

# RAPPORT DE PROJET

**Ingénierie des Systèmes décisionnelles NFE211**  
**- CNAM PARIS**  
**2023 - 2024**

Etude sur les données issues des ventes d'une librairie

Réalisé par : Sarah Bitan

Professeur : Pr Elisabeth METAIS

le cnam

## I. Introduction

La généralisation des entrepôts de données est devenue une pratique courante au sein des entreprises. Cette procédure implique la collecte de données provenant de sources variées telles que les bases de données opérationnelles, les fichiers XML, les ERP, etc. Ces données

sont ensuite soumises à des transformations avant d'être intégrées dans le Datawarehouse, constituant ainsi un projet de Business Intelligence (BI) qui suit un processus en cascade.

Dans le cadre de la validation de l'UE NFE211, j'ai conduit une étude portant sur les données relatives aux ventes d'une librairie. L'objectif de cette étude était d'anticiper et de piloter les préférences littéraires des clients.

Ce projet a été réalisé en deux phases. Dans ce rapport, je vais d'abord aborder la collecte de mes sources de données, en détaillant leurs caractéristiques ainsi que les diverses transformations opérées pour les rendre exploitable dans mon futur Datawarehouse. Ensuite, je vais présenter la structure interne de mon Datawarehouse, mettant particulièrement en avant la table des faits et les dimensions associées à mon jeu de données, à savoir les ventes de la librairie. Je vais également expliciter le modèle dimensionnel que j'ai choisi d'employer.

En ce qui concerne le processus de développement, j'ai opté pour l'utilisation de l'outil Stich Data Loader afin d'assurer l'extraction, la transformation et le chargement des données. De surcroît, j'ai fait le choix d'utiliser le moteur de base de données Oracle11g Express pour stocker les données dans mon Datawarehouse.

Cette approche me permettra d'acquérir des informations pertinentes et exploitables sur les ventes de la librairie, afin d'apporter un éclairage précieux pour la prise de décisions et l'optimisation des activités commerciales.

## II. Librairie La page

La librairie La Page, fondée en 1978 au cœur de South Kensington à Londres, est une référence incontournable pour la communauté francophone de la capitale britannique. Située à proximité du lycée français, de l'Institut Français, ainsi que des prestigieux musées

tels que le Victoria & Albert Museum, le Musée d'Histoire Naturelle et le Musée des Sciences, elle est ancrée au cœur de la vie culturelle de la ville.

En 2008, Isabelle Lemarchand a repris les rênes de la librairie après y avoir travaillé quelques mois. Son ambition était claire : faire de La Page un lieu dédié à la culture francophone. Au fil des années, la librairie a élargi son offre en ouvrant ses rayons à de nombreux nouveaux éditeurs et à des domaines plus variés tels que la bande dessinée, la littérature étrangère, les sciences humaines et la philosophie.

L'entreprise La page était originellement une librairie physique avec plusieurs points de vente. Mais devant le succès de certains de ses produits et l'engouement de ses clients, elle a décidé depuis 2 ans d'ouvrir un site de vente en ligne. La structure a besoin d'aide pour mieux comprendre ses données.

Il est à noter que La Page possède une base de données complète, regroupant des informations sur les transactions, les clients et les livres. Cette précieuse ressource permet d'optimiser la gestion de la librairie et d'offrir un service encore plus personnalisé à la clientèle fidèle de l'établissement.

Nous possédons trois ensembles de données au format CSV provenant de la plateforme Kaggle, datant de l'année 2020, que nous avons récupérés :

Detail   Compact   Column

client_id	sex	# birth
<b>8623</b> unique values	f m	52% 48%
c_4410	f	1967
c_7839	f	1975
c_1699	f	1984
c_5961	f	1962
c_5320	m	1943
c_415	m	1993

Le jeu de données customers.csv contenant 3 variables :

- Le client\_id de chaque client pour l'identifier
- Le sex définit par f ou m pour connaître le genre du client
- Birth qui représente l'année de naissance de chaque client

Detail   Compact   Column



Le jeu de données products.csv contenant 3 variables :

- L'id\_prod contenant l'identifiant de chaque produit
- Le price ou est stocké le prix de chaque produit
- Categ qui prend pour valeurs 0, 1 ou 2 pour définir 3 catégories de livres différentes

