Web Crawling & Web Scraping pour la veille concurrentielle augmentée :
Etude de cas sur les E-commerces

[1]Edouard Ngor SARR, [2]Oumar DIAGNE and All
[1,2] University Assane SECK de Ziguinchor - UASZ
SENEGAL

20-21 Février 2025

Sommaire

- **Contexte**
- **Problématique**
- **Définirons**
  - Web Scraping
  - Web Crawling
  - Veille concurentielle
  - Veille concurentielle augmentée
- **Etude de cas**
  - Architecture
  - Plateforme
  - Resultats

Contexte

- La digitalisation persistante des activités commerciales : **E-commerce**
- **Conséquences :**
  - La mondialisation et la démocratisation des marché
  - L'augmentation exponentielle des prestataires
  - L'augmentation exponentielle des utilisateurs : Les E-clients
  - Accumulation progressive de forte quantités des données collectés : Transactionnelles ou utilisateurs
  - Une rude acticité concurentielle
- Pour les entreprises, exister dans ces situations nécessite
  - Une reforme progressive des processus métiers
  - Une autre vision de la Data : Depuis sa collecte jusqu'à son interpretation
  - **<u>Une bonne connaissance de la concurences</u>**
    - Essentielle pour rester compétitives
    - *Avoir une bonne écoute du marché*

Problématique

- **Comment le web scraping et le web crawling peuvent-ils mener à une veille concurrentielle augmentée ?**
- Cette problématique soulève plusieurs axes de réflexion :
  - **Efficacité et l'efficience**
    - Quels outils et techniques permettent d'automatiser et d'optimiser la fouille et la collecte de données sur la e-commerce concurrence ?
  - **Défis techniques**
    - Comment contourner les protections mises en place par les sites web (CAPTCHA, antibots) et gérer les volumes de données extraites ?
    - S'adapter aux sités de plus en plus en Front End avec le Javascript
  - **Qualité, fiabilité et Utilisation des données**
    - Comment transformer les données brutes en informations exploitables pour une prise de décision stratégique ?
  - **Éthique et légalité**
    - Quelles sont les limites légales et déontologiques du scraping dans le cadre de la veille stratégique ? (RGPD, conditions d'utilisation des sites web, etc.)

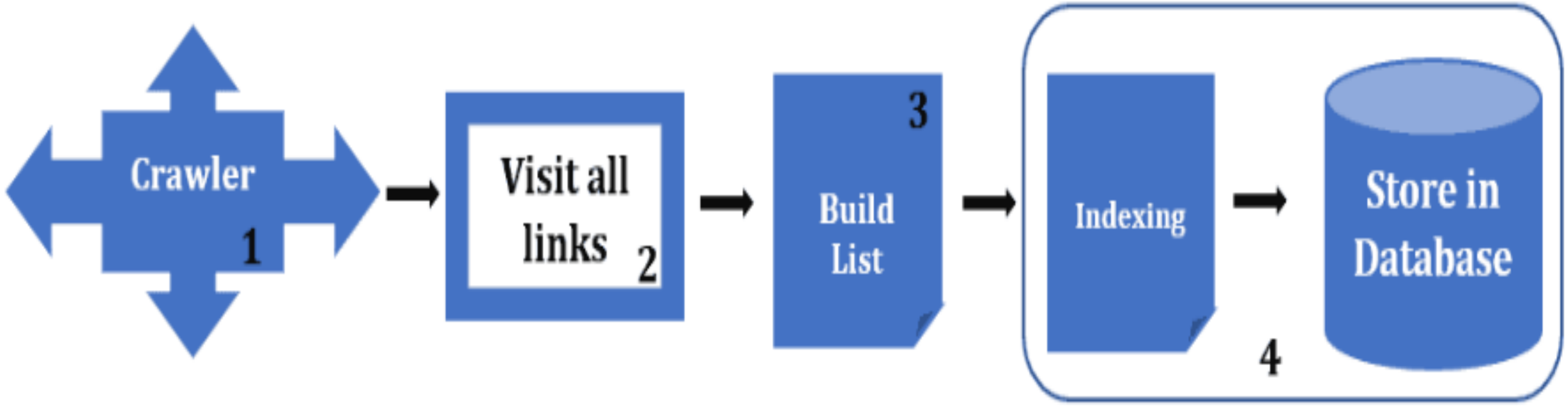
La veille concurentielle

- **Directe et Indirecte**
  - **Veille Concurrentielle Directe :** cible les concurrents offrant des produits et services similaires.
  - **Intelligence Concurrentielle Indirecte :** cible les concurrents offrant des produits ou services qui répondent à des besoins spécifiques ou à des clients similaires.
  - Mixte
- **Active et Passive**
  - Veille Concurrentielle Passive : Implique l'observation et la surveillance des activités et des performances des concurrents sans interagir directement avec eux.
  - Veille Concurrentielle Active : Vise à obtenir des informations spécifiques pour produire des connaissances et guider des actions concrètes.
    - Elle implique une interaction directe avec les concurrents ou leurs ressources en ligne.
  - Mixte

Web Crawling

- Agent logiciel ou bot un programme informatique permettant d'explorer le web, d'analyser sa structure puis de stocker son contenu de manière organisée dans un index.

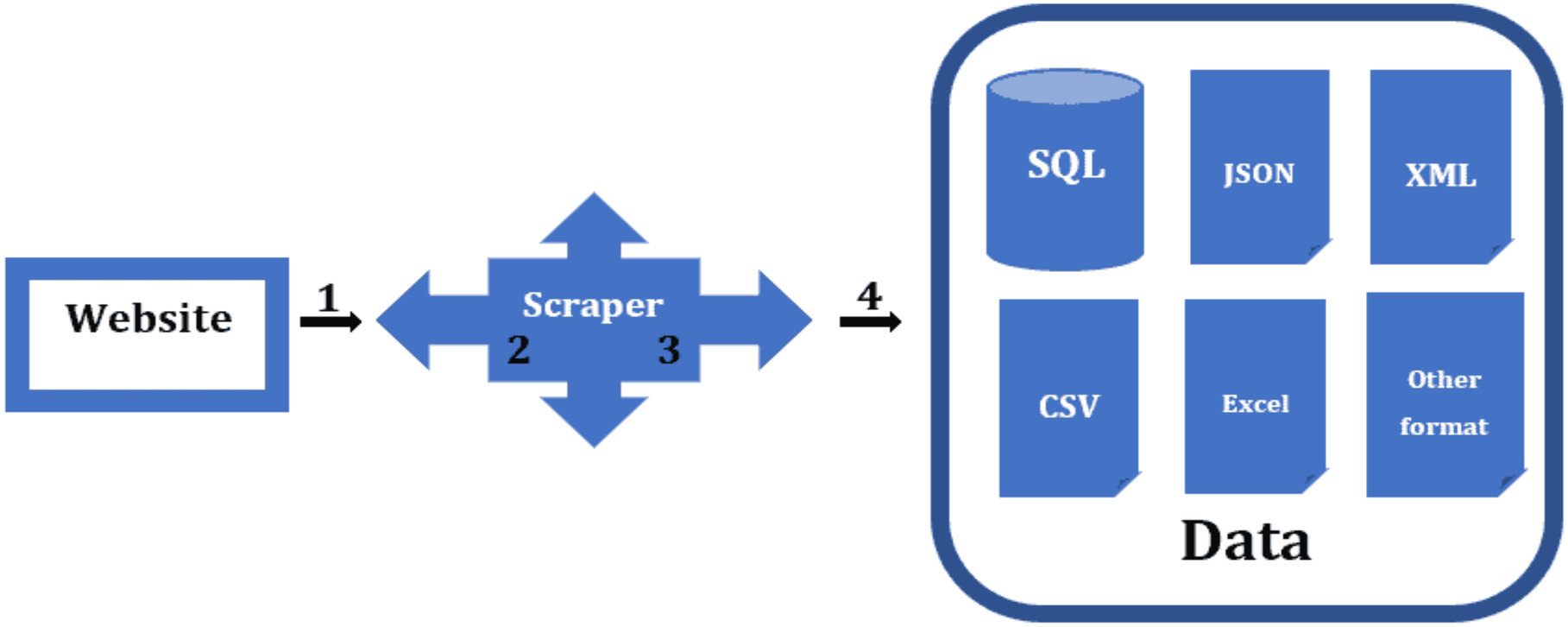
RI = Rechercher + Filtrer + Indexer / Données



