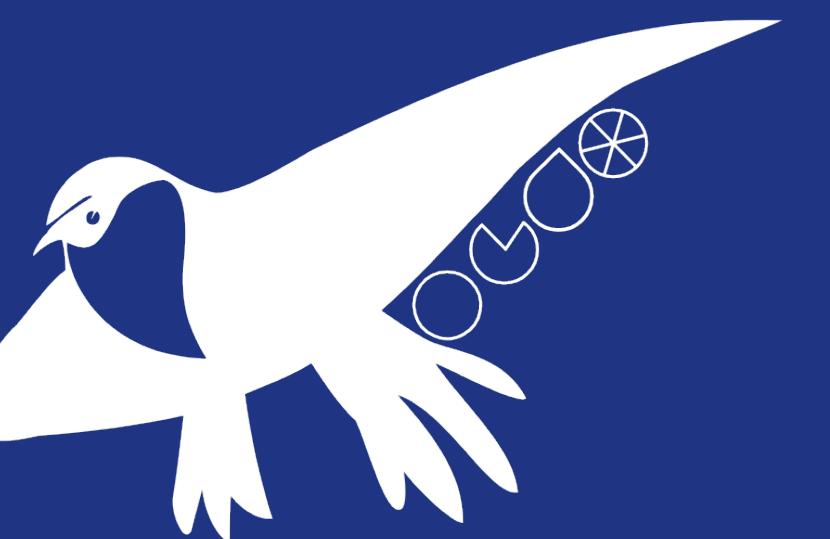




Indexer ses documents bureautique avec la suite Elastic et FSCrawler

David Pilato
Developer / Evangelist, Community
@dadoonet



Please run step 0
before we start!



