الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التعليم العالي والبحث العلمي

جـــــــامعة غـــــردايـــــــة
Université de Ghardaïa
كـــــليـــة العلـــــوم والتكـــنولوجيا
Technologie Faculté des Sciences et de

Mémoire présenté en vue de l'obtention du diplôme de
MASTER II RECHERCHE **en** Informatiques

Spécialité : Systèmes Intelligents pour l'Extraction de Connaissances

Département : **Mathématiques et informatique**

Thème :

Extracteur de données web sur la comparaison de prix

Réalisé par : Hanane Rezzag et **Souheila Bouaroura**

Encadré par : M. Abdelkader ouled Mahrez et Mme. Nacera Brahim

Président :	**M.Kerrache**
Examinateur :	**M. Bouhani**
Examinateur :	**M.Betka**

Soutenu le :23/11/2020

Dédicaces

Je dédie ce modeste travail

A mes très chers parents qui m'a soutenue durant mes études et ne m'a jamais privée de leurs amours, de leurs attentions et de ses encouragements,

A mes chères sœurs et mon cher frère

A mon binôme Souhila

A mes amies et mes collègues en Master 2 et à tous ceux qui m'ont aidé

Hanane

Dédicaces

Je dédie ce modeste travail

A mon très cher marie qui m'a inconditionnellement soutenu durant mon parcours universitaire

A mes très chers enfants et le prince Anis

A mes chers sœurs et frères

A mon binôme Hanane

A mes amies et Mes collègues en Master 2

Ceux qui m'ont aidé et à tous

Souhila

Remerciement

Avec l'aide de Dieu tout puissant, j'ai pu accomplir ce modeste travail.

Je remercie vivement mon encadreur de thèse

Monsieur Abdelkader Ouled Mahrez et madame Nacera Brahim d'avoir accepté d'encadrer ce travail et d'avoir surtout cru au sujet, je le remercie pour leur encouragement, leurs soutiens scientifiques accordés

Tout au long de ce travail. Qu'ils trouvent ici

L'expression de ma profonde gratitude.

Je remercie infiniment tous le cadre enseignant de la spécialité SIEC à l'université de

GHARDAIA, et spécialement ceux m'ayant enseigné cette année, les respectables messieurs : Dr.BELLAOUAR Slimane, Dr.KERRACHE Chaker Abdelaziz, M.Abdelkader BOUHANI et Prof.MOUSSAOUI Abdelouahab pour leur enseignements leur indulgence et leur dévouement.

Je tiens à remercier le cadre administratif du département MI et de la faculté des Sciences et de la Technologie.

Enfin je tiens à remercier toutes les personnes qui ont contribué de près ou de loin à l'accomplissement de ce travail.

Table de matière

ملخص 7

Abstract 8

Résumé 9

Liste des figures 10

Préambule 12

Problématique 12

I. Introduction 14

II. Internet et le web 14

1. Internet 14

2. Le Web 15

3. L'évolution du web 16

III. Le big data 18

1. Le Big Data, c'est quoi ? 18

2. L'analyse de données en masse 19

3. Les évolutions technologiques derrière le Big Data 19

4. Les données massives 21

5. L'avenir du Big Data 22

IV. Conclusion 22

I. Introduction 23

II. Travaux antérieurs 23

III. Recherche d'information 25

1. Introduction 25

2. Définition 25

3. Architecture d'un système de recherche d'information 26

IV. L'extraction de données 27

1. Web crawler 27

2. Web Scraping 30

3. Le web scrapping vs le web crawling 33

4. Différents techniques du web scraping 34

5. Logiciels et outils du web scraping 35

6. Les domaines d'application du Web Scraping 36
7. Défis du Web Scraping 40
V. Conclusion 42
I. Introduction 43
II. E-commerce 43
1. Qu'est-ce que le e-commerce ? 43
2. Différences entre e-commerce et e-business 44
2. Les différentes formes du commerce électronique 44
3. Avantages et inconvénients d'un e-commerce 46
III. Application du scraping dans l'extraction des prix 48
1. L'extraction des prix « Price scraping » 48
2. les principaux avantages des sites de comparaison de prix : 48
2.1 Avantages axés sur le vendeur: 48
3. La relation entre « Price scraping » et le « Dynamic pricing » 49
4. Les outils nécessaires au « Price scraping » 50
5. Comparaison des prix 52
IV. Présentation de l'application My-Price-check vu théorique 53
1. Concept 53
2. Défis et contraintes majeurs de réalisation 53
3. Architecture applicative 54
4. Processus métier de l'application My-Price-Check 55
V. Etude conceptuelle de l'application 58
1. Présentation d'UML 58
1.3. Digramme de cas d'utilisation 62
1.4. Diagramme de séquences 65
1.5. Diagramme de classe 68
VI. Conclusion 69
I. Introduction 71
II. Présentation de l'application My-Price-Check vu technique 72
1. Cheminement fonctionnel 72
2. Environnement de développement 72

3. Algorithme du fonctionnement de My-Price-Check 76
III. Démonstration et résultat de l'application 78
1. L'interface principale 78
2. Recherche 79
3. Résultats de la recherche 79
4. Comparaison et affichage du meilleur prix 80
5. Consultation de produits proposés 80
6. Responsivité du site 82
7. Interfaces administrateur 87
8. Résultat 89
IV. Conclusion 90

ملخص

الموضـوع الـذي نتناولـه هـذه الأطروحـة هـو اسـتخراج بيانـات الويـب أو تجريـف الويـب ، والعمـل المقتـرح مـن خـلال هـذه الأطروحـة هـو الاسـتجابة لمشـكلة محـددة للغايـة وهـي اسـتخراج البيانـات الرقميـة الموجـودة مـن خـلال اقتـراح نظـام تجريـف الويـب الـذي هـو تطبيـق ويب يسمىMy-Price-Check

فـي كثيـر مـن الأحيـان لا تقـدم مواقـع التجـارة الإلكترونيـة نفـس المنـتج ، لـذا لمقارنـة الأسـعار ، يجـب عليـك العثـور علـى نفـس المنـتج. يكمـن التحـدي فـي العثـور علـى نفـس المنتجـات أو علـى الأقـل أكثرهـا تشـابهًا ممكنًـا ، وبالتـالي فـإن التحـدي يكمـن فـي تنفيـذ محـرك بحـث يسـتمد معلومـات حـول المنـتج مـن تـاريخ الموقـع نفسـه مـن أجـل ضـمان مقارنـة متسـقة ومتجانسـة ، في حالة عـدم العثـور علـى منـتج فـي جميـع المواقـع فـي نفـس الوقـت ، سـيقدم النظـام مقارنـة مـع المنتجـات ذات الخصـائص الممكنـة الأكثـر تشـابهًا ويمـنح العميـل إمكانيـة استشـارة المنتجـات واحـدًا تلـو الآخـر فـي مؤسسـاتهم المنزليـة My-Price-Check .هـو تطبيـق ويـب لمقارنـة الأسـعار مـن عمالقـة ويـب التجـارة الإلكترونيـة مثـل Amazon و Ebay و Alibaba وغيرهـا. يتـيح My-price-check تجميـع البيانـات مـن مواقـع الويـب المختلفـة وعـرض خصـائص المنتجـات التـي يطلبهـا العميـل وذلـك مـن خـلال تقـديم أفضـل الأسـعار المتاحـة بـين المتاجر عبر الإنترنت المذكورة أعلاه ، وذلك بفضل برنامج دقيق و محرك بحث مُكيَّف

لضـمان المعالجـة الصـحيحة للبيانـات. مـن الناحيـة الفنيـة الأخـرى ، يتعلـق الأمـر باسـتخراج بيانات الويب "كشط الويب."

الكلمـات المفتاحيـة: اسـتخراج البيانـات ، مقارنـة الأسـعار ، أفضـل الأسـعار ، البحـث عـن المعلومات

، مترجم الأسعار ، متتبع ارتباطات الويب ، أفضل خوارزمية للبحث عن الأسعار.

Abstract

The theme addressed by this thesis is the extraction of web data or web scraping, the work through this same thesis proposed is to respond to a very specific problem which is to extract existing digital data by proposing a web scraping system which is a web application named My-Price-Check.

Very often e-commerce sites do not offer the same product, so to compare prices you have to find the same product.

The challenge is to find the same products or at least the most similar possible, therefore the challenge is to implement a search engine that draws information on the product from the history of the site itself in order to guarantee a consistent and homogeneous comparison, in the event that a product cannot be found in all the sites at the same time the system will offer a comparison with the products with the most similar possible characteristics and give the customer the possibility of consulting the products one by one in their home establishments.

My-Price-Check is a price comparison web application from e-commerce web giants such as Amazon, Ebay, Alibaba and others. My-price-check makes it possible to collect data from different websites and display the characteristics of the products requested by the customer and this by offering him the best prices available among the online shops listed above, this thanks to a meticulous program and an adapted search engine to guarantee the correct processing of the data. In other more technical terms, it is about doing "Web Scraping" extraction of web data.

Keywords; Best prices, Information search, Price compiler, Web crawler, Best price search algorithm.

Résumé

La thématique abordée par ce mémoire est l'extraction des données web ou le web scraping, le travail proposé à travers ce mémoire est de répondre à une problématique bien précise qui est d'extraire les données numériques existantes en proposant un système de web scraping qui est une application web nommé My-Price-Check.

Très souvent les sites e-commerce ne proposent pas le même produit, donc pour comparer des prix il faut trouver le même produit. Le challenge est de trouver les mêmes produits ou à la rigueur les plus similaires possibles, de ce fait le challenge est d'implémenter un moteur de recherche qui puise des informations sur le produit dans l'historique du site lui-même afin de garantir une comparaison cohérente et homogène, dans le cas où un produit est introuvable dans tous les sites en même temps le système va proposer une comparaison avec les produits avec les caractéristiques les plus similaires possibles et donner la possibilité au client de consulter les produit un par un dans leurs établissements d'origine.My-Price-Check est une application web de comparaison de prix à partir des géants du web du e-commerce tel que Amazon, Ebay, Alibaba et autres. My-price-check permet de collecter les données de différents sites web et afficher les caractéristiques des produits demandés par le client et ceci en lui proposant les meilleurs prix disponibles parmi les boutiques en ligne énoncées ci-dessus, ceci grâce à un programme minutieux et un moteur de recherche adapté afin de garantir le bon traitement des données. En d'autres termes plus techniques il s'agit de faire du « Web Scraping » extraction des données web.

Mots clés : Extraction de données, Prix, Comparaison de prix,
Meilleurs prix, Recherche d'information, Compilateur de prix, Robot d'indexation, Algorithme de recherche du meilleur prix.Best prices, Information search, Price compiler, Web crawler, Best price search algorithm.

Liste des figures

FIGURE 1.1 : Architecture en couches du Web sémantique

FIGURE 1.2 : REGLE DES 3V EN BIGDATA

FIGURE 1.3 : REGLE DES 5V EN BIGDATA VU PAR LES ECONOMISTES

FIGURE 2.1 : Architecture générale d'un système de recherche d'information

FIGURE 2.2 : Architecture générale du Crawler

FIGURE 2.3 : Schéma simple du fonctionnement du web scraping

FIGURE 2.4 : Architecture détaillée d'un web scrapper

FIGURE 2.5: Web scraping vs web crawling

FIGURE 3.1 : Processus décrivant le concept du e-commerce

FIGURE 3.2 : Architecture applicative de l'application Price-Check

FIGURE 3.3 : Processus de l'interaction Administrateur-Système

FIGURE 3.4 : Processus de l'interaction Client-Système

FIGURE 3.5 : Les Diagrammes D'uml 2

FIGURE 3.6 : LES DIAGRAMMES UML VU GENERALE

FIGURE 3.7 : Diagramme de cas d'utilisation interaction système

FIGURE 3.8: Diagramme de cas d'utilisation interaction utilisateur

FIGURE 3.9 : Diagramme de séquences Administrateur

FIGURE 3.10 : Diagramme de séquences Système-web_scraping (Recherche produits et calcule des meilleurs prix)

FIGURE 3.11 : Diagramme de séquences Administrateur

FIGURE 4.1 : Algorithme du fonctionnement de My-Price-Check

FIGURE 4.2 : Interface principale du site web (interface client)

FIGURE 4.3 : Exemple de recherche client 'Macbook Pro'

FIGURE 4.4 : élément de résultat de recherche

FIGURE 4.5 : élément de résultat de recherche

FIGURE 4.6 : mise en évidence du meilleur prix d'un produit recherché

FIGURE 4.7 : Redirection vers le site Amazon pour consulter le produit

FIGURE 4.8 : Redirection vers le site Craigslist pour consulter le produit (le meilleur prix)

FIGURE 4.9 : Rendu visuel de l'interface du menu principal de l'application sur mobile

FIGURE 4.10 : Rendu visuel de l'interface bas de page du menu principal de l'application sur mobile

FIGURE 4.11 : Exemple de recherche de produit 'Playstation 4' sur mobile

FIGURE 4.12 : Rendu visuel de résultat de recherche sur mobile

FIGURE 4.13 : Rendu visuel de résultat de recherche sur mobile

FIGURE 4.14 : interface d'authentification administrateur

FIGURE 4.15 : interface administrateur

FIGURE 4.16 : interface administrateur consultation d'historique de recherche

FIGURE 4.17 : interface administrateur vu mobile

FIGURE 4.18 : fichier .CSV du résultat de recherche

Introduction générale

Préambule

Les données sont des éléments essentiels de toute recherche, qu'elles soient académiques, marketing ou scientifiques, cependant ces données sont entrain de connaitre une croissance exponentielle cela crée une certaine difficulté pour gérer l'ensemble de ces données.

Là est apparu l'intelligence artificielle qui a pour but de concevoir des systèmes qui répondent aux besoins humains et résoudre des problèmes de la vie courante de l'être humain dans divers domaines. Et c'est aussi faciliter les tâches répétitives qui nous font souvent perdre du temps.

L'explosion quantitative des données numériques a obligé les chercheurs à trouver de nouvelles manières de voir et d'analyser le monde. Il s'agit de découvrir de nouveaux ordres de grandeur concernant la capture, la recherche, le partage, le stockage, l'analyse et la présentation des données. Ainsi est né le « Big Data ». Il s'agit d'un concept permettant de stocker un nombre indicible d'informations sur une base numérique.

Il est aussi question de la façon de la collecte de données et comment les extraire

Problématique

L'extraction des données c'est de collecter des données provenant d'une ou plusieurs sources de données, notamment les sites web qui contiennent d'innombrable quantité de données.

Cette action peut être difficile car les données peuvent être réparties sur plusieurs pages dans différentes sections et aussi la plupart des sites web ne permettent pas d'enregistrer une copie des données affichées sur leur site web sur votre stockage local alors l'utilisateur est obligé de consulter manuellement les données affichées par les sites web et les recopier dans un fichier local dans son ordinateur.

C'est un travail très fastidieux et prend beaucoup de temps.

La solution est la technique de 'web scraping

Notre travail consiste à développer une application qui permet d'extraire les prix de différents sites web puis les comparer et afficher le meilleur prix.

Introduction générale

Structure du document

Nous avons structuré ce travail comme suit :

- **Chapitre I** : dans ce chapitre nous allons présenter quelques notions sur l'Internet et le Web.
- **Chapitre II** : dans ce chapitre nous allons détailler la technologie du web scraping **Chapitre III** : dans ce chapitre nous allons aborder le monde du e-commerce et l'apllication du web scraping dans l'extraction des prix pour ensuite entamer notre application en commençant par l'étude conceptuelle
- **Chapitre IV** : dans ce dernier chapitre, nous allons concrétiser notre conception et partie théorique en procédant à la réalisation d'une application web pour l'extraction et la comparaison de prix baptisé My-Price-Check.

Chapitre I : I ’évolution du web et des données

I. Introduction

La fin du vingtième siècle est caractérisée par le développement rapide des technologies de l'information, l'informatique a connu un grand développement, notamment avec la vulgarisation de l'utilisation des réseaux sociaux. En particulier internet.

Ce chapitre aura pour objectif de présenter quelques notions sur l’Internet et le Web

II. Internet et le web

1. Internet

1.1. Historique

Dans les années 1960, Internet est issu du réseau ARPANET par l’**ARPA** (Advanced Research Project Agency) pour le département américain de la défense. En 1969, le réseau Arpanet a pu envoyer son premier message d'un appareil à un autre où le premier appareil est situé au l'Université de California Research Laboratory, tandis que le deuxième appareil est situé à l'Université de Stanford. **[1]**

À la fin de 1969, le réseau Arpanet comptait 4 universités Américaines. Nous pouvons résumer tout ce qui précède aux dates suivantes :

- En 1974, Le départ a été avec l'Américain Vinton Gray Cerf et son ami Robert Elliott Khan lorsqu'ils ont publié leurs recherches sur les protocoles réseau d'échange de paquets : le protocole **TCP / IP**
- En 1975, un premier réseau TCP/IP l'université de Stanford et celle du London College été établit puis en 1977, il a été élargi pour inclure un troisième réseau situé en Norvège.
- En 1980 - Développement de **FTP** (File Transfer Protocol) et **UDP** (User Datagram Protocol)
- En février 1981, David Mills a développé le NTP : Computer Clock Sync Protocol.
- En mars 1982, le département américain de la Défense a accepté d'exploiter le protocole TCP / IP sur ses réseaux militaires
- Le 1er janvier 1983, ARPANET a converti tous ses protocoles de communication en un groupe TCP / IP.
- En 1983, Paul Mockapetris a écrit la première implémentation de **DNS**

- En 1984, L’émergence du Cisco : Cisco est l'une des principales sociétés de l'industrie Internet qui offrant une gamme de produits répondant aux besoins de l'utilisateur.
- En 1989, Tim Berners-Lee a proposé de "base de données hypertexte puis a mis au point l’hypertexte pour relier des documents entre eux
- En 1990, dévelepment du premier serveur HTTP (HyperText Transfer Protocol) et le premier navigateur web et le World Wide Web en 1991 par Tim Berners-Lee
- En 1991, Paul Jardetzky et Quentin Stafford-Fraser a développé Webcam dans Laboratoire Informatique de l'Université de Cambridge
- En 1993, dévelepment du premier navigateur web, Mosaïc
- En 1994, Tim Berners-Lee fonde le w3c.
- En 1998, Tim Berners-Lee émet l'idée d'utiliser le web pour aussi lier des données naissait alors le web sémantique.
- En 1998, création de Google.
- En 2005, lancement de la plateforme vidéo YouTube
- En Juillet 2006, twitter fait son apparition sur Internet
- Septembre 2006, ouverture au grand public du réseau social Facebook, puis le réseau s'est développé en Internet que nous utilisons aujourd'hui.

1.2. Définition

L’Internet est un réseau informatique mondial composé d'un groupe de réseaux nationaux, régionaux et privés qui utilisent le même protocole de communication : TCP / IP (Transmission Control Protocol / Internet Protocol) Autrement dit, Internet est le réseau des réseaux où des millions de personnes du monde entier se connectent pour communiquer, échanger des informations, partager des données et des programmes.
[2]

2. Le Web

La définition d'Internet en tant que système d’information mondial permet de comprendre que le Web, ou pour être complet le World Wide Web (www) n'est en fin de compte qu’une application / utilisations d’Internet qui vise à favoriser l’échange des documents et de données tout comme l'e-mail (e-mail), la messagerie instantanée (IM), transfert de fichiers vers ou depuis un serveur (transfert FTP) ou encore systèmes de partage de fichiers peer-to-peer, etc.

Chapitre I : I ’évolution du web et des données

Le World Wide Web est naît quinze ans après L’internet, il est inventé en 1990 par Tim Berners-Lee, chercheur au CERN à Genève, en Suisse.

Le Web est le service qui permet de consulter des informations provenant d'Internet sous la forme de pages qui sont placées sur des sites et qui peuvent être recherchées à l'aide d'un navigateur Web

Grâce au web, Internet est devenue accessible à tous, plus développé et plus étendu. En effet, Le Web est très utilisé par Internet, ce qui a causé une confusion entre les termes. **[3]**

Pour conclure **Internet** est le réseau, l'infrastructure. Le **Web** est un service sur ce réseau. Internet ne serait jamais devenu ce qu'il est aujourd'hui sans l’évènement du World Wide Web et le web n'aurait jamais pu exister sans Internet.

3. L’évolution du web

3.1. Avant le WEB

Dans les années 1980-1990 : premiers ordinateurs lourds, lents et très coûteux utilisés par quelques privilégiés. Seuls moyens de communication sont mails en mode texte, transferts de fichiers, chat et quelques newsgroups. Interaction réduite et diffusion massive d'informations vers le plus grand nombre quasi inexistantes. Mais tout a bien changé depuis : par la démocratisation des micro-ordinateurs et augmentation de bande passante sur les réseaux. **[4]**

Le World Wide Web a été créé pour accéder aux données n’importe où, n'importe quand sous forme de langage hypertexte interconnecté.

Au cours du temps, la structure et les usages d'Internet et donc du Web ont évolué selon nos pratiques sociales et commerciales. Le web a été caractérisé par une évolution continue du contenu et de l'apparence des pages web. **[4]**

Cette évolution peut se résumer dans les générations suivantes :

3.2. Le Web 1.0

Web 1.0 ou ce que l'on appelle le web traditionnel (années 1990) représentant un réseau de communication de l'information.

Web 1.0 est un réseau statique, encodé en HTML. Ces pages ne sont pas interactives mais plutôt un espace d'informations dans lequel les éléments importants, appelés ressources, sont identifiés par des identifiants universels appelés URI (Uniform Resource Identifier).

3.3. Le web 2.0

Le web 2.0, appelée également le web social (Depuis l'an 2000). Ici, le web est dynamique et collaboratif, permettant aux internautes d’échanger d’informations ou contenus (textes, vidéos, images ...), participer dans la création du contenu via les wikis, les blogs ou les réseaux sociaux où ils communiquent, créent et partagent des liens (par exemple Facebook, LinkedIn, Snapchat, YouTube. etc.).

Aussi avec le web 2.0, il est possible de développer des applications très spécifiques grâce à des fournisseurs de services comme Flickr, Instagram, ou encore Google. **[4] [5]**

3.4. Le web 3.0

Le web sémantique (web des données), est un réseau qui tente de comprendre les données, vise à créer des relations logiques entre les données, à organiser les informations disponibles et à rendre les ressources sur le web compréhensibles et utilisables, selon le contexte et les besoins de chaque utilisateur, comme une base de données géante où les ressources sont décrites et la relation entre elles est également un réseau de données qui relie de plus en plus le monde réel est virtuel, si le Modèle RDF (Resource Description Framework) permet de décrire les ressources Web et partager leurs métadonnées, les ontologies sont utilisées pour créer, échanger, intégrer, fusionner, étendre les propos des objets du domaine concerné.

Tim Berners-Lee (l’inventeur du World Wide Web) a proposé une architecture en couches pour le web sémantique qui est souvent représenté à l'aide de diagrammes, avec de nombreuses variations depuis.

3.5. Le web 4.0

La quatrième génération du web est ultra-intelligente, est une interface informatique entre objets communicants et personnes Où les plateformes web peuvent personnaliser leurs interfaces selon les habitudes de chaque utilisateur. C’est-à-dire, si vous visitez amazon.com plus d'une fois, il vous reconnaîtra et vous fournira des conseils pertinents et personnalisés.

Le web 4.0 est une version caractérisée par :

- Interactions naturelles entre les utilisateurs et les machines (reconnaissance vocale par ex)
- Ubiquité : interaction n'importe où avec une architecture informatique distribuée Attentives aux utilisateurs et aux objets par des capteurs (caméras, radars...) **[4] [5]**

III. Le big data

L’explosion quantitative des données numériques a obligé les chercheurs à trouver de nouvelles manières de voir et d’analyser le monde. Il s’agit de découvrir de nouveaux ordres de grandeur concernant la capture, la recherche, le partage, le stockage, l’analyse et la présentation des données. Ainsi est né le « Big Data ». Il s’agit d’un concept permettant de stocker un nombre indicible d’informations sur une base numérique. Selon les archives de la bibliothèque numérique de l’Association for Computing Machinery (ou ACM) dans des articles scientifiques concernant les défis technologiques à relever pour visualiser les « grands ensembles de données », cette appellation est apparue en octobre 1997. **[6]**

1. Le Big Data, c’est quoi ?

Inventé par les géants du web, le Big Data se présente comme une solution dessinée pour permettre à tout le monde d’accéder en temps réel à des bases de données géantes. Il vise à proposer un choix aux solutions classiques de bases de données et d’analyse (plate-forme de Business Intelligence en serveur SQL…).

Littéralement, ces termes signifient méga données, grosses données ou encore données massives. Ils désignent un ensemble très volumineux de données qu’aucun outil classique de gestion de base de données ou de gestion de l’information ne peut vraiment travailler. En effet, nous procréons environ 2,5 trillions d’octets de données tous les jours. Ce sont les informations provenant de partout : messages que nous nous envoyons, vidéos que nous publions, informations climatiques, signaux GPS, enregistrements transactionnels d’achats en ligne et bien d’autres encore. Ces données sont baptisées Big Data ou volumes massifs de données. Les géants du Web, au premier rang desquels Yahoo (mais aussi Facebook et Google), ont été les tous premiers à déployer ce type de technologie.

Cependant, aucune définition précise ou universelle ne peut être donnée au Big Data. Etant un objet complexe polymorphe, sa définition varie selon les communautés qui s’y intéressent en tant qu’usager ou fournisseur de services. Une approche transdisciplinaire permet d’appréhender le comportement des différents acteurs : les concepteurs et fournisseurs d’outils (les informaticiens), les catégories d’utilisateurs (gestionnaires, responsables d’entreprises, décideurs politiques, chercheurs), les acteurs de la santé et les usagers.

Le big data ne dérive pas des règles de toutes les technologies, il est aussi un système technique dual. En effet, il apporte des bénéfices mais peut également générer des inconvénients. Ainsi, il sert aux spéculateurs sur les marchés financiers, de manière autonome avec, à la clé, la constitution des bulles hypothétiques.

L’arrivée du Big Data est maintenant présentée par de nombreux articles comme une nouvelle révolution industrielle semblable à la découverte de la vapeur (début du 19e

siècle), de l’électricité (fin du 19e siècle) et de l’informatique (fin du 20e siècle). D’autres, un peu plus mesurés, qualifient ce phénomène comme étant la dernière étape de la troisième révolution industrielle, laquelle est en fait celle de « l’information ». Dans tous les cas, le Big Data est considéré comme une source de bouleversement profond de la société. **[6]**

2. L’analyse de données en masse

Selon le Gartner, ce concept regroupe une famille d’outils qui répondent à une triple problématique dite règle des 3V. Il s’agit notamment d’un Volume de données considérable à traiter, une grande Variété d’informations (venant de diverses sources, non-structurées, organisées, Open…), et un certain niveau de Vélocité à atteindre, autrement dit de fréquence de création, collecte et partage de ces données. **[6]**

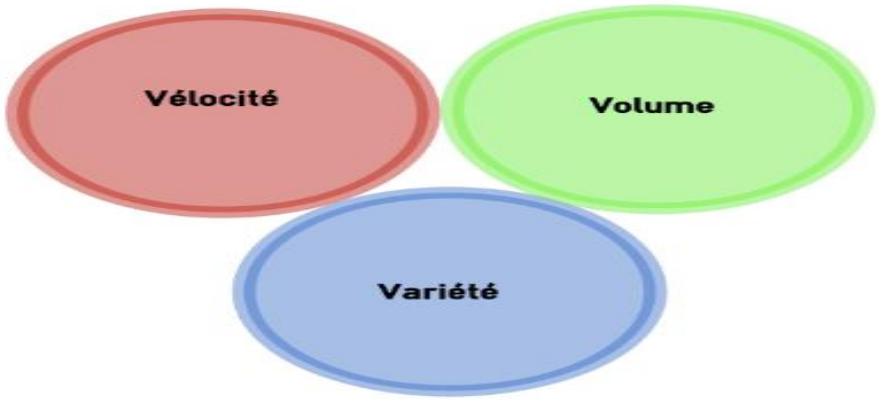


FIGURE 1.2 : REGLE DES 3V EN BIGDATA

3. Les évolutions technologiques derrière le Big Data

Les créations technologiques qui ont facilité la venue et la croissance du Big Data peuvent globalement être catégorisées en deux familles : d’une part, les technologies de stockage, portées particulièrement par le déploiement du Cloud Computing. D’autre part, l’arrivée de technologies de traitement ajustées, spécialement le développement de nouvelles bases de données adaptées aux données non-structurées (Hadoop) et la mise au point de modes de calcul à haute performance (MapReduce).

