Université Abdelmalek Essaâdi
ENSA Tétouan

2èME année Cycle d'Ingénieur – BDIA
Module : Business intelligence

Analyse des Avis Clients Amazon : Une Approche Business Intelligence

Préparé par :

BOULAALAM YASSINE
BOUSKINE OTHMANE
BENADIR MARYAME
TARRE BASMA
BOUHAFA TAHA

Encadré par:

Prof.Faouzi Marzouki

TABLE DE MATIERE

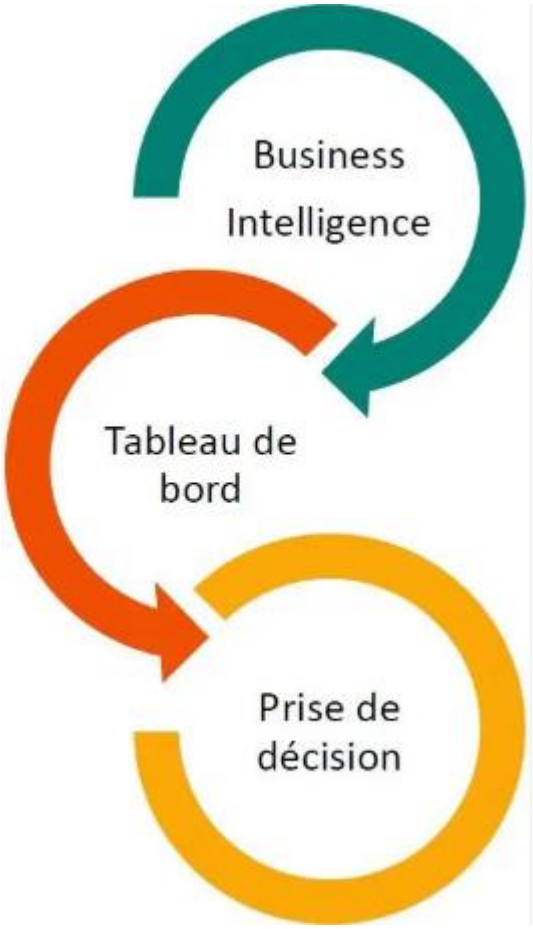
I.INTRODUCTION

Contexte et Problématique

Dans un environnement commercial de plus en plus concurrentiel, les entreprises s'appuient sur des avis clients pour comprendre les besoins, les attentes et les perceptions de leur public. La plateforme **Amazon Fine Food Reviews** fournit une grande quantité d'avis d'utilisateurs sur divers produits alimentaires, créant une opportunité unique pour extraire des informations stratégiques. Cependant, cette abondance de données peut être difficile à exploiter efficacement sans une analyse appropriée et des outils dédiés.

La problématique principale que nous adressons est :

Comment exploiter les avis clients pour fournir des insights exploitables permettant d'améliorer les performances des produits et de la stratégie commerciale d'une entreprise ?



Les entreprises doivent répondre à plusieurs questions clés :

- Quels produits reçoivent les meilleurs gains et pourquoi ?
- Comment les différents aspects des avis, comme leur longueur ou leur sentiment, influencent-ils la perception des clients ?
- Existe-t-il des tendances dans les avis qui pourraient guider la stratégie marketing ou les efforts de développement produit ?

Objectifs de l'étude

L'objectif de ce projet est de développer une **solution de Business Intelligence (BI)** pour analyser et visualiser les avis clients afin d'aider les décideurs à prendre des décisions éclairées. Plus précisément, nous cherchons à :

1. **Transformer les données brutes en informations stratégiques** :
    - Identifier les produits populaires et moins performants.
    - Comprendre l'impact de la longueur des avis sur leur score ou leur utilité.
    - Analyser les sentiments exprimés dans les avis pour une meilleure compréhension de la satisfaction client.

2. **Mettre en place une architecture BI robuste** :
    - Utiliser des outils comme MySQL pour le stockage et la manipulation des données.
    - Exploiter Pentaho pour l'ETL (Extraction, Transformation, Chargement) et le nettoyage des données.
    - Visualiser les résultats dans Power BI pour faciliter leur interprétation.
3. **Proposer une solution argumentée et adaptée** :
    - Démontrer l'optimalité de notre approche.
    - Offrir une analyse critique de la pertinence de la solution pour répondre à la problématique.

Terminologie et Concepts Clés de la BI

Business Intelligence (BI)

La Business Intelligence se réfère à l'ensemble des méthodes, processus, architectures et technologies qui transforment les données brutes en informations exploitables pour la prise de décisions stratégiques. Elle repose sur trois étapes principales :

- **Extraction des données** : Collecter les données provenant de sources.
- **Analyse** : Appliquer des techniques pour découvrir des tendances et des modèles.
- **Visualisation** : Présenter les résultats sous une forme compréhensible pour les décideurs.

Concepts spécifiques appliqués à ce projet

- **Entrepôt de données (Data Warehouse)** : Structure centralisée pour stocker les données nettoyées et agrégées.
- **ETL (Extraction, Transformation, Chargement)** : Processus qui permet de préparer les données pour l'analyse BI.
- **Indicateurs clés de performance (KPIs)** : Métriques permettant de mesurer l'efficacité de l'entreprise à répondre à ses objectifs (exemple : score moyen par produit, ratio d'utilité des avis).
- **Visualisation interactive** : Utilisation de tableaux de bord dynamiques dans Power BI pour explorer les données de manière intuitive.

Portée et Contributions du Projet

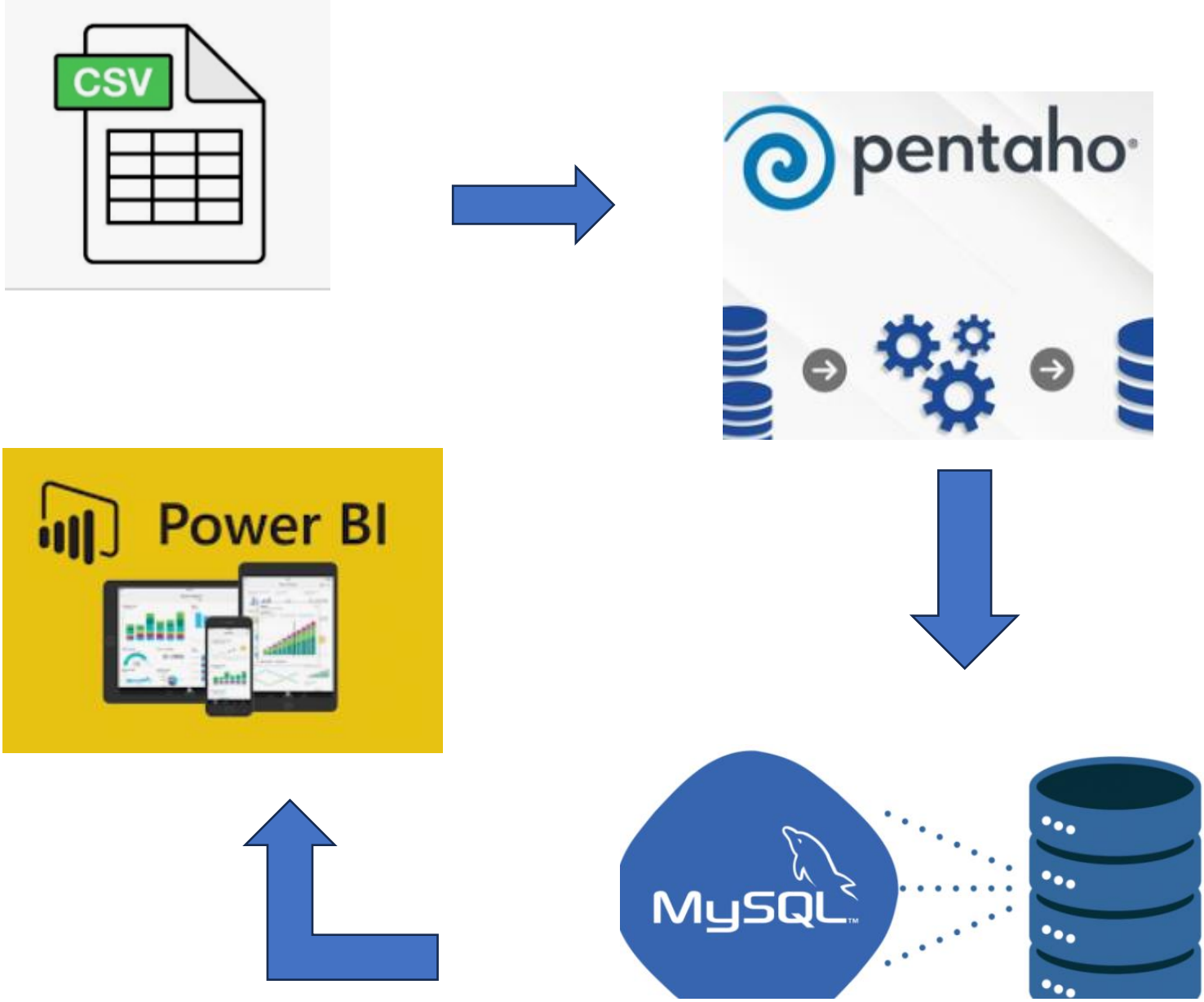
Ce projet contribue à :

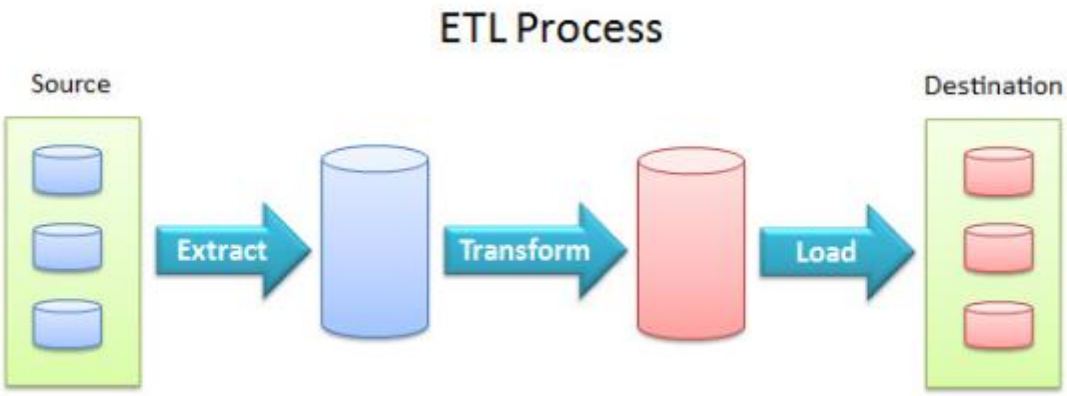
- Offrir une solution intégrée d'analyse des avis clients pour répondre à des problèmes stratégiques concrets.

- Montrer l'impact de la BI dans le contexte de la satisfaction client et du développement produit.
- Mettre en œuvre une architecture moderne et efficace en utilisant des outils reconnus dans le domaine de la BI.

