Projet Big Data Healthcare

Cloud Healthcare Unit (CHU)

LIVRABLE 1

Référentiel de Données

Modèle conceptuel et architecture de l'entrepôt

Équipe Projet :
Nejma MOUALHI | Brieuc OLIVIERI | Nicolas TAING

Formation : CESI FISA A4
Année universitaire : 2025-2026

Date : 10/10/2025

Table des matières

1 Introduction

1.1 Contexte du projet CHU

Le groupe hospitalier **Cloud Healthcare Unit (CHU)** souhaite se doter d'un système décisionnel capable d'exploiter efficacement la grande quantité de données issues de ses systèmes médicaux et administratifs.

Ces données, générées quotidiennement par les soins, les consultations, les hospitalisations, la satisfaction des patients et les registres de décès, constituent une ressource stratégique pour l'amélioration des soins et la performance des établissements. Cependant, leur volume, leur hétérogénéité et leur caractère sensible rendent leur exploitation complexe. Les infrastructures relationnelles classiques ne suffisent plus à absorber, traiter et analyser ces flux massifs tout en garantissant la conformité réglementaire.

Dans ce contexte, le projet **CHU – Cloud Healthcare Unit** vise à concevoir une infrastructure Big Data évolutive, sécurisée et performante, capable d'intégrer, de transformer et d'exposer les données de santé dans un cadre analytique cohérent. Cette architecture servira de base à la mise en place d'un entrepôt de données (Data Warehouse) dédié à la décision médicale et stratégique.

1.2 Objectif global du système décisionnel Big Data

L'objectif global est de mettre en œuvre un écosystème décisionnel complet permettant :

— d'intégrer des sources multiples (base PostgreSQL, fichiers CSV historiques, exports FTP) ;

— de garantir la sécurité et la confidentialité des données conformément au RGPD et à la certification HDS ;

— de modéliser les données sous une forme décisionnelle (faits et dimensions) afin de permettre une analyse par axes : temps, diagnostic, professionnel, établissement, etc. ;

— d'assurer des performances d'accès optimales grâce à un stockage distribué et des transformations massivement parallèles (Hadoop / Hive / Spark) ;

— et enfin, de favoriser la visualisation et la restitution des indicateurs via des outils comme Power BI.

1.3 Enjeux du secteur médical

Le secteur de la santé est aujourd'hui confronté à quatre enjeux majeurs :

Le volume et la variété les établissements de santé génèrent des données massives, hétérogènes, issues de multiples systèmes (soins, satisfaction, mortalité, administratif).

La confidentialité les données de santé relèvent de la catégorie des données sensibles (article 9 du RGPD). Leur stockage et leur traitement nécessitent des mesures strictes de pseudonymisation, de traçabilité et d'hébergement certifié HDS.

La performance les utilisateurs attendent un accès rapide à l'information pour la prise de décision. Cela suppose un système capable de gérer des requêtes complexes sur de grands volumes en temps quasi réel.

L'évolutivité et la gouvernance les besoins analytiques changent avec le temps. L'architecture doit rester ouverte, extensible et bien documentée pour permettre de nouvelles analyses sans remise en cause du modèle existant.

1.4 But du livrable

Ce premier livrable a pour but de définir le référentiel de données du futur système décisionnel du CHU. Il s'agit de la phase de conception conceptuelle, centrée sur :

— la définition de l'architecture Big Data et du pipeline d'intégration (ETLT) adapté aux données médicales ;

— la modélisation conceptuelle des données sous forme de faits et dimensions alignés sur les besoins des utilisateurs ;

— et la description des flux d'alimentation nécessaires à la constitution du futur entrepôt.

Ce travail jette les bases du modèle physique et de la phase d'implémentation à venir dans le livrable 2, garantissant la continuité entre la conception et la réalisation du système décisionnel.

2 L'architecture Big Data

2.1 Choix de l'approche : ETLT

2.1.1 Justification du modèle

Le choix du pipeline **ETLT (Extract – Transform – Load – Transform)** s'impose dans le contexte d'un projet manipulant des **données médicales sensibles**. Contrairement à un **ETL classique**, où toutes les transformations précèdent le chargement, ou à un **ELT** où les données sont chargées brutes dans le cluster avant tout traitement, le modèle **ETLT** introduit une première transformation de **conformité** avant le stockage.

Cette étape intermédiaire répond à deux impératifs :

1. **Conformité réglementaire (RGPD / HDS)** : les données de santé ne peuvent pas être stockées dans leur forme brute, même temporairement, dans un environnement partagé. Une première transformation (T_1) est donc réalisée avant le chargement dans HDFS pour pseudonymiser, minimiser et normaliser les données sensibles.
2. **Performance et évolutivité** : la deuxième transformation (T_2) est réalisée directement dans le cluster Big Data, à grande échelle, pour préparer les données au modèle décisionnel. Ce découplage permet de séparer les traitements **sécuritaires** (avant le stockage) des traitements **analytiques** (après le stockage).

