Captación y procesamiento BDD DIEE, Documentación.

Directorio de Empresas y Establecimientos

INEC – 2013/ 10/ 17

Contenido

INTRODUCCIÓN.

El Directorio de empresas (DIEE) se compone de diferentes bases de datos, partiendo del Censo Económico (CENEC), el SRI, IESS, Superintendencia de Compañías y Bases de Datos con ciertas variables investigadas por el call center del Directorio de empresas. Esta información es complementada y validada en menor proporción con matrices de equivalencias de variables codificadas de diferentes maneras entre la fuente de información y el proveedor; como también se cuenta además esporádicamente con variables que son revisadas por encuestas económicas internas del INEC.

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 lo realizado en 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
- Heramienta BI: Pentaho Data Integration.
- Oracle Express Edition 10g

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:

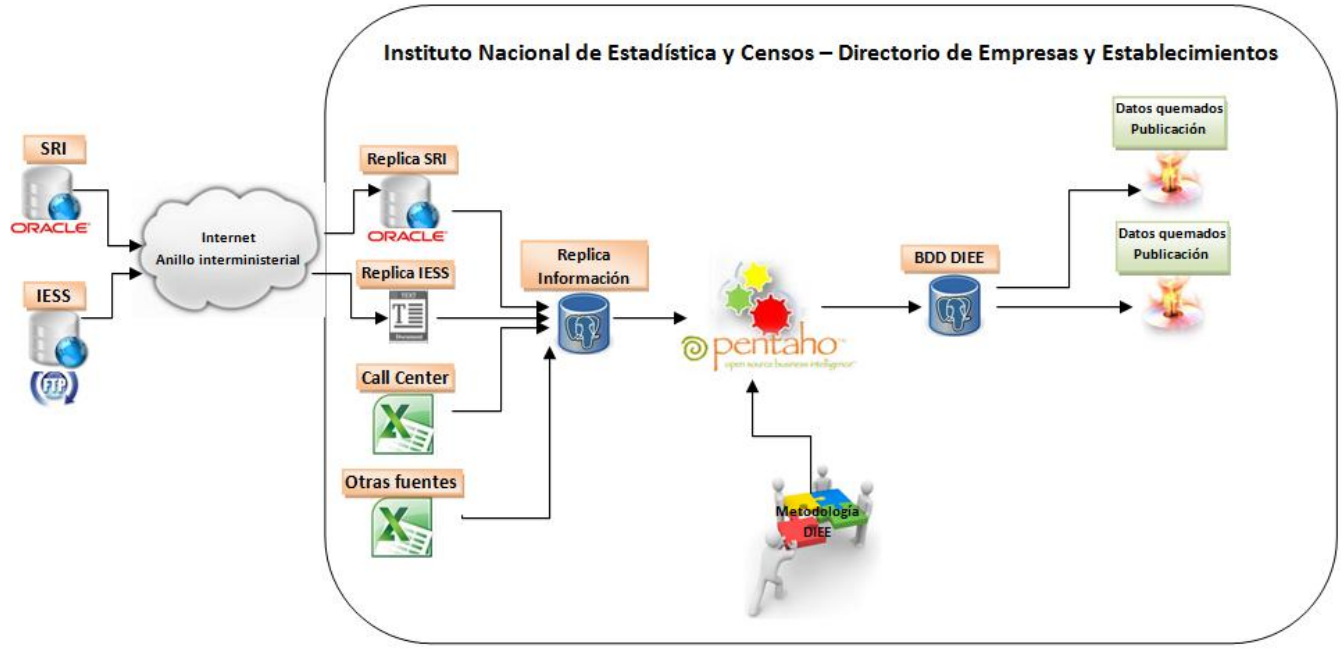


Gráfico N: 1

Captación:

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 ETLs y pequeños programas con la interacción de aplicaciones como Excel.

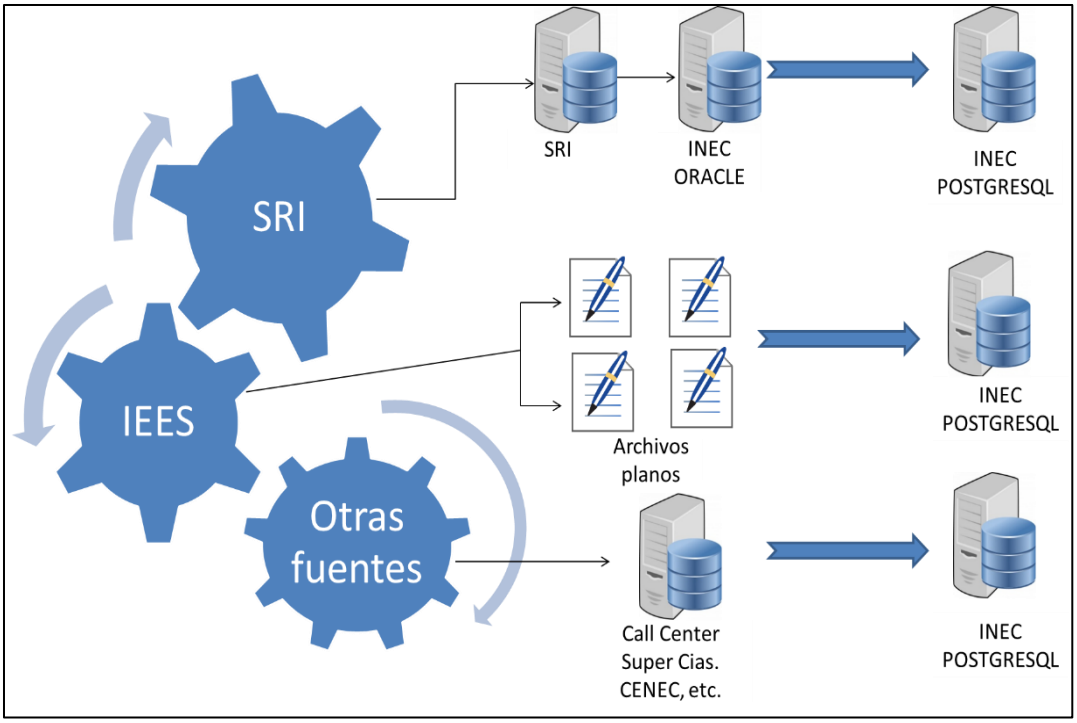


Grafico N: 2

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:

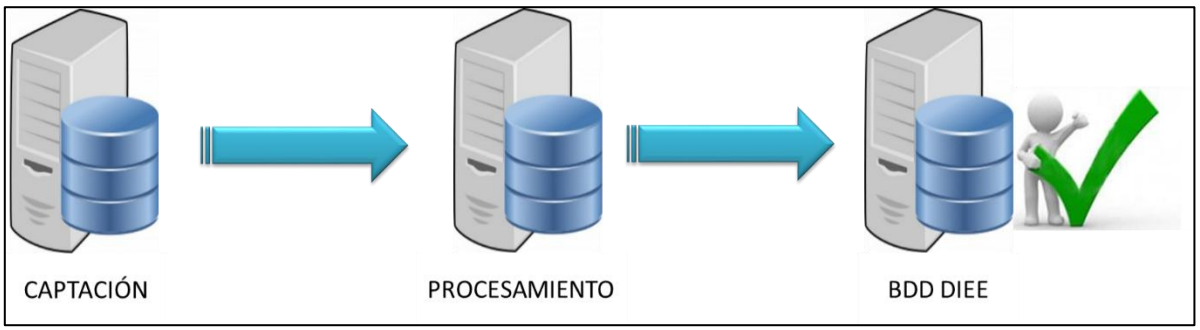


Grafico N: 3

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, mas 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 4 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.
- Se almacena la prioridad dependiendo de la fuente.
- 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.

[1] **NOTA**. En muchas de estas reglas se desconoce si la actualización más reciente corresponde al repositorio de contribuyente o al repositorio de Establecimiento

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.

Grafico 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:

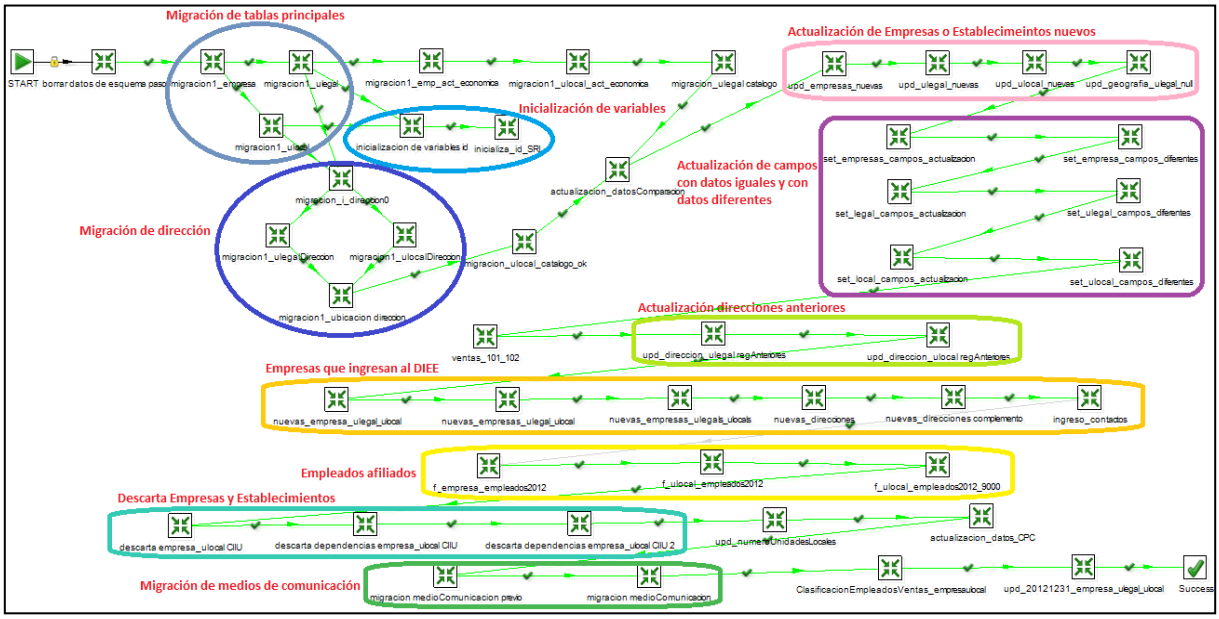


Grafico N: 5

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:

Grafico N: 6

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

2. Borrar datos del esquema paso.

Grafico 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:

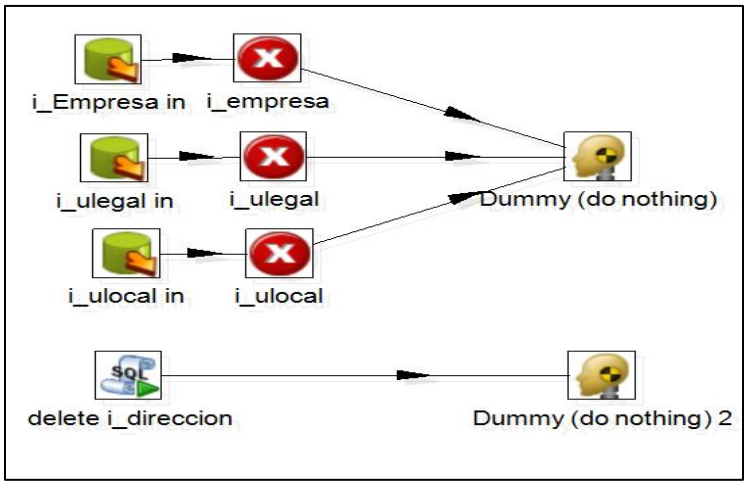


Grafico N: 8

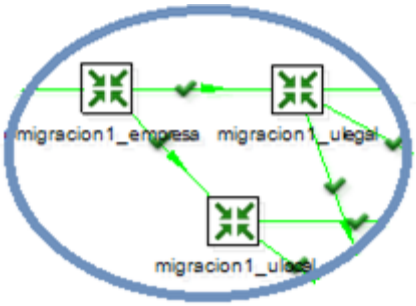
Los Objetos “i_Empresa in”, “i_ulegal in”, “i_ulocal in” tienen como objetivo borrar los datos de empresa cuando la variable fecha_desde no sea nula.

El objeto “delete i_direccion” ejecuta script´s donde borra la información de las tablas:

- paso.i_direccion;
- paso.i_ubicacion_direccion;
- paso.i_unidad_legal_direccion;
- paso.i_unidad_local_direccion;

Migración de tablas principales

En la migración de las tablas principales intervienen las transformaciones:



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*.

