



Islas Canarias
Del 15 al 19 de noviembre de 2021

istac

INSTITUTO CANARIO
DE ESTADÍSTICA

XXI Jornadas de Estadística de las Comunidades Autónomas

Automatización de hilos de computación

Esther Torres Medina, ISTAC
José Carlos Gil Carreras, EVM
Miguel Núñez Pereyra, EVM
Idafen Santana Pérez, ULPGC
Rita Díaz Adán, ARTE

<https://www.linkedin.com/company/istac>



@istac_es



Índice

- Objetivos
- Nuestra infraestructura de datos
- Proceso y herramientas
- Lo que está por venir
- Casos de uso
 - Demandantes de empleo
 - Calidad del aire

Objetivos

- Optimizar la gestión de nuestro tiempo y recursos para dedicarlos a:

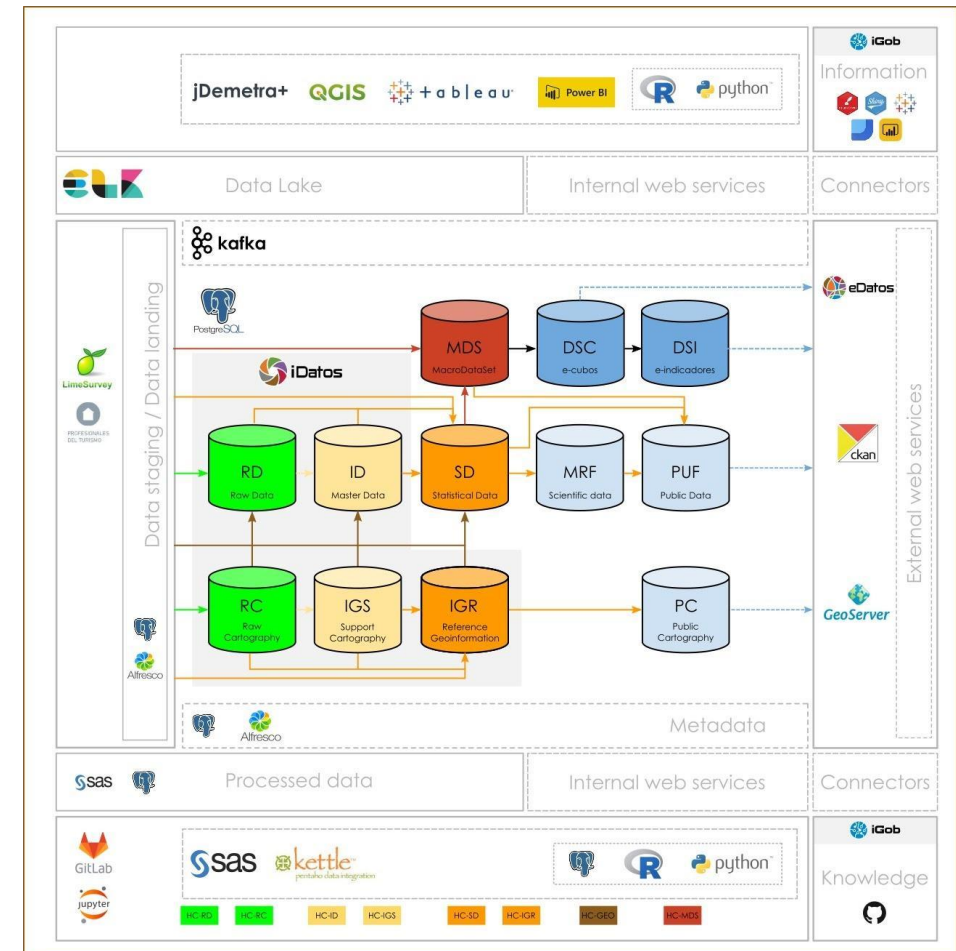
Mejorar el servicio al usuario

- Más tiempo para:
 - Analizar demandas internas y externas.
 - Proveer de nuevos datos.
 - Mejorar la calidad de los datos y el acceso a los mismos.



