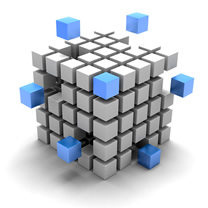


|  |
| --- |
| Master 2 MIAGE Apprentissage |
| Projet Entrepôt de données |
| **MARECHAL Nicolas & MOREAU Pierrick** |

|  |
| --- |
| Université Evry Val d’Essonne  19/05/2015 |



# 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)

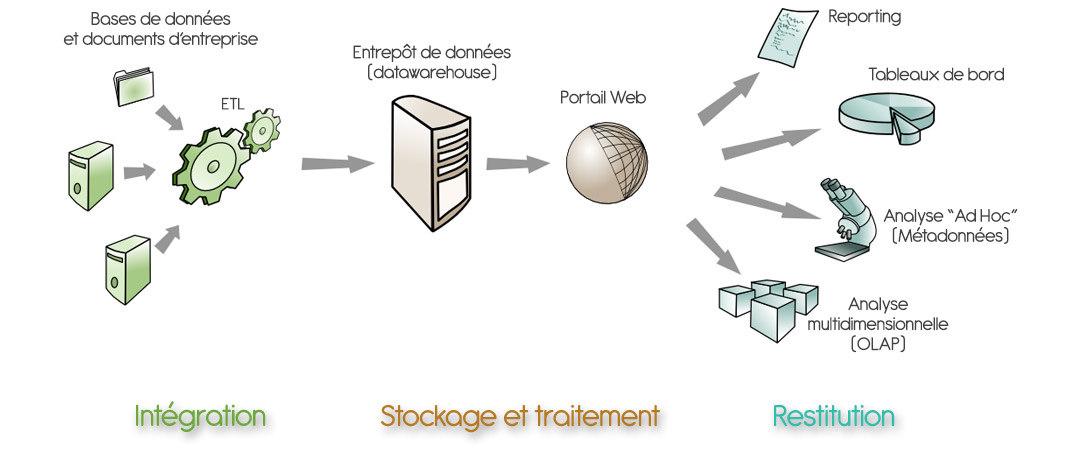
1. **Pourquoi Pentaho ?**

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**

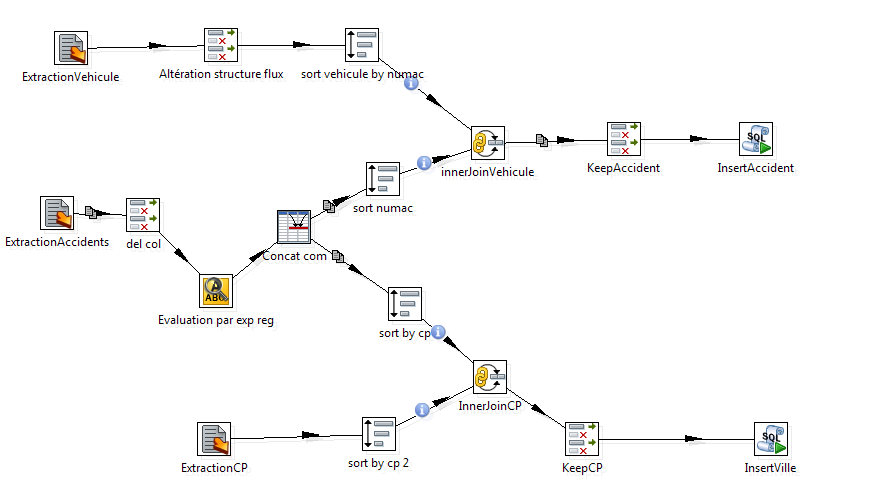
Pour créer notre ETL, il faut alors sélectionner un nombre lambda de transformations et opérations sur les données puis les ordonner afin d’optimiser le temps d’exécution.

