50 milliards d'euros, une deadline, zéro régression : migrer un service de calcul critique vers R avec l'IA

Vincent Guyader[1] Matthieu de Canteloube[2]

Résumé (max 300 mots)

La migration de systèmes statistiques legacy vers R représente un défi croissant pour les institutions publiques confrontées à la fin de vie de logiciels propriétaires. L'ATIH (Agence Technique de l'Information sur l'Hospitalisation) fait face à ce défi dans un contexte critique : son service de calcul, qui orchestre la distribution de 50 milliards d'euros de l'ONDAM hospitalière aux établissements de santé français, repose sur une soixantaine de procédures legacy à migrer vers R d'ici fin 2026.

Theodo et ThinkR ont co-conçu une méthode de migration en trois étapes, adossée à deux outils open source développés pour l'occasion. DATALineage analyse statiquement le code legacy pour cartographier les flux de données et extraire les règles métier. Cette représentation structurée est transmise à Claude Code, qui réécrit chaque procédure en R dans un cadre contraint : architecture cible (conteneurs Docker sur Kubernetes), conventions de code maintenables, et boucle de rétroaction automatique pilotée par {datadiff}. Ce package R open source, surcouche à {pointblank} permet la comparaison des sorties de chaque procédure originale et de son équivalent R, offrant à l'IA un signal précis pour s'auto-corriger. La même validation est ensuite appliquée à l'échelle de l'ensemble des établissements de santé français pour garantir la non-régression en production.

Nous présenterons la méthode complète, les retours d'expérience des premières migrations, et les métriques obtenues sur les données réelles du parc hospitalier français.

Mots-clefs : migration, IA générative, non-régression, legacy

Développement

Contexte et enjeux

L'ATIH gère le PMSI (Programme de Médicalisation des Systèmes d'Information), pilier du financement hospitalier français depuis la mise en place de la T2A.

En 2022, l'agence décide de migrer l'intégralité de son service de calcul.

Le code à migrer est un corpus de procédures statistiques complexes, accumulées sur plusieurs décennies, mêlant logique métier, macros imbriquées et transformations de données non documentées.

[1]ThinkR, vincent@thinkr.fr
[2]Theodo, matthieu.de-canteloube@theodo.com

Theodo apporte son expérience des migrations assistées par l'IA ; ThinkR apporte l'expertise R et la méthodologie de validation. Ensemble, nous avons conçu un pipeline reproductible en trois étapes articulées autour de deux outils open source.

Étape 1 - DATALineage : cartographier l'existant

Avant toute réécriture, il faut comprendre ce que fait réellement le code. DATALineage est un outil d'analyse statique développé sur-mesure pour décrypter des bases de code legacy volumineuses et peu documentées. Il extrait les flux de données (tables en entrée et en sortie, transformations intermédiaires) ainsi que les règles métier portées par les macros et les conditions. Le résultat est une cartographie structurée, lisible à la fois par un humain et par un modèle de langage, qui constitue le matériau de base transmis à l'étape suivante.

Étape 2 - Claude Code : réécriture contrainte

Claude Code traduit chaque procédure en R à partir du lineage extrait. Le modèle n'opère pas en mode libre : il reçoit un cahier des charges précis incluant les contraintes d'infrastructure (conteneurs Docker, déploiement Kubernetes), les conventions de code attendues par l'ATIH pour garantir la maintenance du code en autonomie, et les règles de format des sorties. La clé de la méthode est la boucle d'auto-correction : {datadiff} fournit à Claude Code un feedback quantitatif et précis sur chaque tentative, lui permettant d'itérer jusqu'à convergence sans intervention humaine.

Étape 3 - datadiff : non-régression à deux niveaux

{datadiff} est un package R open source qui compare deux jeux de données via des règles YAML configurables : tolérance numérique, normalisation de texte, validation du nombre de lignes, gestion des valeurs manquantes. Pendant le développement, il sert de boucle de rétroaction pour l'IA. En production, la même batterie de comparaisons est appliquée sur les données réelles de l'ensemble des établissements de santé français, garantissant que la procédure R reproduit exactement, à la tolérance définie, le comportement de l'original.

Résultats et perspectives

À la date de soumission, les procédures sont en cours de migration. Les métriques de non-régression sur données réelles sont encourageantes. Les Rencontres R seront l'occasion de présenter les résultats complets du déploiement à l'échelle, notamment le gain de productivité observé par rapport à une migration manuelle, ainsi que les cas limites rencontrés (règles métier ambiguës, comportements non documentés) et la manière dont la méthode les traite.

Les deux outils présentés, DATALineage et {datadiff}, sont open source et réutilisables dans tout contexte de migration de code statistique legacy, quelle que soit l'institution concernée.