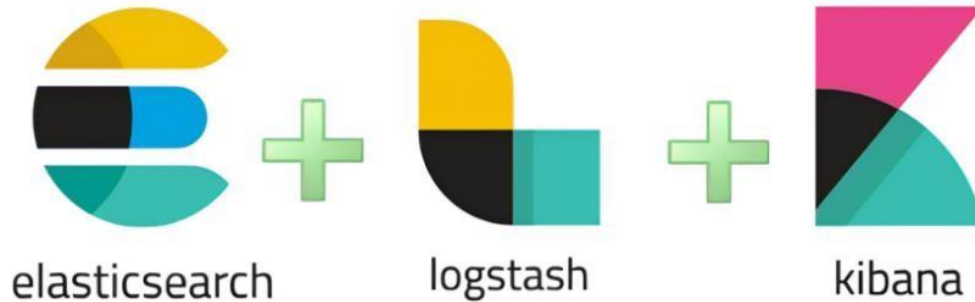


## Chapitre3 ElasticSearch

**Mr DIATTARA Ibrahima**



# Elasticsearch

Elasticsearch est un **moteur de recherche** et d'analyse distribué en temps réel, Il est utilisé pour:

- Recherche full text
- Recherche structurée
- L'analyse

<b>Relational DB</b>	Base de données	Tables	Lignes	Colonnes
<b>Mongo DB</b>	Base de données	Collections	Documents	Champs
<b>Elasticsearch</b>	Index	Types	Documents	Champs

Un Document possède un type (qui définit son mapping) et chaque document est relié par un id

# Démarrage

Deux ports par défaut:

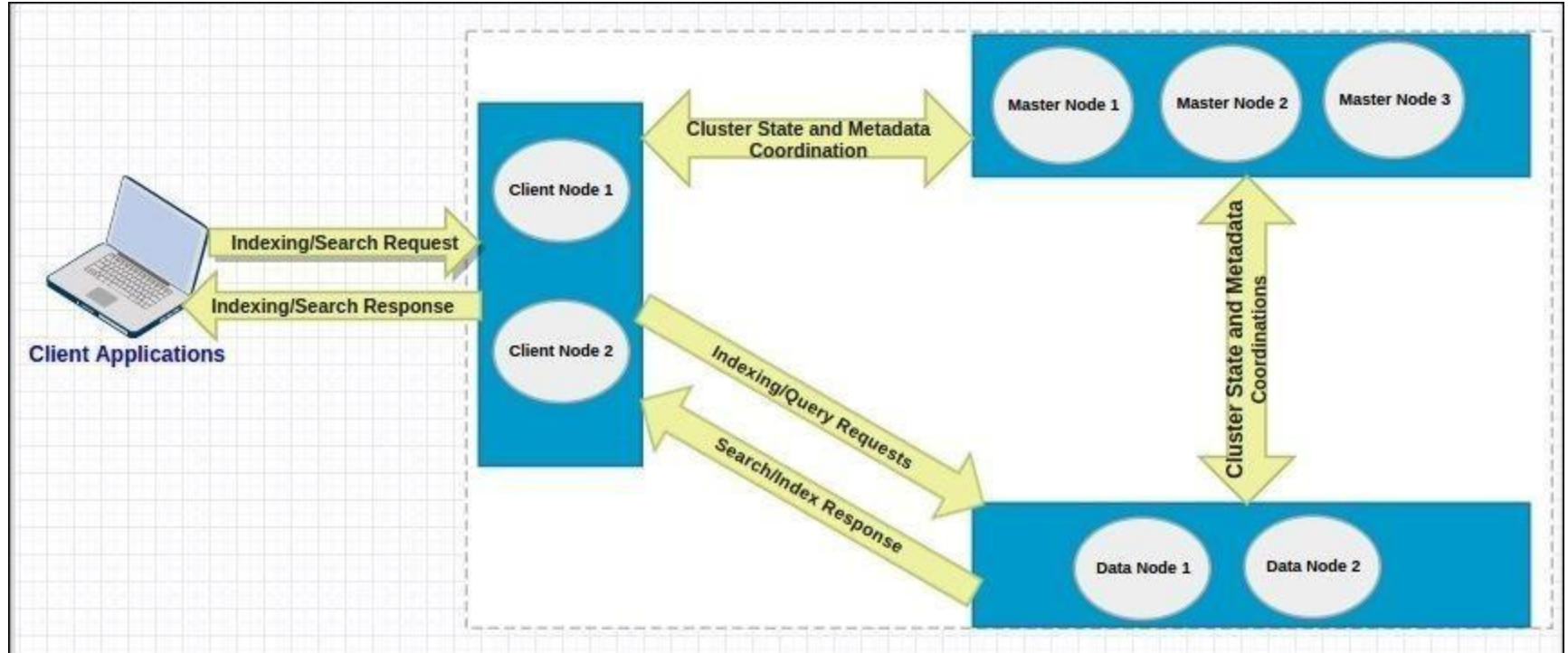
- 9200 ⇒ http (requête / ingestion)
- 9300 ⇒ transport (inter-node communications)

```
← → ↻ ⚠ Not secure | http://mosef02.westeurope.cloudapp.azure.com:9200

{
  "name" : "elk-mosef1",
  "cluster_name" : "elasticsearch",
  "cluster_uuid" : "HPfHb2S2RZ6YBjwZtX1NoQ",
  "version" : {
    "number" : "7.16.3",
    "build_flavor" : "default",
    "build_type" : "deb",
    "build_hash" : "4e6e4eab2297e949ec994e688dad46290d018022",
    "build_date" : "2022-01-06T23:43:02.825887787Z",
    "build_snapshot" : false,
    "lucene_version" : "8.10.1",
    "minimum_wire_compatibility_version" : "6.8.0",
    "minimum_index_compatibility_version" : "6.0.0-beta1"
  },
  "tagline" : "You Know, for Search"
}
```

