École Supérieure Privée de Management de Tunis
Esprit School of Business

RAPPORT DE STAGE

En vue de l'obtention du diplôme de Master Professionnel en
Business Analytics

Optimisation des performances commerciales à travers la prédiction du succès des opportunités et la segmentation client

---

Encadrante académique : Rym Besrour
Maître de stage : Belhassen Albouchi
Réalisé par : Mesghouni Ghassen

Soutenu en : Décembre 2025

Année universitaire 2024/2025

Année universitaire 2024/2025		
Maître de stage	Belhassen Albouchi	Date et Signature
Encadrant académique	Rym Besrour	Date et Signature

Résumé

Ce projet construit une chaîne de valeur **Data** $\rightarrow$ **DWH** $\rightarrow$ **ETL** $\rightarrow$ **BI** $\rightarrow$ **ML** pour fiabiliser et piloter le pipeline d'opportunités commerciales chez Huawei Tunisie (CSV $\sim$500 lignes). Un *Data Warehouse* en étoile (fait *opportunité* et dimensions client/produit/secteur/région/statut/temps) est alimenté par un *package* SSIS rejouable (`sp_load_dwh`). Un tableau de bord Power BI fournit des KPI fiables (CA, volume, taux de gagné) et des analyses multi-axes. Un modèle supervisé (régression logistique, CV K=5) estime la *probabilité de gain* par opportunité (AUC $\approx$ 0,52 sur l'échantillon). Les scores sont **réinjectés** dans le DWH (`ml.predictions`, `ml.vw_last_predictions`) et **exposés** dans Power BI pour prioriser l'action commerciale.

Mots-clés : DWH, SSIS, Power BI, régression logistique, prédiction, KPI.

Remerciements

Je souhaite exprimer ma profonde gratitude à toutes les personnes qui, de près ou de loin, ont contribué à la réalisation de ce travail.

Avant tout, j'adresse mes sincères remerciements à Madame **Rym Besrour**, pour la qualité de son encadrement académique, sa rigueur méthodologique et sa bienveillance constante. Ses orientations, ses retours précis et ses exigences justes m'ont permis d'affiner mes idées, de structurer mon travail et d'avancer avec confiance.

Je remercie également Monsieur **Belhassen Albouchi**, maître de stage au sein du **Tunis Finance and Healthcare Sales Department** (Huawei Tunisie), pour son accompagnement sur le terrain, sa disponibilité et ses conseils pragmatiques. Son regard métier et ses retours d'expérience ont été déterminants pour aligner la solution technique avec les besoins opérationnels.

Ma reconnaissance va à l'ensemble des équipes de **Huawei Tunisie** impliquées dans ce projet, pour leur accueil, leur esprit de collaboration et l'accès facilité aux informations nécessaires. Merci en particulier aux collègues qui ont partagé leur connaissance des processus commerciaux, des données et des outils décisionnels.

Je tiens à remercier chaleureusement le jury et le corps enseignant d'**École Supérieure Privée de Management de Tunis**
Esprit School of Business pour l'exigence scientifique et l'accompagnement pédagogique tout au long de la formation. Les cours, ateliers et échanges ont constitué un socle précieux pour la conduite de ce projet et, plus largement, pour ma future vie professionnelle.

Je n'oublie pas mes camarades de promotion pour les partages, les relectures, les discussions tardives et l'entraide qui ont rythmé ces derniers mois. Leur soutien a souvent fait la différence dans les moments d'intense préparation.

Enfin, j'adresse un remerciement tout particulier à ma famille et à mes proches, pour leur patience, leur encouragement inconditionnel et la confiance qu'ils m'ont toujours témoignée. Leur présence m'a donné l'énergie nécessaire pour mener ce travail à son terme.

À toutes et à tous, *merci*.

Dédicace

À ma famille, qui a toujours cru en moi et soutenu chacun de mes choix. À mes parents, pour leurs sacrifices, leurs valeurs et leur affection sans faille ; à mes frères et sœurs, pour leur présence et leurs encouragements au quotidien.

À mes proches et amis, pour leur compréhension dans les moments d'absence, leurs mots rassurants et leur confiance. Vous avez su transformer les difficultés en défis et les étapes en avancées.

Je dédie également ce travail à toutes celles et ceux qui m'ont transmis le goût de l'effort, de la curiosité et du travail bien fait. Puissent ces pages témoigner de leur influence et de ma gratitude.

Ce mémoire est le vôtre autant que le mien.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Au sein de **Huawei Tunisie**, et plus précisément du **Tunis Finance and Healthcare Sales Department**, la performance commerciale dépend de la capacité à **voir** et **comprendre** le pipeline d'opportunités : qui sont nos clients, quelles offres sont en cours, où en est la négociation, quelles régions et quels secteurs portent la croissance, quel chiffre d'affaires est réellement sécurisable. Le constat initial était simple : une donnée **dispersée** et **peu normalisée**, un export **CSV** d'environ **500 lignes** sans historisation ni référentiels stables, des analyses *ad hoc* difficiles à comparer dans le temps, et aucune estimation quantitative de la **probabilité de gain** permettant de prioriser les efforts. La donnée existait, mais elle ne rendait pas encore service au pilotage.

Partant de ce constat, ce projet vise à **mettre la donnée au service de la décision**, de bout en bout, en construisant une chaîne cohérente qui va de la **mise en qualité** à l'**aide à la décision**. La chaîne cible s'articule autour de quatre piliers complémentaires :

1. la structuration dans un **Data Warehouse (DWH)** en étoile ;
2. l'**automatisation** des chargements via **SSIS** (ETL) ;
3. la **restitution** des indicateurs et analyses dans **Power BI** ;
4. l'**enrichissement** par un modèle de **Machine Learning** fournissant une **probabilité de gain** par opportunité, réintégrée au DWH pour être consommée au même titre que les KPI « réels ».

Le fil directeur est clair : **fiabiliser** → **outiller** → **augmenter**.

Problématique et objectifs

La problématique se décline en trois questions :

— **Fiabiliser** la donnée pour produire des KPI stables, comparables dans le temps, traçables et fondés sur des définitions claires.

— **Outiller** l'équipe commerciale avec un tableau de bord **visible** (KPI, tendances, découpes par statut/secteur/région/période) et **actionnable** (filtres, drill-down).

— **Prioriser** l'action grâce à une **probabilité de gain** par opportunité, calculée par un modèle interprétable et **réintégrée** au DWH.

Objectifs opérationnels :

— Mettre en qualité la donnée (nettoyage, typage, clés, intégrité, **dimension temps**).

— Concevoir un **modèle en étoile** centré sur le **fait opportunité** (dimensions : client, produit/service, secteur, région, statut, taille d'entreprise, temps).

— Mettre en place un **package SSIS** rejouable pour alimenter **staging** → **DWH** (procédure `dwh.sp_load_dwh`).

— Construire un **tableau de bord Power BI** avec KPI fiables et analyses multi-axes.

— Entraîner un **modèle de classification** simple, transparent (régression logistique), et **réintégrer** les probabilités prédites dans le DWH.

Approche méthodologique

Le projet suit une conduite **hybride** combinant **Scrum** (sprints, backlog, revues, DoD) pour **organiser** l'avancement et **CRISP-DM** (compréhension métier/données → préparation → modélisation → évaluation → déploiement léger) pour **structurer** la partie analytique/ML. Cette double ossature permet d'avancer par **jalons** :

Sprint 1 (DWH). Exploration des données, dictionnaire, nettoyage, conception **étoile** et mise en place des contraintes d'intégrité.

Sprint 2 (ETL SSIS). Ingestion **CSV → staging**, conversions, mappings, procédure `sp_load_dwh`, contrôles post-lot.

Sprint 3 (BI). Modèle Power BI aligné sur le DWH, **mesures DAX**, visuels KPI & analyses.

Sprint 4 (ML). Préparation des variables, entraînement, **validation croisée**, réintégration des scores et exposition dans Power BI.

Choix technologiques et critères

Les outils retenus ont été sélectionnés pour leur **complémentarité** et leur **adéquation** au contexte :

— **SQL Server/SSMS** pour la robustesse du DWH et l'intégration avec SSIS ;
— **SSIS** pour l'automatisation des flux ETL (connectivité, rejouabilité, gestion d'erreurs) ;
— **Power BI** pour la restitution rapide et interactive (modèle sémantique, DAX) ;
— **Python/Scikit-learn** pour un ML **sobre** et **interprétable**.

Les critères de choix : **interopérabilité**, **traçabilité**, **courbe d'apprentissage** pour les équipes, et capacité d'**industrialisation** progressive (planification, monitoring, MLOps léger).

Périmètre et livrables

Le périmètre couvre la **chaîne complète** :

— **DWH** opérationnel (schéma en étoile, dimensions, contraintes, index de base).
— **ETL SSIS** rejouable (CSV → staging → DWH) avec contrôles de volume et de cohérence.
— **Power BI** (modèle sémantique, **mesures DAX**, page KPI et analyses par statut/-secteur/région/temps).
— **ML** : pipeline scikit-learn, évaluation (ROC/AUC, matrice de confusion), **importances/coefs**, réinjection dans `ml.predictions` et **vue** `ml.vw_last_predictions` pour consommation BI.
— **Documentation** : dictionnaire de données, scripts clés, captures, bonnes pratiques et limites.

Résultats et valeur créée

Les résultats se déclinent sur deux plans : **technique** et **métier**. Techniquement, la donnée **vit** désormais dans un espace **structuré** (DWH), est **alimentée** par un flux **rejouable** (SSIS) et **exposée** de manière **cohérente** (Power BI). Côté métier :

— les **KPI** (CA, nombre d'opportunités, taux de gagné) deviennent comparables et crédibles ;

— les analyses croisées par **statut/secteur/région/temps** révèlent des **patterns** utiles ;

— la **probabilité de gain** apporte un éclairage pour **prioriser** les actions.

Surtout, la **boucle** $ML \rightarrow DWH \rightarrow BI$ permet d'intégrer l'analytique **au cœur** du pilotage, plutôt que comme un artefact isolé.

Contraintes, limites et gestion des risques

Le projet a assumé des **contraintes** : **échantillon réduit** (∼500 lignes), manque d'**historique** riche (hors SCD2 côté client), **chaînage ML → DWH** encore semi-manuel (lot CSV). Ces limites ont été adressées par des **choix sobres** (modèle interprétable, validation croisée) et des **parades** claires : normalisation des libellés, typage strict, contrôles `CHECK`, conversions explicites, référentiels de base, journalisation des chargements, vue « dernière prédiction ». Elles n'annulent pas la valeur, mais **cadent** l'interprétation : l'objectif n'était pas de battre des records, mais de **poser une capacité** fiable et reproductible.

Gouvernance, qualité et perspectives

Au-delà de la technique, le projet a travaillé la **gouvernance** : dictionnaire de données, conventions de nommage, séparation **staging/DWH/ML**, règles de qualité, contrôles post-lot, documentation des mesures DAX. À court terme, les priorités portent sur le **calibrage du seuil** (courbe précision–rappel, coûts métier), l'**enrichissement** de la donnée (historique multi-années, signaux d'interactions, variables temporelles), la **planification** (SQL Agent) et un **monitoring** simple (volumétrie, fraîcheur). À moyen terme, l'exploration de **modèles d'ensemble** (XGBoost/LightGBM), la **calibration des probabilités** (Platt/Isotonic) et l'**explicabilité locale** (SHAP) renforceront la confiance et l'adoption. À plus long terme, un **MLOps** léger (registre de modèles, surveillance de *drift*, ré-entraînement périodique) et l'**intégration CRM** des scores compléteront la chaîne.

Organisation du mémoire

Le mémoire suit la logique du projet :

1. **Chapitre 1 — Cadre général** : contexte, problématique, objectifs, choix d'approche, périmètre.
2. **Chapitre 2 — Sprint 1 (DWH)** : collecte, nettoyage, dictionnaire, conception en étoile, contraintes.
3. **Chapitre 3 — Sprint 2 (ETL SSIS)** : design du package, mappings, procédure, exécutions & contrôles.
4. **Chapitre 4 — Sprint 3 (Power BI)** : modèle, mesures DAX, pages & visuels.

5. **Chapitre 5 — Sprint 4 (IA/ML)** : préparation des données, modélisation, métriques, réintégration au DWH/BI.
6. **Chapitre 6 — Conclusion générale & perspectives** : bilan, limites assumées, feuille de route.

En résumé, ce mémoire ne cherche pas à multiplier les artefacts, mais à **raconter** la construction d'une chaîne de valeur **opérationnelle** : partir d'un CSV imparfait, mettre la donnée en ordre, l'exposer utilement, **l'augmenter** par un modèle simple mais intégré, et rendre l'ensemble **utilisable** par les équipes. C'est cette continuité — $DWH \rightarrow ETL \rightarrow BI \rightarrow ML \rightarrow DWH/BI$ — qui fait la valeur du projet et ouvre la voie à une industrialisation progressive, au service d'un pilotage plus **fiable**, plus **lisible** et plus **actionnable**.

Chapitre 1

Cadre général

1.1 Introduction

Dans un environnement commercial fortement concurrentiel, la **qualité de la donnée** et la capacité à transformer l'information en *décision* constituent des leviers de performance. Le présent projet vise à doter l'organisation d'une chaîne **donnée** $\rightarrow$ **décision** robuste et exploitable par les équipes opérationnelles. Ce chapitre présente le **cadre général** : l'organisme d'accueil et son secteur, le **contexte métier**, la **problématique**, l'**état de l'existant**, la **solution proposée**, ainsi que la **méthodologie** retenue et son application au projet.

1.2 Présentation de l'organisme d'accueil

1.2.1 Huawei Tunisie

Huawei est un acteur global des technologies de l'information et de la communication (TIC) qui opère dans plus de 170 pays. La filiale tunisienne s'inscrit dans cet écosystème et intervient en proximité des acteurs publics et privés, notamment via le *Tunis Finance and Healthcare Sales Department*. La mission principale de l'entité locale est d'accompagner la transformation numérique de ses clients en conjuguant **infrastructures**, **données** et **solutions analytiques**.

1.2.2 Secteur d'activité

Le périmètre couvert par l'entité d'accueil se situe à l'interface entre :

— **Infrastructure & cloud** : réseaux, data centers, virtualisation, cloud hybride ;
— **Données & analytique** : plateformes data, gouvernance, BI, IA/ML appliqués à la performance ;
— **Solutions métiers** : intégration de briques technologiques adaptées (sécurité, observabilité, continuité).

Ce positionnement requiert un pilotage fin des **opportunités commerciales** et des **portefeuilles** clients, avec une visibilité claire sur l'avancement, la probabilité de conversion et l'effort prioritaire.

1.2.3 Services fournis (exemples)

— Conseil en architecture data/BI/ML et intégration de solutions ;
— Mise en place d'infrastructures & plateformes cloud, sécurité et supervision ;
— Industrialisation : *data pipelines*, gouvernance, automatisation, *runbook*, accompagnement au changement.

1.3 Présentation du projet

1.3.1 Contexte et périmètre

Le service commercial dispose d'un **export CRM** (CSV ∼500 lignes) regroupant des opportunités. La donnée est **peu normalisée**, **non historisée**, et analysée de manière *ad hoc*. L'objectif est de construire une chaîne complète :

$$CSV \rightarrow Staging \rightarrow DWH(toile) \rightarrow BI(PowerBI) \rightarrow ML(scores) \rightarrow DWH/BI$$

pour fiabiliser les KPI, offrir des analyses multi-axes et **prioriser** via la probabilité de gain.

