



Indexing your office documents with Elastic and FSCrawler

David Pilato

Developer / Evangelist, Community

[@dadoonet](https://twitter.com/dadoonet)

Apache Tika - a content analysis toolkit

The Apache Tika™ toolkit detects and extracts metadata and text from over a thousand different file types (such as PPT, XLS, and PDF). All of these file types can be parsed through a single interface, making Tika useful for search engine indexing, content analysis, translation, and much more. You can find the latest release on the [download page](#). Please see the [Getting Started](#) page for more information on how to start using Tika.


The [Parser](#) and [Detector](#) pages describe the main interfaces of Tika and how they work.

If you're interested in contributing to Tika, please see the [Contributing](#) page or send an email to the [Tika development list](#).


Tika is a project of the [Apache Software Foundation](#) , and was formerly a subproject of [Apache Lucene](#) .

Latest News

11 February 2022: Apache Tika Release

Apache Tika 1.28.1 has been released! This release includes security related dependency upgrades. Please see the [CHANGES.txt](#)  file for the full list of changes in the release and have a look at the [download page](#) for more information on how to obtain Apache Tika 1.28.1. **Note:** The Apache Tika PMC has set September 30, 2022 as the End Of Life for the Tika 1.x branch. The PMC will make security fixes for the 1.x branch until that date.

07 February 2022: Apache Tika Release

Apache Tika 2.3.0 has been released! This release includes several security upgrades in dependencies, including an upgrade to log4j2 (version 2.17.1). This release also includes a non-trivial upgrade to Apache POI 5.2.0 (TIKA-3164); users will observe significantly more logging from the POI parsers. Please see the [CHANGES.txt](#)  file for the full list of changes in the release and have a look at the [download page](#) for more information on how to obtain Apache Tika 2.3.0.

Apache Tika

- [Introduction](#)
- [Download](#)
- [Contribute](#)
- [Mailing Lists](#)
- [Tika Wiki](#) 
- [Issue Tracker](#) 
- [Security](#)

Documentation

- [Apache Tika 2.3.0](#)
 - [Getting Started](#)
 - [Supported Formats](#)
 - [Parser API](#)
 - [Parser 5min Quick Start Guide](#)
 - [Content and Language Detection](#)
 - [Configuring Tika](#)
 - [Usage Examples](#)
 - [API Documentation](#)
- [Apache Tika 1.28.1](#)
- [Apache Tika 2.2.1](#)
- [Apache Tika 1.28](#)
- [Apache Tika 2.2.0](#)
- [Apache Tika 2.1.0](#)
- [Apache Tika 2.0.0](#)
- [Apache Tika 1.27](#)
- [Apache Tika 1.26](#)
- [Apache Tika 1.25](#)
- [Apache Tika 1.24.1](#)
- [Apache Tika 1.24](#)
- [Apache Tika 1.23](#)
- [Apache Tika 1.22](#)
- [Apache Tika 1.21](#)
- [Apache Tika 1.20](#)
- [Apache Tika 1.19.1](#)
- [Apache Tika 1.19](#)
- [Apache Tika 1.18](#)
- [Apache Tika 1.17](#)
- [Apache Tika 1.16](#)
- [Apache Tika 1.15](#)
- [Apache Tika 1.14](#)
- [Apache Tika 1.13](#)
- [Apache Tika 1.12](#)
- [Apache Tika 1.11](#)
- [Apache Tika 1.10](#)
- [Apache Tika 1.9](#)
- [Apache Tika 1.8](#)
- [Apache Tika 1.7](#)
- [Apache Tika 1.6](#)
- [Apache Tika 1.5](#)
- [Apache Tika 1.4](#)
- [Apache Tika 1.3](#)
- [Apache Tika 1.2](#)
- [Apache Tika 1.1](#)

33 D I 0031 I I E3 B I

Apache Tika 2.3.0

the full list of changes in the release and have a look at the download page for more information on how to obtain

Apache Tika 2.3.0. Please see the CHANGES.txt file for the full list of changes in the release and have a look at the download page for more information on how to obtain

Please note that Apache Tika is able to detect a much wider range of formats than those listed below, this page only documents those formats from which Tika is able to extract metadata and/or textual content.

