05/11/2022

BIG DATA

Livrable 1 : Référentiel de données

MEMBRES U GROUPE :

GUELLATI Mohamed (Chef De Groupe)

ALI ZOUAOUI Zakaria

ZAMOUCHE Nadir

El Mejor Groupe

Table des matières :

Liste des figures

Contexte du projet :

Comme pour l'ensemble du secteur hospitalier, le groupe CHU (Cloud Healthcare Unit) a pris conscience de l'intérêt - voire de la nécessité - d'une transformation digitale majeure. Et votre service est sollicité pour l'aider à mettre en place son propre entrepôt de données qui permettra au groupe d'exploiter la quantité considérable de données générées par les systèmes de gestion de soins ainsi que les systèmes FTP.

Le but est de pouvoir répondre à la variété des besoins et exigences d'accès et d'analyses des utilisateurs. Pour cela, le groupe CHU attend :

- Une solution complète en termes de modèles, d'outils et d'architecture qui permette d'extraire et de stocker les données, pour pouvoir ensuite les explorer et les visualiser suivant différents critères
- Une solution d'intégration de données des fichiers distribués dans une source unique persistante
- Recenser les besoins des utilisateurs (praticiens, chef d'établissement) en termes d'analyse des données pour l'exploitation directe sur le suivi des patients au niveau national et à long terme
- Des préconisations en termes d'outillage d'intégration, de stockage et de logiciel de visualisation adapté ainsi que l'exploration de données en toute sécurité afin de favoriser une meilleure prise de décision.

Introduction :

Dans cette première partie du projet il s'agit de réaliser un modèle conceptuel des données et jobs nécessaires pour alimenter le schéma décisionnel.

Nous allons atteindre cet objectif en utilisant les données disponibles et en répondant au mieux aux besoins des utilisateurs.

Les données :

Il s'agit de données provenant d'historiques des systèmes de gestion des soins (incluant les fichiers de satisfaction et de décès sur FTP). Ils contiennent des données d'exploitation sur plusieurs années. L'infrastructure de données adoptée devra être gouvernée avec une haute sécurité vu la sensibilité des informations qui seront traitées.

Voici un bref descriptif des sources de données mises à disposition :

- Une BDD PostgreSQL qui gère les soins-medico-administratives des patients
- Une BDD exportée en csv sur la gestion des établissements hospitaliers de France
- Des fichiers plats sur des notes de satisfactions émises par des patients sur différents établissements de santé
- Ainsi que des fichiers qui exposent le répertoire de décès en France

Besoins utilisateurs :

Une première consultation a permis de mettre en évidence le type d'analyses souhaitées par les utilisateurs :

- 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/consultation 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

I. A prendre en considération :

Talend : Talend est un ETL (Extract Transform and Load) qui permet d'extraire des données d'une source, de modifier ces données, puis de les recharger vers une destination. La source et la destination des données peuvent être une base de données, un service web, un fichier csv.

Hadoop est un framework open source qui repose sur Java. Hadoop prend en charge le traitement des données volumineuses (Big Data) au sein d'environnements informatiques distribués. Hadoop fait partie intégrante du projet Apache parrainé par l'Apache Software Foundation.

Un job Talend est la représentation graphique d'un ou plusieurs composants reliés entre eux. Il regroupe un ensemble de tâches et permet d'exécuter des processus de flux de données.

Le modèle en étoile est une représentation fortement dénormalisée qui assure un haut niveau de performance des requêtes même sur de gros volumes de données

Les Business Models Talend permettent à toutes les parties prenantes d'un projet d'intégration de données de représenter graphiquement leurs besoins sans avoir à se soucier de leur implémentation technique. Grâce aux Business Models ainsi élaborés, le service informatique de l'entreprise peut ensuite mieux comprendre ces besoins et les traduire en processus techniques. Un Business Model intègre généralement les systèmes et les processus déjà en place dans l'entreprise, ainsi que ceux dont elle aura besoin à l'avenir.

Cloudera : Cloudera est une entreprise Américaine basée en Californie, elle se consacre au développement d'une solution Big Data basée historiquement sur le framework distribué Hadoop et qui est en train de se réorienter vers le Cloud. Cloudera développe depuis plus d'un an, sa solution dans les Cloud publiques AWS, Azure et GCP.

Vm (Machine virtuelle) : Une machine virtuelle, ou « virtual machine », est « le client » créé dans un environnement informatique, « l'hôte ». Plusieurs machines virtuelles peuvent coexister sur un seul hôte.

