Cours : **Architecture Bigdata**

Projet M2

Durée : 1 mois

Date de début : 11/03/2025

Date de fin : 11/04/2025

Rendu :

- ***Un document PDF présentant tout le projet***
- ***Un fichier PowerPoint de présentation***
- ***Une vidéo montrant les composants et le process***

En groupe de 3

Barème

Présentation des documents	Déroulés présentation vidéo	Contenu du projet	Touche personnelle
2	3	12	3

Enoncé :

Soit une structure disposant d'un système de vente des livres.

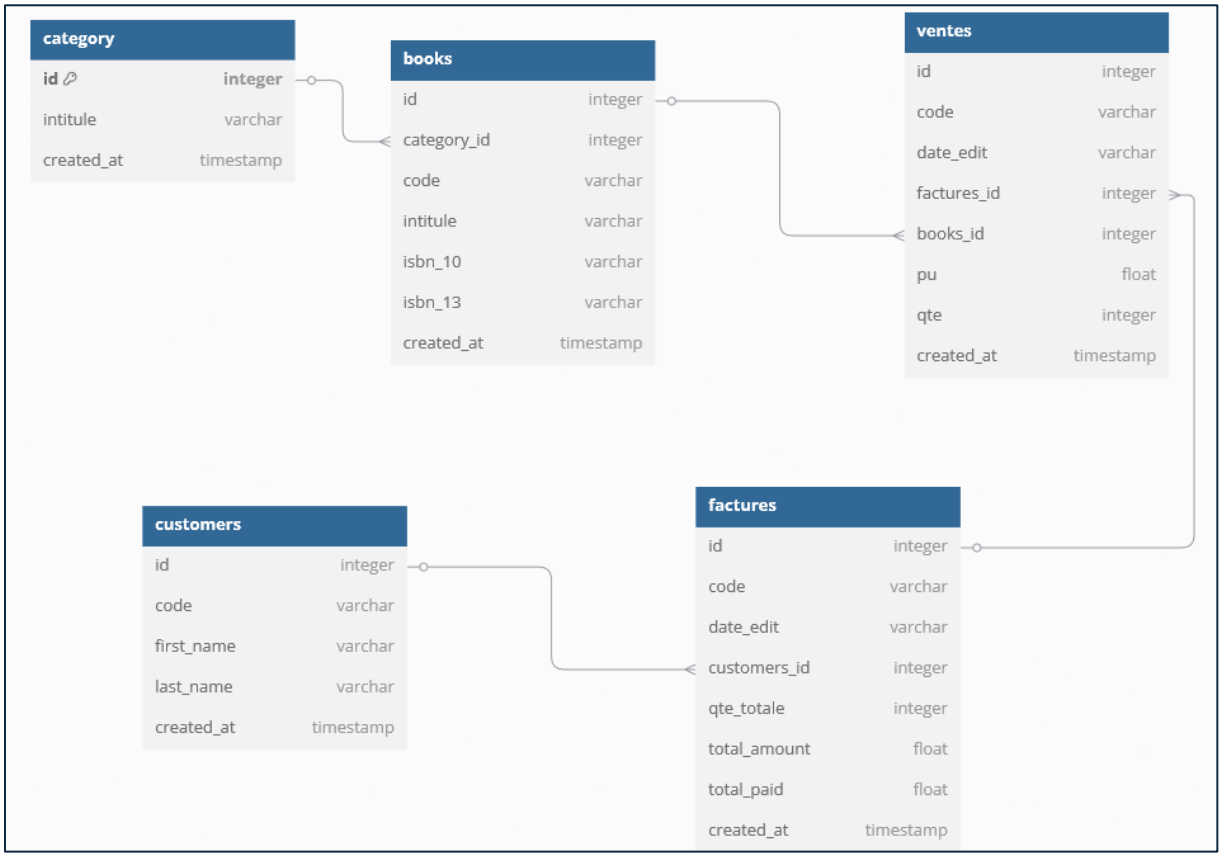
L'objectif est de faire les transformations nécessaires afin de produire des OBT des ventes.

