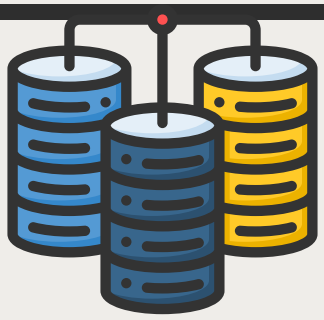
Livrable1



MATHEO PINGET | EVAN JOASSON| ALBAN CALVO

FISA 24-27

SOMMAIRE

1. Contexte

Le secteur de la santé connaît aujourd'hui une transformation numérique profonde, portée par l'essor des données médicales et par la nécessité d'améliorer la qualité des soins. Le volume, la variété et la valeur des données issues des systèmes hospitaliers, des dispositifs médicaux connectés et des plateformes administratives représentent une source considérable d'informations exploitables. Ces données, lorsqu'elles sont correctement intégrées et analysées, peuvent contribuer à une meilleure compréhension des parcours patients, à l'optimisation de la gestion hospitalière et à la prise de décision éclairée.

Cependant, l'exploitation de ces masses de données reste un défi majeur pour les établissements de santé. En effet, la diversité des formats (bases de données relationnelles, fichiers CSV, fichiers plats sur FTP, etc.) et la dispersion des sources complexifient leur intégration. De plus, les contraintes de performance, de sécurité et de confidentialité imposent la mise en place d'infrastructures robustes, évolutives et conformes aux exigences du domaine médical.

C'est dans ce contexte que s'inscrit le projet Cloud Healthcare Unit (CHU). À l'image de nombreux acteurs du secteur hospitalier, le groupe CHU souhaite amorcer une transformation digitale en développant son propre entre entrepôt de données.

Ce dernier doit permettre d'unifier, stocker et exploiter les données issues de différentes sources — bases médico-administratives, fichiers de satisfaction, répertoire des décès, gestion des établissements hospitaliers — afin de produire des analyses pertinentes et des indicateurs de performance fiables.

2. Demandes du livrable

Objet du livrable

Ce livrable formalise le référentiel de données du projet CHU et pose les bases du schéma décisionnel à partir duquel seront réalisées les analyses métiers (suivi patient, hospitalisations, satisfaction, décès, etc.). Il couvre à la fois la modélisation (axes et mesures), la définition des traitements d'alimentation (jobs ETL/ELT) et la description de l'architecture de l'entrepôt de données.

Contenu attendu

- **Modèle conceptuel des données (MCD)**
  - Identification des axes d'analyse (temps, établissement, patient, professionnel, diagnostic, localisation, sexe, tranche d'âge…).
  - Définition des mesures (taux de consultation, taux d'hospitalisation, nombre de décès, scores de satisfaction, indicateurs dérivés).
  - Dictionnaire de données (entités, attributs, clés, règles d'intégrité et de confidentialité).

- **Jobs nécessaires pour alimenter le schéma décisionnel**
  - Spécification des flux d'intégration depuis les sources (PostgreSQL soins médico-administratifs, CSV établissements, fichiers plats satisfaction, répertoire des décès).
  - Conception des processus d'extraction, de transformation et de chargement (nettoyage, normalisation, historisation, déduplication, gestion des référentiels, contrôles de qualité).
  - Orchestration et planification (dépendances, fréquence, surveillance, journalisation, reprise sur incident).

- **Description de l'architecture de l'entrepôt de données**
  - Schéma cible (zones staging, intégration/ODS, data warehouse et datamarts).
  - Patrons de modélisation décisionnelle (étoile/flocon, gestion des slowly changing dimensions).
  - Exigences non fonctionnelles : sécurité & conformité (anonymisation/pseudonymisation, accès, traçabilité), scalabilité, coût-efficacité, performance (indexation, partitionnement), gouvernance (catalogue, lignage, qualité).

3. Modélisation des différents axes d'analyse ainsi que les mesures

3.1. Objectif de la modélisation

La modélisation a pour objectif de structurer les données issues des systèmes sources hétérogènes afin de permettre une analyse décisionnelle fiable, cohérente et performante.
Elle repose sur un modèle en étoile dans lequel les tables de faits centralisent les indicateurs quantitatifs et les tables de dimensions fournissent les axes d'analyse contextuels.