- [Supported Document Formats](#)
 - [HyperText Markup Language](#)
 - [XML and derived formats](#)
 - [Microsoft Office document formats](#)
 - [OpenDocument Format](#)
 - [iWorks document formats](#)
 - [WordPerfect document formats](#)
 - [Portable Document Format](#)
 - [Electronic Publication Format](#)
 - [Rich Text Format](#)
 - [Compression and packaging formats](#)
 - [Text formats](#)
 - [Feed and Syndication formats](#)
 - [Help formats](#)
 - [Audio formats](#)
 - [Image formats](#)
 - [Video formats](#)
 - [Java class files and archives](#)
 - [Source code](#)
 - [Mail formats](#)
 - [CAD formats](#)
 - [Font formats](#)
 - [Scientific formats](#)
 - [Executable programs and libraries](#)
 - [Crypto formats](#)
 - [Database formats](#)
 - [Natural Language Processing](#)
 - [Image and Video object recognition](#)

Parsing a stream

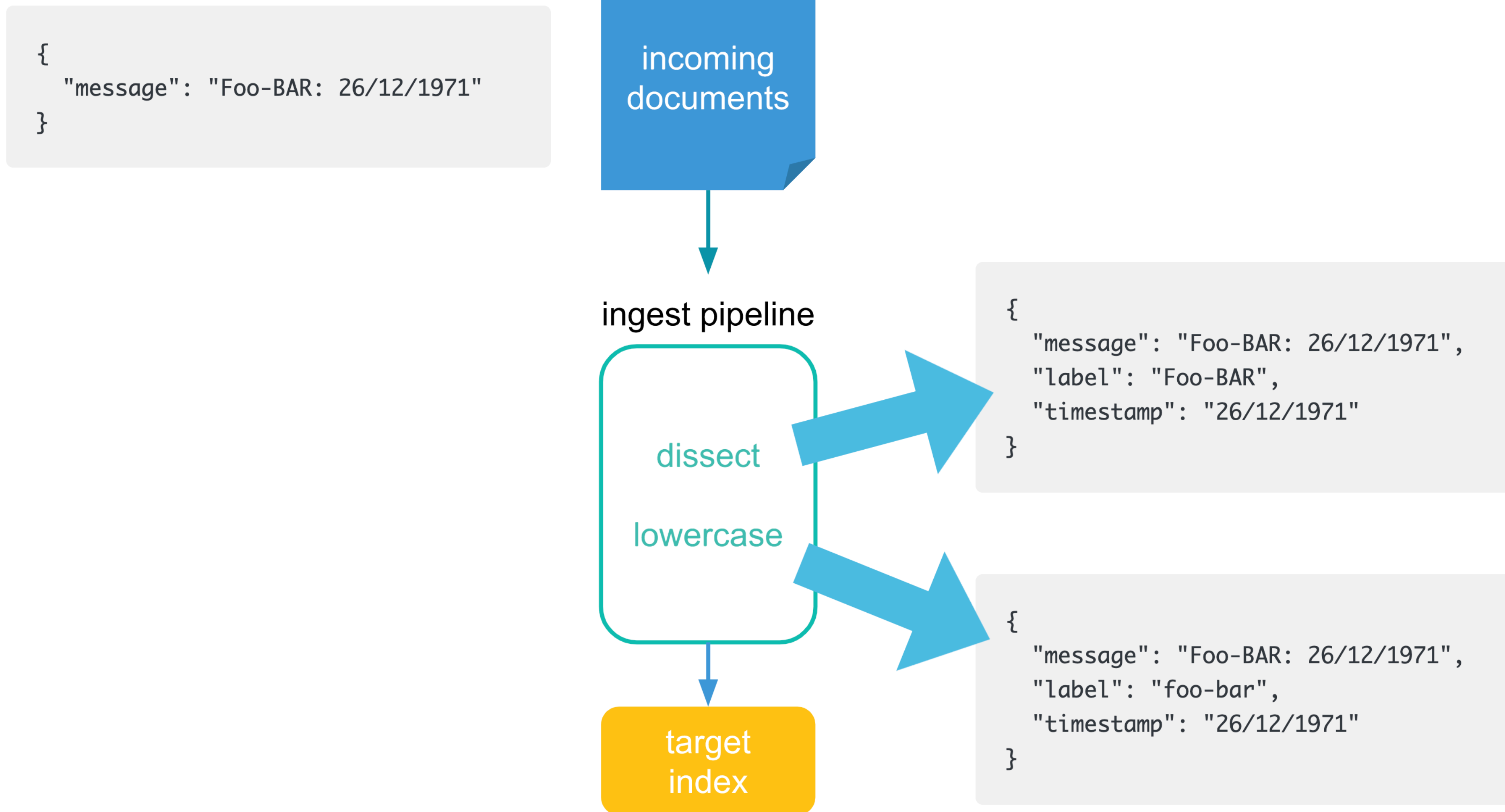
and getting content and metadata

```
static void extractTextAndMetadata(InputStream stream) throws Exception {
    BodyContentHandler handler = new BodyContentHandler();
    Metadata metadata = new Metadata();
    try (stream) {
        new DefaultParser().parse(stream, handler, metadata, new ParseContext());
        String extractedText = handler.toString();
        String title = metadata.get(TikaCoreProperties.TITLE);
        String keywords = metadata.get(TikaCoreProperties.KEYWORDS);
        String author = metadata.get(TikaCoreProperties.CREATOR);
    }
}
```



