BeETLe:
Herramienta ETL geo-espacial libre

[I] Arévalo, Juan ; [I] Martinez Izquierdo, Cesar; [I] Simonazzi Domínguez, Walter

[I] European Topic Centre on Land Use and Spatial Information, European Environment Agency
Universidad Autónoma de Barcelona, Campus Bellaterra, Barcelona, España

RESUMEN

Las herramientas ETL (Extract, Transform, Load – extraer, transformar, cargar) permiten modelizar flujos de datos, facilitando la ejecución automática de procesos repetitivos. El intercambio de información entre dos modelos de datos heterogéneos es un claro ejemplo del tipo de tareas que pueden abordarse con software ETL. El proyecto Kettle es una herramienta ETL con licencia LGPL (Library General Public License) que utiliza técnicas de computación grid (ejecución paralela y distribuida) para poder procesar grandes cantidades de datos en un tiempo reducido. Kettle combina una potente ejecución en modo servidor con una intuitiva herramienta de escritorio para modelar los procesos y configurar los parámetros de ejecución. GeoKettle es una extensión de Kettle, que añade la posibilidad de tratar datos con componente geográfica, si bien está limitado a datos vectoriales y a ciertas operaciones espaciales muy concreta. El Centro Temático Europeo de Usos del Suelo e Información Espacial (ETC-LUSI) está impulsando un proyecto complementario, llamado BeETLe, que pretende ampliar drásticamente las capacidades de análisis y transformación espacial de GeoKettle. Para ello se ha elegido el proyecto Sextante, una librería de análisis espacial que incluye más de doscientos algoritmos ráster y vectoriales. La intención del proyecto BeETLe es integrar el conjunto de algoritmos de Sextante en GeoKettle, de forma que estén disponibles como transformaciones de GeoKettle. Las principales características de la herramienta BeETLe incluyen: automatización de procesos de análisis espacial o de transformaciones repetitivas de datos espaciales, ejecución paralela y distribuida (grid computing), capacidad para procesar grandes cantidades de datos sin limitaciones de memoria, y soporte de datos ráster y vectorial. Los usuarios actuales de Sextante descubrirán que BeETLe les propone una forma de trabajo sencilla e intuitiva, que añade a Sextante toda la potencia que ofrecen las herramientas ETL para procesar y transformar información en bases de datos.

Palabras clave: *SIG, ETL geo-espacial, procesado en paralelo, servicios IDE, BeETLe, INSPIRE, automatización*

INTRODUCCIÓN

Qué es la tecnología ETL

La tecnología ETL (Extract, Transform and Load, en sus siglas en inglés) es el proceso mediante el cual las organizaciones mueven datos desde múltiples fuentes, los reformatean y limpian, y finalmente cargan en una base de datos, siendo su objetivo en ultima instancia el de analizar la información y apoyar un determinado proceso de negocio. La implementación de esta tecnología supone la realización de los 3 pasos anteriormente mencionados y por este orden:

Extracción: que trataría de juntar datos múltiples y fuentes de datos heterogéneas. Estas fuentes pueden ser bases de datos operacionales pero también pueden ser ficheros en varios formatos, pudiendo ser internos de la organización o externos.

Transformación: modificación del dato, del formato de la fuente de datos origen al formato de almacén de datos final. Esto incluye varios aspectos: limpieza (que consiste en eliminar errores e inconsistencias), conversión a un formato estándar, integración con el resto de datos de diferentes fuentes y las modificaciones necesarias a nivel de esquema para poderlos introducir en el almacén de datos final.

Carga: alimentación del almacén de datos con el dato transformado. Esto incluye la automatización de la actualización del almacén de datos con una determinada frecuencia por ejemplo, semanalmente, o diariamente o incluso casi en tiempo real.

Por tanto, entre las tareas típicas que una herramienta ETL puede abordar tenemos por ejemplo las siguientes:

- Coger datos desde una fuente, transformarlos y volcarlos a otro soporte.
- Leer y escribir de cualquier base de datos, fichero Excel, Access, etc.
- Operaciones a nivel de Base de Datos: operar con los campos renombrando, normalizando, calculando campos en función de otros, mapeando valores, realizando búsquedas auxiliares en bases de datos, normalizando/desnormalizando los datos, etc.

