SAE 4.01 : développement avec une base de données et visualisation

Compétences ciblées :
– Développer — c'est-à-dire concevoir, coder, tester et intégrer — une solution informatique pour un client.
– Proposer des applications informatiques optimisées en fonction de critères spécifiques : temps d'exécution, précision, consommation de ressources..
~~– Installer, configurer, mettre à disposition, maintenir en conditions opérationnelles des infrastructures, des services et des réseaux et optimiser le système informatique d'une organisation~~
– Concevoir, gérer, administrer et exploiter les données de l'entreprise et mettre à disposition toutes les informations pour un bon pilotage de l'entreprise
– Satisfaire les besoins des utilisateurs au regard de la chaîne de valeur du client, organiser et piloter un projet informatique avec des méthodes classiques ou agiles
– Acquérir, développer et exploiter les aptitudes nécessaires pour travailler efficacement dans une équipe informatique

Objectifs et problématique professionnelle :
La problématique professionnelle est d'améliorer une base de données existante, du point de vue de la qualité, la performance et la sécurité. Il sera également nécessaire de proposer des outils de visualisation des données. L'ensemble sera mis en œuvre au sein d'une équipe, dans une démarche de développement itérative ou incrémentale.

Descriptif générique :
En partant d'une application existante et de sa base de données, l'équipe devra en évaluer les performances, la qualité du modèle, détecter les éventuelles failles de sécurité, afin d'améliorer l'ensemble de ces points. Elle devra également proposer des outils de visualisation des données pour les utilisateurs. En outre, l'impact environnemental de la solution devra être pris en compte

Les livrables attendus généralement dans le monde professionnel sont :
– Documents de suivi du projet
– Jeux d'essais
– Modèle et script optimisé de la base de données
– Outils de visualisation proposés
– Modèles IA développés
– Compte-rendu sur les optimisations réalisées et sur la sécurité
– Revue finale du projet

Ressources mobilisées et combinées :
– R4.01 | Architecture logicielle
– R4.02 | Qualité de développement
– R4.03 | Qualité et au-delà du relationnel
– R4.05 | Anglais
– R4.06 | Communication interne
~~– R4.Admin.08 | Cryptographie et sécurité~~
~~– R4.Admin.09 | Réseau avancé~~
– R4.Admin.10 | Analyse et visualisation avancée des données

– R4.Admin.11 | Management avancé des systèmes d'information

Sujet 2023 :

https://fr.openfoodfacts.org/

Open Food Facts (OFF) est un projet collaboratif dont le but est de constituer une base de données libre et ouverte sur les produits alimentaires commercialisés dans le monde entier. Cette base est manuellement alimentée par des volontaires. Open Food Facts est disponible via un site web ou des applications pour mobiles. L'association 1901 qui gère cette plateforme dispose actuellement d'une base de données relationnelle ne contenant qu'une seule table. Les données disponibles dans cette table représentent 7 Go de données. Après suppression des produits doublons, suppression des produits pour lesquels les informations essentielles sont absentes et suppression des colonnes dont l'information n'est plus jugée comme pertinente, ces données resprésentent 326,6 Mo, accessibles au format CSV à https://drive.google.com/file/d/1XQBMKRM56GSTA-9_CPl4vdgiWeql4wWE/view?usp=sharing

OFF souhaite faire évoluer sa base de données PostgreSQL de façon à stocker, pour chaque produit, les volumes de ventes observés à différentes périodes pour différents distributeurs. **Vous devrez tout d'abord constituer à partir du fichier plat un modèle relationnel OLTP** afin de limiter la redondance et le volume des données. En ce qui concerne **les chiffres de vente, ceux-ci devront être simulés pour les besoins de cette SAE**. En termes de volumétrie, vous veillerez à ce que chaque produit dispose en moyenne d'une centaine de chiffres de ventes sur les trois dernières années, tous distributeurs confondus, et ce en appliquant une ou plusieurs lois normales. Attention : les valeurs manquantes ne devront pas être imputées.

Vous mettrez en place un workflow Knime permettant d'explorer les données de cette base. Vous analyserez ensuite les données de cette base sous la forme de rapports analytiques PowerBI réalisés en anglais. Vous analyserez également les performances de ces rapports et pourrez proposer une nouvelle architecture que vous pourrez mettre en place.

Une fois ceci fait, OFF souhaite également que **vous mettiez en place un modèle IA sous Knime qui puisse automatiquement calculer le Nutri-Score d'un produit**, i.e. sa qualité nutritionnelle, en fonction de ses caractéristiques. Ce modèle permettra ainsi de détecter d'éventuelles erreurs de saisie (valeurs et grades du Nutri-Score). L'utilisation de la formule de référence n'est en effet pas possible, car les données sur les quantités de fibres et de sodium contenues dans les produits référencés ne sont pas souvent disponibles(données

non fournies par le fabricant/vendeur ou non saisies par les volontaires). Pour information, la formule du Nutri-Score a été proposée par l'EREN, une équipe de recherche publique française sur la nutrition, dirigée par le professeur Serge Hercberg. Il est basé sur le score nutritionnel de la FSA créé par la Food Standards Agency du Royaume-Uni.

Un deuxième besoin, pour lequel les techniques d'IA seront mobilisées, est la **création automatique d'un Open Fact Food score via un modèle dédié. Ce score, que vous construirez sous forme de grades (nombre à définir)**, doit proposer un équivalent au Nutri-Score qui soit établi uniquement en fonction des données disponibles. **Un dernier modèle IA devra enfin être construit de façon à attribuer automatiquement ce score OFF à chaque saisie de produit.** Une attention particulière sera portée au fait que les modèles IA devront être développés en maîtrisant la consommation de ressources : seule l'utilisation d'un PC standard, et non l'usage d'une station de calcul intensif, sera autorisée. Par ailleurs, vous veillerez à adopter des plans d'expérience qui soient les plus frugaux possible.

Groupes : 4 à 5 étudiants

Planning :

semaine	**base de données**	**rapports Power BI**	**IA Knime**	**autonomie**
s10	P. Colin (4h)	V. Couturier (4h)	N. Méger (4h)	8h
s11	P. Colin (6h)	V. Couturier (4h)	N. Méger (4h)	10h
s12	P. Colin (2h)	V. Couturier (8h)	N. Méger (8h)	8h
s13		V. Couturier (8h)	N. Méger (8h)	12h
s14		V. Couturier (6h)	N. Méger (6h)	8h

+ soutenance finale en s14, 2 heures : P.Colin, V. Couturier, N. Méger

Livrables :
- Documents de suivi du projet (cycle de développement, GANTT final)
- Document technique :
  - modèle et script optimisé de la base OLTP en expliquant les optimisations,
  - modèle en étoile expliqué et choix d'architecture,
  - modèles IA produits en expliquant la démarche de développement et d'optimisation de ces modèles (données d'apprentissage, types de modèles, paramétrages, performances atteintes).
- Livrables techniques : script des bases, rapports power BI, workflows Knime
- Soutenance finale du projet dont une conclusion en anglais
- Abstract d'1/2 page présentant le sujet (thème et résultats)

Date limite de remise : 7 avril 2023

PostgreSQL :

- Vous pourrez utiliser les instances locales de PostgreSQL disponibles sur les machines de l'IUT
- OU utiliser le serveur PostgreSQL sur OVH (mais vous ne pourrez pas vous y connecter via PowerBI car pas de https) géré par Luc Damas
- OU configurer un serveur PostgreSQL sur Azure (voir fichier explicatif)

Vous ne pourrez pas utiliser Alwaysdata.com car limité à 100 Mo.