

Buscadores

Sesion II

Agenda

- Elasticsearch. Introducción
- Bundle ELK
- Comparativa Elastic vs SolR

- Utiliza Lucene como motor de indexación
- Permite también búsquedas muy variadas de texto completo
- Soporta configuración en cluster con servicios de alta disponibilidad
- La gestión se realiza a través de servicios RESTful
- El documento a ingestar deberá estar, principalmente, en formato JSON
- Posicionado en muchos casos como sistema de búsqueda en entornos de tiempo real

<https://www.elastic.co/>

Ecosistema Elastic

THE ELASTIC STACK

Elasticsearch + Kibana

Meet the open source tools that power experiences from the search for life on Mars to finding the best sushi in your neighborhood. Learn more about [the Elastic Stack](#).



Elasticsearch

Elasticsearch is a distributed, JSON-based search and analytics engine.

[Learn more](#)



Kibana

Kibana is the window into the Elastic Stack. Explore your data and manage the stack.

[Learn more](#)



Beats

Beats is a platform for lightweight shippers that send data from edge machines.

[Learn more](#)



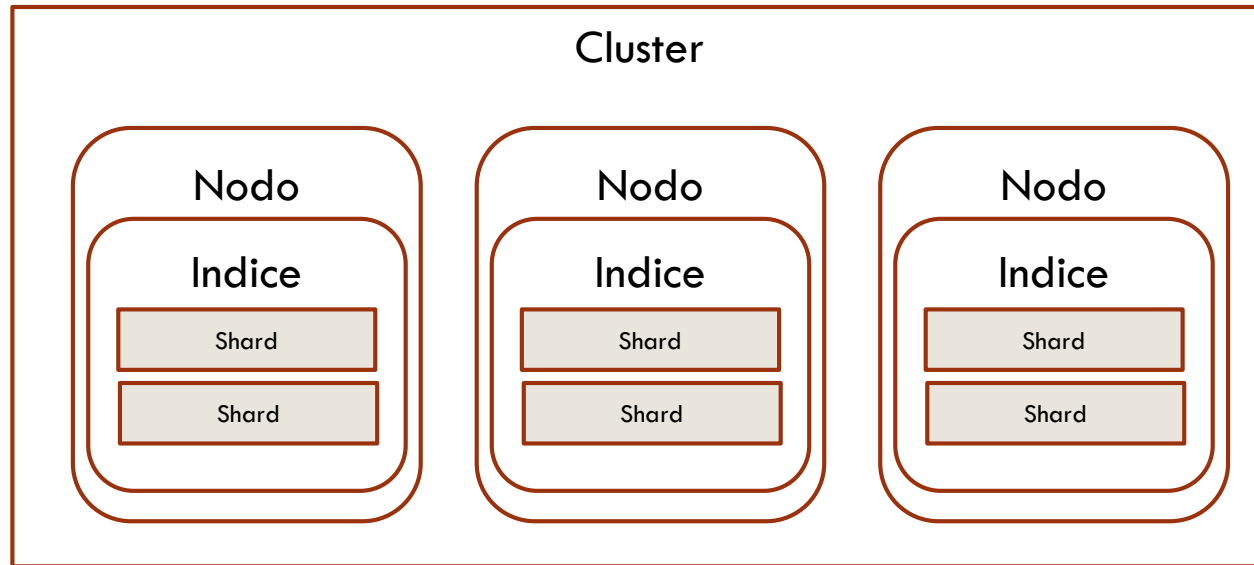
Logstash

Logstash is a dynamic data collection pipeline with an extensible plugin ecosystem.

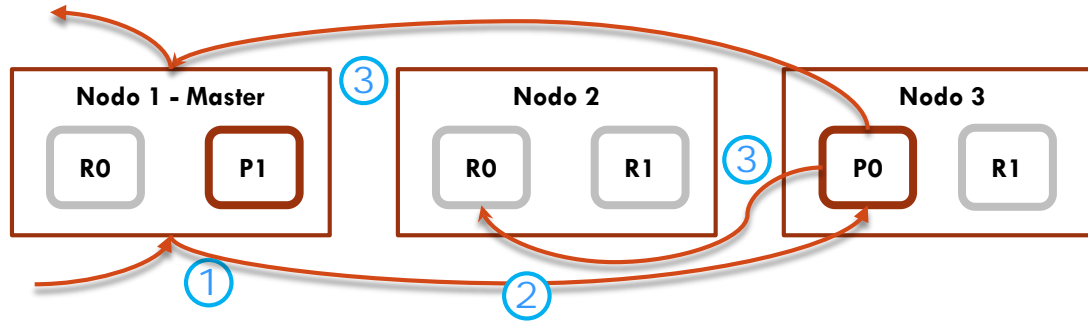
[Learn more](#)

