

Elastic Search: Load Data[1-3]

Ways to Collect and Index Big Data

Weimao Ke

Drexel University

Table of contents

1. Bulk APIs
2. Beats and Logstash
3. Kibana XLSX CSV Import

Bulk APIs

Bulk API

Performs multiple indexing or delete or update operations in a single API call.

POST `/_bulk`

POST `/<index>/_bulk`

Example:

```
1  POST _bulk
2  { "index" : { "_index" : "test", "_id" : "1" } }
3  { "field1" : "value1" }
4  { "delete" : { "_index" : "test", "_id" : "2" } }
5  { "create" : { "_index" : "test", "_id" : "3" } }
6  { "field1" : "value3" }
7  { "update" : { "_id" : "1", "_index" : "test" } }
8  { "doc" : { "field2" : "value2" } }
```

Index many documents in one request:

```
1  POST /courses/_bulk
2  { "index" : { "_id" : "1" } }
3  { "number" : "INFO624", "title":"IR Systems" }
4  { "index" : { "_id" : "2" } }
5  { "number" : "INFO634", "title":"Data Mining" }
6  { "index" : { "_id" : "3" } }
7  { "number" : "INFO659", "title":"Data Analytics" }
8  ...
```

Much more efficient than multiple POST requests.

But you still need to prepare your data in the JSON format.

Beats and Logstash

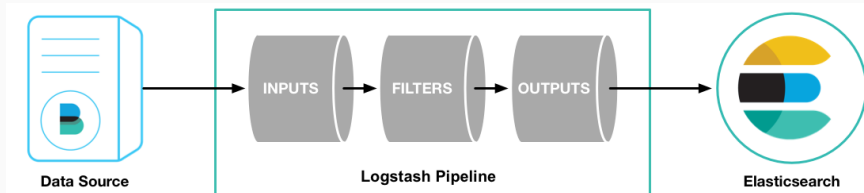
Beats are data shippers (agents) to send data to Elasticsearch

- Beats are lightweight data shipper with small footprint.
- Beats for different data: log files (Filebeat), availability (Heartbeat), network traffic (Packetbeat), etc.
- Each Beat can be installed on your servers to send (and index) ongoing, operational data.

Logstash

Logstash is a data collection engine:

- More powerful than Beats with input, filter, and output filters;
- Real-time pipelines to collect, enrich, and transform data from various sources;
- Originally designed for log collection but is well beyond that.



Logstash can collect data from:

- Logs and metrics
- The World Wide Web, based on pages and service APIs
- Data stores and stream, SQL, NoSQL, etc.
- Sensors and IoT

Logstash can clean and transform data:

- Unstructured to structured data (Grok)
- Geo decoding (location), date normalization, fingerprinting (anonymizing) sensitive information, etc.
- Codecs for common structures

Ongoing collection of big data from various sources.

Kibana XLSX CSV Import

Kibana plugin: XLSX CSV Import

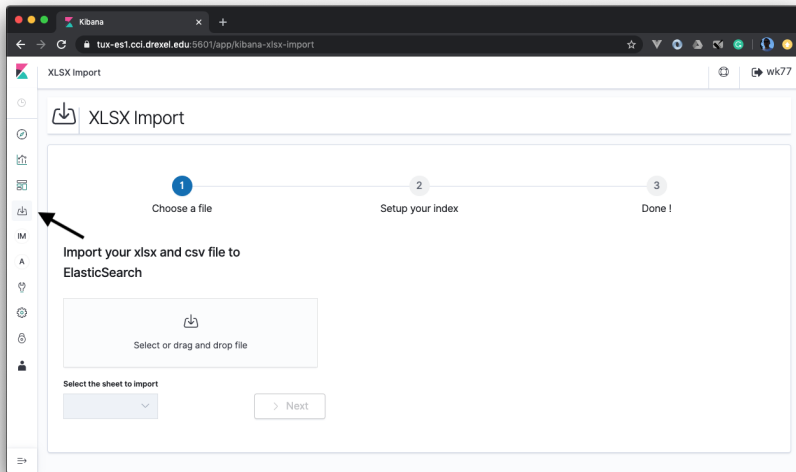
- Web interface via Kibana;
- Imports data from a XLSX (Excel) or CSV format;
- Indexes data to Elasticsearch;
- Supports basic mappings and good for smaller datasets;

Available at:

<https://github.com/Orange-OpenSource/kibana-xlsx-import>

XLSX CSV Import

XLSX CSV Import interface (1):



XLSX CSV Import

XLSX CSV Import interface (2):

The screenshot shows the Kibana XLSX Import interface. The browser address bar indicates the URL is `tux-es1.cci.drexel.edu:5601/app/kibana-xlsx-import`. The page title is "XLSX Import".

The main heading is "Import your xlsx and csv file to Elasticsearch". Below this, there is a file upload area showing a file named "fake.xlsx" with a "Remove" link. Below the upload area, there is a section titled "Select the sheet to import" with a dropdown menu showing "Sheet1" and a "> Next" button.

