Big Data for Labour Market Intelligence

Day 1, Session 1

Analysis of demand based on data from online job vacancies sources

Alessandro Vaccarino – Mauro Pelucchi

22 November 2021

Topics

1. Goal & context
2. Challenges
    1. Stakeholders
    2. The functional architecture
    3. Data ingestion techniques
    4. Data processing pipeline
    5. Classification techniques
3. Outputs
4. Outcomes

Topics

1. **Goal & context**
2. Challenges
    1. Stakeholders
    2. The functional architecture
    3. Data ingestion techniques
    4. Data processing pipeline
    5. Classification techniques
3. Outputs
4. Outcomes

Context

Continuously evolving Labour Market:

- Digitalization of professions
- Relevance of Soft skills
- Internationalisation
- New professions and skills emerging
- Smart and Remote working
- Impact of Covid-19 pandemic
- …

We need *something* that can help us monitor and analyze **how** LM is evolving, to support Decision Makers taking **the right decisions at the right time**

What we have / what we need

We already have **official statistics**, that are:

- *Representative*
- *Strong* in terms of value

But we can benefit of **additional, complementary information** that could be:

- *Fast*, to track what's happening now (e.g. Covid-19 Impact analysis)
- *Granular* and *adherent* to real and current market terms, to capture emerging trends analyzing what companies are actually looking for

How to find a similar, complementary source of information?
Using **Web Labour Market**

Why Web Labour Market

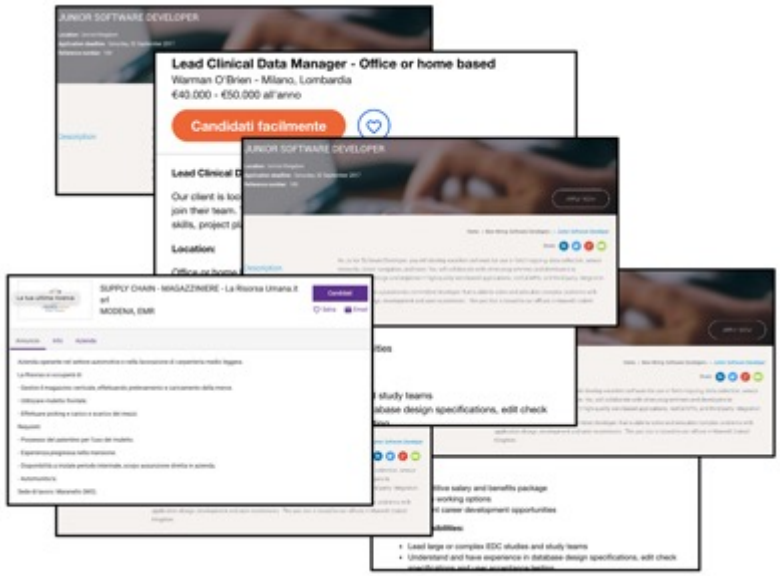
It's the exact representation of what companies are looking in a given period:

- Up to date: companies publish an announcement when they actually need to hire
- Detailed: an announcement describes as well as possibile the specific need, in terms of:
  - Profession needed
  - Requirements (skills, experience, educational level,…)
  - Working context (place, contract, sector, working hours,…)
- Adherent to reality: market terms are used, both for occupation and skills. This helps identify emerging terminology adopted by Market

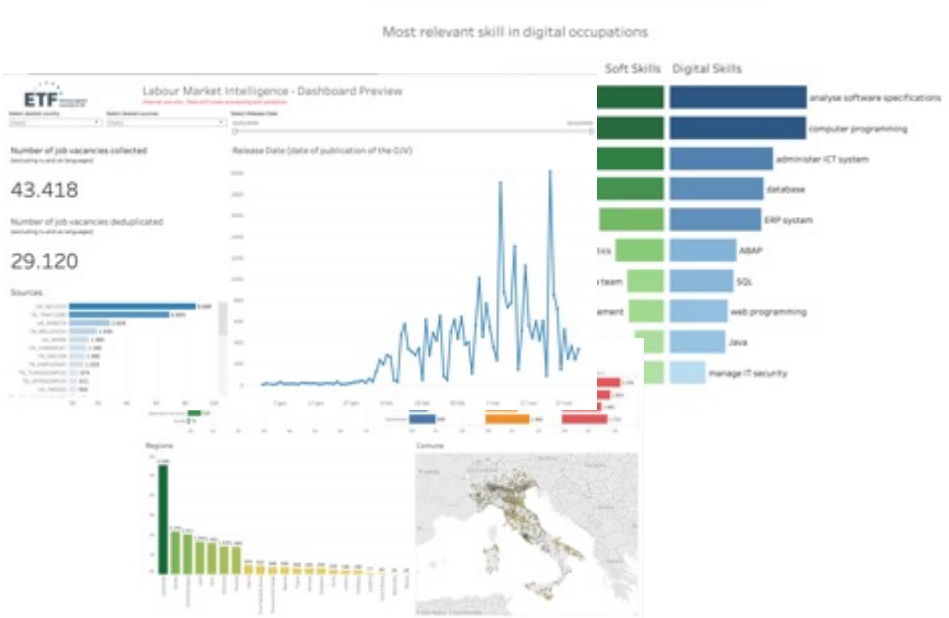
It would be great to use those information in addition to better and deeper understand how Labour Market is evolving in a given country, even compared to other countries

Our Goal

Transform Online Job Advertisments…



…in insights and analytics

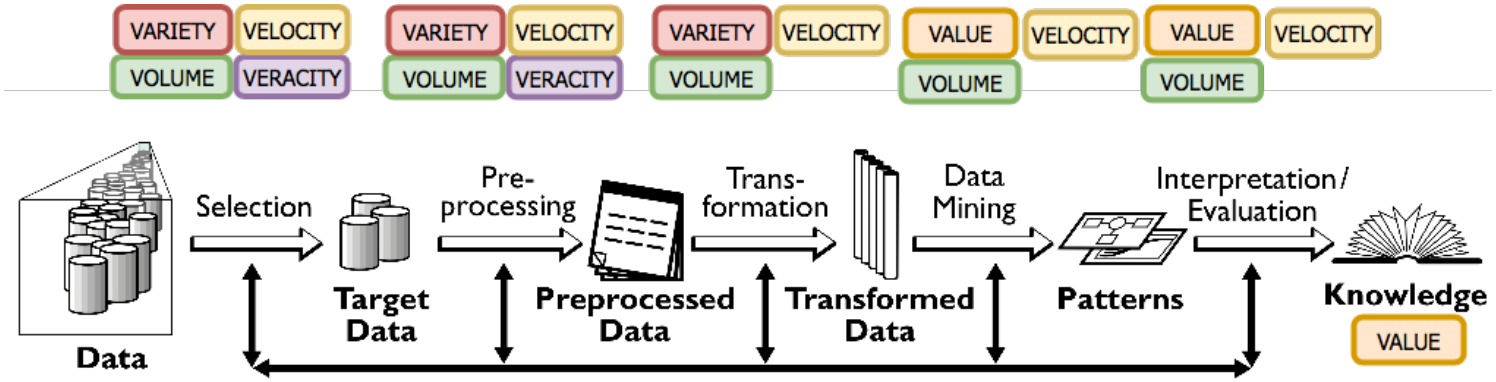


Challenges

- Handle a huge **amount** of near real time data
- Data coming from web → Need to detect and reduce **noise**
- **Multi language** environment
- Need to relate to **classification standards**
- Find a way to **summarize and present** a wide and complex scenario

Methodological background

KDD – Fayyad, 1997



VARIETY

Unstructured data (plain text to be processed)

