UNIVERSITAS
NUSA MANDIRI

MODUL AJAR MATA KULIAH

DATA WAREHOUSE DAN BUSINESS INTELLIGENCE

Program Studi Sistem Informasi
Fakultas Teknologi Informasi
Universitas Nusa Mandiri
TA. 2023-2024

Modul Ajar Mata Kuliah Data Warehouse dan Business Intelligence
TA. 2023-2024

Tim Penyusun:
Ami Rahmawati, M.Kom

KATA PENGANTAR

Puji syukur kami panjatkan ke hadirat Tuhan Yang Maha Esa, karena atas rahmat dan karunia-Nya, modul ajar mata kuliah "Data Warehouse dan Business Intelligence" ini dapat disusun dan diselesaikan dengan baik. Modul ini dirancang untuk membantu mahasiswa dalam memahami konsep, teori, dan aplikasi praktis terkait pengelolaan data serta pemanfaatan teknologi Business Intelligence (BI) dalam mendukung pengambilan keputusan bisnis yang efektif.

Perkembangan teknologi informasi yang pesat telah mengubah cara bisnis beroperasi dan berinteraksi dengan data. Data Warehouse (DW) sebagai fondasi dari sistem BI memungkinkan perusahaan untuk mengonsolidasi dan mengelola data dari berbagai sumber secara efisien. Dengan modul ini, mahasiswa diharapkan dapat menguasai teknik perancangan dan implementasi DW, serta memahami bagaimana BI dapat digunakan untuk menganalisis data secara mendalam guna menghasilkan wawasan yang berharga bagi organisasi. Modul ini mencakup berbagai topik, mulai dari pengenalan dasar tentang konsep DW dan BI, proses ETL (*Extract, Transform, Load*), hingga eksplorasi berbagai alat dan teknik BI yang digunakan dalam analisis data.

Kami menyadari bahwa modul ini masih memiliki kekurangan dan terbuka untuk perbaikan. Oleh karena itu, kami sangat mengharapkan masukan dan saran konstruktif dari para pembaca, dosen, dan mahasiswa demi kesempurnaan modul ini di masa yang akan datang. Akhir kata, kami mengucapkan terima kasih kepada semua pihak yang telah berkontribusi dalam penyusunan modul ini. Semoga modul ajar ini dapat memberikan manfaat yang optimal bagi mahasiswa dalam mengembangkan keterampilan dan pengetahuan di bidang Data Warehouse dan Business Intelligence.

Jakarta, 29 Maret 2024

Penyusun

DAFTAR ISI

Cover
Kata Pengantar 1
Daftar Isi 2

Modul 1: Data Warehouse Dan Business Intelligence I 3
Modul 2: Data Warehousing 11
Modul 3: Business Performance Management 23
Modul 4: BPM Methodologies 25
Modul 5: Data Mining 26
Modul 6: Metode Learning Algoritma Data Mining 35
Studi Kasus: Klasifikasi Pada Data Lung Cancer 39

Daftar Pustaka 42

MODUL 1: Data Warehouse dan Business Intelligence I

Kompetensi: Mahasiswa mampu memahami konsep Business Intelligence dan memyimpulkan hubungan DSS dan BI

Definisi Business Intelligence, Arsitektur Business Intelligence, Komponen Business Intelligence, Siklus Analisis Business Intelligence, Pengembangan Sistem Business Intelligence, Contoh Penggunaan Business Intelligence

Definisi Business Intelligence

Data Warehousing Institute mendefinisikan Business Intelligence sebagai: "Proses, teknologi, dan alat yang diperlukan untuk mengubah data menjadi informasi, informasi menjadi pengetahuan, dan pengetahuan menjadi rencana yang mendorong tindakan bisnis yang menguntungkan. Intelijen bisnis mencakup data warehouse, business analytic tools, dan manajemen konten/pengetahuan."

Secara konseptual, data, informasi, dan pengetahuan memiliki definisi sebagai berikut:

a. **Data adalah** kumpulan elemen nilai atau fakta yang digunakan untuk menghitung, menalar, atau mengukur yang disimpan dalam berbagai bentuk.
b. **Informasi adalah** hasil pengumpulan dan pengorganisasian data sedemikian rupa sehingga membentuk hubungan antar item data, sehingga memberikan konteks dan makna.
c. **Pengetahuan adalah** konsep pemahaman informasi berdasarkan pola yang dikenali dengan cara yang memberikan wawasan terhadap informasi.

1. Mengubah Informasi Menjadi Pengetahuan

Proses mengubah data menjadi informasi dapat diringkas sebagai proses menentukan data apa yang akan dikumpulkan dan dikelola serta dalam konteks apa. Contoh mengubah informasi menjadi pengetahuan adalah proses perancangan database yang memodelkan sekumpulan entitas di dunia nyata, seperti penjualan produk, yang digunakan untuk menyimpan data penjualan antara lain data produk yang terjual, jumlah penjualan, tanggal transaksi, lokasi penjualan, dan banyak lagi. Lalu selanjutnya data tersebut diubah menjadi informasi seperti tren penjualan, sehingga membuat pengelolaan dan akses data lebih efisien.

Dalam setiap bit data, seperti nama produk, tidak memiliki nilai yang berarti. Akan tetapi setelah menetapkan bit data mana yang akan digunakan untuk mengonfigurasi deskripsi suatu pihak, serta membuat instance dan mengisi instance tersebut dengan nilai data terkait, maka data menjadi sebuah bagian informasi yang bermakna.

Aspek BI ini melibatkan infrastruktur pengelolaan dan penyajian data, yang menggabungkan platform perangkat keras, sistem basis data relasional atau jenis lainnya, dan perangkat lunak terkait. Aspek ini juga mencakup alat query dan reporting yang menyediakan akses ke data. Bagian dari proses ini tidak dapat dilakukan tanpa para ahli di bidang pengelolaan data yang mengintegrasikan dan mengoordinasikan teknologi ini.

2. Mengubah Informasi Menjadi Pengetahuan

Mengubah informasi menjadi pengetahuan adalah proses transformasi informasi yang dikumpulkan menjadi pemahaman yang lebih mendalam, relevan, dan bermakna untuk mendapatkan insight yang berharga dan digunakan untuk membuat keputusan dan tindakan yang lebih baik.

Dari contoh diatas, melalui analisis data penjualan, maka dapat mengidentifikasi tren penjualan yang berkelanjutan seperti peningkatan atau penurunan dari waktu ke waktu. Pengetahuan yang dihasilkan dari tren ini dapat membantu dalam perencanaan strategi penjualan jangka panjang dan pengambilan keputusan terkait stok dan produksi.

