



Search Solutions 2022

Simplifying NLP researchers work with
Datafari Open Source

JULIEN MASSIERA & CÉDRIC ULMER | WORKSHOP

Description

- ▶ The objective of this workshop is to be able to use Datafari to extract text of various file types and apply NLP methods on them
 - ▶ Understand Datafari and its architecture
 - ▶ Install Datafari
 - ▶ Perform text extraction on a file set
 - ▶ Use NLP methods on files using a Spacy server
- ▶ Duration : 3 hours
- ▶ Requirements :
 - ▶ Laptop CPU 4 cores, at least 12GB of RAM, 20GB free disk space, with a Debian 10/11 or Ubuntu 20.04/22.04 OS (can be a VM)

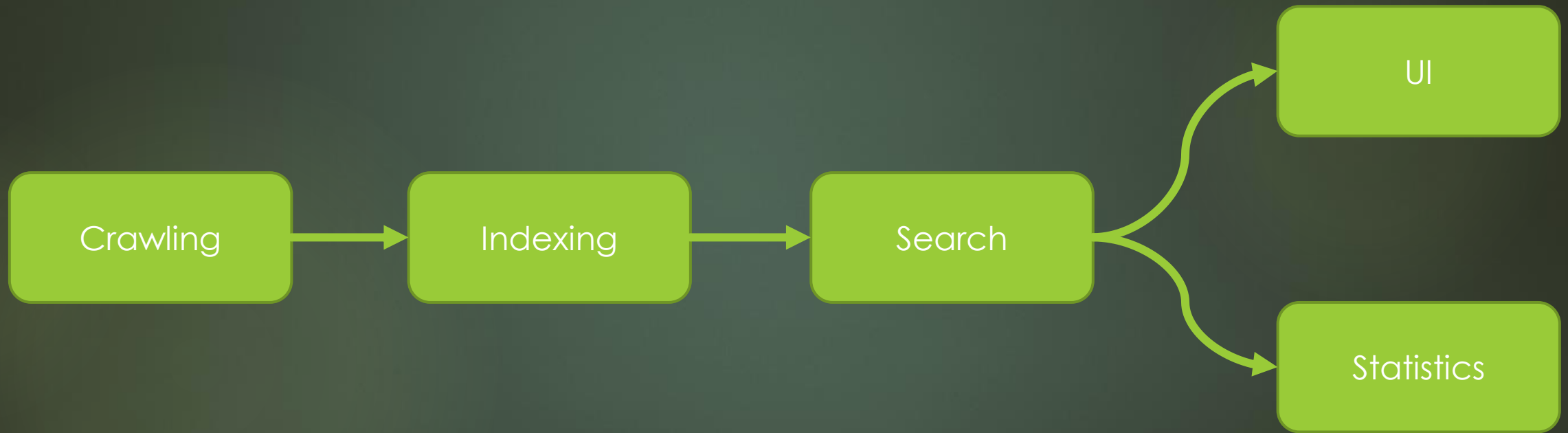
Enterprise Search Solution - Définition

An enterprise search application enables employees to find all the information that the company possesses without the need to know where the information is stored [securely].

Source: Enterprise search, Martin White

Enterprise Search Solution – How ?

4



Datafari

One solution to rule them all !

