## 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](https://ec.europa.eu/info/business-economy-euro/economic-and-fiscal-policy-coordination/eu-economic-governance-monitoring-prevention-correction/european-semester/european-semester-your-country/spain/europe-2020-targets-statistics-and-indicators-spain_en#employment-rate)

* + **Empresas Infomediarias**, 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](https://ec.europa.eu/eurostat/web/microdata/european-union-labour-force-survey) 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](http://uis.unesco.org/en/topic/international-standard-classification-education-isced).csv | Clasificación normalizada internacional de educación  (ISCED-F 2013) y el campo de estudio | [MECD](http://www.educacionyfp.gob.es/servicios-al-ciudadano-mecd/estadisticas/educacion/universitaria/estadisticas/anexos.html) |
| 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](https://datos.gob.es/es/catalogo?q=Encuesta%2Bde%2BInserci%C3%B3n%2BLaboral%2Bde%2BGraduados%2BUniversitarios%2B&tags=Mercado%2Blaboral%2By%2Bsalarios&theme_es=Empleo&tags=Labour%2Bmarket%2Band%2Bwages&administration_level=E&tags=Estad%C3%ADsticas&sort=metadata_modified%2Bdesc). Datos  de egresados según situación profesional en 2014. | datos.gob.es (*OpenData*) |
| educ\_uoe\_grad01.xls | Comparación internacional del número de egresados universitarios entre España y otros  Países. | Eurostat. [Estadísticas](https://ec.europa.eu/eurostat/web/education-and-training/data/database) [Europeas](https://ec.europa.eu/eurostat/web/education-and-training/data/database) |
| edat\_lfse\_03.xls | Comparación internacional del porcentaje de egresados universitarios jóvenes (Entre 20 y 29 años) con estudios superiores completos. | Eurostat. [Estadísticas](https://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&pcode=t2020_41&plugin=1) [Europeas](https://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&pcode=t2020_41&plugin=1) |

1. [La reutilización de información ocupa a 5.200 profesionales de 535 empresas en España](https://www.red.es/redes/es/actualidad/magazin-en-red/la-reutilizaci%C3%B3n-de-informaci%C3%B3n-ocupa-5200-profesionales-de-535-empresas)

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 adecue 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.

## 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.

## 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:

1. Tipo de Universidad
2. Tipo de Universidad (Privadas o Públicas) y Universidad.
3. Tipo de Universidad y Modalidad (Presencial, Especial, No Presencial).
4. Universidad y Ramas de Enseñanza.
5. 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:
6. Sexo.
7. 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:

1. Tipo de Universidad (P.Ej: Universidades Privadas).
2. Modalidad.(P.Ej: No Presencial)
3. Universidad (P.Ej: Oberta de Catalunya).
4. Rama de enseñanza. (P.Ej: Ciencias Sociales y Jurídicas)
5. 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.

1. 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 Públicas” |
| 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**](http://uis.unesco.org/en/topic/international-standard-classification-education-isced)**.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\_uoe\_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\_lfse\_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 **🡪** fichero anual  **🡪** Tipo Universidad  **🡪** Modalidades  🡪 32 Universidades  **🡪** 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 **🡪** fichero anual  **🡪** Tipo Universidad  **🡪** Modalidades  🡪 50 Universidades  **🡪** 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  **🡪** fichero anual  **🡪**88 Ámbitos Estudio  🡪 2 Sexo  **🡪**4 Rangos Edad | Niveles Académicos : 2 datos | 1 fichero  x 88 Ámbitos  x 2 Sexo  x 4 Edades  x 2 datos = 1408 |
| 03003 🡪 1 fichero anual  🡪 2 Tipo Universidad  🡪 2 Sexo  🡪 5 Ramas | Situaciones Laborales : 3 datos | 1 fichero  x 2 Tipos  x 2 Sexo  x 5 Ramas  x 3 datos = 60 |
| educ\_uoe\_grad01  **🡪** ficheros anuales  **🡪1**2 Países | Cursos académicos: 5 datos | 1 ficheros  x 12 Países  x 5 datos = 60 |
| edat\_lfse\_03  **🡪** fichero anual  **🡪1**2 Países | Cursos académicos: 5 datos | 1 fichero  x 12 Países  x 5 datos = 60 |
| [ISCED\_2013](http://uis.unesco.org/en/topic/international-standard-classification-education-isced)  **🡪** fichero anual | Titulaciones 🡪 162 datos | 1 fichero  x 162 datos= 162 |
|  | TOTAL | =11.590 |

1. Análisis Funcional

A continuación, se propone el tipo de arquitectura para la factoría de información que mejor se adecua 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:

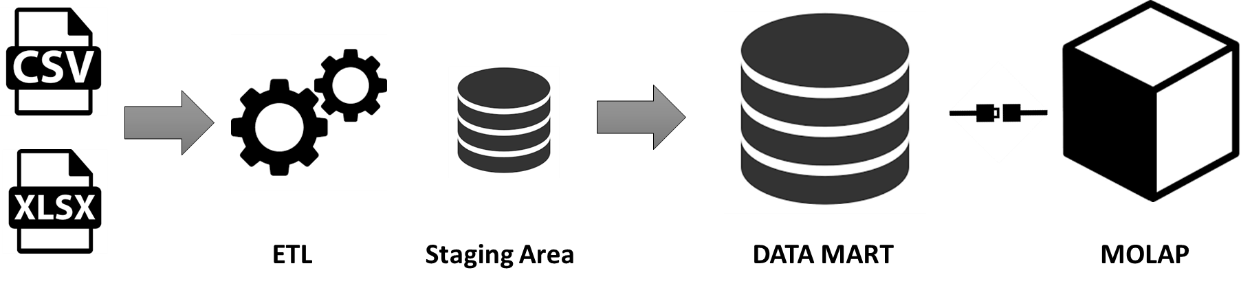
* *Staging Area* (opcional): En el caso de tener múltiples fuentes (ficheros, bases de datos, servicios RSS….) es conveniente cargarlas para consolidar la información en una estructura de carga intermedia que puede ser creada en la misma base de datos.

Esta área del DW también puede servir para entender, simplificar y consolidar el proceso ETL.

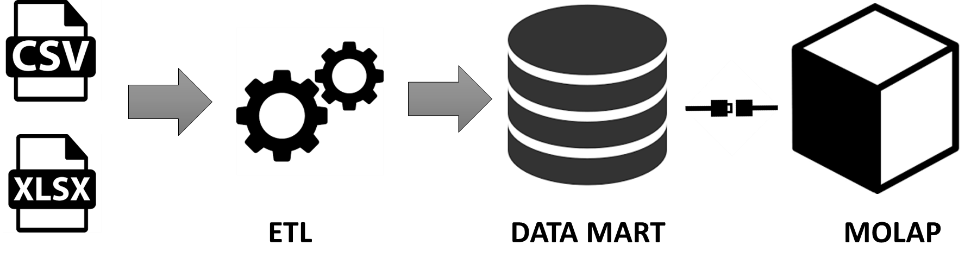
* *Data Mart* de personas egresadas: al centrarnos en una única área temática como es el análisis de egresados universitarios, es más correcto considerar que se está creando un *data mart* en lugar de un almacén de datos corporativo.
* *MOLAP*: a partir de la información del *datamart* se creará un cubo 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*.



1. 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:

**FACT\_PEGR\_EVOL**

DIM\_UNIVERSIDAD

DIM\_TIPO\_UNIV

DIM\_ANIO

DIM\_RAMA

DIM\_MODALIDAD

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:

**FACT\_PEGR\_PERFIL**

DIM\_SEXO

DIM\_EDAD

DIM\_ANIO

DIM\_RAMA

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:

**FACT\_PEGR\_INSERTADAS**

DIM\_SEXO

DIM\_TIPO\_UNIV

DIM\_ANIO

DIM\_RAMA

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:

**FACT\_PEGR\_COMPARATIVA**

DIM\_PAISES

DIM\_ANIO

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.



**FACT\_PEGR\_EVOL**

DIM\_UNIVERSIDAD

DIM\_TIPO\_UNIV

DIM\_ANIO

DIM\_RAMA

DIM\_MODALIDAD

PERSONAS\_EGRESADAS

COD\_ANIO

ANIO

COD\_UNIVERSIDAD

UNIVERSIDAD

COD\_TIPO\_UNIV

TIPO\_UNIVERSIDAD

COD\_RAMA

RAMA DE ENSEÑANZA

COD\_MODALIDAD

MODALIDAD

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:

**FACT\_PEGR\_PERFIL**

DIM\_SEXO

DIM\_EDAD

DIM\_ANIO

DIM\_RAMA

PERSONAS\_EGRESADAS

COD\_ANIO

ANIO

COD\_SEXO

SEXO

COD\_EDAD

INT\_EDAD

COD\_RAMA\_N4

ÁMBITO DE ENSEÑANZA

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:

**FACT\_PEGR\_INSERTADAS**

DIM\_SEXO

DIM\_TIPO\_UNIV

DIM\_ANIO

DIM\_RAMA

PEGR\_TRABAJANDO

PEGR\_DESEMPLEADAS

PEGR\_INACTIVAS

COD\_SEXO

SEXO

COD\_ANIO

ANIO

COD\_TIPO\_UNIV

TIPO\_UNIVERSIDAD

COD\_RAMA

RAMA DE ENSEÑANZA

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:

**FACT\_PEGR\_COMPARATIVA**

DIM\_PAISES

DIM\_ANIO

PERSONAS\_EGRESADAS

PORCENTAJE\_PEGR\_JÓVENES

COD\_PAIS

PAIS

COD\_ANIO

ANIO

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 *Foreing 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 |

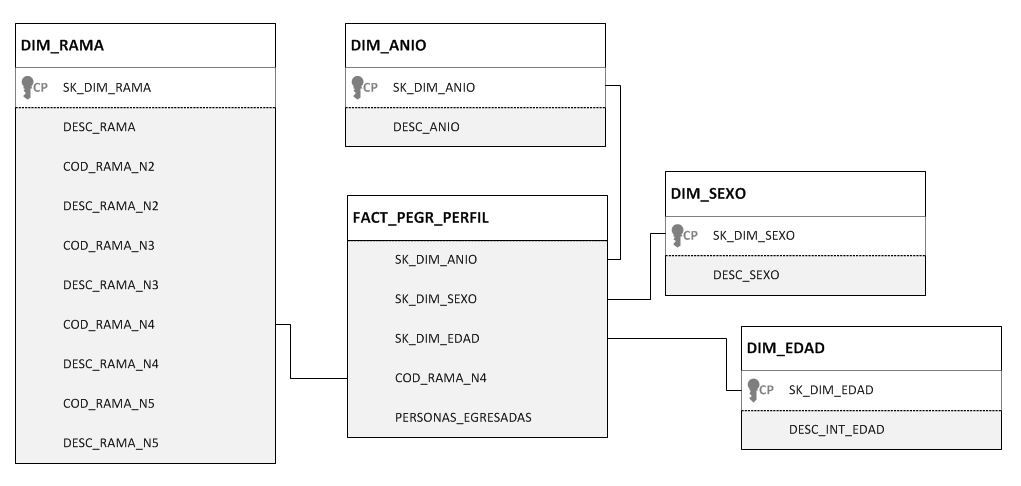
En la siguiente imagen se muestra el diseño del modelo físico[[1]](#footnote-1) para la tabla de hechos FACT\_PEGR\_EVOLUTIVO.



* 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 |

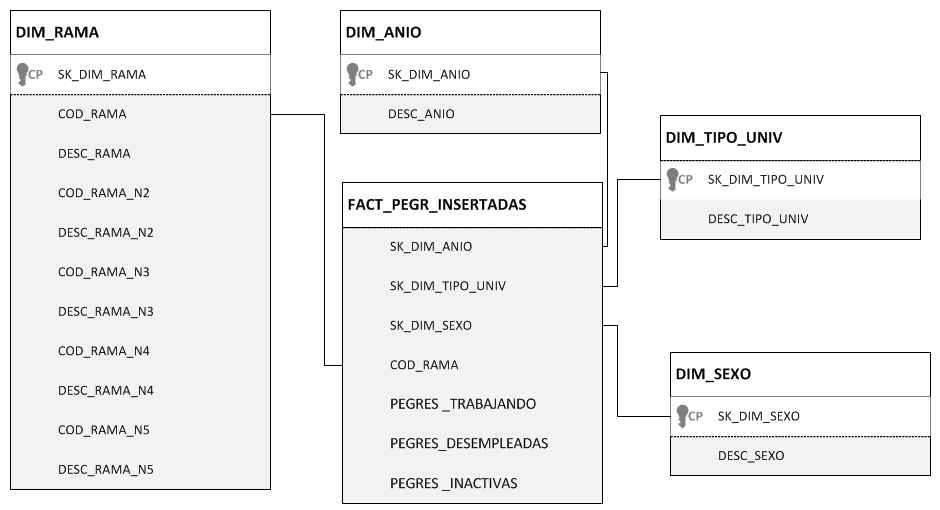
El diseño del modelo físico para la tabla de hechos FACT\_PEGR\_PERFIL:



* 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 |

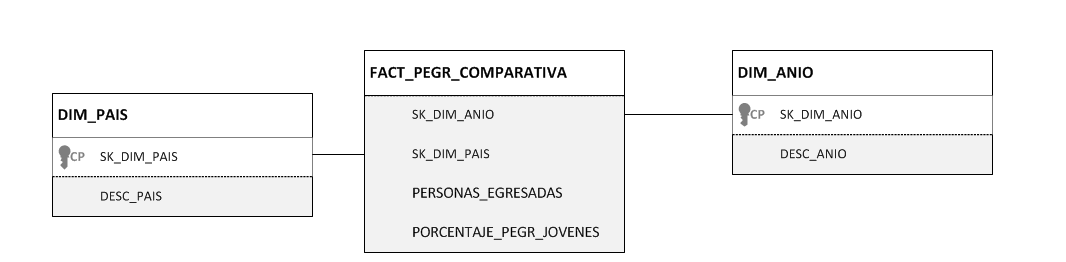
El diseño del modelo físico para la tabla de hechos FACT\_PEGR\_INSERTADAS:



* 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.

1. **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 crearan 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 cargaran a través de un script sql, las dimensiones DIM\_RAMA y DIM\_UNIVERSIDAD así como las tablas de hechos se cargaran a través de flujos de PDI.

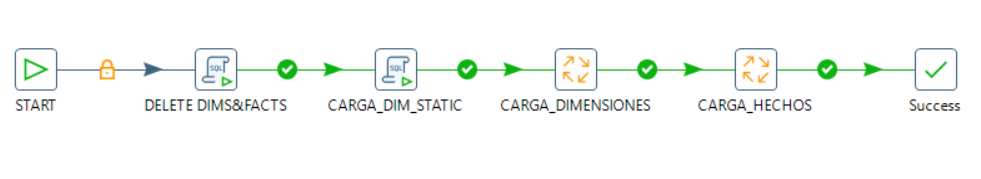
Se definirá un directorio para el manejo de errores en la carga de datos sobre la tabla destino (F:\Pentaho\PRA2\error) y otro directorio para almacenar los archivos correspondientes a las fuentes de datos originales y algunos archivos intermedios obtenidos a partir de las mismas fuentes (F:\Pentaho\PRA2\fuentes)

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 que 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*.

