



Hak cipta dan penggunaan kembali:

Lisensi ini mengizinkan setiap orang untuk menggubah, memperbaiki, dan membuat ciptaan turunan bukan untuk kepentingan komersial, selama anda mencantumkan nama penulis dan melisensikan ciptaan turunan dengan syarat yang serupa dengan ciptaan asli.

Copyright and reuse:

This license lets you remix, tweak, and build upon work non-commercially, as long as you credit the origin creator and license it on your new creations under the identical terms.

**IMPLEMENTASI WEB SCRAPING PADA WEBSITE
EDUESIA.COM UNTUK PENGUKUR KESENJANGAN
JUMLAH MAHASISWA PERGURUAN TINGGI DI
INDONESIA**

SKRIPSI



**Diajukan Guna Memenuhi Persyaratan Memperoleh
Gelar Sarjana Komputer (S.Kom.)**

Yuri Pramana

11110310002

**PROGRAM STUDI SISTEM INFORMASI
FAKULTAS TEKNOLOGI INFORMASI DAN KOMUNIKASI
UNIVERSITAS MULTIMEDIA NUSANTARA**

2015

HALAMAN PENGESAHAN SKRIPSI

IMPLEMENTASI WEB SCRAPING PADA WEBSITE DALAM KESENJANGAN PERGURUAN TINGGI DI INDONESIA

Oleh:

Yuri Pramana

Telah diujikan pada hari Senin, 10 Agustus 2015

Pukul 13.00 s/d 14.30 dan dinyatakan lulus
dengan susunan penguji sebagai berikut

Ketua Sidang

Penguji

Johan Setiawan, S.Kom., M.M., M.B.A.

Marcelli Indriana, S.Kom., M.Sc.

Pembimbing Skripsi

Wira Munggana, S.Si., M.Sc

Ketua Program Studi Sistem Informasi

Wira Munggana, S.Si., M.Sc

PERNYATAAN TIDAK MELAKUKAN PLAGIAT

Dengan ini saya:

Nama : Yuri Pramana
NIM : 11110310002
Program Studi : Sistem Informasi

Dengan ini saya menyatakan bahwa skripsi ini adalah karya ilmiah saya sendiri, dan saya tidak melakukan plagiat. Semua kutipan karya ilmiah orang lain atau lembaga lain yang dirujuk dalam skripsi ini telah saya sebutkan sumber kutipannya serta saya cantumkan di Daftar Pustaka.

Jika di kemudian hari terbukti ditemukan kecurangan atau penyimpangan baik dalam pelaksanaan maupun dalam penulisan skripsi, saya bersedia menerima konsekuensi dinyatakan TIDAK LULUS untuk mata kuliah skripsi yang telah saya tempuh.

Tangerang, 7 Juli 2015

Yuri Pramana

IMPLEMENTASI WEB SCRAPING PADA WEBSITE EDUESIA.COM UNTUK PENGUKUR KESENJANGAN JUMLAH MAHASISWA PERGURUAN TINGGI DI INDONESIA

ABSTRAKSI

Oleh : Yuri Pramana

Pendidikan merupakan aset penting bagi masa depan masing-masing individu, masyarakat, negara, dan dunia. Namun, tidak semua orang bisa mendapatkan kesempatan yang sama karena kendala dibidang ekonomi dan geografis yang memicu kesenjangan dalam penilaian publik terhadap kualitas intelektual seseorang. Berdasarkan data yang diberikan oleh UNESCO, Indonesia tercatat diurutan ke 4 dari negara ASEAN lainnya, dalam *Gross Enrolment Ratio*. Proses penelitian ini ditujukan untuk membentuk gambaran kesenjangan edukasi pada tingkat perguruan tinggi setiap daerah yang ada di Indonesia. Penelitian ini dibuat dengan teknik *Web scraping*, dan menghasilkan sebuah website yang menyajikan informasi yang valid dan akurat serta analisis rasio *Gross Enrolment Ratio* untuk memperlihatkan tingkat partisipasi pada tingkat perguruan tinggi.

