MapReduce

MapReduce for $200

- Nous voulons développer un programme pour compter les mots de milliers de documents organisés dans des dossiers avec la première lettre du nom du fichier («A», «B», «C»…). Vous avez 27 travailleurs. Décrivez l'implémentation MapReduce pour ce problème.

MapReduce for $200

- Nous voulons développer un programme pour compter les mots de milliers de documents organisés dans des dossiers avec la première lettre du nom du fichier («A», «B», «C»…). Vous avez 27 travailleurs. Décrivez l'implémentation MapReduce pour ce problème.

1) Envoyez un dossier à chaque travailleur.
    1) Faites attention! Tous les dossiers ne contiennent pas la même quantité de fichiers! Vous pouvez attribuer une taille appropriée aux travailleurs ou utiliser moins de travailleurs et envoyer certaines des lettres les moins populaires à un seul travailleur.
2) Les tâches map comptent les mots des fichiers par fichier.
3) Les tâches reduce somment les résultats par dossier et pour tout le système de fichiers.
    1) Vous pouvez avoir plusieurs niveaux de tâches reduce et l'entrée d'une tâche reduce peut être la sortie d'une tâche reduce précédente.

MapReduce for $400

- Nous avons la base de données de l'ARC pour l'impôt sur le revenu. L'ARC a plusieurs serveurs de données à travers le pays. Nous voulons savoir qui a payé le plus d'impôts en 2019. Décrivez l'implémentation de MapReduce pour ce problème.

MapReduce for $400

- Nous avons la base de données de l'ARC pour l'impôt sur le revenu. L'ARC a plusieurs serveurs de données à travers le pays. Nous voulons savoir qui a payé le plus d'impôts en 2019. Décrivez l'implémentation de MapReduce pour ce problème.

1. Chaque serveur de données peut exécuter une réplique de l'algorithme maximal.
2. Les tâches map consistent à trouver le maximum dans un seul serveur.
3. Les tâches reduce consistent à trouver le maximum par province et une autre tâche reduce pour trouver le maximum global du pays.

MapReduce for $600

- Nous voulons trouver le joueur avec la plus grande moyenne de buts par match dans toute l'histoire de la NHL. Chaque équipe conserve ses propres statistiques. (Ne considérez pas les équipes qui n'existent pas actuellement). Décrivez l'implémentation MapReduce pour ce problème.

MapReduce for $600

- Nous voulons trouver le joueur avec la plus grande moyenne de buts par match dans toute l'histoire de la NHL. Chaque équipe conserve ses propres statistiques. (Ne considérez pas les équipes qui n'existent pas actuellement). Décrivez l'implémentation MapReduce pour ce problème.

1. Le serveur de données de chaque équipe exécutera une réplique de l'algorithme.
2. Les tâches map consistent à trouver la moyenne par équipe.
3. Les tâches reduce trouveront le maximum sur les moyennes renvoyées.
4. Et la réponse est… Mike Bossy (0,762) des New York Islanders (1977-1987).

MapReduce for $800

- Chaque université détient toutes les thèses soutenues dans leur bibliothèque respective. Nous souhaitons rechercher dans le monde entier des thèses contenant le terme «DevOps» dans leur titre. Décrivez l'implémentation MapReduce pour ce problème.

MapReduce for $800

- Chaque université détient toutes les thèses soutenues dans leur bibliothèque respective. Nous souhaitons rechercher dans le monde entier des thèses contenant le terme «DevOps» dans leur titre. Décrivez l'implémentation MapReduce pour ce problème.

1. Le serveur de données de chaque université exécutera une réplique de l'algorithme grep / search.
2. Les tâches map sont les recherches avec le string «DevOps» sur le titre.
3. Les tâches reduce regrouperont simplement les résultats individuels des tâches map dans une liste.

MapReduce for $1000

- Nous voulons une liste complète de toutes les villes, villages et communautés du monde classés par population par ordre décroissant. Les serveurs de recensement de chaque pays contiennent des données démographiques. Décrivez l'implémentation MapReduce pour ce problème.

MapReduce for $1000

- Nous voulons une liste complète de toutes les villes, villages et communautés du monde classés par population par ordre décroissant. Les serveurs de recensement de chaque pays contiennent des données démographiques. Décrivez l'implémentation MapReduce pour ce problème.

1. Chaque serveur de recensement peut exécuter une réplique de l'algorithme de tri.
2. Les tâches map stockeront les villes par pays.
3. La tâche reduce agrégera la liste totale par ordre décroissant.


DAILY
DOUBLE

MapReduce for $2000

- Nous voulons une liste complète de toutes les villes, villages et communautés du monde classés par population par ordre décroissant. Les serveurs de recensement de chaque pays contiennent des données démographiques. Décrivez l'implémentation MapReduce pour ce problème.
- Question bonus: quel est l'algorithme de tri le plus adapté à ce cas?