VELOCITY

Real time data

VOLUME

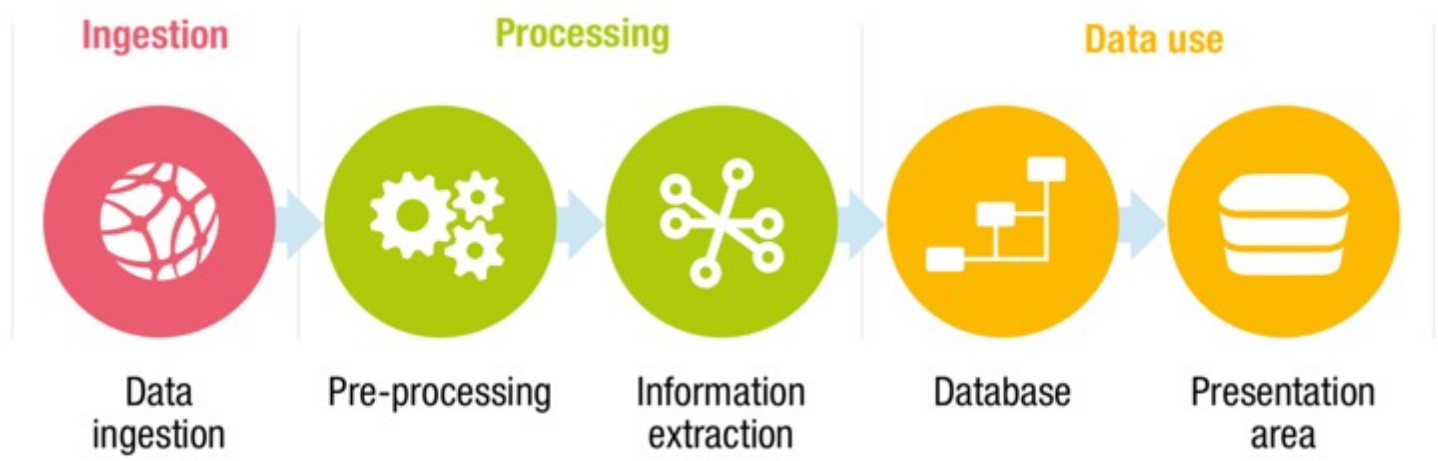
Huge amount of data (Terabytes)

VERACITY

Data is noisy, uncontrolled

Our Approach

KDD 4 LMI



Topics

1. Goal & context
2. Challenges
    1. **Stakeholders**
    2. The functional architecture
    3. Data ingestion techniques
    4. Data processing pipeline
    5. Classification techniques
3. Outputs
4. Outcomes

Stakeholders

Project Leader

Key Users

Domain Experts

End Users

Project leader

- ETF
  - Lead the project with the steering committee
  - Define the scope of the project
  - Define key organizations
  - Maintain relations with EU stakeholders
  - Provide advice

Key Users

- ETF, EMSIBG
  - Define requirements
  - Monitor quality of the project
  - Provide input to the development of the project
  - Manage the landscaping
  - Validate overall data flow and methodology

Domain Experts

- International Country Experts
  - Provide the knowledge and expertise
  - Execute the landscaping
  - Understand the language/terms of their context
  - Evaluate the accuracy of the results
  - Test the product
  - Provide feedback

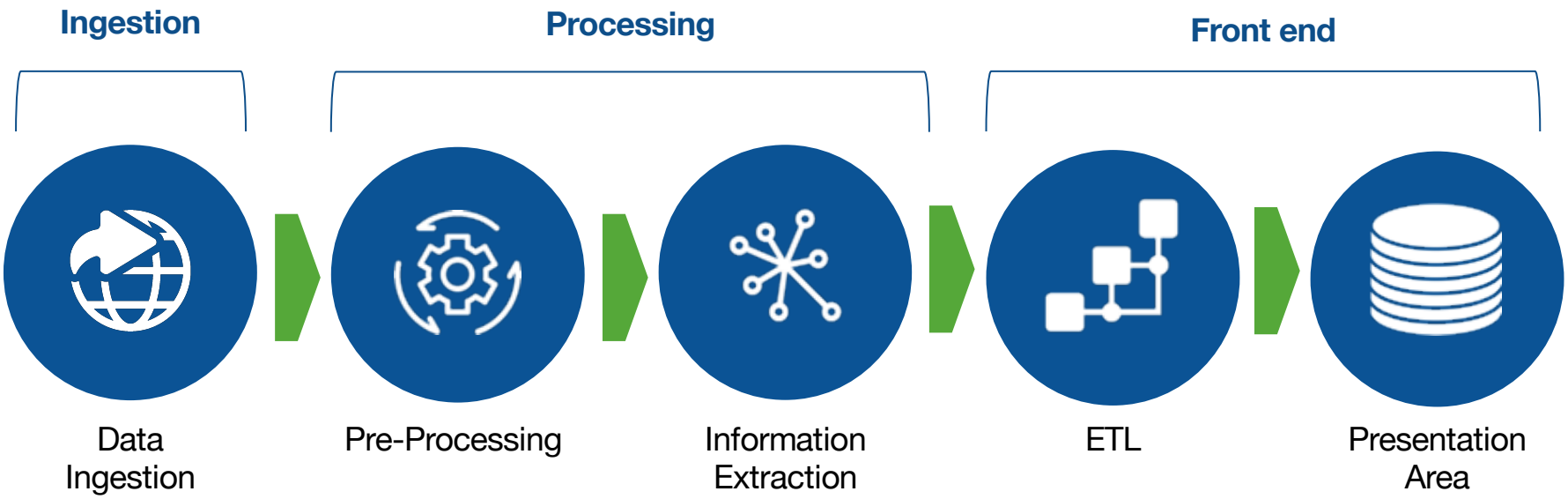
End Users

- Decision Makers and Business Users
  - (Visual) Explore dataset, analysis and aggregate data
  - Define new analysis processes
  - Produce Data storytelling
  - Make decisions by exploring data
- Data Scientists
  - Apply new machine learning models and AI techniques
  - Extract new insights from the data
  - Apply advanced data modelling to the dataset
- Data Analysts
  - Interprets data and turns it into information
  - Identifying patterns and trends
  - Extract and analyze aggregate data
  - Publish and share their analysis

