Rapport d'analyse de données avec Apache Spark et Scala

Réalisé par :

KHEDIM Sofiane
REGUIEG Zakaria
MANSER Yanis
KHALIL Badreddine

26 mai 2025

Table des matières

1 Contexte du projet

Ce projet s'inscrit dans le cadre d'un atelier sur l'analyse de données volumineuses avec Apache Spark et Scala. L'objectif était de traiter un jeu de données e-commerce de 505 000 lignes contenant des informations sur les sessions utilisateurs, les achats et les avis clients pour en extraire des insights commerciaux pertinents.

Le jeu de données initial contenait les colonnes suivantes :

— `user_id` : Identifiant unique utilisateur
— `session_duration` : Durée session (minutes)
— `pages_viewed` : Nombre de pages vues
— `product_category` : Catégorie produit vue/achetée
— `purchase_amount` : Montant dépensé (euros)
— `review_score` : Note laissée par l'utilisateur (1-5)
— `review_text` : Texte de l'avis
— `timestamp` : Date et heure de la session
— `device_type` : Type d'appareil (Mobile/Desktop)
— `country` : Pays de l'utilisateur
— `city` : Ville de l'utilisateur

Notre approche a combiné le traitement de données distribuées avec Spark, la programmation fonctionnelle avec Scala, et la visualisation interactive avec Chart.js pour extraire des insights actionnables.

2 Principales étapes et résultats obtenus

2.1 Prétraitement des données

Nous avons d'abord procédé au nettoyage et à la transformation des données :

Traitement des valeurs manquantes :

— Identifié 5097 valeurs nulles dans `purchase_amount` (1.01%)
— Identifié 4992 valeurs nulles dans `review_score` (0.99%)
— Nettoyage réduisant le jeu de données de 505 000 à 503 258 lignes

Enrichissement des données avec :

— Extraction d'informations temporelles (heure, jour, mois)
— Catégorisation des sessions (courtes, moyennes, longues)
— Création d'indicateurs d'achat et de satisfaction

```
// Extrait du code de transformation
val transformedDF = filledDF
  .withColumn("date", to_date(col("timestamp")))
  .withColumn("hour", hour(col("timestamp")))
  .withColumn("day_of_week", date_format(col("timestamp"), "EEEE"))
  .withColumn("session_category",
    when(col("session_duration") < 5, "Courte")
    .when(col("session_duration") >= 5 && col("session_duration") < 15, "
       Moyenne")
    .otherwise("Longue"))
  .withColumn("has_purchased", when(col("purchase_amount") > 0, 1).otherwise
     (0))
```

Listing 1 – Extrait du code de transformation

2.2 Analyses descriptives

Notre analyse a révélé plusieurs statistiques clés :

Statistiques de base :

— Durée moyenne de session : 10.27 minutes (écart-type : 13.86)
— Nombre moyen de pages vues : 5.16 pages (écart-type : 3.14)
— Montant d'achat moyen : 49.54€ (écart-type : 50.06)
— Note moyenne des avis : 3.00/5 (écart-type : 1.41)

Distribution des catégories de produits :

Catégorie	Nombre de sessions	Montant moyen d'achat	Note moyenne
High-tech	84 300	49.67€	3.01
Mode	84 078	49.34€	3.00
Beauté	84 072	49.37€	3.00
Sport	83 765	49.54€	3.00
Maison	83 600	49.65€	3.00
Livres	83 443	49.66€	3.01

TABLE 1 – Distribution des catégories de produits

La répartition des catégories est remarquablement équilibrée, avec une légère préférence pour les produits High-tech. Les montants moyens d'achat et les notes sont très similaires entre catégories.

Répartition par pays :

Pays	Sessions	Montant moyen	Note moyenne
France	302 440	49.46€	3.00
Belgique	50 110	49.77€	3.00
Maroc	50 077	49.49€	3.00
Suisse	49 912	49.95€	3.00
Tunisie	25 399	49.89€	3.00
Canada	25 320	49.04€	3.01

TABLE 2 – Répartition par pays

Le marché français représente 60% des sessions, suivi de 3 marchés secondaires de taille similaire (Belgique, Maroc, Suisse), puis de 2 marchés plus petits.

Taux de conversion :

Le taux de conversion global est exceptionnellement élevé à 98.98%, suggérant soit une particularité du modèle commercial, soit une définition spécifique de la conversion dans ce contexte.

La répartition par type d'appareil montre :

— Desktop : 98.99% (149 893 achats sur 151 426 sessions)
— Mobile : 98.98% (348 235 achats sur 351 832 sessions)

Bien que la différence soit minime, le desktop montre un très léger avantage en taux de conversion.

