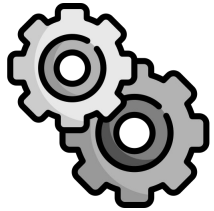
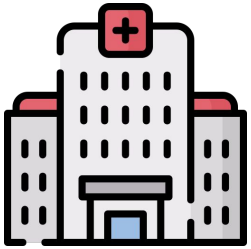


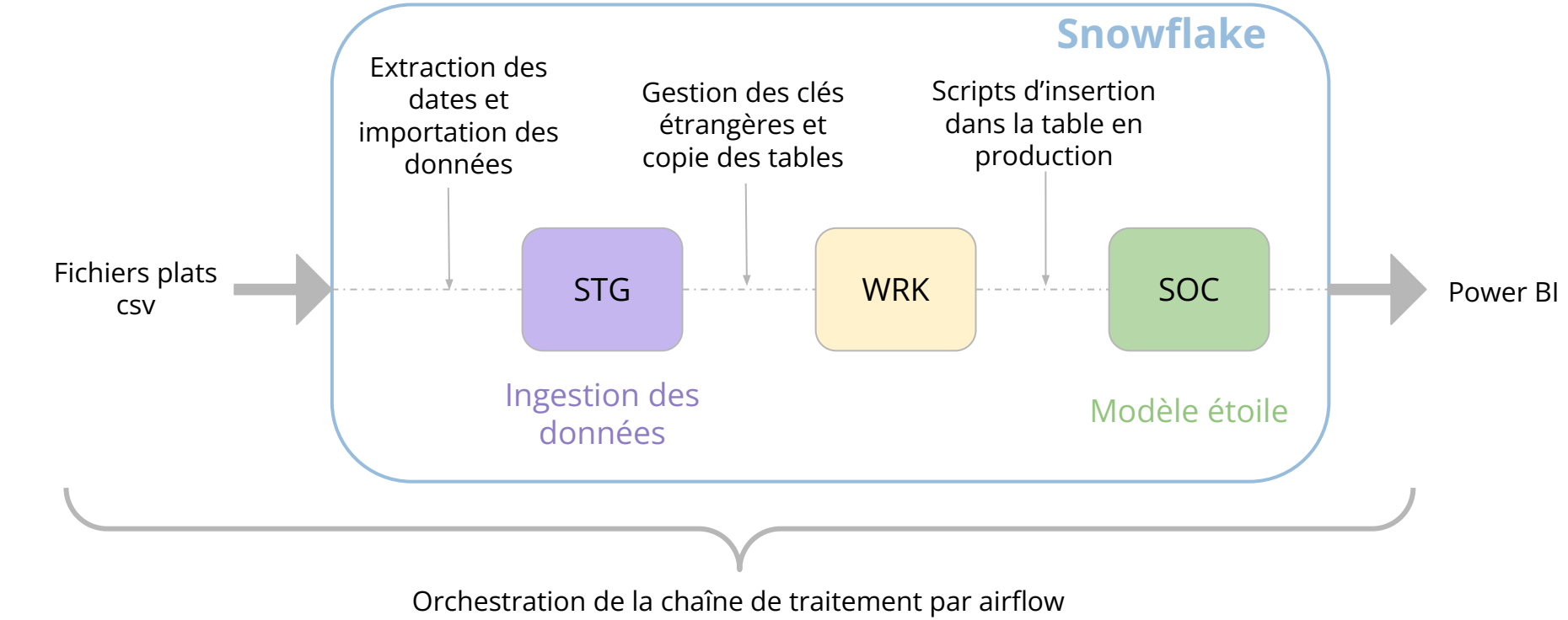
Projet NF26 en collaboration avec Smart Teem

