Devoxx France 2023

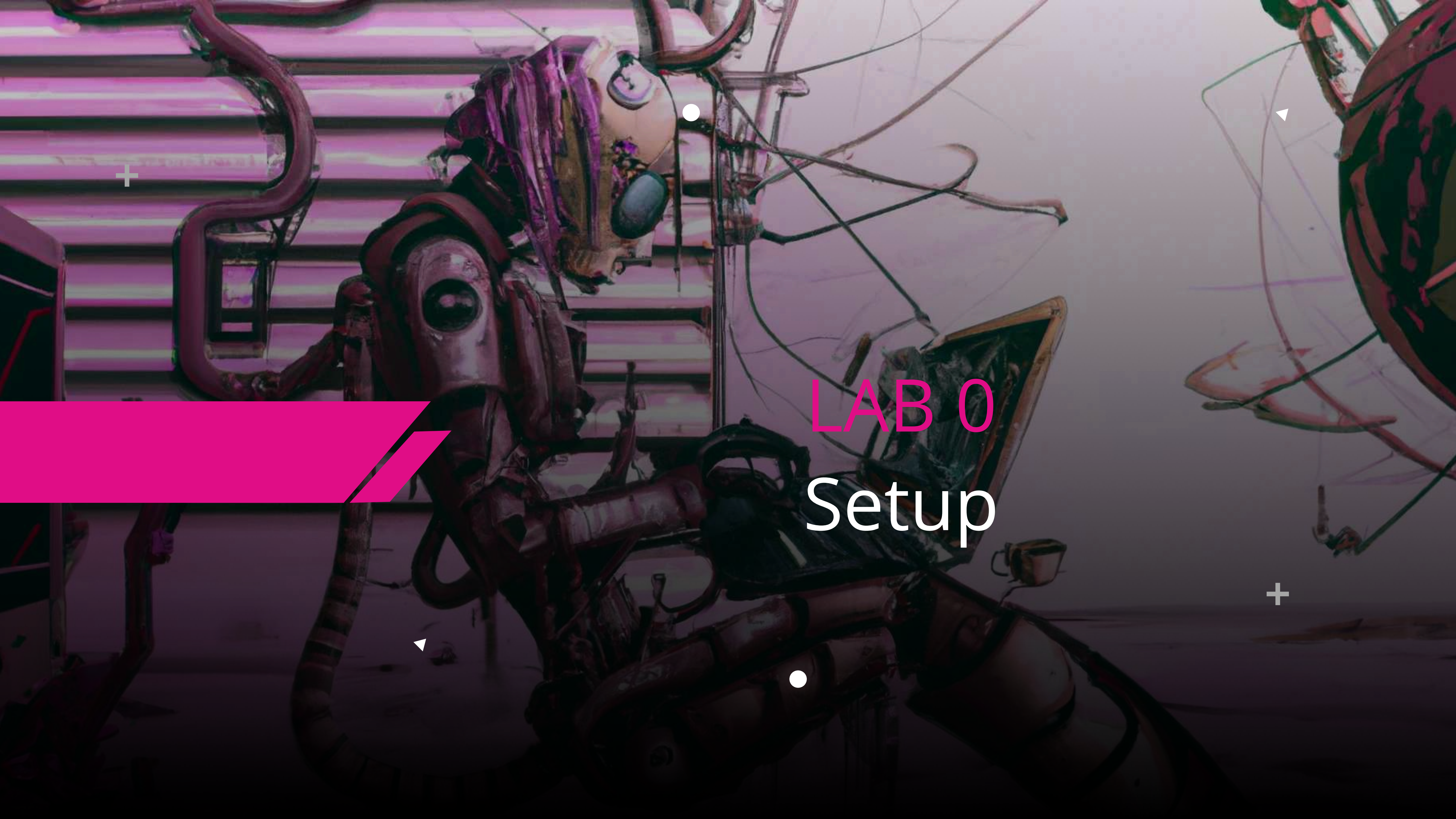
Please run step 0 before we start!

elastic

Un moteur de recherche de documents d'entreprise

Maha ALSAYASNEH (@MahaALSayasneh)
David PILATO (@dadoonet)

https://github.com/dadoonet/DevoxxFR-2023



LAB 0

Setup

Devoxx France 2023

SUMMARY

Elasticsearch
Basics

Ingest Pipelines
Modify documents on the fly

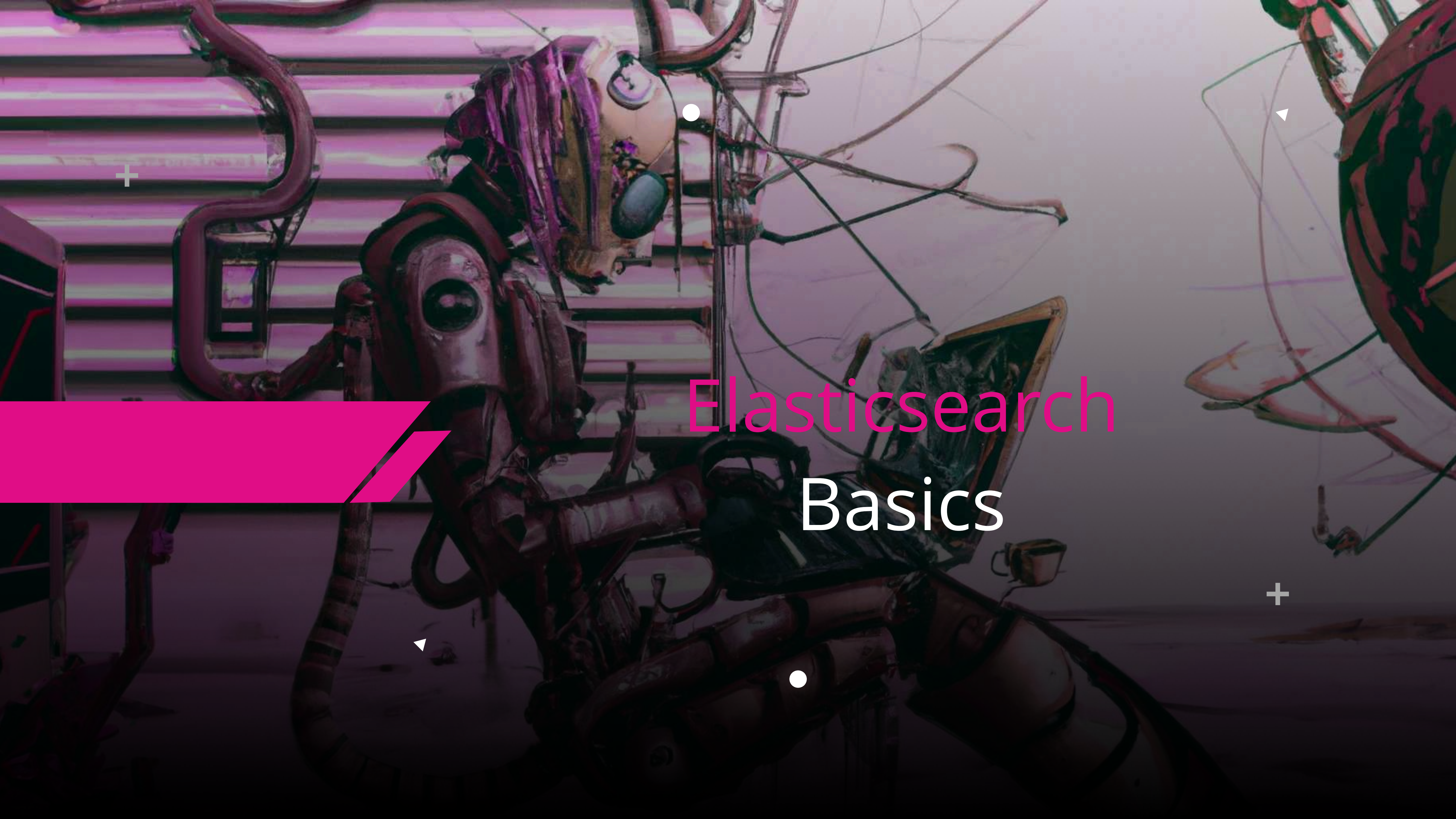
Apache Tika
& Ingest Attachment Processor