Aspek BI ini melibatkan komponen analytics, seperti online analytical processing (OLAP), data quality, data profiling, business rule analysis, predictive analysis, dan jenis data mining lainnya.

3. Mengubah Pengetahuan Menjadi Actionable Plans

Mampu mengambil tindakan berdasarkan kecerdasan yang telah dipelajari adalah poin kunci dari setiap strategi BI. Melalui tindakan tersebut sponsor manajemen senior dapat melihat laba atas investasi yang sebenarnya untuk belanja teknologi informasi (TI) mereka.

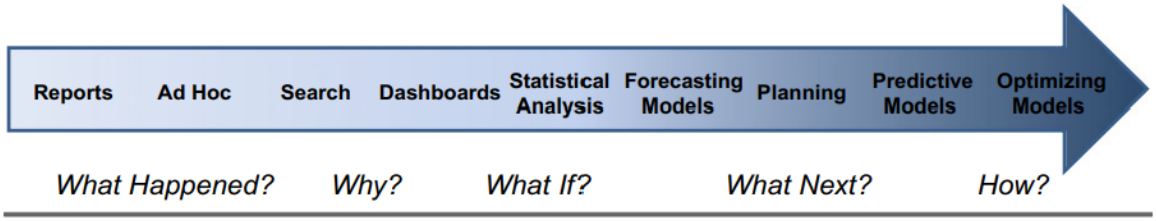
Program BI memberikan manfaat yang meningkatkan efisiensi bisnis, meningkatkan penjualan, memberikan penargetan pelanggan yang lebih baik, mengurangi biaya layanan pelanggan, mengidentifikasi penipuan, dan secara umum meningkatkan keuntungan sekaligus mengurangi biaya.

Pengetahuan yang ditemukan tidak akan bernilai jika tidak ada tindakan yang menghasilkan nilai yang dapat diambil sebagai konsekuensi dari perolehan pengetahuan tersebut. Kemitraan bisnis-teknologi harus bekerja sama tidak hanya untuk bertindak berdasarkan kecerdasan yang ditemukan, namun juga melakukannya secara tepat waktu.

Konsep **'business intelligence' and 'analytics'** mencakup alat dan teknik yang mendukung kumpulan komunitas pengguna di seluruh organisasi, sebagai hasil dari pengumpulan dan pengorganisasian berbagai sumber data untuk mendukung manajemen dan pengambilan keputusan pada tingkat operasional, taktis, dan tingkat strategis.

Organisasi yang telah mematangkan program pergudangan datanya, memungkinkan penggunanya mengekstrak pengetahuan yang dapat ditindaklanjuti dari aset informasi perusahaan dan dengan cepat mewujudkan nilai bisnis.

Ada berbagai jenis kemampuan analitis yang disediakan oleh BI, dan semuanya membantu memberikan jawaban atas serangkaian pertanyaan yang semakin berharga. Pertanyaan-pertanyaan ini semakin kompleks dan menambah nilai kumulatif yang lebih besar.



Business intelligence dapat didefinisikan sebagai seperangkat model matematika dan metodologi analisis yang memanfaatkan data yang tersedia untuk menghasilkan informasi dan pengetahuan yang berguna untuk proses pengambilan keputusan yang kompleks.

Business intelligence harus menghasilkan keputusan yang efektif dan tepat waktu, di mana dalam organisasi yang kompleks, baik pemerintah maupun swasta, keputusan dibuat secara terus-menerus. Keputusan-keputusan tersebut mungkin lebih atau kurang penting, mempunyai dampak jangka panjang atau pendek dan melibatkan orang-orang dan peran di berbagai tingkat hierarki.

Sebagian besar knowledge workers mengambil keputusan dengan menggunakan metodologi yang mudah dan intuitif yang mempertimbangkan elemen spesifik seperti pengalaman, pengetahuan tentang domain aplikasi, dan informasi yang tersedia. Pendekatan ini mengarah pada gaya pengambilan keputusan yang stagnan dan tidak sesuai dengan kondisi tidak stabil yang disebabkan oleh perubahan lingkungan ekonomi yang sering dan cepat.

Tujuan utama business intelligence systems adalah untuk memberikan alat dan metodologi kepada knowledge workers yang memungkinkan mereka membuat keputusan yang efektif dan tepat waktu.

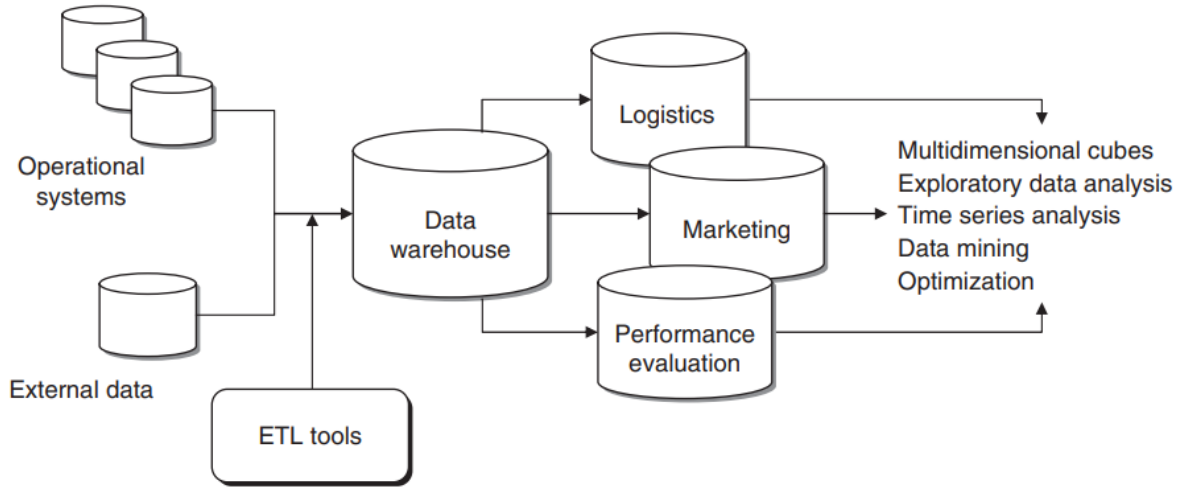
Keputusan yang efektif. Penerapan metode analisis yang ketat memungkinkan pengambil keputusan untuk mengandalkan informasi dan pengetahuan yang lebih dapat diandalkan. Hasilnya, mereka mampu membuat keputusan yang lebih baik dan menyusun rencana aksi yang memungkinkan tujuan mereka tercapai dengan cara yang lebih efektif.

