Seminar Nasional Informatika 2010 (semnasIF 2010) *ISSN: 1979-2328*
UPN "Veteran" Yogyakarta, 22 Mei 2010

ANALISIS FAKTOR-FAKTOR YANG MEMPENGARUHI PROSES ETL PADA DATA WAREHOUSE

Armadyah Amborowati
STMIK AMIKOM Yogyakarta
Jl. Ring Road Utara, Condong Catur, Yogyakarta Telp (0274) 884201
e-mail : armagauthama@yahoo.com

Abstrak
ETL (Extrac Transform Loading) pada proses develop data warehouse merupakan suatu proses yang memakan waktu paling lama. Kesuksesan proses ETL sangat dipengaruhi oleh kualitas data yang ada pada database OLTP. Penelitian ini bertujuan untuk mencari noise-noise yang mungkin timbul pada proses ETL dengan metode pengembangan data warehouse.
Database OTLP yang digunakan untuk penelitian adalah database Perpustakaan STMIK AMIKOM Yogyakarta dan data warehouse yang dibangun berdasarkan tabel fakta transaksi perpustakaan. Dari hasil pengujian yang didapat adalah kegagalan pada proses ETL dari database OLTP ke database data warehouse adalah adanya noise. Setelah dianalisis ternyata noise ada pada tabel pinjam_mhs, yaitu adanya data yang bernilai null pada kolom kd_pinjam_mhs di tabel pinjam_mhs. Sehingga sebelum proses ETL dilakukan perlu adanya proses menghilangkan noise yang ada pada database sumber atau database OLTP.

***Kata Kunci**: OLTP, Extrac, Transform, Loading , Data warehouse*

1. PENDAHULUAN

Data warehouse adalah suatu konsep dan kombinasi teknologi yang memfasilitasi organisasi untuk mengelola dan memelihara data historis yang diperoleh dari sistem atau aplikasi operasional [Ferdiana, 2008]. Dengan *data warehouse* proses untuk pembuatan laporan tidak akan mengganggu sistem OTLP.

Kesuksesan dalam membangun data warehouse sangat dipengaruhi oleh kesuksesan dalam proses ETL dari database OLTP ke database *data warehouse*. ETL (*Extrac Transform Loading*) merupakan suatu proses yang memakan waktu paling lama. Kesuksesan proses ETL sangat dipengaruhi oleh kualitas data yang ada pada database OLTP.

2. TINJAUAN PUSTAKA

Beberapa penelitian terdahulu menunjukkan bahwa *data warehouse* dan *data mining* dibangun untuk dapat memberikan dukungan terhadap pengambil keputusan. Salah satu penelitian mengenai *data warehouse* dilakukan di CV. Andi Offset dimana penelitian tersebut bertujuan untuk mengetahui bagaimanakah rancangan database, pola penjualan buku, serta bentuk representasi data yang dapat membantu pengambilan keputusan dalam melakukan analisis data penjualan yang tersimpan dalam sebuah *data warehouse*. Hasil dari penelitian ini menunjukkan bahwa pola penjualan sebuah produk di dalam CV. Andi Offset mempunyai rentang life cycle selama 2 tahun. Dan dengan representasi data dari aplikasi OLAP, menjadikan informasi-informasi bisnis mudah diinterpretasikan (Srimulyanta, 2006).

Penelitian lainnya yaitu perancangan dan pembuatan aplikasi *cluster analysis* terhadap data transakasi buku di Perpustakaan UK Petra oleh Angelina Sartika Kurniawati (2006). Penelitian ini bertujuan untuk mengetahui hubungan antara pola peminjaman dengan IPK mahasiswa menggunakan metode *cluster analysis.*

