

RAPPORT DE PROJET

Data Warehouse

Analyse des Ventes de Mangas

SAE 302 - Intégration de données dans un datawarehouse

IUT Lumière Lyon 2

Département Science des Données

Février 2026

DELIN Mattéo, ALSHAWWA Tasnim, GROSJEAN Violette

Table des matières

Table des matières.....	2
1. Introduction	3
1.1. Objectifs du projet.....	3
1.2. Questions métier.....	3
2. Sources de données	4
2.1. Description des sources	4
2.2. Structure détaillée des sources.....	4
3. Processus ETL	7
3.1. Initialisation et Zone de Préparation (SAS).....	7
3.2. Transformation et Chargement des Dimensions (SAS vers DWH).....	7
3.3. Chargement de la Table de Faits (FACT)	8
4. Architecture du Data Warehouse	8
4.1. Modèle en flocon (Snowflake Schema)	8
4.2. Schéma relationnel	9
4.3. Justification du modèle en flocon.....	10
5. Description détaillée des tables.....	10
5.1. Table de faits : FACT_Manga_Sales	10
5.2. Tables de dimension.....	11
6. Visualisation et reporting avec Power BI	12
6.1. Mise en place de Power BI	12
6.2. Dashboard développé.....	13
6.3. Analyse des résultats.....	13
6.4. Interactivité et fonctionnalités.....	14
7. Conclusion	14

1. Introduction

Ce projet vise à concevoir et implémenter un Data Warehouse dédié à l'analyse des ventes de mangas. Dans un marché en constante évolution, disposer d'un système d'information décisionnel performant est essentiel pour comprendre les tendances, identifier les auteurs et genres populaires, et optimiser les stratégies commerciales.

Le marché du manga représente aujourd'hui un secteur économique majeur avec des millions de volumes vendus chaque année dans le monde. Les éditeurs, distributeurs et analystes du secteur ont besoin d'outils décisionnels pour comprendre les dynamiques du marché, anticiper les tendances et optimiser leurs stratégies de publication.

1.1. Objectifs du projet

- Centraliser les données de ventes de mangas provenant de différentes sources
- Permettre des analyses multidimensionnelles sur les ventes (par manga, auteur, genre, démographie, éditeur)
- Faciliter la prise de décision stratégique grâce à des tableaux de bord Power BI

1.2. Questions métier

Nous avons structuré notre analyse autour de 5 questions stratégiques auxquelles notre tableau de bord final apporte des réponses chiffrées :

Q1. Performance des Éditeurs :

Au-delà du volume total, quels sont les éditeurs les plus rentables en termes de ventes moyennes par tome ? L'objectif est d'identifier qui détient les licences les plus performantes unitairement.

Q2. Dominance Démographique :

Quelle est la répartition réelle du marché entre les différentes cibles éditoriales (Shonen, Seinen, Shojo) ? Pour savoir où se situe le cœur du marché des best-sellers.

Q3. Impact de la Longueur :

La durée d'une série (classée en Court, Moyen, Long) influence-t-elle la qualité perçue (Score moyen) par les lecteurs ? Une série qui s'étire en longueur perd-elle en qualité ?

Q4. Corrélation Qualité / Ventes :

Existe-t-il un lien direct entre la note critique d'un manga et son volume de ventes total ? Les best-sellers sont-ils nécessairement les mieux notés ?

Q.5 Leaders du marché (Auteurs) :

Qui sont les auteurs incontournables qui cumulent le plus de ventes historiques ?

2. Sources de données

2.1. Description des sources

Les données proviennent de deux datasets publics issus de Kaggle, basés sur les données de [MyAnimeList](#) :

Source 1	Kawaii Dataset
URL	https://www.kaggle.com/datasets/joshjms/kawaii
Format	3 fichiers CSV (manga.csv, genre.csv, manga_genre.csv)
Volume	17 562 entrées de mangas
Contenu	Métadonnées complètes : titres (japonais et anglais), auteurs, genres multiples, scores d'évaluation, nombre de chapitres, statut de publication

Source 2	Best Selling Manga
URL	https://www.kaggle.com/datasets/drahulsingh/best-selling-manga
Format	CSV
Volume	205 entrées (top des best-sellers)
Contenu	Données commerciales : volumes vendus en millions, ventes moyennes par tome, éditeurs, cibles démographiques (Shonen, Seinen, Shoyo, etc.)