3. Migracion1_empresa

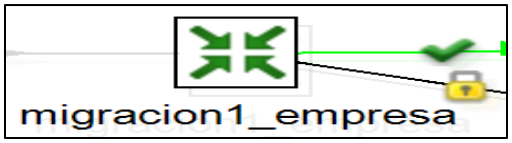


Grafico N: 9

Dentro de esta transformación existen los siguientes objetos:

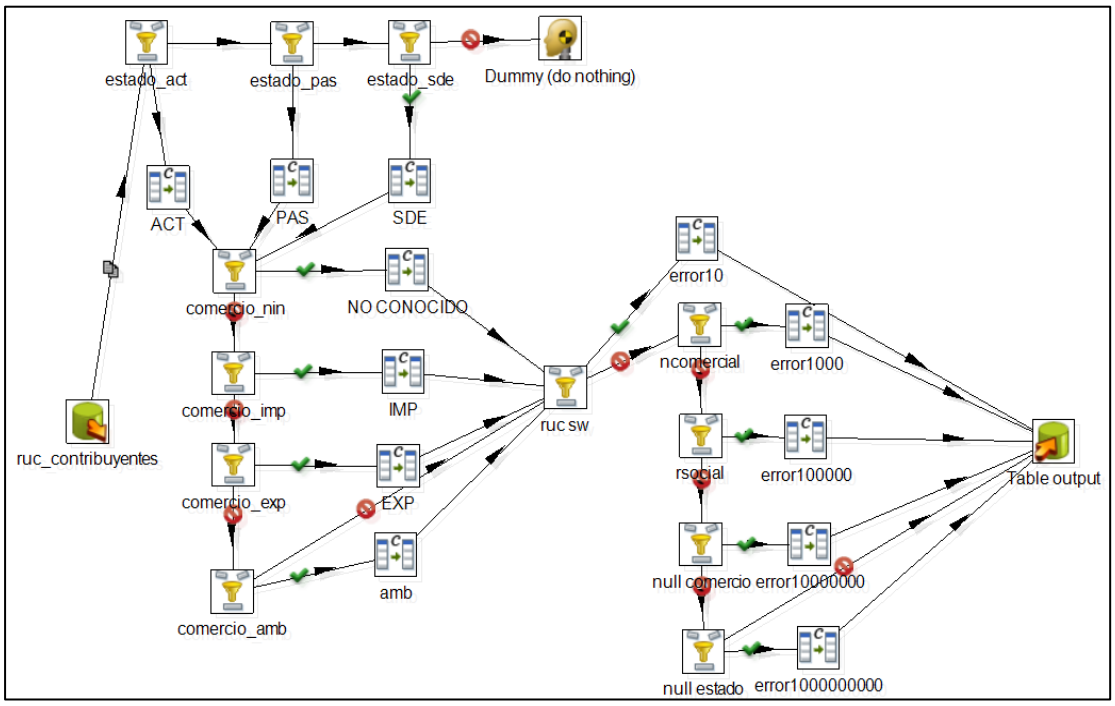


Grafico N: 10

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	SDE	3

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

- Se valida que tanto razón social como 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 la razón social 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.

4. Migracion1_ulegal

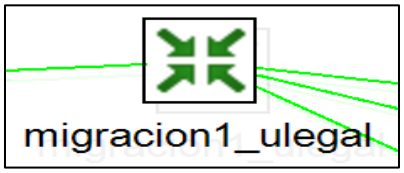


Grafico N: 11

Dentro de esta transformación existen los siguientes objetos:

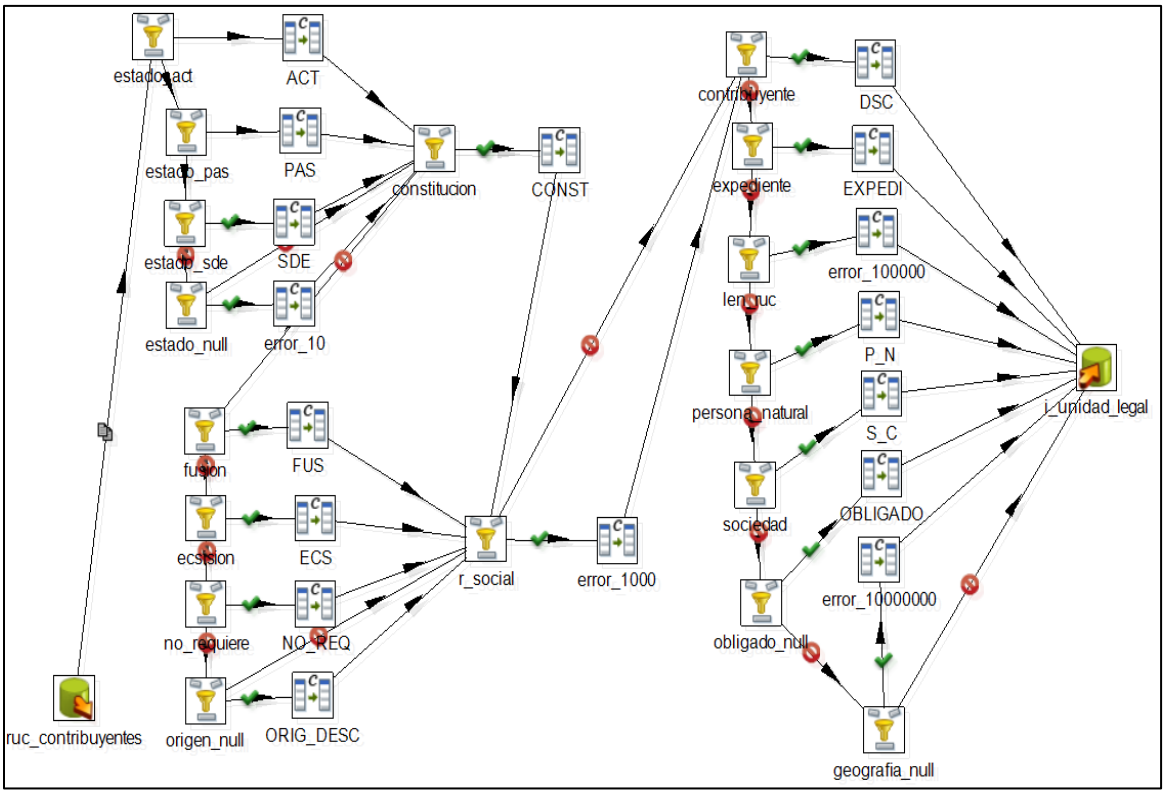


Grafico N: 12

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

5. Migracion1_ulocal

Grafico N: 13

La transformación de unidad local tiene los siguientes objetos:

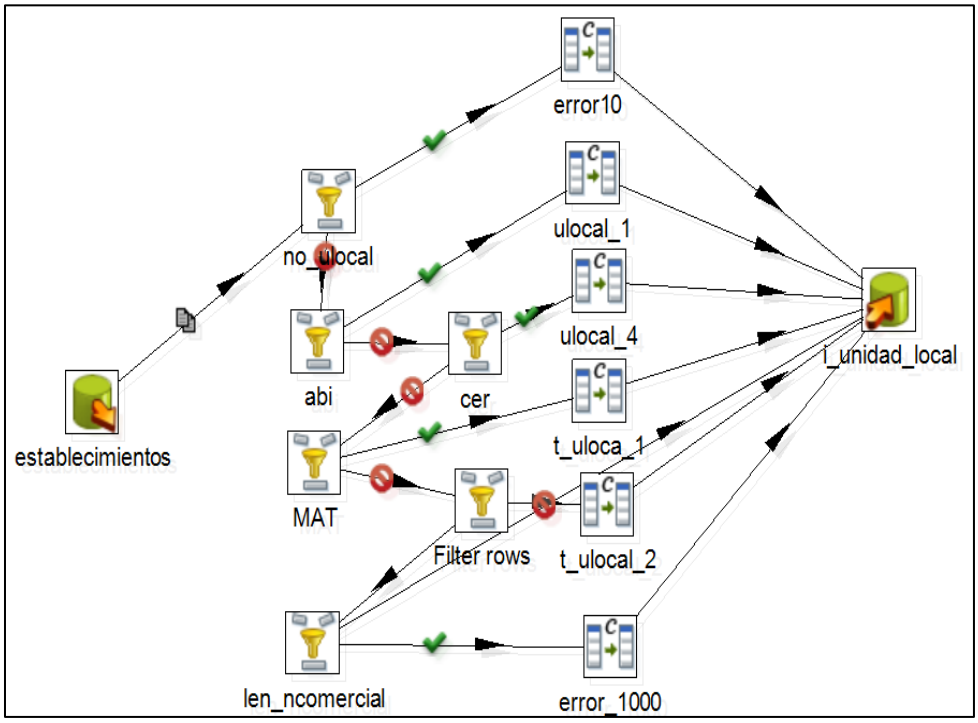


Grafico N: 14

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

6. Migracion1_emp_act_economica

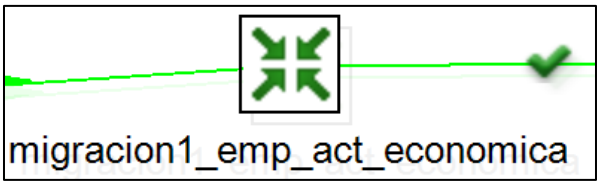


Grafico N: 15

La transformación de migracion1_emp_act_economica tiene los siguientes objetos:

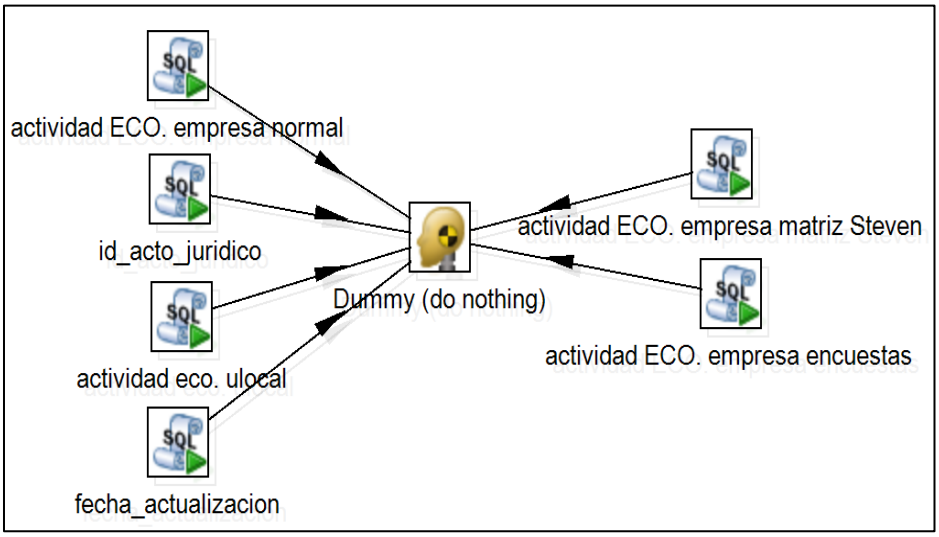


Grafico N: 16

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. ulocal": se definen los establecimientos únicos y para estos se baja la actividad económica directamente desde contribuyente. Para los demás establecimientos se pasa la actividad desde establecimientos.
- "fecha_actualizacion": trasforma y pasa la fecha de actualización a unidad legal desde contribuyentes.
- "actividad ECO empresa matiz Steven": En el DIEE se trabajó con una matriz de conversión propia para las actividades que no tienen

un mapeo en la matriz de CIIU 3 a CIIU 4; en esta fase se hace el llamado a esta matriz propia y se asigna las actividades económicas para las empresas que tienen las actividades en esta matiz.

