# Set up Tika on Elastic Search

**In command prompt execute this commands**

# change directory

cd /usr/share/elasticsearch

# Install ElasticSearch Mapper Attachments Plugin

sudo ./bin/plugin install elasticsearch/elasticsearch-mapper-attachments/2.4.3

# Restart ElasticSearch Server

sudo service elasticsearch restart

# Check list of plugins

sudo ./bin/plugin –list

**Add new attachment field "file"**

curl –XPUT <http://vehicles-search.cloudapp.net:9200/demo2/_mapping/doc> -d ‘

{

"doc": {

"properties": {

"body": {"type": "string"},

"caseId": {"type": "string"},

"clientId": {"type": "string"},

"fileType": {"type": "string"},

"id": {"type": "string"},

"kind": {"type": "string"},

"storageLocation": {"type": "string"},

"tags": {"properties": {"privileged": {"type": "boolean"}}},

"title": {"type": "string"},

"userId": {"type": "string"},

"file" :

{

"type": "attachment",

"path": "full",

"fields": {

"file": {"type": "string", "term\_vector":"with\_positions\_offsets", "store": true}

}

}

}

}

**} ‘**

# Search queries for attachments

**Full text search**

<http://ec2-54-165-205-102.compute-1.amazonaws.com:9200/demo2/doc/_search?q=job&fields=title,file,clientId,caseId,isAttachment,id,storageLocation,fileType,userId,title,comments,footnotes,body&_source=false>

**Highlight search**

curl –XPOST <http://ec2-54-165-205-102.compute-1.amazonaws.com:9200/demo2/_search> -d ‘

{

"fields": [],

"query": {

"match": {

"file": "job"

}

},

"highlight": {

"fields": {

"file": {

}

}

}

}’

# Create Create Console app that will read files stored on Azure Blob Storage and process with Tika and then upload the JSON from Tika into Elasticsearch

The algorithm of the program

1. Download file from Azure Blob Storage
2. Convert binary format file into BASE64 string
3. Prepare document for loading in ElasticSearch (Fill ParentID, ID and another fields)
4. Load document in Elasticsearch

Run program (Elastic.Attachments.Index)

* **Elastic.Attachments.Index.exe**

Run process on default container name

* **Elastic.Attachments.Index.exe -c demo2**

Run process on container demo2

* **Elastic.Attachments.Index.exe -u /demo2/demo2/1/1/190/emailAttachments/mymicrosExportSpec109.doc**

Run process on file with local path /demo2/demo2/1/1/190/emailAttachments/mymicrosExportSpec109.doc

Using module (Elastic.Attachments.Core)

Module created with help Autofac Module (<http://docs.autofac.org/en/latest/configuration/modules.html>)

Before use module, you need it initialize

builder.RegisterModule(new ElasticAttachmentsModule());

Than inject service IAttachmentsIndexer.

var indexer = ServiceLocator.Resolve<IAttachmentsIndexer>();

var result = indexer.Index("/demo2/demo2/1/1/190/emailAttachments/mymicrosExportSpec109.doc");

indexer.Index();

indexer.Index(“demo2”);

Settings of modules/program

You can modify settings module/program in app.config

<appSettings>

<add key="StorageConnectionString" value="DefaultEndpointsProtocol=https;AccountName=datageekslaw;AccountKey=t2+vt32W8PfUtHPA+z9inAfHa9BuXyGfhoC/V1GSKRiGj7GWBQknKMVDs6UFVGUaVvPxlxwFCu0zJQQiqF0zkg==" />

<add key="StorageContainerName" value="demo2" />

<add key="searchengine-host" value="ec2-54-165-205-102.compute-1.amazonaws.com:9200" />

<add key="searchengine-basic-auth" value="False" />

<add key="searchengine-user" value="retret" />

<add key="searchengine-password" value="tretretre" />

<add key="searchengine-indexname" value="demo2" />

</appSettings>

* StorageConnectionString – connection string Azure Blob Storage
* StorageContainerName – default container name in Azure Blob Storage
* searchengine-host – host and port of ElasticSearch server
* searchengine-basic-auth – enable/disable basic auth in connections to ElasticSearch
* searchengine-user – username for basic auth
* searchengine-password – password for basic auth
* searchengine-indexname – default index name of ElasticSearch

Third party libraries

* Autofac <http://autofac.org/> - IOC library
* NEST <http://nest.azurewebsites.net/> - ElasticSearch library
* Windows Azure Storage <https://msdn.microsoft.com/library/azure/gg433040.aspx> - Azure Blob Storage library
* Command Line Parser Library <https://github.com/gsscoder/commandline> - for parse command line arguments

**Install libraries:**

* Install-Package Autofac -Version 3.5.2
* Install-Package NEST -Version 1.4.2
* Install-Package WindowsAzure.Storage -Version 4.3.0
* Install-Package CommandLineParser -Version 1.9.71