



DATAFARI

CEDRIC ULMER
FRANCE LABS
DATA SCIENCE MEETUP
15/01/18

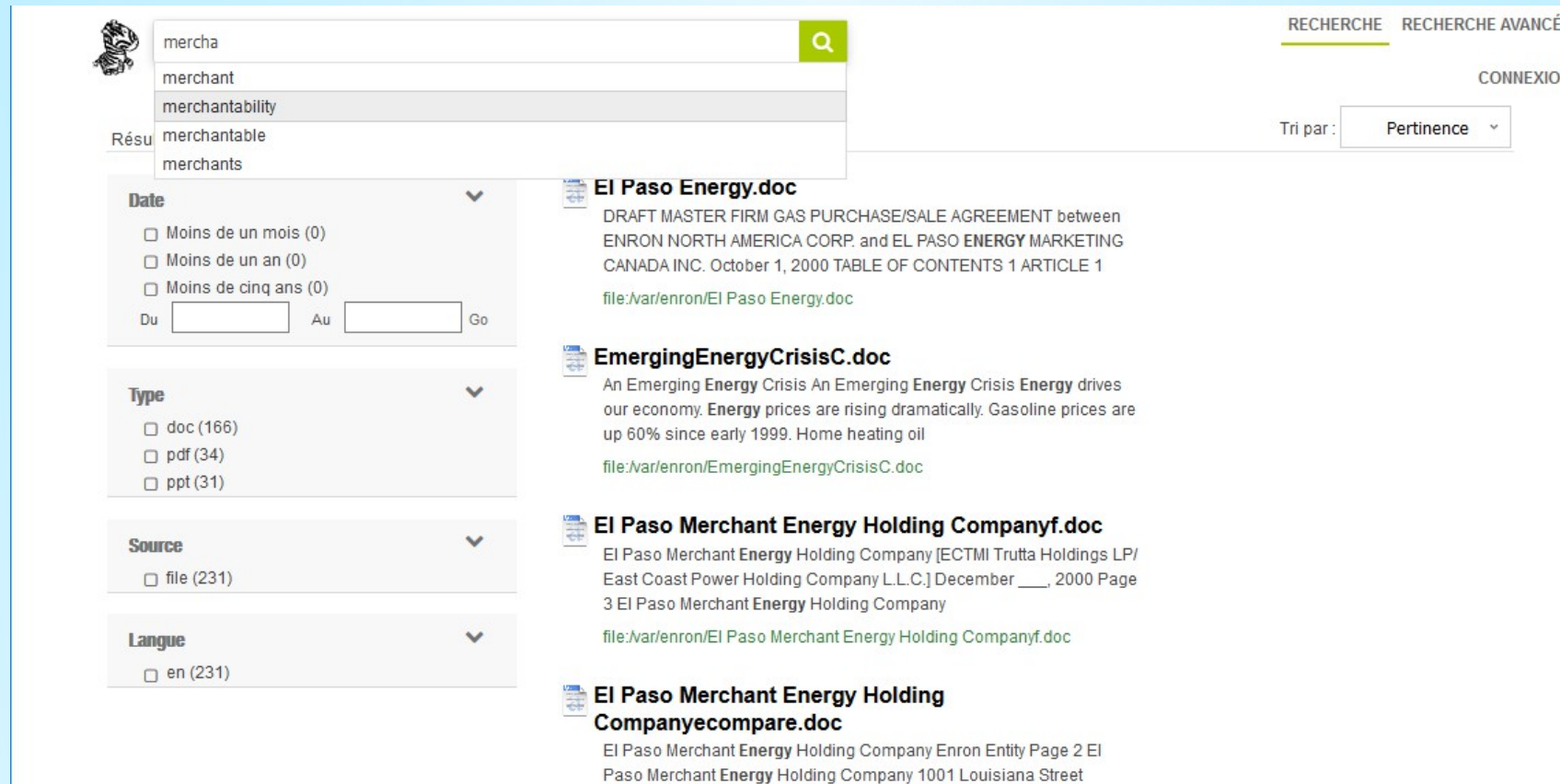
Datafari

 So what is Datafari?

- « Packaged solution » to analyse and search for documents and data
- Can index heterogeneous data formats from multiple sources
- Federated search interface
- Apache v2 licence

Datafari

STOP !!!! SEARCH ?!?!



The screenshot shows a search engine interface with a search bar containing 'mercha'. A dropdown menu lists suggestions: 'mercha', 'merchant', 'merchantability', 'merchantable', and 'merchants'. The search bar has a green magnifying glass icon. To the right of the search bar are links for 'RECHERCHE' (underlined) and 'RECHERCHE AVANCÉE', and a 'CONNEXION' link. Below the search bar is a 'Résumé' section. On the left side, there are four filter sections: 'Date' with checkboxes for 'Moins de un mois (0)', 'Moins de un an (0)', and 'Moins de cinq ans (0)', and input fields for 'Du', 'Au', and 'Go'; 'Type' with checkboxes for 'doc (166)', 'pdf (34)', and 'ppt (31)'; 'Source' with a checkbox for 'file (231)'; and 'Langue' with a checkbox for 'en (231)'. The search results are displayed on the right, showing four documents:

- El Paso Energy.doc**
DRAFT MASTER FIRM GAS PURCHASE/SALE AGREEMENT between ENRON NORTH AMERICA CORP. and EL PASO ENERGY MARKETING CANADA INC. October 1, 2000 TABLE OF CONTENTS 1 ARTICLE 1
[file:/var/enron/El Paso Energy.doc](#)
- EmergingEnergyCrisisC.doc**
An Emerging Energy Crisis An Emerging Energy Crisis Energy drives our economy. Energy prices are rising dramatically. Gasoline prices are up 60% since early 1999. Home heating oil
[file:/var/enron/EmergingEnergyCrisisC.doc](#)
- El Paso Merchant Energy Holding Companyf.doc**
El Paso Merchant Energy Holding Company [ECTMI Trutta Holdings LP/ East Coast Power Holding Company L.L.C.] December __, 2000 Page 3 El Paso Merchant Energy Holding Company
[file:/var/enron/El Paso Merchant Energy Holding Companyf.doc](#)
- El Paso Merchant Energy Holding Companycompare.doc**
El Paso Merchant Energy Holding Company Enron Entity Page 2 El Paso Merchant Energy Holding Company 1001 Louisiana Street

Why Datafari ?

☞ Choice of the Apache Solr and Elasticsearch technologies (more about this later...)

☞ We had three possibilities to answer a customer requirements :

- Use a packaged solution available on the market from a 3rd party
- Starting from Apache Solr or Elasticsearch (or others)
 - Develop, gather necessary components for each customer needs
- Create our own packaged solution (yeah!)

