

# OPTIMISATION DE L'ENTREPÔT DE DONNÉES POUR UN ACTEUR MAJEUR DE L'E-COMMERCE

---

Réalisé par :

Aymane SABRI

Faissal MOUFLLA

Fatima Zahrae BAHADDOU

Yasser AITHNINI

# **CONTENU**

**I.Contexte général et définition des besoins**

**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.OLAP Cube et Test unitaire**

**VII.Conclusion**

# I. CONTEXTE GÉNÉRAL ET DÉFINITION DES BESOINS

XYZ Corp, l'une des principales entreprises de e-commerce en France, fait face à des défis croissants liés à la gestion des données et à l'analyse des ventes. En raison de sa croissance rapide, l'entreprise est confrontée à un volume de données en constante augmentation, ce qui nécessite une révision de ses processus de gestion des données et de son infrastructure d'analyse. Le projet vise à résoudre ces problèmes en mettant en place des solutions modernes pour la collecte, le stockage et l'analyse des données, tout en optimisant la gestion de l'inventaire pour répondre aux besoins croissants de l'entreprise dans le secteur du e-commerce en France.

Pour mieux comprendre ce projet, on a défini les besoins de l'entreprise XYZ :

**1. Gestion des données:** L'entreprise ressent le besoin de gérer efficacement ses données, notamment en ce qui concerne les ventes et l'inventaire. Cela indique une nécessité de collecter, stocker et gérer ces données de manière organisée.

**2. Analyse des ventes:** XYZ Corp souhaite analyser ses ventes de manière plus approfondie. Cela pourrait impliquer l'identification de tendances, la compréhension du comportement des clients, et la prise de décisions éclairées basées sur ces analyses.

**3. Gestion de l'inventaire:** Avec un catalogue de produits en expansion, l'entreprise a besoin d'une solution pour gérer son inventaire de manière efficace. Cela peut inclure le suivi des niveaux de stock, la gestion des réapprovisionnements, et l'optimisation de l'inventaire.

**4. ETL (Extract, Transform, Load):** L'entreprise prévoit de mettre en place un processus ETL optimisé à l'aide de Talend. Cela signifie qu'ils ont besoin d'extraire des données provenant de diverses sources, de les transformer en un format approprié, puis de les charger dans un entrepôt de données.

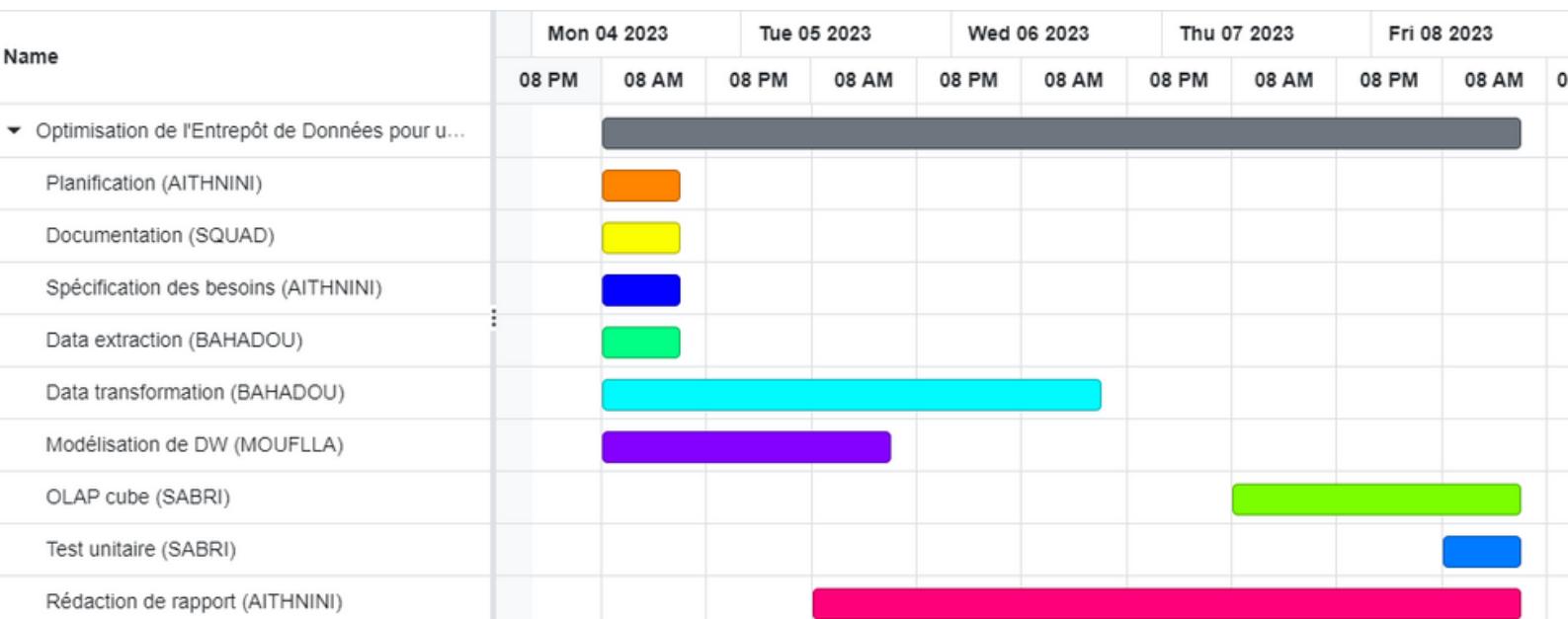
**5. Entrepôt de données:** XYZ Corp a l'intention de mettre en œuvre un entrepôt de données optimisé. Cela servira de référentiel central pour stocker et organiser les données, ce qui facilitera l'analyse et la génération de rapports.

# II. PLANIFICATION DE PROJET

Pour planifier ce projet, on a opté pour la plateforme de gestion de projet sur GitHub ainsi qu'on a créé un diagramme de GANTT.

The screenshot shows a GitHub project board with three columns:

- Todo (2 items):**
  - Draft: Test unitaire
  - Draft: OLAP cube creation
- In Progress (3 items):**
  - Draft: Rédaction de rapport
  - Draft: Data warehouse modelisation
  - Draft: Data loading
- Done (5 items):**
  - Draft: Planification
  - Draft: Documentation
  - Draft: Spécifications de besoins



# 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).

Lors d'une analyse visuelle de notre dataset, nous avons remarqué qu'elle ne contient pas des informations sensibles.

# 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.

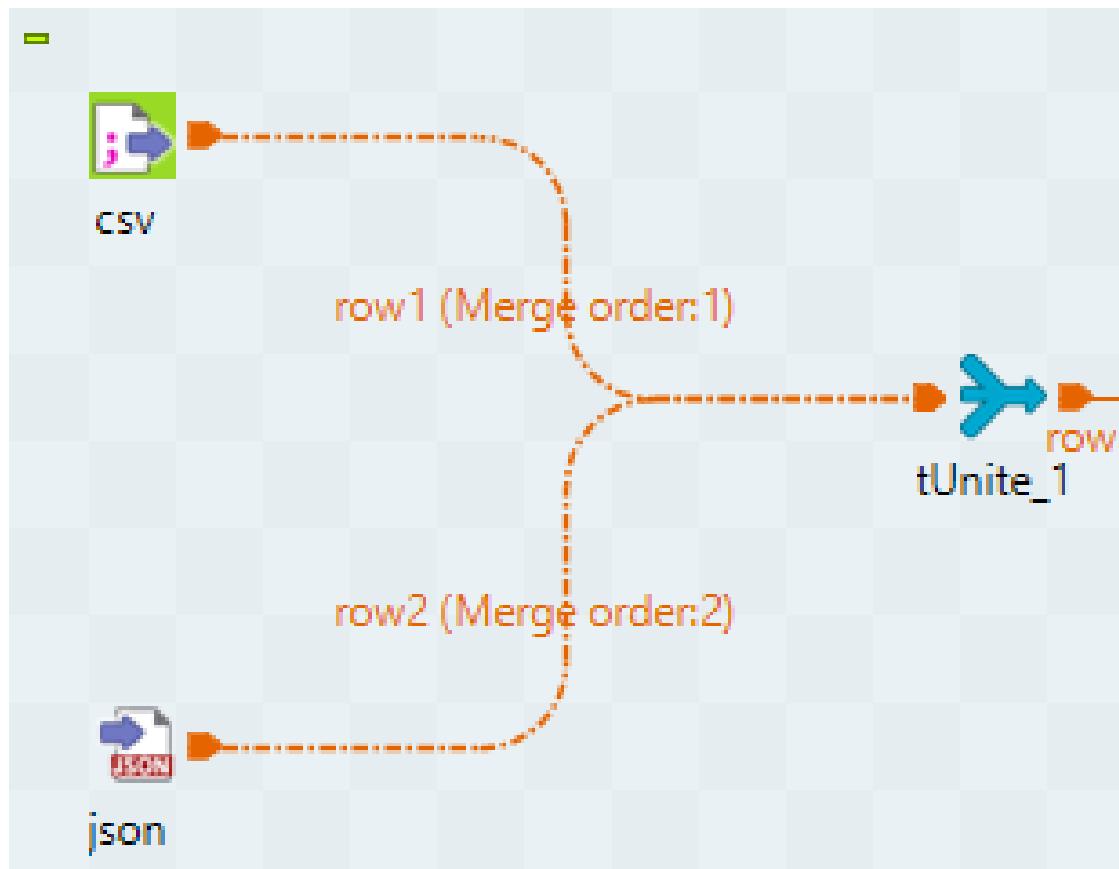
CustomerID	Name	Email	Age	Gender	Location	ProductID	Category	Subcategory	Price	Weight
7659	Megan Myer	donnastephene@yola.net	49	Male	East Anthony	0	Electronics	Smart Watch	680.04	47
275	Luis Freeman	sabrina.little@outlook.com	66	Male	Christina land	13	Electronics	Headphones	473.97	33.2
2985	Willie Robert	vwatson@example.com	64	Male	Davidchester	4	Clothing	T-Shirts	355.64	0
9404	Paul Lopez	williamsgary@outlook.com	31	Male	New William	9	Furniture	Beds	412.55	40.4
5848	Brittany Kent	richardslisa@outlook.com	46	Male	Bettyside	48	Clothing	Jeans	255.42	27.2
3477	Joseph Browne	marc38@example.com	45	Female	Kellyshire	48	Clothing	Jeans	353.01	44.0
7578	Jessica Browne	ohowell@example.com	55	Male	South Nathaniel	4	Clothing	T-Shirts	689.7	18.6
3558	Michael Bonner	rangelashley@outlook.com	42	Female	East Anthony	16	Furniture	Tables	349.35	3.7
2809	Brandon King	edward51@example.com	32	Male	Virginiatown	1	Electronics	Cameras	667.87	31.1
9112	Kenneth Meyer	manuel93@example.com	57	Female	Loristad	6	Furniture	Chairs	832.43	20.1
4861	Jorge Grant	robertschrist@example.com	63	Female	Angelicashire	1	Electronics	Cameras	967.62	18.6
9828	Henry Rodriguez	jonathan80@example.com	48	Female	Thompsonvil	12	Clothing	Dresses	464.37	4.6
3416	David Simms	patrickkaren@example.com	39	Male	South Ryansi	10	Electronics	Laptops	71.99	9
307	Nathan Hunt	ryan66@example.com	52	Male	Jacobsdie	11	Clothing	Hats	801.64	31.3
3908	Tammy Yourterry	terry@exampl.com	23	Male	North Patrick	0	Electronics	Smart Watch	23.55	27.7
348	Marcus Robison	pearsonmeli@example.com	54	Female	Clarkmouth	48	Clothing	Jeans	587.9	25.3
826	Christine Davy	mandy50@example.com	65	Male	Ortegaport	4	Clothing	T-Shirts	550.14	7.7

```
[{"CustomerID":5344,"Name":"Thomas Compton","Email":"rgarrison@example.org","Age":58,"Gender":"Male","ProductID":2,"Category":"Electronics","Subcategory":"Mobile Phones","Price":93.68,"Weight":30,"DateOrder":"2022-04-20","TotalAmount":"\u00a371.71","ShippingMethod":"Standard","Status":"Delivered","RestockThreshold":448,"LastRestockDate":"2021-10-06","WarehouseID":462,"LocationWarehouse":"Newcastle","SupplierID":2,"Company":"Miller-Pearson","LocationSupplier":"Clarkmouth","QualityScore":9.83,"Address": "123 Main Street, Clarkmouth, UK","Postcode": "NG12 3AB","Latitude": 53.5, "Longitude": -2.2}, {"CustomerID":1731,"Name":"Cody Molina","Email":"cody.molina@example.com","Age":35,"Gender":"Female","ProductID":21,"Category":"Furniture","Subcategory":"Tables","Price":279.55,"Weight":17.39,"Dimensions":"33x21x6","OrderID":7444,"DateOrder":"2022-05-01","TotalAmount":"\u00a3996.12","ShippingMethod":"Standard","Status":"Shipped","InventoryID":7444,"LastRestockDate":"2021-08-23","WarehouseID":89,"LocationWarehouse":"Bruceside","Capacity":931,"LocationSupplier":"Clarkmouth","QualityScore":9.83}, {"CustomerID":1731,"Name":"Cody Molina","Email":"cody.molina@example.com","Age":35,"Gender":"Female","ProductID":21,"Category":"Furniture","Subcategory":"Tables","Price":279.55,"Weight":17.39,"Dimensions":"33x21x6","OrderID":1731,"DateOrder":"02/05/2022","TotalAmount":"\u00a3996.12"}]
```

