Projet Entrepôts de données

Pierrick Moreau, Nicolas Marechal

# Introduction

* Le but de ce projet est de monter un système BI basé sur des données libres de droit au choix (wikidata, freebase, data.gouv) que l’on va croiser avec d’autres données puis construire des cubes d’analyse et générer des rapports pertinents avec les outils de notre choix également.
* Le projet dure un peu plus de 2mois au cours duquel nous avons choisi nos données, réfléchi à nos problématique, créer nos bases de production, utiliser un ETL performant pour les extraire vers le format que nous souhaitons puis générer nos cubes d’analyse et nos rapports afin de répondre aux questions que l’on se posait.

# 1ère étape : choix des données

* <https://www.data.gouv.fr/fr/datasets/base-de-donnees-accidents-corporels-de-la-circulation-sur-6-annees/>
* Déjà deux sources de données différentes : accidents et véhicules impliqués
* **Données gouvernementales** : données intègres et complètes puisque ces données sont mises à jour directement par les forces de l’ordre lors de la constatation de l’accident.
* **Aucunes données inutiles**: chaque accident à au moins un blessé.
* **Données récentes** (2006-2011) et incluant tous les départements français (DOM-TOM inclus)
* Données complètes -> de nombreuses problématiques pertinentes possibles

SUJETS CHOISIS :

* Quel est nombre d’accidents par an en France ? Quels départements sont les plus accidentées ?
* Quel est le nombre de morts sur la route ? Quels départements sont les plus meurtriers ?
* Corrélation avec le type de véhicule (moto, scooter, voiture, camion).

# 2ème étape : mise en place de l’ETL (Pentaho)

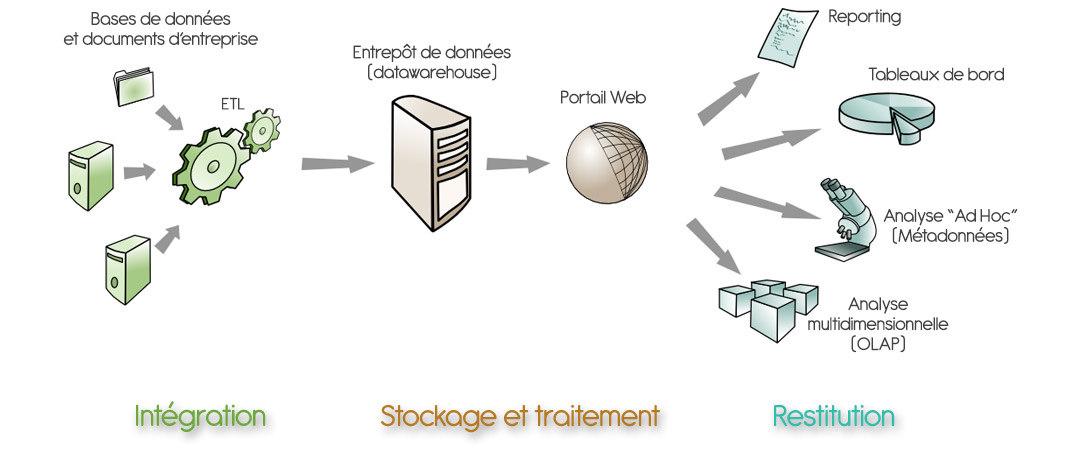
1. **Pourquoi Pentaho ?**

Pentaho est une suite logicielle décisionnelle complète qui permet la création et la diffusion de documents décisionnels à un grand nombre de destinataires via une interface web. Pentaho est entièrement open source, et propose une souscription pour des services éditeurs haut de gamme, incluant le support illimité, différentes garanties.

Pentaho comprend en propre tous les éléments nécessaires à un projet décisionnel, et est le seul éditeur open source capable d'en garantir l'homogénéité dans le temps.

* Un **moteur de reporting**, avec *Pentaho Report* Designer, un logiciel graphique de conception de rapports
* Un **outil de conception de metamodels**, avec *Pentaho Metadata* Designer, outil de définition métier d'une base de données relationnelle, permettant de requêter sur cette base sans connaissance technique.
* Un **moteur d'analyse multidimensionnelle**, avec *Pentaho Analysis* (Mondrian), le serveur OLAP utilisé par toutes les suites décisionnelles concurrentes mais maintenu par Pentaho. Il permet d'analyse un très grand volume de données, toujours très simplement pour l'utilisateur
* Un accès aux documents décisionnels via une **console web pour les utilisateurs, ergonomique, multi-onglets et sécurisée.**
* Une **console d'administration**et de supervision, gérant entre autres la diffusion en masse des rapports par email.
* Un **outil d'alimentation et de transformation**, avec *Pentaho Data Integration* (ex Kettle, Spoon), ETL très simple à prendre en main.
* Un outil de **datamining**, avec *Pentaho Weka*
* Un client MsExcel pour Mondrian
* un connecteur avec Google Maps, ...