En somme, cette étude démontre comment la BI peut transformer les données clients en un avantage concurrentiel majeur pour une entreprise.

II.ARCHITECTURE DES OUTILS UTILISES





plusieurs outils collaborent pour collecter, transformer, stocker et visualiser des données afin de fournir des insights stratégiques.

1. **Source de Données (CSV)** :
    - Les données brutes des avis Amazon Fine Food sont au format CSV. Ce fichier contient les informations initiales nécessaires pour l'analyse.
2. **Transformation des Données (Pentaho)** :
    - **Pentaho** est utilisé comme outil ETL (**Extract, Transform, Load**). Il permet d'extraire les données depuis le fichier CSV, d'appliquer des transformations (nettoyage, calcul de colonnes comme la longueur des avis ou le ratio de pertinence) et de charger les données transformées dans une base de données relationnelle.
3. **Stockage des Données (MySQL)** :
    - Les données transformées sont stockées dans un système de gestion de bases de données MySQL. Cela facilite leur structuration et leur interrogation via des requêtes SQL pour des analyses ultérieures.
4. **Visualisation des Données (Power BI)** :
    - **Power BI** est utilisé pour la phase de reporting et de visualisation. Grâce aux données provenant de MySQL, des tableaux de bord interactifs et des graphiques sont créés pour représenter les métriques clés comme le ratio d'utilité, les scores des avis, et les tendances générales.

Objectif BI :

L'objectif est de transformer les données brutes en **informations exploitables** permettant aux décideurs de mieux comprendre le comportement des utilisateurs, d'identifier les tendances dans les avis et de formuler des stratégies pour améliorer l'expérience client et l'engagement.

III.DATASET

Le dataset **"Amazon Fine Food Reviews"** contient des critiques d'aliments postées par des utilisateurs sur Amazon entre octobre 1999 et octobre 2012. Il se compose de 568 454 critiques, fournies par 256 059 utilisateurs, et couvre 74 258 produits alimentaires différents. Les critiques incluent des informations détaillées sur les utilisateurs, les produits et l'évaluation fournie par les consommateurs. Ce dataset permet d'étudier les comportements des consommateurs, d'analyser les tendances dans les avis sur les produits et de comprendre la relation entre les critiques et les notes des produits.

Les principales colonnes de ce dataset sont les suivantes :

1. **Id** : Identifiant unique pour chaque revue, permettant de différencier les enregistrements dans le dataset.
2. **ProductId** : Identifiant unique du produit alimentaire concerné par la revue.
3. **UserId** : Identifiant unique de l'utilisateur ayant posté la revue.
4. **ProfileName** : Nom du profil de l'utilisateur ayant laissé la critique.
5. **HelpfulnessNumerator** : Nombre d'utilisateurs ayant trouvé la revue utile.
6. **HelpfulnessDenominator** : Nombre total d'utilisateurs ayant évalué l'utilité de la revue (utile ou non utile).
7. **Score** : Note attribuée au produit par l'utilisateur, sur une échelle de 1 à 5.
8. **Time** : Date et heure de publication de la revue, sous forme d'horodatage.
9. **Summary** : Résumé succinct de la critique, offrant un aperçu rapide des opinions exprimées.
10. **Text** : Texte complet de la critique, où l'utilisateur décrit son expérience avec le produit.
11. **State** : Informations géographiques ou autre statut associé à l'utilisateur ou à la revue.

Autres caractéristiques du dataset :

- **Nombre total de produits** : 74 258 produits alimentaires différents.
- **Utilisateurs influents** : 260 utilisateurs ayant laissé plus de 50 critiques chacun, ce qui peut être utile pour des analyses sur l'influence des utilisateurs dans la communauté.
- **Période couverte** : Le dataset contient des critiques entre octobre 1999 et octobre 2012, offrant une perspective sur les tendances à long terme dans les comportements d'achat et les évaluations des consommateurs.

IV.DEMARCHE D'IMPLEMENTATION CHOISIE

Modélisation des données

Pour structurer et optimiser les données pour l'analyse et les rapports, nous utilisons une modélisation en étoile adaptée aux besoins de ce projet. Cette approche organise les données autour d'une table de faits centrale et de plusieurs tables de dimensions. Voici les détails de cette modélisation :

1. Schéma en Étoile

Le modèle de données en étoile est choisi pour sa simplicité et son efficacité dans les entrepôts de données. Il permet une exploration multidimensionnelle rapide des données.

a. Table des Faits : "FactReviews"

Cette table centralise les mesures quantitatives et les indicateurs de performance liés aux critiques. Elle contient les informations suivantes :

- **Reviwim :** Identifiant unique pour chaque critique.
- **StateID** : Identifiant géographique (clé étrangère vers la dimension État).
- **ProductID :** Identifiant du produit (clé étrangère vers la dimension Produit).
- **Quarter** : Trimestre où la critique a été postée (clé étrangère vers la dimension Temps).
- **SentimentDetected :** Sentiment détecté pour la critique (positif, négatif, neutre).
- **HelpfulnessRatio :** Ratio d'utilité calculé **(HelpfulnessNumerator/HelpfulnessDenominator).**
- **WordCount :** Nombre total de mots dans la critique.

b. Tables de Dimensions

Les dimensions fournissent un contexte aux faits et permettent de segmenter les données pour des analyses approfondies.

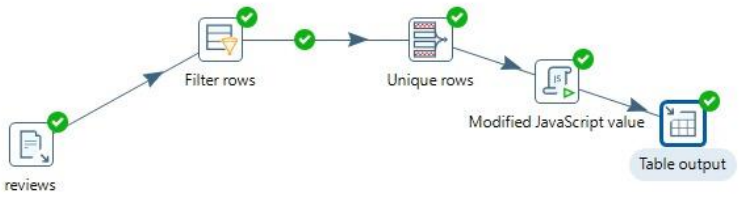
1. **Dimension Temps ("DimTime")**
   Cette table contient les informations temporelles :
   - **QuarterID :** Identifiant du trimestre.
   - **Quarter :** Trimestre de l'année (1, 2, 3, 4).
   - **Year :** Année associée au trimestre.

2. **Dimension Produit ("DimProduct")**
   Cette table décrit les informations sur les produits :
   - **ProductID :** Identifiant unique du produit.
3. **Dimension État ("DimState")**
   Cette table fournit les informations géographiques :
   - **StateID :** Identifiant unique pour chaque état/région.
   - **StateName** : Nom de l'état ou de la région.
4. **Dimension Texte ("DimText")**
   Cette table analyse les mots dans les critiques :
   - **WordID :** Identifiant unique pour chaque mot.
   - **Word :** Le mot extrait des critiques.
   - **Frequency :** Nombre total d'occurrences du mot dans toutes les critiques.

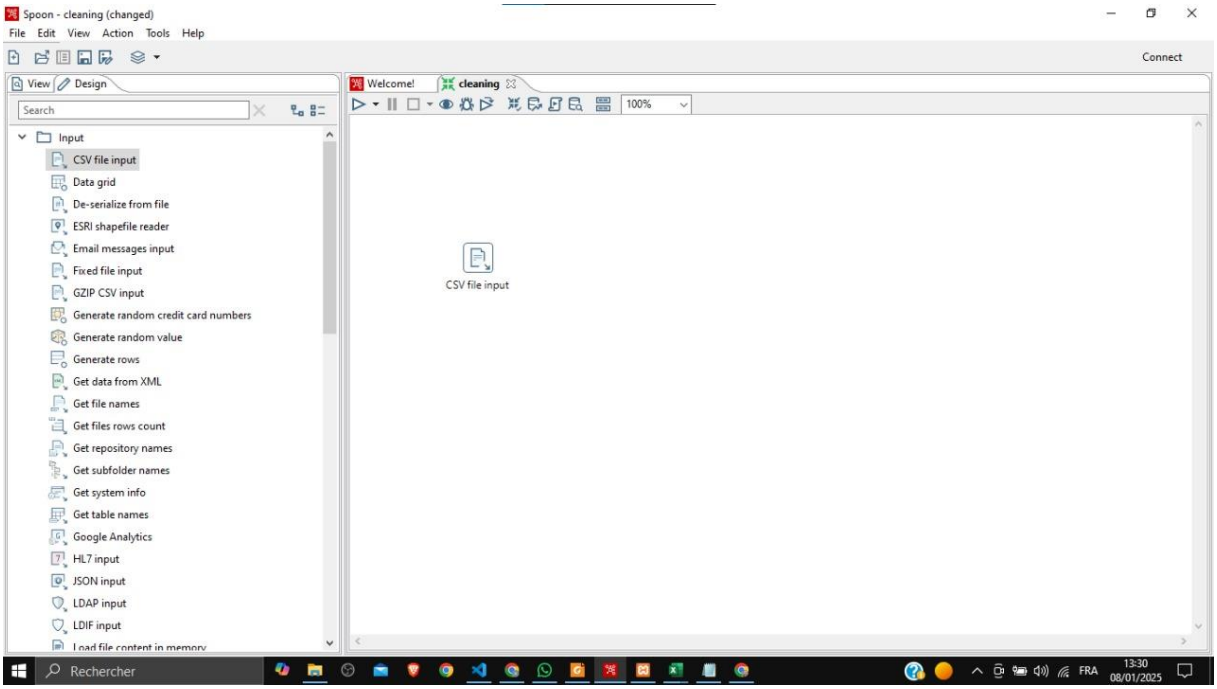
Implémentation des processus ETL

L'implémentation des processus ETL (Extract, Transform, Load) dans ce projet se fait en cinq étapes principales, chacune ayant pour objectif de préparer les données avant leur insertion dans le modèle de données final. Voici les étapes détaillées :

Transformation 1 : Nettoyage et préparation des données à partir du fichier CSV

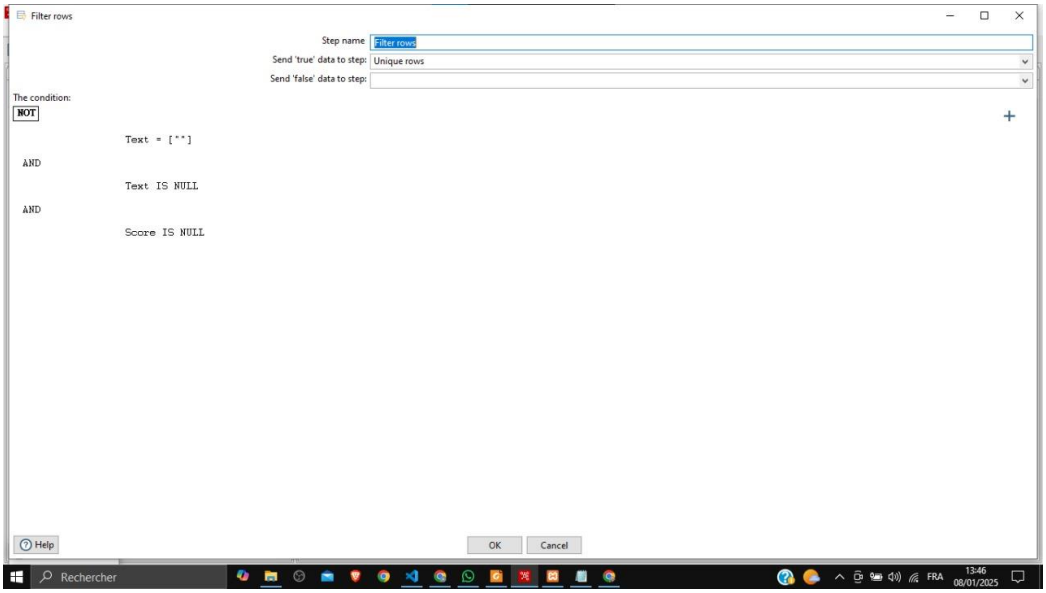


Input : Fichier CSV contenant les critiques de produits.

