Bringing Entreprise Search in the Big Data era with PCU

PCU

Unified, search-first Machine Learning platform targeted at business applications

PCU @ POSS 2017

Marc Dutoo, Smile
Dematerialization track

Overview

- **Why Big Data for Entreprise Search**
- **Demo !**
- **PCU introduction**
- **Questions**

The speaker

- **Marc Dutoo, R&D projects lead at Smile, the leading EU Open Source service provider**
- **PCU project coordinator, Data / API / Cloud expert**

PCU

Why Entreprise Search and Big Data

PCU @ POSS 2017

PARIS
OPEN
SOURCE
SUMMIT

Digital Entreprise ?

Entreprise Search, a powerful asset to make your documents go digital !

But out of fashion ?

+ Big Data

= PCU Entreprise Search open source by SMILE I.T IS OPEN

https://www.smile.eu/fr/technologies/pcu-enterprise-search

images courtesy LinkedIn.com

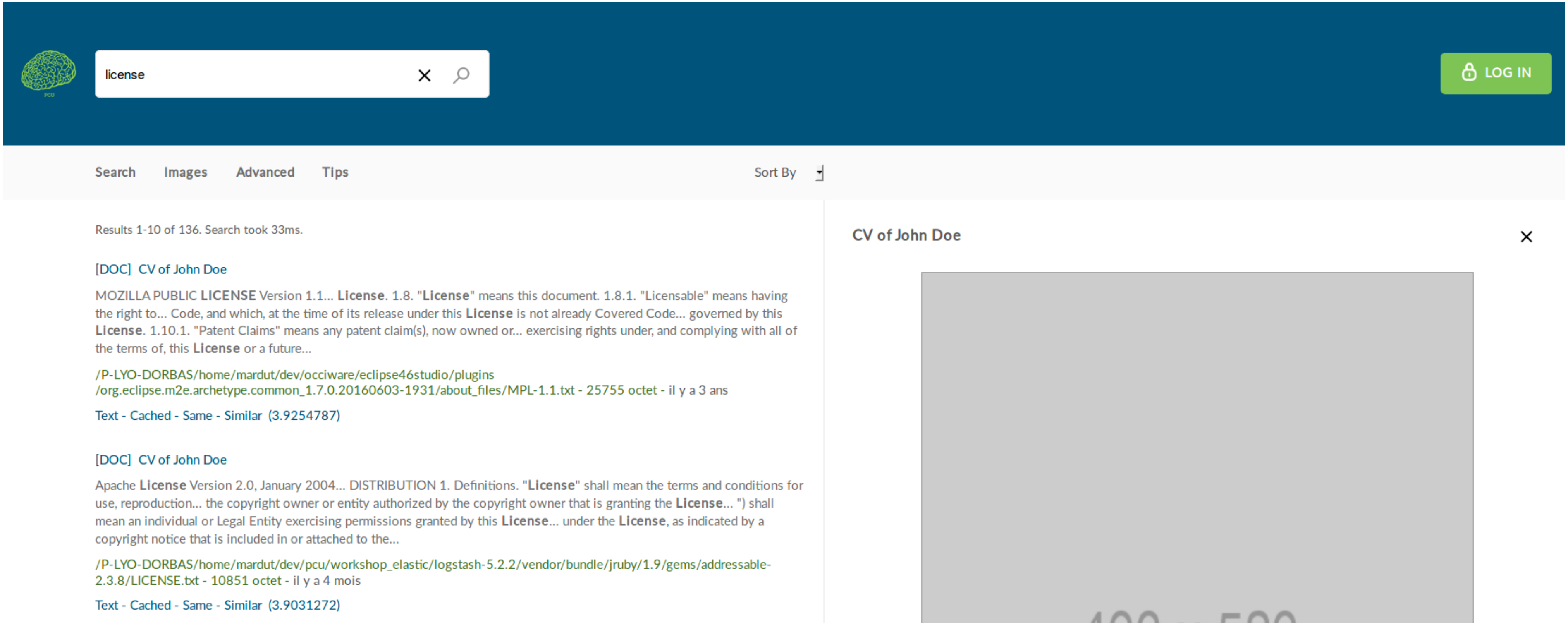
PCU

Demo !