7. Migracion1_ulocal_act_economica

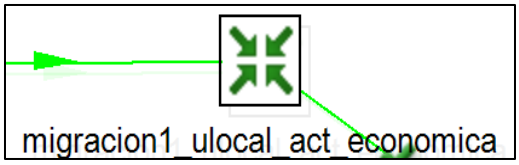


Grafico N: 17

La transformación de migracion1_ulocal_act_economica tiene los siguientes objetos:

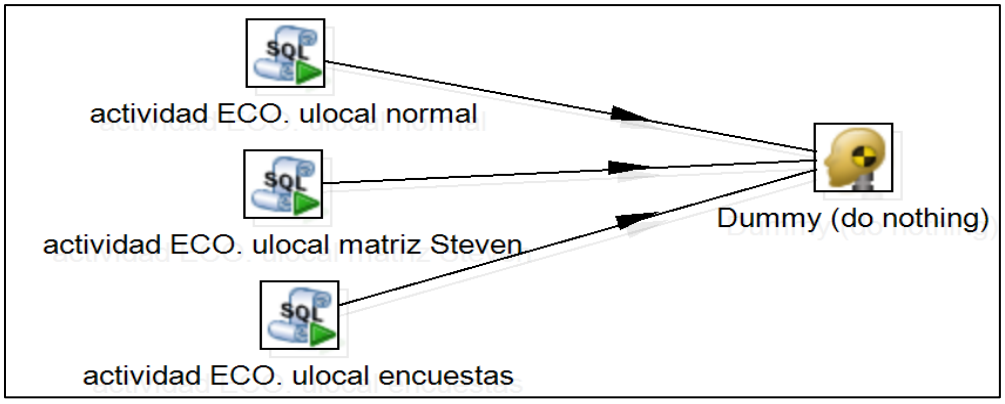


Grafico N: 18

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.

8. Migración_ulegal_catalogo

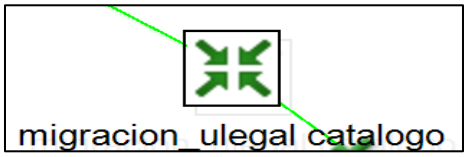


Grafico N: 19

La transformación tiene los siguientes objetos:

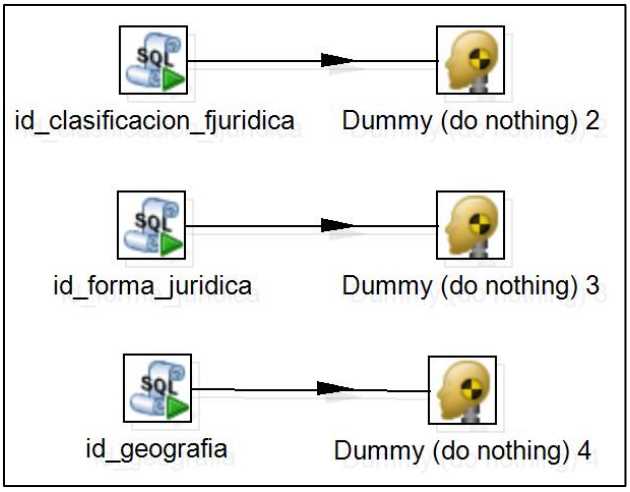


Grafico N: 20

Como se puede observar en el Gráfico N: 20 se tiene también objetos de ejecución de script´s que se detallan a continuación:

- "id_clasificacion_fjuridica": asigna el id clasificación de forma jurídica de SRI a la catalogación del DIEE
- "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
- "forma institucional": Esta clasificación ha sido adherida en la tabla f_unidad_legal, esta solicitud ha sido procesada luego de construido el ETL, por lo que no está aun incluido. 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:

Edit Data - 172.16.2.60 (172.16.2.60:5432) - diee_201310 - catalogo.d_form_institucion

File Edit View Tools Help

No limit

	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

Inicialización de variables

En la inicialización de variables intervienen las transformaciones:

Las que se encargan de llenar la información de los códigos (id) de las tablas principales del esquema *PASO* y de la fuente SRI: que son i_empresa, i_unidad_legal, i_unidad_local y ruc_contribuyentes, ruc_establecimientos, respectivamente, a partir de los datos de la base del DIEE.

9. Inicialización de variables id

Grafico N: 21

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

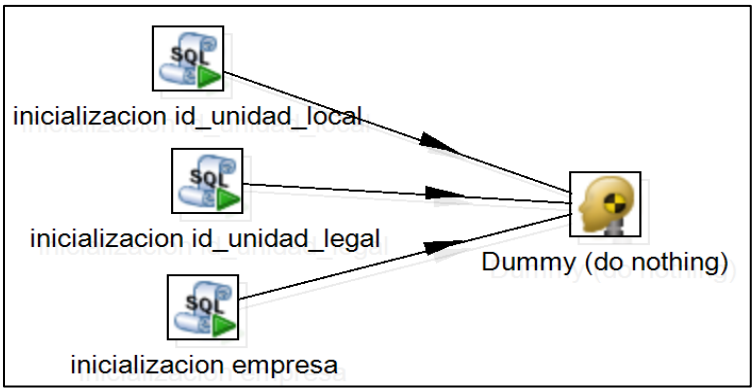


Grafico N: 22

Su objetivo principal como lo dice en el nombre es inicializar las variables en el id propio de cada unidad: empresa, local y legal.

10. Inicializa_id_SRI

Grafico N: 23

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

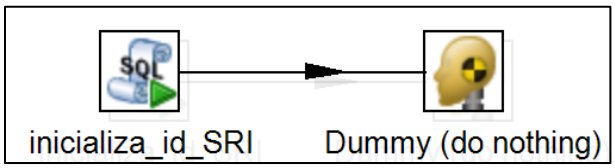
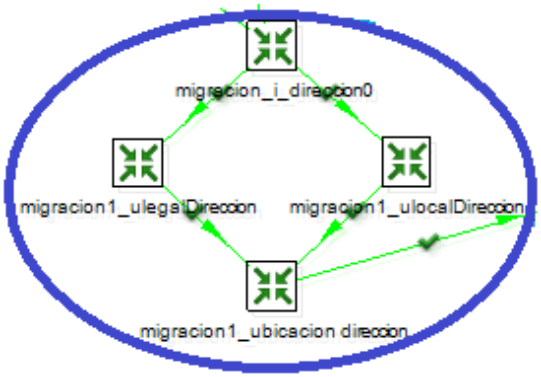


Grafico N: 24

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 del DIEE.

Migración de Dirección

En esta parte intervienen las transformaciones:



La información de dirección proviene de dos tablas del SRI "Contribuyentes" que será la información que pase a las matrices de las empresas y "Establecimientos" que será la información de las demás unidades locales, esta información irá a la tabla f_ubicacion_dirección. El SRI trae la información de dirección por columnas, en el DIEE se las transforma y cataloga para tener en una sola columna toda la información de Dirección.

11. Migración_i_direccion0

Grafico N: 25

La transformación de migración_i_direccion se compone de los siguientes objetos.

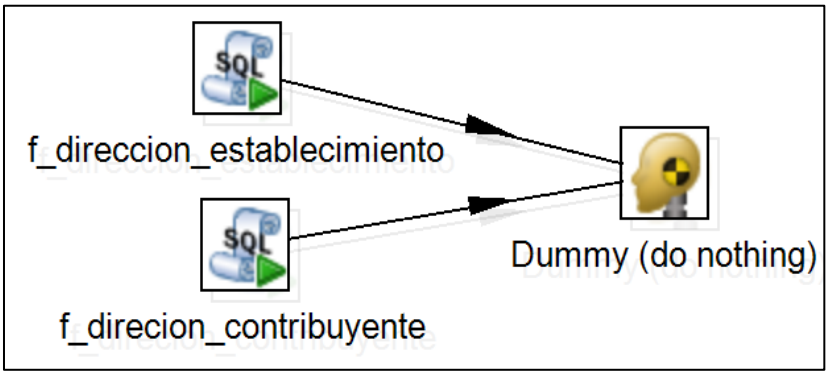


Grafico N: 26

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 llenar los datos en la tabla de i_dirección del esquema *PASO*.

12. Migracion1_ulegalDireccion

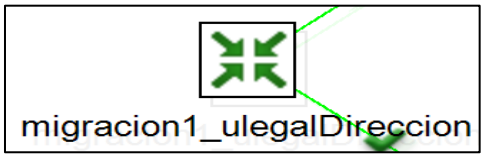


Grafico N: 27

La transformación se compone de dos objetos, estos se encargan de llenar los datos de: id_unidad_legal,numero_ruc, id_direccion en la tabla i_unidad_legal_direccion del esquema *PASO,* datos que son tomados de las tablas: ruc_contribuyentes y f_direccion.

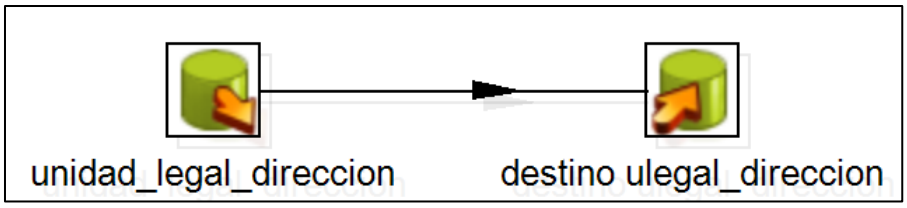


Grafico N: 28

13. Migracion1_ulocalDireccion

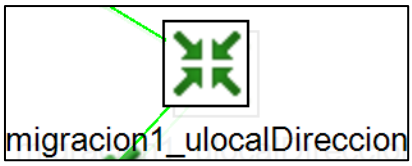


Grafico N: 29

La transformación se compone de dos objetos, estos se encargan de llenar los datos de: id_unidad_local,numero_ruc, numero_unidad_local,

id_direccion en la tabla i_unidad_local_direccion del esquema *PASO,* datos que son tomados de las tablas: ruc_establecimientos y f_direccion.

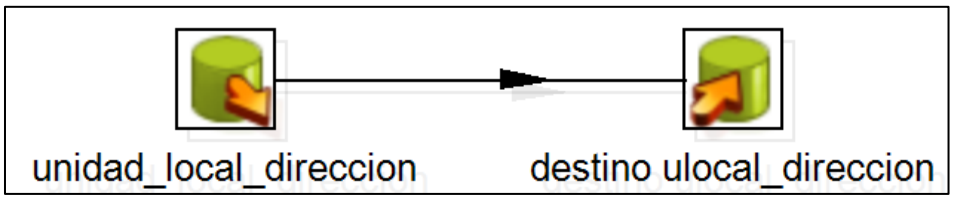


Grafico N: 30

14. Migracion1_ubicacion dirección

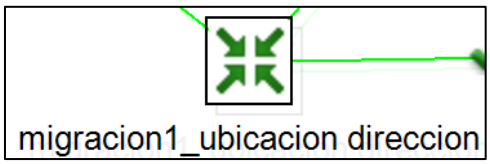


Grafico N: 31

La transformación de migración1_ubicacion_direccion se compone de los siguientes objetos.

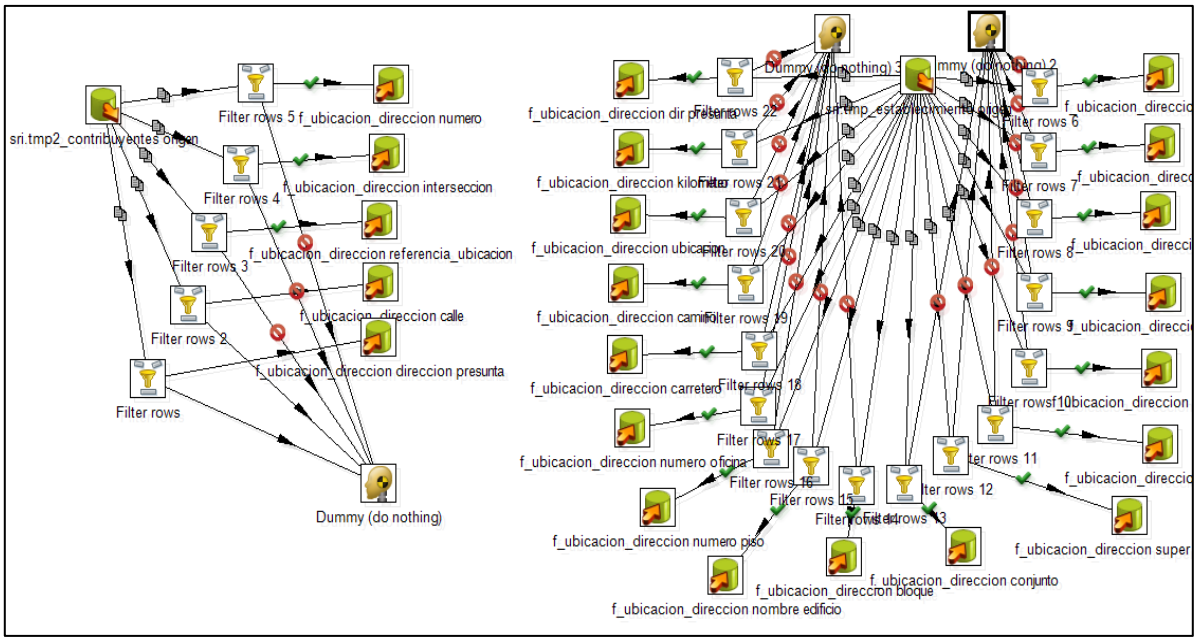


Grafico N: 32

Esta transformación se compone de varios objetos pero con objetivos en común, el de pasar la información de las direcciones de contribuyentes a empresa y de establecimientos a unidad local.