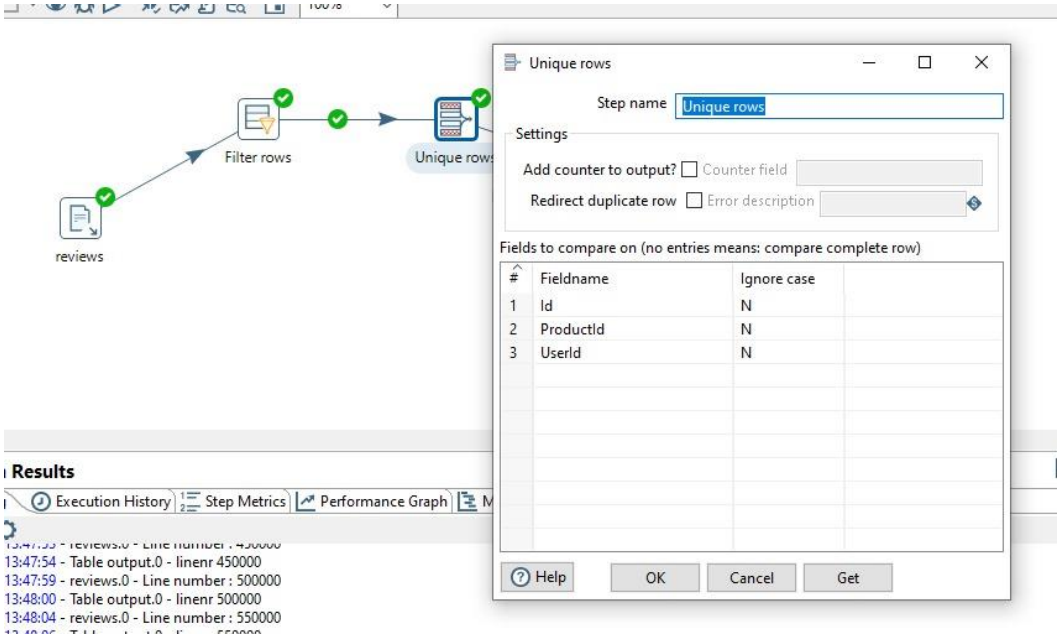


1. **Elimination des valeurs nulles** :

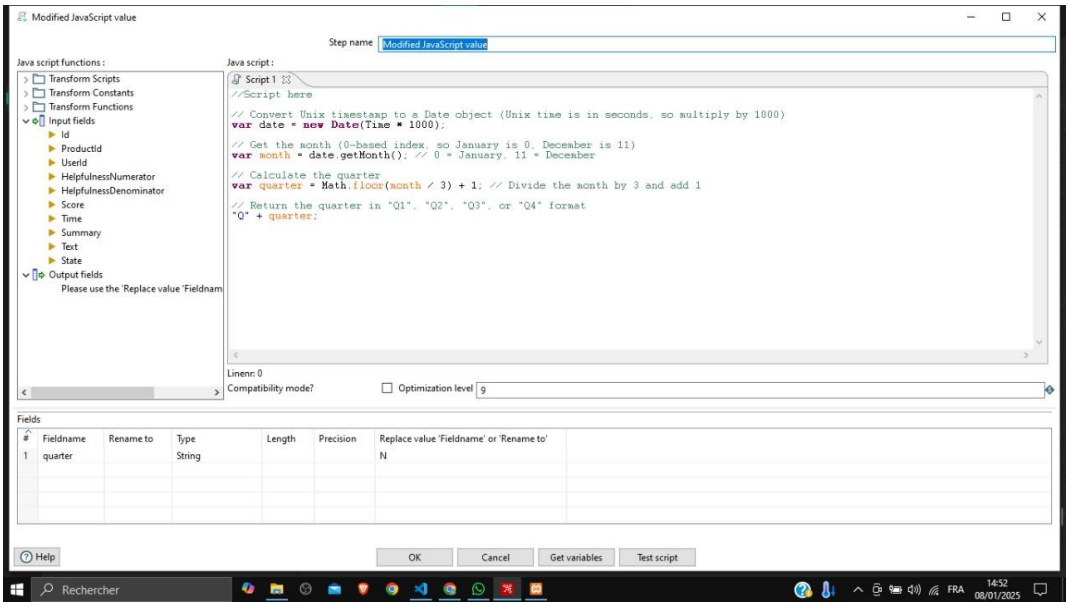
   - Pour les colonnes **Text** et **Score**, les lignes contenant des valeurs nulles sont supprimées. Cela garantit que seules les critiques complètes sont conservées, ce qui est essentiel pour les analyses de sentiment et la qualité des données.



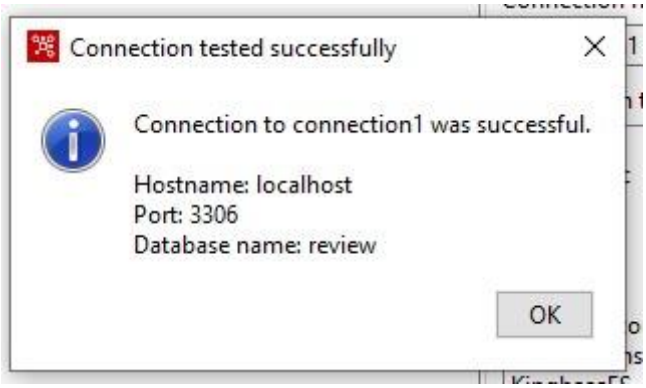
2. **Suppression des doublons** :
    - Les lignes du fichier CSV sont vérifiées et les doublons sont éliminés pour s'assurer que chaque critique est unique. Cette étape est cruciale pour éviter la redondance dans les données.

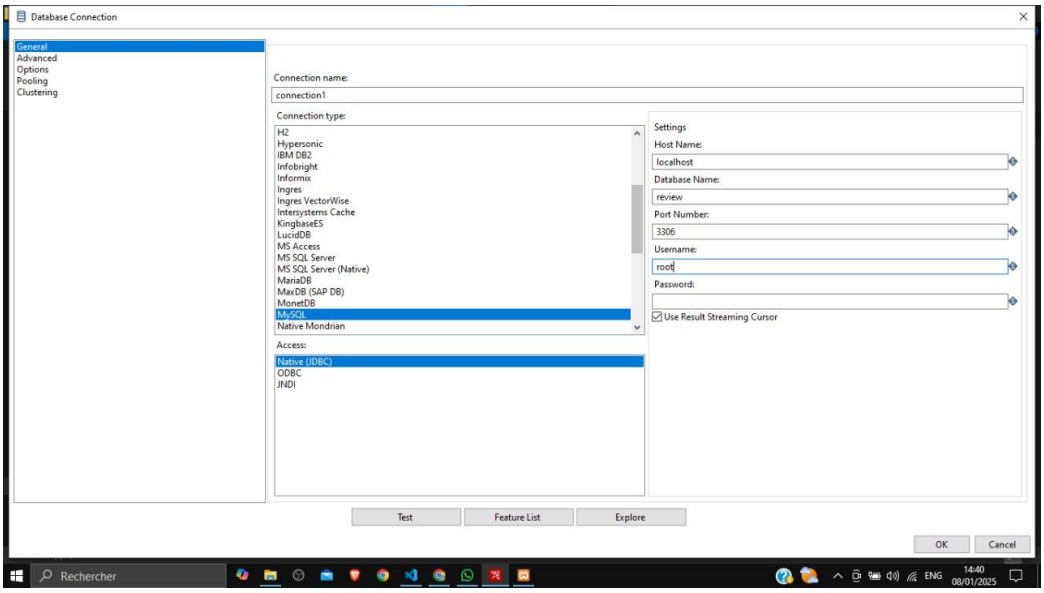


3. **Ajout de la colonne "Quarter"** :
    - Une nouvelle colonne **Quarter** est ajoutée à la table, qui divise les critiques en fonction de la saison de l'année (1 pour le premier trimestre, 2 pour le deuxième trimestre, etc.). Cette étape permet de mieux analyser les critiques selon les saisons et la temporalité.

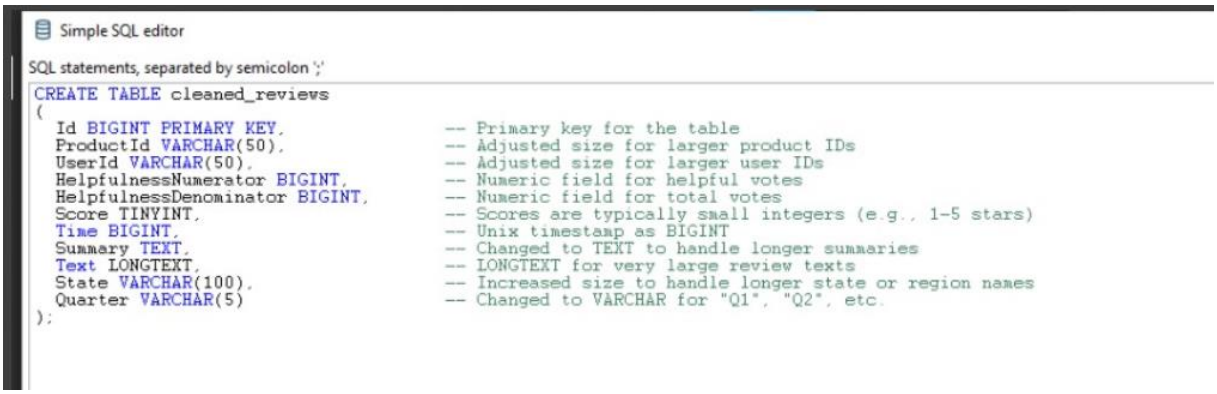
- 

4. **Connexion à la base de données via Pentaho** :
   - Le fichier CSV est déjà connecté à Pentaho, et une table est créée pour stocker le résultat de cette transformation.
   - La table stocke toutes les colonnes du fichier CSV d'origine tout en ajoutant la nouvelle colonne **Quarter** pour identifier la saison des critiques.



