Base de données multimédia

Ce projet consiste à récupérer les données d'un site de vente en ligne et les organiser dans une base de données.
On va récupérer les données du site leboncoin (LBC). Des utilisateurs français postent des annonces en ligne pour vendre des objets ou des services. Il y a une énorme diversité d'objets en vente qui sont catégorisés :

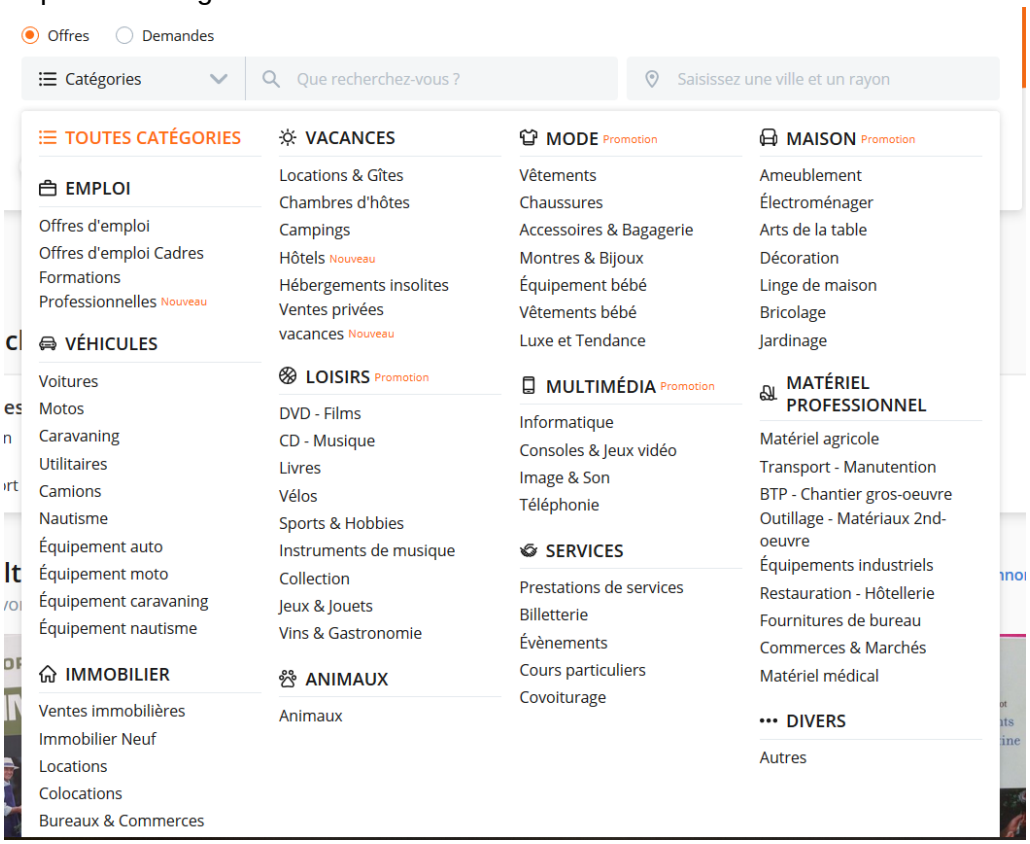


Figure 1 : catégorie des offres en ligne

Comme on peut le voir, le site peut proposer de poster des offres (objets ou services contre de l'argent) ou de poster des demandes (les gens recherches des offres ou services).
Le jour de la réalisation de ce rapport, 33 751 530 annonces étaient en ligne.

LBC est une base de données protégée, producteur et exploitant de sa propre base [1]. Il n'est donc pas possible d'extraire les données de façon simple. Pour cela, il va falloir interroger les pages internet du site LBC et récupérer les données automatiquement. LBC a une architecture basée sur 500 serveurs [2]. Au début du site, LBC utilisait MySQL, puis avec la montée en charge, LBC a migré sur PostGreSQL [2]. Pour l'analyse des données, LBC utilise Amazon Cloud et développe des outils internes à base de solutions open source.

1. Objectif

Ce projet va s'inscrire dans un projet plus global de Big data. Il s'agit de la première brique qui consiste à récupérer, traiter et organiser les données affichées sur chaque annonce en ligne.