Herramientas ETL libres

Aunque existen otros proyectos ETL libres en el mercado además de los mencionados a continuación, sólo haremos referencia a Kettle y Talend, debido principalmente a que han dado lugar a dos proyectos libres de ETL espaciales que han servido como base para definir los objetivos de BeETLe.

Kettle (Pentaho Data Integration) es uno de los productos de la suite de Inteligencia de Negocio de la empresa Pentaho (http://kettle.pentaho.org/) que se utiliza para la integración de datos (ETL). Liberado bajo licencia LGPL, incluye un conjunto de herramientas que se describen a continuación:

- SPOON: permite diseñar de forma gráfica la transformación ETL.
- PAN ejecuta las transformaciones diseñadas con SPOON.
- CHEF permite, mediante una interfaz gráfica, diseñar la carga de datos incluyendo un control de estado de los trabajos.
- KITCHEN permite ejecutar los trabajos batch diseñados con Chef.

Las características más reseñables de Kettle incluyen su interfaz amigable e intuitiva, la posibilidad de ejecutar procesos en paralelo y de lanzar los flujos de trabajo en modo servidor, un sistema de plugins que permite crear nuevas transformaciones con poco esfuerzo, y una comunidad importante de usuarios.

Talend es un proyecto de la empresa Talend Open Data Solutions. Posee una interfaz de usuario bastante atractiva basada en la plataforma Eclipse, aunque su manejo no es tan intuitivo como el del proyecto GeoKettle.

ETL en los procesos SIG

Los Sistemas de Información Geográfica (S.I.G.) tradicionalmente han servido para recoger, almacenar, analizar y manejar información espacial con el objetivo de mejorar la toma de decisiones. Su uso durante varias décadas, ha generado un gran volumen de datos en multitud de formatos, así como flujos de procesos resultado del análisis y tratamiento de la información.

Las herramientas ETL por su parte se utilizan en el mundo de la Inteligencia de Negocio a la hora de extraer, transformar y cargar datos en almacenes, que en última instancia sirven para tomar decisiones empresariales.

Como resultado de la fusión de la tecnología SIG y las herramientas ETL ha surgido una nueva generación de herramientas ETL, son lo que se conoce como **herramientas espaciales ETL**, las cuales proporcionan la funcionalidad de datos tradicional pero además dotándola de la capacidad de manejar datos espaciales. Por tanto estas herramientas hacen posible la conversión de datos mejorando así la interoperabilidad de datos con distintos formatos, así como entre las distintas aplicaciones SIG, y proporcionando además la posibilidad de automatizar flujos de procesado de datos. Gracias a la incorporación de esta tecnología se abre un nuevo camino para las Infraestructuras de Datos Espaciales, ya que facilitará en gran medida la incorporación de multitud de datos heterogéneos en este tipo de plataformas.

Un caso de uso muy interesante, es el del Alto Comisionado de la ONU para Refugiados (ACNUR). Esta organización durante sus operaciones humanitarias, recoge una gran cantidad de datos en muy diversos formatos, la integración de éstos se ha venido realizando de forma manual y en muchas ocasiones no se sacaba el máximo partido a la información. Desde el año 2008, el ACNUR ha mejorado notablemente su organización gracias al uso de una herramienta espacial ETL (Talend Spatial Data Integrator) pudiendo ahora compilar varios datos en distintos formatos Excel, shapes, etc., que necesitan ser actualizados regularmente, almacenándolos en una única base de datos, para posteriormente publicar la información por medio de un visor cartográfico.

Una de las principales características presentes en los ETL de especial interés para el mundo SIG, y en particular de gran interés para nuestro trabajo diario, es la capacidad de computación en paralelo que ofrecen, la cual se basa en la premisa de que "*grandes problemas se pueden dividir en problemas más pequeños*" para después ser resueltos de forma concurrente salvando de esta forma la limitación en hardware que se hacen evidentes a la hora de trabajar con grandes volúmenes de datos. Este paradigma, sin embargo, añade un mayor nivel de complejidad en el desarrollo de un ETL geo-espacial, ya que existen ciertas limitaciones impuestas por los datos espaciales y los procesos SIG que se aplican a ellos que hacen complicada su ejecución en paralelo y que serán explicados mas en detalle en el siguiente punto y para el caso concreto del proyecto BeETLe.

