Présentation de Talend Open Studio

Logiciel Data Integration

Éric Quinton
eric.quinton@irstea.fr

21 octobre 2015

Sommaire

La société

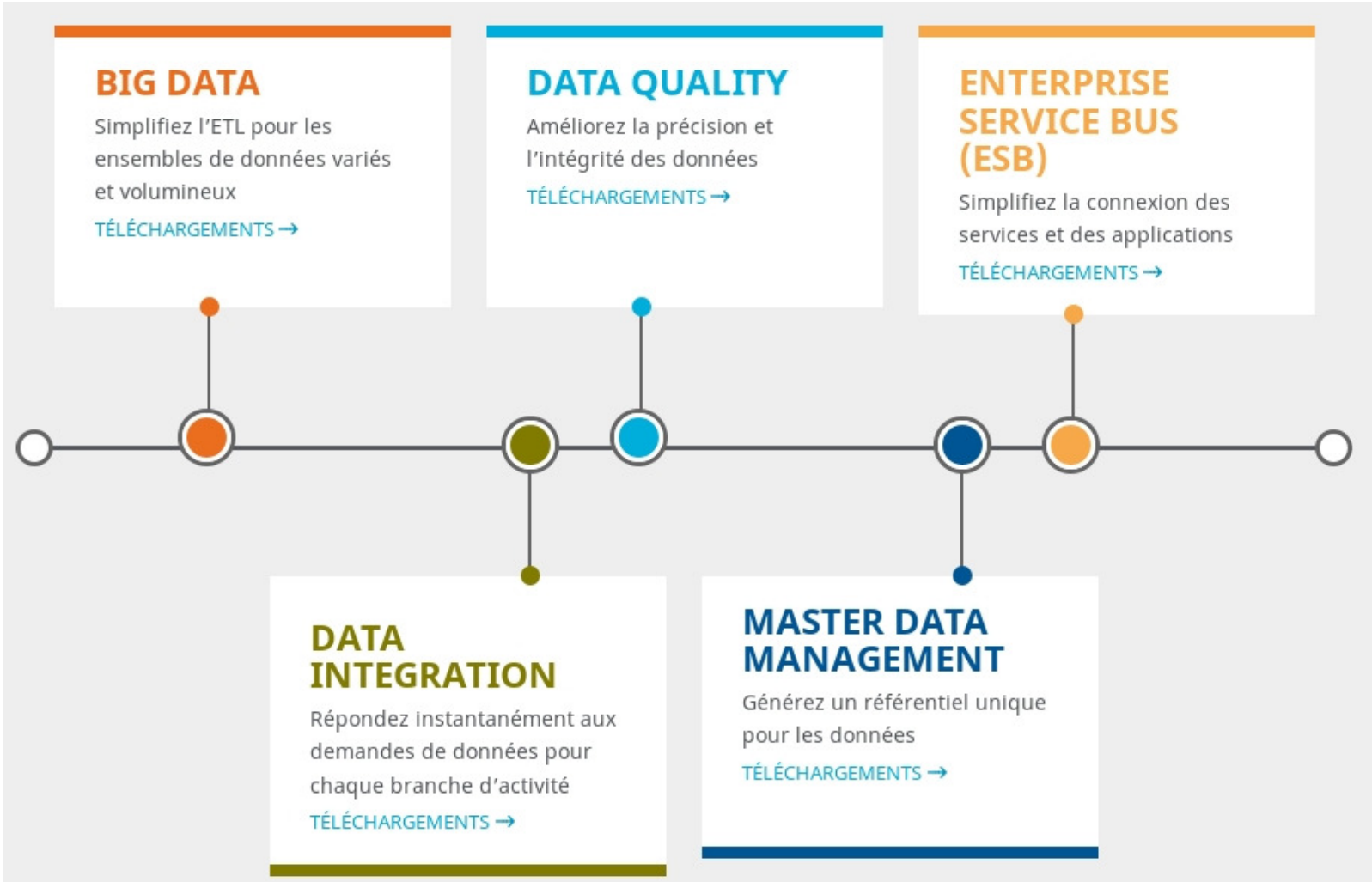
Talend est une société créée en 2006 à Suresnes (France) :

