

# **OPTIMISATION DE L'ETL ET MISE EN ŒUVRE D'UN ENTREPÔT DE DONNÉES POUR LA VISUALISATION DES DONNÉES**

---

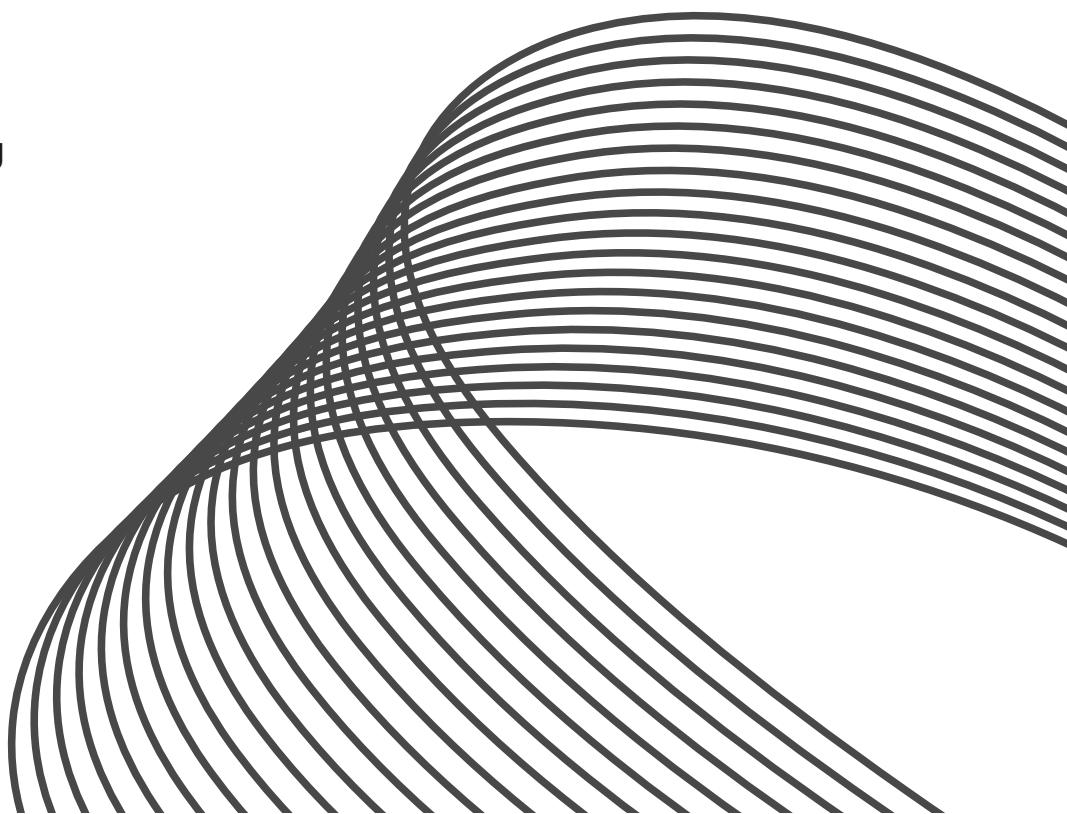
**RÉALISÉ PAR:**

Aymane SABRI

Fatima zahrae BAHADOU

Faissal MOUFLLA

Yasser AITHNINI



# **CONTENU**

**I.Contexte général**

**II.Planification de projet**

**III.Conformité avec le RGPD**

**IV.Répartition des données**

**V.Talend**

**1.Extraction des Données**

**2.Transformation des données**

**3.Chargement dans un data warehouse**

**VI.Visualisations sur Power BI**

**VII.Conclusion**

# I. CONTEXTE GÉNÉRAL

La société X prospère en tant que plateforme e-commerce proéminente, offrant une vaste gamme de produits répartis dans diverses catégories telles que l'électronique, le mobilier, les articles d'épicerie, les vêtements et les livres. À mesure que l'entreprise continue de croître, sa quantité de données a également augmenté de manière significative. Dans le but d'améliorer l'efficacité de ses opérations commerciales et d'obtenir une meilleure compréhension des schémas de comportement de sa clientèle, la société X ambitionne de mettre en place une solution robuste d'entrepôt de données.

En qualité de développeur spécialisé en données, notre rôle principal consistera à concevoir et mettre en œuvre un processus ETL (Extract, Transform, Load) optimisé en utilisant la plateforme Talend. Nous serons chargé de bâtir un entrepôt de données performant, complété par des datamarts ciblés. Les missions centrales qui nous sont confiées englobent notamment :

1. La création d'un flux ETL élaboré qui permettra d'extraire habilement les données pertinentes depuis diverses sources, de les transformer pour qu'elles soient cohérentes et exploitables, puis de les charger de manière efficace dans l'entrepôt de données.

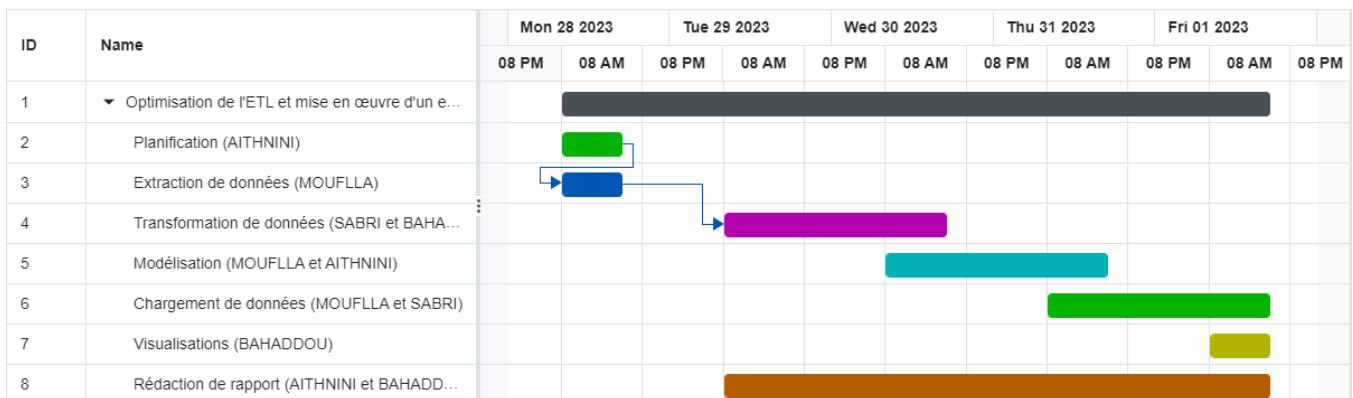
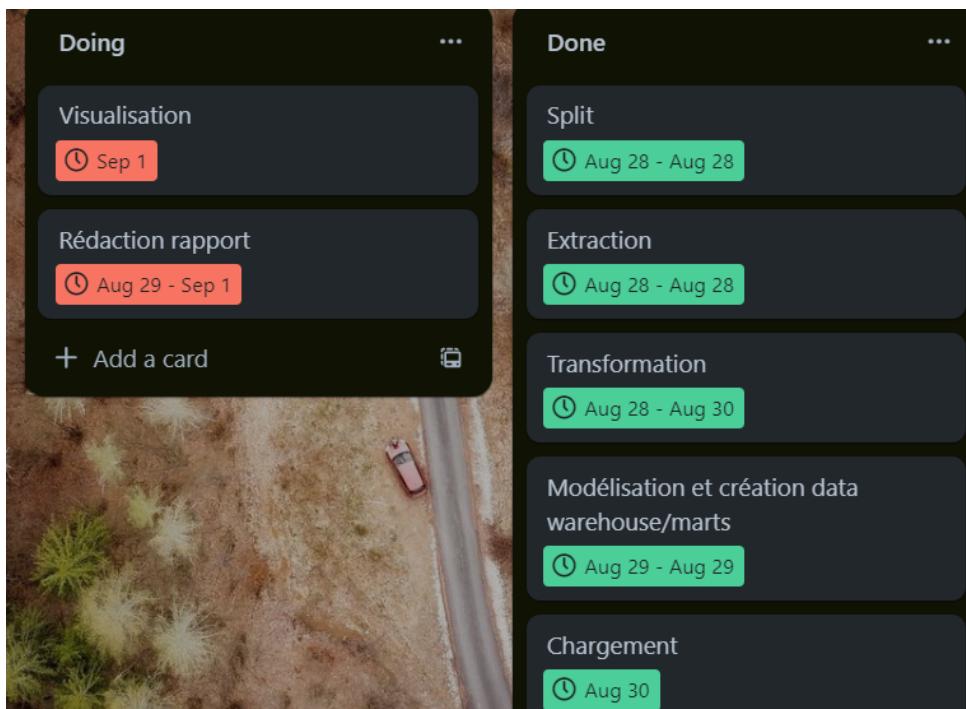
2. L'implémentation d'une structure d'entrepôt de données solide, où les informations seront organisées de manière à favoriser les analyses ultérieures. Cela inclura la modélisation de la base de données pour assurer une récupération des données rapide et optimale.

