



# Projet de Fin d'Etudes

Pour obtenir le diplôme de

Ingénieur en INFORMATIQUE, RÉSEAUX ET MULTIMÉDIA

Spécialité : BIG DATA & BUSINESS INTELLIGENCE

Présenté et réalisé par : **Yasmine Derbel**

---

Suivi de la Mobilité Interne et Gestion des Talents  
avec Machine Learning

---

Président jury :

Mr/Mme

Encadrant académique :

Mr Hamza RAHMANI

Rapporteur :

Mr/Mme

Encadrant Industriel :

Mr Nabih ZINE EL ABIDINE





J'autorise l'étudiant à faire le dépôt de son rapport de stage en vue d'une soutenance.

Encadrant professionnel, **Monsieur Nabih ZINE EL ABIDINE**

**Signature et cachet**

J'autorise l'étudiant à faire le dépôt de son rapport de stage en vue d'une soutenance.

Encadrant académique, **Monsieur Hamza RAHMANI**

**Signature**

# Dédicaces

*Je dédie cet évènement marquant*

*À mes parents, Pour votre amour infini, vos prières silencieuses, vos sacrifices constants et votre soutien inébranlable à chaque étape de ma vie. Vous êtes la source de ma force, de ma motivation et de ma persévérance.*

*À mon frère, Pour tes encouragements sincères, ta présence rassurante et ta complicité fraternelle qui m'ont accompagné tout au long de ce parcours.*

*À mes amis, Pour vos mots réconfortants, votre écoute fidèle et les instants de légèreté partagés qui m'ont aidée à garder le cap. Merci d'avoir été là, à chaque moment important.*

*À mes grands-parents, Merci d'être dans ma vie. Vous êtes la trace de mon appartenance, l'essence de ma bénédiction et la source infaillible de tant d'amour.*

*À toute ma famille, Pour votre affection, votre bienveillance et votre foi inébranlable en moi. Vous êtes les piliers de mon équilibre.*

*À toutes les personnes formidables que j'ai eu la chance de côtoyer durant mon stage, Merci pour votre accueil chaleureux, votre accompagnement bienveillant, vos conseils enrichissants et votre professionnalisme inspirant. Vous avez grandement contribué à faire de cette expérience un souvenir précieux et formateur.*

Yasmine Derbel

# Remerciements

*J'exprime ma gratitude à toutes les personnes qui m'ont aidée à accomplir ma tâche dans de bonnes conditions et qui m'ont accordé toute l'attention nécessaire à l'élaboration du présent travail.*

*Je remercie en particulier Mr. Nabih ZINE EL ABIDINE de m'avoir donné l'opportunité d'intégrer son équipe et de m'y avoir accueillie avec confiance. Je remercie également Mr. Othmane EL GARES, dont l'aide m'a été précieuse. Mes remerciements s'adressent aussi à mon encadrant pédagogique, Mr. Hamza RAHMANI, pour son accompagnement et ses conseils tout au long de ce travail.*

*J'exprime mes vifs remerciements à toute l'équipe Data pour leur accueil chaleureux et leur soutien, tant sur le plan professionnel que personnel, ce qui m'a permis de m'intégrer rapidement.*

*Enfin, je remercie les membres du jury pour le temps qu'ils consacrent à l'évaluation de mon travail.*

# Table des matières

<b>Introduction générale</b>	<b>1</b>
<b>1 Présentation de l'organisme d'accueil et du cadre du projet</b>	<b>2</b>
1.1 Présentation de l'organisme d'accueil . . . . .	3
1.1.1 Talan Group . . . . .	3
1.1.2 Talan Tunisie . . . . .	3
1.1.3 Secteurs d'activités . . . . .	4
1.2 Problématique et contexte général . . . . .	4
1.2.1 Problématique . . . . .	4
1.2.2 Contexte général . . . . .	5
1.3 Etude de l'existant . . . . .	5
1.3.1 Workday . . . . .	5
1.3.2 Oracle HCM Cloud . . . . .	6
1.4 Solution proposée . . . . .	7
1.5 Méthodologie : CRISP-DM . . . . .	8
<b>2 Compréhension du métier</b>	<b>10</b>
2.1 Exploration des besoins métier . . . . .	11
2.1.1 Concepts métier . . . . .	11
2.1.2 Objectifs métier . . . . .	11
2.1.3 Objectifs du projet . . . . .	12
2.1.4 Identification des acteurs . . . . .	12
2.2 Architecture Logique du projet . . . . .	13
2.3 Benchmark des outils . . . . .	14
2.3.1 Les plateformes d'intégration de données . . . . .	14
2.3.2 Les plateformes de gestion de bases de données . . . . .	15
2.4 Architecture Physique . . . . .	17
<b>3 Compréhension des données</b>	<b>19</b>
3.1 Collecte de données . . . . .	20
3.1.1 Source de données . . . . .	20
3.1.2 Critères de choix . . . . .	20

---

3.1.3	Dataset choisi . . . . .	21
3.2	Augmentation et enrichissement de données . . . . .	21
3.2.1	outils et technologies . . . . .	21
3.2.2	Etapas d'enrichissement . . . . .	22
3.3	Description des données . . . . .	23
3.4	Exploration des données . . . . .	24
<b>4</b>	<b>Préparation des données</b>	<b>29</b>
4.1	Modélisation . . . . .	30
4.2	Démarche ETL . . . . .	31
4.2.1	Sources de données . . . . .	32
4.2.2	Couches de traitement . . . . .	32
4.3	Mise en place de l'environnement Snowflake . . . . .	32
4.3.1	Création des bases de données . . . . .	32
4.3.2	Création des schémas . . . . .	33
4.3.3	Création des tables . . . . .	33
4.4	Processus ETL . . . . .	34
4.4.1	Définition . . . . .	34
4.4.2	Informatica Cloud Secure Agent . . . . .	35
4.4.3	Création des connecteurs . . . . .	35
4.4.4	Alimentation de la couche ER . . . . .	37
4.4.5	Alimentation de la couche ES . . . . .	38
4.4.6	Alimentation de la couche ODS . . . . .	39
4.4.7	Alimentation de la couche EA . . . . .	40
<b>5</b>	<b>Modélisation</b>	<b>42</b>
5.1	Module de prédiction . . . . .	43
5.1.1	Définition du besoin . . . . .	43
5.1.2	Random Forest Classifier . . . . .	43
5.1.3	MLP Classifier . . . . .	45
5.1.4	XGBoost Classifier . . . . .	47
5.2	Génération des rapports . . . . .	49
5.2.1	Identification du besoin . . . . .	49
5.2.2	Identification des indicateurs clés de performance . . . . .	50
5.2.3	Technologies et bibliothèques utilisées . . . . .	50

---

5.3	Génération des stratégies de fidélisation . . . . .	51
5.3.1	Identification du besoin . . . . .	52
5.3.2	Rôle du modèle de langage . . . . .	52
5.3.3	Données fournies au LLM . . . . .	52
5.4	Agent conversationnel . . . . .	54
5.4.1	Définition de l'IA agentique . . . . .	54
5.4.2	Identification du besoin et objectifs . . . . .	54
5.4.3	Fonctionnement de l'agent . . . . .	54
5.4.4	Outils utilisés . . . . .	56
<b>6</b>	<b>Evaluation</b>	<b>58</b>
6.1	Evaluation des modèles prédictifs . . . . .	59
6.1.1	Métriques d'évaluation . . . . .	59
6.1.2	RandomForest . . . . .	61
6.1.3	MLP Classifier . . . . .	62
6.1.4	XGBoost . . . . .	64
6.1.5	Benchmark comparatif des modèles . . . . .	65
6.2	Evaluation des modèles génératifs . . . . .	67
6.2.1	Contexte et défis d'évaluation . . . . .	67
6.2.2	Qualité linguistique . . . . .	67
6.2.3	Utilisation des données contextuelles . . . . .	68
<b>7</b>	<b>Déploiement</b>	<b>70</b>
7.1	Présentation de l'environnement de developpement . . . . .	71
7.2	Réalisation . . . . .	71
7.2.1	Interface Rapport . . . . .	72
7.2.2	Interface de prédiction . . . . .	73
7.2.3	Interface des stratégies de fidélisation . . . . .	75
7.2.4	Interface du chatbot conversationnel . . . . .	76
	<b>Conclusion générale</b>	<b>78</b>



# Table des figures

1.1	Présence de TALAN dans le monde . . . . .	3
1.2	Logo TALAN . . . . .	4
1.3	Logo Workday . . . . .	6
1.4	Logo Oracle HCM Cloud . . . . .	6
1.5	CRISP DM . . . . .	9
2.1	Architecture logique du projet . . . . .	13
2.2	Logo Informatica . . . . .	15
2.3	Logo Snowflake . . . . .	17
2.4	Architecture physique du projet . . . . .	17
3.1	Répartition des employés par département . . . . .	25
3.2	distribution de l'âge des employés . . . . .	25
3.3	Répartition des employés par département . . . . .	26
3.4	Répartition de la satisfaction des employés . . . . .	27
3.5	Répartition de la satisfaction des employés . . . . .	27
3.6	Répartition de la satisfaction des employés . . . . .	28
4.1	Modèle en constellation . . . . .	31
4.2	Démarche ETL . . . . .	31
4.3	Bases de données Snowflake . . . . .	33
4.4	Interface de Informatica Cloud Secure Agent . . . . .	35
4.5	Connecteur fichier plat . . . . .	36
4.6	Connecteur Snowflake . . . . .	36
4.7	Configuration de la source . . . . .	37
4.8	Configuration de la destination . . . . .	37
4.9	Premier extrait de la table "ER_EMP" . . . . .	38
4.10	Deuxième extrait de la table "ER_EMP" . . . . .	38
4.11	Alimentation de la table "ES_emp" . . . . .	38
4.12	Premier extrait de la table "ES_EMP" . . . . .	39
4.13	Deuxième extrait de la table "ES_EMP" . . . . .	39
4.14	Alimentation de la table "ODS_sat" . . . . .	40

4.15	Alimentation de la table "fact_mouvement" . . . . .	40
4.16	Alimentation de la table "fact_satisfaction" . . . . .	41
5.1	Logo Python . . . . .	50
5.2	Logo ALTAIR . . . . .	51
5.3	Les variables les plus influentes . . . . .	53
5.4	Flux de traitement de l'agent conversationnel . . . . .	56
5.5	Logo de langchain . . . . .	56
5.6	Logo SQLite . . . . .	57
6.1	Matrice de confusion - Random Forest . . . . .	62
6.2	Matrice de confusion - MLP Classifier . . . . .	63
6.3	Matrice de confusion - XGBoost . . . . .	65
7.1	Logo Streamlit . . . . .	71
7.2	Logo SmartTrack . . . . .	71
7.3	Interface Rapport 1 . . . . .	72
7.4	Interface Rapport 2 . . . . .	73
7.5	Interface prédiction . . . . .	73
7.6	Résultat de la prédiction . . . . .	74
7.7	Filtre sur les prédictions . . . . .	74
7.8	Mise à jour du rapport . . . . .	75
7.9	Interface fidélisation . . . . .	75
7.10	Exemple de stratégie . . . . .	76
7.11	Interface chatbot . . . . .	76
7.12	Exemple d'exécution chatbot . . . . .	77

# Liste des tableaux

1.1	Comparaison entre les solutions traditionnelles et notre solution proposée . . . . .	8
2.1	Principaux acteurs du projet et leurs rôles . . . . .	13
2.2	Benchmark entre Informatica Cloud et Talend . . . . .	15
2.3	Comparaison entre SQL Server et Snowflake . . . . .	16
3.1	Description des colonnes du fichier <code>employee.csv</code> . . . . .	23
3.2	Description des colonnes du fichier <code>survey.csv</code> . . . . .	24
3.3	Description des colonnes du fichier <code>mouvement.csv</code> . . . . .	24
4.1	Organisation des bases et schémas dans Snowflake . . . . .	33
4.2	Répartition des tables par base de données et schéma . . . . .	34
4.3	Transformations appliquées à la table "ER_emp" . . . . .	39
5.1	Avantages et inconvénients de Random Forest . . . . .	44
5.2	Description des hyperparamètres de Random Forest . . . . .	45
5.3	Avantages et inconvénients du MLPClassifier . . . . .	46
5.4	Description des hyperparamètres du MLPClassifier . . . . .	47
5.5	Avantages et inconvénients de XGBoost . . . . .	48
5.6	Description des hyperparamètres de XGBoost . . . . .	49
5.7	Indicateurs clés de performance suivis dans l'analyse RH . . . . .	50
5.8	Comparaison entre Altair et Matplotlib . . . . .	51
6.1	Performances par classe pour le modèle Random Forest . . . . .	61
6.2	Performances par classe pour le modèle MLP Classifier . . . . .	63
6.3	Performances par classe pour le modèle XGBoost . . . . .	64
6.4	Benchmark comparatif des modèles de classification . . . . .	66
6.5	Évaluation de la qualité linguistique des stratégies générées par Mistral . . . . .	68
6.6	Évaluation de l'utilisation des données contextuelles par Mistral . . . . .	69

# Liste des abréviations

—	<b>AgenticAI</b>		<b>Agentic Artificial Intelligence</b>
—	<b>EA</b>	=	<b>Espace Applicatif</b>
—	<b>ER</b>	=	<b>Espace de Reception</b>
—	<b>ES</b>	=	<b>Espace de Simplification</b>
—	<b>ETL</b>	=	<b>Extract Transform Load</b>
—	<b>GenAI</b>	=	<b>Generative Artificial Intelligence</b>
—	<b>IA</b>	=	<b>Intelligence Artificielle</b>
—	<b>LLM</b>	=	<b>Large Language Model</b>
—	<b>ML</b>	=	<b>Machine Learning</b>
—	<b>ODS</b>	=	<b>Operational Data Store</b>
—	<b>SQL</b>	=	<b>Structured Query Language</b>

# Introduction générale

Dans un contexte économique en constante mutation, marqué par l'intensification de la concurrence, la transformation numérique et l'évolution des attentes des talents, les entreprises sont de plus en plus confrontées à des défis complexes en matière de gestion des ressources humaines. La capacité d'une organisation à attirer, fidéliser et faire évoluer ses collaborateurs est aujourd'hui considérée comme un levier stratégique majeur, tant pour sa pérennité que pour son agilité organisationnelle. Dans ce cadre, la donnée occupe une place centrale : l'exploitation intelligente des données RH et l'analyse prédictive deviennent des atouts majeurs pour mieux comprendre les comportements des employés, anticiper les risques et orienter les décisions stratégiques. Parmi les problématiques émergentes, la compréhension fine des dynamiques de mobilité interne et de l'attrition constitue un enjeu central pour les directions des ressources humaines.

C'est dans cette perspective que s'inscrit le présent rapport, qui propose une étude complète et approfondie des mouvements internes des employés et des phénomènes d'attrition, à partir de l'analyse de données issues des ressources humaines. L'objectif est double : d'une part, identifier les talents à fidéliser pour mieux cibler les actions de rétention ; d'autre part, développer des outils prédictifs capables d'anticiper les départs volontaires, afin de favoriser une gestion proactive et stratégique des ressources humaines.

Pour cela, ce travail adopte une approche data-driven, mobilisant des techniques d'intelligence artificielle et de machine learning appliquées à un jeu de données RH simulé. L'application développée s'articule autour de plusieurs modules complémentaires : une phase de prédiction, basée sur des modèles de classification, permet d'anticiper les mouvements des employés ; un générateur de stratégies de fidélisation, reposant sur des modèles de langage, propose des actions ciblées pour retenir les talents à risque ; un chatbot conversationnel facilite l'interaction et l'aide à la décision ; enfin, un dashboard qui permet de visualiser en temps réel quelques indicateurs clés, offrant ainsi un outil complet d'aide au pilotage stratégique des ressources humaines.

# PRÉSENTATION DE L'ORGANISME D'ACCUEIL ET DU CADRE DU PROJET

---

## Plan

1	Présentation de l'organisme d'accueil . . . . .	3
2	Problématique et contexte général . . . . .	4
3	Etude de l'existant . . . . .	5
4	Solution proposée . . . . .	7
5	Méthodologie : CRISP-DM . . . . .	8

## Introduction

Ce premier chapitre pose les fondations stratégiques et techniques de notre projet d'intelligence RH. Après avoir situé le contexte organisationnel de Talan Tunisie où a été effectué notre stage, nous analyserons les défis critiques de la gestion des talents à l'ère de l'IA, avant de positionner notre solution face aux limites des plateformes RH actuelles. Nous finirons par choisir la méthodologie de travail.

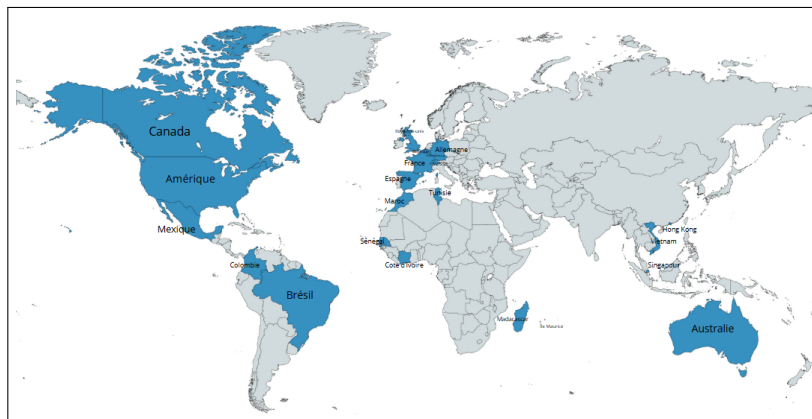
### 1.1 Présentation de l'organisme d'accueil

Dans cette partie, nous allons présenter l'organisme d'accueil où a été réalisé notre projet de fin d'études. Pour ce faire, nous allons commencer par introduire l'entreprise d'accueil et ses secteurs d'activité.

#### 1.1.1 Talan Group

Talan Group est un cabinet international de conseil en innovation et transformation numérique. Fondé en 2002, le groupe accompagne ses clients, issus de secteurs variés dans la réussite de leurs projets technologiques et stratégiques. Présent dans 23 pays à travers le monde, Talan affirme son ambition d'accompagner ses partenaires dans leur transformation à l'échelle internationale.

La figure 6.1 présente la répartition de TALAN à travers le monde.



**FIGURE 1.1 :** Présence de TALAN dans le monde

#### 1.1.2 Talan Tunisie

Talan Tunisie est la filiale du groupe Talan basée à Tunis, fondée en 2008. Elle joue un rôle clé dans le développement des activités du groupe en Afrique et en Europe, en combinant expertise locale et standards internationaux. Depuis sa création, Talan Tunisie s'est imposée comme un centre d'excellence dans les domaines du conseil en transformation digitale, de l'innovation technologique, de l'intégration de solutions et du développement applicatif. Forte d'une équipe de 750 collaborateurs, la

filiale contribue à de nombreux projets internationaux tout en participant activement au rayonnement du savoir-faire tunisien dans les métiers du numérique.

La figure ci-dessous représente le logo de Talan Tunisie.



FIGURE 1.2 : Logo TALAN

### 1.1.3 Secteurs d'activités

Le groupe Talan intervient dans un large éventail de secteurs d'activités, principalement orientés vers la transformation numérique et l'innovation technologique. Ses domaines d'intervention couvrent :

- **Finance et Banque** : accompagnement des institutions financières dans leur modernisation, leur conformité réglementaire, et la mise en place de solutions data/IA.
- **Énergie et Environnement** : optimisation des processus industriels, transition énergétique, gestion intelligente des réseaux.
- **Télécommunications** : accompagnement des opérateurs dans leur transformation réseau, cybersécurité, et gestion des données clients.

## 1.2 Problématique et contexte général

Dans cette partie, nous allons présenter la problématique et le cadre général du projet.

### 1.2.1 Problématique

Dans un contexte économique où le coût du turnover représente en moyenne 150% à 300% du salaire annuel d'un employé selon son niveau hiérarchique, et où 67% des entreprises identifient la rétention des talents comme leur défi RH prioritaire (étude Deloitte 2024), les organisations font face à une équation complexe : comment anticiper et prévenir les départs tout en optimisant la mobilité interne ?

Les approches traditionnelles de gestion RH présentent des limites structurelles critiques :

- **Réactivité insuffisante** : 78% des départs sont détectés moins de 3 mois avant la démission effective



- **Subjectivité des évaluations** : les décisions reposent sur l'intuition managériale plutôt que sur des données factuelles
- **Manque de personnalisation** : les stratégies de fidélisation sont génériques et peu adaptées aux profils individuels

Comment concevoir une architecture data-driven intégrant prédiction, génération et visualisation pour transformer les processus RH traditionnels en un système d'aide à la décision proactif et intelligent ?