2.2. Structure détaillée des sources

Structure de la Source 1 (Kawaii Dataset) :

Le dataset Kawaii est composé de 3 fichiers CSV distincts :

- 1. manga.csv** : Fichier principal avec 13 colonnes contenant les métadonnées descriptives des mangas (id, titres, auteur, chapitres, scores, dates de publication)
- 2. genre.csv** : Table de référence avec 3 colonnes listant tous les genres/thèmes existants (id, name, count)
- 3. manga_genre.csv** : Matrice de 79 colonnes avec des indicateurs binaires (0/1) pour chaque genre, thème et démographie (Action, Romance, Seinen, Shonen, etc.) permettant l'association many-to-many

Fichier 1 : manga.csv (13 colonnes)

Colonne	Type	Description
title	VARCHAR	Titre original japonais
title_english	VARCHAR	Titre traduit en anglais
genres	VARCHAR	Liste de genres séparés par virgules
authors	VARCHAR	Nom(s) des auteurs/mangakas
score	DOUBLE	Note moyenne sur 10 (MyAnimeList)
Chapters	INT	Nombre total de chapitres
status	VARCHAR	État de publication (Finished, Publishing, Hiatus)
volumes	INT	Nombre total de volumes (si terminé)
publishing_start	DATE	Date de début de publication

Fichier 2 : genre.csv (3 colonnes)

Colonne	Type	Description
id	INT	Identifiant unique du genre
name	VARCHAR	Nom du genre (Action, Romance, Seinen, Shounen, etc.)
count	INT	Nombre de mangas associés à ce genre

Fichier 3 : manga_genre.csv (79 colonnes - Matrice binaire)

Colonne	Type	Description
id	INT	Identifiant unique du manga
Action, Adventure, Avant Garde, Award Winning, Boys Love, Comedy, Drama, Fantasy, Girls Love, Gourmet, Horror, Mystery, Romance, Sci-Fi,	BOOLEAN (0/1)	Indicateurs binaires pour chaque genre, thème et démographie. Valeur 1 si le manga appartient au genre, 0 sinon. Inclut : genres (Action, Romance, etc.), thèmes (Isekai, Mecha, etc.) et

Slice of Life, Sports, Supernatural, Suspense, Ecchi, Erotica, Hentai, Adult Cast, Anthropomorphic, CGDCT, Childcare, Combat Sports, Crossdressing, Delinquents, Detective, Educational, Gag Humor, Gore, Harem, High Stakes Game, Historical, Idols (Female), Idols (Male), Isekai, Iyashikei, Love Polygon, Magical Sex Shift, Mahou Shoujo, Martial Arts, Mecha, Medical, Memoir, Military, Music, Mythology, Organized Crime, Otaku Culture, Parody, Performing Arts, Pets, Psychological, Racing, Reincarnation, Reverse Harem, Romantic Subtext, Samurai, School, Showbiz, Space, Strategy Game, Super Power, Survival, Team Sports, Time Travel, Vampire, Video Game, Villainess, Visual Arts, Workplace, Josei, Kids, Seinen, Shoujo, Shounen		démographies (Seinen, Shounen, Shoujo, Josei, Kids)
---	--	---

Structure de la Source 2 (Best Selling Manga) :

Colonne	Type	Description
manga_title	VARCHAR	Titre du manga
no_of_volumes	INT	Nombre de volumes publiés
publisher	VARCHAR	Maison d'édition
demographic	VARCHAR	Cible démographique (Shonen, Seinen, etc.)

approximate_sales_in_million	DOUBLE	Ventes totales approximatives (millions)
average_sales_per_volume	DOUBLE	Ventes moyennes par tome (millions)
serialized	VARCHAR	Période de publication (ex: "1997–present")
author	VARCHAR	Auteur principal

3. Processus ETL

L'architecture de l'ETL repose sur une approche en deux temps : une alimentation d'une zone de préparation (SAS - Staging Area System) suivie de l'alimentation de l'entrepôt de données final (DWH). L'orchestration globale est assurée par des Jobs Pentaho (.kjb) exécutant séquentiellement les transformations (.ktr).

3.1. Initialisation et Zone de Préparation (SAS)

Avant tout chargement, la structure des tables est initialisée au besoin via des scripts SQL exécutés par les transformations `CREATE_SAS_TABLE.ktr` et `CREATE_DWH_TABLE.ktr`.

Extraction vers le SAS :

L'extraction initiale s'effectue à partir des fichiers plats (CSV) du dataset Kawaii (manga.csv, genre.csv, manga_genre.csv) et du dataset Best Selling Manga.