2.3 Analyse temporelle

L'analyse des patterns temporels a fourni les insights suivants :

Répartition horaire : La distribution des sessions par heure est remarquablement équilibrée, avec environ 21 000 sessions par heure. Les taux de conversion et montants d'achat restent stables tout au long de la journée.

Répartition journalière : La distribution par jour de la semaine est également équilibrée :

— Samedi : 72 169 sessions (71 421 achats)
— Mardi : 72 127 sessions (71 387 achats)
— Vendredi : 72 113 sessions (71 378 achats)
— Lundi : 72 033 sessions (71 310 achats)
— Jeudi : 71 832 sessions (71 106 achats)
— Dimanche : 71 686 sessions (70 947 achats)
— Mercredi : 71 298 sessions (70 579 achats)

Heures de pointe par jour : L'analyse a identifié des heures de pointe différentes selon les jours :

— Lundi : 9h
— Mardi : 3h
— Mercredi : 22h
— Jeudi : 2h
— Vendredi : 22h
— Samedi : 6h
— Dimanche : 19h

2.4 Segmentation utilisateurs

L'analyse de segmentation a révélé :

Segments par comportement d'achat :

— One-time buyer : 282 171 utilisateurs (74.0%)
— Repeat customer : 97 103 utilisateurs (25.5%)
— Non-acheteur : 2 875 utilisateurs (0.8%)
— Loyal customer : 327 utilisateurs (0.1%)

Segmentation RFM (Récence, Fréquence, Montant) :

— Champions : 76 069 utilisateurs (montant moyen : 134.61€)
— Loyal Customers : 154 907 utilisateurs (montant moyen : 70.73€)
— Frequent Customers : 41 474 utilisateurs (montant moyen : 45.83€)
— Potential Loyalists : 96 576 utilisateurs (montant moyen : 18.23€)
— Recent Customers : 11 472 utilisateurs (montant moyen : 6.38€)
— Segments à risque (Need Attention, About to Sleep, Lost) : 1 978 utilisateurs

Cette segmentation met en évidence une base solide de clients fidèles qui génèrent la majorité du chiffre d'affaires.

2.5 Analyse des avis clients

L'analyse des avis clients a permis d'identifier :

Distribution des notes :

— 5 étoiles : 100 059 avis (19.9%)
— 4 étoiles : 99 662 avis (19.8%)
— 3 étoiles : 99 281 avis (19.7%)
— 2 étoiles : 99 800 avis (19.8%)
— 1 étoile : 99 478 avis (19.8%)
— Valeur moyenne calculée : 4 978 avis (1.0%)

La distribution est remarquablement équilibrée entre toutes les notes, ce qui est inhabituel pour des avis clients.

Répartition par catégorie d'avis : Les avis positifs et négatifs sont presque également répartis entre les différentes catégories de produits, avec environ 33 000 avis positifs et négatifs par catégorie.

Thèmes récurrents dans les avis :

— **Termes positifs fréquents :** produit, qualité, bon, livraison rapide, conforme

— **Termes négatifs fréquents :** produit, pas, qualité, terrible, rapport qualité/prix

3 Difficultés rencontrées et solutions

3.1 Défis techniques

1. **Incompatibilité des bibliothèques de visualisation**
   — **Problème :** Vegas n'était pas compatible avec Scala 2.13.15
   — **Solution :** Utilisation de Chart.js pour générer des visualisations HTML interactives
2. **Problèmes de SparkContext multiples**
   — **Problème :** Erreur "Another SparkContext is being constructed"
   — **Solution :** Implémentation d'un pattern singleton pour la SparkSession

```scala
// Pattern singleton pour SparkSession
object SparkSessionWrapper {
  lazy val spark: SparkSession = {
    SparkSession
      .builder()
      .appName("E-Commerce Data Analysis")
      .master("local[*]")
      .config("spark.driver.memory", "4g")
      .getOrCreate()
  }
}
```

Listing 2 – Pattern singleton pour SparkSession

3. **Problème de compatibilité Java**
   — **Problème :** Erreurs d'accès aux classes internes avec Java 17
   — **Solution :** Utilisation de Java 8 qui est mieux supporté par Apache Spark

