

Livrable 1



Antoine Sire (chef de projet)

Yann Subts

Tayana Petro

CONTEXTE

Pour continuer dans l'amélioration des performances et la qualité des soins, les praticiens (médecin, personnel infirmier) et administrateurs doivent pouvoir accéder directement aux informations exploitables dans les données médicales.

Il est donc impératif de s'occuper d'informations relatives inexploitées, en effectuant des analyses basées sur des données agrégées, consolidées, historiques et synthétisées.

Le secteur doit s'investir dans le développement des systèmes informatiques évolutifs, qui comprennent un ensemble d'outils et de mécanismes pour charger, extraire et traiter les données médicalement.

Le groupe CHU (Cloud Healthcare Unit) a ainsi pris conscience de l'intérêt, voire de la nécessité d'une transformation digitale majeure.

Notre service est donc sollicité pour l'aider à mettre en place son propre entrepôt de données en répondant à des besoins et exigences d'accès et d'analyses des utilisateurs.

TABLE DES MATIERES

Contexte	2
Introduction	4
Modèle conceptuel des données	5
Développement des jobs d'alimentation	7
Architecture de l'entrepôt de données	10
Conclusion	12
Bilans personnels	13
Webographie	14

INTRODUCTION

Afin de permettre l'exploitation des données générées par les systèmes de gestion de soins, des grandes phases envisagées du projet sont à appliquer.

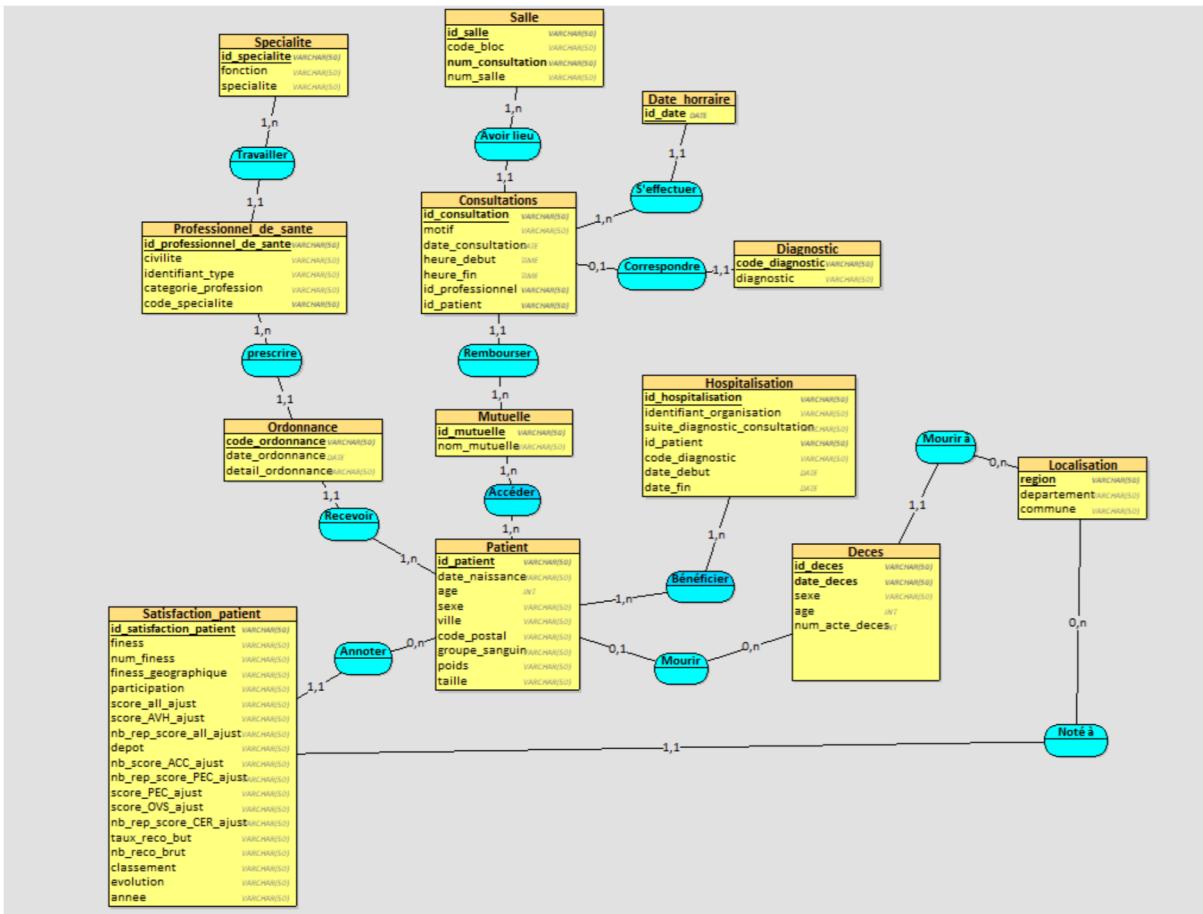
Avec à disposition un environnement virtualisé et des outils nécessaires pour les traitements, nous devons commencer par le référentiel de données dans lequel les différentes missions sont :