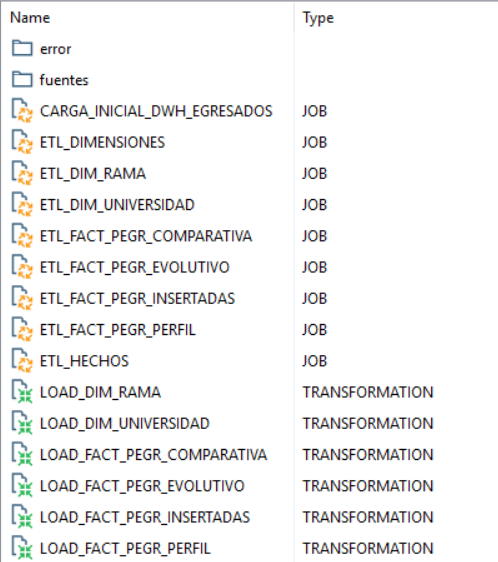
En esta propuesta, se realizará control de errores para las trasformaciones de carga de datos a través de archivos de texto planos almacenados en el entorno de trabajo configurado en PDI, de esta manera el usuario encardo de ejecutar el job principal puede revisar en estos archivos el detalle correspondiente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



## 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]

GO

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

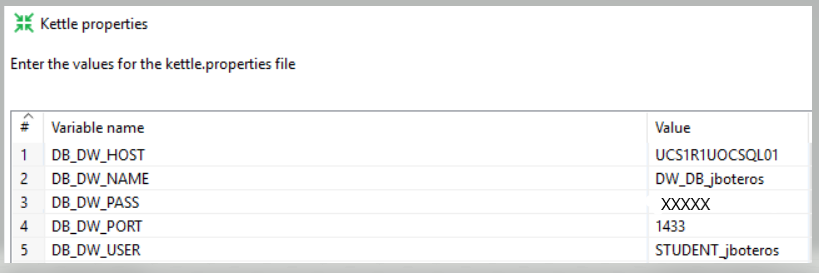
ALTER TABLE [dbo].[FACT\_PEGR\_PERFIL] CHECK CONSTRAINT [FK\_FACT\_PEGR\_PERFIL\_DIM\_SEXO]

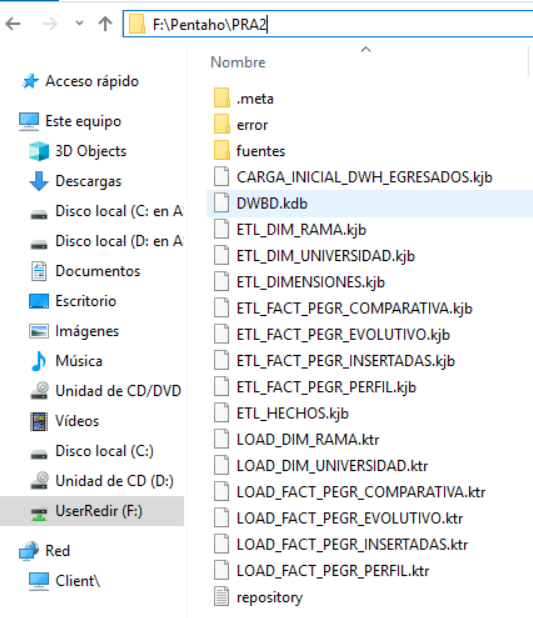
GO

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

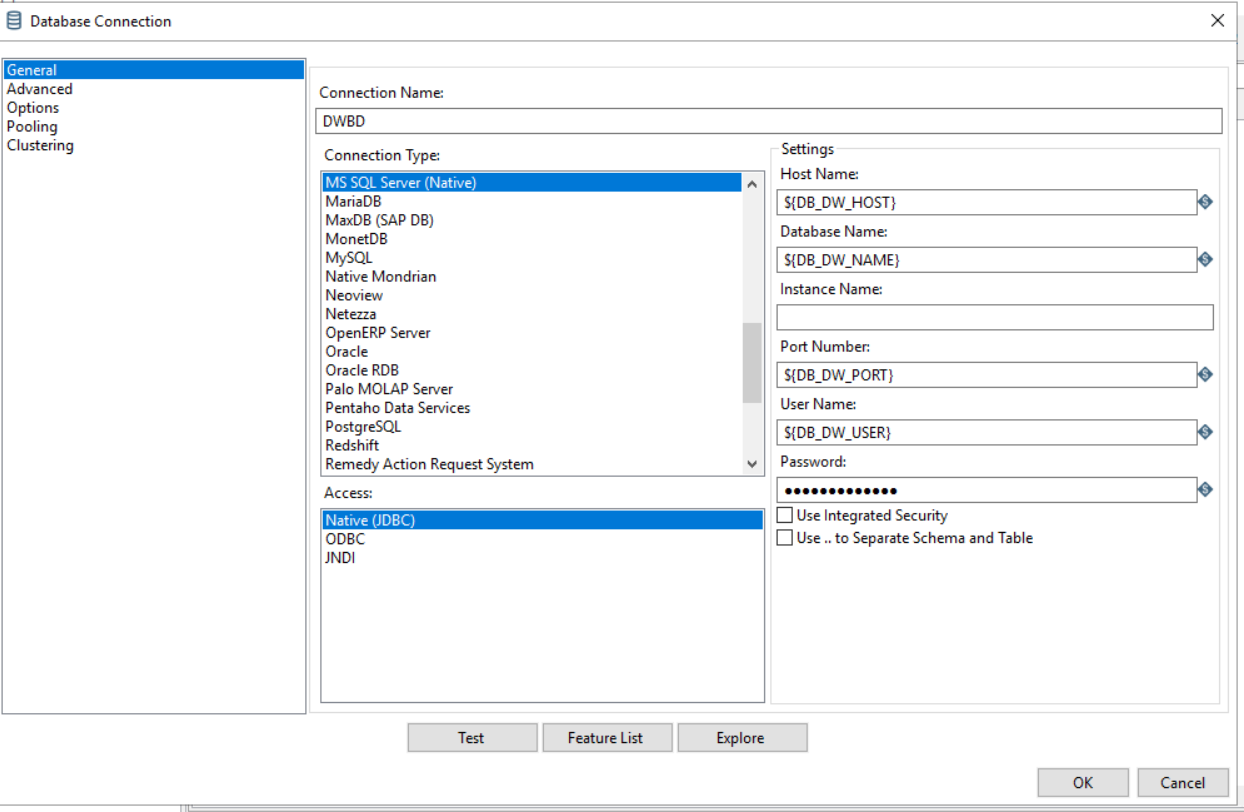
Variables de entorno

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



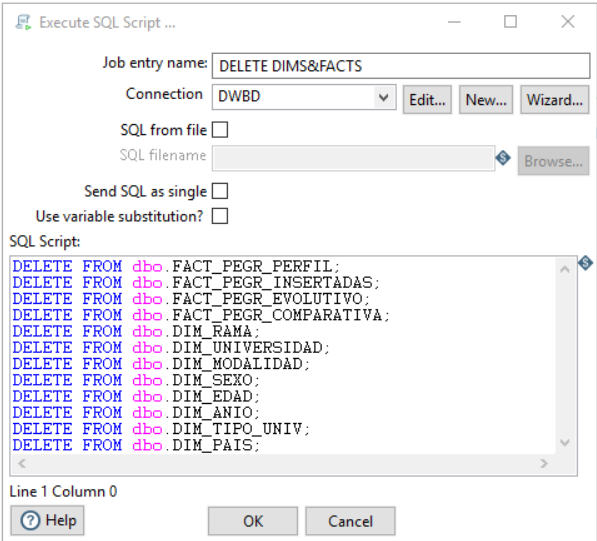


Crear conexión a base de datos **DWBD** utilizando las variables de entorno:

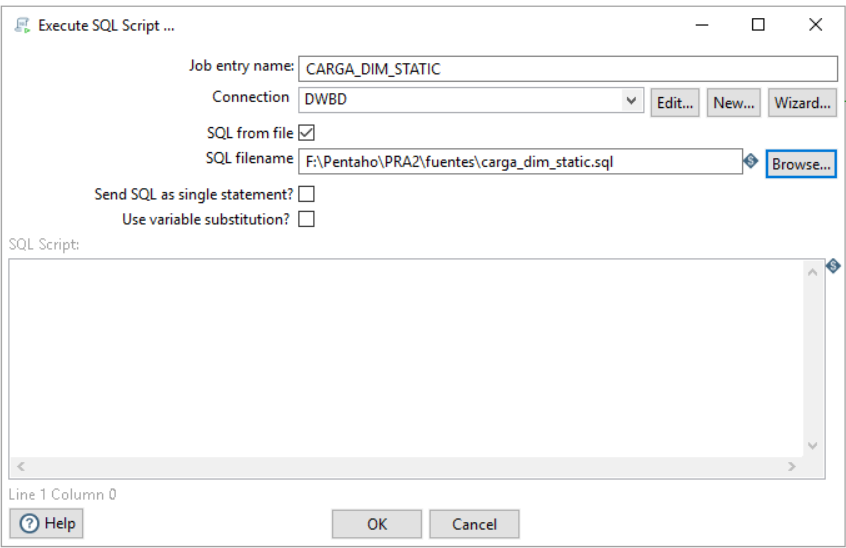


A continuación pasamos a explicar los componentes de la ETL principal y sus flujos de datos:

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



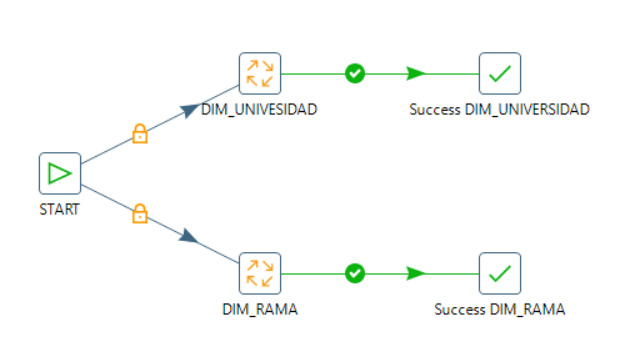
* CARGA\_DIM\_STATIC: scripting sql encargado de cargar las dimensiones DIM\_ANIO, DIM\_EDAD, DIM\_TIPO\_UNIV, DIM\_SEXO, DIM\_PAIS y DIM\_MODALIDAD, a partir de las instrucciones contenidas en el archivo F:\Pentaho\PRA2\fuentes\*carga\_dim\_static.sql*



* CARGA\_DIMENSIONES: job principal encargado de ejecutar el proceso etl en paralelo para las dimensiones DIM\_RAMA y DIM\_UNIVERSIDAD



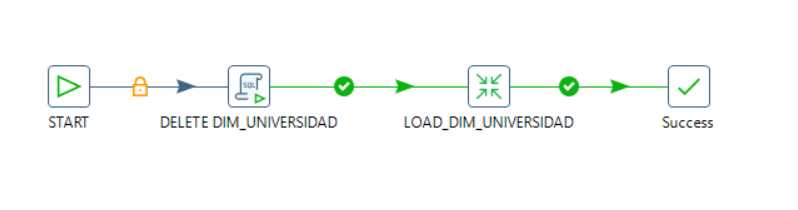
* + ETL\_DIMENSIONES: job encargado de ejecutar el flujo de limpieza y carga de datos en paralelo para las dimensiones DIM\_RAMA y DIM\_UNIVERSIDAD



* + DIM\_UNIVERSIDAD: job encargado de ejecutar de forma individual el flujo de limpieza y carga de la dimensión DIM\_UNIVERSIDAD.

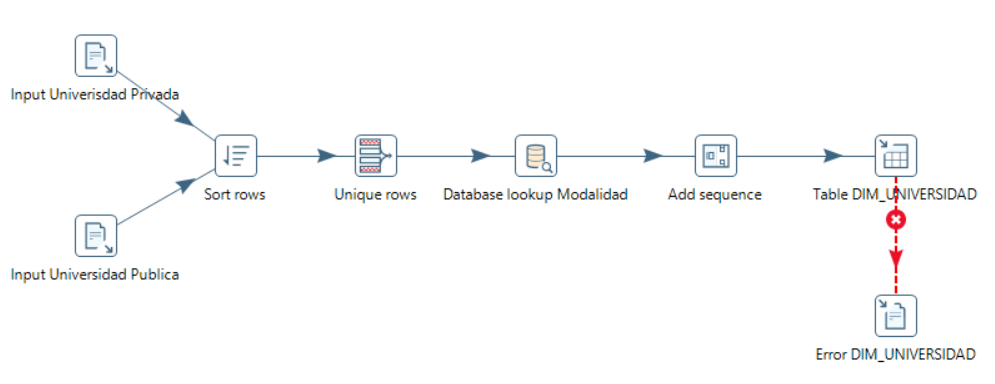


* + ETL\_DIM\_UNIVERSIDAD: job encargado de limpiar la tabla DIM\_UNIVERSIDAD y ejecutar la transformación LOAD\_DIM\_UNIVERSIDAD.

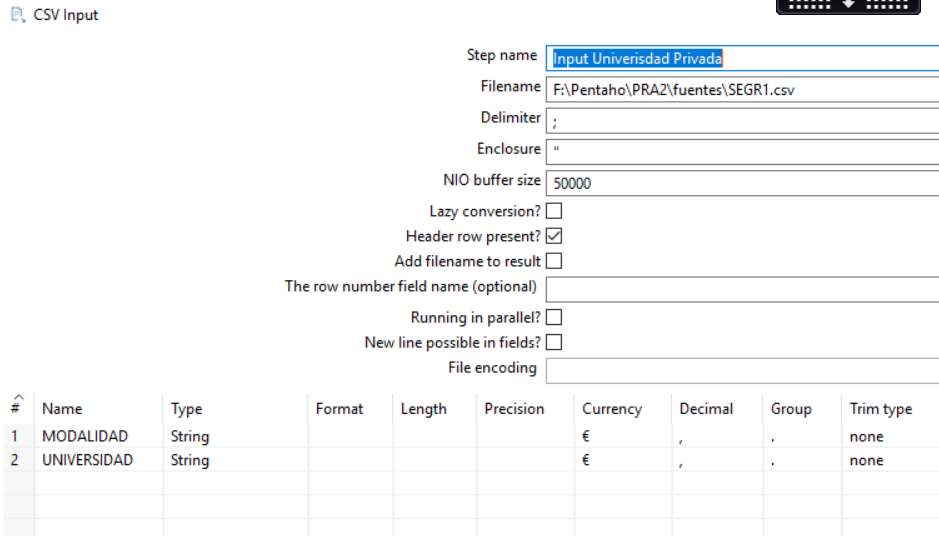


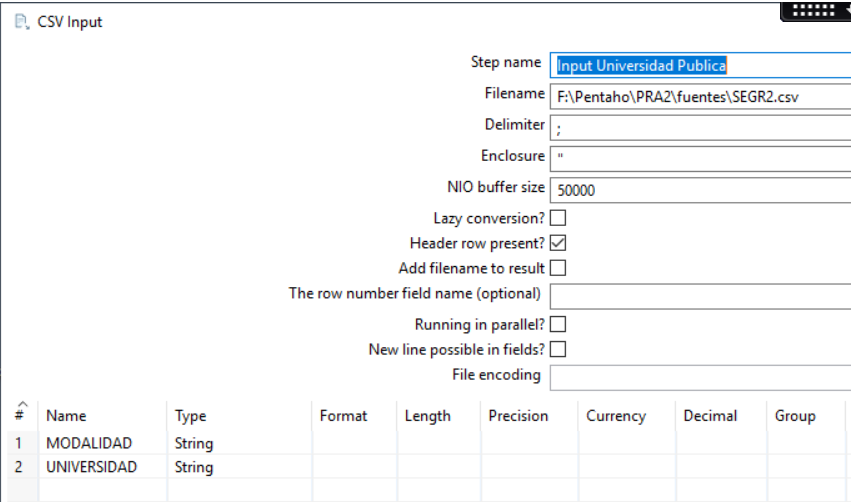
* + LOAD\_DIM\_UNIVERSIDAD: transformación encargada de extraer la información de las fuentes y carga de datos en tabla DIM\_UNIVERSIDAD



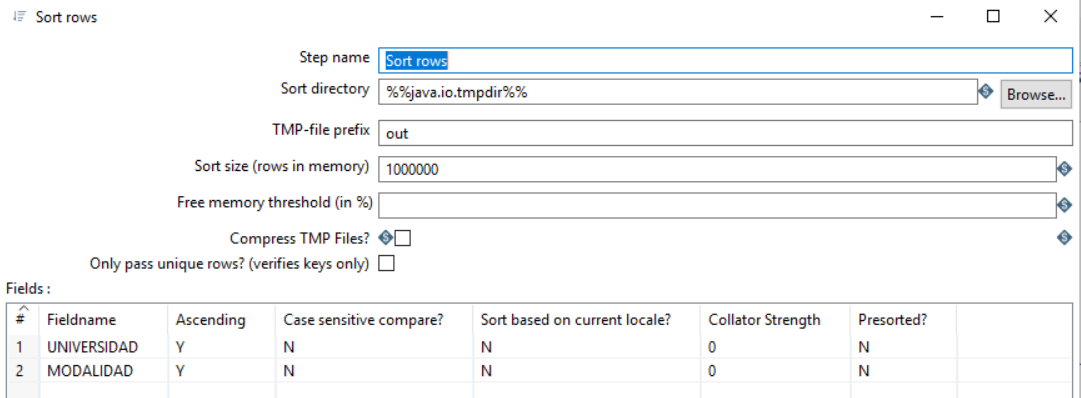


En el primer paso de input, se cargan los campos UNIVERSIDAD y MODALIDAD a partir de los archivos SEGR1.csv y SEGR2.csv que contienen la información correspondiente a los nombres de cada universidad y su respectiva modalidad de impartición

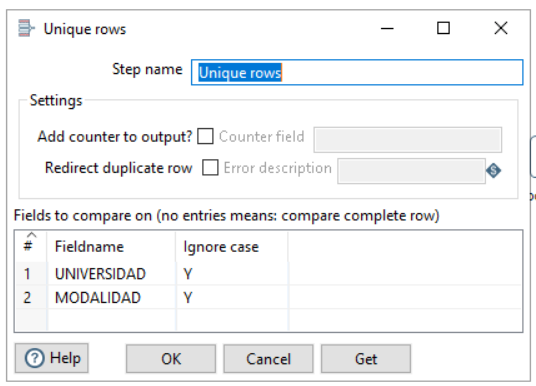




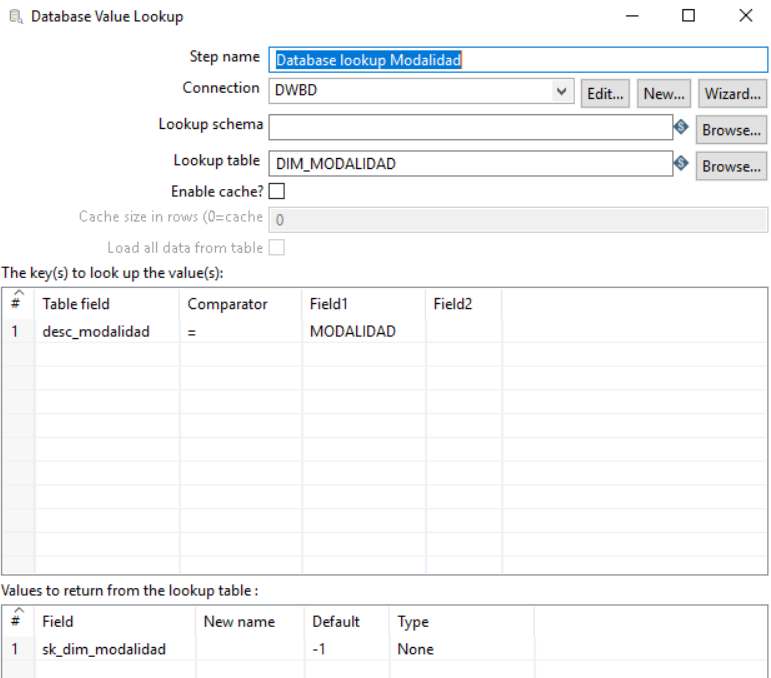
En el paso Sort Rows, se ordena el conjunto de datos obtenido de la unión de ambas fuentes, como prerrequisito para seleccionar las combinaciones únicas UNIVERISAD-MODALIDAD que se cargaran en la tabla destino.



