Installation and Usage Guide for News Crawler

Requirements

Following packages/softwares are required to be installed, for using News Crawler:

S.No.	Software/Package	Version	Remarks
1.	Python	2.7	
2.	Scrapy	1.4.0	
3.	geograpy	0.3.7	Refer URL: https://pypi.python.org/pypi/geograpy
4.	geopy	1.11.0	Refer URL: https://pypi.python.org/pypi/geopy
5.	Firefox	54.0.1	Tested on version 54.0.1, should work on other versions as well.
6.	Elastic Search	5.4.2	
7.	Kibana	5.4.2	

Table 1: Requirements

Directory Structure

Source code directory of News Crawler 'newsCrawler' consists of following file/directories:

1. **File 'config.ini'**
    a. In this file, different parameters are defined which are used for executing the News Crawler.
    b. End user needs to update this file, before executing the News Crawler.
    c. Different parameters defined in config.ini file are described below:

NOTE: While updating the 'config.ini' file, please don't change the parameter names.

S.No.	Config Parameter	Description	Example Values
1.	es_host	Host Name of Elastic Search	localhost
2.	es_port	Port number of Elastic Search	9200
3.	index_name	Name of Elastic Search index to be created, for storing the crawled news.	News
4.	type_name	News Mapping type name as specified in mapping file.	news_type
5.	mapping_file	File in which Index Mapping is described, which is used by News Crawler for creating Elastic Search index.	newsMapping.json
6.	delete_index	Flag which denotes, whether the old 'news' index in Elastic Search is to be deleted or not. NOTE: 1. Once the News Crawler is executed, 'delete_index' is set to zero (by Crawler), so	Possible Values: 0 (i.e. don't delete the index) or 1 (i.e. delete the index if exists).

		that when the crawler executes again, old data is not deleted.	
7.	index_id	Document id from which document insertion is to be started in Elastic Search engine. NOTE: 1. When index is created then 'index_id', should be kept 0. 2. Once News Crawler completes the execution, 'index_id' is set (by Crawler) to total number of documents inserted to elastic search. So, that when crawler executes again documents are not overwritten.	Initially value should be 0.
8.	url_list	CSV file containing list of URLs to be crawled.	urls.csv

Table 2: Configuration Parameters

2. **File 'urls.csv'**
   - This file contains list of News RSS URLs to be crawled.
   - End user can add more RSS URLs of News websites in the original format of file.

3. **File 'newsMapping.json'**
   - This file defines the index mapping, which is used while creating an Elastic Search Index.

4. **File 'configReader.py'**
   - This file reads the 'config.ini' file and make config parameters available to news crawler.

5. **File 'newsCrawler.py'**
   - This file contains implementation of News Crawler using Scrapy library.

6. **File 'main.py'**
   - This is the main executor file of News Crawler, which parsers the command line arguments specified by end user and executes the crawler after every 15 minutes.

7. **Directory 'TimeStamp'**
   - This directory contains 'stamp' files. These stamp files represent maximum publish time of news item, inserted to Elastic Search for each news category.
   - This directory is initially empty when crawler is executed for the First Time. Once the crawler is executed 'stamp' files are written to this directory, corresponding to each news category.
   - These stamp files are used by Crawler in subsequent execution, to ensure that no duplicate news entries are created in Elastic Search.

8. **File ‘geckodriver.log’**
    - This file provides log of Firefox browser, which is launched every time news crawler is executed and closes automatically when crawling completes.

Usage

Follow below steps to execute News Crawler:

1. Run Elastic Search in background. Make sure Elastic Search is up and running.
2. Update the ‘config.ini’ file, for Elastic Search host, port and other parameters as per requirement. Please refer Section **File ‘config.ini’** for description of configuration parameters.
3. [Optional:] Update the file ‘urls.csv’, with additional RSS News Website URLs to be scraped.
4. Change to directory ‘newCrawler’ and execute the following command:

    $ python main.py --AnalysisTime=60

    Where:

    - ‘--AnalysisTime' is a command line argument, which describes total time (in minutes), for which News Crawler is to be executed.
    - For example: ‘--AnalysisTime=60’ means that, News Crawler is executed for 1 hour (i.e. 60 minutes), where given URLs are scraped after every 15minutes to check for latest news.

