

Projet BIG DATA Livrable 1

COURTIES Baptiste

TRONEL-PEYROZ Baptiste

CAVAGNE Théo

SADAoui Eliote

DELPECH Romain

Table des matières

Livrable 1 – Référentiel de données	4
Partie 1 – Introduction et contexte	4
1. Introduction	4
2. Contexte et enjeux.....	4
3. Objectifs du projet.....	5
4. Périmètre du livrable 1	5
Partie 2 – Description des sources de données	6
2.1 Base de données PostgreSQL	6
2.2 Données CSV – Sources externes	7
2.3 Données de satisfaction hospitalière	9
2.4 Synthèse des sources de données	10
2.5 Conclusion	10
Partie 3 – Architecture.....	12
3.1 Schémas d'architecture	12
3.2 Type d'architecture	12
3.3 Source de données	13
3.4 Environnement Big Data.....	13
3.5 Dataviz	13
Partie 4 – MCD et Jobs.....	14
4.1 Analyse des besoins :	14
4.1.1. Domaine des consultations	14
4.1.2. Domaine des hospitalisations	15
4.1.3. Domaine de la mortalité	16
4.1.4. Domaine de la satisfaction.....	16
4.1.5. Synthèse générale	17
4.1.6. Conclusion	18
4.2 Création du MCD	18
4.2 Jobs.....	19
Conclusion	22
Table des figures	23

Livrable 1 – Référentiel de données

Partie 1 – Introduction et contexte

1. Introduction

Le secteur de la santé est aujourd’hui confronté à une transformation numérique majeure, portée par l’explosion des données médicales et administratives. Ces données représentent une source d’information précieuse pour améliorer la qualité des soins, optimiser la gestion hospitalière et renforcer la prise de décision. Cependant, leur volume, leur diversité et leur sensibilité rendent leur exploitation complexe.

Dans ce contexte, le groupe Cloud Healthcare Unit (CHU) souhaite développer une solution décisionnelle centralisée lui permettant de collecter, consolider, stocker et analyser les informations issues de ses différents systèmes de gestion médicale et administrative. L’objectif est de disposer d’un entrepôt de données (Data Warehouse) capable de fournir des analyses fiables, sécurisées et pertinentes à long terme.

2. Contexte et enjeux

Actuellement, les données du CHU sont dispersées dans plusieurs systèmes indépendants :

- Une base de données de gestion des soins.
- Des fichiers CSV décrivant les établissements hospitaliers.
- Des fichiers plats FTP contenant des notes de satisfaction et le répertoire des décès.

Cette fragmentation empêche les utilisateurs (praticiens, responsables médicaux, administrateurs) d’accéder facilement à des informations consolidées. Elle rend également difficile toute analyse globale, qu’il s’agisse du suivi de la fréquentation, de la performance des services ou de la satisfaction des patients.

La mise en place d'un entrepôt de données permettra :

- D'intégrer et centraliser ces données hétérogènes dans une base unifiée.
- De fournir des indicateurs fiables pour la prise de décision.
- D'améliorer la qualité des soins par une meilleure connaissance du parcours patient.
- D'assurer la traçabilité et la sécurité des données sensibles conformément au RGPD.

3. Objectifs du projet

Le projet Cloud Healthcare Unit vise à concevoir et à déployer une infrastructure décisionnelle complète permettant de :

1. Collecter et intégrer les données provenant des différentes sources (bases, fichiers, FTP).
2. Modéliser et structurer ces données selon une architecture décisionnelle adaptée (Data Warehouse / Datamart).
3. Mettre à disposition des tableaux de bord et indicateurs répondant aux besoins métiers : taux de consultation, hospitalisation, satisfaction et mortalité.
4. Garantir la sécurité, la confidentialité et l'évolutivité de la plateforme.

4. Périmètre du livrable 1

Ce premier livrable porte sur la phase de conception du projet. Il comprend :

- La modélisation conceptuelle des données (MCD).
- La définition du modèle décisionnel.
- La description des flux d'alimentation ETL.
- La proposition d'architecture technique adaptée.

Ce référentiel servira de fondation pour la mise en œuvre physique du système dans les livrables suivants.

Partie 2 – Description des sources de données

L'ensemble du projet repose sur un écosystème de données hétérogènes, combinant des bases relationnelles (PostgreSQL), des fichiers plats (CSV) et des données (XLSX) issues de la satisfaction hospitalière. Ces différentes sources constituent le socle du futur entrepôt de données du Centre Hospitalier Universitaire (CHU).

