



Étude de cas

Distribution des données avec Elasticsearch

Loïc Herman et Jonas Troeltsch

11 juin 2024

HEIG-VD — MAC

Plan

- 1 Un système distribué par défaut
 - 1.1 Architecture
 - 1.2 Mécanismes : disaster recovery
 - 1.3 Mécanismes : gestion des nœuds et équilibrage
 - 1.4 Mécanismes : gestion des répliques
 - 1.5 Tolérance au partitionnement
 - 1.6 Service discovery
- 2 Une refonte récente en prévention des divergences
 - 2.1 Un besoin de restructuration
 - 2.2 Ingestion des documents
 - 2.3 Système de prévention
 - 2.4 Problèmes évités
- 3 Bibliographie

Un système distribué par défaut

Architecture

Elasticsearch utilise une architecture sans partage, un cluster contiendra un nœud master et un système de coordination entre les nœuds pour coordonner les opérations.

- Chaque nœud gère ses processeurs, RAM et disques
- Un nœud contient des shards et fait automatiquement partie d'un cluster
- Chaque shard est une instance d'Apache Lucene

Les nœuds et leurs responsabilités

- **Nœud master** : création/suppression d'indexes, management du cluster, connaît l'emplacement des documents
- **Nœud de données** : indexation, recherche, suppression et autres opérations liée aux documents
- **Nœud de coordination** : agit en tant que gestionnaire de travail, distribuant les requêtes entrantes aux nœuds appropriés et répondant au client

Partitionnement et réplique

- Chaque index est séparé équitablement en shards sur les nœuds
- Une shard primaire peut avoir des répliques qui ne font que des lectures

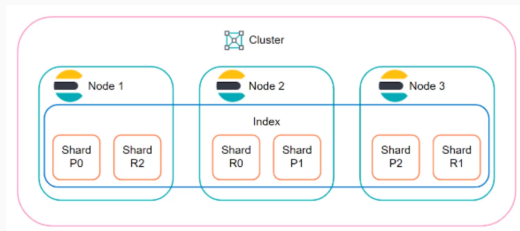


Figure 1 : Shards et répliques dans un cluster

Elasticsearch définit trois agrégations pour représenter l'état du système :

1. **RED** : Certains nœuds possèdent des shards qui ne sont pas assignées,
2. **YELLOW** : Les shards sont toutes assignées et prêtes, mais pas les replicas,
3. **GREEN** : Toutes les shards et les replicas sont assignées et prêtes à l'emploi.

Elasticsearch utilise un modèle de partitionnement par hachage de l'ID pour déterminer sur quelle shard un document doit aller.

$$S_n = \text{hash}(\text{doc}_{\text{id}}) \bmod N_{\text{shards}}$$

Alternative : architecture stateless

- Introduite en 2022
- Séparation du stockage des données de l'indexation et de la recherche.
- 2 types de nœuds
 - nœud d'index
 - nœud de recherche
- Permet l'utilisation d'un service de stockage externe (S3, GCS, Azure Blob Storage)

Alternative : architecture stateless

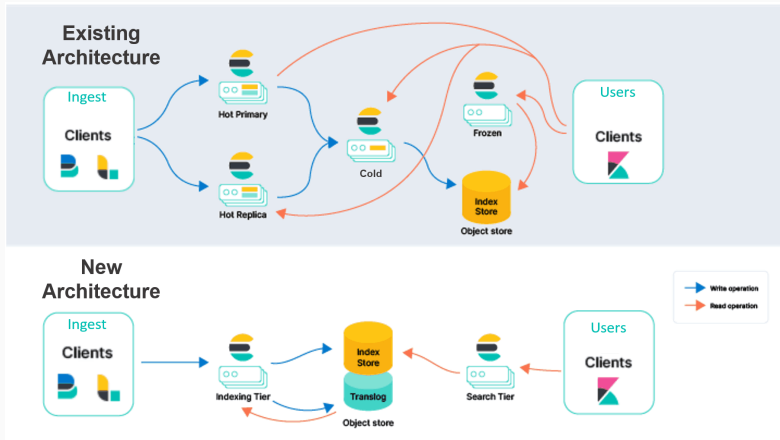


Figure 2 : Comparaison entre l'architecture à disques partagés et l'architecture classique

