Geeks México

BLOG DE PROGRAMACIÓN EN ESPAÑOL SOBRE JAVA, FRAMEWORKS, BASES DE DATOS, CÓMPUTO EN LA NUBE, ETC. EN ESPAÑOL Y EN INGLÉS.

Aprende a configurar un cluster de Solr,

indexar documentos y realizar búsquedas en Español !

HACE 3 DÍAS DEJA UN COMENTARIO

Solr es un motor de búsqueda de código abierto del proyecto Lucene que expone sus servicios a través de una interfaz REST, este permite almacenar(indexar) documentos vía JSON,XML, CSV o binarios a través de HTTP con un increíble performance, en este post explicaremos como instalarlo, configurarlo, navegar en el, indexar documentos y realizar búsquedas sobre los mismos.

A continuación se presentan algunos casos de uso de Solr en algunas de las compañías más importantes del mundo:

- Instagram: Utiliza Solr para generar las búsquedas geo referenciadas.
- WhiteHouse.gov: El sitio web de la casa blanca funciona utilizando Drupal y Solr.
- Netflix: Utiliza Solr para la búsqueda de películas.

Paso 1 Descarga de Solr

En este post descargaremos la versión 7.2.1, pero puedes encontrar cualquier versión en el siguiente enlace http://archive.apache.org/dist/lucene/solr/ (http://archive.apache.org/dist/lucene/solr/), una vez que decidimos la versión a instalar ejecutaremos los siguientes comandos para descargarla e instalarla:

```
wget http://archive.apache.org/dist/lucene/
unzip solr-7.2.1.zip
cd solr-7.2.1
```

Con lo anterior descargaremos Solr, lo descomprimiremos y accederemos a su directorio.

Paso 2 Iniciar el servidor

Una vez que entendemos lo que significan los directorios el siguiente paso será iniciar el servidor de Solr, para hacerlo ejecutaremos el siguiente comando:

Linux o MacOS:

```
./bin/solr start -e cloud
```

Windows:

```
bin\solr.cmd start -e cloud
```

En cualquiera de los dos casos veremos una salida como la siguiente:

```
Welcome to the SolrCloud example!

This interactive session will help you laur
To begin, how many Solr nodes would you lik
```

Como se puede ver el prompt nos solicitará el número de nodos que deseamos ejecutar, al final de la línea se puede ver un número 2 lo cual significa que se utilizarán 2 nodos en el cluster, para este ejemplo es suficiente así que solo oprimiremos enter.

El siguiente paso será determinar en que puerto se ejecutará cada uno de los nodos, como se puede ver del lado derecho Solr propondrá los nodos a utilizar, si estas de acuerdo y no tienes ningún otro proceso ejecutándose en esos puertos solo presiona enter dos veces como se muestra a continuación:

```
Please enter the port for node1 [8983]:

Please enter the port for node2 [7574]:
```

Una vez seleccionados los puertos para ambos nodos veremos una salida como la siguiente:

```
Creating Solr home directory /ruta/solr-7.
Cloning /ruta/solr-7.2.1/example/cloud/noc
   /ruta/solr-7.2.1/example/cloud/node2

Starting up Solr on port 8983 using commar
"bin/solr" start -cloud -p 8983 -s "exampl

Waiting up to 180 seconds to see Solr runr
Started Solr server on port 8983 (pid=6481

Starting up Solr on port 7574 using commar
"bin/solr" start -cloud -p 7574 -s "exampl

Waiting up to 180 seconds to see Solr runr
Started Solr server on port 7574 (pid=6491

INFO  - 2018-01-31 10:48:04.558; org.apach

Now let's create a new collection for inde
Please provide a name for your new collect
```

Solr utiliza Zookeeper para funcionar pero como se puede ver no especificamos los detalles de un zookeeper externo, por esto Solr iniciará su propio Zookeeper y conectará los dos nodos que creamos a este.

Una vez hecho esto el servidor de Solr esta listo para usarse y nos solicitará una colección para empezar a indexar documentos en nuestro cluster, para el ejemplo escribiremos la palabra **movies** y presionaremos enter.