Keputusan tepat waktu. Perusahaan beroperasi dalam lingkungan ekonomi yang ditandai dengan meningkatnya tingkat persaingan dan dinamisme yang tinggi. Sebagai konsekuensinya, kemampuan untuk bereaksi dengan cepat terhadap tindakan pesaing dan kondisi pasar yang baru merupakan faktor penting dalam keberhasilan atau bahkan kelangsungan hidup suatu perusahaan.

Sistem intelijen bisnis memberi pengambil keputusan informasi dan pengetahuan yang diambil dari data, melalui penerapan model matematika dan algoritma. Dalam beberapa kasus, aktivitas ini dapat direduksi menjadi penghitungan total dan persentase, yang secara grafis diwakili oleh histogram sederhana, sedangkan analisis yang lebih rumit memerlukan pengembangan model optimasi.

Penerapan sistem intelijen bisnis cenderung mendorong pendekatan ilmiah dan rasional terhadap manajemen perusahaan dan organisasi yang kompleks. Tujuan analisis diidentifikasi dan menentukan indikator kinerja yang akan digunakan untuk mengevaluasi pilihan alternatif. Model matematika kemudian dikembangkan dengan memanfaatkan hubungan antara variabel kontrol sistem, parameter dan metrik evaluasi. Analisis what-if dilakukan untuk mengevaluasi dampak terhadap kinerja yang ditentukan oleh variasi variabel kontrol dan perubahan parameter.

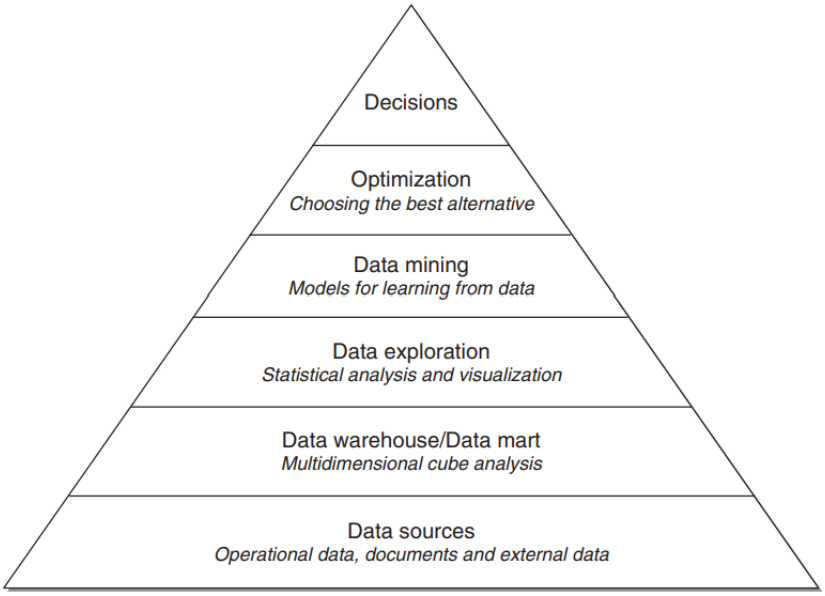
Arsitektur Business Intelligence



Gambar 1. Arsitektur Business Intelligence

1. **Data sources.** Pada tahap pertama, perlu dilakukan pengumpulan dan integrasi data yang disimpan dalam berbagai sumber primer dan sekunder, yang asal dan jenisnya heterogen. Sumbernya sebagian besar terdiri dari data milik sistem operasional, namun mungkin juga mencakup dokumen tidak terstruktur, seperti email dan data yang diterima dari penyedia eksternal.
2. **Data warehouses dan data marts.** Dengan menggunakan alat ekstraksi dan transformasi yang dikenal sebagai extract, transform, load (ETL), data yang berasal dari berbagai sumber disimpan dalam database yang dimaksudkan untuk mendukung analisis intelijen bisnis.
3. **Business intelligence methodologies.** Data akhirnya diekstraksi dan digunakan untuk memberi masukan pada model matematika dan metodologi analisis yang dimaksudkan untuk mendukung pengambil keputusan.

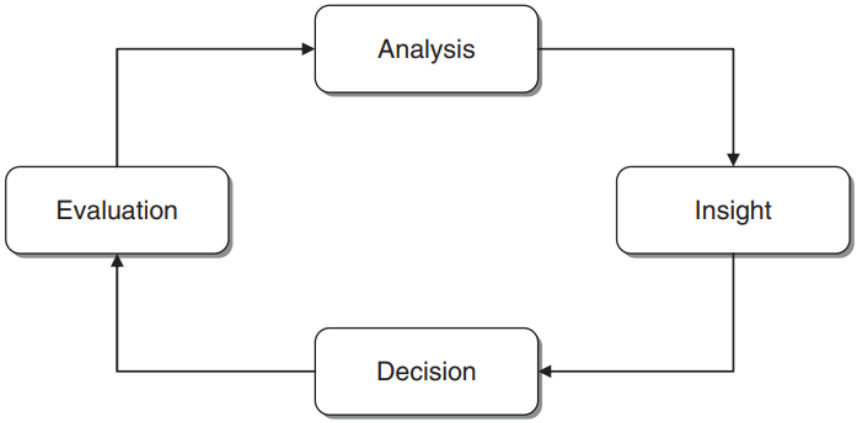
Komponen Business Intelligence



Gambar 2. Komponen Business Intelligence

1. **Data exploration.** Pada tingkat ketiga piramida kita menemukan alat untuk melakukan analisis intelijen bisnis pasif, yang terdiri dari sistem query dan reporting, serta metode statistik. Hal ini disebut sebagai metodologi pasif karena pengambil keputusan diminta untuk menghasilkan hipotesis sebelumnya atau menentukan kriteria ekstraksi data, dan kemudian menggunakan alat analisis untuk menemukan jawaban dan mengkonfirmasi wawasan awal mereka.
2. **Data mining.** Tingkat keempat mencakup metodologi intelijen bisnis aktif, yang tujuannya adalah mengekstraksi informasi dan pengetahuan dari data. Berbeda dengan alat yang dijelaskan pada tingkat piramida sebelumnya, model aktif tidak memerlukan pengambil keputusan untuk merumuskan hipotesis sebelumnya untuk kemudian diverifikasi. Tujuannya adalah untuk memperluas pengetahuan para pengambil keputusan.
3. **Optimization.** Dengan naik satu tingkat dalam piramida, telah menemukan model pengoptimalan yang memungkinkan dalam menentukan solusi terbaik dari serangkaian tindakan alternatif, yang biasanya cukup luas dan terkadang bahkan tidak terbatas.
4. **Decisions.** Terakhir, puncak piramida berhubungan dengan pilihan dan pengambilan keputusan tertentu, dan dalam beberapa hal mewakili kesimpulan alami dari proses pengambilan keputusan. Bahkan ketika metodologi intelijen bisnis tersedia dan berhasil diadopsi, pilihan keputusan tetap ada pada pengambil keputusan, yang juga dapat memanfaatkan informasi informal dan tidak terstruktur yang tersedia untuk mengadaptasi dan memodifikasi rekomendasi dan kesimpulan yang dicapai melalui penggunaan model matematika.

