Projet Big Data Analytics : Analyse de la Clientèle d'un Concessionnaire Automobile pour la Recommandation de Modèles de Véhicules

(Par Nicolas PASQUIER pour la partie Analyse de Données,
Gabriel MOPOLO et Sergio SIMONIAN pour la partie Gestion des Données,
Marco WINKLER pour la partie Data Visualisation, deadline : 4 Janvier 2021)

1. Contexte et Objectifs

La réalisation de ce projet va permettre d'évaluer les cours de *Big Data* (Pr. Gabriel MOPOLO et Sergio SIMONIAN), de *Data Visualisation* (Pr. Marco WINCKLER), le cours *Data Analytics & Machine Learning* (Pr. Nicolas PASQUIER) et voir même le cours de programmation WEB/Mobile (G. GALLI, …).

Lire attentivement la section 5.

Contexte du Projet

Vous avez été contacté par un concessionnaire automobile afin de l'aider à mieux cibler les véhicules susceptibles d'intéresser ses clients. Pour cela il met à votre disposition :

- Son catalogue de véhicules
- Son fichier clients concernant les achats de l'année en cours
- Un accès à toutes les informations sur les immatriculations effectuées cette année
- Une brève documentation des données
- Un vendeur (voir son interview ci-dessous)

Votre client sera satisfait si vous lui proposez un moyen afin :

- Qu'un vendeur puisse en quelques secondes évaluer le type de véhicule le plus susceptible d'intéresser des clients qui se présentent dans la concession
- Qu'il puisse envoyer une documentation précise sur le véhicule le plus adéquat pour des clients sélectionnés par son service marketing (voir ci-dessous)

Documentation des Données

Les fichiers de données à votre disposition vous sont décrits dans les tables ci-dessous. Pour chaque attribut du fichier, vous sont donnés son nom, son type (numérique, caractères, catégoriel ou booléen) sa description et son domaine de valeurs.

Certains attributs peuvent comporter des valeurs manquantes ou incorrectes (erreur de saisie par exemple). Celles-ci sont représentées par une cellule vide ou bien contenant une valeur hors du domaine de valeurs de la variable (valeurs « ? », « » ou « N/D » par exemple).

Immatriculations.csv : informations sur les immatriculations effectuées cette année

Attribut	Type	Description	Domaine de valeurs
Immatriculation	caractères	Numéro unique d'immatriculation du véhicule	Texte au format « 9999 AA 99 »
Marque	caractères	Nom de la marque du véhicule	Audi, BMW, Dacia, Daihatsu, Fiat, Ford, Honda, Hyundaï, Jaguar, Kia, Lancia, Mercedes , Mini, Nissan, Peugeot, Renault, Saab, Seat, Skoda, Volkswagen, Volvo
Nom	caractères	Nom du modèle de véhicule	S80 T6, Touran 2.0 FSI, Polo 1.2 6V, New Beatle 1.8, Golf 2.0 FSI, Superb 2.8 V6, Toledo 1.6, 9.3 1.8T, Vel Satis

			3.5 V6, Megane 2.0 16V, Laguna 2.0T, Espace 2.0T, 1007 1.4, Primera 1.6, Maxima 3.0 V6, Almera 1.8, Copper 1.6 16V, S500, A200, Ypsilon 1.4 16V, Picanto 1.1, X-Type 2.5 V6, Matrix 1.6 FR-V 1.7, Mondeo 1.8, Croma 2.2, Cuore 1.0, Logan 1.6 MPI, M5, 120i, A3 2.0 FSI, A2 1.4
Puissance		Puissance en chevaux Din	[55, 507]
Longueur		Catégorie de longueur	courte, moyenne, longue, très longue
NbPlaces	numérique	Nombre de places	[5, 7]
NbPortes	numérique	Nombre de portes	[3, 5]
Couleur	catégoriel	Couleur	blanc, bleu, gris, noir, rouge
Occasion	booléen	Véhicule d'occasion ?	true, false
Prix	numérique	Prix de vente en euros	[7500, 101300]