AI on Files
Infererence Processor

FSCrawler
You know, for files…

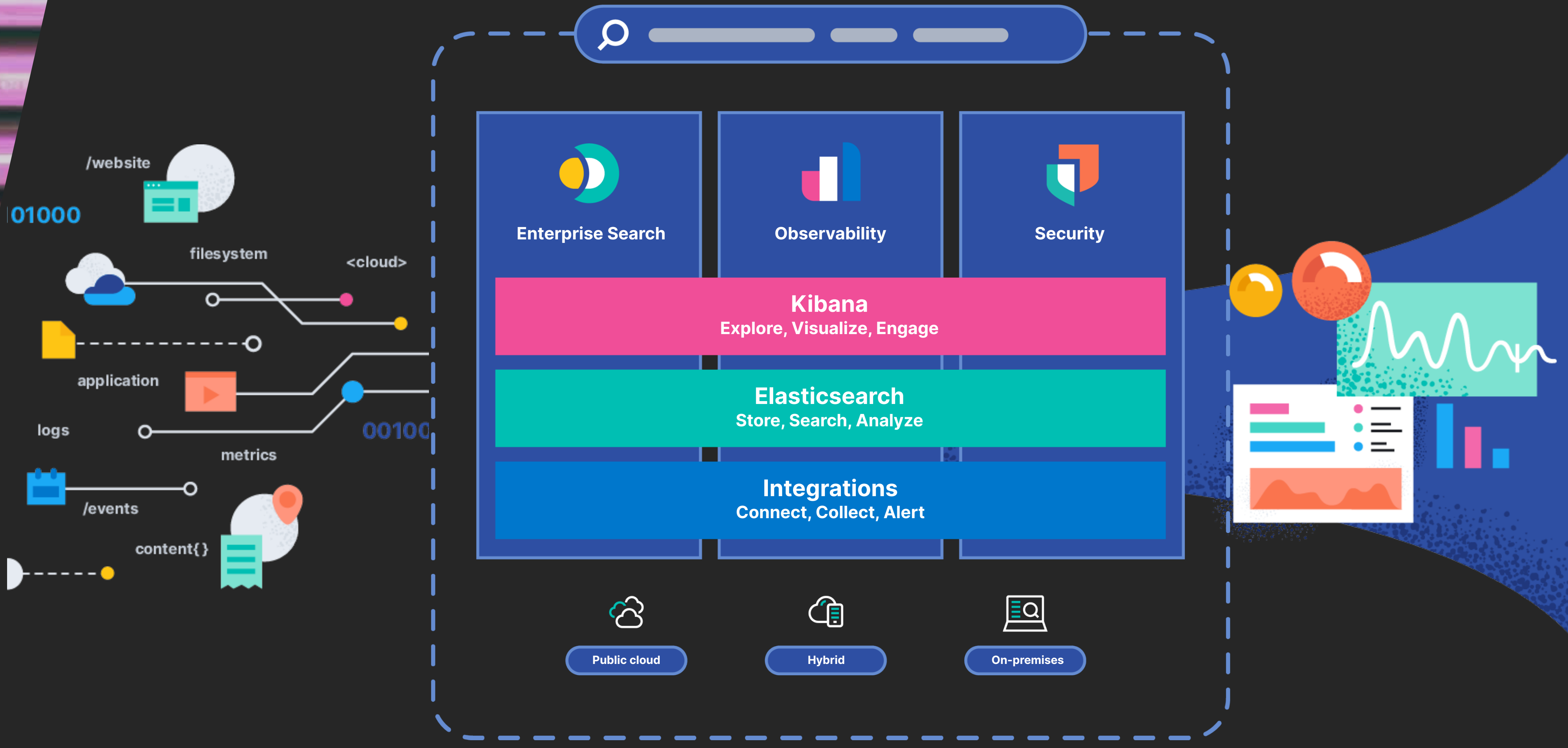
Workplace Search
Even better with a UI

THANKS



Elasticsearch

Basics



Enterprise Search

Observability

Security

Kibana
Explore, Visualize, Engage

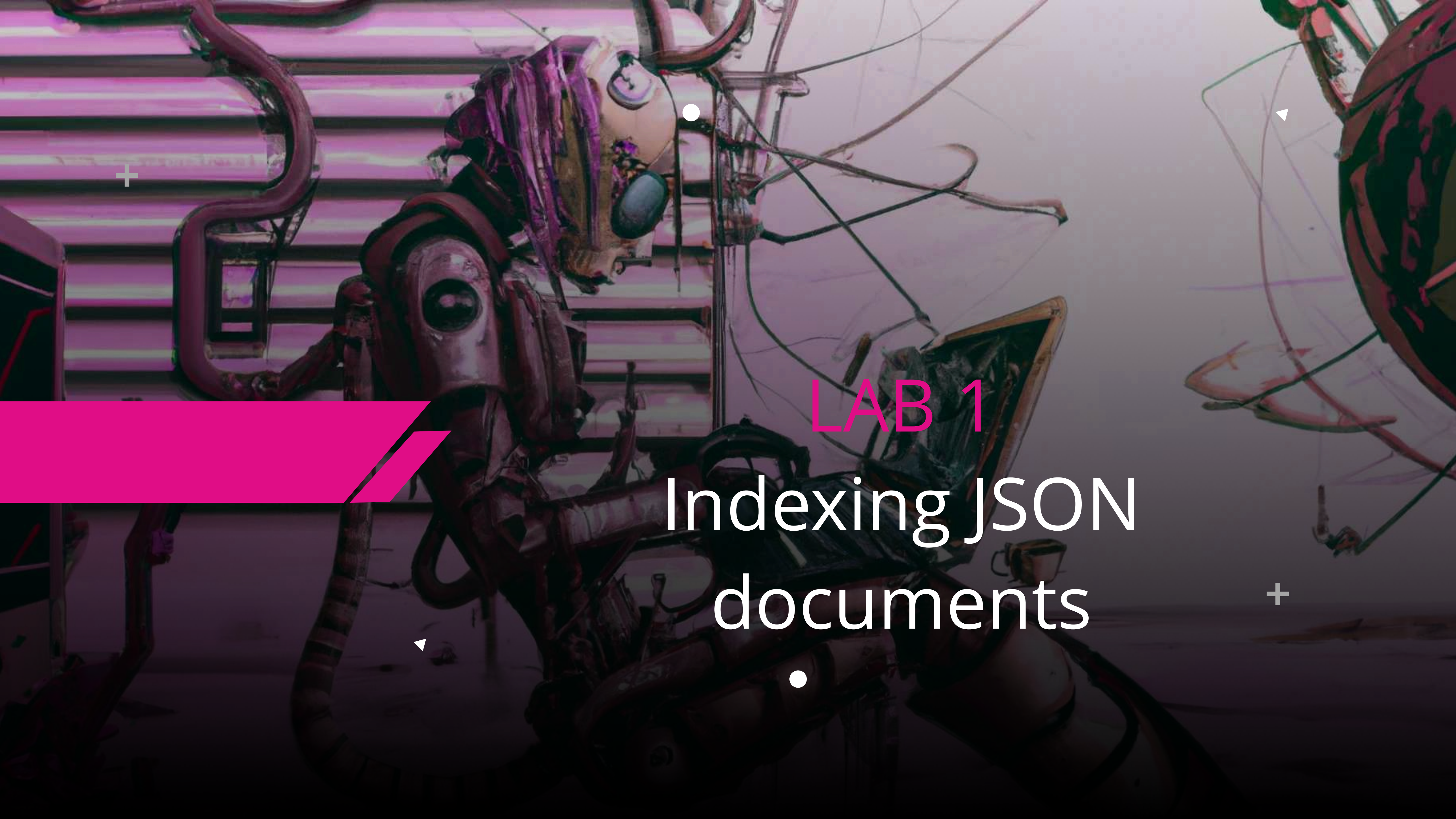
Elasticsearch
Store, Search, Analyze