El siguiente paso será elegir el número de shards a utilizar por nuestra colección, el default son 2 (Esto es

porque definimos 2 nodos en el cluster) así que utilizaremos 2 para el ejemplo y presionaremos enter.

```
How many shards would you like to split mov
```

La siguiente pregunta será ¿Cuántas replicas por shard deseamos crear?, una replica es una copia del indice, esto se utiliza para la tolerancia a fallas, de nuevo el default es 2 así que oprimiremos enter para continuar.

```
How many replicas per shard would you like
```

Una vez definidas las réplicas Solr nos preguntará por la configuración a utilizar para la colección movies que creamos.

```
Please choose a configuration for the movie
_default or sample_techproducts_configs [_d
```

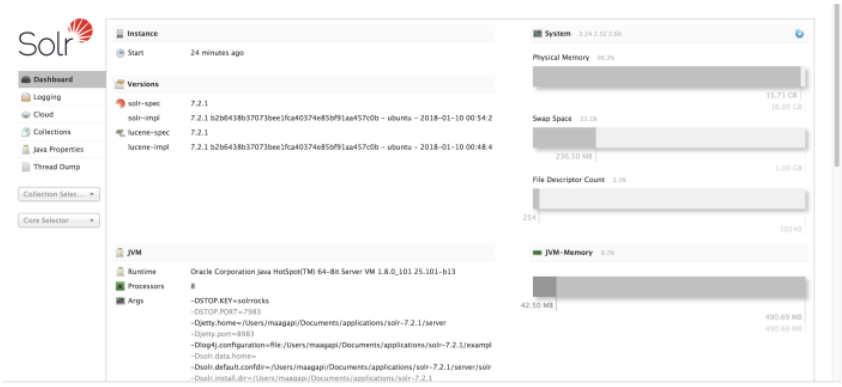
En este caso seleccionaremos _default para utilizar la configuración por default de solr y presionaremos enter, esto mostrará la siguiente salida:

```
Created collection 'movies' with 2 shard(s)

Enabling auto soft-commits with maxTime 3 s

POSTing request to Config API: http://local
{"set-property":{"updateHandler.autoSoftCom
Successfully set-property updateHandler.aut

SolrCloud example running, please visit: ht
```

Lo anterior nos indica que la colección movies fur creada de forma correcta en 2 shards y con 2 réplicas.

Solr provee una interfaz gráfica para administrar el cluster, en este caso creamos dos nodos, esto significa que tendremos dos url's para administrar nuestro cluster, estas son http://localhost:8983/solr/ (http://localhost:8983/solr/#/) y http://localhost:7574/solr/ (http://localhost:7574/solr/) , como se puede ver la única

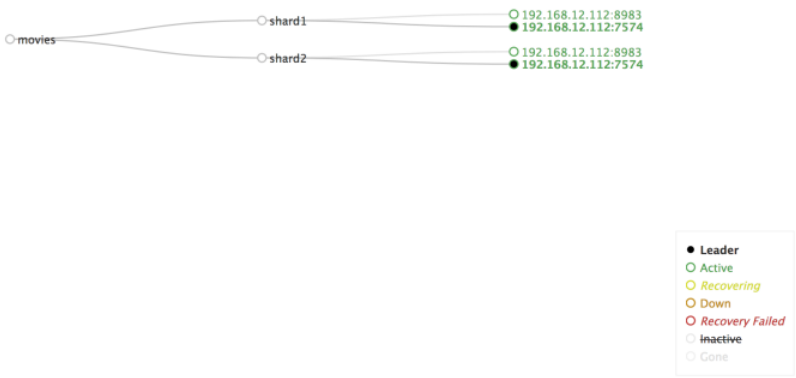
diferencia entre una y otra es el puerto en el que se ejecutan, los puertos son los que definimos previamente cuando creamos los nodos, una vez que accedemos a alguna de las urls desde nuestro navegador veremos una interfaz como la que se muestra en la siguiente imagen:



Con esto ya estas listo para empezar a indexar datos !.

