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:

Nous allons collecter les donées dépuis notre plateforme. (Dans notre cas: Un script de géneration des données.)

Exigences Techniques:

Utilisation de technologies adaptées au traitement des données au niveau du cloud. Respect des bonnes pratiques de sécurité pour la protection des données.

Python:

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

Spark dans Azure Databricks:

Spark est une technologie puissante qui supporte le traitement des données à grande échelle. Le service Databricks 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 Power BI:

Azure Power BI est une solution unifiée d'analyse en libre-service et d'entreprise qui permet aux utilisateurs de visualiser leurs données, de partager des informations au sein de leur organisation et de les intégrer dans leur application ou leur site Web.

Git / Github:

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

Jira:

Etant unesolution optimale pour planifier le projet.

RGPD:

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

Contraintes et Limitations:

Gestion des imprévu:

Suite au changement majeur demandé par le client suite à l'utilisation des services Azure HDInsight et Azure Zeppelin à cause de leurs disponibilité avec notre souscription, nous avons opté à l'utilisation de services similaires qui sont disponible pour nous avec un coût bénefique. Les services de remplacement sont Azure Databricks et Azure Power BI.

Ce changement a impacté le planning de notre travail du fait que l'on a pris le temps de se documenter aux services et comment les configurer en assurant la connectivité entre eux.

Contraintes Temporelles (Délais):

Le projet doit être complété selon la durée définie par les parties prenantes qui est d'une semaine allant du 18/12/2023 au 22/12/2023

Contraintes Budgétaires (Coûts):

Les coûts sont limités au budget préalablement alloué pour le projet. Ce budget est de 250.25 Dirhams :

- Azure Event Hub : 6.37 Dirhams pour 40 heures
- Azure Databricks : 192.92 Dirhams pour 40 heures
- Azure Power BI : 50.96 Dirhams pour 5 heures

Pour plus d'informations, vous trouverez un document excel concernant les détails des couts pour chaque services.

Dépôt github : https://github.com/Sefdine/realtime-log-analysis/tree/services/pricing

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