

SKRIPSI

**PENGEMBANGAN PERANGKAT LUNAK PENGUMPUL
BIG DATA BERBASIS WEB, PRAOLAH DAN
PENYIMPANAN PADA SISTEM TERDISTRIBUSI HADOOP**



Roy Vito Joshua

NPM: 2012730019

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

UNDERGRADUATE THESIS

**WEB BASED BIG DATA COLLECTOR SOFTWARE
DEVELOPMENT, PRE-PROCESSING AND STORING ON
HADOOP DISTRIBUTED SYSTEM**



Roy Vito Joshua

NPM: 2012730019

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

LEMBAR PENGESAHAN



PENGEMBANGAN PERANGKAT LUNAK PENGUMPUL BIG DATA BERBASIS WEB, PRAOLAH DAN PENYIMPANAN PADA SISTEM TERDISTRIBUSI HADOOP

Roy Vito Joshua

NPM: 2012730019

Bandung, 11 Desember 2018

Menyetujui,

Pembimbing

Dr. Veronica Sri Moertini

Ketua Tim Penguji

Rosa De Lima, M.Kom.

Anggota Tim Penguji

Dr.rer.nat. Cecilia Esti Nugraheni

Mengetahui,

Ketua Program Studi

Mariskha Tri Adithia, P.D.Eng



PERNYATAAN

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

PENGEMBANGAN PERANGKAT LUNAK PENGUMPUL BIG DATA BERBASIS WEB, PRAOLAH DAN PENYIMPANAN PADA SISTEM TERDISTRIBUSI 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 11 Desember 2018



Roy Vito Joshua
NPM: 2012730019

ABSTRAK

Internet merupakan salah satu kebutuhan yang penting dalam hidup manusia. Dalam kehidupan sehari-hari setiap orang menggunakan internet untuk berbagai keperluan seperti, mengakses atau membagikan informasi, berkomunikasi, melakukan jual-beli barang atau jasa, belajar-mengajar, dan lain-lain. Konten-konten tersebut dapat diakses melalui *website*. Setiap *website* menawarkan berbagai macam informasi maupun produk kepada penggunanya.

Website dibagi menjadi beberapa jenis yaitu, *online shop*, media sosial, *e-learning*, dan lain-lain. Dari beberapa jenis *website* tersebut, banyak data-data yang berguna untuk dikumpulkan menjadi kumpulan data yang sangat besar dalam kata lain *big data*. Terutama data-data dari *website* seperti media sosial dan *online shop* cukup banyak dan bervariasi.

Dari salah satu sosial media yaitu twitter, memberikan akses API yang cukup lengkap sesuai dengan data-data yang tersebar secara publik melalui antarmuka twitter. Namun, untuk *online shop* seperti TripAdvisor dan masih banyak lainnya tidak memberikan akses API. Maka dari itu, dibutuhkan *web crawler* untuk mendapatkan data-datanya.

Data-data yang diambil dari sosial media dan *online shop* dapat diolah dan dianalisis sehingga menghasilkan pengetahuan yang berharga. Hal tersebut akan tercapai jika data-data yang diambil relevan dengan tujuan data tersebut diambil. Salah satu contohnya adalah tren yang bisa diambil dari *hashtag* terbanyak yang dikandung setiap *tweet*. Contoh lainnya adalah sentimen ulasan produk atau jasa yang jika dikumpulkan dengan banyak dapat mencerminkan baik buruknya suatu produk atau jasa.

Kata-kata kunci: Media Sosial, *Online Shop*, API, *Web Crawler*

ABSTRACT

Nowadays internet has become one of many basic human needs. So many people use internet for their necessity such as, accessing and sharing information, communicating, product and services sales, education, and so on. These type of contents can be accessed through website. Every website also offer various type of information or even a produk to the consumers.

Website is divided into two types such as, online shop, social media, e-learning, and so forth. From the types mentioned before, so much data useful data to be collected into massive data collection, in other word big data. Especially data from website like social media and online shop which contains so big and various form of data.

One of many social media accross the internet which is twitter, has been giving pretty good access to their API based on their own public data. However, online shop TripAdvisor and many others don't give API access. Therefore, web crawler is needed to collect their data without the use of API.

Data which are collected from social media and online shop can be processed and analyzed so that can result valuable knowledge. It can be achieved if the data that will be collected are relevant with the purpose of why the data are collected. One of many examples is trends which can be collected from the most plenty hashtags contained on tweets. The other example is the sentiment from consumer reviews of product or services when they are gathered can reflect the overall quality.