L’accès à ces données via une base de données permettra ainsi d’interroger la BDD et d’en tirer de la valeur (statistiques, Machine Learning, savoirs, etc).
Ainsi sur la page internet en annexe, dont voici la capture d’écran (une partie) :

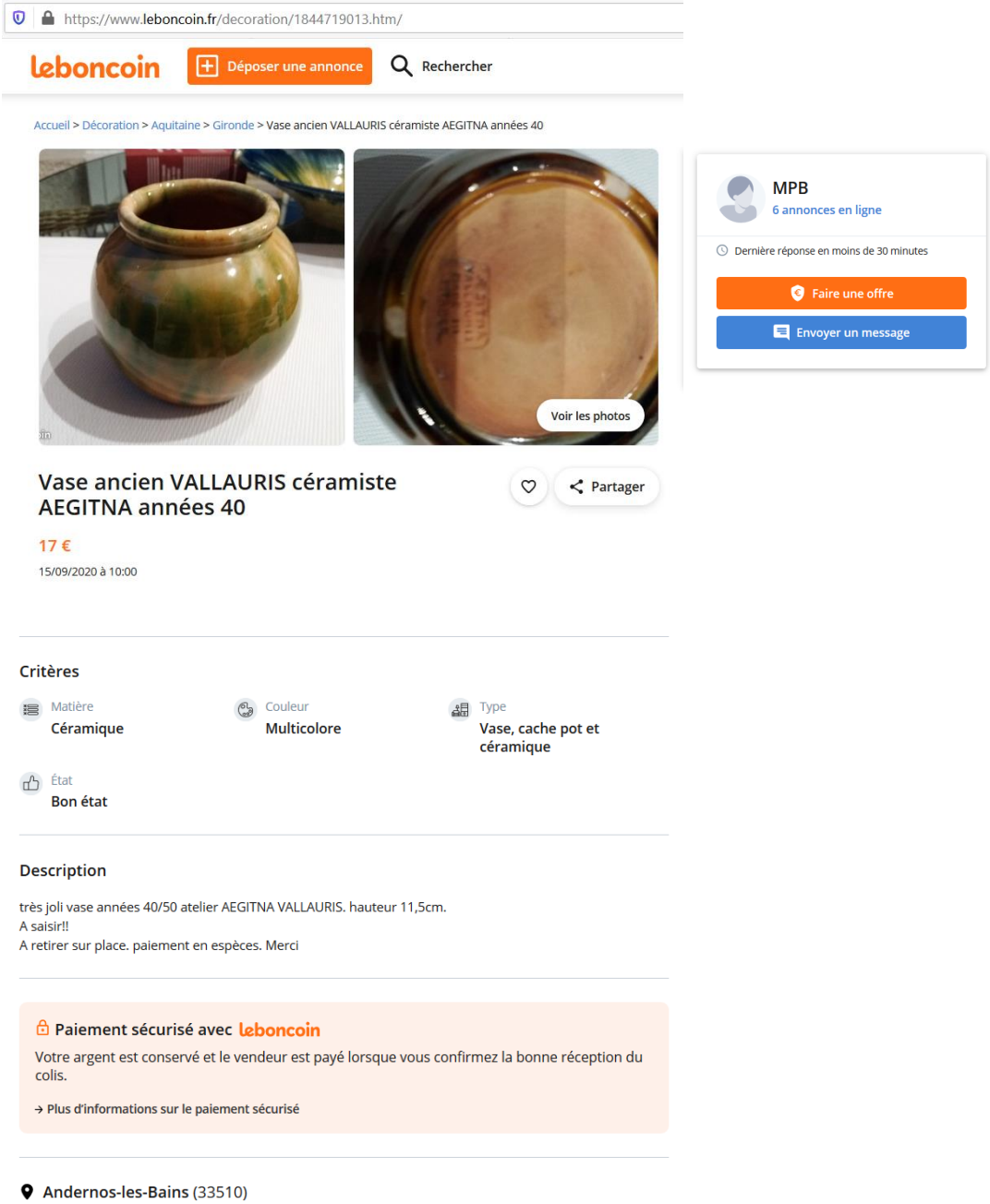


Figure 2 : Exemple d’une annonce LBC

Sur cette page, on récuperera les données suivantes :

- URL ;
- Catégorie (Décoration) ;
- Région et département (Aquitaine et Gironde) ;
- Titre ;
- Prix ;
- Date de publication ;
- Critères (différents selon les catégories) ;
- Description ;
- Code Postal ;

- Identifiant du vendeur ;
- Nombre d'annonces du vendeur.