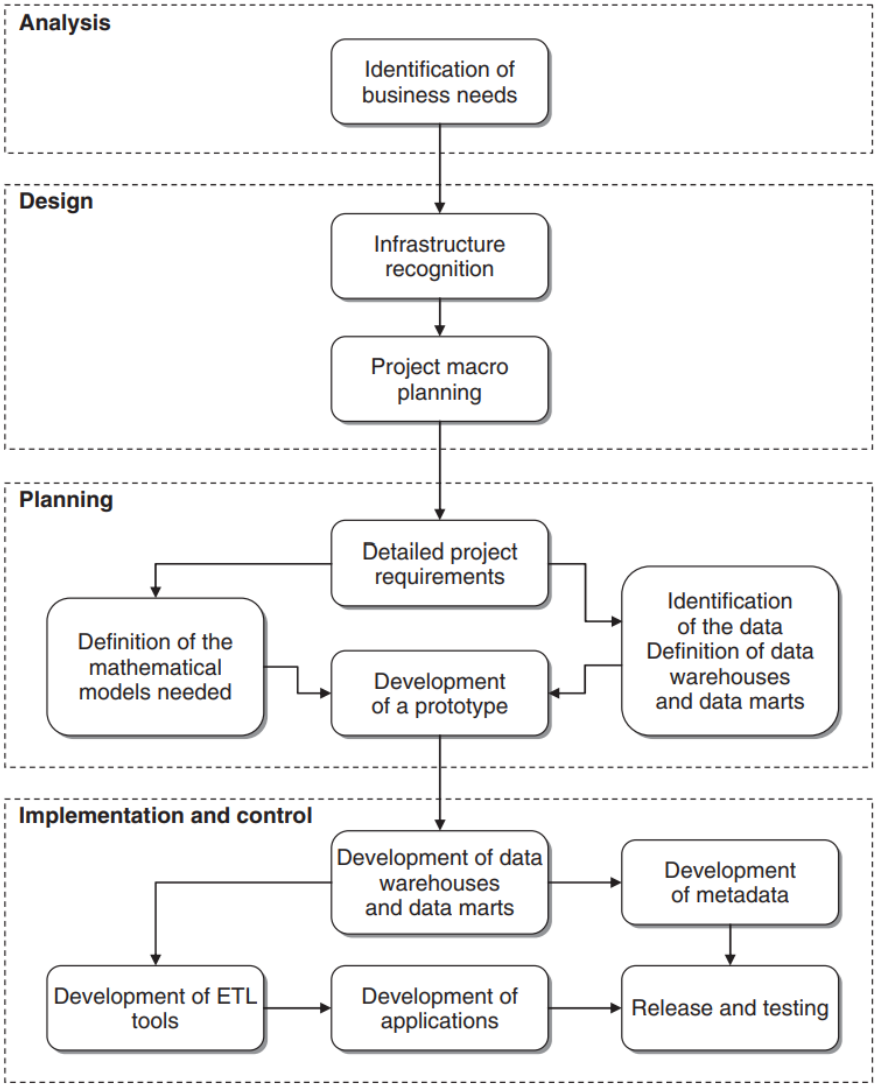
Siklus Analisis Business Intelligence



Gambar 3. Siklus Analisis Business Intelligence

1. **Analysis.** Selama tahap analisis, penting untuk mengenali dan menguraikan masalah yang ada secara akurat. Pengambil keputusan kemudian harus dapat memahami dan menginterpretasikan informasi yang diamati atau dianalisis, kemudian mengidentifikasi faktor-faktor penting yang dianggap paling relevan.
2. **Insight.** Fase kedua memungkinkan pengambil keputusan untuk memahami masalah yang dihadapi dengan lebih baik dan lebih mendalam, seringkali pada tingkat sebab-akibat.
3. **Decision.** Selama fase ketiga, pengetahuan yang diperoleh sebagai hasil dari fase insight diubah menjadi keputusan dan selanjutnya menjadi tindakan. Ketersediaan metodologi intelijen bisnis memungkinkan fase analisis dan insight dijalankan lebih cepat sehingga keputusan yang lebih efektif dan tepat waktu dapat diambil dan lebih sesuai dengan prioritas strategis organisasi tertentu.
4. **Evaluation.** Terakhir, fase keempat dari siklus intelijen bisnis melibatkan pengukuran dan evaluasi kinerja.

Pengembangan Sistem Business Intelligence



Gambar 5. Pengembangan Sistem Business Intelligence

1. **Analysis.** Pada tahap pertama, kebutuhan organisasi sehubungan dengan pengembangan sistem intelijen bisnis harus diidentifikasi secara cermat. Fase awal ini umumnya dilakukan melalui serangkaian wawancara terhadap knowledge workers yang melakukan berbagai peran dan aktivitas dalam organisasi. Penting untuk menguraikan dengan jelas tujuan umum dan prioritas proyek, serta menetapkan biaya dan manfaat yang diperoleh dari pengembangan sistem intelijen bisnis.
2. **Design.** Fase kedua mencakup dua sub-fase dan bertujuan untuk mendapatkan rencana sementara dari keseluruhan arsitektur, dengan mempertimbangkan perkembangan apa pun dalam waktu dekat dan evolusi sistem dalam jangka menengah.
   a. Perlu dilakukan penilaian terhadap infrastruktur informasi yang ada.
   b. Dengan menggunakan metodologi manajemen proyek klasik, rencana proyek akan ditetapkan, mengidentifikasi fase pengembangan, prioritas, waktu dan biaya pelaksanaan yang diharapkan, serta peran dan sumber daya yang diperlukan.