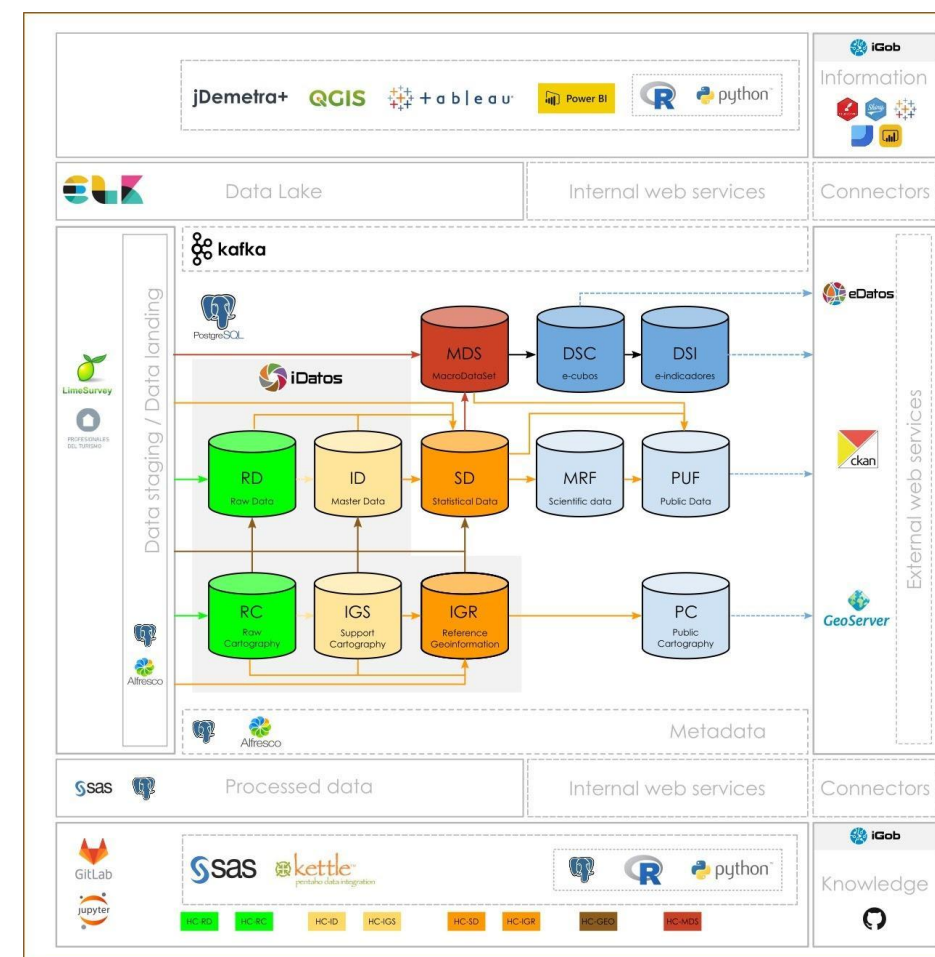
Nuestra infraestructura de datos

- Data landing.
- Bloque privado de microdatos:
 - Datos: en crudo, datos maestros y datos estadísticos.
 - Cartografía: en crudo, de soporte y de referencia.
- Bloque de datos públicos:
 - Cubos e indicadores.
 - Microdatos: Análisis científico y Públicos.
 - Cartografía pública.



Nuestra infraestructura de datos

Automatización de hilos de computación para la carga del banco de datos



Proceso

Hemos definido



Pentaho Data Integration

Buenas prácticas

BLOQUE	CAMPO	TYPE	OBLIGATORI	VERSIÓN	DESCRIPCIÓN
Identificadores	uuid	uuid	Sí		UUID de la variable
Identificadores	luid	integer	Sí		LUID de la variable
Catalogación	block_1	varchar (36)	No		Bloque de variables
Catalogación	block_2	varchar (36)	No		Bloque de variables
Informática	field_name	varchar (30)	Sí		Nombre de la variable
Informática	tech_type	varchar(100)	Sí		Tipo de datos en base de datos
Informática	variable_order	smallint	Sí		Número de orden a utilizar para que las columnas queden en la tabla
Descriptores	label	varchar(100)	Sí		Etiqueta de la variable
Descriptores	concept	varchar(100)	No		
Descriptores	description	text	No		
Descriptores	derivation_description	text	No		
Descriptores	annotations	text	No		
Descriptores	var_statistical_type	varchar(15)	Sí		

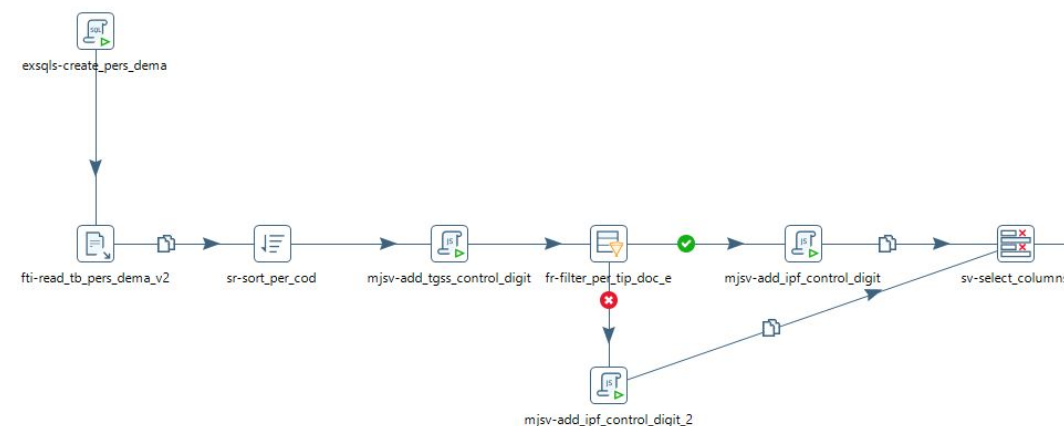
- Normalización de definición de diseños de registros:
 - Información a recoger de cada variable para una documentación completa.
 - Normalización de clasificaciones.
 - Versionado.
 - ...