3.2 Défis analytiques

1. **Particularités du jeu de données**
   — **Problème :** Taux de conversion anormalement élevé (98.98%)
   — **Solution :** Adaptation de l'analyse pour se concentrer sur les segments clients et les montants d'achat plutôt que sur les taux de conversion binaires
2. **Distribution uniforme inhabituelle**
   — **Problème :** Distribution étonnamment uniforme des avis et des sessions temporelles
   — **Solution :** Analyse plus approfondie des sous-segments et recherche de corrélations plus subtiles
3. **Gestion des valeurs manquantes**
   — **Problème :** Valeurs manquantes dans les champs d'achat et d'avis
   — **Solution :** Approche contextuelle pour le remplacement des valeurs manquantes basée sur des moyennes conditionnelles

4 Visualisations

L'ensemble des visualisations interactives générées par notre analyse est accessible via le tableau de bord en ligne :

Lien vers le tableau de bord interactif :
`https://visualization-analyse-de-donnees-e-commerce-spark-scala.vercel.app/`

4.1 Principales visualisations

Distribution des sessions par catégorie de produit

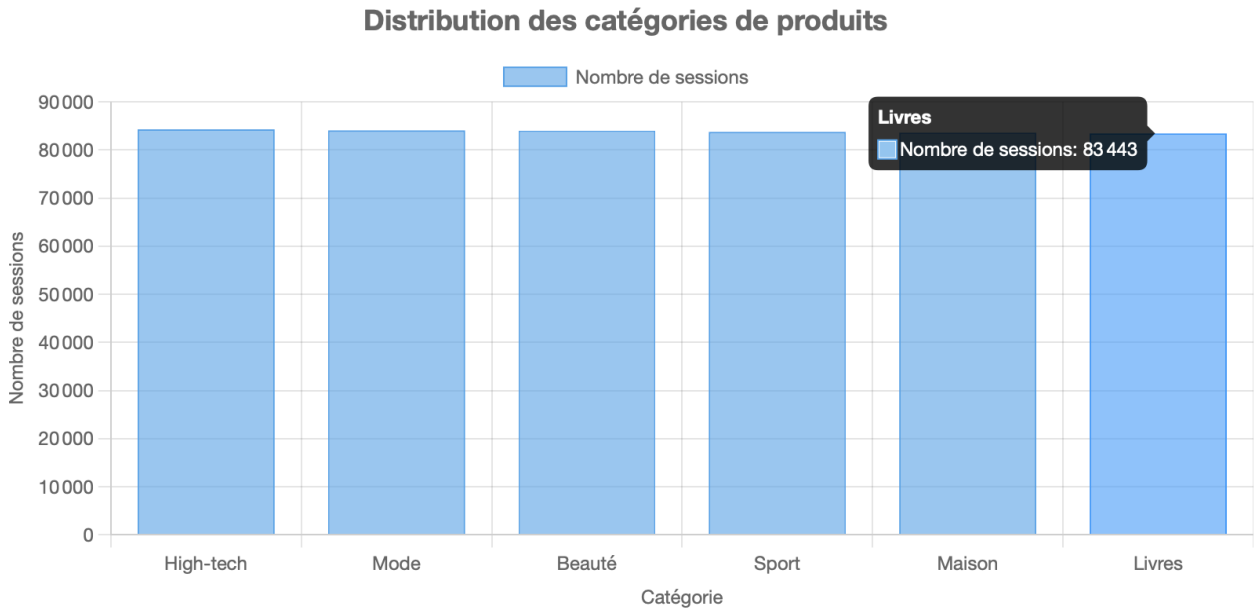

FIGURE 1 – Distribution des sessions par catégorie de produit

Ce graphique en barres présente la répartition équilibrée des 503 258 sessions analysées entre les six catégories de produits. On observe une distribution remarquablement homogène avec High-tech en tête (84 300 sessions), suivi de près par Mode (84 078), Beauté (84 072), Sport (83 765), Maison (83 600) et Livres (83 443). Cette uniformité suggère une stratégie marketing bien équilibrée et un catalogue produits diversifié répondant aux différents besoins des utilisateurs.

