

Caso Práctico: Sistema Integrado de Egresados Universitarios.

JUAN PABLO BOTERO SUAZA

Universitat Oberta de Catalunya

Índice

- Introducción
- Contexto
- Usuarios potenciales
- Fuentes de datos
- Análisis de requerimientos
- Análisis de fuentes de datos.
- Análisis funcional.
- Diseño del modelo conceptual, lógico y físico del almacén de datos
- Metadatos
- Revisar el diseño conceptual, lógico y físico
- Diseño del proceso ETL
- Implementación del proceso ETL
- Carga de datos
- Explotación de datos
- Programas
- Bibliografía

Introducción

Este material, titulado “**Sistema Integrado de Egresados Universitarios**”, está creado para practicar el diseño e implementación del núcleo de toda herramienta de inteligencia de negocio: el almacén de datos.

El diseño, desarrollo e implantación de un sistema de *Data Warehouse (DW)* en cualquier organización supone llevar a cabo un proyecto que puede durar meses, o incluso años, en función del alcance del proyecto, de la naturaleza y el grado de madurez de la organización, así como de la participación de equipos multidisciplinares que van implementando diferentes proyectos en un proceso de mejora continua del almacén.

El objetivo de este caso no es desarrollar un almacén de datos que dé respuesta a todas las necesidades, sino entender y utilizar las metodologías para desarrollar este tipo de proyectos en un contexto real, pasando por todas las fases que comprenden proyectos de esta tipología:

1. **Diseño e implementación:** consiste en desarrollar e implementar un almacén de datos que permita la gestión de la información disponible.
2. **Carga:** implica diseñar e implementar los procesos de carga de datos necesarios para disponer de información en el almacén de datos implementado.
3. **Explotación:** conlleva la creación de informes, elementos de análisis multidimensional, cuadros de mando, etc. para la explotación de la citada información.

Con el fin de poder desarrollar un proyecto lo más específico posible, el estudiante tendrá que afrontar el reto de desarrollar un almacén de datos que sólo describe parte de los servicios que se pueden ofrecer, en base a los datos tratados en el caso y que formarían parte de un sistema real.

A partir de unas necesidades de negocio acotadas, el estudiante deberá adquirir un conocimiento básico del entorno tecnológico, de los procesos de negocio, de las necesidades existentes y definir una propuesta adecuada que responda a ellas.

Mediante el desarrollo del caso, el estudiante se va a encontrar con los problemas, dudas y dificultades que se plantean en un proyecto de estas características.

Contexto

En el marco de un nuevo modelo de desarrollo socioeconómico basado en la inversión en capital humano y bienes intangibles, el conocimiento, es decir, la educación, la investigación y la innovación, se han convertido en un importante motor de crecimiento y prosperidad, dada su capacidad para crear valor económico.

Una de las principales consecuencias del proceso de adaptación de la economía a las exigencias de una sociedad global y del conocimiento es que ha puesto el foco de atención en la educación. En este sentido, la Comisión Europea advierte de que, en un futuro muy próximo, el 90% de las ofertas de trabajo en suelo europeo requerirán personal cualificado o muy cualificado.

Dentro de la Estrategia Europa 2020, se proponen tres prioridades que se refuerzan mutuamente:

- Crecimiento inteligente: desarrollo de una economía basada en el conocimiento y la innovación como impulsores del crecimiento futuro.
- Crecimiento sostenible: promoción de una economía que haga un uso más eficaz de los recursos, que sea más verde y competitiva.
- Crecimiento integrador: fomento de una economía con alto nivel de empleo que tenga cohesión social y territorial.

Para la consecución de estos objetivos, se establecen dos indicadores importantes:¹

- El 75 % de la población de entre 20 y 64 años debería estar empleada.
- Al menos el 40 % de la generación más joven debería tener estudios superiores completos.

En este contexto la universidad influye en todos los ámbitos de la sociedad y está llamada a jugar un papel fundamental en cuanto que es proveedora de capital humano cualificado, generadora de nuevo conocimiento y transmisora de dicho conocimiento al ámbito productivo y a la sociedad en su conjunto.

La Comisión Europea, para el cumplimiento de los objetivos marcados en la Estrategia 2020, precisa de la construcción de un sistema integrado con información de la educación terciaria y en concreto de información sobre los estudiantes egresados² que permita la recogida, procesamiento y análisis de la misma, como apoyo a la toma de decisiones de los usuarios potenciales y posterior difusión de datos a la sociedad, atendiendo a los principios del movimiento de datos abiertos (**open data**). De esta manera, el sistema desarrollado cumplirá con el principio de transparencia en el tratamiento de información pública y cuyo uso contribuirá a la innovación y servicio a la ciudadanía aportándoles valor en forma de conocimiento útil.

El desarrollo del sistema integrado de personas egresadas requiere de:

- Diseño y Construcción del *data warehouse* que permita la integración de datos de diferentes fuentes con datos de graduados universitarios y el mercado de trabajo para su posterior análisis.
- Diseño e Implementación de los procesos de carga iniciales e incrementales al *data warehouse*.
- Implantación de un sistema de Inteligencia de Negocio de apoyo a la toma de decisiones a los usuarios potenciales, que permita analizar los graduados

universitarios.

Sin lugar a dudas, estos datos de egresados universitarios serán una herramienta imprescindible para la sociedad en general y en particular servirá a los estudiantes para ayudarles a decidir los estudios que van a cursar.

Usuarios potenciales

Como fase inicial del diseño del sistema integrado de egresados universitarios realizaremos el análisis de requerimientos teniendo en cuenta quienes serán los usuarios potenciales. Tendremos que tener en cuenta que el sistema responde a sus necesidades y genera información útil.

Los usuarios finales que harán uso del sistema son:

- **El Espacio Europeo de Educación Superior (EEES)**, organización que establece el sistema de ordenación de las enseñanzas universitarias oficiales de la Unión Europea e introduce procesos para asegurar la calidad de las titulaciones universitarias. Utilizará el sistema integrado de egresados universitarios para evaluación del cumplimiento de las medidas marcadas por la normativa europea en materia de educación terciaria. Además, le permitirá proporcionar información relacionada con el número de egresados por diferentes características a los estudiantes europeos que lo soliciten.
- **La Agencia de Evaluación de la Calidad y Acreditación (ANECA)**, como organismo cuyo objetivo es contribuir a la mejora de la calidad del sistema de educación superior, hará uso del sistema integrado de egresados universitarios para recabar información que le permita desarrollar actividades de evaluación, certificación y acreditación.
- **Las Comunidades Autónomas**. El sistema integrado de información de egresados permitirá a los órganos de evaluación propios extraer información útil en relación a las características de los egresados de las universidades de su ámbito territorial, además de contribuir a la mejora de la calidad dado que tendrá un conocimiento en la misma materia de universidades de otras comunidades y así poder realizar comparativas.
- **Estudiantes**. Con la información proporcionada por el sistema integrado de egresados universitarios, los estudiantes y sus familiares dispondrán de conocimiento para el apoyo en la elección del tipo de universidad en relación con el porcentaje de egresados que se han incorporado al mundo laboral en función del área de conocimiento, la rama de enseñanza y ámbito de estudio que estudiaron.
- **Instituciones Universitarias**. Las universidades tendrán una herramienta que les permita mejorar sus programas y titulaciones en función del número de egresados insertados.

¹ Objetivos Europa 2020: estadísticas e indicadores para España

- **Empresas Infomedias**, cuya actividad empresarial está basada en desarrollar aplicaciones, productos y servicios que reutilizan la información pública y privada. Según un estudio del Observatorio Nacional de las Telecomunicaciones y Sociedad de la Información (ONTSI)³, las empresas reutilizadoras, en 2015 generaron un volumen de negocio de entre 600 y 750 millones de euros en España. El estudio apunta al desarrollo de servicios para las ciudades inteligentes y el *social data*, los proyectos relacionados con *Big Data* y la disponibilidad de datos en tiempo real como las grandes oportunidades para el crecimiento y consolidación de su sector.

Fuentes de datos

Uno de los objetivos de este caso de estudio es integrar las fuentes de datos proporcionadas para poder realizar diferentes tipos de análisis. En concreto, disponemos de información de tres fuentes de datos: MECD, Eurostat y OpenData.

Utilizaremos datos extraídos de la web del Ministerio de Educación, Cultura y Deporte (MECD) que aporta información estadística muy completa del sistema universitario. Otra de las fuentes, son datos de la web de Eurostat con información que permite realizar una comparativa entre egresados universitarios españoles y de otros países. Y por último, utilizaremos datos abiertos (open data), un conjunto de datos sobre la situación laboral de los egresados universitarios.

Los ficheros con los datos están agregados a nivel de año. La relación de ficheros que utilizaremos para la carga inicial del *data warehouse* son:

²Se considera estudiante egresado a aquel que ha completado con éxito todos los créditos del plan de estudios en el que está matriculado.

Nombre fichero	Descripción	Fuente
SEGR1.csv	Series de estudiantes egresados en universidades privadas por curso académico, modalidad de impartición, rama de enseñanza y universidad. Cursos 2009-2017.	MECD
SEGR2.csv	Series de egresados de universidades públicas por curso académico, modalidad de impartición, rama de enseñanza y universidad. Cursos 2009-2017.	MECD
ISCED_2013.csv	Clasificación normalizada internacional de educación (ISCED-F 2013) y el campo de estudio	MECD
grad_5sc.csv	Perfil de Egresados de grado y master por Ámbito de Enseñanza, Sexo y Grupos de Edad. Curso 2016-2017	MECD
03003.xls	Encuesta Inserción Laboral. Datos de egresados según situación profesional en 2014.	datos.gob.es (OpenData)
educ_uee_grad01.xls	Comparación internacional del número de egresados universitarios entre España y otros Países.	Eurostat. Estadísticas Europeas
edat_lfse_03.xls	Comparación internacional del porcentaje de egresados universitarios jóvenes (Entre 20 y 29 años) con estudios superiores completos.	Eurostat. Estadísticas Europeas

² La reutilización de información ocupa a 5.200 profesionales de 535 empresas en España

Se tendrá en cuenta que con frecuencia anual recibiremos los datos de las personas egresadas correspondientes a posteriores cursos académicos y por tanto, se realizarán cargas incrementales para la integración de esos datos en el *data warehouse* para su análisis en nuestro sistema.

Enunciado

1) Diseño del *Data Warehouse*

A partir del análisis del contexto del caso y de las fuentes de datos disponibles, el estudiante deberá diseñar y proponer un almacén de datos que ofrezca soporte al funcionamiento del sistema integrado de estudiantes egresados universitarios.

Mediante la metodología de diseño de un DW, el estudiante debe llevar a cabo:

- **El análisis de requerimientos:** como resultado se generará un documento que describa las preguntas a las que el sistema dará respuesta para los usuarios potenciales del mismo.
- **El análisis de fuentes de datos:** se deben revisar las fuentes de datos proporcionadas, qué tipo de información contienen, cuál es su formato, y qué cantidad representan para la carga inicial.
- **El análisis funcional:** se debe de proponer el tipo de arquitectura para la factoría de información que mejor se aadecue al proyecto (por ejemplo, si es necesario un *data mart* operacional o una estructura de carga intermedia).
- **Diseño del modelo conceptual, lógico y físico del almacén de datos:** se deben identificar, diseñar e implementar las tablas de hecho, las dimensiones y atributos que describen la información.

Para este apartado, el estudiante debe preparar un documento en que se expliquen las secciones anteriores.

Se deberá de tener en cuenta que para el desarrollo del DW, es preciso definir correctamente los hechos (*facts*), dimensiones de análisis (*dimensions*) y los atributos que nos permitan tener el nivel de granularidad suficiente para la medida y presentación de los objetivos que se definan en el análisis de requerimientos.

2) Carga de datos

En la segunda parte del caso práctico se realizará el desarrollo de los procesos de carga del DW. A partir del análisis de las fuentes de datos, el estudiante debe diseñar los procesos que permitan la carga de los datos desde los ficheros al sistema utilizando las herramientas indicadas.

Se deberá tener en cuenta que ésta es una carga inicial de almacén de datos, por lo que se espera que, teniendo en cuenta el período que comprenden los datos y la cantidad de los mismos, el estudiante haga estimaciones sobre las necesidades de arquitectura del almacén de datos (por ejemplo, tiempo de carga, estimaciones de crecimiento, ...).

El estudiante por lo tanto debe:

- Revisión del diseño conceptual, lógico y físico del modelo multidimensional realizado en el punto anterior (*1-Diseño de Data warehouse*).
- Identificar y diseñar los procesos necesarios para la extracción, transformación y carga de datos
- Implementar los procesos mediante las herramientas de diseño proporcionadas.
- Realizar la carga de datos de forma efectiva.

3) Explotación de datos

Por último, el estudiante debe diseñar un modelo MOLAP (*Multidimensional On Line Analytical Processing*) para el análisis multidimensional de la información disponible en el DW que permita:

- Responder a las preguntas planteadas en la toma de requerimientos.
- Realizar un informe que permita analizar la tendencia del número de egresados universitarios (grados y masters) desde 2009 hasta 2017 por:
 - a. Tipo de Universidad
 - b. Tipo de Universidad (Privadas o Públicas) y Universidad.
 - c. Tipo de Universidad y Modalidad (Presencial, Especial, No Presencial).
 - d. Universidad y Ramas de Enseñanza.
 - e. Tipo de Universidad y Ramas de Enseñanza y Modalidad
- Realizar un informe que permita caracterizar a las personas egresadas del curso académico 2016-2017 por:
 - a. Sexo.
 - b. Intervalos de Edad
- Analizar la incorporación de los graduados universitarios del curso 2009-2010 al mercado laboral en 2014.
- Comparativa internacional del conjunto de egresados.

Se deja en manos del estudiante enriquecer el sistema con otras herramientas/visualizaciones que estime oportunas.

Diseño y Construcción del *Data Warehouse*

Autora: Nerea Sevilla Marchena

A partir del análisis del contexto del caso y de las fuentes de datos disponibles, el estudiante deberá diseñar y proponer un almacén de datos que ofrezca soporte al funcionamiento del sistema integrado de estudiantes egresados universitarios.

1. Análisis de requerimientos

El análisis de requerimientos se basa en identificar las necesidades específicas que tiene una particular organización respecto al análisis de la información. Normalmente en esta fase, se debe ser previsor y pensar más allá de las necesidades actuales para poder cubrir las futuras.

La necesidad principal de la Comisión Europea es disponer de información integrada sobre los egresados para su análisis y difusión mediante herramientas de inteligencia de negocio que le permita por un lado, la mejora en la toma de decisiones tanto a instituciones universitarias como a la sociedad y por otro, el cumplimiento de principios de transparencia y eficiencia.

En este caso, el objetivo es diseñar un almacén de datos que lo permita, a través de un proyecto, que incluye la creación e implementación de un modelo relacional, el diseño e implementación de procesos ETL, el diseño e implementación del modelo OLAP y, por último, el diseño de las consultas establecidas en el enunciado.

A continuación, identificamos las siguientes necesidades de información:

- Conocer la evolución temporal del número de egresados en el sistema educativo universitario.
- Esta evolución debe poderse analizar desde diferentes perspectivas:
 - a. Tipo de Universidad (P.Ej: Universidades Privadas).
 - b. Modalidad.(P.Ej: No Presencial)
 - c. Universidad (P.Ej: Oberta de Catalunya).
 - d. Rama de enseñanza. (P.Ej: Ciencias Sociales y Jurídicas)
 - e. Ámbito de Estudio. (P.Ej: Ciencias de la educación)
- Conocer el perfil de los estudiantes egresados en el curso académico 2016-2017, en términos de características personales como sexo y edad.

- Analizar la incorporación de los graduados universitarios del curso 2009-2010 al mercado laboral en 2014.
- Realizar la comparativa entre egresados universitarios en España y otros países.

Si se tiene en cuenta toda esta información, el sistema podrá responder a múltiples preguntas y de esta manera, cubrir las necesidades de los usuarios potenciales.

De forma específica, se pide que el sistema debe como mínimo ser capaz de dar respuesta a las siguientes preguntas:

- Top 10 de universidades con mayor número de egresados.
- Ranking de ramas de conocimiento con mayor número de estudiantes egresados insertados.
- Evolución en los últimos 5 años del número de egresados universitarios por tipo de universidad.
- Evolución del número de egresados universitarios por modalidad de impartición y rama de conocimiento.
- Ranking de ámbitos de estudios con mayor número de egresados menores de 25 años en el curso académico 2016-2017.
- Ranking de edad con mayor número de personas egresadas en el curso académico 2016-2017.
- Ámbito de estudios con mayor número de mujeres egresadas.
- Ámbito de estudios con mayor número de hombres egresados.
- Tipo de universidad y rama de conocimiento con menor número de estudiantes egresados insertados.
- Ranking de países con mayor porcentaje de estudiantes jóvenes con estudios superiores completos.

2. Análisis de fuentes de datos

En este apartado se deben revisar las fuentes de datos proporcionadas, qué tipo de información contienen, cuál es su formato, y qué datos deben ser cargados.

Veamos a continuación un análisis detallado por cada tipo de formato.

Ficheros Planos

Los ficheros planos desde el origen vienen con las siguientes características:

- Formato: CSV
- Primera línea con etiquetas de los campos.

- Separador de campos: Punto y coma (;
- Campos de Texto: Entre comillas ("")

SEGR1.csv: Series de número de egresados Universidades Privadas, modalidad, universidad y rama de enseñanza. Cursos 2009-2017.

Nombre Campo	Tipo	Ejemplo
TIPO_UNIVERSIDAD	Texto	"Universidades Privadas"
MODALIDAD	Texto	"No Presencial"
UNIVERSIDAD	Texto	"Oberta de Catalunya"
RAMA_ENSEÑANZA	Texto	"Ciencias Sociales y Jurídicas"
EGR_C16_17	Numérico	3364
EGR_C15_16	Numérico	3129
EGR_C14_15	Numérico	3355
EGR_C13_14	Numérico	2718
EGR_C12_13	Numérico	2975
EGR_C11_12	Numérico	3903
EGR_C10_11	Numérico	3073
EGR_C09_10	Numérico	2513

- Total de registros: 160

SEGR2.csv: Series de número de egresados de universidades públicas, modalidad, universidad y rama de enseñanza. Cursos 2009-2017.

Nombre Campo	Tipo	Ejemplo
TIPO_UNIVERSIDAD	Texto	"Universidades Pùblicas"
MODALIDAD	Texto	"No Presencial"
UNIVERSIDAD	Texto	"Nacional de Educación a Distancia"
RAMA_ENSEÑANZA	Texto	"Ciencias Sociales y Jurídicas"
EGR_C16_17	Numérico	4482
EGR_C15_16	Numérico	5640
EGR_C14_15	Numérico	5245
EGR_C13_14	Numérico	5442
EGR_C12_13	Numérico	5491
EGR_C11_12	Numérico	4663
EGR_C10_11	Numérico	4967
EGR_C09_10	Numérico	3695

Total de registros: 250

ISCED_2013.csv: Clasificación normalizada internacional de educación (ISCED-F 2013). Clasificación del campo de estudio a 5 Niveles.

Nombre Campo	Tipo	Ejemplo
COD_RAMA	Texto	“2”
NOM_RAMA	Texto	“Ciencias Sociales y Jurídicas”
COD_RAMA_N2	Texto	“03”
NOM_RAMA_N2	Texto	“Ciencias sociales, periodismo y documentación”
COD_RAMA_N3	Texto	“031”
NOM_RAMA_N3	Texto	“Ciencias sociales y del comportamiento”
COD_RAMA_N4	Texto	“0311”
NOM_RAMA_N4	Texto	“Economía”
COD_RAMA_N5	Texto	“031101”
NOM_RAMA_N5	Texto	“Economía”

Total de registros: 162

grad_5sc.csv: Perfil de Egresados de grado y master por Ámbito de Enseñanza (4º Nivel de la Clasificación ISCED), Sexo y Grupos de Edad. Curso 2016-2017.

Nombre Campo	Tipo	Ejemplo
COD_AMBITO	Texto	“0311 - Economía”
SEXO	Texto	“Mujeres”
EDAD	Texto	“De 25 a 30 años”
NUM_EGR_NV1	Numérico	249
NUM_EGR_NV2	Numérico	214

Total de registros: 703

Ficheros Excel

03003.xls: Titulados universitarios según su situación laboral en 2014 por sexo, tipo de universidad y rama de conocimiento del curso académico 2009-2010.

Nombre Campo	Tipo	Ejemplo
TIPO_UNIVERSIDAD	Texto	“Universidades Públicas”
SEXO	Texto	“Mujeres”
RAMA_ENSEÑANZA	Texto	“Ciencias de la salud”
TRABAJANDO	Numérico	14648
EN DESEMPLEO	Numérico	2490

INACTIVO	Numérico	1128
----------	----------	------

Total de registros: 20

De este fichero Excel no se cargarán datos sobre totales por tipo universidad, por sexo y rama de conocimiento porque son campos redundantes, y se pueden calcular.

educ_uee_grad01.xls: Comparación del número de egresados universitarios entre España y otros Países de los cursos 2013-2017.

Nombre Campo	Tipo	Ejemplo
GEO/TIME	Texto	“Finland”
2013	Numérico	52730
2014	Numérico	53878
2015	Numérico	56829
2016	Numérico	56066
2017	Numérico	56136

Total de registros: 12

Este fichero Excel, contiene 3 hojas (Data, Data2 y Data3), el total de personas egresadas, mujeres egresadas y hombres egresados respectivamente. Sólo se cargará la información de la primera hoja con el total de egresados por curso académico ya que es suficiente para el análisis comparativo entre países que se solicita en los requerimientos.

edat_ifse_03.xls: Porcentaje de egresados universitarios jóvenes (Entre 20 y 29 años).con estudios superiores completos

Nombre Campo	Tipo	Ejemplo
GEO/TIME	Texto	“Finland”
2013	Numérico	50,6
2014	Numérico	50,8
2015	Numérico	54,4
2016	Numérico	53,9
2017	Numérico	53,6

Total de registros: 12

La unidad es número de personas egresadas por cada mil habitantes.

Este fichero Excel, contiene 2 hojas (Data y Data2), con el total de personas egresadas por cada mil habitantes y jóvenes egresados de entre 20 y 29 años por cada mil habitantes, dado que en los requerimientos sólo se hace referencia al análisis sobre jóvenes egresados, y para no complicar la práctica, sólo se tendrá en cuenta la información de la segunda hoja (Data2).

En los proyectos de diseño de factoría de información corporativa existe una primera fase en la que se realiza una carga inicial, y a posteriori, una segunda fase para realizar las cargas incrementales de los datos nuevos que nos van llegado.

Una estimación del volumen de datos de nuestro almacén para la carga de los datos que disponemos sería:

Fuente de datos	Valores a almacenar	Total registros
SEGR1 <input type="checkbox"/> fichero anual <input type="checkbox"/> Tipo Universidad <input type="checkbox"/> Modalidades <input type="checkbox"/> 32 Universidades <input type="checkbox"/> 5 Ramas	Cursos académicos: 8 datos	1 fichero x 1 Tipo Univ x 3 Modalidades x 32 Universidades x 5 Ramas x 8 datos = 3840
SEGR2 <input type="checkbox"/> fichero anual <input type="checkbox"/> Tipo Universidad <input type="checkbox"/> Modalidades <input type="checkbox"/> 50 Universidades <input type="checkbox"/> 5 Ramas	Cursos académicos: 8 datos	1 fichero x 1 Tipo Univ x 3 Modalidades x 50 Universidades x 5 Ramas x 8 datos = 6000
grad_5sc <input type="checkbox"/> fichero anual <input type="checkbox"/> 88 Ámbitos Estudio <input type="checkbox"/> 2 Sexo <input type="checkbox"/> 4 Rangos Edad	Niveles Académicos : 2 datos	1 fichero x 88 Ámbitos x 2 Sexo x 4 Edades x 2 datos = 1408
03003 <input type="checkbox"/> 1 fichero anual <input type="checkbox"/> 2 Tipo Universidad <input type="checkbox"/> 2 Sexo <input type="checkbox"/> 5 Ramas	Situaciones Laborales : 3 datos	1 fichero x 2 Tipos x 2 Sexo x 5 Ramas x 3 datos = 60
educ_uee_grad01 <input type="checkbox"/> ficheros anuales <input type="checkbox"/> 12 Países	Cursos académicos: 5 datos	1 ficheros x 12 Países x 5 datos = 60
edat_lfse_03 <input type="checkbox"/> fichero anual <input type="checkbox"/> 12 Países	Cursos académicos: 5 datos	1 fichero x 12 Países x 5 datos = 60
ISCED_2013 <input type="checkbox"/> fichero anual	Titulaciones <input type="checkbox"/> 162 datos	1 fichero x 162 datos= 162
	TOTAL	=11.590

3. Análisis Funcional

A continuación, se propone el tipo de arquitectura para la factoría de información que mejor se adapta al proyecto.

Para ello vamos a considerar los requisitos funcionales y establecer para ellos una prioridad asociada que podrá ser exigible (E) o deseable (D). En el contexto de esta actividad, los requerimientos exigibles son aquellos que demanda el enunciado y los deseables son aquellos que complementan la actividad.

Por otro lado, en términos de la escala de prioridades asignamos una prioridad de 1 a 4 siendo 1 completamente prioritario para la actividad y 4 no prioritario para la actividad.

A continuación, se describen los requerimientos funcionales para el diseño de una factoría de información para nuestra organización, bajo las consideraciones del enunciado:

#	Requerimiento	Prioridad	Exigible / Deseable
1	Se extraerá de forma adecuada la información de las fuentes de datos (considerando sólo la información relevante).	1	E
2	Se creará un almacén de datos.	1	E
3	Se cargará la información de las personas egresadas en el almacén de datos.	1	E
4	Se creará un modelo OLAP para consultas multidimensionales de los usuarios.	2	E
5	Se crearán los informes estáticos solicitados.	2	E
6	Se redactará un manual de carga de datos incremental	3	D

Cabe comentar que en un caso genérico real podemos encontrar también otros requerimientos funcionales:

- Creación de procesos de calidad de datos.
- Creación de *data marts* (si se analizan otras áreas).
- Automatizar cada proceso de carga de *data marts* (según sus necesidades).
- Creación de procesos de cargas totales e incrementales.
- Creación de un repositorio de metadatos de gestión del almacén de datos, así como de los procesos ETL.

Así mismo, dado que estos sistemas frecuentemente forman parte de la implementación de un sistema de inteligencia de negocio, la lista de requerimientos funcionales sería mucho mayor.

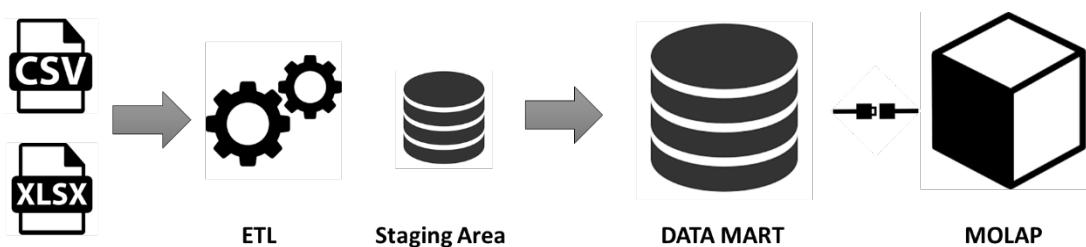
En términos de la arquitectura funcional tenemos los siguientes elementos:

1. Las fuentes de datos están compuestas por los ficheros planos csv y hojas de cálculo xls obtenidos de las webs del Ministerio de Educación (MECD), Eurostat y *OpenData*. Estos ficheros contienen información sobre la evolución de las personas egresadas desde 2012 a 2017, características personales como el sexo y edad de egresados universitarios en el curso 2016-2017, inserción al mercado laboral en 2014 de los egresados del curso 2009-2010 y datos comparativos entre España y otros países de los egresados entre 2012 y 2017.
2. La arquitectura de la factoría de información de personas egresadas puede estar formada por varios elementos alojados en la misma máquina:
 - *Staging Area* (opcional): En el caso de tener múltiples fuentes (ficheros, bases de datos, servicios RSS....) es conveniente cargarlas para consolidar la información en una estructura de carga intermedia que puede ser creada en la misma base de datos.