De la información de 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.

Todas las variables del SRI anteriormente detalladas, en la base del DIEE son agrupadas 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:

Edit Data - 172.16.2.60 (172.16.2.60:5432) - diee_201310 - catalo

File Edit View Tools Help

No limit

	id_tipo_direccion [PK] serial	es sr	texto character varying(25)
1	1	1	CALLE PRINCIPAL
2	2	1	NUMERO
3	3	1	INTERSECCION
4	4	1	VIA, CARRETERO, CAMINO
5	5	1	KILOMETRO
6	6	1	URBANIZACION
7	7	1	CONJUNTO
8	8	1	BLOQUE
9	9	1	NOMBRE EDIFICIO
10	10	1	NUMERO DE PISO
11	11	1	NUMERO DE OFICINA
12	12	1	CIUDADELA
13	13	1	BARRIO
14	14	1	SUPER MANZANA
15	15	1	REFERENCIA UBICACION
16	16	1	DEPARTAMENTO
17	17	1	MANZANA
18	18	1	DIRECCION PRESUNTA
19	19	1	LOTE
20	20	1	SECTOR

Grafico N: 33

15.Migracion_ulocal_catalogo_ok

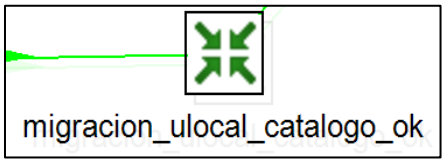


Grafico N: 34

La transformación de migración_ulocal_catalogo_ok se compone de los siguientes objetos.

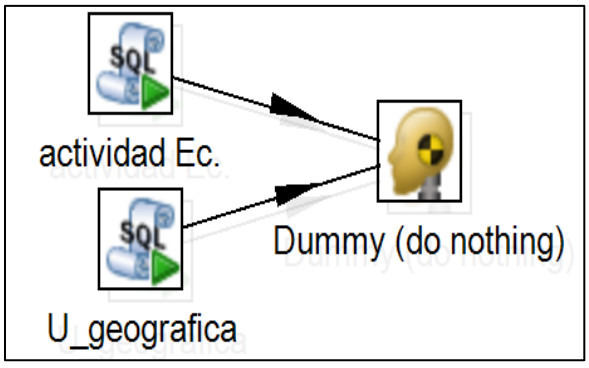


Grafico N: 35

Esta transformación tiene dos script que ejecutan lo siguiente:

- "Actividad Ec.": según la actividad económica que tiene la unidad local se actualiza su id_actividad_economica desde el respectivo catálogo de actividad económica
- "U_geografia": de la misma manera, se actualiza el id_geografía para unidad local desde el catálogo de geografía.

16. Actualización datosComparacion

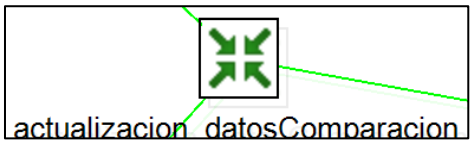


Grafico N: 36

La transformación contiene los siguientes objetos.

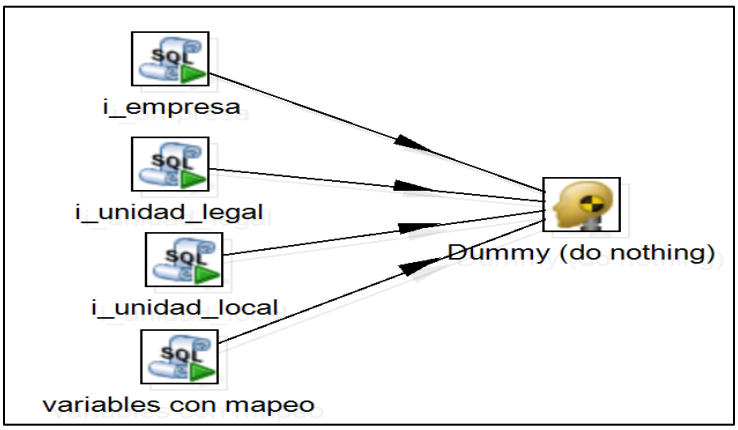
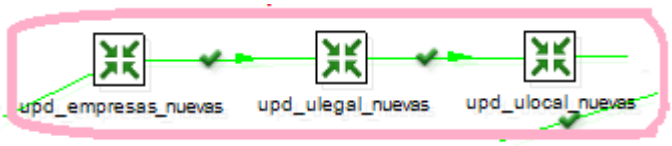


Grafico N: 37

La principal función de los scripts es pasar la información de la última base congelada al esquema paso (tablas que contienen información actualizada a 31 de diciembre del 2012) esto con la finalidad de comparar que campos se han cambiado y poder actualizarlos.

Actualización de Empresas o Establecimientos nuevos

En esta parte intervienen las transformaciones:



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*.

17.Upd_empresas_nuevas

Grafico N: 38

Actualización de empresas que ingresan al directorio.

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

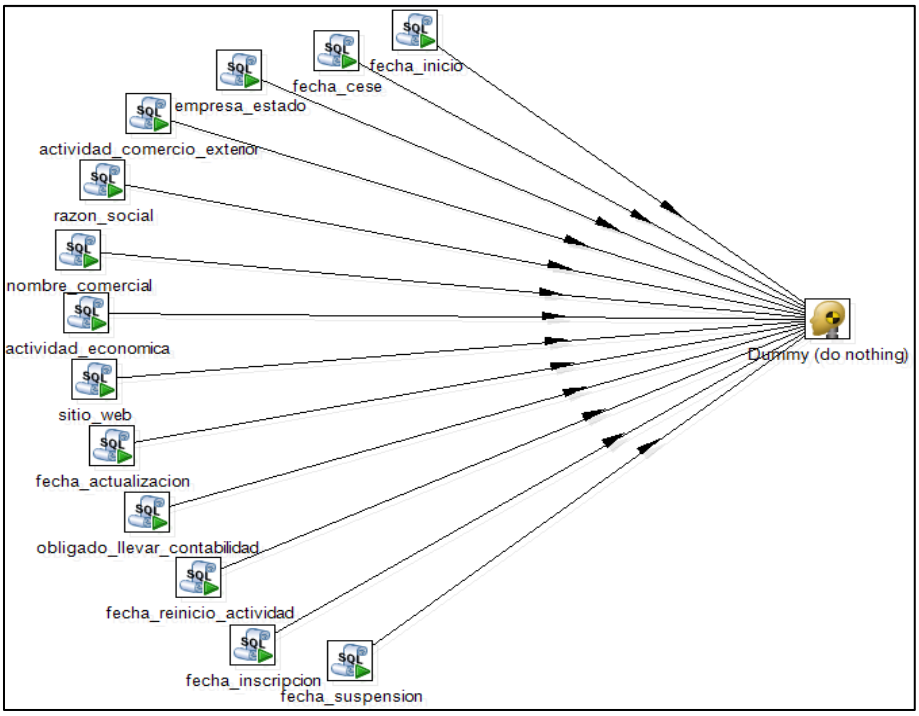


Grafico N: 39

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

18. Upd_ulegal_nuevas

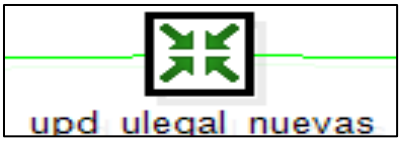


Grafico N: 40

La transformación tiene la siguiente transformación:

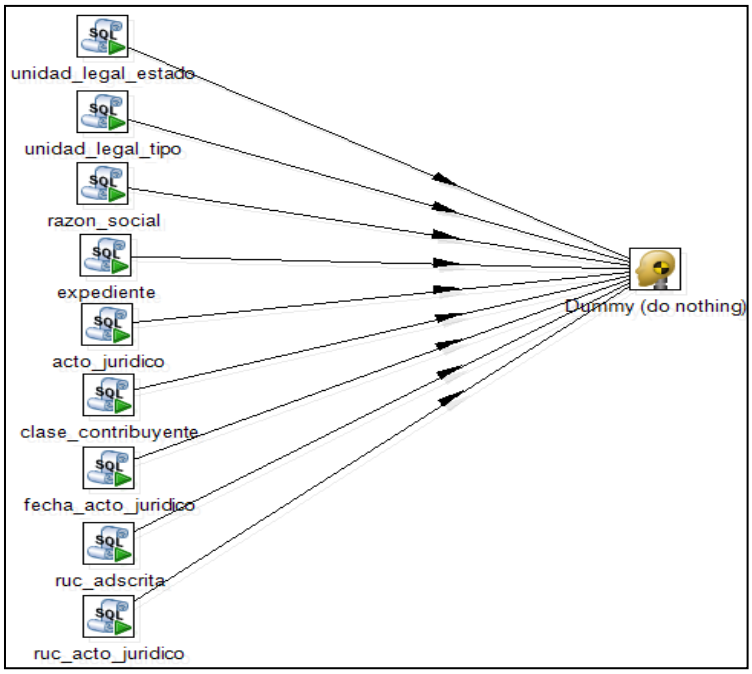


Grafico N: 41

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

19. Upd_ulocal_nuevas

Grafico N: 42

La transformación que corresponde a la siguiente transformación:

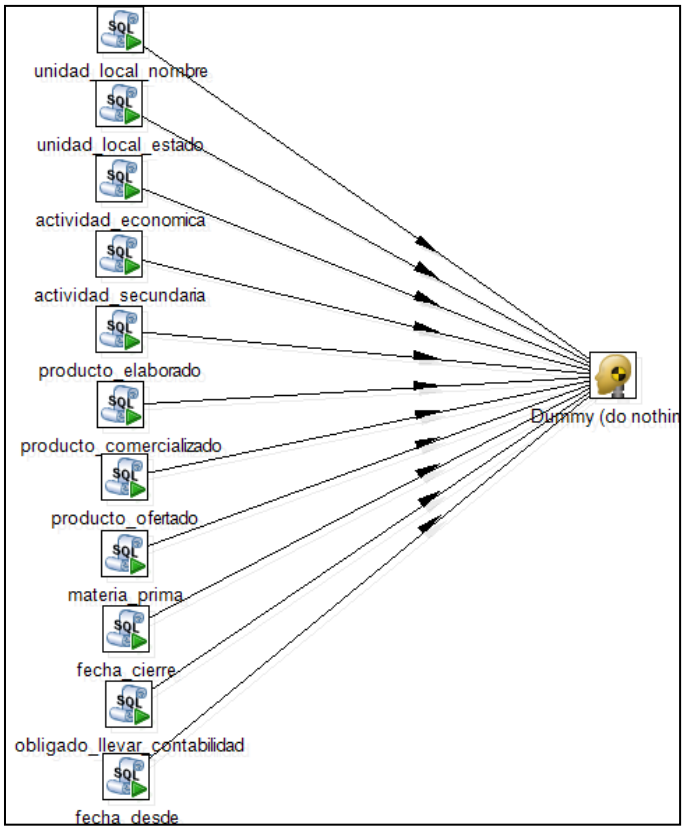


Grafico N: 43

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

20.Upd_geografia_ulegal_null

Grafico N: 44

La transformación contiene los siguientes objetos:

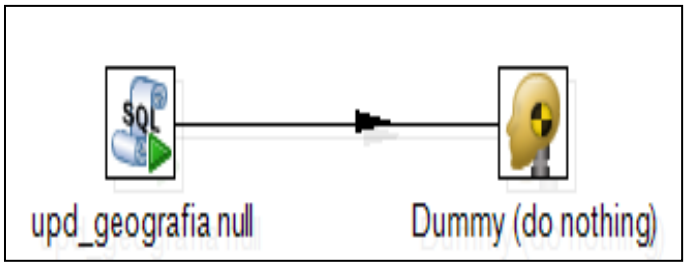
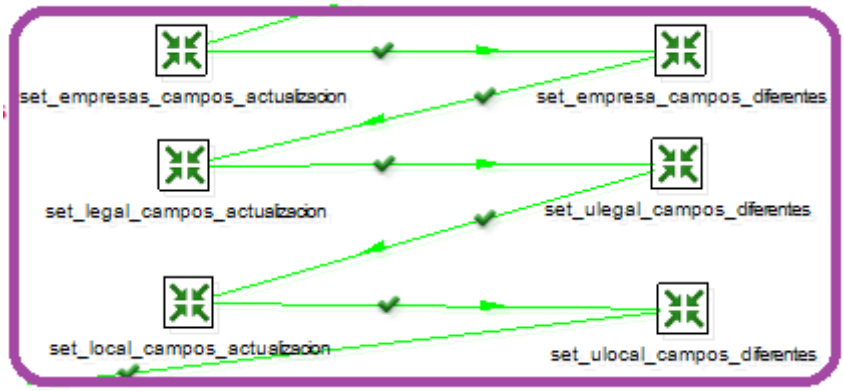