Il existe plusieurs solutions qui peuvent entrer en jeu pour optimiser les temps de traitement sur des bases de données géantes à savoir les bases de données NoSQL (comme

MongoDB, Cassandra ou Redis), les infrastructures du serveur pour la distribution des traitements sur les nœuds et le stockage des données en mémoire :

La première solution permet d’implémenter les systèmes de stockage considérés comme plus performants que le traditionnel SQL pour l’analyse de données en masse (orienté clé/valeur, document, colonne ou graphe).

La deuxième est aussi appelée le traitement massivement parallèle. Le Framework Hadoop en est un exemple. Celui-ci combine le système de fichiers distribué HDFS, la base NoSQL HBase et l’algorithme MapReduce.

Quant à la dernière solution, elle accélère le temps de traitement des requêtes.

Parmi les utilisateurs les plus enthousiastes du Big Data, on retrouve les gestionnaires et les économistes. Ces derniers définissent ce phénomène par la règle des 5V (Volume, Velocity, Variety, Veracity, Value). **[7]**

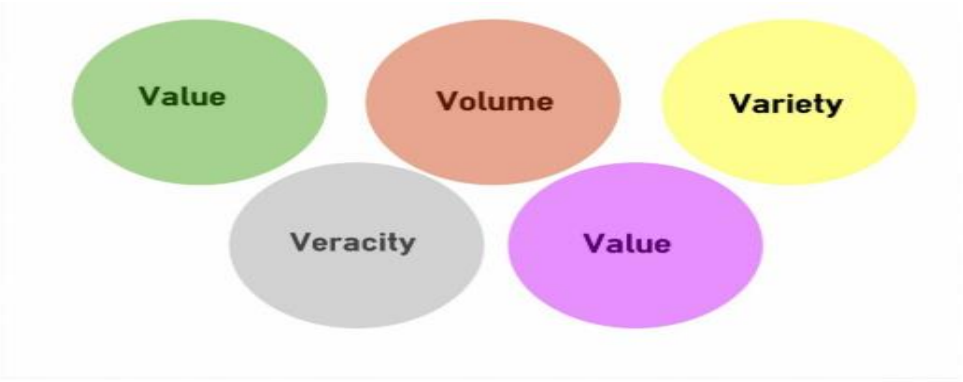


FIGURE 1.3 : REGLE DES 5V EN BIGDATA VU PAR LES ECONOMISTES

Le volume

Le volume correspond à la masse d’informations produite chaque seconde. Selon des études, pour avoir une idée de l’accroissement exponentiel de la masse de données, on considère que 90 % des données ont été engendrées durant les années où l’usage d’internet et des réseaux sociaux a connu une forte croissance. L’ensemble de toutes les données produites depuis le début des temps jusqu’à la fin de l’année 2008, conviendrait maintenant à la masse de celles qui sont générées chaque minute **[6]**

La vélocité

La vélocité équivaut à la rapidité de l’élaboration et du déploiement des nouvelles données. Par exemple, si on diffuse des messages sur les réseaux sociaux, ils peuvent devenir « viraux » et se répandre en un rien de temps. Il s’agit d’analyser les données au

décours de leur lignée (appelé parfois analyse en mémoire) sans qu’il soit indispensable que ces informations soient entreposées dans une base de données. **[6]**

La variété

Seulement 20% des données sont structurées puis stockées dans des tables de bases de données relationnelles similaire à celles utilisées en gestion comptabilisée. Les 80% qui restent sont non-structurées. Cela peut être des images, des vidéos, des textes, des voix, et bien d’autres encore… La technologie Big Data, permet de faire l’analyse, la comparaison, la reconnaissance, le classement des données de différents types comme des conversations ou messages sur les réseaux sociaux, des photos sur différents sites etc. Ce sont les différents éléments qui constituent la variété offerte par le Big Data. **[6]**

La véracité

La véracité concerne la fiabilité et la crédibilité des informations collectées. Comme le Big Data permet de collecter un nombre indéfini et plusieurs formes de données, il est difficile de justifier l’authenticité des contenus, si l’on considère les post Twitter avec les abréviations, le langage familier, les hashtags, les coquilles etc. Toutefois, les génies de l’informatique sont en train de développer de nouvelles techniques qui devront permettre de faciliter la gestion de ce type de données notamment par le W3C. **[6]**

La valeur

La notion de valeur correspond au profit qu’on puisse tirer de l’usage du Big Data. Ce sont généralement les entreprises qui commencent à obtenir des avantages incroyables de leurs Big Data. **[6]**

4. Les données massives

Un marché mondial en plein épanouissement

D’ici quelques années, le marché du big data va se mesurer en centaines de milliards de dollars. C’est un nouvel eldorado pour le business.

D’après le calcul effectué par le cabinet Vanson Bourne, dans le monde, l’ensemble des dépenses consacrées au Big data, dans les budgets IT des grandes entreprises, devrait représenter un quart du budget total IT. Le Cap Gemini a aussi commandité une étude Le résultat a montré que plus de 61% des entreprises sont conscientes de l’utilité du Big Data en tant que “moteur de croissance à part entière”. De ce fait, on lui accorde beaucoup plus d’importance que leurs produits et services existants. Cette même étude a encore indiqué que 43% d’entre elles se sont déjà réorganisées ou se restructurent présentement pour exploiter le potentiel du Big Data. **[6]**

5. L’avenir du Big Data

Etant une tendance lourde, le Big Data n’est pas une mode. Dans le domaine de l’usage, il satisfait une nécessité de travailler la donnée plus profondément, pour créer de la valeur, conjointement à des aptitudes technologiques qui n’existaient pas dans le passé. Cependant, compte tenu de l’évolution des technologies qui ne semble pas vouloir s’estomper, on ne peut pas alors parler d’une norme véritable ou de standards dans le domaine du Big data.

Beaucoup d’applications du Big Data n’en sont qu’à leurs préludes et on peut s’attendre à voir apparaître des utilisations auxquelles on ne s’attend pas encore aujourd’hui. En quelque sorte, le Big Data est un tournant pour les organisations au moins aussi important qu’internet en son temps. Chaque entreprise doit donc s’y mettre dès maintenant. Dans le cas contraire, il y a un risque qu’elle se rendent comptent d’ici quelques années qu’elles se sont faites dépasser par la concurrence. Les gouvernements et les organismes publics se penchent également sur la question à travers l’open data. **[6]**

IV. Conclusion

Dans ce chapitre, nous avons abordé les termes Internet et le Web, expliqué la différence Entre eux ainsi que les étapes de développement du Web, nous avons par ailleurs défini le big data et énumérer les différents challenges que fait face le monde informatique afin de subvenir à des besoins futurs, et notamment la monté en charge de la quantité de données.

Le chapitre suivant sera consacré à l’initiation d’une technique d’extraction de données à savoir : le web scraping.

Chapitre II : Initiation au web scraping

I. Introduction

Après la naissance du World Wide Web en 1989, le nombre de sites web, c'est-à-dire le nombre de pages web a eu une croissance exponentielle qui dépasse aujourd'hui 5,61[1] milliards de pages, donc le volume de données est très important, ces données sont variés (texte, images, vidéo ...) et renouvelable, ce qui fait que l'accès à ces données représente un challenge, qui nécessite des techniques diffèrent des méthodes traditionnelles comme le copier-coller, le capture d'écran ou encore accéder au bases de données à travers les requêtes. Ces techniques ne permettent pas d'extraire les informations d'une façon optimale. Là intervient une technique intéressante, efficace, très rapide et prometteuse On parle d'une technologie baptisée le « **web scraping** », dans ce chapitre nous allons détailler cette technologie.

II. Travaux antérieurs

L'extraction des données web ou le web scraping été abordé par de nombreux chercheurs ce processus n'est pas difficile, mais il existe plusieurs problèmes qui doivent être résolus :

Tout d'abord, les mises à jour fréquentes des structures de données, des sources de données et de la distribution Web rendent le coût de développement, de maintenance et de surveillance d'un web scraper plus élevé que le coût de construction d'un web scraper. Ce problème a déjà été étudié afin de le résoudre. Guojun et ses collègues (2017) étudient un moyen d'améliorer le crawler pour l'extraction de données de pages web dynamiques. **[8]**

Chaulagain et ses collègues. (2017, ont également essayée de résoudre le problème discuté dans l'article de Guojun et al. (2017). En construisant un web scraper intelligent pour obtenir des données à partir de sources dont la structure des documents est variante. Le XPath a été utilisé pour extraire les données pertinentes **[9]**

1 : https://www.worldwidewebsize.com/

Chapitre II : Initiation au web scraping

Selon Rajapriya (2014), « crawler » sont un élément essentiel des moteurs de recherche et leur mise en œuvre est essentielle afin de collecter toutes les informations nécessaires car les données sur le web sont représentées par différentes méthodes et structures et pour pouvoir explorer la plupart des données, il est nécessaire de comprendre et d'étudier tous ces types de structures et la possibilité d'extraire des informations. Il a également mentionné certains problèmes rencontrés lors de l'exécution crawlers tels que la recherche d'adresses en double url. **[10]**

Les premiers moteurs d'achat sont apparus au milieu des années 1990. Ces moteurs d'achat étaient à l'origine appelés "robots de prix" ou "moteurs de comparaison de prix" en ce sens qu'ils permettaient aux consommateurs de trouver le meilleur prix pour un livre ou un CD. **[11]**

BargainFinder est le premier agent d'achat d'intelligence artificielle, connu sous le nom de « robot de prix ». Le but de BargainFinder était de trouver le meilleur prix pour un produit spécifique ; trouver spécifiquement le meilleur prix sur les CD de magasins comme CDNow.com. Il développe par Bruce Krulwich en 1995 alors qu'il travaillait pour Arthur Andersen. Il existe des agents commerciaux tels que mySimon.com, Django ou EvenBetter.com qui permet aux consommateurs de rechercher sur le Web un produit complètement spécifique, puis de planifier Sites où le produit peut être acheté et à quel prix **[12]**

Un exemple qui illustre la comparaison des prix est l'étude faite par Vemula Satyanarayana, Rahul Kumar Behera et Gaurav Kumar de l'Université « KLEF » Département d'Electronique et de Génie Informatique pour Comprendre le comportement des achats en ligne pour les clients en inde et donner un aperçu du développement de l'Inde dans le domaine des affaires en ligne. L'examen a permis de tirer des conclusions à partir d'informations réelles, qui fourniront des données précieuses aux détaillants en ligne pour améliorer la méthodologie d'achat en ligne en Inde. Ce qui permet aux clients d'acheter un produit à un prix inférieur et de vérifier le prix de ce produit. Ils voient qu'il existe de nombreux facteurs qui poussent les gens à magasiner sur le Web tels que : Commodité, Information, Rentabilité et efficacité, Produits et services disponibles. Ils ont proposé un site web, qui permet de vendre un produit à un client et les clients peuvent également vendre leur propre produit en fournissant une page de revente aux clients. Pour afficher leur produit usagé avec son détail et si un client voit le poste et veut acheter peut contacter la personne qui a posté **[13]**

Chapitre II : Initiation au web scraping

III. Recherche d'information

1. Introduction

Historiquement l'attention a été portée d'abord sur les documents textuels. Les pages Web, les e-mails, les articles scientifiques, les livres et les articles de presse en sont que quelques exemples. Tous ces documents ont une certaine structure telle que le titre, la date, l'auteur etc.

Cependant la différence triviale entre un document et un enregistrement de bases de données est que la plupart de l'information du document est sous forme de texte qui est non structuré. Pour illustrer la différence en prend l'exemple de deux données le numéro de compte et le solde de celui-ci, les deux ont une structure bien définie (un entier à six chiffres pour le numéro de compte et un nombre réel avec deux chiffres après la virgule) dès lors il est facile de comparer deux valeurs ou faire des opérations arithmétiques. Prenons maintenant l'exemple d'articles de journaux, ceux-là ont bel et bien une certaine structure (Titre, Date etc.) Mais le contenu principal est l'article lui-même. S'il on venait à stocker ces articles dans une base de données un large champ sera attribué au texte de l'article mais ce texte n'a pas de structure ou décomposition. Afin de répondre aux besoins en informations des utilisateurs on traite leurs requêtes qui sont souvent simpliste (quelques mots), parfois vagues et imprécises. Pour ce faire on aura besoins d'algorithmes capables de comprendre et décider si l'information souhaité par l'utilisateur est contenue dans ce champ ou dans le texte de l'article, or il est plus difficile de définir le sens d'une phrase ou d'un paragraphe que de manipuler des numéros de comptes et des soldes La compréhension et la modélisation de la manière avec laquelle les gens comparent le texte, et la conception d'algorithmes qui font cette tâche est au cœur de la recherche d'information.

De plus en plus, les applications de recherche d'information impliquent des documents multimédias tels que les images la vidéo ou l'audio. Ces supports ont un contenu qui, comme le texte, est difficile à décrire et à comparer. Dans la recherche de documents non textuels on se référait aux descriptions textuelles de leur contenu plutôt que sur le contenu lui-même, mais actuellement des progrès sont en cours en ce qui concerne les techniques de comparaison directe d'images, par exemple.

2. Définition

Un système de recherche information est un système qui permet de retrouver une information pertinente par rapport à une requête dans une grande collection de documents

3. Architecture d'un système de recherche d'information

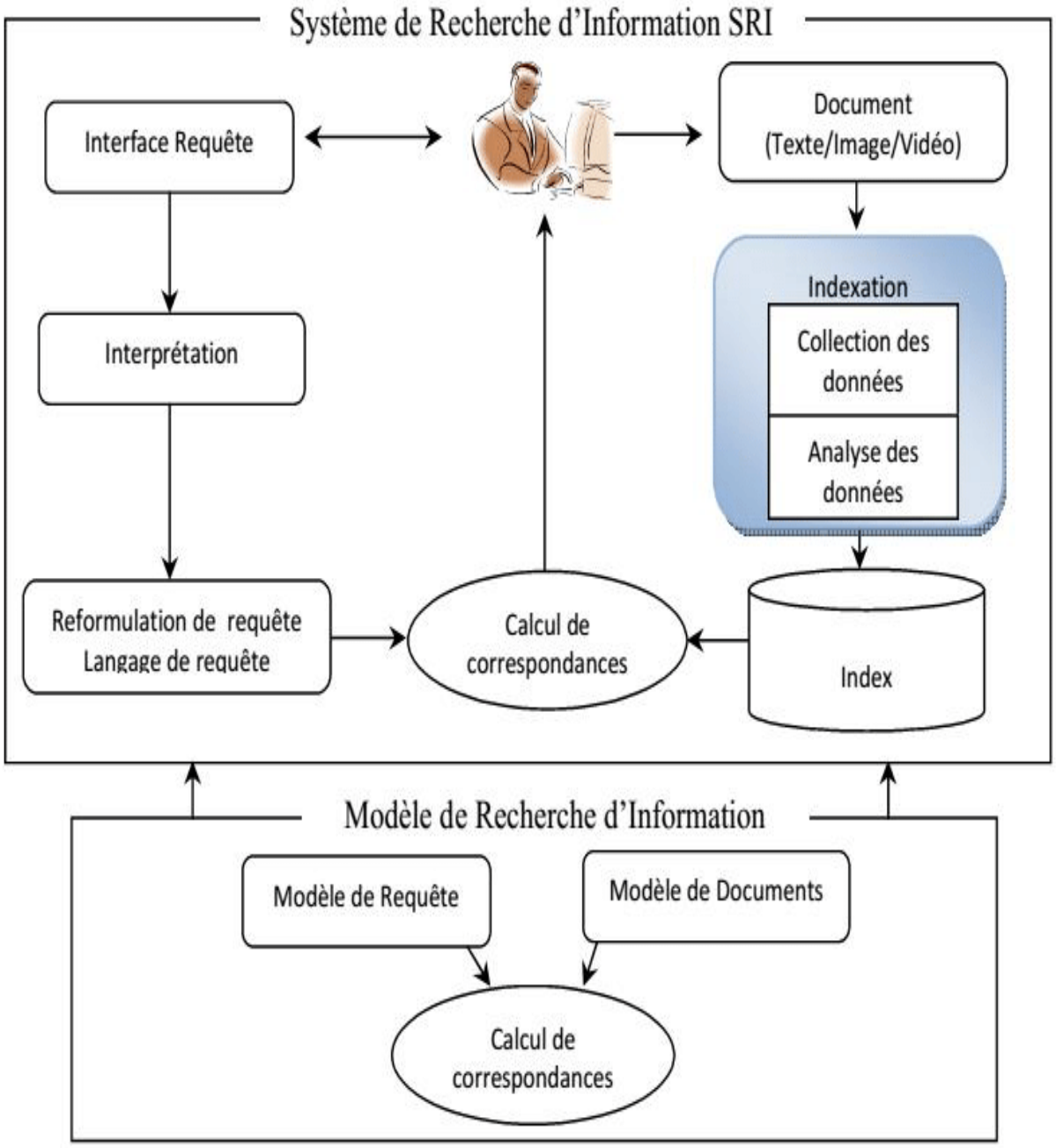


FIGURE 2.1 : Architecture générale d'un système de recherche d'information **[14]**

Chapitre II : Initiation au web scraping

IV. L'extraction de données

1. Web crawler

1.1. Définition

Crawler, robot d'indexation ou spider bot, désignent dans le monde de l'informatique un robot d'indexation. Concrètement, il s'agit d'un système qui explore le Web afin d'analyser le contenu des pages web visités et les stocker de manière organisée dans un index.

Le crawler parcourt donc, en permanence, de façon autonome et automatique, les différents sites et pages Internet à la recherche de contenus nouveaux ou d'éventuelles mises à jour de contenus déjà explorés par le passé. **[15]**

1.2. Fonctionnement

Le fonctionnement basique d'un robot de crawl rappelle celui de logiciel malveillant, comme les malwares, mais contrairement à ces derniers, leur passage n'est pas nuisible pour le site. Derrière cette activité se cache une autre mission : celle d'indexer les pages Web en fonction de la qualité des contenus (évaluée selon des critères paramétrés en amont) et, ainsi, aider les moteurs de recherche à classer les pages Internet dans l'affichage des résultats. **[15]**

Le crawler commence à partir d'une URL et identifie les liens hypertexte (HTML ou XML) sur la page Web qui ont été récupérés à partir de l'adresse et visite fréquemment des liens hypertexte pour récupérer une nouvelle page Web. Le système ne prend pas en compte les relations entre les pages Web.

En raison des quantités de données que le crawler doit traiter, les crawlers distribués ont été développés afin d'améliorer les performances des crawlers normaux en utilisant une architecture distribuée.

1.3. Architecture du Crawler

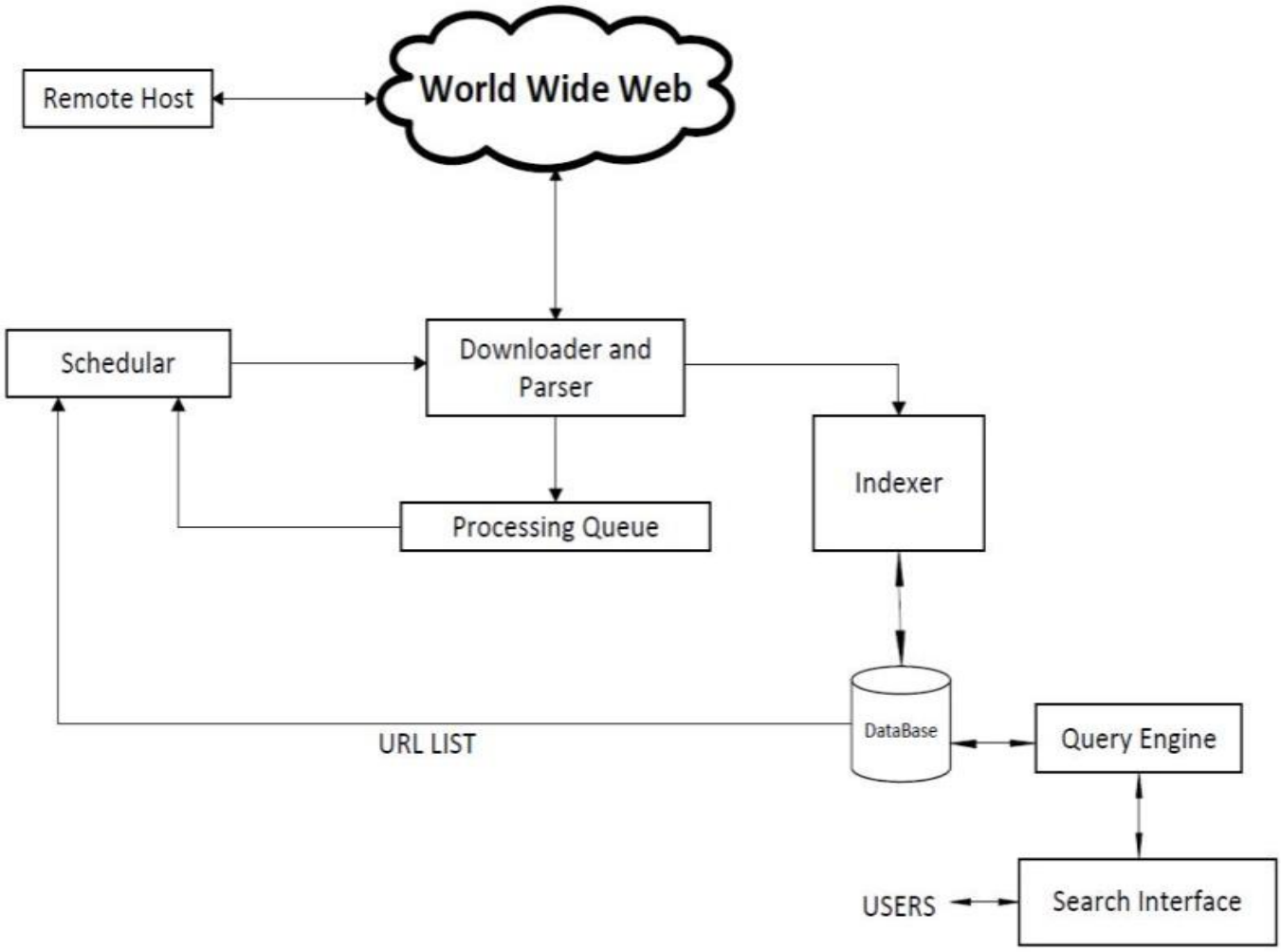


FIGURE 2.2 : Architecture générale du Crawler **[16]**

Un robot d'exploration **(Crawler)** de par son architecture exécute un certain nombre d'étapes prédéfinies l'une après l'autre. En règle générale, le crawler commence par cibler les différentes **URL** d'un site Web (sur internet **WWW**) une par une depuis un serveur distant **(Remote host),** les url's seront après téléchargées et analysées pour être exécutées par les modules **(Downloader/ Parser)** les résultats obtenus sont mis en attente dans une file d'attente de traitement **(Processing queue)** les url's seront mis en attente avant d'être exécutées, cette file d'attente est gérée par l'unité de programmation **(Scheduler)** chaque url transmis et exécuté par le système de web crawling enregistre les résultats dans un index par le module **(Indexer)** puis stockés dans une base de données **(BDD)**,

La base de données transfert périodiquement la liste des url's

Les données recueillis seront accessibles via le moteur de requête **(Query engine)** à la demande du client pour qu'à la fin les résultats de recherche seront affichés pour le client via l'interface de recherche **(Search interface)**

L'apparence de cet index évoqué ci-dessus dépend de l'algorithme spécifique, par exemple, l'algorithme de Google a spécifié l'ordre dans lequel les résultats apparaissent pour une requête de recherche spécifique.

Les crawlers peuvent être programmés pour parcourir le Web avec des objectifs déterminés. Ils sont actifs en permanence et visitent les pages selon les instructions qui leur sont données.

L'un des crawlers les plus connus est celui utilisé par Google pour son moteur de recherche, Googlebot. Avant lui, le moteur de recherche AltaVista utilisait le **crawler** Scooter pour effectuer cette même mission. Les crawlers de recherche suivent plusieurs chemins pour parvenir jusqu'aux documents à explorer. Soit ils partent des résultats déjà existants dans les moteurs de recherche, soit ils suivent une liste, soit ils obéissent à des soumissions ponctuelles, soit ils suivent les liens qu'ils rencontrent au fil de leur exploration.

Chaque moteur a ses propres règles. Les référenceurs s'intéressent beaucoup au fonctionnement des moteurs de recherche parce qu'il leur donne des pistes pour faire indexer les pages web importantes. **[17].**

Dans le cas des sites web statiques traditionnels, le client rend simplement le document récupéré sur le serveur, qui est également utilisé par le système de crawling pour extraire les hyperliens. Dans les pages web dynamiques modernes utilisant les méthodes **AJAX**, la page comprend également un code qui peut être exécuté par le client pour envoyer des requêtes asynchrones au serveur ou pour changer l'état de la page visualisée par le client. Ce code est généralement géré par un moteur JavaScript capable d'exécuter du JavaScript. Les informations récupérées par les appels asynchrones ou le code JavaScript exécuté par le navigateur sont utilisés pour manipuler l'état du site web. L'état de la page est représenté par un Document Object Model (**DOM**) qui est un arbre ordonné qui est encodé avec du **HTML**. Grâce à ce mécanisme, les pages web dynamiques font rentrer plus d'informations dans la page que ce qui est récupéré lors de la requête initiale, ce qui fait que le système d'exploration ne manque pas d'informations potentiellement vitales ou d'hyperliens.

Le problème résolu par le crawler web dynamique est donc d'atteindre tous les états potentiels de l'application web afin d'effectuer les tâches de crawling traditionnelles **[17].**

Chapitre II : Initiation au web scraping

2. Web Scraping

2.1. Définition

Le terme «scraping» désigne l'extraction des données, et le web scraping peut être défini comme un processus ou une technologie utilisé pour collecter ou extraire une grande quantité de données à partir d'un ou plusieurs sites web en un temps court et il est largement reconnu comme une technique efficace et puissante pour la collecte de données hétérogènes et volumineuses via un programme, un autre site web ou un script, ce processus effectuée peut être manuel par un utilisateur, à l'aide d'un bot ou d'un crawler (robot d'indexation).

Une interface **API** fait office d'intermédiaire entre le script d'exécution du programme de scraping et le site ciblé à l'extraction de données, ceci afin de traiter manuellement les données recueillis à partir des sites web.

Le programme Web Scraper est connecté au site via le protocole http, il récupère la page et extrait les données de cette page et les échange entre plusieurs pages de sites Web en fonction des exigences d'extraction de données, Lorsque les données sont extraites, elles seront ensuite exportées dans différents formats tels que CSV et JSON selon les besoins. **[18] [19] [20]**

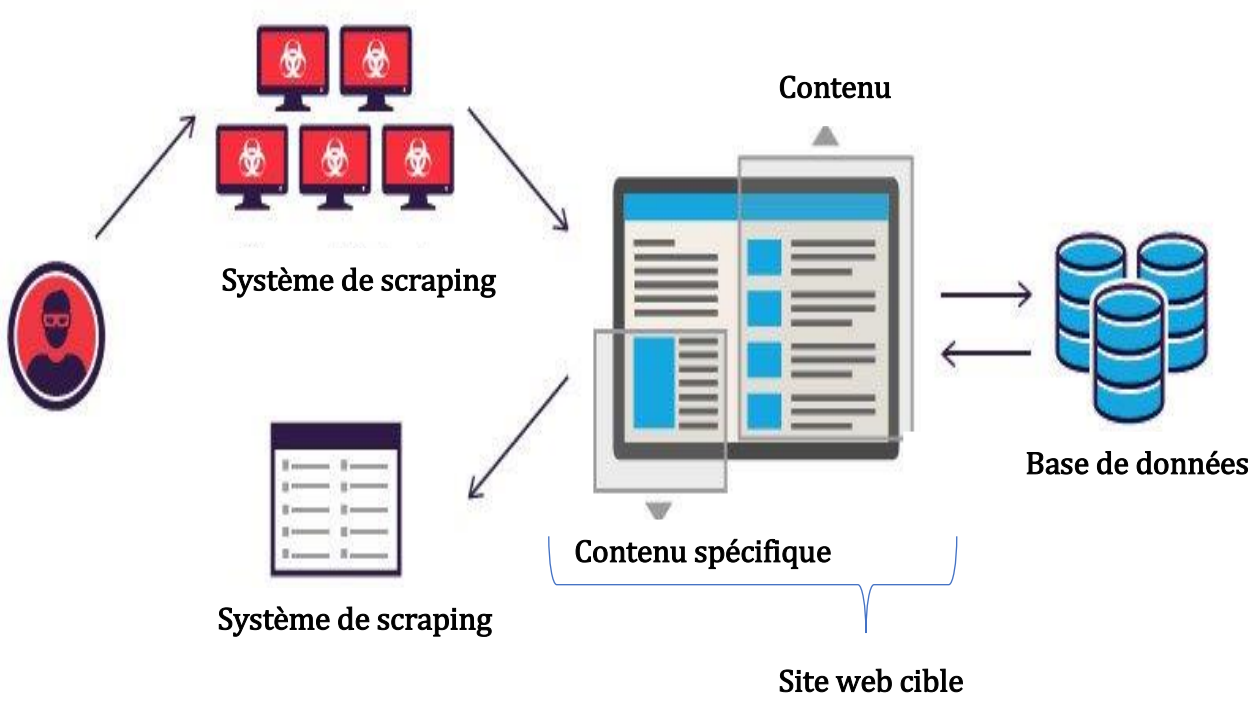


FIGURE 2.3 : Schéma simple du fonctionnement du web scraping **[21]**

2.2. Fonctionnement du web scraping

2.2.1. Demande-Réponse

La première étape simple de tout programme de scraping Web (également appelé « racleur ») consiste à demander au site Web cible le contenu d'une URL spécifique. En retour, le grattoir obtient les informations demandées au format HTML. Le HTML est le type de fichier utilisé pour afficher toutes les informations textuelles sur une page Web. **[22]**

2.2.2. Analyser et extraire

En termes simples, HTML est un langage de balisage avec une structure simple. En ce qui concerne l'analyse, elle s'applique généralement à n'importe quel langage informatique. C'est le processus qui consiste à prendre le code sous forme de texte et à produire une structure en mémoire que l'ordinateur peut comprendre et utiliser.