3.2. Axes d'analyse (dimensions)

Les principales dimensions identifiées pour les analyses métiers sont :

- **dim_patient** : informations démographiques et médicales du patient (âge, sexe, groupe sanguin, date de naissance, etc.).
- **dim_praticien** : caractéristiques du professionnel de santé (nom, spécialité, mode d'exercice, établissement associé).
- **dim_etablissement** : référentiel FINESS et SIRET des établissements hospitaliers (adresse, type, région, département, statut juridique).
- **dim_diagnostic** : classification des pathologies selon la nomenclature CIM-10 (code, libellé, catégorie).
- **dim_temps** : hiérarchie temporelle pour l'analyse (jour, mois, trimestre, année, semaine ISO, jour ouvré).
- **dim_localisation** : découpage géographique (commune, département, région, pays).
- **bridge_activite_praticien_etablissement** : table de liaison entre praticiens et établissements, permettant l'analyse multi-sites.

3.3. Mesures (faits)

Les mesures sont regroupées dans plusieurs tables de faits, liées aux événements cliniques et administratifs :

- **fact_consultation** : nombre de consultations, durée, diagnostics associés, motifs.
- **fact_hospitalisation** : nombre d'hospitalisations, durée moyenne de séjour, réadmissions, codes diagnostics principaux.
- **fact_deces** : nombre de décès, répartition par âge, sexe, région, causes.
- **mart_satisfaction_region** : taux de satisfaction par région et indicateurs d'accueil, PEC, recommandation.
- **mart_deces_region** : taux de mortalité régionale et évolution temporelle.

Les indicateurs calculés incluent :

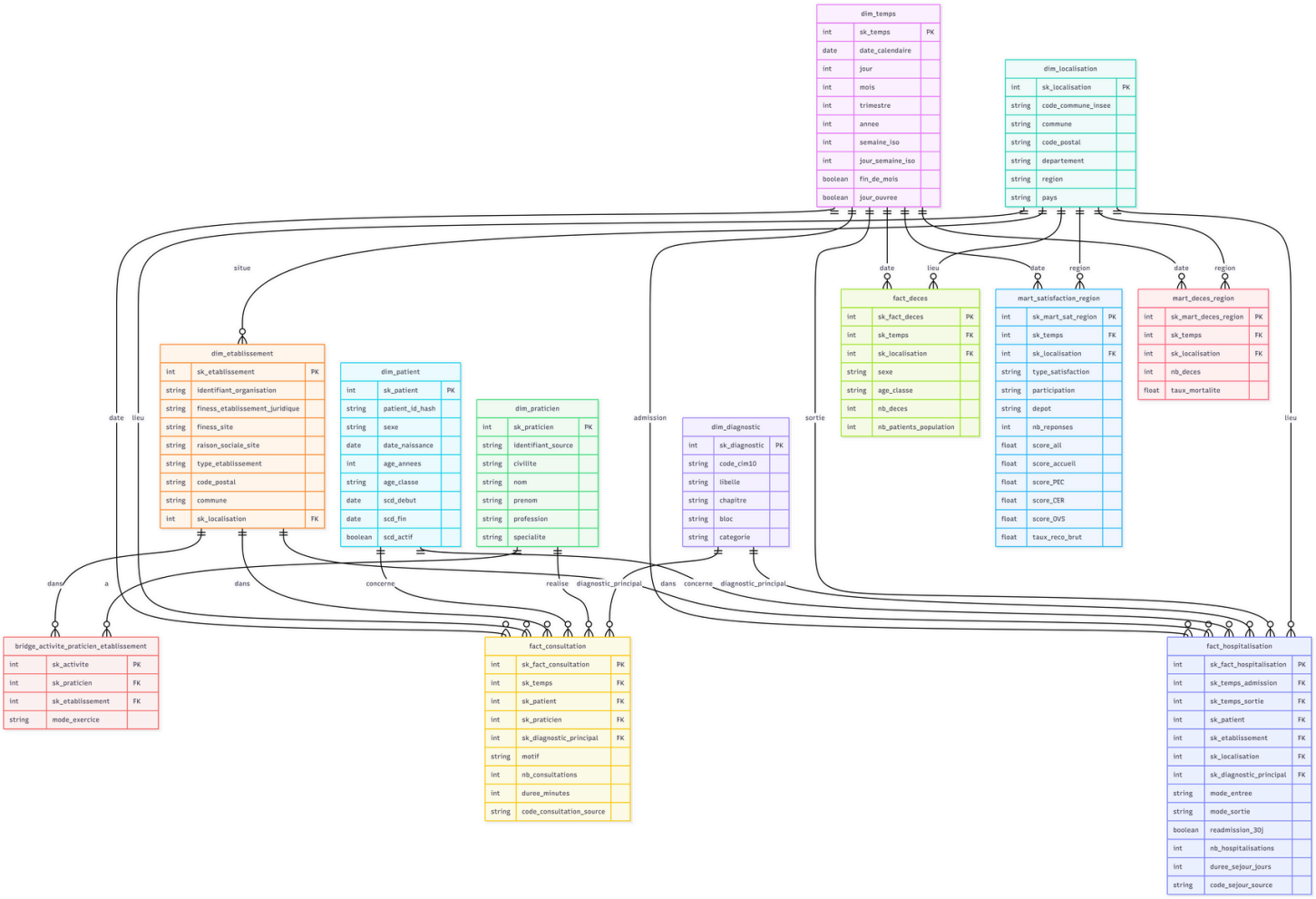
- Taux de consultation par praticien, établissement ou diagnostic.
- Taux d'hospitalisation global et par catégorie de pathologie.
- Taux de mortalité par tranche d'âge, sexe et région.
- Scores de satisfaction moyens et ajustés par région.
- Durée moyenne de séjour et indicateurs de réadmission.