Why Datafari ?

🗨️ Problems with 3rd party proprietary solutions:

- Black box
- Roadmap not clear
- Durability (bankrupt, acquisition...)

🗨️ Problems with 3rd party open source solutions:

- Lack of technical documentation
- Difficulty to setup an understandable debug environment
- Delay in the embedded components updates: In particular Solr or ES
- License issues (mostly copyleft/viral ones)
- Lack of resilience from the makers

=> Required us to develop our own solution to better address our customer needs

Why Datafari

💡 Idea:

- Gather the best of both worlds :
 - The “packaged” aspect of existing solutions
 - Many functionalities
 - All in one
 - The flexibility of a solution based on Solr and ES
- All of that with an Apache v2 licence 😊
- Focus on Enterprise Search:
 - Admin for search experts
 - Admin for search admin
 - Eased AD/LDAP management
 - Search and data analytics

Solr 

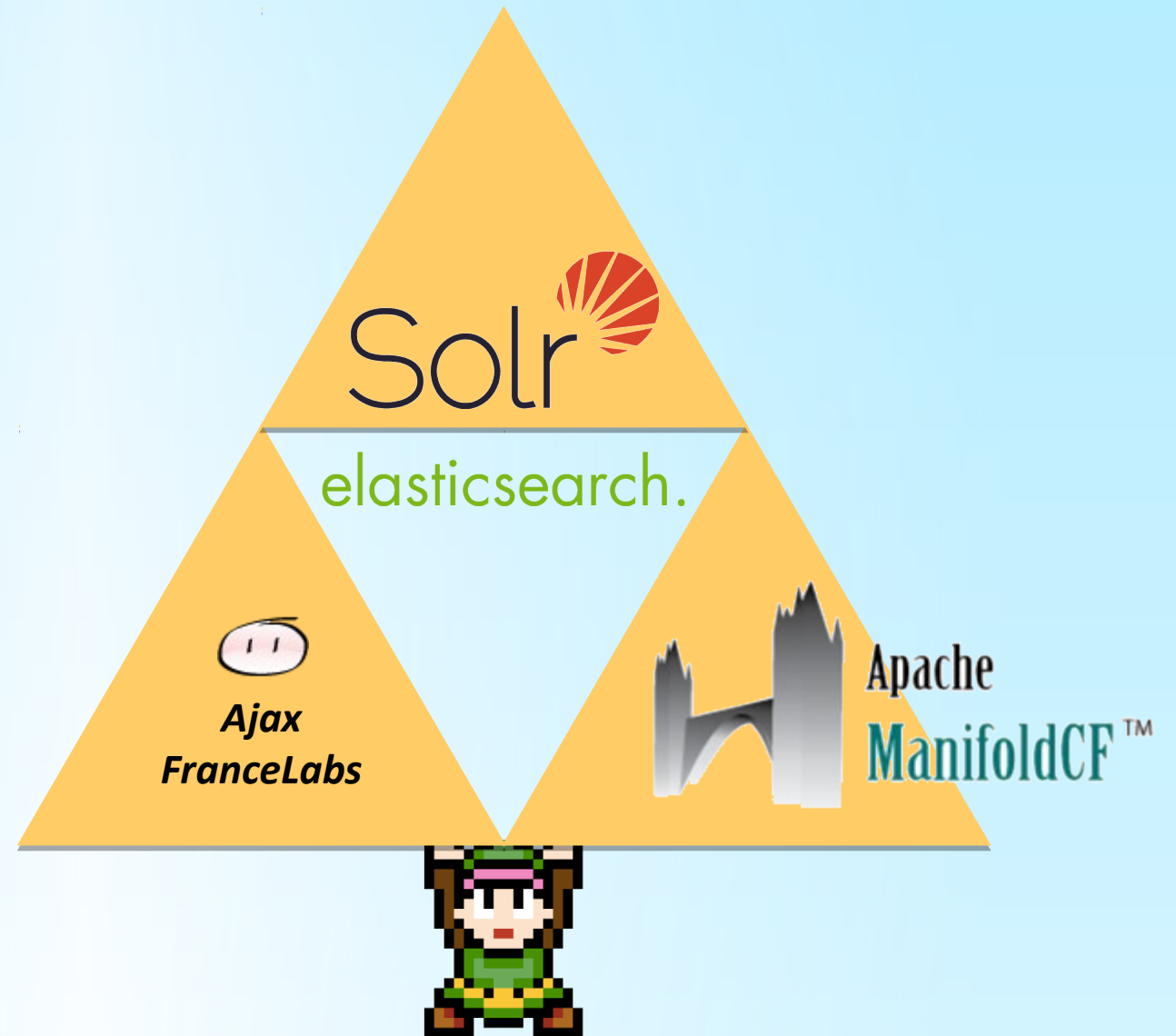


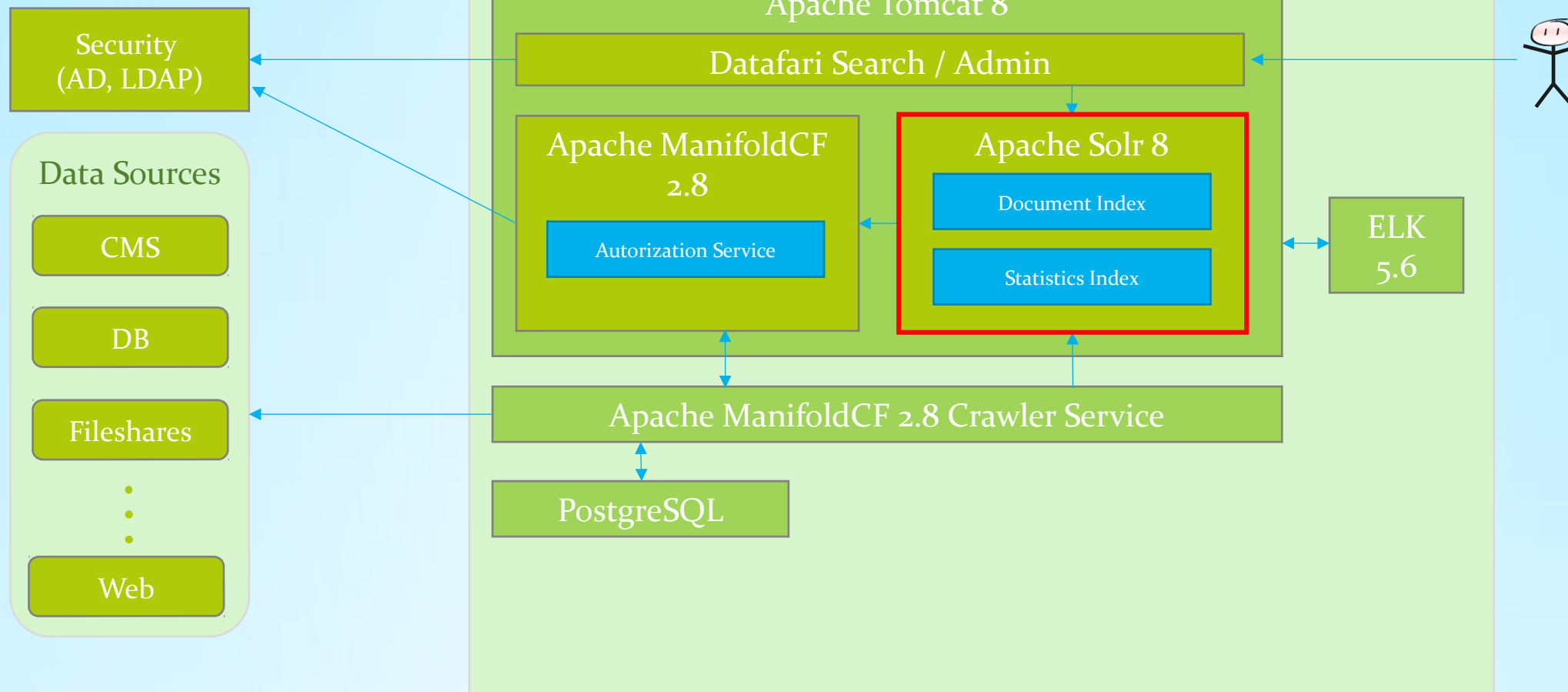
Datafari 

Datafari

 Based on 4 building blocks:

- Apache Solr
 - The heart of the search engine
- Apache Manifold CF
 - Crawling documents
- Ajax FranceLabs
 - UI
- Elasticsearch
 - Data analytics





Apache Solr

- ☺ Lucene based Full text search engine
- ☺ Apache Top Level project
- ☺ Large community (users/devs)
- ☺ Efficient/Reliable
- ☺ Scalable (with Solrcloud)
 - High availability
 - Queries
 - Index volume



Apache Solr

- ☞ Standalone application
- ☞ REST APIs XML/HTTP
 - Indexing
 - Querying
- ☞ Caching
- ☞ Web admin interface
- ☞ Configuration through XML config files or APIs



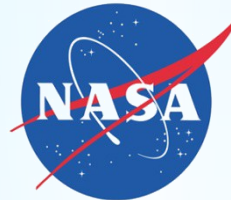
Apache Lucene/Solr – Some refs



Instagram



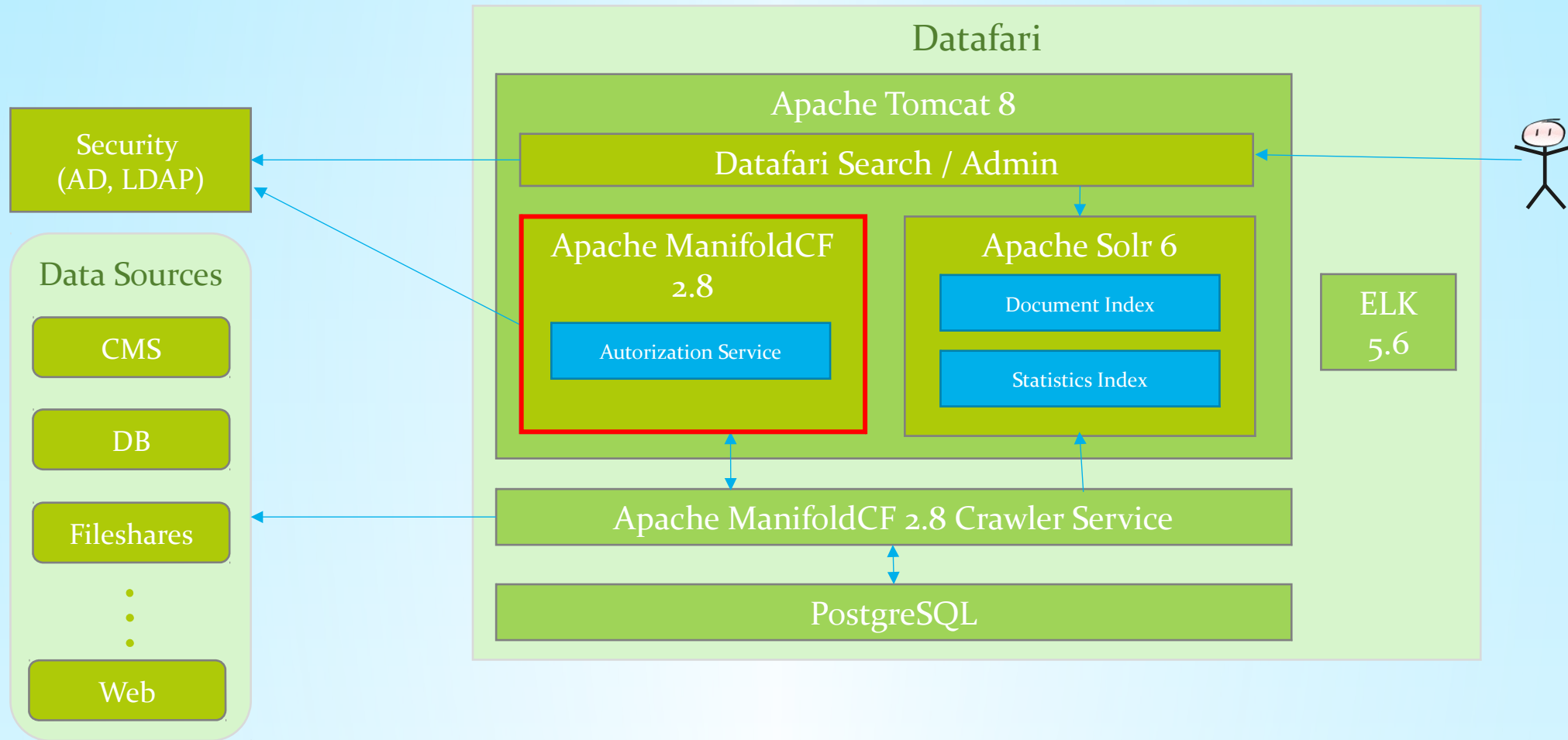
at&t



Apache Solr for Datafari

- 🗉 Search core of Datafari
- 🗉 Preconfigured index for rich documents
 - Language detection
 - Standard facets
 - Autocomplete
 - Spellchecker





Apache Manifold CF

- 🗨 Framework for data crawling
- 🗨 Management of incremental crawling
- 🗨 Authorization management
- 🗨 Programmable crawls (time windows, loads, regex...)