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 :



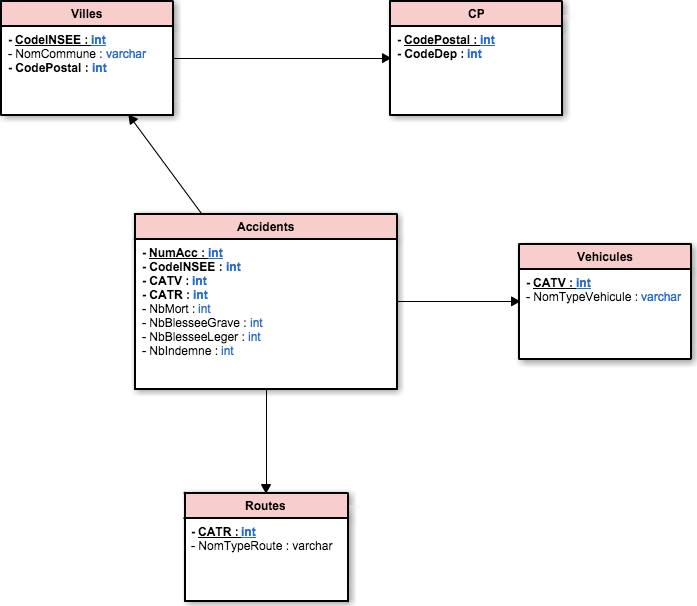
*ATTENTION : afin d’uniformiser certains champs, nous avons ajouté les etapes* ***‘del col’****,* ***‘concat com’*** *et l’evaluation par reg exp permettant, respectivement, de supprimer les colonnes inutiles (permettant ainsi l’optimisation de notre ETL) et la concaténation entre les champs com (code 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, après 45minutes d’attente et 663 348 insertions dans la table de faits, on obtient donc notre datawarehouse complet 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

