Hoja de respuestas

Módulo	Inteligencia de negocio y Visualización
Nombre y apellidos	Lucía López Fuentes
Fecha entrega	15/03/2023

Un análisis de los datos de origen en el que se detallen los campos que se encuentran en los ficheros.
Se van a cargar a staging y al data mart, así como un diagrama entidad-relación del modelo.

¿Qué datos se usarán?

Se dispone de dos ficheros CSV (WWBICountry.csv y WWBIData.csv), de los cuales una servirá para obtener la métrica.

WWBIData
Columnas:
Country Name Country Code Indicator Name Indicator Code a2000 a2001 a2002 a2003 a2004
a2005 a2006 a2007 a2008 a2009 a2010 a2011 a2012 a2013 a2014 a2015 a2016

WWBICountry
Columnas:
Country Code Short Name Table Name Long Name 2-alpha code Currency Unit Special Notes Region Income Group WB-2 code National accounts base year National accounts reference
year SNA price valuation Lending category Other groups System of National Accounts Alternative conversion factor PPP survey year Balance of Payments Manual in use External debt Reporting status System of tradeGovernment Accounting concept IMF data dissemination standard Latest population census Latest household survey Source of most recent Income and expenditure data Vital registration complete Latest agricultural census Latest industrial data Latest trade data

En este punto, se tiene que realizar la extracción de los ficheros CSV a una base de datos de staging, usando procesos de PDI.
Se debe crear la base de datos de staging y sus tablas, así como los procesos de carga.
Primer paso: crear base de datos de staging y sus tablas, adjuntar el script de creación de las tablas.

Se crea la base de datos de staging (STG_WWBI) y las tablas STG_COUNTRY y STG_DATA donde cargaremos los datos en bruto en formato string, es el primer paso para realizar nuestra ETL.

```
USE [STG_WWBI]
GO

SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[STG_COUNTRY](
    [Country Code] [varchar](250) NULL,
    [Short Name] [varchar](250) NULL,
    [Table Name] [varchar](250) NULL,
    [Long Name] [varchar](250) NULL,
    [2-alpha code] [varchar](250) NULL,
    [Currency Unit] [varchar](250) NULL,
    [Special Notes] [varchar](2500) NULL,
    [Region] [varchar](250) NULL,
    [Income Group] [varchar](250) NULL,
    [WB-2 code] [varchar](250) NULL,
    [National accounts base year] [varchar](250) NULL,
    [National accounts reference year] [varchar](250) NULL,
    [SNA price valuation] [varchar](250) NULL,
    [Lending category] [varchar](250) NULL,
    [Other groups] [varchar](250) NULL,
    [System of National Accounts] [varchar](250) NULL,
    [Alternative conversion factor] [varchar](250) NULL,
    [PPP survey year] [varchar](250) NULL,
    [Balance of Payments Manual in use] [varchar](250) NULL,
    [External debt Reporting status] [varchar](250) NULL,
    [System of trade] [varchar](250) NULL,
    [Government Accounting concept] [varchar](250) NULL,
    [IMF data dissemination standard] [varchar](250) NULL,
    [Latest population census] [varchar](250) NULL,
    [Latest household survey] [varchar](250) NULL,
    [Source of most recent Income and expenditure data] [varchar](250) NULL,
    [Vital registration complete] [varchar](250) NULL,
    [Latest agricultural census] [varchar](250) NULL,
    [Latest industrial data] [varchar](250) NULL,
    [Latest trade data] [varchar](250) NULL
) ON [PRIMARY]
GO
```

```
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[STG_DATA](
    [Country Name] [varchar](250) NULL,
    [Country Code] [varchar](250) NULL,
    [Indicator Name] [varchar](250) NULL,
    [Indicator Code] [varchar](250) NULL,
    [a2000] [varchar](250) NULL,
    [a2001] [varchar](250) NULL,
    [a2002] [varchar](250) NULL,
    [a2003] [varchar](250) NULL,
    [a2004] [varchar](250) NULL,
    [a2005] [varchar](250) NULL,
    [a2006] [varchar](250) NULL,
    [a2007] [varchar](250) NULL,
    [a2008] [varchar](250) NULL,
    [a2009] [varchar](250) NULL,
    [a2010] [varchar](250) NULL,
    [a2011] [varchar](250) NULL,
    [a2012] [varchar](250) NULL,
    [a2013] [varchar](250) NULL,
    [a2014] [varchar](250) NULL,
    [a2015] [varchar](250) NULL,
    [a2016] [varchar](250) NULL
) ON [PRIMARY]
GO
```

Respecto los procesos de carga, responder las siguientes preguntas:

ecution Results

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

Nombre paso	Numero Copia	Le?do	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo
CSV_COUNTRY	0	0	115	116	0	0	0	0	Finalizado	0.1s
CSV_DATA	0	0	10005	10006	0	0	0	0	Finalizado	0.1s
STG_DATA Mapping	0	10005	10005	0	0	0	0	0	Finalizado	0.2s
STG_COUNTRY Mapping	0	115	115	0	0	0	0	0	Finalizado	0.1s
STG_COUNTRY	0	115	115	0	115	0	0	0	Finalizado	0.3s
STG_DATA	0	10005	10005	0	10005	0	0	0	Finalizado	0.7s

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

Se cargan en el staging 115.

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

De un total de 10006 filas de entrada se cargan en el staging 10005.