- 400 salariés ;
- siège en France et en Californie ;
- croissance annuelle du chiffre d'affaires > 100 %

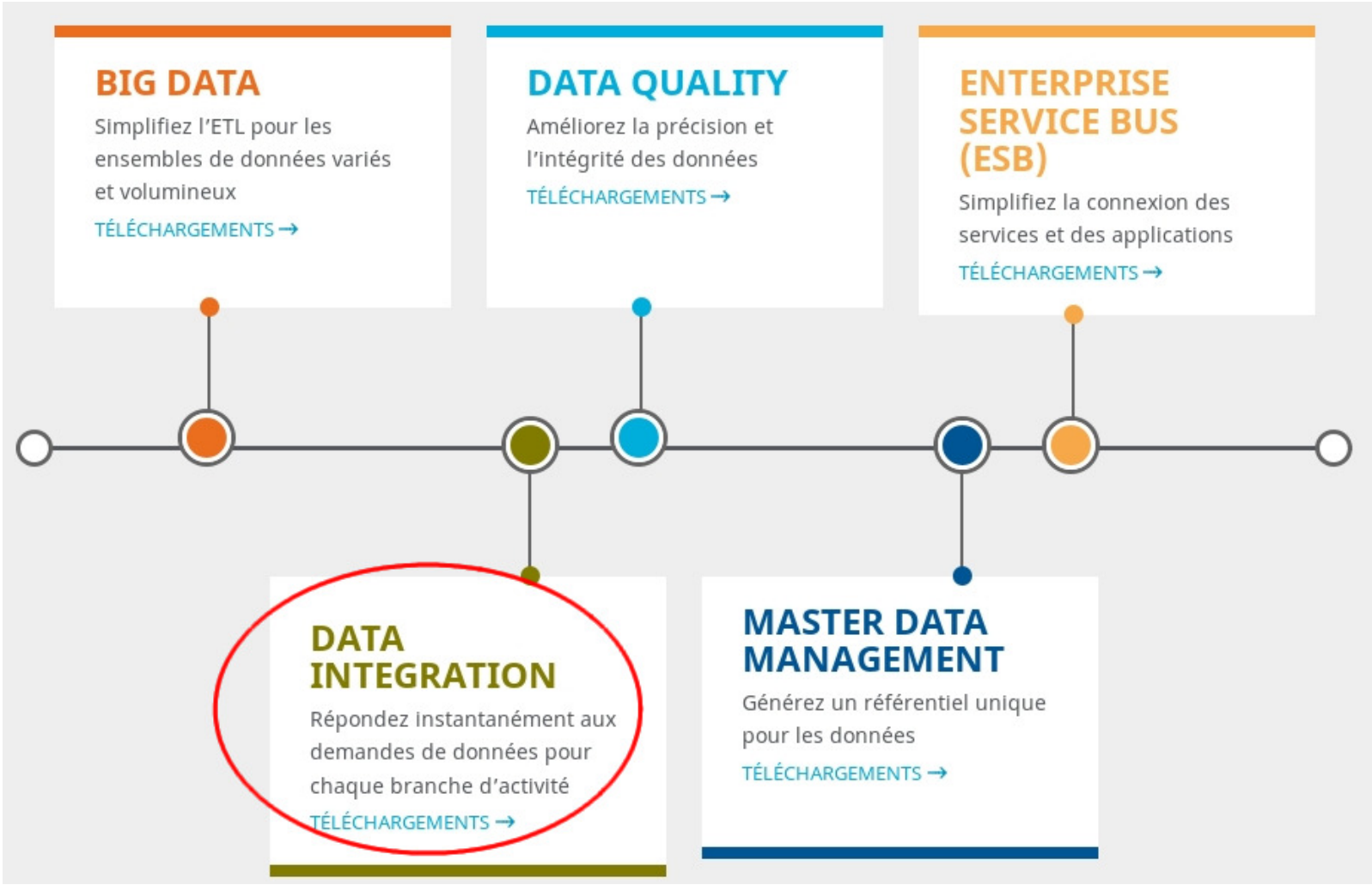
Deux familles de logiciels :

- des produits professionnels, avec support et travail collaboratif ;
- des produits Open Source, parfois limités en matière d'automatisation ;
  - regroupés sous l'appellation TOS (Talend Open Studio).

