REPUBLIQUE DU SENEGAL

Un Peuple, Un But, Une foi



UNIVERSITE GASTON BERGER

L'excellence au service du développement



Rigueur et persévérance

DATA ENGINEERING

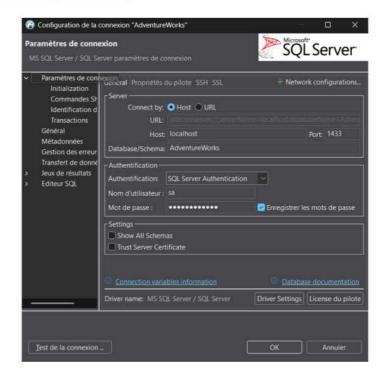
Final Projet
Spark-ETL-Power BI

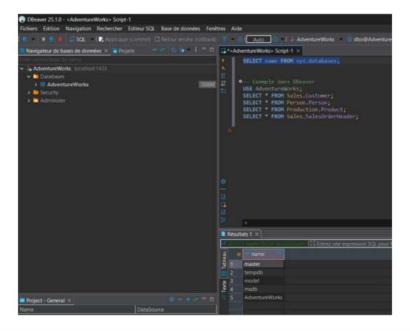
Version	Date	Auteur	Commentaires
0.1	15.06.2025	Mouhamadou Mourtala MBOW Momar Talla Salla	Version initiale

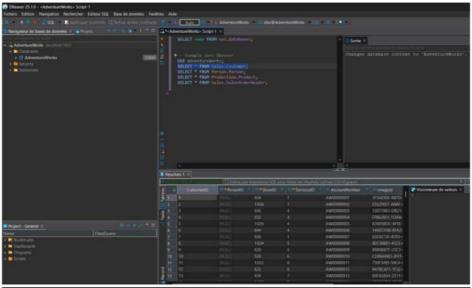
I. Etape 1 : Préparation de l'environnement :

- 1. Cloner le dépôt GitHub et lancer le conteneur MSSQL :
- → git clone https://github.com/sopeKhadim/etl.git
- → cd etl
- → docker-compose up -d

- → DBeaver : Pour visualiser la base AdventureWorks
 - Configuration dans DBeaver :
 - Type de base : SQL Server
 - Host: localhost
 - Port: 1433
 - Database : AdventureWorks (ou master au début)
 - User : sa
 - Password : password123?







→ Téléchargeons et Copions les drivers mssql-jdbc-12.10.0.jre11.jar et postgresql-42.7.2.jar dans ton dossier etl/libs

- → Lançons le conteneur PostgreSQL
- → docker run --name some-postgres -e POSTGRES PASSWORD=postgres -p 5432:5432 -d postgres

```
DELL XPS@iPadPro MINGW64 /c/Rep_Travail/projetFinalDataEngineer/etl (master)

$ docker run --name some-postgres -e POSTGRES_PASSWORD=postgres -p 5432:5432 -d postgres
Unable to find image 'postgres:latest' locally
latest: Pulling from library/postgres

$ f4ce9941f6e3: Pull complete

$ b47a445a47f0: Pull complete

$ e928d00623a6e: Pull complete

$ e928d00623a6e: Pull complete

$ e9549sc11b3: Pull complete

$ da667da3f26b: Pull complete

$ da667da3f26b: Pull complete

$ da667da3f26b: Pull complete

$ d64e8f1b2b243: Pull complete

$ d64e8f1b2b243: Pull complete

$ 64e8f1b2b243: Pull complete

$ 663e696408a9b37: Pull complete

$ 603e69fcdd8e: Pull complete

$ 03e69fcdd8e: Pull complete

$ 03
```

→ Créons une base de données **PostgreSQL** nommée **adw_dw** en se connectant au conteneur pour y charger les données transformées :

```
DELL XPS@iPadPro MINGw64 /c/Rep_Travail/projetFinalDataEngineer/etl (master)

$ docker exec -it some-postgres psql -U postgres
psql (17.5 (Debian 17.5-1.pgdg120+1))
Type "help" for help.

postgres=# CREATE DATABASE adw_dw;

\q
CREATE DATABASE
```

