# **Cahier des Charges – Version 2**

# **Projet de Migration et Valorisation de Données Assistée par Intelligence Artificielle (EasyStrat)**

**Date de révision:** 4 juin 2025

## **1. Contexte du Projet**

### **1.1 Présentation de l'entreprise**

Easystrat est une entreprise spécialisée dans le développement de logiciels et la centralisation de données, visant à accompagner les organisations dans leur transformation numérique et l'optimisation de leurs systèmes d'information.

Secteur d'activité: IT, Développement de logiciels, Conseil en transformation digitale.

### **1.2 Contexte Général du Projet : La Problématique des Systèmes Hérités**

De nombreuses PME et grandes entreprises sont aujourd'hui confrontées à des systèmes d'information hérités (legacy). Ces systèmes, souvent développés avec des technologies obsolètes, non maintenus par les éditeurs, ou reposant sur des architectures fermées, entravent considérablement l'agilité, l'innovation et la transformation numérique. La dépendance à des bases de données inaccessibles, des formats de données propriétaires, ou des API limitées et complexes complique drastiquement les projets de modernisation et de valorisation des données. La multiplication de logiciels spécialisés non intégrés génère des silos d'information, des redondances, et des risques de perte ou d'incohérence des données.

### **1.3 Enjeux Stratégiques**

Les enjeux liés à ces systèmes legacy sont multiples et critiques pour les entreprises :

* **Coûts de maintenance élevés:** Les compétences techniques nécessaires sont rares et coûteuses, et l'infrastructure vieillissante engendre des frais importants.
* **Risques de sécurité accrus:** Les systèmes obsolètes présentent souvent des vulnérabilités de sécurité non corrigées.
* **Blocage de l'innovation:** L'incapacité à intégrer de nouvelles technologies (Cloud, IA, Big Data) freine la compétitivité. L'accès limité ou inexistant à la documentation des bases de données, ou les restrictions imposées par les éditeurs, empêchent l'exploitation des données.
* **Qualité des données dégradée:** La dispersion des données, les formats hétérogènes et les difficultés d'extraction et de consolidation nuisent à la fiabilité des informations et donc à la prise de décision.
* **Perte d'agilité métier:** Les processus rigides des anciens systèmes ne permettent pas de s'adapter rapidement aux évolutions du marché.

Il est donc urgent de mettre en place une stratégie pour moderniser ces systèmes, centraliser les données, en assurer la gouvernance et faciliter la migration vers des architectures modernes et évolutives. Cela permettra aux entreprises de tirer pleinement parti des avantages de la digitalisation. De plus, les logiciels développés aujourd'hui seront les systèmes legacy de demain ; une approche pérenne est donc essentielle.

### **1.4 Le Défi Particulier de l'Accès aux Données des Systèmes Sources**

Un obstacle majeur réside dans les difficultés d'accès aux données des systèmes existants :

* **Restrictions contractuelles ou techniques** des éditeurs limitant l'accès direct aux bases de données.
* **API inexistantes, incomplètes, mal documentées ou très restrictives** (ex: limitation du nombre d'enregistrements par requête, complexité des méthodes d'authentification et de requêtage).
* **Hétérogénéité des formats et des structures de données** rendant l'automatisation de l'extraction complexe.
* **Déploiement dans des environnements réseau fermés** (intranets, sous-réseaux isolés) empêchant la communication avec des systèmes externes.

## **2. Vision et Objectifs du Projet EasyStrat**

### **2.1 Vision à Long Terme**

La vision d'EasyStrat est de devenir un acteur de référence dans la modernisation des systèmes d'information en proposant une plateforme innovante qui simplifie et accélère la migration, la centralisation, et la valorisation des données grâce à une utilisation avancée de l'Intelligence Artificielle. Nous visons à développer une expertise "deep tech" unique dans l'interopérabilité des données, permettant de "déverrouiller" le potentiel des systèmes hérités.

