# SKRIPSI

# *WEB CRAWLING* TERDISTRIBUSI PADA LINGKUNGAN HADOOP

**Gabriella**

**NPM: 2014730013**

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

UNDERGRADUATE THESIS

# DISTRIBUTED WEB CRAWLING ON HADOOP ENVIRONMENT


**Gabriella**

**NPM: 2014730013**

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

# LEMBAR PENGESAHAN

# *WEB CRAWLING* TERDISTRIBUSI PADA LINGKUNGAN HADOOP


**Gabriella**

**NPM: 2014730013**


**Bandung, 31 Mei 2018**

**Menyetujui,**

**Pembimbing**

**Gede Karya, M.T., CISA, IPM**

**Ketua Tim Penguji**

**Vania Natali, M.T.**

**Anggota Tim Penguji**

**Pascal Alfadian, M.Comp.**

**Mengetahui,**

**Ketua Program Studi**

**Mariskha Tri Adithia, P.D.Eng**

# PERNYATAAN

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

## *WEB CRAWLING* TERDISTRIBUSI PADA LINGKUNGAN HADOOP

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 31 Mei 2018


Gabriella
NPM: 2014730013

## ABSTRAK

*Web crawler* merupakan program yang melakukan *web scanning* dan *data indexing* dengan metode tertentu. Salah satu metode *crawling* yakni *distributed web crawling* yang memanfaatkan banyak agen *crawler* dengan tujuan mempercepat proses *crawling*. Data hasil *crawling* dapat berukuran besar, atau yang sering disebut sebagai *big data*, sehingga diperlukan media penyimpanan yang dapat mengolah *big data*. Untuk mengatasi masalah ini, dapat digunakan Hadoop ecosystem, seperti: HDFS dan HBase. Oleh karena itu, pada penelitian ini dikembangkan aplikasi *distributed web crawler* pada lingkungan Hadoop *Cluster* (HBase). Penelitian juga mencakup ekperimen untuk menjawab pertanyaan: (1) bagaimana pengaruh banyaknya *crawler* terhadap kecepatan *crawling*; dan (2) bagaimana kecepatan pemrosesan data menggunakan Hadoop.

Untuk menyelesaikan permasalahan ini, dilakukan analisis dan perancangan arsitektur dan algoritma *distributed crawling*. Setelah itu, dikembangkan aplikasi *web crawler* yang terdiri atas situs induk, *server crawler*, dan agen *crawler* terdistribusi. Berikutnya, disiapkan Hadoop *Cluster* (HDFS dan HBase) untuk pengujian dan pengumpulan URL *seed* untuk data uji. Langkah terakhir yang dilakukan yakni melaksanakan pengujian dan eksperimen performa *distributed crawler*.

Pengujian dilakukan untuk memastikan perangkat lunak telah berfungsi sebagaimana mestinya. Pengujian ini dilakukan berdasarkan *usecase* dan skenario yang diperoleh dari analisis kebutuhan perangkat lunak. Berdasarkan pengujian fungsional tersebut, didapati perangkat lunak sudah berjalan sebagaimana mestinya. Eksperimen dilakukan untuk melihat performa *crawler*; dilakukan dengan menjalankan aplikasi *web crawler* pada empat komputer yang tergabung dalam *cluster* Hadoop; satu *master* dan tiga *slave*, dimana masing-masing komputer menjalankan lima agen *crawler*. Ada dua eksperimen yang dilakukan, yakni eksperimen *crawling* dan eksperimen *searching*. Eksperimen *crawling* dilakukan untuk melihat bagaimana pengaruh banyaknya *crawler* dan *node* komputer pada *cluster* Hadoop yang digunakan terhadap kecepatan *crawling*. Eksperimen *searching* dilakukan untuk melihat bagaimana pengaruh banyaknya *node* komputer pada *cluster* Hadoop yang digunakan terhadap kecepatan pencarian.

