IMF Business School

Hoja de respuestas

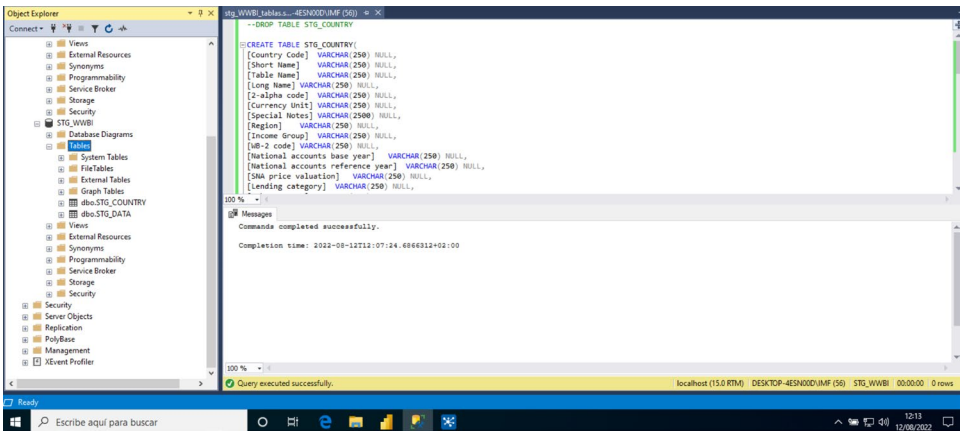
Módulo	INTELIGENCIA DE NEGOCIO Y VISUALIZACIÓN
Nombre y apellidos	SARA SIRVIENTE ALONSO
Fecha entrega	

1. ¿Qué datos se usarán?

Los datos empleados en el desarrollo de este caso práctico pertenecen a la base de datos del Grupo World Bank, específicamente el conjunto de datos sobre indicadores de la burocracia mundial, donde se recoge información sobre el empleo y los salarios del sector público.

La base de datos contiene dos archivos, uno que recoge la información del país los atributos (nombre, unidad monetaria, año, préstamos, censo de población, datos industriales de comercio, etc) y su descripción correspondiente. El otro .csv es donde encuentran los datos por año. Los datos presentan una cobertura temporal de 16 años (2000-2016) y un intervalo temporal de 1 año. El paí, el código del país y el nombre del indicador

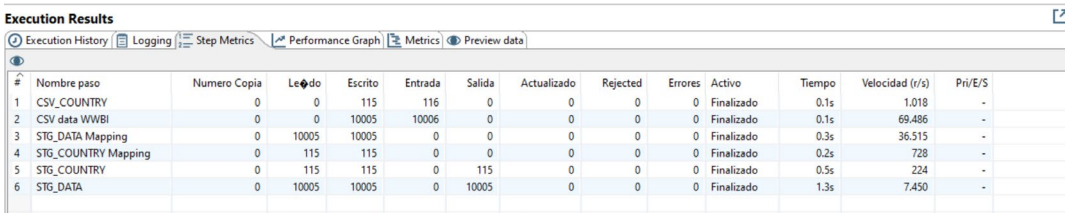
2. En este punto, se tiene que realizar la extracción de los ficheros CSV a una base de datos de *staging*, usando procesos de PDI.



Script de creación de tablas adjunto (Anexo 1). Para la carga, se procede a usar SPOON.

a. ¿Cuántas filas se han cargado en la tabla de staging País?

