Big Data for Labour Market Intelligence

Jour 1
Présentation du système et résultats

Alessandro Vaccarino - Mauro Pelucchi

Juin 2021

Thèmes

1. Objectif et contexte
2. Défis
    1. Parties prenantes
    2. L'architecture fonctionnelle
    3. Techniques d'ingestion de données
    4. Pipeline de traitement des données
    5. Techniques de classification

Thèmes

1. **Objectif et contexte**
2. Défis
    1. Parties prenantes
    2. L'architecture fonctionnelle
    3. Techniques d'ingestion de données
    4. Pipeline de traitement des données
    5. Techniques de classification

Contexte

Un marché du travail **en constante évolution** :

- Numérisation des professions
- Pertinence des compétences non techniques
- Internationalisation
- Émergence de nouvelles professions et compétences
- Travail flexible et à distance
- Impact de la pandémie de Covid-19
- …

Nous avons besoin de *quelque chose* qui puisse nous aider à surveiller et à analyser l'évolution du marché du travail, pour aider les décideurs à prendre **les bonnes décisions au bon moment**

Ce que nous avons / ce dont nous avons besoin

Nous avons déjà des **statistiques officielles**, qui sont :

- *Représentatives*
- *Fortes* en termes de valeur

Mais nous pouvons bénéficier d'**informations supplémentaires et complémentaires** qui pourraient être :

- *Rapides*, pour suivre ce qui se passe maintenant (par exemple, l'analyse de l'impact de la Covid-19)
- *Granulaires* et *conformes* aux termes réels et actuels du marché, afin de saisir les tendances émergentes et d'analyser ce que les entreprises recherchent réellement.

Comment trouver une source d'information similaire et complémentaire ?

En utilisant le **marché du travail en ligne**

Pourquoi le marché du travail en ligne

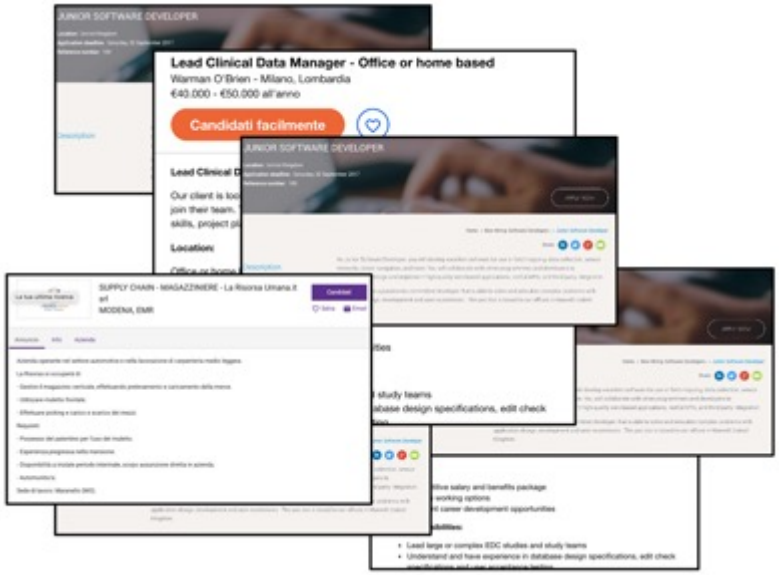
C'est la représentation exacte de ce que les entreprises recherchent à une période donnée :

- À jour : les entreprises publient une annonce lorsqu'elles ont réellement besoin de recruter.
- Détaillée : une annonce décrit aussi bien que possible le besoin spécifique, en termes de :
  - Profession requise
  - Exigences (compétences, expérience, niveau d'éducation...)
  - Contexte de travail (lieu, contrat, secteur, horaires de travail...)
- Conforme à la réalité : les termes du marché sont utilisés, tant pour la profession que pour les compétences. Cela permet d'identifier la terminologie émergente adoptée par le marché

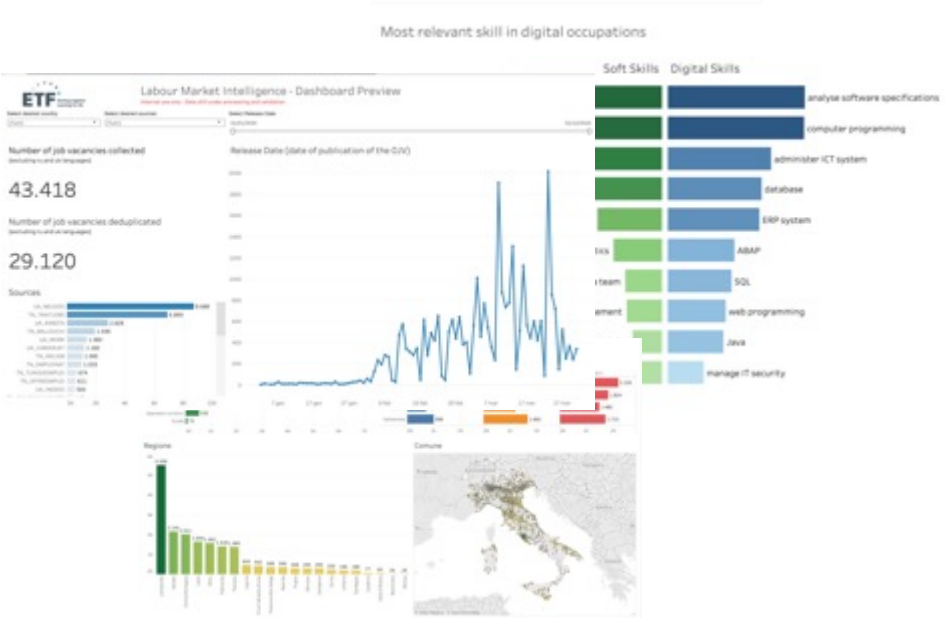
Il serait formidable d'utiliser ces informations en plus de mieux et plus profondément comprendre l'évolution du marché du travail dans un pays donné, même par rapport à d'autres pays

Notre objectif

Transformer les annonces d'emploi en ligne...



...en connaissance et analyses

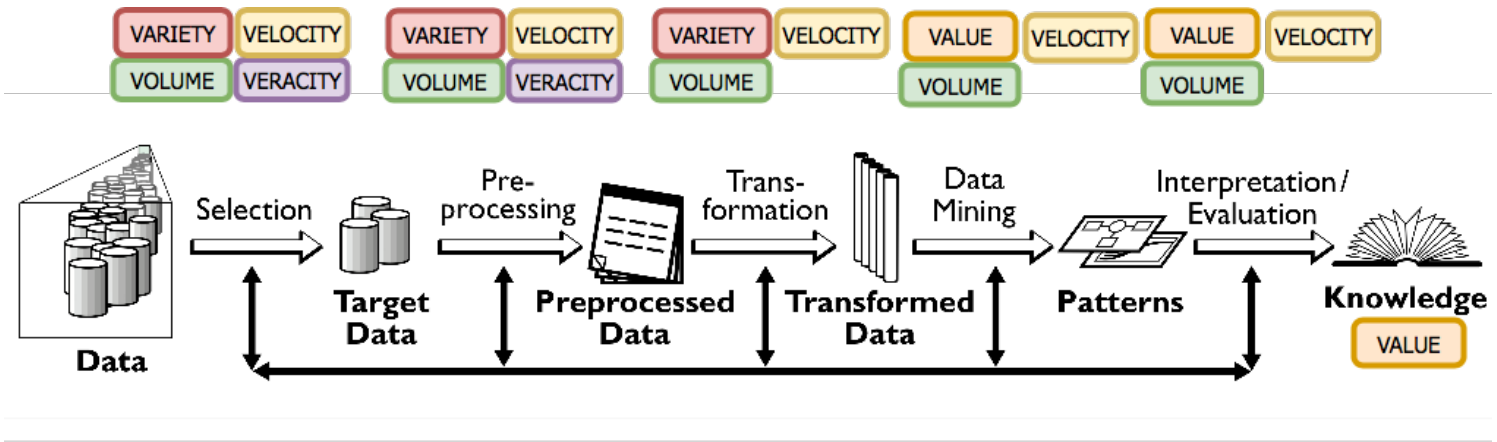


Défis

- Traiter une énorme **quantité** de données en temps quasi réel
- Données provenant du web → Nécessité de détecter et de réduire le **bruit**
- Environnement **multilingue**
- Nécessité de se référer aux **normes de classification**
- Trouver un moyen de **résumer et de présenter** un scénario vaste et complexe

Contexte méthodologique

KDD - Fayyad, 1997



VARIETY

Données non structurées (texte brut à traiter)

Données en temps réel

VELOCITY

Quantité énorme de données (téraoctets)

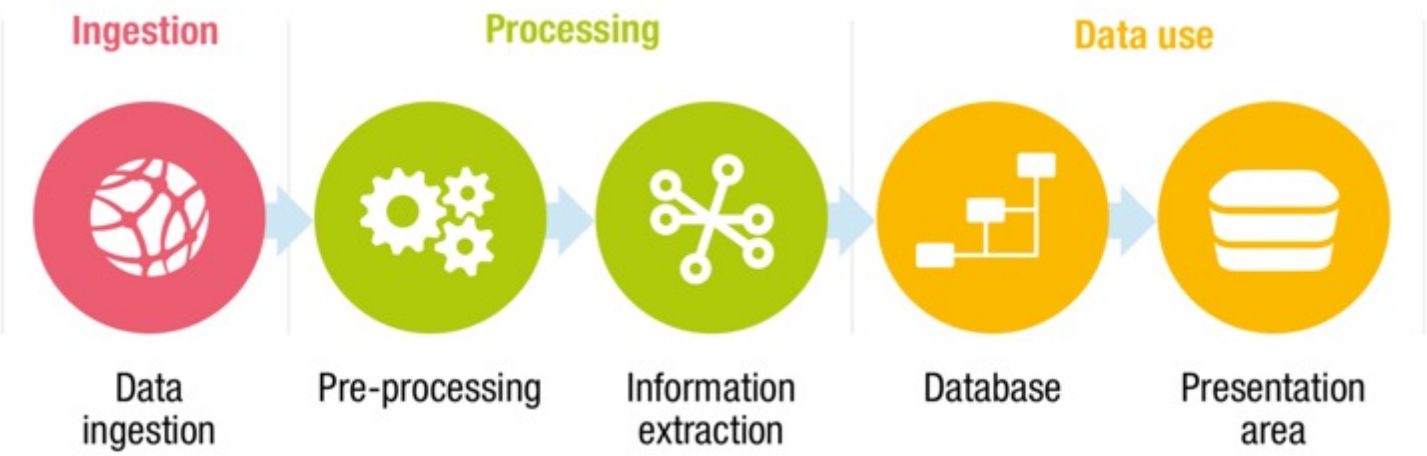
VOLUME

Les données sont surchargées, non contrôlées

VERACITY

Notre approche

KDD 4 LMI



Quelques résultats

- Skillspanorama - Les compétences dans les offres d'emploi en ligne
  - https://skillspanorama.cedefop.europa.eu/en/indicators/skills-online-vacancies
- Skills OVATE
  - https://www.cedefop.europa.eu/en/data-visualisations/skills-online-vacancies
- ETF - Big Data pour l’information sur le marché du travail
  - Tunisie
  - Ukraine

Thèmes

1. Objectif et contexte
2. Défis
    1. **Parties prenantes**
    2. L'architecture fonctionnelle
    3. Techniques d'ingestion de données
    4. Pipeline de traitement des données
    5. Techniques de classification

Parties prenantes

Chef de projet

Utilisateurs clés

Experts dans le domaine

Utilisateurs finaux

Chef de projet

- ETF
  - Dirige le projet avec le comité de pilotage
  - Définit la portée du projet
  - Définit les organisations clés
  - Entretient des relations avec les parties prenantes de l'UE
  - Fournit des conseils

Utilisateurs clés

- ETF, Burning Glass
  - Définissent les besoins
  - Contrôlent la qualité du projet
  - Contribuent à l'élaboration du projet
  - Gèrent les inventaires
  - Valident le flux de données global et la méthodologie