Afin d'avoir une vision complète et détaillée de la structure des différentes sources de données, un document visuel a été élaboré sous Draw.io. Ce fichier présente l'ensemble des tables issues de la base PostgreSQL ainsi que les fichiers CSV externes, avec pour chacun :

- Les champs disponibles.
- Les types de données et identifiants.
-  [Lien](#) vers le descriptif complet des tables et fichiers (Draw.io).

Ce référentiel visuel sert de support de référence pour la phase de conception et garantit une compréhension globale et homogène des données du CHU avant leur intégration dans l'entrepôt décisionnel.

2.1 Base de données PostgreSQL

La base PostgreSQL regroupe les principales informations hospitalières structurées. Elle contient plusieurs tables interrelées autour du patient, des consultations, des diagnostics et des professionnels de santé.

Tables principales :

Table	Description
Patient	Contient les informations personnelles et administratives des patients (nom, prénom, sexe, adresse, âge, numéro de sécurité sociale, groupe sanguin, taille, poids, etc.).
Consultation	Regroupe les données relatives aux consultations : identifiant patient, professionnel de santé, mutuelle, motif, code diagnostic et date de consultation.
Diagnostic	Liste les diagnostics avec leur code et libellé descriptif.

Table	Description
Prescription	Fait le lien entre une consultation et les médicaments prescrits via le code CIS.
Medicaments	Contient les informations réglementaires et pharmaceutiques sur les médicaments : dénomination, forme, voie d'administration, statut de commercialisation, titulaire, etc.
Professionnel_de_sante	Données sur les praticiens (nom, prénom, civilité, spécialité, profession, type d'identifiant, etc.).
Specialite	Répertorie les spécialités médicales et leur code associé.
Mutuelle	Liste les mutuelles partenaires avec leur nom et adresse.
Adher	Fait le lien entre un patient et sa mutuelle.
Salle	Informations relatives aux salles et blocs où se déroulent les consultations (numéro, étage, code bloc, etc.).
Laboratoire	Informations sur les laboratoires associés à certains diagnostics.
AAAA / Date	Tables de référence mineures, servant au stockage d'identifiants, de dates ou de codes internes.

Cette base constitue la source principale des données médicales et administratives. Elle permettra d'assurer la cohérence entre les différentes dimensions (patient, consultation, diagnostic, professionnel, médicament...).

2.2 Données CSV – Sources externes

Plusieurs fichiers CSV viennent enrichir le périmètre de la base relationnelle. Ils apportent des données complémentaires sur les décès, les établissements de santé et les hospitalisations.

a) Données de décès

Fichier	Colonnes principales
Deces.csv	nom, prénom, sexe, date_naissance, code_lieu_naissance, lieu_naissance, pays_naissance, date_deces, code_lieu_deces, numéro_acte_deces

Description :

Ces données proviennent du répertoire national des décès. Elles permettent de suivre les statistiques de mortalité et de croiser les informations avec les patients de la base principale afin d'étudier la survie et la prise en charge médicale.

b) Données des établissements de santé

Fichier	Colonnes principales
Etablissement_sante.csv	adresse, code_postal, commune, email, raison_sociale, identifiant_organisation, siret_site, téléphone, type_voie, voie, pays
Activite_professionnel_sante.csv	identifiant, civilité, catégorie_professionnelle, nom, prénom, commune, profession, spécialité, type_identifiant

Description :

Ces fichiers contiennent des informations sur les structures de soins et les professionnels affiliés.

Ils permettent d'identifier les établissements partenaires du CHU, les lieux de consultation externes, et les professionnels de santé intervenants.

c) Données d'hospitalisation

Fichier	Colonnes principales
Hospitalisation.csv	Num_Hospitalisation, Id_patient, identifiant_organisation, Code_diagnostic, Suite_diagnostic_consultation, Date_Entree, Jour_Hospitalisation

Description :

Ce fichier retrace les séjours hospitaliers : durée, diagnostic associé, et établissement d'accueil.

Il sert à analyser l'activité hospitalière et à relier les hospitalisations aux diagnostics et consultations précédentes.

2.3 Données de satisfaction hospitalière

La base “Satisfaction” regroupe un volume important de fichiers CSV contenant des enquêtes de satisfaction patients collectées dans différents services hospitaliers. Chaque fichier correspond à une enquête ou un indicateur de qualité.

Contenu général :

- Identifiants anonymisés de patients ou d'enquêtes
- Notes globales et par critère (accueil, soins, propreté, restauration, écoute, information médicale, etc.)
- Commentaires textuels libres
- Dates et lieux d'enquête
- Données de contexte (type de service, durée du séjour, motif de passage, etc.)