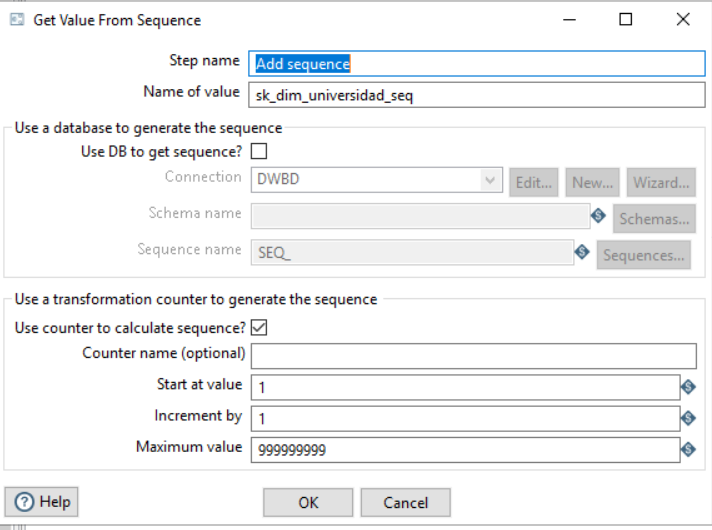
En el paso Unique Rows, se obtiene los valores únicos UNIVERSIDAD-MODALIDAD del subconjunto de datos anterior, dejando así las combinaciones necesarias para la carga de datos.



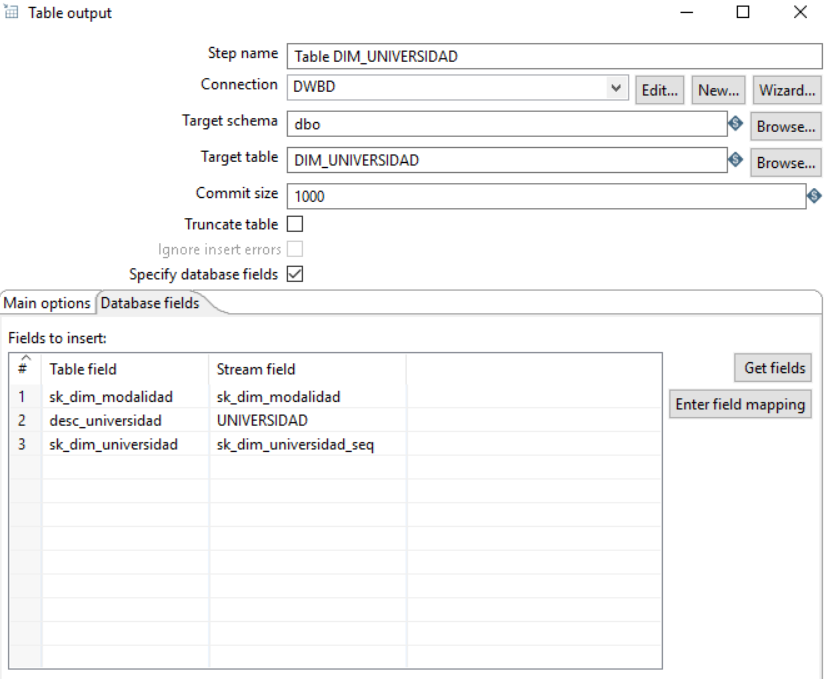
Como en la definición del modelo físico, la dimensión DIM\_UNIVERSIDAD tiene una llave de integridad referencial hacia la dimensión DIM\_MODALIDAD, es necesario ejecutar un proceso de lookup sobre esta última, para obtener así el valor de la clave primaria correspondiente a la modalidad de impartición 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.



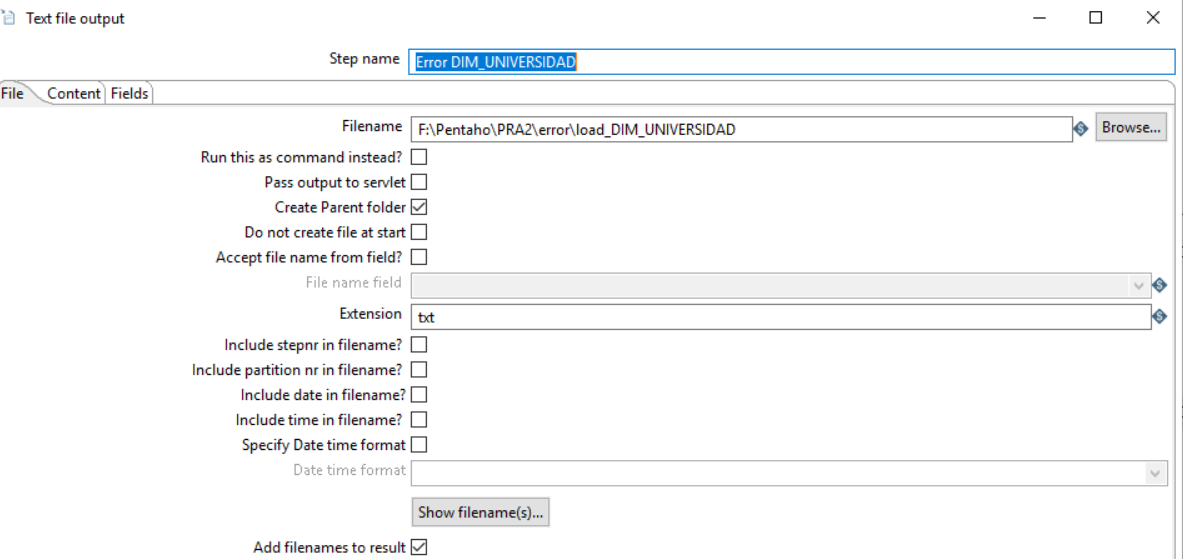
En el paso Add Sequence, se generan de forma automática los valores de secuencia que harán parte de la clave primaria en la dimensión DIM\_UNIVESIDAD, iniciando a partir del valor 1 y continuando con incrementos de 1 unidad: 1,2,3,4…, como se había definido en la fase de diseño las secuencias no se generan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



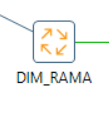
Por último en el paso Table DIM\_UNIVERSIDAD, se mapean los valores obtenidos de las fases anteriores con los respectivos campos a poblar en la tabla DIM\_UNIVERSIDAD, y se especifica un flujo alterno ante falla, con la salida de error a un archivo de texto plano.



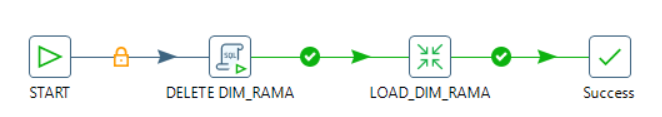




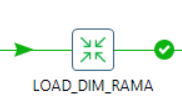
* + DIM\_RAMA: job encargado de ejecutar de forma individual el flujo de limpieza y carga de la dimensión DIM\_RAMA.

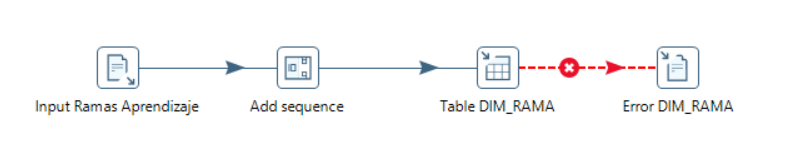


* + ETL\_DIM\_RAMA: job encargado de limpiar la tabla DIM\_RAMA y ejecutar la transformación LOAD\_DIM\_RAMA.

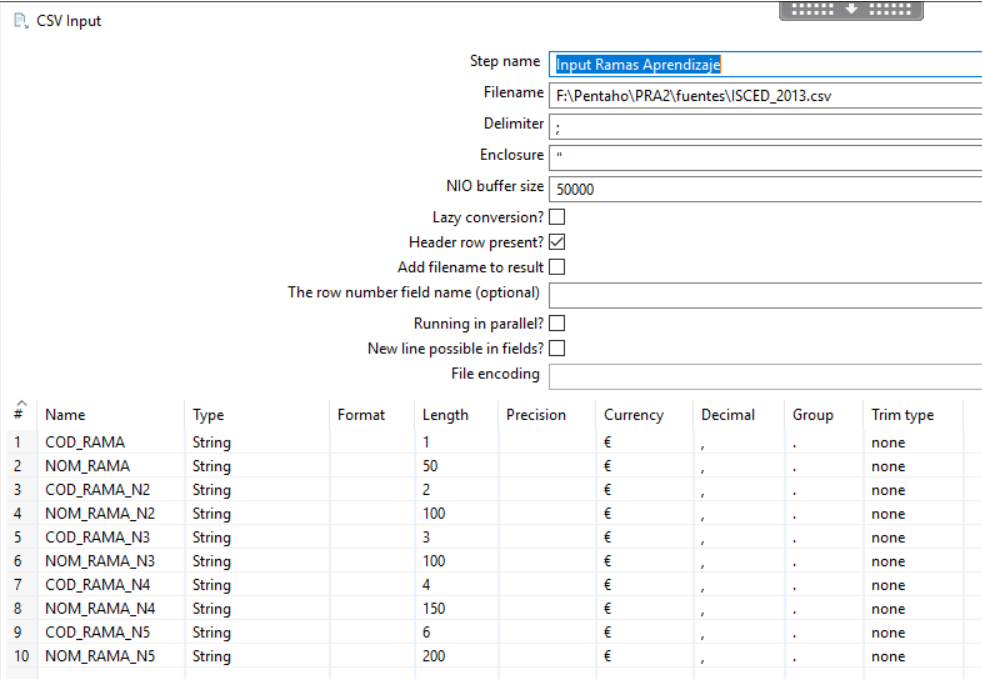


* + LOAD\_DIM\_RAMA: transformación encargada de extraer la información de las fuentes y carga de datos en tabla DIM\_RAMA

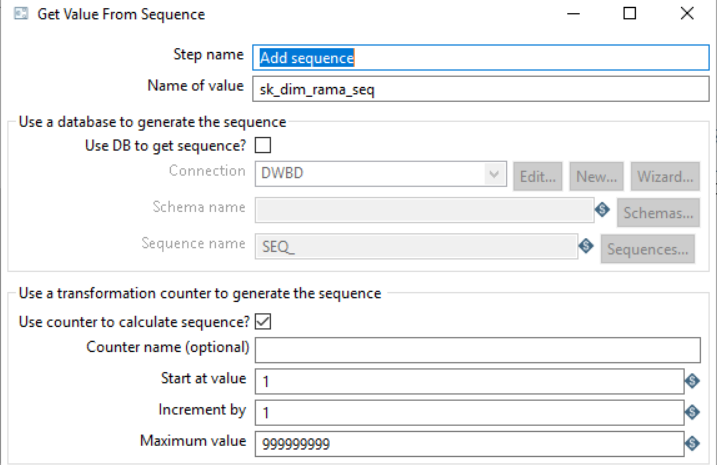




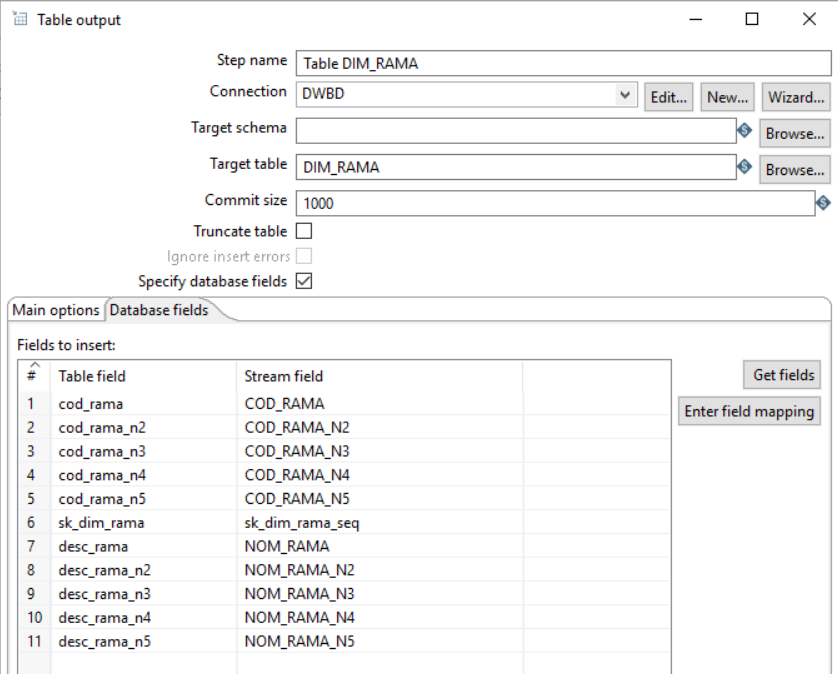
En el primer paso de input, se cargan los campos planos tal cual se encuentran en el archivo ISCED\_2013.csv y acorde a las definiciones de tipo de dato y tamaño definido en el modelo físico para la tabla DIM\_RAMA



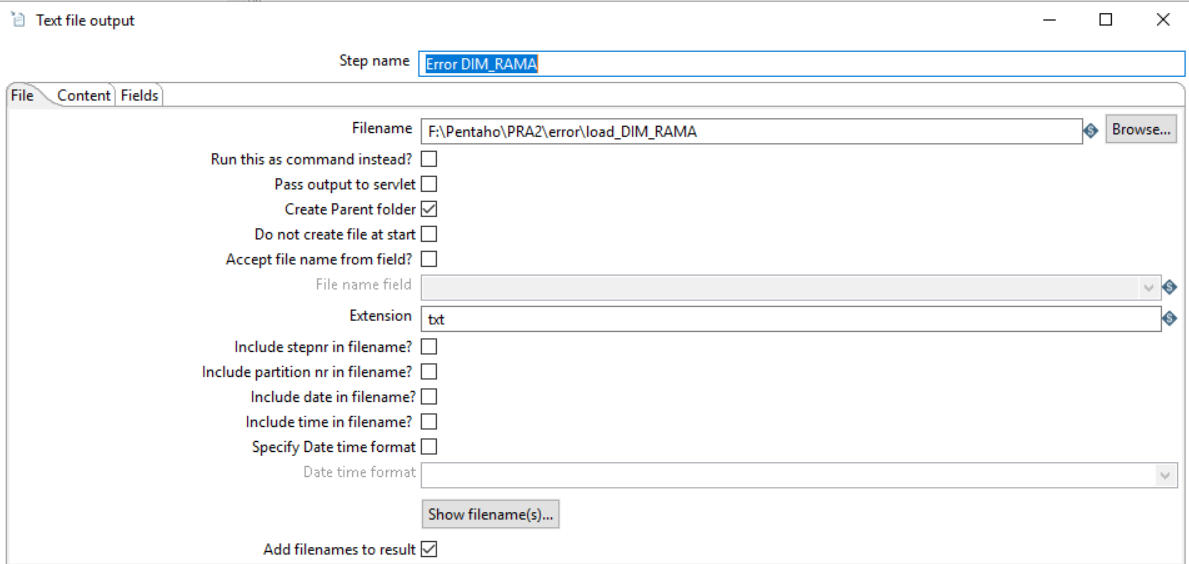
En el paso Add Sequence, se generan de forma automática los valores de secuencia que harán parte de la clave primaria en la dimensión DIM\_RAMA, iniciando a partir del valor 1 y continuando con incrementos de 1 unidad: 1,2,3,4…, como se había definido en la fase de diseño las secuencias no se generan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



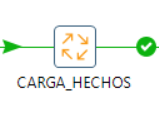
Por último en el paso Table DIM\_RAMA, se mapean los valores obtenidos de las fases anteriores con los respectivos campos a poblar en la tabla DIM\_RAMA, y se especifica un flujo alterno ante falla, con la salida de error a un archivo de texto plano.