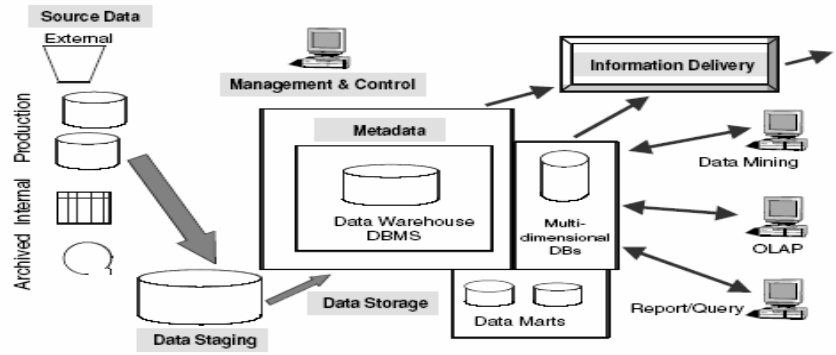
Data warehouse adalah suatu paradigma baru dilingkungan pengambilan keputusan strategik. *Data warehouse* bukan suatu produk tetapi suatu lingkungan dimana user dapat menemukan informasi strategik [Poniah, 2001, h.14]. *Data warehouse* adalah kumpulan data-data logik yang terpisah dengan database operasional dan merupakan suatu ringkasan.

Data warehouse mengandung beberapa elemen penting antara lain [Mallach, 2000,h.473]:

1. Sumber data yang digunakan oleh *data warehouse*, database transaksional dan sumber data eksternal.
2. Proses ETL (*Extraction, Transformation, Loading*) dari sumber data ke database *data warehouse*.
3. Membuat suatu ringkasan atau *summary* terhadap *data warehouse* misalkan dengan menggunakan fungsi agregat.
4. Metadata.
   Metadata mengacu data tentang data. Metadata menguraikan struktur dan beberapa arti tentang data, dengan demikian mendukung penggunaan efektif atau tidak efektif dari data.
5. Database *data warehouse*.
   Database ini berisi data yang detail dan ringkasan data dari data yang ada di dalam *data warehouse*. Karena *data warehouse* tidak digunakan dalam proses transaksi individu, maka databasenya tidak perlu

diorganisasikan untuk akses transaksi dan untuk pengambilan data, melainkan dioptimisasikan untuk pola akses yang berbeda di dalam analisis.

6. *Query Tools* yaitu dengan OLAP (*Online Analytical Processing*) dan data mining. Tool untuk *query* ini meliputi antarmuka pengguna akhir dalam mengajukan pertanyaan kepada database, dimana proses ini disebut sebagai *On-line Analytical Processing* (OLAP). Tool ini juga terdiri dari tool otomatis yang menemukan pola-pola di dalam data, yang sering disebut sebagai *data mining*. *Data warehouse* harus memiliki salah satu dari kedua tipe ini atau malah kedua-duanya.
7. User.
   Pengguna yang memanfaatkan *data warehouse* tersebut.



Gambar 1. Arsitektur *Data Warehouse*
Sumber: Poniah, 2001, h. 29.

Pada sistem OLTP (*Online Transactional Processing*) digunakan suatu teknik pemodelan data yang disebut sebagai E-R (*Entity-Relationaship*). Pada *data warehouse* digunakan teknik pemodelan data yang disebut *dimensional modelling technique.* Pemodelan dimensional adalah suatu model berbasis pemanggilan yang mendukung akses *query* volume tinggi. *Star Schema* adalah alat dimana pemodelan dimesional diterapkan dan berisi sebuah tabel fakta pusat. Tabel fakta berisi atribut deskriptif yang digunakan untuk proses *query* dan *foreign key* untuk menghubungkan ke tabel dimensi. Atribut analisis keputusan terdiri dari ukuran performa, metrik operasional, ukuran agregat, dan semua metrik yang lain yang diperlukan untuk menganalisis performa organisasi. Tabel fakta menunjukkan apa yang didukung oleh *data warehouse* untuk analisis keputusan. Tabel dimensi mengelilingi tabel fakta pusat. Tabel dimensi berisi atribut yang menguraikan data yang dimasukkan dalam tabel fakta. Tabel dimensi menunjuk bagaimana data akan dianalisis.

