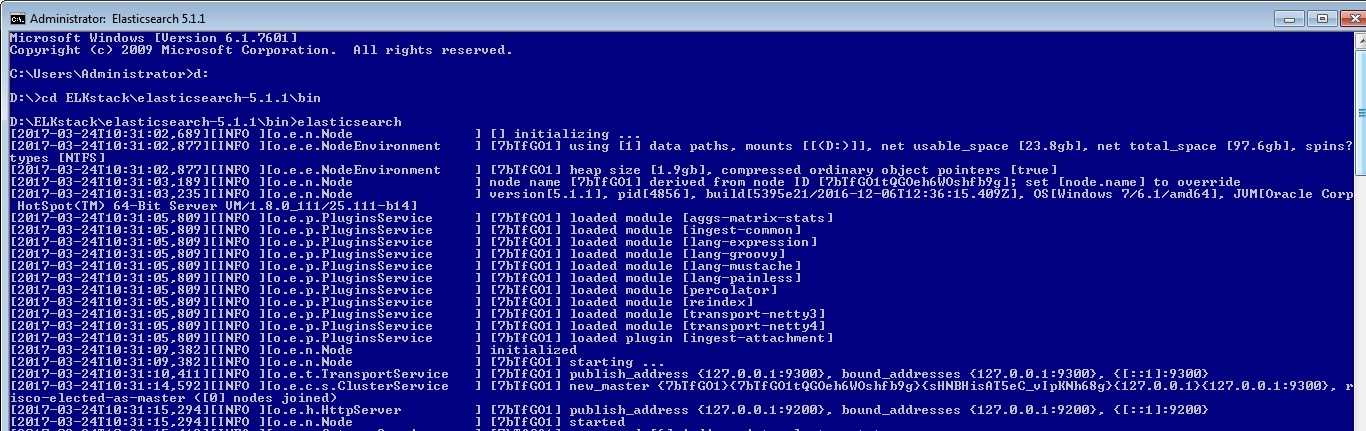
**ELASTICSEARCH(Indexing pdf file,Doc file)**

**PREREQUISITES –**

1. Elasticsearch 5.2
2. FScrowler-2.3
3. KIBANA 5.2
4. Set JAVA\_HOME (required java 1.8 and above)

**STEP 1** – Start Elsticsearch

/bin>elasticsearch



**STEP 2** –Download Fscrawler from following <https://oss.sonatype.org/content/repositories/snapshots/fr/pilato/elasticsearch/crawler/fscrawler/2.3-SNAPSHOT/>

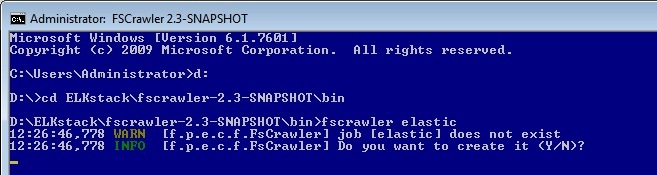
File Name- [fscrawler-2.3-20170418.072434-25.zip](https://oss.sonatype.org/content/repositories/snapshots/fr/pilato/elasticsearch/crawler/fscrawler/2.3-SNAPSHOT/fscrawler-2.3-20170418.072434-25.zip)

**STEP3** -Start Fscrawler

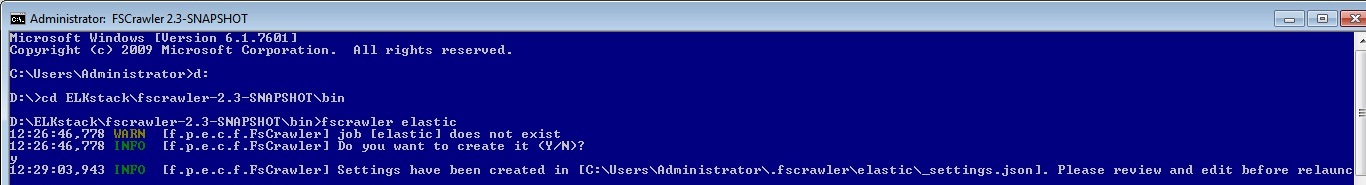
/bin>fscrawler job\_name

FS crawler will read a local file(default to- C:\Users\Administrator\.fscrawler\elastic\\_setting.json)

 If the file does not exist, FS crawler will propose to create your first job.



