

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Meningkatnya kebutuhan akan informasi mendorong manusia untuk mengembangkan teknologi-teknologi baru agar pengolahan data dan informasi dapat dilakukan dengan mudah dan cepat. Salah satu teknologi yang sedang berkembang dengan pesat saat ini adalah teknologi informasi komputer[1]. Untuk mengolah data dan informasi tersebut dibutuhkan yang namanya teknologi *internet*, dengan adanya *internet* akan mempermudah seseorang untuk mendapatkan atau bertukar informasi dan mempercepat proses pengolahan data agar dapat diolah menjadi sebuah informasi. Dengan meningkatnya kecanggihan teknologi *internet* maka kegiatan jual beli pun dapat dilakukan secara *online*, dan pada zaman sekarang sudah banyak toko dan para penjual yang sudah dapat memasarkan produknya secara online melalui *e-commerce* dan *e-marketplace*.

Situs-situs *e-commerce* dan *e-marketplace* yang ada di dunia internet sekarang ini memberikan dan menawarkan setiap produk dengan harga dan deskripsi produk yang berbeda-beda. Jumlah deskripsi produk dan jenis produk ini tentunya jumlahnya sangat banyak dan beraneka ragam sehingga untuk mencari produk yang diinginkan menghabiskan waktu yang sangat lama, Untuk mengatasi masalah tersebut dalam mencari produk yang diinginkan secara detail dengan waktu yang efisien dapat diatasi dengan pengambilan informasi menggunakan *web scraping*.

Web Scraping adalah proses pengambilan sebuah dokumen-semi terstruktur dari internet, umumnya berupa halaman-halaman Web dalam bahasa *markup* seperti HTML (*Hypertext Markup Language*) atau XHTML (*Extensible HyperText Markup Language*), dan menganalisis dokumen tersebut untuk diambil data tertentu dari halaman tersebut untuk digunakan bagi kepentingan lain. Bahasa *markup* adalah bahasa komputer yang menggunakan *tag* untuk mendefinisikan elemen dalam dokumen. (Matthew Turland, 2010). Dilakukannya kegiatan Web Scraping biasanya adalah untuk Kepentingan pelaku seperti mencari sebuah produk secara

detail dengan waktu yang cepat, menggali konten sebuah situs web secara cepat perhalaman.

Dengan adanya teknik *web scraping* ini tentu saja sudah banyak beberapa orang yang menggunakan teknik ini untuk mencari informasi dan detail yang lebih cepat. Tetapi ada juga yang berdampak buruk terhadap beberapa penyedia informasi tertentu yang tidak ingin informasi, data, dan kontennya yang diambil dengan cara *scraping* ini karena untuk kepentingan persaingan bisnis yang dimana data dapat diambil untuk dijadikan sebuah informasi yang memiliki kepentingan dari pelaku *web scraping* ini, dan sebagai contoh ada beberapa kasus yang melibatkan *bot* yang diprogram untuk kecepatan mengambil sebuah produk secara cepat di dunia *e-commerce* dan *e-marketplace* untuk mengambil beberapa produk yang memiliki promo diskon terutama promo *flash sale* yang dimana *bot* ini sudah disiapkan untuk mengambil produk tersebut ketika baru pertama kali rilis sehingga menyebabkan produk *flash sale* tersebut habis terjual secara cepat, bahkan pengguna biasa yang ingin membeli produk tersebut tidak memiliki kesempatan untuk membeli karena kalah cepat dengan *bot-bot* yang telah dirancang oleh pelaku *web scraping* ini. Karena kasus yang ada tersebut maka dibuatlah sebuah metode sistem proteksi situs web dari *web scraping* ini untuk dapat menghalangi sebuah *robot* untuk masuk kedalam sistem situs web.

CAPTCHA atau Captcha (Completely Automated Public Turing test to Tell computer and Human Apart) adalah sebuah sistem tes uji tanggapan (*challenge-response test*) yang biasanya digunakan di dunia perkomputeran untuk melakukan verifikasi manusia, apakah yang melakukan akses atau meminta request ke sebuah sistem adalah seorang manusia sungguhan atau hanya komputer(*bot*). Dengan adanya CAPTCHA ini, dapat melindungi sistem situs web penyedia informasi dari *bot*, *crawler*, *scraper*, *spammer*, bahkan melindungi konten web.

Dalam penelitian ini, penulis mengembangkan sebuah metode untuk mencegah dan mengatasi masalah tersebut, yaitu Metode bagaimana memproteksi situs web dari teknik *web scraping* ini. Penulis dapat menggunakan metode *human verification* untuk mencegah bot masuk kedalam sebuah situs web, kemudian metode manajemen *IP address*, proteksi *javascript*, blokir alamat *IP* yang

mencurigakan, membatasi *request* untuk halaman web yang dilindungi,, dan manajemen *traffic data* situs web.

