RAG et implementations possibles

Anas Lahrouchi, Pradityo WicaKsono
UFR IM2AG,
Université de Grenoble Alpes,
Anas.Lahrouchi@etu.univ-grenoble-alpes.fr

30 octobre 2023

1 Pourquoi a-t-on besoin d'un RAG ?

1.1 C'est quoi un RAG ?

Les grands modèles de langage (LLM) peuvent être inconsistant. En général, ils ont des réponses qui correspondent exactement aux requêtes, parfois ils ont des réponses qui sont compléments incorrects, on appelle ces cas-là des "hallucinations". Ce problème provient du fait que le dataset d'entrainement du LLM n'est pas à jour et que le modèle ne sait pas s'il a une réponse ou non.

Retreival Augmented Generation (RAG) est un Framework d'intelligence artificielle qui vise à améliorer la qualité des réponses générées par les LLM en fournissant un contexte se basant sur des données externes avec chaque requête pour compléter la représentation internes des informations du LLM. Implémenter RAG dans un système LLM a deux bénéfices : il assure que le modèle à accès aux faits actuelles les plus précises et que l'utilisateur a accès aux sources du modèle assurant ainsi que la validité et la précision des réponses peuvent être vérifier [1].

1.2 Comment fonctionne un LLM avec un RAG ?

Un RAG prend comme entrée la requête de l'utilisateur, il récupère un ensemble de documents pertinents et justificatifs depuis une base de connaissance. Les documents sont ensuite concaténés et fournis avec la requête en tant que contexte à LLM, ce dernier génère la réponse à la requête de l'utilisateur [2].

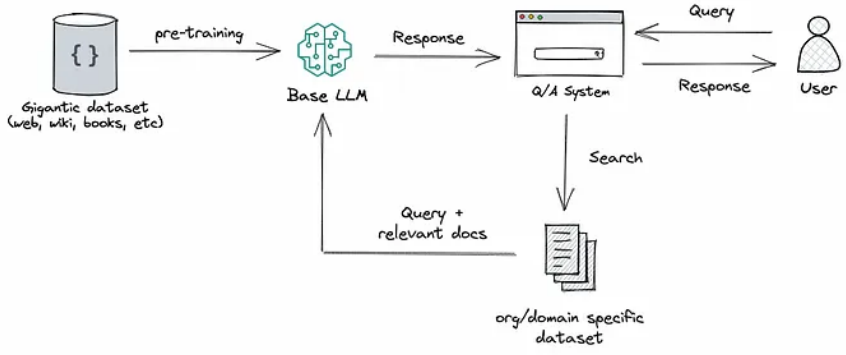

FIGURE 1 – une explication simple du Framework RAG

1. MARTINEAU 2023.
2. HOTZ 2023.

2 Quelles sont les implementations possibles de ce concept ?

Le Framework RAG a différentes implémentations, ca dépend des exigences et contraintes de l'application en question. Durant ce travail, Nous discutrons les implémentations suivantes :

2.1 RAG basée sur une récupération par mot clés

La première approche que nous allons discuter, la recherche par mots clés ou son extension la recherche par texte, est utilisée dans différents systèmes informatiques, tel que les navigateurs internet, les sites e-Commerce, les systèmes de gestion de fichiers. Il existe plusieurs implémentations de la recherche par mots clés, nous allons voir le cas du système open-source appelé ElasticSearch. ElasticSearch est un moteur de recherche et d'analyse distribué qui a comme noyau Apache Lucene développé en Java. Il nous permet de stocker, chercher et analyser des gros volumes de data rapidement et en temps réel. Alors comment fonctionne Elastic search ?

2.1.1 les stuctures de données utilisées

Les document JSON sont est l'unité informatique basique d'ElasticSearch. Un document peut être comparé à une ligne dans une base relationnelle. Chaque Document a un identifiant unique et un type de donnée décrivant son contenu.

Un index est une collection de documents aux caractéristique similaires, il s'agit du plus haut niveau d'entité sur laquelle on peut effectuer des requêtes. ElasticSearch utilise **des index inversés**. Ce mécanisme est à la source du fonctionnement de tous les moteurs de recherche, et associe un mapping du contenu à son emplacement dans un document ou un ensemble de documents. Cette structure de données hashmap-like permet de se diriger d'un mot vers un document [3].

2.1.2 Un système distibué

Un cluster est constitué d'un ou plusieurs instances de node connectées. Un node est un serveur qui appartient à un cluster, il stock les données, il les indexe et il effectue des recherches. Un index peut etre sous-divisé en fragment(shards), chaque fragment est un index indépendant et pleinement fonctionnel pouvant être hébergé sur n'importe quel noeud au sein d'un cluster [5].

3. GOPALAKRISHNAN 2023.
5. GOPALAKRISHNAN 2023.

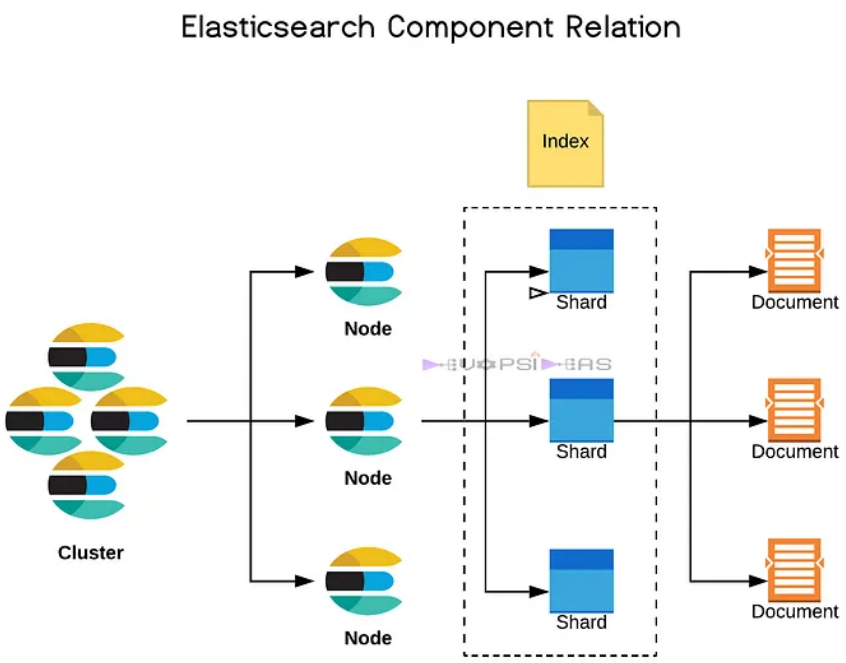

FIGURE 2 – Les composants du syste ElasticSearch [4]

Références

DEVOPSIDEAS (2022). *Different Elasticsearch components and what they mean*. `https://devopsideas.com/different-elasticsearch-components-and-what-they-mean-in-5-mins/`. Vue : 2023-10-27.

GOPALAKRISHNAN, Jay (2023). *Elasticsearch : What it is, How it works, and what it's used for*. `https://www.knowi.com/blog/what-is-elastic-search/`. Vue : 2023-10-27.

HOTZ, Heiko (2023). *RAG vs Finetuning — Which Is the Best Tool to Boost Your LLM Application ?* `https://towardsdatascience.com/rag-vs-finetuning-which-is-the-best-tool-to-boost-your-llm-application-94654b1eaba7`. Vue : 2023-10-27.

MARTINEAU, Kim (2023). *What is retrieval-augmented generation ?* `https://research.ibm.com/blog/retrieval-augmented-generation-RAG`. Vue : 2023-10-27.

Table des figures

6. DEVOPSIDEAS 2022.