Componentes

- La clusterización la construimos en función de nodos, cada uno de los cuales contiene los índices con su correspondiente partición (opcional). El número de particiones deberá ser igual al número de nodos y el número de réplicas $n-1$ si queremos alta disponibilidad total

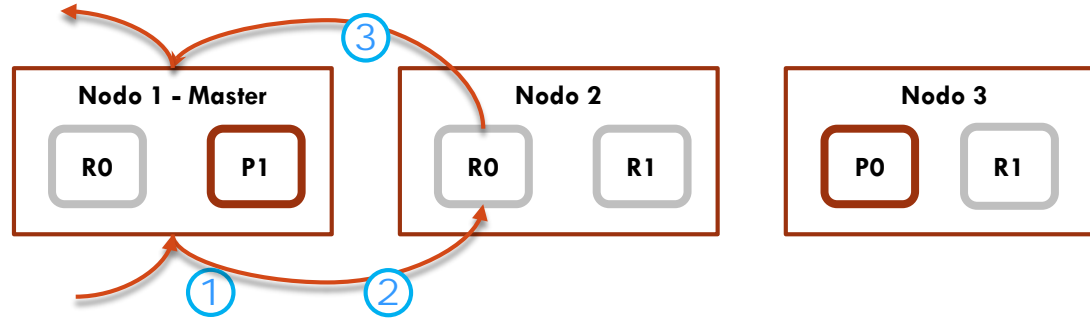


Crear, indexar y borrar un documento



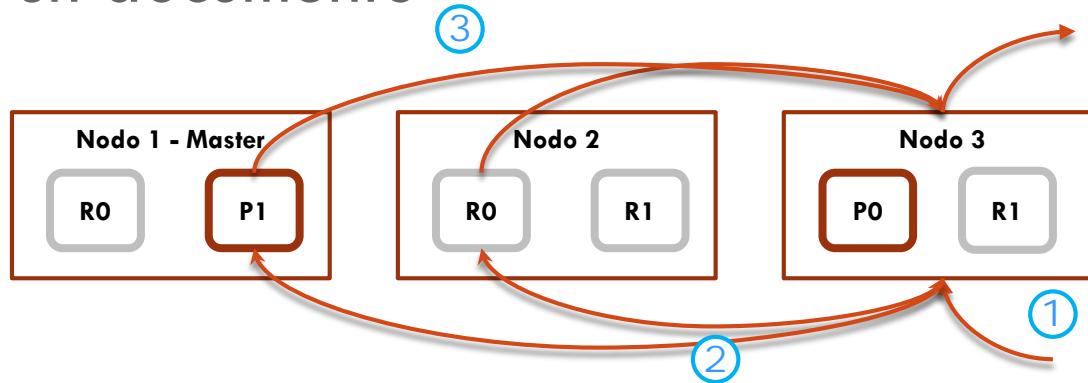
1. Se lanza una solicitud al Nodo1 de tipo CRUD
2. El nodo utiliza el `_id` del documento para determinar que el documento está en la partición/índice 0 y que ésta está alojada en el Nodo 3 (donde está la primera copia si esta disponible el nodo)
3. El Nodo 3 realiza la tarea de CRUD en su partición y si termina bien notifica en paralelo a los nodos 1 y 2 que deben realizarla. Una vez todos ellos devuelven su conformidad el Nodo 3 responde al nodo coordinador (el Nodo 1) y éste a su vez lo deberá hacer al cliente

Recuperar un documento



1. El cliente envía una solicitud al Nodo 1
2. El nodo utiliza el `_id` del documento para determinar que el documento está en la partición/índice 0. Aunque está disponible en todos los nodos selecciona el Nodo 2 para recuperar la información
3. El Nodo 2 devuelve el documento al Nodo 1 y éste lo devuelve a su vez al cliente