### **2.2 Objectifs Généraux du Projet**

1. **Développer la plateforme EasyStrat:** Une solution modulaire pour l'extraction, la transformation, la centralisation (dans un entrepôt de données moderne et éco-efficient) et la gouvernance des données issues de systèmes hétérogènes.
2. **Mettre en œuvre une stratégie IA en deux phases distinctes et complémentaires:**
   * **Phase 1 (Court/Moyen Terme):** Fournir une valeur ajoutée immédiate aux utilisateurs via un assistant IA (basé sur RAG) interagissant avec les données centralisées, et collecter en parallèle les informations nécessaires pour la phase 2.
   * **Phase 2 (Moyen/Long Terme - R&D Deep Tech):** Développer une IA avancée pour l'interopérabilité des données, capable d'analyser, de comprendre et d'automatiser la migration des schémas et des données de systèmes legacy complexes.

### **2.3 Objectifs Spécifiques – Phase 1 : Assistant IA (RAG) & Centralisation Initiale**

* **Extraction et Centralisation:** Mettre en place les mécanismes d'extraction (connecteurs, adaptateurs API) pour récupérer les données des systèmes sources identifiés (ex: iCar pour Willermin) et les centraliser dans l'entrepôt de données EasyStrat.
* **Entrepôt de Données Éco-Friendly:** Concevoir et implémenter un data warehouse optimisé en termes de performance et d'efficience énergétique.
* **Assistant IA (RAG):**
  + Développer un assistant conversationnel permettant aux utilisateurs finaux (ex: commerciaux, RH) d'interroger en langage naturel les données centralisées.
  + Fournir des réponses pertinentes, des synthèses, et des recommandations basées sur les données disponibles.
* **Collecte de Données pour la Phase 2:** Durant les missions de centralisation et de mise en place du RAG, collecter et structurer des informations sur les architectures, les schémas de données, les API des systèmes sources rencontrés. Ces informations constitueront le corpus d'entraînement pour l'IA d'interopérabilité.

### **2.4 Objectifs Spécifiques – Phase 2 : IA pour l'Interopérabilité et la Migration Avancée (R&D)**

* **Analyse et Compréhension de Schémas:** Développer des modèles d'IA (probablement basés sur du Machine Learning/Deep Learning) capables d'analyser la structure (schémas, relations, types de données) des bases de données sources, même avec une documentation limitée.
* **Mapping et Transformation Automatisés:** Créer une IA capable de proposer et d'automatiser le mapping des données sources vers le schéma cible de l'entrepôt EasyStrat, et de gérer les transformations nécessaires.
* **Facilitation de la Migration:** Réduire significativement le temps, le coût et la complexité des projets de migration de données grâce à cette automatisation intelligente.
* **Axe de R&D "Deep Tech":** Positionner cette phase comme le cœur de l'innovation d'EasyStrat, justifiant des efforts de recherche et développement conséquents et des collaborations avec des laboratoires spécialisés (ex: Mines d'Alès, Lirmes).

### **2.5 Impact Attendu sur le Business et les Opérations**

* **Réduction des coûts et des délais** de migration de données.
* **Amélioration de la qualité et de la gouvernance** des données centralisées.
* **Prise de décision éclairée** grâce à un accès facilité à l'information et aux analyses de l'assistant IA.
* **Agilité métier accrue** par la modernisation des systèmes et la flexibilité de la plateforme EasyStrat.
* **Ouverture à l'innovation** en facilitant l'intégration de nouvelles technologies sur la base de données centralisées.
* **Réduction de l'empreinte environnementale** grâce à un stockage et un traitement optimisés des données.

## **3. Cas d'Usage Cibles**

### **3.1 Assistant Commercial pour Concessionnaire Automobile (Ex: Willer**min**)**

* **Description:** Un vendeur utilise l'assistant IA EasyStrat (RAG) pour obtenir rapidement des informations sur les véhicules en stock, les options disponibles, les fiches techniques, l'historique d'un véhicule d'occasion, ou pour configurer un véhicule selon les désirs d'un client.
* **Données sources:** Base de données iCar (ventes, véhicules, interventions), catalogues constructeurs.
* **Bénéfices:** Gain de temps, pertinence des informations, amélioration de l'expérience client, aide à la vente.

