# Caso práctico: Almacén de datos para el análisis de la cobertura por inmunización

# PRA2 - Carga y explotación del *data warehouse*

## Presentación

Los objetivos de esta actividad son los siguientes:

- Identificar y desarrollar los procesos de carga del almacén de datos, así como su carga efectiva.
- Implementar un modelo OLAP para el análisis multidimensional de la información para responder a las preguntas definidas en la toma de requerimientos.

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. Tras la carga efectiva de los datos en el almacén de datos mediante los procesos ELT desarrollados, se debe implementar un cubo multidimensional para la explotación de la información como apoyo a la toma de decisiones de los usuarios potenciales.

## Se pide:

La segunda parte de la práctica consiste en lo siguiente:

- Identificación de los procesos de ETL (extracción, transformación y carga de datos) hacia el almacén de datos.
- Diseño y desarrollo de los procesos ETL mediante las herramientas de diseño proporcionadas.
- Implementación con trabajos de los procesos ETL para su carga efectiva planificada.
- Diseño de un modelo OLAP (*multidimensional on line analytical processing*) para el análisis multidimensional de la información disponible en el almacén de datos que permita realizar:
    - Análisis evolutivo de la tendencia de la cobertura de inmunización.
    - Análisis de la tendencia de los diez últimos años de la cobertura de inmunización de España.

- Análisis de los países miembros que consiguen el cumplimiento en 2018 del objetivo 1 para la meta 2015 propuesta por el GVAP (Plan de Acción Mundial sobre Vacunas).
- *Ranking* de países de la OMS (Organización Mundial de la Salud) que llegan al cumplimiento del objetivo 1 para la meta 2020 propuesta por el GVAP.
- *Ranking* de países con mayor porcentaje de población objetivo vacunada.
- Evolución de los casos de muerte por sarampión en los países de Europa.
- Planteamiento de otros análisis que puedan enriquecer el sistema con otras herramientas o visualizaciones que se estimen oportunas.

## Descripción

Se evaluará la implementación sobre la máquina virtual, por ese motivo se debe desarrollar la PRA2 en ella.

La solución de la PRA2 debe incluir lo siguiente:

- Descripción de todas las acciones realizadas en el proceso.
- Capturas o *script* de creación de las tablas en SQL Server.
- Capturas de pantalla de todas las partes del ETL, sus características y su correspondiente explicación.
- Capturas de pantalla que demuestren la correcta ejecución del ETL y el tiempo de ejecución.
- Capturas que demuestren la correcta carga de datos (datos cargados en la base de datos).
- Capturas que demuestren la correcta definición de las vistas y cubos OLAP, así como las dimensiones, relaciones y jerarquías
- Capturas que muestren la correcta definición de las consultas y la visualización del resultado de las explotaciones.

## Guía de muestra

Con el fin de ayudar a alcanzar los objetivos planteados de la PRA2, se desarrolla esta guía de muestra. La guía servirá para realizar alguna de las tareas anteriormente descritas; es decir, el diseño y el desarrollo de los procesos ETL, la carga efectiva al almacén de datos de cobertura de inmunización y el diseño de explotaciones de un modelo OLAP.

Para la entrega de la segunda práctica, PRA2, el estudiante, siguiendo la guía de muestra, deberá realizar los puntos solicitados para completar la carga y explotación de los datos del almacén de cobertura de inmunización.

# I. 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 basándose en 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 dos bloques:

- **Bloque IN**: procesos de carga de los datos desde las fuentes a las tablas intermedias en el área intermedia (*staging area*). Estos procesos se distinguen por el prefijo: IN_ en el nombre.
- **Bloque TR**: procesos de transformación para la carga de datos desde las tablas intermedias a nuestro almacén según el modelo multidimensional diseñado. Se diferencian los procesos ETL de transformación para la carga de dimensiones de los procesos de transformación para la carga de las tablas de hechos. Estos procesos se distinguen con el prefijo TR_ en el nombre.