¿Cuántas transformaciones has usado para realizar la carga?

1 Por cada CSV ya que la cantidad de datos es relativamente pequeña.

¿Qué objetos has usado en estas transformaciones?

CSV file input: nos viene marcado por el tipo de fichero que tenemos en origen, el cual nos permite cargar este tipo de ficheros.

Salida Tabla: es un paso que nos permite la carga de la información a la base de datos

Mapping: nos permite el mapeado del nombre de los campos entre el INPUT-OUTPUT

¿Has usado el componente Start?

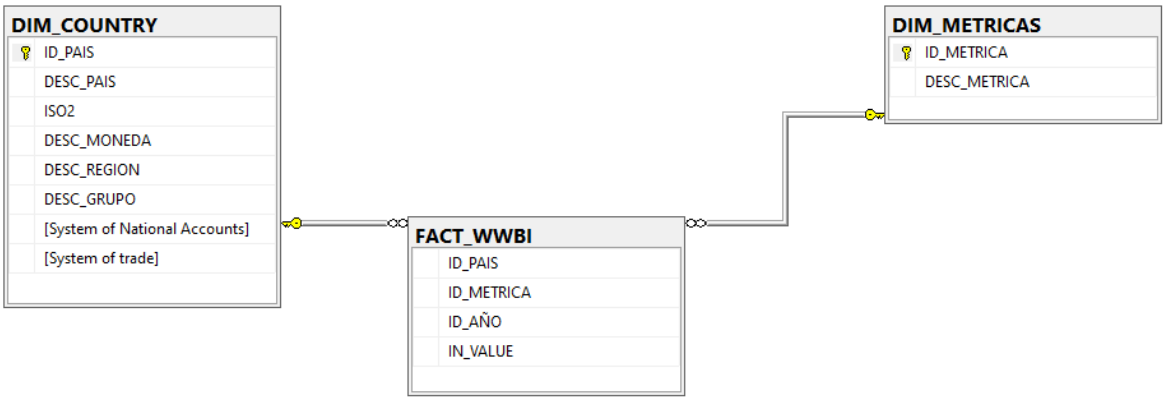
Sí pero más adelante para la elaboración del Job, aquí no ya que es una transformación.

En este punto se tiene que realizar las transformaciones y carga de los datos desde la base de datos de staging al data warehouse.

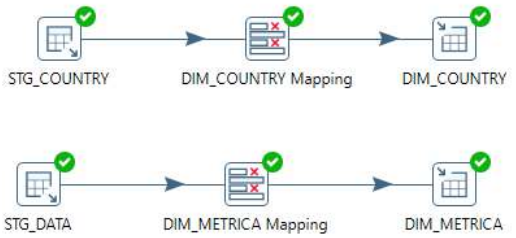
Para ello, se debe crear una base de datos data warehouse y sus tablas. Estas se cargarán usando PDI.

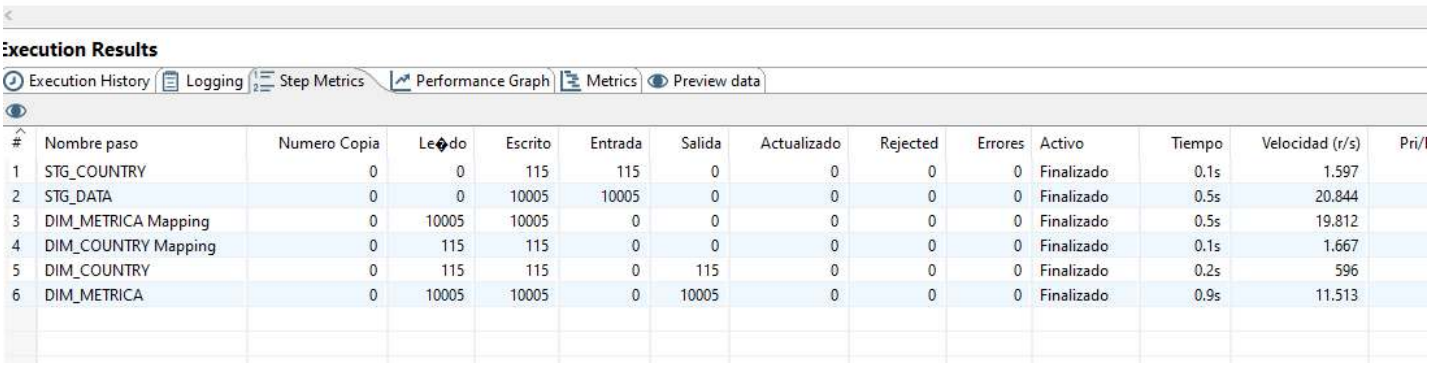
Crear el datamart WWBI y las tablas. Adjuntar scripts de creación de tablas.

```
CREATE TABLE [dbo].[DIM_COUNTRY](
    [ID_PAIS] [varchar](3) NOT NULL,
    [DESC_PAIS] [varchar](250) NULL,
    [ISO2] [varchar](250) NULL,
    [DESC_MONEDA] [varchar](250) NULL,
    [DESC_REGION] [varchar](250) NULL,
    [DESC_GRUPO] [varchar](250) NULL,
    [System of National Accounts] [varchar](250) NULL,
    [System of trade] [varchar](250) NULL,
PRIMARY KEY CLUSTERED
(
    [ID_PAIS] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON, OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]
GO


SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[DIM_METRICAS](
    [ID_METRICA] [varchar](25) NOT NULL,
    [DESC_METRICA] [varchar](1000) NULL,
PRIMARY KEY CLUSTERED
(
    [ID_METRICA] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = ON, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON, OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]
GO


SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[FACT_WWBI](
    [ID_PAIS] [varchar](3) NULL,
    [ID_METRICA] [varchar](25) NULL,
    [ID_AÑO] [int] NULL,
    [IN_VALUE] [float] NULL
) ON [PRIMARY]
GO
```



Realizar las transformaciones necesarias para cargar el datamart usando PDI.