### **3.2 Assistant RH pour la Gestion Quotidienne**

* **Description:** Un responsable RH interroge l'assistant IA pour connaître le solde de congés d'un employé, obtenir des informations sur la législation du travail (potentiellement enrichi par une veille web sécurisée), ou identifier des profils internes pour une nouvelle mission.
* **Données sources:** SIRH, base de données paie, documents légaux internes, CVthèque.
* **Bénéfices:** Accès rapide à l'information RH, aide à la conformité, optimisation de la gestion des talents.

### **3.3 Migration de Données de Systèmes Propriétaires Complexes**

* **Description:** Une entreprise souhaite migrer les données d'un ancien CRM propriétaire vers la plateforme EasyStrat. L'IA d'interopérabilité (Phase 2) analyse le schéma du CRM, propose un mapping vers l'entrepôt EasyStrat, et automatise une grande partie du processus de migration et de transformation des données.
* **Données sources:** Base de données du CRM propriétaire.
* **Bénéfices:** Migration accélérée, réduction des erreurs manuelles, préservation de l'intégrité des données.

### **3.4 Classification Automatique et Gouvernance des Données**

* **Description:** Lors de la centralisation, l'IA d'EasyStrat analyse le contenu des données pour les classifier automatiquement (ex: données personnelles, sensibles, confidentielles) et aider à la mise en place de règles de gouvernance et d'accès.
* **Données sources:** Toutes données ingérées dans EasyStrat.
* **Bénéfices:** Conformité RGPD facilitée, meilleure gestion des risques, sécurité des données renforcée.

## **4. Architecture et Spécifications de la Solution EasyStrat**

### **4.1 Architecture Générale**

La plateforme EasyStrat sera conçue selon une architecture modulaire, potentiellement microservices, pour assurer flexibilité, scalabilité et maintenabilité. Elle comprendra :

* **Module d'Ingestion & Connecteurs:** Pour l'accès aux diverses sources de données.
* **Module de Traitement & Transformation (ETL/ELT):** Intégrant les capacités de l'IA d'interopérabilité (Phase 2).
* **Entrepôt de Données Centralisé (Data Warehouse):** Optimisé et sécurisé.
* **Module IA - Assistant RAG:** Moteur de recherche sémantique et LLM.
* **Module IA - Interopérabilité:** Modèles de ML/DL pour l'analyse de schémas.
* **Module de Gouvernance des Données:** Catalogue de données, gestion des accès, lignage.
* **Couche d'API:** Pour l'intégration avec des applications tierces et les interfaces utilisateur.
* **Interfaces Utilisateur:** Tableau de bord d'administration, interface de l'assistant IA, outils de visualisation.

*(Le schéma fonctionnel de la page 4 du document original de Baurice reste une bonne base de départ pour illustrer ce flux).*

### **4.2 Module d'Ingestion & Connecteurs**

* **Connecteurs standards:** Pour bases de données relationnelles (PostgreSQL, MySQL, Oracle, SQL Server), NoSQL (MongoDB), fichiers plats (CSV, XLSX, JSON, XML), services cloud.
* **Développement de connecteurs spécifiques / Adaptateurs API:** Pour les systèmes legacy ou propriétaires lorsque les API sont la seule voie d'accès. Ce développement pourra nécessiter une expertise ponctuelle.
* **Capacités de découverte de schémas préliminaires.**

### **4.3 Module de Traitement & Transformation (ETL/ELT)**

* **Fonctions classiques:** Nettoyage, dédoublonnage, validation, enrichissement, formatage.
* **Intégration de l'IA d'Interopérabilité (Phase 2):**
  + Analyse des schémas sources.
  + Proposition de mapping de données intelligent.
  + Automatisation des transformations complexes.
  + Détection d'anomalies et d'incohérences.
* **Gestion des Data Catalogs et suppression des doublons.**

### **4.4 Entrepôt de Données Centralisé**

* **Technologies cibles:** Évaluation de solutions comme PostgreSQL, Snowflake, Redshift, CockroachDB, en tenant compte des performances, de la scalabilité, des coûts et de l'efficience énergétique ("éco-friendly").
* **Optimisation:** Partitionnement, indexation, compression.
* **Sécurité:** Cryptage des données au repos et en transit.