3. La conception de datamarts spécifiques qui se concentreront sur des domaines d'activité clés. Ces datamarts offriront une vue spécialisée des données pour différents services au sein de l'entreprise, favorisant ainsi une meilleure prise de décision.

Notre travail sera essentiel pour permettre à la société X de tirer parti de ses données de manière stratégique, en transformant des données brutes en informations exploitables. Ce projet contribuera à améliorer la visibilité sur les performances de l'entreprise, à mieux répondre aux besoins des clients et à prendre des décisions éclairées basées sur des données concrètes.

## ***II. PLANIFICATION***

Pour planifier ce projet, nous avons utilisé la plateforme Trello ainsi qu'on a réalisé un diagramme de GANTT.



## **III . CONFORMITÉ AVEC LE RGPD**

Les données sont désignées comme l'or noir du 21ème siècle. L'exploitation de ces datas a pris une ampleur déterminante aussi bien pour les administrations et organisations publiques que pour les entreprises privées qui ne peuvent plus se permettre de passer à côté de ce supercarburant...Or ce sont les données à caractère personnel qui permettent aux entreprises d'obtenir des détails précis sur les utilisateurs et leur comportement en ligne. C'est pourquoi nos données doivent être préservées par des règles et des lois.

Par conséquent, la protection des données implique bien plus qu'une simple protection contre la collecte et l'utilisation abusive de nos données. Elle implique une protection contre la manipulation, les inégalités et la discrimination.

Une donnée personnelle s'assimile à toute information se rapportant à une personne physique identifiée ou identifiable. Ainsi, une personne peut être identifiée de façon directe par son nom et son prénom, ou de façon indirecte par un identifiant, un numéro, une donnée biométrique, plusieurs éléments spécifiques propres à son identité physique, physiologique, économique ou encore culturelle.

L'identification d'une personne physique peut ainsi être effectuée :

À partir d'une seule donnée, ce qui peut correspondre à un numéro de sécurité sociale par exemple,

À partir du croisement d'un ensemble de données (une personne de sexe féminin vivant à telle adresse, née tel jour, abonnée à tel service de diffusion et militant pour telle cause par exemple).

Après des discussions entre les membres de notre squad, nous avons décidé de crypter les valeurs dans les colonnes suivantes : CreditCard et FullName. Ces colonnes là contiennent des informations sensibles. Nous avons aussi crée des login pour la base données pour garantir la sécurité de ces données même-ci elles sont déjà cryptées.

# ***IV. RÉPARTITION DES DONNÉES***

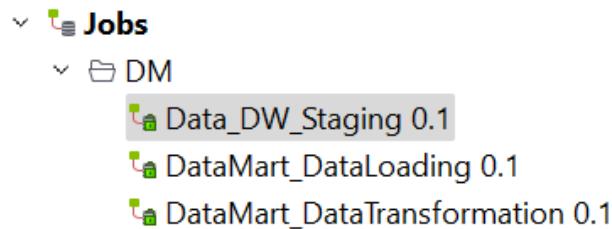
Nous avons utilisé le script de data split que nous avons réalisé lors d'un ancien challenge. Nous avons divisé les données en deux parties : 50% sous format csv et 50% sous format json. On a dû randomiser les lignes avant de diviser les lignes.

ID	Name	Age	Gender	City	Country	CreditCard	FullAddress	UserSignUpD	UserLastPurc	UserStatus
602	Matthew Rollins	-3	M	Lake Sylvia	Kenya	2.7056E+15	247 Victoria Ct	8/31/2022	11/1/2022	Churned
772	Jessica Buchan	67	M	Ericstad	Iraq	2.2655E+15	974 Hopkins St	1/11/2023	3/24/2023	Churned
846	Danielle Bell	150	F	West Ericach	Montserrat	2.2912E+15	99052 Ryan Ln	2/1/2023	7/12/2023	Active
1090	William Bell	-5	F	South Scott	Qatar	2.7204E+15	374 Troy Loc	1/28/2022	1/1/2023	Churned
1365	Dennis Henderson	150	M	South Laura	Russian Fed	2.2371E+15	65243 Kimbe	11/21/2021	10/7/2022	Churned
581	Katelyn Gonzales	2	M	North Jillches	Costa Rica	5.4334E+15	1305 Steve V	1/6/2022	7/1/2022	Churned
1474	Maria Georgiou	39	F	North Timot	Albania	2.7004E+15	47840 Serran	7/27/2023	8/22/2023	Active
788	Katie Craig	-2	M	New Michael	Romania	2.2495E+15	41990 Dustin	1/28/2022	4/14/2023	Churned
1388	Carrie Wright	10	F	Lauramouth	Hungary	2.2537E+15	935 Frank Ca	12/13/2021	4/2/2023	Churned
376	Kim Knight	-4	F	North Teresa	Luxembourg	5.1026E+15	6002 Wilson	7/10/2023	8/18/2023	Active
314	Rachel Brown	150	F	Lake Tyler	Benin	2.2518E+15	USS Reynolds	9/13/2022	11/30/2022	Churned
442	Dwayne Hill	2	F	Mariabury	Comoros	2.2858E+15	35369 Ross H	6/15/2023	8/23/2023	Active
319	Jennifer Simms	40	F	Andersenbu	Kuwait	2.2599E+15	60203 Fowle	8/4/2022	4/20/2023	Churned
1478	Amber Taylor	-4	F	South Kenne	Sweden	5.372E+15	737 Norris Hi	7/30/2023	8/11/2023	Active

