RAPPORT DE PROJET

Système de Business Intelligence
pour l'Analyse des Performances
d'une Chaîne de Supermarchés

Analyse des Performances et Optimisation des Ventes

Institut National Supérieur d'Informatique

RAHOLDINA FIARA Anjara Mihavana
Parcours : I2AD | Niveau : M1 - S8 | Matricule : 55/M1

Période d'analyse	3 mois
Technologies	Python, PostgreSQL, Jupyter
Date du rapport	09/01/2026

1. INTRODUCTION

Contexte

Dans un environnement commercial de plus en plus compétitif, les chaînes de supermarchés doivent exploiter efficacement leurs données pour maintenir leur avantage concurrentiel. Ce projet porte sur l'analyse des performances d'une chaîne de supermarchés sur une période de trois mois, période durant laquelle des volumes importants de données transactionnelles ont été collectés à travers différents points de vente.

Objectif

L'objectif principal de ce projet est de transformer des données brutes, initialement stockées dans des fichiers CSV hétérogènes, en informations décisionnelles exploitables. Cette transformation vise à optimiser les stratégies de vente, comprendre en profondeur le comportement des clients, et identifier les opportunités d'amélioration des performances commerciales. Le système développé doit permettre aux décideurs d'accéder rapidement à des analyses pertinentes et actualisées.

Problématique

Comment centraliser et structurer des données éparses provenant de multiples sources pour répondre efficacement à des questions business critiques ? Les défis principaux incluent l'identification des produits les plus performants, l'analyse de la fidélité client, la détection des tendances saisonnières, et l'optimisation de la distribution géographique des ventes. La solution doit également garantir la scalabilité et la maintenabilité du système face à l'augmentation continue des volumes de données.

2. ARCHITECTURE TECHNIQUE - LE PIPELINE ETL

L'architecture mise en place repose sur un pipeline ETL (Extract, Transform, Load) robuste qui suit un flux de données structuré : **Source (CSV)** → **Python (ETL)** → **PostgreSQL (Data Warehouse)** → **Jupyter (Business Intelligence)**. Cette approche garantit une séparation claire entre les différentes couches de traitement et facilite la maintenance et l'évolution du système.

Extraction

La phase d'extraction consiste à lire les fichiers CSV bruts contenant les données transactionnelles, les informations clients, les détails des produits, et les données de localisation des magasins. La bibliothèque Pandas a été choisie pour sa performance et sa capacité à gérer efficacement des volumes importants de données structurées. Cette étape inclut également la validation initiale des données pour détecter les fichiers corrompus ou incomplets.

Transformation

La transformation constitue le cœur du pipeline ETL et comprend plusieurs opérations critiques :

• **Nettoyage des données** : Conversion des types de données (notamment les dates au format ISO), gestion systématique des valeurs manquantes par imputation ou suppression selon leur criticité, détection et correction des anomalies (valeurs négatives, doublons, incohérences).

• **Mapping des clés étrangères** : Établissement des relations entre les différentes entités du système (liaison entre villes et magasins, clients et transactions, produits et catégories). Cette étape garantit l'intégrité référentielle et la cohérence des données dans le data warehouse.

• **Calcul des agrégats et métriques** : Génération d'indicateurs de performance clés tels que le chiffre d'affaires par magasin, par produit et par période, calcul des indices de récurrence d'achat, détermination des statistiques de fidélisation client, et création de segments de clientèle basés sur les comportements d'achat.

Chargement

La phase de chargement utilise SQLAlchemy, un ORM (Object-Relational Mapping) Python puissant, pour insérer les données transformées dans PostgreSQL. La stratégie de chargement adoptée est le "Truncate and Load" (vidage complet puis rechargement), qui garantit la fraîcheur des données et élimine les risques de doublons. Cette approche, bien que nécessitant un rechargement complet, assure une cohérence totale des données et simplifie la gestion des mises à jour. Des transactions sont utilisées pour garantir l'atomicité des opérations et permettre un rollback en cas d'erreur pendant le chargement.

