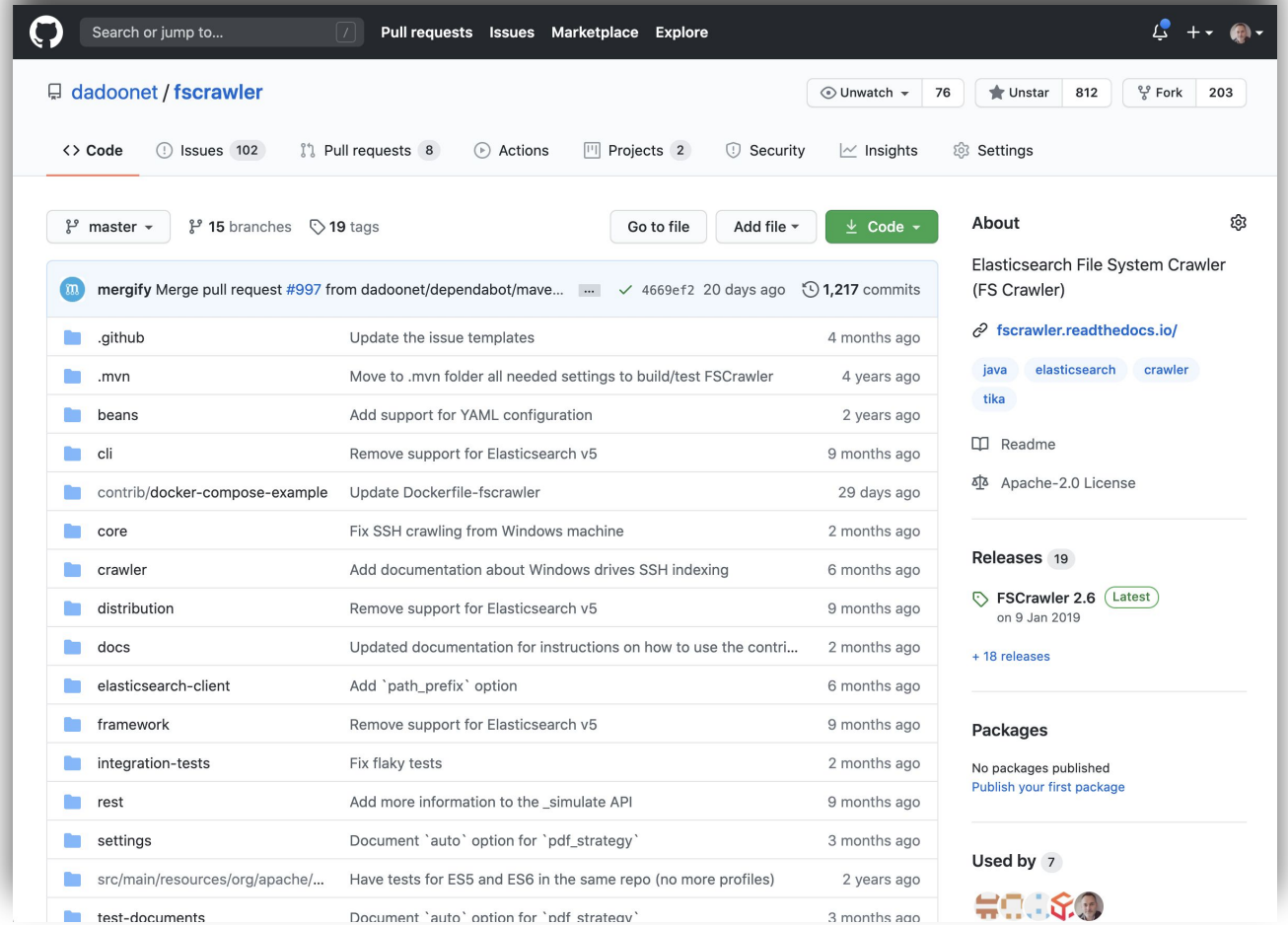
FSCrawler: you know, for files!

