Objet : Proposition d'utilisation du script Python dans un pipeline ETL pour la récupération de données

Cher Sam,

Voici le mail demandé pour l'établissement d'un pipeline ETL(Extract, Transform, Load) construit grâce à mon programme. Ce dernier extrait les informations, organise et stocke les informations du site web "http://books.toscrape.com/" de manière structurée.

Étapes du pipeline ETL :

1. **Extraction :** Le script visite et extrait les catégories, les pages et les détails demandés des produits du site grâce à plusieurs méthodes de scrapping. Ces méthodes sont divisées en 3 classes distinctes pour une lecture et compréhension plus simple et efficace du script général. De plus, les images de chaque livres sont téléchargées.
2. **Transformation :** Les données sont transformées dans un second temps en un format adapté, avec un nettoyage des caractères spéciaux. Les informations sont organisées dans des fichiers CSV séparés par catégories. Les images sont quant à elles convertis en format JPEG.
3. **Chargement :** Les données sont ensuite écrites dans leurs fichiers CSV respectifs pour une analyse future. Les images sont également téléchargées et disponibles localement, nommées par le titre du produit.

Utilisation pratique :

1. **Installation :** Assurez-vous d'avoir installé l'environnement virtuel requis à l'aide du fichier requirements.txt
2. **Exécution :** Lancez le script Python main.py dans votre environnement de développement.
3. **Résultats :** Les données extraites seront stockées dans des fichiers CSV organisés par catégorie, et les images des livres seront téléchargées dans le dossier "Books image". Quelques “print()” permettent de suivres l'avancé du programme
4. **Analyse :** Les fichiers CSV peuvent être analysés à l'aide d'outils tels qu'Excel ou Pandas.

Je serais ravi de discuter davantage de cette initiative et de son potentiel d'application dans le cadre de nos projets actuels. N'hésitez pas à me faire part de vos commentaires ou de vos questions.

Cordialement,

THIERRY
Edwin