Finalement, Pentaho possède un nombre important d’applications permettant un large panel de possibilités quant à l’intégration des données mais aussi à l’analyse multidimensionnelle et à la rédaction de rapport. Egalement, il est possible de le connecter avec google Maps ce qui est intéréssant au vue du sujet que nous avons choisi. C’est pourquoi nous avons décidé d’utiliser Pentaho.



1. **Comment mettre en place Pentaho ?**

La première chose à faire est d’installer **Pentaho Data Integration**, si possible la version complète afin de pouvoir utiliser **Spoon** ou **Kettle**, des briques permettant la connexion à une ou plusieurs BDD et les transformations à appliquées sur les données afin de générer notre datawarehouse. (Disponible sur Sourceforge - <http://sourceforge.net/projects/pentaho/> ).

Une fois installé, Il faut alors aller dans le dossier C\Pentaho\design-tools\data-integration et lancer le fichier de commande : Spoon. Celui-ci lancera Pentaho Data Integration et vous pourrez alors commencer l’intégration de vos données dans l’ETL.

1. **Création de l’ETL**

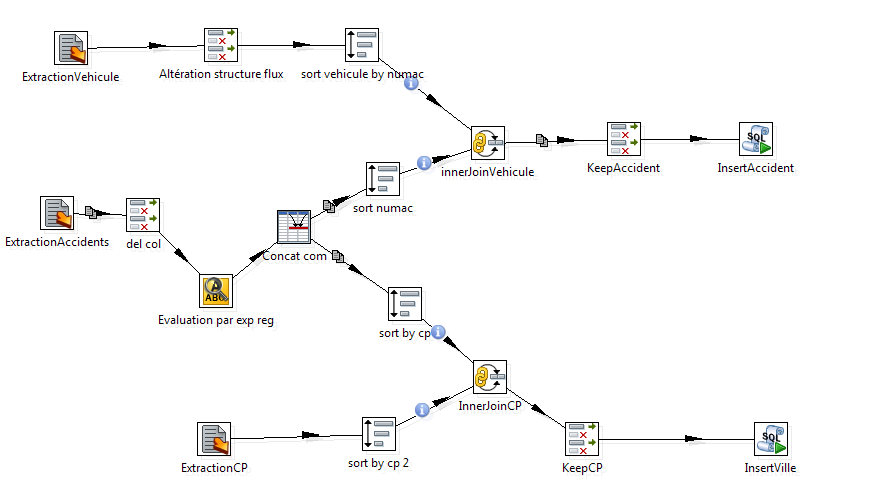
Pour créer notre ETL, il faut alors sélectionner un nombre lambda de transformations et opérations sur les données puis les ordonner afin d’optimiser le temps d’exécution.

Tout d’abord, étant donné que nos données sont issues de fichiers CSV, nous devons les extraire via la transformation associée au format CSV.

Puis une fois nos données extraites, nous devons faire une jointure à partir d’une clef de référence. La jointure entre les accidents et les véhicules se fait via le « numAcc » où numéro d’accidents. La jointure entre les accidents et la localisation se fait via l’attribut « com » (numéro INCEE de la commune où a eu lieu l’accident)

Ensuite, une fois les tables jointes, on peut effectuer nos requêtes d’insertion dans notre datawarehouse hébergé sur MySQL. Pour cela, nous avons créé une connexion MySQL et nous avons utilisé cette connexion dans une transformation de type « exécution de scripts SQL dynamique »

Ainsi, notre ETL est terminé et est modélisé de la manière suivante :