Kata kunci — **Data Collection, UMN, Web Crawling, Data Cleansing, Web Scraping**

IMPLEMENTATION OF WEB SCRAPING ON EDUESIA.COM FOR MEASURING THE GAP NUMBER OF COLLEGE STUDENTS IN INDONESIA

2 ABSTRACT

By: Yuri Pramana

Education is an important asset for the future of each individual, community, nation, and world. However, not everyone can have the same opportunity as the field of economic and geographic constraints that trigger gaps in the public assessment of the quality of one's intellectual. Based on data provided by UNESCO, Indonesia recorded no. 4 from other ASEAN countries, in the level of *Gross Enrolment Ratio*. The research process is intended to form a picture of educational inequality at the college level in every region in Indonesia. This study was made with Web scraping's techniques and produce a website that presents a valid and accurate information and analysis of the ratio of *Gross Enrolment Ratio* to show the level of participation at the college level.

Keywords — Data Collection, UMN, Web Crawling, Data Cleansing, Web Scraping

UMN

Kata Pengantar

Di setiap tarikan nafas, langkah kaki, dan desiran aliran darah kita, sudah seharusnya kita selalu mengucapkan syukur atas kemudahan dan kenikmatan dalam mencapai tujuan hidup. Begitu pula dengan penulis yang saat ini telah menyelesaikan skripsi yang berjudul “Implementasi Web Scraping Pada Website Dalam Kesenjangan Perguruan Tinggi Di Indonesia” yang di tujukan kepada Program Studi Sistem Informasi, Fakultas Teknologi Informasi dan Komunikasi, Universitas Multimedia Nusantara.

Penulis menyadari bahwa tanpa bantuan dan dukungan dari berbagai pihak. Oleh karena itu penulis ingin menyampaikan rasa terimakasih kepada:

1. Wira Mungguna, S. Si., M.Sc. selaku Ketua Program Studi Sistem Informasi di Universitas Multimedia Nusantara dan sebagai Dosen Pembimbing yang selalu memberikan banyak dukungan dalam pembuatan skripsi dan penulisan laporan ini.
2. IR. Raymond Sunardi Oetama, M.C.I.S yang telah memberikan saran, dukungan, dan inspirasi dalam pembuatan dan penulisan laporan skripsi ini
3. Pak Feris Thia, dan tim IBICC yang turut memberikan wawasan mengenai *Pentaho Data Integration*.
4. Wisnu Satyagraha, Theodora Giovani dan Fededi selaku teman penulis dan rekan riset yang melewati suka duka bersama selama pembuatan riset dan penulisan laporan skripsi.
5. Yonathan Hadiputra, Gustave Lyman, Stefanus Hendri Jason Japutra, Aril Rulif, Agustyan Hidayat, Ivan Dermawan, Granodio Pratama, Reisha Pahlevi, dan teman-teman yang turut membantu memberikan saran, semangat, dan mengibur penulis selama pengerjaan laporan ini.

6. Orang Tua, saudara, dan keluarga yang selalu memberikan dukungan doa dan kepercayaan selama pengerjaan laporan skripsi ini.
7. Kepada semua pihak yang telah membantu memberikan dukungan dan tidak dapat disebutkan satu per satu

Penulis menyadari bahwa banyak kekurangan pada laporan skripsi. Semoga isi dari laporan skripsi ini dapat memberikan wawasan dan inspirasi bagi pembaca dan pihak-pihak yang membutuhkan

