Projet : LiChess Statistical Thinking

LiChess est une application Web totalement gratuite qui rassemble des joueurs d'échecs et leur permet de disputer des parties en ligne mais aussi d'apprendre à jouer. Les données sur les parties et les joueurs sont enregistrées et disponibles.

À partir des données de LiChess, on souhaite concevoir un système d'aide à la décision pour les joueurs d'échecs. Ce système doit exploiter les données des parties existantes pour fournir au joueur des indications sur le coup joué (nombre de parties gagnantes ou perdantes et nombre de coups, probabilité du coup, etc.) et une recommandation sur les différents coups suivants possibles. Le système devrait être aussi capable de disputer des parties contre des humains avec ou sans l'assistance d'un moteur de jeu comme StockFish. Une autre fonctionnalité dérivée serait de détecter des joueurs non humains lors des parties en ligne (ce qui constitue une triche). En effet, les coups proposés par les moteurs s'éloignent quelques fois de la forme habituelle des coups joués par les humains, permettant ainsi de détecter des anomalies ou des régularités. La fonctionnalité de détection pourrait servir à améliorer le moteur proposé, en le rendant "plus réaliste" et plus proche des coups habituellement joués par les humains.

Le projet global[1] consiste à réaliser une application qui met en avant les possibilités offertes par l'exploitation des données de LiChess. Les principales fonctionnalités sont :

— établir des statistiques générales sur les joueurs, les parties ;
— associer des métriques à des coups ;
— établir des recommandations de coups à partir d'un état de l'échiquier.

La collecte et le stockage des données sont donc les éléments essentiels de l'application. Les données sont récupérables en ligne.

Gestion des données

Dans un premier temps, les données des LiChess devront être insérées dans une base de données relationnelle, dont le schéma doit permettre de supporter les différents fonctionnalités de l'application.

Pour cela, les fichiers bruts devront être étudiés afin d'identifier les différents attributs et de concevoir un schéma adapté. Un programme ETL (*Extract Transform Load*) devra ensuite être conçu pour l'insertion des données. Ce programme doit pouvoir être réutilisé facilement en cas de nouvelles données disponibles.

Consultation des statistiques générales

Un utilisateur qui possède un compte sur l'application doit pouvoir s'authentifier et consulter des statistiques sur les parties, les joueurs, sélectionner une partie et dérouler les coups et également avoir des statistiques sur chaque coup de la partie. L'affichage doit pouvoir se faire sous forme textuelle ou graphique.

La problématique principale induite par le volume des données est l'organisation d'une structure de données pour permettre de calculer rapidement les indicateurs. En effet, le volume total des parties jouées dépasse plusieurs To (dans un premier temps on se concentrera sur quelques Go). Une problématique secondaire concerne la scalabilité c'est-à-dire la capacité du système à absorber les nouvelles données sans compromettre la durée d'exécution du calcul des indicateurs.

En outre, une spécificité de ce projet concerne la partie utilisateur qui ne modifie pas les données mais nécessite des interrogations rapides. Il conviendra donc de faire une typologie des requêtes à exécuter pour satisfaire les besoins, d'étudier leur durée d'exécution et de déterminer puis tester une architecture technique et logicielle capable de fournir des données rapidement à la partie *front*.

1. qui se déroule avec une première phase sur le S1 de master 1 et une seconde phase sur le S2 du master 1.

Travail à réaliser pour le premier semestre

Sur le premier semestre, le travail se concentrera principalement sur le stockage, consultation des données et le calcul d'indicateurs. Il s'agit de construire le socle technique de l'application qui sera utilisé ensuite pour les autres fonctionnalités du projet. Pour cela, les principales tâches à réaliser sont les suivantes :

- concevoir le schéma de stockage des données, nécessitant une phase d'étude des données disponibles ainsi que des données nécessaires pour l'application ;
- développer un programme ETL pour construire la base de données ;
- développer la partie serveur, qui permettra aux utilisateurs d'accéder aux différentes fonctionnalités de l'application via une interface Web.

Liens et documentation utiles

- Données de LiChess : `https://database.lichess.org/`
- Examples de projets et documentation : `https://lichess.org/source`
- `https://github.com/jcw024/lichess_database_ETL`
- `https://blog.scottlogic.com/2017/09/01/apache-spark-meets-chess.html`