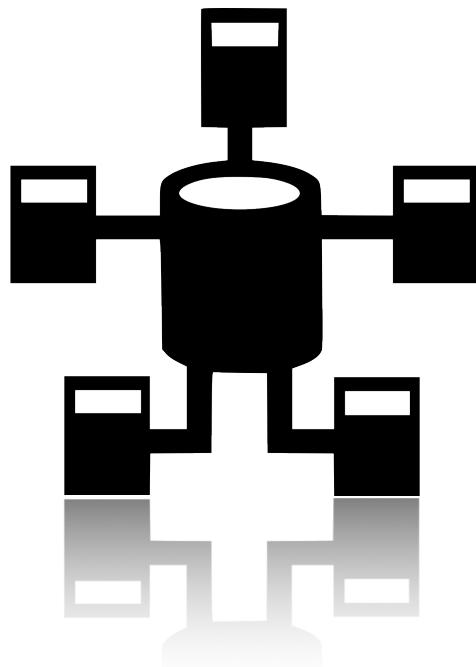




INSTITUT NATIONAL DE STATISTIQUE ET D'ÉCONOMIE APPLIQUÉE

DataWarehouse

Rapport du projet



Réalisée par :

Habiba Ezzagrani

Encadrée par :

Mme Imane Hilal

Année Universitaire:

2022-2023

Table de Matières

Introduction	3
CHAPITRE 1: DataWarehouse	4
Le but de DataWarehouse :	4
BD vs DataWarehouse :	4
ETL :	5
CHAPITRE 2 : Talend	6
Historique :	6
Définition :	6
Pourquoi Talend ?	7
CHAPITRE 3 : Outils Utilisées	7
TALEND :	7
SqlServer :	8
Docker :	8
Excel :	8
Azure Data studio :	9
Kaggle :	9
CHAPITRE 4 : Conception	9
Source :	9
Schéma en étoile :	10
La table de fait :	10
CHAPITRE 4: Implémentation	11
Intégration des données :	11
Extraction et le Mapping :	12
CONCLUSION	16

Introduction

L'entreposage de données est le stockage électronique sécurisé d'informations par une entreprise ou une autre organisation. Toutes les entreprises modernes ont à leur disposition des données, ces données sont stockées dans différents endroits. L'objectif de l'entreposage de données est de créer un trésor de données historiques qui peuvent être récupérées et analysées pour fournir un aperçu utile des opérations de l'organisation. L'entreposage de données est un élément essentiel de l'intelligence d'affaires. Les gestionnaires doivent avoir accès à des informations actuelles et historiques pour remplir leurs fonctions de gestion. Ces données sont stockées dans différents endroits et pour gérer ces derniers on aura besoin des mécanismes pour les réaliser. Pour cela, on a choisi l'ETL comme mécanisme. L'entrepôt de données, ou le DataWarehouse, est une collection de données provenant de toutes les sources de l'entreprise dans un lieu unique. L'accès à des informations fiables et précises est essentiel pour la gestion des Bibliothèques.

CHAPITRE 1: DataWarehouse

Le but de DataWarehouse :

Un entrepôt de données est un endroit pratique pour créer et stocker des métadonnées Améliorer la qualité des données en nettoyant les données lors de leur importation dans l'entrepôt de données (fournissant des données plus précises) et en fournissant des codes et des descriptions cohérents Les rapports utilisant l'entrepôt de données ne seront pas affectés par nouvelles versions de logiciels d'application.

BD vs DataWarehouse :

Les bases de données fournissent des données en temps réel, tandis que les entrepôts stockent les données auxquelles accéder pour les grandes requêtes analytiques. L'entrepôt de données est un exemple de système OLAP ou de système de réponse aux requêtes de base de données en ligne. OLTP est un système de modification de base de données en ligne, par exemple, ATM.Les différences sont remarquables au niveau :

Stockage vs analyse : Une base de données est conçue principalement pour enregistrer des données. Un entrepôt de données, d'autre part, est conçu principalement pour analyser les données. Une base de données est normalement optimisée pour effectuer des opérations de lecture-écriture de transactions ponctuelles. Il n'est pas conçu pour effectuer de grandes requêtes analytiques de la même manière qu'un entrepôt de données.

Collecte vs catégorie : Alors qu'une base de données est une collecte de données axée sur les applications, un entrepôt de données est plutôt axé sur une catégorie de données. Une base de données est normalement limitée à une seule application, ce qui signifie qu'une base de données équivaut

habituellement à une application ; elle cible habituellement un processus à la fois. Un entrepôt de données, d'autre part, stocke les données d'un nombre quelconque d'applications. Un entrepôt de données comprend un nombre infini d'applications et cible autant de processus que nécessaire.

