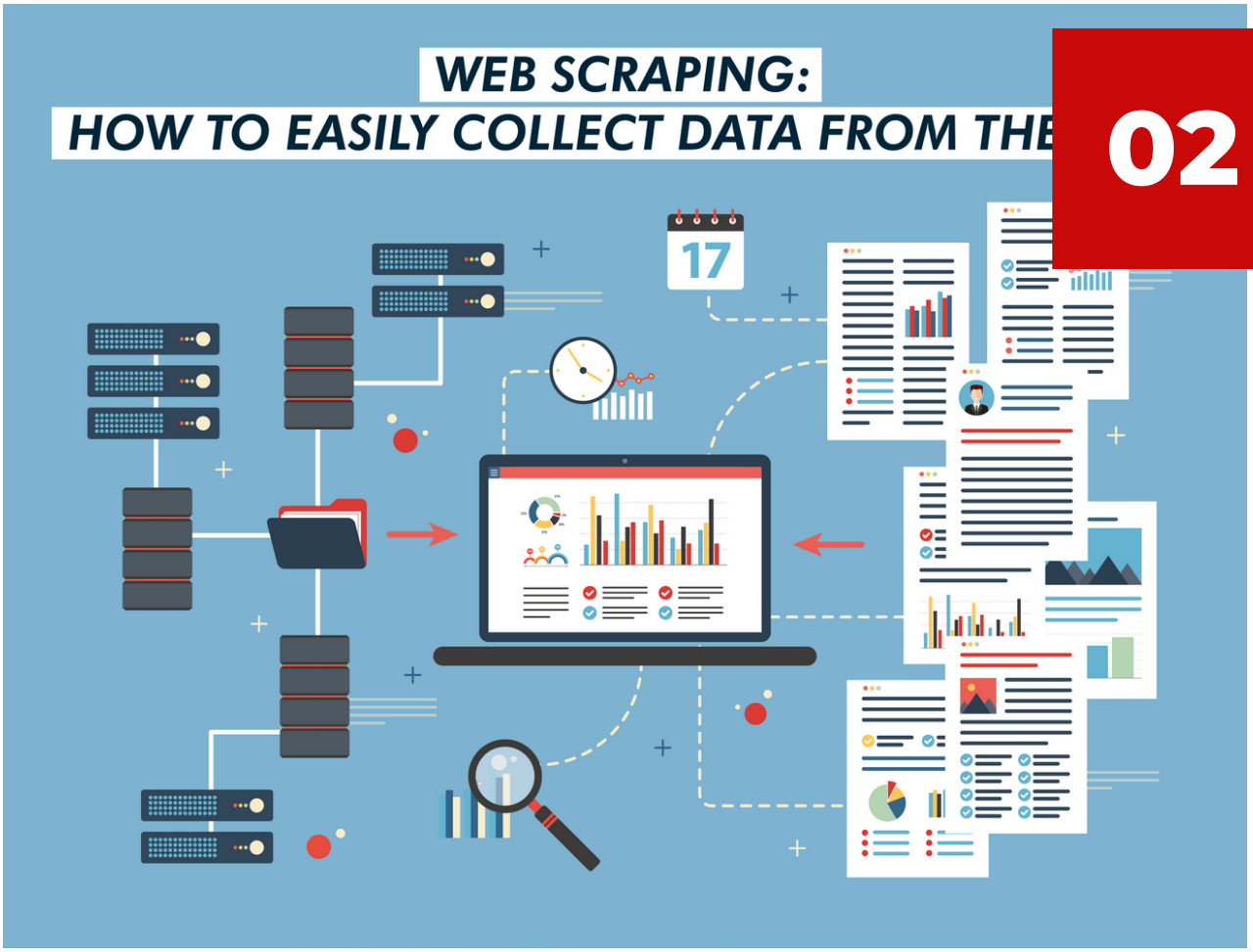
COLLECTE WEB

RAPPORT COLLECTE WEB

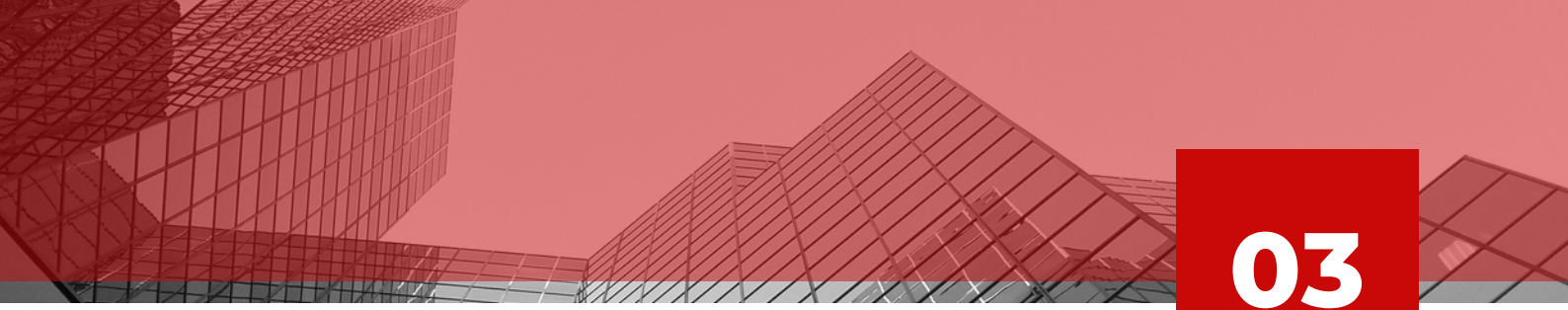
Thomas BELAID
Patrick CHEN
BUT SD - 2 FA VCOD