<http://mosef02.westeurope.cloudapp.azure.com:9200>

# Types des noeuds Elasticsearch



# Types des noeuds Elasticsearch

## Master Node

- responsable des actions légères à l'échelle du cluster, telles que la création ou la suppression d'un index, le suivi des nœuds faisant partie du cluster et le choix des shards à allouer à quels nœuds.
- Ne stocke pas de données
- Config : (elasticsearch.yml)
  - Elasticsearch version 7
    - `node.roles: [ master ]`

## Data Node

- Les nœuds "data" effectuent des opérations liées aux données telles que CRUD, recherche et agrégations.
- stocke les données
- Config: (elasticsearch.yml)
  - Elasticsearch version 7
    - `node.roles: [ data ]`

# Types des noeuds Elasticsearch

## Client Node (coordinating node)

- Routage des requetes et load balancer
- Ne stocke pas de données
- Config : (elasticsearch.yml)
  - Elasticsearch version 7
    - node.roles: [ ]

PS : d'autres rôles ont été ajouté récemment dans les dernières versions d'Elasticsearch , tels que : ml , ingest ...

<https://www.elastic.co/guide/en/elasticsearch/reference/current/modules-node.html>

# Exemple

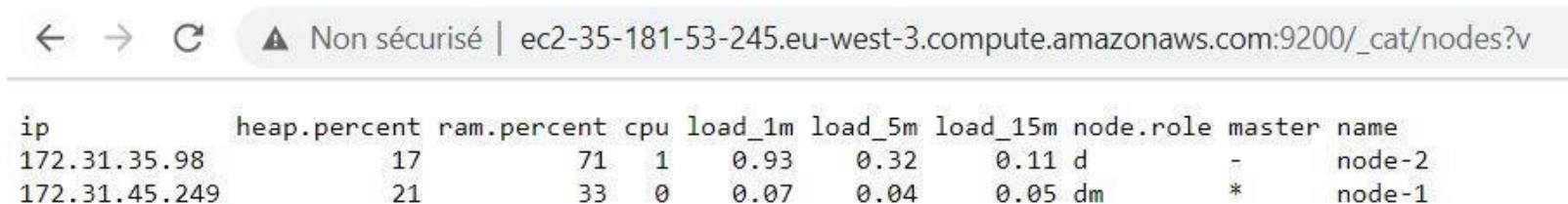
Comment lister les noeuds du cluster Elasticsearch?

~ l'API `/_cat/nodes?v`



A screenshot of a web browser window. The address bar shows a URL starting with 'http://mosef02.westeurope.cloudapp.azure.com:9200/\_cat/nodes?v'. The page content displays a table with 11 columns: ip, heap.percent, ram.percent, cpu, load\_1m, load\_5m, load\_15m, node.role, master, and name. There is one row of data.

ip	heap.percent	ram.percent	cpu	load_1m	load_5m	load_15m	node.role	master	name
10.0.0.9	34	97	2	0.03	0.09	0.04	cdfhilmrstw	*	elk-mosef1



A screenshot of a web browser window. The address bar shows a URL starting with 'ec2-35-181-53-245.eu-west-3.compute.amazonaws.com:9200/\_cat/nodes?v'. The page content displays a table with 11 columns: ip, heap.percent, ram.percent, cpu, load\_1m, load\_5m, load\_15m, node.role, master, and name. There are two rows of data.

ip	heap.percent	ram.percent	cpu	load_1m	load_5m	load_15m	node.role	master	name
172.31.35.98	17	71	1	0.93	0.32	0.11	d	-	node-2
172.31.45.249	21	33	0	0.07	0.04	0.05	dm	*	node-1

# Index Elasticsearch

Les **documents JSON** sont stockés dans un ou plusieurs **index** Elasticsearch.