TOS : un ensemble de logiciels dédiés aux flux de données

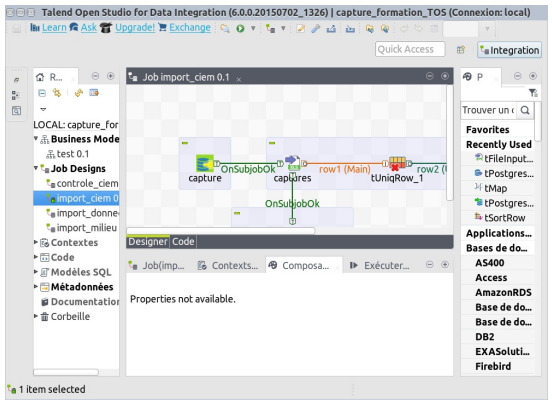


TOS : un ensemble de logiciels dédiés aux flux de données



Talend Open Studio

- Une plate-forme basée sur Eclipse RCP (Rich Client Platform) :
  - possibilité de gérer des connexions multiples, notamment pour travailler en équipe (disponible dans la version payante) ;
  - possibilité de gérer des projets multiples ;
  - une ergonomie classique pour les logiciels basés sur Eclipse RCP :



- Un arbre à gauche, avec les objets manipulés ;
- une fenêtre centrale, pour dessiner les opérations ;
- un arbre à droite, avec les objets à insérer ;
- des boites de paramètres ou d'exécution, en bas

Principes généraux

- des composants graphiques sont assemblés :
  - soit par des flux (entrée, sortie)
  - soit par des événements (traitement ok, ko)

Principes généraux

- des composants graphiques sont assemblés :
  - soit par des flux (entrée, sortie)
  - soit par des événements (traitement ok, ko)
- TOS génère, puis exécute, du code Java :
  - le code peut être visualisé (pratique pour déboguer) ;
  - il faut faire attention au typage (int vs Integer, conversions de types) ;
  - les tests utilisent la notation ternaire : *condition* ? **ok** : **ko**

Principes généraux

- des composants graphiques sont assemblés :
  - soit par des flux (entrée, sortie)
  - soit par des événements (traitement ok, ko)
- TOS génère, puis exécute, du code Java :
  - le code peut être visualisé (pratique pour déboguer) ;
  - il faut faire attention au typage (int vs Integer, conversions de types) ;
  - les tests utilisent la notation ternaire : *condition* ? **ok** : **ko**
- il est possible d'exporter un job pour l'exécuter en ligne de commande Java.

Il n'est pas nécessaire de connaître Java

mais avoir quelques notions est un plus !

TOS Data Integration - principes

- Il manipule des matrices de données sous forme de flux :
  - données en lignes, attributs en colonnes ;
  - une itération par ligne ;
    - il n'est pas adapté aux formulaires Excel (A4, B8, F15, A12 à récupérer, p. e.) ;
    - il vaut mieux recourir à un logiciel de pré-traitement qui normalisera les données

TOS Data Integration - principes

- Il manipule des matrices de données sous forme de flux :
  - données en lignes, attributs en colonnes ;
  - une itération par ligne ;
    - il n'est pas adapté aux formulaires Excel (A4, B8, F15, A12 à récupérer, p. e.) ;
    - il vaut mieux recourir à un logiciel de pré-traitement qui normalisera les données
- il peut associer des données de sources multiples :
  - bases de données ;
  - fichiers textes : Excel, CSV, XML ;

TOS Data Integration - principes

- Il manipule des matrices de données sous forme de flux :
  - données en lignes, attributs en colonnes ;
  - une itération par ligne ;
    - il n'est pas adapté aux formulaires Excel (A4, B8, F15, A12 à récupérer, p. e.) ;
    - il vaut mieux recourir à un logiciel de pré-traitement qui normalisera les données
- il peut associer des données de sources multiples :
  - bases de données ;
  - fichiers textes : Excel, CSV, XML ;
- les données sont associées en jointures internes (INNER JOIN) ou externes (OUTER JOIN), gérées par le logiciel
  - possibilité de traiter différemment les données non comprises dans une jointure interne
  - intéressant pour détecter des anomalies...

