

Bringing Enterprise 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





PCU

Overview

- **Why Big Data for Enterprise 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 Enterprise Search and Big Data

PCU @ POSS 2017

PARIS
**OPEN
SOURCE**
SUMMIT



PCU

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



But out of fashion ?



Big Data



PCU Enterprise Search



open source

by



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



PCU

Demo !

PCU @ POSS 2017

PARIS
OPEN
SOURCE
SUMMIT



Entreprise Search (d  mo)



× 🔍

 LOG IN

Search Images Advanced Tips

Sort By 

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)

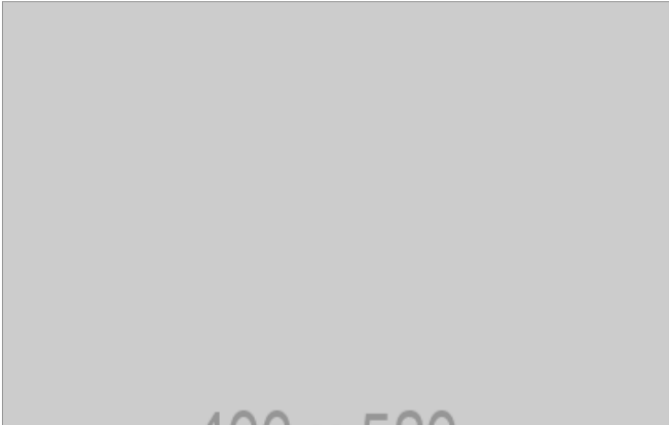
[DOC] CV of John Doe

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...

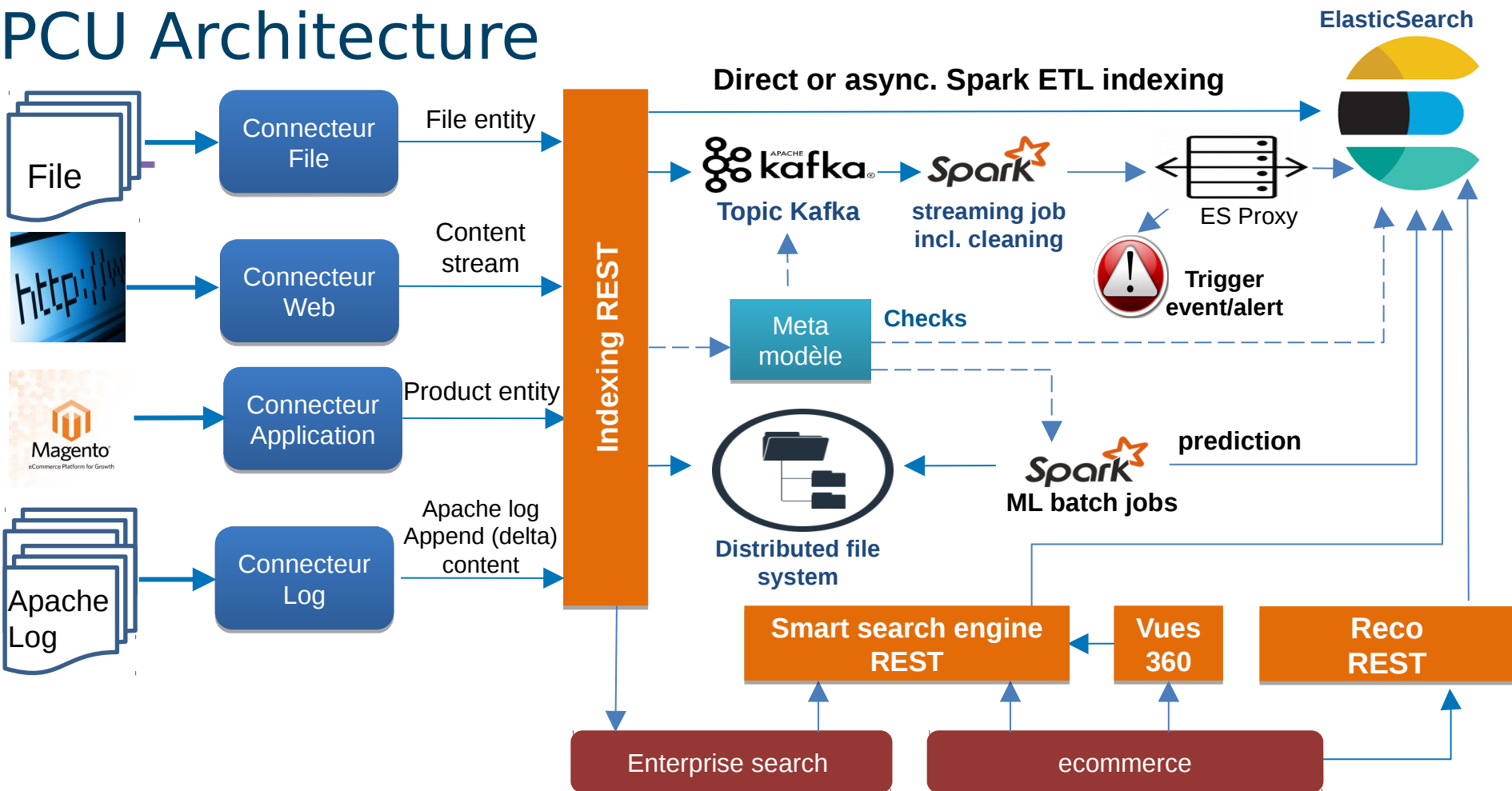
/P-LYO-DORBAS/home/mardut/dev/pcu/workshop_elastic/logstash-5.2.2/vendor/bundle/jruby/1.9/gems/addressable-2.3.8/LICENSE.txt - 10851 octet - il y a 4 mois