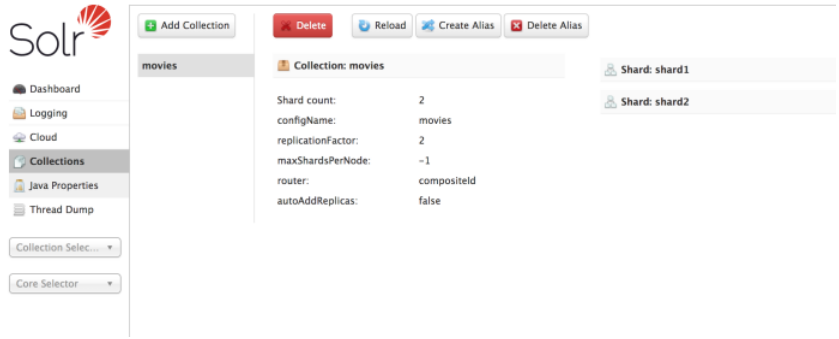
Paso 3 Analizando nuestro cluster

Una vez que ejecutamos los pasos anteriores ya tenemos un cluster ejecutándose en nuestra máquina, si presionamos la pestaña Cloud veremos lo siguiente:



Esto muestra que tenemos una colección llamada movies distribuida en 2 shards y cada uno contiene 2 replicas de la información.

Si accedemos a la pestaña Collections y seleccionamos la colección movies veremos lo siguiente:



Esta es la información de nuestra colección movies, ahora el siguiente paso será indexar datos en este cluster.

Paso 4 Indexar información

En este ejemplo indexaremos información sobre películas en nuestro cluster, para esto utilizaremos la información que proporciona https://movielens.org/ (https://movielens.org/), este sitio se encarga de mostrar películas, ratings y recomendaciones. Para utilizar su información accederemos al sitio https://grouplens.org/datasets/movielens/ (https://grouplens.org/datasets/movielens/) y descargaremos el dataset llamado ml-latest-small.zip (http://files.grouplens.org/datasets/movielens/ml-latest-small.zip), en mi caso descargué el archivo en el directorio de sonar con los siguientes comandos:

```
1 wget http://files.grouplens.org/datasets/mo
2 unzip ml-latest-small.zip
```

Esto descargará el dataset y lo descomprimirá, el zip contiene los siguientes archivos:

- README.txt
- links.csv
- **movies.csv**
- ratings.csv
- tags.csv

El archivo que indexaremos en nuestro cluster de Solr será movies.csv utilizando el siguiente comando:

```
./bin/post -c movies /tu_ruta/ml-latest-sma
```

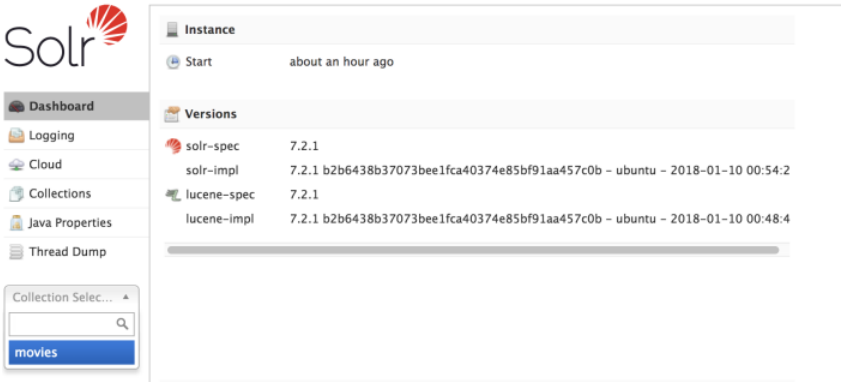
Con esto tendremos una salida como la siguiente:

```
SimplePostTool version 5.0.0
Posting files to [base] url http://localhos
Entering auto mode. File endings considered
POSTing file movies.csv (text/csv) to [base
1 files indexed.
COMMITting Solr index changes to http://loc
Time spent: 0:00:02.314
```