Tangerang 1 Juni 2015

Yuri Pramana

UMN

Daftar Isi

HALAMAN PENGESAHAN SKRIPSI.....	ii
PERNYATAAN TIDAK MELAKUKAN PLAGIAT	iii
ABSTRAKSI	iv
ABSTRACT.....	v
Kata Pengantar	vi
Daftar Isi.....	viii
Daftar Gambar.....	x
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Batasan Masalah.....	3
1.4 Tujuan Penelitian.....	4
1.5 Kegunaan Penelitian.....	4
1.6 Sistematika Penulisan.....	5
BAB II Landasan Teori.....	6
2.1 Data dan Informasi	6
2.1.1 Data	6
2.1.2 Informasi	6
2.2 <i>Bussiness Intelligence</i>	7
2.3 Data Collection.....	7
2.4 Web Based Application.....	8
2.5 Web Server	9
2.6 PHP.....	9
2.6.1 Kelebihan PHP	10
2.7 XAMPP	10
2.7.1 Kelebihan XAMPP	11
2.8 <i>Database</i>	12
2.9 Pengertian MySQL.....	12
2.9.1 Kelebihan MySQL	13
2.10 Pentaho Data Intergration.....	13

2.10.1	Kelebihan <i>Pentaho Data Integration</i>	14
2.11	Microsoft Excel	15
2.12	Bootstrap	15
BAB III Metodologi Penelitian		16
3.1	Profil Pendidikan	16
3.2	System Development Life Cycle (SDLC)	17
3.3	Ukuran kesuksesan website	24
Bab IV Analisis dan Pembahasan		26
4.1	Fase <i>Requirements Planning</i>	26
4.2	Fase <i>User Design</i>	27
4.2.1	Halaman Index	28
4.2.2	Halaman Search_Page	29
4.3	Fase <i>Construction</i>	31
4.3.1	Tools	31
4.3.2	Flow Data	33
4.3.3	<i>Work Flow</i> Pengunggahan Data	46
4.3.4	Rasio	47
4.4	Fase Cutover	51
4.5	Demo Portal	53
Bab V Kesimpulan dan Saran		56
5.1	Kesimpulan	56
5.2	Saran	57
Daftar Pustaka		58
LAMPIRAN		60

Daftar Gambar

Gambar 1.1	Gross Enrolment Ratio ASEAN.....	1
Gambar 3.1	Siklus Metode Rapid Application Development.....	17
Gambar 3.2	Diagram Konteks.....	19
Gambar 3.3	diagram level 1 website.....	21
Gambar 3.4	Struktur tabel pada database dengan ERD	23
Gambar 4.1	Halaman Index	28
Gambar 4.2	Halaman Search_Page pada website	29
Gambar 4.3	fitur pada tabel pada search_page	30
Gambar 4.4	Proses pengolahan data	32
Gambar 4.5	Skema proses pengambilan data	33
Gambar 4.6	Tampilan pada website http://ban-pt.kemdiknas.go.id/	34
Gambar 4.7	Source code dari hasil pencarian.....	35
Gambar 4.8	Tampilan pada Filter Row	36
Gambar 4.9	Hasil dari Filter Row	37
Gambar 4.10	Proses dalam penghilangan noise data	37
Gambar 4.11	kolom yang diambil dari http://ban-pt.kemdiknas.go.id	38
Gambar 4.12	sample data yang menjadi duplikat	39
Gambar 4.13	Data yang telah dicleansing dan di sinkronisasi.....	40
Gambar 4.14	Stage 1	41
Gambar 4.15	Field pada select value, stage 1	42
Gambar 4.16	Join Rows	43
Gambar 4.17	Hasil dari Stage 1	43
Gambar 4.18	Stage 2	44
Gambar 4.19	select value pada stage 2	44
Gambar 4.20	Join Rows Stage 2	45
Gambar 4.21	Hasil pada database	46
Gambar 4.22	Work flow pengunggahan data	46
Gambar 4.23	Gross Enrolment Ratio	48
Gambar 4.24	Hasil perhitungan rasio GER perprovinsi di Indonesia.....	50
Gambar 4.25	Sampling pengujian validasi data perguruan tinggi	51
Gambar 4.26	Sampling pengujian validasi data penduduk.....	52
Gambar 4.27	Halam Index portal.....	53
Gambar 4.28	Slide pada pencarian.....	54
Gambar 4.29	halaman pencarian.....	54