Detail   Compact   Column

4 of 4 columns ▾

<b>A</b> id_prod	<b>A</b> date	<b>A</b> session_id	<b>A</b> client_id
<b>3267</b> unique values	 1Mar21      1Mar23	<b>342316</b> unique values	c_1609 c_6714 Other (644857)
0_1518	2022-05-20 13:21:29.043970	s_211425	c_103
1_251	2022-02-02 07:55:19.149409	s_158752	c_8534
0_1277	2022-06-18 15:44:33.155329	s_225667	c_6714
2_209	2021-06-24 04:19:29.835891	s_52962	c_6941

Le jeu de données transactions.csv contenant 4 variables :

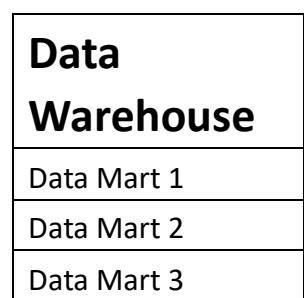
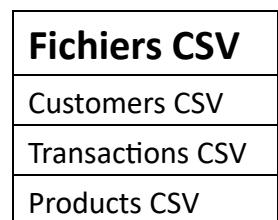
- L'id\_prod qui est l'identifiant de chaque vente
- La date qui est la date de la transaction sous format YY/MM/DD hh:mm:ss tt
- Le session\_id qui est l'identifiant de chaque session
- Le client\_id qui est l'identifiant de chaque client

### III. Architecture technique de mon Data Warehouse (DWH)

#### 1. Présentation de mon Data Warehouse

L'architecture technique d'un Data Warehouse est cruciale pour garantir l'efficacité, la robustesse et la performance du système.

Voici une description générale des composants et de l'organisation de notre architecture de Data Warehouse :



Architecture fonctionnelle du Datawarehouse

- **Sources de Données**

L'intégration des données dans mon Data Warehouse provient principalement des transactions des clients enregistrées lors de chaque vente effectuée sur le site de la librairie La Page.

Étant donné les contraintes d'accès à la base de données opérationnelle, j'ai opté pour la collecte des données émises à chaque transaction sur le site en ligne. Ces données sont collectées sous forme de fichiers csv.

En résumé, le processus d'alimentation de mon Data Warehouse repose sur la collecte des données de chaque vente, qui constituent une source fiable et pertinente pour suivre les transactions des clients à la librairie La Page. Cette démarche permet d'obtenir des données exploitables et structurées, prêtes à être intégrées dans le système de stockage et d'analyse de mon Data Warehouse.

- **Zone de Staging**

Cette zone est utilisée pour stocker temporairement les données brutes extraites des sources. Elle sert de zone de transit avant que les données ne soient transformées et chargées dans le Data Warehouse proprement dit.

- **ETL (Extract, Transform, Load)**

C'est le processus d'extraction, de transformation et de chargement des données du système source vers le Data Warehouse. Les outils ETL (comme Talend, Informatica, etc.) sont utilisés pour automatiser ces opérations.

- **Data Warehouse**

Le Data Warehouse est le cœur du système. Il est conçu pour stocker les données organisées et structurées de manière à faciliter l'analyse. On distingue deux zones principales

**Zone des Faits** : C'est là où sont stockées les mesures numériques, telles que les ventes, les quantités, etc.

**Dimensions** : Ce sont les éléments qui permettent de décrire les faits. Par exemple, pour les ventes, les dimensions peuvent inclure la date, le produit, le client, etc.

- **Moteur de Base de Données**

C'est le logiciel qui gère le stockage, la récupération et la mise à jour des données dans le Data Warehouse. Il peut s'agir de bases de données relationnelles comme Oracle, MySQL, SQL Server, etc., ou de technologies NoSQL comme MongoDB, Cassandra, etc.

- **Outils d'Accès et d'Analyse**

Ces outils permettent aux utilisateurs d'interroger, d'analyser et de visualiser les données stockées dans le Data Warehouse. Il peut s'agir de solutions comme Tableau, Power BI, QlikView, etc.

- **Couches de Sécurité et Gestion des Accès**

Ces couches garantissent que seules les personnes autorisées ont accès aux données et qu'elles ne peuvent voir que les informations auxquelles elles sont habilitées.

- **Monitoring et Gestion des Performances**

Des outils de surveillance sont utilisés pour suivre les performances du Data Warehouse, détecter les goulots d'étranglement et optimiser les requêtes.

- **Système d'Archivage**

Les données historiques peuvent être déplacées vers un système d'archivage pour libérer de l'espace dans le Data Warehouse principal tout en permettant un accès ultérieur si nécessaire.

