



PIPELINE ETL POUR LA PLATEFORME COURSERA

Présenté par:

Islem Farhane

Mohamed Aziz Benlazreg

Rym Mathlouthi

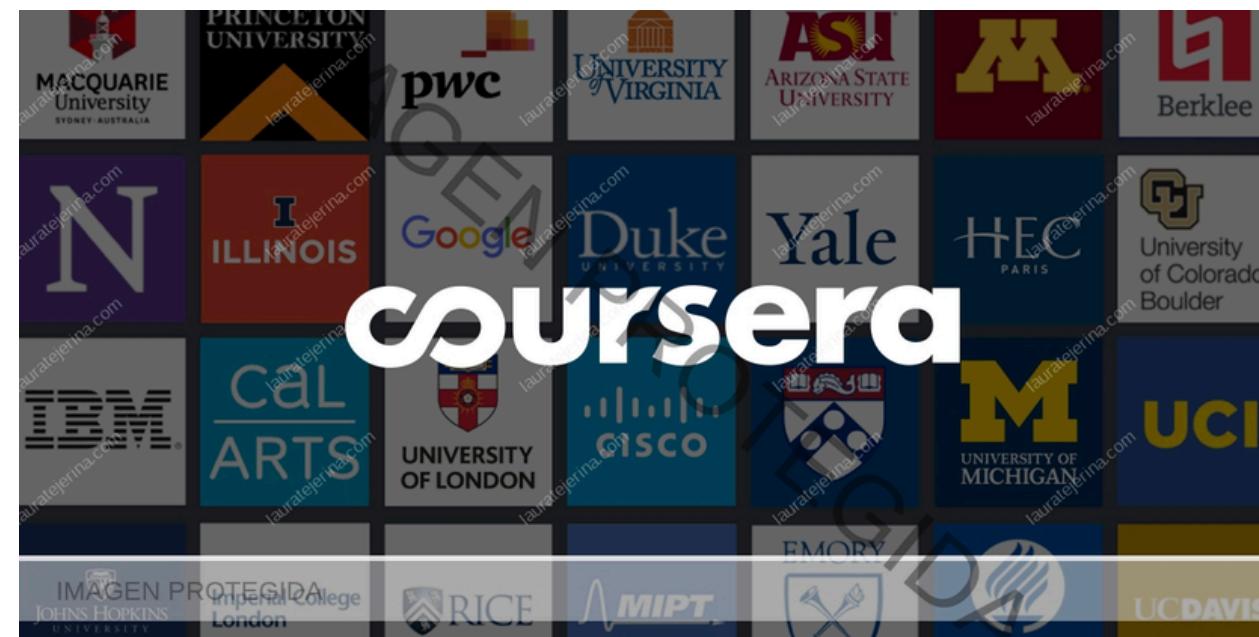
TABLE DES MATIÈRES

- 01** Introduction
- 02** Contexte du projet
- 03** Technologies utilisées
- 04** Expérimentation et résultats
- 05** Analyse des données
- 06** Conclusion

INTRODUCTION

- La capacité à gérer efficacement les données est devenue cruciale pour les entreprises et les organisations.
- Les pipelines ETL jouent un rôle essentiel dans ce processus, permettant d'extraire des données de différentes sources, de les transformer selon les besoins et de les charger dans des entrepôts de données .

CONTEXTE DU PROJET



01

l'organisme à qui s'adresse l'entrepôt

Réalisation d'entrepôts de données à partir des données provenant de la chaîne YouTube et la page Facebook de Coursera.

02

Extract Transform Load (ETL)

- Extraction et prétraitement des données .
- la définition des modèles multidimensionnels.
- Chargement des données dans les entrepôts correspondants.

03

Exploration

Les données sont prêtes pour une analyse approfondie afin de générer des insights précieux sur les interactions des utilisateurs de Coursera sur les plateformes de médias sociaux.

TECHNOLOGIES UTILISÉES



01

Export Comments

En utilisant l'outil "Export Comments", on peut extraire les commentaires des publications YouTube et des Facebook de Coursera de manière simple et efficace.

02

Python

Utiliser Python pour effectuer le prétraitement des données extraites. Cela peut inclure le nettoyage des commentaires , la normalisation des données..

03

Talend

Utiliser Talend pour concevoir et implémenter le processus de chargement des données dans notre entrepôt de données.

04

phpMyAdmin

Utilisez phpMyAdmin pour gérer notre base de données MySQL où on va stocker les données prétraitées.

EXPÉRIMENTATION ET RÉSULTATS

01

Création de base de données :

Modele en étoile :

The screenshot shows the MySQL Workbench interface for a database named 'datalake_model_etoile'. The main window displays a list of tables in a grid format. The columns include 'Table' (with checkboxes), 'Action' (with icons for Parcourir, Structure, Rechercher, Insérer, Vider, Supprimer), 'Lignes' (number of rows), 'Type' (InnoDB), 'Interclassement' (utf8mb4_general_ci), 'Taille' (size), and 'Perte' (lost data). There are six tables listed: dim_date, dim_date_yt, dim_post, dim_video, fait_fb, and fait_yt. Below the table list, it says '6 tables' and 'Somme'. At the bottom, there are buttons for 'Tout cocher' (check all) and 'Avec la sélection' (with selection), and links for 'Imprimer' and 'Dictionnaire de données'.

Table	Action	Lignes	Type	Interclassement	Taille	Perte
dim_date	Parcourir, Structure, Rechercher, Insérer, Vider, Supprimer	0	InnoDB	utf8mb4_general_ci	16,0 kio	-
dim_date_yt	Parcourir, Structure, Rechercher, Insérer, Vider, Supprimer	0	InnoDB	utf8mb4_general_ci	16,0 kio	-
dim_post	Parcourir, Structure, Rechercher, Insérer, Vider, Supprimer	0	InnoDB	utf8mb4_general_ci	16,0 kio	-
dim_video	Parcourir, Structure, Rechercher, Insérer, Vider, Supprimer	0	InnoDB	utf8mb4_general_ci	16,0 kio	-
fait_fb	Parcourir, Structure, Rechercher, Insérer, Vider, Supprimer	0	InnoDB	utf8mb4_general_ci	16,0 kio	-
fait_yt	Parcourir, Structure, Rechercher, Insérer, Vider, Supprimer	0	InnoDB	utf8mb4_general_ci	16,0 kio	-
6 tables	Somme	0	InnoDB	utf8mb4_general_ci	96,0 kio	0 o

EXPÉRIMENTATION ET RÉSULTATS

Modele en constellation :

The screenshot shows the MySQL Workbench interface for a database named "datalake_model_const". The main window displays a list of tables with their details:

Table	Action	Lignes	Type	Interclassement	Taille	Perte
dim_date	Parcourir Structure Rechercher Insérer Vider Supprimer	0	InnoDB	utf8mb4_general_ci	16,0 kio	-
dim_post	Parcourir Structure Rechercher Insérer Vider Supprimer	0	InnoDB	utf8mb4_general_ci	16,0 kio	-
fait_fb	Parcourir Structure Rechercher Insérer Vider Supprimer	0	InnoDB	utf8mb4_general_ci	32,0 kio	-
fait_yt	Parcourir Structure Rechercher Insérer Vider Supprimer	0	InnoDB	utf8mb4_general_ci	32,0 kio	-
video	Parcourir Structure Rechercher Insérer Vider Supprimer	0	InnoDB	utf8mb4_general_ci	16,0 kio	-
5 tables	Somme	0	InnoDB	utf8mb4_general_ci	112,0 kio	0 o

At the bottom, a creation dialog for a new table is open, showing fields for "Nom de table" (with a placeholder) and "Nombre de colonnes" (set to 4), with a "Créer" button.

Modele en galaxie :

Serveur : 127.0.0.1 > Base de données : datalake_model_galaxy

Structure SQL Rechercher Requête Exporter Importer Opérations Privilèges Procédures stockées Évènements

Filtres

Contenant le mot :

