracion |
| IN\_LLAMADAS\_112 | Llamadas al 112 por ámbito geográfico y tipología (accidentes de tráfico, civismo, incendios, asistencia sanitaria, seguridad...) | rows.xml | SGT\_Llamadas112 |

Tabla - Procesos ETL Bloque IN.

## Bloque TR

Respecto al bloque TR tenemos tanto los procesos para dotar de datos a las dimensiones como a los hechos.

### Dimensiones

Los procesos ETL que se encargar de añadir la información a la dimensiones son los siguientes:

|  |  |  |  |
| --- | --- | --- | --- |
| **Nombre del ETL** | **Descripción** | **Tabla de origen** | **Tabla de destino (dimensión)** |
| TR\_DIM\_FECHA | Carga y transformación de la dimensión temporal. |  |  |
| TR\_DIM\_AMBITO\_GEOGRAFICO | Carga y transformación de la dimensión con los datos de los ámbitos geográficos. |  |  |
| TR\_DIM\_GRUPO\_EDAD | Carga y transformación de la dimensión con los datos de los grupo de edad. |  |  |
| TR\_DIM\_MEDICION | Carga y transformación de la dimensión con los datos de las mediciones. |  |  |
| TR\_DIM\_TIPOLOGIA | Carga y transformación de la dimensión con los datos de la tipología. |  |  |

Tabla - Procesos ETL Bloque TR Dimensiones.

### Hechos

Respecto a los hechos tenemos los siguientes procesos de carga:

|  |  |  |
| --- | --- | --- |
| **Nombre del ETL** | **Descripción** | **Tabla de origen** |
| TR\_FACT\_LLAMADAS112 | Carga y transformación de la tabla de hechos Fact\_Llamadas112. | STG\_Llamadas112 |
| TR\_FACT\_MEDICIONES | Carga y transformación de la tabla de hechos Fact\_Mediciones |  |

Tabla - Procesos ETL Bloque TR Hechos.

|  |
| --- |
| Diseño y desarrollo de los procesos ETL |

En este apartado, se deben diseñar los procesos de carga identificados en el punto anterior con la herramienta de diseño proporcionada. En este caso es Pentho Data Integration (PDI).

## Creación de tablas

El primer paso para la implementación de los procesos ETL consiste en la creación de las tablas. Esto se llevará a cabo una única vez, mediante *scripts*, sobre la base de dastos proporcionada (en nuestro caso: SQL Server). Se deberán crear las tablas intermedias y las tablas del modelo dimensional de la solución oficial, es decir, las dimensiones y las tablas de hechos. Para hacerlo, deben utilizarse los *scripts* facilitados junto a la solución de la PRA1.

### Tablas del área intermedia (*staging area*)

Lo primero que vamos a hacer es la creación de las tablas intermedias:

#### Tabla intermedia STG\_Denuncias\_Infracciones

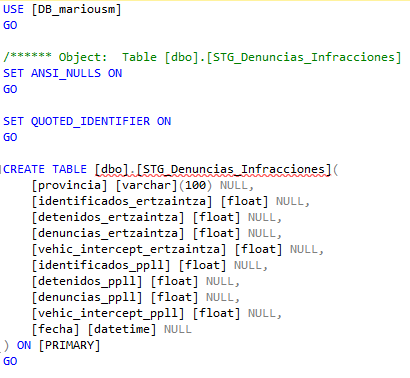


Ilustración - STG\_Denuncias\_Infracciones.

#### Tabla intermedia STG\_Poblacion

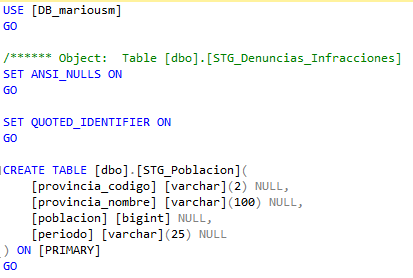


Ilustración - STG\_Poblacion.

#### Tabla intermedia STG\_Llamadas112

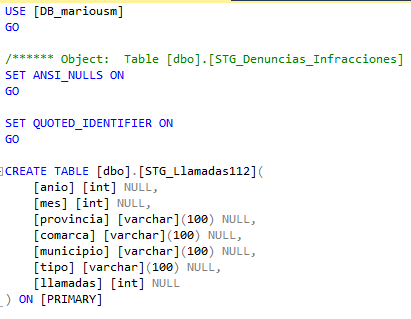


Ilustración - STG\_Llamadas112.

#### Tabla intermedia STG\_Movilidad

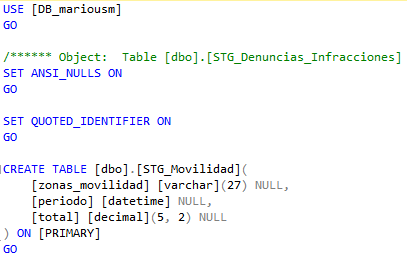


Ilustración - STG\_Movilidad.