Esta área del DW también puede servir para entender, simplificar y consolidar el proceso ETL.
 - *Data Mart* de personas egresadas: al centrarnos en una única área temática como es el análisis de egresados universitarios, es más correcto considerar que se está creando un *data mart* en lugar de un almacén de datos corporativo.
 - *MOLAP*: a partir de la información del *datamart* se creará un cubo multidimensional.

En nuestro caso, en la arquitectura funcional se ha optado por usar un área intermedia (*Staging Area*), incluida dentro de la misma base de datos, cuyos objetos se identificarán con un prefijo en los nombres.

El siguiente gráfico resume los elementos de la arquitectura para esta actividad:



También sería correcta utilizar una arquitectura sin *Staging Area*.



4. Diseño del modelo conceptual, lógico y físico del almacén de datos.

Diseño Conceptual

Para el correcto desarrollo del DW es preciso definir los hechos (*facts*), dimensiones de análisis (*dimensions*), las métricas y los atributos que nos permitan tener el nivel de granularidad suficiente para presentación de los objetivos que se han definido en el análisis de requerimientos y de las fuentes de datos.

Del análisis de las fuentes de datos se determinan que disponemos de datos anuales, correspondientes a datos acumulados de cada curso académico y que los hechos son:

- Personas egresadas
- Personas egresadas insertadas, que la situación laboral es trabajando.

Teniendo en cuenta los requerimientos solicitados, los hechos identificados se analizarán para resolver cuatro necesidades principales de los usuarios:

1. Análisis temporal de las personas egresadas.
2. Caracterización de las personas egresadas.
3. Análisis de las personas egresadas insertadas (cuya situación laboral es trabajando).
4. Comparativa de las personas egresadas entre España y otros países.

El análisis temporal de las personas egresadas, determina el diseño de la primera tabla de hechos:

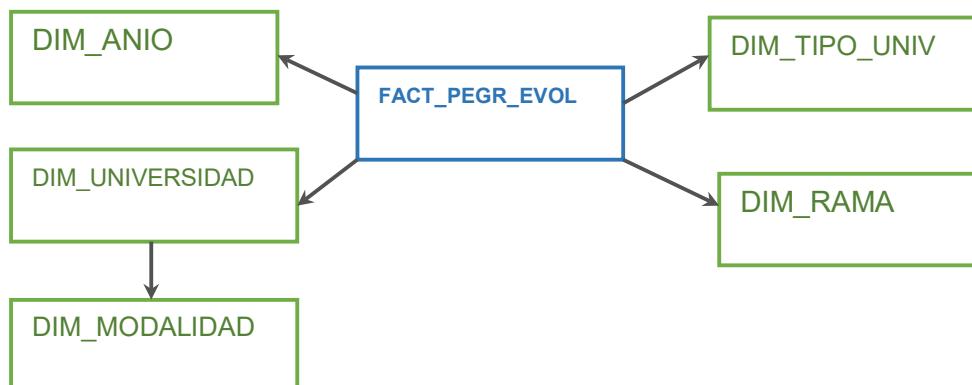
Tabla de hechos	Descripción
Fact_PersonasEgresadas_Evolutivo	Recoge el número de egresados por curso académico.

Esta tabla de hechos, además de la métrica del número personas egresadas, se analizará desde diferentes perspectivas:

Dimensiones	Descripción
Año	Número de personas egresadas por Curso Académico
TipoUniversidad	Número de egresados por Tipo de Universidad
Universidad	Número de personas egresadas por Universidad
Rama Enseñanza	Número de personas egresadas por Rama Enseñanza (Primer Nivel de Clasificación ISCED).

A partir de las dimensiones y la tabla de hechos identificados, se construye el modelo conceptual, siendo tanto las dimensiones como los hechos, entidades independientes que forman parte de nuestro modelo de estrella/copo de nieve.

El diseño conceptual para esta tabla de hechos y sus dimensiones es:



Para la caracterización de las personas egresadas, identificamos una segunda la tabla de hechos:

Tabla de hechos	Descripción
Fact_PersonasEgresadas_Perfil	Recoge el número de egresados del curso 2016-2017.

Esta tabla de hechos, almacenará la métrica del número de personas egresadas que será un campo calculado de la suma de las personas egresadas de grado (NUM_EGR_NV1) más las personas egresadas de master (NUM_EGR_NV2) y será analizada desde diferentes perspectivas:

Dimensiones	Descripción
Año	Número de personas egresadas por Curso Académico
Sexo	Número de personas egresadas por Sexo
Edad	Número de personas egresadas por intervalo de Edad

Rama Enseñanza	Número de personas egresadas por Rama Enseñanza, utilizando el nivel, Ámbito de Enseñanza, que corresponde al 4º Nivel de jerarquía de la dimensión Rama Enseñanza.
----------------	---

En nuestro caso tendremos un modelo en estrella conceptual para cada tabla de hechos con dimensiones comunes como la dimensión tiempo y rama de enseñanza. El diseño conceptual para esta tabla de hechos y sus dimensiones es:



La tabla de hechos que permitirá obtener información sobre la situación laboral en 2014 de las personas egresadas en el curso académico 2009-2010 es:

Tabla de hechos	Descripción
Fact_PersonasEgresadas_Insertadas	Recoge el número de egresados insertados en 2014.

Esta tabla de hechos contendrá 3 métricas, número de estudiantes egresados trabajando, desempleados e inactivos) y será analizada desde diferentes perspectivas:

Dimensiones	Descripción
Año	Número de personas egresadas insertadas por Curso Académico
TipoUniversidad	Número de egresados insertados por Tipo de Universidad
Sexo	Número de personas egresadas insertados por Sexo
Rama de Enseñanza	Número de personas egresadas insertados por Rama de Enseñanza (Primer Nivel de Clasificación ISCED)

El diseño conceptual para esta tabla de hechos y sus dimensiones es:



Por último, la tabla de hechos que recoge la información de Eurostat con datos de personas egresados de otros países en cuatro años académicos y que permitirá realizar un análisis comparativo es:

Tabla de hechos	Descripción
Fact_PersonasEgresadas_Comparativa	Recoge el número de egresados por países.

Contendrá dos métricas, el número de egresados y porcentaje de egresados jóvenes con estudios superiores o terciarios. Esta tabla será analizada desde diferentes perspectivas:

Dimensiones	Descripción
Año	Número de personas egresadas por Curso Académico
País	Número de personas egresadas por País

El diseño conceptual para esta tabla de hechos y sus dimensiones es:



Diseño Lógico

Una vez obtenido el modelo conceptual del almacén de datos de personas egresadas, compuesto de cuatro tablas de hechos y ocho dimensiones; vamos a detallar las métricas de cada una de las tablas de hechos y sus atributos para diseñar el modelo lógico.

A continuación, se muestra una tabla con las métricas que contiene cada tabla de hechos que compone el modelo lógico del almacén de personas egresadas.

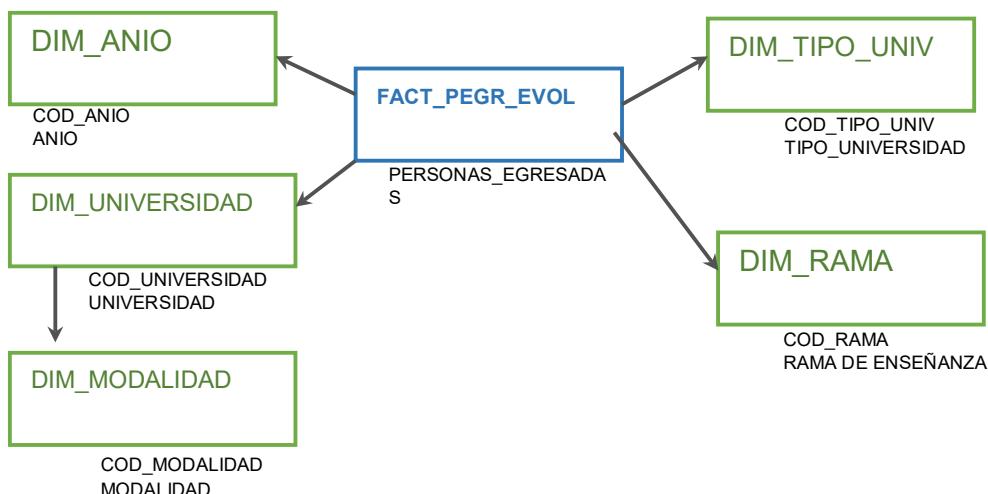
Tabla de Hechos	Métricas
FACT_PEGR_EVOLUTIVO	PERSONAS_EGRESADAS
FACT_PEGR_PERFIL	PERSONAS_EGRESADAS
FACT_PEGR_INSERTADAS	PEGRES_TRABAJANDO PEGRES_DESEMPLEADAS PEGRES_INACTIVAS
FACT_PEGR_COMPARATIVA	PERSONAS_EGRESADAS PORCENTAJE_EGR_JÓVENES

Lo siguiente que veremos son los atributos que contiene cada tabla de hechos. Los atributos junto con las métricas, nos permitirán realizar los diferentes análisis de los requerimientos.

En la siguiente tabla, se muestran los atributos descriptores con las referencias a sus dimensiones de la tabla de hechos FACT_PEGR_EVOLUTIVO:

Dimensiones	Atributos descriptores
DIM_ANIO	COD_ANIO
DIM_TIPO_UNIV	COD_TIPO_UNIV
DIM_UNIVERSIDAD	COD_UNIVERSIDAD
DIM_RAMA	COD_RAMA

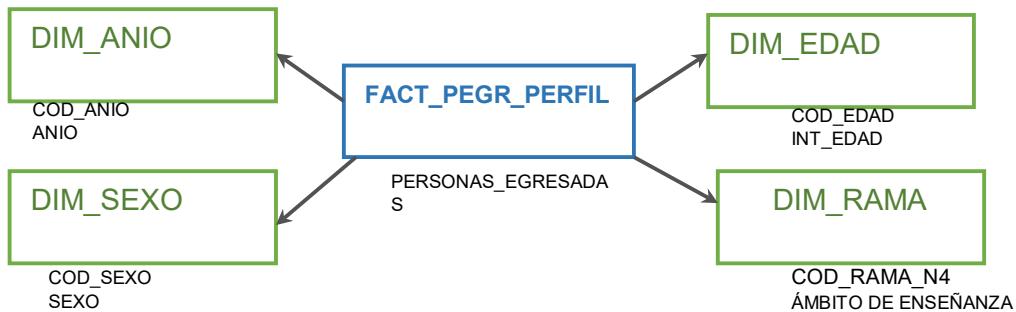
En la siguiente imagen se muestra el diseño del modelo lógico propuesto para la tabla de hechos FACT_PEGR_EVOLUTIVO.



Los atributos descriptores de la tabla de hechos FACT_PEGR_PERFIL:

Dimensiones	Atributos descriptores
DIM_ANIO	COD_ANIO
DIM_SEXO	COD_SEXO
DIM_EDAD	COD_EDAD
DIM_RAMA	COD_RAMA_N4

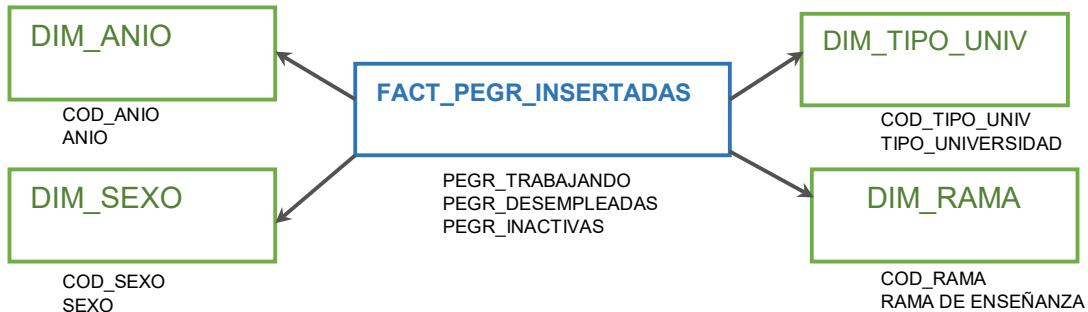
El diseño del modelo lógico propuesto para la tabla de hechos FACT_PEGR_PERFIL sería:



Los atributos descriptores a las dimensiones en la tabla de hechos FACT_PEGR_INSERTADAS:

Dimensiones	Atributos descriptores
DIM_ANIO	COD_ANIO
DIM_TIPO_UNIV	COD_TIPO_UNIV
DIM_SEXO	COD_SEXO
DIM_RAMA	COD_RAMA

En la imagen se muestra su correspondiente modelo lógico:



Por último, en la siguiente tabla vemos los atributos de la tabla de hechos FACT_PEGR_COMPARATIVA.

Dimensiones	Atributos descriptores
DIM_ANIO	COD_ANIO
DIM_PAIS	COD_PAIS

Y su modelo lógico sería:



Diseño Físico

Para el correcto diseño físico del almacén debemos tener en cuenta diversos aspectos:

- El tipo de base de datos con el que trabajemos, puesto que cada una de ellas tiene su particularidad.
- El diseño físico debe estar orientado a generar un buen rendimiento en el procesamiento de consultas.
- La definición de los procesos de administración del DW.
- La revisión periódica del diseño físico inicial para validar que continúa dando respuesta a las necesidades del cliente.

Una vez determinados qué tablas de hechos, dimensiones, métricas y atributos existen en nuestro modelo, podemos determinar también las claves foráneas que debe incluir el modelo físico. En este paso es necesario tener en cuenta el tamaño adecuado de los atributos (por ejemplo, qué longitud tiene una cadena). También es relevante acordarse de crear correctamente las claves primarias, claves foráneas y disparadores (por ejemplo, para actualizar de forma automática las claves primarias).

Dado que nuestro modelo de almacén está compuesto de más de una tabla de hechos, también debemos revisar las dimensiones que hemos definido en el diseño conceptual y lógico de cada *fact* aplicando una visión conjunta del modelo. Esto nos permitirá definir dimensiones comunes, como año, tipo de universidad o rama de enseñanza y así simplificar el modelo final y conseguir un rendimiento óptimo en la ejecución de los análisis.

Como es lógico, primero se crean las tablas de dimensiones y posteriormente las tablas de hechos. De esta forma creamos cada una de las tablas de nuestro almacén de datos.

Dimensiones

Las dimensiones del modelo podrán ser que referenciadas en las tablas de hecho utilizando sus claves primarias o *primary key* (CK/PK). El modelo físico de las dimensiones es:

- DIM_ANIO: Corresponde a la dimensión temporal de nuestro almacén. Disponemos de datos anuales, correspondientes a datos acumulados de cada curso académico. En nuestro caso es muy simple. Normalmente es bastante más complicada, ya que puede incluir días, días de la semana, festivos, semestres, cuatrimestres, etc. La clave primaria será el año de carga de datos y una descripción para describir los cursos académicos a los que pertenecen los datos.

Nombre Campo	Tipo	Tamaño	Ejemplo
SK_DIM_ANIO (CP/PK)	Numérico	4	2011
DESC_ANIO	Texto	50	'2009-2010'

- DIM_TIPO_UNIV: Contiene los valores de los tipos de universidades existentes (1. Universidades Públicas; 2. Universidades Privadas)

Nombre Campo	Tipo	Tamaño	Ejemplo
SK_DIM_TIPO_UNIV (CP/PK)	Numérico	1	2
DESC_TIPO_UNIV	Texto	30	'Universidades Privadas'

- DIM_RAMA: La dimensión rama, contiene información sobre la clasificación normalizada internacional de educación (ISCED-F 2013) a 5 Niveles.

Nombre Campo	Tipo	Tamaño	Ejemplo
SK_DIM_RAMA (CP/PK)	Numérico	5	1
COD_RAMA	Texto	1	1
DESC_RAMA	Texto	50	'Artes y Humanidades'
COD_RAMA_N2	Texto	2	'01'
DESC_RAMA_N2	Texto	100	'Educación'
COD_RAMA_N3	Texto	3	'011'
DESC_RAMA_N3	Texto	100	'Educación'
COD_RAMA_N4	Texto	4	'0111'
DESC_RAMA_N4	Texto	150	'Ciencias de la educación'
COD_RAMA_N5	Texto	6	'011101'
DESC_RAMA_N5	Texto	200	'Pedagogía'

- DIM_MODALIDAD: Contiene las modalidades de impartición de las universidades (1. Presencial; 2. No Presencial; 3.Especiales)

Nombre Campo	Tipo	Tamaño	Ejemplo
SK_DIM_MODALIDAD (CP/PK)	Numérico	1	3
DESC_MODALIDAD	Texto	50	'Especiales'

- DIM_UNIVERSIDAD: Contiene la información de cada una de las universidades sobre la que se recoge información.

Según el análisis de las fuentes de datos, se ha considerado que la modalidad de impartición es una propiedad de cada universidad, por esta razón en nuestro modelo será un atributo de la dimensión de universidad en lugar de ser un atributo de la tabla de hechos.

Si las necesidades de los usuarios requieren realizar análisis sobre la información de la modalidad de impartición, podrá modificarse el modelo y utilizar la dimensión modalidad en las tablas de hechos.

Nombre Campo	Tipo	Tamaño	Ejemplo
SK_DIM_UNIVERSIDAD (CP/PK)	Numérico	3	40
DESC_UNIVERSIDAD	Texto	100	'Internacional de Andalucía'
SK_DIM_MODALIDAD (CA/FK)	Numérico	1	3

El atributo SK_DIM_MODALIDAD, hace referencia al tipo de modalidad de impartición de cada universidad y cuyos valores se encuentran en la tabla creada para la dimensión modalidad (DIM_MODALIDAD). Es por esta razón, que se crea una restricción de Clave Ajena o *Foreign Key* en el atributo SK_DIM_MODALIDAD en la dimensión universidad. Y así garantizar la integridad referencial.

- DIM_SEXO: Contiene los valores correspondiente al sexo de las personas egresadas (1.Mujeres y 2.Hombres)

Nombre Campo	Tipo	Tamaño	Ejemplo
SK_DIM_SEXO (CP/PK)	Numérico	1	1
DESC_SEXO	Texto	10	'Hombres'

- DIM_EDAD: Contiene los valores correspondientes al intervalo de edad de las personas egresadas (1. Menos de 25 años; 2. De 25 a 30 años; 3. De 31 a 40 años y 4. Más de 40 años)

Nombre Campo	Tipo	Tamaño	Ejemplo
SK_DIM_EDAD (CP/PK)	Numérico	1	3
DESC_INT_EDAD	Texto	50	'De 31 a 40 años'

- DIM_PAIS: Contiene la relación de los nueve países con los que se realizarán el análisis comparativo de personas egresadas.

Nombre Campo	Tipo	Tamaño	Ejemplo
SK_DIM_PAIS (CP/PK)	Numérico	3	11
DESC_PAIS_ES	Texto	50	'Suiza'
DESC_PAIS_EN	Texto	200	'Switzerland'

Tablas de Hechos

El modelo físico de las tablas de hechos consistirá en la creación de las tablas cuyos campos serán claves foráneas a las dimensiones del modelo estrella del su diseño lógico y de las métricas.

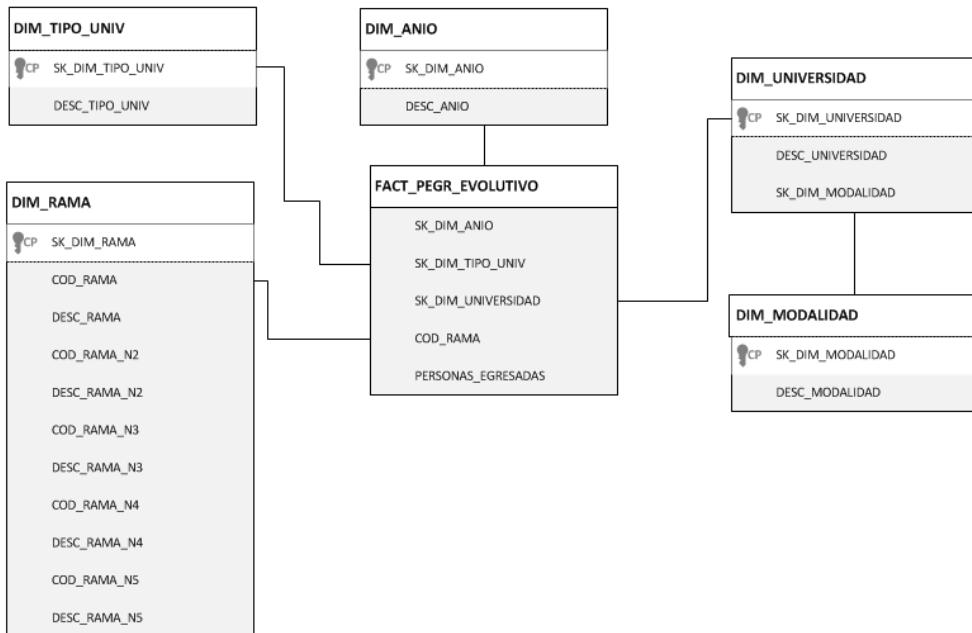
El modelo físico de las tablas de hechos del almacén integrado de personas egresadas está compuesto de las siguientes tablas:

- FACT_PEGR_EVOLUTIVO: Es la tabla física que contendrá la información que permitirá realizar el análisis evolutivo de las personas egresadas, concretamente tendrá los siguientes campos:

Nombre Campo	Tipo	Tamaño	Ejemplo
SK_DIM_ANIO	Numérico	5	2017
SK_DIM_TIPO_UNIV	Numérico	1	1
SK_DIM_UNIVERSIDAD	Numérico	3	9
COD_RAMA	Numérico	2	2
PERSONAS_EGRESADAS	Numérico	8	1351

En la siguiente imagen se muestra el diseño del modelo físico¹ para la tabla de hechos FACT_PEGR_EVOLUTIVO.

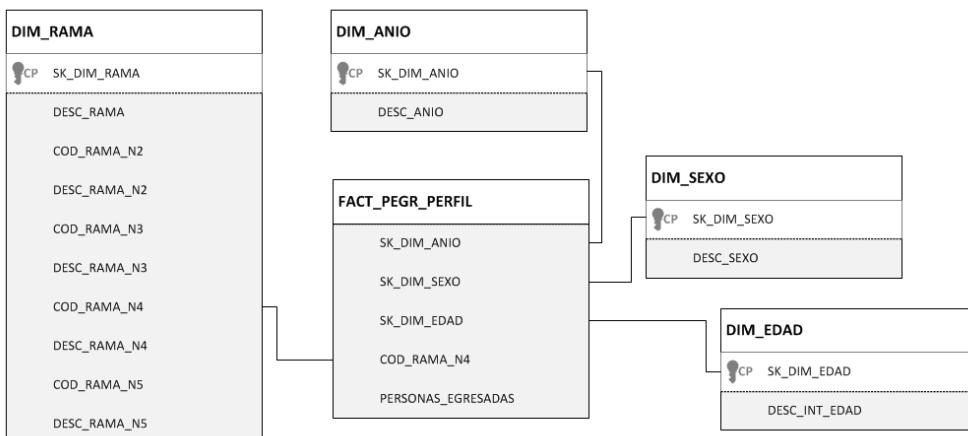
¹ El diseño se ha realizado utilizando la herramienta Microsoft Visio.



- **FACT_PEGR_PERFIL**: Es la tabla física que contendrá la información que permitirá realizar el análisis de la caracterización por sexo y edad de las personas egresadas, concretamente está compuesta de los siguientes campos:

Nombre Campo	Tipo	Tamaño	Ejemplo
SK_DIM_ANIO	Numérico	5	2018
SK_DIM_SEXO	Numérico	1	1
SK_DIM_EDAD	Numérico	1	2
COD_RAMA_N4	Texto	4	'0111'
PERSONAS_EGRESADAS	Numérico	8	320

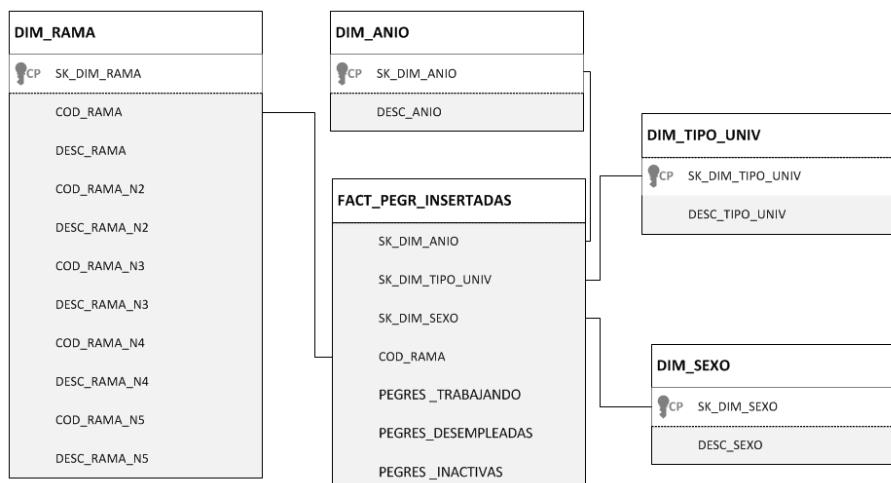
El diseño del modelo físico para la tabla de hechos FACT_PEGR_PERFIL:



- FACT_PEGR_INSERTADAS: Es la tabla física que contiene información sobre la inserción laboral de las personas egresadas en el curso académico 2009-2010, sus campos son:

Nombre Campo	Tipo	Tamaño	Ejemplo
SK_DIM_ANIO	Numérico	5	2011
SK_DIM_TIPO_UNIV	Numérico	1	2
SK_DIM_SEXO	Numérico	1	1
COD_RAMA	Numérico	1	1
PEGR_TRABAJANDO	Numérico	8	2674
PEGR_DESEMPLEADOS	Numérico	8	1122
PEGR_INACTIVOS	Numérico	8	458

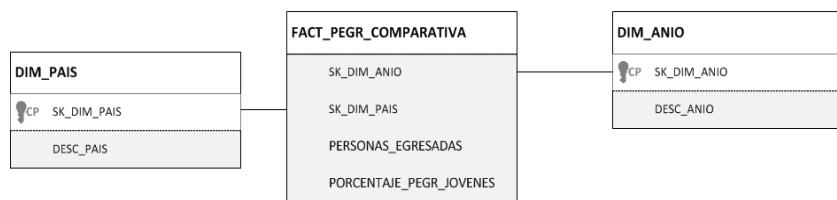
El diseño del modelo físico para la tabla de hechos FACT_PEGR_INSERTADAS:



- FACT_PEGR_COMPARATIVA: Es la tabla que contiene información del número personas egresadas y el porcentaje de jóvenes con estudios superiores completos por países, sus campos son:

Nombre Campo	Tipo	Tamaño	Ejemplo
SK_DIM_ANIO	Numérico	5	2013
SK_DIM_PAIS	Numérico	3	2
PERSONAS_EGRESADAS	Numérico	8	52730
PORCENTAJE_PEGR_JOVENES	Numérico	8	51

El diseño del modelo físico para la tabla de hechos FACT_PEGR_COMPARATIVA se muestra en la siguiente figura:



Carga de datos

Autor: Juan Pablo Botero Suaza

1) Revisar el diseño conceptual, lógico y físico

En este apartado se deberían incorporar aquellas diferencias detectadas al comparar nuestro modelo de *data warehouse* con el que se propone en la solución oficial del caso.