Text - Cached - Same - Similar (3.9031272)

CV of John Doe



PCU Architecture





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

- **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 via à via de schémas centralisés



PCU

Tomorrow : PCU Introduction

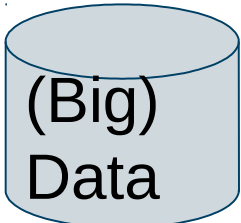
PCU @ POSS 2017

PARIS
**OPEN
SOURCE**
SUMMIT



PCU

PCU - the problem



=



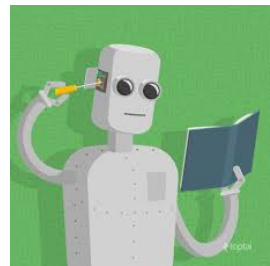
BUT you need...



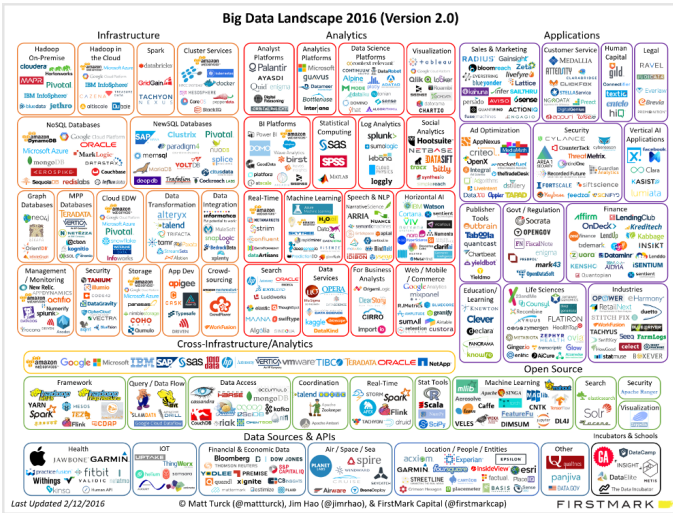
Data scientist



Devops



Machine Learning



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



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



PCU