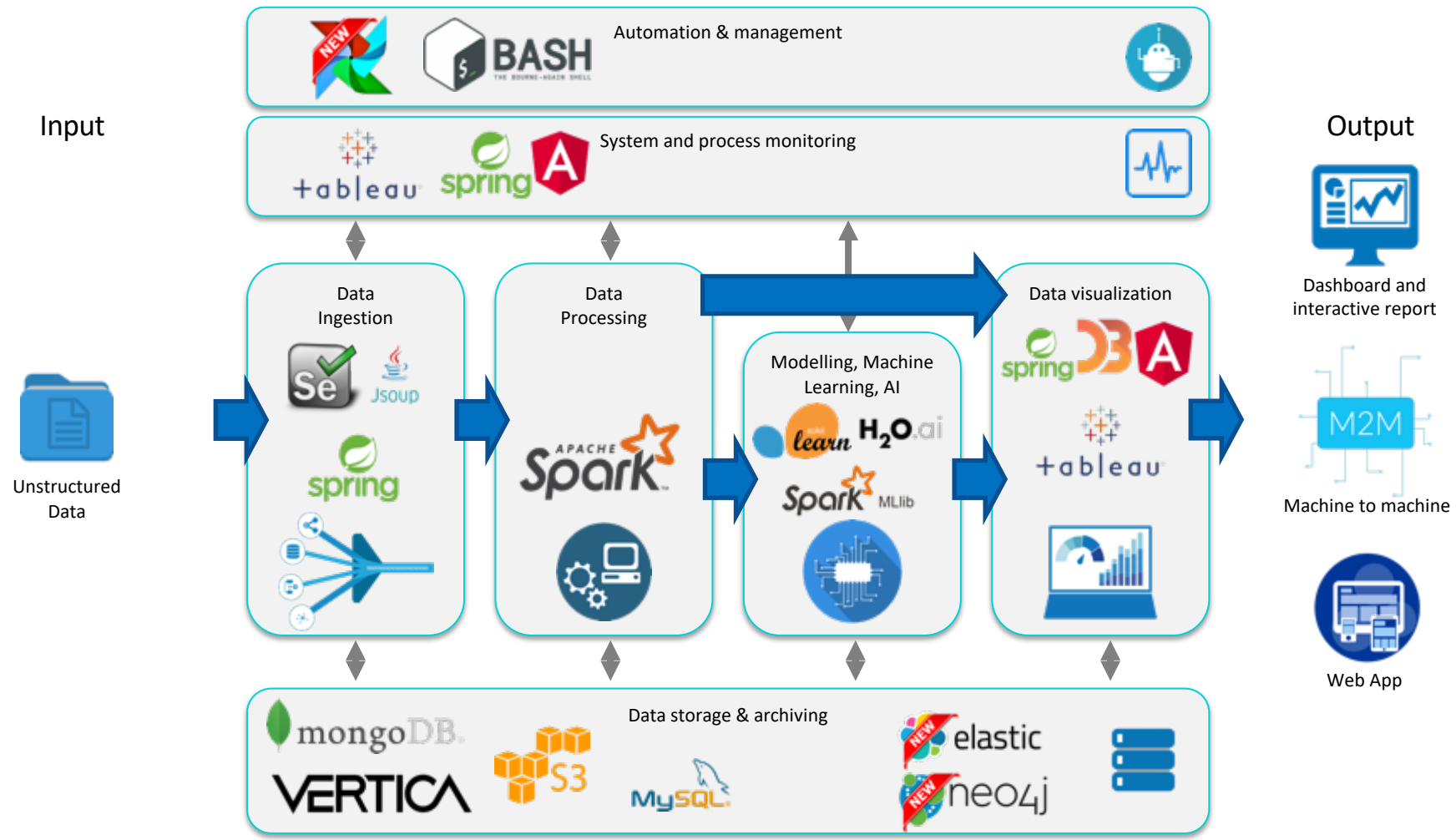
Topics

1. Goal & context
2. Challenges
    1. Stakeholders
    2. **The functional architecture**
    3. Data ingestion techniques
    4. Data processing pipeline
    5. Classification techniques
3. Outputs
4. Outcomes

Overall Data Flow



Technology view



Key components

- Data ingestion: collect raw data from OJV in both structured and unstructured (raw text) formats
- Data processing: classify data through machine learning techniques
- Data analysis: extract information from data and make it available through visualization

Infrastructure Challenges

- Manage multiple **parallel ingestion** activities
- Availability of **high performance** computational infrastructure **at a glance**
- **High memory** requirements
- High **storage** volumes to store source and staging data
- Big data environment
- **Scalable** architecture

Topics

1. Goal & context
2. Challenges
    1. Stakeholders
    2. The functional architecture
    3. **Data ingestion techniques**
    4. Data processing pipeline
    5. Classification techniques
3. Outputs
4. Outcomes

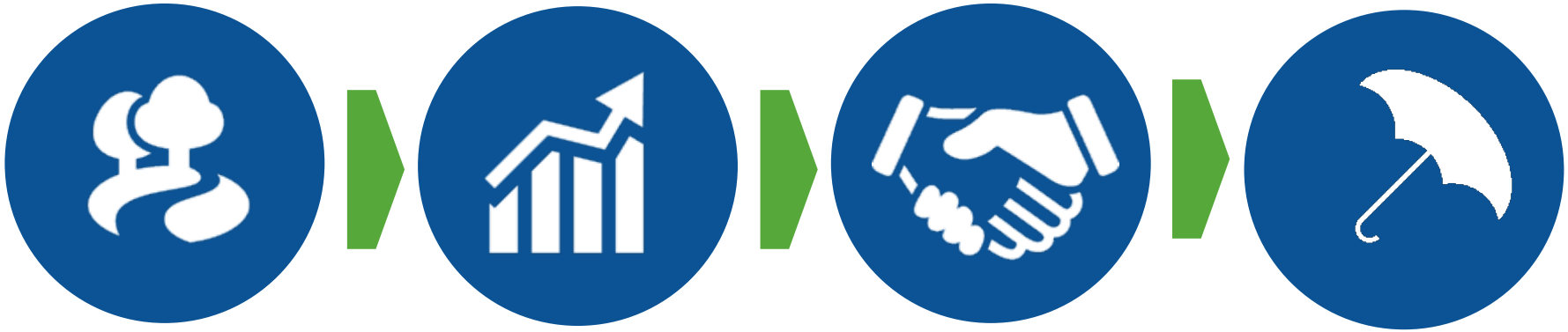
Landscaping

A **Landscaping activity** is performed to produce a list of **sources** (web portals) that are relevant for the Web Labour Market in a given country.

A Country Expert **validates** this list, that will become the initial step of the LMI System