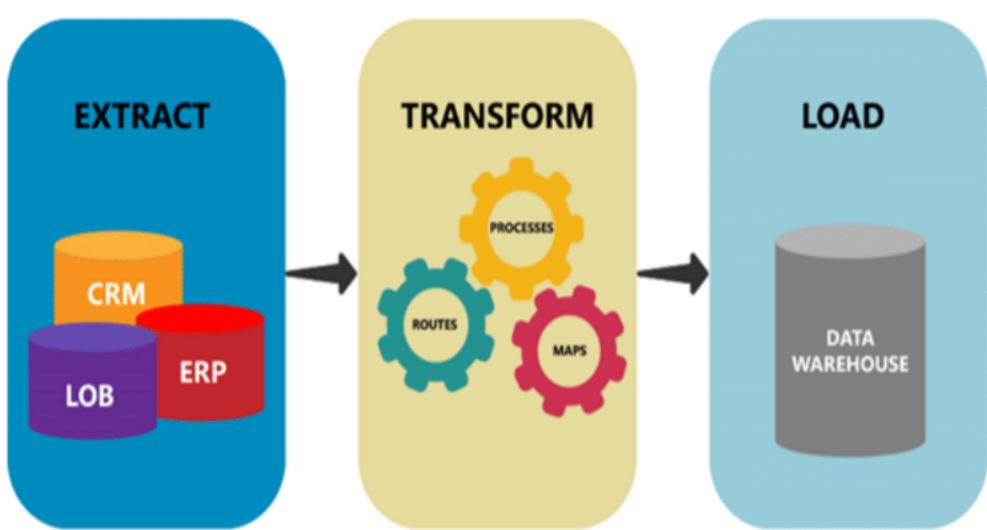
Fournisseur de données vs source d'analyse de données : L'une des différences pratiques entre une base de données et un entrepôt de données est que le premier est un fournisseur de données en temps réel, tandis que le second est davantage une source d'analyse des données à mesure qu'elles sont enregistrées. Toutes les données peuvent être extraites d'un entrepôt de données pour être analysées chaque fois que cela est nécessaire.

Rapidité de stockage vs temps d'analyse : Une base de données comporte généralement des tables complexes parce que les données sont organisées de telle sorte qu'aucun élément n'est dupliqué. Cette structure organisationnelle permet un traitement et un stockage très efficaces des données; une réponse est très rapide. Un entrepôt de données, par contre, n'est pas conçu pour des transactions rapides, mais plutôt pour améliorer les requêtes analytiques, ce qui est obtenu en utilisant moins de tables et une structure plus simple.

ETL :

L'ETL est un type d'intégration de données qui fait référence aux trois étapes (extraction, transformation, load) utilisé pour mélanger des données provenant de plusieurs sources. Il est souvent utilisé pour construire un entrepôt de données.

Au cours de ce processus, les données sont prises (extraites) d'un système source, converties (transformé) dans un format qui peut être analysé, et stockées (chargé) dans un entrepôt de données ou autre système. L'extraction, la charge, la transformation (ELT) est une approche alternative mais connexe conçue pour pousser le traitement vers la base de données afin d'améliorer les performances.



CHAPITRE 2 : Talend

Historique :

Talend (Prononciation : TAL-end) est un éditeur de logiciel spécialisé dans l'intégration de données. La société a été créée en 2005 à Suresnes², et dispose d'un siège administratif³ à Redwood City (Californie) et de filiales en Amérique du Nord, en Europe et en Asie, ainsi qu'un réseau mondial de partenaires techniques et de service⁴. Le 20 mars 2021, Talend est racheté par la société américaine de capital-investissement Thoma Bravo LP pour 2,4 milliards de dollars⁵.

Définition :

Talend est une plateforme d'intégration logicielle open source qui vous aide à transformer sans effort ces données en informations commerciales. La demande toujours croissante de Talend Certification aujourd'hui est la preuve de sa valeur sur le marché. A travers ce blog sur ce qu'est Talend, je vais vous présenter Talend ETL Tool.

Pourquoi Talend ?

Talend vous aide à mettre des données en bonne santé à la disposition de toute votre organisation pour changer la manière de prendre des décisions. Appuyez-vous sur une plateforme unifiée, capable de répondre à vos besoins en matière de données, quelle que soit l'échelle ou la complexité, pour améliorer vos résultats.

CHAPITRE 3 : Outils Utilisées

TALEND :

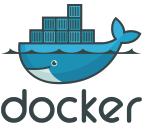
Talend est un outil de gestion de code pour les applications open source. Il propose divers logiciels et services de traitement et de gestion des données, l'intégration dans le stockage en nuage des applications d'entreprise, la qualité des données et le Big Data. Le premier fournisseur open source commercial d'applications d'intégration de données a été Talend, qui a été lancé sur le marché en 2005. Il s'agit de Talend open studio, maintenant connu sous le nom de Talend Open Studio for data intégration, qui a été publié par Talend en octobre 2006. Depuis, un nombre de marchandises ont été libérées et sont utilisées très favorablement sur le marché. Il est considéré comme la plate-forme d'intégration cloud et Big Data leader de la prochaine génération. Cela aide les entreprises à prendre des décisions en temps réel et est davantage alimenté par les résultats. Il s'agit principalement d'un outil ETL qui vous permet de gérer facilement les étapes impliquées dans le processus ETL, de la configuration du travail à l'exécution du chargement des données ETL du système cible. Pour réaliser le mappage entre le périphérique source et le périphérique cible, vous

pouvez utiliser l'interface utilisateur graphique de Talend Open Studio pour faire glisser et déposer le composant souhaité depuis la palette.

SqlServer :

 Microsoft SQL Server est un système de gestion de base de données relationnelle développé par Microsoft. En tant que serveur de base de données, il s'agit d'un produit logiciel dont la fonction principale est de stocker et de récupérer les données demandées par d'autres applications logicielles, qui peuvent s'exécuter soit sur le même ordinateur, soit sur un autre ordinateur sur un réseau (y compris Internet). Microsoft commercialise au moins une douzaine d'éditions différentes de Microsoft SQL Server, destinées à différents publics et pour des charges de travail allant de petites applications mono-machine à de grandes applications Internet avec de nombreux utilisateurs simultanés.

