Programmation Big Data

Présentation du module

Sylvain Tenier

Département TIC - Esigelec

Février 2016

Objectifs

1. Programmation concurrente
2. Programmation distribuée évolutive
3. SQL sur systèmes distribués
4. Systèmes “In memory”

Prérequis

1. Tronc commun : JAVA, SQL, gestion de projet
2. TIC : Git
3. Dominante : Threads JAVA

Déroulé type d'une séance

1. Présentation et restructuration (30 min)
   - Présentation de la séance
   - Restructuration des séances précédentes
     - Envoi de vos questions par mail avant la séance
2. Réalisation (3h)
   - Conception (groupe de 4)
     - Groupes différents des autres modules !
   - Réalisation
   - Création des slides de restitution
3. Restitution (30min)
   - Projet Git à jour sur Github
   - Présentation orale de 5 min maximum (bibliographie, description de l'approche retenue, répartition des taches, résultats obtenus, perspectives)
   - Tableau comparatif des performances

Dataset 1 : MovieLens

- Le groupe de recherche GroupLens a collecté plusieurs millions de notes attribuées à des films par des utilisateurs du site Movielens
- Vous devez exploiter ces données pour répondre aux questions suivantes :
  1. Quelle est la proportion de films ayant 1, 2, 3, 4, 5 étoiles ?
  2. Quel utilisateur a noté le plus de films ?
  3. Quel est le film le plus populaire ?
- Vous répondrez à ces questions en utilisant une technologie différente à chaque séance

Dataset 2 : Yelp

- Yelp est un service permettant de trouver et noter des services de proximité. La société a mis à disposition un “ddataset académique” comprenant les évaluations des 250 sociétés les plus proches de 30 universités :
  - `https://www.yelp.com/academic_dataset`
- Vous devez exploiter ces données pour répondre aux questions suivantes :
  1. Quelle est la proportion d'établissements par nombre d'étoiles ?
  2. Pour chaque catégorie (useful, funny, cool), quel utilisateur a apporté le plus de votes ?
  3. Quel est le restaurant le mieux noté pour chaque université ?
- Vous répondrez à ces questions en utilisant une technologie différente à chaque séance