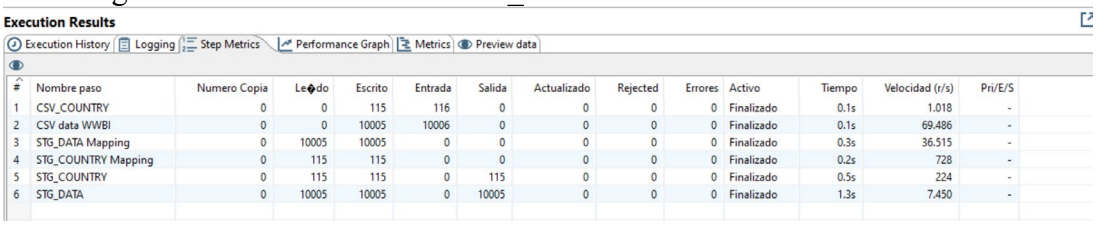
Se han cargado un total de 115 filas, lo que es lo mismo la base de datos contiene información de 115 países. Ver fila STG_COUNTRY



#	Nombre paso	Numero Copia	Le◆do	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo	Velocidad (r/s)	Pri/E/S
1	CSV_COUNTRY	0	0	115	116	0	0	0	0	Finalizado	0.1s	1.018	-
2	CSV data WWBI	0	0	10005	10006	0	0	0	0	Finalizado	0.1s	69.486	-
3	STG_DATA Mapping	0	10005	10005	0	0	0	0	0	Finalizado	0.3s	36.515	-
4	STG_COUNTRY Mapping	0	115	115	0	0	0	0	0	Finalizado	0.2s	728	-
5	STG_COUNTRY	0	115	115	0	115	0	0	0	Finalizado	0.5s	224	-
6	STG_DATA	0	10005	10005	0	10005	0	0	0	Finalizado	1.3s	7.450	-

b. ¿Cuántas filas se han cargado en la tabla de staging de datos?

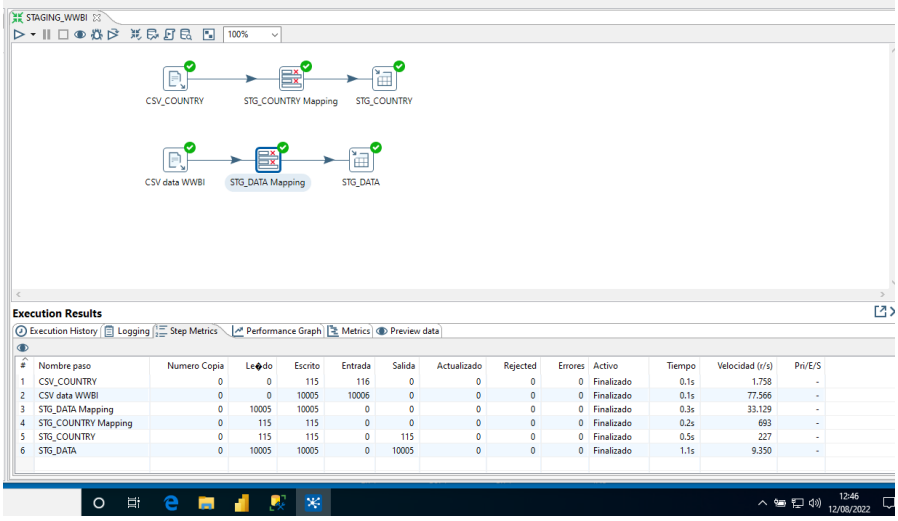
Se han cargado 10005 filas. Ver fila STG_DATA



#	Nombre paso	Numero Copia	Le◆do	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo	Velocidad (r/s)	Pri/E/S
1	CSV_COUNTRY	0	0	115	116	0	0	0	0	Finalizado	0.1s	1.018	-
2	CSV data WWBI	0	0	10005	10006	0	0	0	0	Finalizado	0.1s	69.486	-
3	STG_DATA Mapping	0	10005	10005	0	0	0	0	0	Finalizado	0.3s	36.515	-
4	STG_COUNTRY Mapping	0	115	115	0	0	0	0	0	Finalizado	0.2s	728	-
5	STG_COUNTRY	0	115	115	0	115	0	0	0	Finalizado	0.5s	224	-
6	STG_DATA	0	10005	10005	0	10005	0	0	0	Finalizado	1.3s	7.450	-

c. ¿Cuántas transformaciones has usado para realizar la carga?