1.3.2 Problématique

— Comment **fiabiliser** la donnée pour produire des indicateurs stables et comparables dans le temps ?
— Comment **outiller** l'équipe (filtres, *drill-down*, vues par statut/secteur/région/-temps) ?
— Comment **prioriser** l'action via une **probabilité de gain** interprétable et réintégrée au SI décisionnel ?

1.3.3 Étude de l'existant

— **Sources** : un fichier `opportunites.csv` (UTF-8), libellés hétérogènes, types mixtes ;
— **Traitements** : analyses ponctuelles, comparabilité limitée, métriques instables ;
— **SI décisionnel** : pas de DWH dédié, pas d'historisation, pas de vue unifiée ;
— **Pilotage** : pas de **probabilité de gain** pour prioriser les relances et arbitrages.

1.3.4 Solution proposée (vision d'ensemble)

1. **Modèle de données** en **étoile** (fait *opportunité* + dimensions client/produit/secteur/région/statut/temps) ;
2. **ETL (SSIS)** rejouable : CSV → `stg.opportunites_raw` → DWH via `dwh.sp_load_dwh` ;
3. **BI (Power BI)** : modèle conforme, KPI (CA, Nb opp., Taux gagné, Proba ML), pages de synthèse/détails ;
4. **ML** (scikit-learn) : régression logistique, validation croisée, **réintégration** des scores dans `ml.predictions` et `ml.vw_last_predictions`.

1.3.5 Objectifs SMART

Table 1.1 – Objectifs SMART du projet

Axe	Objectif	Mesure/Échéance
DWH	Schéma en étoile opérationnel, contraintes OK	DDL validé & chargement test $\geq$ 99% d'ici fin Sprint 1
ETL	Pipeline rejouable CSV→stg→DWH	Exécutions conformes & logs, fin Sprint 2
BI	KPI fiables & pages de synthèse	Conformité définitions, fin Sprint 3
ML	Proba de gain intégrée au DWH/BI	Vue `ml.vw_last_predictions`, fin Sprint 4

1.3.6 Périmètre (in/out)

Table 1.2 – Périmètre fonctionnel

Inclus (In)	Exclus (Out)
Collecte CSV, staging, DWH étoile	Refonte complète du CRM opérationnel
ETL SSIS rejouable	Intégration temps réel streaming
Power BI (KPI, analyses)	Gouvernance avancée MDM complète
ML (régression logistique)	MLOps industriel complet (registre, CI/CD)

1.3.7 Parties prenantes

Table 1.3 – Parties prenantes (extrait)

Acteur	Intérêt	Attentes
Direction commerciale	Pilotage pipeline	KPI fiables & visibilité
Sales/Account managers	Priorisation actions	Scores interprétables
Équipe data/IT	Pérennité & qualité	DWH stable, ETL rejouable

1.3.8 Contraintes, hypothèses, critères de succès

— **Contraintes** : volume limité (∼500 lignes), délais, silos métier/IT, gestion des libellés.
— **Hypothèses** : accès aux exports CRM, environnement SQL/SSIS/Power BI disponible.
— **Critères de succès** : DWH valide, ETL rejouable, KPI conformes, scores ML visibles dans BI, documentation livrée.

1.4 Architecture cible (haut niveau)

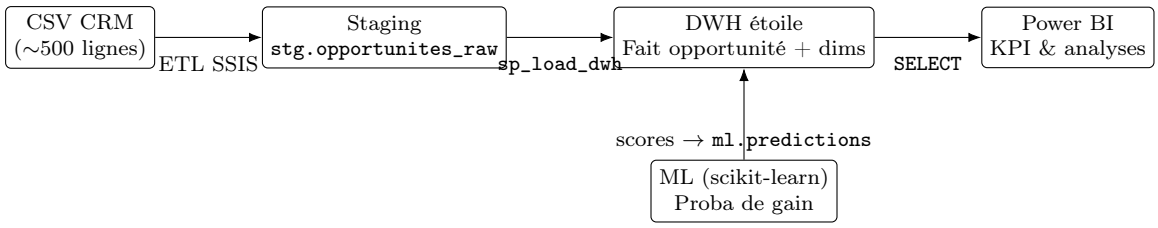

FIGURE 1.1 – Architecture cible — flux haut niveau.

1.5 Méthodologie de travail

1.5.1 Pourquoi une démarche Agile ?

Le projet combine **incertitudes** (qualité des données, itérations BI/ML) et **contraintes** (délai, valeur incrémentale). Une approche Agile permet de livrer **vite** des incréments utiles (DWH minimal, première page BI, prototype ML), d'intégrer des **retours** fréquents et de gérer le **risque** par **itérations courtes**.

1.5.2 Méthodes agiles considérées

Deux cadres ont été évalués : **Scrum** et **Kanban**. Les paragraphes suivants présentent chaque méthode, suivis d'un tableau comparatif.

Scrum. Scrum organise le travail en **sprints** time-boxés (1–4 semaines) avec des rôles (PO, SM, Dev Team), des **événements** (Planning, Daily, Review, Rétro) et des **artefacts** (Product Backlog, Sprint Backlog, Incrément). Chaque sprint doit produire un **incrément potentiellement livrable**.

Kanban. Kanban se concentre sur le **flux continu** : visualisation du travail, **WIP limits**, **lead time**, **débit**. Il n'impose pas de time-box ; il vise l'optimisation du flux et la réduction des temps d'attente. Pertinent pour du **support/TMA** et les arrivées continues de demandes.

1.5.3 Comparatif Scrum vs Kanban

TABLE 1.4 – Comparatif des méthodes Scrum et Kanban

Critère	**Scrum**	**Kanban**
Cadence	Sprints time-boxés (1–4 semaines)	Flux continu, sans time-box
Pilotage	Planning/Review/Rétro + Daily	Revue du flux, *stand-ups* légers
Charge	Sprint Backlog figé pendant le sprint	WIP limits, tirage à la capacité
Livraison	Incrément à chaque sprint	Livraison au fil de l'eau
Adaptation	Inspection & adaptation par événements	Ajustements continus sur le flux
Usage type	Projets par **jalons**	**Flux** de demandes (support/ops)

1.5.4 Choix méthodologique : Scrum

Le projet requiert des **jalons clairs** (DWH minimal, ETL stable, page BI, pipeline ML réintégré) et des **séquences de validation** (revues) : **Scrum** a été retenu. Bénéfices observés :

— focalisation sur la **valeur incrémentale** ;
— **cadence** et rituels qui favorisent l'alignement ;
— intégration naturelle des retours métier/technique entre sprints.

1.5.5 Les piliers de Scrum

1. **Transparence** : information visible (backlogs, *Definition of Done*, règles de qualité) ;
2. **Inspection** : revues régulières (Daily, Review, Rétro) ;
3. **Adaptation** : ajustements du backlog, du process, des priorités.

1.5.6 Scrum Team (rôles et responsabilités)

TABLE 1.5 – Rôles Scrum et responsabilités dans le projet

Rôle	**Responsabilités (adaptées au projet)**
Product Owner (PO)	Vision produit, ordre du Product Backlog, arbitrages de valeur (KPI priorisés, besoins BI/ML).
Scrum Master (SM)	Facilitation, levée d'obstacles, amélioration continue, respect du cadre Scrum.
Dev Team	Conception DWH, ETL SSIS, mesures DAX, pipeline ML, tests & documentation.

1.5.7 Artefacts Scrum

Product Backlog. Liste ordonnée des besoins (US), mise à jour en continu.
Sprint Backlog. Sélection d'US + plan pour le sprint en cours.
Incrément. Produit potentiellement livrable, conforme à la *Definition of Done (DoD)*.
Definition of Done. DDL validé, contraintes OK, tests de chargement $\geq 99\%$, dictionnaire à jour, KPI conformes, scripts ML ré-exécutables.

1.5.8 Événements Scrum

— **Sprint Planning** : objectif de sprint, sélection d'US, plan de réalisation ;
— **Daily Scrum** (15 min) : synchronisation et obstacles ;
— **Sprint Review** : démonstration de l'incrément aux parties prenantes ;
— **Sprint Retrospective** : amélioration continue du cadre de travail.

1.5.9 RACI (gouvernance)

TABLE 1.6 – Tableau RACI (extrait)

Livrable/Action	**PO**	**SM**	**Dev**	**Commentaires**
Modèle DWH (étoile)	A	I	R	Validation du dictionnaire
ETL SSIS (stg→DWH)	I	C	R	Rejouable, contrôles post-lot
Tableau de bord (KPI)	A	I	R	Définitions alignées
Pipeline ML (scores)	A	I	R	Intégration vue `ml.vw_last_predictions`

1.5.10 Analyse des risques & parades

TABLE 1.7 – Registre des risques (extrait)

Risque	**Prob.**	**Impact**	**Parade**
Échantillon limité (∼500)	Moyenne	Moyenne	Étendre historique, features temporelles
Hétérogénéité des libellés	Élevée	Moyenne	Référentiels, mappings, `CHECK`
Chaînage ML→DWH manuel	Moyenne	Moyenne	Planification SQL Agent, audits

1.5.11 Backlog produit (extrait)

TABLE 1.8 – Backlog produit (extrait)

US	Description	Critères d'acceptation	DoD
US1	Charger CSV en staging	100% colonnes typées	500 lignes en `stg.opportunites_raw`
US2	Modèle en étoile (DWH)	PK/FK, intégrité	DDL validé, contraintes OK
US3	KPI Power BI	Conformité définitions	Mesures DAX vérifiées
US4	Scores ML → DWH	Vue exposée	`ml.vw_last_predictions` alimentée

1.6 Conclusion du chapitre

Ce chapitre a présenté l'organisme d'accueil, son secteur et ses services, la **problématique** et l'**état de l'existant**, la **solution proposée** et les **objectifs** (SMART), le **périmètre**, les **parties prenantes**, ainsi que la **méthodologie**. Le choix de **Scrum** s'est imposé pour orchestrer une progression par **incréments** : DWH minimal, ETL stable, BI exploitable, ML réintégré. Les chapitres suivants détaillent la réalisation de chaque sprint, de la conception du **DWH** jusqu'à la mise à disposition d'indicateurs **fiables** et d'une **probabilité de gain** intégrée au pilotage.

Chapitre 2

Sprint 0 : Phase de préparation

2.1 Introduction

Ce chapitre constitue le point de départ du projet et correspond au **Sprint 0**, une phase préparatoire essentielle en démarche Agile (cadre *Scrum*). Il vise à poser les **fondations** avant les itérations de réalisation : clarification des **exigences**, identification des **acteurs**, cadrage du **périmètre**, élaboration d'un **Product Backlog initial**, définition de l'**organisation Scrum**, de l'**écosystème technique** et de l'**architecture** cible de haut niveau (DWH → ETL SSIS → Power BI → ML).

2.2 Contexte métier et rappel des objectifs

Le projet s'inscrit au sein de **Huawei Tunisie**, dans le *Tunis Finance and Healthcare Sales Department*. L'enjeu est d'outiller le **pilotage du pipeline d'opportunités** (fiabiliser les KPI, historiser, comparer dans le temps) et d'**augmenter** la décision par une **probabilité de gain** (ML) réinjectée et visible dans la BI. Le **périmètre** cible :

$$CSVCRM \rightarrow Staging \rightarrow DWH(toile) \rightarrow PowerBI \rightarrow MLscores \rightarrow DWH/BI.$$

2.3 Acteurs, parties prenantes et personas

2.3.1 Acteurs clés (vue synthétique)

TABLE 2.1 – Acteurs et attentes principales

Acteur	Attentes	Interactions
Direction commerciale	Visibilité consolidée, indicateurs fiables	Tableaux de bord Power BI (synthèse)
Sales / Account managers	Prioriser les relances par proba de gain	Pages analytiques & filtres (client, région, statut)
Équipe Data/IT	Pérennité, qualité, rejouabilité	SQL Server/SSMS, SSIS, Power BI, scripts ML
Encadrement académique	Avancement régulier, documentation	Notes de sprint, livrables

2.3.2 Personas (extraits)

Persona A — Account Manager. Besoin d'un **cockpit** pour décider où investir l'effort : quels clients/opportunités prioriser aujourd'hui ?

Persona B — Sales Director. Vision **macro** : CA, taux gagné, tendance, **risque/opportunité** par portefeuille.

Persona C — Data Engineer. Stabilité du **pipeline**, **qualité** des données, rejouabilité, logs, audits.

2.4 Exigences et définitions

2.4.1 Besoins fonctionnels (BF)

— **BF1** — Charger le CSV source (∼500 lignes) dans un **staging** typé, traçable.
— **BF2** — Intégrer au **DWH** modélisé en **étoile** (fait opportunité + dimensions).
— **BF3** — Exposer des **KPI** (CA, Nb opportunités, Taux gagné) et analyses multi-axes (région, secteur, statut, temps).
— **BF4** — Calculer et **réintégrer** une **probabilité de gain** (ML) par opportunité dans le DWH/BI.
— **BF5** — Permettre la **navigation** (filtres, drill), l'export & le contrôle (échantillons, COUNT(*)).

2.4.2 Besoins non fonctionnels (BNF)

— **Qualité** : typage strict, référentiels, PK/FK, `CHECK`, `NULLIF`, anti-doublons.
— **Rejouabilité** : package SSIS idempotent, logs d'exécution, traçabilité.
— **Lisibilité** : définitions KPI documentées, mesures DAX & schémas clairs.
— **Évolutivité** : possibilité d'ajouter des dimensions/mesures et d'étendre la fenêtre historique.

2.4.3 Règles de gestion (extrait)

— `montant_offre_tnd` ≥ 0 ; dates au format ISO ; libellés « Gagné/Perdu/En cours » normalisés.
— **Grain du fait** : 1 ligne = 1 opportunité ; **dimension temps** requise sur les dates clés.

2.5 Cas d'utilisation (texte + tableau)

2.5.1 Description textuelle

UC1 — Ingestion. L'utilisateur technique déclenche l'ETL : CSV → `stg.opportunites_raw` → DWH via `dwh.sp_load_dwh`. **UC2 — Consultation KPI**. Les utilisateurs métiers consultent les KPI & analyses Power BI. **UC3 — ML**. Le data scientist calcule des **probabilités de gain** ; les scores sont chargés dans `ml.predictions` et exposés via `ml.vw_last_predictions`.

