Captación y Procesamiento BDD DIEE, Documentación.

Directorio de Empresas y Establecimientos

INEC – 2014/ 12/ 22

Contenido

INTRODUCCIÓN.

El Directorio de Empresas (DIEE) se compone de diferentes bases de datos, entre las principales se tiene: el SRI, IESS, Superintendencia de Compañías, bases de datos con ciertas variables investigadas por el call center del Directorio de Empresas e información obtenida de ciertas encuestas internas del INEC como son: el Censo Económico (CENEC), la encuesta Exhaustiva, ACTI, Ambientales e Industriales.

Esta información es complementada y validada en menor proporción con matrices de equivalencias de variables codificadas de diferentes maneras entre la fuente de informacióny el proveedor.

La información por cada fuente se obtiene de diferentes maneras; es decir que se tiene diferentes formatos o diferentes motores de bases de datos, diferentes modos de transmisión; es por eso que se hace sustancial la intervención de procesos ETL´s que se encargan de transformar a toda la información y llevarla a la lógica definida en el DIEE.

Una vez conseguido que la información este consolidada, el DIEE procede a realizar análisis de la información y posteriormente se realiza una publicación.

El presente documento tiene la finalidad de proporcionar una idea clara de cómo se realiza el proceso de captación y procesamiento de la información para la construcción de la Base de Datos (BDD) del DIEE.

A través del documento se explicará paso a paso cada fase de transformación de la información para tener una base de datos depurada y lista para ser analizada y publicada.

Como se explicó anteriormente cada fuente de información viene al DIEE de diferentes maneras como por ejemplo:

- SRI: Base de datos en Oracle.
- IESS: Base de datos en Archivos de Texto.
- Call Center: Archivos Excel.

- Superintendencia de Compañías: Archivos Excel.

Es por esto que para cada fuente de información se lleva un tratamiento diferente porque además de ser diferentes en formato son diferentes en contenido.

Las herramientas de software con las que el DIEE trabaja son:

- Motor de Base de Datos: PostgresSQL 9.2.
- Heramienta BI: Pentaho Data Integration.
- Oracle Express Edition 10g.
- SQL Power DQguru

Con esta pequeña introducción se da una idea de cómo es la captación y el procesamiento en el DIEE, el cual tiene un orden secuencial para llegar a su objetivo final que es la base de datos depurada.

PROCESO DE CONSTRUCCIÓN DE BASE DE DATOS DEL DIRECTORIO DE EMPRESAS

La construcción de la BDD del DIEE se compone de varias fases, en el Gráfico N: 1 se las expone de manera general:

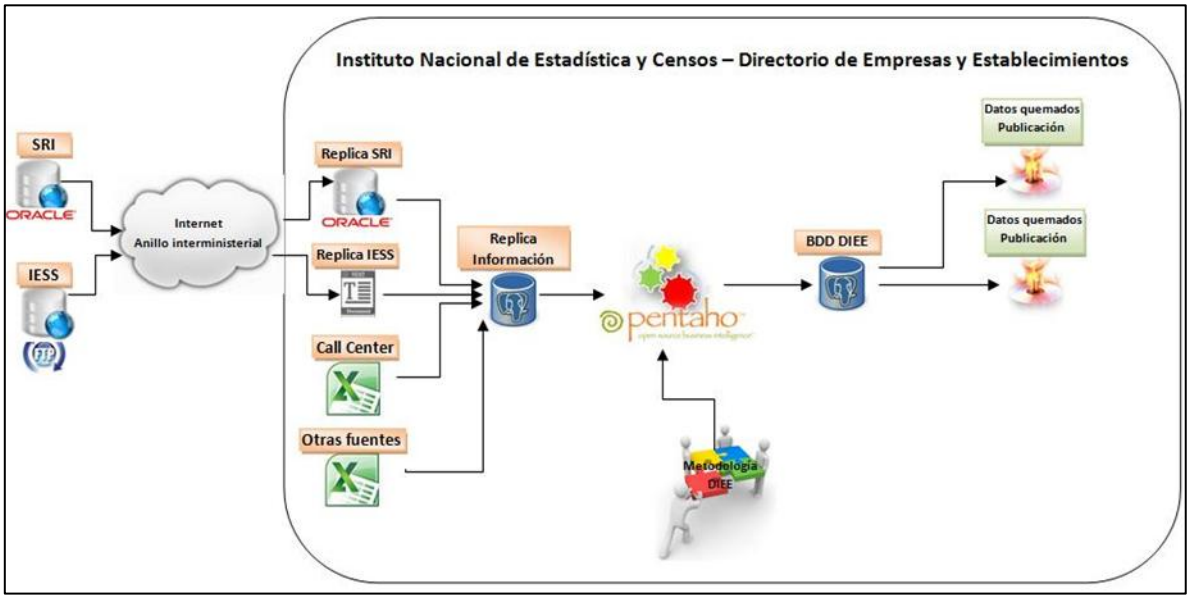


Gráfico N: 1.

El propósito de la captación de información es crear un entorno de Base de datos similar al que se tiene el proveedor de información en su Base, para ello es necesario conocer:

- Medio de comunicación a usar
- Variables a recibir
- Formato de información enviada por el proveedor
- Formato de cada variable enviada
- Volumen de información
- Frecuencia de transmisión.

Una vez identificados con claridad estos datos, se procede a diseñar y desarrollar el mecanismo de transmisión de información; sea este por uso de herramientas propias del motor de Bases de Datos, uso de Herramientas externas para tratamiento de información como ETL´s y pequeños programas con la interacción de aplicaciones como Excel.

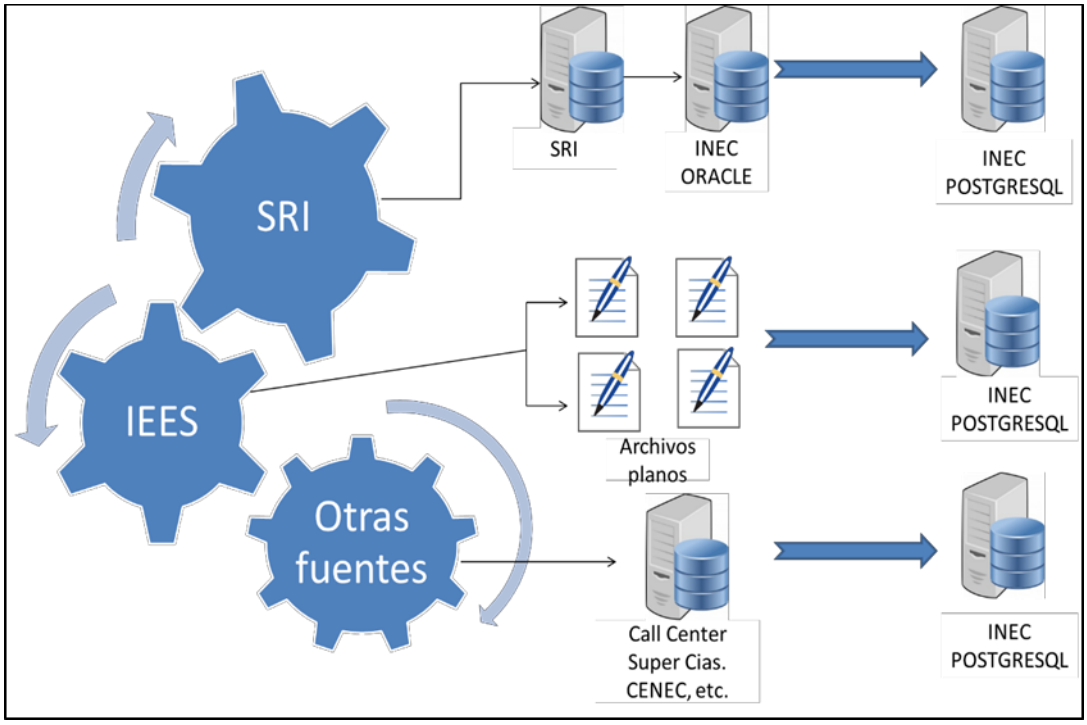


Gráfico N: 2.

Dependiendo de las herramientas usadas para la trasmisión de información generará inconvenientes en el proceso de captación. Es también importante resaltar que los problemas generados en la captación muchas veces no son identificados en el momento mismo en que este se realiza, sino en la fase de procesamiento.

IESS. Información recibida por el Directorio mensualmente, en archivos de texto, cada columna es separada por caracteres especiales, de manera que se generan varios errores al tener más caracteres o menos caracteres separadores.

SRI. La información llega al Directorio diariamente, mediante herramientas del Oracle (vistas materializadas), por lo que la posibilidad de error es mínima. El problema generado a partir de este modo de transmisión de

información requiere de disponer el motor de BDD ORACLE, pero al usar software libre existe limitante de espacio a 5GB.

Call Center. La información es recolectada por las personas que trabajan en el call center a través de un sistema, pero también esta recolección se hace en varis archivos de Excel; Al ser de esta manera existen muchos inconvenientes para la subida de esta información,

Otras fuentes de información. La información es obtenida en archivos de Excel, por lo que no presentan normalización o estandarización de estos datos. La complejidad de subida de información es igualmente alta.

PROCESAMIENTO

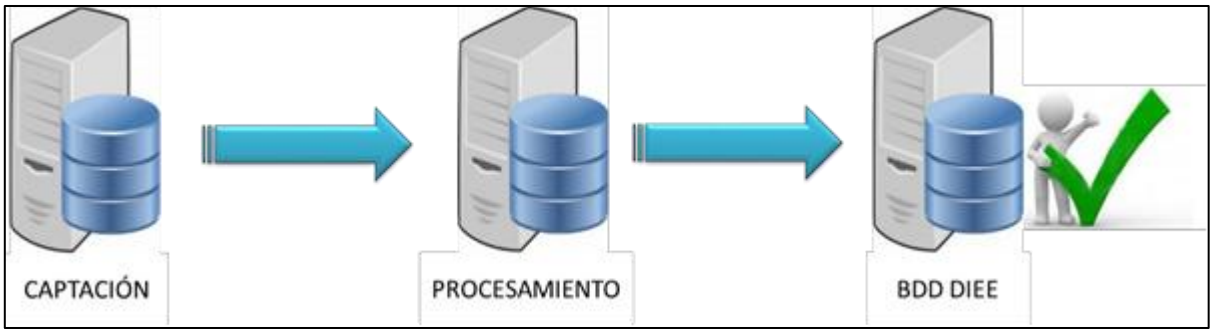


Gráfico N: 3.

Para realizar este proceso es necesario contar con documentos como:

- Matriz de prioridades.
- Matriz de reglas.
- Plan de validación y tabulación.

Estos documentos proporcionan la guía para el desarrollo de los procesos de Extracción de información desde la BDD replicada, Transformación de acuerdo a la lógica que se maneje internamente en la institución y Carga de datos a la estructura de BDD del Directorio de Empresas (ETL).

En los procesos desarrollados en Pentaho, los tres documentos no son plasmados de forma explícita o separada, más si es posible separar los ETL's que gobiernan la migración de cada fuente de información.

Catálogos de información.

En los procesos ETL plasmados en la herramienta de Software Pentaho, existen diferentes grados de dificultad.

Estos a su vez adquieren más complejidad si la variable ha sido revisada por el call center o validada por algún otro medio, ya que en los procesos debe también considerarse actualizaciones solo para registros de menor prioridad de fuente de información.

Actividad Económica.

Una de las variables que presenta mayor complicación es la actividad económica, esto debido a algunas consideraciones:

- El SRI maneja diferente versión de CIIU
- Existen muchos códigos CIIU3 que no tienen mapeo directo a CIIU4; tienen una relación de un código CIIU3 a varios códigos CIIU4.
- El repositorio en el que reporta el SRI, la información de actividad económica a nivel de Establecimiento, no permite la identificación de una actividad económica principal por cada establecimiento.

Para tratar estos inconvenientes se han creado varias reglas en el DIEE, y muchas de ellas a partir del caso que se ha presentado.

- Para generar el mapeo de CIIU3 a CIIU4 se ha solicitado la matriz de equivalencias de estas 2 versiones de CIIU.
- Se ha definido pasar la actividad económica principal de empresa a establecimiento, para el caso de contribuyentes con establecimientos únicos.
- Se han creado matrices de equivalencia a diferentes niveles.
- Al no existir un campo que señale la actividad económica por cada uno de los establecimientos, se ha definido un valor ordinal para las actividades económicas y asignando a la primera actividad como la actividad del establecimiento. 1

Direcciones

La estructura que al momento presenta la BDD del DIEE, adaptada de acuerdo a sus necesidades; es así que tenemos 2 repositorios para este fin, en los cuales se almacena la siguiente información:

- Datos generales de la dirección y se distingue el estado, la fuente, origen.
- Detalles de la dirección: calle principal, calle secundaria, número, etc.

Estos datos igualmente son validados y actualizados dependiendo de la fuente de información.

Fechas de actualización o cambios.

Se realiza actualización de las variables recibidas del SRI dependiendo de la fecha de actualización y tomando en cuenta la matriz de prioridades. Tomando en cuenta que en contribuyentes es el único que indica la fecha de actualización, esta fecha es tomada para la actualización de establecimientos.

Con la actualización de variables, se actualiza también las variables de control.