Un **index** est l'équivalent d'**une table** SQL.

Chaque index peut être raffiné avec la notion de **type**, correspondant à une sous-catégorie de l'index que l'on pourra spécifier au besoin. Tous les types d'un index partagent le même **schéma** de documents JSON.



# Les Shards

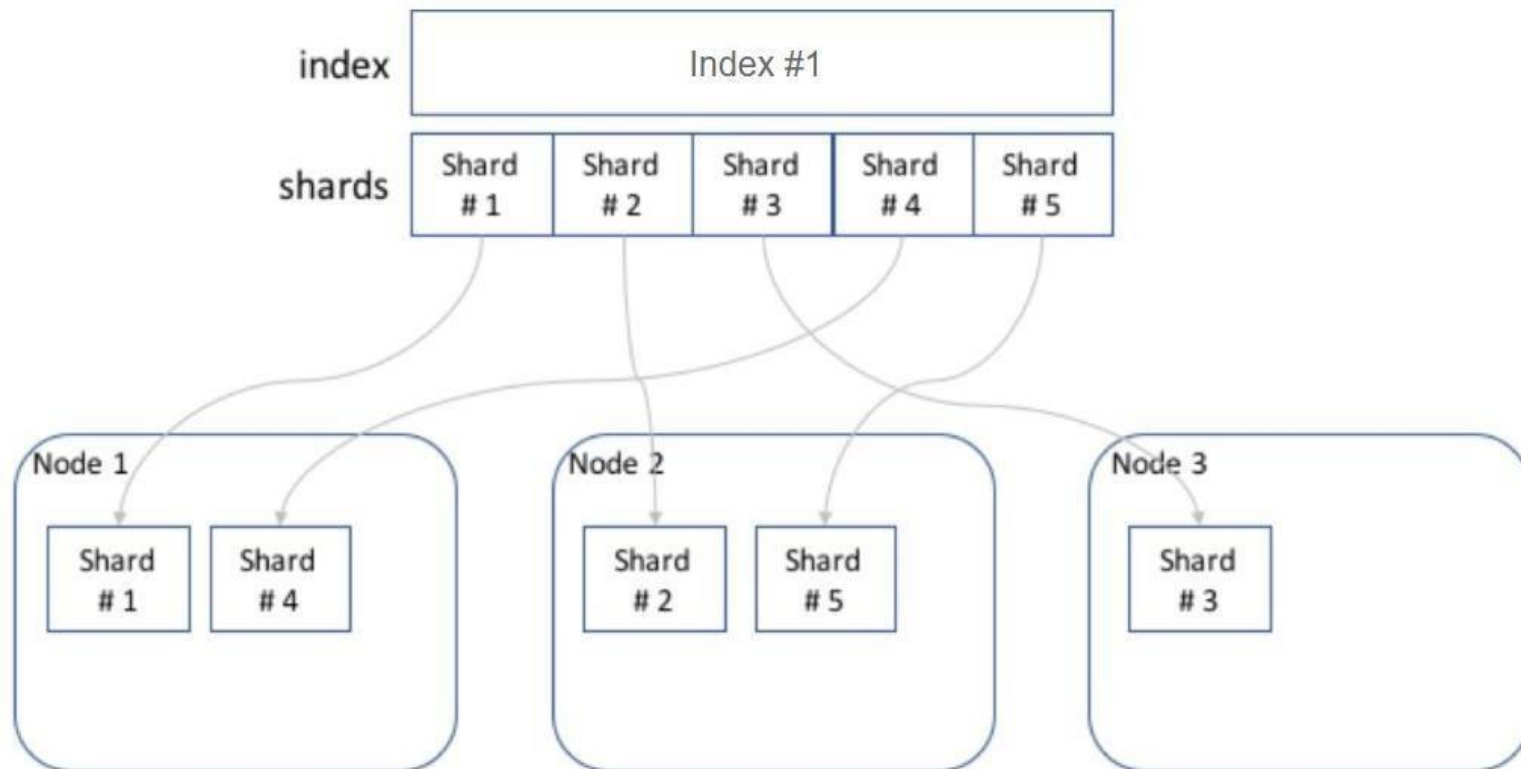
Un index Elasticsearch est **partitionné en un ou plusieurs shards**