PostgreSQL est un système de gestion de base de données relationnelle orienté objet puissant et open source qui est capable de prendre en

charge en toute sécurité les charges de travail de données les plus complexes.

II. Modélisation :

Après avoir pris le temps de réfléchir, nous avons choisi la modélisation en étoile pour notre projet de modélisation de données, car elle présentait plusieurs avantages. Le principal avantage est que chaque table de dimension est liée à la table de faits par une seule et unique relation, ce qui facilite grandement les requêtes et améliore leur temps d'exécution. Étant donné que nous travaillons avec de gros volumes de données, c'est un avantage crucial.

Les tables de faits contiennent les données que nous souhaitons afficher dans les rapports d'analyse, sous forme de métriques. Les données de la table de faits sont agrégées à partir des tables de dimensions qui leur sont associées.

Les tables de dimensions sont utilisées pour décrire les données que nous voulons stocker dans notre Data Lake.

Schéma modélisation :

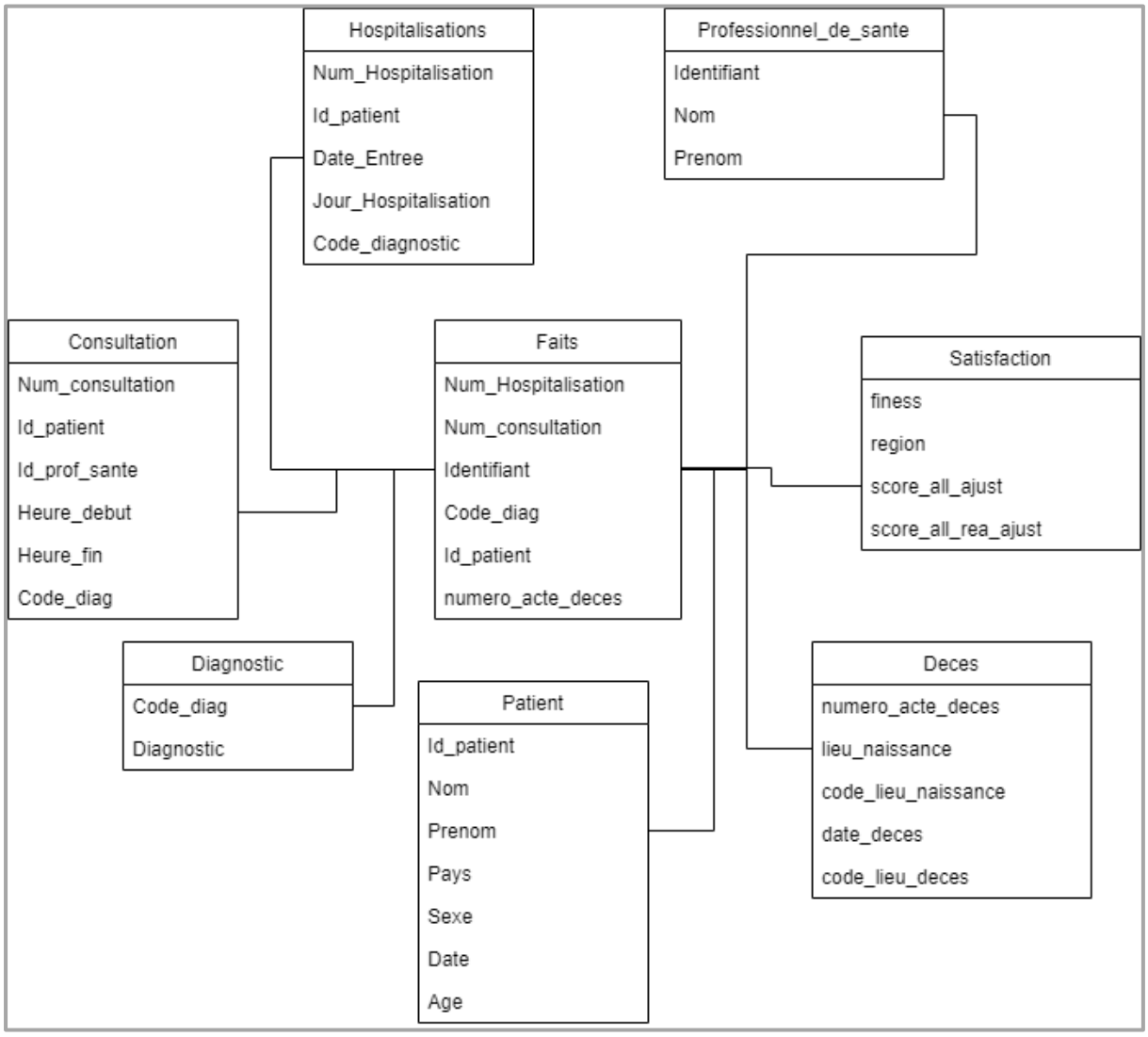

