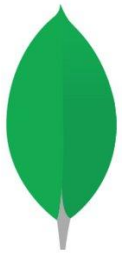
Search Engines

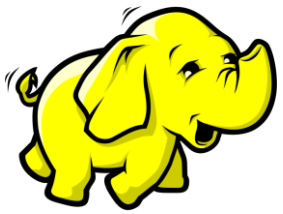
Michael Enudi

Journey through the world of databases and data engineering





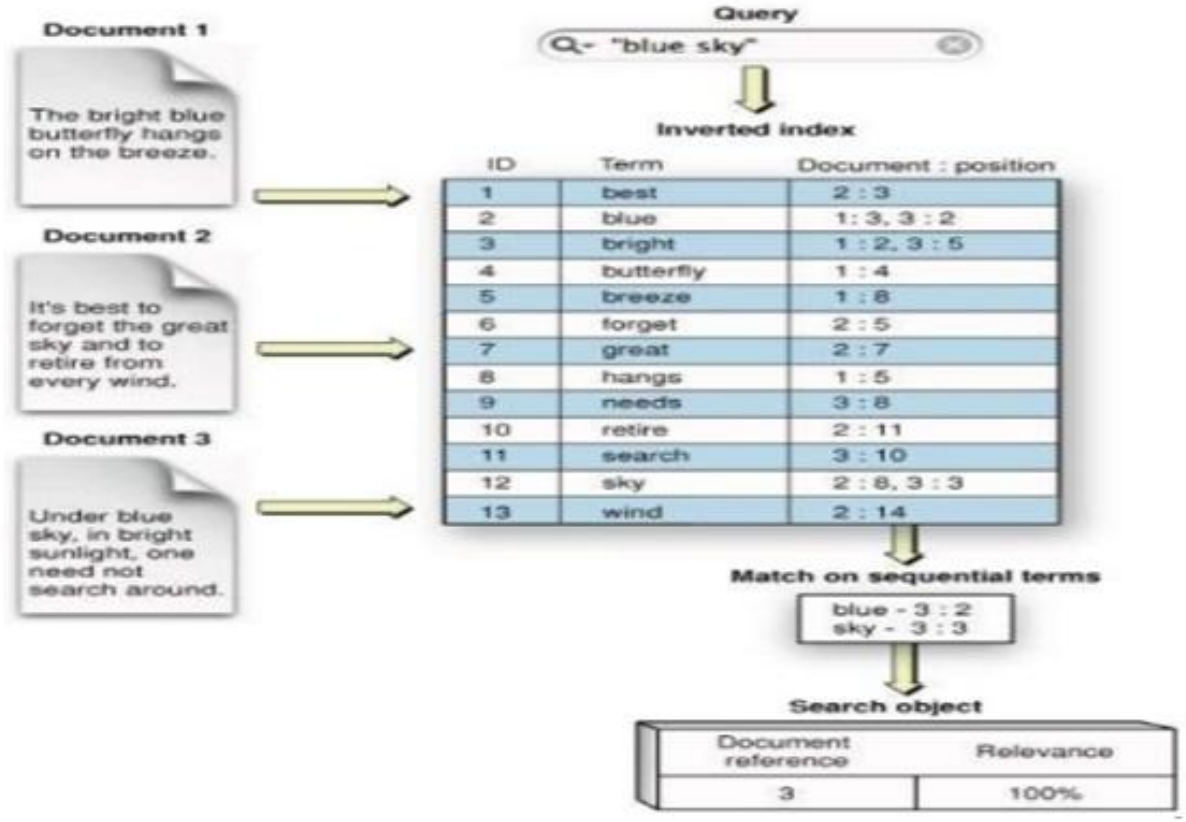



Scope

- Introduction to Search Engine Stores
- Use Case: Elasticsearch
- Modelling for Elasticsearch
- CRUD in Elasticsearch
- Data Analysis in Elasticsearch
- The Elastic Stack (ELK)
- Lab: UFO Sighting in Elasticsearch
- Search Engines: Wrap Up.