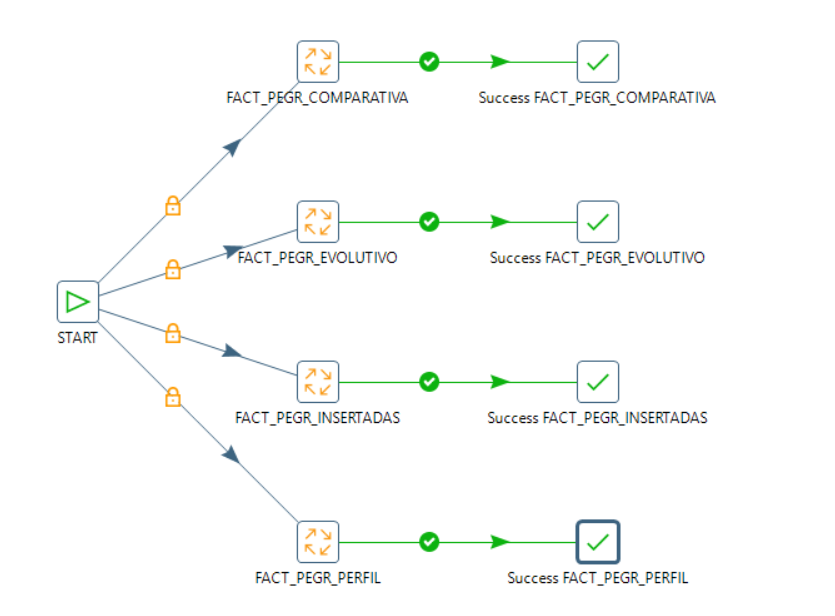




* CARGA\_HECHOS: job principal encargado de ejecutar el proceso etl en paralelo para las 4 tablas de hecho definidas en el diseño lógico y físico del DW.



* + 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.

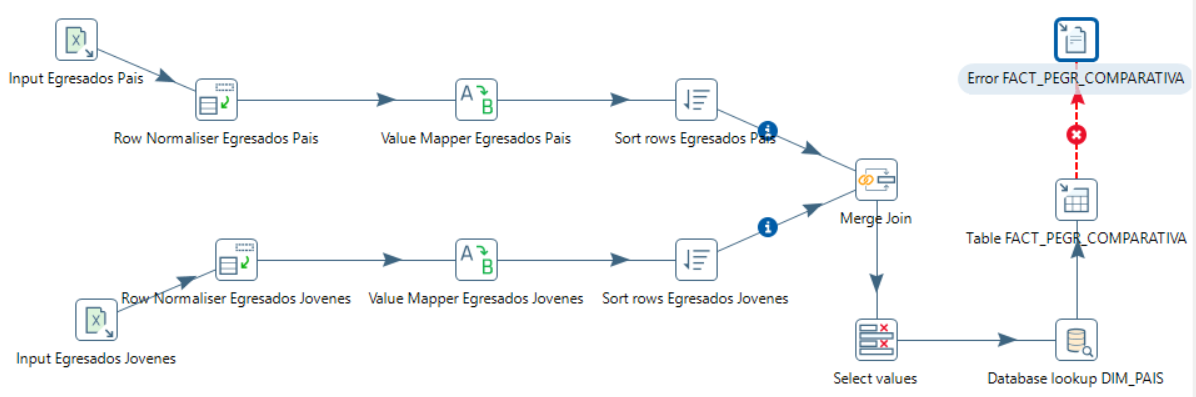


* + ETL\_FACT\_PEGR\_COMPARATIVA: job encargado de limpiar la tabla FACT\_PEGR\_COMPARATIVA y ejecutar la transformación LOAD\_FACT\_PEGR\_COMPARATIVA.

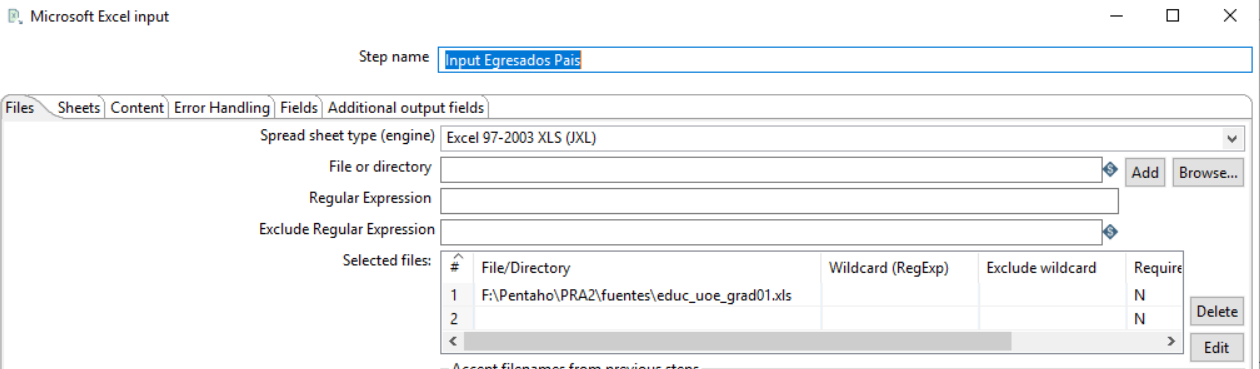


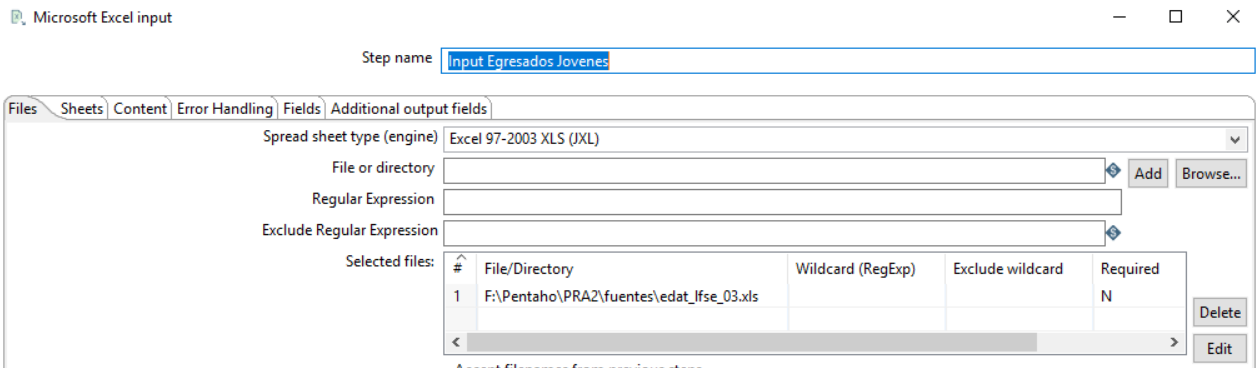
* + LOAD\_FACT\_PEGR\_COMPARATIVA: transformación encargada de extraer la información de las fuentes y carga de datos en tabla FACT\_PEGR\_COMPARATIVA.



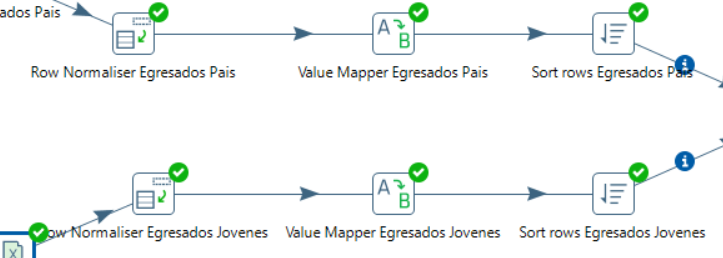


En el paso de input de datos, se utiliza la función para leer archivos .xls, uno de ellos referenciando la cantidad total de egresados por país, educ\_uoe\_grad01.xls , y el otro referencia el archivo con los porcentajes de egresados jóvenes por cada uno de los países, edat\_lfse\_03.xls

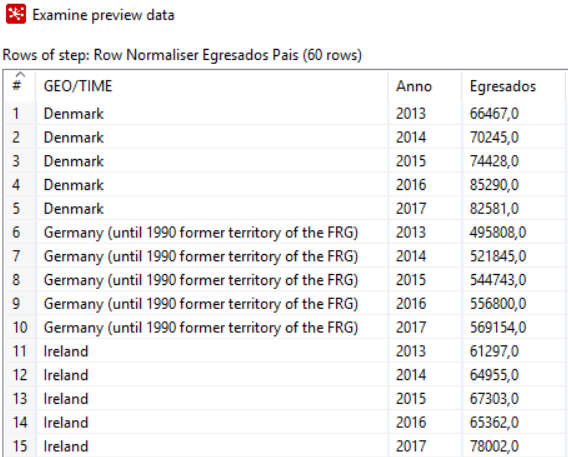




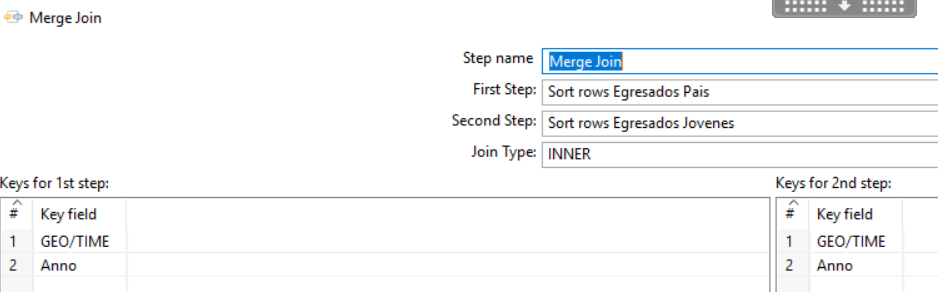
En los siguientes 3 pasos, se aplica la misma lógica sobre cada uno de los archivos fuente, una paso de normalización de manera que se obtengan los valores objetivo del total y el porcentaje de egresados en forma columnar agrupados por país y año, una función de value mapper para reemplazar los valores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



Ejemplo step de normalización fuente egresados por país y año



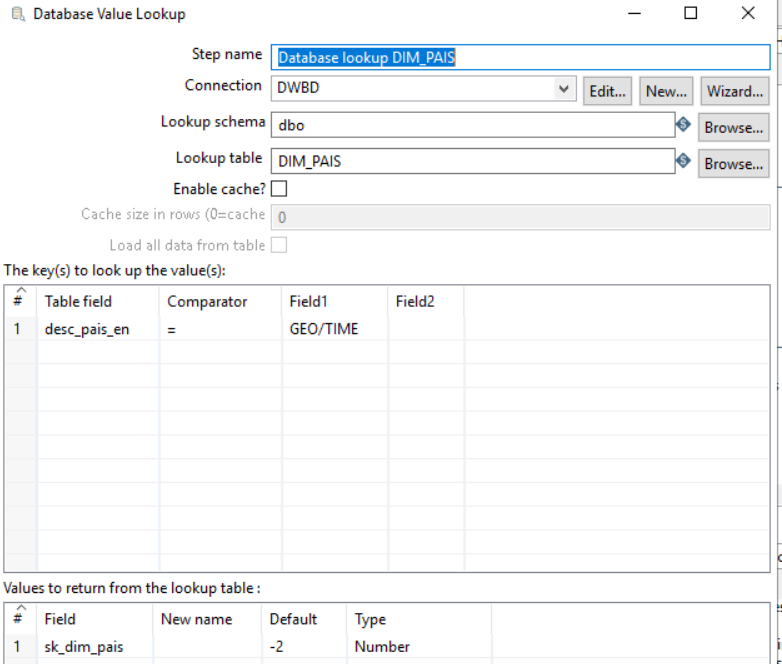
En el paso de merge join se unifican las fuentes de datos por los campos país y año, como resultado se consolidan las métricas de total de egresados y el porcentaje de jóvenes egresados en un solo flujo de datos



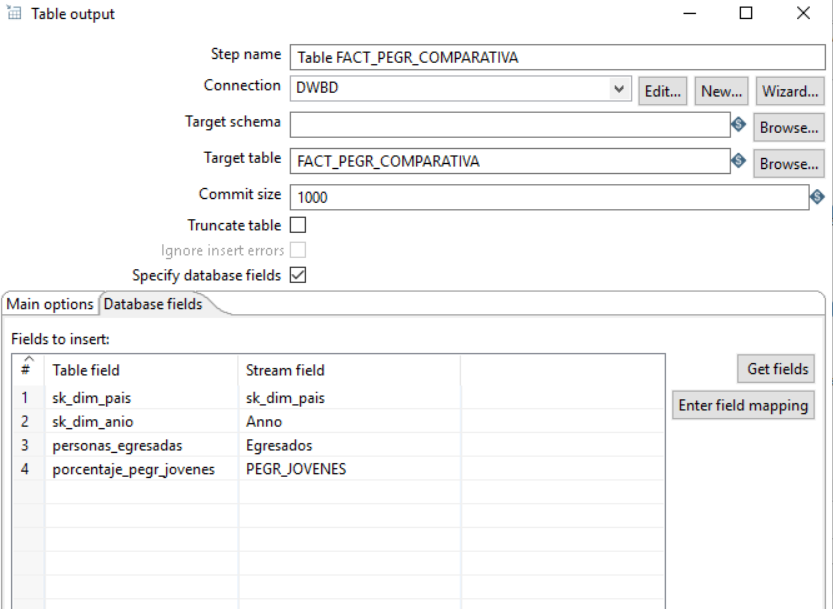
Luego de esto, se seleccionan los valores necesarios para la carga final de datos en la tabla de hecho con los respectivos tipos de datos especificados para el modelo físico

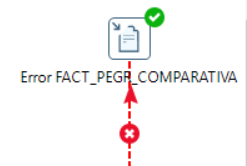


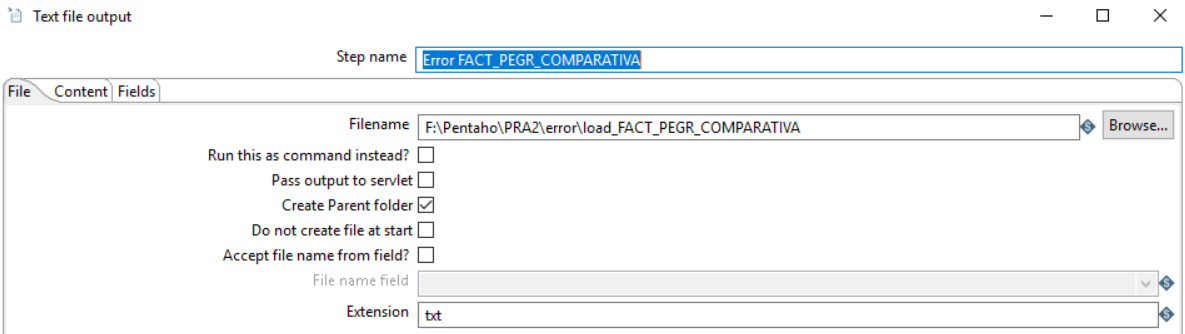
Como en la definición del modelo físico, la tabla de hecho tiene referencias foráneas a las dimensiones DIM\_PAIS y DIM\_ANIO, se utiliza el paso Database lookup DIM\_PAIS, para obtener así el valor de la clave primaria correspondiente al país para una tupla dada. Como se observa en la definición, si 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.



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.

