ElasticSearch для .NET

Татьяна Ёлкина
Старший программист
Интерфакс

elasticsearch

Обо мне:

Старший инженер-программист
Интерфакс
@Y0Tata ytata.512@gmail.com
facebook.com/y0tata

О чем поговорим:

- Какие задачи решали?
- Какие есть варианты решения подобных задач?
- Что такое ElasticSearch? Когда и зачем создавался?
- Как это работает?
- Про запросы и анализаторы
- Стек ELK. Что это такое?
- NEST. Примеры использования

Задача:

- Нужен быстрый и качественный поиск по постоянно растущему числу документов
- Нужна возможность изменения конфигурации “на лету”
- Нужно удобство использования ....

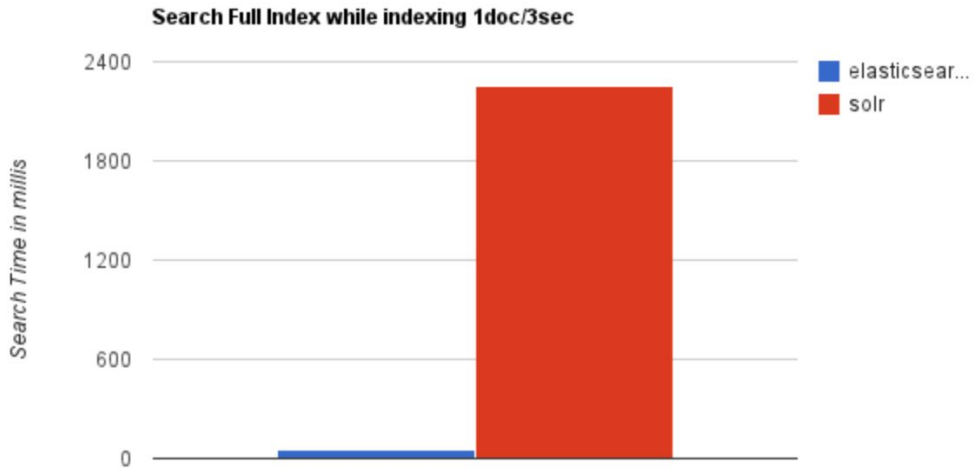
Какие продукты есть сейчас

- ElasticSearch (Lucene, Java)
- Solr (Lucene, Java)
- Sphinx (C++)
- Xapian (C++)

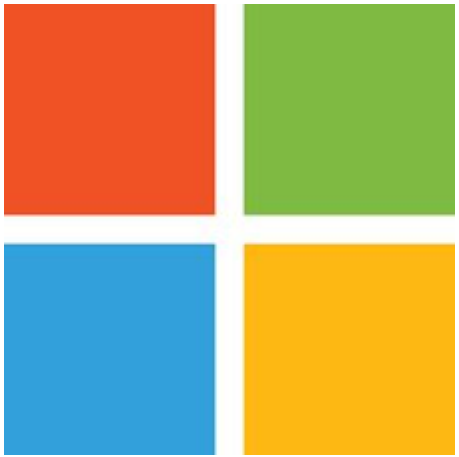
Почему выбрали ElasticSearch

- Поддержка шардинга и репликации на лету
- Написан на Java и базируется на Lucene
- Real-time индексация
- *Статья Ryan Sonnek*

ES.Преимущества производительности



Кто использует ещё...



ElasticSearch. История

Shay Banon Создал в 2010.

2012 год создана компания ElasticSearch BV для платной поддержки пользователей.

В 2014 году привлечены инвестиции.

