

BAB I

PENDAHULUAN

Pada bab ini menjelaskan mengenai latar belakang permasalahan, rumusan masalah, tujuan dan manfaat, batasan masalah, metodologi penelitian, dan sistematika penulisan dalam penelitian ini.

1.1. Latar Belakang

Web Crawler dikenal juga sebagai *web spider*. *Web crawler* merupakan program yang digunakan suatu *search engine* untuk menghimpun informasi yang diperlukan dari *World Wide Web* secara otomatis sesuai dengan aturan yang telah ditetapkan oleh pengguna [1]. Proses yang dijalankan *web crawler* adalah mengekstraksi URL utama untuk menghasilkan daftar link halaman yang akan dikunjungi, kemudian mengunjungi semua halaman yang terdaftar, selanjutnya mengidentifikasi semua hyperlink yang terdapat pada halaman itu, lalu menambahkannya ke daftar link halaman yang selesai dikunjungi [2].

Web crawler dasar dengan program *single-thread* hanya menggunakan satu thread dalam proses penghimpunan informasi yang banyak dari pengguna. Oleh sebab itu *web crawler* dengan *single-thread* tidak mampu untuk memenuhi permintaan pengguna informasi baik dari segi jumlah maupun kualitasnya [1]. Kekurangan ini dapat diatasi dengan menerapkan teknik pemrograman *multi-threaded* dan mendistribusikan program *web crawler* ke masing – masing server. *Multi-threaded* merupakan salah satu teknik dalam pemrograman komputer yang bertujuan untuk membuat proses berjalan secara bersamaan [3]. Dengan teknik ini, dimungkinkan untuk membuat *web crawler* mampu menangani permintaan informasi yang banyak dari pengguna dalam satu waktu.

Proses ekstraksi isi halaman *website* sangat bergantung pada struktur dari masing – masing *website* yang akan di-*crawl*, jadi bot yang digunakan untuk proses ekstraksi harus menyesuaikan struktur *website*-nya. Struktur *website* dikelompokkan menjadi 4, yaitu struktur linear, non-linear, hirarki, dan hybrid [4].

Data yang tersimpan dari hasil *crawling* cukup besar sehingga akan membuat server bekerja lebih berat dan besar kemungkinan resiko kegagalan akses oleh pengguna dari *search engine* yang menggunakan data tersebut. Implementasi *database* terdistribusi dapat menghasilkan kinerja yang baik dalam hal ketersediaan data. Replikasi *database* dapat memungkinkan pembagian beban akses ke *server*, sehingga dapat

meminimalisir resiko kegagalan akses oleh pengguna *search engine* [5]. Replikasi database merupakan suatu teknik yang dapat digunakan untuk optimasi akses data dan memberikan toleransi kesalahan dengan cara menyimpan salinan data di beberapa lokasi [6].

Replikasi database dapat menggunakan 3 cara yaitu *snapshot replication*, *merging replication*, dan *transactional replication*. *Snapshot replication* digunakan untuk penyalinan data pada database server ke database lain di server yang sama atau berbeda. *Merging replication* digunakan untuk penggabungan data dari dua atau lebih database yang digabungkan menjadi satu database. *Transactional replication* digunakan untuk penyalinan secara lengkap dari database master, kemudian mendapat pembaruan berkala pada perubahan yang terjadi pada database master [7].

Penelitian ini akan menggunakan cara atau teknik *transactional replication*, karena distribusi data yang diharapkan selalu melakukan pembaruan secara berkala dari database master dengan kapasitas data yang cukup besar. Data pada masing – masing server juga harus sinkron data pada server A harus sama dengan server B apabila server A terdapat perubahan maka server B menerima pembaruan, begitupun sebaliknya. Untuk mengatasi hal tersebut pada penelitian ini lebih cocok memanfaatkan komponen *master to master* daripada *master to slave* yang telah disediakan oleh database yang akan digunakan. Karena jika menggunakan master to slave hanya pada database master yang dapat memberikan pembaruan data dan mereplikasi ke database slave, database slave hanya dapat menerima pembaruan dari database master [8].