Para esta implementación se ha modificado parte del *diseño físico y lógico* presentado en el anterior apartado:

- Las claves foráneas relacionadas a la dimensión DIM_RAMA desde las 4 tablas de hechos, referenciaran la clave primaria sk_dim_rama y no los atributos de código correspondientes a cada uno de los niveles.
- Se incluye el atributo sk_dim_anio_laboral en la tabla de hecho FACT_PEGR_INSERTADAS, como clave foránea a la dimensión DIM_ANIO, de manera que pueda capturarse el año de análisis de inserción laboral, adicional al año de promoción de los egresados que ya se encontraba en el diseño inicial.
- El formato definido para los códigos de nivel en la DIM_RAMA, si bien continuara de tipo cadena de caracteres, no utilizara el carácter '0' antecedido al código original de la fuente de datos, para efectos de facilitar las operaciones de manipulación y comparación sobre este tipo de datos.

2) Diseño del proceso de ETL

Consideraciones iniciales

El proceso de ETL diseñado para esta implementación tiene como objetivo realizar la **carga inicial** de los datos en el modelo físico definido, las cargas incrementales no hacen parte de este primer alcance, más sin embargo puede ser considerado en una fase posterior como requisito exigible.

El diseño propuesto de carga para cada una de las dimensiones y hechos, está definido bajo el siguiente patrón, haciendo uso de los flujos de ejecución para transformaciones (transformations) y trabajos (jobs) que ofrece la herramienta Pentaho Data Integrator:

- 1 flujo de transformación para la carga de datos en la tabla física a partir de la fuente origen acorde a las definiciones realizadas en la fase de análisis para cada uno de los archivos de datos.
- 1 flujo de trabajo para eliminar los datos de la tabla a cargar, y posterior ejecución de la transformación de carga de datos. De esta forma se garantiza la repetitividad del proceso de forma individual en caso de falla o cambios en las fuentes de datos correspondientes a la carga inicial sobre la tabla física.

No se hará uso de un modelo de staging al no considerarse necesario para el caso de uso de la carga inicial de datos y por los volúmenes de datos a procesar, donde en cierta medida es relativamente fácil determinar la cantidad de datos procesados y cargados. Las reglas de consistencia e integridad de datos harán parte de los flujos de transformación mencionados anteriormente.

En cuanto a la creación de las claves primarias para las dimensiones, se hará uso de secuencias numéricas no autogeneradas en base de datos, algunas de ellas se crearán como parte de los flujos de transformación y otras de ellas mediante ejecución de scripts sql. Para las tablas de hechos, la clave primaria estará definida por la composición de las claves foráneas a las respectivas llaves primarias de las dimensiones correspondientes, por lo tanto la integridad de las mismas se mantendrá a nivel de las restricciones establecidas sobre el modelo físico.

Las dimensiones que presentan atributos estáticos, poco variables o de bajo volumen como DIM_SEXO, DIM_PAIS, DIM_ANIO se cargarán a través de un script sql, las dimensiones DIM_RAMA y DIM_UNIVERSIDAD así como las tablas de hechos se cargarán a través de flujos de PDI.

Se definirá un directorio para el manejo de errores en la carga de datos sobre la tabla destino (F:\Pentaho\PRA2\error) y otro directorio para almacenar los archivos correspondientes a las fuentes de datos originales y algunos archivos intermedios obtenidos a partir de las mismas fuentes (F:\Pentaho\PRA2\fuentes)

Se debe tener en cuenta que uno de los objetivos del proceso de carga es la automatización y por ello solo se deben hacer cambios manuales cuando sea estrictamente necesario, la automatización puede prevenir errores ocasionados por la manipulación manual de los datos.

Caracterización del proceso ETL

El job principal que se encargará del proceso de carga inicial es el siguiente:

CARGA_INICIAL_DWH_EGRESADOS



- START: Inicio del trabajo
- DELETE DIMS&FACTS: limpiar las tablas de dimensiones y luego las de hechos.
- CARGA_DIM_STATIC: carga las dimensiones DIM_ANIO, DIM_EDAD, DIM_TIPO_UNIV, DIM_SEXO, DIM_PAIS, DIM_MODALIDAD a partir de script sql.
- CARGA_DIMENSIONES: carga de dimensiones DIM_RAMA y DIM_UNIVERSIDAD en el DW.
- CARGA_HECHOS: carga de hechos en el DW.
- Success: Finalización del trabajo.

A su vez cada uno de estos trabajos estará compuesto de otros trabajos y transformaciones más granulares que se encargan de actividades específicas de manera que podamos dar cierto grado de modularidad al diseño y sea más fácil realizar cambios o ajustes en etapas posteriores de así requerirse. La nomenclatura utilizada para el nombramiento de los componentes de la ETL está basada en la operación a ejecutar y sobre qué objeto, por ejemplo:

Transformación: *LOAD_DIM_RAMA*, carga los datos a la dimensión *DIM_RAMA*.

Trabajo: *ETL_DIM_RAMA*, primero limpia y después carga los datos a la dimensión *DIM_RAMA*.

En esta propuesta, se realizará control de errores para las trasformaciones de carga de datos a través de archivos de texto planos almacenados en el entorno de trabajo configurado en PDI, de esta manera el usuario encargado de ejecutar el job principal puede revisar en estos archivos el detalle correspondiente a alguna falla generada durante dicha ejecución.

Estructura componentes del proceso ETL

Teniendo en cuenta estos requisitos se han definido los siguientes componentes para implementar el caso de uso propuesto:

- **DELETE DIMS&FACTS:** limpiar las tablas de dimensiones y luego las de hechos a través de un scripting SQL. Asegura la repetitividad del proceso de carga inicial completo.
- **CARGA_DIM_STATIC:** carga las dimensiones DIM_ANIO, DIM_EDAD, DIM_TIPO_UNIV, DIM_SEXO, DIM_PAIS, DIM_MODALIDAD a partir de scripting sql.
- **CARGA_DIMENSIONES:** job principal encargado de ejecutar el proceso etl en paralelo de las dimensiones DIM_RAMA y DIM_UNIVERSIDAD
 - **ETL_DIMENSIONES:** job encargado de ejecutar el flujo de limpieza y carga de datos en paralelo para las dimensiones DIM_RAMA y DIM_UNIVERSIDAD
 - **DIM_UNIVERSIDAD:** job encargado de ejecutar de forma individual el flujo de limpieza y carga de la dimensión DIM_UNIVERSIDAD.
 - **ETL_DIM_UNIVERSIDAD:** job encargado de limpiar la tabla DIM_UNIVERSIDAD y ejecutar la transformación LOAD_DIM_UNIVERSIDAD.
 - **LOAD_DIM_UNIVERSIDAD:** transformación encargada de extraer la información de las fuentes y carga de datos en tabla DIM_UNIVERSIDAD
 - **DIM_RAMA:** job encargado de ejecutar de forma individual el flujo de limpieza y carga de la dimensión DIM_RAMA.
 - **ETL_DIM_RAMA:** job encargado de limpiar la tabla DIM_RAMA y ejecutar la transformación LOAD_DIM_RAMA
 - **LOAD_DIM_RAMA:** transformación encargada de extraer la información de las fuentes y carga de datos en tabla DIM_RAMA

- CARGA_HECHOS

- ETL_HECHOS: job encargado de ejecutar el flujo de limpieza y carga de datos en paralelo para las tablas de hechos
 - FACT_PEGR_COMPARATIVA: job encargado de ejecutar de forma individual el flujo de limpieza y carga de la tabla de hecho FACT_PEGR_COMPARATIVA.
 - ETL_FACT_PEGR_COMPARATIVA: job encargado de limpiar la tabla FACT_PEGR_COMPARATIVA y ejecutar la transformación LOAD_FACT_PEGR_COMPARATIVA.
 - LOAD_FACT_PEGR_COMPARATIVA: transformación encargada de extraer la información de las fuentes y carga de datos en tabla FACT_PEGR_COMPARATIVA
- FACT_PEGR_EVOLUTIVO: job encargado de ejecutar de forma individual el flujo de limpieza y carga de la tabla de hecho FACT_PEGR_EVOLUTIVO.
- ETL_FACT_PEGR_EVOLUTIVO: job encargado de limpiar la tabla FACT_PEGR_EVOLUTIVO y ejecutar la transformación LOAD_FACT_PEGR_EVOLUTIVO
 - LOAD_FACT_PEGR_EVOLUTIVO: transformación encargada de extraer la información de las fuentes y carga de datos en tabla FACT_PEGR_EVOLUTIVO
- FACT_PEGR_INSERTADAS: job encargado de ejecutar de forma individual el flujo de limpieza y carga de la tabla de hecho FACT_PEGR_INSERTADAS.
- ETL_FACT_PEGR_INSERTADAS: job encargado de limpiar la tabla FACT_PEGR_INSERTADAS y ejecutar la transformación LOAD_FACT_PEGR_INSERTADAS
 - LOAD_FACT_PEGR_INSERTADAS: transformación encargada de extraer la información de las fuentes y carga de datos en tabla FACT_PEGR_INSERTADAS

- FACT_PEGR_PERFIL: job encargado de ejecutar de forma individual el flujo de limpieza y carga de la tabla de hecho FACT_PEGR_PERFIL.
- ETL_FACT_PEGR_PERFIL: job encargado de limpiar la tabla FACT_PEGR_PERFIL y ejecutar la transformación LOAD_FACT_PEGR_PERFIL
 - LOAD_FACT_PEGR_PERFIL: transformación encargada de extraer la información de las fuentes y carga de datos en tabla FACT_PEGR_PERFIL

Estructura de proyecto

Name	Type
error	
fuentes	
CARGA_INICIAL_DWH_EGRESADOS	JOB
ETL_DIMENSIONES	JOB
ETL_DIM_RAMA	JOB
ETL_DIM_UNIVERSIDAD	JOB
ETL_FACT_PEGR_COMPARATIVA	JOB
ETL_FACT_PEGR_EVOLUTIVO	JOB
ETL_FACT_PEGR_INSERTADAS	JOB
ETL_FACT_PEGR_PERFIL	JOB
ETL_HECHOS	JOB
LOAD_DIM_RAMA	TRANSFORMATION
LOAD_DIM_UNIVERSIDAD	TRANSFORMATION
LOAD_FACT_PEGR_COMPARATIVA	TRANSFORMATION
LOAD_FACT_PEGR_EVOLUTIVO	TRANSFORMATION
LOAD_FACT_PEGR_INSERTADAS	TRANSFORMATION
LOAD_FACT_PEGR_PERFIL	TRANSFORMATION

3) Implementación del proceso ETL

El procesos inicia con la creación del modelo físico a partir de la definición obtenida en la fase de análisis, para ello se hará uso de scripts de sql que contienen los DDL necesarios para persistir el modelo en el respectivo SGBD, que para nuestro caso es, Microsoft SQL Server.

Creación del modelo multidimensional

En este paso crearemos las restricciones que se han diseñado y que son propias del modelo multidimensional, las claves primarias de las dimensiones y las foráneas de las tablas de hechos.

Tablas de Dimensiones

DIM_ANIO

```
CREATE TABLE [dbo].[DIM_ANIO](
[sk_dim_anio] [numeric](4, 0) NOT NULL,
[desc_anio] [varchar](50) 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
```

DIM_EDAD

```
CREATE TABLE [dbo].[DIM_EDAD](
[sk_dim_edad] [numeric](1, 0) NOT NULL,
[desc_int_edad] [varchar](50) NOT NULL,
CONSTRAINT [PK_DIM_EDAD] PRIMARY KEY CLUSTERED
(
[sk_dim_edad] 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
```

DIM_MODALIDAD

```
CREATE TABLE [dbo].[DIM_MODALIDAD](
[sk_dim_modalidad] [numeric](1, 0) NOT NULL,
[desc_modalidad] [varchar](50) NOT NULL,
CONSTRAINT [PK_DIM_MODALIDAD] PRIMARY KEY CLUSTERED
(
[sk_dim_modalidad] 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
```

DIM_PAIS

```
CREATE TABLE [dbo].[DIM_PAIS](
[sk_dim_pais] [numeric](3, 0) NOT NULL,
[desc_pais_es] [varchar](50) NOT NULL,
[desc_pais_en] [varchar](200) NOT NULL,
CONSTRAINT [PK_DIM_PAIS] PRIMARY KEY CLUSTERED
(
[sk_dim_pais] 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
```

DIM_TIPO_UNIV

```
CREATE TABLE [dbo].[DIM_TIPO_UNIV](
[sk_dim_tipo_univ] [numeric](1, 0) NOT NULL,
[desc_tipo_univ] [varchar](30) NOT NULL,
CONSTRAINT [PK_DIM_TIPO_UNIV] PRIMARY KEY CLUSTERED
(
[sk_dim_tipo_univ] 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
```

DIM_RAMA

```
CREATE TABLE [dbo].[DIM_RAMA]{
[sk_dim_rama] [numeric](5, 0) NOT NULL,
[cod_rama] [varchar](1) NOT NULL,
[desc_rama] [varchar](50) NOT NULL,
[cod_rama_n2] [varchar](2) NOT NULL,
[desc_rama_n2] [varchar](100) NOT NULL,
[cod_rama_n3] [varchar](3) NOT NULL,
[desc_rama_n3] [varchar](100) NOT NULL,
[cod_rama_n4] [varchar](4) NOT NULL,
[desc_rama_n4] [varchar](150) NOT NULL,
[cod_rama_n5] [varchar](6) NOT NULL,
[desc_rama_n5] [varchar](200) NOT NULL,
CONSTRAINT [PK_DIM_RAMA] PRIMARY KEY CLUSTERED
(
[sk_dim_rama] 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
```

DIM_UNIVERSIDAD

```
CREATE TABLE [dbo].[DIM_UNIVERSIDAD]{
[sk_dim_universidad] [numeric](3, 0) NOT NULL,
[desc_universidad] [varchar](100) NOT NULL,
[sk_dim_modalidad] [numeric](1, 0) NOT NULL,
CONSTRAINT [PK_DIM_UNIVERSIDAD] PRIMARY KEY CLUSTERED
(
[sk_dim_universidad] 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
```

Tablas de Hechos

FACT_PEGR_COMPARATIVA

```
CREATE TABLE [dbo].[FACT_PEGR_COMPARATIVA]([sk_dim_anio] [numeric](4, 0) NOT NULL, [sk_dim_pais] [numeric](3, 0) NOT NULL, [personas_egresadas] [numeric](8, 0) NULL, [porcentaje_pegr_jovenes] [numeric](3, 1) NULL, CONSTRAINT [PK_FACT_PEGR_COMPARATIVA] PRIMARY KEY CLUSTERED () [sk_dim_anio] ASC, [sk_dim_pais] 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
```

FACT_PEGR_EVOLUTIVO

```
CREATE TABLE [dbo].[FACT_PEGR_EVOLUTIVO]([sk_dim_anio] [numeric](4, 0) NOT NULL, [sk_dim_tipo_univ] [numeric](1, 0) NOT NULL, [sk_dim_universidad] [numeric](3, 0) NOT NULL, [cod_rama] [numeric](5, 0) NOT NULL, [personas_egresadas] [numeric](8, 0) NOT NULL, CONSTRAINT [PK_FACT_PEGR_EVOLUTIVO] PRIMARY KEY CLUSTERED () [sk_dim_anio] ASC, [sk_dim_tipo_univ] ASC, [sk_dim_universidad] ASC, [cod_rama] 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
```

FACT_PEGR_INSERTADAS

```
CREATE TABLE [dbo].[FACT_PEGR_INSERTADAS]([sk_dim_anio] [numeric](4, 0) NOT NULL,[sk_dim_tipo_univ] [numeric](1, 0) NOT NULL,[sk_dim_sexo] [numeric](1, 0) NOT NULL,[cod_rama] [numeric](5, 0) NOT NULL,[pegr_trabajando] [numeric](8, 0) NOT NULL,[pegr_desempleados] [numeric](8, 0) NOT NULL,[pegr_inactivos] [numeric](8, 0) NOT NULL,[sk_dim_anio_laboral] [numeric](4, 0) NOT NULL,CONSTRAINT [PK_FACT_PEGR_INSERTADAS] PRIMARY KEY CLUSTERED( [sk_dim_anio] ASC,[sk_dim_tipo_univ] ASC,[sk_dim_sexo] ASC,[cod_rama] 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
```

FACT_PEGR_PERFIL

```
CREATE TABLE [dbo].[FACT_PEGR_PERFIL]([sk_dim_anio] [numeric](4, 0) NOT NULL,[sk_dim_sexo] [numeric](1, 0) NOT NULL,[sk_dim_edad] [numeric](1, 0) NOT NULL,[cod_rama] [numeric](5, 0) NOT NULL,[personas_egresadas] [numeric](8, 0) NOT NULL,CONSTRAINT [PK_FACT_PEGR_PERFIL] PRIMARY KEY CLUSTERED( [sk_dim_anio] ASC,[sk_dim_sexo] ASC,[sk_dim_edad] ASC,[cod_rama] 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
```

Restricciones de Integridad

```
ALTER TABLE [dbo].[DIM_UNIVERSIDAD] WITH NOCHECK ADD CONSTRAINT  
[FK_DIM_UNIVERSIDAD_DIM_MODALIDAD] FOREIGN KEY([sk_dim_modalidad])  
REFERENCES [dbo].[DIM_MODALIDAD] ([sk_dim_modalidad])  
GO  
ALTER TABLE [dbo].[DIM_UNIVERSIDAD] CHECK CONSTRAINT  
[FK_DIM_UNIVERSIDAD_DIM_MODALIDAD]  
GO  
ALTER TABLE [dbo].[FACT_PEGR_COMPARATIVA] WITH CHECK ADD  
CONSTRAINT [FK_FACT_PEGR_COMPARATIVA_DIM_ANIO] FOREIGN  
KEY([sk_dim_anio])  
REFERENCES [dbo].[DIM_ANIO] ([sk_dim_anio])  
GO  
ALTER TABLE [dbo].[FACT_PEGR_COMPARATIVA] CHECK CONSTRAINT  
[FK_FACT_PEGR_COMPARATIVA_DIM_ANIO]  
GO  
ALTER TABLE [dbo].[FACT_PEGR_COMPARATIVA] WITH CHECK ADD CONSTRAINT  
[FK_FACT_PEGR_COMPARATIVA_DIM_PAIS] FOREIGN  
KEY([sk_dim_pais])  
REFERENCES [dbo].[DIM_PAIS] ([sk_dim_pais])  
GO  
ALTER TABLE [dbo].[FACT_PEGR_COMPARATIVA] CHECK CONSTRAINT  
[FK_FACT_PEGR_COMPARATIVA_DIM_PAIS]  
GO  
ALTER TABLE [dbo].[FACT_PEGR_EVOLUTIVO] WITH CHECK ADD CONSTRAINT  
[FK_FACT_PEGR_EVOLUTIVO_DIM_ANIO] FOREIGN KEY([sk_dim_anio])  
REFERENCES [dbo].[DIM_ANIO] ([sk_dim_anio])  
GO  
ALTER TABLE [dbo].[FACT_PEGR_EVOLUTIVO] CHECK CONSTRAINT  
[FK_FACT_PEGR_EVOLUTIVO_DIM_ANIO]  
GO  
ALTER TABLE [dbo].[FACT_PEGR_EVOLUTIVO] WITH CHECK ADD CONSTRAINT  
[FK_FACT_PEGR_EVOLUTIVO_DIM_RAMA] FOREIGN KEY([cod_rama])  
REFERENCES [dbo].[DIM_RAMA] ([sk_dim_rama])  
GO  
ALTER TABLE [dbo].[FACT_PEGR_EVOLUTIVO] CHECK CONSTRAINT  
[FK_FACT_PEGR_EVOLUTIVO_DIM_RAMA]  
GO  
ALTER TABLE [dbo].[FACT_PEGR_EVOLUTIVO] WITH CHECK ADD CONSTRAINT  
[FK_FACT_PEGR_EVOLUTIVO_DIM_TIPO_UNIV] FOREIGN  
KEY([sk_dim_tipo_univ])  
REFERENCES [dbo].[DIM_TIPO_UNIV] ([sk_dim_tipo_univ])  
GO
```

```
ALTER TABLE [dbo].[FACT_PEGR_EVOLUTIVO] CHECK CONSTRAINT
[FK_FACT_PEGR_EVOLUTIVO_DIM_TIPO_UNIV]
GO
ALTER TABLE [dbo].[FACT_PEGR_EVOLUTIVO] WITH CHECK ADD CONSTRAINT
[FK_FACT_PEGR_EVOLUTIVO_DIM_UNIVERSIDAD] FOREIGN
KEY([sk_dim_universidad])
REFERENCES [dbo].[DIM_UNIVERSIDAD] ([sk_dim_universidad])
GO
ALTER TABLE [dbo].[FACT_PEGR_EVOLUTIVO] CHECK CONSTRAINT
[FK_FACT_PEGR_EVOLUTIVO_DIM_UNIVERSIDAD]
GO
ALTER TABLE [dbo].[FACT_PEGR_INSERTADAS] WITH CHECK ADD
CONSTRAINT [FK_FACT_PEGR_INSERTADAS_DIM_ANIO] FOREIGN
KEY([sk_dim_anio])
REFERENCES [dbo].[DIM_ANIO] ([sk_dim_anio])
GO
ALTER TABLE [dbo].[FACT_PEGR_INSERTADAS] CHECK CONSTRAINT
[FK_FACT_PEGR_INSERTADAS_DIM_ANIO]
GO
ALTER TABLE [dbo].[FACT_PEGR_INSERTADAS] WITH CHECK ADD
CONSTRAINT [FK_FACT_PEGR_INSERTADAS_DIM_RAMA] FOREIGN
KEY([cod_rama])
REFERENCES [dbo].[DIM_RAMA] ([sk_dim_rama])
GO
ALTER TABLE [dbo].[FACT_PEGR_INSERTADAS] CHECK CONSTRAINT
[FK_FACT_PEGR_INSERTADAS_DIM_RAMA]
GO
ALTER TABLE [dbo].[FACT_PEGR_INSERTADAS] WITH CHECK ADD
CONSTRAINT [FK_FACT_PEGR_INSERTADAS_DIM_SEXO] FOREIGN
KEY([sk_dim_sexo])
REFERENCES [dbo].[DIM_SEXO] ([sk_dim_sexo])
GO
ALTER TABLE [dbo].[FACT_PEGR_INSERTADAS] CHECK CONSTRAINT
[FK_FACT_PEGR_INSERTADAS_DIM_SEXO]
GO
ALTER TABLE [dbo].[FACT_PEGR_INSERTADAS] WITH CHECK ADD
CONSTRAINT [FK_FACT_PEGR_INSERTADAS_DIM_TIPO_UNIV] FOREIGN
KEY([sk_dim_tipo_univ])
REFERENCES [dbo].[DIM_TIPO_UNIV] ([sk_dim_tipo_univ])
GO
ALTER TABLE [dbo].[FACT_PEGR_INSERTADAS] CHECK CONSTRAINT
[FK_FACT_PEGR_INSERTADAS_DIM_TIPO_UNIV]
```

```
ALTER TABLE [dbo].[FACT_PEGR_PERFIL] WITH CHECK ADD CONSTRAINT
[FK_FACT_PEGR_PERFIL_DIM_ANIO] FOREIGN KEY([sk_dim_anio])
REFERENCES [dbo].[DIM_ANIO] ([sk_dim_anio])
GO
ALTER TABLE [dbo].[FACT_PEGR_PERFIL] CHECK CONSTRAINT
[FK_FACT_PEGR_PERFIL_DIM_ANIO]
GO
ALTER TABLE [dbo].[FACT_PEGR_PERFIL] WITH CHECK ADD CONSTRAINT
[FK_FACT_PEGR_PERFIL_DIM_EDAD] FOREIGN KEY([sk_dim_edad])
REFERENCES [dbo].[DIM_EDAD] ([sk_dim_edad])
GO
ALTER TABLE [dbo].[FACT_PEGR_PERFIL] CHECK CONSTRAINT
[FK_FACT_PEGR_PERFIL_DIM_EDAD]
GO
ALTER TABLE [dbo].[FACT_PEGR_PERFIL] WITH CHECK ADD CONSTRAINT
[FK_FACT_PEGR_PERFIL_DIM_RAMA] FOREIGN KEY([cod_rama])
REFERENCES [dbo].[DIM_RAMA] ([sk_dim_rama])
GO
ALTER TABLE [dbo].[FACT_PEGR_PERFIL] CHECK CONSTRAINT
[FK_FACT_PEGR_PERFIL_DIM_RAMA]
GO
ALTER TABLE [dbo].[FACT_PEGR_PERFIL] WITH CHECK ADD CONSTRAINT
[FK_FACT_PEGR_PERFIL_DIM_SEXO] FOREIGN KEY([sk_dim_sexo])
REFERENCES [dbo].[DIM_SEXO] ([sk_dim_sexo])
GO
ALTER TABLE [dbo].[FACT_PEGR_PERFIL] CHECK CONSTRAINT
[FK_FACT_PEGR_PERFIL_DIM_SEXO]
GO
```

Creación del proceso de extracción, transformación y carga (ETL)

Variables de entorno

Se anexan los datos de conexión a la base de datos SQL Server como variables de entorno en el archivo kettle.properties, además se define el repositorio de trabajo sobre el directorio F:\Pentaho\PRA2

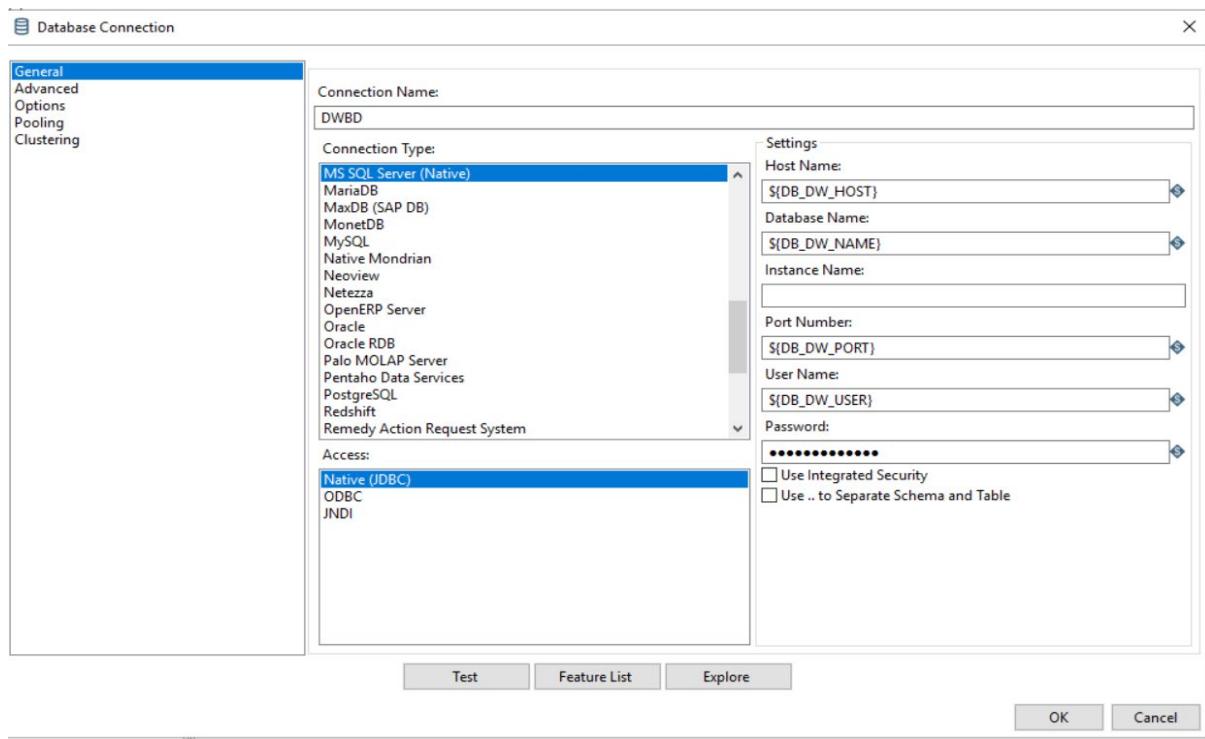
The screenshot shows two windows side-by-side. The top window is titled 'Kettle properties' and contains a table of environment variables:

#	Variable name	Value
1	DB_DW_HOST	UCS1R1UOCSQL01
2	DB_DW_NAME	DW_DB_jboteros
3	DB_DW_PASS	XXXXXX
4	DB_DW_PORT	1433
5	DB_DW_USER	STUDENT_jboteros

The bottom window is a file browser showing the contents of the 'F:\Pentaho\PRA2' directory. The 'DWBD.kdb' file is selected.