В 2015 изменение названия на Elastic

Актуальная версия 6.1.2

ElasticSearch - основные преимущества

Open Source

Полнотекстовый поиск

Распределенность

Отказоустойчивость

Документ-ориентируемость

Schema-free

Ответ в режиме реального времени

ElasticSearch - основные преимущества

Open Source

Полнотекстовый поиск

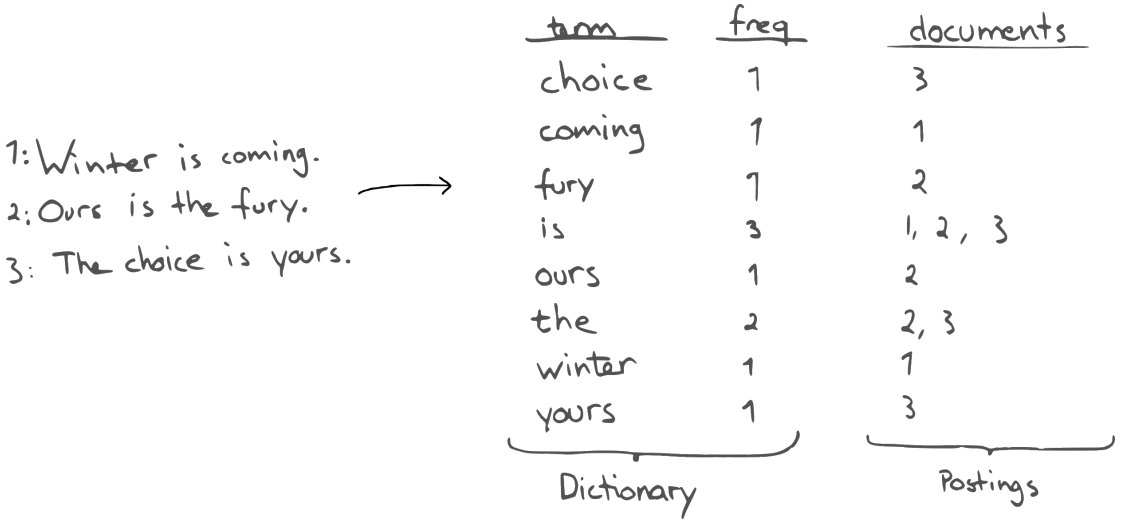
Распределенность

Отказоустойчивость

Документ-ориентируемость

Schema-free

Ответ в режиме реального времени



ElasticSearch - основные преимущества

Open Source

Полнотекстовый поиск

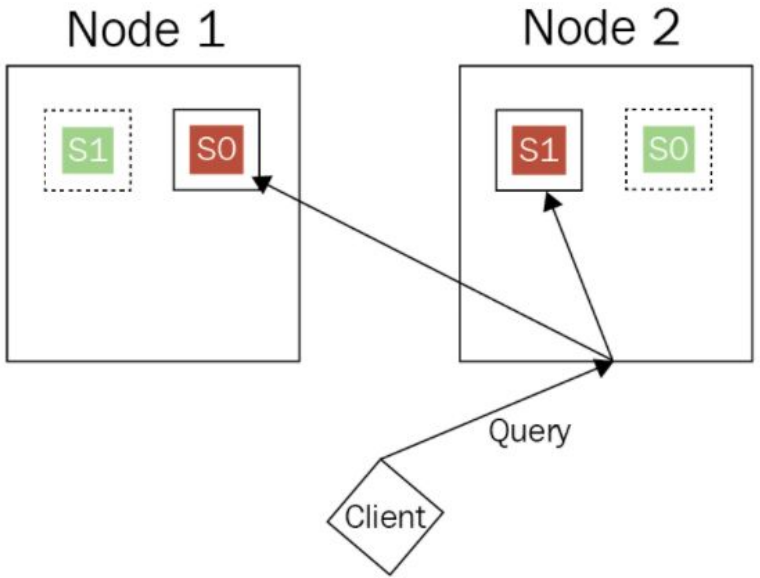
Распределенность

Отказоустойчивость

Документ-ориентируемость

Schema-free

Ответ в режиме реального времени



ElasticSearch - основные преимущества

Open Source

Полнотекстовый поиск

Распределенность

Отказоустойчивость

Документ-ориентируемость

Schema-free

Ответ в режиме реального времени

ElasticSearch - основные преимущества

Open Source

Полнотекстовый поиск

Распределенность

Отказоустойчивость

Документ-ориентируемость

Schema-free

Ответ в режиме реального времени

ElasticSearch - основные преимущества

Open Source

Полнотекстовый поиск

Распределенность

Отказоустойчивость

Документ-ориентируемость

Schema-free

Ответ в режиме реального времени

ElasticSearch - основные преимущества

Open Source

Полнотекстовый поиск

Распределенность

Отказоустойчивость

Документ-ориентируемость

Schema-free

Ответ в режиме реального времени

ElasticSearch - терминология

- **Нода -** процесс JVM, запущенный на каком-то сервере
- **Индекс** - набор документов
- **Шард** - часть индекса
- **Реплика** - копия шарда

ElasticSearch - терминология. Цвета кластера

- **Красный** - часть индекса недоступна вообще
- **Желтый** - какие-то реплики либо находятся в состоянии миграции, либо не прикреплены к нодам
- **Зеленый** - доступно требуемое количество реплик каждого шарда, каждого индекса