Description synthétique :

Ces données sont destinées à l'évaluation qualitative des services et à l'analyse du ressenti patient.

Elles présentent une grande volumétrie et une forte granularité. Seules les variables pertinentes seront retenues lors de la phase d'intégration (note globale, service, date, durée du séjour...).

2.4 Synthèse des sources de données

Type de source	Format	Nombre estimé de fichiers	Thématique principale
Base PostgreSQL	Relationnelle (SQL)	12 tables	Données médicales et administratives
Fichiers CSV – Décès	CSV	1	Statistiques de mortalité
Fichiers CSV Établissements de santé	CSV	2	Structure et professionnels de santé
Fichiers CSV Hospitalisation	CSV	1	Suivi des séjours hospitaliers
Fichiers CSV – Satisfaction	CSV multiples	~10+	Enquêtes patients et indicateurs de qualité

2.5 Conclusion

L'ensemble de ces sources forme une base de connaissance riche, mêlant données relationnelles, administratives et qualitatives. Elles constitueront le socle du projet d'entrepôt de données du CHU, permettant d'analyser :

- L'activité médicale.
- La performance hospitalière.
- La satisfaction patient.
- L'amélioration continue de la qualité des soins.

Partie 3 – Architecture

3.1 Schémas d'architecture

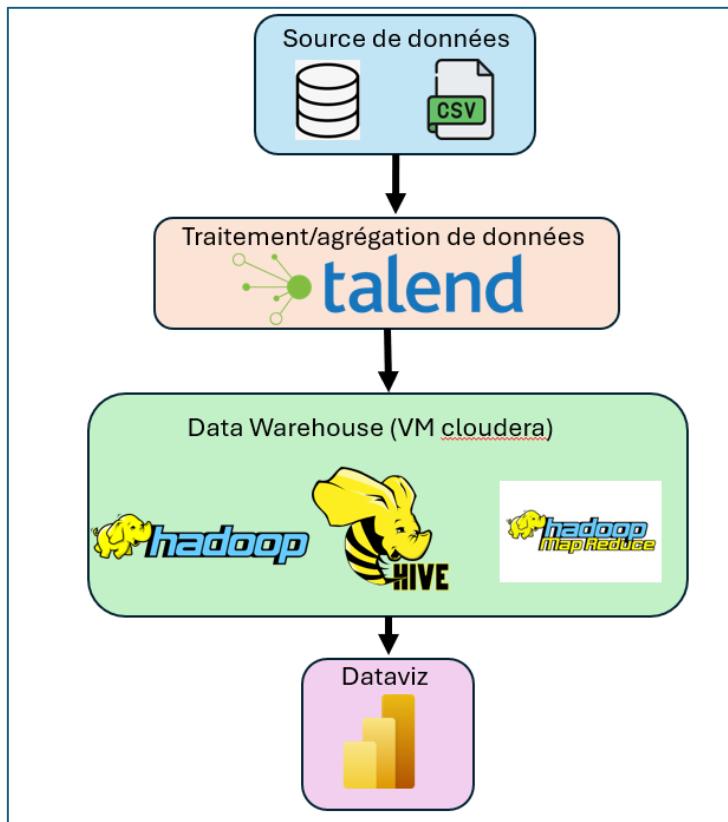


Figure 1 Schéma de l'architecture

3.2 Type d'architecture

L'architecture mise en place s'appuie sur une approche ETL (Extract – Transform – Load), c'est-à-dire un processus d'extraction, de transformation et de chargement des données. Cette méthode a été choisie car nous traitons des données personnelles, ce qui nécessite un contrôle rigoureux sur chaque étape du traitement afin de garantir la conformité et la sécurité des informations, notamment vis-à-vis du RGPD.

3.3 Source de données

Les sources de données utilisées sont variées : elles proviennent à la fois de fichiers CSV contenant des données brutes exportées depuis différents systèmes, et d'une base de données relationnelle utilisée pour compléter et enrichir ces informations. Ces données sont ensuite traitées dans un processus ETL orchestré par Talend. Cet outil permet d'automatiser l'extraction des données depuis les différentes sources, de les transformer (par des opérations de nettoyage, de filtrage, d'agrégation ou encore d'anonymisation), puis de les charger dans l'entrepôt de données.

3.4 Environnement Big Data

Une partie de ces traitements est exécutée au sein de la plateforme Cloudera, afin de profiter de son environnement Big Data basé sur Hadoop. Cette plateforme fournit une infrastructure distribuée et performante pour le stockage et le traitement de gros volumes de données.

