

Manejo del entorno ELK

Máster en Data Science. URJC

Maria Cruz Gálvez Ortiz

Resumen de la práctica

Se ha tomado un fichero de datos sobre películas de la red, formato texto, y usando Logstash se han leído e introducido en Elasticsearch como documentos json creando un índice. Una vez comprobado que el índice se ha creado y Elasticsearch ha mapeado los documentos, se han realizado consultas sencillas referentes a varios campos. Por último se ha ejecutado Kibana, donde se ha creado el índice asociado al de Elasticsearch y además de algunas consultas simples en la ventana de consultas, se han usado un par de visualizaciones para ver la distribución de alguna de los campos de los documentos.

1. Primera parte: recogida de datos y generación de documentos e índices necesarios

Se ha seleccionado un fichero de datos en formato cvs, "The Movies Dataset", movies_metadata.csv. Tomado de <https://www.kaggle.com/rounakbanik/the-movies-dataset/version/7>, Los datos dan información sobre películas, clasificación, género, compañía productora, país de producción, título, lenguaje, una puntuación de popularidad, etc. Nombre de los 24 campos:

"adult", "belongs_to_collection", "budget", "genres", "homepage", "id", "imdb_id", "original_language", "original_title", "overview", "popularity", "poster_path", "production_companies", "production_countries", "release_date", "revenue", "runtime", "spoken_languages", "status", "tagline", "title", "video", "vote_average", "vote_count"

Se ha usado Logstash para crear documentos y un índice a través de un fichero de configuración con la estructura de entrada del fichero, filtros y salida a Elasticsearch. La salida también se pide por terminal para controlar cuando se cargaba el fichero.

Fichero de datos: movies_metadata.csv

Fichero de configuración de Logstash: practica.elk.mov.conf

```
input {
  file {
    path => "/Users/mcz/master/segundot/recuperacioninfo/practicas/
    practica2/movies_metadata.csv"
    type => "csv"
    sincedb_path => "./sincedb"
    start_position => beginning
  }
}
filter {
  csv {
    #add mapping columns name correspondily values assigned
    columns => ["adult","belongs_to_collection","budget","genres","homepage",
    "id","imdb_id","original_language","original_title","overview","popularity",
    "poster_path","production_companies","production_countries","release_date",
    "revenue","runtime","spoken_languages","status","tagline","title","video",
    "vote_average","vote_count"]
  }
}
```

```

        convert => {
            "budget" => "integer"
            "popularity" => "integer"
            "vote_average" => "integer"
            "vote_count" => "integer"
            "runtime" => "integer"
        }

        separator => ","
        remove_field => ["message", "path", "host"]
    }

#Remove first header line to insert in elasticsearch
    if [ID] =~ "ID"
    {
    drop {}
    }
    }
    output {
        elasticsearch {
            index => "moviesf"
            hosts => ["localhost:9200"]
        }
    }
#Console Out put
stdout
{
    codec => rubydebug
}
}

```

Ejecución: para ello se ejecuta primero ElasticSearch (bin/elasticsearch desde el directorio que contiene Elastic), y luego Logstash desde el directorio que lo contiene y especificando el camino al fichero de configuración:

```
bin/logstash -f /path/practica.elk_mov_red.conf
```

Se puede comprobar que el índice se ha creado listando los índices que contiene ElasticSearch:

```
curl -X GET "localhost:9200/_cat/indices?v"
```

Salida:

```

yellow open      moviesf                m0eoch40SFGSAgP5CYWdQw   5      1      45554
              0      71.1mb              71.1mb

```

2. Segunda parte: búsquedas.

Se han realizado 5 consultas sobre los datos, escritas en ficheros ejecutables desde terminal. Son:

- Consulta 1: consulta1_movies.sh, consulta que proporciona el mapeado los datos ElasticSearch. Fichero con la salida de la consulta: outputconsulta1_movies.json
Código:

```
curl -XGET 'localhost:9200/moviesf/_mapping?pretty'
```

