Большие данные для аналитики рынка труда

День 1, занятие 1

Анализ потребностей на основе данных из онлайн-вакансий

Алессандро Ваккарино — Мауро Пелуччи

22 ноября 2021

Темы

1. Цель и контекст
2. Задачи
    1. Участники
    2. Функциональная архитектура
    3. Методы приема данных
    4. Конвейер обработки данных
    5. Методы классификации
3. Выходные продукты
4. Результаты

Темы

1. **Цель и контекст**
2. Задачи
    1. Участники
    2. Функциональная архитектура
    3. Методы приема данных
    4. Конвейер обработки данных
    5. Методы классификации
3. Выходные продукты
4. Результаты

Контекст

Постоянно развивающийся рынок труда:

- Цифровизация профессий
- Востребованность межличностных умений
- Интернационализация
- Появление новых профессий и умений
- Работа в режиме гибкого рабочего времени и дистанционная работа
- Влияние пандемии COVID-19
- …

Нужно *что-то*, что будет помогать нам анализировать рынок труда и наблюдать за тем, **как** он развивается, а также помогать ответственным лицам **принимать нужные решения в нужное время**

Что у нас есть / что нам нужно

У нас уже есть данные **официальной статистики**, которые являются:

- *репрезентативными*
- *значимыми* в плане ценности

Но нам пригодилась бы и **дополнительная информация**, которая могла бы быть:

- *Оперативной* — для отслеживания текущих событий (например, для анализа влияния COVID-19)
- *Подробной* и *соответствующей* реальным и текущим рыночным условиям — для выявления новых тенденций и анализа реальных потребностей компаний

Как найти такой дополнительный источник информации?
На **онлайн-рынке труда**!

Почему именно онлайн-рынок труда?

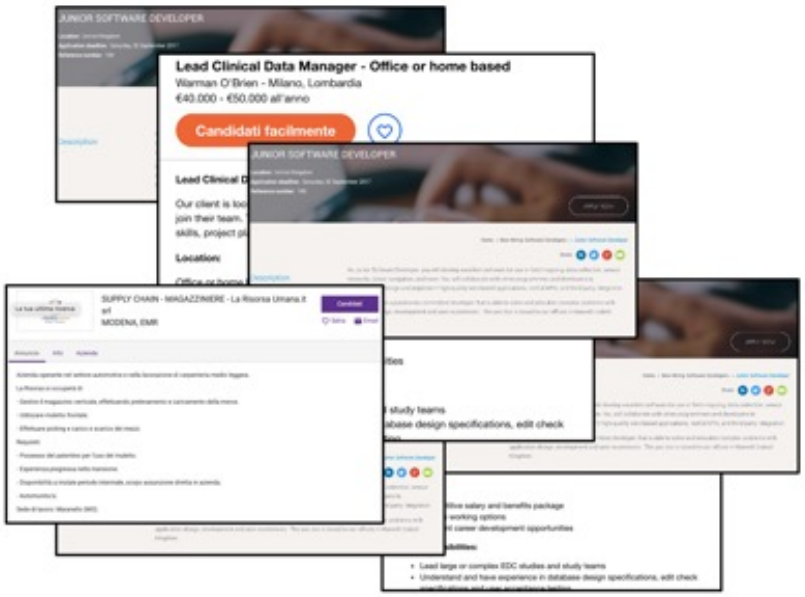
Он точно отражает то, что именно нужно компаниям в указанный период времени:

- Актуальность: компании размещают объявления тогда, когда они действительно ищут новых сотрудников
- Подробность: в объявлении максимально четко описывается конкретная потребность, в частности:
  - необходимая специальность
  - требования (умения, опыт, уровень образования и т. п.)
  - контекст работы (местоположение, тип договора, отрасль, график работы и т. п.)
- Соответствие реальности: используется принятая на рынке терминология — как в отношении профессии, так и в отношении умений. Это позволяет выявить новые термины, которые входят в обиход на рынке труда

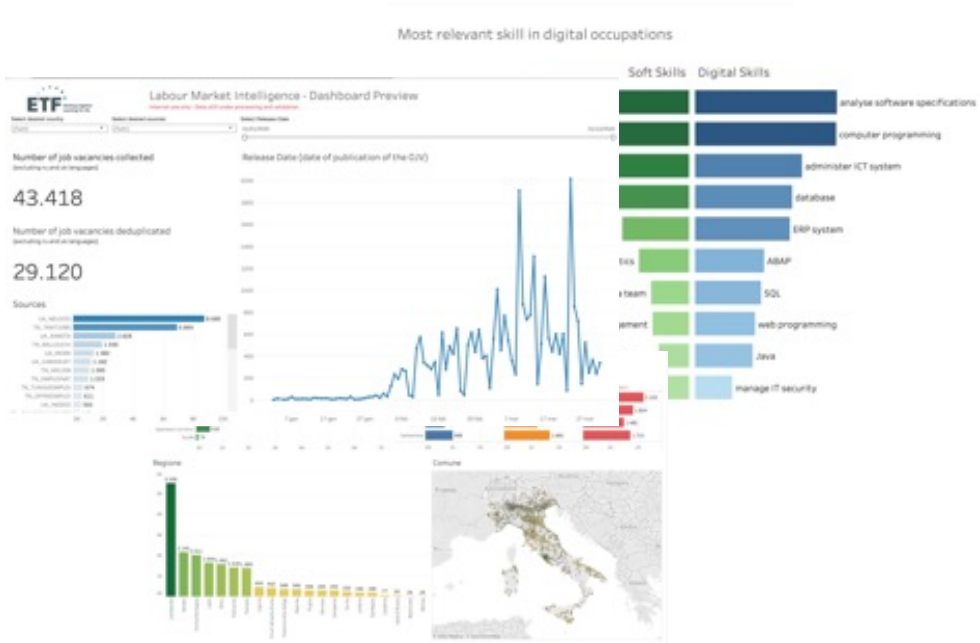
Было бы хорошо использовать эту дополнительную информацию, чтобы лучше и глубже понять, как развивается рынок труда в конкретной стране, и даже в сравнении с другими странами