- **Plan de Reprise d'Activité (PRA) et Sauvegarde**

Des procédures et des mécanismes sont en place pour garantir la disponibilité des données en cas de sinistre ou de panne.

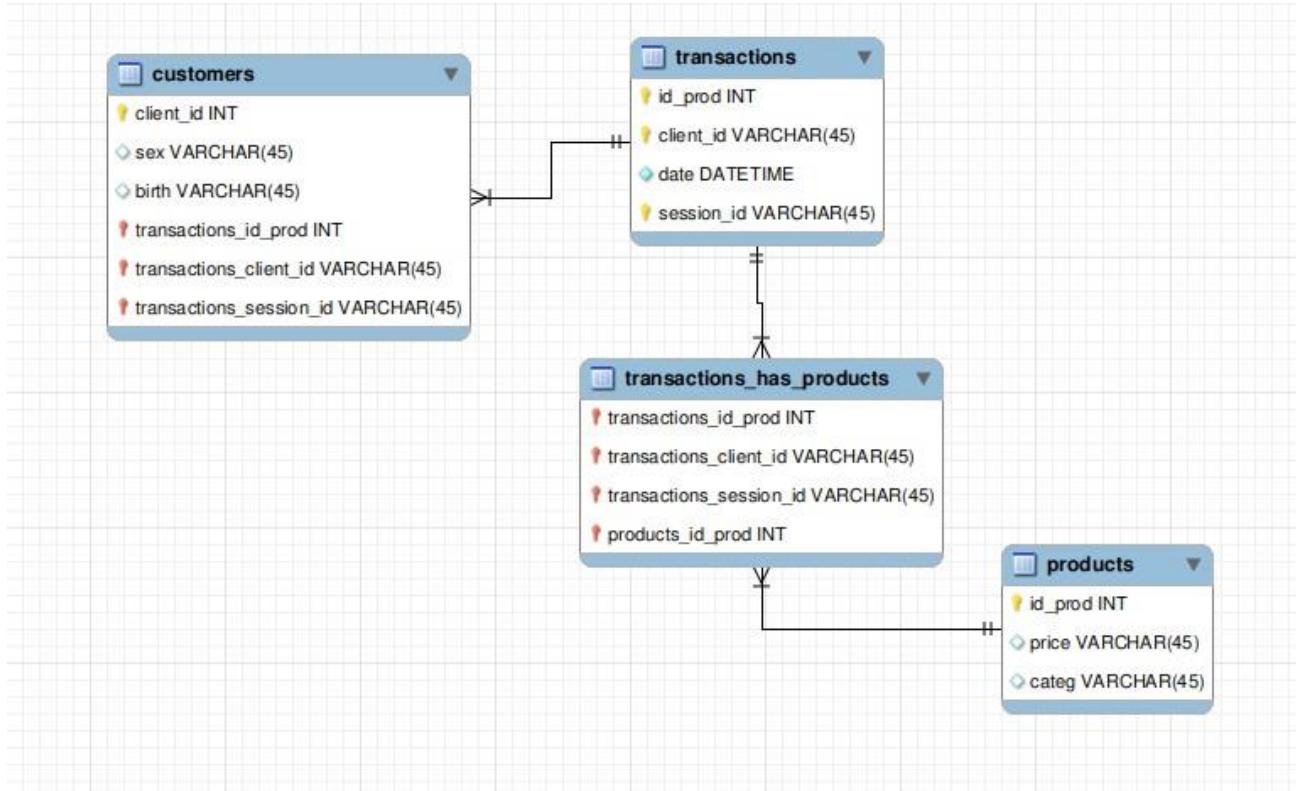
## 2. Model Multidimensionnel (OLAP)

Un modèle OLAP (OnLine Analytical Processing) est une approche informatique qui permet d'interagir avec des données multidimensionnelles de manière rapide et flexible. Il est spécialement conçu pour l'analyse et la visualisation de grandes quantités de données, ce qui en fait un outil essentiel pour les entreprises et les organisations qui doivent traiter des ensembles de données complexes.

# IV. Conception et développement

## 1. Création de la BDD en SQL

Pour la conception de notre base de données, nous avons opté pour MySQL, un système de gestion de base de données relationnelle réputé pour sa fiabilité et sa performance. Afin de structurer efficacement nos données, nous avons formulé des requêtes SQL pour créer et définir nos tables.