Docker :

 Docker est un ensemble de produits de plate-forme en tant que service (PaaS) qui utilisent la virtualisation au niveau du système d'exploitation pour fournir des logiciels dans des packages appelés conteneurs. Le service a des niveaux gratuits et premium. Le logiciel qui héberge les conteneurs s'appelle Docker Engine. Il a été lancé pour la première fois en 2013 et est développé par Docker.

Excel :

 Microsoft Excel est une feuille de calcul développée par Microsoft pour Windows, macOS, Android et iOS. Il comporte des capacités de calcul ou de calcul, des outils graphiques, des tableaux croisés dynamiques et un langage de programmation macro appelé Visual Basic pour Applications (VBA). Excel fait partie de la suite logicielle Microsoft Office.

Azure Data studio :



Microsoft Azure Data Studio est un outil multiplateforme gratuit qui peut être utilisé pour gérer SQL Server, Azure SQL Database et Azure SQL Data Warehouse. Azure Data Studio offre désormais un moyen de créer des visualisations rapides pour vos données. Cette extension est utile lorsque vous essayez d'examiner les données et de comprendre ce qui se passe. Nous utilisons une technologie appelée **SandDance** de Microsoft Research, qui peut générer des visualisations sur place des données.

Kaggle :



Kaggle, une filiale de Google LLC, est une communauté en ligne de scientifiques des données et de praticiens de l'apprentissage automatique. Kaggle permet aux utilisateurs de trouver et de publier des ensembles de données, d'explorer et de créer des modèles dans un environnement de science des données basé sur le Web, de travailler avec d'autres scientifiques des données et des ingénieurs en apprentissage automatique, et de participer à des concours pour résoudre les défis de la science des données.

CHAPITRE 4 : Conception

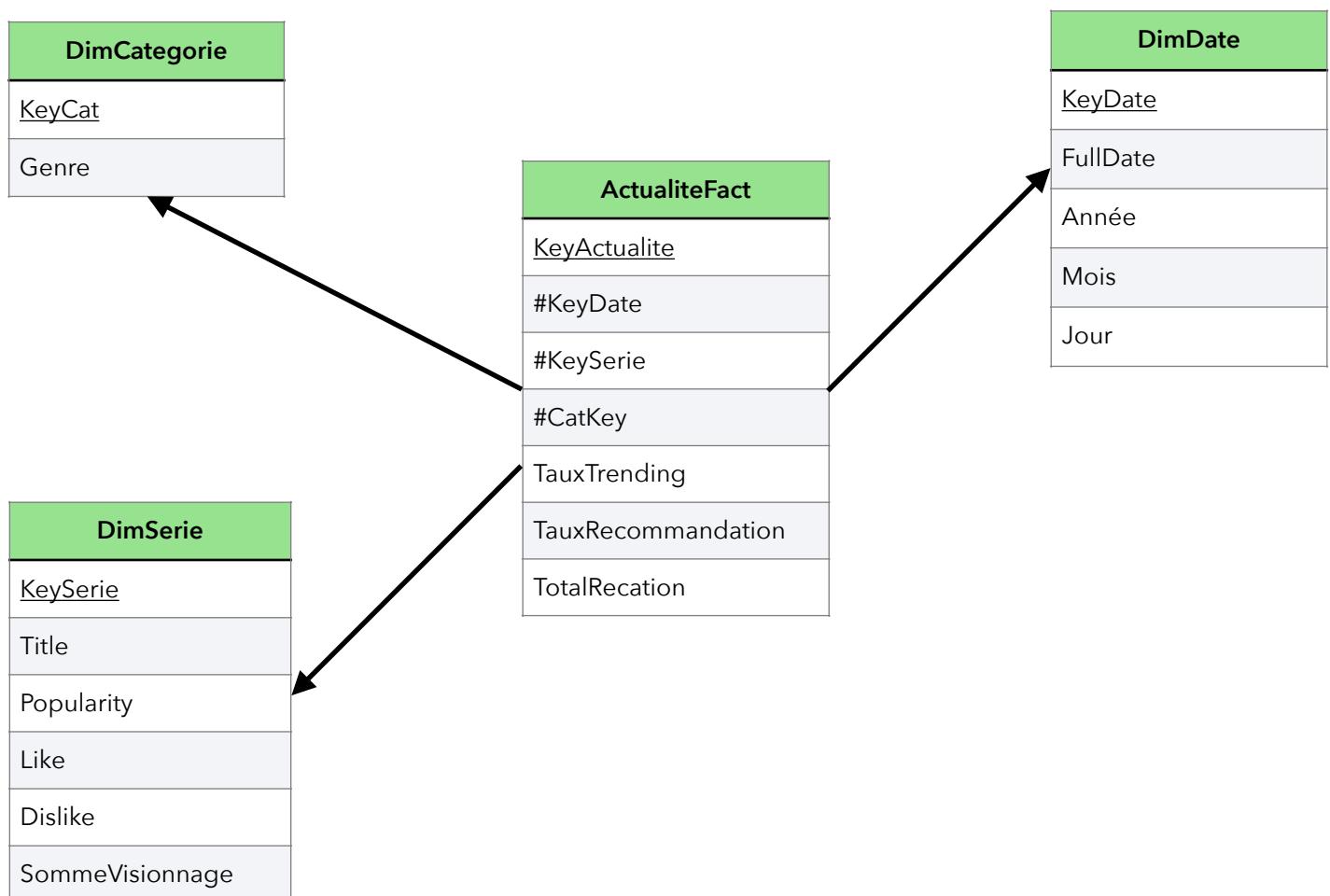
Source :

Je me suis servie de Kaggle pour collecter my DataSet , cette DataSet est lié aux séries originales sur Netflix .