Buscar en un documento



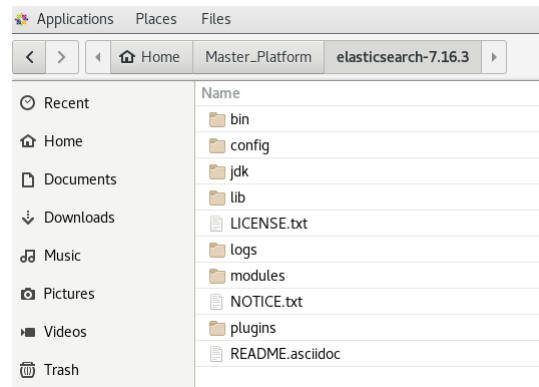
1. El cliente envía una solicitud de búsqueda al Nodo 3. Éste crea a su vez una búsqueda “vacía” en una cola de búsquedas
2. El Nodo 3 reenvía la solicitud a un nodo primario o replica de cada partición/índice. Cada nodo ejecuta la búsqueda en local y lo añade a una cola
3. Cada partición devuelve el doc id y ordena los documentos en su cola de prioridad. El Nodo 3 recopila toda la información recibida y devuelve la información ya ordenada

Elasticsearch for Apache Hadoop

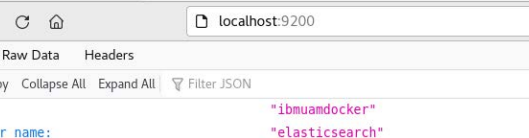
- Existe una versión específica para Apache Hadoop
- Es una implementación que modifica los procesos de lectura y escritura para poder utilizar HDFS y MapReduce
- La recuperación de información no es tan directa con GET/POST sino procesos java adecuadamente configurados

Vamos a utilizarlo

- Empecemos descargando el código binario para nuestra plataforma. <https://www.elastic.co/downloads/elasticsearch>
 - Windows: MSI
 - Linux: targz
- Lo descomprimos y lo desempaquetamos en el directorio que queramos



- Ejecutamos elasticsearch. Aparecerán muchos mensajes que incluirán dirección y puerto por el que acceder: localhost:9200.



The screenshot shows a web browser window with the address bar displaying 'localhost:9200/'. The browser's developer tools or a similar interface shows the response in JSON format. The JSON object contains the following fields and values:

```
{
  "name": "ibmuamdocker",
  "cluster_name": "elasticsearch",
  "cluster_uuid": "PL5A6BzXQ82WeBtVZWcl9Q",
  "version": {
    "number": "7.16.3",
    "build_flavor": "default",
    "build_type": "tar",
    "build_hash": "4e6e4eab2297e949ec994e688dad46290d018022",
    "build_date": "2022-01-06T23:43:02.825887787Z",
    "build_snapshot": false,
    "lucene_version": "8.10.1",
    "minimum_wire_compatibility_version": "6.8.0",
    "minimum_index_compatibility_version": "6.0.0-beta1"
  },
  "tagline": "You Know, for Search"
}
```

Prestemos atención a los mensajes

- En algunas ocasiones la versión de java instalada
- Valores por defecto de `ulimits` para el número de ficheros abiertos (suele estar por debajo del valor necesario)
- Dirección y puerto de consulta
- Las opciones de arranque están en `elasticsearch.yml`. Algunas vienen por defecto y hay que modificarlas explícitamente. Por ejemplo, podemos habilitar seguridad o indicar que el cluster es un único nodo:

```
xpack.security.enabled: false  
discovery.type: single-node
```
- Y estado del cluster

Veamos si va todo bien

Sintaxis: <HTTP Verb>/<Index>/<ID>

➤ Desde línea de comando:

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

➤ Desde navegador:

➤ https://localhost:9200/_cat

➤ https://localhost:9200/_cat/health

```
[umaster@ibmuamdocker bin]$ curl -X GET "localhost:9200/_cat/health?v"
epoch      timestamp cluster      status node.total node.data shards pri relo init unassign pending_tasks max_task_wait_time active_shards_percent
1642675981 10:53:01 elasticsearch green          1          1      3  3    0    0        0          0              -             100.0%
[umaster@ibmuamdocker bin]$
```

