BAB II
LANDASAN TEORI

II.1 Data, Information dan Knowledge

II.1.1 Data

Data adalah informasi yang telah diubah ke dalam suatu bentuk yang lebih tepat untuk diubah atau diproses [www.SearchData Management.com, 2007].

Data adalah sesuatu yang belum mempunyai arti bagi penerimanya dan masih memerlukan adanya suatu pengolahan.

Data dapat direpresentasikan sebagai suatu fakta, angka atau teks yang dapat diproses oleh komputer sehingga menghasilkan informasi. Informasi merupakan pola, asosiasi, atau relasi yang diperoleh dari data.

II.1.2 Information

Information adalah perangsang yang artinya dalam beberapa konteks bagi penerimanya sehingga mendorong untuk melakukan sesuatu [www.SearchDataManagement.com, 2007].

Informasi merupakan hasil pengolahan dari sebuah model, formasi, organisasi, ataupun suatu perubahan bentuk dari data yang memiliki nilai tertentu, dan bisa digunakan untuk menambah pengetahuan bagi yang menerimanya [http://en.wikipedia.org/wiki/Data, 2007].

Definisi dari Informasi adalah kumpulan data yang sudah diolah sehingga dapat dipergunakan sesuai dengan kebutuhan atau keperluan penggunaan informasi tersebut.

Dalam hal ini, data bisa dianggap sebagai obyek dan informasi adalah suatu subyek yang bermanfaat bagi

penerimanya. Informasi juga bisa disebut sebagai hasil pengolahan ataupun pemrosesan data.

Suatu informasi dapat dikatakan berkualitas bila memenuhi tiga sifat berikut ini (*Burch dan Grudnitski : 1986*):

1. Akurat.

   Informasi tersebut harus bebas dari kesalahan-kesalahan. Informasi harus secara jelas, karena ketidakakuratan informasi akan mengakibatkan keputusan yang tidak tepat.

2. Tepat pada waktunya.

   Suatu informasi yang sudah terlambat tidak akan mempunyai nilai lagi, karena informasi merupakan landasan didalam pengambilan keputusan. Dewasa ini mahalnya informasi disebabkan informasi tersebut didapatkan secara cepat dan akurat, untuk melakukannya diperlukan teknologi yang canggih yaitu komputer.

3. Relevansi.

   Setiap orang mengambil suatu tindakan atau keputusan memerlukan informasi yang berbeda-beda, sehingga informasi dikatakan relevan jika informasi tersebut diberikan kepada orang-orang yang betul-betul membutuhkan. Nilai suatu informasi ditentukan oleh 2 hal yaitu manfaat dan biaya.

II.1.3 Knowledge

Knowledge (pengetahuan) merupakan konversi dari informasi yang berupa pola historis dan trend masa yang akan datang yang dapat digunakan untuk menyelesaikan suatu masalah.

Informasi yang berguna yang berasal dari data-data mentah kemudian diolah sehingga menghasilkan suatu *knowledge*. *Knowledge* digunakan untuk membantu dalam menyelesaikan masalah serta dalam pengambilan keputusan.

II.2 Basis Data

Berikut adalah beberapa definisi dari basis data :

a Anthony J. Fabbri dan A. Robert Schwab dalam bukunya, *Practical Database Management*, menyatakan:
"Basis data adalah sistem berkas terpadu yang dirancang terutama untuk meminimalkan pengulangan data".

b C.J. Date, melalui bukunya *An Introduction to Database Systems*, mengatakan:
"Basis data dapat dianggap sebagai tempat untuk sekumpulan berkas data terkomputasi".

c Himpunan kelompok data (arsip) yang saling berhubungan yang diorganisasi sedemikian rupa agar kelak dapat dimanfaatkan kembali dengan cepat dan mudah [Widjajanto, 2004].

d Kumpulan data yang saling berhubungan yang disimpan secara bersama sedemikian rupa dan tanpa pengulangan (redundansi) yang tidak perlu, untuk memenuhi berbagai kebutuhan [Widjajanto, 2004].

e Kumpulan *data* yang saling berhubungan yang disimpan dalam media penyimpan elektronis [Widjajanto, 2004].