- La modélisation des différents axes d'analyse ainsi que les mesures,
- Le développement des jobs d'alimentation du schéma décisionnel,
- La description de l'architecture de l'entrepôt de données.

Nous allons donc présenter ces différentes étapes qui nous aideront à étudier en profondeur les données importantes ainsi que l'étude de l'architecture à mettre en place.

MODELE CONCEPTUEL DES DONNEES

Un modèle conceptuel de données (MCD) est une [représentation graphique](#) de haut niveau qui permet d'écrire des données utilisées par le système d'information, et de le décrire à l'aide d'objets définis nommés « entité ».



Le modèle conceptuel des données ci-dessus contient des entités (ex : patient) dans lesquelles se trouvent des informations concernant cette entité (date de naissance, age...) qui vont permettre de définir nos différentes données au sein de notre BDD.

Des liaisons permettent de renforcer les tables avec des cardinalités représentant le nombre maximum et minimum de possibilités que chaque classe contient dans la relation liant les objets entre eux.

Par exemple, pour la classe ordonnance, un professionnel de santé ou plusieurs (1,n) ne peuvent prescrire que une et une seule même ordonnance à un patient (1,1).

Ainsi, nous avons cinq dimensions :

- Localisation, qui concerne l'établissement de santé
- Patient, équipé des informations concernant l'individu patient
- Consultation, qui contient les informations relatives aux consultations
- Temps, concernant les dates et heures importantes (consultation, hospitalisation...)
- Professionnel de santé (médecin...)

DEVELOPPEMENT DES JOBS D'ALIMENTATION

L'alimentation d'entrepôts de données implique l'extraction et la migration de données de notre base de données vers un ou plusieurs systèmes afin d'être analysés.

Ici, nous avons créé des jobs d'alimentation qui sont des représentations graphiques d'un ou plusieurs composants reliés entre eux. Ils regroupent un ensemble de tâches et permettent d'exécuter des processus de flux de données.

Sur l'outil de conception visuelle de flux de traitement de données Talend, nous avons créé les différents jobs qui correspondent aux entités qui ont été rentrées précédemment dans notre MCD.

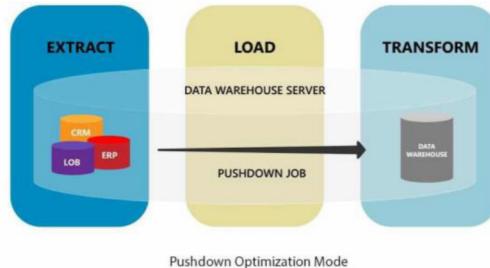
On commence par alimenter en récupérant les données dans les clusters HDFS, qui correspondent à des systèmes de fichiers distribués dans lesquels sont traitées les données.

Ainsi, on obtient des métadonnées qui étaient anciennement des fichiers csv.

Pour ce faire, on effectue un premier job GetData où on a pris chaque source de données à notre disposition.



Cette phase correspond donc à l'Extract de la technologie informatique ELT qui correspond à cette illustration :



Après avoir configuré les paramètres et sélectionné le fichier d'entrée, les données lues ont été envoyées vers une sortie standard puis le composant **tLogRow** nous a permis d'afficher les résultats dans la console Run.

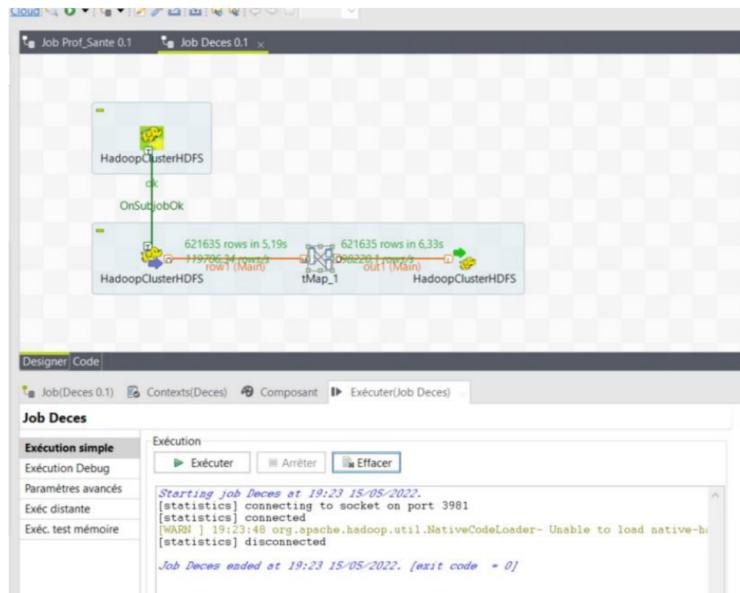
Voici le résultat du composant **tLogRow** pour le job de récupération des salles.

