Ecole Nationale des Sciences Appliquées d'Al Hoceima Université Abdelmalek Essaadi

ETAT D'AVANCEMENT - PROJET ANALYSE DE RECHERCHE MAROCAINE À L'ÉCHELLE MONDIALE

Membres de group:
BACHIRI JAWAD
BOUKAYOUA LOUBNA
EL BOUTAHERI NAJMA

Encadré par :
Professeur: Anass El heddadi

Année Universitaire: 2024/2025

Table de contenu:

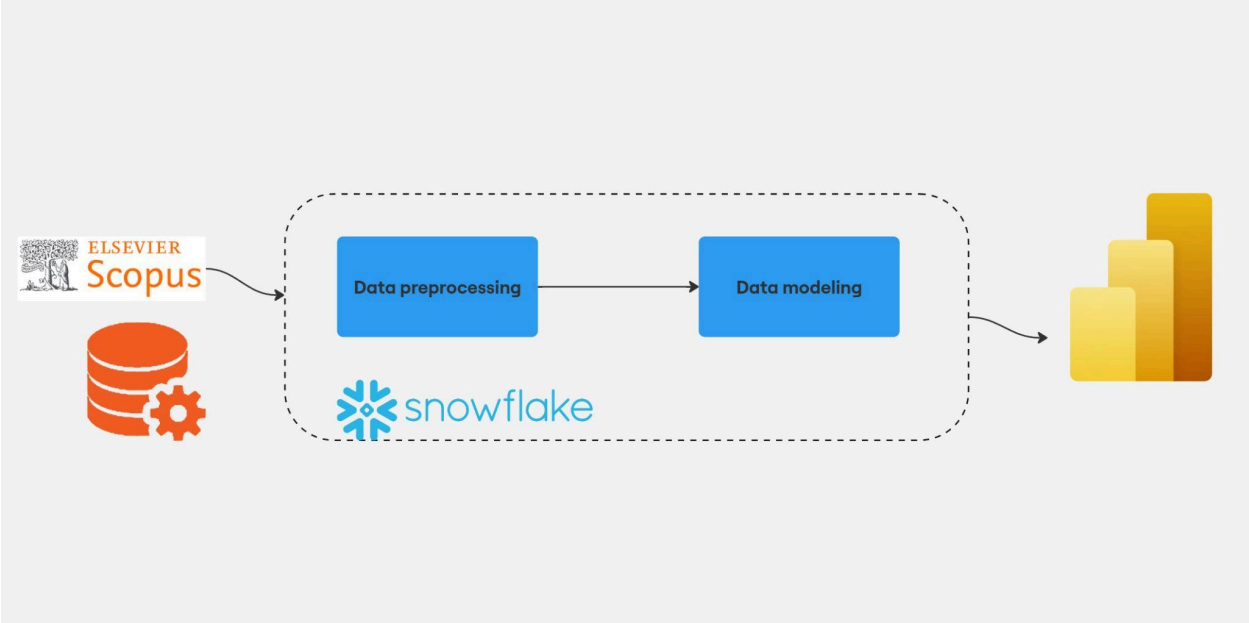
Introduction:

Le projet consiste en la création d'un pipeline d'analyse de données intégrant la plateforme Snowflake et des outils de visualisation avancés. L'objectif est de transformer des données brutes provenant de bases comme Elsevier Scopus en visualisations exploitables via Power BI. Ce rapport présente l'architecture mise en place, le traitement des données, et l'avancement dans la création du modèle de données.

1. Architecture de projet:

L'architecture du projet repose sur une chaîne de traitement robuste :

- **Source des données :** Les données brutes sont extraites de la plateforme Elsevier Scopus.
- **Stockage et traitement :** Snowflake est utilisé pour stocker et traiter les données.
- **Pré Processing et modélisation :** Les données sont prétraitées avant d'être modélisées dans un format exploitable.
- **Visualisation :** Les résultats sont visualisés à l'aide de Power BI.



2. Traitement de données:

Pour garantir la validité et le nettoyage des données, nous avons suivi plusieurs étapes appliquées à la majorité des colonnes. Les principales actions réalisées sont les suivantes :

1. **Élimination des valeurs nulles** : Suppression des entrées contenant des valeurs manquantes.
2. **Élimination des doublons** : Identification et suppression des lignes en double dans les données.
3. **Transformation des types de données** : Conversion des types de données pour certaines colonnes afin de garantir leur cohérence.