Datafari – an open source core

6

- ▶ Apache v2 Licence



Datafari – Main components

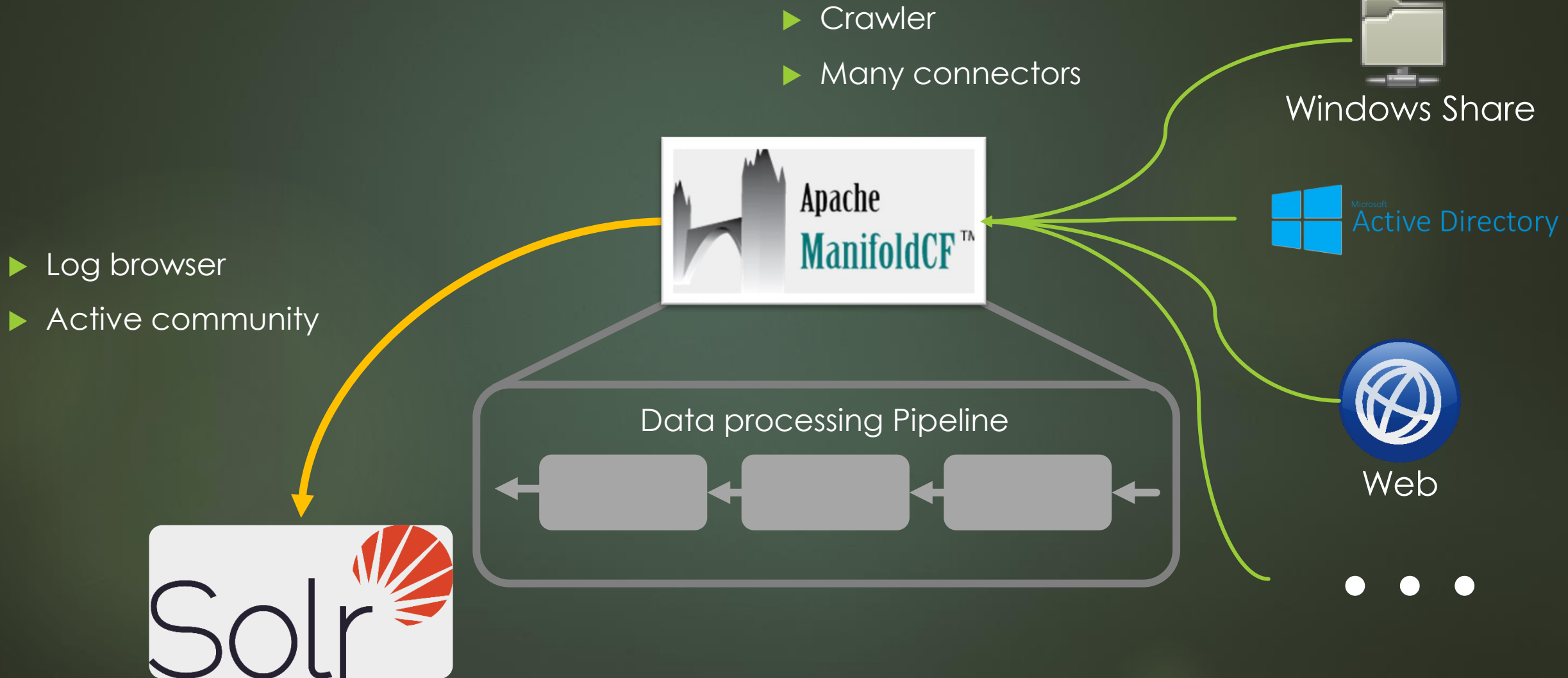
7



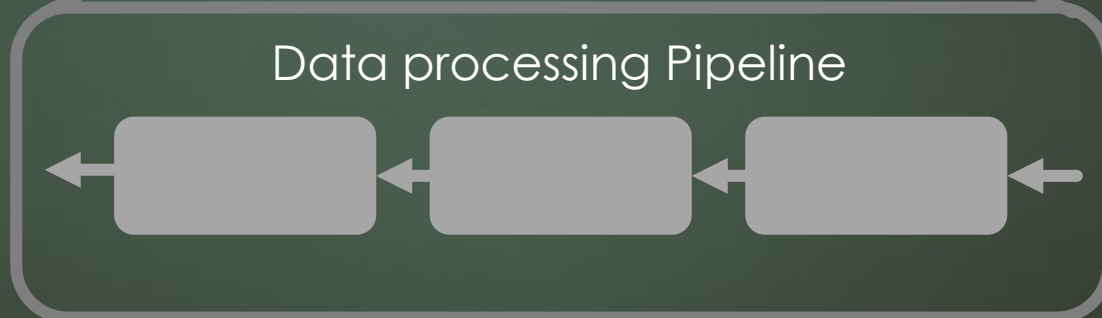
A simplified and unique admin UI to:

- ▶ Install
- ▶ Manage
- ▶ Use

Datafari - ManifoldCF



- ▶ Crawler
- ▶ Many connectors

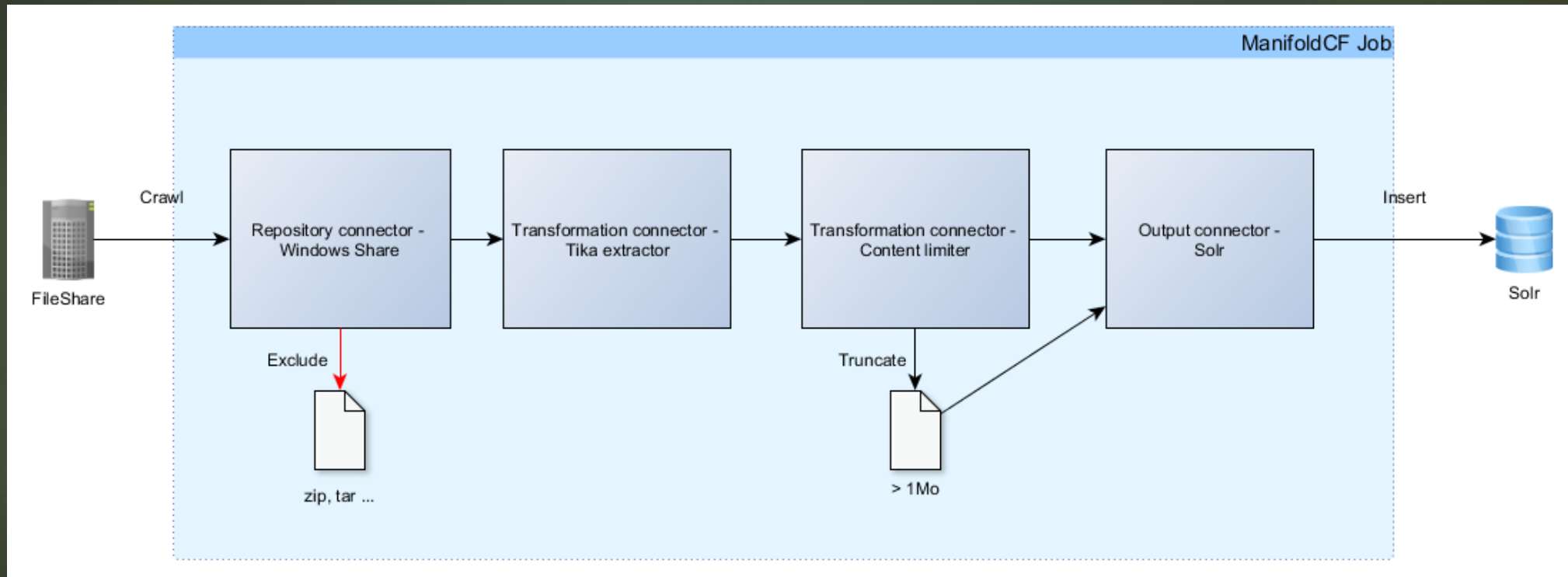


Datafari - ManifoldCF

- ▶ Used to configure a document processing pipeline between an input named « Repository connector » and an output named « Output connector ».
- ▶ Between these two connectors, documents are processed by another connector type named « Transformation connector »
- ▶ A repository connector is specific to a document source type (ex: Windows Share), same for an output connector (ex: Solr)