Source selection strategy

4 Processing Steps



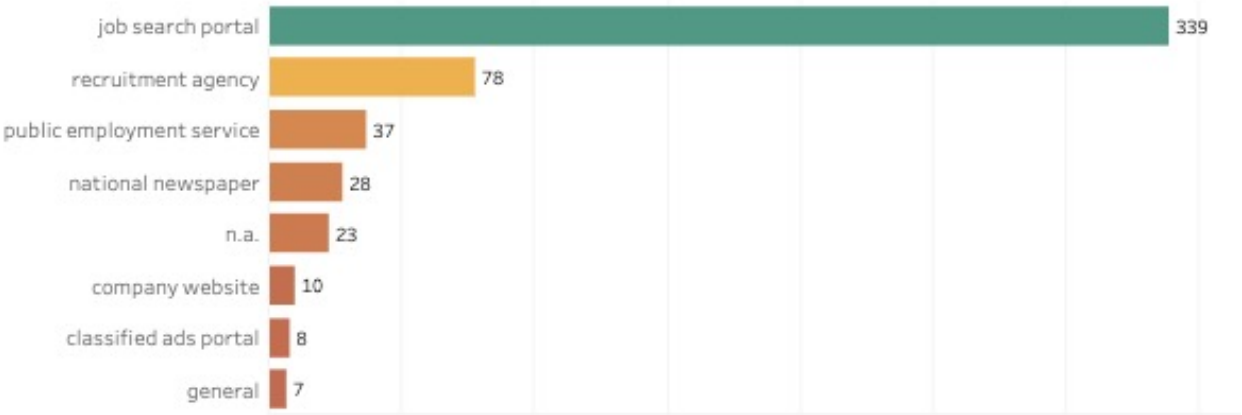
Source selection in landscaping

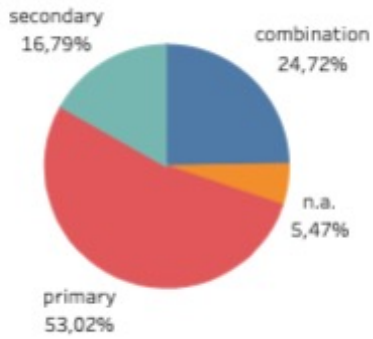
Augmentation

Agreements

Coverage

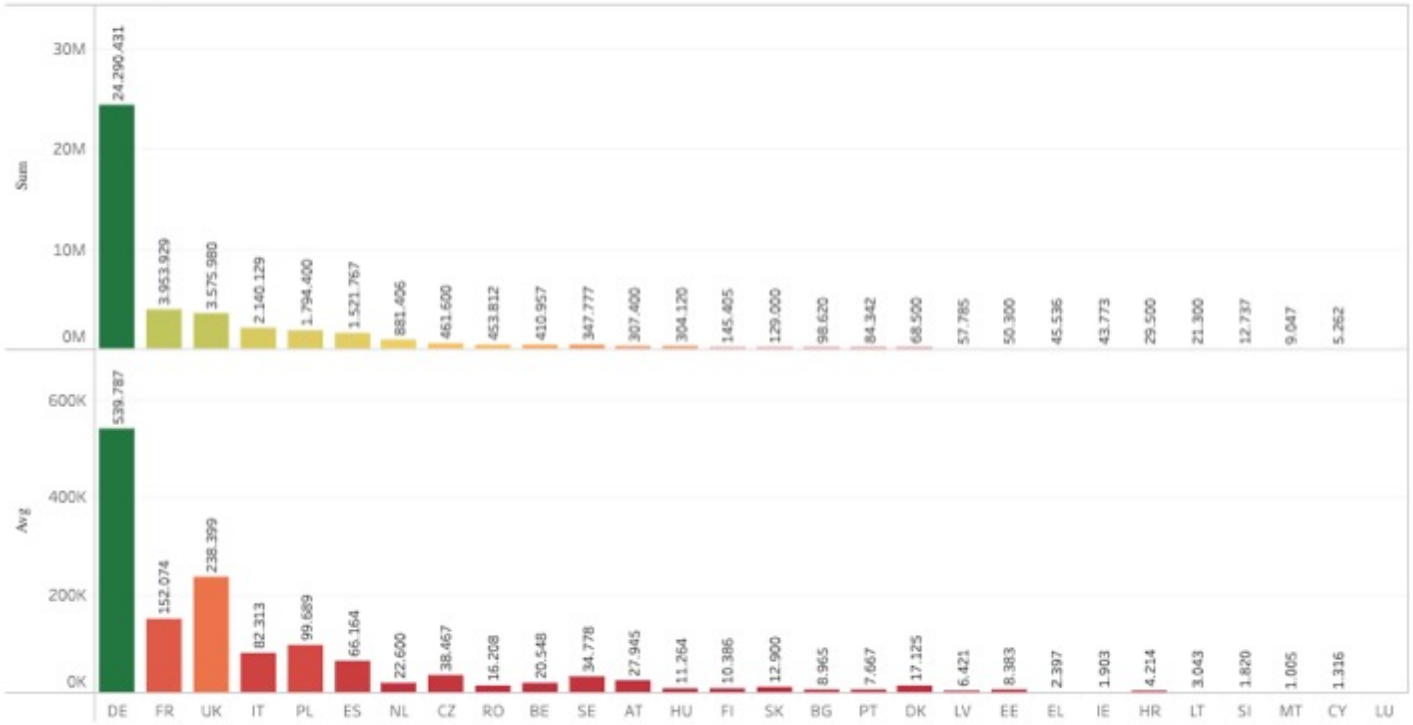
Sites by type of operator

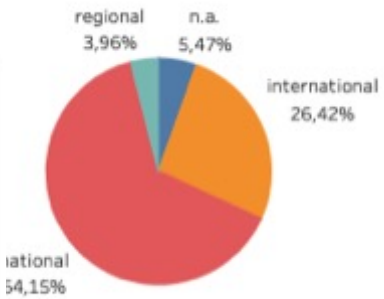


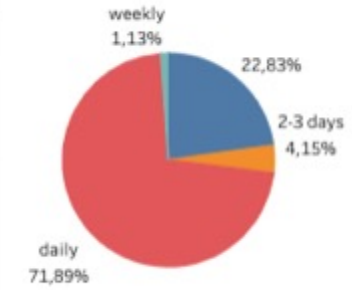


Vacancy volume by country

(estimated by ICE)







Augmentation

We analysed the results of the landscaping activity

- Completing the mapping of transnational sources
- Adding further transnational sources

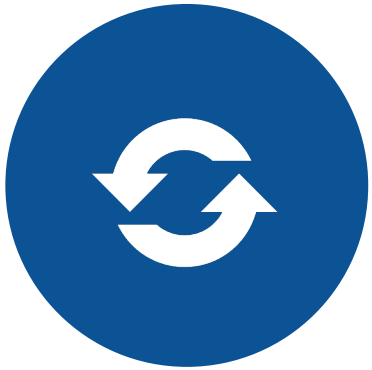
In order to define

- a priority list to define agreements
- a relevance order to realize data ingestion channels

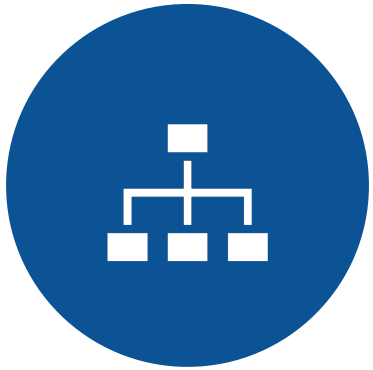
Relevance and ranking of sources

Volume

Type of
web portal



Data
Update



Structured
Data

Data Ingestion phase

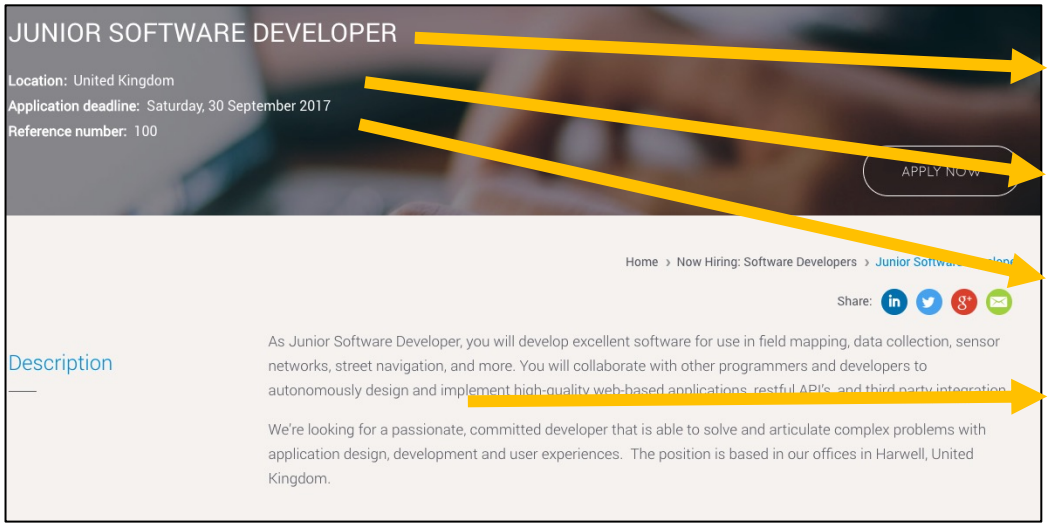
The process of obtaining and importing data from web portals and storing them in a Database

Focus on volumes

Coverage augmentation & maximization

Direct agreements with the most relevant sources

An example



Title: Junior Software Developer

Area: United Kingdom

Time: Saturday, 30 September 2017

Description: As Junior Software Developer, you will develop excellent software for use …

Topics

1. Goal & context
2. Challenges
    1. Stakeholders
    2. The functional architecture
    3. Data ingestion techniques
    4. **Data processing pipeline**
    5. Classification techniques
3. Outputs
4. Outcomes

Data Pre-Processing – Challenges & Definitions

The process of cleaning ingested data and dedupicating OJVs, to guarantee that analytical phase'll work on data at the highest quality possible

Language detection

Noise reduction

OJVs Deduplication

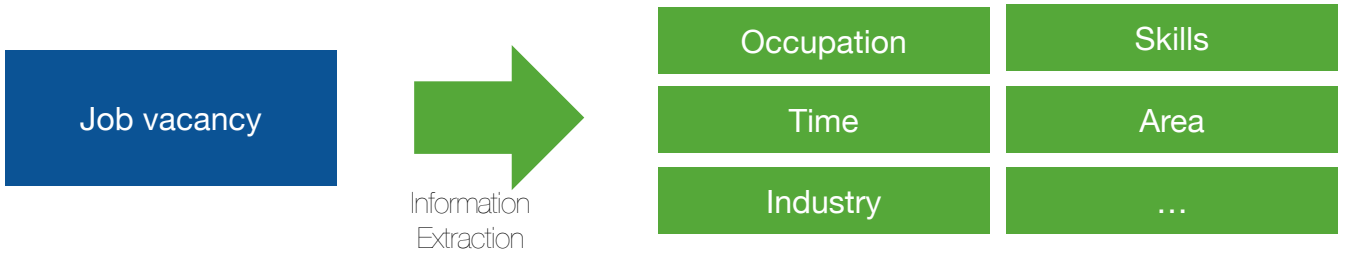
Topics

1. Goal & context
2. Challenges
    1. Stakeholders
    2. The functional architecture
    3. Data ingestion techniques
    4. Data processing pipeline
    5. **Classification techniques**
3. Outputs
4. Outcomes

Data Classification

- Goal:
  - Extract and structure information from data, to be provided to the presentation layer
- Challenges:
  - Handle massive amount of heterogeneous data written in different languages
- Approach:
  - Develop an adaptable framework, language dependent, tailored on different information features. Some relevant challenges:
    - **Occupation** feature classification: combined methods such as Machine Learning, Topic Modeling and Unsupervised Learning
    - **Skill** feature classification: another different combined methods, such as Text Analysis with corpus based or Knowledge based similarity
- Features:
  - Guarantee Explainable information extraction, logging classification methods and relevant features.

