



Mini Projet

Contexte et objectif

La fluctuation des prix des produits multimédia (Ordinateurs, Téléphones, Accessoires multimédia, Télévisions, Jeux vidéo, etc.) et électroménagers (Réfrigérateurs, Congélateurs, Climatiseurs, Cuisinières, Machines à laver, Micro-ondes, Machines à café, etc.) est un enjeu économique majeur au Sénégal. Les données publiées sur des sites de commerce en ligne peuvent être exploitées pour analyser les tendances de prix.

L'objectif de ce projet, est d'analyser les tendances des prix des produits multimédia et électroménagers partir de données des sites de vente en ligne. Vous devez, construire une chaîne Big Data complète, depuis la collecte jusqu'à l'analyse en suivant le pipeline ci-dessous :

Web Scraping → Pentaho (ETL) → Cassandra → Spark SQL →
Résultats analytiques

La collecte de données peut se faire à partir des sites de vente en ligne les plus connus (voir le tableau ci-dessous).

1	EXPAT-DAKAR	https://www.expat-dakar.com/
2	JUMIA	https://www.jumia.sn/
3	SHOP ME AWAY	https://www.shopmeaway.com/
4	MANO JIA	https://www.manojia.com/
5	DAKARMARKET	https://dakarmarket.sn/

Technologies imposées

- **Web Scraping** : Python (BeautifulSoup / Scrapy)
- **ETL** : Pentaho Data Integration (Kettle)
- **Stockage Big Data** : Apache Cassandra
- **Traitement & Analyse** : Apache Spark SQL

Déroulement du travail

Collecte de données

À partir de sites de vente en ligne, collectez les informations suivantes :

- Nom du produit
- Catégorie



- Prix
- Vendeur
- Date de collecte

Les données collectées seront stockées dans un fichier CSV ou JSON.

Modélisation et ETL

Proposer un schéma Cassandra pour le stockage des données collectées. Ensuite, proposez un Job ETL dans Pentaho pour le nettoyage et le chargement des données dans Cassandra.

Analyse avec Spark SQL

Connecter Spark à Cassandra puis effectuez au moins trois requêtes analytiques.
Par exemple :

- Le prix moyen par produit ;
- Comparaison des prix par vendeur ou par catégorie.

Livrables

1. Script de web scraping
2. Fichier CSV ou JSON
3. Job Pentaho (ETL)
4. Schéma Cassandra
5. Scripts Spark SQL
6. Rapport (5–10 pages) qui doit contenir ce qui listé ci-dessus, ainsi que les résultats et l'analyse des résultats obtenus avec les scripts Spark SQL
7. Présentation synthétique de 5 à 10 minutes

Consignes

- Le travail est à faire individuellement.
- Vous disposez de 10 jours pour rendre le travail.
- Une date vous sera communiquée pour la présentation de vos travaux.