21/12/2024

BELASSEL Meryem
NICOLLE Thomas

Rapport d'analyse : Exercice 2

Introduction

Le but de cet exercice est de développer des programmes MapReduce en Java pour analyser des données issues du site internet Amazon. Les trois requêtes traitées sont les suivantes :

1. Calculer pour chaque catégorie de produit le nombre de commentaires et la moyenne des étoiles attribuées.
2. Ordonner ces résultats par nombre de commentaires, dans un ordre décroissant.
3. Identifier le type des 50 produits les plus achetés parmi les clients ayant acheté au moins 10 articles.

Build et run les programmes

Il faut tout d'abord définir 4 variables d'environnement :

```
export JAVA_HOME=/usr/gide/jdk-1.8
export PATH=${JAVA_HOME}/bin:${PATH}
export PDSH_RCMD_TYPE=ssh
export HADOOP_CLASSPATH=${JAVA_HOME}/lib/tools.jar
```

Mettre les fichiers AmazonReview*.tsv dans Hadoop :

```
bin/hdfs dfs -put
/home4/tn837970/Documents/M1/Systeme_distribue/TP_Note/AmazonRev/*tsv
/user/tn837970/input
```

Compiler notre programme :

```
bin/hadoop com.sun.tools.javac.Main ExoX.java
```

Créer le .jar :

```
jar cf ExoX.jar ExoX*.class
```

Exécuter le programme avec Hadoop :

```
bin/hadoop jar ExoX.jar ExoX /user/tn837970/input /user/tn837970/output
```

21/12/2024

BELASSEL Meryem
NICOLLE Thomas

Requête 1 : Calcul du nombre de commentaires et de la moyenne des étoiles par catégorie

Fonctionnement du Mapper

Le Mapper extrait les informations suivantes des entrées :

- Catégorie du produit (« product_category »).
- Contenu du commentaire (« review_body »).
- Note (« star_rating »).

Pour chaque entrée, le Mapper produit :

- Une clé : *product_category*.
- Une valeur :
    - *1* pour compter un commentaire, 0 si aucun commentaire.
    - *rating:X* pour indiquer une note, où *X* correspond à la note des étoiles.

Fonctionnement du Reducer

Le Reducer reçoit les pairs clef/valeur du mapper et effectue les calculs suivants :

1. La somme des étoiles pour calculer la moyenne.
2. Le total des commentaires.

Résultat :

Nous obtenons un résultat en 7 minutes environ, cela inclut le processus de map et de reduce. C'est un temps raisonnable compte tenu de la quantité de données en entrée (~51Go).

- Clé : *product_category*.
- Valeur : Nombre de commentaires et moyenne des étoiles.

```
Apparel average rating: 4.105233511471771       number of comments: 5906455
Automotive      average rating: 4.246323856774859       number of comments: 3515165
Baby    average rating: 4.163213810967948       number of comments: 1758834
Beauty  average rating: 4.187224158954813       number of comments: 5115718
Books   average rating: 4.208745941316642       number of comments: 3944321
Camera  average rating: 4.128970967897998       number of comments: 1818952
Digital_Ebook_Purchase  average rating: 4.262505033002482       number of comments: 6350583
Digital_Music_Purchase  average rating: 4.63857186586893        number of comments: 1796750
Digital_Software        average rating: 3.5393303553935973      number of comments: 102084
Digital_Video_Download  average rating: 4.209602709078981       number of comments: 5115244
Digital_Video_Games     average rating: 3.8531262248076406      number of comments: 145431
Electronics     average rating: 4.035706373046583       number of comments: 3105328
Furniture       average rating: 4.083950558058681       number of comments: 792121
Gift Card       average rating: 4.731363105858364       number of comments: 149086
Grocery average rating: 4.312220392628272       number of comments: 2402476
Health & Personal Care  average rating: 4.1617661068312914      number of comments: 5332520
Home    average rating: 4.052316890881913       number of comments: 2007
```

BELASSEL Meryem
NICOLLE Thomas

Requête 2 : Classement par nombre de commentaires

Fonctionnement du Mapper

Le comportement est identique à celui de la Requête 1. La différence principale réside dans la logique de tri qui est gérée par le Reducer.

Fonctionnement du Reducer

Les données sont organisées dans une *TreeMap*, triée par ordre décroissant du nombre de commentaires.