Experts dans le domaine

- Experts pays internationaux
  - Fournissent les connaissances et l'expertise
  - Exécutent l’inventaire
  - Comprennent la langue/les termes de leur contexte
  - Évaluent l'exactitude des résultats
  - Testent le produit
  - Fournissent des informations

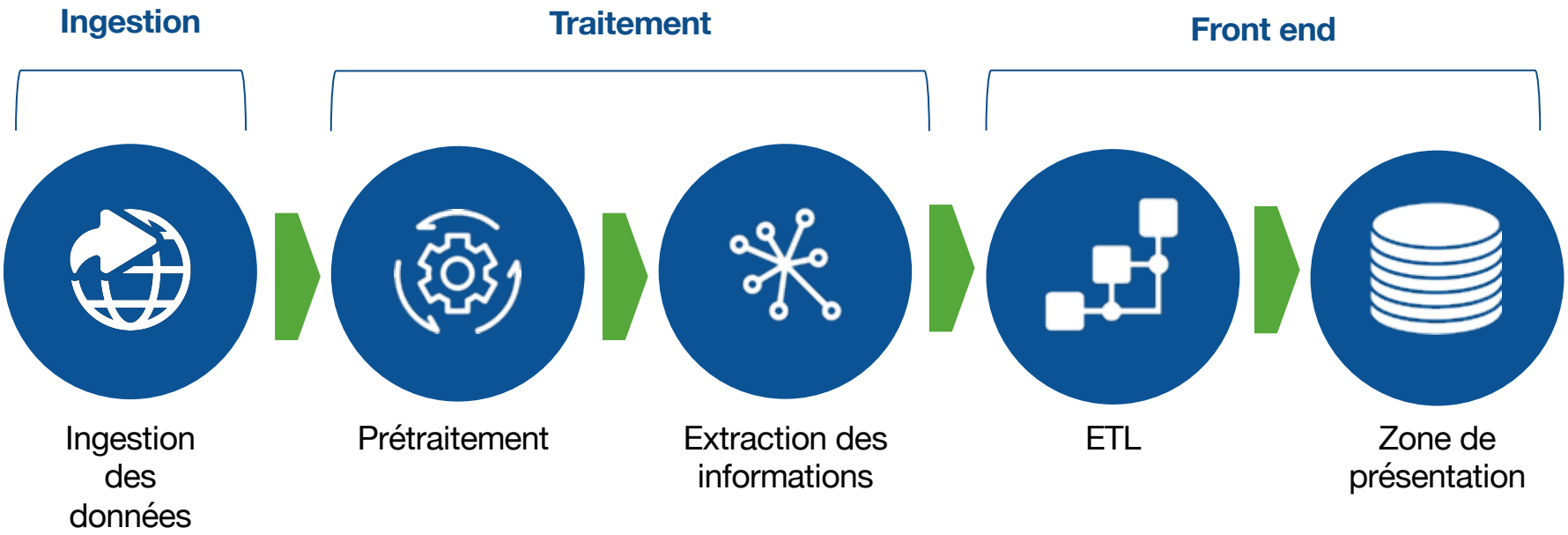
Utilisateurs finaux

- Décideurs et utilisateurs professionnels
  - (Visuel) Explorent l'ensemble des données, les analyses et les données agrégées
  - Définissent de nouveaux processus d'analyse
  - Produisent une narration de données
  - Prennent des décisions en explorant les données
- Data Scientists
  - Appliquent de nouveaux modèles d'apprentissage automatique et des techniques d'IA
  - Extraient de nouvelles informations à partir des données
  - Appliquent une modélisation avancée des données à l'ensemble des données.
- Analystes de données
  - Interprètent les données et les transforment en informations
  - Identifient les modèles et les tendances
  - Extraient et analysent des données agrégées
  - Publient et partagent leur analyse

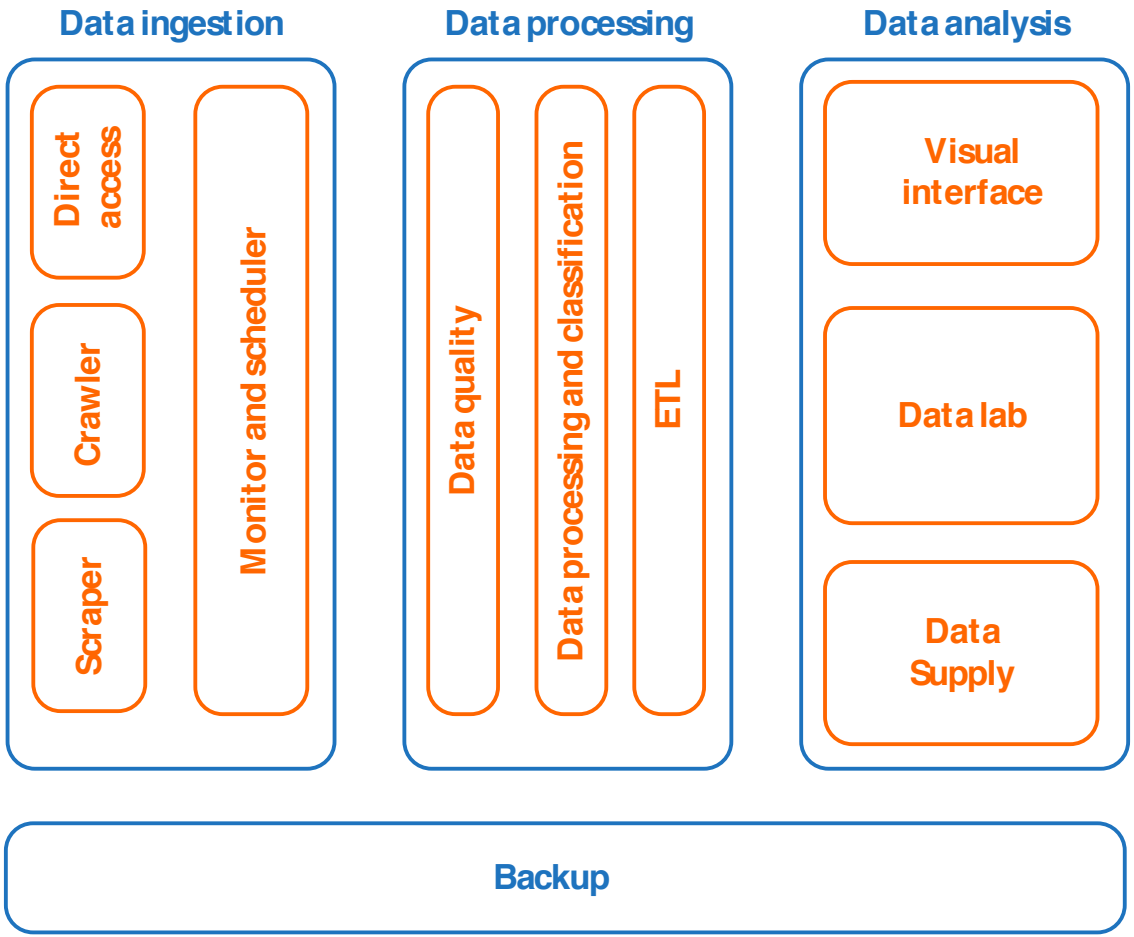
Thèmes

1. Objectif et contexte
2. Défis
    1. Parties prenantes
    2. **L'architecture fonctionnelle**
    3. Techniques d'ingestion de données
    4. Pipeline de traitement des données
    5. Techniques de classification