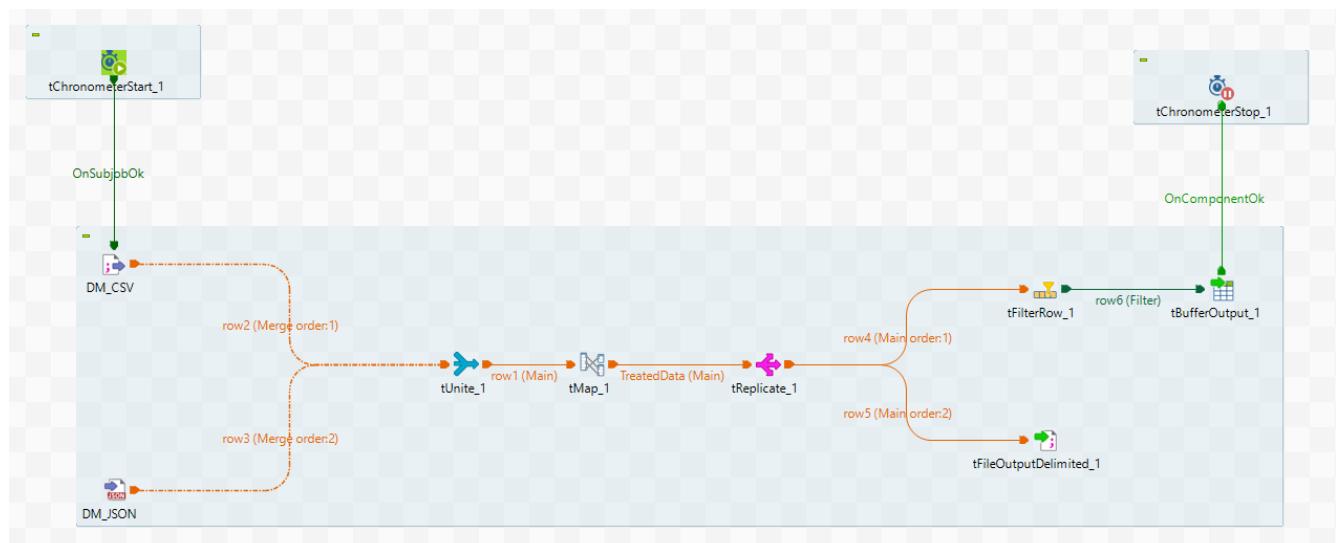
# V. TALEND

## 1. Extraction des Données

Voici l'architecture de nos Jobs :



On a commencé par le chargement des fichiers (fichier JSON, et fichier CSV).



Après l'extraction, nous avons fusionné les données en utilisant le composant tUnité. Pour achever cette fusion, le nom des colonnes dans les fichier csv et json devaient être les mêmes.

## 2. Transformation des données

Ensute on a commencé la transformation dans laquelle on a :

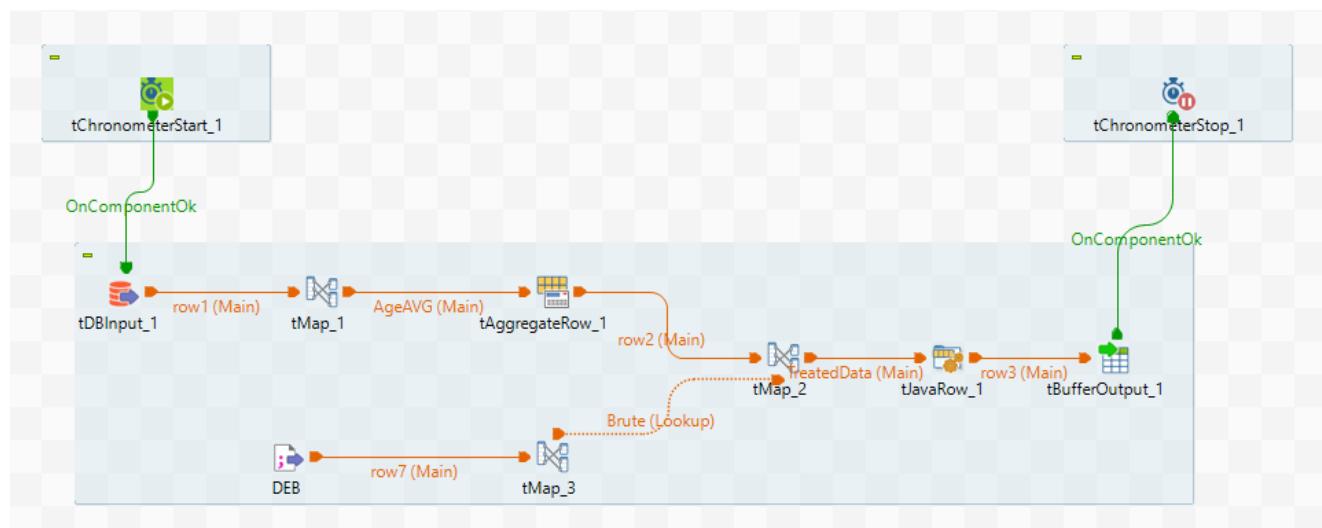
Appliqué la valeur absolue sur les colonnes qui contiennent des valeurs négatives.

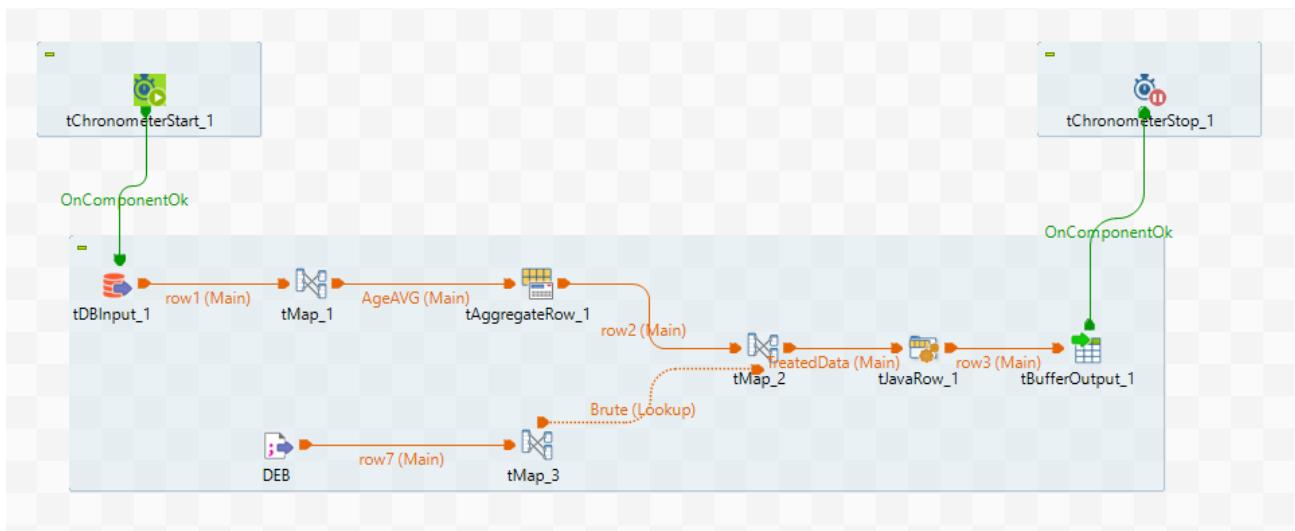
Toutes les colonnes de type string on été mise en majuscule pour faciliter la visibilité.

Pour les valeurs d'âge qui sont supérieures à 100 ou inférieures à 13, nous les avons remplacées par 0.

Dans la colonne ProductName on a remplacé les valeurs NAN et les valeurs vides par les valeurs de Subcategory.