02

SOMMAIRE

- Introduction du Sujet
- Explication des Etapes
- Difficultés rencontrés



03

INTRODUCTION

Le projet vise à recueillir des données et des informations relatives à des entreprises. Nous disposions d'une base de données comprenant l'ensemble des numéros SIREN, ainsi qu'une sous-base constituée de 1000 de ces numéros pour effectuer des tests. Nous devions utiliser plusieurs techniques afin d'arriver à nos fins, notamment l'utilisation d'API adresse et chromedriver.

Afin de pouvoir avoir le plus d'information, nous avons du utiliser une autre base de donnée qui nous permet d'avoir accès au nom de l'entreprise

L'objectif est de développer un code en Python permettant, dans un premier temps, d'extraire les coordonnées des entreprises, puis d'obtenir certaines informations telles que le site web et le numéro de téléphone.

Les livrables attendus comprennent :
Le code Python, présenté via un notebook
Un rapport détaillé
Date de remise : 08/03/2024

STRATÉGIE DE TRAVAIL

04

Découpage des fichiers

Pour pouvoir importer nos données de manière efficace, nous avons dû opter pour une approche de découpage des fichiers, suivie de leur concaténation, en raison de leur volumétrie importante (plusieurs millions de lignes). Ainsi, nous avons procédé à la segmentation du fichier StockUniteLégale et du fichier StockEtablissement.

Traitement par partition

Pour optimiser le traitement de nos données, nous avons mis en place une stratégie consistant à diviser notre vaste base de données en partitions. Concrètement, cela se traduit par la création de sections comprenant un nombre défini de lignes en amont (nous avons choisi 100 lignes comme valeur optimale). L'objectif est d'effectuer tous les traitements nécessaires sur chaque partition avant de les agréger et de les intégrer à un nouveau fichier CSV. Cette approche nous permet d'améliorer l'efficacité de nos opérations de traitement, en rendant le processus plus gérable et en facilitant la manipulation des données à chaque étape du traitement.

Recherche des coordonnées

Pour localiser précisément les coordonnées des entreprises, notre approche a consisté à tirer parti de l'AIP adresse. À cet effet, nous avons mis en place une variable dénommée 'adresse', résultant de la concaténation astucieuse de plusieurs autres variables pertinentes.

STRATÉGIE DE TRAVAIL

05

Suite recherche des coordonnées

Cette démarche sophistiquée nous permet, par le biais d'un code HTML dédié, de récupérer et d'exploiter efficacement les informations d'adresse associées aux entreprises dans notre ensemble de données. Nous avons du sélectionner la partie essentielle dans le html fourni par l'API afin de pouvoir par la suite stocker la longitude et la latitude dans des variables dans le CSV final

Web scapring

Afin d'obtenir des informations essentielles telles que les numéros de téléphone ou les sites web, nous avons automatisé nos recherches via Google Maps en utilisant chromedriver. Pour garantir l'exhaustivité des données, nous avons croisé notre base de données StockEtablissement avec celle de StockUniteLegales pour obtenir les noms des entreprises correspondantes. Les résultats de nos recherches ont été stockés dans les colonnes téléphone et site web.

DIFFICULTÉS RENCONTRÉS

06

Concernant ce projet nous avons eu deux grandes difficultés concernant :

Taille des fichiers

Nous avons rencontré un problème lors de l'importation des données en raison de la taille importante du fichier. Pour surmonter cette difficulté, nous avons élaboré une stratégie de partitionnement afin de permettre l'importation par étapes. Cette stratégie nous a permis de diviser le fichier volumineux en plusieurs dataframes gérables. Ensuite, nous avons utilisé une boucle pour fusionner progressivement ces dataframes par petits morceaux.
Cette approche nous a permis de traiter efficacement les données en évitant les problèmes de mémoire liés à la manipulation de fichiers de grande taille. De plus, elle nous a permis d'optimiser les performances en réduisant le temps nécessaire pour l'importation et le traitement des données.

Web scraping

Après l'étape de fusion et de recherche des coordonnées via l'API adresse. La partie Web scrapping nous a posé quelques problèmes. Certaines informations précisées dans les fichiers "StockEtablissement_utf.csv" ainsi que "StockUniteLegale_utf8.csv" semblent approximatives, rendant les recherches googles parfois peu fiables.

Axe d'amélioration

Pour compenser le manque d'informations, nous aurions pu remédier à cela en effectuant des recherches variées, en sélectionnant différentes colonnes du fichier pour chaque recherche, et en croisant ces différentes sources d'information.