#### Tabla intermedia STG\_Evitar\_Aglomeracion

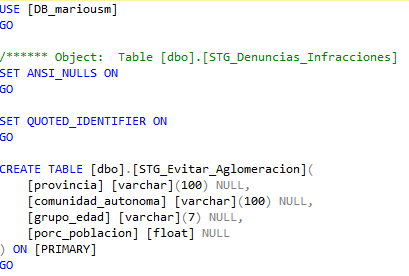


Ilustración - STG\_Evitar\_Aglomeracion.

Comprobamos que todas las tabla intermedias se han creado correctamente:

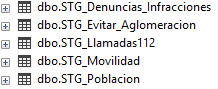


Ilustración - Tablas de staging area.

### Tablas de las dimensiones

Lo segundo que debemos de hacer es la creación de las tablas de dimensiones:

#### Tabla dimensión DIM\_Ambito\_Geografico

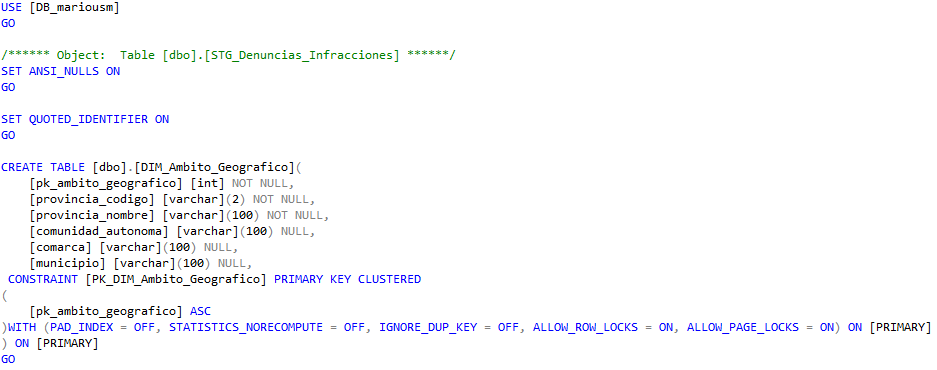


Ilustración - DIM\_Ambito\_Geografico.

#### Tabla dimensión DIM\_Fecha

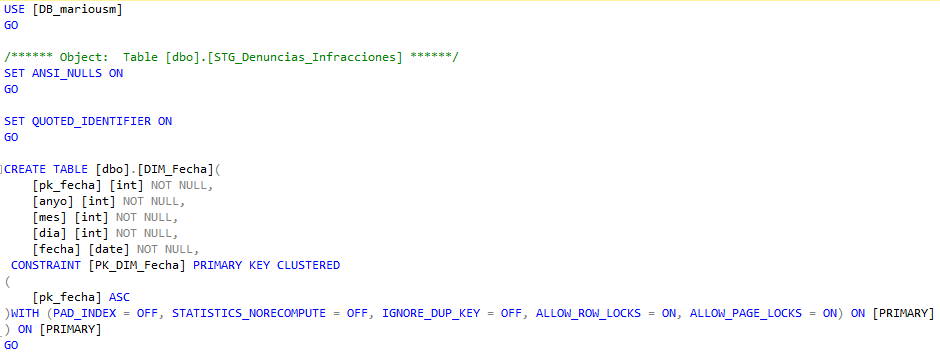


Ilustración - DIM\_Fecha.

#### Tabla dimensión DIM\_Grupo\_Edad

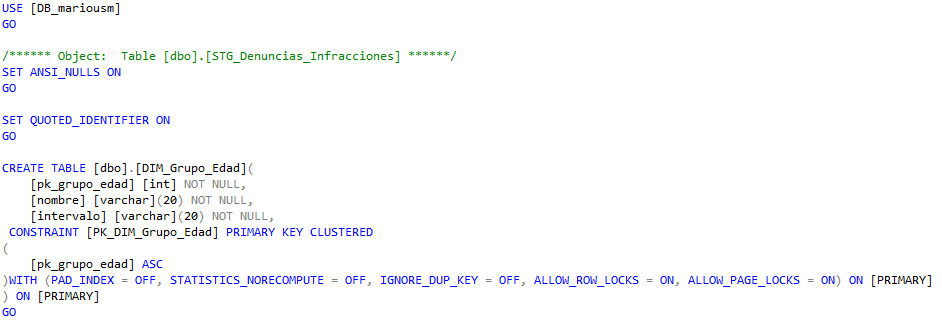


Ilustración - DIM\_Grupo\_Edad.

