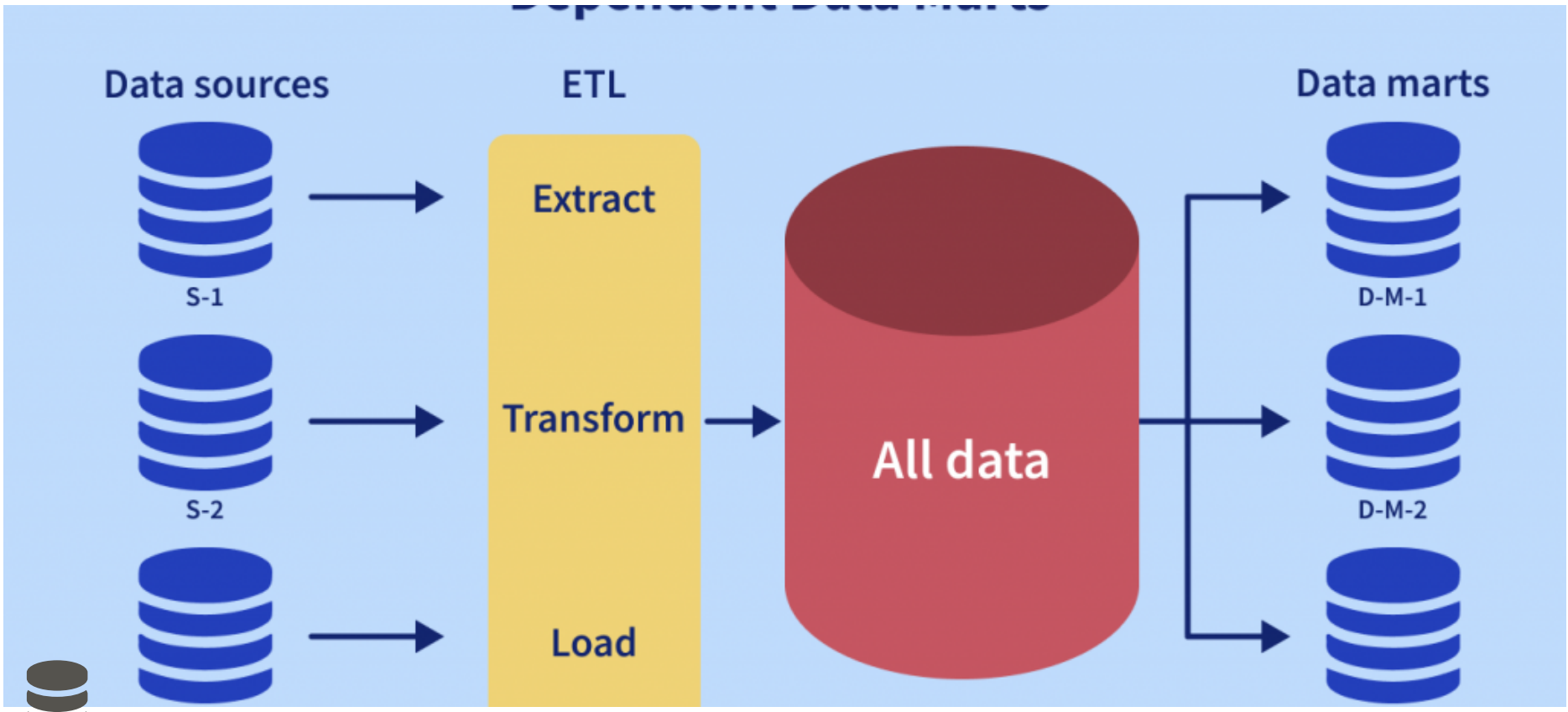


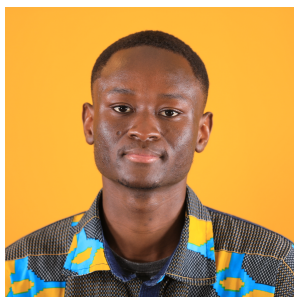
Optimisation de l'ETL et mise en œuvre d'un entrepôt de données pour la visualisation des données

Scenario :

La société X, une plateforme e-commerce florissante, propose une large gamme de produits dans différentes catégories, tels que l'électronique, le mobilier, les épiceries, les vêtements et les livres. Avec la croissance de l'entreprise, les données générées ont également augmenté. Afin d'optimiser les opérations commerciales et de mieux comprendre le comportement des clients, l'équipe Data de la société X est chargée de mettre en place une solution solide d'entrepôt de données.

L'équipe Data, composée de :






TARIFI Hicham	YASSINE Essadi	YONLI Fidel	HARRATI Yassine
Chef de projet	Data Engineer	Data Analyst	Data Analyst

Objectifs :

- **S'assurer que toutes les données sont traitées conformément au RGPD.**
- **Choix des outils et technologies qui seront utilisés pour la collecte, le stockage, le traitement, et la mise à disposition des données.**
- **Implémentation d'un Job ETL optimisé et la gestion des erreurs du job.**
- **Construction de la zone de staging.**
- **Modélisation de l'entrepôt de données.**
- **Création de 2 data marts :**

    - **Etudier la démographie des utilisateurs, les tendances d'inscription , analyse la rétention et le taux de désabonnement des utilisateurs.**
    - **Analyser les ventes mensuelles/trimestrielles/annuelles, les produits et catégories les plus vendus, et les tendances des ventes dans le temps.**
- **Gestion de l'accès basé sur les rôles aux data marts.**
- **Visualisation des données résultant du datamarts.**

Livrables :

- **Un rapport qui décrit les détails techniques du projet, y compris le workflow ETL, les schémas de données, les transformations, les connexions aux sources de données, le schéma de l’entrepôt de données, les datamarts aussi que la visualisation.**
- **Code source de ETL.**
- **Fichier .pbix.**

Les étapes suivies durant le projet :

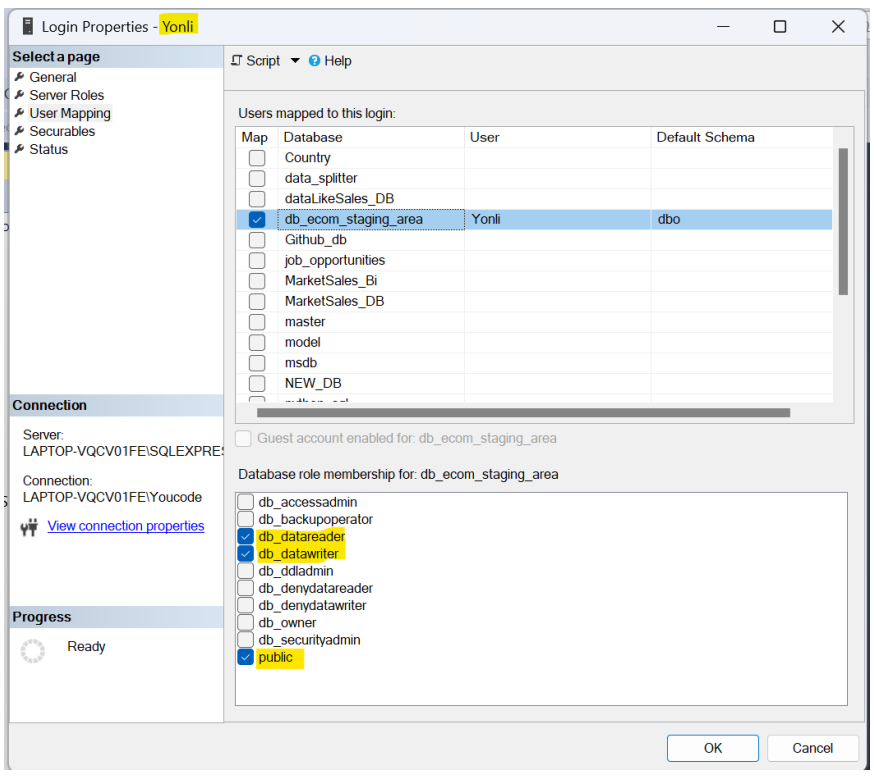
- **Choix des outils et technologies.**

Talend : un outil d'intégration de données puissant et polyvalent conçu pour faciliter l'extraction, la transformation et le chargement (ETL) des données.