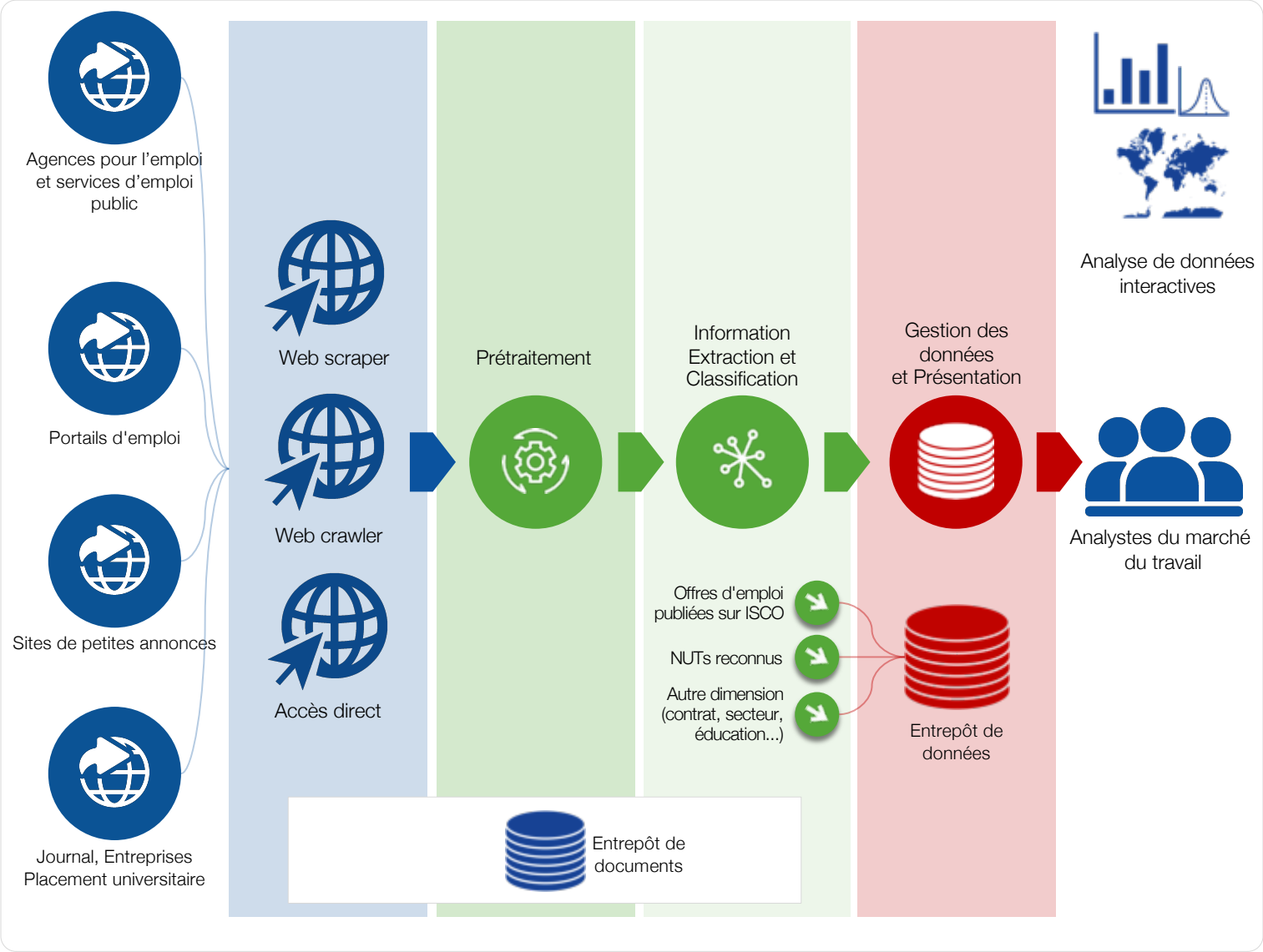
Flux de données global



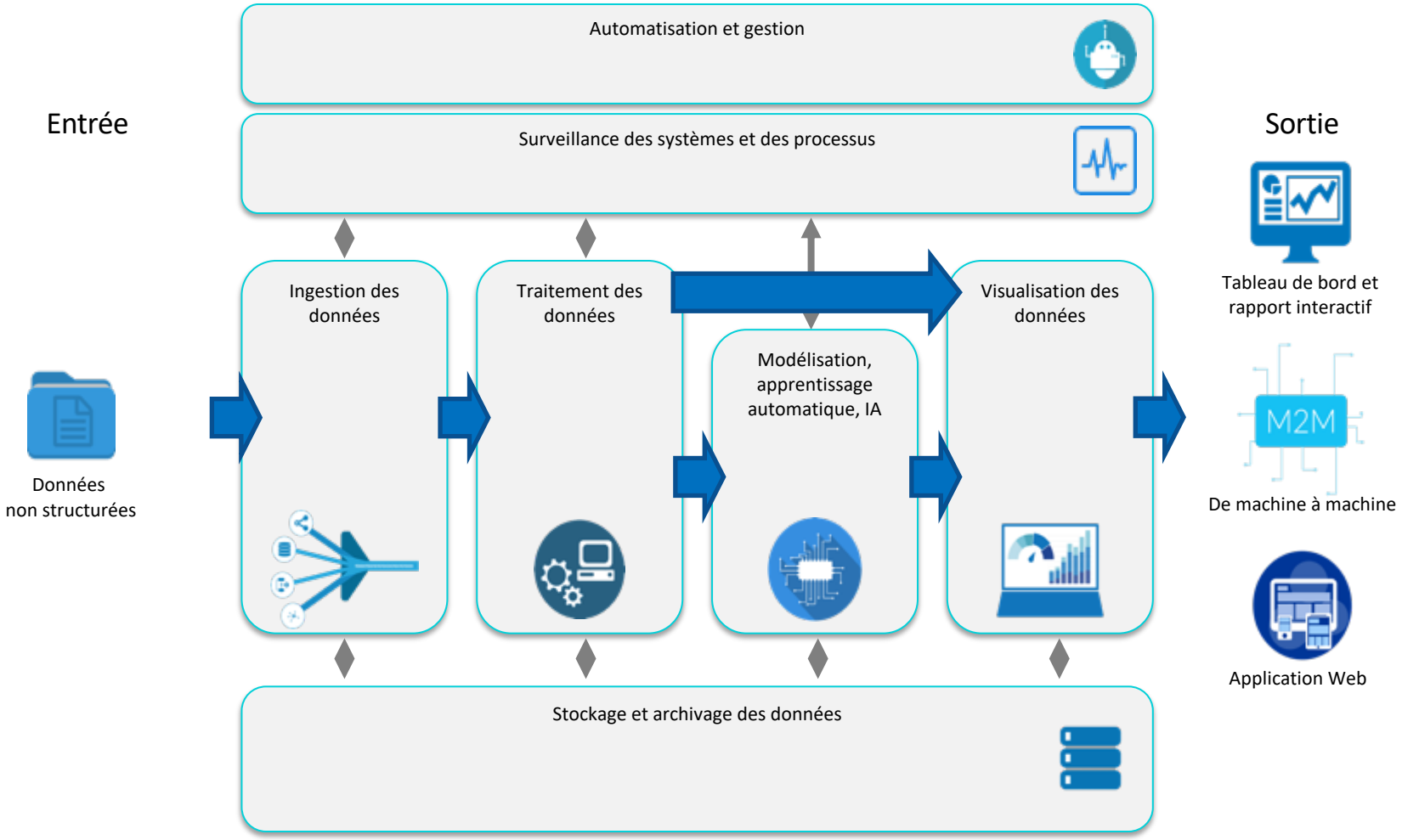
Architecture conceptuelle



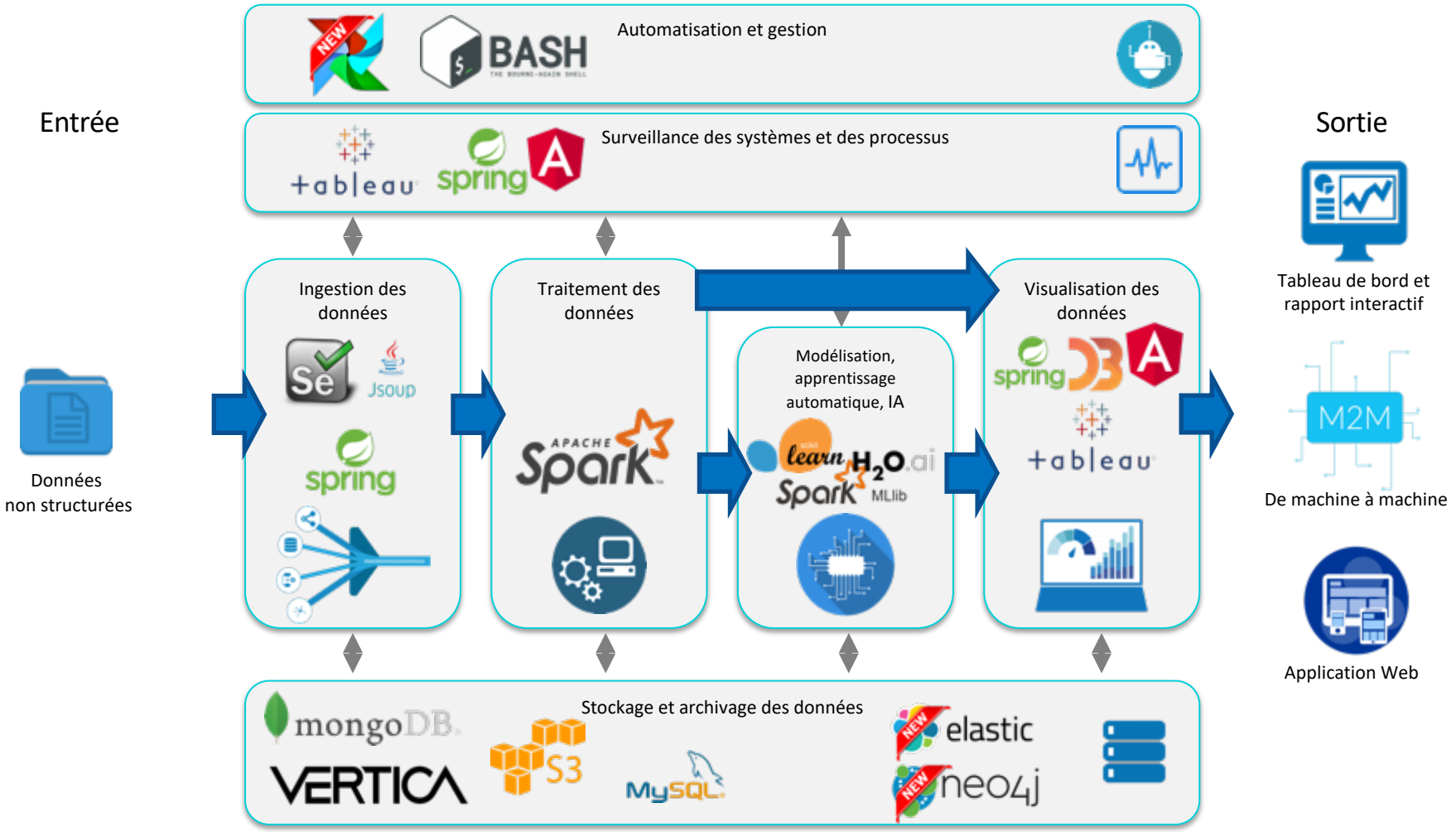
Vue logique


Agences pour l'emploi et services d'emploi public
Portails d'emploi
Sites de petites annonces
Journal, Entreprises Placement universitaire
Web scraper
Web crawler
Accès direct
Prétraitement
Information Extraction et Classification
Offres d'emploi publiées sur ISCO
NUTs reconnus
Autre dimension (contrat, secteur, éducation...)
Gestion des données et Présentation
Entrepôt de données
Analyse de données interactives
Analystes du marché du travail
Entrepôt de documents

Vue physique



Vue de la technologie



Projets de conception clés

- Microservices
- Subdivision des composantes
  - Spécialisation des composantes
  - Petites applications
  - Portabilité
  - Réutilisation
  - Maintenance
- Développement
  - Performance

Composantes clés

- Ingestion des données : collecte des données brutes des offres d'emploi en ligne dans des formats structurés et non structurés (texte brut).
- Traitement des données : classement des données grâce à des techniques d'apprentissage automatique.
- Analyse des données : extraction des informations des données et les rendre disponibles par la visualisation.
- Sauvegarde : stockage des données dans un environnement sûr pour permettre une restauration à chaud et à froid.

Les défis de l'infrastructure

- Gestion de plusieurs activités d'**ingestion parallèles**
- Disponibilité de l'infrastructure de calcul **haute performance en un coup d'œil**
- Besoins **élevés** en termes de **mémoire**
- Volumes de **stockage** élevés pour stocker les données sources et les données intermédiaires.
- Environnement Big Data
- Architecture **évolutive**

Flux Big Data

Exigences de qualité

Conception de microservices

Composantes par définition

Défis en matière d'infrastructures

Contexte

Maniabilité

Surveillance

Évolutivité

Mises à jour

Embarquement

Microservices de prétraitement

Détecteur de langues	Filtre de spams	Filtre de postes non vacants	Stemmer
Composant de déduplication	Composante N-gramme	Nettoyeur de texte	Merge Vacancy
Transformateur TF-IDF	Document2Vec	Tokenizer	Suppresseurs de mots vides

Microservices de classification

Classificateur de compétences

Classificateur de professions

Classificateur de secteurs

Classificateur d'exigences en matière d'éducation

Détecteur d'horaires de travail

Détecteur de contrat

Détecteur de lieux

Extracteur de salaire

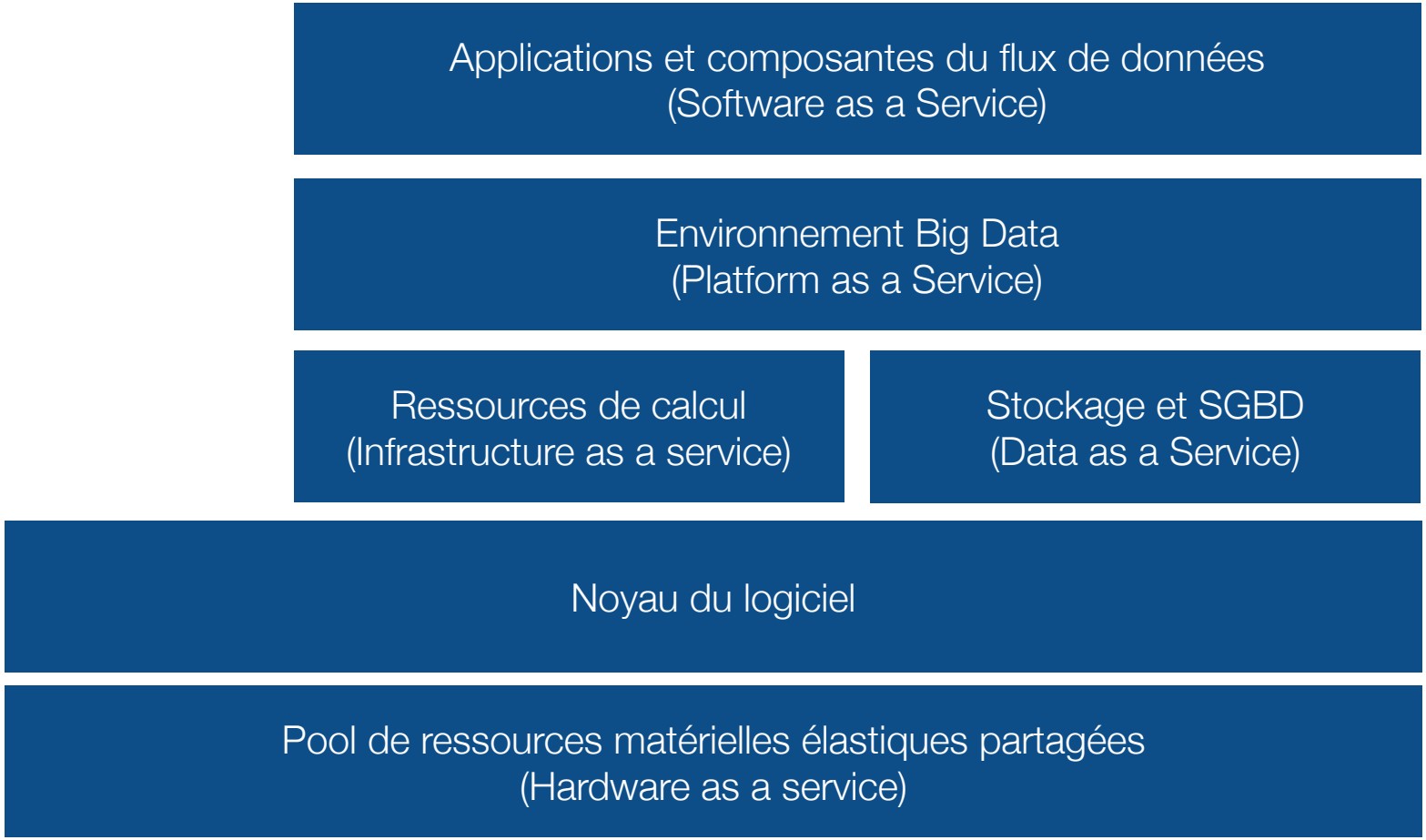
Extracteur d'expérience

Extracteur de dates

Exigences technologiques

1. Services sur demande
2. Accès au réseau
3. Mise en commun des ressources
    1. Gouvernance
4. Elasticité rapide
5. Mesure des services
    1. Qualité des données
    2. Performance
6. Portabilité (sur site et différents services en nuage)
7. Polyglotte
    1. Langages de programmation informatique
    2. Technologies

Vue organique



Récapitulatif et mots-clés

- Composantes clés et flux de données
  - Ingestion, traitement, classification, présentation
- Subdivision en composantes et microservices
- Pile hétérogène et big data
  - Selenium, Hadoop, Spark, Sklearn, Spark
- Environnement évolutif
  - Cloud

Des questions ?