Execution Results

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

#	Nombre paso	Numero Copia	Le♦do	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo	Velocidad (r/s)	Pri/
1	STG_COUNTRY	0	0	115	115	0	0	0	0	Finalizado	0.1s	1.597	
2	STG_DATA	0	0	10005	10005	0	0	0	0	Finalizado	0.5s	20.844	
3	DIM_METRICA Mapping	0	10005	10005	0	0	0	0	0	Finalizado	0.5s	19.812	
4	DIM_COUNTRY Mapping	0	115	115	0	0	0	0	0	Finalizado	0.1s	1.667	
5	DIM_COUNTRY	0	115	115	0	115	0	0	0	Finalizado	0.2s	596	
6	DIM_METRICA	0	10005	10005	0	10005	0	0	0	Finalizado	0.9s	11.513	

Responder las siguientes preguntas:

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

Utilizando el DBMS se crea la base de datos de staging (DWH_WWBI), donde alojaremos dos tablas de dimensiones y una tabla de hechos. Se decide crear estas dos dimensiones porque la de país nos viene dado por el propio archivo de datos, por otro lado, la

dimensión métrica la creamos para enriquecer nuestro modelo multidimensional y explotar mejor nuestros procesos.

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

Entrada Tabla: nos permite la carga información desde una BBDD mediante una consulta SQL.
Salida Tabla: es un paso que nos permite la carga de la información a la BBDD.
Mapping: nos permite el mapeado del nombre de los campos entre el INPUT-OUTPUT.
Normalización de fila: pasa los datos que tenemos en WIDE format a LONG format, en este caso lo hacemos con los años.

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

10005 filas

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

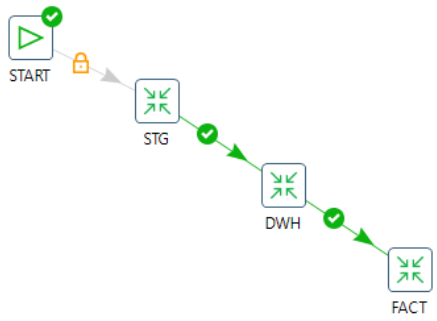
Como hemos realizado el cambio de WIDE format a LONG format se han multiplicado el número de filas por 17.

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

Execution Results

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

#	Nombre paso	Numero Copia	Le◆do	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo	Velocidad (r/s)
1	STG_DATA	0	0	10005	10005	0	0	0	0	Finalizado	0.0s	208.438
2	Normalizacion de Fila	0	10005	170085	0	0	0	0	0	Finalizado	2.4s	72.162
3	FACT_DATA Mapping	0	170085	170085	0	0	0	0	0	Finalizado	2.5s	67.817
4	FACT_DATA	0	170085	170085	0	170085	0	0	0	Finalizado	2.7s	64.062



cution Results

History | Logging | Job metrics | Metrics

bajo / Entrada de Trabajo	Comentario	Resultado	Raz�n	Nombre Fichero	N�m	Fecha registro
ETL_WWBI						
Trabajo: ETL_WWBI	Start of job execution		start			2023/03/14 09:47:26
START	Start of job execution		start			2023/03/14 09:47:26
START	Job execution finished	Exito			0	2023/03/14 09:47:26
Trabajo: ETL_WWBI	Job execution finished	Exito	finished		0	2023/03/14 09:47:26

¿Se ha usado una transformación o una tarea?
¿Por qué?
Una tarea, porque es lo necesario para conseguir un proceso automatizado.

¿Qué tipo de objetos se han usado?

Start: Es la primera entrada del job, la cual nos permite programar el calendario del día y hora de ejecución
Transformaciones: Cargamos los ficheros con las transformaciones que se van a ejecutar.

Responder las siguientes preguntas realizando consultas SQL:
¿Cuántos países pertenecen a cada grupo de ingresos (income group)?

```
USE [STG_WWBI]
GO

SELECT [Income Group]
      ,COUNT([Short Name]) Países
      FROM [STG_WWBI].[dbo].[STG_COUNTRY]
      GROUP BY [Income Group]
```

90 %

Results | Messages

	Income Group	Países
1	High income	21
2	Low income	27
3	Lower middle income	37
4	Upper middle income	30

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

```
SELECT DISTINCT
COUNT ([Indicator Name]) Metricas
FROM [STG_WWBI].[dbo].[STG_DATA]
GROUP BY [Indicator Name]
```