2 types de shards:

- **Primary** : shard primaire
- **Replicat** : shard répliat (une copie du shard primaire)

Le fait d'avoir des shards “replicat” nous permet d'avoir un cluster Elasticsearch **résilient aux pannes**

# Index / Shard



# Définition

## ❑ Index

- Un peu comme une base de données sur un SGBDR relationnel
- Une collection de document qui ont tous un points commun (

## ❑ Type Mapping

- Le mapping est similaire au schéma du type
- Le mapping peut être défini manuellement, mais aussi généré automatiquement quand les documents sont indexés

## ❑ Shard

- Découper un index en plusieurs parties pour y distribuer les documents
- C'est l'équivalent des partitions dans un SGBDR
- Nos documents sont stockés et indexés dans les Shards, mais nous ne nous adressons pas directement à eux : nos applications s'adressent à un index

## ❑ Réplica

- Recopie d'un shard en une ou plusieurs copie dans l'ensemble du cluster
- Un Shard replica : est une copie d'un Shard primaire (similaire au RAID 1)

## ❑ Alias

- C'est l'équivalent d'une vue dans le monde SGBDR
- Un alias ElasticSearch peut être configuré de manière à pointer vers un ou plusieurs indexes d'un cluster tout en spécifiant des filtres ou des clés de routage

# Exemple

Comment lister les index Elasticsearch dans l'ordre décroissant ?

~ l'API `/_cat/indices?v`

[http://mosef02.westeurope.cloudapp.azure.com:9200/\\_cat/indices?v=true&s=store.size:desc](http://mosef02.westeurope.cloudapp.azure.com:9200/_cat/indices?v=true&s=store.size:desc)

← → ↻ ⚠ Not secure   http://mosef02.westeurope.cloudapp.azure.com:9200/_cat/indices?v=true&s=store.size:desc									
health	status	index	uuid	pri	rep	docs.count	docs.deleted	store.size	pri.store.size
yellow	open	maas_cheikhyakhoub_one_shot	kE4c5IxuT6yTnz69g7V2bQ	1	1	32000	96000	42.9mb	42.9mb
green	open	.geopip_databases	Tn5GovQoTdamj-PAF2RzUA	1	0	41	38	38.9mb	38.9mb
yellow	open	ibrahima_camara_streaming	Zn04GwGrRjue9IdTQkCfDA	2	1	77184	0	25mb	25mb
yellow	open	eyabenalaya_oneshot	p06qSwizQbqD9D6jBbGgXg	3	2	32000	9500	14.8mb	14.8mb
green	open	khadija_projet_kibana	EPNGCGlCS1Cm0xJlQmkmqQ	1	0	32000	7276	13.8mb	13.8mb
yellow	open	kane_oumar_streaming	vil7QWfoT1KRDWhHEnrTLw	3	1	32001	0	12.8mb	12.8mb
yellow	open	kane_oumar_one_shot	VD81re3QQOesFkmbAluANQ	3	2	32000	0	12.5mb	12.5mb
yellow	open	youssoupe_one_shot	LdaWjtnCTS6CjBzCMyyxSg	3	2	30070	3284	12.2mb	12.2mb
yellow	open	papabagaye_streaming	-ELxF93pSYy6-cDvCRoMiA	3	2	32022	0	12.2mb	12.2mb
yellow	open	ngomstreaming	pnHtCm4d75aXaMasePk1lg	3	2	32001	0	12mb	12mb
yellow	open	samsidine_projet_kiban	S5pkXw0lQBmIQgtAtMe-9g	3	2	32000	0	12mb	12mb
yellow	open	papabagaye_one-shot	Nq1BYJ-7TTK4fZ8f44-PEg	3	2	32000	0	12mb	12mb
yellow	open	ngomoneshot	MLX0RA0kTEys7FXuyKAagg	3	2	32000	0	11.9mb	11.9mb
yellow	open	ibrahima_camara_projet_kibana	OJ3BktwJTUa8NbR6weLu0w	2	1	32000	0	11.9mb	11.9mb
yellow	open	eyabenalaya_index	8d1gbxehQxGeBcmX8qhvG	1	1	32000	0	11.6mb	11.6mb
green	open	abdoukarim_projet_kibana	C80diXTmQFmkb8KVOKIpbA	1	0	32000	0	11.2mb	11.2mb
yellow	open	moustapha_ndiaye_one-shot	cBZF6nLNRcOA24E5_L7bdA	3	2	32000	0	11.1mb	11.1mb
yellow	open	maas_cheikhyakhoub_streaming	JrV5FFtAT2-Zb8PG-890xQ	1	1	32004	0	11.1mb	11.1mb
yellow	open	papasambadia_streaming	wo2XAXA2Qf094m0ToCY8zA	1	1	32001	0	11.1mb	11.1mb
yellow	open	khadim_mbacke_ndiaye_oneshoot	HR70Bv2aTG55Ag1W744bgw	1	1	32000	0	11.1mb	11.1mb
yellow	open	papasambadia_projet_kibana	mkJYyU8UTE-8A1NKI450vw	1	1	32000	0	11.1mb	11.1mb
yellow	open	khalifa	HjxV1hTwSu-ywIf1g15F0Q	3	1	30072	0	10.9mb	10.9mb