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1 title	release_year	score	number_of_main_genre	main_genre	Rating	ratingLevel		original_lan	popularity	SommTauVisionnage	DateOfTrending	nombreLike	nombreDislike	
2 Moon: Python's Flying Circus	1969	8.8	4	comedy	GB	18+	0	English	10.104	10/1/2022	17890	1950		
3 Knight Rider	1982	8.9	4	sci-fi	US	18+	0	English	46.933	10/1/2022	7579	678		
4 Seinfeld	1989	8.9	9	comedy	US	18+	0	Parents strongly cautioned. May be unsuitable for children ages 14 and under.	95.667	01/05/2019	9865	876		
5 Star Trek: Deep Space Nine	1993	8.1	7	sci-fi	US	16+	1	Parents strongly cautioned. May be unsuitable for children ages 14 and under.	1.511.996	7/22/2020	12345	9987		
6 Neon Genesis Evangelion	1995	8.5	1	sci-fi	JP	16+	1	Parental guidance suggested. May not be suitable for all children.	195.038	20/08/2020	6789	1322		
7 StarGate SG-1	1997	8.4	10	sci-fi	US	17+	1	Parental guidance suggested. May not be suitable for children ages 14 and under.	100.235	3/9/2021	5938	567		
8 Cowboy Bebop	1999	8.9	1	western	JP	16+	1	For mature audiences. May not be suitable for children 17 and under.	117.893	4/04/2014	547	987		
9 One Piece	1999	8.8	21	action	JP	18+	0	Parents strongly cautioned. May be unsuitable for children ages 14 and under.	103.762	4/42/2014	9876	654		
10 Gilmore Girls	2000	8.2	8	comedy	US	16+	0	For mature audiences. May not be suitable for children 17 and under.	101.767	1/15/2016	34567	432		
11 Trainwreck	2001	8.6	12	comedy	CA	16+	1	Parental guidance suggested. May not be suitable for children ages 14 and under.	89.893	1/30/2022	5678	190		
12 House	2004	8.4	6	comedy	US	18+	1	Parental guidance suggested. May not be suitable for all children.	71.587	3/20/2019	56789	3950		
13 Chappelle's Show	2005	8.8	3	comedy	US	18+	0	Parents strongly cautioned. May be unsuitable for children ages 14 and under.	71.587	27/1/2020	7890	7893		
14 The 4400	2006	7.3	4	sci-fi	US	16+	0	Parents strongly cautioned. May be unsuitable for children ages 14 and under.	71.587	16/09/2019	12358	345		
15 Avatar: The Last Airbender	2005	9.3	3	sci-fi	US	18+	0	Parents strongly cautioned. May be unsuitable for children ages 14 and under.	280.696	1/18/2020	34567	876		
16 DARTH NOTE	2005	9.1	1	sci-fi	JP	18+	1	Language and violence.	138.869	2/3/2018	8765	654		
17 Game of Thrones	2011	8.7	6	comedy	US	18+	0	Language and violence.	280.696	1/4/2018	45758	1714		
18 Breaking Bad	2008	9.5	5	drama	US	16+	0	Parents strongly cautioned. May be unsuitable for children ages 14 and under.	10.638	127	01/01/2020	3456	763	
19 Community	2009	8.5	6	comedy	US	16+	0	For mature audiences. May not be suitable for children 17 and under.	59.795	1609	2/8/2021	98765	7864	
20 Don't Be a Hater	2010	8.7	6	drama	GB	18+	1	Parents strongly cautioned. May be unsuitable for children ages 14 and under.	71.587	3/23/2015	7654	678		
21 House of Cards	2013	8.7	3	drama	JP	18+	1	This movie has not been rated.	71.587	3/20/2019	9876	989		
22 Call the Midwife	2012	8.5	11	drama	GB	18+	0	Suitable for all ages.	53.144	10/10/2022	45675	786		
23 Attack on Titan	2013	8.7	4	sci-fi	JP	18+	0	Suitable for all ages.	10.34	761	19/9/2021	3764	523	
24 BoJack Horseman	2014	8.8	6	drama	US	18+	0	This movie has not been rated.	10.34	820	10/02/2016	67644	653	
25 Better Call Saul	2015	8.8	6	comedy	US	7+	0	This movie has not been rated.	277.186	18/1	12/1/2017	9877	345	
26 Arrested Development	2003	8.7	5	sci-fi	US	7+	0	No mature audiences. May not be suitable for children 17 and under.	48.446	640	1/10/2017	45675	111	
27 Arrested Development	2017	8.7	3	drama	CA	16+	0	Suitable for all ages.	64.215	299	04/01/2017	45783	100	
28 Cobra Kai	2018	8.6	5	action	US	18+	0	For mature audiences. May not be suitable for children 17 and under.	13.838	147	05/07/2018	127890	7890	
29 Koto Factory	2019	9.3	2	drama	IN	16+	1	Parental guidance suggested. May not be suitable for all children.	71.952	556	02/05/2020	45670	678	
30 The Last Dance	2019	8.5	1	documentary	US	18+	0	Parents strongly cautioned. May be unsuitable for children ages 14 and under.	18.331	331	01/11/2021	67545	789	
31 Arcane	2021	9.1	1	drama	US	18+	0	Parents strongly cautioned. May be unsuitable for children ages 14 and under.	32.411	400	12/09/2021	1710	345	
32 Heartstopper	2022	8.9	1	drama	GB	18+	1	For mature audiences. May not be suitable for children 17 and under.	64.764	1064	01/11/2021	56789	564	