Le Data Warehouse repose donc sur plusieurs composants de l'écosystème Hadoop :

- HDFS (Hadoop Distributed File System) pour le stockage distribué,
- MapReduce pour le traitement parallèle des données,
- Hive pour la création de tables logiques et l'interrogation des données via un langage proche du SQL.

Ainsi, les données nettoyées et consolidées sont centralisées et stockées dans Cloudera, ce qui permet une gestion efficace, scalable et sécurisée de l'ensemble du patrimoine de données.

3.5 Dataviz

Enfin, la couche de datavisualisation est assurée par Power BI. Cet outil permet de se connecter directement à l'entrepôt de données pour produire des tableaux de bord interactifs et des rapports analytiques. Il facilite l'exploration visuelle des données et aide les utilisateurs à mieux comprendre les résultats issus du traitement.

Partie 4 – MCD et Jobs

4.1 Analyse des besoins :

L'étude des besoins exprimés permet de dégager plusieurs axes d'analyse et d'identifier les mesures à calculer dans le futur entrepôt de données.

L'objectif global est de suivre l'activité des patients et des établissements de santé, selon différents points de vue : la fréquence des consultations, le taux d'hospitalisation, la répartition des diagnostics, la satisfaction des usagers et la mortalité.

Rappel des besoins :

- Taux de consultation des patients dans un établissement X sur une période de temps Y.
- Taux de consultation des patients par rapport à un diagnostic X sur une période de temps Y.
- Taux global d'hospitalisation des patients dans une période donnée Y.
- Taux d'hospitalisation des patients par rapport à des diagnostics sur une période donnée.
- Taux d'hospitalisation par sexe, par âge.
- Taux de consultation par professionnel.
- Nombre de décès par localisation (région) et sur l'année 2019.
- Taux global de satisfaction par région sur l'année 2020.

4.1.1. Domaine des consultations

Les premiers besoins portent sur le taux de consultation des patients, soit selon certains critères comme le diagnostic, le professionnel de santé ou l'établissement.

Ces indicateurs permettent d'évaluer l'activité médicale, la charge de travail des professionnels et la répartition des actes selon les pathologies.

Les mesures principales associées à ce domaine sont :

- Le nombre total de consultations réalisées sur une période donnée.
- Le taux de consultation, c'est-à-dire le rapport entre le nombre de consultations et le nombre total de patients suivis.

Pour ces analyses, les axes d'étude nécessaires sont :

- La dimension temporelle, pour suivre l'évolution des consultations dans le temps (jour, mois, année).
- La dimension établissement, afin d'observer les différences entre structures.
- La dimension patient, qui permet de caractériser la population (âge, sexe...).
- La dimension professionnel de santé, pour mesurer l'activité individuelle ou par spécialité.
- La dimension diagnostic, utile pour distinguer les pathologies les plus consultées.

4.1.2. Domaine des hospitalisations

Le second ensemble d'indicateurs concerne les taux d'hospitalisation, aussi bien globalement que par pathologie, par sexe ou par âge.

Ces besoins visent à mesurer la charge hospitalière et à identifier les groupes de patients les plus concernés par certaines maladies ou traitements.

La mesure correspondante est :

- La durée du séjour.

Les dimensions associées sont similaires à celles des consultations :

- La dimension diagnostic, afin d'étudier les hospitalisations selon la cause médicale.
- La dimension temporelle, pour suivre l'évolution des consultations dans le temps (jour, mois, année).
- La dimension patient, qui permet de caractériser la population (âge, sexe...).

Ces analyses combinent généralement le temps, l'âge, le sexe, et la pathologie comme principaux axes d'exploration.

4.1.3. Domaine de la mortalité

Le besoin exprimé autour du nombre de décès par localisation correspond à une analyse de la mortalité par région et par période (par exemple sur l'année 2019).

Cette analyse permet de suivre les tendances régionales et temporelles de la mortalité, en lien éventuel avec les hospitalisations ou les pathologies traitées.

La mesure pertinente est le nombre de décès.

Les dimensions nécessaires sont :

- La dimension région, pour la localisation géographique.
- La dimension temps, pour l'évolution annuelle.
- La dimension patient.

4.1.4. Domaine de la satisfaction

Enfin, le besoin relatif au taux global de satisfaction par région vise à mesurer la qualité perçue des soins à travers les enquêtes ou indicateurs recueillis.

L'objectif est de comparer les performances entre régions ou établissements et de suivre leur évolution dans le temps.

Les mesures associées sont :

- Le score moyen de satisfaction.
- Le taux de réponse (proportion de patients ayant participé à l'enquête).