3. **Planning.** Tahap perencanaan mencakup sub-fase dimana fungsi sistem intelijen bisnis didefinisikan dan dijelaskan secara lebih rinci. Selanjutnya, data yang ada serta data lain yang mungkin diambil secara eksternal akan dinilai. Hal ini memungkinkan struktur informasi arsitektur intelijen bisnis, yang terdiri dari data warehouse dan mungkin beberapa data mart, untuk dirancang. Bersamaan dengan pengenalan data yang tersedia, model matematika yang akan diadopsi harus ditentukan, memastikan ketersediaan data yang diperlukan untuk memenuhi setiap model dan memverifikasi bahwa efisiensi algoritma yang akan digunakan akan memadai untuk besarnya dampak yang dihasilkan. Setelah itu, maka akan membuat prototipe sistem, dengan biaya rendah dan kemampuan terbatas, untuk mengetahui terlebih dahulu adanya perbedaan antara kebutuhan aktual dan spesifikasi proyek.
4. **Implementation and control.** Fase terakhir terdiri dari lima subfase utama. Pertama, data warehouse dan setiap data mart spesifik dikembangkan. Ini mewakili infrastruktur informasi yang akan mendukung sistem intelijen bisnis. Untuk menjelaskan arti data yang terdapat dalam data warehouse dan transformasi yang diterapkan terlebih dahulu pada data primer, arsip metadata harus dibuat. Selain itu, prosedur ETL ditetapkan untuk mengekstrak dan mengubah data. data yang ada di sumber utama, memuatnya ke dalam data warehouse dan data mart. Langkah selanjutnya ditujukan untuk mengembangkan aplikasi intelijen bisnis inti yang memungkinkan dilakukannya analisis yang direncanakan. Terakhir, sistem dirilis untuk pengujian dan penggunaan.

Contoh Penggunaan Business Intelligence - Meningkatkan Pendapatan

1. **Targeted marketing.** Hal ini melibatkan informasi tentang minat, kebiasaan, preferensi, dan kebutuhan pelanggan dapat membantu dalam mempersonalisasi pesan pemasaran dan menawarkan produk atau layanan yang sesuai dengan kebutuhan mereka.
2. **Cross-selling and up-selling.** Hal ini melibatkan analisis riwayat transaksi dan kecenderungan pelanggan, mengevaluasi pola keberhasilan, dan mencari perilaku umum untuk menemukan kesamaan antara profil pelanggan dan produk serta peluang untuk melakukan penjualan tambahan atau penjualan produk kelas atas.
3. **Market development.** Aktivitas ini memungkinkan analis untuk mengevaluasi karakteristik demografi individu di wilayah tertentu sebagai cara memahami ketergantungan geografis untuk mendorong program pemasaran yang lebih efisien yang ditargetkan pada profil geo-demografis berdasarkan lokasi fisik.
4. **Loyalty management.** Menentukan kapan pelanggan akan mengakhiri hubungan mereka memberikan peluang besar untuk membangun kembali hubungan dengan memberikan penawaran terpisah untuk tetap tinggal. Jenis analisis ini membantu mengurangi kehilangan pelanggan sekaligus memberikan profil yang dapat membantu menghindari pelanggan yang memiliki kecenderungan untuk membelot.

MODUL 2: Data Warehousing

Kompetensi: Mahasiswa mampu menjelaskan kembali pengertian dari konsep DW, standard konsep arsitektur DW, mengetahui resiko implementasi DW dan penanganannya dan memahami proses ETL.
Definisi Data Warehouse, Tujuan Data Warehouse, Pendekatan Desain Data Warehouse, Komponen Data Warehouse, Arsitektur Data Warehouse, Skema Data Warehouse

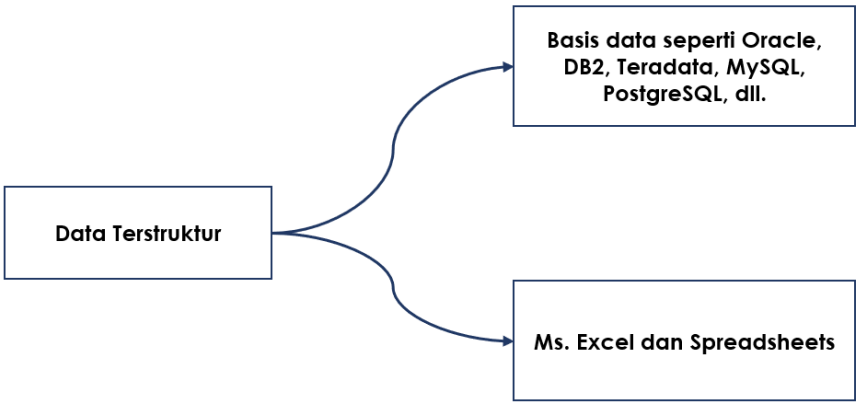
Definisi Data Warehouse

Data warehouse adalah suatu sistem penyimpanan data dari berbagai sumber yang biasanya diterapkan dalam sebuah perusahaan atau institusi tertentu sebagai bahan pertimbangan manajemen.

Data Warehouse merupakan kombinasi dari dua komponen utama: Basis data pendukung keputusan terintegrasi dan program perangkat lunak terkait, yang digunakan untuk mengumpulkan, membersihkan, mengubah, dan menyimpan data dari berbagai sumber operasional dan eksternal.

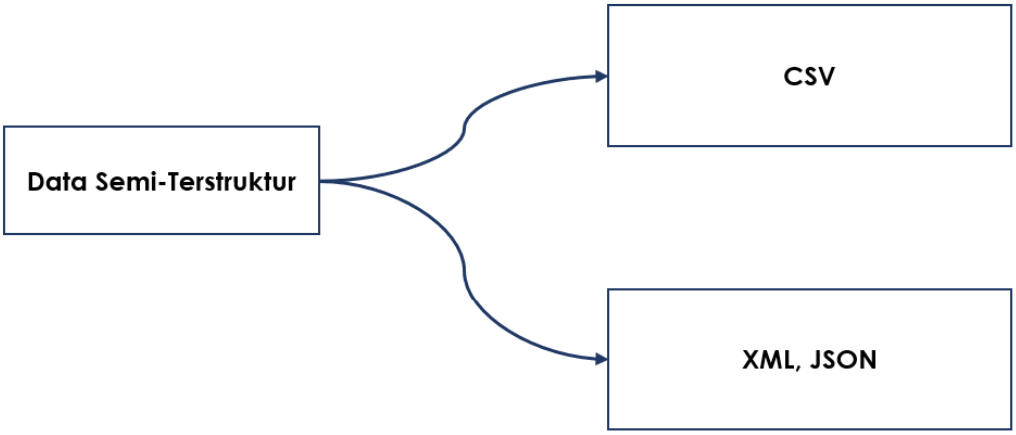
Data warehouse memiliki makna sebagai kegiatan mengumpulkan data dari berbagai sumber yang kemudian diolah menjadi informasi yang lebih mudah dibaca dan dipahami untuk kepentingan strategi bisnis. Data warehouse mencakup penyimpanan berbagai informasi dalam jumlah besar yang dirancang untuk query dan analisis dalam menjalankan bisnis.