Beberapa pendekatan dalam perancangan suatu basis data antara lain :

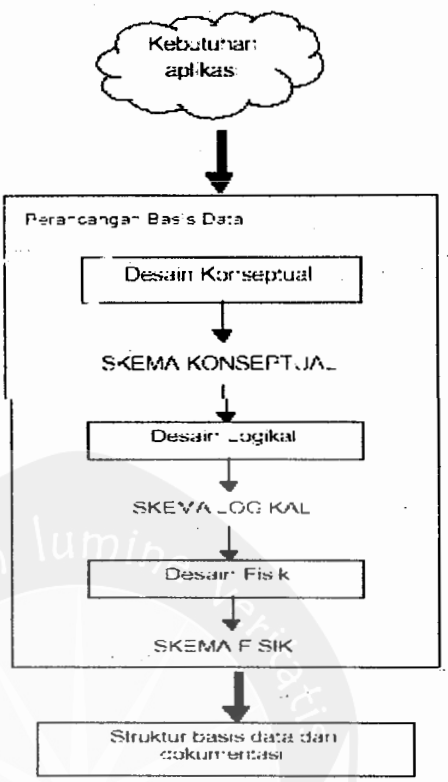
a Bottom-Up

Dari seluruh atribut yang dibutuhkan kemudian dikelompokkan menjadi table, sering disebut Normalisasi.

b Top Down

Dimulai dengan pemodelan data dan hubungan antar data (*relationship*) dan entitas hingga penentuan atribut yang sesuai. Proses ini merupakan pemodelan Entity Relationship Diagram / E-R Model.

Metodologi untuk perancangan basis data adalah sebagai berikut :



Gambar 2.1 Tahapan Perancangan Basis Data [Widjajanto, 2004].

a Perancangan Konseptual

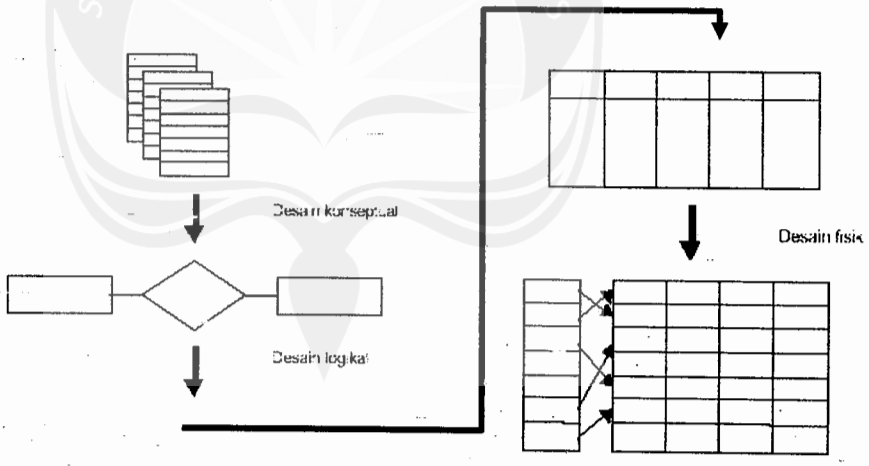
Merepresentasikan kebutuhan suatu aplikasi dalam bentuk deskripsi yang formal dan lengkap, tetapi tergantung pada sistem basis data tertentu. Dihasilkan skema konseptual yang dikenal dengan model data konseptual yang memungkinkan kita untuk mendeskripsikan organisasi data pada level abstraksi yang tinggi (E-R Model).

b Perancangan Logikal

Pengubahan dari skema konseptual menjadi model data yang akan diterapkan pada sistem manajemen basis data yang tersedia. Dihasilkan skema logikal yang dikenal dengan model data logikal yang mana akan merepresentasikan data tanpa tergantung pada detail fisik maupun DBMS yang digunakan untuk implementasi (Normalisasi).

c Perancangan Fisik

Skema logikal yang telah diperoleh dilengkapi dengan detail implementasi fisik (organsisasi file dan *indexing*) suatu DBMS. Hasil dari tahap ini adalah skema fisik yang disebut model data fisik. Model ini tergantung pada DBMS yang akan digunakan.