Type “Y” to create index.It will give output as follows



It will create \_setting.json file in path “C:\Users\Administrator\.fscrawler\elastic\\_setting.json

Edit \_setting.json

{

"name" : "deepak",

"fs" : {

"url" : "/tmp/es", Change to C:/tmp/< name of file

"update\_rate" : "15m",

"excludes" : [ "~\*" ],

"json\_support" : false,

"filename\_as\_id" : false,

"add\_filesize" : true,

"remove\_deleted" : true,

"add\_as\_inner\_object" : false,

"store\_source" : false,

"index\_content" : true,

"attributes\_support" : false,

"raw\_metadata" : true,

"xml\_support" : false,

"index\_folders" : true,

"lang\_detect" : false,

"continue\_on\_error" : false

},

"elasticsearch" : {

"nodes" : [ {

"host" : "127.0.0.1",

"port" : 9200,

"scheme" : "HTTP"

} ],

"type" : "doc",

"bulk\_size" : 100,

"flush\_interval" : "5s"

},

"rest" : {

"scheme" : "HTTP",

"host" : "127.0.0.1",

"port" : 8080,

"endpoint" : "fscrawler"

}

}

**STEP 4** - Create a directory named /tmp/es or c:\tmp\es, add some files you want to index in it and start again:

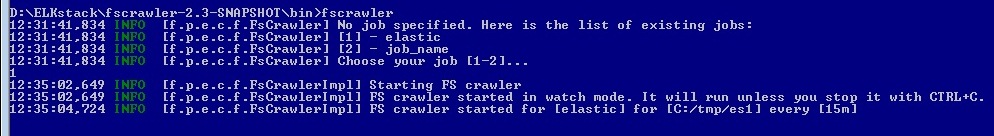
/bin>fscrawler

If you did not create the directory, FS crawler will complain until you fix it:

18:30:34,683 WARN [f.p.e.c.f.FsCrawlerImpl] Error while indexing content from /tmp/es: /tmp/es doesn't exists.

Again Run

/bin>fscrawler



It will ask to select your job. Select your job and press enter.

**Crawler Option-**

By default, FS crawler will read your file from **C:/tmp/es** every **15** minutes. You can change those settings by modifying **~/.fscrawler/{job\_name}/\_settings.json** file where {job\_name} is the name of the job you just created.

{

"name" : "job\_name",

"fs" : {

"url" : "C:/temp/elastic",

"update\_rate" : "15m"

}

}

-You can change also update\_rate to watch more or less frequently for changes.

-You can also change a elasticsearch port

"elasticsearch" : {

"nodes" : [ {

"host" : "127.0.0.1",

"port" : 9200,

"scheme" : "HTTP"

} ]