ingest-attachment plugin extracting from BASE64 or CBOR

An ingest pipeline



ingest-attachment processor plugin

using Tika behind the scene



[Products](#) [Customers](#) [Learn](#) [Company](#) [Pricing](#)

Docs

[Elasticsearch Plugins and Integrations \[8.0\]](#) » [Ingest Plugins](#) » **Ingest Attachment Processor Plugin**

[« Ingest Plugins](#)

[Using the Attachment Processor in a Pipeline »](#)

Ingest Attachment Processor Plugin



The ingest attachment plugin lets Elasticsearch extract file attachments in common formats (such as PPT, XLS, and PDF) by using the Apache text extraction library [Tika](#).

You can use the ingest attachment plugin as a replacement for the mapper attachment plugin.

The source field must be a base64 encoded binary. If you do not want to incur the overhead of converting back and forth between base64, you can use the CBOR format instead of JSON and specify the field as a bytes array instead of a string representation. The processor will skip the base64 decoding then.

Installation



This plugin can be installed using the plugin manager:

```
sudo bin/elasticsearch-plugin install ingest-attachment
```

The plugin must be installed on every node in the cluster, and each node must be restarted after installation.

Demo



<https://cloud.elastic.co>



FSCrawler

You know, for files...





Search or jump to...

Pull requests Issues Marketplace Explore



dadoonet / fscrawler Public

Unpin Unwatch 75 Fork 250 Starred 1k

Code Issues 109 Pull requests 13 Discussions Actions Projects 2 Security Insights Settings

master 21 branches 22 tags

Go to file Add file Code

mergify[bot] Merge pull request #1393 from dadoonet/dependabot... 27b054f 7 days ago 1,918 commits

.github	Change milestone to 2.10	2 months ago
.mvn	Move to .mvn folder all needed settings to build/test FSCrawler	5 years ago
3rdparty	Fixing all tests	13 days ago
beans	Clean up Json util classes	27 days ago
cli	Fix --trace and --debug modes	11 days ago
contrib	Fix Workplace Search Client tests	26 days ago
core	Allow switching between nodes and retry if node is failing	11 days ago
crawler	prepare for next development iteration	2 months ago
distribution	Add curl	11 days ago
docs	Update documentation	11 days ago
elasticsearch-client	Make the error message clearer when failing to connect	11 days ago