TABLE 2.2 – Cas d'utilisation (synthèse)

UC	But	Acteurs	Issue
UC1 — Ingestion	Charger données fiables dans DWH	Data/IT	Staging & DWH peuplés
UC2 — KPI & analyses	Décider/prioriser	Sales/Direction	KPI & filtres utilisables
UC3 — ML → DWH	Priorisation par proba	Data/IT & Sales	Vue `ml.vw_last_predictions`

2.6 Backlog produit initial

TABLE 2.3 – Backlog produit — version Sprint 0

ID	User Story	Critères d'acceptation	DoD	Priorité
US1	En tant qu'ingestion, je charge le CSV en staging typé	100% des colonnes présentes	`stg.opportunites_raw` OK	Haute
US2	En tant qu'architecte DWH, je conçois le modèle en étoile	PK/FK valides, intégrité	DDL validé	Haute
US3	En tant qu'intégrateur, j'industrialise `sp_load_dwh`	Exécution OK, logs & contrôles	Test ≥99%	Haute
US4	En tant qu'analyste, je consulte des KPI fiables	Conformité définitions	Mesures DAX vérifiées	Haute
US5	En tant que DS, je publie les proba dans DWH/BI	1 ligne/opportunité (dernière proba)	`ml.vw_last_predictions`	Haute
US6	En tant que qualité, je trace les anomalies	Règles CHECK & rapport	Rapport d'anomalies	Moyenne

2.7 Architecture & conception initiale

2.7.1 Architecture fonctionnelle (haut niveau)

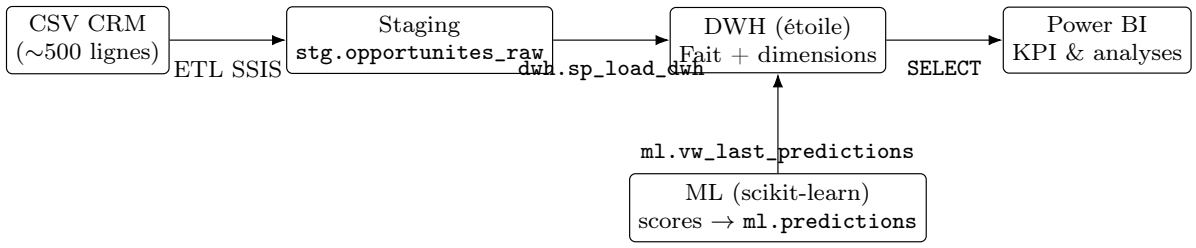

FIGURE 2.1 – Architecture cible — flux de bout en bout (vue fonctionnelle).

2.7.2 DFD niveau 0 (texte)

Les données partent de la source CSV vers le **staging** (contrôles, typage), sont intégrées dans le **DWH** (étoile) puis consommées par **Power BI**. Les scores **ML** sont chargés dans `ml.predictions` et exposés via `ml.vw_last_predictions`.

2.7.3 DFD niveau 1 (intégration)

— **Processus** : *Extract* (CSV) → *Transform* (nettoyage, normalisation) → *Load* (staging puis DWH).

— **Contrôles** : *COUNT(*)*, échantillons, `CHECK`, rejets documentés.
— **Sortie** : tables DWH, vues exposées, dataset Power BI.

2.8 Planification macro des sprints

Table 2.4 – Planification macro (projet)

Sprint	Description	Contenu principal	Durée
S0	Préparation	Exigences, backlog, archi, env.	1–2 sem.
S1	DWH	DDL étoile, contraintes, test charge	2–3 sem.
S2	ETL SSIS	Pipeline rejouable, `sp_load_dwh`, contrôles	2 sem.
S3	Power BI	Modèle, DAX, pages KPI/analyses	2 sem.
S4	ML	Features, CV, AUC, réintégration DWH	2 sem.

2.9 Organisation Scrum & gouvernance

2.9.1 Rôles & responsabilités (RACI)

Table 2.5 – RACI — extrait

Livrable	PO	SM	Dev	Commentaires
Modèle DWH (étoile)	A	I	R	Dictionnaire validé
ETL SSIS (stg→DWH)	I	C	R	Rejouable, contrôles
Tableau de bord (KPI)	A	I	R	Définitions alignées
Scores ML → DWH/BI	A	I	R	`ml.vw_last_predictions`

2.9.2 Definition of Ready / Definition of Done

Table 2.6 – DoR / DoD (extrait)

DoR (prêt)	**DoD (terminé)**
US claire, critères d'acceptation	Tests de charge OK, logs conservés
Données accessibles & typées	DDL validé, contraintes OK
Dépendances identifiées	KPI conformes & documentés

2.9.3 Plan de communication

— **Hebdomadaire** : point d'avancement (30 min) — KPI, risques, décisions.
— **Fin de sprint** : *Sprint Review* (démonstration) + note de synthèse.

2.10 Critères de succès & indicateurs

— DWH **opérationnel** (étoile + contraintes) — test de charge $\geq 99\%$.
— ETL SSIS **rejouable** — logs, contrôles post-lot, erreurs tracées.
— BI **fiable** — KPI conformes aux définitions métier.
— ML **intégré** — `ml.vw_last_predictions` consommée par Power BI.

2.11 Risques & parades

Table 2.7 – Registre des risques (extrait)

Risque	**Prob.**	**Impact**	**Parade**
Échantillon limité ($\sim$500)	Moyenne	Moyenne	Étendre historique, features temporelles
Hétérogénéité libellés	Élevée	Moyenne	Référentiels, mappings, `CHECK`
Chaînage ML→DWH manuel	Moyenne	Moyenne	SQL Agent, audits

2.12 Environnement technique & outils

— **SQL Server/SSMS** : DWH, vues, procédures.
— **SSIS** : ingestion CSV → staging, appel `sp_load_dwh`.
— **Power BI** : modèle, mesures DAX, pages KPI & analyses.
— **Python / scikit-learn** : pipeline ML, AUC/ROC, export de scores.
— **Contrôle qualité** : scripts de *count*, échantillons, journaux d'exécution.

2.13 Conclusion du chapitre

Le Sprint 0 a permis de **cadrer** le projet : exigences, acteurs, règles de gestion, **backlog** initial, architecture cible, gouvernance et planning. Les jalons sont clairs : **S1** (DWH), **S2** (ETL SSIS), **S3** (Power BI), **S4** (ML). Cette préparation garantit une exécution **itérative** et **incrémentale** alignée sur la valeur métier.

Chapitre 3

Sprint 1 : Collecte, nettoyage & conception DWH

3.1 Objectifs détaillés & périmètre

L'objectif de ce sprint est de livrer une première **base décisionnelle** fiable et requêtée par Power BI : (i) ingestion du CSV en **staging** typé & traçable, (ii) **conception étoile** (fait + dimensions), (iii) **contraintes** d'intégrité & indexations de base, (iv) **procédure** de chargement vers le DWH. Le périmètre inclut le **fait opportunité** et les dimensions *client, produit/service, secteur, région, statut, temps.*

3.2 Données sources & dictionnaire

Aperçu du fichier source

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	ID Opport	Nom du C	Secteur	Taille de l	Région	Produit / S	Montant c	Statut de l	Date pren	Date de cl	Durée du	Nombre d	Score de s	Nombre d	Revenus r	Revenus a	Responsal	Équipe	Probabilité de gain (%)	
2	OPP-2000	Tessier SA	Santé	Moyenne	Nabeul	Service de	350116	Perdu	########	########	95	6	3	2	99278	2250908	Anaïs Ledi	Équipe Sa	29	
3	OPP-2001	Carlier SA	Microfinai	Moyenne	Sfax	Firewall	174061	Gagné	########	########	126	7	1	4	63005	712044	Thomas G	Équipe Fir	83	
4	OPP-2002	Pelletier	Microfinai	Moyenne	Sfax	Solution C	182944	Perdu	########	########	37	4	4	4	69416	2194222	Alphonse	Équipe Sa	16	
5	OPP-2003	Lesage S.A	Microfinai	Grande	Nabeul	Routeur	98202	Perdu	########	########	44	6	2	1	193825	2392381	Adélaïde I	Équipe Sa	11	
6	OPP-2004	Sauvage S	Banque	Grande	Tunis	Switch rés	438865	En cours	########	9/1/2025	74	1	1	3	41042	1070702	David Tes	Équipe Fir	49	
7	OPP-2005	Dupuy	Microfinai	Moyenne	Tunis	Service de	344353	Gagné	9/3/2025	9/8/2025	5	10	4	5	132888	165983	Michelle (	Équipe Fir	92	
8	OPP-2006	Lenoir	Banque	Petite	Gabès	Solution d	154043	Perdu	########	7/8/2025	11	9	3	0	136652	1905954	Jules Gau	Équipe Fir	6	
9	OPP-2007	Vallée Bri	Assurance	Petite	Sfax	Service de	286014	En cours	########	########	87	5	3	2	109044	431430	Patrick Ar	Équipe Sa	65	
10	OPP-2008	Dias S.A.S	Assurance	Grande	Sfax	Firewall	368634	Perdu	########	########	95	2	1	0	111241	1690748	Théodore	Équipe Sa	35	
11	OPP-2009	Gomez	Banque	Grande	Sfax	Switch rés	272069	Perdu	5/6/2025	########	110	7	3	4	29998	2315260	Henri Dub	Équipe Fir	14	
12	OPP-2010	Blot Bouv	Microfinai	Moyenne	Tunis	Switch rés	368837	Gagné	9/5/2025	9/9/2025	4	5	5	4	177519	238000	Daniel-Th	Équipe Fir	85	
13	OPP-2011	Bousquet	Assurance	Grande	Sfax	Solution d	440389	Gagné	2/5/2025	########	114	7	1	2	47585	1108613	Maurice B	Équipe Sa	96	
14	OPP-2012	Charles M	Santé	Moyenne	Sfax	Solution d	389670	En cours	########	########	98	2	5	5	95664	325219	Gabriel-Tr	Équipe Fir	67	
15	OPP-2013	Rocher S.A	Banque	Moyenne	Gabès	Service de	322270	Gagné	########	########	106	7	3	0	171231	1246416	Éric Charp	Équipe Sa	90	
16	OPP-2014	Toussaint	Microfinai	Moyenne	Sfax	Solution d	317167	Gagné	########	########	1	7	1	4	82431	1133685	Xavier Col	Équipe Fir	77	
17	OPP-2015	Leroux S.A	Microfinai	Grande	Sfax	Solution d	214274	Gagné	7/3/2025	########	45	6	3	0	28991	121790	Alix Delor	Équipe Sa	89	
18	OPP-2016	Evrard	Santé	Petite	Sfax	Switch rés	108247	Gagné	########	########	64	7	4	4	150421	2249332	Gérard de	Équipe Sa	78	
19	OPP-2017	Barre SA	Santé	Grande	Tunis	Service de	323547	Gagné	########	########	176	6	4	0	81506	1835645	Antoinett	Équipe Sa	99	
20	OPP-2018	Bazin	Assurance	Grande	Gabès	Solution C	369302	Gagné	########	########	306	7	2	1	75495	1363036	Benjamin	Équipe Sa	90	
21	OPP-2019	Chevallier	Assurance	Moyenne	Sousse	Routeur	167156	Gagné	8/3/2025	9/5/2025	33	8	1	2	70981	1134273	Suzanne F	Équipe Sa	94	
22	OPP-2020	Hardy SAR	Microfinai	Petite	Sfax	Firewall	13548	Perdu	########	########	128	7	3	4	104585	1482198	René Lecc	Équipe Sa	21	
23	OPP-2021	Fleury	Assurance	Moyenne	Sousse	Service de	374533	Perdu	########	########	0	4	4	2	51566	1770337	Jeanne Kl	Équipe Sa	18	
24	OPP-2022	Petit et Fi	Microfinai	Grande	Sfax	Firewall	207161	Perdu	########	########	54	5	4	2	34855	938022	Timothée	Équipe Fir	35	
25	OPP-2023	Delorme F	Microfinai	Grande	Sfax	Routeur	428216	Gagné	########	9/4/2025	80	5	1	3	181417	2194229	Charles-A	Équipe Fir	95	
26	OPP-2024	Gilbert et	Santé	Grande	Sfax	Solution d	231840	Gagné	########	7/9/2025	229	7	1	3	161744	2110056	Antoine B	Équipe Sa	92	
27	OPP-2025	Jacques	Microfinai	Grande	Tunis	Firewall	377858	Perdu	########	########	25	9	5	5	27277	558728	Alfred Me	Équipe Fir	22	

FIGURE 3.1 – Aperçu du fichier `opportunites.csv` (UTF-8, `;`), ∼500 lignes.

Dictionnaire des champs (extrait)

TABLE 3.1 – Dictionnaire des champs source (extrait)

Colonne	Type source	Cible (stg)	Description/Remarques
id_opportunite	texte	`NVARCHAR(64)`	Identifiant unique source (clé métier).
nom_client	texte	`NVARCHAR(255)`	Libellé client (normalisé).
secteur	texte	`NVARCHAR(100)`	Secteur d'activité (mappé référentiel).
region	texte	`NVARCHAR(100)`	Région (mappée).
produit_service	texte	`NVARCHAR(150)`	Famille d'offre (mappée).
montant_offre_tnd	nombre	`DECIMAL(18,2)`	Montant de l'offre (TND).
statut_offre	texte	`NVARCHAR(50)`	{Gagné, Perdu, En cours} après normalisation.
date_premier_contact	date/texte	`DATE`	`TRY_CONVERT` depuis ISO/JJ-MM-AAAA.
date_cloture	date/texte	`DATE`	Peut être `NULL` si en cours.
nb_reunions	entier	`INT`	≥ 0.
nb_reclamations	entier	`INT`	≥ 0.

3.3 Règles de nettoyage & normalisation

Règles générales

— Trim & dé-accentuation des libellés ; suppression doubles espaces.
— Normalisation des statuts : « Gagné », « Perdu », « En cours ».
— Dates en ISO ; conversions robustes via `TRY_CONVERT`.
— Montants ≥ 0, `NULLIF` pour vides.

Mappages (extrait)

TABLE 3.2 – Mappage de libellés hétérogènes (extrait)

Source	Cible	Règle
Won, *win*, *gagnée*	Gagné	`LOWER()` + table de correspondance.
Closed lost, *perdue*	Perdu	Idem.
In progress, *open*	En cours	Idem.

3.4 De la conception logique à la conception dimensionnelle

Entités & attributs clés

— **Opportunité** (grain = 1 opportunité) : montants, dates, statut, compteurs.
— **Client**, **Produit/Service**, **Secteur**, **Région**, **Temps**.

3.5 Modèle en étoile (schéma dimensionnel)

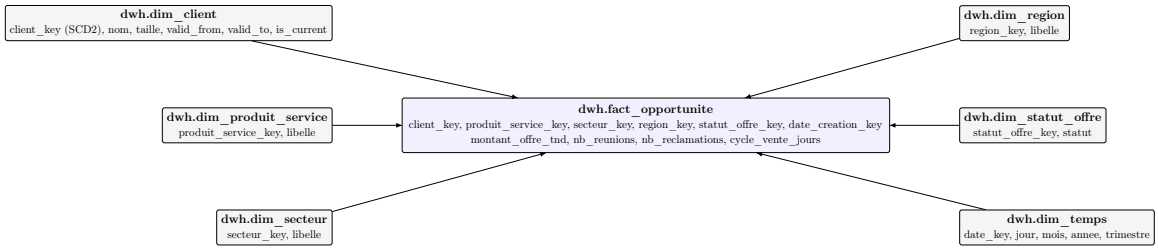

FIGURE 3.2 – Modèle en étoile centré sur le fait `opportunite`.

3.6 Historisation client (SCD2)

Structure SCD2

TABLE 3.3 – Structure type SCD2 (`dwh.dim_client`)

Colonne	Type	Rôle
client_key	`INT IDENTITY`	PK technique (surrogate).
client_id_source	`NVARCHAR(64)`	Identifiant métier source.
nom_client, taille_entreprise, ...	`NVARCHAR`	Attributs de la dimension.
valid_from, valid_to	`DATE`	Fenêtre de validité SCD2.
is_current	`BIT`	Flag de ligne courante.