L'analyse HTML consiste essentiellement à intégrer du code HTML et à extraire des informations pertinentes telles que le titre de la page, les paragraphes de la page, les en-têtes de la page, les liens, le texte en gras, etc. **[22]**

2.2.3. Télécharger les données

La dernière partie est l'endroit où vous téléchargez et enregistrez les données dans un CSV, JSON ou dans une base de données afin qu'elles puissent être récupérées et utilisées manuellement ou utilisées dans tout autre programme.

Avec cela, vous pouvez extraire des données spécifiques du Web et les stocker généralement dans une base de données locale centrale ou une feuille de calcul pour une récupération ou une analyse ultérieure. **[22]**

2.3. Architecture du web scraping

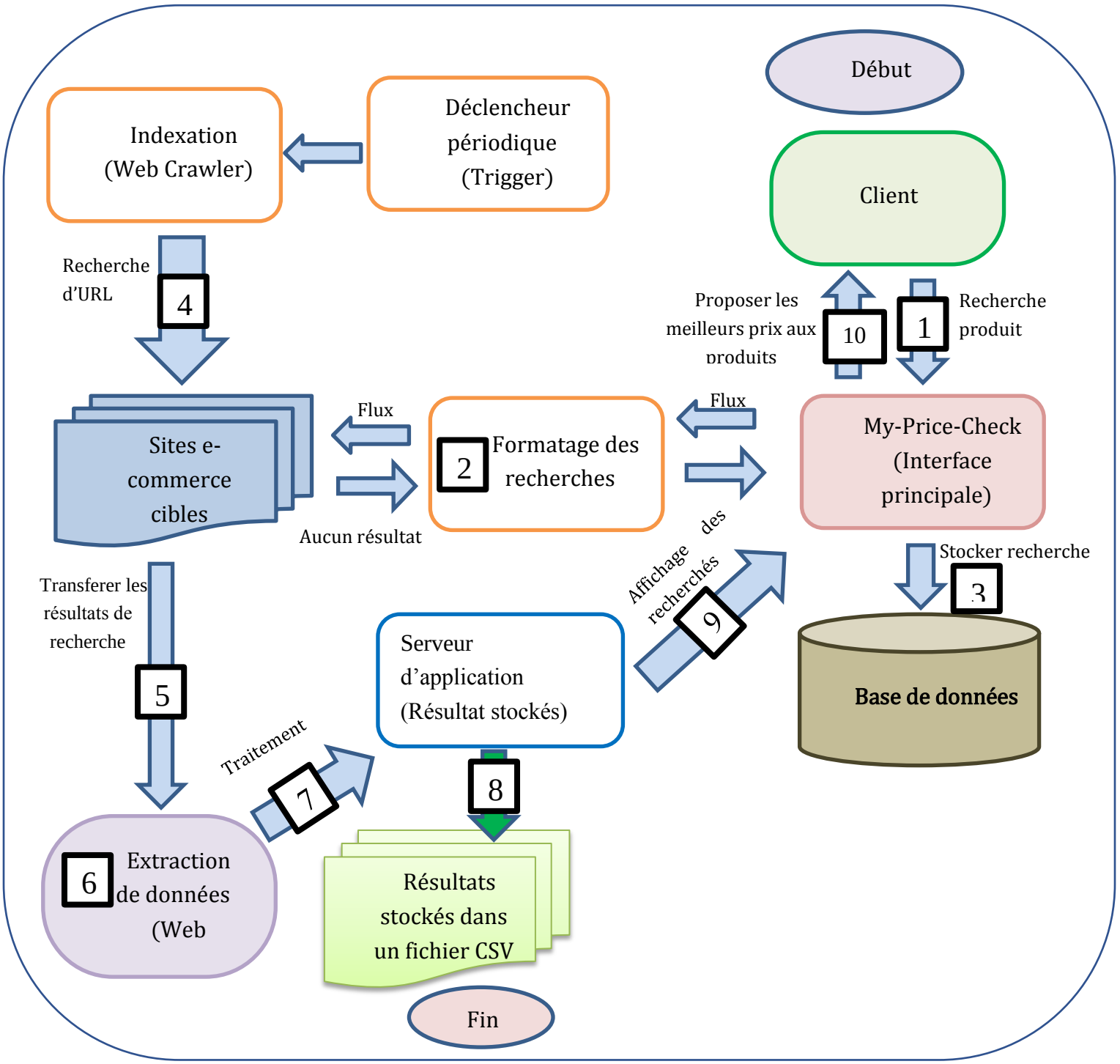


FIGURE 2.4 : Architecture détaillée d'un web scraper

3. Le web scrapping vs le web crawling

Le Web Crawling est effectué d'une manière différente avec des résultats différents par rapport au Web scraping.

La figure ci-dessous illustre ces deux mécanismes : Le processus du « Web crawling » consiste à parcourir le Web et à l'indexer en suivant des hyperliens. Par contre le "Web scraping" est le processus d'extraction d'informations structurées d'une page web, et pour faire du "web scraping", il faut faire un certain degré de "web crawling" pour se déplacer sur les sites web **[23].**

Pour recapitaliser le web crawling et le web scraping sont deux techniques d'extraction de données mais de manière différente et évidemment donc avec des résultats différents, cependant dans la majorité des cas le web crawling est inclus dans le processus du scraping. **[23]**

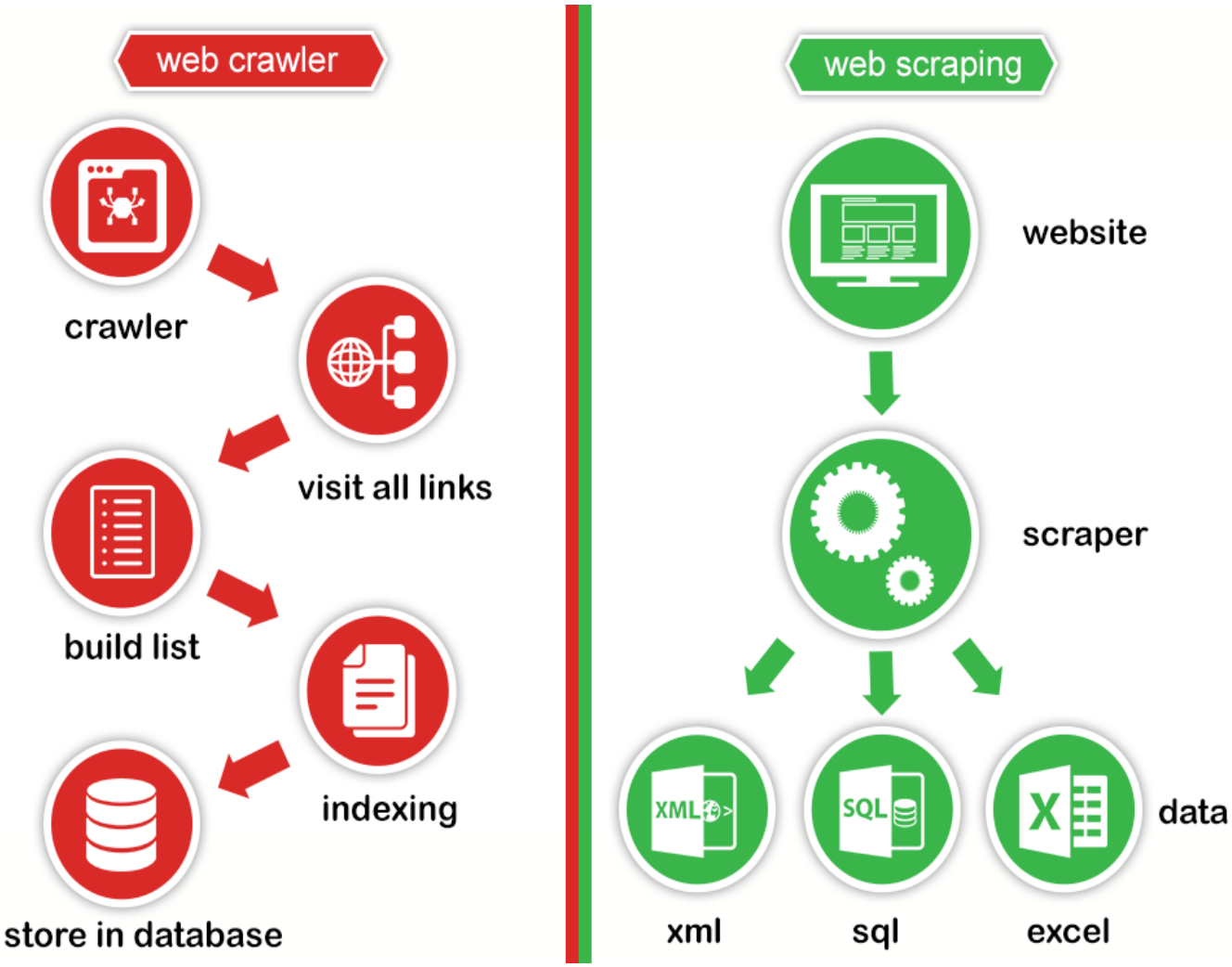


FIGURE 2.5: Web scraping vs web crawling **[23]**

Chapitre II : Initiation au web scraping

4. Différents techniques du web scraping

4.1. Accès au site Web

Dans ce cas le web scraper crée une connexion au site Web cible via HTTP.Il s'agit d'un protocole HTTP (HyperText Transfer Protocol) qui coordonne les paramètres de réponse aux demandes entre le client, et transfert ces demandes généralement un navigateur Web et un serveur Web. Dans le protocole HTTP, la « méthode » la plus courante est la méthode GET, souvent utilisée pour les demandes de ressources tandis que la méthode POST est utilisée pour envoyer des formulaires et télécharger des fichiers, à travers ces méthodes le protocole crée un flux de données qui par la suite est récupéré et stocker afin d'utiliser les données recueillis. Donc c'est du web scraping via site web. **[19]**

4.2. Analyse HTML et extraction du contenu

Lorsque le document HTML est récupéré, le Web scraper peut extraire le contenu qui l'intéresse. La signification de l'analyse HTML est de charger le HTML, extraire et traiter les informations pertinentes telles que le titre de l'en-tête, les actifs de la page, les sections principales et plus tard, enregistrer le fichier traité.

Entre autres cibler les parties du document converties en HTML pour en extraire des informations pertinentes. Pour cela il existe des bibliothèques d'analyse HTML fonctionnant sur la structure Document Object Model des pages Web et des langages basés sur des sélecteurs, tels que XPath et La syntaxe du sélecteur CSS (Cascading Style Sheets).

4.3. Analyse du modèle d'objet de document (DOM)

DOM (**D**ocument **O**bject **M**odel) est une interface permettant de structurer des documents HTML et XML, cette interface offre aux programmeurs un accès aux différents composants d'un projet Web, pour faciliter l'insertion, la suppression ou l'édition des contenus, d'attributs et de styles.

L'Analyse DOM peut être intégrée à un navigateur web tel que Google Chrome ou Mozilla, cette analyse en fonction des programmes est capable de récupérer des parties des pages tel que le contenu dynamique généré par les scripts côté client.

4.4. Plateformes d'agrégation verticales

Sont des plates-formes de récolte de données verticales spécifiques. Ces plateformes Créent et surveillent un grand groupe de « robots » pour des secteurs spécifiques opérant sur n'importe quel site, la plateforme est robuste si elle récupère un nombre considérable de champs et peut évoluer jusqu'à des centaines ou des milliers de sites **[24]**

Chapitre II : Initiation au web scraping

5. Logiciels et outils du web scraping

Dans l'extraction des données ou " Web scraping "La plupart des sites web n'offrent pas d'API (interface de programmation d'applications) ou de moyens pour les utilisateurs d'enregistrer des données, si bien qu'au début l'utilisateur était obligé de copier et de coller des données dans Excel ou un autre programme et de les stocker manuellement

Mais aujourd'hui Il existe plusieurs façons d'effectuer du Web scraping et les développeurs ont pu développer leurs propres scrapers web qui répondent à leurs besoins en utilisant principalement des langages de programmation tel que python ainsi que d'autres bibliothèques, parmi les outils et les logiciels les plus connus :

5.1. Scrapy

Scrapy est un Framework d'exploration Web open source écrit en Python, il a été publié pour la première fois le 2 juin 2008 et la dernière version est Scrapy 0.24 et est compatible avec Python 2.7. Scrapy écrit du code pour les développeurs Python qui cherchent à créer des robots d'exploration web évolutifs où les robots sont définis comme des classes héritant de Classe 'BaseSpider', qui définit un ensemble d'URL de départ" et une fonction 'parse' appelée à chaque itération Web. Les pages Web sont automatiquement analysées et le contenu Web sont extraits à l'aide d'expressions XPath.

Scrapy est idéal car il s'agit d'un Framework basé sur Python qui fournit des outils à la fois pour le scraping et crawling. Il peut être considéré comme le meilleur outil de Web scraping Python pour les nouvelles applications et le plus simple à apprendre et à mettre en œuvre.**[25] [26]**

5.2. Beautiful Soup

Est une bibliothèque Python qui permet d'extraire des données de fichiers HTML et XML. Elle fonctionne avec un analyseur de documents web pour fournir des moyens de navigation, de recherche et de modification de l'arbre d'analyse. Elle permet généralement aux programmeurs d'économiser leurs temps **[27]**

Les non-développeurs peuvent toujours extraire des données des sites web manuellement en utilisant des outils de Web scraping qui peuvent être installés localement. Bien que ces outils ne soient pas idéaux pour les sites web complexes ou à fort contenu en JavaScript. On trouve par exemple : ScrapeSimple, Import.io, Mozenda, ParseHub, Octoparse, DiffBot

5.3. ScrapeSimple

C'est le service idéal pour extraire des sites web est aussi simple que de remplir un formulaire avec des instructions pour le type de données que vous souhaitez Il vous suffit de leur indiquer les informations que vous souhaitez obtenir de quels sites et ils concevront un web scraper personnalisé qui vous fournira les informations

périodiquement (quotidiennement, hebdomadairement, mensuellement ou autre) au format CSV directement dans votre boîte de réception. **[26]**

5.4. Import.io

Est un outil web permettant d'extraire des données d'un site web sans écrire de code fondé en mars 2012. Pour l'extraction des données, l'utilisateur entre l'URL et l'application extrait automatiquement les données dont il a besoin. Une fois l'extraction des données terminée, l'ensemble de données extraites est stocké sur le serveur en nuage d'Import.io et est ensuite téléchargé au format CSV, Excel, JSON **[24].**

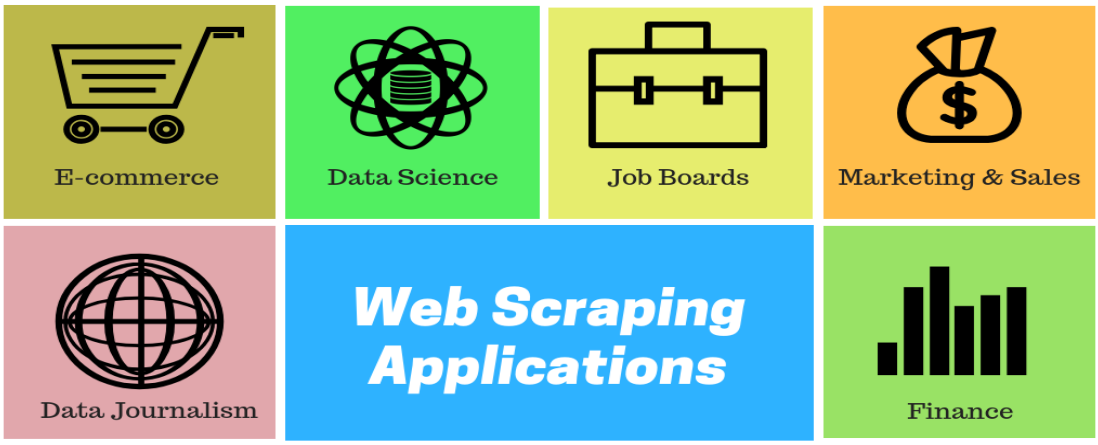
5.5. Mozenda

Est un web scraper puissant. Il peut extraire des données de sites web ainsi que des PDF. Il exécute votre projet de scraping sur leur environnement sur le cloud. L'utilisateur peut également combiner des données provenant de plusieurs sources en un seul ensemble de données **[28]**

6. Les domaines d'application du Web Scraping

Le web scraping peut être utilisé dans divers domaines allant du l'apprentissage automatique vers la médecine ou encore le e-business. On parle de web scraping dans n'importe quel domaine où la collecte des données est requise**.[29]**

Nous énumérer quelques-unes de ses applications



6.1. Surveillance des prix

Dans les affaires de commerce électronique, les entreprises utilisent des prix compétitifs comme stratégie. Afin de réussir dans une telle entreprise, vous devez suivre la stratégie de prix des concurrents. Selon les données de tarification, vous pouvez décider de votre

propre tarification. Vous seriez surpris de voir à quel point le web scraping peut vous aider à avoir un avantage sur les autres en matière de surveillance des prix.

Le prix est le facteur décisif dans les entreprises comme le commerce électronique. Les entreprises de commerce électronique souhaitent suivre les prix de leurs concurrents et présenter leurs prix en conséquence pour obtenir un avantage stratégique.

De plus, ce n'est pas une affaire ponctuelle. Les prix changent constamment et les entreprises de commerce électronique ont besoin de mises à jour en temps réel sur les changements de prix qui se produisent sur les sites Web de leurs concurrents.

C'est là que le scraping Web peut vous donner un avantage considérable. Avec l'aide du web scraping, vous pouvez gratter les prix sur une base continue et suivre les stratégies de prix de vos concurrents.

6.2. *Génération de leads*

Pour toute entreprise, le marketing est d'une importance capitale. Pour le marketing, vous devez avoir les coordonnées de ceux à qui vous envoyez votre matériel marketing. C'est ce qu'est la génération de leads. Avec l'aide du web scraping, vous pouvez obtenir un nombre incroyablement grand de données à partir desquelles vous pouvez générer d'innombrables prospects. Voici comment cela fonctionne :

- Quand vous pensez accélérer votre campagne marketing, quelle est la première chose dont vous avez besoin ? Leads, bien sûr !
- Vous en avez besoin en vrac - des milliers et des milliers d'identifiants de courrier électronique, de numéros de téléphone, etc.
- Le web scraping peut extraire ces identifiants de messagerie et numéros de téléphone avec une précision chirurgicale. Ce n'est pas simplement précis, mais aussi rapide comme l'éclair. Vous l'obtenez en une fraction du temps que vous prendriez pour le faire manuellement.
- Vous l'obtenez également au format CSV ou dans un format que vous pouvez facilement utiliser pour un traitement ultérieur. Vous pouvez également l'intégrer dans vos outils de vente ou d'automatisation.

6.3. *Analyse concurrentielle*

À une époque de compétition acharnée, vous devez très bien connaître vos concurrents et comprendre leurs stratégies, leurs forces et leurs faiblesses. Pour ce faire, vous avez besoin de beaucoup de données. C'est là que le web scraping peut vous aider. Voici comment cela fonctionne :

- Vous auriez certainement besoin de faire une analyse concurrentielle de temps en temps. Mais les données dont vous avez besoin sont dispersées ici, là et partout. Comment y accédez-vous ?
- C'est là que le web scraping peut créer un avantage pour vous. Vous pouvez rapidement extraire les données dont vous avez besoin de plusieurs sources et les exploiter pour une analyse concurrentielle.
- Plus les outils de scraping Web sont rapides et efficaces, meilleure sera l'analyse concurrentielle. C'est si simple !

6.4. *Récupération des images et description du produit*

Chaque nouvelle entreprise de commerce électronique a besoin de descriptions de produits et d'images de milliers de produits qui doivent être affichées sur le site Web. Comment rédiger des descriptions de produits et créer de nouvelles images pour le grand nombre de produits du jour au lendemain ? Le raclage Web peut également vous aider:

- Disons que vous créez une entreprise de commerce électronique. Vous aurez besoin d'images et de descriptions de produits de centaines de milliers de produits, n'est-ce pas ?
- Vous pouvez, bien sûr, demander à quelqu'un de le copier et de le coller manuellement à partir d'un autre site de commerce électronique. Cela prendra probablement une éternité pour le faire. Au lieu de cela, le web scraping peut automatiser le processus d'extraction des images et de la description du produit et peut terminer la tâche en un rien de temps !
- Donc, si vous voulez gérer une entreprise de commerce électronique, le web scraping en fait en quelque sorte partie intégrante.

6.5. *Analyse en temps réel*

- L'analyse en temps réel signifie simplement que les données sont analysées juste après leur disponibilité. Elle diffère des analyses de type batch, car les analyses de style batch peuvent prendre des heures ou des délais pour traiter les données et produire des informations.
- Par rapport à cela, les analyses en temps réel peuvent produire des informations sans délai.
- Les institutions financières utilisent des analyses en temps réel pour la notation du crédit afin de prendre des décisions concernant l'octroi ou l'arrêt du crédit.

6.6. Analyse prédictive

- L'analyse prédictive est un processus d'analyse des données existantes afin de déterminer des modèles et de prédire les résultats ou les tendances futurs. L'analyse prédictive ne peut pas prévoir avec précision le futur, mais il s'agit de prévoir quelles sont les probabilités.
- Outre les autres domaines, l'analyse prédictive trouve son application dans le monde des affaires. L'analyse prédictive est utilisée pour étudier et comprendre le comportement des clients, les produits et diverses autres choses pour déterminer les risques et les opportunités.
- Cependant, comme il est évident, il s'agit d'une sorte d'analyse qui se déroule sur la base d'une grande quantité de données existantes.
- C'est pourquoi le web scraping a gagné en importance car il permet d'extraire et de mettre à disposition de grandes quantités de données qui peuvent ensuite être utilisées dans l'analyse prédictive. En d'autres termes, le web scraping est primordial pour l'analyse prédictive.

6.7. Académique

- Le monde académique dépend beaucoup des données. Le travail académique tourne largement autour de l'un ou l'autre type d'information.
- Qu'il s'agisse d'une mission d'enseignement ou d'un projet de recherche, les universitaires doivent se procurer des données puis les traiter afin d'obtenir les informations nécessaires.
- Grâce au scraping Web, il leur est désormais extrêmement facile d'extraire et de traiter les données dont ils ont besoin.

6.8. Journalisme de données

- Comme le Nom indique que c'est une sorte de journalisme qui utilise des données pour renforcer les reportages.
- L'utilisation d'infographies ou de graphiques est un exemple typique de la façon dont les données sont tissées dans ces histoires.
- La raison pour laquelle les données leur importent beaucoup est que les données donnent de la crédibilité aux arguments et aux affirmations formulés dans les histoires.
- Il est également utile car il permet aux lecteurs de comprendre des sujets complexes de manière visuelle.
- Le scraping Web est ici très pratique car il rend les données disponibles en premier lieu et permet au journaliste de créer un impact grâce à l'utilisation créative des données.

Chapitre II : Initiation au web scraping

7. Défis du Web Scraping

Bien que le scraping Web soit facile à certains égards, il est assez difficile à d'autres égards. Voici les principaux défis qu'on peut rencontrer : **[30]**

7.1. Modifications fréquentes de la structure

Une fois que vous avez configuré votre grattoir, vous pouvez penser que tout est prêt. Mais vous pourriez être surpris ici. Les changements de structure peuvent représenter un défi pour vos plans de scraping Web :

- Il est évident que les sites Web doivent continuer à mettre à jour leur interface utilisateur et d'autres fonctionnalités pour améliorer la perception de l'utilisateur et l'expérience numérique globale.
- En effet, cela signifierait de nombreux changements structurels sur le site Web. Mais cela bouleverserait vos plans car vous avez configuré un robot en gardant à l'esprit ses éléments de code existants. Cela signifie que les grattoirs devraient également être changés.
- Par conséquent, vous devrez continuer à mettre à jour ou à modifier votre grattoir de temps en temps car le moindre changement sur le site Web cible peut faire planter votre grattoir ou au moins vous donner des données incomplètes et inexactes.
- Faire face aux changements constants et à la mise à jour du site Web cible est un défi majeur dans le scraping Web.

7.2. Pièges HoneyPot

Les sites Web qui stockent des données sensibles et précieuses mettront naturellement en place un mécanisme pour protéger également leurs données. De tels mécanismes peuvent contrecarrer vos efforts de scraping Web et vous laisser vous demander ce qui n'a pas fonctionné. Les HoneyPots sont un tel piège :

- Les HoneyPots sont des mécanismes de détection des robots d'exploration ou des grattoirs.
- Il pourrait être là sous la forme de liens « cachés » mais peut être extrait par des grattoirs / araignées.
- Ces liens auraient probablement un style CSS défini à afficher : aucun. Ils peuvent être mélangés en ayant la couleur de l'arrière-plan ou même être déplacés hors de la zone visible de la page.
- Dès que votre robot d'exploration visite un tel lien, votre adresse IP peut être signalée pour une enquête plus approfondie ou même être instantanément bloquée.

- L'autre moyen utilisé pour détecter les robots d'exploration est d'ajouter des liens avec des arborescences de répertoires infiniment profondes.
- Cela signifie qu'il faudrait limiter le nombre de pages récupérées ou limiter la profondeur de parcours.

7.3. Technologies anti-scraping

Les sites Web ayant de gros morceaux de données qu'ils ne veulent partager avec personne n'essaieraient d'utiliser des technologies anti-grattage. Si vous n'en avez pas conscience, vous pouvez finir par être bloqué. Voici tout ce que vous devez savoir :

- Les sites Web tels que LinkedIn, Stubhub et Crunchbase qui craignent le grattage agressif ont naturellement tendance à utiliser de puissantes technologies anti-grattage qui peuvent déjouer toute tentative d'exploration.
- Ces sites Web utilisent des algorithmes de codage dynamique pour empêcher l'accès aux robots et mettre en œuvre des mécanismes de blocage IP même si l'on se conforme aux pratiques légales de scraping Web.
- Il est assez difficile d'éviter de se bloquer et nécessite de trouver une solution qui puisse fonctionner face à de tels mécanismes anti-grattage. Développer un tel outil qui peut fonctionner contre toute attente est extrêmement long et pour ne pas dire, coûteux !

