AUTOMATIZACIÓN DE HILOS DE COMPUTACIÓN

Esther Torres Medina
Instituto Canario de Estadística, ISTAC
etormed@gobiernodecanarias.org

José Carlos Gil Carreras
Ecosistemas Virtuales y Modulares S.L., EVM
josecarlos.gil@evm.net

Miguel Núñez Pereyra
Ecosistemas Virtuales y Modulares S.L., EVM
miguel.nunez@evm.net

Idafen Santana-Pérez
Universidad de Las Palmas de Gran Canaria, ULPGC
idafen.santana@ulpgc.es

Rita Díaz Adán
Arte Consultores Tecnológico S.L., ARTE
rdiaada@arte-consultores.com

PALABRAS CLAVE
ETL, Transformación, Carga, Cómputo, Automatización

RESUMEN

En la actualidad, la extracción de la información a partir de diferentes fuentes de datos está adquiriendo cada vez más relevancia. En el Instituto Canario de Estadística (ISTAC) hemos abordado un proceso de industrialización de la carga de datos para ser capaces de dar respuesta a las crecientes demandas, tanto del personal propio como del resto de organismos del gobierno, ciudadanos y empresas.

Partíamos de cargas realizadas de forma semiautomatizada, en las que varias personas tenían que coordinarse para ejecutarlas y validarlas. Esta manera de realizarlas consumía una cantidad considerable de tiempo y tenía que ser llevada a cabo por personal técnico que pudiese resolver las incidencias.

La nueva estrategia consiste en la automatización de dichas cargas, realizando simultáneamente recodificaciones y validaciones, para que, desde el momento de su llegada, los datos estén a disposición de los técnicos estadísticos en el repositorio.

De cara a implementar esta estrategia, la herramienta elegida por el Instituto ha sido PDI (Pentaho Data Integration), herramienta de Hitachi que cuenta con versiones gratuitas. Los desarrolladores pueden programar un flujo de datos, de forma modular, que transformará los datos de entrada en tablas en el repositorio. El desarrollo de los flujos y transformación de datos lo hacemos con la interfaz visual Spoon (Kettle) y la planificación y ejecución automática con la herramienta Pentaho Server.

Esta forma de abordar las cargas de datos ha supuesto un avance importante en el trabajo pues nos ha permitido abordar la carga de históricos.

Además, nos apoyamos en todo un proceso de estandarización del procedimiento para llevarlas a cabo: definición de diseños de registros con un formato estándar propio; uso de clasificaciones estandarizadas; normalización en la nomenclatura de las variables, los tipos de datos a usar, etc.

Si bien esta estrategia aporta beneficios, cuyo impacto ya se ha visto reflejado en la gestión de datos de ISTAC, también introduce nuevos retos. Adaptar esta forma de gestión y procesamiento de datos a los diferentes casos de uso, con distintas formas de recepción y almacenamiento, análisis y depuración de errores, así como la gestión de roles, entre otros casos, requiere de un diseño y planificación coherentes. En un futuro, la idea es no sólo automatizar las cargas sino poder llegar a crear hilos computacionales genéricos (validaciones, imputaciones, análisis de calidad...) que nos permitan la publicación de datos de una forma sistematizada siguiendo unos criterios de calidad sólidos y bien establecidos.

Por último, comentar que en el proceso de industrialización resulta fundamental disponer de una herramienta que permita llevar a cabo un exhaustivo control sobre el inventariado, la meta-documentación, las planificaciones y las ejecuciones de los procesamientos estadísticos. De cara a cumplir con todos estos requisitos, el ISTAC ha desarrollado un sistema denominado “Consola ETL” que actualmente se encuentra en fase de pruebas.