

Karakteristik Sumber Data untuk Data Warehouse

Djoni Darmawikarta

djoni_darmawikarta@yahoo.ca

Lisensi Dokumen:

Copyright © 2004 IlmuKomputer.Com

Seluruh dokumen di IlmuKomputer.Com dapat digunakan, dimodifikasi dan disebarkan secara bebas untuk tujuan bukan komersial (nonprofit), dengan syarat tidak menghapus atau merubah atribut penulis dan pernyataan copyright yang disertakan dalam setiap dokumen. Tidak diperbolehkan melakukan penulisan ulang, kecuali mendapatkan ijin terlebih dahulu dari IlmuKomputer.Com.

Data Warehouse (DW) menyimpan data yang berasal dari satu atau lebih sumber. DW tidak menciptakan data baru. Data yang disimpan didalam DW sering diolah sebelum disajikan untuk end-user, misalnya diringkas (summary) sebelum dicetak sebagai laporan.

Tulisan ini membahas 6 karakteristik penting dari sumber data yang perlu diperhatikan waktu merancang dan membangun DW.

ARAH

- Pemilik sumber data mengirim ke DW.
- Pengelola DW mengambil data dari sumber sesuai jadwal yang sudah disetujui bersama pemiliknya.

Pendekatan kedua lebih populer, karena pemilik data pada prinsipnya cukup memberi ijin dan menyetujui kapan pengelola DW boleh mengambil data yang diperlukan.

Pada pendekatan pertama, pemilik data terbebani untuk membuat program baru dan/atau menjalankan job untuk mengambil data yang diinginkan untuk disimpan di DW, dimana proyek DW mungkin tidak ada hubungan dengan aplikasi yang menjadi sumber data atau bukan prioritasnya.

Saya lebih menyukai pendekatan pertama, dengan dua alasan utama:

1. Pemilik data yang paling tahu seluk beluk datanya.
2. Menggalakkan spirit atau kultur “berbagi data” (information sharing) didalam perusahaan.

Dengan sponsor pimpinan yang efektif, pendekatan pertama pembangunan DW akan memberikan hasil yang lebih besar nilainya – nilai “corporate information sharing culture”.

ASAL

- Dari dalam; ini merupakan sumber data utama untuk DW, berasal dari biasanya aplikasi-aplikasi untuk operasi, transaksi, dan administrasi perusahaan.
- Dari luar; misalnya data yang diperoleh/dibeli/berlangganan dari biro statistik, seperti demografis konsumen; dari kode pos dari kantor pos, misalnya untuk segmentasi geografis pasar; standar ISO, seperti kode mata-uang dan kode negara.

FREKUENSI

- Satu kali (one time); diperlukan:
 1. Waktu awal/pertama kali mengisi DW, bila sejarah data sebelum tanggal mulai ini perlu/diinginkan disimpan di DW.
 2. Untuk recovery, dari backup, bila ada masalah yang mengharuskan DW dihapus seluruh/sebagian datanya dan dikembalikan ke status data sebelum terjadi masalah.
 3. Untuk menggabungkan data (merge), bila dua atau lebih DW dijadikan satu.
- Berkala (regular); sesuai jadwal yang dirancang, untuk menambahkan data baru dan perubahan (incremental update). Beberapa data (tabel database) mungkin lebih efektif/efisien diupdate seluruh datanya (total refresh) daripada incremental.

FORMAT

Beberapa format sumber data yang populer (paling sering ditemui) adalah:

- File (text, flat); biasanya hasil perubahan data yang nara sumbernya memiliki format bukan relational database. Flat file paling populer karena kesederhanaannya memudahkan pengambilan (extract) dan pengolahannya (transform) dengan kecepatan tinggi. Banyak piranti end-user, terutama OLAP (OnLine Analytical Processing) dan Data Mining, yang menganjurkan input datanya berupa flat file.
Oracle menyediakan utility SQL*Loader untuk mengambil flat file dan menyimpannya di tabel database. Sejak release 9, disediakan fungsi untuk “melihat” flat file sebagai tabel database (disebut *external table*)
- Relational database; terutama bila database system-nya sama dengan yang digunakan di DW, pengambilan, pengolahan dan penyimpanannya ke DW dapat menggunakan teknik dan program yang sama dengan proses ETL yang lain (Extract, Transform, Load).
- ODBC; fasilitas yang “membungkus” sumber data dan formatnya sehingga “dilihat” sebagai tabel database.

KESERAGAMAN

Bila data berasal dari berbagai sumber, dengan struktur dan nama-nama field yang berbeda, juga kadang mengandung makna berbeda, mereka perlu diseragamkan sebelum disimpan terintegrasi ke DW tidaklah sederhana. Selain kompleksitas teknis, proses penyeragaman membutuhkan persetujuan bersama seluruh komunitas end-user.

Misalnya kasus *nama*, apakah perlu dipisah menjadi beberapa bagian, seperti *nama panggilan*, *nama tengah*, dan *nama keluarga*? Apakah nama tengah boleh/cukup satu karakter (singkatan)? Contoh lain, *jumlah penjualan*. Apakah dinyatakan dalam satu mata uang? Ini nilai penjualan bersih, atau termasuk komisi dan diskon?

KEBERSIHAN

Menjaga kualitas sumber data merupakan proses dan tanggung jawab paling penting dan rumit perancangannya dalam membangun DW. Idealnya, semua data yang tersimpan di DW harus bersih, untuk menjamin kepercayaan bagi end-user, terutama bila data di DW menjadi dasar pengambilan keputusan para pimpinan.

Apa yang harus dilakukan bila misalnya nama seorang pelanggan yang data disumbernya disimpan dua kali dengan ejaan berbeda? Memiliki dua alamat rumah yang berbeda? Apakah keduanya benar, karena memang sudah berubah (*legitimate changes*)?

Mengambil sumber data dan menyimpannya kedalam DW, dengan benar, konsisten, pada waktunya, serta sesuai persetujuan bersama pemilik dan seluruh komunitas end-user, merupakan ukuran kualitas dan kehandalan DW yang maha penting.