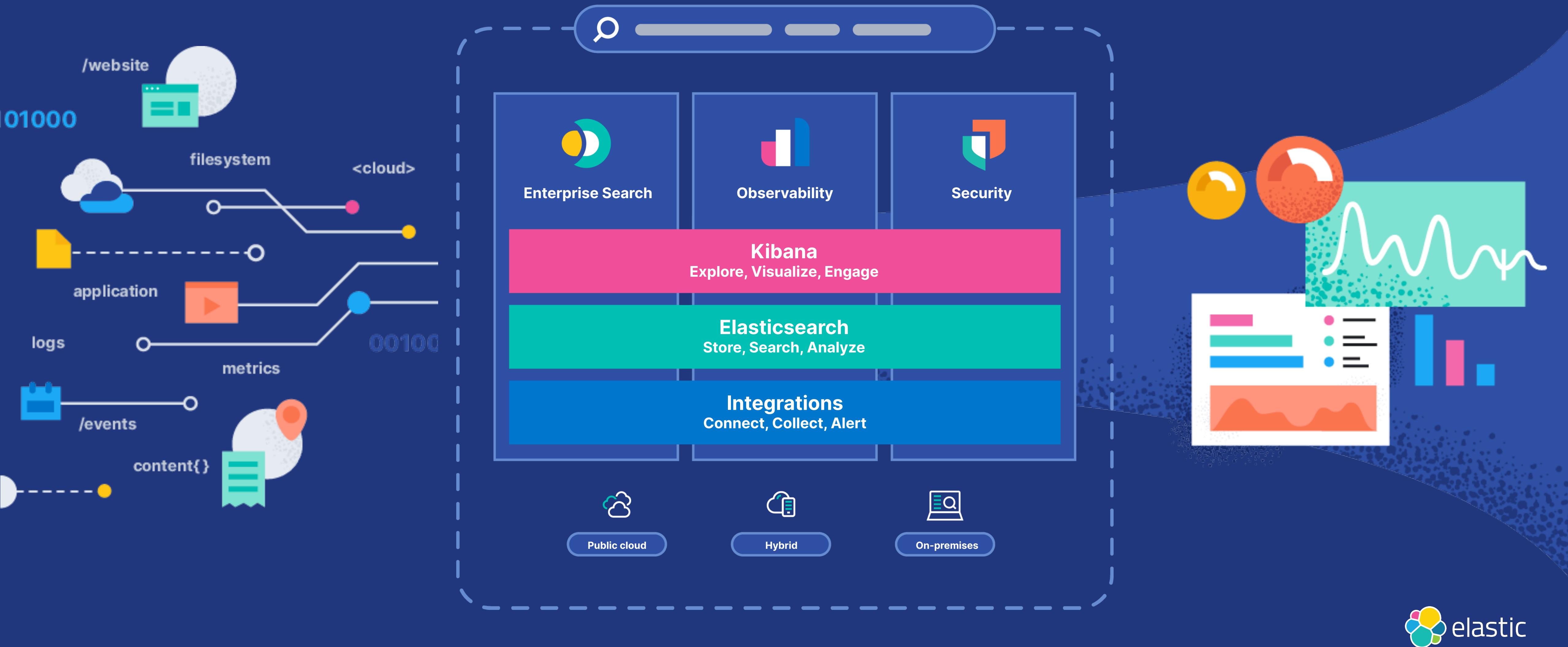
<https://github.com/dadoonet/JDLL>

Lab 0 setup



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

The Elastic Search Platform





Lab 1

indexing json documents



**ingest-attachment processor
extracting from
BASE64 or CBOR**



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.

For more in-depth documentation, see our [wiki](#), especially for [tika-server](#).

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

2 May 2022: Apache Tika Release

Apache Tika 2.4.0 has been released! This release includes new mime detection for http-responses, frictionless data packages, DGN files and others. Added basic parsers for WARC and WACZ. Added configuration for metadata write filters, custom content handler decorators and bus JARs for standard providers.

Apache Tika

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

Documentation

- [Apache Tika 2.4.0](#)
 - Getting Started
 - Supported Formats
 - Parser API
 - Parser 5min Quick Start Guide
 - Content and Language Detection
 - Configuring Tika
 - Usage Examples
 - API Documentation
- [Apache Tika 1.28.2](#)
- [Apache Tika 2.3.0](#)