### 1.2.2 Contexte général

L'économie mondiale connaît une transformation profonde du capital humain. Selon le World Economic Forum, d'ici 2027, 50% des compétences actuelles seront devenues obsolètes, accélérant la nécessité d'adaptation des organisations. Dans le même temps, la Génération Z, qui représentera 27% de la population active en 2025, redéfinit les attentes en matière de travail : quête de sens, équilibre vie professionnelle/personnelle, et une mobilité accrue, avec un changement d'emploi en moyenne tous les 2,3 ans. Cette instabilité croissante s'inscrit dans un contexte de guerre des talents où 73% des entreprises déclarent avoir des difficultés de recrutement, tandis que le coût d'acquisition d'un nouveau talent a bondi de 67% depuis 2020. Face à ces mutations générationnelles, à l'inertie technologique de certaines fonctions RH et aux opportunités offertes par l'intelligence artificielle, notre projet vise à opérer un véritable changement de paradigme : passer d'une gestion des talents réactive et descriptive à un pilotage proactif, prédictif et data-driven.

## 1.3 Etude de l'existant

Dans le but de positionner le projet proposé par rapport aux solutions actuellement disponibles sur le marché, une étude comparative a été menée sur deux plateformes RH majeures : Workday et Oracle HCM Cloud. Ces solutions figurent parmi les plus utilisées à l'échelle internationale pour la gestion des talents, l'analyse RH et l'optimisation des processus internes.

### 1.3.1 Workday

Workday est une solution cloud de gestion du capital humain (HCM), largement adoptée par les grandes entreprises. Elle propose une suite intégrée couvrant la gestion des ressources humaines, la paie, la planification du personnel, et l'analytique RH. Workday se distingue par son interface conviviale, sa flexibilité, et sa capacité à centraliser les données RH dans une plateforme unique.

L'image suivante présente le logo de Workday :



FIGURE 1.3 : Logo Workday

#### 1.3.1.1 Fonctionnalités pertinentes

- **Analyse des talents** : suivi des compétences, des performances et des potentiels de mobilité.
- **Mobilité interne** : suggestion d'opportunités de carrière en interne, accompagnement des plans de succession.
- **Workday People Analytics** : module d'analyse avancée permettant d'explorer les tendances RH (taux de départ, diversité, satisfaction).
- **Rapports personnalisables et tableaux de bord** : pour le suivi en temps réel des indicateurs RH clés.

#### 1.3.1.2 Limites de Workday

- Les modèles prédictifs sur les départs sont limités et peu personnalisables.
- Absence d'un moteur intelligent de recommandations personnalisées basé sur des LLMs.
- Pas d'assistant conversationnel avancé pour accompagner la prise de décision RH.
- L'intégration de la stratégie de fidélisation n'est pas automatisée ni générative.

### 1.3.2 Oracle HCM Cloud

Oracle HCM Cloud est une plateforme complète de gestion des ressources humaines proposée par Oracle. Elle couvre l'ensemble du cycle de vie des employés : recrutement, gestion de la performance, rémunération, planification des talents, etc. Elle repose sur l'IA et l'automatisation pour optimiser les processus RH.

L'image suivante présente le logo de Oracle HCM Cloud :



FIGURE 1.4 : Logo Oracle HCM Cloud

### 1.3.2.1 Fonctionnalités pertinentes

- **Oracle Talent Management** : suivi des compétences, mobilité interne, planification de la relève.
- **Oracle AI Apps for HCM** : modules utilisant l'intelligence artificielle pour détecter les risques de départ et suggérer des actions.
- **Tableaux de bord analytiques** : visualisation dynamique des indicateurs de performance RH.
- **Oracle Digital Assistant** : chatbot intégré pour les tâches simples (congrés, absences, suivi RH).

### 1.3.2.2 Limites de Oracle HCM Cloud

- Le chatbot reste limité à des tâches administratives et ne propose pas d'analyse stratégique ou de recommandations personnalisées.
- L'analyse prédictive existe, mais elle est orientée détection et non génération proactive de solutions (comme des stratégies de fidélisation).
- Les fonctionnalités d'IA sont souvent peu transparentes pour une adaptation fine aux besoins spécifiques.
- Absence d'intégration directe de modèles de langage génératifs dans l'aide à la décision RH.

## 1.4 Solution proposée

La solution proposée est une plateforme intégrée d'intelligence RH qui transforme les données en actions stratégiques grâce à une architecture modulaire combinant prédiction, génération et visualisation. Elle dispose d'une architecture innovante à quatre piliers :

- **Module prédictif avancé** : Contrairement aux solutions existantes qui se limitent à des analyses rétrospectives, notre solution implémente un système de classification multiclasse capable de distinguer trois trajectoires professionnelles (stabilité, mobilité interne, départ).
- **Générateur de stratégies personnalisées** : Là où les solutions actuelles proposent des recommandations génériques prédéfinies, notre système exploite le modèle Mistral pour générer des stratégies de fidélisation contextualisées et actionnables. Cette approche générative s'appuie sur trois niveaux d'information (prédiction, variables influentes, médianes comparatives) pour formuler des recommandations adaptées au profil spécifique de chaque employé à risque.
- **Tableau de bord analytique dynamique** : Au-delà des KPIs statiques, notre solution offre une visualisation interactive des tendances RH par département, permettant d'identifier les zones

de tension et d'anticiper les besoins en recrutement ou en mobilité. L'utilisation d'Altair garantit une expérience utilisateur fluide et intuitive, accessible aux décideurs non techniques.

- **Agent conversationnel RH** : Contrairement aux chatbots RH traditionnels limités aux questions administratives, notre agent conversationnel permet d'interroger directement les données prédictives en langage naturel, démocratisant ainsi l'accès à l'intelligence RH pour l'ensemble des managers opérationnels.

Afin de mieux mettre en évidence la valeur ajoutée de notre solution, le tableau ci-dessous compare ses principales caractéristiques avec celles des approches traditionnelles en gestion des talents.

Aspect	Solutions traditionnelles	Solution proposée
Prédiction	Binaire (reste/part)	Multiclasse (3 trajectoires)
Recommandations	Génériques, prédéfinies	Personnalisées, génératives
Intégration	Modules séparés	Plateforme unifiée

**TABLEAU 1.1** : Comparaison entre les solutions traditionnelles et notre solution proposée

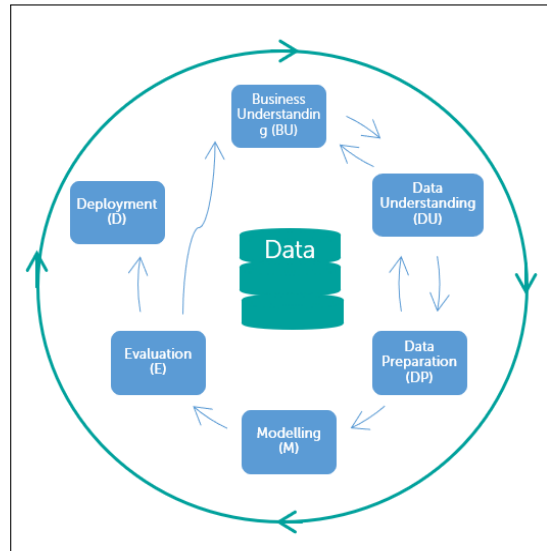
## 1.5 Méthodologie : CRISP-DM

Cross Industry Standard Process for Data Mining est un modèle de processus d'exploration de données qui décrit une approche communément utilisée pour résoudre les problèmes du domaine de l'analyse, de l'extraction et des sciences des données. Elle se compose de six étapes principales :

- **Compréhension du métier** : analyse des objectifs RH et des besoins en matière de mobilité interne.
- **Compréhension des données** : étude des sources de données disponibles (profils, postes, compétences, etc.).
- **Préparation des données** : nettoyage, transformation et sélection des données pertinentes pour l'analyse.
- **Modélisation** : application d'algorithmes de machine learning adaptés aux objectifs du projet.
- **Évaluation** : analyse des performances du modèle et vérification de sa pertinence métier.
- **Déploiement** : réflexion sur l'intégration du prototype dans un environnement RH opérationnel.

Cette démarche a permis d'assurer une progression logique, tout en maintenant un lien constant avec les enjeux métiers du projet.

La figure 1.5 présente les étapes de CRISP DM.



**FIGURE 1.5 : CRISP DM**

## Conclusion

Ce chapitre a posé les bases stratégiques du projet en mettant en lumière les limites des solutions RH actuelles, ce qui justifie pleinement la création de notre solution. En s'appuyant sur la méthodologie CRISP-DM, adaptée au contexte de Talan Tunisie, nous disposons d'un cadre structuré pour concrétiser cette vision. Le chapitre suivant approfondira l'analyse métier en précisant les besoins fonctionnels des RH et en définissant une architecture technique adaptée à nos quatre modules.

# COMPRÉHENSION DU MÉTIER

---

## Plan

1	Exploration des besoins métier . . . . .	11
2	Architecture Logique du projet . . . . .	13
3	Benchmark des outils . . . . .	14
4	Architecture Physique . . . . .	17

## Introduction

Ce chapitre est consacré à la compréhension du métier, étape essentielle pour identifier avec précision les besoins fonctionnels du projet. Une fois ces exigences clarifiées, nous définirons l'architecture logique qui structurera l'ensemble du système. Enfin, un benchmark des outils existants sera réalisé afin de sélectionner les technologies les plus adaptés aux contraintes et objectifs identifiés.

### 2.1 Exploration des besoins métier

Dans cette partie, nous allons définir quelques concepts clés puis fixer les objectifs du projet.

#### 2.1.1 Concepts métier

Avant d'aborder les aspects techniques de l'analyse, il est essentiel de définir les principaux concepts métiers qui structurent notre problématique. Ces notions permettent de mieux saisir les enjeux liés à la gestion des ressources humaines et à l'optimisation des performances organisationnelles.

- **Data RH** : Il s'agit de l'ensemble des données issues des ressources humaines (recrutement, évolution, départs, absences, évaluations, etc.). L'analyse de ces données permet de mieux comprendre les dynamiques internes et d'orienter les décisions stratégiques.
- **Rétention** : Ce terme désigne la capacité d'une organisation à fidéliser ses employés et à limiter les départs volontaires. Une bonne rétention est souvent le reflet d'un climat de travail sain, de perspectives d'évolution claires et d'une politique RH efficace.
- **Rotation** : Le taux de rotation du personnel mesure le nombre de départs et d'arrivées au sein de l'entreprise sur une période donnée. Un taux de turnover élevé peut indiquer des problèmes d'organisation, de management ou d'attractivité de l'environnement de travail.
- **Attrition** : L'attrition désigne la réduction naturelle des effectifs, généralement liée à des départs non remplacés (retraites, démissions, fins de contrat non renouvelées). Contrairement à la rotation, l'attrition entraîne une perte nette de ressources humaines. Un taux d'attrition élevé peut signaler un désengagement progressif des collaborateurs ou une politique de réduction des effectifs.

#### 2.1.2 Objectifs métier

Le projet a pour ambition de doter l'organisation d'un outil décisionnel puissant, conçu pour offrir aux décideurs une vision claire, factuelle et stratégique des dynamiques internes liées aux ressources humaines. Il vise à remplacer les décisions basées sur l'intuition par des décisions appuyées sur l'analyse

de données concrètes et des indicateurs fiables. Les objectifs principaux du projet sont les suivants :

- Collecter et structurer les données RH disponibles,
- Identifier et analyser les facteurs clés influençant la mobilité interne et les départs,
- Prédire les départs volontaires et anticiper les besoins de mobilité,
- Générer des rapports personnalisés synthétisant les causes des départs et les leviers de fidélisation,
- Visualiser les indicateurs RH essentiels à travers un tableau de bord interactif,
- Intégrer la solution de manière fluide dans les processus RH existants.

### **2.1.3 Objectifs du projet**

Le projet vise à doter l'organisation d'un outil décisionnel permettant une meilleure compréhension et anticipation des dynamiques internes liées aux ressources humaines. Les objectifs principaux se déclinent comme suit :

- Collecte des données RH
- Analyse des facteurs influençant la mobilité interne et les départs
- Prédiction des départs et des besoins en mobilité interne.
- Génération de rapports résumant les causes et les stratégies de fidélisation
- Visualisation des indicateurs RH nécessaires
- Intégration de la solution dans les processus RH.

### **2.1.4 Identification des acteurs**

Un acteur représente un rôle joué par une entité externe (utilisateur humain, dispositif matériel ou autre système) qui interagit directement avec le système étudié.

Ce tableau présente les acteurs identifiés :



Acteur	Rôle
<b>Responsables RH</b>	<ul style="list-style-type: none"><li>- Analyser les indicateurs de mobilité, satisfaction et attrition</li><li>- Identifier les profils à risque et suivre les dynamiques internes</li><li>- Mettre en œuvre les stratégies de fidélisation générées par l'outil</li><li>- Utiliser les rapports et dashboards pour orienter les politiques RH</li></ul>
<b>Managers opérationnels</b>	<ul style="list-style-type: none"><li>- Visualiser les indicateurs RH de leur périmètre (équipe, service)</li><li>- Anticiper les départs ou mobilités au sein de leur équipe</li><li>- Interagir avec le chatbot pour obtenir des analyses ciblées</li><li>- Soutenir l'application des actions de fidélisation recommandées</li></ul>

TABLEAU 2.1 : Principaux acteurs du projet et leurs rôles

## 2.2 Architecture Logique du projet

Cette section présente l'organisation fonctionnelle du système, en décrivant les flux de données, les modules d'analyse et les interactions entre les composants.

La figure ci-dessous illustre l'architecture logique du projet.

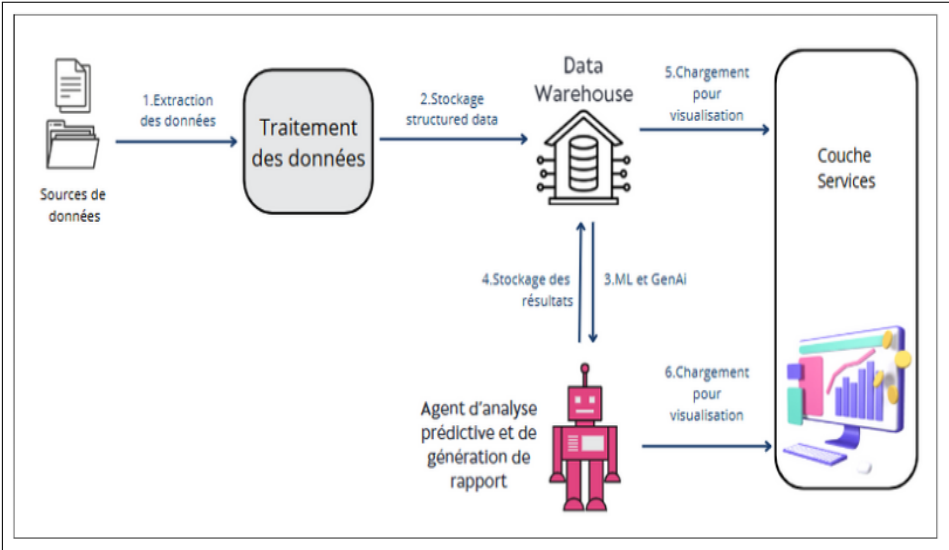


FIGURE 2.1 : Architecture logique du projet

Sur cette architecture logique, nous pouvons observer les différentes étapes du processus de traitement des données, de l'extraction à la visualisation.

- **Etape 1 : Extraction des données** : Le processus débute par l'extraction des données à partir de diverses sources de données internes ou externes à l'entreprise.
- **Etape 2 : Traitement des données** : Une phase de traitement est effectuée afin de nettoyer, normaliser et structurer les données collectées. Cette étape assure la qualité, la cohérence et l'harmonisation des informations avant leur exploitation analytique.
- **Etape 3 : Stockage des données structurées dans le Data Warehouse** : Les données traitées sont ensuite transférées dans un entrepôt de données (Data Warehouse). Cette infrastructure centralisée permet de stocker durablement les données structurées, facilitant ainsi leur accessibilité.
- **Etape 4 : Application des modèles de Machine Learning et de GenAI** : Un agent d'analyse prédictive et de génération de rapport, intégrant des algorithmes de machine learning et des techniques de génération de langage par intelligence artificielle (GenAI), est connecté au Data Warehouse. Il utilise les données disponibles pour effectuer des prédictions (attrition, mobilité interne) et formuler des recommandations RH personnalisées.
- **Etape 5 : Stockage des résultats analytiques** : Les résultats issus du module prédictif sont renvoyés vers le Data Warehouse, où ils sont stockés à part dans des tables spécifiques pour être ultérieurement exploités.
- **Etape 6 : Visualisation via la couche services** : Enfin, les données analysées ainsi que les rapports générés sont chargés dans la couche services, qui correspond à l'interface utilisateur. Celle-ci permet une visualisation interactive des résultats, destinée aux gestionnaires RH et décideurs, à travers des tableaux de bord dynamiques, graphiques explicatifs et exports automatisés.

## 2.3 Benchmark des outils

Afin d'identifier les solutions techniques les plus adaptées aux besoins du projet on va effectuer en premier lieu un benchmark des plateformes d'intégration de données et en second lieu des plateformes de gestion de bases de donnée.

### 2.3.1 Les plateformes d'intégration de données

Les plateformes d'intégration de données, comme Informatica ou Talend, sont des outils conçus pour faciliter le processus d'ETL. Leur rôle est de connecter différentes sources de données, de transformer ces données selon les besoins métiers (nettoyage, standardisation, enrichissement), puis de les charger dans un entrepôt ou un système cible. Ces plateformes sont essentielles pour garantir la qualité, la cohérence et la centralisation des données, permettant ainsi leur exploitation efficace dans des analyses, des tableaux de bord ou des modèles prédictifs.

Le tableau ci-dessous présente une comparaison entre les outils Informatica et Talend selon plusieurs critères clés, afin d'évaluer leur pertinence dans le cadre de notre projet.

Critère	Informatica Cloud	Talend
<b>Coût</b>	Licence commerciale à coût élevé.	Version gratuite (Open Studio), offre entreprise plus économique.
<b>Facilité d'utilisation</b>	Interface intuitive avec glisser-déposer.	Interface plus technique, moins accessible aux non-développeurs.
<b>Performance</b>	Très performante sur des volumes massifs, optimisée pour les processus ELT.	Bonne performance, dépend des réglages et de l'environnement.
<b>Connecteurs et Compatibilité</b>	Large catalogue de connecteurs : ERP, cloud, bases de données.	Nombreux connecteurs, certaines limitations en version gratuite.
<b>Apprentissage</b>	Facile pour les utilisateurs métiers, formation utile pour les cas complexes.	Prise en main plus technique, compétences en Java recommandées.
<b>Cloud / Big Data</b>	Intégration native avec les plateformes cloud, compatible Hadoop et Snowflake.	Très bon support Big Data (Spark, Hadoop) via Talend Data Fabric.

**TABLEAU 2.2** : Benchmark entre Informatica Cloud et Talend

En raison de sa robustesse, de la richesse de ses connecteurs natifs et de sa facilité d'intégration dans un environnement d'entreprise, Informatica a été retenu comme outil d'intégration de données pour ce projet.

La figure 2.2 présente le logo de Informatica Cloud



**FIGURE 2.2** : Logo Informatica

### 2.3.2 Les plateformes de gestion de bases de données

Les plateformes de gestion de bases de données (SGBD) permettent de stocker, organiser, interroger et sécuriser des données de manière structurée. Elles sont au cœur des systèmes d'information et jouent un rôle essentiel dans le bon fonctionnement des applications métiers. Ces plateformes peuvent être relationnelles (comme SQL Server, PostgreSQL ou MySQL) ou cloud natives (comme Snowflake,

BigQuery, Redshift), selon l'environnement et les besoins. Elles facilitent l'accès aux données pour les analystes, développeurs et outils BI, tout en garantissant la performance, la cohérence et la sécurité des opérations sur les données.

Le tableau suivant présente une comparaison entre SQL Server et Snowflake selon trois critères clés : le coût, la scalabilité et la performance, afin de déterminer la solution la plus adaptée aux besoins du projet.

Critère	SQL Server	Snowflake
<b>Coût</b>	<ul style="list-style-type: none"> <li>— Coût fixe lié aux serveurs et à la maintenance</li> <li>— Coût additionnel pour montée en charge ou haute disponibilité</li> </ul>	<ul style="list-style-type: none"> <li>— Paiement à l'usage</li> <li>— Aucun coût d'infrastructure initial</li> <li>— Économie pour les charges variables ou les usages intermittents</li> </ul>
<b>Scalabilité</b>	<ul style="list-style-type: none"> <li>— Scalabilité verticale</li> <li>— Limites physiques selon l'infrastructure</li> </ul>	<ul style="list-style-type: none"> <li>— Scalabilité horizontale automatique</li> <li>— Ajout ou retrait de clusters à la demande</li> <li>— Conçu pour une montée en charge illimitée</li> </ul>
<b>Performance</b>	<ul style="list-style-type: none"> <li>— Bonne performance sur de petits/moyens volumes si bien optimisé</li> </ul>	<ul style="list-style-type: none"> <li>— Haute performance native</li> <li>— Optimisation automatique des ressources selon la charge</li> </ul>

**TABLEAU 2.3 :** Comparaison entre SQL Server et Snowflake

Au vu de sa flexibilité, de ses performances en environnement cloud, et de son modèle de facturation à l'usage, le choix s'est porté sur Snowflake pour l'implémentation de l'entrepôt de données du projet.

