

Compte-Rendu TP 2 : Technologie du Big Data

Réalisé par : Mouad KHAZNAOUI, Badreddine BELACHKAR, le 17 janvier 2024

**Introduction :**

Ce rapport expose les travaux pratiques effectués lors de l'utilisation de la librairie MRJobs pour le traitement de données massives avec Hadoop. Divisé en deux phases, le premier volet consistait en des tests de la librairie, notamment l'exécution d'exemples visant à compter les mots d'un livre selon divers modes tels que le mode inline, local et hadoop.

La seconde partie du travail pratique s'est concentrée sur l'exploration et l'analyse des données de ventes du fichier ventes/purchases.txt. L'objectif était de recueillir des informations significatives et de calculer des statistiques pertinentes.

Ce travail porte sur l'interrogation du fichier de ventes pour résoudre des questions spécifiques en utilisant la librairie MRJob. Les questions portaient sur le nombre d'achats par catégorie, la somme totale dépensée par catégorie, les dépenses par ville et moyen de paiement… Chaque question a été traitée à travers des scripts Python distincts, dont les résultats ont été rapportés sur l'ensemble des données, sauf dans les cas extrêmes où la capacité de la machine ne le permettait pas.

**Exercice 1 : Questionner un fichier de ventes**

**Question 1 : Quel est le nombre d’achats effectués pour chaque catégorie d’achat ?**

Le script que nous avons utilisé est un script MapReduce qui se base sur la Library MRJob, pour répondre à la question "Quel est le nombre d’achats effectués pour chaque catégorie d’achat ?" en suivant un modèle algorithmique où la fonction de mappage extrait la catégorie d'achat de chaque transaction, émettant des paires clé-valeur indiquant une occurrence d'achat. La fonction de réduction agrège ces occurrences par catégorie, calculant le nombre total d'achats dans chaque catégorie.

**Résultat obtenu :**

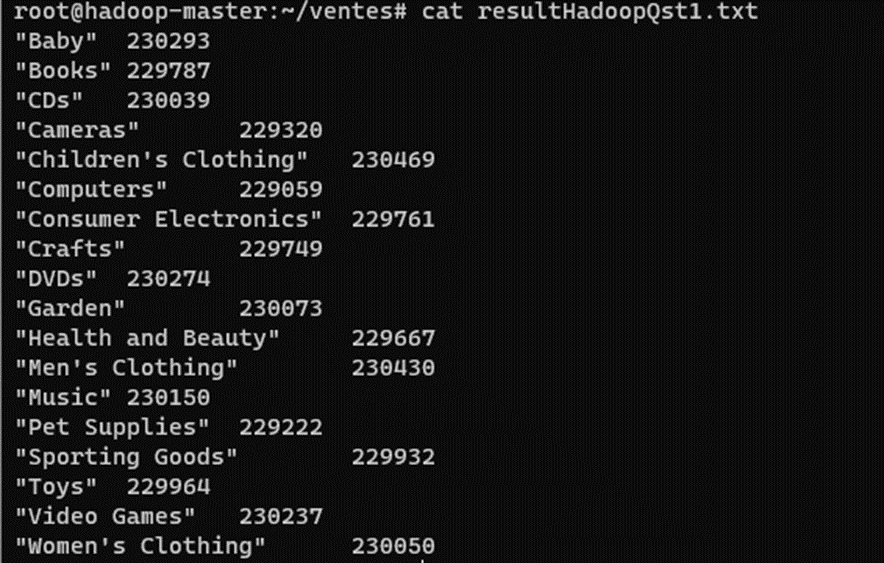


Figure 1 - le nombre d’achats effectués pour chaque catégorie d’achat

**Question 2 : Quelle est la somme totale dépensée pour chaque catégorie d’achat ?**

Dans cette question, la fonction de mappage décortique subtilement la catégorie d'achat et le montant de chaque transaction, émettant des paires clé-valeur pour illustrer le montant dépensé dans chaque catégorie. De manière similaire, la fonction de réduction agrège ces montants par catégorie, permettant ainsi de calculer la somme totale dépensée dans chaque catégorie d'achat.

