HULBEE ENTERPRISE SEARCH

Installation and setup manual

Modified: 17-Dec-2015

Version: 1.7.15

Table of Contents

1 Architecture

1.1 Hulbee Enterprise Search

Hulbee Enterprise Search (HES)[1] is designed to work with MS Active Directory users. The current version is intended to conduct a search in file storage, taking into account users' file permissions. Users should at least have a file read permission to see files in the search results.

Supported file formats[2]:

File type	Extensions	Text extraction	Meta tags extraction	Attached files extraction
Text	txt, rtf, doc/dot, odt, wri, sxw	✔		
	docx/docm/dotx	✔	✔	
Publication	pdf	✔	✔	
	xps	✔		
Hypertext	html, htm, xml	✔		
	mht, shtml	✔	✔	
Tables	xsl, xslt, xls, ods, csv	✔		
	xlsx	✔	✔	
Presentations	pptx	✔	✔	
	ppt, pps, odp	✔		
Graphics	bmp, jpg/jpeg, png, jfif, tif, tiff, jpe		✔	
E-mail	msg, eml	✔	✔	✔
Archives	zip, rar, 7zip			✔
Media	avi, mp3, mp4, wav, m4a, wma, wmv,ogg, flac, mkv, ape, mpc		✔	
Source Code and Scripting	cs, vb, js, csproj, h, c, cpp, vbs,vcproj, vbproj, pl, sql, bat, cmd	✔		
	css	✔	✔	

Any modern browser (Mozilla Firefox, Chrome, etc.) with the opened link to the search engine in intranet can be used as a user interface. The search and user settings modification are performed here. Using admin panel, administrators can also adjust various settings.

To open search-found files a user needs to run a computer under his/her own real account in one local network with file storage. HES Desktop Manager utility, which opens files, should be installed on the client

---

[1] Some of the modules have Swisscows or SES in the filename or texts. It is the old name of the project and is synonymous for Hulbee Enterprise Search or HES.

[2] The possibility of the search system not only to index file metadata (file name, path, size, creation date, modification date), but also to work with its content – to extract the text and/or meta tags, and/or unpack files, containing other files.

computer. A user needs to possess the necessary rights for software installation, or to turn to the administrator for help.

HES software complex consists of two main parts:

- Indexstore (Linux server with installed and configured Elasticsearch).
- Processing server (Windows server with all other components).

In case of small filestores, these servers might be combined into the single Windows Server Machine. Large filestores may require mounting the Elasticsearch cluster containing a number of servers.

1.2   Swisscows Company Search

The present configuration is the appliance server, containing all necessary components of Hulbee Enterprise Search. Taking into consideration the fact that it contains the full-featured Windows Server 2012 R2 Standard, it can be used not only for search, but also for deployment of services based on MS Active Directory[3] and storage of data in medium-sized enterprises[4], which still use peer networks.

Swisscows Company Search contains versions notable both for Hardware capabilities and for limitation of number of users.

2 Initial Installation

2.1   Full installation

2.1.1   Indexstore

Elasticsearch v.1.7 is used as an indexstore. Install it, using directions from the manufacturer website [See chapter 5]. Elasticsearch could be installed both on the computer, containing Processing Server components, and on individual computer. If you install Elasticsearch on individual computer you may use operating system different from Windows, but supporting JRE (GNU\Linux, Solaris, etc).

In case of extremely high loads, it's recommended to use a cluster containing a number of Elasticsearch servers.

After installation in accordance with Elastic website recommendations, it's necessary to make post-installation settings in Elasticsearch configuration (elasticsearch.yml file). Add the following lines at the end of the file:

```
script.inline: on
script.indexed: on
```

Notice! Index may contain confidential data. To prevent leaks of such data please disallow all TCP/IP connections for all of the components except Application Server.

---

[3] You may find links to introductory articles about Active Directory in chapter 5 - Useful links.
[4] Remember that you need to adjust and regularly back up your data. You may find the general information about data back-up in chapter 5 - Useful links

2.1.2 Processing server

2.1.2.1 Hardware requirements for HES Processing Server

Component	Minimum	Recommended
Processor Cores	4	>=8
Memory	16 GB	64 GB
Hard disks and available storage space	256 GB	512 GB
Network adapter speed (to filestorage and indexstorage)	1 Gb/s	>=10 Gb/s

2.1.2.2 Software requirements for HES Processing Server

Install Windows Server 2012 R2 Standard with the latest updates and the following components:

NetFx4ServerFeatures	IIS-RequestFiltering	IIS-ISAPIExtensions	IIS-WebServerManagementTools
NetFx4	IIS-StaticContent	IIS-ISAPIFilter	
NetFx4Extended-ASPNET45	IIS-DefaultDocument	IIS-ASPNET45	IIS-ManagementConsole
IIS-WebServerRole	IIS-DirectoryBrowsing	IIS-HealthAndDiagnostics	WCF-Services45
IIS-WebServer	IIS-HttpErrors	IIS-HttpLogging	WCF-TCP-PortSharing45
IIS-CommonHttpFeatures	IIS-ApplicationDevelopment	IIS-Performance	
IIS-Security	IIS-NetFxExtensibility45	IIS-HttpCompressionStatic	

Before installation you can execute the following command:

```
>Dism /Online /Enable-Feature /FeatureName:NetFx4ServerFeatures
    /FeatureName:NetFx4 /FeatureName:NetFx4Extended-ASPNET45
    /FeatureName:IIS-WebServerRole /FeatureName:IIS-WebServer
    /FeatureName:IIS-CommonHttpFeatures /FeatureName:IIS-Security
    /FeatureName:IIS-RequestFiltering /FeatureName:IIS-StaticContent
    /FeatureName:IIS-DefaultDocument /FeatureName:IIS-DirectoryBrowsing
    /FeatureName:IIS-HttpErrors /FeatureName:IIS-ApplicationDevelopment
    /FeatureName:IIS-NetFxExtensibility45 /FeatureName:IIS-ISAPIExtensions
    /FeatureName:IIS-ISAPIFilter /FeatureName:IIS-ASPNET45 /FeatureName:IIS-
    HealthAndDiagnostics /FeatureName:IIS-HttpLogging /FeatureName:IIS-
    Performance /FeatureName:IIS-HttpCompressionStatic /FeatureName:IIS-
    WebServerManagementTools /FeatureName:IIS-ManagementConsole
    /FeatureName:WCF-Services45 /FeatureName:WCF-TCP-PortSharing45 /All
```

You may also do it, using system applet "Turn Windows features on or off".

2.1.2.3   Domain and Users setup

1. Join server into your AD.
2. Create a user **hes_user**.
3. Grant permissions for **reading** and **list folder contents** in the filestorage for **hes_user**.
4. Add hes_user to the local administrator group in the HES processing server.

2.1.2.4   IFilters setup

Install the following IFilters to get more exact text extraction from MS Office and PDF documents:

- MS Office: http://www.microsoft.com/en-US/download/details.aspx?id=17062 with service pack http://support.microsoft.com/kb/2687447. Install 64-bit versions.
- PDF iFilter 64 11.0.01: http://www.adobe.com/support/downloads/detail.jsp?ftpID=5542.

2.1.2.5   HES unpacking

HES applications pack has a name like **HES.1.7.XX.XXXXX.zip (XX being the numbers of your specific version)**. It contains the following components:

- Helpers
  - ConfigTransformationHelper
- Services
  - Ses.Services.Crawler
  - Ses.Services.IndexCleaner
- Utilities
- Web
  - Ses.Web
- Ses.Setup.*

Unpack it to the some folder, for example C:\HES (i.e. C:\HES\Web\, C:\HES\Services\, C:\HES\Utilities\, etc).

