Université Assane SECK
de Ziguinchor

UFR Sciences et Technologies
Département Informatique
Master 2 Informatique
Spécialité : *Génie Logiciel*

★★★ Rapport de projet ★★★

Analyse de la fluctuation des prix des produits multimédias et électroménagers au Sénégal à l'aide d'une chaine Big Data

★★★★★

Module : Big Data

Fait le :
Ziguinchor, Déc. 2025

Mis en lumière par :
El Hadji Abdou DRAME

Sous la direction de :
Pr. Marie DIOP NDIAYE

Année académique 2024 - 2025

TABLE DES ILLUSTRATIONS

TABLE DES MATIERES

INTRODUCTION

Avec l'essor des nouvelles technologies au Sénégal, le commerce en ligne prend une place importante pour les consommateurs. Aujourd'hui, il est possible d'acheter des produits multimédias et électroménagers tels que les téléphones, ordinateurs ou cuisiniers sur différentes plateformes d'e-commerce.

Cependant, nous notons que les prix varient suivant les différentes plateformes, et parfois même pour les produits. Ces fluctuations des produits rendent la comparaison difficile pour nous consommateurs et compliquent l'analyse du marché de ces produits.

A l'égard de ce constat, les technologies Big Data proposent des solutions intéressantes pour collecter, stocker et analyser de grandes quantités de données issues du web. C'est dans ce contexte que s'inscrit ce projet, dont l'objectif est d'analyser la fluctuation des prix des produits multimédias et électroménagers vendus en ligne au Sénégal à partir de données réelles.

Pour atteindre cet objectif, nous avons mis en place une chaîne Big Data complète allant du web scraping jusqu'à l'analyse des données avec Spark SQL.

I. PRESENTATION GENERALE DU PROJET

Ce projet a été réalisé dans le cadre du mini-projet du module Big Data en Master 2 Génie Logiciel. L'objectif de ce projet est de mettre en pratique les notions enseignées en classe.

Les données ont été récoltées à partir de sites de vente en ligne populaires au Sénégal, tels que:

- Jumia Sénégal
- ManoJia
- Expat-Dakar

Les technologies utilisées sont :

- **Python** pour le web scraping
- **Pentaho Data Integration** pour le traitement ETL[1]
- **Apache Cassandra** pour le stockage des données

[1] ETL signifie Extract, Transform, Load (Extraire, Transformer, Charger). Il permet de récupérer des données depuis différentes sources, de les nettoyer et de les structurer avant de les charger dans une base de données.

- **Apache Spark SQL** pour l'analyse et les requêtes analytiques

II. COLLECTE DES DONNEES – WEB SCRAPING

II.1. Principe du Web Scraping

Le web scraping est une technique qui permet d'extraire automatiquement des informations à partir de pages web. Il consiste à envoyer une requête HTTP à un site, récupérer le code HTML de la page, puis analyser ce code pour extraire les données souhaitées.

Dans ce projet, le web scraping permet de collecter des informations telles que le nom du produit, la catégorie, le prix, le vendeur et la date de collecte.

II.2. Outils utilisés

La collecte des données a été réalisée avec Python en utilisant les bibliothèques suivantes :

- **Requests** pour récupérer le contenu HTML des pages
- **BeautifulSoup** pour analyser le HTML et extraire les informations

Ces bibliothèques sont simples à utiliser et largement utilisées pour le web scraping.

II.3. Données de collecte

Chaque site possède une structure HTML différente. L'analyse de ces sites a permis d'identifier les produits disponibles et d'en extraire les informations nécessaires. Les données suivantes ont ainsi été collectées :

- Nom du produit
- Catégorie
- Prix
- Vendeur
- Date de collecte

Les données ont ensuite été stockées dans un fichier CSV, qui servira d'entrée pour le processus ETL.

