

Oefeningen MapReduce

Versie 20190416

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 $\langle g, v \rangle$ 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 $\langle T, \langle x, y \rangle \rangle$ hebben, met $T = R$ of $T = S$.
4. Beschrijf de constructie van een inverted index met behulp van M/R.

Voorbeeld input

$\langle doc1, \text{daar sta je dan je zag dit moment al zo vaak in je dromen} \rangle$

$\langle doc2, \text{daar gaat ze en zoveel schoonheid heb ik nooit verdiend} \rangle$

Voorbeeld output (niet gesorteerd op term)

$\langle al, [doc1] \rangle$

$\langle daar, [doc1, doc2] \rangle$

...

5. Breid de inverted index van 4 uit met positionele informatie:

Voorbeeld output (niet gesorteerd op term)

$\langle je, [\langle doc1, [3, 5, 13] \rangle] \rangle$

$\langle daar, [\langle doc1, [1] \rangle, \langle doc2, [1] \rangle] \rangle$