He usado una transformación en la que he cargado las dos tablas, la de País y la de Datos. Primero se crea el archivo de entrada, con la base de datos correspondiente, luego la salida en tabla (seleccionando la tabla creada en SQL server previamente)



#	Nombre paso	Numero Copia	Le◆do	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo	Velocidad (r/s)	Pri/E/S
1	CSV_COUNTRY	0	0	115	116	0	0	0	0	Finalizado	0.1s	1.758	-
2	CSV data WWBI	0	0	10005	10006	0	0	0	0	Finalizado	0.1s	77.566	-
3	STG_DATA Mapping	0	10005	10005	0	0	0	0	0	Finalizado	0.3s	33.129	-
4	STG_COUNTRY Mapping	0	115	115	0	0	0	0	0	Finalizado	0.2s	693	-
5	STG_COUNTRY	0	115	115	0	115	0	0	0	Finalizado	0.5s	227	-
6	STG_DATA	0	10005	10005	0	10005	0	0	0	Finalizado	1.1s	9.350	-

d. ¿Qué objetos has usado en estas transformaciones?

e.

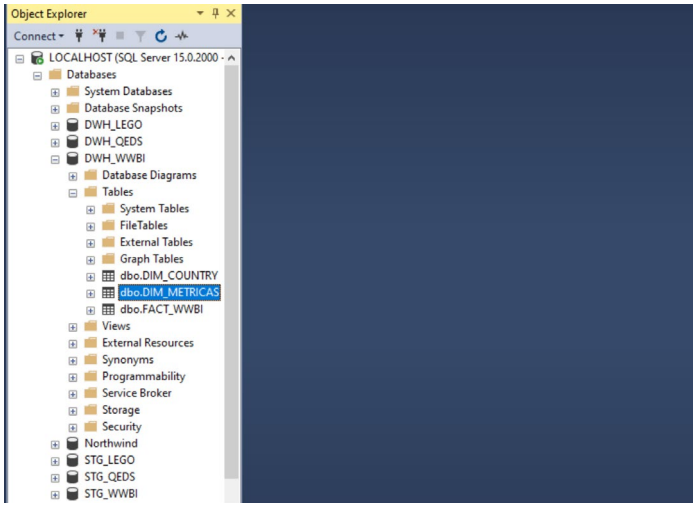
Se ha usado un objeto tipo mapping

Se ha realizado un mapeo para vincular los campos que queremos que se haga la transformación, es decir que source con que target. Y este paso se ha incluido entre el origen y el destino, para que la asignación de las columnas sea correcta.

f. ¿Has usado el componente Start?

No, el objeto Start se utiliza en los trabajos (job). El trabajo siempre debe de empezar por este objeto. Cómo en este punto del ejercicio no se nos ha pedido aún realizar trabajos no se hacen por el momento. El trabajo lo definiremos una vez tengamos todo diseñada área staging, tabla de hechos y de dimensiones para el datamart.

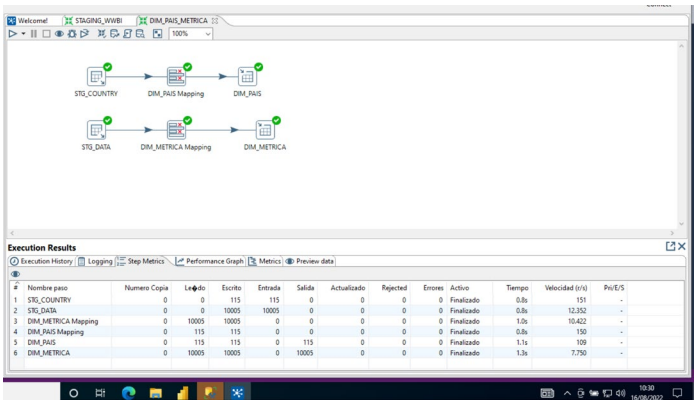
3. **En este punto se tiene que realizar las transformaciones y carga de los datos desde la base de datos de staging al data warehouse.**
   Para ello, se debe crear una base de datos data warehouse y sus tablas. Estas se cargarán usando PDI.