## REQUETES SQL:

### Table Customers

```
CREATE TABLE customers (
    `client_id` INT NOT NULL,
    `sex` VARCHAR(45) NULL,
    `birth` VARCHAR(45) NULL,
    `transactions_id_prod` INT NOT NULL,
    `transactions_client_id` VARCHAR(45) NOT NULL,
    `transactions_session_id` VARCHAR(45) NOT NULL,
    PRIMARY KEY (`client_id`, `transactions_id_prod`, `transactions_client_id`,
    `transactions_session_id`),
    INDEX `fk_customers_transactions1_idx` (`transactions_id_prod` ASC,
    `transactions_client_id` ASC, `transactions_session_id` ASC) VISIBLE,
    CONSTRAINT `fk_customers_transactions1`
        FOREIGN KEY (`transactions_id_prod`, `transactions_client_id`, `transactions_session_id`)
        REFERENCES `mydb`.`transactions` (`id_prod`, `client_id`, `session_id`))
```

### Table Transactions:

```
CREATE TABLE transactions (
    `id_prod` INT NOT NULL,
    `client_id` VARCHAR(45) NOT NULL,
    `date` DATETIME NOT NULL,
    `session_id` VARCHAR(45) NOT NULL,
    PRIMARY KEY (`id_prod`, `client_id`, `session_id`))
```

Table Products:

```
CREATE TABLE products (
    `id_prod` INT NOT NULL,
    `price` VARCHAR(45) NULL,
    `categ` VARCHAR(45) NULL,
    PRIMARY KEY (`id_prod`))
```

Table transactions\_has\_products:

```
CREATE TABLE transactions_has_products (
    `transactions_id_prod` INT NOT NULL,
    `transactions_client_id` VARCHAR(45) NOT NULL,
    `transactions_session_id` VARCHAR(45) NOT NULL,
    `products_id_prod` INT NOT NULL,
    PRIMARY KEY (`transactions_id_prod`, `transactions_client_id`, `transactions_session_id`,
    `products_id_prod`),
    INDEX `fk_transactions_has_products_products1_idx` (`products_id_prod` ASC) VISIBLE,
    INDEX `fk_transactions_has_products_transactions_idx` (`transactions_id_prod` ASC,
    `transactions_client_id` ASC, `transactions_session_id` ASC) VISIBLE,
    CONSTRAINT `fk_transactions_has_products_transactions`
        FOREIGN KEY (`transactions_id_prod` , `transactions_client_id` ,
    `transactions_session_id`)
        REFERENCES `mydb`.`transactions`(`id_prod` , `client_id` , `session_id`),
    CONSTRAINT `fk_transactions_has_products_products1`
        FOREIGN KEY (`products_id_prod`)
        REFERENCES `mydb`.`products`(`id_prod`))
```

## 2.Réalisation de l'ETL

ETL signifie Extraction, Transformation et Chargement (en anglais, Extract, Transform, Load). C'est un processus essentiel dans le domaine de la gestion de données et de l'informatique décisionnelle (BI).

Voici ce que chaque étape représente :

- **Extraction (Extract)** : Dans cette étape, les données sont extraites de diverses sources de données. Ces sources peuvent être des bases de données, des fichiers plats, des API, des services cloud, etc. L'objectif est de récupérer les données nécessaires pour les analyser ou les traiter ultérieurement.
- **Transformation (Transform)** : Les données extraites ne sont souvent pas directement utilisables telles quelles. La phase de transformation implique le nettoyage, la normalisation, la structuration et la modification des données pour les rendre cohérentes et exploitables. Cela peut impliquer des opérations telles que la fusion de données, le filtrage, le calcul de nouvelles métriques, etc.
- **Changement (Load)** : Une fois les données extraites et transformées, elles sont chargées dans un entrepôt de données ou une base de données spécialement conçue pour l'analyse. Cette base de données est souvent optimisée pour la lecture et est utilisée par les outils de business intelligence pour générer des rapports et des visualisations.

Le processus ETL est essentiel car il permet de préparer et de transformer les données brutes en un format utilisable pour l'analyse et la génération de rapports.

Cela est particulièrement crucial dans les environnements où les données proviennent de sources hétérogènes et doivent être consolidées pour une analyse cohérente.

En résumé, ETL est un processus clé pour préparer les données en vue de l'analyse et de la prise de décision dans le domaine de la gestion de données et de la business intelligence.

