

Introduction à GeoKettle un outil ETL spatial open source

par Etienne Dubé et
Thierry Badard
{etienne.dube,thierry.badard}@scg.ulaval.ca

Groupe de recherche GeoSOA

<http://geosoa.scg.ulaval.ca>

Université Laval
Québec (Québec), Canada

Tutoriel présenté à OGRS 2009, Nantes
Juillet 2009



Plan

- Historique et introduction
- Fonctionnalités de base de Kettle
- Fonctionnalités spatiales de GeoKettle
- Exemples avec GeoKettle
- Conclusion

Historique

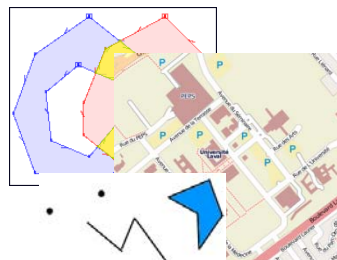
- 2006-2007 – équipe GeoSOA (Université Laval) – projets de recherche sur la géomatique décisionnelle en mobilité :
 - Absence d'outils BI (ETL et OLAP) supportant nativement les données spatiales
 - Inventaire de l'existant : outils open source de Pentaho (ETL : Kettle, OLAP : Mondrian)
 - But : étendre ces logiciels pour qu'ils prennent en charge les données spatiales

Historique (suite)

- Ainsi est né GeoKettle : une version «spécialisée» de Kettle (Pentaho Data Integration)
- Mai 2008 – première version diffusée en *open source* : 2.5.2-20080531
- Novembre 2008 – 3.1.0-20081103
- Juin 2009 – 3.2.0-20090609



+



extensions
spatiales

=



GeoKettle

Qu'est-ce qu'un ETL?

- Une catégorie de logiciel utilisé pour peupler les bases de données ou entrepôts de données, à partir d'une ou plusieurs sources de données.
- ETL:
 - **Extract** – extraction depuis les sources
 - **Transform** – transformation des données, afin de corriger les erreurs, les restructurer (changements aux schémas), les rendre conformes à des standards définis, etc.
 - **Load** – chargement des données vers les BD cibles
- Un ETL peut gérer autant l'**insertion** de nouvelles données que la **mise à jour** de données existantes.

Pourquoi utiliser un ETL?

- **Automatisation de traitements de données** complexes et répétitifs, sans avoir à faire de code sur mesure
- **Conversion** entre formats de fichiers
- **Migration des données** d'un SGBD à un autre
- **Diffusion de données** vers plusieurs SGBD
- Alimentation d'un **entrepôt de données décisionnelles**
- etc.

Introduction aux fonctionnalités de base de Kettle

Pentaho Data Integration (Kettle)

- Logiciel ETL libre (LGPL) développé en Java.
- Conçu à l'origine par Matt Casters (www.ibridge.be).
- Disponible en LGPL depuis décembre 2005.
- Acquis par Pentaho Corp. (entreprise en BI *open source*) en avril 2006.
- Fonctionne sous Windows, Linux, Mac OS X ainsi que toute autre plateforme supportant Java et SWT.
- Version actuelle: 3.2.0
- Sur le web:
<http://kettle.pentaho.org>



Pentaho Data Integration (Kettle)

- **Orienté sur les métadonnées.**
- **Exécution directe des transformations** (sans génération de code).
- **Support de BD:** MySQL, PostgreSQL, Oracle, DB2, MS SQL Server, etc. (37 au total)
- **Lecture/écriture** de divers **formats de fichiers:** texte, Excel, Access, DBF, XML, etc.
- **Étapes de transformations** variées: jointures, calculs, filtrage, dénormalisation/normalisation, validation, *scripting* (JavaScript), etc.
- Support natif pour les **dimensions changeantes, type 1 et 2** (*slowly changing dimensions* – R. Kimball)