Figure 1 : *Modèle en étoile*

III. Création des jobs et importation dans HDFS

Le Studio Talend est un outil qui va nous permettre de créer un Job en utilisant des composants techniques de la Palette que l'on dépose dans une zone graphique. Ensuite, on relie ces composants entre eux.

Pour sa part, HDFS est un système de fichiers distribué qui gère des volumes de données importants en s'appuyant sur du matériel de base. Il est principalement utilisé pour étendre un cluster Apache Hadoop vers des centaines voire des milliers de nœuds. Parmi les composants principaux d'Apache Hadoop, on trouve HDFS, MapReduce et YARN. Toutefois, il ne faut pas confondre ou remplacer HDFS par Apache HBase, qui est une base de données non relationnelle orientée colonnes, qui s'appuie sur HDFS et qui peut mieux répondre aux besoins des données en temps réel grâce à son moteur de traitement en mémoire.

Job consultation :

On a connecté la base données PostgreSQL dans Talend de ce fait on a créé le job consultation

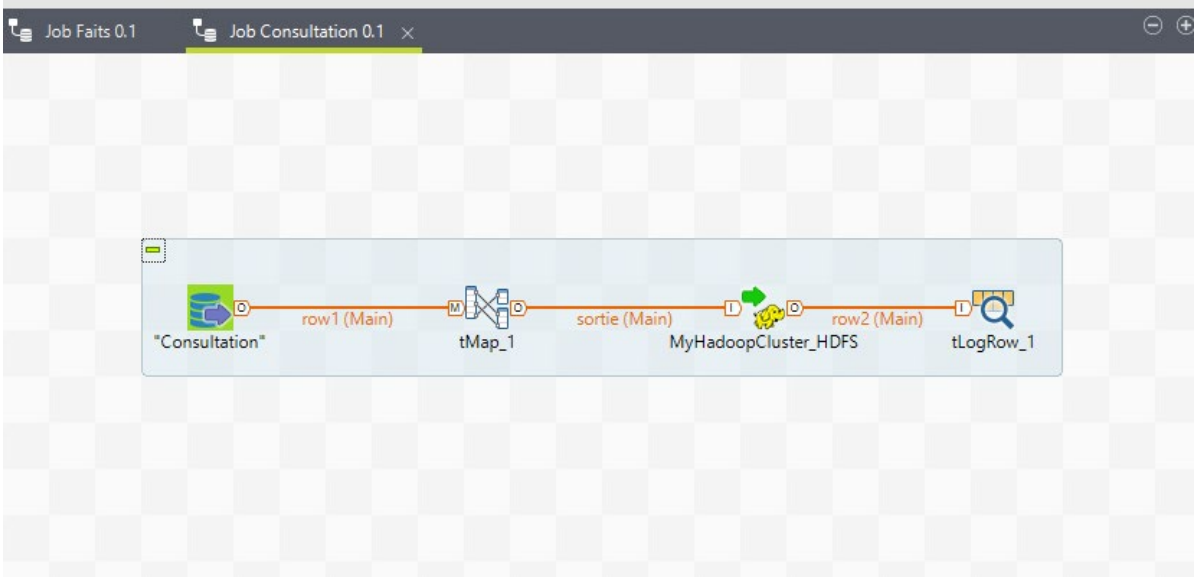

Figure 2 : *Job Consultation*

Job Hospitalisation :