Un système distribué par défaut

Mécanismes : disaster recovery

Elasticsearch implémente un modèle simple de disaster recovery

- **Panne d'un nœud data** : une des répliques de chaque shards primaires sera élue primaire
- **Panne d'un nœud master** : élection d'un nouveau nœud master par les nœuds éligibles
- **Panne d'un cluster entier** : possibilité de failover sur un cluster de copie qui reprend la charge

Elasticsearch implémente un modèle simple de disaster recovery

- **Panne d'un nœud data** : une des répliques de chaque shards primaires sera élue primaire
- **Panne d'un nœud master** : élection d'un nouveau nœud master par les nœuds éligibles
- **Panne d'un cluster entier** : possibilité de failover sur un cluster de copie qui reprend la charge

Elasticsearch implémente un modèle simple de disaster recovery

- **Panne d'un nœud data** : une des répliques de chaque shards primaires sera élue primaire
- **Panne d'un nœud master** : élection d'un nouveau nœud master par les nœuds éligibles
- **Panne d'un cluster entier** : possibilité de failover sur un cluster de copie qui reprend la charge

Un système distribué par défaut

Mécanismes : gestion des nœuds et équilibrage

Ajout d'un nœud à un cluster

En cas d'ajout d'un nœud, Elasticsearch va automatiquement commencer à équilibrer la charge de données entre tous les nœuds présents dans le cluster.

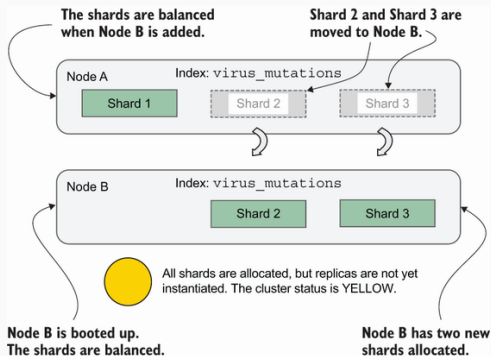


Figure 3 : Équilibrage automatique lors de l'ajout d'un nœud

Déploiement des répliques

Une fois les shards « primaires » initialisées, les répliques peuvent être ajoutées.

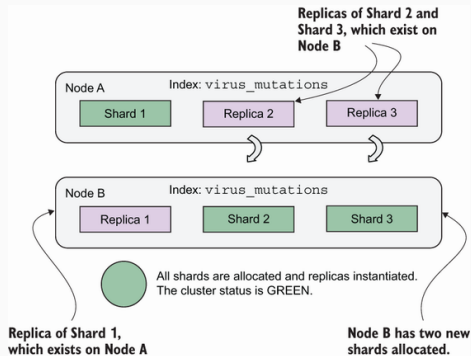


Figure 4 : Allocation des répliques avec le nœud additionel

Un système distribué par défaut

Mécanismes : gestion des répliques

Initialisation d'une nouvelle réplique

Lors de la création d'une nouvelle réplique, un processus standard est utilisé :

1. Snapshot de la shard « primaire » avec le *sequence number* associé
2. Copie de la snapshot dans la réplique
3. Mise à jour de la copie selon la différence du *sequence number* actuel de la shard « primaire »

Initialisation d'une nouvelle réplique

Lors de la création d'une nouvelle réplique, un processus standard est utilisé :

1. Snapshot de la shard « primaire » avec le *sequence number* associé
2. Copie de la snapshot dans la réplique
3. Mise à jour de la copie selon la différence du *sequence number* actuel de la shard « primaire »

Initialisation d'une nouvelle réplique

Lors de la création d'une nouvelle réplique, un processus standard est utilisé :

1. Snapshot de la shard « primaire » avec le *sequence number* associé
2. Copie de la snapshot dans la réplique
3. Mise à jour de la copie selon la différence du *sequence number* actuel de la shard « primaire »

Ajout de shards et répliques

- Il n'est **pas possible** d'augmenter le nombre de shards d'un index **actif**.
- ⇒ Le nombre de shards doit donc être spécifié **à la création** d'un index (par défaut 1 shard et 1 réplique).
- Il est néanmoins possible de changer le nombre de répliques de chaque shards via la configuration de l'index.
- ⇒ Les nouvelles répliques seront automatiquement **dispersées** dans les nœuds du cluster de sorte à garantir une **tolérance aux pannes**.