Integrations
Connect, Collect, Alert

Public cloud

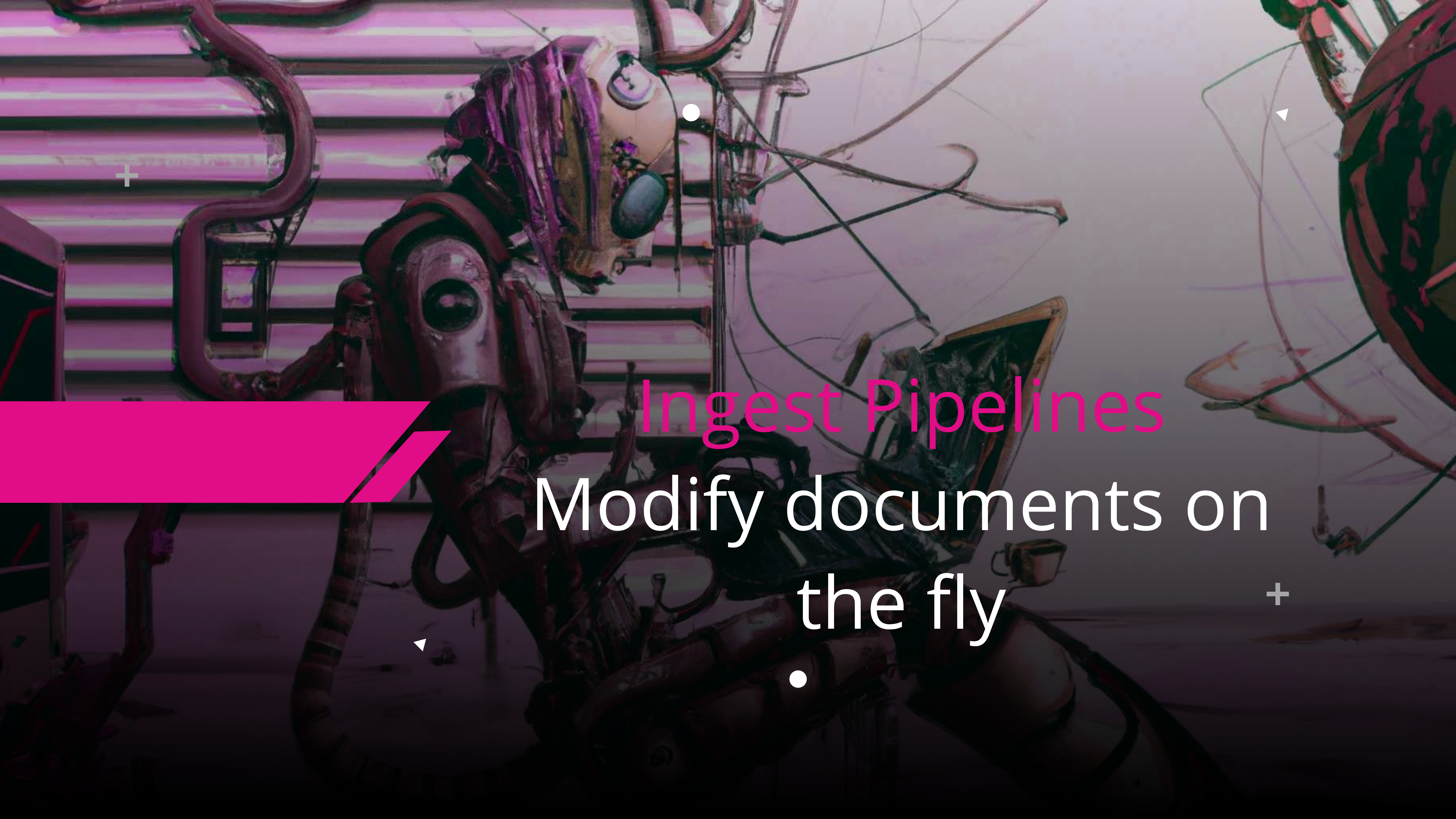
Hybrid

On-premises



LAB 1

Indexing JSON documents

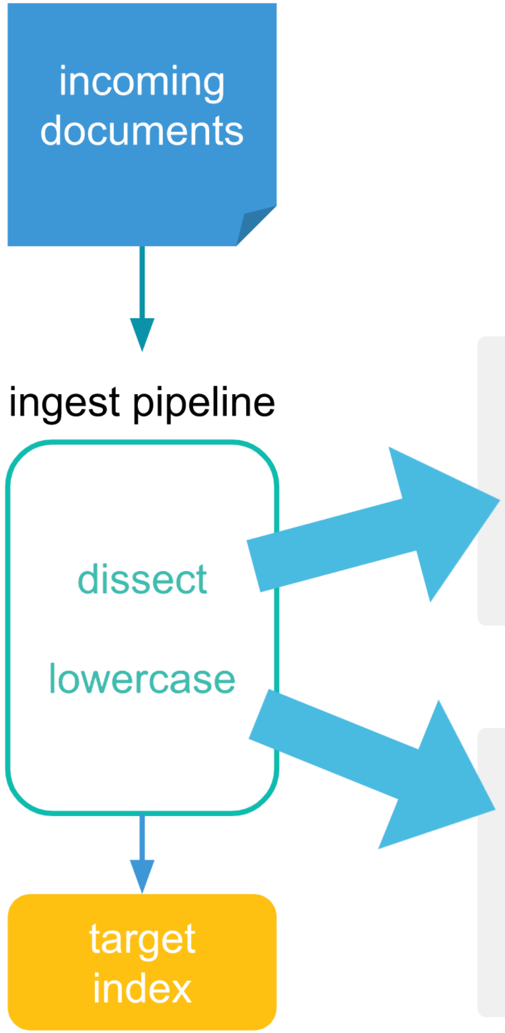


Ingest Pipelines

Modify documents on the fly

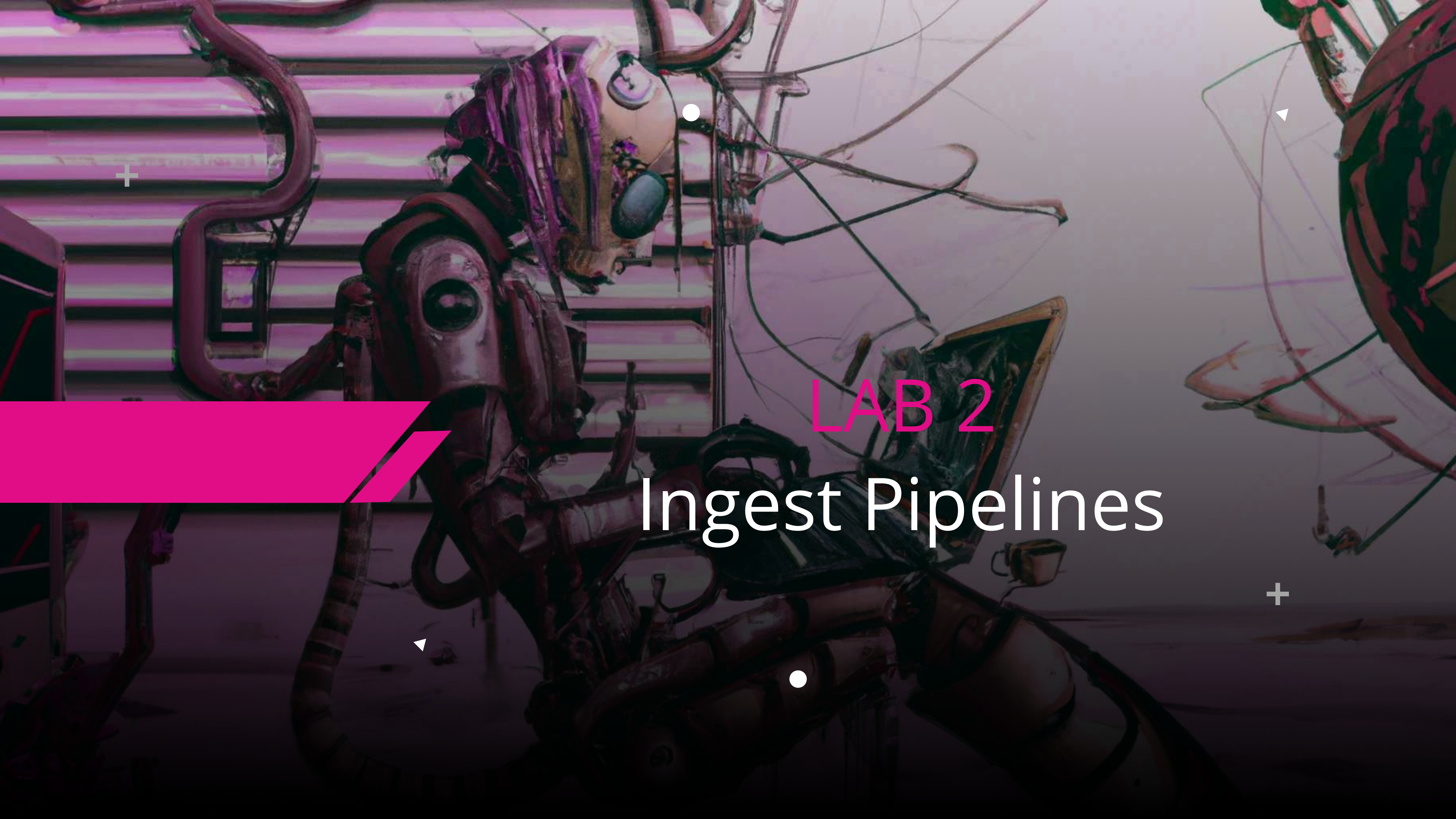
an ingest pipeline

```
{
  "message": "Foo-BAR: 26/12/1971"
}
```