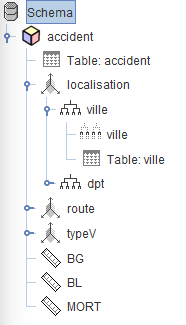
## Etape 3 : Dimensions

* 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>

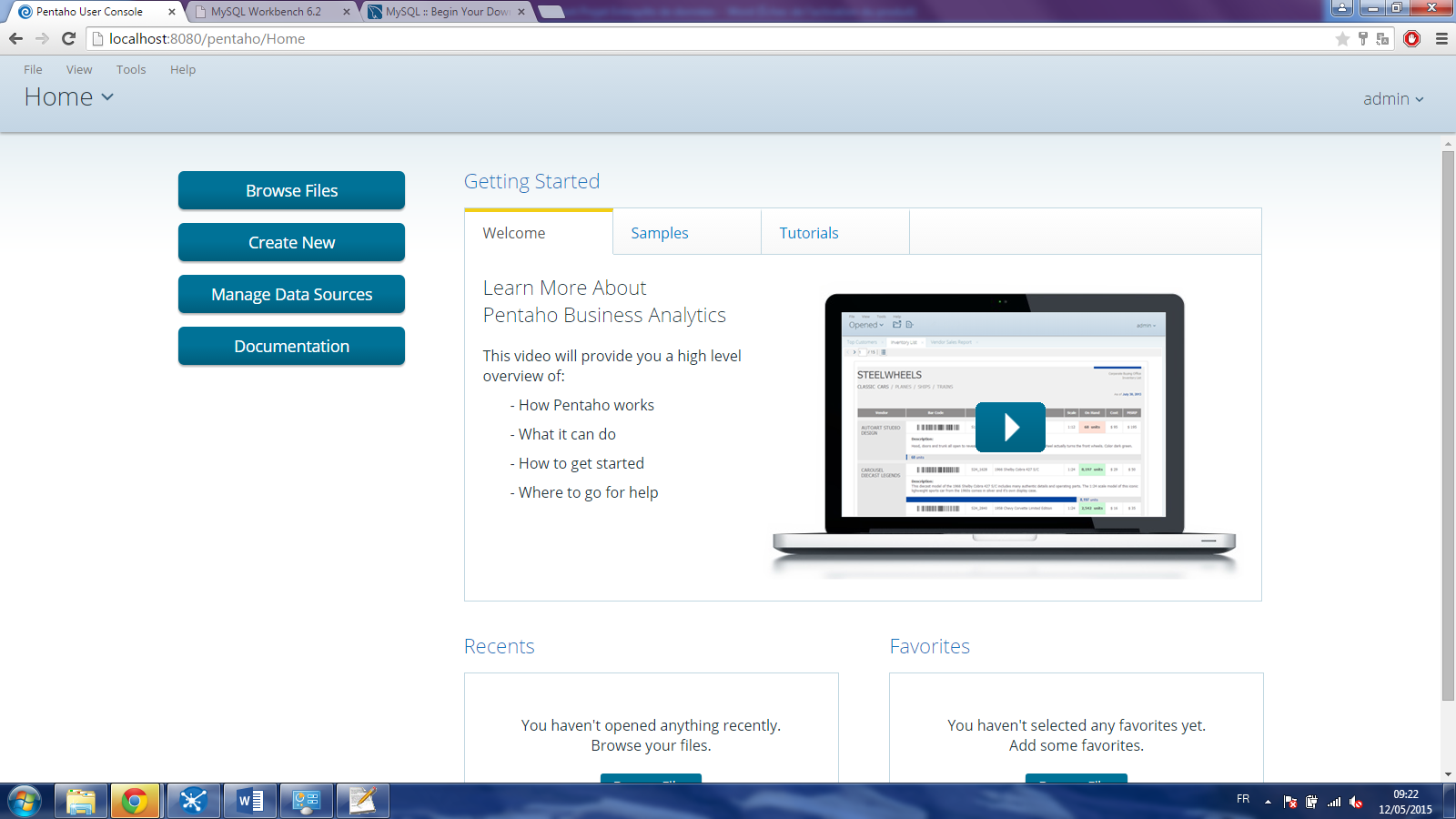
</Cube>

</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.

Ensuite, nous pouvons créer nos rapports via l’outil « Pentaho Report Designer ». Pour cela, on recréer la connexion vers notre datawarehouse puis on créer un ensemble de requêtes statiques nous permettant de générer les données utiles à afficher dans le rapport :

* « All accidents » : permettant de récupérer l’ensemble des informations sur la table de fait
* « All Accidents by com » : permettant de récupérer l’ensemble des accidents selon la commune
* « All Accidents by dpt » : permettant de récupérer l’ensemble des accidents selon le département
* « count all accident » : compteur d’accident général
* « count all accident by com » : compteur d’accidents par commune
* « count all accident by dpt » : compteur d’accidents par département
* « count all accident by catr » : compteur accidents par type de route
* « count all accident by catv » : compteur accident par type de vehicle

Example “Count all accident by catv” :

SELECT type, sum(nbMort), sum(nbBlesseGrave), sum(nbBlesseLeger), sum(nbIndemes)

FROM ACCIDENT A

INNER JOIN VEHICULE V ON A.ctav = V.catv

GROUP BY A.ctav

Une fois les requêtes prêtes, nous pouvons alors créer nos rapports avec le même outil. Pour cela, on peut insérer des graphiques, des textes comme ceci :

* Concernant notre sujet 1, à savoir *« Quel est nombre d’accidents par an en France ? Quels départements sont les plus accidentées ? »,* nous avons choisi d’afficher le nombre d’accidents en entête afin que celui-ci soit bien visible. Puis étant donné qu’afficher une centaine de colonnes dans un graphique serait illisible, nous avons pris la décision d’afficher les cinq départements plus accidentés et les cinq moins accidentés dans un BAR CHART.
* Concernant notre sujet 2, à savoir *«*Quel est le nombre de morts sur la route ? Quels départements sont les plus meurtriers ?*»,* nous avons affiché le nombre de mort en entête afin que celui-ci soit bien visible. Puis étant donné qu’afficher une centaine de colonnes dans un graphique serait illisible, nous avons pris la décision d’afficher les cinq départements les plus mortelles et les cinq départements moins accidentés dans un BAR CHART.
* Enfin le sujet 3 sera composé d’une étude plus détaillé concernant le nombre de mort (selon la route et selon le type de véhicule) dans des BAR CHART.

