Research Fair Unisri 2019 **P- ISSN: 2550-0171**
Vol 3, Number 1, Januari 2019 **E- ISSN: 2580-5819**

PROSES ETL (*EXTRACT TRANSFORMATION LOADING*) *DATA WAREHOUSE* UNTUK PENINGKATAN KINERJA BIODATA DALAM MENYAJIKAN PROFIL MAHASIWA DARI DIMENSI ASAL SEKOLAH

Studi Kasus: Biodata Mahasiswa UKDW

Yetli Oslan[1] **dan Harianto Kristanto**[2]

[1,2]*Prodi Sistem Informasi, Fakultas Teknologi Informasi - Universitas Kristen Duta Wacana*

[1]yetli@staff.ukdw.ac.id, [2]harianto@staff.ukdw.ac.id

ABSTRAK

Biodata mahasiswa merupakan salah satu data utama dalam basis data akademik di sebuah perguruan tinggi. Biodata mahasiswa diisi sejak yang bersangkutan menjalani proses mendaftar sebagai calon mahasiswa, lalu dilengkapi lagi saat registrasi sebagai mahasiswa baru, dan terus diperbaharui lagi sampai dengan mahasiswa tersebut lulus dan berubah status menjadi alumni. Salah satu atribut dalam biodata mahasiswa adalah atribut yang menyimpan informasi tentang asal SMA dari seorang mahasiswa. Informasi tentang asal SMA menjadi salah satu dimensi yang banyak digunakan program studi dalam melakukan evaluasi diri. Persoalan yang terjadi, data yang diisikan pada atribut ini banyak yang 'kotor', bias, dan tidak valid sehingga tidak mampu memenuhi kebutuhan pandangan pemakai atas profil mahasiswa.

Metode penelitian yang dipakai adalah mengumpulkan dan mendokumentasikan biodata mahasiswa. Langkah selanjutnya adalah menyusun berbagai kemungkinan pandangan pemakai atas profil mahasiswa, kemudian menyusun skema multidimensional database dan melakukan ETL biodata mahasiswa, diakhiri dengan mengukur kinerja dari biodata dalam memenuhi kebutuhan pandangan pemakai atas profil mahasiswa.

Penelitian ini merekomendasikan agar dilakukan perubahan antar muka pada sistem pencatatan biodata, dengan mengganti *text box* menjadi *combo box* yang terhubung dengan tabel Ref_Sekolah.dbf. Perlu ditambahkan tabel untuk menyimpan kategori sekolah berdasarkan pengelompokkan yang diinginkan, sehingga memperluas dimensi analisis atas profil mahasiswa.

Kata kunci: ***ETL*, kinerja biodata mahasiswa, dimensi asal sekolah**

PENDAHULUAN

ETL merupakan singkatan dari *extract, transform, load* secara sederhana didefinisikan sebagai set proses untuk mendapatkan data dari OLTP *(on-line transaction processsing)* masuk ke *Data Warehouse*. *Extract* adalah kegiatan membuat resume. Sebagai contoh, meresume berapa banyak mahasiswa yang berasal dari daerah Yogyakarta, berapa banyak mahasiswa yang berasal dari luar Yogyakarta. *Transform* adalah kegiatan melakukan transformasi, misal asal sekolah yang berisi DIY, ditransformasikan menjadi Pulau Jawa, dalam hal ini data asli biodata mahasiswa tidak mengandung atribut tentang pulau. Field pulau ditransformasikan dari data referensi yang menyatakan bahwa DIY ada di pulau Jawa.

Proses *extract* dan *transform* data inilah yang menjadi inti pembahasan, yaitu bagaimana keduanya dapat dilakukan dengan benar dan memenuhi keinginan pandangan para pemakai sistem. Sekaligus juga akan diukur tingkat validitas biodata yang telah ada pada perguruan tinggi dapat ditransformasikan dengan sempurna. Tentunya ketidak sempurnaan dari biodata akan diperbaiki dengan usulan teknik pembetulan data

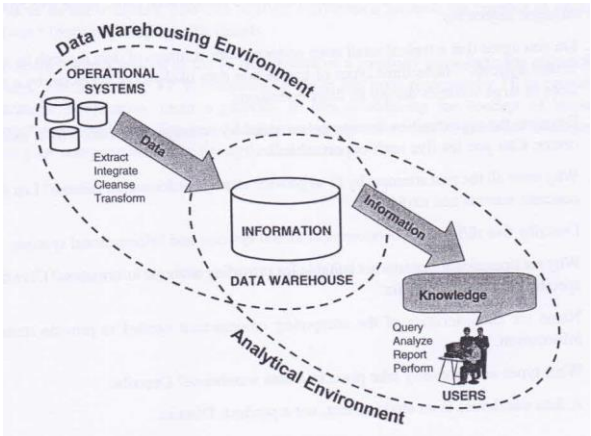
OLTP nya yaitu perbaikan teknis penangkapan dan pemasukkan data.

