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 conprobado 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 selecionado 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, compañia 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             mOeoch40SFGSAgP5CYWdQw   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 palabla "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 palabla "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 selecionamos 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_laguages:(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 esta España, donde sólo el 41% se produce unicamente por España. Ver figura 2.
   - Visualización. Diagrama de tarta para determinar los 20 lenguajes predominantes en la lista de películas, siendo el inglés el dominantes. Se matiza el diagrama con las películas donde alguno de los lenguajes sea el español, donde en el 46% es leguaje 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

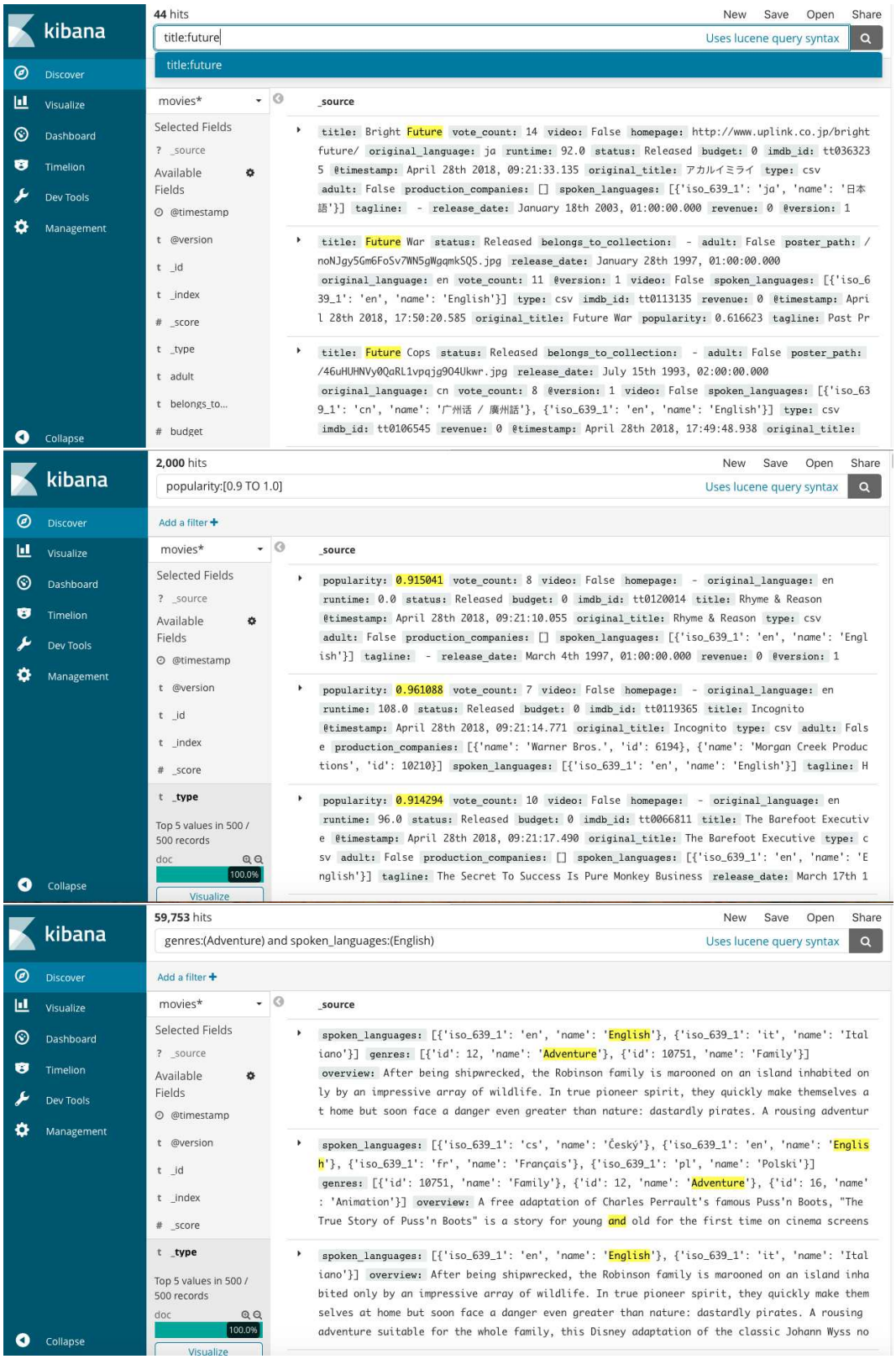

Figure 1: Consultas: documentos donde el título contiene la palabra "future", documentos donde el valor de popularidad está entre 0.9 y 1.0, y documentos donde el género es de aventuras y el lenguaje original el inglés.