Thèmes

1. Objectif et contexte
2. Défis
    1. Parties prenantes
    2. L'architecture fonctionnelle
    3. **Techniques d'ingestion de données**
    4. Pipeline de traitement des données
    5. Techniques de classification

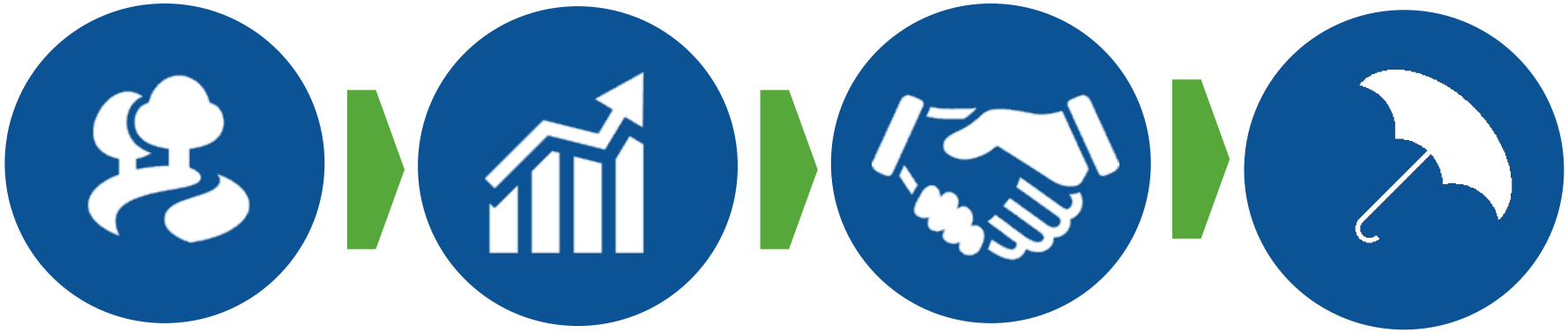
Inventaire

Une **activité d’inventaire** est menée pour produire une liste de **sources** (portails web) pertinentes pour le marché du travail en ligne dans un pays donné.

Un expert national **valide** cette liste, qui deviendra l'étape initiale du système d'information sur le marché du travail.

Stratégie de sélection des sources

4 étapes de traitement



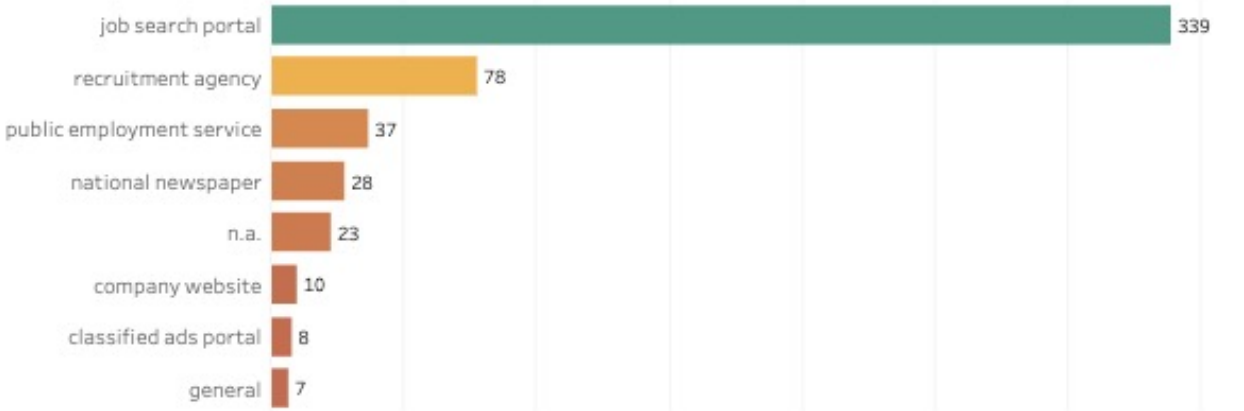
Sélection des sources lors de l'inventaire

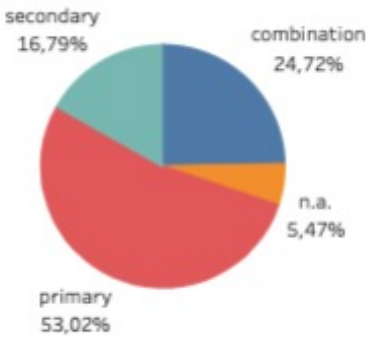
Augmentation

Accords

Couverture

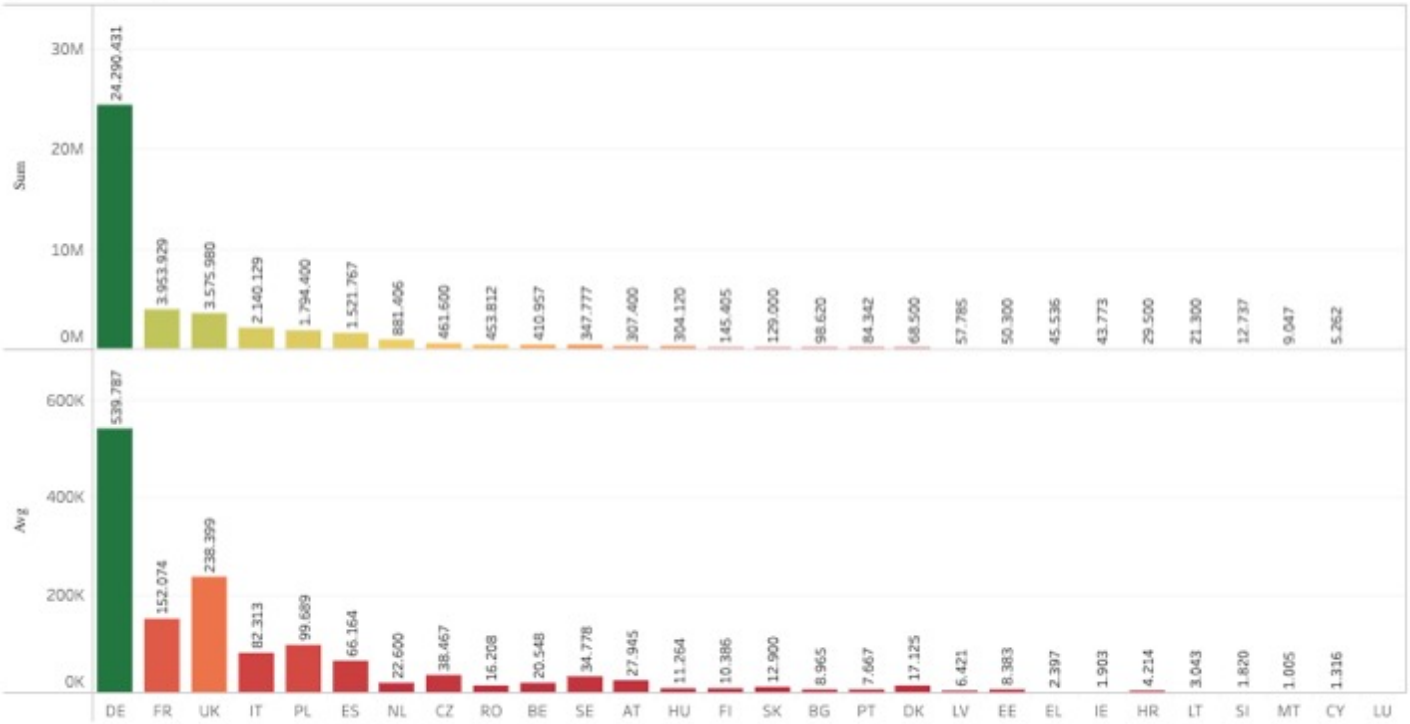
Sites by type of operator

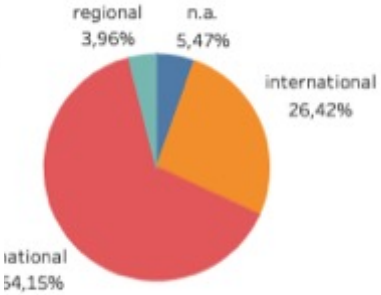


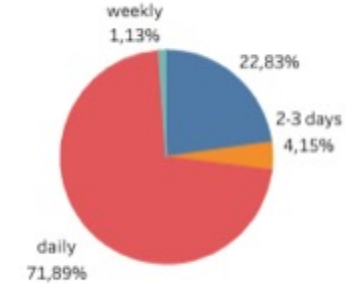


Vacancy volume by country

(estimated by ICE)







Augmentation

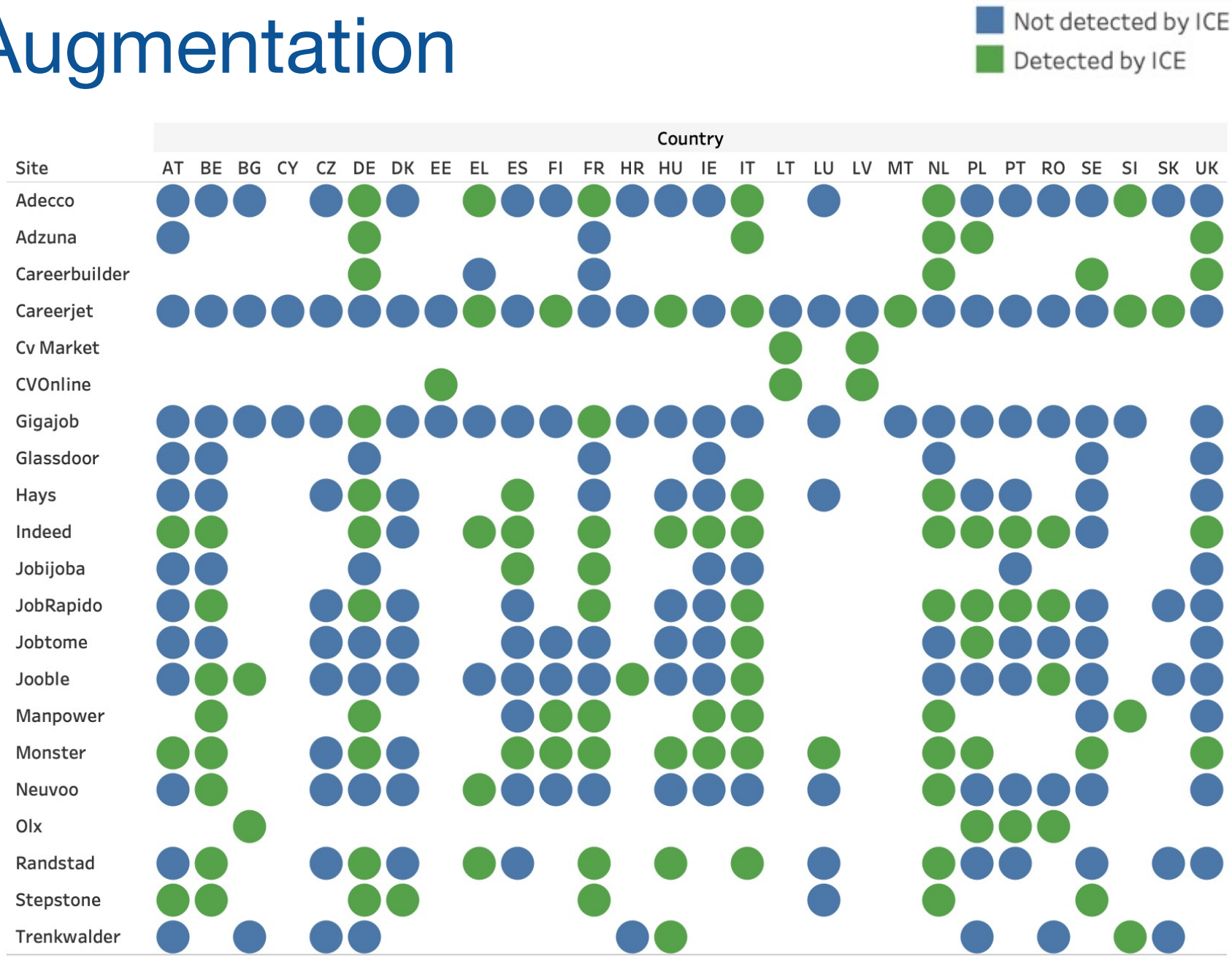
Nous avons analysé les résultats de l'activité d’inventaire