- 

Output : Une table contenant toutes les colonnes du fichier CSV d'origine, avec l'ajout de la colonne **Quarter** pour chaque ligne.

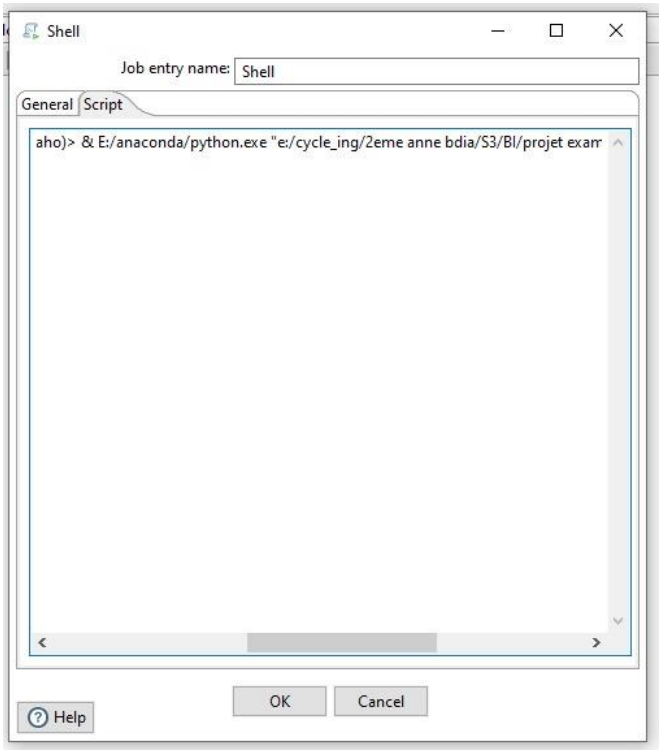


Id	ProductId	UserId	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text	State	Quarter
1	B001E4KFG0	A3SGXH7AUHU8GW	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned dog f...	Texas	2
2	B00813GRG4	A1D87F6ZCVE5NK	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanuts.....	Ohio	3
3	B000LQOCH0	ABXLMWJIXXAIN	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a few ce...	Illinois	3
4	B000UA0QIQ	A395BORC6FGVXV	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient in Ro...	Ohio	2
5	B006K2ZZ7K	A1UQRSCLF8GW1T	0	0	5	1350777600	Great taffy	Great taffy at a great price. There was a wide as...	California	4
6	B006K2ZZ7K	ADT0SRK1MGOEU	0	0	4	1342051200	Nice Taffy	I got a wild hair for taffy and ordered this five ...	Ohio	3
7	B006K2ZZ7K	A1SP2KVKFXXRU1	0	0	5	1340150400	Great! Just as good as the expensive brands!	This saltwater taffy had great flavors and was ver...	California	2
8	B006K2ZZ7K	A3JRGQVEQN31IQ	0	0	5	1336003200	Wonderful, tasty taffy	This taffy is so good. It is very soft and chewy....	Georgia	2
9	B000E7L2R4	A1MZYO9TZK0BBI	1	1	5	1322006400	Yay Barley	Right now I'm mostly just sprouting this so my cat...	Michigan	4
10	B0017IAPVA	A21BT40VZCCYT4	0	0	5	1351209600	Healthy Dog Food	This is a very healthy dog food. Good for their di...	Florida	4
11	B0001PB9FE	A3HDKO7OW0QNK4	1	1	5	1107820800	The Best Hot Sauce in the World	I don't know if it's the cactus or the tequila or ...	Georgia	1

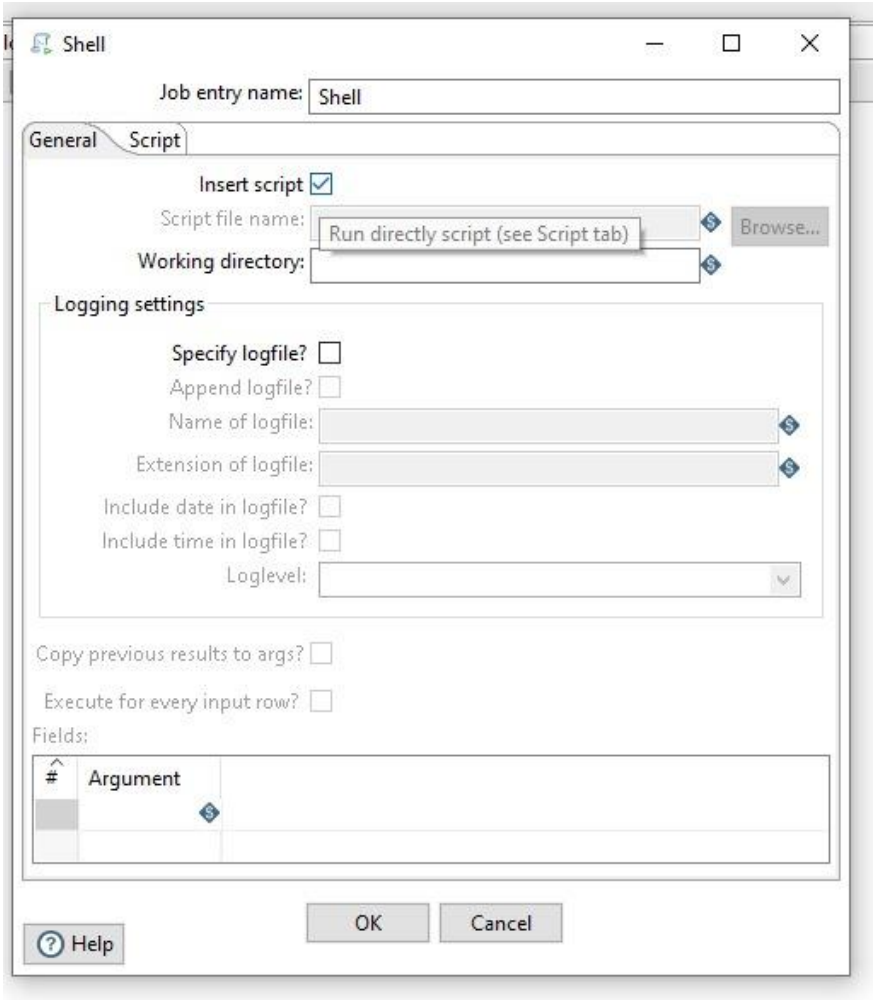
Job 1 : Analyse de sentiment à l'aide d'un script Shell

Input : Table récemment créée contenant les critiques et les métadonnées (incluant la colonne **Quarter**).

1. **Exécution du script de détection de sentiment** :
    - Un job Pentaho est créé pour exécuter un script Shell (.sh) via Pentaho. Ce script appelle un fichier Python (. Py) qui effectue une **analyse de sentiment** sur les critiques dans la colonne **Text**.



Le fichier Python détecte le sentiment (positif, négatif ou neutre) pour chaque critique



sentiment_analysis.py 3

C: > Users > dell > Desktop > sentiment_analysis.py > ...

```python
import mysql.connector
from nltk.sentiment import SentimentIntensityAnalyzer
from nltk import download

# Download VADER lexicon
download('vader_lexicon')

# Initialize VADER sentiment analyzer
sia = SentimentIntensityAnalyzer()

# Function to determine sentiment using VADER
def get_sentiment(text):
    polarity = sia.polarity_scores(text)['compound']  # Get compound score
    if polarity > 0.5:
        return 'positive'
    elif polarity > 0.2:
        return 'little positive'
    elif polarity < -0.5:
        return 'negative'
    elif polarity < -0.2:
        return 'little negative'
    else:
        return 'neutral'

# Connect to the MySQL database
db = mysql.connector.connect(
    host="localhost",
    user="root",
    password="",
    database="review"
)

cursor = db.cursor()
```

```python
# Add Sentiment column if it doesn't exist
try:
    cursor.execute("ALTER TABLE cleaned_reviews ADD COLUMN Sentiment VARCHAR(20)")
    db.commit()
except:
    pass  # Ignore if the column already exists

# Process reviews in batches
batch_size = 50000
offset = 0

while True:
    # Fetch a batch of reviews
    fetch_query = f"SELECT Id, Text FROM cleaned_reviews WHERE Sentiment IS NULL LIMIT {batch_size} OFFSET {offset}"
    cursor.execute(fetch_query)
    reviews = cursor.fetchall()

    # Break if no more rows are left
    if not reviews:
        break

    # Process the batch
    for review in reviews:
        review_id, text = review
        sentiment = get_sentiment(text)
        cursor.execute("UPDATE cleaned_reviews SET Sentiment = %s WHERE Id = %s", (sentiment, review_id))

    # Commit after each batch
    db.commit()
```

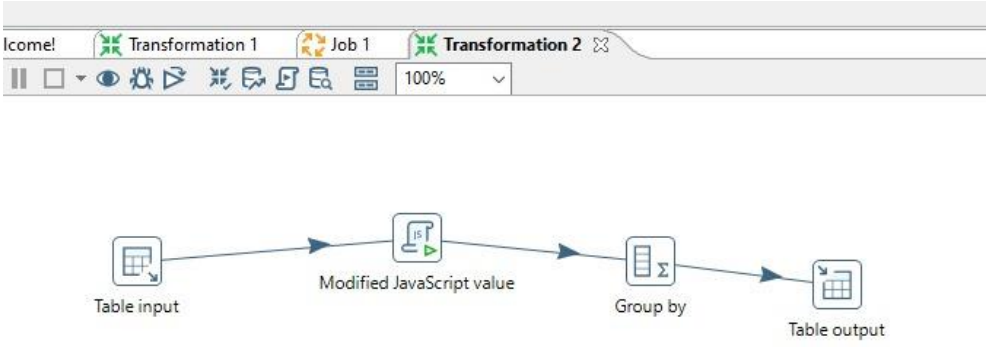
```
    # Commit after each batch
    db.commit()

    # Move to the next batch
    offset += batch_size
    print("batch done")
# Close the connection
cursor.close()
db.close()

print("Sentiment analysis completed and database updated.")
```

2. **Ajout de la colonne "SentimentDetected"** :
    - Une nouvelle colonne **SentimentDetected** est ajoutée à la table, indiquant le sentiment détecté pour chaque revue.