L'objectif de cette brique est donc de mettre à disposition dans une base de données toutes les annonces accessibles en temps réel. Cette BDD sera enrichie tous les jours avec les nouvelles annonces et mise à jour avec un suivi des annonces.

2. Enjeux

Cette brique est essentielle pour comprendre les comportements d'achat et la réalisation de statistiques (prix de vente, catégories en forte demande, etc). En effet, pour acquérir du savoir dans ce domaine, il est nécessaire de le tirer de données et donc d'accéder à cette donnée. Il est évident que LBC est propriétaire de ces données et ne les donne pas gratuitement. Il est donc obligatoire de récupérer ces données. La réalisation de cette brique nous permet donc de gagner de la donnée, qui est aujourd'hui l'or numérique.

La donnée permet en effet de créer de la valeur [3]. L'enjeu est donc le gain en données pour créer de la valeur (voir point 3 – Usages).

D'un point de vue plus humain, on y gagne aussi en compétences personnelles :

- Gestion de projet :
    - Méthode Agile ;
    - Travail en équipe ;
    - Pouvoir mener et terminer un projet en le définissant et en présentant un produit fini.
- Compétences techniques :
    - Langage de programmation orienté objet (ex : Python)
    - Logiciel ETL - Extract, Transform, Load (ex : Talend Open Studio)
    - Base de données noSQL (ex : MongoDB)
    - Machine Learning (ex : Python et librairies scikit-learn, TensorFlow)
    - Statistiques (ex : Python et librairies pandas, numpy, maths, etc ou R).

3. Usages

Cette brique s'inscrirait dans un projet plus global qui souhaite utiliser les données recueillies pour monter un modèle de Machine Learning (IA). Cependant, ces données sont un puits de valeur immense qui permettent de répondre à de petites interrogations comme à de plus grandes questions.

Pour les petites interrogations (statistiques), il suffira d'interroger la BDD via MongoDB :

- Durée de vie des annonces ;
    - Cette question peut être déclinée selon plusieurs critères :
        - Catégorie ;
        - Géographie (région, département, ville) ;
        - Utilisateur ;
        - Prix ;
        - Qualité de l'annonce (* voir la partie sur les questions de ML).
- Domaines et objets les plus vendus ou les plus demandés ;
    - Il y a une valeur à tirer (investissements).
- Définir l'intérêt des photos et des descriptions (textes) pour la vente des objets.

- Définir les comportements de vente :
    - Nombre d'annonces par utilisateur
    - Diversité des objets vendus ;
    - Prix des objets (Objets plus ou moins cher que la moyenne).
- Tirer de la valeur sur la géographie :
    - Prix selon la géographie ;
    - Nombre d'annonces selon la géographie ;
    - Catégories selon la géographie.

Il y a d'autres très nombreuses questions statistiques à se poser, qui peuvent mener à des questions plus globales et complexes.

Ainsi pour les interrogations plus complexes, cela demandera de développer un modèle de ML :

- Définir la qualité d'une annonce :
    - Traitement du texte – NLP Natural Language Processing sur le texte de description de l'annonce ;
    - Traitement d'image – Computer Vision sur les images postées.
- Définir l'intérêt d'une annonce.
- Repérer la fraude :
    - Prix ?
    - Photo « volées ».
    - Profil utilisateur suspect

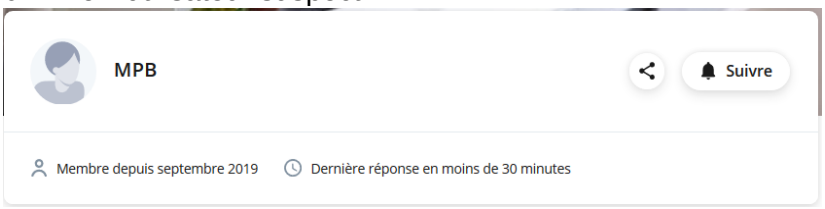


Figure 3 : Exemple d'un profil utilisateur
On peut récupérer le nom, l'identifiant (via l'URL), la date d'inscription, le nombre d'annonces).

4. Fonctions

Pour cette brique, il est nécessaire :

- D'automatiser le processus de récupération des données (en continu) pour alimenter notre BDD (scraping) ;
- De nettoyer la donnée ;
- D'organiser la donnée (dictionnaire json) ;
- De charger la donnée dans la base de donnée ;
- De mettre à jour la BDD en interrogeant les URL enregistrées pour constater la durée de vie des annonces (on enregistrera 2 champs : booléen encore_en_ligne et date date_derniere_consultation) ;
- D'interroger la BDD.