7.4. Qualité des données

Il existe différentes façons d'obtenir des données, mais ce qui compte, c'est la précision et la propreté des données. Ainsi, vous pourrez peut-être extraire des données Web, mais cela peut ne pas être d'une grande utilité s'il y a des erreurs ou si les données sont incomplètes. Voici ce que vous devez garder à l'esprit lors de la recherche de données :

- En fin de compte, vous avez besoin de données propres et prêtes à l'emploi. Par conséquent, la qualité des données est le critère le plus important d'un point de vue commercial.
- Vous souhaitez utiliser les données pour une décision commerciale particulière et pour cela, vous avez besoin de données de haute qualité sur une base cohérente. En particulier, lorsque vous récupérez les données à grande échelle, cela est encore plus critique car vous ne pouvez pas vous permettre d'obtenir des données inexactes à la fin du processus.
- La qualité des données déterminera si le projet verra le jour ou si vous devrez le mettre de côté et renoncer à votre avantage concurrentiel.

- À moins que vous ne puissiez trouver un moyen d'obtenir des données de haute qualité sur une base cohérente, vos mécanismes de scraping Web ne seront pas très fructueux et utiles

V. Conclusion

Dans ce chapitre, nous avons présenté des recherches au sujet de web scraping. Ensuite nous avons défini le processus de web scraping et web crawling avec leurs architectures respectives nous avons aussi comparé ces deux techniques. Par ailleurs nous avons cité les différentes techniques et les outils utilisés pour web scraping et enfin nous avons donné des exemples des applications du web scraping

Le chapitre suivant sera consacré à l'une de ces applications : « le e-commerce et le web scraping »

I. Introduction

Nous avons mentionné précédemment que le web scraping est mis en œuvre dans pratiquement tous les domaines, l'un des domaines les plus mis en avant est sans doute le e-commerce, pour rappel le e-commerce se réfère à la transaction de biens et services entre un acheteur et un vendeur.

Dans ce chapitre, nous traiterons l'extraction des prix des sites d'achat et de vente les plus connus et traiterons par ailleurs la comparaison des prix afin de proposer au client le meilleur prix pour sa recherche.

II. E-commerce

1. Qu'est-ce que le e-commerce ?

Le e-commerce, commerce électronique ou commerce en ligne est

« L'échange pécuniaire de biens, de services et d'informations par l'intermédiaire des réseaux informatiques, notamment Internet. »

En d'autres termes, il s'agit d'un commerce traditionnel doté d'un système qui gère les paiements grâce à des moyens électroniques. **[31]**

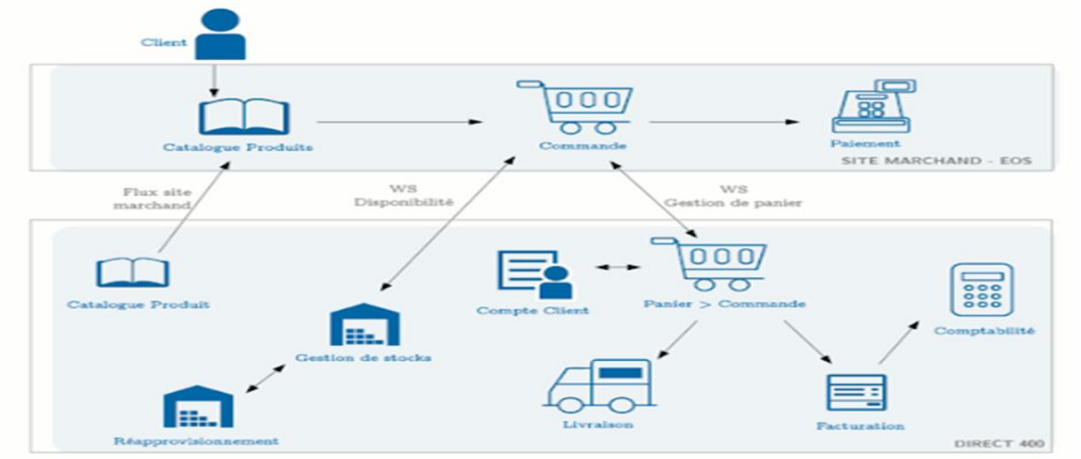


FIGURE 3.1 : Processus e-commerce **[32]**

2. Différences entre e-commerce et e-business

E-commerce et e-business sont deux concepts que l'on a tendance à confondre. En réalité, ils n'ont rien à voir l'un avec l'autre.

Le e-commerce se réfère seulement à la transaction de biens et services entre un acheteur et un vendeur. Le e-business, quant à lui, fait référence à l'ensemble du processus à mettre en œuvre afin de gérer un commerce en ligne. **[31]**

À l'intérieur du e-business, on trouve par exemple (termes en anglais) :

- Le inbound marketing
- Les promotions
- La gestion du stock
- Le SEO
- L'email marketing

Comme on le voit, le concept de e-business est bien plus large que celui de e-commerce. En tant que type de modèle d'entreprise, le e-commerce fait partie du e-business. **[31]**

2. Les différentes formes du commerce électronique

Le monde du e-commerce est très vaste et comprend plusieurs modèles distincts. Il est possible de faire 2 classifications :

- Une classification globale basée sur le modèle commercial (qui vend et qui achète)
- Une autre basée sur le modèle commercial. **[33]**

2.1. Selon le profil commercial

Chaque commerce se dirige à un type de client spécifique. Cela nous permet de distinguer les types d'e-commerce suivants :

- **B2B (*Business-to-Business*) :** entreprises dont les clients finaux sont d'autres entreprises ou d'autres organisations. Par exemple, une boutique de matériel de construction qui s'adresse à des architectes ou des décorateurs d'intérieurs.
- **B2C (*Business-to-Consumer*) :** entreprises qui vendent directement les produits ou services aux consommateurs finaux. C'est la forme la plus courante, et l'on dénombre une multitude d'exemples dans les domaines de la mode, de l'électronique, etc.
- **C2B (*Consumer-to-Business*) :** portails sur lesquels les consommateurs proposent un produit ou un service que les entreprises peuvent acquérir. Il s'agit des portails d'emploi freelance classiques comme Elance, Odesk, Humaniance, etc.
- **C2C (*Consumer-to-Consumer*) :** entreprises qui facilitent la vente de produits entre consommateurs. Les exemples les plus parlants sont eBay, Leboncoin ou tout portail de vente d'articles d'occasion entre particuliers.

2.2. Selon le modèle commercial

Le monde en ligne est un secteur encore peu mâture. Les changements technologiques sont constants et les nouveaux commerces en ligne répondent aux nouvelles nécessités.

On peut différencier le type de e-commerces en fonction des revenus générés ou de la façon dont a lieu l'échange entre l'acheteur et le vendeur :

- **Boutique en ligne disposant de ses propres produits** : c'est la première chose à laquelle on pense quand on pense au e-commerce. Les mêmes caractéristiques qu'une boutique physique, dans une version en ligne.
- **Dropshipping** : pour le client, il semble s'agir d'un e-commerce normal. La différence vient du fait que c'est un tiers qui envoie le produit, et non pas le vendeur. Pour en savoir proche sur le dropshipping, nous vous conseillons de lire notre guide complet consacré au dropshipping (en français).

- **E-commerce d'affiliation** : les commerces d'affiliation vont plus loin encore que le dropshipping. Dans ce cas, non seulement la boutique n'envoie pas le produit, mais la vente n'a pas lieu sur sa plateforme. Le e-commerce redirige le client vers une autre boutique qui lui paie une commission une fois la vente conclue. L'affiliation avec Amazon est la plus fréquente. Par exemple : Biodegradable.es. Si ces modèles vous intéressent, n'hésitez pas à lire nos articles (en anglais) consacrés spécifiquement au marketing d'affiliation ou encore à la vente sans stock.

- **Adhésion** : ce type de e-commerce cherche à ce que ses clients effectuent des achats récurrents. Le moyen privilégié pour les obtenir est à travers d'un abonnement périodique (hebdomadaire, mensuel, trimestriel, etc.). Ce type d'adhésions est actuellement en vogue avec les « boîtes surprises ». Il s'agit d'une boîte envoyée chaque mois (ou à une autre fréquence) et qui contient certains produits
- **Marketplace** : une marketplace est une boutique regroupant plusieurs boutiques. Il s'agit d'un site web sur lequel différents vendeurs proposent leurs produits. Amazon est l'exemple de marketplace par excellence : plusieurs entreprises mettent leurs produits en vente sur la plateforme en échange d'une commission reversée à Amazon.
- **Services** : un e-commerce ne vend pas forcément des produits. Formations, conseils, coaching et, de manière générale, tout temps échangé contre de l'argent. C'est une bonne option viable pour commencer sans prendre de risque.

3. Avantages et inconvénients d'un e-commerce

Pourquoi les e-commerces sont-ils devenus si nombreux sur Internet en si peu de temps ?

3.1. Avantages

- **Davantage de clients** : ni une boutique locale ni une entreprise implantée dans plusieurs villes ne peut atteindre autant de personnes qu'un e-commerce. Pouvoir acheter et vendre depuis n'importe quel endroit du globe élargit considérablement le public cible et permet d'obtenir davantage de clients.
- **Pas d'horaires fixe** : à l'inverse des boutiques traditionnelles, qui sont rarement ouvertes 24/24h, le e-commerce n'a pas d'horaires. Le site web reste ouvert et accessible au public toute la journée et le client peut donc faire ses achats à n'importe quelle heure.
- **Moindres coûts** : pouvoir se passer d'un établissement physique permet de réduire les coûts par rapport au fonctionnement d'un commerce traditionnel. Et si le e-commerce fonctionne en mettant en contact des fournisseurs avec des acheteurs, il n'y aura même pas de frais de production (cas du dropshipping).
- **Davantage de marge** : la réduction des coûts et l'augmentation du nombre de clients permettent d'atteindre une plus grande marge qu'avec un commerce traditionnel, même en baissant les prix. On vend davantage et on gagne plus d'argent.
- **Scalabilité** : dans un e-commerce, vous pouvez vendre à une ou mille personnes en même temps. Dans une entreprise physique, il y a toujours une limite au nombre de

clients que vous pouvez servir à la fois ; dans le commerce électronique, la limite est votre capacité d'attirer des visiteurs. Et bien sûr, celle de votre serveur informatique.

3.2. *Inconvénients / défis*

- **Manque de confiance** : bien que les passerelles et les moyens de paiement aient fait d'énormes progrès et soient aujourd'hui aussi sûrs que dans les boutiques physiques, beaucoup de personnes continuent de ne pas faire entièrement confiance aux transactions en ligne. Pour les aider à faire davantage confiance, il est possible d'utiliser un certificat SSL (https) qui encrypte l'information transférée, ainsi que d'autres certificats qui permettent de garantir la sécurité du client.
- **Produits et services que l'on ne peut ni voir ni toucher** : en tant que clients, nous aimons avoir la sensation de faire un bon achat. Nous aimons voir le produit et le toucher pour nous rendre compte de sa qualité et cela ne peut pas se faire dans un e-commerce. Comment surmonter cet inconvénient ? Grâce à des fiches produits complètes, comprenant des images, des vidéos et une description très détaillée du produit.

- **Connexion Internet indispensable** : c'est évident, mais afin de vendre et d'acheter sur internet, un dispositif connecté à internet est nécessaire. Cela ne concerne pas la majorité des activités en ligne, mais peut représenter un problème pour certains secteurs où le public cible est plus âgé ou moins familiarisé avec les nouvelles technologies.
- **Difficultés techniques** : faire face à des thématiques inconnues est le quotidien des entrepreneurs, que ce soit hors ligne ou en ligne. Dans le cas d'un e-commerce, la partie technologique requiert un minimum de connaissances technologiques, dont tout le monde ne dispose pas. La meilleure façon de résoudre cette difficulté est de déléguer cette partie, bien que cela ait évidemment un coût.
- **Concurrence** : la barrière d'entrée économique pour créer un e-commerce n'est pas aussi élevée que pour un commerce physique. La concurrence est donc plus importante, et il faut se montrer plus compétent que les autres.
- **Temps pour obtenir des résultats** : quand un commerce physique ouvre ses portes, les clients qui passent devant le voient. Obtenir de la visibilité pour un commerce en ligne est plus difficile qu'il n'y paraît. En effet, vous pouvez avoir un très bon produit et être présent sur une bonne plateforme, mais si vous ne travaillez pas pour gagner en visibilité, personne ne vous remarquera.

III. Application du scraping dans l'extraction des prix

Avec différents sites Web de e-commerce, tels que eBay, Amazon, Alibaba, Walmart, Jumia, etc. Les clients ont une connaissance limitée des tendances des produits cibles. Ces sites Web ont souvent des tarifs différents pour le même produit. Donc Trouver le meilleur prix pour un produit donné devient difficile en raison d'une variété de sites Web d'achat. Les clients doivent rechercher manuellement différents sites Web en ligne afin de trouver un prix optimal pour un produit cible par conséquent, un outil spécifique est nécessaire pour montrer les tendances d'un produit particulier sur les marchés en ligne et les sites de commerce électronique. **[18]**

1. L'extraction des prix « Price scraping »

Est une technique permettant de naviguer sur des sites web et de récupérer les prix des concourants en fonction d'une technologie pour extraire depuis ces sites les différentes données affectant le prix.

2. les principaux avantages des sites de comparaison de prix :

2.1 Avantages axés sur le vendeur:

Tout d'abord, les sites de comparaison de prix pour les achats en ligne offrent aux entreprises la possibilité d'élargir leurs canaux de vente. En outre, les fournisseurs ont la possibilité d'utiliser les données de l'agrégateur pour fixer des prix compétitifs, ainsi que de recevoir du trafic supplémentaire du public cible.

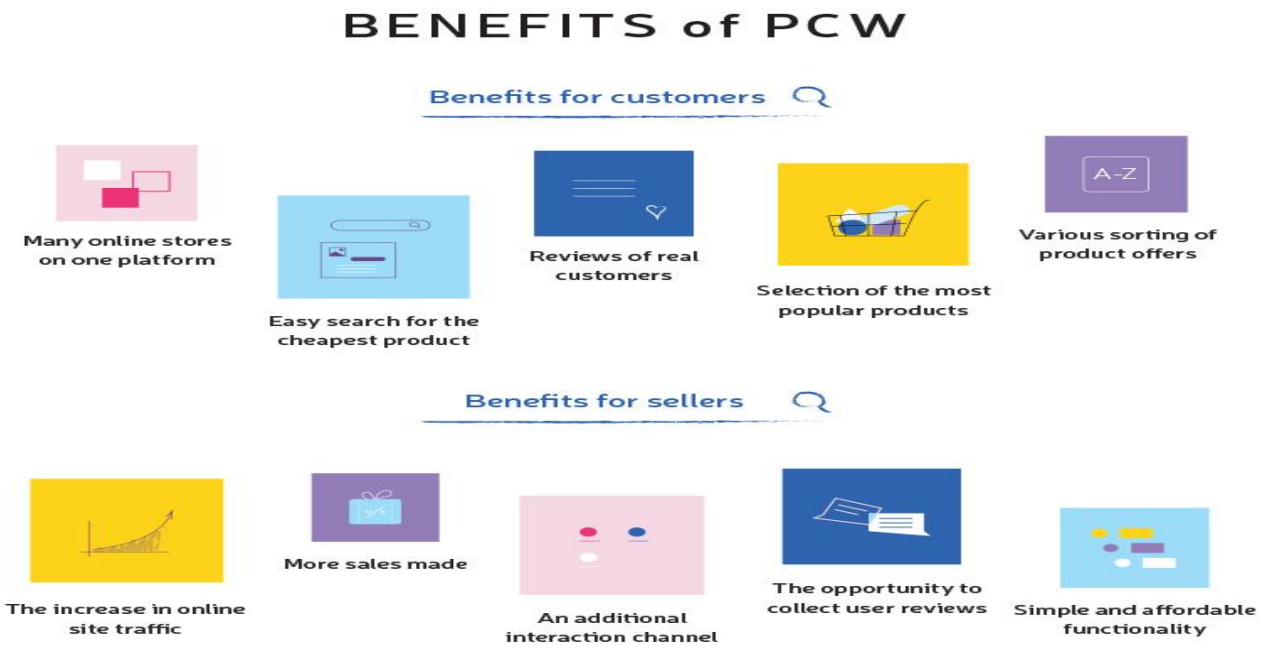
Ainsi, les principaux avantages comprennent:

- L'augmentation rapide du trafic sur le site en ligne;
- Plus de ventes réalisées;
- Un canal supplémentaire d'interaction avec le public;
- L'opportunité de recueillir des avis d'utilisateurs et de travailler sur le développement de la ressource;
- Fonctionnalité simple et abordable qui permet de travailler avec un site Web de comparaison de prix sans aide tierce.

2.2 Avantages axés sur le client:

Un bon site de comparaison de prix est particulièrement nécessaire pour ceux qui font constamment des achats en ligne et ne veulent pas passer trop de temps à essayer de trouver «où est le moins cher». Ainsi, les avantages comprennent:

- la possibilité de voir un grand nombre de boutiques en ligne sur une seule plateforme;
- une réponse rapide à la question de savoir où est le produit le moins cher;
- beaucoup d'avis de vrais clients;
- sélection des produits les plus populaires (ils sont généralement placés en haut des listes);
- différents tri d'offres de produits.



3. La relation entre « Price scraping » et le « Dynamic pricing »

L'extraction des prix « **Price scraping** » est l'un des éléments qui permettent d'élaborer la stratégie de « **dynamic pricing** » **[34]**

Cette technique consiste à :

- Identifier une recherche donnée
- Formatter la recherche pour utilisation
- Stocker les recherches des clients pour la traçabilité
- Localiser les sites cibles de cette recherche via URL
- Convertir les pages en données HTML
- Repérer les éléments d'articles contenant les données relatives aux produits recherchés

- Extraire les prix des produits disponibles
- Faire de même pour tous les sites cibles
- Collecter les URL's, titres, image et prix des produits
- Calculer le meilleur prix (prix minimale)
- Proposer le meilleur prix au client et lui donner le choix de consulter les produits disponibles

La tarification dynamique « **dynamic pricing** » est une stratégie basée sur des algorithmes qui permettent de collecter et analyser les données des concurrents concernant un article spécifique, suivre les prix et l'évolution de l'offre et de demande ainsi que le niveau de stocks, et cela a pour but de fixer les bons prix aux bons moments.

Pour résumer, le « **dynamic pricing** » est une stratégie globale de repérage des meilleurs prix à des moment opportuns, tandis que « **Price scraping** » c'est la technique concrète d'extraction des prix et les comparer. Finalement on dira que ces deux techniques se complètent

4. Les outils nécessaires au « Price scraping »

"Price scraping "ou l'extraction de prix est un programme informatique qui peut être effectué dans plusieurs langages de programmation informatique, on prendra comme exemple le langage Python car nous avons programmé notre application web de « Price scraping » avec ce langage. Nous allons brièvement aborder les quatre aspects de notre projet afin d'énumérer l'ensemble des outils nécessaire au **Price Scrapping** (Ces aspects là nous allons les définir plus profondément par la suite)

4.1. Processus

Avant d'aborder la partie technique il est essentiel d'évoquer la partie stratégie et conception, Effectivement conception et stratégie sont le point de départ de notre application et sans doute la partie la plus importante. De ce fait nous avons commencé par élaborer le processus métier de notre application web avec la méthode **BPMN (**Business Process Model and Notation en anglais) en utilisant le l'outil Heflo. **[35]** nous allons définir la méthode dans la section suivante.

4.2. Algorithme

Après avoir défini le processus métier il nous est devenu plus clair et plus simple de mettre en place notre algorithme de recherche et d'extraction des prix et calculer le meilleur prix pour le présenter au client.

4.3. Conception

L'étape qui suit de celle de la définition de l'algorithme est de définir la conception en utilisant les diagrammes de la méthode **UML**. Avec l'outil dédié **StarUML**

4.4. Développement

La dernière phase de notre projet est de concrétiser notre travail théorique en passant par le développement du système en question, et dans ce cas plusieurs outils sont nécessaires à savoir : (l'ensemble de ces outils vont défini en détail dans le prochain chapitre)

- Le langage Python pour le développement backend
- Le langage SQL pour la base de données
- Les langages HTML, CSS, JavaScript pour le développement frontend
- Outil de développement IDE Visual Studio Code
- Le Framework Django
- La bibliothèque requests (Pour les récupérer les requêtes entre serveurs et pour gérer les interactions client-serveur)
- La bibliothèque Quote_Plus (Pour le formatage des recherches afin de les intégrer aux URL's cible des sites)
- La bibliothèque BeautifulSoup Bs4 (Responsable de l'extraction des pages web et de les convertir en HTML)
- Compilateur de prix (la partie du programme que nous avons réalisé responsable de la collecte, analyse et traitement des prix, effectivement ce compilateur extrait et collecte les données des prix et vérifie les correspondances avec les produits concernés et compare les prix des produits disponible ou les produits similaires pour ainsi calculer et proposer les meilleurs prix au client)

Tous ces éléments sont des éléments principaux pour le développement d'une stratégie de « **dynamic pricing** » qui consiste à ajuster les prix aux variations de demande et pour qu'à la fin proposer au client les meilleurs prix et lui permettre de choisir ce qui lui convient.

5. Comparaison des prix

La comparaison des prix est un concept qui extrait les prix correspondant à un article servant surtout à aider et clarifier au client les différents prix proposés afin qu'il puisse aisément choisir ce qui lui convient. Et le but de notre travail est de concevoir et réaliser un système de comparaison de prix à partir des géants du web du e-commerce tel que Amazon, Ebay, Alibaba et autres. Ce système permet de collecter les données de différents sites web et afficher les caractéristiques des produits demandés par le client et ceci en lui proposant les meilleurs prix disponibles parmi les boutiques en ligne énoncées ci-dessus, ceci grâce à un programme minutieux et un moteur de recherche adapté afin de garantir le bon traitement des données. En d'autres termes plus techniques il s'agit de faire du « Web Scraping » extraction des données web.

Un système de comparaison de prix est généralement un système composé de :

- Une plateforme électronique sur web (Site web)
- Un serveur de données
- Un serveur d'applications
- Un gestionnaire de panier d'achat
- Un programme de « web scraping » (module d'extraction)
- Un compilateur de prix (Pour la collecte et traitement des prix extraits)
- Un programme de calcule et de comparaison
- Un programme de redirection vers les liens cibles des articles proposés

Le système crée un panier virtuel portable qui permet à l'internaute de naviguer sur les marchands rivaux sans avoir l'apparence de quitter le site du commerçant d'accueil et permet en outre à l'acheteur de retourner rapidement sur le site web du commerçant hôte pour acheter des articles dans le Panier d'achat.

IV. Présentation de l'application My-Price-check vu théorique

1. Concept

My-Price-Check est une application web de comparaison de prix à partir des géants du web du e-commerce tel que Amazon, Ebay, Alibaba et autres. My price check permet de collecter les données de différents sites web et afficher les caractéristiques des produits demandés par le client et ceci en lui proposant les meilleurs prix disponibles parmi les boutiques en ligne énoncées ci-dessus, ceci grâce à un programme minutieux et un moteur de recherche adapté afin de garantir le bon traitement des données. En d'autres termes plus techniques il s'agit de faire du « Web Scraping » extraction des données web.

2. Défis et contraintes majeurs de réalisation

Très souvent les sites e-commerce ne proposent pas le même produit, donc pour comparer des prix il faut trouver le même produit !

Le challenge était de trouver les mêmes produits ou à la rigueur les plus similaires possibles, de ce fait on a dû implémenter un moteur de recherche qui puise des informations sur le produit dans l'historique du site lui-même afin de garantir une comparaison cohérente et homogène, dans le cas où un produit est introuvable dans tous les sites en même temps le système va proposer une comparaison avec les produits avec les caractéristiques les plus similaires possibles et donner la possibilité au client de consulter les produit un par un dans leurs établissements d'origine

L'extraction de données (le web scraping) est une pratique qui n'est pas interdite en soi mais qui est quand même dépréciée fortement qui est même considérée comme enfreinte à la juridiction de quelques géants du web donc il fallut faire preuve de raison et avoir une bonne justification, dans un cas comme le nôtre c'est un cadre avec des intentions de recherches et de développement donc utilisé à petite échelle et déni de mauvaises intentions.

3. Architecture applicative

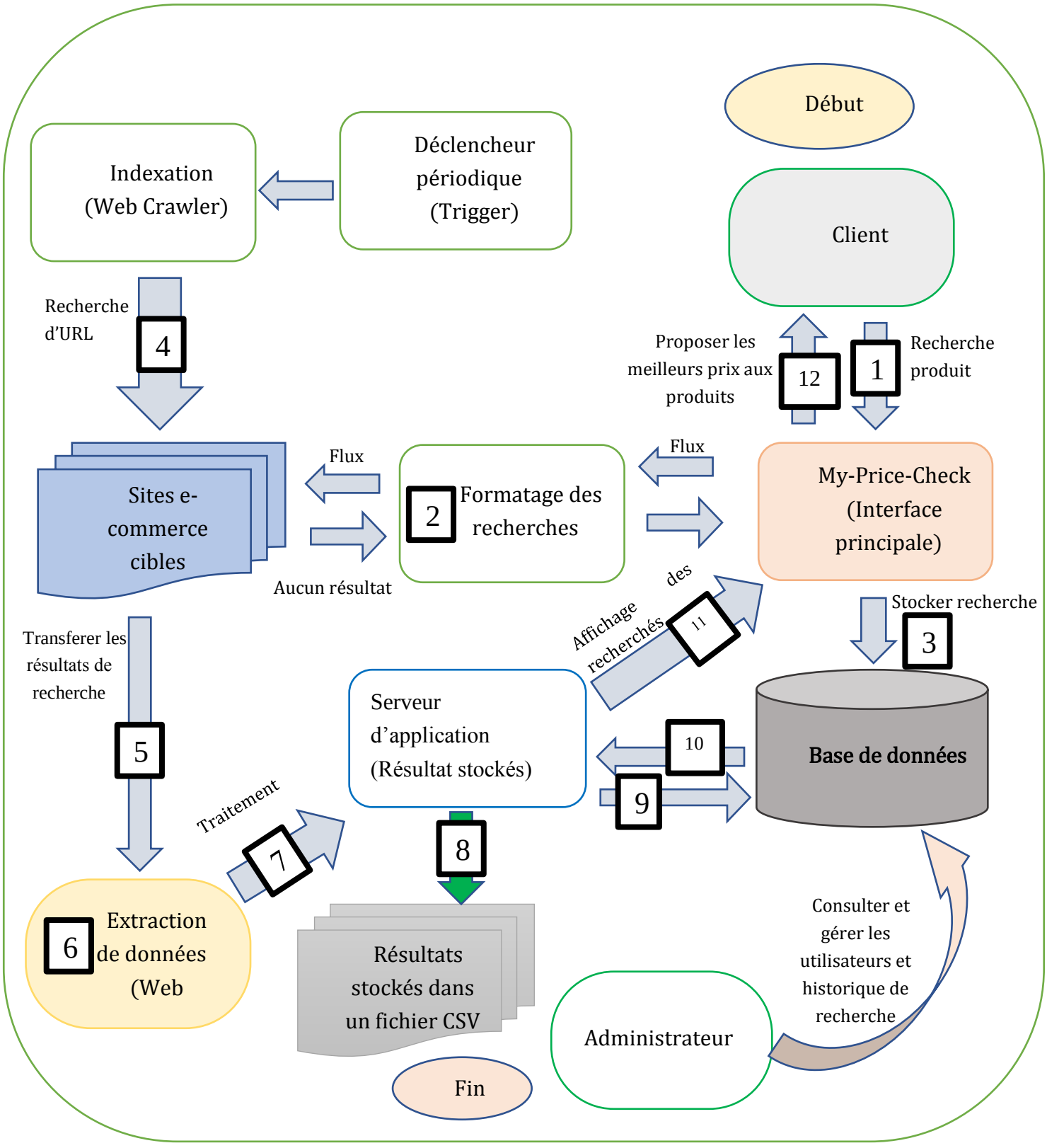


FIGURE 3.2 : Architecture applicative de application Price-Check

4. Processus métier de l'application My-Price-Check

My Price Check permet au client d'obtenir le meilleur prix de ce qu'il recherche.
L'application donnera des prix différents des sites les plus connus comme Amazon, ebay, Alibaba, et autres, comparera les prix, affichera tous les prix et lui permettra de visiter chaque lien de produit qu'il souhaite choisir.

- L'utilisateur entre sur le site et tape le nom du produit qu'il souhaite
- Après avoir cliqué sur le bouton de recherche, les produits disponibles dans des sites e-commerce internationaux avec leurs prix s'afficheront
- Devant lui, le client verra également le meilleur prix (le plus bas)
- Le client peut choisir le lien qui a été fourni par le système et il peut même choisir les autres liens.