III. MODELISATION ET ETL DES DONNEES

III.1. Modélisation Cassandra

Pour le stockage des données, nous avons choisi **Apache Cassandra**, une base de données NoSQL conçue pour gérer de grands volumes de données et offrir de bonnes performances en lecture.

Contrairement aux bases de données relationnelles classiques, Cassandra repose sur une **modélisation orientée requêtes**. Le schéma de données a donc été conçu en fonction des analyses prévues avec **Spark SQL**, notamment le calcul des prix moyens par catégorie et par vendeur.

La table principale, nommée *produits*, est stockée dans le keyspace[2] *ecommerce*. Elle permet de conserver l'historique des prix des produits en tenant compte de la catégorie, du vendeur et de la date de collecte.

Afin de faciliter le déploiement et la configuration de l'environnement, **Apache** Cassandra a été installé à l'aide de **Docker**. Cette approche permet d'obtenir un environnement stable, reproductible et facilement portable, ce qui est particulièrement adapté dans le cadre de ce projet.

Une fois le schéma de stockage défini et la base Cassandra mise en place, l'étape suivante consiste à traiter et charger les données à l'aide d'un processus ETL.

III.2. Processus ETL

Après la collecte des données et la définition du schéma Cassandra, un processus ETL a été mis en place avec Pentaho Data Integration pour nettoyer, structurer et charger les données.

Le fichier **produits.csv**, issu du web scraping, a d'abord été traité pour corriger les incohérences: conversion des prix en valeurs numériques et normalisation des champs textuels (Prix et nom du produit). Un filtrage a ensuite permis de séparer les lignes complètes des lignes incomplètes. Les données complètes ont été conservées dans un fichier CSV pour insertion dans Cassandra, tandis que les anomalies ont été sauvegardées dans un fichier séparé pour suivi.

[2] Un **keyspace** est l'équivalent d'une base de données dans les systèmes relationnels. Il contient plusieurs tables et définit les règles de réplication et de durabilité des données

Enfin, les données valides ont été préparées selon le schéma du espace ***ecommerce***, garantissant la qualité des informations stockées et facilitant les analyses ultérieures avec Spark SQL. L'exécution du job ETL s'est déroulée avec succès, confirmant la bonne intégration des différentes étapes du processus.

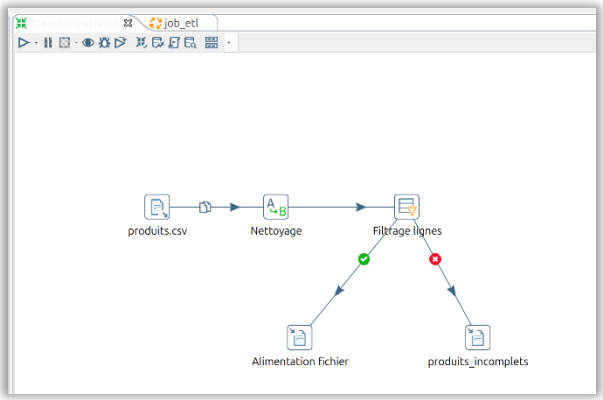


Figure 1: Schéma de la transformation des données

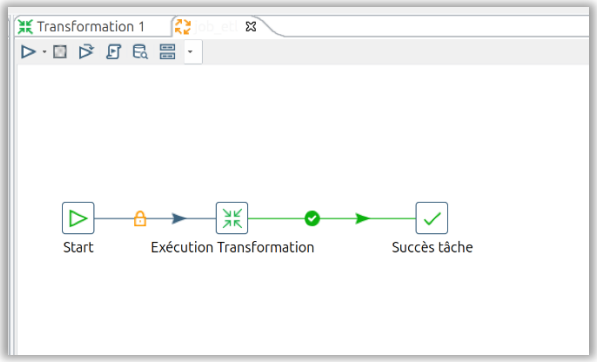


Figure 2: Schéma du Job ETL

IV. Analyse des données avec Spark SQL

IV.1. Connexion Spark – Cassandra

Apache Spark a été connecté à Cassandra à l'aide du connecteur Spark-Cassandra. Les données ont été chargées sous forme de DataFrame[3], ce qui permet d'exécuter facilement des requêtes analytiques avec Spark SQL.