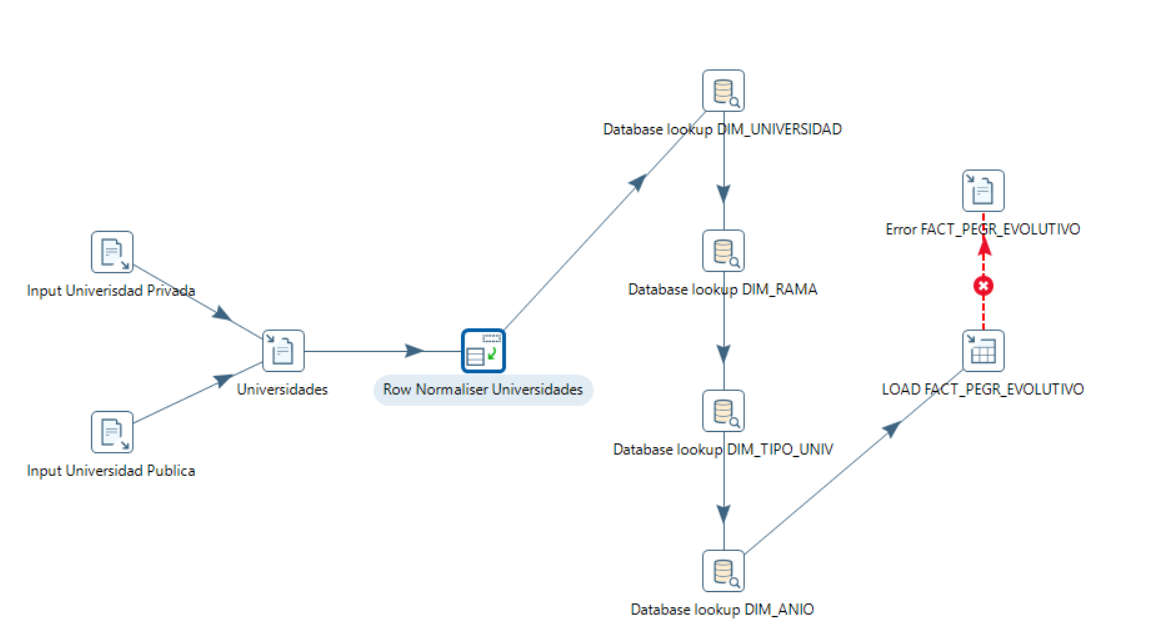


* + ETL\_FACT\_PEGR\_EVOLUTIVO: job encargado de limpiar la tabla FACT\_PEGR\_EVOLUTIVO y ejecutar la transformación LOAD\_FACT\_PEGR\_EVOLUTIVO

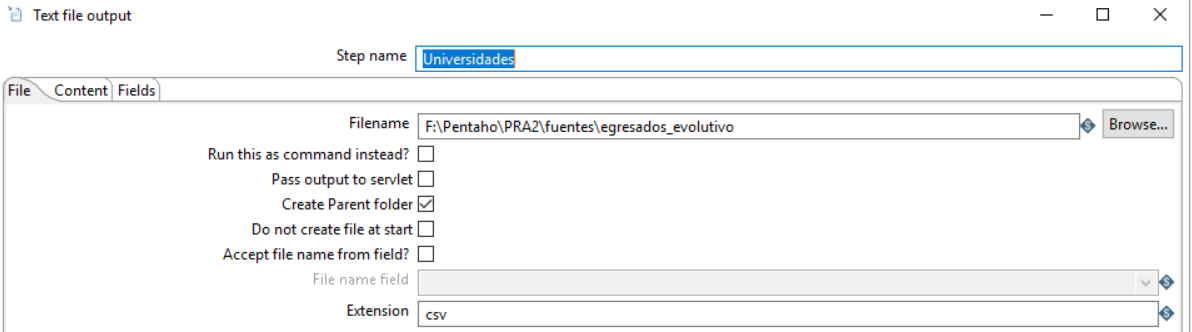


* + LOAD­\_FACT\_PEGR\_EVOLUTIVO: transformación encargada de extraer la información de las fuentes y carga de datos en tabla FACT\_PEGR\_EVOLUTIVO

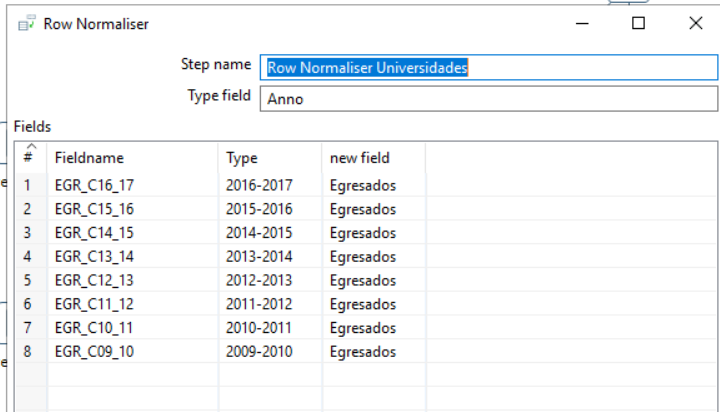




Como parte del procesamiento de las fuentes de entrada, se unifica la información de los archivos SEGR1.csv y SEGR2.csv (incluyendo todos los campos), en un solo archivo llamado egresados\_evolutivo.csv, de esta manera facilitamos la lectura de la información y las transformaciones necesarias en las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



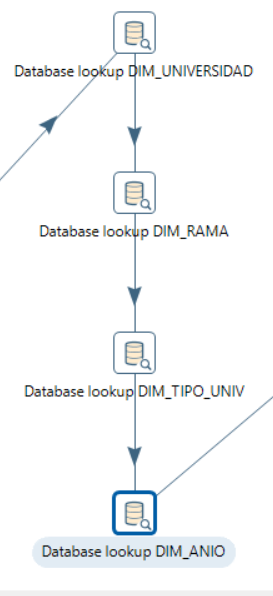
Homologo al paso realizado en la tabla de hechos anterior, se aplica una función de normalización, de forma que los datos queden agrupados verticalmente por año y el número total de egresados



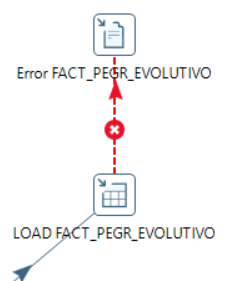
Ejemplo normalización por año y egresados



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.



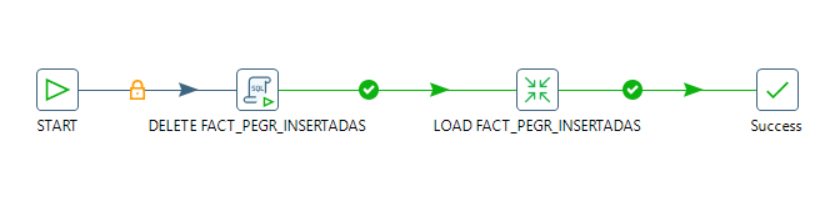
Por último en el paso Table FACT\_PEGR\_EVOLUTIVO, se mapean los valores obtenidos de las fases anteriores con los respectivos campos a poblar en la tabla FACT\_PEGR\_EVOLUTIVO, y se especifica un flujo alterno ante falla, con la salida de error a un archivo de texto plano utilizando el mismos patrón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



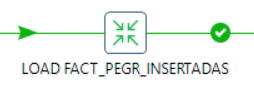
* + FACT\_PEGR\_INSERTADAS: job encargado de ejecutar de forma individual el flujo de limpieza y carga de la tabla de hecho FACT\_PEGR\_INSERTADAS.

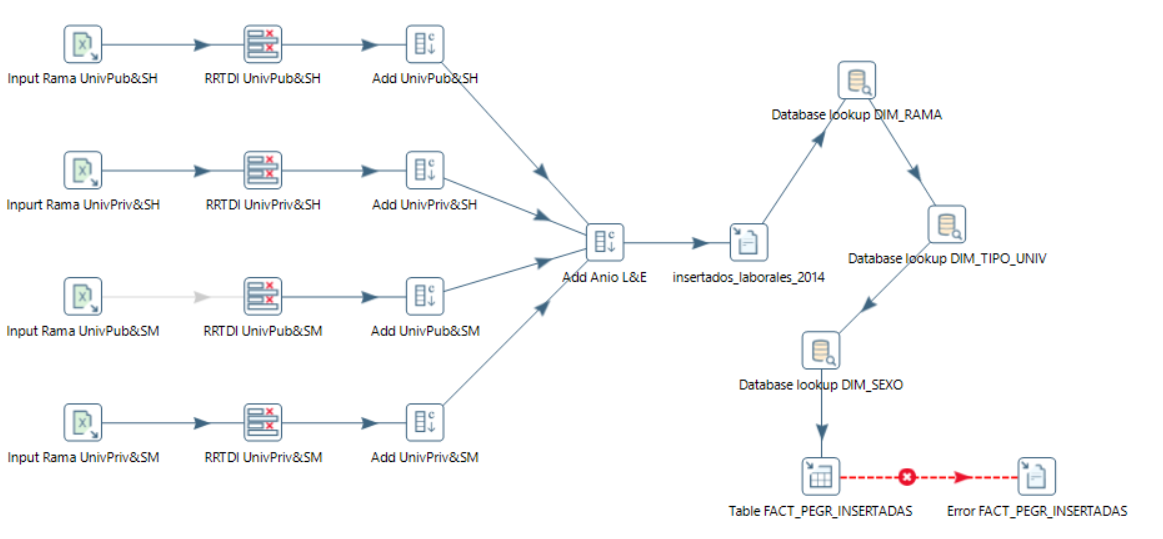


* + ETL\_ FACT\_PEGR\_INSERTADAS: job encargado de limpiar la tabla FACT\_PEGR\_INSERTADAS y ejecutar la transformación LOAD\_ FACT\_PEGR\_INSERTADAS



* + LOAD\_ FACT\_PEGR\_INSERTADAS: transformación encargada de extraer la información de las fuentes y carga 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:

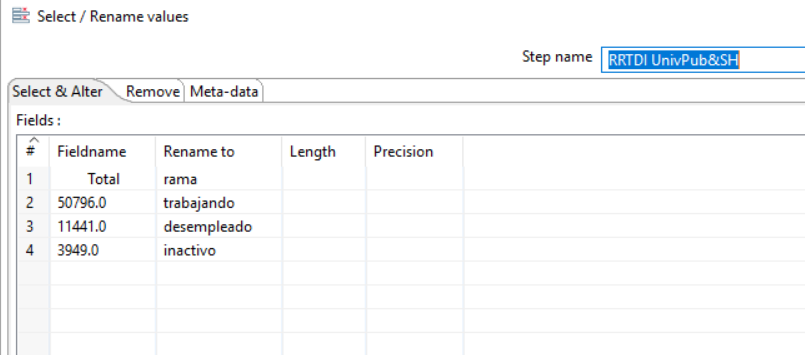
El campo Total, se renombra a **rama**, que corresponde a la descripción de la rama de estudio

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

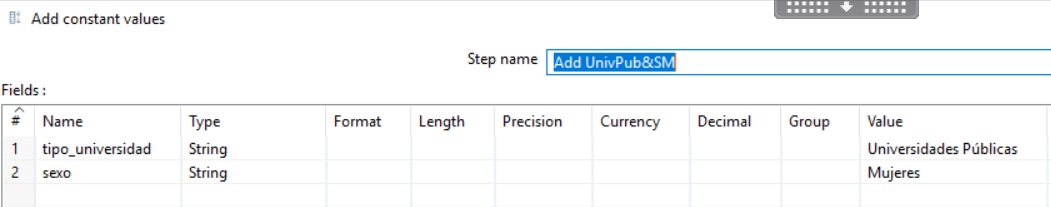
El campo 3 correspondiente al valor numérico sumarizado de la columna Trabajando, se renombra por **trabajando**

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

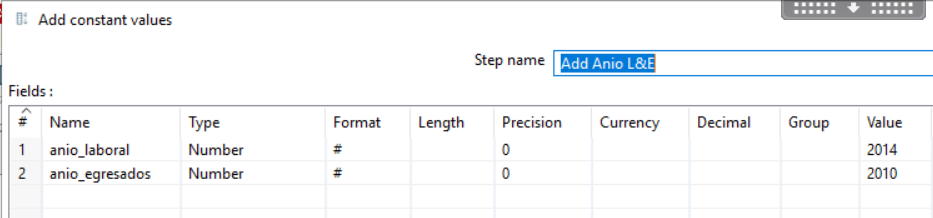
El campo 5 correspondiente al valor numérico sumarizado de la columna Inactivo, se renombra por **inactivo**



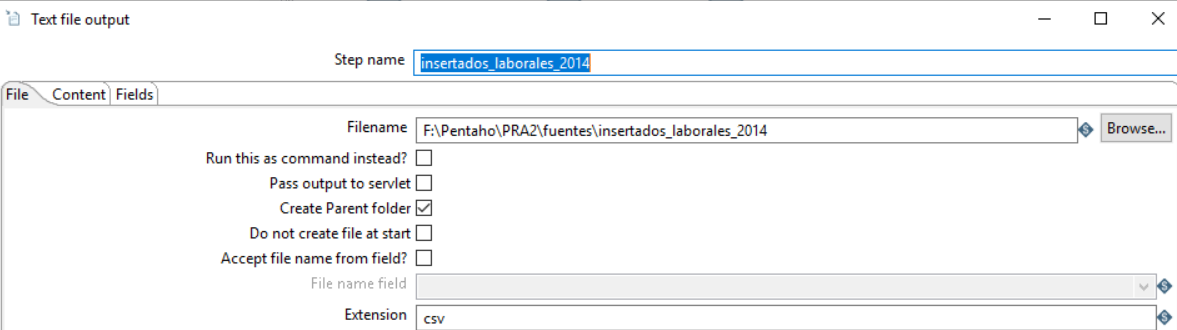
Luego de este paso, se aplican funciones de Add Constant a cada bloque de procesamiento, de manera que se incluyan como campos los valores correspondientes al Sexo y el Tipo de Universidad correspondientes a cada una de las secciones



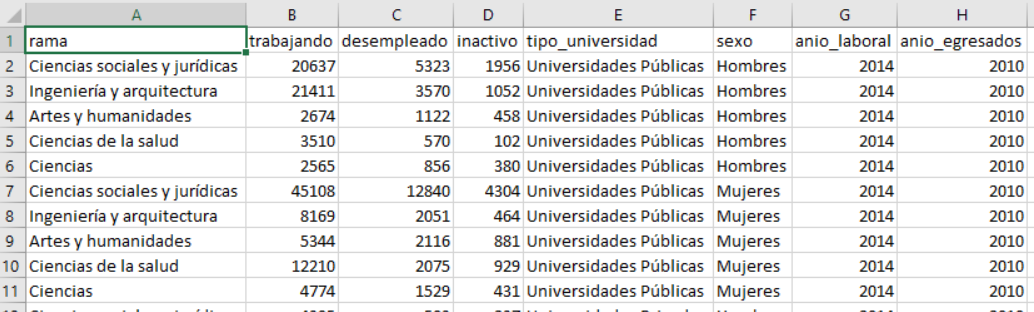
En el siguiente paso se aplica un step más de Add Constant, pero sobre la unión de los 4 bloques de datos, adicionando así el año de inserción laboral 2014 y el año de promoción de los egresados 2010 los cuales son comunes a todos los bloques de información



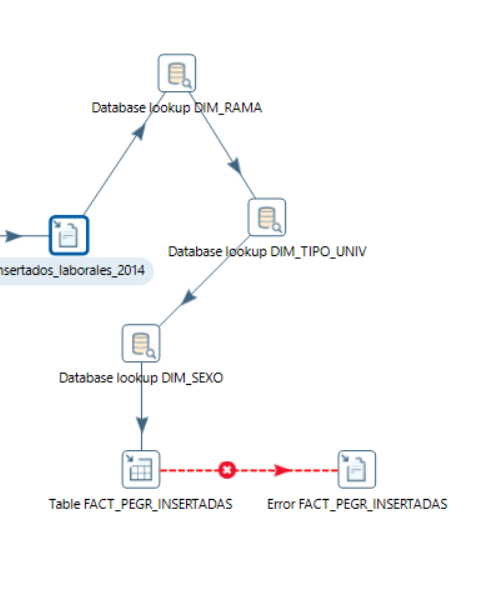
El conjunto de datos obtenido de las transformaciones anteriores, se persiste en un archivo físico insertados\_laborales\_2014.csv, el cual será la base de datos a cargar sobre la tabla de hecho física, previo a las respectivas operaciones de database lookup.



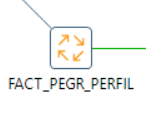
Muestra del archivo csv obtenido

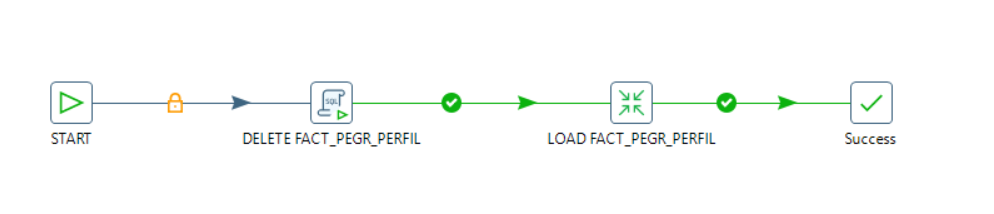


Siguiendo el mismo patrón de los procesos de carga anteriores, aplicamos las operaciones de databse lookup sobre las dimensiones DIM\_RAMA, DIM\_TIPO\_UNIV y DIM\_SEXO para obtener la información de las llaves foráneas a poblar en la tabla de hechos, aplicar las mapeos necesarios sobre la tabla física FACT\_PEGR\_INSERTADAS y por último se establece el control de errores sobre un archivo de texto plano.