Nous allons tout d'abord présenter le processus de l'interaction Administrateur-Système avec le modèle BPMN

4.1. C'est quoi BPMN ?

Business Process Model and Notation (BPMN en anglais), c'est-à-dire « modèle de procédé métier et notation », est un modèle de processus d'affaires et une notation pour décrire les chaînes de valeur et les activités métier d'une organisation sous forme d'une représentation graphique standardisée

BPMN a été développé au départ par la Business Process Management Initiative (BPMI), et est maintenu par l'Object Management Group (OMG) depuis la fusion de ces deux Consortium en juin 2005. La version actuelle de BPMN est la 2.0.2 et date de 2013. Elle est depuis juillet 2013 une norme internationale ISO/CEI 19510 **[35]**

4.2. Processus de l'interaction Administrateur-Système

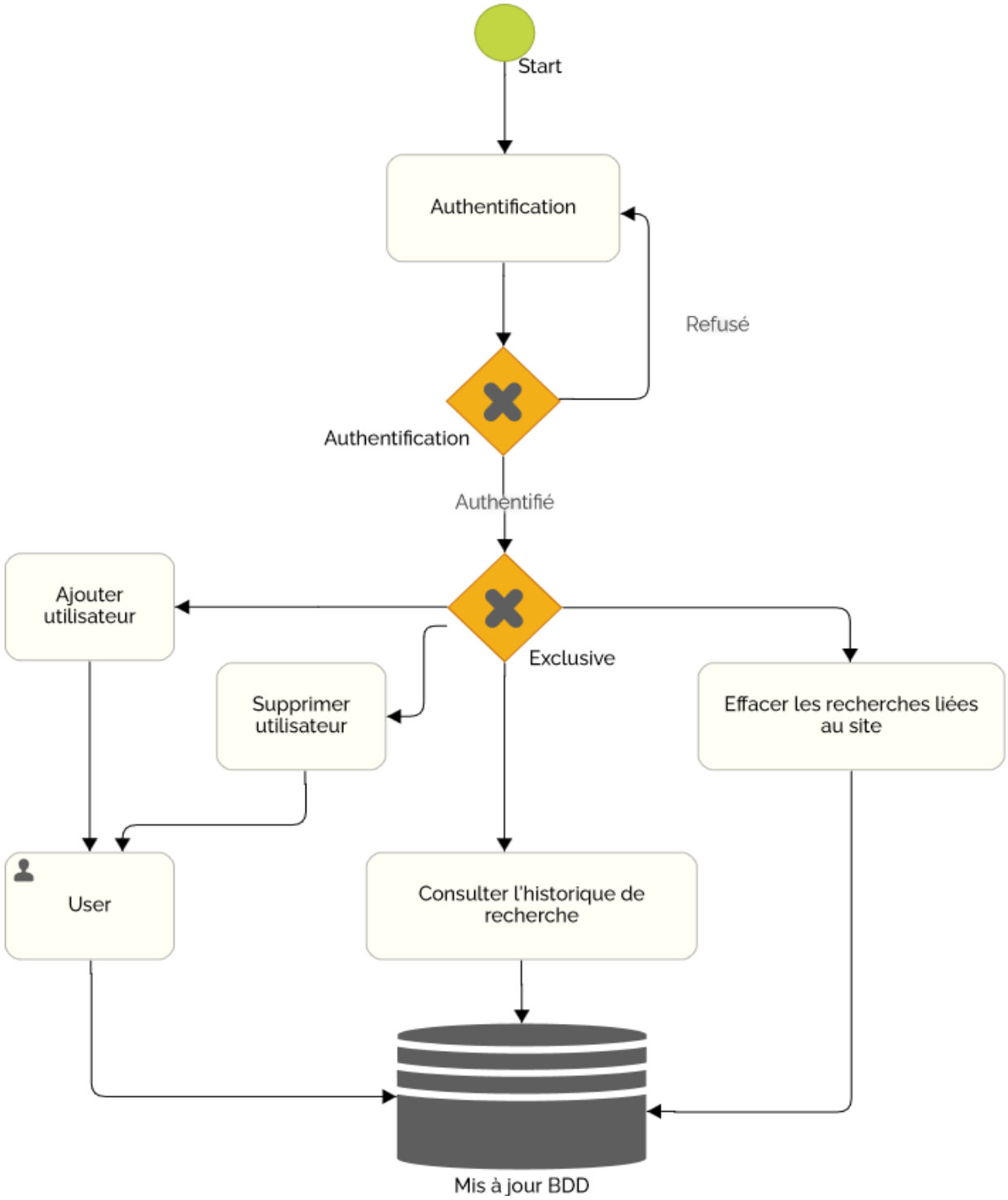


FIGURE 3.3 : Processus de l'interaction Administrateur-Système

4.3. processus d'interaction Client-Système :

Voici ci-dessous le processus de l'interaction Client-système avec tout le cheminement d'une requête client.

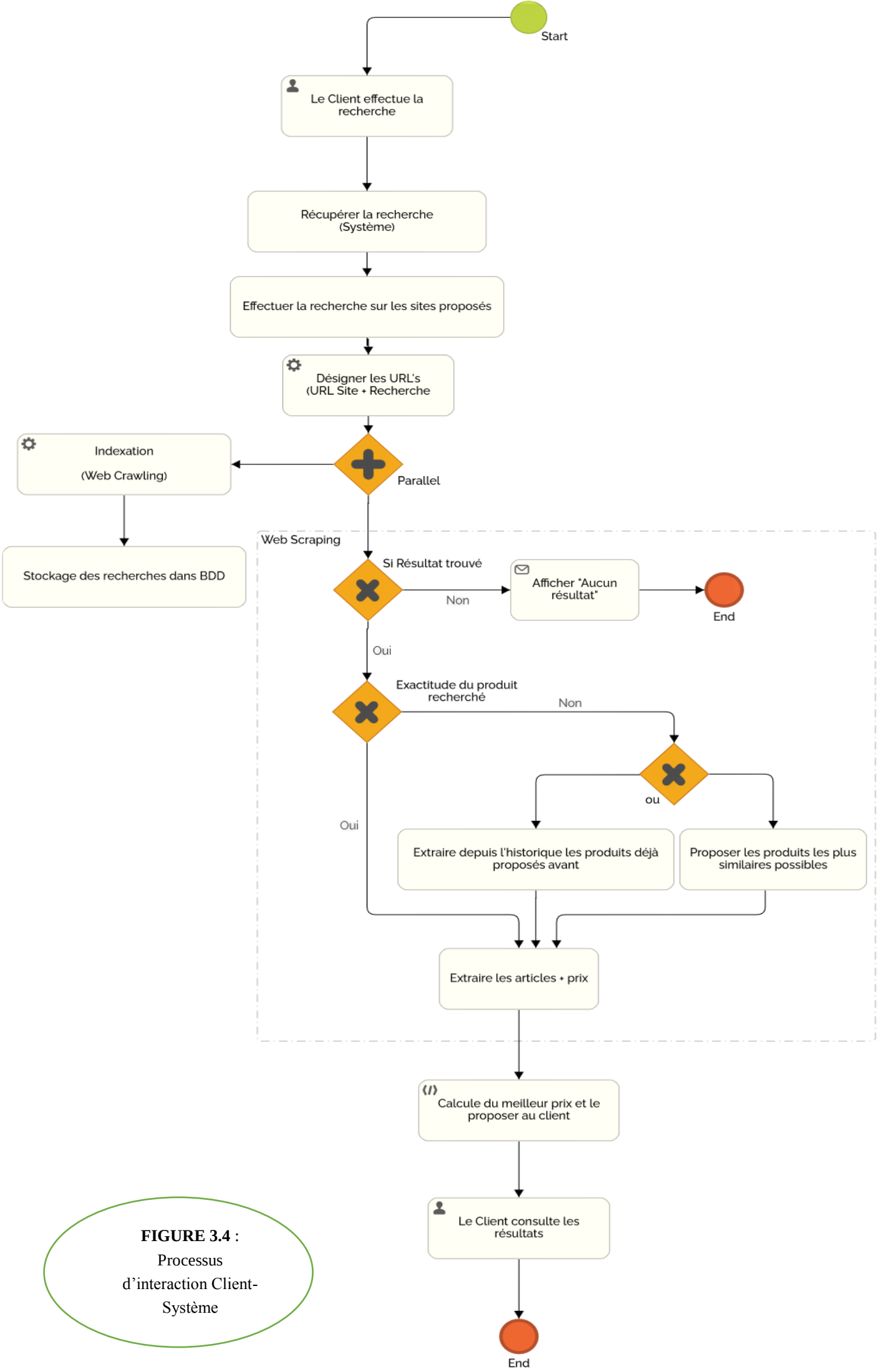


FIGURE 3.4 : Processus d'interaction Client-Système

V. Etude conceptuelle de l'application

La conception consiste à déterminer de façon détaillée et précise, ce dont le système est en mesure de faire. Dans ce chapitre, nous présentons les phases de GLOBALE conception de la démarche proposée dans le développement de l'application web à l'aide du langage UML.

1. Présentation d'UML

UML (en anglais Unified Modeling Language, « langage de modélisation unifié ») UML se définit comme un langage de modélisation graphique et textuel destiné à comprendre et à décrire des besoins, spécifier et documenter des systèmes, esquisser des architectures logicielles, concevoir des solutions et communiquer des points de vue. C'est un langage formel et normalisé et qui grâce à sa représentation graphique, permet de concevoir des solutions et facilite leur compréhension. **[36]**

Son caractère polyvalent et sa souplesse ont en fait un langage de modélisation universel et un standard incontournable pour les activités d'analyse et de conception, et permet notamment de :

- Comprendre et décrire les besoins
- Spécifier un système
- Établir l'architecture logicielle

Face à la diversité des formalismes utilisés par les méthodes d'analyse et de conception objet, UML représente un réel facteur de progrès par l'effort de normalisation.

En effet, UML est issu de la fusion de trois méthodes qui ont le plus influencé la modélisation objet au milieu des années 90 : Booch Grady Booch, OMT (Object Modelling Technique) de James Rumbaugh et OOSE (Object Oriented Software Engineering) d'Ivar Jacobson.

UML est à présent un standard défini par l'OMG (Object Management Group).

UML est fondé sur des concepts orientés objets. **[37]**

UML normalises les concepts objet, sa notion graphique permet d'exprimer une solution objet, ce qui simplifie la comparaison et l'appréciation des solutions.

UML cadre l'analyse objet, il permet non seulement de représenter les concepts objets, mais il sous-entend une démarche d'analyse qui permet de reproduire une solution objet de manière itérative, grâce aux diagrammes, qui supportent l'abstraction.

Un diagramme UML est une représentation graphique, et à chaque vue correspondent des diagrammes qui sont répartis selon leurs aspects statiques ou dynamiques : **[38]**

1.1. Les diagrammes d'UML

UML dans sa version 2, propose treize diagrammes complémentaires qui permettent la modélisation d'un projet tout au long de son cycle de vie. Dans le schéma ci-dessous, l'illustration de ces diagrammes.

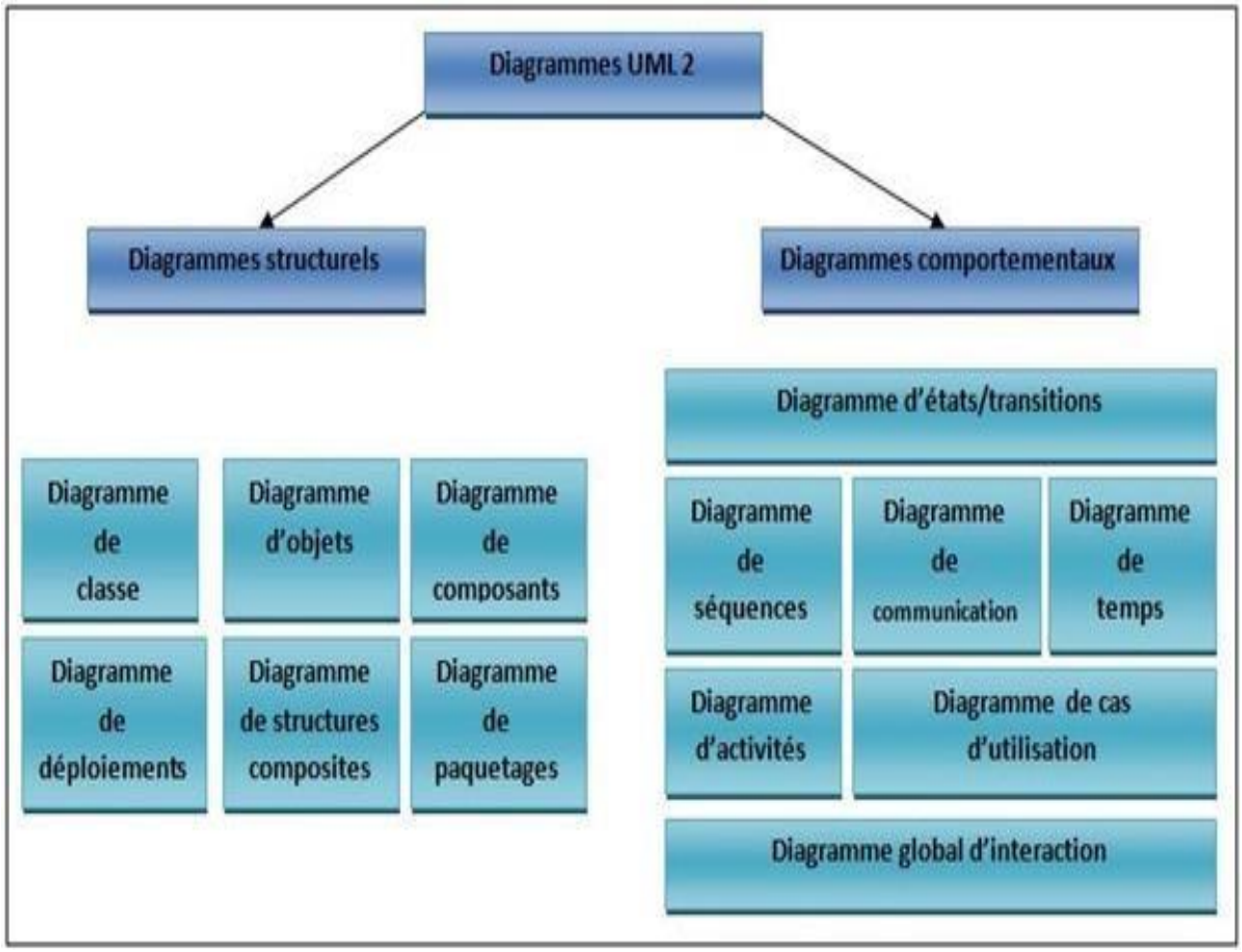


FIGURE 3.5 : Les Diagrammes D'uml 2

Chaque diagramme étant dédié à la représentation des concepts particuliers d'un système logiciel. Les treize types de diagrammes UML, sont répartis en deux catégories :

1.1.1. Diagrammes structurels

Les six diagrammes structurels se présentent comme suit : **[36]**

- Diagramme de classes : il représente l'architecture conceptuelle d'un système à travers une simulation centrée sur les concepts de classes et d'association.

- Diagramme d'objet : montre les instances des éléments structurels et leurs liens à l'exécution et permet d'éclairer un diagramme de classes.

- Diagramme de composants : indique les structures complexes avec leurs interfaces fournies et requises.

- Diagramme de déploiement : définit le déploiement physique des objets sur les ressources matérielles.

- Diagramme de paquetage : spécifie l'organisation logique du modèle et les relations entre paquetage.

- Diagramme de structure composite : désigne l'organisation interne d'un élément statique complexe et décrit les collaborations d'instances.

1.1.2. Diagrammes comportementaux

Les sept diagrammes comportementaux sont les suivants : **[36]**

- Diagrammes de cas d'utilisation : représente la structure des grandes fonctionnalités nécessaires aux utilisateurs du système.

- Diagramme d'activité : montre l'enchaînement des actions et décisions au sein d'une activité et représente graphiquement le comportement d'un cas d'utilisation.

- Diagramme d'états-transitions : décrit le comportement interne d'un objet à l'aide d'un automate d'état finis.

- Diagramme de séquence : montre la séquence verticale des messages passés entre objets au sein d'une interaction.

- Diagramme de communication : désigne la communication entre objets dans le plan au sein d'une interaction.

- Diagramme globale d'interaction : fusionne les diagrammes d'activité et de séquence pour combiner des fragments d'interaction avec des décisions et des flots.

- Diagramme de temps : représente les états et les interactions d'objets dans un contexte où le temps a une forte influence sur le comportement du système.

1.2. Les points forts d'UML

- UML permet une visualisation complète d'un système grâce à ses différents diagrammes.

- Notations claires avec une syntaxe très riche et une sémantique précise, d'où la réduction des ambigüités.

- UML est caractérisé par sa souplesse, sa cohérence et sa performance. En effet, il est utilisé pour des projets innovants, complexes et accélère leur déroulement.

- UML est basé sur les principes de méta-modèle, c'est l'une de ses véritables forces.

- UML formalise l'expression des contraintes en utilisant un langage textuel <OCL> (Object Constraint Language) afin de compléter les diagrammes. **[38]**

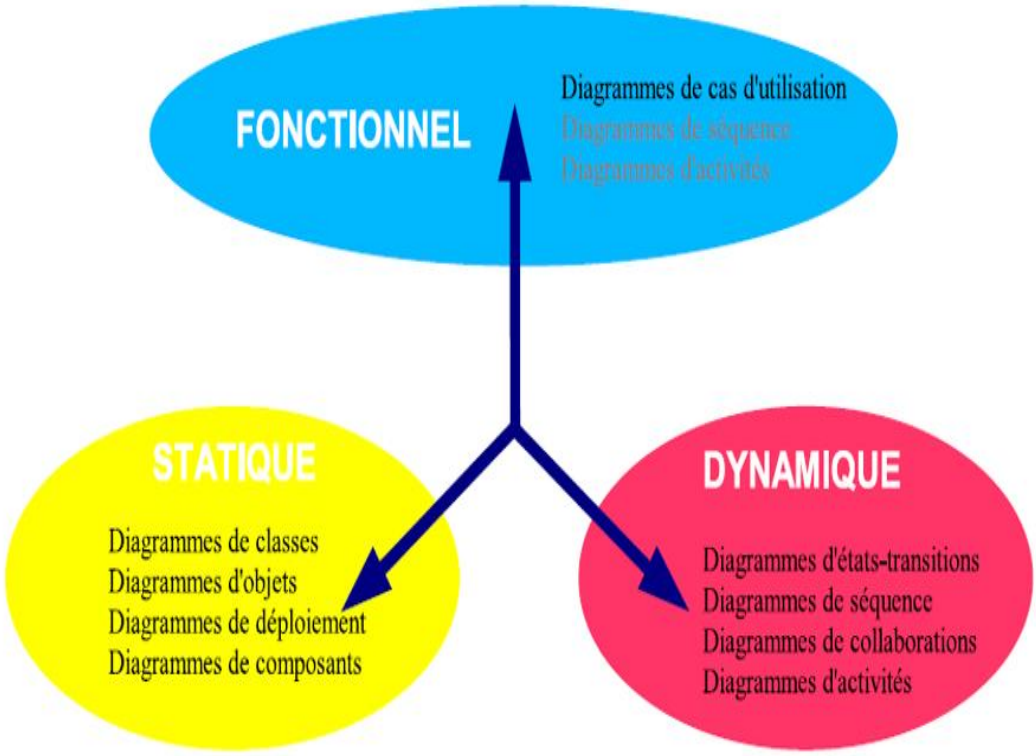


FIGURE 3.6 : Les Diagrammes Uml Vu Generale

1.3. Digramme de cas d'utilisation

Un diagramme de cas d'utilisation capture le comportement d'un système, d'un sous-système, d'une classe ou d'un composant tel qu'un utilisateur extérieur le voit.

Les cas d'utilisation constituent un moyen de recueillir et de décrire les besoins des acteurs du système. Ils permettent également d'exprimer le besoin des utilisateurs d'un système.

Ils sont donc une vision orientée utilisateur de ce besoin au contraire d'une vision informatique.

Les rôles des diagrammes de cas d'utilisation sont :

- Recueillir, analyser et organiser les besoins.
- Recenser les grandes fonctionnalités d'un système.

Il s'agit alors de la première étape UML pour la conception d'un système.

1.3.1. Cas d'utilisation interaction système

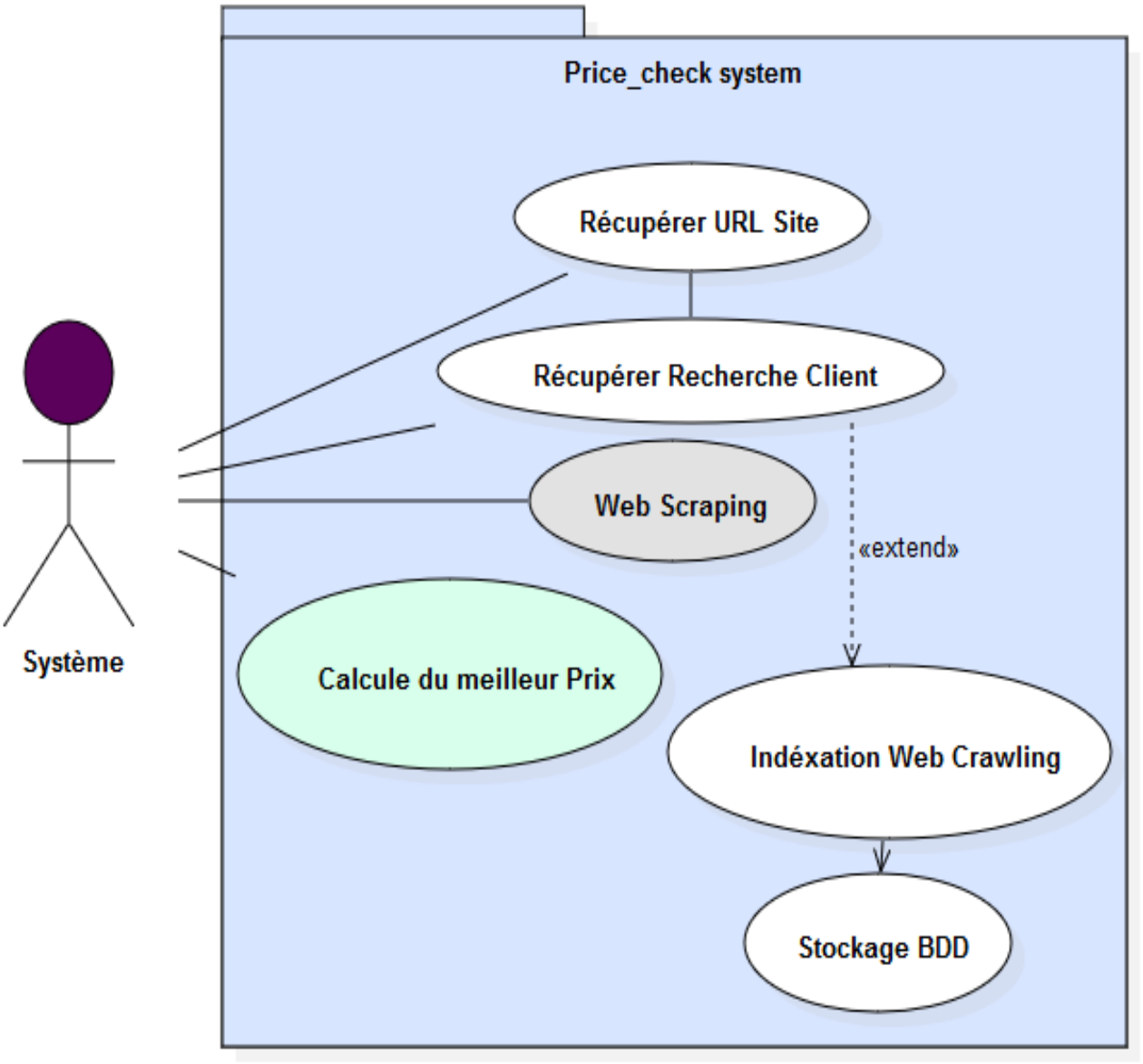


FIGURE 3.7 : Diagramme de cas d'utilisation interaction système

1.3.2. Cas d'utilisation interaction utilisateur

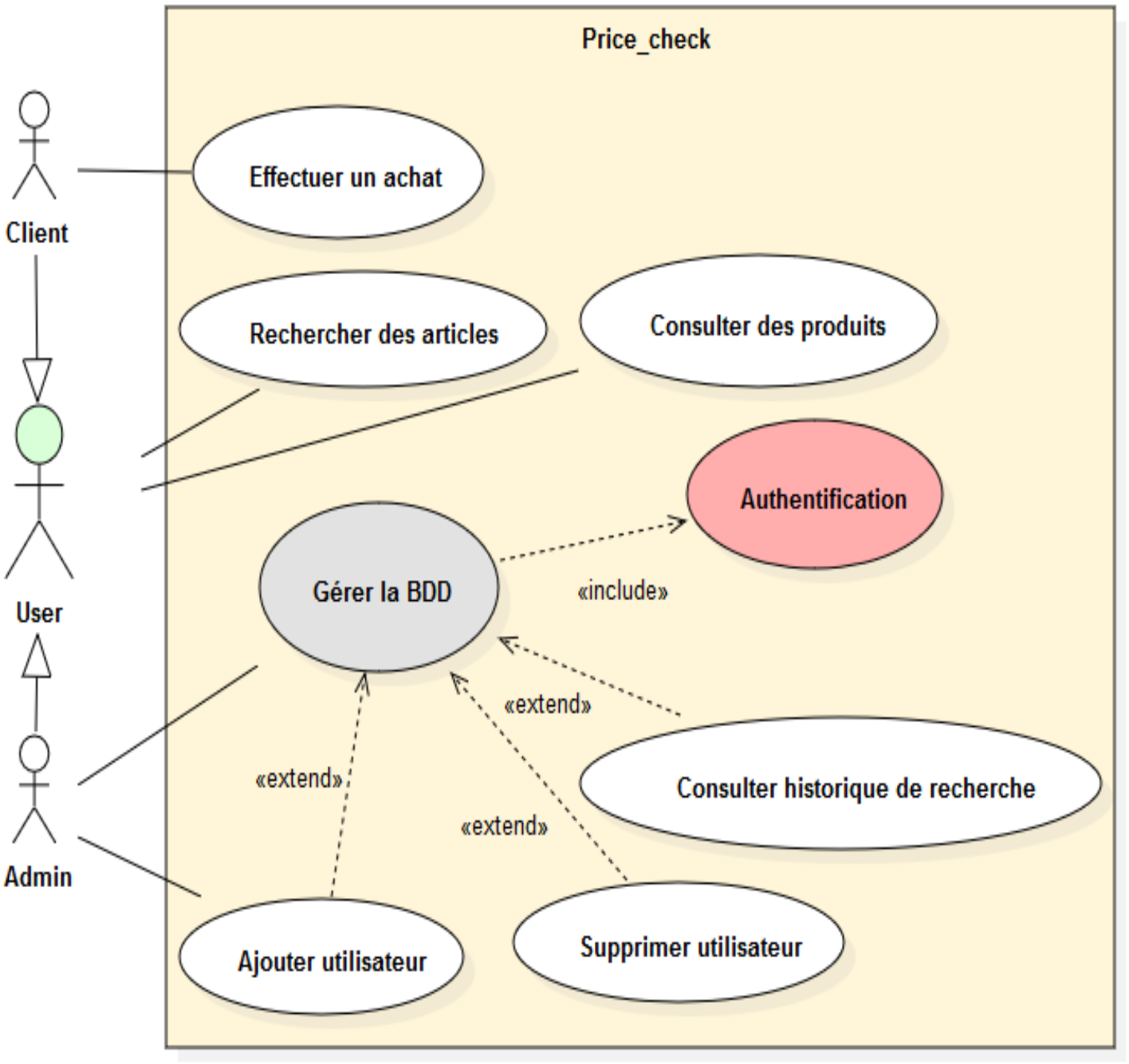


FIGURE 3.8: Diagramme de cas d'utilisation interaction utilisateur

1.4. Diagramme de séquences

Il s'agit d'une explication détaillée d'un cas d'utilisation. Les principales informations contenues dans un diagramme de séquence sont les messages échangés entre les lignes de vie, présentés dans un ordre chronologique.

Un diagramme de séquences est un diagramme d'interaction qui expose en détail la façon dont les opérations sont effectuées : quels messages sont envoyés et quand ils le sont. L'accent est mis sur la communication.
Les diagrammes de séquences sont organisés en fonction du temps qui s'écoule.