Schéma en étoile :

Le modèle de données en étoile. Le modèle de données en étoile doit son nom à sa forme. Ce modèle de conception privilégie l'approche utilisateur, l'orientation métier. La table de référence contient les faits. Les faits ou mesures sont les données chiffrées (du type résultats par secteur). Les tables satellites correspondent aux dimensions. Le schéma suivant Contient nos dimensions plus d'une table de fait qui contient tous les Id pour les rassembler.



La table de fait :

La table de fait a une vision précise quand elle est lié avec les dimensions , cette table de fait essaie de nous montrer les séries les plus regardées et recommandées parmi les spectateurs sur Netflix dans certaines

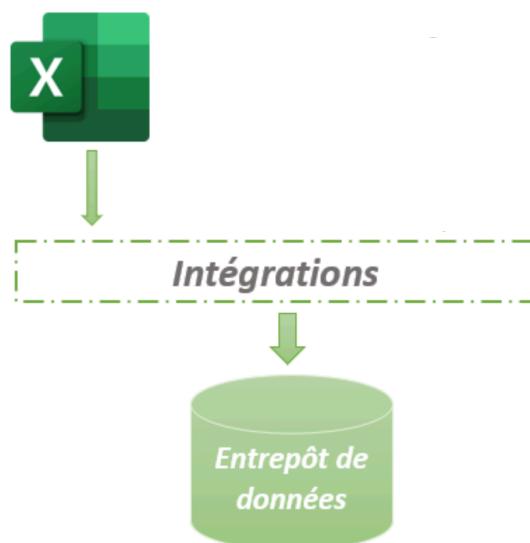
catégories dans une date précise. Pour bien comprendre ce que cette table de fait vise , **des questions rhétoriques** s'impose :

- ✓ Quelles sont les séries qui ont beaucoup de réaction d'après les Watchers ?
- ✓ Quelles sont les séries les plus recommandées ?
- ✓ Quelles sont les séries en tendance dans une période précise ?

CHAPITRE 4: Implémentation

Intégration des données :

Dans cette partie ,j'ai récupéré les données **INPUT** d'Excel et les stocker dans une autre base de donnée **OUTPUT** dans 'SqlServer' appelée **netflix** :



La Connexion avec SqlServer :

La connexion est faite dans la métadonnées/Connexions de bases de données. Puis , j'ai importer les dimensions après les mapper automatiquement à travers TMAP vers la base de données OUTPUT .

Fichier Excel :

L'importation du fichier Excel source est faite dans la métadonnées/Fichier Excel .

Extraction et le Mapping :

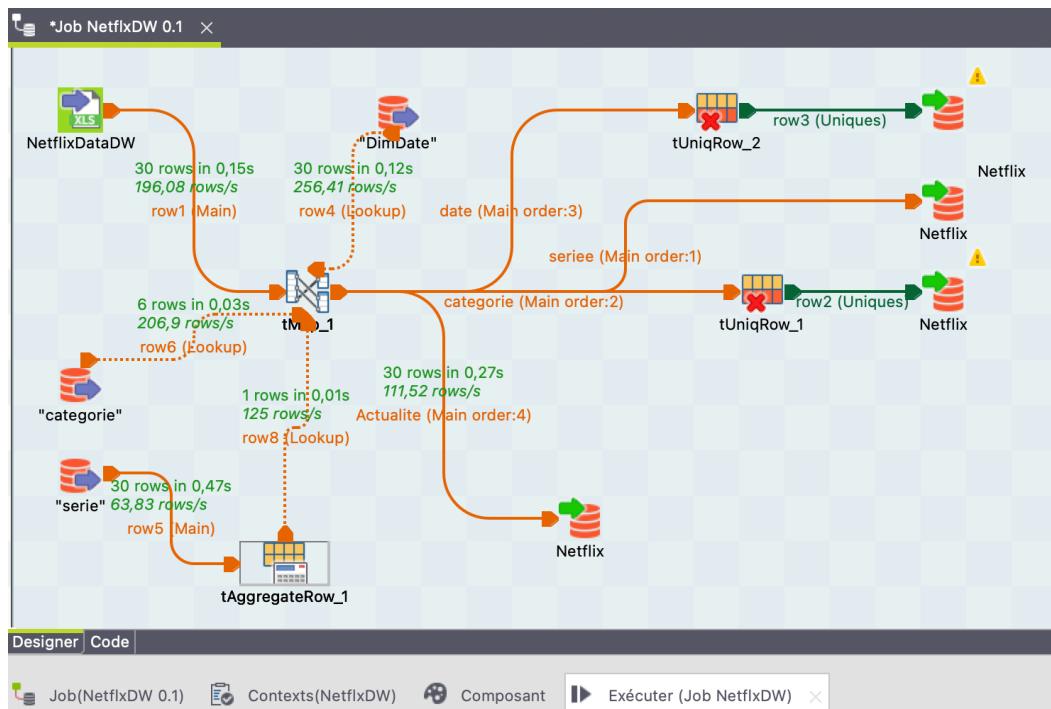
Tous le travail est faite dans le job .