Existen también registros con fechas de cierre y/o fechas de cese superior a la fecha de corte, pero estos datos no son considerados, ya que a la fecha de corte estos tienen aún el estado anterior.

Empleados y Ventas

Esta información es particular, ya que existe de varios años por cada empresa en el caso de ventas, y de empleados existe tanto a nivel de empresas como de establecimientos.

Para el caso de empleados, al reportar la información mensualmente e indicar los datos diariamente por cada establecimiento, estos deben ser promediados previo a la subida de información a la BDD de establecimiento, mientras que a la BDD de empresa sube la información como una suma. En lo referente a información de subida a establecimientos, se llega a concentrar la información en establecimientos

matriz de aquellos cuyo número de establecimiento no concuerde con lo expuesto en la BDD del DIEE.

En el caso de ventas solo se realiza un filtro de la información de ventas, para subirla cada año.

Medios de comunicación

Los medios de comunicación han recibido tratamiento (transformación) de acuerdo a lo expuesto en el manual de tabulación.

La BDD de medios de comunicación principalmente recibe información de la BDD del CENEC y SRI y en mínima proporción del CALL CENTER. En este punto ha sido necesario descartar registros al no acogerse a las reglas que deben cumplir los teléfonos para ser tomados en cuenta.

Registros nuevos

Los registros nuevos obtenidos del corte anual, se ingresan con normalidad, sin recibir actualización alguna; excepto la actualización solicitada bajo demanda (fuente de validación CALL CENTER).

DETALLES DE PROCESAMIENTO

La herramienta Pentaho es la que juega uno de los roles más importantes en esta fase debido a que aquí se trabaja con procesos ETL´s.

ETL: Siglas en inglés que significan: Extraer, Transformar y Cargar, por ello se dice que un ETL es el proceso que permite mover datos desde múltiples fuentes, limpiarlos y cargarlos en otra base de datos, data mart, o data warehouse para apoyar un proceso de negocio.

Job: Es el conjunto de objetos que conforman una transformación.

Gráfico N: 4.

El siguiente gráfico muestra de manera general como se realiza las "transformaciones" en Pentaho. Es necesario indicar que se tienen:

- Job. Puede tener "n" cantidad de transformaciones
- Transformaciones. Puede tener "n" cantidad de scripts y objetos para realizar cambios y adaptaciones de información.
- Scripts. Líneas de código creadas con un propósito especial

A continuación se indica el aspecto que presenta el JOB en la herramienta de software:

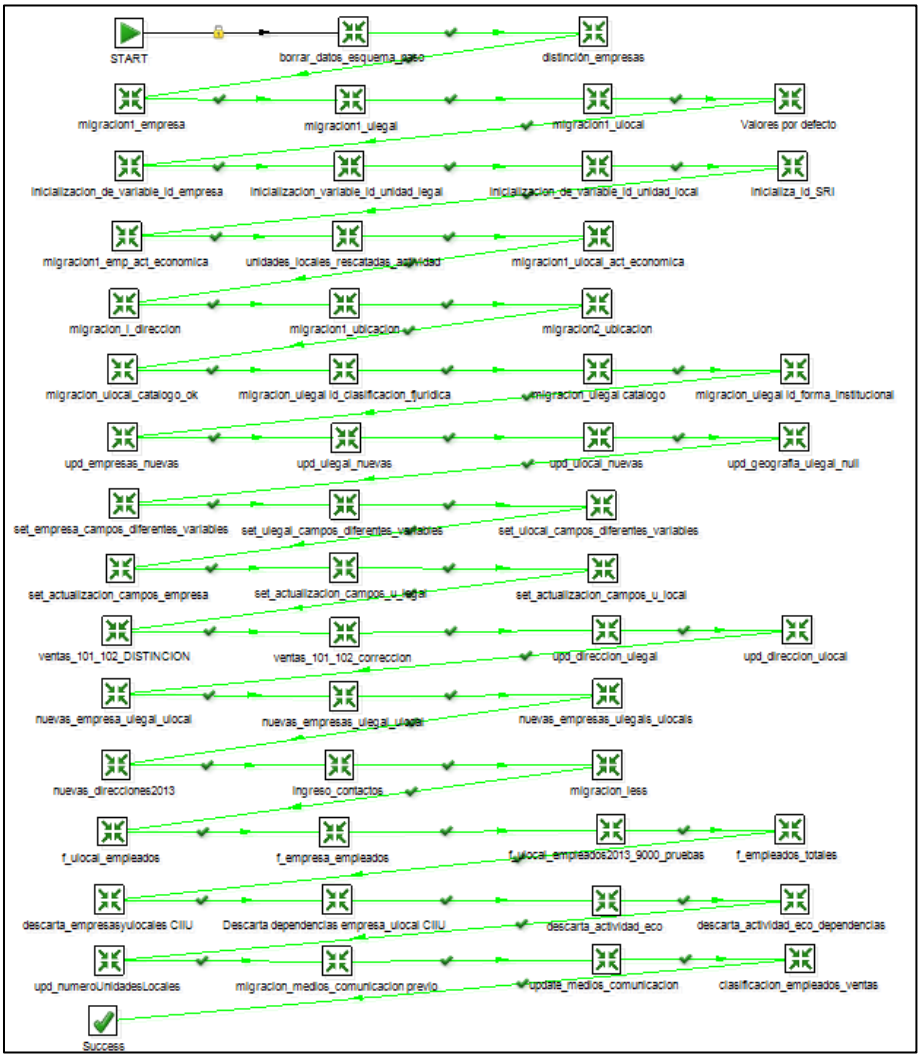


Gráfico N: 5.

En esta sección se irá explicando que se ejecuta en cada transformación y cuál es su finalidad.

Para comenzar a trabajar con las transformaciones, previamente se crea un esquema alterno PASO que contiene las tablas principales de DIEE como son f_empresa llamada en "paso" como i_empresa, i_unidad_local, i_unidad_legal, etc, su objetivo es actuar como puente de la información antes de llegar a la base final, ya que existen transformaciones que no se pueden ejecutar directamente en la base final.

1. **Start.**

Gráfico N: 6.

El objeto "START" tiene como objetivo darle comienzo a la ejecución de todas las transformaciones.

2. **Borrar datos del esquema paso.**

Gráfico N: 7.

En el DIEE se crea una base de datos alterna PASO donde se va a trabajar y se ejecutarán todos los cambios, esto con la finalidad de pasar a la base de datos oficial ya los datos reales y sin fallos.

Dentro de la transformación (Grafico N: 7) existen los siguientes objetos:

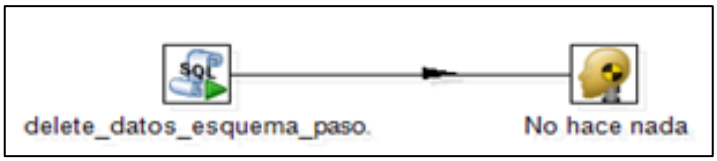


Gráfico N: 8.

El objeto “delete_datos_esquema_paso ejecuta script´s donde borra la información de las tablas:

- paso.i_empresa.
- paso.i_unidad_legal.
- paso.i_unidad_local.
- paso.i_direccion.
- paso.i_ubicacion_direccion.

3. Distinción de empresas.

Gráfico N: 9.

Dentro de esta transformación existen los siguientes objetos:

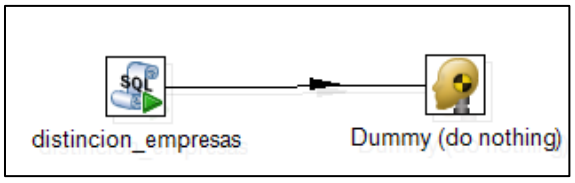


Gráfico N: 10.

El objeto “distinción_empresas” tiene como objetivo marcar las empresas nuevas que ingresarán al directorio y las empresas que ya existen en el mismo.

Migración de tablas principales

En la migración de las tablas principales intervienen las transformaciones:

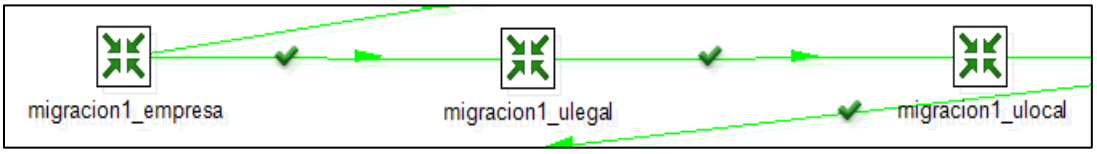


Gráfico N: 11.

Las que se encargan de migrar determinadas variables de las tablas: ruc_contribuyentes y ruc_establecimientos de la fuente SRI a las tablas: i_empresa, i_unidad_local, i_unidad_legal del esquema PASO.

4. Migracion1 empresa.

Gráfico N: 12.

Dentro de esta transformación existen los siguientes objetos:

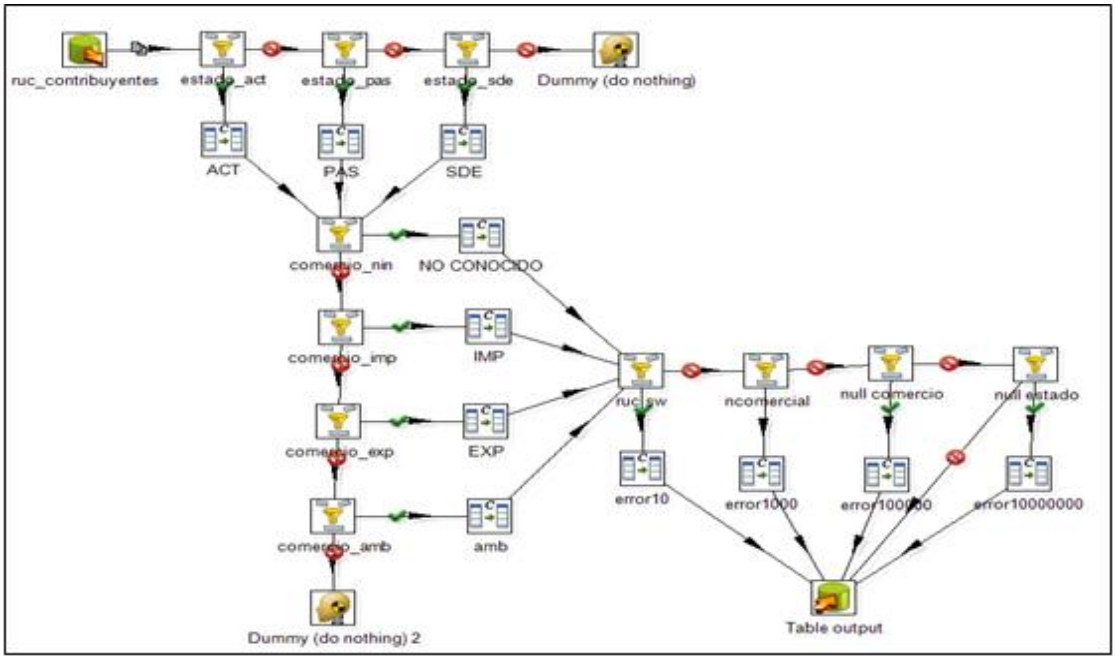


Gráfico N: 13.

Esta transformación tiene como principal objetivo el migrar los datos desde la fuente SRI a partir de la tabla ruc_contribuyentes hacia la tabla i_empresa del esquema PASO, con los debidos cambios como por ejemplo:

Se transforma el catálogo de estados de empresa que tiene el SRI al catálogo que tiene el DIEE y de la misma manera se hace para comercio exterior. Quedando de la siguiente manera:

SRI		DIEE
ESTADO_PERSONA_NATURAL	ESTADO_SOCIEDAD	ID_EMPRESA_ESTADO
ACT	ACT	1
PAS	PAS	2
SDE		3

SRI	DIEE
COMERCIO_EXTERIOR	ID_ACTIVIDAD_COMERCIO_EXTERIOR
NULL	99
IMP	01
EXP	02
AMB	03

- Se valida que nombre comercial tengan una longitud de mínimo tres caracteres.
- Cuando las trasformaciones encuentran que hay valores que no existen como por ejemplo en los catálogos o el nombre tiene menos de tres caracteres les cataloga con error a las empresas y se las tiene identificadas, se emite un reporte para su posterior análisis.

5. Migracion1 ulegal.

Gráfico N: 14.