# Exemple

Comment lister les shards Elasticsearch ?

~ l'API `/_cat/shards?v`

← → ↻ ⚠ Non sécurisé | ec2-15-236-206-197.eu-west-3.compute.amazonaws.com:9200/\_cat/shards?v

index	shard	prirep	state	docs	store	ip	node
kibana_sample_data_flights	0	p	STARTED	13059	5.5mb	172.31.45.249	ip-172-31-45-249.eu-west-3.compute.internal
ilm-history-3-000001	0	p	STARTED			172.31.45.249	ip-172-31-45-249.eu-west-3.compute.internal
.kibana-event-log-7.10.0-000001	0	p	STARTED	6	27.5kb	172.31.45.249	ip-172-31-45-249.eu-west-3.compute.internal
.kibana_1	0	p	STARTED	95	10.4mb	172.31.45.249	ip-172-31-45-249.eu-west-3.compute.internal
.apm-custom-link	0	p	STARTED	0	208b	172.31.45.249	ip-172-31-45-249.eu-west-3.compute.internal
.async-search	0	p	STARTED	0	231b	172.31.45.249	ip-172-31-45-249.eu-west-3.compute.internal
.kibana_task_manager_1	0	p	STARTED	5	120.5kb	172.31.45.249	ip-172-31-45-249.eu-west-3.compute.internal
.apm-agent-configuration	0	p	STARTED	0	208b	172.31.45.249	ip-172-31-45-249.eu-west-3.compute.internal

# Comment vérifier l'état du cluster Elasticsearch ?

~ l'API `/_cluster/health?pretty`

← → ↻ ⚠ Non sécurisé | ec2-35-181-53-245.eu-west-3.compute.amazonaws.com:9200/\_cluster/health?pretty

```
{
  "cluster_name" : "formation_big_data",
  "status" : "green",
  "timed_out" : false,
  "number_of_nodes" : 2,
  "number_of_data_nodes" : 2,
  "active_primary_shards" : 10,
  "active_shards" : 20,
  "relocating_shards" : 0,
  "initializing_shards" : 0,
  "unassigned_shards" : 0,
  "delayed_unassigned_shards" : 0,
  "number_of_pending_tasks" : 0,
  "number_of_in_flight_fetch" : 0,
  "task_max_waiting_in_queue_millis" : 0,
  "active_shards_percent_as_number" : 100.0
}
```

← → ↻ ⚠ Non sécurisé | 20.101.123.129:8082/\_cluster/health?pretty

```
{
  "cluster_name" : "elasticsearch",
  "status" : "yellow",
  "timed_out" : false,
  "number_of_nodes" : 1,
  "number_of_data_nodes" : 1,
  "active_primary_shards" : 50,
  "active_shards" : 50,
  "relocating_shards" : 0,
  "initializing_shards" : 0,
  "unassigned_shards" : 38,
  "delayed_unassigned_shards" : 0,
  "number_of_pending_tasks" : 0,
  "number_of_in_flight_fetch" : 0,
  "task_max_waiting_in_queue_millis" : 0,
  "active_shards_percent_as_number" : 56.81818181818182
}
```

# Status du cluster Elasticsearch

- **Green** : tous les shards primaires et répliqués sont assignés à des data nodes
- **Yellow** : un ou plusieurs shards répliqués sont non assignés
- **Red** : un ou plusieurs shards primaires sont non assignés

Indexation



# Exercice

1 Coder un script python qui permet de stocker le document JSON suivant dans un index "prenom\_chiffre" (exemple : ibrahima\_1)