# V. TALEND

## 1.Extraction des Données

Nous avons extrait les données depuis les deux fichiers csv et json, puis nous avons fusionné ces données avec le composant tUnite? Pour achever cette fusion, les noms des colonnes et leurs types devaient être les mêmes.



## 2. Transformation des données

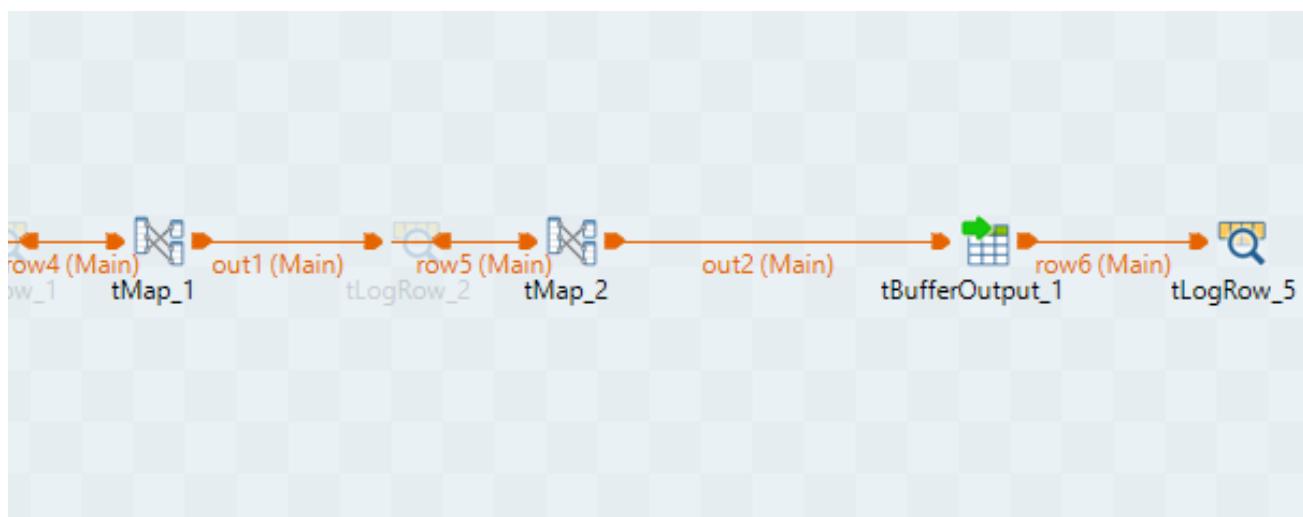
The screenshot shows the Talend Data Integration interface with three main panels:

- row4:** A list of columns from the input dataset, including CustomerID, Name, Email, Age, Gender, Location, ProductID, Category, Subcategory, Price, Weight, Dimensions, OrderID, DateOrder, TotalAmount, ShippingMethod, Status, InventoryID, and QualityScore.
- Var:** A configuration panel for variables. It contains two rows:
  - Expression: StringHandling.LEFT(row4.TotalA...) Type: double Variable: taux
  - Expression: StringHandling.CHANGE(row4.T... Type: String Variable: TotalAmountNum...
- out1:** A list of columns for the output dataset, which includes all the columns from the input dataset except QualityScore, plus additional calculated columns: Math.abs(row4.Age), row4.Gender, row4.Location, row4.ProductID, row4.Category, row4.Subcategory, row4.Price, row4.Weight, row4.Dimensions, row4.Dimensions.matches(".\*?\d+.\*"), row4.Dimensions.matches(".\*?\d+x(\\"...), row4.Dimensions.matches(".\*x(\\"d+).\*"), row4.OrderID, row4.DateOrder.charAt(2) == '-' ? Talen..., Var.TotalAmountNumber, and row4.ShippingMethod.

Nous avons commencé la transformation par unifier le format de la date car on avait des dates en formats différents. Nous avons aussi appliqué la valeur absolue sur la colonne d'âge, ainsi qu'on a rempli les valeurs nulles dans la colonne QualityScore par 0 pour qu'on puisse la remplir par la moyenne dans la transformation suivante. La colonne Category avait des caractères spéciaux, donc on a dû les supprimer. La colonne Totalamount avait des montants en devise différentes, on a dû les unifier en usd.

row5	Var	out2
Column		Column
CustomerID		CustomerID
Name		Name
Age		Age
Gender		Gender
Location		Location
ProductID		ProductID
Category		Category
Subcategory		Subcategory
Price		Price
Weight		Weight
Dimensions		Dimensions
OrderID		OrderID
DateOrder		DateOrder
TotalAmount		TotalAmount
ShippingMethod		ShippingMethod
Status		Status
InventoryID		InventoryID
StockLevel		StockLevel
RestockThreshold		RestockThreshold

Dans la deuxième étape de transformation, nous avons supprimé les valeurs moins de 13 et plus que 87 dans la colonne Age. Nous avons changé le type des valeurs dans la colonne QualityScore en Float (double).



DateOrder	row4.Dimensions	Dimensions
TotalAmount	row4.Width	Lenght
ShippingMethod	row4.Height	Width
Status	row4.Depth	Height
InventoryID	row4.OrderID	OrderID
StockLevel	row4.DateOrder	DateOrder
RestockThreshold	Float.valueOf(row4.TotalAmount)	TotalAmount
LastRestockDate	row4.ShippingMethod	ShippingMethod
WarehouseID	row4.Status	Status
LocationWarehouse	row4.InventoryID	InventoryID
Capacity	row4.StockLevel	StockLevel
SupplierID	row4.RestockThreshold	RestockThreshold
Company	row4.LastRestockDate	LastRestockDate
LocationSupplier	row4.WarehouseID	WarehouseID
QualityScore	row4.LocationWarehouse	LocationWarehouse
	row4.Capacity	Capacity
	row4.SupplierID	SupplierID
	row4.Company	Company
	row4.LocationSupplier	LocationSupplier
	row4.QualityScore == 0.0 ? row3.Quali...	QualityScore

row3

Clé d'expr.