Gambar 2.2 Hasil dari Perancangan Basis Data [Widjajanto, 2004].

Tujuan dibangun suatu Basis data :

a Kecepatan dan kemudahan (*Speed*).

b Efisiensi ruang penyimpanan (*Space*).

c Keakuratan (*Accuracy*).

d Ketersediaan (*Availability*).

e Kelengkapan (*Completeness*).

f Keamanan (*Security*).

g Kebersamaan pemakai (*Shareability*).

II.3 *Data Base Management System*

Pengelolaan suatu basis data secara fisik dilakukan oleh sebuah perangkat lunak yang disebut *Data Base Management System* (DBMS).

Database Management System (DBMS) adalah himpunan dari inter-relasi data, disebut *database*, dan seperangkat program untuk memanage dan mengakses data *(Han* dan *Kamber, 2001)*.

Contoh DBMS :

- Untuk kelas sederhana : dBase III+, dBase IV, FoxBase, Ms Access, dan Borland Paradox.
- Untuk kelas kompleks : Borland-Interbase, Ms SQL Server, CA-Open Ingres, Oracle, Informix dan Sybase.

II.4 *Knowledge Data Discovery (KDD)*

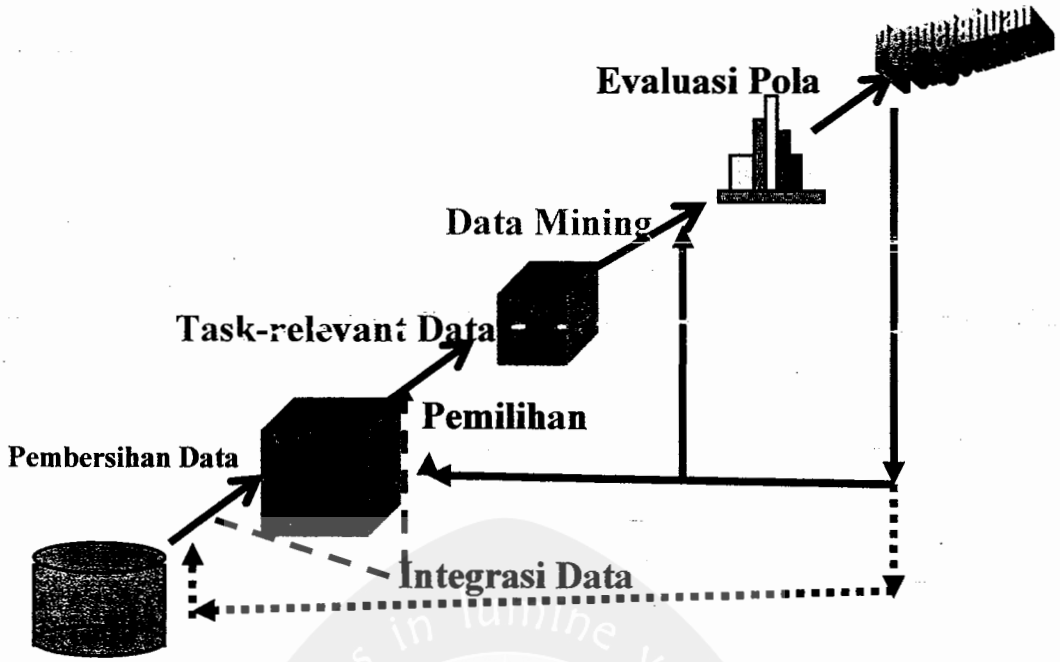
II.4.1 Definisi.