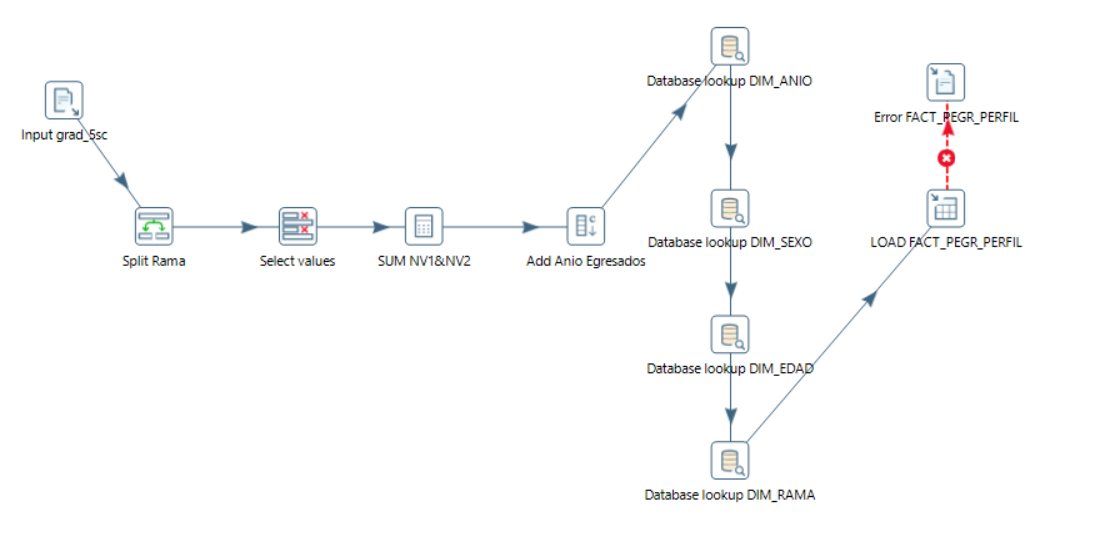


* + FACT\_PEGR\_PERFIL: job encargado de ejecutar de forma individual el flujo de limpieza y carga de la tabla de hecho FACT\_PEGR\_PERFIL.

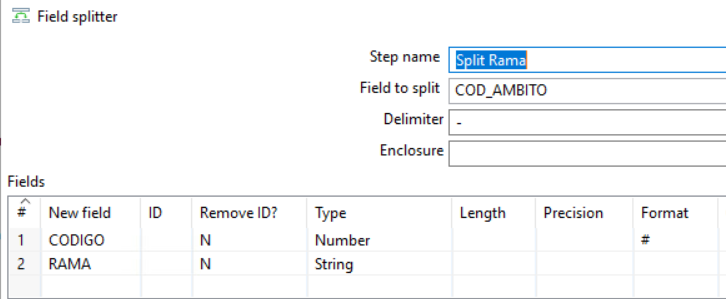


* + ETL\_ FACT\_PEGR\_PERFIL: job encargado de limpiar la tabla FACT\_PEGR\_PERFIL y ejecutar la transformación LOAD\_ FACT\_PEGR\_PERFIL
  + LOAD\_FACT\_PEGR\_PERFIL: transformación encargada de extraer la información de las fuentes y carga de datos en tabla FACT\_PEGR\_PERFIL





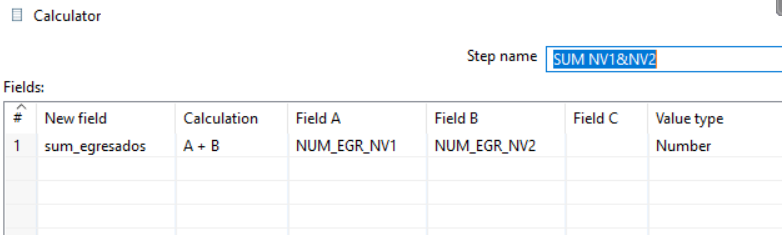
Para la carga de la tabla de hecho, se utilizará la información contenida en el archivo grad\_5sc.xls, el cual tiene la particularidad de que el código y la descripción de la rama de estudio esta concatenada en el campo COD\_AMBITO, por lo tanto, se aplica función de Split para obtener por separado cada una de las variables.



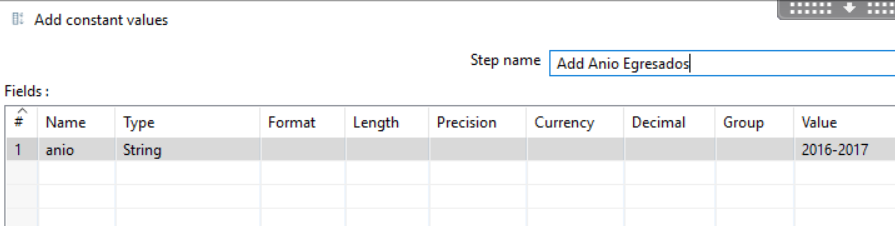
Como se había especificado en la sección de diseño, no se utilizaran valores ‘0’ antecedidos del código de la rama, por lo tanto al aplicar el tipo de dato Number sobre el nuevo campo CODIGO, este es removido automáticamente y en el paso siguiente de Select Values, se cambia su tipo de datos a String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



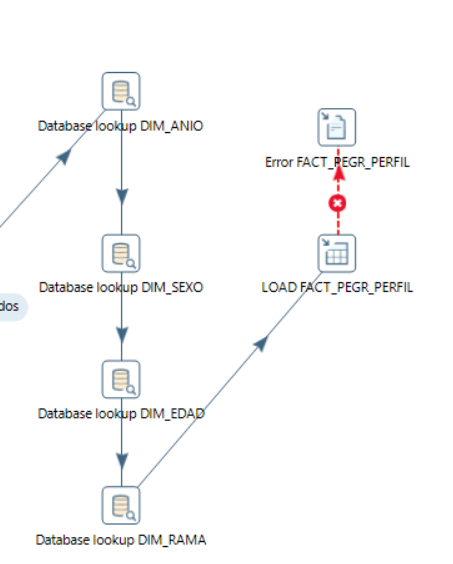
En las especificaciones de diseño se indica que el valor total de egresados corresponderá a la suma de los campos NUM\_EGR\_NV1 y NUM\_EGR\_NV2, por lo que se aplica un step Calulator para obtener dicho valor



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



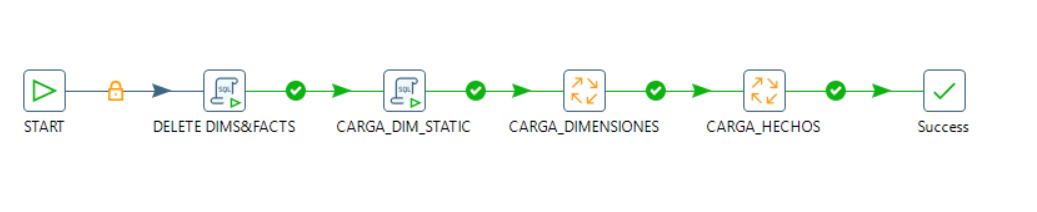
Siguiendo el mismo patrón de los procesos de carga anteriores, aplicamos las operaciones de databse lookup sobre las dimensiones DIM\_RAMA, DIM\_ANIO, DIM\_EDAD y DIM\_SEXO para obtener la información de las llaves foráneas a poblar en la tabla de hechos, aplicar las mapeos necesarios sobre la tabla 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.

## 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 Univerisdad Privada.0 - Header row skipped in file 'F:\Pentaho\PRA2\fuentes\SEGR1.csv'

2019/12/17 23:12:39 - Input Univerisdad 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 Univerisdad Privada.0 - Header row skipped in file 'F:\Pentaho\PRA2\fuentes\SEGR1.csv'

2019/12/17 23:12:43 - Input Univerisdad 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 - Inpurt 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])

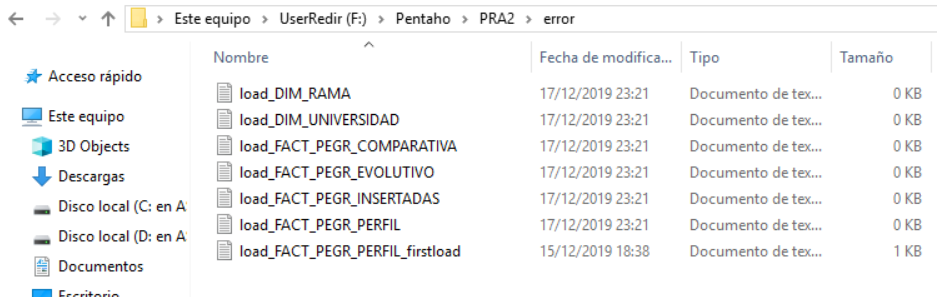
2019/12/17 23:12:46 - CARGA\_INICIAL\_DWH\_EGRESADOS - Finished job entry [CARGA\_DIM\_STATIC] (result=[true])

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.

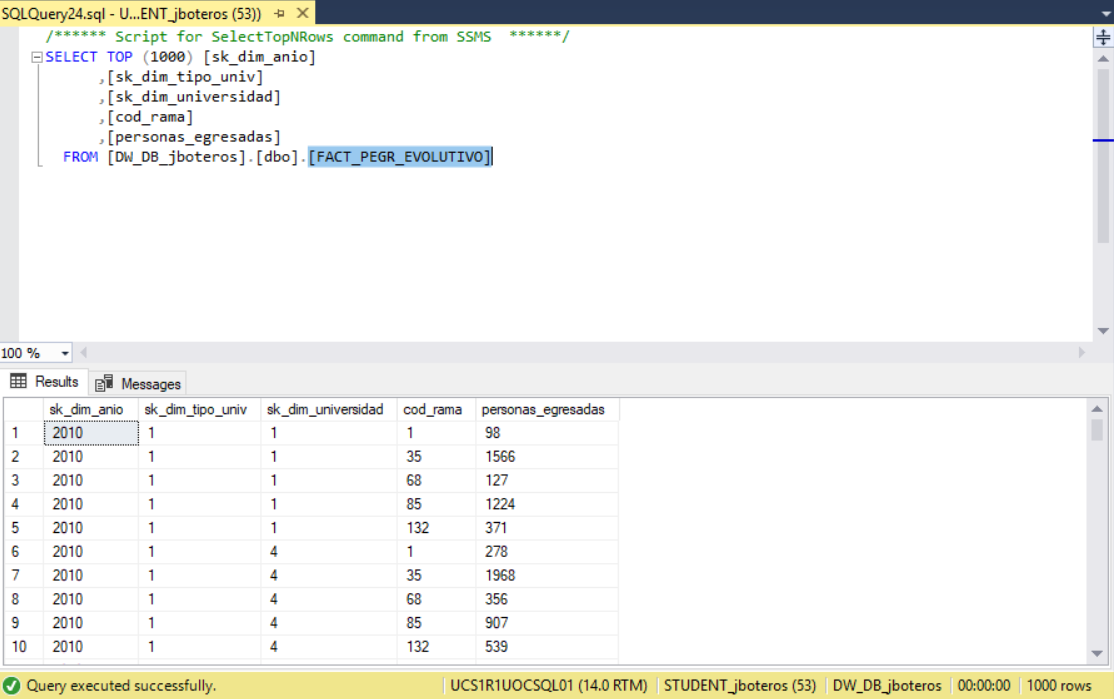
Además podemos observar que en el directorio de errores, todos los archivos de salida tienen 0 bytes, por lo tanto el proceso finalizo OK