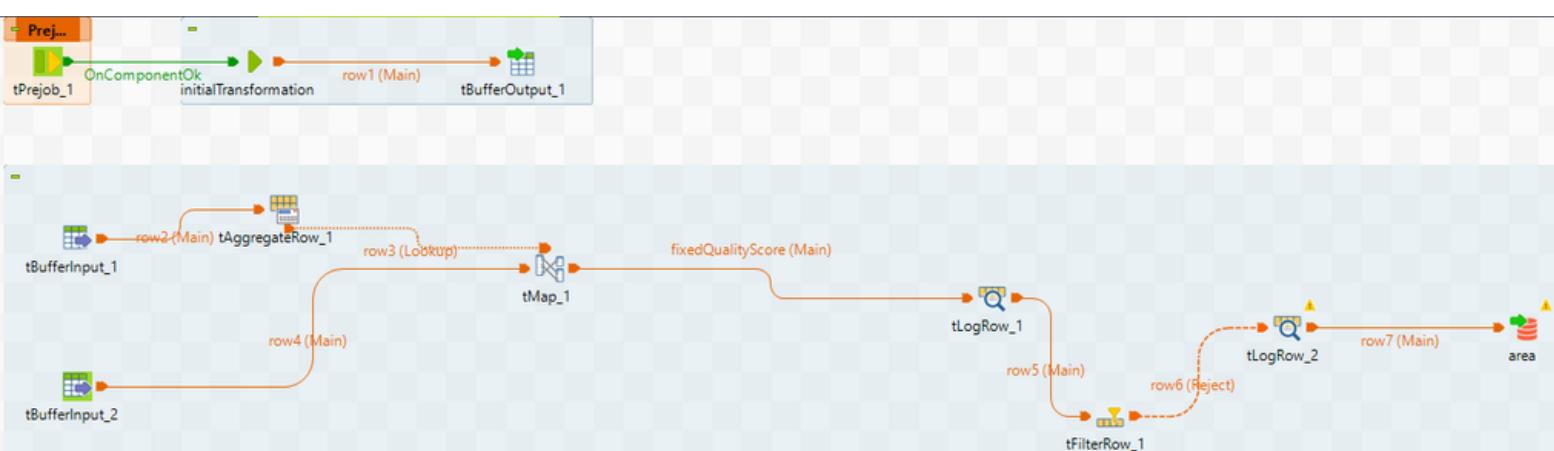
row4.Company

Column

Company

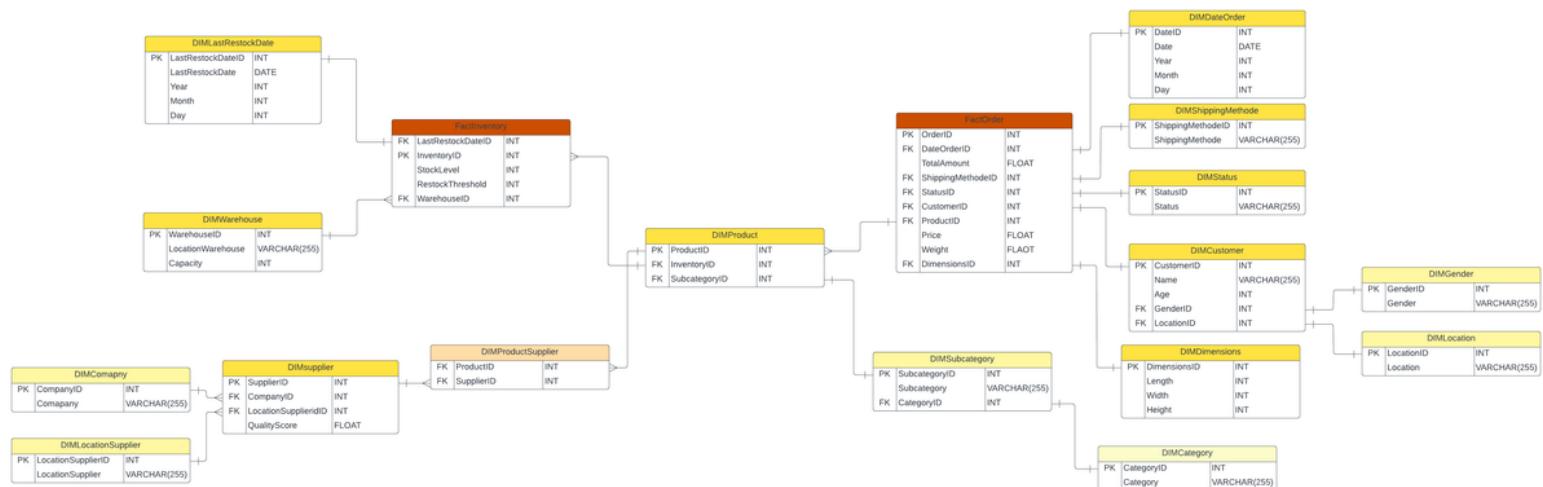
QualityScore

Dans la troisième étape de transformation, nous avons remplacé les valeurs '0' dans la colonne qualityscore par la moyenne à l'aide du composant tAggregate qui nous permet de faire des agrégations sur des colonnes numériques précises.



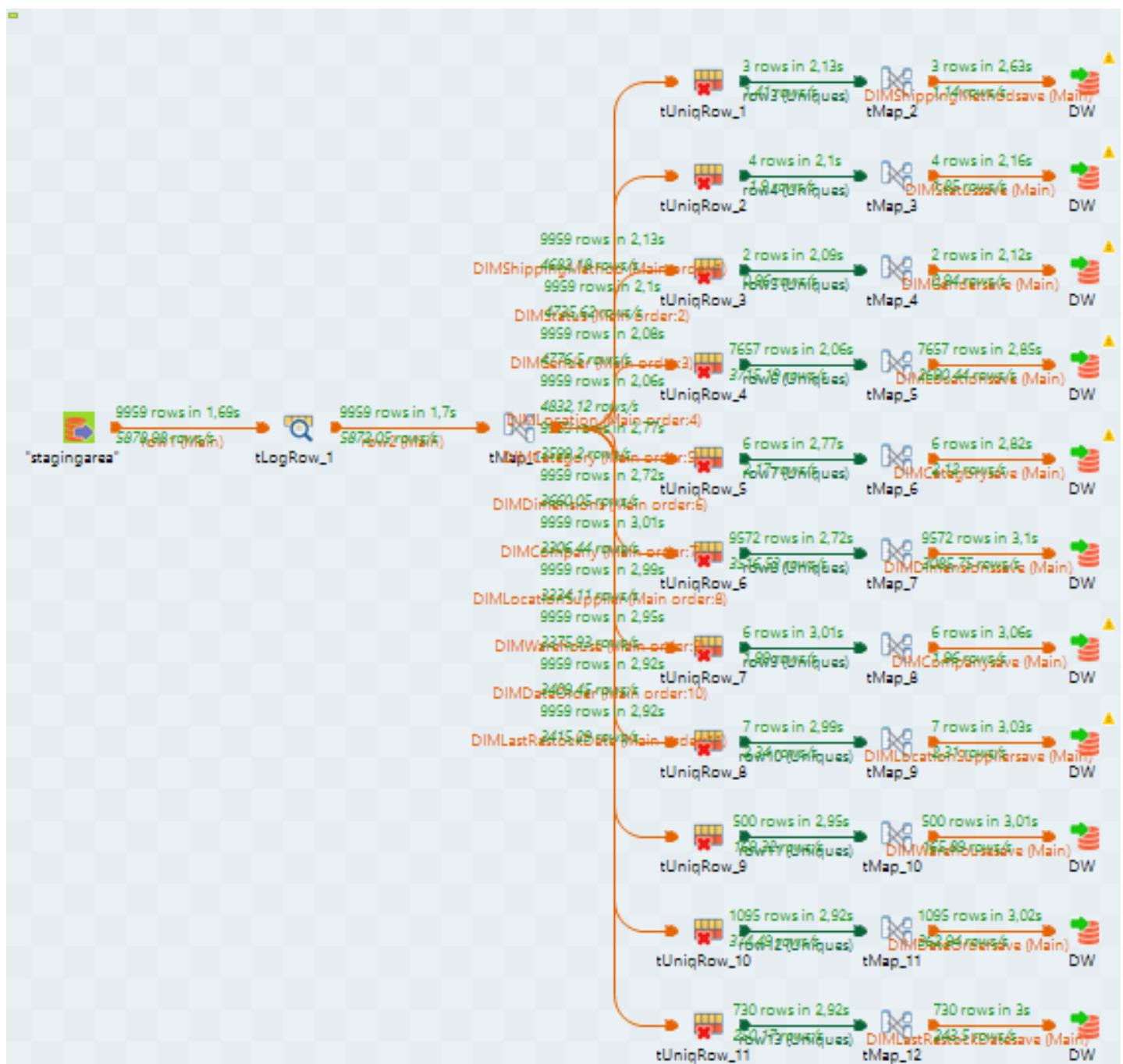
### 3. Chargement dans un data warehouse

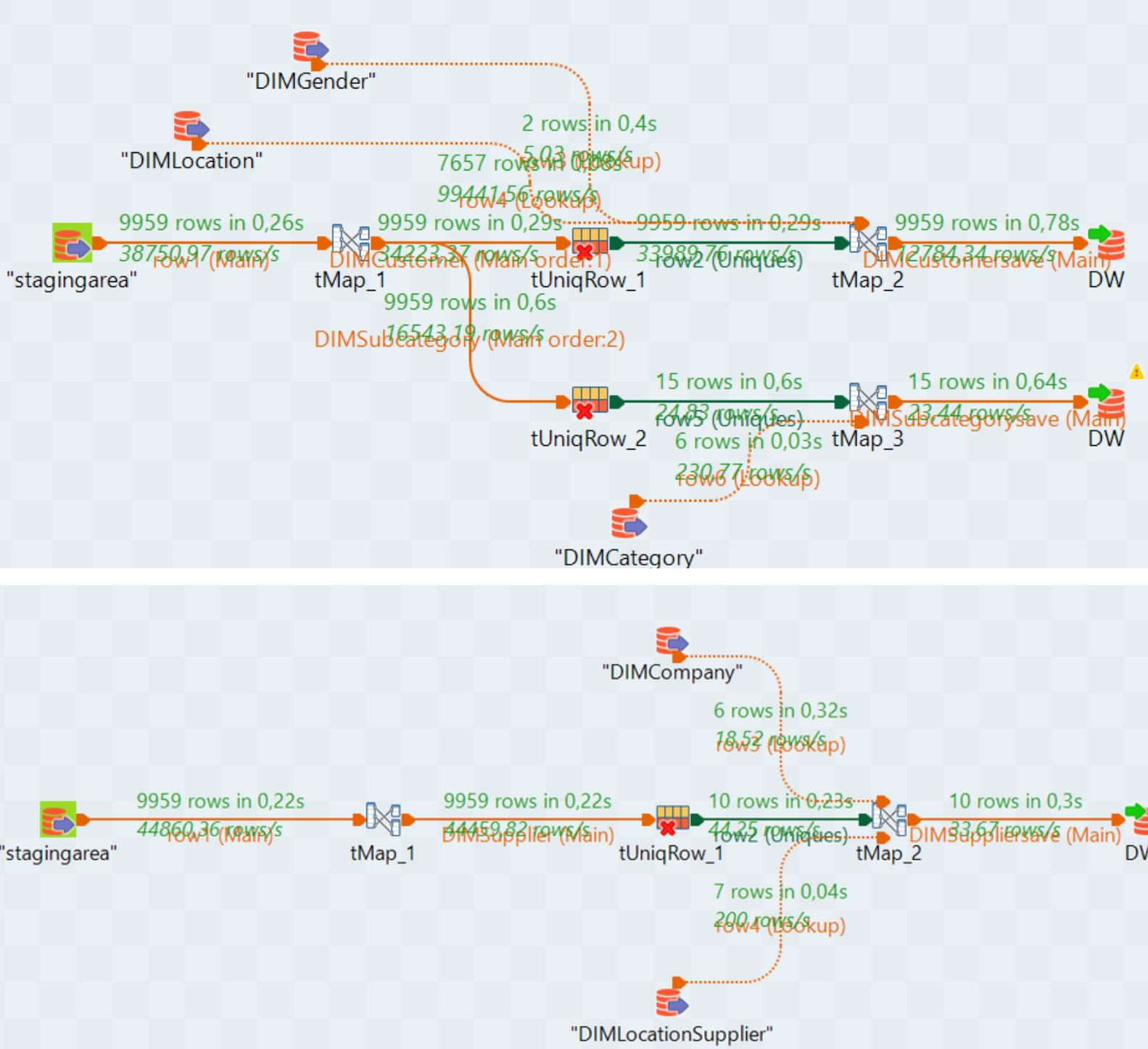
Avant de charger les données dans un data warehouse, nous avons réalisé la conception ci-dessous :



