epsi®
l'école d'ingénierie informatique

Pro Alterna®

CERTIFICATION PROFESSIONNELLE N° 36921 EXPERT INGENIERIE DES DONNEES

BLOC 3 – Construction d'une plateforme Big Data permettant la collecte, l'assemblage, le traitement et le stockage des données générées par les systèmes d'une entreprise

Cahier des Charges de la MSPR « Conception de l'architecture et construction de l'infrastructure d'une plateforme Big Data sécurisée à partir d'une situation réelle ou reconstituée »

COMPÉTENCES ÉVALUÉES :

- Concevoir une architecture de collecte et de restitution de données robuste, évolutive, sécurisée et utilisant l'intelligence artificielle (machine learning) afin d'améliorer en continu sa capacité à prédire les besoins Data des experts métiers utilisateurs.
- Concevoir une architecture de stockage de données (data warehouse, data lake…) permettant de répondre aux besoins Data des experts métiers et respectant la politique de sécurité des données définie par le/la RSSI.
- Définir les processus de collecte et d'intégration de données par lot ou en streaming afin de favoriser la vitesse d'intégration et la volumétrie de données dans le respect de l'architecture définie.
- Mettre en place un système d'ingestion de données structurées et non structurées afin de permettre la manipulation et l'accès aux données ainsi que l'authentification des utilisateurs.
- Développer une solution de migration inter-systèmes et multi-environnements à l'aide d'un outil de son choix afin de permettre l'intégration de données diverses et l'interopérabilité des différentes sources de donnée.
- Développer un pipeline de données et/ou un pipeline ETL prenant en compte l'environnement technologique déployé (infrastructure, services, applications…) dans le respect du cahier des charges de la solution proposée.
- Créer un lac de données (data lake) afin de collecter des données brutes dans le respect de l'architecture de collecte des données définie dans la solution proposée.
- Créer un entrepôt unique à partir du référentiel de données établi pour centraliser les informations stratégiques de l'entreprise et répondre rapidement aux besoins métiers.
- Déployer un processus de collecte, stockage et traitement de données selon une approche ETL (Extract-Transform-Load) ou une approche ELT afin de permettre l'extraction, le stockage et le traitement des données de manière optimale et adaptée aux besoins utilisateurs métiers.

PHASE 1 : PRÉPARATION DE CETTE MISE EN SITUATION PROFESSIONNELLE RECONSTITUÉE

- Durée de préparation :
  - *30 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

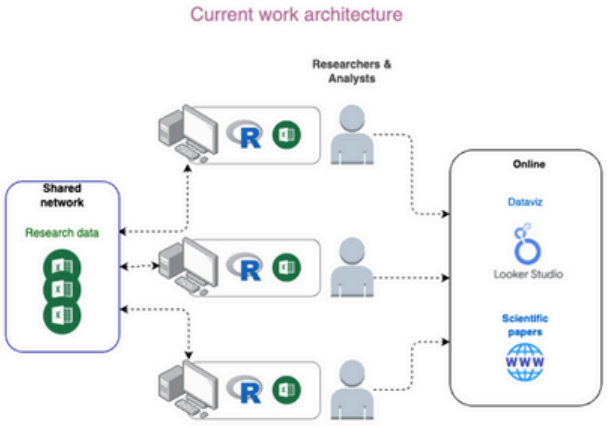
Deux sources de données intéressent GoodAir :

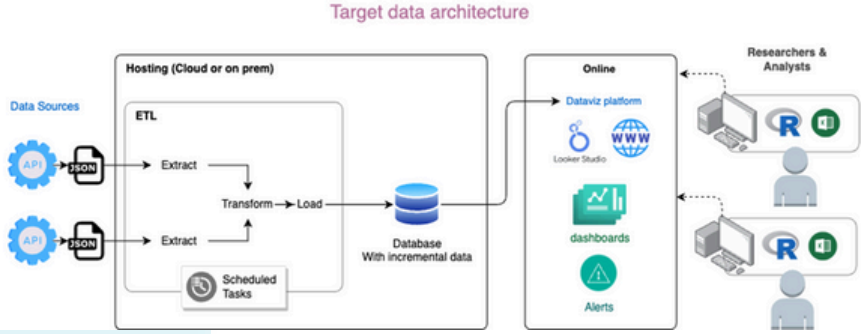
- Les données relatives à la qualité de l'air[1] : https://aqicn.org/json-api/doc/
- Les données météorologiques1 : https://openweathermap.org/api

2.2 Besoin

GoodAir souhaite utiliser ces données afin de mettre à disposition de ses chercheurs un certain nombre de rapports dans les principales villes de France.