La figure 2.3 présente le logo de Snowflake



FIGURE 2.3 : Logo Snowflake

## 2.4 Architecture Physique

Suite au benchmark réalisé dans la section précédente, les outils Snowflake pour l'entrepôt de données et Informatica pour l'intégration ont été retenus comme les plus adaptés aux besoins du projet. L'architecture physique présentée ci-après reflète ces choix technologiques en intégrant concrètement ces solutions dans l'infrastructure cible. Elle décrit la mise en œuvre réelle du système, incluant les composants matériels, logiciels, les flux de données, ainsi que les interactions entre les différentes couches du projet. Elle s'articule autour de plusieurs couches fonctionnelles interconnectées, comme l'illustre la figure ci-dessous :

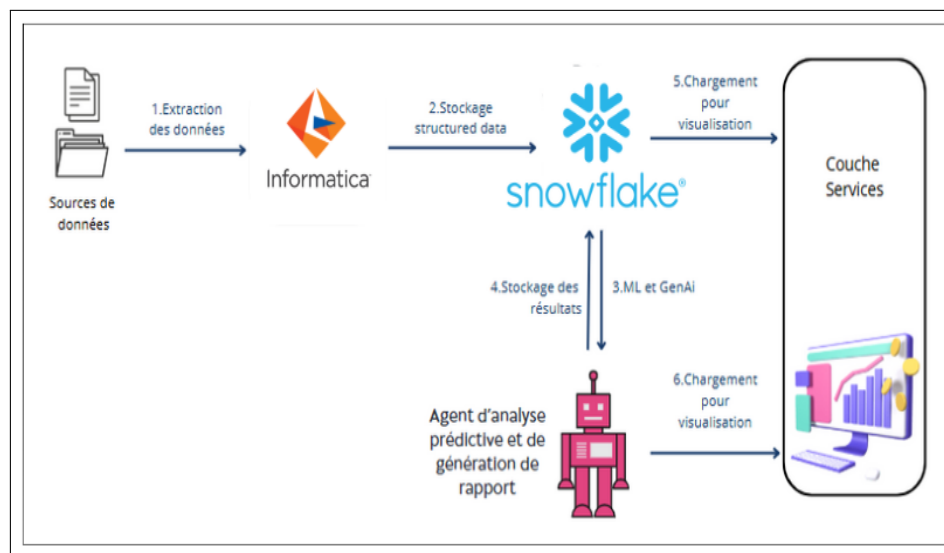


FIGURE 2.4 : Architecture physique du projet

- **Couches sources** : données RH extraites de systèmes internes (SIRH, fichiers Excel, outils collaboratifs) ou externes (open datasets, APIs sectorielles).
- **Couche d'intégration** : traitement des données brutes via des workflows Informatica. Cette couche permet le nettoyage, la normalisation, la transformation et l'enrichissement initial avant stockage.
- **Entrepôt de données (DWH)** : stockage structuré dans Snowflake, organisé en schéma en étoile ou flocon selon les besoins analytiques. Trois zones logiques y sont maintenues : zone de staging, zone d'intégration et zone de présentation.

- **Couche analytique** : déploiement des modèles de machine learning (XGBoost) et des modèles de génération (Mistral / LLM via API) pour la prédiction de la mobilité interne et la génération de recommandations RH.
- **Couche de restitution** : interface utilisateur (type Streamlit, Power BI ou application web) permettant la visualisation des données traitées sous forme de dashboards dynamiques, graphiques interactifs et rapports téléchargeables.

## Conclusion

Ce chapitre a permis de poser les fondations du projet en définissant précisément les besoins fonctionnels liés à la gestion RH, notamment la rétention et la mobilité interne. Après avoir modélisé l'architecture logique, un benchmark technologique a été mené pour comparer plusieurs solutions d'intégration et de stockage de données. L'architecture physique finale garantit une solution robuste, scalable et performante, en cohérence avec les exigences du projet.

# COMPRÉHENSION DES DONNÉES

---

## Plan

1	Collecte de données . . . . .	20
2	Augmentation et enrichissement de données . . . . .	21
3	Description des données . . . . .	23
4	Exploration des données . . . . .	24

## Introduction

Ce chapitre marque une étape clé de notre démarche : enrichir un dataset RH initialement limité afin de mieux refléter les dynamiques organisationnelles réelles. Face aux contraintes de notre jeu de données, nous avons adopté une approche innovante alliant intelligence artificielle générative et techniques d’augmentation de données. Après avoir défini notre stratégie de collecte et nos critères de sélection, nous présentons notre méthodologie d’enrichissement, suivie d’une exploration approfondie destinée à identifier les patterns et biais structurants pour la modélisation prédictive.

### 3.1 Collecte de données

La phase de collecte des données constitue un socle essentiel du projet, car la qualité des analyses et la fiabilité des prédictions reposent en grande partie sur la pertinence du jeu de données initial. Cette section présente la source de données sélectionnée, les critères ayant guidé ce choix, ainsi que les caractéristiques du dataset finalement retenu pour l’expérimentation.

#### 3.1.1 Source de données

Afin de disposer d’un jeu de données à la fois riche, structuré et accessible, la plateforme Kaggle a été retenue comme source principale. Réputée pour la diversité et la qualité de ses datasets dans le domaine de la science des données, Kaggle met à disposition des corpus spécialisés, souvent élaborés par des experts ou des institutions de référence. Elle constitue un environnement idéal pour les expérimentations liées à l’analyse prédictive et à l’apprentissage automatique.

#### 3.1.2 Critères de choix

Le choix de ce dataset repose sur des considérations à la fois méthodologiques et fonctionnelles :

- **Pertinence** : le dataset devait permettre d’étudier les causes de l’attrition et de la mobilité interne des employés, en lien avec les problématiques RH ciblées.
- **Structure réaliste** : les données devaient refléter des environnements organisationnels crédibles, avec des variables alignées sur celles utilisées dans les systèmes RH réels.
- **Variété des dimensions analysables** : un bon niveau de granularité était requis, intégrant à la fois des variables individuelles, professionnelles et comportementales.



### 3.1.3 Dataset choisi

À l'issue de cette démarche de sélection, le dataset "IBM HR Analytics Employee Attrition Performance" a été retenu. Ce jeu de données, bien que fictif, a été élaboré par des data analysts d'IBM afin de simuler des cas concrets de gestion des ressources humaines. Il comprend 1 470 enregistrements et 35 variables, couvrant une large gamme d'informations sur les employés : données démographiques, informations contractuelles, satisfaction au travail, performance, mobilité, etc. Sa structuration rigoureuse, combinée à sa pertinence thématique, en fait un support parfaitement adapté aux objectifs du projet : modéliser le risque de départ, identifier les facteurs explicatifs et générer des stratégies personnalisées de fidélisation.

## 3.2 Augmentation et enrichissement de données

Face aux limites structurelles du dataset initial — notamment l'absence d'informations sur les mouvements internes des employés et une taille d'échantillon restreinte — une démarche d'enrichissement des données a été engagée. Cette section présente d'une part les outils technologiques mobilisés, et d'autre part les étapes successives du processus mis en œuvre pour accroître la richesse et la robustesse analytique du corpus.

### 3.2.1 outils et technologies

L'enrichissement du jeu de données repose sur l'utilisation combinée de deux types de technologies issues de l'intelligence artificielle : la génération de texte par modèles de langage et la génération de données tabulaires par réseaux antagonistes génératifs conditionnels.

- **Générative AI (GenAI)** : il s'agit d'un ensemble de techniques d'intelligence artificielle visant à créer de nouvelles données (textes, images, tableaux) à partir d'exemples existants. Dans le cadre de ce projet, la GenAI est utilisée pour simuler des parcours professionnels internes et des raisons de départ plausibles à partir des profils initiaux.
- **Large Language Models (LLM)** : ces modèles de traitement du langage naturel, tels que LLaMA 3, sont capables de comprendre un contexte donné et de générer des textes cohérents en réponse à des prompts spécifiques. Leur capacité à simuler des comportements humains et organisationnels réalistes en fait un outil pertinent pour enrichir qualitativement les données RH.
- **Generative Adversarial Networks (GAN)** : les GANs sont des modèles de deep learning composés de deux réseaux (générateur et discriminateur) s'opposant pour produire des données

synthétiques réalistes. Ils sont à l'origine de nombreuses avancées dans la génération de données simulées.

- **CTGAN (Conditional Tabular GAN)** : variante spécialisée des GANs, CTGAN est conçu pour générer des données tabulaires incluant des variables catégorielles. Il permet de contrôler la distribution des données générées à travers des conditions explicites, rendant les échantillons synthétiques plus représentatifs et exploitables dans des contextes d'apprentissage supervisé avec déséquilibre des classes.

### 3.2.2 Etapes d'enrichissement

Le processus d'enrichissement a été réalisé en deux étapes complémentaires, articulant génération qualitative et expansion quantitative :

- **Étape 1 : Simulation de parcours internes et de raisons de départ**

Afin d'introduire des dimensions organisationnelles et comportementales absentes du jeu de données initial, des scénarios de mobilité interne (promotions, changements de département) ainsi que des motifs de départ (insatisfaction, conflits, stagnation) ont été simulés. Cette génération a été effectuée à l'aide du modèle LLaMA 3, exécuté localement via la plateforme Ollama, à partir de prompts calibrés sur les variables RH disponibles. Ce procédé a permis d'obtenir des données textuelles cohérentes, adaptées aux réalités du monde du travail, et contextualisées en fonction du profil de chaque employé.

- **Étape 2 : Génération de données synthétiques avec CTGAN**

À partir du dataset ainsi enrichi qualitativement, une seconde phase a consisté à augmenter le volume de données par génération synthétique. Le modèle CTGAN a été entraîné pour produire de nouveaux enregistrements d'employés présentant des profils crédibles et diversifiés. Cette technique permet d'augmenter la taille du corpus tout en respectant les distributions statistiques originales, notamment dans le cas de variables déséquilibrées comme l'attrition volontaire. Le dataset final, à la fois élargi et raffiné, a offert une base solide pour l'entraînement des modèles prédictifs.

En combinant une approche narrative contextualisée (via LLM) à une génération probabiliste contrôlée (via CTGAN), cette méthodologie offre à la fois une richesse descriptive et une profondeur statistique accrues. Elle permet de construire un corpus de données représentatif des réalités organisationnelles, renforçant ainsi la fiabilité des analyses prédictives tout en maintenant une forte applicabilité opérationnelle dans un environnement RH réel.

### 3.3 Description des données

À la suite du processus de collecte et d'enrichissement des données, nous avons pu constituer un ensemble structuré de trois fichiers CSV contenant des informations complémentaires sur les employés : `employee.csv`, `survey.csv` et `mouvement.csv`

➤ **`employee.csv`** : Ce fichier contient les caractéristiques sociodémographiques et professionnelles des employés. Il constitue la base principale sur laquelle sont croisées les autres sources de données.

longtable

**TABLEAU 3.1** : Description des colonnes du fichier `employee.csv`

Colonne	Description	Type
EmployeeNumber	Identifiant unique de l'employé	Numérique
Age	Âge de l'employé	Numérique
BusinessTravel	Fréquence des déplacements professionnels	Catégoriel
DailyRate	Taux journalier	Numérique
Department	Département d'affectation actuel	Catégoriel
DistanceFromHome	Distance domicile-travail (en km)	Numérique
Education	Niveau d'éducation (codé de 1 à 5)	Numérique
EducationField	Spécialité ou domaine d'étude	Catégoriel
Gender	Sexe (M/F)	Catégoriel
HourlyRate	Taux horaire	Numérique
JobLevel	Niveau hiérarchique (1 à 5)	Numérique
JobRole	Intitulé du poste occupé	Catégoriel
MaritalStatus	Statut marital	Catégoriel
MonthlyIncome	Salaire mensuel brut	Numérique
MonthlyRate	Taux mensuel	Numérique
NumCompaniesWorked	Nombre d'entreprises précédentes	Numérique
OverTime	Heures supplémentaires effectuées (Yes/No)	Catégoriel
PercentSalaryHike	Pourcentage d'augmentation de salaire	Numérique
StandardHours	Heures de travail standard (valeur constante)	Numérique
TotalWorkingYears	Nombre total d'années d'expérience	Numérique
TrainingTimesLastYear	Formations suivies l'an dernier	Numérique
YearsAtCompany	Ancienneté dans l'entreprise (années)	Numérique
YearsInCurrentRole	Ancienneté dans le poste actuel	Numérique

Colonne	Description	Type
YearsSinceLastPromotion	Années depuis la dernière promotion	Numérique
YearsWithCurrManager	Ancienneté avec le manager actuel	Numérique
YOB	Année de naissance	Numérique
AgeGroup	Catégorie d'âge (ex : 20–30, 30–40)	Catégoriel

➤ **survey.csv** : Ce fichier regroupe les informations issues d'un questionnaire RH renseigné par les employés, mesurant leur implication, satisfaction et perception de l'environnement de travail.

Colonne	Description	Type
EmployeeNumber	Identifiant de l'employé (jointure avec employee.csv)	Numérique
JobInvolvement	Implication dans le travail (score de 1 à 4)	Numérique
EnvironmentSatisfaction	Satisfaction vis-à-vis de l'environnement de travail	Numérique
JobSatisfaction	Satisfaction par rapport au poste occupé	Numérique
PerformanceRating	Évaluation de la performance (1 à 4)	Numérique
RelationshipSatisfaction	Qualité des relations professionnelles	Numérique
WorkLifeBalance	Équilibre vie professionnelle / personnelle	Numérique

**TABLEAU 3.2** : Description des colonnes du fichier **survey.csv**

➤ **mouvement.csv** : Ce fichier retrace les mouvements internes des employés dans l'organisation, notamment les mutations, promotions ou départs.

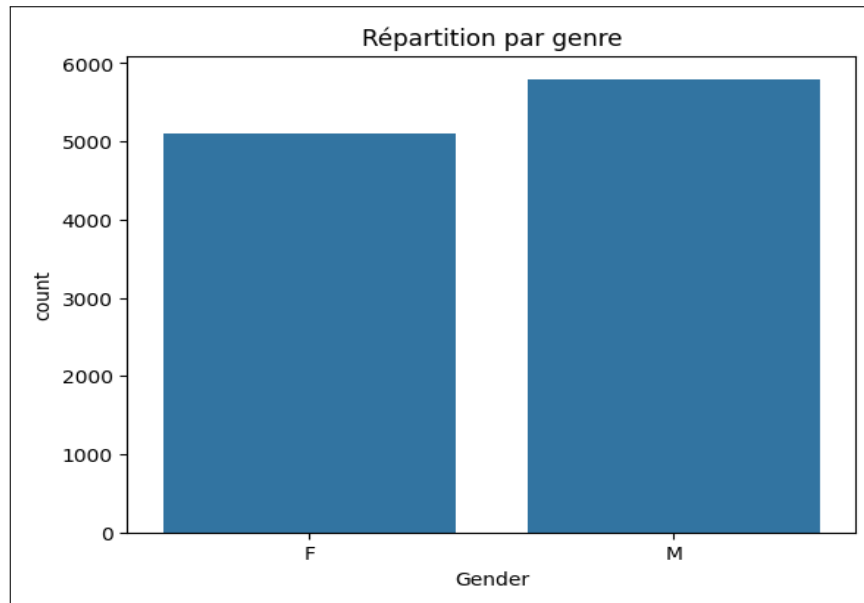
Colonne	Description	Type
TypeMouvement	Nature du mouvement (Entrée, Sortie, Mutation)	Catégoriel
Motif	Raison du mouvement (ex : démission, promotion)	Catégoriel
AncienDepartement	Département d'origine	Catégoriel
NouveauDepartement	Département de destination	Catégoriel
AnneeMouvement	Année du mouvement	Numérique
EmployeeNumber	Identifiant de l'employé concerné	Numérique

**TABLEAU 3.3** : Description des colonnes du fichier **mouvement.csv**

### 3.4 Exploration des données

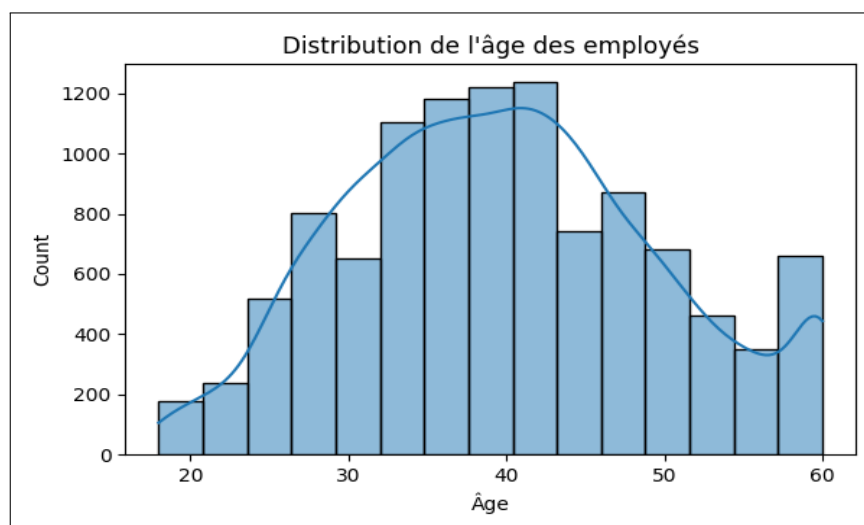
La répartition des employés selon le genre est relativement équilibrée, avec une légère majorité d'hommes. Ce constat témoigne d'un certain respect de la parité au sein de l'entreprise, ce qui est un

point positif en matière de diversité. Toutefois, cette analyse globale mériterait d'être complétée par une étude croisée, notamment pour évaluer la présence de biais dans la répartition par département, la rémunération moyenne, ou encore l'accès aux postes de responsabilité selon le genre. Une telle analyse permettrait d'identifier des inégalités structurelles invisibles dans les agrégats globaux. La figure ci-dessous représente cette répartition :



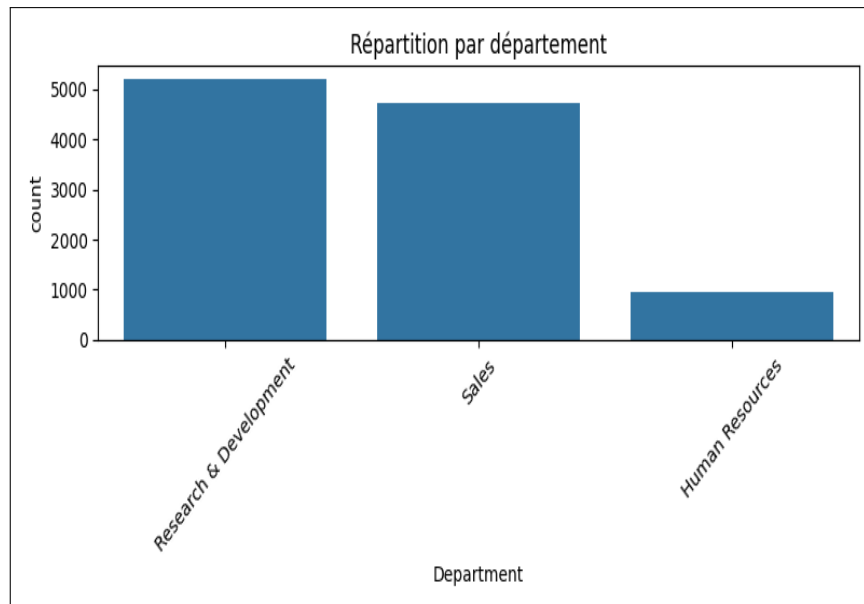
**FIGURE 3.1 :** Répartition des employés par département

L'histogramme ci-dessous montre une répartition centrée autour de 38 à 42 ans, ce qui indique une population majoritairement adulte et expérimentée. La distribution est assez équilibrée, légèrement asymétrique à droite, ce qui signifie qu'il y a un peu plus de jeunes que de seniors. Ce type de répartition est favorable à des analyses de mobilité ou de rétention, car elle inclut à la fois des profils juniors, en transition, et des profils seniors plus stabilisés.



**FIGURE 3.2 :** distribution de l'âge des employés

Le graphique ci-dessous montre une forte concentration des effectifs dans deux départements principaux : Research Development, avec un peu plus de 5 000 employés. Sales, avec environ 4 700 employés. En comparaison, le département Human Resources est très peu représenté. Cette distribution peut indiquer des priorités organisationnelles ou un déséquilibre à prendre en compte dans les analyses de satisfaction et de turnover : les départements surreprésentés auront plus de poids dans les tendances globales.