Наша цель

Превратить онлайн-объявления о работе…



…в статистику и аналитику

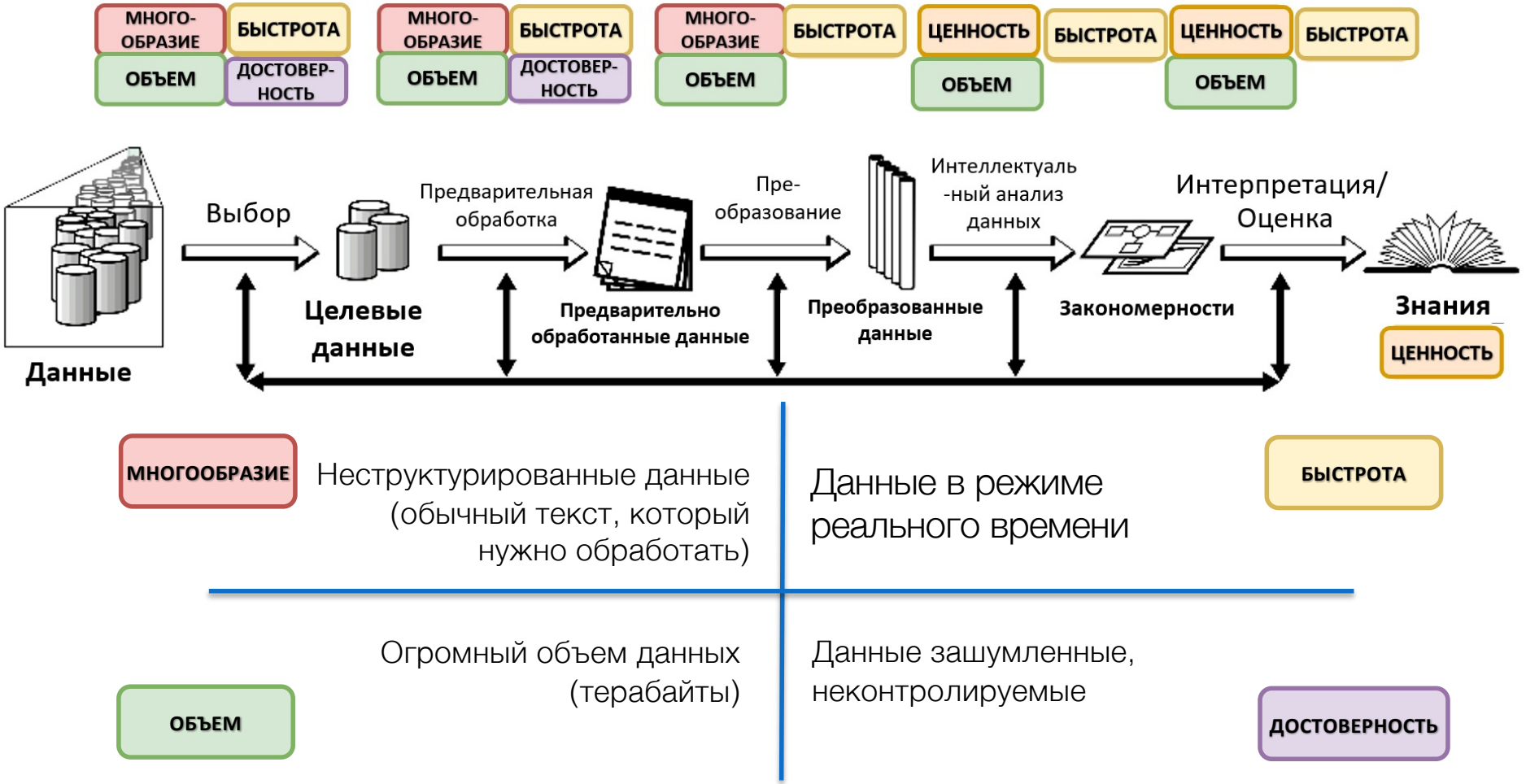


Задачи

- Обработка громадного **объема** данных, поступающих практически в режиме реального времени
- Данные из Интернета → Необходимость выявить **шум** и уменьшьить его количество
- **Многоязычная** среда
- Необходимость привязки к **стандартам классификации**
- Найти способ **сформулировать и представить** широкий и комплексный сценарий

Методологическая основа

ОЗБД (обнаружение знаний в базах данных) — Файяд, 1997



Наш подход

Обнаружение знаний в базах данных (ОЗБД) для аналитики рынка труда (АРТ)

Темы

1. Цель и контекст
2. Задачи
    1. **Участники**
    2. Функциональная архитектура
    3. Методы приема данных
    4. Конвейер обработки данных
    5. Методы классификации
3. Выходные продукты
4. Результаты

Участники

Руководитель проекта

Ключевые пользователи

Специалисты предметной области

Конечные пользователи

Руководитель проекта

- ЕФО
  - Руководит проектом вместе с координационным комитетом
  - Определяет сферу охвата проекта
  - Определяет ключевые организации
  - Поддерживает отношения с заинтересованными сторонами проекта из ЕС
  - Консультирует

Ключевые пользователи

- ЕФО, EMSIBG
  - Определяют требования
  - Следят за качеством выполнения проекта
  - Вносят вклад в развитие проекта
  - Руководят процессом изучения ландшафта
  - Проверяют весь поток данных и методологию

Специалисты предметной области

- Международные страновые эксперты
  - Привносят знания и опыт
  - Выполняют изучение ландшафта
  - Понимают язык и терминологию в своем контексте
  - Оценивают точность результатов
  - Тестируют продукт
  - Предоставляют обратную связь