### **4.5 Modules d'Intelligence Artificielle**

#### **4.5.1 Assistant IA (RAG - Phase 1)**

* **Architecture:**
  + LLM (Large Language Model): Sélection d'un modèle adapté (ex: Mistral pour des raisons de souveraineté et performance, ou autres modèles open source/propriétaires selon budget et besoins).
  + Base de données vectorielle: Pour stocker les embeddings des données centralisées et des documents de contexte.
  + Mécanisme de "Retrieval": Algorithmes pour retrouver les informations les plus pertinentes à injecter dans le prompt du LLM.
* **Sources de données pour le RAG:** Données de l'entrepôt EasyStrat, documentation technique, bases de connaissance métier.
* **Hébergement du LLM:**
  + **Cloud:** Plus simple à mettre en œuvre initialement, mais pose des questions de coût à l'usage et de confidentialité des données.
  + **On-premise / Cloud Privé:** Plus complexe et coûteux à mettre en place (GPU), mais offre un meilleur contrôle sur la sécurité et la confidentialité. Cette option sera privilégiée pour les clients manipulant des données très sensibles.
* **Gestion des droits d'accès:** Le RAG devra respecter les habilitations des utilisateurs pour ne fournir que les informations auxquelles ils ont droit.

#### **4.5.2 IA pour l'Interopérabilité et la Migration (R&D - Phase 2)**

* **Approche d'apprentissage:**
  + Apprentissage supervisé sur des paires (schéma source, schéma cible mappé) si des données d'entraînement peuvent être constituées.
  + Apprentissage non supervisé ou auto-supervisé pour la découverte de structures et de sémantique dans les schémas sources.
  + Utilisation de techniques de NLP pour analyser la documentation des API ou les noms de colonnes/tables.
* **Données d'entraînement:** Schémas de bases de données réelles, documentations d'API, mappings existants, collectés lors des projets de Phase 1.
* **Objectifs de performance:** Taux d'automatisation de la reconnaissance de schémas, précision du mapping, réduction du temps d'intervention humaine.

### **4.6 Interfaces Utilisateur (UX/UI)**

* **Tableau de bord d'administration EasyStrat:** Pour configurer les sources de données, superviser les migrations, gérer les utilisateurs et la sécurité.
* **Interface de l'assistant IA:** Intuitive (type chatbot), potentiellement intégrable dans les outils existants des clients.
* **Outils de Data Discovery et de visualisation** pour les Data Analysts et Business Users.
* L'expérience utilisateur sera une priorité pour faciliter l'adoption.

## **5. Spécifications Techniques Détaillées**

### **5.1 Technologies Envisagées**

* **Backend:** Python (Flask/Django), Java (Spring Boot), Node.js. Python est privilégié pour son écosystème IA/ML.
* **Frontend:** React, Angular, Vue.js.
* **Bases de données (pour EasyStrat):** PostgreSQL, MongoDB.
* **Frameworks IA/ML:** TensorFlow, PyTorch, scikit-learn, Hugging Face Transformers.
* **Orchestration & Conteneurisation:** Docker, Kubernetes.
* **Cloud Providers (pour hébergement optionnel):** AWS, Azure, GCP, fournisseurs européens.

### **5.2 Gestion des Données (Sources et Entrepôt)**

* **Sources de données:** Internes (bases de données, fichiers), externes (API, open data).
* **Type et volume de données:** De quelques Go à plusieurs To (ex: 1Go à 67TB mentionné). La solution doit être scalable.
* **Stratégie de nettoyage et d'annotation des données:** Processus semi-automatisé avec validation humaine. L'annotation sera cruciale pour l'entraînement de l'IA d'interopérabilité.

### **5.3 Performance et Scalabilité**

* **Temps de réponse maximal attendu pour l'assistant RAG:** Quelques secondes pour une interaction typique (objectif < 5s, idéalement < 2s).
* **Précision et robustesse des modèles IA:** Objectif > 90-95% pour les tâches de classification ou de mapping critiques, avec des mécanismes de détection d'incertitude.
* **Exigences de scalabilité:** La plateforme doit pouvoir gérer un nombre croissant d'utilisateurs, de sources de données et de volumes de données. Les modèles IA devront être monitorés pour le "concept drift" et le "data drift" et ré-entraînés si nécessaire.