[3] Structure de données tabulaire utilisée dans Apache Spark, similaire à une table de base de données ou à une feuille Excel. Elle organise les données en lignes et colonnes et permet d'effectuer facilement des opérations d'analyse à l'aide de Spark SQL

IV.2. Requêtes analytiques réalisées

Plusieurs requêtes ont été réalisées afin d'extraire des indicateurs pertinents, notamment :

- Le prix moyen par vendeur
- Le prix moyen par catégorie
- Le nombre de produits par catégorie

IV.3. Résultats obtenus

- Prix moyen par vendeur

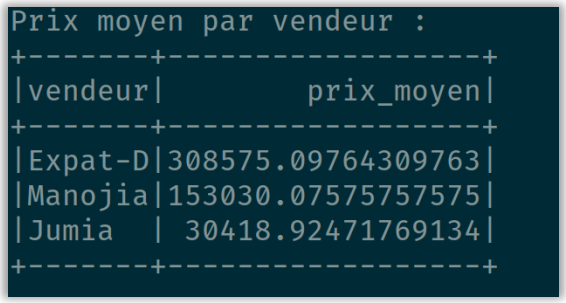


Figure 3: Résultat - Prix moyen par Vendeur

- Prix moyen par catégorie

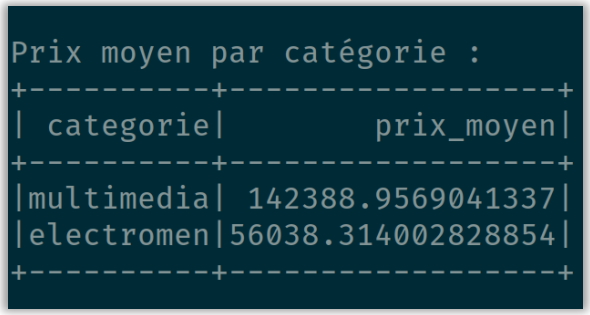


Figure 4: Résultat - Prix moyen par catégorie

- Nombre de produits par catégorie

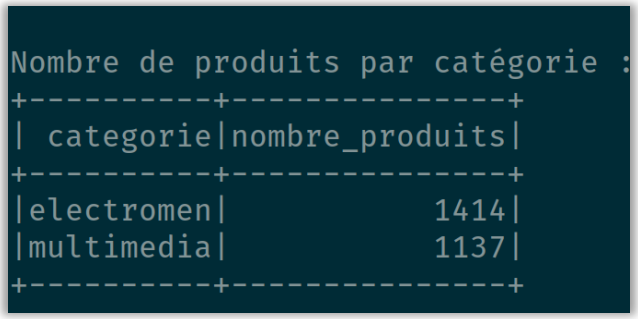


Figure 5: Résultat - Nombre de produits par catégorie

V. Interprétation et analyse des résultats

L'analyse des résultats met en évidence des écarts de prix significatifs entre les vendeurs (*Figure 3: Résultat - Prix moyen par Vendeur*). La plateforme **Expat-Dakar** affiche les prix moyens les plus élevés, tandis que **Jumia** propose les prix les plus bas. **ManoJia** se positionne de manière intermédiaire entre ces deux extrêmes.

Ces différences peuvent s'expliquer par le modèle économique et le positionnement commercial de chaque plateforme. Jumia adopte une stratégie basée principalement sur le volume de ventes, en proposant des prix bas afin d'attirer un large public, parfois au détriment de la qualité ou de la durabilité des produits. À l'inverse, Expat-Dakar met davantage en avant des produits souvent plus coûteux, destinés à une clientèle spécifique disposant d'un pouvoir d'achat plus élevé. ManoJia se situe dans une logique intermédiaire, combinant accessibilité des prix et offre relativement diversifiée.