Ajout de shards et répliques

- Il n'est **pas possible** d'augmenter le nombre de shards d'un index **actif**.
- ⇒ Le nombre de shards doit donc être spécifié **à la création** d'un index (par défaut 1 shard et 1 réplique).
- Il est néanmoins possible de changer le nombre de répliques de chaque shards via la configuration de l'index.
- ⇒ Les nouvelles répliques seront automatiquement **dispersées** dans les nœuds du cluster de sorte à garantir une **tolérance aux pannes**.

Ajout de shards et répliques

- Il n'est **pas possible** d'augmenter le nombre de shards d'un index **actif**.
- ⇒ Le nombre de shards doit donc être spécifié à la **création** d'un index (par défaut 1 shard et 1 réplique).
- Il est néanmoins possible de changer le nombre de répliques de chaque shards via la configuration de l'index.
- ⇒ Les nouvelles répliques seront automatiquement **dispersées** dans les nœuds du cluster de sorte à garantir une **tolérance aux pannes**.

Ajout de shards et répliques

- Il n'est **pas possible** d'augmenter le nombre de shards d'un index **actif**.
- ⇒ Le nombre de shards doit donc être spécifié à la **création** d'un index (par défaut 1 shard et 1 réplique).
- Il est néanmoins possible de changer le nombre de répliques de chaque shards via la configuration de l'index.
- ⇒ Les nouvelles répliques seront automatiquement **dispersées** dans les nœuds du cluster de sorte à garantir une **tolérance aux pannes**.

Un système distribué par défaut

Tolérance au partitionnement

Garanties au sens du théorème de CAP

- Elasticsearch se caractérise par la réplication et la coordination entre les nœuds d'un cluster.
 - Les données sont toujours répliquées sur plusieurs shards, idéalement selon la configuration sur plusieurs nœuds.
- ⇒ La tolérance au partitionnement de l'infrastructure est toujours garantie.

Garanties au sens du théorème de CAP

- Elasticsearch se caractérise par la réplication et la coordination entre les nœuds d'un cluster.
- Les données sont toujours répliquées sur plusieurs shards, idéalement selon la configuration sur plusieurs nœuds.

⇒ La tolérance au partitionnement de l'infrastructure est toujours garantie.

Garanties au sens du théorème de CAP

⇒ La tolérance au partitionnement de l'infrastructure est toujours garantie.

- Toutefois, Elasticsearch traite de deux manières différentes les accès :

1. **En cas d'écriture** : la cohérence est privilégiée à la disponibilité.

- Une écriture acquittée ne devrait jamais être perdue
- Mais, un nœud rompu du système aura toujours les écritures bloquées

⇒ La disponibilité en écriture n'est pas garantie en cas de partition du système

2. **En cas de lecture** : la disponibilité est privilégiée à la cohérence

⇒ Il y a une plus forte garantie en disponibilité à la lecture, en échange d'une pénalité de cohérence

Garanties au sens du théorème de CAP

⇒ La tolérance au partitionnement de l'infrastructure est toujours garantie.

- Toutefois, Elasticsearch traite de deux manières différentes les accès :

1. **En cas d'écriture** : la cohérence est privilégiée à la disponibilité.

- Une écriture acquittée ne devrait jamais être perdue
- Mais, un nœud rompu du système aura toujours les écritures bloquées

⇒ La disponibilité en écriture n'est pas garantie en cas de partition du système

2. **En cas de lecture** : la disponibilité est privilégiée à la cohérence

⇒ Il y a une plus forte garantie en disponibilité à la lecture, en échange d'une pénalité de cohérence

Garanties au sens du théorème de CAP

- ⇒ La tolérance au partitionnement de l'infrastructure est toujours garantie.
- Toutefois, Elasticsearch traite de deux manières différentes les accès :
 1. **En cas d'écriture** : la cohérence est privilégiée à la disponibilité.
 - ⇒ La disponibilité en écriture n'est pas garantie en cas de partition du système
 2. **En cas de lecture** : la disponibilité est privilégiée à la cohérence
 - Selon la configuration, une recherche peut privilégier de retourner un vieux résultat plutôt que de retourner une erreur
 - ⇒ Il y a une plus forte garantie en disponibilité à la lecture, en échange d'une pénalité de cohérence

Garanties au sens du théorème de CAP

- ⇒ La tolérance au partitionnement de l'infrastructure est toujours garantie.
- Toutefois, Elasticsearch traite de deux manières différentes les accès :
 1. **En cas d'écriture** : la cohérence est privilégiée à la disponibilité.
 - ⇒ La disponibilité en écriture n'est pas garantie en cas de partition du système
 2. **En cas de lecture** : la disponibilité est privilégiée à la cohérence
 - Selon la configuration, une recherche peut privilégier de retourner un vieux résultat plutôt que de retourner une erreur
 - ⇒ Il y a une plus forte garantie en disponibilité à la lecture, en échange d'une pénalité de cohérence