Dentro de esta transformación existen los siguientes objetos:

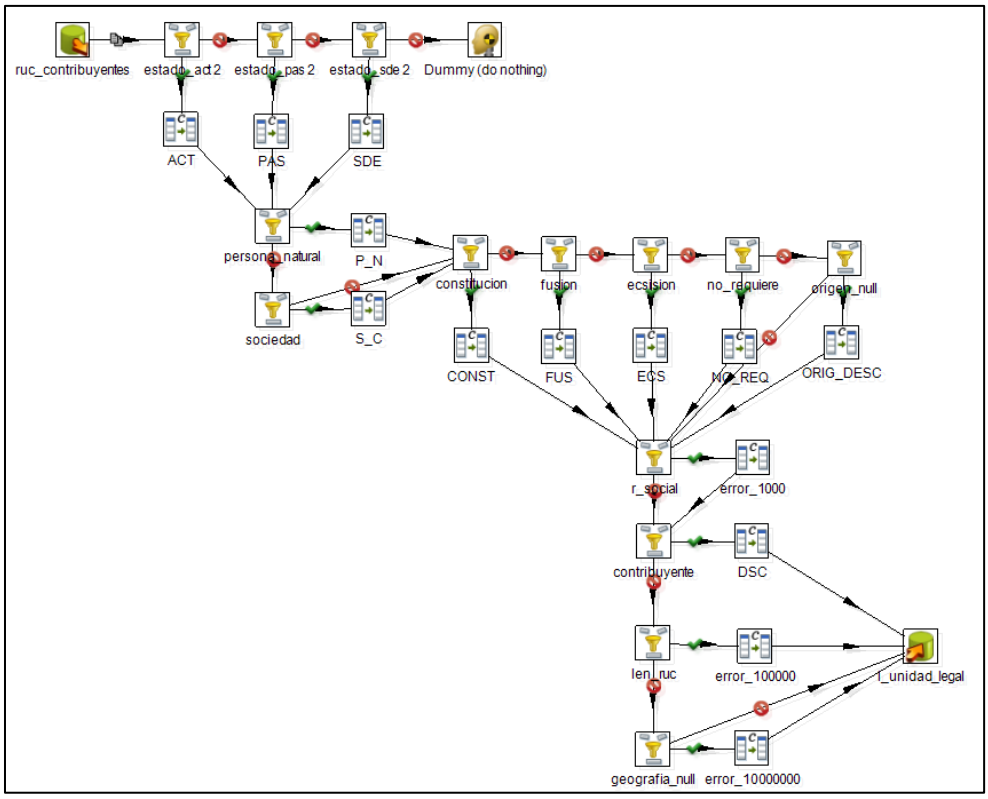


Gráfico N: 15.

Al igual que la migración de empresa esta busca pasar los datos que corresponden a la parte legal que está en la tabla ruc_contribuyentes del SRI, a la tabla i_unidad_legal del esquema PASO.

En este proceso ETL tenemos objetos que transforman y validan información:

- Transforma los estados de empresas.
- Transforma el acto jurídico.
- Valida la clase de contribuyente.
- Pasa la información del expediente.
- Valida el largo de la razón social y del ruc.
- De la misma manera se marca a las empresas cuando tengan algún tipo de error, y se procederá a reportar cuales son los errores encontrados.

6. Migracion1 ulocal.

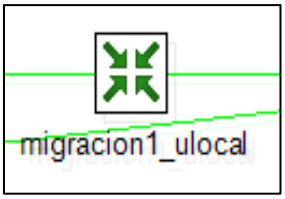


Gráfico N: 16.

La transformación de unidad local tiene los siguientes objetos:

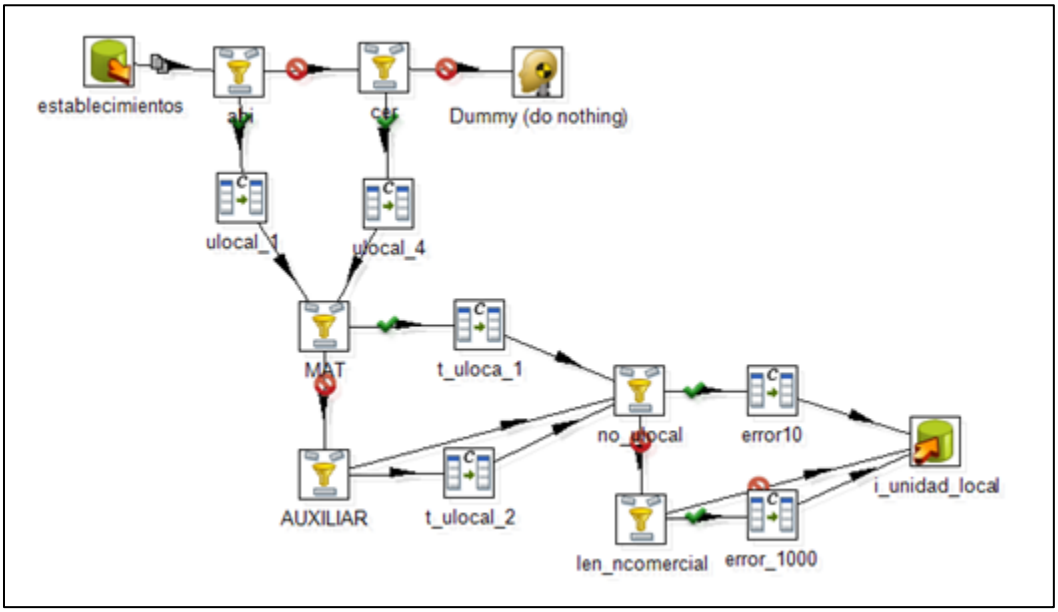


Gráfico N: 17.

Lo que se busca en unidad local es:

- Validar el número de la unidad local.
- Transformar el estado de la unidad local.
- Transformar el tipo de unidad local.
- Validar el largo del nombre comercial.
- De la misma manera los errores que bote la transformación se los procederá a reportar.

7. Valores por defecto.

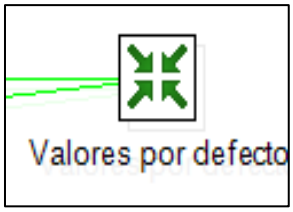


Gráfico N: 18.

La transformación de valores por defecto tiene los siguientes objetos:

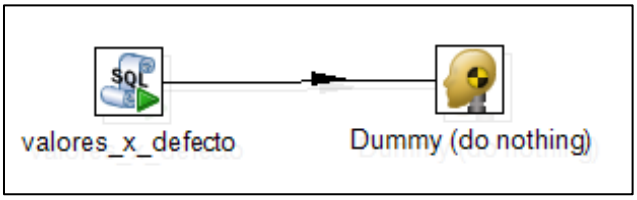


Gráfico N: 19.

La transformación busca llenar los valores que se han definido en el Plan de Validación y tabulación como los "valores por defecto", es decir:

Por ejemplo si en la variable "Nombre comercial", el valor es nulo, esta transformación automáticamente llena ese campo con el valor "-1", así para variables como: fechas, actividad secundaria, producto elaborado, etc. Todas dependiendo del valor que se haya definido en el Plan de Validación y Tabulación.

Inicialización de variables.

En la inicialización de variables intervienen las transformaciones:

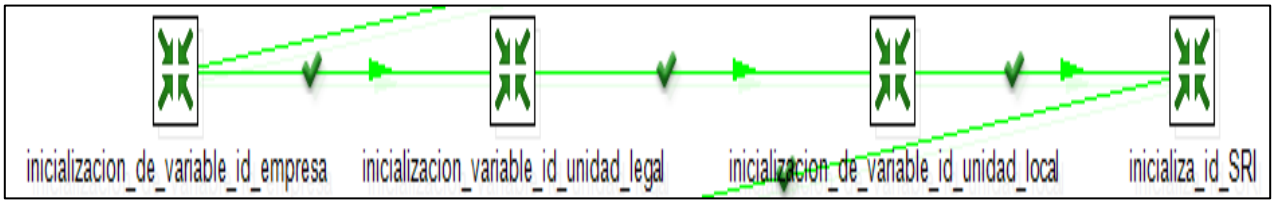


Gráfico N: 20.

Las que se encargan de llenar la información de los códigos (id) de las tablas principales del esquema PASO y de la fuente SRI:

- i_empresa.
- i_unidad_legal.
- i_unidad_local.
- ruc_contribuyentes.
- ruc_establecimientos.

Respectivamente, a partir de los datos de la base del DIEE.

8. Inicialización de variable id_empresa.

Gráfico N: 21.

La transformación de inicialización de variables id empresa se compone de los siguientes objetos:

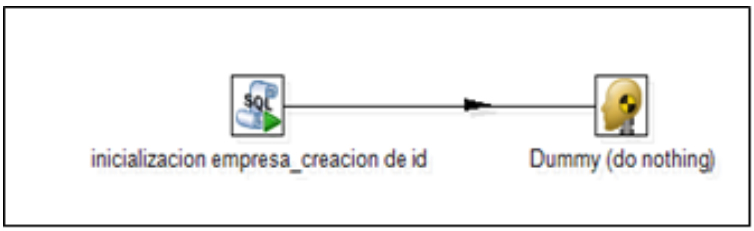


Gráfico N: 22.

Su objetivo principal como lo dice en el nombre es inicializar las variables en el id propio de empresa.

9. Inicialización de variables id_unidad_legal.

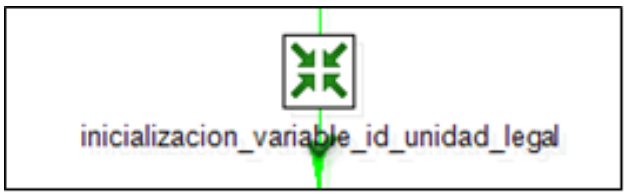


Gráfico N: 23.

La transformación de inicialización de variables id unidad_legal se compone de los siguientes objetos:

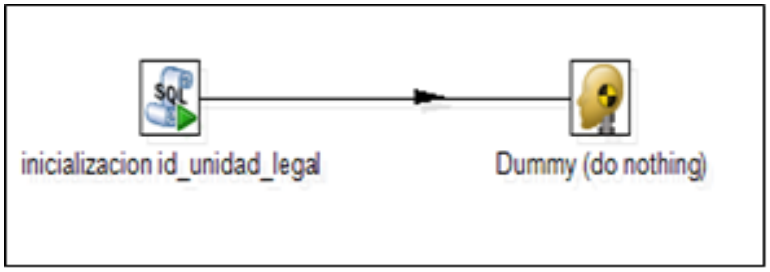


Gráfico N: 24.

Su objetivo principal como lo dice en el nombre es inicializar las variables en el id propio de unidad legal.

10. Inicialización de variables id_unidad_local.

Gráfico N: 25.

La transformación de inicialización de variables id_unidad_local se compone de los siguientes objetos:

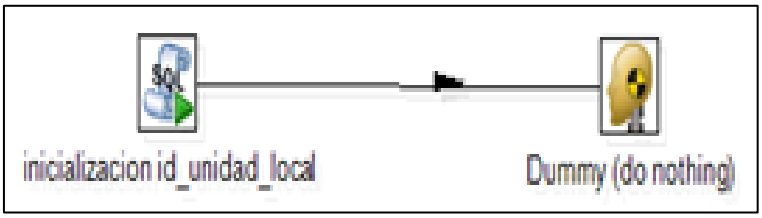


Gráfico N: 26.

Su objetivo principal como lo dice en el nombre es inicializar las variables en el id propio de unidad local.

11. Inicializa_id_SRI.

Gráfico N: 27.

La transformación de inicialización de variables id se compone de los siguientes objetos:

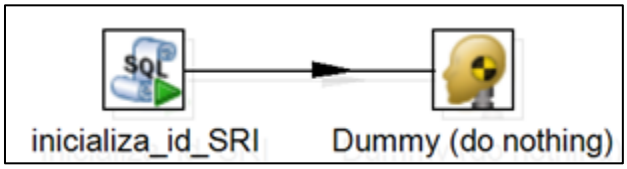


Gráfico N: 28.

Esta transformación tiene la finalidad de llenar los campos de: id_empresa, id_unidad_legal, id_unidad_local que previamente se han creado en las tablas: ruc_contribuyentes y ruc_establecimientos del SRI, el llenado se hace con los datos de las tablas de PASO.

Actividad Económica.

Gráfico N: 29.

Estas transformaciones se encargan de ingresar el id_actividad_economica según las matrices de correspondencias que se tienen en el DIEE.

12. Migracion1_emp_act_economica.

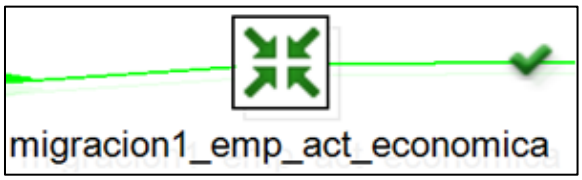


Gráfico N: 30.

La transformación de migracion1_emp_act_economica tiene los siguientes objetos:

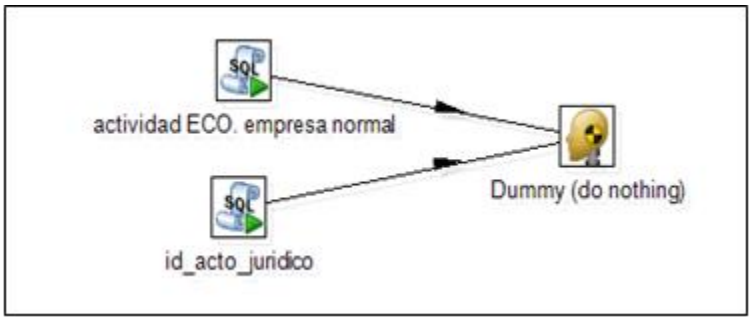


Gráfico N: 31.

La transformación se compone de objetos que ejecutan script´s, cada script tiene su objetivo:

- "actividad ECO.empresa normal": mapeo para conversión automática de actividad económica de versión CIIU3 a CIIU4.
- "Id_acto_juridico": hace una actualización del ruc_acto_juridico cuando es constitución, escisión o fusión.
- "actividad ECO empresa matiz Steven": En el DIEE se trabajó con una matriz de conversión propia para las actividades que no tienen.
- Finalmente se asigna id´s de actividad económica a 4 y a 2 dígitos según las matrices que el DIEE posee.