Below the selection area, there is a table preview showing the first five rows of data. The table has 20 columns: `uuid`, `ord_i...`, `author`, `publi...`, `title`, `text`, `lang...`, `craw...`, `site...`, `coun...`, `dom...`, `thre...`, `spa...`, `main...`, `repli...`, `parti...`, `likes`, `com...`, `shares`, and `type`.

uuid	ord_i...	author	publi...	title	text	lang...	craw...	site...	coun...	dom...	thre...	spa...	main...	repli...	parti...	likes	com...	shares	type
90a...	0	Alex...	201...	The...	The...	engl...	201...	amt...	US		The...	0	http...	0	1	0	0	0	bs
513...	0	Alex...	201...	Opi...	Opi...	engl...	201...	amt...	US		Opi...	0	http...	0	1	0	0	0	bs
731...	0	Alex...	201...	US ...	US ...	engl...	201...	amt...	US		US ...	0	http...	0	1	0	0	0	bs
3e0...	0	Alex...	201...	Is A...	Is A...	engl...	201...	amt...	US		Is A...	0	http...	0	1	0	0	0	bs
1f9...	0	Alex...	201...	Tru...	Tru...	engl...	201...	amt...	US		Tru...	0	http...	0	1	0	0	0	bs

XLSX CSV Import

XLSX CSV Import interface (3):

The screenshot shows the Kibana XLSX Import interface. The browser address bar indicates the URL is `tux-es1.cci.drexel.edu:5601/app/kibana-xlsx-import`. The interface has a sidebar on the left with icons for various Kibana features. The main content area is titled 'XLSX Import' and shows a progress bar with three steps: 1. Choose a file (completed), 2. Setup your index (current step), and 3. Done! (pending).

Index name

Name the elasticsearch index that will be created. If the index is already existing, documents will be added or updated according to the chosen docID

Custom docID

example rendering

Import will provide a unique document identifier linked to the line number of the imported file. You can customize this doc ID using special reserved variables : `{_uid}` for an auto-generated identifier, `{_importedLine}` for the current line number, or `{<column-name>}` to access a value of the imported line.

Removing columns

XLSX CSV Import

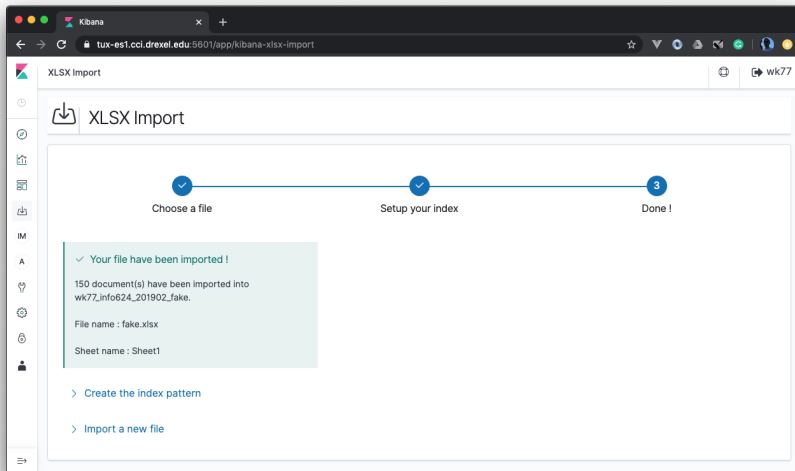
XLSX CSV Import interface (4):

The screenshot displays the Kibana XLSX Import interface. At the top, a toggle switch is set to 'On' with the label 'Configure your own mapping'. Below this, a table lists fields from the imported file and allows mapping them to Elasticsearch types or writing custom JSON mappings.

Fields	Type	Advanced JSON
uuid	Text	<pre>{ "type": "text", ... }</pre>
ord_in_thread	Float	<pre>{ "type": "float", ... }</pre>
author	Text	<pre>{ "type": "text", ... }</pre>
published	Text	<pre>{ "type": "text", ... }</pre>

XLSX CSV Import

XLSX CSV Import interface (5):



XLSX CSV Import

Recommended steps:

1. Create an index with proper mappings;
2. Then import data from XLSX or CSV into the index;

Example mappings:

```
PUT /index_to_host_data
{
  "mappings": {
    "properties": {
      "age":    { "type": "integer" },
      "email": { "type": "keyword" },
      "name":   { "type": "text", "analyzer": "standard" }
    }
  }
}
```

References

- [1] elastic.co. Elasticsearch reference: Beats getting started.
<https://www.elastic.co/guide/en/beats/libbeat/current/getting-started.html>, . Accessed: 2020-1-16.
- [2] elastic.co. Elasticsearch reference [7.5]: Document apis.
<https://www.elastic.co/guide/en/elasticsearch/reference/current/docs.html>, . Accessed: 2020-1-16.
- [3] elastic.co. Elasticsearch reference: Logstash introduction.
<https://www.elastic.co/guide/en/logstash/current/introduction.html>, . Accessed: 2020-1-16.