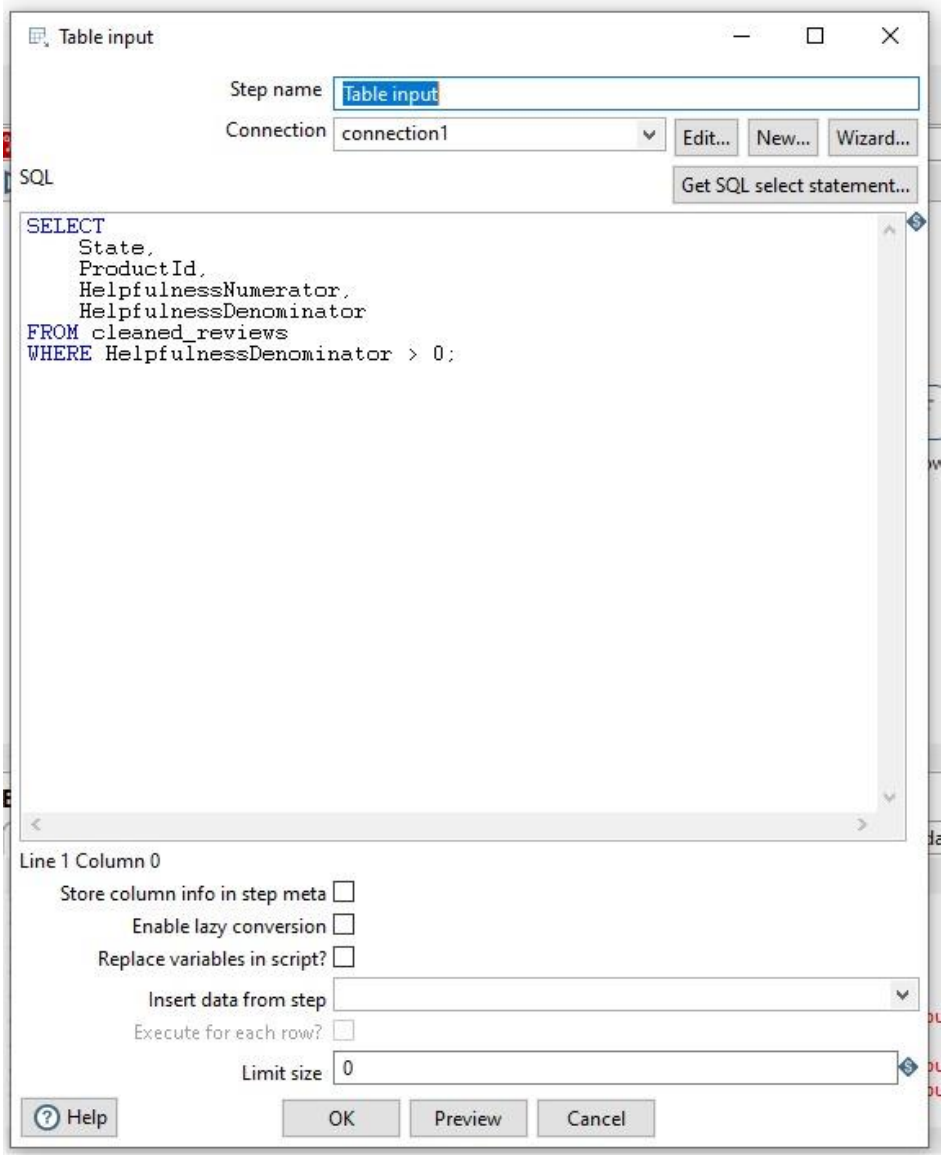
Output : La même table, mais avec l'ajout de la colonne **SentimentDetected**, qui contient l'analyse de sentiment pour chaque critique.

Text	State	Quarter	Sentiment
This is a high quality oolong tea, very tasty, I a...	California	2	positive
Arrived promptly in 2 days. The tea was as advert...	Florida	1	positive
I have been drinking oolong tea for a few years no...	Pennsylvania	1	positive
This product is OK, but truthfully I've had much m...	California	1	positive

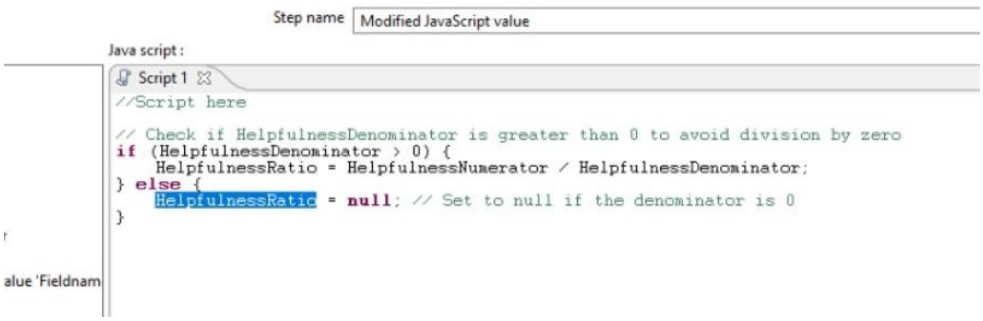
Transformation 2 : Calcul du ratio d'utilité des critiques et agrégation par produit et état



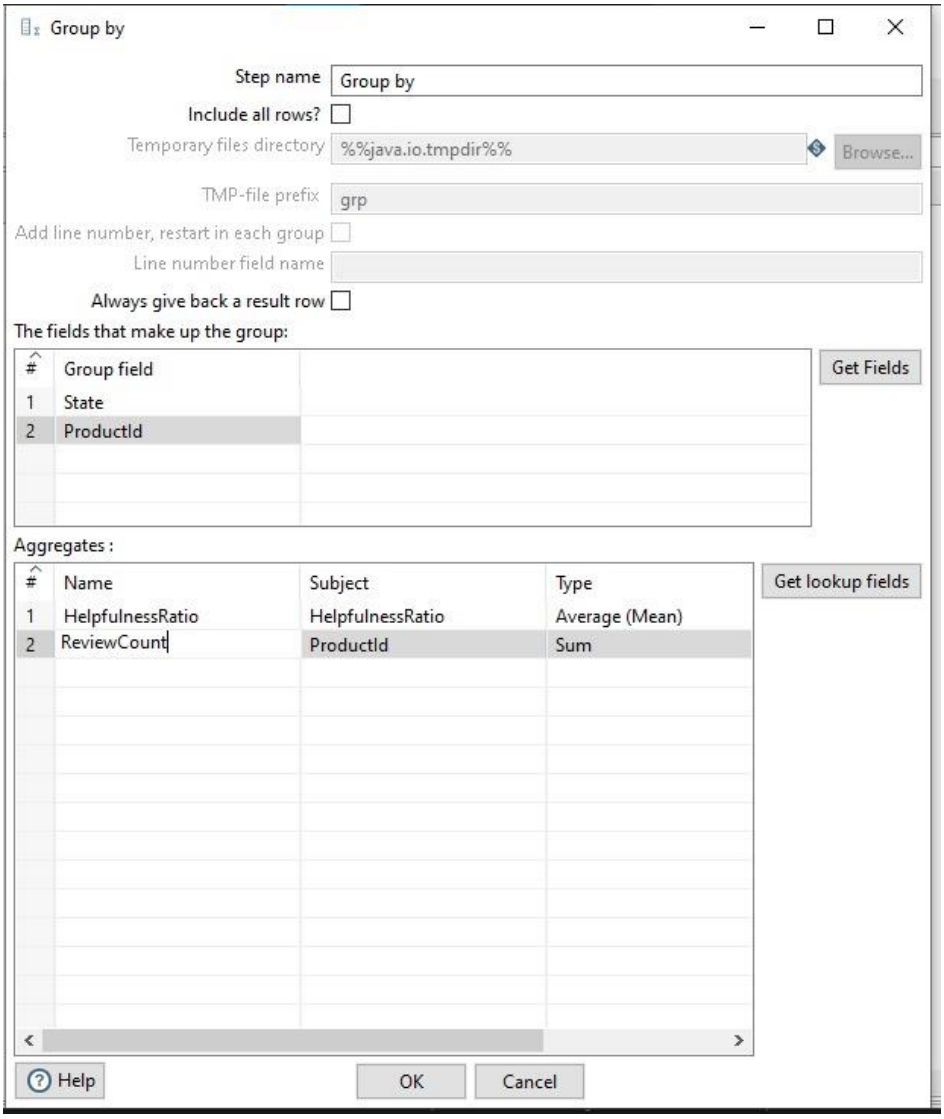
Input : Table modifiée récemment (avec les colonnes **State**, **ProductId**, **HelpfulnessNumerator**, **HelpfulnessDenominator**, et **SentimentDetected**).



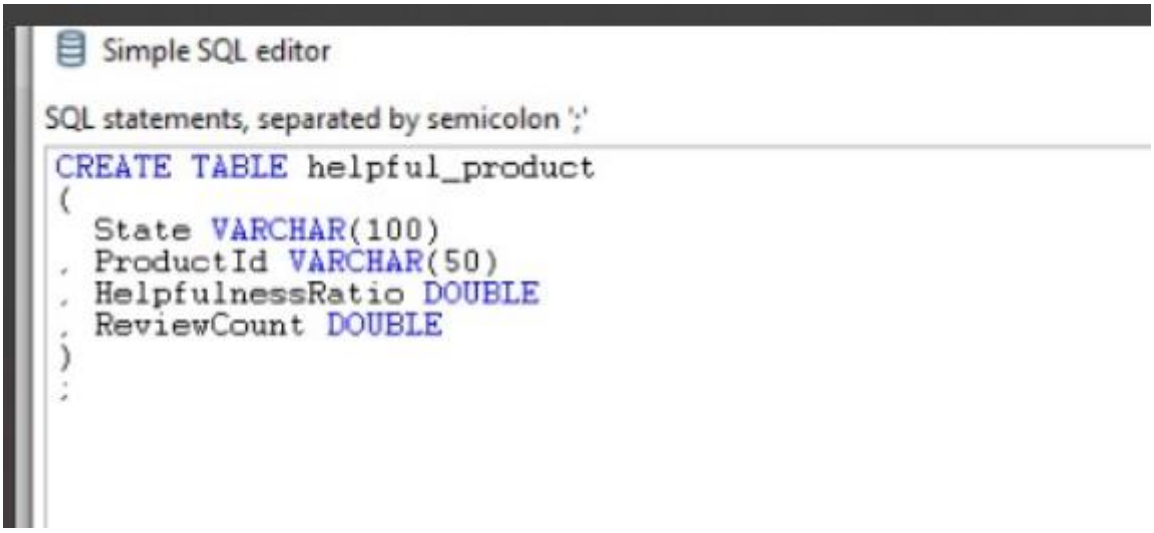
1. **Calcul du ratio d'utilité** :
    - Une nouvelle colonne **Ratio** est ajoutée à la table. Cette colonne est calculée en divisant la valeur de **HelpfulnessNumerator** par **HelpfulnessDenominator**, ce qui donne une mesure de l'utilité perçue de chaque critique.



2. **Agrégation des données** :
    - Les données sont agrégées par **State** et **ProductId**.
    - Deux agrégations sont appliquées :
        - **Moyenne du ratio** (avg(ratio)) : Calcule la moyenne du ratio d'utilité pour chaque combinaison d'état et de produit.
        - **Somme des produits** (sum(ProductId)) : Calcule le nombre total de produits pour chaque combinaison d'état et de produit.