Script de creación de tablas adjunto (Anexo 2). Para la carga, se procede a usar SPOON.

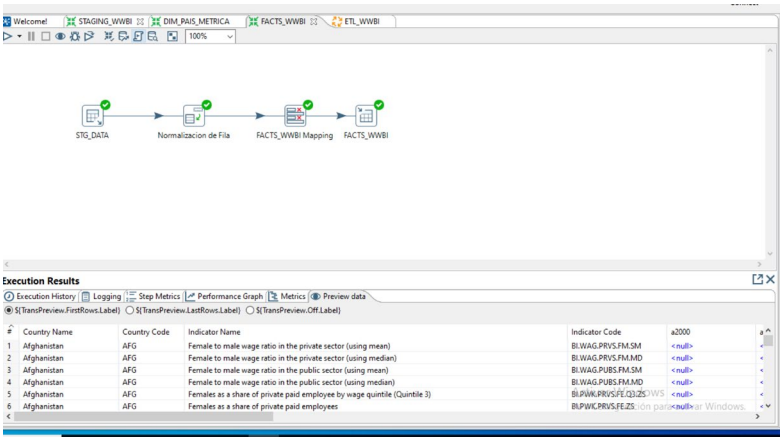
A. ¿Cómo se ha cargado la tabla "dim_metrica"? ¿Cuál es su origen?

La tabla de DIM_Metrica se ha cargado a partir de la tabla de datos de la base de datos de staging. Se decide usar las columnas de "Indicator code" e "Indicator name" ya que estas columnas contienen información variada que permitirá desarrollar el informe.
Es decir, esta tabla tiene su origen de la base de datos de staging STG_DATA.



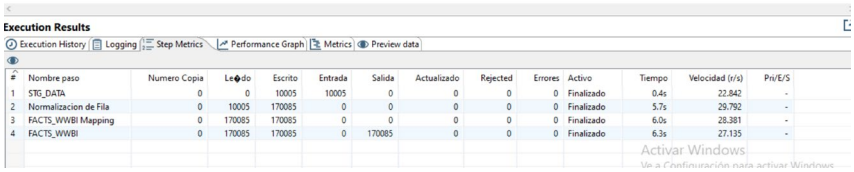
B. ¿Qué componentes se han usado para crear la tabla de hechos?

Para la tabla de hechos, se han utilizo id_pais , id_metrica, id_año e id_value. Para poder crear las columnas de id_año e id_value y cargarlas con la base de datos de staging, previamente se ha generado una normalización de la columna año y valor.



C. ¿Cuántas filas se han cargado en la tabla de hechos?

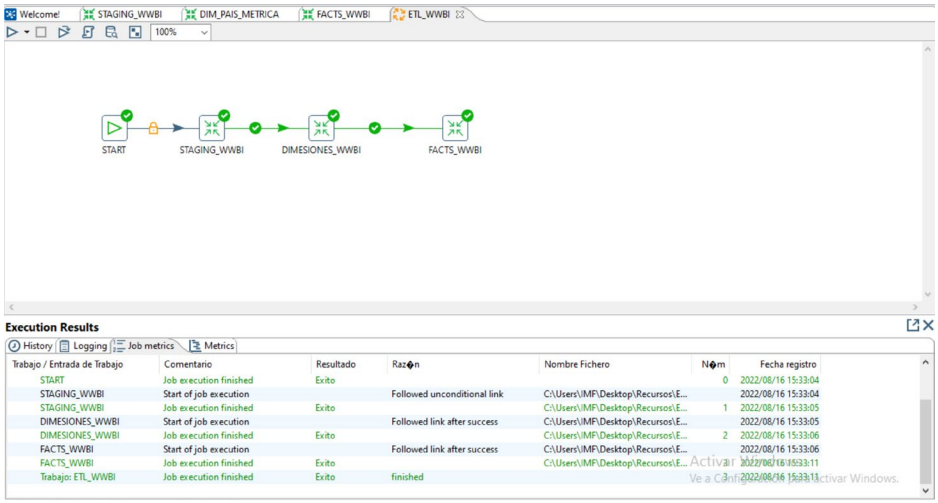
Se han cargado 170085 filas en la tabla de hechos, ya que se partía de1005 pero se han multiplicado por 17 a la hora de normalizar las columnas para ponerlas como filas.