About

Elasticsearch File System Crawler (FS Crawler)

fscrawler.readthedocs.io/

java elasticsearch crawler tika

- Readme
- Apache-2.0 License
- Code of conduct
- 1k stars
- 75 watching
- 250 forks

Releases 4

v2.9 Latest 42 seconds ago



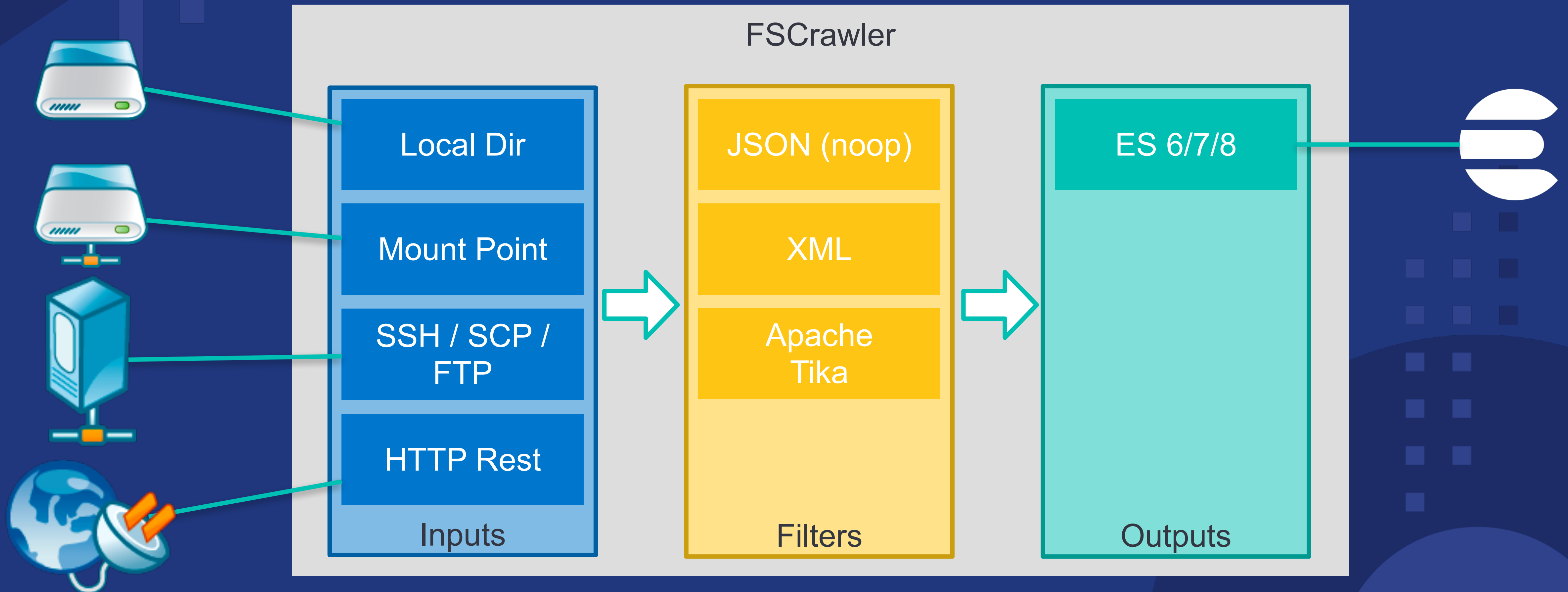
Disclaimer

This project is a community project.
It is not officially supported by Elastic.
Support is only provided by FSCrawler community
on discuss and stackoverflow.