Apache Manifold CF

☺ Many off the shelf connectors:

- FileShare (Samba)
- JDBC
- Website
- Alfresco
- CMIS
- Sharepoint
- Mail
- Dropbox
- LDAP/AD

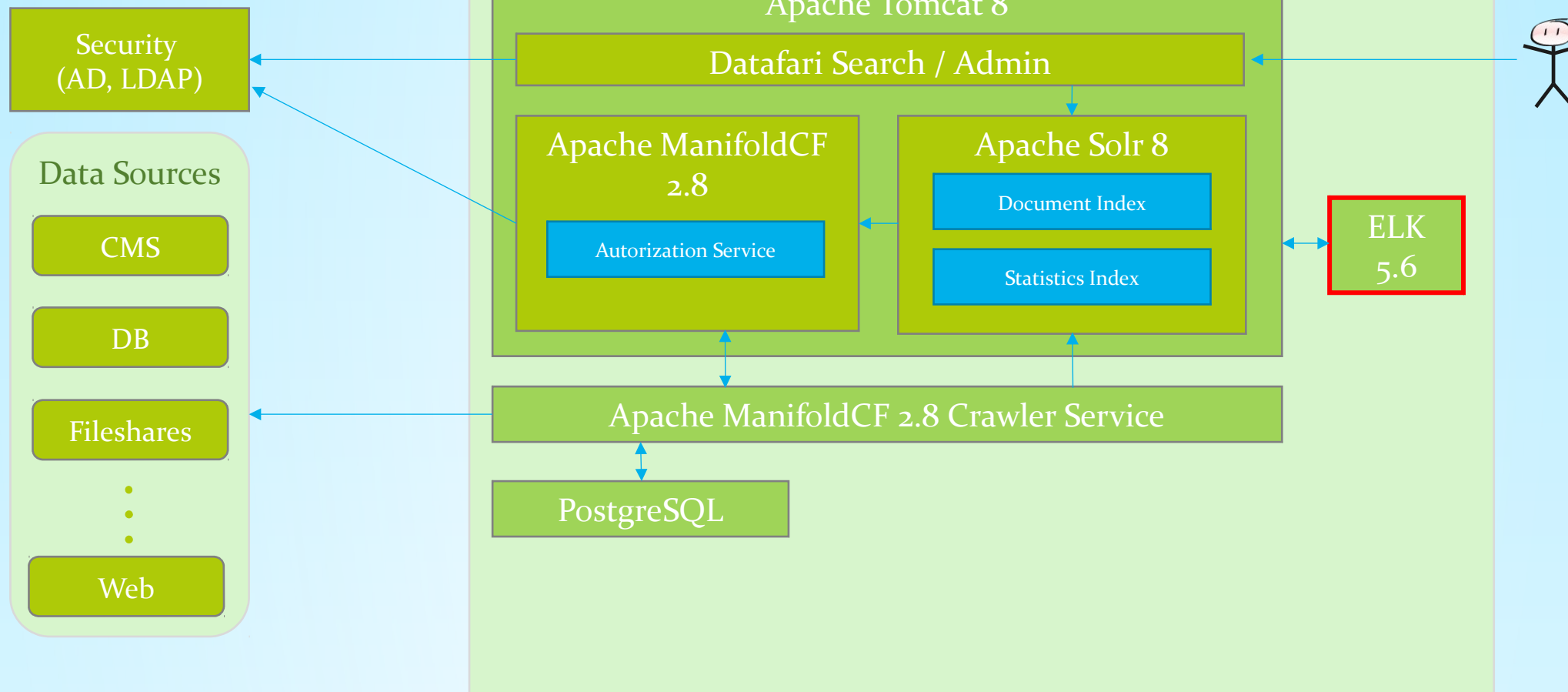


Apache Manifold CF for Datafari

- ☞ Manages data crawling
- ☞ Manages authentication
- ☞ Preconfigured integration with our Solr

Datafari 





Elasticsearch/Logstash/Kibana

- 🗨️ Elasticsearch (hence Lucene) based Logs analyser
- 🗨️ Large community (users/devs)
- 🗨️ Efficient/Reliable
- 🗨️ Scalable
 - High availability
 - Queries
 - Index volume