3. Modélisation des différents axes d'analyse ainsi que les mesures

3.4. Schéma décisionnel (modèle en étoile)

Le modèle décisionnel global suit une architecture étoile :

- **Dimensions** : Patient, Praticien, Établissement, Diagnostic, Temps, Localisation.
- **Faits** : Consultation, Hospitalisation, Décès.
- **Les data marts** régionaux dérivent de ces faits pour les analyses de satisfaction et de mortalité.



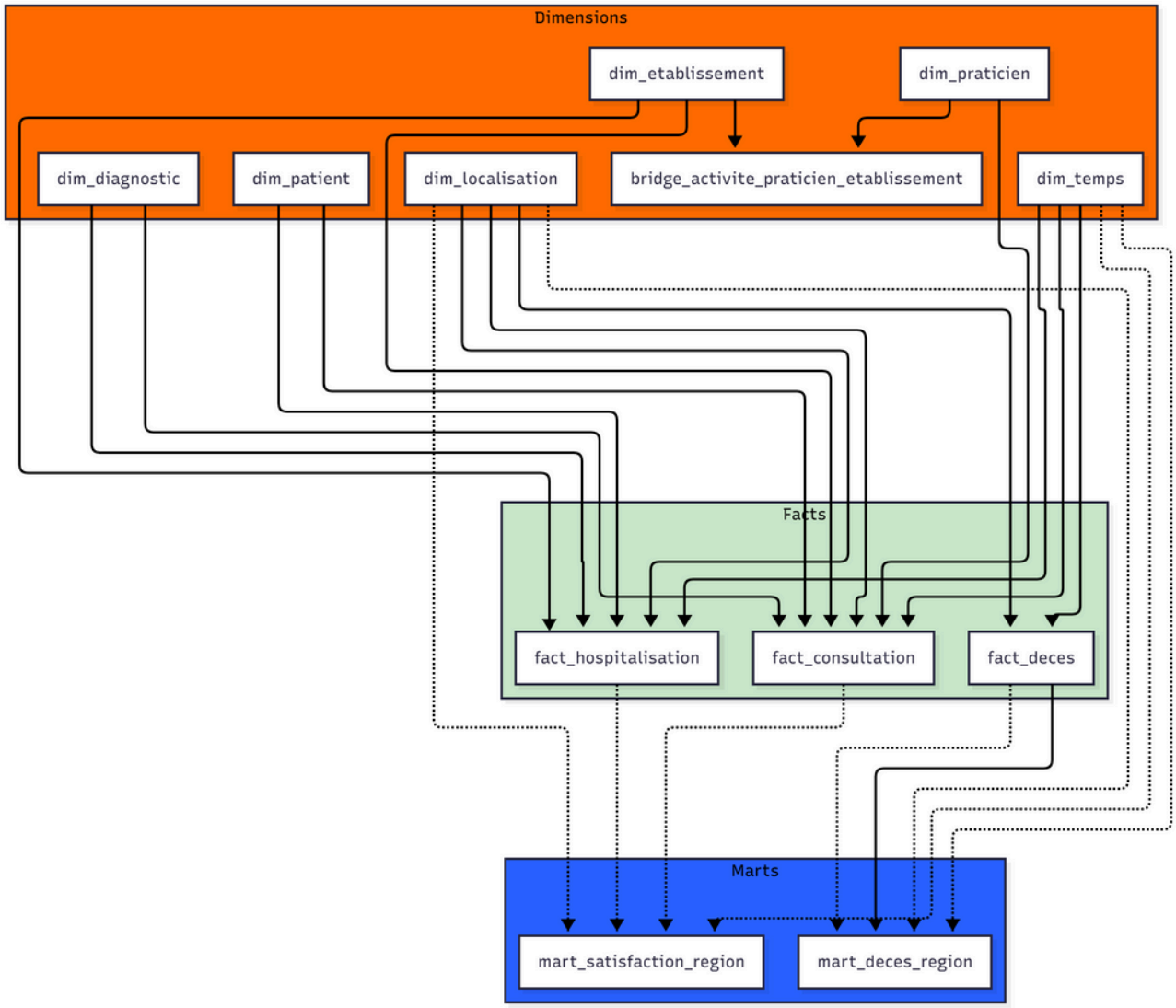
lien du modèle en étoile

3. Modélisation des différents axes d'analyse ainsi que les mesures

3.4. Schéma décisionnel (modèle en étoile)

Le modèle décisionnel global suit une architecture étoile :

- **Dimensions** : Patient, Praticien, Établissement, Diagnostic, Temps, Localisation.
- **Faits** : Consultation, Hospitalisation, Décès.
- **Les data marts** régionaux dérivent de ces faits pour les analyses de satisfaction et de mortalité.



lien du diag dim fact mart

4. Développement des jobs d'alimentation du schéma décisionnel

4.1. Objectif

Les jobs d'alimentation constituent la chaîne automatisée permettant d'extraire, transformer et charger (ETL) les données issues des systèmes sources vers l'entrepôt décisionnel.
Ils garantissent la qualité, la cohérence, la sécurité et la traçabilité des informations à travers toutes les zones du data lake.

4.2. Politique de transformation et d'alimentation du système décisionnel

4.2.1 Introduction

