

Data Mining

Menggali Informasi yang Terpendam

Yudho Giri Sucahyo
sucahyoy@cs.curtin.edu.au



Lisensi Dokumen:

Copyright © 2003 IlmuKomputer.Com

Sehuruh dokumen di **IlmuKomputer.Com** dapat digunakan, dimodifikasi dan disebarakan 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**.

Ketika saya mengikuti program orientasi mahasiswa baru pasca sarjana di Curtin University of Technology, saya berkenalan dengan seorang mahasiswi asal Australia. Dia mengambil program Master di bidang Jaringan Komputer dan telah menyandang gelar MCSE (*Microsoft Certified Systems Engineer*). Dia lalu bertanya pada saya, “*Apa topik penelitian Anda?*”, saya menjawab. “*Data Mining*”. Dia kemudian memberi komentar kepada saya, “*Oh.... itu bagus sekali.... Anda tepat sekali mengambil topik itu disini, karena kita punya pertambangan emas yang besar sekali di Kalgoorlie (Kalgoorlie berada 600 km di sebelah timur Perth dan Curtin University mempunyai cabang kampus disana)*”.

Data Mining memang salah satu cabang ilmu komputer yang relatif baru. Dan sampai sekarang orang masih memperdebatkan untuk menempatkan *data mining* di bidang ilmu mana, karena *data mining* menyangkut *database*, kecerdasan buatan (*artificial intelligence*), statistik, dsb. Ada pihak yang berpendapat bahwa *data mining* tidak lebih dari *machine learning* atau analisa statistik yang berjalan di atas *database*. Namun pihak lain berpendapat bahwa *database* berperan penting di *data mining* karena *data mining* mengakses data yang ukurannya besar (bisa sampai terabyte) dan disini terlihat peran penting *database* terutama dalam optimisasi *query*-nya.

Lalu apakah *data mining* itu? Apakah memang berhubungan erat dengan dunia pertambangan.... tambang emas, tambang timah, dsb. Definisi sederhana dari *data mining* adalah ekstraksi informasi atau pola yang penting atau menarik dari data yang ada di *database* yang besar. Dalam jurnal ilmiah, *data mining* juga dikenal dengan nama *Knowledge Discovery in Databases* (KDD).

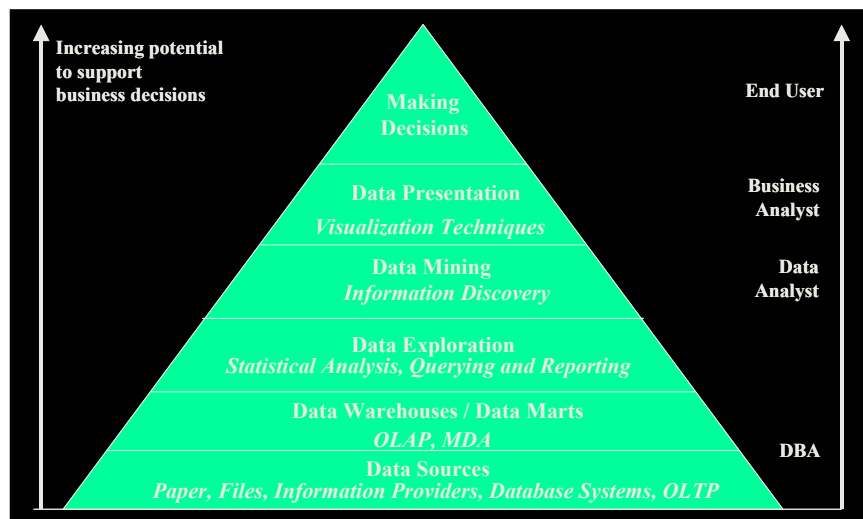
Kehadiran *data mining* dilatar belakang dengan problema *data explosion* yang dialami akhir-akhir ini dimana banyak organisasi telah mengumpulkan data sekian tahun lamanya (data pembelian, data penjualan, data nasabah, data transaksi dsb.). Hampir semua data tersebut dimasukkan dengan

menggunakan aplikasi komputer yang digunakan untuk menangani transaksi sehari-hari yang kebanyakan adalah OLTP (*On Line Transaction Processing*). Bayangkan berapa transaksi yang dimasukkan oleh *hypermarket* semacam Carrefour atau transaksi kartu kredit dari sebuah bank dalam seharinya dan bayangkan betapa besarnya ukuran data mereka jika nanti telah berjalan beberapa tahun. Pertanyaannya sekarang, apakah data tersebut akan dibiarkan menggunung, tidak berguna lalu dibuang, ataukah kita dapat me-‘nambang’-nya untuk mencari ‘emas’, ‘berlian’ yaitu informasi yang berguna untuk organisasi kita. Banyak diantara kita yang kebanjiran data tapi miskin informasi.

Jika Anda mempunyai kartu kredit, sudah pasti Anda bakal sering menerima surat berisi brosur penawaran barang atau jasa. Jika Bank pemberi kartu kredit Anda mempunyai 1.000.000 nasabah, dan mengirimkan sebuah (hanya satu) penawaran dengan biaya pengiriman sebesar Rp. 1.000 per buah maka biaya yang dihabiskan adalah Rp. 1 Milyar!! Jika Bank tersebut mengirimkan penawaran sekali sebulan yang berarti 12x dalam setahun maka anggaran yang dikeluarkan per tahunnya adalah Rp. 12 Milyar!! Dari dana Rp. 12 Milyar yang dikeluarkan, berapa persenkah konsumen yang benar-benar membeli? Mungkin hanya 10 %-nya saja. Secara harfiah, berarti 90% dari dana tersebut terbuang sia-sia.