90 %

Results | Messages

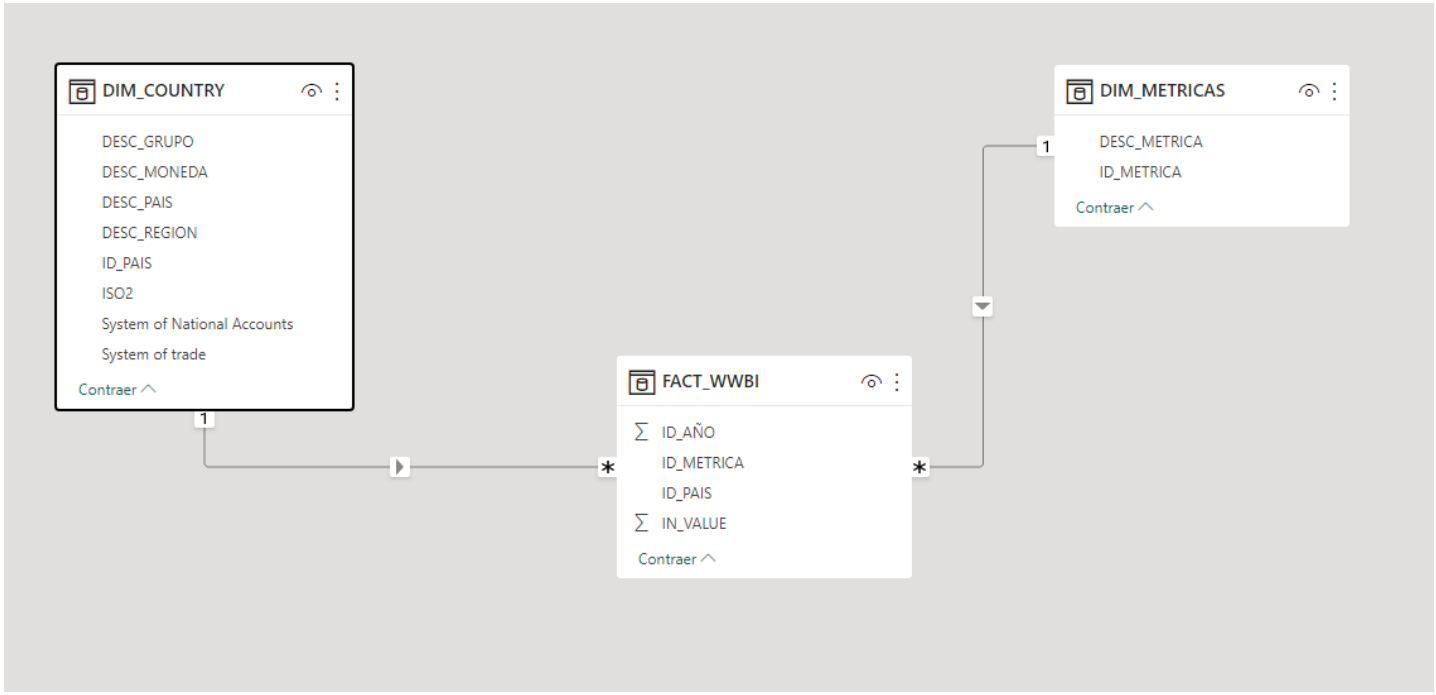
	Metricas
1	115

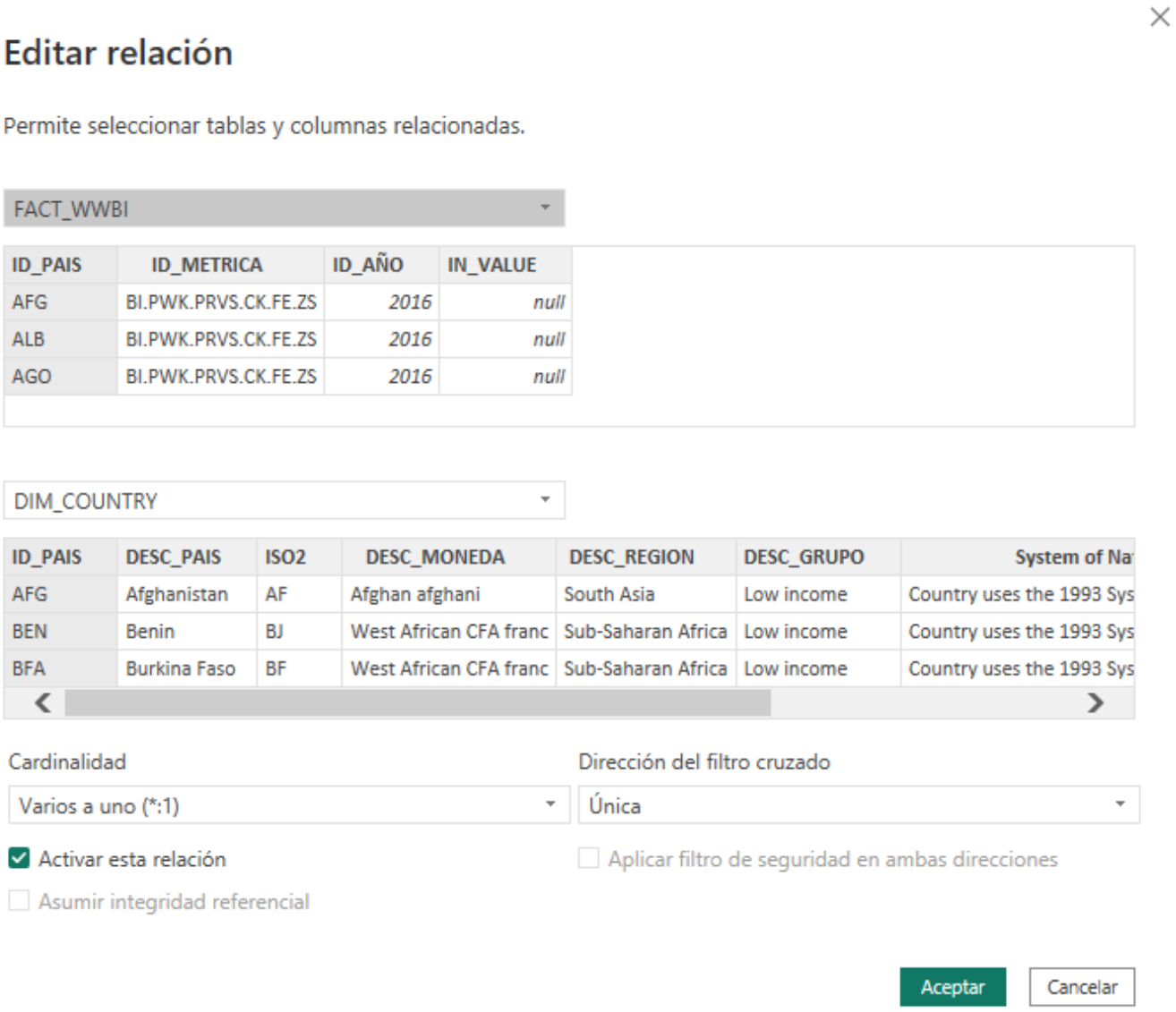
Crear un informe en Power Bi accediendo a la información del datamart recién cargado.
Indicar la estructura del modelo de datos. Definir las tablas, sus relaciones y cardinalidades