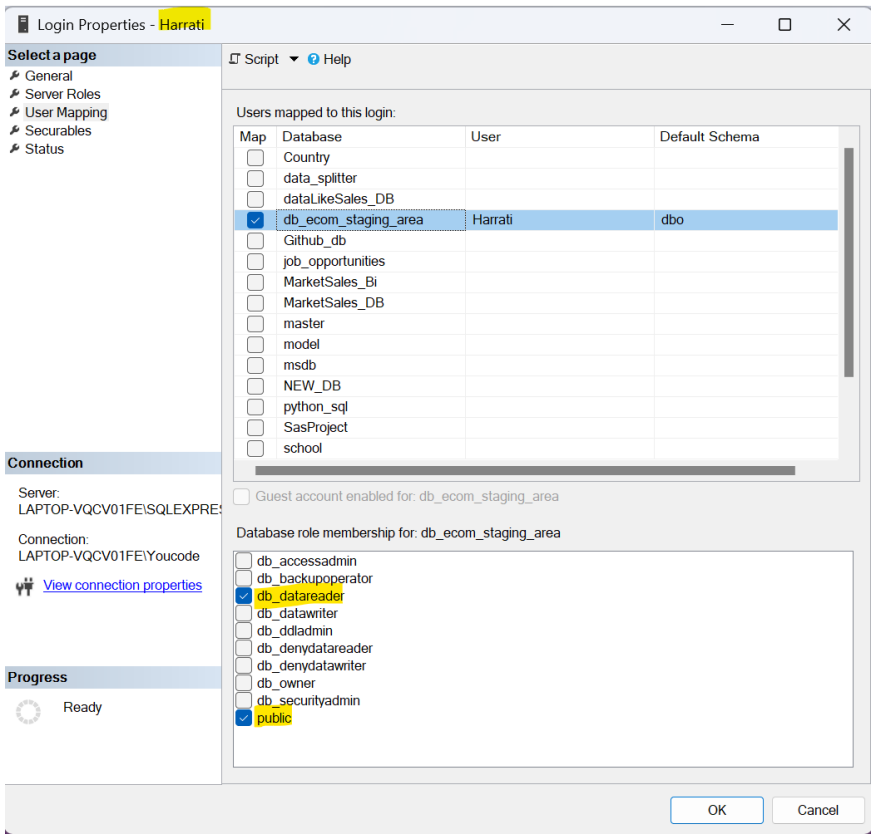
SQL Server : un système de gestion de base de données relationnelle (SGBDR) développé par Microsoft.

Power BI : une suite d'outils d'analyse de données et de visualisation développée par Microsoft. Elle permet aux utilisateurs de transformer les données brutes en informations visuelles significatives.

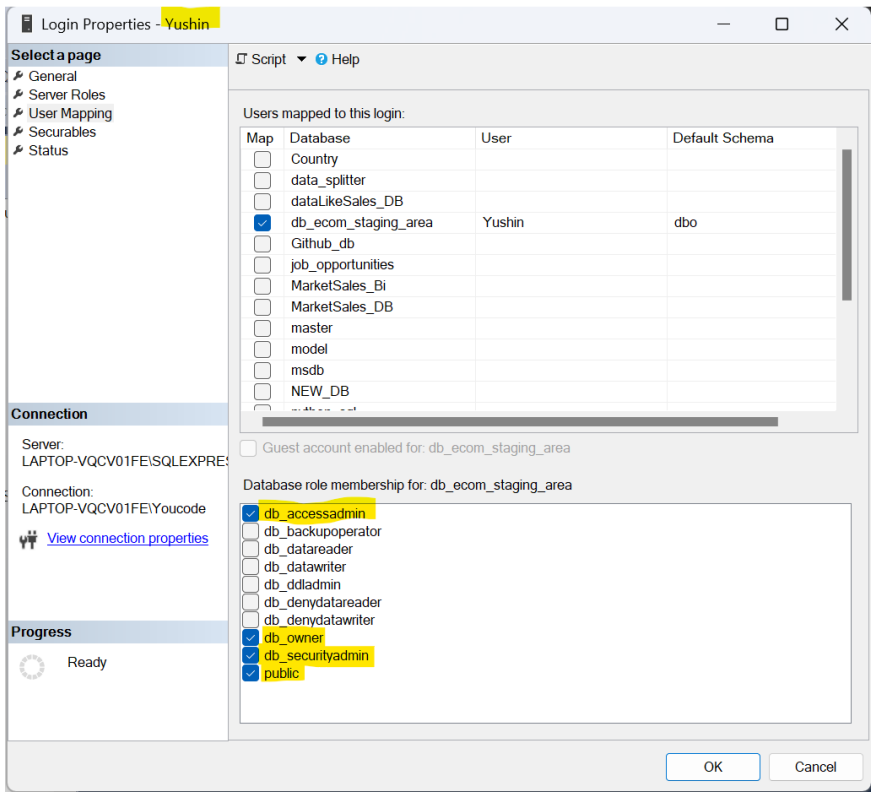
- **Gestion d'accès a la base de données.**