- [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](#)
- [Apache Tika 1.23](#)
- [Apache Tika 1.22](#)
- [Apache Tika 1.21](#)
- [Apache Tika 1.20](#)
- [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](#)
- [Apache Tika 1.0](#)

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.

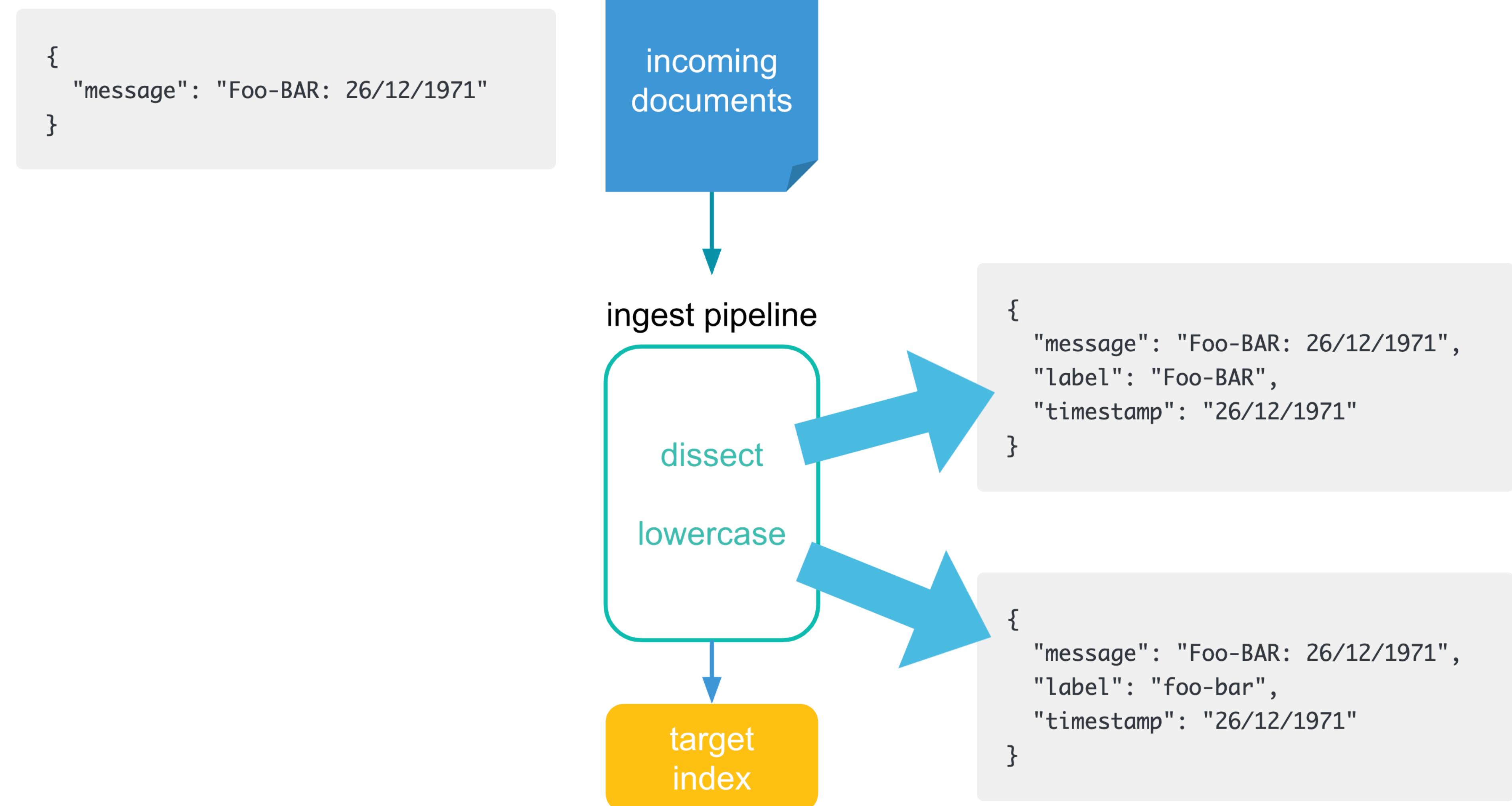
- [Supported Document Formats](#)
 - [HyperText Markup Language](#)
 - [XML and derived formats](#)
 - [Microsoft Office document formats](#)
 - [OpenDocument Format](#)
 - [iWorks document formats](#)
 - [WordPerfect document formats](#)
 - [Portable Document Format](#)
 - [Electronic Publication Format](#)
 - [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);  
    }  
}
```

An ingest pipeline



ingest-attachment processor

using Tika behind the scene

Ingest pipelines
Example: Parse logs
Enrich your data
Processor reference
Append
Attachment
Bytes
Circle
Community ID
Convert
CSV
Date
Date index name
Dissect
Dot expander
Drop
Enrich
Fail
Fingerprint
Foreach

Example

If attaching files to JSON documents, you must first encode the file as a base64 string. On Unix-like systems, you can do this using a `base64` command:

```
base64 -in myfile.rtf
```

The command returns the base64-encoded string for the file. The following base64 string is for an `.rtf` file containing the text `Lorem ipsum dolor sit amet`:

```
e1xydGYxXGFuc2kNCkxvcmVtIGlwc3VtIGRvbG9yIHNpdCBhbWV0DQpccGFyIH0=.
```

Use an attachment processor to decode the string and extract the file's properties:

```
PUT _ingest/pipeline/attachment
{
  "description" : "Extract attachment information",
  "processors" : [
    {
      "attachment" : {
        "field" : "data",
        "remove_binary": false
      }
    }
  ]
}
PUT my-index-000001/_doc/my_id?pipeline=attachment
{
  "data": "e1xydGYxXGFuc2kNCkxvcmVtIGlwc3VtIGRvbG9yIHNpdCBhbWV0DQpccGFyIH0=.
```



On this page

[Using the attachment processor in a pipeline](#)

Example

[Exported fields](#)

[Use the attachment processor with CBOR](#)

[Limit the number of extracted chars](#)

[Using the attachment processor with arrays](#)

Most Popular

VIDEO
[Get Started with Elasticsearch](#)

VIDEO
[Intro to Kibana](#)

VIDEO
[ELK for Logs & Metrics](#)



Lab 2

ingest attachment



FSCrawler
You know, for files...

 Search or jump to... / Pull requests Issues Marketplace Explore

[dadoonet/fscrawler](#) Public

Unpin Unwatch 74 Fork 263 Starred 1.1k

Code Issues 112 Pull requests 11 Discussions Actions Projects 2 Security 1 Insights Settings

⚠ We found potential security vulnerabilities in your dependencies.

Only the owner of this repository can see this message. See Dependabot alerts

master ▾ 17 branches 22 tags Go to file Add file ▾ Code ▾