Pada akhirnya juga akan diberikan beberapa cara untuk meningkatkan validitas data melalui proses transformasi biodata sehingga memenuhi keinginan pandangan pemakai untuk kebutuhan analisis biodata berdasarkan asal sekolah mahasiswa. Hasil ini dapat menjadi inspirasi bagi para analis dan programmer dalam memperbaiki kesalahan-kesalahan yang mungkin terkandung pada kumpulan data yang akan dianalisis.

Dalam penelitian ini digunakan data mahasiswa UKDW sebanyak 5 angkatan yaitu dari tahun 2011-2015. Data referensi asal sekolah diambil dari Dinas Pendidikan di Indonesia dan beberapa website yang dapat dipertanggungjawabkan kebenarannya.

1. KAJIAN LITERATUR DAN PENGEMBANGAN HIPOTESIS

Konsep pengolahan data pada penelitian ini dapat dilihat pada gambar 1, yaitu adanya sistem On Line Transaction Processsing (OLTP) yaitu saat data diinputkan dalam database, dan setelah proses ETL dilakukan, maka Data Warehouse terbentuk. Proses lanjutan dari Data Warehouse adalah proses penggunaan data tersebut yang disebut sebagai sistem On Line Analytical Processsing (OLAP) [Ponniah,2010]

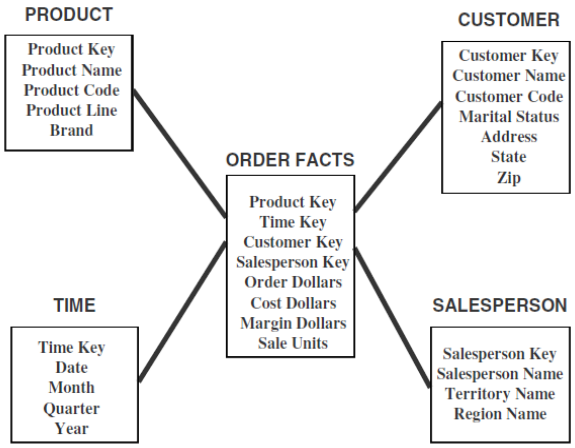


Gambar 1: Lingkungan OLTP – Data Warehouse - OLAP
(Sumber: buku Paulry Ponniah gambar 1-10)

Pembahasan mengenai rancangan data warehouse dibahas paling banyak pada buku Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals. Paulraj Ponniah bab 11 [Ponniah,2010], rancangan model star dan snowflake akan banyak dipakai untuk menampilkan lebih ringkas permasalahan yang diteliti dengan gambar-gambar.

Rancangan model star akan dipakai nantinya untuk menunjukkan bagaimana hubungan antara tabel Biodata Mahasiswa dengan tabel Dimensi Propinsi, Kabupaten, Kecamatan asal sekolah.

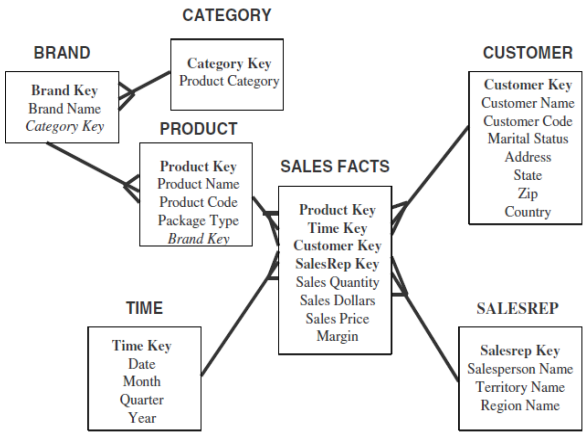
Demikian pula model Snowflake dipakai untuk menunjukkan bagaimana hubungan antara tabel Biodata Mahasiswa dengan tabel-tabel Demografi mahasiswa yang dibagi dalam kategori kategori tertentu (kategori SMA berasaskan keagamaan)



dan sejenisnya.