Conectamos Power BI con la BBDD local que tenemos y extraemos las 2 tablas de dimensiones y la tabla de hechos.

Tablas:

DIM_METRICAS: Es una tabla que almacena la información de los diferentes KPI'S.
DIM_COUNTRY: En esta tabla almacenamos toda la información relativa a los diferentes países.
FACT_WWBI: La tabla de hechos que recoge los valores de los diferentes KPI's en los años en los que se han medido.

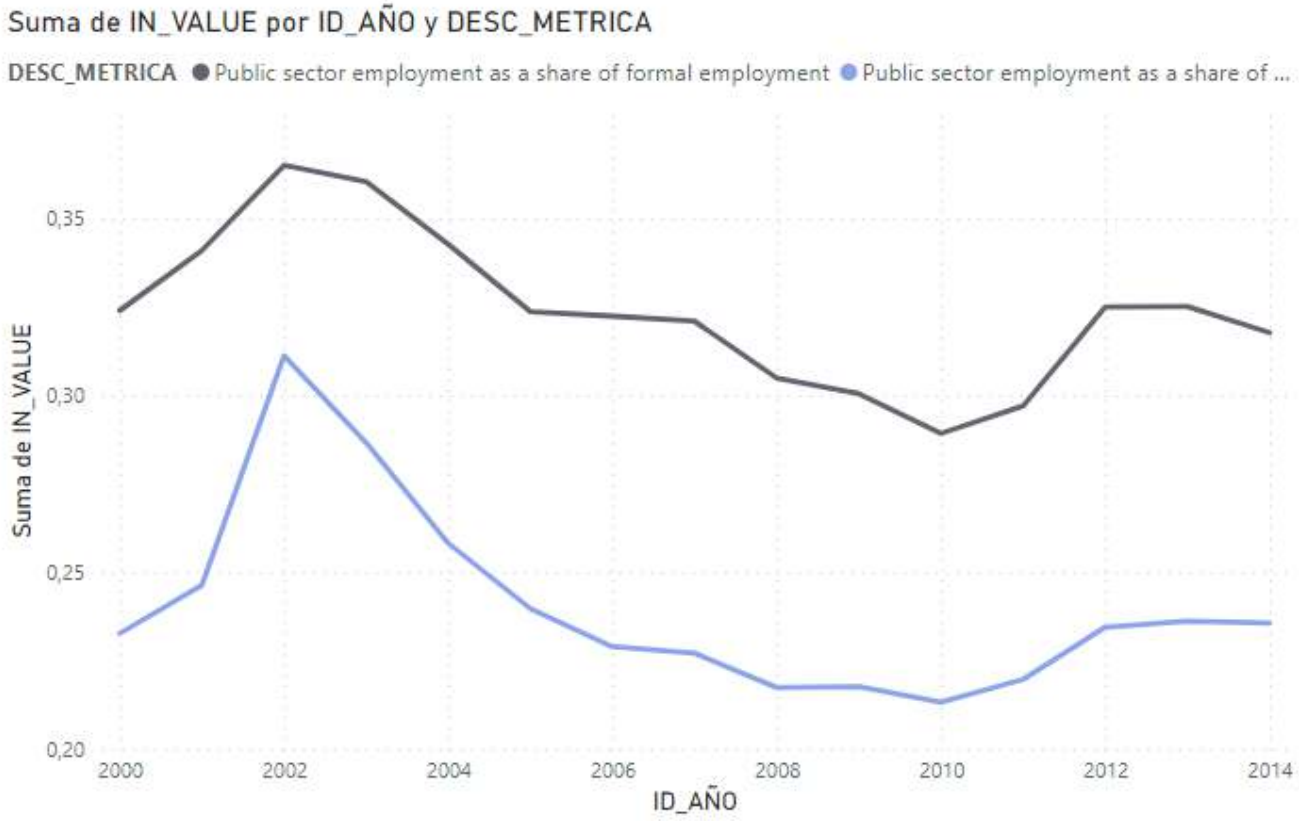

DIM_COUNTRY
DESC_GRUPO
DESC_MONEDA
DESC_PAIS
DESC_REGION
ID_PAIS
ISO2
System of National Accounts
System of trade
Contraer
1
DIM_METRICAS
DESC_METRICA
ID_METRICA
Contraer
1
FACT_WWBI
ID_AÑO
ID_METRICA
ID_PAIS
IN_VALUE
Contraer
*
*