- Compléter la cartographie des sources transnationales
- Ajouter d'autres sources transnationales
- Ajouter l'ensemble complet des sources EURES

Afin de définir

- une liste de priorités pour définir les accords
- un ordre de pertinence pour réaliser des canaux d'ingestion de données

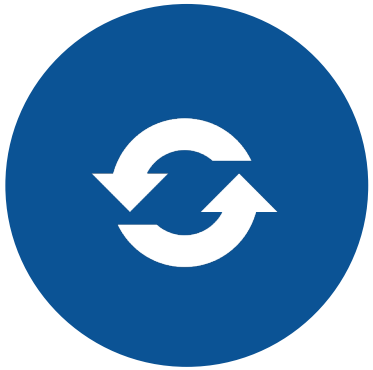
Augmentation



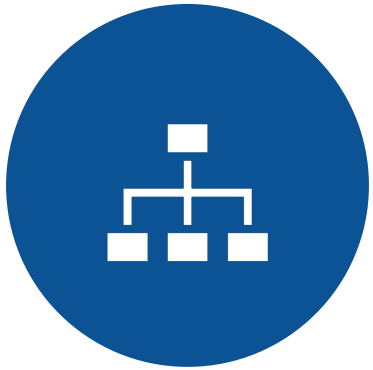
Pertinence et classement des sources

Volume

Type de portail web



Mise à jour des données



Données structurées

Phase d'ingestion des données

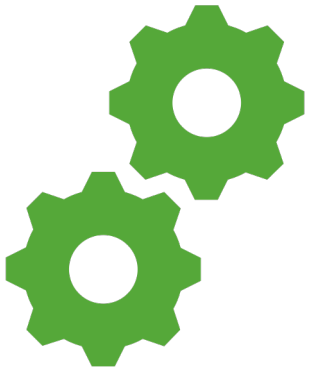
Le processus d'obtention et d'importation de données à partir de portails web et leur stockage dans une base de données.

Se concentrer sur les volumes

Augmentation et maximisation de la couverture

Accords directs avec les sources les plus pertinentes

Les défis de l'ingestion

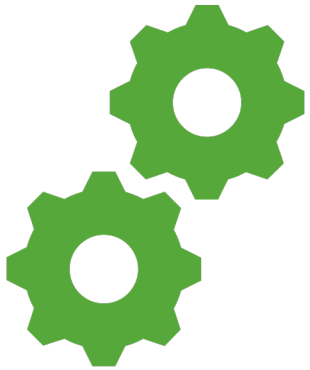


Robustesse du processus

Qualité des données collectées

Évolutivité et gouvernance

Les défis de l'ingestion



1. Robustesse

Problème : problèmes techniques potentiels lors de la collecte de données à partir d'une source (indisponibilité, blocage, changements dans la structure des données).

Risque : perte de données

Solution : la redondance

- Avoir les sites les plus importants (par le volume et/ou la couverture) ingérés à partir de deux sources ou plus
- Éviter la perte de données en cas de problèmes avec une source.
- Collecter des données à partir de sources primaires et secondaires

Les défis de l'ingestion

2. Qualité

Problème : nécessité d'obtenir des données aussi propres que possible, en détectant les données structurées lorsqu'elles sont disponibles.

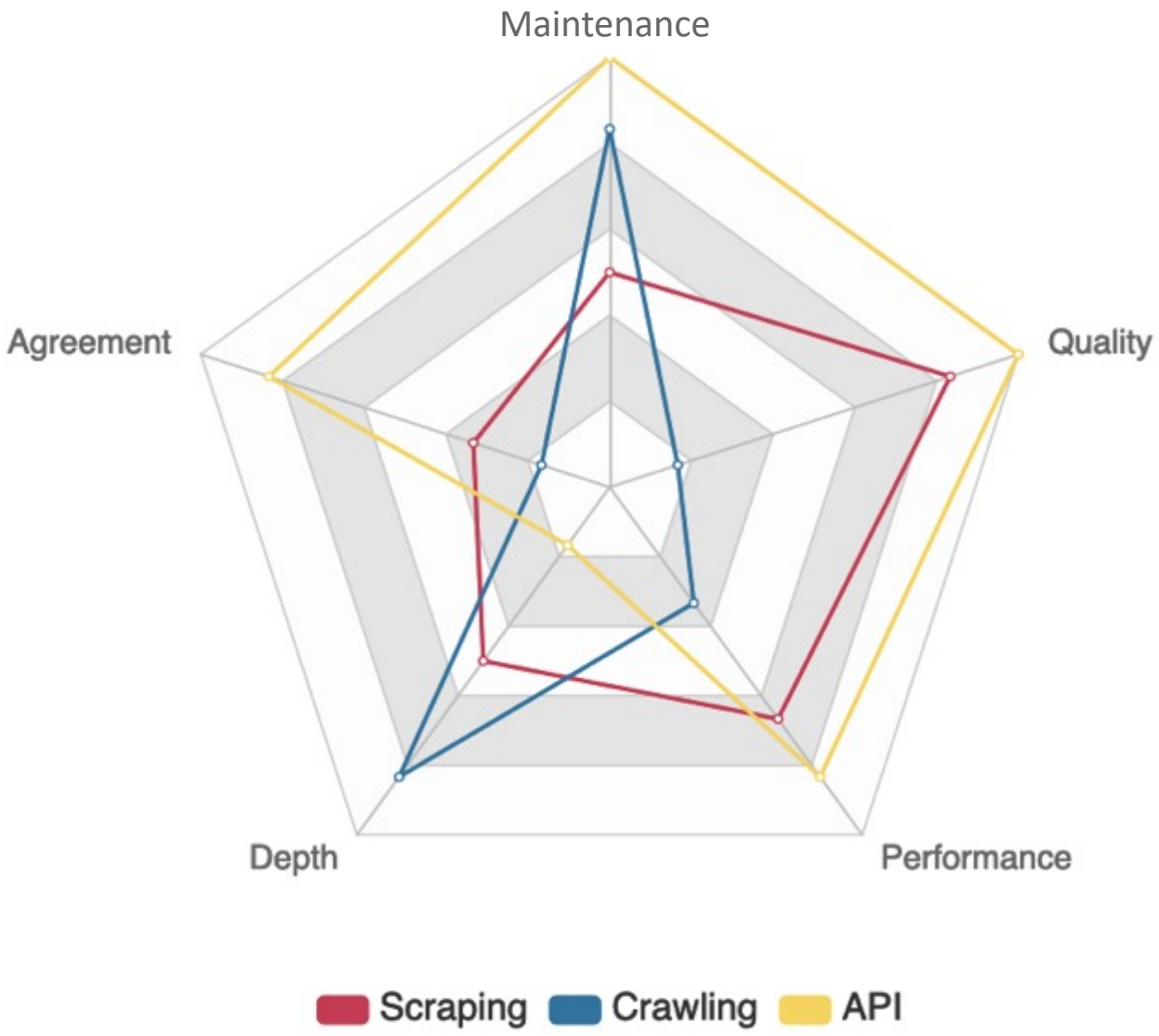
Risque : perte de qualité

Solution : une ingestion sur mesure. Nous collectons les données selon une approche spécifique basée sur la source unique :

- API
- Scraping
- Crawling

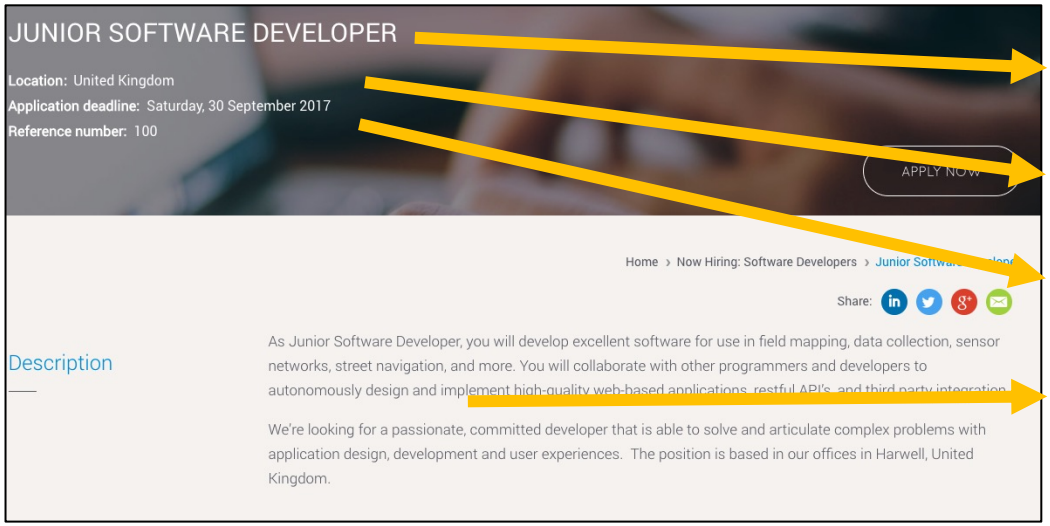
Défis de l'ingestion - Qualité

- API : lorsqu'elles sont disponibles (accords), nous collectons principalement des données structurées à partir de portails Web.
  - Avantages : Très haute qualité (la plupart des champs sont structurés)
  - Inconvénients : besoin d'un accord, pas toujours disponible
- Scraping : si l'API n'est pas réalisable et que la structure du portail web est cohérente, nous développons un scraper personnalisé qui extrait les données structurées/non structurées des pages.
  - Avantages : Haute qualité (nombreux champs structurés)
  - Inconvénients : développement spécifique au portail Web
- Crawling : si la structure des pages du portail web n'est pas cohérente, nous ingérons les données en utilisant une approche de crawling polyvalente.
  - Avantages : Qualité inférieure (pas de champs structurés)
  - Inconvénients : Approche rapide et polyvalente


Maintenance
Quality
Performance
Depth
Agreement
Scraping
Crawling
API

Scraping - Un exemple

Le web scraping consiste en un « grattage » de données utilisé pour extraire des données structurées de sites web.



