SKRIPSI

STUDI DAN EKSPLORASI MEMPEKERJAKAN *WEB CRAWLER* APACHE NUTCH

AGINA RINDA

NPM: 2014730062

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2018

UNDERGRADUATE THESIS

STUDY AND EXPLORATION TO EMPLOY APACHE NUTCH WEB CRAWLER

AGINA RINDA

NPM: 2014730062

DEPARTMENT OF INFORMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY
2018

LEMBAR PENGESAHAN

STUDI DAN EKSPLORASI MEMPEKERJAKAN *WEB CRAWLER* APACHE NUTCH

AGINA RINDA

NPM: 2014730062

Bandung, 10 Desember 2018

Menyetujui,

Pembimbing

Gede Karya, M.T., CISA, IPM

Ketua Tim Penguji

Luciana Abednego, M.T.

Anggota Tim Penguji

Raymond Chandra Putra, M.T.

Mengetahui,

Ketua Program Studi

Mariskha Tri Adithia, P.D.Eng

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

STUDI DAN EKSPLORASI MEMPEKERJAKAN *WEB CRAWLER* APACHE NUTCH

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 10 Desember 2018

AGINA RINDA
NPM: 2014730062

ABSTRAK

Web crawler merupakan perangkat lunak yang dapat melakukan *web scanning* dan *indexing* untuk membantu *search engine* dalam mengumpulkan informasi yang dibutuhkan oleh manusia dengan waktu yang singkat. *Web crawler* Apache Nutch merupakan salah satu *web crawler* yang dapat digunakan untuk melakukan penjelajahan ke banyak URL. *Web crawler* Apache Nutch yang menjadi objek penelitian dijalankan di atas lingkungan terdistribusi Hadoop. Hadoop merupakan *framework* yang dapat menyimpan jumlah data yang sangat besar, salah satunya data yang dihasilkan dari hasil *crawling*. Pada lingkungan Hadoop, media penyimpanan yang dapat digunakan oleh *web crawler* Apache Nutch sebagai media penyimpanan dan membantu dalam mengolah data yang berukuran besar adalah *Not Only SQL* (NoSQL) HBase.

Pengembangan *web crawler* dapat dilakukan dengan dua cara, yaitu cara pertama dengan mengembangkan aplikasi *web crawler* sendiri menggunakan algoritma sesuai dengan teknik *web crawling* tertentu atau cara kedua dengan memanfaatkan *web crawler* yang bersifat *open source*, salah satunya *web crawler* Apache Nutch. Pada penelitian ini, *web crawler* yang digunakan adalah *web crawler* Apache Nutch yang dipekerjakan melalui Nutch REST API. Untuk dapat mempekerjakan *web crawler* Apache Nutch melalui Nutch REST API dibangun aplikasi Java sebagai REST API Client yang diberi nama Agen *Crawler* yang mengimplementasikan pemanggilan Nutch REST API dan juga aplikasi situs induk J2EE yang berfungsi untuk mengakses konten hasil *crawling*.

Pengujian dilakukan dengan dua jenis pengujian, yaitu pengujian fungsional terhadap situs induk J2EE dan eksperimen performa *web crawler* Apache Nutch. Pengujian fungsional dilakukan terhadap situs induk J2EE untuk memastikan bahwa semua fungsi dan fitur yang ada pada situs induk berjalan dengan semestinya. Pengujian performa dilakukan pada *web crawler* Apache Nutch untuk mendapatkan performa *web crawler* Apache Nutch dalam proses *crawling* di atas lingkungan terdistribusi Hadoop. Pengujian performa dilakukan menggunakan empat komputer dan dilakukan secara bertahap dengan penambahan jumlah *web crawler* Apache Nutch dan penambahan *region server*. Berdasarkan hasil pengujian performa, didapatkan bahwa semakin banyak *region server* yang digunakan, maka semakin banyak URL yang dapat dilakukan *crawl* oleh *web crawler* Apache Nutch, dan waktu pencarian kata terhadap konten hasil *crawling* juga semakin cepat.