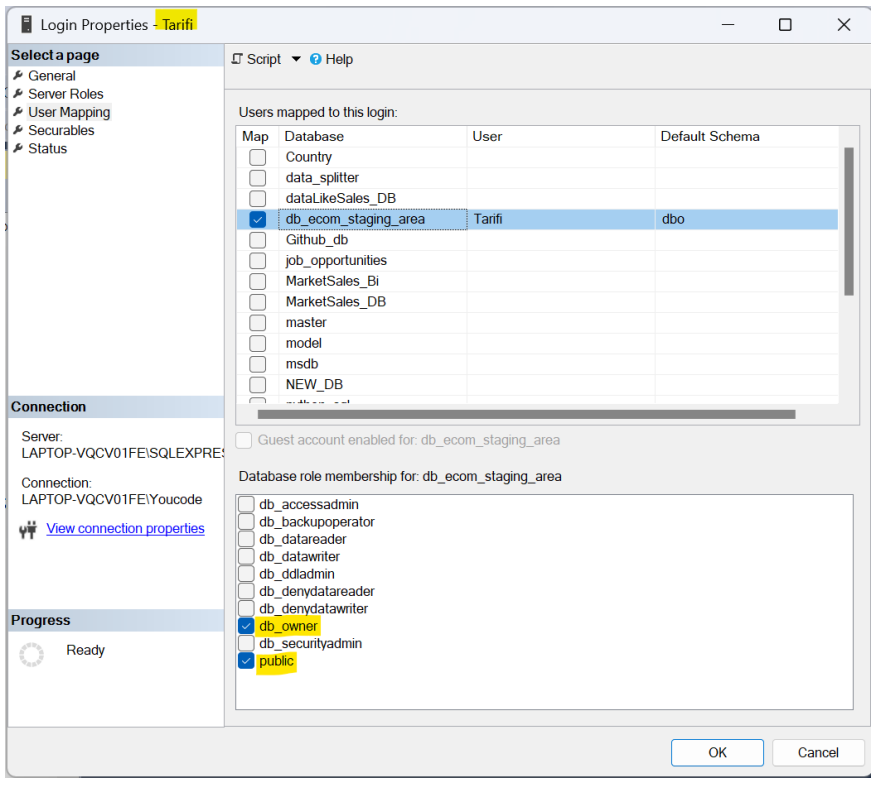
Nous avons créé un utilisateur nommé "**Yonli**" qui bénéficie d'un accès à la base de données "**db_ecom_staging_area**" avec des privilèges lui permettant de consulter, d'ajouter, de supprimer et de modifier des données dans toutes les tables.



Nous avons créé un utilisateur nommé "**Harrati**" qui bénéficie d'un accès à la base de données "**db_ecom_staging_area**" avec des privilèges lui permettant seulement de consulter les données dans toutes les tables.

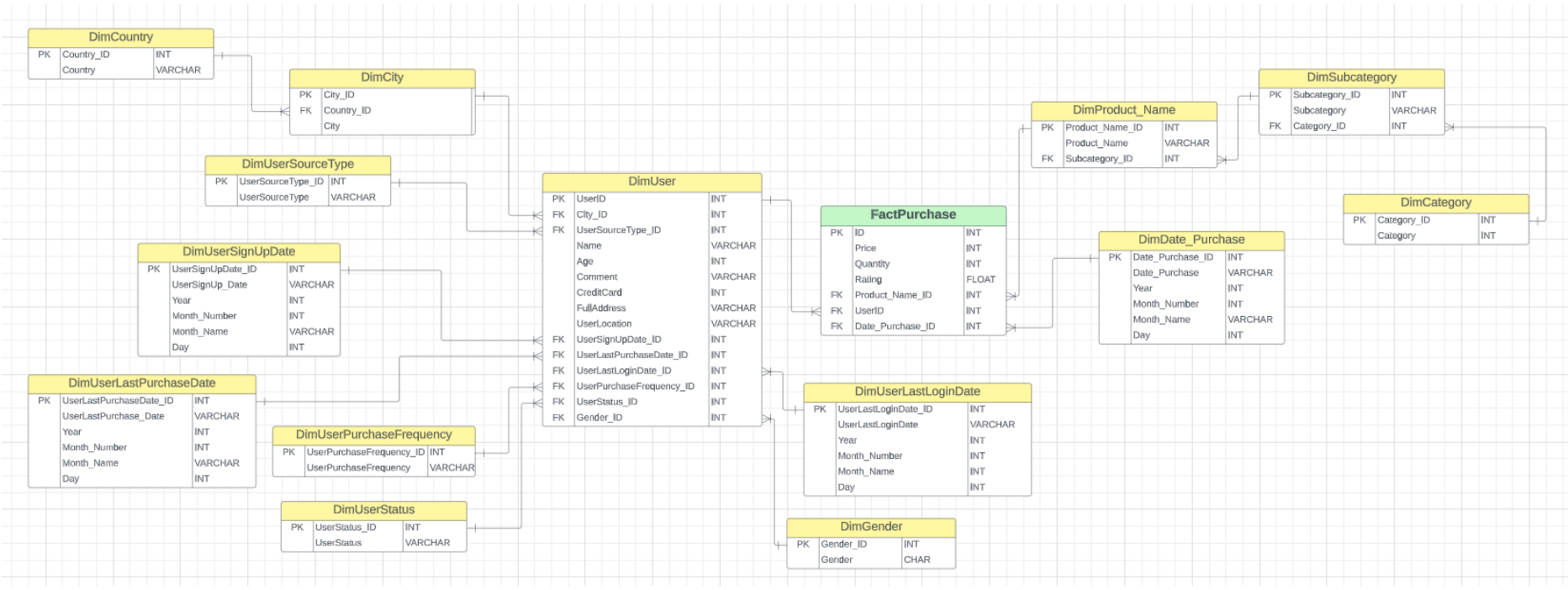


Nous avons créé un utilisateur nommé "**Yushin**" qui bénéficie d'un accès totale à la base de données "**db_ecom_staging_area**".



Nous avons créé un utilisateur nommé "**Tarifi**" qui bénéficie d'un accès à la base de données "**db_ecom_staging_area**" avec des privilèges lui permettant d'effectuer toutes les activités de configuration et de maintenance sur la base de données.

- **Modélisation de l'entrepôt de données**



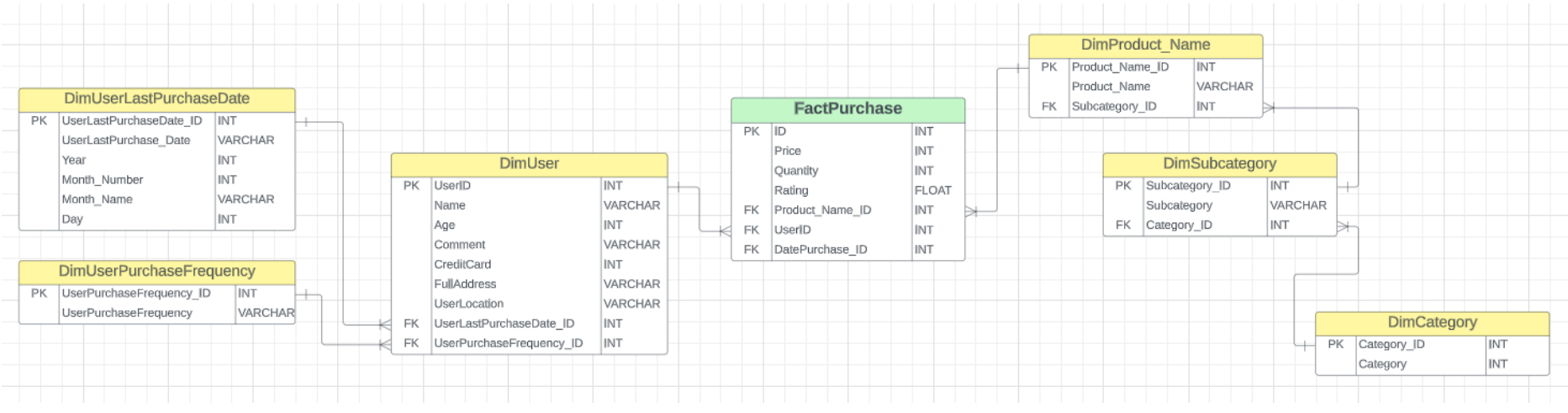
Pour la modélisation de notre entrepôt de données, notre décision s'est portée sur l'utilisation d'une seule table de faits "**FactPurchase**", qui sera directement liée à deux dimensions clés :

- La dimension "**User**", regroupant toutes les informations relatives aux utilisateurs. Cette dimension englobera des attributs tels que l'identifiant de l'utilisateur, les données démographiques, les préférences, l'historique d'achat, etc. Elle permettra d'analyser le comportement des utilisateurs, d'identifier les segments de clientèle et de personnaliser les offres en fonction des caractéristiques des utilisateurs.
- La dimension "**Purchase**", englobant toutes les informations liées aux ventes. Cette dimension comprendra des attributs comme l'identifiant de la transaction, la date et l'heure de l'achat, les produits achetés, les quantités, les prix, les remises, les modes de paiement, etc. Elle permettra d'analyser les tendances d'achat, les performances des produits, les revenus générés, ainsi que de mesurer l'efficacité des stratégies de vente.