Les axes d'analyse nécessaires sont :

- Le temps.
- La région et l'établissement, pour localiser les résultats.

4.1.5. Synthèse générale

L'ensemble de ces besoins montre qu'il est nécessaire de concevoir plusieurs tables de faits, chacune correspondant à un domaine d'activité :

- Une table F_Consultation pour l'activité médicale courante.
- Une table F_Hospitalisation pour les séjours en établissement.
- Une table F_Décès pour les données de mortalité.
- Une table F_Satisfaction pour les indicateurs de qualité perçue.

Ces tables de faits seront reliées à un ensemble commun de dimensions :

- D_Date, pour structurer l'analyse temporelle.
- D_Etablissement, pour décrire le contexte organisationnel.
- D_Patient, pour représenter les caractéristiques démographiques.
- D_Diagnostic, pour identifier les pathologies.
- D_ProfessionnelSante, pour les analyses par acteur.
- D_Région, pour la localisation géographique.

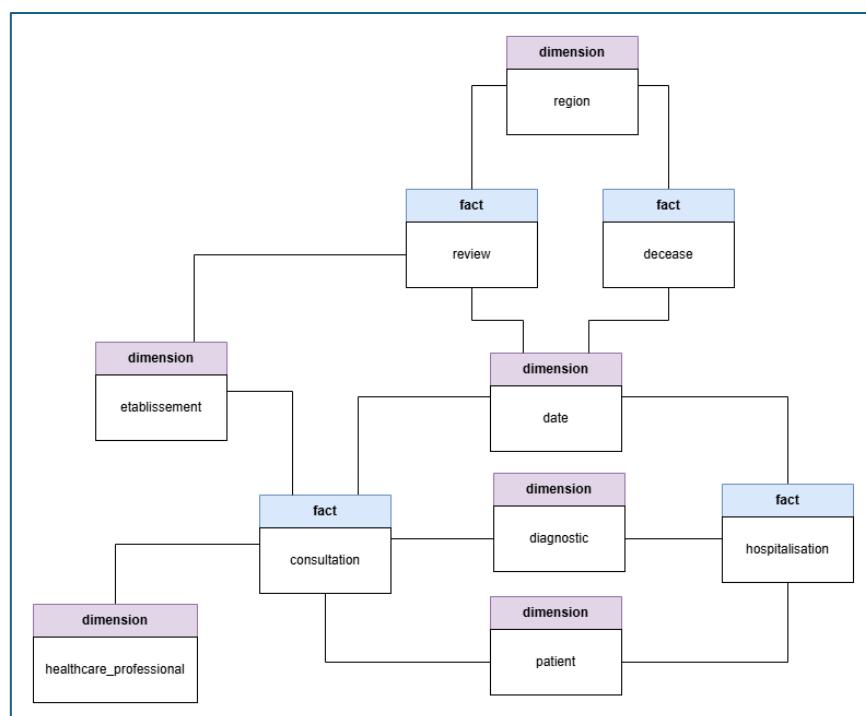


Figure 2 MCD simplifié

4.1.6. Conclusion

L'analyse des besoins montre que le système décisionnel doit permettre d'étudier à la fois l'activité médicale, la performance des établissements, les profils de patients et la qualité des soins.

Les mesures principales portent sur les volumes d'actes (consultations, hospitalisations, décès), leurs taux relatifs, ainsi que sur la satisfaction des patients.

Les dimensions identifiées (temps, établissement, patient, diagnostic, région, professionnel) offriront une grande souplesse d'analyse pour croiser ces indicateurs selon les différents contextes temporels et géographiques.

4.2 Création du MCD

Grâce au travail d'analyse du besoin et d'analyse des sources de données, nous avons pu créer un modèle conceptuel de données regroupant et normalisant les informations sources dans différentes tables de faits (f_) et de dimensions (d_).

Voici notre MCD, reprenant les informations que l'on a qualifié de plus importantes pour la structure de notre base de données.