Catalogue.csv : catalogue de véhicules

Attribut	Type	Description	Domaine de valeurs
Marque	caractères	Nom de la marque du véhicule	Audi, BMW, Dacia, Daihatsu, Fiat, Ford, Honda, Hyundaï, Jaguar, Kia, Lancia, Mercedes , Mini, Nissan, Peugeot, Renault, Saab, Seat, Skoda, Volkswagen, Volvo
Nom	caractères	Nom du modèle de véhicule	S80 T6, Touran 2.0 FSI, Polo 1.2 6V, New Beatle 1.8, Golf 2.0 FSI, Superb 2.8 V6, Toledo 1.6, 9.3 1.8T, Vel Satis 3.5 V6, Megane 2.0 16V, Laguna 2.0T, Espace 2.0T, 1007 1.4, Primera 1.6, Maxima 3.0 V6, Almera 1.8, Copper 1.6 16V, S500, A200, Ypsilon 1.4 16V, Picanto 1.1, X-Type 2.5 V6, Matrix 1.6 FR-V 1.7, Mondeo 1.8, Croma 2.2, Cuore 1.0, Logan 1.6 MPI, M5, 120i, A3 2.0 FSI, A2 1.4
Puissance	numérique	Puissance en chevaux Din	[55, 507]
Longueur	catégoriel	Catégorie de longueur	courte, moyenne, longue, très longue
NbPlaces	numérique	Nombre de places	[5, 7]
NbPortes	numérique	Nombre de portes	[3, 5]
Couleur	catégoriel	Couleur	blanc, bleu, gris, noir, rouge
Occasion	booléen	Véhicule d'occasion ?	true, false
Prix	numérique	Prix de vente en euros	[7500, 101300]

Clients_N.csv[1] : fichier clients concernant les achats de l'année en cours

Attribut	Type	Description	Domaine de valeurs
Age	numérique	Age en années du clients	[18, 84]
Sexe	catégoriel	Genre de la personne	M, F

1 Le numéro N (1, 2, 3, ...) du fichier Clients_N.csv à utiliser dépend de votre numéro de groupe.

Taux	numérique	Capacité d'endettement du client en euros (30% du salaire)	[544, 74185]
SituationFamiliale	catégoriel	Situation familiale du client	Célibataire, Divorcée, En Couple, Marié(e), Seul, Seule
NbEnfantsAcharge	numérique	Nombre d'enfants à charge	[0, 4]
2eme voiture	booléen	Le client possède déjà un véhicule principal ?	true, false
Immatriculation	caractères	Numéro unique d'immatriculation du véhicule	Texte au format « 9999 AA 99 »

Marketing.csv : clients sélectionnés par le service marketing

Attribut	Type	Description	Domaine de valeurs
Age	numérique	Age en années du clients	[18, 84]
Sexe	catégoriel	Genre de la personne	M, F
Taux	numérique	Capacité d'endettement du client en euros (30% du salaire)	[544, 74185]
SituationFamiliale	catégoriel	Situation familiale du client	Célibataire, Divorcée, En Couple, Marié(e), Seul, Seule
NbEnfantsAcharge	numérique	Nombre d'enfants à charge	[0, 4]
2eme voiture	booléen	Le client possède déjà un véhicule principal ?	true, false

Informations Données par le Concessionnaire

L'interview du gestionnaire de la concession automobile nous a permis de définir le contexte et les objectifs de l'application :

« Les différents véhicules de notre catalogue répondent à des besoins différents. Certains sont petits afin de mieux circuler en ville, d'autres ont de l'espace pour transporter toute une famille tandis que certains sont plus puissants et destinés à une clientèle plus fortunée. Nous souhaitons définir différentes catégories de véhicules afin de mieux comprendre les désirs des clients et proposer aux nouveaux clients le véhicule le plus adapté à leurs besoins. ».

La réalisation de ce projet va nécessiter la mise en œuvre des notions acquises durant les cours concernant les thématiques des Big Data, de la Data Visualisation et du Data Mining & Machine Learning.