Document JSON:

Warning Vous devez cree dans un premier temps votre mapping sinon location sera un type text

[https://github.com/idiattara/Spark\\_DIATTARA/blob/main/map](https://github.com/idiattara/Spark_DIATTARA/blob/main/map)

[https://github.com/idiattara/Spark\\_DIATTARA/blob/main/crate\\_mapping.py](https://github.com/idiattara/Spark_DIATTARA/blob/main/crate_mapping.py)

[https://github.com/idiattara/Spark\\_DIATTARA/blob/main/post\\_elastic.py](https://github.com/idiattara/Spark_DIATTARA/blob/main/post_elastic.py)

```
{
  "location": "14.76, -14.76",
  "typeproduit": "electronique",
  "prix": 220,
  "agent_timestamp": datetime.utcnow().strftime('%Y-%m-%dT%H:%M:%SZ')
}
```

2 consulter la data

[http://clustersdaelatsic.eastus.cloudapp.azure.com:9200/index\\_name/\\_search?pretty](http://clustersdaelatsic.eastus.cloudapp.azure.com:9200/index_name/_search?pretty)

3 Récupérer le schéma de votre index avec la requête:

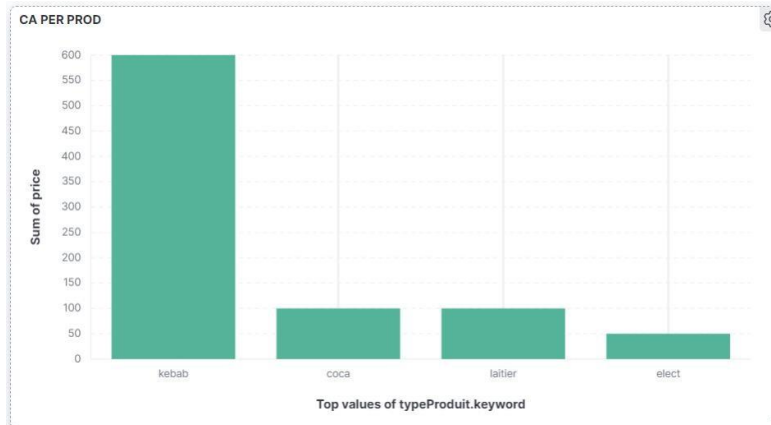
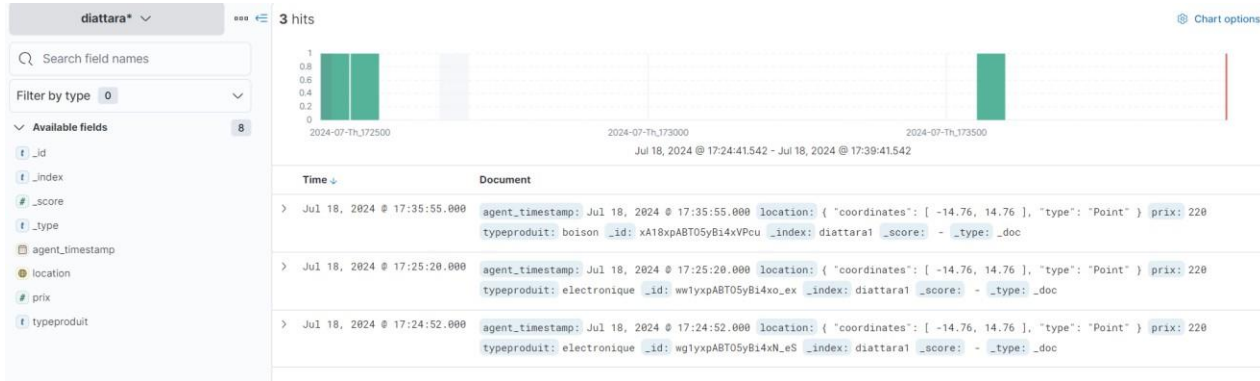
[http://clustersdaelatsic.eastus.cloudapp.azure.com:9200/ibrahima\\_1/\\_mapping?pretty](http://clustersdaelatsic.eastus.cloudapp.azure.com:9200/ibrahima_1/_mapping?pretty)

4 Sous kibana créer:

Un discover data

Un dashboard avec un map(chiffre d'affaire par location) et un viz bar ca par typedeproduit

# Resultat Exercice



# Solution Index mapping

- Le **schéma** de données correspond à **un mapping**. Mais concrètement, qu'est-ce que c'est ?
- **Lucene** a besoin, pour effectuer des **recherches**, de savoir comment lire les données.
- Si le schéma n'est pas défini, le mapping sera la structure du premier document inséré
- adresse : [http://mosef02.westeurope.cloudapp.azure.com:9200/my-index-01/\\_search?pretty](http://mosef02.westeurope.cloudapp.azure.com:9200/my-index-01/_search?pretty)

*PS: Le paramètre pretty permet de présenter le résultat de manière présentable.*

# Index mapping

il n'est pas possible de modifier le mapping d'un index une fois qu'il a été instancié (après la première importation). Il faut soit le supprimer, soit en créer un nouveau.