Introduction to Search Engines



elasticsearch

lucene

- open-source, horizontally scalable search engine that is built on top of Apache Lucene
- most popular search engine today
- has extensive aggregations query support
- developed by Elastic NV and offered as part of their ELK stack
- schema less database that also support schema definition
- AP from CAP theory perspective, with eventual consistency
- no single point of failure
- uses the Rest API over HTTP for client communication
- great support for Geospatial use case.
- high profile users include GitHub, Foursquare, Stackoverflow.

Elasticsearch Model

```
{
    "mappings" : {
        "properties" : {
            "datetime" : {"type": "date", "store": true, "format": "M/d/yyyy HH:mm"},
            "city" : {"type":"keyword", "store": true},
            "state" : {"type":"keyword", "store": true},
            "country" : {"type":"keyword", "store": true},
            "shape" : {"type":"keyword", "store": true},
            "duration_in_sec" : {"type" : "integer","store": true},
            "duration_in_hrs_mins" : {"type" : "text", "store" : false, "index":"false"},
            "comments" : {"type":"text", "store": true, "index":"true", "analyzer" : "snowball"},
            "date_posted" : {"type" : "date", "store" : true, "format":"M/d/yyyy", "ignore_malformed": "true"},
            "location" : {"type" : "geo_point", "ignore_malformed" : true}
        }
    },
    "settings" : {
        "number_of_shards" : 2,
        "number_of_replicas" : 1,
        "index" : {
            "analysis" : {
                "analyzer" : {
                    "keyword_analyzer" : {
                        "type" : "custom",
                        "tokenizer" : "keyword",
                        "filters" :  ["lowercase"]
                    }
                }
            }
        }
    }
}
```

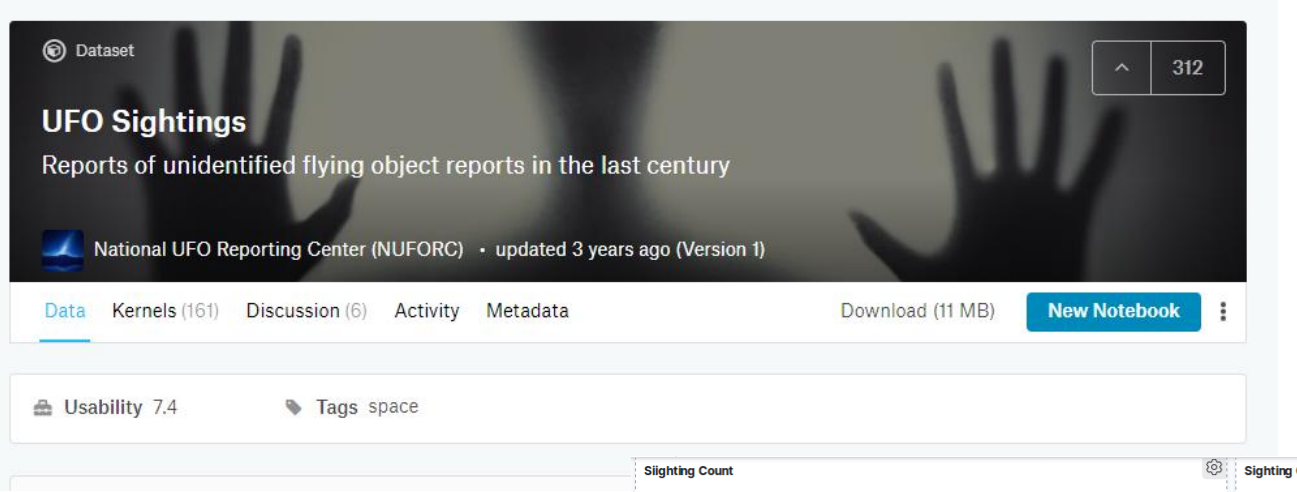
- ❑ Index
- ❑ Type
- ❑ Mapping
- ❑ Document
- ❑ Field
- ❑ Node
- ❑ Cluster
- ❑ Shard
- ❑ Replica
- ❑ Analyzer
  - ❑ Character Filter
  - ❑ Tokenizers
  - ❑ Token Filters

an open-source, horizontally scalable search engine database built on top of Apache Lucene