25/03/2021

Laetitia Richard
Consultante

David Pilato
Developer | Evangelist


Search or jump to...
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

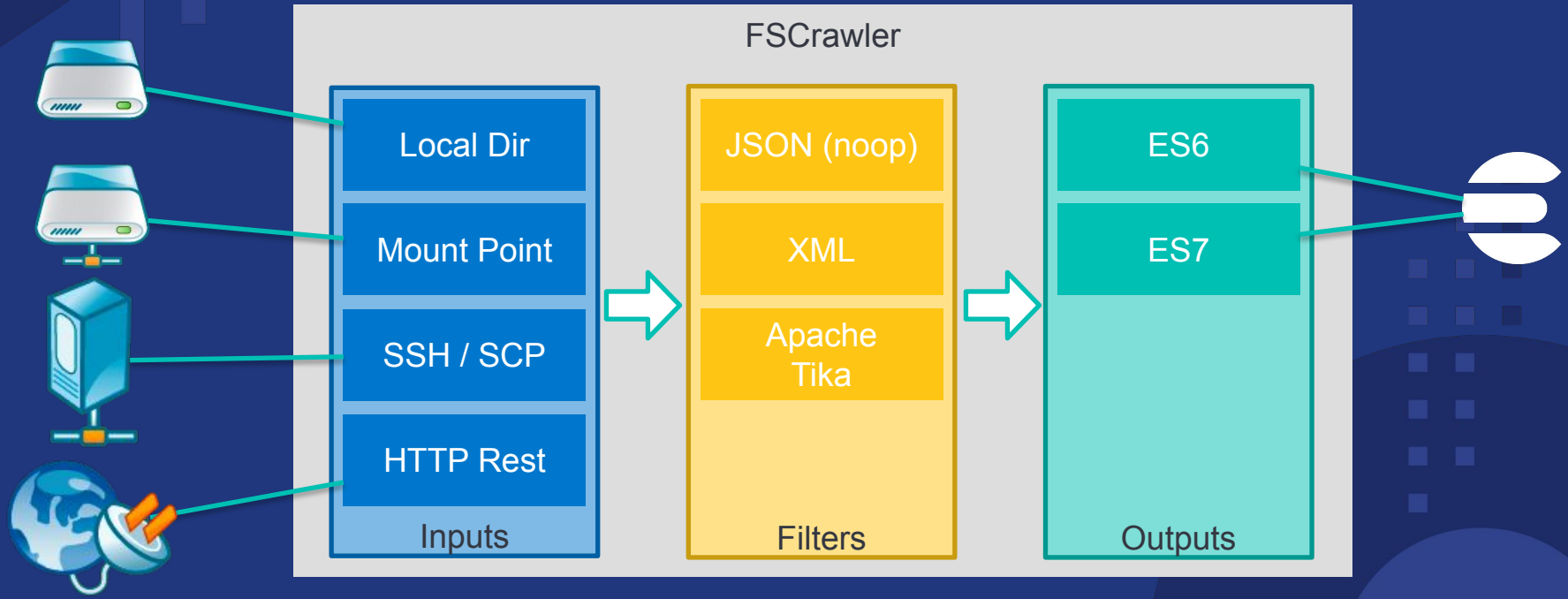
This project is a community project.

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



elastic

FSCrawler

Key Features

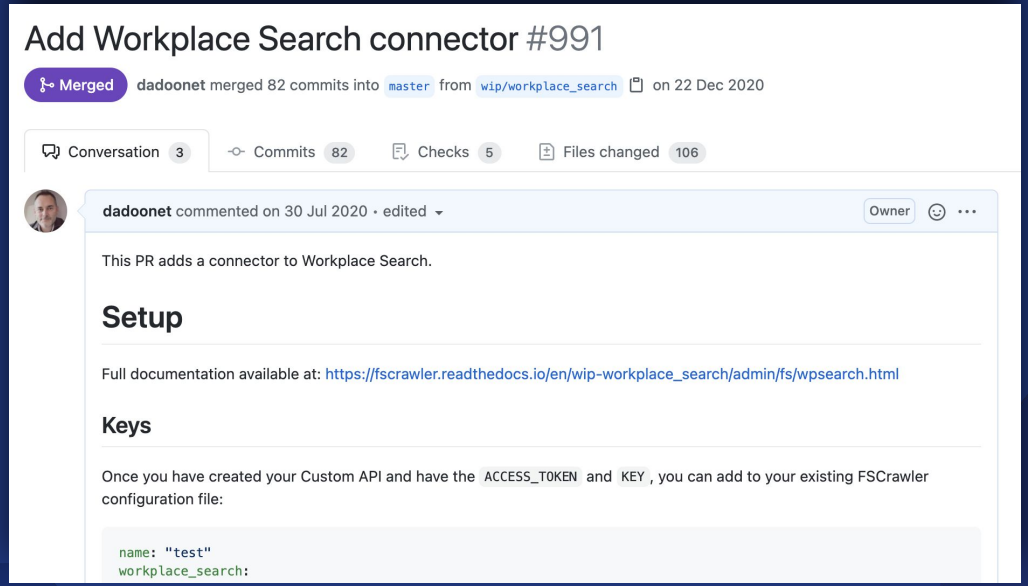
- Much more formats than ingest attachment plugin
- OCR (Tesseract)
- Much more metadata than ingest attachment plugin
  (See https://fscrawler.readthedocs.io/en/latest/admin/fs/elasticsearch.html#generated-fields)
- Language detection