En reliant la table de faits "**FactPurchase**" à ces deux dimensions clés, nous pourrons effectuer des analyses croisées entre les différents aspects des utilisateurs et des ventes. Cela nous offrira une vision globale et approfondie de notre activité, nous permettant de prendre des décisions éclairées, d'identifier des opportunités de croissance et d'améliorer l'expérience client.

- **Modélisation des data marts**

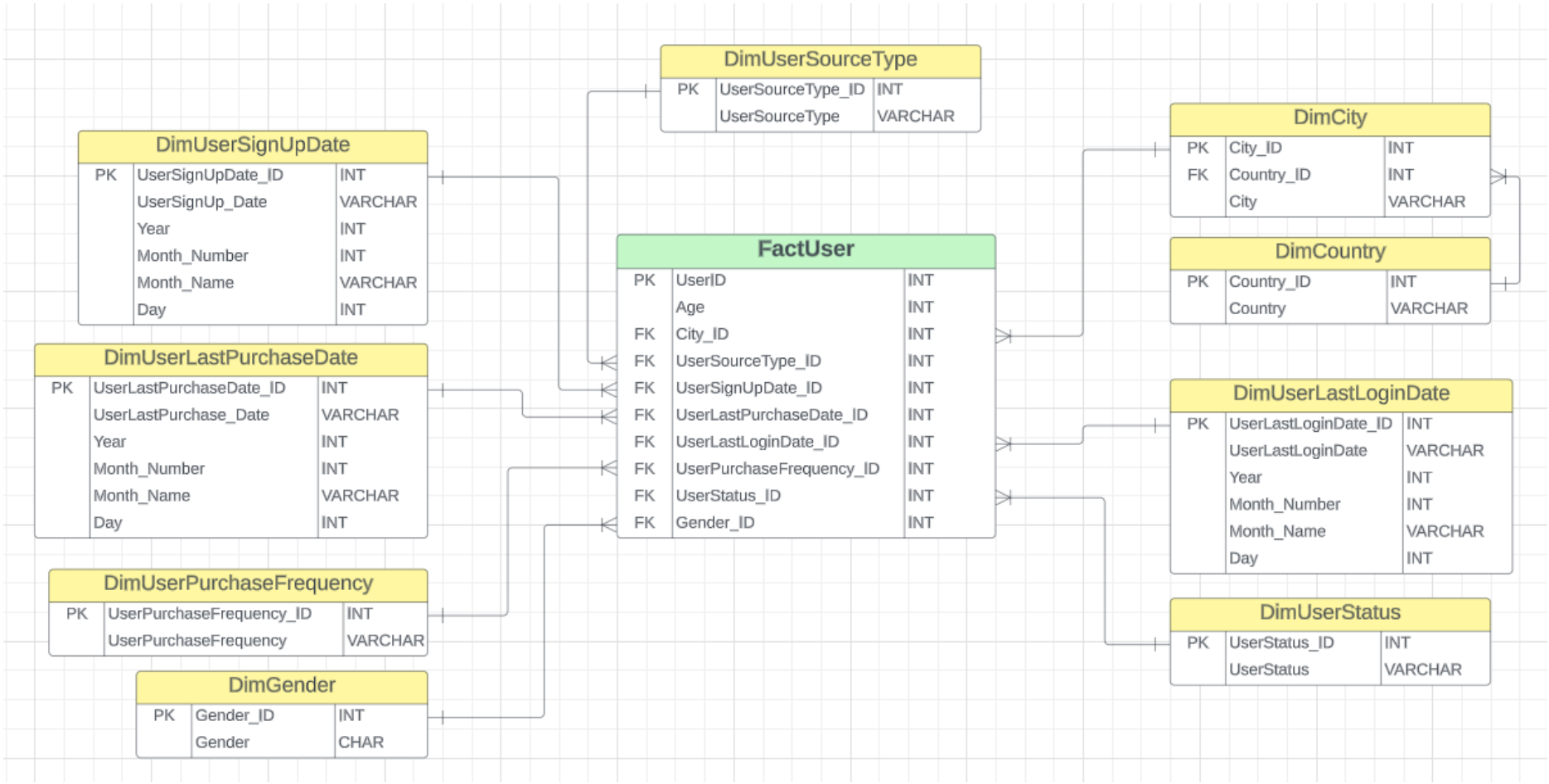
1. Data mart **”Purchase”**



Pour une vision plus précise des ventes dans notre entrepôt de données, nous avons créé une modélisation spécifique à la dimension "**Purchase**". Cette approche permet d’avoir des analyses approfondies des ventes mensuelles, trimestrielles et

annuelles, des produits et catégories les plus vendus, ainsi que des tendances de vente dans le temps. En se concentrant sur cette dimension clé, nous examinons en détail les performances par période, identifions les pics d'activité, les saisons favorables et les variations saisonnières.