**Résultat obtenu :**

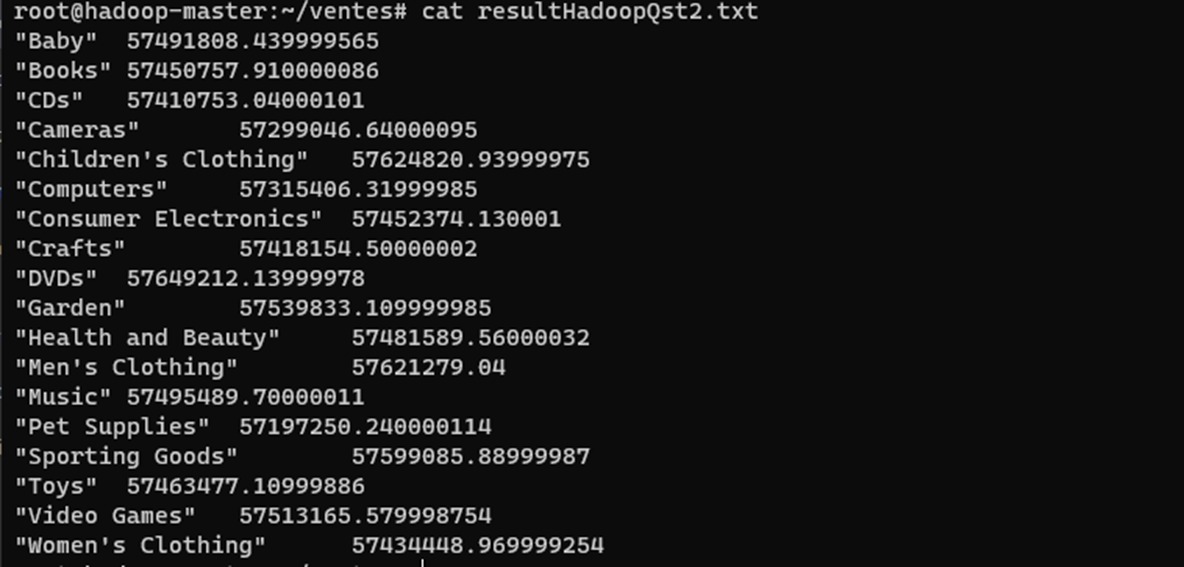
****

Figure 2 - la somme totale dépensée pour chaque catégorie d’achat

**Question 3 : Quelle somme est dépensée dans la ville de San Francisco dans chaque moyen de paiement ?**

Ici, la fonction de mappage extrait la ville d'achat, le montant de la vente, et le moyen de paiement associé à chaque transaction, émettant la ville comme clé et un tuple contenant le montant de la vente et le moyen de paiement comme valeur. La fonction de réduction agrège ensuite ces montants pour chaque moyen de paiement, fournissant une perspective détaillée des dépenses selon les différents moyens de paiement dans chaque ville. Le script donne les moyens de paiement pour chaque ville.

Pour afficher uniquement la ligne correspondant à San Francisco dans le fichier resultHadoopQst3.txt, nous avons utilisé la commande ‘**grep’** :

grep "San Francisco" resultHadoopQst3.txt

**Résultat obtenu :**

****

Figure 3 - La somme dépensée dans la ville de San Francisco dans chaque moyen de paiement

**Question 4 : Dans quelle ville la catégorie Women’s Clothing a permis de générer le plus d’argent Cash ?**

Ici, la fonction de mappage extrait la catégorie d'achat, la ville, la somme dépensée et le moyen de paiement associé à chaque transaction. Il émet ensuite la combinaison (catégorie, ville, paiement) comme clé et la somme dépensée comme valeur. La fonction de réduction agrège ces montants pour chaque combinaison, et le résultat final est émis uniquement si la catégorie est "Women's Clothing" et le moyen de paiement est "Cash". Le script permet de fournir l'argent généré en espèces pour chaque ville.

Nous avons extrait la ville qui a généré le montant maximum en utilisant la commande ‘**sort’** dans le terminal :

sort -t$'\t' -k2,2nr resultHadoopQst4P.txt | head -n 1

**Résultat obtenu :**

****

Figure 4 - La ville ou la catégorie Women’s Clothing a permis de générer le plus d’argent Cash

**Question 5 : Ajoutez une requête originale et complexe (i.e. nombre de STEPS > 1) de votre choix sur ce fichier. Dans le CR, prenez le temps de m’expliquez l’objectif de la requête, de présenter le code et le résultat obtenu : tableau, graphique…**

**Requête choisie : Analyse des ventes par catégorie et par ville à partir des données d'achats (STEPS =2)**L'objectif de cette requête est d'analyser comment les différents types de catégories d'achat sont répartis dans chaque ville. L'analyse des ventes par catégorie et par ville prend une importance clé dans le monde des affaires. C'est un peu comme un outil de détective pour comprendre ce que les gens aiment acheter dans chaque coin du monde. Imaginez que chaque ville a ses propres petites préférences en matière de produits. En regardant ces données, les entreprises peuvent ajuster leur offre et leurs promotions pour coller au plus près des goûts locaux. C'est un peu comme personnaliser une playlist musicale pour chaque ville, mais avec des produits. Cela aide les entreprises à être plus malines sur où vendre quoi, à éviter les tracas logistiques, et à être plus proches de leurs clients. En fin de compte, c'est comme donner aux entreprises un super pouvoir pour mieux comprendre leurs marchés et faire des choix plus futés.