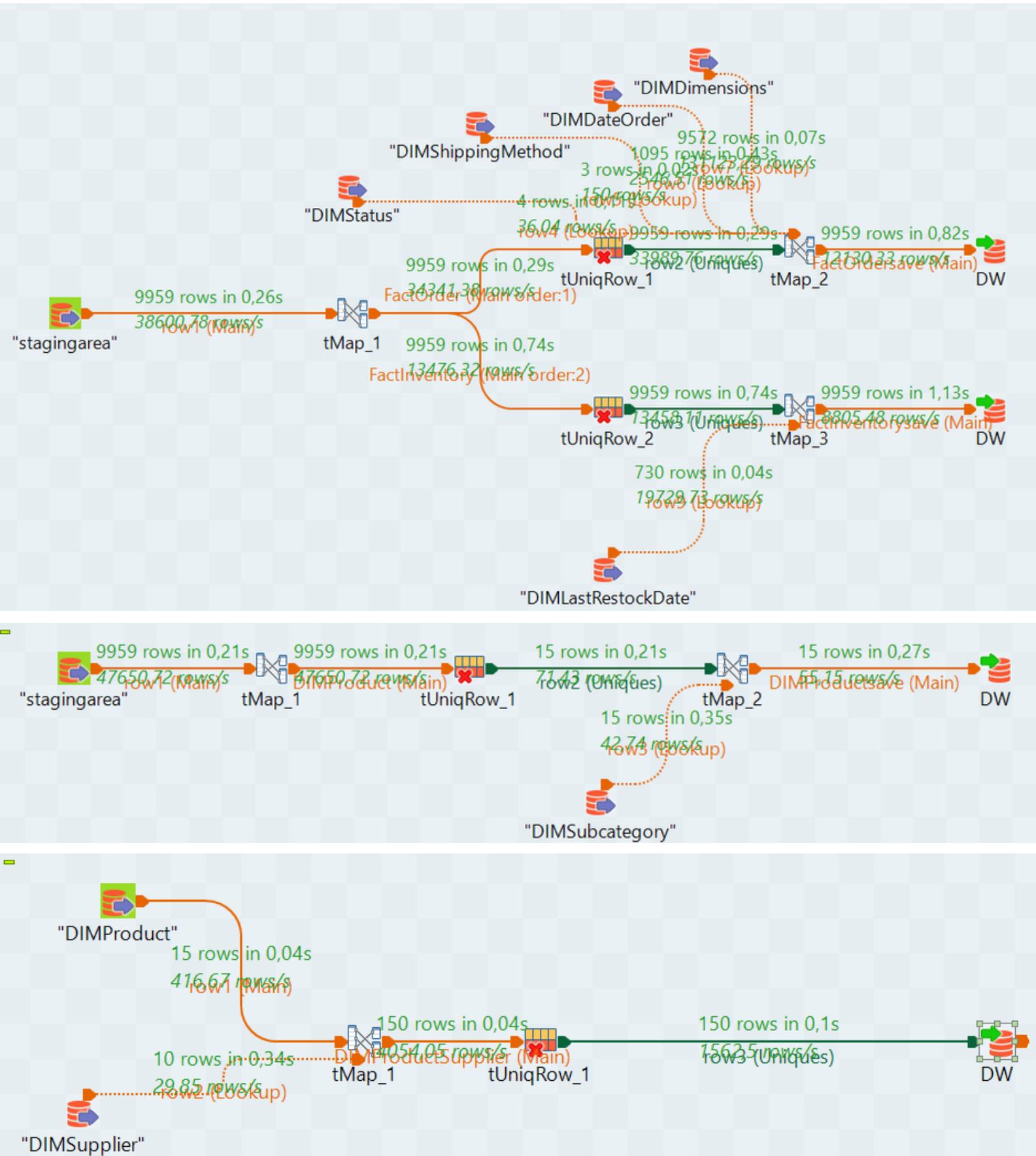
Ce schéma s'appelle schéma fact constellation, il se caractérise par avoir 2 ou plus de tables de fait (fact tables).

Nous avons séparé les dimension à l'aide de composant tMap. Nous avons dû suivre un ordre précis dans cette création, nous avons commencé par les dimensions externes, puis les internes, puis les tables de fait et enfin les tables associatives.





Nous avons aussi implémenté la 'staging area' avant la création de chaque groupe de dimensions.



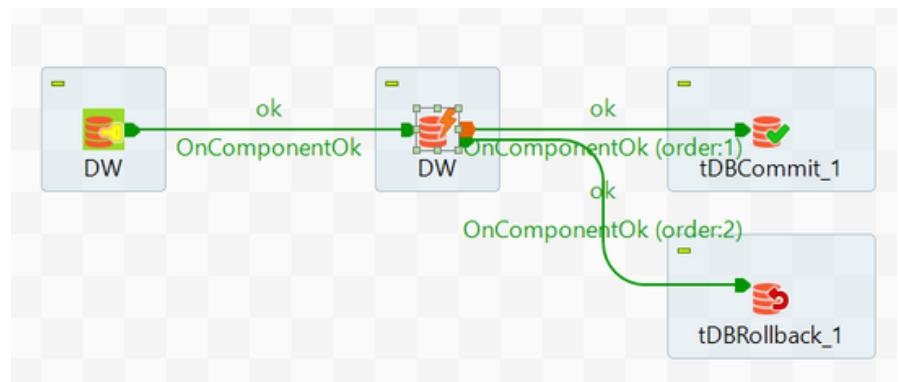
Nous avons créé les clés étrangères pour établir les relations entre les dimensions.

```
ALTER TABLE DIMCustomer
ADD CONSTRAINT fk_DIMCustomer_DIMGender
FOREIGN KEY (GenderID)
REFERENCES DIMGender(GenderID)
ON DELETE CASCADE;
```

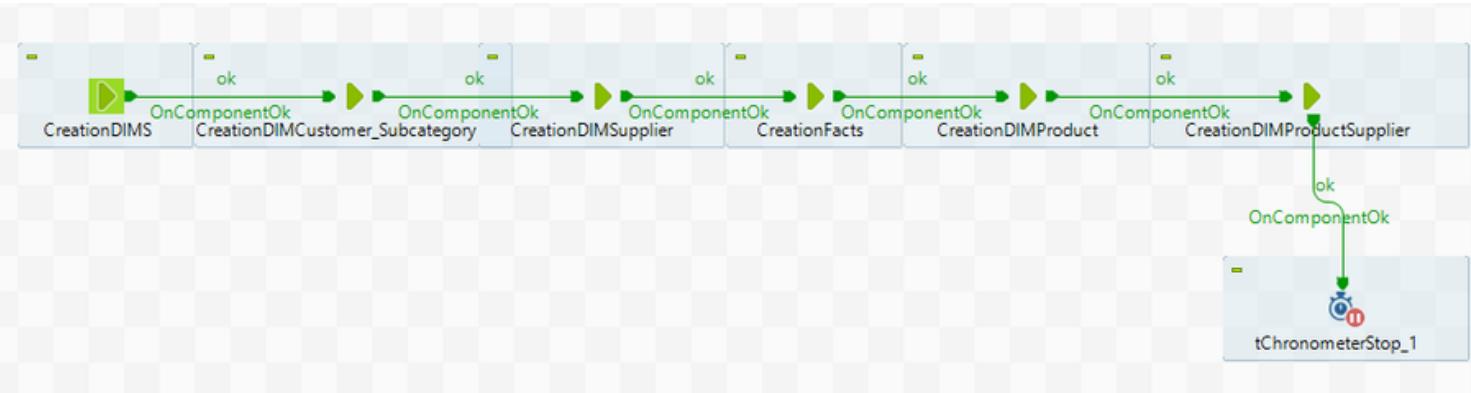
```
ALTER TABLE DIMCustomer
ADD CONSTRAINT fk_DIMCustomer_DIMLocation
FOREIGN KEY (LocationID)
REFERENCES DIMLocation(LocationID)
ON DELETE CASCADE;
```

```
ALTER TABLE FactOrder
ADD CONSTRAINT fk_FactOrder_DIMDateOrder
FOREIGN KEY (DateOrderID)
REFERENCES DIMDateOrder(DateOrderID)
ON DELETE CASCADE;
```

```
ALTER TABLE FactOrder
ADD CONSTRAINT fk_FactOrder_DIMShippingMe
```



Voici une vue générale sur le job et les sous job de notre projet :



tChronometerStop\_1 ] 5124 milliseconds  
statistics] disconnected

Le temps d'exécution est minimal.

```
CREATE INDEX IX_FactOrder_CustomerID ON FactOrder (CustomerID);
CREATE INDEX IX_FactOrder_ProductID ON FactOrder (ProductID);
CREATE INDEX IX_FactOrder_ShippingMethodID ON FactOrder (ShippingMethodID);
CREATE INDEX IX_FactOrder_StatusID ON FactOrder (StatusID);
CREATE INDEX IX_FactOrder_DimensionsID ON FactOrder (DimensionsID);
CREATE INDEX IX_FactOrder_DateOrderID ON FactOrder (DateOrderID);

CREATE INDEX IX_FactInventory_InventoryID ON FactInventory (InventoryID);
CREATE INDEX IX_FactInventory_WarehouseID ON FactInventory (WarehouseID);
CREATE INDEX IX_FactInventory_LastRestockID ON FactInventory (LastRestockDateID);

CREATE INDEX IX_DIMProductSupplier_ProductID ON DIMProductSupplier (ProductID);
CREATE INDEX IX_DIMProductSupplier_SupplierID ON DIMProductSupplier (SupplierID);
```

Les indexes créés sur les tables FactOrder, FactInventory et DIMProductSupplier servent à optimiser les performances des requêtes dans des scénarios spécifiques. Par exemple, ils permettent une récupération plus rapide des commandes par client, produit, méthode d'expédition, statut et date, améliorant ainsi l'efficacité des opérations liées aux clients et aux commandes. De même, les indexes sur les colonnes liées à l'inventaire améliorent la vitesse des requêtes d'inventaire, en particulier en ce qui concerne les identifiants d'inventaire uniques, les associations d'entrepôts et les dates de réapprovisionnement. Enfin, les indexes sur les identifiants de produit et de fournisseur accélèrent les recherches liées aux relations produit-fournisseur, simplifiant les tâches liées aux fournisseurs et aux produits. Ces indexes améliorent collectivement la réactivité de la base de données et les performances de l'application.

```

-- Create filegroups for DataCommerceDB database
ALTER DATABASE DataCommerceDB
ADD FILEGROUP DimDateOrder_Historical;

ALTER DATABASE DataCommerceDB
ADD FILEGROUP DimDateOrder_Recent;

-- Create files within filegroups (update paths accordingly)
ALTER DATABASE DataCommerceDB
ADD FILE (
    NAME = DimDateOrder_Historical_File,
    FILENAME = 'C:\Users\Youcode\Desktop\DataCommerceDB_DimDateOrder_Historical.ndf'
) TO FILEGROUP DimDateOrder_Historical;

ALTER DATABASE DataCommerceDB
ADD FILE (
    NAME = DimDateOrder_Recent_File,
    FILENAME = 'C:\Users\Youcode\Desktop\DataCommerceDB_DimDateOrder_Recent.ndf'
) TO FILEGROUP DimDateOrder_Recent;

-- Create a partition function for DataCommerceDB database
CREATE PARTITION FUNCTION DateOrderPartitionFunction(DATE)
AS RANGE LEFT FOR VALUES ('2021-01-01');

-- Create a partition scheme for DataCommerceDB database
CREATE PARTITION SCHEME DateOrderPartitionScheme
AS PARTITION DateOrderPartitionFunction
TO (
    DimDateOrder_Historical,
    DimDateOrder_Recent
);

-- Create the partitioned table without a primary key constraint
CREATE TABLE DataCommerceDB.dbo.DimDateOrder_Partitioned (
    DateOrderID INT,
    DateOrder DATE,
    Year INT,
    Month INT,
    Day INT
) ON DateOrderPartitionScheme(DateOrder);

```

Ce script SQL effectue plusieurs tâches pour configurer la partition de la table nommée DimDateOrder dans la base de données DataCommerceDB. Tout d'abord, il crée deux groupes de fichiers (DimDateOrder\_Historical et DimDateOrder\_Recent) et spécifie les chemins des fichiers pour chacun d'eux. Ensuite, il définit une fonction de partition qui détermine comment les données seront divisées en fonction d'une colonne de date. Ensuite, il crée un schéma de partition qui associe les partitions aux groupes de fichiers précédemment créés. Enfin, il crée une table partitionnée (DimDateOrder\_Partitioned) basée sur ce schéma de partition, ce qui permet de stocker et de récupérer efficacement des données en fonction de plages de dates, optimisant ainsi les performances des requêtes.

```

-- Create individual partitions for DataCommerceDB database
CREATE TABLE DataCommerceDB.dbo.DimDateOrder_Partitioned_2020 (
    DateOrderID INT,
    DateOrder DATE,
    Year INT,
    Month INT,
    Day INT
) ON DateOrderPartitionScheme(DateOrder);

CREATE TABLE DataCommerceDB.dbo.DimDateOrder_Partitioned_2021 (
    DateOrderID INT,
    DateOrder DATE,
    Year INT,
    Month INT,
    Day INT
) ON DateOrderPartitionScheme(DateOrder);

CREATE TABLE DataCommerceDB.dbo.DimDateOrder_Partitioned_2022 (
    DateOrderID INT,
    DateOrder DATE,
    Year INT,
    Month INT,
    Day INT
) ON DateOrderPartitionScheme(DateOrder);

CREATE TABLE DataCommerceDB.dbo.DimDateOrder_Partitioned_2023 (
    DateOrderID INT,
    DateOrder DATE,
    Year INT,
    Month INT,
    Day INT
) ON DateOrderPartitionScheme(DateOrder);