Keywords: Social Media, Online Shop, API, Web Crawler

Teknik Informatika Unpar dan diri sendiri

KATA PENGANTAR

Puji dan syukur kehadiran Tuhan Yang Maha Esa atas berkat rahmat serta kasih-Nya sehingga penulis dapat menyelesaikan tugas akhir dengan judul **Pengembangan Perangkat Lunak Pengumpul Big Data Berbasis Web, Praolah Dan Penyimpanan Pada Sistem Terdistribusi Hadoop** dengan baik. Dalam penyelesaian skripsi ini, penulis mendapat bantuan dari berbagai pihak. Maka dari itu, penulis mengucapkan terima kasih kepada semua pihak yang membantu terselesaikannya tugas akhir ini, di antaranya:

1. Ayah dan Ibu atas jasa-jasanya, kesabaran, doa, dan cinta kasih kepada penulis.
2. Ibu Veronica Sri Moertini sebagai dosen pembimbing yang telah membimbing penulis hingga dapat menyelesaikan tugas akhir ini.
3. Ibu Rosa De Lima dan Ibu Cecilia Esti Nugraheni sebagai dosen penguji yang telah membantu dalam menguji tugas akhir ini.
4. Theodorus Kurniawan dan Herfan Heryandi yang membantu penulis dalam mengembangkan diri.
5. Rekan-rekan administrator lab komputasi FTIS yang membantu penulis dalam menyiapkan lingkungan sistem untuk perangkat lunak di komputer lab skripsi.
6. Teman-teman Teknik Informatika Unpar angkatan 2012 yang telah berbagi ilmu dan memberikan semangat kepada penulis.
7. Pihak-pihak lain yang belum disebutkan, yang berperan dalam penyelesaian tugas akhir ini.

Akhir kata, penulis berharap agar tugas akhir ini bermanfaat bagi pembaca yang hendak melakukan penelitian dan pengembangan yang terkait dengan tugas akhir ini.

Bandung, Desember 2018

Penulis

DAFTAR ISI

KATA PENGANTAR	xv
DAFTAR ISI	xvii
DAFTAR GAMBAR	xix
DAFTAR TABEL	xxi
1 PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Tujuan	2
1.4 Batasan Masalah	2
1.5 Metodologi	2
1.6 Sistematika Pembahasan	3
2 LANDASAN TEORI	5
2.1 <i>Big Data</i>	5
2.2 Twitter API	6
2.2.1 Pendaftaran Twitter API	6
2.2.2 Standard Search API	7
2.2.3 <i>Tweet Objects</i>	8
2.3 <i>Library</i> Abraham TwitterOAuth	10
2.3.1 TwitterOAuth	10
2.4 <i>CSS Selector</i>	11
2.5 <i>Library Goutte Web Crawler</i>	12
2.6 Hadoop	12
2.6.1 Definisi Hadoop	12
2.6.2 <i>Hadoop Distributed File System</i>	13
2.6.3 <i>Hadoop MapReduce</i>	14
3 ANALISIS	17
3.1 Analisis Deskripsi Perangkat Lunak	17
3.2 Analisis Arsitektur Perangkat Lunak	18
3.3 Analisis <i>Use Case</i> dan Skenario	18
3.4 Analisis Data	23
3.4.1 Analisis Data Twitter	23
3.4.2 Analisis Data TripAdvisor	26
4 PERANCANGAN	31
4.1 Perancangan Antarmuka	31
4.1.1 Perancangan Antarmuka <i>User</i>	31
4.1.2 Perancangan Antarmuka <i>Admin</i>	34

4.2	Perancangan Diagram Kelas Perangkat Lunak Berbasis Java	35
4.3	Perancangan <i>Flow Chart</i> Pengumpul Data	37
4.4	Perancangan <i>Flow Chart</i> Praolah Data	38
5	IMPLEMENTASI DAN EKSPERIMEN	41
5.1	Implementasi	41
5.1.1	Lingkungan Implementasi	41
5.1.2	Hasil Implementasi	42
5.2	Eksperimen	58
5.2.1	Tujuan Eksperimen	58
5.2.2	Cara Melakukan Eksperimen	58
6	KESIMPULAN DAN SARAN	69
6.1	Kesimpulan	69
6.2	Saran	69
	DAFTAR REFERENSI	71
	A KODE PROGRAM MODEL	73
	B KODE PROGRAM VIEW	85
	C KODE PROGRAM CONTROLLER	119