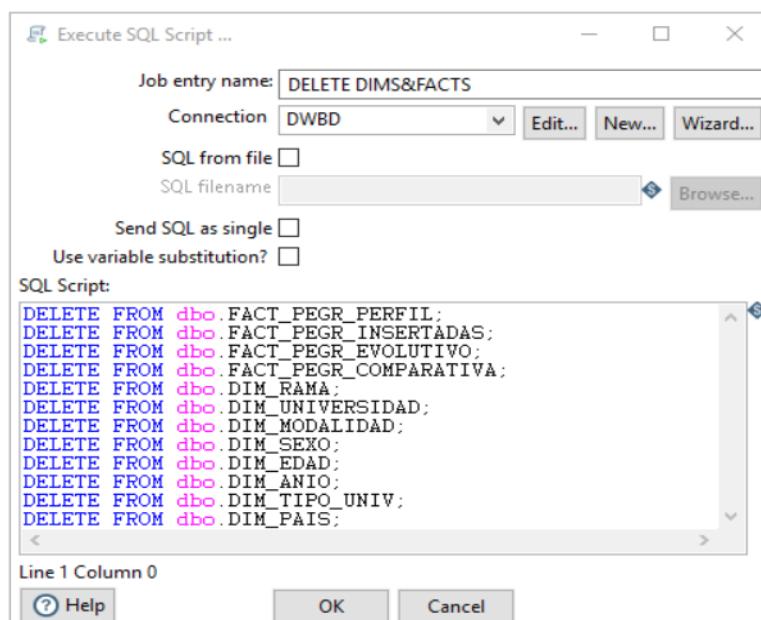
Nombre
.meta
error
fuentes
CARGA_INICIAL_DWH_EGRESADOS.kjb
DWBD.kdb
ETL_DIM_RAMA.kjb
ETL_DIM_UNIVERSIDAD.kjb
ETL_DIMENSIONES.kjb
ETL_FACT_PEGR_COMPARATIVA.kjb
ETL_FACT_PEGR_EVOLUTIVO.kjb
ETL_FACT_PEGR_INSERTADAS.kjb
ETL_FACT_PEGR_PERFIL.kjb
ETL_HECHOS.kjb
LOAD_DIM_RAMA.ktr
LOAD_DIM_UNIVERSIDAD.ktr
LOAD_FACT_PEGR_COMPARATIVA.ktr
LOAD_FACT_PEGR_EVOLUTIVO.ktr
LOAD_FACT_PEGR_INSERTADAS.ktr
LOAD_FACT_PEGR_PERFIL.ktr
repository

Crear conexión a base de datos DWBD utilizando las variables de entorno:

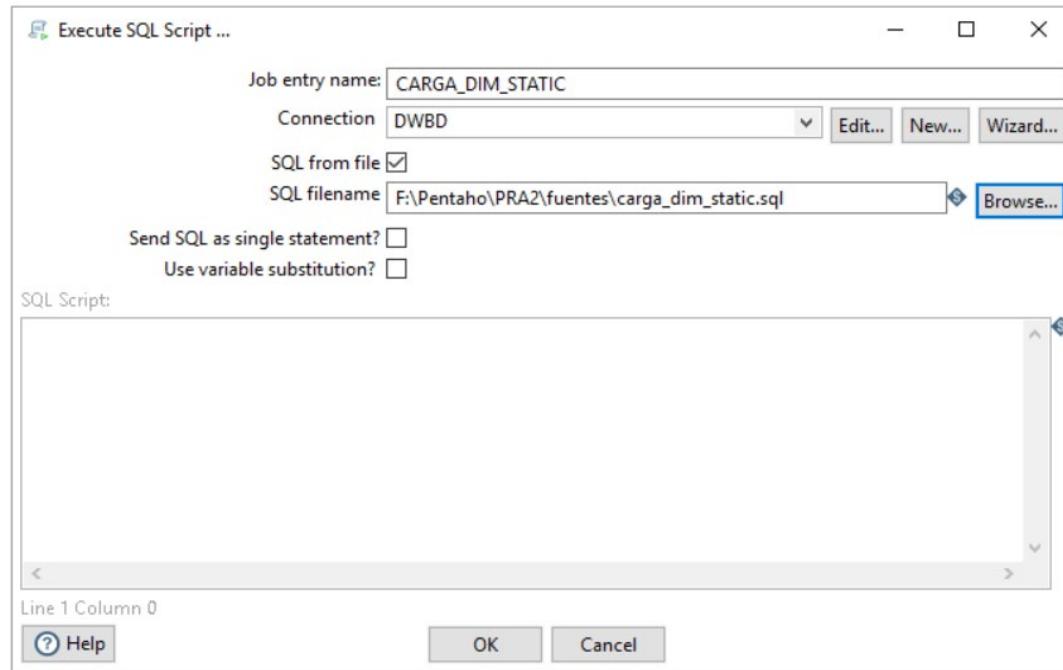


A continuación pasamos a explicar los componentes de la ETL principal y sus flujos de datos:

- **DELETE DIMS&FACTS:** scripting sql encargado de limpiar la información de las tablas del modelo.



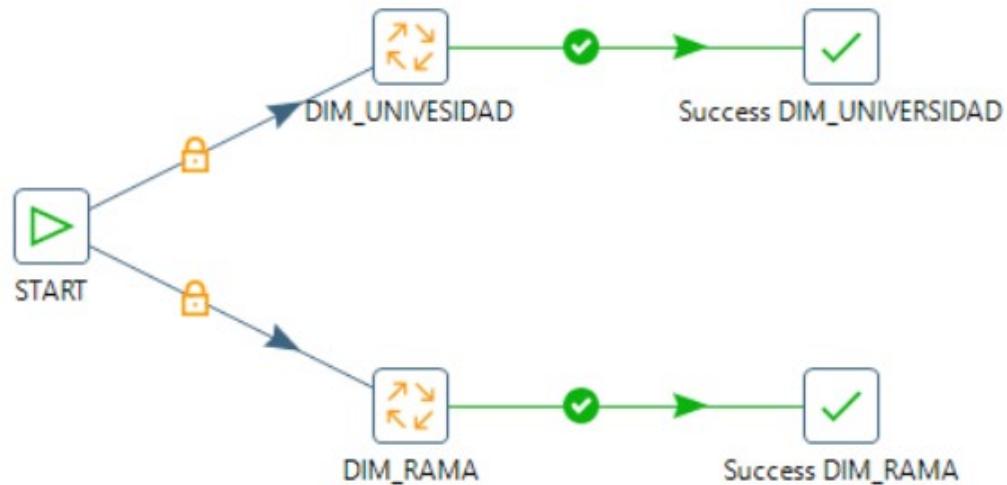
- CARGA_DIM_STATIC: scripting sql encargado de cargar las dimensiones DIM_ANIO, DIM_EDAD, DIM_TIPO_UNIV, DIM_SEXO, DIM_PAIS y DIM_MODALIDAD, a partir de las instrucciones contenidas en el archivo F:\Pentaho\PRA2\fuentes\carga_dim_static.sql



- CARGA_DIMENSIONES: job principal encargado de ejecutar el proceso etl en paralelo para las dimensiones DIM_RAMA y DIM_UNIVERSIDAD



- ETL_DIMENSIONES: job encargado de ejecutar el flujo de limpieza y carga de datos en paralelo para las dimensiones DIM_RAMA y DIM_UNIVERSIDAD



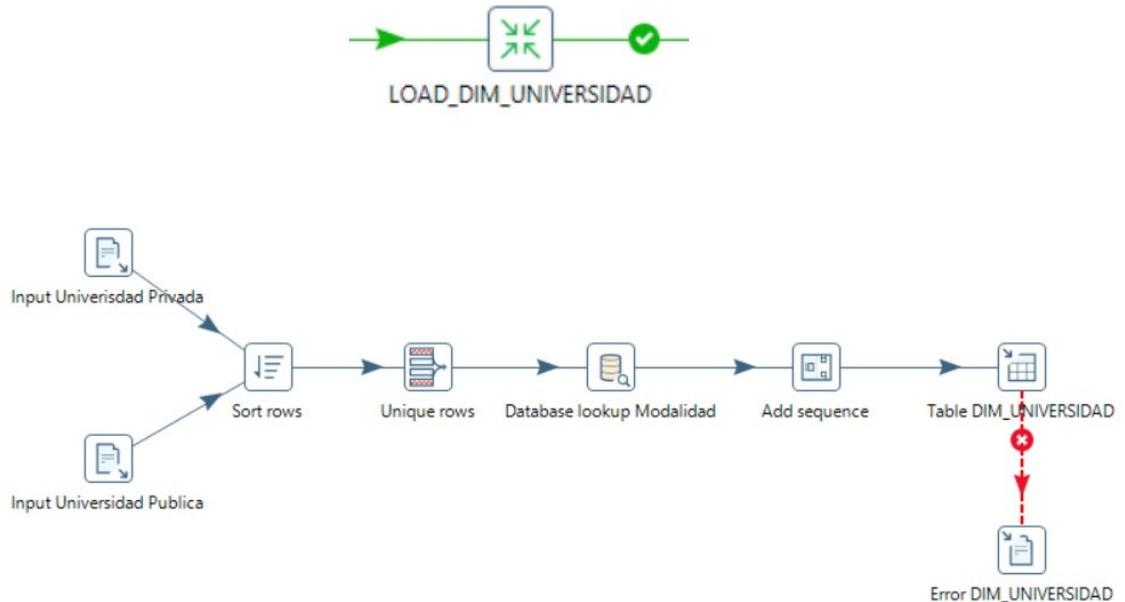
- DIM_UNIVERSIDAD: job encargado de ejecutar de forma individual el flujo de limpieza y carga de la dimensión DIM_UNIVERSIDAD.



- ETL_DIM_UNIVERSIDAD: job encargado de limpiar la tabla DIM_UNIVERSIDAD y ejecutar la transformación LOAD_DIM_UNIVERSIDAD.



- LOAD_DIM_UNIVERSIDAD: transformación encargada de extraer la información de las fuentes y carga de datos en tabla DIM_UNIVERSIDAD



En el primer paso de input, se cargan los campos UNIVERSIDAD y MODALIDAD a partir de los archivos SEGR1.csv y SEGR2.csv que contienen la información correspondiente a los nombres de cada universidad y su respectiva modalidad de impartición

CSV Input

Step name: Input Universidad Privada

Filename: F:\Pentaho\PRA2\fuentes\SEGR1.csv

Delimiter: ;

Enclosure: "

NIO buffer size: 50000

Lazy conversion?

Header row present?

Add filename to result

The row number field name (optional):

Running in parallel?

New line possible in fields?

File encoding:

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	MODALIDAD	String				€	,	.	none
2	UNIVERSIDAD	String				€	,	.	none

CSV Input

Step name	Input Universidad Publica
Filename	F:\Pentaho\PRA2\fuentes\SEGR2.csv
Delimiter	;
Enclosure	"
NIO buffer size	50000
Lazy conversion?	<input type="checkbox"/>
Header row present?	<input checked="" type="checkbox"/>
Add filename to result	<input type="checkbox"/>
The row number field name (optional)	
Running in parallel?	<input type="checkbox"/>
New line possible in fields?	<input type="checkbox"/>
File encoding	

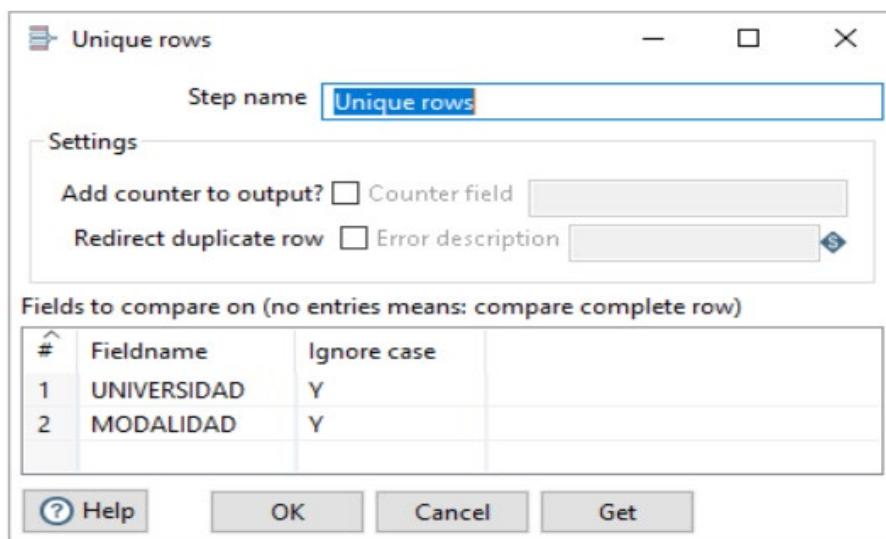
#	Name	Type	Format	Length	Precision	Currency	Decimal	Group
1	MODALIDAD	String						
2	UNIVERSIDAD	String						

En el paso Sort Rows, se ordena el conjunto de datos obtenido de la unión de ambas fuentes, como prerequisito para seleccionar las combinaciones únicas UNIVERSIDAD-MODALIDAD que se cargarán en la tabla destino.

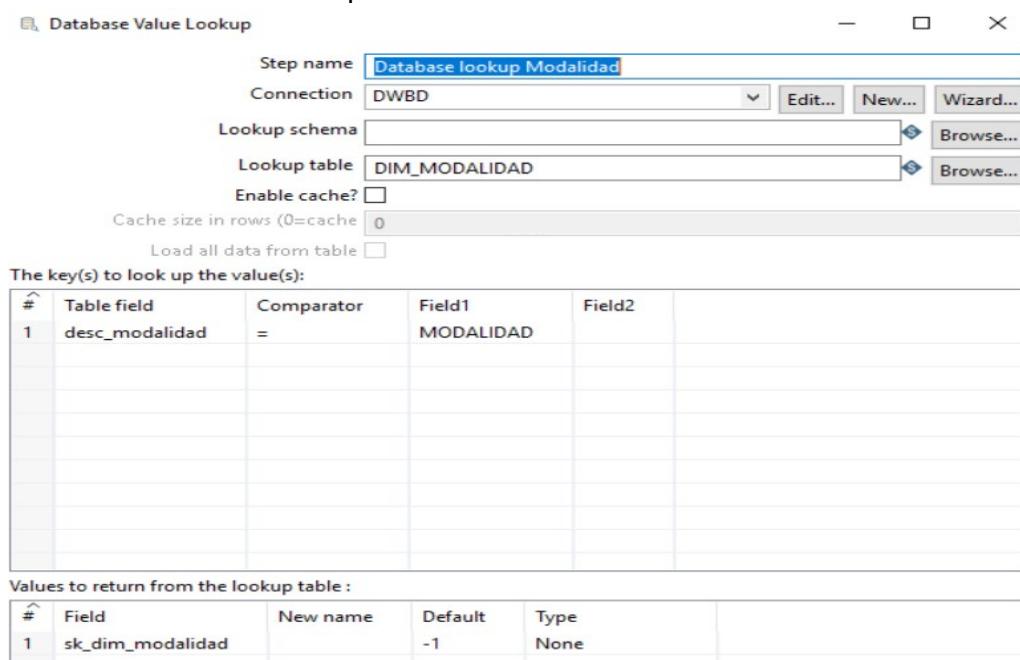
Sort rows

Step name	Sort rows					
Sort directory	%%java.io.tmpdir%% <input type="button" value="Browse..."/>					
TMP-file prefix	out					
Sort size (rows in memory)	1000000					
Free memory threshold (in %)						
Compress TMP Files?	<input type="checkbox"/>					
Only pass unique rows? (verifies keys only)	<input type="checkbox"/>					
Fields :						
#	Fieldname	Ascending	Case sensitive compare?	Sort based on current locale?	Collator Strength	Presorted?
1	UNIVERSIDAD	Y	N	N	0	N
2	MODALIDAD	Y	N	N	0	N

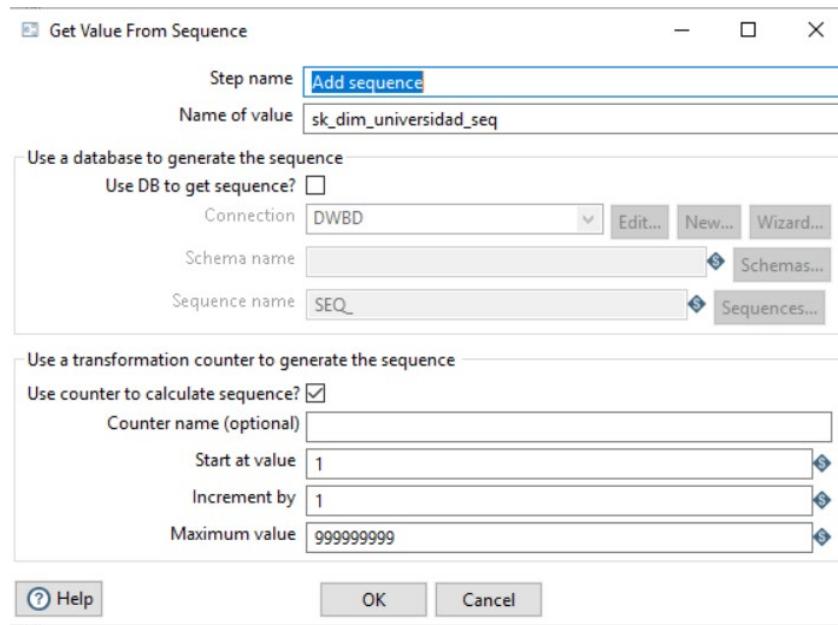
En el paso Unique Rows, se obtiene los valores únicos UNIVERSIDAD-MODALIDAD del subconjunto de datos anterior, dejando así las combinaciones necesarias para la carga de datos.



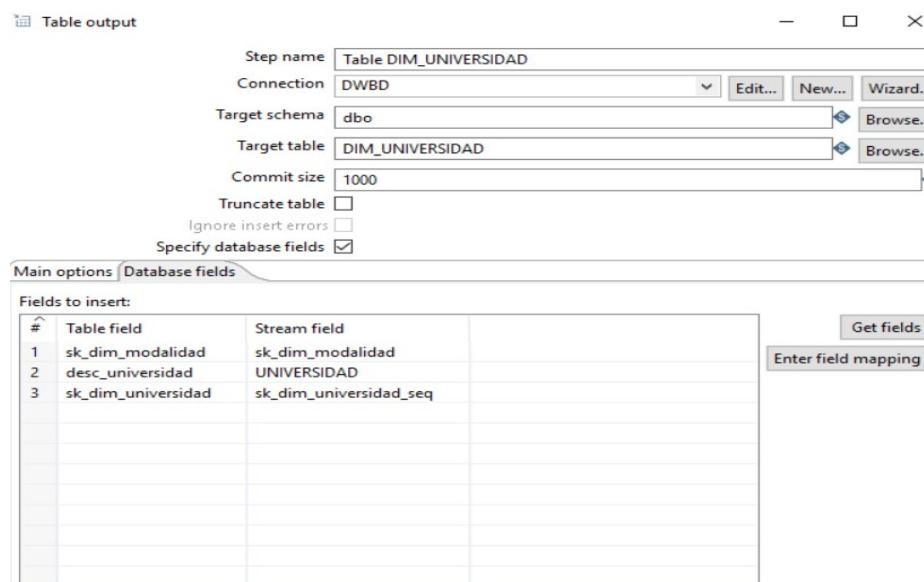
Como en la definición del modelo físico, la dimensión DIM_UNIVERSIDAD tiene una llave de integridad referencial hacia la dimensión DIM_MODALIDAD, es necesario ejecutar un proceso de lookup sobre esta última, para obtener así el valor de la clave primaria correspondiente a la modalidad de impartición para una universidad dada. Como se observa en la definición, si el valor no es encontrado, se asigna un valor de -1 como valor por defecto, de esta forma podemos identificar errores de integridad en los datos al momento de la carga final dado que todas las secuencias validas de llave primaria en todas las dimensiones comienzan a partir de 1.

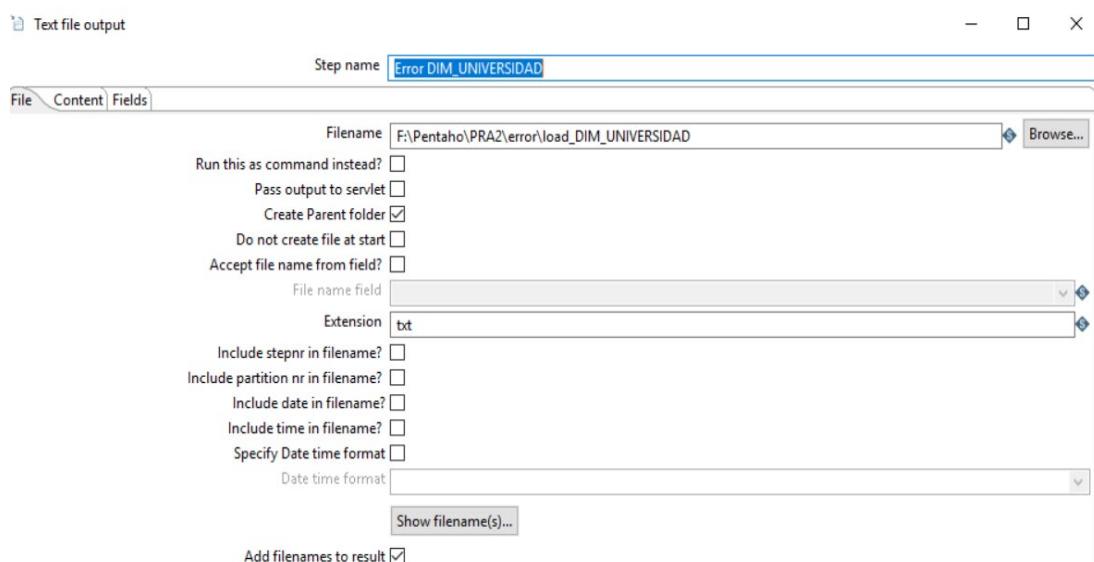


En el paso Add Sequence, se generan de forma automática los valores de secuencia que harán parte de la clave primaria en la dimensión DIM_UNIVERSIDAD, iniciando a partir del valor 1 y continuando con incrementos de 1 unidad: 1,2,3,4..., como se había definido en la fase de diseño las secuencias no se generan en base de datos para evitar la dependencia con el gestor de bases de datos.



Por último en el paso Table DIM_UNIVERSIDAD, se mapean los valores obtenidos de las fases anteriores con los respectivos campos a poblar en la tabla DIM_UNIVERSIDAD, y se especifica un flujo alterno ante falla, con la salida de error a un archivo de texto plano.





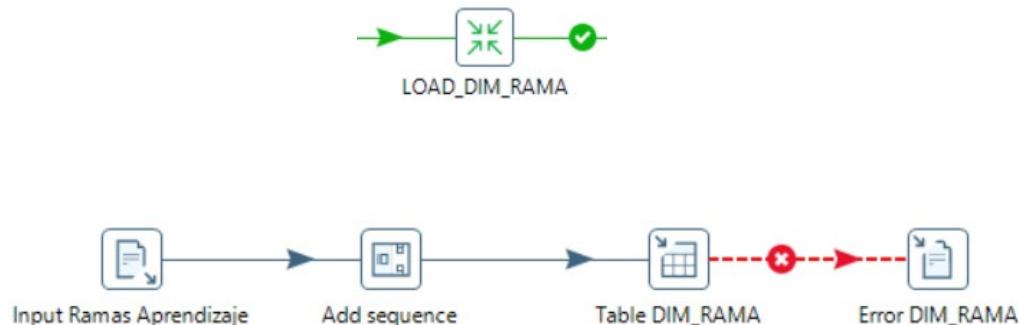
- **DIM_RAMA:** job encargado de ejecutar de forma individual el flujo de limpieza y carga de la dimensión DIM_RAMA.



- **ETL_DIM_RAMA:** job encargado de limpiar la tabla DIM_RAMA y ejecutar la transformación LOAD_DIM_RAMA.