Sin embargo hemos de decir que esta característica, unida a la obtención de una única herramienta SIG en la que definir, describir, ejecutar y guardar un flujo de trabajo SIG, han sido los principales motores que nos han llevado a llevar a cabo este proyecto.

ETL geo-espaciales libres que han surgido de Kettle y Talend:

Existen dos proyectos de referencia en el mundo de los ETL geo-espaciales que han surgido como la integración de un proyecto ETL libre y librerías geo-espaciales libres. Estos dos proyectos se conocen con los nombres Talend SDI y GeoKettle.

Ambos persiguen el mismo objetivo: consolidarse como herramientas ETL geo-espaciales libres, aprovechando las características ofrecidas por los ETL genéricos, extendiéndolas para ser usadas contra datos SIG. Estos dos proyectos han servido de punto de partida para el proyecto BeETLe, en lo que se refiere al análisis de la funcionalidad que ofrecen confrontándola con la funcionalidad necesaria para cubrir nuestras necesidades particulares. A continuación se hace una breve mención a las características de ambos:

- **Talend Spatial Data Integrator,** es una herramienta ETL con capacidad geoespacial. Basado en Talend Open Studio (TOS), incluye componentes geo-espaciales específicos, todos ellos desarrollados por la empresa CamptoCamp. Permite la lectura y escritura de distintos formatos SIG, manipular entidad y crear entidades, publicar metadatos, etc. Liberado bajo licencia GPL.
-
- **Geokettle** es la versión de Pentaho Data Integration (Kettle) dotada con capacidad para tratar datos espaciales. Esta herramienta ha sido desarrollada por el grupo de investigación de GeoSOA del Departamento de Geomática de la Universidad de Laval en Canadá. Liberada bajo licencia GNU Lesser General Public License (LGPL).

En la siguiente tabla resumen, se enumeran las principales características de estos proyectos, comparándolas con las ofrecidas por la herramienta propietaria FME de Safe Software, que es la herramienta ETL espacial de referencia en el mercado SIG:

	GEOKETTLE	TALEND / SDI	FME (SAFE SOFTWARE)
Tipo de Licencia	LGPL	GPL V.2	Comercial
Número de formatos SIG soportados	4	8	> 200
Lenguaje de Programación y librerías	Java, JTS GeoTools	Java, JTS, GeoTools	?
Soporte ráster	NO	SI	SI
Soporte vector	SI	SI	SI
Operaciones de análisis vectorial	> 25	29	224
Operaciones de análisis ráster	No	No	46
Ejecución paralela y distribuida	Si	Sí	?
Visor Cartográfico integrado	No	Si	Sí

Sextante como librería SIG para construir un ETL espacial:

Aunque no se trate de un herramienta ETL, Sextante ofrece la posibilidad de dotar a estas herramientas de la capacidad necesaria para convertirse en ETL espaciales. De hecho, ha sido integrada con éxito en proyectos muy diversos como gvSIG, GeoTools o el framework WPS 52ºN. Por la importancia que cobra Sextante en el proyecto BeETLe, describiremos brevemente las características principales que aporta:

- Una gran cantidad de algoritmos de análisis y gestión de datos (>200).
- Acesso a datos ráster, vectoriales y tabulares, utilizando librerías auxiliares de acceso a datos (como GeoTools o gvSIG).
- Facilidad en la creación de nuevos procesos SIG, debido a una API muy simple que permite crear algoritmos sin preocuparse del acceso a datos ni de la interfaz gráfica.
- Facilidad de integración con otras herramientas, debido a su cuidado diseño modular y extensible.

Proyecto BeETLe. Motivación y principales características del proyecto

La motivación principal para llevar a cabo el proyecto BeETLe, ha sido la necesidad de solventar los problemas a los que nos venimos enfrentando desde hace ya varios años y que pasamos a listar a continuación y de forma breve:

- Manejo de grandes volúmenes de datos difícilmente gestionables por las herramientas SIG convencionales.
- Imposibilidad de definir flujos de trabajo en un entorno único que se ejecutaran de forma automática o semi-automática, lo que nos fuerza a la utilización de distintas herramientas SIG y no SIG en un mismo flujo de trabajo, con los problemas que este hecho introduce en cuanto cambios de formato entre paso y paso, disponibilidad de distintos entornos de ejecución, etc.
- Imposibilidad de ejecutar procesos en paralelo que nos permitieran simplificar los problemas debidos al manejo de grandes volúmenes de información SIG.