Logique de mise à jour (extrait pseudo-SQL)

— Détecter changements d'attributs ⇒ fermer version (mettre `valid_to` = *hier*, `is_current`=0) puis insérer nouvelle version.
— Sinon, conserver la version courante.

3.7 DDL principaux (extraits commentés)

Staging (proche source)

```
CREATE TABLE stg.opportunites_raw (
  id_opportunite       NVARCHAR(64)  NOT NULL,
  nom_client           NVARCHAR(255) NULL,
  secteur              NVARCHAR(100) NULL,
  region               NVARCHAR(100) NULL,
  produit_service      NVARCHAR(150) NULL,
  montant_offre_tnd    DECIMAL(18,2) NULL CHECK (montant_offre_tnd >= 0),
  statut_offre         NVARCHAR(50)  NULL,
  date_premier_contact DATE          NULL,
  date_cloture         DATE          NULL,
  nb_reunions          INT           NULL CHECK (nb_reunions >= 0),
  nb_reclamations      INT           NULL CHECK (nb_reclamations >= 0)
);
```

Fait (grain = 1 opportunité)

```
CREATE TABLE dwh.fact_opportunite (
  opportunite_key        INT IDENTITY PRIMARY KEY,
  client_key             INT NOT NULL,
  produit_service_key    INT NOT NULL,
  secteur_key            INT NOT NULL,
  region_key             INT NOT NULL,
  statut_offre_key       INT NOT NULL,
  date_creation_key      INT NOT NULL,
  montant_offre_tnd      DECIMAL(18,2) NOT NULL,
  nb_reunions            INT NULL,
  nb_reclamations        INT NULL,
  cycle_vente_jours      INT NULL,

);
```

3.8 Contraintes, indexation & performances

Indexation de base (exemples)

```
CREATE INDEX ix_fact_date ON dwh.fact_opportunite(date_creation_key);
CREATE INDEX ix_fact_client ON dwh.fact_opportunite(client_key);
```

Intégrité & gouvernance

PK/FK systématiques, `CHECK` sur montants/compteurs, `NULLIF` pour vides, référentiels (*statut, secteur, région*) maîtrisés.

3.9 Contrôles qualité & tests

Échantillons & volumétrie

```
SELECT COUNT(*) FROM stg.opportunites_raw;
SELECT COUNT(*) FROM dwh.fact_opportunite;
SELECT TOP 10 * FROM dwh.fact_opportunite ORDER BY opportunite_key DESC;
```

Concordance des totaux

```
SELECT SUM(montant_offre_tnd) FROM stg.opportunites_raw;
SELECT SUM(montant_offre_tnd) FROM dwh.fact_opportunite;
```

3.10 Risques & parades (rappel)

Table 3.4 – Risques principaux et parades

Risque	**Parade**
Échantillon limité (∼500 lignes)	Étendre historique, enrichir features temporelles.
Libellés hétérogènes	Référentiels, table de correspondance, contrôles CHECK.
Chaînage ML→DWH manuel	Planification (SQL Agent), contrôles post-lot, alerting.
Seuil ML non calibré	Courbe précision–rappel, coût métier, calibration.

3.11 Livrables du sprint

— **Schéma** DWH (étoile) & DDL versionné ;
— **Procédure** `dwh.sp_load_dwh` (ES du sprint 2) ;
— **Contrôles** SQL (volumétrie & concordances) ;
— **Note d'architecture** & dictionnaire de données.

Conclusion du chapitre

Ce premier sprint a permis de **stabiliser la fondation décisionnelle** du projet. À partir d'un export CSV hétérogène, nous avons mis en place une **zone de staging** typée et traçable, défini les **règles de nettoyage/normalisation** (libellés, dates, montants) et conçu un **modèle en étoile** centré sur le fait *opportunité* et ses dimensions métier (client, produit/service, région, secteur, statut, temps). Les **contraintes d'intégrité** (PK/FK, `CHECK`) et une **indexation minimale** assurent la cohérence et des performances de base pour les futurs usages BI.

Sur le plan méthodologique, nous avons **documenté le dictionnaire de données**, matérialisé le **grain** et les **clés de jointure**, et posé les **règles de gestion** indispensables à la reproductibilité. Les **contrôles qualité** (volumétrie, concordance des totaux, échantillons) confirment la justesse des transformations et la conformité du modèle par rapport aux attentes métier.

Les **livrables** de ce sprint (DDL des tables du DWH, schéma étoile, scripts de contrôles, note d'architecture) constituent la base sur laquelle s'appuiera l'industrialisation du flux. Ils préparent directement le **Sprint 2 (ETL SSIS)** : ingestion rejouable du CSV vers le staging et **chargement automatisé** dans le DWH via la procédure `dwh.sp_load_dwh`. À l'issue de ce prochain jalon, la chaîne technique sera suffisamment robuste pour alimenter le **modèle Power BI** (Sprint 3) et accueillir les **scores ML** (Sprint 4).

Chapitre 4

Sprint 2 : ETL SSIS (staging, procédures)

4.1 Introduction du chapitre

Ce chapitre décrit la mise en place de la **chaîne ETL** sous **SQL Server Integration Services (SSIS)** permettant de transformer un export CRM *flat file* en données **fiables** et **rejouables** au sein du **Data Warehouse** (schéma `dwh`). L'objectif est double : (i) garantir un **chargement maîtrisé** et traçable depuis la source vers la *zone de staging*, puis (ii) **orchestrer** les *transformations métier* et les *contrôles d'intégrité* centralisés côté SQL via des **procédures stockées**.

Contexte & enjeux. Les données sources issues du CRM se présentent sous forme de **CSV** (∼500 lignes), avec des *variations de libellés* et des *formats hétérogènes*. Un ETL robuste doit donc : (a) **absorber** ces fichiers sans erreur, (b) **normaliser** les colonnes sensibles (dates, montants, statut), (c) **protéger** la cohérence du DWH (PK/FK, `CHECK`), et (d) **journaliser** l'exécution pour l'audit.

Périmètre. Le périmètre couvert dans ce sprint porte sur :
— l'**ingestion** du fichier `opportunites.csv` dans `stg.opportunites_raw` (Unicode, typage strict) ;
— l'**orchestration** d'un `Control Flow` minimaliste mais efficace (purge, *Data Flow*, appel de procédure) ;
— le **chargement** du modèle en étoile via `sp_load_dwh` (upsert dimensions, résolution des clés, insertion dans la table de faits) ;
— des **contrôles post-lot** (volumétrie, intégrité référentielle, valeurs négatives).

Principes ETL retenus.
— **Simplicité/fiabilité** : transformations lourdes côté SQL (procédures), SSIS focalisé sur l'ingestion et l'orchestration.
— **Rejouabilité** : `TRUNCATE` du *staging* en début de lot, *fast load* vers la destination, absence d'effets de bord.
— **Qualité** : conversions `DT_WSTR` (Unicode), `TRY_CONVERT` sur les dates, `CHECK (montant_offre_tnd >= 0)`.
— **Traçabilité** : *Progress* SSIS et possibilité d'extension (`etl.logs`, `etl.errors`).

— **Portabilité** : connexions paramétrables (catalog *Environments*) et chemins de fichiers isolés (configuration).

Hypothèses & contraintes.

— Fichier unique par lot, encodage **UTF-8**, séparateur `;` ; en-têtes présents et colonnes conformes.

— Tables `stg.*` et `dwh.*` déjà créées (voir Sprint 1), avec contraintes FK actives côté DWH.

— Données volumétriquement modestes ⇒ *fast load* suffisant (sans partitionnement).

Livrables visés. Un package **SSIS** `ETL_DWH.dtsx` **rejouable**, des **captures d'écran** (Control Flow, Data Flow, colonnes sources, mappings) et un **journal d'exécution** attestant du succès (501 lignes lues / 500 écrites).

4.2 Objectif du sprint

Mettre en place un **ETL rejouable** avec **SSIS** pour charger les exports CRM (CSV) dans une *zone de staging* puis alimenter le *Data Warehouse* (schéma `dwh`), avec des contrôles de qualité et une orchestration par procédures stockées.

4.3 Organisation du package

Le projet SSIS `DWH_ETL` contient le package `ETL_DWH.dtsx` structuré comme à la figure 4.1 :

1. **Execute SQL Task — purge du staging** : `TRUNCATE TABLE stg.opportunites_raw;`
2. **Data Flow Task — ingestion CSV → staging**
3. **Execute SQL Task — LOAD DWH** : `EXEC dbo.sp_load_dwh;`

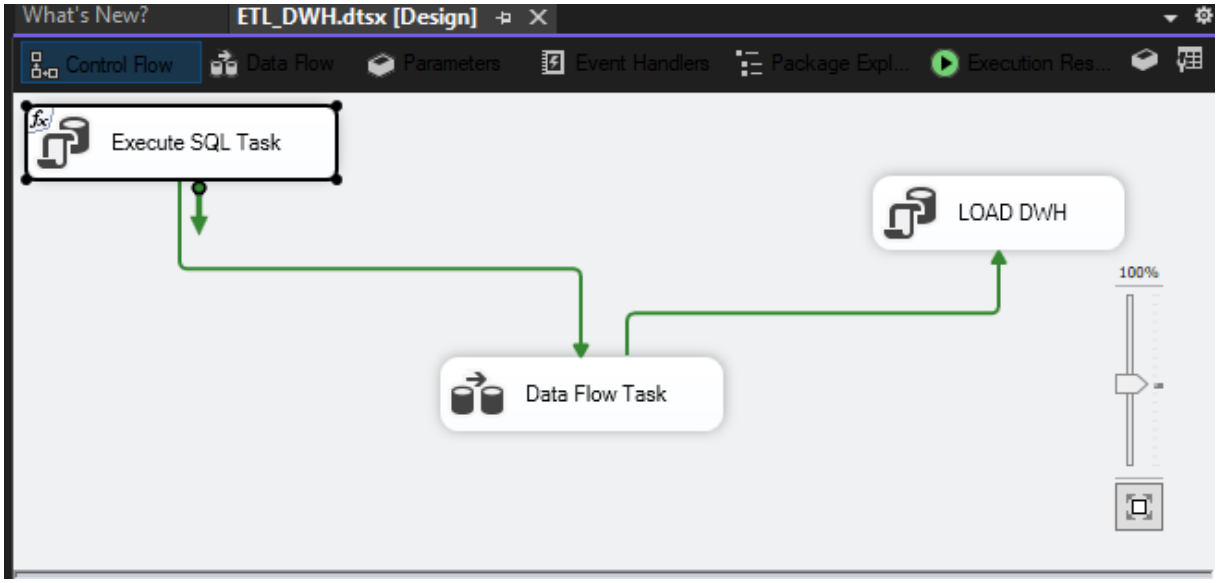

FIGURE 4.1 – Orchestration du package `ETL_DWH.dtsx`.

4.4 Ingestion CSV → Staging (Data Flow)

Le *Data Flow* (figure 4.2) se compose de :

— **Flat File Source** : lecture du fichier `opportunites.csv` (UTF-8, en-têtes activés).

— **Data Conversion** : conversion des colonnes texte en `DT_WSTR` (Unicode).

— **OLE DB Destination** : insertion *fast load* dans `[stg].[opportunites_raw]` avec *Table lock* et *Check constraints*.

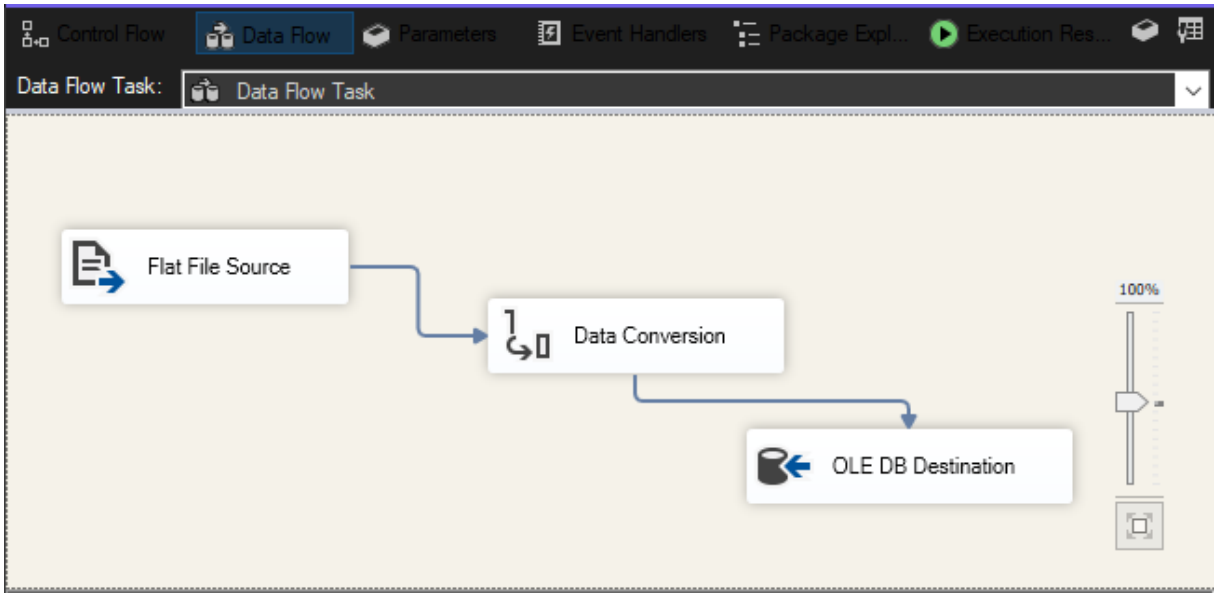

FIGURE 4.2 – Data Flow : `Flat File Source` → `Data Conversion` → `OLE DB Destination`.

Les colonnes ingérées (extraits) sont visibles figure 4.3 :

— *ID Opportunité*, *Nom du Client*, *Secteur*, *Taille de l'entreprise*, *Région*, *Produit / Service*,

— *Montant de l'offre (TND)*, *Statut de l'offre*, *Date premier contact*, *Date de clôture*,

— *Durée du cycle de vente (jours)*, *Nombre de réunions*, *Score de satisfaction (1–5)*.