La politique de transformation et d'alimentation définit l'ensemble des processus ETL (Extract, Transform, Load) nécessaires à l'intégration des données dans le système décisionnel.

Elle s'appuie sur deux moteurs principaux :

- Talend : extraction, typage, nettoyage, anonymisation et orchestration.
- Apache Spark : enrichissement, transformation distribuée et modélisation analytique.

L'ensemble des traitements est exécuté sur la Cloudera Data Platform (CDP), qui garantit la scalabilité, la sécurité et la gouvernance des données sensibles.

Les objectifs principaux sont :

- Assurer la traçabilité complète des flux.
- Maintenir la qualité et la cohérence des informations.
- Garantir la fiabilité du modèle décisionnel basé sur le schéma en étoile (zone Gold).

4.2.2 Architecture globale du pipeline de données

Le pipeline se compose de plusieurs zones logiques correspondant à des niveaux progressifs de traitement :

1. **Zone Bronze – Staging typé (/bronze/)**
   - Normalisation et typage homogène des données.
   - Conversion des formats (SQL, CSV) vers Parquet.
   - Anonymisation SHA256 des identifiants sensibles.
   - Ajout de colonnes techniques :
   - _ingestion_date, _source, _hash_id.
2. **Zone Silver – Traité (/silver/)**
   - Jointures multi-sources (patients, diagnostics, satisfaction, décès, établissements).
   - Nettoyage, dédoublonnage, validation chronologique.
   - Enrichissement sémantique (âge, sexe, région, code diagnostic normalisé).
   - Contrôle de qualité et cohérence inter-domaines.