13. Unidades locales rescatadas actividad.

Gráfico N: 32.

La transformación tiene los siguientes objetos:

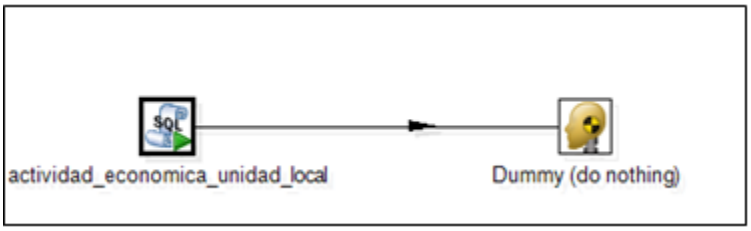


Gráfico N: 33.

Aquí se busca identificar a los establecimientos únicos, únicos por estado y bajar la actividad de empresa directo al establecimiento.

Se identifican los establecimientos que tienen una sola actividad para poder convertir su actividad directamente con las matrices de conversión.

14. Migracion1_ulocal_act_economica.

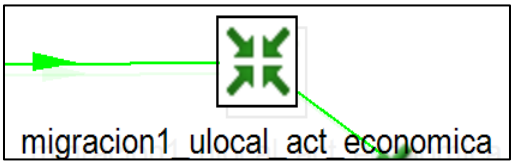


Gráfico N: 34.

La transformación de migracion1_ulocal_act_economica tiene los siguientes objetos:

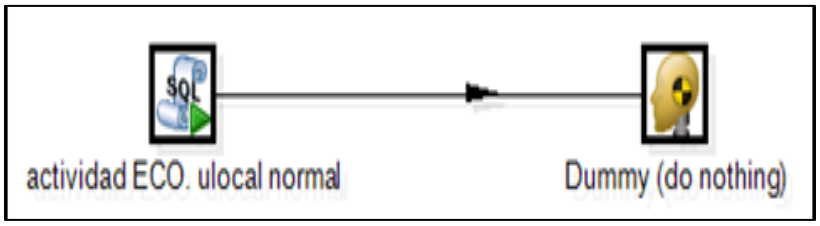


Gráfico N: 35.

Al igual que la anterior transformación este se compone básicamente del mapeo normal de actividades económicas, la aplicación de la matriz propia del DIEE y el llenado de las demás actividades que vienen de otras fuentes diferentes al SRI.

Migración de Dirección.

En esta parte intervienen las transformaciones:

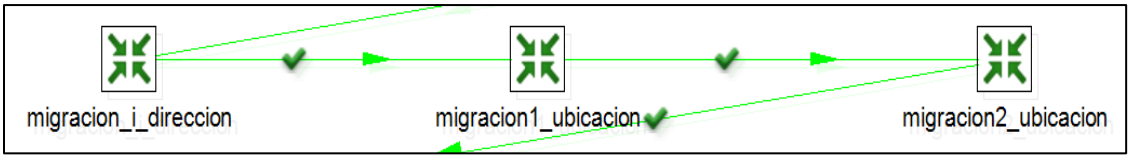


Gráfico N: 36.

15. Migración_i_direccion.

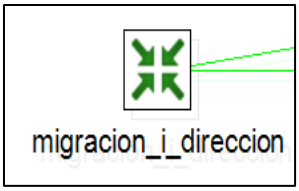


Gráfico N: 37.

La transformación de migración_i_direccion se compone de los siguientes objetos.

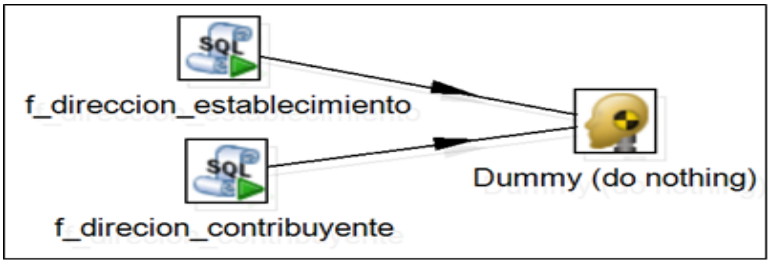


Gráfico N: 38.

Estos tienen la función de extraer los datos desde las tablas de ruc_contribuyentes y ruc_establecimientos del SRI, en lo referente a dirección y se asocia el tipo de vía y el tipo de zona a la cual pertenece la dirección de cada empresa y establecimiento, para llenar para llenar con éstos datos la tabla de i_dirección del esquema PASO.

16. Migracion1_ubicacion.

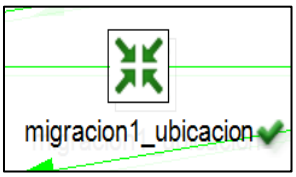


Gráfico N: 39.

La transformación de migración1_ubicacion se compone de los siguientes objetos.

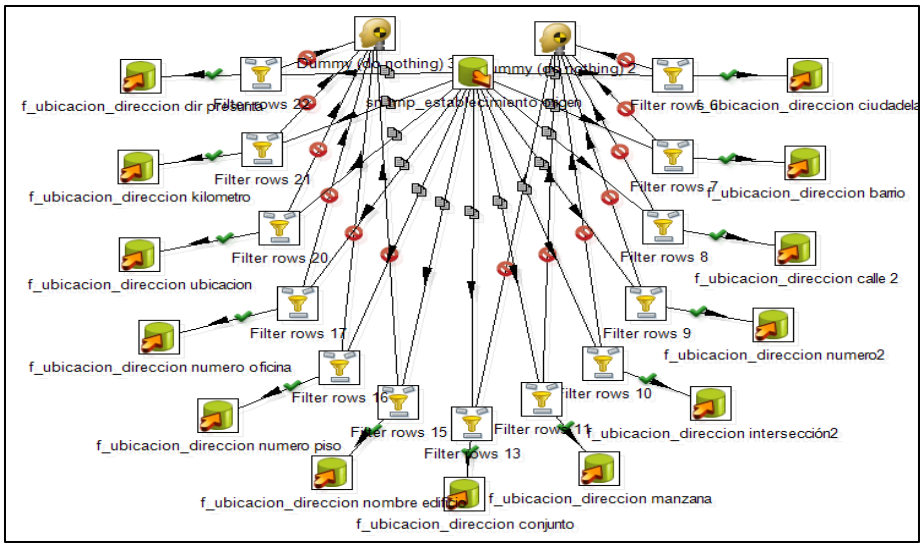


Gráfico N: 40.

17. Migracion2_ubicacion.

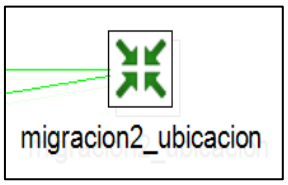


Gráfico N: 41.

La transformación de migración2_ubicacion se compone de los siguientes objetos:

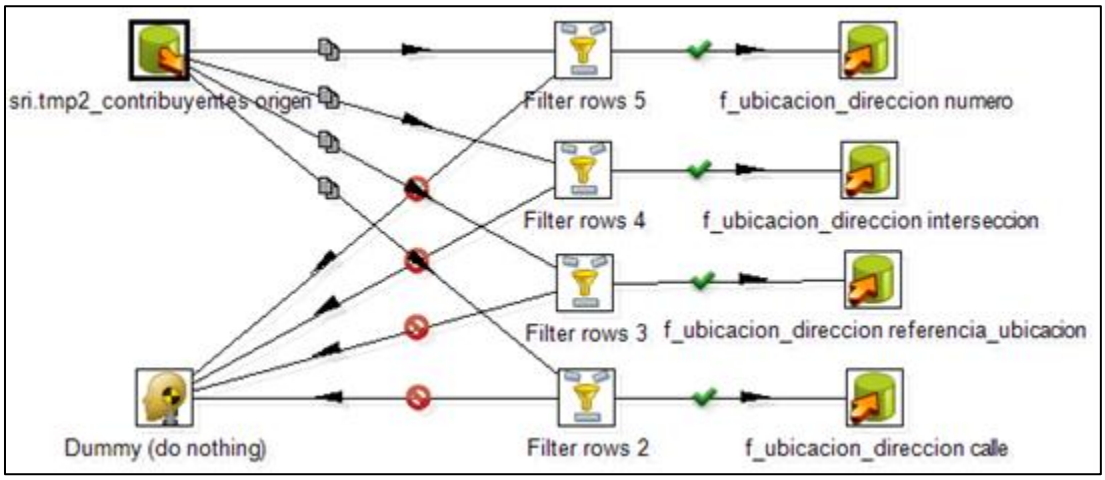


Gráfico N: 42.

Las transformaciones de los gráficos 74 y 49 se compones de varios objetos pero con objetivos en común, pasar la información de las direcciones de contribuyentes a empresa y de establecimientos a unidad local.

La fuente SRI se tiene toda la información referente a direcciones en una sola tabla, en donde se detallan los siguientes datos;

Para empresas: calle, número, intersección, referencia_ubicacion.

Para establecimientos: barrio, ciudadela, conjunto, bloque, calle, interseccion, nombre_edificio, numero, numero_oficina, manzana, supermanzana, kilometro, carretero, camino, numero_piso, direccion_presunta, referencia_ubicacion.

De los cuales se han agrupado en 12 variables para el mejor manejo de las direcciones, por ello se manejan las siguientes variables: calle_final,

numero, interseccion_final, kilometro, conjunto, nombredificio_bloque, numero_piso, numero_oficina, ciudadela, barrio, referencia_ubicacion, manzana_supermanzana, según los nombres se puede apreciar que las variables:

nombredificio_bloque está concatenando la información de nombre_edificio y bloque; manzana_supermanzana, concatena las variables manzana y supermanzana; y las variables: calle_final e interseccion_final se componen adicionalmente de la información de los datos de las variables de camino y carretero, dependiendo de la información que se tenga en estas 4 variables. (Ver documento: "Resoluciones sobre Direcciones")

A su vez, las variables del SRI anteriormente detalladas, en la base del DIEE son almacenadas en una sola variable, en la tabla f_ubicacion_direccion, con el nombre de: descripción, pero adicional a esto existe en la tabla la variable id_tipo_direccion, la cual indica el tipo de dato que se tiene, según el siguiente catálogo:

18. Migración ulocal catalogo ok.

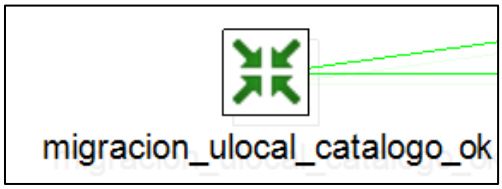


Gráfico N: 43.

La transformación de migración_ulocal_catalogo_ok se compone de los siguientes objetos.

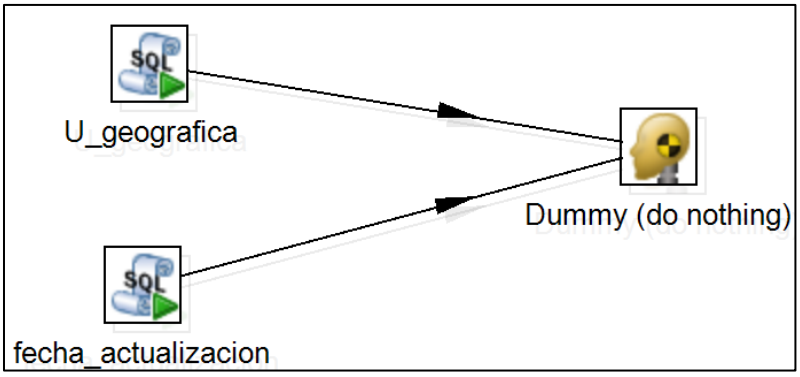


Gráfico N: 44.

Esta transformación tiene dos script que ejecutan lo siguiente:

- "U_geografia": se actualiza el id_geografia para unidad local desde el catálogo de geografía.
- Se actualiza la fecha de actualización, para el respectivo llenado de variables de control.

19. Migracion1_ulegal_clasificacion_fjuridica.

Gráfico N: 45.

La transformación tiene los siguientes objetos:

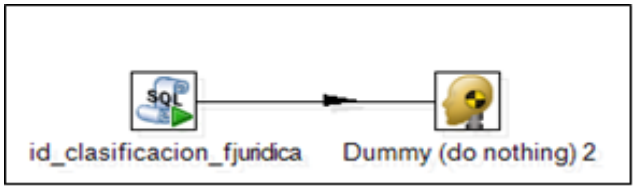


Gráfico N: 46.

Llena el campo "id_clasificacion_fjuridica": asigna el id clasificación de forma jurídica de SRI a la catalogación del DIEE.

20. Migracion u_legal_catalogo.

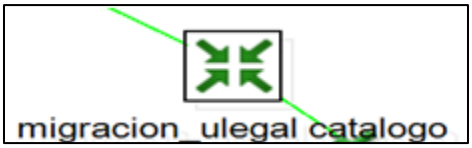


Gráfico N: 46.