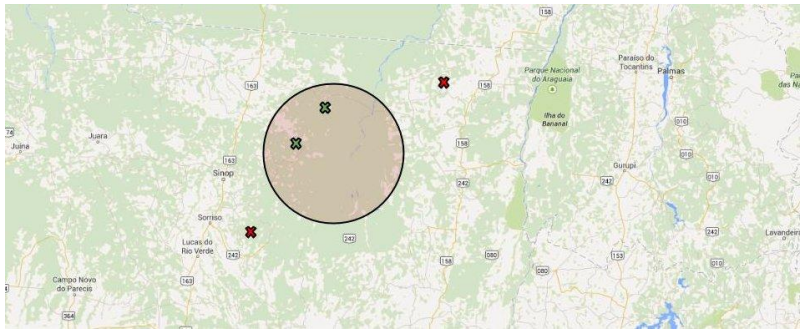
ElasticSearch - терминология

- **Стемминг** — это нахождение основы слова для заданного исходного слова. Основа необязательно совпадает с морфологическим корнем слова.
- **Лемматизация** — приведение слова к нормальной (словарной) форме. Для существительных это именительный падеж и единственное число.
- **Корпус** — в лингвистике корпусом называется совокупность текстов, собранных в соответствии с определенными принципами, размеченных по определенному стандарту и обеспеченных специализированной поисковой системой. Это может быть и разделение по стилям и жанрам, разделение по эпохе написания, по форме написания.
- **Параллельный корпус** — это один или более текстов на двух языках, сопоставленные между собой парами, когда в каждой паре оба предложения несут один и тот же смысл.
- **Стоп-слова, или шумовые слова,** — предлоги, суффиксы, междометия, цифры, частицы и подобное. Общие шумовые слова всегда исключаются из поискового запроса (кроме поиска по строгому соответствию поисковой фразы), также они игнорируются при построении инвертированного индекса.
- **N-грамма** — последовательность из n элементов. С семантической точки зрения это может быть последовательность звуков, слогов, слов или букв.

ElasticSearch. Запросы

- Обычные запросы (Term, Match/Not Match, Prefix, Bool, Wildcard, Range)
- Географические запросы (Geo-Shape, Geo-Bounding, Geo-Distance, Geo-Polygon)
- Boosting запросы
- Span запросы
- Нечеткие запросы
- Aggregations
- Nested Queries

ElasticSearch. Geo-Distance




```
{
    "filtered" : {
        "query" : {
            "field" : { "text" : "restaurant" }
        },
        "filter" : {
            "geo_distance" : {
                "distance" : "12km",
                "pin.location" : {
                    "lat" : 40,
                    "lon" : -70
                }
            }
        }
    }
}
```

ES. Примеры запросов

```
GET /_search

{    " query": {

          "query_string" : {
                "fields" : ["content", "name"],
                 "query" : "this AND that"
}}}
```

ES. Пример ответа

ElasticSearch. Анализаторы

Изнутри каждый анализатор представляет собой своеобразный конвейер, состоящий из нескольких обработчиков:

- **Символьной фильтрации**
- **Токенизации**
- **Фильтрации полученных токенов**

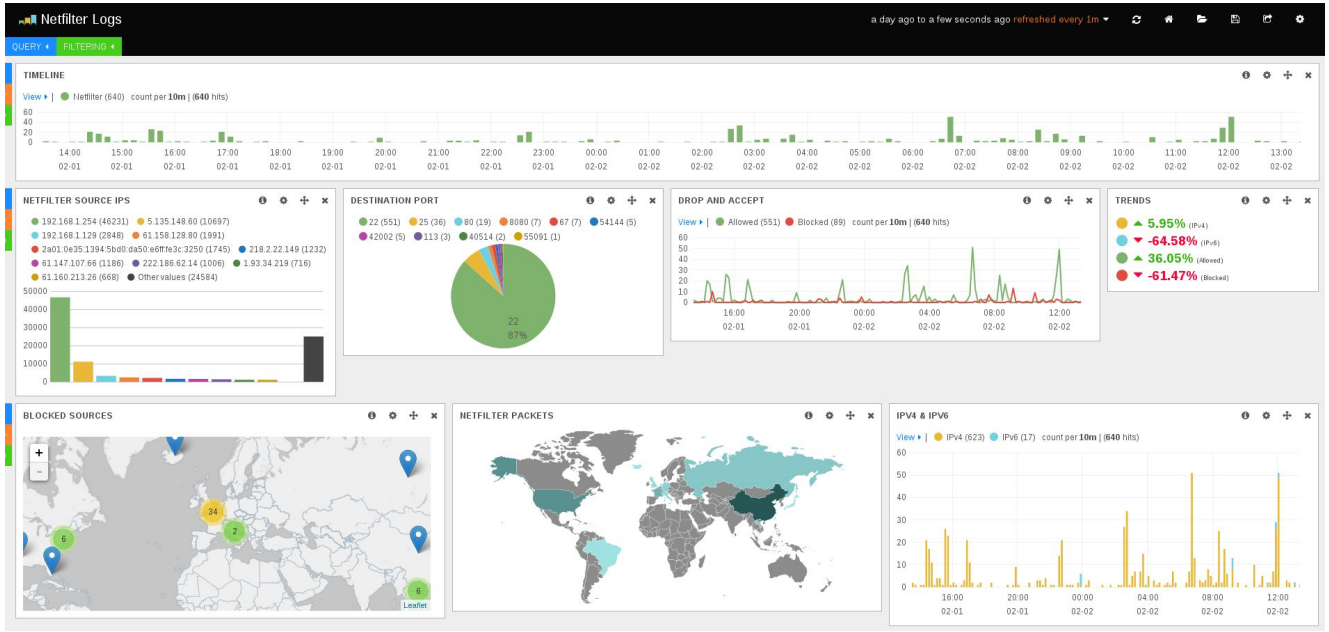
ElasticSearch. Анализаторы

```
"settings": {
        "analysis": {
            "analyzer": {
                "case_insensitive": {
                "tokenizer": "keyword",
                "filter": ["lowercase"]
            }
            }
        }
    }
```