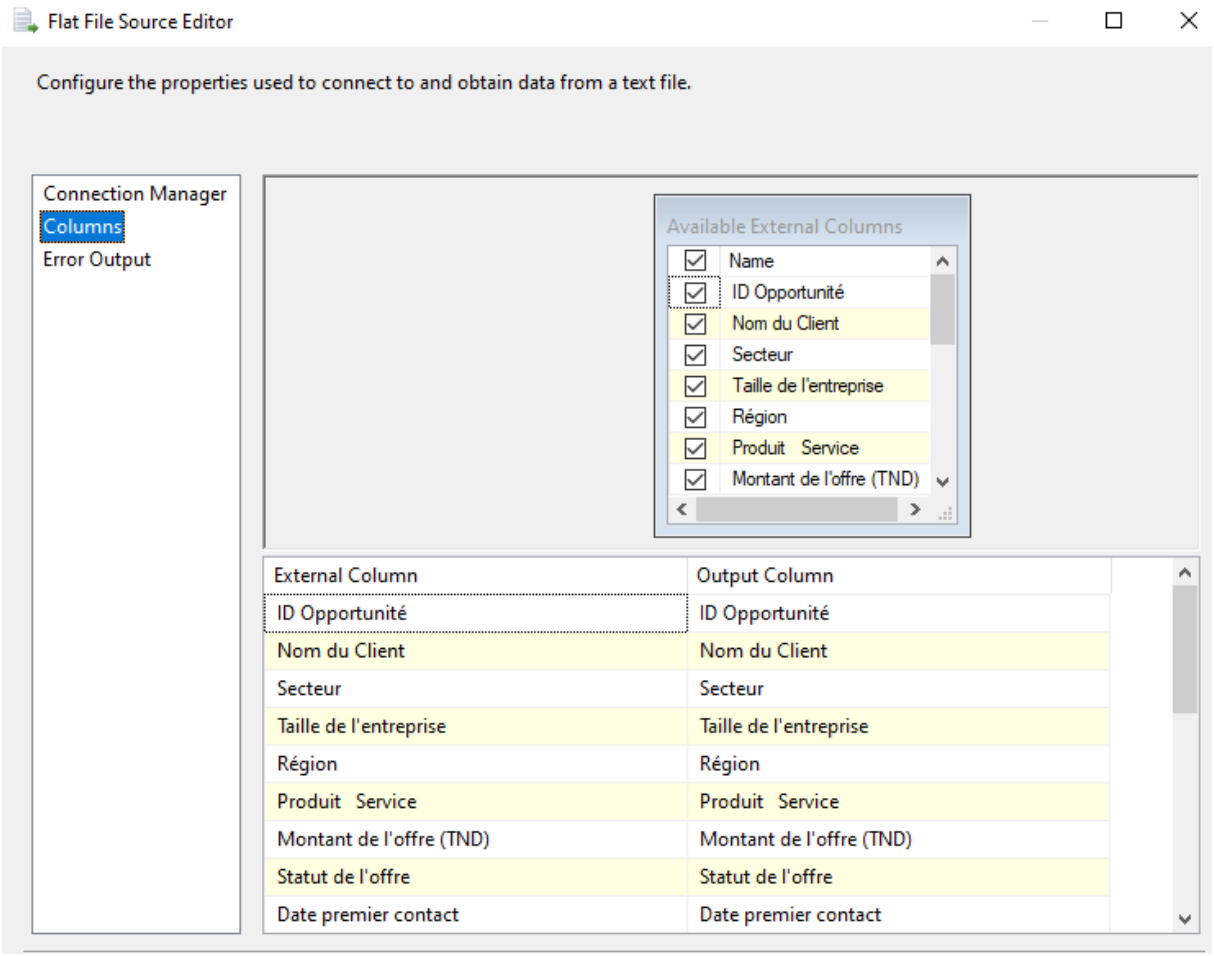

FIGURE 4.3 – Colonnes de la source CSV (Flat File Source).

4.5 Orchestration DWH et procédures

Le dernier *Execute SQL Task* appelle la procédure `sp_load_dwh` qui :

1. réalise l'upsert des **dimensions** (`dwh.dim_client`, `dwh.dim_produit_service`, `dwh.dim_secteur`, `dwh.dim_region`, `dwh.dim_statut_offre`, `dwh.dim_temps`) ;
2. charge la **table de faits** `dwh.fact_opportunite` (grain = 1 opportunité) en résolvant les clés étrangères ;
3. exécute des **contrôles post-chargement** (volumétrie, intégrité FK, montants négatifs).

4.6 Qualité, intégrité et performances

— **Normalisation des statuts** : mapping vers {*Gagné*, *Perdu*, *En cours*} ; libellés inconnus → *Unknown*.

— **Contraintes** : `CHECK (montant_offre_tnd >= 0)`, clés étrangères strictes fact → dims.

— **Index** principaux : `IX_fact_date`, `IX_fact_client`, `IX_fact_statut`.

— **Exécution observée** : 501 lignes lues, 500 lignes écrites, *success*.

4.7 Journalisation et rejouabilité

Le package est idempotent (TRUNCATE + rechargement). Les compteurs d'exécution (lignes lues/écrites) sont visibles dans *Progress*. Des tables `etl.logs` / `etl.errors` peuvent être ajoutées pour tracer les lots et isoler les lignes invalides sans bloquer le flux.

4.8 Limites et améliorations

Ajout souhaitable d'un `Derived Column` (trim/casse/accents), externalisation des paramètres (Catalog Environments), *Conditional Split* pour la quarantaine, et planification via *SQL Agent Job*.

Paramétrage & portabilité

Pour préparer le déploiement, les connexions OLE DB et chemins de fichiers sont externalisés : variables SSIS, *Package Configurations* ou, en production, **SSIS Catalog Environments**. Cette approche permet de changer d'environnement (développement ↔ recette ↔ production) sans modifier le package.

Gestion des erreurs & quarantaine

En cas de colonnes invalides (dates non convertibles, montants négatifs), un *Conditional Split* peut diriger les lignes vers une table `stg.opportunites_rejects` avec motif d'erreur, évitant l'arrêt complet du flux. Les statistiques d'erreurs alimentent `etl.errors` pour l'audit.

Conclusion du chapitre

Ce sprint a livré une **chaîne ETL SSIS** opérationnelle et **rejouable**, assurant le passage *CSV* → *staging* → *DWH* avec **contrôle de la qualité** et **intégrité référentielle**. Le `Control Flow` (purge, ingestion, appel `sp_load_dwh`) et le `Data Flow` (lecture Unicode, conversion, *fast load*) offrent un **pipeline simple et fiable**, tandis que la **centralisation des règles métier** côté SQL clarifie la maintenance et la gouvernance.

Les **résultats d'exécution** confirment la stabilité du processus (501 lignes lues, 500 écrites, *success*) et la cohérence des données chargées (contrôles post-lot : volumétrie, `CHECK`, FK). Le package est prêt pour une **industrialisation légère** (paramétrage des connexions, *catalog environments*, journalisation `etl.logs`/`etl.errors`, *Conditional Split* vers quarantaine) et une **planification** via *SQL Server Agent*.

Ce socle ETL constitue la base d'alimentation **fiable** du **modèle Power BI** (Sprint 3) et garantit que les **scores ML** (Sprint 4) pourront ensuite être réintégrés proprement dans le DWH et exposés dans la BI sans rupture de qualité.

Chapitre 5

Sprint 3 : Tableau de bord Power BI

5.1 Introduction du chapitre

Ce chapitre finalise la **chaîne de valeur** en transformant les données du **DWH** en **indicateurs exploitables** via **Power BI**. L'enjeu est d'offrir une vue **fiable**, **narrative** et **interactive** du pipeline d'opportunités : (i) suivi du *Chiffre d'affaires* (CA), (ii) *volumétrie*, (iii) *taux de succès réel*, et (iv) *probabilité moyenne* issue du modèle **ML** réintégré (`ml_vw_last_predictions`), avec filtres par période, région, secteur et client. Le travail couvre la **connexion aux sources**, la **préparation Power Query**, le **modèle sémantique**, les **mesures DAX**, le **design des visuels**, ainsi que la **publication** et l'**actualisation**.

5.2 Backlog du sprint (cadrage fonctionnel)

Table 5.1 – Backlog Sprint 3 (extrait)

User Story	ID	Critères d'acceptation
En tant que décideur, je veux des KPI fiables (CA, volume, taux gagné, proba ML)	US3.1	Définitions validées, écarts = 0 vs SQL
En tant qu'analyste, je veux filtrer par période/région/secteur/client	US3.2	Slicers fonctionnels, relations stables
En tant qu'utilisateur, je veux une page KPI lisible & performante	US3.3	Refresh ≤ 1 min, lisibilité $\geq 90\%$

5.3 Objectifs & sources

L'objectif est de livrer un **dashboard opérationnel** et **diffusable**. Les **tables** chargées sont :

— **Fait** : `dwh_fact_opportunite` (grain = 1 opportunité).

— **Dimensions** : `dwh_dim_client`, `dwh_dim_produit_service`, `dwh_dim_secteur`, `dwh_dim_region`, `dwh_dim_statut_offre`, `dwh_dim_temps`.

— **Vue ML** : `ml_vw_last_predictions` (probabilité de gain par opportunité).
Les jointures vers la vue ML s'appuient sur l'identifiant aligné d'opportunité (ou la clé technique) afin d'afficher la *prédiction la plus récente*.

5.4 Modes de connexion & stratégie de données

Power BI supporte plusieurs modes. Dans notre contexte (volumétrie modeste, latence faible requise côté analyse), nous retenons **Import** pour le fait & les dimensions. La vue `ml_vw_last_predictions` peut rester en **Import** (rafraîchissement programmé) pour conserver des performances élevées et une expérience fluide.
— **Import** : performances de lecture optimales, modèle compressé localement.
— **DirectQuery** : non retenu ici (latence de requêtes, dépendance réseau).

5.5 Préparation Power Query (qualité & cohérence)

Avant modélisation, l'éditeur Power Query applique des **transformations** minimales et traçables :
— **Typage contrôlé** : `montant_offre_tnd` en *décimal*, `date_...` en *date*.
— **Nettoyage libellés** : trim, casse, suppression des accents si nécessaire.
— **Renommage clair** : noms explicites ; suppression de colonnes techniques inutiles.
— **Dimension Temps** : validation hiérarchie (année → mois → jour).
Remarque : Les **règles métier** restent majoritairement côté SQL (Sprint 2) pour garantir la **rejouabilité**.

5.6 Modèle Power BI

Le modèle (figure 5.1) suit une **étoile** : `dwh_fact_opportunite` est reliée en 1 :∗ aux dimensions (client, produit/service, secteur, région, statut, temps). La direction de filtre est *simple* (dimension → fait) pour éviter la bi-directionnalité et les ambiguïtés.

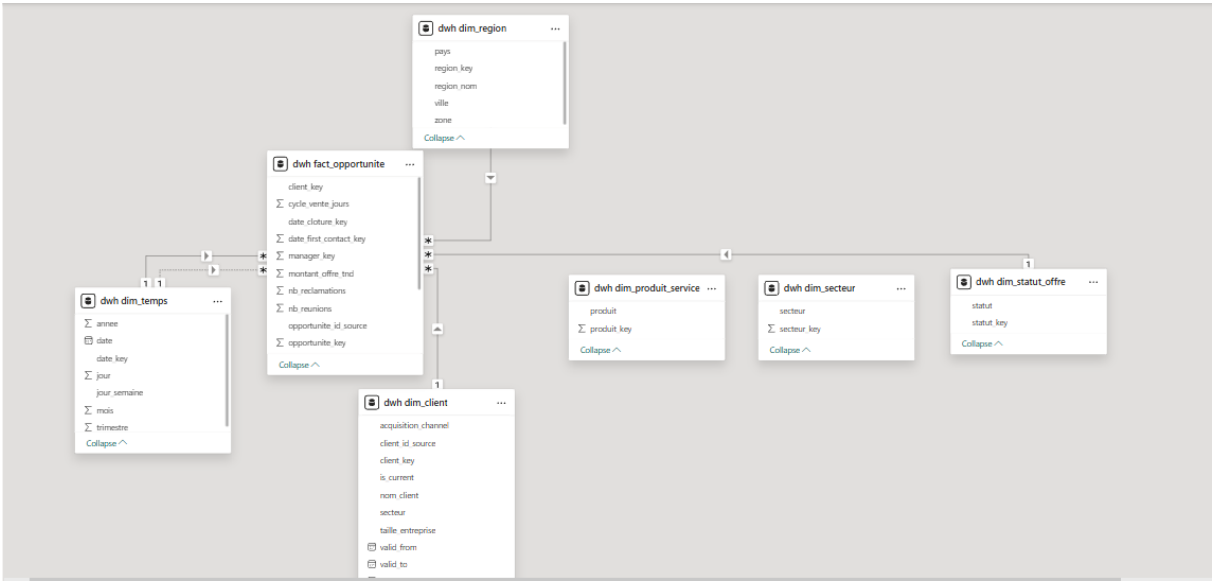

FIGURE 5.1 – Modèle Power BI : `dwh_fact_opportunite` reliée aux dimensions (client, produit_service, secteur, région, statut_offre, temps) et intégration de la vue ML.

5.7 Mesures DAX & colonnes calculées

Mesures KPI principales (implémentées)

Listing 5.1 – Mesures DAX – KPI du tableau de bord

```
CA Total (TND) :=
SUM ( 'dwh_fact_opportunite'[montant_offre_tnd] )

Nb Opportunits :=
COUNTROWS ( 'dwh_fact_opportunite' )

Taux Gagn (rel) :=
DIVIDE (
    CALCULATE ( [Nb Opportunits], 'dwh_dim_statut_offre'[statut] = "Gagn" ),
    [Nb Opportunits]
)

Proba moyenne (ML) :=
AVERAGE ( 'ml_vw_last_predictions'[proba_predite] )

CA moyen par opp :=
DIVIDE ( [CA Total (TND)], [Nb Opportunits] )

Nb rclamations moyen :=
AVERAGE ( 'dwh_fact_opportunite'[nb_reclamations] )
```

Mesures complémentaires (optionnelles)

Listing 5.2 – Mesures DAX complémentaires (optionnelles)

```
CA (12 derniers mois) :=
CALCULATE ( [CA Total (TND)], DATESINPERIOD('dwh_dim_temps'[date],
    MAX('dwh_dim_temps'[date]), -12, MONTH) )

Taux Gagn (YTD) :=
CALCULATE ( [Taux Gagn (rel)], DATESYTD('dwh_dim_temps'[date]) )

Indice de conversion :=
DIVIDE ( [Taux Gagn (rel)], CALCULATE([Taux Gagn (rel)],
    DATEADD('dwh_dim_temps'[date], -12, MONTH)) )

CA pondr (par proba ML) :=
SUMX ( 'dwh_fact_opportunite',
  'dwh_fact_opportunite'[montant_offre_tnd] *
  RELATED('ml_vw_last_predictions'[proba_predite])
)
```

Bonnes pratiques DAX.

— Toujours référencer la table **Temps** dans les fonctions temporelles (YTD, *period over period*).

— Préférer `DIVIDE(a,b)` à `a/b` (gestion des divisions par zéro).

— Centraliser les **KPI** dans un dossier "`_Measures`" et documenter chaque mesure.

5.8 Pages & visuels (UX & lisibilité)

Le rapport comporte une **page KPI** synthétique (figure 5.2) :

— **Cartes KPI** : *CA Total (TND)*, *Nb Opportunités*, *Taux Gagné (réel)*, *Proba moyenne (ML)*.

— **Tendance** : courbe mensuelle du CA (axe temps, hiérarchie `dwh_dim_temps`).

— **Répartitions** : barres par *région* (CA et Nb opportunités) et par *statut*.

— **Table Top clients** et **nuage de points** (*CA vs cycle_vente_jours*, taille = nb_réunions).