D. ¿Por qué se han multiplicado el número de filas de la tabla de hechos?

Al haber pivotado las columnas las filas (1005) se han multiplicado por 17 (tenemos 17 años, del 2000 al 2016 ambos inclusive) por lo que obtenemos un numero final de filas mucho mayor (170085).

4. Crear la tarea que permita cargar todo el datamart desde los orígenes > staging > datamart.



A. ¿Se ha usado una transformación o una tarea?

Se ha usado una tarea o trabajo (job)

B. ¿Por qué?

Desde Spoon hemos generado los paso para la extracción, transformación y cargas de los datos, sincronizadas mediante saltos que indican el orden de los pasos. A través de un trabajo, se consigue agrupar estos pasos y que se desarrollen todos a partir de un único ejecutable.

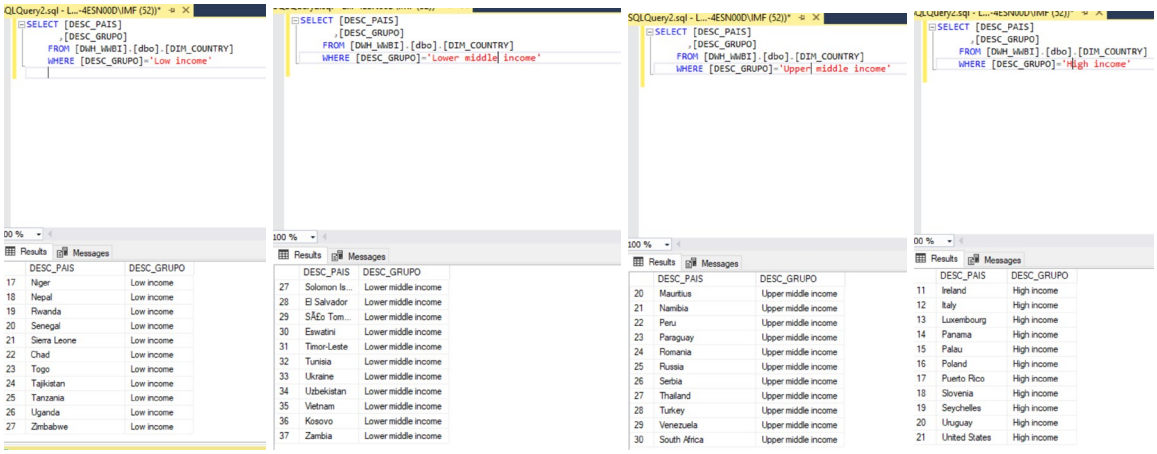
C. ¿Qué tipo de objetos se han usado?

Primero se ha utilizado un objeto Start, el cuál no necesita configuración. El resto de los objetos utilizados son transformaciones. Se configura cada transformación de manera independiente indicando la ruta de dónde esta el fichero de esa transformación que se ha creado previamente.