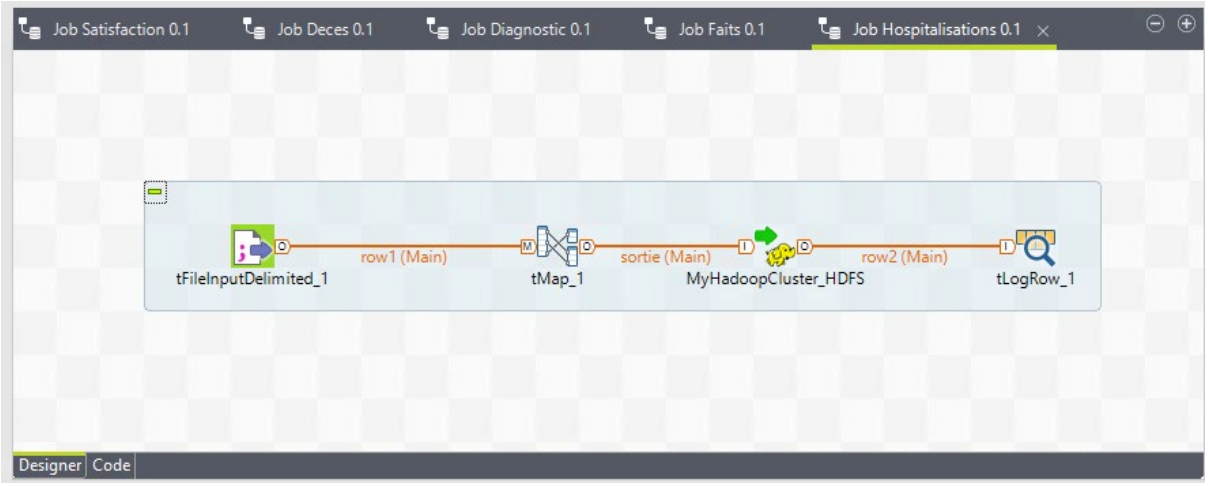

Figure 3 : *Job Hospitalisation*

En utilisant le composant TMap de Talend, nous avons extrait les colonnes "Num_Hospitalisation", "Id_patient", "Date_Entree", « Jour_Hospitalisation » et «Code_diagnostic » à partir du fichier CSV "Hospitalisation". Nous avons ensuite stocké ces données dans le cluster HDFS (Hadoop Distributed File System) afin de les préparer pour les étapes ultérieures de traitement et d'analyse. Cette étape de stockage dans le cluster HDFS nous permet de rendre les informations relatives aux hospitalisations disponibles pour une analyse ultérieure.

Job décès :

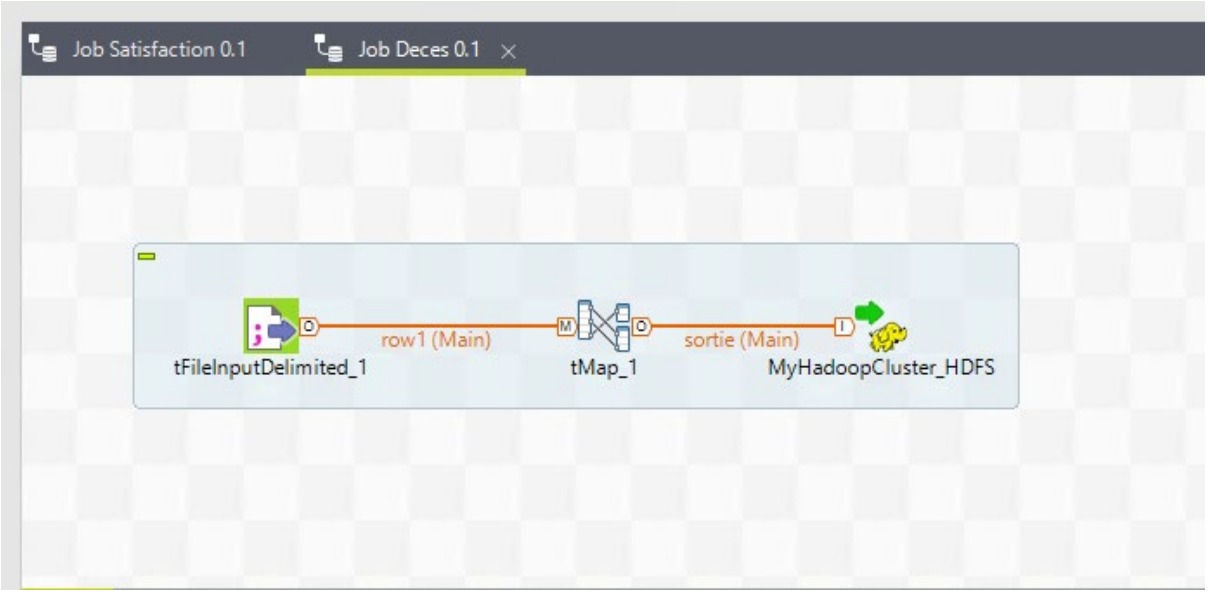

Figure 4 : *Job décès*

Nous avons effectué une analyse du nombre de décès par localisation (région) en 2019 en utilisant le fichier CSV "Décès". À l'aide du composant TMap de Talend, nous avons sélectionné les colonnes pertinentes pour notre analyse, telles que "numero_acte_deces", "lieu_naissance", "code_lieu_naissance", "date_deces" et "code_lieu_deces". Ensuite, nous avons stocké ces données dans le cluster HDFS (Hadoop Distributed File System) pour les rendre accessibles pour les étapes ultérieures de traitement et d'analyse des données. Ce processus nous a permis de préparer les données requises pour calculer le nombre de décès par localisation (région) pour l'année 2019.