<http://discuss.elastic.co/>
<https://stackoverflow.com/questions/tagged/fscrawler>

FSCrawler

Architecture



FSCrawler

Key Features

- Much more formats than ingest attachment plugin
- OCR (Tesseract)
- Much more metadata than ingest attachment plugin
(See <https://fscrawler.readthedocs.io/en/latest/admin/fs/elasticsearch.html#generated-fields>)
- Language detection

Documentation

- <https://fscrawler.readthedocs.io/>
- <https://fscrawler.readthedocs.io/en/latest/user/tutorial.html>
- <https://fscrawler.readthedocs.io/en/latest/user/formats.html>
- <https://fscrawler.readthedocs.io/en/latest/admin/fs/index.html>



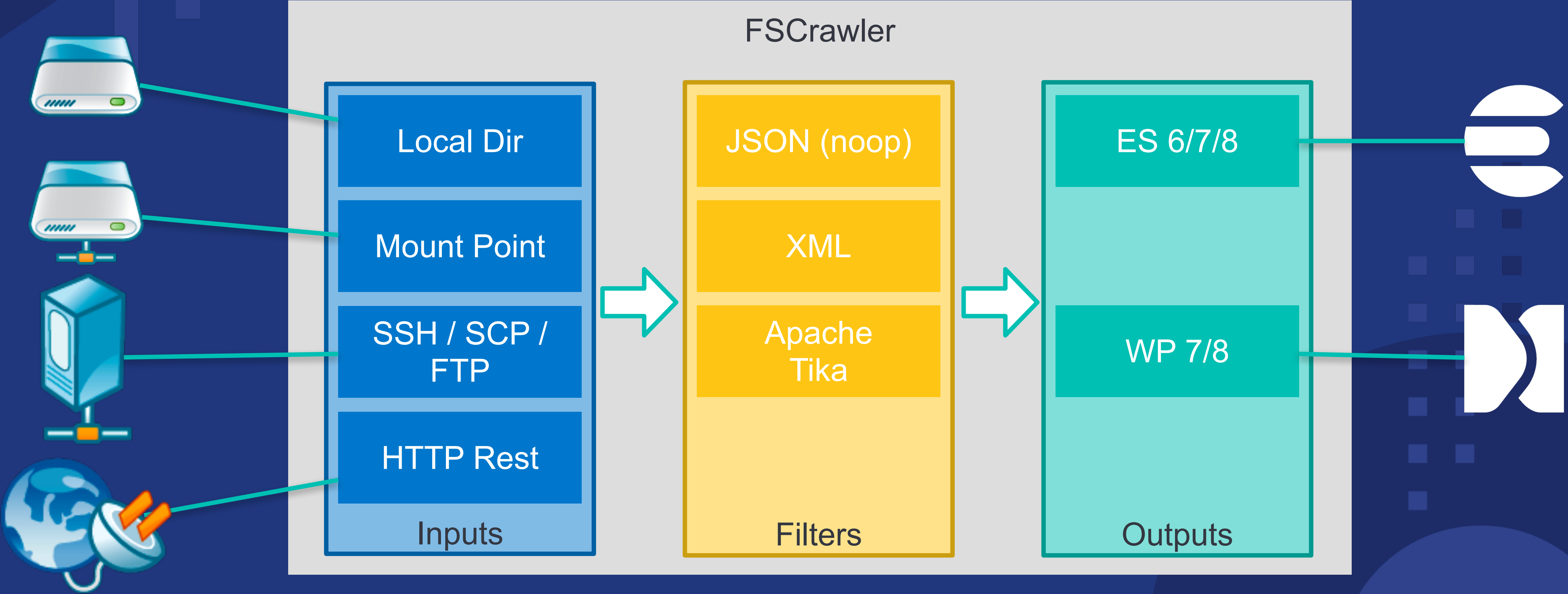
FSCrawler

even better with a UI



FSCrawler

Architecture



Demo



<https://cloud.elastic.co>



Thanks!

PR are warmly welcomed!

<https://github.com/dadoonet/fscrawler>