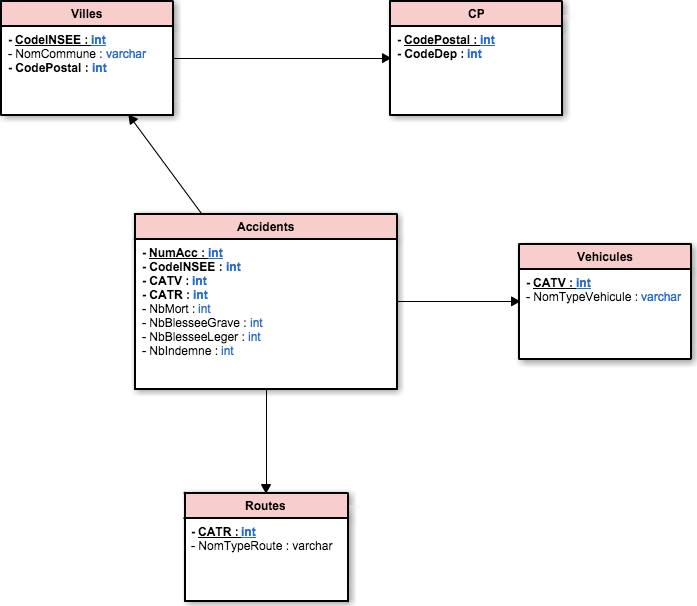
*ATTENTION : afin d’uniformiser certains champs, nous avons ajouté les etapes* ***‘del col’****,* ***‘concat com’*** *et l’evaluation par reg exp permettant, respectivement, de supprimer les colonnes inutiles (permettant ainsi l’optimisation de notre ETL) et la concaténation entre les champs com (code INSEE commune) et dep (numéro de département et normalisé le format du departement, rendant plus performant les jointures qui suivent.*

**‘InsertVéhicule’** insère les accidents dans la table de faits de notre datawarehouse.

**‘ExportLocalisation’** insère les données de localisation dans la table de Dimension Localisation.

1. **Description du datawarehouse obtenu**

Une fois notre ETL exécuté, on obtient donc notre datawarehouse ci-dessous :



La table Accidents représente la table de faits. Cette table contient l’ensemble des mesures qui nous serviront pour nos cubes OLAP et des liens vers les trois dimensions (Localisation, Type de Véhicule, Type de routes)

Nous avons choisi de le modéliser en flocon afin d’avoir plus de liberté quant au choix du degré de précision de la dimension ‘Localisation’. Notre datawarehouse est maintenant prêt à être exploité.

# 3ème étape : modélisation des cubes de données

Pour modéliser les cubes OLAP, il faut utiliser **‘Pentaho schema *workbench’*** comme suit :

## Etape 1 : Cube

* Création du cube

## Etape 2 : Faits

* Définir la table de faits

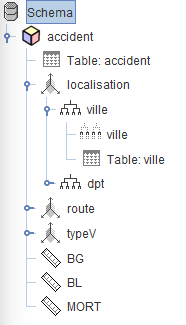
## Etape 3 : Dimensions

* Définir chacune des dimensions
* Lié la table de la base de données correspondantes à la dite dimension
* Définir la hiérarchie (différents niveaux de « zoom »)

## Etape 4 : Mesures

* Définir les différentes mesures que nous souhaitons effectuer sur le cube de l’étape 1. (Sum, Count, min, max…)

Voici le schéma crée :



Au final, nous obtenons un fichier XML permettant de définir la structure du cube OLAP et ses différentes caractéristiques :

<Schema name="New Schema1">

<Cube name="accident" visible="true" cache="true" enabled="true">

<Table name="accident" alias="">

</Table>

<Dimension type="StandardDimension" visible="true" foreignKey="com" name="localisation">

<Hierarchy name="ville" visible="true" hasAll="true" primaryKey="com">

<Table name="ville" alias="">

</Table>

<Level name="ville" visible="true" column="com" uniqueMembers="false">

</Level>

</Hierarchy>

<Hierarchy name="dpt" visible="true" hasAll="true" primaryKey="dpt">

<Table name="ville" alias="">

</Table>

<Level name="dpt" visible="true" column="dpt" uniqueMembers="false">

</Level>

</Hierarchy>

</Dimension>

<Dimension type="StandardDimension" visible="true" foreignKey="catr" name="route">

<Hierarchy name="route" visible="true" hasAll="true" primaryKey="catr">

<Table name="route" alias="">

</Table>

<Level name="route" visible="true" column="catr" uniqueMembers="false">

</Level>

</Hierarchy>

</Dimension>

<Dimension type="StandardDimension" visible="true" foreignKey="catv" name="typeV">

