**REPUBLIC OF CAMEROON**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**Peace-Work-Fatherland**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**MINISTRY OF HIGHER EDUCATION**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**UNIVERSITY OF NGAOUNDERE**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**REPUBLIQUE DU CAMEROUN**

**\*\*\*\*\*\*\*\*\*\*\*\*\***

**Paix-Travail-Patrie**

**\*\*\*\*\*\*\*\*\*\*\*\***

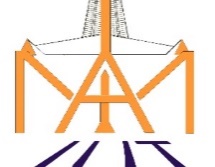
**MINISTERE DE L’ENSEIGNEMENT SUPERIEUR**

**\*\*\*\*\*\*\*\*\*\*\*\*\***

**UNIVERSITE DE NGAOUNDERE**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\***





INSTITUT UNIVERSITAIRE DE TECHNOLOGIE DE NGAOUNDERE

BP: 455 NGAOUNDERE Tel: 697 51 88 51/ /6 74 91 60 57

E-mail: iut@gmail.cm

DEPARTEMENT DE GENIE INFORMATIQUE

=

**Parcours : Génie Logiciel**

**TPE : BASES DE DONNÉES NON STRUCTURÉES**

**(NOSQL)**

Thème : Faire une implémentation de Bases de Données Orientées Objets Index Inversé

**Membres du groupe**

1. AMADOU JEAN ESAIE
2. MOHAMADOU BELLO
3. ILYASSOU ISSA

Enseignant : Dr. DASSI

Enseignant : M. HASSANA BAOURO

**Année Académique : 2022 / 2023**

# SOMMAIRE

[SOMMAIRE 0](#_Toc128527006)

[INTRODUCTION 1](#_Toc128527007)

[I. HISTORIQUE ET ÉVOLUTION 1](#_Toc128527008)

[II. ARCHITECTURE GLOBALE 1](#_Toc128527009)

[1. Définition de l'index inversé 1](#_Toc128527010)

[2. Qu'est-ce que l'index inversé ? 1](#_Toc128527011)

[3. Index inversé : exemple 2](#_Toc128527012)

[4. Index inversé et Index : à ne pas confondre 3](#_Toc128527013)

[5. Comment fonctionne l'Index inversé ? 3](#_Toc128527014)

[III. MODE DE DÉPLOIEMENT : LE CAS ELASTICSEARCH 4](#_Toc128527015)

[1. Qu'est-ce qu'un index Elasticsearch ? 4](#_Toc128527016)

[2. À quoi sert Logstash ? 4](#_Toc128527017)

[3. À quoi sert Kibana ? 5](#_Toc128527018)

[IV. AVANTAGES ET LIMITES 5](#_Toc128527019)

[1. Avantages 5](#_Toc128527020)

[2. Limites 5](#_Toc128527021)

[CONCLUSION 5](#_Toc128527022)

[RÉFÉRENCES 6](#_Toc128527023)

# INTRODUCTION

Un index inversé est une structure de données qui permet à un moteur de recherche de répertorier le contenu de son index classique de façon à pouvoir identifier rapidement lors d'une requête les pages web sur lesquelles les mots utilisés se trouvent. Une correspondance est donc faite entre chaque expression ou mot et sa position dans chaque document de l'index qui le contient. Cette stratégie d'indexation permet un gain de temps considérable.

# HISTORIQUE ET ÉVOLUTION

# ARCHITECTURE GLOBALE

## Définition de l'index inversé

L'index inversé fait partie des ressources qui permettent aux moteurs de recherche de fonctionner efficacement. Il est indispensable pour un traitement rapide des données lors d'une recherche plein texte.

## Qu'est-ce que l'index inversé ?

En informatique, on parle d'**index inversé** lorsqu'un contenu est répertorié ou classé de manière à ce que l'on puisse obtenir pour chaque mot trouvé dans un ensemble de pages la liste complète des documents où il est présent. Il est basé sur le même principe que l'index terminologique qu'on trouve dans les ouvrages scientifiques et littéraires, en complément de la table des matières.

La recherche d'un mot dans un long document prend du temps, même lorsque cette tâche est effectuée par un programme informatique. Imaginez à présent le temps que cela prendrait pour un moteur de recherche d'identifier des mots ou des expressions parmi plusieurs milliards de pages web. La lenteur du processus affecterait l'expérience utilisateur. C'est pour éviter cela que l'index inversé a été créé. Cette structure de données permet une **recherche rapide de l'information**, quelle que soit la fréquence du mot ou de l'expression recherchée dans la base de données. Google, Bing, Yahoo! et tous les principaux moteurs de recherche possèdent un index inversé.