Méthode : Les transformations (ex: `LOAD_SAS_MANGA.ktr`, `LOAD_SAS_GENRE.ktr`) utilisent l'étape "CSV file input" pour lire les fichiers sources.

Chargement SAS :

Ces données sont insérées brutes (ou avec un typage minimal) dans les tables de la zone de préparation (SAS) via l'étape "Table output". Cela permet de découpler la source du traitement complexe.

3.2. Transformation et Chargement des Dimensions (SAS vers DWH)

Contrairement à une extraction directe depuis les CSV, les dimensions sont alimentées en lisant les données depuis les tables du SAS (étape "Table Input" exécutant des requêtes SQL `SELECT colonne1, colonne2... FROM ...`).

Nettoyage et Normalisation :

Les opérations de nettoyage (suppression des nulls, formatage des dates, standardisation des chaînes) sont effectuées lors du passage du SAS vers les dimensions.

Dimensions Simples (Auteur, Éditeur, Genre, Démographie) :

Les données sont extraites du SAS, dédoublonnées via les étapes “Sort rows” et “Unique rows”, puis chargées dans les tables de dimension correspondantes (DIM_Author, DIM_Publisher, etc.).

Dimension Complexe (DIM_Manga) :

La transformation `LOAD_DIM_Manga.ktr` effectue la jointure des données techniques (du de la table `manga` stocké dans le SAS) avec les données de ventes (de `best_selling` stocké dans le SAS).

Des étapes de “Database lookup” sont utilisées pour récupérer les clés étrangères ou vérifier l'existence des enregistrements.

3.3. Chargement de la Table de Faits (FACT)

Le chargement de la table de faits `FACT_Manga_Sales` intervient en dernière étape, une fois toutes les dimensions à jour.

Récupération des Clés (Surrogate Keys) :

La transformation `LOAD_FACT_Manga.ktr` lit les données consolidées depuis le SAS. Elle utilise des étapes de “Database lookup” (Recherche dans base de données) pour interroger les tables de dimension (DIM_Manga, DIM_Author, etc.) et récupérer les clés techniques (`manga_id`, `author_id`) correspondant aux clés métier.

Calcul des Métriques :

Les mesures comme le score, les ventes approximatives et le nombre de volumes sont formatées et insérées avec les clés étrangères dans la table de faits.

4. Architecture du Data Warehouse

4.1. Modèle en flocon (Snowflake Schema)

Pour ce projet, nous avons adopté une architecture en flocon (Snowflake Schema). Ce modèle est une variation du schéma en étoile où certaines dimensions sont normalisées en sous-dimensions. Dans notre cas, la dimension Genre est normalisée et reliée à la dimension Manga plutôt qu'à la table de faits, créant ainsi la structure caractéristique du flocon.

La Table de Faits :

`FACT_Manga_Sales` contient les métriques numériques (ventes, scores, volumes, chapitres) ainsi que les clés étrangères vers toutes les dimensions de premier niveau.

Les Dimensions de premier niveau :

DIM_Manga, DIM_Publisher, DIM_Author et DIM_Demographic sont directement reliées à la table de faits par des relations Many-to-One.

La Dimension de second niveau :

La relation entre DIM_Manga et DIM_Genre est de type Plusieurs-à-Plusieurs (N,N) : un titre peut posséder plusieurs genres et inversement. Pour gérer cette cardinalité dans un schéma en étoile, nous avons implémenté une table associative nommée Manga_Genre.