Outils de base de Kettle

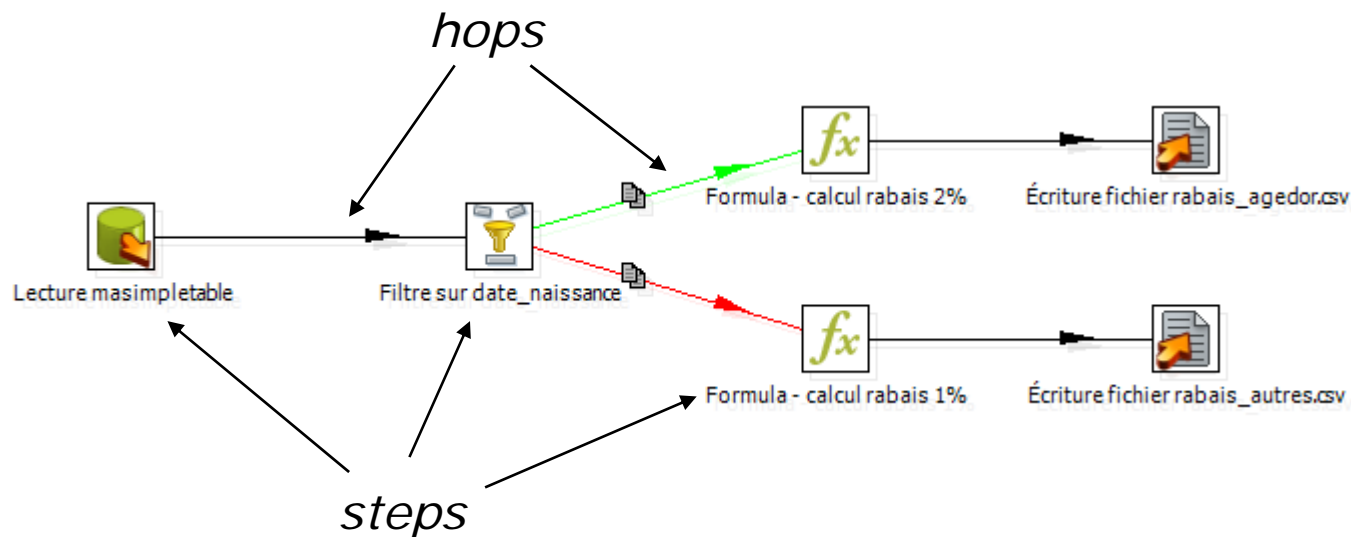
- ***Spoon*** : environnement graphique pour l'édition des *transformations* et des *jobs*
- ***Pan*** : lancement en ligne de commande des transformations
- ***Kitchen*** : lancement en ligne de commande des jobs
- ***Carte*** : serveur web pour exécution à distance des transformations et jobs

Installation

- Prérequis : Java JRE version 5 ou plus récente
- Décompresser le contenu de l'archive (distribution binaire) dans un répertoire
- Lancement de Spoon : *spoon.sh* (Linux / UNIX) ou *spoon.bat*

Transformations (1/3)

- Les **processus ETL de base** sont nommés *transformations*.
- **Étapes** de transformation: *steps* (étapes)
- **Liens** entre les étapes: *hops* (liens)
- **Exécution parallèle** (*threads*) des steps.

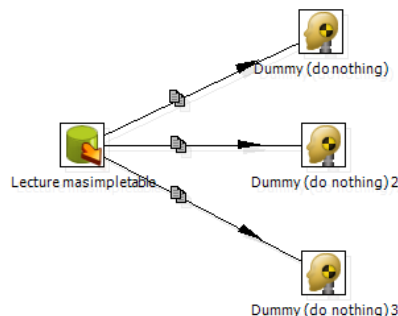


Transformations (2/3)

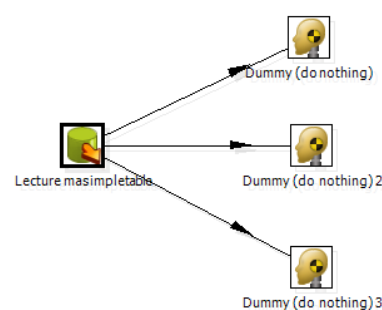
- Les *steps* comportent des **paramètres configurables** (double-clic pour ouvrir la boîte de dialogue) :
 - connexion à la BD
 - nom de fichier à ouvrir
 - critères de filtrage
 - code source d'un script (JavaScript)
 - etc.
- **Catégories de steps** :
 - entrées
 - sorties
 - transformations
 - contrôle de flux
 - scripting
 - etc.