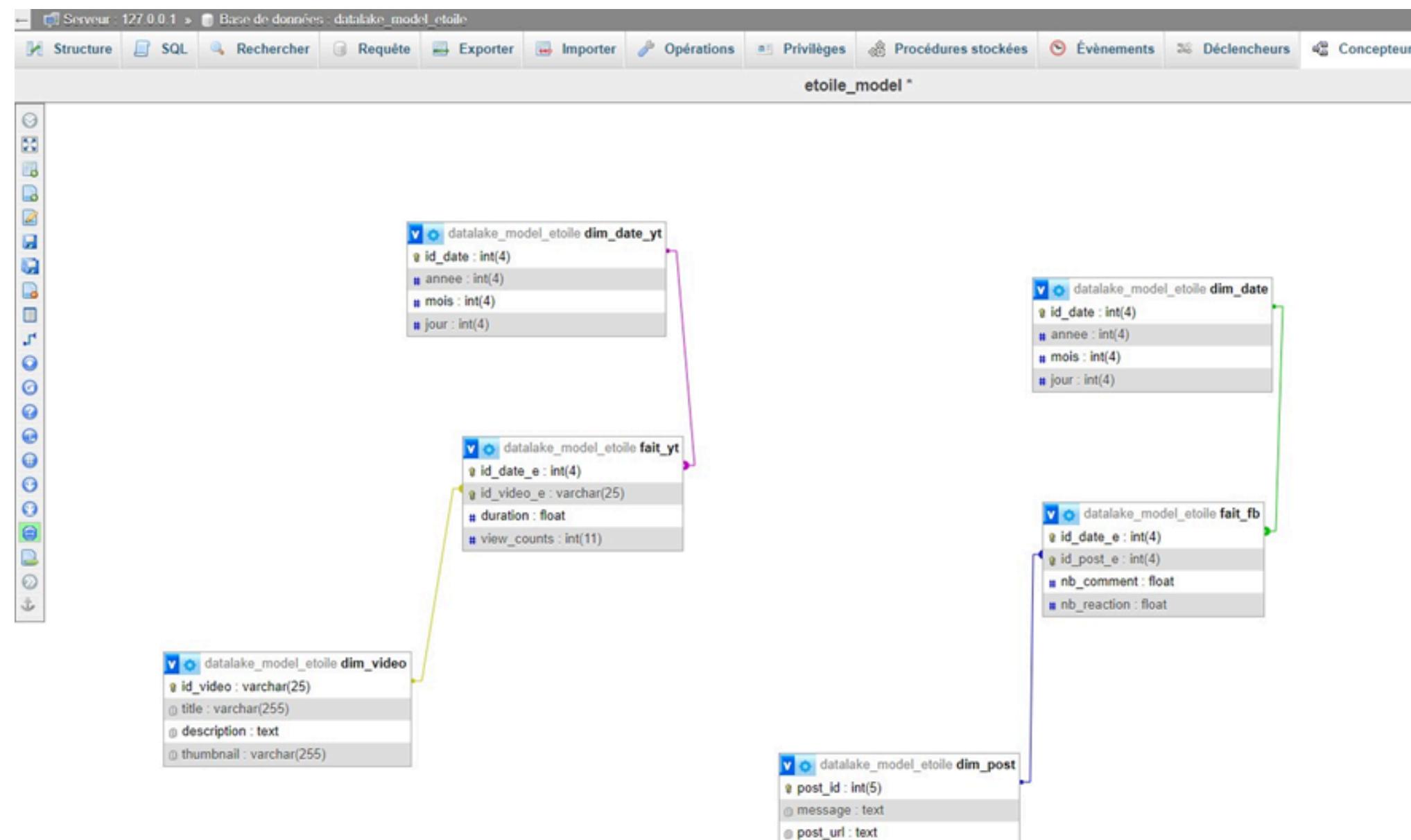
Table	Action	Lignes	Type	Interclassement	Taille	Perte
dim_date	Parcourir Structure Rechercher Insérer Vider Supprimer	0	InnoDB	utf8mb4_general_ci	16,0 kio	-
dim_post	Parcourir Structure Rechercher Insérer Vider Supprimer	0	InnoDB	utf8mb4_general_ci	16,0 kio	-
fait_fb	Parcourir Structure Rechercher Insérer Vider Supprimer	0	InnoDB	utf8mb4_general_ci	16,0 kio	-
fait_yt	Parcourir Structure Rechercher Insérer Vider Supprimer	0	InnoDB	utf8mb4_general_ci	16,0 kio	-
galaxy_fb	Parcourir Structure Rechercher Insérer Vider Supprimer	0	InnoDB	utf8mb4_general_ci	48,0 kio	-
galaxy_yt	Parcourir Structure Rechercher Insérer Vider Supprimer	0	InnoDB	utf8mb4_general_ci	48,0 kio	-
video	Parcourir Structure Rechercher Insérer Vider Supprimer	0	InnoDB	utf8mb4_general_ci	16,0 kio	-
7 tables	Somme	0	InnoDB	utf8mb4_general_ci	176,0 kio	0 o

Avec la sélection :

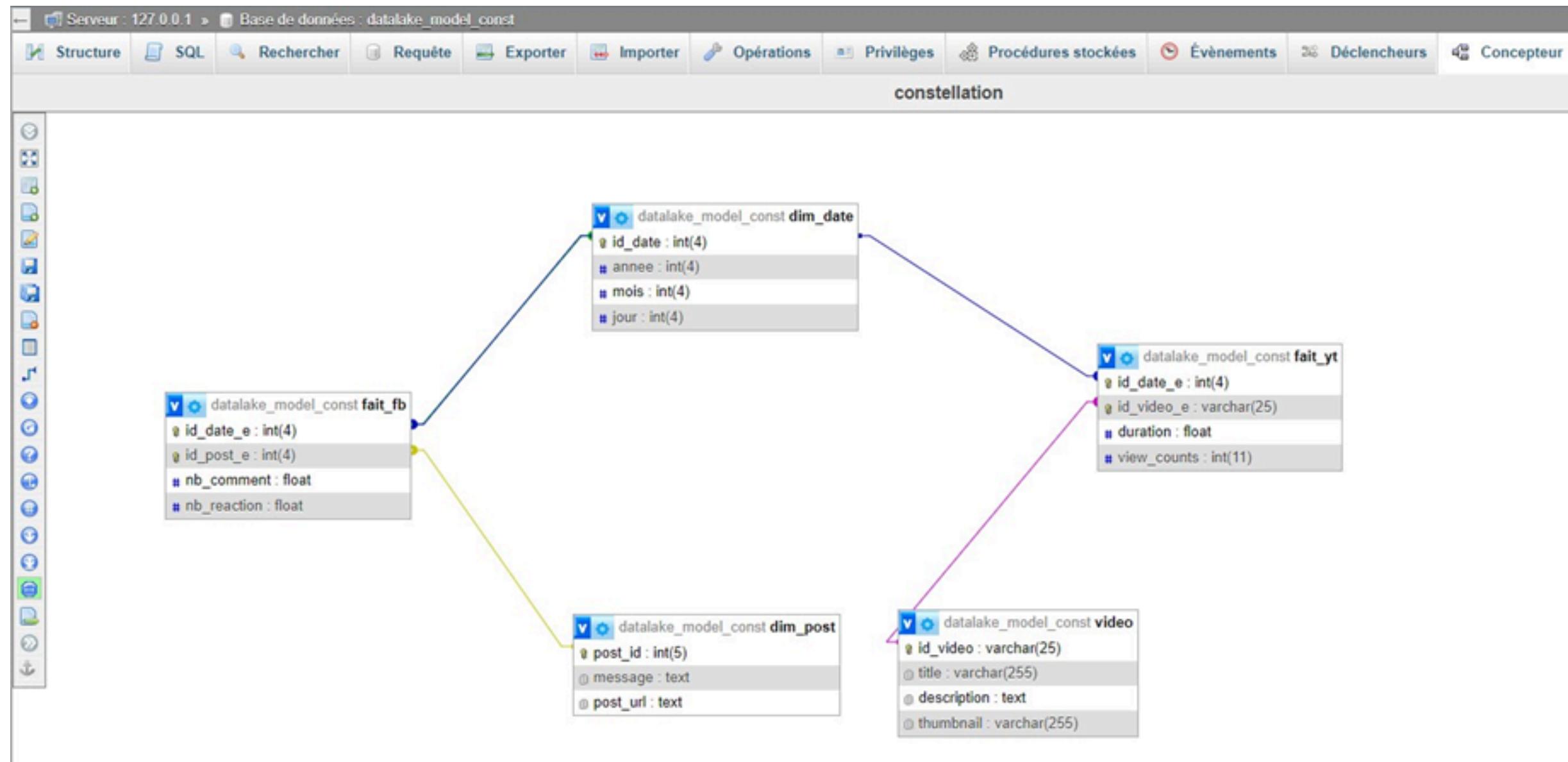
↑ Tout cocher Imprimer Dictionnaire de données