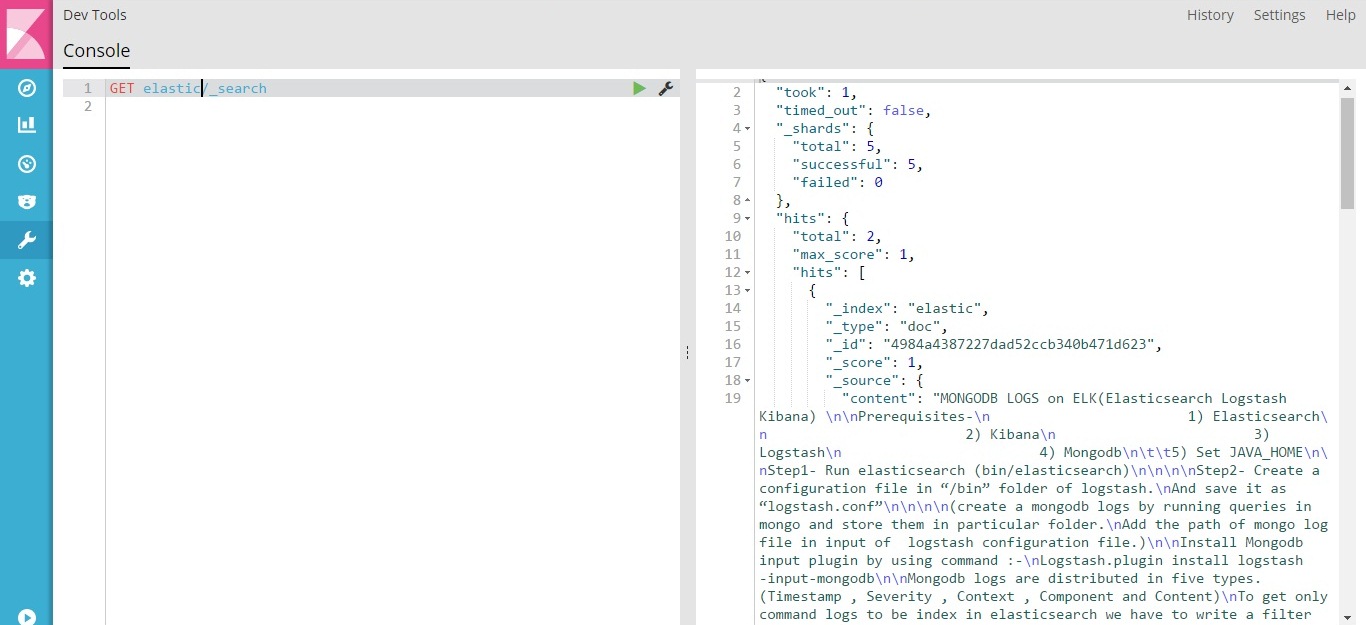
(For this step first you have change elasticsearch port in elasticsearch.yml and kibana.yml)