Web Crawling

- Processus de récupération, de structuration et de stockage de données depuis les pages web. Objectif : Constituer un corpus de données propre prêt à l'emplois ;

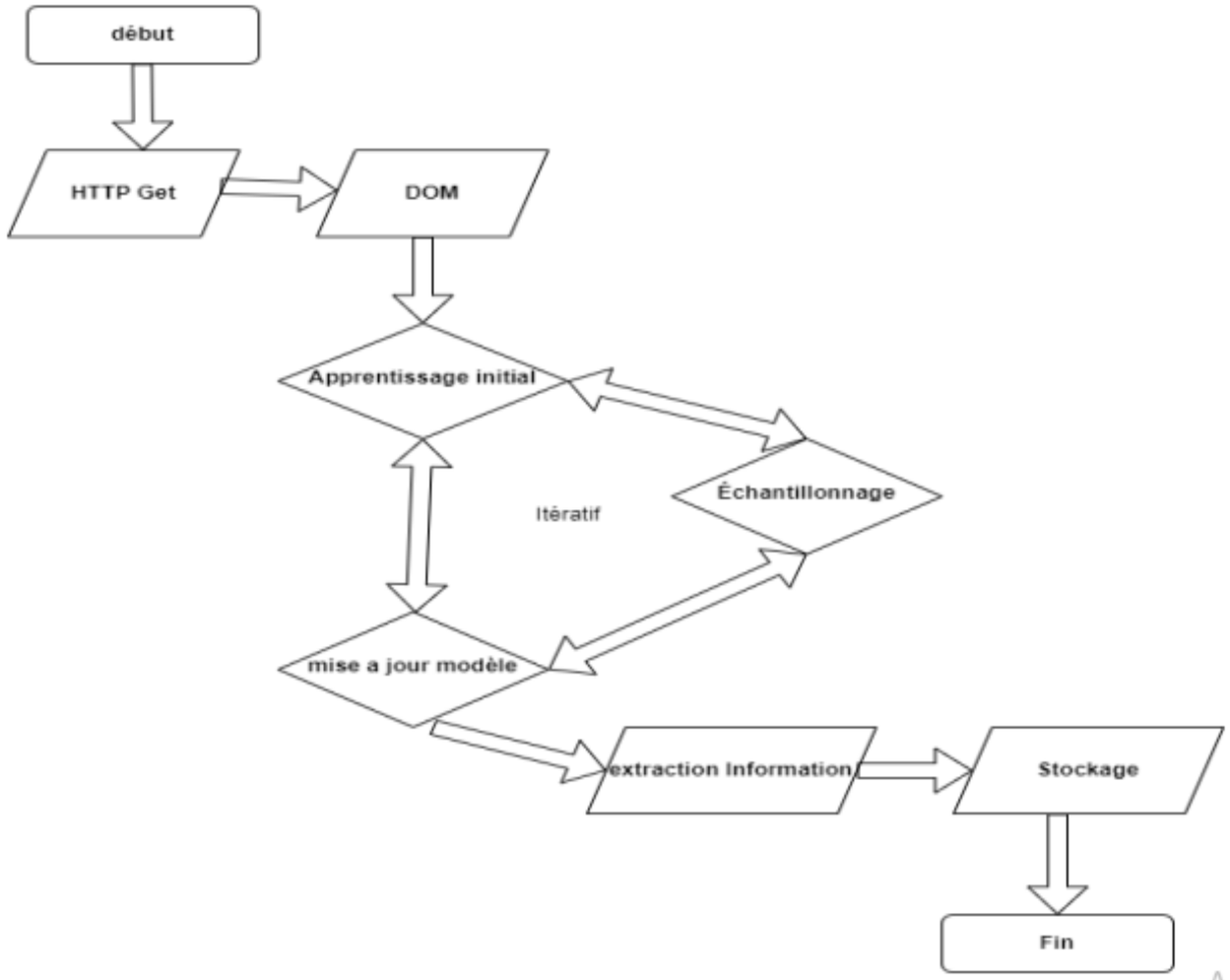
Acquisition de données = RI + collecter + Structurer + Stoker



