PageRank 2025-2026

Pascal Molli - Large Scale Data Management

Homework

Bonjour,

Je veux avoir une comparaison des performances sur pagerank, entre une implantation une implantation PySpark DataFrame et une implantation PySpark RDD.

Je veux plusieurs configurations de cluster (Attention, vous êtes visiblement limité à 32 vCPU au total):

- 2 nœuds,
- 4 nœuds,
- 6 nœuds (mais gardez le même hardware CPU/RAM par nœud , sinon les résultats ne sont pas comparables).

Les données sont les pages Wikipedia :

- https://databus.dbpedia.org/dbpedia/generic/wikilinks/2022.12.01/wikilinks_lang=en.ttl.bz2 (1,8Go compressé)
- **!! Attention !!, c'est déjà gros pour un petit cluster. Commencez avec un 10% des données et vérifiez votre expérience.**

Les résultats doivent être présentés sur un github ou gitlab avec le code source et les résultats d'exp dans le README. Je veux voir quelle est l'entité avec le plus grand pagerank ie. le centre de Wikipedia …

Le rendu est donc une URL par groupe de 3 maximum.
!! N'OUBLIEZ PAS VOS NOMS avec votre URL !!

Faites attention au partitionnement des données (voir l'article NSDI), je veux que vous évitiez le shuffle pour pagerank/neighbours.