Pour mettre en œuvre notre processus ETL, nous opterons pour l'outil ELT Talend Open Studio for Data. Ce logiciel offre aux utilisateurs la possibilité de concevoir et de développer des processus d'intégration de données grâce à une interface graphique intuitive.

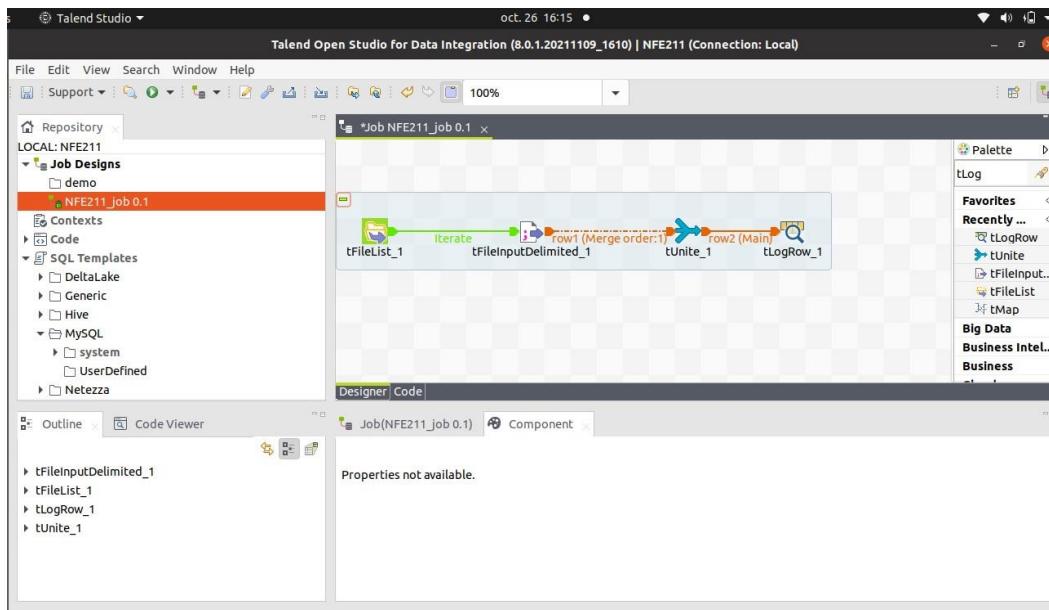
Il prend en charge une diversité de sources et de destinations de données, et propose des composants préconfigurés qui simplifient grandement le processus d'intégration.

Nous souhaitons créer un job dans Talend Open Studio qui réalise une jointure entre nos trois fichiers CSV et génère en sortie le fichier final.

Pour ce faire, nous ajoutons les composants suivants dans l'espace de modélisation graphique : un tFileDialog, un tFileInputDelimited, un tUnite et un tLogRow.

Nous configurons la destination de nos fichiers dans les paramètres de tFileList et nous spécifions les caractères de notre fichier CSV, notamment le séparateur "," , dans les paramètres de tFileInputDelimited.

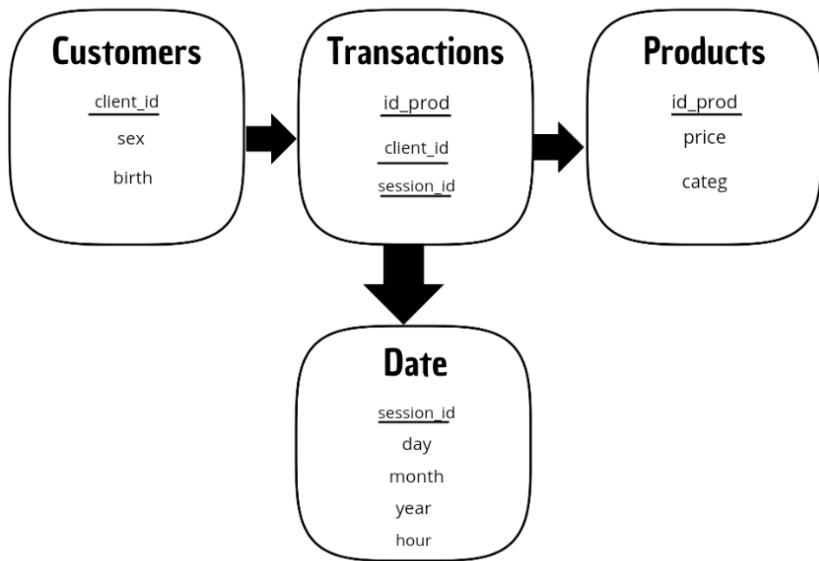
Ensuite, nous établissons une connexion entre le composant tFileList et le tFileInputDelimited en utilisant un lien Iterate, et nous connectons les autres composants à l'aide de liens Row main.



