Nom : Parent

Prénom : Jasmin

Groupe : 00263

Cours: BD4

TP : TP1: Mise en place d’un processus ETL avec python

Collège de Bois de Boulogne

Introduction

Ce document fait l'objet d'un cas d'implémentation de processus ETL (Extract-Transform-Load) avec le langage de programmation Python 3. Trois sources de données différentes seront traitées pour en extraire les données brutes d'une clientèle fictive. Les sources sont composées d'un fichier au format JSON, un fichier au format CSV ainsi qu'une table de base de données SQL. Une fois les données traitées pour toute incohérences, elles seront insérées dans une base de données relationnelle MySQL. Cet exercice a pour but de se familiariser avec le processus en question ainsi que le langage Python.

Caractéristiques des données

- Les données source contiennent 6 informations à propos de clients, exemple : Id, nom, prénom, etc…
- En plus de ces 6 valeurs, la table de destination gardera « source_name » et « source_id ».
- Les valeures nulles sont remplacées par « unknown ».
- Les sexes sont normalisés vers « F » ou « M » (ou « unknown »).

id	source_id	source_name	first_name	last_name	email	gender	ville
31	5244	week_cust.csv	Judy	Crawford	jcrawfordu@t-online.de	F	unknown
32	5245	week_cust.csv	Ralph	Chavez	rchavezv@cpanel.net	M	Arizona
33	5246	week_cust.csv	Johnny	Little	jlittlew@hud.gov	M	New York

Table destination dans phpMyAdmin

Caractéristiques du processus

- Un module « mod_db » pour les fonctions de bases de données :
    - Connect : Établie une connexion à une base de donnée MySQL
- Module « mod_data » qui contient les fonctions de manipulation de données :
    - Extraction: Une fonction pour chaque type de source (csv, json et table) qui prend en entrée le chemin d'un fichier et renvoie les données structurées et normalisées.
    - Insertion: Fonction prenant des données structurées en entrée pour les insérer dans la DB.
    - Normalisation : Une fonction qui prend le genre en entrée et renvoie la version normalisée.
- Module contrôleur « mod_control ».
- Toute les fonctions ont un docstring la décrivant ainsi que tous les paramètres et leur type.
- Conçu pour être facilement simple, lisible et maintenable.

Étapes de l'exécution du processus

1. Des connexions avec les bases de données source et destination sont instanciées.
2. La table de destination est créée si ce n'est pas déjà le cas.
3. Les données sont extraites et traitées pour chaque source.
4. Le data est inséré dans la table de destination.
5. Les connexions et les fichiers ouverts sont fermés, fin de l'exécution
6. Vérification que la table de destination ne contient pas de valeurs nulles et que le nombre de rows est celui attendu.

Conclusion

Avec l'aide de la librairie pymsql pour gérer la création d'une connexion de base de données, le processus est simple à écrire en Python. Les modules de base csv et json rendent la manipulation des fichiers très facile.
Python semble donc offrir une solution simple et rapide pour mettre en place un processus ETL.