A continuación, se identifican algunos de los procesos que forman parte de cada uno de los bloques de actualización:

### Bloque IN (de las fuentes a las tablas intermedias)

| Nombre ETL | Descripción |
|---|---|
| IN_SCOB_OMS | Carga de los datos de la fuente coverage_estimates_series.xls correspondiente a la cobertura de inmunización por años, por región, por países y vacunas a la tabla intermedia IN_SCOB_OMS. |
| IN_REGION_OMS | Carga de los registros de la fuente REGION.json a la tabla intermedia IN_REGION_OMS. |

### Bloque TR (poblar las tablas de nuestro almacén)

El bloque TR_ de procesos ETL para poblar el modelo multidimensional del almacén tiene dos partes diferenciadas. Los procesos de carga y transformación de las dimensiones y los de las tablas de hechos. El orden de ejecución es importante para que la carga de datos sea correcta. Las dimensiones se cargarán primero y después las tablas de hechos para no tener errores en la carga.

Algunos de los procesos del bloque TR de carga y transformación de las dimensiones son los siguientes:

| Nombre ETL | Descripción |
|---|---|
| TR_DIM_CAUSAS_MORTANDAD | Carga y transformación de la dimensión causas de mortandad (DIM_ CAUSAS_MORTANDAD). |
| TR_DIM_PAISES | Carga y transformación de la dimensión países (DIM_PAISES). |

Y el proceso del bloque de carga y transformación de la tabla de hechos TR_FACT_COBERTURA:

| Nombre ETL | Descripción |
|---|---|
| TR_FACT_COBERTURA | Carga y transformación de la tabla de hechos FACT_COBERTURA. |

El estudiante, para resolver este punto, deberá completar la identificación de los procesos en cada uno de los bloques (IN y TR) que desarrollará para la carga de dimensiones y tablas de hechos del modelo multidimensional del almacén de datos de la cobertura de inmunización.

## II. Diseño y desarrollo de los procesos ETL

En este apartado, se deben diseñar los procesos de carga identificados en el punto anterior mediante la herramienta de diseño proporcionada: Pentaho Data Integration (PDI).

### Creación de tablas

El primer paso para la implementación de los procesos ETL consiste en la creación de las tablas. Este se llevará a cabo una única vez, mediante *scripts* sobre la base de datos 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.

A continuación se muestra el *script* de creación de una de las dimensiones del modelo multidimensional definido en la solución oficial de la PRA1. En concreto, indicamos en la guía, a modo de ejemplo, el *script* de creación de la dimensión temporal DIM_ANIO.

#### Tabla de dimensión la temporal DIM_ANIO

```
CREATE TABLE [dbo].[DIM_ANIO](
                [SK_DIM_ANIO] [numeric](4, 0) NOT NULL,
                [DESC_ANIO] [varchar](50) NOT NULL,
 CONSTRAINT [PK_DIM_ANIO] PRIMARY KEY CLUSTERED
(
                [SK_DIM_ANIO] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF,
ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]
GO
```

En el aula, al estudiante se le proporciona el *script* **Script_DIM_ANIO.sql** de creación de la dimensión temporal.

El resto de tablas (intermedias, dimensiones y tablas de hechos) del modelo de cobertura de inmunización tendrán que ser creadas por el estudiante y, en este apartado, incluir todos los *scripts* de creación.

Una vez que ya hayamos implementado el modelo físico del almacén, el siguiente paso es diseñar los procesos ETL de cada uno de los bloques (IN y TR) que permitirán poblar las tablas intermedias del área intermedia (*staging area*) y las tablas de dimensiones y de hechos del *data mart* que hemos diseñado.

## Bloque IN

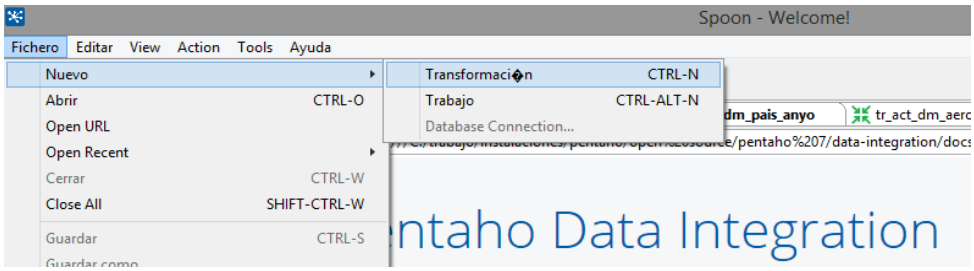
### Transformación IN_COB_SERIES

A continuación, se describe el desarrollo de la transformación IN_COB_SERIES (identificada en el primer punto de la guía), mediante spoon, para la carga de la fuente coverage_estimates_series.xls con los datos por años de la cobertura de inmunización a la tabla intermedia IN_COB_OMS del *staging area*, que ha tenido que ser creada en la base de datos analítica y cuyo *script* se habrá escrito en el apartado de creación.

