TP Hadoop M2GI

Antoine Mercier--pronchery

Juan Pablo Cadavid

---

1 Prise en main

1.1 Exécution locale

1. Que signifie Map Input records ? Et map output records ?

Map Input records représente le nombre de ligne du fichier texte d'entré (vide ou avec texte).
Map output record représente le nombre de mots identifié dans le fichier texte d'entré.

2. Quel est le lien entre Map output records et reduce input records

Map output records étant le compte total de mot du document, reduce input record récupère cette taille pour les opérations du reduce.

3. Que signifie Reduce input groups

Reduce Input group est le nombre de groupe identifié au sein du document ou le nombre de mot unique du document texte.

1.2 Premier contact avec HDFS

Notre login sur hdfs est merciant.
Le chemin de mon dossier sur hdfs est : hdfs://NameNode:9000/user/merciant/

1.3 Exécution sur le cluster

On observe dans la trace 5 split, ce nombre correspond aux partition que hadoop a réalisé du fichier texte en entré.

1.4 Combiner

1. On peut voir que les champs Combine input records et Combine output records possède des valeurs, l'algorithme est donc passé par le combiner.
2. Si on le compare à l'exécution précédente on peut voir qu'il y a beaucoup moins de lecture et écriture. Aussi le nombre d'entré de reduce input record est beaucoup plus bas que lorsque le combiner était absent.
   Enfin on peut voir que le temps pour exécuter la tâche reduce a été réduit.
3. Une fois le jeu de données récupéré nous l'avons importé dans libreoffice et avons traité les résultats à l'aide de la fonction MAX (Line1:Line2) (Line1 étant la première ligne contenant les occurrence de mot et line2 le nombre maximum de ligne d'une feuille de calcul). On obtient alors que le mot le plus dit est répété 16 757 fois et avec un CTL+F de 16 757 on trouve qu'il s'agit de "de".
4. On peut voir qu'il y a trois output différents,puisque les combiner ont divisé le travail en trois sous tâche. Ainsi on obtient trois partition de l'occurrence des mots dans les misérables.

2 Top-tags Flickr par pays

2.1 Map et Reduce

Pour commencer chaque passage dans le map traite une nouvelle image FlickR.
J'ai donc stocké les différentes informations de ces entré dans un tableau, chaque attribut étant séparé par une tabulation.

Puisque les champs peuvent être nul je vérifie si la photo possède une longitude et une latitude avant d'essayer de trouver un pays correspondant.

Si un pays est trouvé, pour chacun des tag associé sur cette photo je crée un couple clé valeur avec le code du pays et le tag.

Une fois le map effectué, je comptabilise pour chaque pays le nombre d'occurence de chaque tag en parcourant la liste des couple clé valeurs et en incrémentant le compteur pour les tags déjà rencontré.

La librairie google MinMaxPriorityQueue ne stocke que les k tags les plus populaires, on insère donc un par un les tag et leurs quantité trié auparavant. Ne reste donc que les 3 tag les plus populaire par pays.

On finit par écrire dans le fichier ces trois tag les plus populaire pour ce fichier test. On obtient:

```
AG	3 الطوارق
AG	3 الهقار
AG	3 تمنراست
BN	ghana 7
BN	lab 5
BN	africa 2
ML	mali 15
ML	niger 11
ML	desierto 10
UV	africa 10
UV	burkina faso 9
UV	burkina-faso 9
```

2.2 Combiner

Pour pouvoir utiliser le combiner, le type des données intermédiaire devrait être “Text, StringAndInt”.
Le combiner prend en entrée un Text, Text (pays, tag) et fournira un Text, StringAndInt (pays, (tag, occurrence)).

Le combiner s’occupera désormais d’agréger les différents tag et leurs occurrences. Ainsi le travail du reducer sera seulement de récupérer la liste de tag d'occurrence, de l’ajouter au tableau MinMax et d’obtenir une sortie.

Les tags les plus utilisé en France sont :

Pays,	Tag,	Occurrence
FR,	france,	563
FR,	spain,	113
FR,	europe,	75
FR,	españa,	70
FR,	bretagne,	67

Dans le reducer, nous avons une structure en mémoire dont la taille dépend du nombre de tags distincts : on ne le connaît pas a priori, et il y en a potentiellement beaucoup. Est-ce un problème ?

Cela devrait normalement ne pas poser de problème, Hadoop est conçu pour supporter des données sur de grandes échelles et est capable de s'adapter à la taille du Dataset d’entré pour traiter les demandes d’opérations. (Scalabilité)

3 Top-tags Flickr par pays, avec mémoire limitée

Question préliminaire:
Le rôle du job 1 sera de rechercher les tag et les agréger dans la structure de donnée. Dans son reducer on fera le même travail que précédemment, c'est à dire totaliser les occurrences de tag par pays. Ce job est pratiquement inchangé par rapport à la question précédente.

Le rôle du job 2 sera de réaliser un tri sur les données entrante, ainsi il ne sera pas nécessaire de faire de traitement dans le map, le travail se fera surtout dans le reduce ou le job s'occupera du tri des données du job1.

3.1 Passes MapReduce en chaîne

La fonction reduce n'a pas à effectuer de transformation, c'est à dire elle n'as pas besoin de recomposer une MinMaxPriorityQueue, elle peut donc se concentrer pleinement sur le tri des données existantes.

Avec le découpage en groupe, en fonction du timing de l'arrivée des données, on pourrait avoir des résultats différents, les résultats ex aequo seront placé dans leurs ordres d'arrivés.