FACT_WWBI to DIM_METRICAS
La tabla de hechos con la dimensión se relaciona mediante el campo ID_METRICA de las dos tablas y tienen
una cardinalidad N:1 FACT_WWBI to COUNTRY
La tabla de hechos con la dimensión se relaciona mediante el campo ID_COUNTRY de las dos tablas y tienen una
cardinalidad N:1

Crear las siguientes visualizaciones, adjuntar comentarios de por qué se eligió cada tipo de visualización, así como capturas de pantalla con los gráficos.
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.
Las métricas que se han de usar son las siguientes:
Public sector employment as a share of paid employment".
"Public sector employment as a share of formal employment".

1.¿Qué tipo de gráfico se ha usado y por qué?



El gráfico utilizado es el de línea, ya que se puede apreciar correctamente las tendencias existentes de las métricas solicitadas con el paso del tiempo.

¿Qué campo se ha usado para filtrar los datos?

Se ha filtrado el país mediante el campo DESC_PAIS (Argentina)
DESC_METRICA ("Public sector employment as a share of paid employment", "Public sector employment as a share of formal employment")

¿Qué campo se ha usado para el eje de la gráfica?

Para el eje X se ha empleado el campo ID_AÑO para expresar una serie temporal.

Para el eje Y se ha empleado la suma de IN_VALUE para expresar la totalidad de los valores del mismo año.

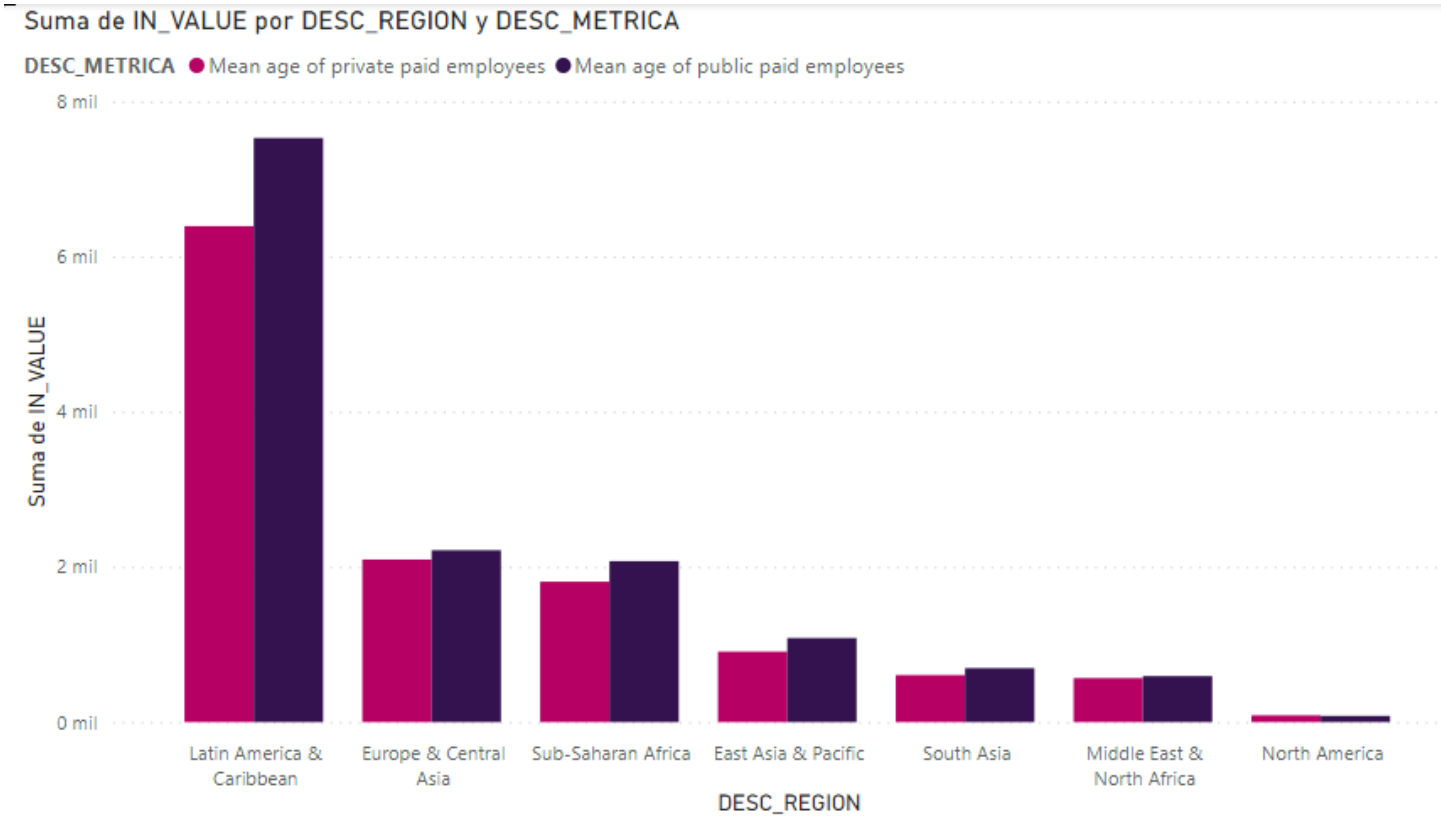
¿Y en la leyenda?
Se ha usado en leyenda el campo DESC_METRICA

¿Qué campo se usó para mostrar como valores?