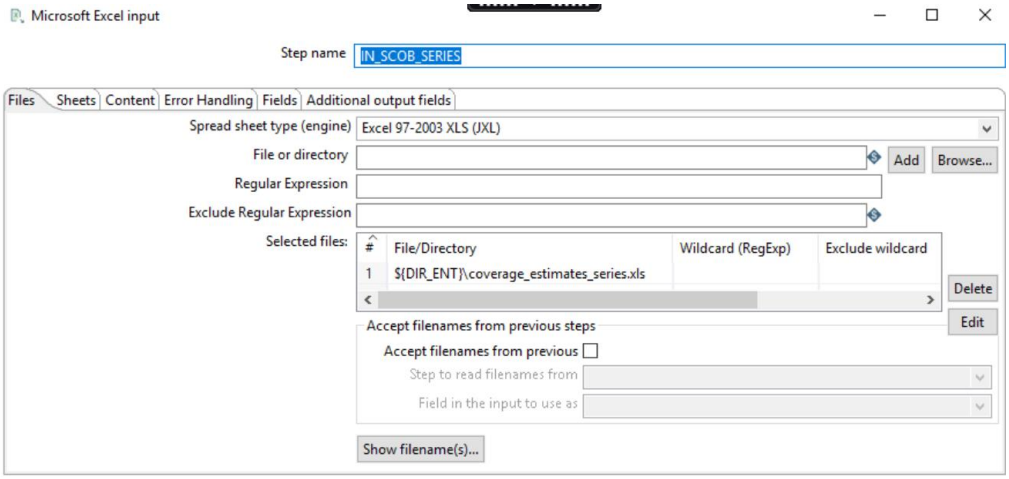
Para este caso práctico hemos utilizado fuentes externas que emplearemos para descubrir conocimiento realizando análisis de los datos. No se utilizan fuentes operacionales, en cuyo caso hay una etapa muy importante de preparación de las fuentes para dejarlas listas para su tratamiento con la herramienta ETL. Es muy habitual manipular los ficheros, concretamente los de tipo xls, realizando manualmente una serie de acciones de preparación antes de su procesamiento.

En concreto, para que se pueda cargar la fuente coverage_estimates_series.xls, se eliminará la última fila (53649) que aparece después de la tabla de datos, con información del significado de la columna asterisco.

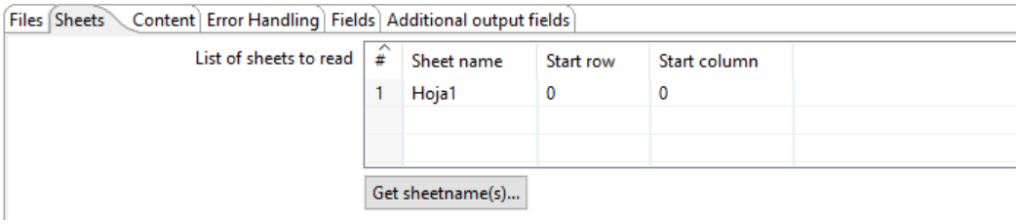

La transformación IN_COB_SERIES contiene cuatro transformaciones: lectura del fichero xls, operaciones con cadenas, ordenar filas y carga a la tabla intermedia IN_COB_OMS.




Este es el primer paso de la transformación. Como se trata de un fichero xls, utilizaremos como entrada el tipo Microsoft Excel input. En la pestaña «Ficheros (Files)» añadimos el fichero Excel desde donde extraemos los datos, para ello utilizamos la variable de entorno DIR_ENT.

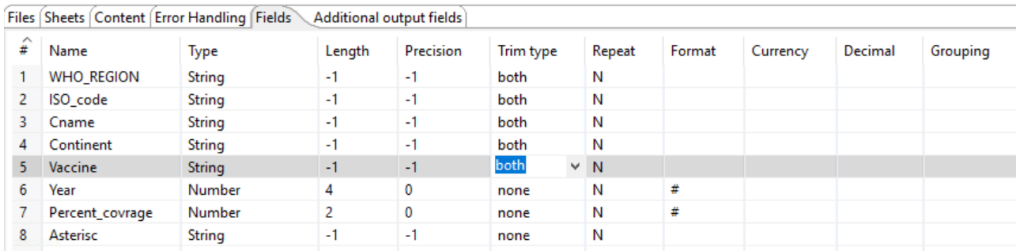



En la pestaña «Hojas (Sheets)», indicamos el nombre de la hoja de cálculo que queremos procesar e indicamos desde qué fila del archivo se comienzan a leer los datos; en nuestro caso, después de la preparación la fila 0.