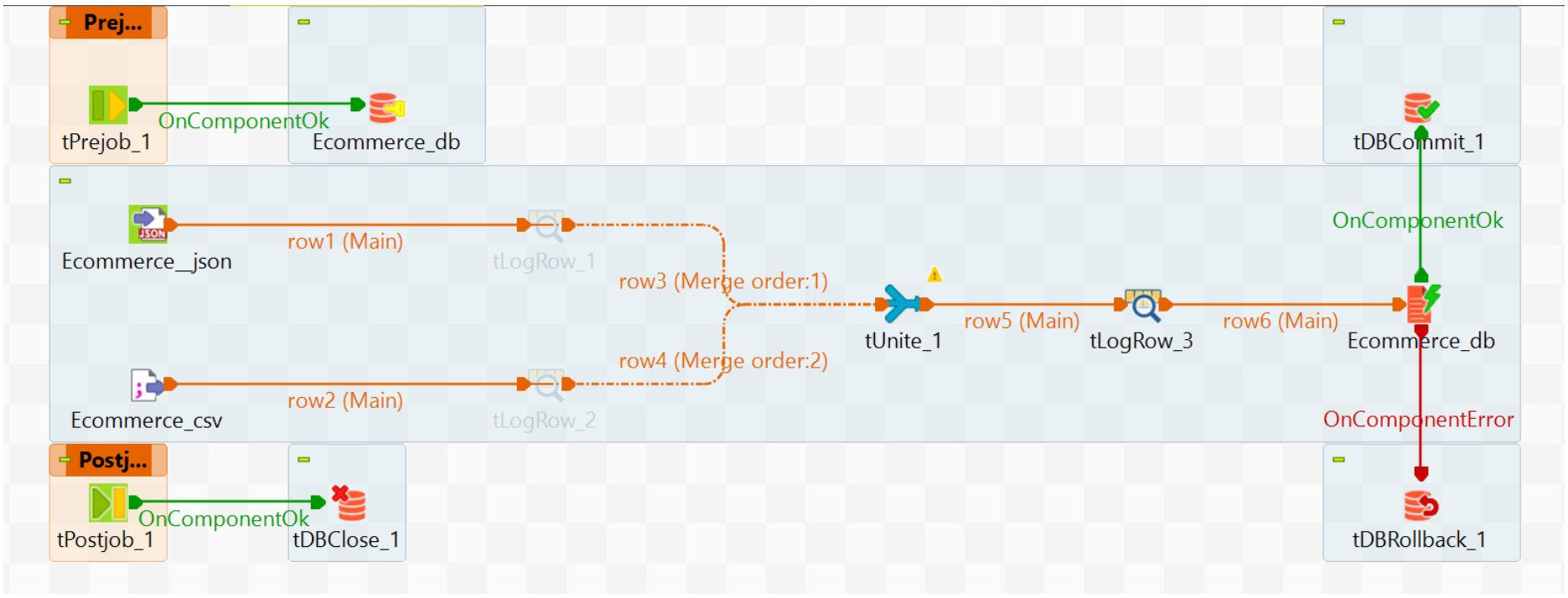
2. Data mart “**User**”



Cette modélisation met en avant la dimension "**User**" pour des analyses approfondies de la démographie des utilisateurs, des tendances d'inscription, de la rétention et du taux de désabonnement. En examinant les caractéristiques démographiques telles que l'âge, le sexe et la localisation, nous comprenons mieux notre base d'utilisateurs. L'analyse des tendances d'inscription nous permet d'identifier les périodes et canaux d'acquisition les plus efficaces. En étudiant la rétention, nous comprenons combien de temps les utilisateurs restent actifs et les facteurs influençant leur fidélité. Enfin, en examinant le taux de désabonnement, nous identifions les raisons et prenons des mesures pour réduire les désabonnements.

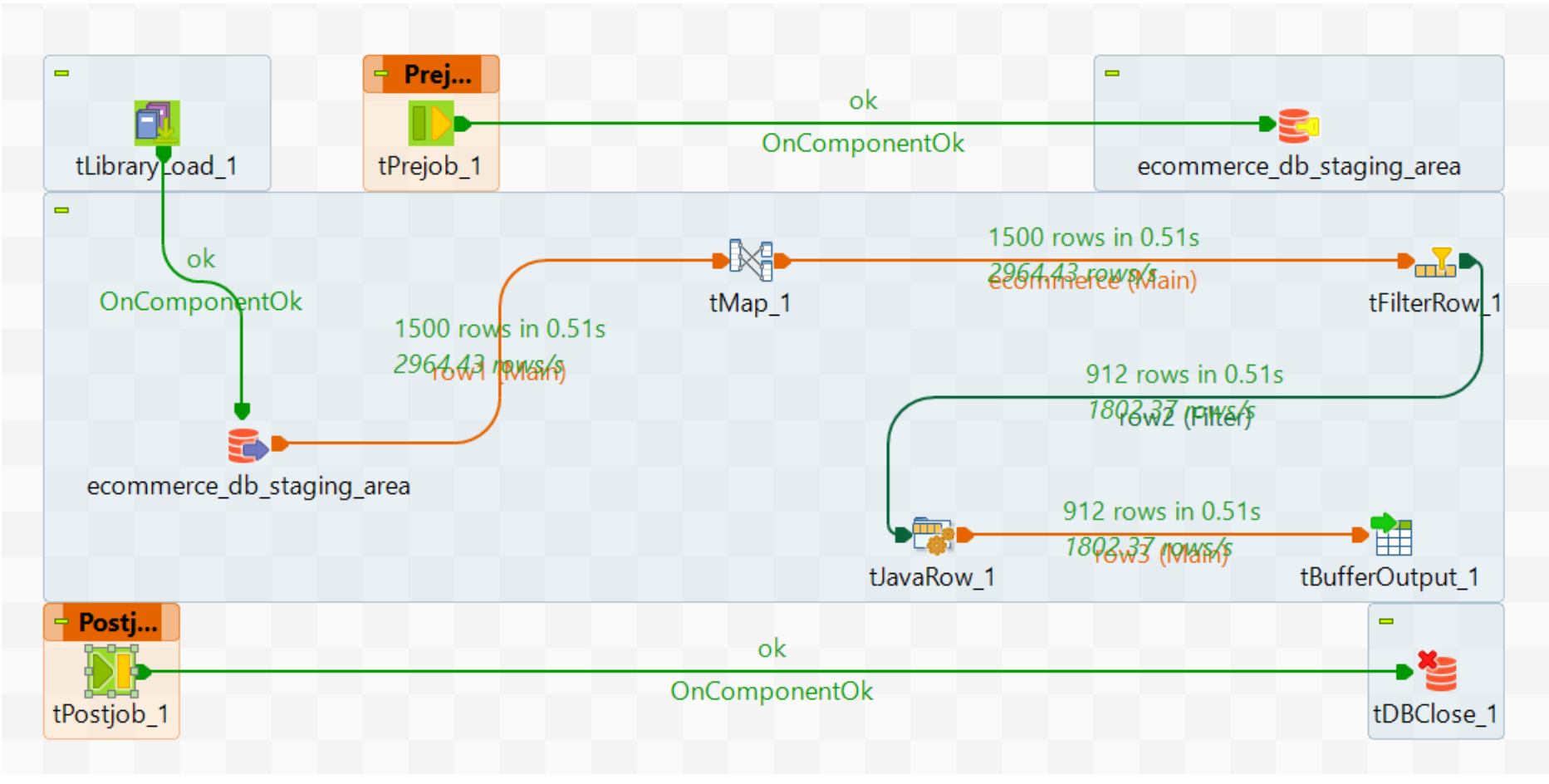
- **Le processus ETL**

1. **Extracttion des donnees |Staging Area|**



À partir de cette capture d'écran, nous avons utilisé l'outil "**tUnite**" pour fusionner les deux fichiers que nous avions précédemment séparés à l'aide d'un script Python. Après avoir concaténé les deux fichiers, nous les avons insérés dans notre base de données "**db_ecom_staging_area**" afin d'effectuer une transformation et de les préparer pour le chargement dans notre entrepôt de données.

2. Transformation des données



Dans le cadre de cette mission, nous avons mis en place un processus de transformation avec Talend, englobant des étapes cruciales telles que le chiffrement, la gestion des erreurs, et bien plus encore. Pour commencer, notre processus démarre en extrayant les données depuis "**Staging area**". Ensuite, nous utilisons le composant “**tMap**” pour traiter les valeurs manquantes et rectifier les erreurs typographiques.