2.1.2.6   Initial setup

Run the ses.setup.exe utility in the root folder of unpacked distributive.
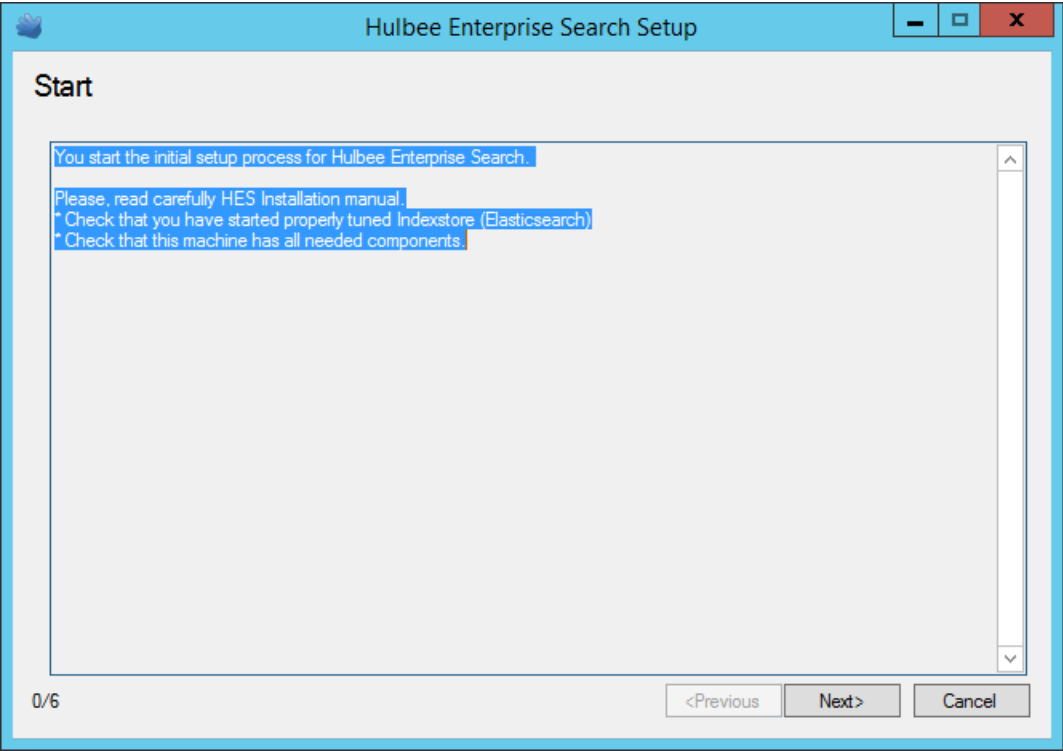
Fig. 1. HES Setup – start of Setup.

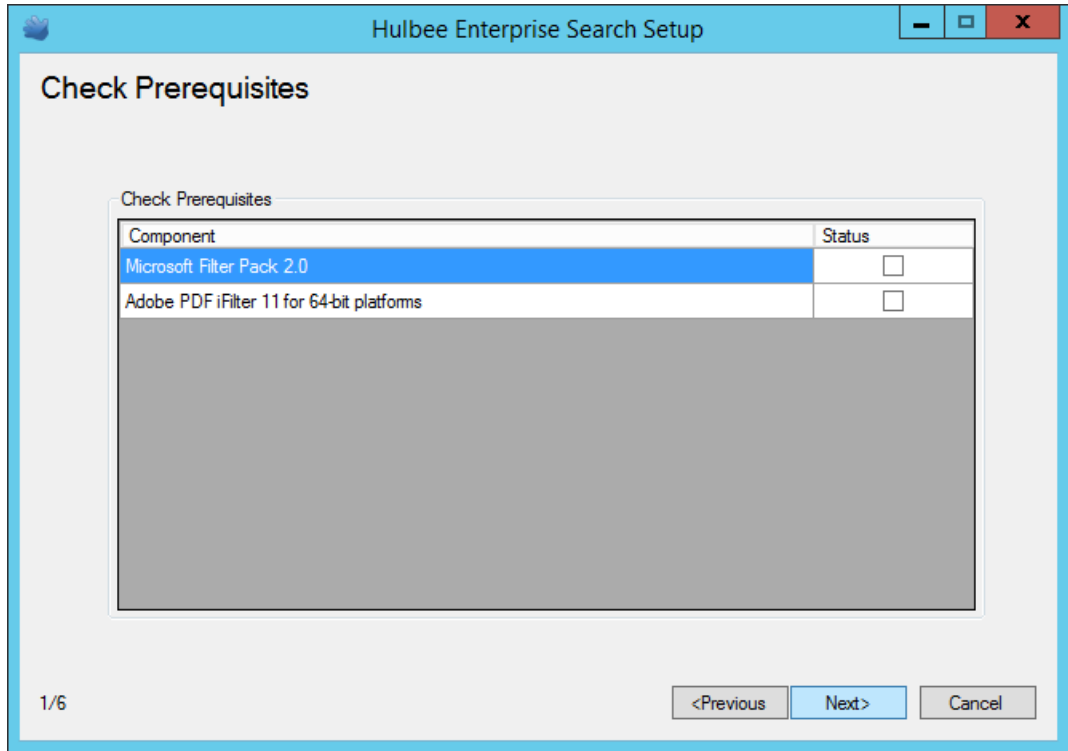Click the "Next" button to take step 1.

1. Step 1.

Fig. 2. Step 1 of HES Setup.

At the stage of optional prerequisites check, missing components should be specified. The setup can be continued, but the lack of IFilter could have an impact on the quality of extraction of the text from some formats (MS Office and PDF).

2. Step 2.

Enter the valid url to the Elasticsearch server and the index name (default value is "hes"). URL should have http protocol prefix, proper IP or domain name and port.
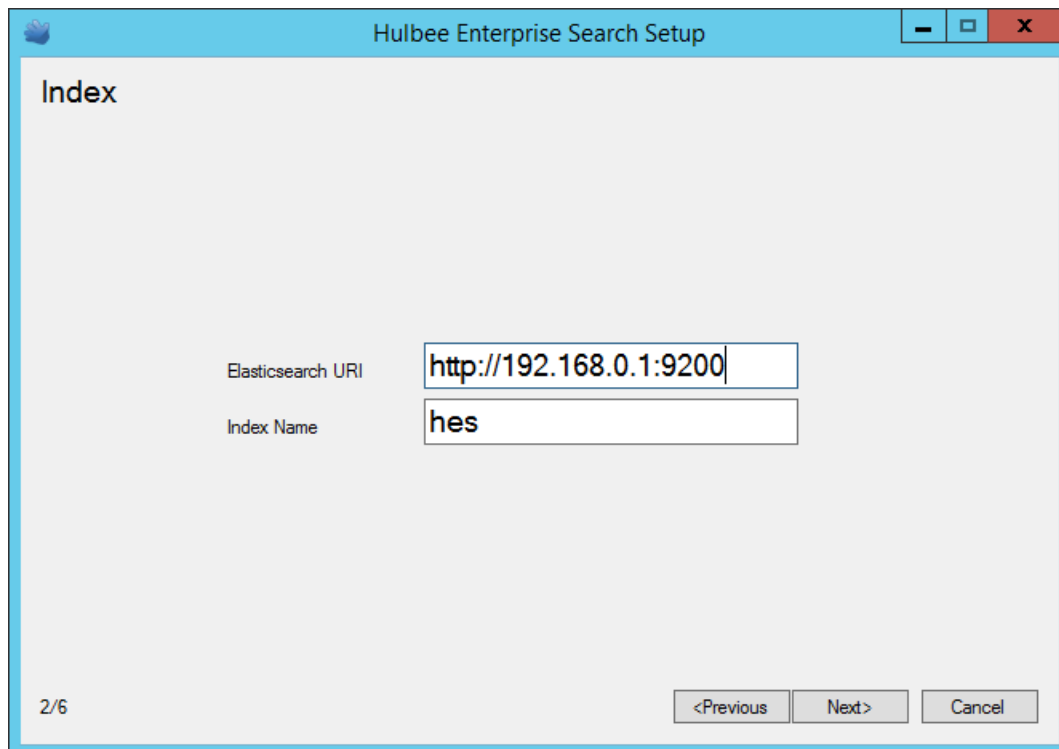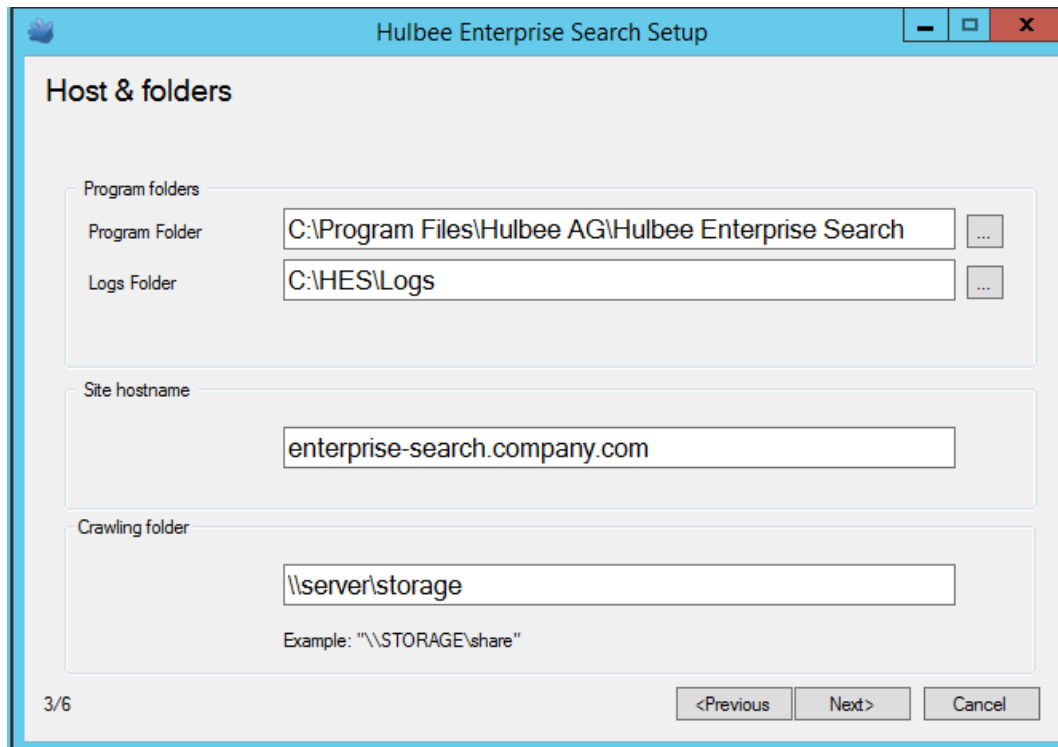
Fig. 3. Step 2 of HES Setup.