- LOAD_DIM_RAMA: transformación encargada de extraer la información de las fuentes y carga de datos en tabla DIM_RAMA

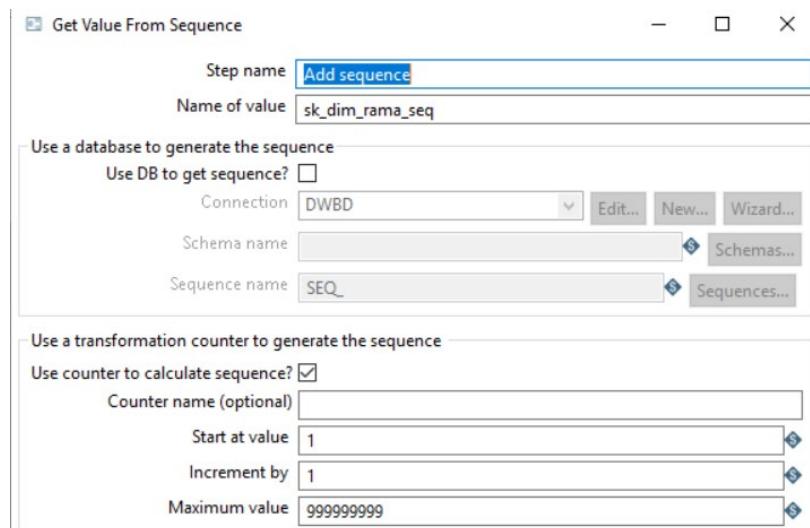


En el primer paso de input, se cargan los campos planos tal cual se encuentran en el archivo ISCED_2013.csv y acorde a las definiciones de tipo de dato y tamaño definido en el modelo físico para la tabla DIM_RAMA

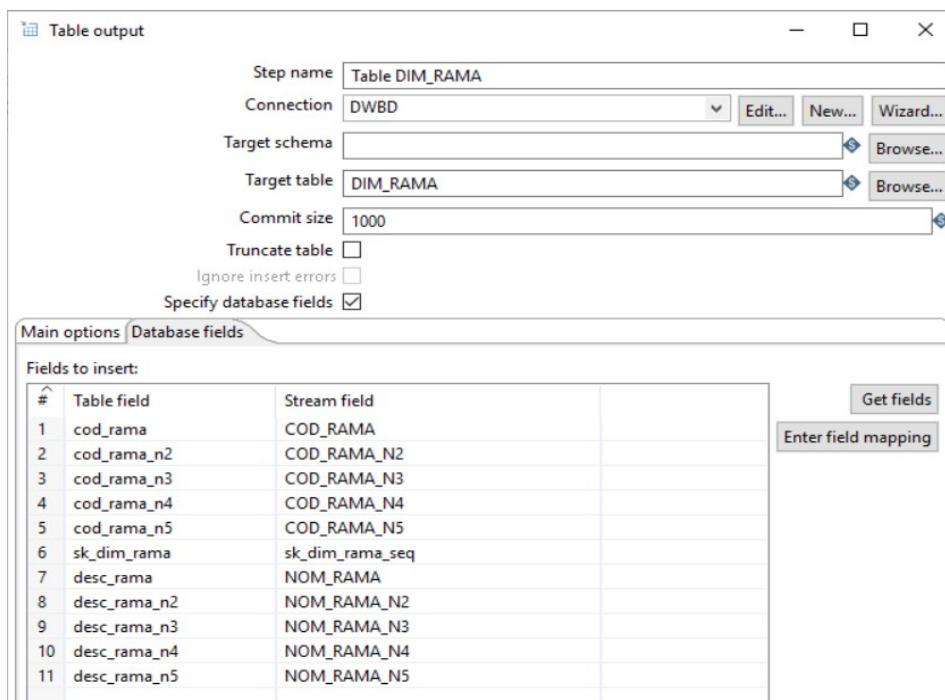
The screenshot shows the configuration for the 'Input Ramas Aprendizaje' step. The 'Step name' is set to 'Input Ramas Aprendizaje'. The 'Filename' is 'F:\Pentaho\PRA2\fuentes\ISCED_2013.csv'. The 'Delimiter' is ';' and the 'Enclosure' is '\"'. The 'NIO buffer size' is '50000'. Under 'Advanced' settings, 'Header row present?' is checked, and 'File encoding' is set to 'UTF-8'. Below these settings is a table mapping fields:

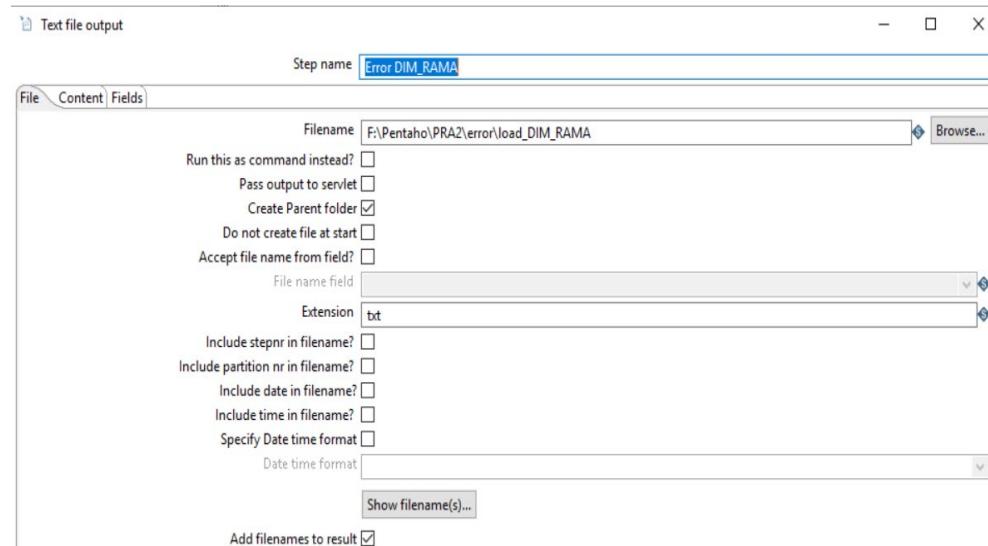
#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	COD_RAMA	String		1		€	,	.	none
2	NOM_RAMA	String		50		€	,	.	none
3	COD_RAMA_N2	String		2		€	,	.	none
4	NOM_RAMA_N2	String		100		€	,	.	none
5	COD_RAMA_N3	String		3		€	,	.	none
6	NOM_RAMA_N3	String		100		€	,	.	none
7	COD_RAMA_N4	String		4		€	,	.	none
8	NOM_RAMA_N4	String		150		€	,	.	none
9	COD_RAMA_N5	String		6		€	,	.	none
10	NOM_RAMA_N5	String		200		€	,	.	none

En el paso Add Sequence, se generan de forma automática los valores de secuencia que harán parte de la clave primaria en la dimensión DIM_RAMA, iniciando a partir del valor 1 y continuando con incrementos de 1 unidad: 1,2,3,4..., como se había definido en la fase de diseño las secuencias no se generan en base de datos para evitar la dependencia con el gestor de bases de datos.



Por último en el paso Table DIM_RAMA, se mapean los valores obtenidos de las fases anteriores con los respectivos campos a poblar en la tabla DIM_RAMA, y se especifica un flujo alterno ante falla, con la salida de error a un archivo de texto plano.

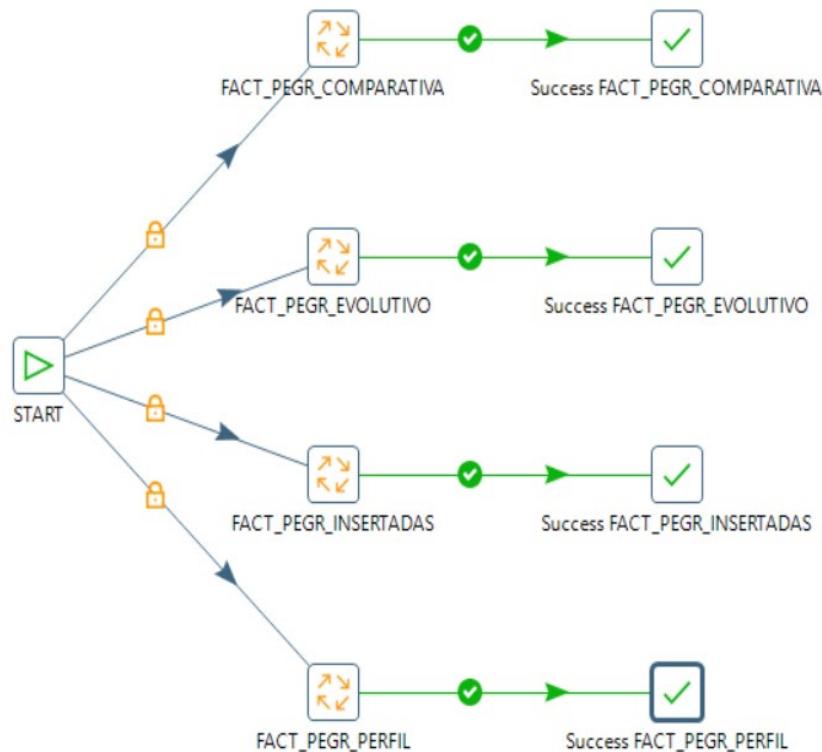




- CARGA_HECHOS: job principal encargado de ejecutar el proceso etl en paralelo para las 4 tablas de hecho definidas en el diseño lógico y físico del DW.



- ETL_HECHOS: job encargado de ejecutar el flujo de limpieza y carga de datos en paralelo para las tablas de hechos



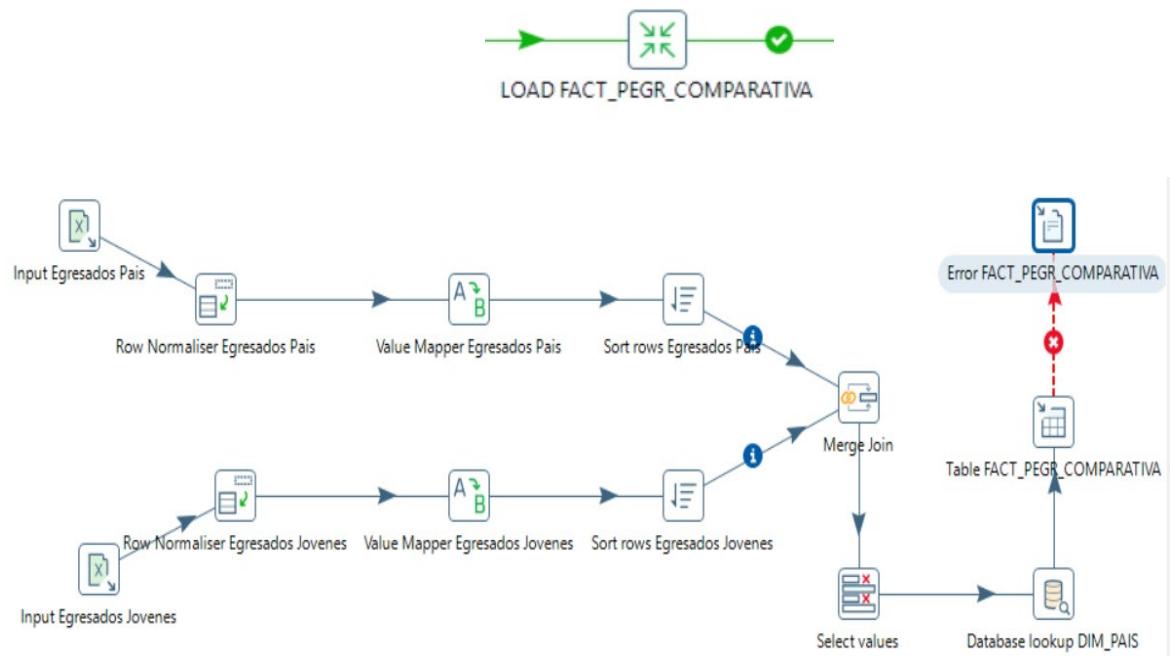
- FACT_PEGR_COMPARATIVA: job encargado de ejecutar de forma individual el flujo de limpieza y carga de la tabla de hecho FACT_PEGR_COMPARATIVA.



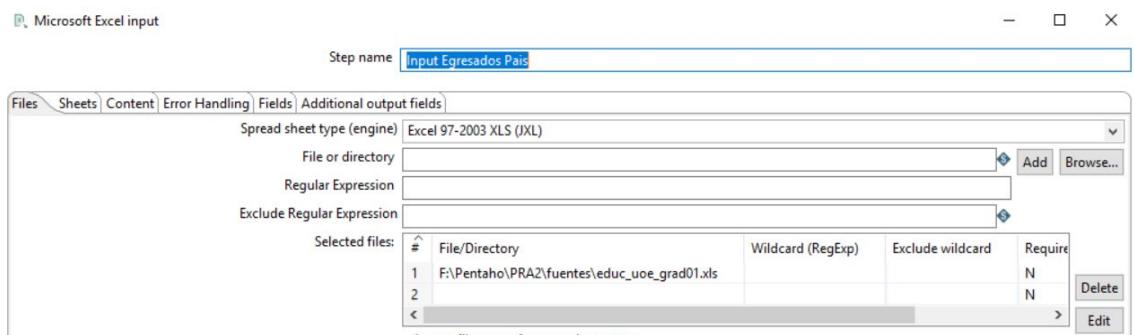
- ETL_FACT_PEGR_COMPARATIVA: job encargado de limpiar la tabla FACT_PEGR_COMPARATIVA y ejecutar la transformación LOAD_FACT_PEGR_COMPARATIVA.

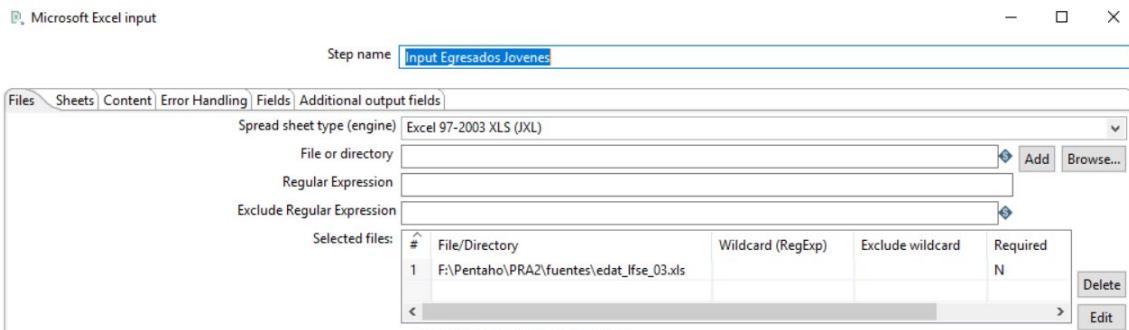


- LOAD_FACT_PEGR_COMPARATIVA: transformación encargada de extraer la información de las fuentes y carga de datos en tabla FACT_PEGR_COMPARATIVA.

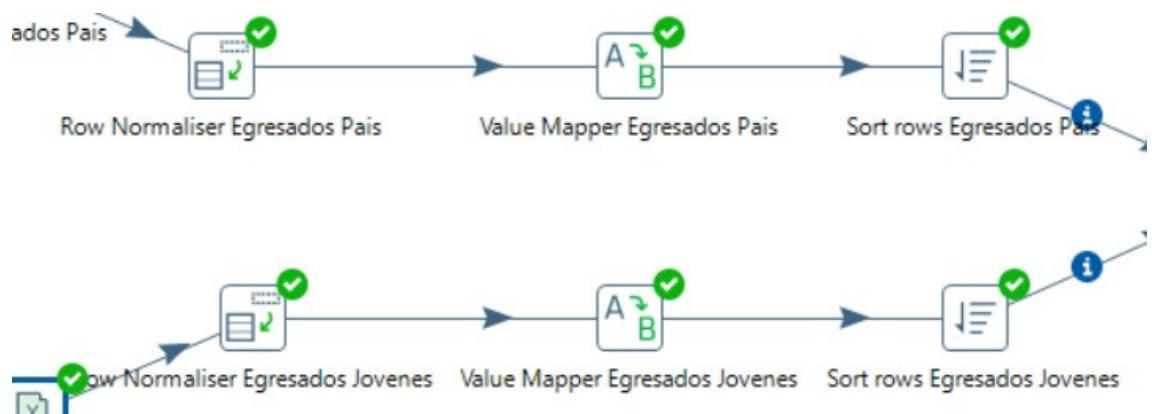


En el paso de input de datos, se utiliza la función para leer archivos .xls, uno de ellos referenciando la cantidad total de egresados por país, educ_uee_grad01.xls , y el otro referencia el archivo con los porcentajes de egresados jóvenes por cada uno de los países, edat_lfse_03.xls





En los siguientes 3 pasos, se aplica la misma lógica sobre cada uno de los archivos fuente, una paso de normalización de manera que se obtengan los valores objetivo del total y el porcentaje de egresados en forma columnar agrupados por país y año, una función de value mapper para reemplazar los valores no conocidos de los valores numéricos (:) por el valor -1, y por último un paso de ordenamiento por país y año, como prerequisito para la función de merge de las dos fuentes de datos en solo una.



Ejemplo step de normalización fuente egresados por país y año

Examine preview data

Rows of step: Row Normaliser Egresados Pais (60 rows)

#	GEO/TIME	Anno	Egresados
1	Denmark	2013	66467,0
2	Denmark	2014	70245,0
3	Denmark	2015	74428,0
4	Denmark	2016	85290,0
5	Denmark	2017	82581,0
6	Germany (until 1990 former territory of the FRG)	2013	495808,0
7	Germany (until 1990 former territory of the FRG)	2014	521845,0
8	Germany (until 1990 former territory of the FRG)	2015	544743,0
9	Germany (until 1990 former territory of the FRG)	2016	556800,0
10	Germany (until 1990 former territory of the FRG)	2017	569154,0
11	Ireland	2013	61297,0
12	Ireland	2014	64955,0
13	Ireland	2015	67303,0
14	Ireland	2016	65362,0
15	Ireland	2017	78002,0

En el paso de merge join se unifican las fuentes de datos por los campos país y año, como resultado se consolidan las métricas de total de egresados y el porcentaje de jóvenes egresados en un solo flujo de datos



Step name Merge Join	
First Step: Sort rows Egresados Pais	
Second Step: Sort rows Egresados Jovenes	
Join Type: INNER	
Keys for 1st step:	
#	Key field
1	GEO/TIME
2	Anno
Keys for 2nd step:	
#	Key field
1	GEO/TIME
2	Anno

Luego de esto, se seleccionan los valores necesarios para la carga final de datos en la tabla de hecho con los respectivos tipos de datos especificados para el modelo físico

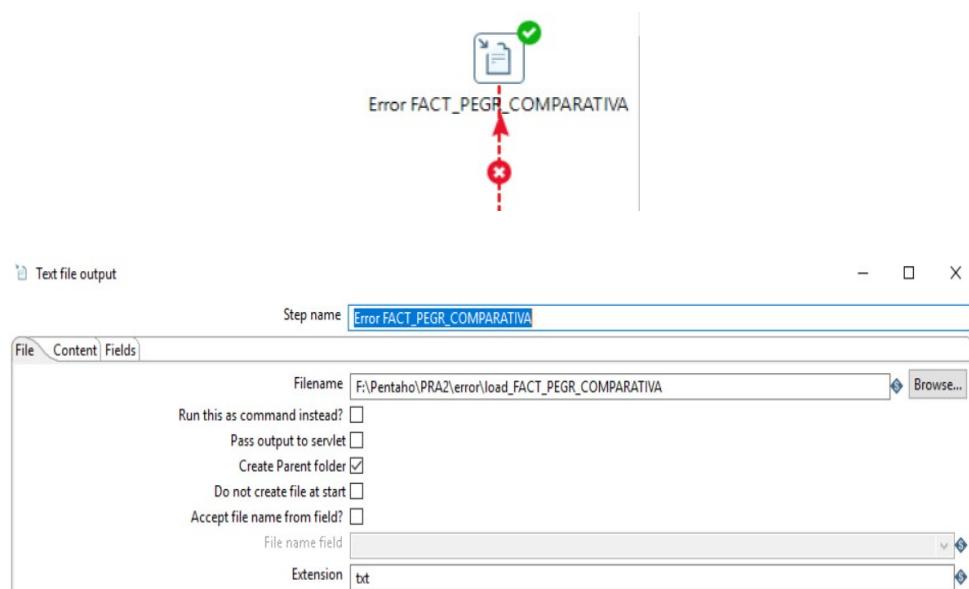
Select / Rename values							
Step name Select values							
Select & Alter / Remove / Meta-data							
Fields to alter the meta-data for :							
#	Fieldname	Rename to	Type	Length	Precision	Binary to Normal?	Format
1	PEGR_JOVENES		Number			N	,#
2	GEO/TIME		String			N	
3	Egresados		Number	0	N		#
4	Anno		Number	0	N		#

Como en la definición del modelo físico, la tabla de hecho tiene referencias foráneas a las dimensiones DIM_PAIS y DIM_ANIO, se utiliza el paso Database lookup DIM_PAIS, para obtener así el valor de la clave primaria correspondiente al país para una tupla dada. Como se observa en la definición, si el valor no es encontrado, se asigna un valor de -1 como valor por defecto, de esta forma podemos identificar errores de integridad en los datos al momento de la carga final dado que todas las secuencias validas de llave primaria en todas las dimensiones comienzan a partir de 1. Como para el año ya contamos con los valores numéricos que hacen parte de la llave primaria en DIM_ANIO, no es necesario hacer un lookup sobre esta dimensión.

Database Value Lookup

Step name	Database lookup DIM_PAIS			
Connection	DWBD	<input type="button" value="Edit..."/>	<input type="button" value="New..."/>	
Lookup schema	dbo	<input type="button" value="Browse..."/>		
Lookup table	DIM_PAIS	<input type="button" value="Browse..."/>		
Enable cache?	<input type="checkbox"/>			
Cache size in rows (0=cache)	0			
Load all data from table	<input type="checkbox"/>			
The key(s) to look up the value(s):				
#	Table field	Comparator	Field1	Field2
1	desc_pais_en	=	GEO/TIME	
Values to return from the lookup table:				
#	Field	New name	Default	Type
1	sk_dim_pais		-2	Number

Por último en el paso Table FACT_PEGR_COMPARATIVA, se mapean los valores obtenidos de las fases anteriores con los respectivos campos a poblar en la tabla FACT_PEGR_COMPARATIVA, y se especifica un flujo alterno ante falla, con la salida de error a un archivo de texto plano.



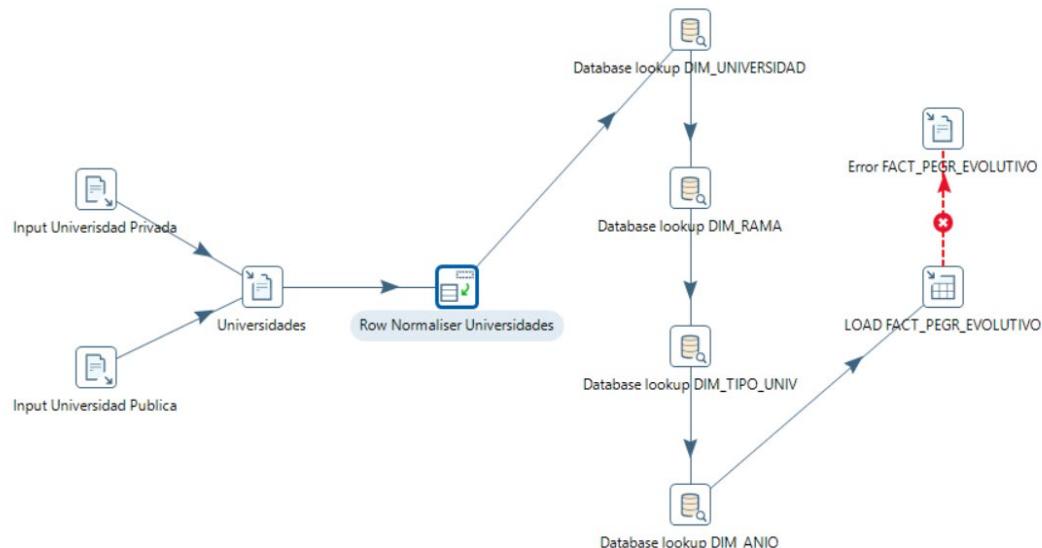
- FACT_PEGR_EVOLUTIVO: job encargado de ejecutar de forma individual el flujo de limpieza y carga de la tabla de hecho FACT_PEGR_EVOLUTIVO.



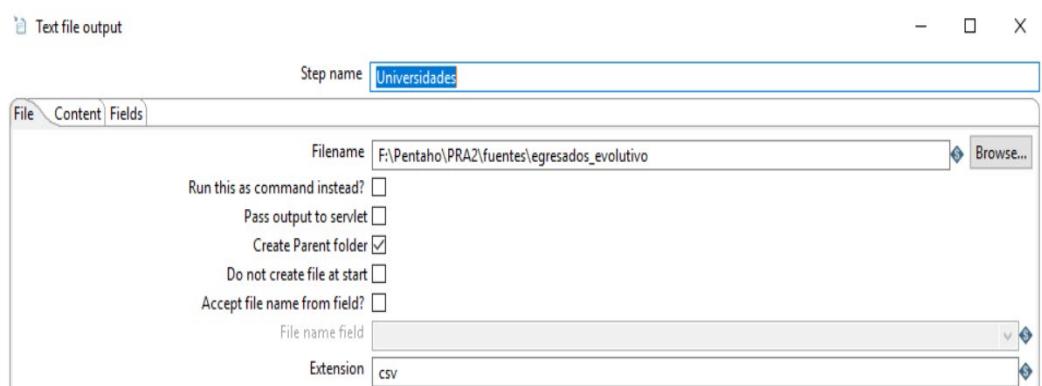
- ETL_FACT_PEGR_EVOLUTIVO: job encargado de limpiar la tabla FACT_PEGR_EVOLUTIVO y ejecutar la transformación LOAD_FACT_PEGR_EVOLUTIVO



- LOAD_FACT_PEGR_EVOLUTIVO: transformación encargada de extraer la información de las fuentes y carga de datos en tabla FACT_PEGR_EVOLUTIVO



Como parte del procesamiento de las fuentes de entrada, se unifica la información de los archivos SEGR1.csv y SEGR2.csv (incluyendo todos los campos), en un solo archivo llamado egresados_evolutivo.csv, de esta manera facilitamos la lectura de la información y las transformaciones necesarias en las fases posteriores



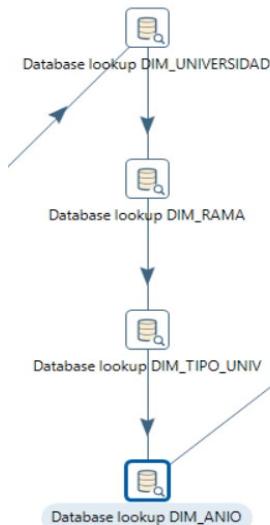
Homologo al paso realizado en la tabla de hechos anterior, se aplica una función de normalización, de forma que los datos queden agrupados verticalmente por año y el número total de egresados

#	Fieldname	Type	new field
1	EGR_C16_17	2016-2017	Egresados
2	EGR_C15_16	2015-2016	Egresados
3	EGR_C14_15	2014-2015	Egresados
4	EGR_C13_14	2013-2014	Egresados
5	EGR_C12_13	2012-2013	Egresados
6	EGR_C11_12	2011-2012	Egresados
7	EGR_C10_11	2010-2011	Egresados
8	EGR_C09_10	2009-2010	Egresados

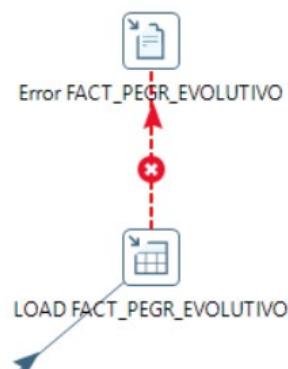
Ejemplo normalización por año y egresados

#	TIPO_UNIVERSIDAD	MODALIDAD	UNIVERSIDAD	RAMA_ENSEÑANZA	Anno	Egresados
1	Universidades Públicas	Presencial	A Coruña	Ciencias Sociales y Jurídicas	2016-2017	1567
2	Universidades Públicas	Presencial	A Coruña	Ciencias Sociales y Jurídicas	2015-2016	1669
3	Universidades Públicas	Presencial	A Coruña	Ciencias Sociales y Jurídicas	2014-2015	1662
4	Universidades Públicas	Presencial	A Coruña	Ciencias Sociales y Jurídicas	2013-2014	1915
5	Universidades Públicas	Presencial	A Coruña	Ciencias Sociales y Jurídicas	2012-2013	1712
6	Universidades Públicas	Presencial	A Coruña	Ciencias Sociales y Jurídicas	2011-2012	1434
7	Universidades Públicas	Presencial	A Coruña	Ciencias Sociales y Jurídicas	2010-2011	1629
8	Universidades Públicas	Presencial	A Coruña	Ciencias Sociales y Jurídicas	2009-2010	1566
9	Universidades Públicas	Presencial	A Coruña	Ingeniería y Arquitectura	2016-2017	886
10	Universidades Públicas	Presencial	A Coruña	Ingeniería y Arquitectura	2015-2016	1187
11	Universidades Públicas	Presencial	A Coruña	Ingeniería y Arquitectura	2014-2015	1349
12	Universidades Públicas	Presencial	A Coruña	Ingeniería y Arquitectura	2013-2014	1179
13	Universidades Públicas	Presencial	A Coruña	Ingeniería y Arquitectura	2012-2013	1233
14	Universidades Públicas	Presencial	A Coruña	Ingeniería y Arquitectura	2011-2012	1278
15	Universidades Públicas	Presencial	A Coruña	Ingeniería y Arquitectura	2010-2011	1151
16	Universidades Públicas	Presencial	A Coruña	Ingeniería y Arquitectura	2009-2010	1224
17	Universidades Públicas	Presencial	A Coruña	Artes y Humanidades	2016-2017	77
18	Universidades Públicas	Presencial	A Coruña	Artes y Humanidades	2015-2016	87

En las fases posteriores de Database lookup, se ejecuta el proceso de recuperación de las claves primarias para las dimensiones DIM_UNIVERSIDAD, DIM_RAMA, DIM_TIPO_UNIV y DIM_ANIO que son almacenarlas en los respectivos campos de clave foránea en la tabla de hechos, con la misma regla por defecto, valor -1 para las claves no recuperadas y controlar así los errores de integridad referencial para datos no existentes.