Job diagnostic :

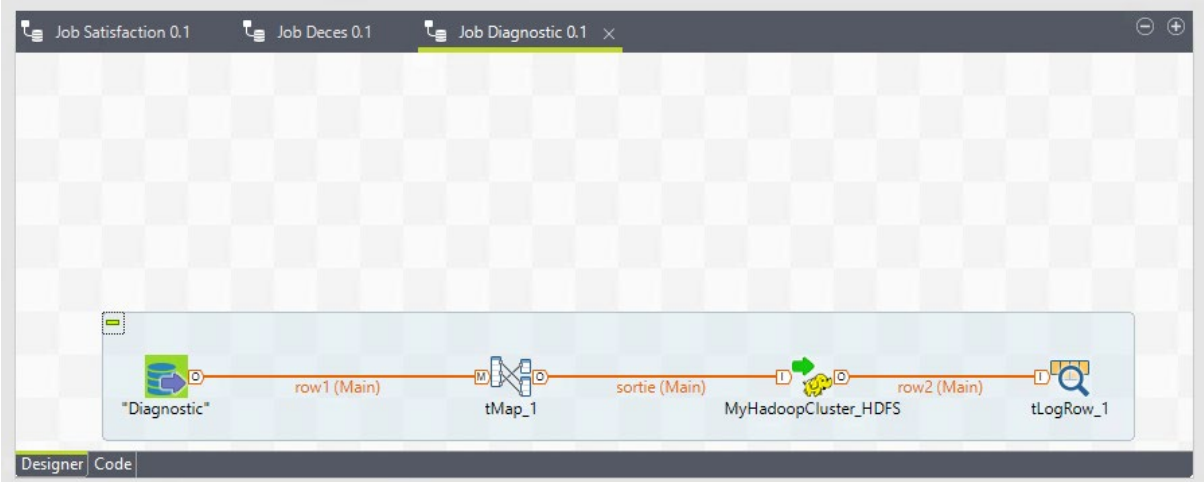

Figure 5 : *Job diagnostic*

On a connecté la base données PostgreSQL dans Talend de ce fait on a créé le job diagnostic.

On a pris la table diagnostic , on l'a mis dans un tmap ensuite on a extrait le « Code_diag » et le « Diagnostic» pour enfin charger la sortie dans le HDFS .

Job Patient :

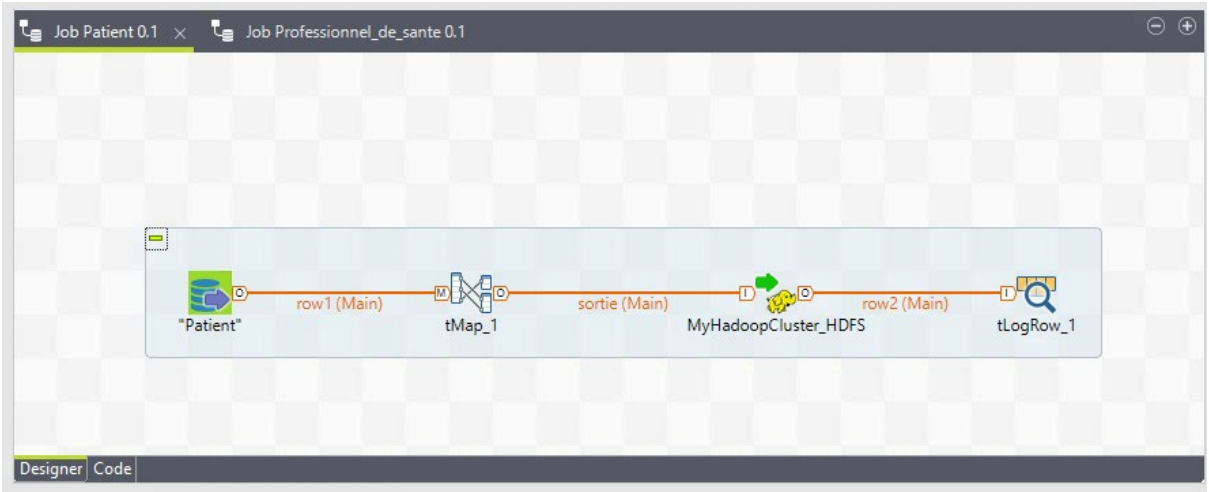

Figure 6 : *Job patient*

Nous avons utilisé la table "Patient" comme source de données pour effectuer cette tâche. En utilisant le composant TMap de Talend, nous avons choisi les