L'ETLT permet donc de **garantir la conformité** sans sacrifier la **scalabilité**. C'est une approche hybride, combinant la rigueur du monde décisionnel et la puissance du Big Data distribué.

2.1.2 Description du flux ETLT

Le pipeline se compose de quatre étapes principales :

E – Extract L'extraction regroupe les données provenant de plusieurs sources :

— **PostgreSQL** : données médico-administratives des patients et des consultations.

— **Fichiers CSV** : exports des établissements hospitaliers, enquêtes de satisfaction, hospitalisations et registre des décès.

Les outils recommandés sont **Apache Sqoop** (pour les bases relationnelles) et **Apache Flume** ou **NiFi** (pour les flux de fichiers CSV). Les données extraites sont ensuite validées (intégrité, schéma, doublons) avant de passer à l'étape suivante.

T_1 – Transform Conformité Cette première transformation intervient **avant le stockage dans HDFS**. Elle est dédiée à la **sécurité et la conformité** :

— **Pseudonymisation** : application d'un hachage salé sur les identifiants patients et professionnels de santé ;

— **Minimisation** : suppression des champs inutiles à l'analyse (adresse, téléphone, numéro de sécurité sociale, email) ;

— **Normalisation** : uniformisation des formats de dates, codes et unités ;

— **Contrôles** : validation du type, contraintes, dictionnaires (sexe, région, spécialité, diagnostic).

Les données ainsi transformées sont considérées comme *pseudonymisées*, donc stockables dans un environnement Big Data certifié.

L – Load Le chargement consiste à **insérer les données pseudonymisées dans le Data Lake (HDFS)**. Elles sont déposées dans la **zone Bronze (Landing)** sous leur forme semi-brute, accompagnées de métadonnées de traçabilité (source, date, checksum, version). Les fichiers sont stockés au format **Parquet** ou **ORC**, optimisés pour le stockage colonne et la compression.

Chaque ingestion est historisée pour permettre la reconstitution d'un état passé (audit trail).

T_2 – Transform Métier Cette deuxième transformation est exécutée directement dans le **cluster Big Data** (via **Spark** ou **Hive**) et correspond à la phase de **préparation analytique** :

— Nettoyage et harmonisation (référentiels communs, codes régionaux, tables FINESS) ;

— Conformation des dimensions (patient, professionnel, diagnostic, établissement, temps, satisfaction) ;

— Agrégations et calculs des indicateurs dans les **tables de faits** ;

— Génération du **modèle décisionnel en constellation** stocké dans la **zone Gold** du Data Lake.

Ce jeu de données devient la base du futur **entrepôt de données (Data Warehouse)**, exploitable via HiveQL, Power BI ou Tableau.

2.1.3 Respect du RGPD et de la certification HDS

L'architecture garantit la conformité réglementaire selon deux axes :

Technique

— Pseudonymisation avant stockage (aucune donnée directement identifiable dans HDFS) ;

— HDFS configuré avec chiffrement au repos et journalisation d'accès ;

— Contrôle d'accès granulaire via **Apache Ranger** et authentification Kerberos ;

— Hébergement sur une infrastructure **certifiée HDS**.

Organisationnel

— Politique de minimisation : seules les données strictement nécessaires aux analyses sont conservées ;

— Gestion des droits d'accès par rôle (RBAC) : distinction Data Engineer / Data Scientist / Analyste ;

— Journalisation des traitements et conservation limitée selon la durée légale ;

— Documentation et traçabilité complète des flux (Airflow + logs centralisés).

2.2 Description des couches de l'architecture

2.2.1 Sources de données

TABLE 1 – Récapitulatif des sources de données

Source	Description	Format	Volume estimé
PostgreSQL – Consultation	Données des consultations : date, diagnostic, professionnel, patient	Relationnel	∼1M lignes
PostgreSQL – Patient	Informations patients (pseudonymisées)	Relationnel	∼100K lignes
CSV – Établissements	Référentiel national des hôpitaux (FINESS)	CSV	24 colonnes
CSV – Hospitalisations	Données d'hospitalisation (durée, service, diagnostic)	CSV	Variable
CSV – Décès	Registre national des décès	CSV	Variable
CSV – Satisfaction	Enquêtes de satisfaction (2014–2020)	CSV (27 fichiers)	Variable

2.2.2 Zone Bronze (Landing)

La **zone Bronze** correspond au **niveau d'atterrissage des données pseudonymisées**. Les fichiers y sont stockés tels quels, sans modification de structure, afin de garantir la **traçabilité** et la **reproductibilité** des traitements. Chaque dataset comporte un fichier de métadonnées (JSON ou Avro) décrivant :

— la source d'origine ;

— le schéma de données ;

— la date et l'heure d'ingestion ;

— le hash de contrôle et l'identifiant de version.

Cette zone est considérée comme **non exploitable directement** par les analystes — elle sert de référence brute pour les traitements en Silver.