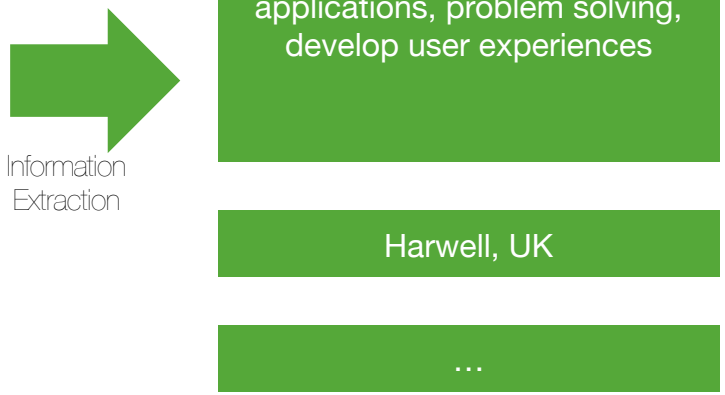
Data Classification - An example



Junior Software Developer

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.



Topics

1. Goal & context
2. Challenges
    1. Stakeholders
    2. The functional architecture
    3. Data ingestion techniques
    4. Data processing pipeline
    5. Classification techniques
3. **Outputs**
4. Outcomes

DataLab Model

2 Tables

FT Document

1 Row for each:
- General_ID → Key
- OJV
- Source
- Place

Why? Because, for each OJV, we can detect a multi-place Vacancy

e.g. «Software Developer in London / Liverpool»

FT Skill Analysis

1 Row for each:
- General_ID → Key
- OJV
- Source
- Place
- Skill

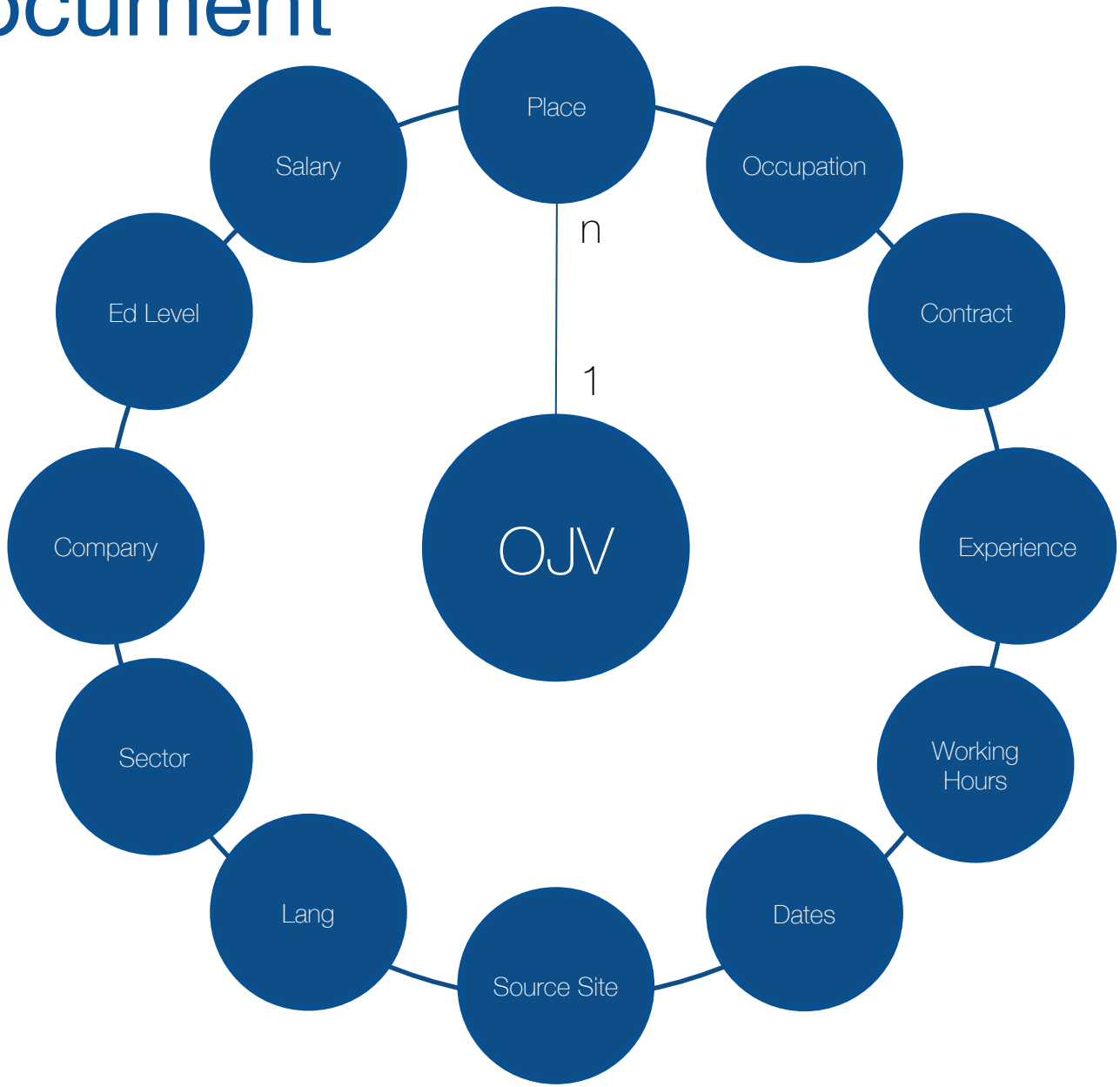
Why? Because, for each OJV, we can obviously detect multiple skills

e.g. «Software Developer in London / Liverpool, with customer orientation culture, that speaks english and tolerates stress»