Indicamos, en la siguiente pestaña, que existe una fila de encabezados de los campos.


Le indicamos que recupere los campos que vamos a tratar mediante el botón «Get fields from header row» y completamos la definición de los campos, especificando la precisión de los campos numéricos y eliminando espacios en ambos lados de los campos de tipo *string*.




Para realizar una visualización previa de los datos que se cargarían se utiliza el botón «Previsualizar (Preview rows)».

Rows of step: IN_SCOB_SERIES (1000 rows)

| # | WHO_REGION | ISO_code | Cname | Continent | Vaccine | Year | Percent_covrage | Asterisc |
|---|---|---|---|---|---|---|---|---|
| 1 | EMR | AFG | Afghanistan | Asia | BCG | 1980 | 33 | <null> |
| 2 | EMR | AFG | Afghanistan | Asia | BCG | 1981 | 8 | <null> |
| 3 | EMR | AFG | Afghanistan | Asia | BCG | 1982 | 10 | <null> |
| 4 | EMR | AFG | Afghanistan | Asia | BCG | 1984 | 11 | <null> |
| 5 | EMR | AFG | Afghanistan | Asia | BCG | 1985 | 17 | <null> |
| 6 | EMR | AFG | Afghanistan | Asia | BCG | 1986 | 18 | <null> |
| 7 | EMR | AFG | Afghanistan | Asia | BCG | 1987 | 27 | <null> |
| 8 | EMR | AFG | Afghanistan | Asia | BCG | 1988 | 40 | <null> |
| 9 | EMR | AFG | Afghanistan | Asia | BCG | 1989 | 38 | <null> |
| 10 | EMR | AFG | Afghanistan | Asia | BCG | 1990 | 30 | <null> |
| 11 | EMR | AFG | Afghanistan | Asia | BCG | 1991 | 21 | <null> |
| 12 | EMR | AFG | Afghanistan | Asia | BCG | 1993 | 44 | <null> |
| 13 | EMR | AFG | Afghanistan | Asia | BCG | 1994 | 15 | <null> |
| 14 | EMR | AFG | Afghanistan | Asia | BCG | 1996 | 47 | <null> |

El siguiente paso de la transformación es asegurar la calidad de los datos mediante la normalización de los valores de los campos *string*, para ello convertiremos los datos de las fuentes origen en mayúsculas mediante el componente «String_operations».

String operations

Step name: String_operations

The fields to process:

| # | In stream field | Out stream field | Trim type | Lower/Upper | Padding | Pad char | Pad Length | InitCap |
|---|---|---|---|---|---|---|---|---|
| 1 | WHO_REGION | | none | upper | none | | | N |
| 2 | ISO_code | | none | upper | none | | | N |
| 3 | Cname | | none | upper | none | | | N |
| 4 | Continent | | none | upper | none | | | N |
| 5 | Vaccine | | none | upper | none | | | N |

El siguiente paso de la transformación sería la ordenación ascendente por los campos región, iso_code del país miembro, nombre del país y vacuna. Para ello utilizaremos el componente «Ordenar filas (Sort rows)» de las posibles transformaciones disponibles.

Sort rows

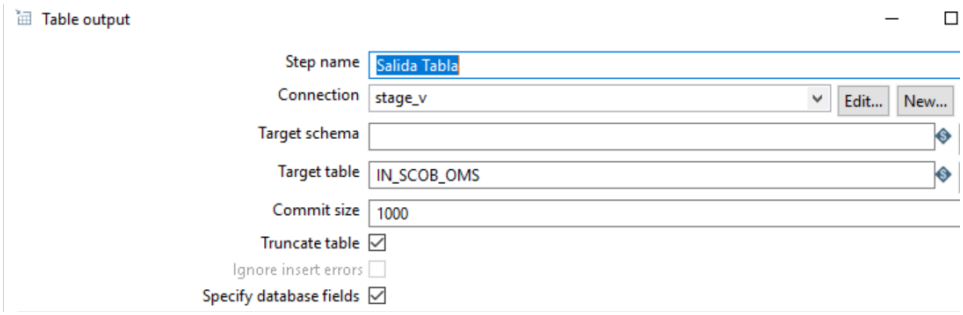
Step name: Ordenar filas
Sort directory: %%java.io.tmpdir%% Browse...
TMP-file prefix: out
Sort size (rows in memory): 1000000
Free memory threshold (in %):
Compress TMP Files?
Only pass unique rows? (verifies keys only)