2.2.3 Zone Silver (Integration)

La **zone Silver** regroupe les données **nettoyées, harmonisées et jointes** à partir des différentes sources Bronze. Les opérations typiques sont :

— correspondance entre les identifiants patients/professionnels ;

— enrichissement à l'aide des référentiels (FINESS, géographie, spécialités) ;

— détection et suppression des doublons ;

— validation des clés étrangères.

Cette zone fournit des tables prêtes à être modélisées sous forme de faits et dimensions. Elle constitue l'espace de travail privilégié des **Data Engineers**.

2.2.4 Zone Gold (Data Warehouse)

La **zone Gold** représente la partie **décisionnelle** du Data Lake. Les données y sont organisées selon un **modèle en constellation** :

— **Tables de faits** : Consultation, Hospitalisation, Décès, Satisfaction ;

— **Dimensions communes** : Temps, Patient, Professionnel, Établissement, Diagnostic, etc.

Ces tables sont stockées au format **ORC** (optimisé pour le requêtage SQL via Hive) et indexées pour accélérer les analyses. Cette zone alimente directement les outils de **Business Intelligence** comme Power BI ou Tableau.

2.2.5 Outils préconisés

TABLE 2 – Stack technologique préconisée

Couche	**Outil**	**Rôle principal**
Extraction	Apache Sqoop	Import depuis PostgreSQL
Extraction	Apache Flume / NiFi	Ingestion des fichiers CSV
Transformation	Apache Spark	Exécution des jobs T_1 et T_2
Stockage	HDFS	Data Lake distribué
Data Warehouse	Apache Hive	Requêtage SQL et stockage ORC
Orchestration	Apache Airflow / NiFi	Séquencement, logs et monitoring
Sécurité	Apache Ranger	Contrôle d'accès granulaire, RBAC
Visualisation	Power BI / Tableau	Tableaux de bord interactifs

2.3 Analyse des données sources

2.3.1 Validation des volumes

L'analyse des données réelles confirme les estimations du projet :

TABLE 3 – Volumes réels des données sources

Source	Volume réel	Commentaire
PostgreSQL – Patient	100 000 lignes	Volume idéal pour tests et prototypage
PostgreSQL – Consultation	1 027 157 lignes	Confirme le challenge Big Data (1M+)
PostgreSQL – Professionnel	1 048 575 lignes	Base complète des professionnels
PostgreSQL – Diagnostic	15 490 codes	Référentiel CIM-10 complet
CSV – Satisfaction	27 fichiers	Période 2014-2020, structure évolutive
CSV – Décès	Volume important	Fichier national, plusieurs Mo

Ces volumes justifient pleinement l'approche Big Data : avec plus d'un million de consultations et la nécessité de croiser plusieurs sources hétérogènes, les technologies distribuées (Hadoop/Spark) s'imposent pour garantir la scalabilité.

2.3.2 Schéma UML des données PostgreSQL

L'analyse de la base de données PostgreSQL révèle un modèle relationnel classique qu'il faudra transformer en modèle dimensionnel :

FIGURE 1 – Schéma UML des tables PostgreSQL sources

2.3.3 Justification des choix architecturaux

Cette analyse des données réelles confirme nos choix techniques :

Nécessité de la pseudonymisation Les tables contiennent des données hautement sensibles (nom, prénom, email, téléphone, numéro de sécurité sociale). La transformation T_1 doit impérativement les pseudonymiser avant stockage dans HDFS.

Complexité des jointures Le modèle relationnel impose de multiples jointures (Patient↔Consultatio
Le modèle dimensionnel en constellation simplifiera grandement les requêtes analytiques.

Hétérogénéité des formats PostgreSQL utilise des types `date` et `time` normalisés, mais les CSV utilisent des formats texte variés. La couche Silver d'intégration est cruciale pour harmoniser ces formats.

Volume justifiant le Big Data Avec 1M+ de consultations et la perspective de croissance, les jointures sur des tables relationnelles classiques deviendraient rapidement prohibitives. Le stockage colonnaire (ORC) et le parallélisme (Spark) sont nécessaires.