Titre : Développeur de logiciels junior

Zone : Royaume-Uni

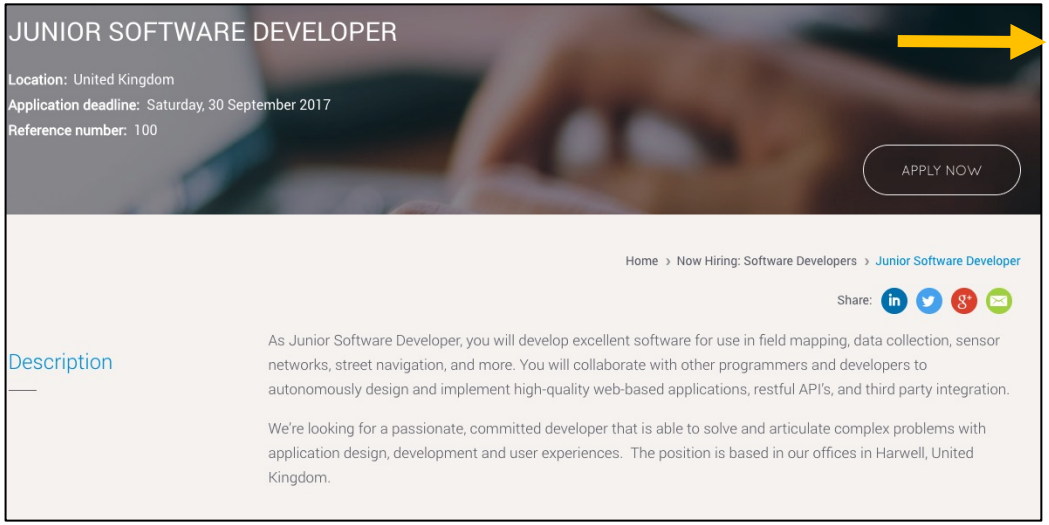
Date : Samedi 30 septembre 2017

Description : En tant que développeur de logiciels junior, vous développerez d'excellents logiciels destinés à être utilisés...

Crawling - Un exemple

Un Web crawler est un robot qui parcourt systématiquement les portails Web dans le but de télécharger toutes leurs pages.

Le crawling est le moyen le plus courant d'obtenir massivement des informations sur Internet : les spiders des moteurs de recherche (par exemple GoogleBot)



Page Web :

```
<!DOCTYPE html>
  <br />
    <meta name="title" content="Junior
Software Developer" />
  </head>
  <body>
    <bip>En-tête>
      <h2>Développeur de logiciels junior</h2>
      <div><div>Location</div>Royaume-
Uni</div>.
    …
    </header>
    <div><div>Description</div>.
    <span>En tant que développeur de logiciels
junior, vous développerez d'excellents
logiciels pour une utilisation...
```

Les défis de l'ingestion

3. Évolutivité et gouvernance

Problème : nécessité de gérer un environnement Big Data réel et complexe, en se connectant simultanément à des milliers de sites web.

Risque : Perte de contrôle du processus et perte d'offres d'emploi en ligne en raison de la lenteur du processus.

Solution :

- Une infrastructure évolutive
- Un outil personnalisé de suivi et de gouvernance

Défis de l'ingestion - Scaling

Nous avons développé une solution basée sur des microservices, qui crée et supprime des « ordinateurs de navigation virtuels » selon les besoins. Chaque ordinateur dispose de plusieurs navigateurs qui peuvent émuler la navigation humaine sur le web.

Les principales différences avec un véritable ordinateur sont les suivantes :

1. Ils n'ont pas de moniteur, mais enregistrent les pages sur notre lac de données.
2. Nous pouvons augmenter ou diminuer les effectifs selon les besoins

Récapitulatif et mots-clés

- Inventaire, sélection des sources et augmentation
- Approche sur mesure
  - Composantes API, Scraping, Crawling
- Focus sur la quantité
  - Scaling et collecte en temps réel
- Suivi en temps réel des données collectées

Des questions ?

Thèmes

1. Objectif et contexte
2. Défis
    1. Parties prenantes
    2. L'architecture fonctionnelle
    3. Techniques d'ingestion de données
    4. **Pipeline de traitement des données**
    5. Techniques de classification

Prétraitement des données - Défis et définitions

- Objectif :
  - Alimenter la phase d'extraction des informations avec des données appropriées
- Les défis :
  - Mesurer, contrôler et améliorer la qualité des données, afin de maximiser l'exhaustivité, la cohérence, la complexité, l'actualité et la périodicité.
- Approche :
  - Développer un pipeline multi-phases, axé sur :
    - Détection des postes vacants : analyse des pages du site web afin de sélectionner uniquement le contenu relatif aux postes vacants.
    - Déduplication : détecter les postes vacants dupliqués pour obtenir une seule entité de poste vacant.
    - Détection des dates : identifier les dates de publication et d'expiration grâce à l'analyse de la description des postes vacants.
    - Durée de la vacance : méthode pour définir la date d'expiration, lorsqu'elle n'est pas explicitement disponible.
- Caractéristiques :
  - Garantir la qualité des données pendant toutes les phases de traitement

Prétraitement des données - Défis et définitions

Le processus de nettoyage des données ingérées et la déduplication des offres d'emploi en ligne, pour garantir que la phase analytique travaillera sur des données de la meilleure qualité possible.

Détection de la langue

Réduction du bruit

Déduplication des offres d'emploi en ligne

Étapes de prétraitement

Fusion

Nettoyage

Traitement de texte
et résumé

Prétraitement des données
La détection de la langue

- Pourquoi :
  - Chaque langue a des mots-clés différents, des mots vides...
  - Elle peut refléter des cultures et des scénarios de marché du travail différents...
  - ... Il est donc fondamental de classer la langue de l'offre d'emploi en ligne, donc d'utiliser le pipeline de classification le plus approprié.
- Comment :
  - Nous avons formé pour chaque langue (60+) un classificateur spécifique basé sur le corpus Wikipedia.
  - Les modèles obtenus sont très précis (~99 % de précision) et rapides à adopter dans le pipeline.
- Ce que nous obtenons :
  - Une classification rapide et forte de la langue utilisée dans chaque offre d'emploi en ligne.
  - Un moyen d'archiver les offres d'emploi en ligne pour lesquelles nous n'avons pas de pipeline de classification.

Prétraitement des données

Comment traiter le bruit ?

- Dans un environnement de Big Data, nous devons faire face au bruit.
  - Pourquoi ? Parce que les informations sont collectées sur le web, l'un des endroits les plus bruyants jamais connus.
- Tout d'abord, nous devons maîtriser le type de bruit auquel nous devons faire face... :
  - Pages web explicitement non liées aux offres d'emploi en ligne :
    - Pages de réseaux sociaux
    - Pages d'actualités
    - Pages de politique de confidentialité
    - ...
  - Des pages web déguisées en offres d'emploi en ligne :
    - Cours de formation
    - CVs
    - Services de conseil
    - ...

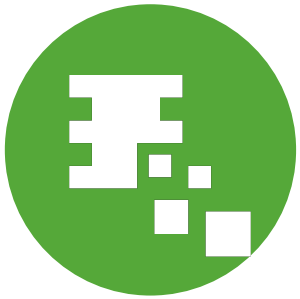
- ...Ensuite, nous devons détecter et gérer les offres d'emploi en ligne dupliquées :
  - En général, un poste vacant est publié sur plusieurs portails
  - Si nous les traitons de manière distincte, nous surestimons la demande de main-d'œuvre.
  - Il s'agit donc de détecter les OJV dupliqués et de fusionner les informations provenant de ces OJV en un seul.

Prétraitement des données
Détection du bruit – Comment ?

- Approche en 2 étapes :
  - Approche de l'apprentissage automatique
    - Pour chaque langue, nous avons formé un classificateur Naïve Bayes avec plus de 20 000 pages web :
      - » 10 000 pages réelles liées aux offres d'emploi en ligne
      - » 10 000 pages web non liées aux offres d'emploi en ligne
    - Précision de ~99 %.
    - Rapide à former et à utiliser
    - Une approche similaire à celle d'un système de « détection des spams par courrier électronique ».
  - Approche de la correspondance floue
    - Utilisée pour détecter les pages Web « similaires aux offres d'emploi en ligne », mais liées à des offres de formation, des services de conseil...
    - Examine l'en-tête et le corps de la page publicitaire pour détecter des mots clés (en fonction de la langue) qui peuvent nous aider à l'étiqueter comme une page « non liée aux offres d'emploi en ligne ».

Mais, avant de commencer la phase de déduplication d'offres d'emploi en ligne, nous devons nettoyer le texte pour le simplifier et le consolider...

Prétraitement des données
Phase de déduplication



Annonces d'emploi

Déduplication physique ou fuzzy matching

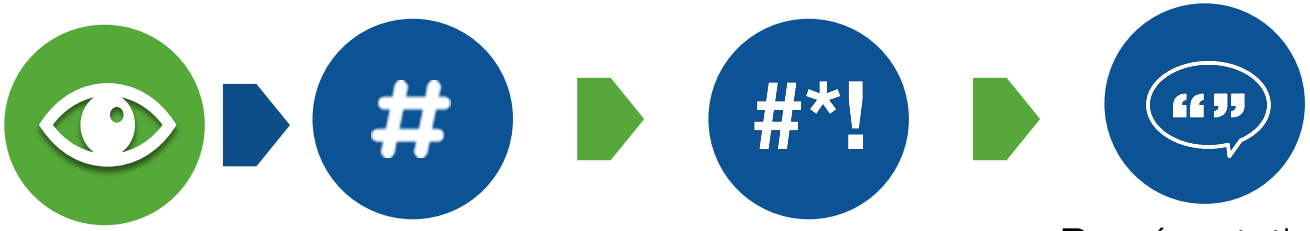
Réalisée sur la partie description (ou contenu) de l'offre d'emploi.

Correspondance des métadonnées

Utilisation des métadonnées provenant des portails d'emploi pour supprimer les doublons d'offres d'emploi sur les sites web des agrégateurs (par exemple, l'id de référence, l'url de la page).

Traitement et résumé de texte

La phase de traitement et de résumé du texte vise à réduire le texte pour améliorer le processus de classification des offres d'emploi selon les normes européennes.



JUNIOR SOFTWARE DEVELOPER

Location: United Kingdom
Application deadline: Saturday, 30 September 2017
Reference number: 100

Description

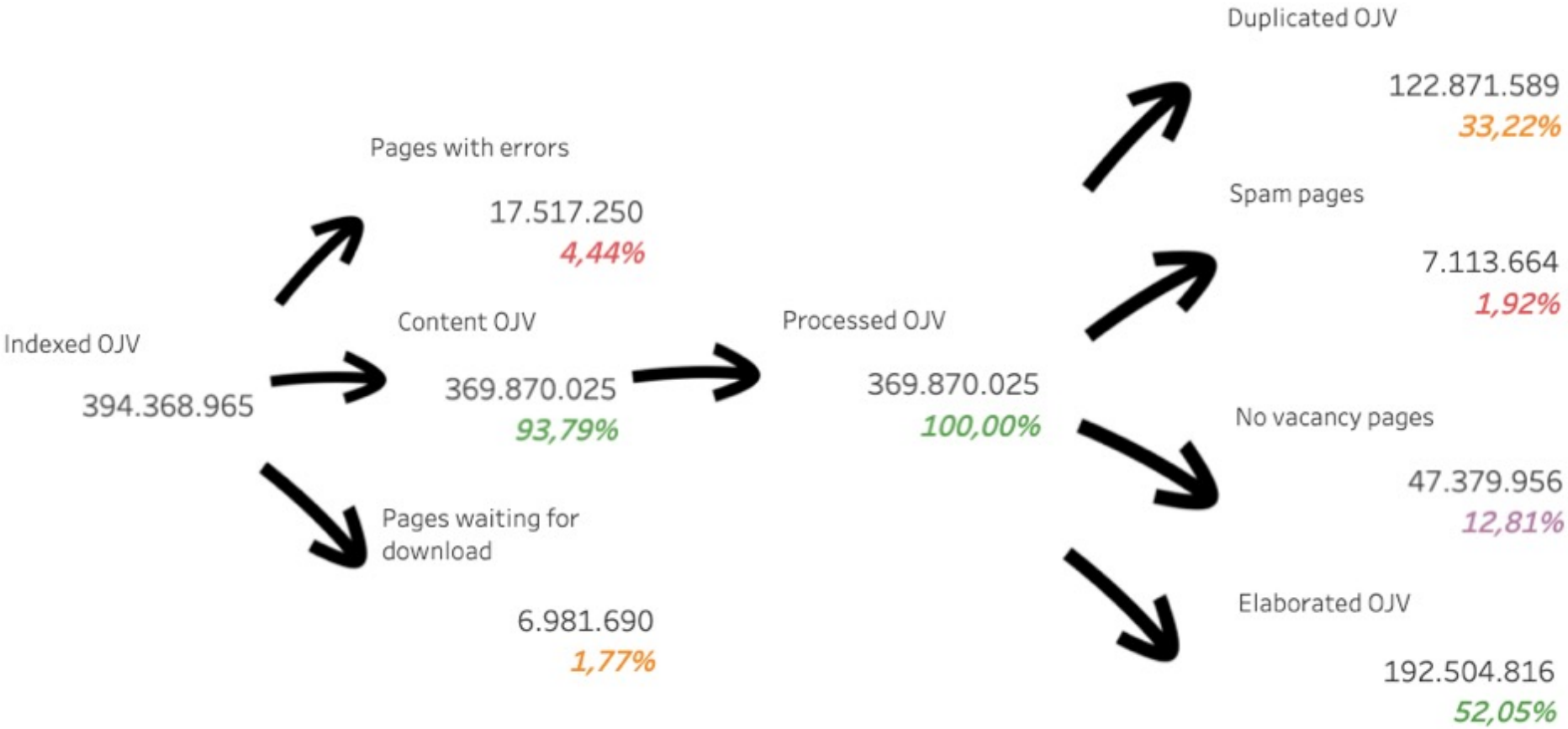
As Junior Software Developer, you will develop excellent software for use in field mapping, data collection, sensor networks, street navigation, and more. You will collaborate with other programmers and developers to autonomously design and implement high-quality web-based applications, restful API's, and third party integration.