2. Architecture du Projet Big Data Voitures et Activités Attendues par G. MOPOLO et B. RENAUT (à rendre 4 Janvier 2021)

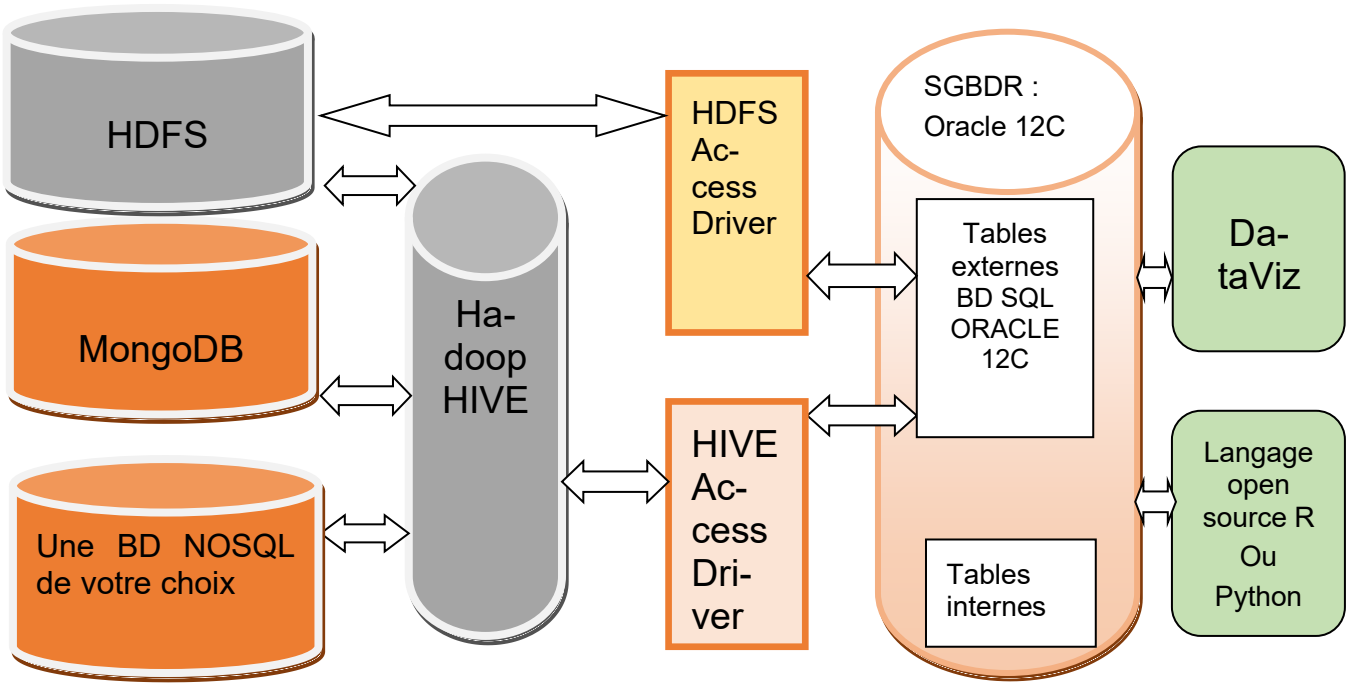
Le projet d'analyse de données peut se faire uniquement avec les fichiers CSV fournis. Toutefois, afin de valider l'écosystème BIG DATA Hadoop et la stratégie de construction de lacs de données, vous devez mettre en œuvre les deux architectures ci-dessous (pas d'accès aux fichiers CSV directement) :

- L'architecture DATA LAKE HADOOP qui consiste à s'appuyer les tables externes pour accéder aux données de sources hétérogènes (MongoDB, une 2ème BD NoSQL de votre choix, Hadoop HDFS, Oracle SQL). L'Accès aux données pour les parties Data Visualization et Data Mining & Machine Learning se fera via le langage SQL interrogeant des tables externes et internes
- L'architecture HDFS HADOOP qui consiste à accéder aux fichiers directement via le système de fichiers HADOOP. L'Accès aux données pour les parties Data Visualization et Data Mining & Machine Learning se fera en accédant directement aux fichiers.

Architecture 1 : DATA LAKE HADOOP avec tables externes

L'architecture DATA LAKE HADOOP qui consiste à s'appuyer les tables externes pour accéder aux données de sources hétérogènes (MongoDB, une 2ème BD NoSQL de votre choix, Hadoop HDFS, Oracle SQL).

L'Accès aux données pour les Data Visualization et Data Analysis with R se fera via le langage SQL interrogeant des tables externes et internes.



L'organisation des données se fera comme suit :

- Une ou plusieurs de vos sources de données devront être chargées sur la base NOSQL de votre choix
- Une ou plusieurs de vos sources de données devront être chargées sur la base MongoDB
- Une ou plusieurs de vos sources de données doivent être des fichiers hadoop HDFS
- Un de vos fichiers CLIENTS sera chargé sur la base Oracle SQL comme table interne.