Конечные пользователи

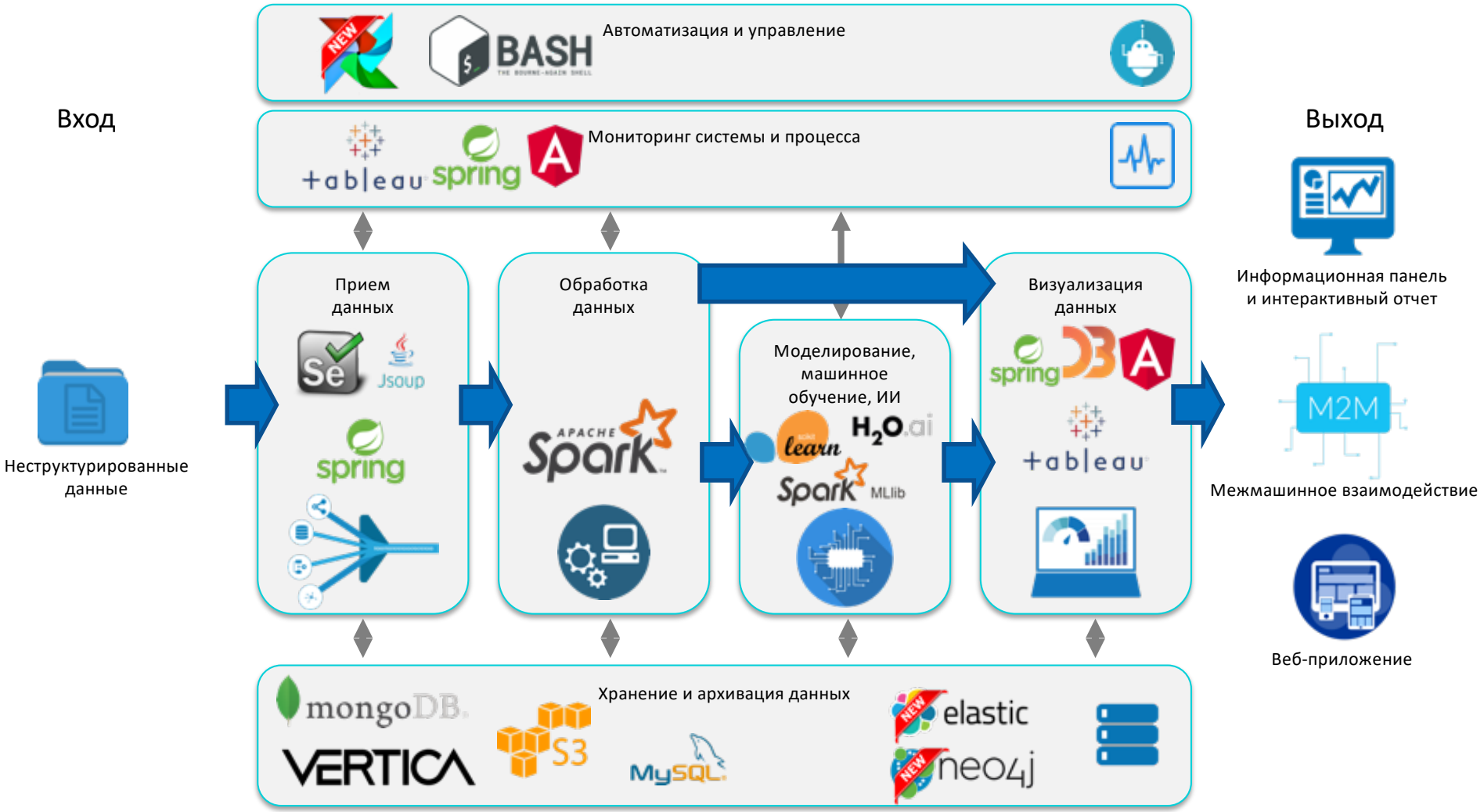
- Лица, ответственные за принятие решений, и корпоративные пользователи
  - (визуально) изучают набор данных, данные анализа и сводные данные
  - определяют новые процессы анализа
  - осуществляют сторителлинг данных
  - принимают решения на основе изучения данных
- Специалисты по обработке и анализу данных
  - применяют новые модели машинного обучения и методы ИИ
  - извлекают из данных новые сведения
  - применяют к наборам данных расширенное моделирование
- Аналитики данных
  - интерпретируют данные и превращают их в информацию
  - выявляют закономерности и тенденции
  - извлекают и анализируют сводные данные
  - публикуют и распространяют результаты своего анализа

Темы

1. Цель и контекст
2. Задачи
    1. Участники
    2. **Функциональная архитектура**
    3. Методы приема данных
    4. Конвейер обработки данных
    5. Методы классификации
3. Выходные продукты
4. Результаты

Общий поток данных

Технологическое представление



Ключевые компоненты

- Прием данных: сбор первичных данных из онлайн-вакансий в структурированном и неструктурированном (необработанный текст) виде
- Обработка данных: классификация данных при помощи методов машинного обучения
- Анализ данных: извлечение информации из данных и ее распространение посредством визуализации

Задачи, связанные с инфраструктурой

- Управление множеством параллельных процессов приема данных
- Доступность высокопроизводительной вычислительной инфраструктуры в любой момент
- Высокие требования к объему оперативной памяти
- Большие объемы хранилища для хранения исходных и промежуточных данных
- Среда больших данных
- Масштабируемая архитектура

Темы

1. Цель и контекст
2. Задачи
    1. Участники
    2. Функциональная архитектура
    3. **Методы приема данных**
    4. Конвейер обработки данных
    5. Методы классификации
3. Выходные продукты
4. Результаты

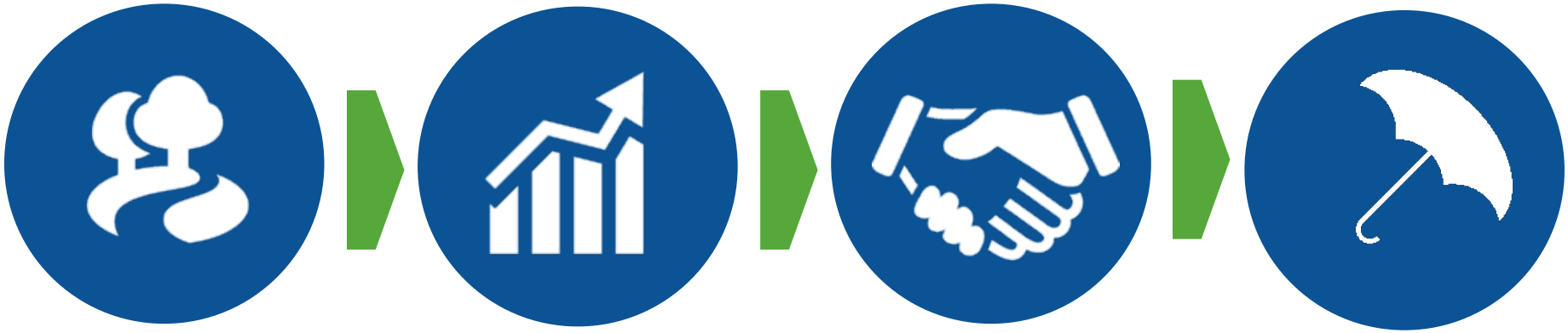
Изучение ландшафта

Работа по **изучению ландшафта** позволяет составить список **источников** (веб-порталов), актуальных для онлайн-рынка труда в определенной стране.

Этот список **утверждается** экспертом по данной стране и становится начальным этапом создания системы аналитики рынка труда (АРТ).

Стратегия выбора источников

4 этапа обработки



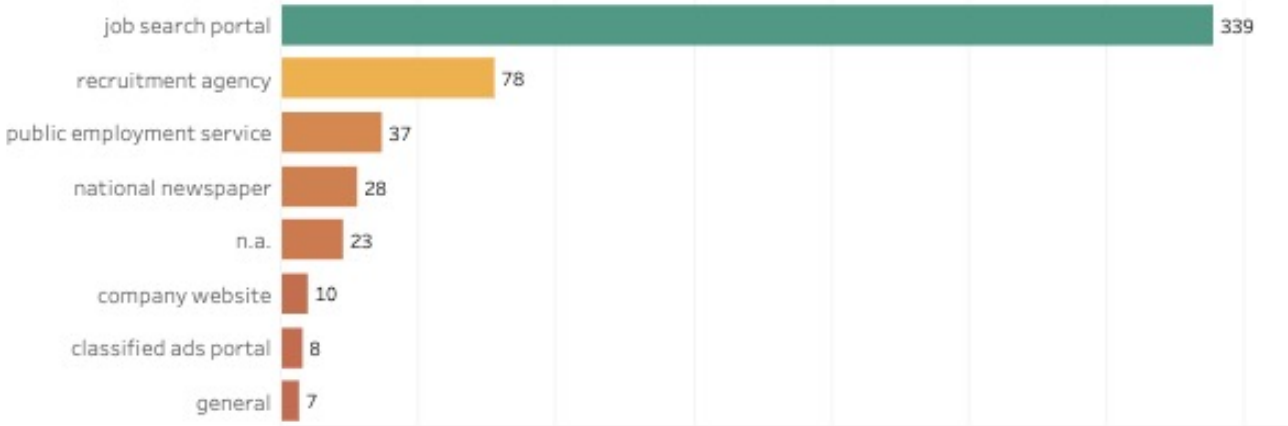
Выбор источников в рамках изучения ландшафта

