Université de Caen Normandie

IUT Grand Ouest Normandie

Pôle de Caen

Bachelor Universitaire de Technologie

Science des Données

BUT Science des Données
Deuxième année

Rapport de Projet

Talend Open Studio

Auteur

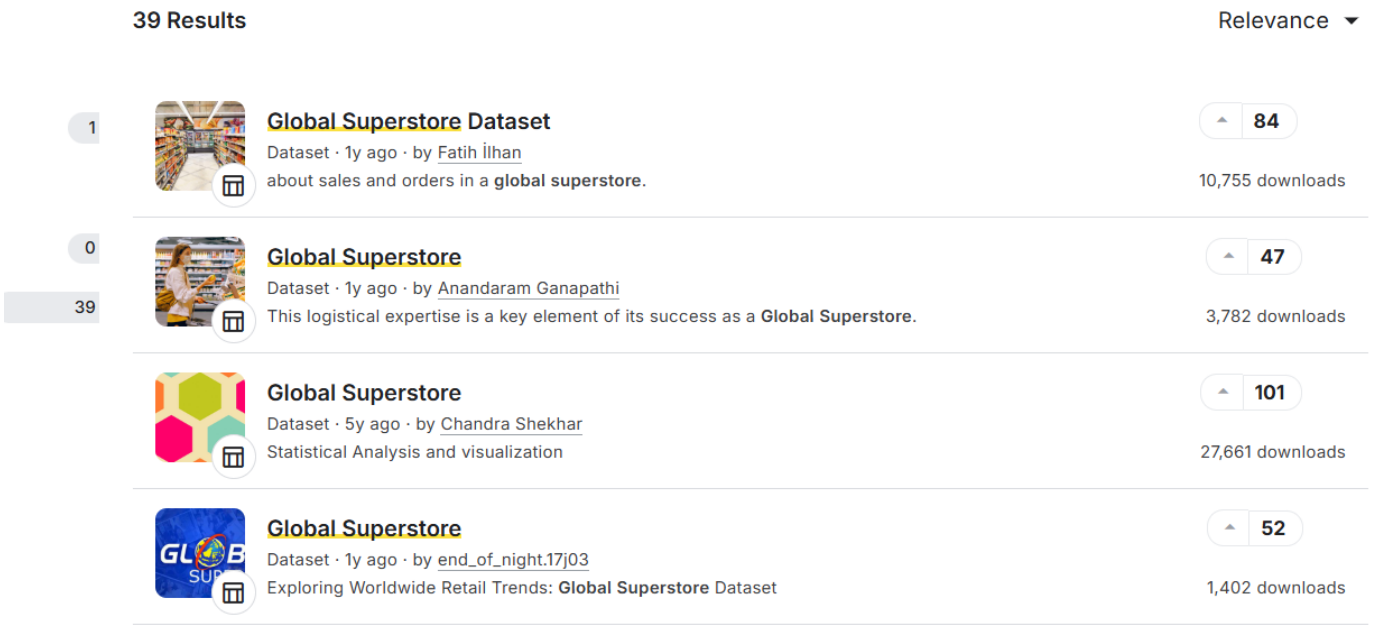
Sidy Babacar Diop
Courteney Saint-Hubert

Année universitaire 2024-2025

SOMMAIRE

Introduction

Pendant ce projet, nous avons utilisé Talend Open Studio pour traiter et analyser un fichier de données, le nettoyer, le transformer et enfin, le charger dans une base de données relationnelle. L'objectif est d'exploiter les fonctionnalités mises à notre disposition par le logiciel Talend afin d'automatiser le processus ETL (Extract, Transform, Load). Ce projet repose sur l'utilisation du jeu de données Global Superstore qui nous vient de Kaggle.



1 - Présentation des données

Le fichier source utilisé est Global_Superstore2.csv, contenant des informations sur les ventes, les clients et les produits.

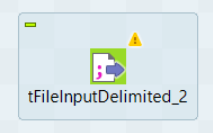
Les colonnes qui vont nous intéresser ici sont :

— Order ID : Identifiant unique de la commande
— Order Date : Date de la commande
— Ship Date : Date d'expédition
— Ship Mode : Mode de livraison
— Customer ID : Identifiant client
— Customer Name : Nom du client
— Segment : Type de client
— Country : Pays de l'achat
— Product ID : Identifiant du produit
— Product Name : Nom du produit
— Category : Catégorie du produit
— Quantity : Quantité vendue
— Sales : Montant des ventes
— Profit : Bénéfice généré
— Region : Région de la vente

2 - Nettoyage et transformation des données

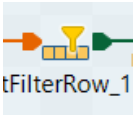
2.1 Importation des données

Pour commencer notre projet, il nous faut importer les données du fichier Global_Superstore2.csv dans Talend Open Studio à l'aide du composant tFileInputDelimited. Ce composant permet de lire les fichiers CSV en définissant le séparateur adéquat , ou ; selon le format du fichier(nous allons utiliser la virgule) tout en spécifiant les types de colonnes correspondants. Une fois l'importation effectuée , le schéma des données a été défini pour garantir une bonne manipulation des champs.



