



Indexing your office documents with Elastic and FSCrawler

David Pilato
Developer / Evangelist, Community
@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

29 March 2021: Apache Tika Release

Apache Tika 1.26 has been released! This release includes improved performance in the ForkParser, improved detection and parsing of XPS files and numerous bug fixes and 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 1.26.

16 January 2021: Apache Tika 2.0.0-ALPHA Release

Apache Tika 2.0.0-ALPHA has been released! This ALPHA release includes a major refactoring of the modules to enable more fine-grained selection of resources, among many other refactorings. Note: There may still be breaking changes before the 2.0.0-BETA and 2.0.0 releases. Please see the [CHANGES.txt](#) file for a list of major changes in the release. Please follow [Migrating to Tika 2.x](#) for updates on migrating to Tika 2.x. See the download page for more information on how to obtain Apache Tika 2.0.0-ALPHA.

30 November 2020: Apache Tika Release

Apache Tika 1.25 has been released! This release includes detection of new file types (parquet, bplist, hprof and flat ODF), new parsers for XLZ, IDML and MIF and flat ODF files, and a critical fix to a license inconsistency in Adobe's xmpcore dependency. This release also includes numerous bug fixes and 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 1.25.

Apache Tika

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

Documentation

- [Apache Tika 1.26](#)
 - [Getting Started](#)
 - [Supported Formats](#)
 - [Parser API](#)
 - [Parser 5min Quick Start Guide](#)
 - [Content and Language Detection](#)
 - [Configuring Tika](#)
 - [Usage Examples](#)
 - [API Documentation](#)
 - [REST API Documentation \(Miredot\)](#)
- [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](#)

The Apache Software Foundation

[About](#)
[License](#)
[Security](#)