Garanties au sens du théorème de CAP

- ⇒ La tolérance au partitionnement de l'infrastructure est toujours garantie.
- Toutefois, Elasticsearch traite de deux manières différentes les accès :
 1. **En cas d'écriture** : la cohérence est privilégiée à la disponibilité.
 - ⇒ La disponibilité en écriture n'est pas garantie en cas de partition du système
 2. **En cas de lecture** : la disponibilité est privilégiée à la cohérence
 - ⇒ Il y a une plus forte garantie en disponibilité à la lecture, en échange d'une pénalité de cohérence
- On peut malgré tout définir Elasticsearch comme étant un système principalement Consistent-Partition (CP) au sens de la conjecture de Brewer, car la pénalité de cohérence en lecture est inférée par le blocage des écritures en cas de partitionnement.

Un système distribué par défaut

Service discovery

Elasticsearch prend en charge plusieurs types de *service discovery* :

- **Seed-based discovery** : une liste d'adresses IP et/ou une liste de noms de domaines dont les adresses liées seront résolues unes-à-unes
- **Cloud-based discovery** : des plugins sont disponibles pour connecter le service discovery avec les services de cloud providers (AWS EC2, Azure Classic, Google Compute Engine)

Ces listes d'adresses seront ensuite évaluées et si un nœud elastic est présent alors il sera rajouté dans le cluster avec la gestion d'état partagée.

Une refonte récente en prévention des divergences

Un besoin de restructuration

Un petit peu d'histoire

La version 7.0 d'Elasticsearch a été une version plus que majeure, comprenant une série d'améliorations conséquentes quant au système de **coordination des clusters** et l'ajout de *log sequence numbers* sur toutes les opérations d'écriture.

Ces changements ont par la suite permis d'offrir le système de réindexation permettant de cloner un index vers un autre en appliquant éventuellement une mutation.

Dans les composantes internes, cela a pu permettre d'optimiser le temps de récupération des shards désynchronisées en permettant d'éviter de devoir sourcer les fichiers sous-jacents.

Ces changements ont aussi pu permettre de limiter l'impact qu'avaient anciennement les partitionnements réseau, qui pouvaient causer des **divergences**, **pertes d'écriture** et des **dirty reads**.

Nous allons donc nous concentrer plus en détail sur la façon dont ces changements ont pu permettre d'améliorer la gestion de ces cas d'erreurs, principalement dans le cadre des divergences.

Une refonte récente en prévention des divergences

Ingestion des documents

Un traitement en deux parties

Elasticsearch possède deux pipelines d'accès aux données :

1. **Les écritures** passent via chaîne de nœuds par un processus synchrone avant d'être traité indépendamment dans la shard primaire et ensuite, parallèlement, ses répliques,
2. **Les lectures** se font en parallèle sur chaque shard qui contient les données pertinentes, chaque nœud du système peut recevoir une demande et sera responsable d'agréger les résultats retournés par toutes les shards pertinentes

Pipeline d'ingestion

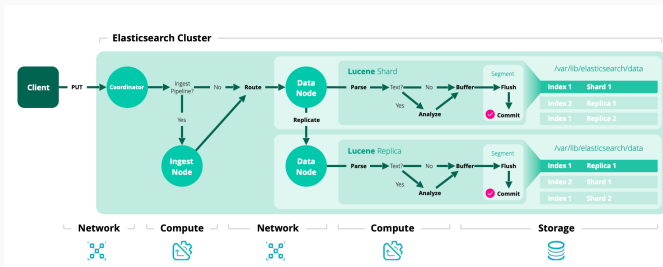


Figure 5 : Étapes lors de la réception des documents

Les opérations sont d'abord validées par la shard « primaire » routée pour le document, ensuite l'opération est effectuée par cette dernière. Les répliques vont ensuite effectuer la même opération en parallèle pour conserver les copies.

Une refonte récente en prévention des divergences

Système de prévention

Il existe donc trois opérations « modificatrices » sur un index :

1. **Indexation** : ajouts de nouveaux documents
2. **Updation** : mise à jour des documents
3. **Deletion** : suppression de documents

Jusqu'à la version 7.0, Elasticsearch ne possédait pas de système pour tracer l'ordre d'exécution de ces trois types d'opérations lors de la copie dans les répliques.