Grafico N: 45

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:



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

21. Set_empresas_campos_actualizacion

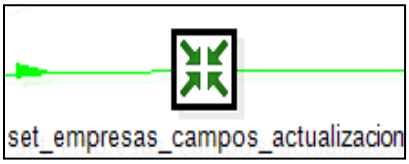


Grafico N: 46

La transformación contiene los siguientes objetos:

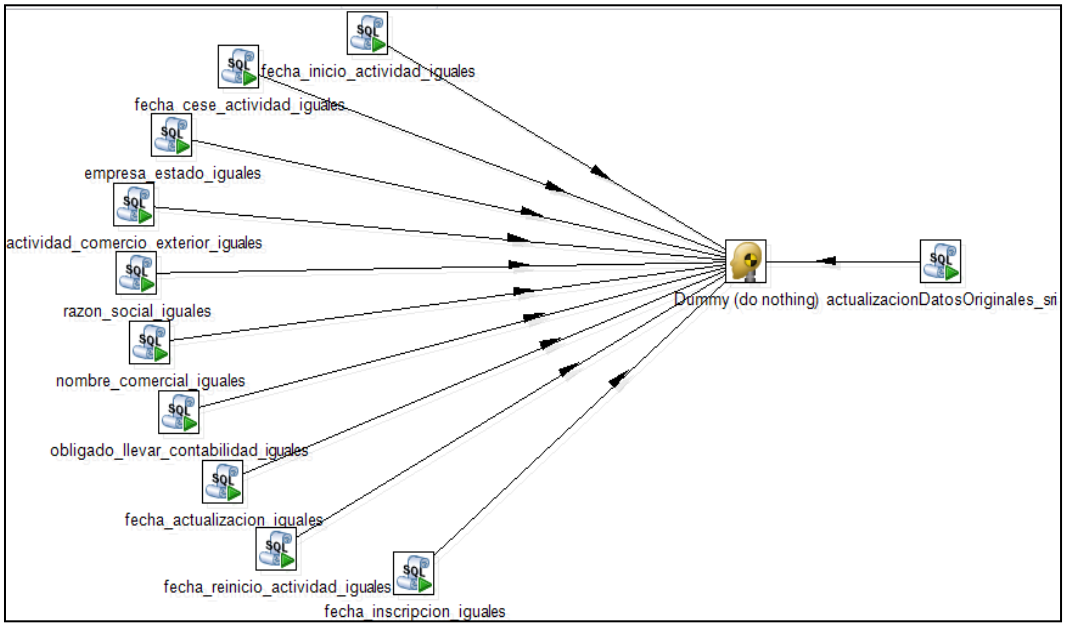


Grafico N: 47

Esta transformación pasa los datos de las variables de control idénticamente desde la tabla f_empresa hasta *PASO* i_empresa, siempre y cuando la información pertenezca a los años anteriores y que las variables indicadas en la transformación no hayan cambiado, esto se lo hace vs la captación del SRI.

Previo a este proceso, el objeto del costado derecho del Gráfico N: 46 realiza:

Actualización de datos originales: la información de la fuente SRI llega solo una descripción de los campos de comercio exterior y estado de la empresa, en esta parte de la transformación se codifica a estos dos campos para fines de administración.

Para comercio exterior la codificación queda de la siguiente manera:

comercio exterior	
código	**descripción SRI**
00	NIN
01	IMP
02	EXP
03	AMB

Para el estado de la empresa la codificación es la siguiente:

estado de empresa	
código	**descripción SRI**
1	ACT
2	PAS
3	SDE

22.Set empresa campos diferentes

Grafico N: 48

La transformación contiene los siguientes objetos:

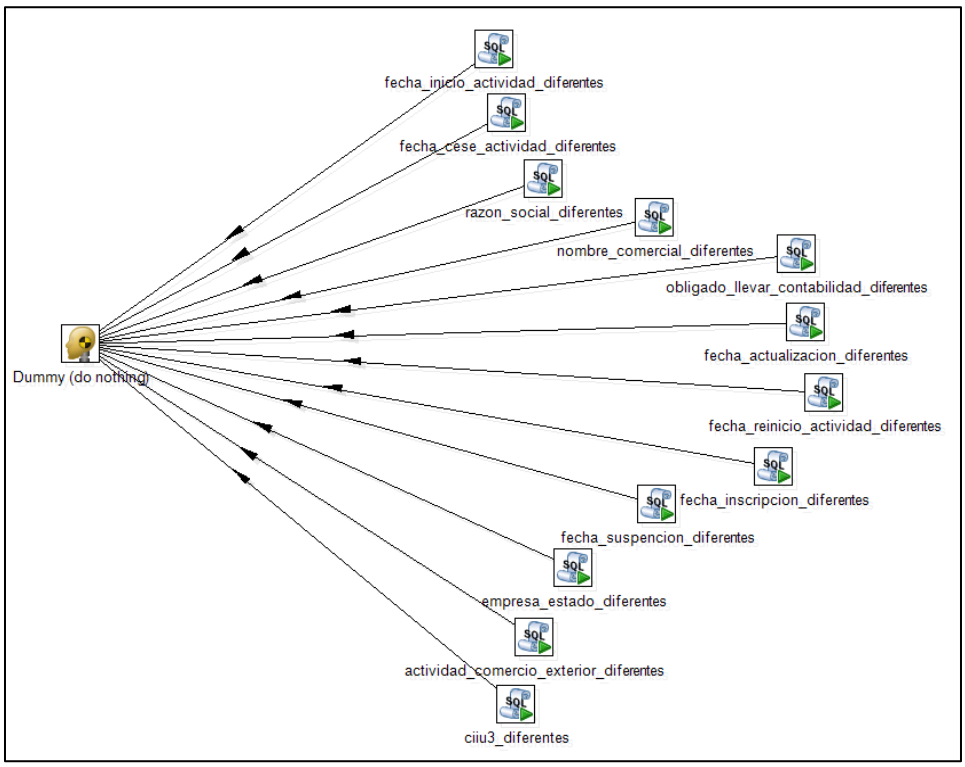


Grafico N: 49

Cuando se tienen variables diferentes entre la tabla de empresas del esquema *PASO* y la tabla de la base del DIEE, se cambia los valores según la información que nos proporciona la fuente y se actualiza también sus respectivas variables de control en la tabla f_empresa.

23.Set_legal_campos_actualizacion

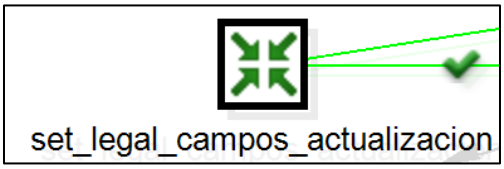


Grafico N: 50

La transformación contiene los siguientes objetos:

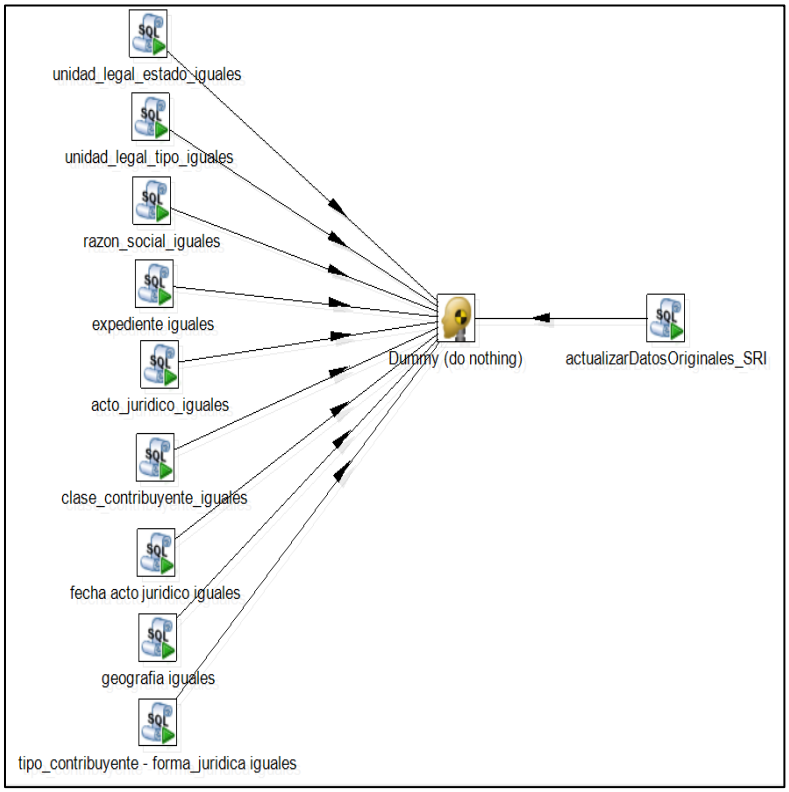


Grafico N: 51

Esta transformación actualiza la información que se tiene de años anteriores, cuando las variables indicadas en la transformación tienen la misma información en la tabla de unidad legal en el esquema *PASO* y en la de la base del DIEE, la información de las variables de control van a ser las mismas que están en la tabla f_unidad_legal.

Previo a esto, el objeto del costado derecho del Gráfico N: 46 realiza una:

Actualización de datos originales: la información de la fuente SRI llega solo una descripción de los campos de estado de la unidad legal y tipo de unidad legal, en esta parte de la transformación se codifica a estos campos para fines de administración. La codificación interna para el estado de unidad legal es la misma que para el estado de empresa y para tipo de unidad legal se tiene: *1* para Persona Natural y *2* para Persona Jurídica.

24. Set_ulegal_campos_diferentes

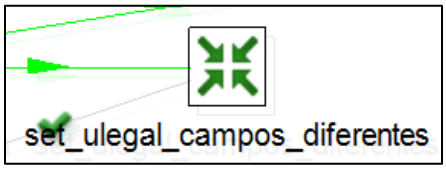


Grafico N: 52

La transformación contiene los siguientes objetos:

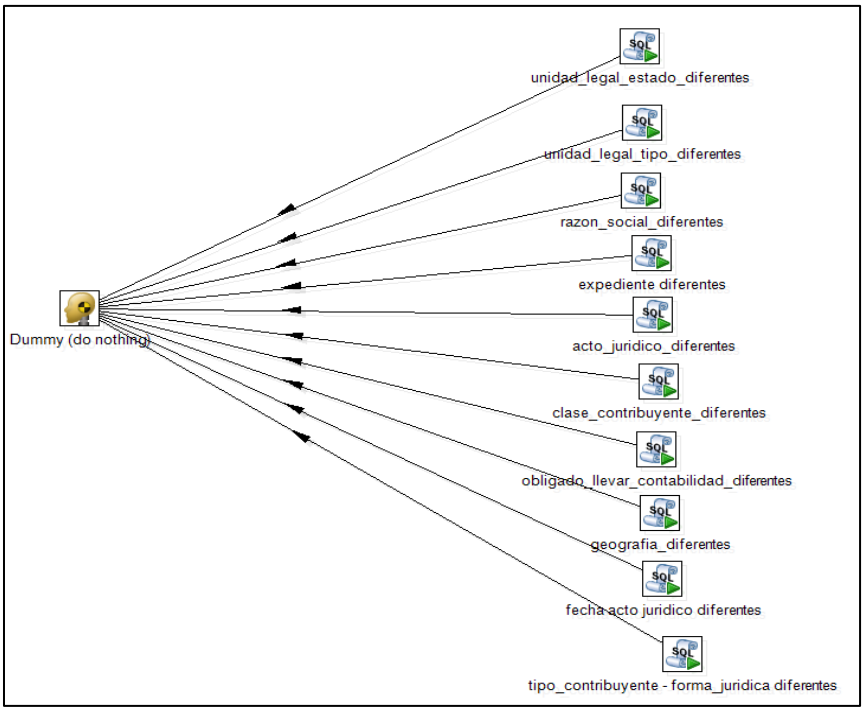


Grafico N: 53

Cuando se tienen variables diferentes entre la tabla de unidad legal del esquema PASO y la de la base del DIEE, se cambia los valores según la información que nos proporciona la fuente y se actualiza también sus respectivas variables de control en la tabla f_unidad_legal.