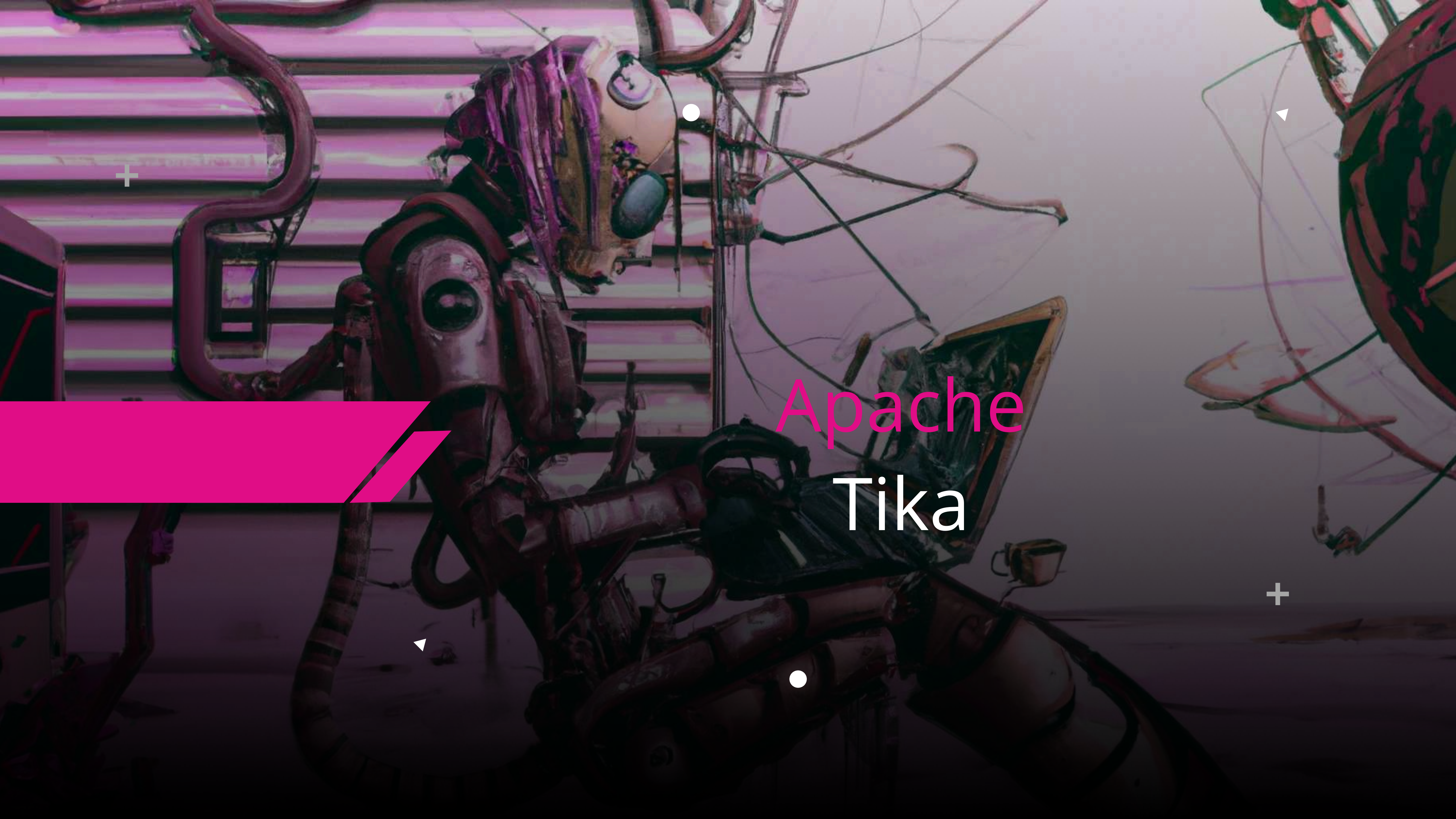
```
{
  "message": "Foo-BAR: 26/12/1971",
  "label": "Foo-BAR",
  "timestamp": "26/12/1971"
}
```

```
{
  "message": "Foo-BAR: 26/12/1971",
  "label": "foo-bar",
  "timestamp": "26/12/1971"
}
```

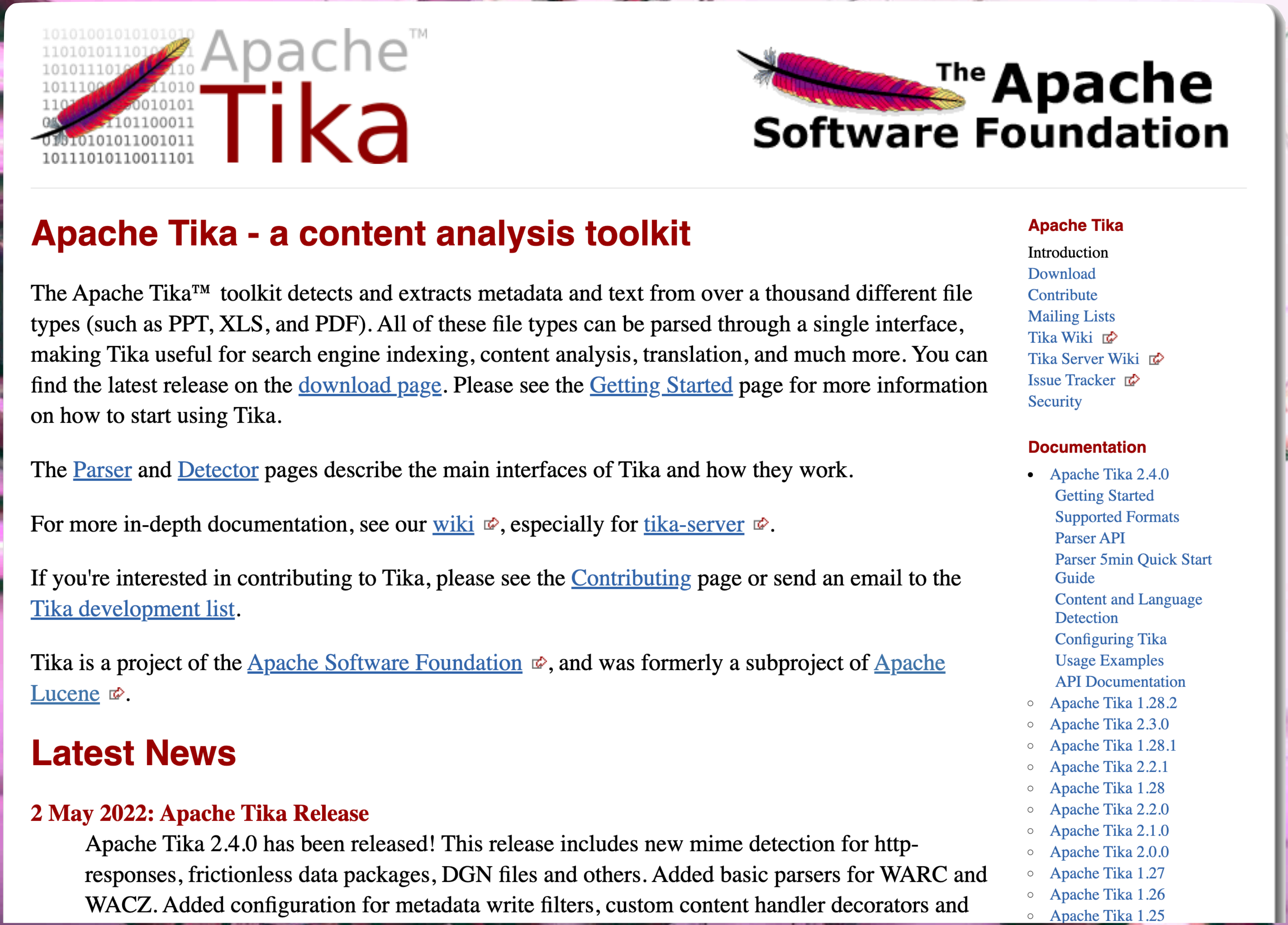


LAB 2

Ingest Pipelines



Apache Tika



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

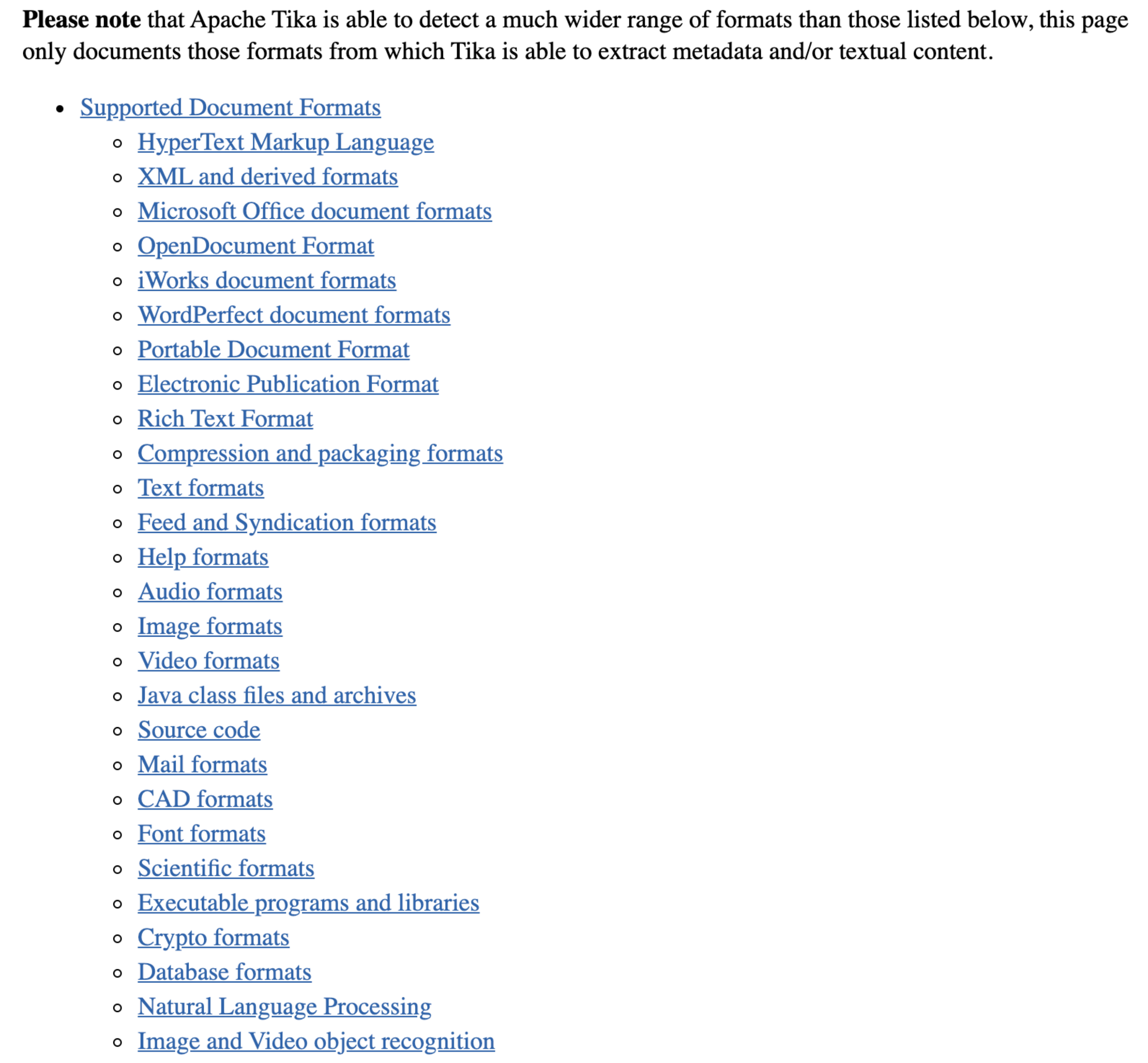
Apache Tika 2.4.0 has been released! This release includes new mime detection for http-responses, frictionless data packages, DGN files and others. Added basic parsers for WARC and WACZ. Added configuration for metadata write filters, custom content handler decorators and