02 Conception des modeles :

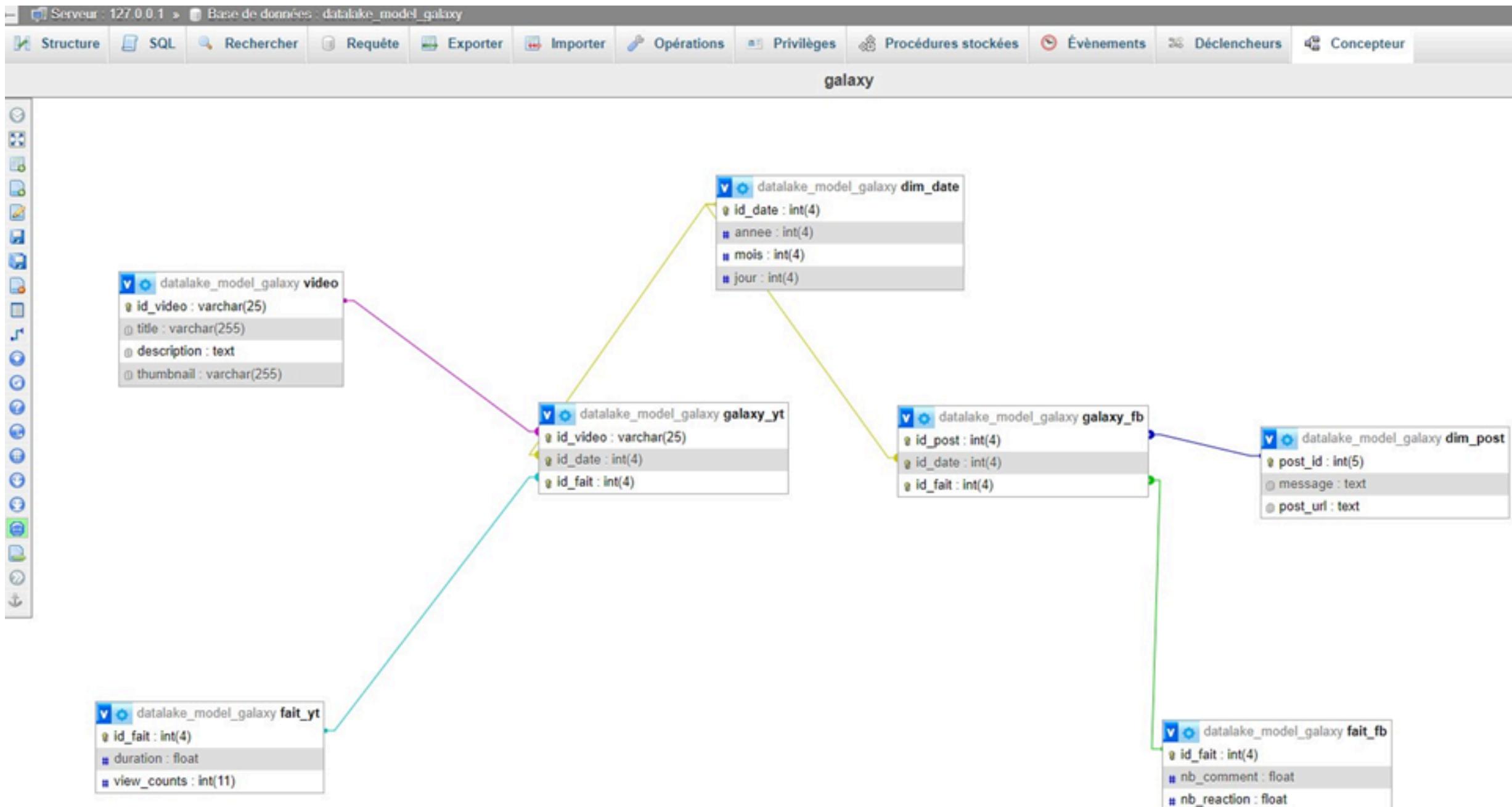
Modele en étoile :



Modele en constellation :

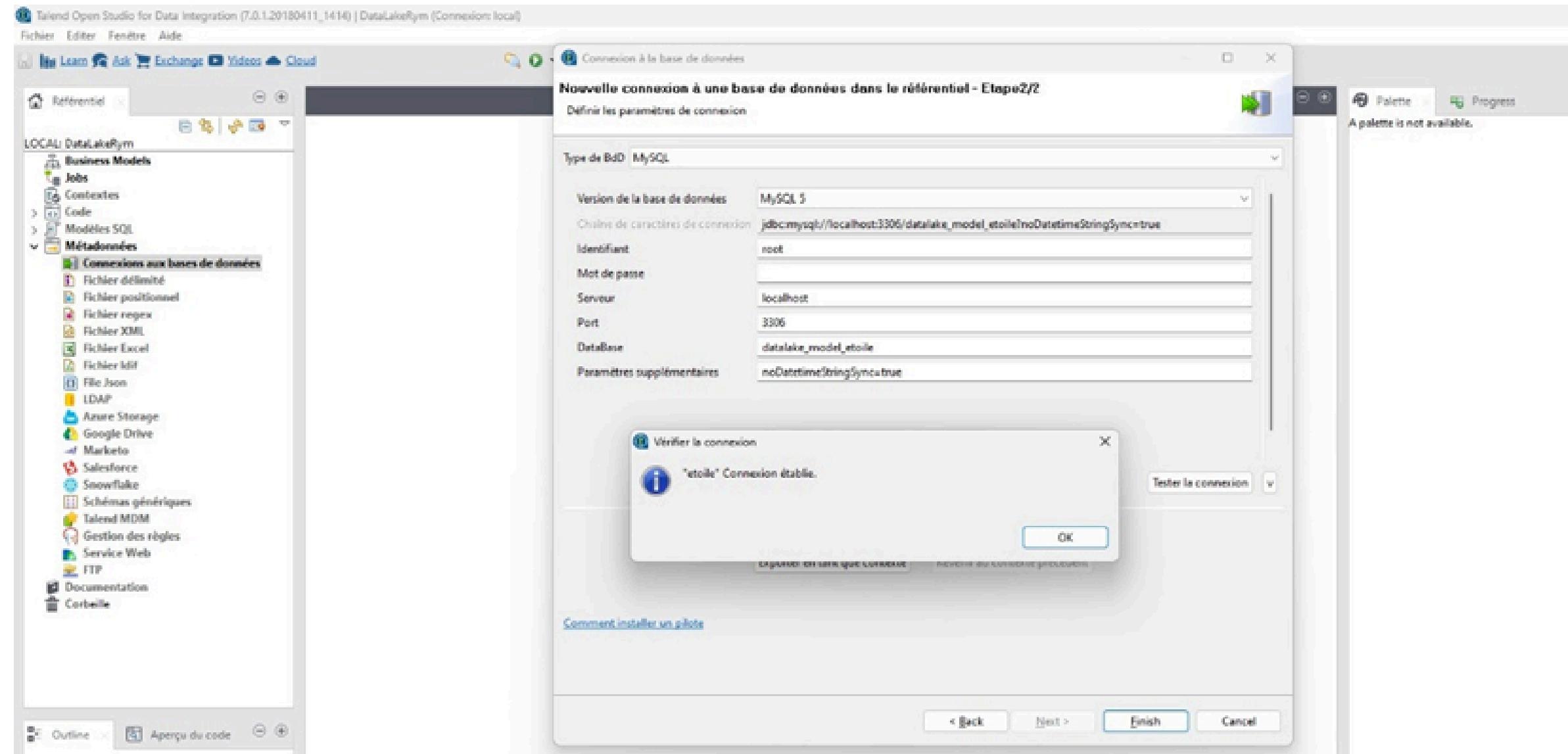


Modele en galaxie :