Gambar 2: Dimentional Modelling – Star Scheme untuk data contoh Order
(Sumber: buku Paulry Ponniah gambar 11-1 hal 227)

Sebuah awal skema adalah paradigma pemodelan di mana data warehouse berisi data tunggal yang besar sebagai pusat Tabel Fakta dan set Tebel Dimensi yang kecil-kecil. Tabel Fakta berisi data ringkasan rinci dan punya kunci utama yang memiliki satu kunci penghubung ke setiap dimensi. Setiap tuple di tabel fakta terdiri dari fakta atau topik yang menarik, dan dimensi yang menunjukkan fakta. Tabel dimensi terdiri dari kolom yang sesuai dengan atribut dimensi. Star Schema biasanya dirancang untuk satu data mart [Manjunath,2011]



Gambar 3: Dimentional Modelling – Snowflake Scheme untuk data contoh Produk
(Sumber: buku Paulry Ponniah gambar 11-8 hal 236)

ETL adalah fungsi integrasi data yang melibatkan penggalian data dari sumber luar (sistem operasional), mengubahnya sesuai dengan kebutuhan bisnis, dan akhirnya loading ke sebuah gudang data. Untuk mengatasi masalah tersebut, perusahaan menggunakan ekstrak, transform dan load (ETL) teknologi, yang meliputi membaca data dari sumbernya, membersihkannya dan format itu seragam, dan kemudian menulis ke repositori target untuk dieksploitasi. Data yang digunakan dalam proses ETL bisa datang dari sumber manapun: mainframe aplikasi, aplikasi ERP, alat CRM, flat file atau Excel spreadsheet. [Vishal,2010; da Silva 2012]. Ada banyak masalah untuk menerapkan proses ETL yang efisien dan dapat diandalkan yaitu: [Vishal,2010]

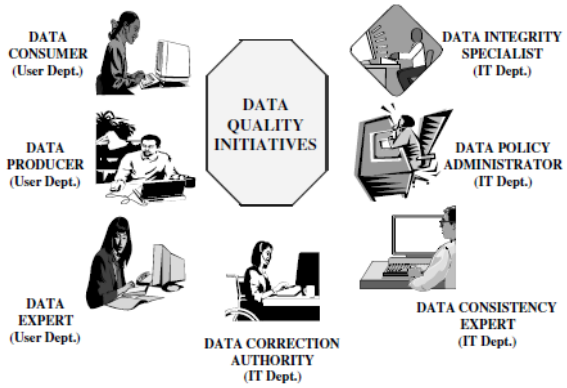
- tantangan teknis memindah, mengintegrasikan, dan mengubah data dari lingkungan yang berbeda
- Tidak konsisten, sulit untuk mempertahankan aturan bisnis
- Kurangnya paparan aturan bisnis ke pengguna akhir
- Sumber data tertentu yang penting ternyata hilang atau sumber data tidak bersih
- Kinerja query yang jelek

Transformasi data merupakan kegiatan mengkonversi data legacy atau host format (sumber asal) ke format Warehouse dengan menggunakan langkah-langkah seperti: [Esmail,2014]

- Memilih hanya kolom tertentu untuk dimuat/diambil;
- Terjemahan kode nilai;
- Encoding bebas-bentuk nilai-nilai;
- Menurunkan nilai yang dihitung baru (menghitung data berasal /derived;
- Pemilahan atau pengurutan;
- Menggabungkan data (joining data) dari berbagai sumber;
- Menghasilkan nilai pengganti-kunci;
- Memisahkan kolom menjadi beberapa kolom;
- Agregasi;
- Menerapkan bentuk sederhana, validasi data yang kompleks

Fungsi ETL membentuk kembali data yang relevan dari sistem sumber menjadi informasi yang berguna untuk disimpan di gudang data. Tanpa fungsi-fungsi ini, tidak akan ada informasi strategis dalam gudang data. Jika sumber data yang diambil dari berbagai sumber tidak dibersihkan, diekstraksi dengan benar, diubah dan diintegrasi dengan cara yang tepat, proses query yang merupakan tulang punggung dari data warehouse tidak bisa terjadi. [Vishal 2010].
Konsep data quality diambil dari buku dari *Data Warehousing Fundamentals: For IT Professionals*, Paulray Ponniah, edisi kedua, John Wiley &Sons, Inc Publication, tahun 2010 yang memperlihatkan proses Extraction Transformation Loading (ETL) dan Data Quality.