Secara tradisional, data warehouse berfokus pada data terstruktur yaitu data yang mempunyai struktur tetap dan dapat disusun dalam baris dan kolom.



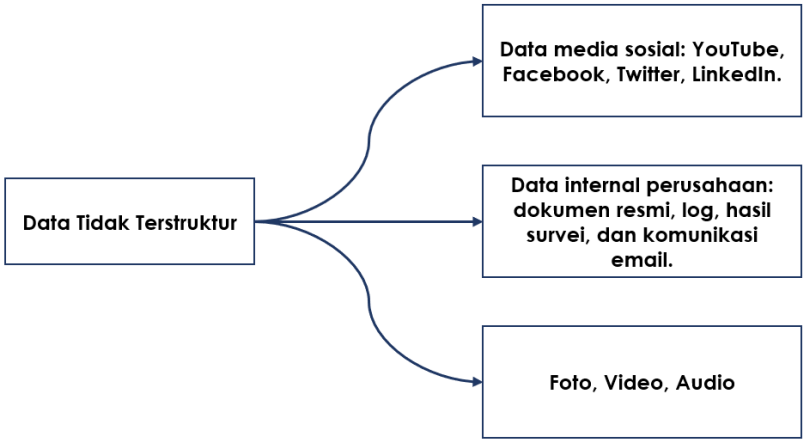
Gambar 6. Jenis Data Terstruktur

Selain itu, data warehouse juga mencakup data semi-terstruktur, yang didefinisikan sebagai elemen elektronik yang disusun sebagai entitas semantik tanpa keterhubungan atribut yang diperlukan. Data semi-terstruktur tidak berada dalam database relasional tetapi memiliki beberapa properti organisasi yang membuatnya lebih mudah untuk dianalisis.



Gambar 7. Jenis Data Semi-Terstruktur

Selanjutnya data warehouse mencakup data tidak terstruktur, yang mengacu pada data yang tidak ditentukan sebelumnya melalui model data.



Gambar 8. Jenis Data Tidak Terstruktur

Tujuan Data Warehouse

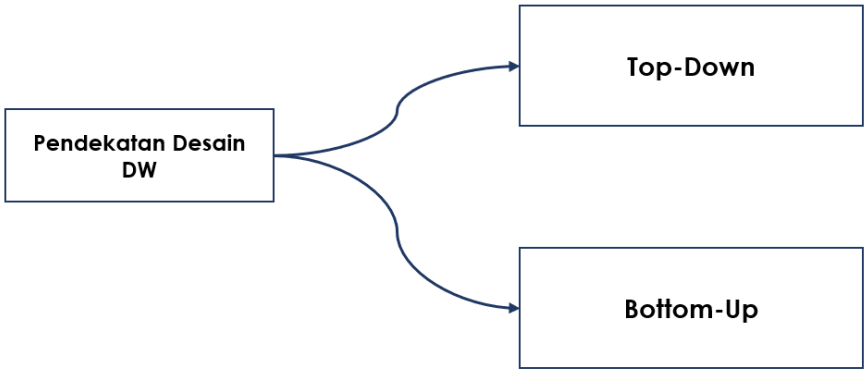
Data warehouse harus membuat informasi organisasi mudah diakses: Isi data warehouse harus dapat dimengerti. Data harus intuitif dan jelas bagi pengguna bisnis, bukan hanya pengembang. Tools yang mengakses data warehouse harus sederhana dan mudah digunakan, serta harus dapat mengembalikan hasil kueri kepada pengguna dengan waktu tunggu minimal.

Data warehouse harus menyajikan informasi organisasi secara konsisten: Data yang ada di gudang harus kredibel. Data harus dikumpulkan secara hati-hati dari berbagai sumber di seluruh organisasi dan terjamin kualitasnya, serta dirilis hanya jika data tersebut layak untuk dikonsumsi pengguna.

Data warehouse harus adaptif dan tahan terhadap perubahan: Data warehouse harus dirancang untuk menangani perubahan yang tidak dapat dihindari. Perubahan pada data warehouse harus dilakukan dengan baik, artinya perubahan tersebut tidak membuat data atau aplikasi yang ada menjadi tidak valid.

Data warehouse harus menjadi benteng aman yang melindungi aset informasi organisasi atau perusahaan. Data warehouse harus berfungsi sebagai landasan untuk pengambilan keputusan yang lebih baik: Data warehouse harus memiliki data yang tepat di dalamnya untuk mendukung pengambilan keputusan.

Pendekatan Desain Data Warehouse



Gambar 9. Pendekatan Desain Data Warehouse

1. Top-Down

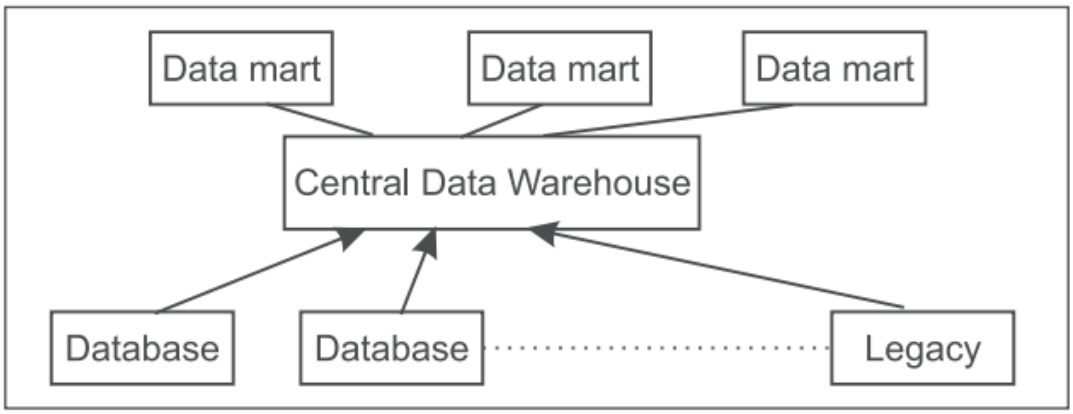
Persyaratan pengguna di tingkat organisasi yang berbeda digabungkan sebelum proses desain dimulai, dan satu skema untuk seluruh data warehouse dibangun. Kemudian, data mart terpisah disesuaikan dengan karakteristik masing-masing area bisnis atau proses. Top-down didesain secara inheren, bukan gabungan dari data mart yang berbeda. Serta aturan dan kontrol terpusat akan tetapi membutuhkan waktu lebih lama untuk membangun bahkan dengan metode berulang.

