Elastic Search: Load Data[1–3]

Ways to Collect and Index Big Data

Weimao Ke

Drexel University

Table of contents

Bulk APIs

Bulk API

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

```
POST /_bulk
```

```
POST /<index>/_bulk
```

Example:

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

Bulk API

Index many documents in one request:

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

Much more efficient than multiple POST requests.

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

Beats and Logstash

Beats

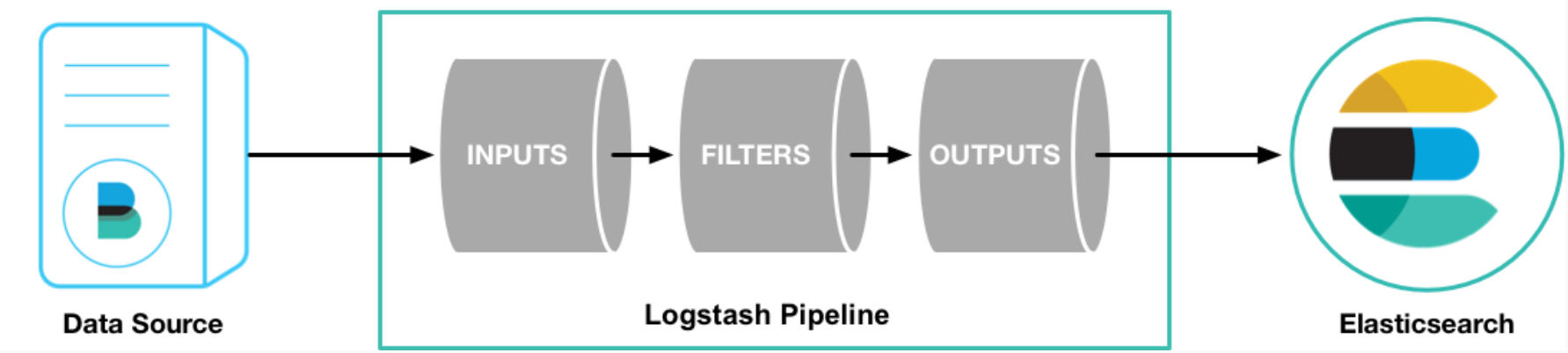
Beats are data shippers (agents) to send data to ElasticSearch

- Beats are lightweight data shipper with small footprint.
- Beats for different data: log files (Filebeat), availability (Hearbeat), 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