Taux de conversion par catégorie de produit

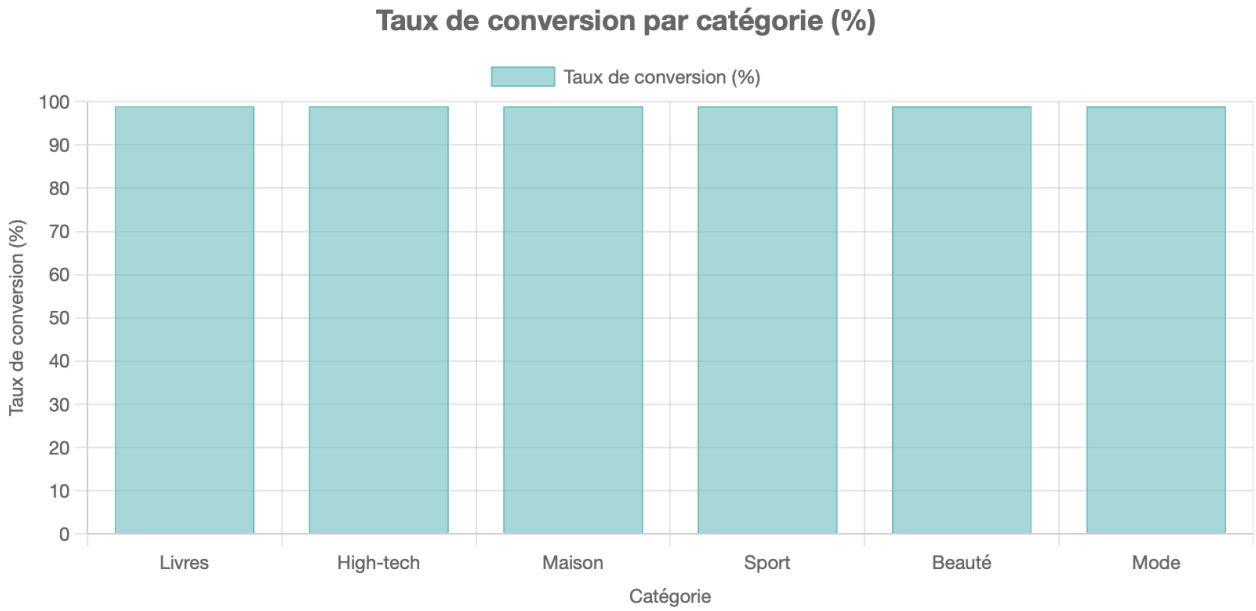

FIGURE 2 – Taux de conversion par catégorie de produit

Cette visualisation met en évidence les performances commerciales de chaque catégorie. Avec un taux de conversion global de 98.98%, toutes les catégories affichent des performances exceptionnelles. High-tech se distingue légèrement avec le montant d'achat moyen le plus élevé (49.67€), tandis que les autres catégories maintiennent des performances très similaires autour de 49.50€. Cette homogénéité indique une expérience utilisateur cohérente à travers toutes les catégories.