Knowledge Data Discovery adalah proses pencarian otomatis data dalam volume luas untuk menemukan pola dengan menggunakan tools seperti klasifikasi, asosiasi, cluster, dan sebagainya [http://en.Wikipedia.org/Wiki/Data_mining, 2007].

II.4.2 Proses KDD

Dalam buku "*Data Mining Concepts and Techniques*", proses Knowledge *Data Discovery* adalah sebagai berikut :

1. *Data Cleaning*
   Membersihkan *noisy* pada data dan data yang tidak konsisten.
2. *Data Integration*
   Mengkombinasikan banyak *datasource*.
3. *Data Selection*
   Data yang relevan untuk dilakukan analisis pada basis data.
4. *Data Transformation*
   Data ditransformasikan/dikonsolidasikan dalam bentuk yang siap di lakukan *mining*.
5. *Data Mining*
   Proses esensial dimana metode intelegensi diterapkan untuk mengekstrak pola data.
6. *Pattern Evaluation*
   Identifikasi pola yang menarik yang merepresentasikan pengetahuan dengan menggunakan ukuran tertentu.
7. *Knowledge Information*
   Visualisasi menggunakan teknik representasi pengetahuan.



Gambar 2.3 *Data Mining* merupakan bagian dari KDD [Purba, 2006].

II.5 Data Mining

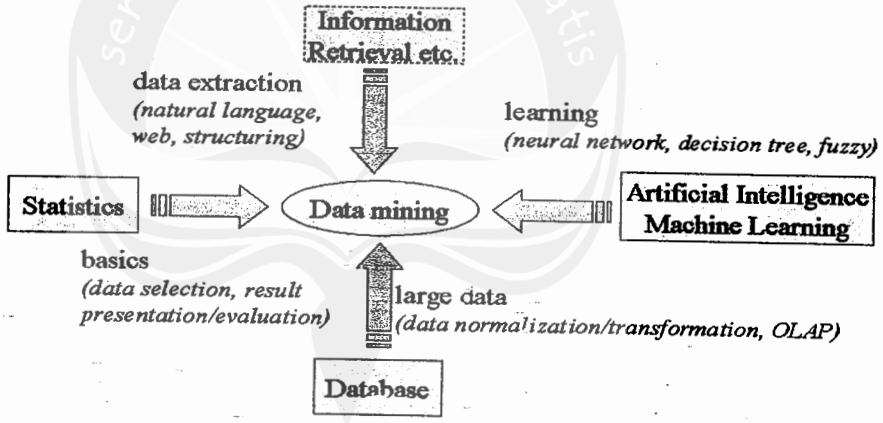
Perkembangan teknologi khususnya teknologi informasi (*Information Technology*) dewasa ini sudah sangat menjamur. Seiring dengan perkembangan tersebut, permasalahan yang ada juga semakin kompleks.

Dalam perkembangan teknologi informasi tersebut juga berkembang ilmu yang menjadikan perkembangan tersebut semakin dahsyat. Ilmu-ilmu yang sudah ada dewasa ini diantaranya adalah *Artificial Intelligence, Statistics, Information System, Computers Graphics, Data Structure, Database Management System,* dan sebagainya.

Permasalahan yang ada dan sekarang ini sering muncul ialah bukan pada bagaimana kita mendapatkan

data, akan tetapi apa yang akan kita lakukan terhadap data yang ada dalam jumlah yang besar atau dengan kata lain kita memiliki segudang data tetapi pengetahuan yang diperoleh sedikit atau kita melakukan pencarian pengetahuan dengan menemukan pola-pola yang menarik dari data-data tersebut.

Data Mining bukanlah suatu bidang yang sama sekali baru. Salah satu kesulitan untuk mendefinisikan *Data Mining* adalah kenyataan bahwa *Data Mining* mewarisi banyak aspek dan teknik dari bidang-bidang ilmu yang sudah mapan terlebih dulu. Gambar berikut menunjukkan bahwa *Data Mining* memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (*artificial intelligent*), *machine learning*, *statistic*, *database* dan juga *information retrieval*.