**FIGURE 3.3 :** Répartition des employés par département

L'analyse des salaires moyens par département révèle une forte disparité entre les différentes divisions de l'entreprise. Le département Sales se distingue avec un salaire mensuel moyen nettement supérieur aux autres, atteignant environ 1,3 million d'unités monétaires. En comparaison, les départements Research Development et Human Resources affichent des niveaux de rémunération beaucoup plus bas et relativement proches. Cette différence peut s'expliquer par des politiques de rémunération basées sur la performance commerciale, la présence de postes stratégiques, ou encore par une dynamique métier propre au domaine des ventes. Toutefois, un tel écart soulève des questions d'équité interne qui peuvent influencer la satisfaction et la fidélisation des talents.

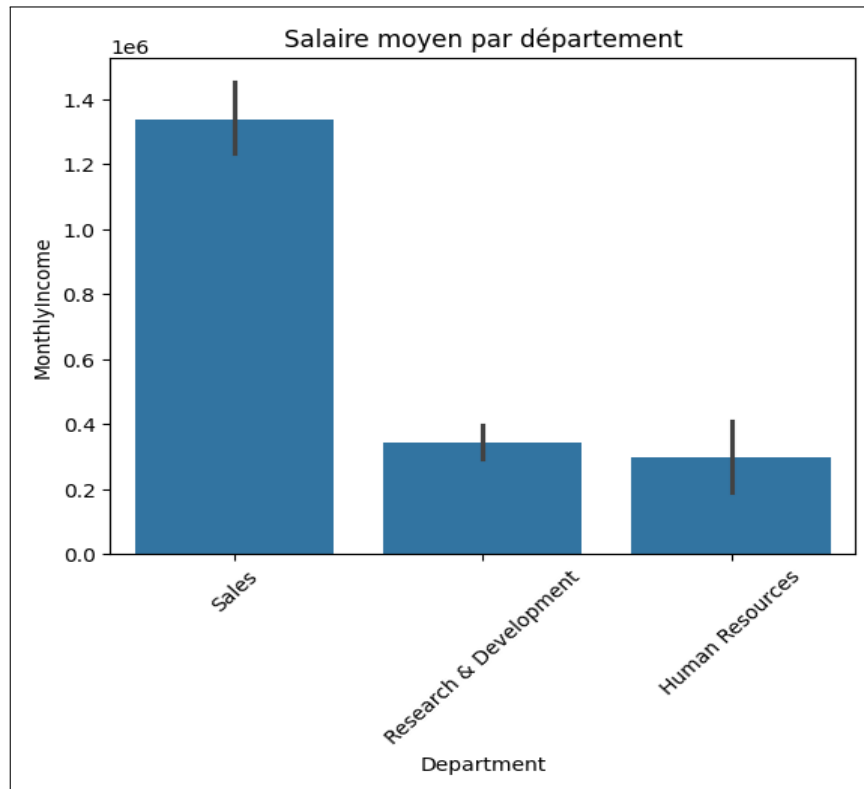


FIGURE 3.4 : Répartition de la satisfaction des employés

L'étude de la variable JobLevel met en évidence une structure hiérarchique pyramidale au sein de l'organisation. Une majorité des employés occupe des postes de niveau 1 et 2, indiquant une forte concentration dans les fonctions d'exécution ou de début de carrière. À l'opposé, les niveaux 4 et 5 sont très faiblement représentés, ce qui est cohérent avec la présence limitée de postes managériaux ou stratégiques dans une structure classique. Ce déséquilibre suggère que les opportunités de promotion ou d'évolution professionnelle sont peu nombreuses, ce qui peut générer une stagnation de carrière pour une large part des effectifs.

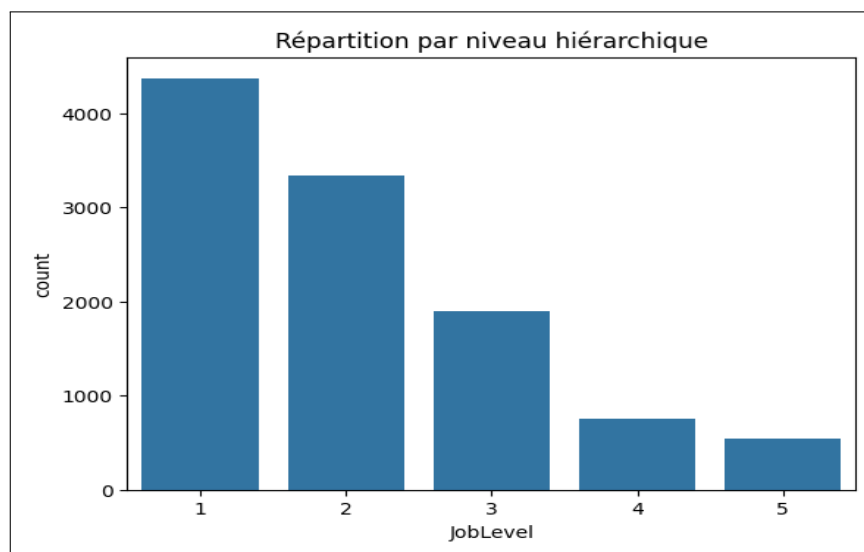


FIGURE 3.5 : Répartition de la satisfaction des employés

La figure ci-dessous montre que la majorité des employés se situent dans les niveaux de satisfaction 3 et 4, ce qui est positif globalement. Cependant, environ un tiers des effectifs ont une satisfaction inférieure ou égale à 2, ce qui représente un groupe à surveiller. Ces niveaux de satisfaction peuvent être croisés avec d'autres variables (départements, âge, ancienneté) pour identifier les poches de risque en matière de rétention.



**FIGURE 3.6 :** Répartition de la satisfaction des employés

## Conclusion

Ce chapitre a validé la faisabilité et la valeur ajoutée de notre stratégie d'enrichissement, combinant LLaMA 3 et CTGAN pour transformer un dataset limité en un corpus étendu et plus représentatif. L'analyse exploratoire a mis en évidence des déséquilibres structurels (poids de certains départements, écarts salariaux, corrélations satisfaction-département) essentiels pour orienter la modélisation. Le chapitre suivant se consacrera à leur transformation en un modèle dimensionnel optimisé, appuyé par une architecture ETL robuste.



---

# PRÉPARATION DES DONNÉES

---

## Plan

- 

1	Modélisation . . . . .	30
2	Démarche ETL . . . . .	31
3	Mise en place de l'environnement Snowflake . . . . .	32
4	Processus ETL . . . . .	34

## Introduction

Ce chapitre présente la manière dont les données ont été structurées et intégrées au sein de l'architecture du projet. Elle se compose de deux parties complémentaires : la modélisation des données, qui définit la structure logique et les relations entre les entités RH, et la mise en œuvre du processus ETL, chargé d'extraire, transformer et charger les données dans l'entrepôt. Ces étapes sont essentielles pour assurer la cohérence, la qualité et l'exploitabilité des données tout au long du cycle analytique.

### 4.1 Modélisation

Après la phase de collecte et d'enrichissement des données, nous avons conçu un **modèle en constellation** pour structurer les informations de manière efficace et analytique. Ce type de modélisation a été retenu en raison de la complexité des analyses RH à mener, nécessitant plusieurs axes d'observation croisés. Le modèle repose sur deux tables de faits : **fact\_mouvement** et **fact\_satisfaction**, qui permettent de répondre aux deux volets principaux du projet.

- La table **fact\_mouvement** centralise les informations relatives aux parcours internes des employés, tels que les changements de poste, de département ou de localisation. Elle intègre également des indicateurs comme le taux de rétention, la distance domicile-travail et les motifs de mouvement.
- La table **fact\_satisfaction** regroupe quant à elle les données liées à la satisfaction et à l'engagement des employés : satisfaction au travail, équilibre vie professionnelle/personnelle, implication, etc. Cette table est essentielle pour anticiper les départs et comprendre les leviers de fidélisation.

Les deux tables de faits partagent trois dimensions communes :

- **dim\_calendar** : pour assurer une analyse temporelle (par jour, mois, trimestre, etc.).
- **dim\_employee** : pour relier chaque événement aux caractéristiques personnelles et professionnelles de l'employé (âge, ancienneté, promotion, etc.).
- **dim\_department** : pour rattacher chaque donnée à une unité organisationnelle, ce qui permet une analyse par service ou entité.

La table **fact\_mouvement** est également liée à trois dimensions supplémentaires :

- **dim\_job** : pour identifier le poste concerné lors d'un changement (niveau hiérarchique, rôle).
- **dim\_education** : pour intégrer le niveau d'études et le domaine de formation, éléments pouvant influencer les parcours internes.
- **dim\_income** : pour suivre l'évolution des rémunérations (revenu mensuel, augmentation, taux de hausse), et évaluer leur rôle dans les décisions de mobilité ou de départ.

Ce modèle en constellation permet une analyse multidimensionnelle des comportements RH, tout en garantissant la souplesse et la performance des requêtes dans l'entrepôt de données.

L'image ci-dessous présente la modélisation

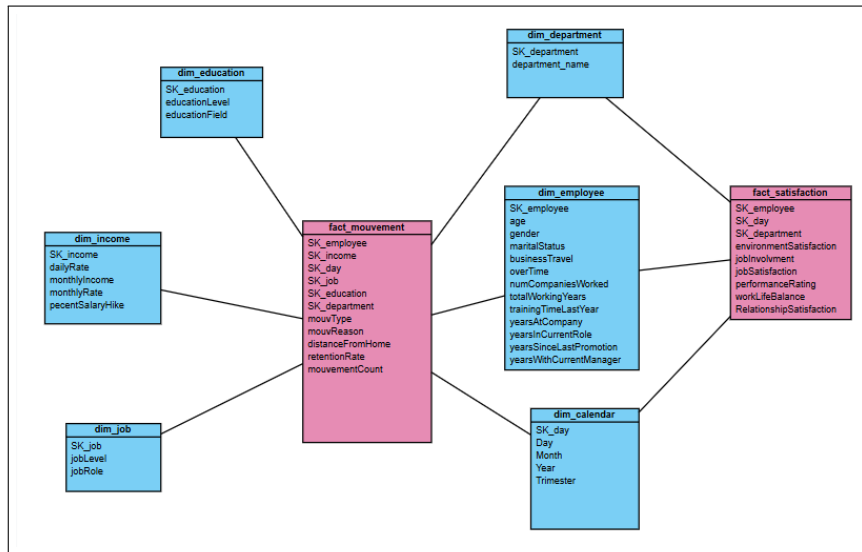


FIGURE 4.1 : Modèle en constellation

## 4.2 Démarche ETL

Le schéma ci-dessus décrit la démarche du processus d'intégration des données mis en place à l'aide de l'outil Informatica, avec stockage sur la plateforme cloud Snowflake. Le processus suit une approche classique en couches successives, chacune remplissant un rôle spécifique dans la chaîne de traitement.

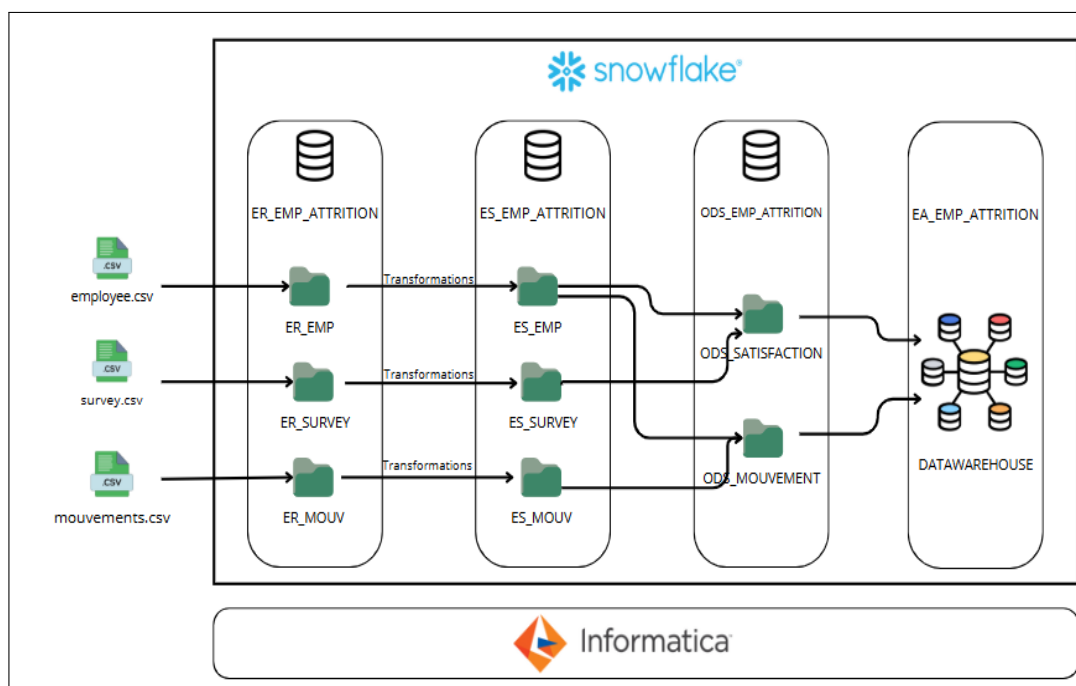


FIGURE 4.2 : Démarche ETL

### 4.2.1 Sources de données

Les données brutes sont issues de trois fichiers :

- `employee.csv` : informations personnelles et professionnelles des employés.
- `survey.csv` : résultats des enquêtes internes (satisfaction, engagement).
- `mouvements.csv` : historique des mobilités internes.

### 4.2.2 Couches de traitement

- **ER** : cette couche sert à réceptionner les données brutes dans leur état initial. Aucun traitement n'est effectué à ce stade. Elle assure la traçabilité complète des fichiers sources et permet de conserver un historique brut pour vérification ou audit.
- **ES** : les données sont ici nettoyées, typées, standardisées et validées. C'est une étape de pré-traitement essentielle où les valeurs manquantes sont gérées, les formats harmonisés, et les jeux de données structurés de manière homogène.
- **ODS** : cette couche regroupe les données issues de différentes sources après transformation. Les données sont intégrées par sujet fonctionnel, comme :
  - `ODS_SATISFACTION` : croisement des données employé et enquête.
  - `ODS_MOUVEMENT` : croisement des données employé et mobilité.
- **EA** : cette couche constitue l'environnement où les données, une fois nettoyées et intégrées, sont structurées pour répondre aux besoins analytiques et opérationnels du projet. Le Data Warehouse constitue le cœur de cette couche : il centralise les données consolidées, historisées et enrichies, prêtes à être interrogées.

## 4.3 Mise en place de l'environnement Snowflake

Avant de procéder à l'intégration des données, il est nécessaire de préparer l'environnement de stockage dans Snowflake. Cette étape consiste à créer les bases, les schémas et les tables nécessaires pour accueillir les données issues des sources.

### 4.3.1 Création des bases de données

Pour chaque couche de traitement, une base de données distincte a été créée dans Snowflake, comme le montre l'image ci-dessous. Cette structuration permet une meilleure organisation des données et facilite leur gestion tout au long du processus d'intégration.









NAME ↑	SOURCE	OWNER
 EA_EMP_ATTRITION	Local	 ACCOUNTADMIN
 ER_EMP_ATTRITION	Local	 ACCOUNTADMIN
 ES_EMP_ATTRITION	Local	 ACCOUNTADMIN
 ODS_EMP_ATTRITION	Local	 ACCOUNTADMIN

FIGURE 4.3 : Bases de données Snowflake

### 4.3.2 Création des schémas

Pour chaque base de données créée, un schéma a été défini afin d'organiser logiquement les objets (tables, vues, etc.) associés à cette couche. Les schémas permettent de structurer les données au sein d'une base de manière claire, facilitant ainsi leur gestion, leur sécurisation et la séparation des différentes étapes du traitement des données.

Dans ce projet, chaque couche fonctionnelle dispose de sa propre base de données, et un schéma unique y est défini pour organiser les objets de manière cohérente.

Le tableau ci-dessous illustre cette structuration, en présentant pour chaque couche de traitement la base de données correspondante ainsi que le schéma associé.

Couche de traitement	Base de données	Schéma associé
ER	ER_EMP_ATTRITION	ERschema
ES	ES_EMP_ATTRITION	ESschema
ODS	ODS_EMP_ATTRITION	ODSschema
EA	EA_EMP_ATTRITION	EAschema

TABLEAU 4.1 : Organisation des bases et schémas dans Snowflake

### 4.3.3 Création des tables

Après avoir mis en place les bases de données et les schémas nécessaires, l'étape suivante consiste à créer les tables qui accueilleront les données RH. Cette phase est essentielle pour structurer les informations de manière cohérente et adaptée aux traitements à venir.

Le tableau ci-dessous présente, pour chaque base de données et schéma, la liste des tables créées.

Base de données	Schéma	Tables
ER_EMP_ATTRITION	ERschema	ER_emp ER_survey ER_mouvement
ES_EMP_ATTRITION	ESschema	ES_emp ES_survey ES_mouvement
ODS_EMP_ATTRITION	ODSschema	ODS_satisfaction ODS_mouvement
EA_EMP_ATTRITION	EAschema	fact_mouvement fact_satisfaction dim_departement dim_employee dim_calendar dim_job dim_income dim_education

**TABLEAU 4.2** : Répartition des tables par base de données et schéma

## 4.4 Processus ETL

Tout au long de cette section nous allons d’abord commencer par définir le processus ETL puis passer à la réalisation.

### 4.4.1 Définition

**ETL** est un processus clé dans la gestion des données, qui consiste à :

- **Extraire** les données depuis diverses sources (fichiers, bases de données, systèmes internes...).
- **Transformer** ces données pour les nettoyer, les enrichir et les structurer selon les besoins du système cible.
- **Charger** les données transformées dans un entrepôt de données (Data Warehouse) ou une base de destination.

Ce processus permet de centraliser des données hétérogènes, de les fiabiliser, puis de les rendre exploitables pour l’analyse, la visualisation ou les modèles prédictifs.

### 4.4.2 Informatica Cloud Secure Agent

L'Informatica Cloud Secure Agent est un composant logiciel installé localement qui assure la liaison sécurisée entre les services cloud d'Informatica et les sources de données internes de l'entreprise. Il permet d'exécuter localement des tâches d'intégration, comme l'extraction, la transformation et le chargement (ETL) des données, tout en garantissant que les données sensibles restent protégées dans l'environnement sur site. Grâce à lui, les entreprises peuvent connecter leurs systèmes locaux et cloud de manière fluide, fiable et sécurisée, sans exposer directement leurs bases de données à Internet.

La capture ci-dessous montre l'interface de l'agent et que tous les services sont prêts :

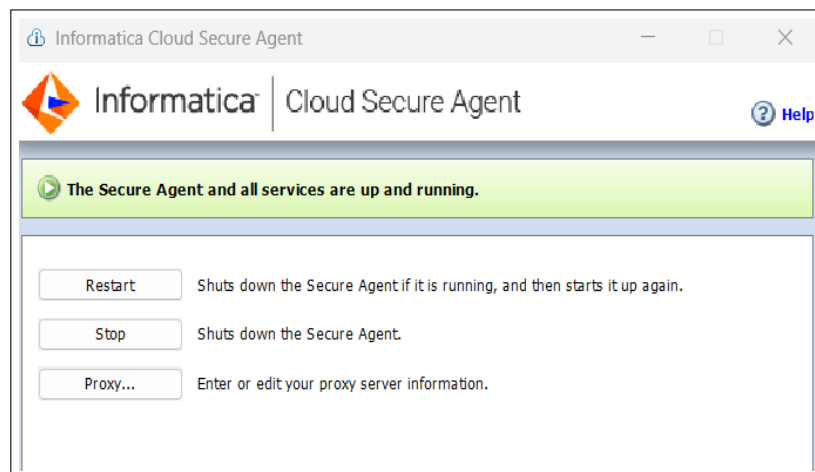


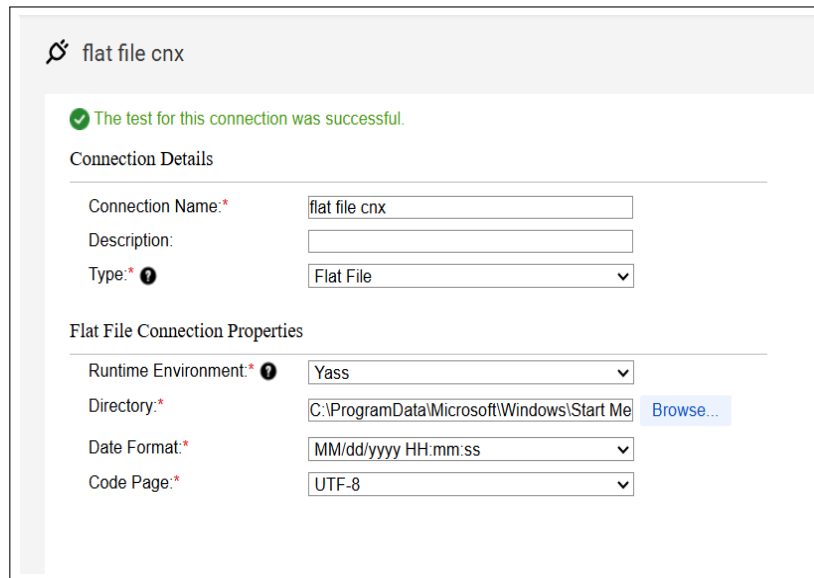
FIGURE 4.4 : Interface de Informatica Cloud Secure Agent

### 4.4.3 Création des connecteurs

Afin de connecter efficacement les sources de données aux destinations, il est nécessaire de créer au préalable les connecteurs appropriés. Dans notre cas, deux connecteurs seront requis : un premier pour les fichiers plats, qui permettra de lire les données depuis des fichiers locaux ou partagés, et un second pour Snowflake, afin de charger les données transformées dans l'entrepôt cloud cible.