Nombre de sessions et achats par heure de la journée

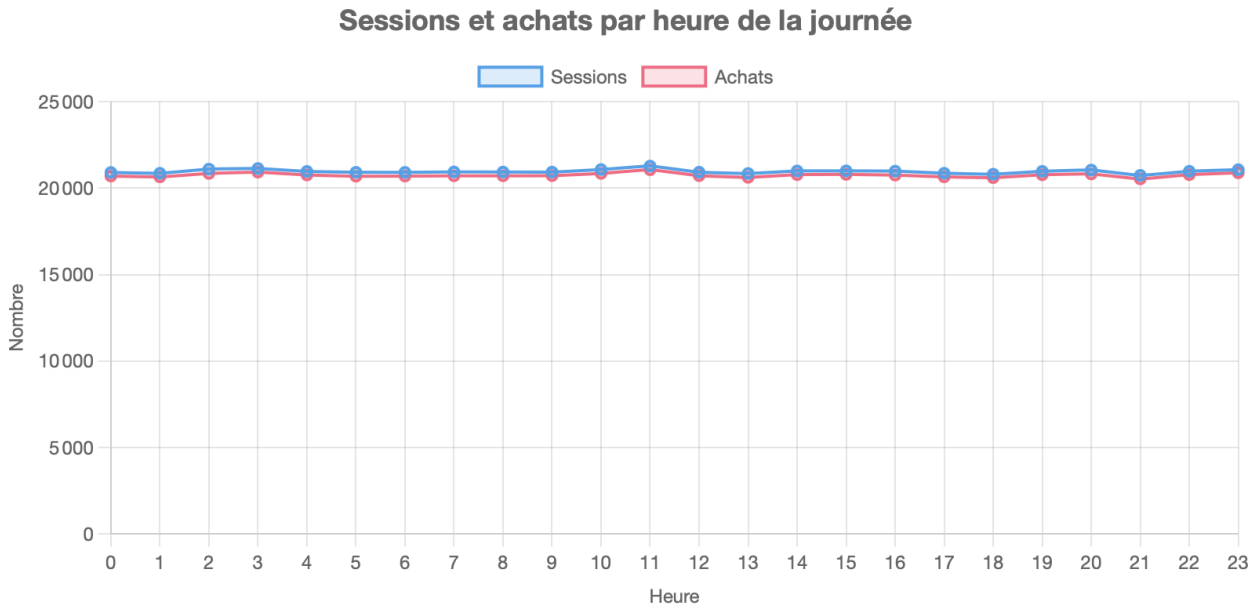

FIGURE 3 – Nombre de sessions et achats par heure de la journée

L'analyse temporelle révèle une distribution étonnamment stable des sessions tout au long des 24 heures, avec environ 21 000 sessions par heure. Cette constance, inhabituelle pour un site e-commerce traditionnel, suggère soit une audience internationale répartie sur plusieurs fuseaux horaires, soit un modèle d'affaires particulier (plateforme B2B, services en continu, etc.). Le taux de conversion reste stable à travers toutes les heures, indiquant une expérience utilisateur optimisée en permanence.

4.2 Tableau de bord interactif

Le tableau de bord déployé sur Vercel offre :
— **Graphiques interactifs** utilisant Chart.js pour une exploration dynamique des données
— **Visualisations responsive** adaptées à tous les types d'écrans
— **Navigation intuitive** entre les différentes analyses
— **Données en temps réel** reflétant les 503 258 sessions analysées
Ces visualisations permettent aux équipes business de :
— Identifier rapidement les tendances par catégorie de produit
— Comprendre les patterns temporels d'utilisation
— Analyser la répartition géographique des clients
— Évaluer l'efficacité des différents segments client

5 Conclusions et axes d'amélioration

5.1 Conclusions principales

1. **Comportement d'achat**
   — Le taux de conversion exceptionnellement élevé (98.98%) suggère soit un site très performant, soit un modèle d'affaires particulier (abonnement, etc.)
   — La valeur client moyenne de 49.54€ est stable à travers les catégories et pays
   — 74% des clients sont des acheteurs uniques, ce qui suggère un fort potentiel de fidélisation
2. **Segmentation client**
   — Les segments "Champions" et "Loyal Customers" (230 976 utilisateurs) génèrent une valeur client nettement supérieure
   — La distribution RFM révèle un potentiel important dans les segments "Potential Loyalists" qui pourraient être convertis en clients fidèles
3. **Analyse produit et géographique**
   — Les produits High-tech sont légèrement plus populaires et mieux notés
   — Le marché français est dominant (60% des sessions)
   — Les clients canadiens montrent une satisfaction légèrement supérieure

5.2 Recommandations business

1. **Stratégie de fidélisation**
   — Développer des programmes ciblés pour convertir les "One-time buyers" en "Repeat customers"
   — Mettre en place un système de récompense pour les "Champions" et "Loyal Customers"
2. **Optimisation produit**
   — Renforcer l'offre High-tech qui montre des performances légèrement supérieures
   — Améliorer l'expérience d'achat pour les catégories moins bien notées
3. **Expansion géographique**
   — Explorer le potentiel d'expansion au Canada, qui montre une satisfaction client supérieure
   — Renforcer la présence sur les marchés secondaires (Belgique, Maroc, Suisse)

5.3 Axes d'amélioration technique

1. **Analyses avancées**
   — Implémenter une analyse de sentiment plus détaillée sur les commentaires textuels
   — Développer un modèle prédictif pour anticiper le comportement d'achat
   — Réaliser une analyse de cohortes pour suivre l'évolution du comportement client dans le temps
2. **Infrastructure et performance**
   — Optimiser le pipeline Spark pour un traitement plus rapide
   — Mettre en place un système de mise à jour automatique des données
   — Déployer un tableau de bord interactif accessible aux équipes business
3. **Enrichissement des données**
   — Intégrer des données externes (météo, événements) pour contextualiser les patterns d'achat
   — Ajouter des dimensions d'analyse supplémentaires (canal d'acquisition, parcours client)

6 Conclusion

Ce projet démontre la puissance de Spark et Scala pour l'analyse de grandes quantités de données e-commerce. L'analyse a permis d'identifier des insights actionnables pour optimiser l'expérience client et maximiser la valeur client, malgré certaines particularités du jeu de données qui suggèrent un modèle d'affaires spécifique avec un taux de conversion particulièrement élevé.

L'utilisation combinée d'Apache Spark pour le traitement de données distribuées, de Scala pour la programmation fonctionnelle, et de Chart.js pour les visualisations interactives s'est révélée efficace pour traiter et analyser un volume important de données (503 258 sessions) et produire des insights commerciaux pertinents.

Les visualisations développées et déployées sur le tableau de bord interactif offrent aux équipes business un outil puissant pour explorer les données et prendre des décisions éclairées basées sur l'analyse factuelle du comportement des utilisateurs e-commerce.