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

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

XLSX CSV Import

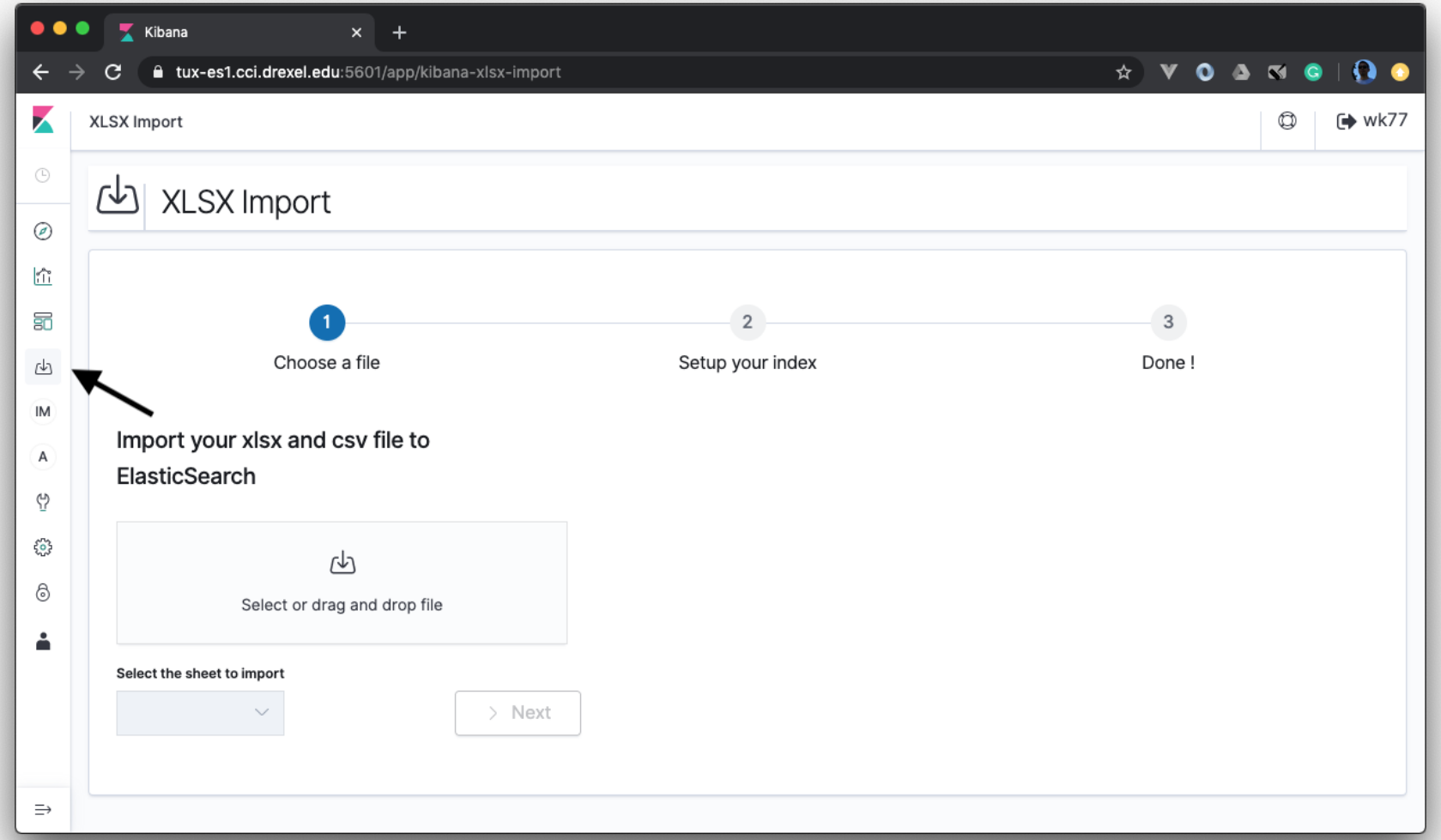