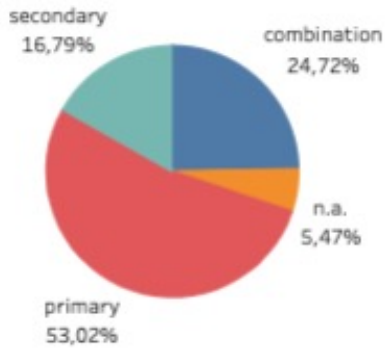
Расширение

Соглашения

Охват

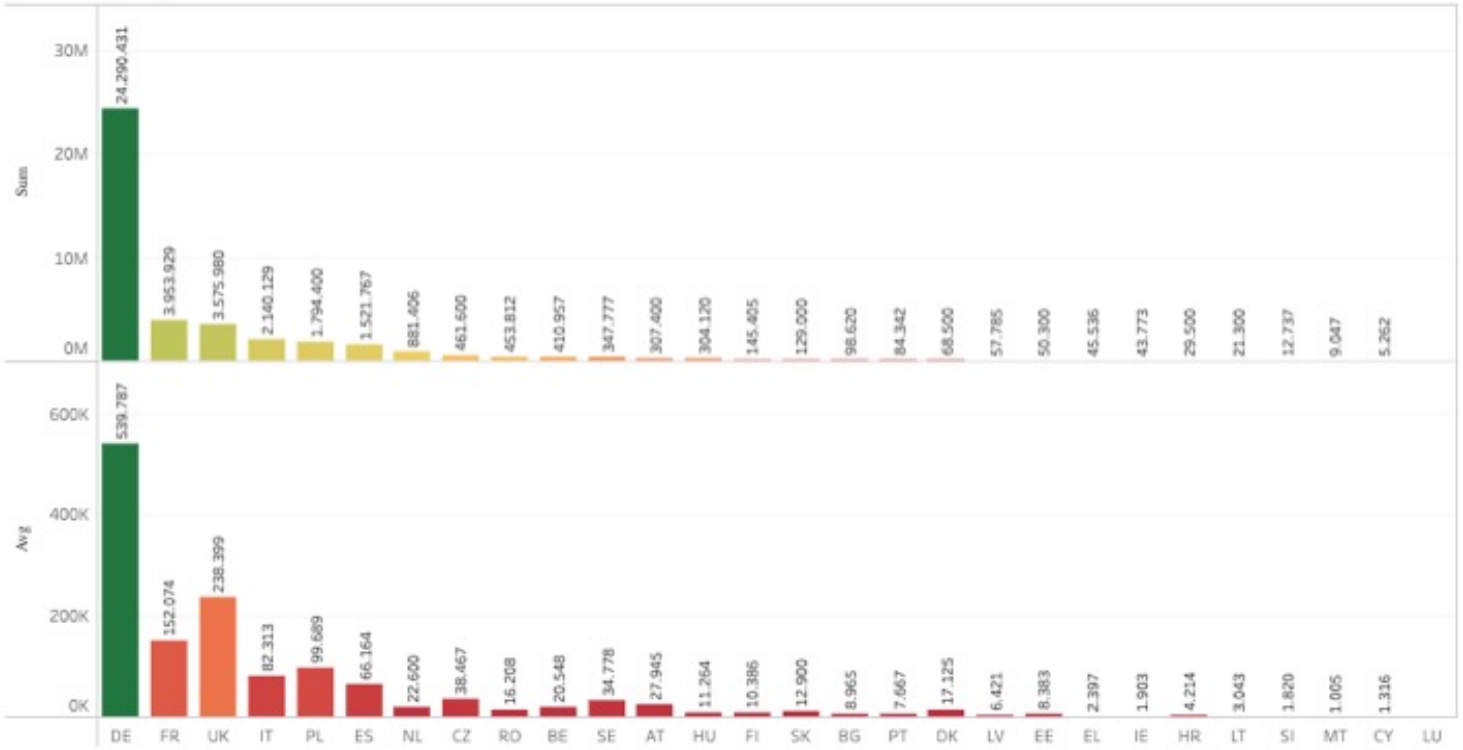
Sites by type of operator

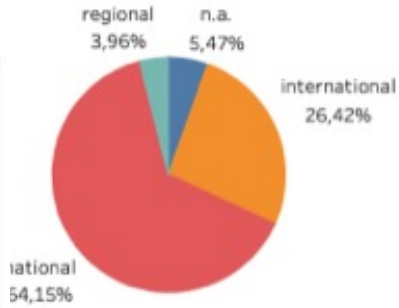


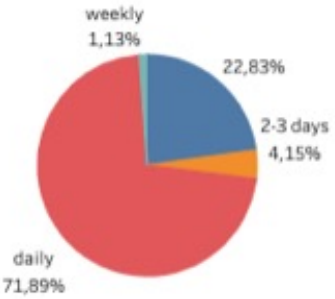


Vacancy volume by country

(estimated by ICE)







Расширение

Мы проанализировали результаты изучения ландшафта

- Сопоставили транснациональные источники
- Добавили дополнительные транснациональные источники

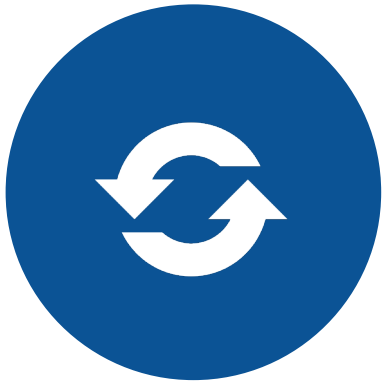
Чтобы

- составить список приоритетности для определения соглашений
- определить порядок значимости для реализации каналов приема данных

