

Setting up Ensembl Solr Search on an Ubuntu external server

22nd November 2013
(edited 10th July 2018)

Background

Ensembl uses Solr Apache as its search engine. We provide the indexes and configuration files on our FTP site, and these instructions as to how to deploy the search for your own mirror. These instructions relate to Solr version 3.6.1.

There are a large number of records in the index for a solr release, so for both maintenance and performance we separate these into nine shards. As of Ensembl release 93 (July 2018) the human variation indexes are split into three shards for performance reasons. A mirror site can have one or all of these installed:

- (i) ensembl_ga: genomic alignments for all species
- (ii) ensembl_variation_mouse: variation database records for mouse
- (iii) ensembl_variation_cow: variation database records for cow
- (iv) ensembl_variation_other: variation database records for all other species
- (v) ensembl_core - records for all other datatypes (gene IDs, clone IDs, GeneTRee IDs, Regulatory features, etc)
- (vi) ensembl_docs - search robot crawled content from the site
- (vii) ensembl_variation_human_1: variation database records (rsIDs, associated phenotypes and disease, etc) for human
- (viii) ensembl_variation_human_2: variation database records (rsIDs, associated phenotypes and disease, etc) for human
- (ix) ensembl_variation_human_3: variation database records (rsIDs, associated phenotypes and disease, etc) for human

There are two request handlers for each shard - 'ensembl' which returns just the records for that shard, and 'ensemblshards' which returns records across all. This is useful for debugging, but in practise we just use the 'ensemblshards' request handler for the ensembl-core shard.

Requirements

Server Requirements

Ubuntu Linux Server (Ubuntu 12.04.3) 32Gb RAM. Approx 600Gb Disk (in our experience performance is greatly improved if you can use SSD rather than magnetic)

Software

Java JDK
Tomcat 7
Solr 3.6.1
Web client

Components

1) Java JDK

You need to a JDK on the server. To check if you have one installed:

```
$ java -version
```

If you have you should see something like this

```
java version "1.7.0_25"  
OpenJDK Runtime Environment (IcedTea 2.3.12) (7u25-2.3.12-4ubuntu3)  
OpenJDK 64-Bit Server VM (build 23.7-b01, mixed mode)
```

If you need to install a Java JDK we recommend OpenJDK:

```
$ sudo apt-get install openjdk-7-jdk  
  
# check  
$ sudo dpkg --get-selections | grep jdkapt-cache search jdk  
$ java -version
```

2) Tomcat 7

Get the software:

```
$ sudo aptitude update  
  
$ sudo aptitude install tomcat7  
$ sudo aptitude install tomcat7-admin  
$ sudo aptitude install tomcat7-docs
```

Add new user tomcat7:

```
$ sudo adduser tomcat7 -G tomcat  
$ sudo passwd tomcat7  
  
$ sudo chgrp -R tomcat7 /etc/tomcat7  
$ sudo chmod -R g+w /etc/tomcat7  
$ sudo chown -R tomcat7:tomcat7 /etc/tomcat7
```

Add md5 to server.xml file (in /etc/tomcat7 replace Line ~124):

```
<Realm className="org.apache.catalina.realm.UserDatabaseRealm" resourceName="UserDatabase" digest="md5" />
```

Create md5 password:

This allows you to secure the Solr Admin GUI. The example below uses an md5sum of 'solr', but you should choose something a bit more secure. Add the following to tomcat-users.xml in /etc/tomcat7, replacing the username and password with your own.

```
<tomcat-users>
  <role rolename="manager-gui"/>
  <role rolename="manager-script"/>
  <role rolename="manager"/>
  <role rolename="admin-gui"/>
  <role rolename="solrAdmin"/>
  <user username="solr" password="9f79967efee8ad705016101de169736f" roles="manager-gui, manager, admin-gui,
solrAdmin"/>
</tomcat-users>
```

Create a new file in tomcat bin directory called setenv.sh:

```
$ sudo vi /usr/share/tomcat7/bin/setenv.sh
```

Add the following:

```
JAVA_OPTS="-d64 -Xms26G -Xmx26G -XX:+UseConcMarkSweepGC -XX:+UseParNewGC -
XX:+CMSIncrementalMode -XX:-UseLoopPredicate -XX:MaxPermSize=256M -server"
```

You may need to tweak these settings, but we have found that the memory settings and garbage collection settings perform well.