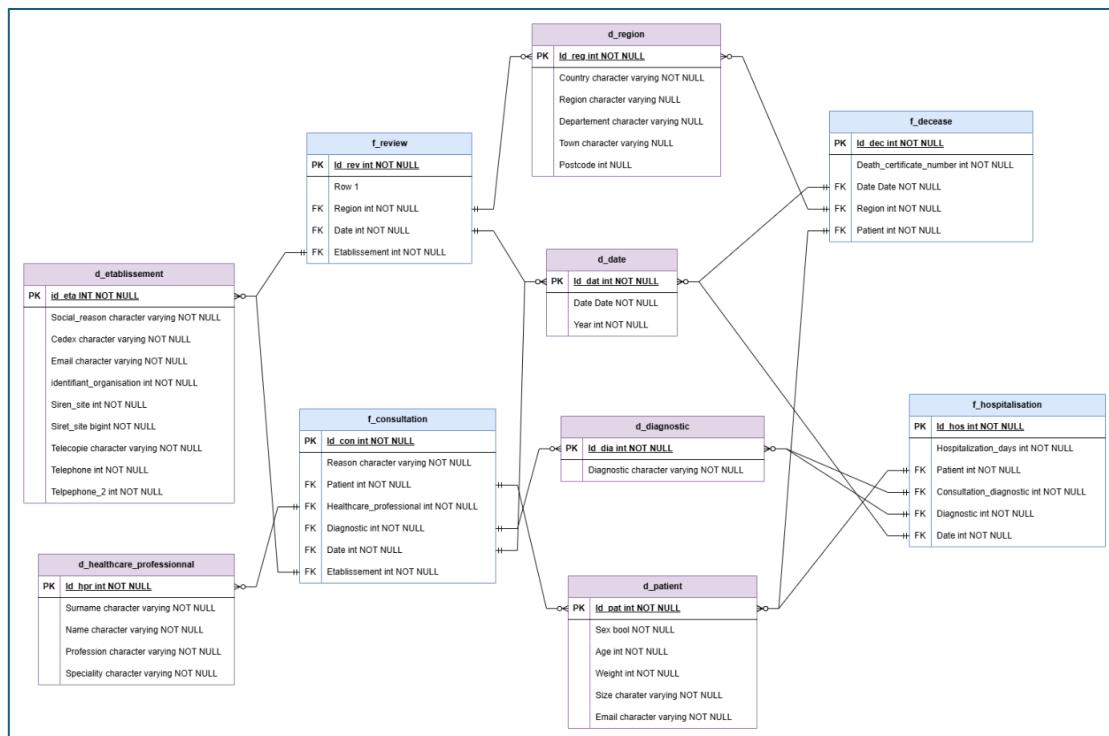


Figure 3 MCD complet

On remarquera que nous avons pris la décision de simplifier quelques relations qui étaient présentes dans les données sources, notamment la relation entre un professionnel de santé et un établissement. Nous avons fait ce choix car, du point de vue des besoins exprimés par le corps médical (voir [4.1 Analyse des besoins](#)), il n'est pas nécessaire de faire apparaître cette relation dans l'entrepôt de données.

4.2 Jobs

Le Job Talend ci-dessous a pour objectif de nettoyer et d'harmoniser les champs de dates issus du fichier « deces.csv », afin de garantir la cohérence des formats avant leur intégration dans l'entrepôt de données.

Cette étape de préparation est indispensable, car les fichiers sources peuvent contenir des formats de dates hétérogènes, comme dd/MM/yyyy, yyyy-MM-dd, dd-MM-yy... Ce qui empêche toute agrégation temporelle fiable lors de l'analyse.

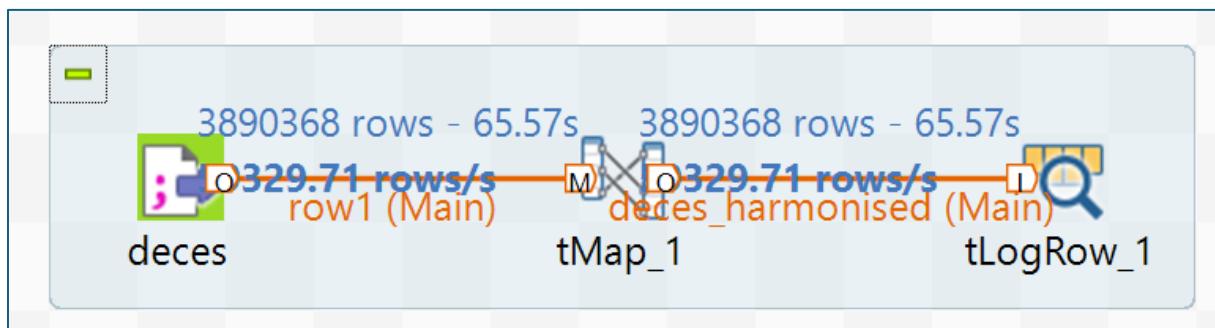


Figure 4 Job Talend

Le schéma du Job se compose de trois composants principaux :

- tFileInputDelimited (“deces”) : lit le fichier brut contenant les informations de décès (nom, prénom, date de naissance, date de décès, lieu, etc.).
- tMap (“tMap_1”) : transforme et harmonise les valeurs, notamment les dates, à l'aide d'une expression Java.
- tLogRow (“deces_harmonised”) : Il permet ici de visualiser les données transformées pour contrôler avant de charger.

