# Ejercicio 2 – Configuración de índices

|  |  |
| --- | --- |
| Apellidos, Nombre: | Campos de Padua, Alfonso |
| Puntuación Final: |  |

## Pregunta 1 [Puntuación 2]

Crear un índice y su mapping para documentos que lleven el formato mostrado en el siguiente ejemplo

{

"source": "MySource-10.134.4.94-DBName",

"lastExecDate": "2015-04-22 10:45:45",

"currentDate": "2015-04-22 08:52:25",

"instant": "2015-04-22T08:52:42Z",

"host": "EMEOE3402.dir.svc.accenture.com",

"serviceName": "System Center Core Monitoring",

"counterName": "agent processor utilization",

"category": "agent processor utilization",

"component": "",

"value": "0"

}

Asegurarse que el campo host no es analizado.

PUT dblog

{

"settings" : {},

"mappings" : {

"log" : {

"properties" : {

"source" : {"type" : "string" },

"lastExecDate" : {"type" : "date", "format": "yyyy-MM-dd HH:mm:ss"},

"currentDate" : {"type" : "date", "format": "yyyy-MM-dd HH:mm:ss" },

"instant" : {"type" : "date", "format": "yyy-MM-dd'T'HH:mm:ss'Z'" },

"host" : {"type" : "string", "index": "not\_analyzed" },

"serviceName" : {"type" : "string" },

"counterName" : {"type" : "string" },

"category" : {"type" : "string" },

"component" : {"type" : "string" },

"value" : {"type" : "integer" }

}

}

}

}

## Pregunta 2 [Puntuación 1]

##### Configurar el índice para que elimine todo documento al cabo de 1 minuto de haberlo dado de alta.

POST dblog/\_close

PUT dblog/\_settings

{

"index": {

"\_ttl" : {

"enabled" : true,

"default" : "1m"

}

}

}

POST dblog/\_open

## Pregunta 3 [Puntuación 1]

##### Obtener los settings actuales del índice que acabamos de crear

GET dblog/\_settings

### Pregunta 4 [Puntuación 1]

Mostrar todos los índices que tiene el ElasticSearch actual

Parece que el comando más útil para ver esta información es: GET \_cat/indices?v

## Pregunta 4 [Puntuación 5]

### Implementar un analizador que permita eliminar cualquier rastro de etiquetas HTML de un documento y nos retorne el contenido del mismo

#### Eliminar todos los tags HTML del documento suministrado.

#### Particionar el texto haciendo uso de espacios en blanco

#### Eliminar los caracteres de puntuación

#### Mover todos los tokens a lowercase.

#### Eliminar stopwords tradicionales

#### Para reducir la complejidad del ejercicio, únicamente tener en cuenta los siguientes tags de HTML

#### <em></em>

#### <br>

#### <p></p>

#### <HTML></HTML>

#### El analyzer debe ser capaz de retornar el texto, palabra a palabra, sin los tags HTMLs proporcionados en un documento como el siguiente

Seguro que hay más formas de implementarlo, pero esta parece bastante compacta y funciona al probarlo con GET /html/\_analyze?analyzer=html2info

Creo un índice de ejemplo, aunque el analizador anida dentro como se puede ver:

PUT /html

{

"settings": {

"analysis": {

"analyzer": {

"html2info": {

"type": "custom",

"char\_filter": "html\_strip",

"tokenizer": "whitespace",

"filter": [ "lowercase", "stop" ]

}}

}}}