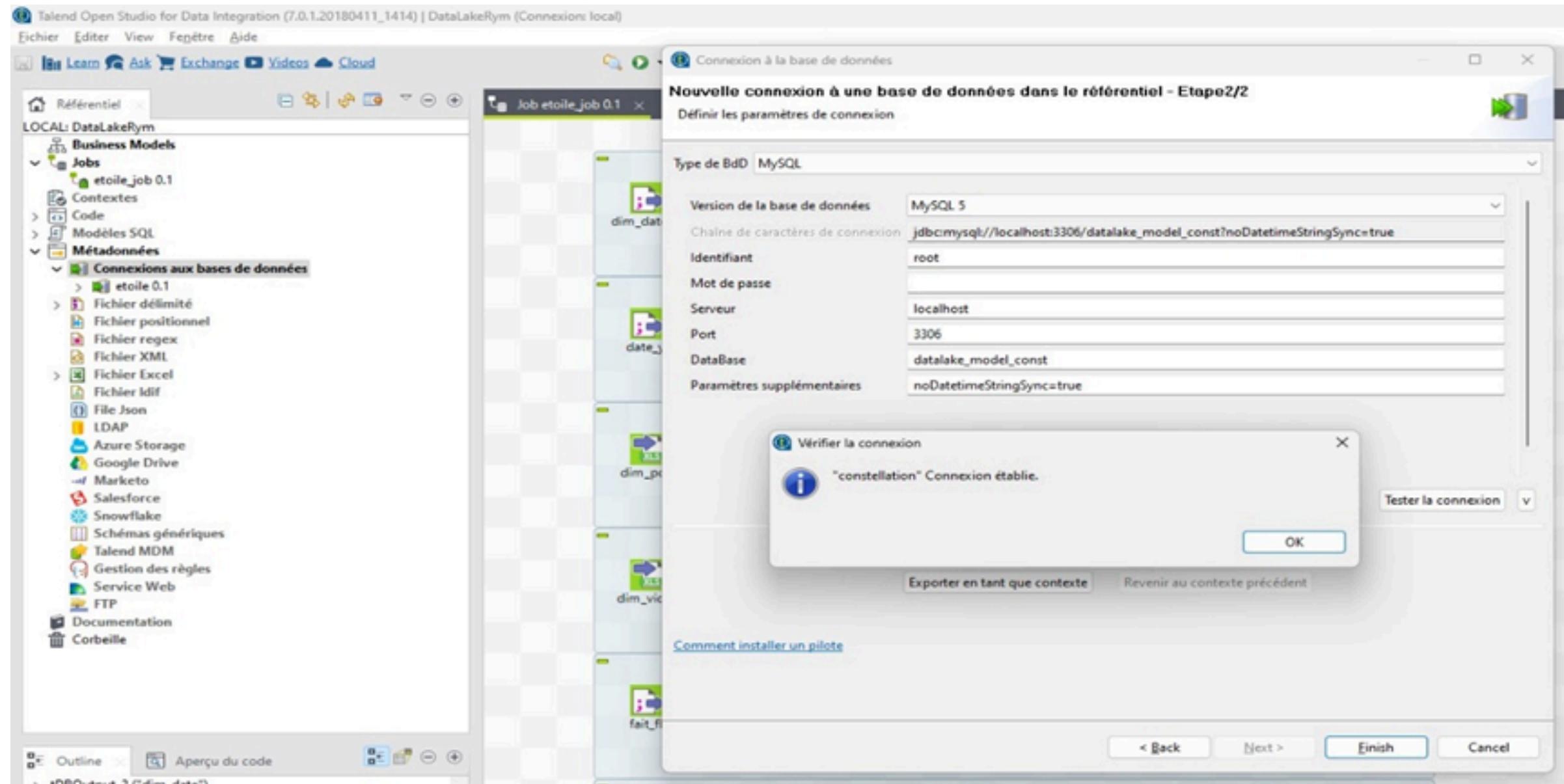


03 Connexion à la base de données :

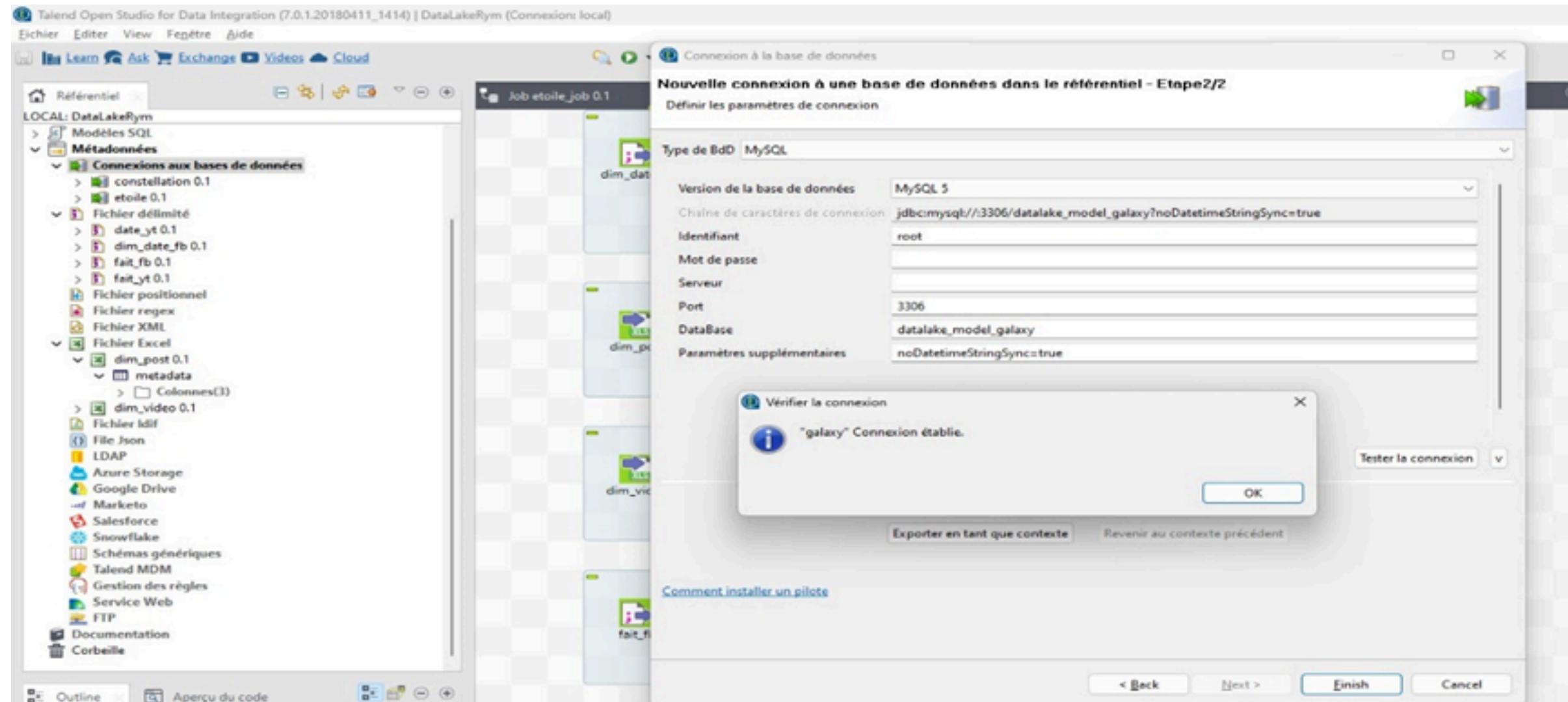
Modele en étoile :



Modele en constellation :



Modele en galaxie :

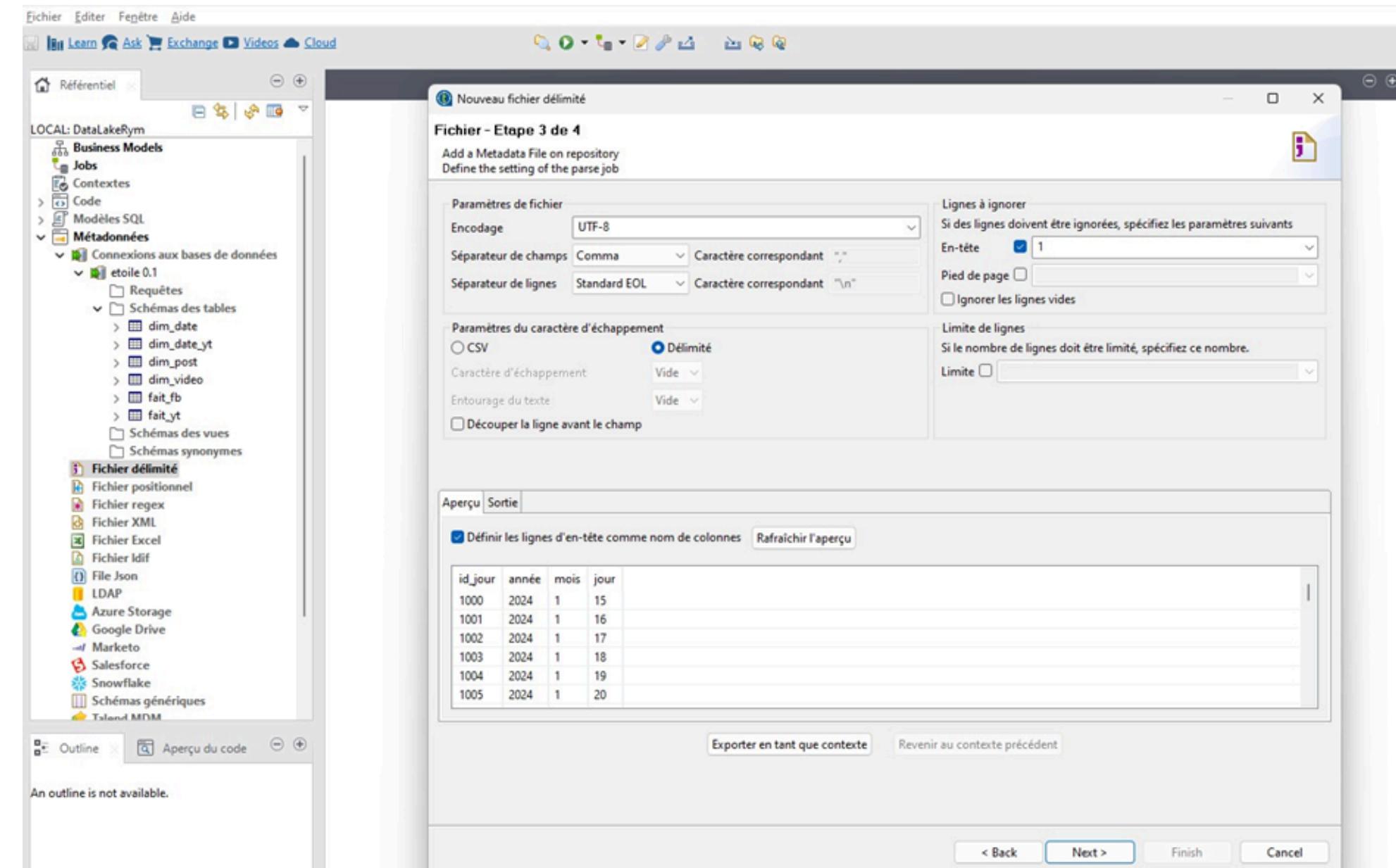


04

Chargement des données CSV :