PCU @ POSS 2017

PARIS OPEN SOURCE SUMMIT

Entreprise Search (démo)



Results 1-10 of 136. Search took 33ms.

[DOC] CV of John Doe

MOZILLA PUBLIC **LICENSE** Version 1.1... **License**. 1.8. "**License**" means this document. 1.8.1. "Licensable" means having the right to... Code, and which, at the time of its release under this **License** is not already Covered Code... governed by this **License**. 1.10.1. "Patent Claims" means any patent claim(s), now owned or... exercising rights under, and complying with all of the terms of, this **License** or a future...

/P-LYO-DORBAS/home/mardut/dev/occiware/eclipse46studio/plugins
/org.eclipse.m2e.archetype.common_1.7.0.20160603-1931/about_files/MPL-1.1.txt - 25755 octet - il y a 3 ans

Text - Cached - Same - Similar (3.9254787)

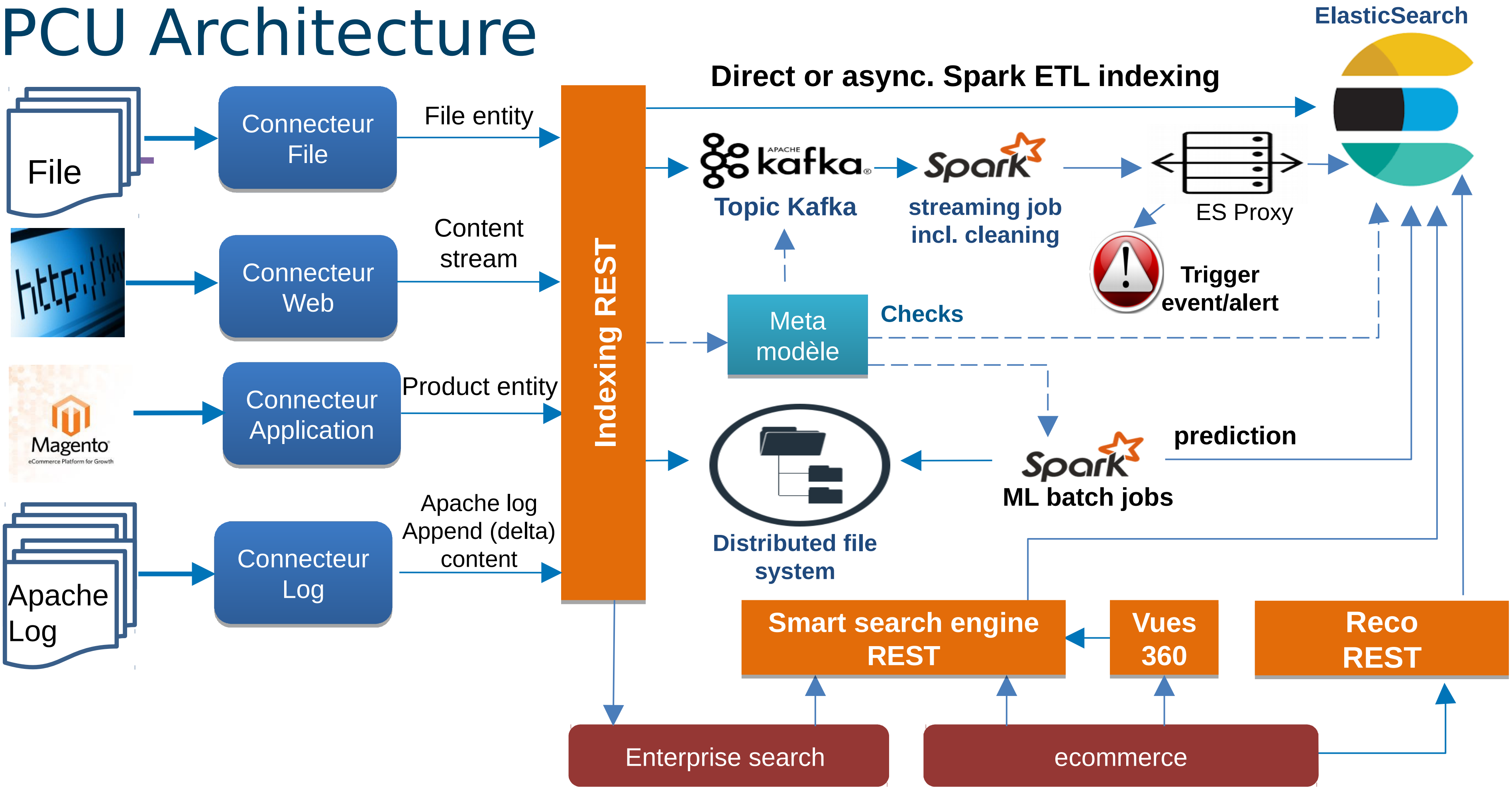
[DOC] CV of John Doe

Apache **License** Version 2.0, January 2004... DISTRIBUTION 1. Definitions. "**License**" shall mean the terms and conditions for use, reproduction... the copyright owner or entity authorized by the copyright owner that is granting the **License**... ") shall mean an individual or Legal Entity exercising permissions granted by this **License**... under the **License**, as indicated by a copyright notice that is included in or attached to the...

/P-LYO-DORBAS/home/mardut/dev/pcu/workshop_elastic/logstash-5.2.2/vendor/bundle/jruby/1.9/gems/addressable-2.3.8/LICENSE.txt - 10851 octet - il y a 4 mois