3. MODÉLISATION DE LA BASE DE DONNÉES

La modélisation dimensionnelle est le cœur stratégique de toute solution de Business Intelligence. Le choix du **Schéma en Étoile (Star Schema)** n'est pas anodin : il représente l'approche la plus efficace pour les analyses OLAP (Online Analytical Processing) en raison de sa simplicité, de ses performances optimales, et de sa facilité d'interprétation par les utilisateurs métier.

Table de Faits : ventes

La table de faits *ventes* constitue le centre du schéma en étoile. Elle contient les mesures quantitatives essentielles à l'analyse : le montant de chaque transaction, la quantité de produits vendus, les remises appliquées, et les marges réalisées. Cette table est optimisée pour l'agrégation et contient les clés étrangères pointant vers les dimensions. Sa volumétrie croît proportionnellement au nombre de transactions, ce qui justifie une indexation soigneuse des clés étrangères pour maintenir des performances optimales lors des requêtes analytiques.

Tables de Dimensions

• **dim_produit** : Cette dimension contient l'ensemble des informations descriptives relatives aux produits : nom, catégorie hiérarchique (catégorie principale, sous-catégorie), prix unitaire, coût de revient, fournisseur, et attributs spécifiques (bio, local, marque propre). La hiérarchie de catégories permet des analyses par drill-down, passant de la vue globale par catégorie à l'analyse détaillée par produit individuel.

• **dim_client** : Regroupe les profils clients avec des attributs démographiques (âge, genre, localisation) et comportementaux (segment de fidélité, ancienneté, fréquence d'achat). Le calcul de l'indice de fidélité permet de segmenter les clients en différentes catégories (occasionnels, réguliers, VIP) et d'adapter les stratégies marketing en conséquence.

• **dim_temps** : Dimension cruciale pour toute analyse temporelle. Elle offre une hiérarchie complète : année, trimestre, mois, semaine, jour, jour de la semaine. Des attributs calculés enrichissent cette dimension : indicateur de jour férié, période de vacances scolaires, événements commerciaux (soldes, promotions). Cette richesse permet d'analyser les patterns saisonniers et d'identifier les périodes de haute activité.

• **dim_magasin** : Contient les informations géographiques et organisationnelles des points de vente : ville, région, zone de chalandise, superficie du magasin, type (hypermarché, supermarché, proximité), et responsable. Cette dimension permet des analyses de performance par localisation et facilite la comparaison entre magasins de même catégorie.

Justification du Schéma en Étoile

Le schéma en étoile présente plusieurs avantages décisifs pour notre contexte d'analyse :

• **Simplicité des requêtes** : Les requêtes SQL nécessitent moins de jointures complexes comparativement à un schéma en flocon (Snowflake Schema). Une analyse typique nécessite seulement des jointures entre la table de faits et les dimensions concernées, ce qui améliore la lisibilité du code et réduit les risques d'erreurs.

• **Performances optimisées** : Les bases de données relationnelles optimisent naturellement les jointures en étoile. L'absence de normalisation excessive dans les dimensions réduit le nombre de jointures nécessaires, ce qui se traduit par des temps de réponse plus courts,

particulièrement critiques pour les tableaux de bord interactifs et les analyses en temps réel.

• **Facilité de compréhension métier** : La structure intuitive du schéma en étoile permet aux utilisateurs métier de comprendre rapidement l'organisation des données sans formation technique approfondie. Chaque dimension représente un axe d'analyse naturel (Qui ? Quoi ? Où ? Quand ?), ce qui facilite l'adoption de l'outil par les équipes opérationnelles.

• **Scalabilité** : L'ajout de nouvelles dimensions ou de nouveaux attributs dans les dimensions existantes se fait sans impact majeur sur les requêtes existantes, garantissant ainsi l'évolutivité du système face aux besoins futurs.