Les données doivent être accessibles au niveau de la base SQL via des tables externes et internes. Vous répartissez vos données comme vous le souhaitez.

Le chargement des données dans les différentes bases de données doit se faire :

- Via des outils HADOOP tel que SQOOP ou des programmes java ou tout autre utilitaire (pour char-

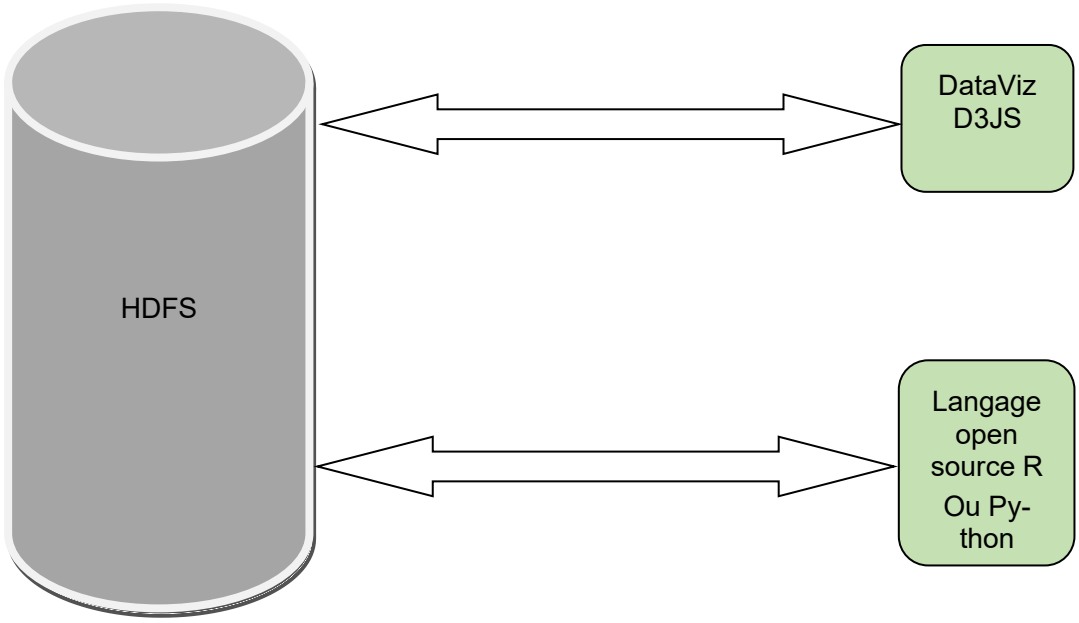
ger les données dans les bases NoSQL par exemple)
- Via Sqlloader ou autre (pour charger les données dans la base SQL)
- Il faut considérer les données chargées comme si c'étaient des données brutes.

L'accès aux données pour l'analyse avec R se fera via des requêtes SQL (cela concerne l'activité demandée par Nicolas PASQUIER et Marco WINCKLER).

Architecture 2 : HDFS HADOOP uniquement sans tables externes

L'architecture HDFS HADOOP qui consiste à accéder aux fichiers directement via le système de fichiers HADOOP HDFS.

L'Accès aux données pour la Data Visualization et la Data Analyse avec R se fera en accédant directement aux fichiers.



Vous devez charger tous vos fichiers dans hadoop hdfs et procéder à la visualisation et à la création de vos matrices en vue de l'analyse par la suite.

Travail à rendre sur l'ingénierie des données dans les architectures 1 et 2

- Un document contenant :
  - Le descriptif du projet.
  - L'architecture 1 DATA LAKE HADOOP avec la ventilation des données par base ou système.
  - L'architecture 2 HDFS HADOOP avec les données chargées.
- Scripts pour l'architecture 1
  - Le script de création de documents et de chargement de données dans MongoDB, le script de création des tables externes MongoDB dans HIVE et enfin le script de création des tables externes correspondantes dans Oracle NoSQL.
  - Le script de création de tables dans Oracle NoSQL, le script de création des tables externes Oracle NoSQL dans HIVE et enfin le script de création des tables externes correspondantes dans Oracle SQL
  - Le script de création des tables externes correspondantes Hadoop HDFS dans Oracle SQL.
  - Le programme java de chargement des données dans la base Oracle NoSQL.
  - Le script de chargement des données dans Hadoop HDFS.
  - Toute autre information justifiant de votre travail.
- Scripts pour l'architecture 2
  - Le script de chargement des données dans Hadoop HDFS.
  - Scripts de comparaison des performances entre l'Architecture 1 et l'Architecture 2. Montrer

les avantages et inconvénients de chacune de ces architectures.