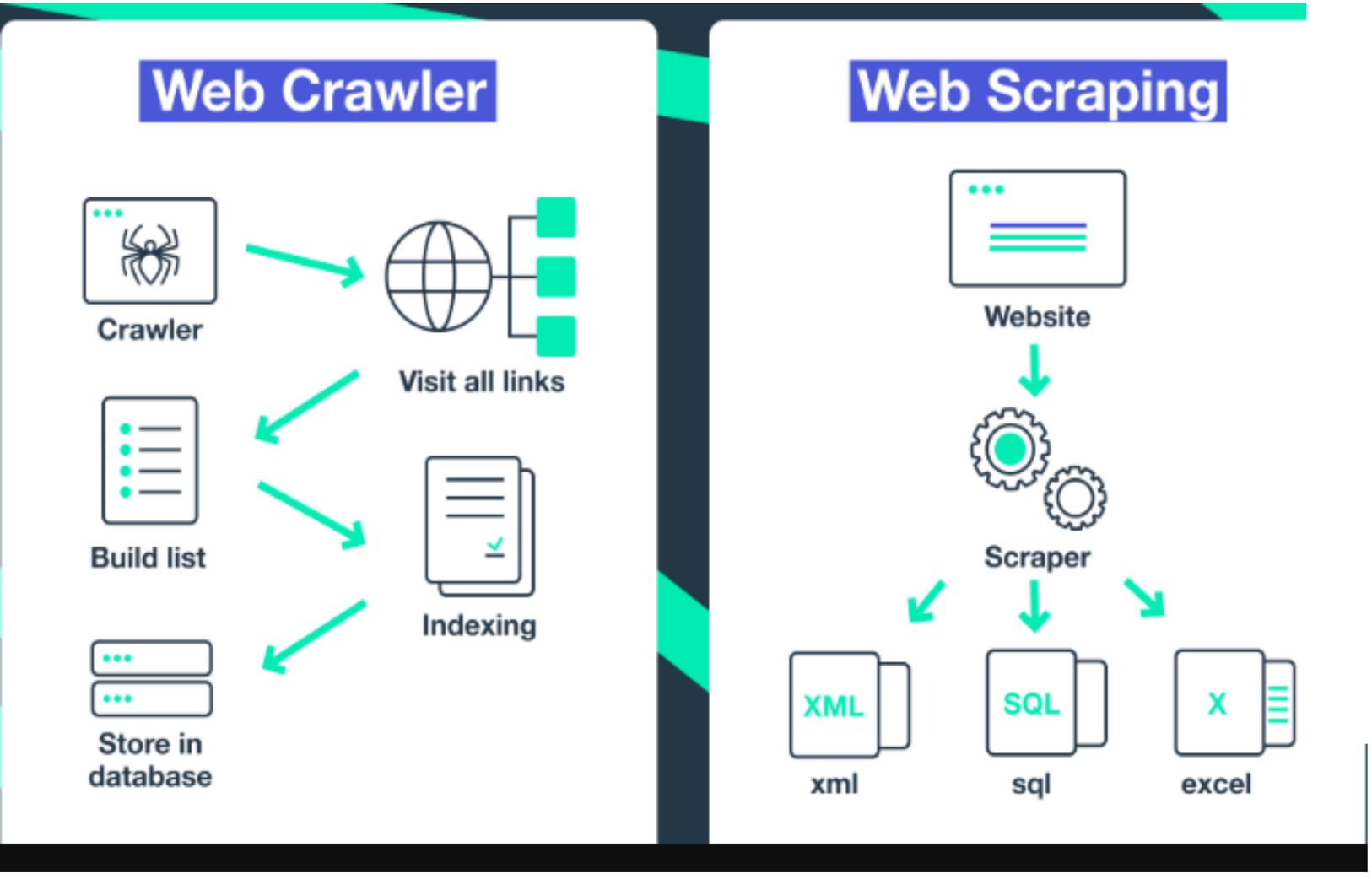
Web Scraping

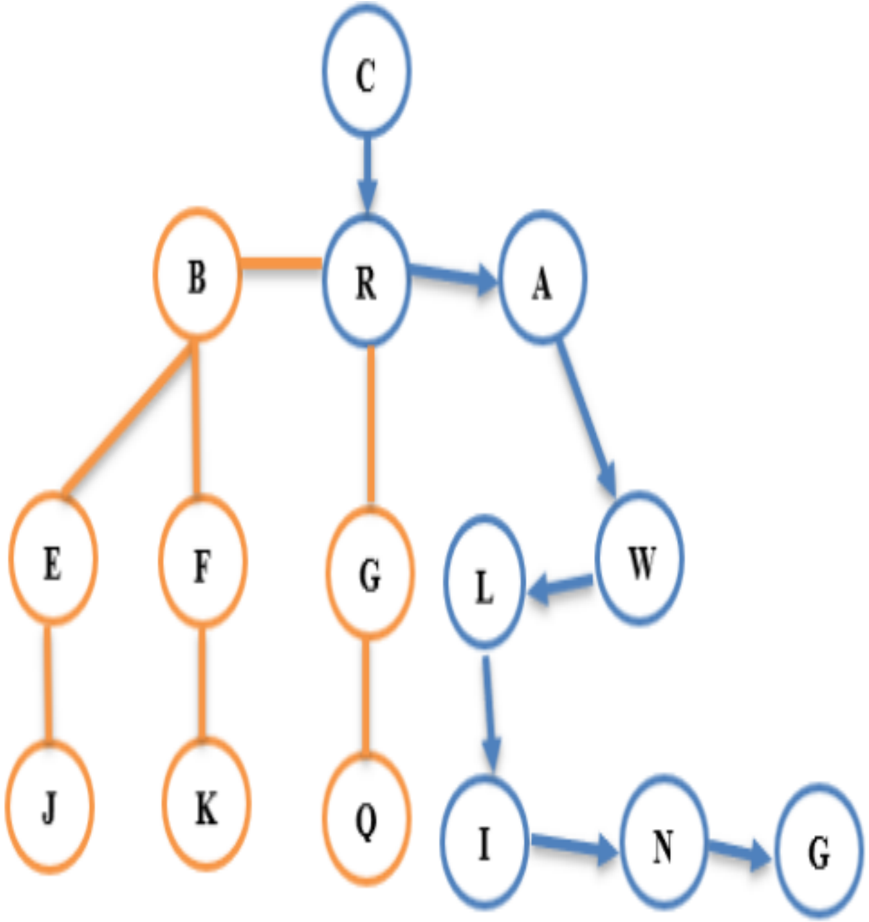
Méthode de Web Scraping	Avantages	Inconvénients
Web Scraping basé sur les expressions régulières	- Simple à mettre en œuvre pour extraire des données structurées simples	- Vulnérable aux changements de structure
Web Scraping basé sur le DOM	- Adapté pour les pages dynamiques	- Nécessite des connaissances sur les structures web
Web Scraping basé sur les APIs	- Généralement plus stable	- Dépend de l'existence d'une API - Peut avoir des limitations d'utilisation (quota, authentification)
Web Scraping basé sur le machine learning, sur l'intelligence artificielle	- Capacité à extraire des données complexes - S'adapte aux variations dans la structure des données	- Nécessite une expertise en machine learning - Peut être coûteux en termes de temps et de ressources

Langage de programmation	Bibliothèque/Framework	Avantages	Inconvénients
Java	Jsoup	- Analyse HTML facile - Gère bien les HTML malformés	- Ne prend pas en charge les pages JavaScript complexes
Python	Scrapy	- Rapide et évolutif - Excellent pour le scraping à grande échelle	- Courbe d'apprentissage plus lent
JavaScript	Puppeteer	- Idéal pour le contenu dynamique - Prise en charge de navigateur sans tête	- Consomme beaucoup de ressources
PHP	Goutte	- Léger pour les tâches simples	- Limitées pour les projets complexes