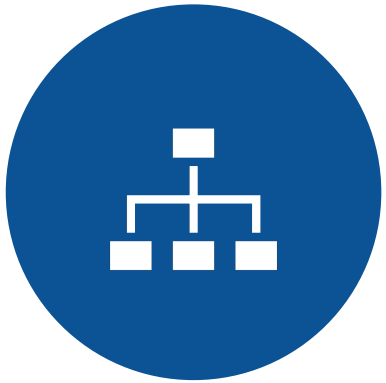
Значимость и ранжировка источников

Объем

Тип
веб-портала



Обновление
данных



Структурированные
данные

Фаза приема данных

Процесс получения и импорта данных с веб-порталов

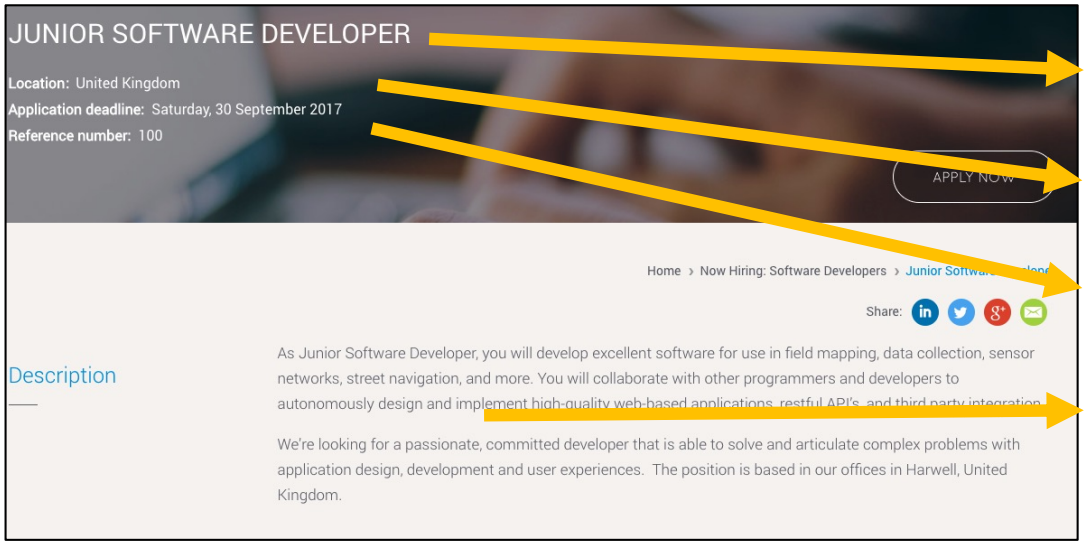
и их помещение в базу данных

Акцент на
объемах

Расширение и
и максимизация
охвата

Прямые соглашения
с наиболее
значимыми
источниками

Пример



Должность: Младший разработчик ПО

Регион: Великобритания

Время: Суббота, 30 сентября 2017 г.

Описание: В качестве младшего разработчика ПО вы будете разрабатывать прекрасное программное обеспечение для использования…

Темы

1. Цель и контекст
2. Задачи
    1. Участники
    2. Функциональная архитектура
    3. Методы приема данных
    4. **Конвейер обработки данных**
    5. Методы классификации
3. Выходные продукты
4. Результаты

Предварительная обработка данных — задачи и определения

Процесс очистки принимаемых данных и дедупликация онлайн-вакансий, чтобы во время аналитической фазы обрабатывались данные как можно более высокого качества

Определение языка

Снижение шума

Дедупликация онлайн-вакансий

Темы

1. Цель и контекст
2. Задачи
    1. Участники
    2. Функциональная архитектура
    3. Методы приема данных
    4. Конвейер обработки данных
    5. **Методы классификации**
3. Выходные продукты
4. Результаты

Классификация данных

- Цель:
  - Извлечь и структурировать информацию из данных для передачи на уровень представления
- Задачи:
  - Обработка огромного массива неоднородных данных на разных языках
- Подход:
  - Разработать адаптируемую схему с учетом языка, подстроенную к различным особенностям информации. Некоторые актуальные задачи:
    - Классификация по **профессии**: комбинированные методы, такие как машинное обучение, тематическое моделирование, обучение без учителя
    - Классификация по **профессиональным умениям**: другие различные комбинированные методы, такие как анализ текста с учетом сходства на основе корпуса или знаний
- Особенности:
  - Гарантировать извлечение объяснимой информации, регистрацию методов классификации и соответствующие особенности.

Классификация данных — пример

Младший разработчик ПО

В качестве младшего разработчика ПО, вы будете разрабатывать прекрасное программное обеспечение для использования в сфере сопоставления полей, сбора данных, сенсорных сетей, уличной навигации, и многих других. Вы будете сотрудничать с другими программистами и разработчиками в целях самостоятельного проектирования и внедрения высококачественных веб-приложений, API, соответствующих ограничениям REST, и обеспечения интеграции сторонних решений.

Мы ищем увлеченного, преданного делу разработчика, умеющего решать и формулировать сложные задачи в сфере проектирования приложений, разработки и взаимодействия с пользователем. Работа в нашем офисе в Харвелле, Великобритания.

Темы

1. Цель и контекст
2. Задачи
    1. Участники
    2. Функциональная архитектура
    3. Методы приема данных
    4. Конвейер обработки данных
    5. Методы классификации
3. **Выходные продукты**
4. Результаты

Модель DataLab

2 таблицы

Документ FT

1 строка для каждого элемента:

- Основной идентификационный номер → Ключ
- Онлайн-вакансия
- Источник
- Местоположение

Почему? Потому что для каждой онлайн-вакансии
мы можем обнаружить указание нескольких местоположений,

например «Разработчик ПО в Лондоне / Ливерпуле»

Анализ проф. умений FT

1 строка для каждого элемента:

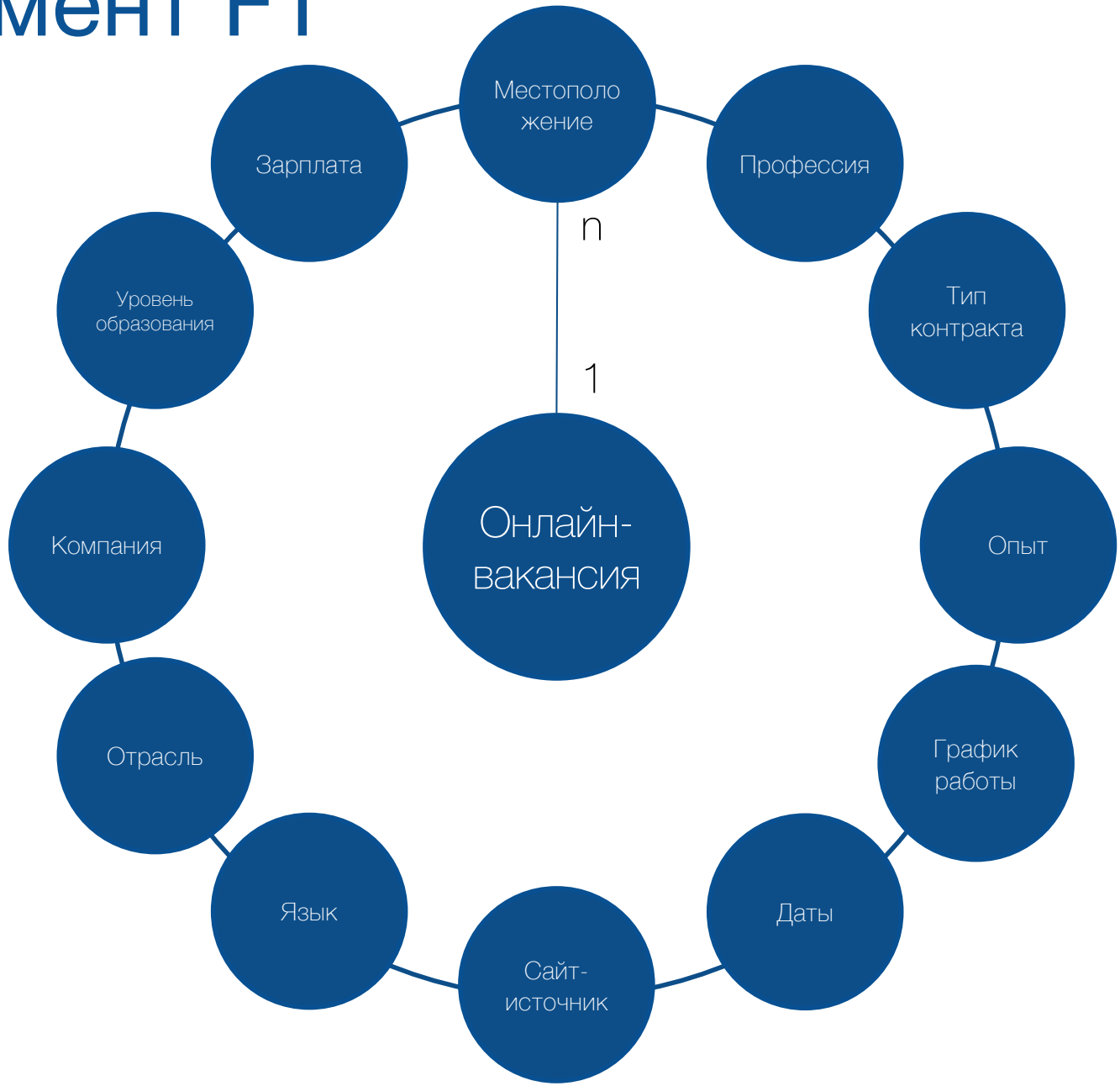
- Основной идентификационный номер → Ключ
- Онлайн-вакансия
- Источник
- Местоположение
- Профессиональное умение

Почему? Потому что для каждой онлайн-вакансии мы, очевидно, можем выявить
множество профессиональных умений, например «Разработчик ПО в Лондоне / Ливерпуле,
способный ориентироваться на клиентов, владеющий английским языком и обладающий стрессоустойчивостью».