4. Développement des jobs d'alimentation du schéma décisionnel

4.2. Politique de transformation et d'alimentation du système décisionnel

4.2.2 Architecture globale du pipeline de données

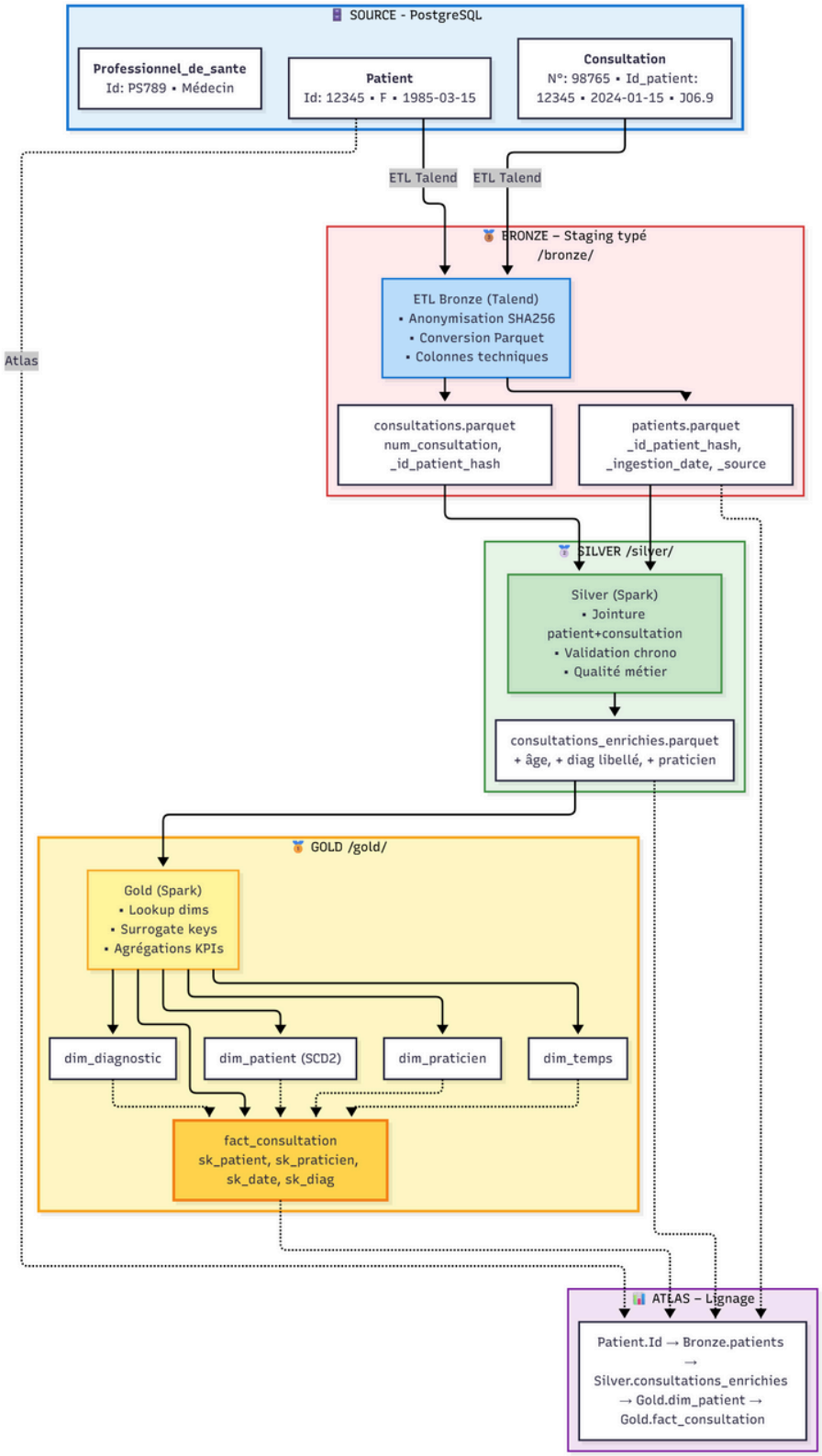
3. **Zone Gold – Analytique (/gold/)**
   - Modélisation en étoile.
   - Création des tables de faits et dimensions.
   - Agrégations et calculs d'indicateurs (KPI).
   - Génération des data marts thématiques (satisfaction, décès).
   - Optimisation via Iceberg, partitionnement et transactions ACID.
4. **Zone d'exposition SQL (Hive / Impala)**
   - Publication des vues analytiques.
   - Connexion Power BI en mode DirectQuery ou Import.
5. **Couche gouvernance et sécurité**
   - Apache Ranger : gestion des droits d'accès, masquage et audit.
   - Apache Atlas : catalogue, lignage et documentation.

4.3. Politique de transformation et d'alimentation du système décisionnel

Étape	Nom du job	Outil	Source	Cible	Fréquence	Objectif principal
1	job_extraction_etl	Talend	PostgreSQL / FTP / CSV	/bronze/	Quotidien	Extraction, typage, anonymisation
2	silver_builder.py	Spark (PySpark)	Bronze	/silver/	Quotidien	Jointures, enrichissements, qualité
3	gold_star_model.py	Spark (PySpark)	Silver	/gold/	Quotidien	Modélisation, agrégations, marts
4	refresh_views.sql	Hive / Impala	Gold	Exposition SQL	À la demande	Actualisation des vues d'analyse