- **Guía para los desarrolladores.**
 - **Nomenclatura.**
 - **Registro de log.**
 - **Versionado.**
 - **Compartir/Reutilizar.**
 - **...**

[illegible]

Resultado

uid	luid	block_1	block_2	field_name	tech_type	variable_order	label	concept
uid	integer	varchar (36)	varchar (36)	varchar (30)	varchar(100)	smallint	varchar(100)	varchar(100)
		1 CLAVE		UUID	uid		1 Identificador Único Universal	
		2 CLAVE		LUID	integer		2 Identificador Único Local	
		3 CLAVE		STD	uid		3 Identificador de esquema y tabla	
		4 CLAVE		MARCATIEMPO	date(8)		4 Sello de tiempo de creación de la observación	
		5 IDENTIFICACION	GEOGRÁFICAS	PMH_DOMIRES_CPPO	varchar(2)		5 ID - Código de provincia de residencia (raw)	ISTAC.CST_ISTAC
		6 IDENTIFICACION	GEOGRÁFICAS	PMH_DOMIRES_CMUN	varchar(3)		6 ID - Código de municipio de residencia (raw)	ISTAC.CST_ISTAC
		7 IDENTIFICACION	GEOGRÁFICAS	PMH_DOMIRES_CODMUN_RC	varchar(15)		7 ID - Código de municipio de residencia (recodificado)	ISTAC.CST_ISTAC
		8 IDENTIFICACION	IDENTIFICADORES	PMH_ID_NOMBRE	varchar(20)		8 ID - Nombre del habitante	
		9 IDENTIFICACION	IDENTIFICADORES	PMH_ID_PARTICULA_APELLIDO1	varchar(6)		9 ID - Partícula del primer apellido	
		10 IDENTIFICACION	IDENTIFICADORES	PMH_ID_APELLIDO1	varchar(25)		10 ID - Primer apellido	



Automatización de hilos de computación

Instituto Canario de Estadística, ISTAC

Diseño de la ETL

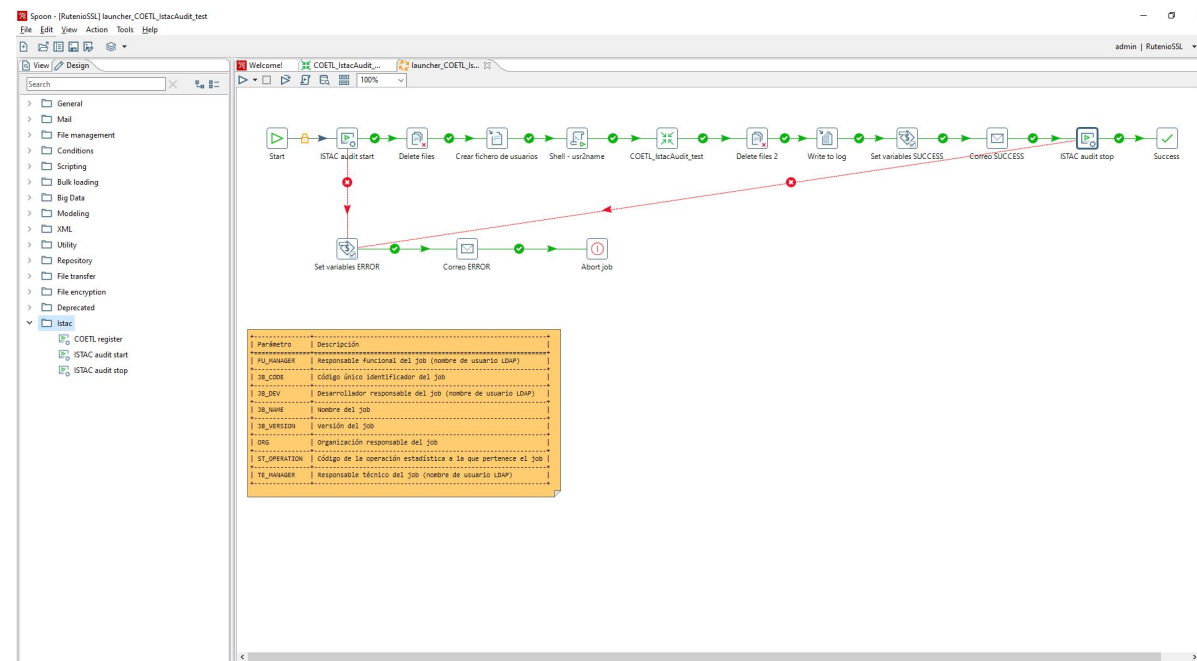
Ajustándose a las normas antes mencionadas, se hace el diseño del procedimiento de carga de los datos en nuestro banco de datos.