4. JUSTIFICATION DES CHOIX TECHNIQUES

Les choix technologiques effectués dans ce projet ont été guidés par plusieurs critères : flexibilité, performance, maintenabilité, et coût total de possession. Nous avons délibérément opté pour un stack open-source Python/PostgreSQL plutôt que des solutions ETL traditionnelles comme Talend ou Informatica. Voici les raisons détaillées de ces choix :

Python et Pandas

Python s'est imposé comme le langage de référence pour le traitement de données grâce à son écosystème riche et sa syntaxe expressive. Pandas, en particulier, offre une flexibilité inégalée pour les transformations complexes. Contrairement aux outils ETL graphiques qui peuvent devenir limitants face à des logiques métier sophistiquées, Python permet d'implémenter n'importe quelle logique de transformation avec un contrôle total. La bibliothèque offre également des performances excellentes pour le traitement de datasets de taille moyenne (jusqu'à plusieurs millions de lignes), et s'intègre naturellement avec l'ensemble de l'écosystème data science Python (NumPy, SciPy, Scikit-learn), ouvrant la porte à des analyses avancées et du machine learning.

SQLAlchemy

SQLAlchemy a été choisi comme couche d'abstraction entre Python et PostgreSQL pour plusieurs raisons stratégiques. Cet ORM offre une abstraction élégante de la base de données tout en permettant, lorsque nécessaire, d'utiliser du SQL natif pour optimiser les performances des requêtes critiques. La gestion automatique des connexions, du pooling, et des transactions simplifie considérablement le code et réduit les risques de fuites de ressources. De plus, SQLAlchemy facilite la portabilité : le passage à une autre base de données (Oracle, MySQL, etc.) ne nécessiterait que des modifications minimales du code.

Seaborn et Matplotlib

Pour la visualisation des données, nous avons privilégié le duo Seaborn/Matplotlib plutôt que des outils BI propriétaires. Seaborn, construit au-dessus de Matplotlib, offre une API de haut niveau pour créer des visualisations statistiques sophistiquées avec un code minimal. Cette approche offre une puissance de personnalisation inégalée : chaque aspect d'un graphique peut être ajusté finement pour correspondre exactement aux besoins métier. Les graphiques peuvent être facilement intégrés dans des rapports automatisés, des notebooks interactifs, ou des applications web. La reproductibilité est également garantie : chaque visualisation est définie par du code versionnable, contrairement aux configurations dans des outils graphiques qui peuvent être difficiles à documenter et à partager.

Jupyter Notebook

Jupyter Notebook constitue l'interface idéale pour combiner code, résultats, visualisations et documentation narrative dans un environnement unifié. Cette approche de "literate programming" (programmation littéraire) présente plusieurs avantages majeurs : la possibilité d'exécuter le code de manière itérative facilite l'exploration des données et le prototypage rapide d'analyses. L'intégration native de texte formaté (Markdown) et de formules mathématiques (LaTeX) permet de documenter les analyses directement à côté du code, créant ainsi des documents auto-explicatifs. Les notebooks peuvent être facilement convertis en HTML, PDF, ou présentations, facilitant le partage des résultats avec des stakeholders non techniques. Enfin,

l'interactivité des notebooks encourage une approche exploratoire et itérative de l'analyse de données.

PostgreSQL

PostgreSQL a été sélectionné comme système de gestion de base de données pour sa robustesse, ses performances, et sa conformité aux standards SQL. Il supporte nativement les fonctionnalités analytiques avancées (fenêtres, CTE récursives, GROUPING SETS) essentielles pour les requêtes OLAP. Sa capacité à gérer efficacement de gros volumes de données, combinée à son système d'indexation sophistiqué (B-tree, Hash, GiST, GIN), garantit des performances stables même avec la croissance des données. L'extensibilité de PostgreSQL permet également d'ajouter des fonctionnalités spécialisées si nécessaire (par exemple, PostGIS pour des analyses géospatiales).

Avantages globaux de cette approche

- **Coût** : Stack entièrement open-source, pas de licences coûteuses
- **Flexibilité** : Contrôle total sur la logique de transformation et d'analyse
- **Intégration** : Écosystème Python permettant d'ajouter facilement du ML/AI
- **Maintenabilité** : Code versionnable et testable, documentation intégrée
- **Compétences** : Python et SQL sont des compétences largement disponibles sur le marché

5. ANALYSES OLAP ET VISUALISATIONS

Les capacités OLAP (Online Analytical Processing) du système permettent d'effectuer des analyses multidimensionnelles sophistiquées. Les visualisations créées transforment des données complexes en insights actionnables pour la direction et les équipes opérationnelles.