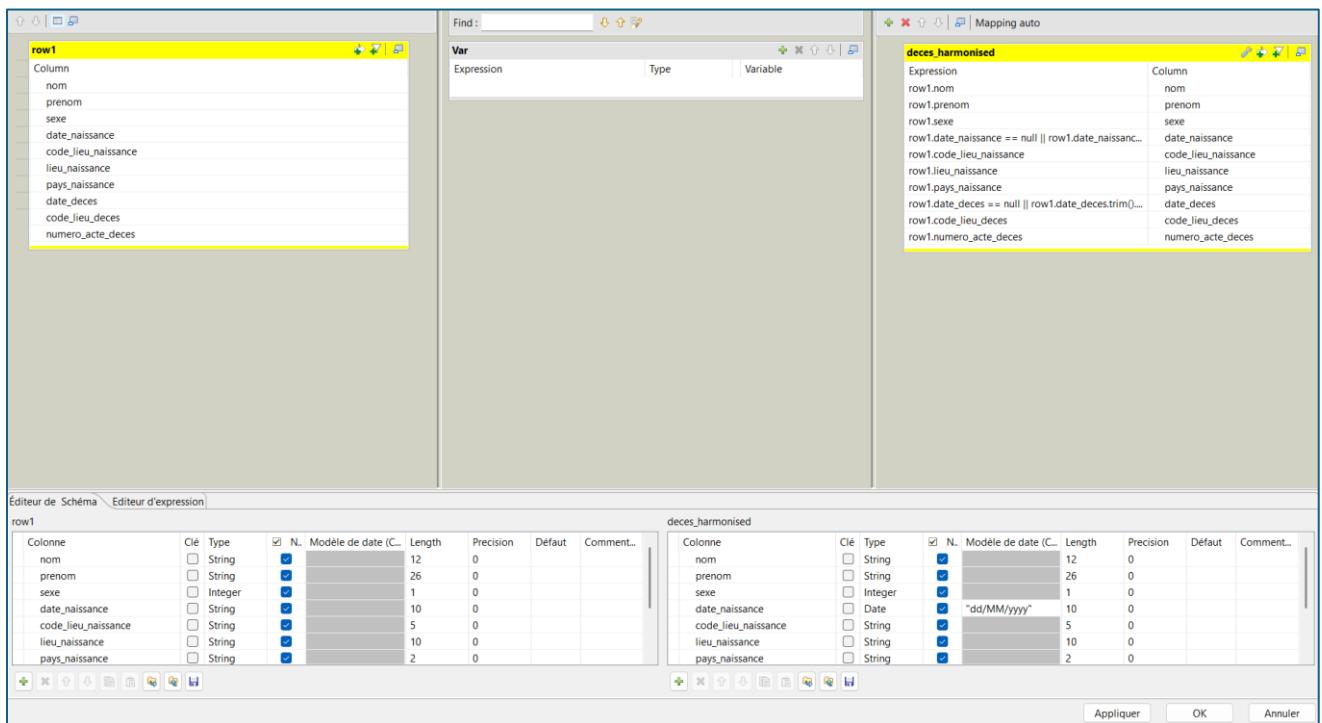


Figure 5 TMap du Job

Le cœur du Job réside dans le composant tMap, qui contient une expression de transformation complexe appliquée aux colonnes date_naissance et date_deces.

Cette expression a pour rôle de :

- Convertir les séparateurs - en / pour unifier la syntaxe ;
- Gérer les formats partiels (yyyy, yyyy-MM) en les complétant automatiquement pour former une date complète (yyyy/01/01 ou yyyy/MM/01) ;
- Rejeter les valeurs non valides ou vides en les remplaçant par null.

L'expression utilisée est la suivante :

```

Editeur de Schéma | Editeur d'expression
row1.date_naissance == null || row1.date_naissance.trim().isEmpty()
? null
:(
  !row1.date_naissance.replaceAll("-", "/").matches("^\\d{4}/\\d{2}/\\d{2}$")
  ? (
    row1.date_naissance.replaceAll("-", "/").matches("^\\d{4}$")
    ? TalendDate.parseDate("yyyy/MM/dd", row1.date_naissance.replaceAll("-", "/") + "/01/01")
    : (
      row1.date_naissance.replaceAll("-", "/").matches("^\\d{4}/\\d{2}$")
      ? TalendDate.parseDate("yyyy/MM/dd", row1.date_naissance.replaceAll("-", "/") + "/01")
      : null
    )
  )
  : TalendDate.parseDate("yyyy/MM/dd", row1.date_naissance.replaceAll("-", "/"))
)

```

Figure 6 Expression pour la colonne date_naissance

Cette logique vérifie d'abord si la valeur est vide ou nulle.