HADOOP MAP REDUCE (cours Sergio SIMONIAN)

Histoire / Detail

Après avoir construit votre DATA LAKE le Concessionnaire vous appelle et vous fait part que certaines données étaient perdues avant votre intervention – notamment les détails sur l'émission CO2 / le coût d'énergie / la valeur de Bonus/Malus pour la taxation par marque et modelé de voiture. Il est possible que ses données seraient utiles pour améliorer la qualité de vos modelés prédictives. En cherchant sur Internet vous avez trouvé un fichier CO2.csv. C'est une autre base des données qui a certaines informations qui peuvent vous aider mais elle n'est pas parfaite. Elle ne contient pas tous les marques et modelés des voitures qui sont dans le catalogue du Concessionnaire. De plus le format de stockage est différent (la marque et le modelé sont dans une même colonne), il y a des valeurs manquant (colonne Bonus/Malus) et des valeurs erronés (colonne Bonus/Malus par exemple contient '-6 000€ 1' a la place de '-6 000€').

Le but est d'écrire un programme map/reduce avec Hadoop ou Spark qui va permettre d'adapter le fichier-CO2.csv pour intégrer ses informations complémentaires dans la ou les tables catalogue du Concessionnaire (ajouter des colonnes "Bonus / Malus", "Rejets CO2 g/km", "Cout Energie").

Notes :

- Les modelés des voitures du fichier CO2.csv n'ont pas beaucoup des valeurs en commun par rapport à la table catalogue – on voudrait utiliser une valeur moyenne d'émission CO2 (de même pour les autres colonnes : "Bonus / Malus", "Cout Energie") pour la marque de voiture concerne.
- Pour les marques de voitures qui ne sont pas dans le fichier CO2.csv on voudrait insérer la moyenne d'émission CO2 (de même pour les autres colonnes) de tous les marques de véhicules qui sont présent des deux côtés.
- Pour l'import / export des données vous pouvez utiliser des connecteurs Hadoop vu dans le cours 3 ou/et découvrir l'outil Hadoop Sqoop (http://sqoop.apache.org/) qui est projet Apache simplifiant cette tâche.
- Le fichier CO2.csv se trouve avec les autres ressources disponibles.

Travail à rendre :

- Code source du programme map/reduce utilise pour l'adaptation du fichier CO2.csv et son intégration dans la table catalogue.
- Un rapport au format PDF décrivant votre démarche pour l'adaptation du fichier CO2.csv et son intégration dans la table catalogue incluant les commandes / programmes que vous avez exécutés et leur résultat / output

3. Techniques de Data Visualisation et Activités Attendues par M. WINCKLER (à rendre 4 Janvier 2021)

Il faudrait analyser les données du domaine d'application et les catégoriser :

- Décrire la chaîne de traitement (« visualisation pipeline ») que, à partir de donnés bruts, permet de créer une représentation graphique et interactive avec l'ensemble de données.

- Décrire les utilisateurs visés ;

- Décrire les objectifs de visualisation et le tâches utilisateurs ;

- Développer de (au moins) trois techniques de visualisations avec l'API D3JS selon les contraintes suivantes :

    - Utiliser une base de données multidimensionnelles (minimum 5 attributs différents) ;
    - Prévoir de l'interaction avec l'ensemble des données (ex. navigation, sélection, filtres, etc.) ;
    - Deux niveaux de visualisation, c'est-à-dire une vision globale (« overview ») plus contexte ;
    - Donner la possibilité de charger des ensembles de données indépendants de l'application

- Faire un prototype exécutable avec D3JS qui démontre que l'utilisateur peut explorer l'ensemble des données pour atteindre les buts fixés.

4. Analyse des Données par les Techniques de Data Mining, Machine Learning et Deep Learning et Activités Attendues par N. PASQUIER et A. TEMIN (à rendre 4 Janvier 2021)

Voici la description d'un processus générique possible pour réaliser cette analyse. Ce processus peut être étendu et particularisé en utilisant d'autres étapes ou techniques, afin par exemple d'optimiser l'approche ou de vérifier la cohérence des résultats obtenus durant les différentes étapes par exemple.