- Consulta 2: consulta2_movies.sh, consulta que proporciona tres campos, "title", "original_language" y "popularity", de los 100 primeros documentos. Fichero con la salida de la consulta: outputconsulta2_movies.json

Código:

```
#!/bin/bash
# -*- ENCODING: UTF-8 -*-
curl -XGET 'localhost:9200/moviesf/_search?pretty' -H 'Content-Type:
application/json' -d'
{
  "_source": ["title","original_language","popularity"],
  "query": { "match_all": {} },
  "size": 10
},
```

- Consulta 3: consulta3_movies.sh, consulta que proporciona tres campos, "title", "original_language" y "popularity", de documentos donde el título de la película contenga la palabra "Future", hasta 10 documentos. Fichero con la salida de la consulta: outputconsulta3_movies.json

Código:

```
#!/bin/bash
# -*- ENCODING: UTF-8 -*-
curl -XGET 'localhost:9200/moviesf/_search?pretty' -H 'Content-Type:
application/json' -d'
{
  "_source": ["title","original_language","popularity"],
  "query": { "match_phrase": { "title": "Future" } },
  "size": 10
},
```

- Consulta 4: consulta4_movies.sh, consulta que proporciona tres campos, "title", "original_language" y "popularity", de documentos donde el título de la película contenga tanto la palabra "future" como la palabra "woman", hasta 40 documentos. Fichero con la salida de la consulta: outputconsulta4_movies.json

Código:

```
#!/bin/bash
# -*- ENCODING: UTF-8 -*-
curl -XGET 'localhost:9200/moviesf/_search?pretty' -H 'Content-Type:
application/json' -d'
{
  "_source": ["title","original_language","popularity"],
  "query": {
    "bool": {
      "must": [
        { "match": { "title": "future" } },
        { "match": { "title": "woman" } }
      ]
    }
  },
  "size": 40
},
```

```
}  
,
```

- Consulta 5: consulta5_movies.sh, consulta que proporciona los documentos agrupados por país de producción("production_countries"). Fichero con la salida de la consulta: outputconsulta5_movies.json
Código:

```
#!/bin/bash  
# -*- ENCODING: UTF-8 -*-  
curl -XGET 'localhost:9200/moviesf/_search?pretty' -H 'Content-Type:  
application/json' -d'  
{  
  "size":0,  
  "aggs": {  
    "group_by_production_countries": {  
      "terms":{  
        "field": "production_countries.keyword"}}  
    }  
  }  
}'
```

3. Tercera parte: visualización con Kibana.

Se ejecuta Kibana desde el directorio que lo contiene: ./bin/kibana. En la interfaz de Kibana del navegador seleccionamos crear un índice nuevo "movies" y lo asociamos al índice metido en Elasticsearch (moviesf). No se ha asociado un campo temporal en este caso.

- Se han realizado consultas sencillas en la interfaz de manera similar a las realizadas en la segunda parte. Ver figura 1.

Documentos donde el título contiene la palabra "future":
title: future

Documentos donde el valor de popularidad está entre 0.9 y 1.0:
popularity:[0.9 TO 1.0]

Documentos donde el género es de aventuras y el lenguaje original el inglés:
genres:(Adventure) and spoken_languages:(English)

- Visualización. Diagrama de tarta para determinar los 20 países que son mayores productores de películas en los datos, Estados Unidos sale como el mayor productor. Se matiza el diagrama con las películas donde entre los productores está España, donde sólo el 41% se produce únicamente por España. Ver figura 2.
- Visualización. Diagrama de tarta para determinar los 20 lenguajes predominantes en la lista de películas, siendo el inglés el dominante. Se matiza el diagrama con las películas donde alguno de los lenguajes sea el español, donde en el 46% es lenguaje único. Ver figura 3.
- Visualización. Diagrama de barras sencillo donde se dibujan los tiempos de duración de las películas, hasta 20, siendo lo más usual que duren 90 minutos y el segundo más usual 100 minutos. Cero: 0 minutos evidentemente serán datos que no tienen la duración. Ver figura 4.

