

Cahier de Charges : Analyse de données de ventes de voitures

Objectif du Projet :

Le projet vise à implémenter un pipeline de données de ventes de voitures afin de conseiller à l'achat de voitures.

Portée du Projet :

- Ciblage des sources de données
- Traitement
- Modélisation et stockage
- Visualisation et analyse.

Méthodologies utilisées :

SCRUM, CRISP-DM.

Exigences de Données :

Les données seront collectées depuis deux sources majeures :

- DVM-CAR : Un ensemble de données à grande échelle pour les applications automobiles.
- Marketcheck : Données de marché pour les chercheurs et les développeurs.

Exigences Techniques :

Utilisation de technologies adaptées au traitement des données. Respect des bonnes pratiques de sécurité des données.

RGPD :

Nous allons effectuer nos transformations conformément aux normes et règlements du RGPD et de la loi 09-08 du dahir marocain.

Python :

Suite à sa grande popularité ainsi que sa communauté, nous utilisons Python comme langage de programmation.

HDFS :

Pour sa puissance et sa variété de stockage de données brutes. Nous l'utilisons comme "staging area".

SQL Server :

L'entrepôt de données sera mis en place avec SQL Server.

Airflow :

Le pipeline doit être optimisé en batch et en temps réel.

Power BI :

Faire des analyses et visualisations avec Power BI pour sa grande connectivité avec SQL Server et son interactivité.

Docker :

Nous utiliserons Docker pour la configuration de l'environnement, ce qui facilitera le déploiement.

Git / Github :

Nous assurons un système de gestion de version avec Git et Github.

Github Actions :

Pour assurer une continuité dans l'intégration et le déploiement, nous utiliserons Github Actions.

Contraintes et Limitations :**Contraintes Temporelles (Délais) :**

Le projet doit être complété selon le temps imparti qui est d'un mois.

Contraintes Budgétaires (Coûts) :

Les coûts sont limités au budget préalablement alloué pour le projet.

Calendrier et Jalons :

- 01/01/2024 (Sprint 1) : Business Case, validation du cahier de charges, mise en place des règles de gouvernance, conformité au RGPD et ciblage et extraction des données.
- 08/01/2024 (Sprint 2) : Validation des données, Configuration de l'environnement, implémentation du pipeline.
- 19/01/2024 (Sprint 3) : Analyse et visualisation. Déploiement du pipeline.

Perspectives et Améliorations Futures :

Intégrer un système de traitement des données en temps réel pour conseiller aux clients à partir des voitures mises en vente récemment.

Date de publication : 01/01/2024

Signé : John Doe