Les regles appliquer en respectant les normes du **RGPD** :

1. Remplacement des valeurs négatives par leur valeur absolue dans toutes les colonnes numériques telles que "Age", "Quantity", "Price", "Rating".
2. Substitution des valeurs nulles dans la colonne "**Product_Name**" par la valeur de la colonne "**SubCategory**".
3. Suppression des lignes contenant les informations d'utilisateurs âgés de 6 à 12 ans, et remplacement des valeurs de la colonne "**Age**" inférieures à 6 ou supérieures à 100 ans par l’esperance d’age “**87 ans”**.
4. Pour garantir la sécurité des données sensibles des utilisateurs, nous avons choisi de chiffrer les informations présentes dans les colonnes “**CreditCard**” et “**FullAddress**” en utilisant la bibliothèque "**org.jasypt.encryption.pbe.StandardPBEStringEncryptor**", cette derniere nous l’avons importée dans le composant "**tJavaRow**".

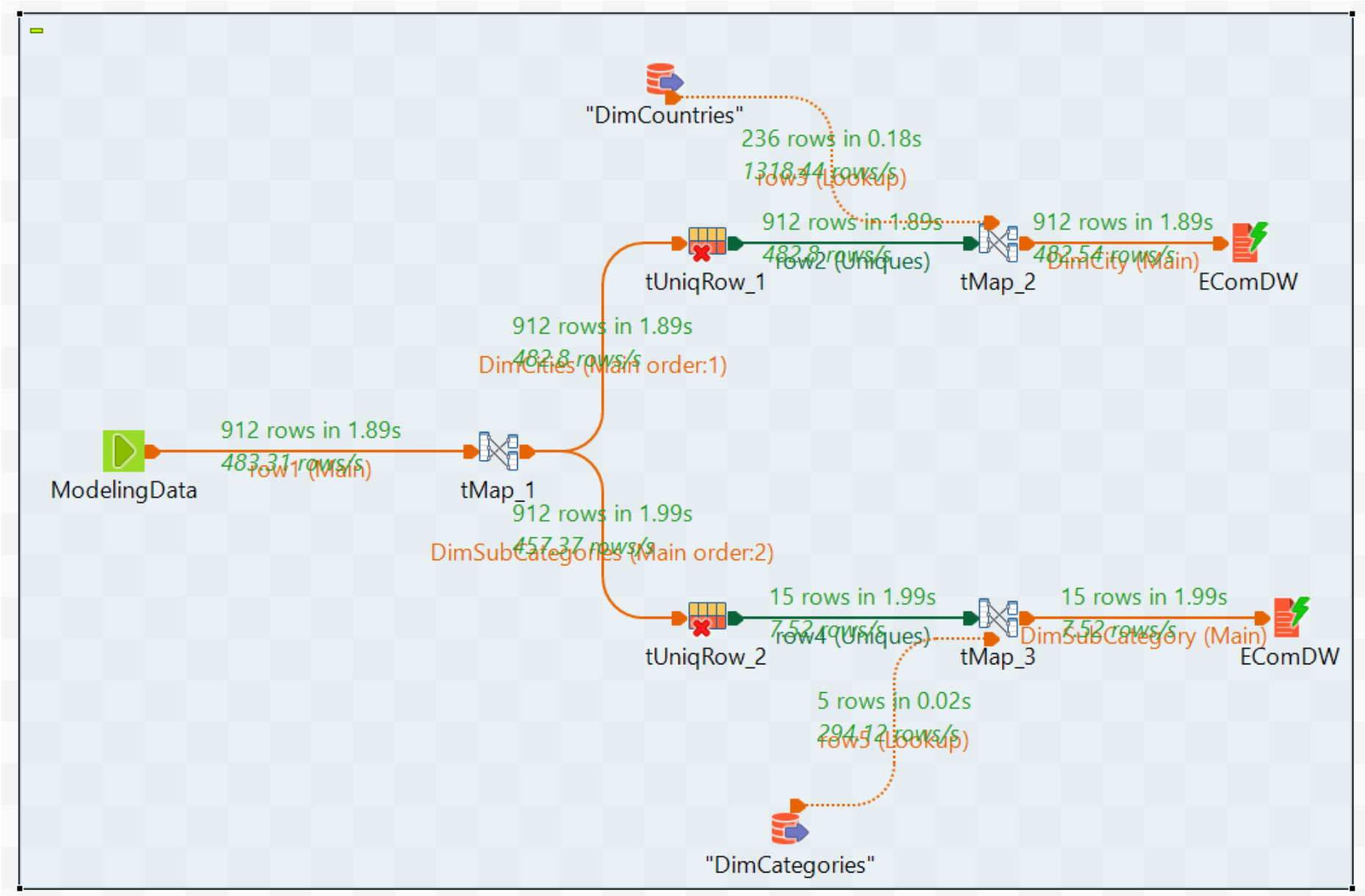
3. Chargement des donnees.

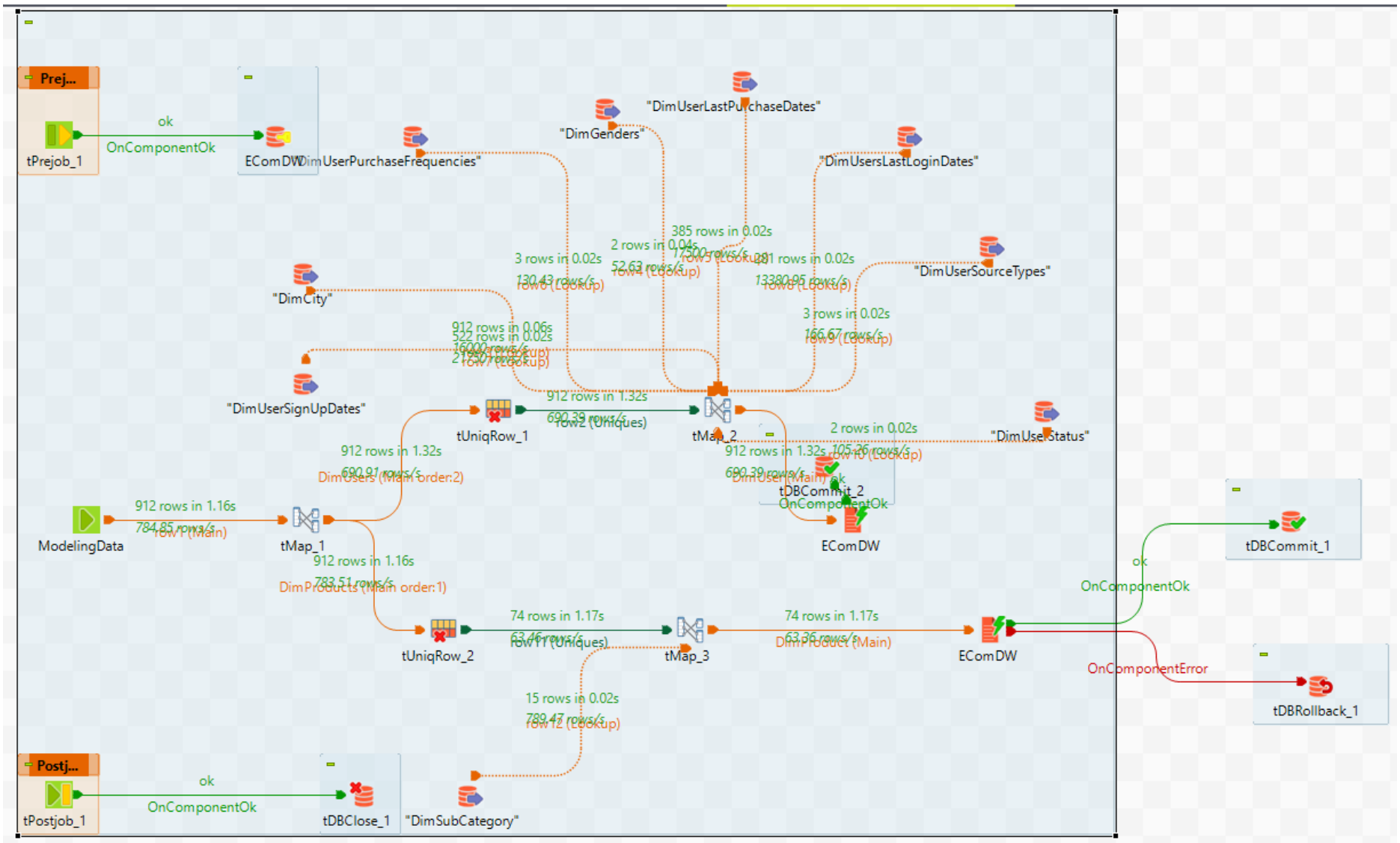
1. chargement des dimensions extérieures.