Se introducen los pasos necesarios para registrar el log e implementar la funcionalidad requerida a la ETL.

Todo esto lo hace el desarrollador en su entorno local.

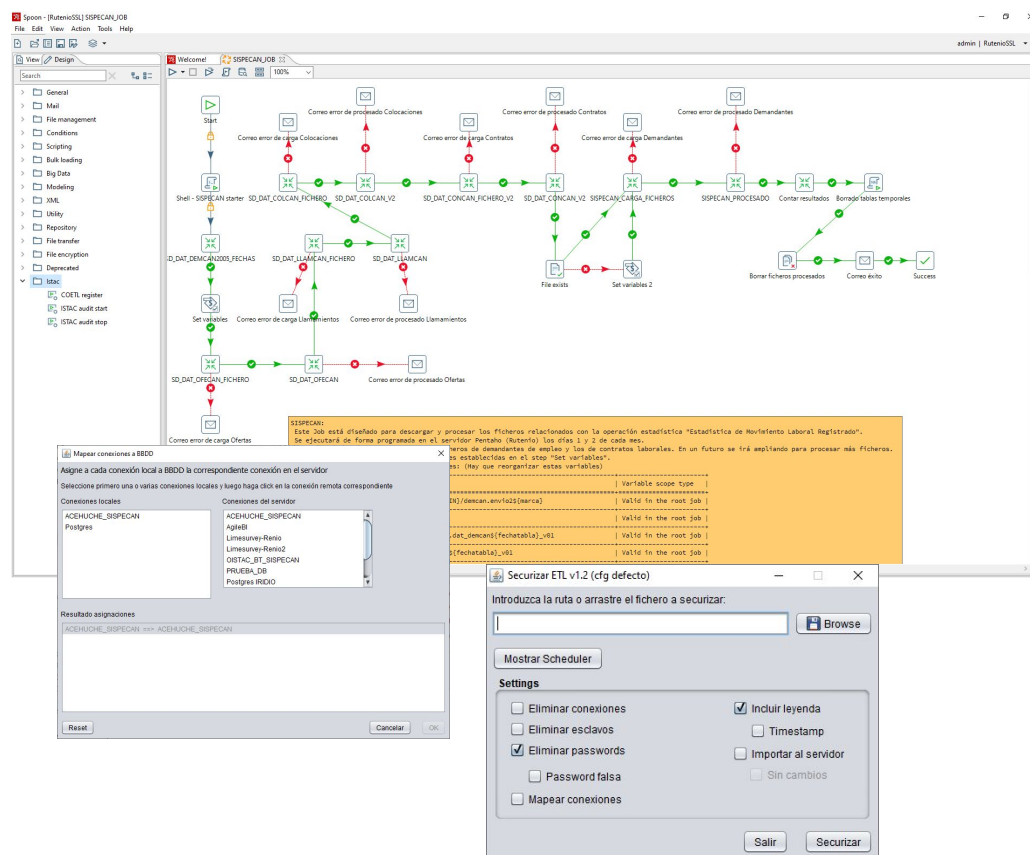
Herramientas:

- Cliente local de PDI - Spoon 8.3 (Hitachi)
- Apache Hop 1.0



Integración en Pentaho Server

- Securizar la ETL: eliminar credenciales y datos privados del desarrollador.
- Mapear conexiones: pasar de las conexiones privadas del desarrollador a las conexiones del servidor de Pentaho.
- Cambiar rutas locales por rutas accesibles desde el servidor, parametrización,...
- Añadir notificaciones a usuarios.
- Pruebas finales en el entorno del servidor.



Herramientas:

- Spoon 8.3 - Conectado al repositorio del servidor.
- Securización de ETL (Desarrollada por ISTAC)
- Mapeo de conexiones (Desarrollada por ISTAC)

Planificación y ejecución en Pentaho Server

File View Tools Help

Schedules ▾

Manage Schedules

Schedule Name	Repeats	Source File	Out
PentahoSystemVersionCheck	Every day at 12:50:42 PM	PentahoSystemVersionCheck	-
launcher_ib_sce_management_data_download	Day 1 of every month at 11:00:00 AM	/home/admin/DESCARGAS/SISPECAN_GESTION/launcher_ib_sce_m...	/h
launcher_tgss	Day 1 of every month at 8:00:00 AM	/home/admin/DESCARGAS/TGSS/launcher_tgss	/h
CONVOCATORIAS_MENSUALES_JOB	Day 2 of every month at 6:50:00 AM	/home/admin/DESCARGAS/SUBVENCIONES/CONVOCATORIAS_MEN...	/h
SUBVENCIONES_MENSUALES_JOB	Day 2 of every month at 6:30:00 AM	/home/admin/DESCARGAS/SUBVENCIONES/SUBVENCIONES_MENS...	/h
SISPECAN_JOB	Day 1 of every month at 10:00:00 AM	/home/admin/DESCARGAS/DEMCAN/SISPECAN_JOB	/h
SISPECAN_JOB	Day 2 of every month at 6:00:00 AM	/home/admin/DESCARGAS/DEMCAN/SISPECAN_JOB	/h
launcher_tgss_v3	Day 1 of every month at 8:30:00 AM	/home/admin/DESCARGAS/TGSS_V3/launcher_tgss_v3	/h

Blockout Times

No blockout times for scheduling have been created.