Hasil dari eksperimen performa *web crawler* Apache Nutch tersebut dibandingkan dengan hasil eksperimen performa yang sudah dilakukan dari penelitian *web crawler* lainnya, yaitu terhadap penelitian *incremental web crawler, focused web crawler*, dan *distributed web crawler*. Berdasarkan analisis perbandingan terhadap hasil eksperimen *crawling* penelitian *web crawler* tersebut, didapatkan kesimpulan bahwa untuk performa *focused web crawler* pada 3 *region server* lebih baik dibandingkan *web crawler* lainnya. Tetapi pada saat *region server* ditambah menjadi 5, *web crawler* Apache Nutch mengungguli *web crawler* lainnya (jika dihitung menggunakan rumus *growth*).

Kata-kata kunci: *Web crawler* Apache Nutch, Hadoop, HBase, Nutch REST API, *region server*

ABSTRACT

Web crawler is a software which can perform web scanning and indexing to help search engine to collect information needed by humas in a short time. Apache Nutch web crawler is a web crawler that can be used to crawl many URLs. The Apache Nutch web crawler can be employed via the command line, or through REST API.

The development of web crawlers can be done in two ways, including by developing the web crawler application using an algorithm according to each web crawling technique or by utilizing an open source web crawler, such as Apache Nutch web crawler developed by the Apache Software Foundation. In this study, the web crawler used was the Apache Nutch web crawler that was employed through REST API. To be able to employ the Apache Nutch web crawler through the REST API, Java applications are built as a REST API Client which is named Crawler Agent that implements the calling of the REST API and the parent site that is built using Java Platform, Enterprise Edition (J2EE) which used to access crawled content. The Apache Nutch web crawler that is the object of the research is run on top of the Hadoop distributed environment. Hadoop is a framework that can store a very large amount of data, one of which is data generated from results of crawling process. In the Hadoop environment, storage media that can be used by Apache Nutch web crawler as a storage medium and helps in processing large sized data is the Not Only SQL (NoSQL) HBase.

There are two types of tests performed, including functional testing of the J2EE parent site and performance testing of the Apache Nutch web crawler. Functional testing is carried out on the J2EE parent site to ensure that all functions and features of the parent site are running properly. Performance testing is done on the Apache Nutch web crawler to get Apache Nutch web crawler performance in the crawling process on top of the Hadoop distributed environment. Performance testing was carried out using four computers and carried out in stages by using five Apache Nutch web crawler agents without any region server, ten Apache Nutch web crawler agents and one region server, fifteen Apache Nutch web crawler agents and two region server, and twenty Apache Nutch web crawler agents and three region server. Based on the results of the performance test, the more region servers are used, the more URLs that can be crawled by Apache Nutch web crawler, and the time to search for the results of the crawled content is also getting faster.

The results of the Apache Nutch web crawler performance experiments are then compared with the results of the performance experiments that have been carried out from other *web crawler* research, namely incremental web crawler, focused web crawler, and distributed web crawler. Based on the results of a comparative analysis of each of the experimental results of each web crawler, focused web crawler better than other web crawlers on 3 region server. But when the region server added to 5, Apache Nutch web crawler better than other web crawlers (if calculated using the growth formula).

Keywords: Apache Nutch web crawler, Hadoop, HBase, Nutch REST API, region server

Dipersembahkan untuk diri sendiri dan Orangtua penulis

KATA PENGANTAR

Puji dan syukur penulis panjatkan kepada Tuhan Yesus Kristus yang telah memberikan berkat dan rahmat-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul "Studi dan Eksplorasi Mempekerjakan Web Crawler Apache Nutch". Adapun tugas akhir ini disusun untuk memenuhi salah satu persyaratan untuk menyelesaikan pendidikan di Fakultas Teknologi Informasi dan Sains pada Program Studi Teknik Informatika di Universitas Katolik Parahyangan Bandung. Dalam proses penyusunan skripsi, penulis banyak mendapat kesempatan untuk menambah ilmu dan mempelajari hal-hal baru, serta mendapatkan bantuan baik secara langsung maupun tidak langsung dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis ingin menyampaikan terima kasih kepada:

1. Orangtua penulis yaitu Adres Ginting dan Julita Depari, abang penulis Adnanta Rio, serta adik penulis Arnov Toto yang memberikan motivasi dan menjadi penyemangat untuk penulis dalam menyelesaikan skripsi ini.
2. Bapak Gede Karya, M.T., CISA, IPM selaku pembimbing yang telah memberikan banyak masukan dan arahan selama pengerjaan skripsi ini.
3. Ibu Luciana Abednego, M.T., dan Bapak Raymond Chandra Putra, M.T., selaku dosen penguji yang telah memberikan kritik dan saran yang membangun dalam penulisan skripsi ini.
4. Muhammad Hilman selaku teman penulis dari awal mulai perkuliahan yang selalu mendengarkan keluh kesah dan memberikan dukungan penuh dalam proses menyelesaikan skripsi ini.
5. Gaby Chairunnissa, Desti Asihanti Saputra, dan Adam Ghaffar selaku sahabat SMA penulis yang selalu memberikan dukungan serta kepercayaan kepada penulis untuk tetap berjuang menyelesaikan skripsi ini.
6. Melinda Nur Abianti, Reza Reynaldi Hasan Haznam, dan Sapta Hadi Kesuma selaku bagian dari kelompok belajar Catatan Anak Sukses (CAS) yang senantiasa mengajarkan, memberikan motivasi, dan membantu penulis dalam melewati masa-masa perkuliahan.
7. Daud Andrew Gorgha Hutasoit sebagai teman yang selalu mendukung dan memberikan motivasi agar penulis dapat menyelesaikan skripsi ini.
8. Pihak-pihak lain yang belum disebutkan, yang telah memberikan bantuan dalam penyusunan skripsi.

Semoga Tuhan memberikan balasan yang berlipat dari segala bentuk kebaikan yang telah diberkan oleh pihak tersebut. Penulis berharap dengan adanya skripsi ini dapat memberikan manfaat baik untuk pembelajaran dan peneltian selanjutnya. Akhir kata, penulis memohon maaf apabila terdapat kesalahan dan kekurangan pada penulisan skripsi ini.

Bandung, Desember 2018

Penulis

DAFTAR ISI

DAFTAR GAMBAR

[1] https://github.com/rishal-hurbans/The-REST-Architectural-Style (diakses pada 4 Oktober 2018)

DAFTAR TABEL

BAB 1

PENDAHULUAN

Bab ini berisi latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi, dan sistematika pembahasan.

1.1 Latar Belakang

Informasi yang ada di internet begitu masif jumlah dan jenisnya. Milyaran manusia terus berinteraksi di internet dan terus meningkat kebutuhannya akan informasi yang ada di internet, yang mana hal ini didukung oleh kemajuan teknologi informasi dan komunikasi. Setiap informasi yang ada pada internet secara umum dapat diakses oleh manusia, namun karena banyaknya jumlah informasi yang ada maka diperlukan bantuan mesin pencari. Kinerja tinggi dari mesin pencari dalam menelusuri banyaknya informasi di internet sangat dibutuhkan, supaya menghasilkan hasil pencarian yang sesuai dan dalam waktu yang singkat. Untuk itu mesin pencari perlu dibantu oleh sejenis robot yang disebut dengan *web crawler* yang dapat melakukan *scanning* terhadap informasi yang tersimpan pada internet.

Web crawler (atau dikenal juga sebagai *web spider*, *web robot*, *bot*, *crawl*, atau *automatic indexer*) merupakan perangkat lunak yang dengan metode tertentu melakukan *web scanning* dan *indexing* dari data yang dicari sesuai dengan kebutuhan manusia [1]. Pada dasarnya, *web crawler* merupakan salah satu komponen utama mesin pencari yang digunakan untuk mempercepat proses pengumpulan informasi yang nantinya akan diakses oleh manusia. Selain *web crawler*, komponen lain dari mesin pencari, yaitu *indexing system* yang bertanggung jawab dalam membuat indeks sehingga proses pencarian akan lebih cepat dan *search system* yang bertanggung jawab dalam proses pencarian yang dibutuhkan oleh manusia. Ada berbagai jenis *web crawler*, antara lain *incremental web crawler*, *focused web crawler*, dan *distributed web crawler* yang melakukan penjelajahan terhadap seluruh halaman *web* dengan algoritma yang berbeda-beda.