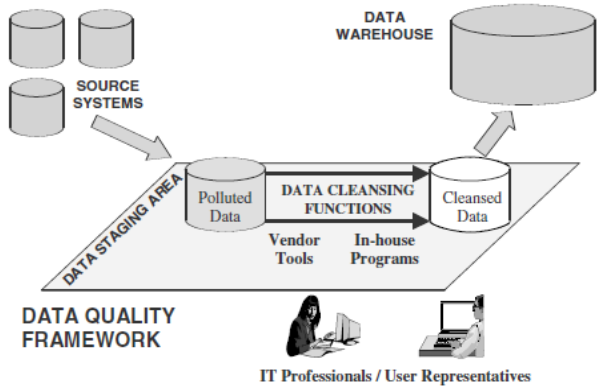
Proses pemasukkan data yang sudah terlanjur ada dalam database masih dapat dibersihkan (*cleansing*) dengan konsep yang dibahas dalam buku ini. Secara garis besar kualitas data dapat dilakukan atas inisiatif dari beberapa pemeran/ *role*. Pada penelitian ini difokuskan pada pemeran *Data Correction Authority* yang dilakukan oleh bagian Teknologi Informasi. Dapat dilihat pada gambar 4 dibawah ini.



Gambar 4: Data quality: participant and roles
(Sumber: buku Paulry Ponniah gambar 13-6)

Tugas bagian *Data Correction Authority* adalah menerapkan satu perangkat program-program yang dibangun khusus *(tools)* untuk pembersihan data ini. Pada penelitian ini akan dibuat khusus satu program untuk mengatasi masalah pembersihan data SMA Biodata Mahasiswa.
Proses ini dapat juga sebagai proses pemurnian data (purification), yang secara gambaran umum dapat dilihat pada gambar 5

Gambar 5: Overall data purification
(Sumber: buku Paulry Ponniah gambar 13-7)

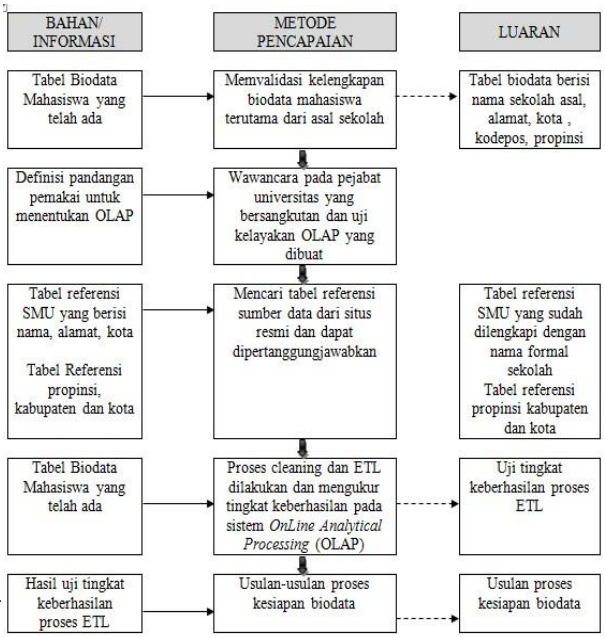


Proses ETL memerlukan kualitas data yang bersih *(clean)*, hal-hal yang membuat data tidak bersih diantaranya adalah: [Sweety,2012]

- Karena kesalahan manusia
- Kesalahan pada keterbatasan aplikasi
- Keterbatasan pada penerimaan data yang dimasukkan
- Data tidak kotor tetapi tidak memiliki format data teknis yang tepat sesuai kebutuhan aturan aplikasi
- Data duplikasi dalam sumber data pada saat integrasi dari sistem yang heterogen
- Data yang dimasukkan benar, tetapi tidak diperbarui sesuai jadwal waktu sehingga menjadi salah
- Kurangnya validasi data sebagai waktu data yang masuk dalam aplikasi tersebut dengan aturan software
- Verifikasi terbatas yang dilakukan ketika data dilewatkan melalui beberapa filter integrasi dari sistem heterogen
- Aturan pemetaan data logis yang tidak benarKurangnya pelatihan untuk orang yang memasukkan data ke dalam database sesuai aturan regulasi untuk melengkapi persyaratan masukan.