Kibana plugin: XLSX CSV Import

- 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`

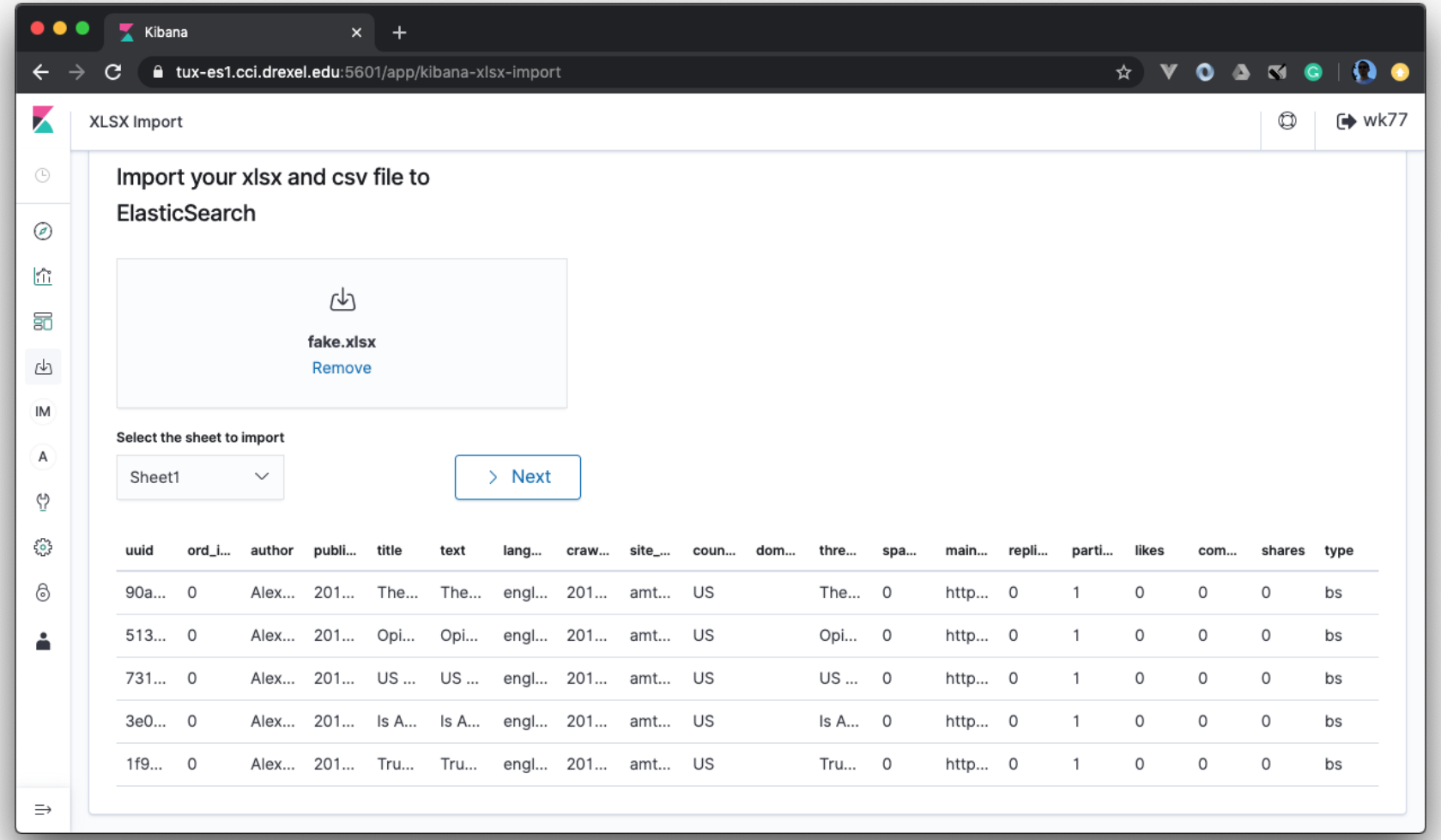
XLSX CSV Import

XLSX CSV Import interface (1):