201707 1059225112 Bloc-F Etagé-2 F-2
201708 1059225113 Bloc-B Etagé-2 B-2
201709 1059225114 Bloc-F Etagé-3 F-3
201710 1059225115 Bloc-C Etagé-0 C-0
201711 1059225116 Bloc-D Etagé-0 D-0
201712 1059225117 Bloc-E Etagé-2 E-2
201713 1059225118 Bloc-A Etagé-2 A-2
201714 1059225119 Bloc-D Etagé-1 D-1
201715 1059225120 Bloc-A Etagé-3 A-3
201716 1059225121 Bloc-B Etagé-0 B-0
201717 1059225122 Bloc-B Etagé-0 B-0
201718 1059225123 Bloc-F Etagé-0 F-0
201719 1059225124 Bloc-E Etagé-1 E-1
201720 1059225125 Bloc-F Etagé-2 F-2
201721 1059225126 Bloc-E Etagé-3 E-3
201722 1059225127 Bloc-A Etagé-0 A-0
201723 1059225128 Bloc-B Etagé-3 B-3
201724 1059225129 Bloc-C Etagé-3 C-3
201725 1059225130 Bloc-B Etagé-0 B-0
201726 1059225131 Bloc-A Etagé-0 A-0
201727 1059225132 Bloc-D Etagé-0 D-0
201728 1059225133 Bloc-D Etagé-1 D-1
201729 1059225134 Bloc-A Etagé-2 A-2
201730 1059225135 Bloc-A Etagé-1 A-1
201731 1059225136 Bloc-D Etagé-0 D-0
201732 1059225137 Bloc-F Etagé-3 F-3
201733 1059225138 Bloc-E Etagé-2 E-2
201734 1059225139 Bloc-B Etagé-3 B-3

[statistics] disconnected
Job Salle ended at 22:46 15/05/2022. [exit code = 0]

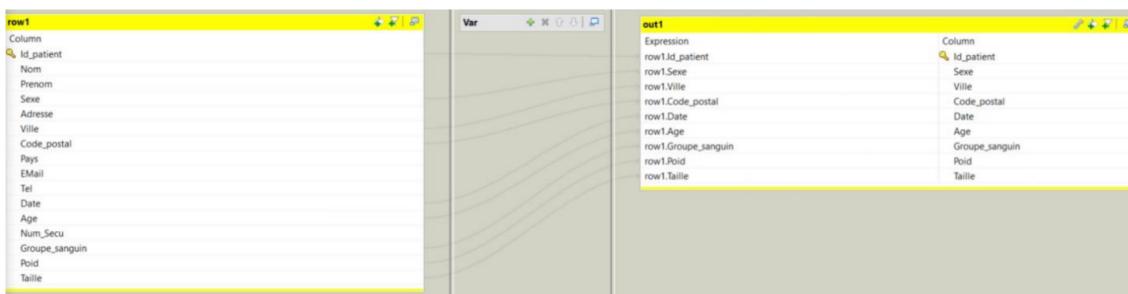
En exécutant un job, le contenu du fichier apparaît dans la console de la vue Run comme vous pouvez constater ci-dessous.

Exemple création job « décès »



Pour enrichir notre job et récupérer uniquement les données qui nous intéressent, nous avons utilisé le composant **tMap** qui nous permet de faire la sélection des données et de gérer les entrées et sorties multiples. Grâce au composant **tMap**, on peut optimiser les temps de traitement entre les données.

Cette étape correspond à la phase « Transform » de la technologie ELT.



A gauche, on retrouve les informations concernant le patient, et à droite se trouvent les informations transformées à utiliser.

Ainsi, les jobs d'alimentation nous permettront d'alimenter notre Data Warehouse.

ARCHITECTURE DE L'ENTREPOT DE DONNEES

Un entrepôt de données est un référentiel centralisé utilisé pour collecter, ordonner et stocker des informations provenant de bases de données opérationnelles organisées en tableaux et colonnes.