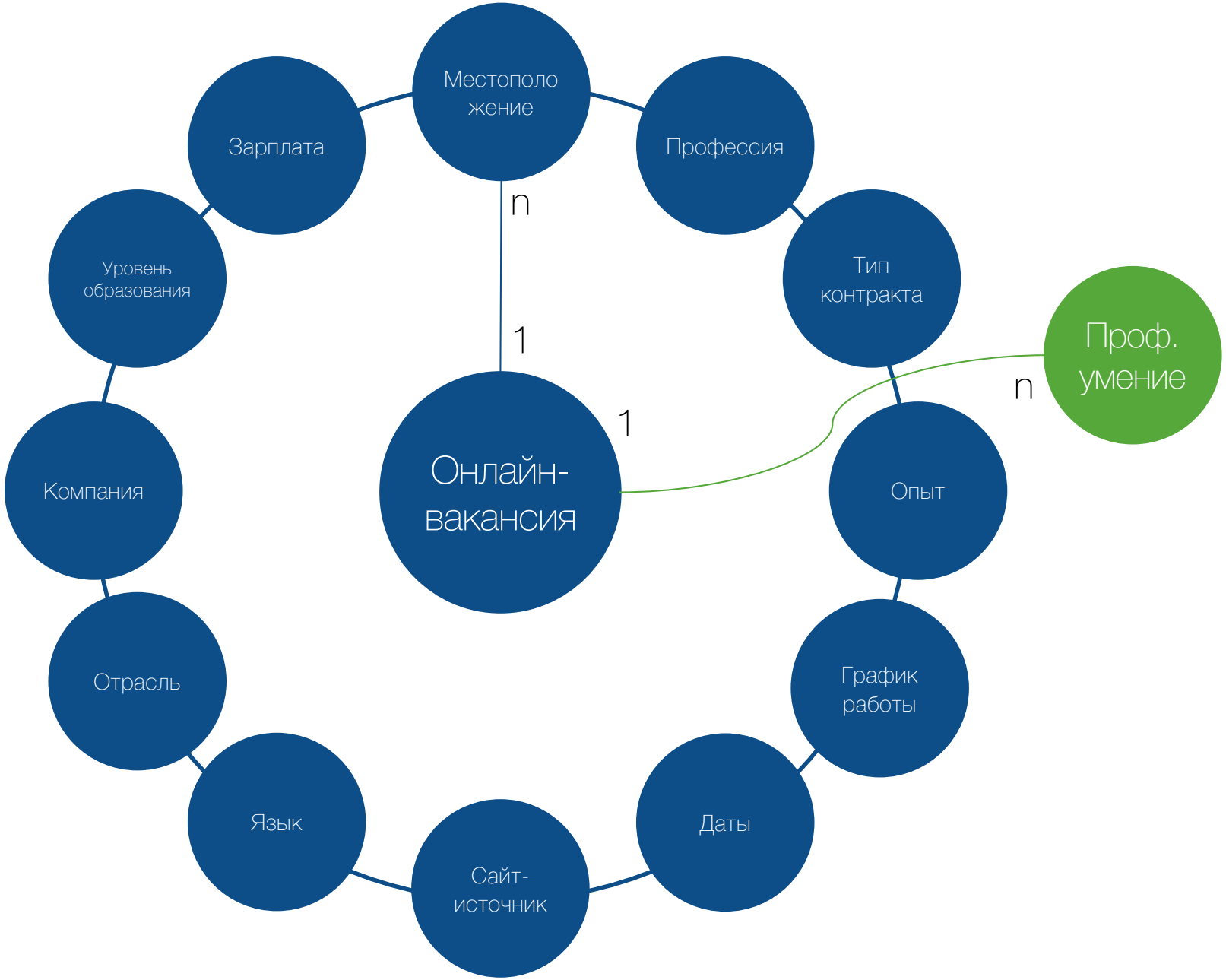
Модель DataLab

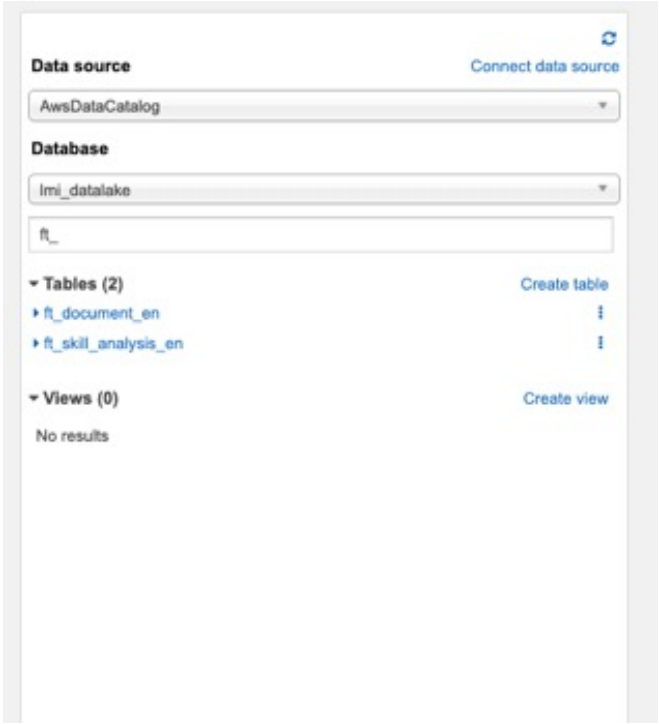
Таким образом, чтобы получить
определенное количество вакансий,
нужно каждый раз вычислять
уникальный основной
идентификационный номер