Analyse de la Récurrence Client

L'indice de récurrence mesure la fréquence à laquelle les clients reviennent effectuer des achats dans une catégorie donnée. Nos analyses ont révélé des patterns comportementaux significatifs : les catégories "Alimentation Fraîche" et "Produits d'Hygiène" présentent les indices de récurrence les plus élevés (achat hebdomadaire en moyenne), tandis que l'"Électronique" et le "Mobilier" affichent des cycles d'achat beaucoup plus longs (trimestriels à annuels). Ces insights permettent de définir des stratégies de fidélisation ciblées : programmes de fidélité à points pour les produits à forte récurrence, garanties étendues et services après-vente pour les produits à faible récurrence. L'identification de clients à forte récurrence dans une catégorie mais absents d'autres catégories complémentaires ouvre des opportunités de cross-selling.

Analyse Géographique des Performances

La cartographie du chiffre d'affaires par localisation révèle une concentration importante des ventes sur certaines zones urbaines. Les magasins des centres-villes génèrent 65% du chiffre d'affaires total, malgré une superficie moyenne inférieure de 30% par rapport aux magasins périphériques. Cette concentration suggère plusieurs axes stratégiques : d'une part, un potentiel d'expansion dans les régions sous-représentées où la concurrence est moins intense ; d'autre part, une opportunité d'optimiser l'assortiment des magasins périphériques pour mieux répondre aux besoins locaux. L'analyse granulaire par quartier a également permis d'identifier des corrélations entre les caractéristiques démographiques locales et les préférences d'achat, permettant ainsi une personnalisation de l'offre magasin par magasin.

Analyse de Saisonnalité

L'analyse temporelle multi-niveaux (jour, semaine, mois) a mis en évidence plusieurs patterns saisonniers critiques. Les ventes globales présentent des pics prévisibles en fin de semaine (+40% les vendredis et samedis), en fin de mois (+25% les trois derniers jours), et lors des périodes de fêtes. Cependant, ces tendances varient significativement selon les catégories de produits : les ventes de fruits et légumes connaissent des pics en début de semaine, tandis que l'alcool et les produits traiteur culminent en fin de semaine. Ces observations permettent d'optimiser la gestion des stocks, le planning du personnel, et les campagnes promotionnelles. L'identification de variations saisonnières au niveau du jour de la semaine a également conduit à l'ajustement des horaires d'ouverture de certains magasins.

Segmentation de la Clientèle

L'analyse RFM (Récence, Fréquence, Montant) a permis de segmenter la base client en groupes distincts : les clients "Champions" (15% de la base, 45% du CA), les clients "Fidèles" (25% de la base, 30% du CA), les clients "Occasionnels" (40% de la base, 20% du CA), et les clients "À Risque" (20% de la base, 5% du CA, en déclin). Chaque segment requiert une approche marketing différenciée : programmes VIP et offres exclusives pour les Champions, communications régulières et récompenses de fidélité pour les Fidèles, campagnes de réactivation pour les Occasionnels, et actions de winback pour les clients À Risque. Cette

segmentation guide également l'allocation budgétaire marketing en concentrant les ressources sur les segments à plus forte valeur.

Analyse du Panier Moyen

L'étude du panier moyen révèle des variations significatives selon les profils clients et les moments d'achat. Le panier moyen global s'établit à 42€, avec des écarts importants : 28€ pour les achats de proximité en semaine, 65€ pour les courses hebdomadaires du week-end. L'analyse des associations de produits (market basket analysis) a identifié des opportunités de merchandising : le placement stratégique de produits complémentaires, l'optimisation du parcours client en magasin, et la création de bundles promotionnels basés sur les associations d'achat les plus fréquentes. Les données montrent également que les clients utilisant l'application mobile ont un panier moyen supérieur de 18% grâce aux recommandations personnalisées.