- → Lançons un conteneur **Spark** avec montage du volume
- → docker run -it --rm -v ~/etl:/app -p 4040:4040 bitnami/spark bash

```
jarvis@iPadPro:/mnt/c/Rep_Travail/projetFinalDataEngineer$
jarvis@iPadPro:/mnt/c/Rep_Travail/projetFinalDataEngineer$
jarvis@iPadPro:/mnt/c/Rep_Travail/projetFinalDataEngineer$
jarvis@iPadPro:/mnt/c/Rep_Travail/projetFinalDataEngineer$
docker run -it --rm -v ~/etl:/app -p 4040:4040 bitnami/spark bash
spark 18:59:14.50 INFO ==> Welcome to the Bitnami spark container
spark 18:59:14.50 INFO ==> Subscribe to project updates by watching https://github.com/bitnami/containers
spark 18:59:14.50 INFO ==> Did you know there are enterprise versions of the Bitnami catalog? For enhanced secure software supply chair
e https://www.arrow.com/globalecs/na/vendors/bitnami/ for more information.
spark 18:59:14.51 INFO ==>
spark@a63b7476f033:/opt/bitnami/spark$ __
```

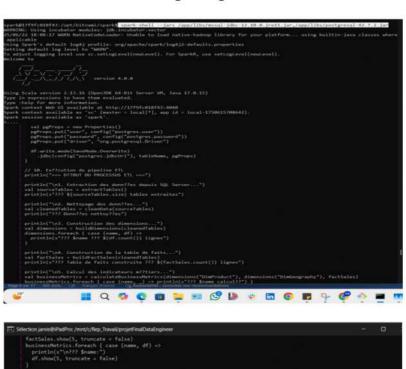
- → Vérifions dans le conteneur que les JAR sont bien là :
- → ls /app/libs

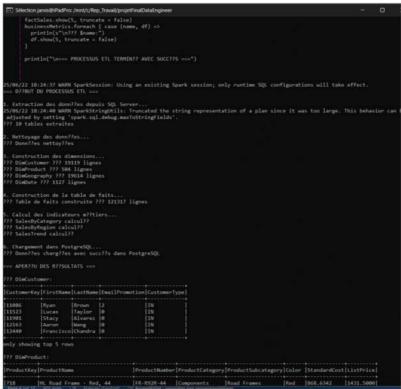
```
spark@37329236615f:/opt/bitnami/spark$ ls /app/libs
mssql-jdbc-12.10.0.jre11.jar postgresql-42.7.2.jar
spark@37329236615f:/opt/bitnami/spark$ _
```

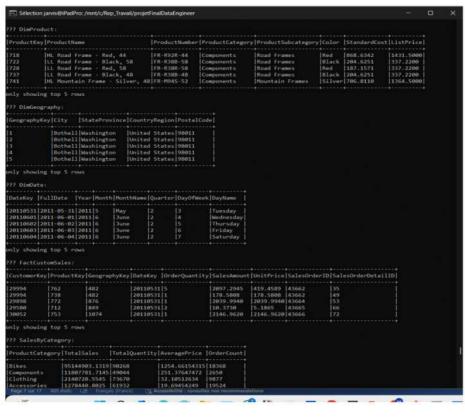
- → Lancer spark-shell avec le bon chemin vers le JAR :
- → spark-shell --jars /app/libs/mssql-jdbc-12.10.0.jre11.jar,/app/libs/postgresql-42.7.2.jar

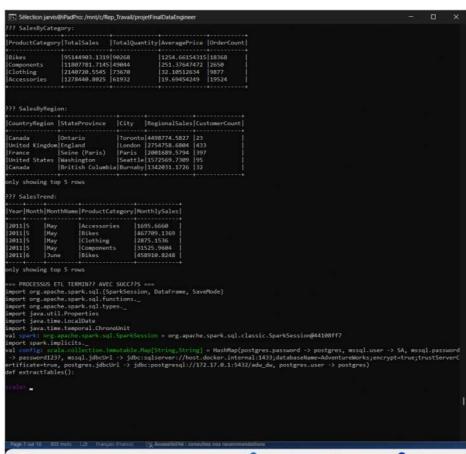
II. <u>Code ETL avec Spark : Script Scala/Spark pour extraire, transformer et charger les données transformées dans PostgreSQL</u>

→ Collons le code ETL dans le prompt scala>

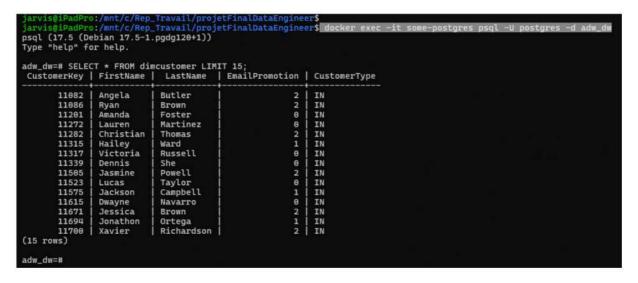








- → Voici des commandes SQL à exécuter dans notre psql (PostgreSQL) pour vérifier l'existence et voir un aperçu de toutes les tables supposées chargées par notre ETL.
- → Connectons-nous dans notre base PostgreSQL et faisons des requêtes:
- → docker run --name some-postgres -e POSTGRES_PASSWORD=postgres -p 5432:5432 -d postgres
- → docker exec -it some-postgres psql -U postgres -d adw dw
- → SELECT * FROM dimcustomer LIMIT 15;



