

Job market

Cursus concerné : Data Engineer

Difficulté : 9/10

Description détaillée :

Ce projet a pour but de mettre en avant vos compétences de Data Engineer. Vous allez regrouper des informations sur les offres d'emplois et les compagnies qui les proposent. À la fin du projet, vous aurez une meilleure vision du marché de l'emploi : quels secteurs recrutent le plus, quelles compétences sont requises, quelles sont les villes les plus actives etc ...

Etape	Description	Objectif	Modules / Masterclass / Templates	Conditions de validation du projet
1	Récolte des données	<p>Les sources :</p> <ol style="list-style-type: none"> 1. API The Muse (lien) : vous devrez vous créer un compte ; vous trouverez des offres d'emplois ainsi que des informations sur les compagnies 2. API Adzuna (lien) : clé API requise ; vous trouverez des offres d'emplois, des données historiques sur les salaires 3. Scraper Welcome To The Jungle (démonstration ici), LinkedIn etc ... <p>Difficultés :</p> <ol style="list-style-type: none"> 1. Les attributs ne sont pas identiques entre les sources 2. Les langues sont différentes, vous pouvez choisir de ne garder que les données en anglais (avec spacy) 	<p>Utilisation de la librairie requests ou de l'outil Postman (pour tester)</p> <p>Techniques de web scraping : 133</p> <p>Transformations des données : 101</p>	Fichier en format tableau avec le nom de la source, la technique utilisée, un échantillon des données
2	Architecture des données	<p>Vous pouvez créer un data lake avec les données recueillies afin d'historiser vos extractions.</p> <p>Vous pouvez créer une base de données NoSQL et/ou SQL selon les données (entreprises, offres d'emplois, métiers etc...) et selon les requêtes que vous allez faire</p>	<p>142 - SQL</p> <p>Hbase</p> <p>Elasticsearch</p> <p>MongoDB</p>	<p>Diagramme UML avec les attributs des collections/tables/index</p> <p>Justification du choix de(s) SGBD choisi(s)</p>
3	Consommation de la donnée	<p>Analyse sur le nombre d'offres par entreprise</p> <p>Détecter le secteur d'activité qui recrute le plus</p> <p>Trouver l'emploi de ses rêves : lieu ; techno ; secteur ; niveau (senior etc...)</p>	<p>111</p> <p>114</p> <p>104</p> <p>116</p>	Des statistiques alimentées par le(s) base(s) de données

			Dash Elasticsearch	
4	Déploiement	Il faut créer une API pour requêter le(s) base(s) de données. Faire un conteneur Docker de chaque composant du projet (BDD, API) et faire un docker-compose fonctionnel.	FastAPI ou Flask Docker	Un fichier yaml du docker-compose une api
5	Automatisation des flux	ÉTAPE FACULTATIVE Il faut récupérer les données des sources selon un rythme bien défini pour l'envoyer aux différents consommateurs de la donnée.	Airflow	Fichier python du DAG
6	Soutenance	Démonstration de leur appli et explication du raisonnement effectué lors de leur projet.	X	Soutenance Rapport

Pour aller plus loin :

- vous pouvez récupérer des données sur les villes comme le coût de la vie, la qualité de vie etc ([lien](#)) ;
- vous pouvez extraire des données supplémentaires sur les compagnies sur ce site ([lien](#)) ;
- vous pouvez créer votre propre moteur de recherche d'emploi en combinant les qualifications requises, le cadre de vie, la qualité de l'entreprise, le type de contrat