We're looking for a passionate, committed developer that is able to solve and articulate complex problems with application design, development and user experiences. The position is based in our offices in Harwell, United Kingdom.

En tant que ⟨Développeur de logiciels⟩ junior, vous développerez un excellent ⟨logiciel⟩ destiné à la ⟨cartographie⟩, à la ⟨collecte de données⟩, aux ⟨réseaux de capteurs⟩, à la ⟨navigation routière⟩, etc. Vous ⟨collaborerez⟩ avec d'autres ⟨programmeurs⟩ et ⟨développeurs⟩ pour concevoir ⟨de manière autonome⟩ et mettre en œuvre des ⟨applications web⟩ de qualité, ⟨API⟩ restful et tiers ⟨intégration⟩.

Nous recherchons un ⟨développeur⟩ passionné et engagé, capable de ⟨résoudre⟩ et d'articuler des ⟨problèmes complexes⟩ avec ⟨conception d'application⟩, ⟨développement⟩ et ⟨expériences utilisateurs⟩.
Le poste est basé dans nos bureaux à ⟨Harwell⟩ , ⟨Royaume-Uni⟩.

Pré-traitement des données – Résultats
Le bruit



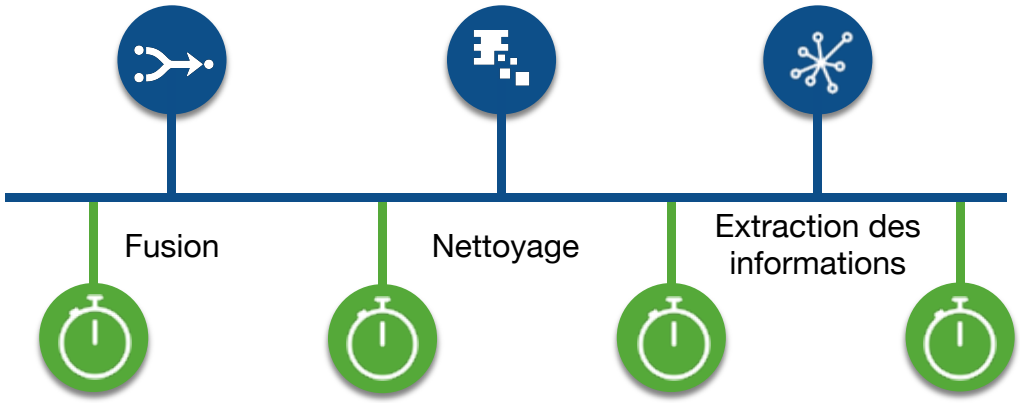
Prétraitement des données
Que faire du bruit ?

Nous ne supprimons pas physiquement le bruit

Nous les recueillons pour suivre l'ensemble du processus et en assurer le suivi :

- Type de bruit → Identifier le besoin de développer un processus de contrôle de qualité plus approfondi.
- Tendances du bruit→ Détecter les sources qui augmentent/diminuent le bruit et les traiter.
- Objectifs analytiques → Analyser les environnements culturels spécifiques à chaque pays, comme l'utilisation du portail des offres d'emploi en ligne pour promouvoir les cours de formation.
- Suivi → Suivre l'ensemble du processus

Comment assurer le suivi de la qualité des données ?



Portiques de contrôle de la qualité en continu

Récapitulatif et mots-clés

- Priorité à la qualité
  - Comment supprimer le bruit ?
  - Activités de déduplication
- Le défi des langues
  - Composante adaptée à chaque langue
- Suivi de la qualité des données
  - Contrôle continu de la qualité et contrôles

Des questions ?

Thèmes

1. Objectif et contexte
2. Défis
    1. Parties prenantes
    2. L'architecture fonctionnelle
    3. Techniques d'ingestion de données
    4. Pipeline de traitement des données
    5. **Techniques de classification**

Content

379.794.151

Processed

379.794.151

Elaborated

199.008.930

Contract

Structured fields collected
(% over total elaborated OJV)

23,75%

Total fields extracted
(% over total elaborated OJV)

49,00%

Method

Educational level

Structured fields collected
(% over total elaborated OJV)

5,45%

Total fields extracted
(% over total elaborated OJV)

77,63%

Method

Experience

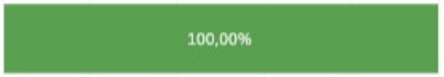
Structured fields collected
(% over total elaborated OJV)

3,86%

Total fields extracted
(% over total elaborated OJV)

35,49%

Method



Occupation

Structured fields collected
(% over total elaborated OJV)

5,21%

Total fields extracted
(% over total elaborated OJV)

76,09%

Method

Industry

Structured fields collected
(% over total elaborated OJV)

25,36%

Total fields extracted
(% over total elaborated OJV)

78,37%

Method

Salary

Structured fields collected
(% over total elaborated OJV)

13,71%

Total fields extracted
(% over total elaborated OJV)

18,24%

Method

Skill

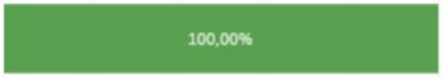
Structured fields collected
(% over total elaborated OJV)

49,94%

Total fields extracted
(% over total elaborated OJV)

62,52%

Method



Working hours

Structured fields collected
(% over total elaborated OJV)

18,21%

Total fields extracted
(% over total elaborated OJV)

43,11%

Method

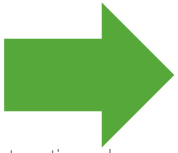
Feature extraction (Equal)
Feature extraction (Similarity)

Classification des données

- Objectif :
  - Extraire et structurer les informations des données, pour les fournir à la couche de présentation.
- Les défis :
  - Traiter des quantités massives de données hétérogènes écrites dans différentes langues.
- Approche :
  - Développer un cadre adaptable, dépendant de la langue, adapté aux différentes caractéristiques de l'information. Quelques défis pertinents :
    - Classification des caractéristiques **professionnelles** : méthodes combinées telles que l'apprentissage automatique, la modélisation thématique et l'apprentissage non supervisé.
    - Classification des caractéristiques des **compétences** : autres méthodes combinées, telles que l'analyse de texte avec corpus ou la similarité basée sur la connaissance.
- Caractéristiques :
  - Garantir l'extraction d'informations explicables, les méthodes de classification des journaux et les caractéristiques pertinentes.

Classification des données – Un exemple

Offre d'emploi



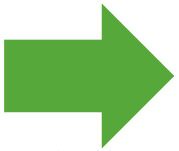
Extraction des informations

Profession	Compétences
Temps	Zone
Industrie	…

Développeur de logiciels junior

En tant que développeur de logiciels junior, vous développerez d'excellents logiciels destinés à être utilisés pour la cartographie sur le terrain, la collecte de données, les réseaux de capteurs, la navigation dans les rues, etc. Vous collaborerez avec d'autres programmeurs et développeurs afin de concevoir et de mettre en œuvre de manière autonome des applications Web de haute qualité, des API restful et l'intégration de tiers.

Nous recherchons un développeur passionné et engagé, capable de résoudre et d'articuler des problèmes complexes de conception d'applications, de développement et d'expériences utilisateur. Le poste est basé dans nos bureaux à Harwell, au Royaume-Uni.



Extraction des informations

2512 - Développeur de logiciels

Compétences : développement de logiciels, mise en œuvre d'applications web, résolution de problèmes, développement d'expériences utilisateur.

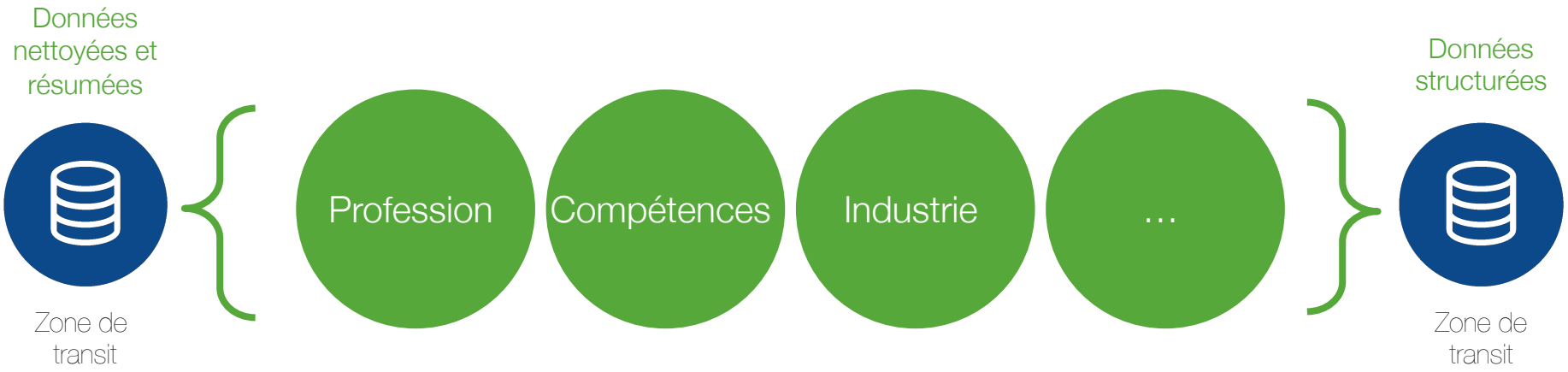
Harwell, Royaume-Uni

…

Extraction et classification des informations Information sur le marché du travail en temps réel

