Zaragoza
AYUNTAMIENTO

Nº Expte.: 183028-18

OFICINA TÉCNICA DE PARTICIPACIÓN, TRANSPARENCIA Y GOBIERNO ABIERTO

- ***Sobre el proceso de crawling. Si el proceso actual está hecho también en java, entiendo que toda la parte de crawling y extracción de datos podría reutilizarse, y solamente habría que rehacer la parte que lo almacena en Apache Solr para que lo almacene en Elasticsearch (añadiendo la información adicional de número de visitas y si está adaptado o no a móvil). ¿Es correcto?***

  Es correcto

- ***Otro tema que nos preocupa es el rendimiento. En el caso del proceso de crawling, ¿se está reindexando diariamente toda la información, o se dispone de información de los contenidos nuevos, los modificados, los eliminados, etc. desde la indexación anterior, de forma que solo se reindexan los contenidos que han cambiado? Esto es importante, porque si estamos hablando de 250.000 documentos, y ahora para cada uno de ellos antes de reindexarlo hay que realizar dos llamadas adicionales (una para obtener el número de visitas y otra para obtener si está adaptado a móvil o no), esto puede ser un problema si hay que reindexar cada día los 250.000 documentos, puesto que el proceso de indexación se puede alargar mucho.***

  El proceso de crawling de la sede electrónica se ejecuta por la noche una vez por semana

- ***Y en el caso de la hemeroteca estamos hablando de 750.000 ficheros, que habrá que tratar (bien el archivo tiff o bien el archivo ALTO XML) antes de indexarlos. El proceso de indexación de toda esta información será muy costoso en tiempo y tendrá grandes requerimientos de memoria. En caso de que el nuevo motor de indexación requiriera más RAM que el actual, mayor número de nodos de Elasticsearch, etc. para soportar todo esto, entiendo que podrá ampliarse la infraestructura actual. ¿Es correcto?***

  Sí, la infraestructura podría ampliarse.

- **Y finalmente, ¿existe algún requerimiento en cuanto a la ejecución del proceso de crawling y de indexación de la hemeroteca, en cuanto al tiempo máximo en el que debe ejecutarse y número de veces al día? Un proceso tan pesado como este desde nuestro punto de vista debe ejecutarse en una máquina separada y en horario de baja carga, puesto que estamos hablando de un proceso que va a tardar varias horas en ejecutarse y va a tener un uso intensivo de RAM y CPU**

Zaragoza
AYUNTAMIENTO

Se asume que es un proceso batch pesado, cuya ejecución se realizará desde el servidor de trabajos batch y se coordinara con el equipo técnico municipal.

El proceso de indexación de la hemeroteca solo se ha de ejecutar una vez, puesto que los contenidos no van a cambiar.

I.C. de Zaragoza a 27 de noviembre de 2018

El Jefe de la Unidad de
Gestión de la Web Municipal

Fdo.: Víctor Morlán Plo

La jefa de la Oficina Técnica de
Participación, Transparencia y
Gobierno Abierto,

Fdo.: Mª Jesús Fernández Ruíz