Gambar 2.4 Hubungan *Data Mining* dengan ilmu lainnya.

Data Mining adalah ekstraksi atau penambangan pengetahuan dari data yang besar *(Han and Kamber, 2001)*.

Beberapa teknik yang sering disebut-sebut dalam literatur *Data Mining* seperti klasifikasi, *neural network*, dan *genetic algorithm*. Statistik memberikan

kontribusi pada *Data Mining* dengan teknik-teknik untuk menyeleksi data dan evaluasi hasil *Data Mining* selain teknik-teknik *Data Mining* seperti klasterisasi.

Yang membedakan persepsi terhadap *Data Mining* adalah perkembangan teknik-teknik *Data Mining* untuk aplikasi pada *database* skala besar. Sebelum populernya *Data Mining*, teknik-teknik tersebut pada umumnya diterapkan untuk data skala kecil saja. Selain itu beberapa teknik dari bidang *database* untuk transformasi data juga merupakan bagian integral dari proses *Data Mining*.

Akhir-akhir ini ada beberapa bidang ilmu seperti *information retrieval* yang juga terlibat dalam proses *Data Mining* untuk mengekstrak sumber data bagi *Data Mining* seperti teks dan website.

Walaupun *Data Mining* memiliki sumber dari beberapa bidang ilmu, *Data Mining* berbeda dalam beberapa aspek dibandingkan dengan bidang ilmu yang lain yakni sebagai berikut :

- Statistik : model statistik dipersiapkan oleh para ahli statistik, sedangkan *Data Mining* mengembangkan statistik untuk menangani data berjumlah besar secara otomatis.
- *Expert system* (Sistem Pakar) : model pada *expert system* dibuat berupa aturan-aturan berdasar pada pengalaman-pengalaman para ahli.
- *Data Warehouse (DWH)* : sering terjadi kerancuan antara *Data Mining* dan *Data Warehouse* karena keduanya sering

dipakai bersamaan. Pada umumnya *data warehouse* lebih merujuk pada tempat untuk menyimpan data yang terkonsolidasi sedangkan *data mining* bisa dianggap sebagai perkakas untuk menganalisa otomatis nilai dari data tersebut.

- *OLAP (On-Line Analytical Processing)* : seperti *data warehouse*, OLAP juga sering dibahas bersama *Data Mining*. Tetapi OLAP memiliki tujuan untuk memastikan hipotesa yang sudah diformulasikan terlebih dulu oleh penggunanya.

Tujuan dari *data mining* pada dasarnya adalah melakukan ekstraksi informasi level tinggi pada sekumpulan data atau dengan kata lain mencari pola yang sering muncul, asosiasi, korelasi, atau struktur sebab musabab diantara himpunan item-item atau objek-objek dalam database transaksi, database relasional, dan penyimpanan informasi lainnya.

Proses data dalam *data mining* yaitu *preprocessing*, *data mining*, dan *post processing*. Proses data sebelum dilakukan *mining* harus dilakukan *preprocessing*. Hal ini dikarenakan, dalam kenyataannya data mungkin tidak lengkap, *noisy*, dan *inconsistent*. Data yang lebih baik akan menghasilkan *data mining* yang lebih baik.

Data *preprocessing* membantu didalam memperbaiki presisi dan kinerja data mining dan mencegah kesalahan didalam *data mining*.

II.5.1 Fungsionalitas *Data Mining*

Fungsionalitas *data mining* digunakan untuk menspesifikasikan bermacam-macam pola yang ditemukan dalam melakukan *data mining task*. Secara umum *data mining task* diklasifikasikan menjadi 2 kategori yaitu *descriptive* dan *predictive*. *Descriptive* melakukan karakterisasi secara umum terhadap data dalam *database*. *Predictive* menjaga performansi interferensi pada data yang kemudian digunakan untuk melakukan prediksi.