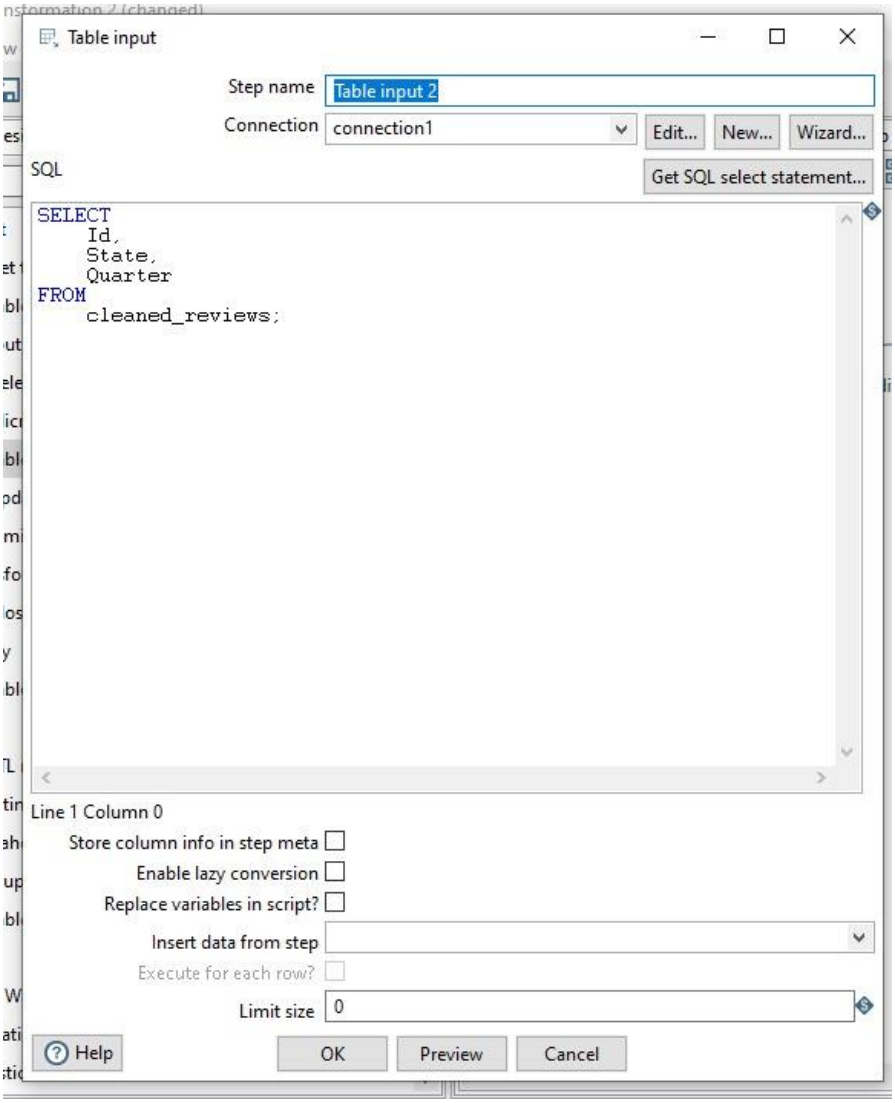
Output : Une table agrégée, avec les colonnes **State**, **ProductId**, la moyenne du **Ratio**, et la somme des **ProductId**, fournissant une vue consolidée de l'utilité des critiques par produit et par état.



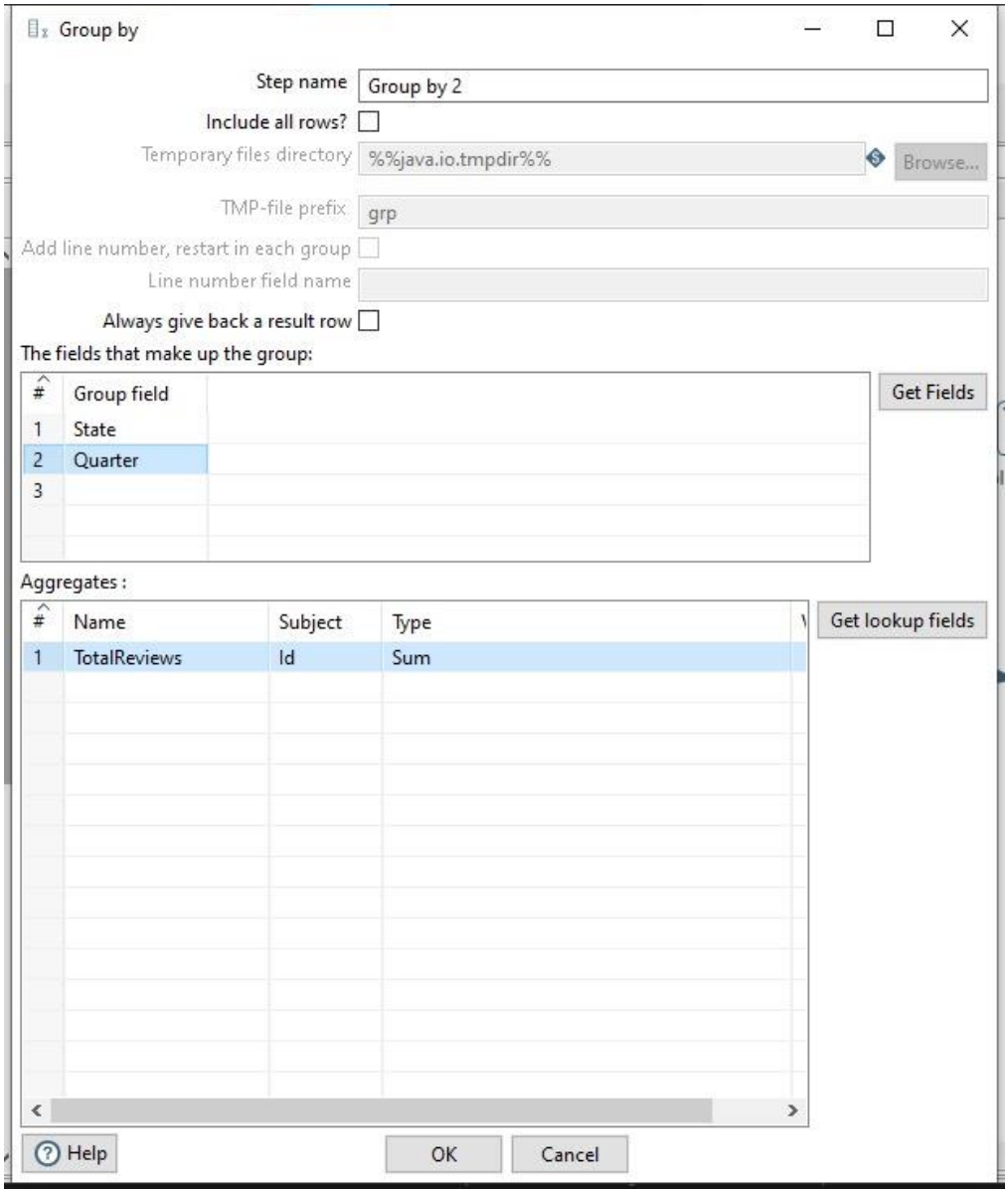
State	ProductId	HelpfulnessRatio	ReviewCount
California	7310172001	0.789642857	7
California	7310172101	0.919116666	6
California	B00002N8SM	0.75	1
California	B00004CI84	0.51285	12
California	B00004CXX9	0.50155	18
California	B00004RAMS	0.875	4
California	B00004RAMX	0.5	1
California	B00004RAMY	0.5	4
California	B00004RBDU	0.952385714	7
California	B00004RBDW	1	1
California	B00004RBDZ	1	2
California	B00004RYGX	0.7694875	8
California	B00004S1C5	1	2
California	B00004S1C6	0.92725	2
California	B0000537KC	0.5	1

Transformation 3 : Agrégation des critiques par état et trimestre

Input : Table contenant les colonnes suivantes :



- **Id review**
- **State**
- **Quarter**

1. **Calcul du total des critiques** :
    - On calcule le **nombre total de critiques** en fonction des colonnes **state** et **quarter**.
2. **Groupement par état et trimestre** :
    - Les données sont **groupées** par les colonnes **state** et **quarter** pour obtenir un total consolidé des critiques pour chaque combinaison.
    - 

Output : Une nouvelle table contenante :

- **State** : La région ou l'état associé à la critique.
- **Quarter** : Le trimestre où la critique a été enregistrée.
- **Total reviews** : La somme totale des critiques pour chaque état et trimestre.

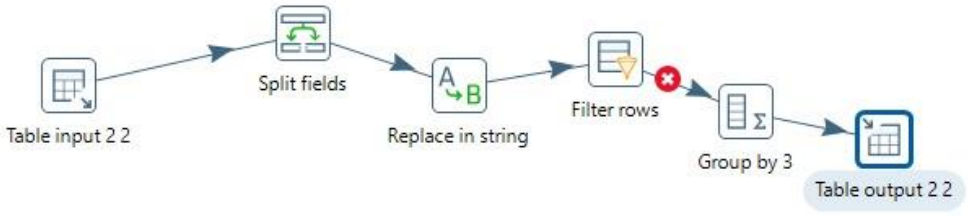
Simple SQL editor

SQL statements, separated by semicolon ';'

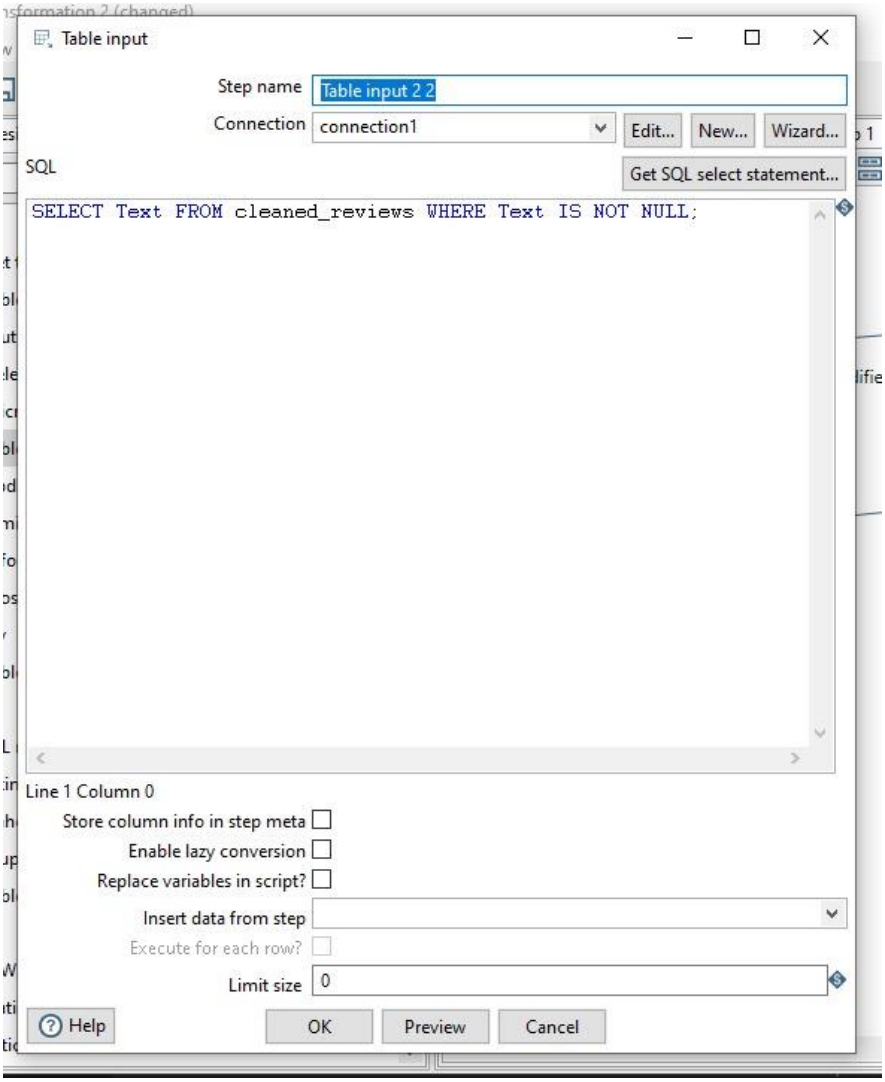
```
CREATE TABLE `state_quarter_reviews `
(
  State VARCHAR(100)
, Quarter VARCHAR(5)
, TotalReviews INT
)
;
```