Une gestion paramétrée des données

- Les données (entrantes ou sortantes) sont décrites sous la forme de métadonnées ;
- Les variables utilisées dans les traitements sont regroupées sous l'appellation de **Contextes** ;
  - un même jeu de variables peut être décrit dans des contextes différents
  - les connexions aux bases devraient être traitées sous la forme de contextes :
    - développement, pré-production, production...
    - c'est la seule possibilité pour définir un schéma autre que le schéma *public*

Les composants graphiques

- des composants dédiés aux flux entrants ou sortants :
  - lecture - écriture des principales bases de données : *tPostgresqlRow*, *tPostgresqlOutput*...
    - avec ou sans support des transactions
  - lecture - écriture de fichiers textes (excel, xml, csv, json...)

Les composants graphiques

- des composants dédiés aux flux entrants ou sortants :
  - lecture - écriture des principales bases de données : *tPostgresqlRow*, *tPostgresqlOutput*...
    - avec ou sans support des transactions
  - lecture - écriture de fichiers textes (excel, xml, csv, json...)
- des composants de transformation :
  - *tMap* : mixer plusieurs flux et en créer de nouveaux. C'est LE composant à connaître
  - *tSortRow*, *tFilterRow*, *tUniqRow*... pour trier, filtrer, conserver les valeurs uniques...
  - *tNormalize* et *tDenormalize*, pour transformer des données présentées en lignes en colonnes et inversement

Les composants graphiques

- des composants dédiés aux flux entrants ou sortants :
  - lecture - écriture des principales bases de données : *tPostgresqlRow*, *tPostgresqlOutput*...
    - avec ou sans support des transactions
  - lecture - écriture de fichiers textes (excel, xml, csv, json...)
- des composants de transformation :
  - *tMap* : mixer plusieurs flux et en créer de nouveaux. C'est LE composant à connaître
  - *tSortRow*, *tFilterRow*, *tUniqRow*... pour trier, filtrer, conserver les valeurs uniques...
  - *tNormalize* et *tDenormalize*, pour transformer des données présentées en lignes en colonnes et inversement
- des composants pour visualiser et déboguer :
  - *tLogRow* : affiche les infos traitées

Les composants graphiques

- des composants dédiés aux flux entrants ou sortants :
  - lecture - écriture des principales bases de données : *tPostgresqlRow*, *tPostgresqlOutput*...
    - avec ou sans support des transactions
  - lecture - écriture de fichiers textes (excel, xml, csv, json...)
- des composants de transformation :
  - *tMap* : mixer plusieurs flux et en créer de nouveaux. C'est LE composant à connaître
  - *tSortRow*, *tFilterRow*, *tUniqRow*... pour trier, filtrer, conserver les valeurs uniques...
  - *tNormalize* et *tDenormalize*, pour transformer des données présentées en lignes en colonnes et inversement
- des composants pour visualiser et déboguer :
  - *tLogRow* : affiche les infos traitées
- des composants d'orchestration :
  - *tFileList*, pour traiter une liste de fichiers ou dossiers

Extension spatiale pour Talend

Il est possible d'intégrer des composants géographiques à Talend :

- fonctionne en s'appuyant sur GeoTools, GDAL et *Java Topology Suite*
- support des principaux formats SIG : Postgis, ESRI, MIF/MID, GPX, KML, Oracle Spatial, OGR Vector formats
- possibilité d'alimenter des serveurs géographiques :
  - Standards OGC : CSW (envoi de métadonnées), WFS
  - Geographic information metadata standard ISO19139
  - GeoServer REST API