BAHAN DAN METODE

Penelitian ini dilakukan dengan tahapan seperti terlihat dalam bagan alir dibawah ini:



P- ISSN: 2550-0171
E- ISSN: 2580-5819

Gambar 6: Bagan alir pentahapan dan capaian kegiatan penelitian

Data yang digunakan pada penelitian ini berasal dari biodata mahasiswa UKDW angkatan tahun 2011-2015 yang dapat diperoleh dari PUSPINdIKA UKDW dan juga data referensi dari sumber Deparmenen Pendidikan tentang asal sekolah. Sedang metode penelitian yang dipakai adalah wawancara untuk penentuan pandangan pemakai dalam menentukan OLAP yang akan diciptakan, serta melakukan ETL secara berulang ulang dalam mengukur sampai sesuai dengan pandangan pemakai.

Data yang diperlukan dalam penelitian ini adalah **Biodata_Gab** sebagai **tabel fakta** dan **Ref_Sekolah** sebagai **tabel dimensi.** Tabel **Biodata_Gab.dbf** berisi biodata mahasiswa seperti terlihat pada Gambar 7 berikut ini:

Prodi	Record Awal	Angkatan 2011-2015
01	1232	254
11	3305	502
12	1096	340
21/61	1444	410
24/62	293	95
22/71	4271	780
23/72	929	345
31	948	261
41	578	489
Gab		3476

Struktur Biodata.dbf

Fields | Indexes | Table

Name	Type	Width
nim	Character	8
nama	Character	30
kelamin	Character	6
agama	Character	8
slta	Character	25
jurusan	Character	10
lainlain	Character	40
no_ijazah	Character	15
alam_slta	Character	90
kota_slta	Character	30
kd_kotasek	Character	5
prop_slta	Character	30
kd_propsek	Character	5

Gambar 7: Struktur Biodata_Gab.dbf

Tabel **Ref_sekolah** berisi informasi sekolah yang ada di Indonesia yang diolah dari situs http://psma.kemdikbud.go.id dengan sampel data seperti yang terlihat pada Gambar 8 berikut ini:

Gambar 8: Sampel Data Ref_Sekolah.dbf

DAFTAR SEKOLAH MENENGAH ATAS (SMA)
BERDASARKAN DATA TAHUN 2015/2016

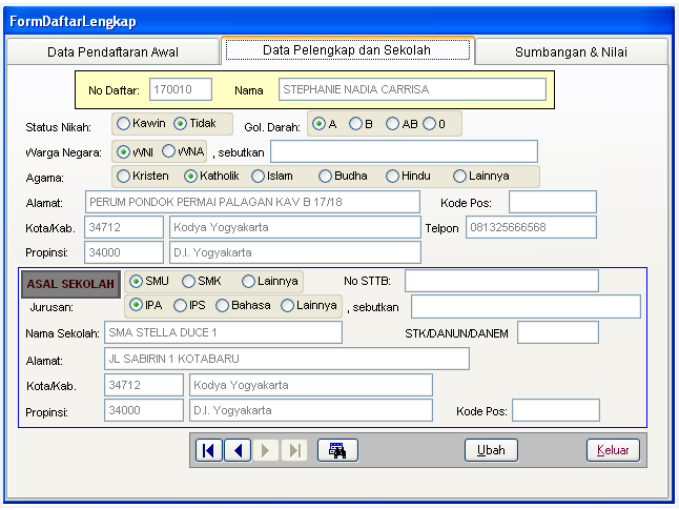
Prop. D.K.J. Jakarta
Kab. Kepulauan Seribu

No.	NPSN	Nama Sekolah	Status	Alamat	Kode Pos	Telp	Fax
1	20107185	SMAN 69 JAKARTA	NEGERI	PULAU PRAMUKA RT 003/05	14530	2170611769	

Prop. D.K.J. Jakarta
Kota Jakarta Pusat

No.	NPSN	Nama Sekolah	Status	Alamat	Kode Pos	Telp	Fax
1	20100194	SMA MAHATMA GANDHI SCHOOL	SWASTA	JL. TABING BLOK B -16 NO. 3	10720	216542241	
2	20100216	SMAN 1 JAKARTA	NEGERI	JL. BUDI UTOMO NO. 7	10710	3865001	
3	20100217	SMAN 10 JAKARTA	NEGERI	JL. MANGGA BESAR XIII	10730	216590192	
4	20100218	SMAN 20 JAKARTA	NEGERI	JL. KREKOT BUNDER III/I	10710	3440021	
5	20100219	SMAN 24 JAKARTA	NEGERI	JL. LAPANGAN TEMBAK NO. 1 SENAYAN	10270	215736984	
6	20100221	SMAN 25 JAKARTA	NEGERI	JL. AM. SANGAJI NO. 22-24	10130	6331921	
7	20100223	SMAN 27 JAKARTA	NEGERI	JL. MARDANI RAYA JAKARTA	10560	4245969	