Dengan adanya metode-metode tersebut mungkin dapat mengurangi bahkan mencegah *bot* untuk masuk kedalam sebuah situs web yang tidak ingin dimasuki oleh *bot*. Karena metode untuk melakukan verifikasi Manusia atau tidak biasanya hanya dapat dilakukan oleh manusia yang asli karena manusia memiliki otak untuk berfikir sedangkan *bot* hanya mengandalkan mesin untuk bekerja, dan juga melakukan manajemen *web server* juga sangatlah penting agar pemilik atau penyedia informasi dapat membatasi akses *bot* untuk masuk dan mengambil data-data yang ada.

1.2. Rumusan Masalah

Ada beberapa rumusan masalah untuk penelitian ini, diantaranya adalah sebagai berikut :

1. Sistem Verifikasi manusia seperti apa yang cocok untuk melakukan verifikasi terhadap pengunjung web ?.
2. Bagaimana metode dan implementasi dalam sebuah situs web untuk mencegah akses *bot* ?.
3. Bagaimana cara melindungi konten web agar tidak dapat diambil dan disalahgunakan oleh *bot*?
4. Dapatkah kita melindungi situs web secara menyeluruh dengan metode-metode yang ada untuk melindungi situs web dari web scraping?

1.3. Tujuan dan Manfaat

Ada beberapa tujuan dan manfaat untuk penelitian ini, diantaranya adalah sebagai berikut :

1. Mengetahui Sistem Verifikasi manusia yang cocok untuk melakukan verifikasi terhadap pengunjung web.

2. Melakukan Metode dan implementasi ke sebuah situs web untuk mencegah *bot* web scraping.
3. Dapat melindungi konten web agar tidak dapat diambil dan disalahgunakan oleh pengunjung situs web dan *bot*.
4. Dapat mengetahui apakah situs web dapat dilindungi secara menyeluruh dengan metode-metode perlindungan situs web dari web scraping.

1.4. Batasan Masalah

Ada beberapa batasan masalah untuk penelitian tentang metode proteksi situs web dari scraping data ini:

1. Situs Web yang dijadikan bahan penelitian adalah situs web buatan sendiri.
2. Pembatasan akses *bot* dapat berdampak pada *bot crawler* pengumpul data milik google agar dapat menampilkan situs web penulis paling relevan dengan kata kunci.
3. Karena situs Web yang dijadikan bahan penelitian adalah situs web buatan sendiri maka penulis harus membuat sebuah *enviroemnt* situs web sehingga dapat dilakukan uji coba untuk proteksi keamanan dari *bot* yang didampingi oleh penguji kerentanan situs web terhadap *bot*.
4. Manajemen *traffic data*, alamat *IP* hanya dapat dilakukan ketika sudah memiliki web server dan admin situs web yang memiliki akses penuh.
5. Metode *human verification* menggunakan CAPTCHA sudah ada beberapa pihak yang dapat melakukan *bypass*, maka dari itu penulis berusaha untuk membuat sistem verifikasi manusia dengan sistem CAPTCHA yang lebih efektif.

1.5. Sistematika Penulisan Tugas Akhir

Sistematika penulisan yang digunakan dalam Tugas Akhir ini adalah :

BAB I PENDAHULUAN

Dalam BAB I berisi tentang latar belakang, rumusan masalah, batasan masalah, dan sistematika penulisan.

BAB II LANDASAN TEORI

Dalam BAB II berisi mengenai dasar-dasar teori berupa pengertian dari web scraping, CAPTCHA, *Javascript*, pemanfaatan *javascript* untuk CAPTCHA, membatasi akses *right click* dan *inspect element*, Sistem *javascript browser Checker*, *nodeJS*, *Robot.txt*, Blokir IP mencurigakan.

BAB III ANALISIS DAN PERANCANGAN SISTEM

Dalam BAB III berisi tentang gambaran umum sistem, analisis kebutuhan sistem, perancangan sistem, dan diagram pengujian. Pada gambaran umum sistem akan dijelaskan bagaimana sistem secara umum bekerja. Analisis kebutuhan sistem menjelaskan spesifikasi yang dibutuhkan untuk sistem dengan parameter kebutuhan data, *software*, *hardware*, dan *brainware*. Perancangan sistem akan dijelaskan secara umum bagaimana sistem bekerja melalui ketika metode-metode akan diimplementasikan ke situs web.

BAB IV HASIL DAN ANALISIS

Dalam BAB IV berisi tentang pengujian yang bertujuan untuk mengetahui implementasi metode dan hasil dari metode, beserta cara kerjanya.

BAB V KESIMPULAN DAN SARAN

Dalam BAB V berisi kesimpulan yang diambil dari hasil penelitian yang dilakukan dan saran atau masukan untuk penelitian selanjutnya.