2.2 Nettoyage des données

Le nettoyage des données est une étape essentielle pour garantir l'intégrité des données.tégrité des informations. Dans ce projet, les valeurs nulles ou invalides ont été supprimées grâce au composant tFilterRow. Les dates (Order Date, Ship Date) ont été formatées correctement.

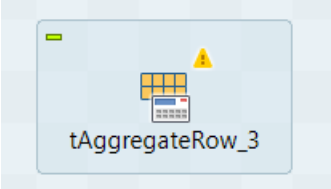


2.3 Transformation des données

Une fois les données nettoyées, elles ont été transformées pour enrichir l'analyse. Une nouvelle colonne, Sales_Category, a été ajoutée via tMap. Un autre traitement a consisté à calculer le total des ventes et la marge bénéficiaire pour chaque client. Pour cela, nous avons utilisé le composant tAggregateRow, qui a permis de regrouper les ventes et les profits par Customer ID. Une nouvelle colonne Marge_bénéficiaire a été créée avec l'expression suivante :

(Profit / Sales * 100)

Enfin, un filtrage a été appliqué pour sélectionner uniquement les transactions appartenant à la région West. Ceci a été réalisé avec tFilterRow, qui a permis d'exclure les autres régions et de concentrer l'analyse sur une zone spécifique.

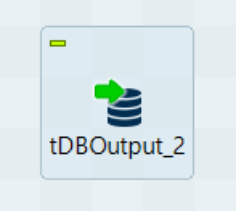


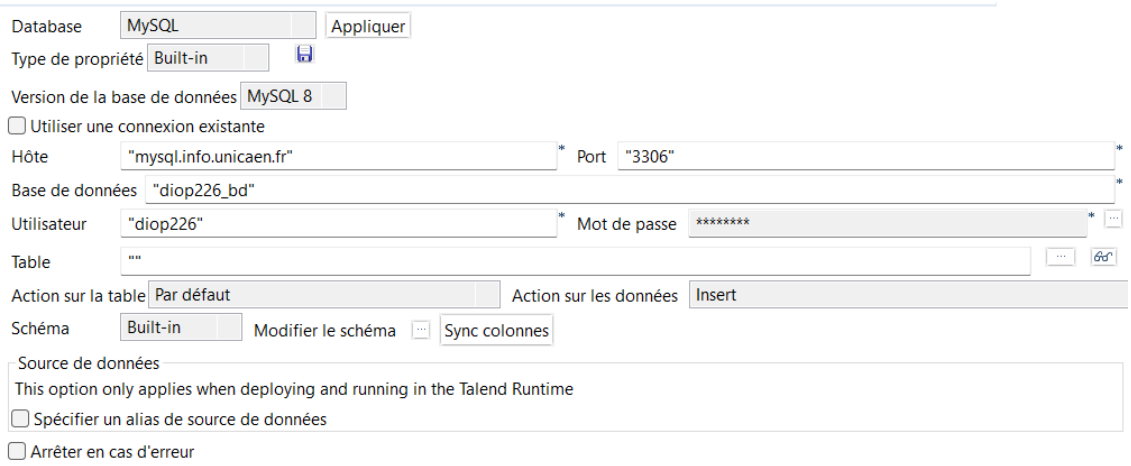
3 - Chargement des données dans une base de données

Une fois les données transformées, elles ont été chargées dans une base de données MySQL pour une exploitation ultérieure. La connexion à la base de données a été réalisée à l'aide du composant tDBOutput avec les paramètres suivants :

Une fois les données transformées, elles ont été chargées dans une base de données MySQL pour une exploitation ultérieure. La connexion à la base de données a été réalisée à l'aide du composant tDBConnection avec les paramètres suivants :

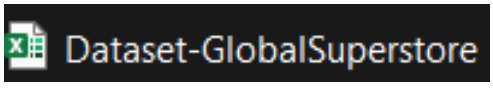
- Type : mysql
- Host : mysql.info.unicaen.fr
- Port : 3306
- Base de données : diop226_bd
- Utilisateur : diop226
- Mot de passe : sécurisé





4 - Exportation des données transformées

En complément, les données nettoyées et transformées ont été sauvegardées sous différents formats pour une exploitation ultérieure. Un fichier CSV a été généré à l'aide de tFileOutputDelimited, tandis qu'un fichier Excel contenant les résultats agrégés a été produit avec tFileOutputExcel. Nous avons choisi de récupérer le fichier au format .xlsx, donc nous avons utilisé tFileOutputExcel sous le nom de Dataset-GlobalSuperstore qui comporte 4 feuilles différentes contenant nos 4 tables.



5 - Conclusion

Ce projet a permis d'automatiser le traitement des données de vente en utilisant Talend Open Studio. Grâce aux transformations effectuées, nous avons pu :

— Nettoyer et structurer les données pour une meilleure analyse.
— Catégoriser les ventes selon un seuil défini.
— Stocker les informations en base de données pour une utilisation future.
— Générer des rapports exploitables sous format Excel.

L'utilisation de composants tels que tMap, tFilterRow, tAggregateRow et tDBOutput a facilité la mise en place d'un processus ETL efficace. Ce projet constitue une base solide pour l'analyse et la gestion des ventes d'une entreprise.