Text - Cached - Same - Similar (3.9031272)

PCU Architecture



ElasticSearch

Direct or async. Spark ETL indexing

File

Connecteur File

File entity

Connecteur Web

Content stream

Magento

eCommerce Platform for Growth

Connecteur Application

Product entity

Apache Log

Connecteur Log

Apache log Append (delta) content

Indexing REST

APACHE kafka

Topic Kafka

Spark

streaming job incl. cleaning

ES Proxy

Trigger event/alert

Meta modèle

Checks

Distributed file system

Spark

ML batch jobs

prediction

Smart search engine REST

Vues 360

Reco REST

Enterprise search

ecommerce

PCU

Entreprise Search (WP7), avec Spark ETL indexing (WP2) (démo)

- **Entreprise Search : “qui peut le plus peut le moins”**
  - le produit d’appel “pied dans la porte” de PCU pour élargir son audience au-delà des early adopters
- **… MAIS pas seulement !**
  - plus tard, il héritera des fonctionnalités de recherche intelligentes mises au point pour le e-commerce (en “trickle-down”)
  - dès à présent, il bénéficie de l’intégralité de l’architecture de PCU, qu’il valide
    - **pipeline d’indexation alternatif sur YAML-configured Spark ETL** :
      - à la volée (mode streaming), configuration simple (YAML)
      - asynchrone et scalable grâce à Kafka (files partitionnées)
      - transformation de données en Spark (tout comme le ML)
    - validation des données vis à vis de schémas centralisés