4. Développement des jobs d'alimentation du schéma décisionnel

4.4. Exemple de traitement



lien du schéma de l'exemple

5. Description de l'architecture de l'entrepôt de données

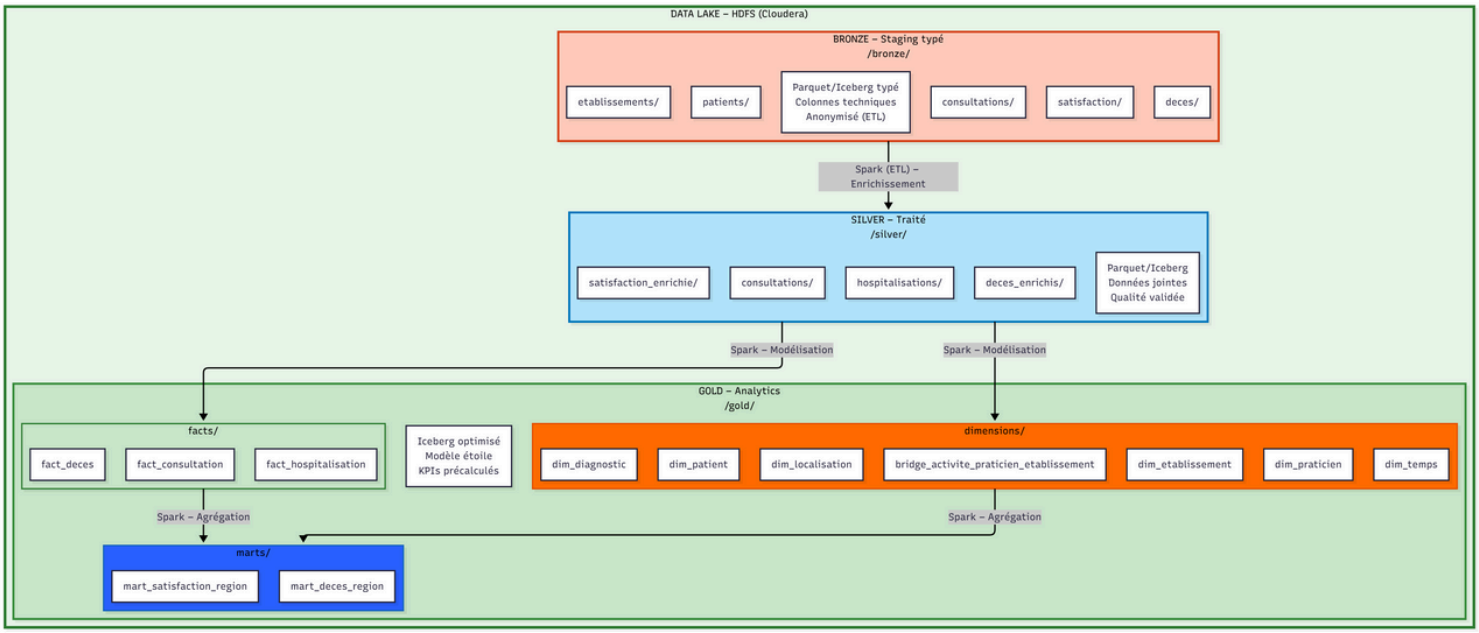
5.1. Présentation générale

L'entrepôt de données du projet Cloud Healthcare Unit (CHU) repose sur une architecture moderne de type Data Lakehouse, déployée sur la Cloudera Data Platform (CDP).

Cette architecture combine la flexibilité du Data Lake, qui permet de stocker de grands volumes de données hétérogènes, et la structuration analytique du Data Warehouse, indispensable à la production d'indicateurs fiables et à la prise de décision.

Elle vise à :

- Centraliser les données issues de multiples sources (PostgreSQL, CSV, FTP).
- Uniformiser, nettoyer et sécuriser les données sensibles (patients, praticiens, établissements, décès, satisfaction).
- Offrir une base analytique cohérente et évolutive pour le reporting et les tableaux de bord (Power BI).
- Garantir la gouvernance, la traçabilité et la sécurité des données conformément aux exigences du RGPD et du secteur médical.



lien du schéma data lake

5. Description de l'architecture de l'entrepôt de données

5.2. Architecture fonctionnelle

L'architecture de l'entrepôt de données s'articule autour de cinq zones logiques correspondant à un pipeline de traitement progressif :

1. Zone de sources

Contient les données brutes provenant de :

- PostgreSQL : base médico-administrative (patients, consultations, diagnostics, professionnels, prescriptions).
- Fichiers CSV : établissements de santé, activité des praticiens.
- Serveur FTP : fichiers d'enquêtes de satisfaction (ESATIS, IQSS) et répertoire des décès.

Ces données sont extraites quotidiennement via des connecteurs Talend et injectées dans la zone de staging.

2. Zone Bronze – Staging typé (/bronze/)

Objectif : préparer et normaliser les données avant intégration.

- Extraction via Talend (connecteurs JDBC, FTP, CSV).
- Conversion des formats (SQL/CSV → Parquet).
- Typage homogène des colonnes (string, date, int).
- Anonymisation des identifiants sensibles (patients, praticiens) avec SHA256.
- Ajout de colonnes techniques : _ingestion_date, _source, _hash_id.
- Stockage dans HDFS (format Parquet, partitions par année/mois/source).

