epsi®
l'école d'ingénierie informatique

Pro Alterna®

CERTIFICATION DÉVELOPPEUR EN INTELLIGENCE ARTIFICIELLE ET DATA SCIENCE RNCP 36581

BLOC DE COMPÉTENCES 1 : Créer un modèle de données d'une solution I.A en utilisant des méthodes de Data sciences

Cahier des Charges de la MSPR : Développement et déploiement d'une application dans le respect du cahier des charges Client // Création d'un backEnd métier permettant le nettoyage et la visualisation des données.

COMPÉTENCES ÉVALUÉES :

- Définir les sources et les outils nécessaires pour permettre de collecter les données
- Recueillir de manière sécurisée les informations à partir de sources adaptées (sources hétérogènes, internes fournies par le client ou externes accessibles en Open Data) permettant de définir les données à collecter pour réaliser l'architecture de données
- Paramétrer les outils afin d'importer les données de manière automatisée et sécurisée
- Analyser, nettoyer, trier et s'assurer de la qualité des données afin de les rendre exploitables pour la solution I.A, en utilisant des outils d'analyse et de visualisation des données et se basant sur des approches de la Data science.
- Construire la structure de stockage des données (modèle de données) qui répond au mieux au besoin d'analyse.
- Représenter graphiquement les relations entre les données afin de les visualiser en créant des tableaux de bord accessibles à tout public garantissant ainsi l'accessibilité numérique.
- Exploiter de manière automatisée et analyser les informations recueillies dans les structures de stockage des données (requêtage ou interrogation) afin de répondre aux exigences de la solution IA définit dans le cahier des charges.

PHASE 1 : PRÉPARATION DE CETTE MISE EN SITUATION PROFESSIONNELLE RECONSTITUÉE

- Durée de préparation :
  - 19 *heures*
- Mise en œuvre :
  - *Travail d'équipe constituée de 4 apprenants-candidats (5 maximum si groupe impair)*
- Résultat attendu :
  - *Préparation du jeu de données d'entraînement-validation-test*
  - *Modélisation et création des bases de données*
  - *Création du script/requête de récupération des infos complémentaires d'une espèce*
  - *Création du premier script/requête d'écriture des données recueillies*

PHASE 2 : PRÉSENTATION ORALE COLLECTIVE + ENTRETIEN COLLECTIF

- **Durée totale par groupe** : 50 mn se décomposant comme suit :
  - 20 mn de soutenance orale par l'équipe.
  - 30 mn d'entretien collectif avec le jury (questionnement complémentaire).
  - Objectif : mettre en avant et démontrer que les compétences visées par ce bloc sont bien acquises.

- **Jury d'évaluation** : 2 personnes (binôme d'évaluateurs) par jury – Ces évaluateurs ne sont pas intervenus durant la période de formation et ne connaissent pas les apprenants à évaluer.

I - CONTEXTE

Au vu de la gestion de la récente pandémie du COVID 19 et de la montée du virus de la variole du singe (MPOX), l'Europe ainsi que le gouvernement Américain ont décidé de créer une nouvelle division de recherche au sein de l'OMS. Cette division portera ses recherches sur la détection et prévention des pandémies.

Organisation mondiale de la Santé

Cette nouvelle division du laboratoire a pour objectif de développer une plateforme permettant de collecter, nettoyer, analyser et visualiser les données historiques sur les pandémies.

Cette plateforme doit permettre aux chercheurs et aux décideurs de visualiser des indicateurs clés aux travers de différents tableau, de comprendre les dynamiques des pandémies passées à l'aide de graphique détaillés, et de formuler des hypothèses pour des modèles prédictifs.

Actuellement, la division ne compte qu'une seule personne à son actif. Cette personne se concentre actuellement sur la mise en place de l'infrastructure et n'a donc pas de temps pour pouvoir développer une quelconque solution.

II– EXPRESSION DES BESOINS

Vous travaillez pour le groupe ANALYZE IT, spécialisé dans les données de santé et dépêché par l'OMS en tant que prestataire de service pour répondre à son besoin de création de Système d'Information.

Votre équipe et vous-même avez la charge de développer une solution permettant d'ingérer des données provenant de différentes sources, telles que les bases de données publiques de santé, les archives hospitalières, les publications scientifiques, etc.

Les données devront être au format JSON ou CSV. Les données peuvent être des pandémies de votre choix, une justification sera attendue.

En complément, vous aurez la charge de nettoyer, trier, et assurer la qualité des données, en tenant compte des différentes sources hétérogènes. Il serait appréciable d'avoir une solution générique, le minimum attendu dans un premier temps est une solution permettant de nettoyer et trier des données sur deux sources de données de votre choix que vous justifierez.

Enfin, vous avez la charge de la construction d'une structure de stockage adaptée pour l'analyse des données, permettant une interrogation rapide et flexible de ces dernières. Vous mettrez également en place une solution de lecture de données pour l'équipe de développement de la division et créerez des tableaux de bord interactifs pour la visualisation et l'extraction des indicateurs clés de performance, tels que les taux de transmission, le taux de mortalité, etc.

III- LIVRABLES ATTENDUS

1. Un modèle de données au format Merise avec MCD, MLD et MPD ou format UML.
2. Une base de données relationnelle (script de création).
3. Mise en place d'une solution ETL. Les jobs mis en place devront permettre de :
   a. Extraire les données des fichiers Json et CSV
   b. Nettoyer, agréger, normaliser et supprimer les doublons de données
   c. Charger les données dans la base de données mis en place
4. Une API afin de pouvoir lire, modifier, supprimer et ajouter des données. L'API doit être flexible et permettre d'obtenir qu'une partie des informations si le développeur le souhaite. L'API ne sera utilisée que par un seul développeur pour le moment.
5. Une documentation d'API de type OPEN API Spécification.
6. Un tableau de bord interactifs réalisés avec un outil de datavisualisation, permettant l'exploration et l'exportation des données historiques des pandémies. Mise en place de filtre afin de fluidifier la lecture de ses tableaux. Les filtres mis en place seront à justifier dans une documentation.
7. Une documentation présentant la solution ETL ainsi que sa justification.
8. Une documentation détaillée sur le processus de collecte et de nettoyage des données.
9. Un benchmark des solutions (Comparaison des performances des outils de collecte et de visualisation des données utilisés).

Ce projet tend à évoluer, aussi, vous présenterez les différents jalons du projet sous forme de GANTT ou outils similaire en prenant en considération les évolutions du projet, autrement dit, votre gestion de projet n'est pas définitive et évoluera en fonction des besoins de l'OMS. Il est indispensable de remonter aux commanditaires du projet les différentes avancées. Vous adopterez les méthodes agiles pour construire cette partie.

IV-DATASET KAGGLE

- https://www.kaggle.com/datasets/imdevskp/corona-virus-report?resource=download
- https://www.kaggle.com/datasets/josephassaker/covid19-global-dataset
- https://www.kaggle.com/datasets/utkarshx27/mpox-monkeypox-data

WEBOGRAPHIE

- Documentation Pandas : https://pandas.pydata.org/pandas-docs/stable/
- Documentation PostgreSQL : https://www.postgresql.org/docs/
- Documentation Power BI : https://docs.microsoft.com/fr-fr/power-bi/
- Kaggle Dataset : https://www.kaggle.com/datasets/
- Guide ETL open source Apache Hop : https://hop.apache.org/manual/latest/getting-started/
- Guide ETL Talend : https://www.talend.com/fr/resources/guide-etl/
- Comparatif ETL vs ELT : https://www.talend.com/fr/resources/elt-vs-etl/