Стек ELK

- ElasticSearch
- Logstash (Inputs(rabbitmq, syslog, file, redis), Filters, Outputs(file, elasticsearch..))
- Kibana
- ...

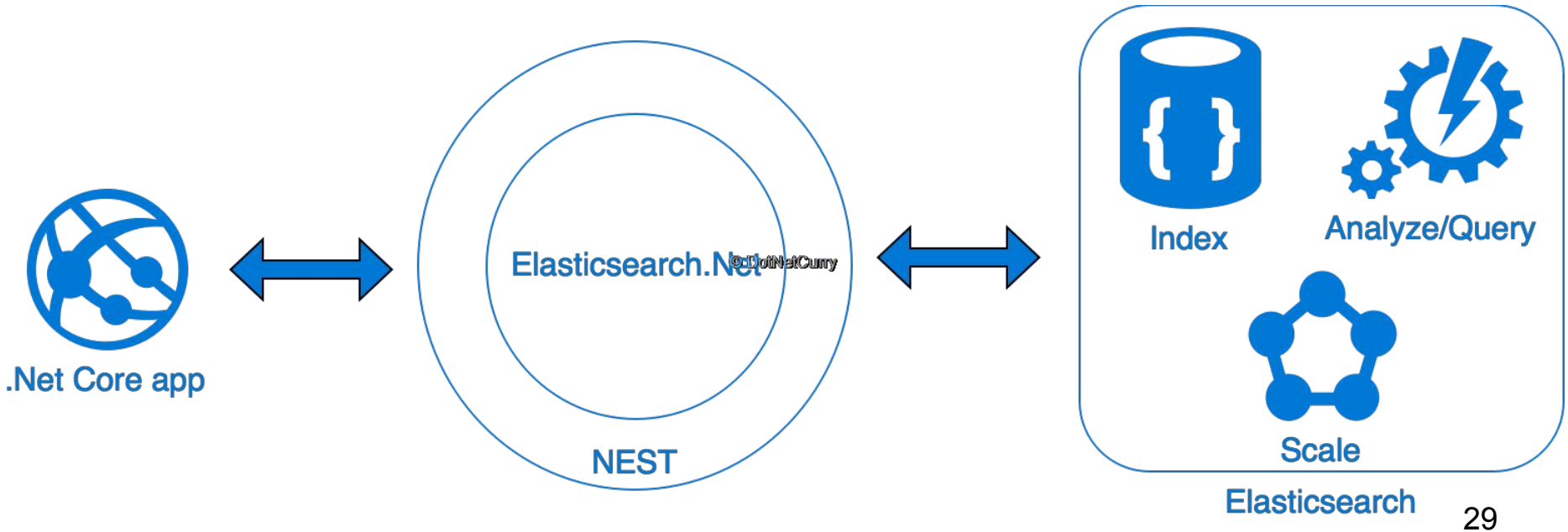
Kibana


Netfilter Logs
a day ago to a few seconds ago refreshed every 1m
QUERY
FILTERING
TIMELINE
NETFILTER SOURCE IPS
DESTINATION PORT
DROP AND ACCEPT
TRENDS
5.95% (IPv4)
-64.58% (IPv6)
36.05% (Allowed)
-61.47% (Blocked)
BLOCKED SOURCES
NETFILTER PACKETS
IPV4 & IPV6

NEST - ...

“NEST is a high level client that has the advantage of having mapped all the request and response objects, comes with a strongly typed query DSL that maps 1 to 1 with the Elasticsearch query DSL, and takes advantage of specific .NET features such as covariant results. NEST internally uses, and still exposes, the low level Elasticsearch.Net client.” www.elastic.co/

NEST. Схема


.Net Core app
Elasticsearch.Net
NEST
Index
Analyze/Query
Scale
Elasticsearch

Демо

Эпилог.