Cahier de Charge : Analyse Temps Réel des Logs d'Application avec Apache Spark sur Azure HDInsight

Objectif du Projet:

L'entreprise souhaite surveiller et optimiser la performance de son application ainsi que d'assurer une expérience fluide aux utilisateurs

Portée du Projet:

Implémenter un système pour :

- Surveiller et optimiser la performance de l'application
- Détecter les anomalies
- Analyser le comportement des utilisateurs

Méthodologies utilisées:

CRISP-DM.

Exigences de Données:

Collecte des données depuis notre plateforme. Création d'une API contenant 3 points d'entrées à savoir:

- Movies : Données liées aux films
- Users : Données liées aux utilisateurs
- Ratings : Intérations entre les utilisateurs et les films

Exigences Techniques:

Utilisation de technologies adaptées au développement web et au traitement des données. Respect des bonnes pratiques de sécurité pour l'authentification des utilisateurs et la protection des données.

RGPD:

Nous allons effectuer nos transformations conformément aux normes et réglements du RGPD et de la loi 30-09 du dahir marocain.

Python:

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

Flask:

Étant un framework de Python, nous utiliserons Flask pour la création de l'API.

Spark dans Azure HDInsight:

Spark est une technologie puissante qui supporte le traitement des données à grande échelle. Le service HDInsight dans Azure nous permettra de l'utiliser dans le cloud.

Azure Event Hub:

Le service Event Hub sera utilisé pour l'ingestion des données.

Azure Zeppelin:

Azure Zeppelin est un outil open source qui permet une analyse de données interactive et basée sur les données. Il permet de faire une visualisation intéractive.

Git / Github:

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

Contraintes et Limitations:

Contraintes Temporelles (Délais):

Le projet doit être complété selon la durée définie par les parties prenantes.

Contraintes Budgétaires (Coûts):

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

Calendrier et Jalons:

- 18/12/2023 : Gouvernance des données, planification et conception.
- 19/12/2023 : Mise en place de l'environement avec Azure.
- 20/12/2023 : Développement du Pipeline de Données.
- 21/12/2023 : Analyse et Visualisation.
- 22/12/2023 : Fermeture et livraison du projet.

Perspectives et Améliorations Futures:

Optimiser les requetes d'intéraction avec la base de données, revoir le matériel réseau pour minimiser la latence.

Date de publication : 18/12/2024
Signé : Sefdine Nassuf