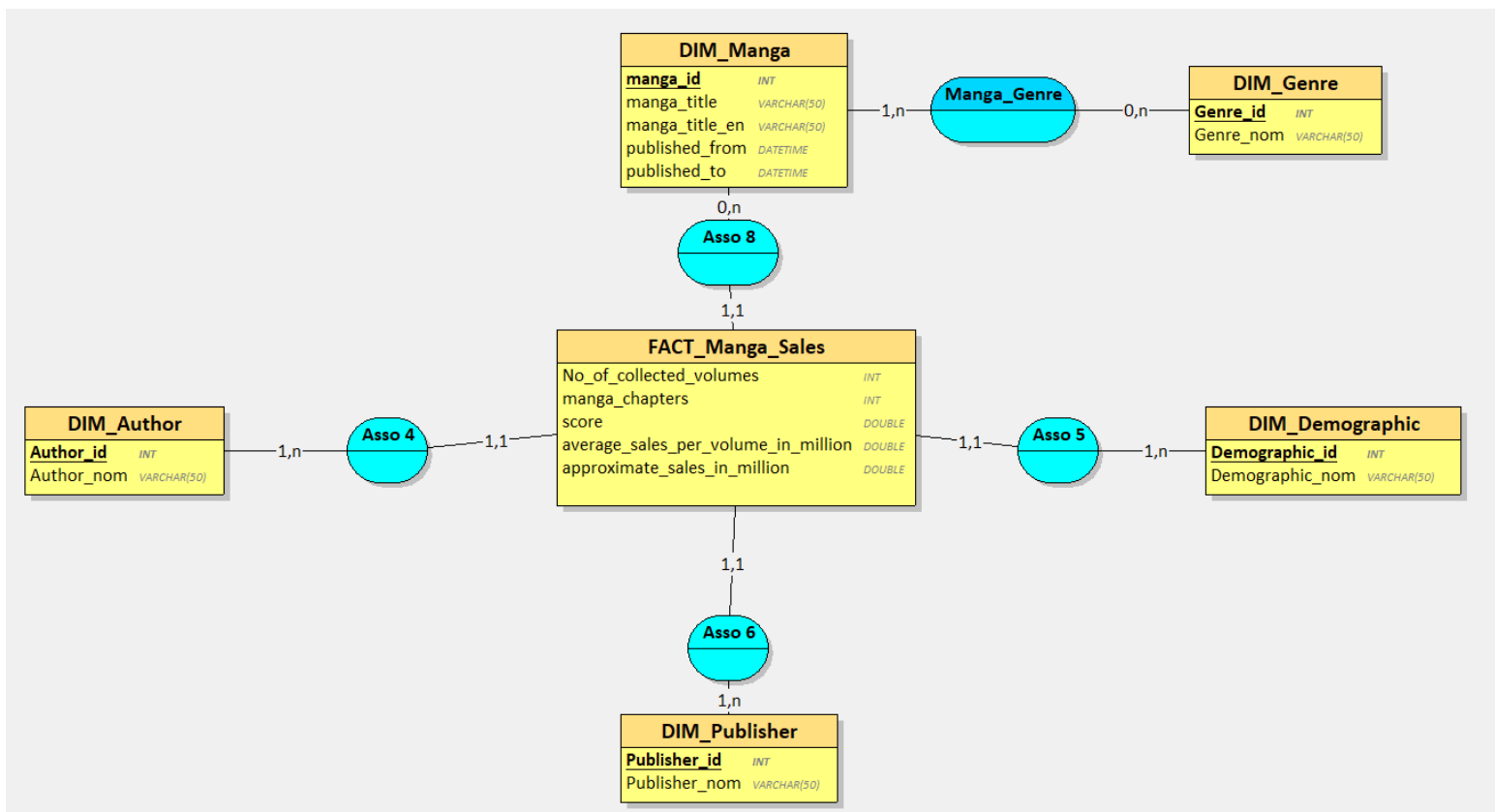
Rôle technique : Elle sert de pivot en stockant uniquement les paires de clés primaires (manga_id, genre_id).

Intégrité des données : Cette structure évite de lier le genre directement à la table de faits (FACT_Manga_Sales). Une liaison directe provoquerait une multiplication artificielle des lignes de ventes pour les mangas multi-genres, faussant ainsi les agrégats financiers (doublons de chiffre d'affaires).

Granularité : La table de faits reste ainsi focalisée sur l'unité de vente (le Manga), tandis que la table Manga_Genre permet de ventiler ces ventes par thématique lors des jointures analytiques.

4.2. Schéma relationnel

Schéma du modèle en flocon



Relations Facts ↔ Dimensions :

La table FACT_Manga_Sales possède quatre clés étrangères : manga_id, Author_id, Publisher_id et Demographic_id. Chaque relation est de type Many-to-One (plusieurs enregistrements de ventes peuvent référencer le même manga, le même auteur, etc.).

4.3. Justification du modèle en flocon

Normalisation et Intégrité (3NF) :

Les genres (Action, Aventure, Fantasy, Romance, etc.) sont des libellés textuels répétitifs. En les isolant dans DIM_Genre, on évite la redondance et on facilite les mises à jour. Un changement de libellé se fait en un seul endroit.

Optimisation des requêtes analytiques :

Les filtres par genre sont fréquents dans les analyses. Avec une table dédiée indexée, les jointures sont efficaces même sur de gros volumes. PostgreSQL optimise automatiquement ces jointures.