3. Step 3.

Fig. 4. Step 3 of HES Setup.

In most cases, you only need to change "**Site hostname**" and "**Crawling folder**". "**Program folder**" and "**Logs folder**" can be left with default values. Use UNC format for "**Crawling folder**" name.

4. Step 4.
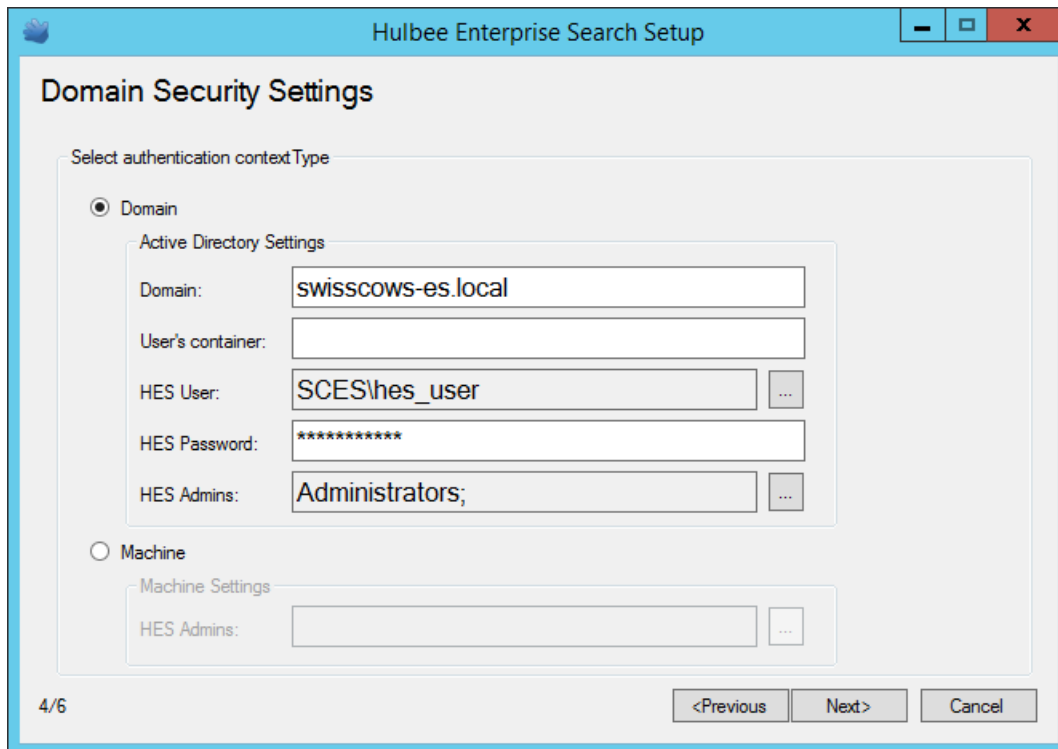
Fig. 5. Step 4 of HES Setup.

"**Domain**", "**HES User**" / "**Password**" fields correspond to the user "**hes_user**" which should be created in step 2.1.2.3 of the current manual.

User's container field can be left as it is in the most cases.

To select User or Admin group, use the button to the right of the entry field, which opens the system dialog.

Machine option is intended for test installations and special configurations, not reviewed in the present manual.

5. Step 5.
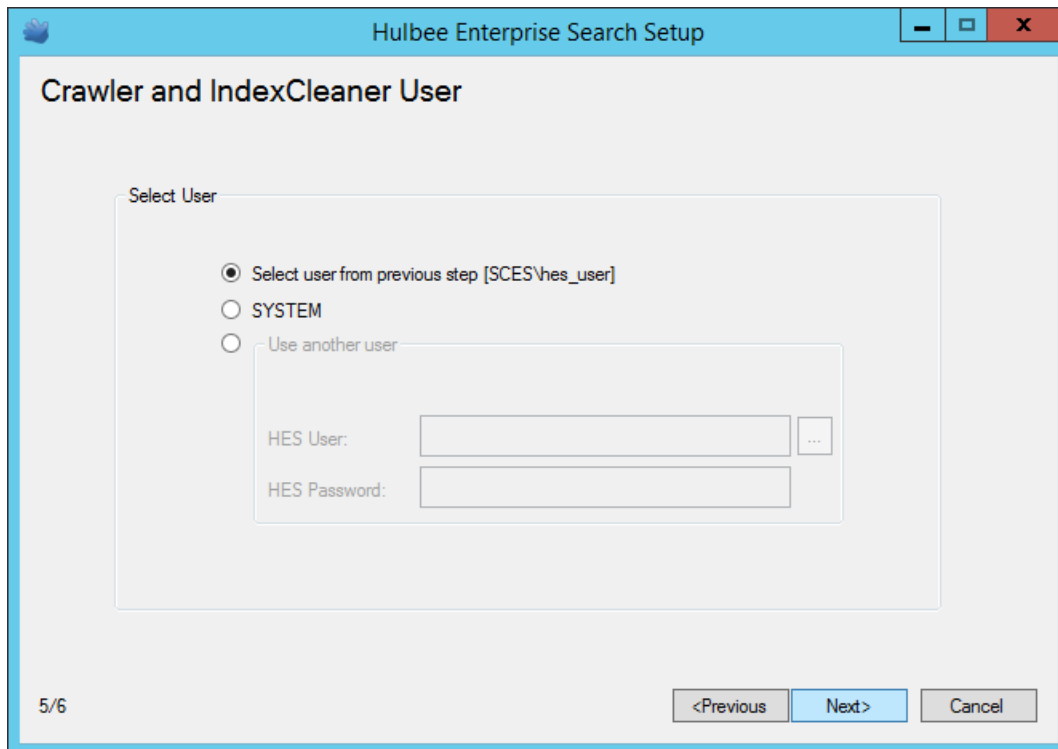
Fig. 6. Step 5 of HES Setup.

Use the "**hes_user**" for the configuration with separated filestore. Do not forget to add it in the Processing server before.