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

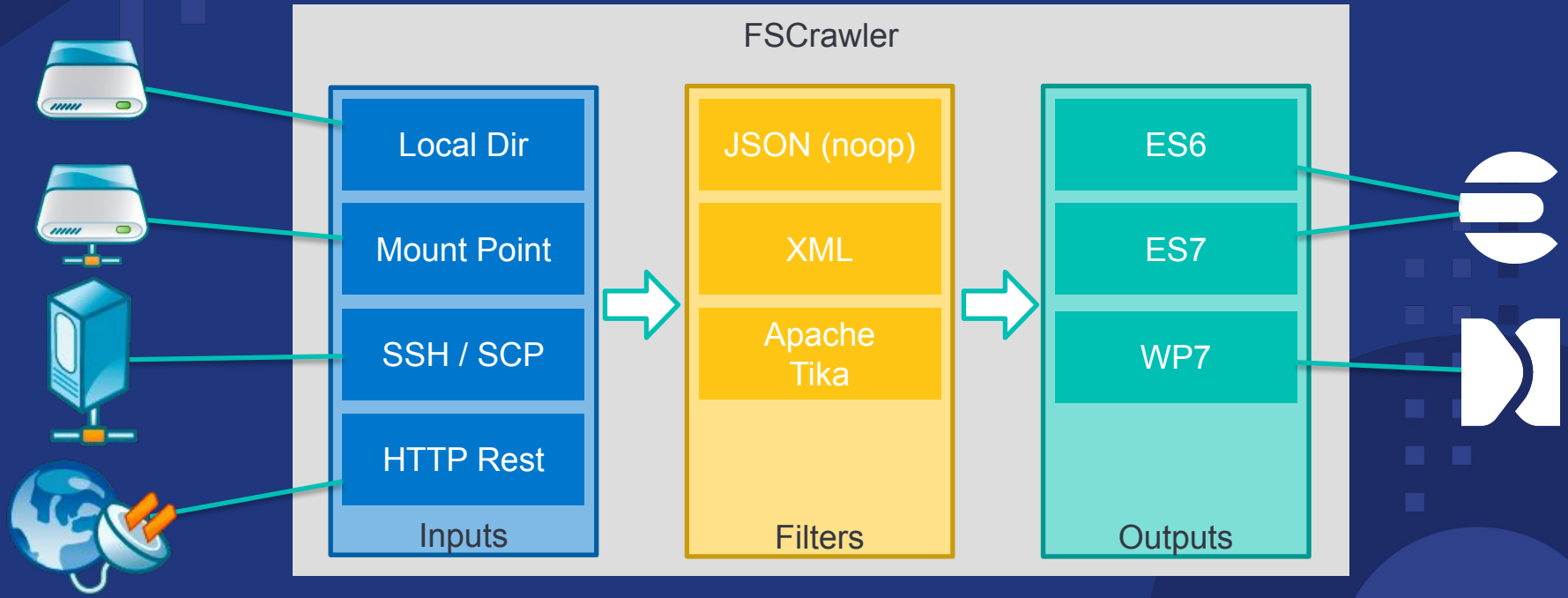
Workplace Search integration



elastic

FSCrawler

Architecture



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

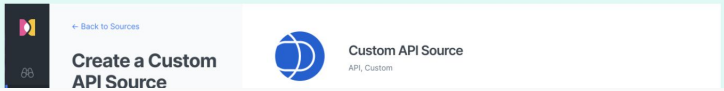
To easily start locally with Workplace Search, follow the steps:

- 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.