Il a donc fallu rajouter un système de ***log sequence numbers*** à ces opérations d'écriture pour pouvoir détecter les différences de traitement en cas de partitionnement.

Il existe donc trois opérations « modificatrices » sur un index :

1. **Indexation** : ajouts de nouveaux documents
2. **Updation** : mise à jour des documents
3. **Deletion** : suppression de documents

Jusqu'à la version 7.0, Elasticsearch ne possédait pas de système pour tracer l'ordre d'exécution de ces trois types d'opérations lors de la copie dans les répliques.

Il a donc fallu rajouter un système de *log sequence numbers* à ces opérations d'écriture pour pouvoir détecter les différences de traitement en cas de partitionnement.

Définition d'une séquence

Deux nombres sont associés à chaque opération :

- **Terme primaire** : Incrémenté à chaque nouvelle élection d'une shard primaire, déterminé par le nœud master.
- **Numéro de séquence (seq#)** : Incrémenté par la shard primaire pour chaque opération.

Protocole d'ordonnancement de deux opérations :

- $o1 < o2$ si $s1.seq\# < s2.seq\#$ ou ($s1.seq\# == s2.seq\#$ et $s1.term < s2.term$).

Définition d'une séquence

Variables utilisées pour déterminer si une shard est correctement synchronisée :

- **Local checkpoint#** : Plus haut seq# pour lequel tous les seq# inférieurs ont localement été traités.
- **Global checkpoint#** : Plus haut seq# traité sur toutes les répliques actives.

Chaque shard maintient en mémoire et dans les métadonnées de chaque commit de Lucene ces deux numéros, permettant de gérer la coordination lors de la réplication des données.

Une refonte récente en prévention des divergences

Problèmes évités

Indexation pendant un partitionnement

Lorsqu'une shard « primaire » est isolée du cluster, elle continue à indexer localement.

L'isolement est découvert uniquement lors de la tentative de réplication et la shard ne va donc jamais acquitter l'écriture du document, en attendant une résolution en amont. Celle-ci peut engendrer des problèmes d'incohérences si la résolution ne peut plus déterminer l'ordre d'exécution à utiliser.

Solution : L'infrastructure mise en place permet une identification unique des documents en utilisant les champs de terme primaire et de numéro de séquence, permettant de mieux résoudre un conflit lié à un partitionnement réseau.

Désynchronisation des répliques

Lorsqu'une shard primaire échoue, une réplique sera promue en primaire. Lors de la promotion, les autres répliques présentes pourraient devenir désynchronisées si des opérations en cours d'envoi ont été perdues lors de la partition.

Problème : Ces désynchronisations ne sont pas corrigées lors de la promotion de la shard primaire, mais sont différées jusqu'à la relocalisation des répliques (nécessitant une relecture des fichiers sous-jacents), rendant la période de désynchronisation indéfinie.

Solution : Les numéros de séquence permettent d'identifier les divergences entre répliques au niveau du document, facilitant la synchronisation efficace des répliques restantes avec le nouveau primaire.



Discovery and cluster formation.

<https://l8n.ch/q4wP0>.

[Dernière consultation le 09.06.2024].



Reading and writing documents.

<https://l8n.ch/LH3G4>.

[Dernière consultation le 09.06.2024].



Scalability and resilience : clusters, nodes, and shards.

<https://l8n.ch/JA9Tc>.

[Dernière consultation le 09.06.2024].



Set up a cluster for high availability.

<https://l8n.ch/RyX7W>.

[Dernière consultation le 09.06.2024].



Size your shards.

<https://l8n.ch/cxRJm>.


[Dernière consultation le 09.06.2024].




C. Gormley and Z. Tong.

Elasticsearch : The Definitive Guide.

O'Reilly Media, Inc., 1st edition, 2015.

 M. Kleppmann.
Designing Data-Intensive Applications : The Big Ideas Behind Reliable, Scalable, and Maintainable Systems.
O'Reilly Media, Inc., 1st edition, 2017.

 M. Konda.
Elasticsearch in Action, Second Edition.
In Action. Manning, 2nd edition, 2023.

“You allow things to be inconsistent and then you find ways to compensate for mistakes, versus trying to prevent mistakes altogether.”

— Eric Brewer