Se rellena los valores con el campo IN_VALUE.

Evaluar la edad media de los empleados del sector privado y público por región. Las métricas que se han de utilizar son las siguientes:

"Mean age of private paid employees".
"Mean age of public paid employees".



¿Qué tipo de gráfico se ha usado y por qué?

Se ha utilizado el gráfico de columnas agrupadas para poder comprar entre los trabajadores públicos y los privados.

¿Qué campo se ha usado para filtrar los datos?

Se ha filtrado la métrica por el campo DESC_METRICA (“Mean age of private paid employees”, “Mean age of public paid employees”)

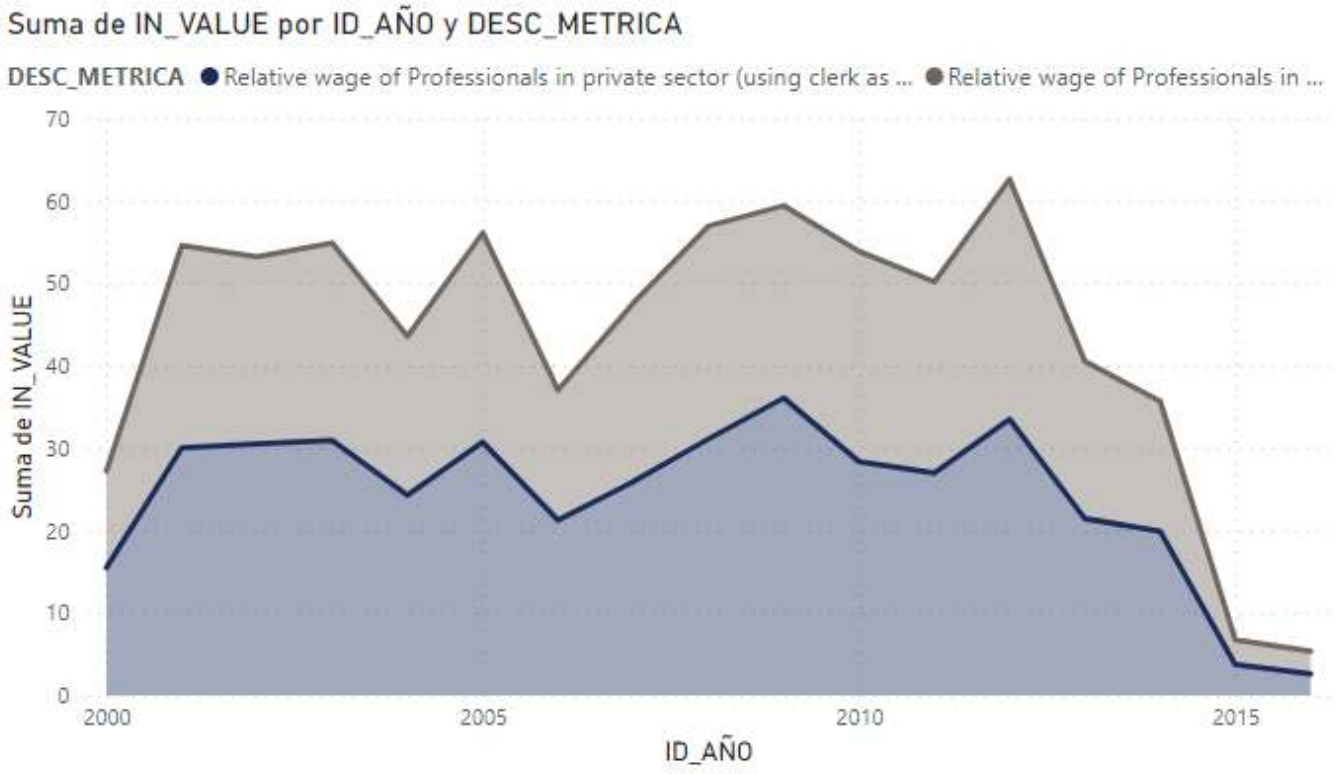
¿Qué campo se ha usado para el eje de la gráfica?

Para el eje X se ha usado el campo DESC_REGION, para agrupar por regiones.
Para el eje Y se ha empleado el promedio de IN_VALUE para expresar el promedio de cada métrica en la misma región para poder compararlo con el sector público vs privado.

¿Y en la leyenda?

DESC_METRICA

Realizar una gráfica del 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:
“Relative wage of technicians in private sector (using clerk as reference)”.
“Relative wage of technicians in public sector (using clerk as reference)”



¿Qué tipo de gráfico se ha usado y por qué?

Se ha usado un gráfico de áreas apiladas porque se ha querido representar una variable cuantitativa continua en una escala temporal y obtener una idea de su peso respecto la totalidad.

