

	Université de Corse - Pasquale PAOLI	
	Diplôme : M1 DE et DFS	2025-2026
	EC : Datawarehouse TP3 – Data warehouse et analyses décisionnelles Analyse des données de la recherche Enseignant : Evelyne VITTORI	

Dans la première partie du projet, vous avez conçu une base relationnelle PostgreSQL décrivant l'écosystème de la recherche et les métadonnées des projets, contrats, publications et jeux de données.

Dans la deuxième partie, vous avez construit une base NoSQL MongoDB permettant de stocker les données effectives des jeux de données scientifiques (fichiers, mesures, logs, versions, annotations).

Dans cette troisième partie, vous allez construire un Data Warehouse alimenté par un processus ETL (Extract, Transform, Load) défini sous Pentaho Data Integration (PDI). Le Data Warehouse servira ensuite à construire un cube OLAP sous icCube afin de répondre à des questions analytiques sur l'activité de recherche.

Le Data Warehouse doit être alimenté par au minimum une **source** :

- la **base relationnelle PostgreSQL** (métadonnées : projets, contrats, publications, datasets),
- la **base NoSQL MongoDB** (contenus effectifs et métadonnées techniques des datasets),
- ou un **fichier CSV** issu d'une plateforme d'open data (optionnel).

Objectifs Clés du TP

1. **Définir un objectif analytique** simple à atteindre grâce aux données du Data Warehouse.
2. **Concevoir le schéma du Data Warehouse** sous forme de schéma en étoile comprenant une table de faits et des tables dimensions pertinentes (2 dimensions)
3. **Créer un processus ETL** sous Pentaho DI pour intégrer les données des bases PostgreSQL, MongoDB ou CSV dans le Data Warehouse.
4. **Construire un cube OLAP** sous icCube pour répondre à l'objectif analytique défini, en exploitant les données intégrées.
5. **Réaliser deux requêtes analytiques avancées** en MDX en lien avec les objectifs définis pour explorer les données agrégées.

Critères d'évaluation

1. Pertinence de l'objectif analytique défini en Partie 1.
2. Qualité du schéma du Data Warehouse et justifications.
3. Cohérence et efficacité du processus ETL.
4. Pertinence du cube OLAP et des requêtes proposées.
5. Compréhension globale de la démarche.

Rendus

Date rendu : à préciser

Ce travail doit être réalisé en groupes de 4 étudiants.

Le rendu sera effectué dans l'onglet travaux sur l'ENT sous la forme d'un **seul fichier archive (zip ou rar)** contenant l'ensemble des fichiers demandés :

- Rapport synthétique contenant :
 - Description détaillée de vos objectifs d'analyse.
 - Schéma de votre Data warehouse avec justification de vos choix
 - Description de votre processus ETL avec captures d'écran des étapes sous pentaho.
 - Description et justification du cube OLAP créé.
 - Bilan personnel critique sur la solution proposée
- Scripts SQL et MDX
 - Création du schéma du DW
 - Requêtes analytiques associées au cube OLAP

Soutenance Orale

- Date prévue : vendredi 9 janvier 2026
- Présentation de 10 minutes avec démonstration

Partie 1 - Définition de l'objectif d'analyse

Avant toute implémentation technique, vous devrez définir un **objectif analytique** précis : Quelle question de pilotage ou d'évaluation scientifique souhaitez-vous poser ?

Exemples :

- *Quels laboratoires produisent le plus de datasets déposés?*
- *Quel est le délai moyen entre la collecte de données et leur dépôt par discipline ?*
- *Quelle volumétrie totale de données (octets) est produite par type de format ?*

Vous choisirez et formulerez clairement votre **objectif analytique** sous la forme d'un ensemble de questions.

Partie 2 - Conception du Data Warehouse

Concevez un **schéma en étoile** pour structurer les données dans le Data Warehouse en fonction de vos objectifs.

Ce schéma devra inclure au moins :

- Une **table de faits** qui contiendra les informations clés permettant de répondre aux objectifs analytiques définis ex. F_Fichier, F_Publication, F_Dataset,..
- Deux **dimensions** pour fournir des axes d'analyse (temps, Projet, Discipline, Laboratoire, Format, etc..)

Implémentez ensuite ce schéma sous PostgreSQL.

Partie 3 - Cration du processus ETL avec Pentaho DI

Définissez et implémentez un **processus ETL sous Pentaho DI** pour extraire les données de PostgreSQL et MongoDB (et/ou d'éventuelles autres sources externes), les transformer et les charger dans le Data Warehouse.

Partie 4 - Cration d'un cube OLAP et Requetes analytiques

- Définissez un **cube OLAP** sous IcCube, structuré en fonction des objectifs analytiques définis en phase 1.
 - Créez des **mesures** et des **dimensions** pour permettre des analyses croisées pertinentes.
 - Définissez au moins **2 requêtes analytiques** exploitant le Data Warehouse et le cube OLAP. Ces requêtes doivent être en lien direct avec les objectifs définis et inclure des calculs d'agrégation.

Exemple de requêtes :

- classer les projets selon le délai moyen entre création et dépôt..
 - obtenir des totaux par laboratoire, par discipline et par format de fichier, en une seule requête.
 - croiser deux dimensions (laboratoire \times type de fichier).