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 : création de la base de production

1. **Insertion des données dans la base de production**
2. Filtrage des données afin de supprimer les informations inutiles ou incomplètes
   1. Suppression de colonne (V1, V2, gps, lat, long, coderivoli, distancemetre, PR, PR1)

# 3ème étape : mise en place de l’ETL (Pentaho)

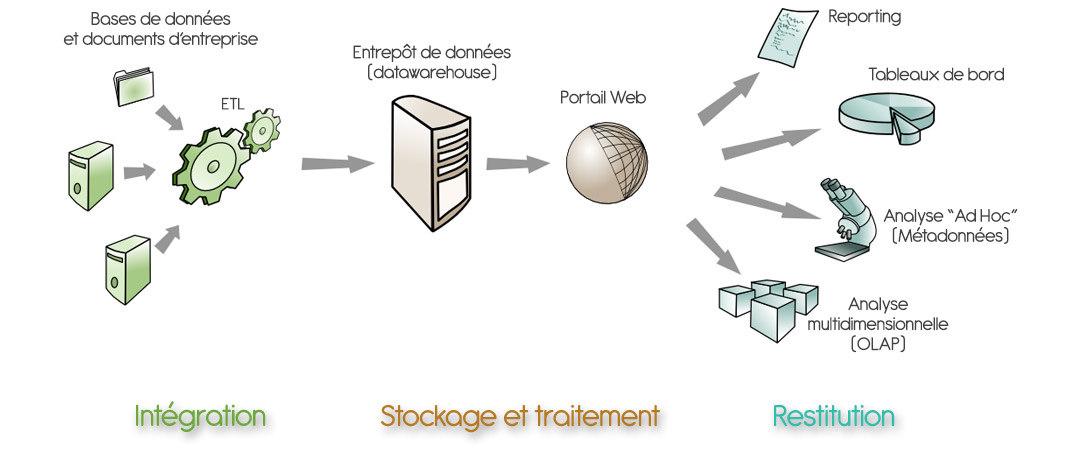
1. **Pourquoi Pentaho ?**

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 ?**

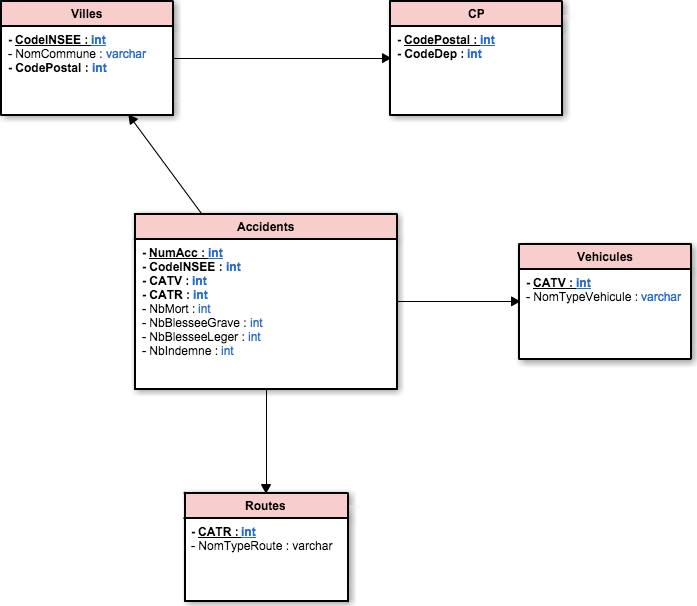
La première chose à faire est d’installer **Pentaho Data Integration**, si possible la version complète afin de pouvoir utiliser **Spoon** ou **Kettle**, des briques permettant la connexion à une ou plusieurs BDD et les transformations à appliquées sur les données afin de générer notre datawarehouse. (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**

…

1. **Description du datawarehouse obtenu**



# 4ème étape : modélisation des cubes de données

1. Nombres d’accidents par départements
2. Nombres de morts sur la route par départements
3. Pourcentage selon le type de véhicule et département

# 5ème étape : exploitation des cubes et mise en place du reporting

1. Nombres d’accidents par départements
2. Nombres de morts sur la route par départements
3. Pourcentage selon le type de véhicule et département

# Résultat

1. Nombres d’accidents par départements
2. Nombres de morts sur la route par départements
3. Pourcentage selon le type de véhicule et département

# Conclusion