25. Set_local_campos_actualizacion

Esta transformación actualiza la información que se tiene de años anteriores, cuando las variables indicadas en la transformación tienen la misma información en la tabla de unidad legal en el esquema *PASO* y en la de la base del DIEE, las variables de control van a ser las mismas que están en la tabla f_unidad_local.

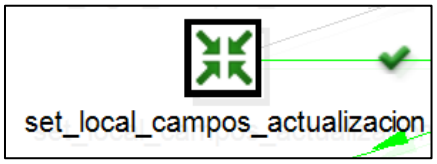


Grafico N: 54

La transformación contiene los siguientes objetos

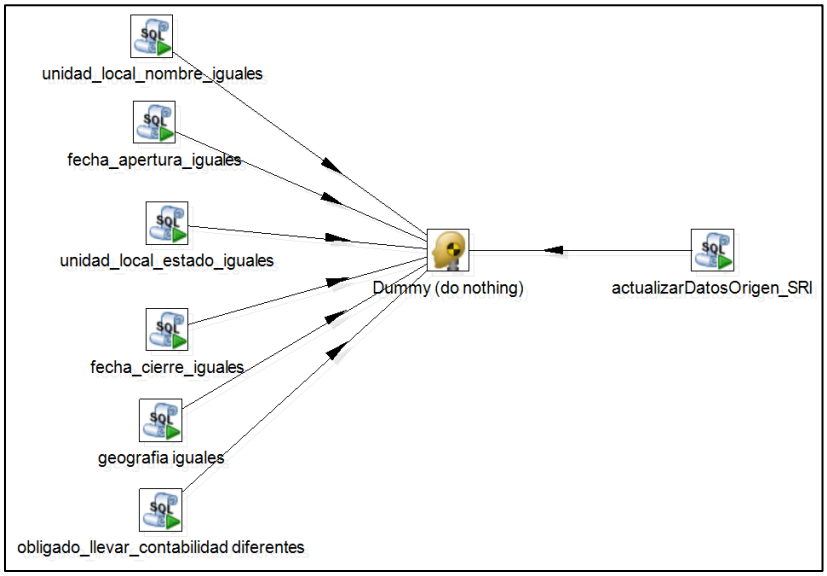


Grafico N: 55

Esta transformación trabaja con la información que se tiene de años anteriores, y su función es que cuando las variables indicadas en la transformación tienen el mismo nombre, las variables de control van a ser las mismas que están en la tabla f_unidad_local.

26.Set_ulocal_campos_diferentes

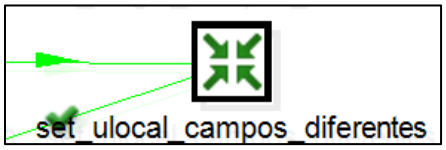


Grafico N: 56

La transformación contiene los siguientes objetos:

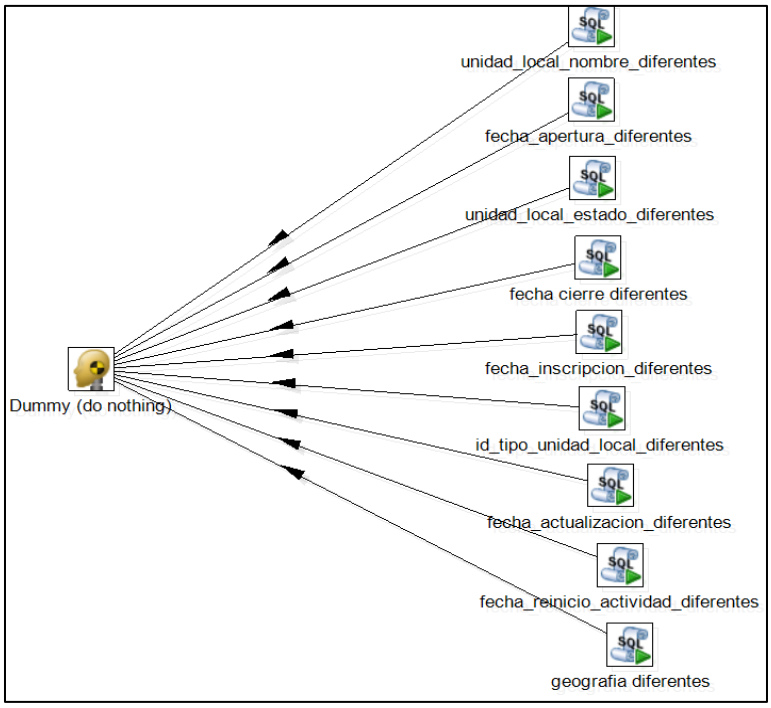


Grafico N: 57

Cuando se tienen variables diferentes entre la tabla de unidad legal del esquema PASO y la de la base del DIEE, se cambia los valores según la información que nos proporciona la fuente y se actualiza también sus respectivas variables de control de la tabla f_unidad_local.

27.Ventas_101_102

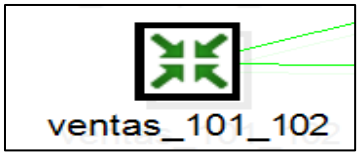


Grafico N: 58

La transformación contiene los siguientes objetos:

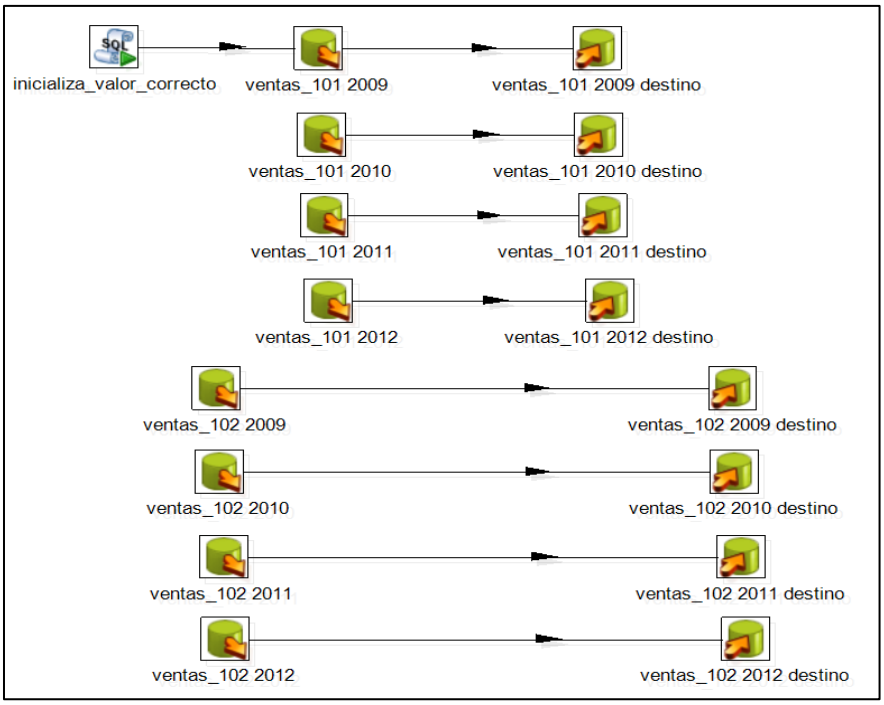


Grafico N: 59

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

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 este año, al directorio del 2012. La información que se utiliza para extraer los registros de ventas que el SRI proporciona, corresponde a las siguientes tablas:

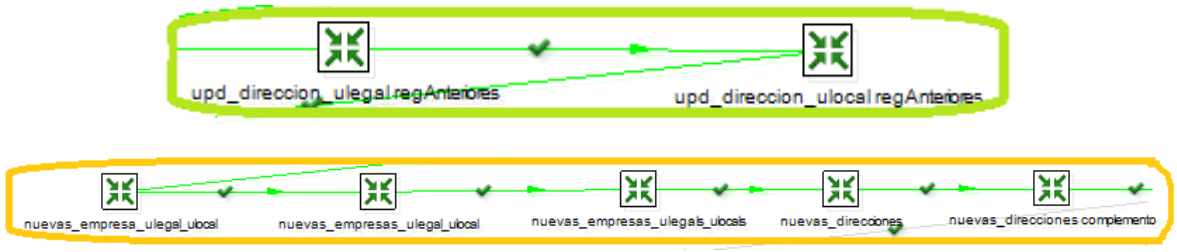
- owb_mvw_ine_anexo3_estruc_f101
- owb_mvw_ine_anexo3_estruc_f102

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: 57, 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:



En esta sección se actualizan los datos de dirección de las empresas que ya existían en el DIEE, para esto se iguala dato a dato y se distingue cuáles han cambiado. Los datos que se distinguen como diferentes se los actualiza tanto el dato como de sus variables de control. Y los registros nuevos se migran del esquema *PASO* a la tabla definitiva de la base del DIEE.

28. Upd_direccion_ulegal regAnteriores

Grafico N: 60

La transformación contiene los siguientes objetos:

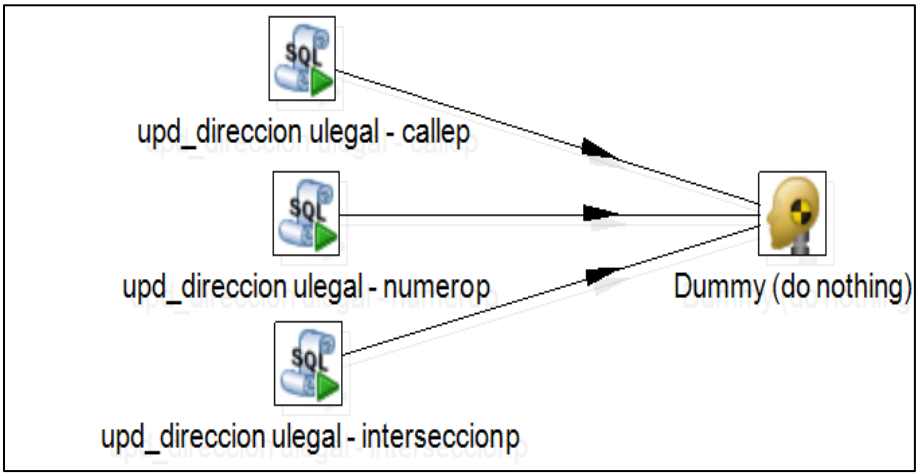


Grafico N: 61

En esta transformación se actualizan, cuando son diferentes, las variables de dirección que determinan si la dirección ha cambiado, las mismas que son: calle, número e intersección, y sus respectivas variables de control en la tabla ubicacion_direccion del esquema *PASO*, a partir de la información proporcionada por la fuente SRI, donde se almacena la descripción de todas las direcciones, en este caso para unidad_legal.

29.Upd_direccion_ulocal regAnteriores

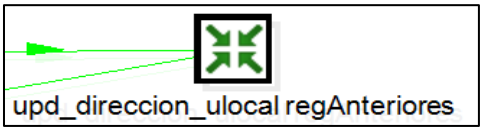


Grafico N: 62

La transformación contiene los siguientes objetos:

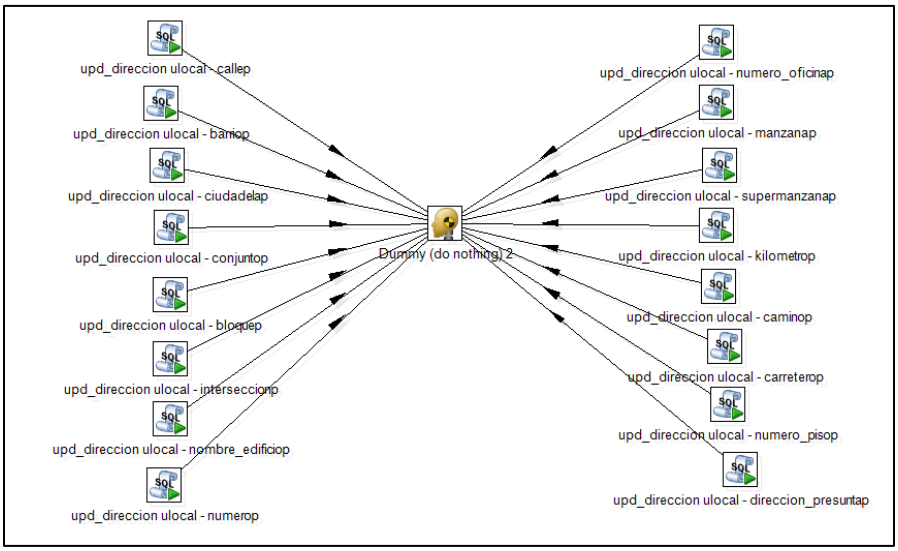


Grafico N: 63

Se actualizan las variables referentes a dirección cuando la información de la tabla ubicación_dirección del esquema *PASO* es diferente a la proporcionada por la fuente SRI. Si se tiene este caso se actualizan también las variables de control, para direcciones de unidad_local.

