eBook Gratuit

APPRENEZ mapreduce

eBook gratuit non affilié créé à partir des **contributeurs de Stack Overflow.**

#mapreduc

e

Table des matières

À propos

You can share this PDF with anyone you feel could benefit from it, downloaded the latest version from: mapreduce

It is an unofficial and free mapreduce ebook created for educational purposes. All the content is extracted from Stack Overflow Documentation, which is written by many hardworking individuals at Stack Overflow. It is neither affiliated with Stack Overflow nor official mapreduce.

The content is released under Creative Commons BY-SA, and the list of contributors to each chapter are provided in the credits section at the end of this book. Images may be copyright of their respective owners unless otherwise specified. All trademarks and registered trademarks are the property of their respective company owners.

Use the content presented in this book at your own risk; it is not guaranteed to be correct nor accurate, please send your feedback and corrections to info@zzzprojects.com

Chapitre 1: Démarrer avec mapreduce

Remarques

Cette section fournit une vue d'ensemble de ce qu'est mapreduce et pourquoi un développeur peut vouloir l'utiliser.

Il devrait également mentionner tous les grands sujets dans mapreduce, et établir un lien avec les sujets connexes. La documentation de mapreduce étant nouvelle, vous devrez peut-être créer des versions initiales de ces rubriques connexes.

Examples

Installation ou configuration

Mapreduce fait partie de Hadoop. Ainsi, lorsque Apache Hadoop (ou toute distribution de Hadoop est installée), MR est automatiquement installé.

MapReduce est la structure de traitement de données sur HDFS (système de fichiers distribué Hadoop). Les travaux MR peuvent être écrits en utilisant Java, Python, Scala, R, etc.

Qu'est-ce que mapreduce fait et comment?

Mapreduce est un modèle de programmation permettant de traiter des (très) grandes quantités de données.

Le «HPC» traditionnel (High Performance Computing) accélère les gros calculs sur des quantités de données relativement importantes en créant un ensemble d'ordinateurs hautement connectés (utilisant des fonctions réseau extrêmement rapides et un accès rapide au stockage partagé, mémoire partagée) pour gérer les problèmes informatiques. nécessitent généralement des calculs pour avoir accès aux données des autres. Un exemple classique est la prévision météorologique.

Mapreduce, quant à lui, excelle dans le traitement de calculs relativement petits et indépendants sur d'énormes quantités de données. Pour rendre cela possible, les données sont réparties sur plusieurs ordinateurs (en raison de la quantité de données), et le calcul souhaité est divisé en une phase qui peut être effectuée sur chaque bit de données indépendamment (la phase «carte»). Les résultats de ces calculs indépendants sont ensuite rassemblés et une seconde partie des calculs est effectuée pour combiner tous ces résultats individuels dans le résultat final (la phase de «réduction»).

Exemple: Compter les votes

Imaginez que vous ayez un très grand nombre de votes pour compter, et que vous ayez un peu

de travail pour compter chaque vote (par exemple, à partir de l'image numérisée, quelle case a été cochée).

Dans ce cas, une implémentation de mapreduce:

Etape 1: 'Spread'

Répartissez les images sur les ordinateurs disponibles.

Etape 2: 'Carte'

Sur chaque ordinateur, pour chaque image:

- prendre 1 des images copiées sur cet ordinateur en entrée
- savoir quelle case a été cochée
- sortir le numéro (ou code ou nom) de l'article voté pour

Notez que le travail peut démarrer dès qu'un ordinateur obtient 1 image à travailler. Il n'est pas nécessaire que tous ces ordinateurs interagissent pour effectuer leur travail. Il n'est donc pas nécessaire de les interconnecter rapidement, d'avoir une mémoire partagée ou un espace disque partagé.

Étape 3: "Rassembler"

Rassemblez toutes ces sorties sur 1 ordinateur.

Étape 4: Réduire

Comptez le nombre de votes pour chaque numéro (ou code ou nom).

Cet exemple très basique montre également comment d'autres optimisations sont souvent possibles. Dans ce cas, l'étape de réduction proprement dite peut clairement être effectuée partiellement sur chaque ordinateur, puis une réduction finale peut être effectuée sur un ordinateur central. Cela réduira la quantité de travail sur l'ordinateur exécutant l'étape de réduction et limitera la quantité de données à transporter sur le réseau.

Exemple: Compter les votes - optimisé (en utilisant le combineur)

Etape 1: 'Spread'

Comme précédemment: répartissez les images sur les ordinateurs disponibles.

Etape 2: 'Carte'

Comme précédemment: Sur chaque ordinateur, pour chaque image:

- prendre 1 des images copiées sur cet ordinateur en entrée
- savoir quelle case a été cochée
- sortir le numéro (ou code ou nom) de l'article voté pour

Étape 3: «Rassembler» localement

Rassemblez toutes les sorties d'un ordinateur sur l'ordinateur lui-même.

Étape 4: Réduire localement

Comptez le nombre de votes de chaque nombre (ou code ou nom) dans les résultats locaux et générez ces comptes.

Étape 5: «Rassembler» globalement

Rassemblez toutes les sorties du local réduit sur 1 ordinateur.

Étape 6: "Réduire" globalement

Résumez les comptes de votes faits localement pour chaque nombre (ou code ou nom).

Notez qu'à l'étape 3, il n'est *pas nécessaire* d'attendre **tous les** résultats dans les cas suivants:

- si cela devient trop pour les ordinateurs ressources locales comme le stockage / la mémoire
- si le coût du travail à refaire lorsqu'un ordinateur tombe en panne est jugé trop long pour attendre tous les résultats locaux
- si le réseau est maintenant libre de transporter des résultats intermédiaires

le rassemblement local et la réduction locale peuvent être effectués sur les résultats obtenus jusqu'à présent sur l'ordinateur local, et cela peut être fait à tout moment.

L'étape de réduction locale est appelée étape de combinaison. Ceci est une étape facultative utilisée pour améliorer les performances.

Lire Démarrer avec mapreduce en ligne: https://riptutorial.com/fr/mapreduce/topic/2192/demarrer-avec-mapreduce

Crédits

S. No	Chapitres	Contributeurs
1	Démarrer avec mapreduce	Ani Menon, Community, Legolas