6. Step 6.
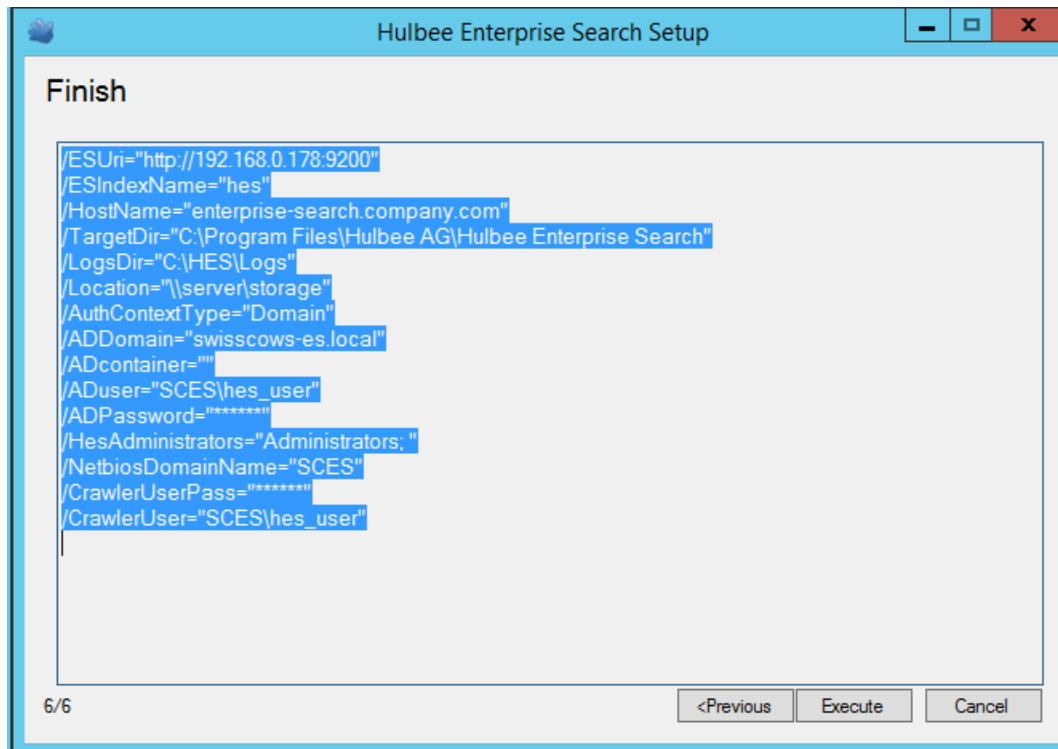
Fig. 7. Step 5 of HES Setup.

You can see the selected options and run setup with "Execute" button.
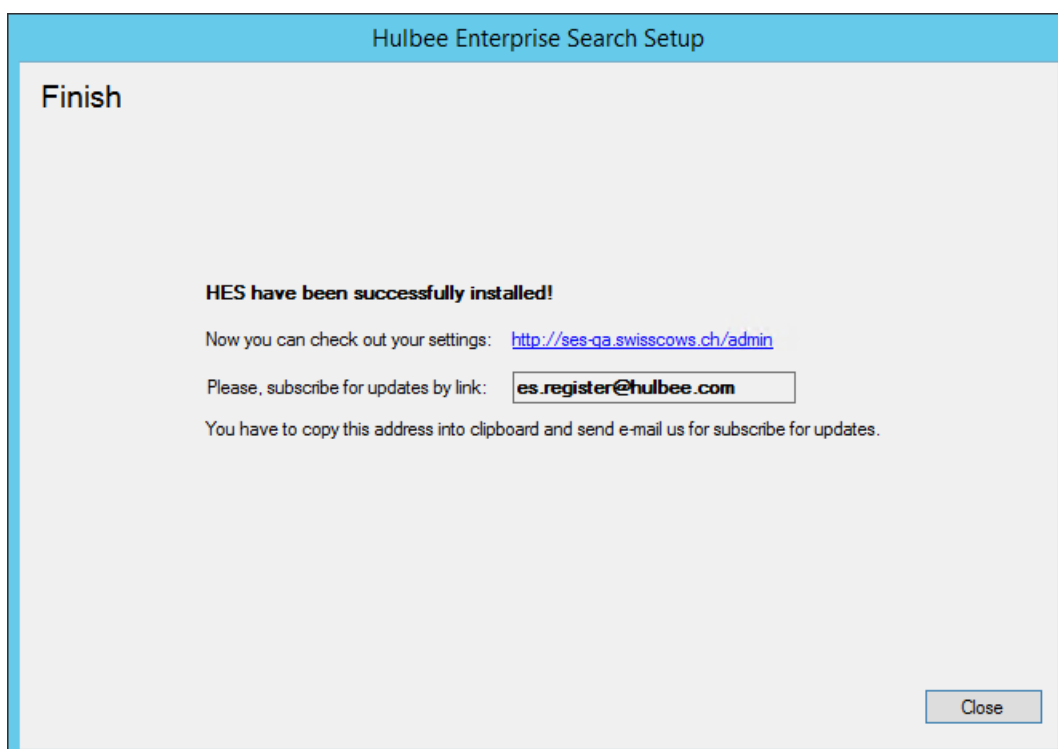
7. Final step.

Fig. 8. HES Setup – Setup finish.

After successful setup the final window with the link to the admin panel and mail address appears. You may subscribe for updates, sending a letter with the title "subscribe" to the above mentioned address.

2.2 Swisscows Company Search

This configuration comes with all pre-installed components, necessary for installation and work of HES. So, you need to take the following steps.

2.2.1 Domain setting up

First of all, plan the appliance server usage scenario. Depending on it, the following standard scenarios are available:

- Joining the Swisscows Company Search Server to an existing domain.
- Creation of a new domain based on Microsoft Active Directory in Windows Server 2012 R2.

Clicking the links in chapter 5 "Useful links" you will find a useful background information on the work with Active Directory.

In the process of distribution, administrator may sign in to the server, using pre-installed Administrator account with password "Admin123".

Notice! You must change administrator's password.

When the server will operate in the domain, add users to the domain, as indicated in chapter 2.1.2.3 Domain and Users setup.

2.2.2 File storage

The typical size of search index is up to 10% of binary document size (approximately, as it essentially depends on real database set). So, the appliance server may also be used as network file storage.

For example, if you suppose that within several years the file volume won't exceed 1 TB, and your configuration contains 2 TB of disk space, it's quite safe to organize not only appliance server based search, but also file storage. This scenario will also decrease network load during intensive indexing.

If you suppose that volumes will be higher, you need to store files on server with sufficient storage capacity disk array, or to set additional disks in Appliance Server.

Notice! The file storage should be available at local network and should use the same Active Directory as other HES parts (hes_user).
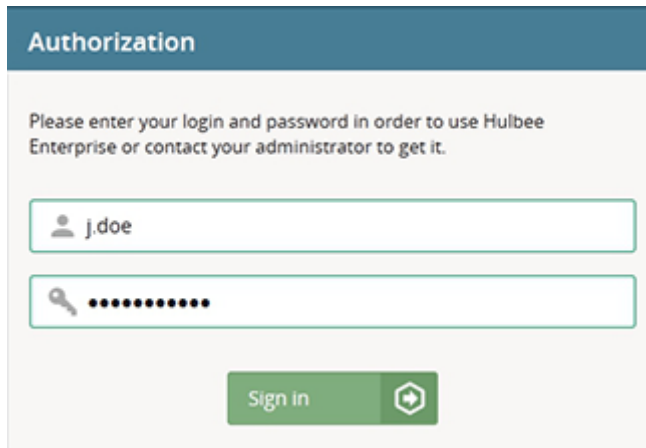
2.2.3 Initial setup

After you've planned the Appliance Server usage scenario in the field of Active Directory use and file storage location scenario, you need to follow the setup steps in accordance with 2.1.2.6 Initial setup.