Transformations (3/3)

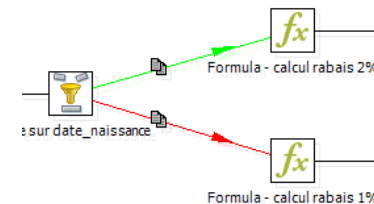
- **Les hops lient les steps les uns aux autres**, afin de définir les flux de données.
- **Pour créer un hop** : glisser-déplacer d'un step à l'autre avec le bouton du milieu (ou Maj.+ bouton gauche)
- Dans un hop:
 - les données circulent de la sortie d'un step à l'entrée du step suivant, rangée par rangée
 - la **définition des champs** (nombre, noms, types) est **toujours identique** d'une rangée à l'autre
- Types de hops:



copie

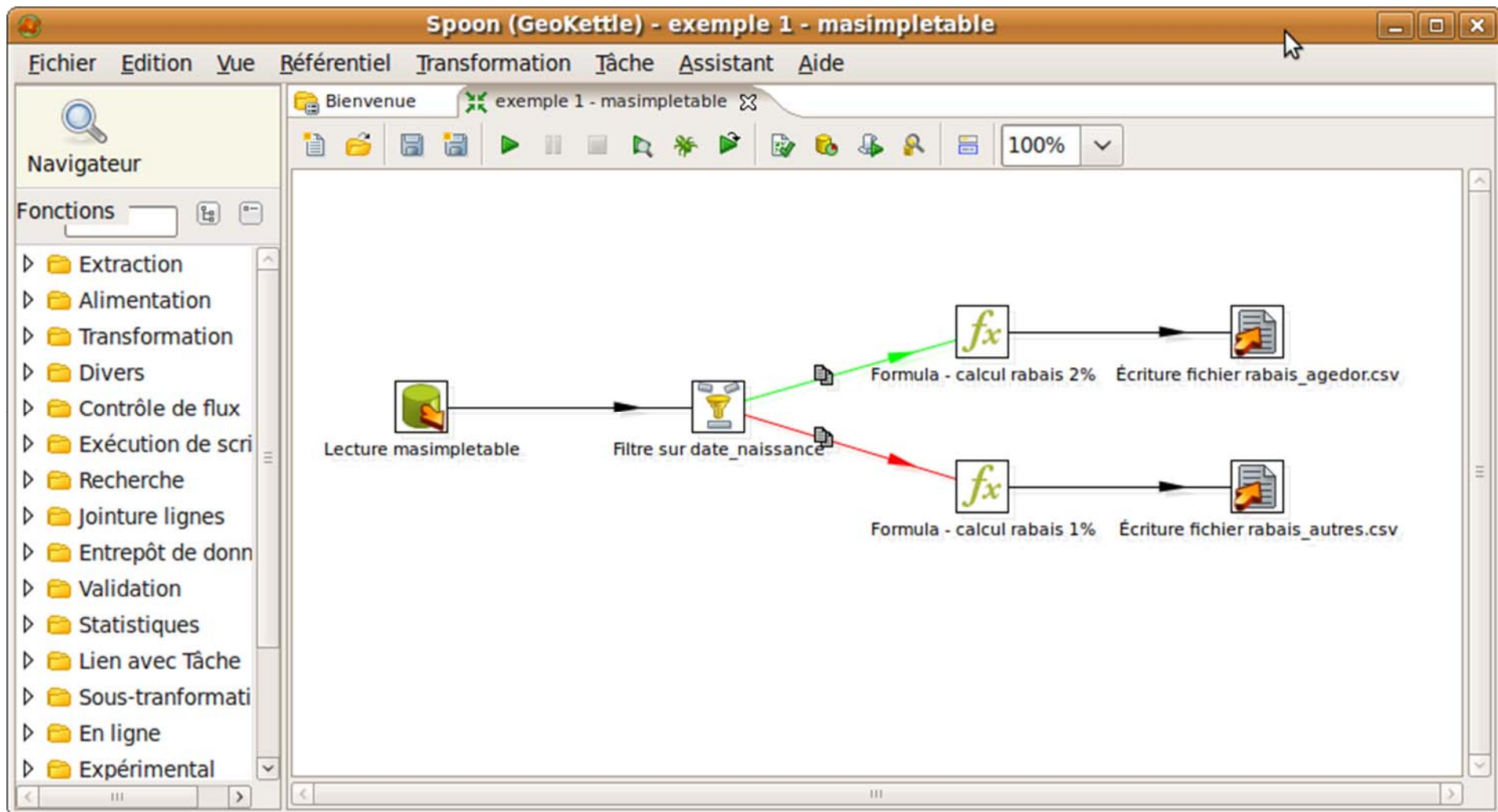


distribution



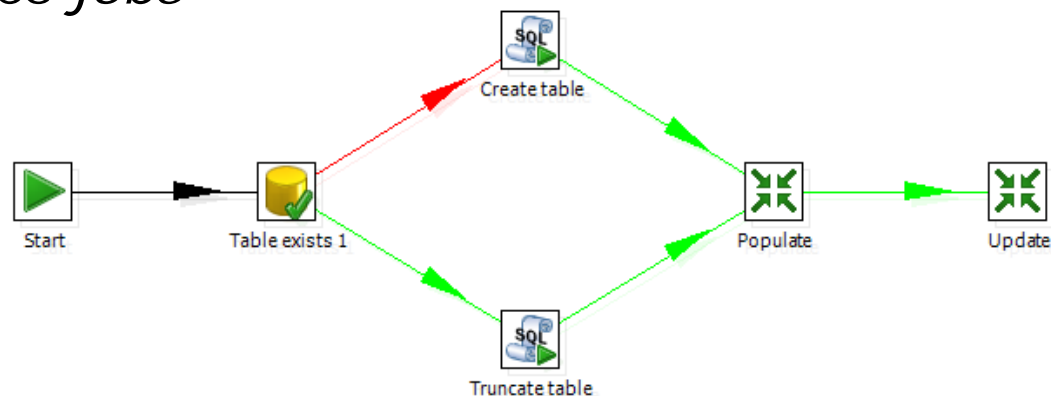
sortie
conditionnelle

Exemple 1 – Transformation simple



Jobs

- Un **job** (tâche) définit une **série de tâches à exécuter séquentiellement**. Ces tâches peuvent être :
 - des transformations
 - des requêtes SQL
 - des manipulations de fichiers (copie, suppression, téléchargement, etc.)
 - des tests conditionnels
 - des scripts (*shell*, JavaScript)
 - des envois / réceptions de mails
 - d'autres *jobs*
 - etc.



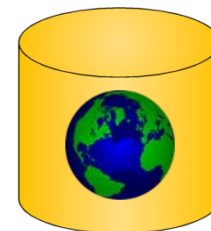
Le référentiel

- Les *transformations* et *jobs* sont normalement stockés dans des **fichiers XML** (.ktr/.kjb)
- Alternative : stockage d'un **référentiel dans une BD**