Menurut Rainardi, 2008, ETL adalah suatu proses mengambil dan mengirim data dari data sumber ke *data warehouse*. Dalam proses pengambilan data, data harus bersih agar didapat kualitas data yang baik. Contohnya ada nomor telepon yang invalid, ada kode buku yang tidak eksis lagi, ada beberapa data yang *null*, dan lain sebagainya. Pendekatan tradisional pada proses ETL mengambil data dari data sumber, meletakan pada *staging area*, dan kemudian mentransform dan meng-*load* ke data warehouse.

Kualitas data merupakan hal terpenting yang harus diperhatikan dalam membangun data warehouse karena kualitas data mempengaruhi proses ETL. Pada proses ETL jika pada data terjadi suati noice maka proses ETL adan gagal.

Kualitas data dapat dilihat dari beberapa parameter, yaitu [melisadata.com, 2010]:

1. Akurat (*accurate*)
   Ketika melihat record alamat konsumen, maka alamat harus mengandung kota, kode pos. Jika konsumen memiliki suatu bisnis maka alamat konsumen juga berisikan alamat atau lokasi dari bisnisnya.
2. Tepat waktu (*Up to date*)
   Selalu memberikan informasi terbaru jika terjadi proses perubahan.
3. Komplet (*complete*)
   Setiap data harus berisikan informasi penting, misalkan untuk proses surat-menyurat. Misalkan nama apartemen, no apartemen, jalan, kode pos, dan jika dibutuhkan denah alamatnya atau rutenya.
4. Tidak redudansi (*not redud*ant)
   Misalkan hanya ada satu record per contact untuk setiap alamat dalam surat menyurat.
5. Standar (*standardized*)
   Setiap record harus standar dalam pemberian nama, proses pembacaan, dan singkatan.

3. METODE PENELITIAN

3.1 Bahan Penelitian

Bahan penelitian yang digunakan dalam penelitian ini adalah sebagai berikut.

1. Data yang diperoleh di lapangan melalui pengamatan dan wawancara langsung di departemen IC (*Inovation Center*) dan Staf Perpustakaan STMIK AMIKOM Yogyakarta, yang terdiri dari:
    1) Hubungan antar tabel pada database OLTP STMIK AMIKOM Yogyakarta.
    2) Data-data pada setiap tabel.
2. Informasi mengenai *data warehouse* dan proses analisis melalui studi literatur.

3.2 Alat Penelitian

Alat penelitian yang digunakan pada penelitian ini adalah sebagai berikut.

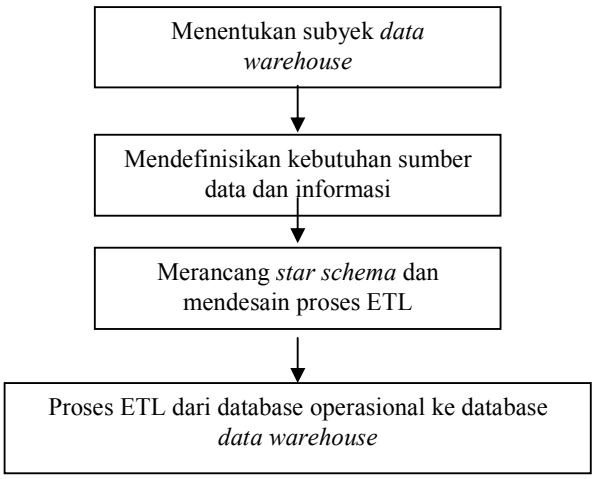
1. Hardware
   Komputer dengan spesifikasi sesuai untuk menjalankan software dalam penelitian.
2. Software
    1) DBDesigner 4
    2) DBMS SQL Server 2005

3.3 Jalan Penelitian

Adapun jalan penelitian yang dilakukan dalam penelitian ini adalah sebagai berikut.