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

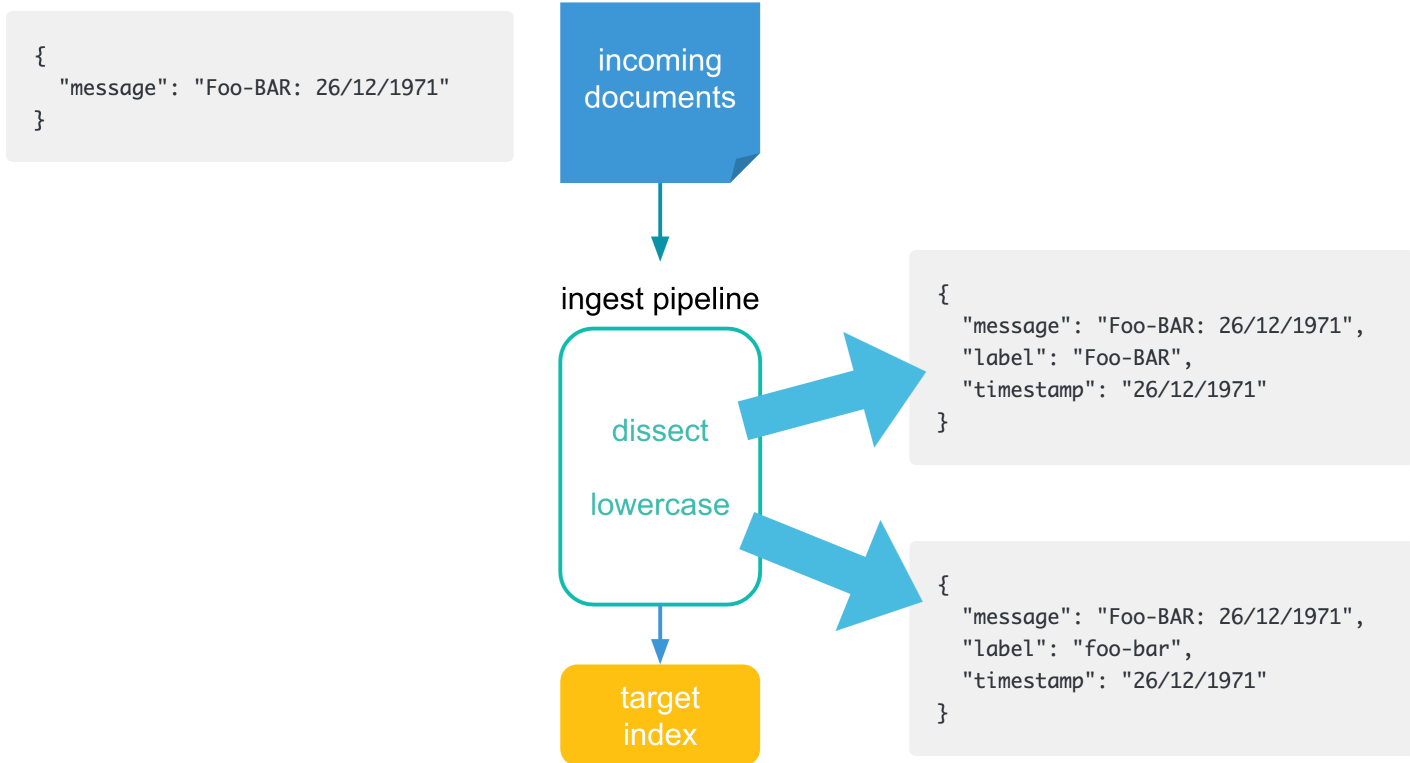
```
static void extractTextAndMetadata(InputStream stream) throws Exception {
    BodyContentHandler handler = new BodyContentHandler();
    Metadata metadata = new Metadata();
    try (stream) {
        new DefaultParser().parse(stream, handler, metadata, new ParseContext());
        String extractedText = 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 \[7.12\]](#) » [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.

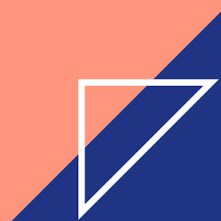
Installation



This plugin can be installed using the plugin manager:

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





Demo



<https://cloud.elastic.co>





FSCrawler

You know, for files...



[Pull requests](#)
[Issues](#)
[Marketplace](#)
[Explore](#)

[dadoonet / fscrawler](#)

Unwatch 76
Unstar 812
Fork 203

[Code](#)
[Issues 102](#)
[Pull requests 8](#)
[Actions](#)
[Projects 2](#)
[Security](#)
[Insights](#)
[Settings](#)

master
15 branches
19 tags

Go to file
Add file
Code

mergify Merge pull request #997 from dadoonet/dependabot/mave...
4669ef2 20 days ago 1,217 commits

.github	Update the issue templates	4 months ago
.mvn	Move to .mvn folder all needed settings to build/test FSCrawler	4 years ago
beans	Add support for YAML configuration	2 years ago
cli	Remove support for Elasticsearch v5	9 months ago
contrib/docker-compose-example	Update Dockerfile-fscrawler	29 days ago
core	Fix SSH crawling from Windows machine	2 months ago
crawler	Add documentation about Windows drives SSH indexing	6 months ago
distribution	Remove support for Elasticsearch v5	9 months ago
docs	Updated documentation for instructions on how to use the contri...	2 months ago
elasticsearch-client	Add `path_prefix` option	6 months ago
framework	Remove support for Elasticsearch v5	9 months ago