3 Checking installation

To start searching your documents, enter the URL to the WEB-application. This URL was adjusted under 2.1.2.6, Step 3:

http://**enterprise-search.company.com/**

If client's account is not recognized automatically, please enter login (including NetBIOS domain name) and password into the authorization form:

Fig. 9. Authorization form.

After the login the start page will appear:

14/26
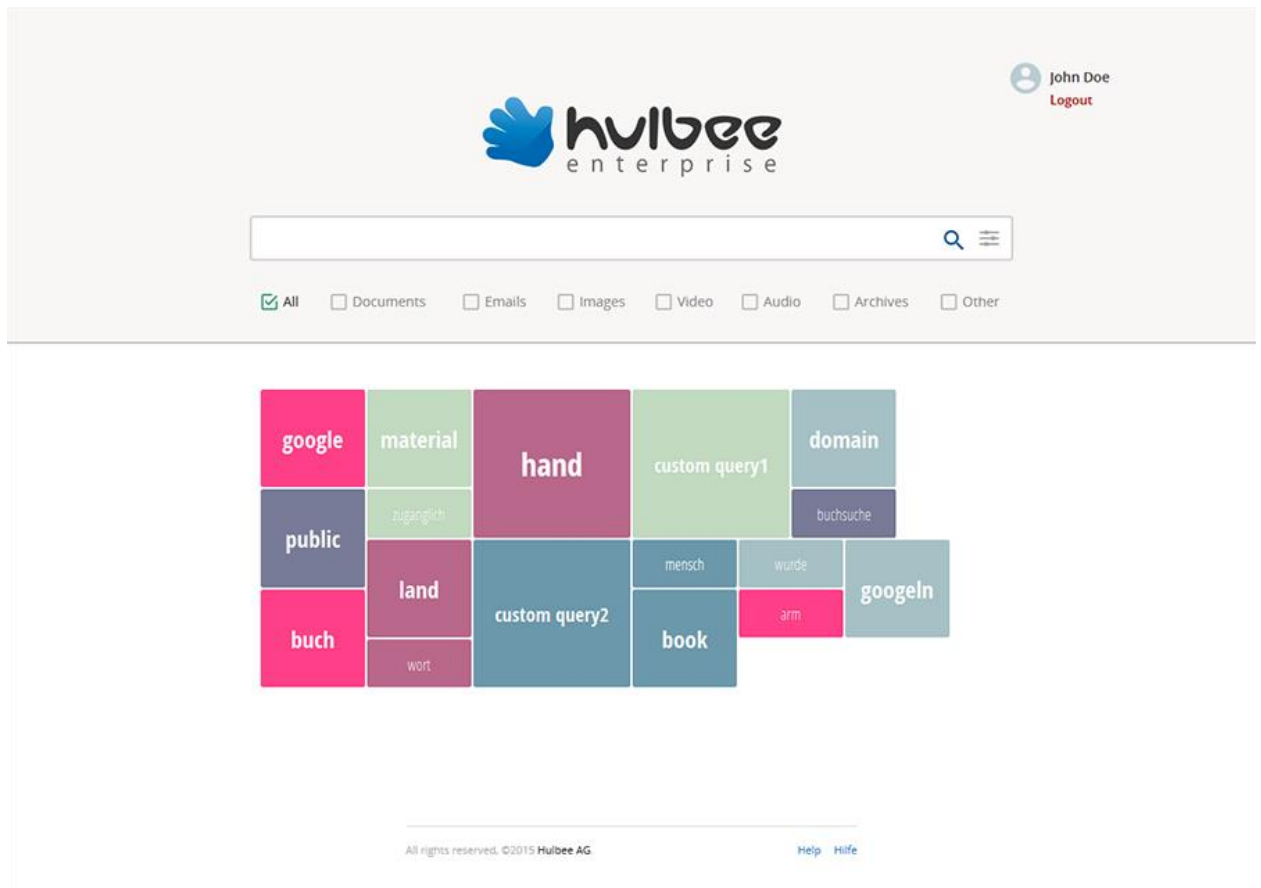
Fig. 10. HES start page.

Search query should be entered here in order to get documents with the relevant information.

4 Admin page setup

The remaining setup steps can be done using admin panel. Please follow the next link:

http://**enterprise-search.company.com**/admin

Instead of enterprise-search.company.com enter the link, used at step 3 of setup and depicted on Fig. 4.

4.1   Restore options

If you have already installed and tuned instance of HES, you can restore previously backed up settings. Select "Export/Import" area and:

1.   Click "Select file" button and select the back-up file.
2.   Click "Import" button.

This operation allows restoring global and personal users settings. In other words, visible data that can be changed with the help of admin panel and user's account. Documents search index is not included in settings – it will be updated with the help of Crawling service.

4.2   Administration Area

Pages of admin panel contain navigation area and workspace. Navigation area contains links. Following these links, you will be shifted to different subsections of admin panel. It's the same for all pages. The workspace contains elements for options edit and display of system operation information.
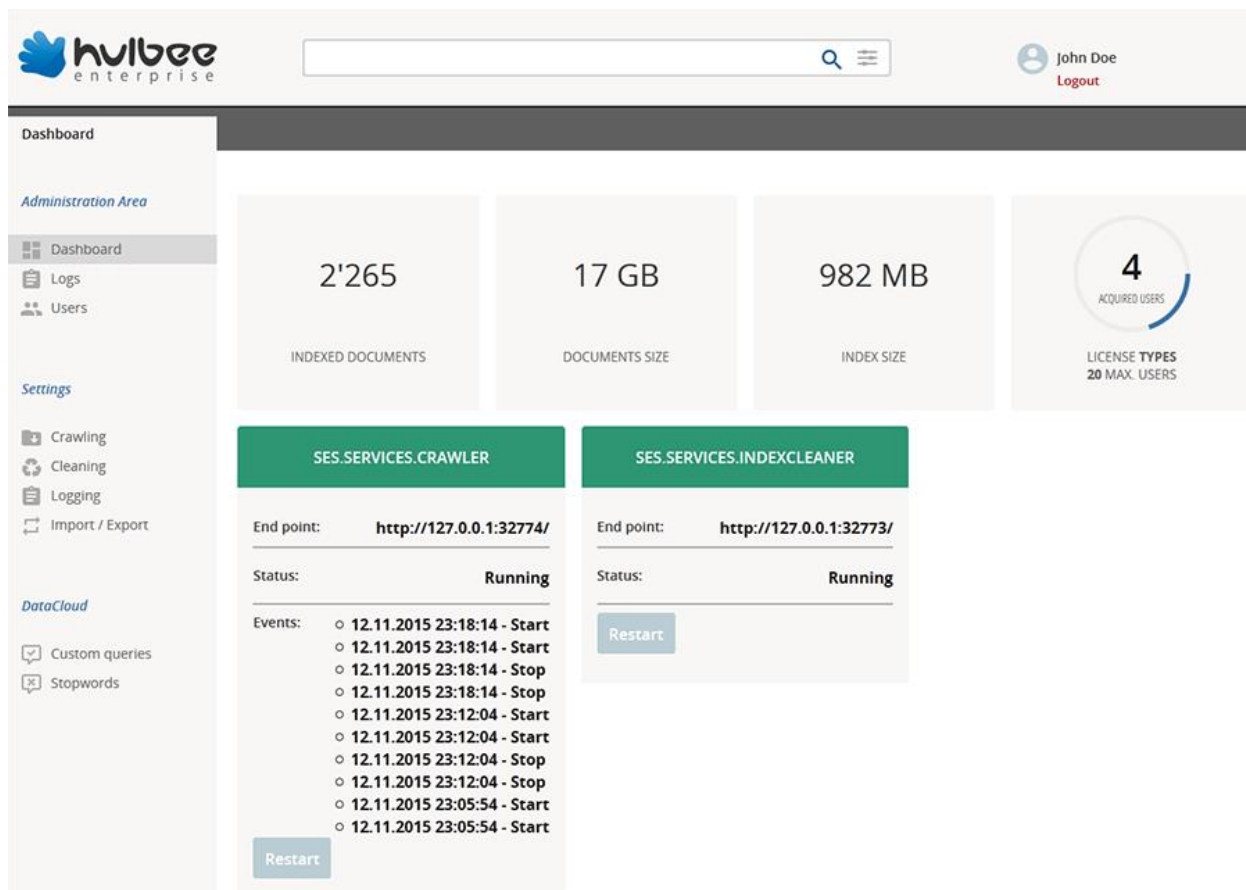

Fig. 11. Administration area. General view.