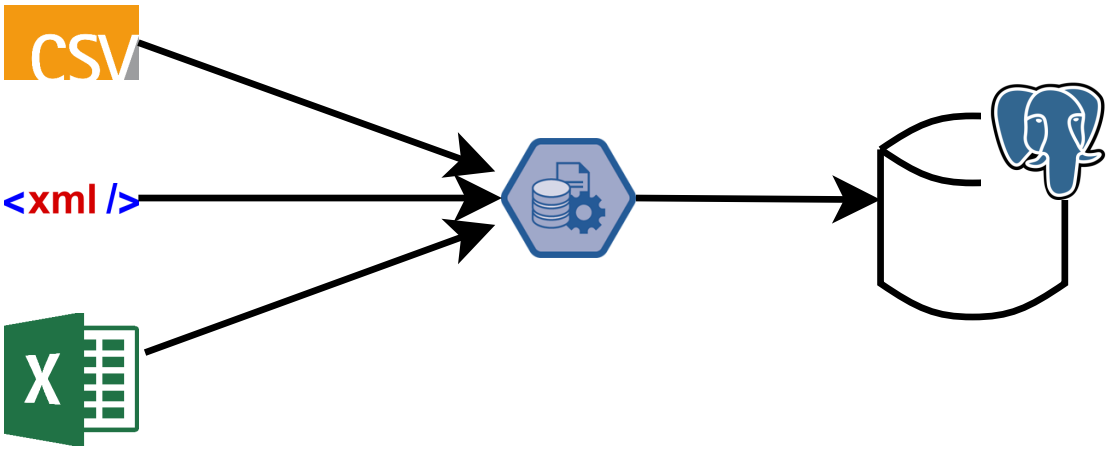
Plus d'informations sur :

`https://talend-spatial.github.io/`

Quelques exemples de cas d'utilisation

- Créer ou alimenter une base de données
- Synchroniser des bases de données
- Extraire les métadonnées de couches géographiques
- Transformer les données pour alimenter des bases infocentres
- ...

Créer ou alimenter une base de données

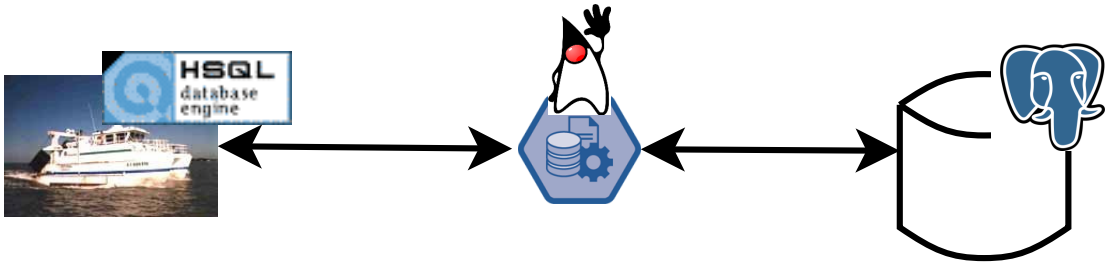


Attention à la qualité des données

- Libellés différents pour le même contenu ;
- formats mal maîtrisés (texte dans des champs numériques ou des dates...)
- ...

Toujours réaliser des tests dans une plate-forme dédiée avant de passer en production