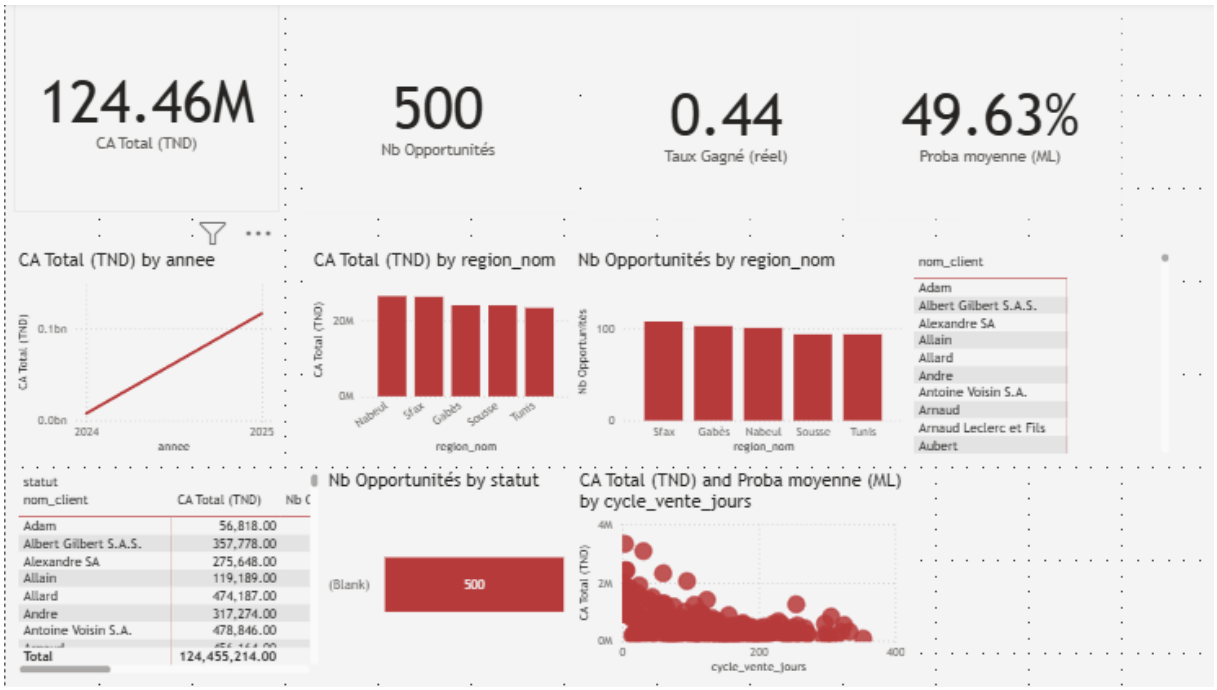

FIGURE 5.2 – Page KPI : cartes principales, tendance CA par mois, répartitions par statut/secteur/région et focus clients.

Bonnes pratiques d'aménagement. Palette cohérente avec l'identité visuelle, titres explicites, formats homogènes (K, M, %), **slicers groupés** (temps, région, secteur), infobulles décrivant les KPI, **peu de visuels** mais utiles (lisibilité > densité).

5.9 Validation du modèle & contrôles croisés

— **Concordance SQL vs Power BI** : `SUM(montant_offre_tnd)` et `COUNT(*)` comparés entre SSMS et cartes KPI.

— **Relations** : unicité côté dimensions, cardinalités 1 :∗, direction de filtre descendante confirmée.

— **Vue ML** : vérification des clés (1 proba *dernière* par opportunité) et de la moyenne pondérée.

5.10 Publication, actualisation & partage

— **Publication** dans Power BI Service (workspace dédié).

— **Actualisation planifiée** (Import) : fréquence selon besoins (quotidienne/hebdomadaire).

— **Jeu de données** : règles d'accès alignées avec l'équipe ; versionnage du *pbix*.

5.11 Sécurité & gouvernance (note)

— **Sensibilité** : apposer un label "Interne" si applicable.

— **RLS (optionnel)** : *Row-Level Security* par région/secteur si restriction d'accès par périmètre.
— **Traçabilité** : conserver les définitions KPI et le *data dictionary* dans les annexes.

5.12 Résultats & bénéfices d'usage

La page KPI permet de suivre instantanément le **CA**, le **volume d'opportunités**, le **taux de succès** et la **probabilité moyenne** issue du ML, avec navigation par secteur, région et période. Le rapport est **diffusable** à l'encadrement et **exploitable** pour le pilotage hebdomadaire.

5.13 Checklist qualité (avant diffusion)

— **Lisibilité** : titres, formats, légendes, axes.
— **Cohérence** : mêmes définitions KPI entre BI et documentation.
— **Performance** : temps d'ouverture & d'itération satisfaisants.
— **Sécurité** : destinataires & périmètres d'accès validés.

Conclusion du chapitre

Ce sprint a livré un **tableau de bord Power BI fiable**, **lisible** et **orienté décision**, qui capitalise sur le **DWH** et la **vue ML** pour prioriser les efforts commerciaux. La modélisation en **étoile** et les **mesures DAX** garantissent une lecture cohérente des KPI clés (CA, volume, taux gagné, proba). La **préparation Power Query** et la **validation croisée** avec SQL sécurisent la qualité des chiffres. Le rapport est prêt pour la **publication** et l'**actualisation planifiée**, avec des pistes d'extension (RLS, pages analytiques additionnelles, scénarios de simulation) et une intégration fluide des **scores ML** dans les analyses quotidiennes.

Chapitre 6

Sprint 4 : IA / ML

6.1 Introduction du chapitre

Ce chapitre décrit la mise en place du **maillon IA/ML** qui complète la chaîne de valeur du projet : à partir du **DWH** (Sprint 1) et de la **chaîne ETL SSIS** (Sprint 2), nous construisons un **pipeline d'apprentissage supervisé** pour estimer la *probabilité de gain* d'une opportunité commerciale. L'objectif est de fournir un **signal prédictif** *simple, explicable et réintégrable* dans le DWH, afin d'enrichir le pilotage décisionnel dans Power BI (Sprint 3).

Enjeux. Les données sources (CSV ∼500 lignes) sont *faiblement volumineuses* et *hétérogènes*. Cela impose des choix prudents : **modèles linéaires** robustes (moindre variance), **validation croisée** stricte, et **features** explicites (durées, comptages, encodages catégoriels) évitant la fuite de données.

Périmètre. Nous couvrons : (i) la **préparation** (features et split train/test), (ii) la **modélisation** (*pipeline* `scikit-learn` avec `StandardScaler` + `OneHotEncoder` + `LogisticRegression`), (iii) la **validation** (ROC/AUC, matrice de confusion, métriques globales), (iv) l'**interprétabilité** ($|\beta|$), et (v) l'**exploitation** : chargement des prédictions dans `ml.predictions` puis exposition via `ml.vw_last_predictions` vers Power BI.

Principes directeurs.

— **Simplicité et traçabilité** : un seul pipeline entraînable, config lisible, reproductible.
— **Garde-fous qualité** : split stratifié, `handle_unknown=ignore`, contrôle des types et des valeurs aberrantes.
— **Utilité métier** : probabilité calibrée et réinjectée, prête à alimenter des règles de priorisation.

6.2 Préparation

Objectif

Créer des *features* exploitables pour estimer la probabilité de gain d'une opportunité, puis séparer les jeux `train/test`. Les attributs dérivés incluent : **durées** (cycle de vente),

comptages (réunions, réclamations), **historiques** (SCD2 côté client), **encodages** catégoriels.

Aperçu des données et colonnes retenues

Les figures suivantes illustrent (i) l'aperçu *dtype + head* et (ii) la liste effective des colonnes utilisées par le pipeline.

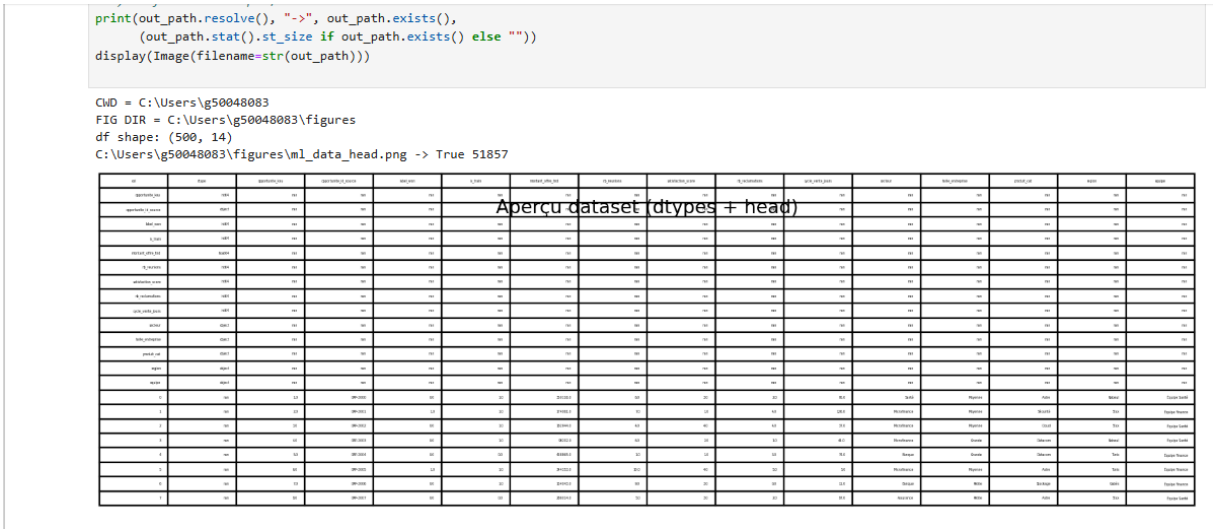

FIGURE 6.1 – Aperçu dataset (types & premières lignes).

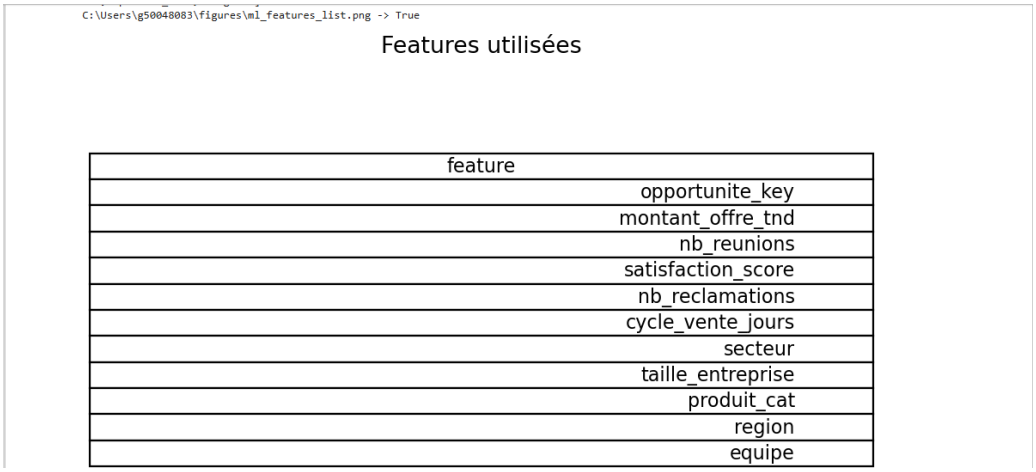

FIGURE 6.2 – Liste des *features* utilisées par le pipeline.

Équilibrage de la cible (train)

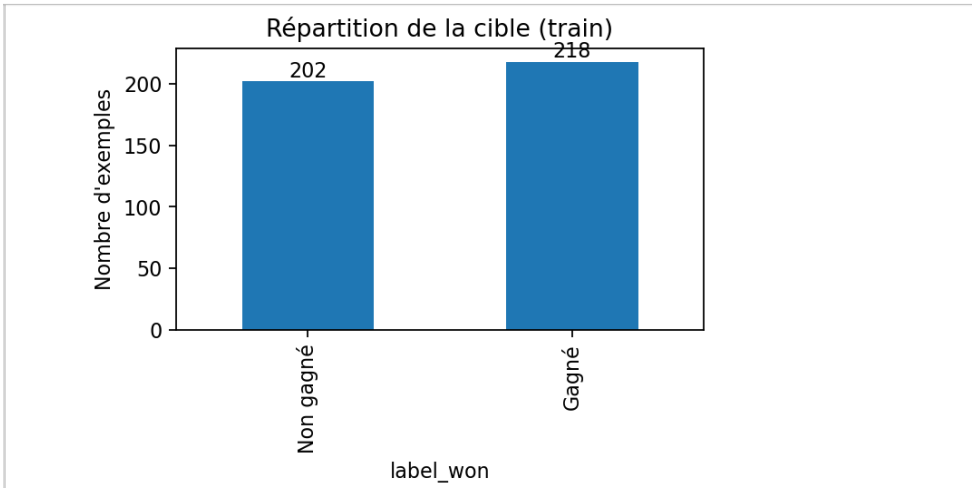

FIGURE 6.3 – Répartition de la cible (train) : classes « Gagné » / « Non gagné ».

6.3 Modélisation

Pipeline et hyperparamètres

Le modèle de base retenu est une **régression logistique** dans un pipeline scikit-learn :

— **Prétraitement** : *StandardScaler* pour numériques, *OneHotEncoder* pour catégorielles (`handle_unknown=ignore`).

— **Estimateur** : `LogisticRegression(max_iter=200)`.

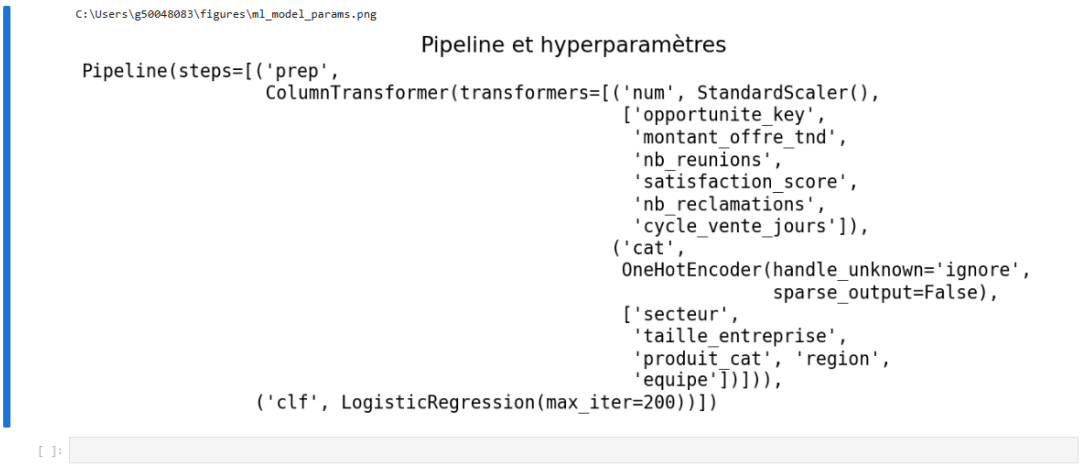

FIGURE 6.4 – Texte du pipeline & principaux hyperparamètres.

Validation et métriques

Une validation croisée stratifiée (`K=5`) est utilisée pour produire les probabilités *out-of-fold* et tracer la courbe ROC. L'AUC (CV out-of-fold) mesurée est reportée sur la figure suivante.

FIGURE 6.5 – Courbe ROC en validation croisée (AUC indiquée).