La transformación tiene los siguientes objetos:

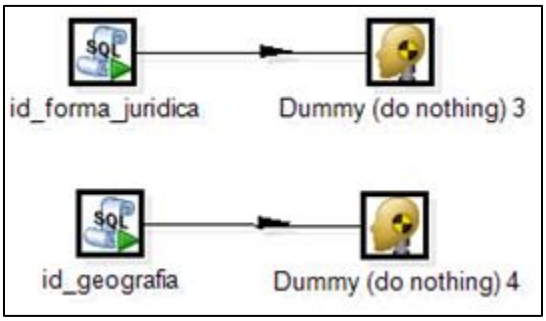


Gráfico N: 47.

Como se puede observar en el Gráfico N: 39 tiene también objetos de ejecución de script´s que se detallan a continuación:

- "id_forma_juridica": asigna el id forma jurídica de SRI a la catalogación del DIEE.
- "id_geografia": asigna el id de geografía de SRI a la catalogación del
- DIEE.

21. Migracion ulegal_id_forma_juridica.

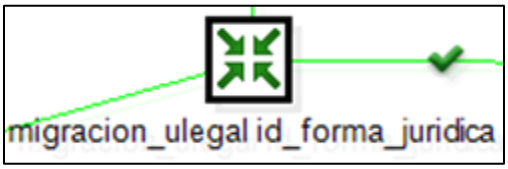


Gráfico N: 48.

La transformación tiene los siguientes objetos:

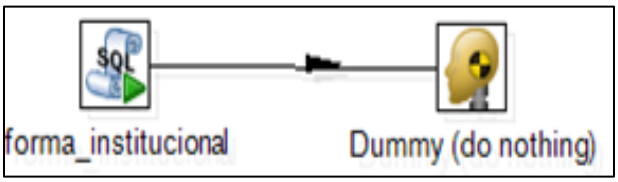
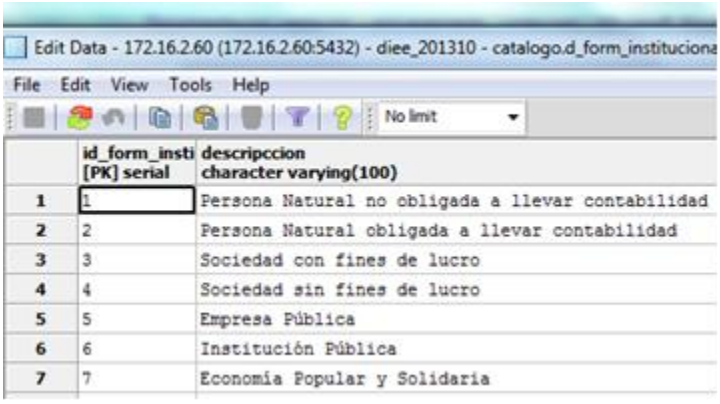


Gráfico N: 49.

Forma institucional": El propósito de esta variable es obtener información más desagregada de las empresas y establecimientos y obtener mejores resultados en la fase de análisis de la información. La catalogación de la variable en mención es la siguiente:



	id_form_insti [PK] serial	descripccion character varying(100)
1	1	Persona Natural no obligada a llevar contabilidad
2	2	Persona Natural obligada a llevar contabilidad
3	3	Sociedad con fines de lucro
4	4	Sociedad sin fines de lucro
5	5	Empresa Pública
6	6	Institución Pública
7	7	Economía Popular y Solidaria

Gráfico N: 50.

Actualización de Empresas o Establecimientos nuevos.

En esta parte intervienen las transformaciones:

Gráfico N: 51.

En las que se establecen las variables de control tanto para empresas como para establecimientos en las tablas de: i_empresa, i_unidad_legal, i_unidad_local del esquema PASO.

22. Upd_empresas_nuevas.

Gráfico N: 52.

Actualización de empresas que ingresan al directorio.

La transformación contiene los siguientes objetos que se encargan del siguiente proceso:

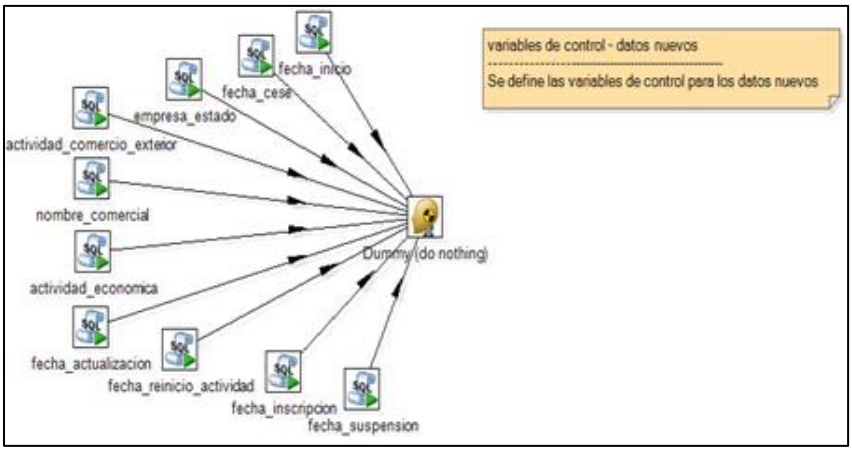


Gráfico N: 53.

En cada objeto se definen las diferentes variables de control, como son: registro, registro_fecha, fuente y fuente_fecha, fecha_desde para las variables de empresas que ingresan al directorio.

23. Upd_ulegal_nuevas.

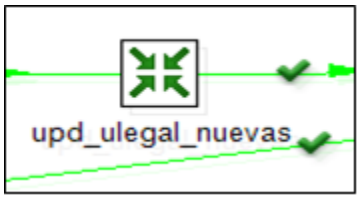


Gráfico N: 54.

La transformación tiene la siguiente transformación:

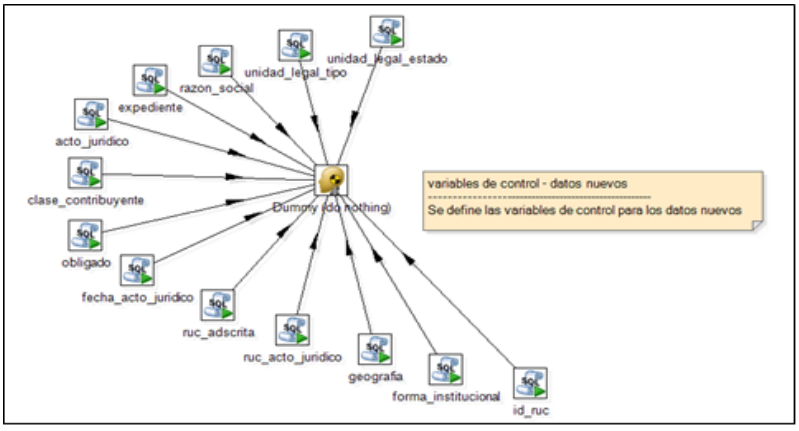


Gráfico N: 55.

Al igual que la anterior transformación se definen las diferentes variables de control de unidad legal que ingresan al directorio.

24. Upd_ulocal_nuevas.

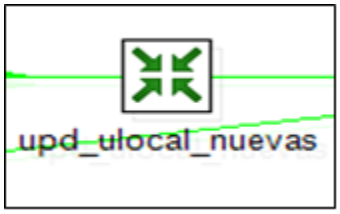


Gráfico N: 57.

La transformación que corresponde a la siguiente transformación:

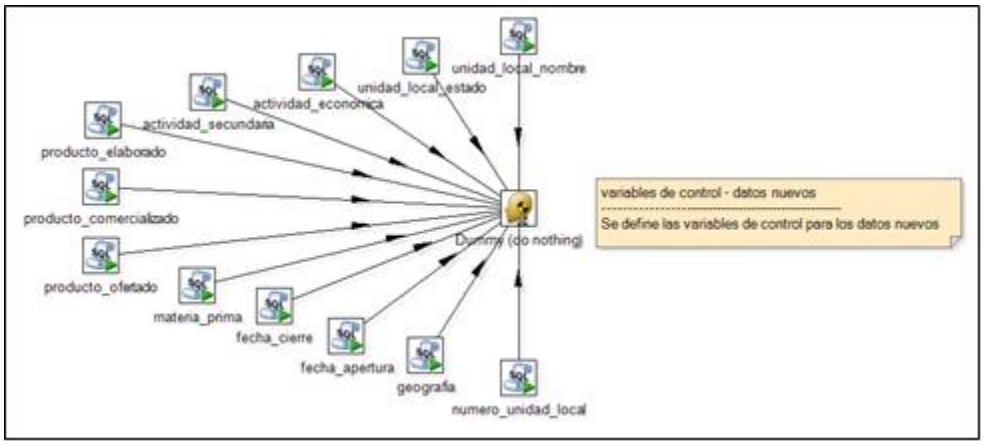


Gráfico N: 58.

En esta transformación se definen las diferentes variables de control de las unidades locales que ingresan al directorio.

25. Upd_geografia_ulegal_null.

Gráfico N: 59.

La transformación contiene los siguientes objetos:

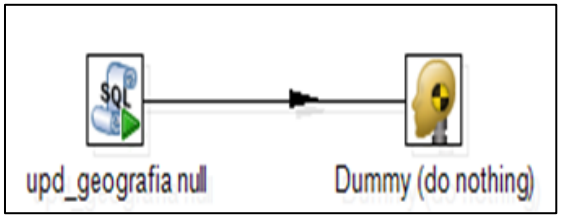


Gráfico N: 60.

Cuando no se tiene dato de geografía en unidad legal, esta transformación se encarga de extraer dicha información, a partir de la geografía existente en el establecimiento matriz de la empresa en cuestión.

Actualización de campos con datos iguales y diferentes.

Para esta actualización intervienen las transformaciones:

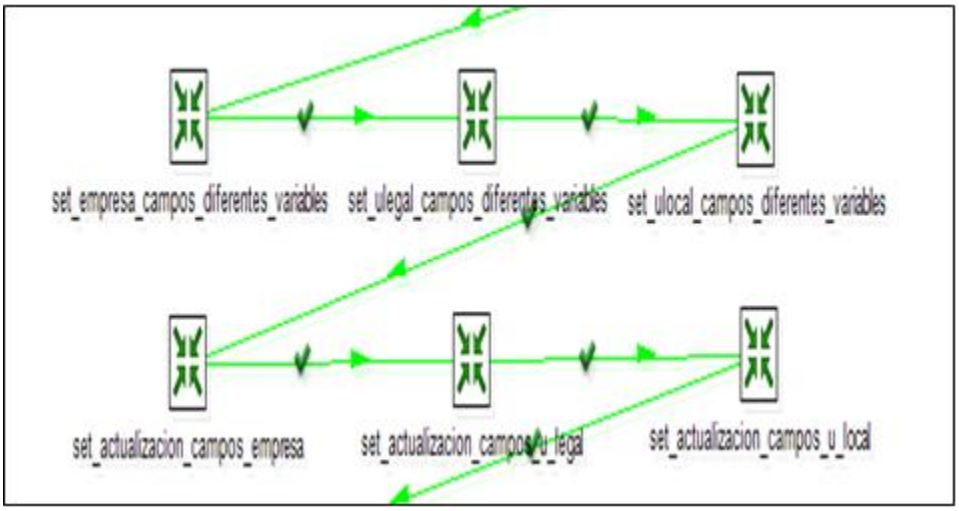


Gráfico N: 61.

En las que se actualiza determinado campo y sus respectivas variables de control, se actualiza con la información de la fuente SRI si los datos son diferentes.

26. Set_empresas_campos_diferentes_variables

Gráfico N: 62.

La transformación contiene los siguientes objetos:

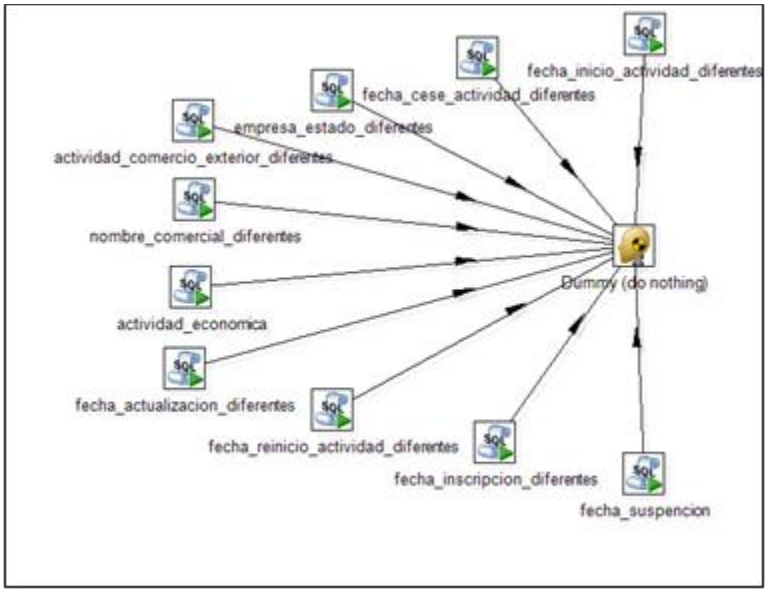


Gráfico N: 63.

Esta transformación actualiza los datos de las variables de control cuando haya existido una actualización de cualquier dato que posea variable de control.

27. Set_ulegal_campos_diferentes_variables.

Gráfico N: 63.