Por último en el paso Table FACT_PEGR_EVOLUTIVO, se mapean los valores obtenidos de las fases anteriores con los respectivos campos a poblar en la tabla FACT_PEGR_EVOLUTIVO, y se especifica un flujo alterno ante falla, con la salida de error a un archivo de texto plano utilizando el mismo patrón de secuencias anteriores para esta misma fase



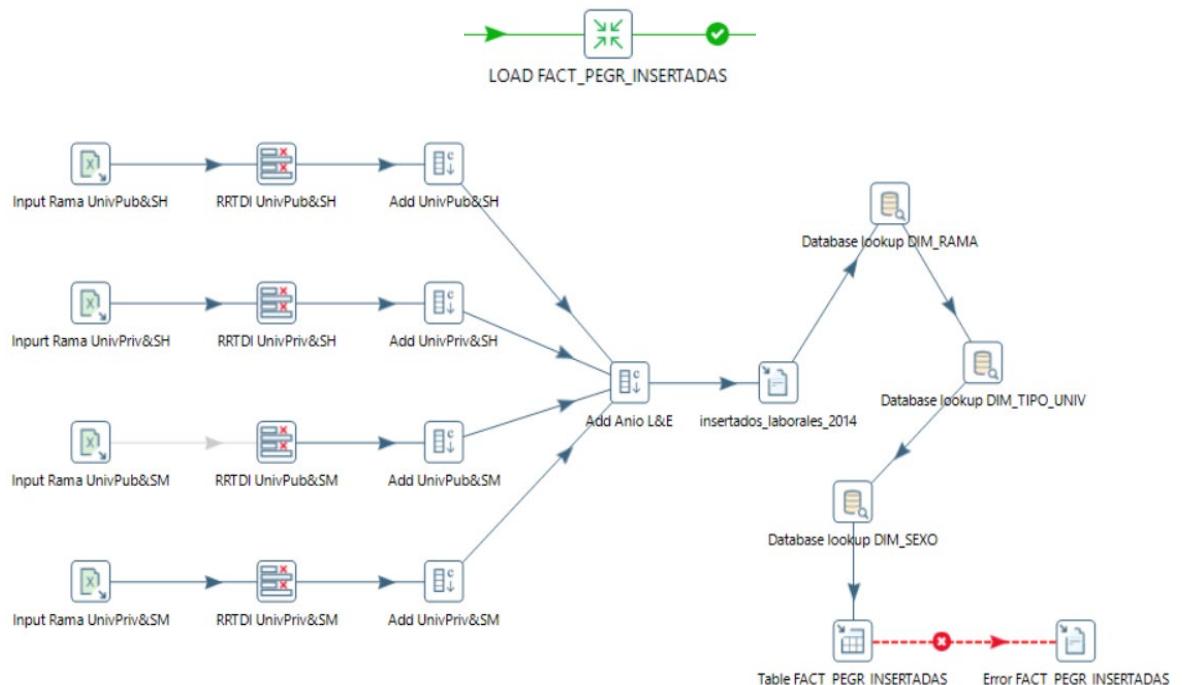
- FACT_PEGR_INSERTADAS: job encargado de ejecutar de forma individual el flujo de limpieza y carga de la tabla de hecho FACT_PEGR_INSERTADAS.



- ETL_FACT_PEGR_INSERTADAS: job encargado de limpiar la tabla FACT_PEGR_INSERTADAS y ejecutar la transformación LOAD_FACT_PEGR_INSERTADAS



- LOAD_FACT_PEGR_INSERTADAS: transformación encargada de extraer la información de las fuentes y cargar de datos en tabla FACT_PEGR_INSERTADAS



La estrategia utilizada para procesar el archivo 03003.xlsx, correspondiente a los datos de inserción laboral en 2014 para los egresados 2009-2010, fue segmentar la información por los bloques correspondientes a los niveles mínimos de agregación en la tabla dinámica, partiendo del hecho que todos los totales obtenidos por las agrupaciones de Ambos Sexos, Sexo y Tipo Universidad, pueden obtenerse ejecutando operaciones de agregación sobre los niveles de menor granularidad

Hombres-Universidad Pública-Rama: filas de la 39 a la 44

Hombres-Universidad Privada-Rama: filas de la 46 a la 51

Mujeres-Universidad Pública-Rama: filas de la 61 a la 66

Mujeres-Universidad Privada-Rama: filas de la 68 a la 73

De esta forma se definió la selección de datos en cada una de las fases de input en el anterior diagrama.

Como resultado de la fase de selección, se obtienen encabezados correspondientes a los totales de cada bloque, por cada una de las variables categóricas Trabajando, En desempleo, Inactivo, al igual que el campo Total con la sumatoria de egresados acorde a cada sección en la tabla dinámica. Para normalizar los encabezados para cada bloque, se aplica la función de Select/Rename values de la siguiente forma:

El campo Total, se renombra a **rama**, que corresponde a la descripción de la rama de estudio

El campo 2 correspondiente al valor numérico summarizado de la columna Total se elimina

El campo 3 correspondiente al valor numérico summarizado de la columna Trabajando, se renombra por **trabajando**

El campo 4 correspondiente al valor numérico summarizado de la columna En Desempleo, se renombra por **desempleado**

El campo 5 correspondiente al valor numérico summarizado de la columna Inactivo, se renombra por **inactivo**

#	Fieldname	Rename to	Length	Precision
1	Total	rama		
2	50796.0	trabajando		
3	11441.0	desempleado		
4	3949.0	inactivo		

Luego de este paso, se aplican funciones de Add Constant a cada bloque de procesamiento, de manera que se incluyan como campos los valores correspondientes al Sexo y el Tipo de Universidad correspondientes a cada una de las secciones

The screenshot shows the 'Add constant values' step configuration in the Pentaho Data Integration interface. The step name is 'Add UnivPub&SM'. The 'Fields' table contains two rows:

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Value
1	tipo_universidad	String							Universidades Públicas
2	sexo	String							Mujeres

En el siguiente paso se aplica un step más de Add Constant, pero sobre la unión de los 4 bloques de datos, adicionando así el año de inserción laboral 2014 y el año de promoción de los egresados 2010 los cuales son comunes a todos los bloques de información

The screenshot shows the 'Add constant values' step configuration in the Pentaho Data Integration interface. The step name is 'Add Anio L&E'. The 'Fields' table contains two rows:

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Value
1	anio_laboral	Number	#		0				2014
2	anio_egresados	Number	#		0				2010

El conjunto de datos obtenido de las transformaciones anteriores, se persiste en un archivo físico insertados_laborales_2014.csv, el cual será la base de datos a cargar sobre la tabla de hecho física, previo a las respectivas operaciones de database lookup.

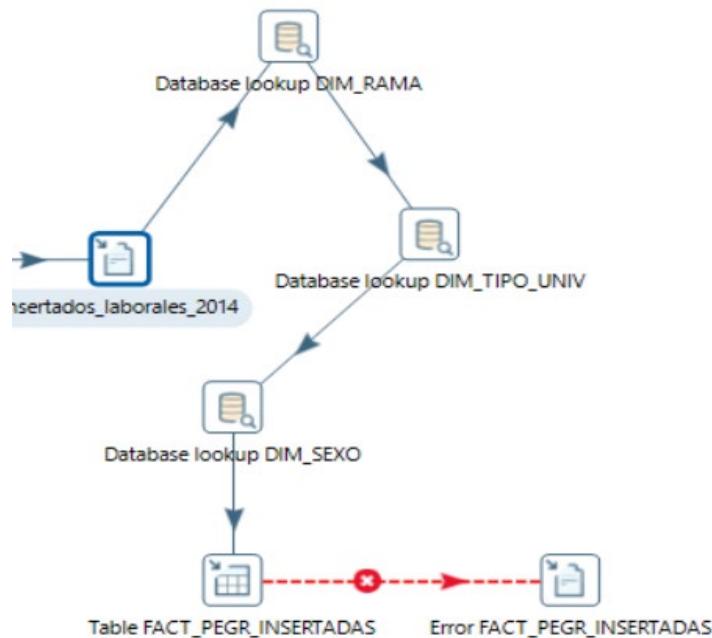
The screenshot shows the 'Text file output' step configuration in the Pentaho Data Integration interface. The step name is 'insertados_laborales_2014'. The 'Content' tab is selected. The 'File' section contains the following settings:

- Filename: F:\Pentaho\PRA2\fuentes\insertados_laborales_2014
- Run this as command instead?
- Pass output to servlet
- Create Parent folder
- Do not create file at start
- Accept file name from field?
- File name field: (empty)
- Extension: csv

Muestra del archivo csv obtenido

A	B	C	D	E	F	G	H
rama	trabajando	desempleado	inactivo	tipo_universidad	sexo	anio_laboral	anio_egresados
2 Ciencias sociales y jurídicas	20637	5323	1956	Universidades Públicas	Hombres	2014	2010
3 Ingeniería y arquitectura	21411	3570	1052	Universidades Públicas	Hombres	2014	2010
4 Artes y humanidades	2674	1122	458	Universidades Públicas	Hombres	2014	2010
5 Ciencias de la salud	3510	570	102	Universidades Públicas	Hombres	2014	2010
6 Ciencias	2565	856	380	Universidades Públicas	Hombres	2014	2010
7 Ciencias sociales y jurídicas	45108	12840	4304	Universidades Públicas	Mujeres	2014	2010
8 Ingeniería y arquitectura	8169	2051	464	Universidades Públicas	Mujeres	2014	2010
9 Artes y humanidades	5344	2116	881	Universidades Públicas	Mujeres	2014	2010
10 Ciencias de la salud	12210	2075	929	Universidades Públicas	Mujeres	2014	2010
11 Ciencias	4774	1529	431	Universidades Públicas	Mujeres	2014	2010

Siguiendo el mismo patrón de los procesos de carga anteriores, aplicamos las operaciones de database lookup sobre las dimensiones DIM_RAMA, DIM_TIPO_UNIV y DIM_SEXO para obtener la información de las llaves foráneas a poblar en la tabla de hechos, aplicar las mapeos necesarios sobre la tabla física FACT_PEGR_INSERTADAS y por último se establece el control de errores sobre un archivo de texto plano.



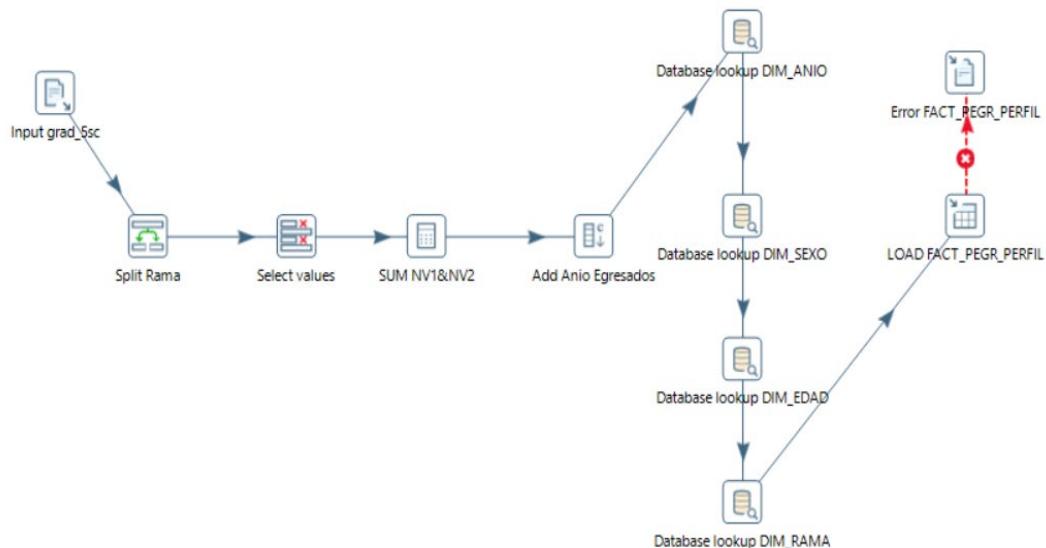
- FACT_PEGR_PERFIL: job encargado de ejecutar de forma individual el flujo de limpieza y carga de la tabla de hecho FACT_PEGR_PERFIL.



- ETL_FACT_PEGR_PERFIL: job encargado de limpiar la tabla FACT_PEGR_PERFIL y ejecutar la transformación LOAD_FACT_PEGR_PERFIL



- LOAD_FACT_PEGR_PERFIL: transformación encargada de extraer la información de las fuentes y carga de datos en tabla FACT_PEGR_PERFIL



Para la carga de la tabla de hecho, se utilizará la información contenida en el archivo grad_5sc.xls, el cual tiene la particularidad de que el código y la descripción de la rama de estudio están concatenada en el campo COD_AMBITO, por lo tanto, se aplica función de Split para obtener por separado cada una de las variables.

Field splitter

Step name: **Split Rama**

Field to split: **COD_AMBITO**

Delimiter: **-**

Enclosure:

#	New field	ID	Remove ID?	Type	Length	Precision	Format
1	CODIGO		N	Number			#
2	RAMA		N	String			

Como se había especificado en la sección de diseño, no se utilizarán valores '0' antecedidos del código de la rama, por lo tanto al aplicar el tipo de dato Number sobre el nuevo campo CODIGO, este es removido automáticamente y en el paso siguiente de Select Values, se cambia su tipo de datos a String para homologarlo como el valor esperado en el modelo físico de la tabla de hecho.

Select / Rename values

Step name: **Select values**

Select & Alter | Remove | Meta-data

Fields to alter the meta-data for :

#	Fieldname	Rename to	Type	Length	Precision	Binary to Normal?
1	CODIGO		String			N

En las especificaciones de diseño se indica que el valor total de egresados corresponderá a la suma de los campos NUM_EGR_NV1 y NUM_EGR_NV2, por lo que se aplica un step Calculator para obtener dicho valor.

Calculator

Step name: **SUM NV1&NV2**

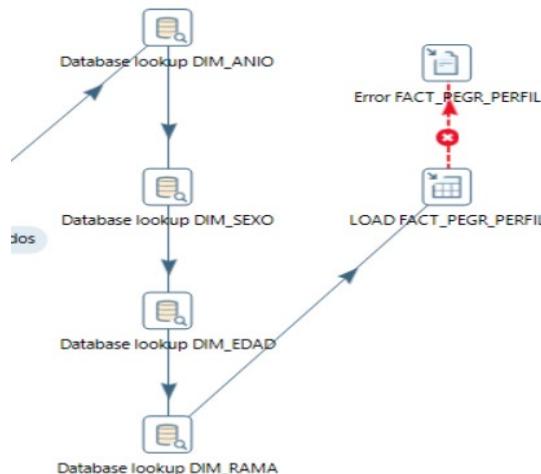
Fields:

#	New field	Calculation	Field A	Field B	Field C	Value type
1	sum_egresados	A + B	NUM_EGR_NV1	NUM_EGR_NV2		Number

Posteriormente, se adiciona el año de promoción de los egresados como una constante con valor “2016-2017”

Add constant values									
Step name Add Anio Egresados									
Fields :									
#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Value
1	anio	String							2016-2017

Siguiendo el mismo patrón de los procesos de carga anteriores, aplicamos las operaciones de database lookup sobre las dimensiones DIM_RAMA, DIM_ANIO, DIM_EDAD y DIM_SEXO para obtener la información de las llaves foráneas a poblar en la tabla de hechos, aplicar las mapeos necesarios sobre la tabla física FACT_PEGR_PERFIL y por último se establece el control de errores sobre un archivo de texto plano.



Nota: durante la fase de test para esta transformación, se identificó que el código de rama N4: 1029; *Saneamiento y seguridad laboral (otros estudios)*, no existe en la fuente de datos ISCED_2013.csv, que es el maestro de ramas de estudio utilizado. Por esta razón, se procedió a agregar manualmente dicho registro sobre el archivo ISCED_2013.csv y así garantizar la correcta carga de la totalidad de los datos.

4) Carga de los datos

Procedamos con la ejecución del job principal e identifiquemos el tiempo de ejecución total empleado para llevar a cabo la carga de datos inicial sobre todo el modelo físico



Luego de la ejecución, observamos que el tiempo total de carga oscila entre 8 y 10 segundos, a continuación una muestra del log de ejecución:

```
2019/12/17 23:12:38 - Spoon - Starting job...
2019/12/17 23:12:38 - CARGA_INICIAL_DWH_EGRESADOS - Start of job execution
2019/12/17 23:12:38 - CARGA_INICIAL_DWH_EGRESADOS - Starting entry [DELETE DIMS&FACTS]
2019/12/17 23:12:38 - CARGA_INICIAL_DWH_EGRESADOS - Starting entry [CARGA_DIM_STATIC]
2019/12/17 23:12:38 - CARGA_INICIAL_DWH_EGRESADOS - Starting entry [CARGA_DIMENSIONES]
2019/12/17 23:12:38 - CARGA_DIMENSIONES - Using run configuration [Pentaho local]
2019/12/17 23:12:38 - ETL_DIMENSIONES - Starting entry [DIM_RAMA]
2019/12/17 23:12:38 - DIM_RAMA - Using run configuration [Pentaho local]
2019/12/17 23:12:38 - ETL_DIM_RAMA - Starting entry [DELETE DIM_RAMA]
2019/12/17 23:12:38 - ETL_DIM_RAMA - Starting entry [LOAD_DIM_RAMA]
2019/12/17 23:12:38 - LOAD_DIM_RAMA - Loading transformation from repository [LOAD_DIM_RAMA] in directory []
2019/12/17 23:12:38 - LOAD_DIM_RAMA - Using run configuration [Pentaho local]
2019/12/17 23:12:38 - LOAD_DIM_RAMA - Using legacy execution engine
2019/12/17 23:12:38 - LOAD_DIM_RAMA - Dispatching started for transformation [LOAD_DIM_RAMA]
2019/12/17 23:12:39 - Table DIM_RAMA.0 - Connected to database [DWBD] (commit=1000)
2019/12/17 23:12:39 - Input Ramas Aprendizaje.0 - Header row skipped in file 'F:\Pentaho\PRA2\fuentes\ISCED_2013.csv'
2019/12/17 23:12:39 - Input Ramas Aprendizaje.0 - Finished processing (I=163, O=0, R=0, W=162, U=0, E=0)
2019/12/17 23:12:39 - Add sequence.0 - Finished processing (I=0, O=0, R=162, W=162, U=0, E=0)
2019/12/17 23:12:39 - Table DIM_RAMA.0 - Finished processing (I=0, O=162, R=162, W=162, U=0, E=0)
2019/12/17 23:12:39 - ETL_DIM_RAMA - Starting entry [Success]
2019/12/17 23:12:39 - ETL_DIM_RAMA - Finished job entry [Success] (result=[true])
2019/12/17 23:12:39 - ETL_DIM_RAMA - Finished job entry [LOAD_DIM_RAMA] (result=[true])
2019/12/17 23:12:39 - ETL_DIM_RAMA - Finished job entry [DELETE DIM_RAMA] (result=[true])
2019/12/17 23:12:39 - ETL_DIMENSIONES - Starting entry [Success DIM_RAMA]
2019/12/17 23:12:39 - ETL_DIMENSIONES - Finished job entry [Success DIM_RAMA] (result=[true])
2019/12/17 23:12:39 - ETL_DIMENSIONES - Finished job entry [DIM_RAMA] (result=[true])
2019/12/17 23:12:39 - ETL_DIMENSIONES - Starting entry [DIM_UNIVESIDAD]
2019/12/17 23:12:39 - DIM_UNIVESIDAD - Using run configuration [Pentaho local]
2019/12/17 23:12:39 - ETL_DIM_UNIVERSIDAD - Starting entry [DELETE DIM_UNIVERSIDAD]
2019/12/17 23:12:39 - ETL_DIM_UNIVERSIDAD - Starting entry [LOAD_DIM_UNIVERSIDAD]
2019/12/17 23:12:39 - LOAD_DIM_UNIVERSIDAD - Loading transformation from repository [LOAD_DIM_UNIVERSIDAD] in directory []
2019/12/17 23:12:39 - LOAD_DIM_UNIVERSIDAD - Using run configuration [Pentaho local]
2019/12/17 23:12:39 - LOAD_DIM_UNIVERSIDAD - Using legacy execution engine
2019/12/17 23:12:39 - LOAD_DIM_UNIVERSIDAD - Dispatching started for transformation [LOAD_DIM_UNIVERSIDAD]
2019/12/17 23:12:39 - Table DIM_UNIVERSIDAD.0 - Connected to database [DWBD] (commit=1000)
```

```
2019/12/17 23:12:39 - Input Universidad Privada.0 - Header row skipped in file 'F:\Pentaho\PRA2\fuentes\SEGR1.csv'
2019/12/17 23:12:39 - Input Universidad Privada.0 - Finished processing (I=161, O=0, R=0, W=160, U=0, E=0)
2019/12/17 23:12:39 - Input Universidad Publica.0 - Header row skipped in file 'F:\Pentaho\PRA2\fuentes\SEGR2.csv'
2019/12/17 23:12:39 - Input Universidad Publica.0 - Finished processing (I=251, O=0, R=0, W=250, U=0, E=0)
2019/12/17 23:12:39 - Sort rows.0 - Finished processing (I=0, O=0, R=410, W=410, U=0, E=0)
2019/12/17 23:12:39 - Unique rows.0 - Finished processing (I=0, O=0, R=410, W=82, U=0, E=0)
2019/12/17 23:12:39 - Database lookup Modalidad.0 - Finished processing (I=82, O=0, R=82, W=82, U=0, E=0)
2019/12/17 23:12:39 - Add sequence.0 - Finished processing (I=0, O=0, R=82, W=82, U=0, E=0)
2019/12/17 23:12:39 - Table DIM_UNIVERSIDAD.0 - Finished processing (I=0, O=82, R=82, W=82, U=0, E=0)
2019/12/17 23:12:39 - ETL_DIM_UNIVERSIDAD - Starting entry [Success]
2019/12/17 23:12:39 - ETL_DIM_UNIVERSIDAD - Finished job entry [Success] (result=[true])
2019/12/17 23:12:39 - ETL_DIM_UNIVERSIDAD - Finished job entry [LOAD_DIM_UNIVERSIDAD] (result=[true])
2019/12/17 23:12:39 - ETL_DIM_UNIVERSIDAD - Finished job entry [DELETE DIM_UNIVERSIDAD] (result=[true])
2019/12/17 23:12:39 - ETL_DIMENSIONES - Starting entry [Success DIM_UNIVERSIDAD]
2019/12/17 23:12:39 - ETL_DIMENSIONES - Finished job entry [Success DIM_UNIVERSIDAD] (result=[true])
2019/12/17 23:12:39 - ETL_DIMENSIONES - Finished job entry [DIM_UNIVESIDAD] (result=[true])
2019/12/17 23:12:39 - CARGA_INICIAL_DWH_EGRESADOS - Starting entry [CARGA_HECHOS]
2019/12/17 23:12:39 - CARGA_HECHOS - Using run configuration [Pentaho local]
2019/12/17 23:12:39 - ETL_HECHOS - Starting entry [FACT_PEGR_COMPARATIVA]
2019/12/17 23:12:39 - FACT_PEGR_COMPARATIVA - Using run configuration [Pentaho local]
2019/12/17 23:12:39 - ETL_FACT_PEGR_COMPARATIVA - Starting entry [DELETE FACT_PEGR_COMPARATIVA]
2019/12/17 23:12:39 - ETL_FACT_PEGR_COMPARATIVA - Starting entry [LOAD FACT_PEGR_COMPARATIVA]
2019/12/17 23:12:39 - LOAD FACT_PEGR_COMPARATIVA - Loading transformation from repository [LOAD_FACT_PEGR_COMPARATIVA] in directory []
2019/12/17 23:12:39 - LOAD FACT_PEGR_COMPARATIVA - Using run configuration [Pentaho local]
2019/12/17 23:12:39 - LOAD FACT_PEGR_COMPARATIVA - Using legacy execution engine
2019/12/17 23:12:39 - LOAD_FACT_PEGR_COMPARATIVA - Dispatching started for transformation [LOAD_FACT_PEGR_COMPARATIVA]
2019/12/17 23:12:39 - Table FACT_PEGR_COMPARATIVA.0 - Connected to database [DWBD] (commit=1000)
2019/12/17 23:12:42 - Input Egresados Jovenes.0 - Finished processing (I=12, O=0, R=0, W=12, U=0, E=0)
2019/12/17 23:12:42 - Input Egresados Pais.0 - Finished processing (I=12, O=0, R=0, W=12, U=0, E=0)
2019/12/17 23:12:42 - Row Normaliser Egresados Jovenes.0 - Finished processing (I=0, O=0, R=12, W=60, U=0, E=0)
2019/12/17 23:12:42 - Row Normaliser Egresados Pais.0 - Finished processing (I=0, O=0, R=12, W=60, U=0, E=0)
2019/12/17 23:12:42 - Value Mapper Egresados Jovenes.0 - Finished processing (I=0, O=0, R=60, W=60, U=0, E=0)
2019/12/17 23:12:42 - Value Mapper Egresados Pais.0 - Finished processing (I=0, O=0, R=60, W=60, U=0, E=0)
2019/12/17 23:12:42 - Sort rows Egresados Jovenes.0 - Finished processing (I=0, O=0, R=60, W=60, U=0, E=0)
2019/12/17 23:12:42 - Sort rows Egresados Pais.0 - Finished processing (I=0, O=0, R=60, W=60, U=0, E=0)
2019/12/17 23:12:43 - Merge Join.0 - Finished processing (I=0, O=0, R=120, W=60, U=0, E=0)
2019/12/17 23:12:43 - Select values.0 - Finished processing (I=0, O=0, R=60, W=60, U=0, E=0)
2019/12/17 23:12:43 - Database lookup DIM_PAIS.0 - Finished processing (I=60, O=0, R=60, W=60, U=0, E=0)
2019/12/17 23:12:43 - Table FACT_PEGR_COMPARATIVA.0 - Finished processing (I=0, O=60, R=60, W=60, U=0, E=0)
2019/12/17 23:12:43 - ETL_FACT_PEGR_COMPARATIVA - Starting entry [Success]
2019/12/17 23:12:43 - ETL_FACT_PEGR_COMPARATIVA - Finished job entry [Success] (result=[true])
2019/12/17 23:12:43 - ETL_FACT_PEGR_COMPARATIVA - Finished job entry [LOAD FACT_PEGR_COMPARATIVA] (result=[true])
2019/12/17 23:12:43 - ETL_FACT_PEGR_COMPARATIVA - Finished job entry [DELETE FACT_PEGR_COMPARATIVA] (result=[true])
2019/12/17 23:12:43 - ETL_HECHOS - Starting entry [Success FACT_PEGR_COMPARATIVA]
2019/12/17 23:12:43 - ETL_HECHOS - Finished job entry [Success FACT_PEGR_COMPARATIVA] (result=[true])
2019/12/17 23:12:43 - ETL_HECHOS - Finished job entry [FACT_PEGR_COMPARATIVA] (result=[true])
2019/12/17 23:12:43 - ETL_HECHOS - Starting entry [FACT_PEGR_EVOLUTIVO]
2019/12/17 23:12:43 - FACT_PEGR_EVOLUTIVO - Using run configuration [Pentaho local]
2019/12/17 23:12:43 - ETL_FACT_PEGR_EVOLUTIVO - Starting entry [DELETE FACT_PEGR_EVOLUTIVO]
```