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

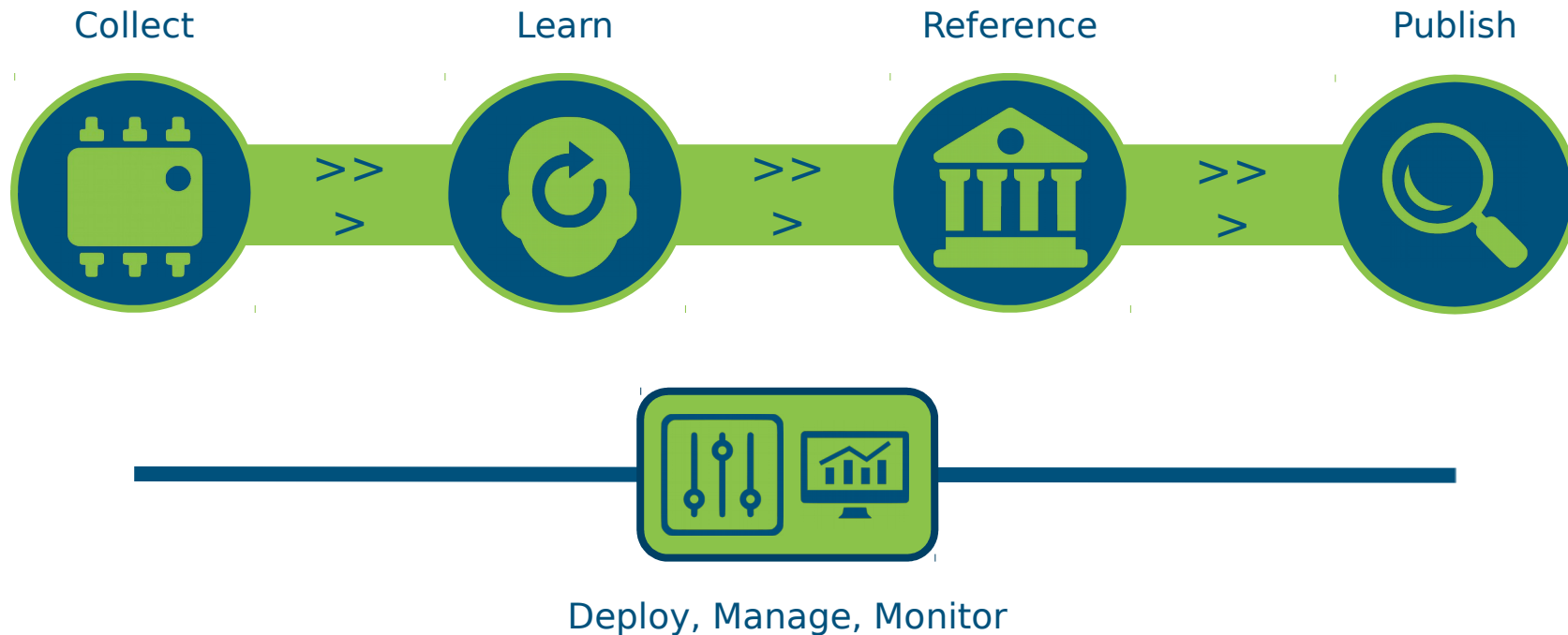
Cluster : System@tic





PCU

Overall architecture





Target outputs

● Generic platform



A distributed streaming platform



elastic

- 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...

Questions ?

<https://pcu-consortium.github.io/>
<https://twitter.com/PCUConsortium>

Contact : marc.dutoo@smile.fr

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



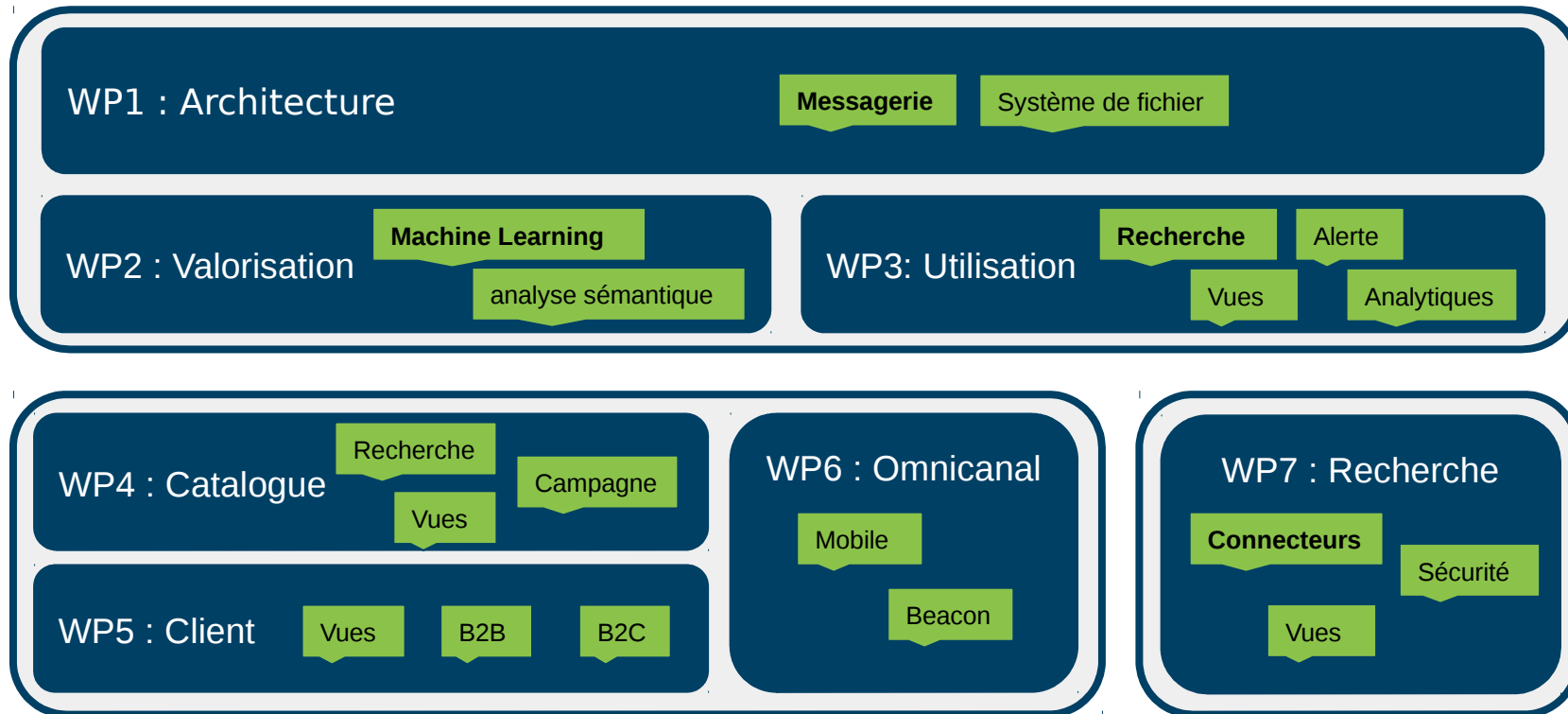
Thanks for
your attention !

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





Vue d'ensemble des Work Packages techniques





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