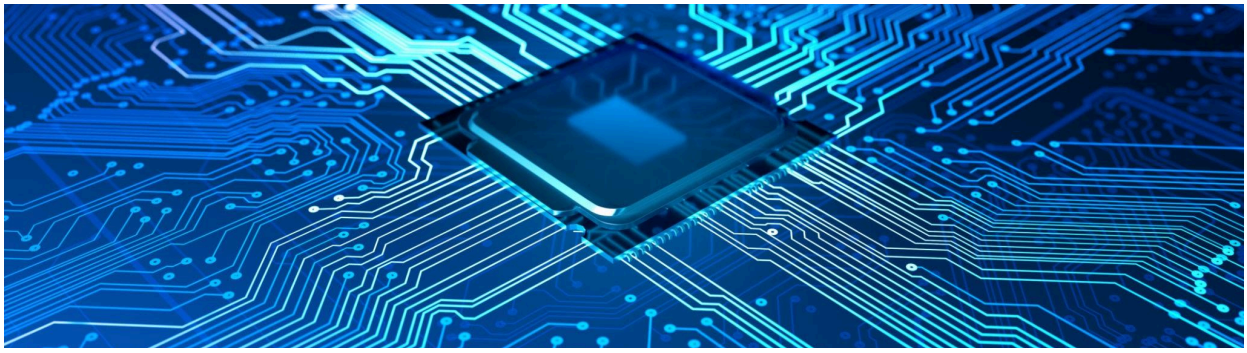

Status Not started

Cahier des charges du projet : Moteur d'Indexation et de Recherche de Documents Local (Enterprise Search)

Introduction

Ce présent cahier des charges a pour objectif de présenter le cadre général du projet ainsi que les éléments nécessaires à sa compréhension. Ainsi il décrit les besoins identifiés et la solution à proposer dans le cadre de ce projet.

1. Contexte

L'évolution rapide des technologies de stockage de l'information pourvoit les entreprises de volume gigantesque de données stockées dans des fichiers ou dans de gigantesques bases de données. Ainsi, il devient très vite difficile de retrouver des informations précises dans ces grands volumes de données.

Dans ce cadre, il devient utile d'utiliser de disposer d'outils informatique très rapide capable de scanner ces gros volumes de données à la recherche d'informations précises.

Ainsi l'introduction d'un outil de recherche pourrait contribuer à rendre la recherche de fichiers spécifiques plus rapide en offrant un gain en temps pour la recherche de fichiers.

2. Problématique :

En entreprise ou dans les structures, on connaît vite une augmentation exponentielle du volume des données. On fait rapidement face à des difficultés pour retrouver des fichiers. Dès lors, une question essentielle se pose : comment faciliter la recherche d'informations spécifiques sur de gros volumes de données.

3.Objectif général

Ce projet vise à terme de concevoir un moteur de recherche haute performance capable d'indexer et de retrouver des données textuelles dans une arborescence de fichiers avec des spécifications bien définies tout en garantissant une confidentialité des données utilisées et une consommation de ressource infime.

4.Objectif spécifiques

Les objectifs visés tout au long de ce projet sont de :

- Réduire le temps des recherches à l'ordre de 200 ms
- Créer un système de stockage des index directement sur le disque dure de l'ordinateur
- Avoir un exécutable unique minimisant au maximum les dépendances.
- Concevoir un logiciel avec une architecture modulaires(le logiciel doit être flexible)

5. Analyse et critique de l'existant

Solution	Points Forts	Points Faibles / Critiques

Windows Search / macOS Spotlight	Intégré à l'OS, interface visuelle.	Parfois très lent, indexation opaque, consomme beaucoup de CPU/Disque de manière imprévisible, peu flexible pour les développeurs.
Grep / Ack / Ag (The Silver Searcher)	Extrêmement puissant, pas d'indexation préalable.	Lent sur de très gros volumes (doit lire chaque fichier à chaque fois), pas adapté pour des recherches instantanées sur des To de données.
ElasticSearch (Local)	Puissance industrielle, fonctionnalités de recherche avancées.	Trop lourd pour un usage individuel ou une PME (nécessite Java, beaucoup de RAM), complexe à configurer et à maintenir.

6. Solutions proposées

1. Approche linéaire : scanner chaque fichier jusqu'à trouver le ou les fichiers recherchés
    - Avantages : Simple à coder,
    - Inconvénients : Devient vite inutilisable sur de gros volumes de données.
2. Base de données sql : stocker le texte des fichiers directement dans une base de donnée
    - Avantages : Gestion et requête sql facile à faire,
    - Inconvénients : Dépendance externe lourde, performances limitées par rapport à une structure de données optimisée "maison".
3. Index Inversé Custom (Structure de données dédiée) : Créer une table de hachage où chaque mot pointe vers une liste de documents et de positions
    - Avantages : Vitesse absolue, contrôle total sur la mémoire, apprentissage technique maximal,
    - Inconvénients : Plus complexe à implémenter (gestion des collisions, sérialisation sur disque).

7. Solution choisie

La solution choisie est celle de l'index inversé custom car elle permet d'exploiter d'autres horizons du langage et offre un défi algorithmique.

8. Résultats attendus

A la fin de ce projet, on s'attend à avoir un système disposant de :

- Un index stocké directement sur le disque dur;
- Une interface en ligne de commande CLI ou une interface graphique permettant d'utiliser l'outil;
- Une fonction de l'application permettant d'identifier les fichiers modifier et de les indexer;
- Un rapport sur les fichiers analysés etc… .

Conclusion

En entreprise plus précisément dans les PME, des difficultés sont rencontrées en matière de recherche sur des volumes gigantesques de données rendant difficile la recherche de fichiers précis. La mise en place d'une solution rapide et légère permettra de réduire les coûts en temps et la complexité de la recherche de fichiers spécifiques. Cette initiative vise à simplifier et à optimiser la recherche de fichier.