➤ **Connecteur Fichier Plat** : Le connecteur "flat file cnx" est destiné à accéder à des fichiers plats (CSV, TXT, etc.) stockés localement. Il est configuré avec le répertoire source, le format de date utilisé dans les fichiers (MM/dd/yyyy HH:mm:ss) ainsi que le jeu de caractères (UTF-8). Ce connecteur joue un rôle fondamental dans la lecture des fichiers sources, souvent utilisés comme point de départ des flux d'intégration de données.

La capture suivante illustre la création du connecteur pour fichiers plats, spécifiant le chemin d'accès aux fichiers, le format de date et le jeu de caractères choisi.



flat file cnx

✓ The test for this connection was successful.

**Connection Details**

Connection Name:\* flat file cnx

Description:

Type:\* Flat File

**Flat File Connection Properties**

Runtime Environment:\* Yass

Directory:\* C:\ProgramData\Microsoft\Windows\Start Me [Browse...](#)

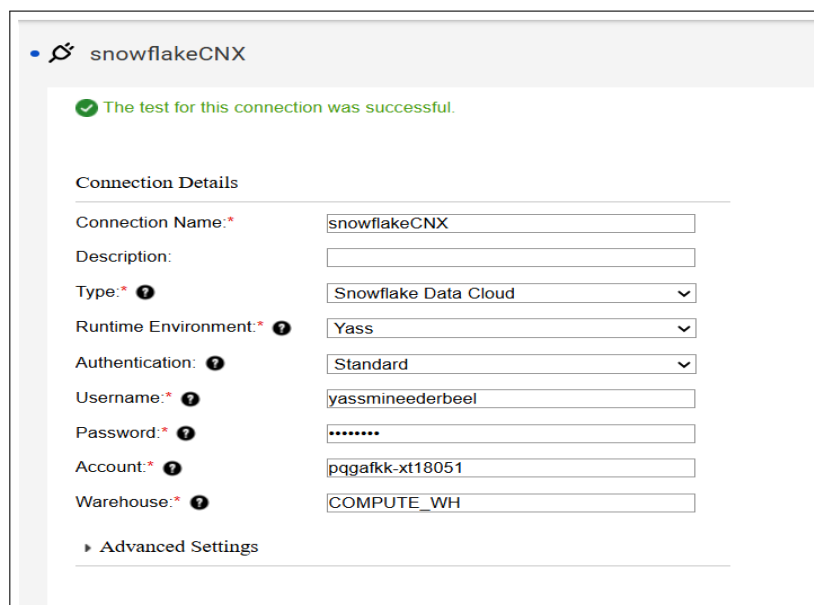
Date Format:\* MM/dd/yyyy HH:mm:ss

Code Page:\* UTF-8

FIGURE 4.5 : Connecteur fichier plat

➤ **Connecteur Snowflake** : Le connecteur SnowflakeCNX permet d'établir une liaison sécurisée entre Informatica Cloud et l'entrepôt de données cloud Snowflake. Il est configuré avec les informations d'authentification de l'utilisateur, le compte Snowflake, ainsi que le warehouse COMPUTE\_WH, qui servira à exécuter les requêtes et charges de données. Ce connecteur est essentiel pour charger les données transformées depuis les sources vers Snowflake dans le cadre du processus d'intégration.

L'image ci-dessous montre la configuration du connecteur Snowflake, incluant les paramètres d'authentification et le warehouse utilisé pour le chargement des données.



snowflakeCNX

✓ The test for this connection was successful.

**Connection Details**

Connection Name:\* snowflakeCNX

Description:

Type:\* Snowflake Data Cloud

Runtime Environment:\* Yass

Authentication:\* Standard

Username:\* yassmineederbeel

Password:\* .....

Account:\* pggafkk-xt18051

Warehouse:\* COMPUTE\_WH

► Advanced Settings

FIGURE 4.6 : Connecteur Snowflake



#### 4.4.4 Alimentation de la couche ER

Dans ce qui suit, nous allons illustrer le processus d'alimentation de la couche ER en prenant comme exemple le chargement de la table "ER\_EMP" à partir du fichier source "employee.csv".

L'image ci-dessous illustre la configuration de la source de données utilisée. Il s'agit d'un fichier local, connecté à l'aide du connecteur intitulé "flat file cnx".

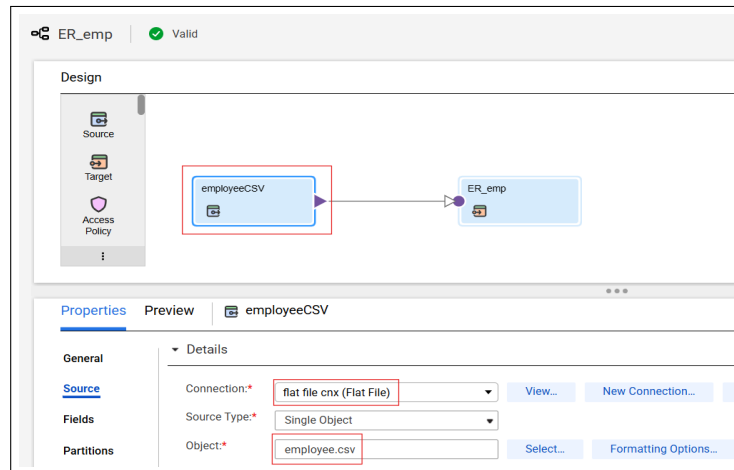


FIGURE 4.7 : Configuration de la source

La seconde image présente la cible du flux, qui correspond à une table brute dans Snowflake nommée "ER\_EMP", située dans le schéma "ERSchema". Cette table appartient à l'espace "ER\_EMP\_ATTRITION". Les données issues du fichier "employee.csv" y sont insérées directement sans transformation, via l'opération Insert. La connexion utilisée est "snowflakeCNX", précédemment définie lors de la configuration de l'environnement Snowflake.

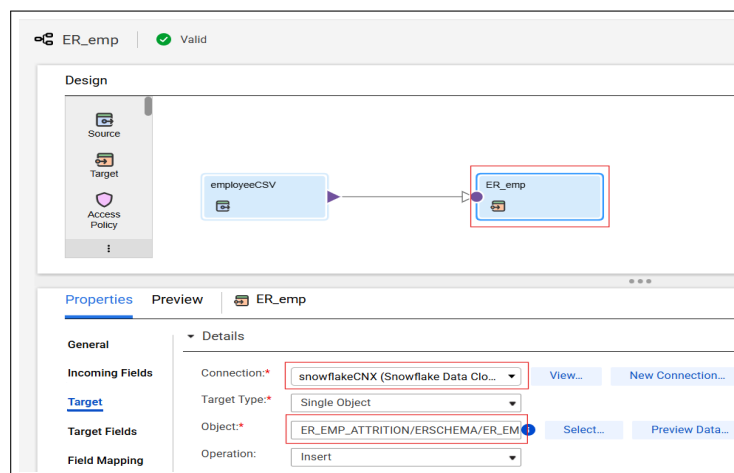


FIGURE 4.8 : Configuration de la destination

Ci-dessous quelques extraits de la table "ER\_EMP" de la couche ER dans Snowflake. Cette couche, rappelons-le, stocke les fichiers bruts sans transformations.

	BUSINESS TRAVEL	DAILY RATE	DEPARTMENT	DISTANCE FROM HOME	EDUCATION
1	Travel_Rarely	1102	Sales	1	2
2	Travel_Frequently	279	Research & Development	8	1
3	Travel_Rarely	1373	Research & Development	2	2
4	Travel_Frequently	1392	Research & Development	3	4
5	Travel_Rarely	591	Research & Development	2	1
6	Travel_Frequently	1005	Research & Development	2	2
7	Travel_Rarely	1324	Research & Development	3	3
8	Travel_Rarely	1358	Research & Development	24	1
9	Travel_Frequently	216	Research & Development	23	3
10	Travel_Rarely	1299	Research & Development	27	3

FIGURE 4.9 : Premier extrait de la table "ER\_EMP"

	EDUCATION FIELD	GENDER	HOURLY RATE	JOB LEVEL	JOB ROLE	MARITAL STATUS
1	Life Sciences	Female	94	2	Sales Executive	Single
2	Life Sciences	M	61	2	Research Scientist	Married
3	Other	M	92	1	Laboratory Technician	Single
4	Life Sciences	F	56	1	Research Scientist	Married
5	Medical	Male	40	1	Laboratory Technician	Married
6	Life Sciences	Mal	79	1	Laboratory Technician	Single
7	Medical	Female	81	1	Laboratory Technician	Married
8	Life Sciences	Male	67	1	Laboratory Technician	Divorced
9	Life Sciences	Male	44	3	Manufacturing Director	Single
10	Medical	Mal	94	2	Healthcare Representative	Married

FIGURE 4.10 : Deuxième extrait de la table "ER\_EMP"

#### 4.4.5 Alimentation de la couche ES

Dans ce qui suit, nous allons illustrer le processus d'alimentation de la couche ES en prenant comme exemple le chargement de la table "ES\_EMP" à partir de la table "ER\_EMP".

La figure ci-dessous illustre le mapping mis en place pour charger les données depuis la table "ER\_EMP", appliquer les transformations nécessaires à l'aide du composant "Expression", puis charger les données transformées dans la table "ES\_EMP".

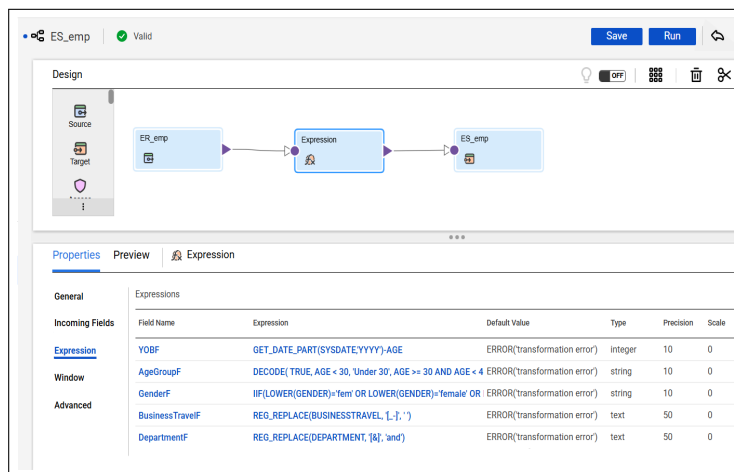


FIGURE 4.11 : Alimentation de la table "ES\_emp"

Le tableau ci-dessous résume les transformations effectuées :

Colonne	Transformation appliquée
<b>YOB</b>	Calcul de l'année de naissance en fonction de l'âge de l'employé
<b>AgeGroup</b>	Regroupement des individus par tranches d'âge (ex. : Under 30, 30-40, etc.)
<b>Gender</b>	Uniformisation des différentes écritures du genre (ex. : fem, female, F → Female)
<b>BusinessTravel</b>	Nettoyage des caractères spéciaux dans les intitulés de déplacements professionnels
<b>Department</b>	Standardisation des noms de départements en supprimant/remplaçant les caractères spéciaux

**TABLEAU 4.3** : Transformations appliquées à la table "ER\_emp"

Ci-dessous quelques extraits de la table "ES\_EMP" de la couche ES dans Snowflake. Cette couche, rappelons-le, stocke les données nettoyées, standardisées et enrichies.

	YOB	AGEGROUP	BUSINESSTRAVEL	DAILYRATE	DEPARTMENT
1	1984	40-50	Travel Rarely	1102	Sales
2	1976	40-50	Travel Frequently	279	Research and Development
3	1988	30-40	Travel Rarely	1373	Research and Development
4	1992	30-40	Travel Frequently	1392	Research and Development
5	1998	Under 30	Travel Rarely	591	Research and Development
6	1993	30-40	Travel Frequently	1005	Research and Development
7	1966	50+	Travel Rarely	1324	Research and Development
8	1995	30-40	Travel Rarely	1358	Research and Development
9	1987	30-40	Travel Frequently	216	Research and Development
10	1989	30-40	Travel Rarely	1299	Research and Development

**FIGURE 4.12** : Premier extrait de la table "ES\_EMP"

	EDUCATIONFIELD	GENDER	HOURLYRATE	JOBLEVEL	JOBROLE	MARITALSTATUS
1	Life Sciences	F	94	2	Sales Executive	Single
2	Life Sciences	M	61	2	Research Scientist	Married
3	Other	M	92	1	Laboratory Technician	Single
4	Life Sciences	F	56	1	Research Scientist	Married
5	Medical	M	40	1	Laboratory Technician	Married
6	Life Sciences	M	79	1	Laboratory Technician	Single
7	Medical	F	81	1	Laboratory Technician	Married
8	Life Sciences	M	67	1	Laboratory Technician	Divorced
9	Life Sciences	M	44	3	Manufacturing Director	Single
10	Medical	M	94	2	Healthcare Representative	Married

**FIGURE 4.13** : Deuxième extrait de la table "ES\_EMP"

#### 4.4.6 Alimentation de la couche ODS

Le mapping présenté ci-dessous permet d'alimenter la table "ODS\_sat" à partir de deux sources standardisées : "ES\_emp" et "ES\_survey". Afin d'éviter les conflits de noms lors de la jointure, un composant "Expression" est appliqué à la table "ES\_survey" pour renommer les champs en doublon avec "ES\_emp". Une fois cette étape de préparation effectuée, un composant "Joiner" est utilisé pour

effectuer la jointure entre les deux sources sur un champ commun qui est le numéro de l'employé. Le résultat de cette jointure est ensuite chargé dans la table cible "ODS\_sat" intégrant ainsi les informations RH et les retours d'enquête dans une vision unifiée des données.

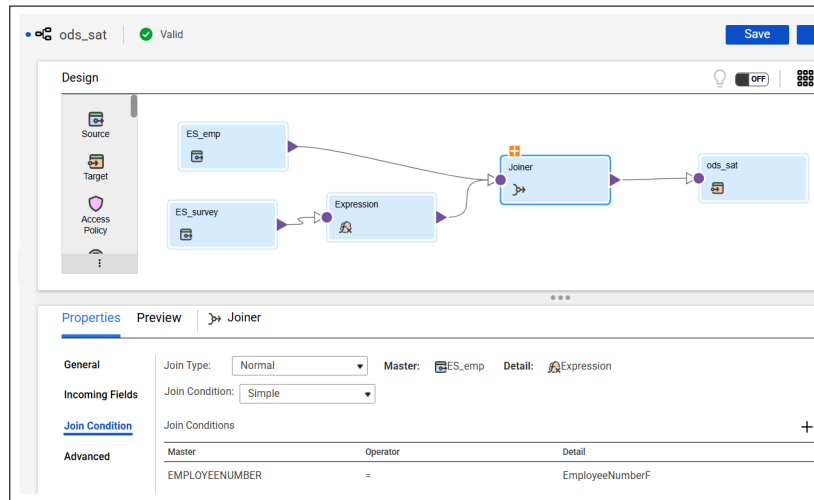


FIGURE 4.14 : Alimentation de la table "ODS\_sat"

#### 4.4.7 Alimentation de la couche EA

La dernière étape du processus ETL consiste à alimenter la couche EA, qui correspond à notre entrepôt de données. Contrairement aux couches précédentes, cette phase a été réalisée de manière directe via des requêtes SQL exécutées dans Snowflake, sans utiliser de flux ETL automatisé.

Ce choix a été motivé par le besoin de contrôle fin sur les jointures et la gestion des clés substitutives (SK) nécessaires à la modélisation. Pour chaque table de dimension et table de faits, des instructions requêtes ont été soigneusement construites, en s'appuyant sur les données sources de la couche ODS.

L'exemple ci-dessous illustre l'alimentation de la table "fact\_mouvement" :

```
INSERT INTO ea_emp_attrition.easchema.fact_mouvement (
  SK_employee, SK_income, SK_day, SK_job, SK_education, SK_department,
  movvType, movvReason, distanceFromHome
)
SELECT
  e.SK_employee,
  i.SK_income,
  c.SK_day,
  j.SK_job,
  ed.SK_education,
  d.SK_department,
  m.TypeMouvement,
  m.Motif,
  m.distanceFromHome
FROM
  ods_emp_attrition.odsschema.ods_mouv m
JOIN ea_emp_attrition.easchema.dim_employee e ON e.EmployeeNumber = m.EmployeeNumber
JOIN ea_emp_attrition.easchema.dim_income i ON i.DAILYRATE = m.DAILYRATE AND i.HOURLYRATE =
m.HOURLYRATE AND i.MONTHLYINCOME = m.MONTHLYINCOME AND i.MONTHLYRATE = m.MONTHLYRATE
JOIN ea_emp_attrition.easchema.dim_job j ON j.JOBROLE = m.JOBROLE AND j.JOBLEVEL = m.JOBLEVEL
JOIN ea_emp_attrition.easchema.dim_education ed ON ed.EDUCATIONFIELD = m.EDUCATIONFIELD AND
ed.EDUCATION = m.EDUCATION
JOIN ea_emp_attrition.easchema.dim_department d ON d.DEPARTMENT_NAME = m.DEPARTMENT
JOIN ea_emp_attrition.easchema.dim_calendar c ON c.year = m.ANNEEMOUEMENT;
```

FIGURE 4.15 : Alimentation de la table "fact\_mouvement"

L'exemple ci-dessous illustre l'alimentation de la table "fact\_satisfaction" :

```
INSERT INTO ea_emp_attrition.easchema.fact_satisfaction (
    SK_employee, SK_department, SK_day, environmentSatisfaction,
    jobSatisfaction, performanceRating, workLifeBalance, relationshipSatisfaction, jobInvolvement
)
SELECT
    e.SK_employee,
    d.SK_department,
    c.SK_day,
    m.environmentSatisfaction,
    m.jobSatisfaction,
    m.performanceRating,
    m.workLifeBalance,
    m.relationshipSatisfaction,
    m.jobInvolvement
FROM
    ods_emp_attrition.odsschema.ods_sat m
JOIN ea_emp_attrition.easchema.dim_employee e ON e.EmployeeNumber = m.EMPLOYEEENUMBER
JOIN ea_emp_attrition.easchema.dim_department d ON d.department_name = m.DEPARTMENT
JOIN ea_emp_attrition.easchema.dim_calendar c ON c.year = m.yob;
```

**FIGURE 4.16 :** Alimentation de la table "fact\_satisfaction"

## Conclusion

Ce chapitre a détaillé les différentes étapes de traitement et de structuration des données, depuis leur collecte initiale jusqu'à leur intégration dans l'entrepôt. La modélisation en constellation, avec deux tables de faits distinctes, offre une vision analytique fine des mouvements internes et de la satisfaction des employés. Enfin, l'intégration des données via un processus ETL structuré assure la cohérence, la qualité et la disponibilité des informations pour les analyses futures.

# MODÉLISATION

---

## Plan

1	Module de prédiction . . . . .	43
2	Génération des rapports . . . . .	49
3	Génération des stratégies de fidélisation . . . . .	51
4	Agent conversationnel . . . . .	54

## Introduction

La phase de modélisation constitue le cœur technique de notre projet. Cette étape transforme les données préparées lors de l'étape précédente en modèles prédictifs opérationnels capables de répondre aux objectifs métier définis.

### 5.1 Module de prédiction

Dans cette section, nous allons commencer par identifier le problème puis passer à la présentation des trois modèles testés lors de la phase de prédiction.

#### 5.1.1 Définition du besoin

Dans ce projet, la prédiction vise à estimer la trajectoire probable d'un employé à partir de ses caractéristiques personnelles, professionnelles et historiques. Il s'agit d'un problème de classification multiclasse, où chaque individu est assigné à l'une des trois classes suivantes :

- **Stabilité** : l'employé est susceptible de rester à son poste actuel.
- **Sortie** : l'employé présente un risque de départ de l'entreprise.
- **Mouvement interne** : l'employé pourrait changer de poste ou de service en interne.

La classification multiclasse est un type de tâche supervisée où le modèle apprend à prédire une catégorie parmi plusieurs (plus de deux). Contrairement à la classification binaire, elle nécessite une gestion spécifique des déséquilibres et des métriques adaptées.