## Les strings:

le type "**string**" est divisé en deux nouveaux types:

- **Text** : qui doit être utilisé pour la recherche full-text
- **Keyword** : qui doit être utilisé pour la recherche par mot-clé

et pour les agrégations (count, ...).

```
"actors" : {  
  "type" : "text",  
  "fields" : {  
    "keyword" : {  
      "type" : "keyword",  
      "ignore_above" : 256  
    }  
  }  
},  
"directors" : {  
  "type" : "text",  
  "fields" : {  
    "keyword" : {  
      "type" : "keyword",  
      "ignore_above" : 256  
    }  
  }  
},  
}
```

# Index template

Un "**template**" est un moyen d'indiquer à Elasticsearch **comment configurer un index lors de sa création**.

Les "templates" sont configurés **avant la création de l'index**, puis lorsqu'un index est créé manuellement ou via l'indexation d'un document, les paramètres du "template" sont utilisés comme base pour la création de l'index.

# Indexation des données dans Elasticsearch

## Bulk API

Permet d'effectuer **plusieurs opérations d'indexation** ou **de suppression** en un seul appel d'API. Cela peut **augmenter** considérablement la **vitesse d'indexation**.

Exemple : (Dataset : [https://github.com/idiattara/data-ELK/blob/main/movies\\_elastic.json](https://github.com/idiattara/data-ELK/blob/main/movies_elastic.json))

```
curl -XPUT -H "Content-Type: application/json" http://mosef02.westeurope.cloudapp.azure.com:9200/_bulk --data-binary @movies_elastic.json
```

←	→	↻	⚠ Non sécurisé   20.101.123.129:8082/_cat/indices?v						
health	status	index	uuid	pri	rep	docs.count	docs.deleted	store.size	pri.store.size
yellow	open	toto	JQJMTT7kR8a9nGX63DUAdQ	1	1	1	0	4.5kb	4.5kb
yellow	open	fournisseur_20211210_064746	-JzmNZIYSnejApcOf2tGzA	1	1	4	0	5.4kb	5.4kb
yellow	open	fournisseur_20211210_035543	CV6PBRTuRjQNL5YwdFCNKA	1	1	4	0	5.4kb	5.4kb
yellow	open	fournisseur_20211210_041641	FIVI9DfmRrGSGbJF20m-8w	1	1	4	0	5.4kb	5.4kb
green	open	.kibana_task_manager_7.16.0_001	BNnW1UwSTsS50SVxRBeyWg	1	0	17	64411	6.4mb	6.4mb
yellow	open	fournisseur_20211209	a4qnXTzkSjWf23IB9TyquQ	1	1	4	0	5.4kb	5.4kb
green	open	.kibana_7.16.0_001	sNhppQe7TxqOMVo3nziBA	1	0	25	2	2.3mb	2.3mb
yellow	open	fournisseur_20211210_030332	hgFBfQhTSwa1iHfSb3GgcA	1	1	4	0	5.4kb	5.4kb
yellow	open	fournisseur_20211210_070247	mXr77pkxSAMhad4mh7JppA	1	1	4	0	5.4kb	5.4kb
yellow	open	fournisseur_20211210_050739	bJ5TOy4gQ4mtqFh7c1MJjg	1	1	4	0	5.4kb	5.4kb
yellow	open	fournisseur_20211210_030335	SGWG7F2WSuiZ6oL8o9SCDw	1	1	4	0	5.4kb	5.4kb
yellow	open	fournisseur_20211210_032217	19YKf6goTZiBHC294t-Bfg	1	1	4	0	5.4kb	5.4kb
yellow	open	fournisseur_20211210_073719	HZpYd4FpRSa6G4EEsNWNwg	1	1	4	0	5.4kb	5.4kb
yellow	open	fournisseur_20211210_055608	u8B_NaGcQuuuuFzc1JDkPg	1	1	4	0	5.4kb	5.4kb
yellow	open	fournisseur_20211210_063150	AUjK3f2wTImgbD93P_lmsg	1	1	4	0	5.4kb	5.4kb
yellow	open	my-index-01	vbwvgx5MSCWjvoOzuRB-zQ	1	1	1	0	4.8kb	4.8kb
yellow	open	movies	Dng5wyuxSmaqyUeWvXM8qg	1	1	4849	0	4mb	4mb

# Reindex API

Permet de copier des documents d'un index/alias source vers un index/alias destination.