colonnes pertinentes pour notre analyse, à savoir "Id_patient", "Nom ", "Prenom ", "Age", "Sexe", «Adresse » , « Ville », « Code_postal », « Pays », « Email », « Tel », « Date », « Num_Secu », « Groupe_sanguin », « Poid » et « Taille » . Ensuite, nous avons stocké ces données dans le cluster HDFS (Hadoop Distributed File System) pour les préparer en vue des étapes ultérieures de traitement et d'analyse des données. Ce processus nous a permis de préparer les informations relatives aux patients et de les rendre disponibles pour une analyse ultérieure.

Job Professionnel santé :

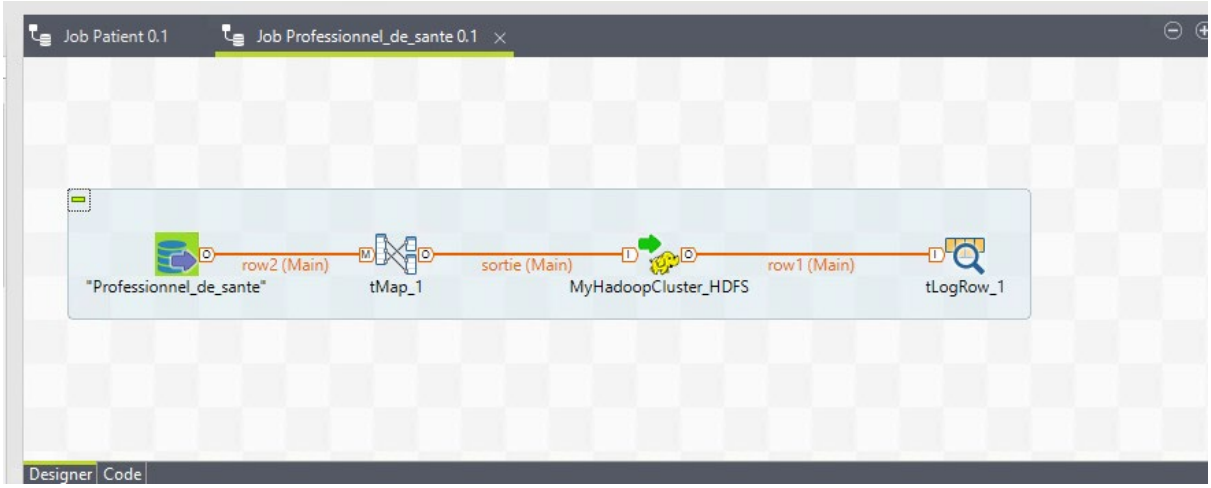

Figure 7 : *Job professionnel sante*

Pour cette tâche, nous avons utilisé la table "Professionnel" comme source de données. À l'aide du composant TMap de Talend, nous avons choisi les colonnes pertinentes pour notre analyse, à savoir "Identifiant", « Civilite », « Categorie_professionnel » "Nom", "Prenom" et "Profession", « Type_identifiant », « Code_specialite » . Ensuite, nous avons stocké ces données dans le cluster HDFS (Hadoop Distributed File System) pour les préparer en vue des étapes ultérieures de traitement et d'analyse des données. Ce processus nous a permis de préparer les informations relatives aux professionnels et de les rendre accessibles pour une analyse ultérieure.