Note - you may also need to define JAVA_HOME in if Tomcat does not run. As an example, when installing the OpenJDK the standard path to be added to the setenv.sh file is:

```
export JAVA_HOME=/usr/lib/jvm/java-1.7.0-openjdk-amd64
export PATH=${PATH}:${JAVA_HOME}/bin
```

Start/ Stop Tomcat:

```
$ source /usr/share/tomcat7/bin/setenv.sh
$ sudo /etc/init.d/tomcat7 stop
$ sudo /etc/init.d/tomcat7 start
```

Log file:

```
$ sudo tail -f /var/log/tomcat7/catalina.out
```

Docbase:

```
docBase="/usr/share/tomcat7-admin/manager"
```

View Tomcat in the browser:

http://your-server-instance:8080/

3) Solr

Configuration files:

Download the solr config tar ball (*ensembl-solr.tar.gz*) from the Ensembl FTP Site and unzip it in the `/www/` directory

```
$ sudo mkdir /www
# download
$ sudo tar -xvf ensembl-solr.tar.gz
```

Indexes:

The indexes are provided on the Ensembl FTP site as one tar ball for each shard. Copy them to the index directory (`/www/www-live/indexes/ensembl`), unzip and set permissions. Please note that this could take some time as the indexes are currently in the order of 750Gb.

```
$ sudo mkdir -p /www/www-live/indexes/ensemble
$ cd /www/www-live/indexes/ensembl
# download
# for the tarball of each shard you can compare it's md5sum with that in the md5sum download file,
for example
$ md5sum ensembl_core_74.tar.gz
$ grep core indexes.md5
# extract
$ sudo tar -xvf ensembl-solr.tar.gz
$ sudo chown -R tomcat7:tomcat7 /www/
```

You should see a structure similar to the following

```
ubuntu@domU-12-31-39-06-2A-0D:/www/www-live/indexes/ensembl$ ls -ld *74
drwxr-xr-x 4 tomcat7 tomcat7 4096 Dec  2 18:26 ensembl_core_74
drwxr-xr-x 4 tomcat7 tomcat7 4096 Nov 19 16:06 ensembl_docs_74
drwxr-xr-x 4 tomcat7 tomcat7 4096 Dec  2 18:26 ensembl_genomic_alignment_74
drwxr-xr-x 5 tomcat7 tomcat7 4096 Dec  2 18:23 ensembl_variation_74
```

Update July 2018 – since the above screenshot was prepared we have more shards and they are no longer versioned in the configuration:

```
ensembl_core
ensembl_docs
ensembl_ga
ensembl_variation_core
ensembl_variation_human_1
ensembl_variation_human_2
ensembl_variation_human_3
ensembl_variation_mouse
ensembl_variation_other
```

The locations of the indices are defined by the 'data.dir' in the solr.properties file for each shard. For example the downloaded solrcore.properties file in /www/solr/sanger/ensembl_core/conf/ contains:

```
MASTER_CORE_URL=http://localhost:8080/solr-sanger/
SHARD_CORE_URL=localhost:8080/solr-sanger/
data.dir=/www/www-live/indexes/ensembl/ensembl_core
enable.master=false
enable.slave=true
REPLICATION_INTERVAL=00:00:00
```

Dictionary files:

dict.txt, dict2.txt and dict3.txt are used for auto completion and autosuggestion. Since they are only needed in the ensembl_core configuration directory, the ensembl-solr.tar.gz tarball only contains them in this location. This does lead to warning in the server logs; if you want to prevent these copy the dict* files to each shard, for example:

```
$ cd /www/solr/sanger/ensembl_core/conf/
$ cp -p dict*.txt ../../ensembl_ga/conf
```

Deploying Solr to Tomcat:

solr-sanger.xml file is the context file for deploying to Tomcat. It points to the secure Solr war file in the same directory, and sets the solr /home parameter to /www/solr/sanger.

```
$ cd /www/www-live/war/  
$ cat solr-sanger.xml  
<Context path="/solr-sanger" docBase="/www/www-live/war/apache-solr-3.6.1-secure-admin.war" debug="0"  
crossContext="true">  
  <Environment name="solr/home" type="java.lang.String" value="/www/solr/sanger" override="true"/>  
</Context>
```

Copy the solr-sanger.xml file to the tomcat directory where it gets picked up by the server.



```
$ cd /www/www-live/war/  
$ sudo cp solr-sanger.xml /etc/tomcat7/Catalina/localhost/
```

If you monitor the catalina.out file, you should see some activity representing the deployment of Solr:

```
$ sudo tail -f /var/log/tomcat7/catalina.out
```

If all is well you should be able to view tomcat in the browser at <http://your-server-instance:8080/manager/html>.

Note - the username and password are the ones that you set up earlier.