Le Job :

La transformation des données source dans le fichier Excel vers la bd output 'Netflix' est faite à travers tMap , et pour éviter les doublons , on utilise tUniqueRow , puis on repose les dimensions comme des données INPUT pour mapper la table de fait .le tAggregateRow est utilisé pour calculer la somme d'une colonne pour le calcul des taux .

The screenshot shows the Talend Designer interface with the 'Composant' tab selected. A 'tAggregateRow_1' component is currently configured. The 'Paramètres simples' (Simple Parameters) panel on the left lists 'Paramètres simples', 'Paramètres avancés', 'Paramètres dynamiques', 'Vue', and 'Documentation'. The main configuration area has two tabs: 'Schéma' (Schema) and 'Opérations' (Operations). In the 'Schéma' tab, 'Group by' is set to 'Colonne de sortie' (Output Column) and 'Position de la colonne d'entrée' (Input Column Position) is empty. Below this are buttons for adding (+), removing (-), moving up (^), and moving down (^). In the 'Opérations' tab, there are two rows of operations. The first row has 'Colonne de sortie' (popularity, SommeTauxVisionnage), 'Fonction' (somme, somme), and 'Position de la colonne d'entrée' (popularity, SommeTauxVisionnage). The second row has a checked checkbox 'Ignorer les valeurs nulles' (Ignore null values) and a checked checkbox 'Valeurs par défaut' (Default values). Below the operations are buttons for adding (+), removing (-), moving up (^), and moving down (^).



Job NetflixDW

Exécution simple

Exécution en mode Debug

Paramètres avancés

Cible d'exécution

Exécution pour la mémoire

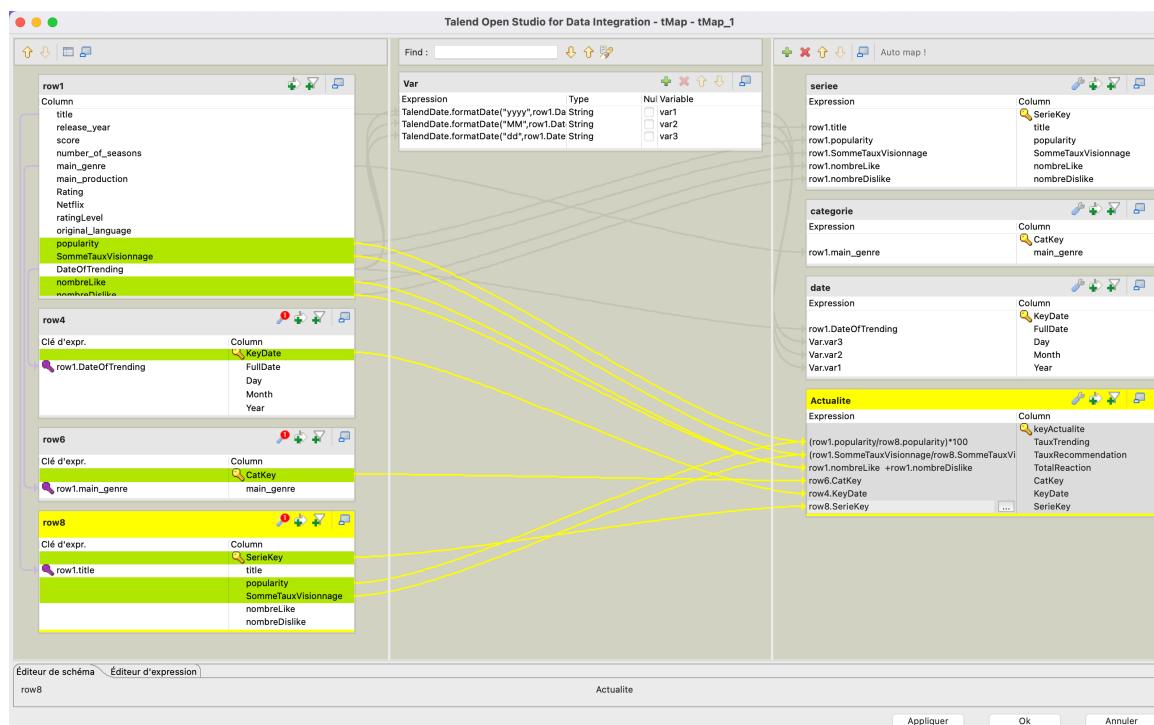
Exécution

Exécuter Arrêter Effacer

```
Démarrage du Job NetflixDW à 16:54 19/01/2023.
[statistics] connecting to socket on port 3901
[statistics] connected
multiple points
[statistics] disconnected

Job NetflixDW terminé à 16:54 19/01/2023. [Code de sortie = 0]
```

Voici le tMap qui montre tous les extractions



Comme on voit dans ce screen dans Azure Studio, les dimensions et la table de faite sont créées sans aucune erreur et voici la table de fait .