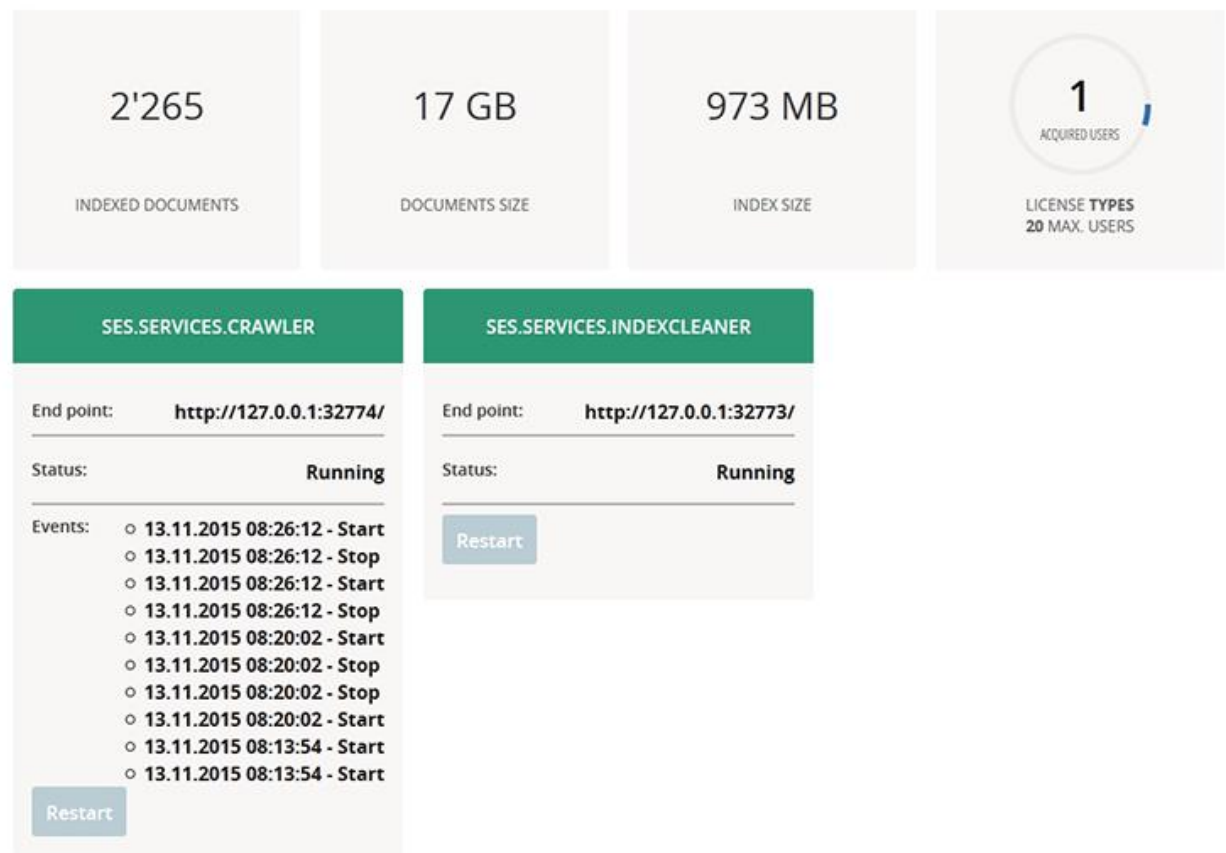Here are the screenshots with areas, which change when you follow different links of admin panel.

4.2.1 Dashboard

Fig. 12. Dashboard page.

You can see different kinds of statistics on this page:

- INDEXED DOCUMENTS. The number of indexed documents. It can differ from the total number of files in storage, as a part of documents doesn't need to be processed because of different filters (by size, by extension). At the same time, archives and mail messages may amount to more than one file.
- DOCUMENTS SIZE. The overall size of documents in index. The sum of their binary sizes is meant. It can differ from the space occupied in the file storage by the same reasons as for preceding item.
- INDEX SIZE. The size of Elasticsearch index on the disk.
- LICENSE {id}. This widget shows license type, the number of users who are already using the service and the number of users who potentially can begin using it.
- Service widgets. Provide an opportunity to check if crawling service and index cleanup are functioning. Here you also can restart service after parameters change.

4.2.2   Logs

Fig. 13. Logs page.

This page shows error messages, warning messages and other information. You may filter event type by ticking appropriate items in the drop-down list Events.
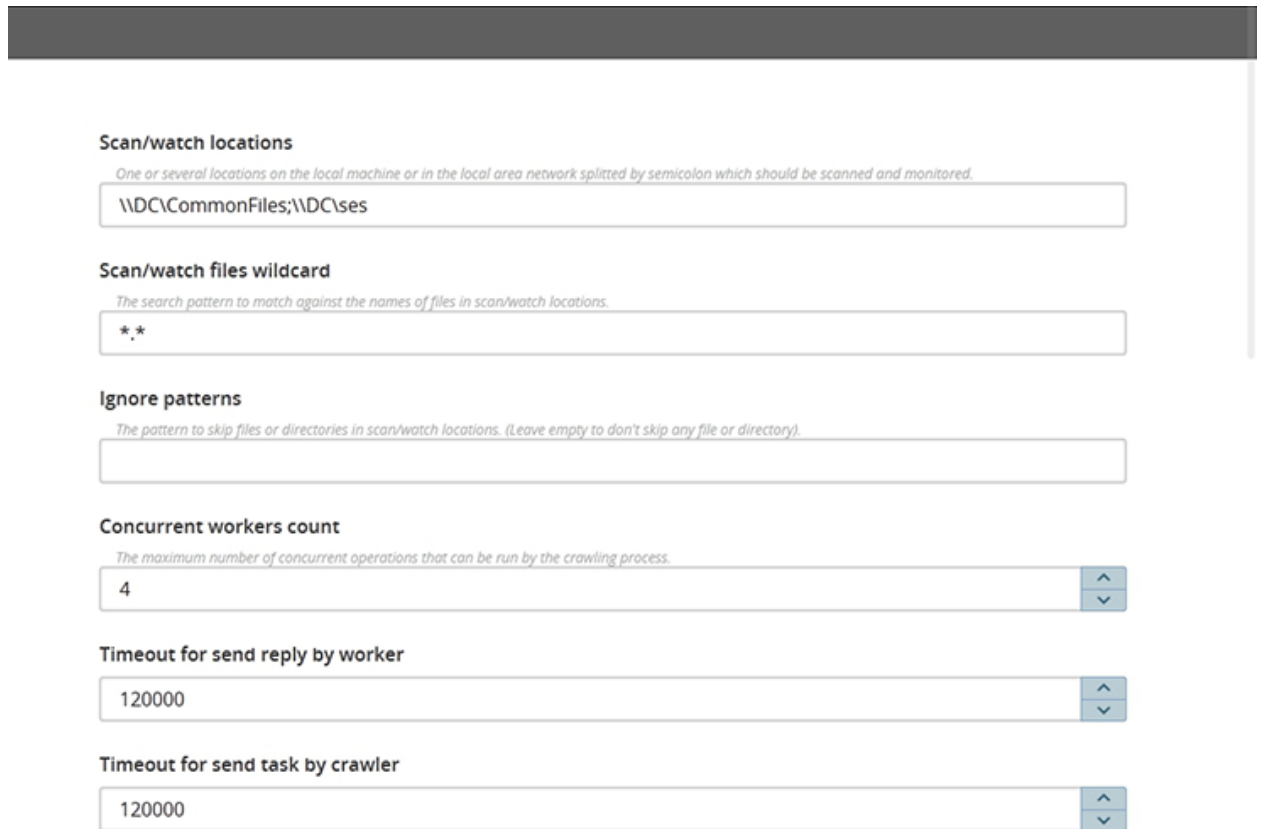
4.2.3   Users

Fig. 14. Users page.