4. Lista ficheros que componen la práctica:

1. Fichero de configuración de Logstash: practica.elk_mov.conf

kibana

Discover

Visualize

Dashboard

Timelion

Dev Tools

Management

44 hits

title:future

Uses lucene query syntax

title:future

movies*

Selected Fields

Available Fields

@timestamp

@version

_id

_index

_score

_type

adult

belongs_to...

budget

_source

title: Bright Future vote_count: 14 video: False homepage: http://www.uplink.co.jp/bright-future/ original_language: ja runtime: 92.0 status: Released budget: 0 imdb_id: tt0363235 @timestamp: April 28th 2018, 09:21:33.135 original_title: アカルイミライ type: csv adult: False production_companies: [] spoken_languages: [{iso_639_1: 'ja', name: '日本語'}] tagline: ~ release_date: January 18th 2003, 01:00:00.000 revenue: 0 @version: 1

title: Future War status: Released belongs_to_collection: ~ adult: False poster_path: /noNJgySGm6Fo5v7WNSgWgmK5Q5.jpg release_date: January 28th 1997, 01:00:00.000 original_language: en vote_count: 11 @version: 1 video: False spoken_languages: [{iso_639_1: 'en', name: 'English'}] type: csv imdb_id: tt0113135 revenue: 0 @timestamp: April 28th 2018, 17:50:20.585 original_title: Future War popularity: 0.616623 tagline: Post Pr

title: Future Cops status: Released belongs_to_collection: ~ adult: False poster_path: /46uHUHNVy0QaRL1vpqjg904Ukwr.jpg release_date: July 15th 1993, 02:00:00.000 original_language: cn vote_count: 8 @version: 1 video: False spoken_languages: [{iso_639_1: 'cn', name: '广州话 / 廣州話'}, {iso_639_1: 'en', name: 'English'}] type: csv imdb_id: tt0106545 revenue: 0 @timestamp: April 28th 2018, 17:49:48.938 original_title:

kibana

Discover

Visualize

Dashboard

Timelion

Dev Tools

Management

2,000 hits

popularity:[0.9 TO 1.0]

Uses lucene query syntax

movies*

Selected Fields

Available Fields

@timestamp

@version

_id

_index

_score

_type

Top 5 values in 500 / 500 records

doc

100.0%

Visualize

_source

popularity: 0.915041 vote_count: 8 video: False homepage: ~ original_language: en runtime: 0.0 status: Released budget: 0 imdb_id: tt0120014 title: Rhyme & Reason @timestamp: April 28th 2018, 09:21:10.055 original_title: Rhyme & Reason type: csv adult: False production_companies: [] spoken_languages: [{iso_639_1: 'en', name: 'English'}] tagline: ~ release_date: March 4th 1997, 01:00:00.000 revenue: 0 @version: 1

popularity: 0.961088 vote_count: 7 video: False homepage: ~ original_language: en runtime: 108.0 status: Released budget: 0 imdb_id: tt0119365 title: Incognito @timestamp: April 28th 2018, 09:21:14.771 original_title: Incognito type: csv adult: False production_companies: [{name: 'Warner Bros.', id: 6194}, {name: 'Morgan Creek Productions', id: 10210}] spoken_languages: [{iso_639_1: 'en', name: 'English'}] tagline: H

popularity: 0.914294 vote_count: 10 video: False homepage: ~ original_language: en runtime: 96.0 status: Released budget: 0 imdb_id: tt0066811 title: The Barefoot Executive @timestamp: April 28th 2018, 09:21:17.490 original_title: The Barefoot Executive type: csv adult: False production_companies: [] spoken_languages: [{iso_639_1: 'en', name: 'English'}] tagline: The Secret To Success Is Pure Monkey Business release_date: March 17th 1

kibana

Discover

Visualize

Dashboard

Timelion

Dev Tools

Management

59,753 hits

genres:(Adventure) and spoken_languages:(English)