2019/12/17 23:12:43 - ETL_FACT_PEGR_EVOLUTIVO - Starting entry [LOAD_FACT_PEGR_EVOLUTIVO]
2019/12/17 23:12:43 - LOAD_FACT_PEGR_EVOLUTIVO - Loading transformation from repository [LOAD_FACT_PEGR_EVOLUTIVO] in directory [/]
2019/12/17 23:12:43 - LOAD_FACT_PEGR_EVOLUTIVO - Using run configuration [Pentaho local]
2019/12/17 23:12:43 - LOAD_FACT_PEGR_EVOLUTIVO - Using legacy execution engine
2019/12/17 23:12:43 - LOAD_FACT_PEGR_EVOLUTIVO - Dispatching started for transformation [LOAD_FACT_PEGR_EVOLUTIVO]
2019/12/17 23:12:43 - LOAD FACT_PEGR_EVOLUTIVO.0 - Connected to database [DWBD] (commit=1000)
2019/12/17 23:12:43 - Input Universidad Publica.0 - Header row skipped in file 'F:\Pentaho\PRA2\fuentes\SEGR2.csv'
2019/12/17 23:12:43 - Input Universidad Publica.0 - Finished processing (I=251, O=0, R=0, W=250, U=0, E=0)
2019/12/17 23:12:43 - Input Universidad Privada.0 - Header row skipped in file 'F:\Pentaho\PRA2\fuentes\SEGR1.csv'
2019/12/17 23:12:43 - Input Universidad Privada.0 - Finished processing (I=161, O=0, R=0, W=160, U=0, E=0)
2019/12/17 23:12:43 - Universidades.0 - Finished processing (I=0, O=411, R=410, W=410, U=0, E=0)
2019/12/17 23:12:43 - Row Normaliser Universidades.0 - Finished processing (I=0, O=0, R=410, W=3280, U=0, E=0)
2019/12/17 23:12:44 - Database lookup DIM_UNIVERSIDAD.0 - Finished processing (I=3280, O=0, R=3280, W=3280, U=0, E=0)
2019/12/17 23:12:44 - Database lookup DIM_RAMA.0 - Finished processing (I=3280, O=0, R=3280, W=3280, U=0, E=0)
2019/12/17 23:12:44 - Database lookup DIM_TIPO_UNIV.0 - Finished processing (I=3280, O=0, R=3280, W=3280, U=0, E=0)
2019/12/17 23:12:44 - Database lookup DIM_ANIO.0 - Finished processing (I=3280, O=0, R=3280, W=3280, U=0, E=0)
2019/12/17 23:12:44 - LOAD FACT_PEGR_EVOLUTIVO.0 - Finished processing (I=0, O=3280, R=3280, W=3280, U=0, E=0)
2019/12/17 23:12:44 - ETL_FACT_PEGR_EVOLUTIVO - Starting entry [Success]
2019/12/17 23:12:44 - ETL_FACT_PEGR_EVOLUTIVO - Finished job entry [Success] (result=[true])
2019/12/17 23:12:44 - ETL_FACT_PEGR_EVOLUTIVO - Finished job entry [LOAD_FACT_PEGR_EVOLUTIVO] (result=[true])
2019/12/17 23:12:44 - ETL_FACT_PEGR_EVOLUTIVO - Finished job entry [DELETE FACT_PEGR_EVOLUTIVO] (result=[true])
2019/12/17 23:12:44 - ETL_HECHOS - Starting entry [Success FACT_PEGR_EVOLUTIVO]
2019/12/17 23:12:44 - ETL_HECHOS - Finished job entry [Success FACT_PEGR_EVOLUTIVO] (result=[true])
2019/12/17 23:12:44 - ETL_HECHOS - Finished job entry [FACT_PEGR_EVOLUTIVO] (result=[true])
2019/12/17 23:12:44 - ETL_HECHOS - Starting entry [FACT_PEGR_INSERTADAS]
2019/12/17 23:12:44 - FACT_PEGR_INSERTADAS - Using run configuration [Pentaho local]
2019/12/17 23:12:44 - ETL_FACT_PEGR_INSERTADAS - Starting entry [DELETE FACT_PEGR_INSERTADAS]
2019/12/17 23:12:44 - ETL_FACT_PEGR_INSERTADAS - Starting entry [LOAD FACT_PEGR_INSERTADAS]
2019/12/17 23:12:44 - LOAD FACT_PEGR_INSERTADAS - Loading transformation from repository [LOAD_FACT_PEGR_INSERTADAS] in directory [/]
2019/12/17 23:12:45 - LOAD FACT_PEGR_INSERTADAS - Using run configuration [Pentaho local]
2019/12/17 23:12:45 - LOAD FACT_PEGR_INSERTADAS - Using legacy execution engine
2019/12/17 23:12:45 - LOAD_FACT_PEGR_INSERTADAS - Dispatching started for transformation [LOAD_FACT_PEGR_INSERTADAS]
2019/12/17 23:12:45 - Table FACT_PEGR_INSERTADAS.0 - Connected to database [DWBD] (commit=1000)
2019/12/17 23:12:45 - Input Rama UnivPriv&SM.0 - Finished processing (I=5, O=0, R=0, W=5, U=0, E=0)
2019/12/17 23:12:45 - Input Rama UnivPub&SH.0 - Finished processing (I=5, O=0, R=0, W=5, U=0, E=0)
2019/12/17 23:12:45 - Input Rama UnivPriv&SH.0 - Finished processing (I=5, O=0, R=0, W=5, U=0, E=0)
2019/12/17 23:12:45 - RRTDI UnivPub&SH.0 - Finished processing (I=0, O=0, R=5, W=5, U=0, E=0)
2019/12/17 23:12:45 - RRTDI UnivPriv&SH.0 - Finished processing (I=0, O=0, R=5, W=5, U=0, E=0)
2019/12/17 23:12:45 - RRTDI UnivPriv&SM.0 - Finished processing (I=0, O=0, R=5, W=5, U=0, E=0)
2019/12/17 23:12:45 - Add UnivPub&SH.0 - Finished processing (I=0, O=0, R=5, W=5, U=0, E=0)
2019/12/17 23:12:45 - Add UnivPriv&SH.0 - Finished processing (I=0, O=0, R=5, W=5, U=0, E=0)
2019/12/17 23:12:45 - Add UnivPriv&SM.0 - Finished processing (I=0, O=0, R=5, W=5, U=0, E=0)
2019/12/17 23:12:45 - Add Anio L&E.0 - Finished processing (I=0, O=0, R=15, W=15, U=0, E=0)
2019/12/17 23:12:45 - insertados_laborales_2014.0 - Finished processing (I=0, O=16, R=15, W=15, U=0, E=0)
2019/12/17 23:12:45 - Database lookup DIM_RAMA.0 - Finished processing (I=15, O=0, R=15, W=15, U=0, E=0)
2019/12/17 23:12:45 - Database lookup DIM_TIPO_UNIV.0 - Finished processing (I=15, O=0, R=15, W=15, U=0, E=0)

2019/12/17 23:12:45 - Database lookup DIM_SEXO.0 - Finished processing (I=15, O=0, R=15, W=15, U=0, E=0)
2019/12/17 23:12:45 - Table FACT_PEGR_INSERTADAS.0 - Finished processing (I=0, O=15, R=15, W=15, U=0, E=0)
2019/12/17 23:12:45 - ETL_FACT_PEGR_INSERTADAS - Starting entry [Success]
2019/12/17 23:12:45 - ETL_FACT_PEGR_INSERTADAS - Finished job entry [Success] (result=[true])
2019/12/17 23:12:45 - ETL_FACT_PEGR_INSERTADAS - Finished job entry [LOAD FACT_PEGR_INSERTADAS] (result=[true])
2019/12/17 23:12:45 - ETL_FACT_PEGR_INSERTADAS - Finished job entry [DELETE FACT_PEGR_INSERTADAS] (result=[true])
2019/12/17 23:12:45 - ETL_HECHOS - Starting entry [Success FACT_PEGR_INSERTADAS]
2019/12/17 23:12:45 - ETL_HECHOS - Finished job entry [Success FACT_PEGR_INSERTADAS] (result=[true])
2019/12/17 23:12:45 - ETL_HECHOS - Finished job entry [FACT_PEGR_INSERTADAS] (result=[true])
2019/12/17 23:12:45 - ETL_HECHOS - Starting entry [FACT_PEGR_PERFIL]
2019/12/17 23:12:45 - FACT_PEGR_PERFIL - Using run configuration [Pentaho local]
2019/12/17 23:12:45 - ETL_FACT_PEGR_PERFIL - Starting entry [DELETE FACT_PEGR_PERFIL]
2019/12/17 23:12:45 - ETL_FACT_PEGR_PERFIL - Starting entry [LOAD FACT_PEGR_PERFIL]
2019/12/17 23:12:45 - LOAD FACT_PEGR_PERFIL - Loading transformation from repository [LOAD_FACT_PEGR_PERFIL] in directory []
2019/12/17 23:12:45 - LOAD FACT_PEGR_PERFIL - Using run configuration [Pentaho local]
2019/12/17 23:12:45 - LOAD FACT_PEGR_PERFIL - Using legacy execution engine
2019/12/17 23:12:45 - LOAD FACT_PEGR_PERFIL - Dispatching started for transformation [LOAD_FACT_PEGR_PERFIL]
2019/12/17 23:12:45 - LOAD FACT_PEGR_PERFIL.0 - Connected to database [DWBD] (commit=1000)
2019/12/17 23:12:45 - Input grad_5sc.0 - Header row skipped in file 'F:\Pentaho\PRA2\fuentes\grad_5sc.csv'
2019/12/17 23:12:45 - Input grad_5sc.0 - Finished processing (I=713, O=0, R=0, W=712, U=0, E=0)
2019/12/17 23:12:45 - Split Rama.0 - Finished processing (I=0, O=0, R=712, W=712, U=0, E=0)
2019/12/17 23:12:45 - Select values.0 - Finished processing (I=0, O=0, R=712, W=712, U=0, E=0)
2019/12/17 23:12:45 - SUM NV1&NV2.0 - Finished processing (I=0, O=0, R=712, W=712, U=0, E=0)
2019/12/17 23:12:45 - Add Anio Egresados.0 - Finished processing (I=0, O=0, R=712, W=712, U=0, E=0)
2019/12/17 23:12:46 - Database lookup DIM_ANIO.0 - Finished processing (I=712, O=0, R=712, W=712, U=0, E=0)
2019/12/17 23:12:46 - Database lookup DIM_SEXO.0 - Finished processing (I=712, O=0, R=712, W=712, U=0, E=0)
2019/12/17 23:12:46 - Database lookup DIM_EDAD.0 - Finished processing (I=712, O=0, R=712, W=712, U=0, E=0)
2019/12/17 23:12:46 - Database lookup DIM_RAMA.0 - Finished processing (I=712, O=0, R=712, W=712, U=0, E=0)
2019/12/17 23:12:46 - LOAD FACT_PEGR_PERFIL.0 - Finished processing (I=0, O=712, R=712, W=712, U=0, E=0)
2019/12/17 23:12:46 - ETL_FACT_PEGR_PERFIL - Starting entry [Success]
2019/12/17 23:12:46 - ETL_FACT_PEGR_PERFIL - Finished job entry [Success] (result=[true])
2019/12/17 23:12:46 - ETL_FACT_PEGR_PERFIL - Finished job entry [LOAD FACT_PEGR_PERFIL] (result=[true])
2019/12/17 23:12:46 - ETL_FACT_PEGR_PERFIL - Finished job entry [DELETE FACT_PEGR_PERFIL] (result=[true])
2019/12/17 23:12:46 - ETL_HECHOS - Starting entry [Success FACT_PEGR_PERFIL]
2019/12/17 23:12:46 - ETL_HECHOS - Finished job entry [Success FACT_PEGR_PERFIL] (result=[true])
2019/12/17 23:12:46 - ETL_HECHOS - Finished job entry [FACT_PEGR_PERFIL] (result=[true])
2019/12/17 23:12:46 - CARGA_INICIAL_DWH_EGRESADOS - Starting entry [Success]
2019/12/17 23:12:46 - CARGA_INICIAL_DWH_EGRESADOS - Finished job entry [Success] (result=[true])
2019/12/17 23:12:46 - CARGA_INICIAL_DWH_EGRESADOS - Finished job entry [CARGA_HECHOS] (result=[true])
2019/12/17 23:12:46 - CARGA_INICIAL_DWH_EGRESADOS - Finished job entry [CARGA_DIMENSIONES] (result=[true])
2019/12/17 23:12:46 - CARGA_INICIAL_DWH_EGRESADOS - Finished job entry [CARGA_DIM_STATIC] (result=[true])
2019/12/17 23:12:46 - CARGA_INICIAL_DWH_EGRESADOS - Finished job entry [DELETE DIMS&FACTS] (result=[true])
2019/12/17 23:12:46 - CARGA_INICIAL_DWH_EGRESADOS - Job execution finished
2019/12/17 23:12:46 - Spoon - Job has ended.

Además podemos observar que en el directorio de errores, todos los archivos de salida tienen 0 bytes, por lo tanto el proceso finalizo OK

Este equipo > UserRedir (F:) > Pentaho > PRA2 > error				
	Nombre	Fecha de modifica...	Tipo	Tamaño
★ Acceso rápido	load_DIM_RAMА	17/12/2019 23:21	Documento de tex...	0 KB
Este equipo	load_DIM_UNIVERSIDAD	17/12/2019 23:21	Documento de tex...	0 KB
3D Objects	load_FACT_PEGR_COMPARATIVA	17/12/2019 23:21	Documento de tex...	0 KB
Descargas	load_FACT_PEGR_EVOLUTIVO	17/12/2019 23:21	Documento de tex...	0 KB
Disco local (C: en A)	load_FACT_PEGR_INSERTADAS	17/12/2019 23:21	Documento de tex...	0 KB
Disco local (D: en A)	load_FACT_PEGR_PERFIL	17/12/2019 23:21	Documento de tex...	0 KB
Documentos	load_FACT_PEGR_PERFIL_firstload	15/12/2019 18:38	Documento de tex...	1 KB
Escrivania				

Comprobación de carga de datos

Veamos los resultados del proceso de carga en algunas tablas físicas del modelo de datos para el DW

FACT_PEGR_EVOLUTIVO

```
SQLQuery24.sql - U...ENT_jboteros (53) # X
/*
***** Script for SelectTopNRows command from SSMS *****/
SELECT TOP (1000) [sk_dim_anio]
,[sk_dim_tipo_univ]
,[sk_dim_universidad]
,[cod_rama]
,[personas_egresadas]
FROM [DW_DB_jboteros].[dbo].[FACT_PEGR_EVOLUTIVO]
```

	sk_dim_anio	sk_dim_tipo_univ	sk_dim_universidad	cod_rama	personas_egresadas
1	2010	1	1	1	98
2	2010	1	1	35	1566
3	2010	1	1	68	127
4	2010	1	1	85	1224
5	2010	1	1	132	371
6	2010	1	4	1	278
7	2010	1	4	35	1968
8	2010	1	4	68	356
9	2010	1	4	85	907
10	2010	1	4	132	539

Query executed successfully. | UCS1R1UOCSQL01 (14.0 RTM) | STUDENT_jboteros (53) | DW_DB_jboteros | 00:00:00 | 1000 rows

DIM_UNIVERSIDAD

SQLQuery25.sql - U...ENT_jboteros (59) ⇐ X SQLQuery24.sql - U...ENT_jboteros (53)

```
***** Script for SelectTopNRows command from SSMS *****/
SELECT TOP (1000) [sk_dim_universidad]
      ,[desc_universidad]
      ,[sk_dim_modalidad]
   FROM [DW_DB_jboteros].[dbo].[DIM_UNIVERSIDAD]
```

100 %

	sk_dim_universidad	desc_universidad	sk_dim_modalidad
1	1	A Coruña	1
2	2	A Distancia de Madrid	2
3	3	Abat Oliba CEU	1
4	4	Alcalá	1
5	5	Alfonso X El Sabio	1
6	6	Alicante	1
7	7	Almería	1
8	8	Antonio de Nebrija	1
9	9	Autónoma de Barcelona	1
10	10	Autónoma de Madrid	1

Query executed successfully. | UCS1R1UOCSQL01 (14.0 RTM) | STUDENT_jboteros (59) | DW_DB_jboteros | 00:00:00 | 82 rows

FACT_PEGR_INSERTADAS

SQLQuery26.sql - U...ENT_jboteros (64) ⇐ X SQLQuery25.sql - U...ENT_jboteros (59) SQLQuery24.sql - U...ENT_jboteros (53)

```
***** Script for SelectTopNRows command from SSMS *****/
SELECT TOP (1000) [sk_dim_anio]
      ,[sk_dim_tipo_univ]
      ,[sk_dim_sexo]
      ,[cod_rama]
      ,[pegr_trabajando]
      ,[pegr_desempleados]
      ,[pegr_inactivos]
      ,[sk_dim_anio_laboral]
   FROM [DW_DB_jboteros].[dbo].[FACT_PEGR_INSERTADAS]
```

100 %

	sk_dim_anio	sk_dim_tipo_univ	sk_dim_sexo	cod_rama	pegr_trabajando	pegr_desempleados	pegr_inactivos	sk_dim_anio_laboral
1	2010	1	1	1	2674	1122	458	2014
2	2010	1	1	35	20637	5323	1956	2014
3	2010	1	1	68	2565	856	380	2014
4	2010	1	1	85	21411	3570	1052	2014
5	2010	1	1	132	3510	570	102	2014
6	2010	1	2	1	5344	2116	881	2014
7	2010	1	2	35	45108	12840	4304	2014
8	2010	1	2	68	4774	1529	431	2014
9	2010	1	2	85	8169	2051	464	2014
10	2010	1	2	132	12210	2075	929	2014

Query executed successfully. | UCS1R1UOCSQL01 (14.0 RTM) | STUDENT_jboteros (64) | DW_DB_jboteros | 00:00:00 | 20 rows

DIM_RAMA

SQLQuery27.sql - U...ENT_jboteros (69) ▶ X SQLQuery26.sql - U...ENT_jboteros (64)) SQLQuery25.sql - U...ENT_jboteros (59))

```
***** Script for SelectTopNRows command from SSMS *****/
SELECT TOP (1000) [sk_dim_rama]
    ,[cod_rama]
    ,[desc_rama]
    ,[cod_rama_n2]
    ,[desc_rama_n2]
    ,[cod_rama_n3]
    ,[desc_rama_n3]
    ,[cod_rama_n4]
    ,[desc_rama_n4]
    ,[cod_rama_n5]
    ,[desc_rama_n5]
FROM [DW_DB_jboteros].[dbo].[DIM_RAMA]
```

Results Messages

	sk_dim_rama	cod_rama	desc_rama	cod_rama_n2	desc_rama_n2	cod_rama_n3	desc_rama_n3	cod_rama_n4	desc_rama_n4
1	1	1	Artes y humanidades	1	Educación	11	Educación	111	Ciencias de la educación
2	2	1	Artes y humanidades	1	Educación	11	Educación	111	Ciencias de la educación
3	3	1	Artes y humanidades	1	Educación	11	Educación	112	Formación de docentes de
4	4	1	Artes y humanidades	1	Educación	11	Educación	113	Formación de docentes de
5	5	1	Artes y humanidades	1	Educación	11	Educación	114	Formación de docentes de
6	6	1	Artes y humanidades	1	Educación	11	Educación	114	Formación de docentes de
7	7	1	Artes y humanidades	1	Educación	11	Educación	114	Formación de docentes de
8	8	1	Artes y humanidades	1	Educación	11	Educación	119	Educación (Otros estudios)
9	9	1	Artes y humanidades	1	Educación	11	Educación	119	Educación (Otros estudios)

Query executed successfully. | UCS1R1UOCSQL01 (14.0 RTM) | STUDENT_jboteros (69) | DW_DB_jboteros | 00:00:00 | 162 rows

FACT_PEGR_PERFIL

SQLQuery28.sql - U...ENT_jboteros (71) ▶ X SQLQuery27.sql - U...ENT_jboteros (69)) SQLQuery26.sql - U...ENT_jboteros (64))

```
***** Script for SelectTopNRows command from SSMS *****/
SELECT TOP (1000) [sk_dim_anio]
    ,[sk_dim_sexo]
    ,[sk_dim_edad]
    ,[cod_rama]
    ,[personas_egresadas]
FROM [DW_DB_jboteros].[dbo].[FACT_PEGR_PERFIL]
```

Results Messages

	sk_dim_anio	sk_dim_sexo	sk_dim_edad	cod_rama	personas_egresadas
1	2017	1	1	1	320
2	2017	1	1	3	292
3	2017	1	1	4	2784
4	2017	1	1	5	2480
5	2017	1	1	8	203
6	2017	1	1	10	1035
7	2017	1	1	11	140
8	2017	1	1	12	588
9	2017	1	1	14	27
10	2017	1	1	15	167

Query executed successfully. | UCS1R1UOCSQL01 (14.0 RTM) | STUDENT_jboteros (71) | DW_DB_jboteros | 00:00:00 | 712 rows

Explotación de datos

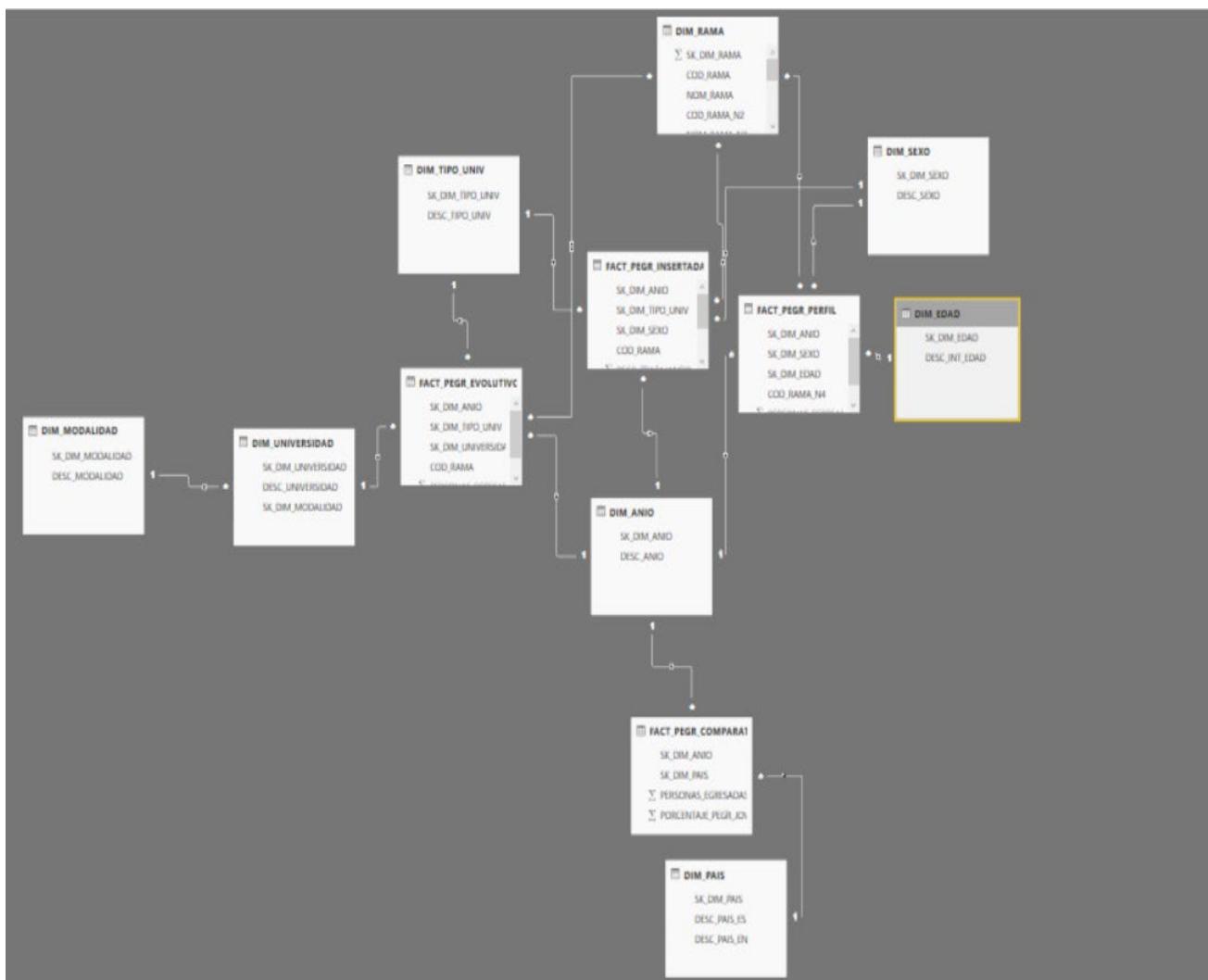
El último paso del proyecto, consiste en diseñar e implementar un modelo OLAP que permita, mediante análisis multidimensionales, responder a las preguntas planteadas en la toma de requerimientos

A continuación, se enumeran las necesidades globales planteadas en la fase de especificación de requerimientos:

- Conocer la evolución temporal del número de egresados en el sistema educativo universitario.
- Esta evolución debe poderse analizar desde diferentes perspectivas:
 - Tipo de Universidad (P.Ej: Universidades Privadas).
 - Modalidad.(P.Ej: No Presencial)
 - Universidad (P.Ej: Oberta de Catalunya).
 - Rama de enseñanza. (P.Ej: Ciencias Sociales y Jurídicas)
 - Ámbito de Estudio. (P.Ej: Ciencias de la educación)
- Conocer el perfil de los estudiantes egresados en el curso académico 2016-2017, en términos de características personales como sexo y edad.
- Analizar la incorporación de los graduados universitarios del curso 2009-2010 al mercado laboral en 2014.
- Realizar la comparativa entre egresados universitarios en España y otros países.

Se toma como base para la construcción de los cubos, el modelo físico con sus datos, entregado en los insumos para la práctica final, export_dw_egr.sql. Para nuestro caso se excluyen las tablas de staging ya que nos son requeridas en esta fase de explotación de los datos.