2. Bottom-Up

Skema terpisah dibuat untuk setiap data mart, dengan mempertimbangkan persyaratan pengguna pengambil keputusan yang bertanggung jawab atas area atau proses bisnis spesifik yang terkait. Nantinya, skema ini digabungkan dalam skema global untuk seluruh data warehouse. Secara inheren bersifat inkremental; dapat menjadwalkan data mart penting terlebih dahulu. Implementasi bagian yang dapat dikelola lebih cepat dan mudah akan tetapi dapat menyerap data yang berlebihan di setiap data mart.

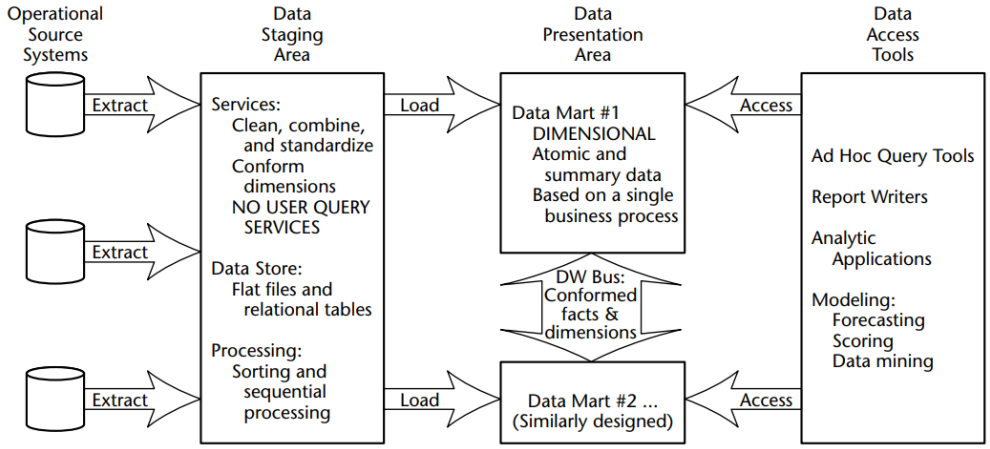
Data Mart

Data mart merupakan data warehouse lokal kecil yang dibangun untuk satu tujuan. Biasanya dibangun untuk memenuhi kebutuhan sekelompok pengguna atau departemen dalam suatu organisasi. Contoh: suatu organisasi dapat memiliki banyak departemen, termasuk keuangan, departemen TI, dan lain-lain. Masing-masing departemen ini dapat memiliki gudang datanya sendiri, yang tidak lain adalah data mart dari departemen tersebut.



Gambar 10. Data Mart

Komponen Data Warehouse



Gambar 11. Komponen Data Warehouse

1. Data Staging Area

Dalam area penyimpanan dan serangkaian proses yang biasa disebut sebagai extract-transformation-load (ETL). Persyaratan arsitektur utama untuk data staging area adalah area tersebut terlarang bagi pengguna bisnis dan tidak menyediakan query dan presentation services. Extraction adalah langkah pertama dalam proses memasukkan data ke dalam lingkungan data warehouse. Mengekstraksi berarti membaca dan memahami data sumber dan menyalin data yang diperlukan untuk data warehouse ke dalam staging area untuk manipulasi lebih lanjut.

Setelah data diekstraksi ke staging area, ada banyak potensi transformasi, seperti pembersihan data (mengoreksi kesalahan ejaan, menyelesaikan konflik domain, menangani elemen yang hilang, atau menguraikan ke dalam format standar), menggabungkan data dari berbagai sumber, menghapus duplikasi data, dan menugaskan warehouse keys.

Data staging area didominasi oleh aktivitas sederhana yaitu sorting dan pemrosesan sekuensial. Dalam banyak kasus, data staging area tidak hanya didasarkan pada teknologi relasional namun mungkin terdiri dari sistem flat file. Langkah terakhir dari proses ETL adalah memuat data. Pemuatan di lingkungan data warehouse biasanya berupa penyajian tabel dimensi dengan kualitas terjamin ke fasilitas pemuatan massal di setiap data mart.

2. Data Presentation

Area presentasi data adalah tempat data diorganisasikan, disimpan, dan disediakan untuk permintaan langsung oleh pengguna, penulis laporan, dan aplikasi analitis lainnya. Data mart adalah bagian dari keseluruhan area presentasi. Dalam bentuknya yang paling sederhana, data mart menyajikan data dari satu proses bisnis. Proses bisnis ini melintasi batas-batas fungsi organisasi.

Industri telah menyimpulkan bahwa pemodelan dimensi adalah teknik yang paling layak untuk mengirimkan data ke pengguna data warehouse. Pemodelan dimensi sangat berbeda dengan bentuk normal ketiga (pemodelan 3NF). Pemodelan 3NF adalah teknik desain yang berupaya menghilangkan redundansi data. Data dibagi menjadi banyak entitas diskrit, yang masing-masing menjadi tabel dalam database relasional. Perbedaan utama antara 3NF dan model dimensi adalah tingkat normalisasi.

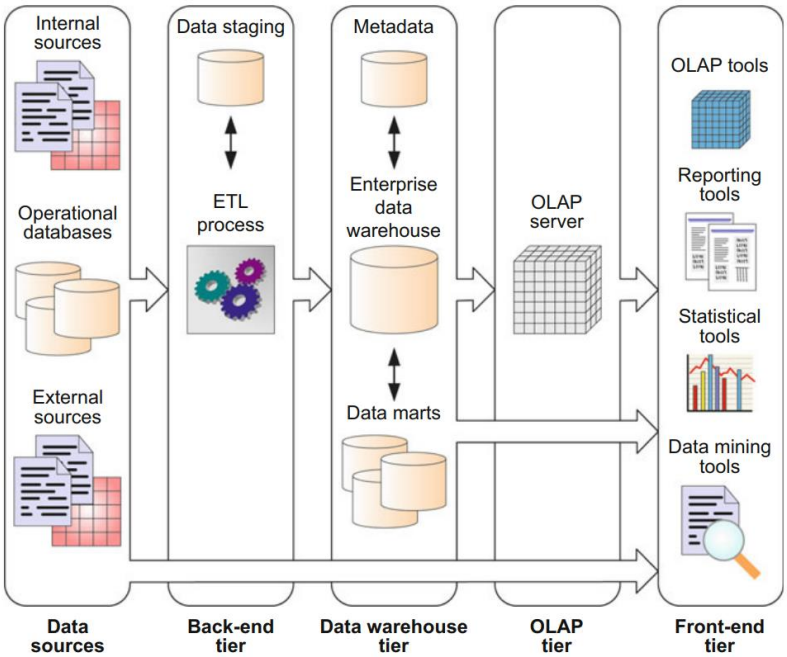
Pemodelan yang dinormalisasi sangat membantu kinerja pemrosesan operasional karena transaksi pembaruan atau penyisipan hanya perlu menyentuh database di satu tempat. Namun, model yang dinormalisasi terlalu rumit untuk kueri data warehouse. Model dimensi berisi informasi yang sama dengan model yang dinormalisasi tetapi mengemas data dalam format yang tujuan desainnya adalah pemahaman pengguna, performa kueri, dan ketahanan terhadap perubahan.