Cette zone correspond aux données brutes typées, conservant leur structure d'origine mais préparées pour l'intégration.

3. Zone Silver – Données traitées (/silver/)

Objectif : fiabiliser et enrichir les données.

Traitements réalisés avec Apache Spark :

- Jointures entre patients, consultations, diagnostics, établissements, satisfaction et décès.
- Nettoyage et dédoublonnage (suppression des doublons et enregistrements incohérents).
- Validation métier : cohérence chronologique (admission < sortie), normalisation des codes diagnostics.
- Enrichissement :
  - Ajout de la dimension géographique (région, département).
  - Ajout de l'âge, de la classe d'âge, du sexe standardisé.
- Vérification de la complétude et qualité métier.

Résultat : Des tables consolidées et propres, prêtes à alimenter la modélisation analytique de la zone Gold.

5. Description de l'architecture de l'entrepôt de données

5.2. Architecture fonctionnelle

4. Zone Gold – Analytique et modélisation (/gold/)

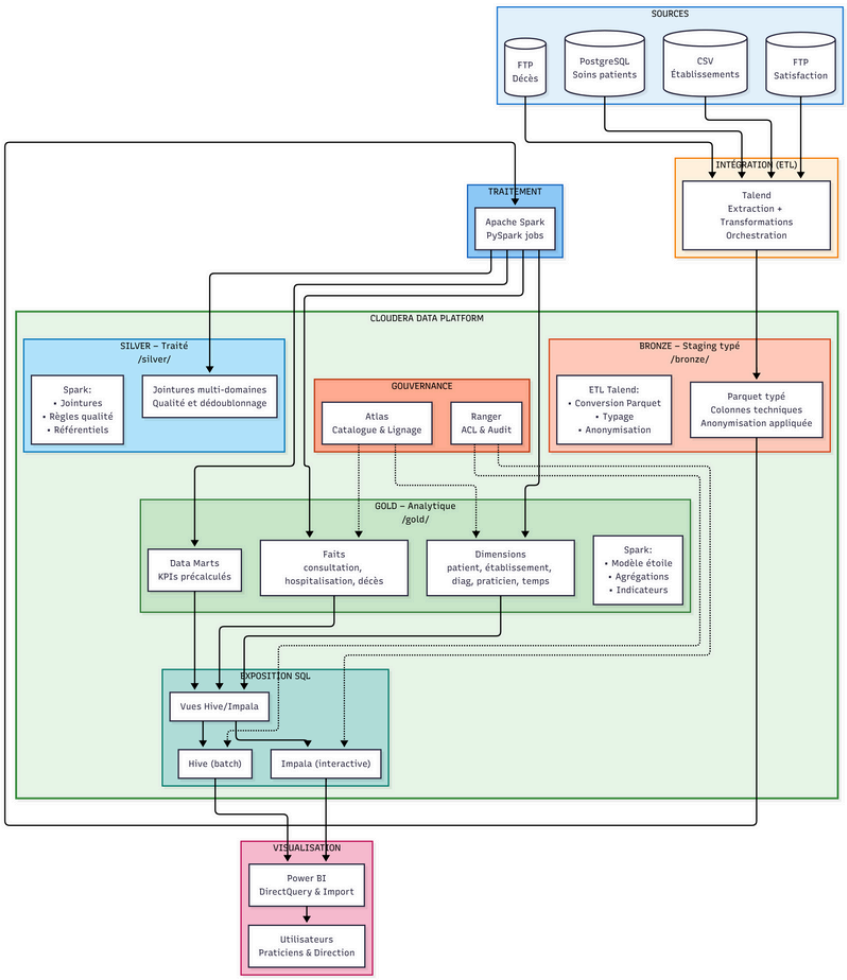
Objectif : construire le modèle décisionnel et les indicateurs métiers.

Cette zone repose sur une modélisation en étoile :

- Tables de dimensions :
  - dim_patient, dim_praticien, dim_etablissement, dim_diagnostic, dim_temps, dim_localisation, bridge_activite_praticien_etablissement.
- Tables de faits :
  - fact_consultation, fact_hospitalisation, fact_deces.
- Data Marts :
  - mart_satisfaction_region, mart_deces_region.