Proses cleaning dalam penelitian ini difokuskan pada pembersihan nama sekolah yang sangat terbuka untuk terjadinya kesalahan saat pemasukkan data. Hal ini ini disebabkan karena antar muka yang digunakan saat input berupa *text box*, seperti terlihat pada Gambar 10 berikut:



Gambar 10: Antar Muka Pengisian Biodata Text Box

Proses *cleaning* atas 3476 data dilakukan dengan langkah-langkah sebagai berikut:

1. Membuat kumpulan unik nama sekolah:

   USE biodata_gab
   index on slta to sekolah_unik UNIQUE

   Hasil: 1857 Record

2. Memperbaiki data yang diduga sama *(manual cleaning)*, dengan memperhatikan dari tulisan nama sekolah dan alamat sekolah

3. Memperbaiki data yang diduga sama dengan menggunakan tabel Ref_Sekolah

Dalam proses *cleaning* ditemukan beberapa kendala yang dapat diselesaikan secara tuntas, melalui langkah-langkah di atas. Namun, ada juga yang tidak dapat diselesaikan karena harus mencari dokumen manual atau bertanya kepada yang bersangkutan. Kendala pertama adalah adanya calon mahasiswa dengan data sekolah yang berasal dari luar negeri. Hal ini dapat diselesaikan atau diabaikan karena jumlah data tidak banyak. Kendala kedua adalah terdapat data sekolah yang tidak terdaftar pada data referensi SMA 2015/2016. Kendala ketiga adalah terdapat data sekolah yang tidak lengkap sehingga sulit untuk diperbaiki. Kendala keempat adalah terdapat data kelengkapan yang tidak sesuai (semisal kode pos yang berbeda/ kabupaten).

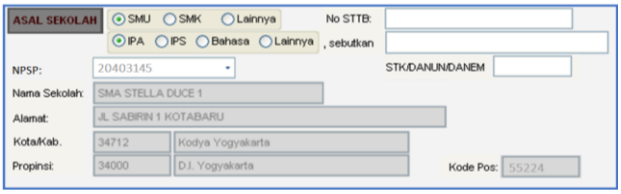
Tabel 1: Hasil Uji Tingkat Keberhasilan Proses ETL

No.	Kendala	Jumlah Kasus	Kasus Selesai	Keberhasilan ETL
1.	Data sekolah yang berasal dari luar negeri	20	20	100%
2.	Data sekolah tidak terdaftar dalam data referensi 2015/2016	10	0	0%
3.	Data sekolah tidak lengkap	75	15	20%
4.	Data kelengkapan tidak sesuai	50	25	50%

HASIL DAN PEMBAHASAN

A. Usulan Proses Kesiapan Biodata

Meskipun proses ETL dapat dikatakan berhasil, namun akan lebih baik jika dilakukan proses perbaikan mulai dari proses pemasukan data. Sumber kesalahan data ditemukan pada antar muka yang digunakan dalam



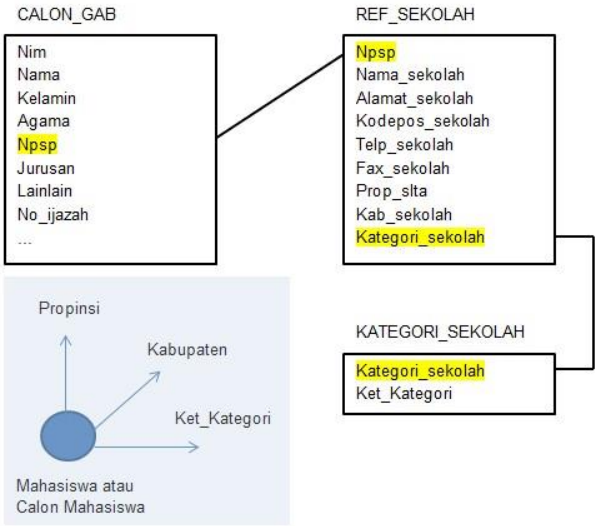
memasukkan data berupa *text box*. Oleh karena itu diusulkan untuk dilakukan perubahan menjadi combo box seperti Gambar 11 berikut ini:

Gambar 11: Antar Muka Pengisian Biodata Combo Box

B. Skema Database untuk Menyajikan Profil Mahasiswa

Dari hasil wawancara dengan para pejabat, diketahui bahwa analisis profil mahasiswa dari dimensi asal sekolah akan sangat bermanfaat untuk pengambilan kebijakan, strategi promosi, maupun pengembangan kurikulum. Mengenal dengan baik profil mahasiswanya, akan membuat proses belajar mengajar menjadi lebih baik. Saat ini kita cuma menyadari profil global mahasiswa, misalnya mahasiswa yang masuk prodi kedokteran lebih pintar dari prodi lainnya. Asumsi ini dibangun karena tingkat seleksi yang ketat, sehingga memang akhirnya yang terpilih adalah calon-calon yang terbaik. Namun, kita tidak dapat juga mengecilkan keberadaan mahasiswa yang pintar yang selalu ada di setiap prodi.

Oleh karena itu, dalam penelitian coba dikembangkan satu skema database berbentuk snowflake, yang dianggap mampu memberi potret lengkap mahasiswa berdasarkan asal sekolahnya saat SMA. Skema tersebut dapat dilihat pada Gambar 11, yang terdiri dari CALON_GAB sebagai tabel fakta, REF_SEKOLAH sebagai tabel dimensi yang kemudian dihubungkan juga dengan tabel KATEGORI_SEKOLAH. Tabel KATEGORI_SEKOLAH direncanakan berisi tentang informasi kategori sekolah, misalnya dapat diisi dengan data seperti 01= sekolah nasionalis, 02= sekolah negeri, 03 = sekolah nasional plus, 04 = sekolah inklusif, dll.



Gambar 11: Skema Database untuk Profil Mahasiswa dari Dimensi Asal Sekolah

KESIMPULAN

Dari hasil penelitian yang telah dilakukan atas kondisi biodata mahasiswa, dapat ditarik beberapa kesimpulan sebagai berikut:

1. Perlu dilakukan perubahan antar muka pada sistem pencatatan biodata, dengan mengganti *text box* menjadi *combo box* yang terhubung dengan tabel Ref_Sekolah.dbf.
2. Perlu ditambahkan tabel untuk menyimpan kategori sekolah berdasarkan pengelompokkan yang diinginkan, sehingga memperluas dimensi analisis atas profil mahasiswa.
3. Validasi data menjadi hal yang sangat perlu dilakukan sejak awal pemasukan data, karena itu tidak cukup menggantungkan pada proses ETL saja, tapi diperlukan perbaikan pada antar muka sebagai 'pintu masuk' data ke tabel.

UCAPAN TERIMA KASIH

Ucapan terima kasih disampaikan kepada semua pihak yang telah memberikan dukungan dalam proses penelitian ini. Secara khusus, kepada Prodi Sistem Informasi UKDW yang memberikan dukungan data pada penelitian ini, dan LPPM UKDW sebagai koordinator kegiatan penelitian di UKDW.

DAFTAR PUSTAKA/RUJUKAN

da Silva M.S. et al; 2012; "A Framework for ETL Systems Development";Journal of Information and Data Management, Vol. 3, No. 3, October 2012, pp 300–315

Esmail Ali, F.S.; (2014); "A Survey of Real-Time Data Warehouse and ETL"; International Scientific Journal of Management Information Systems, 9 (3)

Manjunath T N et al; 2011; "Design and Anlysis of DWH and BI in Education Domain"; IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, pp 545-551

Ponniah, Paulraj; 2010; "Data Warehousing Fundamentals for IT Professional", 2nd ed, John Wiley & Sons Inc

Sweety Patel; 2012; "Requirement to cleanse DATA in ETL process and Why is data cleansing in Business Application?"; International Journal of Engineering Research and Applications (IJERA),Vol. 2, Issue 3, May-Jun 2012, pp.840-842

Vishal Gour et. al. ;2010; "Improve Performance of Extract, Transform and Load (ETL) in Data Warehouse"; (IJCSE) International Journal on Computer Science and Engineering; Vol. 02, No. 03, pp 786-789