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



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

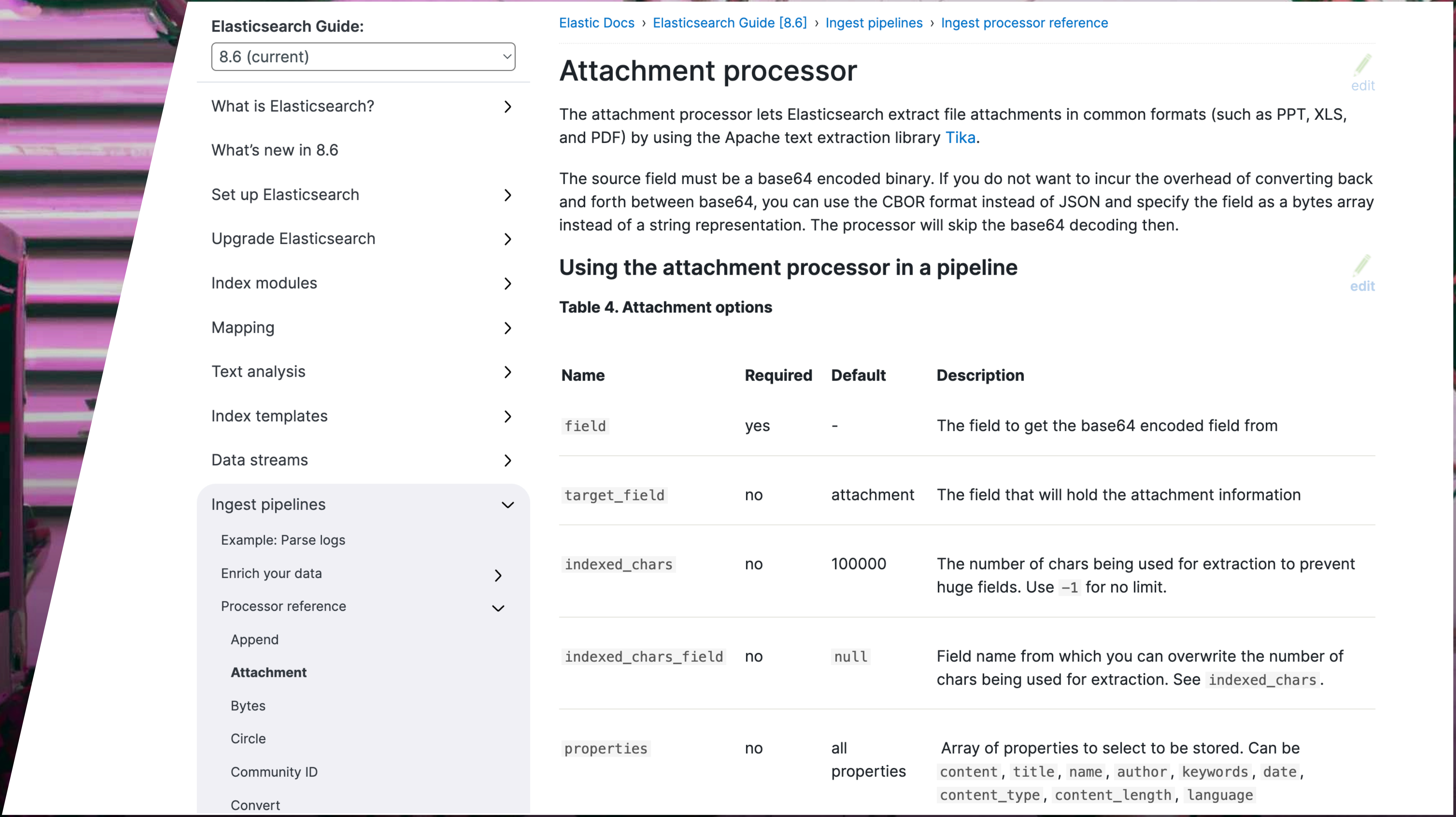
CODE

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 Processor



Attachment processor

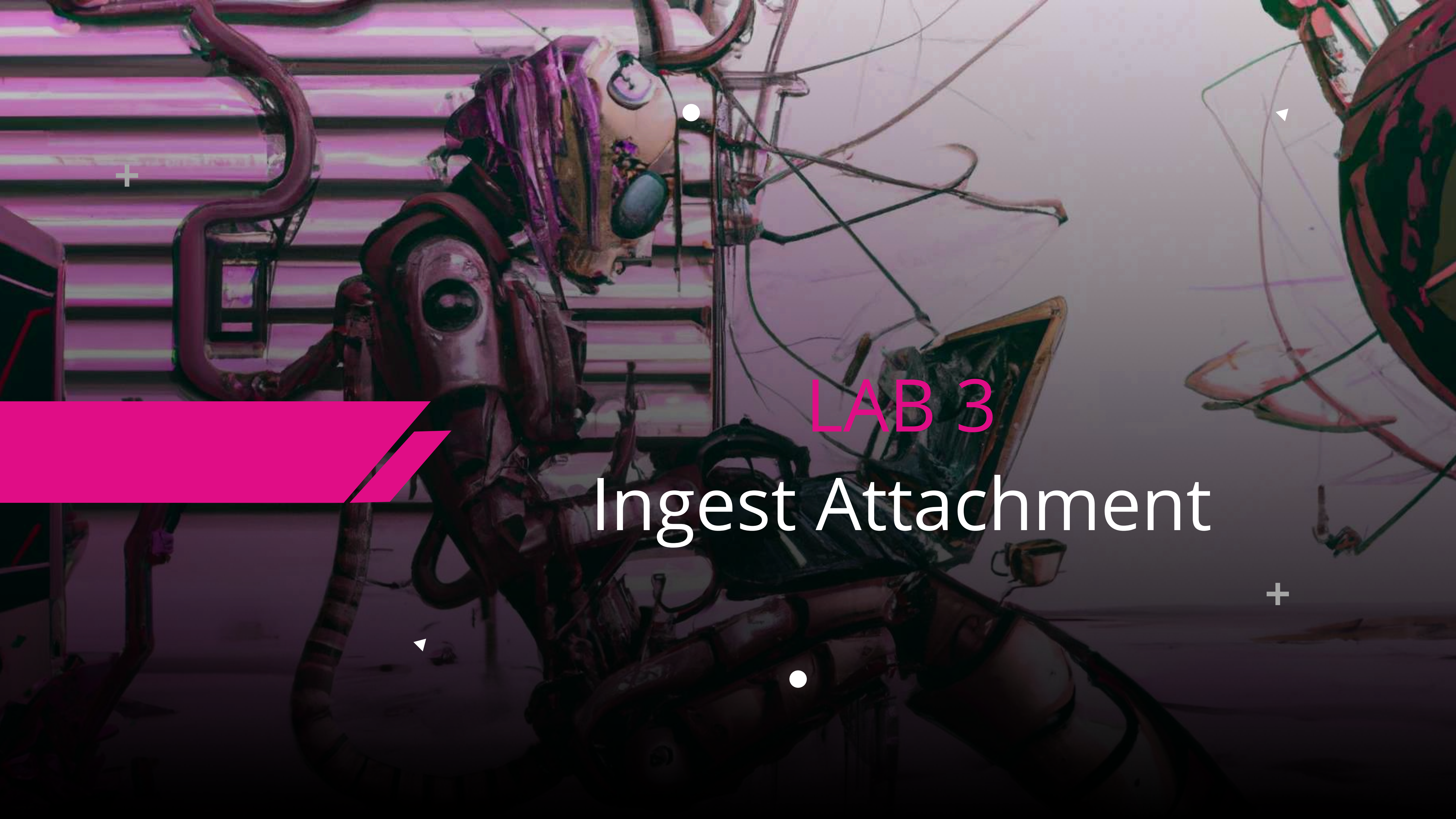
The attachment processor lets Elasticsearch extract file attachments in common formats (such as PPT, XLS, and PDF) by using the Apache text extraction library Tika.