```

Ce code crée des tables partitionnées individuelles pour différentes années (2020, 2021, 2022 et 2023) en fonction de la colonne DateOrder à l'aide du schéma de partition DateOrderPartitionScheme. Chacune de ces tables partitionnées stockera des données pour une année spécifique, ce qui permet un stockage et une récupération efficaces des données en fonction de l'année. Cette stratégie de partitionnement aide à optimiser les performances des requêtes lors de la manipulation de données basées sur la date dans votre table DimDateOrder. Chaque table contiendra le même schéma (colonnes), mais les données seront physiquement séparées par année pour une gestion plus facile et une optimisation des requêtes.

```

-- Insert data from DIMDateOrder into corresponding partitions
-- Insert data for the year 2020
INSERT INTO DataCommerceDB.dbo.DimDateOrder_Partitioned_2020 (DateOrderID, DateOrder, Year, Month, Day)
SELECT
    DateOrderID,
    DateOrder,
    YEAR(DateOrder) AS Year,
    MONTH(DateOrder) AS Month,
    DAY(DateOrder) AS Day
FROM DIMDateOrder
WHERE YEAR(DateOrder) = 2020;

-- Insert data for the year 2021
INSERT INTO DataCommerceDB.dbo.DimDateOrder_Partitioned_2021 (DateOrderID, DateOrder, Year, Month, Day)
SELECT
    DateOrderID,
    DateOrder,
    YEAR(DateOrder) AS Year,
    MONTH(DateOrder) AS Month,
    DAY(DateOrder) AS Day
FROM DIMDateOrder
WHERE YEAR(DateOrder) = 2021;

-- Insert data for the year 2022
INSERT INTO DataCommerceDB.dbo.DimDateOrder_Partitioned_2022 (DateOrderID, DateOrder, Year, Month, Day)
SELECT
    DateOrderID,
    DateOrder,
    YEAR(DateOrder) AS Year,
    MONTH(DateOrder) AS Month,
    DAY(DateOrder) AS Day
FROM DIMDateOrder
WHERE YEAR(DateOrder) = 2022;

-- Insert data for the year 2023
INSERT INTO DataCommerceDB.dbo.DimDateOrder_Partitioned_2023 (DateOrderID, DateOrder, Year, Month, Day)
SELECT
    DateOrderID,
    DateOrder,
    YEAR(DateOrder) AS Year,
    MONTH(DateOrder) AS Month,
    DAY(DateOrder) AS Day
FROM DIMDateOrder
WHERE YEAR(DateOrder) = 2023;

```

Ce code effectue l'insertion de données à partir de la table source DIMDateOrder dans les partitions correspondantes des tables partitionnées pour différentes années (2020, 2021, 2022 et 2023) que vous avez créées. Chaque instruction INSERT INTO copie des données depuis la table source et les mappe vers la table partitionnée appropriée en fonction de l'année extraite de la colonne DateOrder. Les fonctions YEAR, MONTH et DAY sont utilisées pour extraire ces composantes de la colonne DateOrder. Cette approche partitionnée organise et stocke efficacement les données par année, optimisant ainsi les performances des requêtes lors de la manipulation de plages de dates spécifiques.

## SQLQuery1.sql - 17...merceDB (user (65))\* X

```
-- Query data from the partitioned table for the year 2021 and join with other tables
SELECT DDOP.DateOrderID,
       DDOP.DateOrder,
       DDOP.Year,
       DDOP.Month,
       DDOP.Day,
       O.TotalAmount,
       C.Name AS CustomerName
  FROM DataCommerceDB.dbo.DimDateOrder_Partitioned_2021 DDOP
 INNER JOIN FactOrder O ON DDOP.DateOrderID = O.OrderID
 INNER JOIN DIMCustomer C ON O.CustomerID = C.CustomerID
 WHERE DDOP.Year = 2021;
```

73 %

Results Messages

	DateOrderID	DateOrder	Year	Month	Day	TotalAmount	CustomerName
2	3	2021-02-21	2021	2	21	1876,08	Thomas Smith
3	4	2021-03-13	2021	3	13	170,67	Matthew Lee
4	11	2021-09-04	2021	9	4	3996,36	Sara Lawrence
5	14	2021-11-29	2021	11	29	2603,36	Laurie McKee
6	17	2021-07-09	2021	7	9	726	William Hardy
7	18	2021-04-16	2021	4	16	244,74	Steven Thomas
8	25	2021-09-01	2021	9	1	2780,67	Mr. Alec Cunningham
9	28	2021-04-25	2021	4	25	552,15	Catherine Mendez
10	29	2021-02-12	2021	2	12	1533,9	Amanda Black
11	33	2021-03-12	2021	3	12	1251,84	Kevin Price
12	36	2021-05-02	2021	5	2	4679,6	Jason Velasquez
13	44	2021-04-01	2021	4	1	2068,02	Rachel Jenkins
14	46	2021-10-08	2021	10	8	4151,15	Christine Roberts
15	49	2021-08-31	2021	8	31	1724,36	Jaimie Chapman

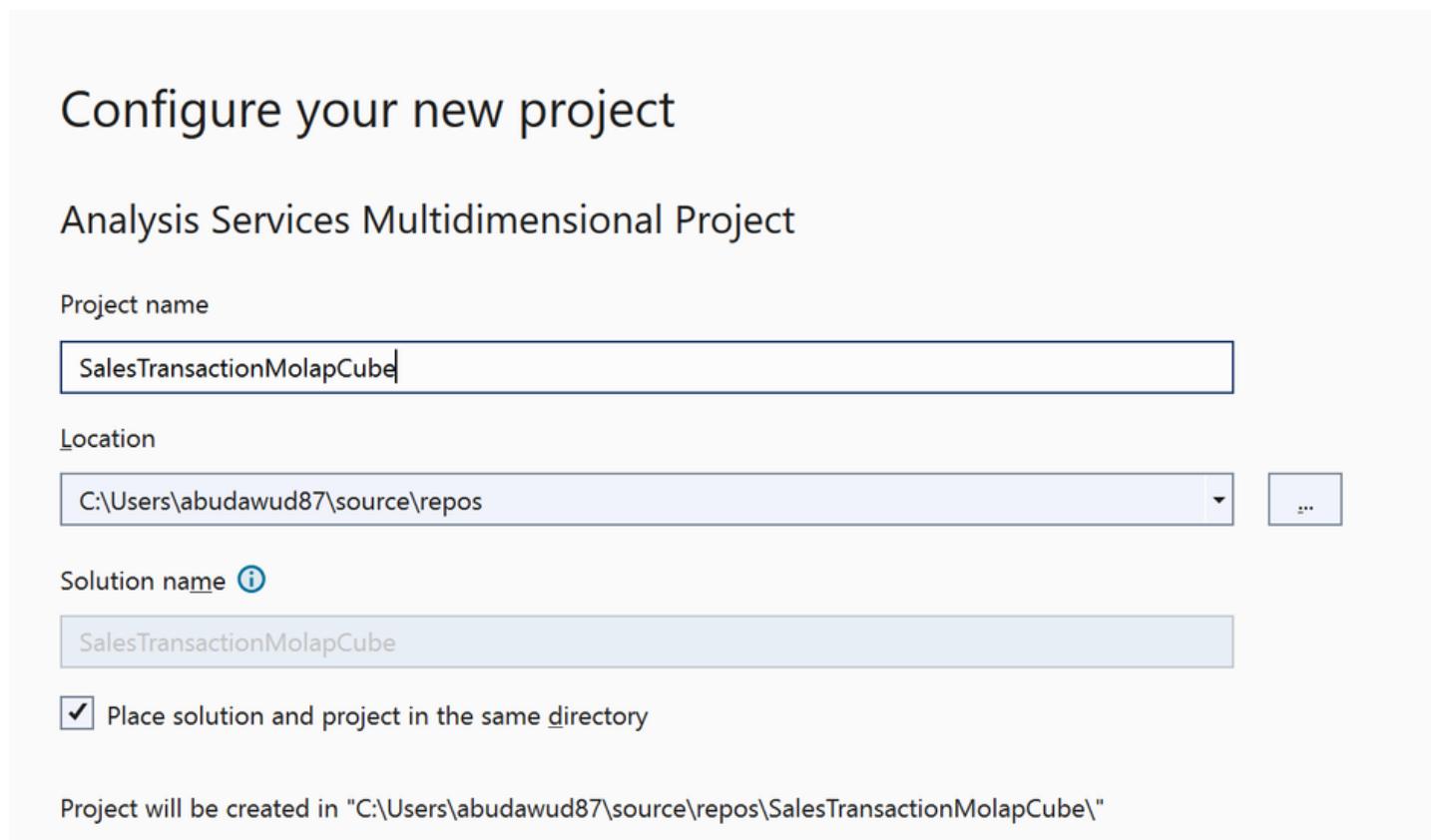
Ce code sélectionne des données de la table partitionnée DimDateOrder\_Partitioned\_2021 et les joint avec les tables FactOrder et DIMCustomer en fonction des conditions spécifiées. Cela permettra de récupérer des données pour l'année 2021 ainsi que les montants totaux et les noms des clients.