Évolutivité :

L'ajout de nouveaux genres ne nécessite qu'une insertion dans DIM_Genre, sans modification de la structure de FACT_Manga_Sales ou de DIM_Manga.

Maintenabilité :

La séparation des responsabilités rend le modèle plus compréhensible. Les développeurs et analystes identifient rapidement que Genre est une propriété du Manga, pas de la vente elle-même.

5. Description détaillée des tables

5.1. Table de faits : FACT_Manga_Sales

Table centrale contenant les métriques de ventes et les clés étrangères vers toutes les dimensions de premier niveau. Chaque ligne représente une mesure de performance pour un manga spécifique.

Colonne	Type	Description
manga_id (FK)	INT	Clé étrangère vers DIM_Manga
Author_id (FK)	INT	Clé étrangère vers DIM_Author
Publisher_id (FK)	INT	Clé étrangère vers DIM_Publisher
Demographic_id (FK)	INT	Clé étrangère vers DIM_Demographic
No_of_collected_volumes	INT	Nombre de volumes collectés/publiés

manga_chapters	INT	Nombre total de chapitres
score	DOUBLE	Score d'évaluation (0-10)
average_sales_per_volume	DOUBLE	Ventes moyennes par volume (millions)
approximate_sales	DOUBLE	Ventes totales approximatives (millions)

Note : Les lignes avec fond orange clair représentent les clés étrangères (FK) vers les tables de dimension.

5.2. Tables de dimension

DIM_Manga

Dimension contenant les informations descriptives sur chaque manga. Cette table possède une clé étrangère vers DIM_Genre, créant la structure en flocon.

Colonne	Type	Description
manga_id (PK)	INT	Clé primaire, identifiant unique
manga_title	VARCHAR(200)	Titre original du manga
manga_title_en	VARCHAR(200)	Titre anglais du manga

DIM_Genre

Dimension de second niveau contenant les genres de mangas. Cette table est normalisée et reliée à DIM_Manga, formant la caractéristique du modèle en flocon.

Colonne	Type	Description
Genre_id (PK)	INT	Clé primaire, identifiant unique
Genre_nom	VARCHAR(100)	Nom du genre (Action, Romance, etc.)

DIM_Author

Dimension contenant les informations sur les auteurs/mangakas. Permet d'identifier les créateurs les plus prolifiques et performants.

Colonne	Type	Description
Author_id (PK)	INT	Clé primaire, identifiant unique
Author_nom	VARCHAR(150)	Nom complet de l'auteur

DIM_Publisher

Dimension référençant les maisons d'édition. Essentielle pour comparer les stratégies éditoriales et parts de marché.

Colonne	Type	Description
Publisher_id (PK)	INT	Clé primaire, identifiant unique
Publisher_nom	VARCHAR(150)	Nom de la maison d'édition

DIM_Demographic

Dimension identifiant la cible démographique (Shonen, Seinen, Shojo, Josei, Kodomo). Cruciale pour les analyses de segmentation marché.

Colonne	Type	Description
Demographic_id (PK)	INT	Clé primaire, identifiant unique
Demographic_nom	VARCHAR(50)	Catégorie démographique cible

Manga_Genre

Table de relation permettant de relier les mangas et leur genre associé.

Colonne	Type	Description
Manga_id	INT	Clé étrangère vers DIM_Manga
Genre_id	INT	Clé étrangère vers DIM_Genre

6. Visualisation et reporting avec Power BI

6.1. Mise en place de Power BI

Connexion au Data Warehouse :

Power BI Desktop a été connecté directement à PostgreSQL en mode DirectQuery. Ce mode garantit que les visualisations affichent toujours les données les plus récentes sans nécessiter de rafraîchissement manuel. La chaîne de connexion pointe vers la base de données PostgreSQL hébergeant notre Data Warehouse.

Détection automatique des relations :

Power BI a automatiquement détecté les relations entre les tables grâce aux clés étrangères définies dans PostgreSQL. Le modèle en flocon (FACT_Manga_Sales → DIM_Manga → DIM_Genre) a été correctement identifié et représenté dans le diagramme de modèle de Power BI.

Création de mesures DAX :

Des mesures DAX (Data Analysis Expressions) ont été créées pour calculer des métriques complexes : ventes moyennes par tome, nombre total de mangas, note moyenne pondérée, et catégorisation de la longueur des séries (Court, Moyen, Long basé sur le nombre de volumes).