Kibana Dashboard

- For visualizing crawled news on Kibana. Visualization JSON file is provided in directory ‘kibanaFiles’. This JSON file contains Dashboard with all the visualizations to be displayed in kibana.
- Once News Crawler has started executing, launch Kibana on any Web Browser and configure the index ‘news’ as shown below:

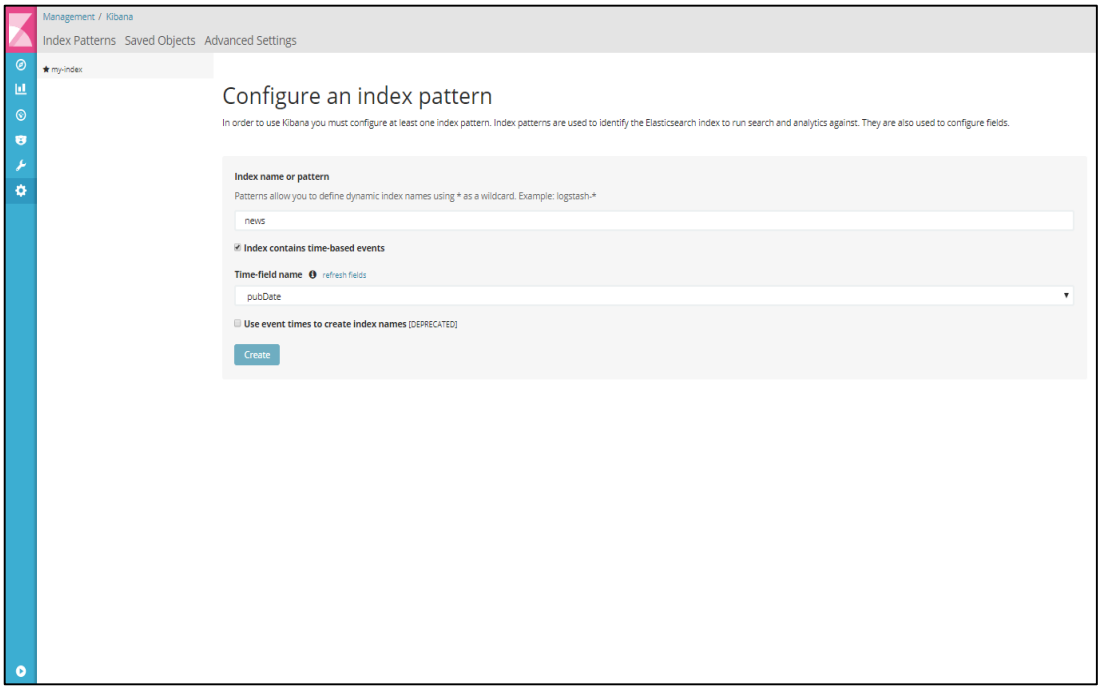

Figure 1

- Once the index ‘news’ has been created, following Index Pattern is displayed:

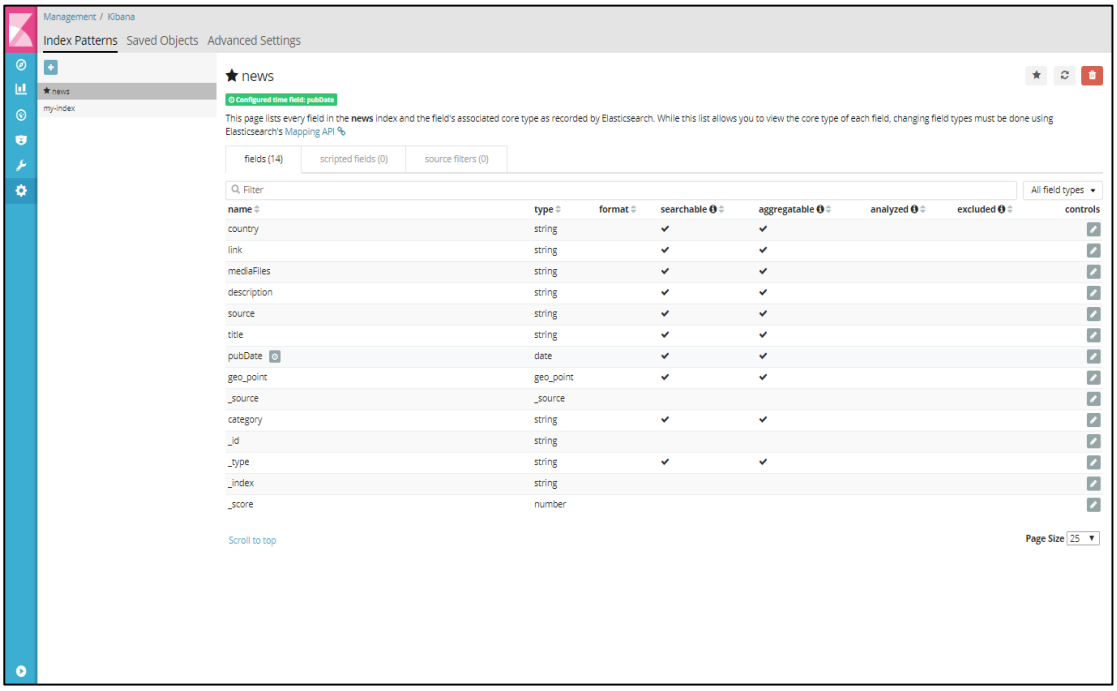

Figure 2

- Update Index format of fields ‘link’, ‘mediaFiles’ and ‘pubDate’ by pressing controls on right side of each field as shown below:

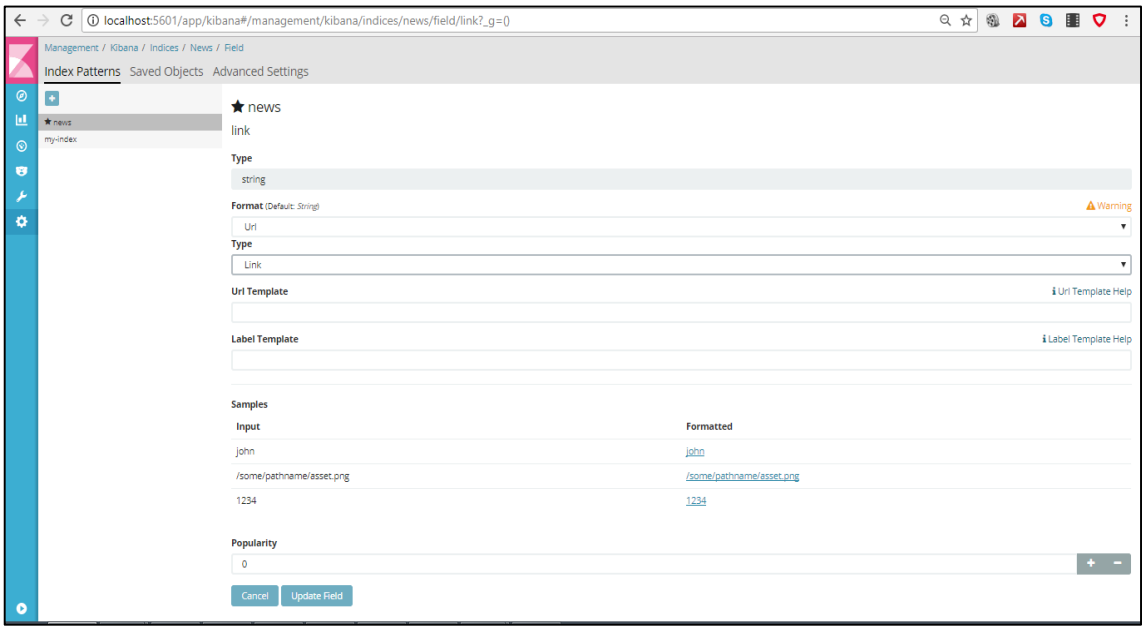

Figure 3: Updating ‘link’ format

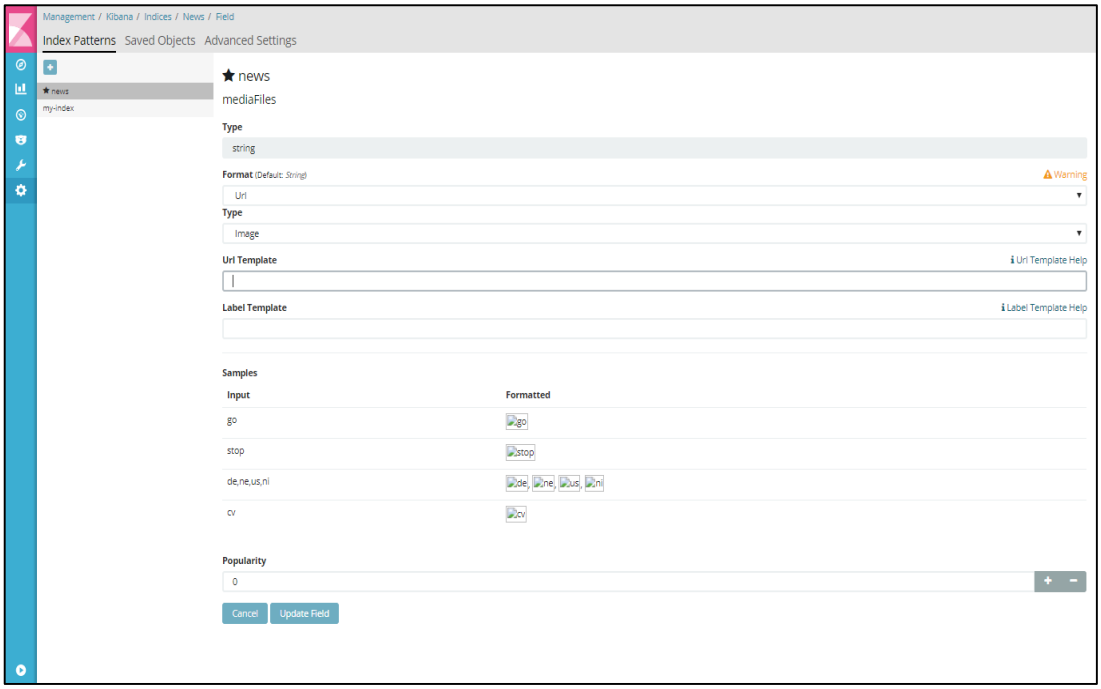

Figure 4: Updating 'mediaFiles' format

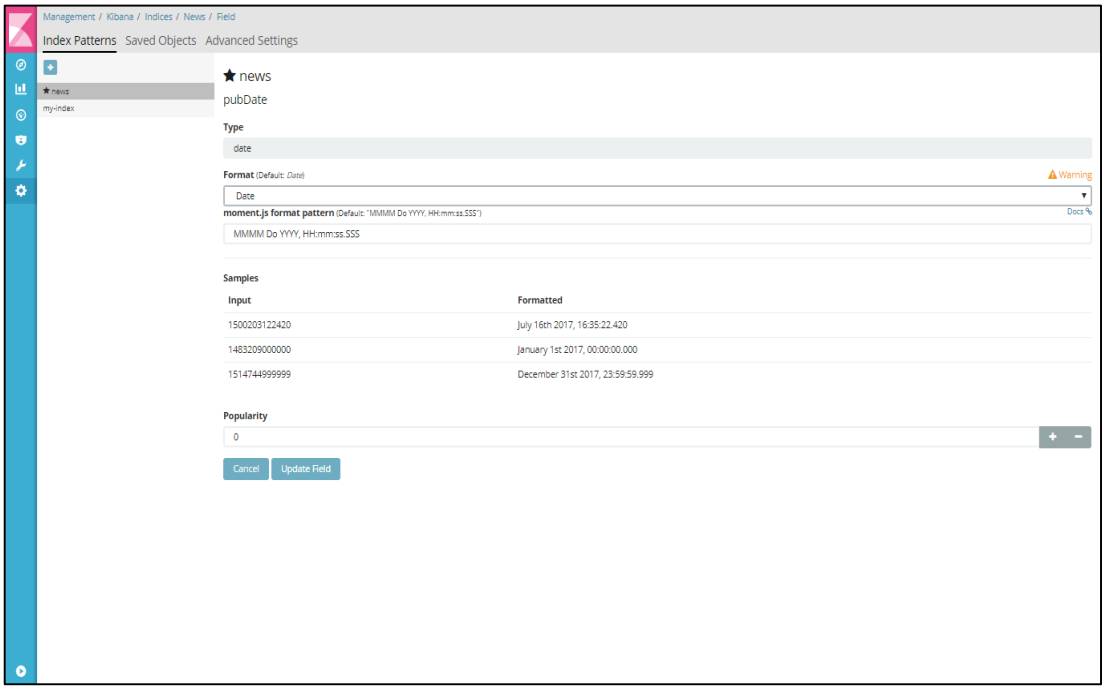

Figure 5: Updating 'pubDate' format

- Once Format of Index fields are updated, now on launched