# VI. OLAP CUBE ET TEST UNITAIRE

## 1. Crédit d'OLAP cube

Pour créer un OLAP cube, nous avons passé par 4 étapes principales :

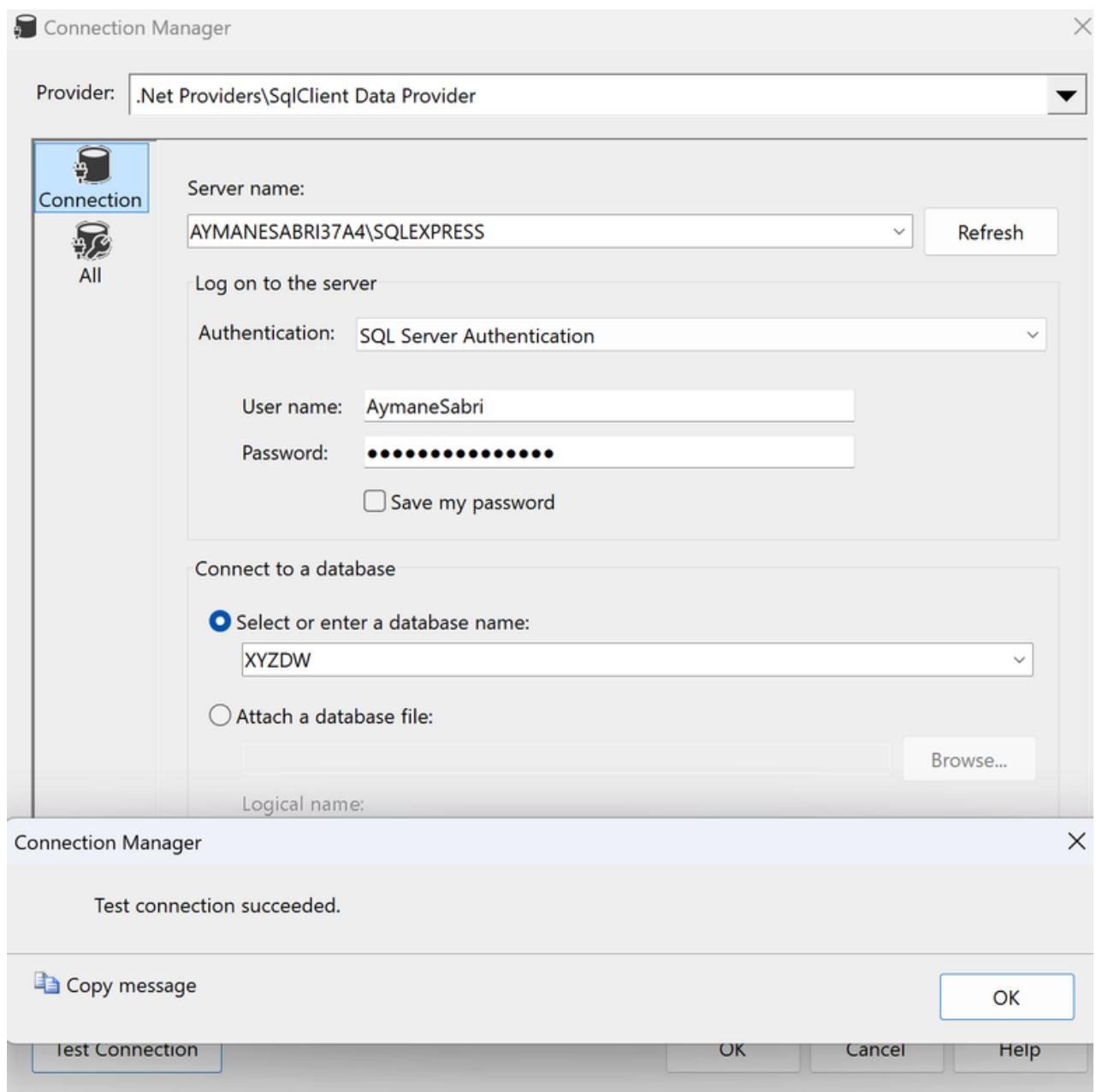
### A. Analysis Services Project Creation

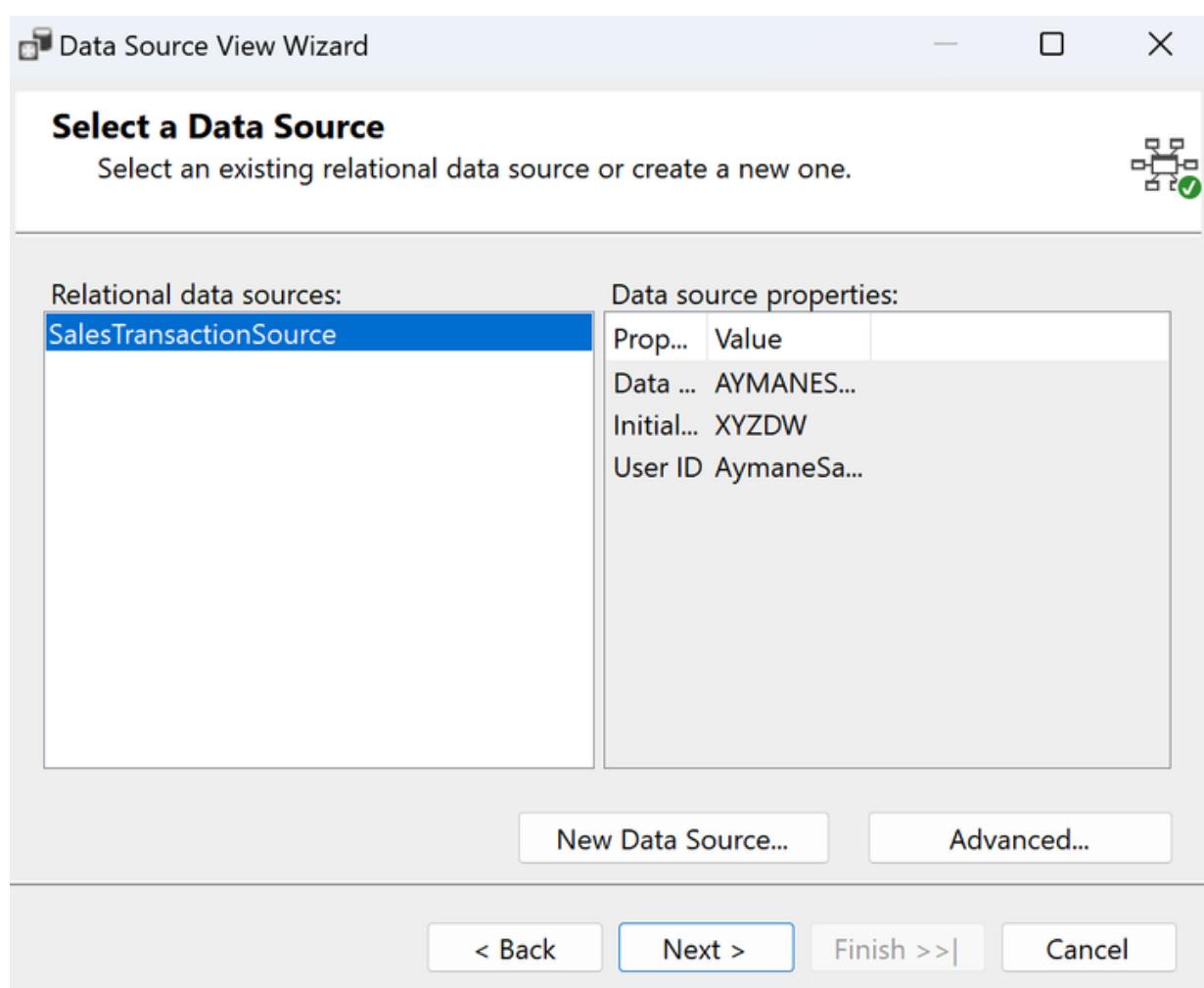
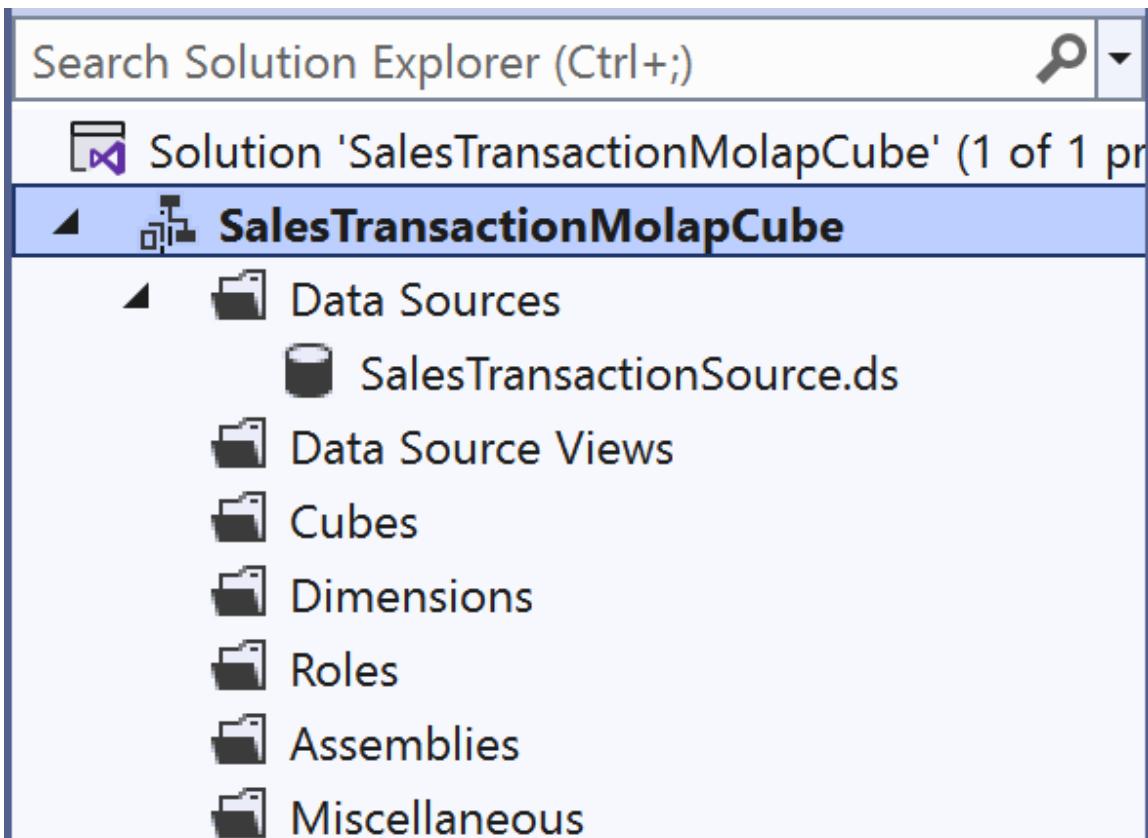


Dans cette étape, nous avons créé un nouveau projet de services d'analyse multidimensionnelle.