Ensuite, selon la longueur du format détecté :

- Si seule l'année est présente (yyyy), la date devient le 1er janvier de cette année (yyyy/01/01).
- Si l'année et le mois sont présents (yyyy/MM), la date devient le premier jour du mois (yyyy/MM/01).
- Si le format complet (yyyy/MM/dd) est présent, il est simplement converti au bon format.

Ainsi, toutes les dates, qu'elles soient initialement complètes ou partielles, sont uniformisées dans le format yyyy/MM/dd.

Ce procédé garantit une cohérence temporelle parfaite, évite les erreurs de parsing lors du chargement et facilite les agrégations par année, mois ou jour dans l'entrepôt de données.

Et de manière analogue pour la date de décès :

```
row1.date_deces == null || row1.date_deces.trim().isEmpty()
? null
: (
    row1.date_deces.replaceAll("-", "/").matches("^\\d{4}/\\d{2}/\\d{2}$")
    ? (
        row1.date_deces.replaceAll("-", "/").matches("^\\d{4}$")
        ? TalendDate.parseDate("yyyy/MM/dd", row1.date_deces.replaceAll("-", "/") + "/01/01")
        : (
            row1.date_deces.replaceAll("-", "/").matches("^\\d{4}/\\d{2}$")
            ? TalendDate.parseDate("yyyy/MM/dd", row1.date_deces.replaceAll("-", "/") + "/01")
            : null
        )
    )
    : TalendDate.parseDate("yyyy/MM/dd", row1.date_deces.replaceAll("-", "/"))
)
```

Figure 7 Expression pour la colonne date_deces

À l'issue de cette étape, le fichier harmonisé (deces_harmonised) contient près de 3,9 millions d'enregistrements traités en moins d'une minute, avec un débit moyen de 330 lignes par seconde.

Ce Job illustre parfaitement le rôle de Talend dans la phase de nettoyage et de normalisation des données : il garantit la qualité et la fiabilité du référentiel temporel avant son intégration dans la plateforme Big Data.

Conclusion

Ce premier livrable marque une étape essentielle dans la conception du système décisionnel du projet Cloud Healthcare Unit (CHU).

À travers l'analyse des sources, la définition de l'architecture et la modélisation des données, nous avons posé les fondations techniques et fonctionnelles de l'entrepôt de données.

Le travail réalisé a permis d'identifier, d'unifier et de structurer les différentes sources d'information issues des systèmes hospitaliers : bases relationnelles, fichiers CSV, enquêtes de satisfaction et données administratives.

Ces données hétérogènes ont été intégrées dans une vision globale grâce à la conception d'un MCD articulé autour de plusieurs tables de faits : consultation, hospitalisation, décès et satisfaction ; et de leurs dimensions associées : patient, établissement, professionnel de santé, diagnostic, région, date.

Cette structure garantit la cohérence, la traçabilité et la pertinence des analyses futures.

L'architecture retenue, basée sur un environnement Big Data (Hadoop/Cloudera) et un processus ETL orchestré par Talend, permettra d'assurer la robustesse et la scalabilité nécessaires au traitement de volumes importants de données.

Elle offre également un cadre sécurisé et conforme au RGPD pour la manipulation d'informations sensibles. Ce livrable a donc permis de transformer un ensemble de données hétérogènes en un référentiel décisionnel cohérent, apte à répondre aux besoins métier du CHU : suivre l'activité médicale, mesurer la performance hospitalière, analyser la satisfaction des patients et étudier la mortalité à différentes échelles temporelles et géographiques.

Les prochaines étapes du projet consisteront à implémenter physiquement ce modèle dans l'environnement Big Data, à développer les jobs d'alimentation complets, puis à mettre en place la couche de visualisation et d'analyse à l'aide d'outils tels que Power BI.

Ces phases permettront de concrétiser le travail de conception mené ici, pour aboutir à une solution décisionnelle opérationnelle et durable au service de la qualité des soins et de la performance hospitalière.

Table des figures

Figure 1 Schéma de l'architecture de l'entrepôt de données	12
Figure 2 MCD simplifié	17
Figure 3 MCD complet	18
Figure 4 Job Talend	19
Figure 5 TMap du Job	20
Figure 6 Expression pour la colonne date_naissance.....	20
Figure 7 Expression pour la colonne date_deces.....	21