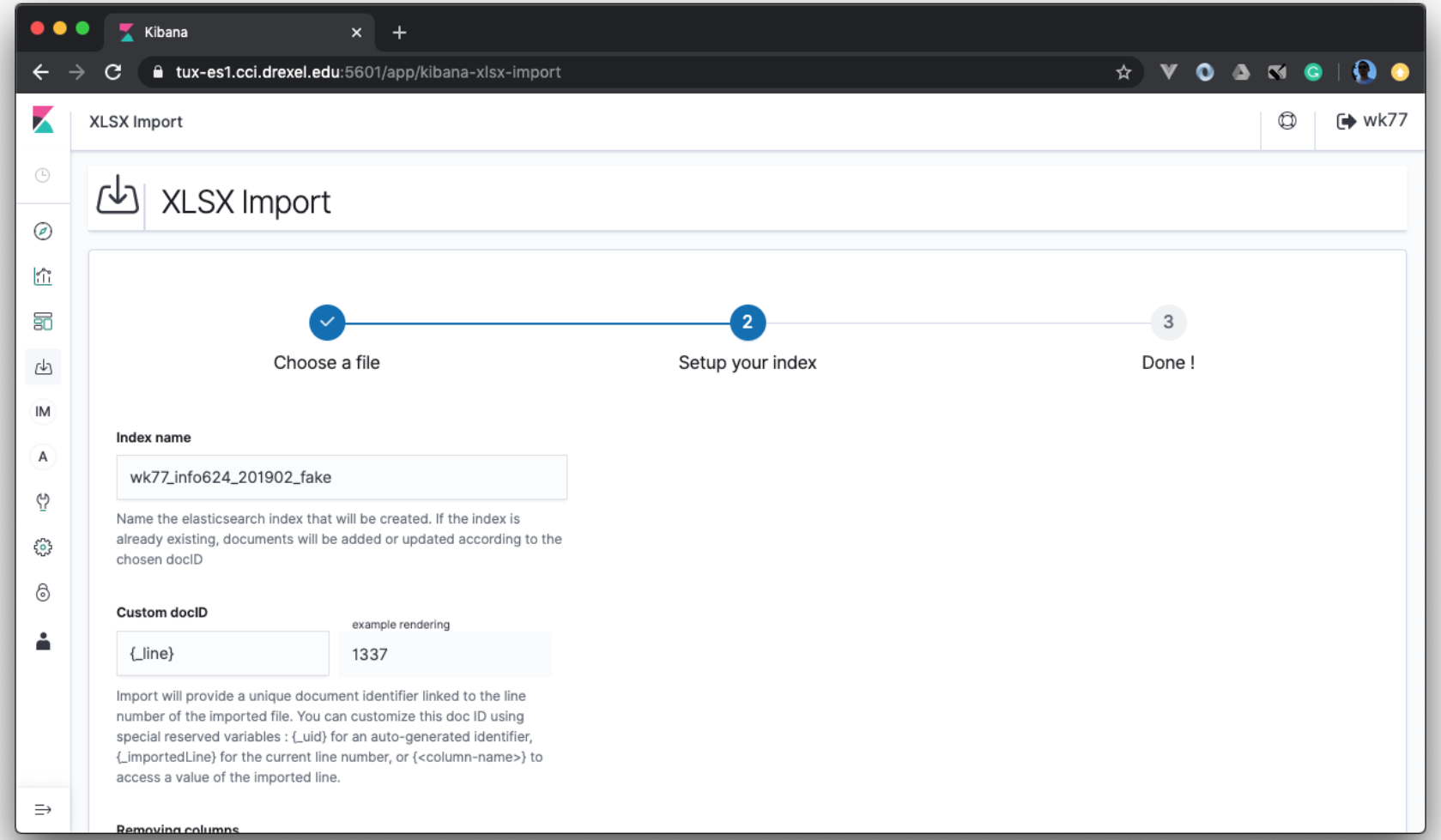
XLSX CSV Import

XLSX CSV Import interface (2):



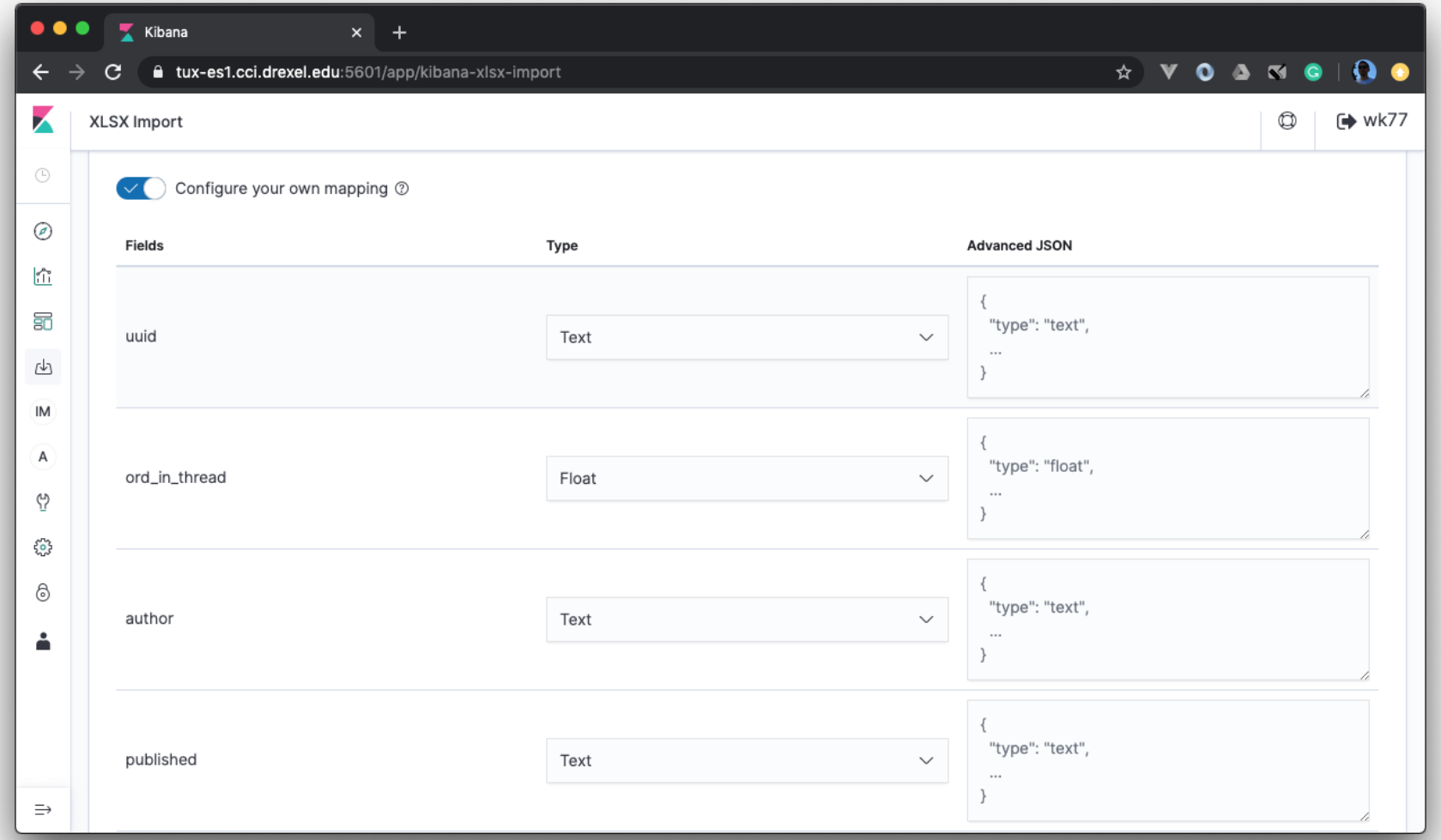
XLSX CSV Import

XLSX CSV Import interface (3):