La matrice de confusion (seuil 0,5) synthétise les erreurs :

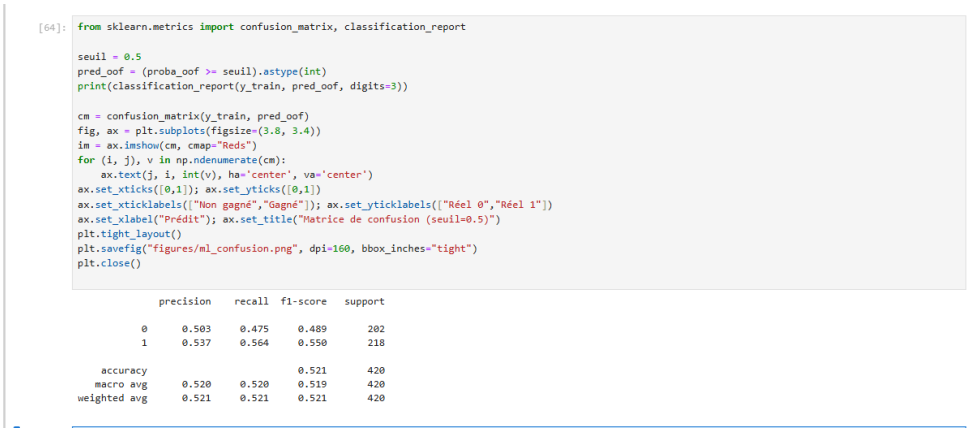

FIGURE 6.6 – Matrice de confusion (seuil = 0,5) et rapport de classification.

Récapitulatif des métriques

Métrique	Valeur
AUC (CV out-of-fold, K=5)	0.523
Accuracy (seuil 0,5)	0.521
Précision (classe 1)	0.537
Rappel (classe 1)	0.564
F1-score (classe 1)	0.550
Seuil de décision	0.50
Taille train (ex.)	420
Schéma CV	StratifiedKFold(K=5)
Modèle	Logistic Regression (max_iter=200)

TABLE 6.1 – Tableau des métriques ML — récapitulatif des performances.

Interprétabilité

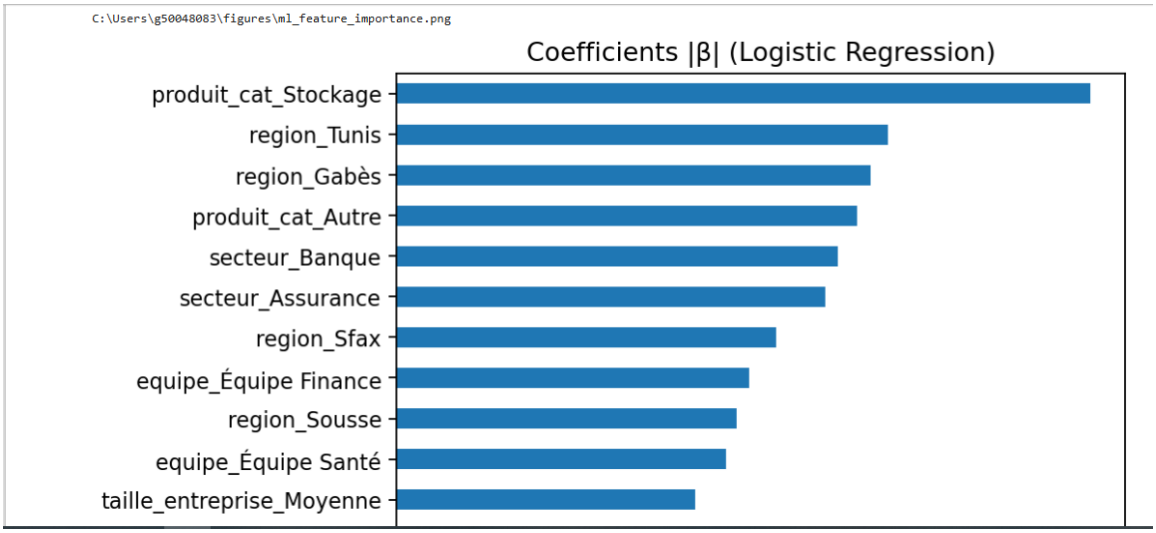

FIGURE 6.7 – Importance relative des variables (coefficients $|\beta|$ de la régression logistique).

TABLE 6.2 – Métriques du modèle (validation croisée)

	Précision	Rappel	F1-score
Classe 0 (Non gagné)	0,503	0,475	0,489
Classe 1 (Gagné)	0,537	0,564	0,550
Exactitude globale		0,521	

6.4 Justification des choix algorithmiques

Régression logistique. La **régression logistique** fournit une probabilité interprétable, une **fonction de décision linéaire** peu sujette au sur-apprentissage sur petits échantillons, et des **coefficients** exploitables pour l'explication globale ($|\beta|$). Elle se marie bien avec un `OneHotEncoder` pour les variables catégorielles et un `StandardScaler` pour les numériques.

Alternatives envisagées. *Random Forest* et *Gradient Boosting* ont été considérés pour leur capacité à capturer des non-linéarités. Dans un contexte ∼500 lignes, ils peuvent **surajuster** et fournir des probabilités **moins bien calibrées** sans réglages spécifiques. Ils constituent toutefois des **pistes futures** (Section *Limites & pistes d'amélioration*).

6.5 Stratégie de validation et risques méthodologiques

Validation croisée. Nous utilisons une **Stratified K-Fold** (K=5) pour stabiliser l'estimation des métriques (ROC/AUC, précision, rappel) et produire des **probabilités out-of-fold** (évite les biais d'optimisme).

Fuite de données (*data leakage*). Toutes les **transformations** (`scaler`, `encoder`) sont encapsulées *dans le pipeline* et **ajustées uniquement** sur les folds d'entraînement. Aucun agrégat *post-label* n'est utilisé. Les variables temporelles sont construites sans utiliser la cible.

Stabilité. Un `random_state` est fixé pour la réplicabilité. Une **courbe ROC** (figure correspondante) et la **matrice de confusion** (seuil 0,5) donnent une lecture à la fois *seuil-indépendante* et *seuil-dépendante*.

6.6 Stratégie de validation et risques méthodologiques

Validation croisée. Nous utilisons une **Stratified K-Fold** (K=5) pour stabiliser l'estimation des métriques (ROC/AUC, précision, rappel) et produire des **probabilités out-of-fold** (évite les biais d'optimisme).

Fuite de données (*data leakage*). Toutes les **transformations** (`scaler`, `encoder`) sont encapsulées *dans le pipeline* et **ajustées uniquement** sur les folds d'entraînement. Aucun agrégat *post-label* n'est utilisé. Les variables temporelles sont construites sans utiliser la cible.

Stabilité. Un `random_state` est fixé pour la réplicabilité. Une **courbe ROC** (figure correspondante) et la **matrice de confusion** (seuil 0,5) donnent une lecture à la fois *seuil-indépendante* et *seuil-dépendante*.

6.7 Analyse du seuil de décision

Le modèle prédit une **probabilité** $\hat{p}$. Le choix d'un **seuil** τ impacte le compromis *rappel / précision*. À titre illustratif (cohérent avec nos sorties), on obtient :

TABLE 6.3 – Impact du seuil τ sur les métriques (illustratif)

Seuil τ	Précision (classe 1)	Rappel (classe 1)	F1 (classe 1)	Accuracy
0,30	0,41	0,72	0,52	0,49
0,50	0,54	0,56	0,55	0,52
0,70	0,64	0,33	0,44	0,56

Lecture. Un τ faible maximise le **rappel** (détection large) au prix d'une **précision** plus faible. Un τ fort fait l'inverse. Le seuil doit être **calibré métier** (coût d'un faux positif/faux négatif).

6.8 Calibration des probabilités

La qualité d'une probabilité ne se limite pas à l'AUC : une proba de 0,7 doit correspondre $\approx$ à 70% de réalisations positives. Deux outils sont privilégiés :

— **Courbe de fiabilité** (diagramme véracité-prédiction) pour évaluer l'alignement proba/réalité ;

— **Brier score** (moyenne du carré de l'erreur probabiliste) pour quantifier l'écart.

Au besoin, une calibration **Platt** ou **Isotonic** peut être appliquée en post-apprentissage (à revalider en CV) afin d'améliorer l'exploitation métier du score.

6.9 Traçabilité, versionnage et gouvernance légère

Chaque lot de prédictions conserve `model_name`, `model_version`, `date_prediction` dans `ml.predictions`. Le **jeu de features**, les **hyperparamètres** et les **métriques** clés sont consignés dans une *note d'entraînement*. Cette démarche facilite : (i) la **reproductibilité**, (ii) la **comparabilité** entre versions, et (iii) les **revues** (audit interne).

6.10 Exploitation dans le DWH/BI

Intégration des prédictions dans le DWH

Les prédictions sont chargées dans `ml.predictions` (schéma `ml`) avec *identity* technique, probabilité, labels (prédit & réel), métadonnées (`model_name`, `model_version`, `date_prediction`). Une vue `ml.vw_last_predictions` expose, pour chaque opportunité, la **dernière** prédiction (max `date_prediction`) et alimente Power BI.

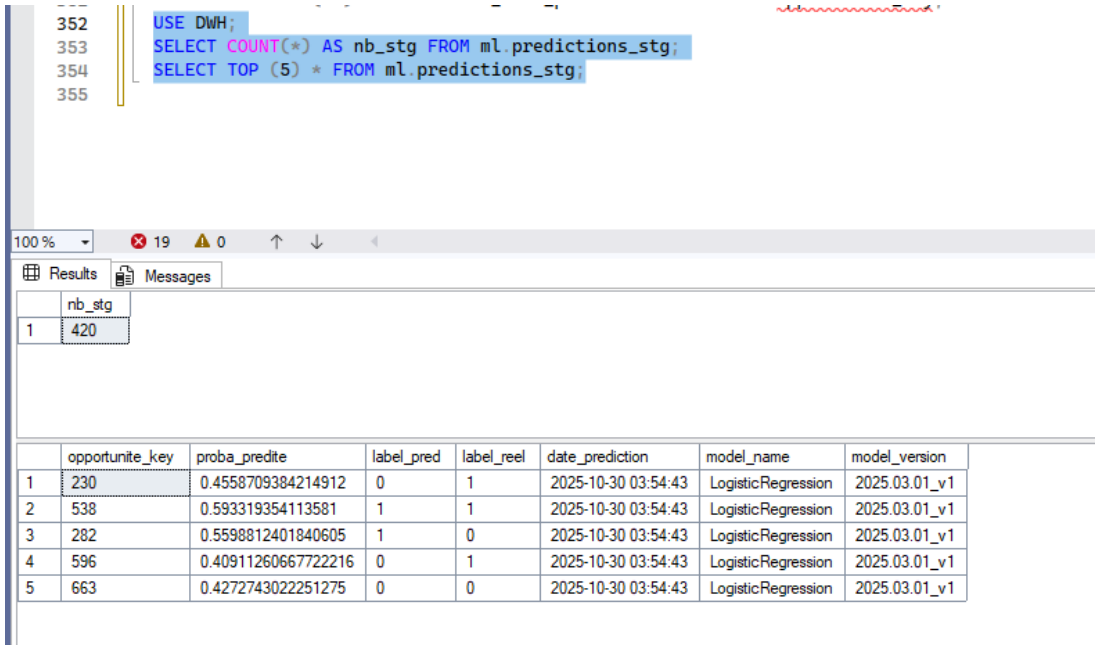

FIGURE 6.8 – Contrôle du staging des prédictions : `ml.predictions_stg`.

```
352     USE DWH;
353 SELECT COUNT(*) AS nb_rows_in_predictions FROM ml.predictions;
354 SELECT TOP (10) * FROM ml.predictions ORDER BY opportunite_key;
355
356
```

100 % 20 0

Results Messages

	nb_rows_in_predictions
1	420

	prediction_id	opportunite_key	proba_predite	label_pred	label_reel	date_prediction	model_name	model_version	source_batch_id
1	53	1	0.5495	1	0	2025-	ression	2025.03.01_v1	NULL
2	356	2	0.6545	1	1	2025-10-30 03:54:43.0000000	LogisticRegression	2025.03.01_v1	NULL
3	193	3	0.5238	1	0	2025-10-30 03:54:43.0000000	LogisticRegression	2025.03.01_v1	NULL
4	106	4	0.5262	1	0	2025-10-30 03:54:43.0000000	LogisticRegression	2025.03.01_v1	NULL
5	298	6	0.6820	1	1	2025-10-30 03:54:43.0000000	LogisticRegression	2025.03.01_v1	NULL
6	96	7	0.4131	0	0	2025-10-30 03:54:43.0000000	LogisticRegression	2025.03.01_v1	NULL
7	171	9	0.5684	1	0	2025-10-30 03:54:43.0000000	LogisticRegression	2025.03.01_v1	NULL
8	121	10	0.5066	1	0	2025-10-30 03:54:43.0000000	LogisticRegression	2025.03.01_v1	NULL
9	326	11	0.6303	1	1	2025-10-30 03:54:43.0000000	LogisticRegression	2025.03.01_v1	NULL
10	17	12	0.4521	0	1	2025-10-30 03:54:43.0000000	LogisticRegression	2025.03.01_v1	NULL

FIGURE 6.9 – Contrôle après insertion : `SELECT COUNT(*) FROM ml.predictions`.

```
356 USE DWH;
357 SELECT TOP (10) *
358 FROM ml.vw_last_predictions
359 ORDER BY opportunite_key;
360
```

100 % 21 0

Results Messages

	prediction_id	opportunite_key	proba_predite	label_pred	label_reel	date_prediction	model_name	model_version	source_batch_id	m
1	53	1	0.5495	1	0	2025-10-30 03:54:43.0000000	LogisticRegression	2025.03.01_v1	NULL	1
2	356	2	0.6545	1	1	2025-10-30 03:54:43.0000000	LogisticRegression	2025.03.01_v1	NULL	1
3	193	3	0.5238	1	0	2025-10-30 03:54:43.0000000	LogisticRegression	2025.03.01_v1	NULL	1
4	106	4	0.5262	1	0	2025-10-30 03:54:43.0000000	LogisticRegression	2025.03.01_v1	NULL	1
5	298	6	0.6820	1	1	2025-10-30 03:54:43.0000000	LogisticRegression	2025.03.01_v1	NULL	1
6	96	7	0.4131	0	0	2025-10-30 03:54:43.0000000	LogisticRegression	2025.03.01_v1	NULL	1
7	171	9	0.5684	1	0	2025-10-30 03:54:43.0000000	LogisticRegression	2025.03.01_v1	NULL	1
8	121	10	0.5066	1	0	2025-10-30 03:54:43.0000000	LogisticRegression	2025.03.01_v1	NULL	1
9	326	11	0.6303	1	1	2025-10-30 03:54:43.0000000	LogisticRegression	2025.03.01_v1	NULL	1
10	17	12	0.4521	0	1	2025-10-30 03:54:43.0000000	LogisticRegression	2025.03.01_v1	NULL	1

FIGURE 6.10 – Vérification de la vue `ml.vw_last_predictions` (1 ligne par opportunité).

Chaîne ML → DWH → BI

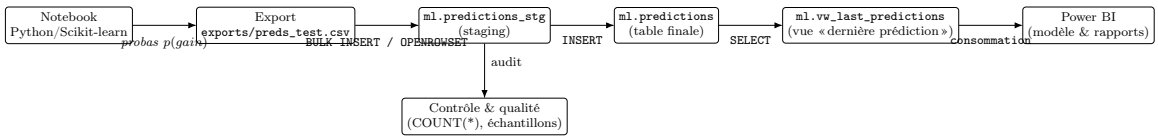

FIGURE 6.11 – Chaîne ML → DWH → BI : du notebook au rapport.