Résultat :

Nous obtenons aussi un résultat en 7 minutes, malgré le trie effectué avec la *TreeMap* une fois que le reduce a été terminé.

- Clé : null (le tri est fait dans la valeur).
- Valeur : Les résultats ordonnés.

```
Wireless        Number of comments: 9024783     Average rating: 3.8921322540386845
PC      Number of comments: 6965595     Average rating: 4.087352910986068
Mobile_Apps     Number of comments: 6507959     Average rating: 4.033718405417121
Digital_Ebook_Purchase  Number of comments: 6350583     Average rating: 4.262505033002482
Video DVD       Number of comments: 6166026     Average rating: 4.31263361523289
Apparel Number of comments: 5906455     Average rating: 4.105233511471771
Music   Number of comments: 5530282     Average rating: 4.435110723105983
Health & Personal Care  Number of comments: 5332520     Average rating: 4.1617661068312914
Beauty  Number of comments: 5115718     Average rating: 4.187224158954813
Digital_Video_Download  Number of comments: 5115244     Average rating: 4.209602709078981
Toys    Number of comments: 4922026     Average rating: 4.214570382196275
Sports  Number of comments: 4854496     Average rating: 4.229267672689399
Shoes   Number of comments: 4374299     Average rating: 4.241352500137737
Books   Number of comments: 3944321     Average rating: 4.208745941316642
Automotive      Number of comments: 3515165     Average rating: 4.246323856774859
Electronics     Number of comments: 3105328     Average rating: 4.035706373046583
Office Products Number of comments: 2644747     Average rating: 4.072483303695968
Pet Products    Number of comments: 2643624     Average rating: 4.143652803878313
Grocery Number of comments: 2402476     Average rating: 4.312220392628272
Outdoors        Number of comments: 2305596     Average rating: 4.239963115827751
```

BELASSEL Meryem
NICOLLE Thomas

Requête 3 : Identification des 50 types de produits les plus achetés

Fonctionnement du Mapper

Le Mapper génère des paires client-produit pour chaque entrée :

- **Clé :** *customer_id*.
- **Valeur :** *product:X*, où *X* représente la catégorie de produit achetée.

Fonctionnement du Reducer

Le Reducer :

1. Compte le nombre total de produits achetés par chaque client.
2. Il s'intéresse uniquement aux clients ayant acheté au moins 10 articles.
3. Agrège ces données par catégorie de produit pour identifier les 50 types de produits les plus populaires.

Les données sont triées par ordre décroissant du nombre d'achats.

Résultat :

Nous obtenons encore un résultat en 7 minutes, ce qui est relativement rapide malgré plusieurs tri (un pour les clients ayant acheté au moins 10 articles et un pour identifier les 50 types de produits les plus populaires).

- Clé : Category.
- Valeur : Quantity.

```
Video DVD       3895992
Wireless        3637736
PC      3271938
Music   3146392
Apparel 2938579
Digital_Ebook_Purchase  2916316
Health & Personal Care  2768009
Beauty  2624735
Mobile_Apps     2572389
Digital_Video_Download  2556047
Toys    2525969
Sports  2385558
Shoes   1875500
Automotive      1771973
Books   1643434
Electronics     1498136
Pet Products    1460136
```

BELASSEL Meryem
NICOLLE Thomas

Problématiques rencontrées et optimisations

Gestion des entrées

- Les premières lignes (en-têtes) sont ignorées.
- Les valeurs nulles ou incorrectes sont filtrées pour prévenir les erreurs d'exécution.

Performances

- L'exécution parallèle via Hadoop permet d'accélérer le traitement des volumes importants de données.
- L'utilisation de structures comme *TreeMap* optimise le tri en mémoire.

Limites

- La mémoire peut être un facteur limitant en cas de traitement de fichiers très volumineux. Cela est arrivé que le processus de map-reduce soit interrompu à cause d'un manque de mémoire tampon.

Conclusion

Les programmes développés répondent aux trois objectifs de l'exercice. L'approche MapReduce permet de traiter efficacement des données massives.

Pour aller plus loin, il serait intéressant d'intégrer des outils de visualisation pour présenter les résultats de manière plus intuitive avec des graphique par exemple, ainsi que de mettre en place des mécanismes d'optimisation lié à la volumétrie des données comme par exemple avec l'utilisation d'un combiner pour accélérer le traitement par le reducer ou encore l'utilisation de plusieurs noeuds pour répartir la charge de stockage.