Demo

FS Settings

```
  "fs": {
    "url": " /PATH/data",
    "update_rate": "1m",
    "includes": ["*.doc",
"*.xls", "*.pdf", "*.ppt",
"*.pptx", "*.docx",
"*.xlsx", "*.odt",
"*.ods", "*.odp",
"*.rtf"],
    "excludes": ["*~*"],
    "indexed_chars":
"10%",
    "filename_as_id":
false,
    "add_filesize": true,
    "remove_deleted":
true,
     "index_content":
true,
      "lang_detect": true
    },
```

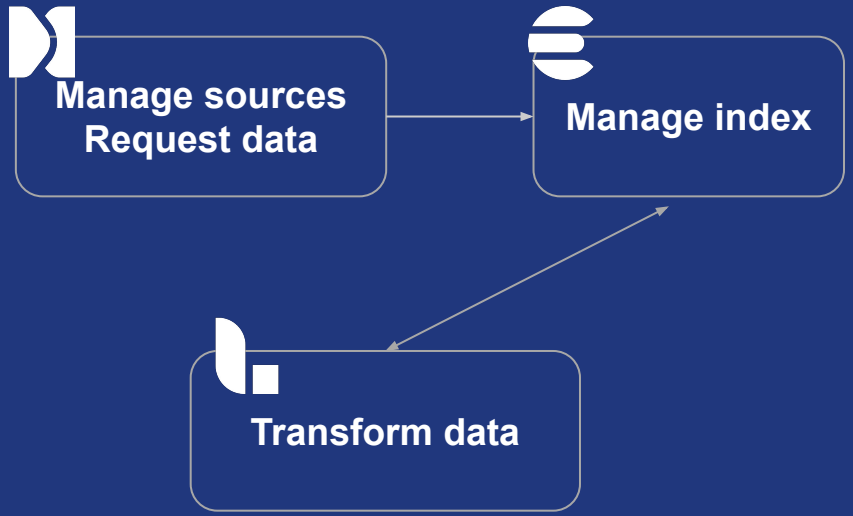
ES Settings

```
 "elasticsearch": {
    "nodes": [
       {
        "cloud_id":
"CLOUD_ID"
      }
    ],
    "username": "USER",
    "password": "PASSWORD"
  },
```

WPS Settings

```
 "workplace_search": {
    "access_token":
"TOKEN",
    "key": "KEY",
    "server":
"https://ID.ent-search.eu-
west-3.aws.elastic-cloud.c
om"
    }
```

Need to enrich your data ?



Input Settings

```
input {
     elasticsearch {
        cloud_id => ["
CLOUD_ID"]
        cloud_auth =>
"USER:PASSWORD"
        index =>
".ent-search-engine-6059f24fff7
092dce460a49b"
        query => '{ "query": {
"query_string": { "query": "*"
} }}'
        size => 100
        scroll => "5m"
        docinfo => true
      }
}
```

Filter Settings

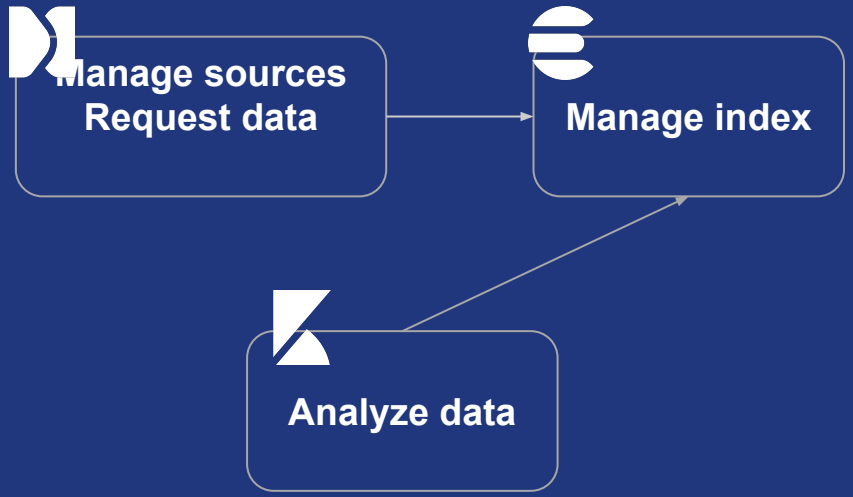
```
filter {
  mutate {
         gsub =>
["body$string", "[\n]", " "]
         gsub =>
["body$string", "[\t]", " "]
         gsub =>
["body$string", "[\r]", " "]
         strip =>
["body$string"]
      }
  if "/PATH/Data/" in
[path$string] {
  grok {
        match =>
["path$string",
"/PATH/Data/%{DATA:customer}
/%{DATA}"]
    }
    }

}
```