NB : ***les composants suivants ne sont pas obligatoires. Dans le cas d'utilisation d'un autre outil ou composant, préciser la raison.***

Composants :

- **Datawarehouse** :
  - **SNOWFLAKE**
  - Cloud : **Free trial (1 mois) ou payant**
  - Contiendra les données **brutes, stagging, warehouse et marts (OBT)**
  - Bien évidement vous pouvez choisir un autre datawarehouse
- **Processing** :
  - DBT
  - Package : **dbt-core, dbt-snowflake**
  - Installation via un environnement **python en local ou DOCKER**
  - Permettra de faire les transformations nécessaires
  - ***NB*** : vous pouvez choisir d'utiliser **SPARK** ou **SNOWPARK** au lieu de DBT
- **Orchestration** :
  - **AIRFLOW**
  - Installation via un environnement **python en local ou DOCKER**
  - Permettra d'orchestrer les différentes transformations

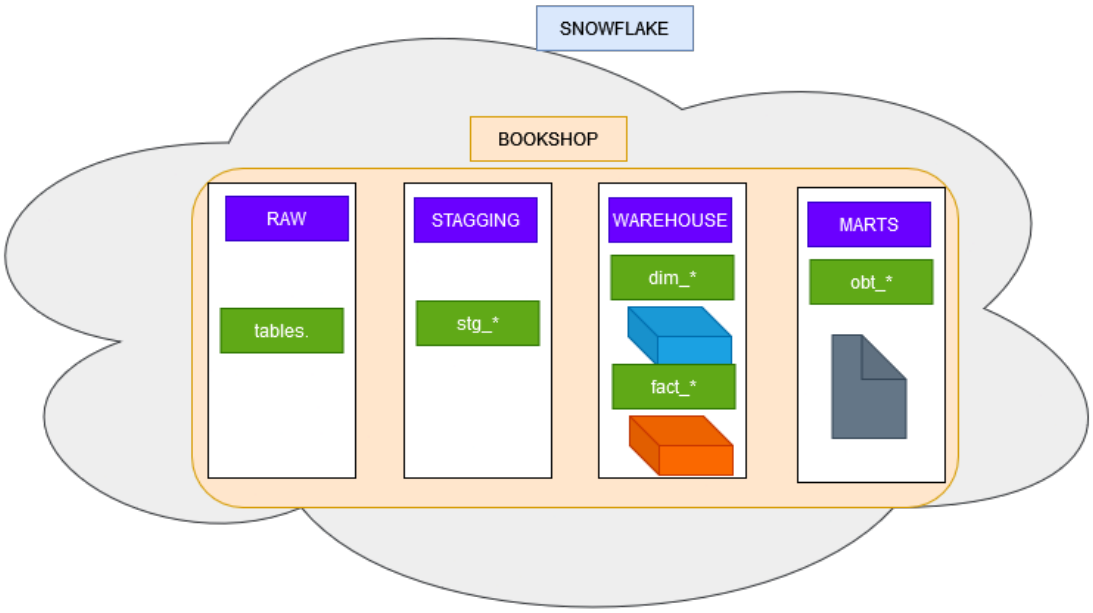
Structure BD :



Scripts SQL :

- Fichier PostgreSQL : **postgres_diagram_bigdata_m2.sql**
- Fichier MySQL : **mysql_diagram_bigdata_m2.sql**
- ***Les champs* ventes.date_edit *et* factures.date_edit *sont aux format* « *YYYYMMDD* »**

Résultat attendu dans SNOWFLAKE :



Steps :