1. Menentukan subyek *data warehouse*.
   Salah satu karakteristik dari *data warehouse* adalah *subject-oriented* sehingga langka awal dalam membuat *data warehouse* adalah menentukan subyeknya.
2. Mendefinisikan kebutuhan
   Mendefinisikan kebutuhan terhadap sumber data yang dibutuhkan oleh *data warehouse* dan informasi yang ingin didapat dari *data warehouse* untuk mendukung keputusan manajemen.
3. Membuat rancangan *star schema* dan mendesain proses ETL.
   Data warehouse menggunakan model data dimensional atau sering disebut sebagai *star schema*. *Star schema* mempunyai dua bagian, yaitu tabel fakta dan tabel dimensi. Setelah *star schema* dibuat maka proses selanjutnya adalah melakukan proses ETL (*Extract, Transformation, Loading*) dari database operasional ke database data warehouse. Sebelum proses ETL ini diproses harus dibuat dulu desain untuk proses ETL-nya.
4. Melakukan proses ETL dari database operasional ke database *data warehouse*.
   Proses ETL secara periodik mengekstrak data dari sistem sumber, mentranformasikannya kesebuah format yang umum, dan kemudian memuatnya kedalam *data store* target, yang umumnya sebuah *data warehouse* atau *data mart*. ETL sangat penting untuk integrasi data dan *data warehousing*.

Jalannya penelitian untuk lebih jelasnya dapat dilihat pada gambar 2. dibawah ini.

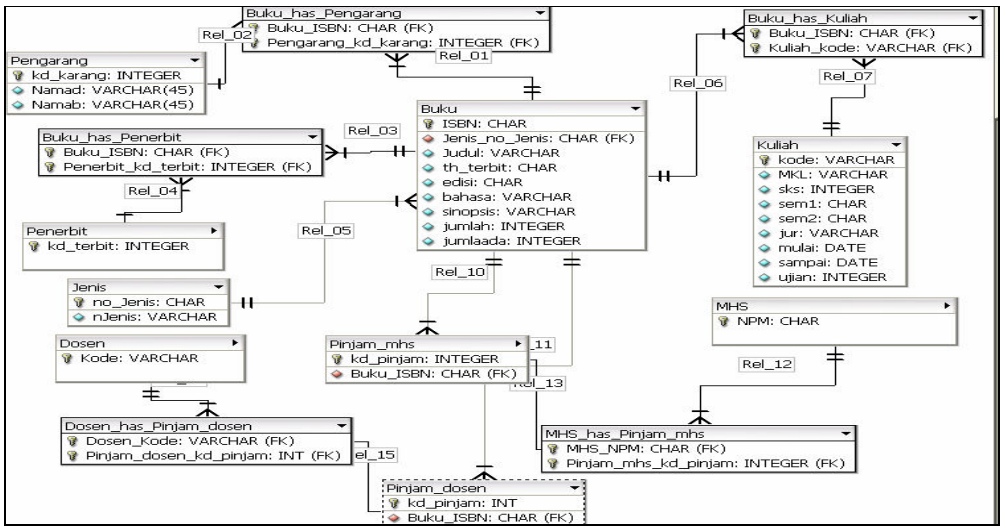


Gambar 2. Jalannya penelitian

4. HASIL DAN PEMBAHASAN

Berdasarkan jalannya penelitian maka dihasilkan:

- S ubyek yang diambil adalah transaksi/ sirkulasi Perpustakaan.
- Mendefinisikan kebutuhan sumber data dan informasi.
  - ❖ Sumber data yang digunakan dalam *data warehouse* adalah data-data dalam database perpustakaan.



