BAB 6

KESIMPULAN DAN SARAN

Pada bab ini diberikan kesimpulan dan saran berdasarkan hasil analisis, perancangan, implementasi, pengujian, dan eksperimen yang telah dilakukan.

6.1 Kesimpulan

Berdasarkan hasil analisis, perancangan, implementasi, dan pengujian yang telah dilakukan, dapat diambil beberapa kesimpulan, yakni:

1. Pembagian tugas dan koordinasi antar agen diatur oleh *server crawler*. *Server* bertugas memasukkan URL ke *frontier* dan agen bertugas memasukkan hasil *crawling* ke *web repository*. Agen juga akan mengekstrak URL dari halaman *web* yang di-*crawl* untuk dikirimkan ke *server* untuk kemudian dimasukkan ke *frontier*. Koneksi antara *server* dan agen dilakukan dengan perantara *thread* yang dibuat oleh *server* setiap ada agen yang berhasil terhubung ke *server*, jadi satu *thread* tepat menangani satu agen. Banyak agen yang terkoneksi ke *server* berbanding lurus dengan kecepatan proses *crawling*. Hal ini dikarenakan pada waktu yang bersamaan agen-agen melakukan *crawling* pada URL-URL yang berbeda.

2. HBase yang berjalan diatas lingkungan *file system* terdistribusi Hadoop dapat memproses data dengan relatif cepat. Banyak komputer yang tergabung dalam *cluster* Hadoop berbanding lurus dengan kecepatan pemrosesan data. Karena tiap agen berjalan pada komputer yang merupakan bagian dari *cluster* Hadoop, maka hasil *crawling* yang disimpan oleh agen akan terdistribusi pada *file system* terdistribusi Hadoop ini.

3. Pada skripsi ini telah berhasil dibangun perangkat lunak *web crawler* dengan metode *distributed web crawling* yang terdiri atas situs induk, *server*, dan agen *crawler*. *Web crawler* yang dibuat memanfaatkan HBase yang merupakan *database* Hadoop sebagai media penyimpanan URL yang akan di-*crawl* maupun hasil *crawling*.

4. Hasil dari eksperimen menunjukkan bahwa banyak agen dan komputer dalam *cluster* Hadoop yang digunakan berbanding lurus dengan kecepatan proses *crawling*. Selain itu, banyak komputer dalam *cluster* Hadoop yang digunakan berbanding lurus dengan kecepatan pencarian.

5. Berdasarkan konfigurasi lingkungan pengujian, koneksi internet juga dapat menjadi faktor yang mempengaruhi kecepatan *crawling*. Selain itu, koneksi dari aplikasi ke HBase juga dapat mempengaruhi kecepatan pemrosesan data.

6.2 Saran

Berdasarkan hasil analisis, perancangan, implementasi, pengujian, dan eksperimen yang telah dilakukan, penulis memberikan beberapa saran sebagai berikut:

1. Agar proses *crawling* dapat lebih cepat dilakukan, maka pengembangan dan eksekusi program *crawling* sebaiknya dilakukan pada komputer dengan spesifikasi yang tinggi. Selain itu, diperlukan pula koneksi internet yang baik dan stabil agar tidak terjadi kendala dalam proses pengambilan konten dan ekstraksi yang dilakukan oleh agen.

2. Agen hendaknya tersebar merata di tiap komputer yang termasuk dalam *cluster* Hadoop. Hal ini dilakukan untuk mengurangi jumlah *resource* komputer yang termakan akibat proses *crawling* yang dilakukan banyak agen.

3. Untuk mencegah terjadinya kondisi *bottleneck* pada *server*, dapat pada *server* itu sendiri dapat diimplementasikan *multi-threading* sehingga koneksi dari banyak agen tidak hanya ditangani oleh satu *server*.

4. Jika ingin meningkatkan kecepatan pemrosesan data pada HBase, maka sebaiknya dilakukan penambahan komputer ke *cluster* Hadoop, agar jumlah *node* pada lingkungan *file system* terdistribusi Hadoop juga bertambah. Semakin banyak *node* pada *file system* terdistribusi Hadoop, makin cepat pula pemrosesan data dapat dilakukan.

DAFTAR REFERENSI

[1] Olston, C. dan Najork, M. (2010) Web crawling. *Found. Trends Inf. Retr.*, **4**, 175–246.

[2] V. Udapure, T., D. Kale, R., dan C. Dharmik, R. (2014) Study of web crawler and its different types, **. 16**, 01–05.

[3] Acharjya, D. dan Mitra, A. (2017) *Bio-Inspired Computing for Information Retrieval Applications*. IGI Global, Hershey, PA, USA.

[4] Beena Mahar, C. K. J. (2015) A comparative study on web crawling for searching hidden web. *International Journal of Computer Science and Information Technologies (IJST)*, **6**, 2159–2163.

[5] Cho, J. dan Garcia-Molina, H. (2002) Parallel crawlers. *Proceedings of the 11th International Conference on World Wide Web*, New York, NY, USA, May 07 - 11 WWW '02, pp. 124–135. ACM.

[6] Sarah, E. (2016) Information overload. *Distillations*, **2**, 26–33.

[7] Erl, T., Khattak, W., dan Buhler, P. (2016) *Big Data Fundamentals: Concepts, Drivers & Techniques*, 1st edition. Prentice Hall Press, Upper Saddle River, NJ, USA.

[8] Holmes, A. (2012) *Hadoop in Practice*. Manning Publications Co., Greenwich, CT, USA.

[9] White, T. (2012) *Hadoop: The Definitive Guide*, 3rd edition. O'Reilly Media, Inc., Sebastopol, CA, USA.