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.

Using the attachment processor in a pipeline

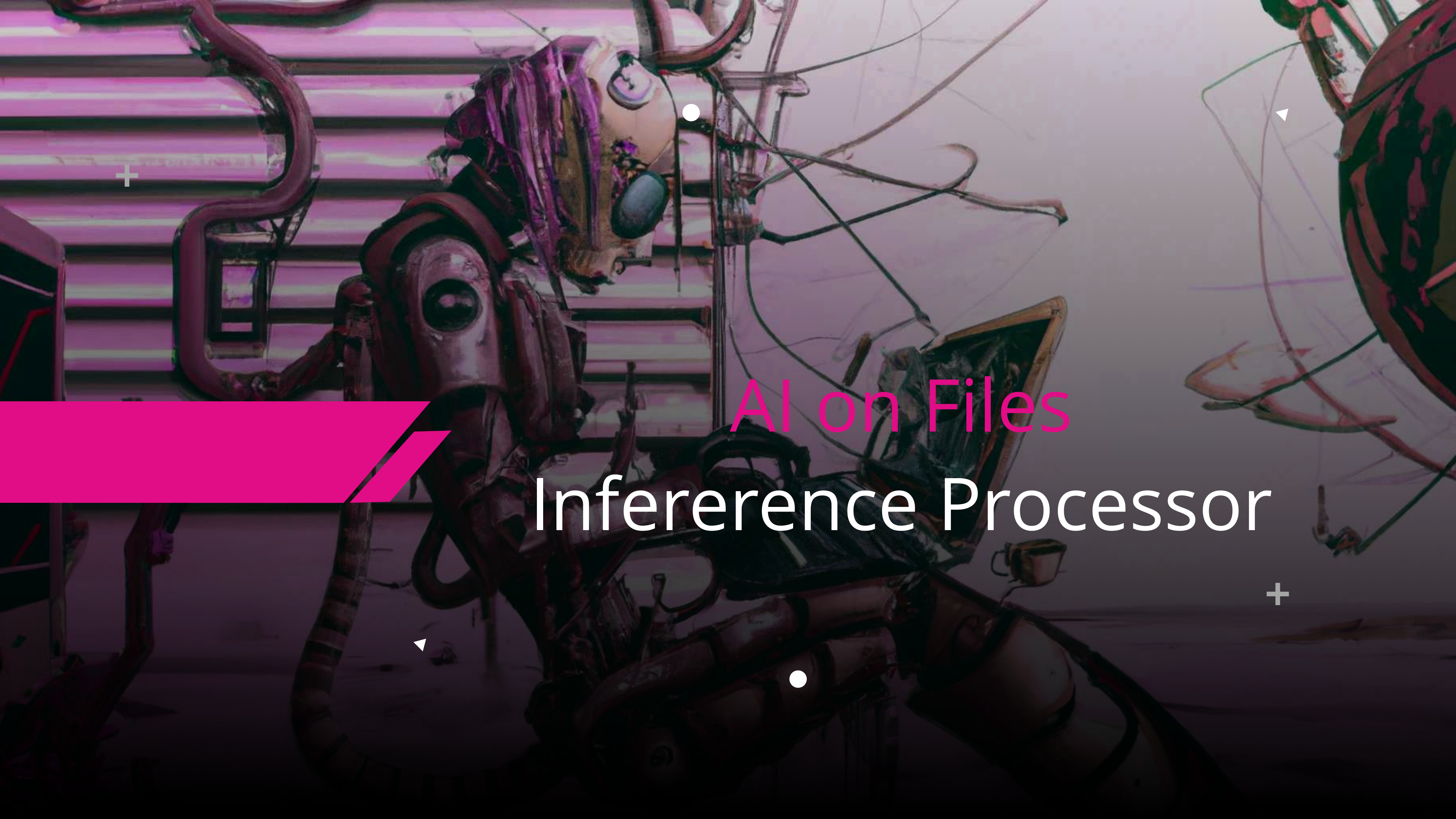
Table 4. Attachment options

Name	Required	Default	Description
`field`	yes	-	The field to get the base64 encoded field from
`target_field`	no	attachment	The field that will hold the attachment information
`indexed_chars`	no	100000	The number of chars being used for extraction to prevent huge fields. Use `-1` for no limit.
`indexed_chars_field`	no	`null`	Field name from which you can overwrite the number of chars being used for extraction. See `indexed_chars`.
`properties`	no	all properties	Array of properties to select to be stored. Can be `content`, `title`, `name`, `author`, `keywords`, `date`, `content_type`, `content_length`, `language`