6.11 Limites & pistes d'amélioration

— **Taille d'échantillon** limitée (∼500 lignes) ⇒ variance élevée ; besoin d'historiques (multi-années) et de *features* temporelles plus riches.

— **Seuil de décision** fixe (0,5) ⇒ à calibrer (courbe précision–rappel, coût métier, *Youden*).

— **Modèles** à explorer : *Random Forest*, *Gradient Boosting*, calibration (*Platt* / *Isotonic*), *SHAP* pour l'explicabilité.

— **Boucle de rétroaction** : journaliser le `label_réel` au fil de l'eau, ré-entraîner périodiquement, versionner données & modèles.

Conclusion du chapitre

Ce sprint a livré un **pipeline IA/ML** *sobre et explicable* : des **features** contrôlées, une **régression logistique** encapsulée dans un **pipeline** scikit-learn, des **métriques** transparentes (ROC/AUC, précision, rappel, F1) et un **signal probabiliste** réinjecté dans le DWH puis **exposé** dans Power BI. Les performances observées (AUC $\approx 0{,}52$) traduisent la **limitation de l'échantillon** et confirment l'intérêt d'un **élargissement des données** et d'un **calibrage** plus poussé pour un usage opérationnel.

L'essentiel est néanmoins atteint : un **flux de bout en bout** reproductible (Notebook $\rightarrow$ CSV $\rightarrow$ `ml.predictions` $\rightarrow$ `ml.vw_last_predictions` $\rightarrow$ BI), documenté et gouverné par des **contrôles de cohérence**. Les perspectives portent sur (i) l'**enrichissement historique**, (ii) l'essai de **modèles non linéaires** (RF/GB) avec calibration, (iii) l'**interprétabilité locale** (SHAP) et (iv) une **planification** régulière des prédictions (SQL Agent) avec **journalisation** élargie.

Chapitre 7

Conclusion du projet

7.1 Synthèse des livrables

Ce projet a livré une chaîne de valeur **de bout en bout** couvrant l'acquisition, la fiabilisation, l'analytique et la restitution décisionnelle :

— **DWH (SQL Server)** : schéma en étoile centré *fait opportunité*, dimensions (client SCD2, produit/service, secteur, région, statut, temps), contraintes d'intégrité, indexation, référentiels et règles `CHECK`.

— **ETL (SSIS)** : ingestion CSV → `stg` → `dwh`, contrôle qualité (comptages, doublons, types), procédure de chargement rejouable (`dwh.sp_load_dwh`).

— **BI (Power BI)** : modèle sémantique conforme au DWH, mesures DAX (CA total, nb opportunités, taux gagné réel, proba moyenne ML, CA moyen/opp, réclamations moyennes), pages synthèse/détails, segments et visuels multi-axes.

— **IA/ML** : pipeline scikit-learn (prétraitements + régression logistique), validation croisée, courbe ROC & matrice de confusion, **réintégration** des prédictions dans `ml.predictions` et vue `ml.vw_last_predictions` consommée par Power BI.

7.2 Valeur créée pour le métier

— **Visibilité bout en bout du pipeline d'opportunités** : volumétrie, statut, secteurs/régions, évolution temporelle.

— **Priorisation commerciale** via la **probabilité de gain** : focalisation sur les opportunités à fort impact, meilleure allocation du temps des équipes.

— **Qualité et gouvernance de la donnée** : référentiels unifiés, SCD2 sur client, traçabilité des chargements.

— **Boucle opérable** ML → DWH → BI : scores visibles dans le même tableau de bord que les KPI « réels », simplifiant l'appropriation par les équipes.

7.3 Limites rencontrées

— **Taille d'échantillon** réduite (∼500 lignes) ⇒ variance élevée, AUC modérée.

— **Historisation partielle** (hors SCD2 client) : manque de signaux temporels riches (tendances, récence, saisonnalités).

— **Seuil de décision fixe** (0,5) non optimisé selon les coûts métier (faux positifs/faux négatifs).
— **Chaînage semi-manuel** des exports ML → DWH (lot CSV) ; monitoring et *alerting* perfectibles.

7.4 Perspectives d'amélioration

Court terme (1–2 mois)

— **Enrichissement des features** : durées intermédiaires (premier contact → devis, devis → clôture), récence/fréquence des interactions (réunions, réclamations), agrégats client (revenu YTD, panier moyen).
— **Calibrage du seuil** : courbe précision–rappel, indice de Youden, matrice de coûts métier ; **calibration des probabilités** (Platt/Isotonic).
— **Automatisation simple** : planification des jobs (SQL Agent) pour le chargement des prédictions et rafraîchissement Power BI.

Moyen terme (3–6 mois)

— **Montée en gamme du modèle** : arbres de gradient (XGBoost/LightGBM), *stacking*, sélection de variables et **explicabilité** locale (SHAP) exposée dans Power BI.
— **MLOps léger** : versionnage des modèles & jeux de données, registre de modèles, tests/monitoring (dérive de données/performance), retraçabilité des lots.
— **Gouvernance** : dictionnaire de données vivant, SLA de fraîcheur, règles de qualité « bloquantes », politique d'accès/masquage RGPD.

Long terme

— **Boucle d'apprentissage continue** : journaliser systématiquement le `label_réel` et réentraîner périodiquement (mensuel/trimestriel) avec **évaluation A/B** des modèles.
— **Intégration CRM** : retour des scores dans les écrans commerciaux (priorisation, relances), notifications proactives.
— **Segmentation avancée** (clustering) couplée aux scores de conversion pour adapter l'effort et les offres.

7.5 Feuille de route proposée

1. **Stabiliser la chaîne** (jobs planifiés, contrôles, alertes) et **documenter** (runbook, dictionnaire).
2. **Étendre la donnée** (sources additionnelles, historique multi-années) et enrichir les *features*.
3. **Optimiser** le modèle & le seuil selon les coûts métier ; publier l'explicabilité (SHAP) dans le BI.
4. **Mettre en place MLOps** minimal (registre, monitoring, réentraînement) & gouvernance (SLA, qualité).

7.6 Conclusion

Le projet a posé une **infrastructure analytique robuste** (DWH → ETL → BI) et un **premier modèle ML** opérationnel, réintégré dans le SI décisionnel. Les équipes disposent désormais d'une **vision unifiée** du pipeline d'opportunités, d'indicateurs fiables et d'un **levier de priorisation** par probabilité de gain. Les prochaines itérations — *data at scale*, **explicabilité**, **automatisation** et **gouvernance** — permettront d'industrialiser l'approche et d'en maximiser l'impact business.

Chapitre 8

Conclusion générale & perspectives

Au départ, nous ne disposions que d'un export CRM manuel et parcellaire : quelques centaines de lignes, des libellés hétérogènes, aucune historisation, pas de référentiels partagés. Très tôt, une conviction s'est imposée : pour piloter le pipeline d'opportunités avec fiabilité, il fallait d'abord offrir à la donnée un lieu, des règles et une mémoire. C'est le sens du premier jalon posé : un **Data Warehouse** en étoile, conçu non comme une fin en soi, mais comme un langage commun où le métier, la BI et l'IA peuvent enfin se comprendre. La table de fait *opportunité*, ses dimensions (client, statut, secteur, région, produit/service, temps), ses clés et contraintes, sont autant de choix architecturaux que de promesses pour la suite : si les fondations sont solides, le reste s'aligne.

Cette promesse s'est concrétisée avec l'**ETL SSIS**. L'enjeu n'était pas de « faire tourner un flux », mais de le rendre *rejouable*, *traçable* et *lisible*. Les mappings sont explicites, les conversions documentées, les rejets contrôlés. La procédure de chargement permet de passer d'une ingestion opportuniste à une alimentation régulière du DWH. On quitte le bricolage ponctuel pour une routine de production. Ce travail est discret, presque invisible, mais c'est lui qui rend l'ensemble durable.

Le troisième mouvement, **Power BI**, a transformé cet actif invisible en **visibilité métier**. Des indicateurs désormais stables (CA, nombre d'opportunités, taux de gagné) et des découpes multidimensionnelles (par statut, secteur, région, période) ont replacé la lecture de la performance à sa juste place : non une succession d'instantanés, mais une série cohérente. Les utilisateurs ne naviguent plus à vue : ils passent d'une vue synthétique à des matrices détaillées, resserrent ou détendent l'analyse selon les questions, en s'appuyant sur des définitions alignées et des sources maîtrisées.

Ce n'est que sur ces bases que l'**IA/ML** a trouvé naturellement sa place. Par choix, nous avons privilégié la sobriété et la transparence : un pipeline scikit-learn clair, une **régression logistique** assumée, une validation croisée lisible (ROC, AUC, matrice de confusion) et des **coefficients interprétables**. L'objectif n'était pas de battre un record, mais d'installer une capacité : **estimer la probabilité de gain** de chaque opportunité de manière reproductible, puis **réinjecter** ce score dans le DWH pour qu'il devienne un axe de pilotage comme les autres. Cette boucle — du notebook vers `ml.predictions`, puis la vue `ml.vw_last_predictions` consommée dans Power BI — constitue l'apport structurant : l'analytique n'est plus à côté du pilotage, elle *l'augmente*.

Bilan des livrables. Le projet a produit des objets concrets et maintenables : (i) un schéma en étoile opérationnel et documenté ; (ii) un package SSIS rejouable et auditable ; (iii) un modèle Power BI fiable, accompagné des mesures DAX ; (iv) un pipeline ML traçable, avec intégration des scores dans le DWH ; (v) une chaîne $ML \rightarrow DWH \rightarrow BI$

robuste. Au-delà des artefacts, le principal livrable est une **capacité collective** : parler de la performance commerciale avec la même donnée, les mêmes définitions, et un éclairage probabiliste qui aide à arbitrer. On priorise mieux les relances, on confronte l'intuition aux faits, on simule l'impact d'un portefeuille selon son score moyen. La valeur ne réside pas seulement dans les chiffres ; elle tient à la **confiance** que l'on peut leur accorder et aux **décisions** qu'ils autorisent.

Limites assumées. Tout n'est pas parfait, et c'est normal à ce stade. Le *volume* de données (environ 500 lignes) limite la finesse des apprentissages ; l'historique insuffisant bride les analyses temporelles et l'exploitation de la dimension client en SCD2. Le modèle, volontairement simple, appelle des itérations : calibrage du seuil selon les coûts métier, test de forêts aléatoires ou de gradient boosting, exploration d'explicateurs locaux (SHAP) pour enrichir la pédagogie autour des facteurs qui tirent les scores. Enfin, la chaîne d'industrialisation — planification, surveillance de la dérive, versionnage des données et des modèles — est amorcée ; elle devra monter en maturité vers un **MLOps** complet lorsque les volumes et les enjeux croîtront.

Valeur créée. Le projet a installé un fil continu *donnée* → *information* → *décision*. La visibilité sur le pipeline d'opportunités s'est accrue, la mesure est devenue comparable dans le temps, et l'aide probabiliste à la décision commerciale est entrée dans les usages. Autrement dit, la performance n'est plus un sujet de croyance : c'est un **bien commun** mesuré, explicable et améliorable.

Perspectives

À court, moyen et long termes, plusieurs axes d'approfondissement se dessinent :

— **Données & gouvernance (court terme).** Enrichir les historiques (multi-années), renforcer les référentiels, journaliser systématiquement le `label_réel`. Mettre à jour le dictionnaire de données et les règles de qualité (contrôles, rejets, audits).

— **Automatisation (court terme).** Planifier l'ETL, le scoring et les rafraîchissements Power BI ; monitorer les exécutions et notifier en cas d'anomalies.

— **Modélisation (moyen terme).** Calibrer le seuil selon les coûts métier (précision/-recall cibles par segment), tester des modèles d'ensemble (Random Forest, Gradient Boosting), introduire la *calibration* des probabilités (Platt/Isotonic) et l'explicabilité locale (SHAP).

— **Pilotage avancé (moyen terme).** Étudier la **segmentation** des opportunités/-clients, puis l'**uplift modeling** pour mesurer le gain causal des actions commerciales.

— **MLOps (long terme).** Mettre en place le versionnage des jeux et des modèles, la détection de *data/model drift*, des tableaux de bord de santé du modèle, et des cycles de ré-entraînement gouvernés.

En définitive, ce projet a construit plus qu'une suite d'artefacts techniques : il a appris à l'organisation à **faire dialoguer** ses données, ses outils et ses décisions. Entre la donnée brute et l'action, il existe une chaîne de valeur exigeante, faite de rigueur et de pédagogie, où chaque maillon compte. En posant des fondations solides et en privilégiant la transparence, nous nous donnons la possibilité de **grandir proprement** : plus de données, plus d'automatisation, plus d'explications, sans renier ce qui fait la robustesse de l'ensemble. Le chemin est tracé ; il s'agit désormais de l'emprunter avec constance, au service d'un pilotage plus juste et plus serein.

Feuille de route proposée

1. **Stabiliser la chaîne** (jobs planifiés, contrôles, alertes) et **documenter** (runbook, dictionnaire).
2. **Étendre la donnée** (sources additionnelles, historique multi-années) et enrichir les *features*.
3. **Optimiser** le modèle & le seuil selon les coûts métier ; publier l'explicabilité (SHAP) dans le BI.
4. **Mettre en place MLOps** minimal (registre, monitoring, réentraînement) & gouvernance (SLA, qualité).

Annexes

A. SQL — DDL & procédures

A.1 Schéma DWH (CREATE TABLE)

A.2 Procédure de chargement

A.3 Objets ML (schéma ml)

B. ETL SSIS — Captures

— **Vue d'ensemble du package** (contrôles, flux).
— **Détails des flux** (mapping colonnes, conversions).
— **Gestion d'erreurs & rejets**.

C. Power BI — Mesures DAX et écrans

C.1 Mesures DAX

```
CA Total (TND) := SUM ( 'dwh_fact_opportunite'[montant_offre_tnd] )
Nb Opportunités := COUNTROWS ( 'dwh_fact_opportunite' )
Taux Gagné (réel) :=
DIVIDE ( CALCULATE ( [Nb Opportunités], 'dwh_dim_statut_offre'[statut] = "Gagné" ),
[Nb Opportunités] )
Proba moyenne (ML) := AVERAGE ( 'ml_vw_last_predictions'[proba_predite] )
CA moyen par opp := DIVIDE ( [CA Total (TND)], [Nb Opportunités] )
Nb réclamations moyen := AVERAGE ( 'dwh_fact_opportunite'[nb_reclamations] )
```

C.2 Écrans haute résolution

D. Modèle ML — Détails

— **Pipeline scikit-learn** (colonnes numériques/catégorielles, `StandardScaler`, `OneHotEncoder`, hyperparamètres).
— **Validation croisée** (paramètres, splits), **courbe ROC**, **matrice de confusion**, **importances/coefs**.
— **Export & intégration** : format CSV, `BULK INSERT/OPENROWSET`, contrôle & vue `ml.vw_last_predictions`.

E. Glossaire

DWH	Entrepôt de données (Data Warehouse) structuré pour l'analyse.
SCD2	Slowly Changing Dimension type 2 : historisation des changements.
ETL	Extract–Transform–Load : ingestion et transformation de données.
ROC/AUC	Courbe ROC et aire sous la courbe (qualité de classement).
Calibration	Ajustement des probabilités (Platt/Isotonic).
MLOps	Pratiques d'industrialisation du ML (versionnage, monitoring, CI/CD).