2.4 Schéma global de l'architecture

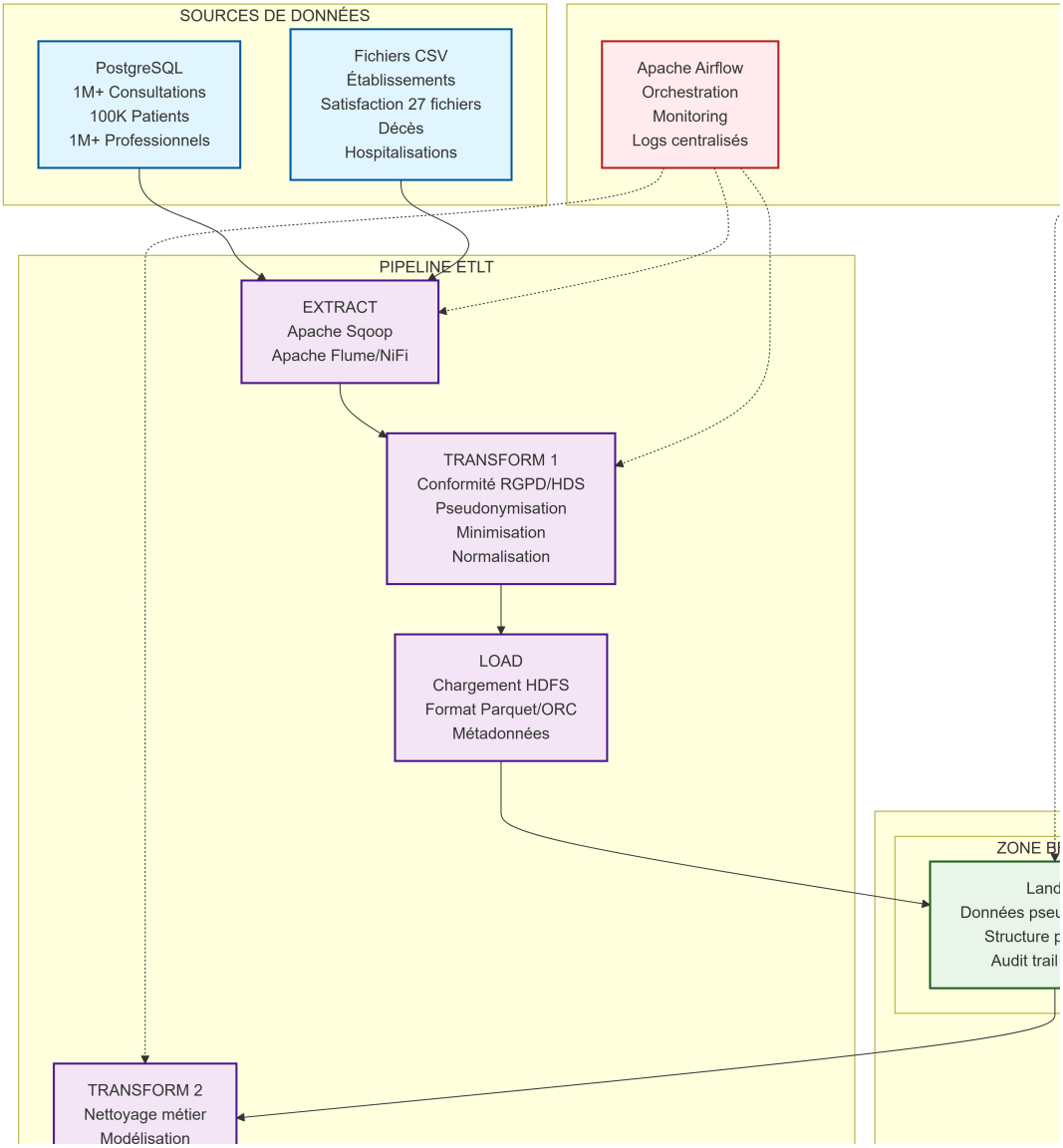

FIGURE 2 – Architecture ETLT avec zones Bronze/Silver/Gold

3 Modèle conceptuel de données

3.1 Identification des axes d'analyse (dimensions)

3.1.1 Dimension Temps

La dimension Temps constitue l'un des axes fondamentaux d'analyse décisionnelle. Elle permet de contextualiser l'ensemble des faits médicaux, administratifs et de satisfaction selon une perspective temporelle. L'analyse temporelle est essentielle pour évaluer l'évolution des taux d'hospitalisation, des consultations ou des décès et pour mettre en évidence des tendances saisonnières, des pics d'activité ou des améliorations des soins sur une période donnée.

Sources de données : horodatages présents dans les systèmes hospitaliers (dates d'admission, de consultation, de diagnostic, de décès, etc.) et fichiers plats / bases fournis par le groupe CHU.

Structure :

— `id_temps` (clé primaire)

— date_complète

— jour, mois, année

— trimestre, semaine, jour_semaine

— indicateurs saisonniers (jour férié, week-end, période estivale, etc.)

Usage analytique : permet de croiser les faits avec la temporalité pour mesurer des tendances (évolution mensuelle du taux d'hospitalisation, périodes de surcharge, évolution de la satisfaction, etc.).

3.1.2 Dimension Patient

La dimension Patient regroupe les informations démographiques et administratives des individus pris en charge par les établissements du groupe CHU. Elle permet une analyse fine de la population soignée selon des critères d'âge, de sexe et de profil sociodémographique.

Sources de données : principalement la base PostgreSQL médico-administrative, enrichie par des fichiers CSV des services d'état civil ou de suivi patient.