#### 5.1.2 Random Forest Classifier

Random Forest est un algorithme d'ensemble basé sur la méthode du bagging, utilisé pour la classification et la régression. Il repose sur la construction de plusieurs arbres de décision indépendants, dont les prédictions sont agrégées pour améliorer la robustesse et la précision du modèle global.

##### 5.1.2.1 Fonctionnement

Random Forest génère plusieurs sous-échantillons du jeu de données d'entraînement par échantillonnage avec remise. Pour chaque sous-échantillon, un arbre de décision est construit. Lors de la prédiction, les sorties des arbres sont combinées (vote majoritaire pour la classification, moyenne pour la régression). À chaque nœud de chaque arbre, un sous-ensemble aléatoire de variables est sélectionné, ce qui introduit de la diversité entre les arbres et réduit le risque de surapprentissage.

### 5.1.2.2 Quand l'utiliser ?

- Lorsque l'on recherche un modèle robuste face au surapprentissage.
- Pour des jeux de données avec de nombreuses variables et peu de traitement préalable.
- Pour des problèmes de classification multiclasse ou de régression.
- Lorsqu'une interprétation partielle via l'importance des variables est suffisante.

### 5.1.2.3 Avantages et inconvénients

Le tableau suivant résume les avantages et les inconvénients du modèle Random Forest Classifier :

Avantages	Inconvénients
<ul style="list-style-type: none"><li>• Très bonne robustesse face au surapprentissage.</li><li>• Nécessite peu de réglage pour de bonnes performances.</li><li>• Fonctionne bien avec des données bruitées ou incomplètes.</li><li>• Peut traiter naturellement les données catégorielles (avec encodage).</li><li>• Fournit une estimation de l'importance des variables.</li></ul>	<ul style="list-style-type: none"><li>• Moins performant que les méthodes de boosting sur certains jeux de données.</li><li>• Moins interprétable qu'un arbre de décision simple.</li><li>• Peut être coûteux en mémoire et en temps avec beaucoup d'arbres.</li><li>• Tendance à créer des modèles plus lents en prédiction.</li></ul>

**TABLEAU 5.1** : Avantages et inconvénients de Random Forest

### 5.1.2.4 Paramètres

Le tableau suivant décrit les hyperparamètres du modèle Random Forest :



Hyperparamètre	Description
<code>n_estimators</code>	Nombre d'arbres dans la forêt.
<code>max_depth</code>	Profondeur maximale de chaque arbre (pour éviter une croissance excessive).
<code>min_samples_split</code>	Nombre minimum d'échantillons requis pour diviser un nœud interne.
<code>min_samples_leaf</code>	Nombre minimal d'échantillons requis dans une feuille.
<code>max_features</code>	Nombre de variables à considérer pour chaque division.
<code>bootstrap</code>	Active ou non l'échantillonnage avec remise (par défaut <code>True</code> ).
<code>class_weight</code>	Poids associés aux classes (utile pour données déséquilibrées).
<code>random_state</code>	Graine aléatoire pour la reproductibilité.

**TABLEAU 5.2** : Description des hyperparamètres de Random Forest

### 5.1.3 MLP Classifier

Le `MLPClassifier` est un réseau de neurones artificiel de type dense, appartenant à la famille des modèles supervisés. Il s'agit d'un modèle non linéaire capable de résoudre des problèmes de classification multiclasse en apprenant des représentations complexes.

#### 5.1.3.1 Fonctionnement

Le MLP est composé de plusieurs couches :

- **Une couche d'entrée**, qui reçoit les caractéristiques du jeu de données,
- **Une ou plusieurs couches cachées**, où s'opèrent des transformations via des neurones artificiels,
- **Une couche de sortie**, qui donne la prédiction.

#### 5.1.3.2 Quand l'utiliser ?

- Lorsque les relations entre les variables sont complexes et non linéaires.
- Dans les contextes où les autres modèles (linéaires, arbres, etc.) atteignent leurs limites de performance.

- Pour des données bruitées ou de grande dimension, où les interactions sont subtiles.
- Dans des cas où l'on souhaite profiter de la capacité d'abstraction des réseaux de neurones.

### 5.1.3.3 Avantages et inconvénients

Le tableau suivant résume les avantages et les inconvénients du modèle MLP Classifier :

Avantages	Inconvénients
<ul style="list-style-type: none"><li>• Grande flexibilité : s'adapte à tout type de données, continues ou catégorielles.</li><li>• Modélisation de relations complexes : capte des patterns non évidents dans les données.</li><li>• Adapté aux données de haute dimension</li><li>• Bonne capacité de généralisation avec une régularisation adaptée.</li></ul>	<ul style="list-style-type: none"><li>• Temps d'entraînement élevé, surtout avec plusieurs couches cachées ou un grand volume de données.</li><li>• Nécessite un réglage minutieux des hyperparamètres.</li><li>• Moins interprétable que des modèles à base d'arbres ou de régressions.</li><li>• Sensible à la normalisation des données : nécessite souvent un prétraitement (MinMaxScaler, StandardScaler).</li></ul>

**TABLEAU 5.3** : Avantages et inconvénients du MLPClassifier

### 5.1.3.4 Paramètres

Le tableau suivant décrit les hyperparamètres du modèle MLP Classifier :

Hyperparamètre	Description
<code>hidden_layer_sizes</code>	Tuple indiquant le nombre de neurones dans chaque couche cachée (ex : (100, 50) pour deux couches).
<code>activation</code>	Fonction d'activation des neurones cachés ('relu', 'tanh', 'logistic', 'identity').
<code>solver</code>	Algorithme d'optimisation ('adam' recommandé pour la plupart des cas).
<code>alpha</code>	Terme de <b>régularisation L2</b> pour éviter le surapprentissage (par défaut 0.0001).
<code>learning_rate</code>	Taux d'apprentissage ('constant', 'invscaling', 'adaptive').
<code>max_iter</code>	Nombre maximal d'itérations pour converger.
<code>early_stopping</code>	Permet d'arrêter l'entraînement si la performance ne s'améliore plus (booléen).
<code>random_state</code>	Fixe la graine aléatoire pour la reproductibilité des résultats.

TABLEAU 5.4 : Description des hyperparamètres du MLPClassifier

#### 5.1.4 XGBoost Classifier

XGBoost (Extreme Gradient Boosting) est un modèle d'arbres de décision basé sur la méthode du gradient boosting. Il s'agit d'un modèle supervisé, puissant pour la classification et la régression, qui combine plusieurs arbres faibles pour former un modèle robuste et précis.

##### 5.1.4.1 Fonctionnement

XGBoost construit les arbres de manière séquentielle, chaque nouvel arbre corrigeant les erreurs des arbres précédents en minimisant une fonction de perte via une optimisation par gradient. Il intègre des techniques avancées comme la régularisation, le traitement efficace des données manquantes et le parallélisme pour accélérer l'entraînement.

##### 5.1.4.2 Quand l'utiliser ?

- Lorsque les données comportent des interactions complexes et non linéaires.
- Pour des jeux de données volumineux ou avec des variables hétérogènes.

- Dans les cas où l'on souhaite un modèle performant avec une bonne capacité de généralisation.
- Lorsque l'importance des variables est nécessaire pour l'interprétation.

#### 5.1.4.3 Avantages et inconvénients

Le tableau suivant résume les avantages et les inconvénients du modèle XGBoost Classifier :

Avantages	Inconvénients
<ul style="list-style-type: none"><li>• Très haute performance sur de nombreux problèmes de classification et régression.</li><li>• Bonne gestion des données manquantes et hétérogènes.</li><li>• Possibilité de régularisation pour réduire le surapprentissage.</li><li>• Rapide grâce à l'optimisation parallèle et au traitement efficace des données.</li><li>• Offre des indicateurs d'importance des variables utiles pour l'interprétation.</li></ul>	<ul style="list-style-type: none"><li>• Modèle plus complexe à paramétrer (nombre d'arbres, profondeur, taux d'apprentissage).</li><li>• Peut être sensible au bruit dans les données si mal régularisé.</li><li>• Moins interprétable qu'un modèle linéaire simple.</li><li>• Nécessite souvent un prétraitement comme l'encodage des variables catégorielles.</li></ul>

TABLEAU 5.5 : Avantages et inconvénients de XGBoost

#### 5.1.4.4 Paramètres

Le tableau suivant décrit les hyperparamètres du modèle XGBoost :

Hyperparamètre	Description
<code>n_estimators</code>	Nombre d'arbres à construire dans le modèle.
<code>max_depth</code>	Profondeur maximale de chaque arbre (contrôle la complexité du modèle).
<code>learning_rate</code>	Taux d'apprentissage ou "shrinkage" utilisé pour réduire l'impact de chaque arbre.
<code>subsample</code>	Proportion des données utilisées pour construire chaque arbre (utilisé pour réduire le surapprentissage).
<code>colsample_bytree</code>	Proportion des variables utilisées pour construire chaque arbre (randomisation des colonnes).
<code>gamma</code>	Seuil minimum de réduction de la perte pour effectuer une nouvelle partition dans un arbre (régularisation).
<code>reg_alpha</code>	Terme de régularisation L1 (lasso) sur les poids.
<code>reg_lambda</code>	Terme de régularisation L2 (ridge) sur les poids.
<code>objective</code>	Fonction objectif à optimiser (ex : <code>'multi:softprob'</code> pour classification multiclasse).
<code>random_state</code>	Graine aléatoire pour assurer la reproductibilité.

TABLEAU 5.6 : Description des hyperparamètres de XGBoost

## 5.2 Génération des rapports

Cette section présente le système de génération de rapports mis en œuvre dans le projet, destiné à fournir aux responsables RH une visualisation claire des mouvements internes et du taux de rétention des employés. Ce module repose sur l'automatisation du suivi des prédictions issues du modèle de machine learning, l'identification des indicateurs clés de performance (KPIs), ainsi que la mise en œuvre de technologies adaptées à la visualisation interactive des données.

### 5.2.1 Identification du besoin

Dans le cadre du suivi de la mobilité interne et de la gestion des talents, il est essentiel pour les responsables RH de disposer d'une vue synthétique, actualisée et exploitable des mouvements des employés, ainsi que du taux de rétention. L'objectif est de permettre une prise de décision rapide, basée

sur les données, à travers une interface ergonomique.

Afin de répondre à ce besoin, un rapport dynamique et automatisé a été mis en place, permettant d’afficher à la fois l’état actuel et l’historique des prédictions effectuées par le modèle de machine learning, sous forme de visualisations interactives.

### 5.2.2 Identification des indicateurs clés de performance

Plusieurs indicateurs clés de performance (KPIs) ont été définis pour suivre l’évolution des mouvements internes des employés. Le tableau suivant présente une synthèse des principaux KPIs utilisés :

KPI	Définition
Total employés	Nombre d’employés analysés à partir du fichier chargé
Taux de rétention	Pourcentage des employés restés dans l’entreprise (c’est-à-dire non classés comme sortants)
Répartition des mouvements	Part respective des mouvements identifiés : <i>Entrée</i> , <i>Interne</i> , et <i>Sortie</i>
Rétention vs Turnover	Représentation synthétique du taux de départ (turnover) face au taux de rétention
Mouvements par département	Distribution des mouvements (en %) pour chaque département

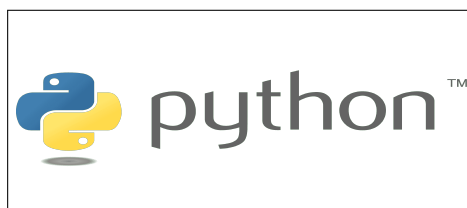
**TABLEAU 5.7** : Indicateurs clés de performance suivis dans l’analyse RH

Ces indicateurs permettent d’évaluer rapidement la stabilité du personnel et d’identifier les zones à risque ou en tension (par exemple, un taux de sortie élevé dans un service spécifique).

### 5.2.3 Technologies et bibliothèques utilisées

Le système de génération de rapports a été développé entièrement en **Python**, un langage largement utilisé en science des données pour sa simplicité, sa richesse en bibliothèques et sa capacité à gérer l’automatisation de tâches analytiques.

L’image ci-dessous présente le logo de Python



**FIGURE 5.1** : Logo Python

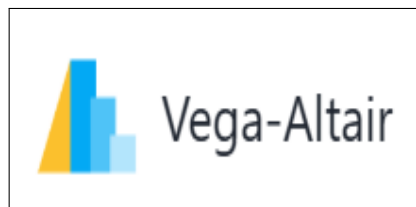
Afin de choisir une bibliothèque de visualisation adaptée à notre besoin de création de rapports interactifs, nous avons comparé deux des principales solutions disponibles en Python : Altair et Matplotlib. Le tableau ci-dessous présente les différences clés entre ces deux outils selon plusieurs critères.

Critère	Altair	Matplotlib
Type de bibliothèque	Déclarative (basée sur Vega-Lite)	Impérative (bas niveau)
Facilité d'utilisation	Syntaxe concise, rapide à prendre en main	Syntaxe plus complexe, nécessite plus de lignes de code
Interactivité	Native et fluide (zoom, survol, filtres)	Limitée, nécessite des extensions (e.g. mpld3, Plotly)

**TABLEAU 5.8 :** Comparaison entre Altair et Matplotlib

Au vu de ses capacités interactives, de sa simplicité d'utilisation et de sa fluidité, Altair a été retenue comme la solution idéale pour la génération dynamique et visuelle des rapports RH dans notre projet. Il s'agit d'une librairie déclarative basée sur Vega-Lite, qui permet de produire des graphiques interactifs et esthétiques de manière concise.

L'image ci-dessous présente le logo de la bibliothèque ALTAIR :



**FIGURE 5.2 :** Logo ALTAIR

L'usage combiné de Python et d'Altair a permis de concevoir une interface réactive, facile à maintenir et offrant une visualisation intuitive des indicateurs RH clés.

### 5.3 Génération des stratégies de fidélisation

Afin de prévenir les départs volontaires et d'optimiser la gestion des talents, il est indispensable de proposer aux responsables RH des recommandations personnalisées, fondées sur l'analyse des données disponibles. Cette section détaille le fonctionnement du module de génération automatique de stratégies de fidélisation, articulé autour de l'exploitation conjointe d'un modèle de classification et

d'un modèle de langage avancé.

### 5.3.1 Identification du besoin

Dans le cadre de la prévention des départs, il est essentiel de proposer des actions concrètes et personnalisées pour retenir les talents à risque. C'est pourquoi une étape clé de notre solution consiste à générer automatiquement des stratégies de fidélisation, à partir des données et des prédictions disponibles. L'objectif est de fournir aux responsables RH des leviers d'action ciblés et exploitables pour chaque employé concerné.

### 5.3.2 Rôle du modèle de langage

Le modèle de langage (LLM) joue un rôle central dans la génération des stratégies de fidélisation, en transformant des données analytiques en recommandations RH concrètes et compréhensibles. Contrairement à un modèle classique de classification, le LLM est capable d'interpréter un ensemble de variables, d'en extraire le sens contextuel, et de proposer une action adaptée sous forme textuelle. Dans notre démarche, nous exploitons le modèle Mistral pour produire automatiquement une stratégie personnalisée pour chaque employé à risque. Pour ce faire, quatre éléments lui sont fournis : le résultat de la prédiction ("Sortie", "Interne", etc.), la probabilité associée, les variables les plus influentes dans la décision (comme le salaire, la satisfaction ou la formation), ainsi que la médiane globale de ces variables au sein de l'entreprise. Ces informations permettent au LLM de formuler une réponse cohérente, contextualisée et exploitable par les RH, en apportant une justification implicite et une orientation stratégique ciblée.

### 5.3.3 Données fournies au LLM

Pour guider le modèle de langage dans la formulation de stratégies pertinentes, quatre types d'informations lui sont systématiquement transmis :

**Le résultat de la prédiction** Le résultat de la prédiction correspond à la catégorie de mouvement anticipé pour chaque employé. Il est obtenu à l'issue du modèle de classification, qui attribue à chaque individu un type de mouvement prédit. Ce résultat détermine l'objectif stratégique de la recommandation.

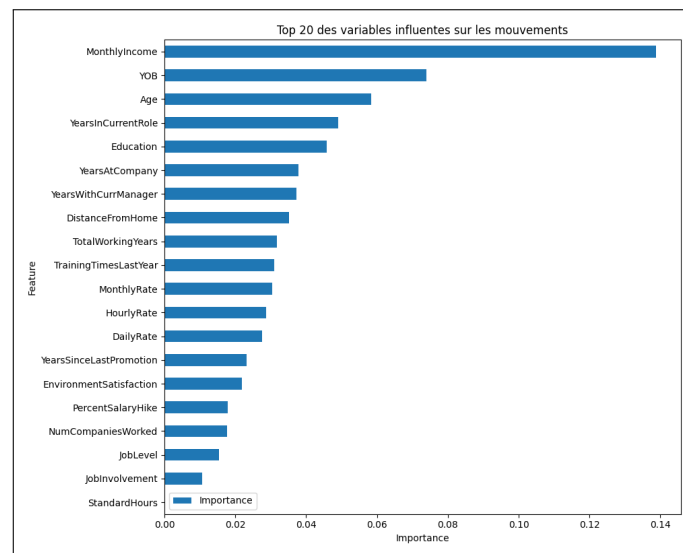
**La probabilité de la prédiction** La probabilité de la prédiction représente le degré de confiance du modèle dans le mouvement prédit. Elle est directement récupérée via la méthode "predict\_proba" du classificateur, et permet d'estimer l'urgence ou la priorité d'intervention.



**Les variables les plus influentes** Ces variables sont identifiées par le modèle de prédiction et jouent un rôle crucial dans l’élaboration des stratégies de fidélisation. Ces variables correspondent aux facteurs qui ont le plus fortement contribué à la décision du modèle concernant le mouvement prédit. En comprenant quels éléments ont pesé le plus dans la prédiction on est en mesure de mieux cerner les causes potentielles d’un départ. Ces variables servent alors de leviers d’action prioritaires pour orienter les recommandations générées par le LLM. En les intégrant dans le prompt, le modèle linguistique est guidé pour produire des suggestions qui répondent directement aux problématiques sous-jacentes.

Ainsi, l’utilisation des variables les plus influentes garantit que les stratégies proposées sont pertinentes, personnalisées et ciblées, car elles prennent appui sur les véritables causes identifiées par le système prédictif, plutôt que sur des hypothèses générales.

L’image ci-dessous montre les variables les plus influentes



**FIGURE 5.3 :** Les variables les plus influentes

**La médiane globale** La médiane globale calculée pour chacune des variables les plus influentes sur l’ensemble du dataset, joue un rôle fondamental dans l’interprétation contextuelle des données de chaque employé. En fournissant un point de référence objectif et stable, elle permet de situer l’individu par rapport au reste du personnel analysé.

Par exemple, si un employé présente un niveau de satisfaction au travail ou un revenu mensuel inférieur à la médiane, cela peut indiquer une source potentielle d’insatisfaction ou un facteur de risque. À l’inverse, des valeurs supérieures à la médiane peuvent être le signe d’une bonne intégration ou d’une reconnaissance adéquate, ce qui oriente naturellement la stratégie vers le maintien ou le renforcement de ces conditions.

L’inclusion de cette comparaison dans le prompt envoyé au modèle de langage offre un contexte supplémentaire, permettant au LLM de mieux comprendre si la situation de l’employé est favorable ou

préoccupante, et donc d'ajuster la tonalité et le contenu de la recommandation.

En résumé, la médiane globale enrichit l'analyse en apportant une dimension comparative, rendant les stratégies générées plus nuancées, personnalisées et ancrées dans la réalité des données.

## 5.4 Agent conversationnel

Dans ce qui suit, nous allons présenter le concept d'agent conversationnel, ses principes de fonctionnement, ainsi que les outils utilisés pour permettre une interaction intuitive et efficace entre les utilisateurs et les données RH.

### 5.4.1 Définition de l'IA agentique

L'IA agentique représente une évolution majeure de l'intelligence artificielle, caractérisée par la capacité des systèmes à agir de manière autonome pour atteindre des objectifs définis. Contrairement aux modèles de langage traditionnels qui se contentent de générer du texte, l'IA agentique peut :

- Planifier une séquence d'actions pour résoudre un problème complexe
- Utiliser des outils externes (bases de données, APIs, calculateurs)
- Raisonner sur les résultats intermédiaires pour ajuster sa stratégie
- Interagir avec l'environnement de manière itérative

### 5.4.2 Identification du besoin et objectifs

Dans le contexte de la démocratisation de l'accès aux données RH, un besoin critique émerge : permettre aux responsables RH et managers opérationnels, souvent non techniques, d'interroger directement les résultats analytiques sans maîtriser SQL ou les outils de Business Intelligence. L'agent conversationnel est un agent SQL conçu pour répondre aux requêtes formulées en langage naturel. Son rôle principal est de :

- Démocratiser l'accès aux résultats de l'analyse RH
- Faciliter la prise de décision par une interaction directe et intuitive
- Fournir des réponses synthétiques, précises et contextualisées.

### 5.4.3 Fonctionnement de l'agent

Cette partie détaille les différentes étapes qui permettent à l'agent de comprendre une question en langage naturel, de la traduire en requête SQL, puis de formuler une réponse claire et exploitable pour l'utilisateur final.