➤ ¿Tenemos índices?: `curl -X GET "localhost:9200/_cat/indices?v"`

```
[umaster@ibmuamdocker bin]$ curl -X GET "localhost:9200/_cat/indices?v"
health status index          uuid                                pri rep docs.count docs.deleted store.size pri.store.size
green  open   .geoip_databases GqPN7GCXT0eoAAiU3eHgDQ             1  0         42             0       40.4mb       40.4mb
```

Creamos un índice(*) y revisamos

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

```
[umaster@ibmuamdocker bin]$ curl -X PUT "localhost:9200/uamcustomer?pretty"
{
  "acknowledged" : true,
  "shards_acknowledged" : true,
  "index" : "uamcustomer"
}
```

```
[umaster@ibmuamdocker bin]$ curl -X GET "localhost:9200/_cat/indices?v"
```

health	status	index	uuid	pri	rep	docs.count	docs.deleted	store.size	pri.store.size
green	open	.geoip_databases	GqPN7GCXT0eoAAiU3eHgDQ	1	0	42	0	40.4mb	40.4mb
yellow	open	uamcustomer	khDNR0TRS0qE2KYzqyIPfw	1	1	0	0	226b	226b

- ¿Por qué el estado es yellow?

(*) la configuración por defecto de Elasticsearch es que si no existe el índice se cree automáticamente

Configuración de Elasticsearch (I)

- Muchos parámetros de la configuración de Elasticsearch se pueden modificar *en vuelo*
- Por ejemplo, la configuración de las réplicas del índice uamcustomer(*):

```
curl -X PUT "localhost:9200/uamcustomer/_settings?pretty" -H  
'Content-Type: application/json' -d
```

```
{
```

```
  "index" : {
```

```
    "number_of_replicas" : 0
```

```
  }
```

```
}'
```

(*) Ejemplos de estos comandos están disponibles en los ficheros de Moodle

Configuración de Elasticsearch (II)

- O, de forma genérica, para cualquier nuevo índice que creemos:

```
curl -X PUT http://localhost:9200/_template/default \-H
'Content-Type: application/json' \
-d \
'{
  "index_patterns": ["*"],
  "order": -1,
  "settings": {
    "number_of_shards": "1",
    "number_of_replicas": "0"
  }
}'
```

(*) Es conveniente hacerlo para evitar el problema del número de réplicas configuradas por defecto

Añadimos un registro y revisamos

```
curl -X PUT "localhost:9200/uamcustomer/_doc/1?pretty" -H 'Content-Type: application/json' -d'
```

```
{
  "SOURCE": "ERP",
  "NAME": "J",
  "LASTNAME": "JACKSON",
  "HOME_ADDRESS": "8388 SOUTH CALIFORNIA ST.",
  "HOME_CITY": "TUCSON",
  "HOME_STATE": "AZ",
  "HOME_ZIPCODE": 85708,
  "HOME_PHONE": "267-3352",
  "OFFICE_ADDRESS": "",
  "OFFICE_CITY": "ALLEN TON",
  "OFFICE_STATE": "MI",
  "OFFICE_ZIPCODE": 48002,
  "OFFICE_AREA_CODE": 810,
  "OFFICE_PHONE": "710-0470",
  "SCNUMBER": "369-98-6555",
  "SSNUMBER": "462-11-4610",
  "BIRTH_DATE": "1953-05-00",
  "GENRE": "F",
  "OTHER": ""
}
```

```
[umaster@ibmuamdocke ficheros]$ sh ./customer_1.sh
{
  "_index" : "uamcustomer",
  "_type" : "_doc",
  "_id" : "1",
  "_version" : 1,
  "result" : "created",
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "failed" : 0
  },
  "_seq_no" : 0,
  "_primary_term" : 1
}
[umaster@ibmuamdocke ficheros]$
```