DAFTAR GAMBAR

2.1	Gambar Pendaftaran Perangkat Lunak di <i>Website</i> Pengembang Twitter	6
2.2	Gambar <i>Keys</i> dan <i>Access Tokens</i> di <i>Website</i> Pengembang Twitter	7
2.3	Gambar Tipe <i>Search Tweets</i> di <i>Website</i> Pengembang Twitter	8
2.4	Gambar <i>Tweet</i> dari <i>Website</i> Twitter	8
2.5	Gambar Tweet Berbentuk JSON dari <i>Website</i> Twitter	9
2.6	Gambar Penulisan CSS	11
2.7	Gambar Diagram Merepresentasikan <i>Namenode</i> dan <i>Datanode</i>	13
2.8	Gambar Contoh Ilustrasi Fungsi <i>Mapper</i>	14
2.9	Gambar Contoh Ilustrasi Fungsi <i>Reducer</i>	15
3.1	Gambar Contoh Ulasan Hotel dari <i>Website</i> TripAdvisor	18
3.2	Gambar Arsitektur Perangkat Lunak Pengumpul <i>Big Data</i> Berbasis <i>Web</i>	18
3.3	Gambar Diagram <i>Use Case</i> Perangkat Lunak Pengumpul <i>Big Data</i> Berbasis <i>Web</i>	19
3.4	Gambar Data JSON dari Twitter yang Terdiri dari <i>Statuses</i> dan <i>Search Metadata</i>	23
3.5	Gambar Data JSON Terperinci dari Variabel <i>Statuses</i>	24
3.6	Gambar Data JSON Terperinci dari Variabel <i>Search_Metadata</i>	24
3.7	Gambar Data JSON Terperinci dari Sebuah Objek <i>Tweet</i>	25
3.8	Gambar Data JSON Terperinci dari Sebuah Objek <i>Entities</i>	26
3.9	Gambar Ulasan Hotel Per Halaman di <i>Website</i> TripAdvisor	26
3.10	Gambar Ulasan Tempat Wisata Per Halaman di <i>Website</i> TripAdvisor	27
3.11	Gambar <i>URL</i> Ulasan Hotel Per Halaman di <i>Website</i> TripAdvisor	27
3.12	Gambar Tampilan Kelas-Kelas Elemen HTML Diakses Menggunakan <i>Chrome DevTools</i>	28
3.13	Gambar Contoh Data Ulasan Hotel Berformat JSON dari <i>Website</i> TripAdvisor	28
4.1	Gambar Antarmuka Halaman <i>Login</i>	31
4.2	Gambar Antarmuka Halaman <i>Signup</i>	32
4.3	Gambar Antarmuka Halaman <i>Home</i>	32
4.4	Gambar Antarmuka Halaman <i>Home</i>	33
4.5	Gambar Antarmuka Halaman <i>Account</i>	33
4.6	Gambar Antarmuka Halaman <i>Login Admin</i>	34
4.7	Gambar Antarmuka Halaman <i>Users</i>	34
4.8	Gambar Antarmuka Halaman <i>Get Data from Twitter</i>	35
4.9	Gambar Antarmuka Halaman <i>Get Data from Twitter</i>	35
4.10	Gambar Diagram Kelas Praolah	36
4.11	Gambar <i>Flow Chart</i> Pengumpulan Data	38
4.12	Gambar <i>Flow Chart</i> Praolah Data	39
5.1	Gambar Halaman <i>Login</i> Khusus untuk Pengguna	42
5.2	Gambar Halaman <i>Signup</i>	43
5.3	Gambar Halaman <i>Home</i>	44
5.4	Gambar Masukan Sumber Data Pada Halaman <i>Home</i>	44
5.5	Gambar Masukan Tipe Ulasan dari TripAdvisor Pada Halaman <i>Home</i>	44