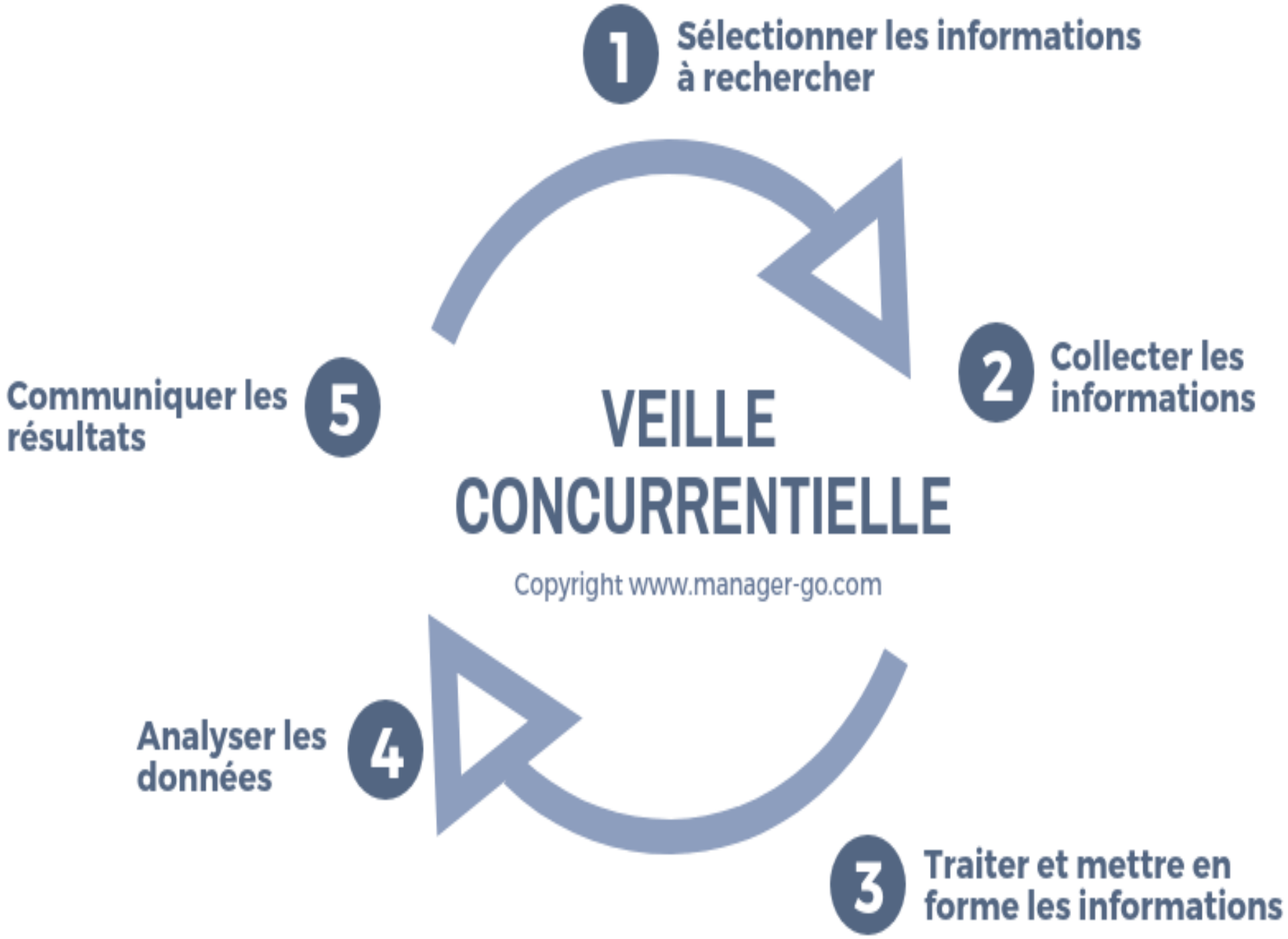
Web scraping et Web Craping





La veille concurrentielle

- **Objectifs:**
  - Rester compétitives dans un marché en constante évolution ;
  - Se doter d'un outil pratique d'aide à la décision ;
  - Développer des stratégies de vente plus rentable ;
  - Identifier ses potentiels clients ;
  - Évaluer les avis des consommateurs sur leurs produits et s'adapter aux tendances ;
  - Ecouter efficacement son écosystème et y répondre favorablement …



Crawling-Scraping et Veille Concurrentiell

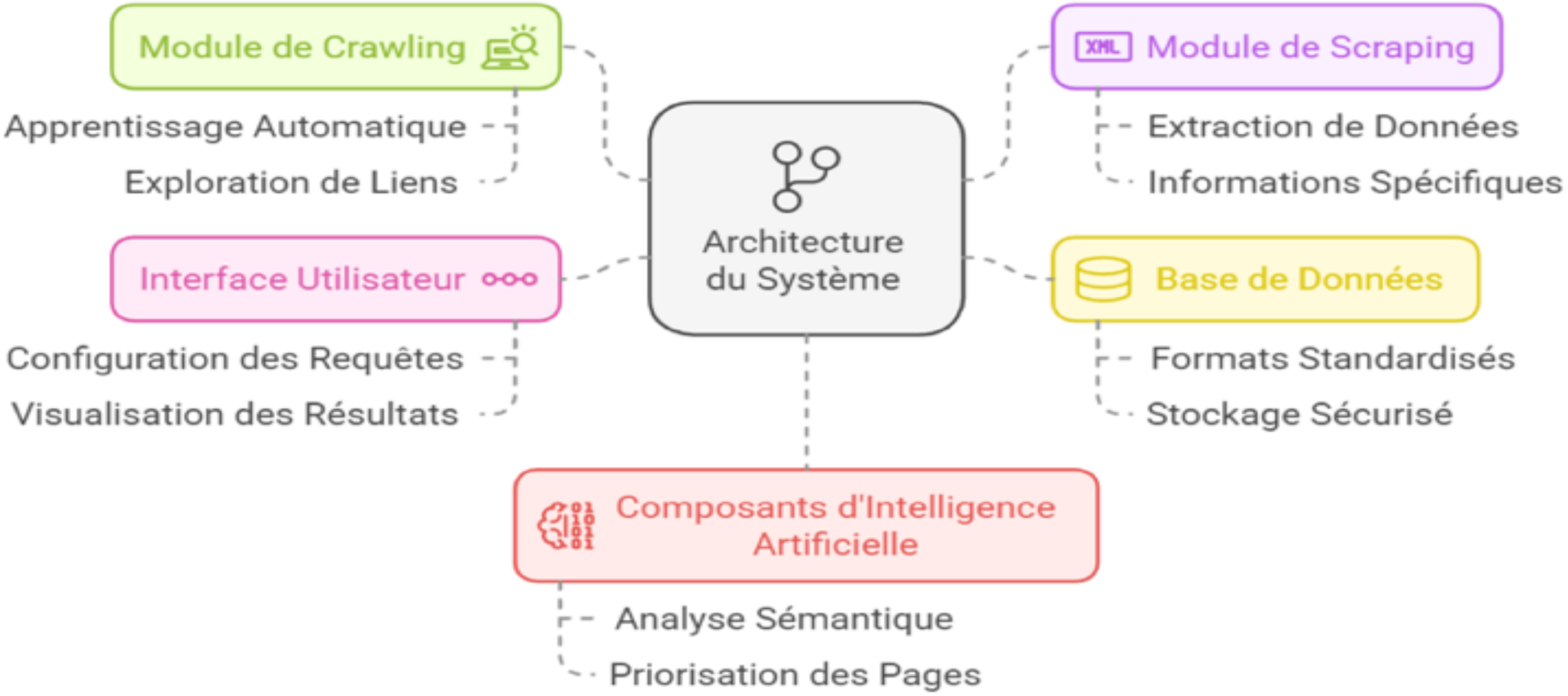
- **Objectifs**
  - Pour une bonne surveillance des tendances du marché ;
  - Pour améliorer l'analyse concurrentielle ;
  - Pour améliorer l'analyse des prix ;
  - Pour augmenter les revenus ;
  - Pour fidéliser les clients ;
  - Pour une meilleure offre de publicité