La transformación contiene los siguientes objetos:

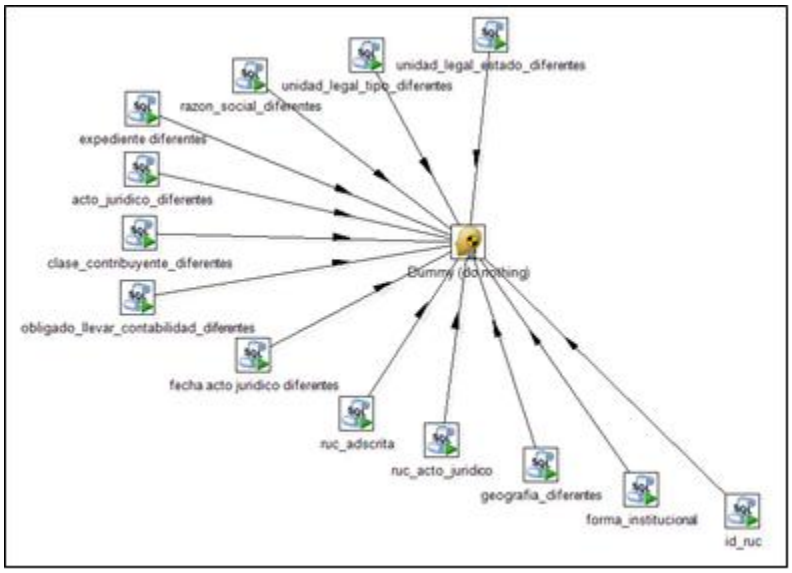


Gráfico N: 64.

Al igual que la anterior transformación esta actualiza los datos de las variables de control cuando haya existido una actualización que provenga de la fuente SRI.

28. Set_ulocal_campos_diferentes_variables.

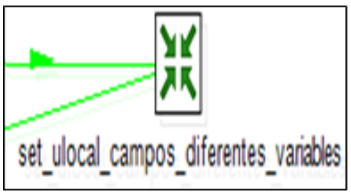


Gráfico N: 64.

La transformación contiene los siguientes objetos:

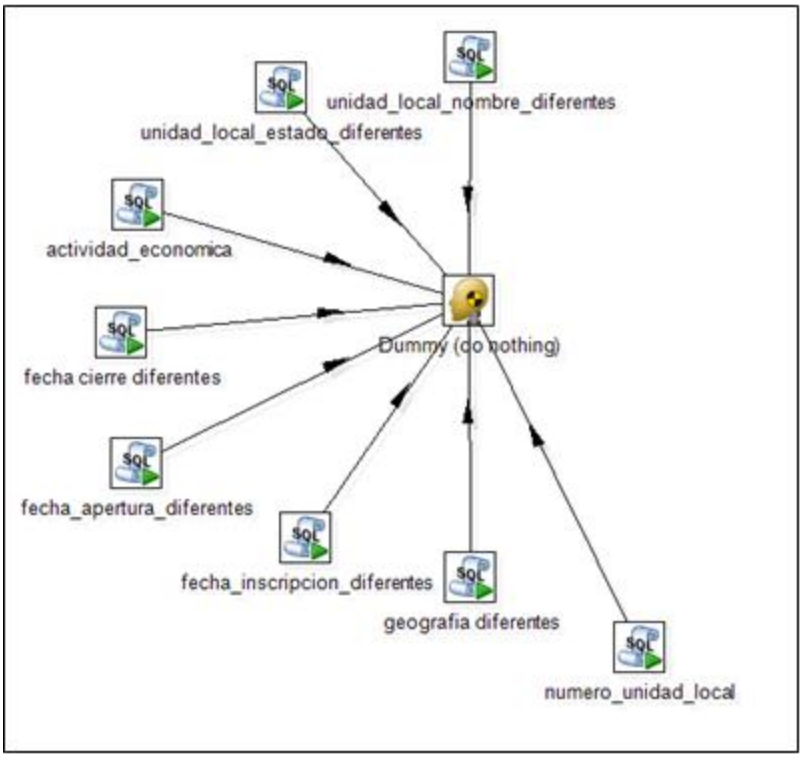


Gráfico N: 65.

Al igual que la anterior transformación esta actualiza los datos de unidad local en sus variables de control cuando haya existido una actualización que provenga de la fuente SRI.

29. Set_actualizacion_campos_empresa.

Gráfico N: 66.

La transformación contiene los siguientes objetos:

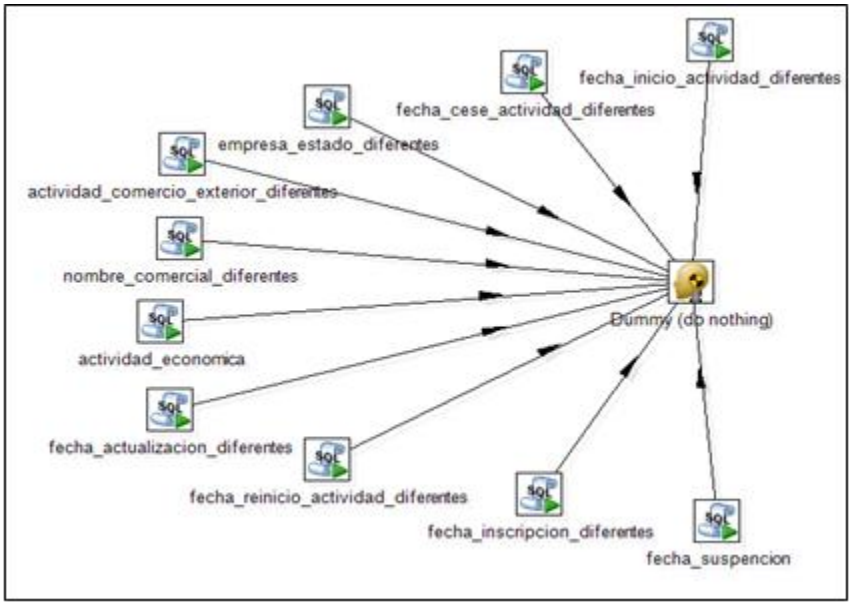


Gráfico N: 67.

Esta transformación actualiza los datos en la tabla de empresa cuando haya existido una actualización, para los casos que pertenecen a la fuente SRI.

30. Set_actualizacion_campos_ulegal.

Gráfico N: 68.

La transformación contiene los siguientes objetos:

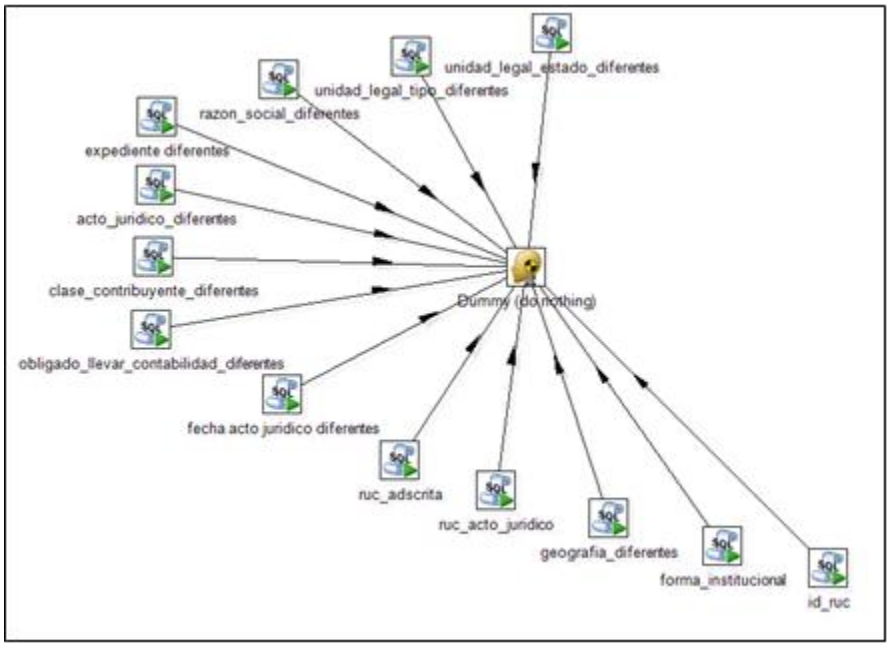


Gráfico N: 69.

Al igual que la anterior transformación esta actualiza los datos en la tabla de unidad_legal cuando haya existido una actualización, para los casos que pertenecen a la fuente SRI.

31. Set_actualizacion_campos_ulocal.

Gráfico N: 70.

La transformación contiene los siguientes objetos:

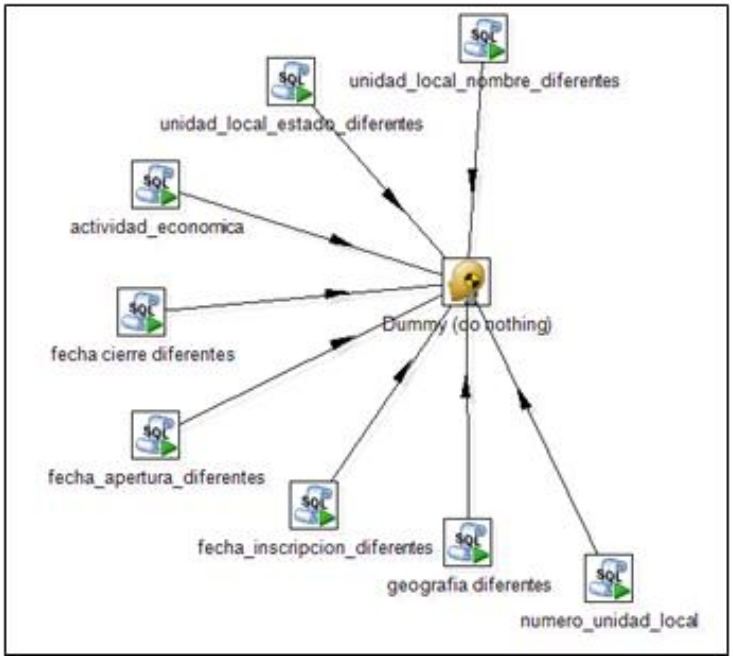


Gráfico N: 71.

De la misma manera la transformación actualiza los datos en la tabla de unidad_local cuando haya existido una actualización, para los casos que pertenecen a la fuente SRI.

Ventas

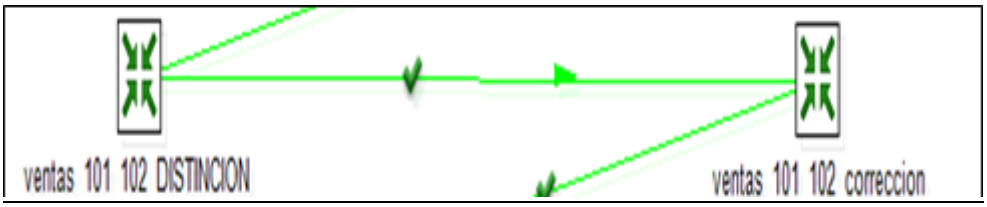


Gráfico N: 72.

En esta sección se trabaja sobre las ventas que reporta el SRI

32. Ventas_101_102_distinción.

La transformación contiene los siguientes objetos:

Gráfico N: 73.

Esta transformación marca que empresas deben estar en el formulario 101 y las que deben estar en el 102, para que al momento de pasar la información al DIEE no exista duplicidad de datos.

33. Ventas_101_102_coreccion.

Gráfico N: 74.

La transformación contiene los siguientes objetos:

Gráfico N: 75.

En primera instancia, la información proporcionada por el SRI llega de la siguiente manera:

26	UTILIDAD_EJE_PATRIMONIO_1760	UTILIDAD DEL EJERCICIO PATRIMONIO	517	517
27	PERDIDA_EJE_PATRIMONIO_1770	PERDIDA DEL EJERCICIO PATRIMONIO	519	519
28	VLN_EAF_TDC_1800	VENTAS NETAS LOCALES EXCLUYE ACTIVOS FIJOS TARIFA DIFERENTE DE CERO	601	601
29	VLN_EAF_TCE_1810	VENTAS NETAS LOCALES EXCLUYE ACTIVOS FIJOS TARIFA CERO	602	602
30	EXPORTACIONES_NETAS_1820	EXPORTACIONES NETAS	603	603
31	OTR_RENTAS_EXENTAS_100_3460	OTRAS RENTAS GRAVADAS	606	606
32	UTILIDAD_VTA_ACT_FIJOS_1860	UTILIDAD VENTA ACTIVOS FIJOS	607	607
33	DIV_PERCIBIDOS_LOCALES_1870	DIVIDENDOS PERCIBIDOS LOCALES	608	608
34	VENTA_NETA_ACTIVOS_FIJOS_1940	VENTA NETA ACTIVOS FIJOS	691	691
35	CTO_IVI_MATERIA_PRIMA_2010	COSTO INVENTARIO INICIAL MATERIA PRIMA	706	706
36	CTO_CLN_MATERIA_PRIMA_2020	COSTO COMPRAS LOCALES NETAS MATERIA PRIMA	707	707

Gráfico N: 76.