Results Messages

keyAc...	TauxTrending	TauxRecom...	TotalReac...	CatKey	KeyDate	SerieKey
1 1	0,801676	4,691558	18890	1	1	1
2 2	12,469678	15,4771	8367	2	2	2
3 3	2,5476327	3,994928	10741	1	3	3
4 4	5,1939034	2,4855633	8111	3	4	4
5 5	2,8290608	5,971768	6245	3	5	5
6 6	3,1420722	6,2085614	1554	4	6	6
7 7	2,7632046	6,7554765	10530	2	7	7
8 8	2,710077	5,068899	34999	1	8	8
9 9	2,3938699	2,2151608	5868	1	9	9
1... 10	1,9063771	4,943628	64679	3	10	10
1... 11	1,9063771	4,943628	15789	1	11	11
1... 12	1,9063771	4,943628	12703	3	12	12
1... 13	7,4749956	1,8087934	35443	3	13	13
1... 14	3,6981113	3,5564792	9419	3	14	14
1... 15	7,482025	2,386263	45912	1	15	15
1... 16	0,2832922	0,19401754	4219	5	16	16
1... 17	1,5923537	2,4580646	106629	1	17	17
1... 18	1,9063771	4,943628	8332	5	18	18
1... 19	1,9063771	4,943628	99574	5	19	19
2... 20	1,3619757	2,253353	46461	5	20	20

netflix

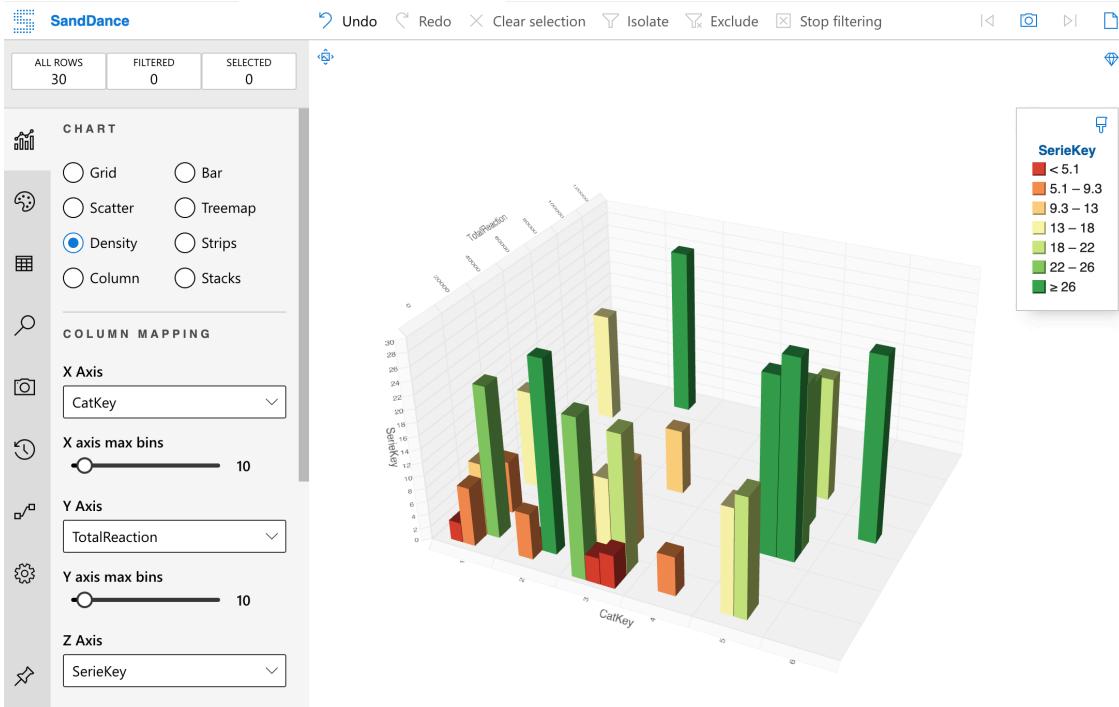
Tables

- > dbo.ActualiteFact
- > dbo.categorie
- > dbo.DimDate
- > dbo.serie
- > Views
- > Synonyms
- > Programmability
- > External Resources
- > Service Broker
- > Storage
- > Security
- > Security
- > Server Objects
- > localhost, <default> (habibaezza...)