Dans le but d'améliorer l'efficacité du processus de chargement de notre entrepôt de données, nous avons pris la décision de fragmenter nos données on commençons par identifier les dimensions extérieures qui ne possèdent pas de clés étrangères, telles que "**Dim_Gender**", "**Dim_Category**", "**Dim_Country**", etc.

Cette approche nous permet de mieux organiser notre entrepôt de données et de simplifier les opérations ultérieures de jointure et d'agrégation. En divisant les données en dimensions distinctes, nous obtenons une meilleure granularité et une plus grande flexibilité pour l'analyse ultérieure.

2. **Chargement des dimensions intérieures.**





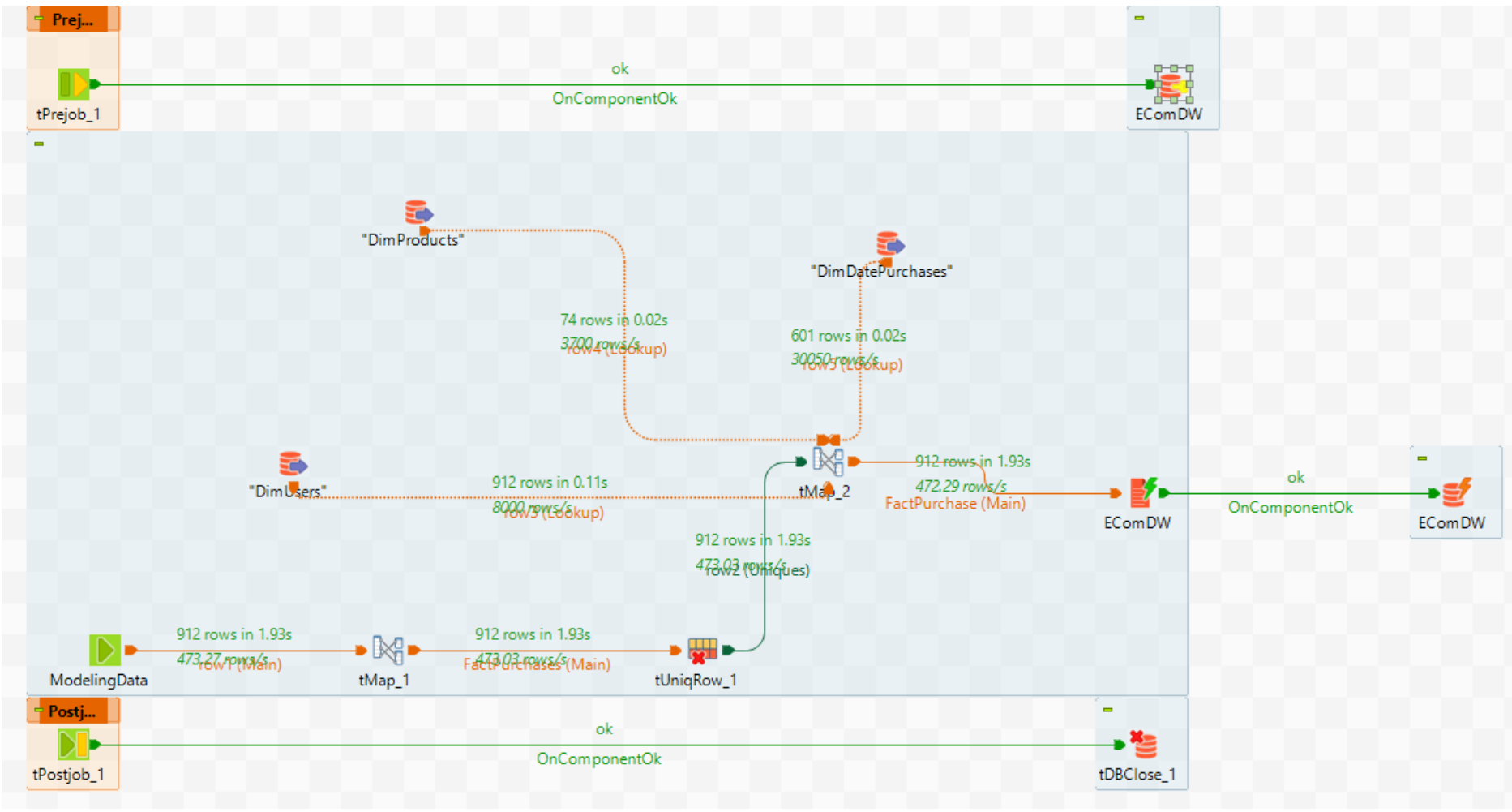
Dans ces deux sous jobs nous avons suivre la meme strategie de chargements des premiers dimensions mais on ajoutons les joitures pour appliquer les relations entre les tables.

Dans ces deux sous-jobs, nous avons suivi la même stratégie de chargement pour les premières dimensions. Cependant, nous avons ajouté des étapes de jointure afin d'appliquer les relations entre les tables.

L'ajout des jointures nous permet d'établir les liens et les relations logiques entre les différentes dimensions. Par exemple, lors du chargement de la dimension "**Dim_Products**", nous avons utilisé une jointure avec la dimension "**Dim_SubCategory**" afin d'associer chaque produit à sa sous-catégorie correspondante. De même, pour la dimension "**Dim_Users**", nous avons effectué une jointure avec la dimension "**Dim_UserSignUpDates**" pour attribuer une date à chaque utilisateur enregistrée.

En appliquant ces jointures, nous créons une structure de données cohérente et complète, qui reflète les relations entre les différentes dimensions. Cela facilite ensuite les requêtes et les analyses ultérieures, car nous pouvons exploiter ces relations pour obtenir des informations plus précises et détaillées.

3. **Chargement de la table de fait “FactPurchase”**



À la fin de notre processus de chargement de données, nous parvenons à créer notre table de fait appelée "**FactPurchase**", qui represente la table centrale de notre entrepot de données.