Los campos que se utiliza para nutrir a la base del DIEE son los que están subrayados con colores: amarillo y lila, los últimos 4 han sido incrementados el último año, al directorio. La información que se utiliza para extraer los registros de ventas que el SRI proporciona, corresponde a las siguientes tablas:

- owb_mvw_ine_anexo3_estruc_f101
- owb_mvw_ine_anexo3_estruc_f102

Que refieren a la información del formulario 101 y 102 respectivamente, donde el formulario 101 contiene las ventas de Personas Jurídicas y el 102 las ventas de las Personas Naturales.

Para la extracción de esta información, la transformación mostrada en el Gráfico N:76 se encarga de pasar de las dos tablas mencionadas de la fuente, la información de ventas tanto del formulario 101 como del 102 por cada año registrado, a la tabla f_empresa_ventas.

Actualización registros Anteriores y Nuevos

Se tiene las siguientes transformaciones:

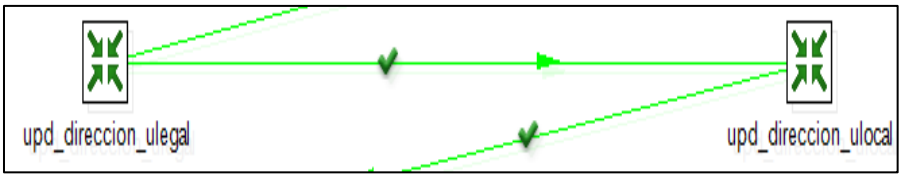


Gráfico N: 77.

34. Upd_direccion_ulegal.

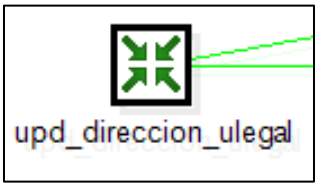


Gráfico N: 78.

La transformación contiene los siguientes objetos:

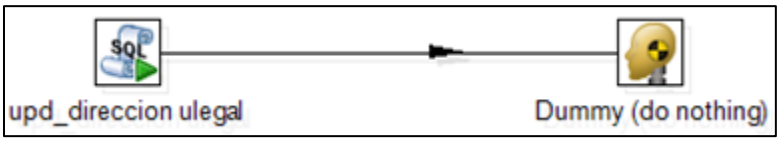


Gráfico N: 79.

En esta transformación se identifican las empresas que hayan cambiado de dirección, lo cual se determina mediante cambios en: calle, número e intersección, cuando existen estos cambios se analiza si la fecha de actualización de la fuente SRI es mayor a la última actualización de la dirección en la base del DIEE, si se tiene este caso se procede a comparar entre la base actual y la del año pasado del SRI para analizar si han existido cambios entre ambos años, caso en el que se procederá a actualizar con estado de "0" las direcciones pertenecientes a las empresas que tuvieron cambios, luego de esto se procede a insertar las nuevas direcciones provenientes del SRI, éstas se registran con estado "1".

35. Upd_direccion_ulocal.

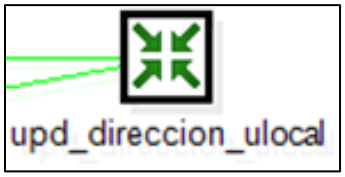


Gráfico N: 80.

La transformación contiene los siguientes objetos:

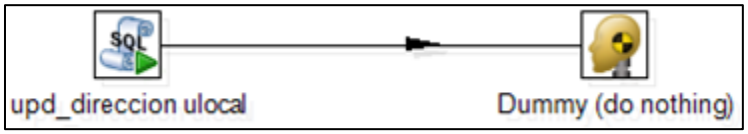


Gráfico N: 81.

Esta transformación realiza el mismo proceso que la transformación upd_direccion_ulegal pero lo realiza para las direcciones de los establecimientos.

Esta transformación transfiere toda la información de la tabla de i_empresa del esquema PASO, de las empresas que ingresan al directorio, a la tabla de f_empresa de la base del DIEE.

36. Nuevas_empresas_ulegal_ulocal.

Gráfico N: 82.

La transformación contiene los siguientes objetos.

Gráfico N: 83.

De la información transferida en la transformación anterior se toma el código creado para empresa (id_empresa), que servirá para asociar a la información a transferir a la tabla f_unidad_legal, esto se realiza para las nuevas empresas a agregarse al directorio.

37. Nuevas_empresa_ulegals_ulocals.

Gráfico N: 84.

La transformación contiene los siguientes objetos:

Gráfico N: 85.

Esta transformación traslada la información únicamente de los nuevos establecimientos a agregarse al directorio, para lo cual, de igual manera que la transformación anterior, se toma el código de la empresa para asociar a la información de cada empresa, para poder pasar la información de los establecimientos a la tabla f_unidad_local.

Nuevas empresas

En esta sección intervienen las siguientes transformaciones.

Gráfico N: 86.

En esta sección se incluye a las empresas nuevas que han nacido durante el año en proceso, para existen tres transformaciones principales.

38. Nuevas_empresas_ulegal_ulocal.

Gráfico N: 87.

La transformación contiene los siguientes objetos:

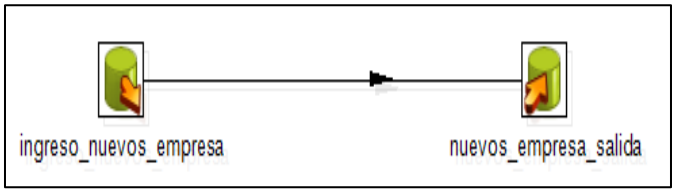


Gráfico N: 88.

Ingresa los nuevos datos desde la tabla i_empresa hasta la tabla f_empresa

39. Nuevas_empresas_ulegals_ulocal.

Gráfico N: 89.

La transformación contiene los siguientes objetos:

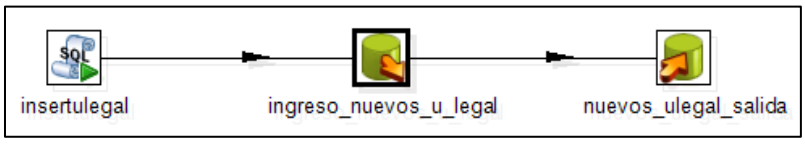


Gráfico N: 90.

En esta sección se ingresa mediante script el id_empresa en unidad_legal y posteriormente ingresa los nuevos datos desde la tabla i_unidad_legal hasta la tabla f_unidad_legal.

40. Nuevas_empresas_ulegals_ulocals.

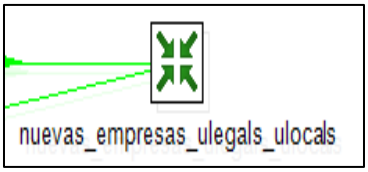


Gráfico N: 91.

La transformación contiene los siguientes objetos:

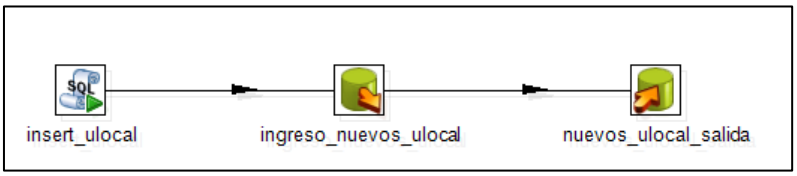


Gráfico N: 92.

En esta sección se ingresa mediante script el id_empresa en unidad_local y posteriormente ingresa los nuevos datos desde la tabla i_unidad_local hasta la tabla f_unidad_local.

Nuevas direcciones

En esta parte están las transformaciones:

- Nuevas_direcciones

Donde se realiza una migración de la información de las tablas de dirección del esquema PASO a las tablas definitivas de la base del DIEE, solamente de los registros nuevos referentes a dirección.

41. Nuevas direcciones.

Gráfico N: 93.

La transformación contiene los siguientes objetos:

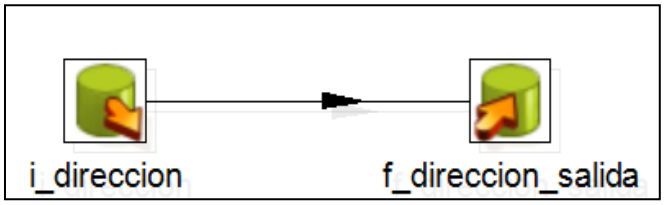


Gráfico N: 94.

Los objetos del Gráfico N: 94, buscan pasar la información de las nuevas direcciones almacenadas en la tabla de i_direccion del esquema PASO a la tabla f_direccion en la base del DIEE.

42. Ingreso_contactos

Gráfico N: 95.

La transformación contiene los siguientes objetos:

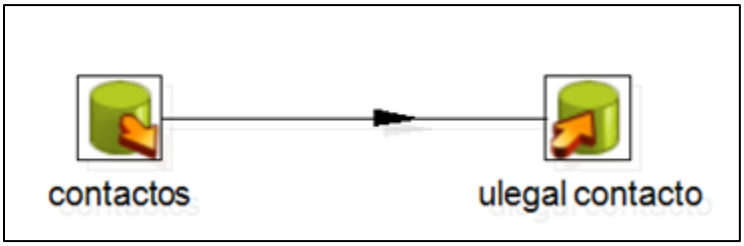


Gráfico N: 96.

43. migracion_iess

Gráfico N: 97.

La transformación contiene los siguientes objetos:

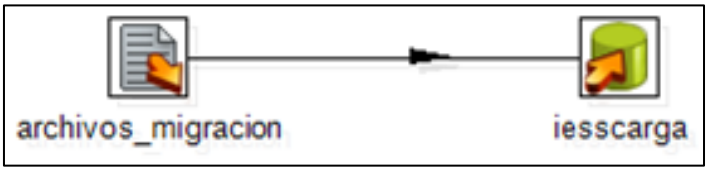


Gráfico N: 98.

Estos objetos se encargan de seleccionar los archivos de texto y migrar en la tabla iesscarga, para que los datos se encuentren en el mismo formato de la base del DIEE y poder procesarlos sin complicaciones. Hasta el momento este es un proceso que necesita ser repetido varias veces, debido a que los archivos de texto que provee la fuente IESS, llegan con un sin número de errores, los mismos que deben ser modificados manualmente hasta que la información esté en óptimas condiciones, para que finalmente pueda ser migrada a PostgreSQL.

44. F_ulocal_empleados

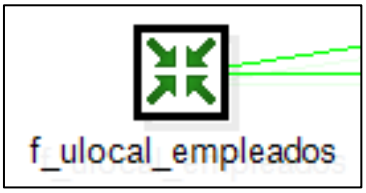


Gráfico N: 99.

La transformación contiene los siguientes objetos:

Gráfico N: 100.

En el Gráfico N: 80, se tiene inicialmente la creación de la tabla de promedios de los empleados, ya que la información llega mensualmente y el DIEE requiere datos anuales, por ello se realiza los respectivos promedios de los mismos y poder tomar la información necesaria de lo que se recibe del IESS con respecto a empleados.

En el objeto llamado iess_carga_promedio se toma la información de: empleados hombres, mujeres y el total de la suma entre ambos datos, esto para empleados de cada unidad local. Finalmente estos valores serán transferidos a la tabla f_unidad_local_empleados de la base del DIEE.

45. F_empresa_empleados

Gráfico N: 101.

La transformación contiene los siguientes objetos:

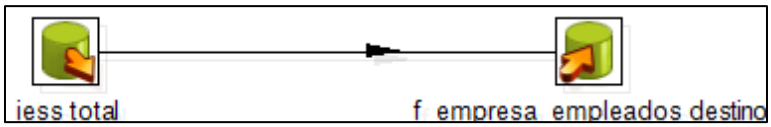


Gráfico N: 102.

En el objeto nombrado como iess total se realiza la sumatoria de los empleados de todos los establecimientos de la empresa, para empleados hombres, mujeres y el total de la suma entre ambas variables, finalmente estos valores serán transferidos a la tabla f_empresa_empleados de la base del DIEE.

46. F_ulocal_empleados_9000

Gráfico N: 103.

La transformación contiene los siguientes objetos:

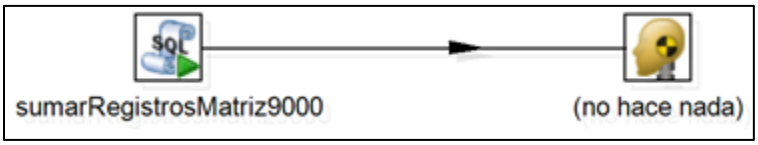


Gráfico N: 104.

En el Gráfico N: 104 se ejecutan una serie de scripts, donde inicialmente se encera la variable id_tipo_unidad_local, para ser nuevamente llenado con la información de la tabla f_unidad_local de la base del DIEE, con el fin de almacenar toda la información que no pertenece a ninguna unidad local en una tabla temporal. Estos datos vienen de la fuente del IESS generalmente con número de unidad local superior o igual a 9000, los empleados que reportan estos casos son sumados a los empleados afiliados de la matriz de la empresa respectiva.

Si se presenta el caso que no existe matriz en la información proporcionada por el IESS, se suma esta la información de establecimientos, con unidad local de 9000, al establecimiento con mayor número de empleados.

47. F_empleados_totales

Gráfico N: 105.

La transformación contiene los siguientes objetos:

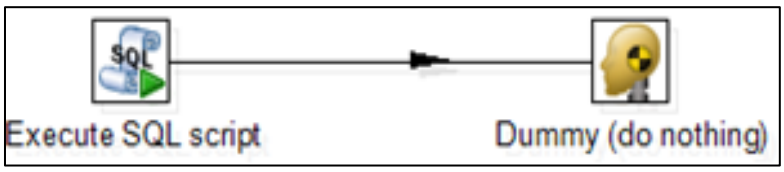


Gráfico N: 106.