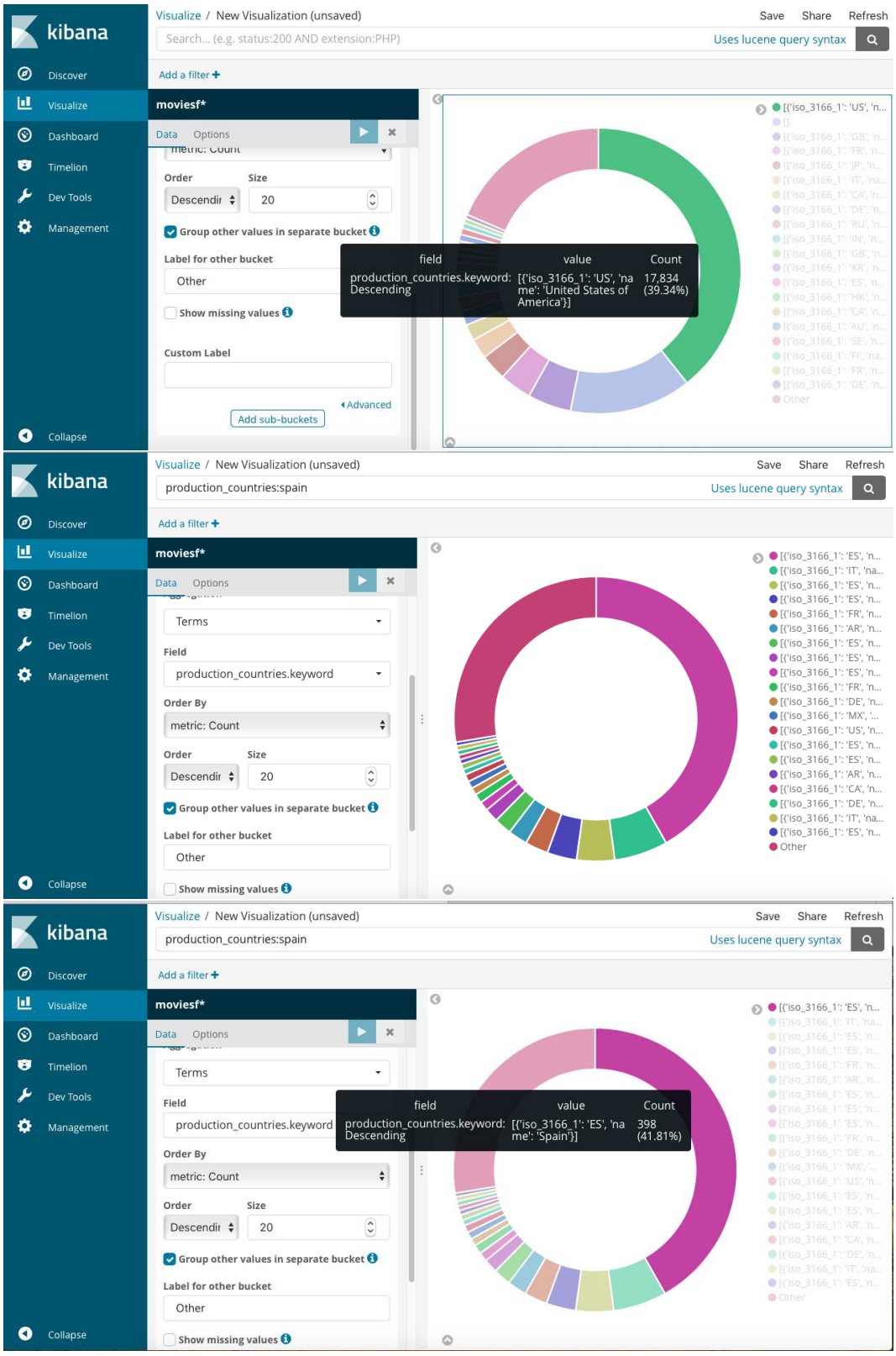

Figure 2: Diagrama de tarta sobre los países productores de películas y las que se producen incluyendo España.

