**ELASTICSEARCH - Analyzers**

**What is an Analyzer?**

**The basic use of analyzers is to split and strip (clean) a text string however we need**

Analyzers are used to create an index of terms. For, examples to index a line, it may be useful to break it into words. So, an analyzer is a mix of tokenizers and filters (eg: whitespace tokenizer - splits line by spaces, lowercase tokenizer - splits line at each lowercase letter; filter is used to filter and convert items; eg: ê, é, è to e.).

The final goal of an analyzer is to convert a line (string) into a series of tokens

Analyzer first sends the input through character filters then through tokenizers and then through token filters.

**How to configure analyzers?**

Analyzers are to be configured through index settings

EG:

PUT /my-index/\_settings  
{  
 "index": {  
 "analysis": {  
 "analyzer": {  
 "customHTMLSnowball": {  
 "type": "custom",   
 "char\_filter": [  
 "html\_strip"  
 ],   
 "tokenizer": "standard",  
 "filter": [  
 "lowercase",   
 "Stop",

“snowball”  
 ]   
 }}}}}

In the above example the analyzer removes all html tags from the input. “Html\_strip” does this.

Converts all tokens to lowercase and removes punctuation, removes words such as ‘the’, ‘and’

**Analyzer types:**

**Standard Analyzer:** It has settings ‘ stopwords ‘ - to set the list of stopwords and ‘ max\_token\_length ‘ - to set maximum length of token (255 is the default).

**Simple Analyzer:**

Uses lowercase tokenizer

**Whitespace Analyzer:**

Uses whitespace tokenizer

**Pattern Analyzer:**

Separates text into terms by a regular expression

Take a look [here](https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-pattern-analyzer.html) for settings and examples

**Language Analyzers** can be used to aim for analyzing specific language text

Take a look [here](https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-lang-analyzer.html) for list of supported languages and different settings

**Creating a custom analyzer:**

To create a custom analyzer we need to set the tokenizers and filters .. that are to be used by the custom analyzer.

Settings:

Tokenizer - name of tokenizer to be used

Filter - list of token filters to be used

Char\_filter - list of character filters to be used

Position\_increment\_gap - no. of positions to increment between value of field

**Mappings: Elasticsearch**

Mapping is defining before indexing what are we going to index. We can tell ES which fields are going to be text fields and which are going to be dates, geo-locations etc.

A mapping type contains Fields(the general fields) and Meta-fields(\_index,\_id,\_source etc).

Fields have different data types like string, date, long, double, bool, ip, geo\_point, geo\_shape etc.

**Dynamic Mapping:**

Dynamic mapping is creating the mapping for new terms as you go.

Remember that you cannot modify an existing mapping, you can only add new ones

For example in index students, in doctype date\_of\_birth, if there is a field called name, you can’t change the mapping of that field, but you can add new field say birth\_day and add a new mapping for it.

So, It is not necessary to define mapping before creation of index, but it is necessary to define mapping before first indexing a term.

Also, remember that fields with same name in same index should have same mapping

**Getting mapping:**

GET index\_name/doc\_type/\_mapping - to get mappings in doctype

GET index\_name/\_mapping - to see all the mappings in index

PUT index/\_mapping/doc\_type

{

"properties": {

"birth\_day": { - to add a mapping for term birth\_day

"type": "date"

}

}

}

**Mapping types:**

string

long, integer, short, byte, double, float

date

boolean

binary

geo\_point, geo\_shape

ip

completion - to provide auto-complete suggestions ( [Example](https://www.elastic.co/guide/en/elasticsearch/reference/current/search-suggesters-completion.html) )

etc...