5. Responder las siguientes preguntas realizando consultas SQL:

A. ¿Cuántos países pertenecen a cada grupo de ingresos (income group)?

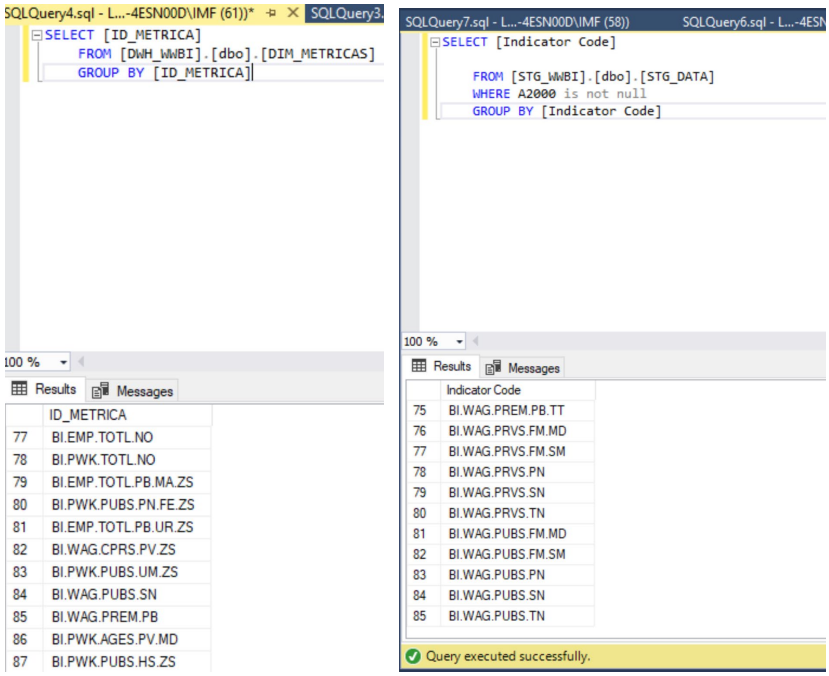
De un total de 115 paises:
Low income: 27 países, lower middle income: 37 países Upper middle income: 30 países y High income: 21 países.
Cómo son pocos valores diferentes (income group) podemos ir viendo uno a uno.



B. ¿Cuántas métricas existen? ¿Y que tengan valor no nulo en el año 2000?

Existen 87 métricas diferentes. Pero, solo 85 que no tengan valor nulo en el año 2000. Esto lo puedo averiguar ya que la dimensión de métricas fue creada sin ningún tipo de filtro, por lo que si le pedimos a través de SQL que nos indique cuantos grupos diferentes hay se obtienen las 87 métricas. Esto también se puede averiguar realizando la misma consulta en vez de a DIM_METRICAS a la tabla de hechos, obteniéndose el mismo resultado, 87 filas diferentes.

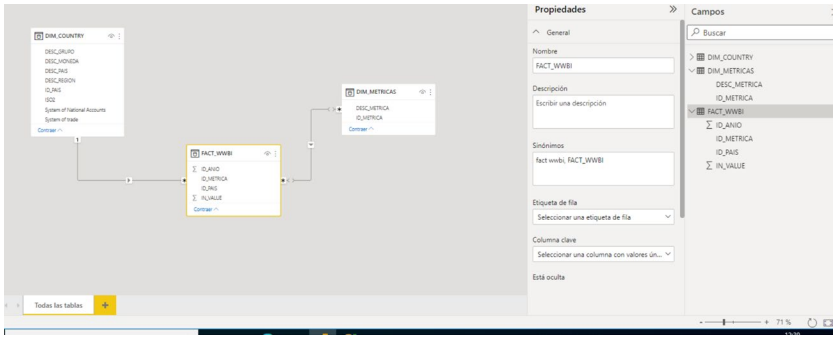
Sin embargo, para poder aplicar el filtro y realizar la consulta para aquellos que tengan valor no nulo en el año 2000, usamos la tabla de datos de la base de staging en la que tenemos el año como columna y no como fila. Se obtienen 85 filas diferentes que tienen valor no nulo en el año 2000.



6. **Crear un informe en Power Bi accediendo a la información del datamart recién cargado. Indicar la estructura del modelo de datos. Definir las tablas, sus relaciones y cardinalidades.**

En la siguiente figura, se puede ver el modelo de datos creado por Power Bi.
Se han cargado las tablas del Datamat y se han realizado las relaciones entre ellas. La primera que se observa es la relación 1-n de DIM_METRICAS a la tabla de hechos (ID_PAIS), esto es así ya que en la tabla de DIM_COUNTRY, "ID_PAIS" es un primary key, lo que indica que no contiene valores duplicados, sin embargo en la tabla de hechos sí estarán los valores duplicados no es (PK) por ello encontramos la conexión 1-n.