5. Dataset

Le dataset existe chez le site qui possède les données. Cependant, elles sont protégées et ne peuvent être récupérées sous forme d'un dictionnaire. Il est donc nécessaire de le constituer en utilisant des outils de scraping et de récupération automatique de données.

6. Scalabilité

N'envisageant pas d'utiliser le Cloud, on souhaite entreposer les données sur le serveur que le CNAM met à disposition (5,23 To de stockage). La scalabilité est une question centrale dans les projets de data science [5]. Ici elle se pose surtout pour les capacités de stockage puisque la récupération de toutes ces données demandera toujours plus de capacité de stockage.

En partant du principe qu'on enregistre environ 3 images pour chaque annonce (environ 50ko par image) et environ 15 champs texte (listés dans la partie 1 Objectif) d'une taille de 100 caractères en moyenne : 1 octet (par caractère) [6] * 100 * 15 = 1500 octets soit 1,5 ko. On retient donc 152 ko par annonce. Environ 800 000 nouvelles annonces étaient publiées chaque jour en 2016 [2]. On a donc besoin de 800 000 * 152 ko = 1,2 Go par jour. A ce rythme, il faut donc imaginer un système capable d'augmenter régulièrement les capacités de stockage ou un nettoyage/réduction des données stockées. En effet, est-il utile de conserver des données datant de plusieurs années quand les comportements changent très rapidement ?

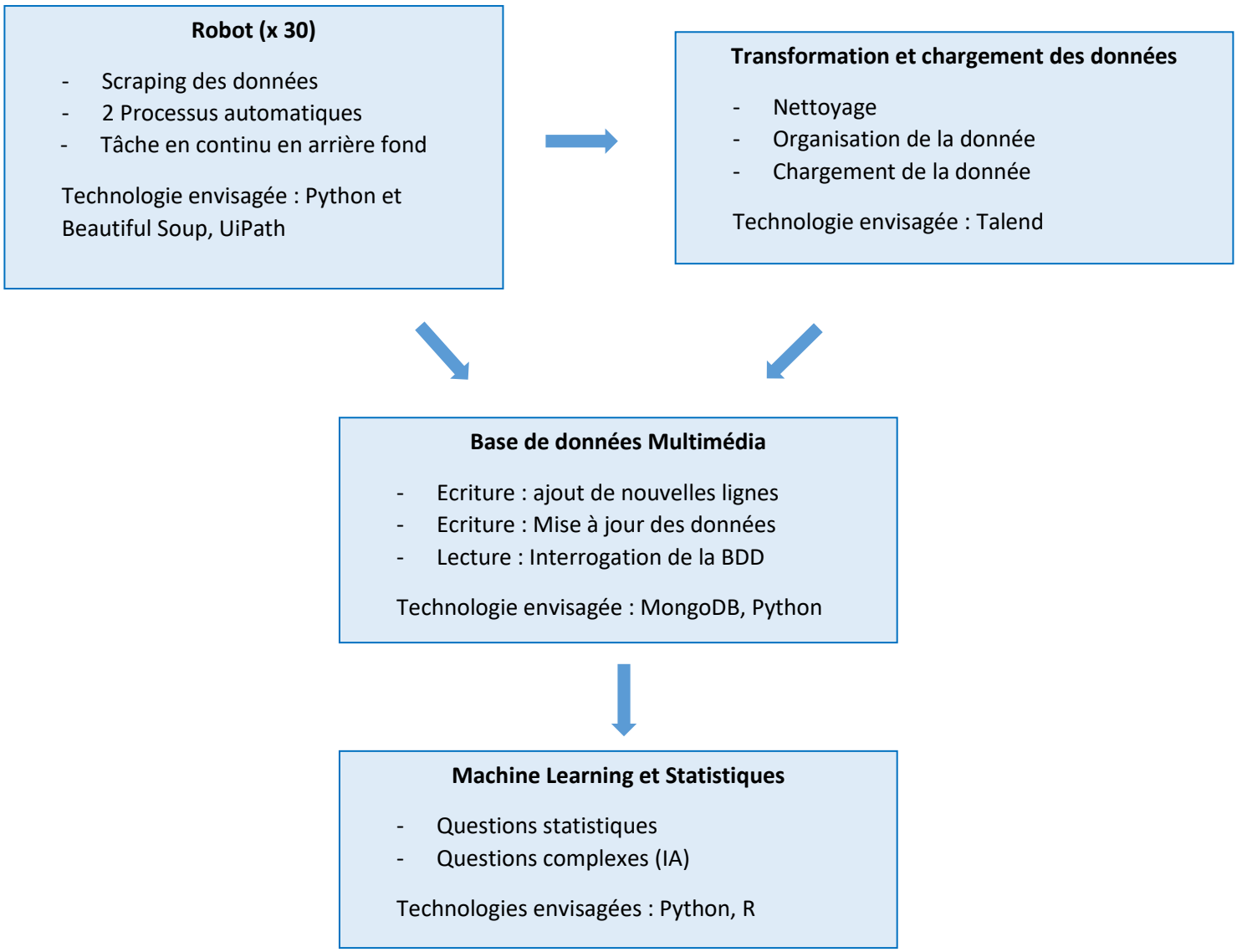
La deuxième interrogation est la capacité à récupérer toutes ces données. En effet, il faut charger chaque jour, 800 000 pages URL pour récupérer la donnée et charger les anciennes pages URL déjà récupérées pour suivre le temps de durée de vie de l'annonce.
En partant du principe qu'il faut 1 seconde pour charger 1 page internet et récupérer la donnée et qu'il y a 86 400 secondes par jour, il faut décliner la récupération de données sur plusieurs machines (au moins 10, sûrement 30). Avec 30 machines, 10 permettent de récupérer les données des nouvelles annonces, les 20 autres permettent de mettre à jour la BDD avec le suivi des annonces.
Pour éviter les problèmes de conflit entre différentes annonces, on propose de créer des robots qui récupèrent la donnée selon la catégorie des annonces :