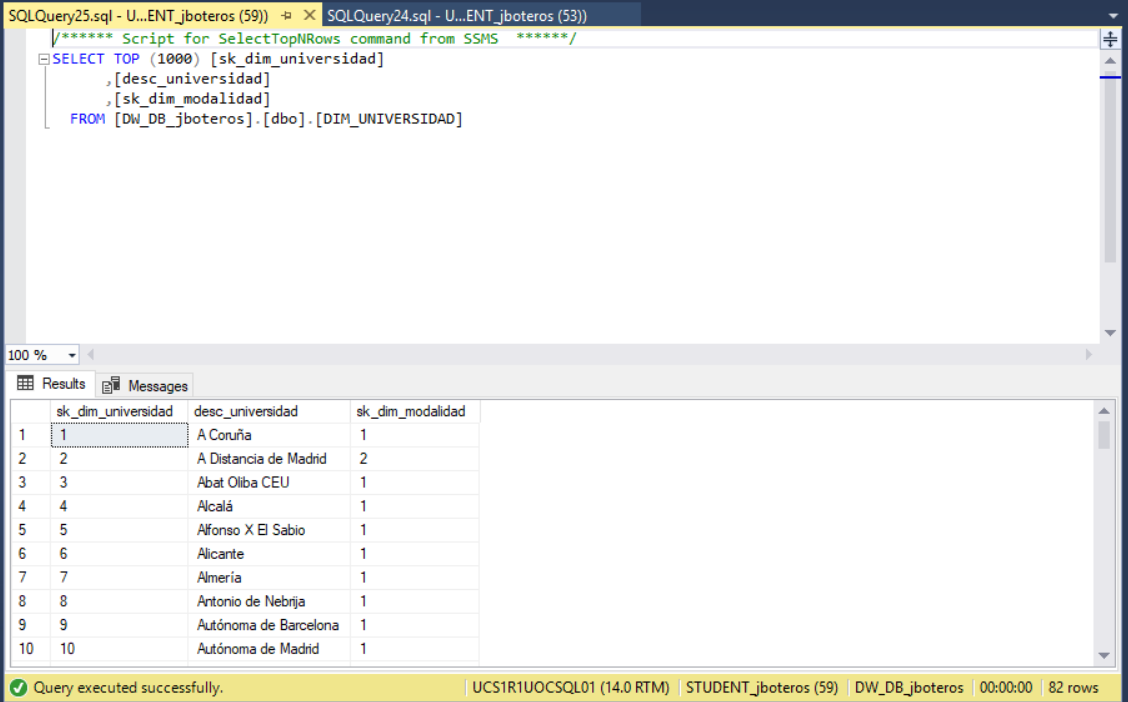
**Comprobación de carga de datos**

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

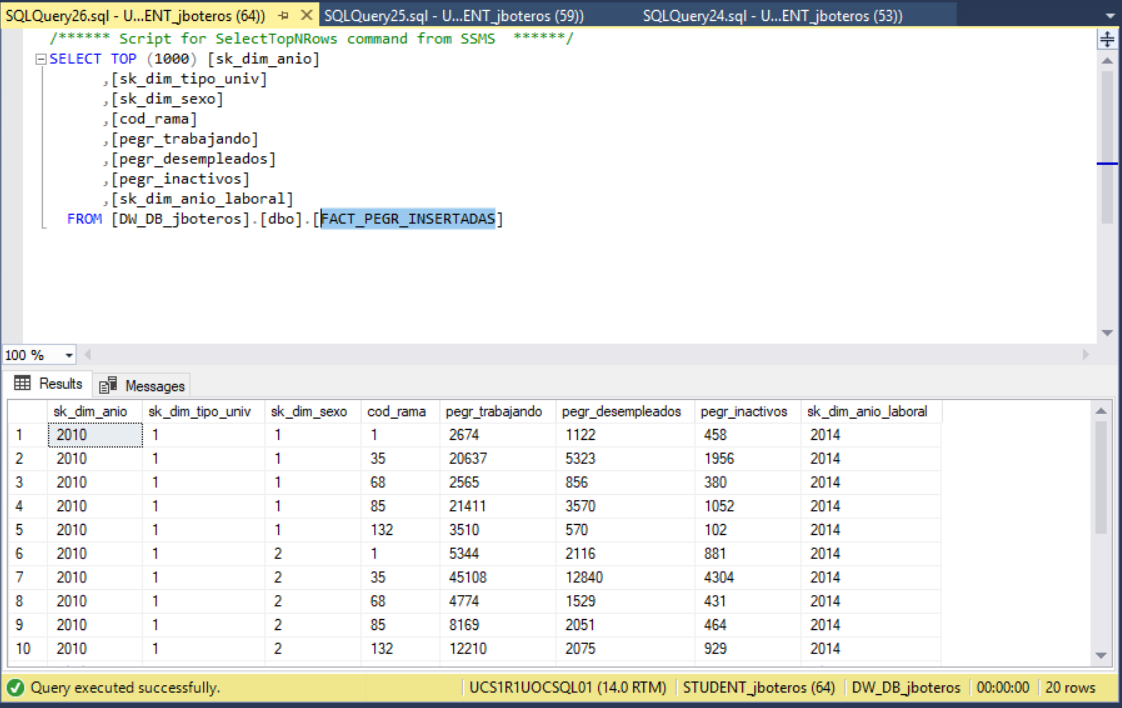
FACT\_PEGR\_EVOLUTIVO



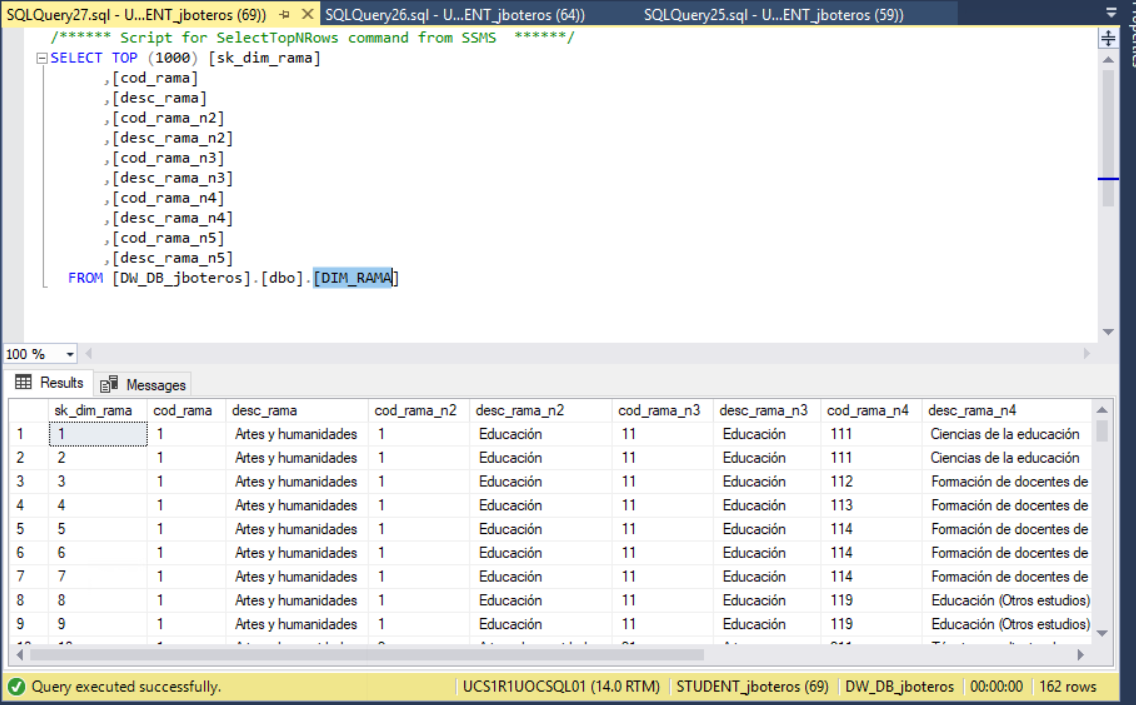
DIM\_UNIVERSIDAD



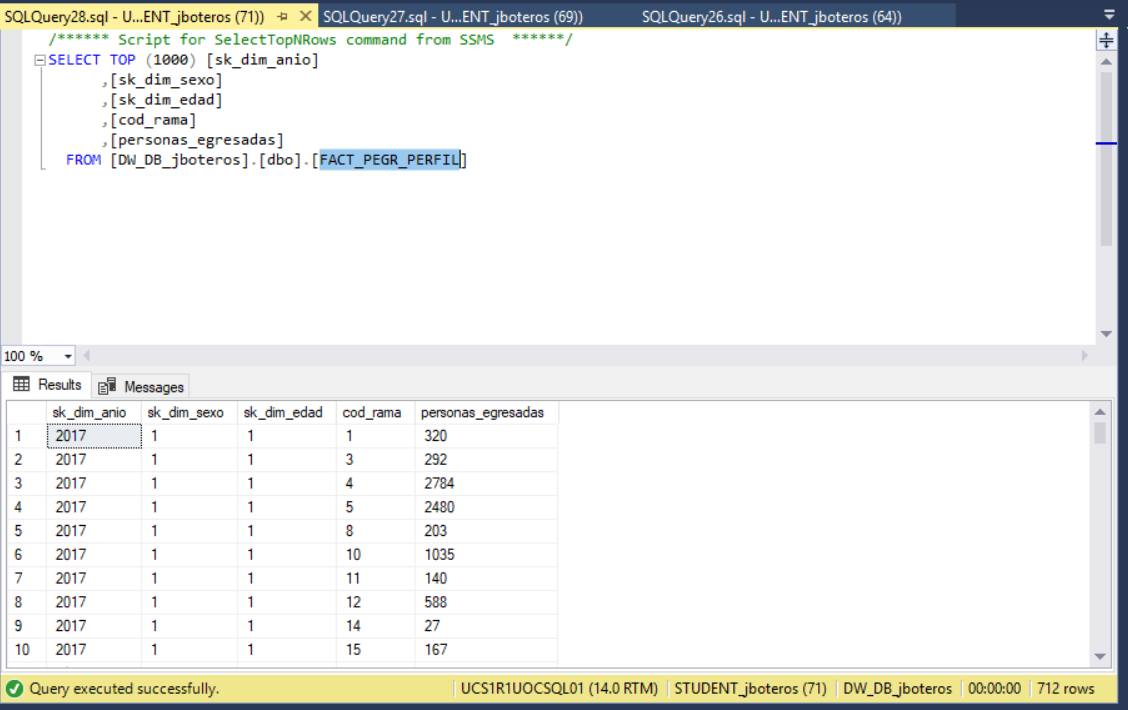
FACT\_PEGR\_INSERTADAS



DIM\_RAMA



FACT\_PEGR\_PERFIL



Explotación de datos

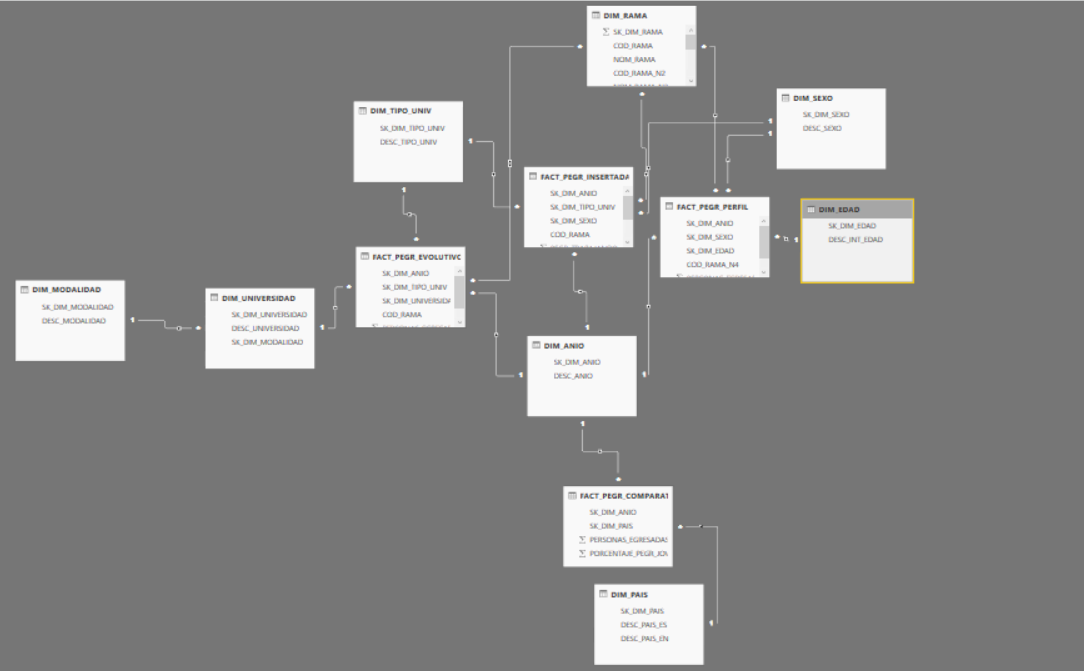
El último paso del proyecto, consiste en diseñar e implementar un modelo OLAP que permita, mediante análisis multidimensionales, responder a las preguntas planteadas en la toma de requerimientos

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:

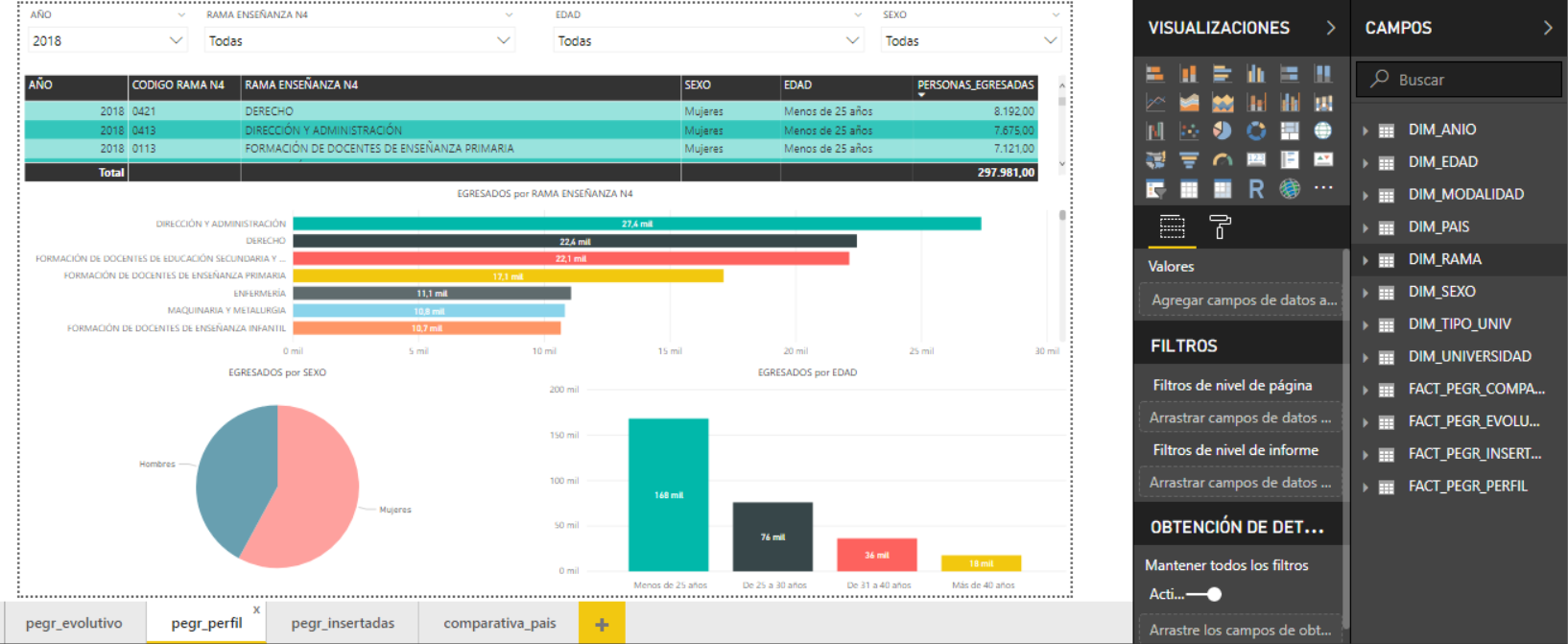
Los informes pueden ser accedidos públicamente en la siguiente URL:

<https://app.powerbi.com/view?r=eyJrIjoiZTY1ZGEwOTUtMTEyNC00YzQzLTllZDEtZjU1NzBmMGFhNzE1IiwidCI6ImFlYzc2MmU0LTNkNTQtNDk1ZS1hOGZlLTQyODdkY2U2ZmU2OSIsImMiOjh9>

EGRESADOS EVOLUTIVO



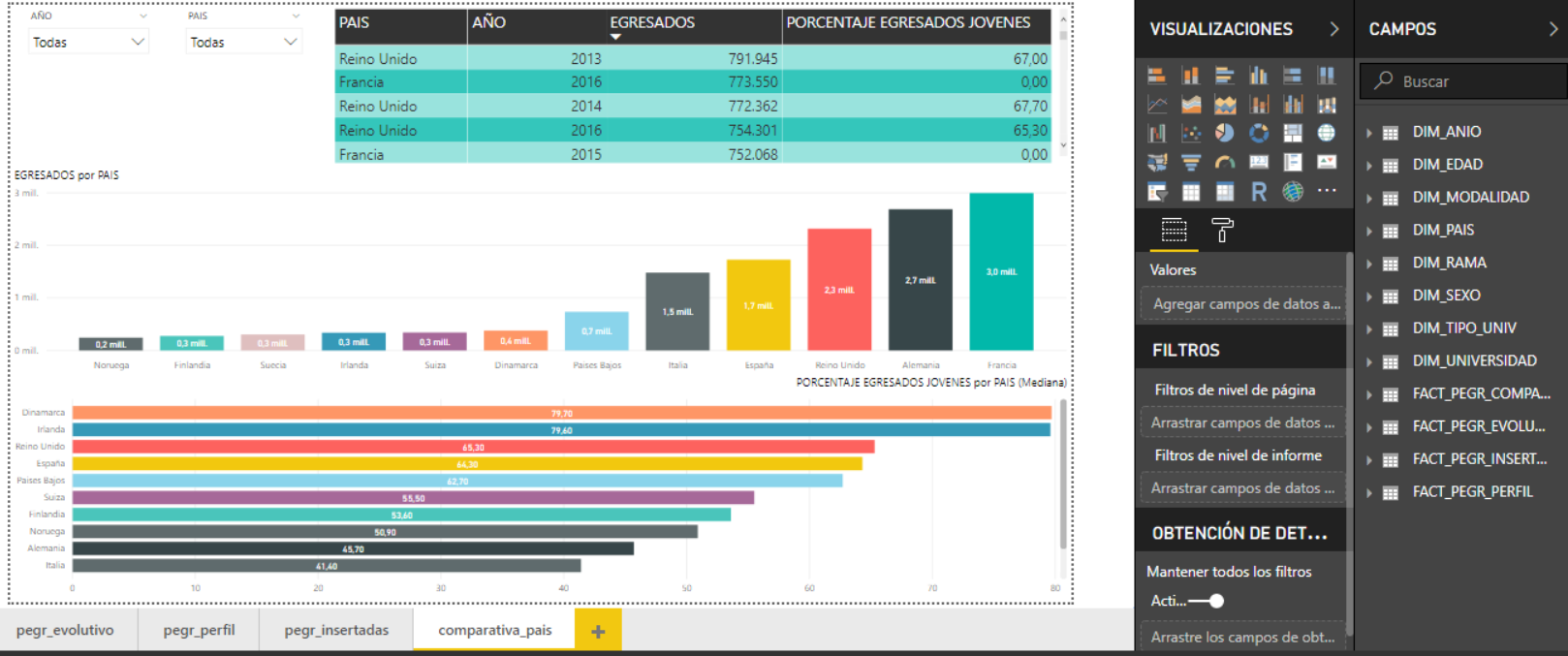
EGRESADOS PERFIL



EGRESADOS INSERTADOS



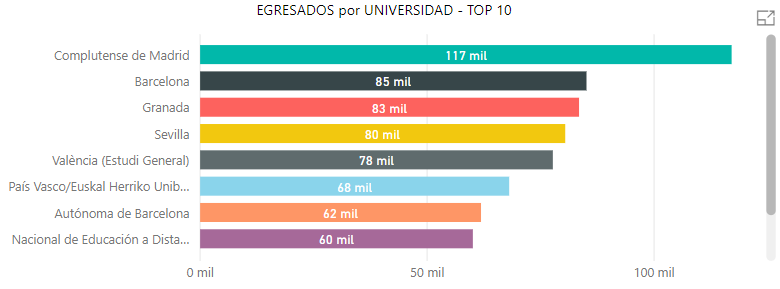
EGRESADOS COMPARATIVA



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

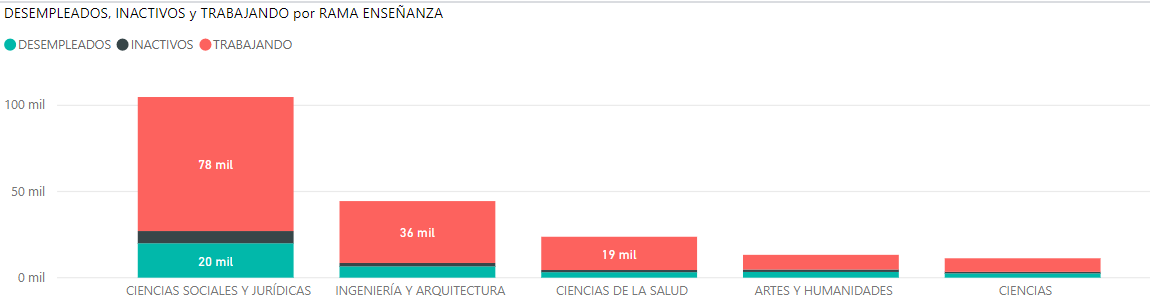
* Top 10 de universidades con mayor número de egresados.