Terdapat dua cara untuk mengembangkan *web crawler*, yaitu cara pertama dengan membuat dan membangun *web crawler* sendiri dengan menggunakan algoritma *crawling* masing-masing *web crawler* dan cara kedua menggunakan dan mempekerjakan *web crawler* yang bersifat *open source*, yaitu *web crawler* Apache Nutch. Cara pertama sudah dilakukan oleh tiga mahasiswa lain, khususnya pada topik skripsi Perangkat Lunak *Incremental Web Crawler* pada Lingkungan Hadoop oleh Melinda Nur Abianti [2], *Focused Web Crawling* pada Lingkungan Hadoop oleh Jovanka Helen Maradenia [3], dan *Web Crawling* Terdistribusi pada Lingkungan Hadoop oleh Gabriella [4]. Pada penelitian ini yang digunakan adalah *web cralwer* Apache Nutch.

Web crawler Apache Nutch merupakan *web crawler* yang dikembangkan oleh Apache Software Foundation yang dapat melakukan *web scanning* dan *crawling* ke berbagai halaman *web* [5]. *Web crawler* Apache Nutch terlebih dahulu harus dikonfigurasi sebelum akhirnya dapat digunakan untuk melakukan proses *crawling*. *Web crawler* Apache Nutch dapat diintegrasikan pada lingkungan terdistribusi Hadoop dan data hasil *crawling* yang dilakukan *web crawler* Apache Nutch dapat disimpan pada basis data, yaitu salah satunya adalah HBase.

Yang dilakukan pada penelitian ini adalah eksplorasi dan pengimplementasian *web crawler* Apache Nutch yang diperintah menggunakan REST API dan mengembangkan perangkat lunak

berbasis *web* yang dibangun menggunakan *Java Platform, Enterpise Edition* (J2EE) untuk memasukkan *seed* URL dan mengakses hasil *crawling*. Pada akhir penelitian ini, hasil eksperimen *crawling web crawler* Apache Nutch dibandingkan dengan hasil eksperimen *crawling* yang telah dilakukan oleh tiga mahasiswa lainnya untuk mengetahui setiap *web crawler* pada saat dijalankan di atas lingkungan terdistribusi Hadoop.

1.2 Rumusan Masalah

Rumusan masalah yang dikaji adalah sebagai berikut:

1. Bagaimana cara mempekerjakan Apache Nutch dari aplikasi yang dibangun menggunakan bahasa pemrograman Java?
2. Metode *crawling* seperti apa yang diimplementasikan pada Apache Nutch?
3. Kinerja *crawling* mana yang lebih baik antara *web crawler* Apache Nutch dibandingkan dengan metode *crawling* lainnya, seperti *incremental web crawler*, *focused web crawler*, dan *distributed web crawler*?

1.3 Tujuan

Berdasarkan rumusan masalah yang telah diuraikan sebelumnya, maka tujuan dari penelitian ini adalah:

1. Mengembangkan perangkat lunak berbasis Java yang dapat mempekerjakan *web crawler* Apache Nutch melalui REST API.
2. Memahami metode *crawling* yang diimplementasikan pada *web crawler* Apache Nutch.
3. Membandingkan hasil eksperimen *web crawler* Apache Nutch dengan metode *crawling* lainnya untuk mengetahui kinerja *web crawler* yang lebih baik.

1.4 Batasan Masalah

Karena fokus penelitian pada skripsi ini adalah mempekerjakan *web crawler* Apache Nutch melalui REST API, maka berikut ini adalah batasan-batasan masalah yang diterapkan pada perangkat lunak yang dikembangkan:

1. Halaman *web* yang dilakukan *crawl* adalah halaman *web* yang kontennya berupa teks.
2. *User* yang bertindak sebagai admin tidak dapat melakukan perubahan pada informasi berupa nama, *username*, dan *password*.
3. Pemanggilan Nutch REST API diimplementasikan sebagai aplikasi Java yang memanggil Nutch REST API khususnya, *generate*, *fetch*, *parse*, dan *update database*.
4. Aplikasi Java yang diimplementasikan sebagai pemanggilan Nutch REST API dijalankan setelah *seed* URL dimasukkan ke dalam basis data.