Uses lucene query syntax

movies*

Selected Fields

Available Fields

@timestamp

@version

_id

_index

_score

_type

Top 5 values in 500 / 500 records

doc

100.0%

Visualize

_source

spoken_languages: [{iso_639_1: 'en', name: 'English'}, {iso_639_1: 'it', name: 'Italiano'}] genres: [{id: 12, name: 'Adventure'}, {id: 10751, name: 'Family'}] overview: After being shipwrecked, the Robinson family is marooned on an island inhabited only by an impressive array of wildlife. In true pioneer spirit, they quickly make themselves at home but soon face a danger even greater than nature: dastardly pirates. A rousing adventure

spoken_languages: [{iso_639_1: 'cs', name: 'Česky'}, {iso_639_1: 'en', name: 'English'}, {iso_639_1: 'fr', name: 'Français'}, {iso_639_1: 'pl', name: 'Polski'}] genres: [{id: 10751, name: 'Family'}, {id: 12, name: 'Adventure'}, {id: 16, name: 'Animation'}] overview: A free adaptation of Charles Perrault's famous Puss'n Boots, 'The True Story of Puss'n Boots' is a story for young and old for the first time on cinema screens

spoken_languages: [{iso_639_1: 'en', name: 'English'}, {iso_639_1: 'it', name: 'Italiano'}] overview: After being shipwrecked, the Robinson family is marooned on an island inhabited only by an impressive array of wildlife. In true pioneer spirit, they quickly make themselves at home but soon face a danger even greater than nature: dastardly pirates. A rousing adventure suitable for the whole family, this Disney adaptation of the classic Johann Wyss no

Figure 1: Consultas: documentos donde el título contiene la palabra "future", documentos donde el valor de popularidad está entre 0.9 y 1.0, y documentos donde el género es de aventuras y el lenguaje original el inglés.