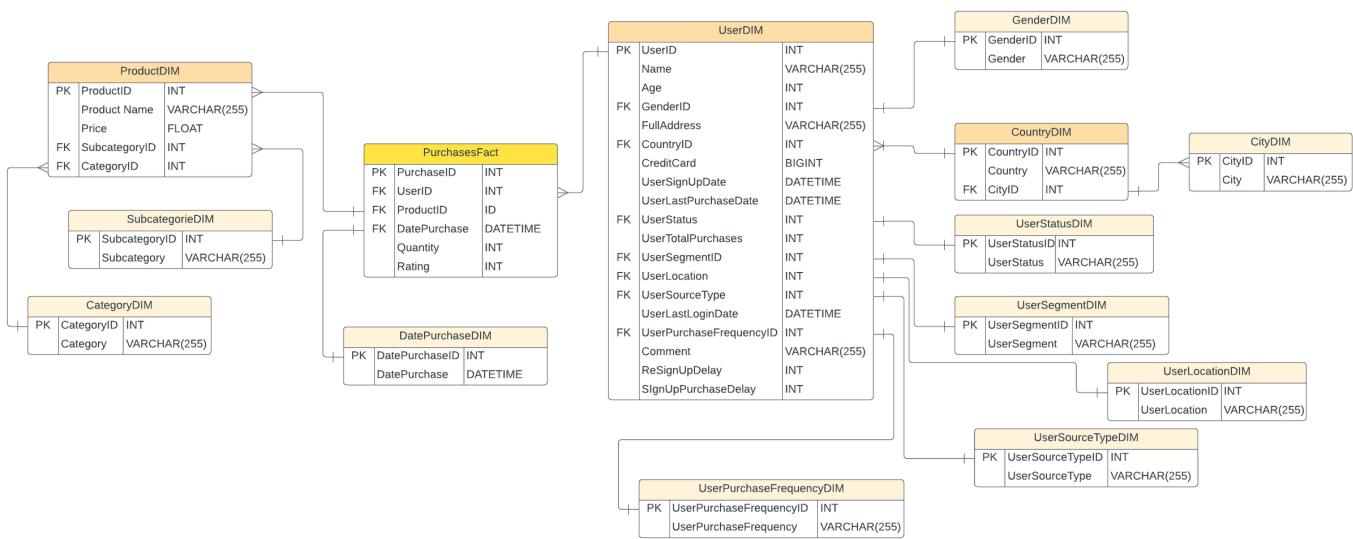
Dans la colonne Quantity on a remplacer les 0 par la moyenne.





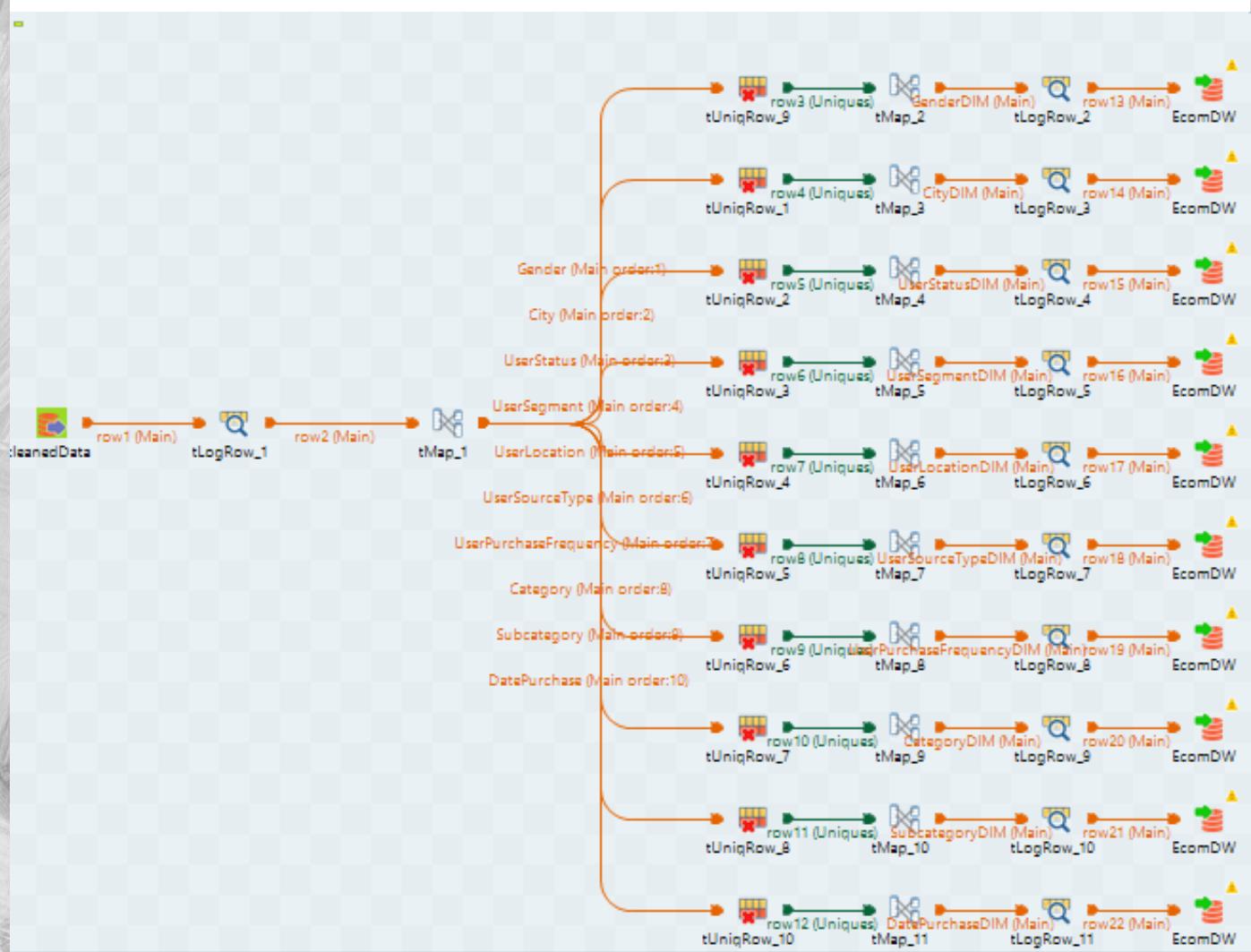
### 3. Chargement dans un data warehouse

Avant d'importer les données dans le data warehouse la première étape consiste à concevoir la modélisation de la base de données. Cela comprend la création de classes, leurs attributs et les relations entre elles. Voici la modélisation conceptuelle de la base de données :