Fields :

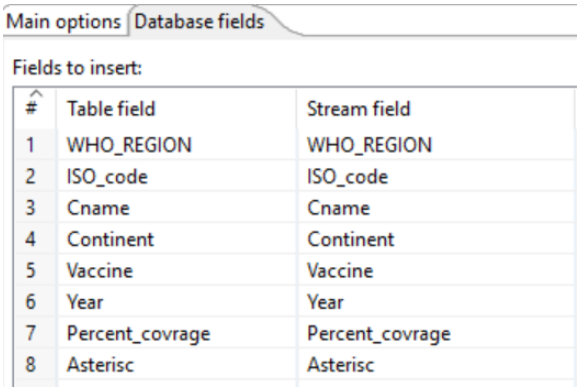
| # | Fieldname | Ascending | Case sensitive compare? | Sort based on current locale? | Collator Strength | Presorted? |
|---|---|---|---|---|---|---|
| 1 | WHO_REGION | Y | N | N | 0 | N |
| 2 | ISO_code | Y | N | N | 0 | N |
| 3 | Cname | Y | N | N | 0 | N |
| 4 | Vaccine | Y | N | N | 0 | N |

Por último, cargamos los datos en la tabla intermedia al *stage*, utilizando el paso Salida Tabla de la carpeta Salida. En este paso se necesita especificar la conexión de base de datos, para ello utilizaremos la variable de entorno STAGE que hemos definido.

El paso de carga de datos a la tabla intermedia del *stage* lo configuraremos como sigue en el menú principal:

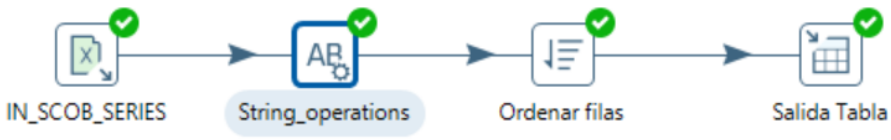



Para dejar la transformación preparada para posibles reprocesos, es necesario realizar un borrado previo para actualizar los datos. Para esto, activaremos el *check* «Truncate table». En los campos de la base de datos:

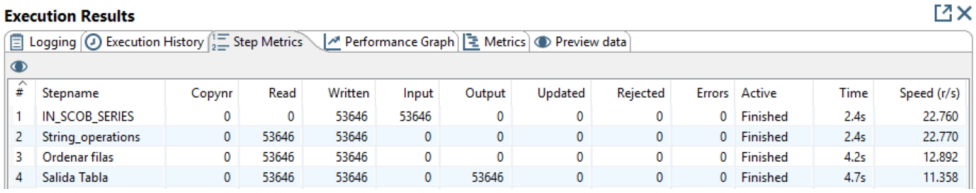



| # | Table field | Stream field |
|---|---|---|
| 1 | WHO_REGION | WHO_REGION |
| 2 | ISO_code | ISO_code |
| 3 | Cname | Cname |
| 4 | Continent | Continent |
| 5 | Vaccine | Vaccine |
| 6 | Year | Year |
| 7 | Percent_covrage | Percent_covrage |
| 8 | Asterisc | Asterisc |

La transformación completa es la siguiente:




El resultado de la ejecución es el siguiente:




| # | Stepname | Copynr | Read | Written | Input | Output | Updated | Rejected | Errors | Active | Time | Speed (r/s) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | IN_SCOB_SERIES | 0 | 0 | 53646 | 53646 | 0 | 0 | 0 | 0 | Finished | 2.4s | 22.760 |
| 2 | String_operations | 0 | 53646 | 53646 | 0 | 0 | 0 | 0 | 0 | Finished | 2.4s | 22.770 |
| 3 | Ordenar filas | 0 | 53646 | 53646 | 0 | 0 | 0 | 0 | 0 | Finished | 4.2s | 12.892 |
| 4 | Salida Tabla | 0 | 53646 | 53646 | 0 | 53646 | 0 | 0 | 0 | Finished | 4.7s | 11.358 |

Como se observa en las métricas se cargan los 53.646 registros del fichero de entrada.