Hacemos un search

```
curl -X GET "localhost:9200/uamcustomer/_doc/1?pretty"
```

http://localhost:9200/uamcustomer/_search?pretty

```
[umaster@ibmuamdocker ficheros]$ curl -X GET "localhost:9200/uamcustomer/_doc/1?pretty"
{
  "_index" : "uamcustomer",
  "_type" : "_doc",
  "_id" : "1",
  "_version" : 1,
  "_seq_no" : 0,
  "_primary_term" : 1,
  "found" : true,
  "_source" : {
    "SOURCE" : "ERP",
    "NAME" : "J",
    "LASTNAME" : "JACKSON",
    "HOME_ADDRESS" : "8388 SOUTH CALIFORNIA ST.",
    "HOME_CITY" : "TUCSON",
    "HOME_STATE" : "AZ",
    "HOME_ZIPCODE" : 85708,
    "HOME_PHONE" : "267-3352",
    "OFFICE_ADDRESS" : "",
    "OFFICE_CITY" : "ALLENTON",
    "OFFICE_STATE" : "MI",
    "OFFICE_ZIPCODE" : 48002,
    "OFFICE_AREA_CODE" : 810,
    "OFFICE_PHONE" : "710-0470",
    "SCNUMBER" : "369-98-6555",
    "SSNUMBER" : "462-11-4610",
    "BIRTH_DATE" : "1953-05-00",
    "GENRE" : "F",
    "OTHER" : ""
  }
}
```



Field	Value
took	9
timed_out	false
shards	{ total: 1, successful: 1, skipped: 0, failed: 0 }
hits	{ total: 1, value: { _index: 'uamcustomer', _type: '_doc', _id: '1', _score: 1, _source: { SOURCE: 'ERP', NAME: 'J', LASTNAME: 'JACKSON', HOME_ADDRESS: '8388 SOUTH CALIFORNIA ST.', HOME_CITY: 'TUCSON', HOME_STATE: 'AZ', HOME_ZIPCODE: 85708, HOME_PHONE: '267-3352', OFFICE_ADDRESS: '', OFFICE_CITY: 'ALLENTON', OFFICE_STATE: 'MI', OFFICE_ZIPCODE: 48002, OFFICE_AREA_CODE: 810, OFFICE_PHONE: '710-0470', SCNUMBER: '369-98-6555', SSNUMBER: '462-11-4610', BIRTH_DATE: '1953-05-00', GENRE: 'F', OTHER: '' } } }

Algunas tareas adicionales

- ¿Como se modifica un registro?:
 - Cargando uno nuevo con el mismo ID (o realizando modificaciones de un campo en concreto)
- ¿Necesito asignar un id a cada documento?
 - No, si no lo sabes dejas a Elasticsearch que genere uno aleatoriamente (en ese caso no es PUT, es POST)
- ¿Cómo borro un índice o un registro?
 - `curl -X DELETE "localhost:9200/uamcustomer"`
 - `curl -X DELETE "localhost:9200/uamcustomer/_doc/1"`

```
{
  "_index": "uamcustomer",
  "_type": "doc",
  "_id": "28dqd34B-3-gjXFLLP0m",
  "_version": 1,
  "result": "created",
  "_shards": {
    "total": 1,
    "successful": 1,
    "failed": 0
  },
  "_seq_no": 4,
  "_primary_term": 1
}
```

Elasticsearch. Principales características



- Datos en tiempo real
- Analítica en tiempo real
- Libre de esquema
- Orientado a documento
- Alta Disponibilidad
- ...