La première étape du processus consiste à utiliser la fonction de mappage pour extraire les informations pertinentes de chaque ligne, notamment la ville d'achat et la catégorie d'achat. Ces informations sont ensuite émises sous la forme d'une paire clé-valeur, où la clé est une combinaison de la ville et de la catégorie, et la valeur est toujours 1. La deuxième étape utilise une fonction de réduction pour agréger ces paires clé-valeur en calculant le total des occurrences pour chaque combinaison de ville et catégorie. Enfin, une deuxième fonction de réduction est utilisée pour pivoter les données, transformant la structure en une table où les colonnes représentent les catégories, les lignes représentent les villes, et les valeurs sont le nombre d'occurrences.

Nous allons créer un tableau qui montre, pour chaque ville, le nombre d'occurrences de chaque type de catégorie.

**Exemple de résultat obtenu pour la ville de San Antonio :**

****

Figure . Exemple de résultat obtenu pour la ville de San Antonio

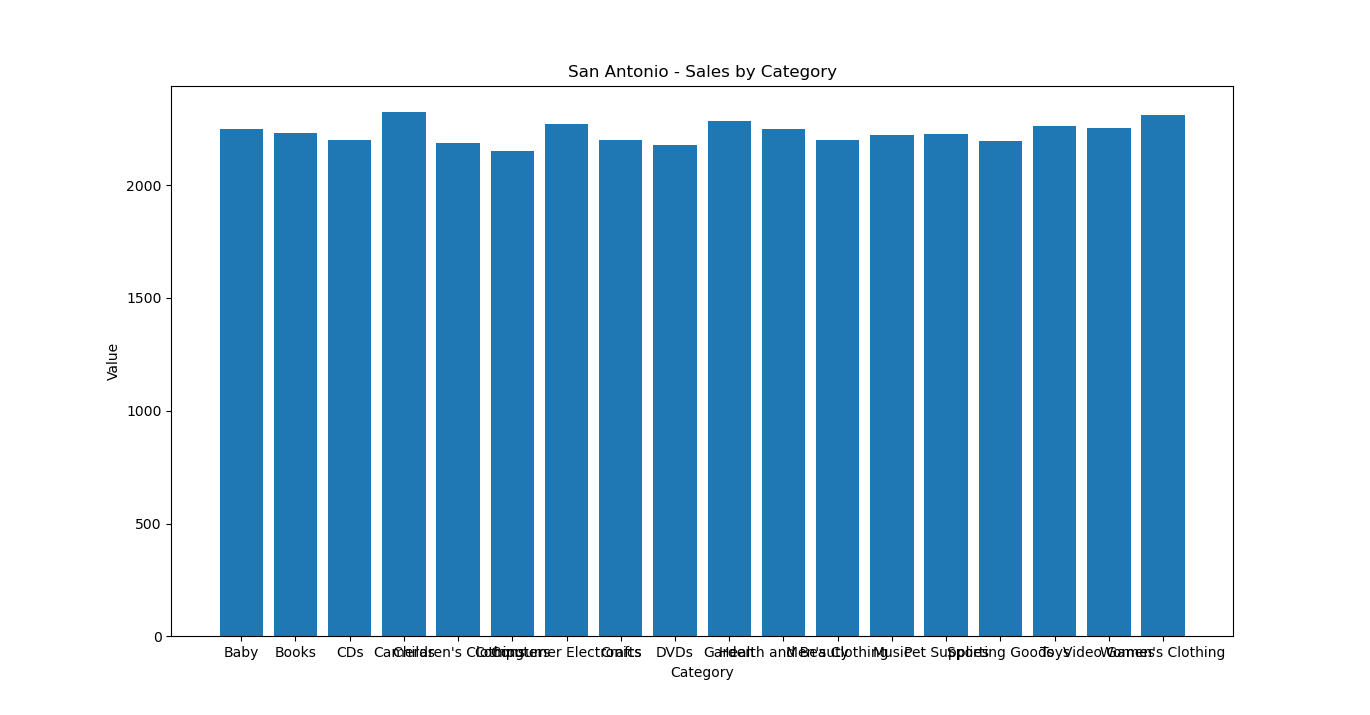
Les résultats peuvent être visualisés sous forme graphique. On peut par exemple tracer la distribution des ventes par catégories pour la ville de San Antonio :