30. Nuevas_empresa_ulegal_ulocal

Grafico N: 64

La transformación contiene los siguientes objetos

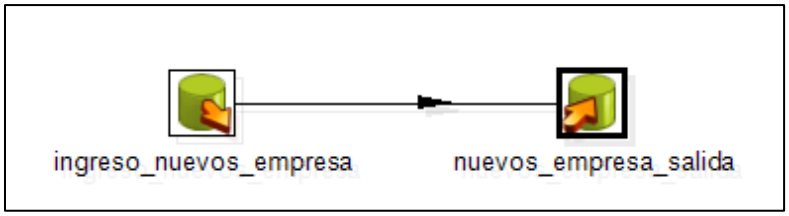


Grafico N: 65

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.

31. Nuevas_empresas_ulegal_ulocal

Grafico N: 66

La transformación contiene los siguientes objetos:

Grafico N: 67

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.

32. Nuevas_empresa_ulegals_ulocals

Grafico N: 68

La transformación contiene los siguientes objetos:

Grafico N: 69

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 direcciones

En esta parte están las transformaciones:

- Nuevas_direcciones
- Nuevas_direcciones_complemento

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

33. Nuevas_direcciones

Grafico N: 70

La transformación contiene los siguientes objetos:

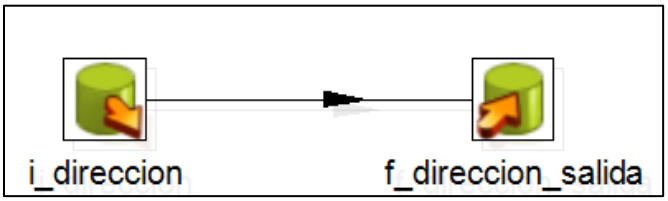


Grafico N: 71

Los objetos del Gráfico N: 70, 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.

34. Nuevas_direcciones_complemento

Grafico N: 72

La transformación contiene los siguientes objetos:

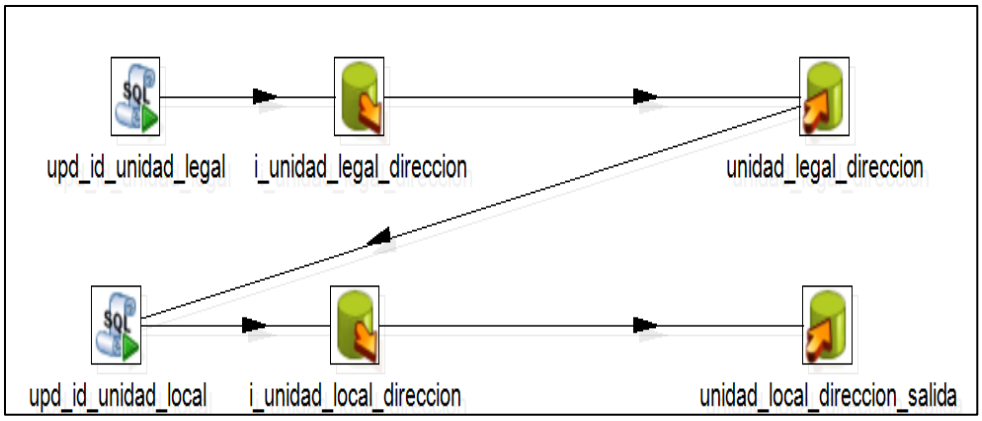


Grafico N: 73

En el Gráfico N: 72, se transfiere la nueva información referente a dirección de las tablas i_unidad_local_direccion y i_unidad_legal_direccion del esquema *PASO* a las tablas de f_unidad_local_direccion y f_unidad_legal_direccion de la base del DIEE, para ello se asocian los códigos de unidad_local y unidad_legal respectivamente y el código de la dirección (id_direccion) de la tabla f_dirección ingresada anteriormente.

35. Ingreso_contactos

Grafico N: 74

La transformación contiene los siguientes objetos:

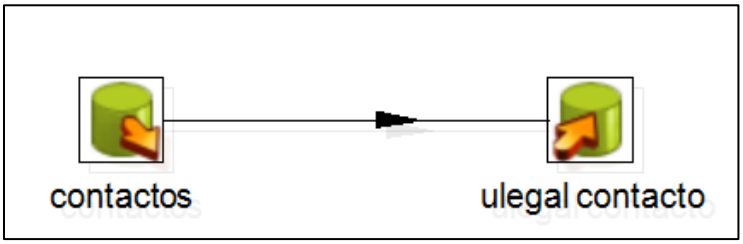


Grafico N: 75

Los objetos del Gráfico N: 74, tienen la finalidad de pasar los datos de contactos desde la tabla de ruc_contadores del SRI a la tabla de f_contacto_ulegal del DIEE, relacionando el número de RUC del contador, con el RUC de unidad legal, para tomar la información correcta, esto se realiza para las nuevas empresas que ingresan al directorio. El nombre del contacto tiene que ir con sus respectivas variables de control, para posibles actualizaciones futuras.

Empleados

En esta parte se tiene las trasformaciones:

Estas transformaciones tienen la tarea de realizar los cálculos para obtener el número de empleados hombres y mujeres por empresa y establecimiento, datos inicialmente almacenados en la tabla de promedios_iess, para que luego esta información se migre a las tablas definitivas de la base del DIEE: f_empresa_empleados, f_unidad_local_empleados.

36. F_empresa_empleados2012

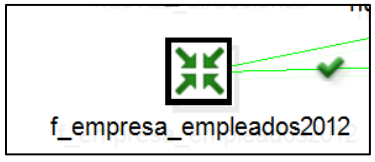


Grafico N: 76

La transformación contiene los siguientes objetos:

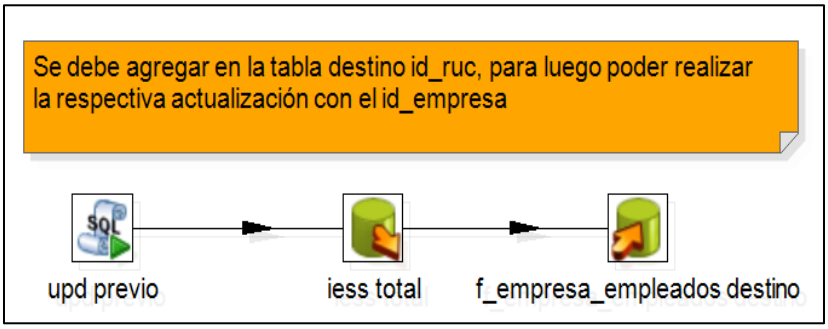


Grafico N: 77

En el Gráfico N: 76, se tiene un paso previo que consiste en encerar la variable id_empresa de la tabla de captación del IESS llamada, promedios_iess, para reemplazarlo por la información que se tiene en la tabla f_empresa de la base del DIEE, adicional a esto como se muestra en la nota, primero hay que agregar los RUC's de las empresas.

En el objeto nombrado como *iess total* se realizan los cálculos para empleados hombres, mujeres y el total de la suma entre ambas variables, esto para empleados de Empresas, finalmente estos valores serán transferidos a la tabla f_empresa_empleados de la base del DIEE.

37.F_ulocal_empleados2012

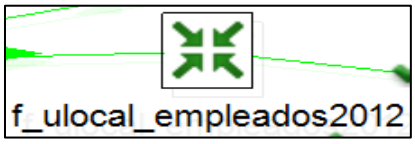


Grafico N: 78

La transformación contiene los siguientes objetos:

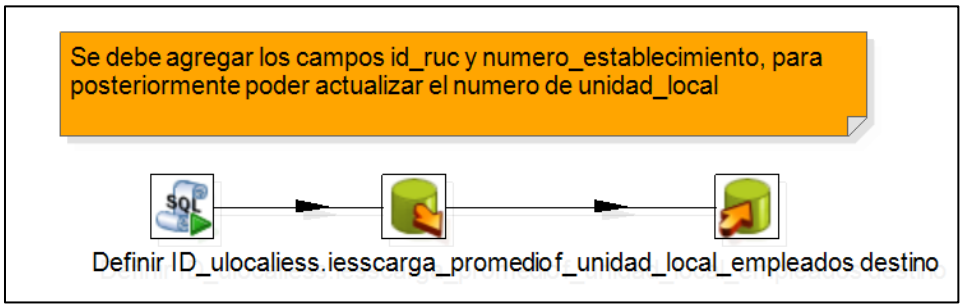


Grafico N: 79

En el Gráfico N: 78, se tiene un paso previo llamado *Definir ID_ulocal* que consiste en encerar la variable id_unidad_local de la tabla de captación del IESS llamada, promedios_iess, para reemplazarlo por la información que se tiene en la tabla f_unidad_local de la base del DIEE, adicional a esto como se muestra en la nota, primero hay que agregar el campo de RUC y número de establecimiento para poder asociar correctamente los registros de 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.

38. F_ulocal_empleados2012_9000

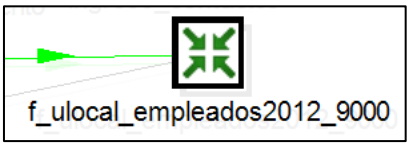


Grafico N: 80

La transformación contiene los siguientes objetos:

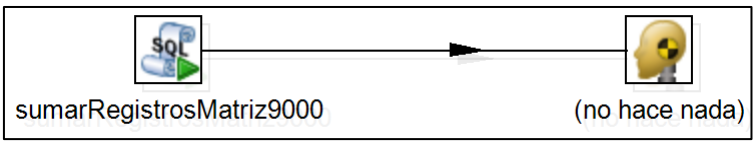


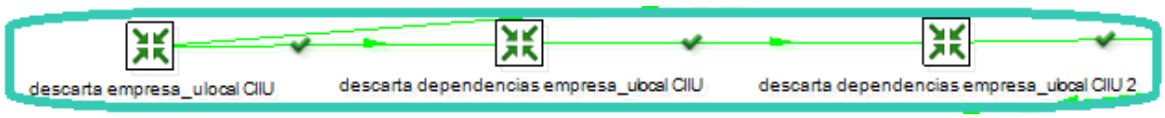
Grafico N: 81

En el Gráfico N: 80 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 crea la matriz y se suma la información de empleados de los establecimientos con número de unidad local 9000.

Descarta empresas y establecimientos

Se tiene las siguientes transformaciones:



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.

39. Descarta_empresa_ulocal CIIU

Grafico N: 82

La transformación contiene los siguientes objetos:

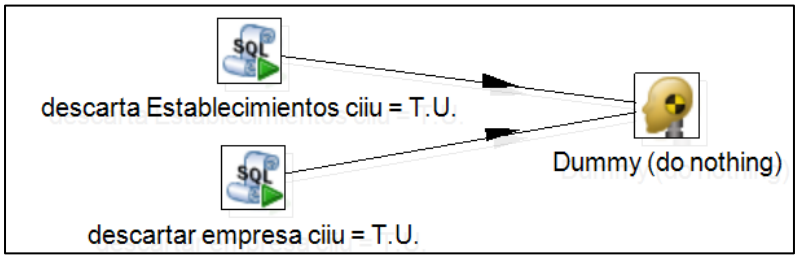


Grafico N: 83

En el Gráfico N: 82 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_2013 con el texto 'NO', de igual manera el campo nota_2013, se actualiza indicando el por qué ha sido descartada la empresa o establecimiento.