Fungsionalitas *data mining* adalah sebagai berikut :

a. *Discovery of Concepts/Class Description (Characterization and Discrimination).*

   Data Characterization adalah rangkuman dari karakteristik atau fitur data secara umum dari kelas target.

   Data Discrimination adalah perbandingan antara fitur secara umum dari kelas target obyek data dengan fitur umum suatu obyek dari kelas lain.

b. *Association Analysis.*

   Menemukan aturan asosiasi berdasarkan atribut dan nilai yang sering muncul bersama pada suatu data.

c. *Classification and Prediction.*

   Mencari sebuah model yang dapat melakukan prediksi pada suatu data baru yang belum pernah ada

d. *Cluster Analysis.*

   Mengelompokan data ke dalam sebuah *cluster* berdasarkan kemiripannya dengan memaksimalkan

kemiripan dalam sebuah *cluster* dan meminimalisasikan kemiripan antar *cluster*.

e. *Outlier Analysis.*

Mencari data obyek yang bersifat anomaly (berbeda dengan sifat umum data).

f. *Evolution Analysis.*

Mencari tren untuk data yang bersifat dinamis/berubah-ubah.

II.6 *Association Rules.*

II.6.1 Definisi

Association rules merupakan aturan dalam bentuk X => Y, dimana X dan Y adalah suatu event/peristiwa. *Association rules* terdiri dari 2 ruas yaitu kanan dan kiri, dimana ruas kiri merupakan deskripsi subset dari populasi sedangkan ruas kanan mendeskripsikan kebiasaan yang tidak biasanya terhadap apa yang dideskripsikan diruas kiri.

Aturan tersebut dinyatakan dengan suatu nilai probabilitas yang disebut *confidence*.

Support, supp(X) dari suatu itemset X adalah rasio dari jumlah transaksi dimana suatu itemset muncul dengan total jumlah transaksi.

Confidence (keyakinan) dari kaidah ($X \Rightarrow Y$) ditulis conf($X \Rightarrow Y$) adalah

- $conf(X \Rightarrow Y) = supp(X \cup Y) \ / \ supp(X)$
- *Confidence* bisa juga didefinisikan dalam terminologi peluang bersyarat

  $conf(X \Rightarrow Y) = P(Y|X) = P(X \cap Y)/P(X)$.

Frequent itemset didefinisikan sebagai itemset dimana support lebih besar atau sama dengan minimum

support yang merupakan ambang yang diberikan oleh user. Dalam setiap item terdapat nilai yang menyatakan banyaknya item terjual. Item dengan jumlah k tersebut disebut k-itemset (k adalah himpunan item yang muncul secara bersama).

Misalnya, terdapat pola dimana 30% transaksi membeli *beer* juga membeli *diapers*, kemudian 2% dari transaksi membeli keduanya. Hal tersebut dapat dikatakan 30% merupakan *Confidence* dan 2% merupakan support dari kaidah asosiasi. Kaidah asosiasi merupakan suatu aturan dimana item-item tertentu hadir secara bersama-sama $X \Rightarrow Y$ dimana $(X \cap Y) \neq \varnothing$.

Secara matematis dapat ditulis sebagai berikut :

$$\textbf{support}(A \Rightarrow B\ [\ s, c\ ]) = p(A \cup B) = \underline{\textbf{support}\ (\{A,B\})}$$

- Support : menunjukkan frekuensi dari kaidah didalam transaksi.

$$\textbf{confidence}(A \Rightarrow B\ [\ s, c\ ]) = p(B|A) = p(A \cup B) / p(A) = \underline{\textbf{support}(\{A,B\})} / \underline{\textbf{support}(\{A\})}$$

- Confidence : menunjukkan persentasi dari transaksi yang memuat A yang juga memuat B.

Association rules mining adalah suatu proses dengan langkah sebagai berikut :

STEP 1: cari frequent itemset (himpunan item-item yang memiliki minimum support).

