EDIT - feature request #4182

use elasticsearch as search engine [DISCUSS]

04/10/2014 08:15 AM - Andreas Kohlbecker

Status:	New		**Start date:**	
Priority:	New		**Due date:**	
Assignee:	Andreas Müller		**% Done:**	0%
Category:	architecture		**Estimated time:**	0:00 hour
Target version:	Unassigned CDM tickets			
Severity:	normal			

Description

Links:

- elasticsearch

- on how to let hibernate send modified entities to a different search backend, slor in this case:
  http://anotherjavaday.blogspot.co.il/2012/01/combining-power-of-hibernate-search-and.html

- https://github.com/55social/hibernate-search-elasticsearch-connector

Questions

- We need to run searches across multiple indexes, is elasticsearch supporting this? Currently we have our own implementation for this: LuceneMultiSearch

- Our own LucenSearch class is capable of returning lucene documents together with cdm entities in the same response. With elastic search this will no longer be possible. So we would have to send two requests: One request to do the search and a second one to fetch cdm entities, if necessary.

- It will be necessary to install a separate Elasticsearch server in parallel to the cdmserver. This makes the initial installation more complex. Can we bundle the cdm server and elastic search together in a single install package. How will this work in workshops. Simple HQL based search as fall back?

History

#1 - 04/10/2014 10:22 AM - Cherian Mathew

I think this is a really good idea. I have been doing some research on this, from the point of of the Data Quality Initiative, but also for the CDM hibernate connection.

It seems that the pushing data from spring-hibernate to elastic search will not be very difficult. In addition to the library you have mentioned, we could also write our own CDM specific Elasticsearch bridge as mentioned here

The above link also mentioned pulling data the other way around, using an Elasticsearch River":http://www.elasticsearch.org/guide/en/elasticsearch/rivers/current/, but there is new project under-development to replace this, called "Gatherer This idea could be useful in certain situations, but probably not for real time editing.

One of the main issues I found with Elasticsearch (related to the CDM) is joins. As mentioned on their page for managing relations the only way to effectively do this is to pre-process the data and store the data as parent-child relationships. This does effect performance and considering the complexity of the CDM, this may not be a good idea for pure CDM entities.

There is already a Elasticsearch Java API for accessing an existing cluster. For the server itself, since Elasticsearch is written in Java and is open source, I'm sure we can find a way to package and launch the Elasticsearch server from within the CDMserver.

From a broader perspective, I think we should separate the so-called 'pure' CDM entities and other DTO objects. We can then index most (if not all) CDM entities in Hibernate Lucene and index all kinds of other objects (only for quick reading) in Elasticsearch. This would be very flexible since we would then do all 'CUD' operations via the CDM and all the 'R' operations via customised DTOs stored in Elasticsearch.

Personally, I am super keen for this idea but it will also be a kind of full-fledged project, so we should discuss this in more detail.

#2 - 10/13/2016 04:36 PM - Andreas Müller

- Description updated

- Target version set to Unassigned CDM tickets

#3 - 11/19/2020 05:16 PM - Andreas Kohlbecker

- Private changed from Yes to No

#4 - 04/20/2021 12:09 PM - Andreas Kohlbecker

- Tags set to search