La table de fait "FactPurchase" constitue le cœur de notre entrepôt de données, car elle contient les mesures et les indicateurs clés liés à nos ventes. En rassemblant les dimensions pertinentes telles que "**Dim_Products**", "**Dim_Users**", "**Dim_DatePurchase**", nous sommes en mesure de fournir une vue globale et intégrée de nos données de ventes.

La création de cette table de fait complète notre processus de chargement de données, en consolidant toutes les dimensions et les mesures clés dans une structure unifiée. Cela nous permet d'effectuer des requêtes et des analyses complexes pour obtenir des informations utiles et des insights précieux sur **Power BI**.

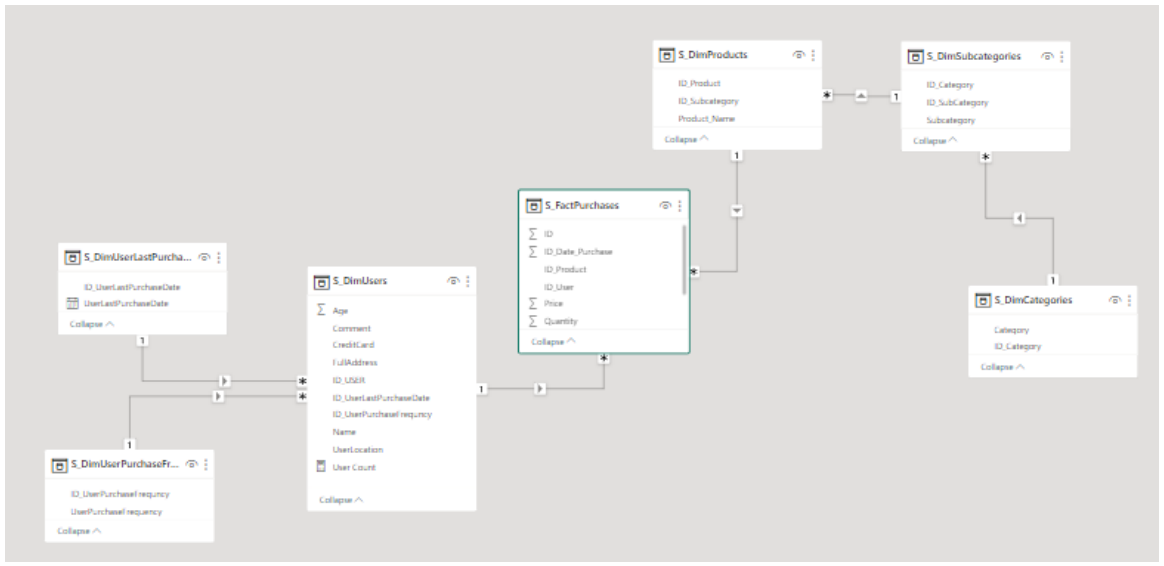
ID	ID_User	Quantity	Rating	ID_Product	ID_Date_Purchase	Price
1	553	1	1	9	427	79
9	333	5	1	41	288	793
11	415	2	1	21	351	580
20	827	10	1	63	558	107
28	545	10	1	42	346	85
42	670	2	1	37	493	384
43	732	1	1	51	60	17
44	783	1	1	3	546	547
47	184	2	1	5	172	62
52	618	5	1	30	76	264
63	800	1	1	40	550	52
66	541	1	1	52	425	207
78	222	4	1	46	206	581

Table de fait “**FactPurchase**”

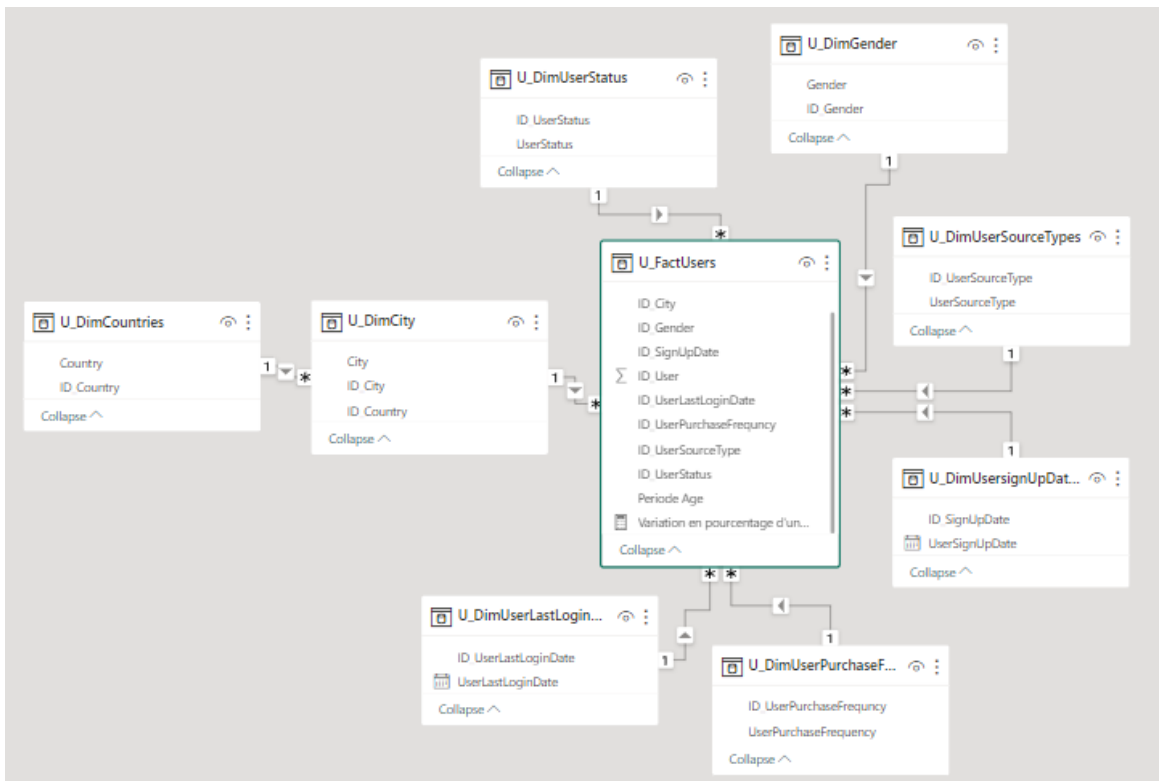
- **Creation des data marts.**

Pour les deux data marts que nous allons utiliser sur Power BI, nous avons décidé de créer une vue pour chaque table afin de bénéficier d'une plus grande flexibilité en matière de sélection des données nécessaires pour de futures analyses.

En créant des vues pour chaque table, nous sommes en mesure de définir des requêtes spécifiques qui extraient uniquement les données pertinentes pour nos analyses sur Power BI. Cela nous permet d'éviter de charger l'intégralité des données de nos tables dans Power BI, ce qui peut être coûteux en termes de performance et de ressources. Au lieu de cela, nous pouvons sélectionner et filtrer les colonnes et les lignes spécifiques qui sont nécessaires pour répondre aux besoins de notre entreprise.



Model sur Power BI du Data mart “**Ventes**”



Model sur Power BI du Data mart “**Utilisateurs**”