1.5 Metodologi

Berikut ini adalah metodologi yang diterapkan dalam penelitian pada skripsi ini:

1. Studi literatur mengenai konsep *web crawling* dan *web crawler*.
2. Studi literatur mengenai sistem terdistribusi Hadoop.
3. Studi literatur mengenai *RESTful web service*.
4. Studi literatur dan eksplorasi mengenai konsep, arsitektur, metode *crawling*, dan REST API Apache Nutch.
5. Instalasi dan konfigurasi *web crawler* Apache Nutch di atas Hadoop.
6. Mengembangkan aplikasi Java yang mengimplementasikan pemanggilan Nutch REST API untuk mempekerjakan *web crawler* Apache Nutch yang sudah dikonfigurasi.
7. Merancang dan mengembangkan perangkat lunak berbasis *web* untuk mengakses hasil *crawling* yang dihasilkan dari aplikasi Java yang mempekerjakan *web crawler* Apache Nutch.
8. Menguji perangkat lunak *web crawler* Apache Nutch baik fungsional maupun performa pada lingkungan terdistribusi Hadoop.
9. Melaporkan hasil pengujian fungsional terhadap situs induk J2EE dan eksperimen terhadap *web crawler* Apache Nutch di atas lingkungan terdistribusi Hadoop.
10. Membandingkan hasil eksperimen *crawling web crawler* Apache Nutch dengan hasil eksperimen *incremental web crawler*, *focused web crawler*, dan *distributed web crawler*.

1.6 Sistematika Pembahasan

Sistematika penulisan dalam skripsi ini adalah sebagai berikut:

- Bab 1 Pendahuluan
  Bab ini berisi latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi, dan sistematika pembahasan.
- Bab 2 Landasan Teori
  Bab ini berisi landasan teori yang mendukung penelitian ini. Landasan teori yang dibahas mencakup teori *information retrieval* yang meliputi teori *web graph* dan teori pemilihan *seed* URL, teori dan dan arsitektur *web crawler* secara umum, fitur *web crawler*, teori *spider trap*, beberapa jenis *web crawler*, teori, Hadoop, HBase, teori *web crawler* Apache Nutch, *RESTful web service*, dan lingkungan pengembangan aplikasi *Java Platform, Enterprise Edition* (J2EE).
- Bab 3 Analisis
  Bab ini berisi analisis algoritma *crawling* dan pencarian kata pada perangkat lunak, diagram *use case* serta skenario untuk menggambarkan kebutuhan atribut dan fungsional yang diharapkan dari perangkat lunak yang dibangun, analisis skema perangkat lunak yang dibangun dan diagram kelas sederhana, hasil eksplorasi mengenai Hadoop dan lingkungannya yang terdiri dari Hadoop, Zookeeper, dan HBase, dan hasil eksplorasi mengenai *web crawler* Apache Nutch.
- Bab 4 Perancangan
  Bab ini berisi perancangan yang dibutuhkan untuk membangun perangkat lunak meliputi perancangan antarmuka, perancangan basis data, dan diagram kelas rinci situs induk J2EE dan aplikasi Java yang mengimplementasi pemanggilan *web crawler* Apache Nutch via Nutch REST API.

- Bab 5 Implementasi
  Bab ini berisi implementasi perangkat lunak yang meliputi lingkungan implementasi perangkat lunak maupun perangkat keras keras, arsitektur perangkat lunak hasil implementasi dari bab 3, hasil implementasi perangkat lunak dan basis data berdasarkan rancangan pada bab 4, pengujian pada perangkat lunak yang dibangun, yaitu pengujian fungsional yang dilakukan untuk memastikan bahwa fitur-fitur pada situs induk J2EE sudah berjalan dengan baik, dan eksperimen performa yang dilakukan untuk mengetahui performa agen *web crawler* Apache Nutch jika dijalankan pada lingkungan terdistribusi Hadoop. Pada bagian akhir bab ini, hasil eksperimen performa *crawling web crawler* Apache Nutch dibandingkan dengan hasil eksperimen performa yang telah dilakukan pada *incremental web crawler*, *focused web crawler*, dan *distributed* web crawler dengan tujuan untuk menentukan *web crawler* yang performanya lebih baik pada lingkungan terdistribusi Hadoop.

- Bab 6 Kesimpulan dan Saran
  Bab ini berisi kesimpulan dan saran berdasarkan hasil analisis, perancangan, implementasi, pengujian, dan eksperimen yang telah dilakukan.