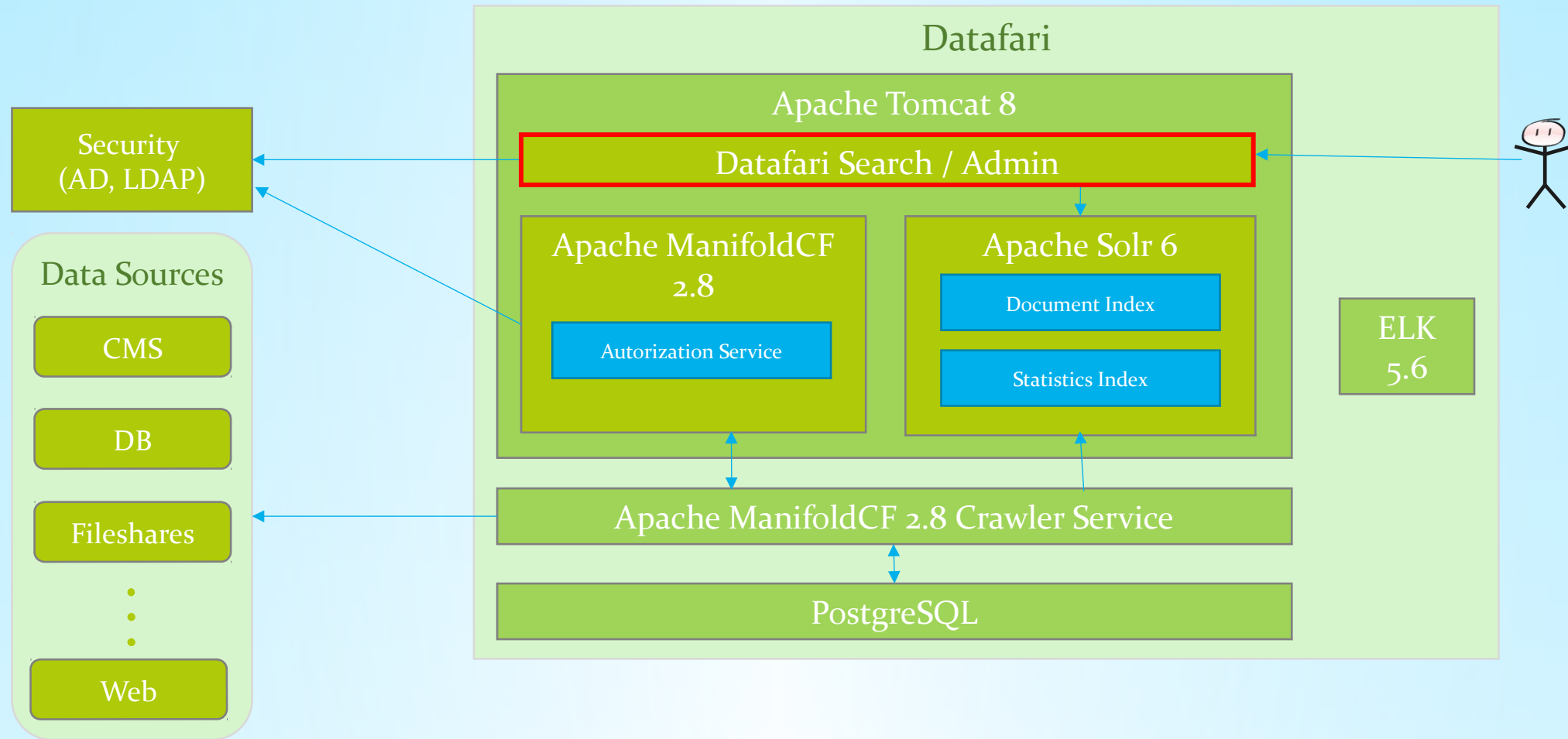


ELK for Datafari

- 🗨 Analytics core for Datafari
- 🗨 Dashboards for users search behavior
- 🗨 Dashboards for analysing the content
- 🗨 Dashboards for monitoring the system