State	Quarter	TotalReviews
North Carolina	2	13545
North Carolina	3	15437
North Carolina	4	13315
Ohio	1	14419
Ohio	2	13407
Ohio	3	15595
Ohio	4	13570
Pennsylvania	1	14280
Pennsylvania	2	13549
Pennsylvania	3	15407
Pennsylvania	4	13656
Texas	1	14350
Texas	2	13593
Texas	3	15584
Texas	4	13337

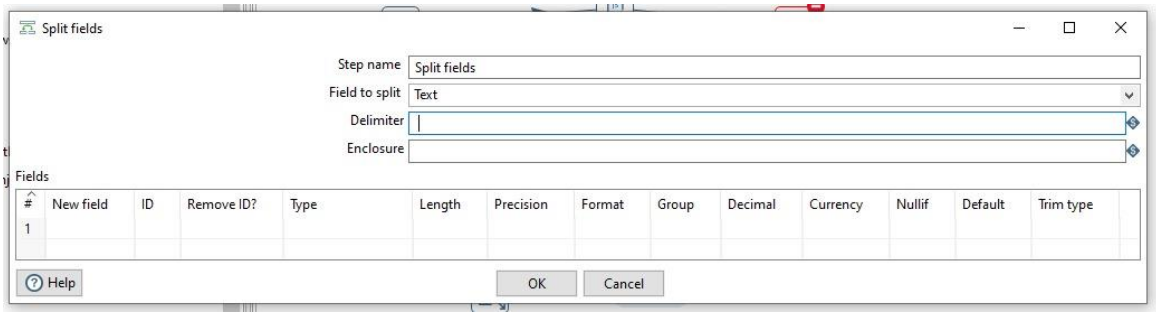
Transformation 4 : Analyse des mots dans les critiques



Input : Table contenant uniquement la colonne **Text** provenant des critiques.

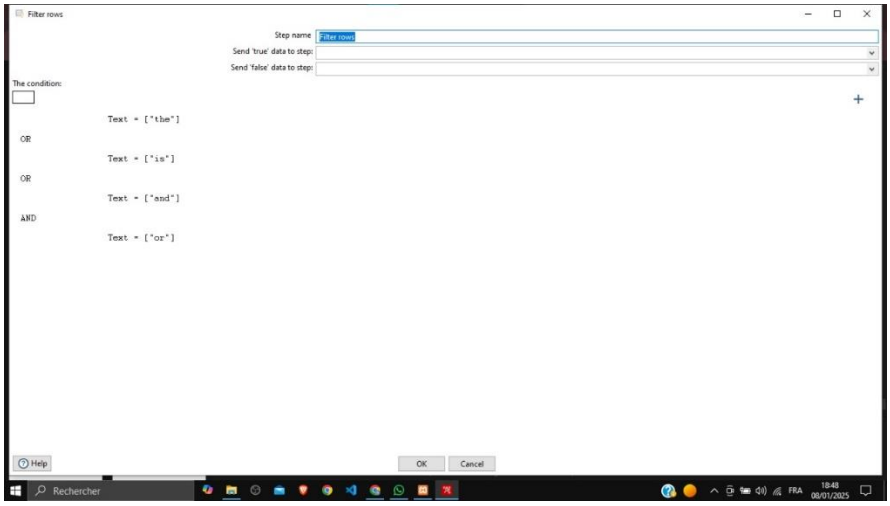


1. **Prétraitement du texte** :
    - **Division du texte** : Les critiques dans la colonne **Text** sont **divisées en mots** individuels à l'aide de l'espace (" ") comme séparateur.

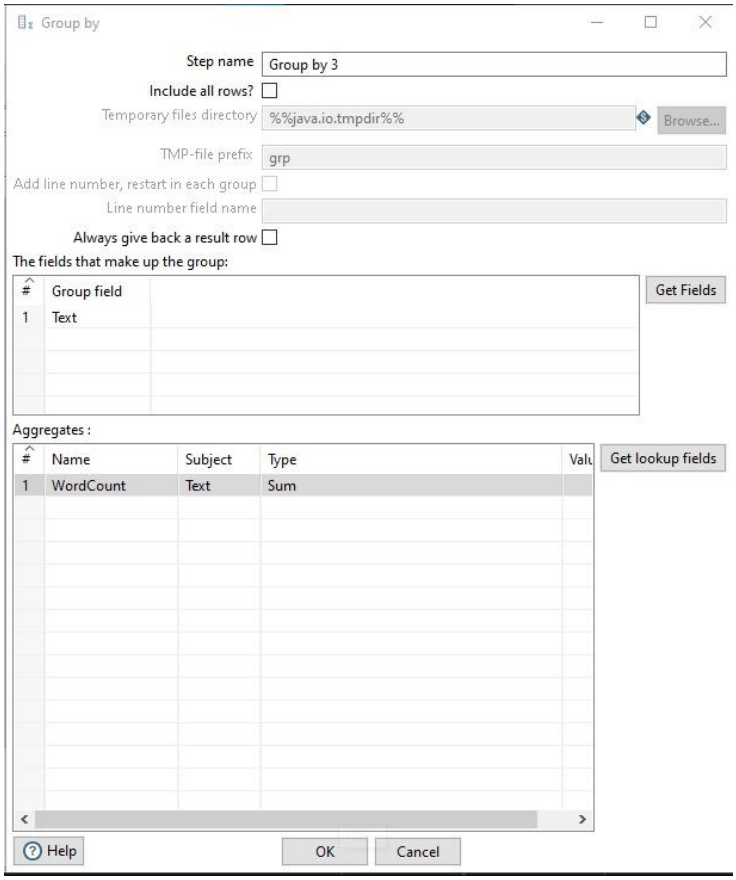


    - **Suppression des mots fréquents** : Les **mots fréquents** ou inutiles (stop words) sont supprimés pour éviter les biais dans l'analyse.

- **Suppression des caractères de ponctuation** : Tous les caractères de ponctuation (par exemple, ".", ",", "!") sont retirés pour nettoyer les données textuelles.

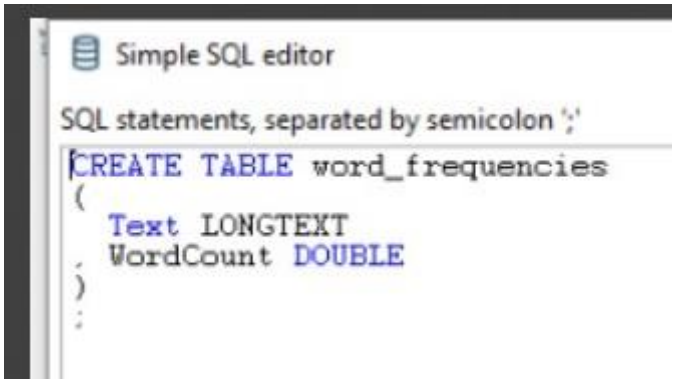


2. **Comptage des mots** :
   - Chaque mot restant est compté (nombre d'occurrences) dans la table.
3. **Groupement par mots** :
   - Les mots sont groupés, et leur somme totale est calculée.



Output : Une nouvelle table contenante :

- **Text** : Les mots distincts extraits des critiques.
- **Wordcount** : Le nombre total d'occurrences pour chaque mot dans les critiques.