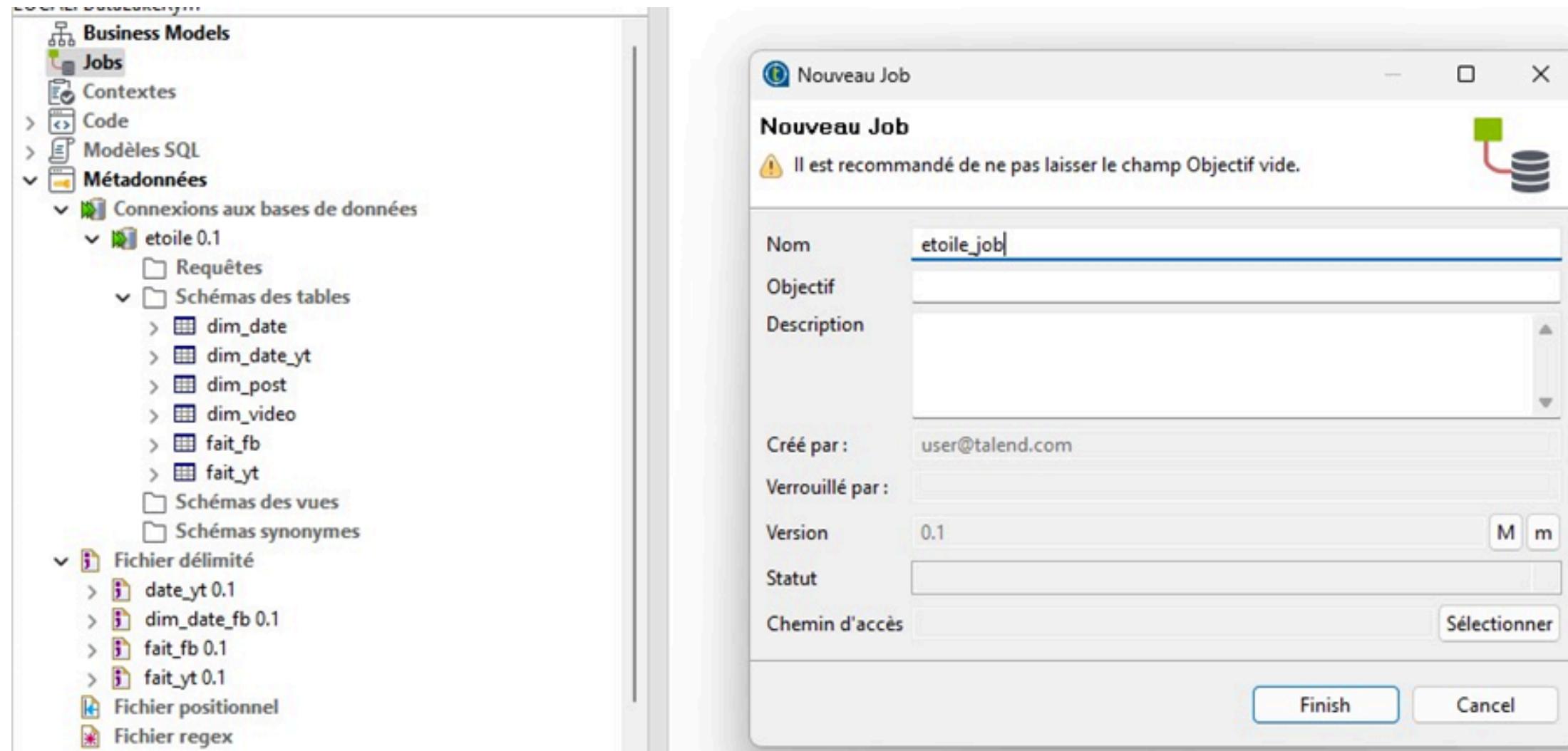
Left sidebar navigation:

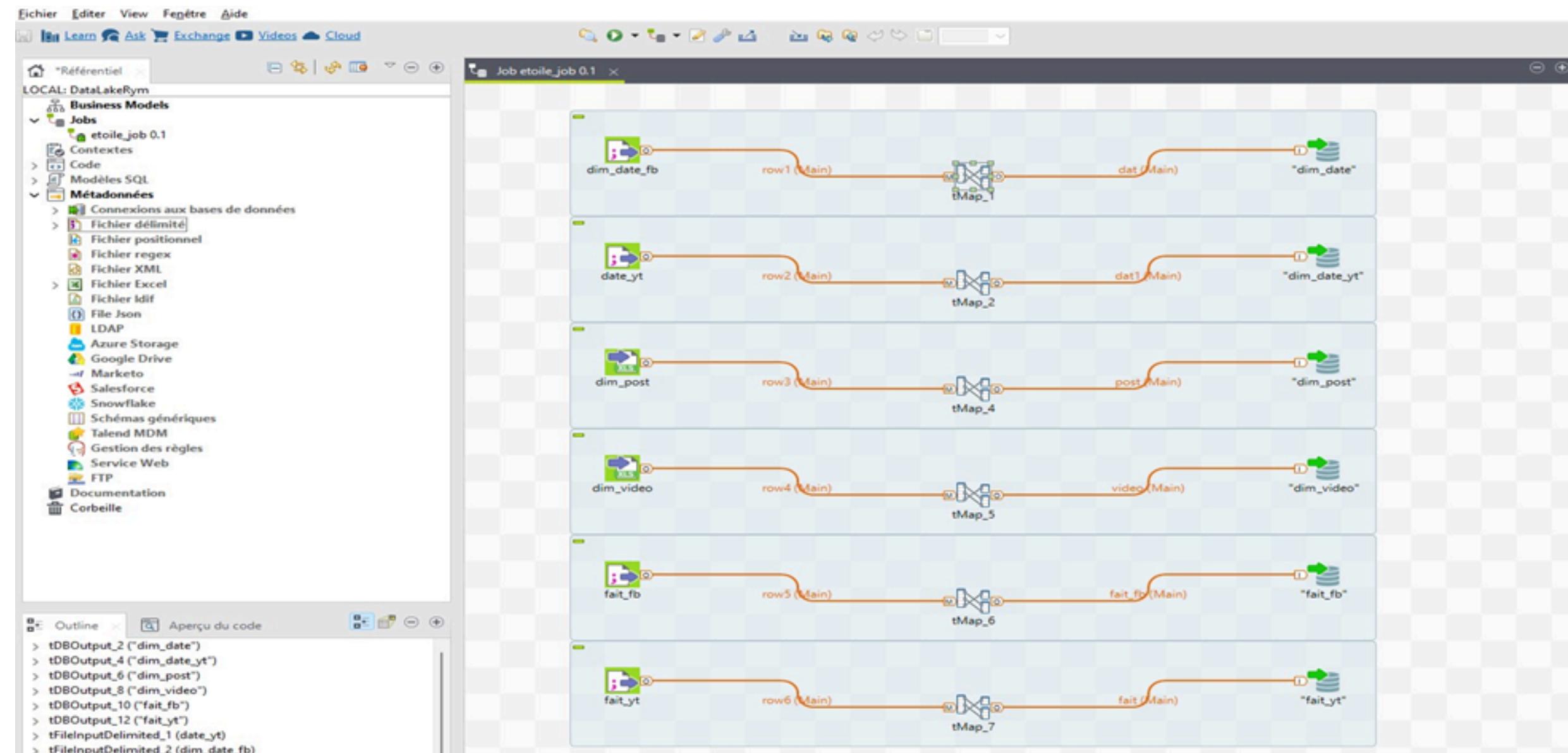
- > Code
- > Modèles SQL
- Métadonnées**
 - > Connexions aux bases de données
 - > etoile 0.1
 - > Requêtes
 - > Schémas des tables
 - > dim_date
 - > dim_date_yt
 - > dim_post
 - > dim_video
 - > fait_fb
 - > fait_yt
 - > Schémas des vues
 - > Schémas synonymes
 - > Fichier délimité
 - > date_yt 0.1
 - > dim_date_fb 0.1
 - > fait_fb 0.1
 - > fait_yt 0.1
 - > Fichier positionnel
 - > Fichier regex
 - > Fichier XML
 - > Fichier Excel
 - > dim_post 0.1
 - > metadata
 - > Colonnes(3)
 - > dim_video 0.1
 - > metadata
 - > Fichier délimité
 - > date_yt 0.1
 - > dim_date_fb 0.1
 - > fait_fb 0.1
 - > fait_yt 0.1



05 Crédit des jobs :

Modèle en étoile :





Talend Open Studio for Data Integration - tMap - tMap_1

Find : Var Mapping auto

row1 dat

Colonne Column Expression

Colonne	Column	Expression
id_jour		row1.id_jour
annee		row1.annee
mois		row1.mois
jour		row1.jour