Structure :

— `id_patient` (clé primaire)

— sexe

— date_naissance / âge

— catégorie d'âge

— situation géographique (région / département)

— statut de vie (vivant / décédé)

Usage analytique : taux d'hospitalisation par tranche d'âge, répartition des consultations par sexe, corrélation âge / satisfaction, segmentation patient.

3.1.3 Dimension Professionnel de santé

La dimension Professionnel décrit le personnel médical et paramédical impliqué dans les soins. Elle offre une vue sur la distribution et l'efficacité des équipes soignantes.

Sources de données : systèmes RH hospitaliers et bases de gestion des ressources humaines du groupe CHU.

Structure :

— `id_professionnel` (clé primaire)

— nom, prénom (pseudonymisés si nécessaire)

— spécialité médicale

— statut (médecin, infirmier, aide-soignant, etc.)

— service / établissement d'affectation

— ancienneté / charge de travail moyenne

Usage analytique : répartition des actes par professionnel, taux de consultation par spécialité, impact de l'expérience sur la satisfaction patient.

3.1.4 Dimension Diagnostic

La dimension Diagnostic regroupe les pathologies identifiées lors des consultations ou hospitalisations. Elle permet d'analyser la répartition et la fréquence des maladies, ainsi que leur lien avec les soins prodigués.

Sources de données : systèmes médicaux du CHU et codes diagnostics (CIM-10 ou équivalents).

Structure :

— `id_diagnostic` (clé primaire)

— code_diagnostic (CIM-10)

— libelle_diagnostic

— catégorie_pathologie

— gravité / type d'intervention associée

Usage analytique : prévalence des pathologies, taux d'hospitalisation par diagnostic, corrélations diagnostic / durée moyenne de séjour.

3.1.5 Dimension Établissement

La dimension Établissement décrit les structures de santé (hôpitaux, cliniques, établissements publics ou privés) rattachées au groupe CHU. Elle est essentielle pour la comparaison inter-établissements et le suivi de la performance hospitalière.

Sources de données : fichier CSV national des établissements (FINESS) et systèmes d'information internes du CHU.

Structure :

— `id_etablissement` (clé primaire)

— nom_etablissement

— type_etablissement (CHU, clinique, centre régional, etc.)

— capacité d'accueil (nombre de lits)

— région / département

— code_Finess
— spécialité principale

Usage analytique : comparaison de fréquentation, identification des zones les plus sollicitées, suivi de la performance opérationnelle par taille ou spécialité.

3.1.6 Dimension Localisation

La dimension Localisation fournit la perspective géographique des analyses. Elle relie les faits aux zones administratives et sanitaires (régions, départements, communes).

Sources de données : fichiers d'état civil, base nationale des établissements hospitaliers, coordonnées géographiques des établissements et des patients.

Structure :

— `id_localisation` (clé primaire)
— région, département, commune
— code_postal
— zone_urbaine / rurale
— coordonnées GPS

Usage analytique : cartographie de la répartition des patients, mortalité ou satisfaction par région, détection de zones sous-dotées en infrastructures.

3.1.7 Dimension Satisfaction

La dimension Satisfaction évalue la perception des patients sur la qualité des soins reçus. Elle est essentielle pour le pilotage de la qualité hospitalière et l'amélioration continue.

Sources de données : fichiers plats de satisfaction patient et enquêtes administrées par les établissements du CHU.

Structure :

— `id_satisfaction` (clé primaire)
— note_globale
— critères (accueil, propreté, qualité du soin, disponibilité du personnel)
— année / période d'enquête
— source de collecte (en ligne, papier, entretien)

Usage analytique : indicateurs de satisfaction par établissement ou région, suivi temporel du ressenti patient, corrélation satisfaction / performance médicale.

3.2 Identification des faits et mesures

3.2.1 Fait Consultation

La table **FAIT_CONSULTATION** centralise les informations relatives aux consultations médicales réalisées dans les établissements du groupe CHU. Elle permet d'analyser l'activité des praticiens, la fréquentation des patients et les tendances de consultation selon différents axes (temps, localisation, diagnostic, professionnel, etc.). Cette table constitue un indicateur opérationnel clé traduisant la demande de soins au sein du réseau hospitalier.

Objectifs analytiques :

— mesurer le taux de consultation par période, région ou diagnostic ;
— suivre la distribution des consultations selon le sexe, l'âge et le profil socio-démographique ;
— évaluer la charge d'activité des professionnels ;
— identifier les évolutions de la demande de soins dans le temps.

Liens dimensionnels :

— `DIM_TEMPS` : date de la consultation ;
— `DIM_PATIENT` : informations démographiques du patient ;
— `DIM_PROFESSIONNEL` : praticien effectuant la consultation ;
— `DIM_ETABLISSEMENT` : lieu de la consultation ;
— `DIM_DIAGNOSTIC` : motif ou résultat de la consultation ;
— `DIM_LOCALISATION` : région / département associés.

Granularité : une ligne = une consultation (horodatage précis, patient, professionnel, diagnostic).

