Sujet de Programmation large échelle PLE 2024-2025

L'objectif de ce projet est de concevoir un système d'analyse de combinaison de cartes pour le jeu vidéo Clash Royale. Dans ce jeu chaque joueur possède un ensemble de cartes et il doit en choisir 8 pour constituer un « deck » qui lui permettra d'affronter un adversaire qui lui aussi possède un deck de 8 cartes.

Nous avons enregistré près de 5.000.000 de parties de ce jeu dans notre master dataset, nous disposons pour chaque partie du deck de chaque joueur ainsi que des informations sur le niveau d'expérience des joueurs mais aussi sur le niveau moyen des cartes des decks.

L'objectif du projet est d'obtenir un système similaire à celui proposé ici: https://royaleapi.com/decks/popular pour aider les joueurs à choisir les meilleurs combinaisons de cartes à jouer.

Le projet se décompose en 2 parties indépendantes.

Partie I (Map/Reduce Hadoop): Data Cleaning

Dans cette partie, nous vous demandons en Hadoop de nettoyer le jeu de donnée pour garantir :

- Que toues les lignes soient bien formattés en JSON
- Que les decks comportent le bon nombre de cartes (exactement 8 pour les deux adversaires)
- Qu'il n'y ait pas de doublon exact
- Que les parties ne sont pas enregistrées plusieurs fois (même joueurs même date).
    - L'ordre des joueurs varie en fonction de si la partie est téléchargée à en prenant comme source le joueur A ou le joueur B
    - L'heure de la partie peut varier à quelques secondes alors pour des partis identiques.

Rendu : Le source de votre projet ainsi qu'un rapport présentant la description et l'analyse de votre programme (ressources nécessaires, scalabilité horizontale, nombre reducer, messages etc…). Vous analyserez les performances de celui-ci sur les différents jeux de données disponible sur le cluster.

Présentation orale (10 décembre) : Analyse et performance de votre solution (5 minutes par groupe)

Partie II (Spark) :

Dans cette partie, nous voulons générer pour chaque decks toutes les combinaisons possibles de 1, 2, 3 ,4, 5, 6, 7 et 8 cartes. Cela représente 255 possibilités pour chaque deck de notre jeu de données.

Pour chacune de ces combinaisons nous voulons calculer les statistiques suivantes :

- Nombre de victoires
- Nombre de parties
- Nombre de joueurs différents (20 max)
- Statistiques des évolutions et des tours.

Votre programme devra générer un fichier JSON avec ces statistiques.

1) Un code effectuant toutes les opérations demandées est à votre disposition (/home/auber/spark_ple sur la gateway). Votre objectif est d'expliquer et étudier son fonctionnement

et de prouver quels sont les performances attendues ainsi que son paramétrage optimal (scalabilité horizontale, transfert réseau et nombre d'exécuteur, mémoire des exécuteurs et du driver).

2) Vous proposerez ensuite des améliorations du code existant pour améliorer ses performances globales et vous comparerez votre nouvelle solution avec celle proposée. Vous avez 100% de liberté, la seule contrainte est que le fichier de départ et le fichier de résultat soient identiques à ceux fournis/obtenus avec le code de d'origine.

Présentation orale (15 minutes) : Analyse et performance de la solution de départ, présentation des améliorations proposées et des mesures de performance de celles-ci.