Users page allows to see authorized users. Here it's also possible to forbid some users to conduct a search in the HES system. In this case, this user won't be counted during verification of the compliance of the number of users with the license conditions. If user-administrator is forbidden from searching, he still has access to the admin panel.

4.3   Settings

4.3.1   Crawling

Fig. 15. Crawling page.

Select "Crawling" area and tune following fields:

- "Scan/watch locations": Check and change if needed one or more locations on the local machine or in the local area network separated by semicolon that should be scanned and monitored. This path should be in the UNC format (for ex. "\\DC\Files").
- "Scan/watch files wildcard": pattern of processed documents. "*.*" is recommended. You may separate patterns by commas. Commonly used for test purposes. Refers only to the file name.
- "Ignore patterns": a pattern, allowing to ignore specific files or folders. Works with the full path. "*.log" will ignore everything ending in ".log". "*\fo*" will ignore all the files and folders beginning with "fo". "*" is the only one wildcard. You may choose several patterns, separated by commas.
- "Concurrent workers count": To get better processing speed, you can enter here a number up to the CPU cores number. For configuration with Elasticsearch, located on another server, it's recommended to assign the number of workers in accordance with the number of cores in the machine with Elasticsearch. But this number shouldn't be greater than the number of cores in the Processing Server. For configuration of Swisscows Company Search and in the cases when all modules are located on one machine, it's recommended to set the number of workers up to the half of the CPU cores number.
- "Timeout for send reply by worker": Time (msec), that a worker spends on a task (one document) processing.

- "Timeout for send reply by crawler": Time (msec), that a worker spends on waiting for a task. If there are no tasks, a worker shuts down to avoid wasting of resources.

Specify quite a little time for the first indexing cycle (about 15 sec or 15000 msec) to process all documents quickly. After the first crawl, you may increase the timeouts (for example, up to 120000 msec) to process heavy files (archives, large documents), which haven't been processed during the first cycle.

Fig. 16. Crawling page. Continuation.

- "Use custom Elasticsearch server": leave this option as it is (in case of typical configuration).
- "Maximum attachment size": maximum size of attached files. Having reached this size, archive or mail message processing stops.
- "Maximum extracted size": maximum size of text in symbols, which undergoes further processing and indexing. Text extracted from document is meant, and not binary file size. Remaining text is cut.
- "Maximum processed common limit": maximum size of file that can be converted. Larger files contain only metadata (name, path, creation and modification date, size).
- "Filters by extensions": selection of converter type depending on extension. As a rule, Ifilter is slower, but it's valid for office documents and pdf documents.
- "Maximum processed limits by extensions": some file types are large and fall within size limitation. But their processing is quite easy, as little part of data is processed. It mainly refers to media files "avi;wmv;mp4;mkv". For this purpose, you may set individual limitations for this group of files. You may enter them in one line, separating extensions by ";" symbol.

Fig. 17. Crawling page. Finish.

After changing these options, click "Save Settings" button. Crawler and IndexCleaner services will use new settings after restart. You may click "Restart" button on Dashboard page. If for some reason the service is not available on the admin panel (no widget), use system applet Services.

4.3.2   Cleaning

Fig. 18. Cleaning page.

Select "Cleaning" area and tune following fields:

- "Batch size": the number of documents to be processed per one iteration. Recommended value – 1000.
- "Iteration timeout": the timeout between iterations. In most cases, 10 seconds is a proper time. When we speak about processing of millions and tens of millions, we can decrease this time (it increases load on Processing server and Indexstore) to delete already deleted documents from index (or in the case of documents storage path change).

After changing these options, click "Save Settings" button. New settings will be used by IndexCleaner service after restart. You may click "Restart" button on Dashboard page. If for some reason the service is not available on the admin panel (no widget), use system applet Services.
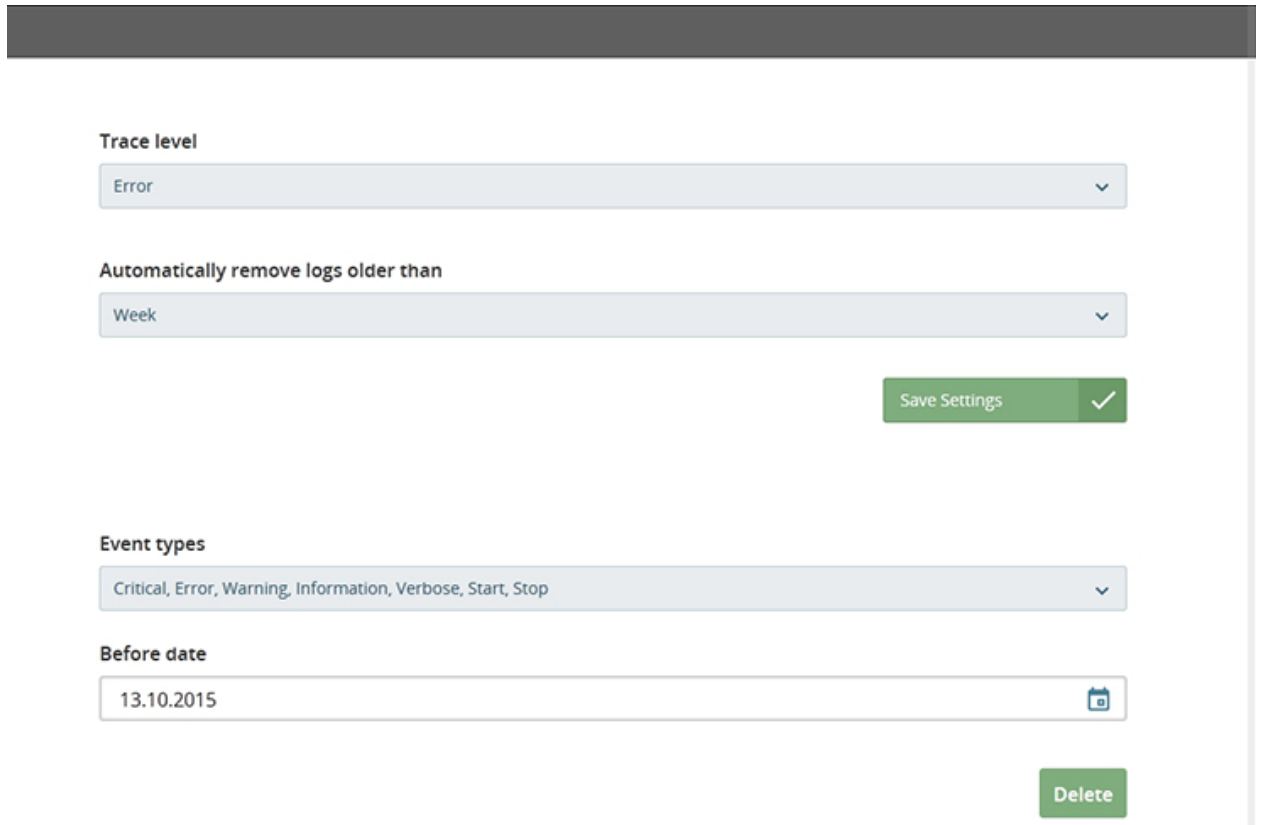
4.3.3   Logging

Trace level

Error

Automatically remove logs older than

Week

Save Settings

Event types

Critical, Error, Warning, Information, Verbose, Start, Stop

Before date

13.10.2015

Delete

Fig. 19. Logging page.