## B.Deploy Data Source & Connection Configuration

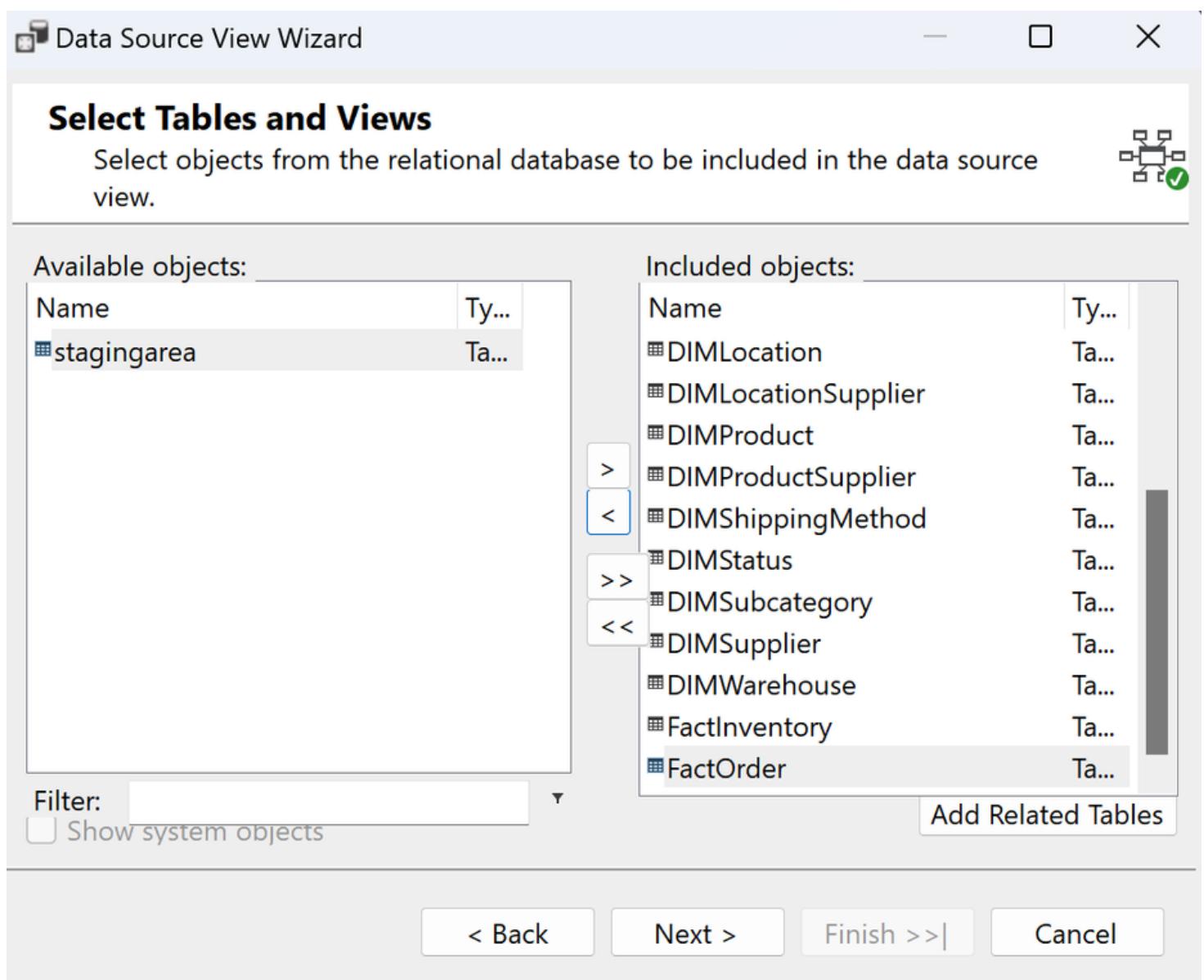
Dans cette étape, nous avons déployé la source de données ainsi que nous avons configuré la connection avec sql server.

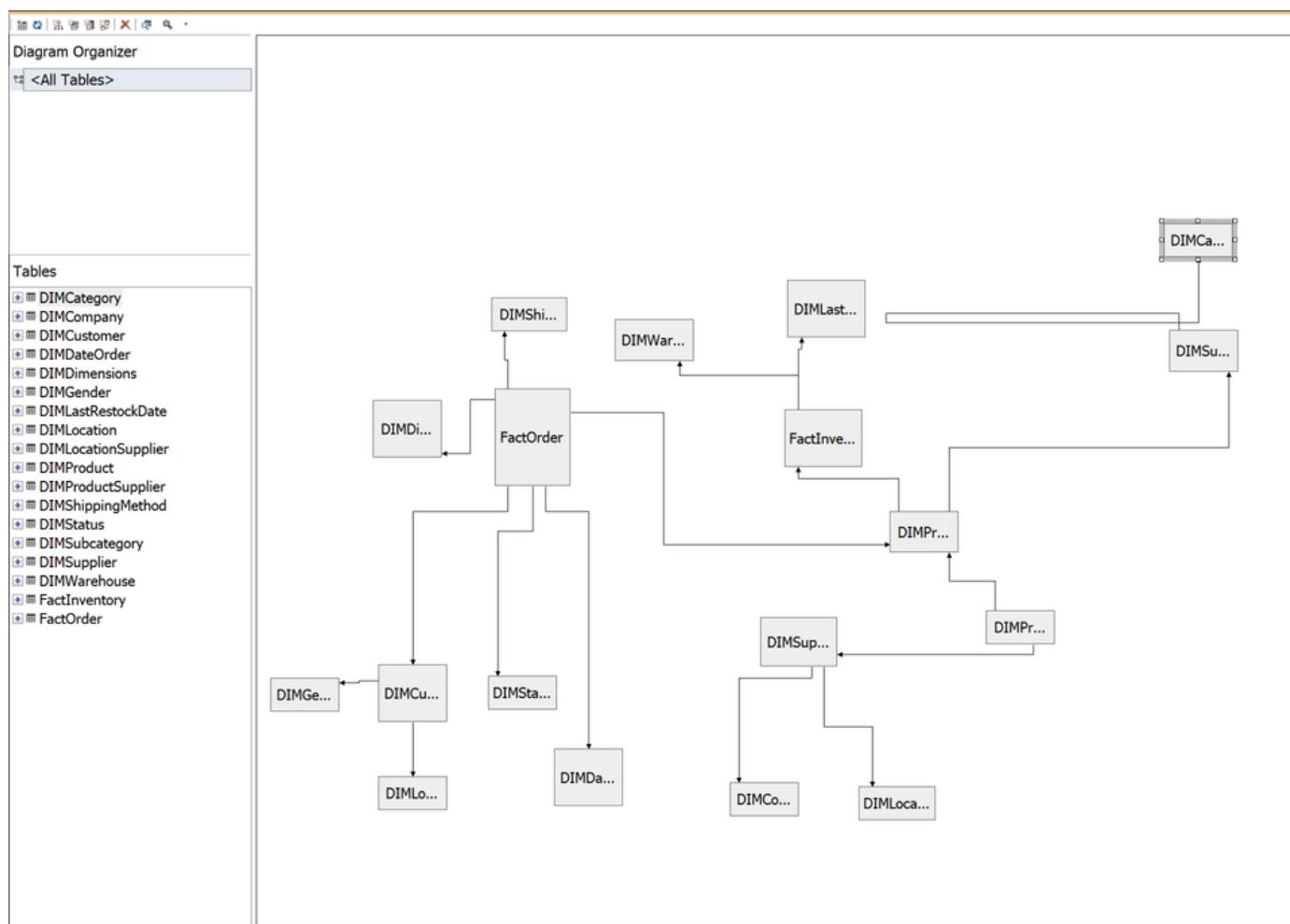
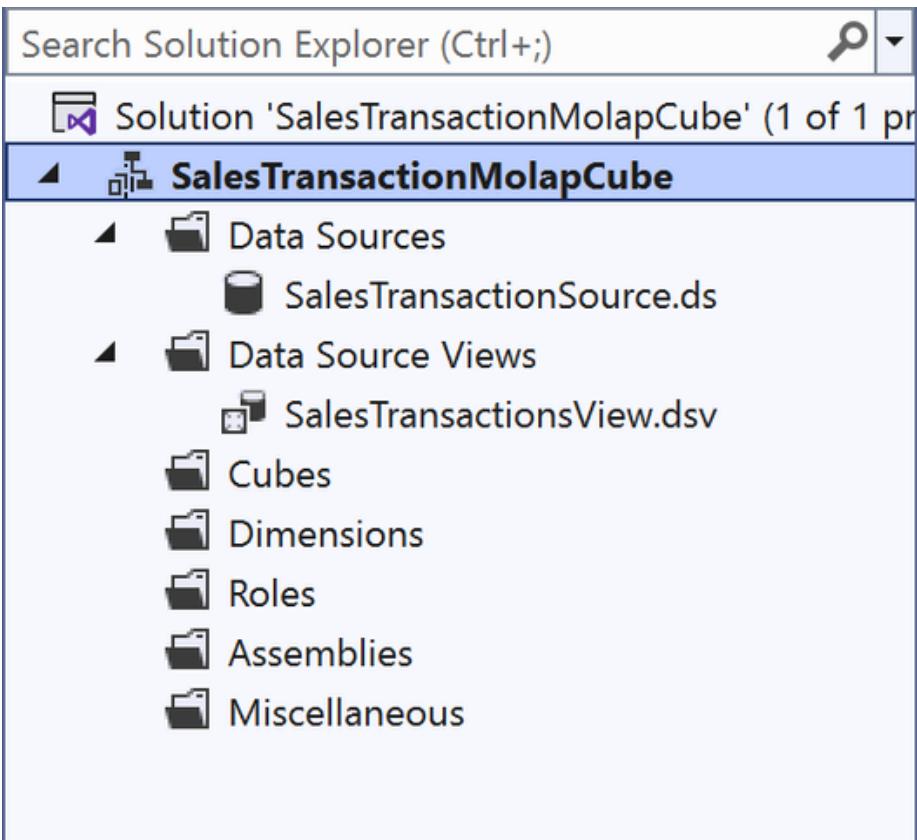




## C.Source View Creation

Dans cette étape, nous avons créé des view des notre source de données. Nous avons sélectionné les table qui feront partie de notre view.





## D.Cube Creation, Verification & Deployment

Dans cette étape, nous avons créé notre cube OLAP en utilisant les tables qu'on avait sélectionné dans l'étape précédente. Nous avons vérifié que les tables ont été importées avec les relations entre eux, et puis on a déployé le cube qui est prêt à être utilisé.

### Select Creation Method

Cubes can be created by using existing tables, creating an empty cube, or generating tables in the data source.



How would you like to create the cube?