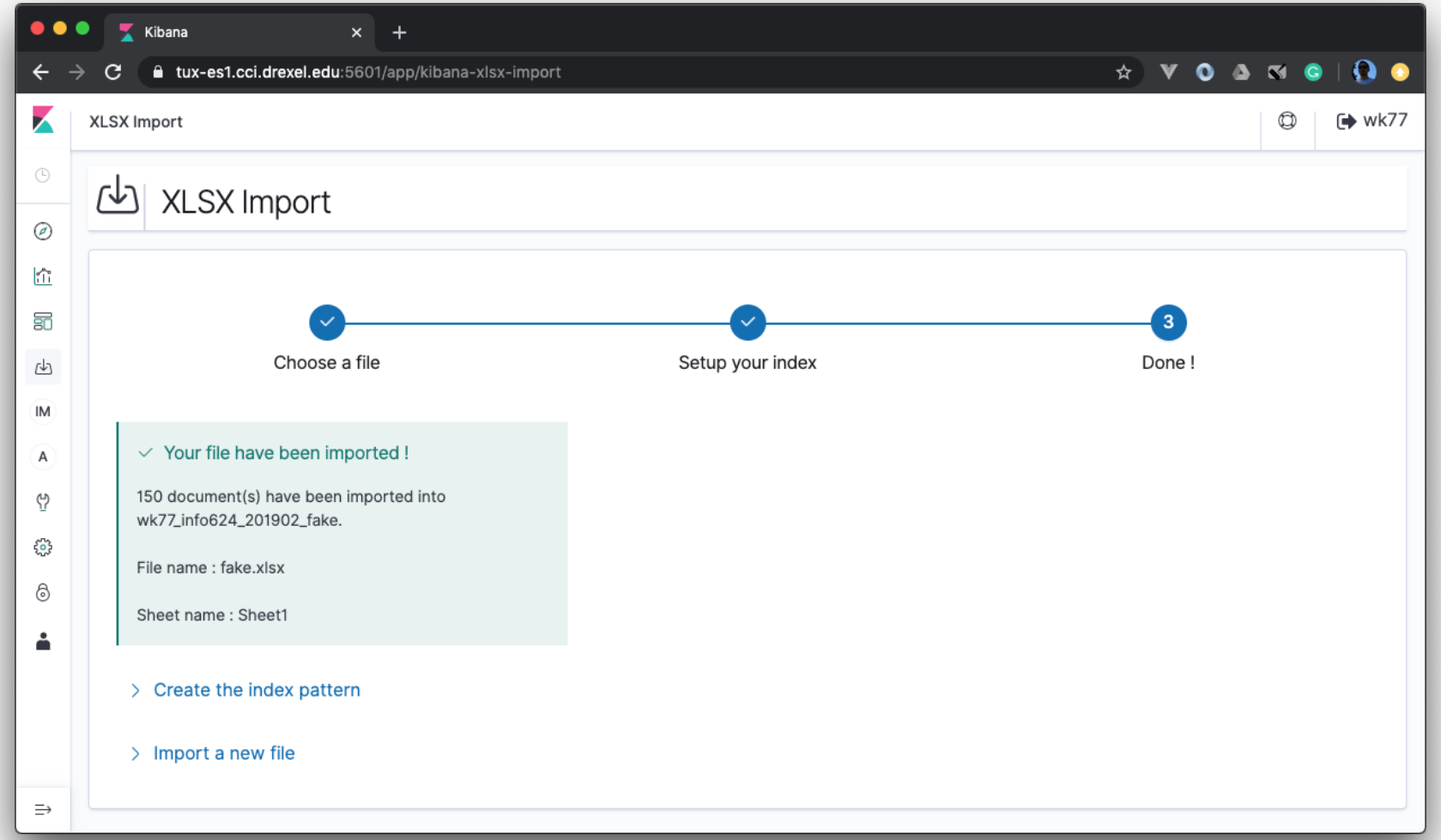
XLSX CSV Import

XLSX CSV Import interface (4):



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.