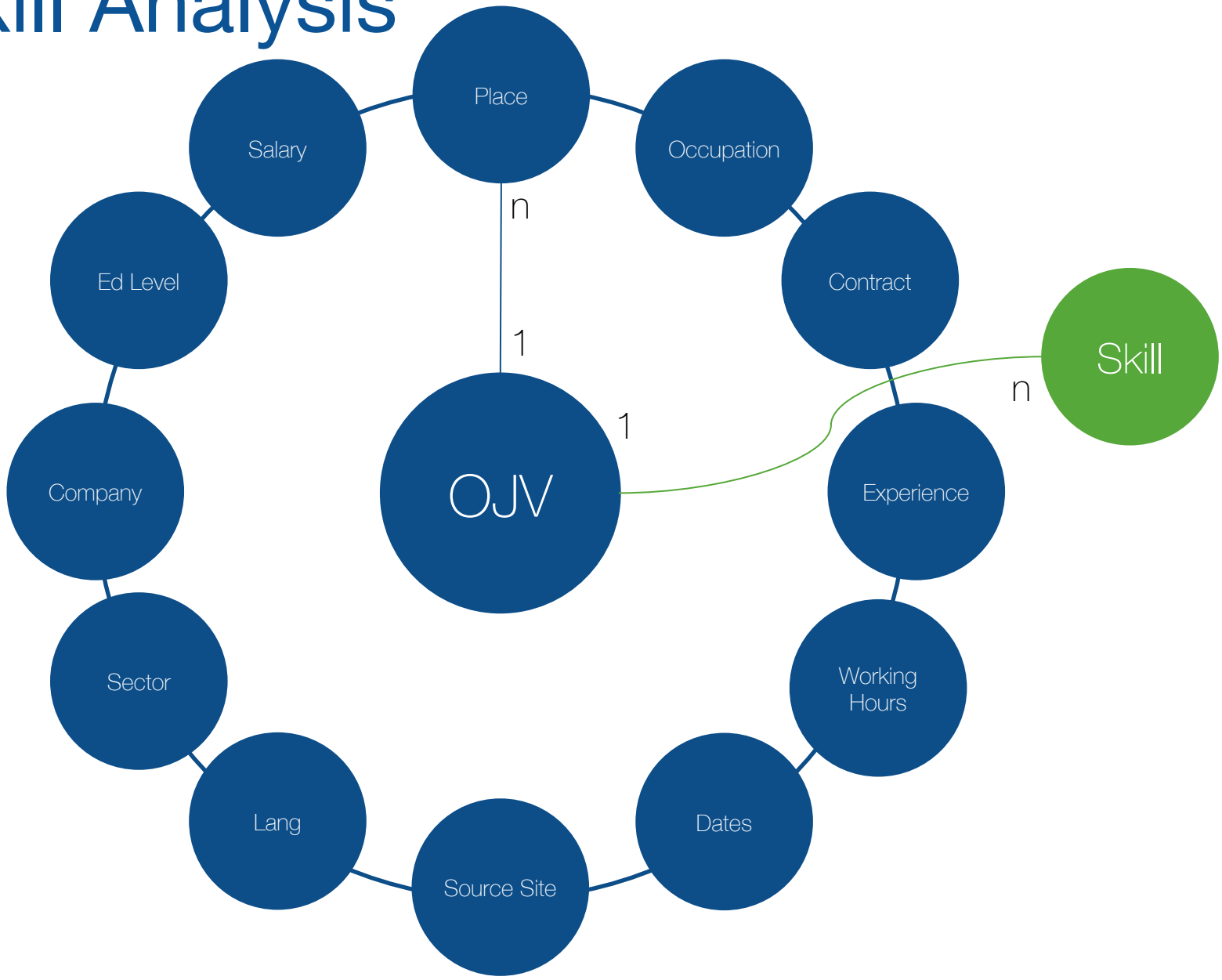
DataLab Model

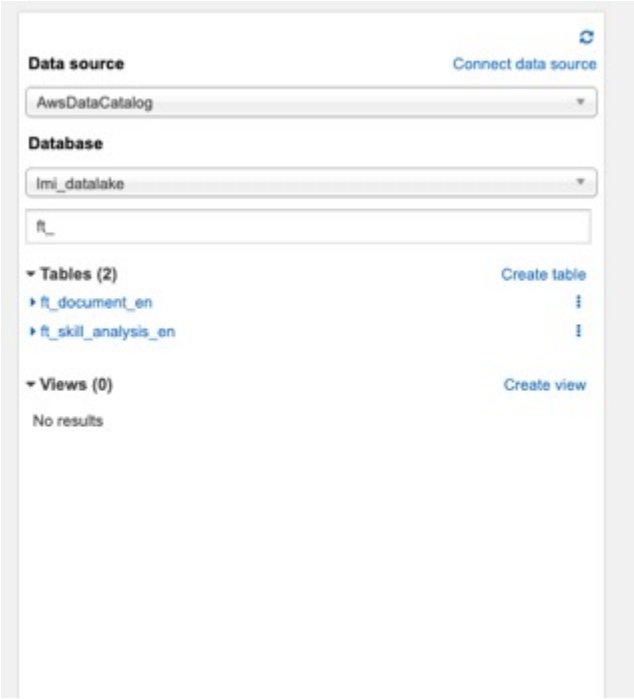
So, to have the number of job vacancies you always have to compute an unique count by General_ID