De forma conjunta estos problemas se traducen en un gasto innecesario en horas de trabajo en conversión de datos y preparación de los distintos entornos de ejecución que por otro lado serían fácilmente solventados por medio de un entorno común que ofreciera la capacidad de distribuir los procesos SIG.

Principales características de los ETL que aprovecha el proyecto BeETLe:

Además de las características que proporcionan los ETL enumeradas al comienzo de este articulo, las principales funcionalidades que proveen las herramientas ETL son la posibilidad de definir, documentar, ejecutar, guardar y recuperar flujos de trabajo SIG bajo un único entorno.

Por otra parte, una de las características más interesantes de las herramientas ETL actuales es la capacidad de ejecutar estos flujos de forma paralela, distribuyendo un mismo proceso en varios entornos hardware que finalmente se consolidan en un resultado común. Uno de los objetivos del proyecto BeETLe es aprovechar la infraestructura de ejecución paralela (*middleware*) que provee Kettle, aplicándola a procesos de análisis y transformación de geodatos, solventando de esta forma uno de los principales problemas al trabajar con datos SIG de gran tamaño.

Existen otras características intrínsecas al proyecto que se derivan de su naturaleza libre y que desde un punto de vista práctico se podrían resumir en la

posibilidad de definir procesos SIG bajo demanda y adaptados a necesidades particulares. Lo que por un lado beneficia a proyecto BeETLe y por otro lado al proyecto Sextante, como proveedor de procesos SIG al ETL espacial.

Tipos de procesamiento en paralelo

En el caso de una herramienta ETL, la paralelización se puede implementar de formas diversas:

- Paralelismo de datos: consiste en dividir los datos en particiones más pequeñas, y procesar cada partición de forma independiente y simultánea. Por ejemplo, varios hilos de una misma tarea podrían cargar diferentes partes de un mismo fichero.
- Paralelismo de segmentación (pipeline): Permite que varias sub-tareas de un flujo de trabajo se ejecuten de forma simultánea sobre diferentes partes de un mismo flujo de datos. Por ejemplo, mientras la sub-tarea B está realizando cálculos con los valores de un registro (R1) de una tabla, la sub-tarea A puede estar leyendo el siguiente registro (R2) en la base de datos, para que pueda ser procesado por B en cuanto concluya su trabajo con el registro R1.
- Paralelismo de componente: Consiste en la ejecución simultánea de diferentes partes del flujo de trabajo, que deben ser independientes entre sí a nivel de datos. Por ejemplo, si la generación de dos tablas (T1 y T2) se realiza partiendo exclusivamente de los datos contenidos en una tercera tabla (T0), sería posible generar T1 y T2 de forma independiente y simultánea.

Estas técnicas pueden aplicarse simultáneamente, y de hecho se aplican en el caso de la herramienta Kettle. El paralelismo de datos y de segmentación se puede dar entre Transformation Steps, mientras que los Jobs ofrecen Paralelismo de componente.

Procesado en paralelo para datos SIG. Problemática

En el caso de los algoritmos SIG, el paralelismo de componente es fácilmente aplicable en cualquier flujo de trabajo, siempre que existan sub-procesos o flujos independientes entre sí.

No obstante, sería mucho más interesante poder aplicar también paralelismo de datos y de segmentación, ya que estos son de utilidad en todos los tipos de flujo de trabajo. Para poder implementar ambos tipos de concurrencia en Sextante, es necesario tener en cuenta la naturaleza diversa de los algoritmos que incluye.