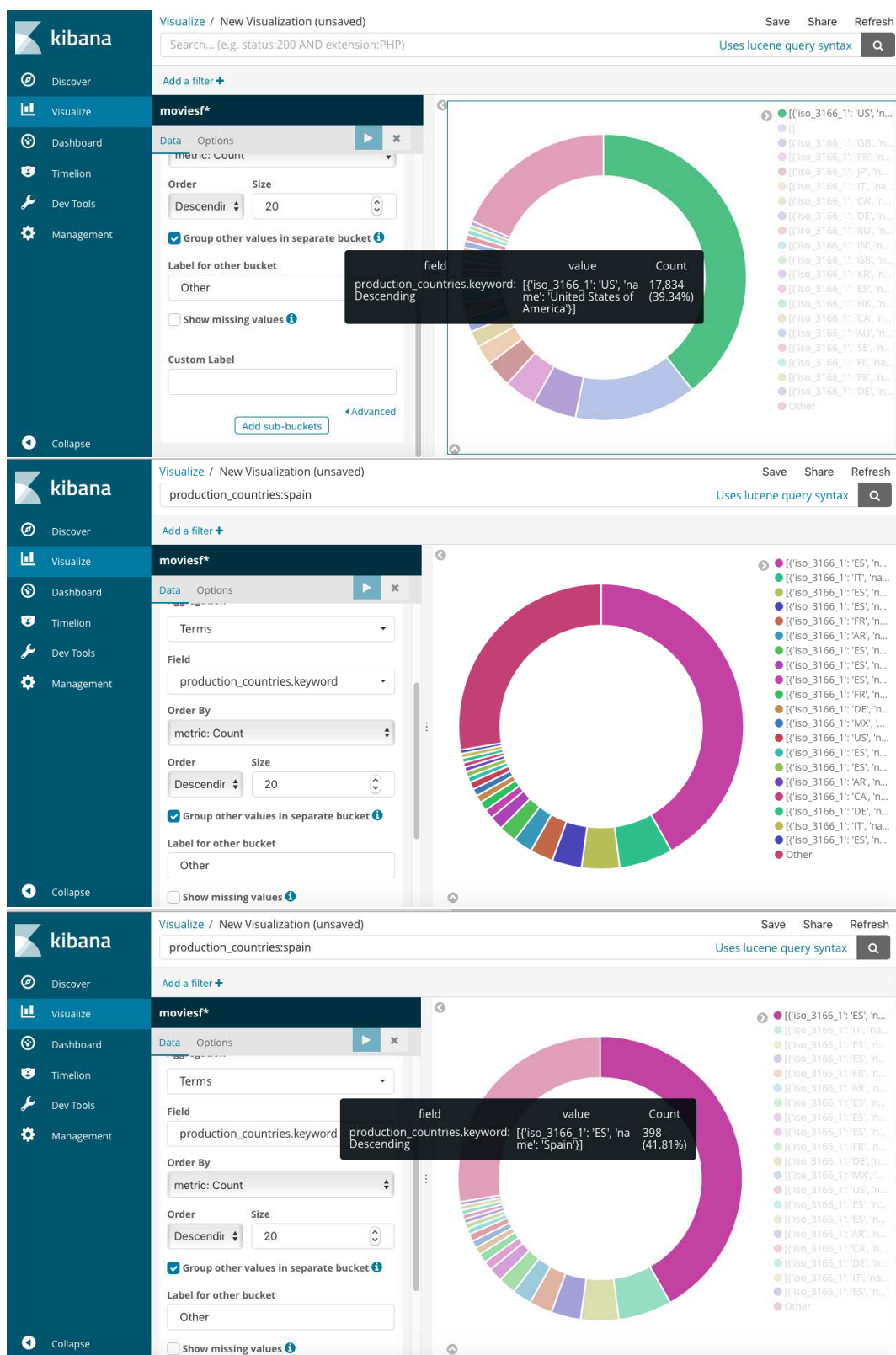


Figure 2: Diagrama de tarta sobre los países productores de películas y las que se producen incluyendo España.