You can change the following parameters on the Logging page.

"Trace level": drop-down list, where we can indicate the events that should be recorded in the diagnostics. Crawler and IndexCleaner services will use new settings after restart.

"Automatically remove logs older than": this option allows you to indicate, how long HES logs should be stored. Both records stored in the IndexStore (they are shown in the admin panel) and text files on the machine with the Processing Server (typical place - c:\hes\logs) are removed. In the course of time they can occupy too much space, that's why it's not recommended to turn this option off (save in exceptional circumstances). IndexCleaner service removes logs. If you want to change a retention period, you need to restart this service.

You can also remove some events, kept in IndexStore, manually. In order to get that done you need to choose types of events that should be removed and date to which it's necessary to remove them. Then click "Delete" button.
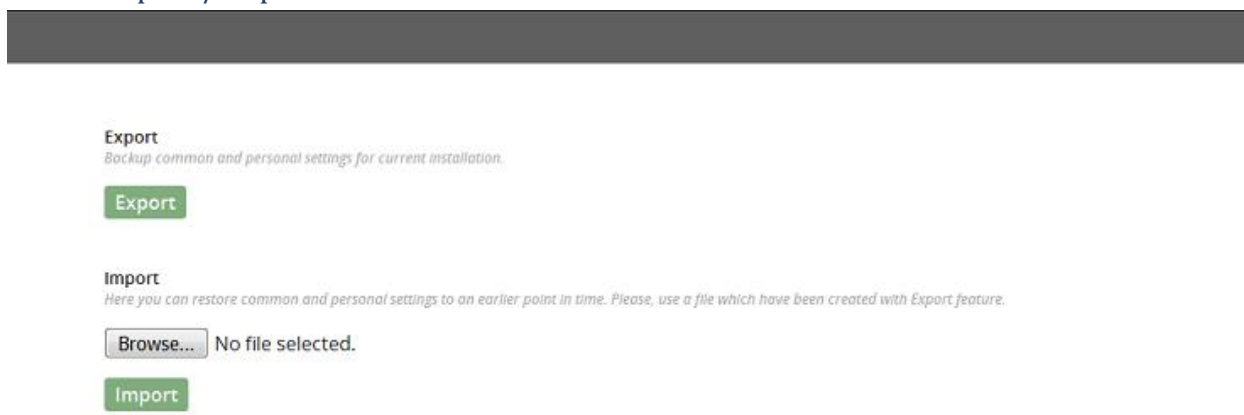
4.3.4 Import / Export

Fig. 20. Import / Export page.

This page contains commands, which provide an opportunity to save HES settings into a file and restore them when necessary. Both common and personal settings are saved. These commands will be useful in the case of system restore after failure, or after update installation.

4.4 Datacloud

These settings are similar to settings in the user account, but they concern all HES users (instead of one).

4.4.1 Custom queries

Fig. 21. Custom queries page.

"Custom queries" option allows to add keywords, later on reflected in DataCloud on home pages of all users. Keywords help to enter typical search queries quickly.

Using keyboard, you may type custom query in the line on the right. Then press "Enter" button and continue entering the next query, if necessary. You may add, edit and delete custom queries. There is also custom query navigation - page scrolling (if you have more than one page).
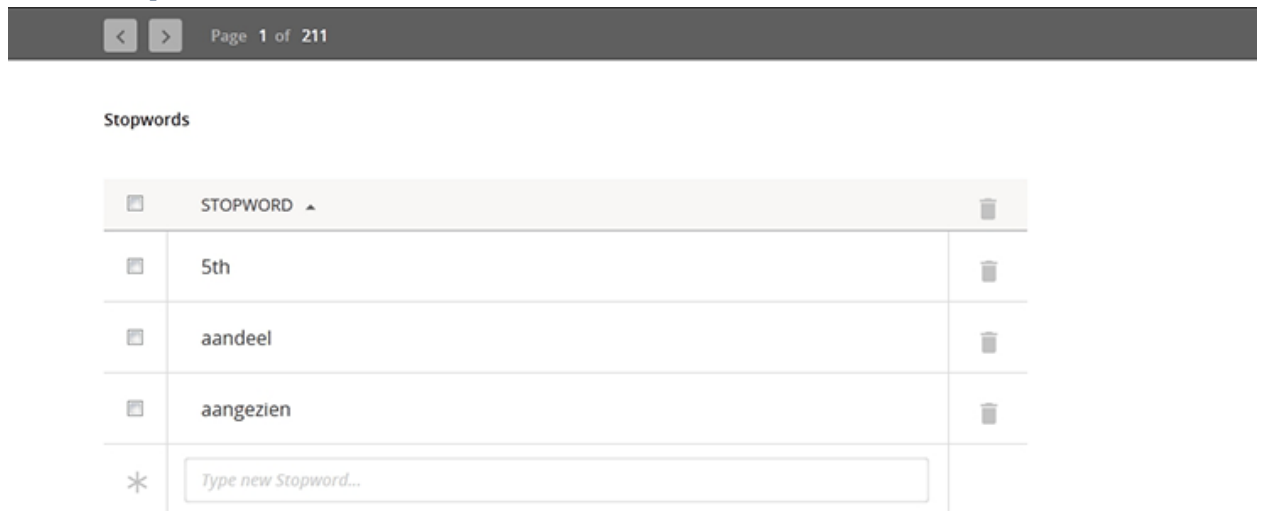
4.4.2 Stopwords

Fig. 22. Stopwords page.

Using "Stopwords" tab administrator can indicate words, which he doesn't want to see in DataCloud, located on the search results page. For example: the name of the user's company. It can be found almost in every document, for which reason it's useless for query specification.

Work with stopwords list is similar to the work with search query list in the "Custom queries" tab.

5 Useful links

Active directory:

- https://technet.microsoft.com/en-us/library/dn283324.aspx - Active Directory Services Overview.
- https://technet.microsoft.com/en-us/library/hh472160.aspx - Deploy Active Directory Domain Services (AD DS) in Your Enterprise.
- https://technet.microsoft.com/en-us/library/jj574166.aspx - Install a New Windows Server 2012 Active Directory Forest (Level 200).

Organization of backup process:

- https://technet.microsoft.com/en-US/library/dn390929.aspx - Windows Server Backup and Storage Pools
- https://en.wikipedia.org/wiki/List_of_backup_software - the list of software for backup (independent vendors, open source).

Elasticsearch

- https://www.elastic.co/downloads/past-releases - download page for Elasticsearch 1.7.*
- https://www.elastic.co/guide/index.html - documentations.

6 Known issues

- Logs folder should have "Full control" permissions for user Everyone. These permissions have been already granted for default folder in Swisscows for Office server, but you need to add them manually if you want to change log's location.
- Uninstalling HES. In case the unsuccessful uninstallation of HES (HES services are still running in the Services applet, a shortcut still present in the Programs and Features applet), try a manual uninstallation scenario:
    1. Stop services Ses.Services.Crawler and Ses.Services.IndexCleaner, using "Services" applet.
    2. Run command console (cmd.exe) as Administrator.
    3. Delete services using the following commands:
        - sc delete Ses.Services.Crawler
        - sc delete Ses.Services.IndexCleaner
    4. If the services are still visible in the "Services", reboot the server.
    5. Delete the folder with installed HES. It is c:\Program Files\Hulbee AG\Hulbee Enterprise Search\ folder by default.
    6. Open "Programs and Features" and delete "Hulbee Enterprise Search" shortcut (press "uninstall" and the delete shortcut option will be proposed after this).
- IIS_IUSRS group should have access to reading files in the crawled storages to enabling the preview for media documents (audio, video, images).