Oefeningen MapReduce

Versie 20250425

1. Slide 7: hoe kunnen we een gedeelte van het werk verschuiven van Reduce naar Map? In hoeverre is er sprake van performance-verbetering?
2. Schrijf pseudocode voor Map en Reduce voor een collectie tupels van de vorm <g,v> die de volgende SQL query representeert:

   SELECT g, SUM(v) FROM Input
   WHERE v >= 100
   GROUP BY g
   HAVING SUM(v) >= 10000
3. Gebruik M/R om de natural join van R(A,B) en S(B,C) te berekenen. Ga ervan uit dat de tuples de vorm <T,<x,y>> hebben, met T = R of T = S.
4. Beschrijf de constructie van een inverted index met behulp van M/R. Ga ervan uit dat elk document in zijn geheel door één Map-worker verwerkt wordt.

   Voorbeeld input

   <*doc1,* daar sta je dan je zag dit moment al zo vaak in je dromen>

   <*doc2:* daar gaat ze en zoveel schoonheid heb ik nooit verdiend>

   Voorbeeld output (niet gesorteerd op term)

   <al, [doc1]>
   <daar, [doc1, doc2]>
   …
5. Breid de inverted index van 4 uit met positionele informatie:

   Voorbeeld output (niet gesorteerd op term)

   <je, [<doc1, [3, 5, 13]>]>
   <daar, [<doc1,[1]>, <doc2,[1]>]>