## **6. Sécurité, Confidentialité et Gouvernance**

### **6.1 Protection des Données et Conformité**

* **Conformité RGPD:** La plateforme devra être conçue en respectant les principes du RGPD dès la conception ("privacy by design").
* **Cryptage:** Données cryptées au repos (AES-256 ou équivalent) et en transit (TLS 1.2+).
* **Anonymisation/Pseudonymisation:** Fonctionnalités pour anonymiser ou pseudonymiser les données lorsque nécessaire, notamment pour les environnements de test ou d'analyse.

### **6.2 Accès et Authentification**

* **Niveaux d'accès granulaires (C1 à C5):** Définition de rôles et de permissions précis pour contrôler l'accès aux données et aux fonctionnalités.
* **Méthodes d'authentification robustes:** Authentification multi-facteurs (MFA/2FA), intégration SSO (Single Sign-On) avec les annuaires d'entreprise (LDAP, Azure AD).
* **Gestion des privilèges pour l'IA:** L'assistant IA devra opérer sous le contexte de l'utilisateur connecté, n'accédant qu'aux données autorisées pour cet utilisateur.

### **6.3 Hébergement des LLM et Données Sensibles**

* Une analyse de risque sera menée pour chaque client afin de déterminer la meilleure stratégie d'hébergement (cloud public, cloud privé, on-premise) en fonction de la sensibilité des données et des exigences réglementaires.
* Pour les données très sensibles, des solutions d'hébergement on-premise ou dans des clouds de confiance certifiés (ex: SecNumCloud en France) seront privilégiées.

### **6.4 Gouvernance des Données Centralisées**

* **Data Catalog:** Catalogue de données intégré pour documenter les métadonnées, la qualité, l'origine et l'utilisation des données.
* **Data Lineage:** Traçabilité des données de la source à la cible et à travers les transformations.
* **Gestion de la qualité des données:** Outils de monitoring et d'alerte.

### **6.5 Sécurité de l'IA "Connectée au Web" (pour veille)**

* Si une fonctionnalité de veille (ex: juridique pour RH) nécessite un accès web, l'IA ou l'agent responsable sera isolé dans un environnement sécurisé (DMZ ou équivalent) pour minimiser les risques. Les flux de données entre cette IA et le reste du système seront strictement contrôlés.

## **7. Stratégie de Recherche et Développement (Deep Tech)**

### **7.1 Axe Principal de R&D**

L'innovation majeure et l'effort principal de R&D se concentreront sur le développement de **l'IA pour l'interopérabilité et l'automatisation de la migration des schémas de données (Phase 2)**. Cela inclut :

* La recherche sur les algorithmes de reconnaissance et de compréhension de structures de données complexes.
* Le développement de techniques de mapping sémantique automatisé.
* La création de modèles capables d'apprendre à partir d'exemples limités et hétérogènes.

### **7.2 Collaboration Scientifique**

Des partenariats stratégiques seront établis avec des laboratoires de recherche spécialisés en IA et en gestion de données (ex: Mines d'Alès, Lirmes) pour :

* Bénéficier d'une expertise de pointe.
* Mener des projets de recherche conjoints.
* Accéder à de jeunes talents (doctorants, stagiaires).

### **7.3 Financement de la R&D**

Une stratégie active de recherche de financements sera mise en place, ciblant :

* Les dispositifs nationaux (Crédit Impôt Recherche - CIR, aides Bpifrance).
* Les appels à projets régionaux et nationaux.
* Les programmes de financement européens (ex: Horizon Europe) axés sur l'IA, la deep tech et la souveraineté numérique.

### **7.4 Propriété Intellectuelle**

Les innovations développées, notamment les algorithmes et modèles d'IA spécifiques, feront l'objet d'une stratégie de protection de la propriété intellectuelle (brevets, droits d'auteur logiciels) pour valoriser les actifs de l'entreprise.