integration-tests	Fix flaky tests	2 months ago
rest	Add more information to the _simulate API	9 months ago
settings	Document `auto` option for `pdf_strategy`	3 months ago
src/main/resources/org/apache/...	Have tests for ES5 and ES6 in the same repo (no more profiles)	2 years ago
test-documents	Document `auto` option for `pdf strategy`	3 months ago

About

Elasticsearch File System Crawler (FS Crawler)

[fscrawler.readthedocs.io/](#)

[java](#)
[elasticsearch](#)
[crawler](#)
[tika](#)

Readme
Apache-2.0 License

Releases 19

FSCrawler 2.6 Latest
on 9 Jan 2019

[+ 18 releases](#)

Packages

No packages published

[Publish your first package](#)

Used by 7



Disclaimer

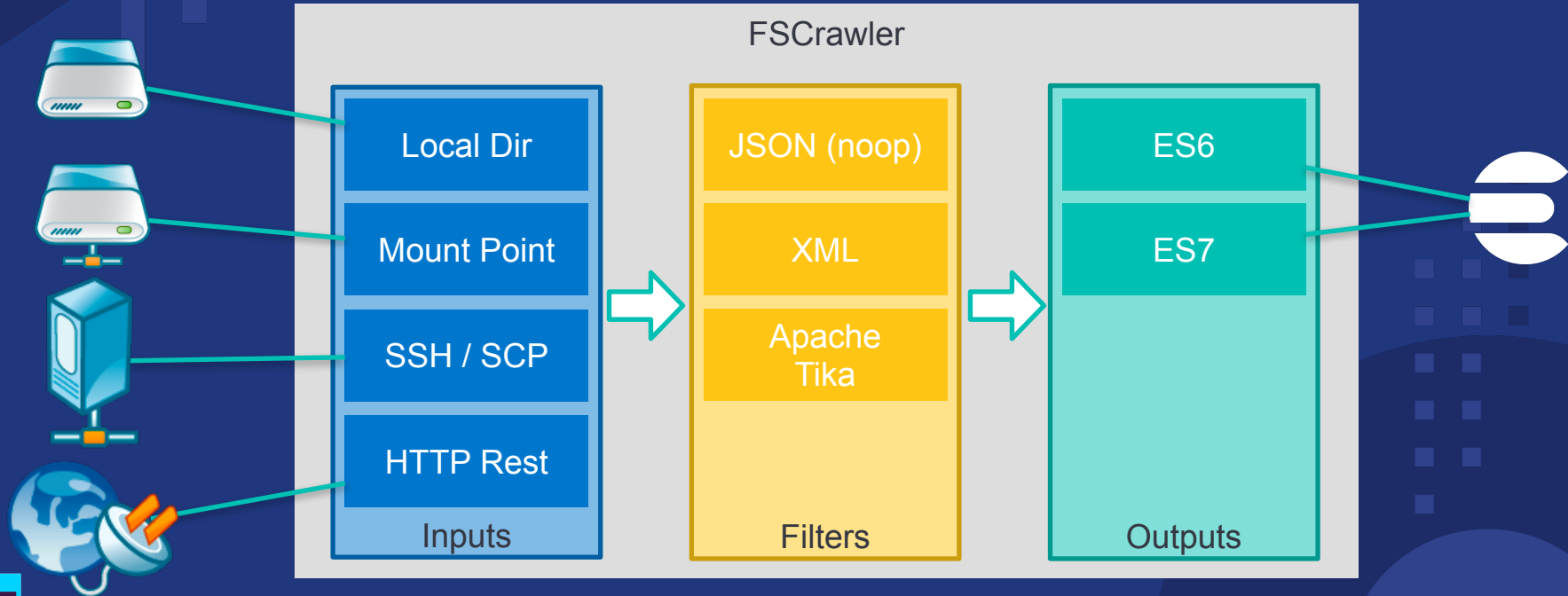
This project is a community project.
It is not officially supported by Elastic.
Support is only provided by FSCrawler community
on discuss and stackoverflow.

<http://discuss.elastic.co/>

<https://stackoverflow.com/questions/tagged/fscrawler>

FSCrawler

Architecture



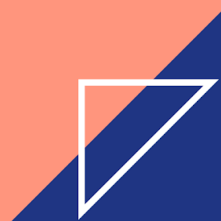
FSCrawler

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>



Demo



<https://cloud.elastic.co>





FSCrawler

even better with a UI



FSCrawler

Workplace Search integration

Add Workplace Search connector #991

 Merged  dadoonet merged 82 commits into `master` from `wip/workplace_search` on 22 Dec 2020

 Conversation **3**

 Commits **82**

 Checks **5**

 Files changed **106**



dadoonet commented on 30 Jul 2020 · edited ▾

Owner 😊 ⋮

This PR adds a connector to Workplace Search.

Setup

Full documentation available at: https://fscrawler.readthedocs.io/en/wip-workplace_search/admin/fs/wpsearch.html

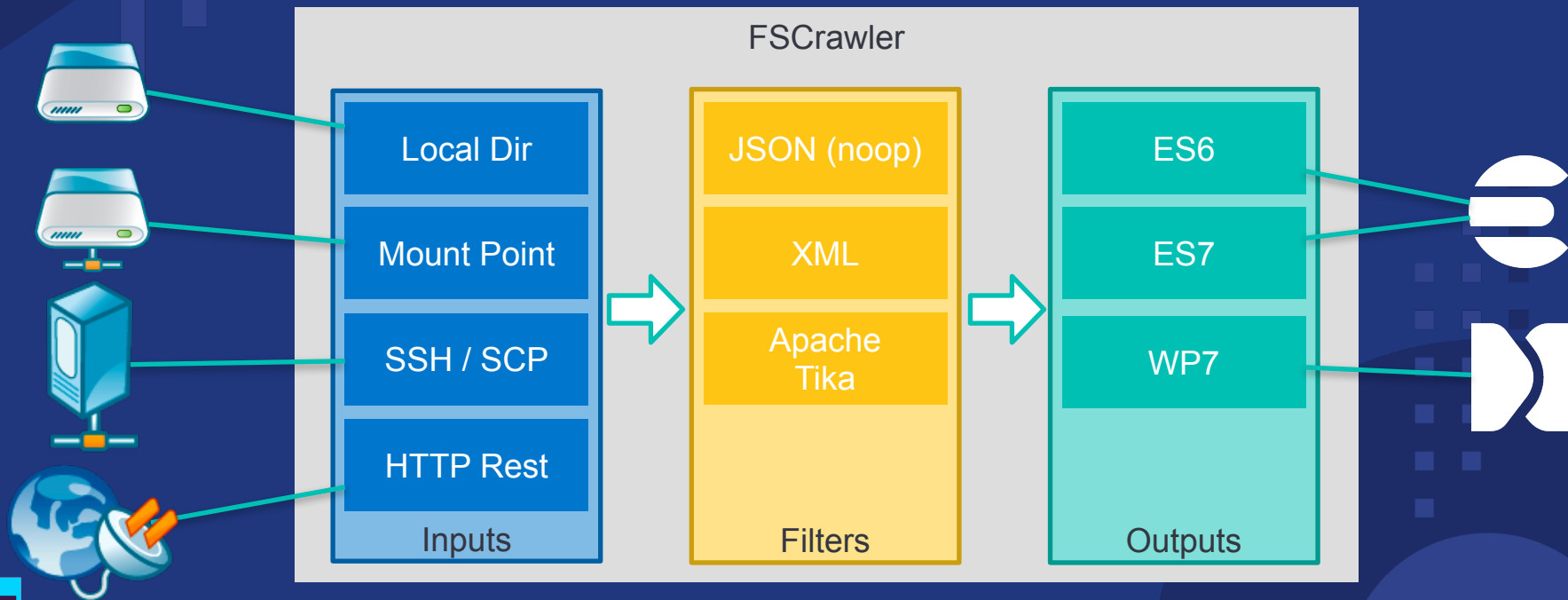
Keys

Once you have created your Custom API and have the `ACCESS_TOKEN` and `KEY`, you can add to your existing FSCrawler configuration file:

```
name: "test"
workplace_search:
```

FSCrawler

Architecture



OCR integration
Starting with a REST gateway
Supported formats
Tips and tricks

ADMINISTRATION GUIDE

Status files
CLI options
JVM Settings
Configuring an external logger
configuration file
Job file specification
The most simple crawler
Local FS settings
SSH settings
Elasticsearch settings

Workplace Search settings

Keys
Server
Running on Cloud
Bulk settings
Documents Repository URL

REST service

DEVELOPER GUIDE

Building the project
Writing documentation
Release the project



Read the Docs [v: wip/workplace_search](#)

Workplace Search settings

New in version 2.7.

FSCrawler can now send documents to [Workplace Search](#).

Note

Although this won't be needed in the future, it is still mandatory to have access to the elasticsearch instance running behind Workplace Search. In this section of the documentation, we will only cover the specifics for workplace search. Please refer to [Elasticsearch settings](#) chapter.

Hint

To easily start locally with Workplace Search, follow the steps:

- Check-out the source code on [GitHub](#):

```
git clone git@github.com:dadoonet/fscrawler.git
cd fscrawler
cd contrib/docker-compose-workplacesearch
docker-compose up
```

This will start Elasticsearch, Kibana (not used) and Workplace Search.

- Wait for it to start and open <http://127.0.0.1:3002/ws>.
- Enter `enterprise_search` as the login and `changeme` as the password.
- Click on "Add sources" button and choose [Custom API](#).
- Name it `fscrawler` and click on "Create Custom API Source" button.
- Copy the "Access Token" value. We will mention it as `ACCESS_TOKEN` for the rest of this documentation.
- Copy the "Key" value. We will mention it as `KEY` for the rest of this documentation.

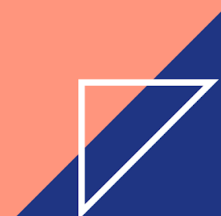


[← Back to Sources](#)

Create a Custom
API Source



Custom API Source
API, Custom



Demo



<https://cloud.elastic.co>





Thanks!

PR are warmly welcomed!

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