How to do a partial solr re-index.

Environment

Alfresco 4.2 and 5.0
Solr
Tomcat

Problem

For a production environment, with large number of documents and many indexes, it is sometime not feasible to do a full re-index.
Since users are using the system, a full re-index will slow down the system and disturb the usage of the system.
Full re-index can take up to 2-3 days depending on the size of the index folder.

Cause

The size of the index folder, will affect the time it takes to re-index.
The hardware and available memory, will have an impact on the time to do the full re-index.
Configuration settings will slow/improve performance of the full or partial re-indexing.

Solution

Option 1: Use the following links to find information on the current status of the Solr indexes:
1. https://localhost:8443/solr/admin/cores?action=SUMMARY&wt=xml
2. https://localhost:8443/solr/admin/cores?action=REPORT&wt=xml
3. https://localhost:8443/solr/admin/cores?action=STATUS&wt=xml

Once the REPORT is generated, this will show the number of missing transactions from the Index:
<long name="Count of missing transactions from the Index">5</long>

If the transaction was missing from SOLR indexes we would see an entry similar to the one below in the report:
<long name="First transaction missing from the Index">3248</long>

If the transaction is missing you can force SOLR to index that transaction with the following URL request, replacing the txid value with the missing transaction id.
https://localhost:8443/solr/admin/cores?action=REINDEX&txid=3248

If multiple transactions are missing then use the FIX report to check and fix all missing transactions and ACLs.
https://localhost:8443/solr/admin/cores?action=FIX
This will take some time depending on how many missing transaction you have. Once the FIX is done, a day needs to be given for Solr to complete the re-index.

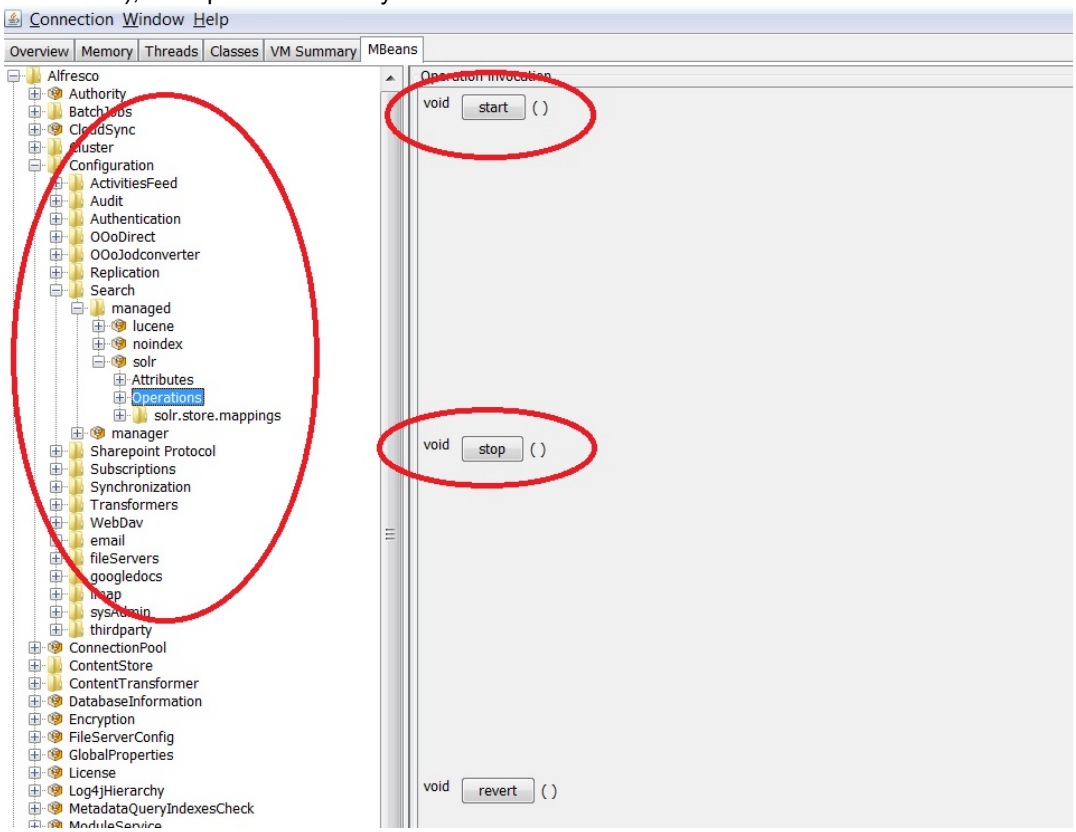
Useful configuration changes to speed up the Solr re-indexing process:
1. Set the corePoolSize to '6': solr/workspace-SpacesStore/conf/solrcore.properties:
alfresco.corePoolSize=6 - This will help with the re-indexing performance.
2. Increase the mergefactor on solrconfig.xml (specially for workspace). The default value is 3 but for big repositories it could be increased to 5, 8 up to 25, depending on the number of documents and indexes.
3. Check that you have this in your JAVA_OPTS: -XX:MaxPermSize=512m (both on Solr and on Alfresco).
4. Make sure you have sufficient over all RAM on the Server and allocated to the JVM, depending on how much you have available.
For more information please take a look at "Alfresco tuning" wiki.
5. Set the MultiThreadedTracking to 'false' before starting the full re-indexing for faster performance:
alfresco.enableMultiThreadedTracking=false
This can be set back to 'true' once the re-indexing is done.

Option 2: Do the full re-index on another path (node), in parallel with the production live system running as usual.
Please follow these steps:
1. install the same version of Solr which you are using in the live environment on a separate Tomcat
2. Point this Solr to the Alfresco instance from the live environment.

3. Run the full re-index from this Solr (make sure this Solr is communicating with Alfresco).
4. Monitor the full re-index process until it is complete (Check the SUMMARY.xml page).
5. Copy the 'index' folder from the \solr\workspace\SpacesStore and \solr\archive\SpacesStore to the live Solr from the live environment and you should have healthy newly rebuild indexes.

i) Before copying the 'index' folder stop both the source and destination Solrs.

ii) Use the the jconsole (please note that you should only stop the solr subsystem - don't stop the tomcat), to stop the Solr subsystem.



iii) In your jconsole, wait till the message 'successfully stopped' appears.

6. copy the 'index' folder to the destination Solr
7. Start solr subsystem from the joconsole (see the image above).

The destination Solr should now have the full re-index indexes and it will continue index new content (documents).

Upload

Related Information

Article: 000004267 - "How to PURGE/INDEX/REINDEX a particular node with SOLR without any access to database"