Datafari 



Datafari Search

 Front-End

 User UI

- AjaxFrance Labs

 Authentication

 Querying the Solr

 Indexing user queries

 Admin UI

- Solr
- ManifoldCF
- Statistics
- Security

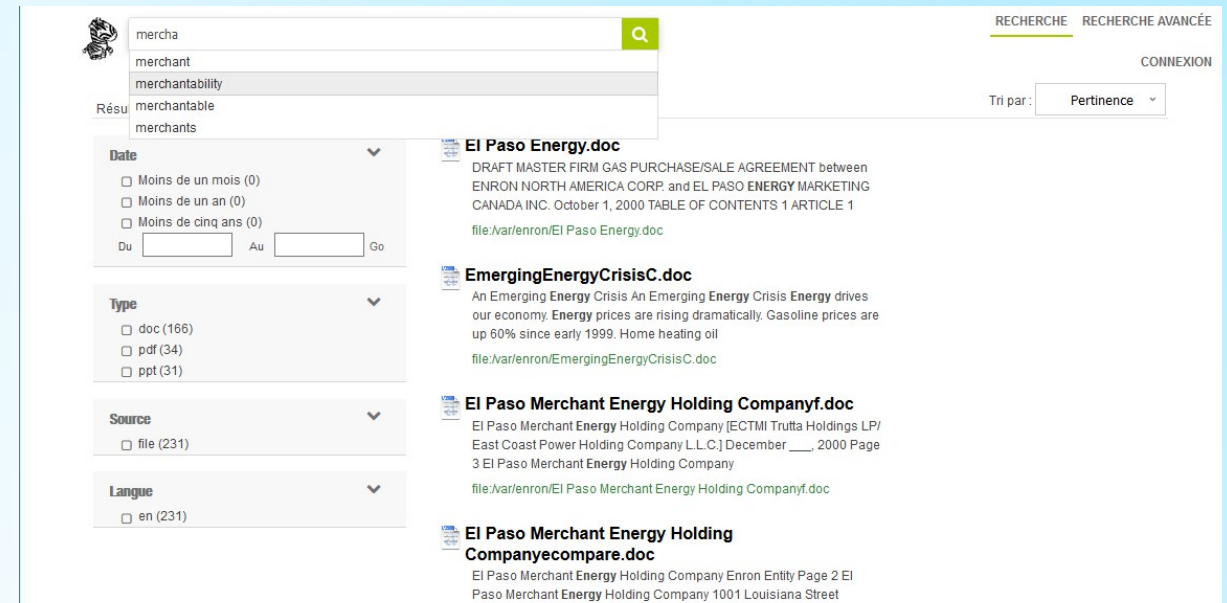
AjaxFranceLabs

🕒 Inspired by AjaxSolr

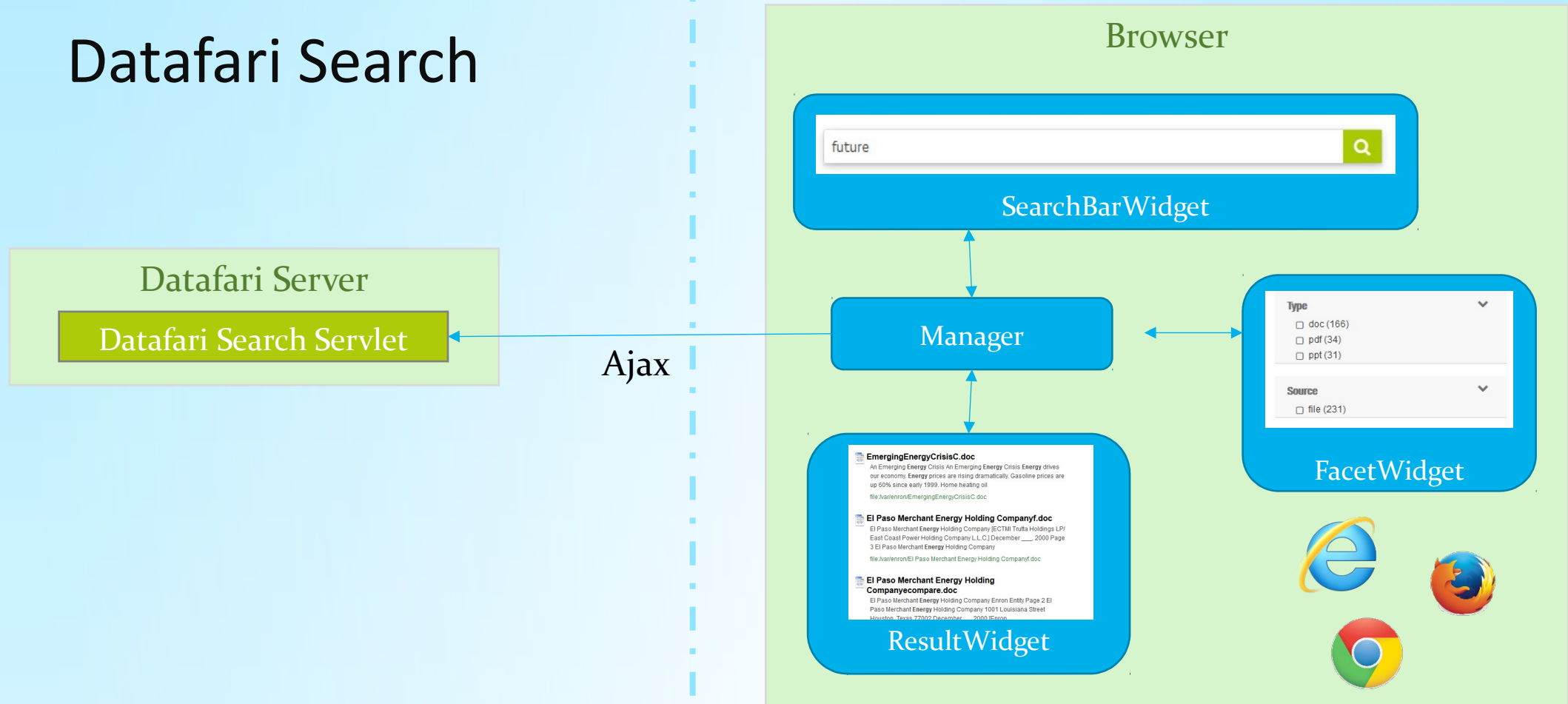
🕒 Javascript/Ajax client

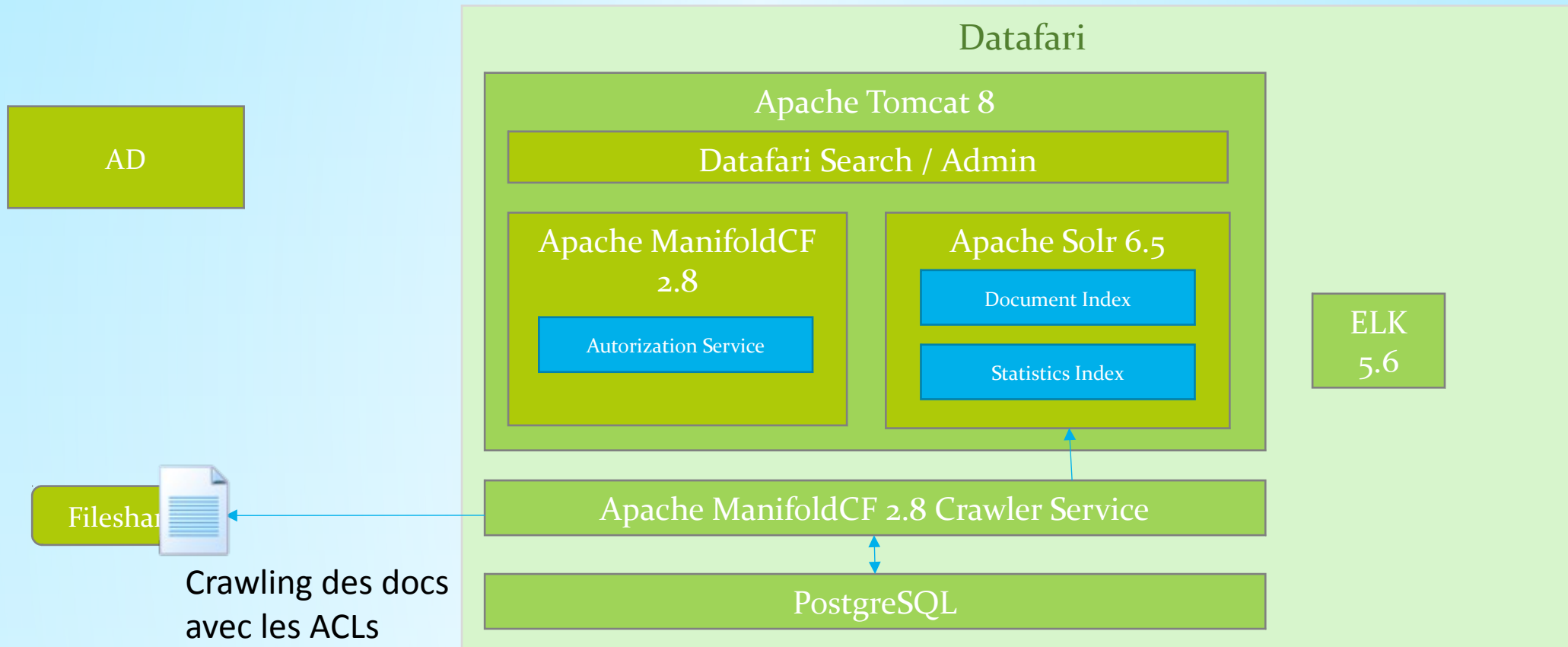
🕒 Provides several components:

- Manager: backend connection
- Widgets
 - Graphical/Logical components
 - (Advanced) Search
 - Facet
 - Geolocalisation (Based on OpenStreetMap)