Job satisfaction :

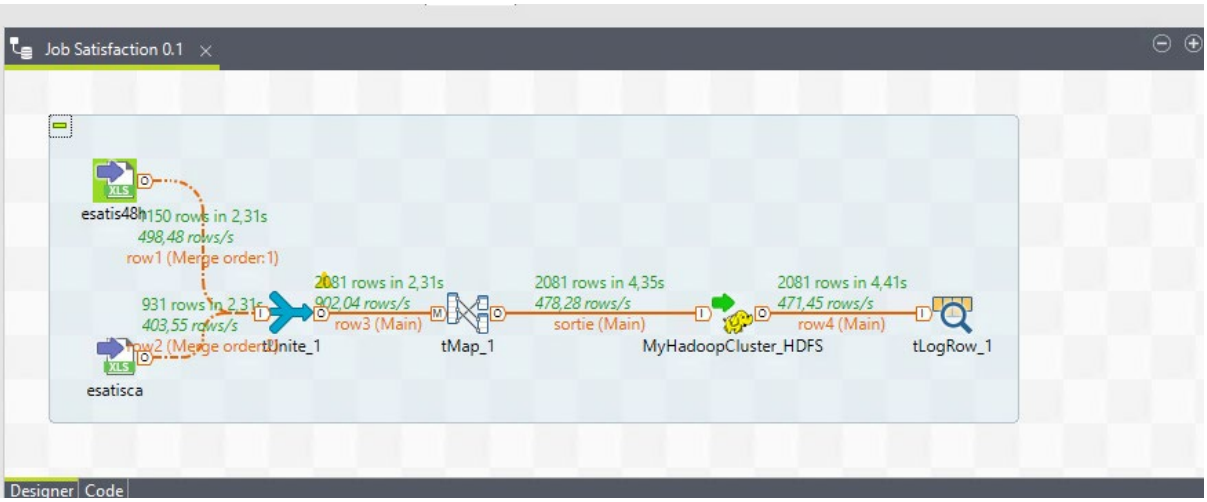

Figure 8: *Job satisfaction*

Nous avons créé un job nommé "satisfaction" dans Talend, en important les fichiers XML de satisfaction. Pour ce faire, nous avons utilisé deux tables : "finess", "region", « score_all_ajust », « score_all_rea_ajust » que nous avons regroupées dans un composant TMap. À partir de ce TMap, nous avons extrait les informations suivantes : l'ID de satisfaction, la région, le code Finess, le score total ajusté et le score ajusté pour la réanimation. Enfin, nous avons chargé les résultats de cette extraction dans le HDFS.

IV. Création de la table des faits

Nous avons utilisé les jobs précédemment créés pour récupérer les IDs correspondant aux différentes mesures, puis nous les avons stockés dans le cluster HDFS. Ainsi, nous avons créé une source de données centrale pour des analyses ultérieures ou des intégrations avec d'autres outils.