1. Avec **Snowflake**, créer une base de données nommée **BOOKSHOP**
2. Avec **snowflake**, créer 04 schémas dans la base de données **BOOKSHOP** : **RAW, STAGGING, WAREHOUSE, MARTS**
3. Dans le schéma **RAW**, créer les tables ci-dessus et les alimenter avec les données de votre choix (***Respecter la structure et le format des données***).
4. Installez en local **DBT** et **AIRFLOW**
5. Transformation **RAW** -> **STAGGING** (préparation des données) avec **DBT** :
    a. Depuis **RAW.ventes**, convertir la colonne **ventes.date_edit** au format **DATE** et déposer le résultat dans la table **STAGGING.stg_ventes**
    b. Depuis **RAW.factures**, convertir la colonne **fatures.date_edit** au format **DATE** et déposer le résultat dans la table **STAGGING.stg_factures**
    c. Depuis **RAW.***, Copier les données des tables **category, books, customers** respectivement vers **STAGGING.stg_category, STAGGING.stg_books, STAGGING.stg_customers**
    d. Finalement le schéma **STAGGING** devrait contenir **05 tables commençant par stg_**
6. Transformation STAGGING -> **WAREHOUSE** avec **DBT** :
    a. Créer **WAREHOUSE.dim_customers, WAREHOUSE.dim_category, WAREHOUSE.dim_books** qui sont des copies respectives de **stg_customers, stg_category, stg_books**
    b. Créer la table **WAREHOUSE.dim_customers** basée sur **STAGGING.stg_customers** avec l'ajout de la colonne « **nom** » **: VARCHAR = first_name + ' ' + last_name** représentant la concaténation des colonnes **first_name et last_name**.
    c. Créer la table **WAREHOUSE.fact_ventes** basée sur **STAGGING.stg_ventes** avec l'ajout des colonnes « **années** » **: INT,** « **mois** » **:VARCHAR('janvier', 'fevrier', …),** « **jour** » **:VARCHAR ('lundi', 'mardi', …)** représentant les extractions du champ **STAGGING.stg_ventes.date_edit.**

d. Créer la table **WAREHOUSE.fact_factures** basée sur **STAGGING.stg_factures** avec l'ajout des colonnes **« années » : INT, « mois » :VARCHAR('janvier', 'fevrier', …), « jour » :VARCHAR ('lundi', 'mardi', …)** représentant l'extraction du champs **STAGGING.stg_factures.date_edit**.
e. Créer les tables **WAREHOUSE.fact_books_annees, WAREHOUSE.fact_books_mois, WAREHOUSE.fact_books_jour** représentant la liste des livres vendu par années, mois et jour
f. Finalement le schéma WAREHOUSE devrait contenir 03 tables commençant par dim_ et 05 tables commençant par fact_

7. Transformation **WAREHOUSE** -> **MARTS** avec **DBT** :
    a. L'objectif est d'avoir une seule table nommée **MARTS.obt_sales** contenant toutes les informations nécessaires pour identifier une ligne de vente.
    b. L'identifiant de la table **MARTS.obt_sales** doit provenir de **WAREHOUSE.fact_ventes**
    c. Champs de la table **MARTS.obt_sales** (les infos proviennent des tables du schéma **WAREHOUSE**) : **fact_ventes[id, annees, mois, jour, pu, qte] + fact_factures[id, code, qte_totale, total_amount, total_paid] + dim_category[intitule] + dim_books[code, intitule, isbn_10, isbn_13] + dim_customers[code, nom]**

Touche Personnelle :

1. **Ingestion** :
    a. Créer une bd locale (**PostgreSQL ou MySQL**)
    b. Insérer ou alimenter les données.
    c. Mettre en place un pipeline d'ingestion (**de votre choix**) pour déposer les données brutes dans **SNOWFLAKE (BOOKSHOP.RAW)**
2. Visualisation au choix (mettre en œuvre une solution) afin de mettre en avant les données **MARTS** et **WAREHOUSE** :
    a. Options possibles de mise en place un système de visualisation :
        i. Application : **streamlit, nicegui, notebook**
        ii. Powerbi, kibana
        iii. Créer un compte sur https://www.tableau.com/ (**payant**) et connecter avec **snowflake**
    b. Ajouter un maximum de graphes dans votre Dashboard

Bonne Chance