Create Blockout Time

Herramientas:

→ Pentaho Data Integration Server CE 8.3 (Hitachi)

Permite

- Planificar la ejecución de los diferentes procesos.
- Seguimiento de los resultados de cada ejecución.
- Cada noche las ETL del servidor son almacenadas en un repositorio GIT.

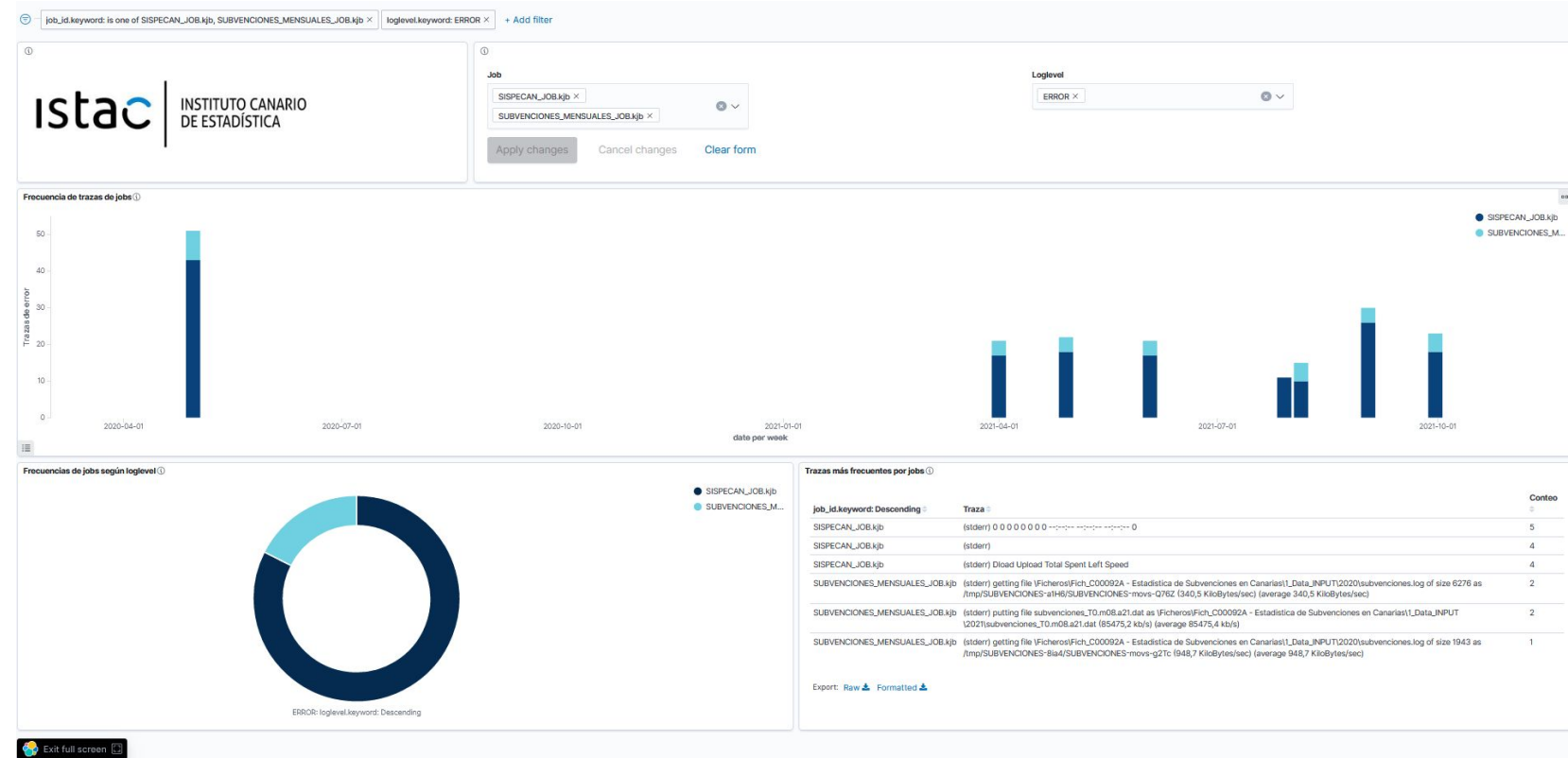
Otras opciones que ofrece el servidor:

- Control de acceso.
 - Gestión de usuarios.
 - Gestión de permisos.
- Ejecución manual de los procesos.