6.2. Dashboard développé

Le dashboard "Performance des ventes de Manga" a été conçu pour répondre directement aux 5 questions métier stratégiques identifiées. Il combine plusieurs types de visualisations pour offrir une vue d'ensemble complète et des analyses détaillées.

Dashboard Power BI - Performance des ventes de Manga



6.3. Analyse des résultats

Réponse à la Question 1 - Performance des Éditeurs :

Le graphique en barres "Top 4 éditeurs les plus rentables" révèle que Shueisha domine avec 57 millions de ventes moyennes par tome, suivi de Kodansha (31M), Shogakukan (23M) et Akita Shoten (7M). Cette visualisation permet d'identifier clairement les éditeurs qui détiennent les licences les plus performantes unitairement.

Réponse à la Question 2 - Dominance Démographique :

Le graphique circulaire montre que le Shonen représente 73,2% des ventes totales, confirmant sa position ultra-dominante. Le Seinen arrive en seconde position avec 16,98%, tandis que le Shojō ne représente que 8,14%. Ces chiffres confirment que le "cœur" du marché des best-sellers se situe clairement dans la démographie Shonen.

Réponse à la Question 3 - Impact de la Longueur :

Le graphique "Moyenne de score par Longueur" révèle que les séries de longueur moyenne (Moyen) obtiennent le meilleur score (8,21), suivies des séries longues (8,06) et courtes (7,87). Contrairement à l'hypothèse initiale, les séries longues ne perdent pas significativement en qualité perçue, avec seulement 0,15 points d'écart avec les séries moyennes.

Réponse à la Question 4 - Corrélation Qualité / Ventes :

Le nuage de points "Répartition des notes par ventes" montre une absence de corrélation forte entre le score et les ventes totales. On observe des mangas avec un score de 9/10 vendant entre 100 et 500 millions d'exemplaires, tandis que certains titres avec des scores autour de 8/10 atteignent également des ventes très élevées. Les best-sellers ne sont donc pas nécessairement les mieux notés.

Réponse à la Question 5 - Leaders du marché (Auteurs) :

Le tableau et le graphique "Top 9 des auteurs ayant faits le plus de vente" identifient clairement les auteurs incontournables. Le tableau détaillé permet de voir non seulement les ventes totales mais aussi les scores moyens de leurs œuvres, offrant une vue complète de leur performance commerciale et critique.

6.4. Interactivité et fonctionnalités

Filtres temporels :

Deux filtres de période permettent d'analyser les ventes sur des plages de dates spécifiques (22/04/1946 à 03/12/2018 dans l'exemple). Cette fonctionnalité est essentielle pour détecter les tendances temporelles.

Filtres démographiques et éditeurs :

Des menus déroulants permettent de filtrer par démographie et par éditeur. Ces filtres sont interconnectés avec toutes les visualisations, permettant des analyses croisées dynamiques.

KPIs en temps réel :

Quatre cartes KPI affichent les métriques clés : 77 mangas analysés, 1,60 millions de ventes moyennes par tome, note moyenne de 8,03/10, et volume total de 4,52K millions de ventes.

Cross-filtering :

Un clic sur n'importe quel élément visuel (barre, segment du graphique circulaire, ligne du tableau) filtre automatiquement toutes les autres visualisations du dashboard, permettant une exploration intuitive des données.

7. Conclusion

Ce projet de Data Warehouse pour l'analyse des ventes de mangas démontre la mise en œuvre complète d'un système décisionnel basé sur un modèle en flocon.

L'architecture adoptée garantit à la fois la normalisation des données (dimension Genre de second niveau) et les performances requises pour les analyses multidimensionnelles complexes.

Le projet a permis de répondre concrètement aux 5 questions métier stratégiques grâce au dashboard Power BI développé. Les résultats révèlent des insights actionnables : la domination écrasante du Shonen (73% des ventes), l'absence de corrélation forte entre qualité perçue et succès commercial, et l'identification des éditeurs les plus performants unitairement.

Perspectives d'évolution :

Enrichissement avec données mensuelles pour analyses temporelles fines, intégration d'algorithmes de machine learning pour prévisions de ventes, automatisation complète du pipeline ETL avec orchestration, extension géographique pour comparer les marchés internationaux.

Ce projet a permis d'acquérir une maîtrise pratique du cycle complet de création d'un Data Warehouse professionnel, de la collecte des sources à la visualisation finale, en passant par la modélisation dimensionnelle, l'ETL, et l'optimisation des performances. Les compétences développées sont directement transférables à tout projet décisionnel d'entreprise.