### 3. Modélisation en étoile

Une modélisation en étoile (star schema) est un schéma de base de données utilisé dans les entrepôts de données pour faciliter les requêtes analytiques.

Elle se compose d'une table centrale de faits entourée de tables de dimensions. Voici une modélisation en étoile basée sur notre base de données :



L'observation dans ce contexte met en évidence la table centrale, à savoir la table "transactions", qui occupe une position centrale dans la structure globale.

Celle-ci est entourée par les tables de dimensions "Customers" et "Products".

Cette disposition suggère une organisation relationnelle où les transactions, représentant probablement des interactions commerciales ou des échanges, sont liées aux informations sur les clients ("Customers") et les détails des produits ("Products"). Ainsi, la table centrale agit comme un pivot central, reliant les aspects transactionnels aux données contextuelles fournies par les tables de dimensions.

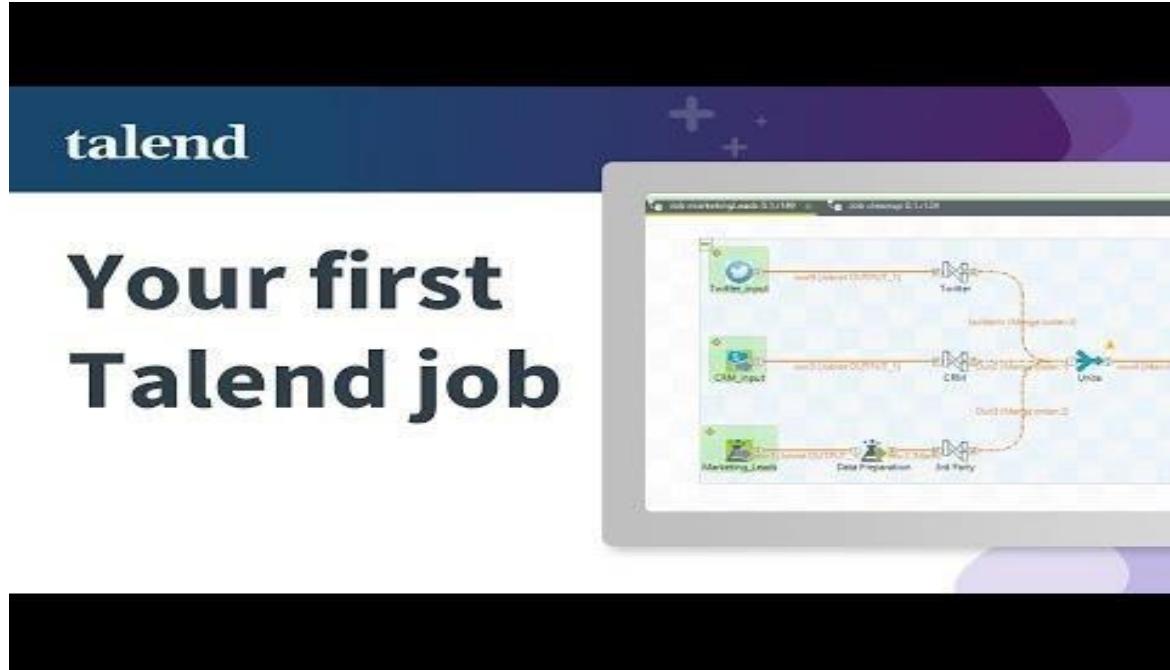
Cette structure est couramment utilisée dans les bases de données relationnelles pour capturer et organiser des informations hétérogènes tout en établissant des relations significatives entre les différentes entités.

## V. Bibliographie

Notre jeu de données : <https://www.kaggle.com/datasets/lolorenzobarcat/data-lapage>

Documentation Talend Open Studio :

<https://www.youtube.com/watch?v=0v8TjYAOS6c>



<https://help.talend.com/r/fr-FR/7.3/processing/textractxmlfield-tmysqlinput-tfileoutputdelimited>  
<https://help.talend.com/r/fr-FR/7.3/processing/textractxmlfield-tmysqlinput-tfileoutputdelimited-extracting-xml-data-from-field-in-database-table-standard-component>

Réalisation des modélisations : <https://app.creately.com/d/KawI01dAXRX/edit>