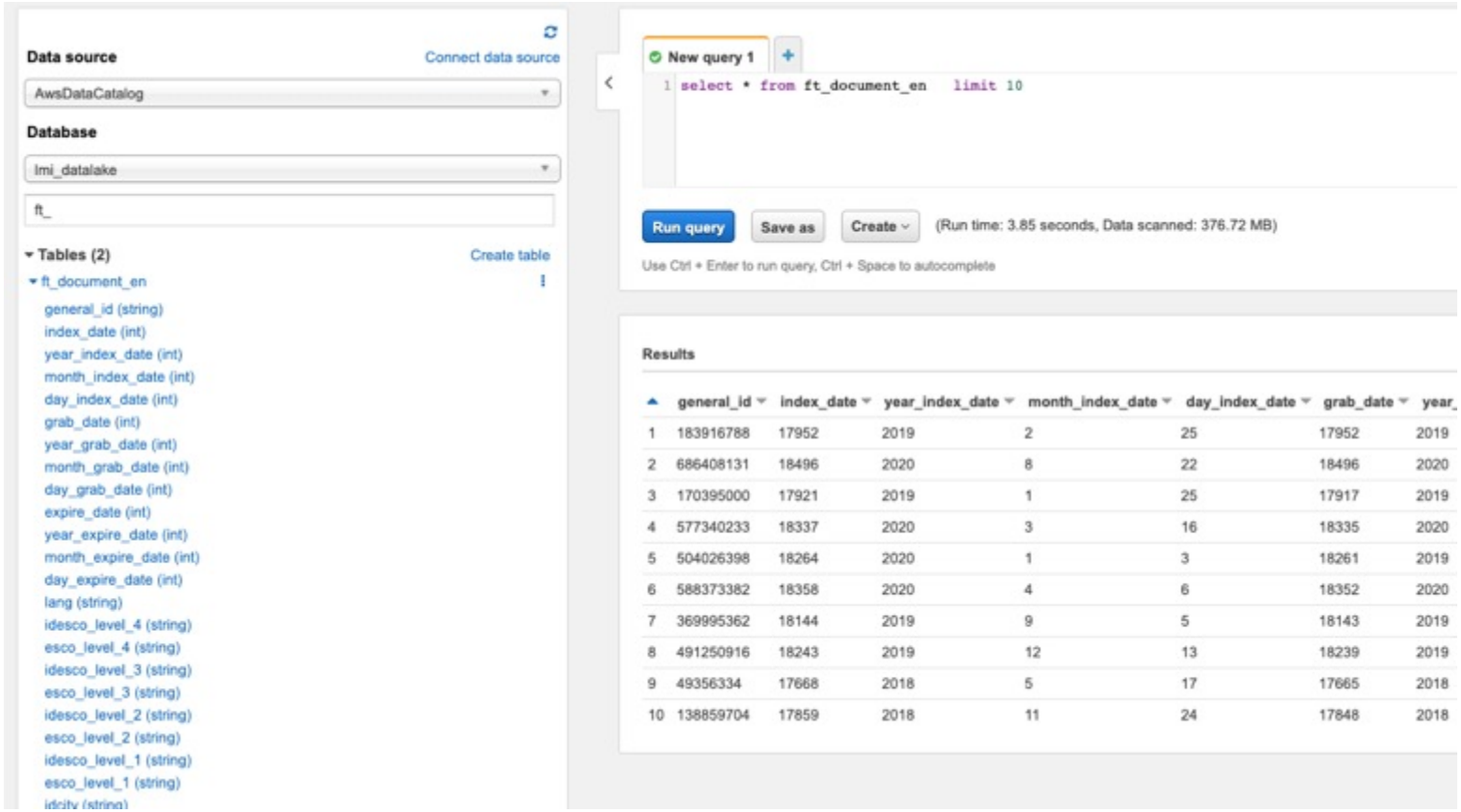
FT Document

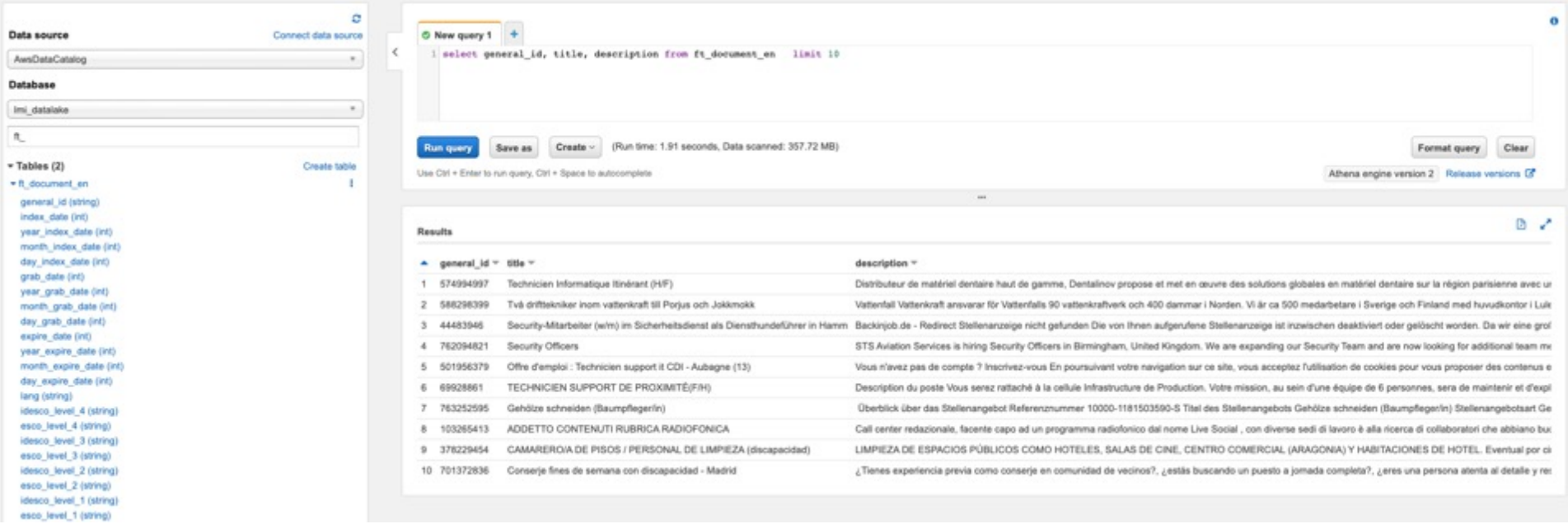

Place
Occupation
Contract
Experience
Working Hours
Dates
Source Site
Lang
Sector
Company
Ed Level
Salary
n
1
OJV

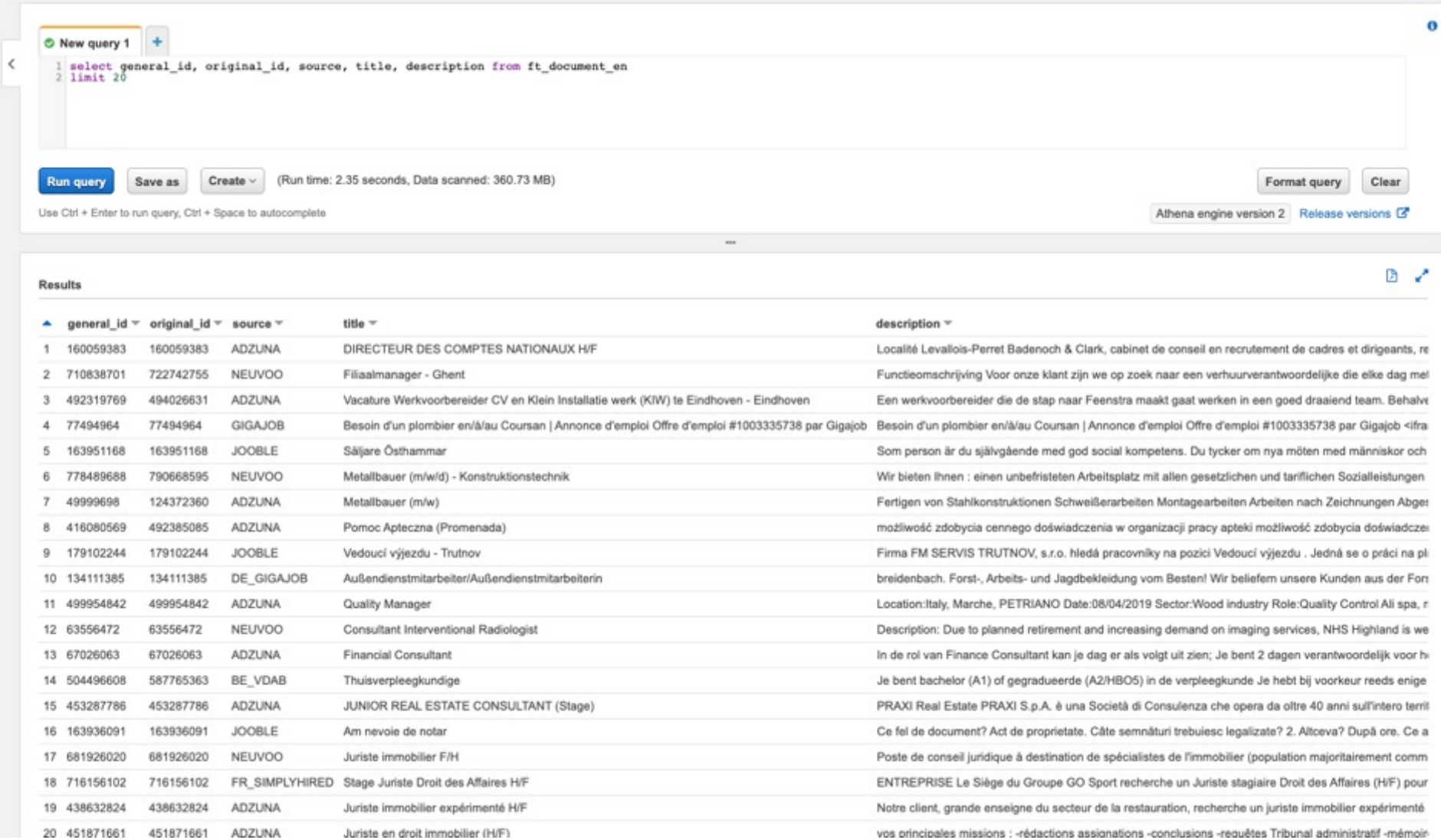
FT Skill Analysis




Data source
Connect data source
AwsDataCatalog
Database
lmi_datalake
ft_
Tables (2)
Create table
ft_document_en
ft_skill_analysis_en
Views (0)
Create view
No results


Data source
Connect data source
AwsDataCatalog
Database
lmi_datalake
ft_
Tables (2)
Create table
ft_document_en
general_id (string)
index_date (int)
year_index_date (int)
month_index_date (int)
day_index_date (int)
grab_date (int)
year_grab_date (int)
month_grab_date (int)
day_grab_date (int)
expire_date (int)
year_expire_date (int)
month_expire_date (int)
day_expire_date (int)
lang (string)
idesco_level_4 (string)
esco_level_4 (string)
idesco_level_3 (string)
esco_level_3 (string)
idesco_level_2 (string)
esco_level_2 (string)
idesco_level_1 (string)
esco_level_1 (string)
idcity (string)
New query 1
1 select * from ft_document_en limit 10
Run query
Save as
Create
(Run time: 3.85 seconds, Data scanned: 376.72 MB)
Use Ctrl + Enter to run query, Ctrl + Space to autocomplete
Results
general_id index_date year_index_date month_index_date day_index_date grab_date year_
1 183916788 17952 2019 2 25 17952 2019
2 686408131 18496 2020 8 22 18496 2020
3 170395000 17921 2019 1 25 17917 2019
4 577340233 18337 2020 3 16 18335 2020
5 504026398 18264 2020 1 3 18261 2019
6 588373382 18358 2020 4 6 18352 2020
7 369995362 18144 2019 9 5 18143 2019
8 491250916 18243 2019 12 13 18239 2019
9 49356334 17668 2018 5 17 17665 2018
10 138859704 17859 2018 11 24 17848 2018


