Agrégation de données et BDD

Projet

Donatien Mottin - dmottin@ece.fr

Contrat de confiance sur l'utilisation de l'IA générative pour ce travail

Niveau d'assistance 2/4

L'IA générative a été utilisée uniquement pour des suggestions ou des corrections mineures, ainsi que pour l'amélioration d'éléments.

Projet

Ce projet peut vous servir de base pour le bloc 1 (compétences C1 à C5) de votre titre RNCP DIA.

Objectif : Mettre en place une chaîne de collecte, nettoyage/harmonisation, et agrégation, puis extraction de données. Sont attendus :

- Étape 1 : Une présentation d'une problématique professionnelle fictive
- Étape 2 : Des scripts de collecte qui récupère de la donnée d'au moins trois sources d'au moins deux types de sources différentes (WebScrapping, Fichier OpenData, API)
- Étape 3 : Des scripts de nettoyages et d'harmonisation des données
- Étape 4 : Un script de conception d'une BDD et d'agrégation des données au sein de cette BDD
- Étape 5 : Des scripts de conception de vue et d'extraction de fichiers plats en vue d'une exploitation des données. Au moins deux vues (et fichiers plats associés) sont attendues.

Le choix de la problématique et des sources de données utilisées sont totalement libres tant qu'elle s'inscrivent dans une thématique professionnelle présentable devant un jury.

L'évaluation se fera via :

- une présentation orale en groupe de 2 à 3
- une mise à disposition de toutes vos sources commentées, sur un Git : ajouter dmottin-ece avec des droits en lecture sur votre projet Git, deux jours avant la date de l'oral. La branche main sera relue. Aucun commit entre cette date et la date de l'oral.

La notation sera effectuée en suivant la grille d'évaluation du titre RNCP compétence C1 à C4 (détaillée ci-après). Les items liés au traitement de la RGPD (dans la compétence C4) ne seront pas pris en compte.

Tout comme le jour du titre RNCP, il est de votre responsabilité de mettre en avant votre acquisition des compétences.

Référentiel RNCP DIA : Compétences C1 à C4

C1 : Automatiser l'extraction de données

Ce bloc détaille les critères de la présentation du projet, et de la collecte (étapes 1 et 2)

1. La présentation du projet et de son contexte est complète : acteurs, objectifs fonctionnels et techniques, environnements et contraintes techniques, budget, organisation du travail et planification.
2. Les spécifications techniques précisent : les technologies et outils, les services externes, les exigences de programmation (langages), l'accessibilité (disponibilité, accès).
3. Le périmètre des spécifications techniques est complet : il couvre l'ensemble des moyens techniques à mettre en oeuvre pour l'extraction et l'agrégation des données en un jeu de données brutes final.
4. Le script d'extraction des données est fonctionnel : toutes les données visées sont effectivement récupérées à l'issue de l'exécution du script.
5. Le script comprend un point de lancement, l'initialisation des dépendances et des connexions externes, les règles logiques de traitement, la gestion des erreurs et des exceptions, la fin du traitement et la sauvegarde des résultats.
6. Le script d'extraction des données est versionné et accessible depuis un dépôt Git.
7. L'extraction des données est faite depuis un mix entre au moins les sources suivantes : un service web (API REST), un fichier de données, un scraping, une base de données et un système big data.

C2 : Développer des requêtes de type SQL d'extraction

Ce bloc détaille les critères de l'étape d'extraction (étape 5)

8. Les requêtes de type SQL pour la collecte de données sont fonctionnelles : les données visées sont effectivement extraites suite à l'exécution des requêtes
9. La documentation des requêtes met en lumière choix de sélections, filtrages, conditions, jointures, etc., en fonction des objectifs de collecte.
10. La documentation explicite les optimisations appliquées aux requêtes .

C3 : Développer des règles d'agrégation de données

Ce bloc détaille les critères de l'étape de nettoyage et d'agrégation (étapes 3 et une partie de la 4)

11. Le script d'agrégation des données est fonctionnel : les données sont effectivement agrégées, nettoyées et normalisées en un seul jeu de données à l'issue de l'exécution du script.
12. Le script d'agrégation des données est versionné et accessible depuis un dépôt Git.
13. La documentation du script d'agrégation est complète : dépendances, commandes, les enchaînements logiques de l'algorithme, les choix de nettoyage et d'homogénéisation des formats données.

C4 : Créer une base de données dans le respect du RGPD

Ce bloc détaille les critères de l'étape de conception de la BDD (étape 4) et du respect de la RGPD (non détaillés ici).

14. La base de données est choisie au regard de la modélisation des données et des contraintes du projet.
15. La reproduction des procédures d'installation décrites (base de données et API) a pour résultat un système conforme aux objets techniques attendus.
16. Le script d'import fourni est fonctionnel : il permet l'insertion des données dans le système mis en place.
17. La documentation technique du script d'import est versionné à la racine du même dépôt Git que celui utilisé pour le script d'import.
18. Les documentations techniques des scripts couvrent les dépendances nécessaires pour la réutilisation des scripts (langages, dépendances externes, etc) et les commandes pour l'exécution des scripts.