Mise en place d'une solution décisionnelle



Introduction

Objectif



Introduction

Gestion de projet

Organisation 

Jira

Microsoft Teams

Développement 

Visual Studio Code

snowflake®

GitHub

Présentation des données

Modèle physique de données du datawarehouse

Données

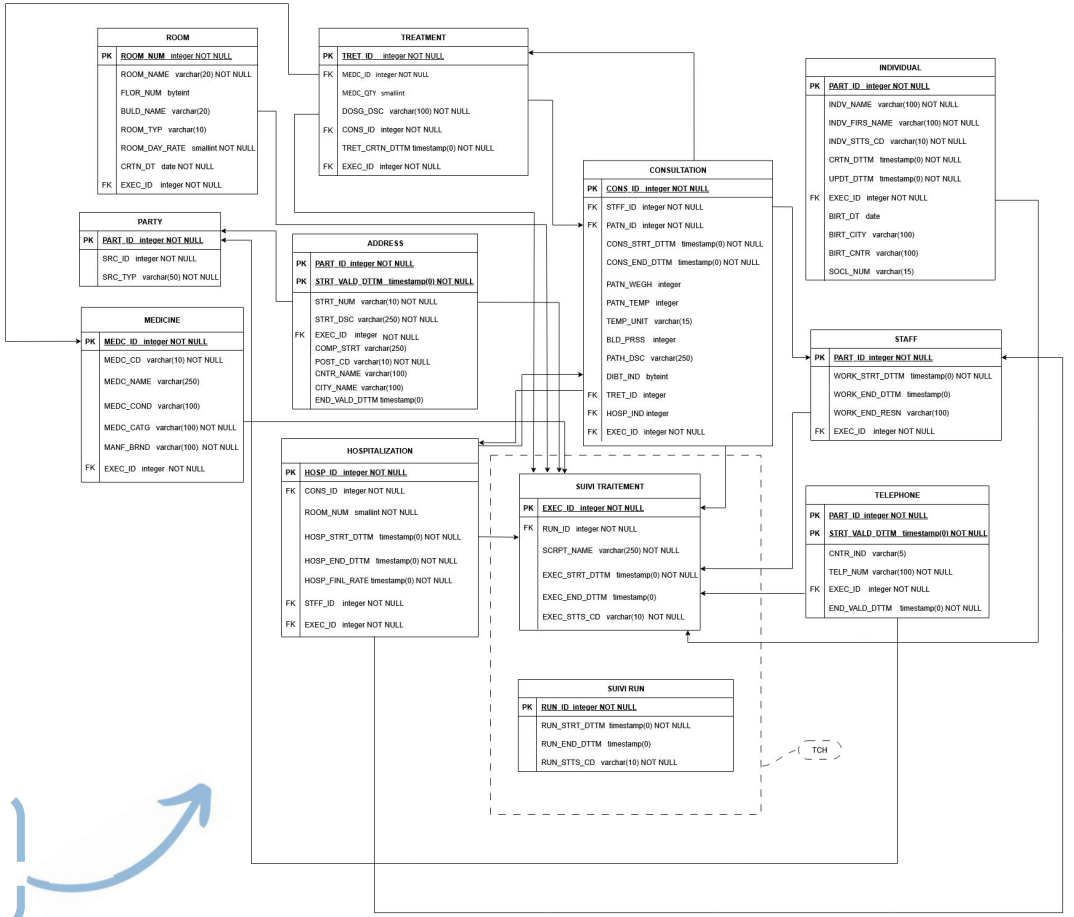
Fichiers .txt journaliers décrivant :

- Chambre,
- Consultation,
- Hospitalisation,
- Médicament,
- Patient,
- Personnel,
- Traitement.

Type de données : nombre, chaîne de caractère, date …

Date dans un format non reconnu par Snowflake

Création d'un modèle physique basé sur les documents fournis



Installation du SID

Création des bases

Appel du script create_database.sql

- ➔ Les bases ne sont pas recréées si elles existent déjà

install_sid.py

- ➔ **Exécution** lancée une seule fois par Airflow
- ➔ **Lancement** des scripts SQL de création des tables

install_sid.log

- ➔ **Exécutions tracées**

Création des tables

A l'aide de l'excel "Hopital Mapping VF.xlsx" :

1. **Création** des tables dans STG (recréés à chaque exécution)
2. **Création** des tables dans SOC (créés une seule fois)
3. **Création** des tables dans WRK (recréés à chaque exécution)
   a. SOC + STG

Ingestion des données

Outil

Script de création de procédures

insert_STG_procedure.sql

- **Instanciation** par airflow pour définir des fonctions d'ingestion simultanée de plusieurs lignes
- **1 procédure** par table de STG
- Permet le **pré-traitement** des données (remplacement des "#" dans les dates par des 0, ajout de valeurs NULL si données d'entrée vide)

Utilise

Dag pour la pipeline ETL

Script d'insertion périodique

launch_load_sid.py

- Permet d'orchestrer la **pipeline de l'ETL** sur un jour donné : **récupération** des données, **transformations** et **ingestion** dans le modèle en production (SOC)
- Vide les tables de **STG** et **WRK** à chaque fois pour éviter les doublons
- Charge les donnés depuis le csv avec un **script python** puis lance les scripts d'alimentation du datawarehouse avec des **requêtes SQL sur SnowFlake**