5.6	Gambar Masukan Kueri Untuk Twitter Pada Halaman <i>Home</i>	45
5.7	Gambar Masukan Bahasa Untuk Pengumpulan Data dari Twitter Pada Halaman <i>Home</i>	45
5.8	Gambar Halaman <i>Search Result</i>	46
5.9	Gambar Halaman <i>Search Result</i> Terperinci	47
5.10	Gambar Halaman <i>MyAccount</i>	48
5.11	Gambar <i>Method AuthenticateTwitter</i> di Kelas Twitter	49
5.12	Gambar <i>Method SearchTweets</i> di Kelas Twitter	49
5.13	Gambar Halaman <i>Login</i> Khusus untuk <i>Admin</i>	50
5.14	Gambar Halaman <i>Main</i>	51
5.15	Gambar Halaman <i>Users-On Wait</i>	52
5.16	Gambar Halaman <i>Users-Approved</i>	53
5.17	Gambar Halaman <i>Users-On hold</i>	54
5.18	Gambar Halaman <i>Query Expiration</i>	55
5.19	Gambar Halaman <i>Run Search Request</i> Twitter	56
5.20	Gambar Halaman <i>Run Search Request Website</i>	57
5.21	Gambar Halaman <i>Export Directories</i>	58
5.22	Gambar Versi Java pada Sistem Operasi	59
5.23	Gambar <i>Node-Node</i> yang Harus Ada di Hadoop	61
5.24	Gambar Antarmuka <i>Hadoop Distributed File System</i>	62
5.25	Gambar Antarmuka <i>Hadoop Cluster</i>	62
5.26	Gambar Membuat Direktori Data Pada <i>Hadoop Distributed File System</i>	62
5.27	Gambar Membuat Direktori Twitter Pada <i>Hadoop Distributed File System</i>	63
5.28	Gambar Mengekspor Direktori Lokal ke <i>Hadoop Distributed File System</i>	63
5.29	Gambar Hasil Hadoop <i>MapReduce</i> Pada Data Twitter	64
5.30	Gambar Kelas <i>Mapper</i> untuk Praolah Twitter	65
5.31	Gambar Kelas <i>Reducer</i> untuk Praolah Twitter	65
5.32	Gambar Kelas <i>Mapper</i> untuk Praolah TripAdvisor	66
5.33	Gambar Kelas <i>Reducer</i> untuk Praolah TripAdvisor	66
5.34	Gambar Membuat Direktori Baru Pada <i>Hadoop Distributed File System</i>	67
5.35	Gambar Mengekspor Direktori Lokal ke <i>Hadoop Distributed File System</i>	67
5.36	Gambar Hasil Hadoop <i>MapReduce</i> Pada Data TripAdvisor	68

DAFTAR TABEL

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Di masa kini, internet merupakan salah satu kebutuhan yang penting dalam hidup manusia. Dalam kehidupan sehari-hari setiap orang menggunakan internet untuk berbagai keperluan seperti, mengakses atau membagikan informasi, berkomunikasi, melakukan jual-beli barang atau jasa, belajar-mengajar, dan lain-lain. Konten-konten tersebut dapat diakses melalui *website*. Setiap *website* menawarkan berbagai macam informasi maupun produk kepada penggunanya.

Website dibagi menjadi beberapa jenis yaitu, *online shop*, media sosial, *e-learning*, dan lain-lain. Namun dalam skripsi ini hanya berfokus pada *online shop* dan media sosial. Di *online shop* dan media sosial terpapar berbagai macam data. Data-data yang terpapar di *website* dapat dikumpulkan bergantung pada akses yang diberikan oleh pemilik *website*.

Beberapa pihak pemilik *website* memberikan akses data-data yang dimiliki, tetapi ada pihak pemilik *website* yang tidak memberikan akses. *Website* yang terbuka terhadap akses data-datanya memberikan akses melalui API atau *library* untuk memudahkan pengambilan data. API adalah singkatan dari *Application Programming Interface*. Dokumentasi penggunaan API dapat ditemukan di masing-masing *website*.

Website yang tidak diberikan API untuk mengakses data-datanya dapat diambil datanya menggunakan *web crawler*. *Web crawler* yang digunakan adalah Goutte. Goutte adalah *library web crawler* yang dibuat menggunakan bahasa PHP, sangat cocok untuk pengembangan perangkat lunak pengumpul data berbasis *web* [1]. Dengan menggunakan *web crawler*, pengguna dapat menentukan data dan melakukan kustomisasi format data yang akan dikumpulkan.

Sumber data yang dikumpulkan adalah dari media sosial Twitter dan *online shop* TripAdvisor. Pihak Twitter menyediakan API berupa *url* yang dapat meminta *request* tertentu dan mengembalikan data dengan format json. Namun untuk memudahkan penggunaan API dari Twitter, tersedia *library* yang membungkus API Twitter bernama Abraham TwitterOAuth. *Library* ini mempermudah penggunaan Twitter API yang harus menggunakan *curl* terlebih dahulu untuk meminta *request* melalui *url* yang diberikan agar diberikan *response* berupa data berbentuk json.

TripAdvisor tidak menyediakan API untuk pengambilan data, tetapi *website* TripAdvisor memungkinkan untuk dilakukan *web crawling*. *Web crawling website* tersebut menggunakan *web crawler* yaitu Goutte. Goutte menjelajahi *url* yang telah diberikan dan mengembalikan *response* HTML. Hasil *response* dapat diseleksi datanya berdasarkan *CSS selector*. Tetapi Goutte hanya dapat mengambil data dari HTML saja, jika ada data yang ditambahkan menggunakan *javascript* tidak dapat diambil oleh Goutte.