1) Analyse exploratoire des données :

   L'analyse exploratoire des données vous permettra d'identifier d'éventuels problèmes dans les données (valeurs incohérentes, codage des valeurs manquantes, etc.) et découvrir d'éventuelles propriétés de l'espace des données (valeurs doublons, variables liées, variables d'importance particulière ou bien inutiles, etc.).

   Appliquez pour cela les différentes méthodes d'analyse exploratoire des données vues en cours (statistiques descriptives, histogrammes, nuages de points, boîtes à moustaches, etc.).

2) Identification des catégories de véhicules :

   Vous devez à partir des informations du *Catalogue* identifier des catégories de véhicules (citadine, routière, sportive, etc.) en fonction de leur taille, puissance, prix, etc. Ces catégories doivent correspondre à diverses besoins de la part des clients (une grande voiture pour les familles nombreuses, une petite voiture pour circuler en ville, etc.).

   Ces catégories de véhicules constitueront les classes à prédire durant les étapes suivantes du processus.

3) Application des catégories de véhicules définies au données des *Immatriculations :*

   Les données d'*Immatriculations* contiennent les informations sur les véhicules vendus cette année. L'objectif est d'attribuer à chacun de ces véhicules la catégorie qui lui correspond en utilisant le modèle définissant les catégories de véhicules généré précédemment.

4) Fusion des données *Clients* et *Immatriculations :*

   Les données *Clients* contiennent les informations sur les clients ayant les véhicules vendus cette année. L'objectif est de faire la fusion entre les données des *Clients* et des *Immatriculations* afin d'obtenir sur une même ligne l'ensemble des informations sur le client (âge, sexe, etc.) et sur le véhicule qu'il a acheté (avec sa catégorie).

   Cet ensemble de données servira lors des étapes suivantes pour l'apprentissage de la catégorie de véhicules (variable cible) la plus adaptée à un client selon ses caractéristiques (variables prédictives).

5) Création d'un modèle de classification supervisée pour la prédiction de la catégorie de véhicules :

   L'objectif de cette étape est de créer à partir du résultat de la fusion précédente un classifieur (modèle de classification supervisée) permettant d'associer aux caractéristiques des clients (âge, sexe, etc.) une catégorie de véhicules.

   Testez les différentes approches et algorithmes (arbres de décision, random forests, support vector machines, réseaux de neurones, deep learning, etc.), avec pour chaque algorithme plusieurs paramétrages testés, afin d'obtenir un classifieur aussi performent que possible.

   L'évaluation et la comparaison des performances de chaque configuration algorithmique (un algorithme et un paramétrage spécifiques) testée sera réalisée grâce aux matrices de confusion et mesures d'évaluation calculées à partir des résultats des tests des classifieurs.

6) Application du modèle de prédiction aux données *Marketing* :

   Les données *Marketing* contiennent les informations sur les clients pour lesquels on souhaite prédire une catégorie de véhicules.

   L'objectif est de prédire pour chacun de ces clients la catégorie de véhicules qui lui correspond le mieux en utilisant le classifieur généré durant l'étape précédente.

Travail à Rendre sur la Partie Data Mining, Machine Learning et Deep Learning

Vous devez déposer dans la boîte de dépôt du projet :

- Un rapport au format PDF décrivant les réalisations pour la gestion et l'analyse des données.
  Ce rapport doit décrire :

- Les choix effectués lors du projet en termes de gestion et d'analyse des données.
- Le(s) processus suivis.
- Les modèles de connaissances générés et l'interprétation de ces modèles.
- Les résultats que vous obtenez pour les clients sélectionnés par le service marketing.

- Les codes sources utilisés pour l'analyse des données et la génération des résultats que vous présentez dans le rapport.

  Cette partie doit contenir l'ensemble des scripts que vous avez créés pour analyser et générer les modèles de connaissances à partir des données.

5. Organisation du projet

Pour réaliser ce projet vous devez former une équipe de 4 étudiants donc 2 étudiants réaliseront l'architecture 1 et les autres travailleront sur la 2 ème architecture proposée. Les données de l'architecture 1 doivent être au final accessible via des tables externes ou internes y compris co2.csv. Pour l'architecture 1 les matrices d'analyses doivent être préparées avec sql.

Une comparaison des résultats des 2 sous-groupes est à faire.