Tomcat Web Application Manager

Message: OK

Manager

List Applications	HTML Manager Help	Manager Help	Server Status
-----------------------------------	-----------------------------------	------------------------------	-------------------------------

Applications

Path	Version	Display Name	Running	Sessions	Commands
/	None specified		true	0	Start <input type="button" value="Stop"/> <input type="button" value="Reload"/> <input type="button" value="Undeploy"/> <input type="button" value="Expire sessions"/> with idle ≥ <input type="text" value="30"/> minutes
/host-manager	None specified	Tomcat Host Manager Application	true	0	Start <input type="button" value="Stop"/> <input type="button" value="Reload"/> <input type="button" value="Undeploy"/> <input type="button" value="Expire sessions"/> with idle ≥ <input type="text" value="30"/> minutes
/manager	None specified	Tomcat Manager Application	true	1	Start <input type="button" value="Stop"/> <input type="button" value="Reload"/> <input type="button" value="Undeploy"/> <input type="button" value="Expire sessions"/> with idle ≥ <input type="text" value="30"/> minutes
/solr-sanger	None specified		true	0	Start <input type="button" value="Stop"/> <input type="button" value="Reload"/> <input type="button" value="Undeploy"/> <input type="button" value="Expire sessions"/> with idle ≥ <input type="text" value="30"/> minutes

Click on the /solr-sanger path to view the Solr Admin front page
(<http://your-server-instance:8080/manager/html>)

Welcome to Solr!

[Admin ensembl core](#)

[Admin ensembl variation other](#)

[Admin ensembl variation human 1](#)

[Admin ensembl variation human 2](#)

[Admin ensembl variation human 3](#)

[Admin ensembl variation cow](#)

[Admin ensembl variation mouse](#)

[Admin ensembl ga](#)


[Admin ensembl docs](#)



Click on one of the shards to access the admin screen for that one:

Solr Admin (ensembl)

ip-10-35-159-61.eu-west-1.compute.internal:8080
cwd=/var/lib/tomcat7 SolrHome=/www/solr/sanger/ensembl_core/
HTTP caching is OFF



Solr	[SCHEMA] [CONFIG] [ANALYSIS] [SCHEMA BROWSER] [REPLICATION] [STATISTICS] [INFO] [DISTRIBUTION] [PING] [LOGGING]
Cores:	[ensembl_core] [ENSEMBL_VARIATION_OTHER] [ENSEMBL_VARIATION_HUMAN] [ENSEMBL_VARIATION_MOUSE] [ENSEMBL_GA] [ENSEMBL_DOCS]
App server:	[JAVA PROPERTIES] [THREAD DUMP]
Make a Query	[FULL INTERFACE]
Query String:	<input type="text" value="**"/> <input type="button" value="Search"/>
Assistance	[DOCUMENTATION] [ISSUE TRACKER] [SEND EMAIL] [SOLR QUERY SYNTAX]
	Current Time: Fri Nov 22 10:48:46 UTC 2013
	Server Start At: Fri Nov 22 08:35:22 UTC 2013

See the Solr site for more details - <http://lucene.apache.org/solr/>

Replication

The configuration of Solr for Sanger is that of a master slave scenario, where we generate the indexes on the master server and then replicate these to the slaves used for production. You can remove replication if it is not needed by editing the solrconfig.xml files in each conf directory:

```
$ cd /www/solr/sanger/ensembl_core/conf

# edit solrconfig.xml to remove this whole section (~ line 1084)
$vi solrconfig.xml
<requestHandler name="/replication" class="solr.ReplicationHandler" >
  <lst name="master">
    <str name="enable">${enable.master}</str>
    <str name="replicateAfter">commit</str>
    <str name="replicateAfter">startup</str>
    <str name="confFiles">schema.xml,solrconfig.xml,stopwords.txt,stopwords_en.txt,dict.txt,dict2.txt</str>
  </lst>
  <lst name="slave">
    <str name="enable">${enable.slave}</str>
    <str name="masterUri">${MASTER_CORE_URL}${solr.core.name}/replication</str>
    <str name="pollInterval">${REPLICATION_INTERVAL}</str>
  </lst>
</requestHandler>
```

Test

Use something like the following URL to check that you are returning the results of search queries:

```
http://your-server-instance:8080/solr-sanger/ensembl_core/ensemblshards?indent=on&version=2.2&
q=%3A*&fq=&start=0&rows=10
```

Note - if you use the admin interface to run queries take care to change the Request Handler from '/select' to either '/ensembl' or '/ensemblshards'.

4) Web client

The code for the web client is supplied in public-plugins/solr which can be obtained via a CVS checkout of the ensembl code. Configuring a mirror web site to use a solrserver installed as described above involves two steps:

(i) Enable the solr plugin in your Plugin.pm:

```
$ cd serverroot/conf

#edit Plugins.pm to include this line:
'EnsEMBL::Solr' => $SiteDefs::ENSEMBL_SERVERROOT.'/public-plugins/solr',
```

(ii) Configure your mirror site to use your solrserver. This is done by editing the SiteDef.pm in your last loaded plugin, probably the one at the top of Plugins.pm

```
$ cd serverroot/my_plugin/conf

# edit SiteDefs.pm to include something like this:
$SiteDefs::ENSEMBL_SOLR_ENDPOINT = "http://your-server-instance:8080/solr-sanger/ensembl_core/ensemblshards";
```