Kibana

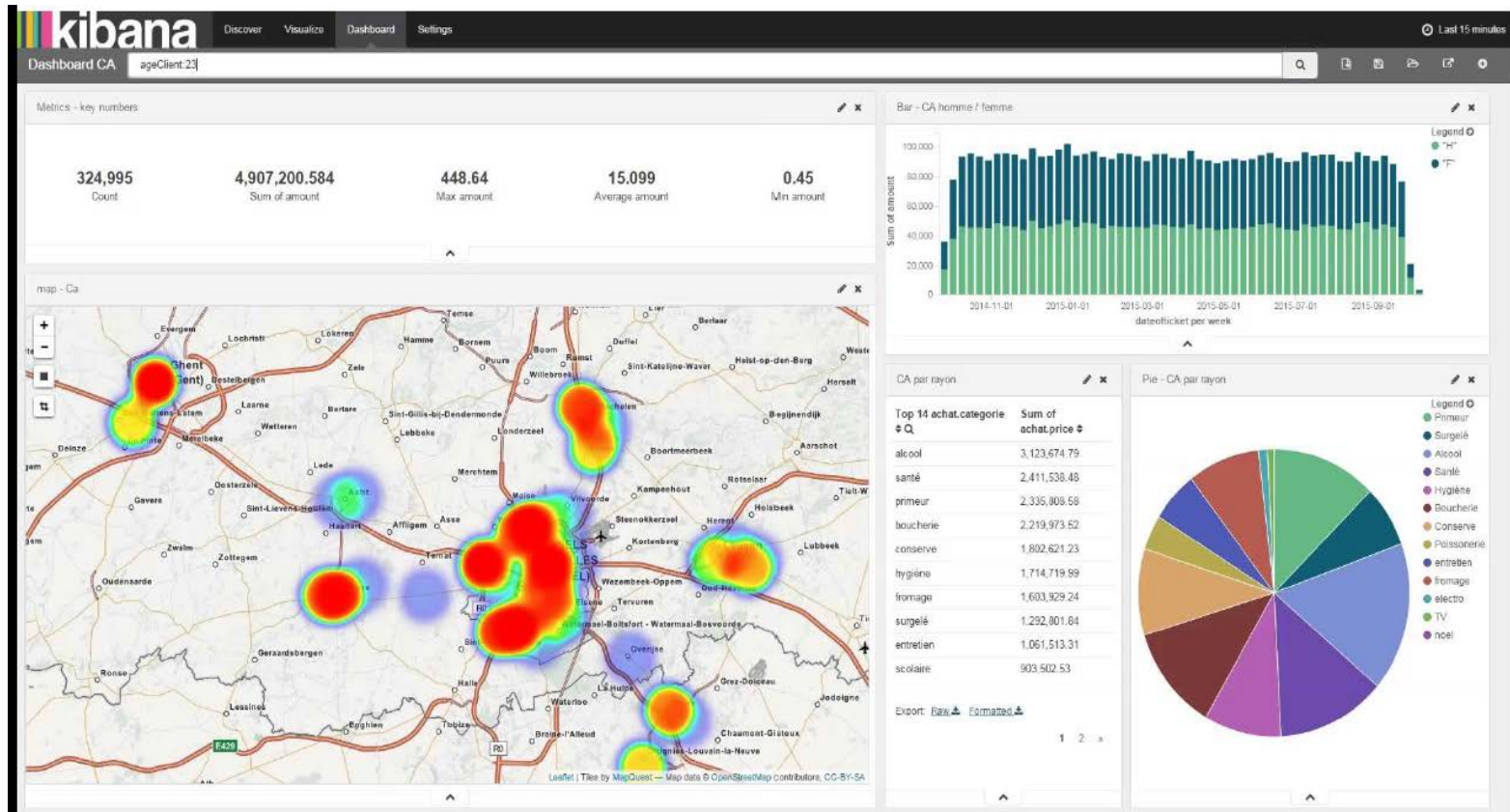
- Es un aplicación/entorno cliente de Elasticsearch
- Permite una visualización y análisis de datos casi en tiempo real
- Es una exploración *interactiva* dado que nos permite cambiar los parámetros de selección/consulta e incluye por defecto el refresco temporal de las gráficas
- Incluye múltiples tipos de visualización
- Su uso no requiere programación o conocer un lenguaje de consulta en la mayoría de los casos

Usamos Kibana

➤ Para gráficos interactivos



Con mucha riqueza visual



Probando Kibana

- Descargamos de <https://www.elastic.co/downloads/kibana>
- La configuración del servicio se localiza en el fichero kibana.yml en el directorio config. La configuración por defecto que viene es válida y aceptable para nuestros propósitos.
- Es importante tener la directiva de seguridad inhabilitada en elasticsearch de forma específica (colocando la correspondiente directiva en el fichero elasticsearch.yml)
- Arrancamos kibana y accedemos por el puerto 5601

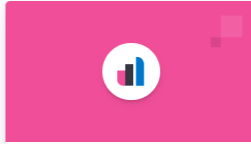
Consola de desarrollo en Kibana

Welcome home



Enterprise Search

Create search experiences with a refined set of APIs and tools.



Observability

Consolidate your logs, metrics, application traces, and system availability with purpose-built UIs.



Security

Prevent, collect, detect, and respond to threats for unified protection across your infrastructure.



Analytics

Explore, visualize, and analyze your data using a powerful suite of analytical tools and applications.