- Véhicules
- Immobilier
- Vacances
- Loisirs
- Animaux et divers
- Mode
- Multimédia
- Services
- Maison
- Matériel Professionnel

Il est possible de pluger davantage de machines (avec des robots) pour récupérer la donnée. Les conflits d'écriture dans la BDD seront évités grâce à l'URL unique des annonces.

7. Architecture

Ce projet utilisera le serveur du CNAM et sera réalisé comme un POC (Proof of Concept). Les services Cloud ne seront pas utilisés car payants. Cependant, des recherches pourront être réalisées pour étudier la possibilité de l'utiliser. Car un serveur a un prix (ici pris en charge par le CNAM). Il sera intéressant de comprendre le fonctionnement du Cloud pour évaluer l'intérêt de l'utiliser (scalabilité gérée par la Cloud). Une courte étude (tarif, gestion, etc) sera réalisée.



Pour les démonstrations, chaque brique sera montrée :

- Scraping des données : On montrera le robot en action sur 3 annonces en montrant le résultat (données obtenues par le robot).
- Transformation des données : On montrera les modules Talend, en montrant le résultat en sortie
- BDD : On montrera comment les données sont organisées et le résultat de quelques requêtes (comment le nombre de vendeurs, les statistiques, etc).

8. Planning

Voici les tâches à réaliser et le temps estimé :

- Installation infrastructure (logiciel) : 2 jours ou 4 demi-journées ;
- Scraping des données : 1 jour ou 2 demi-journées ;
- Transformation et chargement des données : 1 semaine ou 10 demi-journées ;
- BDD MongoDB : 1 semaine ou 10 demi-journées.
- Interrogation BDD : ½ semaine ou 5 demi-journées.

Notes :

- Les catégories d'emploi ne feront pas partie de cette étude.
- Unicité des annonces grâce à l'URL – éviter les doublons entre nouvelles annonces et anciennes annonces grâce aux dates.

[1] https://solwos.com/site-de-petites-annonces-leboncoin-base-de-donnees-protegee/
[2] https://www.base-de-donnees.com/base-de-donnees-lbc/
[3] https://medium.com/databook/des-donn%C3%A9es-%C3%A0-la-cr%C3%A9ation-de-valeur-a5a3f344c845
[4] https://comarketing-news.fr/la-valeur-reelle-de-nos-donnees-personnelles/
[5] https://itsocial.fr/enjeux-it/enjeux-infrastructure/datacenter/quest-scalabilite-5-meilleurs-articles-autour-de-scalabilite/
[6] https://pageperso.lis-lab.fr/stefano.facchini/python/slides8.pdf