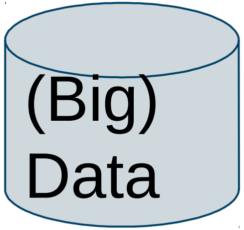
PCU

Tomorrow : PCU Introduction

PCU @ POSS 2017

PARIS OPEN SOURCE SUMMIT

PCU - the problem



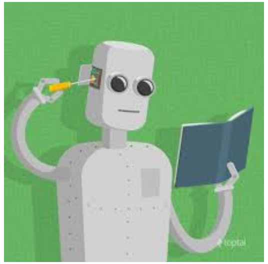
(Big) Data

=

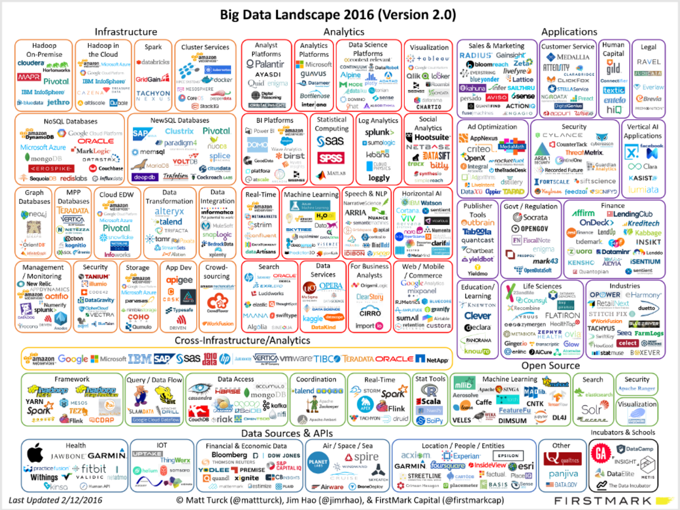
BUT you need...

Data scientist

Devops



Machine Learning



... how can (very) small companies take advantage of it ?

images courtesy thinglink.com, lebigdata.fr, webengage.com, toptal.com, mattturk.com

Factsheet - Unified Knowledge Platform

- 6 partners, 36 man-year over 2017-2019, sponsored by the French ministry of Industry & région Île de France
- In order to democratize Big Data, so that every company will be able to add value to its own core business thanks to its existing data
    - The Big Data / Machine Learning / semantic module to enrich any business application
    - **… Unified, search-first Machine Learning platform targeted at business applications.**
- As showcased in 2 use cases :
    - E-commerce (up to digital in store) & B2B
    - Enterprise search
- Thanks to:
    - A factory of Machine Learning and Semantics-enriched search engines
    - state-of-the-art and new algorithms analyzing user behaviour
    - end-to-end event-driven data processing workflow
    - an open source, best-of-breed, unified, flexible and extensible approach

