Rapport sur le Projet Big Data sur Azure pour "5-A Shop" - Application web d'E-commerce

ELOUARDY Abderrahim
AZEMOUR Amine
ELFAQUIRI Amine
SABRI Aymane
BENAZZOUZ Amine

December 25, 2023

Contents

1 Introduction

1.1 Contexte

Le domaine en constante évolution de l'ingénierie des données occupe une place cruciale dans le paysage technologique contemporain. La capacité à collecter, traiter et analyser d'énormes volumes de données en temps réel offre des opportunités significatives pour comprendre le fonctionnement des applications, améliorer les performances, et prendre des décisions éclairées. C'est dans ce contexte dynamique que le présent projet prend forme, se concentrant sur l'établissement d'une infrastructure robuste utilisant Azure Event Hub, Spark, et d'autres outils associés pour la collecte, le traitement, et l'analyse en temps réel de logs générés par l'application web de "5-A Shop" une plateforme de commerce électronique prospère qui propose une vaste gamme de produits.

1.2 Objectifs du Projet

L'objectif principal est d'assurer une expérience utilisateur transparente, d'optimiser les performances de l'application et de réagir proactivement aux problèmes potentiels en exploitant les données générées par les logs.

1.3 Importance du Projet

La création d'une infrastructure d'ingénierie des données va au-delà de la satisfaction des besoins immédiats d'analyse en temps réel. Elle établit une base solide pour des améliorations continues. En fournissant des insights en temps réel, ce projet contribuera à optimiser les performances de l'application, renforcer la sécurité, anticiper les problèmes potentiels, et faciliter la prise de décisions éclairées basées sur une analyse approfondie des données.

2 Planification et Conception

2.1 Tâches du Projet

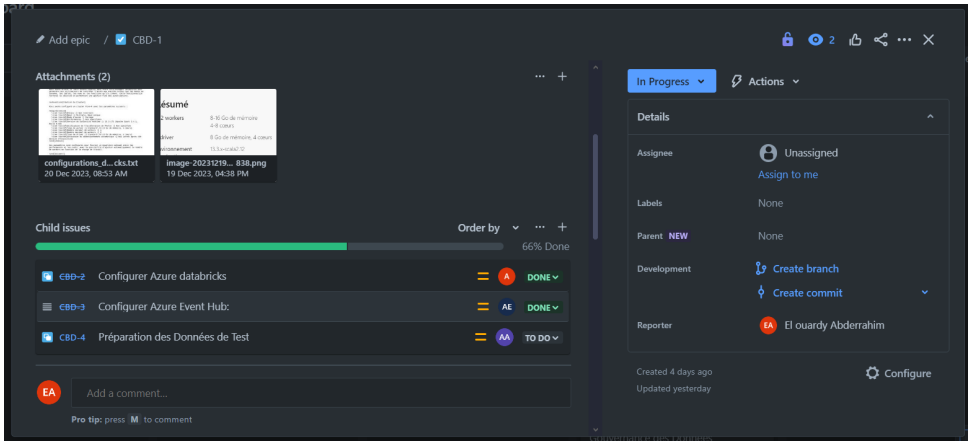

Figure 1: Capture d'écran illustrant exemple de tâche.

Nous avons utilisé la plateforme Jira de manière intensive pour orchestrer la planification et la division des tâches tout au long du projet. L'utilisation de Jira nous a permis d'établir une structure organisée et collaborative, offrant une visibilité claire sur l'ensemble du processus de développement.

2.2 Choix des Technologies

2.3 Architecture du Système

image est un schéma simplifié et structuré illustrant le processus d'analyse en temps réel des logs.

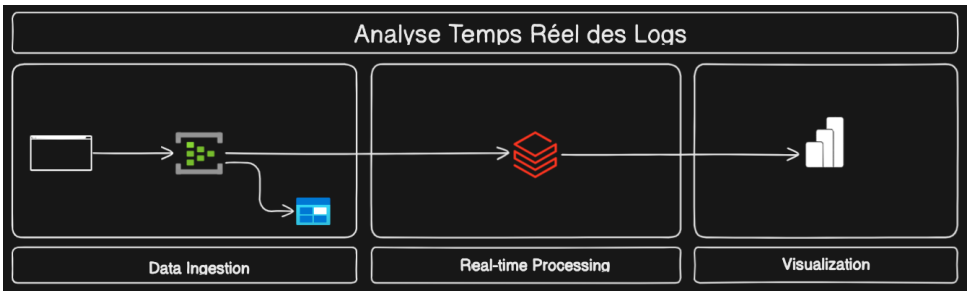

Figure 2: schéma simplifié et structuré illustrant le processus d'analyse en temps réel des logs.