- **Obstacles**
  - Les changements constants dans les pages et les structures du DOM
  - Diversité des données et des sources (répartition des données sur différentes sources)
  - Contourner les mesures et outils anti-scraping imposées par les sites cibles
  - Légalité de la collecte de données sur les pages
  - La qualité des données collectées : incomplètes, inexactes ou obsolètes
  - Coûts de mise en œuvre et de la surveillance pour le web scraping à grande échelle

La veille concurrentielle augmentée

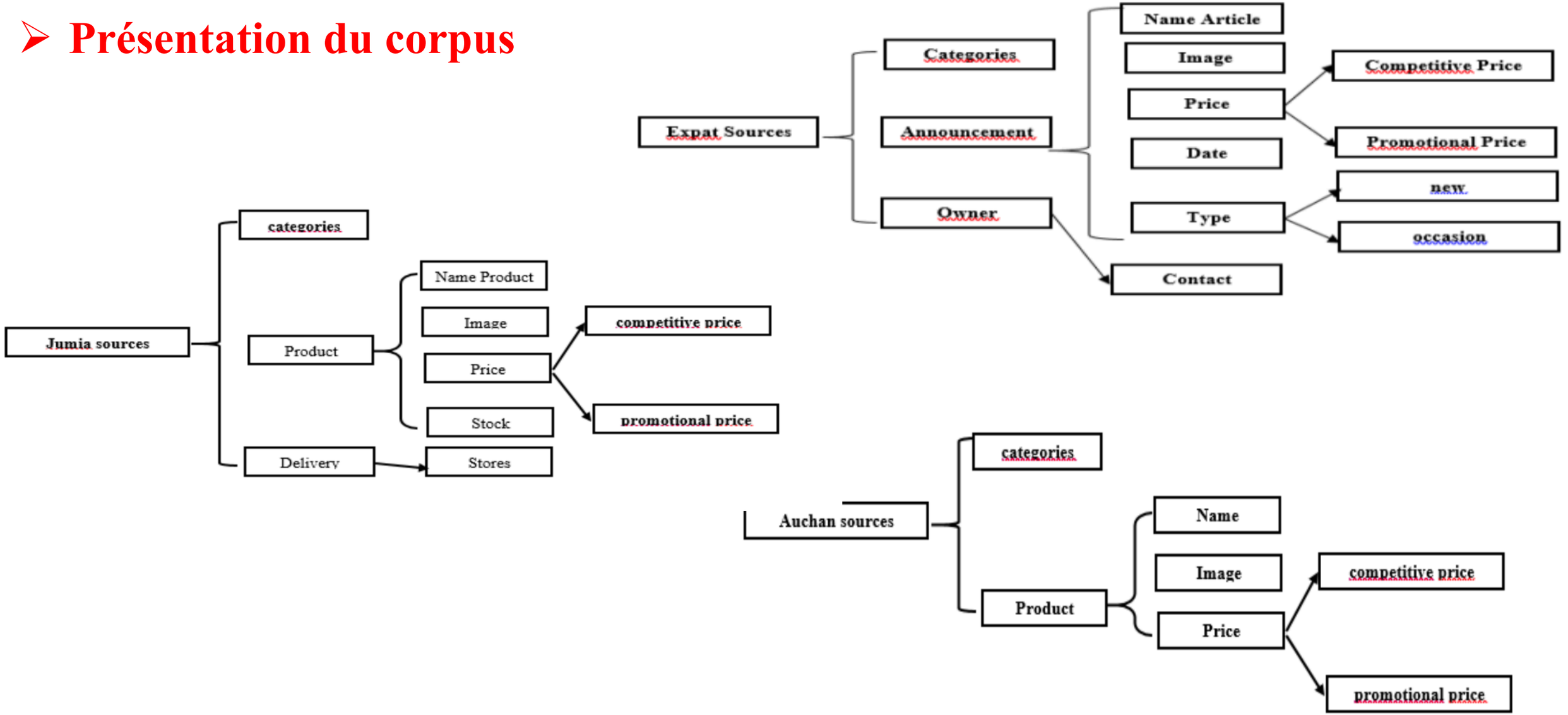
- *La **veille concurrentielle augmentée** fait référence à l'utilisation des technologies avancées (IA, Big Data, Web Scraping, NLP, etc.) pour surveiller et analyser en profondeur les actions des concurrents.*
- Contrairement à la veille traditionnelle, qui repose sur des recherches manuelles et des rapports périodiques, la veille concurrentielle augmentée permet une collecte et une analyse en **temps réel** des données pertinentes.
- **Prise de décision rapide** grâce aux données actualisées.
  - Anticipation des tendances et innovations.
  - Meilleure adaptation des stratégies marketing et commerciales.
  - Optimisation de l'offre produit et des prix.
- **Principaux leviers de la veille concurrentielle augmentée :**
  - Web Crawling & Scraping-& Automatisation
  - Intelligence Artificielle & NLP pour l'analyse descriptive et prédictive
  - Big Data & Data Visualisation pour le Croisement des données de différentes sources et mise en place Tableaux de bord dynamiques avec alertes et indicateurs clés de performance (KPI).