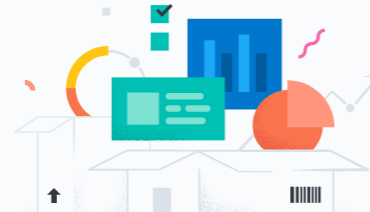
Get started by adding integrations

To start working with your data, use one of our many ingest options. Collect data from an app or service, or upload a file. If you're not ready to use your own data, add a sample data set.

[+ Add integrations](#)

[Try sample data](#)

[Upload a file](#)



Management



Manage permissions

Control who has access and what tasks they can perform.



Monitor the stack

Track the real-time health and performance of your deployment.



Back up and restore

Save snapshots to a backup repository, and restore to recover index and cluster state.



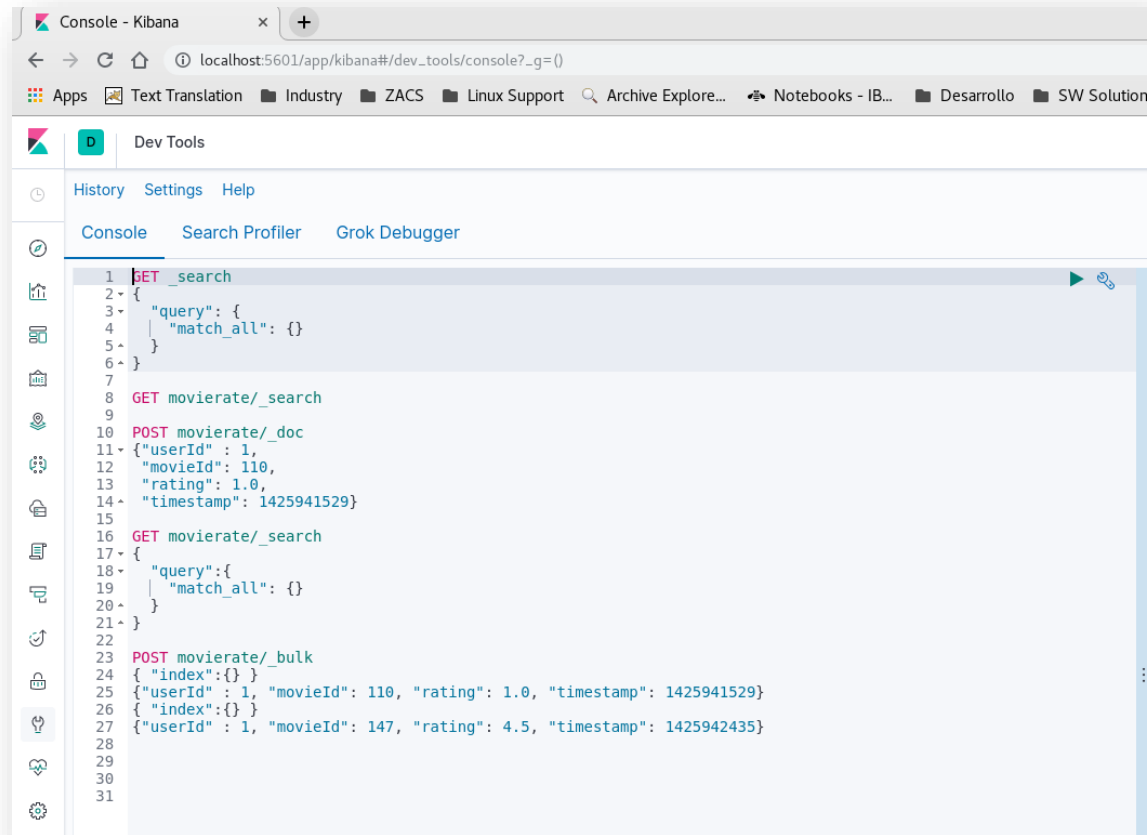
Manage index lifecycles

Define lifecycle policies to automatically perform operations as an index ages.