Algunos algoritmos pueden aplicarse de forma independiente a subconjuntos de los datos de entrada, de forma que el resultado final estaría compuesto por la combinación de los resultados parciales de cada subconjunto. Un ejemplo podría ser el algoritmo de *buffer* (área de influencia), que puede aplicarse de forma independiente a cada geometría de la capa vectorial. Otro ejemplo sería la suma de dos capas ráster, que puede calcularse de forma independiente para subconjuntos (o *tiles*) de las capas de entrada. La unión directa de todos los *tiles* resultado generaría una capa equivalente a la suma completa de las dos capas de entrada, y por tanto estos algoritmos podrían catalogarse como **algoritmos paralelizables directos**, a los que podría aplicarse paralelismo de datos y de segmentación con relativamente poco esfuerzo.

Por otra parte, existen algoritmos que no son directamente paralelizables, ya que la simple combinación de los resultados parciales no es equivalente al resultado de aplicar el algoritmo a la capa entera. Por ejemplo, no es posible aplicar el algoritmo de

reclasificación ráster “Dividir en N clases de igual area” a subconjuntos (*tiles*) de la capa ráster de entrada, ya que la capa resultante de unir todos los *tiles* no contendrá una reclasificación correcta de la capa de entrada. A estos algoritmos los podemos catalogar como **algoritmos secuenciales**.

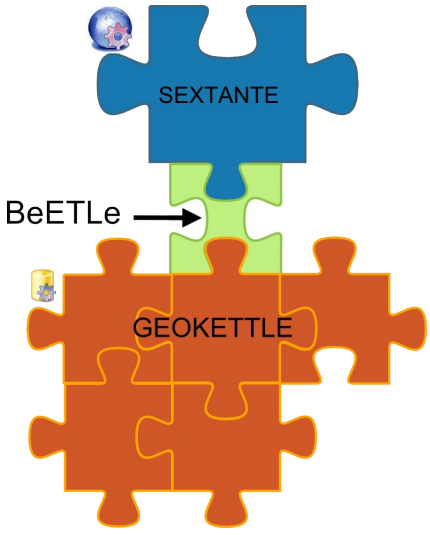
Existen otros algoritmos en los cuales no es posible hacer una simple unión de los resultados parciales para generar la capa resultado, pero que sí son parcialmente paralelizables aplicando un pre-proceso a los datos de entrada o un post-proceso a los datos de salida. Por ejemplo, el algoritmo “tabular área” se puede aplicar por zonas (o *tiles*), pero la simple unión de las tablas resultantes (en el sentido de unión que define SQL) no sería un resultado válido para las capas de entrada completas. En vez de una unión, deberíamos sumar las áreas computadas en cada tabla para cada una de clases o zonas definidas en las mismas. De esta forma, podríamos catalogar estos algoritmos como **algoritmos paralelizables indirectos**. Cuanto más complejo sea el post-proceso o pre-proceso necesario necesario para paralelizar el algoritmo, menor será la mejora de rendimiento obtenida con la paralelización, y viceversa.

Además de la problemática descrita, inherente a la generación de algoritmos SIG paralelos, el proyecto BeETLe necesita enfrentarse a la problemática derivada de las plataformas que utiliza en su implementación (GeoKettle y Sextante).

Por una parte, la API (interfaz de programación de aplicaciones, en sus siglas en inglés) actual de Sextante no contempla la posibilidad de definir algoritmos paralelos, por lo que es necesario adaptarla para que cubra estas necesidades.

Por otra parte, la arquitectura de Sextante está diseñada para que sean los algoritmos los que van pidiendo los datos según los necesitan, mientras que en Kettle las transformaciones reciben los datos en el orden que decide Kettle. Por lo tanto, es necesario adaptar la API de Sextante para que pueda integrarse en Kettle de forma natural.

Características del proyecto BeETLe.



El proyecto BeETLe constituye una potente herramienta ETL espacial, que ofrece multitud de operaciones de análisis ráster y vectorial, combinadas con otras muchas operaciones de carácter no-espacial.