Log de ejecución

Los usuarios tienen acceso al cuadro de mando de Kibana en el que pueden consultar el resultado de la ejecución de los diferentes procesos.

El acceso a este cuadro de mando se puede controlar facilitando a cada usuario el nivel de información adecuado a su perfil.



Lo que está por venir

COETL (COnsola de ETLs)

- Inventario de ETLs metadocumentadas.
- Integrado con un sistema de control de versiones (GIT).
- Control de acceso según roles de usuarios.

eDatos APIs - DIFUSIÓN - PRODUCCIÓN - NORMALIZACIÓN - PLANIFICACIÓN - GESTIÓN

Consola de ETL - Gestión v13.9wps-07 Registro de ETLs + Crear nueva ETL

Registro de ETLs

Código Nombre Organismo responsable Tipo Incluir planificados

Incluir eliminados

CÓDIGO	NOMBRE	ORGANISMO RESPONSABLE	TIPO	PLANIFICACIÓN
launcher_jb_sce_management_data_download	Descarga, almacenamiento en fichero y carga en Postgres de los datos de gestión del SCE	ISTAC	340	No planificado
launcher_tgss	Descarga y carga en Postgres de los datos de Seguridad Social (Afiliados y Cuentas de Cotización)	ISTAC	340	Planificado
CONVOCATORIAS_MENSUALES_JOB	Descarga de los datos de convocatorias de subvenciones	ISTAC	340	No planificado
SUBVENCIONES_MENSUALES_JOB	Descarga de los datos de subvenciones concedidas	ISTAC	340	Planificado
SISPECAN_JOB	Descarga y carga en Postgres de los ficheros estadísticos del SEPE (Hay dos descargas porque se ejecuta los días 1 y 2, el día 2 es una actualización de los datos de un fichero)	ISTAC	340	Planificado
launcher_tgss_v3	Descarga de los datos de Seguridad Social con el nuevo formato de fichero.	ISTAC	340	No planificado

1 - 3 de 3 Tamaño de página 20 « « 1 » »

COETL (cont.)

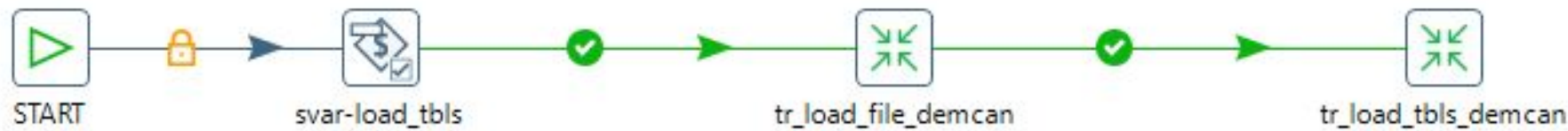
- Permite:
 - Realizar búsquedas.
 - Planificar la ejecución.
 - Lanzar manualmente una ETL.
 - Consultar información básica sobre el estado de las ejecuciones
 - Dar de baja procesos obsoletos.
- Integrada con:
 - PDI.
 - Hop.

The screenshot shows the 'Consola de ETL - Gestión v113-SNAPSHOT' interface. It includes a search bar with filters for 'Código', 'Nombre', 'Organismo responsable', 'Tipo', and 'Incluir planificados'. Below the search bar is a table of ETL jobs with columns 'CÓDIGO', 'NOMBRE', and 'ORGANISMO RESPONSABLE'. The table lists several jobs, including 'launcher_jb_sce_management_data_download', 'launcher_tgss', 'CONVOCATORIAS_MENSUALES_JOB', 'SUBVENCIONES_MENSUALES_JOB', 'SISPECAN_JOB', and 'launcher_tgss_v3'. The interface also features a '+ Crear nueva ETL' button and a 'Reiniciar filtros' button.

CÓDIGO	NOMBRE	ORGANISMO RESPONSABLE
launcher_jb_sce_management_data_download	Descarga, almacenamiento en fichero y carga en Postgres de los datos de gestión del SCE	ISTAC
launcher_tgss	Descarga y carga en Postgres de los datos de Seguridad Social (Afiliados y Cuentas de Cotización)	ISTAC
CONVOCATORIAS_MENSUALES_JOB	Descarga de los datos de convocatorias de subvenciones	ISTAC
SUBVENCIONES_MENSUALES_JOB	Descarga de los datos de subvenciones concedidas	ISTAC
SISPECAN_JOB	Descarga y carga en Postgres de los ficheros estadísticos del SEPE (Hay dos descargas porque se ejecuta los días 1 y 2, el día 2 es una actualización de los datos de un fichero)	ISTAC
launcher_tgss_v3	Descarga de los datos de Seguridad Social con el nuevo formato de fichero.	ISTAC

Casos de uso

ETL del fichero estadístico de demandantes de empleo



JOB DE CARGA

Paso de variables.