Gambar 3. Sumber Data

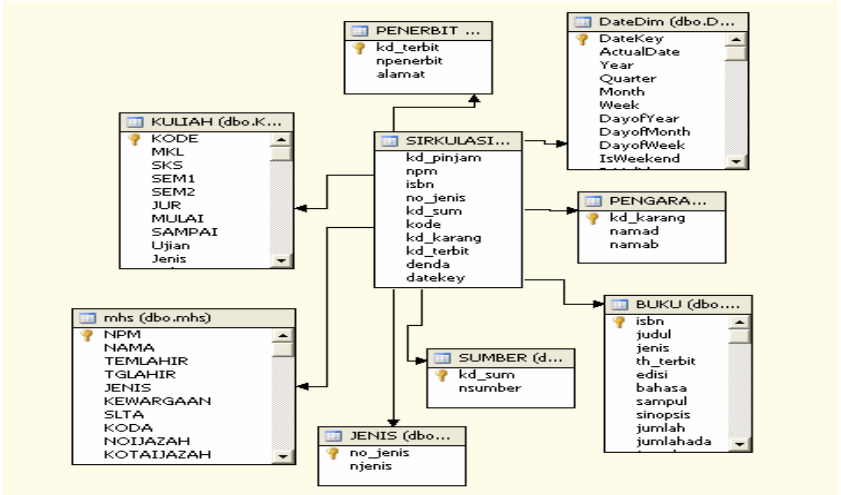
❖ Kebutuhan Informasi

Berdasarkan hasil pengamatan dan wawancara langsung kepada petugas atau manajemen perpustakaan didapat kebutuhan informasi sebagai berikut.

- ➢ Tren terhadap buku-buku yang dipinjam oleh dosen dan mahasiswa.
- ➢ Informasi mengenai jenis buku yang sering dipinjam.
- ➢ Tren proses sirkulasi.
- ➢ Informasi pengenai nama pengarang yang bukunya sering dipinjam.
- ➢ Informasi mengenai nama penerbit yang bukunya sering dipinjam

- Merancang *Star Schema*

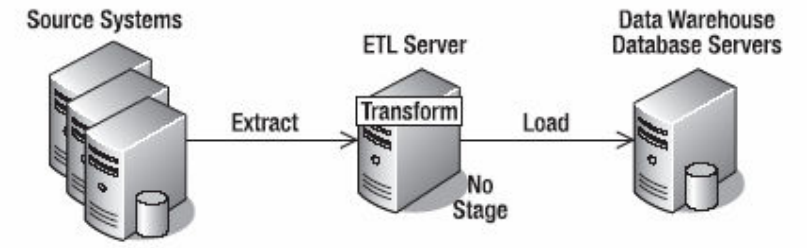
Berdasarkan sumber data dan kebutuhan informasi yang dibutuhkan oleh petugas dan manajemen perpustakaan maka model data dimensional yang dibuat dalam bentuk *star schema* untuk *data warehouse* perpustakaan bisa dilihat pada gambar 4.



Gambar 4. *star schema* untuk data warehouse

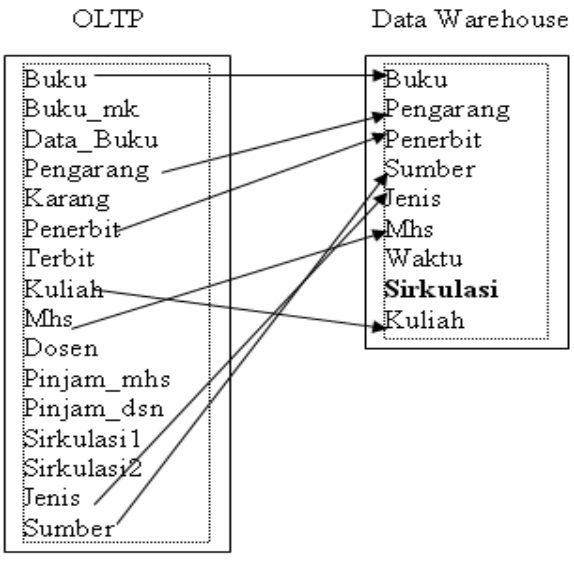
- Proses ETL dari database operasional ke database *data warehouse*

Proses selanjutnya setelah *star schema* dibuat adalah membuat desain proses ETL. Desain proses ETL yang dipakai bisa dilihat pada gambar 5 dibawah ini. Proses ETL mengambil data dari *source systems* menggunakan *query*. ETL berkoneksi dengan source system database dan mengambil data dengan query. Setelah data hasil *query* diambil langkah selanjutnya dilakukan eksekusi proses ETL dan mengirimnya ke database *data warehouse.*



Gambar 5. Desain proses ETL

Sebelum proses ETL dilakukan selanjutnya dilakukan proses mempopulasikan tabel dimensi dan tabel fakta .