Modelo de vista relacional obtenido en PowerBI Desktop para la construcción de los cubos, se agregan las relaciones desde FACT_PEGR_PERFIL y FACT_PEGR_INSERTADAS hacia DIM_RAMA dado que en el script del modelo físico no se encontraban especificadas las llaves foráneas

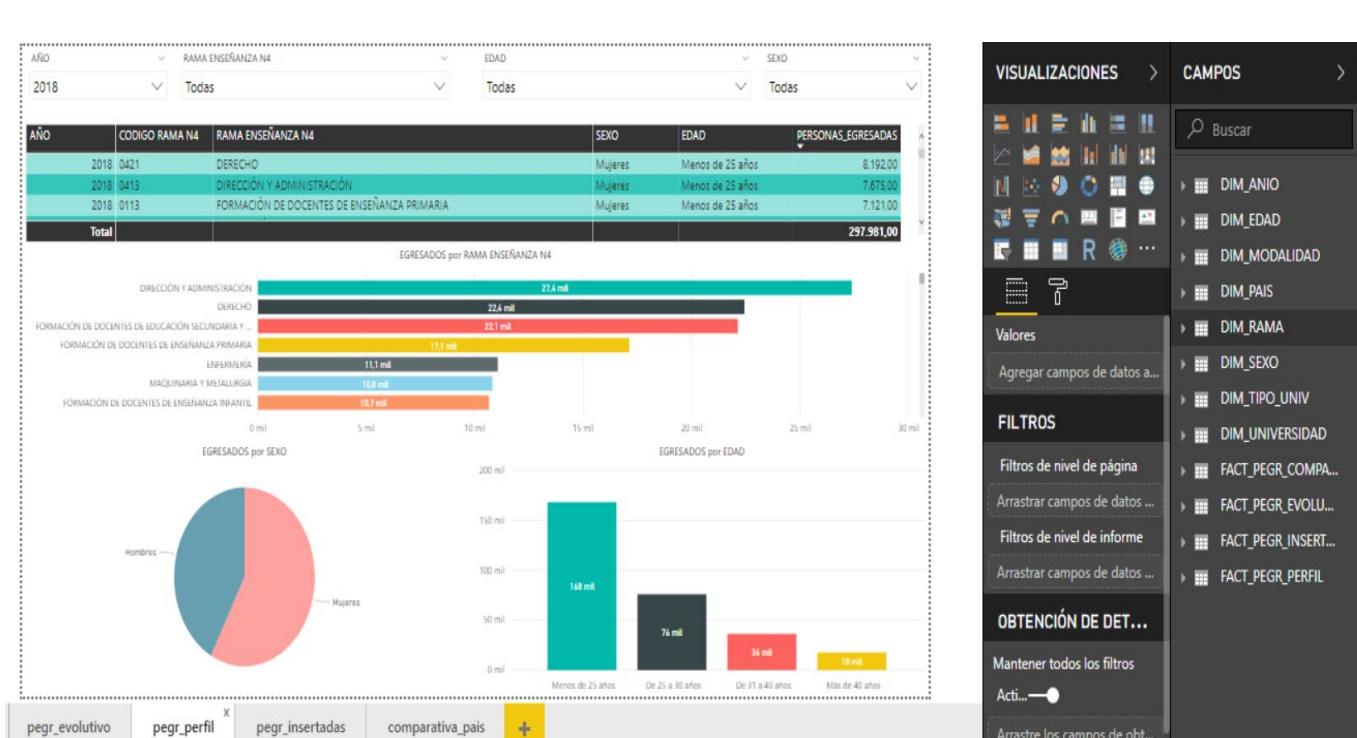
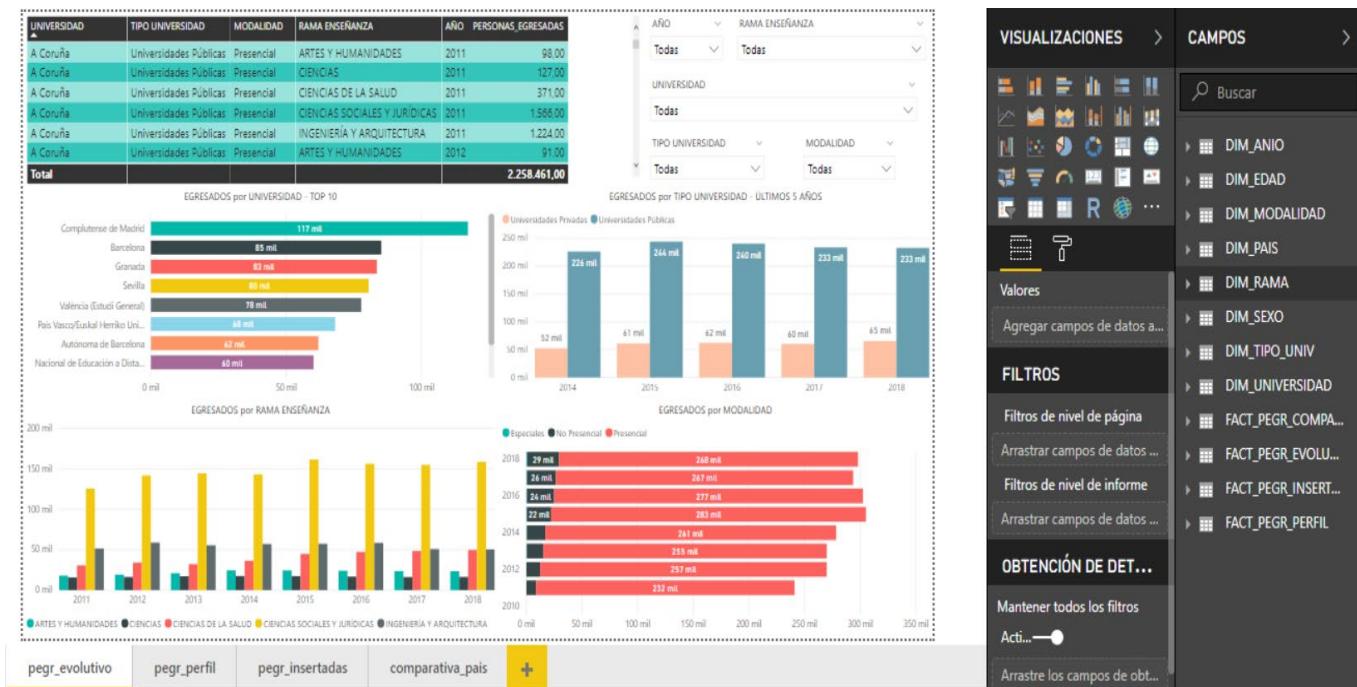


La implementación del modelo OLAP se realizó a través de la herramienta Power BI Desktop, configurando una fuente de datos hacia el modelo relacional contenido en Microsoft SQL Server, obteniendo así un total de 4 vistas, cada una de las cuales hace énfasis a los respectivos **hechos** de interés para nuestro análisis:

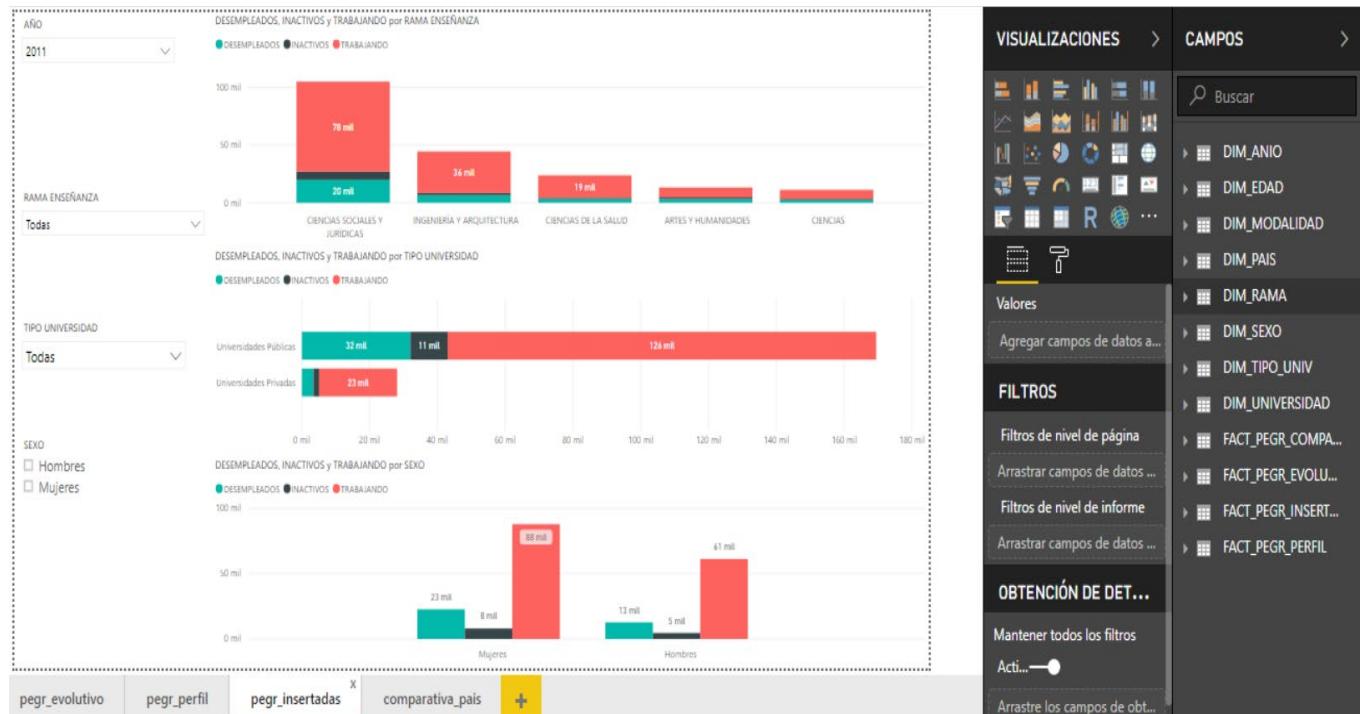
Los informes pueden ser accedidos públicamente en la siguiente URL:

<https://app.powerbi.com/view?r=eyJrljoiZTY1ZGEwOTUtMTEyNC00YzQzLTIIzDEtZjU1NzBmMGFhNzE1IiwidCI6ImFIYzc2MmU0LTNkNTQtNDk1ZS1hOGZLTQyODdkY2U2ZmU2OSIsImMiOjh9>

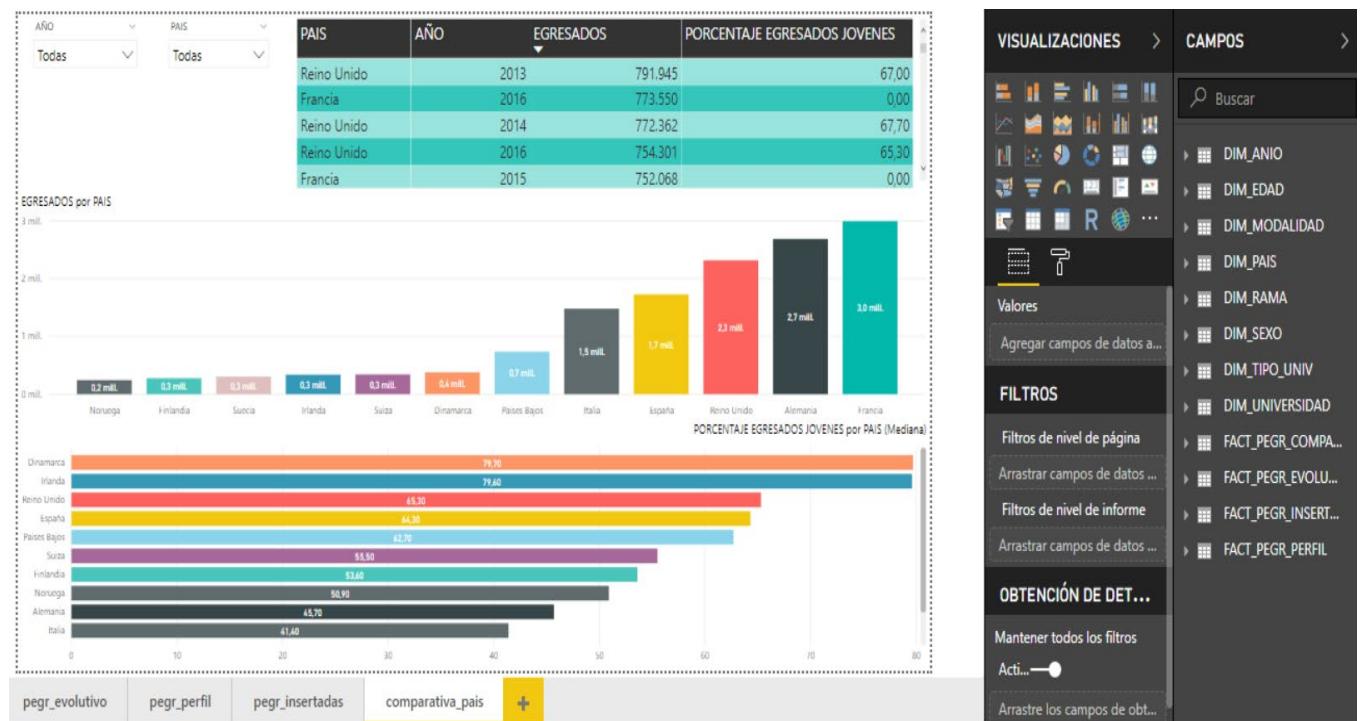
EGRESADOS EVOLUTIVO



EGRESADOS INSERTADOS



EGRESADOS COMPARATIVA



De forma específica, se pide que el sistema debe como mínimo ser capaz de dar respuesta a las siguientes preguntas:

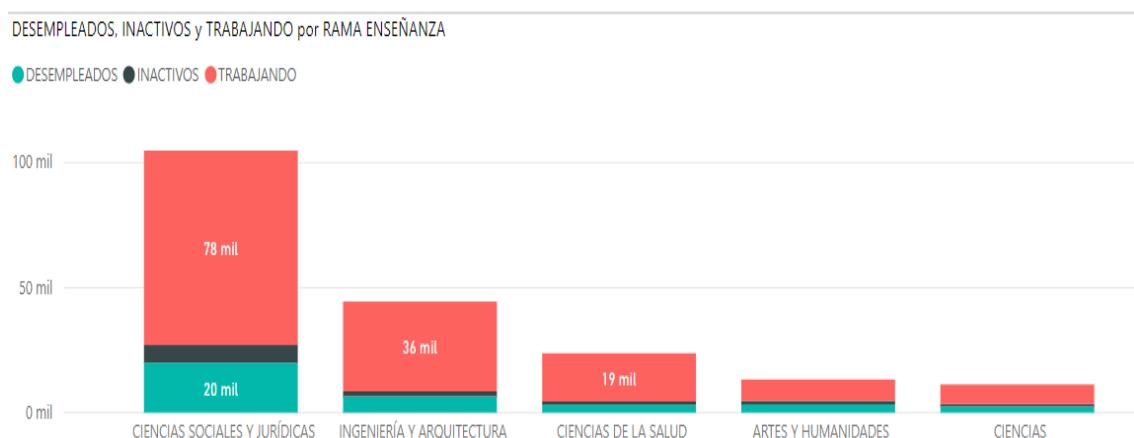
- Top 10 de universidades con mayor número de egresados.

En informe EGRESADOS UNIVERSITARIOS- PERFIL EVOLUTIVO



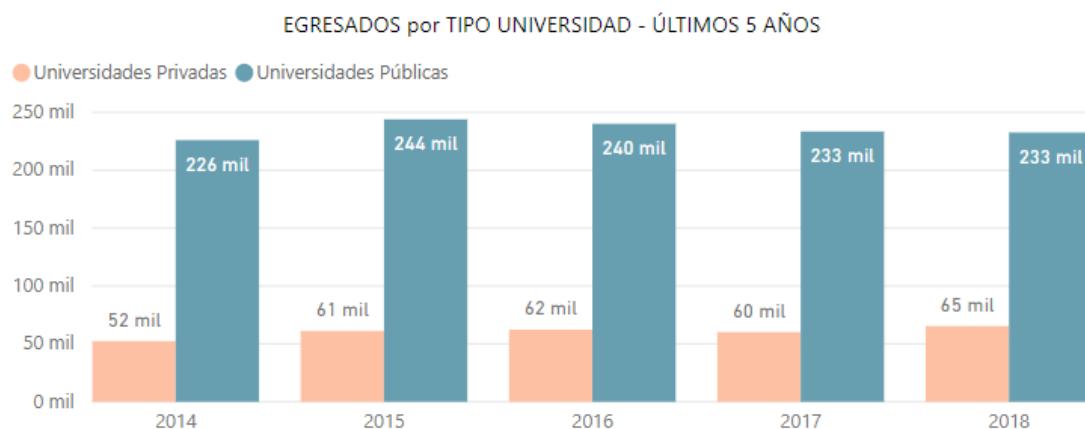
- Ranking de ramas de conocimiento con mayor número de estudiantes egresados insertados.

En informe EGRESADOS UNIVERSITARIOS 2009-2010 INSERCIÓN LABORAL 2014



- Evolución en los últimos 5 años del número de egresados universitarios por tipo de universidad.

En informe EGRESADOS UNIVERSITARIOS- PERFIL EVOLUTIVO

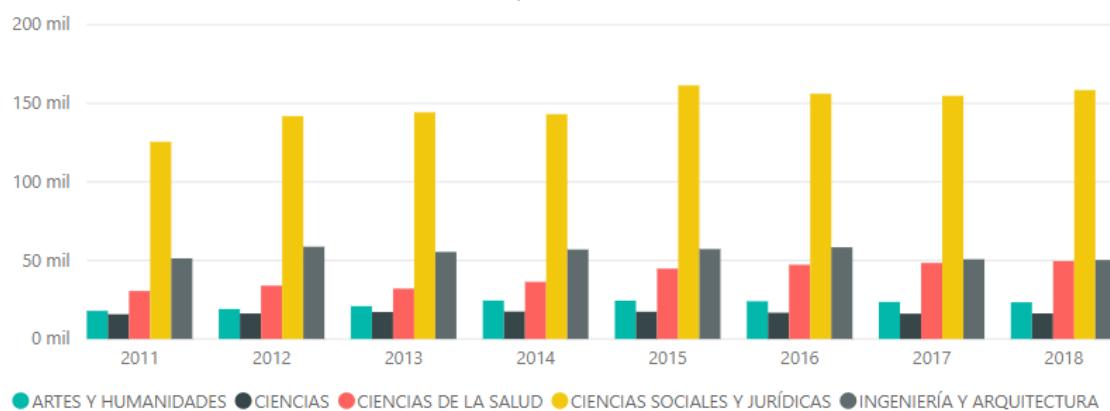


- Evolución del número de egresados universitarios por modalidad de impartición y rama de conocimiento.

En informe EGRESADOS UNIVERSITARIOS- PERFIL EVOLUTIVO

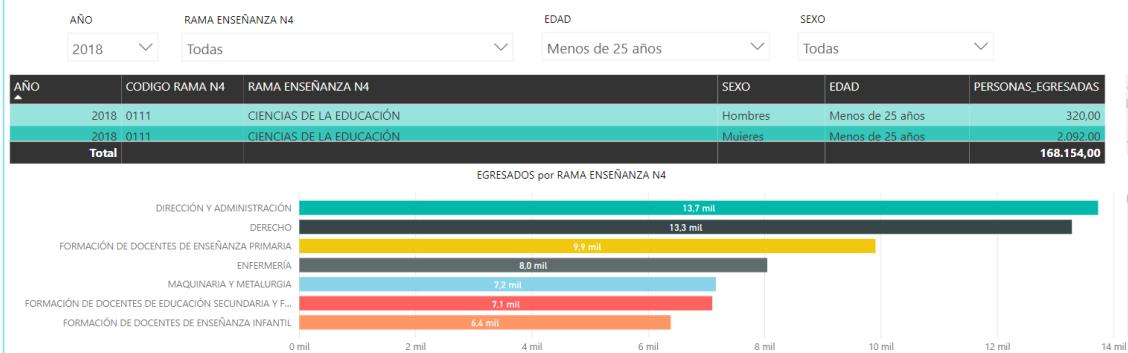


EGRESADOS por RAMA ENSEÑANZA



- Ranking de ámbitos de estudios con mayor número de egresados menores de 25 años en el curso académico 2016-2017.

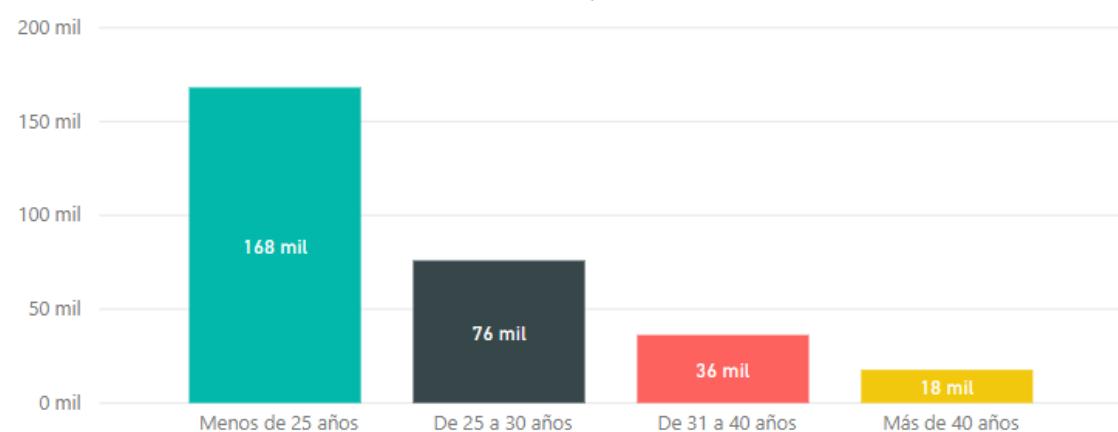
EGRESADOS UNIVERSITARIOS - PERFIL ÁMBITO ENSEÑANZA, EDAD Y GENERO



- Ranking de edad con mayor número de personas egresadas en el curso académico 2016-2017.

EGRESADOS UNIVERSITARIOS – PERFIL AMBITO ENSEÑANZA, EDAD Y GENERO

EGRESADOS por EDAD



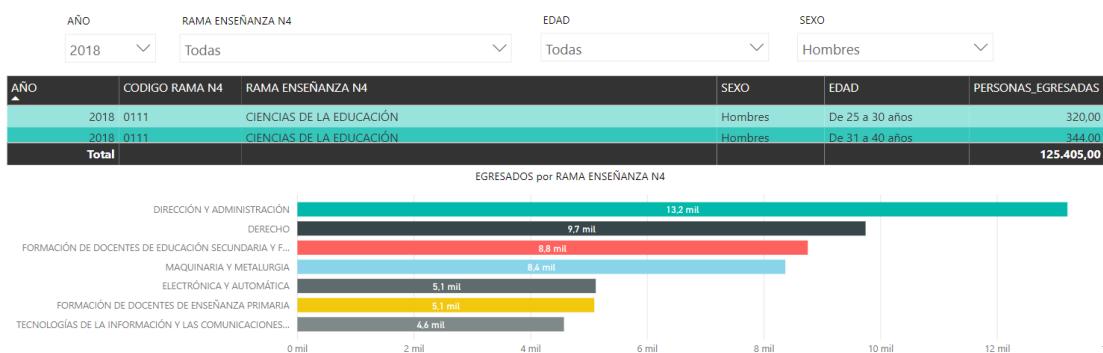
- Ámbito de estudios con mayor número de mujeres egresadas.

EGRESADOS UNIVERSITARIOS - PERFIL ÁMBITO ENSEÑANZA, EDAD Y GENERO



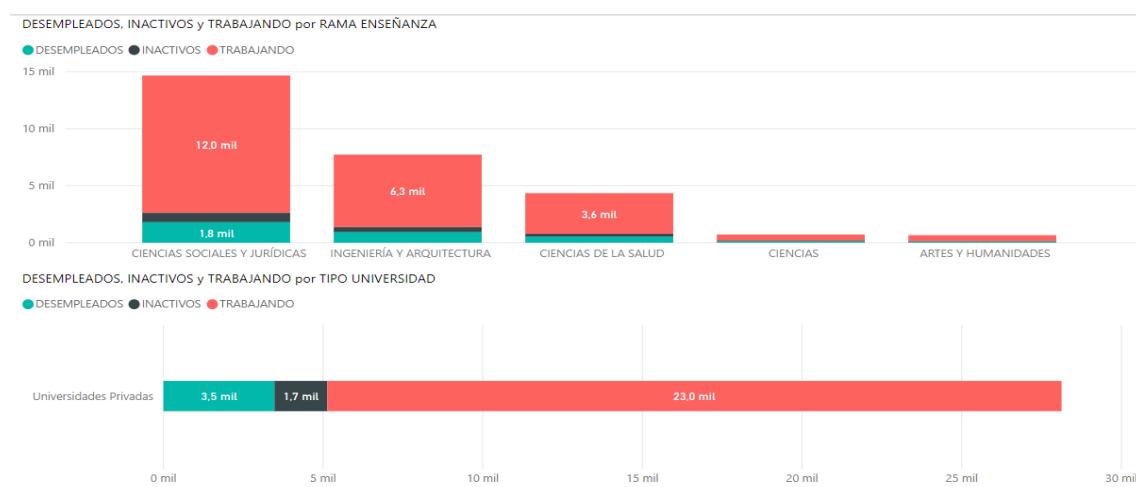
- Ámbito de estudios con mayor número de hombres egresados.

EGRESADOS UNIVERSITARIOS - PERFIL ÁMBITO ENSEÑANZA, EDAD Y GENERO



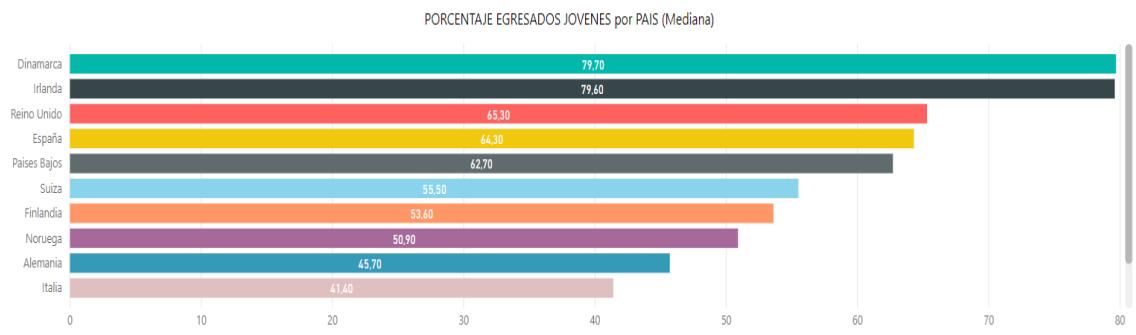
- Tipo de universidad y rama de conocimiento con menor número de estudiantes egresados insertados.

En informe EGRESADOS UNIVERSITARIOS 2009-2010 INSERCIÓN LABORAL 2014



- Ranking de países con mayor porcentaje de estudiantes jóvenes con estudios superiores completos.

En informe EGRESADOS UNIVERSITARIOS – COMPARACIÓN INTERNACIONAL EUROPA



Publicación de los resultados:

En el siguiente repositorio de github se encuentran los recursos necesarios para reproducir los informes obtenidos y el documento completo del proyecto:

<https://github.com/juanpabosu/DatawarehouseUOC>

Programas

Para el presente caso, la UOC proporciona un entorno VDI con todo el software preconfigurado con las siguientes características:

- Sistema operativo: Windows 10
- Base de datos: Base de datos remota Microsoft SQL Server 2016 accesible desde cliente mediante SQL Server Management Studio 17.
- Herramienta para la creación de cubos OLAP: PowerBI Desktop
- Herramienta de diseño de ETLs: Spoon – Pentaho Data Integration 8.0
- Herramienta de creación de informes: PowerBI Desktop

Bibliografía

Material de la asignatura Data Warehouse de la UOC.

Kimball, R. (2013) The Data Warehouse Toolkit. Third Edition New York: John Wiley & Sons Inc.

Inmon W.H., Imhoff Claudia y Sousa Ryan (1998) Corporate Information Factory EEUU: John Wiley & Sons Inc.

Inmon W.H. (1996) Building the Data Warehouse (2^a Ed.). EEUU: John Wiley & Sons Inc.

Inmon, W.H. Strauss, D. Neushloss, G. (2008) DW 2.0: The Architecture for next generation of Data Warehousing. EEUU: Morgan Kaufman Series.

Krish Krishnan (2013) Data Warehousing in the Age of Big Data. The Morgan Kaufmann Series on Business Intelligence

Enlaces a internet

Getting started with SQL Server Analysis Services:

<http://www.mssqltips.com/sqlservertip/1167/getting-started-with-sql-server-analysis-services/>

MSDN Analysis Services tutorial:

<http://msdn.microsoft.com/en-us/library/ms170208%28v=SQL.105%29.aspx>

Tutorial Pentaho Data Integration:

<http://wiki.pentaho.com/display/EAI/Latest+Pentaho+Data+Integration+%28aka+Kettle%29+Documentation>