Dans un premier temps, GoodAir fait appel à vous pour auditer le projet sur les problématiques de modélisation des données, de création des ETLs, ainsi que le stockage en base de données.





III - LES CONTRAINTES IMPOSÉES SONT LES SUIVANTES :

3.1 Accès aux données :

L'API openweathermap propose une formule gratuite dont le nombre d'appels quotidien est limité. Dans le cadre de ce projet, GoodAir souhaite se limiter à cette formule. Le livrable proposé doit donc s'assurer de ne pas dépasser les quotas imposés par l'API.

3.2 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, la modélisation doit permettre de requêter des volumes de données conséquentes en un laps de temps minimal.

3.3 Historisation :

Les deux sources de 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.

3.4 Sécurité

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

Une surveillance de la qualité des données doit pouvoir être mise en place au cours de la chaîne de traitement. Si cela s'avère pertinent, un processus de nettoyage sera mis en place. Le livrable doit aussi être capable d'alerter les équipes en cas de problème sur le pipeline ou la disponibilité des données.

3.6 Gestion de projet

GoodAir souhaite suivre de près l'évolution du projet pour y apporter des précisions ou modifications si nécessaire. Le laboratoire souhaite donc que vous travaillez sur le projet de façon itérative, en proposant un MVP (Minimum Viable Product) qui sera amélioré à chaque itération

III - LES CONTRAINTES IMPOSÉES SONT LES SUIVANTES :

- Réaliser un audit des données à disposition et leur cohérence avec le projet
- Proposer une modélisation normalisée des données à disposition
- Proposer une architecture de collecte et de stockage des données
- Proposer un système d'accès aux données sécurisé, limité pour les utilisateurs autorisés
- Concevoir un pipeline de données permettant de récupérer les données temps réel
- Concevoir un système de capture des données temps réel pour les stocker de façon incrémentale
- Proposer une roadmap réaliste du projet en mode itératif

Les compétences évaluées durant cette MSPR :

- Concevoir une architecture de collecte et de restitution de données robuste, évolutive, sécurisée et utilisant l'intelligence artificielle (machine learning) afin d'améliorer en continu sa capacité à prédire les besoins Data des experts métiers utilisateurs.
- Concevoir une architecture de stockage de données (data warehouse, data lake…) permettant de répondre aux besoins Data des experts métiers et respectant la politique de sécurité des données définie par le/la RSSI.
- Définir les processus de collecte et d'intégration de données par lot ou en streaming afin de favoriser la vitesse d'intégration et la volumétrie de données dans le respect de l'architecture définie.
- Mettre en place un système d'ingestion de données structurées et non structurées afin de permettre la manipulation et l'accès aux données ainsi que l'authentification des utilisateurs.
- Développer une solution de migration inter-systèmes et multi-environnements à l'aide d'un outil de son choix afin de permettre l'intégration de données diverses et l'interopérabilité des différentes sources de donnée.
- Développer un pipeline de données et/ou un pipeline ETL prenant en compte l'environnement technologique déployé (infrastructure, services, applications…) dans le respect du cahier des charges de la solution proposée.
- Créer un lac de données (data lake) afin de collecter des données brutes dans le respect de l'architecture de collecte des données définie dans la solution proposée.
- Créer un entrepôt unique à partir du référentiel de données établi pour centraliser les informations stratégiques de l'entreprise et répondre rapidement aux besoins métiers.
- Déployer un processus de collecte, stockage et traitement de données selon une approche ETL (Extract-Transform-Load) ou une approche ELT afin de permettre l'extraction, le stockage et le traitement des données de manière optimale et adaptée aux besoins utilisateurs métiers.