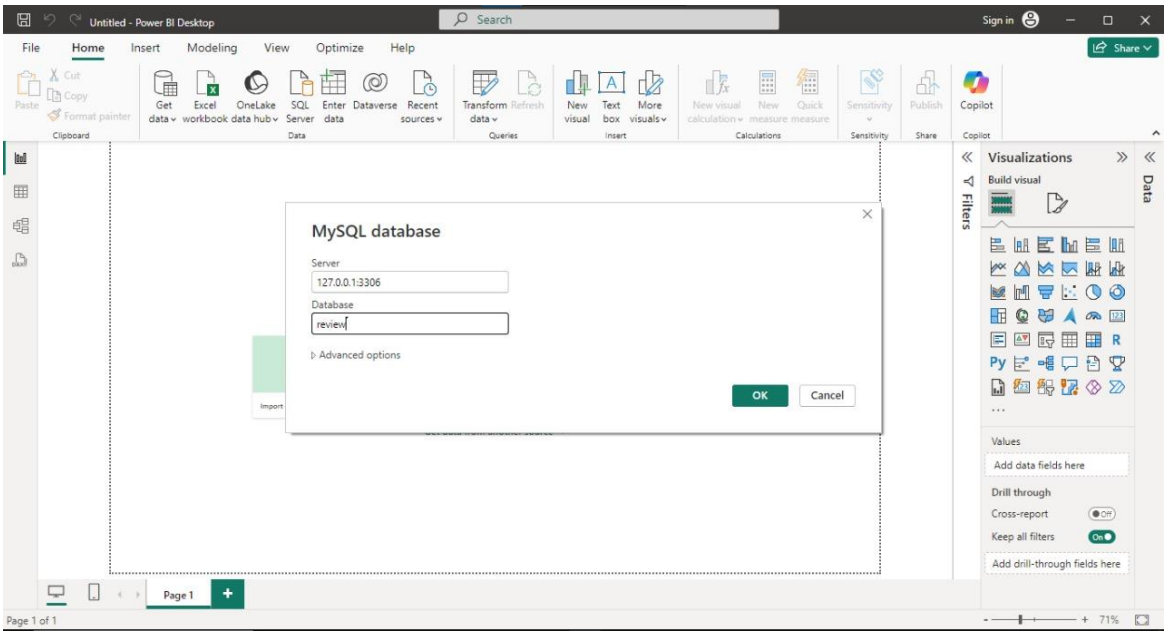


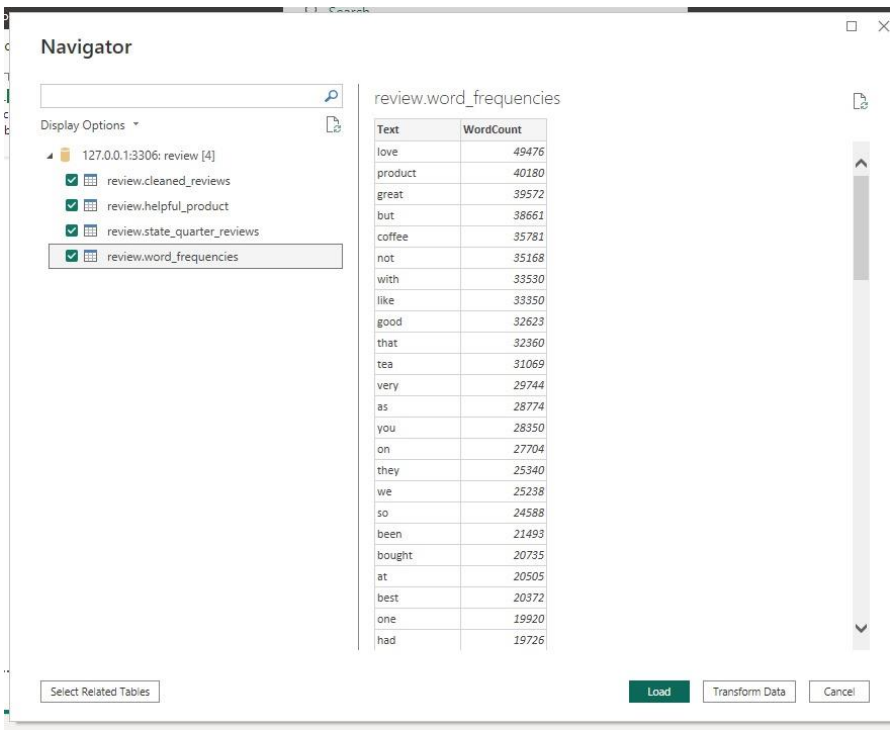
Text	WordCount
love	49476
product	40180
great	39572
but	38661
coffee	35781
not	35168
with	33530
like	33350
good	32623
that	32360
tea	31069
very	29744
as	28774
you	28350
on	27704
they	25340
we	25238
so	24588
been	21493
bought	20735
at	20505
best	20372
one	19920
had	19726
taste	19597

V.RESULTAT OBTENUE

Dans cette partie, nous présentons les différents graphiques générés dans le tableau de bord Power BI, afin de fournir une compréhension claire des résultats obtenus à partir du dataset. Ces visualisations permettent d'extraire des insights clés et de faciliter l'analyse des données.

D'abord se connecter à la database via Power bi :

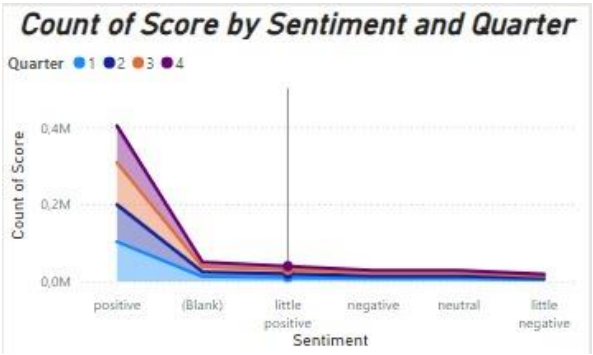




Text	WordCount
love	49476
product	40180
great	39572
but	38661
coffee	35781
not	35168
with	33530
like	33350
good	32623
that	32360
tea	31069
very	29744
as	28774
you	28350
on	27704
they	25340
we	25238
so	24588
been	21493
bought	20735
at	20505
best	20372
one	19920
had	19726

Les différents Graphs choisit :

La carte intitulée **"Somme des Avis Totaux par État"** offre une représentation géographique de l'activité des avis clients à travers les États-Unis. Chaque bulle sur la carte correspond à un état, et la taille de la bulle est proportionnelle au nombre total d'avis soumis depuis cette région. Les états avec des bulles plus grandes indiquent une plus forte implication des clients, mettant en évidence les régions où les utilisateurs sont les plus actifs pour donner leur avis. Cette visualisation met en lumière la densité des avis et peut être utilisée pour identifier les zones clés pour des efforts marketing ciblés, la collecte de retours clients, ou la prise de décisions stratégiques basées sur les tendances géographiques. Les données proviennent du tableau cleaned_reviews, utilisant la colonne State pour les localisations et TotalReviews pour l'agrégation.

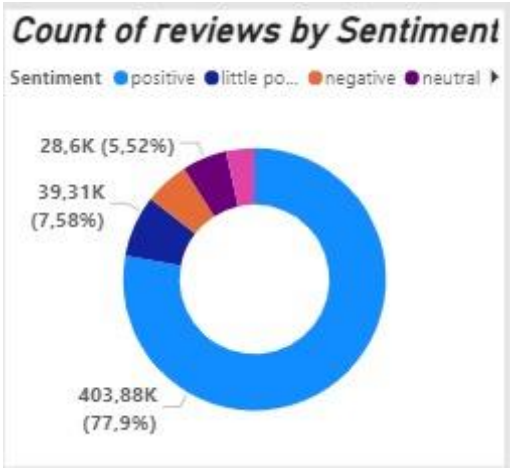


Le graphique intitulé **"Nombre de Scores par Sentiment et Trimestre"** illustre la distribution des scores en fonction des sentiments associés et des trimestres de l'année. L'axe des abscisses (X) représente les différentes catégories de sentiments, telles que **"positive", "neutral", "negative",** etc., tandis que l'axe des ordonnées (Y) affiche le nombre total de scores enregistrés. Les couleurs des courbes différencient les trimestres (1, 2, 3, et 4), permettant de visualiser les variations de sentiments au fil du temps.

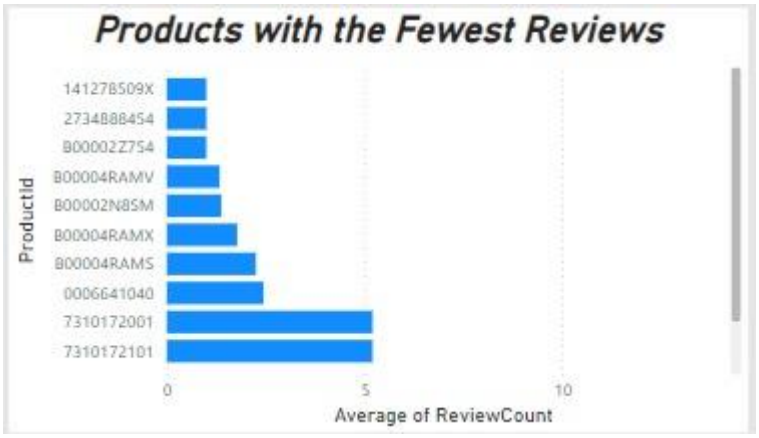
L'image représente une interface de **filtrage dynamique** conçue pour affiner l'analyse des données de manière intuitive et ciblée. Elle comporte trois filtres principaux : le temps, l'état géographique et le trimestre. Le filtre temporel, situé en haut de l'interface, se présente sous forme d'un curseur permettant de sélectionner une plage spécifique de temps. Ce filtre repose sur la colonne Time du jeu de données, permettant à l'utilisateur de se concentrer sur les avis soumis durant une période bien définie En dessous, le filtre par état est proposé sous la forme d'un menu déroulant. Ce filtre, basé sur la colonne State, permet de choisir un ou plusieurs états pour analyser les données de manière régionale. Il offre une approche géographique des avis, en mettant en lumière les particularités et tendances propres à chaque région sélectionnée.

L’image montre deux cartes résumant des métriques clés issues des données analysées. La première carte affiche le nombre total de produits identifiés dans le jeu de données, représenté par la valeur "**74,258K**", ce qui indique une large variété de produits évalués. La deuxième carte présente le nombre total d'avis, soit "**298,40K**", illustrant le volume global de retours des clients. Ces cartes permettent une vue d'ensemble rapide et concise des données, mettant en évidence l'étendue des produits et la participation des utilisateurs via leurs avis.

Ce graphique en anneau illustre la répartition des avis selon les sentiments identifiés dans la colonne Sentiment de la table cleaned_reviews. Les sentiments représentés sont : positif, légèrement positif, neutre, légèrement négatif et négatif. Chaque section de l'anneau montre la proportion des avis appartenant à chacune de ces catégories, offrant une vue d'ensemble sur la tonalité générale des retours des utilisateurs.



Ce **nuage de mots** représente les termes les plus fréquemment utilisés dans les avis, basé sur la colonne WordCount de la table word_frequencies. La taille de chaque mot reflète sa fréquence d'apparition, les mots les plus grands étant ceux qui reviennent le plus souvent. Cette visualisation permet de saisir rapidement les termes clés exprimés par les utilisateurs dans leurs retours, offrant ainsi un aperçu des thèmes ou sujets les plus récurrents dans les avis analysés.



Ce graphique met en évidence les produits ayant reçu le moins d'avis, en se basant sur la colonne ProductId pour identifier les produits et sur la moyenne du nombre d'avis (Average of ReviewCount). Les barres représentent la quantité moyenne d'avis pour chaque produit, permettant de repérer les articles les moins évalués. Cette visualisation est utile pour identifier les produits nécessitant davantage de visibilité ou d'engagement auprès des utilisateurs.

Final Dashboard:

VI.CONCLUSION

Ce projet de Business Intelligence (BI) a permis de transformer des données brutes issues des avis clients en informations exploitables pour une prise de décision éclairée. Grâce à une architecture robuste et cohérente, les données ont été collectées, nettoyées, transformées, stockées et visualisées de manière optimale. Le choix des outils a joué un rôle central dans le succès du projet : Pentaho a facilité l'extraction et la transformation des données via un processus ETL efficace, tandis que MySQL a fourni une solution de stockage centralisé, garantissant une gestion structurée et rapide des données. Enfin, Power BI a permis de créer des tableaux de bord interactifs, mettant en lumière des indicateurs clés comme le ratio d'utilité, les tendances des avis et les comportements des utilisateurs. Cette architecture modulaire offre une grande flexibilité, permettant l'intégration future de nouvelles sources de données et une scalabilité adaptée aux besoins croissants. En rendant les données accessibles et exploitables rapidement, cette solution renforce la capacité de l'entreprise à identifier des tendances, à améliorer l'expérience client et à ajuster ses stratégies de manière proactive, démontrant ainsi la puissance et la pertinence de cette approche BI intégrée.