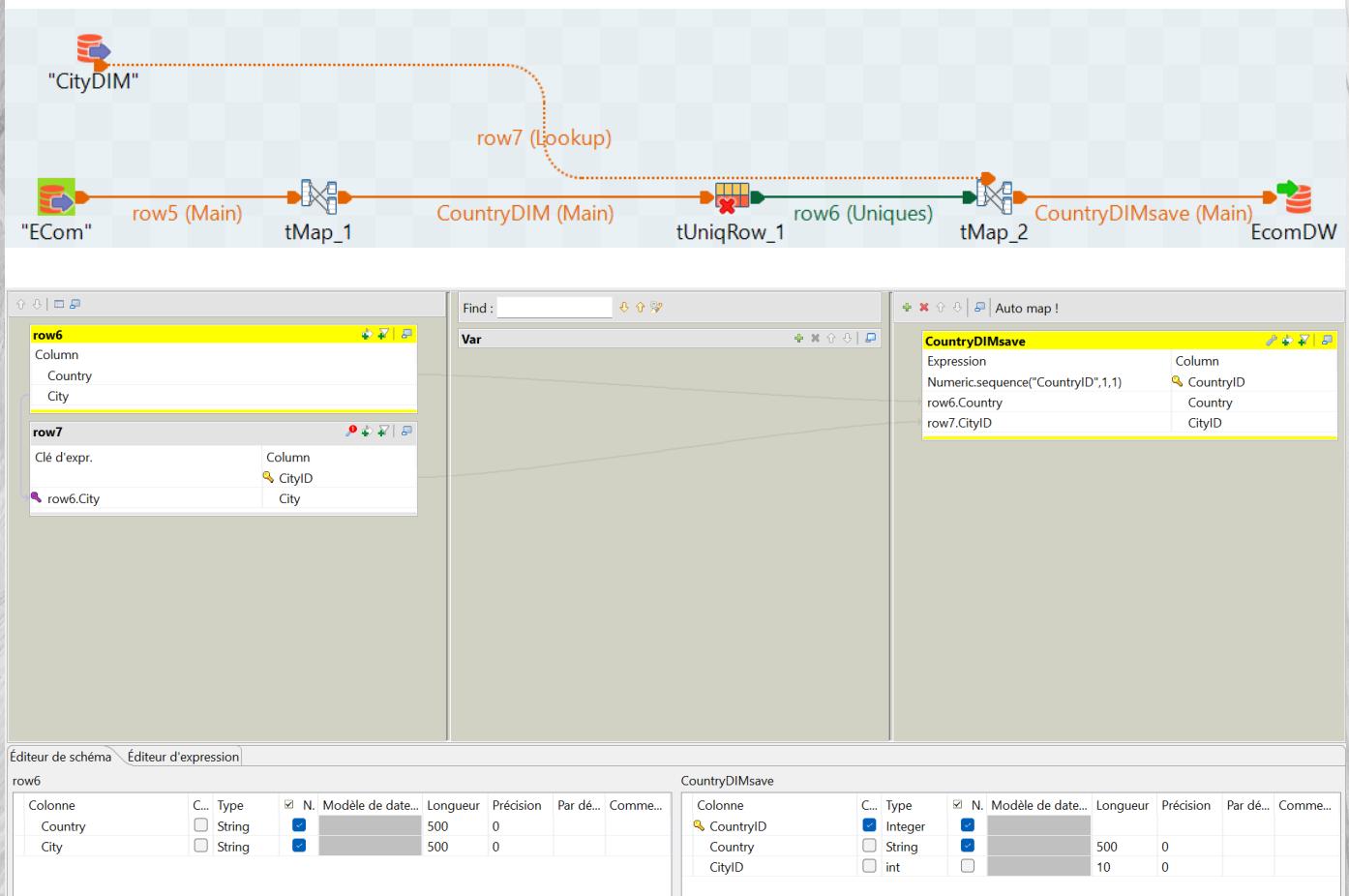


# Création des tables de dimensions

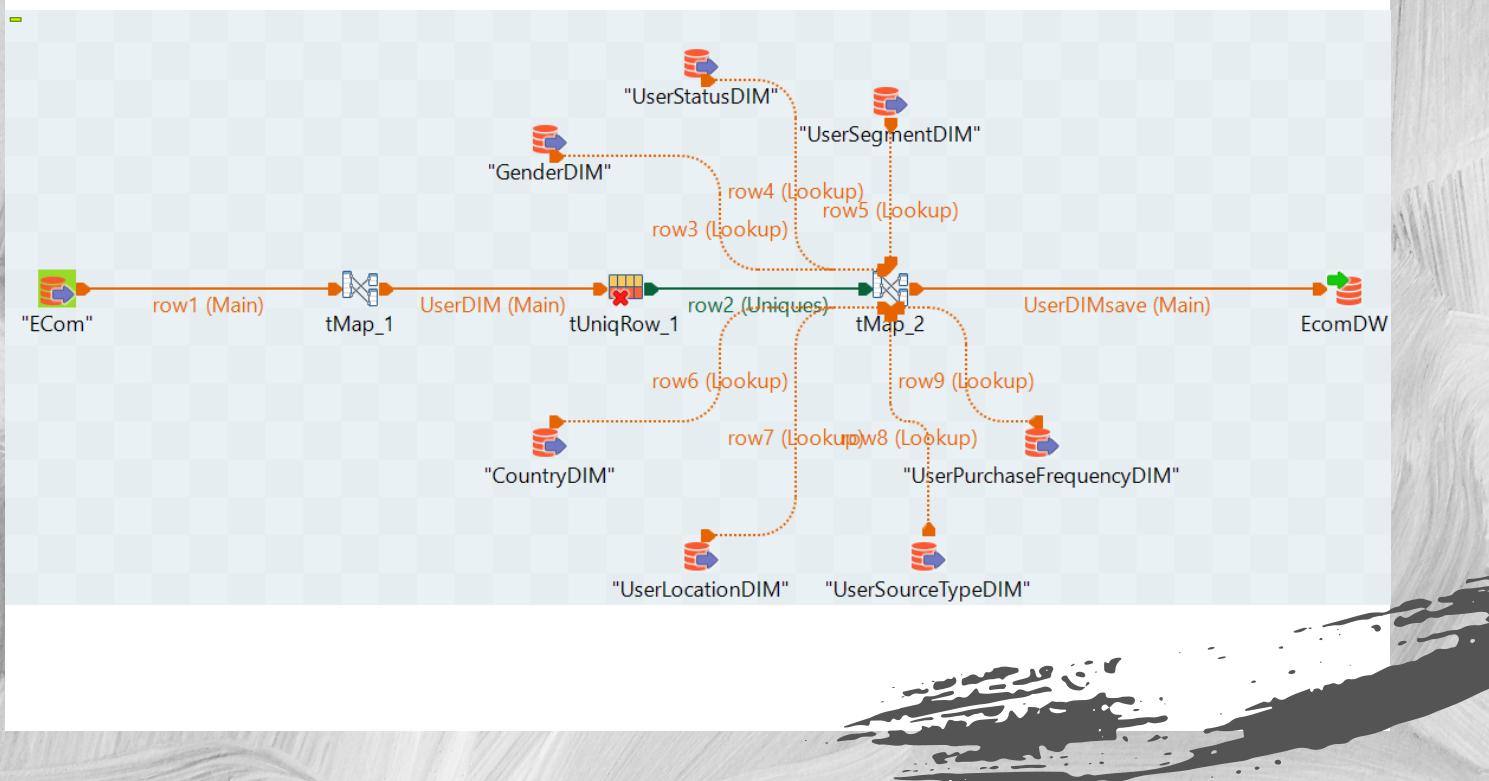
Premièrement on a divisé notre données qui sont déjà nettoyées en des tables, choisir seulement les uniques et générerons les IDS pour chaque table enfin on a les stockées dans le data warehouse.



## DIMCountry :



## DIMUser :



Screenshot of the Talend Data Integration interface showing the schema editor and expression editor.

**Schema Editor:**

- row3:** Clé d'expr. Column GenderID (Gender)
- row4:** Clé d'expr. Column UserStatusID (UserStatus)
- row5:** Clé d'expr. Column UserSegmentID (UserSegment)
- row6:** Clé d'expr. Column CountryID (Country) CityID