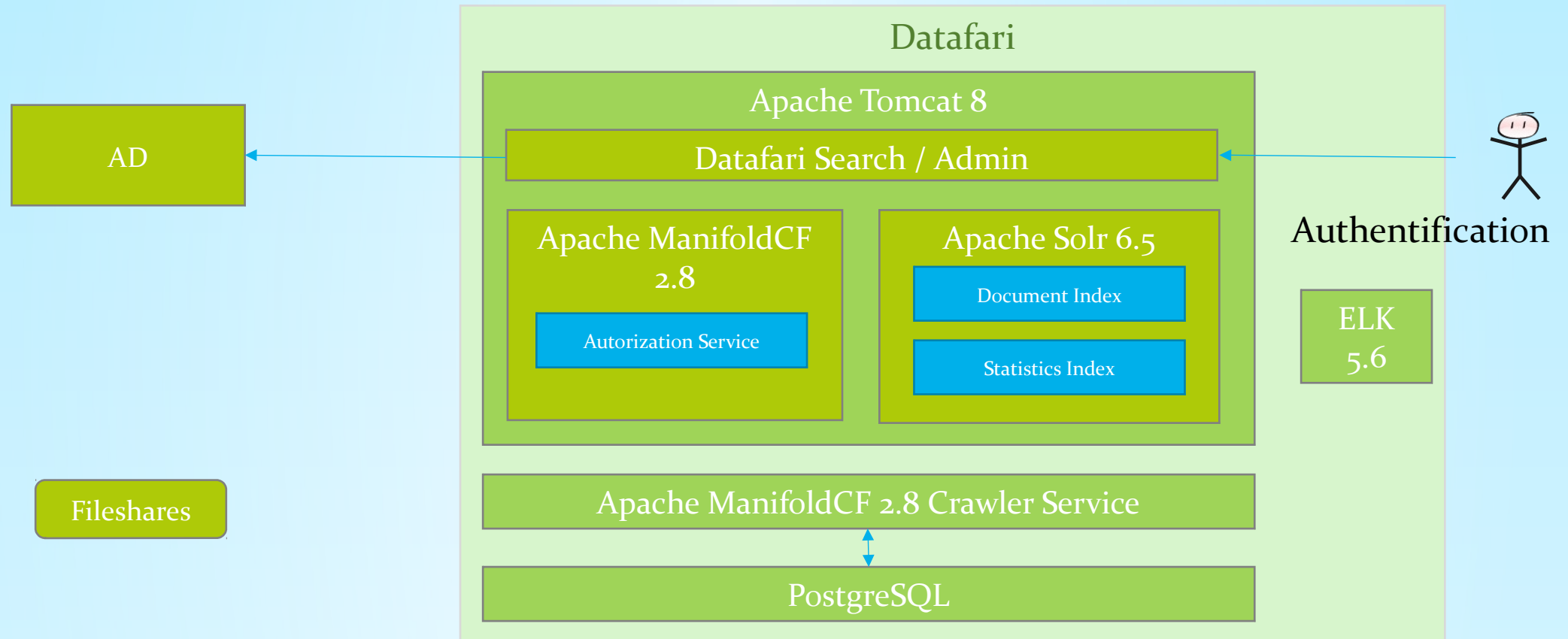


Datafari Search





Phase 1 : Indexation

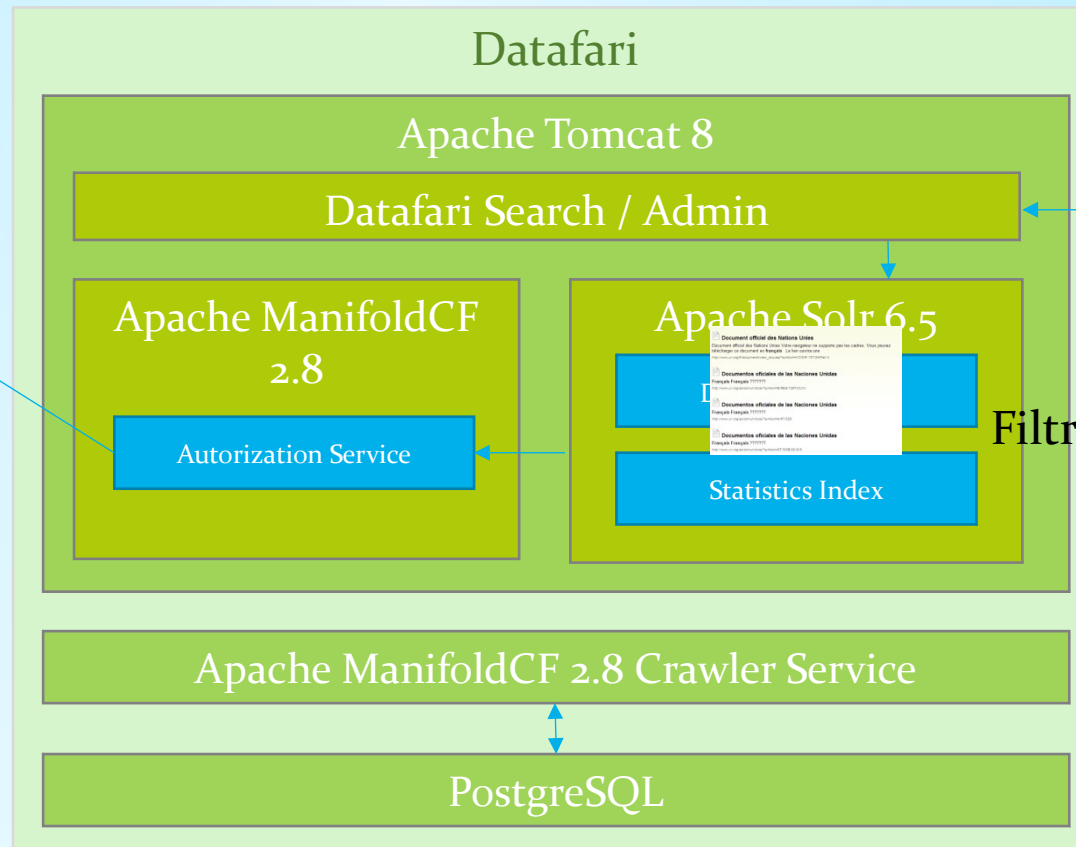


Phase 2 : Requêtage

AUTHORIZED:AD
TOKEN:AD:S-1-5-32-545
TOKEN:AD:S-1-1-0

Récupération des SIDs

Fileshares



Phase 2 : Requêtage

Use case

Nuclear industry maintenance

- Internal data for maintenance :
 - FileShare
 - Files + Metadata + ACLs on Oracle
 - 8M files
- ~ 250 employees