Use existing tables

Create an empty cube

Generate tables in the data source

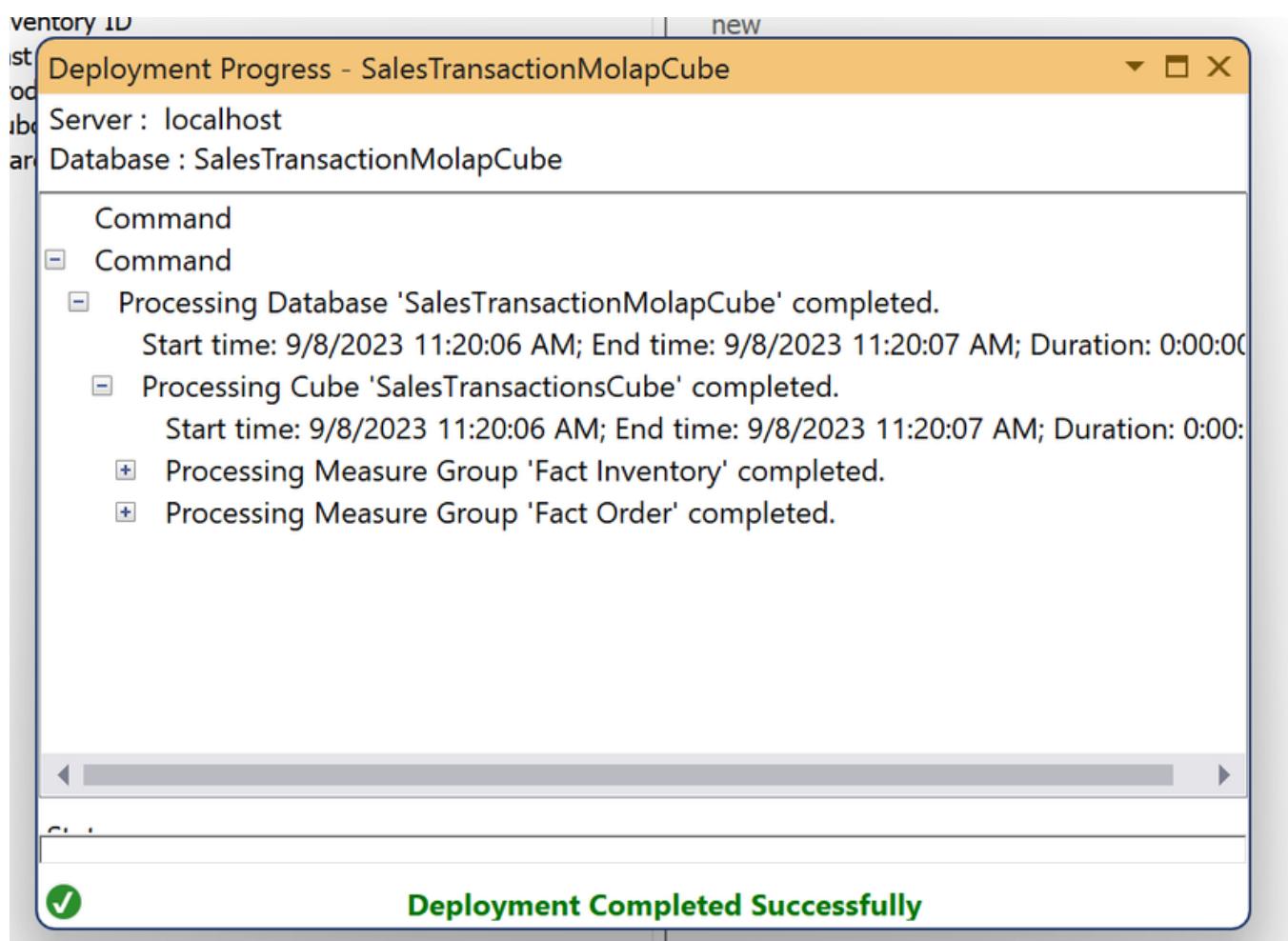
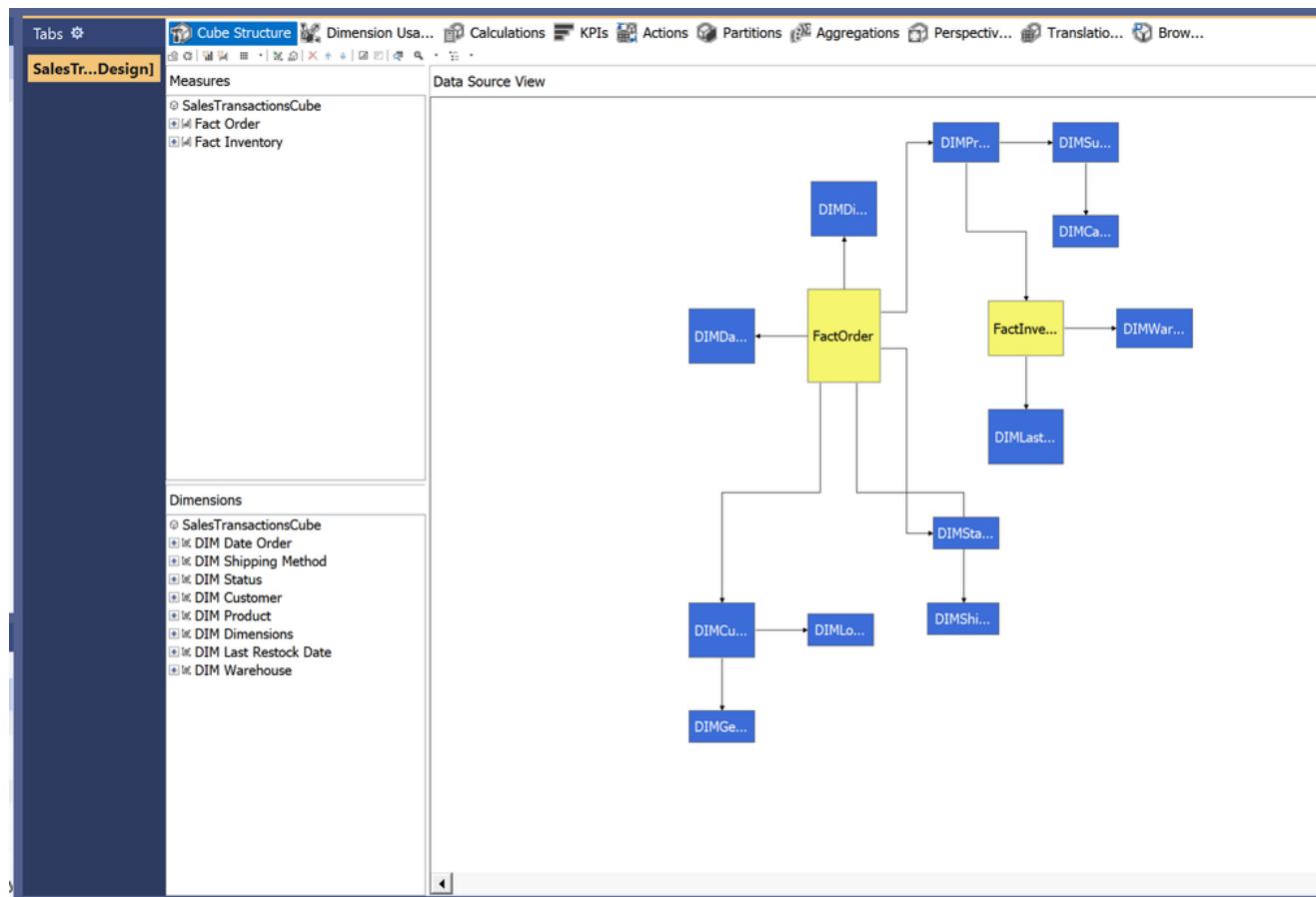
Template:

(None)

Description:

Create a cube based on one or more tables in a data source.

< Back      Next >      Finish >>|      Cancel



## Requêtes MDX :

Nous avons exécuté des requêtes MDX pour vérifier l'intégration des données dans notre cube OLAP.

The screenshot shows the Microsoft Analysis Services Management Studio (SSMS) interface. The left pane is the Object Explorer, displaying the database structure of AYMANESABRI37A4 (SQL Server). The central pane is the MDXQuery1.mdx query editor, showing an MDX script. The right pane is the Results pane, showing the execution results.

**Object Explorer:**

- AYMANESABRI37A4\SQLEXPRESS (SQL Server)
- Databases
- Security
- Server Objects
- Replication
- Management
- XEvent Profiler
- AYMANESABRI37A4 (Microsoft Analysis)
- Databases
- SalesTransactionMolapCube
- Data Sources
- Data Source Views
- Cubes
- SalesTransactionsCube
- Measure Groups
- Fact Inventory
- Fact Order Count
- Restock Threshold
- Stock Level
- Fact Order
- Fact Order Count
- Price
- Total Amount
- Weight
- KPIs
- DIM Customer
- DIM Date Order
- DIM Dimensions
- DIM Last Restock Date
- DIM Product
- DIM Shipping Method
- DIM Status
- DIM Warehouse
- Dimensions
- Mining Structures
- Roles
- Assemblies
- Management

**MDXQuery1.mdx - ...I37A4\abudawud87)\***

**Cube:** SalesTransactionsCube

**Measure Group:** <All>

```
SELECT [Measures].[Fact Order Count] ON COLUMNS
FROM SalesTransactionsCube;

// List All Branches :
SELECT [Dim Product].[Category].[All] ON COLUMNS
FROM SalesTransactionsCube;

// Total Sales by Product Category
SELECT [Measures].[Total Amount] ON COLUMNS,
[Dim Product].[Category ID].[All] ON ROWS
FROM SalesTransactionsCube;

SELECT [Measures].[Stock Level] ON COLUMNS
FROM SalesTransactionsCube;
```

**Results:**

Stock Level
4962332

100 %

Query executed successfully.

## 2. Test unitaire

**Test 1:** Nous avons vérifié s'il existe des valeurs NULL dans la colonne DIMCustomer.Name. S'il trouve des valeurs NULL, il affiche un message indiquant une violation de la règle de qualité des données, en précisant que des valeurs NULL ont été trouvées. S'il n'y a pas de valeurs NULL, il affiche un message indiquant que la règle de qualité des données a été respectée, signifiant ainsi qu'il n'y a pas de valeurs NULL dans la colonne DIMCustomer.Name.

```
-- Test for NULL values in DIMCustomer.Name
IF EXISTS (SELECT 1 FROM DIMCustomer WHERE [Name] IS NULL)
BEGIN
    -- Data quality rule violation: NULL values found in DIMCustomer.Name
    PRINT 'Data quality rule violation: NULL values found in DIMCustomer.Name';
END
ELSE
BEGIN
    -- Data quality rule passed: No NULL values in DIMCustomer.Name
    PRINT 'Data quality rule passed: No NULL values in DIMCustomer.Name';
END
```

73 %

Messages

Data quality rule passed: No NULL values in DIMCustomer.Name

Completion time: 2023-09-08T15:01:32.7494558+01:00

**Test 2 :** Nous avons évalué les performances d'une requête spécifique en activant les options de statistiques SQL Server. Ce script mesure le temps d'exécution de la requête et les performances d'E/S. L'option SET STATISTICS TIME ON enregistre le temps d'exécution de la requête, tandis que SET STATISTICS IO ON surveille le nombre de lectures logiques et physiques effectuées lors de la requête. Ces informations aident les administrateurs de bases de données et les développeurs à optimiser les requêtes pour de meilleures performances.

```
-- Measure query performance
SET STATISTICS TIME ON;
SET STATISTICS IO ON;

-- Your query here (e.g., a complex SELECT query)
SELECT *
FROM FactOrder
WHERE TotalAmount > 1000;

SET STATISTICS TIME OFF;
SET STATISTICS IO OFF;
```

73 %

Results Messages

	OrderID	DateOrderID	TotalAmount	ShippingMethodID	StatusID	CustomerID	ProductID	Price	Weight	DimensionsID
1	2	346	1560,72	3	1	2	2	534,75	40,74	404
2	3	788	1876,08	2	4	3	3	259,52	11,76	5343
3	6	891	3764,04	1	4	6	6	509,78	25	1778
4	7	797	1916,16	3	2	7	3	511,68	28,1	2293
5	8	120	1661,85	3	1	8	8	85,79	4,69	5534
6	9	409	2305,72	2	1	9	9	43,71	33,91	4274
7	10	682	4573,25	3	1	10	10	63,67	28,14	5906
8	11	110	3996,36	2	1	11	11	514,67	24,37	3253
9	13	962	2272,86	1	4	13	13	899,24	12,85	3266
10	14	475	2603,36	1	4	14	3	433,85	42,68	5265
11	15	1064	4143,2	1	2	15	1	47,35	36,32	9112
12	16	469	2313,66	1	2	16	16	120,26	7,19	2867
13	21	215	3661,15	1	1	21	21	541,9	12,54	3429
14	22	604	1422,75	3	4	22	3	146,18	5,01	2671

# VII . CONCLUSION

En conclusion, le projet d'optimisation de la gestion des données et de l'analyse des ventes de XYZ Corp représente une réponse stratégique aux défis croissants auxquels l'entreprise est confrontée dans le secteur du e-commerce en France. Grâce à la mise en place d'infrastructures modernes de gestion des données, de processus ETL efficaces avec Talend, et la création d'un entrepôt de données, XYZ Corp est bien positionnée pour gérer la croissance de sa clientèle et de son catalogue de produits de manière plus efficiente.

L'optimisation de la gestion de l'inventaire et l'amélioration de l'analyse des ventes devraient contribuer à renforcer la compétitivité de l'entreprise sur le marché, en lui permettant de prendre des décisions plus éclairées et de répondre de manière plus proactive aux besoins de ses clients. Ce projet est une étape significative dans la poursuite de l'excellence opérationnelle de XYZ Corp dans le domaine du e-commerce en France, tout en garantissant une expérience client de haute qualité.