#### Tabla dimensión DIM\_Medicion

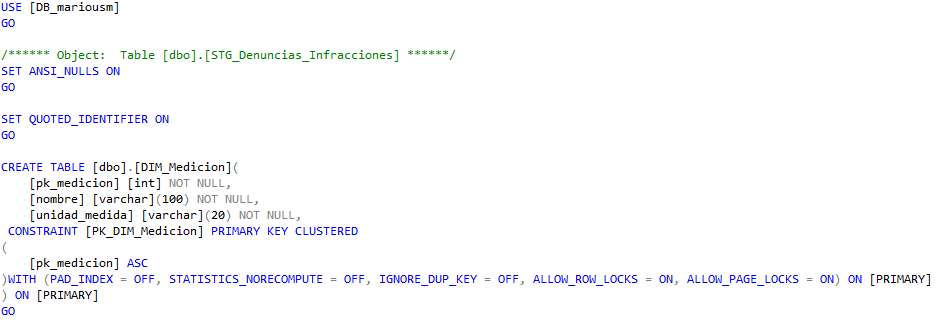


Ilustración - DIM\_Medicion.

#### Tabla dimensión DIM\_Tipologia

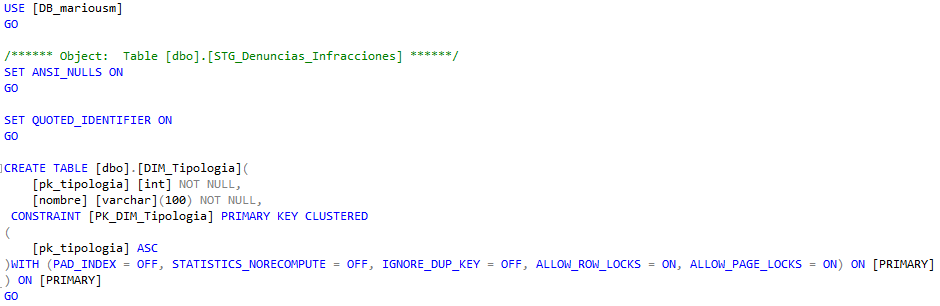


Ilustración - DIM\_Tipologia.

Comprobamos que todas las tablas de dimensiones se han creado correctamente:

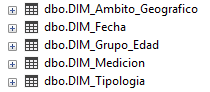


Ilustración - Tablas de dimensiones.

### Tablas de hechos

Finalmente creamos las diferentes tablas de los hechos:

#### Tabla hecho FACT\_Llamadas112

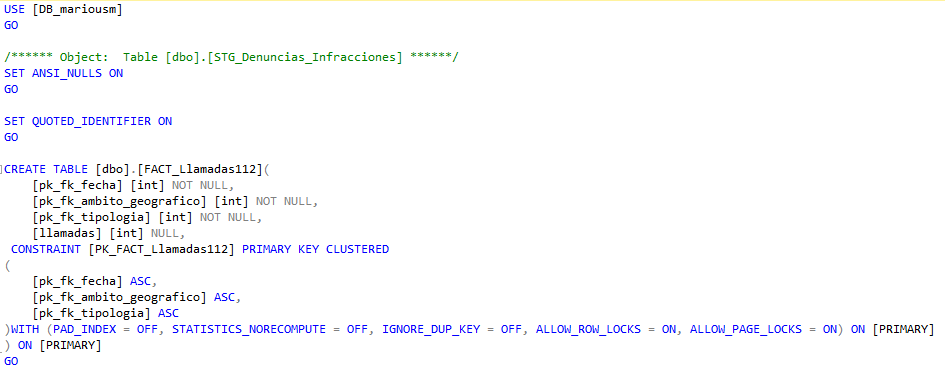


Ilustración - FACT\_Llamadas112.

#### Tabla hecho FACT\_Mediciones

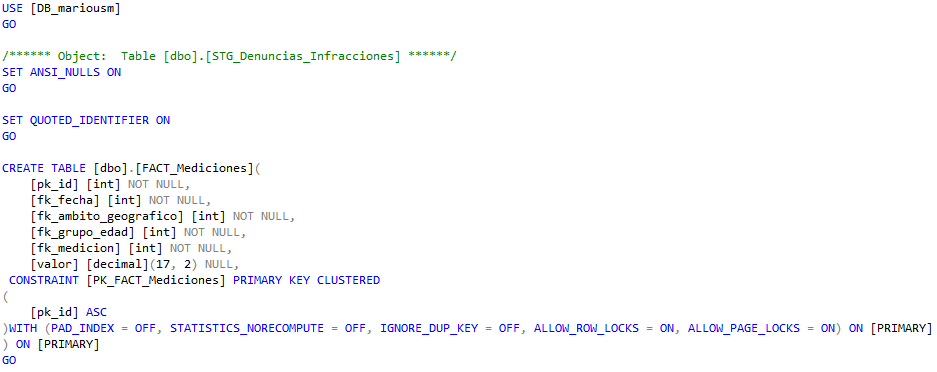


Ilustración - FACT\_Mediciones.

Comprobamos que se han creado todas las tablas correspondientes:



Ilustración - Tablas de hechos.

|  |
| --- |
| Bibliografía |

a