Gambar 6. Mempopulasikan tabel dimensi dan tabel fakta

Pada saat proses ETL dari OLTP ke database data warehouse ada kegagalan proses ETL pada saat meng-*loading* tabel fakta sirkulasi pada database *data warehouse*. Tabel fakta sirkulasi memerlukan data pada tabel OLTP. Setelah dilakukan analisis, kegagalan disebabkan adanya *noise* yang ada ditabel OLTP. Pada gambar 7. dapat dilihat adanya *noise* pada hasil *query*.

```
select * from sirkulasi1 left outer join pinjam_mhs on
sirkulasi1.kd_pinjam_mhs=pinjam_mhs.kd_pinjam_mhs
```

Results | Messages

kd_pinjam_mhs	nim	tgl_pinjam	tgl_kembali	kd_pinjam_mhs	kd_buku
69	01.02.3754	2003-09-13 00:00:00.000	2003-09-13 00:00:00.000	NULL	NULL
86174	03.01.1560	2005-09-17 00:00:00.000	2005-09-17 00:00:00.000	NULL	NULL
199778	06.02.6279	2008-05-17 00:00:00.000	NULL	NULL	NULL
161	02.02.4423	2003-09-15 00:00:00.000	2003-09-15 00:00:00.000	NULL	NULL
78685	04.01.1812	2005-07-06 00:00:00.000	2005-07-06 00:00:00.000	NULL	NULL
151	01.02.3779	2003-09-15 00:00:00.000	2003-09-15 00:00:00.000	NULL	NULL
88277	04.22.0342	2005-09-28 00:00:00.000	2005-09-28 00:00:00.000	NULL	NULL
87073	05.11.0949	2005-09-21 00:00:00.000	2005-09-21 00:00:00.000	NULL	NULL
196809	06.12.2026	2008-04-22 00:00:00.000	NULL	NULL	NULL
196731	07.01.2187	2008-04-22 00:00:00.000	NULL	NULL	NULL
189938	07.12.2334	2008-03-01 00:00:00.000	NULL	NULL	NULL
42	01.02.3498	2003-09-12 00:00:00.000	2003-09-12 00:00:00.000	NULL	NULL
174545	06.02.6386	2007-10-23 00:00:00.000	NULL	NULL	NULL
72334	04.11.0603	2005-05-17 00:00:00.000	2005-05-17 00:00:00.000	NULL	NULL
184719	07.22.0766	2008-01-08 00:00:00.000	NULL	NULL	NULL
86538	04.22.0384	2005-09-19 00:00:00.000	2005-09-19 00:00:00.000	NULL	NULL
184772	05.01.2010	2008-01-09 00:00:00.000	NULL	NULL	NULL
78981	03.01.1617	2005-07-11 00:00:00.000	2005-07-11 00:00:00.000	NULL	NULL
60371	04.01.1773	2005-03-08 00:00:00.000	2005-03-08 00:00:00.000	NULL	NULL

Noise

Gambar 7. Hasil *query* yang menunjukkan adanya *noise*

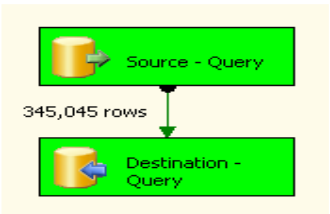
Noise yang ada harus dibersihkan terlebih dahulu agar proses *load*-nya berhasil. Caranya dengan menghapus suatu kolom relationship yang bernilai *null.*

