TP Apache Hop - Traitement de données avec Faker

Clément Agret

Objectifs

Dans ce TP, vous allez apprendre à utiliser Apache Hop pour traiter des données générées avec la bibliothèque Faker en Python. Vous serez guidé pas à pas dans la réalisation des différentes étapes.

1 Génération de données avec Faker

1. Installez Python et la bibliothèque Faker
2. Créez un script Python pour générer un jeu de données de 1000 utilisateurs avec les champs suivants : nom, prénom, email, date de naissance, adresse
3. Exportez les données générées dans un fichier CSV

2 Installation et prise en main d'Apache Hop

1. Téléchargez et installez Apache Hop depuis le site officiel : `https://hop.apache.org/`
2. Lancez Hop et familiarisez-vous avec l'interface graphique
3. Créez un nouveau projet et un nouveau pipeline

3 Importation des données CSV dans Hop

1. Ajoutez une étape "CSV file input" dans votre pipeline

2. Configurez l'étape pour lire le fichier CSV généré précédemment
3. Affichez un aperçu des données importées pour vérifier que tout fonctionne correctement

4 Transformation des données

1. Ajoutez une étape "Split fields" pour séparer le nom et le prénom en deux colonnes distinctes
2. Utilisez une étape "Calculator" pour calculer l'âge de chaque utilisateur à partir de sa date de naissance
3. Filtrez les utilisateurs ayant moins de 18 ans avec une étape "Filter rows"
4. Triez les utilisateurs par ordre alphabétique sur le nom avec une étape "Sort rows"
5. Corrigez les adresses email en minuscules avec une étape "Modify fields"
6. Appliquez un tris sur les adresses email avec une étape "Sort rows"
7. Assurez-vous que les adresse email sont uniques avec une étape "Unique rows"

5 Exportation des données transformées

1. Ajoutez une étape "Text file output" pour exporter les données transformées dans un fichier texte
2. Configurez l'étape pour écrire les champs souhaités, séparés par des tabulations
3. Exécutez le pipeline et vérifiez que le fichier de sortie est correctement généré

Exercices supplémentaires

Exercice 1 : Gestion des doublons

1. Dupliquez aléatoirement certaines lignes dans votre fichier CSV d'entrée pour créer des doublons

2. Ajoutez une étape "Unique rows" dans votre pipeline pour supprimer les doublons
3. Comparez le nombre de lignes avant et après la suppression des doublons

Exercice 2 : Validation des adresses email

1. Ajoutez une étape "Regex evaluation" pour valider le format des adresses email
2. Utilisez une expression régulière appropriée pour vérifier que les adresses email sont de la forme "nom@domaine.com"
3. Filtrez les lignes contenant des adresses email invalides avec une étape "Filter rows"
4. Comptez le nombre d'adresses email valides et invalides avec une étape "Group by"

Exercice 3 : Jointure avec un fichier externe

1. Créez un second fichier CSV contenant des informations complémentaires sur les utilisateurs (par exemple, leur ville et leur pays)
2. Importez ce fichier dans un nouveau pipeline
3. Utilisez une étape "Merge join" pour joindre les données des deux fichiers sur la base de l'email
4. Exportez le résultat de la jointure dans un nouveau fichier texte
5. Vérifiez que les données des deux fichiers ont bien été combinées pour chaque utilisateur

Exercice 4 : Transformation de données

1. Ajoutez une étape "Select values" pour ne conserver que les colonnes pertinentes de votre fichier d'entrée
2. Utilisez une étape "Calculator" pour créer une nouvelle colonne calculée (par exemple, l'âge à partir de la date de naissance)
3. Ajoutez une étape "If field value is null" pour remplacer les valeurs manquantes par une valeur par défaut

4. Exportez le résultat de ces transformations dans un nouveau fichier CSV

Exercice 5 : Agrégation de données

1. Ajoutez une étape "Group by" pour regrouper les utilisateurs par pays
2. Utilisez une étape "Database join" pour récupérer des informations sur chaque pays depuis une base de données externe (par exemple, la population)
3. Calculez des statistiques agrégées pour chaque pays (nombre d'utilisateurs, âge moyen, etc.) avec une étape "Analytic query"
4. Exportez les résultats de l'agrégation dans un nouveau fichier Excel

Exercice 6 : Automatisation du pipeline

1. Paramétrez votre pipeline pour qu'il accepte des arguments en entrée (nom du fichier, date d'exécution, etc.)
2. Ajoutez une étape "Mail" pour envoyer un rapport par email à la fin de l'exécution du pipeline
3. Utilisez une étape "File exists" pour vérifier la présence des fichiers d'entrée avant de lancer le traitement
4. Configurez une tâche planifiée pour exécuter automatiquement votre pipeline tous les jours à une heure donnée
5. Vérifiez que le pipeline s'exécute correctement de manière automatisée et que le rapport est bien envoyé par email