Use case

 IT Consulting company

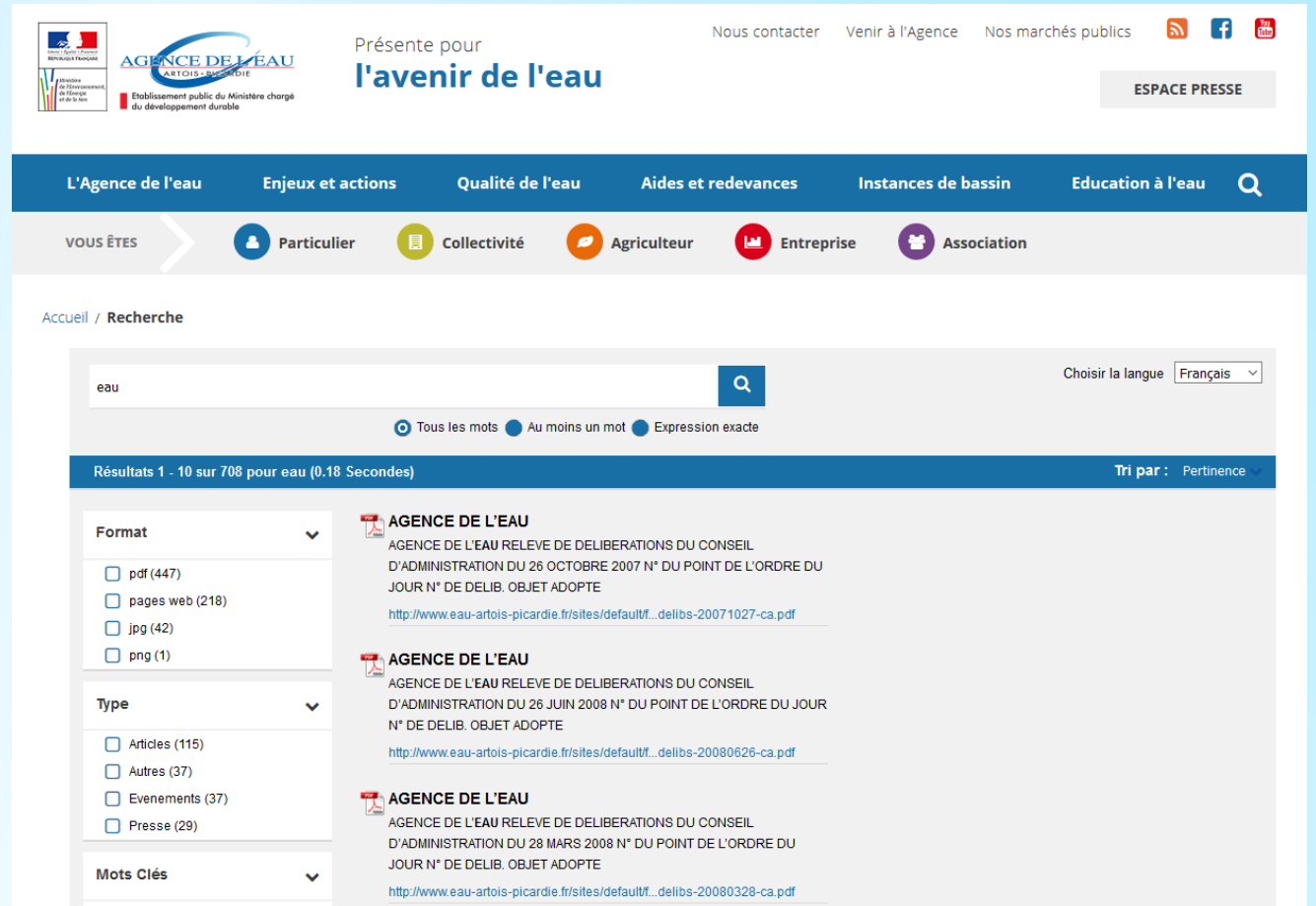
- Indexed Data :
 - Alfresco
 - Wiki, News Liferay
 - OTRS
 - Blog
- Integration in a Liferay portlet UI
- ~ 40 employees



Use case

AEAP

- Indexed data :
 - 1800 documents
 - Web connector (Drupal 7)
- UI on the webpage



The screenshot shows the website of the Agence de l'Eau Artois-Picardie. The header includes the agency's logo, the text "Présente pour l'avenir de l'eau", and navigation links: "Nous contacter", "Venir à l'Agence", "Nos marchés publics", and social media icons. A search bar is located in the top right corner.

The main navigation bar contains the following links: "L'Agence de l'eau", "Enjeux et actions", "Qualité de l'eau", "Aides et redevances", "Instances de bassin", and "Education à l'eau". Below this, a secondary navigation bar shows the user's role: "VOUS ÊTES" followed by icons for "Particulier", "Collectivité", "Agriculteur", "Entreprise", and "Association".

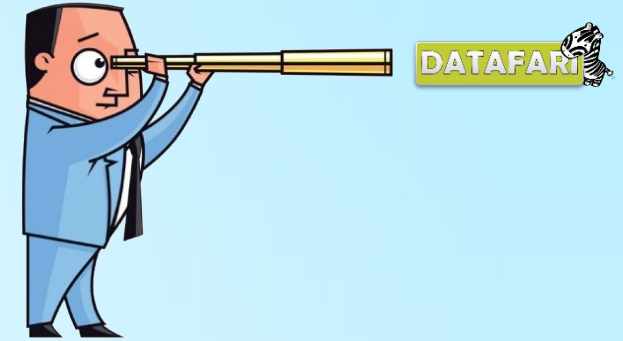
The search results page displays the search term "eau" and the number of results: "Résultats 1 - 10 sur 708 pour eau (0.18 Secondes)". The results are sorted by "Pertinence". The first three results are from the "AGENCE DE L'EAU" and are titled "AGENCE DE L'EAU RELEVÉ DE DÉLIBÉRATIONS DU CONSEIL D'ADMINISTRATION DU 26 OCTOBRE 2007 N° DU POINT DE L'ORDRE DU JOUR N° DE DÉLIB. OBJET ADOPTÉ". Each result includes a PDF icon and a link to the document.

Démo!!!



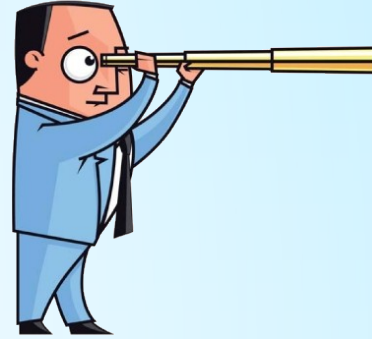
Technical Roadmap (1/2)

- 🗨️ AI applied to relevance
- 🗨️ Contextual search
- 🗨️ RHEL version
- 🗨️ Simplified admin
- 🗨️ Solr 8
- 🗨️ Always more documentation



Technical roadmap (2/2)

- 🗨️ Graph based search
- 🗨️ Smart autocomplete
- 🗨️ Search Cards



Where can I find Datafari

☞ Main hub: <http://www.datafari.com/en>

☞ Source code available on Github:

- <https://github.com/francelabs/datafari>

☞ Install packages for Debian 9 on :

- www.datafari.com

☞ Forum:

- <https://groups.google.com/forum/#!forum/datafari>

☞ Documentation on Confluence

- Technical and functional

☞ Tickets and releases on Jira

Want to follow Datafari ?



@francelabs
#datafari



<https://fr.linkedin.com/company/france-labs>



francelabs



Become a Datafarian ! 😊

💬 We are always open to suggestions

- “What about IA everywhere ...”

💬 Contribution

- What about a Klingon version ?!
- Windows version anyone ?

💬 Most important: your use cases and usage feedback

CONTACT

Don't hesitate to reach out to us for any info

Our corporate website: www.francelabs.com

Email: contact@francelabs.com

Tél: 09 72 43 72 85

Fax: 09 72 29 28 14

Address:

France Labs

Résidence du Grand Large – La Goelette, 2 rue de la Foux, 06800 Cagnes-sur-Mer