Notre Système : Modélisation du Corpus



Notre Système : Modification du Corpus

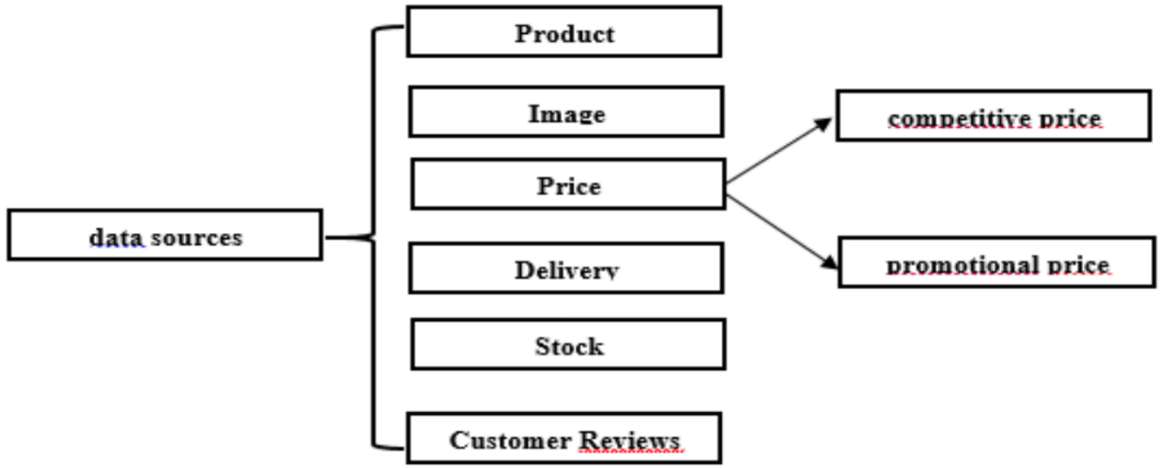
➢ Présentation du corpus



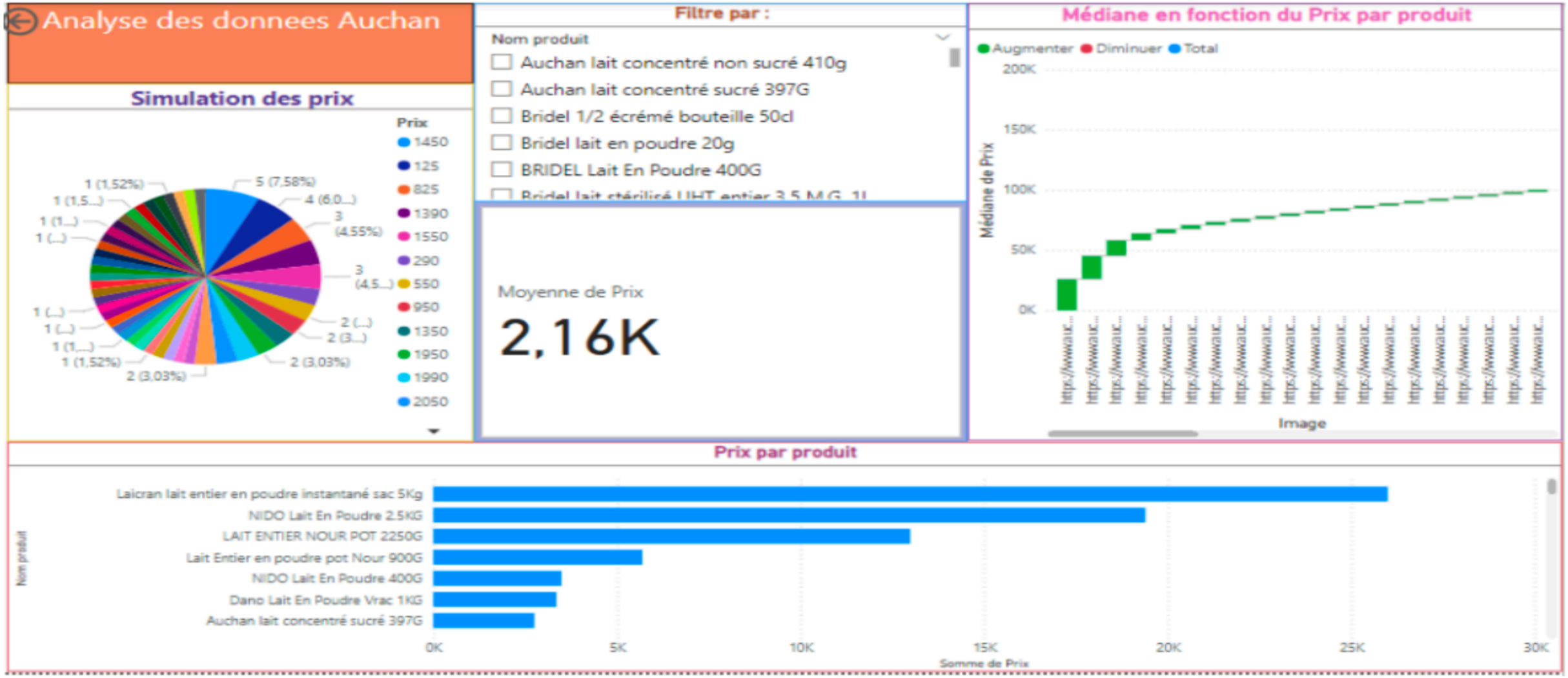
Notre Système : Modification du Corpus

➢ Corpus des données uniforme

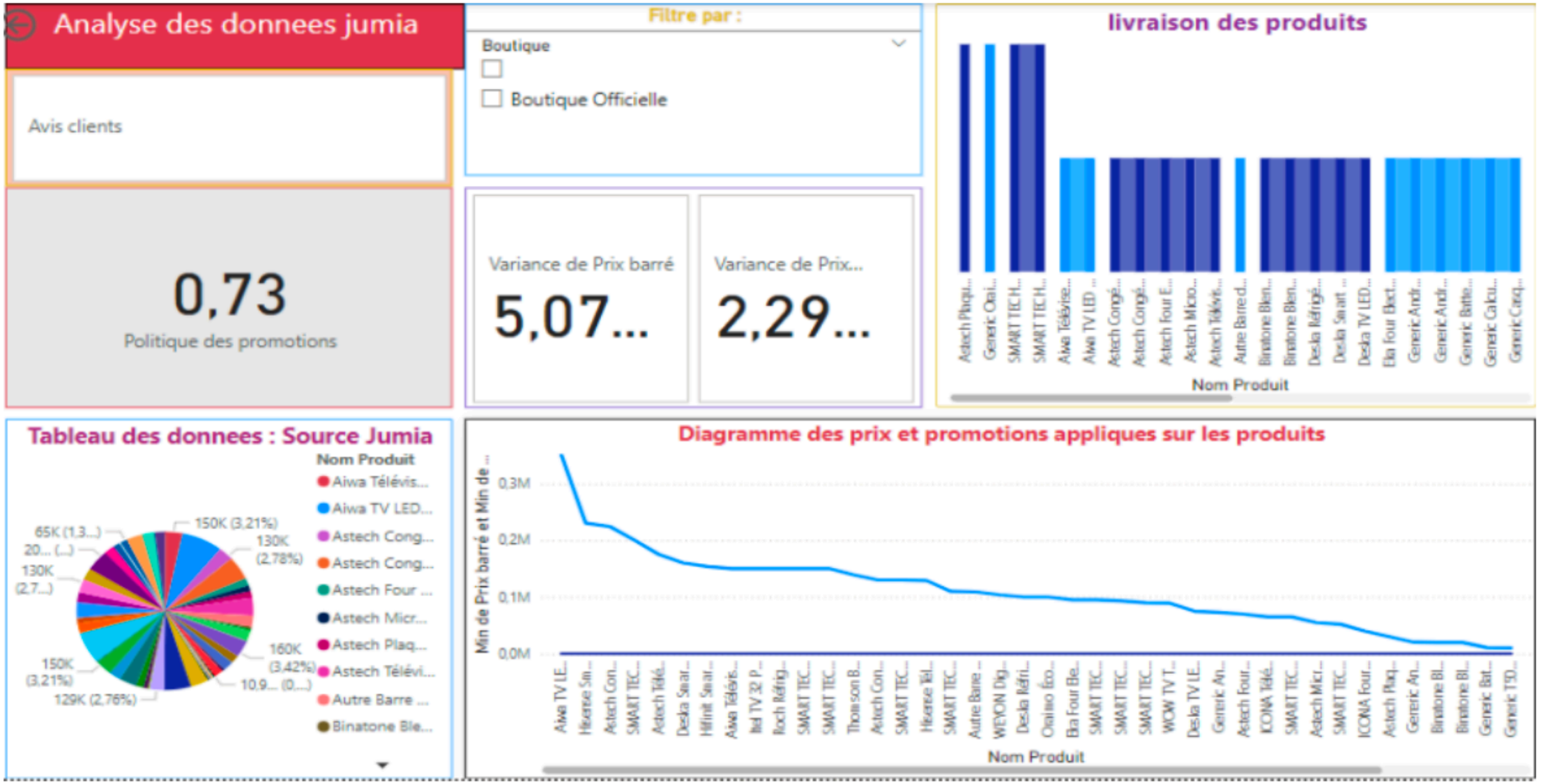
Auchan	Jumia	Expat-Dakar	Final Data
Article	Produit	Announcement	Produit
Nom Article	Nom Produit	Name Article	Nom produit
Prix	Prix	Price	Prix
Flash	Promotions	--	Promotions
Notes et Commande	Notes	--	Avis Client
Livraison	Livraison	Contact	Livraison



