epsi®
l'école d'ingénierie informatique

Pro Alterna®

CERTIFICATION PROFESSIONNELLE N° 36921 EXPERT INGENIERIE DES DONNEES

BLOC 5 – Préparation et Mise à disposition des données d'une plateforme BigData aux équipes utilisatrices (Data scientists, équipe BI, décideurs et experts métiers)

Cahier des Charges de la MSPR « Exploitation d'une plateforme Big Data à partir d'une situation réelle ou reconstituée »

COMPÉTENCES ÉVALUÉES :

- Définir les données de référence de l'entreprise à partir des données utilisées pour créer un référentiel de données afin d'assurer la mise à disposition de données cohérentes aux directions métiers.
- Développer la phase de préparation des données afin de permettre leur chargement en prenant en compte les étapes de découverte, structuration, nettoyage, validation et intégration décrite dans le cahier des charges afin de les rendre consommables par les utilisateurs métiers.
- Assurer la qualité des données en utilisant les outils de gestion de la qualité de données pour garantir l'exactitude, la cohérence, la synchronisation et la traçabilité des données afin de satisfaire les besoins d'accessibilité des utilisateurs métiers.
- Appliquer les procédures de sécurité établies par le / la RSSI de l'entreprise afin d'assurer la confidentialité et la sécurité des données et garantir une mise en conformité avec les obligations légales du RGPD.
- Manipuler les différents services et fonctionnalités des architectures de type Data Lake afin de gérer le cycle de vie des données DLM (Data Life cycle Management).
- Proposer des modèles statistiques et de data science (machine learning) à mettre en pratique aux directions métiers afin de détecter des nouveaux services, anticiper des besoins et résoudre des problématiques métiers de l'entreprise.
- Organiser les sources de données sous forme de résultats exploitables (data visualisation) pour alimenter les outils décisionnels et visualiser les résultats de façon compréhensible permettant d'aider les directions métiers à la prise de décision.
- Développer divers services de stockage, de gestion de systèmes de bases de données, de production d'ensemble de données nettoyées et améliorées pour l'analyse grâce aux langages adaptés et répondant aux besoins afin de mettre en place l'exploitation de données par les différents métiers.

PHASE 1 : PRÉPARATION DE CETTE MISE EN SITUATION PROFESSIONNELLE RECONSTITUÉE

- Durée de préparation :
  - 28 *heures*
- Mise en oeuvre :
  - *Travail d'équipe constituée de 4 apprenants-candidats (5 maximum si groupe impair)*
- Résultat attendu :
  - *Le dossier devra contenir l'ensemble des éléments demandé, en particulier des plans d'actions présentant à une Direction générale les principes adoptés et les principaux parcours usagers.*
  - *Pour votre projet, vous êtes libre d'utiliser les outils vus pendant les cours et/ou les outils utilisés dans votre entreprise.*

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

TotalGreen est une société française travaillant dans le secteur des énergies renouvelables. Afin de développer son pôle de R&D, TotalGreen développe GoodAir : un laboratoire de recherche pour étudier la qualité de l'air et la qualité de l'eau en France.

Ce laboratoire a pour objectif de suivre la qualité de l'air et de l'eau afin de proposer des recommandations à la population, d'étudier les conséquences du changement climatique, et de déterminer des seuils d'alerte. Il pourra mener des recherches scientifiques sur le sujet tout en développant des plateformes de sensibilisation pour le grand public.

Le laboratoire est composé d'une dizaine de chercheurs et d'analystes dans le domaine du climat, de la biologie, et de la météorologie. Comme pour toute recherche, l'équipe du laboratoire a besoin de se reposer sur des données. Celles-ci doivent être fiables, disponibles, et pertinentes. Dans une problématique de limitation des coûts et du temps de collecte, le directeur du laboratoire souhaite se baser sur des sources de données déjà existantes.

Dans ce contexte, GoodAir a besoin de récupérer et stocker un certain nombre d'informations afin de les mettre à disposition de ses chercheurs. Ces données doivent pouvoir être accessibles sur un outil de data visualisation mais aussi exportables pour des études plus avancées. Le laboratoire GoodAir fait appel à vous pour auditer le projet.

II– SPÉCIFICATIONS DU BESOIN

2.1 Données

Deux sources de données intéressent GoodAir :

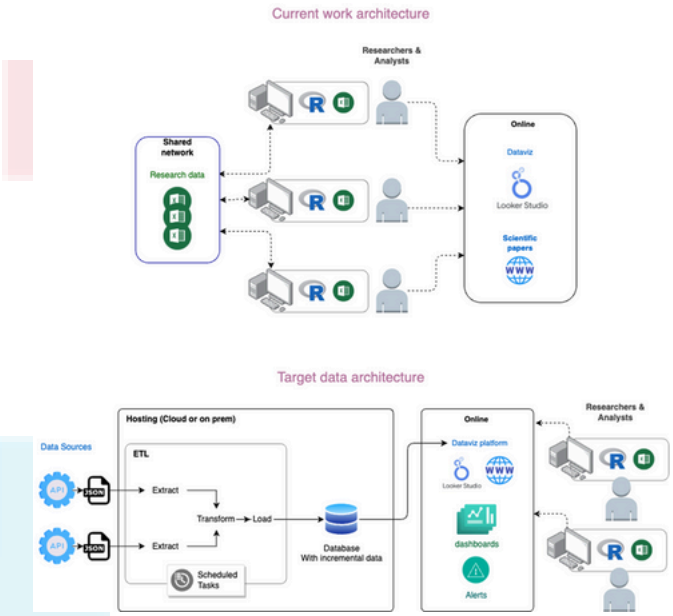
- Les données relatives à la qualité de l'air[1] : https://aqicn.org/json-api/doc/
- Les données météorologiques1 : https://openweathermap.org/api

2.2 Besoin

Good Air souhaite utiliser les données récoltées afin de mettre à disposition de ses chercheurs un certain nombre de rapport dans les principales villes de France. On cherche en effet à valoriser les données pour en extraire des KPI, visualisations, ou dashboards pertinents. GoodAir s'intéresse en particulier à trois types de livrables :

- Des rapports sur les indicateurs principaux de qualité de l'ait, indicateurs météorologiques ;
- Des rapports mettant en lumière les variations extrêmes de façon à alerter les équipes sur une situation anormale ;
- Des rapports mettant en lumière les variables fortement corélées.