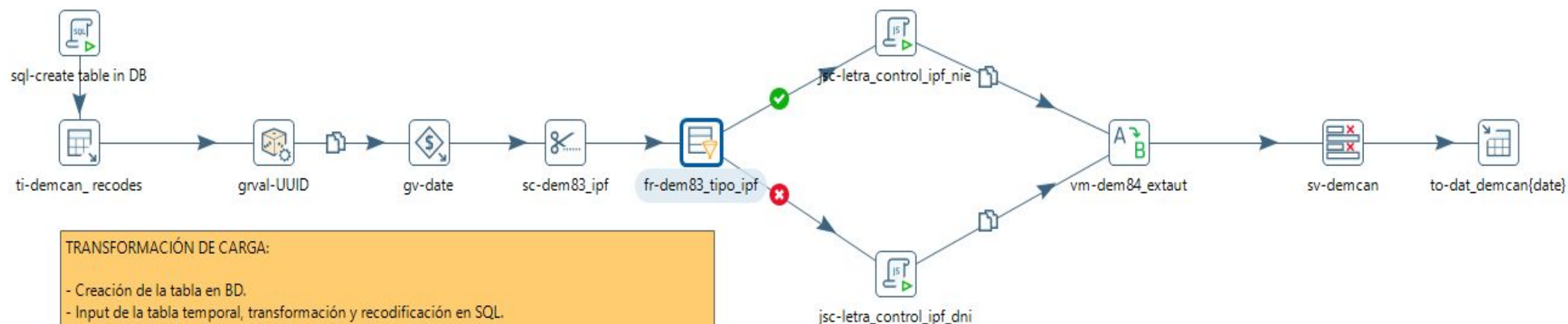
Transformación: carga del Fichero de tabla temporal en BD.

Transformación: de transformación, recodificación de variables y carga de tablas Demcan en BD.

Job de carga de demandantes de empleo

Casos de uso

ETL del fichero estadístico de demandantes de empleo



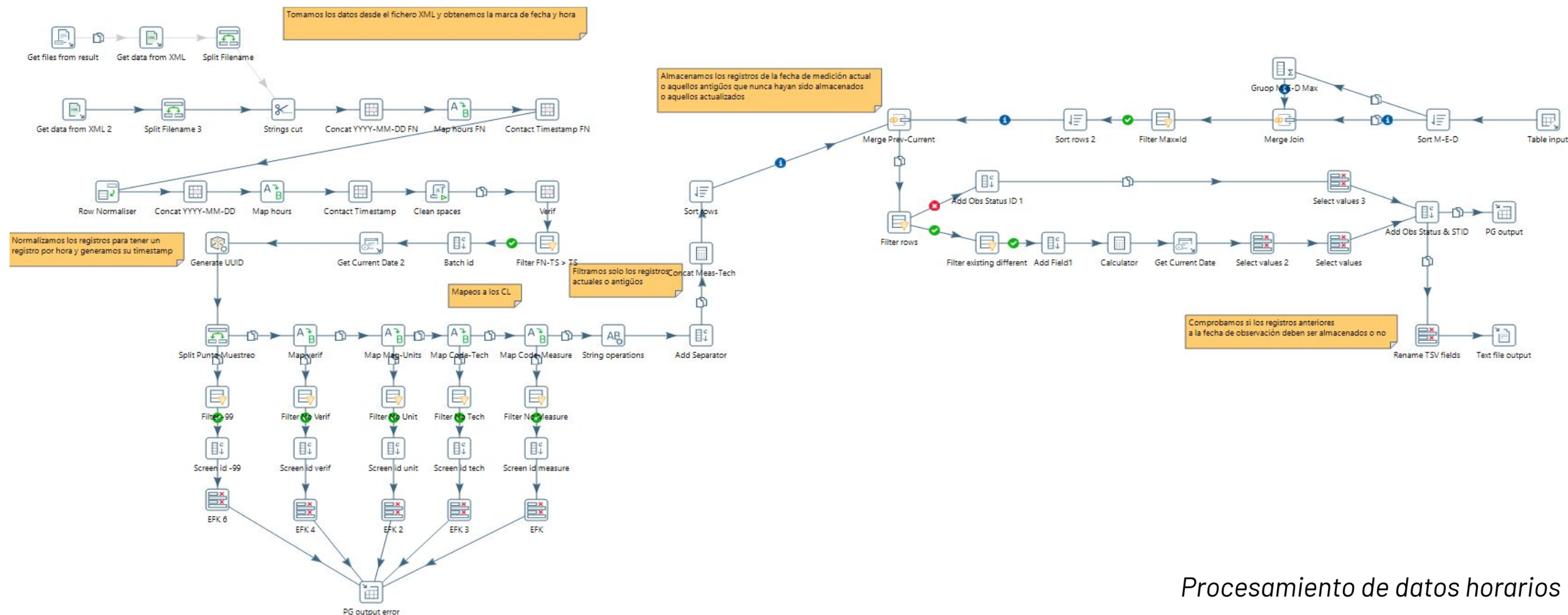
TRANSFORMACIÓN DE CARGA:

- Creación de la tabla en BD.
- Input de la tabla temporal, transformación y recodificación en SQL.
- Generación del UUID.
- Generación y filtro según tipo de IPF.
- Recálculo de la letra del IPF (DNI Y NIE) utilizando un script de algoritmo de cálculo en Javascript.
- Transformación de valores de una variable.
- Selección y orden de variables demcan para la carga en tabla.
- Carga de la tabla en BD Postgres.