Editeur de Schéma Editeur d'expression

row1 dat

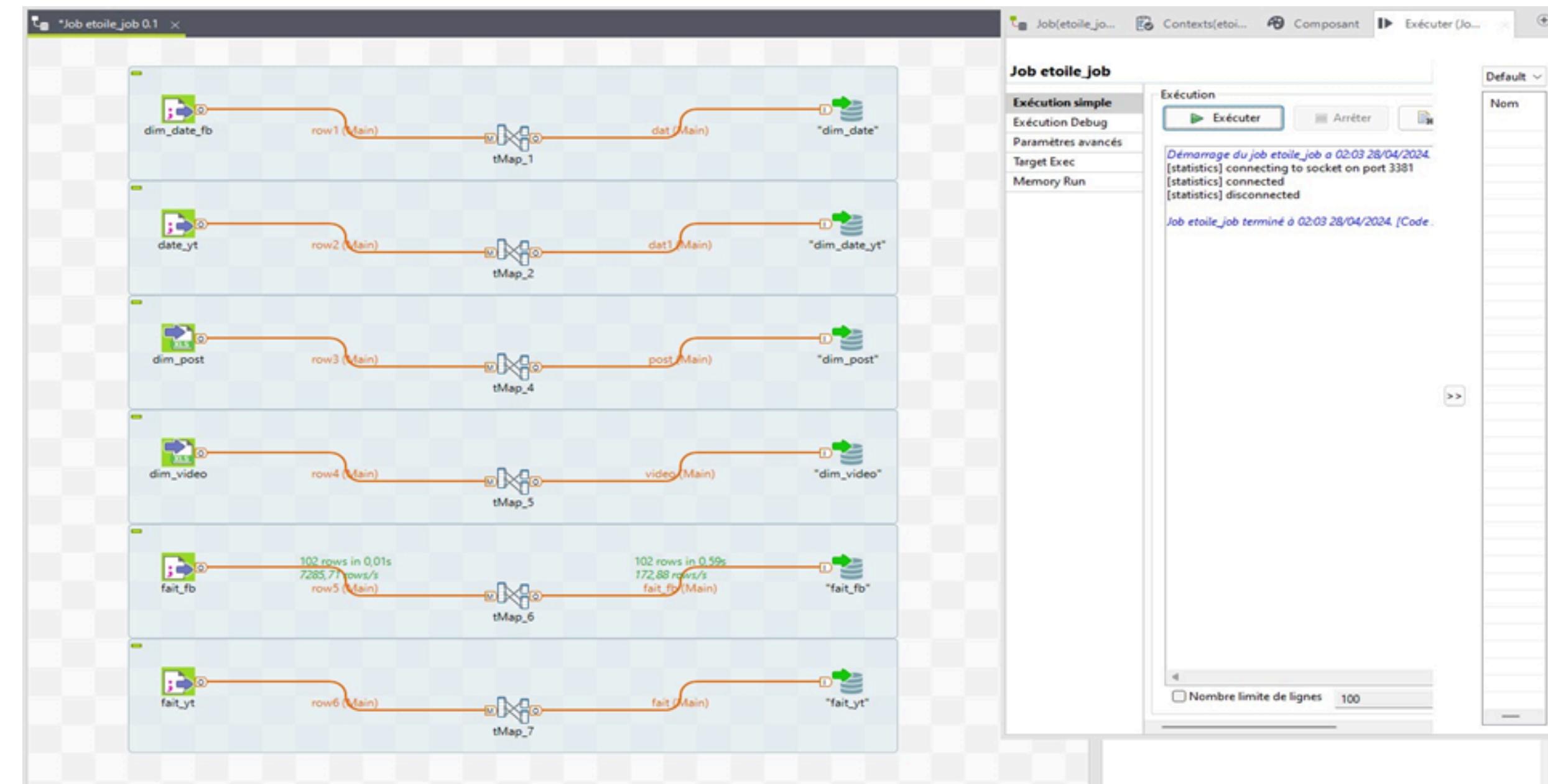
Colonne Clé Type N. Modèle de date (Ctrl+Es... Length Precision Défaut Commentaire

Colonne	Clé	Type	N.	Modèle de date (Ctrl+Es... Length	Precision	Défaut	Commentaire
id_jour		Integer		4	0		
annee		Integer		4	0		
mois		Integer		1	0		
jour		Integer		2	0		

Colonne Clé Type N. Modèle de date (Ctrl+Es... Length Precision Défaut Commentaire

Colonne	Clé	Type	N.	Modèle de date (Ctrl+Es... Length	Precision	Défaut	Commentaire
id_date	✓	int		10	0		
annee		int			10	0	
mois		int			10	0	
jour		int			10	0	

Appliquer OK Annuler



Fin Job

Serveur : 127.0.0.1 » Base de données : datalake_model_etoile

Structure SQL Rechercher Requête Exporter Importer Opérations Priviléges Procédures stockées Évènements Déclencheurs

Filtres

Contenant le mot :

Table	Action	Lignes	Type	Interclassement	Taille	Perte
dim_date	Parcourir Structure Rechercher Insérer Vider Supprimer	102	InnoDB	utf8mb4_general_ci	16,0 kio	-
dim_date_yt	Parcourir Structure Rechercher Insérer Vider Supprimer	100	InnoDB	utf8mb4_general_ci	16,0 kio	-
dim_post	Parcourir Structure Rechercher Insérer Vider Supprimer	102	InnoDB	utf8mb4_general_ci	80,0 kio	-
dim_video	Parcourir Structure Rechercher Insérer Vider Supprimer	100	InnoDB	utf8mb4_general_ci	112,0 kio	-
fait_fb	Parcourir Structure Rechercher Insérer Vider Supprimer	102	InnoDB	utf8mb4_general_ci	32,0 kio	-
fait_yt	Parcourir Structure Rechercher Insérer Vider Supprimer	100	InnoDB	utf8mb4_general_ci	32,0 kio	-
6 tables	Somme	606	InnoDB	utf8mb4_general_ci	288,0 kio	0 o

Tout cocher Avec la sélection :

Imprimer Dictionnaire de données

Modele en constellation :



Fin Job

Serveur : 127.0.0.1 » Base de données : datalake_model_const » Table : fait_yt

Parcourir Structure SQL Rechercher Insérer Exporter Importer Priviléges Opérations Déclencheurs

Affichage des lignes 0 - 24 (total de 100, traitement en 0.0002 seconde(s).)

SELECT * FROM `fait_yt`

Profilage [Éditer en ligne] [Éditer] [Expliquer SQL] [Crée le code source PHP] [Actualiser]

1 > >> | Tout afficher Nombre de lignes : 25 Filtrer les lignes: Chercher dans cette table Trier par clé : Aucun(e)

Options supplémentaires

	id_date_e	id_video_e	duration	view_counts
<input type="checkbox"/> Éditer	1000	VTWPhwmTyYg	3521	1059
<input type="checkbox"/> Éditer	1001	i7ocDOx8bag	189	408
<input type="checkbox"/> Éditer	1002	qFfo4Rq3SGc	405	827
<input type="checkbox"/> Éditer	1003	0YZE4KmW5CE	80	219
<input type="checkbox"/> Éditer	1004	ILFJYU_olw8	61	388
<input type="checkbox"/> Éditer	1005	KGqWufyur_Q	94	396
<input type="checkbox"/> Éditer	1006	YLU7t3AuxSg	71	467
<input type="checkbox"/> Éditer	1007	bPolkr1-bOE	60	587
<input type="checkbox"/> Éditer	1008	m1nNQ-QI2M4	86	545
<input type="checkbox"/> Éditer	1009	oggLJChA5qw	502	832
<input type="checkbox"/> Éditer	1010	2FnCnzM-8k0	62	425
<input type="checkbox"/> Éditer	1011	T-WiwyfR_3c	65	538
<input type="checkbox"/> Éditer	1012	cv72bnKMjls	91	743
<input type="checkbox"/> Éditer	1013	mE6JcoNaJnk	67	313

Serveur : 127.0.0.1 » Base de données : datalake_model_const

Structure SQL Requerre Exporter Importer Opérations Priviléges Procédures stockées Événements Déclencheurs

Filtres

Contenant le mot :

Table	Action	Lignes	Type	Interclassement	Taille	Perte
dim_date	<input type="checkbox"/> Parcourir <input type="checkbox"/> Structure <input type="checkbox"/> Rechercher <input type="checkbox"/> Insérer <input type="checkbox"/> Vider <input type="checkbox"/> Supprimer	102	InnoDB	utf8mb4_general_ci	16,0 kio	-
dim_post	<input type="checkbox"/> Parcourir <input type="checkbox"/> Structure <input type="checkbox"/> Rechercher <input type="checkbox"/> Insérer <input type="checkbox"/> Vider <input type="checkbox"/> Supprimer	102	InnoDB	utf8mb4_general_ci	80,0 kio	-
fait_fb	<input type="checkbox"/> Parcourir <input type="checkbox"/> Structure <input type="checkbox"/> Rechercher <input type="checkbox"/> Insérer <input type="checkbox"/> Vider <input type="checkbox"/> Supprimer	102	InnoDB	utf8mb4_general_ci	32,0 kio	-
fait_yt	<input type="checkbox"/> Parcourir <input type="checkbox"/> Structure <input type="checkbox"/> Rechercher <input type="checkbox"/> Insérer <input type="checkbox"/> Vider <input type="checkbox"/> Supprimer	100	InnoDB	utf8mb4_general_ci	32,0 kio	-
video	<input type="checkbox"/> Parcourir <input type="checkbox"/> Structure <input type="checkbox"/> Rechercher <input type="checkbox"/> Insérer <input type="checkbox"/> Vider <input type="checkbox"/> Supprimer	100	InnoDB	utf8mb4_general_ci	112,0 kio	-
5 tables	Somme	506	InnoDB	utf8mb4_general_ci	272,0 kio	0 o