L'analyse par catégorie (*Figure 4: Résultat - Prix moyen par catégorie*) montre que les produits multimédias présentent des prix moyens plus élevés que les produits électroménagers. Cette différence peut s'expliquer par la nature des produits analysés dans chaque catégorie. Les produits multimédias regroupent souvent des équipements tels que les téléviseurs, les ordinateurs ou les smartphones, dont les prix unitaires sont généralement plus élevés.

À l'inverse, la catégorie électroménagère inclut un grand nombre de produits de petite taille ou d'entrée de gamme (fers à repasser, mixeurs, bouilloires, micro-ondes), ce qui tire le prix moyen vers le bas. Cette différence reflète donc davantage la composition des catégories que des facteurs liés à la production ou à l'importation.

Ces résultats mettent en évidence l'intérêt de l'analyse Big Data pour comprendre les stratégies commerciales des plateformes et les dynamiques de prix sur le marché sénégalais**.**

Dans la suite de ce rapport, nous présentons les principaux apports du projet ainsi que les difficultés rencontrées lors de sa réalisation.

VI. APPORTS ET DIFFICULTES RENCONTREES

VI.1. Apports du projet

Ce projet nous a permis de mettre en place une chaîne Big Data complète, depuis la collecte des données jusqu'à leur analyse. Il a également facilité l'automatisation du processus de collecte et de traitement des données, ce qui réduit les tâches manuelles et les risques d'erreur.

L'utilisation de données réelles issues de sites e-commerce sénégalais a permis d'identifier des tendances de prix concrètes, notamment les différences de prix entre vendeurs et entre catégories de produits. Ce travail a aussi renforcé notre compréhension des technologies Big Data telles que Cassandra, Pentaho et Spark SQL, ainsi que leur intégration dans un même projet.

Enfin, ce projet constitue une bonne initiation à une approche orientée données.

VI.2. Difficultés rencontrées

Plusieurs difficultés ont été rencontrées au cours de la réalisation de ce projet. Tout d'abord, les structures HTML différentes selon les sites web ont rendu le web scraping plus complexe. Chaque site nécessitait une logique d'extraction spécifique, ce qui a demandé du temps d'analyse et plusieurs ajustements dans les scripts Python.

Ensuite, la qualité des données collectées a posé un problème. Certains prix étaient mal formatés, des champs étaient manquants ou certains produits incomplets. Cela a nécessité un travail important de nettoyage et de transformation des données lors de l'étape ETL avec Pentaho.

Par ailleurs, l'utilisation de Pentaho avec Cassandra a présenté des contraintes techniques. Le **plugin Cassandra n'est pas pris en charge nativement** par certaines versions de Pentaho, ce qui a compliqué la connexion et l'insertion des données dans la base. Des solutions alternatives ont dû être mises en place pour assurer le chargement des données.

Des problèmes de compatibilité entre les versions des outils ont également été rencontrés. Certaines versions de Java n'étaient pas compatibles avec Spark, et une version inappropriée de Python a causé des conflits avec le connecteur Cassandra utilisé par Spark. Ces

incompatibilités ont nécessité plusieurs ajustements de configuration et des tests avant d'obtenir un environnement fonctionnel.

Enfin, la **collecte des données sur une période limitée** ne permet pas d'analyser précisément l'évolution des prix dans le temps, ce qui constitue une limite pour des analyses plus approfondies.

Ces apports et difficultés permettent de mieux apprécier les résultats obtenus et ouvrent la voie à une réflexion globale sur les perspectives d'amélioration du projet, qui sera abordée dans la conclusion.

CONCLUSION

Ce projet a permis d'analyser les tendances de prix des produits multimédias et électroménagers sur différentes plateformes de e-commerce au Sénégal, en s'appuyant sur une chaîne Big Data complète. Les résultats obtenus montrent des variations de prix importantes selon les plateformes et les catégories de produits, traduisant des stratégies commerciales distinctes adoptées par les vendeurs.