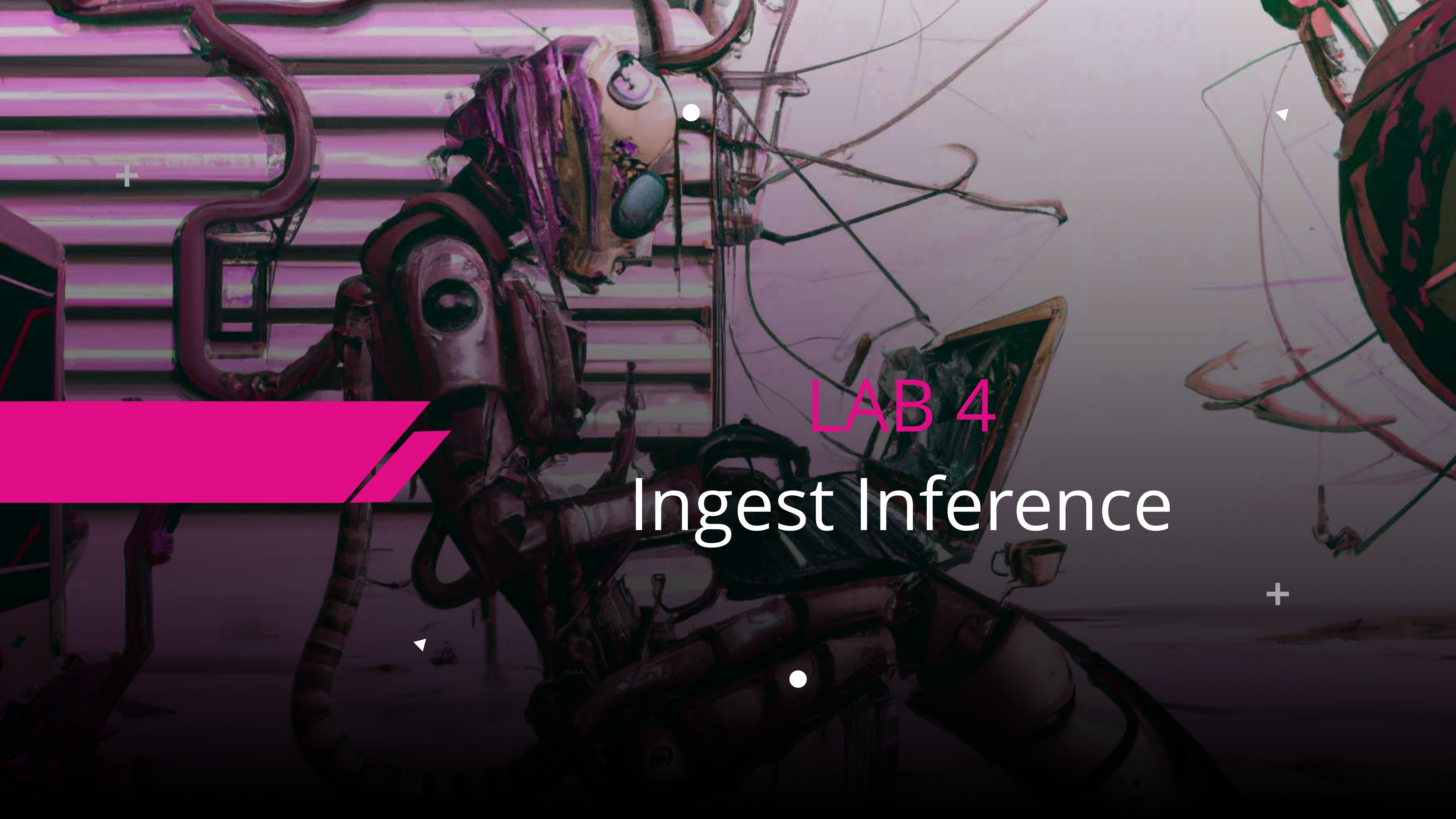
LAB 3

Ingest Attachment



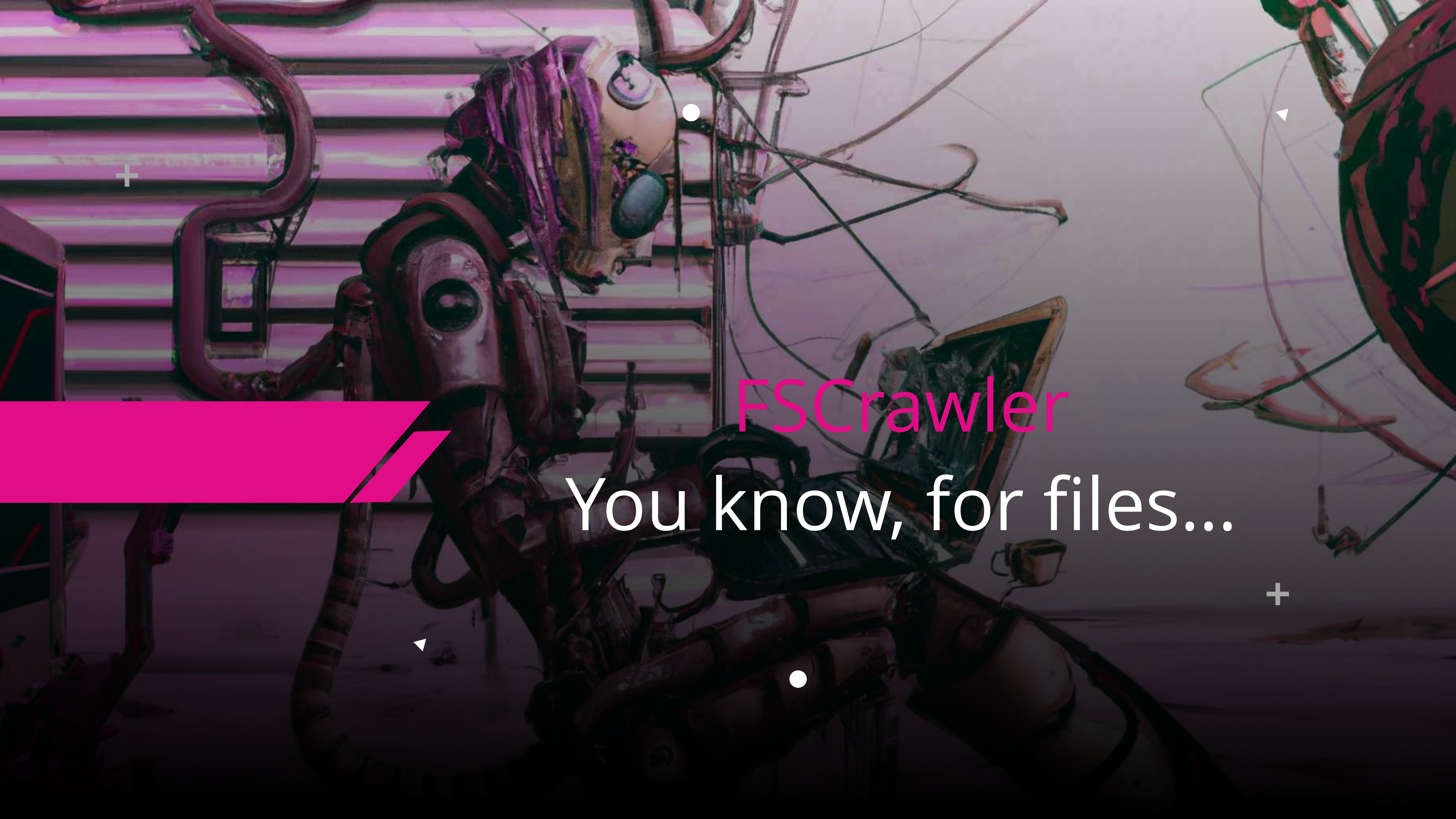
AI on Files

Infererence Processor



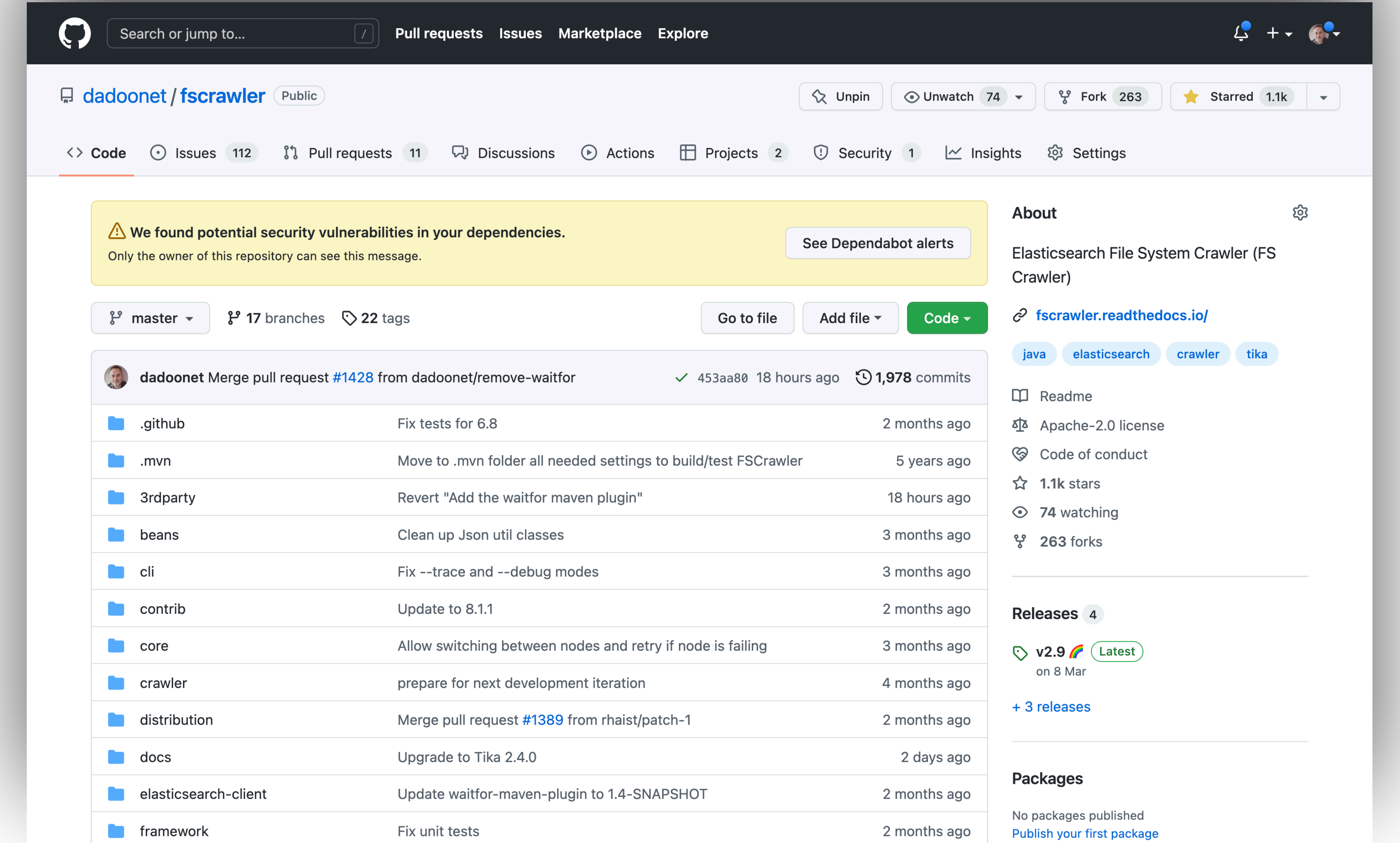
LAB 4

Ingest Inference



FSCrawler

You know, for files...



⚠ We found potential security vulnerabilities in your dependencies.

Only the owner of this repository can see this message.