La ligne de vie

Une ligne de vie représente l'ensemble des opérations exécutées par un objet. Un message reçu par un objet déclenche l'exécution d'une opération. Le retour d'information peut être implicite (cas général) ou explicite à l'aide d'un message retour.

Message synchrone et asynchrone
Dans un diagramme de séquence, deux types de messages peuvent être distingués :

Message synchrone : dans ce cas, l'émetteur reste en attente de la réponse à son message avant de poursuivre ses actions. La flèche avec extrémité pleine symbolise ce type de message. Le message retour peut ne pas être représenté car il est inclus dans la fin d'exécution de l'opération de l'objet destinataire du message.

Message asynchrone : dans ce cas l'émetteur n'attend pas la réponse à son message, il poursuit l'exécution de ses opérations. C'est une flèche avec une extrémité non pleine qui symbolise ce type de message.

L'objectif du diagramme de séquence est de représenter les interactions entre objets en indiquant la chronologie des échanges. Cette représentation peut se réaliser par cas d'utilisation en considérant les différents scénarios associés. **[36]**

1.4.1. Séquence Admin-Système

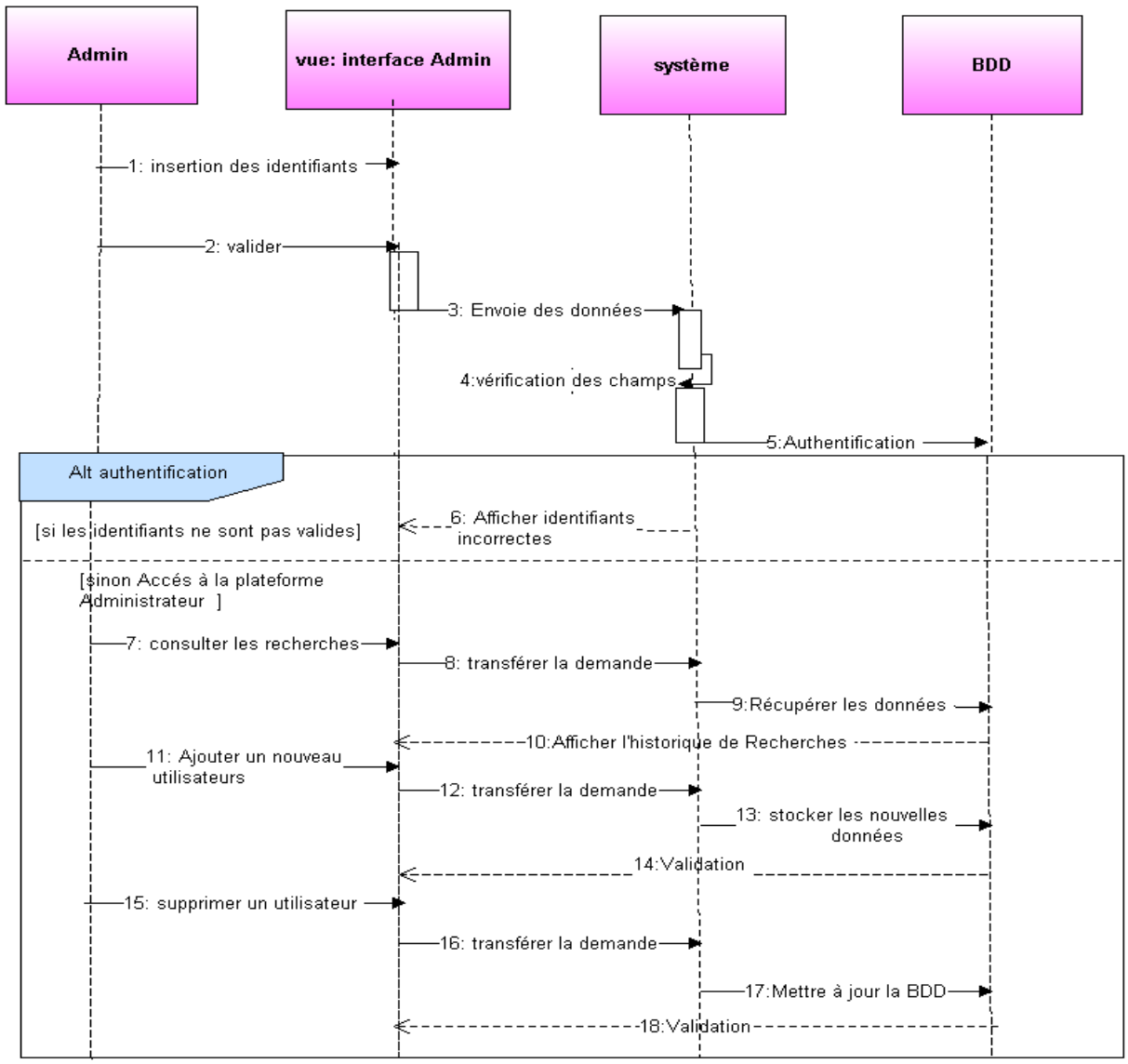


FIGURE 3.9 : Diagramme de séquences Administrateur

1.4.2. Séquence Système-web_scraping

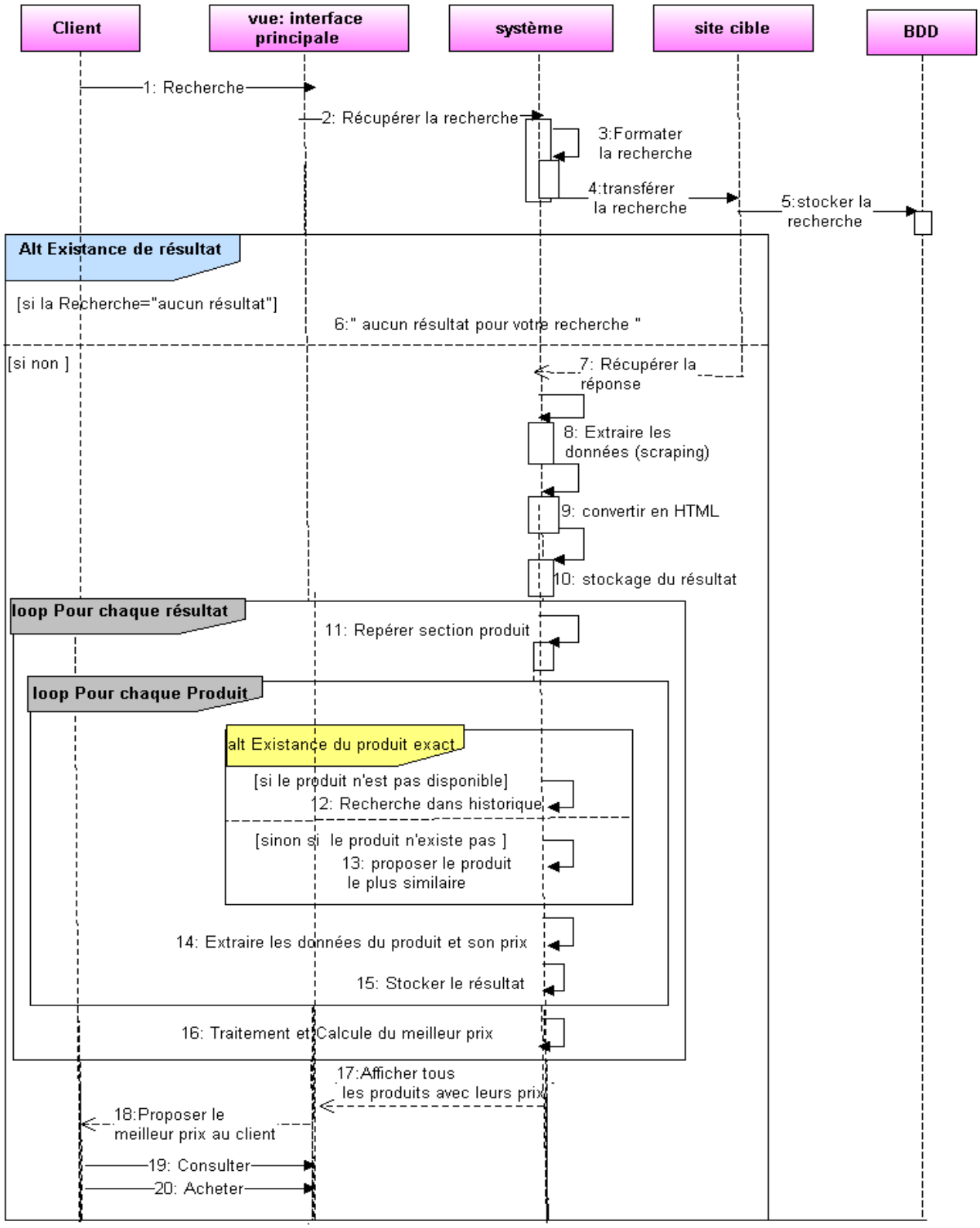


***FIGURE 3.10** : Diagramme de séquences Système-web_scraping (Recherche produits et calcule des meilleurs prix*

1.5. Diagramme de classe

Le diagramme permet de donner la représentation statique du système à développer. Cette représentation est centrée sur les concepts de classe et d'association. Il représente le point central dans un développement orienté objet. En analyse, il a pour objectif de décrire la structure des entités manipulées par les utilisateurs. Et en conception, il représente la structure d'un code orienté objet. **[37]**

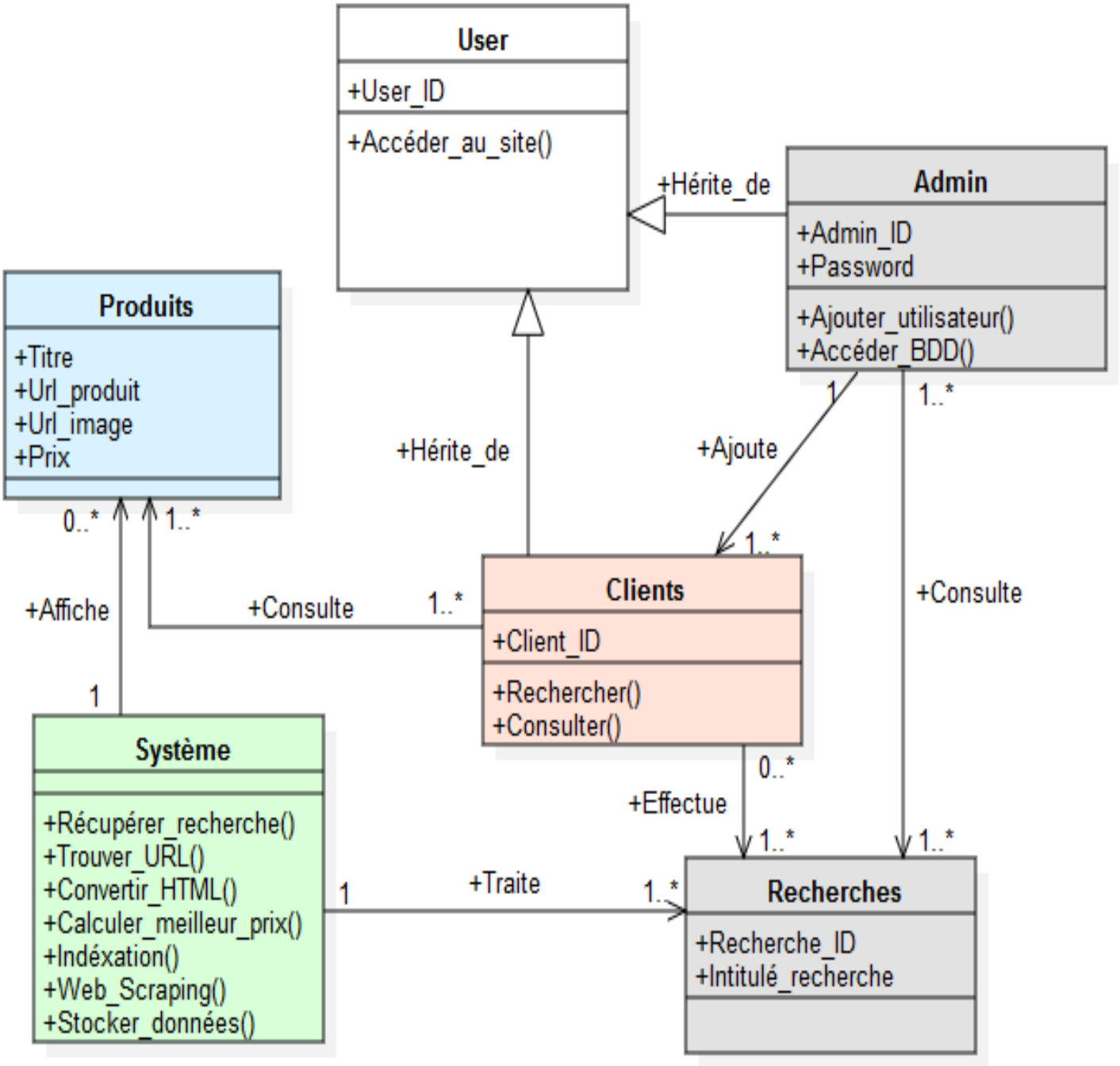


FIGURE 3.11 : Diagramme de séquences Administrateur

VI. Conclusion

Dans ce chapitre, nous avons traité plusieurs aspects du commerce électronique (le e-commerce), nous avons aussi montré l'importance de ce vaste domaine et nous avons ainsi clarifié la relation entre un e-commerce et le client surtout lors de ses achats ou vente en ligne, effectivement nous avons décrit la méthode d'extraction des prix, puis des unités qui composent notre système d'extraction de ces prix pour ainsi faciliter et proposer les meilleurs choix au client.

Enfin nous avons présenté notre système d'extraction de prix avec une étude conceptuelle qui est le point de départ notre application web d'extraction de prix baptisé My-Price-Check.

Dans le chapitre suivant, nous allons entamer la pratique de notre projet et nous allons procéder au développement et implémentation de cette solution informatique qui va extraire les prix de ces produits des principaux sites d'achat et de vente internationaux sur Internet et les comparer pour aboutir au résultat attendu à savoir, proposer au client le meilleur prix des articles qu'il recherche.

Chapitre IV : Réalisation d'une solution de web scraping

I. Introduction

Le Web est une source incroyablement volumineuse et en constante évolution de différents types de données - certaines utiles et d'autres non –

Parfois, nous voulons collecter et stocker ces données pour divers à des fins telles que la recherche ou l'archivage, mais le faire manuellement prendrait probablement plus qu'une vie. À cette fin (et pour éviter les volumes massifs de trafic engendrés par le scraping), certains sites Web et entreprises fournissent des voies publiques (**API**) permettant aux parties intéressées de se connecter et demander des données, par ex. API de Twitter. Cependant, tous les sites Web ne fournissent pas cela, et procure ces interfaces de recherche de données que si les sites concernés considèrent que c'est du web scraping utile et pertinent si non les sites web et entreprises n'approuvent pas l'accès à leurs données et bloquent tout flux entrant.

Le but de cette partie est de mettre au point un système d'extraction de données basé sur la technique du web scraping et aboutir à une solution informatique sous forme d'application web qui finalise au mieux notre travail sur le domaine d'extraction de données, ce système vise aussi à réaliser une structure informatique conforme à l'architecture étudiée et évoquée précédemment mais vise aussi satisfaire les besoins du client et répondre à une problématique en se basant sur l'étude conceptuelle mis en place.

Ce travail de ce dernier chapitre sera l'aboutissement et l'achèvement de notre projet.

II. Présentation de l'application My-Price-Check vu technique

L'application My Price Check est une application d'extraction de prix utilisant la technique du web scraping que nous avons développé pour subvenir aux besoins du client et surtout répondre à notre problématique de départ, nous avons détaillé le coté théorique et conceptuel de l'application dans le chapitre précédent et maintenant dans ce chapitre nous allons aborder l'aspect technique de notre système d'extraction de données.

1. Cheminement fonctionnel

My Price Check est une application web qui permet au client d'obtenir le meilleur prix de ce qu'il recherche l'application donnera des prix différents des sites les plus connus comme Amazon, ebay, Alibaba, etc. et comparera les prix, affichera tous les prix et lui permettra de visiter chaque lien de produit qu'il souhaite choisir.

- L'utilisateur entre sur le site et tape le nom du produit qu'il souhaite
- Après avoir cliqué sur le bouton de recherche, les produits des sites Web internationaux avec leurs prix
- À l'avant, le client verra également le meilleur prix (le plus bas)
- Le client peut choisir le lien qui a été donné par le système et il peut même choisir les autres liens.

2. Environnement de développement

2.1. Langages de programmation

2.1.1. Python

Python est un langage de programmation interprété, multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet. Il est doté d'un typage dynamique fort, d'une gestion automatique de la mémoire par ramasse-miettes et d'un système de gestion d'exceptions, il est ainsi similaire à Perl, Ruby, Scheme, Smalltalk et Tcl.

Le langage Python est placé sous une licence libre et fonctionne sur la plupart des plates-formes informatiques, des smartphones aux ordinateurs centraux, de Windows à Unix avec notamment GNU/Linux en passant par MacOs, ou encore Android, iOS, et peut aussi être traduit en Java ou .NET. Il est conçu pour optimiser la productivité des programmeurs en offrant des outils de haut niveau et une syntaxe simple à utiliser. [**39**]

2.1.2. HTML

Le HyperText Markup Language (HTML) : est le langage de description de la structure des pages Web et c'est aussi un langage de présentation de données ou dans sa dernière version HTML5.

HTML permet d'inclure des ressources multimédias dont des images, des vidéos, et des programmes informatiques, la description de la structure des pages se fait en utilisant le langage de balisage **[40]**

2.1.3. Css

Les feuilles de styles (en anglais "Cascading Style Sheets") (css) sont un langage que publie par **w3c** pour gérer la présentation d'une page Web et pour définir des règles appliquées à une ou plusieurs règles pour les documents HTML sur : le positionnement des éléments, l'alignement, les polices, les couleurs, les marges, l'espacement, les images d'arrière-plan, etc. Avec css, il est possible de changer l'apparence de l'affichage de plusieurs pages dans un seul fichier, ce qui signifie une modification plus grande et plus complète en un temps plus court **[41]**

2.1.4. JavaScripts

JavaScript est un langage de programmation employé dans les pages web interactives Avec les technologies HTML et CSS, permet de créer du contenu mis à jour de façon dynamique JavaScript est parfois considéré comme l'une des technologies cœur du World Wide Web **[42]**

2.1.5. SQL

Le langage SQL (Structured Query Language) est un langage informatique utilisé pour exploiter des bases de données. Il permet de façon générale la définition, la manipulation et le contrôle de sécurité de données.

Créé au début des années 1970 par Donald D. Chamberlin et Raymond F. Boyce, tous deux chez IBM, le langage SQL est aujourd'hui reconnu comme une norme internationale.

Dans la pratique, le langage SQL est utilisé pour créer des tables, ajouter des enregistrements sous forme de lignes, interroger une base de données, la mettre à jour, ou encore gérer les droits d'utilisateurs de cette base de données. Il est bien supporté par la très grande majorité des systèmes de gestion de base de données (SGBD).

Normalisé depuis 1986, le langage est reconnu par la grande majorité des systèmes de gestion de bases de données relationnelles (abrégé **SGBDR**) du marché. **[43]**

2.2. Framework

2.2.1. Django

Django est un Framework open-source de développement Web Python de haut niveau qui encourage un développement rapide qui se base sur une architecture MVT (Model View Template). Django se compose de:

- Un langage de gabarits flexible qui permet de générer du HTML, XML ou autre format texte
- Un contrôleur fourni sous la forme d'un "remapping" d'URL à base d'expressions rationnelles en plus d'Une API d'accès aux données, une interface d'administration fonctionnelle est générée depuis le modèle de données.
- Un système de validation des données qui permet d'afficher des messages d'erreurs automatiques si l'utilisateur entre des données erronées **[44]**

2.3. Technologies et architectures

2.3.1. L'architecture MVT

Le MVT représente une architecture orientée autour de trois pôles : le modèle, la vue et le Template. Elle s'inspire de l'architecture **MVC**, très répandue dans les Framework web. Son objectif est de séparer les responsabilités de chaque pôle afin que chacun se concentre sur ses tâches [**45**]. Le MVT se définit par :

- Le modèle de données (structure des données stockées dans la BDD)
- Les vues (ensemble des scripts définissant le traitement d'arrière-plan)
- Les Templates (les interfaces utilisateurs de traitement externe des données)

2.3.2. BeautifulSoup

Est une bibliothèque Python qui permet d'extraire des données de fichiers HTML et XML. Elle fonctionne avec un analyseur de documents web pour fournir des moyens de navigation, de recherche et de modification de l'arbre d'analyse. Elle permet généralement aux programmeurs d'économiser leurs temps **(défini dans le chapitre2)**

2.3.3. Serveur de données SqlLite3

SQlLite est une bibliothèque écrite en langage C qui propose un moteur de base de données relationnelle accessible par le langage **SQL** (Structured Query Language). Contrairement aux serveurs de bases de données traditionnels, comme **MySQL** sa particularité est de ne pas reproduire le schéma habituel client-serveur

Mais d'être directement intégrée aux programmes et la base de données est intégralement stockée dans un fichier indépendant du logiciel **[46]**

Nous avons privilégié l'utilisation de SqlLite3 au lieu de d'un SGBD plus robuste tel que MySQL ou PostgreSQL car la nature de notre projet ne demande pas un serveur de données de cette ampleur en tous cas pas à cette échelle, SqlLite3 nous suffit de par sa légèreté et sa réactivité, de plus notre système ne se base pas sur un grand flux de données client-serveur_de_données, c'est plutôt l'interaction client-serveur_web qui plus sollicitée.

2.3.4. Heroku

Heroku est une plateforme de cloud computing en tant que service (**PaaS**) supportant plusieurs langages de programmation Java, Node.js, Python, PHP.etc.

Plus précisément c'est un serveur d'application

Les développeurs utilisent Heroku pour héberger, gérer, déployer et mettre à l'échelle des applications, les applications accessibles à distance et prêts à l'exploitation en temps réel [**47]**

Nous avons d'abord choisi d'héberger notre application web pour des besoins de praticité effectivement il est plus pratique d'effectuer nos test dans le web directement et tracer les éventuelles erreurs de développement que sur notre serveur local et de plus c'est une plus-value d'héberger notre système car ça nous permet de faire un travail collaboratif optimal que se soit entre binôme ou il est facile de travailler à distance ou encore lors de phase de consultation avec les éléments extérieurs tel que nos encadrants et cela sans la nécessité de transférer tout le package du projet à chaque fois.

Par ailleurs Héroku est un serveur d'application web avec beaucoup d'atouts intrinsèques notamment :

- Sa réactivité et disponibilité en temps réel
- La possibilité de modifier le projet en plein déploiement
- C'est un serveur web gratuit
- Sa grande compatibilité avec le Framework Django
- Son support performant après déploiement

2.4. Outil de développement

2.4.1. Visual Studio Code

Visual Studio Code est un environnement de développement intégré dédié à la programmation qui de ce fait un éditeur de code extensible, il est développé par Microsoft pour plusieurs systèmes d'exploitation. Les fonctionnalités incluent la prise en charge du débogage, la mise en évidence de la syntaxe, la complétion intelligente du code, la factorisation du code et **Git** intégré pour la traçabilité

Cet éditeur de code a été classé comme l'outil de développement le plus utilisé **[48]**

FIGURE 4.1 : Logo de l'application My-Price-Check

3. Algorithme du fonctionnement de My-Price-Check

4. Pour mieux comprendre le fonctionnement de notre système de web scraping nous avons élaboré un algorithme de recherche du meilleur prix et c'est principalement le fonctionnement de My-Price-Check

```
Début
|   Récupérer la recherche du client
|   Formatter la chaine de caractères de recherche avec « QuotePlus »
|           (Exemple : recherche = « Acer Aspire 5 » => « Acer+Aspire+5 »
|   Pour chaque URL de site web cible
|   |           Attacher la recherche formatté à l'URL
|   |           Effectuer la recherche
|   |           Stocker la recherche dans la BDD
|   |           Si La recherche ne donne pas de résultats Alors
|   |           |       Afficher « Aucun résultat trouvé »
|   |           |       Renvoyer le client à la page de recherche
|   |           Fin Si
|   |           Récupérer la requête réponse
|   |           Extraire les données en texte
|   |           Convertir les données texte en HTML avec le « html.parser »
|   |           Stocker les résultats finaux dans des variable spécifiques
|   |                   (Exemple : « html_data_response »)
|   Fin Pour
|   Pour chaque résultat de recherche html
|   |           Repérer les sections produit dans le document (BeautifulSoup)
|   |           Pour chaque produit
|   |           |       Si le produit exact n'existe pas Alors
|   |           |       |       Rechercher si le produit a déjà été proposé dans
l'historique
|   |           |       |       (le produit existe mais n'est plus en stock)
|   |           |       |       « juste donner une idée du prix au client »
|   |           |       Si Non
|   |           |       |       Proposer les produits les plus similaires possibles
|   |           |       Fin Si
|   |           |       Extraire le titre
|   |           |       Extraire l'image
|   |           |       Extraire le prix
|   |           |       Stocker les résultats
|   |           Fin Pour
|   |           Calculer le meilleur prix
|   Fin Pour
|   Afficher tous les produits trouvés avec les prix respectifs
|   Proposer le meilleur prix au client
Fin
```

FIGURE 4.2 : Algorithme du fonctionnement de My-Price-Check

III. Démonstration et résultat de l'application

1. L'interface principale

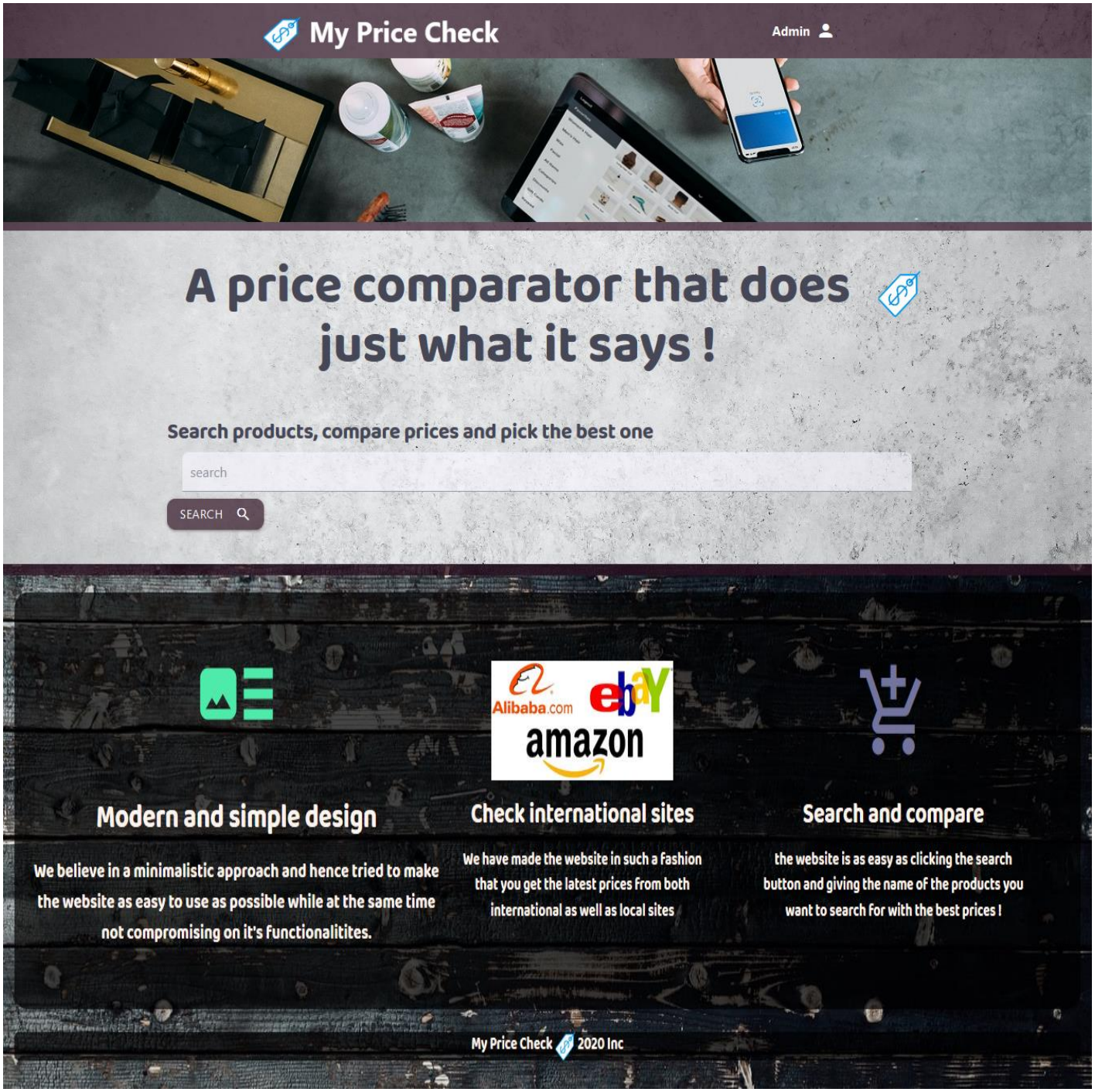


FIGURE 4.3 : Interface principale du site web (interface client)