Partners and stakeholders

Smile : coordinator, architecture, ecommerce

Paris 13 : Machine Learning, semantics

ESILV : pipeline, semantics

Proxem : text & opinion mining, B2B

Wallix : enterprise search experience

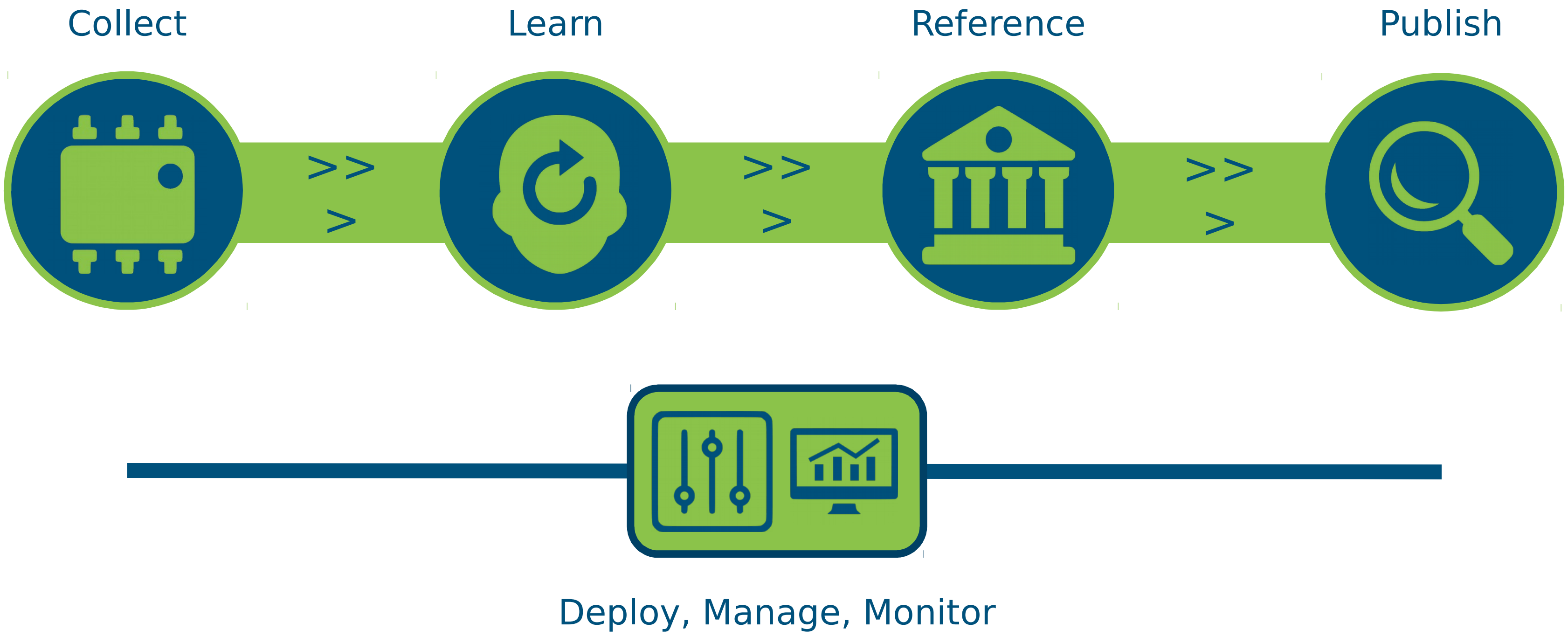
Armadillo : integration & mgmt API & UI

Financial sponsors : BPI, IDF

Cluster : System@tic

PCU

Overall architecture



PCU

Target outputs

- **Generic platform**
  - Unified, flexible, extensible, best-of-breed-based, API-managed
  - Along with a set of standard connectors, data pipeline elements, and Machine Learning (ML) and text mining algorithms
- **Use cases and products**
  - E-commerce (product, deployed at Smile early adopter customers), B2B (deployed at Smile & Proxem)
  - Enterprise Search (product, deployed at each partner's)
- **Open Source Ecosystem**
  - Ties with integrated technical components' communities as well as derived business-specific products
  - Home of platform examples, tryout and adoption

Year 1 outputs

- **Business requirements, up to Machine Learning prototypes**
  - Search, Ecommerce (B2C), CRM (B2B), including 10GB+ data sets
  - Data analysis, up to ML prototypes on Spark + Jupyter : reco, coocs...
- **Architecture and development**
  - State of the art, POCs (ElasticSearch, Solr, Spark), technical architecture
  - Semantic platform architecture, topic detection algorithm
  - YAML-configured ETL pipeline on Spark (prototype)
  - 360 View & A/B testing prototypes
  - Enterprise search demo (API, indexing, crawler, metadata extractor, UI)
- **Project setup**
  - Collaboration, communication
  - Tools : Github, first shared data and Big Data / ML components Cloud, Spark Machine Learning dockerized environment...

PARIS OPEN SOURCE SUMMIT

https://pcu-consortium.github.io/

https://twitter.com/PCUConsortium

Contact : marc.dutoo@smile.fr

https://www.smile.eu/fr/technologies/pcu-enterprise-search

Questions ?

Thanks for your attention !