Mesures principales :

— `nb_consultations` : nombre de consultations ;
— `duree_consultation_moyenne` : durée moyenne par consultation ;
— `taux_consultation_par_diagnostic` ;
— `taux_consultation_par_professionnel` .

3.2.2 Fait Hospitalisation

La table **FAIT_HOSPITALISATION** regroupe les données relatives aux séjours hospitaliers. Elle est destinée aux analyses d'occupation, de durée de séjour et de répartition des pathologies hospitalisées, et sert de référence pour le pilotage des capacités et de la performance clinique.

Objectifs analytiques :

— mesurer le taux d'hospitalisation par période, région et diagnostic ;
— calculer la durée moyenne de séjour (DMS) par pathologie et profil patient ;
— analyser la fréquence d'hospitalisation par sexe et tranche d'âge ;
— estimer le taux d'occupation des lits et services hospitaliers.

Liens dimensionnels :

— `DIM_TEMPS` : date d'entrée, date de sortie ;
— `DIM_PATIENT` : profil du patient hospitalisé ;
— `DIM_DIAGNOSTIC` : motif d'hospitalisation ;
— `DIM_ETABLISSEMENT` : établissement d'accueil ;
— `DIM_PROFESSIONNEL` : médecin référent ;
— `DIM_LOCALISATION` : position géographique de l'établissement.

Granularité : une ligne = un séjour hospitalier (entrant, sortant, diagnostics associés).

Mesures principales :

— `nb_hospitalisations` ;
— `duree_sejour_moyenne` (jours) ;
— `taux_occupation` (lits occupés / lits disponibles) ;
— `taux_hospitalisation_par_diagnostic`.

3.2.3 Fait Satisfaction

La table **FAIT_SATISFACTION** consolide les résultats des enquêtes de satisfaction patient menées après une consultation ou un séjour. Elle sert à piloter la qualité perçue et à orienter les actions d'amélioration.

Objectifs analytiques :

— calculer le taux de satisfaction global par établissement, région et période ;

— analyser les critères de satisfaction (accueil, propreté, qualité du soin, disponibilité du personnel) ;

— suivre l'évolution temporelle des indicateurs de satisfaction ;

— corréler satisfaction, durée de séjour et typologie de pathologie.

Liens dimensionnels :

— `DIM_TEMPS` : période / date d'enquête ;

— `DIM_PATIENT` : répondant (anonymisé) ;

— `DIM_ETABLISSEMENT` : établissement concerné ;

— `DIM_SATISFACTION` : codification des critères ;

— `DIM_LOCALISATION` : zone géographique du répondant.

Granularité : une ligne = une réponse d'enquête (ou agrégation par sondage selon disponibilité).

Mesures principales :

— `note_satisfaction_moyenne` ;

— `taux_reclamation` ;

— `indice_qualite_service` ;

— `nb_enquetes_realisees`.

3.2.4 Fait Décès

La table **FAIT_DECES** centralise les informations issues des registres de décès et permet d'analyser la mortalité selon des critères temporels, géographiques et médicaux. Elle alimente les analyses épidémiologiques et de planification sanitaire.

Objectifs analytiques :

— étudier la répartition des décès par région et période ;

— identifier les diagnostics associés aux décès ;

— mesurer le taux de mortalité hospitalière ;

— croiser les décès avec les caractéristiques patient (âge, sexe, établissement).

Liens dimensionnels :

— `DIM_TEMPS` : date du décès ;

— `DIM_PATIENT` : caractéristiques du défunt ;

— `DIM_ETABLISSEMENT` : lieu du décès (si hospitalier) ;

— `DIM_DIAGNOSTIC` : cause principale du décès ;

— `DIM_LOCALISATION` : région / commune de décès.

Granularité : une ligne = un acte de décès déclaré (avec diagnostics codés si disponibles).

Mesures principales :

— `nb_deces` ;

— `taux_mortalite` (décès / population hospitalisée) ;

— `age_moyen_deces` ;

— `taux_deces_par_diagnostic`.

3.3 Modélisation conceptuelle (diagramme constellation)

3.3.1 Présentation graphique

Le modèle dimensionnel adopte une architecture en constellation avec 4 tables de faits interconnectées par 8 dimensions. Ce schéma permet d'analyser l'activité hospitalière selon différents processus métier tout en partageant des dimensions communes (Temps, Patient, Établissement, Diagnostic).

FIGURE 3 – Schéma en constellation du modèle dimensionnel CHU

Le diagramme illustre les relations entre les 4 tables de faits (Consultation, Hospitalisation, Décès, Satisfaction) et les 8 dimensions identifiées. Les dimensions communes (en bleu) sont partagées entre plusieurs faits, tandis que les dimensions spécifiques (en vert) ne concernent qu'un seul processus métier.