Alimentation d'un datawarehouse

STG → WRK

1. **Sélection** des données de STG
2. **Insertion** dans WRK

WRK → SOC

1. **Sélection** des données de WRK (SOC)
2. **Suppression** des doublons dans SOC coïncidant (en clé primaire) avec WRK
3. **Insertion** dans SOC

WRK (STG) → WRK (SOC)

1. **Sélection** des données de WRK (STG)
2. Respect des règles de gestion du fichier Excel
    a. **Jointures** et **Restrictions**
    b. **Surrogate Key** : Séquence incrémentale selon les valeurs uniques de certains attributs
        i. Window function avec ROW_NUMBER()
3. **Insertion** des données dans WRK (SOC)

Airflow

- Permet l'**automatisation journalière** des tâches pour l'ingestion et le traitement des données

- Connexion possible avec **Snowflake**
  → permet l'injection des données en base depuis Airflow

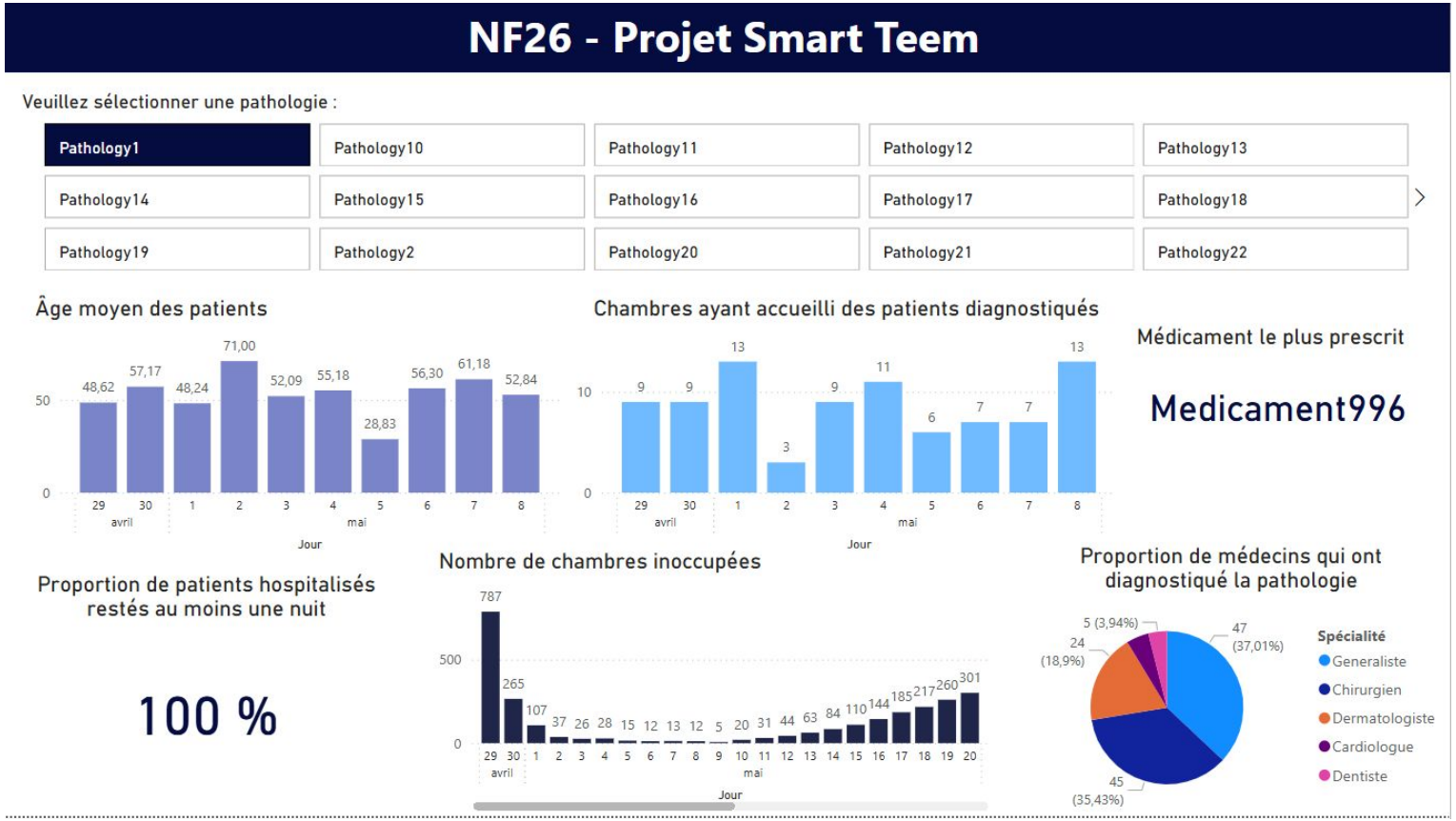
- Utilisation des **DAGs** dans le cadre du projet :
  → un DAG pour la **création** des tables
  → un DAG pour **l'injection** et le **traitement** journaliers des données