2. Recherche

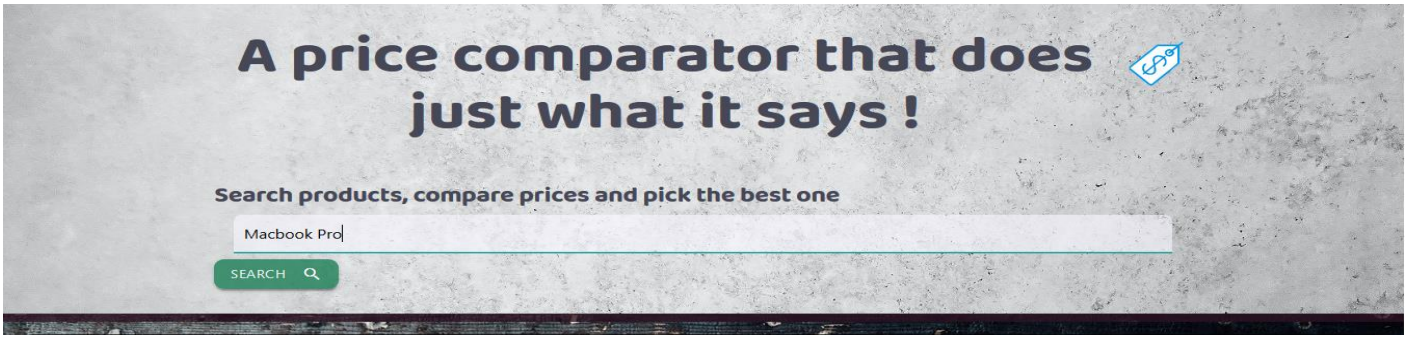
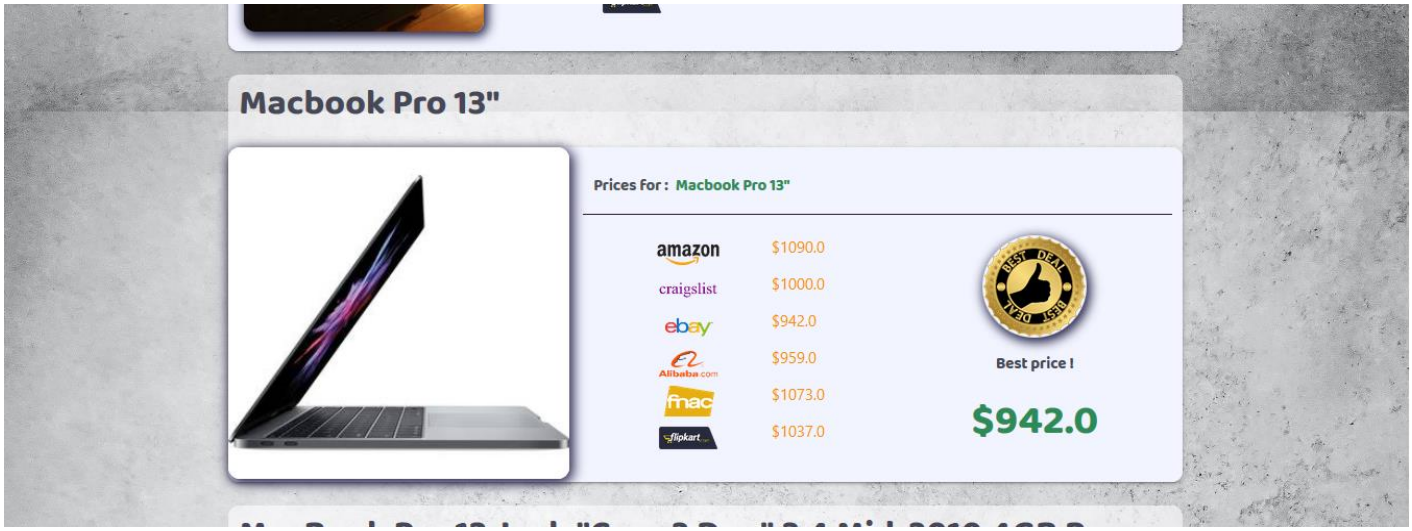


FIGURE 4.4 : Exemple de recherche client 'Macbook Pro'



3. Résultats de la recherche

FIGURE 4.5 : élément de résultat de recherche

FIGURE 4.6 : élément de résultat de recherche

4. Comparaison et affichage du meilleur prix

FIGURE 4.7 : mise en évidence du meilleur prix d'un produit recherché

5. Consultation de produits proposés

Payez jusqu'en 24 fois Voir détails et conditions

Ce produit fonctionne comme neuf et semble neuf. Avec la garantie Amazon Renewed d'1 an.
- Ce produit d'occasion n'est pas certifié Apple, mais a été inspecté, testé et nettoyé par des fournisseurs Amazon certifiés.
- Aucun signe de dommage cosmétique (rayures, bosses, etc.) ne sera visible sur le produit à une distance de 30 centimètres.
- La batterie de ce produit offrira plus de 80% de l'autonomie des batteries neuves.
- Les accessoires peuvent ne pas être d'origine, mais seront compatibles et entièrement fonctionnels. Le produit peut être livré dans une boîte générique.
- Ce produit peut faire l'objet d'un remplacement ou d'un remboursement dans l'année suivant sa réception si vous n'êtes pas satisfait en vertu de la Garantie Amazon Renewed. Voir les conditions ici.

Marque	Apple
Display Size	15 Pouces
Système d'exploitation	Mac OS X
Nombre de coeurs	2
Fabricant de CPU	Intel

Voir plus

FIGURE 4.8 : Redirection vers le site Amazon pour consulter le produit

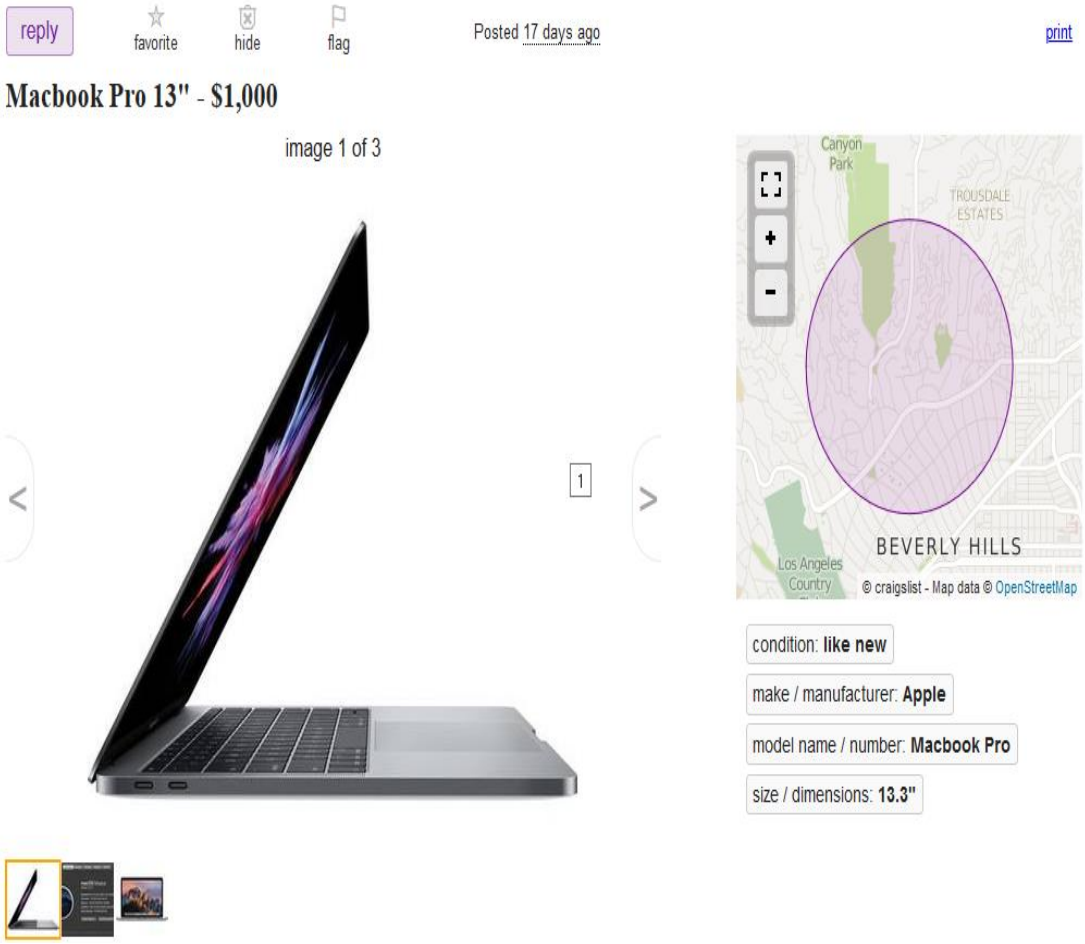


FIGURE 4.9 : Redirection vers le site Craigslist pour consulter le produit (le meilleur prix)

6. Responsivité du site

Le site web a été conçu et modélisé afin qu'il soit utilisé aisément tant sur un support ordinateur que sur un smartphone, effectivement nous avons arrangé le visuel tel que l'affichage rendra correctement sur toutes les plateformes électroniques. Voici ci-dessous quelques interfaces des différentes étapes de recherche d'articles par un client sur appareil mobile (téléphone portable)

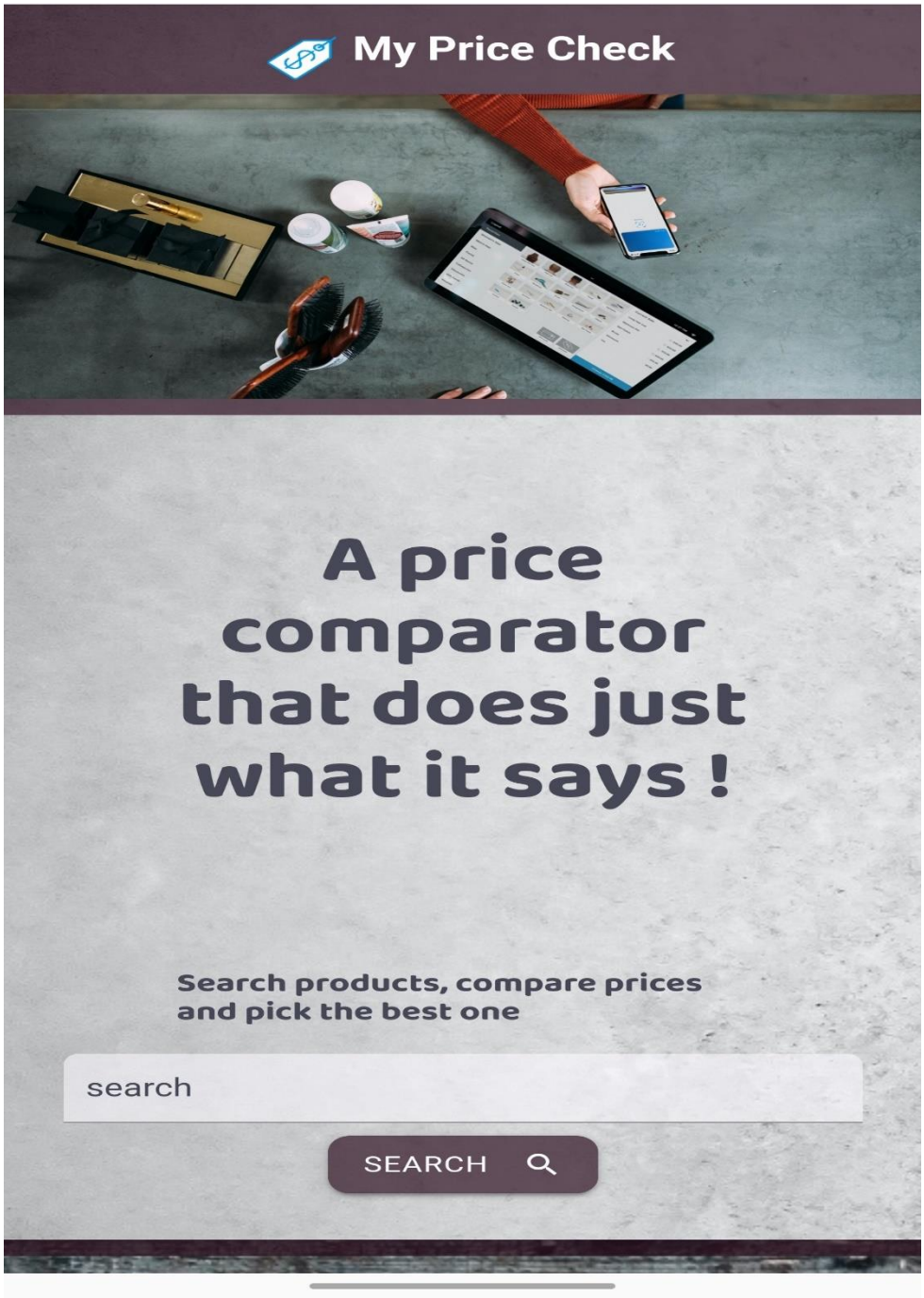


FIGURE 4.10 : Rendu visuel de l'interface du menu principal de l'application sur mobile

FIGURE 4.11 : Rendu visuel de l'interface bas de page du menu principal de l'application sur mobile

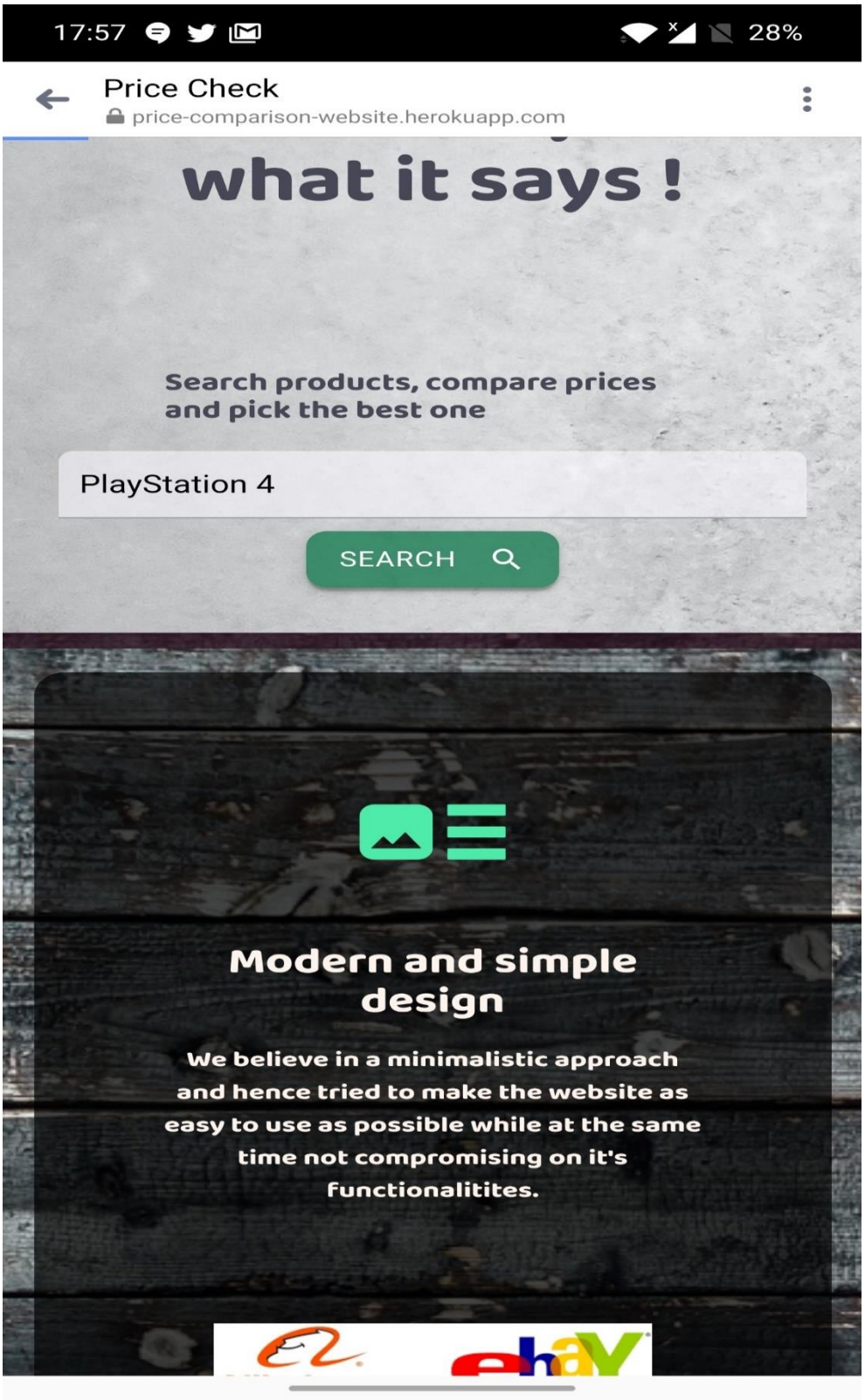


FIGURE 4.12 : Exemple de recherche de produit 'Playstation 4' sur mobile

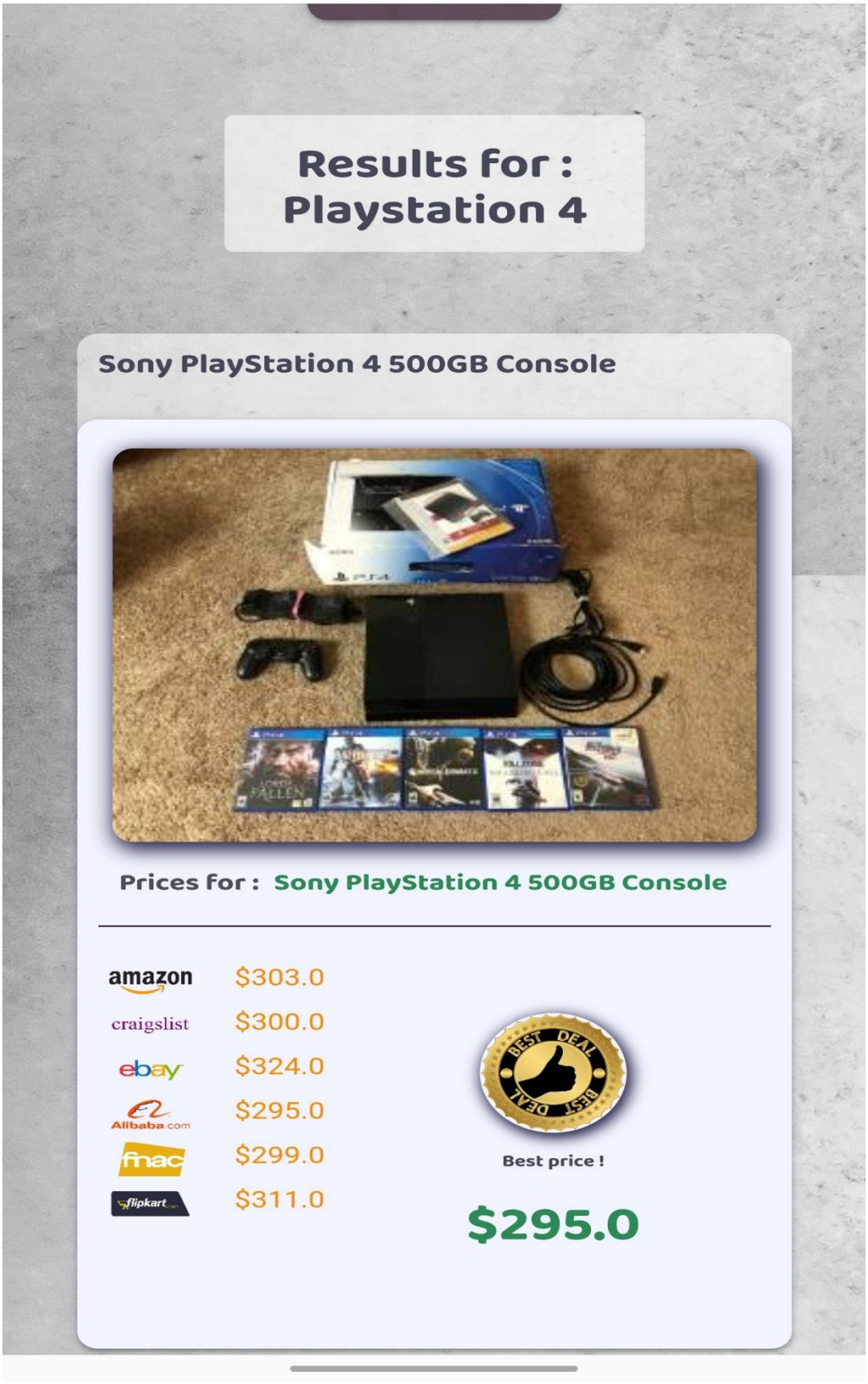


FIGURE 4.13 : Rendu visuel de résultat de recherche sur mobile

FIGURE 4.14 : Rendu visuel de résultat de recherche sur mobile

7. Interfaces administrateur

7.1. Authentification

Nous avons arrangé une interface spéciale administrateur pour qui il puis accéder et gérer les utilisateurs et consulter l'historique de recherches, pour ce il doit s'authentifier.

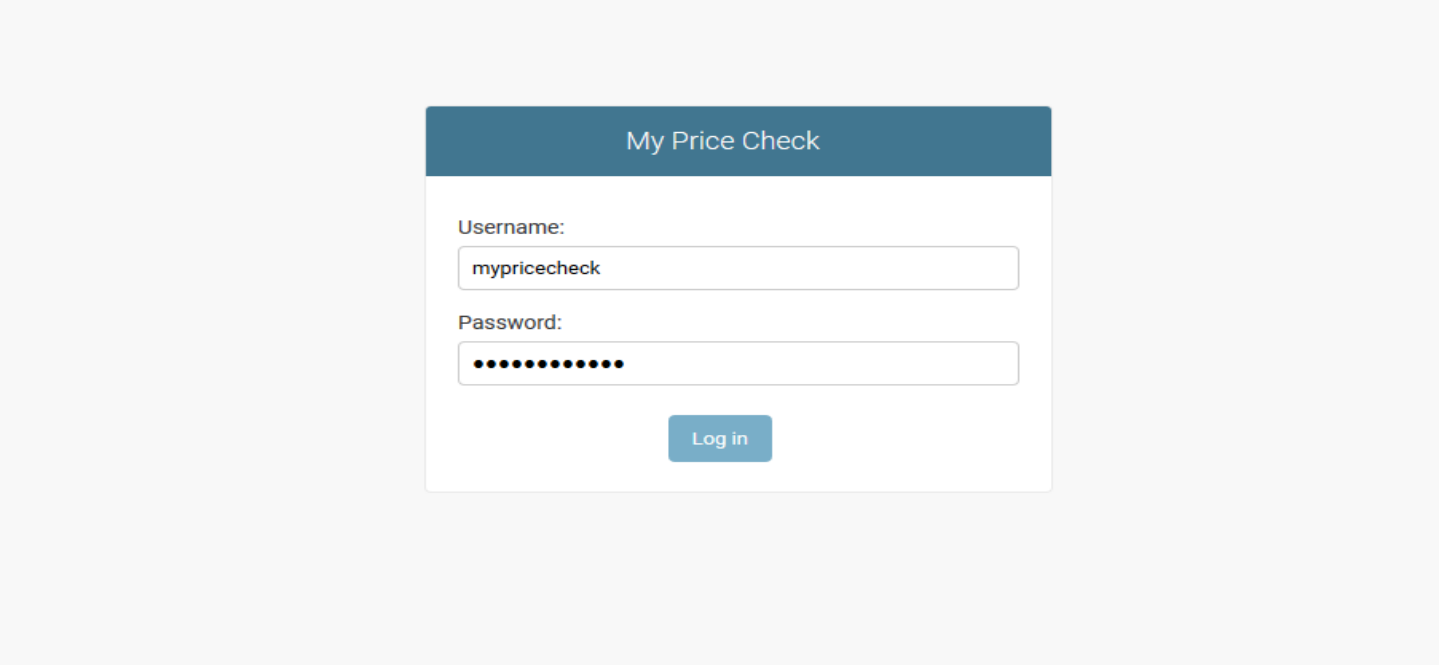


FIGURE 4.15 : interface d'authentification administrateur

7.2. Interface administrateur

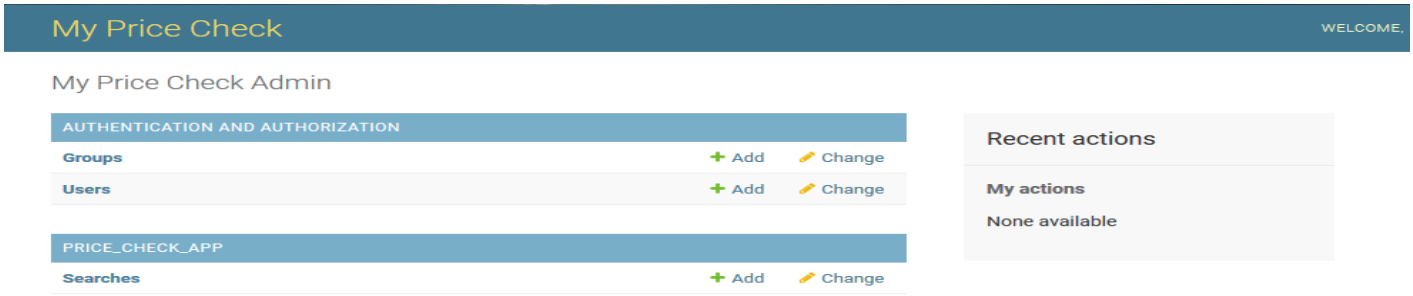


FIGURE 4.16 : interface administrateur

7.3. Consultation d'historique de recherche

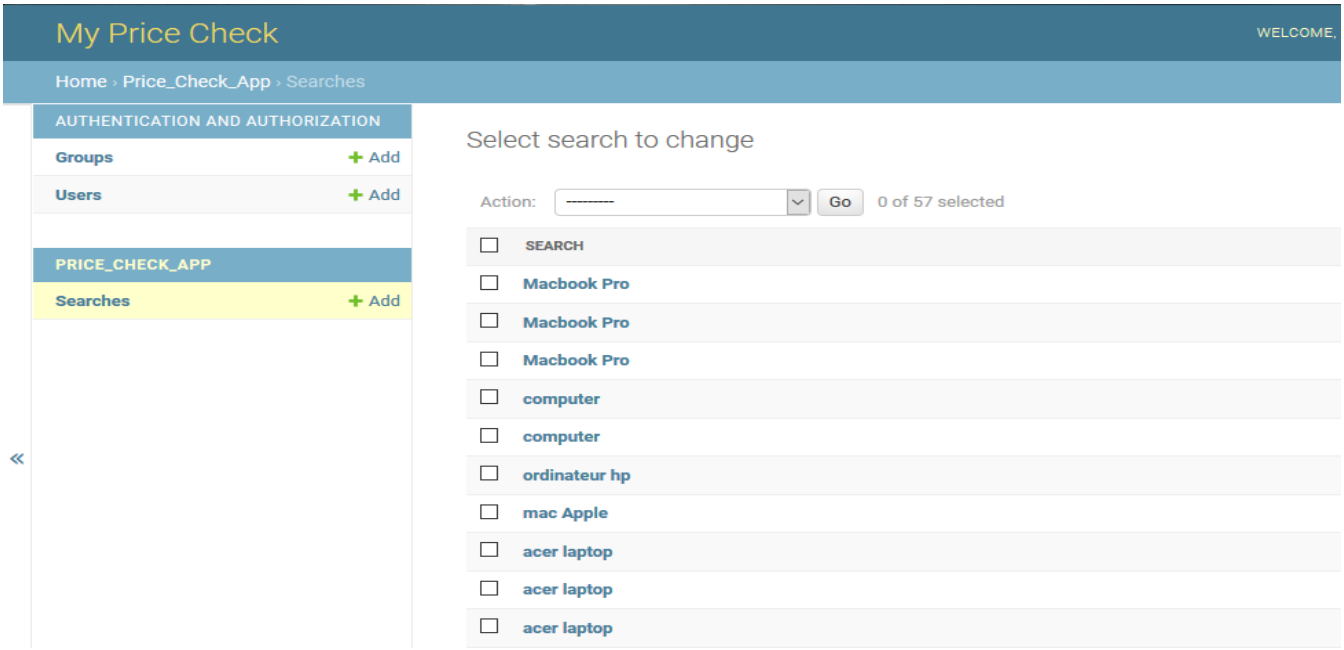
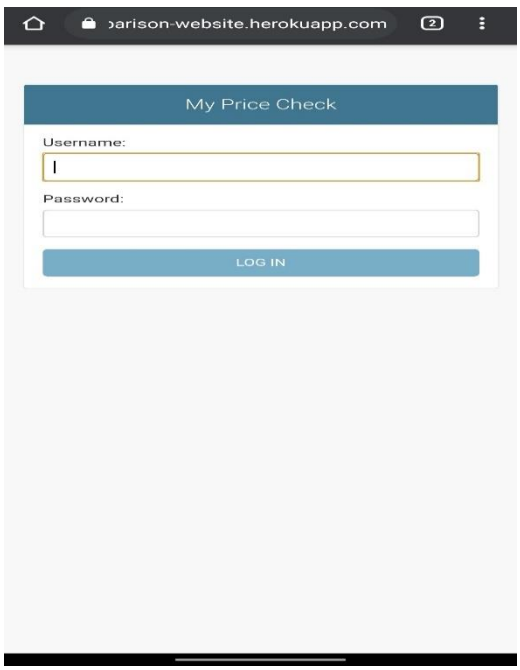