Malgré les difficultés rencontrées, notamment liées à l'hétérogénéité des structures des sites web et à la qualité des données collectées, le projet a démontré l'efficacité des technologies **Apache Cassandra**, **Pentaho Data Integration** et **Apache Spark SQL** pour le traitement, le stockage et l'analyse de données issues du web. L'intégration de ces outils a permis de mettre en place un processus automatisé et cohérent, depuis la collecte jusqu'à l'analyse des données.

En termes de perspectives, ce travail pourrait être amélioré de plusieurs manières. Une collecte des données sur une période plus longue permettrait d'analyser plus finement l'évolution des prix dans le temps et d'identifier des tendances saisonnières. L'intégration de nouvelles plateformes de e-commerce élargirait également le champ de l'étude et rendrait les résultats plus représentatifs du marché sénégalais. Enfin, l'ajout d'outils de visualisation des données, tels que Power BI ou Tableau, faciliterait l'interprétation des résultats et améliorerait la restitution des analyses auprès des décideurs ou des consommateurs.

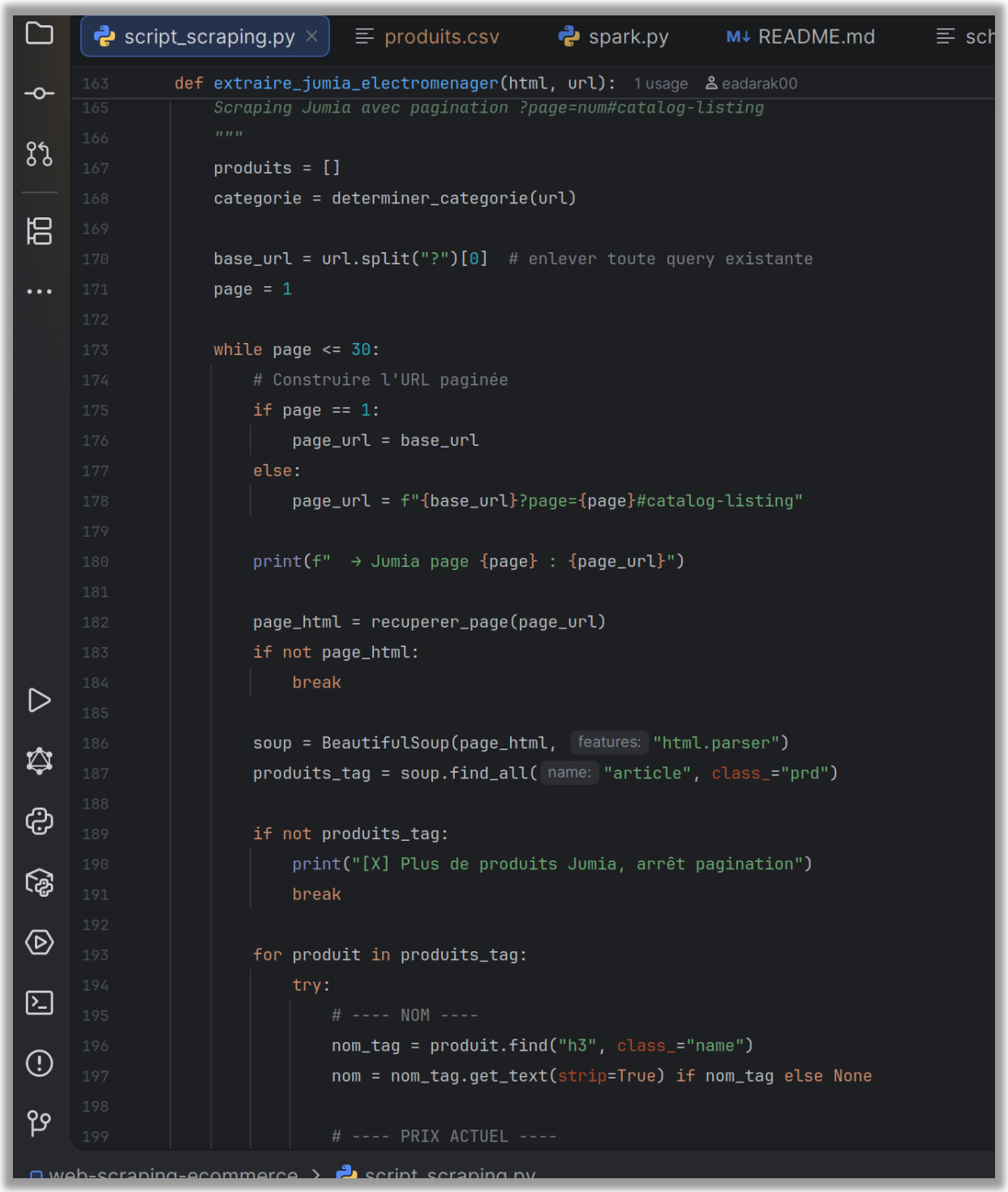
BIBLIOGRAPHIE

1. Apache Cassandra Documentation – https://cassandra.apache.org
2. Apache Spark SQL Documentation – https://spark.apache.org/docs
3. Pentaho Data Integration Documentation – Hitachi – https://docs.pentaho.com/pdia-data-integration/
4. BeautifulSoup Documentation – https://www.crummy.com/software/BeautifulSoup/

ANNEXES

ANNEXE I

Scripts Python (scraping)



ANNEXE II

Script Schéma CQL Cassandra

spark.py schema.cql

```
-- Schema de cassandra

CREATE KEYSPACE IF NOT EXISTS ecommerce
WITH replication = {
    'class': 'SimpleStrategy',