Los objetivos clave de BeETLe son:

- Permitir modelizar flujos de trabajo SIG usando una herramienta ETL
- Procesar correctamente grandes cargas de datos
- Proporcionar soporte vectorial y ráster
- Integrar todos los algoritmos de Sextante como transformaciones de Kettle.
- Permitir la ejecución en paralelo de las operaciones de análisis

A continuación se muestra de forma esquemática las características de BeETLe frente a GeoKettle:

	GEOKETTLE	BEETLE
Tipo de Licencia	LGPL	LGPL
Número de formatos SIG soportados	4	6
Lenguaje de Programación y librerías	Java, JTS GeoTools	Java, JTS, GeoTools, Sextante
Soporte ráster	NO	SI
Soporte vector	SI	SI
Operaciones de análisis vectorial	> 25	> 100
Operaciones de análisis ráster	No	> 100
Ejecución paralela y distribuida	Si	Sí
Visor Cartográfico integrado	No	No

Hoja de ruta

- v0.1 – comienzos de 2010:
- Integración básica: Todos los algoritmos de Sextante disponible como Job Entries de GeoKettle
- Soporte de datos ráster (TIF y ASCII grid)
- Bindings con GeoTools mejorados: más velocidad, mejor manejo de datos de gran tamaño
- Nuevos algoritmos: TabulateArea, FishNetGraticuleBuilder.
- v0.2 – en desarrollo durante 2010:
- Integración avanzada: Algunos algoritmos de Sextante disponibles como Transformation Steps de GeoKettle
- Ejecución en paralelo de algunos algoritmos de Sextante

Futuro del proyecto BeETLe. La Infraestructuras de Datos Espaciales como marco de referencia para el futuro del proyecto.

Como se apuntaba al principio del presente artículo, las motivaciones por las que nos decidimos a abordar el proyecto BeETLe, respondían a necesidades internas de nuestro equipo de trabajo, en lo referente al procesado de datos de gran tamaño, a la problemática debida al uso de varias aplicaciones SIG en un mismo flujo de trabajo y a las limitaciones que nos encontrábamos con las herramientas propietarias de las que hacemos uso. Sin embargo, tomando como base los hitos iniciales que nos marcamos en el proyecto BeETLe, hemos visto que sus capacidades se pueden extender adaptándose cada vez más a las particularidades de nuestro entorno de

trabajo, siendo la principal la *dispersión geográfica del equipo de trabajo del Centro Temático*, característica que añade complejidad al trabajo diario.

Nuestro centro temático se estructura como un consorcio de socios en donde existe un "core team" localizado en Barcelona, y diversas organizaciones que actúan como"partners" que se encuentran distribuidos por diversos países europeos. Poco a poco tanto el core team como los partners están haciendo publica la información que generan través de servicios OGC de visualización y descarga. Este nuevo escenario ha hecho que propongamos marcar como un hito a medio plazo la extensión de las capacidades del proyecto BeETLe de cara a consolidarlo como una herramienta ETL espacial para las infraestructuras de datos espaciales, ofreciendo así una herramienta que permita definir como dato de entrada en un flujo de trabajo un servicio IDE de descarga.

El escenario fuera de nuestro entorno de trabajo es similar. En línea con la implementación de la directiva INSPIRE, los organismos nacionales y europeos que generen y/o custodien y mantengan información geográfica la han de hacer accesible al público en general a través de servicios estandarizados de visualización, y de interés para BeETLe destacamos los servicios de:

a) Descarga, utilizando para ello servicios como los definidos por la Open Geospatial Consortium, Web Feature Service (WFS) o Web Coverage Service (WCS), para datos vectoriales y teselas respectivamente, y

b) De procesado en línea, que exponen el resultado como servicios de descarga WFS o WCS. Por medio de estos servicios se puede estandarizar los procesos SIG a aplicar sobre un determinado dato, y cuyo resultado se publicaría como un servicio de descarga, garantizando así las metodologías de generación de sub-productos de los datos de referencia.

Atendiendo a este escenario, como se ha comentado anteriormente, en el proyecto BeETLe nos planteamos la extensión de su funcionalidad hacia proporcionar la capacidad al usuario de definir como dato de entrada (*Extract*) en un flujo de trabajo SIG:

- un servicio de descarga WFS o WCS.
- servicios de descarga (WFS o WCS) generados como resultado de un servicio asíncrono de procesado en linea (WPS).

La diferencia entre ambas opciones, es que mientras en el primero accederíamos a un dato de referencia, en el segundo accederíamos a un sub-producto de un dato de referencia, el cual se ha generado gracias a la estandarización del proceso por parte del proveedor de la información.