→ SELECT * FROM dimcustomer LIMIT 3;

1	Adjustable Race	AR-5381	Uncategorized	Uncategorized	N/A	0.0000	0.0000
2	Bearing Ball	BA-8327	Uncategorized	Uncategorized	N/A	0.0000	0.0000
3	BB Ball Bearing	BE-2349	Uncategorized	Uncategorized	N/A	0.0000	0.0000
4	Headset Ball Bearings	BE-2908	Uncategorized	Uncategorized	N/A	0.0000	0.0000
316	Blade	BL-2036	Uncategorized	Uncategorized	N/A	0.0000	0.0000

→ SELECT * FROM dimgeography LIMIT 3;

→ SELECT * FROM dimdate LIMIT 3;

DateKey	FullDate	Year	Month	MonthName	Quarter	DayOfWeek	DayName
20120307	2012-03-07	2012	3	March	1	4	Wednesday
20120308	2012-03-08	2012	3	March	1	5	Thursday
20120309 (3 rows)	2012-03-09	2012	3	March	1	6	Friday

→ SELECT * FROM factcustomsales LIMIT 3;

CustomerKey	ProductKey	GeographyKey	DateKey	OrderQuantity	SalesAmount	UnitPrice	SalesOrderID	SalesOrderDetailID
29994	762	482	20110531	5	2097.2945	419.4589	43662	35
29994	738	482	20110531	1	178.5808	178.5808	43662	49
29898	772	876	20110531	1	2039.9940	2039.9940	43664	53
3 rows)								

→ Et cetera

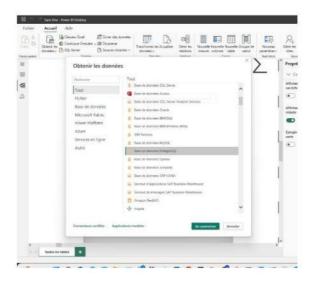
→ \dt

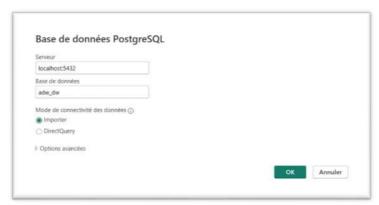
	\dt List of rela	tions		
Schema	ichema Name		Owner	
public	dimcustomer	table	postgres	
public	dimdate	table	postgres	
public	dimgeography	table	postgres	
public	dimproduct	table	postgres	
public	factcustomsales	table	postgres	
public	salesbycategory	table	postgres	
public	salesbyregion	table	postgres	
public	salestrend	table	postgres	

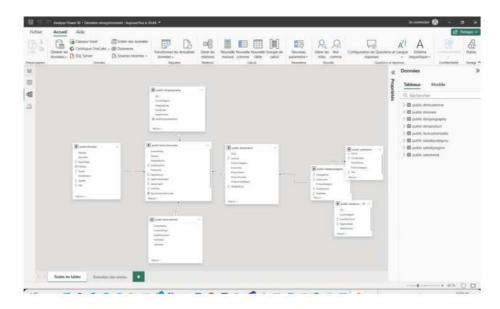
La base de données adw_dw.sql et les vues matérialisées (agrégations)

✓ Maintenant l'extraction, la transformation et le chargement dans PostgreSQL est terminé !!

III. Modèle de Données : Schéma relationnel entre les tables PostgreSQL :







La base de données adw_dw.sql et les vues matérialisées agregation (salesbycategory, salesbyregion, salestrend)

MODÉLISATION EN ÉTOILE RÉUSSIE......

- → On peut récupère les données chargées dans la base de données postgreSQL adw dw après
- → docker exec -t some-postgres pg_dump -U postgres -F p adw_dw > /mnt/c/Rep_Travail/projetFinalDataEngineer/adw_dw.sql

jarvis@iPadPro:/mnt/c/Rep_Travail/projetFinalDataEngineer\$ docker exec -t some-postgres pg_dump -U postgres -F p adw_dw > /mnt/c/Rep_Travail/projetFinalDataEngineer/adw_dw.sql jarvis@iPadPro:/mnt/c/Rep_Travail/projetFinalDataEngineer\$ ls /mnt/c/Rep_Travail/projetFinalDataEngineer/adw_dw.sql /mnt/c/Rep_Travail/projetFinalDataEngineer/adw_dw.sql