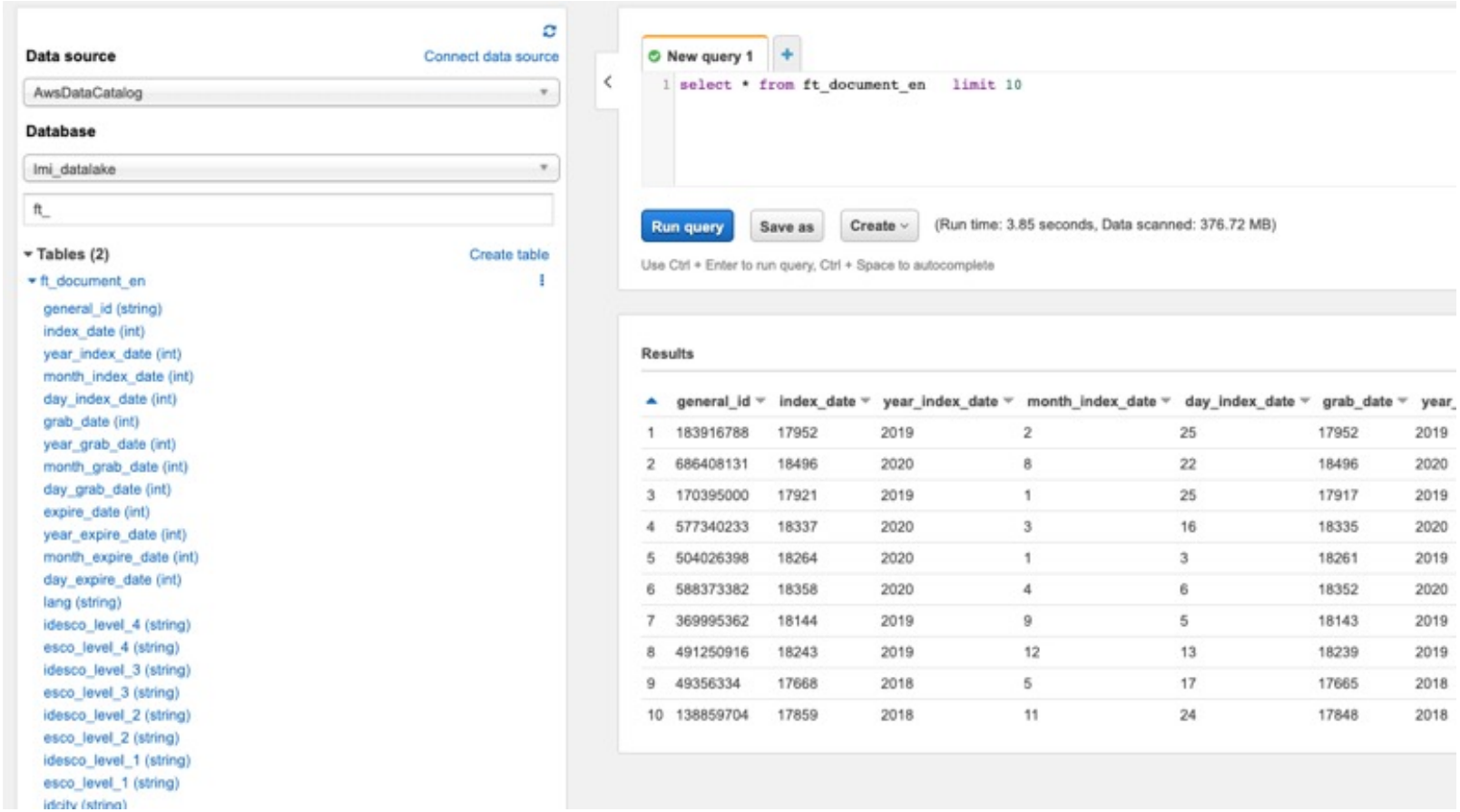
Документ FT

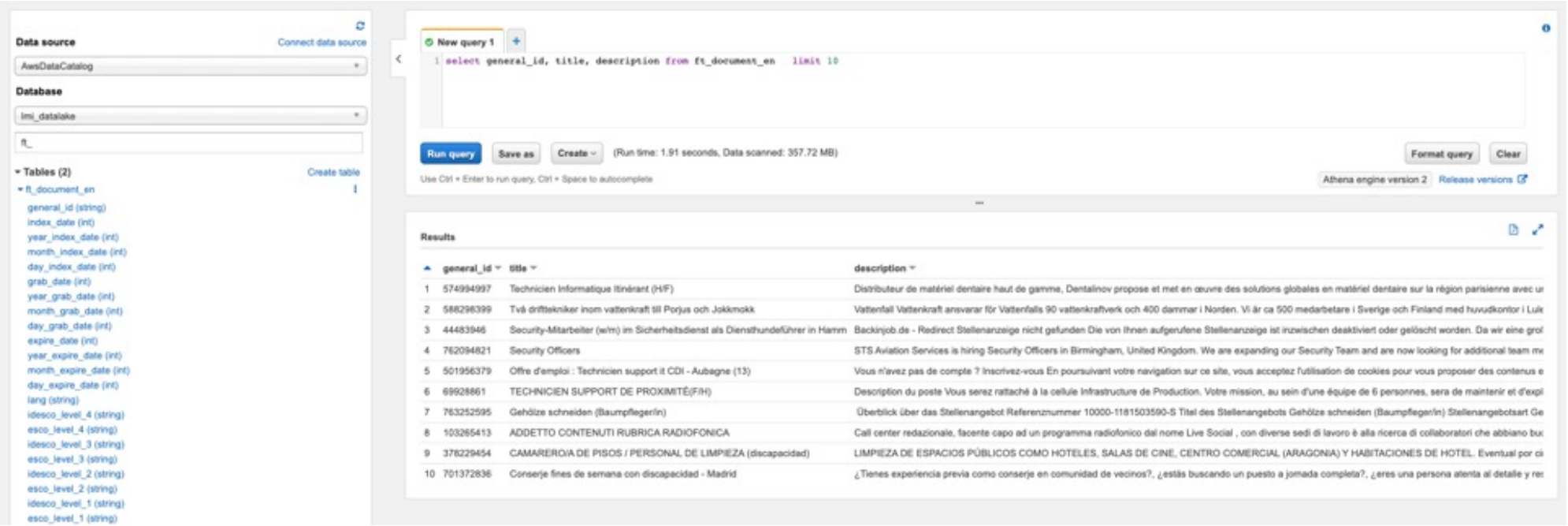

Местополо
жение
Профессия
Тип
контракта
Опыт
График
работы
Даты
Сайт-
источник
Язык
Отрасль
Компания
Уровень
образования
Зарплата
n
1
Онлайн-
вакансия

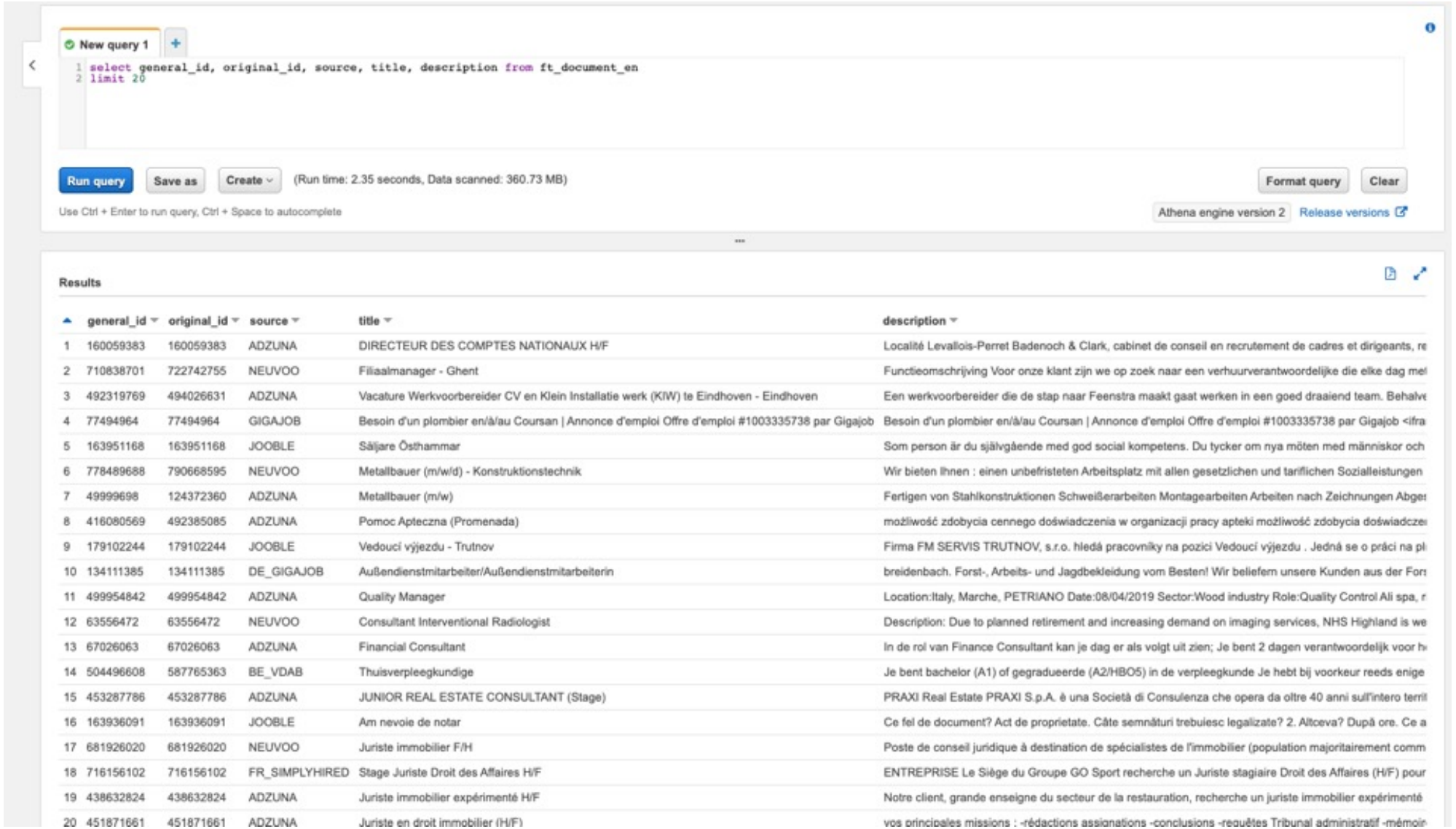
Анализ проф. умений FT




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)
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
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)
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 588298399 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 Luk
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 mel
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 Abge
8	416080569	492385085	ADZUNA	Pomoc Apteczna (Promenada)	możliwość zdobycia cennego doświadczenia w organizacji pracy apteki możliwość zdobycia doświadcze
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 For
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 terri
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

Темы

1. Цель и контекст
2. Задачи
    1. Участники
    2. Функциональная архитектура
    3. Методы приема данных
    4. Конвейер обработки данных
    5. Методы классификации
3. Выходные продукты
4. **Результаты**

Визуализация данных. Просто график?

- Цель:
  - Предоставить нужной заинтересованной стороне нужный инструмент
- Задачи:
  - Найти способ удовлетворить разные потребности с учетом такого значимого объема информации
- Подход:
  - Определить несколько подходов к визуализации и анализу данных:
    - **Инфографика**: привлекательная, статическая и понятная для широкого круга пользователей
    - **Общедоступные порталы для граждан**: просто, быстро и высокоинформативно
    - **Информационная панель**: повышенная информативность, доступна в Интернете, для лиц, отвечающих за принятие решений
    - **Лаборатория для самостоятельного анализа**: доступ к данным, самый высокий уровень информативности, требует особого набора предметных, технических и аналитических навыков

Онлайн-сеанс