 - Les transformations, jobs et paramètres de connexions aux SGBD sont stockés dans une BD dédiée

Fonctionnalités spatiales de GeoKettle

Support spatial intégré

- Intégration cohérente des géométries vectorielles:
 - **Type de données *Geometry*** : géométries vectorielles (**JTS** – modèle point-ligne-polygone)
 - **Conversions** transparentes entre **types de données** :
 - *Geometry* \leftrightarrow *String* : depuis et vers WKT
 - *Geometry* \leftrightarrow *Binary* : depuis et vers WKB
 - Support des **SGBD spatiaux** intégré dans le noyau d'E/S pour SGBD (utilisant JDBC)
 - Tous les *steps* pouvant accéder aux BD **supportent les colonnes géométriques de manière transparente.**



Entrées / sorties

- Lecture/écriture de géométries :

- **SGBD spatiaux:**

- PostgreSQL/PostGIS
- MySQL spatial
- Oracle Spatial / Locator



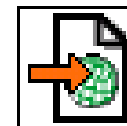
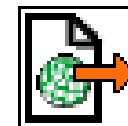
ORACLE

Note : il n'y a pas de *steps* distincts et dédiés spécifiquement aux E/S PostGIS, Oracle Spatial et MySQL, puisque **tous les *steps* de BD existants ont accès aux colonnes géométriques.**

→ Philosophie : ne pas ajouter inutilement de nouveaux *steps* quand ceux existants font déjà le travail tout en étant plus versatiles !