    'replication_factor': 1
};

-- Utilisation du KEYSPACE
USE ecommerce;

-- TABLE Produits
CREATE TABLE IF NOT EXISTS produits(
    categorie TEXT,
    vendeur TEXT,
    nom TEXT,
    date_collection TIMESTAMP,
    prix DOUBLE,

    PRIMARY KEY(
        (categorie),
        vendeur,
        nom,
        date_collection
    )
);


-- Les index facilitant certains analyses
CREATE INDEX IF NOT EXISTS idx_produits_vendeur
ON produits (vendeur);

CREATE INDEX IF NOT EXISTS idx_produits_nom
```

web-scraping-ecommerce > schema.cql

ANNEXE III

Scripts Spark SQL

spark.py

```python
import os
from pyspark.sql import SparkSession
from pyspark.sql.functions import avg, col, count

#  Ajouter le connecteur Cassandra
os.environ["PYSPARK_SUBMIT_ARGS"] = "--packages com.datastax.spark:spark-cassandra-connector_2.13:3.5.0 pyspark-shel

# Créer la SparkSession avec configuration Cassandra
spark = SparkSession.builder \
    .appName("CassandraProduits") \
    .master("local[*]") \
    .config("spark.cassandra.connection.host", "127.0.0.1") \
    .config("spark.cassandra.connection.port", "9042") \
    .getOrCreate()

# Lire la table produits du keyspace ecommerce
df_produits = spark.read \
    .format("org.apache.spark.sql.cassandra") \
    .options(keyspace="ecommerce", table="produits") \
    .load()

# Afficher quelques lignes
print("Données existantes dans produits :")
df_produits.show(10)

# ---------------- ANALYSES ----------------

# Analyse 1 : comparaison des prix par vendeur
print("Prix moyen par vendeur :")
df_produits.groupBy("vendeur") \
    .agg(avg("prix").alias("prix_moyen")) \
    .orderBy(col("prix_moyen").desc()) \
    .show()
```

10:36 LF UTF-8 4 spaces Python 3.12

ANNEXE IV

Extrait du Job Pentaho