El script del Gráfico N: 106 se encarga de realizar una suma entre los empleados hombres y mujeres para obtener el total de empleados tanto de empresas como de establecimientos.

Descarta empresas y establecimientos

Se tiene las siguientes transformaciones:

Gráfico N: 107.

Existen empresas que sus actividades económicas están dentro de las secciones T o U:

Sección T; Actividades de los hogares individuales en calidad de empleadores, Actividades no diferenciadas de los hogares individuales como productores de bienes y servicios para uso propio.

Sección U; Actividades de organizaciones y entidades extraterritoriales. A estas empresas se las descarta del directorio de empresas, este proceso es el que se lleva en las trasformaciones indicadas, y son detalladas a continuación.

48. descarta_empresasyulocales CIIU

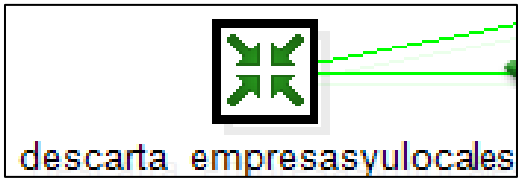


Gráfico N: 108.

La transformación contiene los siguientes objetos:

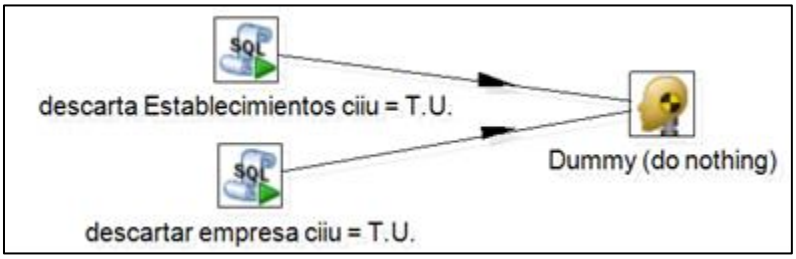


Gráfico N: 109.

En el Gráfico N: 109 se ejecutan scripts para descartar empresas y establecimientos en base al tipo de actividad económica, es decir, cuando la actividad económica pertenece a las secciones T ó U las empresas deben ser descartadas, en este caso se actualiza la variable directorio con el texto 'NO', de igual manera el campo nota, se actualiza indicando el por qué ha sido descartada la empresa o establecimiento, esta actualización se realiza en las tablas de f_unidad_local_estratos y f_empresa_estratos.

49. Descarta dependencias empresa_ulocal CIIU

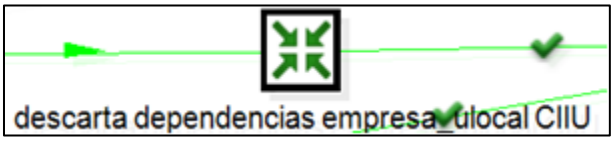


Gráfico N: 110.

La transformación contiene los siguientes objetos:

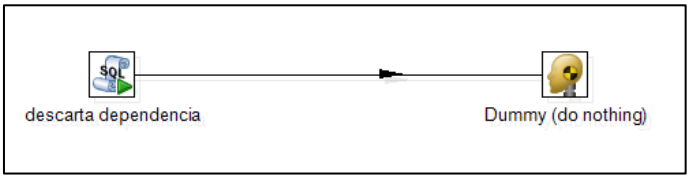


Gráfico N: 111.

El Gráfico N: 111 se procede a descartar las empresas y establecimientos que dependan de las que fueron descartadas en la transformación del Gráfico N: 109 cuando se tiene actividad económica perteneciente a las secciones T ó U, es decir, si se descartó una empresa por este motivo, en esta transformación se descartan todos los establecimientos que pertenecen a esta empresa, y si se descartó un establecimiento, en consecuencia se descartan el resto de establecimientos junto con la empresa a la que pertenece dicho establecimiento. Se actualiza la variable directorio con el texto 'NO', de igual manera el campo nota, se actualiza indicando el por qué ha sido descartada la empresa o establecimiento.

50. Descarta_actividad_eco

Gráfico N: 112.

La transformación contiene los siguientes objetos:

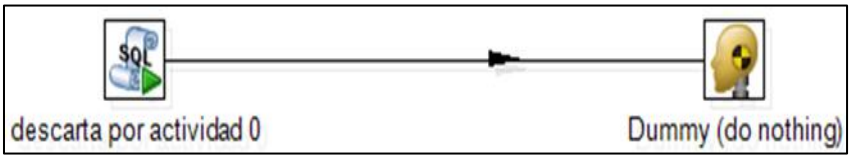


Gráfico N: 113.

El script del gráfico N: 112 se encarga de descartar tanto a empresas como a establecimientos que no tienen correspondencia con ninguna de las matrices para la obtención de la actividad económica, en este caso se actualiza la variable directorio con el texto 'NO', de igual manera el campo nota, se actualiza indicando el por qué ha sido descartada la empresa o establecimiento, esta actualización se realiza en las tablas de f_unidad_local_estratos y f_empresa_estratos.

51. Descarta_actividad_eco_dependencias.

Gráfico N: 114.

La transformación contiene los siguientes objetos:

Gráfico N: 115.

El script del gráfico N: 92 se encarga de descartar tanto a empresas como a establecimientos que dependan de las que fueron descartadas en la transformación del Gráfico N: 90, por no tener correspondencia con ninguna de las matrices para la obtención de la actividad económica, es decir, si se descartó una empresa por este motivo, en esta transformación se descartan todos los establecimientos que pertenecen a esta empresa, y si se descartó un establecimiento, en consecuencia se descartan el resto de establecimientos junto con la empresa a la que pertenece dicho establecimiento. En este caso se actualiza la variable directorio con el texto 'NO', de igual manera el campo nota, se actualiza indicando el por qué ha sido descartada la empresa o establecimiento, esta actualización se realiza en las tablas de *f_unidad_local_estratos y f_empresa_estratos.*

52. Upd_numeroUnidadesLocales.

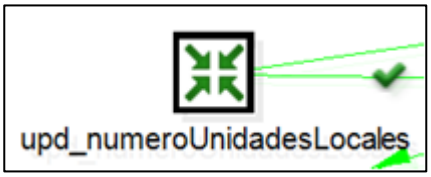


Gráfico N: 116.

La transformación contiene los siguientes objetos:

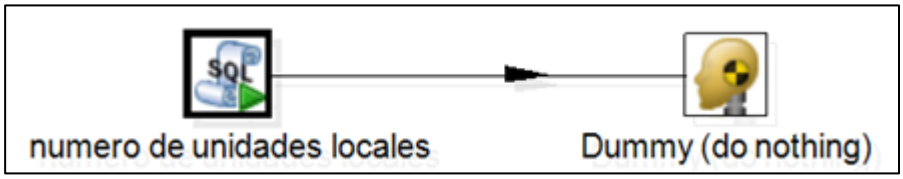


Gráfico N: 117.

En el Gráfico N: 112 se ejecuta el script que realiza un conteo del número de unidades locales activas, para con esta información llenar el campo de numero_unidades_locales en la tabla f_empresa de la base del DIEE, este valor corresponde al número total de establecimientos que tiene cada empresa.

53. Migración medioComunicacion previo.

Gráfico N: 118.

La transformación contiene los siguientes objetos:

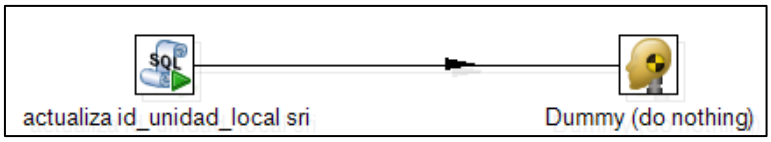


Gráfico N: 119.

El Gráfico N: 114 se encarga de actualizar el código de unidad local (id_unidad_local) de contactos de las fuentes de información CENEC y SRI con el código de la tabla f_unidad_local de la base del DIEE, para ser utilizados en la siguiente transformación.

54. Update_medios_comunicacion

Gráfico N: 120.

La transformación contiene los siguientes objetos

Gráfico N: 121.

En esta transformación se analiza los medios de comunicación que han cambiado y se actualizan si: la fecha de la última actualización es menor que la fecha de actualización del dato del SRI por el cual se va a actualizar la información, se procede a cambiar a estado "0" a los datos anteriores, para no tener duplicidad en la información, y finalmente se procede a agregar la nueva información a la tabla de *f_medio_comunicacion.*

Con los objetos: medio comunicación SRI y medio comunicación, lo que se realiza es la inserción de todos los medios de comunicación nuevos que se han encontrado en la base de la fuente SRI, que deben ingresarse en la tabla de f_medio_comunicacion.

55. Clasificacion Empleados Ventas

Gráfico N: 122.

La transformación contiene los siguientes objetos

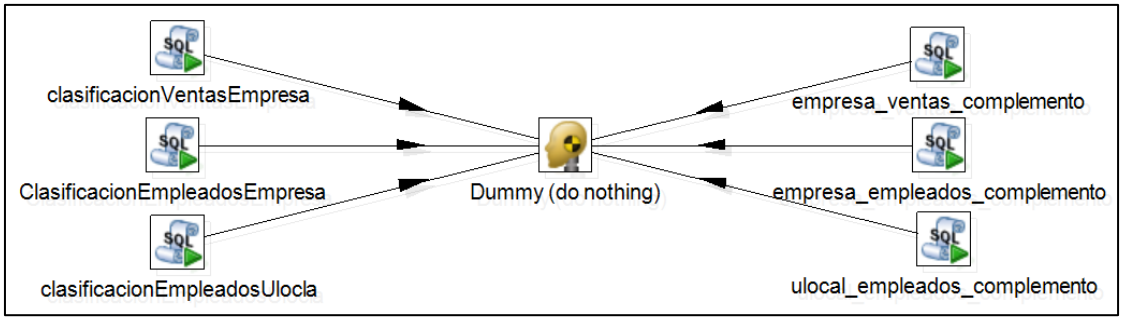


Gráfico N: 123.

En esta transformación se asignan los estratos para empleados y ventas en las empresas según los catálogos de clase de ventas y empleados, que se muestran en las siguientes tablas:

Estratos de ventas:

codigo	estrato	valor_inferior	valor_superior
30	ESTRATO I	0	100000
31	ESTRATO II	100000	1000000
32	ESTRATO III	1000000	2000000
33	ESTRATO IV	2000000	5000000
34	ESTRATO V	5000000	99999999999

Para el Estrato I se considera a todos los casos que tengan forma institucional diferente de 6 (Institución Pública), la clase contribuyente sea igual a RISE y el personal afiliado esté entre 1 y 10.

Estratos de empleados:

codigo	estrato	valor_mímimo	valor_máximo
5	no catalogado		
30	ESTRATO I	1	9
31	ESTRATO II	10	49
32	ESTRATO III	50	99
33	ESTRATO IV	100	199
39	ESTRATO V	200	10000000

La información de los estratos es actualizada en las tablas de: f_empresa_estratos y f_unidad_local_estratos.

CONTEOS:

Para verificar y validar la información que se ha obtenido a partir del procesamiento se procede con conteos establecidos que se tienen en el Plan de Validación y Tabulación del DIEE, si estos conteos están correctos se puede continuar con el congelamiento de la base de datos, caso contrario se analiza cual es el error y se hace un reprocesamiento de la base de datos con la finalidad de corregir el error.

Este proceso se lo realiza hasta tener una base de datos completamente validada y lista para ser publicada.

CONCLUSIONES.

- El proceso de captación desde la fuente SRI puede ser claramente mejorado eliminando la dependencia del motor de BDD o a su vez adquiriendo una licencia de Oracle estándar edition one.
- EL proceso de captación de los datos del IESS tienen muchos inconvenientes por el modo mismo de replicación de información, y mientras este proceso no se cambie difícilmente podrá ser automatizado, por ellos la mejora que puede adaptarse es la adopción de una herramienta de software (moto de BDD) desde el mismo proveedor de información.
- El contar con procesos de calidad de datos puede ayudar a la fase de procesamiento de información, así como a la definición de reglas para el tratamiento de los mismos. Estas herramientas han sido revisadas pero el software libre son bastante limitadas.
- El proceso ETL han sido levantados en algunos puntos de forma paralela a la documentación, esto ha minimizado su facilidad de ser

plasmados en la herramienta de Software, por lo que en futuros procesamientos claramente pueden evidenciarse mejoras en los ETL.

- Existen scripts que han sido creados independientemente de los ETL para las fuentes de datos como la super de companias y el call center, pero esto puede ser fácilmente automatizado si se genera un sistema apegado a las reglas que rigen la BDD del DIEE.
- Los procesos de actualización de información han sido ejecutados y se ha evidenciado tiempos de demora considerables y presentado además otros inconvenientes que pueden ser solucionados al planificarse correctamente los perfiles de usuarios a intervenir y definido el papel que estos tendrán en el proceso.
- Al procesamiento, en cada una de sus fases, se lo puede perfeccionar año tras año, con el fin de que los procesos sean automatizados y así obtener cada vez un producto de mejor calidad, siendo esa la meta a cumplir de este proceso