Les rapports ont été placés dans le dossier « rapport ». Malheureusement, certains rapports demandaient trop de ressources pour nos machines du fait du nombre très élevé de lignes dans notre table de faits (663 348 lignes) et de la complexité des requêtes à exécuter sur ses lignes. En effet, lorsque nous souhaitions grouper nos données selon le département, cela représente une centaine de groupes à créer et renseigner et pour cela, il doit analyser les 600 000 lignes de la table ce qui prend un temps incalculable. Par conséquent, seule deux rapports sont présents (les accidents triées par type de route, les accidents triés par type de vehicule)

# Conclusion

## Analyse des résultats obtenus

* **Sujet 1** : Selon le rapport, il y a 391 178 accidents en France depuis 2005. Soit une moyenne de 78235 accidents par an. A titre de comparaison, la France possède l’un des plus grand nombre d’accidents en Europe.

Si on s’intéresse aux départements les plus accidentés, il est normal d’y retrouver les départements parisiens car ceux-ci représente pas moins de 30% de la population Française.

* **Sujet 2** : d’après les résultats, il y a 40 933 personnes tués sur la route entre 2005 et 2011. Soit une moyenne de 6822 morts par an représentant ainsi le territoire le plus meurtrier excepté la Pologne.
* Si on s’intéresse aux départements les plus meurtries, il est normal d’y retrouver les départements parisiens car ceux-ci représente pas moins de 30% de la population Française.
* **Sujet 3** : d’après les résultats, Il y a beaucoup plus de morts sur des routes départemental en France avec X morts juste devant l’autoroute. En effet, s’il y a bien une chose de vrai sur les autoroutes française, c’est leur qualité et leur maintenance contrairement aux routes départementales.

Si on s’intéresse aux types de véhicule, Il y a beaucoup plus d’accidents entre VL (voiture légère) ce qui est normal aux vue du nombre de voitures en France, loin devant les motos et scooters légers.

## Ce que ce projet nous a apportés

Ce projet nous a permis d’en apprendre plus sur le datawarehouse et comment créer son propre entrepôt de données. Nous nous sommes rendus compte des différentes contraintes liés aux différentes étapes (ETL, report, données).

En effet, lors de l’ETL, nous avons tout fait pour optimiser l’extraction des données et leur transformation afin d’avoir le meilleur résultat possible. Au niveau des données, nous nous sommes assurés de l’intégrité de ces données mais aussi de leur importance. Nous n’avons pas choisi les accidents en France pour rien, c’est un sujet d’actualité qui nous intéressait et qui nous l’espérons vous intéressera aussi. Enfin, nos rapports sont les plus compréhensible possible afin d’être compris par n’importe qui.

## Les difficultés rencontrées

Au cours de ce projet nous avons rencontrés de nombreuses difficultés. En effet, la mise en place d’un datawarehouse implique l’utilisation d’outils que nous ne connaissons pas. Nous avons donc dû apprendre à utiliser Pentaho le plus rapidement possible.

De plus, nous avons dû faire face à des problèmes de liaison avec les données. En effet, nous souhaitons au départ utiliser une base de test sur un dépôt MS SQL Server. Le souci étant que la connexion vers SQL Server n’est pas simple et très peu fiable dû à l’intégration de la sécurité. Par conséquent, nous avons opté pour une solution plus simple : MySQL.

Enfin, nous avons rencontrés de nombreux problèmes quant aux jeux de données choisis et leur jointure. En effet, certains champs n’avaient pas le même format d’un jeu à l’autre. Nous avons dû ajouter au sein de l’ETL des étapes afin d’uniformiser l’ensemble des champs issus de nos trois jeux de données.