FIGURE 4.17 : interface administrateur consultation d'historique de recherche

7.4. Responsivité



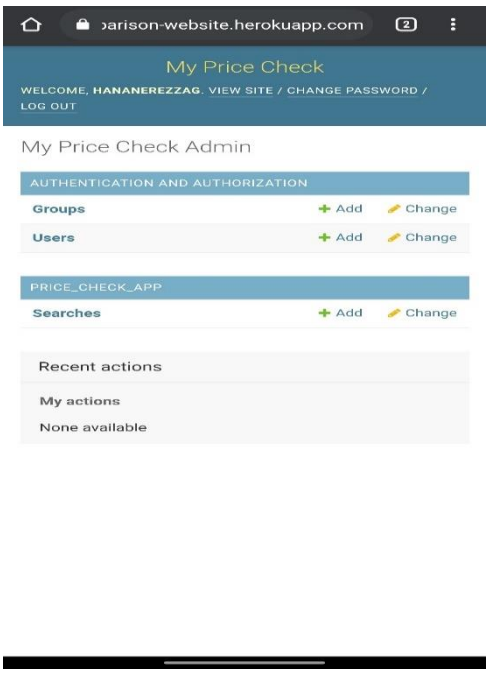


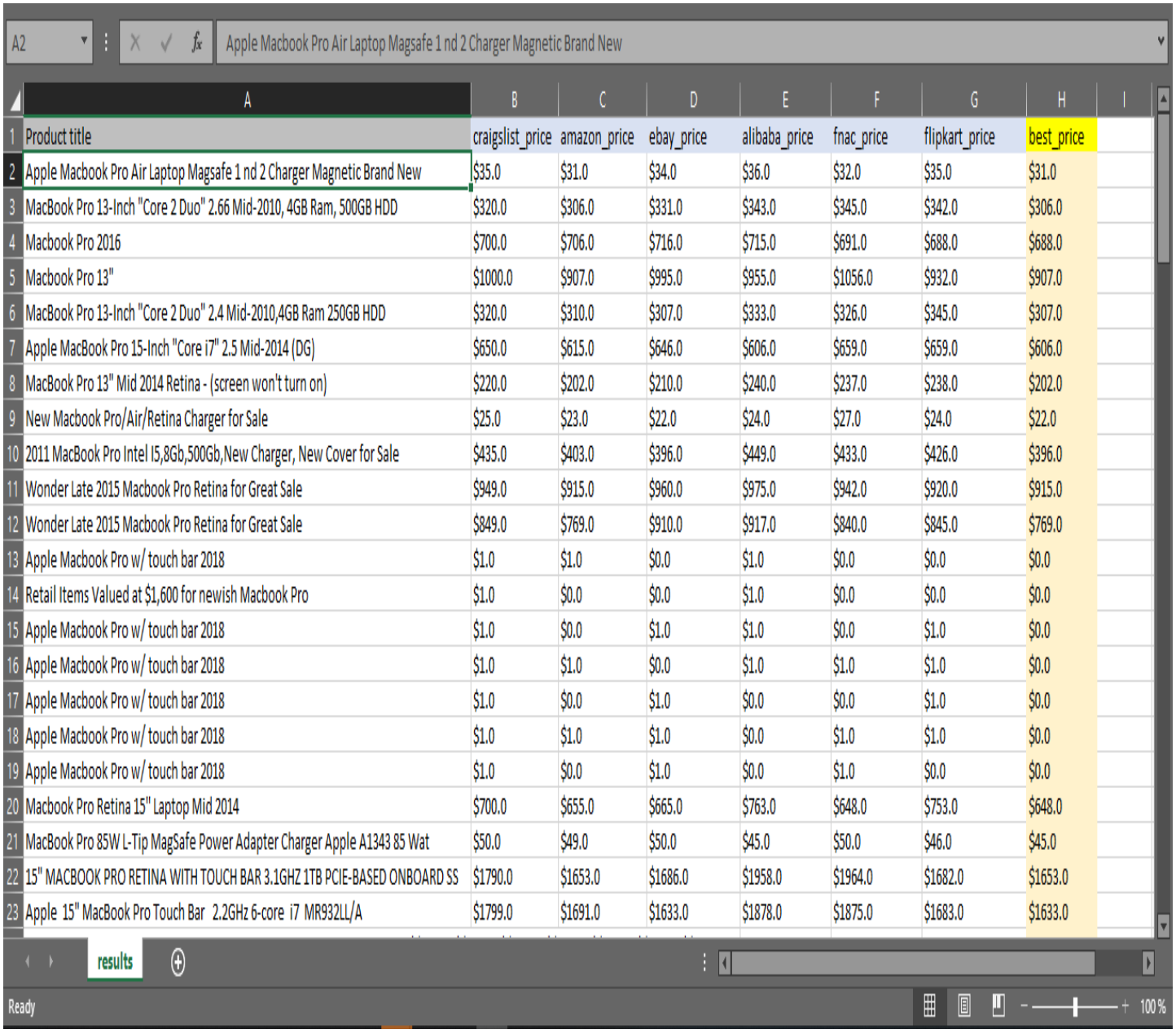
FIGURE 4.18 : interface administrateur vu mobile

8. Résultat

Le résultat est l'aboutissement de la recherche et de notre système d'extraction, effectivement l'application procure à l'administrateur sur son serveur local le résultat de sa recherche sous forme de fichier .CSV

Ce fichier contient tous les produits avec leurs titres, prix (dans tous les sites de e-commerce de l'application) et le meilleur prix pour chaque produit.

Voici ci-dessous un résultat de recherche émit par le système qui est la finalité du web scraping



A2 — Apple Macbook Pro Air Laptop Magsafe 1 nd 2 Charger Magnetic Brand New

	A	B	C	D	E	F	G	H
1	Product title	craigslist_price	amazon_price	ebay_price	alibaba_price	fnac_price	flipkart_price	best_price
2	Apple Macbook Pro Air Laptop Magsafe 1 nd 2 Charger Magnetic Brand New	$35.0	$31.0	$34.0	$36.0	$32.0	$35.0	$31.0
3	MacBook Pro 13-Inch "Core 2 Duo" 2.66 Mid-2010, 4GB Ram, 500GB HDD	$320.0	$306.0	$331.0	$343.0	$345.0	$342.0	$306.0
4	Macbook Pro 2016	$700.0	$706.0	$716.0	$715.0	$691.0	$688.0	$688.0
5	Macbook Pro 13"	$1000.0	$907.0	$995.0	$955.0	$1056.0	$932.0	$907.0
6	MacBook Pro 13-Inch "Core 2 Duo" 2.4 Mid-2010,4GB Ram 250GB HDD	$320.0	$310.0	$307.0	$333.0	$326.0	$345.0	$307.0
7	Apple MacBook Pro 15-Inch "Core i7" 2.5 Mid-2014 (DG)	$650.0	$615.0	$646.0	$606.0	$659.0	$659.0	$606.0
8	MacBook Pro 13" Mid 2014 Retina - (screen won't turn on)	$220.0	$202.0	$210.0	$240.0	$237.0	$238.0	$202.0
9	New Macbook Pro/Air/Retina Charger for Sale	$25.0	$23.0	$22.0	$24.0	$27.0	$24.0	$22.0
10	2011 MacBook Pro Intel I5,8Gb,500Gb,New Charger, New Cover for Sale	$435.0	$403.0	$396.0	$449.0	$433.0	$426.0	$396.0
11	Wonder Late 2015 Macbook Pro Retina for Great Sale	$949.0	$915.0	$960.0	$975.0	$942.0	$920.0	$915.0
12	Wonder Late 2015 Macbook Pro Retina for Great Sale	$849.0	$769.0	$910.0	$917.0	$840.0	$845.0	$769.0
13	Apple Macbook Pro w/ touch bar 2018	$1.0	$1.0	$0.0	$1.0	$0.0	$0.0	$0.0
14	Retail Items Valued at $1,600 for newish Macbook Pro	$1.0	$0.0	$0.0	$1.0	$0.0	$0.0	$0.0
15	Apple Macbook Pro w/ touch bar 2018	$1.0	$0.0	$1.0	$1.0	$0.0	$1.0	$0.0
16	Apple Macbook Pro w/ touch bar 2018	$1.0	$1.0	$0.0	$1.0	$1.0	$1.0	$0.0
17	Apple Macbook Pro w/ touch bar 2018	$1.0	$0.0	$1.0	$0.0	$0.0	$1.0	$0.0
18	Apple Macbook Pro w/ touch bar 2018	$1.0	$1.0	$1.0	$0.0	$1.0	$1.0	$0.0
19	Apple Macbook Pro w/ touch bar 2018	$1.0	$0.0	$1.0	$0.0	$1.0	$0.0	$0.0
20	Macbook Pro Retina 15" Laptop Mid 2014	$700.0	$655.0	$665.0	$763.0	$648.0	$753.0	$648.0
21	MacBook Pro 85W L-Tip MagSafe Power Adapter Charger Apple A1343 85 Wat	$50.0	$49.0	$50.0	$45.0	$50.0	$46.0	$45.0
22	15" MACBOOK PRO RETINA WITH TOUCH BAR 3.1GHZ 1TB PCIE-BASED ONBOARD SS	$1790.0	$1653.0	$1686.0	$1958.0	$1964.0	$1682.0	$1653.0
23	Apple 15" MacBook Pro Touch Bar 2.2GHz 6-core i7 MR932LL/A	$1799.0	$1691.0	$1633.0	$1878.0	$1875.0	$1683.0	$1633.0

results — Ready — 100 %

FIGURE 4.19 : fichier .CSV du résultat de recherche

IV. Conclusion

Dans ce chapitre, nous avons présentés notre application de web scraping ainsi que l'environnement de développement et d'implémentation, en se focalisant sur les techniques et outils de programmation utilisées pour réaliser système qui a pour but d'extraire les prix et les comparer pour les proposer au client.

Enfin, nous avons achevé ce chapitre en faisant une démonstration de notre travail épaulé de quelques interfaces qui montrent la finalité.

Conclusion générale

Nous sommes en train de vivre dans une ère où le web grossit à une vitesse folle et bien entendu, toutes les technologies qui l'entourent évoluent à la même vitesse, effectivement nous l'avons vu lors de notre description de l'évolution des données.

Les données sont aujourd'hui le nerf de la guerre. Les entreprises ont aujourd'hui compris qu'internet peut-être un moyen formidable d'expansion, ça a été par ailleurs notre principale occupation de notre projet et de penser comment extraire et traiter des données spécifiques en utilisant plusieurs techniques.

Le crawling, le scraping, mais surtout le e-commerce et le big data n'en sont qu'au début d'une croissance certaine. Il suffit d'observer le nombre de projet, d'articles et de conférences sur le sujet pour comprendre qu'un phénomène émerge. Le Web n'est, aujourd'hui, certainement pas exploité à son maximum. Cependant, il faut savoir rester prudent car la quantité d'information est telle qu'il serait très facile de s'y perdre. Les futurs outils développés devront faire face à un nombre incalculable de données. On peut d'ores et déjà penser que la mise à disposition de l'information par le biais d'API fournies sera un moyen efficace de gérer cette masse de données. Enfin, un travail est certainement à faire du côté de la réglementation pour clarifier la situation et permettre de se défendre face aux attaques illégales.

Dans le premier chapitre, nous avons abordé les termes Internet et le Web, expliqué la différence Entre eux ainsi que les étapes de développement du Web, nous avons par ailleurs défini le big data et énumérer les différents challenges que fait face le monde informatique afin de subvenir à des besoins futurs, et notamment la monté en charge de la quantité de données.

Nous avons présenté dans le second chapitre des recherches au sujet du web scraping. Ensuite nous avons défini le processus de web scraping et web crawling avec leurs architectures respectives nous avons aussi comparer ces deux techniques. Par ailleurs nous avons cité les différentes techniques et les outils utilisés pour web scraping et enfin nous avons donné des exemples des applications du web scraping

Dans le troisième chapitre, nous avons traité dans un premier temps plusieurs aspects du commerce électronique (le e-commerce), nous avons aussi montré l'importance de ce vaste domaine et nous avons ainsi clarifié la relation entre un e-commerce et le client surtout lors de ses achats ou vente en ligne, effectivement nous avons insisté sur l'application du web scraping et décrit la méthode d'extraction des prix, puis des unités qui composent notre système d'extraction de ces prix pour ainsi faciliter et proposer les meilleurs choix au client.

Conclusion générale

Nous avons présenté dans la seconde partie de ce chapitre notre système d'extraction de prix avec une étude conceptuelle qui est le point de départ notre application web d'extraction de prix baptisé My-Price-Check.

Finalement dans le quatrième et dernier chapitre, nous avons présentés notre application de web scraping ainsi que l'environnement de développement et d'implémentation, en se focalisant sur les techniques et outils de programmation utilisées pour réaliser système qui a pour but d'extraire les prix et les comparer pour les proposer au client.

Enfin, nous avons achevé ce chapitre en faisant une démonstration de notre travail épaulé de quelques interfaces qui montrent la finalité.

- **ARPA** : (Advanced Research Project Agency) est un nom générique du domaine de premier niveau utilisé pour résoudre des problèmes d'infrastructure. « ARPA » est un rétro-acronyme. ARPANET était le prédécesseur d'Internet, construit par la DARPA
- **TCP / IP :** La suite des protocoles Internet est l'ensemble des protocoles utilisés pour le transfert des données sur Internet. Elle est aussi appelée suite TCP/IP, DoD Standard ou bien DoD Model ou encore DoD TCP/IP ou US DoD Model. Elle est souvent appelée TCP/IP, d'après le nom de ses deux premiers protocoles : TCP et IP.
- **FTP** (File Transfer Protocol) : est un protocole de communication destiné au partage de fichiers sur un réseau TCP/IP. Il permet, depuis un ordinateur, de copier des fichiers vers un autre ordinateur du réseau, ou encore de supprimer ou de modifier des fichiers sur cet ordinateur
- **UDP :** (User Datagram Protocol) User Datagram Protocol est un des principaux protocoles de télécommunication utilisés par Internet. Il fait partie de la couche transport du modèle OSI
- **DNS :** Domain Name System, généralement abrégé DNS, qu'on peut traduire en « système de noms de domaine », est le service informatique distribué utilisé pour traduire les noms de domaine Internet en adresse IP
- **HTTP** (HyperText Transfer Protocol) : Hypertext Transfer Protocol est un protocole de communication client-serveur développé pour le World Wide Web. HTTPS est la variante sécurisée par l'usage des protocoles Transport Layer Security. HTTP est un protocole de la couche application
- **CERN :** L'Organisation européenne pour la recherche nucléaire
- **WWW** : World Wide Web
- **ACM :** l'Association for Computing Machinery (ou ACM)
- **HDFS** : Hadoop Distributed File System (HDFS)
- **Hadoop** : est un framework libre et open source écrit en Java destiné à faciliter la création d'applications distribuées
- **W3C** : Le World Wide Web Consortium, abrégé par le sigle W3C, est un organisme de standardisation à but non lucratif, fondé en octobre 1994 chargé de promouvoir la compatibilité des technologies du World Wide Web telles que HTML5, HTML, XHTML, XML, RDF

- **IDC** : International Data Conseil, un cabinet de conseil dans le marketing et les services
- **XPath** : est un langage de requête pour localiser une portion d'un document XML.
- **XML** : L'Extensible Markup Language, généralement appelé XML, « langage de balisage extensible
- **DOM** : Le Document Object Model est une interface de programmation normalisée par le W3C, qui permet à des scripts d'examiner et de modifier le contenu du navigateur web
- **API** : une interface de programmation d'application ou interface de programmation applicative est un ensemble normalisé de classes, de méthodes, de fonctions et de constantes qui sert de façade par laquelle un logiciel offre des services à d'autres logiciels
- **CSV** : Comma-separated values, connu sous le sigle CSV, est un format texte ouvert représentant des données tabulaires sous forme de valeurs séparées par des virgules.
- **JSON** : JavaScript Object Notation est un format de données textuelles dérivé de la notation des objets du langage JavaScript. Il permet de représenter de l'information structurée
- **Url** : est une chaîne de caractères uniforme qui permet d'identifier une ressource du World Wide Web par son emplacement
- **Bot** : est un agent logiciel automatique ou semi-automatique qui interagit avec des serveurs informatiques. Un bot se connecte et interagit avec le serveur comme un programme client utilisé par un humain, d'où le terme « bot », qui est la contraction par aphérèse de « robot ».
- **IP** : Une adresse IP est un numéro d'identification qui est attribué de façon permanente ou provisoire à chaque périphérique relié à un réseau informatique qui utilise l'Internet Protocol
- **AJAX :** Ajax est une architecture informatique qui permet de construire des applications Web et des sites web dynamiques interactifs sur le poste client
- **Scheduler :** (ordonnanceur en FR) Dans les systèmes d'exploitation, l'ordonnanceur désigne le composant du noyau du système d'exploitation choisissant l'ordre d'exécution des processus sur les processeurs d'un ordinateur.

- **L'inbound marketing** : est une stratégie marketing visant à faire venir le client à soi plutôt que d'aller le chercher avec les techniques de marketing traditionnelles de type outbound marketing

- **CEO :** L'optimisation pour les moteurs de recherche, appelé également référencement naturel ou organique en français, inclut l'ensemble des techniques qui visent à améliorer le positionnement d'une page, d'un site ou d'une application web dans la page des résultats de recherche d'un moteur de recherche
- **TLS :** La Transport Layer Security ou « Sécurité de la couche de transport », et son prédécesseur la Secure Sockets Layer ou « Couche de sockets sécurisée », sont des protocoles de sécurisation des échanges par réseau informatique, notamment par Internet.
- **HEFL :** est un outil complet qui va vous permettre de tirer le meilleur parti des technologies BPM. Notre logiciel de modélisation des processus métiers vous permet d'analyser, concevoir, documenter, optimiser et automatiser vos processus dans un même outil.
- **BPMN :** Business Process Model and Notation, c'est-à-dire « modèle de procédé d'affaire et notation », est une méthode de modélisation de processus d'affaires pour décrire les chaînes de valeur et les activités métier d'une organisation sous forme d'une représentation graphique
- **StarUML** : StarUML est un logiciel de modélisation UML
- **Requests** : est une bibliothèque HTTP Python, publiée sous la licence Apache 2.0. L'objectif du projet est de rendre les requêtes HTTP plus simples et plus conviviales
- **ISO :** L'Organisation internationale de normalisation, généralement désigné sous son sigle : ISO, choisi pour être le sigle identique dans toutes les langues
- **Perl, Ruby, Scheme, Smalltalk et Tcl :** des langages de programmation orientées objets
- **.NET :** Framework est un cadriciel pouvant être utilisé par un système d'exploitation Microsoft Windows et Microsoft Windows
- **iOS :** anciennement iPhone OS le « i » de iOS étant pour iPhone d'où la minuscule, est le système d'exploitation mobile développé par Apple pour plusieurs de ses appareils
- **Linux** : est, au sens restreint, le noyau Linux, et au sens large, tout système d'exploitation fondé sur le noyau Linux. Cet article couvre le sens large. Linux est un

système d'exploitation de type Unix. Le noyau Linux a été créé en 1991 par Linus Torvalds. C'est un logiciel libre

- **macOS** : est un système d'exploitation partiellement propriétaire développé et commercialisé par Apple depuis 1998, dont la version la plus récente est macOS Catalina lancée le 7 octobre 2019. Il fait partie des systèmes d'exploitation d'Apple
- **SGBD** : (Système de Gestion de Bases de Données) est un logiciel qui stocke des données de façon organisées et cohérentes. Un SGBDR (Système de Gestion de Bases de Données Relationnelles) est le type particulier de SGBD
- **BDD** : Une base de données, permet de stocker et de retrouver des données brutes ou de l'information
- **MySQL** : est un système de gestion de bases de données relationnelles. Il est distribué sous une double licence GPL et propriétaire
- **PaaS** : Platform as a service, ou Plate-forme en tant que service, est l'un des types de cloud computing, principalement destiné aux développeurs ou aux entreprises de développement, où : l'entité cliente maintient les applications proprement dites
- **Git** : est un logiciel de gestion de versions décentralisé. C'est un logiciel libre créé par Linus Torvalds, auteur du noyau Linux, et distribué selon les termes de la licence publique générale GNU

Références

[1] https://histoire-internet.vincaria.net/p/documents- historiques.html

[2] https://vbdl.wordpress.com/2018/07/14/407/

[3] https://www.loukam.net/TECHNOLOGIE_Services_Web_Chapitre1.pdf

[4] Nasreddine Bouhaï, Imad Saleh : Internet des objets : Evolutions et innovations

[5] Karan Patel: Incremental Journey for World Wide Web: Introduced with Web 1.0 to Recent Web 5.0 – A Survey Paper - October 2013.

[6] https://www.lebigdata.fr/definition-big-data.

[7] https://fr.blog.businessdecision.com/bigdata/2017/07/big-data-potentiel-archives-entreprises/

[8] ZHENG Guojun, JIA Wenchao, SHI Jihui, SHI Fan, ZHU Hao, LIU Jiang: Design and Application of Intelligent Dynamic Crawler for Web Data Mining- (2017) [https://libgen.lc/scimag/ads.php?doi=10.1109/YAC.2017.7967575]

[9] Ram Sharan Chaulagain, Santosh Pandey, Sadhu Ram Basnet, Subarna Shakya: Cloud Based Web Scraping for Big Data Applications - November 2017 [https://libgen.lc/scimag/ads.php?doi=10.1109%2FSmartCloud.2017.28&downloadname=]

[10] Rajapriya: A literature survey on web crawlers mai 2014

[11] Bruce Krulwich: Information integration agents: BargainFinder and NewsFinder - janvier 1996

[12] Ganesh Iyer et Amit Pazgal: Internet Shopping Agents: Virtual Co-Location and CompetitionSeptember,2001 [http://faculty.haas.berkeley.edu/giyer/index_files/agents.pdf]

[13] Vemula Satyanarayana, Rahul Kumar Behera, Gaurav Kumar: International Journal of Engineering & Technology: Online Price Comparator and Reselling Website - (2018)

[14] Mawloud Mosbah. Mesures de Distance dans le Contexte de la Recherche d'Images par le Contenu (CBIR). Recherche d'information [cs.IR]. Université 20 août 1955 Skikda (Algérie), 2017. Français. fftel-02948637f

[15] https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1203499-crawler-definition-traduction-et-acteurs/.

[16] Focused Web Crawler with Page Change Detection Policy Swati Mali, VJTI, Mumbai B.B. Meshram VJTI, Mumbai

[17] How to Build a Web Scraper for Social Media: Oskar Lloyd, Christoffer Nilsson - juin 2019

[18] Habib Ullah, Zahid Ullah, Shahid Maqsood, and Abdul Hafeez: Web Scraper Revealing Trends of Target Products and New Insights in Online Shopping Websites- Janvier 2018.

[19] Daniel Glez-Pen‹a, Analia Lourenco, Hugo Lopez-Fernandez, Miguel Reboiro-Jato et Florentino Fdez-Riverola: Web scraping technologies in an API world - Mars 2013.

Références

[20] Vlad Krotov, Leiser Silva: Legality and Ethics of Web Scraping -September 2018.

[21] https://www.antevenio.com/fr/quest-ce-que-le-web-scraping-et-a-quoi-sert-il/

[22] L. Grechanik. The Complete Beginner's Guide to web Scraping, 05/2020.

[23] http://prowebscraping.com/web-scraping-vs-web-crawling

[24] S Sirisuriya: AComparative Study on Web Scraping - November2015

[25] Choosing Scrapy -Daniel Myers -James W. McGuffee-October2015. [https://www.researchgate.net/publication/314179276_Choosing_Scrapy]

[26] https://www.scraperapi.com/blog/the-10-best-web-scraping-tools/

[27] Leonard Richardson : Beautiful Soup Documentation- Décembre 2019 [https://www.crummy.com/software/BeautifulSoup/bs4/doc/.]

[28] Anand V. Saurkar, Kedar G. Pathare, Shweta A. Gode : An Overview On Web Scraping Techniques And Tools- avril 2018

[29]- https://celadonsoft.com/ai-ml/complete-beginners-guide-to-web-scraping

[30] A Brief History of Web Crawlers Seyed M. Mirtaheri, Mustafa Emre Dinc¸t¨urk, Salman Hooshmand, Gregor V. Bochmann, Guy-Vincent JourdanSchool of Electrical Engineering and Computer ScienceUniversity of OttawaOttawa, Ontario, Canada

[31] https://www.doofinder.com/fr/blog/qu-est-ce-que-le-ecommerce

[32] Alain Laidet, E-commerce Paris : 30000 visiteurs autour de l'innovation, Classe Export, no 191, septembre 2010, page 40.

[33] https://booweb.eu/e-commerce/

[34] https://www.minderest.com/fr/blog/2019/06/12/qu%E2%80%99est-ce-que-le-price-scraping-utilisation-et-strat%C3%A9gie-pour-votre-e-commerce

[35] Patrice Briol, BPMN, the Business Process Modeling Notation Pocket Handbook. LuLu.com, 2008. (ISBN 978-1-4092-0299-8)

[36] Grady Booch, James Rumbaugh, Ivar Jacobson, Le guide de l'utilisateur UML, 2000 (ISBN 2-212-09103-6)

[37] Franck Barbier, UML 2 et MDE [archive], Ingénierie des modèles avec études de cas, 2009 (ISBN 978-2-10-049526-9)

[38] Laurent Audibert, UML 2, De l'apprentissage à la pratique (cours et exercices) [archive], Ellipses, 2009 (ISBN 978-2729852696)

[39] Doug Hellmann, *The Python Standard Library by Example*, Addison-Wesley, 2017

[40] https://www.w3.org/standards/webdesign/htmlcss

[41] https://www.futura-sciences.com/tech/definitions/internet-css-4050/

Références

[42]https://developer.mozilla.org/fr/docs/Learn/JavaScript/First_steps/What_is_Javascript

[43] *The Art of SQL* - Stéphane Faroult - O'Reilly, 2006

[44] http://dictionnaire.sensagent.leparisien.fr/Django%20(framework)/fr-fr/

[45] https://openclassrooms.com/fr/courses/4425076-decouvrez-le-framework-django/4631014-decouvrez-larchitecture-mvt

[46] https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1203607-sqlite-definition/

[47] https://www.heroku.com/

[48] http://dictionnaire.sensagent.leparisien.fr/VISUAL%20STUDIO/fr-fr/