<Hierarchy name="vehicule" visible="true" hasAll="true">

<Table name="vehicule" alias="">

</Table>

<Level name="vehicule" visible="true" column="catv" uniqueMembers="false">

</Level>

</Hierarchy>

</Dimension>

<Measure name="BG" column="nbBlesseGrave" aggregator="sum" visible="true">

</Measure>

<Measure name="BL" column="nbBlesseLeger" aggregator="sum" visible="true">

</Measure>

<Measure name="MORT" column="nbMort" aggregator="sum" visible="true">

</Measure>

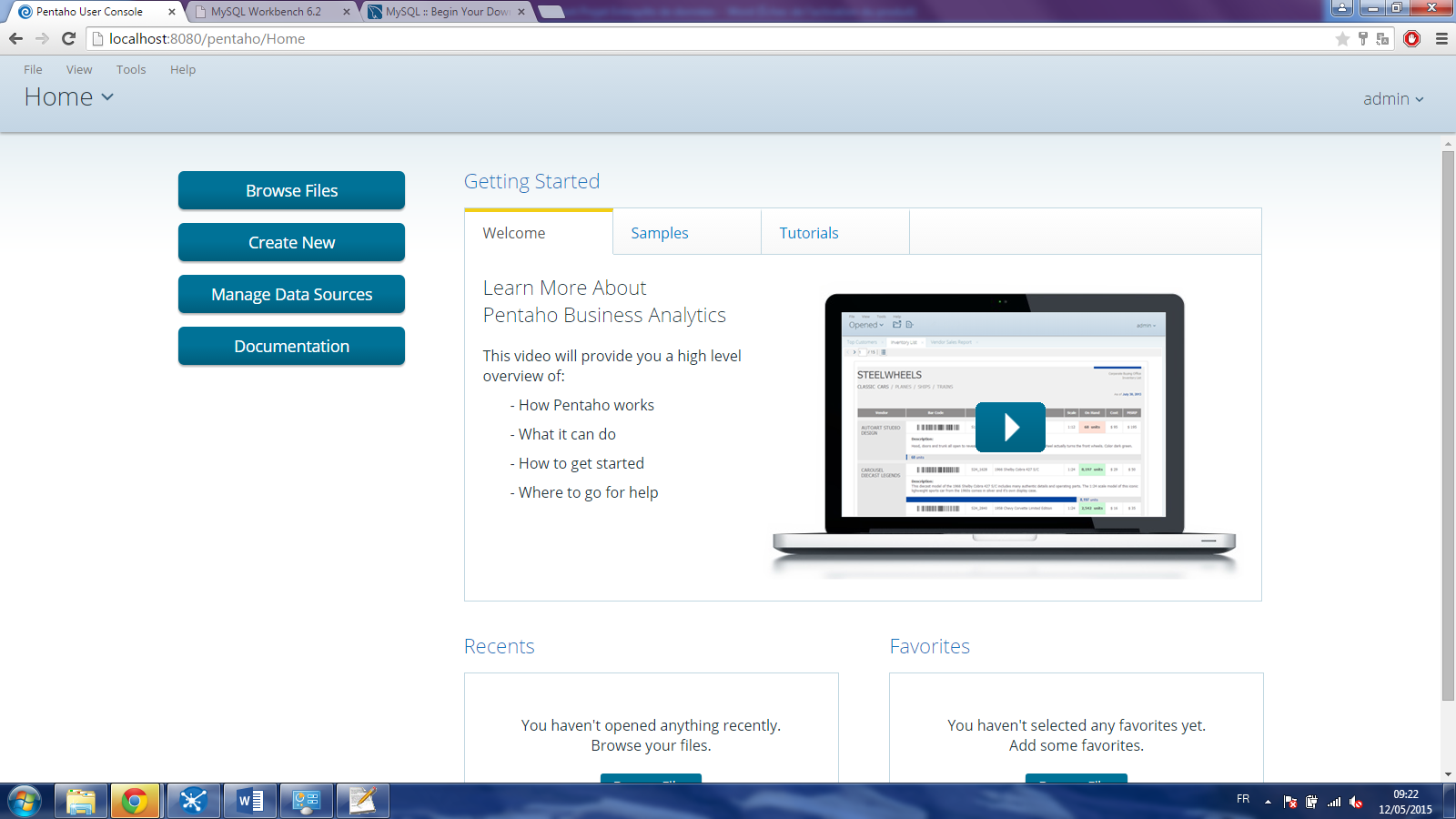
</Cube>

</Schema>

Ensuite, nous publions ce cube OLAP sur le BI Server afin de pouvoir l’utiliser dans notre exploitation afin de générer nos rapports.

# 4ème étape : exploitation des cubes et mise en place du reporting

Afin de mettre en place notre reporting et d’exploiter notre cube, nous devons lancer notre BI Server via la commande start-pentaho.bat. Une fois démarré, nous pouvons accéder à notre interface de reporting via l’adresse : <http://localhost:8080/pentaho/Home>



Ensuite, nous pouvons nous rendre dans l’espace de Management des Data Sources, créer une nouvelle data Source pointant vers notre datawarehouse (MySQL)

Puis nous pouvons relier notre Data source à notre cube OLAP via la table de faits et enfin utiliser notre data source afin de créer nos rapports via des actions sur le cube :

# Conclusion