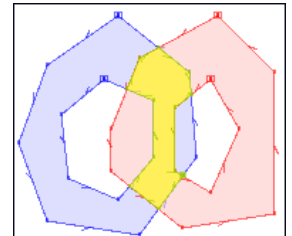
- Formats de **fichiers SIG:**

- ShapeFile
- GML (en développement)



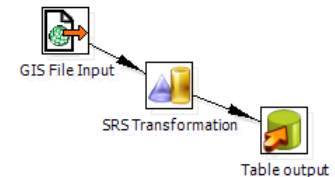
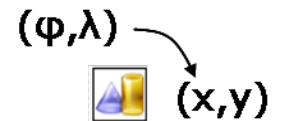
Analyses spatiales

- *Scripting* sur objets géométriques en langage JavaScript.
- Fonctions d'analyse spatiale:
 - **Prédicats topologiques** : *intersects*, *touches*, *within*, etc.
 - exploitables à partir des *steps* de filtrage et de jointure
 - **Fonctions spatiales** : *union*, *intersection*, *length*, *buffer*, etc., soit toutes celles offertes par la librairie JTS
 - accessibles en *JavaScript*



SRS et projections

- Gestion native des **systemes de référence spatiaux** (SRS) dans les métadonnées des champs *Geometry* (bibliothèque **GeoTools – referencing**)
- **Reprojection** / changement de système de référence spatiale
- Affectation d'un SRS
- **Lecture et écriture** des métadonnées de **SRS** :
 - Lecture des SRS depuis les sources de données : SGBD et Shapefile (.prj)
 - Validation du SRS lors de l'insertion de données dans PostGIS et Oracle
 - Écriture du fichier .prj lors de la création d'un Shapefile



Exemples avec GeoKettle

Exemple 2 – Lecture Shapefile

The screenshot displays the Spoon (GeoKettle) application window titled "Spoon (GeoKettle) - exemple 2 - lecture shapefile". The interface includes a menu bar with "Fichier", "Edition", "Vue", "Référentiel", "Transformation", "Tâche", "Assistant", and "Aide". Below the menu is a toolbar with various icons and a "100%" zoom level. The main workspace shows a data flow diagram with three steps: "GIS File Input (RRN)", "Set SRS (epsg:4140)", and "Insertion dans table (rm_quebec)".

On the left side, there is a "Navigateur" (Navigator) panel with a search icon and a "Palette de création" (Creation Palette) with a paintbrush icon. Below these is a "Transformations" tree view showing a folder structure:

- Transformations
 - exemple 2 - lecture shape
 - Connexions
 - Étapes
 - Liens
 - Schémas partitionner
 - Serveurs esclave
 - Schémas grappe PDI