**Compréhension du langage naturel** L’agent analyse la question utilisateur pour identifier :

- **L’intention** : que veut savoir l’utilisateur ?
- **Les entités** : quelles tables, colonnes ou valeurs sont concernées ?
- **Le contexte** : période temporelle, conditions spécifiques, agrégations nécessaires

**Mapping Sémantique** : L’agent établit une correspondance entre les termes du langage naturel et le schéma de base de données :

- **Synonymes** : employés -> table employees
- **Relations** : identification des jointures nécessaires entre tables
- **Contraintes** : application des filtres et conditions

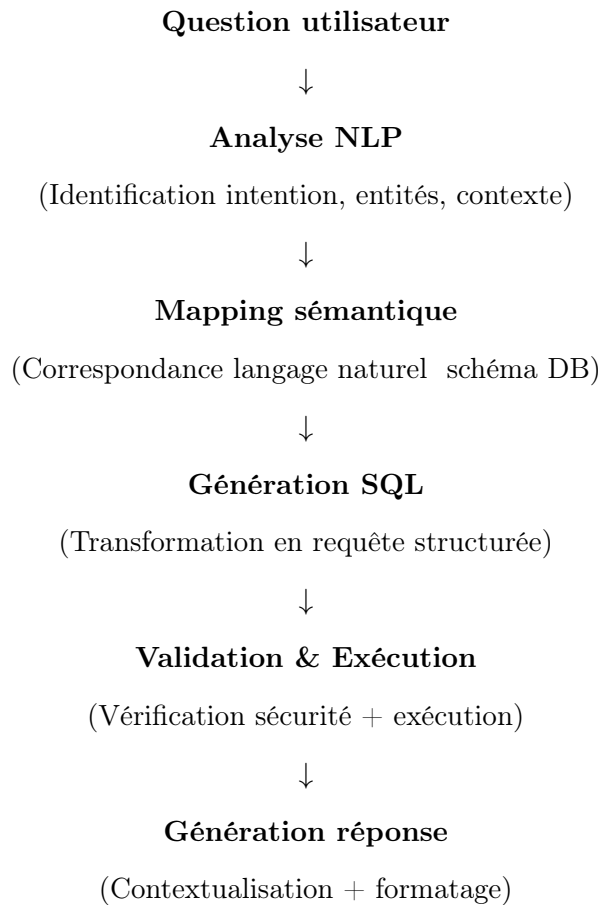
**Génération de requête SQL** : Transformation de l’intention en requête SQL syntaxiquement correcte.

**Validation syntaxique et sémantique** : vérification de la structure SQL et de la cohérence avec le schéma de données.

**Exécution sécurisée de la requête SQL** : Une fois la requête générée, elle est exécutée via un outil sécurisé qui vérifie l’absence d’opérations destructrices (DELETE, DROP, etc.) et s’assure que seules les tables et colonnes autorisées sont utilisées. Le résultat brut est récupéré sous forme tabulaire (DataFrame), généralement converti en Markdown pour une lisibilité optimale.

**Génération de la réponse interprétée** : Le résultat SQL est ensuite transmis au LLM qui, en se basant sur le contexte de la question et la structure des données, génère une réponse formulée en langage naturel. Cette réponse vise à être compréhensible, contextualisée et directement exploitable par un utilisateur non technique.

La figure suivante résume les étapes de fonctionnement de l’agent :



**FIGURE 5.4 :** Flux de traitement de l’agent conversationnel

#### 5.4.4 Outils utilisés

La partie suivante présente les outils mobilisés pour concevoir l’agent conversationnel SQL.

**Langchain** Framework central de l’agent, Langchain permet de chaîner les différentes étapes du raisonnement (analyse de la requête, appel aux outils, gestion du prompt) et d’orchestrer les interactions entre le LLM et la base de données.

La figure suivante présente le logo de langchain :



**FIGURE 5.5 :** Logo de langchain

**SQLite + SQLAlchemy** Une base de données relationnelle légère (SQLite) est utilisée pour stocker les informations RH, avec SQLAlchemy comme couche d’abstraction pour interagir avec Langchain et valider les requêtes.

La figure suivante présente le logo de SQLite :



**FIGURE 5.6 :** Logo SQLite

**Ollama + Mistral** Ollama permet d'exécuter localement le modèle Mistral, un LLM performant spécialisé en génération de texte. Il est utilisé pour interpréter la requête utilisateur et formuler des réponses RH contextualisées.

## Conclusion

Ce chapitre a détaillé la phase de modélisation de notre solution de gestion des talents, couvrant le développement d'un modèle prédictif pour anticiper les départs d'employés, d'un modèle génératif afin de générer des stratégies de fidélisation personnalisées et enfin la création d'un agent conversationnel pour démocratiser l'accès aux données RH.

# EVALUATION

---

## Plan

1	Evaluation des modèles prédictifs . . . . .	59
2	Evaluation des modèles génératifs . . . . .	67

## Introduction

L'évaluation constitue l'étape décisive qui valide la pertinence et l'efficacité de notre solution. Ce chapitre adopte une approche bidimensionnelle inédite : d'une part, l'évaluation quantitative rigoureuse des modèles prédictifs à travers des métriques adaptées au contexte multiclasse des mouvements RH ; d'autre part, l'assessment qualitatif des stratégies de fidélisation générées par Mistral, combinant critères linguistiques et pertinence métier.

### 6.1 Evaluation des modèles prédictifs

Cette première section présente l'évaluation comparative rigoureuse des trois algorithmes de classification implémentés.

#### 6.1.1 Métriques d'évaluation

L'évaluation des modèles prédictifs dans le contexte de la classification multiclasse nécessite un ensemble de métriques complémentaires, chacune apportant un éclairage spécifique sur les performances du modèle.

##### 6.1.1.1 Précision globale(Accuracy)

L'accuracy représente la proportion d'observations correctement classées par rapport au nombre total d'observations :

$$\text{Accuracy} = \frac{\text{Nombre de prédictions correctes}}{\text{Nombre total de prédictions}}$$

Cette métrique fournit une vue d'ensemble des performances du modèle. Cependant, dans le contexte de données déséquilibrées, où certaines classes sont majoritaires, l'accuracy peut être trompeuse. Un modèle prédisant systématiquement la classe majoritaire pourrait afficher une accuracy élevée tout en étant inefficace pour détecter les cas critiques de départ.

##### 6.1.1.2 F1-score pondéré

Le F1-score pondéré constitue notre métrique principale d'évaluation. Il calcule le F1-score pour chaque classe, puis effectue une moyenne pondérée par le nombre d'échantillons de chaque classe :

$$\text{F1-score pondéré} = \sum_{i=1}^K w_i \cdot \text{F1}_i$$

Cette approche permet de tenir compte du déséquilibre des classes tout en donnant plus

d'importance aux classes les mieux représentées, ce qui correspond à notre objectif de performance globale équilibrée.

#### 6.1.1.3 Précision

La précision mesure la proportion de prédictions positives correctes parmi toutes les prédictions positives pour une classe donnée :

$$\text{Précision} = \frac{\text{Vrai Positifs}}{\text{Vrai Positifs} + \text{Faux Positifs}}$$

Une précision élevée pour la classe "Sortie" signifie que lorsque le modèle prédit un départ, cette prédiction est généralement correcte, minimisant ainsi les fausses alertes.

#### 6.1.1.4 Rappel

Le rappel indique la proportion de cas positifs réels correctement identifiés :

$$\text{Rappel} = \frac{\text{Vrai Positifs}}{\text{Vrai Positifs} + \text{Faux Négatifs}}$$

Un rappel élevé pour la classe "Sortie" garantit que la plupart des employés réellement à risque de départ sont identifiés, ce qui est crucial pour une intervention préventive efficace.

#### 6.1.1.5 F1 score

Le F1-score combine précision et rappel en une seule métrique harmonique :

$$F1\text{-score}_i = 2 * \frac{\text{Precision}_i * \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

Cette métrique est particulièrement pertinente pour évaluer l'équilibre entre la détection des cas à risque et la minimisation des fausses alertes.

#### 6.1.1.6 Matrice de confusion

La matrice de confusion fournit une vue détaillée des erreurs de classification en présentant, pour chaque classe réelle, la distribution des prédictions effectuées. Elle permet d'identifier les confusions spécifiques entre classes.

#### 6.1.1.7 Macro et Micro moyennes

- **Macro-moyenne** : calcule la métrique pour chaque classe puis effectue une moyenne non pondérée



$$\text{Macro-F1} = \frac{1}{N} \sum_{i=1}^N F1_i$$

- **Micro-moyenne** : agrège les contributions de toutes les classes pour calculer la métrique globale

$$\text{Micro-F1} = \frac{2 * \text{Micro-Precision} * \text{Micro-Recall}}{\text{Micro-Precision} + \text{Micro-Recall}}$$

### 6.1.2 RandomForest

L'évaluation du modèle Random Forest sur notre jeu de données RH révèle les performances suivantes :

#### 6.1.2.1 Métriques globales

- **Accuracy globale** : 61.7%
- **F1-score pondéré** : 60.2%
- **F1-score macro** : 68.2%
- **Précision macro** : 69.5%
- **Rappel macro** : 68.1%

#### 6.1.2.2 Analyse détaillée par classe

Le tableau suivant résume les performances du modèle RandomForest par classe :

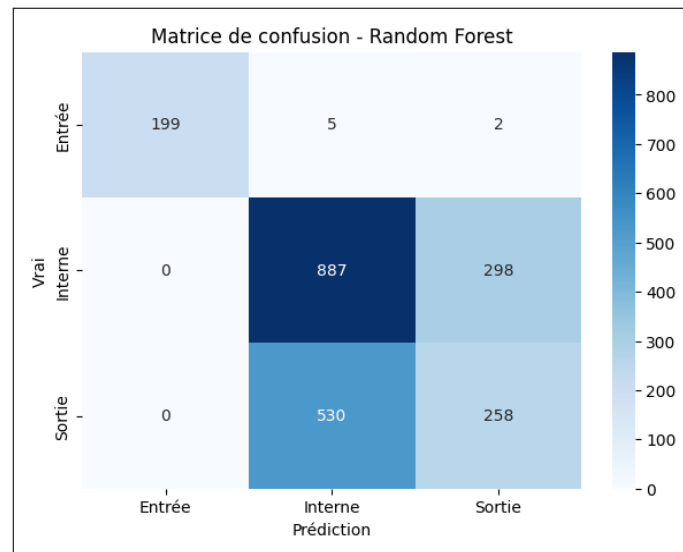
Métrique	Entrée (Stabilité)	Interne (Mouvement)	Sortie (Départ)
Précision	100%	62%	46%
Rappel	97%	75%	33%
F1-score	98%	68%	38%

**TABLEAU 6.1** : Performances par classe pour le modèle Random Forest

Le modèle Random Forest affiche d'excellentes performances pour la classe "Entrée", avec des scores presque parfaits, ce qui montre sa capacité à identifier les profils stables. Les résultats pour la classe "Interne" sont modérés mais équilibrés, bien que le modèle ait tendance à confondre certains départs avec des mouvements internes. Enfin, les performances pour la classe "Sortie" restent faibles, ce qui suggère une difficulté à détecter correctement les signaux de départ des employés.

### 6.1.2.3 Matrice de confusion

L'image ci-dessous présente la matrice de confusion du modèle Random Forest :



**FIGURE 6.1 :** Matrice de confusion - Random Forest

La matrice de confusion révèle que Random Forest excelle dans l'identification des profils stables ("Entrée") sans aucune confusion. Cependant, le modèle présente une confusion majeure "Sortie" → "Interne" avec 530 cas (67% des erreurs), suggérant une similarité comportementale entre ces catégories. L'asymétrie des erreurs (seulement 298 cas dans l'autre sens) indique une approche conservatrice du modèle dans ses prédictions de départ, privilégiant la prudence au risque de sous-estimer les cas de sortie réels.

### 6.1.3 MLP Classifier

L'évaluation du modèle MLP Classifier sur notre jeu de données RH révèle les performances suivantes :

#### 6.1.3.1 Métriques globales

- **Accuracy globale** : 55.5%
- **F1-score pondéré** : 55.0%
- **F1-score macro** : 61.1%
- **Précision macro** : 63.5%
- **Rappel macro** : 59.4%

#### 6.1.3.2 Analyse détaillée par classe

Le tableau suivant résume les performances du modèle MLP Classifier par classe :

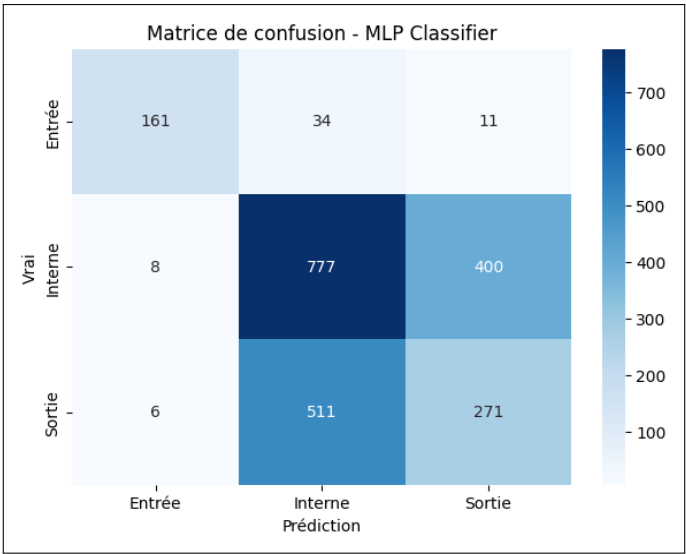
Métrique	Entrée (Stabilité)	Interne (Mouvement)	Sortie (Départ)
Précision	92%	59%	40%
Rappel	78%	66%	34%
F1-score	85%	62%	37%

**TABLEAU 6.2 :** Performances par classe pour le modèle MLP Classifier

Le modèle MLP Classifier présente des performances contrastées selon les classes. Pour la classe "Entrée", il affiche une précision élevée (92%) mais un rappel plus modéré (78%), indiquant une capacité correcte à identifier les profils stables tout en manquant certains cas. Les résultats pour la classe "Interne" sont équilibrés avec des scores modérés, suggérant une performance acceptable pour détecter les mouvements internes. Cependant, les performances pour la classe "Sortie" demeurent préoccupantes, avec seulement 34% de rappel, ce qui signifie que le modèle échoue à détecter deux tiers des départs réels.

**6.1.3.3 Matrice de confusion**

L'image ci-dessous présente la matrice de confusion du modèle MLP Classifier :



**FIGURE 6.2 :** Matrice de confusion - MLP Classifier

La matrice de confusion du MLP Classifier montre une bonne discrimination de la classe "Entrée" (161/206 correctes) malgré quelques confusions mineures. Le problème principal réside dans la confusion "Sortie" → "Interne" avec 511 cas (65% des erreurs), révélant les difficultés du réseau de neurones à distinguer ces profils. Contrairement aux autres modèles, MLP Classifier présente une confusion bidirectionnelle significative (400 cas "Interne" → "Sortie"), indiquant une zone d'incertitude

importante entre ces deux catégories.

### 6.1.4 XGBoost

L'évaluation du modèle XGBoost sur notre jeu de données RH révèle les performances suivantes :

#### 6.1.4.1 Métriques globales

- **Accuracy globale** : 61.9%
- **F1-score pondéré** : 60.9%
- **F1-score macro** : 68.9%
- **Précision macro** : 70.0%
- **Rappel macro** : 68.6%

#### 6.1.4.2 Analyse détaillée par classe

Le tableau suivant résume les performances du modèle XGBoost par classe :

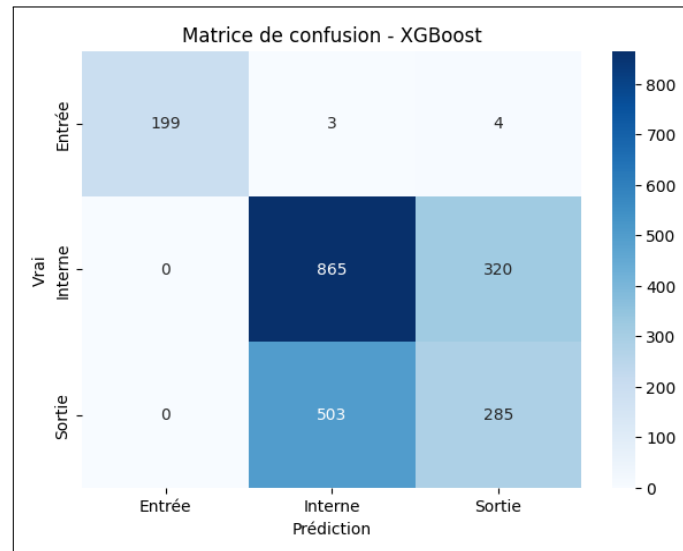
Métrique	Entrée (Stabilité)	Interne (Mouvement)	Sortie (Départ)
Précision	100%	63%	47%
Rappel	97%	73%	36%
F1-score	98%	68%	41%

**TABEAU 6.3** : Performances par classe pour le modèle XGBoost

Le modèle XGBoost démontre d'excellentes performances pour la classe "Entrée", atteignant une précision parfaite (100%) et un rappel très élevé (97%), confirmant sa capacité exceptionnelle à identifier les profils d'employés stables. Pour la classe "Interne", les résultats sont satisfaisants avec un équilibre correct entre précision (63%) et rappel (73%), indiquant une bonne détection des mouvements internes. Cependant, la classe "Sortie" présente des performances limitées avec seulement 36% de rappel, signifiant que le modèle ne parvient à détecter qu'un tiers des départs réels, bien que sa précision de 47% soit légèrement supérieure aux autres modèles.

#### 6.1.4.3 Matrice de confusion

L'image ci-dessous présente la matrice de confusion du modèle XGBoost :

**FIGURE 6.3 :** Matrice de confusion - XGBoost

La matrice de confusion de XGBoost révèle une discrimination quasi-parfaite de la classe "Entrée" (199/206 correctes) avec très peu de confusions. Cependant, le modèle présente la même confusion majeure "Sortie" → "Interne" avec 503 cas (64% des erreurs), montrant que même les techniques avancées de gradient boosting peinent à distinguer ces profils. L'asymétrie des erreurs (320 cas dans l'autre sens) confirme l'approche prudente de XGBoost, privilégiant la minimisation des fausses alertes au détriment de la détection des départs.

### 6.1.5 Benchmark comparatif des modèles

Le tableau suivant présente une synthèse comparative des performances des trois modèles évalués sur notre jeu de données RH :

Métrique	Random Forest	MLP Classifier	XGBoost
Métriques Globales			
Accuracy globale	61.7%	55.5%	<b>61.9%</b>
F1-score pondéré	60.2%	55.0%	<b>60.9%</b>
F1-score macro	<b>68.2%</b>	61.1%	68.9%
Précision macro	69.5%	63.5%	<b>70.0%</b>
Rappel macro	68.1%	59.4%	<b>68.6%</b>
Classe "Entrée" (Stabilité)			
Précision	<b>100%</b>	92%	<b>100%</b>
Rappel	<b>97%</b>	78%	<b>97%</b>
F1-score	<b>98%</b>	85%	<b>98%</b>
Classe "Interne" (Mouvement)			
Précision	62%	59%	<b>63%</b>
Rappel	<b>75%</b>	66%	73%
F1-score	<b>68%</b>	62%	68%
Classe "Sortie" (Départ)			
Précision	46%	40%	<b>47%</b>
Rappel	33%	34%	<b>36%</b>
F1-score	38%	37%	<b>41%</b>
Analyse des Erreurs			
Confusion "Sortie" → "Interne"	530 cas	511 cas	<b>503 cas</b>
Confusion "Interne" → "Sortie"	298 cas	<b>400 cas</b>	320 cas
Erreurs sur "Entrée"	<b>7 cas</b>	45 cas	7 cas

TABLEAU 6.4 : Benchmark comparatif des modèles de classification

À l'issue de cette analyse comparative, **XGBoost** a été retenu comme modèle optimal pour notre système de prédiction des mouvements RH. Cette décision repose sur plusieurs critères déterminants : premièrement, XGBoost affiche les meilleures performances globales avec une accuracy de 61.9% et un F1-score macro de 68.9%. Deuxièmement, et de manière cruciale pour notre contexte métier, XGBoost présente les meilleures performances sur la classe critique "Sortie" avec un F1-score de 41%, une précision de 47% et un rappel de 36%, minimisant ainsi les risques de non-détection des départs imminents. Enfin, le modèle génère le moins de confusions "Sortie" → "Interne", réduisant les erreurs coûteuses en contexte RH.

## 6.2 Evaluation des modèles génératifs

Cette seconde section développe une méthodologie d'évaluation qualitative originale pour les stratégies de fidélisation générées par Mistral.