Data-data yang dikumpulkan secara terus-menerus dari berbagai macam *website* disebut *big data*. *Big data* dalam kata lain adalah kumpulan data-data yang sangat banyak. Data-data yang dikumpulkan akan disimpan ke dalam lingkungan Hadoop sebagai HDFS. Hadoop memiliki sistem *file* yang terdistribusi untuk beroperasi dalam suatu perangkat keras, dinamakan *Hadoop Distributed File System* kepanjangan dari HDFS.

Data-data yang tersimpan di HDFS akan dipraolah menggunakan Hadoop. Hadoop adalah *framework* yang dapat memproses *big data* yang tersebar dan terdistribusi di kelompok komputer

yang memiliki Hadoop. Pada akhirnya, data-data yang telah dipraolah dapat diunduh dan dilihat hasilnya melalui antarmuka yang disediakan oleh Hadoop.

1.2 Rumusan Masalah

Rumusan masalah yang akan dibahas pada skripsi ini:

1. Bagaimana cara mengumpulkan data dari Twitter dan TripAdvisor?
2. Bagaimana cara menggunakan Hadoop untuk menyimpan *file* data ke dalam HDFS dan melakukan praolah data?

1.3 Tujuan

Tujuan-tujuan yang hendak dicapai pada skripsi ini:

1. Membangun perangkat lunak untuk mengumpulkan data-data dari Twitter dan TripAdvisor.
2. Melakukan eksperimen Hadoop dalam menyimpan data-data ke dalam HDFS dan mempraolah data-data.

1.4 Batasan Masalah

Beberapa batasan yang dibuat terkait dengan pengerjaan skripsi ini:

1. Web Crawler hanya dapat mengambil data html, tidak dapat mengambil data yang ditambahkan menggunakan *javascript*.

1.5 Metodologi

Metodologi yang dilakukan pada skripsi ini:

1. Melakukan studi mengenai Twitter API dan *web crawler* untuk mengambil data dari Twitter dan *website* TripAdvisor. CSS Selector digunakan untuk menyeleksi data yang akan diambil dan Chrome DevTools untuk melihat elemen html yang dapat diambil menggunakan CSS *selector*.
2. Melakukan pengumpulan data-data menggunakan *library* Abraham TwitterOAuth dan *web crawler* Goutte.
3. Melakukan studi literatur tentang Hadoop.
4. Merancang perangkat lunak pengumpul data.
5. Mengimplementasikan perangkat lunak berbasis web yang dapat diakses oleh pengguna dan admin.
6. Melakukan eksperimen dari data-data yang telah dikumpulkan menggunakan Hadoop *Map-Reduce*.
7. Membuat dokumentasi.

1.6 Sistematika Pembahasan

Sistematika penulisan setiap bab pada skripsi ini:

1. Bab Pendahuluan
Bab 1 berisi latar belakang, rumusan masalah, tujuan, metodologi, dan sistematika penulisan untuk menyusun skripsi ini.
2. Bab Dasar Teori
Bab 2 berisikan teori-teori yang digunakan dalam pembuatan skripsi ini. Teori yang digunakan yaitu, *Big Data*, Twitter API, *library* Abraham TwitterOAuth, CSS *selector*, *library web crawler* Goutte, dan Hadoop.
3. Bab Analisis
Bab 3 berisikan analisis deskripsi perangkat lunak, analisis arsitektur perangkat lunak, analisis *use case* dan skenario, serta analisis data meliputi, analisis data twitter, dan analisis data tripadvisor.
4. Bab Perancangan
Bab 4 berisikan perancangan antarmuka perangkat lunak, meliputi perancangan antarmuka *user* dan analisis antarmuka *admin*, perancangan diagram kelas perangkat lunak berbasis java, perancangan *flow chart* pengumpul data, serta perancangan *flow chart* praolah data.
5. Bab Implementasi dan Eksperimen
Bab 5 berisikan implementasi dan eksperimen perangkat lunak. Implementasi perangkat lunak berisikan lingkungan implementasi dan hasil implementasi. Eksperimen perangkat lunak berisikan tujuan eksperimen dan cara melakukan eksperimen.
6. Bab Kesimpulan dan Saran
Bab 6 berisikan kesimpulan dari hasil pembangunan perangkat lunak beserta saran untuk pengembangan berikutnya.