L'architecture va permettre d'illustrer la structure générale de son fonctionnement.

On y retrouve :

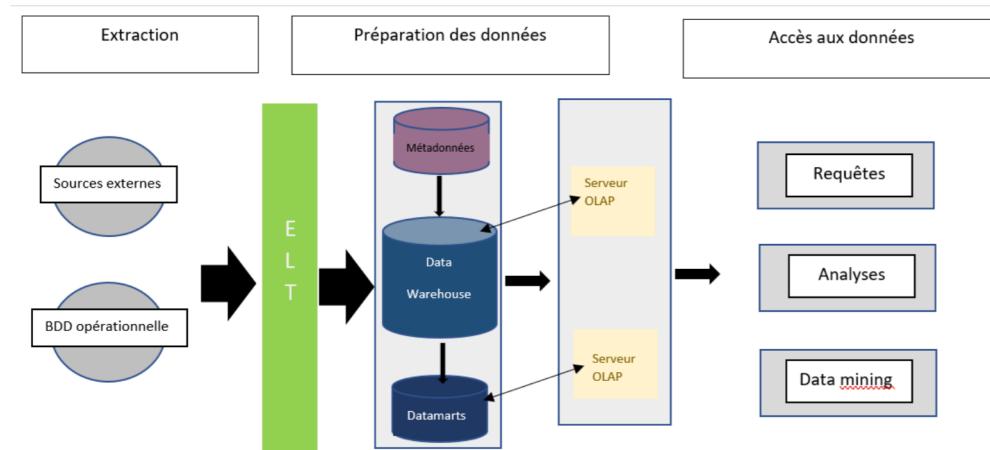
Le processus ELT qui est un concept fondamental de l'entrepôt de données dans lequel les données brutes sont prises en charge, extraites pour l'analyse et transformées en un format qui peut répondre aux besoins opérationnels et chargés dans un Data Warehouse.

ELT signifie :

- Extraction : Les données brutes sont extraites depuis les différentes sources
- Charger (Load) : Les données transformées sont chargées et transférées vers un data warehouse
- Transformer : Les données sont converties en données propres, structurées et vérifiées prêtes à l'emploi

Un **Data Warehouse** est une base de données relationnelle hébergée sur un serveur dans un Data Center ou dans le Cloud. Il recueille des données de sources variées et hétérogènes qui permettent de soutenir l'analyse et faciliter la prise de décision.

Architecture d'entrepôt de données



En première et deuxième partie se trouvent les données issues de différentes sources données qui sont extraites, nettoyées et intégrées dans l'entrepôt de données.

Durant la troisième phase, un serveur OLAP présente les informations demandées par les utilisateurs sous plusieurs formes comme des analyses.

Les métadonnées définissent l'ensemble des données et connexions traitées dans lesquelles se trouvent les informations concernant les fichiers, les bases de données et les systèmes dont on a besoin pour créer nos jobs.

Ces référentiels centraux sont utilisés pour enregistrer les informations sur les patients des différentes unités du domaine médical. On y inclut les données personnelles des patients consolidées dans l'entrepôt de données et connecté via le schéma de la base de données.

L'utilisation de ce type d'architecture permet ainsi une meilleure prise de décision, une analyse des données historiques, des données consolidées provenant de sources différentes, mais aussi de gagner en qualité, cohérence, et précision des données.

CONCLUSION

Après avoir mis en place un modèle conceptuel des données, ainsi qu'une architecture d'entrepôt et un développement des jobs d'alimentations, nous avons pu aboutir à une solution adéquate qui nous permettra ultérieurement de répondre aux besoins et attentes précises de ce projet.

La prochaine étape se base sur une évaluation essentielle des performances par rapport aux différentes tables avec des scripts et des vérifications.

BILANS PERSONNELS

Bilan chef de projet : L'ensemble des tâches a été réparti de manière cohérente. Chaque membre du groupe a permis le bon déroulement de ce premier livrable. Pour le livrable 2, nous continuerons ainsi dans cette lancée afin de ne pas être débordés par la charge de travail.

Antoine SIRE : Pour le premier livrable de ce projet, j'avais la charge de la réalisation des jobs d'alimentation et du MCD.

Yann SUBTS : Lors de ce livrable, je me suis occupé des jobs d'alimentation et au MCD.

Tayana PETRO : Je me suis tournée vers une version du MCD et l'architecture puis j'ai participé aux jobs.

WEBOGRAPHIE

1. Ressources des prosits
2. Corbeilles et Workshops
3. <https://cesi.moodle.fr>
4. [Talend](#)