https://www.smile.eu/fr/technologies/pcu-enterprise-searc

SMILE I.T IS OPEN

UNIVERSITE PARIS 13

DE VINCI RESEARCH CENTER

proxem

armadillo

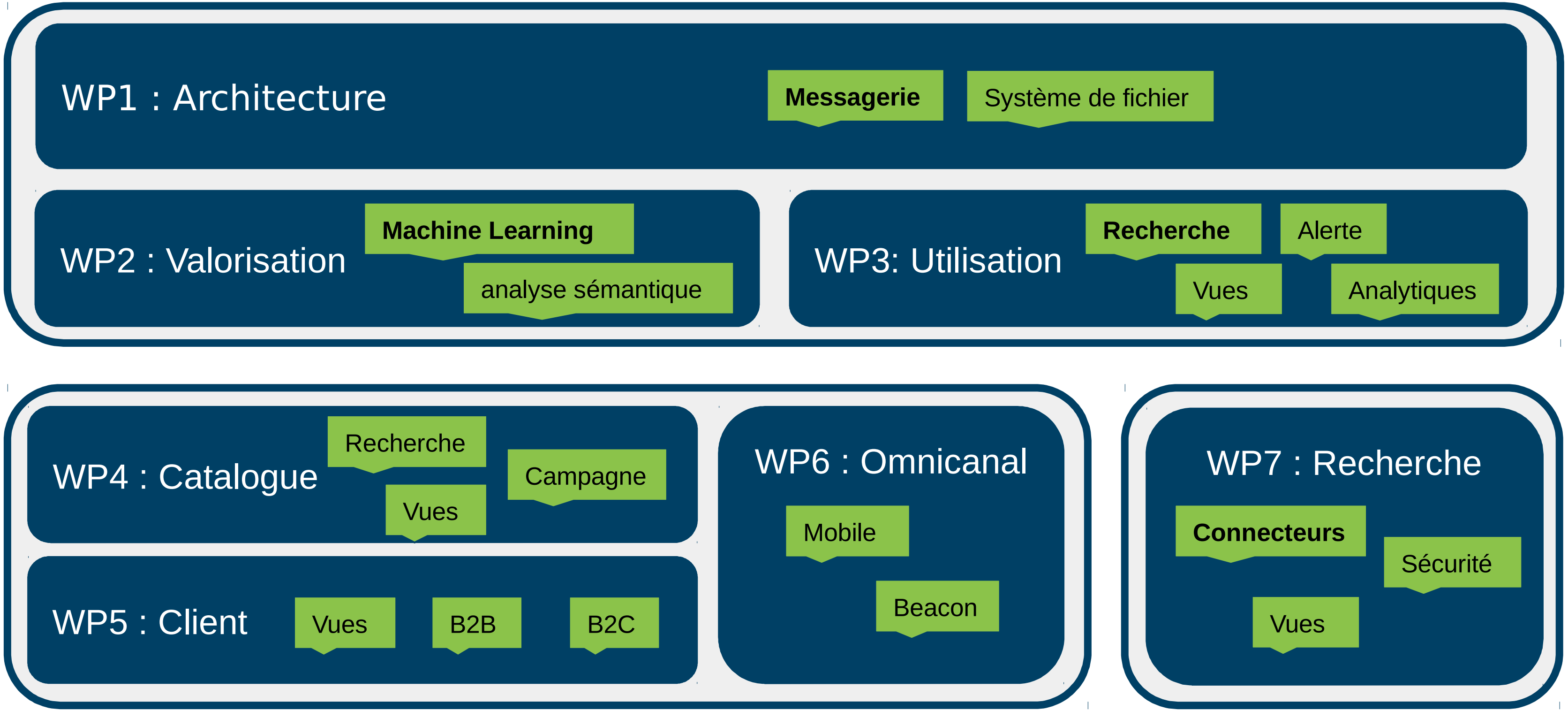
WALLIX TRACE, AUDIT & TRUST

Systematic Paris Region Digital Ecosystem

bpifrance

îledeFrance

Vue d’ensemble des Work Packages techniques



WP1 : Architecture

Messagerie

Système de fichier

WP2 : Valorisation

Machine Learning

analyse sémantique

WP3: Utilisation

Recherche

Alerte

Vues

Analytiques

WP4 : Catalogue

Recherche

Campagne

Vues

WP5 : Client

Vues

B2B

B2C

WP6 : Omnicanal

Mobile

Beacon

WP7 : Recherche

Connecteurs

Sécurité

Vues

PCU

Conclusion

Revue an 1

Conclusion

- **Fait en 2017 :**
    - ○ Besoins prototypés
    - ○ Architecture et fondations R&D
    - ○ Prototypes techniques v1
    - ○ Solution Entreprise Search v1
- **Prévu en 2018 :**
    - ○ Gestion des modèles de données et configuration générique et dynamique
    - ○ Refactoring du framework de connecteurs
    - ○ Algorithmes recherche sémantique, NLP, recommandation