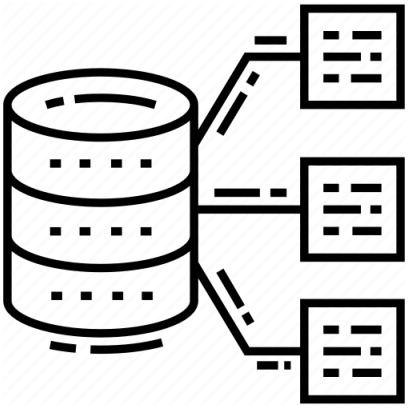
DEPARTEMENT INFORMATIQUE
DE LA FACULTE DES SCIENCES

Quentin Yeche (21520370), Yanis Allouch (21708237)

Rapport du TP Final : Hadoop & Map-Reduce



HMIN122M — Entrepôts de données et Big-Data

Référent: Federico Ulliana et Anne-Muriel Chifolleau

2020

Table des matières

Introduction

Nous étudions et proposons des implémentations en Map/Reduce des requêtes proposées pour notre projet d'entrepôts de données pour l'entreprise Sanofi.
L'environnement de travail est composé de Eclipse qui est utilisé pour le développement et traitement des solutions sous Windows 10 et Ubuntu 20.04 LTS.

Le patch "Hadoop for Windows" proposé sur Moodle à du être appliqué pour le développement Windows.

Ce dernier travail noté va juger de nôtre maîtrise de MapReduce et à pouvoir implémenter une technique de hachage non introduit en cours mais expliqué lors des séances de TP.

Par ailleurs, tout du long de cette seconde partie du semestre nous nous sommes instruits et inspirées du livre faisant foi dans la matière [Whi15], sans compter l'article de recherche ayant fait office d'introduction au MapReduce dans notre cas, publier par Google [DG] ou encore l'article de Cloudera & Teradata [AG12] et enfin le tutoriel quand à l'utilisation de Spark pour notre culture personnel [Tut].

1 Jeu de données

Nous avons utilisé Mockaroo pour créer des données au format CSV adaptées à notre entrepôt de données. Nous avons gardé les fichiers petits en nombre de ligne afin de faciliter le développement.

Voici les fichiers dont nous disposons :
- Tables de faits :
  - Productions.csv (57 lignes)
  - Stocks.csv (30 lignes)
- Dimensions :
  - Produits.csv (10 lignes)
  - Lieux.csv (10 lignes)
  - Dates.csv (20 lignes)

2 Requêtes

Voici les requêtes que nous traitons par la suite.

1. Quels sont les chiffres d'affaires par produit et par lieu ?
2. Quels sont les chiffres d'affaires par produit et par mois
3. Quels sont les chiffres d'affaires par produit et par semaine ?
4. Quels sont les continents qui génèrent les plus gros chiffres d'affaires ?
5. Quels sont les produits dont les stocks connaissent de fortes variations pour une période donnée ?
6. Quels produits, relativement à leur demande moyenne, sont en pénurie ou proche de l'être, par continent ?
7. Quels sont les entrepôts ayant le plus grande nombre de produits différents en stock ?
8. Quelle est le nombre moyen de livraisons par semaine et par continents ?
9. Quel est le taux d'incidence de pertes ?
10. Quelle est la rentabilité de chaque produit pour chaque continent ?

3 Implémentation

On a définis des formats d'entrée/sortie implémentant l'interface *Writable-Comparable* quand cela se trouvait nécessaire ou par commodité.

Les types IN sont *StockKey* et *ProductionsKey* qui peut aussi revêtir le nom de *CompositeKey* qui nous vient des séances de TP, modulo quelques attributs de classe et la logique dans les méthodes *compareTo(), toString()*. Modélisant respectivement la clé pour la table Stocks et Productions.

On ne retrouve qu'un type OUT, *StockValue*.

3.1 Requête 1

Quels sont les chiffres d'affaires par produit et par lieu ?

Le programme est composé de un job constitué de un Map et un Reduce. Le Map consiste à grouper l'identifiant pertinent à la question en clé et de passer tout le reste en valeur d'output du Map.

Le Reduce somme les valeurs associés à la clé d'un produit.

Nous avons décidé de commencer léger, sans jointure, sans travail complexe. Cette première requête s'apprêtant au première requête élaboré dans les premiers TP.

3.2 Requête 2

Quels sont les chiffres d'affaires par produit et par mois ?

On peux voir le problème comme deux tâches de M/R chaînées :

1. group by. Il s'agit de regrouper et de sommer les chiffres d'affaire par produit et par date.
2. join + group by. Il s'agit d'effectuer une jointure avec les dates afin de récupérer le mois et de pouvoir ainsi group by et sommer sur <produit,mois>.

3.3 Requête 3

Quels sont les chiffres d'affaires par produit et par semaine ?

Nous avons deux deux tâches :

1. jointure pour récupérer le numéro de la semaine
2. group by sur <produit,semaine>. On somme les montants de vente.

3.4 Requête 4

Quels sont les continents qui génèrent les plus gros chiffres d'affaires ?