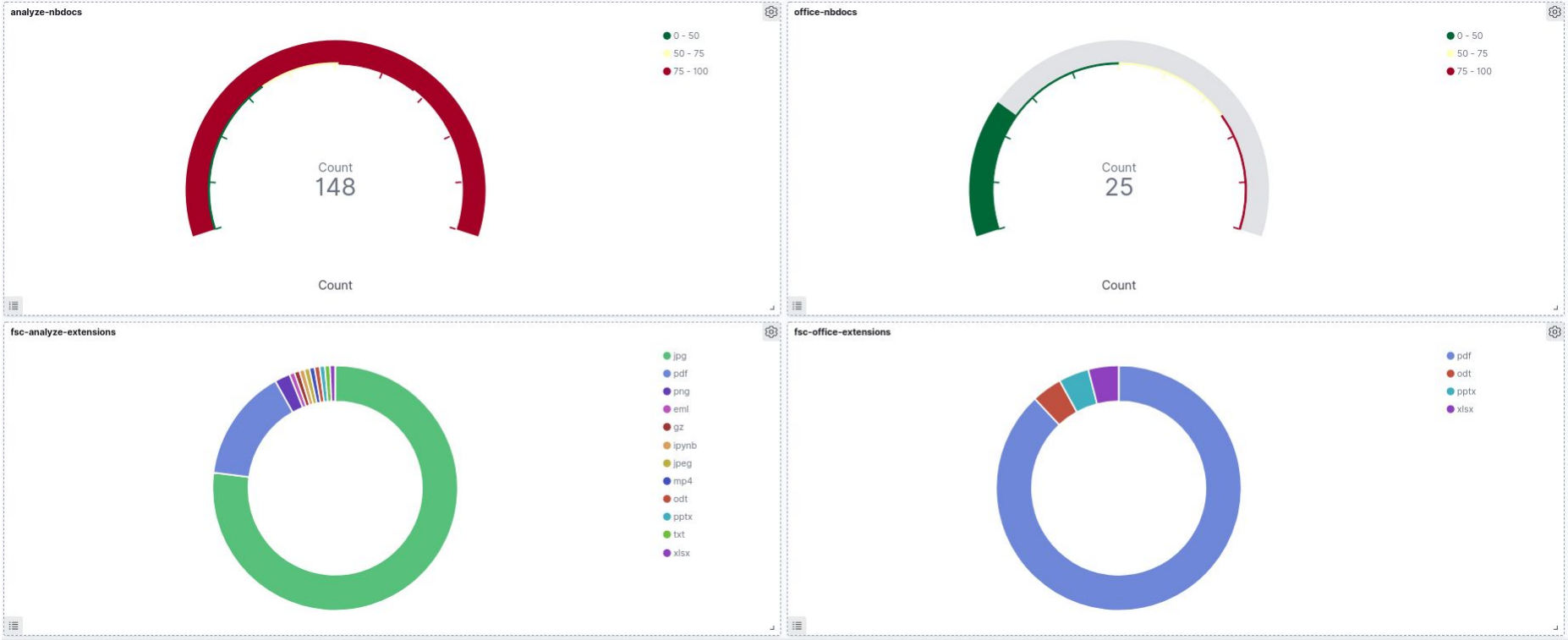
Ouput Settings

```
output {
      elasticsearch {
        cloud_id => ["
CLOUD_ID"]
        cloud_auth =>
"USER:PASSWORD"
        action => "update"
        doc_as_upsert => true
        document_id =>
"%{[@metadata][_id]}"
        #index => "fsc-office"
        index =>
"%{[@metadata][_index]}"
      }
}
```

Need to analyze your data ?



Observe and analyze your data



elastic

Beware of the settings

- FSCrawler with Workplace Search output is **not in watch mode** (you can use systemd)
- To transform your Workplace Search index you will have to **set dynamic mapping to true** first (default is strict)
- If you have other standard Workplace Search connectors, you will have to transform your data in another index because the full sync refresh the content source from scratch

Needs to be done

- New local file crawling implementation (WatchService): #399
- **Docker image**: #820
- Store jobs, configurations, status in Elasticsearch: #717
- Support for plugins (inputs, filters and outputs):
  - refactor with pf4j framework: #1114
  - rsync input: #377
  - Dropbox input: #264
  - S3 input: #263
  - Beats output: #682
- Switch to ECS format for the most common fields: #677
- Extract ACL informations: #464

Thanks!

PR are warmly welcomed!

https://github.com/dadoonet/fscrawler

Bonus slides

FSCrawler

Architecture