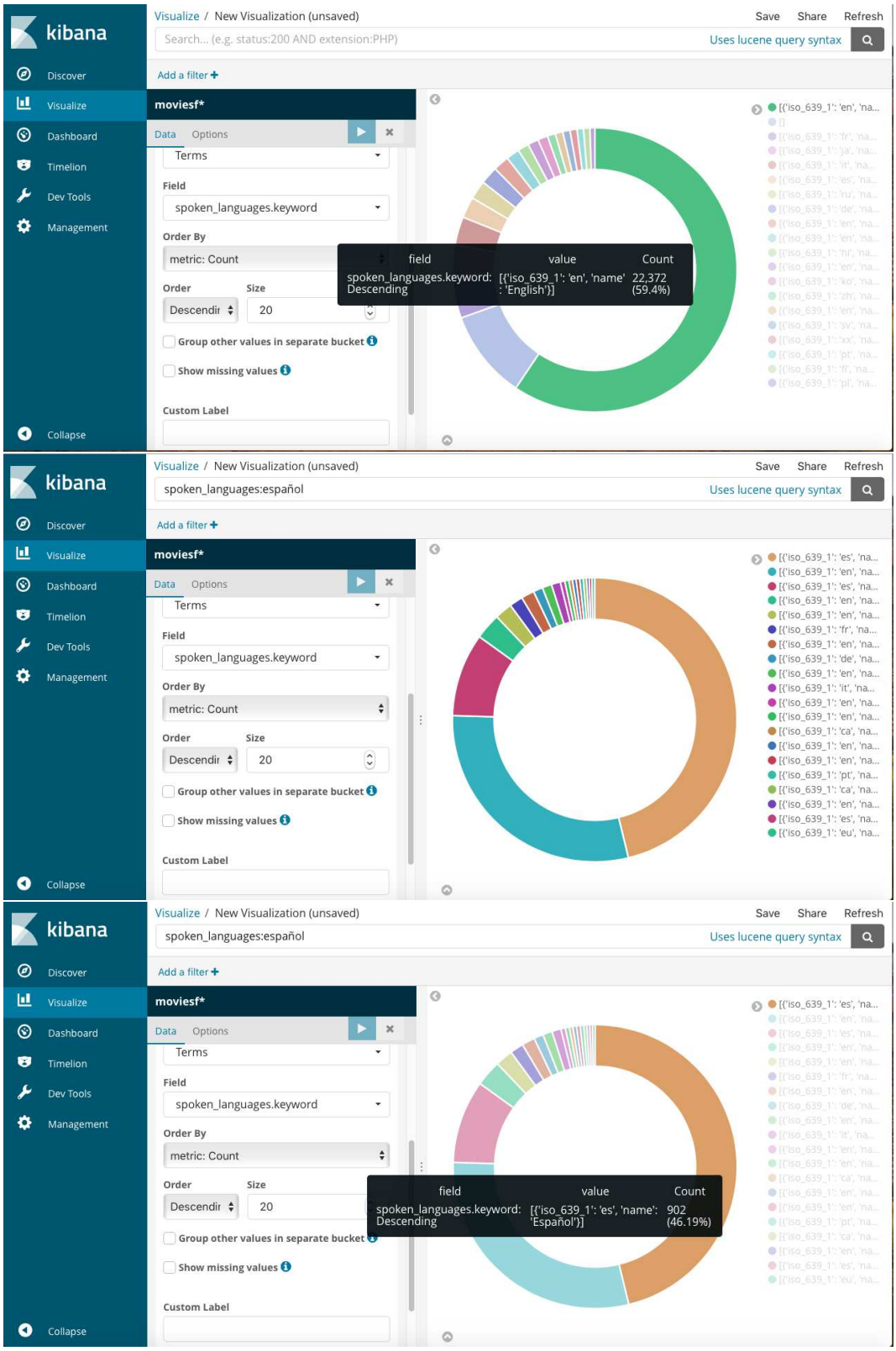

Figure 3: Diagrama de tarta sobre el lenguaje de películas y las que incluyen espaol.

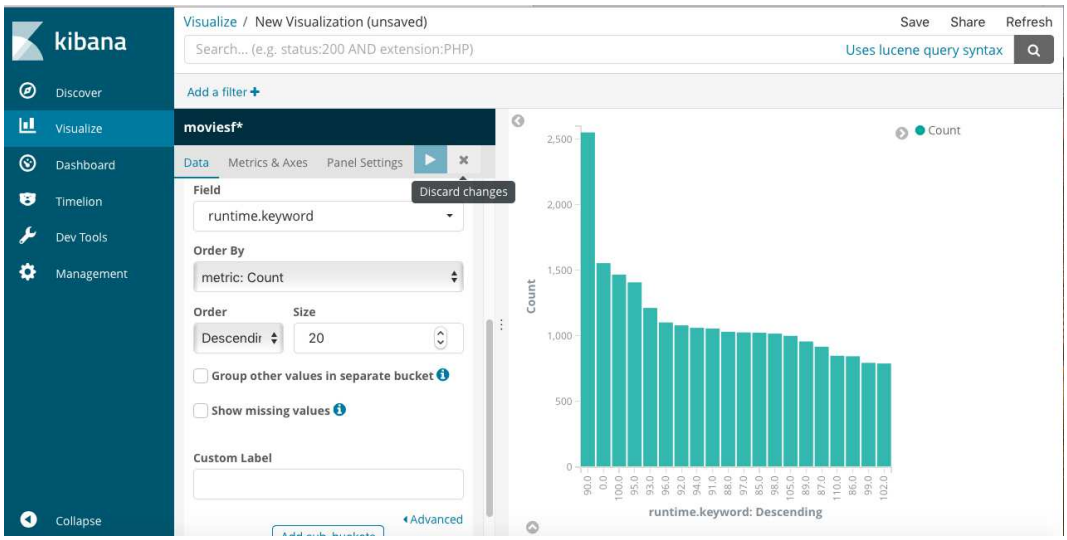

Figure 4: Diagrama de barras de la duración de las películas.

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.