Tugas I: Hadoop

Buatlah program Hadoop/Map Reduce yang membaca file twitter_rv.net, dan menghasilkan hasil perhitungan triangle counting pada data tersebut.

Struktur twitter_rv.net

File twitter_rv.net adalah file teks, yang setiap barisnya berisi pasangan user id – follower, dengan format:

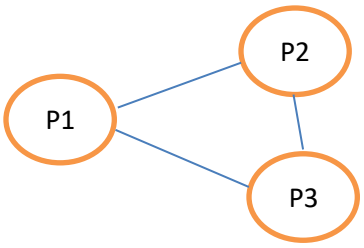
USER \t FOLLOWER \n

Dimana USER adalah user id dari pengguna twitter dan FOLLOWER adalah user id dari pengguna twitter yang mem-follow USER.

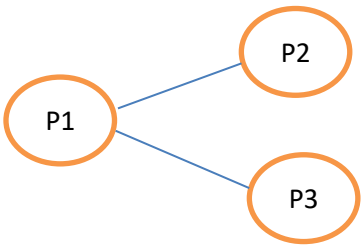
Triangle Counting

Salah satu faktor penting dalam sebuah social network adalah parameter **Clustering Coefficient**, yang menggambarkan seberapa banyak teman dari seseorang yang merupakan teman satu sama lain. Perhitungan Clustering Coefficient dilakukan dengan menghitung triplet, yaitu 3 buah node yang saling berhubungan. Triplet dapat berhubungan satu sama lain, yang disebut sebagai closed triplet (triangle), atau terbuka (open triplet). Clustering Coefficient didefinisikan sebagai perbandingan antara jumlah closed triplet dengan semua triplet. Semakin besar nilai clustering coefficient, semakin tinggi derajat kedekatan sebuah komunitas.

$$Clustering\ Coefficient = \frac{\sum closed\ triplet}{\sum all\ triplet}$$



Gambar 1. Closed Triplet



Gambar 2. Open Triplet

Pada tugas ini, Anda diminta mengembangkan algoritma Triangle Counting, yaitu menghitung jumlah Triangle (closed triplet) pada sebuah social network, yaitu dengan menggunakan data twitter_rv.net.

Algoritma Triangle Counting Serial

Algorithm 1 NodeIterator(V,E)

```
1: T ← 0;
2: for v ∈ V do
3:    for u ∈ Γ(v) do
4:       for w ∈ Γ(v) do
5:          if ((u, w) ∈ E) then
6:             T ← T + 1/2;
7: return T / 3;
```

Pada algoritma 1 di atas, perhitungan Triangle dilakukan dengan iterasi untuk setiap vertex pada graph, dan untuk setiap 2 pasang tetangga dari vertex tersebut, jika saling berhubungan maka akan dihitung sebagai sebuah triangle. Pada algoritma 1, sebuah triangle akan dihitung sebagai 6 triangle, karena dihitung dari 3 vertex dan untuk 2 arah tetangganya (u, w) dan (w, u).

Algoritma 1 di atas dapat dioptimasi dengan hanya menggunakan vertex dengan ide terkecil dalam sebuah triangle, sehingga sebuah triangle tidak dihitung 6 kali.

Algorithm 2 NodeIterator++(V,E)

```
1: T ← 0;
2: for v ∈ V do
3:    for u ∈ Γ(v) and u ≻ v do
4:       for w ∈ Γ(v) and w ≻ u do
5:          if ((u, w) ∈ E) then
6:             T ← T + 1;
7: return T;
```

Algoritma Triangle Counting Map Reduce

Untuk menghitung Triangle Counting dengan Map Reduce, dapat dilakukan dengan strategi 2 fase.

1. Bangkitkan semua kemungkinan pasangan sekuens 3 vertex dengan yang terhubung membentuk path dengan panjang 2
2. Cek apakah pada path yang terbentung, vertex awal dan vertex akhir terhubung, sehingga membentuk triangle.

Sketsa Algoritma Map-Reduce

Pada algoritma Map Reduce di bawah ini, input untuk Map Reduce fase 2 menggunakan gabungan dari input asal dan output dari Map Reduce fase 1.

Algorithm 3 MR-NodeIterator++(V,E)

```
1: Map 1: Input: ⟨(u, v); ∅⟩
2:     if v ≻ u then
3:         emit ⟨u; v⟩
4: Reduce 1: Input ⟨v; S ⊆ Γ(v)⟩
5:     for (u, w) : u, w ∈ S do
6:         emit ⟨v; (u, w)⟩
7: Map 2:
8:     if Input of type ⟨v; (u, w)⟩ then
9:         emit ⟨(u, w); v⟩
10:    if Input of type ⟨(u, v); ∅⟩ then
11:        emit ⟨(u, v); $⟩
12: Reduce 2: Input ⟨(u, w); S ⊆ V ∪ {$}⟩
13:    if $ ∈ S then
14:        for v ∈ S ∩ V do
15:            emit ⟨v; 1⟩
```

Tugas

Hitunglah jumlah triangle berdasarkan data user-follower yang terdapat pada file twitter_rv.net.

File twitter_rv.net dapat diakses pada HDFS pada folder /user/twitter/twitter_rv.net.

Anda harus melakukan preprocessing terlebih dahulu pada data twitter ini dengan mengubah directed edge (a memfollow b) menjadi undirected edge (menambahkan arah sebaliknya (b memfollow a) jika tidak ada pada data asal.

Tugas dikerjakan berkelompok (maksimum 2 orang per kelompok), dan dibolehkan untuk melakukan optimasi algoritma.

Laporan berisi deskripsi algoritma, dan hasil eksekusi, waktu eksekusi dan source code program hadoop anda.

Tugas II: Spark

Buatlah program Spark yang membaca file twitter_rv.net, dan menghasilkan hasil perhitungan jumlah follower unik yang dihitung berdasarkan semua pengikut hingga level ke 2 (follower dari follower).

Struktur twitter_rv.net

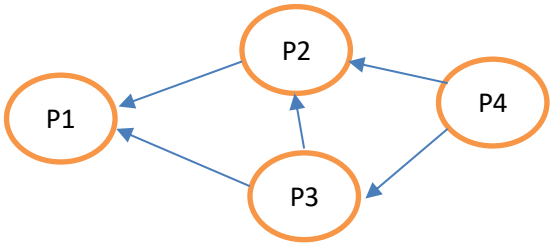
File twitter_rv.net adalah file teks, yang setiap barisnya berisi pasangan user id – follower, dengan format:

USER \t FOLLOWER \n

Dimana USER adalah user id dari pengguna twitter dan FOLLOWER adalah user id dari pengguna twitter yang mem-follow USER.

Contoh

Pada diagram di bawah, jumlah follower unik level ke 2 dari P1 adalah 3, yaitu P2, P3 dan P4. Untuk P2 adalah 2 (P3 dan P4), P3 ada 1 dan P4 ada 0



Tugas

Hitunglah 10 pengguna yang memiliki jumlah follower unik level 2 yang terbanyak.

File twitter_rv.net dapat diakses pada HDFS pada folder /user/twitter/twitter_rv.net. (gunakan HDFS yang sama dengan tugas Hadoop sebelumnya

Laporan berisi deskripsi algoritma, dan hasil eksekusi, waktu eksekusi dan source code program Spark anda.

Link twitter_rv

https://snap.stanford.edu/data/twitter-2010.html

agar diantisipasi ukuran data >5 GB