La base de données adw_dw est sur mon Git

IV. Documentation : Instructions d'utilisation des tableaux de bord Power BI

S Objectif

Ce tableau de bord permet d'analyser les ventes de l'entreprise selon plusieurs dimensions :

- Ventes dans le temps
- · Répartition par catégorie et produit
- Analyse client
- Analyse géographique
- · Performance et comparaison des ventes par période

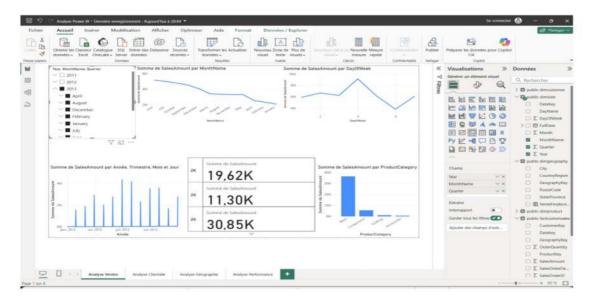
1. Navigation générale

- Le rapport se compose de plusieurs pages (onglets en bas).
- Chaque page contient des graphiques dynamiques et filtres interactifs.

3. Analyse temporelle des ventes

■ Graphique : Évolution des ventes

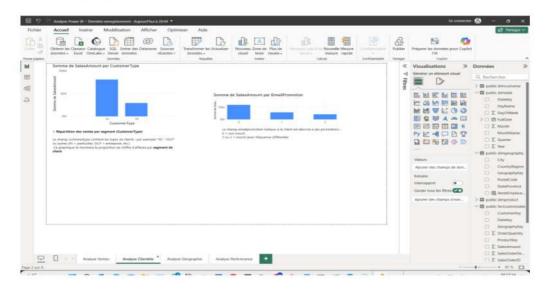
- Permet de suivre les ventes par année, trimestre, mois ou jour.
- Utilise les segments (filtres en haut ou sur le côté) pour :
 - o Sélectionner une année, un mois ou un trimestre spécifique.
 - o Comparer des périodes entre elles.
- Visualise les ventes par catégorie de produit.
- Passe la souris sur un élément pour voir le chiffre d'affaires total, la quantité vendue, et le prix moyen.



22 3. Analyse clientèle

■ Graphique : Répartition des ventes par segment

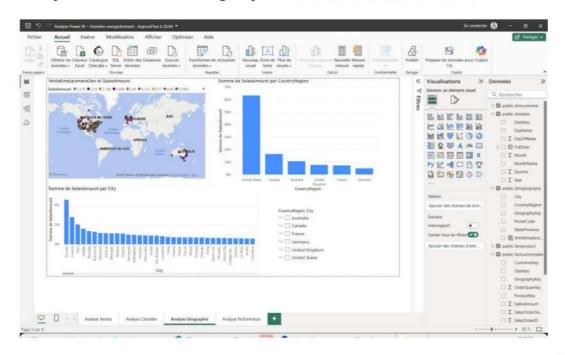
- Affiche la part des ventes par type de client (ex : Regular, VIP...).
- Mesure la fidélité via le nombre de commandes ou le volume d'achat.

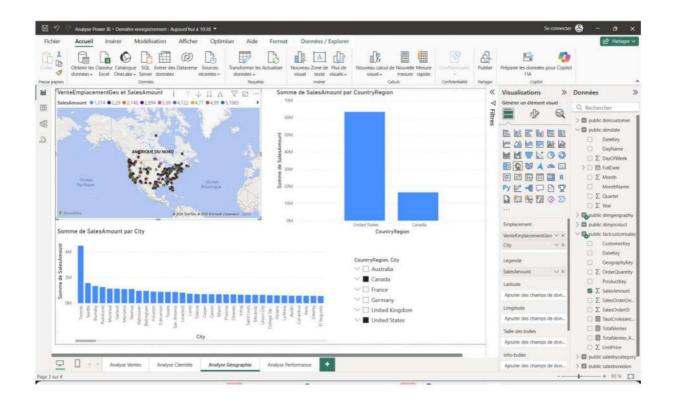


3. 4. Analyse géographique

Q Carte interactive

- Affiche les ventes par région et par ville.
- Utilise la colonne calculée EmplacementGeo pour la géolocalisation.
- Clique sur une ville ou une région pour filtrer tous les autres visuels.





5. Performance commerciale

Ⅲ Graphique : Taux de croissance annuelle

- Compare les ventes de l'année sélectionnée avec l'année précédente.
- Affiche un indicateur de croissance (%) basé sur la formule DAX :
- → TauxCroissance = DIVIDE([TotalVentes] [TotalVentes_AnneePrecedente], [TotalVentes AnneePrecedente])