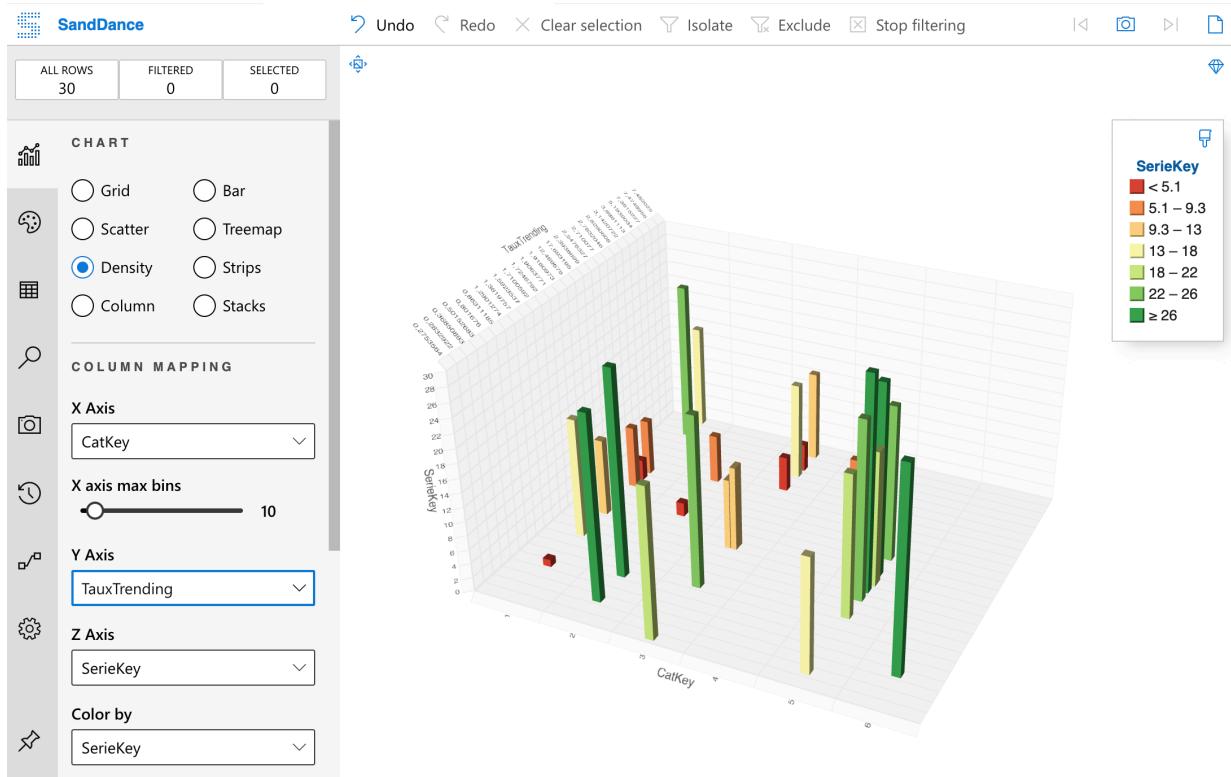
Results grid

J'ai visualisé la DataWarehouse à l'aide de SandDance , ces Charts répondent au questions déjà posées .

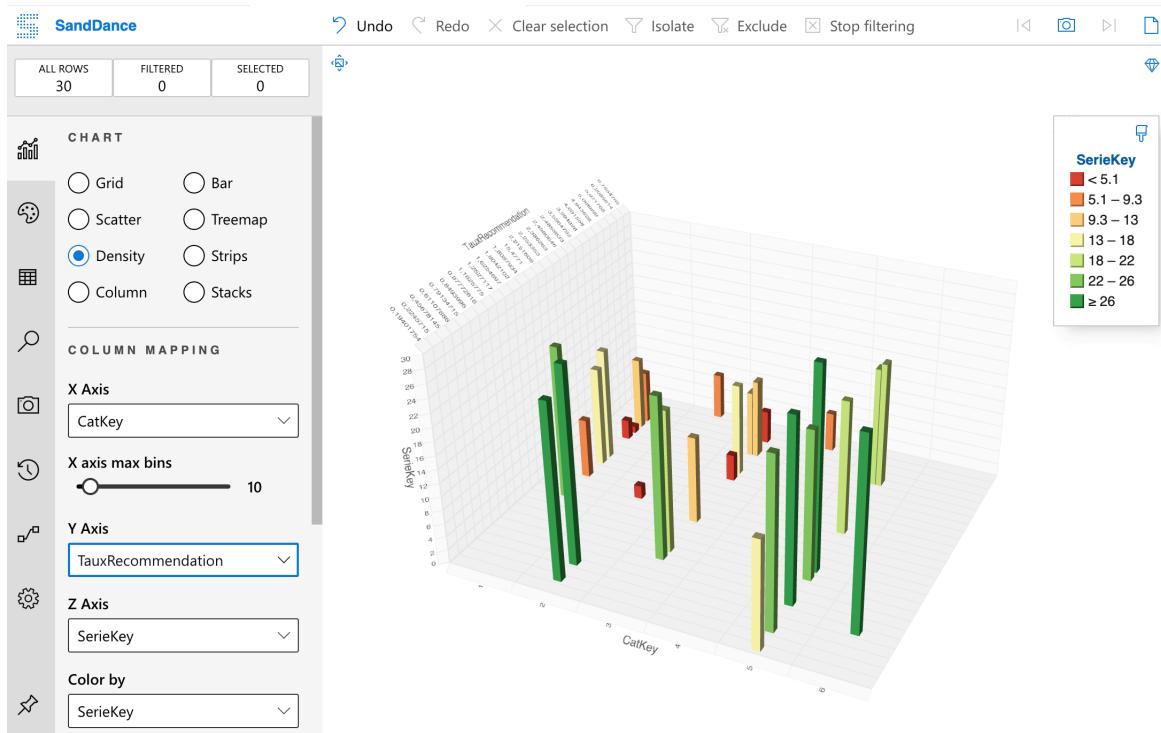
Ce graphe nous montre les séries qui ont eu un grand nombre de réaction d'après les spectateurs dans chaque catégorie



Ce deuxième graphe nous affiche les séries tangence dans des différentes catégories



Ce dernier graphe nous montre les séries les plus recommandées à propos de leur catégories .



CONCLUSION

Tout au long de la préparation de ce projet, nous avons essayé de mettre en pratique les connaissances acquises durant le cours. Comme la totalité des objectifs et fonctionnalités mentionnés sont atteints, d'autres améliorations peuvent être apportées au projet. Dans ce rapport, nous avons présenté la conception et la mise en œuvre d'un entrepôt de données pour les bibliothèques. Il s'agit de la première étape d'un long voyage vers une solution complète d'entrepôt de données.