Suatu *web crawler* perlu melakukan pembaruan isi dari database secara berkala untuk meningkatnya pembaruan dan kualitas informasi pada database yang dimiliki [9]. Pembaruan suatu *website* dapat dilihat dari perubahan yang ada pada halaman utama *website* tersebut. Beberapa cara untuk mengetahui perubahan dari suatu *website* atau *website* yang telah dilakukan pembaruan, yaitu perubahan struktur halaman, perubahan isi teks dan perubahan gambar (*hyperlink*) yang terdapat pada halaman *web* tersebut [10].

Pada penelitian ini, untuk mengetahui perubahan yang terdapat pada halaman *web* dengan cara membandingkan perubahan pada isi atau *content*. Isi atau *content* yang dimaksud adalah *url* atau *link* berita tersebut. Teknik ini diharapkan bisa membandingkan berita yang terbaru dan berita lama. Untuk menghasilkan berita olahraga *up to date* yang

maksimal maka akan dilakukan *crawling* secara berkala setiap 1 jam sekali.

Informasi yang akan dihimpun pada penelitian ini adalah informasi tentang berita olahraga dengan 10 domain yang telah ditentukan yaitu:

1. <http://www.sports.okezone.com/>
2. <http://www.bola.com/>
3. <http://www.cnnindonesia.com/olahraga>
4. <http://www.mediaindonesia.com/news/list/olahraga>
5. <http://www.merdeka.com/olahraga/>
6. <http://www.republika.co.id/kanal/olahraga>
7. <http://www.bbc.com/indonesia/olahraga>
8. <http://www.m.metrotvnews.com/olahraga>
9. <http://www.kompasiana.com/olahraga>
10. <http://www.sport.detik.com/>

Pemilihan 10 domain tersebut berdasarkan *ranking site* teratas di Indonesia dari www.alex.com per tanggal 30 Juni 2017, kemudian dipilih situs – situs yang berhubungan dengan berita olahraga.

Dari beberapa masalah yang telah diuraikan diatas, dalam penelitian ini akan mencoba menerapkan teknik pemrograman *multi-threaded* pada *web crawler* kemudian data hasil *crawling* yang tersimpan ke dalam database didistribusikan dengan cara replikasi dari database master ke master lainnya, dimana komputer server dan komputer master lainnya akan saling berkomunikasi dalam *database* sehingga setiap perubahan pada komputer *master* akan selalu direplikasi pada komputer *master* yang lain begitupun sebaliknya, kemudian menambahkan fitur deteksi pembaruan *website* dengan membandingkan tanggal dan jam *upload* pada berita pada proses pengambilan berita olahraga sehingga berita yang didapatkan nantinya adalah semua berita olahraga yang terbaru dari halaman *web* dengan 10 domain yang telah ditentukan.

Tujuan yang ingin dicapai dari penelitian ini adalah meningkatkan kecepatan proses *web crawler* dengan penerapan teknik pemrograman *multi-thread* dalam menghimpun berita olahraga kemudian data hasil *crawling* didistribusikan dengan teknik replikasi database master *to* master. Kemudian dalam penghimpunan berita olahraga, hanya menghimpun berita olahraga yang terbaru menggunakan teknik membandingkan *url* atau *link* berita pada 10 domain olahraga yang telah ditentukan.

1.2. Perumusan Masalah

Berikut rumusan masalah yang dapat dibuat berdasarkan pemaparan latar belakang di atas:

1. Berapa kecepatan proses *crawling* yang dapat ditingkatkan *web crawler multi-thread* dibandingkan *web crawler single-thread* dalam menghimpun berita olahraga dari 10 domain yang telah ditentukan?
2. Berapa kecepatan proses *crawling* berita olahraga dari 10 domain yang telah ditentukan, setelah penerapan deteksi pembaruan *website* pada *web crawler multi-thread*?
3. Berapa kecepatan proses *crawling* berita olahraga dari 10 domain yang telah ditentukan menggunakan *web crawler multi-thread* dan *web crawler multi-thread* terdistribusi?