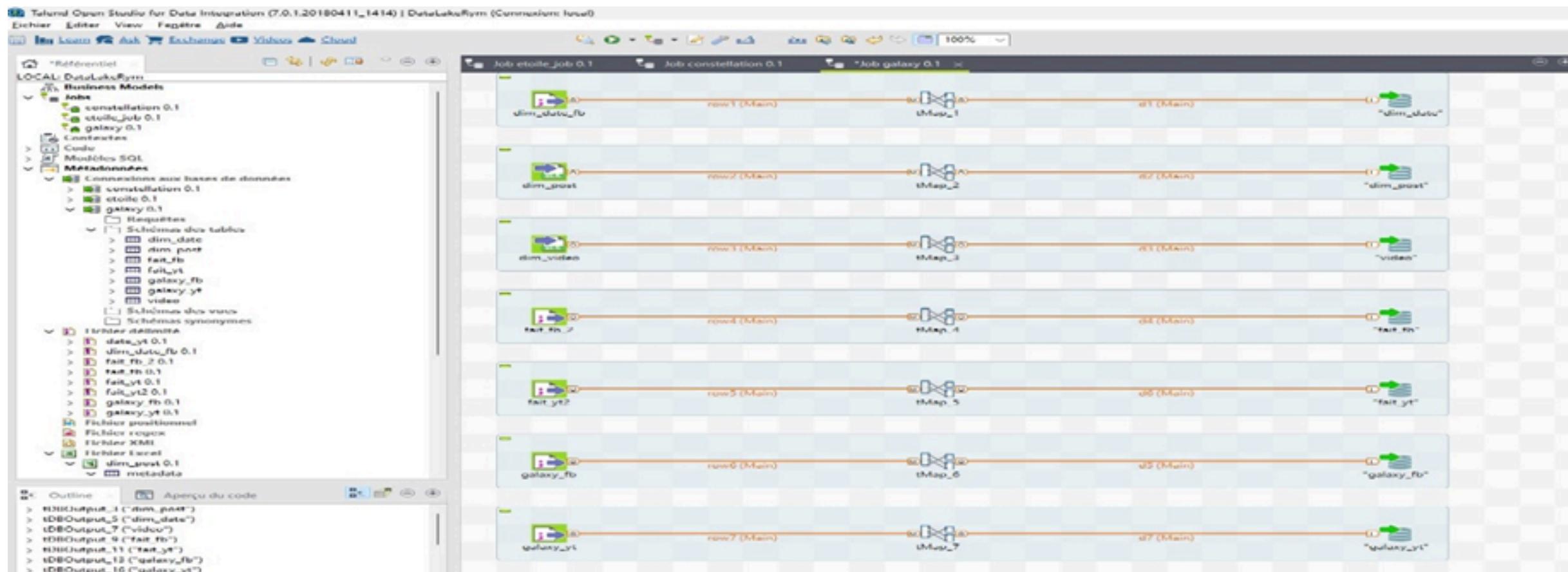
Tout cocher Avec la sélection :

Imprimer Dictionnaire de données

Créer une nouvelle table

Nom de table Nombre de colonnes
 4

Modele en galaxie :



Fin Job

Serveur : 127.0.0.1 > Base de données : datalake_model_galaxy > Table : galaxy_yt

Parcourir Structure SQL Rechercher Insérer Exporter Importer Priviléges Opérations Déclencheurs

Affichage des lignes 0 - 24 (total de 100, traitement en 0.0002 seconde(s).)

SELECT * FROM `galaxy_yt`

Profilage [Éditer en ligne] [Éditer] [Expliquer SQL] [Crée le code source PHP] [Actualiser]

1 > >> Tout afficher Nombre de lignes : 25 Filtrer les lignes: Chercher dans cette table Trier par clé : Aucun(e)

Options supplémentaires

	Éditer	Copier	Supprimer	-Mzm_q8-4R4	1062	1062
				-tzA9i67fyg	1063	1063
				0R0zf_3IVho	1064	1064
				0YZE4KmW5CE	1003	1003
				17_cS92Bqaw	1089	1089
				1xoJfqVhS8Y	1048	1048
				2FdDwY9z9A0	1065	1065
				2FnCnzM-8k0	1010	1010
				2ygS60saP6Q	1034	1034
				39jlxXE92bA	1097	1097
				3HAwmTKLKVI	1090	1090
				3IEZ_txWCCI	1016	1016
				3wXLVstnSro	1017	1017
				3Yrwu4RWrUU	1093	1093

Serveur : 127.0.0.1 > Base de données : datalake_model_galaxy

Structure SQL Rechercher Requête Exporter Importer Opérations Priviléges Procédures stockées

Filtres

Contenant le mot :

Table	Action	Lignes	Type	Interclassement	Taille	Perte
dim_date	Parcourir Structure Rechercher Insérer Vider Supprimer	102	InnoDB	utf8mb4_general_ci	16,0 kio	-
dim_post	Parcourir Structure Rechercher Insérer Vider Supprimer	102	InnoDB	utf8mb4_general_ci	80,0 kio	-
fait_fb	Parcourir Structure Rechercher Insérer Vider Supprimer	102	InnoDB	utf8mb4_general_ci	16,0 kio	-
fait_yt	Parcourir Structure Rechercher Insérer Vider Supprimer	100	InnoDB	utf8mb4_general_ci	16,0 kio	-
galaxy_fb	Parcourir Structure Rechercher Insérer Vider Supprimer	102	InnoDB	utf8mb4_general_ci	48,0 kio	-
galaxy_yt	Parcourir Structure Rechercher Insérer Vider Supprimer	100	InnoDB	utf8mb4_general_ci	48,0 kio	-
video	Parcourir Structure Rechercher Insérer Vider Supprimer	100	InnoDB	utf8mb4_general_ci	112,0 kio	-
7 tables	Somme	708	InnoDB	utf8mb4_general_ci	336,0 kio	0 o

Tout cocher Avec la sélection :

Imprimer Dictionnaire de données

ANALYSE DES DONNÉES



01

Notre objectif principal

Comprendre les préférences et les motivations des utilisateurs de Coursera en examinant les interactions avec les publications sur les réseaux sociaux.

02

La stratégie analytique

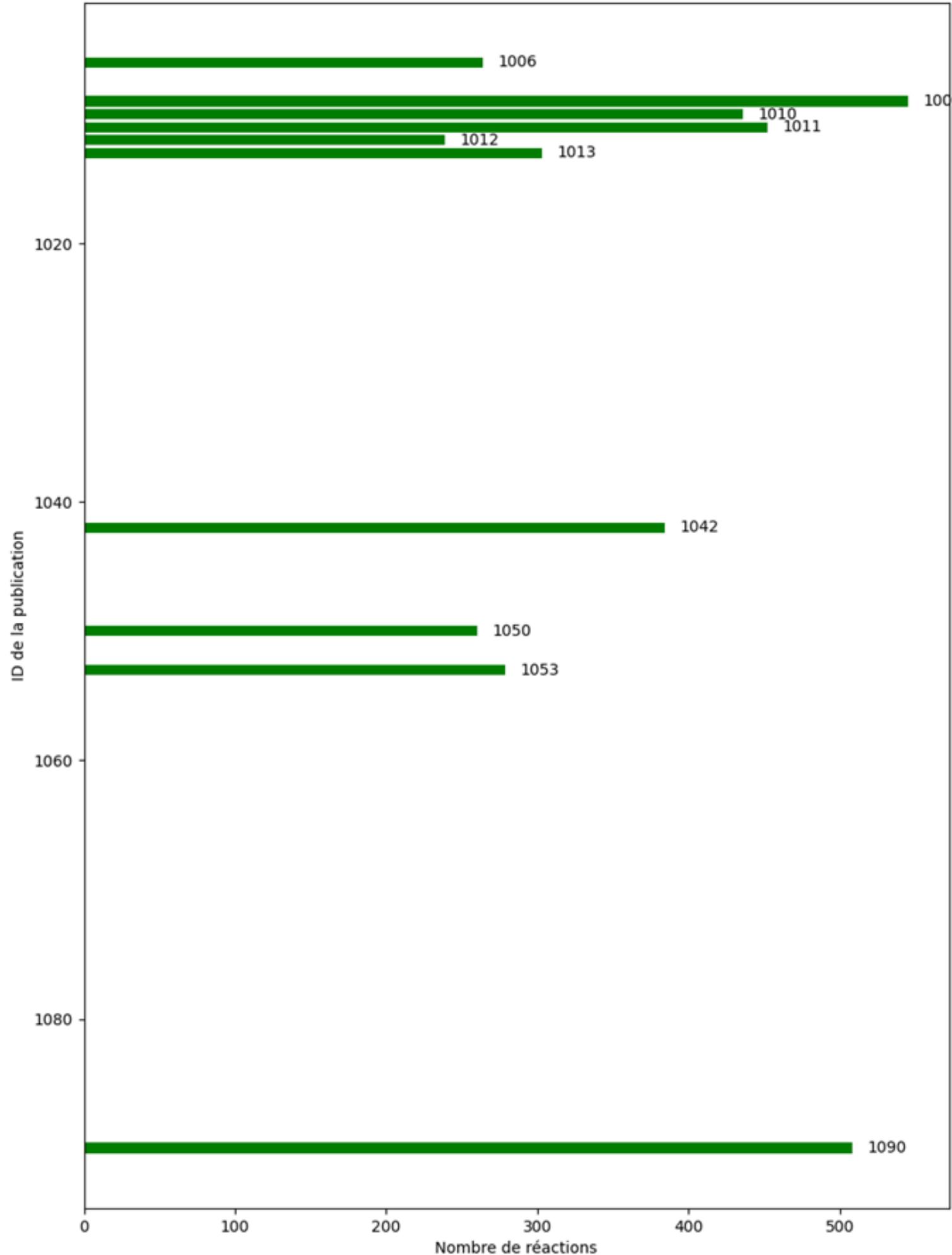
Visualiser les 10 principales publications selon le nombre de commentaires, de réactions et de partages.
Examiner également les 50 mots les plus fréquemment utilisés dans les captions de ces publications.