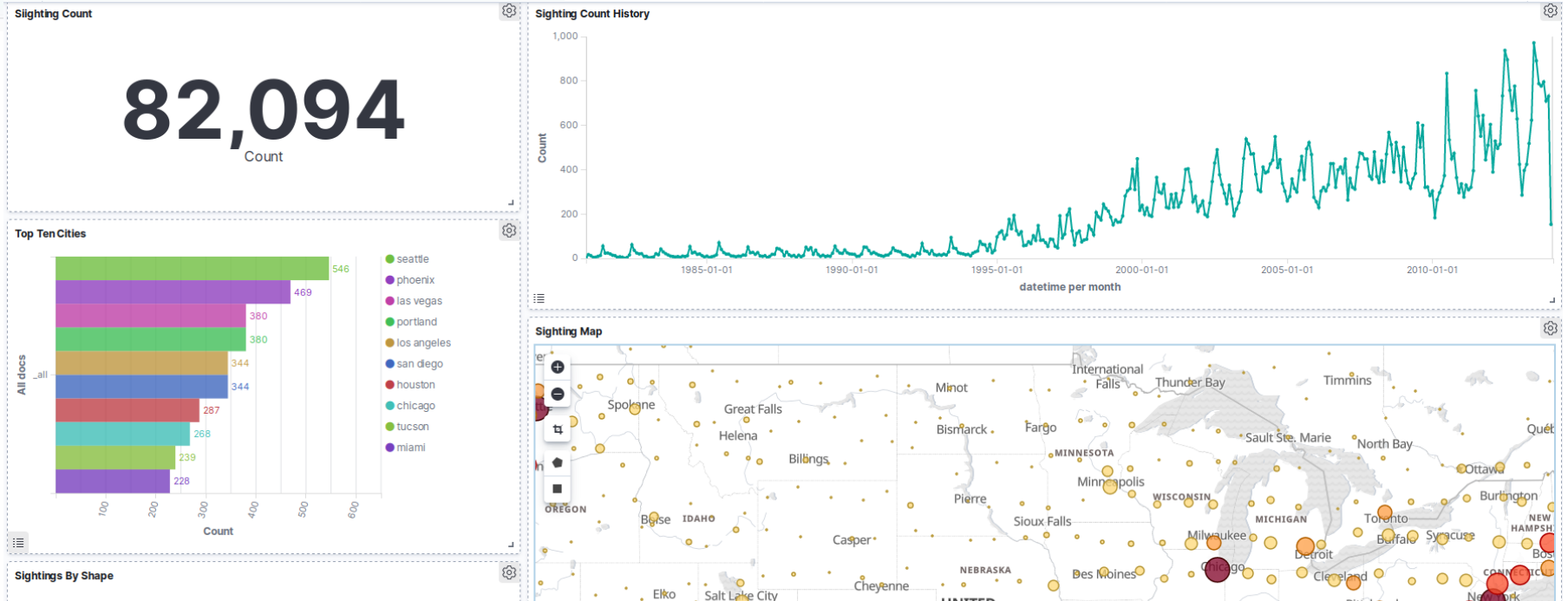
a distributed and extensible tool for data collection, enrichment and transportation pipeline

a visualization and analytics platform for Elasticsearch

UFO Sightings in Elasticsearch and Kibana



- What areas of the country are most likely to have UFO sightings?
- Are there any trends in UFO sightings over time?
- Do they tend to be clustered or seasonal?
- Do clusters of UFO sightings correlate with landmarks, such as airports or government research centers?
- What are the most common UFO descriptions?



https://www.kaggle.com/NUFORC/ufo-sightings

SEARCH ENGINE DATABASE WRAP-UP