1.3. Tujuan dan Manfaat

Adapun tujuan yang ingin dicapai pada penelitian ini, sebagai berikut:

1. Merancang dan mengimplementasikan *web crawler* dengan teknik pemrograman *multi-thread*.
2. Penerapan fitur deteksi pembaruan *website* untuk memperoleh berita olahraga yang *up to date* dan mempersingkat waktu *crawling*, karena dengan penambahan fitur tersebut berita yang sudah disimpan maka tidak perlu disimpan kembali.
3. Melakukan distribusi menggunakan teknik replikasi database pada 2 komputer untuk meningkatkan kecepatan *crawling* berita olahraga

Manfaat yang akan diperoleh dari penerapan beberapa konsep dan fitur pada penelitian ini sebagai berikut:

1. Meningkatkan kecepatan proses *crawling* berita olahraga dengan penerapan konsep *multi-thread*.
2. Memperoleh berita olahraga yang terbaru.
3. Memiliki *backup* data apabila suatu saat server mengalami kerusakan.

1.4. Batasan Masalah

Agar pembahasan menjadi terfokus, maka perlu dibuat batasan masalah sebagai berikut:

1. Penelitian yang dilakukan adalah peningkatan kecepatan proses *crawling* pada *web crawler*.
2. Sistem distribusi yang dilakukan pada penelitian ini pada 2 komputer.

3. Berita atau informasi yang dihimpun adalah berita atau informasi tentang olahraga yang berasal dari 10 domain yang telah ditentukan.
4. Hasil akhir dari penelitian ini berupa data hasil *crawling* dari 10 domain yang telah ditentukan, tersimpan pada *database* dan dapat melihat data hasil *crawling*.

1.5. Metodologi Penelitian

Secara rinci langkah-langkah yang dilakukan dalam penelitian ini adalah sebagai berikut:

1.5.1 Studi Pustaka

Pada tahap ini dilakukan pengumpulan informasi terkait web crawler. Informasi diperoleh dari berbagai literatur meliputi artikel, jurnal, buku, dll.

1.5.2 Analisa dan Perancangan Sistem

Pada tahap ini dilakukan analisa dan perancangan *web crawler multi-thread* yang akan dibangun dengan mengacu pada hasil pemahaman dan studi pustaka pada tahap sebelumnya.

1.5.3 Implementasi Sistem

Pada tahap ini dilakukan implementasi hasil dari analisa dan perancangan sistem ke dalam *web crawler* yang akan dibuat pada penelitian ini.

1.5.4 Uji Coba Sistem

Pada tahap ini sistem diuji coba pada sistem yang telah dibuat. Apakah sudah sesuai dengan tujuan utama pada penelitian ini.

1.5.5 Analisa dan Evaluasi

Pada tahap ini dilakukan analisa untuk mengetahui hasil dari sistem yang telah dibuat. Apakah sudah melampaui target yang ingin dicapai pada penelitian ini.

1.5.6 Dokumentasi

Pada tahap ini dilakukan penyusunan laporan dan dokumentasi dari pelaksanaan secara keseluruhan, mulai dari tahap awal sampai tahap akhir.

1.6. Sistematika Penulisan

Sistematika penulisan yang digunakan pada laporan penelitian ini adalah sebagai berikut:

BAB I : PENDAHULUAN

Pada bagian ini menjelaskan hal-hal yang melatarbelakangi penelitian, rumusan masalah, tujuan, batasan masalah, metodologi penelitian, dan sistematika penulisan.

BAB II : STUDI PUSTAKA

Bagian ini berisi uraian teori-teori penunjang dalam pengerjaan sistem dan penulisan laporan penelitian skripsi ini.

BAB III : PERANCANGAN SISTEM

Bagian ini menguraikan desain atau rancangan dari sistem yang akan dibuat.

BAB IV : IMPLEMENTASI SISTEM

Bagian ini akan menjelaskan uraian tentang implementasi sistem yang telah dibuat.

BAB V : PENUTUP

Bagian ini berisi kesimpulan dari implementasi sistem yang telah dibuat serta saran untuk pengembangan sistem atau penelitian berikutnya.