Persoalan di atas merupakan salah satu persoalan yang dapat diatasi oleh *data mining* dari sekian banyak potensi permasalahan yang ada. *Data mining* dapat menambang data transaksi belanja kartu kredit untuk melihat manakah pembeli-pembeli yang memang potensial untuk membeli produk tertentu. Mungkin tidak sampai presisi 10%, tapi bayangkan jika kita dapat menyaring 20% saja, tentunya 80% dana dapat digunakan untuk hal lainnya.

Lalu apa beda *data mining* dengan *data warehouse* dan OLAP (*On-line Analytical Processing*)? Secara singkat bisa dijawab bahwa teknologi yang ada di *data warehouse* dan OLAP dimanfaatkan penuh untuk melakukan *data mining*. Gambar di bawah menunjukkan posisi masing-masing teknologi:



Gambar 1: *Data mining* dan teknologi *database* lainnya

Dari gambar di atas terlihat bahwa teknologi *data warehouse* digunakan untuk melakukan OLAP, sedangkan *data mining* digunakan untuk melakukan *information discovery* yang informasinya lebih ditujukan untuk seorang *Data Analyst* dan *Business Analyst* (dengan ditambah visualisasi tentunya). Dalam prakteknya, *data mining* juga mengambil data dari *data warehouse*. Hanya saja aplikasi dari *data mining* lebih khusus dan lebih spesifik dibandingkan OLAP mengingat *database* bukan satu-satunya bidang ilmu yang mempengaruhi *data mining*, banyak lagi bidang ilmu yang turut memperkaya *data mining* seperti: *information science* (ilmu informasi), *high performance computing*, visualisasi, *machine learning*, statistik, *neural networks* (jaringan syaraf tiruan), pemodelan matematika, *information retrieval* dan *information extraction* serta pengenalan pola. Bahkan pengolahan citra (*image processing*) juga digunakan dalam rangka melakukan *data mining* terhadap data *image/spatial*.

Dengan memadukan teknologi OLAP dengan *data mining* diharapkan pengguna dapat melakukan hal-hal yang biasa dilakukan di OLAP seperti *drilling/rolling* untuk melihat data lebih dalam atau lebih umum, *pivoting*, *slicing* dan *dicing*. Semua hal tersebut diharapkan nantinya dapat dilakukan secara interaktif dan dilengkapi dengan visualisasi.

Data mining tidak hanya melakukan *mining* terhadap data transaksi saja. Penelitian di bidang *data mining* saat ini sudah merambah ke sistem *database* lanjut seperti *object oriented database*, *image/spatial database*, *time-series data/temporal database*, teks (dikenal dengan nama *text mining*), web (dikenal dengan nama *web mining*) dan *multimedia database*.

Meskipun gaungnya mungkin tidak seramai seperti ketika *Client/Server Database* muncul, tetapi industri-industri seperti IBM, Microsoft, SAS, SGI, dan SPSS terus gencar melakukan penelitian-penelitian di bidang *data mining* dan telah menghasilkan berbagai *software* untuk melakukan *data mining*:

- Intelligent Miner dari IBM. Berjalan di atas sistem operasi AIX, OS/390, OS/400, Solaris dan Windows NT. Dijual dengan harga sekitar US\$60.000. Selain untuk data IBM juga mengeluarkan produk Intelligent Miner untuk teks. *Web site*:
www.software.ibm.com/data/iminer/fortext
www-4.inm.com/software/data/iminer/fordata/index.html
- Microsoft juga telah menambahkan fasilitas *data mining* di Microsoft SQL Server 2000. *Web site*: www.microsoft.com/sql/productinfo/feaover.htm
- Enterprise Miner dari SAS. Berjalan di atas sistem operasi AIX/6000, CMS, Compaq Tru64 UNIX, HP-UX, IRIX, Intel ABI, MVS, OS/2, Open VMS Alpha, Open VMS Vax, Solaris, dan Windows. *Web site*: www.sas.com
- MineSet dari Silicon Graphics. Berjalan di atas sistem operasi Windows 9x/NT dan IRIX. Dijual per *seat* seharga US\$995, server (Windows NT) seharga US\$35.000 dan untuk IRIX dijual US\$50.000. *Web site*: www.sgi.com/software/mineset
- Clementine dari SPSS. Berjalan di atas sistem operasi UNIX dan Windows NT. *Web site*: www.spss.com/software/clementine

Beberapa penelitian sekarang ini sedang dilakukan untuk memajukan *data mining* diantaranya adalah peningkatan kinerja jika berurusan dengan data berukuran *terabyte*, visualisasi yang lebih menarik untuk *user*, pengembangan bahasa *query* untuk *data mining* yang sedapat mungkin mirip dengan SQL. Tujuannya tidak lain adalah agar *end-user* dapat melakukan *data mining* dengan mudah dan cepat serta mendapatkan hasil yang akurat.