03

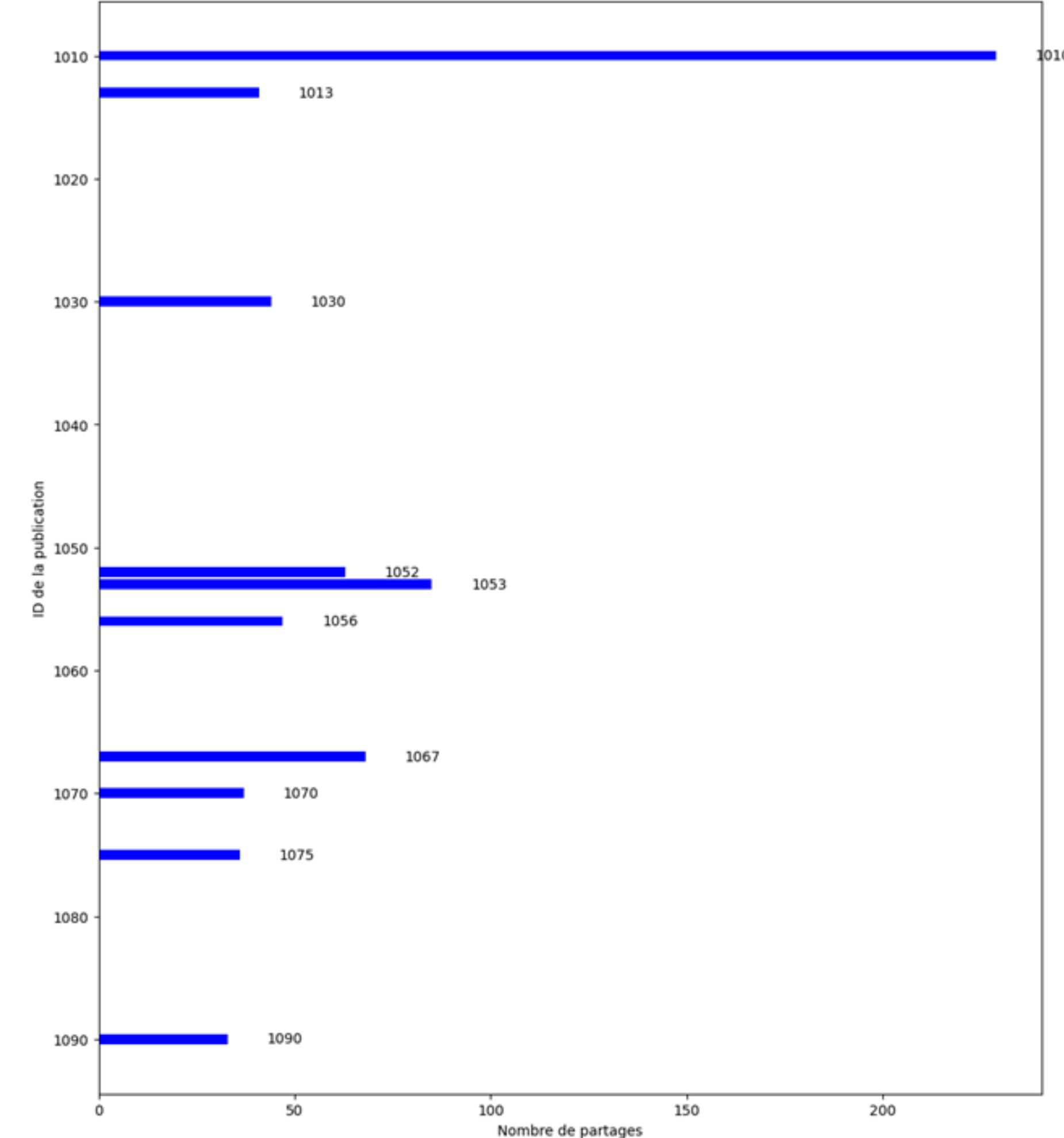
Pourquoi ?

Cette visualisation nous offre un aperçu des thèmes et des sujets qui captent l'attention de notre public et ce qui motive l'audience de Coursera.

Top 10 des publications avec le plus de réactions



Top 10 des publications avec le plus de partages

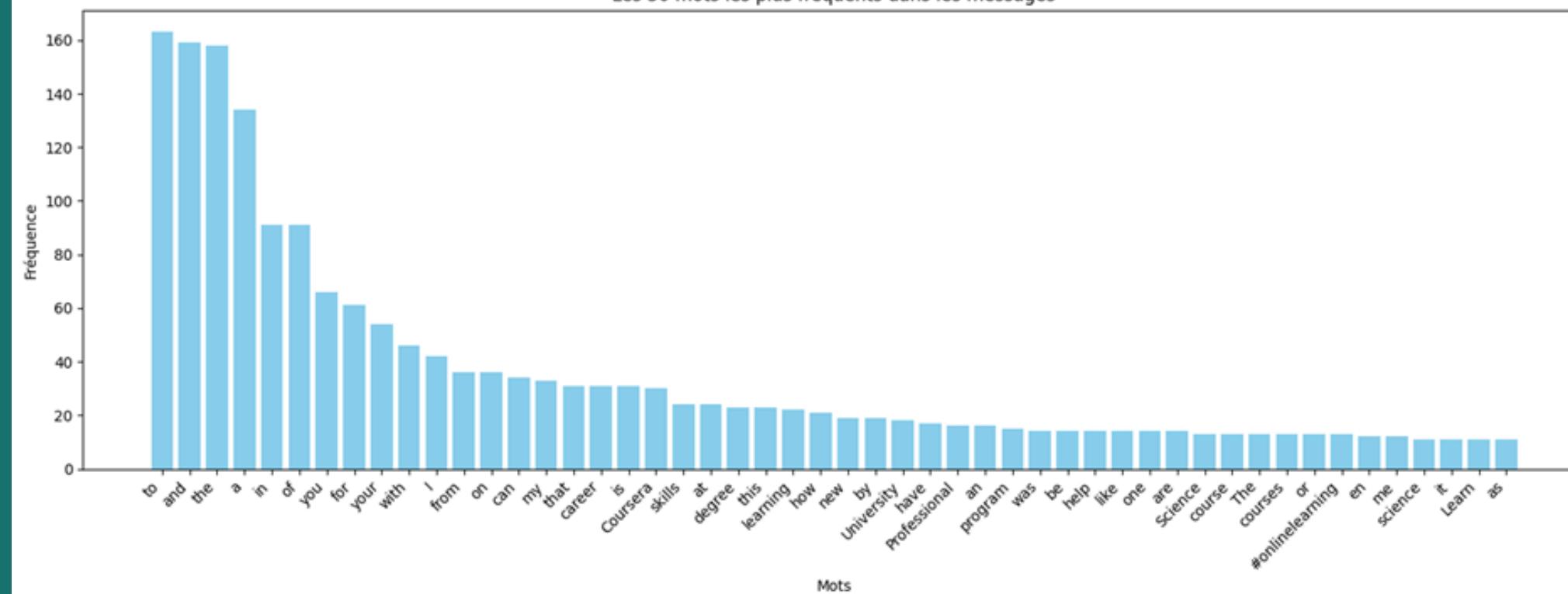


ANALYSE DES DONNÉES

Les mots comme "Coursera", "skills", "learning", "degree", "program", et "online learning" ressortent particulièrement, indiquant une focalisation sur l'éducation, le développement des compétences



Les 50 mots les plus fréquents dans les messages



En mettant en avant ces mots-clés, Coursera cherche à attirer l'attention des étudiants potentiels et à souligner les avantages et les opportunités offerts par sa plateforme.

Ces mots peuvent également jouer un rôle important sur la psychologie des étudiants en les incitant à s'intéresser davantage à l'éducation en ligne.

CONCLUSION