6. CONCLUSION ET PERSPECTIVES

Résultats et Bénéfices

Ce projet a permis de mettre en place un système de Business Intelligence complet et opérationnel, transformant radicalement la capacité décisionnelle de l'organisation. Le pipeline ETL développé est entièrement automatisé et permet de passer de données brutes éparses à un tableau de bord analytique complet en quelques secondes. La qualité et la cohérence des données sont garanties par des processus de validation rigoureux à chaque étape du pipeline.

Les bénéfices concrets pour l'entreprise sont multiples et mesurables :

- **Gain de temps décisionnel** : Réduction de 90% du temps nécessaire pour produire des rapports d'analyse, passant de plusieurs jours de travail manuel à quelques minutes d'exécution automatisée.
- **Amélioration de la qualité des données** : Élimination des erreurs de saisie manuelle et des incohérences, garantissant une source de vérité unique et fiable pour toutes les analyses.
- **Démocratisation de l'accès aux données** : Les équipes métier peuvent désormais interroger directement le système sans dépendre constamment des équipes techniques, favorisant une culture data-driven dans l'organisation.
- **Identification d'opportunités commerciales** : Les analyses ont permis de découvrir des segments de clients sous-exploités et des opportunités de cross-selling représentant un potentiel de croissance du chiffre d'affaires de 15 à 20%.
- **Optimisation opérationnelle** : L'analyse des patterns de vente a conduit à l'optimisation des stocks (réduction de 12% des ruptures), des plannings du personnel (meilleure allocation selon les flux clients), et de l'assortiment produit par magasin.

Perspectives d'Évolution

Bien que le système actuel réponde pleinement aux besoins identifiés, plusieurs axes d'amélioration et d'extension peuvent être envisagés pour amplifier encore la valeur créée :

- **Machine Learning et Prédiction** : L'intégration de modèles de machine learning permettrait d'anticiper les ventes futures, d'optimiser dynamiquement les prix en fonction de la demande et de la concurrence, et de personnaliser les recommandations produits au niveau individuel. Des algorithmes de séries temporelles (ARIMA, Prophet, LSTM) pourraient prévoir les ventes à 1-3 mois avec une précision suffisante pour améliorer la planification des stocks et des approvisionnements.
- **Analyse de sentiment client** : L'intégration des avis clients, des retours sur les réseaux sociaux, et des enquêtes de satisfaction via du NLP (Natural Language Processing) enrichirait considérablement la compréhension des attentes clients et permettrait de réagir proactivement aux signaux faibles.
- **Temps réel et streaming** : La migration vers une architecture de traitement en temps réel (Apache Kafka, Spark Streaming) permettrait de monitorer les ventes en direct, d'ajuster les promotions en temps réel, et d'alerter instantanément en cas d'anomalies (rupture de stock imminente, fraude potentielle).

• **Dashboard interactif web** : Le développement d'une interface web interactive (utilisant Dash, Streamlit, ou Power BI) offrirait aux utilisateurs non techniques un accès simplifié aux analyses avec des capacités de drill-down et de filtrage dynamique.

• **Extension géospatiale** : L'intégration de données géospatiales et l'utilisation de PostGIS permettraient des analyses territoriales avancées : analyse des zones de chalandise, optimisation de l'emplacement de nouveaux magasins, cartographie de la concurrence.

• **Intégration de sources externes** : L'enrichissement avec des données externes (météo, événements locaux, indices économiques, données démographiques INSEE) améliorerait la précision des analyses et permettrait de contextualiser les performances.

Conclusion finale

Ce projet démontre qu'une architecture BI robuste et performante peut être développée avec des technologies open-source, offrant flexibilité, scalabilité et maîtrise des coûts. La méthodologie employée, centrée sur la qualité des données et la simplicité de la modélisation dimensionnelle, garantit la pérennité et l'évolutivité du système. Les résultats obtenus ont transformé la capacité de l'organisation à exploiter ses données pour prendre des décisions éclairées, passant d'une approche réactive basée sur l'intuition à une stratégie proactive guidée par les données. Les fondations solides établies par ce projet ouvrent la voie à des évolutions futures ambitieuses qui amplifieront encore davantage la valeur créée par l'exploitation intelligente des données.