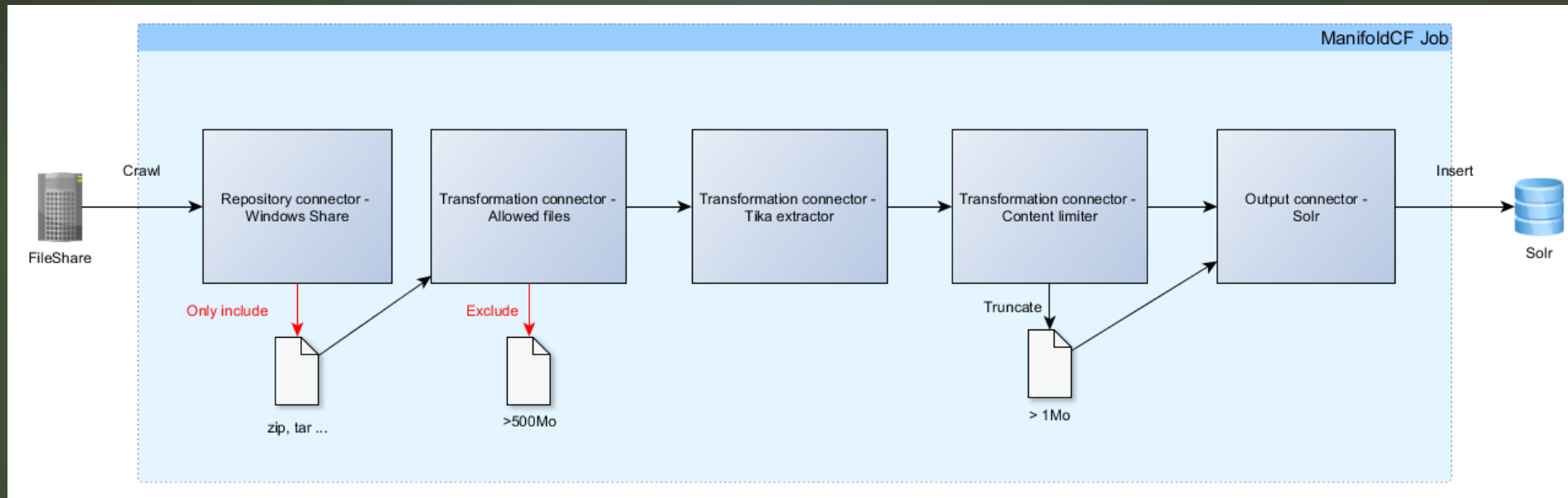
Datafari - ManifoldCF

► MCF pipeline: Example 1



Datafari - ManifoldCF

► MCF pipeline: Example 2



Datafari – ManifoldCF

- ▶ MCF connectors have a global configuration and a job configuration
- ▶ The global configuration is set at the connector level, whereas the job configuration is set at the job level
- ▶ The global configuration is valid for all the jobs using a given connector
- ▶ A job configuration is valid only for job in which it is declared

Datafari – ManifoldCF

13


The screenshot displays the Datafari Connectors Framework interface. The top left features the logo 'Datafari CONNECTORS FRAMEWORK Powered by Apache ManifoldCF'. A main navigation sidebar on the left lists various sections: Outputs, Authorities, Repositories (selected), List Repository Connections, List Notification Connections, Jobs, Status Reports, History Reports, and Miscellaneous. The main content area is titled 'Document Ingestion' and 'Edit Repository Connection - TestRepo_5c9426f500fa4d9'. It contains a configuration form with the following fields:

Name	Type	Throttling	Server
Server:			localhost
Authentication domain (optional):			
User name:			francelabs
Password:		
Connection Timeout			150000
Session Timeout			150000
Socket Timeout			150000
Response Timeout			120000
Resolve Order			LMHOSTS,DNS,WINS
List Count			200
Use SIDS for security:			<input checked="" type="checkbox"/>
Bin Name (optional):			
Maximum version of SMB			2.1

At the bottom of the form, there are 'Save' and 'Cancel' buttons.