Dans le futur, GoodAir pourrait faire appel à vous à nouveau pour réaliser des analyses plus poussées. Le laboratoire s'intéresse par exemple à la réalisation de prédictions sur les prochaines canicules ou la modélisation de la saisonnalité des données liée à la qualité de l'air. Il vous revient d'anticiper ces besoins dans la réalisation de vos livrables.



III - LES CONTRAINTES IMPOSÉES SONT LES SUIVANTES :

3.1 Modélisation :

Les données de ces deux sources doivent être modélisées dans une base de données normalisée. Avec le temps, cette base contiendra de plus en plus de données. Puisqu'elle sera connectée à un outil de data visualisation, le système choisi doit permettre de requêter des volumes de données conséquents en un temps minimum.

3.2 Intégration de données externes

Dans une problématique de conformité avec le RGPD, l'ensemble des traitements et des entrepôts de données doivent être localisés en France ou dans l'Union Européenne. De plus, l'accès aux données doit se faire de façon sécurisée (système d'authentification).

3.4 Qualité et fiabilité

Les deux sources dUne surveillance de la qualité des données doit pouvoir être mise en place au cours de la chaîne de traitement. Le livrable doit aussi être capable d'alerter les équipes (par email ou sur un rapport dédié) en cas de problème sur le pipeline ou la disponibilité des données.
e données sont disponibles à travers des API qui renvoient les informations en temps réel. GoodAir aurait besoin que le livrable soit capable de récupérer ces données chaque heure point en faire une capture et la stocker. Les données ainsi récupérées pourront être directement ajoutées à la base de données.

3.5 Data Viz

Dans une problématique de conformité avec le RGPD, l'ensemble des traitements et des entrepôts de données doivent être localisés en France ou dans l'Union Européenne. De plus l'accès aux données doit se faire de façon sécurisée (système d'authentification).

3.5 Qualité et fiabilité

Good Air souhaite réaliser des rapports sur l'ensemble des données récoltées. Les rapports créés doivent permettre de mettre en regard les informations météorologiques intéressantes et les métriques de qualité de l'air au cours du temps.

Dans cette partie, l'équipe Data aura pour mission d'extraire les données les plus pertinentes afin d'en faire un certain nom de visualisations. Si cela s'avère pertinent, certaines alertes peuvent être mises en place sur les indicateurs principaux en cas de dépassement d'un certain seuil.

III - LES CONTRAINTES IMPOSÉES SONT LES SUIVANTES :

- Développer la phase de préparation des données afin de permettre leur chargement, nettoyage, et validation
- Récolter et intégrer des données complémentaires pertinentes
- Développer une architecture d'intégration, nettoyage, et stockage des données
- S'assurer de la qualité des données livrées aux métiers afin d'en garantir la fiabilité et la disponibilité
- S'assurer de la sécurité des données et de leur cohérence avec le RGPD
- Proposer un outil de dataviz permettant d'accéder aux données
- Anticiper les besoins métiers pour réaliser des visualisations et rapports pertinents
- Proposer des utilisations possibles de ces données dans de la modélisation statistique

Les compétences évaluées durant cette MSPR :

- Définir les données de référence de l'entreprise à partir des données utilisées pour créer un référentiel de données afin d'assurer la mise à disposition de données cohérentes aux directions métiers.
- Développer la phase de préparation des données afin de permettre leur chargement en prenant en compte les étapes de découverte, structuration, nettoyage, validation et intégration décrite dans le cahier des charges afin de les rendre consommables par les utilisateurs métiers.
- Assurer la qualité des données en utilisant les outils de gestion de la qualité de données pour garantir l'exactitude, la cohérence, la synchronisation et la traçabilité des données afin de satisfaire les besoins d'accessibilité des utilisateurs métiers.
- Appliquer les procédures de sécurité établies par le / la RSSI de l'entreprise afin d'assurer la confidentialité et la sécurité des données et garantir une mise en conformité avec les obligations légales du RGPD.
- Manipuler les différents services et fonctionnalités des architectures de type Data Lake afin de gérer le cycle de vie des données DLM (Data Life cycle Management).
- Proposer des modèles statistiques et de data science (machine learning) à mettre en pratique aux directions métiers afin de détecter des nouveaux services, anticiper des besoins et résoudre des problématiques métiers de l'entreprise.
- Organiser les sources de données sous forme de résultats exploitables (data visualisation) pour alimenter les outils décisionnels et visualiser les résultats de façon compréhensible permettant d'aider les directions métiers à la prise de décision.
- Développer divers services de stockage, de gestion de systèmes de bases de données, de production d'ensemble de données nettoyées et améliorées pour l'analyse grâce aux langages adaptés et répondant aux besoins afin de mettre en place l'exploitation de données par les différents métiers.