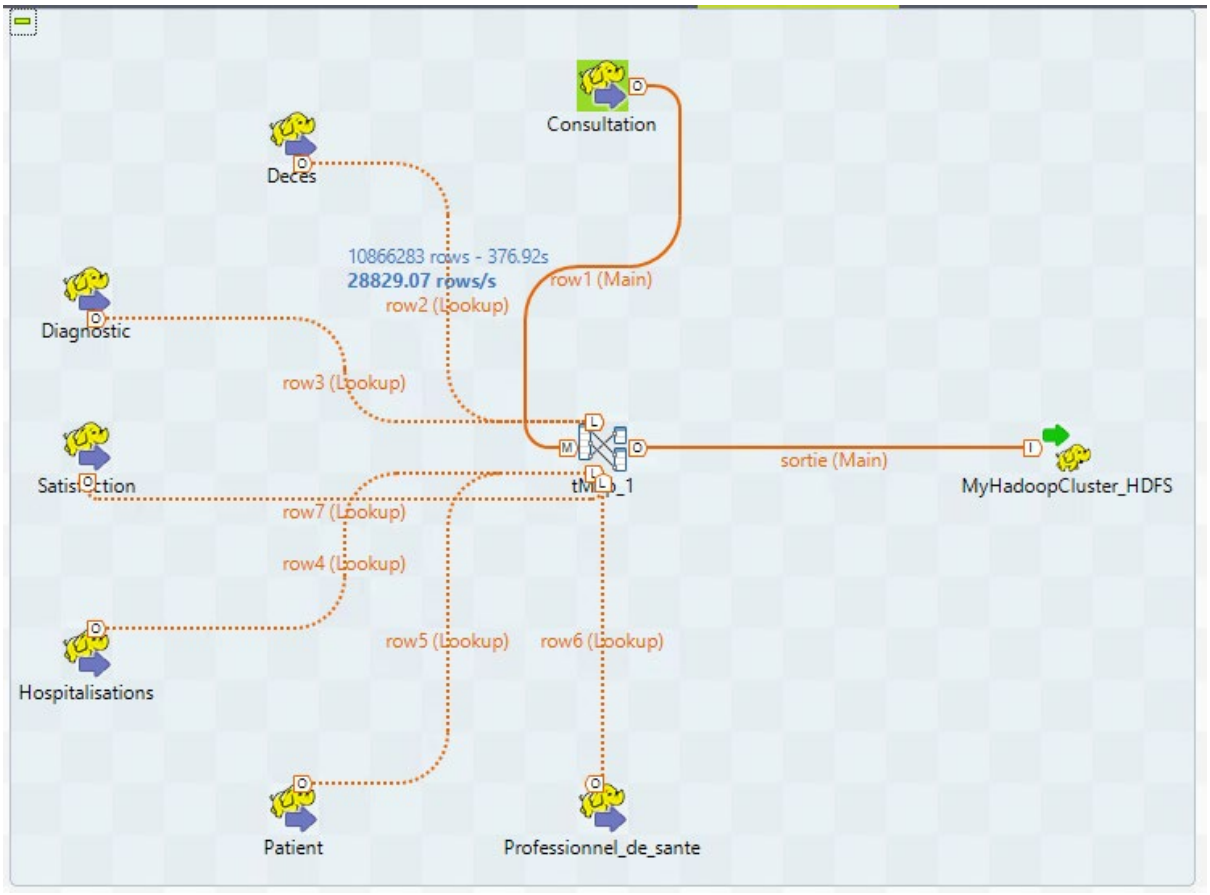

Figure 9 : *Table des faits*

Nous avons regroupé les différentes dimensions à l'aide d'un TMap, extrait les clés correspondantes, puis stocké la sortie dans le HDFS.

V. Architecture business model

Concevoir des Business Model fait partie des bonnes pratiques à adopter en entreprise à un stade très précoce d'un projet d'intégration de données afin d'en assurer le succès. Les Business Model permettent généralement d'identifier et de résoudre rapidement les goulots d'étranglement et autres points faibles du projet à mettre en place, ainsi que de limiter les dépassements de budget, voire de réduire l'investissement initial. Puis, pendant et après la mise en place du projet, les Business Model peuvent être revus et corrigés, si besoin est.

Un Business Model est une vue non technique d'un besoin métier de gestion de flux de données.

Généralement, un Business Model intègre en premier lieu les systèmes stratégiques et étapes d'exécution déjà opérationnels au sein d'une entreprise. Ces systèmes, connexions et autres besoins sont symbolisés dans la perspective Intégration du Talend Open Studio par de multiples formes et liens disponibles dans la Palette. Ils peuvent tous être facilement décrits en utilisant les attributs et outils de formatage du Repository.

Dans l'espace de modélisation graphique de la perspective Intégration du Studio Talend, on dispose de nombreux outils nous permettant de :

- Modéliser nos besoins métier,
- Créer des éléments dans le référentiel de métadonnées et les attribuer à vos objets de Business Model,
- Définir les propriétés d'apparence de vos objets de Business Model.

Voici notre architecture :

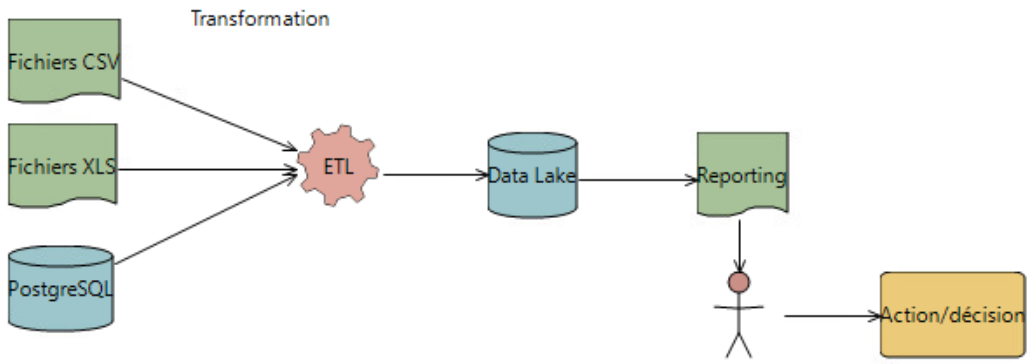

Figure 10 : *Business model*

VI. Conclusion

En résumé, nous avons achevé la conception du modèle conceptuel de données ainsi que la création des jobs nécessaires pour alimenter notre schéma décisionnel. Nous allons maintenant passer à l'étape suivante qui consiste à réaliser le modèle physique de données et à évaluer les performances en termes de temps de réponse pour les requêtes effectuées sur les tables.