MapReduce for $2000

- Nous voulons une liste complète de toutes les villes, villages et communautés du monde classés par population par ordre décroissant. Les serveurs de recensement de chaque pays contiennent des données démographiques. Décrivez l'implémentation MapReduce pour ce problème.
- Question bonus: quel est l'algorithme de tri le plus adapté à ce cas?
- “What is…” mergesort
- En fait, les RDD (ou formats de données similaires) sont des paires clé-valeur, ce qui signifie que le tri peut être très efficace et simple sur les clés.
- Si nous voulons trier par valeur, une solution est avec des «clés composées», où nous pouvons combiner la clé primaire avec la colonne que nous voulons trier, puis trier automatiquement par clé.


NoSQL
JEOPARDY!

NoSQL for $200

- Nous développons une application pour une entreprise commerciale. L'entreprise gère les contrats et les factures des transactions commerciales. Les données doivent être stockées de manière sécurisée et persistante pendant une longue période. L'entreprise fonctionne au niveau mondial avec un grand nombre de transactions.

NoSQL for $200

- Nous développons une application pour une entreprise commerciale. L'entreprise gère les contrats et les factures des transactions commerciales. Les données doivent être stockées de manière sécurisée et persistante pendant une longue période. L'entreprise fonctionne au niveau mondial avec un grand nombre de transactions.
- “What is …”: Document
- Les contrats et les factures sont des fichiers.
- Nous devrions pouvoir récupérer les documents en utilisant des clés ou en interrogeant leurs attributs.
- La quantité de données nécessite une solution NoSQL.
- Autre solution : Wide-Column
  - Par rapport à la sécurité.

NoSQL for $400

- Wikipedia! Développez Wikipedia!

NoSQL for $400

- Wikipedia! Développez Wikipedia!
- “What is …”: Document?
  - Une possibilité
  - On peut traiter chaque page en tant que document
  - Donc, on peut récupérer une page de sa clé ou de ses métadonnées.
- “What is …”: graph
  - Une meilleure solution.
  - Il y a des liens entre les pages. Les liens peuvent aussi représenter des relations complexes (hiérarchie, taxonomie, partonomie).
  - En fait, on a un réseau des pages.

NoSQL for $600

- Nous avons un système de monitoring des ressources cloud. À des intervalles assez fréquents et rapides, le système envoie des mesures (CPU, mémoire, disque, réseau) pour chaque ressource (machine virtuelle).

NoSQL for $600

- Nous avons un système de monitoring des ressources cloud. À des intervalles assez fréquents et rapides, le système envoie des mesures (CPU, mémoire, disque, réseau) pour chaque ressource (machine virtuelle).
- “What is …”: Key-value
- La structure des données est assez simple.
- Nous avons besoin d'une efficacité accrue.
- Il est possible de profiter de la mémoire pour une ingestion de données rapide et efficace.

NoSQL for $800

- Pour le système de monitorage precedent, quelle architecture de traitement des données choisiriez-vous?

NoSQL for $800

- Pour le système de monitorage précèdent, quelle architecture de traitement des données choisiriez-vous?
- “What is …”: Kappa
  - Si on avait seulement l’ingestion des données.
  - Si on utilise une méthode « push » : les ressources envoient leurs mesures au système de monitorage.
- Mais! Si nous avons aussi des analyses au données du monitorage ou c’est le système qui demande les mesures des ressources (méthode « pull »), nous avons des problèmes.
  - On a « l’effet de l’observateur » : en demandant les mesures, le monitorage affecte les mesures elles-mêmes (parce qu’on exécute du code supplémentaire).
  - Dans ce cas, il est mieux d’utiliser l’architecture Lambda.

NoSQL for $1000

- Les bases de données pour les données biologiques existent depuis un certain temps déjà. Ils contiennent des données sur les gènes, les protéines, les organismes. Les entités ont des attributs, mais il est possible de découvrir de nouveaux attributs dans le futur. Diverses analyses et outils existent déjà pour nous aider à étudier le monde naturel.

NoSQL for $1000

- Les bases de données pour les données biologiques existent depuis un certain temps déjà. Ils contiennent des données sur les gènes, les protéines, les organismes. Les entités ont des attributs, mais il est possible de découvrir de nouveaux attributs dans le futur. Diverses analyses et outils existent déjà pour nous aider à étudier le monde naturel.
- “What is …”: Wide-column
- La flexibilité de la structure est réquise.
- Il est possible que des bases des données rélationnelles existent déjà.
- Il est certain que des clients de ces bases de données existent déjà, qui assume l’existence d’un schema.