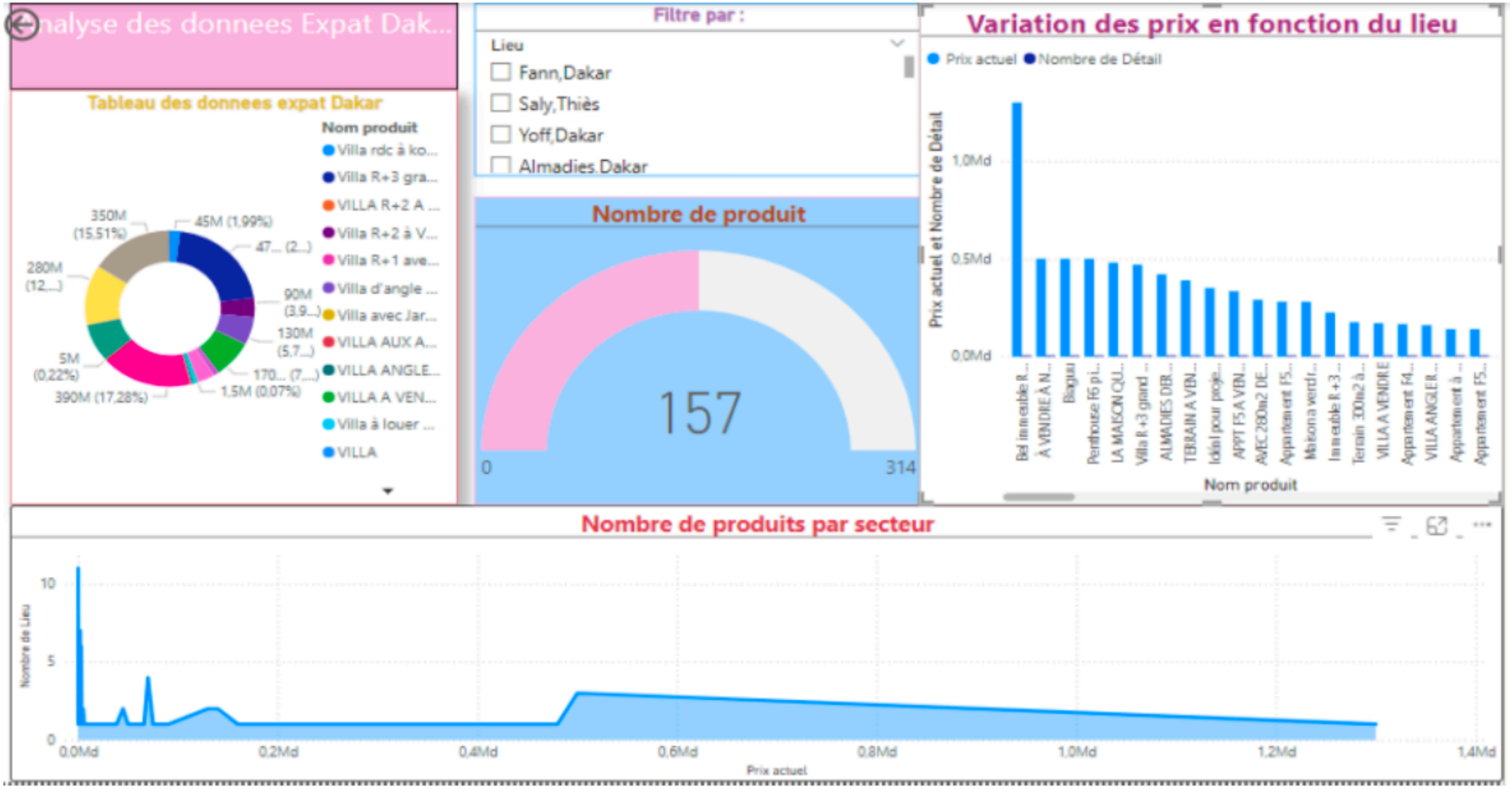
Notre Système : Resultats



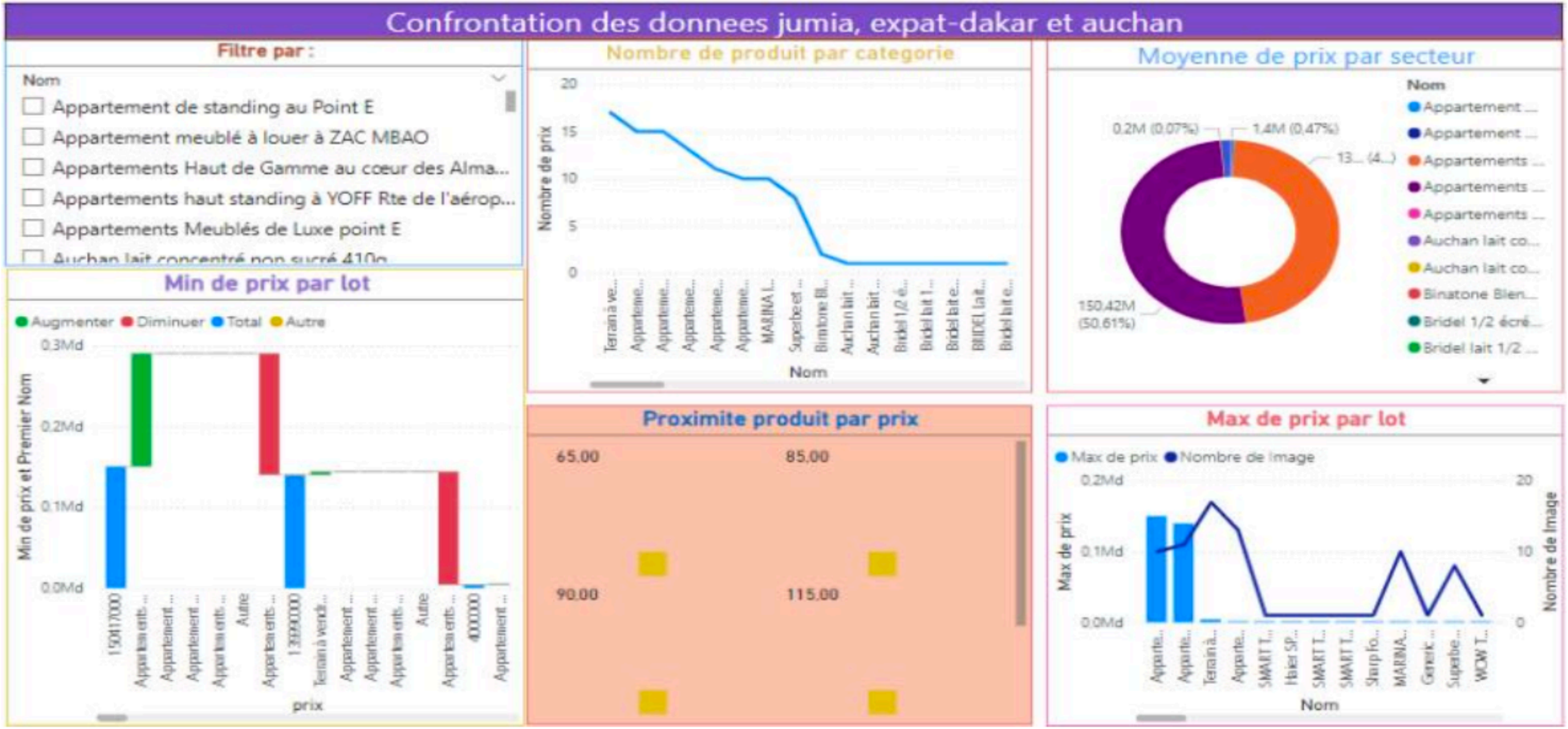
Notre Système : Resultats



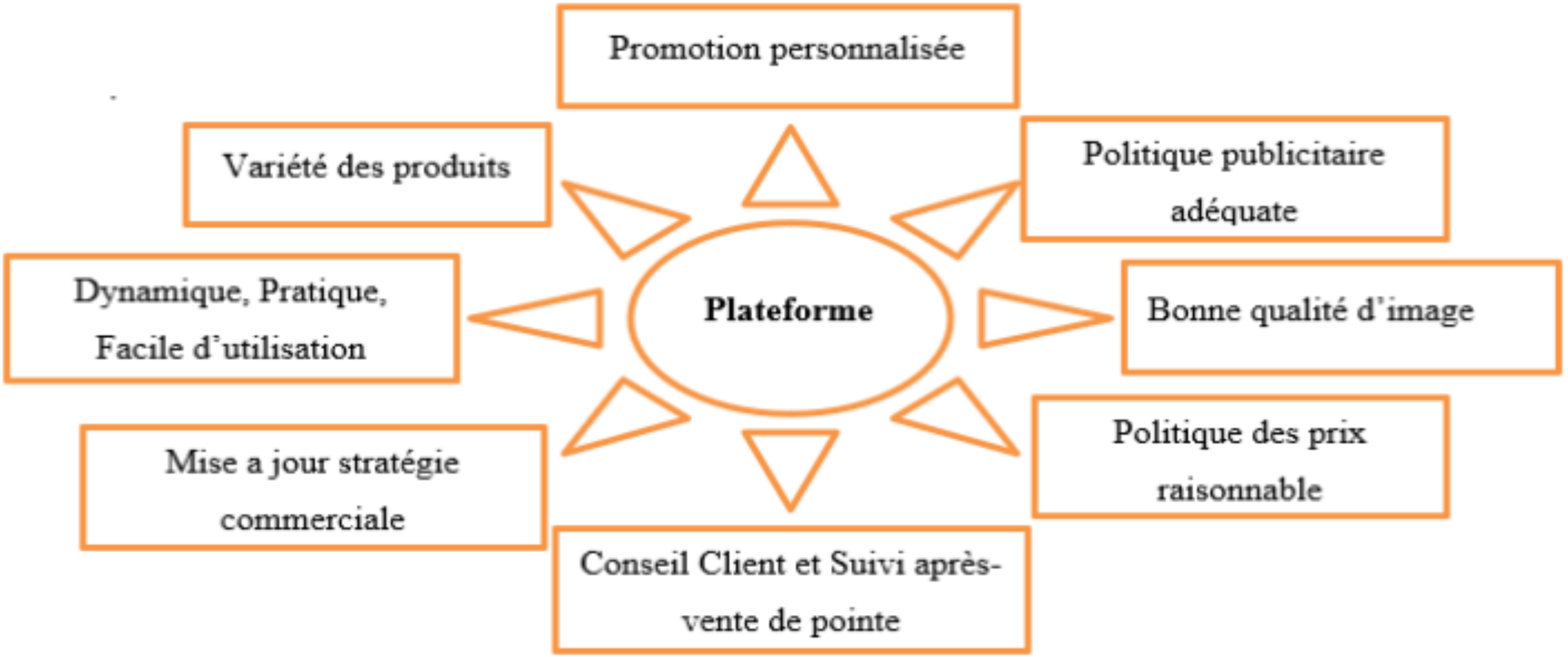
Notre Système : Resultats



Notre Système : Resultats



Orientations commerciales



Merci pour votre attention

Web Scraping & Web Scraping pour la veille concurrentielle augmentée :
Etude de cas sur les E-commerces

[1]Edouard Ngor SARR, [2]Oumar DIAGNE and All
[1,2] University Assane SECK de Ziguinchor - UASZ
SENEGAL

20-21 Février 2025