Pada eksperimen *crawling*, didapatkan rata-rata kecepatan *crawling* bertambah sebesar 1.16 kali untuk setiap penambahan satu *node* dan lima agen. Pada eksperimen *searching*, didapatkan rata-rata peningkatan kecepatan pencarian adalah sebesar 1.19 kali untuk setiap penambahan satu *node*. Berdasarkan kedua hasil eksperimen tersebut, dapat disimpulkan bahwa semakin banyak agen dan komputer dalam *cluster* Hadoop yang digunakan, maka semakin cepat pula proses *crawling*. Selain itu, semakin banyak komputer dalam *cluster* Hadoop yang digunakan, semakin cepat pula waktu pencarian.



**Kata-kata kunci:** *web crawler*, *distributed*, Hadoop

## ABSTRACT

Web crawler is a program that performs web scanning and data indexing with certain methods. One of its method is distributed web crawling that utilizes many crawler agents with the aim of speeding up the crawling process. Data from the crawling result can be large, which is often referred to as big data. Therefore, a storage media which can process a big data is required. To solve this problem, Hadoop ecosystem, such as HDFS and HBase, can be used. On this study, a web crawler application was developed on Hadoop Cluster environment (HBase). The study also includes experiments to answer: how the number of crawlers affects crawling speed; and the speed of data processing on Hadoop environment.

To solve these problems, analysis and design of distributed crawler architecture and algorithm are performed. After that, a web crawler application that consists of main site, crawler server, and distributed crawler agent was developed. Next, Hadoop Cluster (HDFS and HBase) for testing was prepared and the URL seed were collected as a test data. The last step is to conduct testing and performance experiment on the distributed crawler.

Testing is done to ensure the software is working properly. This test is based on the usecase and scenarios obtained from the software requirement analysis. Based on the functional testing, the software is found to be working properly. The experiment was conducted to see the performance of the crawler; done by running the web crawler application on four computers which were part of the Hadoop cluster; one master and three slaves, where each computer run five crawler agents. There are two experiments conducted, namely crawling experiment and searching experiment. Crawling experiment were conducted to see how the number of crawlers and computer nodes on the Hadoop cluster used affects crawling speed. The searching experiment was conducted to see how the number of computer nodes on the Hadoop cluster used affects searching speed.

Based on the crawling experiment result, the average crawling speed increased by 1.16 times for each addition of one node and five agents. Based on the searching experiment result, the average searching speed increased 1.19 times for each addition of one node. Based on these two experiments results, it can be concluded that the more agents and computers in the Hadoop cluster are used, the faster the crawling process can be done. In addition, the more computers in the Hadoop cluster are used, the faster the searching process can be done.



**Keywords:** web crawler, distributed, hadoop

*Dipersembahkan untuk orang tua dan adik penulis*

# KATA PENGANTAR

Puji syukur penulis haturkan pada Tuhan yang Maha Esa atas segala berkat dan rahmat-Nya, sehingga skripsi berjudul "*Web Crawling* Terdistribusi Pada Lingkungan Hadoop" dapat penulis selesaikan dengan baik. Skripsi ini dibuat untuk melengkapi persyaratan kelulusan tingkat sarjana (S-1) di Fakultas Teknologi Informasi dan Sains Jurusan Teknik Informatika Universitas Katolik Parahyangan. Selama pembuatan skripsi ini, penulis mendapatkan bantuan dan dukungan dari berbagai pihak. Maka dari itu, pada kesempatan ini penulis hendak mengucapkan terima kasih kepada:

1. Bapak Gede Karya, M.T., CISA, IPM selaku dosen pembimbing yang telah menyediakan waktu untuk membimbing penulis selama penyusunan skrispi.
2. Ibu Vania Natali, M.T. dan Bapak Pascal Alfadian, M.Comp. selaku dosen penguji yang telah memberi kritik dan saran untuk skripsi ini.
3. Orang tua dan adik penulis yang selalu memberi doa dan dukungan selama penyusunan skrispi.
4. Sahabat dan rekan-rekan penulis yang telah banyak membantu dan memberi dukungan selama penyusunan skrispi.