3 Mise en Place de l'Environnement

3.1 Configuration d'Azure databricks

Table Access Control Nous avons activé le Table Access Control dans notre environnement Databricks pour permettre aux utilisateurs de contrôler l'accès aux entités telles que les bases de données, les tables, les vues et les fonctions qu'ils créent. Cette fonctionnalité renforce la sécurité en permettant une gestion fine des autorisations.

Création du Cluster Nous avons configuré un cluster Five-A avec les paramètres suivants :

- **Policy :** Non restreint
- **Nœud :** Multiple, Nœud unique
- **Mode d'accès :** Partagé
- **Performances :** Non spécifiées
- **Version de Databricks Runtime :** 13.3 LTS (Apache Spark 3.4.1, Scala 2.12)
- **Utilisation de l'accélération de Photon :** Non spécifiée
- **Type de worker :** Standard_F4 (8 Go de mémoire, 4 cœurs)
- **Nombre minimal de workers :** 1

- **Nombre maximal de workers :** 2
- **Type de driver :** Standard_F4 (8 Go de mémoire, 4 cœurs)
- **Activation du dimensionnement automatique :** Oui (arrêt après 120 minutes d'inactivité)

Figure 3: Capture d'écran illustrant résumé après Configuration d'Azure databricks.

Ces paramètres sont configurés pour fournir un équilibre optimal entre les performances et les coûts, avec la possibilité d'ajuster automatiquement le nombre de workers en fonction de la charge de travail.

3.2 Configuration d'Azure Event Hub

Dans une première étape, le "namespace Event Hub" nommé "five-a" a été créé pour établir le contexte de notre infrastructure.

- **Niveau Tarifaire :** Opté pour le niveau standard avec un paiement par unité de débit (TU) mensuel.
- **Unités de Débit :** Au niveau standard, la capacité de débit des Event Hubs est régulée par des unités de débit (UD), permettant l'ingestion jusqu'à 1 Mo par seconde ou 1 000 événements par seconde.

4 Développement du Pipeline de Données

4.1 Description des Transformations dans PySpark

Le script PySpark implémente une série de transformations sophistiquées sur les données en provenance d'Azure Event Hub. Voici une explication détaillée des principales transformations appliquées :

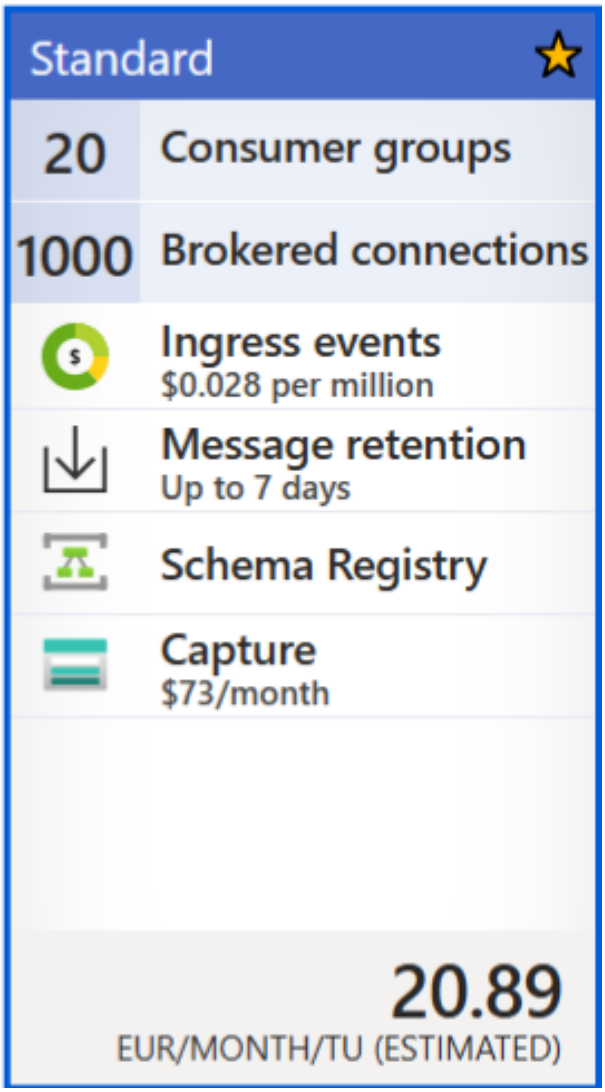

Figure 4: Capture d'écran illustrant la création du namespace Event Hub.