3.3.2 Justification du modèle constellation

Le choix d'une architecture en constellation repose sur l'analyse des processus métier distincts du groupe CHU :

Quatre processus métier indépendants

— **Consultation** : activité ambulatoire des professionnels de santé

— **Hospitalisation** : séjours avec durée et gestion des lits

— **Décès** : mortalité avec contexte géographique

— **Satisfaction** : enquêtes de qualité perçue

Dimensions communes partagées Les dimensions Temps, Patient, Établissement et Diagnostic sont réutilisées par plusieurs faits, ce qui garantit la cohérence des analyses croisées et réduit la redondance.

Dimensions spécifiques métier Certaines dimensions ne concernent qu'un seul processus : Professionnel et Mutuelle pour les consultations, Type_Enquête pour la satisfaction. Cette spécialisation évite la pollution dimensionnelle et simplifie les modèles.

Avantages du modèle constellation

— Séparation logique des processus métier distincts

— Réutilisation optimale des dimensions communes

— Flexibilité pour ajouter de nouveaux faits sans refonte globale

— Performance des requêtes par processus (pas de sur-dimensionnement)

3.3.3 Relations clés entre faits et dimensions

Le tableau suivant récapitule les dimensions liées à chaque table de faits :

TABLE 4 – Récapitulatif des relations faits-dimensions

Table de faits	**Dimensions liées**
FAIT_CONSULTATION	DIM_TEMPS, DIM_PATIENT, DIM_PROFESSIONNEL, DIM_DIAGNOSTIC, DIM_ETABLISSEMENT, DIM_MUTUELLE
FAIT_HOSPITALISATION	DIM_TEMPS (entrée + sortie), DIM_PATIENT, DIM_ETABLISSEMENT, DIM_DIAGNOSTIC
FAIT_DECES	DIM_TEMPS, DIM_PATIENT
FAIT_SATISFACTION	DIM_TEMPS, DIM_ETABLISSEMENT, DIM_TYPE_ENQUETE

Cette structure en constellation permet d'analyser chaque processus métier de manière autonome tout en conservant la cohérence globale grâce aux dimensions partagées.

4 Jobs d'alimentation (vue conceptuelle)

4.1 Objectifs

Cette section présente la conception globale des jobs d'alimentation destinés à automatiser la préparation et le chargement des données dans le futur entrepôt décisionnel. L'objectif est de définir une vision fonctionnelle et séquentielle du traitement des données, avant la phase de développement qui sera réalisée avec Talend Open Studio dans le livrable suivant.

Les principaux objectifs sont :

— Garantir la qualité et la traçabilité des flux de données tout au long du processus

— Assurer la conformité RGPD/HDS dès les premières étapes d'ingestion

— Préparer le futur entrepôt décisionnel et les tables de faits/dimensions nécessaires à la restitution

— Structurer une chaîne d'alimentation claire et rejouable en cas d'erreur ou d'évolution des sources

— Favoriser la standardisation des étapes pour faciliter l'industrialisation future dans Talend

4.2 Description des jobs

La table suivante illustre les principaux jobs d'alimentation prévus dans le cadre du projet. Ils sont organisés en deux grandes phases :

— $\mathbf{T}_1$: Préparation et nettoyage des données sources (qualité, pseudonymisation, normalisation)

— $\mathbf{T}_2$: Construction du modèle décisionnel (dimensions, faits, indicateurs)

Il s'agit ici d'une vue conceptuelle et non technique ; les noms et étapes sont susceptibles d'évoluer lors du développement avec Talend.

TABLE 5 – Liste des jobs d'alimentation

Job	Étape	Rôle principal
J-T_1-Patient	T_1	Pseudonymisation et suppression des informations personnelles (PII) pour conformité RGPD
J-T_1-Consultation	T_1	Normalisation des dates, harmonisation des structures de fichiers et des schémas
J-T_1-Etablissement	T_1	Nettoyage des adresses, validation des identifiants FINESS, contrôle des doublons
J-T_1-Deces	T_1	Pseudonymisation et agrégation des données de mortalité par année et par région
J-T_1-Satisfaction	T_1	Harmonisation des enquêtes de satisfaction (ESATIS48H, ESATISCA 2020) et formats multiples
J-Dim_Temps	T_2	Génération d'un calendrier décisionnel complet (jour, mois, trimestre, année)
J-Dim_Patient	T_2	Construction de la dimension patient à partir des données pseudonymisées
J-Dim_Etablissement	T_2	Mapping des établissements, rattachement régional et fusion des sources
J-Dim_Diagnostic	T_2	Structuration de la hiérarchie CIM-10 pour les diagnostics médicaux
J-Dim_Professionnel	T_2	Regroupement des professionnels de santé et jointure sur spécialité
J-Dim_Specialite	T_2	Intégration du référentiel des spécialités médicales
J-Dim_Mutuelle	T_2	Construction de la dimension mutuelle et adhésions patients
J-Dim_Type_Enquete	T_2	Création de la typologie des enquêtes de satisfaction
J-Fait_Consultation	T_2	Agrégation des consultations selon les axes d'analyse (année, région, diagnostic)
J-Fait_Hospitalisation	T_2	Calcul des durées moyennes de séjour et indicateurs par établissement
J-Fait_Deces	T_2	Croisement localisation / temps pour suivi de la mortalité
J-Fait_Satisfaction	T_2	Pivot et agrégation des indicateurs de satisfaction par établissement et région

4.3 Séquencement et dépendances

L'exécution des jobs suivra une logique séquentielle structurée en deux phases :

Phase T_1 – Préparation des données Ingestion, contrôle de conformité et harmonisation des fichiers sources. Les jobs de cette phase garantissent la qualité et la cohérence des données avant toute intégration. Les jobs T_1 sont indépendants et peuvent s'exécuter en parallèle.