Nous avons trois tâches

1. group by sur l'id de lieu
2. jointure sur les lieux. On écrit en sortie le continent
3. group by sur le continent. On somme les chiffres d'affaires

3.5 Requête 5

Quels sont les produits dont les stocks connaissent de fortes variations pour une période donnée ?

Une seule tâche : group by sur l'id produit, on calcule le quotient (quantité sortie)/(stock). On fait bien attention à ne sélectionner que la période voulue. Dans le reduce on fait la moyenne sur le nombre de valeurs.

3.6 Requête 6

Quels produits, relativement à leur demande moyenne, sont en pénurie ou proche de l'être, par continent ?

Nous avons trois tâches :

1. un group by calculant la demande historique par produit et par lieu. Il s'agit de la moyenne de la quantité sortante. Dans le reduce on écrit également le nombre de valeurs qui ont contribué à cette moyenne
2. jointure intermédiaire sur le lieu et les produits ainsi que les moyennes de produits
3. le reduce s'occupe de faire correspondre les clés et identifiant de la table.

3.7 Requête 7

Quels sont les entrepôts ayant le plus grande nombre de produits différents en stock ?

Une seule tâche : group by entrepôt. Dans le map on sélectionne seulement la date souhaitée. On envoie 1 par produit différent (similaire à wordcount).

3.8 Requête 8

Quelle est le nombre moyen de livraisons par semaine et par continents ?

Nous avons 5 tâches chaînées :

1. group by <lieu,date>. Il s'agit d'un simple wordcount. Il n'est pas strictement nécessaire mais permet de facilement réduire la quantité de données transmise.
2. Jointure + agrégation. La jointure se fait sur le lieu pour obtenir le continent. En sortie de reduce on a <continent,date,somme sur le continent>.
3. Jointure. Cette jointure est faite sur les dates pour obtenir la semaine.
4. group by <semaine,continent>. On calcule simplement la somme.
5. group by continent. On fait la moyenne sur les semaines afin d'avoir le résultat attendu.

3.9 Requête 9

Quel est le taux d'incidence de pertes ?

On définit le taux d'incidence de pertes comme la valeur de (quantité perdue)/(quantité vendue). Une seule tâche suffit :

— Dans le map on transmet <produit,quantite> si le type d'opération correspond (vente ou perte). Il faudra distinguer dans quantité s'il s'agit d'une perte ou d'une vente (nous avons utilisé des entiers positifs ou négatifs pour cela).
— Dans le reduce on somme les quantités perdues et vendues séparément, et on écrit le quotient en sortie.

3.10 Requête 10

Quelle est la rentabilité de chaque produit pour chaque continent ?

Nous définissons la rentabilité comme le rapport des ventes sur les coûts de production. Nous avons besoin de trois tâches :

1. group by sur <produit,lieu>. On somme les coûts d'un côté, les ventes de l'autre.
2. Jointure sur le lieu et le produit. On extrait l'information sur le continent et le nom du produit.
3. troisième et dernier job qui consiste a faire correspondre les identifiant au valeurs textuelles.

4 Hash Join

Nous avons choisi de nous concentrer sur deux requêtes pour le hash join : les requêtes 8 et 10.

4.1 Requête 8

Quelle est le nombre moyen de livraisons par semaine et par continents ?

Le hash join nous permet d'effectuer plusieurs jointures à la fois. Les tâches 2 et 3 peuvent donc être combinées en une seule tâche qui fera la jointure de la sortie de tâche 1 avec les données de Lieux et Dates. Dans le map, comme dans une jointure classique, on doit différencier selon le fichier. Pour une ligne de la sortie de la tâche 1 on peut transmettre la clé composite des images par les fonctions de hachage. Pour une ligne de Lieux ou Dates nous devons envoyer à toutes les valeurs possibles correspondant au partitionnement de la clé manquante.

4.2 Requête 10

Quelle est la rentabilité de chaque produit pour chaque continent ?

De manière analogue a l'analyse précédente, les tâches 2 et 3 peuvent donc être combinées en une seule tâche qui fera la jointure de la sortie de tâche 1 avec les données de Lieux et Produits.

5 Conclusion

Les observations des counters affichés dans la console et du temps d'exécution, nous indiquent aucun changement notables entre une requête join et une requête join utilisant la technique de hachage.

Nos recherches suppose que nous ne possédons ni le matériel, ni les données pour pouvoir identifier les gains de performances mentionnées dans le sujet.

Références

[AG12] Awadallah, A. et Graham, D. « Hadoop and the Data Warehouse When to Use Which ». In : *Marketing Teradata* (2012).

[DG] Dean, J. et Ghemawat, S. *MapReduce : Simplified Data Processing on Large Clusters*. URL : https://research.google/pubs/pub62/. (accès a une copie de 2018).

[Tut] Tutorialspoint. *Apache Spark - Quick Guide*. URL : https://www.tutorialspoint.com/apache_spark/apache_spark_quick_guide.htm. (accès a une copie de 2018).

[Whi15] White, T. *Hadoop : The Definitive Guide*. O'Reillyr, 2015.