**STEP 5** – Start kibana by running command

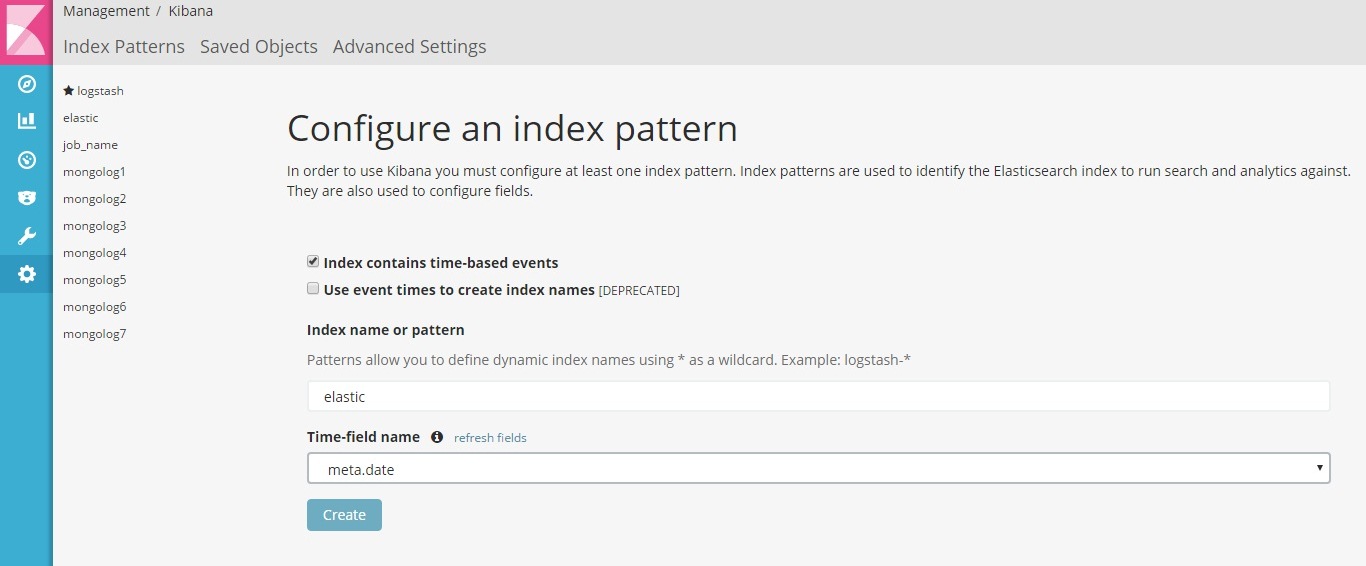
/bin>kibana

Hit “localhost:5601” on browser.

**STEP 6** - Serach for index (click on dev tools window) which you created as your job\_name

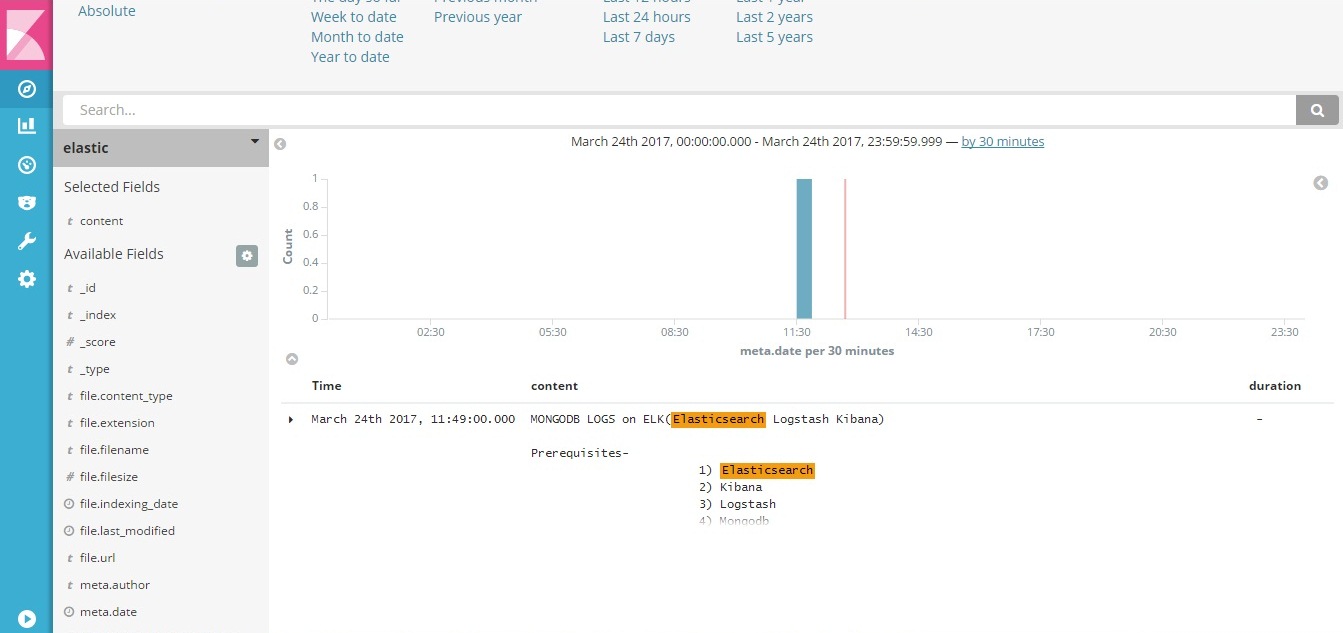


**STEP 7** – Configure Index in kibana (Click on management window to configure your index)



Click on create.

**STEP 8** – Discover Index in kibana



You can index multipal file ( eg. Pdf, Word file) in elasticsearch using Fscrawler.