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

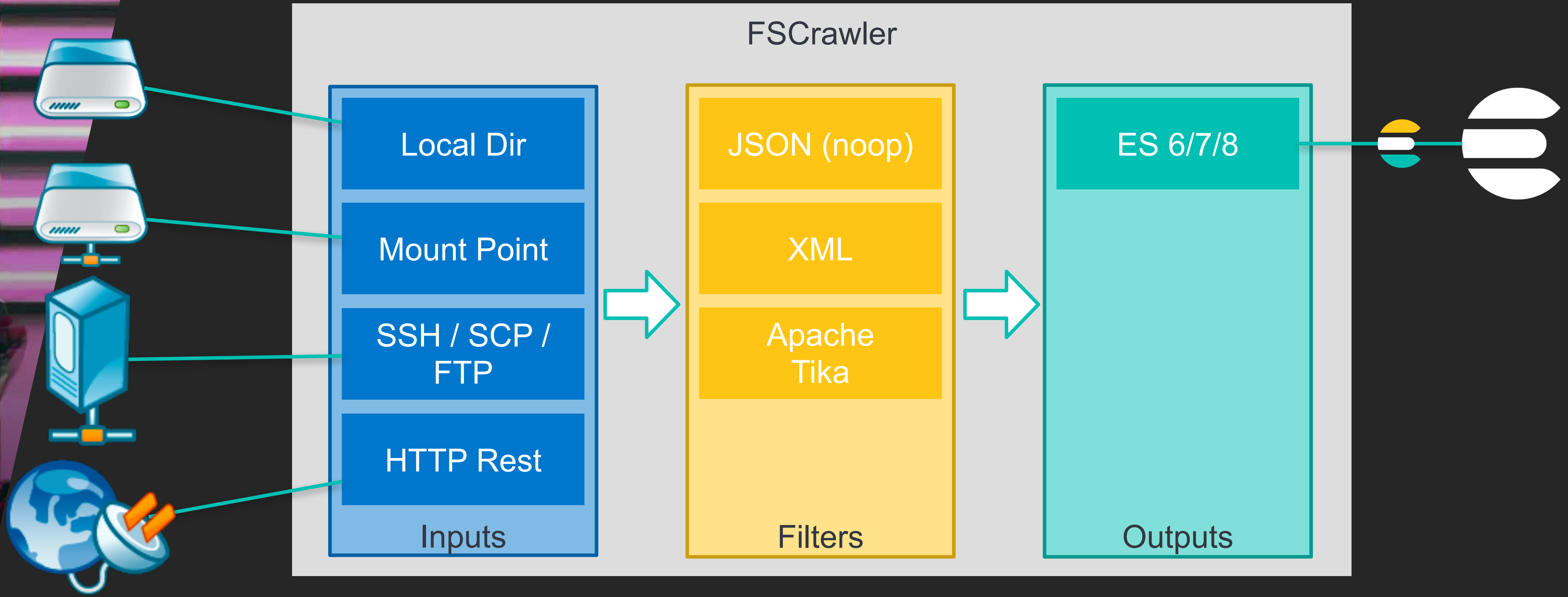
No packages published
Publish your first package

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.

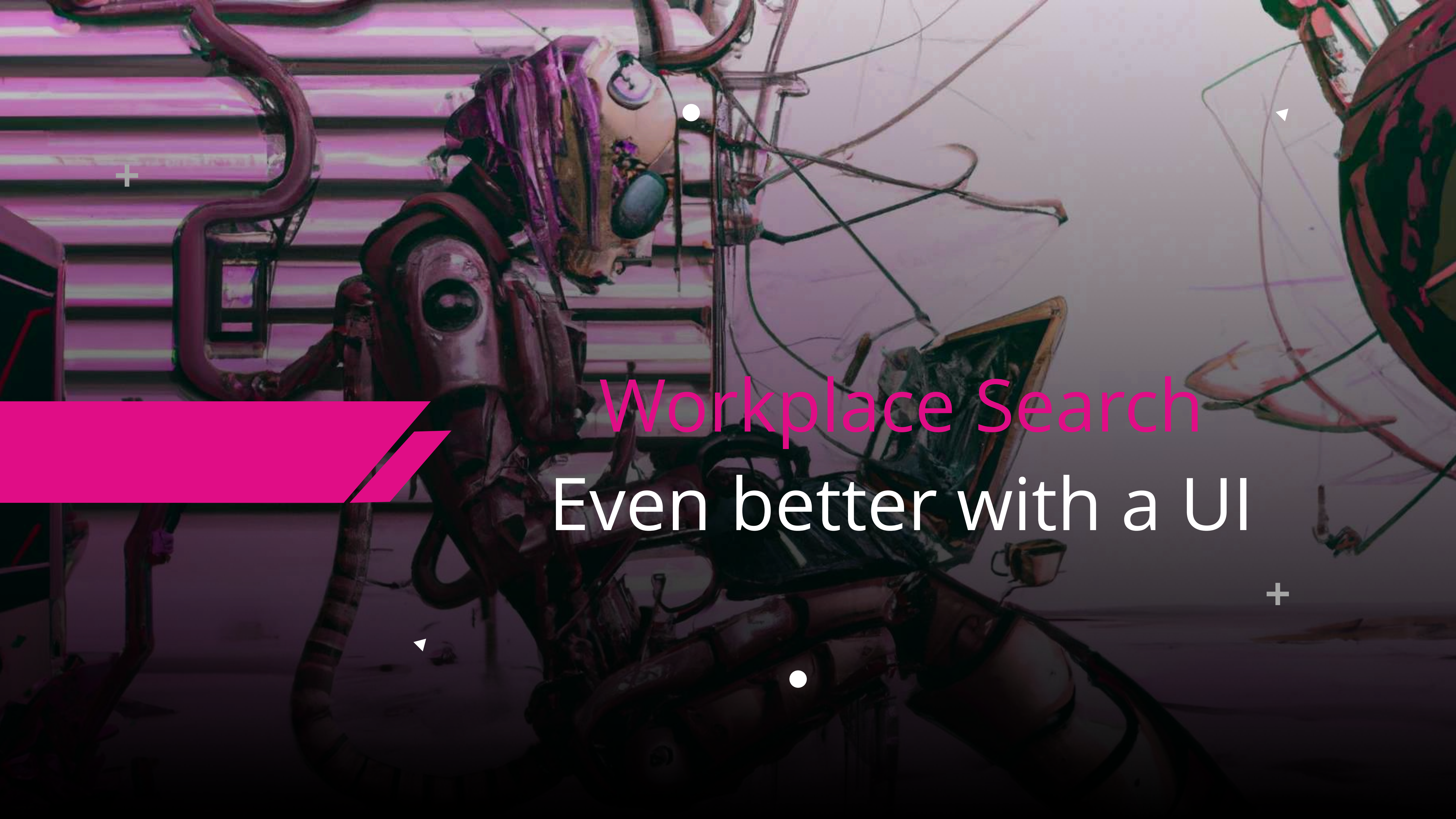
http://discuss.elastic.co/
https://stackoverflow.com/questions/tagged/fscrawler

FSCrawler architecture



LAB 5

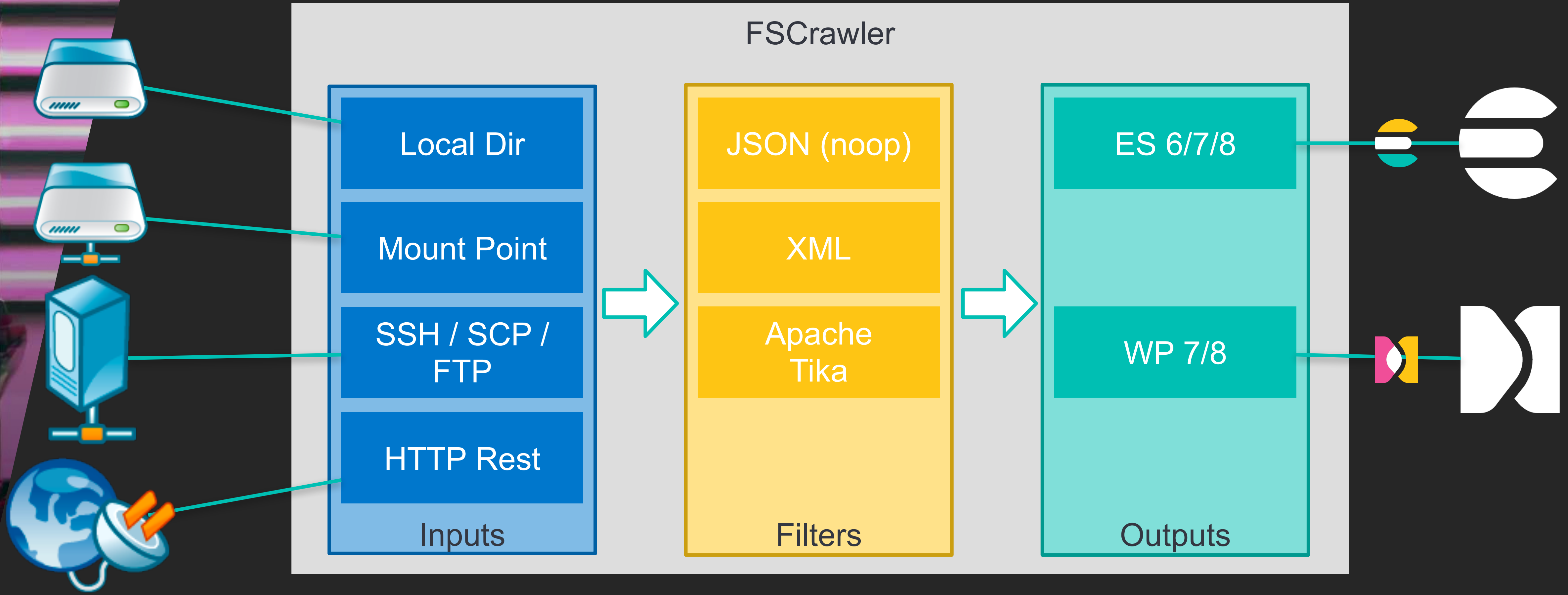
FSCrawler with Elasticsearch



Workplace Search

Even better with a UI

FSCrawler architecture



LAB 6

FSCrawler with Workplace Search

THANKS
FOR WATCHING