Transformación de carga de demandantes de empleo

Casos de uso

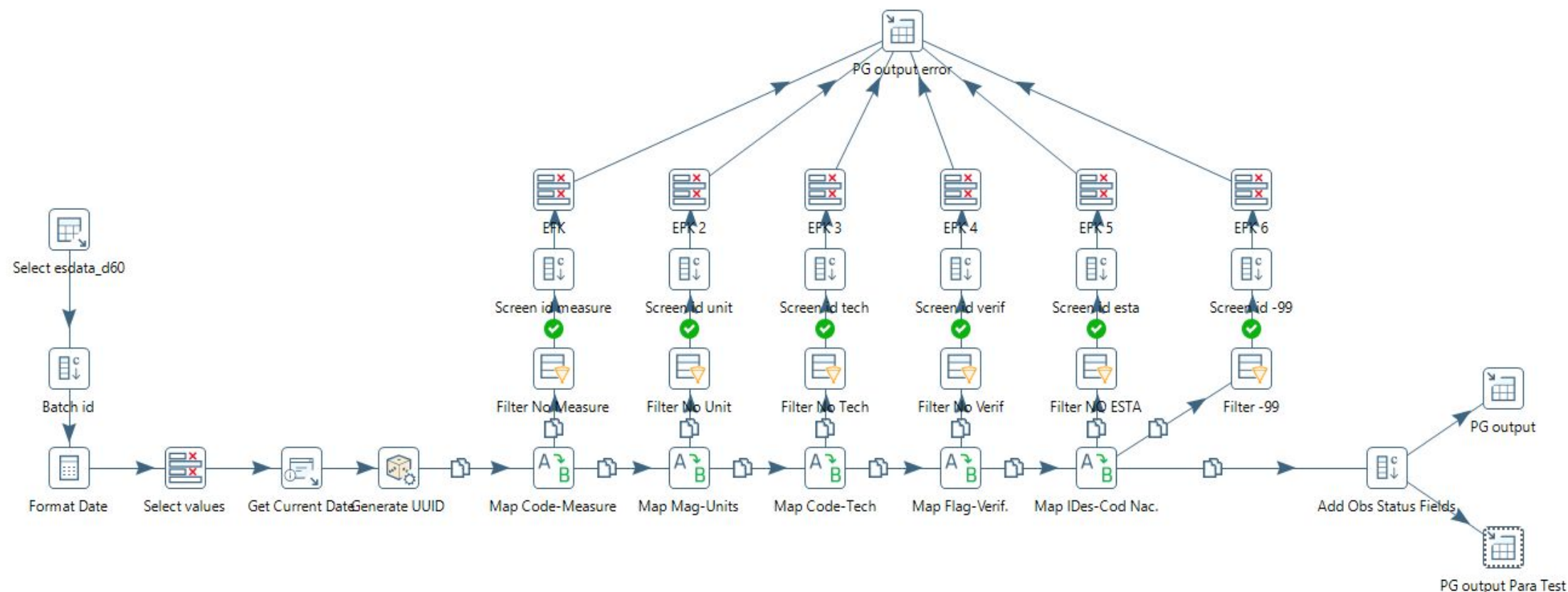
ETL de Calidad del Aire



Procesamiento de datos horarios

Casos de uso

ETL de Calidad del Aire



Procesamiento de datos históricos

Casos de uso

ETL de Calidad del Aire

UUID	SCREEN ID	ETL ID	FECHA
312deea9-aeb8-11e9-b6d8-a1305aee962	ESTA_MISS_ID	T-ETL-HIST-CA-005	2019-07-25 09:42:53.382
312e3cca-aeb8-11e9-b6d8-a1305aee962	TECH_MISS_ID	T-ETL-HIST-CA-005	2019-07-25 09:42:53.384
312fea7c-aeb8-11e9-b6d8-a1305aee962	ESTA_MISS_ID	T-ETL-HIST-CA-005	2019-07-25 09:42:53.393
312fea7d-aeb8-11e9-b6d8-a1305aee962	VERIF_MISS_ID	T-ETL-HIST-CA-005	2019-07-25 09:42:53.393
312fea80-aeb8-11e9-b6d8-a1305aee962	ESTA_MISS_ID	T-ETL-HIST-CA-005	2019-07-25 09:42:53.393

Registro de errores



Islas Canarias
Del 15 al 19 de noviembre de 2021

iGracias!



[Lhttps://www.linkedin.com/company/istac](https://www.linkedin.com/company/istac)



@istac_es