- Disebut trik Apriori: suatu subset tak hampa dari suatu frequent itemset haruslah juga suatu frequent itemset.
- Artinya, jika {AB} adalah suatu frequent itemset, kedua {A} dan {B} harus juga frequent itemsets
- Secara iteratif cari frequent itemsets dengan ukuran dari 1 hingga k (k-itemset)

STEP 2: gunakan frequent itemset untuk membangun kaidah asosiasi.

Association rules didesain untuk membantu dalam menemukan fenomena yang menarik dalam suatu *database*.

II.6.2 Algoritma Apriori

Dalam data mining, sangat banyak teknik algoritma yang digunakan, salah satunya adalah Algoritma Apriori. Prinsip Apriori adalah jika terdapat itemset yang tidak frequent, maka itemset tersebut tidak perlu diekstrak supersetnya sehingga jumlah kandidat yang diperiksa menjadi berkurang.

Algoritma Apriori menggunakan paradigma *generate and test* artinya pembuatan kandidat kombinasi item yang mungkin diuji berdasarkan aturan tertentu kemudian diuji apakah kombinasi item tersebut memenuhi kriteria *Minimum Support*. Kombinasi item (*frequent itemset*) yang memenuhi syarat tersebut kemudian diuji untuk membuat aturan yang memenuhi syarat *Minimum Confidence*.

Inti dari algoritma apriori :

- Gunakan frequent (k - 1)-itemset untuk membangun kandidat frequent k-itemset.
- Gunakan scan database dan pencocokan pola untuk mengumpulkan hitungan untuk kandidat itemsets.

II.7 Visual Basic .Net 2003

Millennium ketiga disambut oleh Microsoft dengan meluncurkan teknologi berbasis .NET. Salah satunya adalah Microsoft Visual Basic .NET.

Visual Basic versi sebelumnya yaitu Visual Basic 6 diluncurkan Microsoft pada tahun 1998 yang kemudian dikembangkan menjadi Microsoft Visual Basic .NET Framework di Orlando, Florida, Amerika Serikat. Pengembangan tersebut dikarenakan keterbatasan Visual Basic versi sebelumnya, lalu diupgrade menjadi VB.NET.

Pada pemrograman *database* disediakan teknologi ADO.NET yang merupakan kumpulan *class* dan berisi komponen untuk melakukan koneksi, akses, dan manipulasi database. Dalam ADO.NET terdapat *provider* data *SQL Server* dan OLE DB. Di VB.NET versi 2003 provider datanya ditambah dengan ODBC dan Oracle.

II.8 Visual C# .NET 2003

Visual C#. NET 2003 juga merupakan salah satu dari bagian *.NET Framework*. Secara global, C#.NET memiliki struktur yang sama dengan VB.NET hanya saja bahasa pemrograman yang digunakan adalah berbeda. VB.NET memiliki bahasa menyerupai VB biasa sedangkan C#.NET memiliki bahasa yang sangat berbeda dengan VB.NET,

bahasa C#.NET menyerupai bahasa C baik struktur, deklarasi dan pengembangannya.

Didalam C#.NET juga terdapat konektivitas ADO.NET yang memungkinkan untuk melakukan koneksi ke *database*, misalnya SQL Server 2000.

Dalam aplikasi ini, C#.NET digunakan untuk membuat file dll (*Dynamic Link Library*) dikarenakan dengan kemudahan dan keluwesannya dibandingkan dengan VB.NET, secara kesuluruhan dalam pengembangan software ini menggunakan VB.NET.

II.9 SQL Server 2000

Sql Server 2000 merupakan tools DBMS (*Database Management System*) dengan kesatuan komponen yang bekerja sama untuk memaintaince *data storage* dan kebutuhan analisis pada data dalam skala kecil atau *enterprise data processing system*.

Fitur *database relational* pada SQL Server 2000 memproteksi integritas data dengan meminimalisasi managemen user yang melakukan modifikasi secara konkuren terhadap *database*.