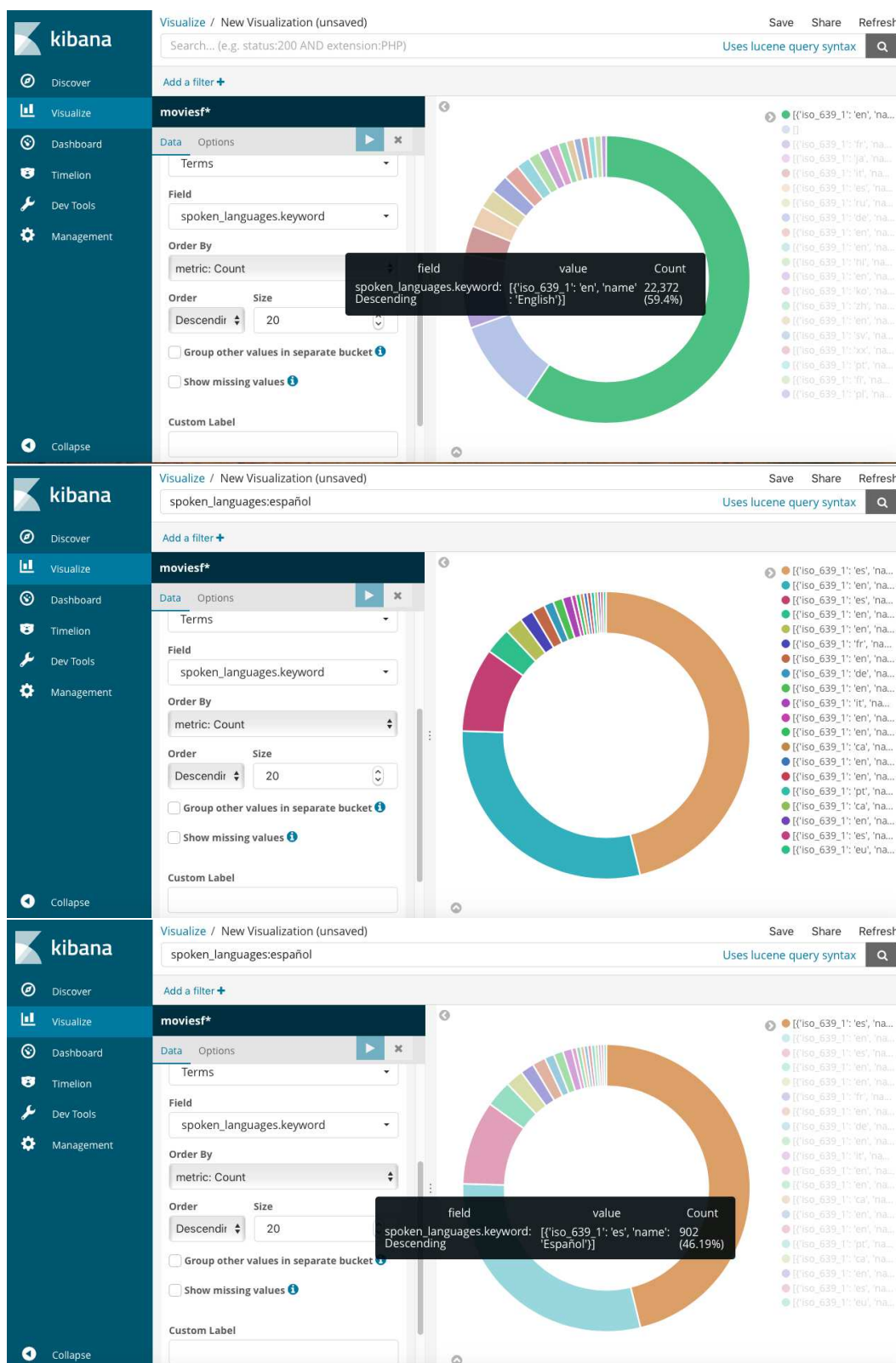


Figure 3: Diagrama de tarta sobre el lenguaje de películas y las que incluyen español.

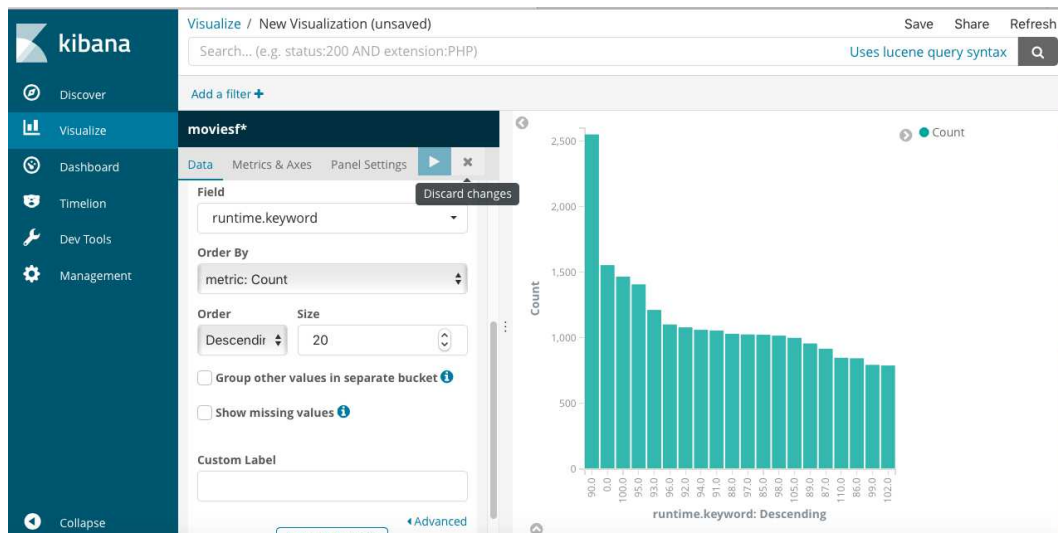


Figure 4: Diagrama de barras de la duración de las películas.

2. Ficheros con las consultas: consulta1_movies.sh, consulta2_movies.sh, consulta3_movies.sh, consulta4_movies.sh y consulta5_movies.sh.
3. Ficheros con la salida de las consultas: outputconsulta1_movies.sh, outputconsulta2_movies.sh, outputconsulta3_movies.sh, outputconsulta4_movies.sh y outputconsulta5_movies.sh.