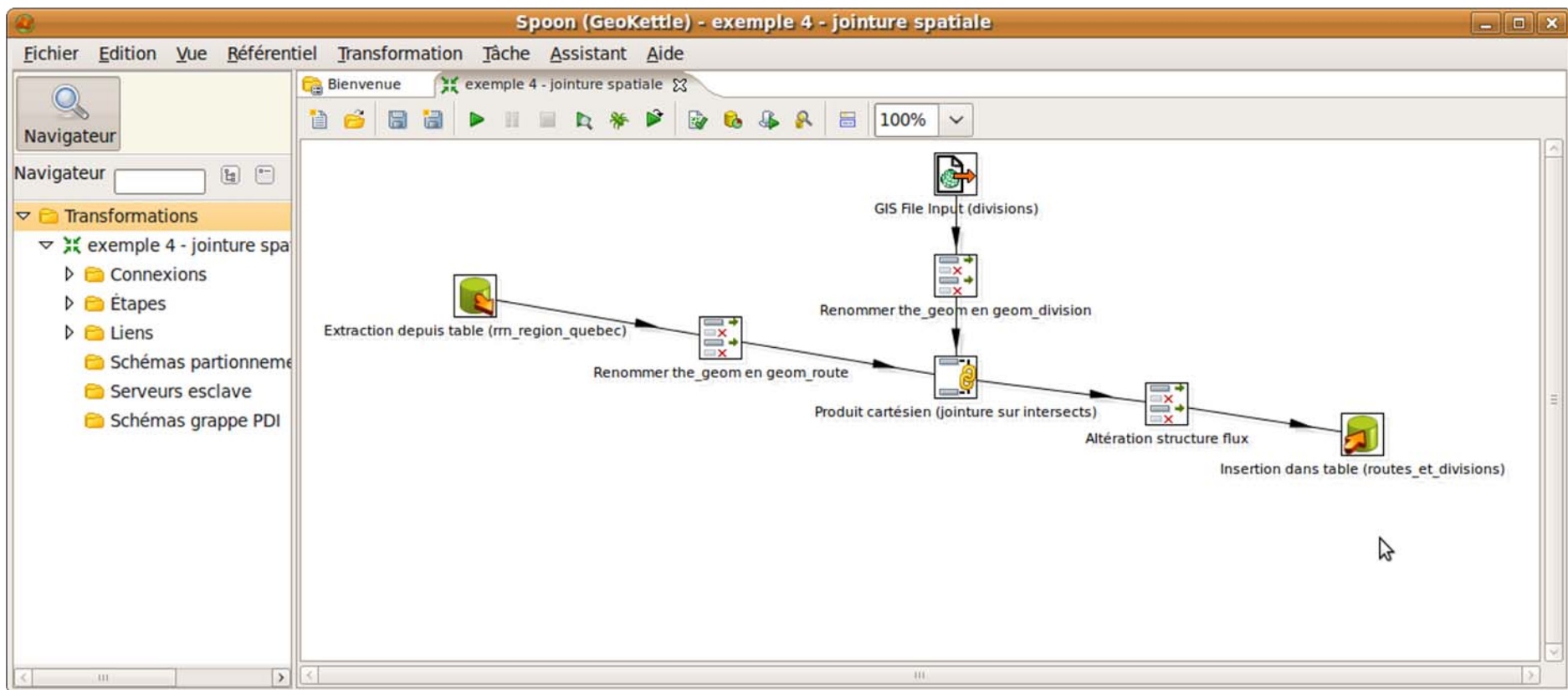
Exemple 3 – Filtrage spatial

The screenshot displays the Spoon (GeoKettle) application window titled "Spoon (GeoKettle) - exemple 3 - filtrage spatial". The interface includes a menu bar with options: Fichier, Edition, Vue, Référentiel, Transformation, Tâche, Assistant, Aide. Below the menu is a toolbar with various icons for file operations and execution. The main workspace shows a workflow diagram with four steps connected by arrows:

- Extraction depuis table (rm)
- Altération structure flux
- Filtrage lignes (intersects)
- Insertion dans table (rm_region_quebec)

The left sidebar contains a "Navigateur" (Navigator) and a "Palette de création" (Creation Palette). Under "Transformations", the current project "exemple 3 - filtrage spatial" is expanded, showing sub-folders: Connexions, Étapes, Liens, Schémas partitionnement, Serveurs esclave, and Schémas grappe PDI.