La otra relación que se presenta es la tabla de DIM_METRICAS con la tabla de hechos a través de la columna "ID_METRICA", esta relación es n-n ya que en ambas tablas los valores de ID_METRICA no son únicos, si no que están repetidos.



7. **Crear las siguientes visualizaciones, adjuntar comentarios de por qué se eligió cada tipo de visualización, así como capturas de pantalla con los gráficos.**

   A. **Evolución en el tiempo del "Empleo del sector público como parte del empleo remunerado" y el "Empleo del sector público como parte del empleo formal" para Argentina.**

   Para poder visualizar esto, se ha utilizado un grafico de series temporales, ya que estos permiten observar la variación de estas dos métricas a lo largo de los diferentes años, es decir, del tiempo. Cuando se quiere evaluar la evolución de una variable a lo largo del tiempo es recomendable utilizar este tipo de gráficos ya que son bastante claros y proporcionan mucha información de manera visual (comportamiento pasado, presente, futuro, tendencias, anomalías anuales, etc).

   Para filtrar los datos, se han utilizado el campo de país, seleccionando únicamente Argentina y descripción de las métricas (DESC_METRICAS) dónde únicamente se seleccionan "Public sector employment as a share of paid employment" y "Public sector employment as a share of formal employment".
   Como eje X se representa el campo: ID_ANIO que es el que contiene la información en el tiempo. Como eje Y, se utiliza el campo: ID_VALUE, que es el campo dónde están contenidos los valores de esas métricas para cada año. Como Leyenda: DESC_METRICAS

   B. **Evaluar la edad media de los empleados del sector privado y público por región.**

Para analizar esto, se utiliza un gráfico de columnas apiladas. Este tipo de gráficos nos permiten visualizar información en barras superpuestas lo que resulta bastante útil para evaluar como son las edades medias de los empleados tanto del sector público como privado por región del mundo.
Para filtrar los datos, se utiliza el campo de descripción de la métrica en la tabla de DIM_METRICAS y se seleccionan las dos métricas que se quieren ("Mean age of private paid employees" y "Mean age of public paid employees".

Como eje X, se utiliza la desc_region de la DIM_COUNTRY, dónde encontramos la información de cada una de las regiones. El eje Y corresponde con el valor promedio de IN_VALUE (la columna de valores) y la leyenda es DESC_METRICA (la columna dónde está la descripción de cada métrica).

C. Realizar una gráficadel promedio del peso relativo de los cargos técnicos en los sectores privados y públicos a lo largo del tiempo.
La gráfica debe permitir ver el total volumen de cada métrica y el total de ambas. Las métricas que se han de usar son las siguientes:

En este caso, se decide usar un gráfico de áreas apiladas, ya que este tipo de gráficos permiten representar el volumen de cada una, pero al ser formato apilado, se puede saber cuánto aporta al total cada valor de la dimensión.

Se utiliza DESC_METRICAS para filtrar los campos y quedarnos solo con "Relative wage of technicians in private sector (using clerk as reference)" y "Relative wage of technicians in public sector (using clerk as reference)". El eje X representa el tiempo (ID_ANIO) y el eje Y los valores (ID_VALUE). En la leyenda representamos el campo DESC_METRICA.

D. Obtener el promedio del peso por región del gasto en empleados públicos respecto al GDP y el gasto público.
Para este caso, se utiliza un grafico de barras agrupadas, ya que así se muestra de una manera clara ambas métricas. Estos tipos de gráficos ayudan a resumir los datos por grupos, en este caso por regiones. Se puede ver el valor promedio de cada métrica para cada región.
Como filtro, se utiliza DESC_METRICA y se seleccionan las dos métricas que pide el ejercicio. Como eje X se utiliza el valor seleccionando promedio (IN_VALUE) y como valor o eje y, se utiliza la DESC_REGION. La leyenda es el campo DESC_METRICA.