[Dev Tools](#)

[Stack Management](#)

Ejemplo



The screenshot shows the Kibana Dev Tools console with the following REST API calls:

```
1 GET _search
2 {
3   "query": {
4     "match_all": {}
5   }
6 }
7
8 GET movierate/_search
9
10 POST movierate/_doc
11 {
12   "userId": 1,
13   "movieId": 110,
14   "rating": 1.0,
15   "timestamp": 1425941529
16 }
17 GET movierate/_search
18 {
19   "query": {
20     "match_all": {}
21   }
22 }
23 POST movierate/_bulk
24 {
25   "index": {}
26 }
27 {
28   "userId": 1, "movieId": 110, "rating": 1.0, "timestamp": 1425941529
29 }
30 {
31   "index": {}
32 }
33 {
34   "userId": 1, "movieId": 147, "rating": 4.5, "timestamp": 1425942435
35 }
36 {
37   "index": {}
38 }
39
```

Otras opciones

```
POST /customer/_doc/1/_update?pretty
{
  "doc": { "name": "Jane Doe" }
}
```

```
POST /customer/_doc/1/_update?pretty
{
  "doc": { "name": "Jane Doe", "age": 20 }
}
```

```
POST /customer/_doc/1/_update?pretty
{
  "script" : "ctx._source.age += 5"
}
```

Carga en batch

```
curl -X PUT "localhost:9200/movierate?pretty"
```

```
curl -X POST -H "Content-Type: application/json" \  
"localhost:9200/movierate/_doc/_bulk?pretty&refresh" \  
--data-binary "@./ratings_total.json"
```

```
[umaster@ibmuamdocke ficheros]$ cat ratings_total.json  
{ "index":{} }  
{"userId" : 1, "movieId": 110, "rating": 1.0, "timestamp": 1425941529}  
{ "index":{} }  
{"userId" : 1, "movieId": 147, "rating": 4.5, "timestamp": 1425942435}
```

Ejemplo de consulta

```
curl -X GET "localhost:9200/movierate/_search/?pretty" \
-H 'Content-Type: application/json' \
-d '{ "query": { "match": {"movieId": 110} } }'
```

```
"hits" : [
  {
    "_index" : "movierate",
    "_type" : "_doc",
    "_id" : "4cdFeH4B-3-gjXFLlP3d",
    "_score" : 1.0,
    "_source" : {
      "userId" : 1,
      "movieId" : 110,
      "rating" : 1.0,
      "timestamp" : 1425941529
    }
  }
]
```

Más consultas

```
GET /bank/_search
{
  "query": { "match_all": {} },
  "_source": ["account_number", "balance"]
}
```

```
GET /bank/_search
{
  "query": { "match": { "account_number": 20 } }
}
```

```
GET /bank/_search
{
  "query": { "match": { "address": "mill" } }
}
```

Más consultas

```
GET /bank/_search
{
  "query": { "match": { "address": "mill lane" } }
}
```

```
GET /bank/_search
{
  "query": { "match_phrase": { "address": "mill lane" } }
}
```

Más consultas

```
GET /bank/_search
{
  "query": {
    "bool": {
      "must": [
        { "match": { "address": "mill" } },
        { "match": { "address": "lane" } }
      ]
    }
  }
}
```

Must o Should o Must not

En resumidas cuentas. SolR vs Elasticsearch

Topic	SolR	Elasticsearch
Indexación	Lucene	Lucene
Arquitectura	Distribuida, escalable, tolerante a fallos. Incluye ZooKeeper	Distribuida, escalable, tolerante a fallos. Solo nodos Elasticsearch
Agrupación	<i>Colección</i>	<i>Índices</i>
Particiones y réplicas	Soportadas	Soportadas
Consultas con o sin enrutamiento	Disponible	Disponible
Discovery	ZooKeeper	Zen Discovery

En resumidas cuentas. SolR vs Elasticsearch

Topic	SolR	Elasticsearch
API	HTTP con Query String	HTTP REST API con Query String
Consultas	Parámetros en la consulta	Objetos JSON
Formato de Datos	Múltiples formatos	JSON in / JSON out
Documento / Batch	Independiente	Distintos tipos de carga
Caching	Disponible	Disponible
Cluster monitoring	Adicional, no incluido por defecto	Adicional, no incluido por defecto
Soporte de Hadoop	Proceso de indexación	Almacenamiento y recuperación de documentos en HDFS/MapReduce

Mas información

- [Web principal](#)
- [Descargas](#)
 - [Elasticsearch](#)
 - [Kibana](#)
- **Soporte Stackoverflow**
 - [Elasticsearch](#)
 - [Kibana](#)