 **dadoonet** Merge pull request #1428 from dadoonet/remove-waitfor ✓ 453aa80 18 hours ago 1,978 commits

 .github	Fix tests for 6.8	2 months ago
 .mvn	Move to .mvn folder all needed settings to build/test FSCrawler	5 years ago
 3rdparty	Revert "Add the waitfor maven plugin"	18 hours ago
 beans	Clean up Json util classes	3 months ago
 cli	Fix --trace and --debug modes	3 months ago
 contrib	Update to 8.1.1	2 months ago
 core	Allow switching between nodes and retry if node is failing	3 months ago
 crawler	prepare for next development iteration	4 months ago
 distribution	Merge pull request #1389 from rhaist/patch-1	2 months ago
 docs	Upgrade to Tika 2.4.0	2 days ago
 elasticsearch-client	Update waitfor-maven-plugin to 1.4-SNAPSHOT	2 months ago
 framework	Fix unit tests	2 months ago

About 

Elasticsearch File System Crawler (FS Crawler)

[fscrawler.readthedocs.io/](#)

java elasticsearch crawler tika

Readme Apache-2.0 license Code of conduct 1.1k stars 74 watching 263 forks

Releases 4

v2.9 Latest on 8 Mar + 3 releases

Packages

No packages published Publish your first package



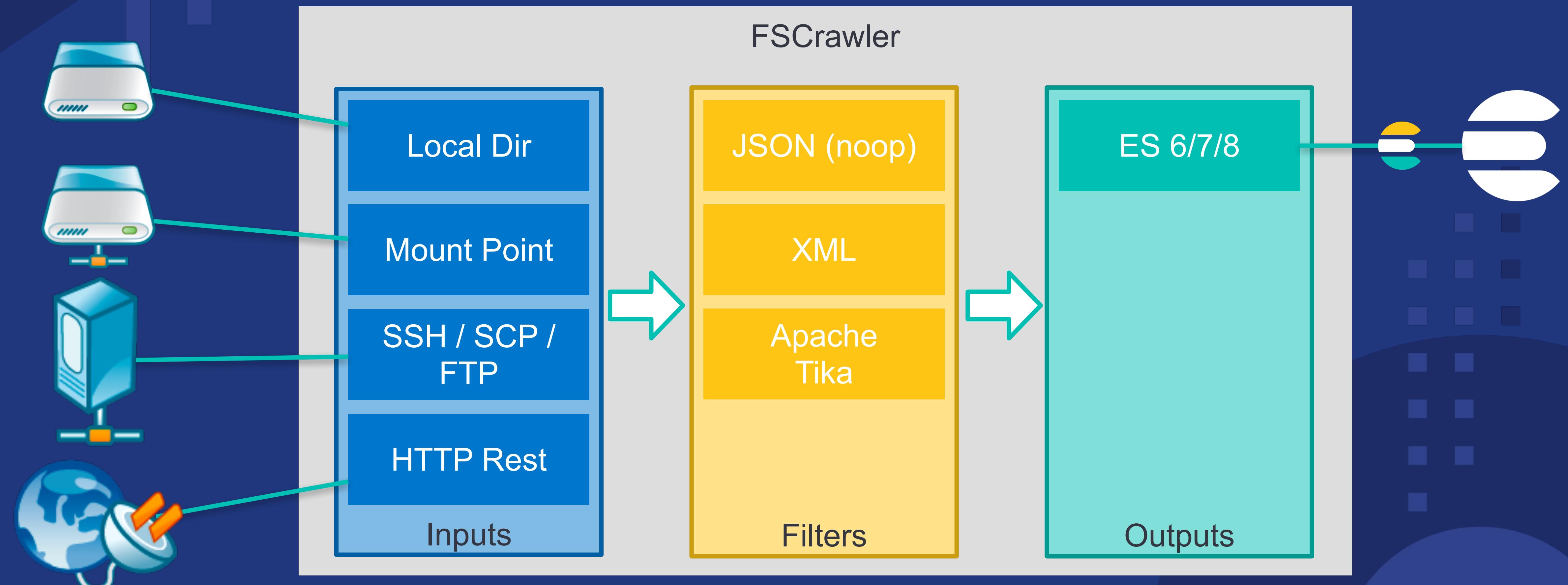
Disclaimer

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





Lab 3

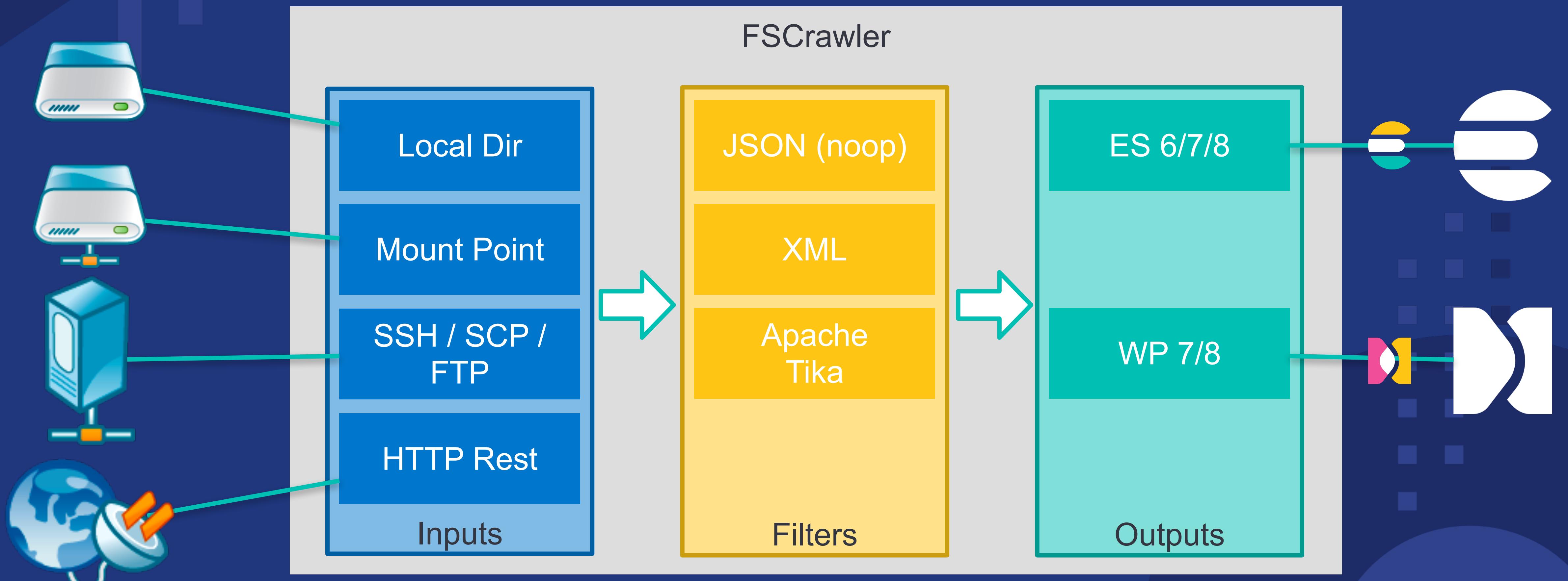
fscrawler



FSCrawler
even better with a UI

FSCrawler

Architecture





Lab 4

workplace search

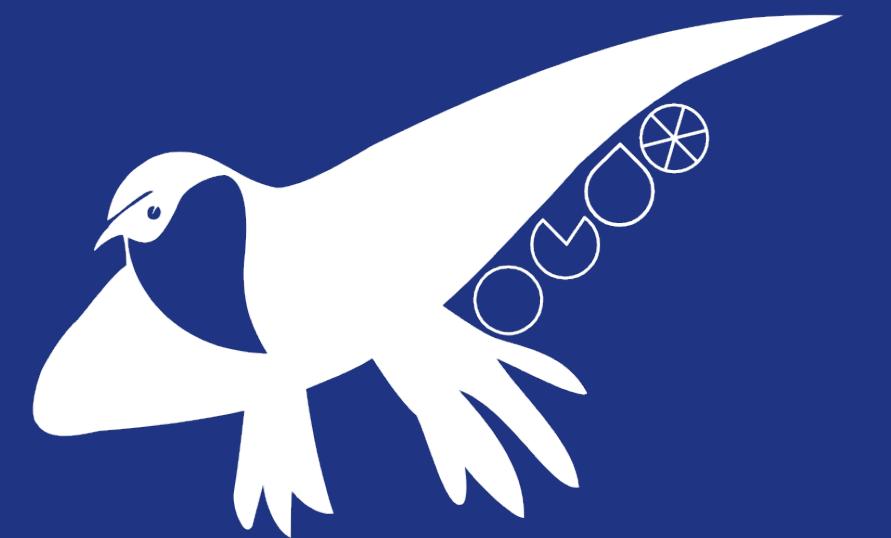
Since
8.2

Network drives connector package

for Enterprise Search

<https://github.com/elastic/enterprise-search-network-drives-connector/>





Thanks!

PR are warmly welcomed!

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