kd_pinjam_mhs	nim	tgl_pinjam	tgl_kembali	kd_pinjam	kd_buku	denda
201904	05.12.1306	6/6/2008		201904	004.2/nUG/a.12	
201904	05.12.1306	6/6/2008		201904	658.403/tUR1/d.44	
201905	05.12.1379	6/6/2008		201905	005.756/mAT/s.4	
201905	05.12.1379	6/6/2008		201905	004.3/tAU/m.9	
201905	05.12.1379	6/6/2008		201905	005.74/sIM/p.5	
201906	06.01.2069	6/6/2008		201906	cd-1924/mIC/06.2	
201906	06.01.2069	6/6/2008		201906	005.369/bUD/a.54	
201906	06.01.2069	6/6/2008		201906	005.74/tUT/m.15	
201907	05.12.1083	6/6/2008		201907	658.403/tUR1/d.21	
50	01.01.1262	9/12/2003	9/12/2003			
198264	06.11.1128	5/6/2008				
49144	02.01.1369	12/13/2004	12/13/2004			
190128	06.11.1018	3/3/2008				
64960	03.11.0319	4/4/2005	4/4/2005			
165114	04.11.0517	6/14/2007				
72801	03.02.5079	5/20/2005	5/20/2005			
88468	02.12.0082	9/29/2005	9/29/2005			
178954	07.01.2182	11/23/2007				
59744	03.01.1672	3/3/2005	3/3/2005			
87668	02.12.0104	9/24/2005	9/24/2005			
182218	07.22.0755	12/17/2007				
122235	05.11.0798	6/14/2006	6/16/2006			
39987	03.11.0165	9/25/2004	9/25/2004			
196676	07.11.1738	4/21/2008				
173354	04.12.1015	9/27/2007				
240	01.02.3523	9/15/2003	9/15/2003			
39988	03.02.5088	9/25/2004	9/25/2004			
197619	06.12.1897	4/29/2008				
83692	04.01.1834	9/5/2005	9/5/2005			
740	01.02.3657	9/19/2003	9/19/2003			

Noise yang harus dihapus

Gambar 8. Noise yang dihapus

Setelah proses penghilangan *noise* selesai. Selanjutnya proses ETL dilakukan kembali. Berikut adalah hasil dari proses ETL.



Gambar 9. Hasil dari proses ETL untuk tabel fakta Sirkulasi

5. KESIMPULAN

Dari hasil pengujian yang didapat adalah kegagalan pada proses ETL dari database OLTP ke database *data warehouse* adalah adanya *noise*. Setelah dianalisis ternyata *noise* ada pada tabel pinjam_mhs, yaitu adanya data yang bernilai *null* pada kolom kd_pinjam_mhs di tabel pinjam_mhs. Sehingga sebelum proses ETL dilakukan perlu adanya proses menghilangkan *noise* yang ada pada database sumber atau database OLTP.

6. DAFTAR PUSTAKA

Han, Jiawei; Kamber, Micheline. 2006. *Data Mining: Concepts and Techniques* . San Fransisco: Morgan Kaufmann.

Hutabarat, Bernaridho I.,2008, Data Warehouse dengan SQL Server 2005. Elex Media Komputindo: Yogyakarta.

Kimball, Ralph; Caserta, Joe. 2004. *The Data Warehouse ETL Toolkit*. New Delhi: Wiley Publishing:.

Mallach, Efrem G.,. 2000. *Decision Support and Data Warehouse Systems*. Singapore: Irwin McGraw Hill.

Melisa Data, 2010, *Scalable Data Quality* [online], http://www.melissadata.com/dqt/whitepaper/scalable-data-quality-whitepaper.pdf tanggal akses 13 Februari 2010.

Ponniah, Paulraj. 2001. Data Warehouse Fundamentals: a Comprehensive Guide for IT Professional. New York : John Wiley & Sons.

Populate Time Dimension of AdventureWorksDW Sample Database and use it in yourDatawarehouse/cube.[Online] http://blogs.msdn.com/azazr/archive/2008/05/09/populate-time-dimension-of-adventureworksdw-sample-database-and-use-it-in-your-datawarehouse-cube.aspx. Tanggal akses 22 Juli 2008, Pukul 10.45.

Rainardi, Vincent, 2008, Building a Data Warehouse with Examples in SQL Server. Apress: New York.

Tang, ZhaoHui; MacLennan, Jamie. 2005. *Data Mining with SQL Server 2005*. Indiana Polis : Wiley Publishing.

Turban, 2005, Decision Support Systems and Intelligent Systems (Sistem Pendukung Keputusan dan Sistem Cerdas) Jilid 1. Andi Offset: Yogyakarta.