**Var:** Find : Auto map !

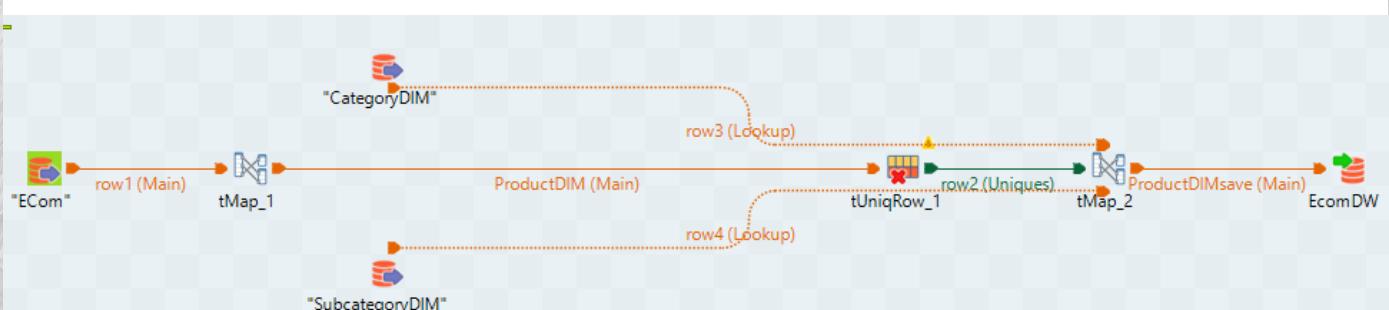
UserDIMsave	Column
Expression	Column
Numeric.sequence("UserID",1,1)	UserID
row2.Name	Name
row2.Age	Age
row3.GenderID	GenderID
row2.FullAddress	FullAddress
row2.CountryID	CountryID
row2.CreditCard	CreditCard
row2.UserSignUpDate	UserSignUpDate
row2.UserLastPurchaseDate	UserLastPurchaseDate
row4.UserStatusID	UserStatusID
row2.UserTotalPurchases	UserTotalPurchases
row5.UserSegmentID	UserSegmentID
row7.UserLocationID	UserLocationID
row8.UserSourceTypeID	UserSourceTypeID
row2.UserLastLoginDate	UserLastLoginDate
row9.UserPurchaseFrequencyID	UserPurchaseFrequencyID
row2.Comment_1	Comment_1
row2.ReSignUpDelay_1	ReSignUpDelay_1

**Table Definitions:**

Colonne	C... Type	N. Modèle de dat...	Longueur	Précision	Par d...	Comme...
Name	String	✓ N.	500	0		
Age	Integer	✓	10	0		
FullAddress	String	✓	500	0		
CreditCard	String	✓	500	0		

Colonne	C... Type	N. Modèle de dat...	Longueur	Précision	Par d...	Comme...
UserID	Integer	✓	500	0		
Name	String	✓	10	0		
Age	Integer	✓	10	0		
GenderID	int					

## DIMProduct :



Screenshot of the Talend Data Integration interface showing the schema editor and expression editor.

**Schema Editor:**

- row2:** Column ProductName, Category, Subcategory, Price, Rating
- row3:** Clé d'expr. Column CategoryID (Category)
- row4:** Clé d'expr. Column SubcategoryID (Subcategory)

**Var:** Find : Auto map !

ProductDIMsave	Column
Expression	Column
Numeric.sequence("ProductID",1,1)	ProductID
row2.ProductName	ProductName
row2.Price	Price
row4.SubcategoryID	SubcategoryID
row3.CategoryID	CategoryID

**Table Definitions:**

Colonne	C... Type	N. Modèle de dat...	Longueur	Précision	Par d...	Comme...
ProductName	String	✓ N.	500	0		
Category	String	✓	500	0		
Subcategory	String	✓	500	0		
Price	Integer	✓	10	0		

Colonne	C... Type	N. Modèle de dat...	Longueur	Précision	Par d...	Comme...
ProductID	Integer	✓	500	0		
ProductName	String	✓	10	0		
Price	Integer	✓	10	0		
SubcategoryID	int					

# FactPurchases :

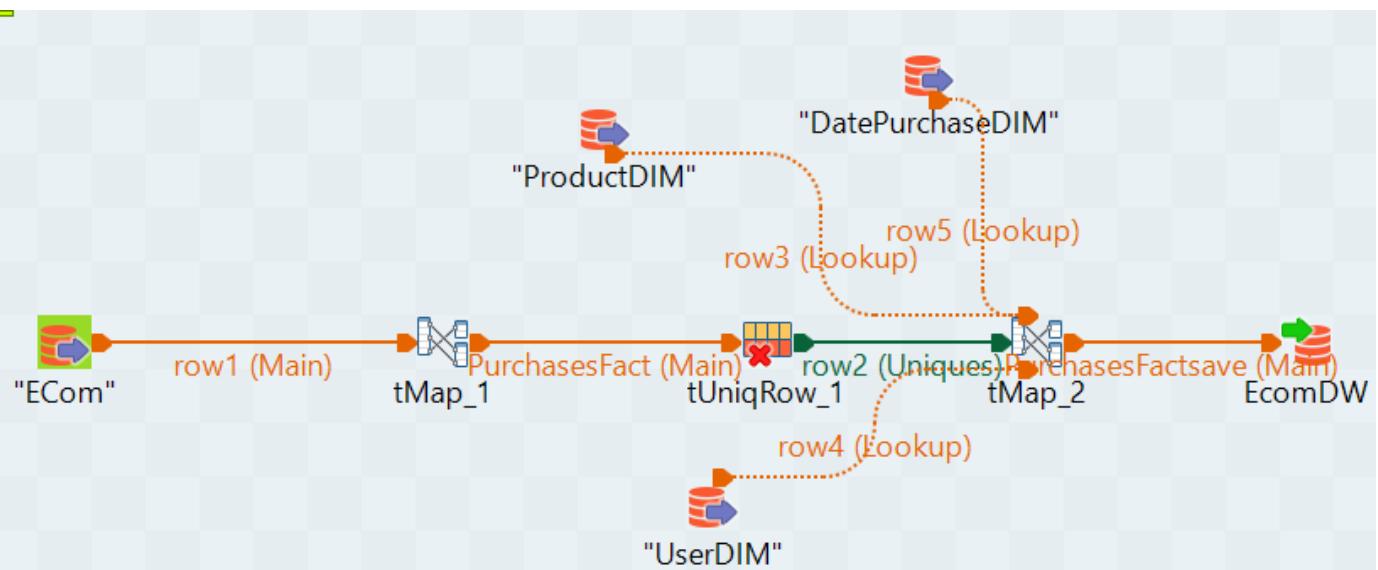


Diagram showing the schema editor and expression editor for the FactPurchases transformation.

**Schema Editor:**

- row2:** Columns: Name, ProductName, DatePurchase, Quantity, Rating.
- row3:** Clé d'expr.: row2.ProductName; Column: ProductID, ProductName, Price, SubcategoryID, CategoryID.
- row4:** Clé d'expr.: row2.Name; Column: UserID, Name.

**Expression Editor:**

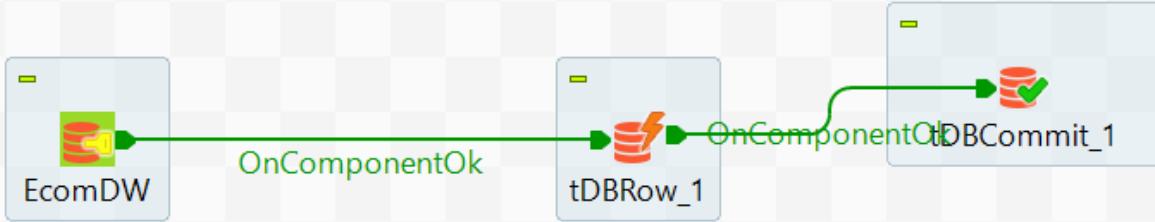
Expression	Column
Numeric.sequence("PurchaseID",1,1)	PurchaseID
row4.UserID	UserID
row3.ProductID	ProductID
row5.DatePurchaseID	DatePurchaseID
row2.Quantity	Quantity
row2.Rating	Rating

**Schema Comparison:**

Éditeur de schéma	Éditeur d'expression
row2	PurchasesFactsave
Colonne	Colonne
Name	PurchaseID
ProductName	UserID
DatePurchase	ProductID
Quantity	DatePurchaseID

Éditeur de schéma	Éditeur d'expression
row2	PurchasesFactsave
Colonne	Colonne
Name	PurchaseID
ProductName	UserID
DatePurchase	ProductID
Quantity	DatePurchaseID

# Création des clés étrangers



**tDBRow\_1(Microsoft SQL Server)**

**Paramètres simples**

Database Microsoft SQL Server Appliquer  
Utiliser une connexion existante Liste des composants tDBConnection\_1 - EcomDW  
Schéma Built-in Modifier le schéma

Nom de la table  
Type de requête Built-in Guess.Ouvr.