¿Qué campo se ha usado para filtrar los datos?

DESC_METRICA (“Relative wage of technicians in private sector (using clerk as reference)”, “Relative wage of technicians in public sector (using clerk as reference))

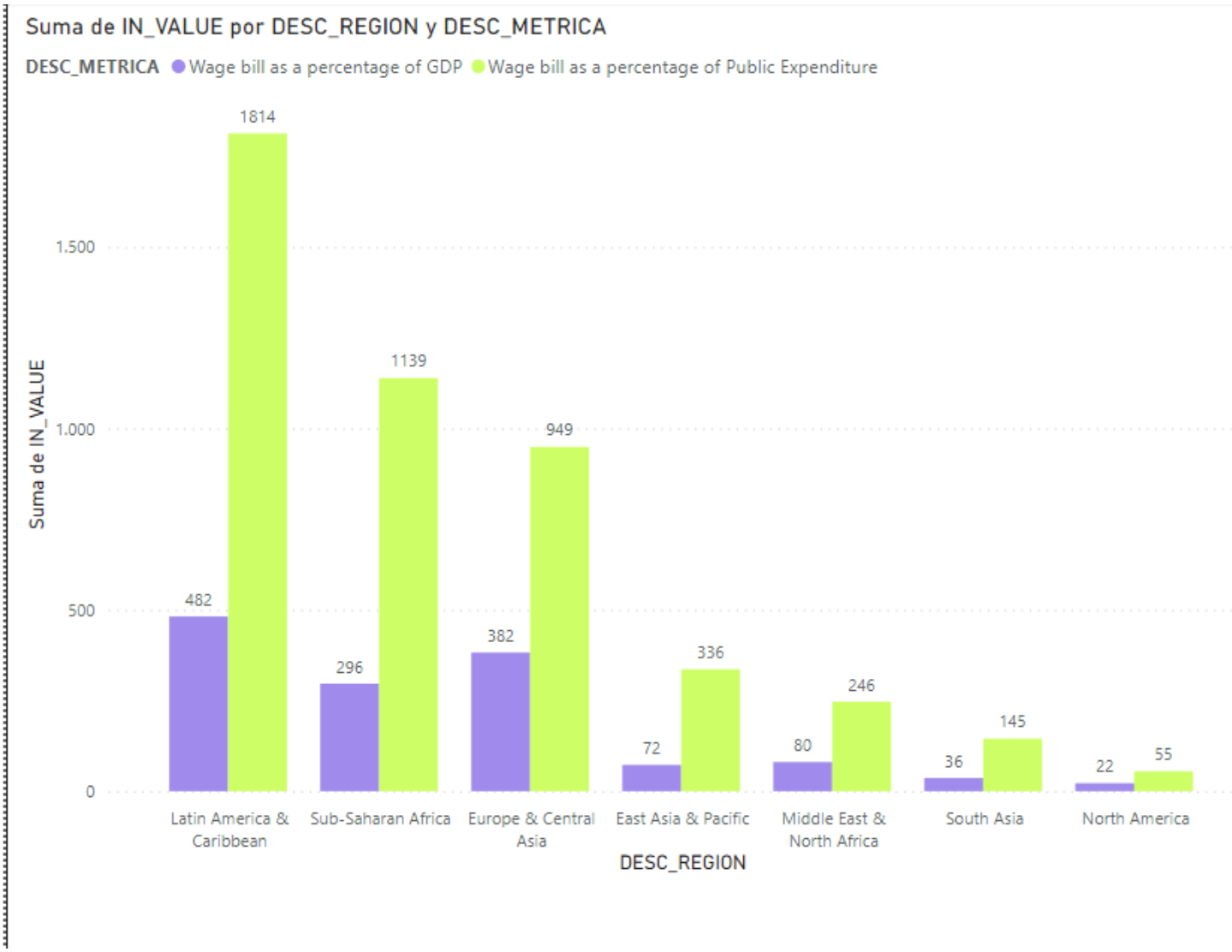
¿Qué campo se ha usado para el eje de la gráfica?

Para el eje X se ha empleado el campo ID_AÑO para expresar una serie temporal.
Para el eje Y se ha empleado el promedio del campo IN_VALUE para expresar el promedio de cada métrica en el mismo año para poder compararlo con el sector público vs privado.

¿Y en la leyenda?

DESC_METRICA

Obtener el promedio del peso por región del gasto en empleados públicos respecto al GDP y el gasto público.
Las métricas que se han de usar son las siguientes:
“Wage bill as a percentage of GDP”.
“Wage bill as a percentage of Public Expenditure”.



¿Qué tipo de gráfico se ha usado y por qué?

Se ha usado un gráfico de columnas agrupadas por que se ha querido comparar el promedio de cada gasto en sus respectivas regiones.

¿Qué campo se ha usado para filtrar los datos?

DESC_METRICA (“Wage bill as a percentage of GDP”, “Wage bill as a percentage of Public Expenditure”)

¿Qué campo se ha usado para el eje de la gráfica?

Para el eje X se ha empleado el campo DESC_REGION, para agrupar por regiones.
Para el eje Y se ha empleado el promedio del campo IN_VALUE para expresar el promedio de cada métrica en la misma región para poder compararlo con el sector público vs privado.

¿Y en la leyenda?

DESC_METRICA

¿Y en los valores?

IN_VALUE