- **Division des données :**Les données brutes, encapsulées dans le corps des messages, sont soigneusement décomposées en plusieurs colonnes distinctes en utilisant le délimiteur —. Chaque composant séparé est attribué à une colonne spécifique, établissant ainsi une structure ordonnée.

- **Conversion des types de données :** Les colonnes subissent des transformations de type pour assurer la cohérence des données. Par exemple, la colonne timestamp est convertie en un type de données temporel, tandis que la colonne 'response time' est extraite en tant que nombre à virgule flottante.

- **Normalisation de la casse :** Pour garantir une uniformité dans la casse, les colonnes "log level", "action" et "http method" sont converties en minuscules.

- **Extraction de sous-chaînes :** Les valeurs des colonnes "endpoint", "referrer", "ip address" et "user agent" sont modifiées pour éliminer les préfixes indésirables, facilitant ainsi une analyse plus précise.

- **Extraction d'informations spécifiques à l'aide d'expressions régulières :**

L'utilisation d'expressions régulières permet d'extraire avec précision le temps de réponse de la colonne "response time".

- **Remplacement de motifs dans les chaînes :** Diverses colonnes, telles que "product id", "cart size" et "checkout status", subissent des remplacements de motifs pour nettoyer et uniformiser les valeurs.
- **Cryptage des données sensibles :** Les informations sensibles, telles que "token" et "server ip", sont sécurisées à l'aide d'une fonction de hachage SHA-256 définie par l'utilisateur (UDF), renforçant ainsi la confidentialité des données.
- **Affichage du résultat final :** Le DataFrame résultant de ces transformations est ensuite affiché, fournissant une visualisation claire et lisible des données préparées pour des analyses ultérieures.

5 Objectifs Analytiques et Visualisations

5.1 Objectifs Analytiques Spécifiques

5.1.1 Phase de Traitement des Données

Dans le cadre du respect des normes de confidentialité, les variables sensibles telles que le token et l'adresse IP du serveur subiront un processus de cryptage avant d'être incluses dans le pipeline d'analyse. Cela garantit la conformité avec le Règlement Général sur la Protection des Données (RGPD), assurant ainsi la confidentialité des informations utilisateur.

5.1.2 Système d'Alerte pour la Détection d'Anomalies

Un système d'alerte a été mis en place pour détecter les anomalies dans les logs en temps réel. Les conditions suivantes ont été définies pour déclencher une alerte:

- Action "Unusual Action": Si une action inattendue est détectée dans les logs, une alerte est déclenchée. Cela peut indiquer un comportement non conforme ou potentiellement malveillant.
- Temps de Réponse Supérieur à 6 Secondes: Si le temps de réponse total dépasse 6 secondes, une alerte est générée. Ce seuil a été défini en considérant les composants individuels tels que le temps de requête de la base de données, le temps de traitement du serveur et la latence du réseau, chacun ayant une limite maximale de 2 secondes.
- Taille de Panier Entre 100 et 1000: Si la taille du panier d'un utilisateur est détectée comme étant comprise entre 100 et 1000, une alerte sera générée. Une taille de panier inhabituellement élevée peut nécessiter une vérification supplémentaire.
- Status_and_Detail avec "500 Internal Server Error": Si le statut de la réponse HTTP est signalé comme "500 Internal Server Error" dans le détail, une alerte sera déclenchée. Cela indique une erreur interne du serveur nécessitant une attention immédiate.

5.1.3 Analyse de Performance

L'analyse de performance inclut une distinction entre la latence provenant du serveur et celle provenant de la base de données. Cela permet d'identifier la source des retards, que ce soit côté serveur ou côté base de données, facilitant ainsi la résolution des problèmes.

5.1.4 Analyse du Comportement des Utilisateurs

Suivi des actions, de la durée des sessions et des modèles de navigation avec des analyses de séquences.

5.1.5 Informations sur le Commerce Électronique

Analyse des vues de produits, des tailles de panier et des statuts de paiement pour comprendre le comportement des consommateurs.

5.1.6 Suivi de l'Utilisation de l'API

Analyse des appels d'API pour comprendre quelles fonctionnalités sont les plus utilisées.

5.1.7 Décisions d'Équilibrage de Charge et de Mise à l'Échelle

Analyse des modèles de trafic pour optimiser l'allocation des ressources.

5.1.8 Audit de Sécurité

Analyse des détails d'authentification et d'autorisation pour surveiller les tentatives de connexion échouées et les adresses IP inhabituelles.

6 Conclusion

Ces mécanismes contribuent à la surveillance proactive du système, permettant une réaction rapide aux situations anormales et facilitant la résolution des problèmes potentiels avant qu'ils n'affectent significativement l'expérience utilisateur.