Phase T_2 – Construction du modèle décisionnel Création des dimensions (temps, établissement, patient) et des faits (hospitalisations, satisfaction). Ces jobs permettent

d'alimenter la base décisionnelle qui servira ensuite à la production d'indicateurs pour Power BI.

Dépendances principales :

— Les jobs T_2 dépendent de la complétion de tous les jobs T_1 correspondants

— Les jobs de construction des dimensions doivent précéder les jobs de construction des faits

— Les jobs de faits peuvent s'exécuter en parallèle une fois toutes les dimensions créées

Un diagramme de séquence ou DAG (Airflow ou Talend) sera développé dans le Livrable 2, afin de visualiser les dépendances entre jobs, leurs ordres d'exécution, et les points de contrôle de qualité.

4.4 Perspective : implémentation sous Talend (Livrable 2)

La prochaine étape du projet consistera à développer ces jobs dans Talend Open Studio, en transformant la conception présentée ici en chaîne d'intégration automatisée. Chaque job sera matérialisé sous forme de flux Talend intégrant :

— des composants de lecture et transformation des données sources

— des contrôles de qualité (cohérence, complétude, unicité)

— des liens logiques entre les tables de faits et de dimensions

— une publication automatisée des jeux validés vers le schéma décisionnel

Cette mise en œuvre permettra de passer d'une architecture théorique à un flux décisionnel complet et opérationnel, garantissant la fiabilité et la reproductibilité des chargements de données.

5 Conclusion et prochaines étapes

5.1 Synthèse du livrable

Ce premier livrable a permis de définir le référentiel de données du système décisionnel CHU. Les principaux livrables sont :

— Une architecture ETLT adaptée aux données médicales sensibles, garantissant la conformité RGPD/HDS
— Un modèle dimensionnel en constellation avec 8 dimensions et 4 tables de faits
— Une identification précise des sources de données (PostgreSQL, CSV, XLSX) avec validation des volumes
— Une conception des jobs d'alimentation T_1 (conformité) et T_2 (transformation métier)
— Une documentation complète des choix techniques et des contraintes réglementaires

Les choix structurants retenus sont :

— Architecture en constellation pour séparer logiquement les 4 processus métier
— Pseudonymisation SHA-256 avant stockage HDFS (étape T_1)
— Enrichissements CIM-10, FINESS et géographiques en T_2
— Données satisfaction limitées à 2020 (ESATIS48H + ESATISCA) pour garantir l'homogénéité

5.2 Prochaines étapes (Livrable 2)

Le livrable 2 portera sur l'implémentation physique du modèle conçu dans ce document :

— Développement des jobs Talend pour les transformations T_1 et T_2
— Implémentation des tables Hive avec partitionnement et bucketing
— Chargement effectif des données sources (PostgreSQL + CSV + XLSX)
— Tests de performance et optimisations
— Validation des temps de réponse sur les requêtes métier
— Mise en place de l'orchestration (Airflow ou Talend) avec DAG complet
— Création des dashboards Power BI pour la restitution

5.3 Risques identifiés et mesures d'atténuation

TABLE 6 – Risques identifiés

Risque	**Impact**	**Mitigation**
Qualité des données sources	Résultats analytiques erronés	Contrôles qualité en T_1, rejets traçables
Performance des jointures	Temps de requête prohibitifs	Partitionnement Hive, format ORC, bucketing
Conformité RGPD	Sanctions réglementaires	Pseudonymisation systématique T_1, audit trail
Évolution des sources	Rupture du pipeline	Tests de non-régression, versioning des schémas

Références

[1] Ralph Kimball et Margy Ross, *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, 3ème édition, Wiley, 2013.

[2] Tom White, *Hadoop : The Definitive Guide*, 4ème édition, O'Reilly Media, 2015.

[3] Bill Chambers et Matei Zaharia, *Spark : The Definitive Guide*, O'Reilly Media, 2018.

[4] Commission Nationale de l'Informatique et des Libertés (CNIL), *Guide du développeur - Conformité RGPD*, https://www.cnil.fr/, 2024.

[5] Agence du Numérique en Santé, *Référentiel de certification Hébergeur de Données de Santé (HDS)*, https://esante.gouv.fr/, 2024.