40. Descarta dependencias empresa_ulocal CIIU

Grafico N: 84

La transformación contiene los siguientes objetos:

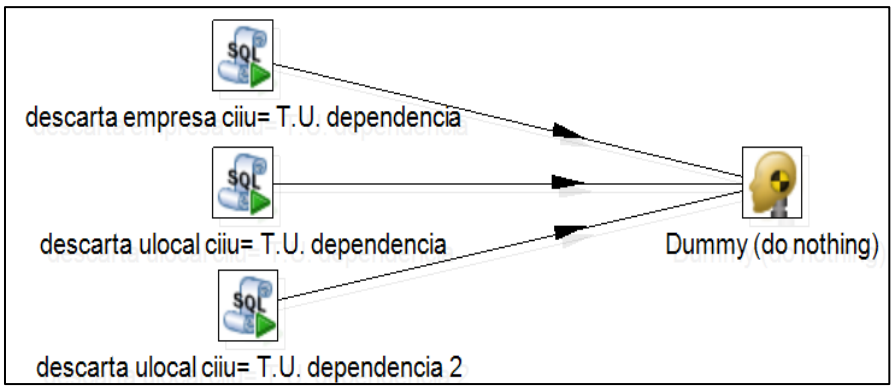


Grafico N: 85

El Gráfico N: 84 se procede a descartar las empresas y establecimientos que dependan de las que fueron descartadas en la transformación del Gráfico N: 81 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_2013 con el texto 'NO', de igual manera el campo nota_2013, se actualiza indicando el por qué ha sido descartada la empresa o establecimiento.

41. Upd_numeroUnidadesLocales

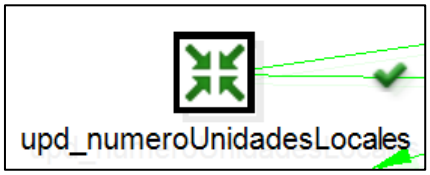


Grafico N: 86

La transformación contiene los siguientes objetos:

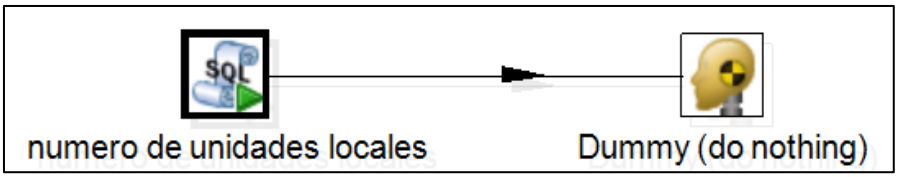


Grafico N: 87

En el Gráfico N: 86 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.

42. Actualización_datos_CPC

Grafico N: 88

La transformación contiene los siguientes objetos:

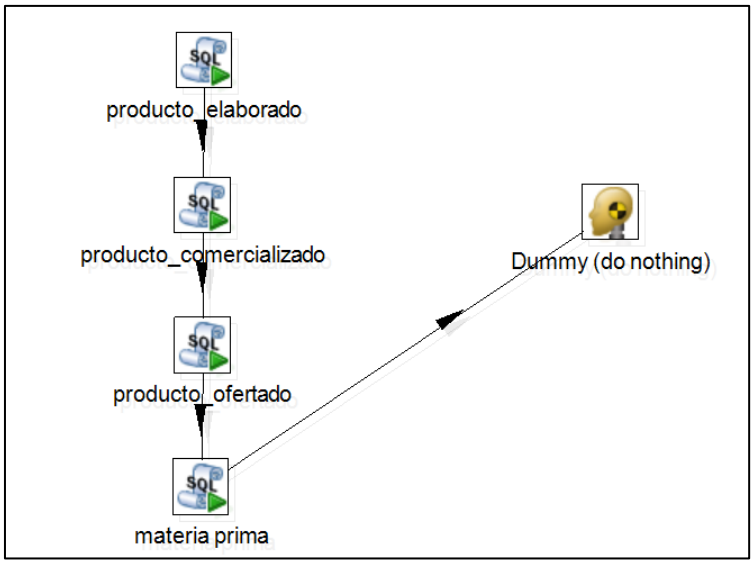


Grafico N: 89

En el Gráfico N: 88 se actualiza la información a nivel de CPC, es decir se actualiza el campo producto_elaborado y sus respectivas variables de control. Si dicho campo no tiene información se coloca el valor de '0000000', y se enceran sus variables de control. Si el campo producto_elaborado ya tiene el valor de '0000000' tan solo se enceran sus variables de control

43.Migracion medioComunicacion previo

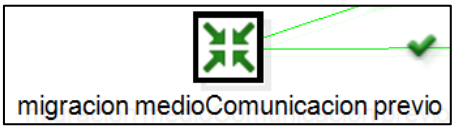


Grafico N: 90

La transformación contiene los siguientes objetos:

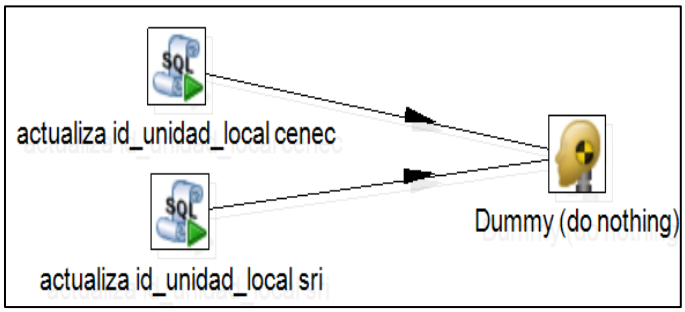


Grafico N: 91

El Gráfico N: 90 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.

44. Migración medioComunicacion

Grafico N: 92

La transformación contiene los siguientes objetos:

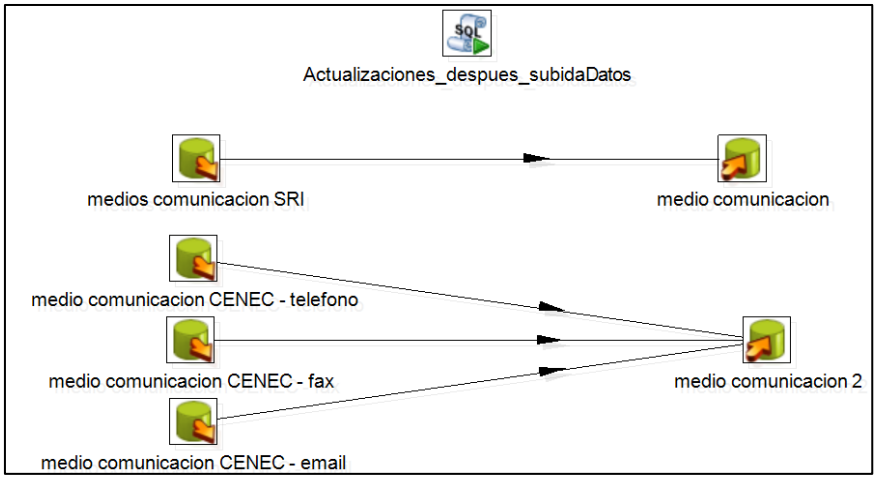


Grafico N: 93

Para procesar la información referente a medios de comunicación, se empieza tomando la información de la Fuente SRI como se puede ver en el Gráfico N: 92, los medios de comunicación del SRI se extrae de la tabla temporal creada previamente en el Gráfico N: 89, llamada tmp_20130704_contactos_sri, se definen también las variables de control y toda esta información se transfiere a la tabla f_medio_comunicación de la base del DIEE.

Luego de la fuente CENEC se extraen 3 tipos de medios de comunicación que son: teléfono, email y fax de la tabla temporal creada previamente también en el Gráfico N: 89, llamada tmp_20130704_contactos_cenec, a esta información se asigna el código del tipo de contacto

(id_tipo_medio_contacto) es decir; 6: para teléfono, 2: para fax y 3: para email y se establecen las variables de control respectivas, para luego transferir esta información a la tabla f_medio_comunicación.

Luego de la subida de toda la información se realiza un proceso de limpieza de los medios de comunicación, donde se aplica una serie de reglas que se detallan en el documento: "*Plan de Validación y Tabulación*".

45. ClasificacionEmpleadosVentas_empresaulocal

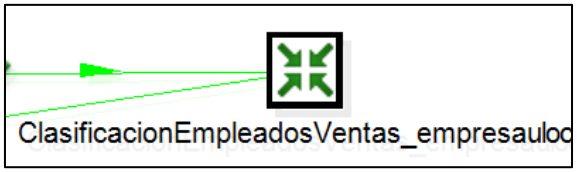


Grafico N: 94

La transformación contiene los siguientes objetos:

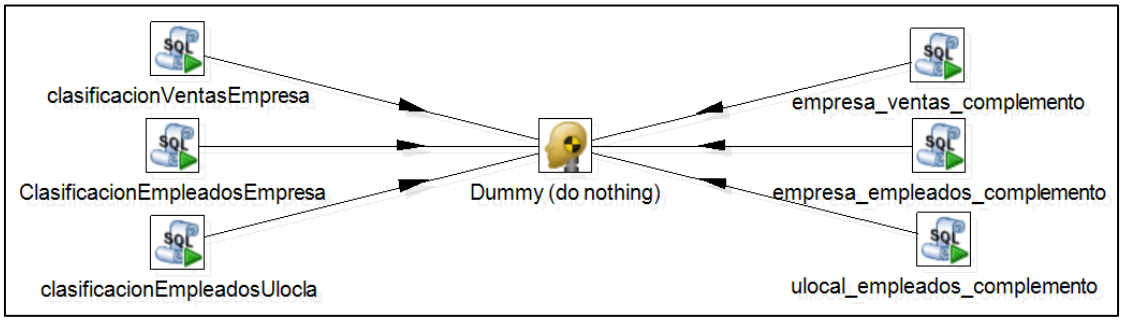


Grafico N: 95

En esta transformación se asignan los estratos para de 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

46.Upd_20131231_empresa_ulegal_ulocal

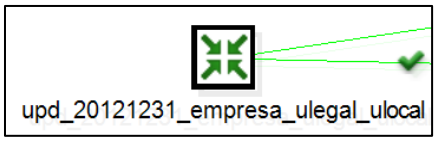


Grafico N: 96

La transformación contiene los siguientes objetos:

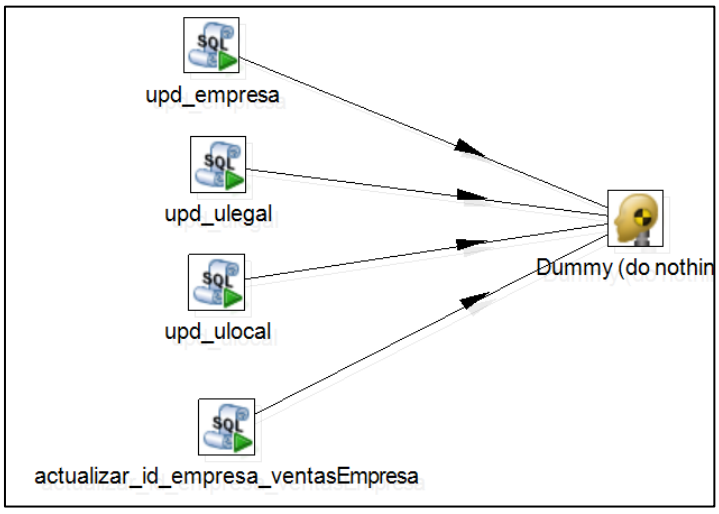


Grafico N: 97

En el Gráfico N: 96 se tiene la actualización de variables significativas de las tablas principales de la base de datos del DIEE, estas tablas son:

- ***f_empresa***, donde se actualizan los datos de las variables que pueden ser cambiadas con regularidad, según se ha establecido en el DIEE, como son: fecha_inicio_actividad, fecha_cese_actividad, fecha_reinicio_actividad, fecha_inscripcion, fecha_actualizacion, razon_social, nombre_comercial, sitio_web, id_empresa_estado, id_actividad_comercio_exterior, id_actividad_economica y obligado_llevar_contabilidad. Estas variables son actualizadas con la información de la tabla i_empresa del esquema *PASO,* pero solo cuando esta información ha sido modificada.

- ***f_unidad_legal,*** de igual manera aquí se actualizan los datos de las variables que cambian con regularidad, según se ha establecido en el DIEE, las cuales son: unidad_legal_tipo, unidad_legal_estado, razon_social, expediente, id_acto_juridico, fecha_acto_juridico, id_clase_contribuyente, ruc_adscrita y ruc_acto_juridico. Estas variables son actualizadas con la información de la tabla i_unidad_legal del esquema *PASO,* cuando dicha información ha sido modificada.

- ***f_unidad_local,*** se actualizan los datos de las variables: unidad_local_nombre, id_unidad_local_estado, id_actividad_economica, fecha_cierre, fecha_apertura y obligado_llevar_contabilidad. Estas variables son actualizadas con la información de la tabla i_unidad_local del esquema *PASO,* cuando dicha información ha sido modificada.

- ***f_empresa_ventas,*** en esta tabla se actualiza únicamente la variable id_empresa para asegurarse que la información de ventas corresponda a la empresa indicada, este dato se actualiza con el id_empresa de la tabla f_empresa de la base del DIEE.

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.