Traitements Spark :

- Création des clés de substitution (surrogate keys).
- Calcul des agrégations et indicateurs (KPI).
- Historisation lente (SCD Type 2 pour les dimensions patients et établissements).
- Optimisation du stockage via Iceberg, partitionnement et Z-ordering.



lien Schema archi total normal

5. Description de l'architecture de l'entrepôt de données

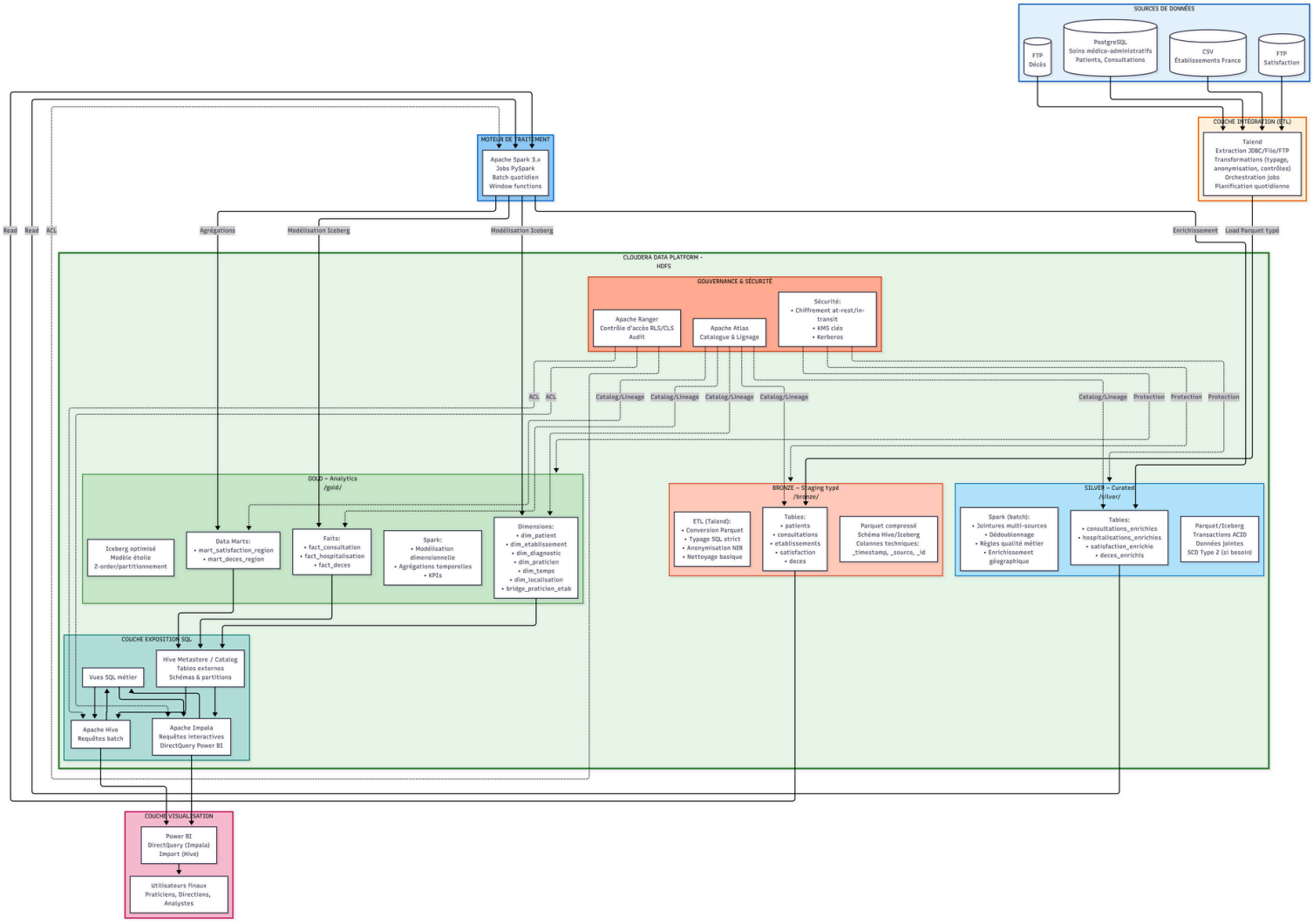
5.2. Architecture fonctionnelle

5. Zone d'exposition SQL et visualisation

Objectif : permettre la consultation et l'analyse des données via des outils de BI.

- Création de vues externes Hive sur les tables Gold (lecture seule).
- Publication via Impala pour les requêtes interactives.
- Connexion directe avec Power BI :
  - Mode DirectQuery : analyses temps réel.
  - Mode Import : analyses historiques volumineuses.

Les utilisateurs finaux (praticiens, responsables hospitaliers, direction) peuvent ainsi explorer les données consolidées et produire des tableaux de bord personnalisés.



lien Schéma archi total détail