Datafari – ManifoldCF

14



Document Ingestion

Edit job - Crawl_TestRepo_5c9426f500fa4d968809986

Name | Connection | Scheduling | Paths | Global Filters | Security | Metadata | Content Length | File Mapping | URL Mapping | Last Access | Metadata expressions

Emptier Filter | Tika server | Metadata Cleaner

Pipeline:

	Stage	Type	Precedent	Description	Connection name
	1.	Repository			TestRepo_5c9426f500fa4d968809986
Delete Insert transformation before Insert output before	2.	Transformation	1.	<input type="text"/>	MetadataAdjuster
Delete Insert transformation before Insert output before	3.	Transformation	2.	<input type="text"/>	EmptierFilter
Delete Insert transformation before Insert output before	4.	Transformation	3.	<input type="text"/>	TikaServerRmetaConnector
Delete Insert transformation before Insert output before	5.	Transformation	4.	<input type="text"/>	MetadataCleaner
Insert transformation before Insert output before	6.	Output	5.	<input type="text"/>	DatafariSolrNoTika

Datafari – ManifoldCF

Name | Connection | Scheduling | **Paths** | Global Filters | Security | Metadata | Content Length | File Mapping | URL Mapping | Last Access | Metadata expressions

Emptier Filter | Tika server | Metadata Cleaner

test

Delete **Insert** Include -- Any file or directory -- matching

Delete 1. Include indexable file(s) matching *

Delete **Insert** Include -- Any file or directory -- matching

Delete 2. Include directory(s) matching *

Add Include -- Any file or directory -- matching

Add (root) + -- Pick a folder -- or type a path: "

Save **Cancel**

A few words about Spacy by Prodigy

- ▶ Open source API specialised in NLP, in particular for transformers
- ▶ Useful to leverage existing algorithms (from Hugging Face for instance)
- ▶ Useful to train your own models
- ▶ Can be combined with FastAPI to expose Spacy as RESTfull server to apply NLP on the text
- ▶ Can be combined with Prodigy for the annotations phase

Workshop

17

- ▶ Install Datafari + JCIFS-NG lib
- ▶ Create simplified filer job
- ▶ Create file system output connector
- ▶ Use the file system output in the filer job
- ▶ Create Spacy FastAPI connector
- ▶ Add entities fields to Solr schema
- ▶ Use Spacy FastAPI connector in the filer job

Bonus – OCR

- ▶ Set-up a new VM/Server with a Tika Server (copy the on from Datafari located in DATAFARI_HOME/tika-server)
- ▶ Install Tesseract and the wanted languages modules
- ▶ Configure the Tesseract and PDF parsers of Tika to fit your needs
- ▶ We have a documentation on how to set-up a Tika server node and how to configure the parsers:
- ▶ Tesseract/Parsers configuration fine tuning is up to you !

Bonus – Tika entity extraction/NLP

19

- ▶ Set-up a new VM/Server with a Tika Server (copy the on from Datafari located in DATAFARI_HOME/tika-server)
- ▶ Follow the Tika documentation, forums and blogs to properly configure Tika
- ▶ Use what you learnt from the training to properly configure your MCF jobs ;)

Bonus – Entity extraction integrated to Datafari

20

- ▶ Pre-configured: phone numbers
- ▶ That you can configure through the administration UI: exact match entity extraction
- ▶ For regex based entity extraction you will need to develop an update processor on your own for the moment

Documentation

- ▶ Confluence:
<https://datafari.atlassian.net/wiki/spaces/DATAFARI/overview>
- ▶ Spacy FastAPI server:
<https://datafari.atlassian.net/wiki/spaces/DATAFARI/pages/2657517573/Setting+up+a+server+to+host+Spacy+for+Named+Entity+Recognition>
- ▶ Spacy FastAPI connector:
<https://datafari.atlassian.net/wiki/spaces/DATAFARI/pages/2469920769/Spacy+Transformation+Connector>

Contact

22

julien.massiera@francelabs.com
cedric.ulmer@francelabs.com

www.francelabs.com
www.datafari.com

Questions ?