- **Gestion des logs** :
  → Automatique
  → Détaillés
  → Pour chaque batch et chaque tâche

DAG (Directed Acyclic Graph)

Graphe orienté et acyclique utilisé pour modéliser des flux de travail

Power BI


NF26 - Projet Smart Teem
Veuillez sélectionner une pathologie :
Pathology1
Pathology10
Pathology11
Pathology12
Pathology13
Pathology14
Pathology15
Pathology16
Pathology17
Pathology18
Pathology19
Pathology2
Pathology20
Pathology21
Pathology22
Âge moyen des patients
48,62
57,17
48,24
71,00
52,09
55,18
28,83
56,30
61,18
52,84
50
0
29 30 1 2 3 4 5 6 7 8
avril mai
Jour
Chambres ayant accueilli des patients diagnostiqués
9 9 13 3 9 11 6 7 7 13
10
0
29 30 1 2 3 4 5 6 7 8
avril mai
Jour
Médicament le plus prescrit
Medicament996
Proportion de patients hospitalisés restés au moins une nuit
100 %
Nombre de chambres inoccupées
787 265 107 37 26 28 15 12 13 12 5 20 31 44 63 84 110 144 185 217 260 301
500
0
29 30 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
avril mai
Jour
Proportion de médecins qui ont diagnostiqué la pathologie
5 (3,94%)
47 (37,01%)
24 (18,9%)
45 (35,43%)
Spécialité
Generaliste
Chirurgien
Dermatologiste
Cardiologue
Dentiste

Power BI

Requêtes

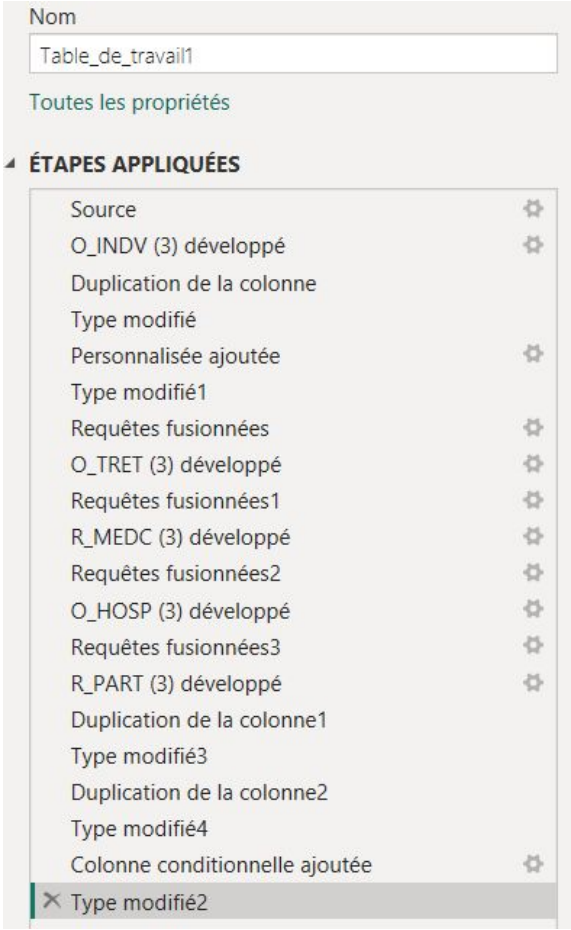
- Transformations de type
- Jointures
- Colonnes calculées (ex : Âge)
- Colonnes conditionnelles (ex : aPasséNuitHosp)

Mesures (DAX)

- Chambre inoccupées
- Médicament le plus prescrit
- Proportion de patients hospitalisés restés au moins une nuit

Tables ajoutées

- Table de travail (issue des requêtes)
- Calendrier (générée depuis Modèle)
- OccupationJourChambre (générée depuis Modèle)



Conclusion

- Apprentissage d'Airflow pour l'orchestration des traitements ETL et Snowflake pour la gestion et le stockage des données

- Gestion de l'entièreté de la chaîne de traitement des données, des fichiers bruts à l'exploitation des données propres avec Power BI

- Acquisition de compétences techniques en Data Engineering ainsi que de compétences en travail d'équipe et en collaboration professionnelle avec l'entreprise Smart Teem