Requête

```
ALTER TABLE CountryDIM
ADD CONSTRAINT fk_CountryDIM_CityDIM
FOREIGN KEY (CityID)
REFERENCES CityDIM([CityID])
ON DELETE CASCADE;

ALTER TABLE UserDIM
ADD CONSTRAINT fk_UserDIM_GenderDIM
FOREIGN KEY (GenderID)
REFERENCES GenderDIM([GenderID])
ON DELETE CASCADE;

ALTER TABLE UserDIM
ADD CONSTRAINT fk_UserDIM_CountryDIM
FOREIGN KEY (CountryID)
REFERENCES CountryDIM([CountryID])
ON DELETE CASCADE;

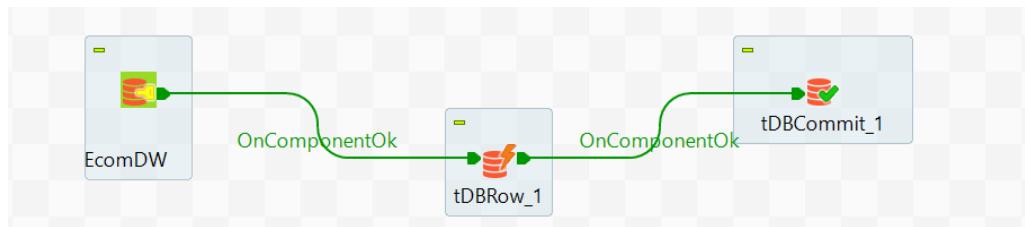
ALTER TABLE UserDIM
ADD CONSTRAINT fk_UserDIM_UserStatusDIM
```

Arrêter en cas d'erreur

This screenshot shows the configuration dialog for the 'tDBRow\_1' component in SSIS. The 'Paramètres simples' tab is selected. It specifies a Microsoft SQL Server database and uses an existing connection named 'tDBConnection\_1 - EcomDW'. The schema is set to 'Built-in'. In the 'Requête' (Query) section, four ALTER TABLE statements are defined to add foreign key constraints for the 'UserDIM', 'CountryDIM', and 'GenderDIM' tables, cascading on delete. A checkbox for stopping on error is present at the bottom.

# Création de Data marts

Nous avons créé 2 data marts : le premier pour les informations sur les utilisateurs et le deuxième est pour les informations sur les ventes. Le choix de créer des data marts à l'aide de vues plutôt que de tables physiques distinctes repose sur plusieurs avantages essentiels. Tout d'abord, l'utilisation de vues permet d'économiser de l'espace de stockage en évitant la duplication des données, ce qui est particulièrement précieux dans un environnement de données en constante évolution. De plus, les vues garantissent une intégration en temps réel, assurant ainsi que les données dans les data marts sont toujours à jour. Cela simplifie également la gestion en réduisant la complexité, tout en offrant un accès facile et sécurisé aux données pour les analystes et les utilisateurs finaux. Les performances sont optimisées grâce à des configurations spécifiques aux vues, et la flexibilité est maintenue pour répondre aux besoins changeants de l'analyse des données. En somme, le recours aux vues pour créer des data marts est une approche efficace, agile et performante dans le domaine de la gestion des données.



### tDBRow\_1(Microsoft SQL Server)

#### Paramètres simples

Paramètres avancés

Paramètres dynamiques

Vue

Documentation

Database Microsoft SQL Server Appliquer

Utiliser une connexion existante Liste des composants tDBConnection\_1 - EcomDW

Schéma Built-in Modifier le schéma

Nom de la table

Arrêter en cas d'erreur

Type de requête Built-in Guess Querr.

Requête

```

CREATE VIEW Utilisateurs AS
SELECT
    Age,
    Gender,
    City,
    Country,
    UserSignUpDate,
    UserLastPurchaseDate,
    UserStatus,
    UserTotalPurchases,
    UserSegment,
    UserSourceType,
    UserLastLoginDate,
    DatePurchaseDIM.DatePurchase,
    UserPurchaseFrequency,
    ReSignUpDelay,
    SignUpPurchaseDelay
FROM UserDIM
INNER JOIN GenderDIM ON UserDIM.GenderID = GenderDIM.GenderID
INNER JOIN CountryDIM ON UserDIM.CountryID = CountryDIM.CountryID
INNER JOIN CityDIM ON CountryDIM.CityID = CityDIM.CityID

```

Arrêter en cas d'erreur

### tDBRow\_1(Microsoft SQL Server)

#### Paramètres simples

Paramètres avancés

Paramètres dynamiques

Vue

Documentation

Database Microsoft SQL Server Appliquer

Utiliser une connexion existante Liste des composants tDBConnection\_1 - EcomDW

Schéma Built-in Modifier le schéma

Nom de la table

Arrêter en cas d'erreur

Type de requête Built-in Guess Querr.

Requête

```

CREATE VIEW Ventes AS
SELECT
    c.City,
    co.Country,
    p.ProductName,
    ccat.Category,
    scat.Subcategory,
    p.Price,
    pf.Quantity,
    dp.DatePurchase,
    pf.Rating
FROM PurchasesFact pf
INNER JOIN CityDIM c ON pf.UserID = c.CityID
INNER JOIN CountryDIM co ON pf.UserID = co.CountryID
INNER JOIN ProductDIM p ON pf.[ProductID] = p.[ProductID]
INNER JOIN CategoryDIM ccat ON p.[CategoryID] = ccat.[CategoryID]
INNER JOIN SubcategoryDIM scat ON p.[SubcategoryID] = scat.[SubcategoryID]
INNER JOIN DatePurchaseDIM dp ON pf.[DatePurchaseID] = dp.[DatePurchaseID];

```

Arrêter en cas d'erreur

Nous avons aussi créé des rôles admin et user. Admin pour les membres de notre squad et user pour les autres utilisateurs.

```
GRANT SELECT , UPDATE , DELETE ON DATABASE::ECommerceSource TO DENGINEERE;
GRANT ALL ON DATABASE::ECommerceSource TO DANALYSTE;
```

A screenshot of the SQL Server Management Studio (SSMS) interface. The title bar shows the connection to 'localhost:ECommerceSource' and the query window title 'SQLQuery\_1 - localhost:ECommerceSource (sa)'. The toolbar includes 'Run', 'Cancel', 'Disconnect', 'Change', 'Database: ECommerceSource', 'Estimated Plan', 'Actual Plan', 'Parse', 'Enable SQLCMD', and 'To Notebook'. The query pane contains the following T-SQL code:

```
1 CREATE LOGIN SP_2_SQ WITH PASSWORD = 'AllahSave.1234/';
2 USE ECommerceSource;
3 CREATE USER SQUADSP2 FOR LOGIN SP_2_SQ ;
```

The 'Messages' pane at the bottom shows the execution results:

```
9:30:47 AM Started executing query at Line 3
Commands completed successfully.
Batch execution time: 00:00:00.013
Total execution time: 00:00:00.013
```