En plus de ces étapes générales, un traitement spécifique a été appliqué à la colonne **"Affiliation"**, car elle nécessitait une attention particulière. Les étapes spécifiques pour cette colonne sont décrites ci-dessous :

1. **Extraction des affiliations marocaines** : Filtrage des affiliations pour conserver uniquement celles ayant le pays égal à "Maroc".
2. **Identification des universités** : Extraction des noms d'universités présents dans les affiliations, en tenant compte du manque d'uniformité dans leur représentation. Pour cela, nous avons créé un dictionnaire contenant plusieurs synonymes du mot "université" dans différentes langues.
3. **Traitement des affiliations non marocaines** : Filtrage des données pour extraire les affiliations où le pays est différent de "Maroc".
4. **Réexécution de l’étape 2** : Application de l'extraction des universités sur les affiliations non marocaines.
5. **Gestion des affiliations inconnues** : Attribution de la valeur "unknown" à toutes les affiliations non reconnues ou manquantes.

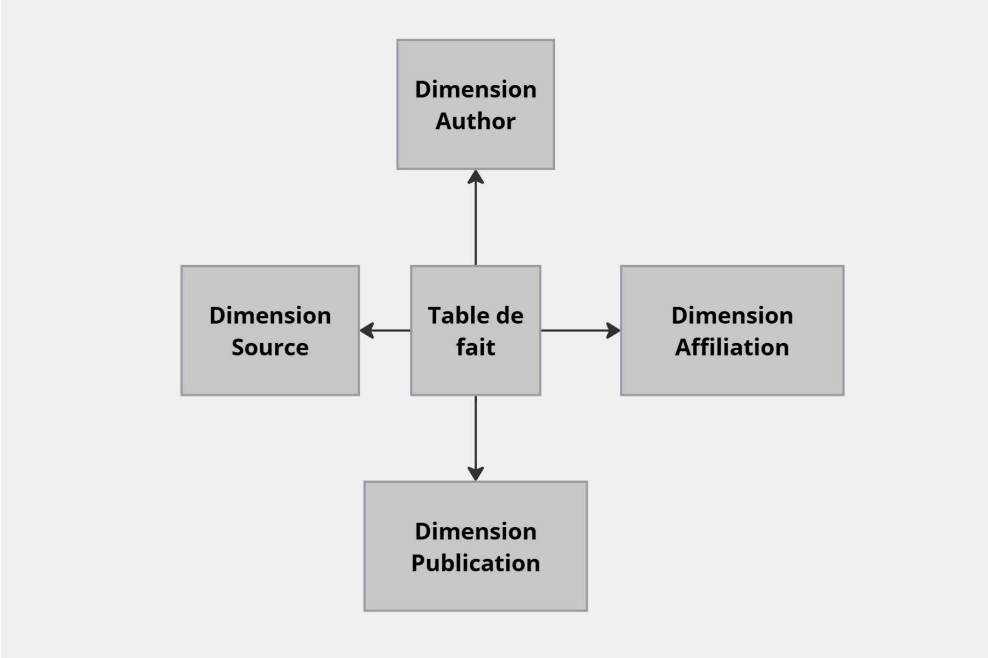
Ces étapes ont permis d'assurer un traitement rigoureux et homogène des données d'affiliation.

3. Création d’un modèle de données:

Le modèle de données repose sur une conception centrée sur les relations entre faits et dimensions :

- **Table de faits :** La *Bridge-table* contient les clés étrangères des dimensions ainsi que les mesures suivantes :
    1. Nombre de publications.
    2. Nombre d’auteurs par publication.
    3. Nombre d’affiliations associées.
- **Tables de dimensions :**
    1. **Table Author** :
    2. **Table Affiliation** :
    3. **Table Source** :
    4. **Table Publication** :

Modèl de données:



conclusion:

L’implémentation de la table de faits et des tables de dimensions est bien avancée. Les données brutes ont été intégrées avec succès dans Snowflake, permettant ainsi de construire une table de faits (Bridge-table) ainsi que des tables de dimensions fonctionnelles. Ces éléments assurent une base solide pour l’analyse et la visualisation des données. Par ailleurs,

les connexions avec Power BI ont été établies, offrant la possibilité de réaliser des visualisations de base. Les prochaines étapes consistent à enrichir les métriques dans la Bridge-table afin de permettre des analyses plus détaillées et granulaires. Il est également prévu d'étendre les informations dans les tables de dimensions pour inclure des données supplémentaires, telles que les citations et les collaborations internationales. Enfin, le travail se concentrera sur la finalisation des tableaux de bord Power BI en y intégrant des visualisations avancées pour maximiser la valeur analytique du projet.