Selon le niveau de précision souhaité, un index inversé peut prendre plusieurs formes. La forme la plus simple est celle qui fournit pour chaque mot d'un corpus la liste des documents dans lesquels il est présent au moins une fois. D'autres formes d'Index donnent en plus de cela **la position du mot ou de l'expression** en question dans les documents où il est présent. Les index inversés les plus avancés prennent également en compte la casse des mots et leur nombre (singulier ou pluriel).

Tout comme un [index classique](https://www.semjuice.com/definition/index), un index inversé contient beaucoup d'informations.

Théoriquement, il a donc besoin d'un espace disque très important qui peut parfois être plus élevé que celui occupé par le contenu indexé lui-même. Mais dans la pratique, avec une technologie similaire à celle du moteur Apache Lucene (entièrement codé en Java), il est possible d'avoir un index significativement plus petit en termes de volume que la base de données à laquelle IL est lié. Dans tous les cas, il faut dédier à l'index inversé au moins 10 % de l'espace requis pour stocker le contenu lui-même.

## Index inversé : exemple

Pour illustrer le concept d'index inversé, prenons l'exemple d'un corpus composé de deux documents (pages web) X et Y.

Le document X a pour contenu : « Le soleil est ardent »

Le document Y a pour contenu « Le soleil se couche à l'ouest »

Dans un index inversé simple, on obtiendra une liste de tous les mots du corpus ainsi que les documents où ils sont présents :

Le {X Y} ; Soleil { X Y }; Est { X }; Ardent { X }; Se { Y }; Couche { Y }; À { Y }; Ouest { Y }.

Si l'index inversé indique également la position des termes sur chaque page web, on aura des données plus précises sur la composition des documents :

Le {X,1 Y,1} (autrement dit, *Le* apparaît en première position dans les documents X et Y); Soleil { X,2 Y,2 }; Est { X,3 } Ardent { X,4 }; Se { Y,3 }; Couche { Y,4 }; À { Y,5 }; Ouest { Y,6 }.

Avec cette structure de données, un moteur peut rapidement identifier une page web pertinente pour les termes de recherche tapés par un internaute et générer des résultats. Comme vous pouvez le voir, la mise en place d'un index inversé n'est pas une mince affaire. **Il faut beaucoup de temps et de ressources pour le constituer et le mettre à jour**. La contrepartie de la rapidité qu'offre cette forme d'indexation est donc un temps d'insertion de données plus long. Il existe cependant certains programmes comme le moteur de recherche open source Apache Lucene qui peuvent implémenter rapidement des index inversés lorsqu'ils sont intégrés à une base de données plein texte.

## Index inversé et Index : à ne pas confondre

Il est très important de bien faire la distinction entre un index inversé et un index tout court, car ces deux expressions désignent deux choses différentes.

L'index d'un moteur de recherche est la base de données contenant l'ensemble des documents ou des pages que ses robots d'indexation crawlent et stockent. **C'est de cette copie du web constamment mise à jour que le moteur tire les informations lorsqu'un internaute lance une requête**. Un moteur de recherche peut avoir plusieurs index. Dans l'architecture de Google par exemple, il existe un index principal correspondant aux pages web les plus qualitatives et un index secondaire qui est réservé au contenu de second choix. Ce dernier n'est sollicité que lorsque l'utilisateur ne trouve pas l'information qu'il recherche parmi les pages « qualitatives ».

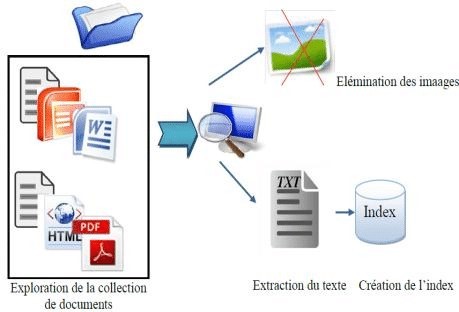
L'index inversé quant à lui est tout simplement un système qui vient compléter l'index classique afin d'accélérer le temps de traitement des documents. Les deux ressources interagissent constamment, mais leurs rôles ne sont pas interchangeables. Le [processus d'indexation](https://www.semjuice.com/blog/comment-fonctionne-lindexation-pages-google-m20) de l'information n'est également pas le même.