## **8. Limites du Projet et Gestion des Risques**

### **8.1 Limites Techniques et Opérationnelles**

* **Dépendance aux API des systèmes sources:** La performance et les capacités d'extraction seront limitées par la qualité, la documentation et les restrictions des API fournies par les éditeurs tiers. Des interventions manuelles ou des développements spécifiques de connecteurs pourront être nécessaires.
* **Accès direct aux bases de données:** De nombreuses applications, notamment propriétaires, ne permettent pas un accès direct à leur base de données. L'IA d'EasyStrat, pour fonctionner de manière optimale dans la compréhension des schémas, préfère un accès direct. Les limitations d'accès impacteront le niveau d'automatisation possible.
* **Environnements réseau isolés:** Le déploiement dans des intranets ou sous-réseaux sans accès externe peut compliquer la communication avec la plateforme EasyStrat (si hébergée en cloud) ou la collecte de données pour l'IA.
* **Complexité intrinsèque de certains systèmes:** Certains systèmes legacy peuvent être si complexes ou si mal documentés que l'automatisation complète de la migration restera un défi, nécessitant une expertise humaine significative.
* **L'IA ne remplacera pas toute intervention humaine:** Elle vise à assister et accélérer, mais une validation et une supervision humaines resteront nécessaires, notamment pour les cas complexes ou les décisions critiques.

### **8.2 Risques du Projet**

* **Difficulté à obtenir des données d'entraînement qualifiées et variées** pour l'IA d'interopérabilité (Phase 2).
* **Complexité et durée du développement des modèles IA "from scratch"** pour l'interopérabilité.
* **Adoption par les utilisateurs:** Nécessité d'une UX soignée et d'une démonstration claire de la valeur ajoutée.
* **Évolution rapide des technologies IA:** Nécessité d'une veille technologique constante et d'une architecture flexible pour intégrer les nouvelles avancées.
* **Sécurité et confidentialité des données:** Risques liés à la manipulation de données sensibles ; une gouvernance stricte est impérative.
* **Dépendance aux financements R&D** pour la réalisation complète de la Phase 2.

### **8.3 Stratégies d'Atténuation des Risques**

* **Approche phasée:** La Phase 1 (RAG) permet de délivrer de la valeur rapidement et de collecter des données pour la Phase 2, réduisant le risque lié à cette dernière.
* **Preuves de Concept (PoC):** Réaliser des PoC sur des périmètres limités pour valider les approches techniques.
* **Veille technologique active et partenariats.**
* **Priorisation de la sécurité et de la conformité** dès la conception.
* **Communication transparente** avec les clients sur les capacités et les limites de la solution.
* **Utilisation initiale de modèles open source ou sur étagère pour le RAG** en attendant le développement de modèles propriétaires ou la sécurisation de budgets pour des modèles plus avancés.

## **9. Livrables du Projet (Principaux)**

### **Phase 1:**

* Plateforme EasyStrat v1 avec modules d'ingestion, de transformation basique, et entrepôt de données fonctionnel.
* Assistant IA (RAG) v1 opérationnel sur un ou plusieurs cas d'usage clients.
* Documentation technique et utilisateur pour la v1.
* Premiers jeux de données (anonymisés) sur les architectures et schémas sources collectés.

### **Phase 2 (R&D):**

* Prototypes et preuves de concept de l'IA d'interopérabilité.
* Modèles IA entraînés pour l'analyse de schémas et le mapping de données.
* Intégration progressive des capacités de l'IA d'interopérabilité dans la plateforme EasyStrat.
* Publications scientifiques et rapports de R&D (si pertinents et en accord avec la stratégie de PI).
* Plateforme EasyStrat v2 avec fonctionnalités avancées de migration assistée par IA.

## **10. Planning Prévisionnel et Budget**

*(Cette section sera développée ultérieurement en collaboration avec les équipes techniques et commerciales. Elle détaillera les jalons, les sprints, les ressources nécessaires et les estimations de coûts pour chaque phase).*

## **11. Équipe Projet et Gouvernance du Projet**

*(Cette section sera développée ultérieurement. Elle identifiera les rôles clés, les responsabilités, et les instances de pilotage du projet).*