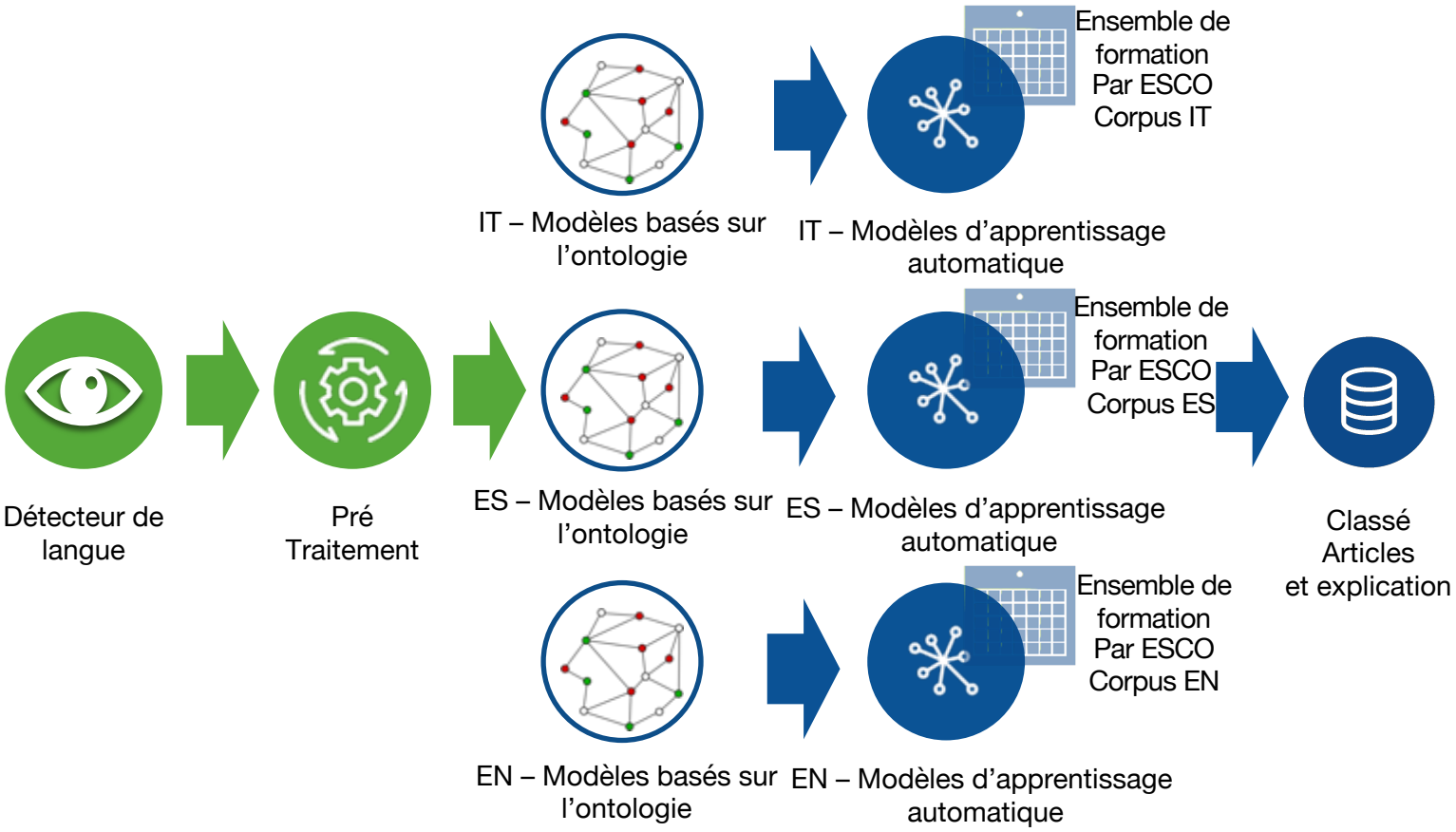
L'extraction d'informations est un domaine du traitement du langage naturel qui traite de la recherche d'informations factuelles dans un texte libre.

Cette tâche utilise des techniques d'apprentissage automatique (apprentissage basé sur l'ontologie, apprentissage supervisé et apprentissage non supervisé) pour faire correspondre les offres d'emploi à des classifications standard.



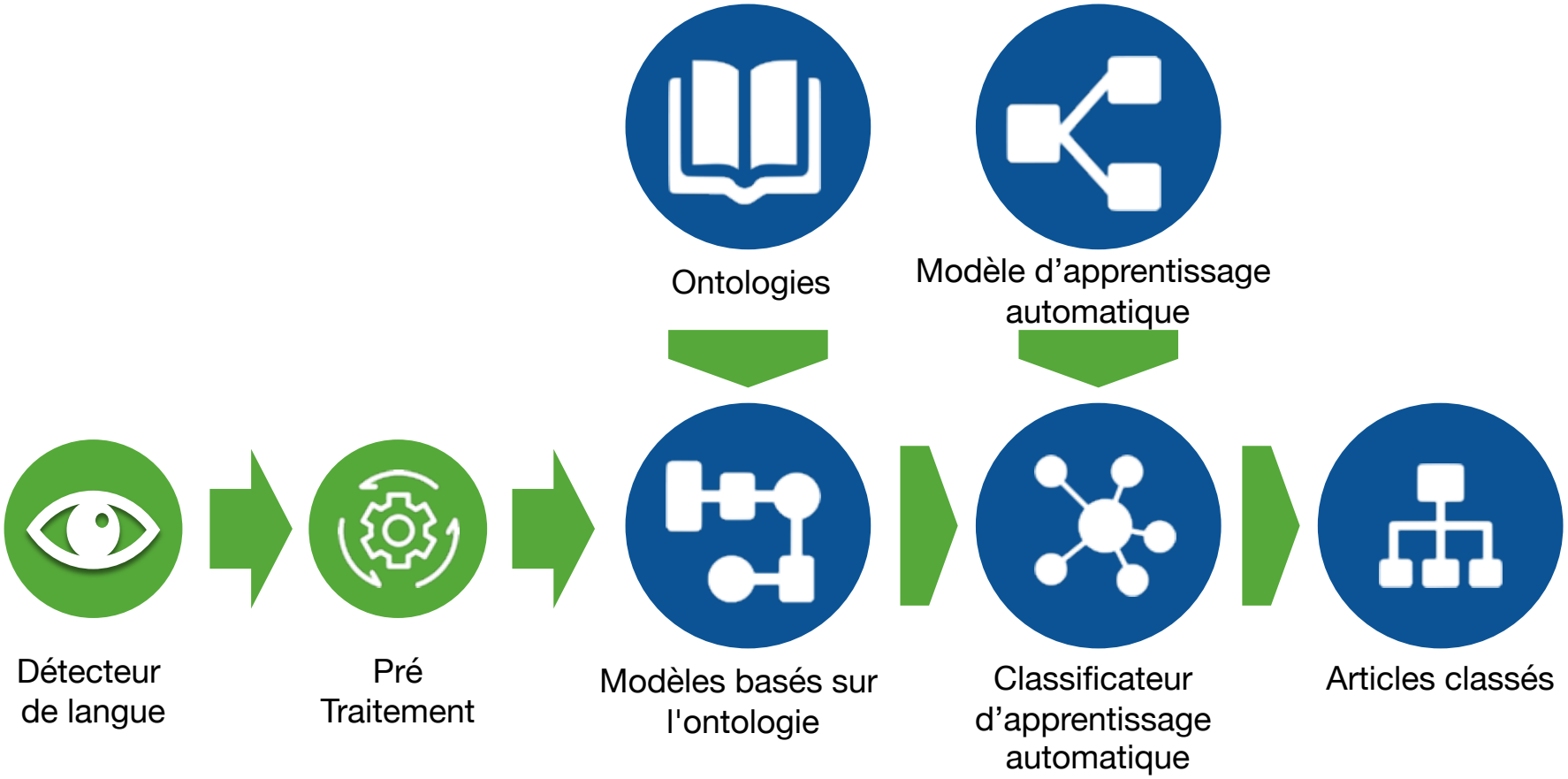
Apprentissage automatique → Apprentissage basé sur l'ontologie, apprentissage supervisé et apprentissage non supervisé, etc.

Classification



Que signifie « modèles basés sur l'ontologie » ?
Comment pouvons-nous utiliser les ontologies pour classer ?

Pipeline des professions



Considérations sur le classificateur de profession

- Apprentissage basé sur l'ontologie + Apprentissage supervisé
  - Ontologie Esco
  - Nouvelles étiquettes issues de la modélisation des thèmes
- Un modèle pour chaque langue
- Données étiquetées par un expert de chaque pays
  - ~100 000 annonces d'emploi (ensemble d'entraînement nettoyé en utilisant notre ontologie)
  - 436 cibles possibles
- Évaluation 20 % des « golden dataset » d'offres d'emploi
  - Précision pondérée ~86 %.
  - ~430 professions détectées

Approches de la similarité des textes

Basées sur les chaînes

Les mesures de similarité des chaînes opèrent sur les séquences de chaînes et la composition des caractères.

Jaro-Winkler, Jaccard, Similarité cosinus

Basées sur les corpus

La similarité basée sur les corpus est une mesure de similarité sémantique qui détermine la similarité entre les mots en fonction des informations obtenues à partir de grands corpus.

Analyse sémantique latente, analyse sémantique explicite, mots similaires à l'aide des CO-occurrences.

Basées sur les connaissances

La similarité basée sur les connaissances est fondée sur l'identification du degré de similarité entre les mots à l'aide d'informations provenant de réseaux sémantiques.

Precision of occupation (overall)

86,66%

Validation Set (overall)

317.864

Validation Set by language

bg	ca	cs	da	de	el	en	es	et	eu	fi	fr	gl	hr	hu	it	lt	lv	nl	pl	pt	ro	sk	sl	sv
5.050	14.210	31.290	6.022	17.420	7.173	35.019	21.680	8.414	196	11.972	39.146	811	4.637	13.813	17.228	7.447	4.443	8.687	10.554	14.678	10.226	3.089	4.576	20.083

Precision of occupation by language

bg	ca	cs	da	de	el	en	es	et	eu	fi	fr	gl	hr	hu	it	lt	lv	nl	pl	pt	ro	sk	sl	sv
85,96%	86,45%	88,96%	96,89%	83,17%	85,02%	82,83%	87,39%	83,63%	68,59%	98,80%	83,89%	83,07%	89,45%	97,40%	81,30%	92,86%	96,95%	93,31%	80,02%	83,01%	95,12%	98,50%	87,78%	78,43%

Precision of occupation (lv1)

Clerical support workers	85,77%
Craft and related trades ..	86,10%
Elementary occupations	86,19%
Managers	86,32%
Plant and machine operat..	86,29%
Professionals	86,61%
Service and sales workers	89,38%
Skilled agricultural, fores..	88,79%
Technicians and associate..	85,54%

Precision of occupation (lv2)

Administrative and comm..	85,06%
Agricultural, forestry and ..	80,82%
Assemblers	84,87%
Building and related trad..	92,30%
Business and administrati..	85,66%
Business and administrati..	80,06%
Chief executives, senior o..	91,36%
Cleaners and helpers	85,11%
Customer services clerks	82,21%
Drivers and mobile plant ..	86,49%
Electrical and electronic t..	74,60%
Food preparation assista..	89,08%
Food processing, wood w..	82,61%
General and keyboard cler..	97,20%
Handicraft and printing w..	89,65%

Precision of occupation (lv3)

Administration professio..	86,21%
Administrative and specia..	84,92%
Agricultural, forestry and ..	80,82%
Animal producers	83,13%
Architects, planners, surv..	87,56%
Artistic, cultural and culin..	91,74%
Assemblers	84,87%
Authors, journalists and li..	90,72%
Blacksmiths, toolmakers ..	86,70%
Building and housekeepin..	90,33%
Building finishers and rel..	95,47%
Building frame and relate..	90,00%
Business services agents	89,57%
Business services and ad..	79,10%
Car, van and motorcycle d..	90,40%

Precision of occupation (lv4)

Accountants	83,60%
Accounting and bookkeepi..	58,14%
Accounting associate prof..	85,65%
Actors	93,41%
Administrative and execu..	84,32%
Advertising and marketin..	65,30%
Advertising and public rel..	71,63%
Aged care services manag..	78,81%
Agricultural and forestry ..	94,55%
Agricultural and industria..	76,49%
Agricultural technicians	81,32%
Air conditioning and refri..	85,95%
Air traffic controllers	84,43%
Air traffic safety electroni..	95,52%
Aircraft engine mechanics..	79,61%

Récapitulatif et mots-clés

- Focus sur le résumé
  - Comment résumer les données et améliorer les résultats de nos analystes de données ?
- Lien vers des taxonomies standard
  - Comparer les données des offres d’emploi en ligne avec d'autres sources
- Les défis des « golden datasets » (cardinalité, qualité et diversité)
- Approches mixtes
  - Apprentissage automatique
  - Apprentissage basé sur l'ontologie
  - Similitude des textes et techniques d'extraction d'informations
- Cycle de vie du modèle

Des questions ?