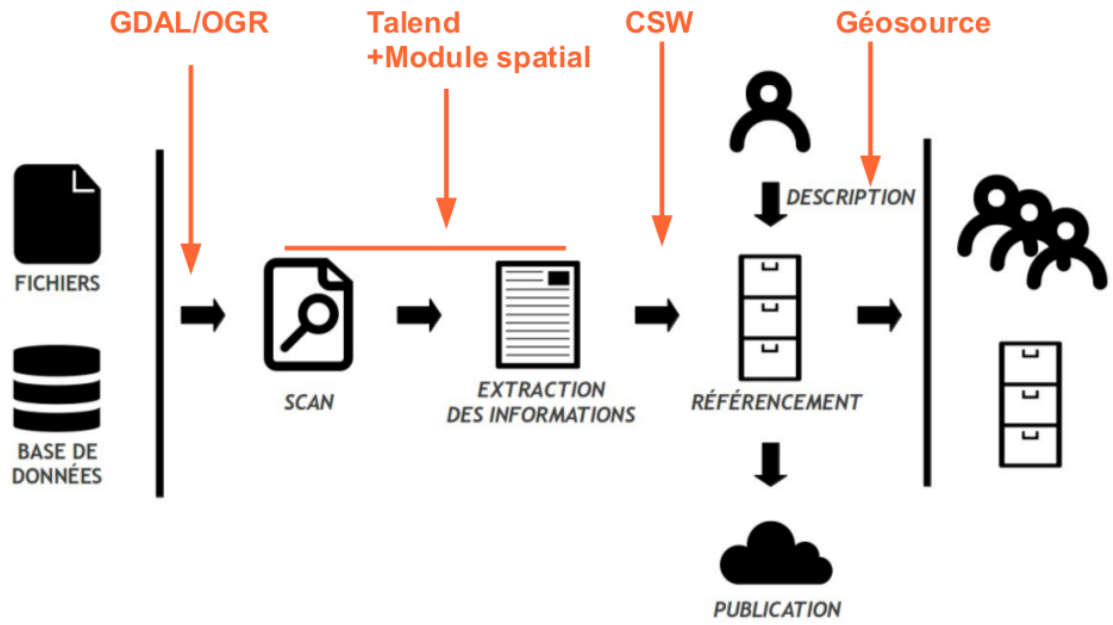
Synchroniser des bases de données



- Les données sont synchronisées entre la base Postgresql et une base embarquée dans le bateau (application Java de saisie)
- 4 scripts créés :
  - création de la base de données embarquée
  - copie des tables de référence ou de paramètres
  - synchronisation des données vers la base embarquée
  - synchronisation depuis la base embarquée vers le serveur
- les scripts sont exécutés en ligne de commande par l'utilisatrice

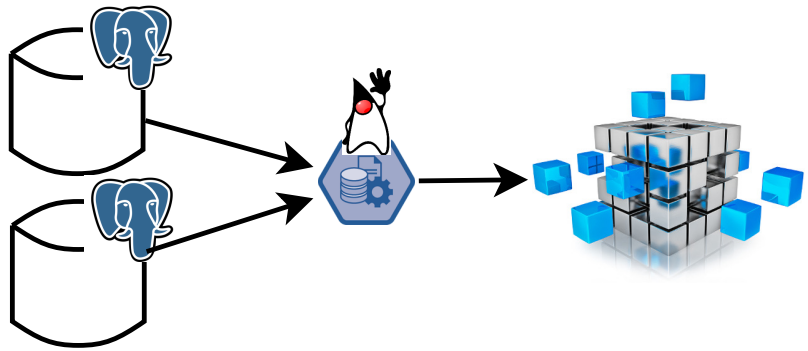
Extraire les métadonnées de couches géographiques

- Projet mené par les Parcs nationaux de France :
  - exporter et publier les métadonnées et les attributs de toutes les couches géographiques



http://forum-tic.espaces-naturels.fr/sites/default/files/fichiers/presentations/bruno_lafage_-pnf-forum_tic_2014.pdf

Transformer les données pour alimenter des bases infocentres



- les informations sont extraites d'une ou plusieurs bases de données relationnelles pour alimenter :
  - un cube Olap ;
  - un moteur noSql ;
  - du web sémantique...
- objectif : représenter les données sous une forme différente
  - adapter les données aux traitements à effectuer ;
  - concilier schéma relationnel et moteurs spécialisés.

De la documentation en ligne

`https://help.talend.com/display/ComposantsTalendOpenStudioGuidedereference60FR/Home`

- l'accès nécessite un enregistrement préalable ;
- il faut être patient...
- mais les composants sont décrits dans le moindre détail, et en français

Ce n’est pas parce que c’est graphique que c’est évident !

- il faut être méticuleux et tester ;
- une plate-forme de test est indispensable ;
- mais... cela reste un outil unique, qui fait gagner un temps énorme lors de la manipulation de données hétérogènes.