Figure . Distribution des ventes par catégories pour la ville de San Antonio

Ou également, on peut tracer la distribution des ventes d’une catégorie (par exemple : books) sur toutes les villes :

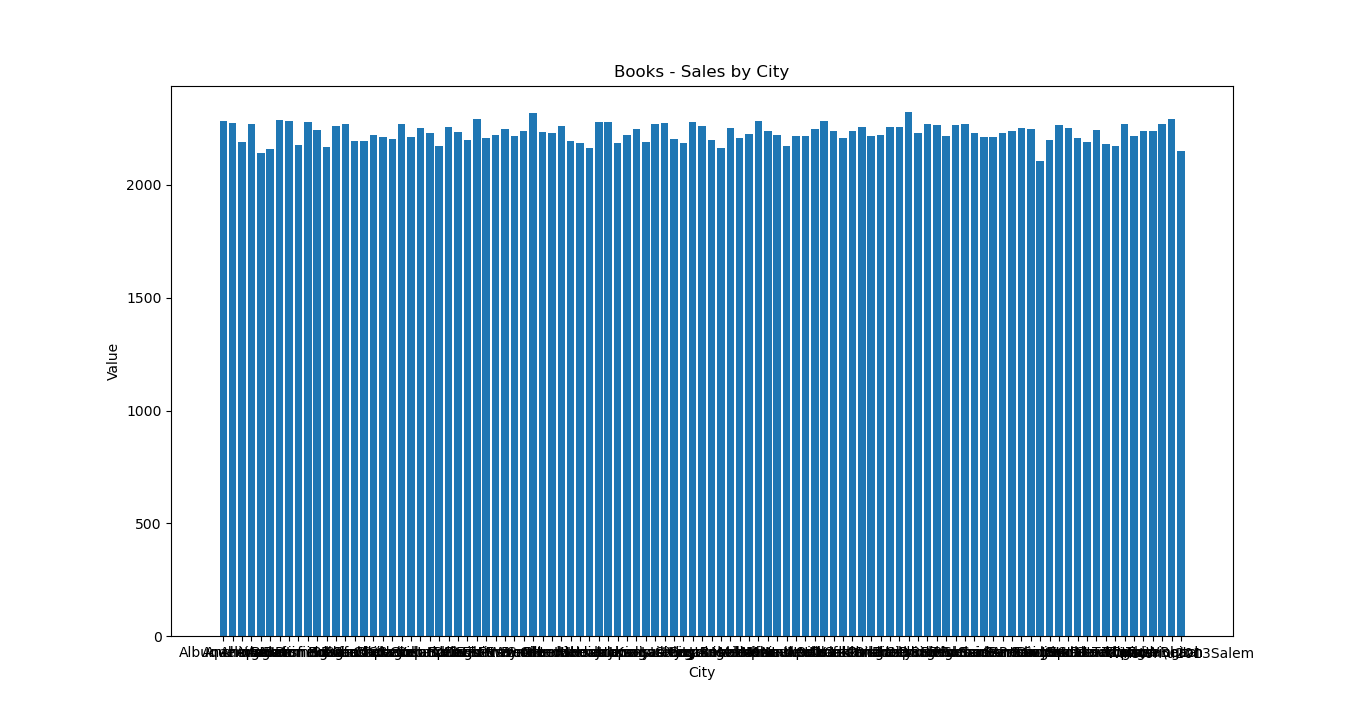


Figure .distribution des ventes pour books par ville

Ces graphes peuvent donner des informations très importants pour prendre des décisions stratégiques telles que les quantités nécessaires dans chaque ville, pour chaque catégorie.

**Conclusion :**

En conclusion, ce TP sur Hadoop/MRJobs a été une immersion enrichissante dans l'univers du traitement de données massives. L'évaluation de la librairie MRJobs, avec ses différentes configurations pour le comptage de mots dans un livre, a posé les bases pour des expériences pratiques plus approfondies. L'exercice 1 axé sur l'analyse du fichier de ventes été fructueux grâce à l'efficacité de MRJobs. Les conseils pratiques pour la préparation des données et la programmation des scripts MapReduce ont simplifié le processus de développement, tandis que l'approche progressive du débogage avec des extraits de données a été judicieuse. La diversité des questions traitées a permis une exploration complète des fonctionnalités de MRJobs, et les résultats ont été présentés de manière détaillée, renforçant notre compréhension des solutions développées. En fin, ce TP a consolidé nos compétences en programmation MapReduce avec MRJobs tout en soulignant l'importance cruciale de la préparation des données et de l'adaptation des solutions, offrant ainsi une perspective concrète sur l'application pratique des outils de traitement de données massives dans le domaine du Big Data.