Projet Hidoop

Projet de données réparties - Réponses HDFS

2020 - 2021

LAPLAGNE Chloé
RAZAFIMANANTSOA Nathan

Ce document présente les corrections et réponses à l'évaluation sur notre première version de HDFS, ainsi que les améliorations apportées.

- Correction de Bugs :

    - Contexte : *en cas de coupure de connexion avec un des serveurs, la suppression mettait bien les métadonnées à jour, laissant les fragments du fichier sur le serveur*.

      Ce bug a été corrigé, les métadonnées ne sont modifiées que si la connexion avec tous les serveurs requis a réussi.

- Pistes d'amélioration :

    - Proposition : *dans HdfsWrite, remplacer le paramètre 'taille d'un chunk' par 'nombre de chunks'*.

      Ici, lorsqu'un utilisateur écrit un fichier il connaît la taille de ce fichier. En revanche, il ne sait pas forcément combien de serveurs sont en ligne à un moment donné et ce nombre peut être variable.
      Indiquer (ici en octets) la taille des chunks permet de mieux se représenter ce qui sera traité par les opérations Map. C'est également le mode de fonctionnement utilisé par Hadoop et cela permet d'adapter par exemple une taille de chunk par défaut pour un format donné.
      Nous avons donc conservé ce choix mais nous avons essayé de le rendre plus clair dans le message d'aide.
      De plus, nous avons enlevé pour simplifier le mode 'distributed' qui permettait de répartir les chunks sur tous les serveurs. Ce mode était en effet surtout utile en debug et n'est plus très pertinent sur un fonctionnement classique.

- Implantation de la fiabilité dans les échanges client / serveur :

    - Mise en place de codes d'erreur codés sur un entier de type long. Ces codes sont échangés entre serveur et client et permettent d'indiquer le problème à l'utilisateur.
    - Affichage des codes d'erreur sur les communications ayant échoué entre client et serveur.