La source et la destination devront être différents

## Exemple :

```
curl -X POST "http://20.101.123.129:8082/\_reindex?pretty" -H 'Content-Type: application/json' -d'
```

```
{  
  "source": {  
    "index": "my-index-000001"  
  },  
  "dest": {  
    "index": "my-new-index-000001"  
  }  
}
```

Requête



# Search API : Exemples

curl -X GET [http://20.101.123.129:8082/movies/\\_search?q=Star+Wars](http://20.101.123.129:8082/movies/_search?q=Star+Wars)

~ retourne les films qui contiennent le mot “Star Wars”

curl -X GET [http://20.101.123.129:8082/movies/\\_search?q=fields.actors:Harrison+Ford](http://20.101.123.129:8082/movies/_search?q=fields.actors:Harrison+Ford)

~ retourne les films dont le champs “fields.actors” contient le mot “Harrison Ford”

curl -X GET [http://20.101.123.129:8082/movies/\\_search?q=fields.actors:Harrison+Ford&size=20](http://20.101.123.129:8082/movies/_search?q=fields.actors:Harrison+Ford&size=20)

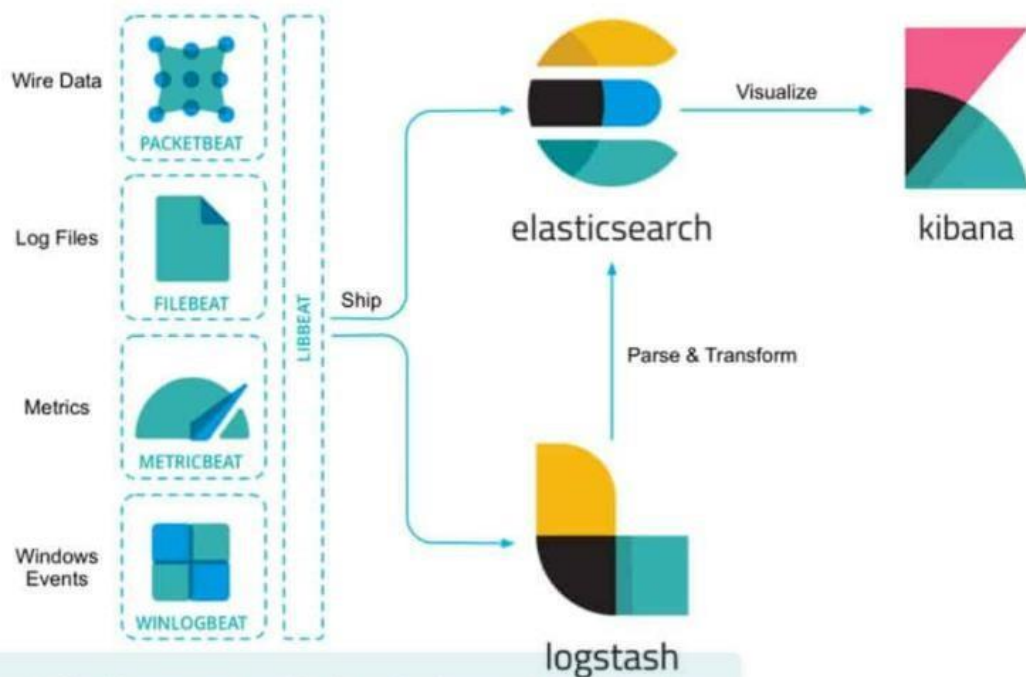
~ retourne 20 films dont le champs “fields.actors” contient le mot “Harrison Ford”

curl -X GET [http://20.101.123.129:8082/movies/\\_search?q=fields.title:Star+Wars%20AND%20fields.directors:George+Lucas](http://20.101.123.129:8082/movies/_search?q=fields.title:Star+Wars%20AND%20fields.directors:George+Lucas)

~ retourne les films dont le champs “fields.title” contient le mot “Star Wars” et le champs “fields.directors” contient le mot “George Lucas”

# ELK

## Elastic (ELK) Stack Architecture



ici pour sélectionner une partie de cette image et la rechercher