Un tercer caso que se esta actualmente valorando sería la posibilidad de extender BeETLe para que permitiera al usuario definir un determinado proceso SIG genérico a partir de los proporcionado por proveedores externos. Dentro del mapa de ruta del proyecto BeETLe se contempla el permitir al usuario la posibilidad de elegir entre los mas de 200 procesos SIG proporcionado por el proyecto Sextante, característica base de este proyecto. Cabe la posibilidad de que existan determinados procesos SIG genérico no disponibles en Sextante, que por el contrario podrían estar siendo ofrecidos por un proveedor externo a través de un servicio WPS. Para este caso concreto, cabrían dos posibilidades:

a) Implementar en Sextante el proceso SIG, haciéndolo así extensible a BeETLe. ó
b) En el caso de ser proporcionado por un proveedor externo como servicio WPS, permitir al usuario definirlo como proceso SIG en el flujo de trabajo de BeETLe.

Es evidente que la opción a) sería la más beneficiosa para ambos proyectos, pero la opción b) sería interesante valorarla dado que aumentaría la flexibilidad de BeETLe como herramienta ETL espacial para las IDEs, y además, desde un punto de vista más práctico, podría ser la solución más rápida ante una demanda urgente. En cualquier caso, la decisión que se tome a este respecto, se tomará en base al análisis de las necesidades internas del consorcio, aunque estamos abiertos a cualquier colaboración ofreciendo nuestro soporte para hacer de BeETLe un entorno ETL espacial libre para las IDEs que cubra y dé respuesta a cada vez más usuarios.

Conclusiones

Los ETL se han postulado como una tecnología con una gran potencialidad de aplicación en el campo de la geomática en lo que se refiere a su capacidad de acceso a múltiples orígenes de datos, la aplicación de transformaciones a los mismos de forma distribuida y a la carga de la información en repositorios comunes o corporativos adecuándolos a un modelo de datos en particular. Esta tecnología proporciona a los SIG la posibilidad de creación de flujos de trabajo estandarizados, lo que permite el mantenimiento de unos estándares de calidad en la información geográfica. Además, las características que presentan en lo referente a la distribución de procesos en varios entornos de ejecución las hacen idóneas para el procesado de datos de gran volumen, aunque a la vez, introducen una problemática difícil de resolver en cuanto a cómo solventar la distribución en varios entornos de ejecución de ciertos procesos espaciales que no se pueden abordar de forma lineal.

En cuanto a su potencialidad en el uso como entorno en donde combinar orígenes de datos internos con servicios IDE, tanto de descarga como de procesado, y combinados con un servicio de descubrimiento, los hace un entorno de trabajo perfecto en donde definir de forma segura flujos de trabajo en donde usamos la información correcta y aplicamos los parámetros correctos para una determinada transformación.

El proyecto BeETLe persigue estos objetivos, e intentará en la medida en que los recursos lo permitan, llevarlos a buen puerto a medio plazo, agradeciendo y animando siempre a las contribuciones de la comunidad SIG.

Referencias

- Blog oficial del proyecto BeETLe: http://beetle-project.blogspot.com/
- Página de BeETLe en OSOR (código fuente, descargas, etc): http://forge.osor.eu/projects/etclusi
- Página web del centro ETC-LUSI: http://etc-lusi.eionet.europa.eu/
- MALINOWSKI, E.; ZIMÁNYI E. (2009), “Advance Data Warehouse Design”. Springer, pp. 55-57
- Definición de Middleware del consorcio OW2: http://middleware.objectweb.org/
- Definición de Middleware en Wikipedia (en inglés): http://en.wikipedia.org/wiki/Middleware
- Introduction to Parallel Computing. *Blaise Barney, 2009:* https://computing.llnl.gov/tutorials/parallel_comp/
- Definición de Computación Paralela en Wikipedia (en inglés): http://en.wikipedia.org/wiki/Parallel_computing
- Paralelismo en herramientas ETL en Wikipedia (inglés): http://en.wikipedia.org/wiki/Extract,_transform,_load#Parallel_processing

Plaça Ferrater Mora 1, 17071 Girona
Tel. 972 41 80 39, Fax. 972 41 82 30
infojornadas@sigte.udg.es http://www.sigte.udg.es/jornadassiglibre/