Kibana go to ‘Saved Objects’ and import the file ‘Visualization.json’ in directory ‘kibanaFiles’.
- Once import is complete, open the News Dashboard, which will be displayed as below:

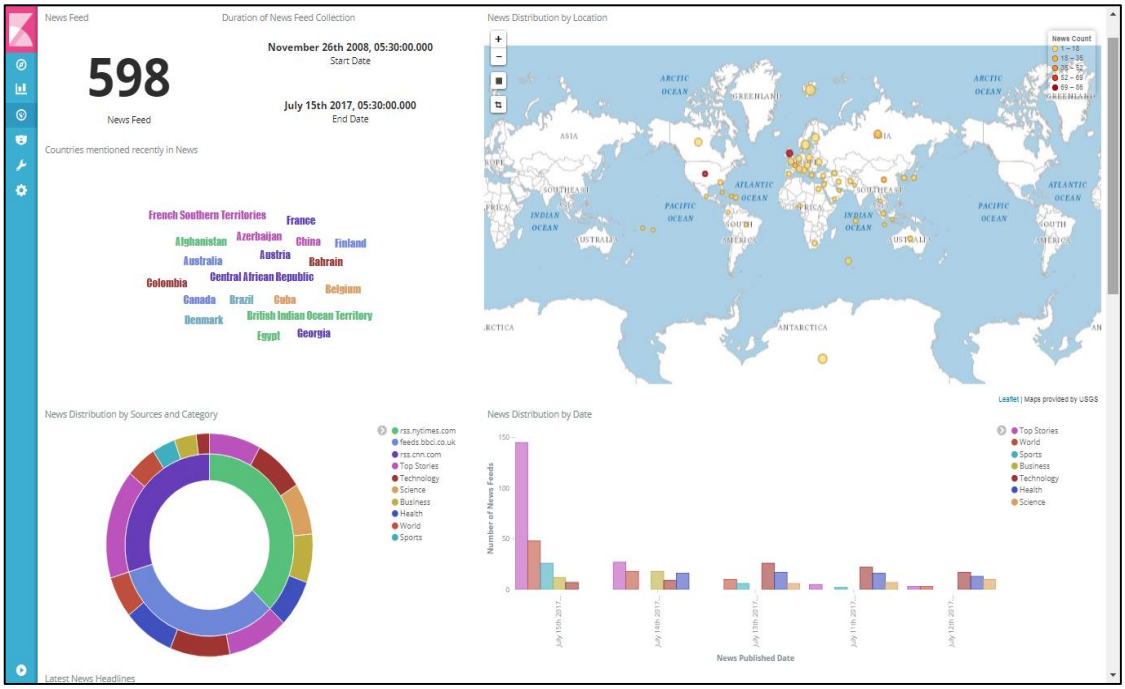

Figure 6: News Dashboard View 1

Figure 7: News Dashboard View 2