Akhir kata, penulis menyadari bahwa skripsi ini belum sempurna. Namun, penulis berharap skripsi ini dapat bermanfaat bagi para pembaca yang ingin menambah wawasan ataupun melakukan penelitian yang serupa.

Bandung, Mei 2018

Penulis

# DAFTAR ISI

# DAFTAR GAMBAR

# DAFTAR TABEL

# BAB 1

# PENDAHULUAN

Pada bab ini dilakukan pembahasan mengenai latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi, dan sistematika pembahasan.

## 1.1 Latar Belakang

Pada jaman sekarang, orang-orang diberi kemudahan dalam mencari informasi dengan adanya *web*. *Web* atau *World Wide Web* merupakan suatu "ruang" informasi yang berisi *resource* dalam berbagai bentuk, misalnya dokumen dan gambar. Tiap *web resource* berada pada suatu tempat yang disebut halaman *web*, dan memiliki alamat atau *Uniform Resource Locator* (URL) sebagai penanda lokasi halaman *web* tersebut. Agar suatu halaman dapat di-*retrieve* ketika melakukan pencarian, *search engine* yang digunakan harus terlebih dahulu mengumpulkan informasi mengenai halaman tersebut. Pengumpulan informasi ini dilakukan dengan menggunakan *web crawler*.

*Web crawler* merupakan program yang melakukan *web scanning* dan *data indexing* dengan metode tertentu. Ada berbagai metode *crawling*, antara lain: *batch and incremental crawling*; *deep crawling*; *distributed crawling*; *focused crawling*; dll. Data yang dihasilkan dari proses *crawling* berupa URL atau halaman-halaman *web*. Pada skripsi ini, dibuat *web crawler* dengan metode *distributed crawling*. *Distributed web crawling* memanfaatkan banyak *crawler* untuk melakukan proses *crawling* dengan tujuan untuk mempercepat proses *crawling*.

Data hasil *crawling* dapat berukuran besar atapun berjumlah banyak. Jenis data seperti ini sering disebut sebagai *big data*. Oleh karena itu, diperlukan suatu media penyimpanan yang dapat memproses atau mengolah *big data*. Hadoop merupakan salah satu *framework* yang digunakan untuk penyimpanan terdistribusi dan pemrosesan *big data*. Inti dari Hadoop terdiri atas bagian penyimpanan yang disebut *Hadoop Distributed File System* (HDFS) dan bagian pemrosesan yang menggunakan model pemrograman MapReduce.

Aplikasi *distributed web crawler* yang dikembangkan memanfaatkan teknologi HBase, yakni *database* Hadoop yang berjalan diatas HDFS. Oleh karena *web crawler* yang dibuat menggunakan metode *distributed web crawling* dan memanfaatkan Hadoop sebagai media penyimpanan data, maka dilakukan pula ekperimen untuk menjawab pertanyaan: (1) bagaimana pengaruh banyaknya *crawler* terhadap kecepatan *crawling*; dan (2) bagaimana kecepatan pemrosesan data menggunakan Hadoop.

## 1.2 Rumusan Masalah

Salah satu masalah dalam *web crawling* adalah proses yang memerlukan waktu yang cukup lama. Untuk itu, pada skripsi ini ingin diselesaikan masalah peningkatan kecepatan *crawling* dengan agen terdistribusi. Oleh karena itu, permasalahan dapat dirumuskan sebagai berikut:

1. Bagaimana mekanisme pembagian tugas dan koordinasi antar agen sehingga bisa meningkatkan kecepatan *crawling*?

2. Bagaimana mekanisme penyimpanan informasi oleh masing-masing agen pada lingkungan *file system* terdistribusi Hadoop?

## 1.3 Tujuan

Tujuan dari skripsi ini adalah sebagai berikut:

1. Memahami mekanisme pembagian tugas dan koordinasi antar agen sehingga bisa meningkatkan kecepatan *crawling*.
2. Memahami mekanisme penyimpanan data oleh masing-masing agen pada lingkungan *file system* terdistribusi Hadoop.
3. Membangun perangkat lunak *web crawler* dengan metode *distributed web crawling* yang memanfaatkan Hadoop sebagai media penyimpanan URL.
4. Melakukan eksperimen untuk melihat pengaruh banyaknya agen dan *node* komputer pada *cluster* Hadoop terhadap kecepatan *crawling* dan pencarian.

## 1.4 Batasan Masalah

Karena fokus penelitian pada skripsi ini adalah mengenai kecepatan *crawling* dan mekanisme penyimpanan hasil *crawling* pada lingkungan Hadoop, maka batasan-batasan berikut diterapkan pada aplikasi yang dikembangkan:

1. Proses *crawling* tidak memproses *crawler policy*. *Crawler policy* memiliki aturan *politeness* yang memungkinkan suatu *web* tidak bisa di-*crawl* oleh *web crawler*. Jika aturan ini diproses maka keakuratan hasil penelitian kecepatan *crawling* dan pemrosesan data dapat berkurang karena bisa saja ada URL yang seharusnya dapat diproses, tetapi karena *policy* tersebut akhirnya URL tidak jadi diproses.
2. Penelusuran URL hanya berdasarkan HTML. Penelusuran *active link* seperti *javascript* tidak akan dilakukan karena memerlukan teknik penelusuran yang berbeda dan tidak terkait fokus penelitian dari skripsi ini.
3. Hasil yang disimpan berupa teks yang terdapat pada halaman *web*. Penyimpanan *tag* diperlukan jika yang dibuat adalah *search engine*, yang berada diluar fokus penelitian pada skripsi ini.

## 1.5 Metodologi

Berikut ini adalah langkah-langkah yang dilakukan dalam membangun perangkat lunak:

1. Studi pustaka tentang konsep *web crawler*.
2. Studi pustaka dan eksplorasi tentang sistem terdistribusi berbasis Hadoop.
3. Studi teknik *distributed crawling*.
4. Studi teknik penyimpanan pada Hadoop.
5. Menganalisis pemilihan penyimpanan pada Hadoop.
6. Mengembangkan perangkat lunak *web crawler* yang terdiri atas situs induk, *server crawler* dan agen *crawler*.
7. Melaksanakan pengujian dan eksperimen performa pada perangkat lunak.

## 1.6 Sistematika Pembahasan

Sistematika pembahasan dalam skripsi ini adalah sebagai berikut:

1. Bab 1 Pendahuluan
   Bab ini membahas tentang latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi, dan sistematika pembahasan.

2. Bab 2 Landasan Teori
   Bab ini membahas tentang teori-teori yang menunjang pengembangan perangkat lunak, antara lain teori *web crawler*, *big data*, Hadoop, Java 2 *Enterprise Edition*, dan *library jsoup*.

3. Bab 3 Eksplorasi dan Analisis
   Bab ini membahas tentang eksplorasi Hadoop, analisis masalah dan usulan solusi, serta analisis kebutuhan perangkat lunak.

4. Bab 4 Perancangan
   Bab ini membahas tentang perancangan *User Interface* dan *database*, serta diagram kelas rinci.

5. Bab 5 Implementasi, Pengujian, dan Eksperimen
   Bab ini membahas implementasi, pengujian, dan eksperimen.

6. Bab 6 Kesimpulan dan Saran
   Bab ini membahas tentang kesimpulan yang didapatkan dan saran yang dapat diberikan berdasarkan hasil analisis, perancangan, implementasi, pengujian, dan eksperimen yang telah dilakukan.