Data source
Connect data source
AwsDataCatalog
Database
imi_datalake
ft_
Tables (2)
Create table
ft_document_en
general_id (string)
index_date (int)
year_index_date (int)
month_index_date (int)
day_index_date (int)
grab_date (int)
year_grab_date (int)
month_grab_date (int)
day_grab_date (int)
expire_date (int)
year_expire_date (int)
month_expire_date (int)
day_expire_date (int)
lang (string)
idesco_level_4 (string)
esco_level_4 (string)
idesco_level_3 (string)
esco_level_3 (string)
idesco_level_2 (string)
esco_level_2 (string)
idesco_level_1 (string)
esco_level_1 (string)
New query 1
1 select general_id, title, description from ft_document_en limit 10
Run query
Save as
Create
(Run time: 1.91 seconds, Data scanned: 357.72 MB)
Format query
Clear
Use Ctrl + Enter to run query, Ctrl + Space to autocomplete
Athena engine version 2
Release versions
Results
general_id
title
description
1 574994997 Technicien Informatique Itinérant (H/F) Distributeur de matériel dentaire haut de gamme, Dentalinov propose et met en œuvre des solutions globales en matériel dentaire sur la région parisienne avec ur
2 586298399 Två driftstekniker inom vattenkraft till Porjus och Jokkmokk Vattenfall Vattenkraft ansvarar för Vattenfalls 90 vattenkraftverk och 400 dammar i Norden. Vi är ca 500 medarbetare i Sverige och Finland med huvudkontor i Lule
3 44483946 Security-Mitarbeiter (w/m) im Sicherheitsdienst als Diensthundeführer in Hamm Backinjob.de - Redirect Stellenanzeige nicht gefunden Die von Ihnen aufgerufene Stellenanzeige ist inzwischen deaktiviert oder gelöscht worden. Da wir eine grol
4 762094821 Security Officers STS Aviation Services is hiring Security Officers in Birmingham, United Kingdom. We are expanding our Security Team and are now looking for additional team me
5 501956379 Offre d'emploi : Technicien support it CDI - Aubagne (13) Vous n'avez pas de compte ? Inscrivez-vous En poursuivant votre navigation sur ce site, vous acceptez l'utilisation de cookies pour vous proposer des contenus e
6 69928861 TECHNICIEN SUPPORT DE PROXIMITÉ(F/H) Description du poste Vous serez rattaché à la cellule Infrastructure de Production. Votre mission, au sein d'une équipe de 6 personnes, sera de maintenir et d'expl
7 763252595 Gehölze schneiden (Baumpfleger/in) Überblick über das Stellenangebot Referenznummer 10000-1181503590-S Titel des Stellenangebots Gehölze schneiden (Baumpfleger/in) Stellenangebotsart Ge
8 103265413 ADDETTO CONTENUTI RUBRICA RADIOFONICA Call center redazionale, facente capo ad un programma radiofonico dal nome Live Social , con diverse sedi di lavoro è alla ricerca di collaboratori che abbiano bu
9 378229454 CAMARERO/A DE PISOS / PERSONAL DE LIMPIEZA (discapacidad) LIMPIEZA DE ESPACIOS PÚBLICOS COMO HOTELES, SALAS DE CINE, CENTRO COMERCIAL (ARAGONIA) Y HABITACIONES DE HOTEL. Eventual por ci
10 701372836 Conserje fines de semana con discapacidad - Madrid ¿Tienes experiencia previa como conserje en comunidad de vecinos?, ¿estás buscando un puesto a jornada completa?, ¿eres una persona atenta al detalle y re



	general_id	original_id	source	title	description
1	160059383	160059383	ADZUNA	DIRECTEUR DES COMPTES NATIONAUX H/F	Localité Levallois-Perret Badenoch & Clark, cabinet de conseil en recrutement de cadres et dirigeants, re
2	710838701	722742755	NEUVOO	Filiaalmanager - Ghent	Functieomschrijving Voor onze klant zijn we op zoek naar een verhuurverantwoordelijke die elke dag met
3	492319769	494026631	ADZUNA	Vacature Werkvoorbereider CV en Klein Installatie werk (KIW) te Eindhoven - Eindhoven	Een werkvoorbereider die de stap naar Feenstra maakt gaat werken in een goed draaiend team. Behalve
4	77494964	77494964	GIGAJOB	Besoin d'un plombier en/à/au Coursan | Annonce d'emploi Offre d'emploi #1003335738 par Gigajob	Besoin d'un plombier en/à/au Coursan | Annonce d'emploi Offre d'emploi #1003335738 par Gigajob <ifra
5	163951168	163951168	JOOBLE	Säljare Östhammar	Som person är du självgående med god social kompetens. Du tycker om nya möten med människor och
6	778489688	790668595	NEUVOO	Metallbauer (m/w/d) - Konstruktionstechnik	Wir bieten Ihnen : einen unbefristeten Arbeitsplatz mit allen gesetzlichen und tariflichen Sozialleistungen
7	49999698	124372360	ADZUNA	Metallbauer (m/w)	Fertigen von Stahlkonstruktionen Schweißerarbeiten Montagearbeiten Arbeiten nach Zeichnungen Abges
8	416080569	492385085	ADZUNA	Pomoc Apteczna (Promenada)	możliwość zdobycia cennego doświadczenia w organizacji pracy apteki możliwość zdobycia doświadczen
9	179102244	179102244	JOOBLE	Vedoucí výjezdu - Trutnov	Firma FM SERVIS TRUTNOV, s.r.o. hledá pracovníky na pozici Vedoucí výjezdu . Jedná se o práci na pl
10	134111385	134111385	DE_GIGAJOB	Außendienstmitarbeiter/Außendienstmitarbeiterin	breidenbach. Forst-, Arbeits- und Jagdbekleidung vom Besten! Wir beliefern unsere Kunden aus der Fors
11	499954842	499954842	ADZUNA	Quality Manager	Location:Italy, Marche, PETRIANO Date:08/04/2019 Sector:Wood industry Role:Quality Control Ali spa, r
12	63556472	63556472	NEUVOO	Consultant Interventional Radiologist	Description: Due to planned retirement and increasing demand on imaging services, NHS Highland is we
13	67026063	67026063	ADZUNA	Financial Consultant	In de rol van Finance Consultant kan je dag er als volgt uit zien; Je bent 2 dagen verantwoordelijk voor h
14	504496608	587765363	BE_VDAB	Thuisverpleegkundige	Je bent bachelor (A1) of gegradueerde (A2/HBO5) in de verpleegkunde Je hebt bij voorkeur reeds enige
15	453287786	453287786	ADZUNA	JUNIOR REAL ESTATE CONSULTANT (Stage)	PRAXI Real Estate PRAXI S.p.A. è una Società di Consulenza che opera da oltre 40 anni sull'intero territ
16	163936091	163936091	JOOBLE	Am nevoie de notar	Ce fel de document? Act de proprietate. Câte semnături trebuiesc legalizate? 2. Altceva? După ore. Ce a
17	681926020	681926020	NEUVOO	Juriste immobilier F/H	Poste de conseil juridique à destination de spécialistes de l'immobilier (population majoritairement comm
18	716156102	716156102	FR_SIMPLYHIRED	Stage Juriste Droit des Affaires H/F	ENTREPRISE Le Siège du Groupe GO Sport recherche un Juriste stagiaire Droit des Affaires (H/F) pour
19	438632824	438632824	ADZUNA	Juriste immobilier expérimenté H/F	Notre client, grande enseigne du secteur de la restauration, recherche un juriste immobilier expérimenté
20	451871661	451871661	ADZUNA	Juriste en droit immobilier (H/F)	vos principales missions : -rédactions assignations -conclusions -requêtes Tribunal administratif -mémoir

Topics

1. Goal & context
2. Challenges
    1. Stakeholders
    2. The functional architecture
    3. Data ingestion techniques
    4. Data processing pipeline
    5. Classification techniques
3. Outputs
4. **Outcomes**

Data Visualization. Just graphic?

- Goal:
  - Provide the right too for the right stackeholder
- Challenges:
  - Find a way to support different needs on such a relevant amount of information
- Approach:
  - Define several data visualization and analysis approaches:
    - **Infographics**: appealing, static and widley understandable
    - **Public portals for citizens:** easy, fast and high level informative
    - **Dashboard:** deeper informative, web based, for decision makers
    - **Self-service analysis labs**: access data, highest informative opportunity, requires domain/technical/analytical skillset

Live session