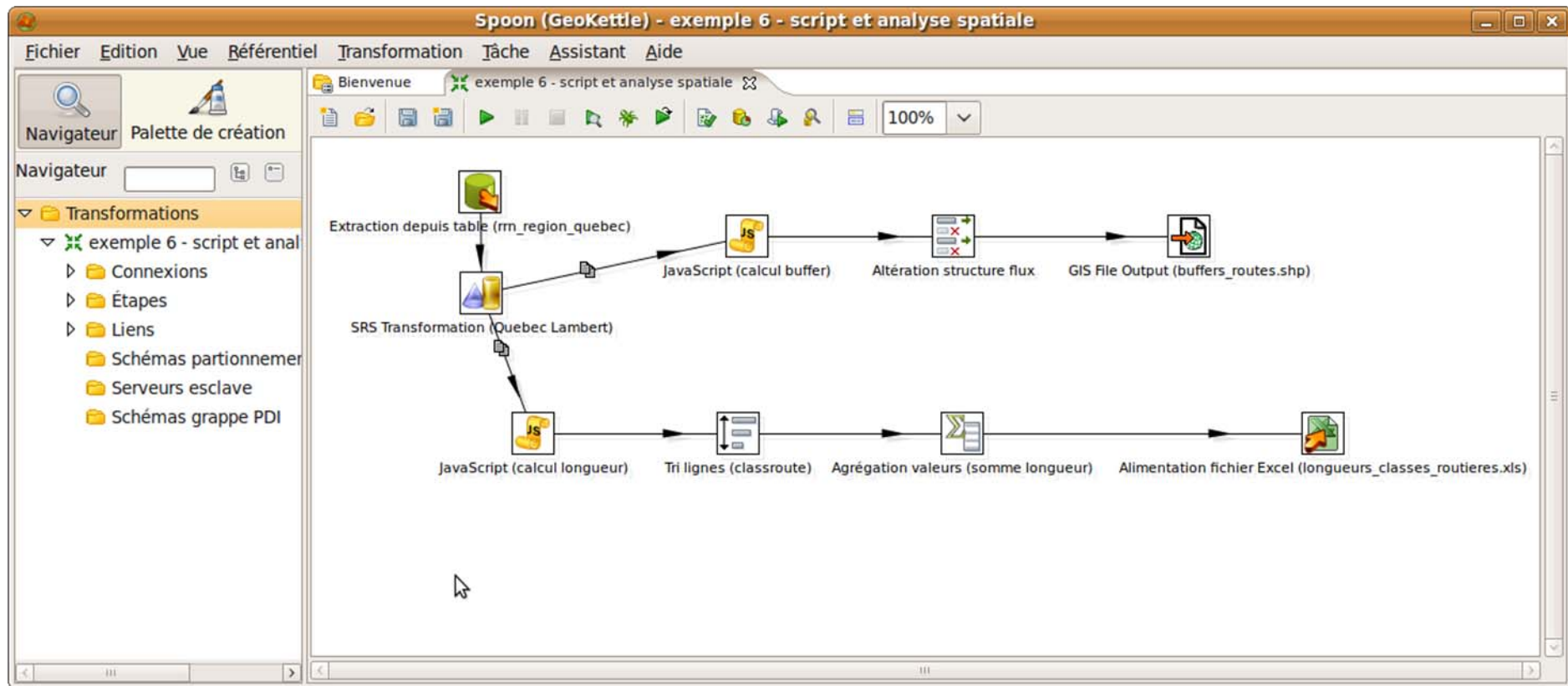
Exemple 4 – Jointure spatiale



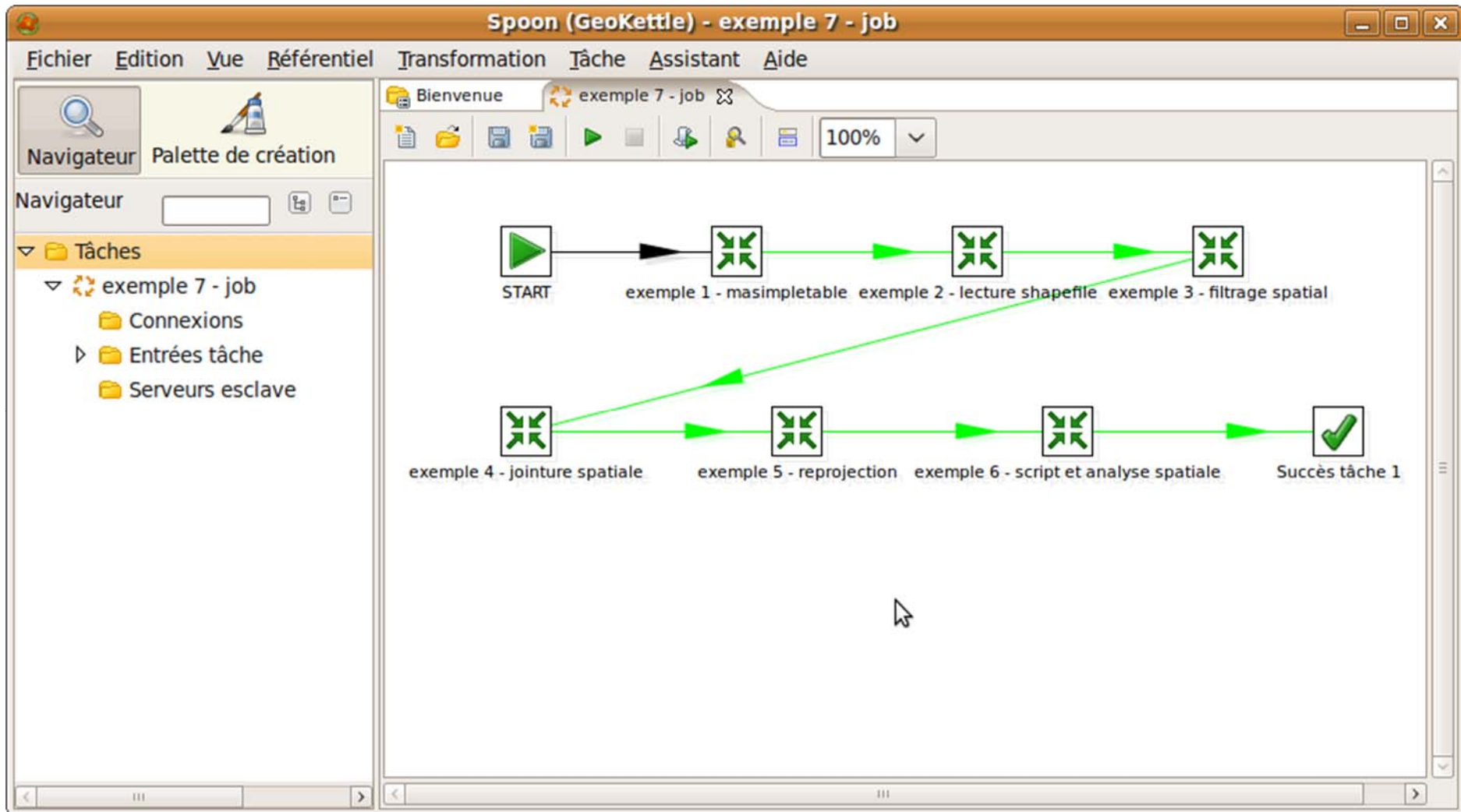
Exemple 5 – Reprojection

The screenshot displays the Spoon (GeoKettle) software interface. The title bar reads "Spoon (GeoKettle) - exemple 5 - reprojection". The menu bar includes "Fichier", "Edition", "Vue", "Référentiel", "Transformation", "Tâche", "Assistant", and "Aide". The left sidebar shows a "Transformations" tree with "exemple 5 - reprojection" expanded, containing sub-items like "Connexions", "Étapes", "Liens", "Schémas partionnemer", "Serveurs esclave", and "Schémas grappe PDI". The main workspace shows a workflow with three steps: "Extraction depuis table", "SRS Transformation (lat./long. vers UTM)", and "GIS File Output (routes_et_divisions_UTM19N.shp)". The interface also features a "Navigateur" and "Palette de création" at the top left, a toolbar with various icons, and a zoom level of 100%.

Exemple 6 – Script et analyse spatiale



Exemple 7 – job



Conclusion

Améliorations à venir

- Prévisualisation cartographique
- Autres formats de fichiers (GML, MapInfo, etc.)
- Accès aux services WFS
- *Steps* avec dialogues pour fonctions d'analyse spatiale (buffer, union, etc.).
- ...

Pour en savoir plus

- GeoKettle sur le web :
 - <http://www.geokettle.org>
- Pentaho Data Integration (Kettle) :
 - <http://kettle.pentaho.org/>
 - La documentation de Kettle est également pertinente pour GeoKettle !
- Projet sur SourceForge :
 - <http://sourceforge.net/projects/geokettle/>
 - Listes de diffusion :
 - geokettle-users
 - geokettle-users-fr (en français)
 - geokettle-devel
 - Accès Subversion
 - Trackers (bugs, suggestions, etc.)
- Vos commentaires, suggestions et contributions sont les bienvenus !