En informe EGRESADOS UNIVERSITARIOS- PERFIL EVOLUTIVO



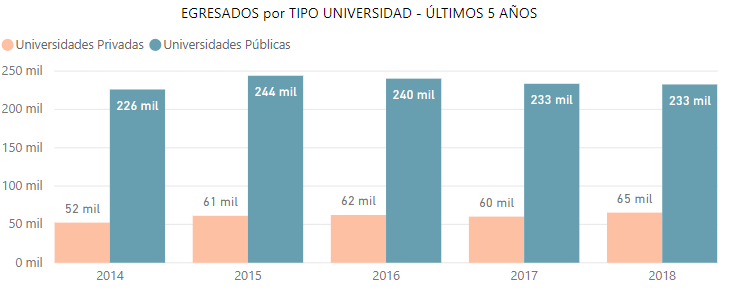
* Ranking de ramas de conocimiento con mayor número de estudiantes egresados insertados.

En informe EGRESADOS UNIVERSITARIOS 2009-2010 INSERCIÓN LABORAL 2014



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

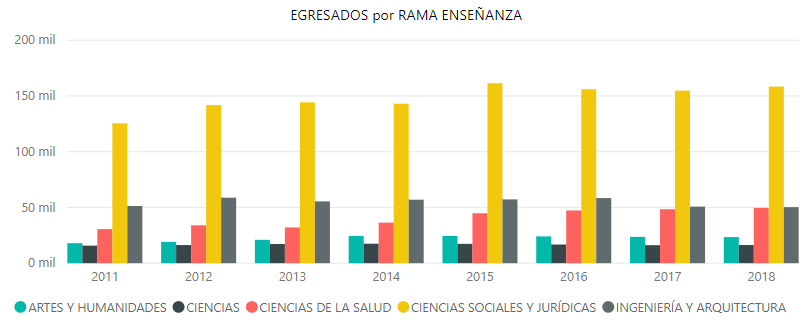
En informe EGRESADOS UNIVERSITARIOS- PERFIL EVOLUTIVO



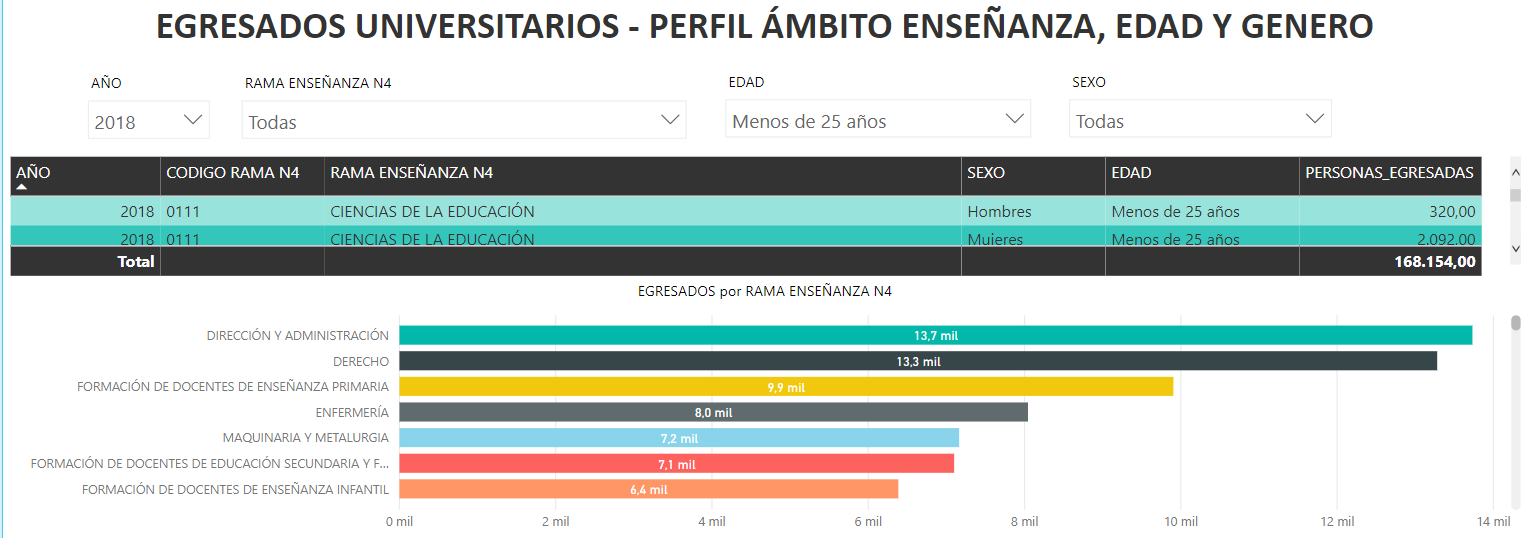
* Evolución del número de egresados universitarios por modalidad de impartición y rama de conocimiento.

En informe EGRESADOS UNIVERSITARIOS- PERFIL EVOLUTIVO



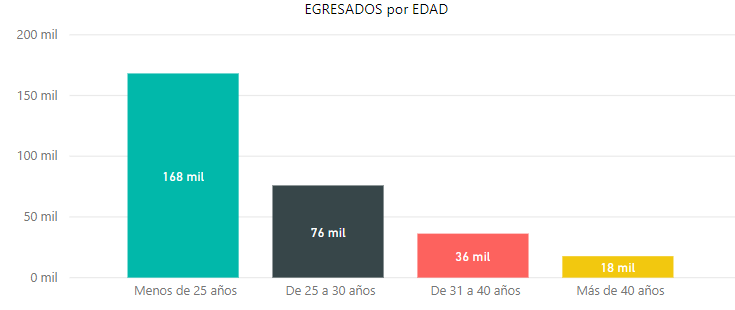


* Ranking de ámbitos de estudios con mayor número de egresados menores de 25 años en el curso académico 2016-2017.

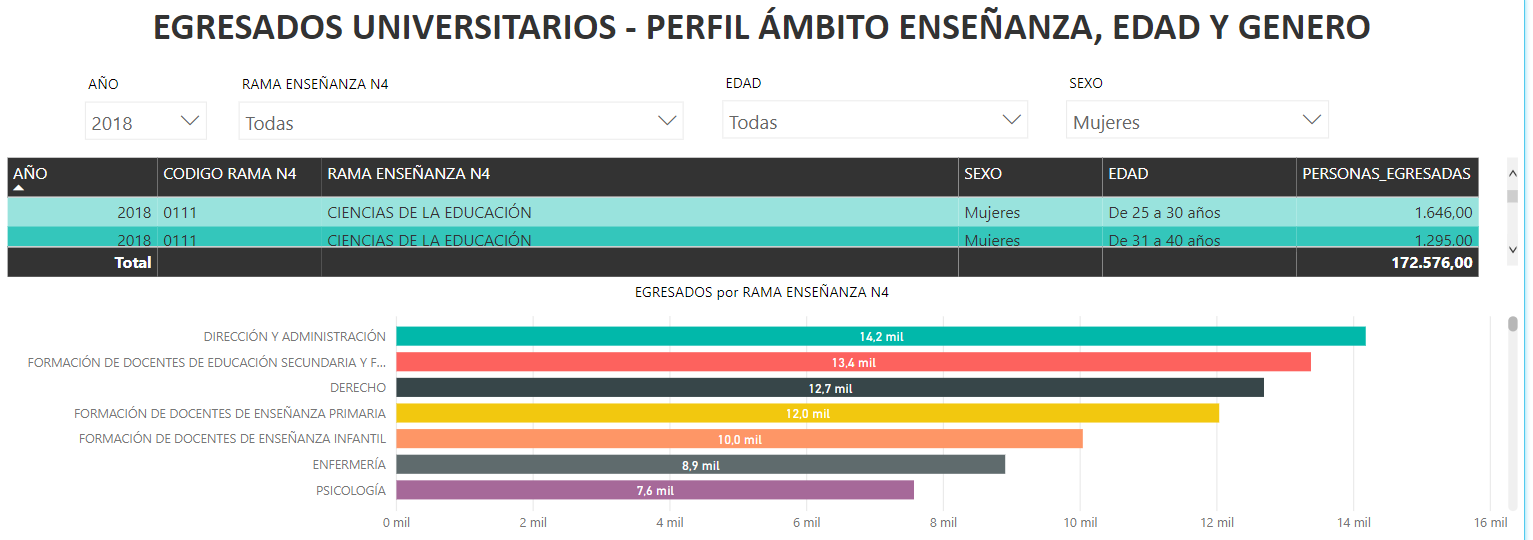


* Ranking de edad con mayor número de personas egresadas en el curso académico 2016-2017.

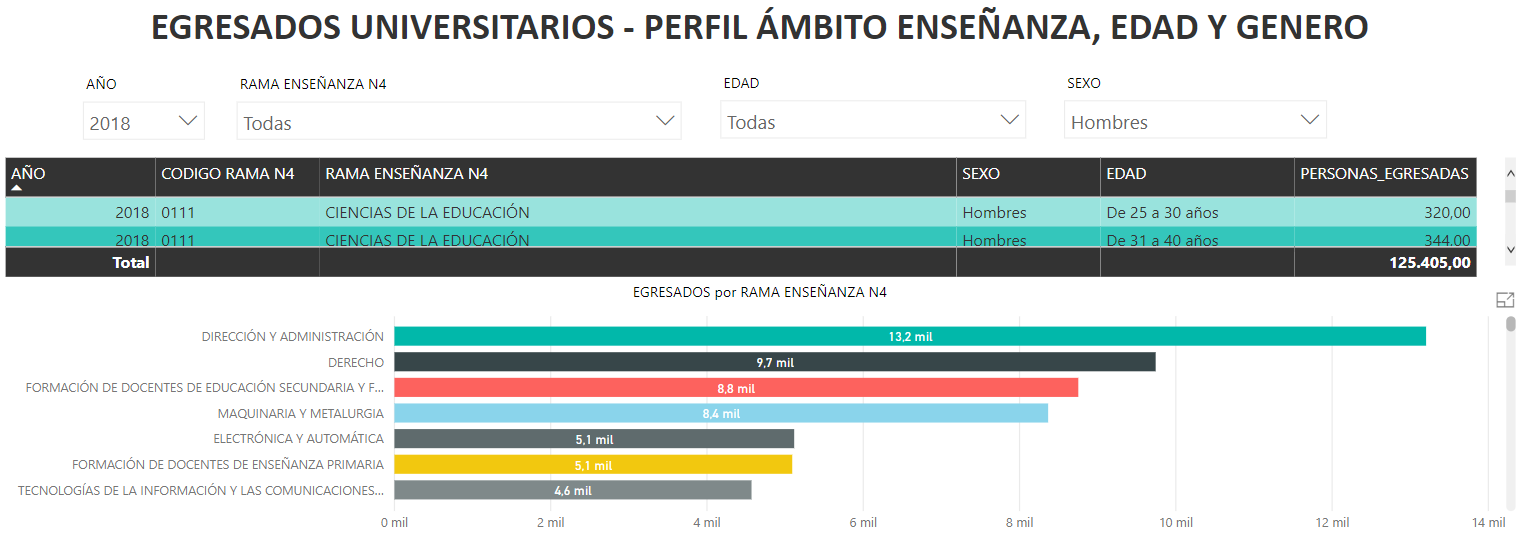
EGRESADOS UNIVERSITARIOS – PERFIL AMBITO ENSEÑANZA, EDAD Y GENERO



* Ámbito de estudios con mayor número de mujeres egresadas.

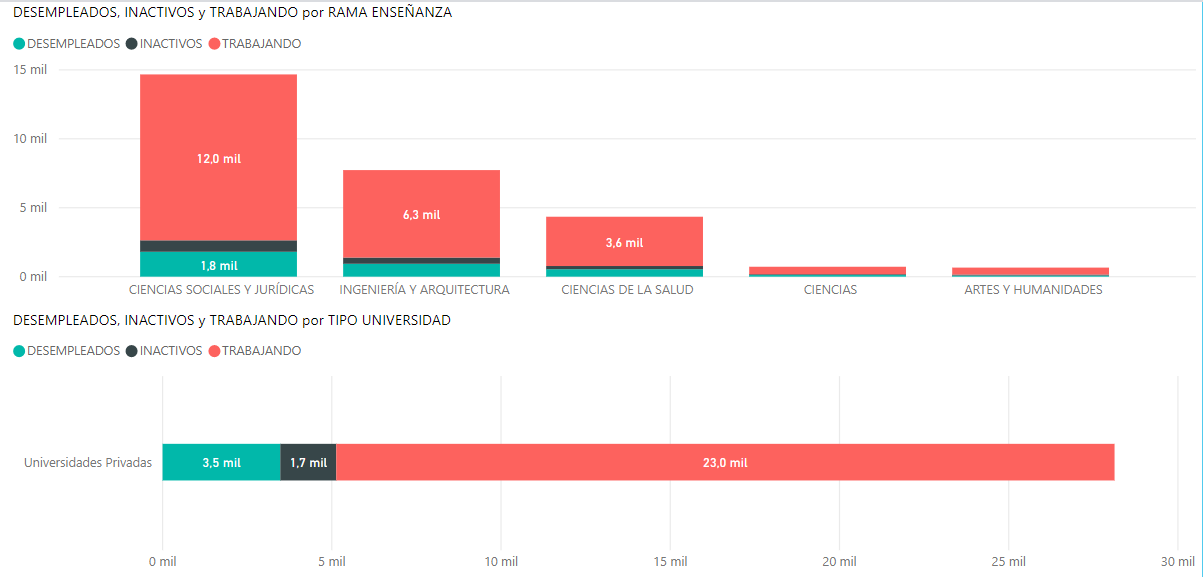


* Ámbito de estudios con mayor número de hombres egresados.



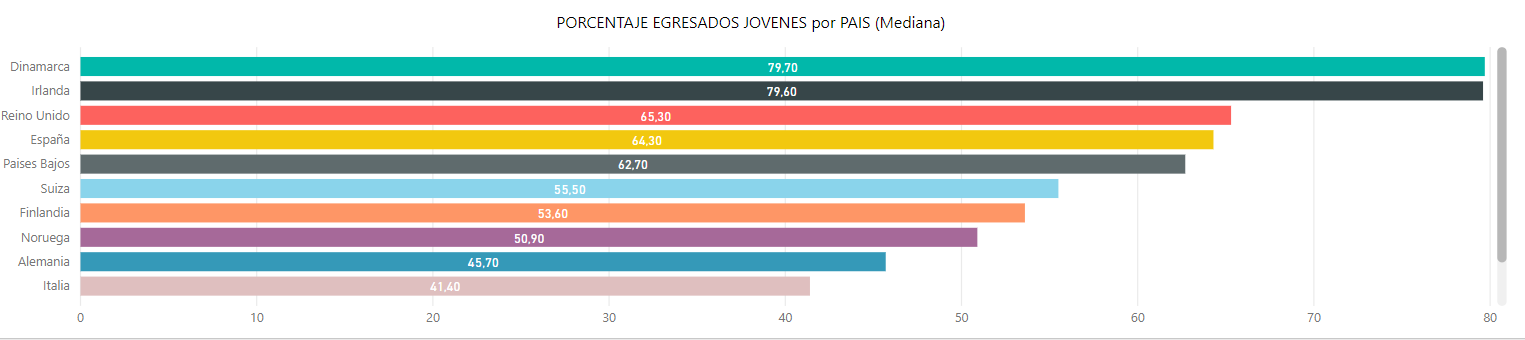
* Tipo de universidad y rama de conocimiento con menor número de estudiantes egresados insertados.

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 (2ª 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-](http://www.mssqltips.com/sqlservertip/1167/getting-started-with-sql-server-analysis-services/) [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-](http://msdn.microsoft.com/en-us/library/ms170208%28v%3DSQL.105%29.aspx) [us/library/ms170208%28v=SQL.105%29.aspx](http://msdn.microsoft.com/en-us/library/ms170208%28v%3DSQL.105%29.aspx)

**Tutorial Pentaho Data Integration:**

[http://wiki.pentaho.com/display/EAI/Latest+Pentaho+Data+Integration+](http://wiki.pentaho.com/display/EAI/Latest%2BPentaho%2BData%2BIntegration%2B%28aka%2BKettle%29%2BDocumentation)

[%28aka+Kettle%29+Documentation](http://wiki.pentaho.com/display/EAI/Latest%2BPentaho%2BData%2BIntegration%2B%28aka%2BKettle%29%2BDocumentation)

1. El diseño se ha realizado utilizando la herramienta Microsoft Visio. [↑](#footnote-ref-1)