### 6.2.1 Contexte et défis d'évaluation

L'évaluation des modèles génératifs dans le contexte de la génération de stratégies de fidélisation RH présente des défis uniques qui diffèrent fondamentalement de l'évaluation des modèles prédictifs. Contrairement aux tâches de classification où une "vérité terrain" objective existe, l'évaluation de la qualité des stratégies générées par Mistral repose sur des critères subjectifs et contextuels.

Le principal défi réside dans l'absence de référence absolue pour juger de la pertinence d'une stratégie de fidélisation. Une recommandation peut être techniquement correcte, linguistiquement fluide, mais inadaptée au contexte organisationnel spécifique ou aux contraintes budgétaires de l'entreprise. De plus, l'efficacité réelle d'une stratégie ne peut être mesurée qu'à long terme, après sa mise en œuvre effective.

Cette complexité nous amène à adopter une approche d'évaluation multidimensionnelle, combinant des critères qualitatifs standardisés, une analyse de la cohérence contextuelle, et une validation de la faisabilité opérationnelle des recommandations générées.

### 6.2.2 Qualité linguistique

Pour évaluer la qualité linguistique des stratégies générées par Mistral, nous analysons plusieurs critères clés qui garantissent la clarté, la cohérence et l'adéquation du texte au contexte professionnel RH

Critère	Description
<b>Fluidité et clarté du texte</b>	Mistral excelle dans la production de textes fluides et compréhensibles. Les stratégies générées présentent une structure logique, un vocabulaire approprié au contexte professionnel, et une progression argumentative cohérente.
<b>Structure et organisation</b>	Les recommandations suivent généralement une structure claire : diagnostic de la situation, identification des leviers d'action, propositions concrètes, et modalités de suivi. Cette organisation facilite la lecture et l'appropriation par les responsables RH.
<b>Ton professionnel approprié</b>	Le modèle maintient un registre de langue adapté au contexte RH, évitant les formulations trop familières ou trop techniques. Le ton reste bienveillant et constructif, approprié pour des recommandations de fidélisation.

**TABLEAU 6.5 :** Évaluation de la qualité linguistique des stratégies générées par Mistral

### 6.2.3 Utilisation des données contextuelles

L'évaluation de l'utilisation des données contextuelles par Mistral se concentre sur sa capacité à intégrer efficacement les informations clés issues des analyses prédictives pour enrichir la pertinence et l'adaptabilité des recommandations.



Critère	Description
<b>Intégration des variables influentes</b>	Mistral démontre une capacité satisfaisante à intégrer les variables les plus influentes identifiées par le modèle prédictif dans ses recommandations. Les stratégies font explicitement référence aux facteurs de risque détectés (satisfaction, rémunération, équilibre vie professionnelle/personnelle).
<b>Exploitation de la médiane globale</b>	Le modèle utilise efficacement les comparaisons avec les médianes globales pour contextualiser la situation de l'employé et justifier ses recommandations. Cette approche comparative enrichit la pertinence des stratégies proposées.
<b>Prise en compte de la probabilité de prédiction</b>	L'urgence et l'intensité des recommandations s'adaptent généralement au niveau de risque indiqué par la probabilité de départ. Les stratégies pour les employés à très haut risque sont plus détaillées et proposent des actions plus immédiates.

TABLEAU 6.6 : Évaluation de l'utilisation des données contextuelles par Mistral

## Conclusion

L'évaluation croisée de nos deux volets a confirmé la pertinence de la solution SmartTrack. Le modèle XGBoost s'est distingué par ses performances prédictives, atteignant un F1-score macro de 68,9%, tandis que le modèle Mistral a démontré sa capacité à générer des stratégies de fidélisation à la fois fluides sur le plan linguistique et cohérentes sur le plan contextuel. Le chapitre suivant présentera l'intégration opérationnelle de ces modèles dans une interface utilisateur intuitive.

# DÉPLOIEMENT

---

## Plan

1	Présentation de l'environnement de developpement . . . . .	71
2	Réalisation . . . . .	71

## Introduction

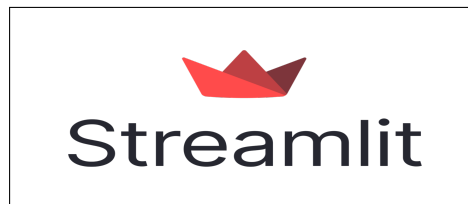
Ce chapitre représente le fruit de notre travail acharné, dans lequel nous décrirons en détail les environnements de travail ainsi que les interfaces de la solution.

### 7.1 Présentation de l'environnement de developpement

L'interface utilisateur du projet a été développée à l'aide de Streamlit, un framework open-source en Python conçu spécifiquement pour la création rapide d'applications web interactives, notamment dans les domaines de la data science et de l'intelligence artificielle. Sa simplicité de déploiement, sa compatibilité avec des bibliothèques de traitement de données telles que Pandas ou NumPy, ainsi que sa capacité à intégrer des modèles de machine learning en font un outil adapté à des prototypes de solutions digitales dans le domaine des services architecturaux.

Streamlit permet une intégration directe du code Python dans une interface visuelle, facilitant ainsi l'expérimentation et l'itération dans le processus de développement. L'architecture de l'application repose sur un enchaînement d'interfaces permettant d'entrer des données, de visualiser des résultats analytiques et de simuler des scénarios dans le cadre d'une stratégie d'entrée sur le marché des services d'architecture.

L'image ci-dessous présente le logo de Streamlit :



**FIGURE 7.1 :** Logo Streamlit

### 7.2 Réalisation

Cette section présente la phase finale de la réalisation du projet. Elle met en évidence les interfaces qui ont été conçues et intégrées dans l'application SmartTrack.

L'image ci-dessous présente le logo de notre application SmartTrack ainsi que notre slogan :



**FIGURE 7.2 :** Logo SmartTrack

### 7.2.1 Interface Rapport

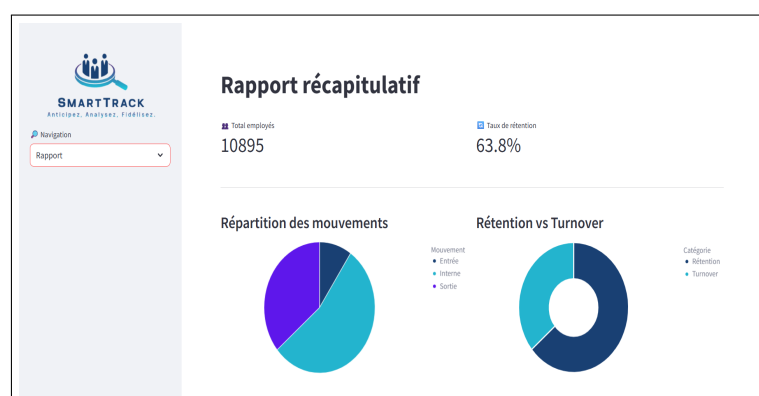
L'interface Rapport joue un rôle central dans le suivi analytique et stratégique des mouvements RH au sein de l'entreprise. Elle permet d'accéder à une vue synthétique et visuelle des principaux indicateurs de mobilité. Grâce à l'intégration de graphiques interactifs, cette section facilite l'identification des tendances et la détection des départements à risque, permettant aux responsables RH de disposer d'éléments objectifs pour piloter leurs décisions. En résumé, il s'agit d'un tableau de bord dynamique qui transforme les données de prédiction en indicateurs exploitables pour une gestion proactive des talents.

Sur cette première interface on voit le menu de navigation à gauche avec les différentes fonctionnalités du système. L'utilisateur peut ainsi explorer dynamiquement chaque module de l'outil RH.

Au moment de son lancement l'interface Rapport récapitulatif charge les prédictions historiques.

On y trouve :

- Le nombre total d'employés analysés (10895 dans l'exemple), mis en valeur par une icône illustrative et un affichage numérique clair.
- Le taux de rétention, présenté ici à 63,8%, ce qui reflète la part des collaborateurs ayant conservé leur poste sans départ prédictif.
- Un graphique circulaire de répartition des mouvements (Entrée, Interne, Sortie), illustrant la dynamique du personnel au sein de l'organisation.
- Un second graphique donut mettant en opposition la rétention et le turnover, permettant de visualiser rapidement l'équilibre entre stabilité et départs.



**FIGURE 7.3 :** Interface Rapport 1

La figure ci-dessous présente la suite du rapport avec un graphique en barres horizontales qui montre la répartition des mouvements par département. Chaque barre horizontale représente un département, et sa composition montre la proportion relative de chaque type de mouvement. Cela

permet d'identifier visuellement les départements les plus exposés à la fuite de talents ou, à l'inverse, ceux bénéficiant d'une bonne stabilité.

Enfin, un bouton de téléchargement permet d'exporter les données agrégées, facilitant leur exploitation dans un rapport ou une présentation stratégique.

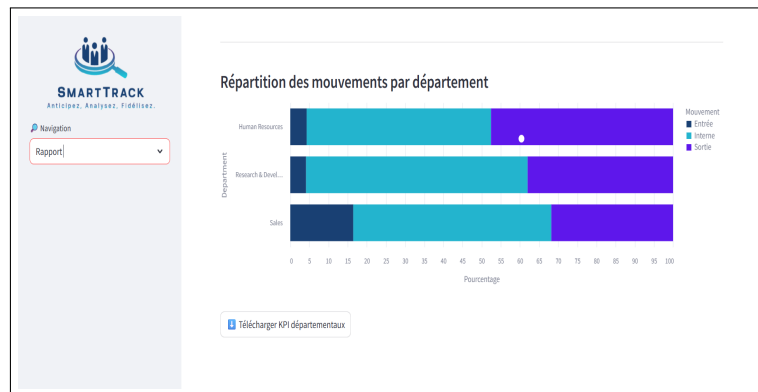


FIGURE 7.4 : Interface Rapport 2

### 7.2.2 Interface de prédiction

L'interface suivante correspond à la section "Prédiction des mouvements" de l'application SmartTrack. Son objectif principal est de permettre à l'utilisateur de charger un fichier CSV contenant les données RH des employés, afin de lancer une analyse prédictive.

L'utilisateur peut soit glisser-déposer un fichier CSV dans la zone prévue à cet effet, soit cliquer sur "Browse files" pour le sélectionner manuellement.



FIGURE 7.5 : Interface prédiction

Cette interface affiche les résultats des prédictions générées après le chargement du fichier de données. Une fois l'analyse terminée, un message de confirmation apparaît en haut de l'écran pour indiquer que les résultats sont disponibles.

L'interface présente un tableau interactif listant les employés avec leurs informations (nom, prénom, département) ainsi que la prédiction du type de mouvement (Entrée, Interne ou Sortie) et la probabilité associée à cette prédiction. L'utilisateur peut également appliquer des filtres dynamiques selon plusieurs critères : nom, prénom, identifiant, département ou type de mouvement.

Cette page permet ainsi une analyse fine et ciblée des prédictions, facilitant l'identification rapide des employés à risque de départ ou de mobilité.

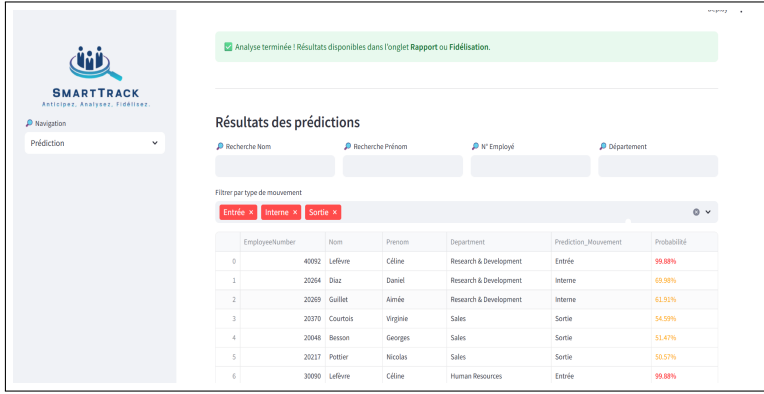


FIGURE 7.6 : Résultat de la prédiction

Dans cette interface, un exemple concret d'utilisation des filtres a été appliqué afin d'illustrer la flexibilité de l'outil. Plus précisément, un filtre sur le type de mouvement a été défini pour ne conserver que les employés prédits comme sortants, combiné à un filtre sur le département "Sales". Cette double sélection permet d'isoler rapidement les cas à risque dans une unité spécifique, facilitant ainsi la prise de décision ciblée et la mise en œuvre de stratégies de fidélisation adaptées à ce contexte.

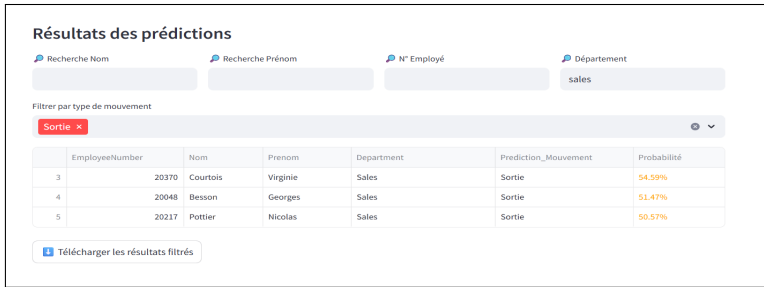


FIGURE 7.7 : Filtre sur les prédictions

Cette interface illustre la mise à jour dynamique du rapport récapitulatif après l'exécution d'une nouvelle analyse prédictive. On observe une augmentation du nombre total d'employés analysés (passant à 10916) ainsi qu'une légère amélioration du taux de rétention (63.9%). Ces ajustements reflètent l'intégration des dernières données dans le tableau de bord, permettant ainsi un suivi en temps réel des évolutions du capital humain. Les visualisations s'actualisent automatiquement, offrant aux décideurs RH une vision consolidée et à jour.

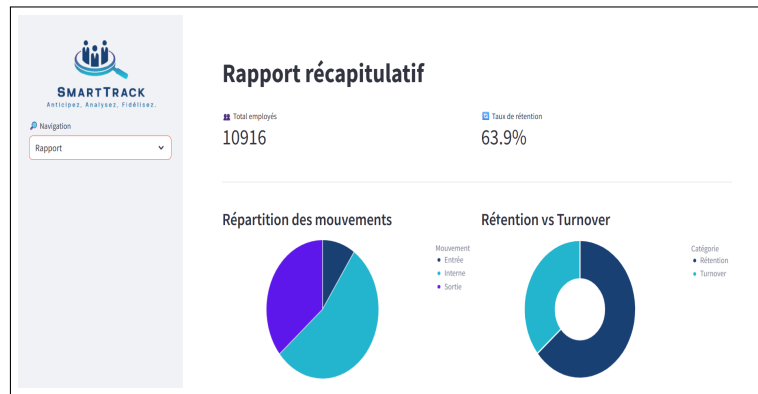


FIGURE 7.8 : Mise à jour du rapport

### 7.2.3 Interface des stratégies de fidélisation

Dans cette interface dédiée à la fidélisation des employés à risque, seule la liste des employés prédits pour une sortie est affichée, ce qui permet de concentrer l'attention sur les cas critiques. L'utilisateur peut sélectionner un collaborateur via le menu déroulant, identifié par son matricule, son nom, et son département. Une fois l'employé choisi, l'interface génère automatiquement la stratégie de rétention personnalisée associée, générée par le LLM à partir des facteurs influents détectés. Ce fonctionnement offre un accès ciblé et rapide aux recommandations RH concrètes pour chaque profil à risque.



FIGURE 7.9 : Interface fidélisation

L'interface illustrée ci-dessus présente un exemple concret de stratégie de fidélisation générée pour un employé identifié comme étant à risque de départ. Après sélection de l'individu (ici Georges Besson du département Sales), une analyse textuelle personnalisée est affichée. Cet affichage synthétique rend les résultats exploitables directement par les responsables RH.



FIGURE 7.10 : Exemple de stratégie

### 7.2.4 Interface du chatbot conversationnel

L'interface Chatbot RH conversationnel joue un rôle central dans la démocratisation de l'accès aux données RH. Elle permet aux utilisateurs, même non techniques, d'interagir avec le système via des questions en langage naturel, et d'obtenir en retour des réponses précises, contextualisées et générées automatiquement à partir des résultats prédictifs et des indicateurs clés. Cette interface contribue à fluidifier la prise de décision RH, tout en offrant un canal interactif et intuitif pour explorer les résultats.

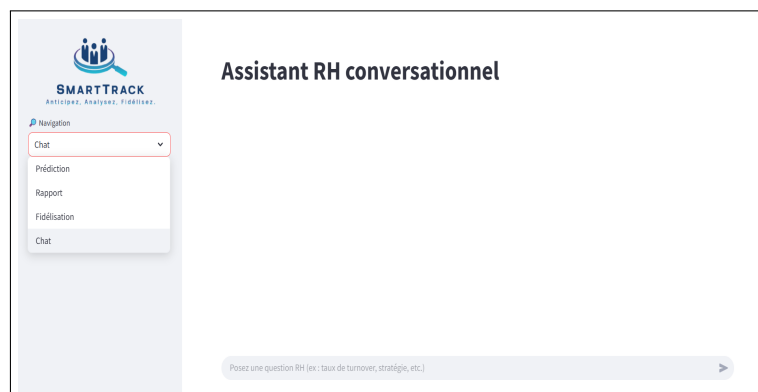
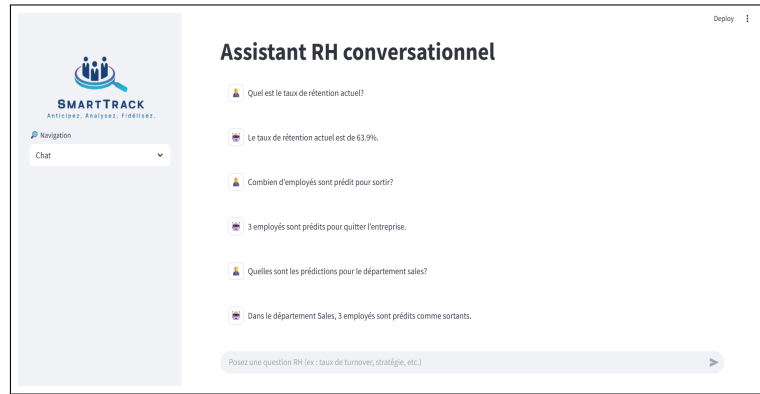


FIGURE 7.11 : Interface chatbot

Dans la capture ci-dessous, on observe une série de questions posées par l'utilisateur à l'assistant RH, comme : « Quel est le taux de rétention actuel ? » ou encore « Combien d'employés sont prédits pour sortir ? ». Le chatbot fournit des réponses claires, chiffrées et directement issues des dernières prédictions. Par exemple, il indique que 3 employés sont prédits comme sortants et que le taux de rétention actuel est de 63.9%. Ce type de fonctionnalité permet aux responsables RH de consulter en temps réel des métriques essentielles sans consulter manuellement les tableaux ou rapports.





**FIGURE 7.12 :** Exemple d'exécution chatbot

## Conclusion

Ce chapitre a démontré la faisabilité opérationnelle de SmartTrack à travers une interface utilisateur complète et intuitive. L'implémentation Streamlit offre aux responsables RH un accès fluide aux quatre fonctionnalités clés. Cette réalisation concrète valide notre approche de bout en bout, de la collecte des données à l'aide à la décision.

# Conclusion générale

Dans le cadre de l'obtention du diplôme d'ingénieur en Big Data et Business Intelligence, ce projet s'inscrit dans un contexte économique où l'économie mondiale traverse une transformation profonde du marché du travail. Cette mutation s'accompagne d'une guerre des talents intensifiée et d'un retard critique de la fonction RH : seulement 31% des entreprises ont modernisé leurs outils, générant un décalage où 68% des décisions RH reposent encore sur l'intuition plutôt que sur l'analyse factuelle.

Face à ces enjeux, ce projet a développé SmartTrack, une solution d'intelligence RH intégrée transformant les données en actions stratégiques. Notre démarche a permis de concevoir une architecture modulaire innovante combinant prédiction multiclasse (68.9% de précision avec XGBoost), génération de stratégies personnalisées via Mistral, visualisation dynamique des KPIs et assistance conversationnelle. L'originalité réside dans l'enrichissement des données via LLaMA 3 et CTGAN, transformant un dataset limité en corpus représentatif, et dans l'évaluation bidimensionnelle alliant métriques quantitatives et grilles qualitatives.

Au-delà des aspects techniques, ce projet démontre que la transformation digitale des RH peut véritablement révolutionner la prise de décision stratégique. En réconciliant données quantitatives et recommandations qualitatives, SmartTrack illustre le potentiel de l'IA générative pour humaniser la technologie au service du capital humain, préfigurant l'avenir des ressources humaines : data-driven, prédictive et profondément centrée sur l'expérience collaborateur.

Ce travail ouvre des perspectives prometteuses avec l'ajout de nouvelles fonctionnalités comme le matching intelligent des profils pour optimiser les affectations internes, le recrutement prédictif pour identifier les candidats les plus adaptés, ou encore la recommandation automatique de formations personnalisées.