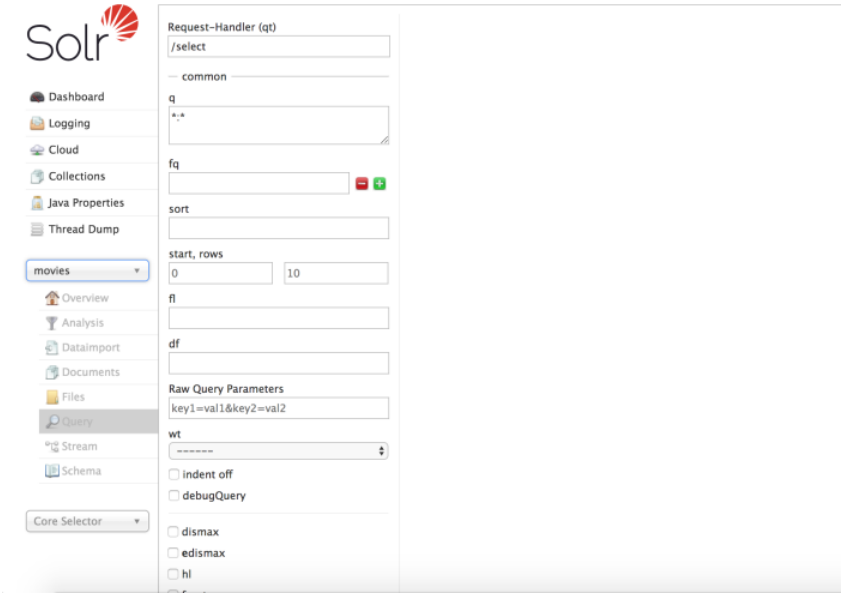
Con esto nuestro document estará indexado de forma correcta en nuestro cluster.

Paso 5 Búsqueda de información

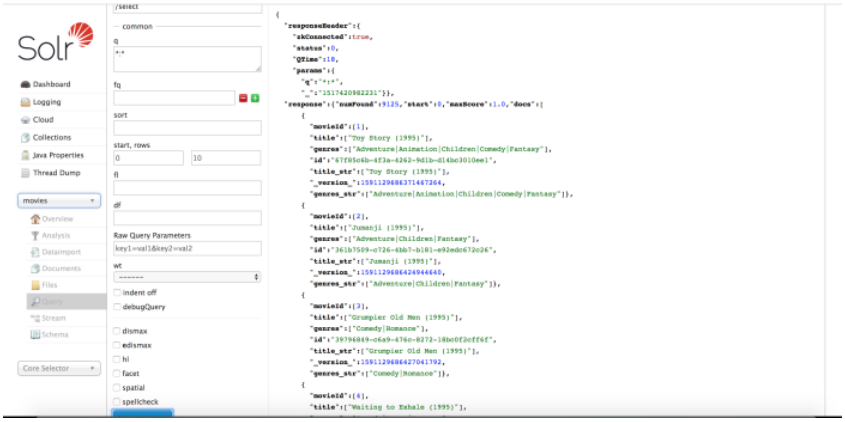
Una vez que la información fue indexada correctamente, el siguiente paso será realizar búsquedas, esto se puede hacer desde la interfaz de usuario, para hacerlo seleccionaremos la colección movies en el primer selector de la izquierda, como se muestra en la siguiente imagen:



Esto mostrará un menú en el que seleccionaremos la opción query, esto nos mostrará la siguiente ventana:



El query por default mostrará toda la información almacenada en la colección, para hacer una prueba solo presiona el botón Execute Query, este mostrará una salida como la siguiente:



Como se puede ver la respuesta contiene la información de las películas que indexamos, esta contiene un atributo llamado numFound el cual contiene 9125 registros que son los que contiene el archivo de excel.

El valor **"QTime":18** significa que desde que el servidor recibió nuestra solicitud hasta que dio la respuesta solo tomó 18 milisegundos lo cuál es un performance increíble.

Más adelante crearemos un post para explicar como ejecutar consultas más complejas a la información.

Si te gusta el contenido y quieres enterarte cuando realicemos un post nuevo síguenos en nuestras redes sociales https://twitter.com/geeks_mx (https://twitter.com/geeks_mx) y https://www.facebook.com/geeksJavaMexico/ (https://www.facebook.com/geeksJavaMexico/).


Wizink
Tu banco senzillo