En el aula, al estudiante se le proporciona la transformación IN_COB_SERIES.ktr para la carga de la tabla intermedia IN_COB_OMS.

Para cargar los datos de la PARA2, el estudiante deberá diseñar todos los procesos ETL en cada uno de los bloques (IN y TR).

# III. Implementación de trabajos con procesos ETL

Teniendo en cuenta los bloques de procesos ETL implementados:

- **Bloque IN_:** procesos ETL de transformación y carga al área intermedia.
- **Bloque TR_DIM**: procesos ETL de transformación y carga de dimensiones.
- **Bloque TR_FACT**: procesos ETL de transformación y carga de hechos.

El estudiante, en este punto, para realizar la carga efectiva de los datos, debe diseñar los trabajos (*jobs*), mediante PDI, que van a permitir la ejecución secuencial de todos los procesos ETL incluidos en cada bloque definido.

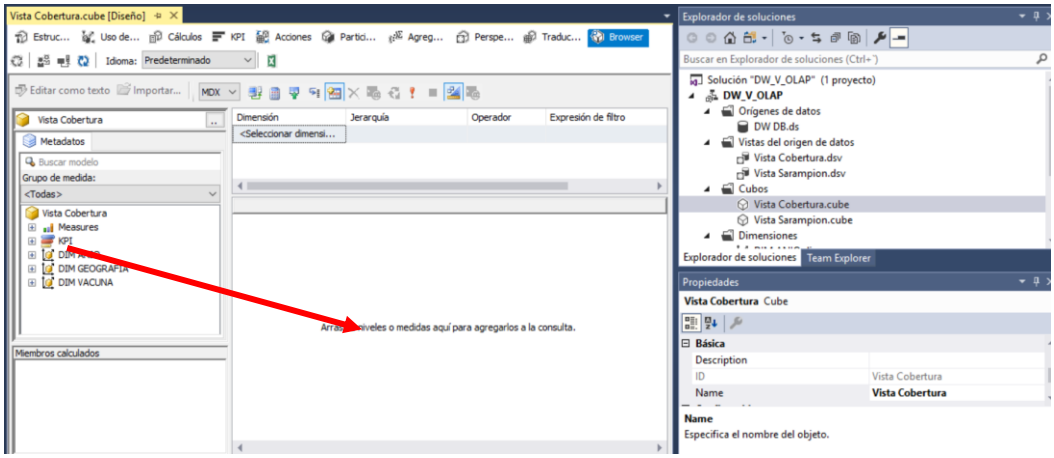
# IV. Diseño del modelo OLAP

El diseño del cubo se realizará creando un proyecto multidimensional y de minería de datos en Visual Studio. Habrá que definir las vistas, los cubos, las dimensiones y las jerarquías necesarias para realizar las explotaciones solicitadas en el enunciado de la PRA2.