A screenshot of the SQL Server Management Studio (SSMS) interface. The title bar shows the connection to 'localhost:ECommerceSource' and the query window title 'SQLQuery\_1 - localhost:ECommerceSource (sa)'. The toolbar includes 'Run', 'Cancel', 'Disconnect', 'Change', 'Database: ECommerceSource', 'Estimated Plan', 'Actual Plan', 'Parse', 'Enable SQLCMD', and 'To Notebook'. The query pane contains the following T-SQL code:

```
1 USE ECommerceSource;
2
3 CREATE ROLE DENGINEERE;
4
5 ALTER ROLE DENGINEERE ADD MEMBER YASSERDENGINEERE;
6 ALTER ROLE DENGINEERE ADD MEMBER ZAHRADENGINEERE;
7 ALTER ROLE DENGINEERE ADD MEMBER MOUFALLAENGINEERE;
8 ALTER ROLE DENGINEERE ADD MEMBER AYMANEDENGINEERE;
9
10 CREATE ROLE DANALYSTE;
11
12 ALTER ROLE DANALYSTE ADD MEMBER PDANALYSTE;
13
```

The 'Messages' pane at the bottom shows the execution results:

```
11:12:19 AM Started executing query at Line 12
Commands completed successfully.
Batch execution time: 00:00:00.007
Total execution time: 00:00:00.007
```

## Database Role - DENGINEERE (Preview)

### General

Name

Owner  [Browse...](#)

#### Owned Schemas

Select	Schema
<input type="checkbox"/>	db_accessadmin
<input type="checkbox"/>	db_backupoperator
<input type="checkbox"/>	db_datareader
<input type="checkbox"/>	db_datawriter
<input type="checkbox"/>	db_ddladmin
<input type="checkbox"/>	db_denydatareader
<input type="checkbox"/>	db_denydatawriter
<input type="checkbox"/>	db_owner
<input type="checkbox"/>	db_securityadmin
<input type="checkbox"/>	dbo

#### Members

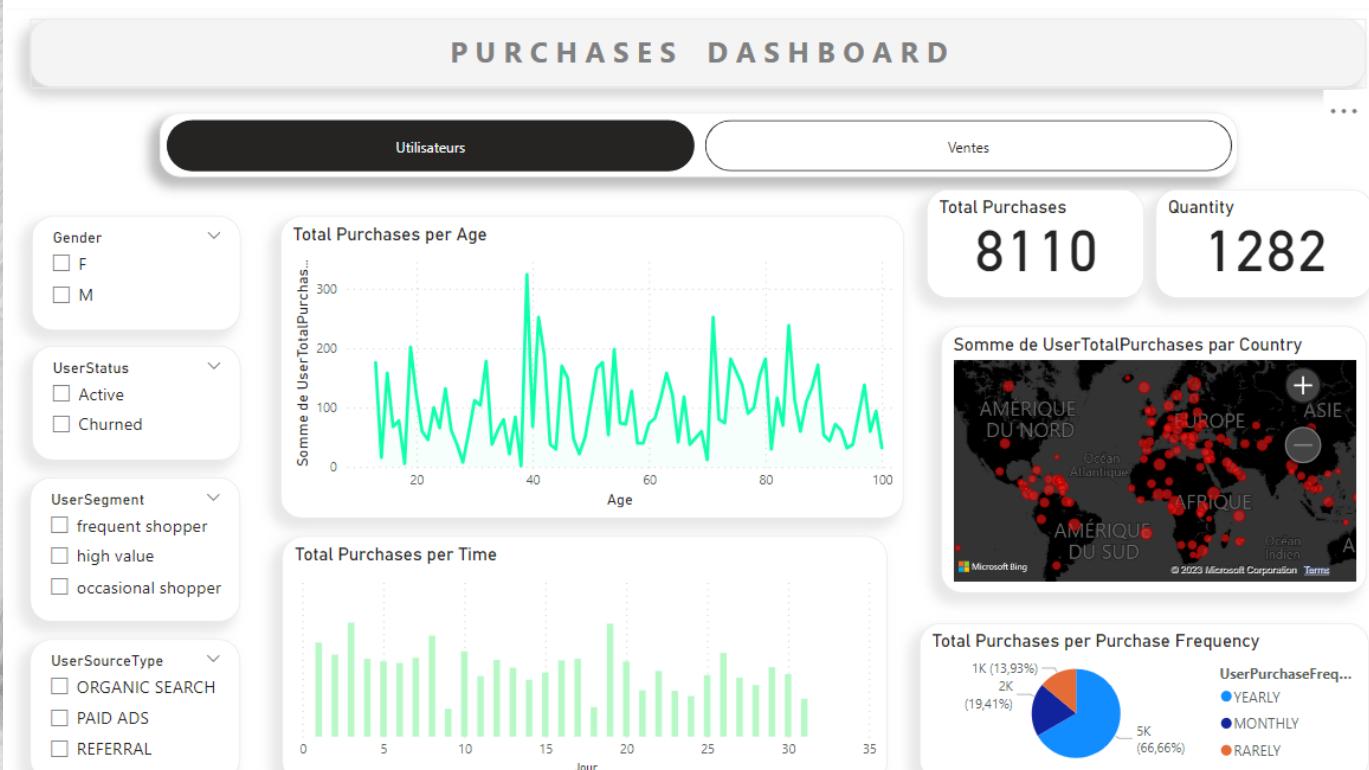
Name
AYMANEDENGINEERE
MOUFALLAENGINEERE
YASSERDENGINEERE
ZAHRADENGINEERE

[Add...](#) [Remove](#)

#### Securables

# VI. VISUALISATIONS SUR POWER BI

Nous nous sommes basées sur les data marts qu'on a créés pour visualiser les données. Voici les visualisations qu'on a réalisé pour le data mart des Utilisateurs.



# **VII. CONCLUSION**

Ce projet visait à mettre en place un entrepôt de données robuste pour la société X, une plateforme e-commerce en pleine croissance. L'objectif principal était de fournir une infrastructure solide pour analyser les ventes, les comportements des utilisateurs et les tendances des produits, afin d'aider les parties prenantes à prendre des décisions éclairées.

Dans le cadre de ce projet, plusieurs étapes clés ont été accomplies. Tout d'abord, une attention particulière a été portée à la conformité au règlement général sur la protection des données (RGPD), avec l'élaboration d'un plan détaillé pour garantir que toutes les données sont traitées en toute conformité.

L'implémentation d'un processus ETL optimisé à l'aide de Talend a été une étape cruciale pour assurer l'efficacité du traitement des données. Ce processus a permis de diviser le jeu de données en formats CSV, JSON, facilitant ainsi la gestion des données tout au long du projet.

La création d'une zone de staging a permis de préparer les données avant leur intégration dans l'entrepôt principal, en nettoyant, en transformant et en validant les données.

La modélisation de l'entrepôt de données a été réalisée avec soin, comprenant les schémas de données, les transformations et les relations nécessaires. Deux data marts distincts ont été créés pour répondre aux besoins spécifiques des parties prenantes.

La gestion de l'accès basé sur les rôles aux data marts a été mise en place pour garantir la sécurité des données, restreignant l'accès aux données sensibles aux utilisateurs autorisés.

Enfin, la visualisation des données résultantes des data marts a été réalisée à l'aide de Power BI, offrant ainsi des insights significatifs aux parties prenantes.

En conclusion, ce projet a réussi à mettre en place un entrepôt de données conforme au RGPD, efficace dans le traitement des données, bien structuré dans sa modélisation, sécurisé dans son accès et informatif dans ses visualisations. Il offre à la société X un outil puissant pour analyser ses opérations et prendre des décisions stratégiques basées sur des données fiables. Ce projet démontre l'importance d'une approche méthodique et rigoureuse dans la gestion des données pour soutenir la croissance et la réussite d'une entreprise.