## Comment fonctionne l'Index inversé ?

Lorsqu'une requête est lancée par un utilisateur en quête d'information, les moteurs de recherches sollicitent systématiquement leur index inversé pour identifier en un laps de temps très court la liste des documents (de l'index) dans lesquels le mot-clé tapé est présent. Ensuite, un classement est réalisé grâce aux algorithmes en charge du calcul de pertinence, ce qui aboutit à la génération d'une page de résultats contenant l'information recherchée. **Les pages sont trouvées et affichées par ordre de pertinence décroissant**.

Lorsque la recherche d'une information porte sur une expression contenant plusieurs mots, le processus devient beaucoup plus complexe. Pour commencer, les moteurs identifient dans leur index inversé les pages web dont le texte contient chaque mot de l'expression recherchée et les confrontent pour trouver leur intersection (c'est-à-dire les documents dans lesquels tous les mots de l'expression sont présents). Si les termes constituant l'expression sont rares, alors cette approche suffit pour trouver une information pertinente, car l'intersection sera petite. Par contre, lorsque ce sont des termes fréquents, beaucoup d'autres analyses doivent être effectuées par les moteurs afin que des pages pertinentes pour la recherche puissent être identifiées. Dans certains cas, plusieurs index inversés peuvent même être utilisés lors d'une recherche.

Enfin, un index inversé peut également garder en mémoire tampon les requêtes les plus fréquentes afin de réduire le temps de traitement lorsqu'il est sollicité pour ce type de recherches.



Architecture globale de MORIC A.1 L'indexation Le processus de l’indexation permet d’explorer la collection de document et par la suite le traitement de chaque document. Pour chaque document on procède l’élimination des images et l’extraction du texte afin de le décomposer sous forme de tokens (termes) qui facilitent la création de l’index. Chaque document est analysé et indexé et stocké dans l’Index.[5]

# MODE DE DÉPLOIEMENT : LE CAS ELASTICSEARCH

## Qu'est-ce qu'un index Elasticsearch ?

Un index Elasticsearch est une collecte de documents en lien les uns avec les autres. Elasticsearch stocke des données sous forme de documents JSON. Chaque document met en corrélation un ensemble de clés (les noms des champs ou des propriétés) avec leurs valeurs correspondantes (chaînes, nombres, booléens, dates, choix de valeurs, géolocalisations ou autres types de données).

Elasticsearch utilise une structure de données appelée index inversé et conçue pour faire des recherches full-text très rapides. Un index inversé liste chaque mot unique qui apparaît dans n'importe quel document et identifie tous les documents dans lesquels chaque mot apparaît.

Au cours du processus d'indexation, Elasticsearch stocke des documents et fabrique un index inversé pour permettre de rechercher les données du document en temps quasi réel. L'indexation commence par l'index API, avec lequel vous pouvez ajouter ou mettre à jour un document JSON dans un index spécifique.

## À quoi sert Logstash ?

Logstash, l'un des produits de base de la Suite Elastic, sert à agréger et à traiter des données pour les envoyer dans Elasticsearch. Logstash est un pipeline côté serveur en open source destiné au traitement des données, qui vous permet d'ingérer simultanément des données provenant de multitude de sources, puis de les transformer avant qu'elles soient indexées dans Elasticsearch.

## À quoi sert Kibana ?

Kibana est un outil de visualisation et de gestion de données pour Elasticsearch qui intègre des histogrammes en temps réel, des graphes linéaires, des camemberts et des cartes. Kibana intègre également des applications comme Canvas, qui permet aux utilisateurs de créer des infographies dynamiques personnalisées basées sur leurs données, et sur Elastic Maps pour visualiser des données géospatiales.

# AVANTAGES ET LIMITES

## Avantages

## Limites

L'implémentation de la méthode index inversé est onéreuse car elle est un peu lente au moment de l’indexation et qu’elle implique de devoir refaire le travail d’indexation si on ajoute/supprime des documents. Donc, même si on réalise une économie de temps pour les recherches, l’indexation prend un peu plus de temps, et il faut reconstituer l’index si on supprime ou si on ajoute des documents.

# CONCLUSION

Nous avons appris à construire un index inversé, et nous avons réalisé une petite implémentation en Javascript. L’index inversé est particulièrement utile lorsque l’on souhaite indexer des documents et faciliter ainsi les recherches.

# RÉFÉRENCES

* <https://mylittleneuron.com/2023/01/21/gestion-des-donnees-lindex-inverse/>
* [simon\_chabrol](https://mylittleneuron.com/author/simonchabrol/) [Non classé](https://mylittleneuron.com/category/non-classe/) 21 janvier 2023 4 Minutes
* https://www.elastic.co/fr/what-is/elasticsearch