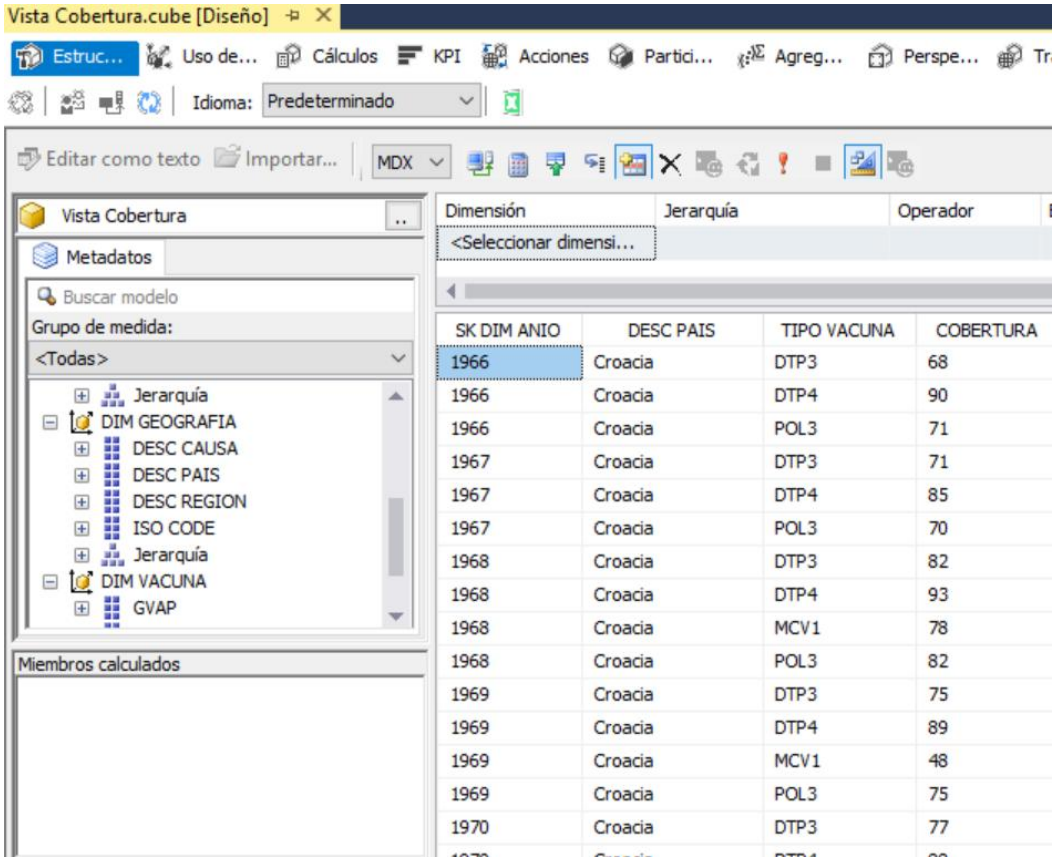
A continuación, se muestra un ejemplo de explotación de datos que se puede realizar tras la implementación de la solución. La manera de acceder al visor OLAP es entrando en la pestaña “Browser” en cada cubo creado.

**Análisis evolutivo de la cobertura de inmunización de los países miembros desde 1966 hasta 2018**

La primera explotación que vamos a realizar es el análisis evolutivo de la cobertura de inmunización desde 1966 hasta 2018 de los países. Para diseñar el informe, abriremos el cubo cobertura que hemos diseñado y nos posicionaremos en la pestaña «Browser», que es la que nos permite crear los informes utilizando las medidas y dimensiones disponibles.




Para mostrar la tendencia de la evolución de la cobertura a lo largo de los años, arrastraremos la medida COBERTURA al área de trabajo y también arrastraremos el atributo SK_DIM_ANIO de la dimensión DIM_ANIO, el atributo DESC_PAIS de la dimensión DIM_GEOGRAFIA y TIPO_VACUNA de la DIM_VACUNA. Una vez que tengamos las dimensiones y medidas que necesitamos para crear el informe, se visualizarán los datos en el área de trabajo.




| SK DIM ANIO | DESC PAIS | TIPO VACUNA | COBERTURA |
|---|---|---|---|
| 1966 | Croacia | DTP3 | 68 |
| 1966 | Croacia | DTP4 | 90 |
| 1966 | Croacia | POL3 | 71 |
| 1967 | Croacia | DTP3 | 71 |
| 1967 | Croacia | DTP4 | 85 |
| 1967 | Croacia | POL3 | 70 |
| 1968 | Croacia | DTP3 | 82 |
| 1968 | Croacia | DTP4 | 93 |
| 1968 | Croacia | MCV1 | 78 |
| 1968 | Croacia | POL3 | 82 |
| 1969 | Croacia | DTP3 | 75 |
| 1969 | Croacia | DTP4 | 89 |
| 1969 | Croacia | MCV1 | 48 |
| 1969 | Croacia | POL3 | 75 |
| 1970 | Croacia | DTP3 | 77 |

En el análisis se observa que desde 2010 a 2011 la tendencia de la cobertura de inmunización es positiva.

El estudiante, en la PRA2, tomando como modelo el diseño del cubo cobertura de la guía deberá completar la fase de explotación de datos que permita realizar los análisis solicitados en el enunciado.

## Formato y fecha de entrega

La entrega final de esta actividad debe realizarse enviando un único mensaje al buzón Registro de AC del apartado Evaluación del aula. Adjunto en el mensaje se enviará un único archivo en formato Word o PDF con la solución de la PRA2. El nombre del archivo debe ser la composición del nombre de usuario y _BDA_PRA2 (por ejemplo: si el nombre de usuario es bantich, el nombre del archivo debe ser bantich_BDA_PRA2.pdf).

**La fecha máxima de entrega es el 04/06/2020 a las 23.59 horas.**