3. Data Access Tools

Komponen utama terakhir dari lingkungan data warehouse adalah data access tools. Komponen tersebut merujuk pada beragam kemampuan yang dapat diberikan kepada pengguna bisnis untuk memanfaatkan area presentasi untuk pengambilan keputusan analitik.

Data access tools bisa sesederhana alat kueri ad hoc atau serumit aplikasi penambangan data atau pemodelan. Contoh alat kueri ad hoc adalah mengeksplorasi data kemudian membuat visualisasi data interaktif dan menyajikan insight yang ditemukan secara langsung.

Arsitektur Data Warehouse



Gambar 12. Arsitektur Data Warehouse

1. Tingkat Back-end

Tingkat back-end terdiri dari alat ekstraksi, transformasi, dan pemuatan (ETL), yang digunakan untuk memasukkan data ke dalam data warehouse dari database operasional dan sumber data lainnya, yang dapat berasal dari internal atau eksternal organisasi, dan area data staging yang merupakan database perantara tempat semua proses integrasi dan transformasi data dijalankan sebelum data dimuat ke dalam data warehouse. Dalam tingkat back-end terdapat proses ETL yaitu:

a. Extraction, mengumpulkan data dari berbagai sumber data yang heterogen. Sumber-sumber ini mungkin berupa database operasional, namun bisa juga berupa file dalam berbagai format; dapat berasal dari internal organisasi atau eksternal.
b. Transformation, mengubah data dari format sumber data ke format gudang. Hal ini mencakup beberapa aspek: pembersihan, yang menghilangkan kesalahan dan ketidakkonsistenan dalam data dan mengubahnya menjadi format standar.
c. Loading, hal ini juga mencakup memuat data ke dalam data warehouse, yaitu menyebarkan pembaruan dari sumber data ke gudang data pada frekuensi tertentu untuk menyediakan data terkini untuk proses pengambilan keputusan.

Proses ETL biasanya memerlukan area data staging, yaitu database tempat data yang diekstraksi dari sumber mengalami modifikasi berturut-turut hingga akhirnya siap dimuat ke dalam data warehouse.

2. Tingkat Data Warehouse

Tingkat data warehouse terdiri dari enterprise data warehouse dan/atau beberapa data mart dan repositori metadata yang menyimpan informasi tentang data warehouse dan isinya. Dalam tingkat data warehouse terdapat komponen enterprise data warehouse, beberapa data mart, dan repositori metadata.

Enterprise data warehouse terpusat dan mencakup seluruh organisasi, sedangkan data mart adalah data warehouse khusus yang ditargetkan ke area fungsional atau departemen tertentu dalam suatu organisasi.

Metadata secara tradisional diklasifikasikan menjadi metadata teknis dan bisnis. Metadata bisnis menjelaskan arti (atau semantik) data dan aturan organisasi, kebijakan, dan batasan yang terkait dengan data. Sedangkan Metadata teknis menggambarkan bagaimana data disusun dan disimpan dalam sistem komputer serta aplikasi dan proses yang memanipulasi data tersebut.

Contoh penggunaan metadata antara lain, metadata digunakan untuk menyimpan konfigurasi dan definisi alur kerja (*workflows*) yang didefinisikan oleh pengguna. Ini mencakup informasi tentang langkah-langkah yang harus dijalankan, ketergantungan antara langkah-langkah, dan parameter yang diperlukan untuk menjalankan alur kerja.

Metadata dapat menyimpan informasi tentang status dan riwayat pekerjaan yang dijalankan. Ini mencakup waktu mulai dan selesai, status keberhasilan atau kegagalan, serta metrik dan log yang dihasilkan selama proses eksekusi.

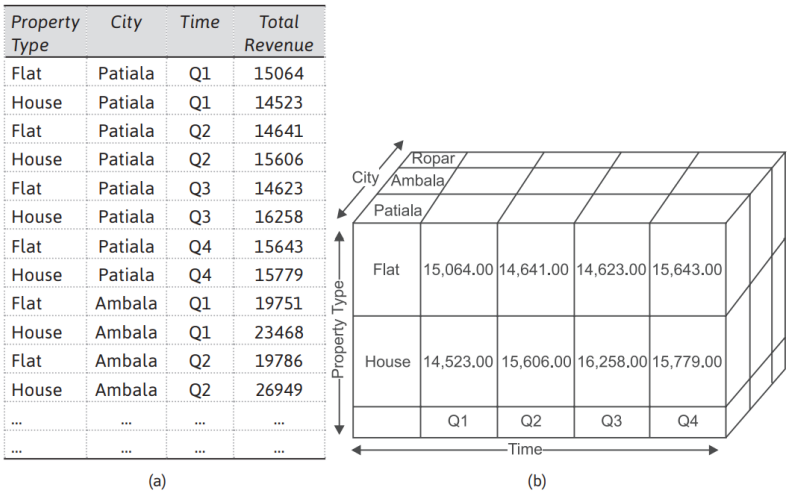
Metadata dapat digunakan untuk menyimpan log dan metrik yang dihasilkan selama eksekusi alur kerja. Ini termasuk log pesan, waktu eksekusi, jumlah rekaman diproses, dan metrik lainnya yang dapat digunakan untuk memantau kinerja dan performa alur kerja.

3. Tingkat OLAP

Tingkat OLAP terdiri dari server OLAP, yang menyediakan tampilan data multidimensi, terlepas dari cara sebenarnya data disimpan dalam sistem yang mendasarinya.

OLAP mengumpulkan informasi dari berbagai sistem dan memberikan informasi/pandangan yang diringkas kepada manajemen. OLAP menangani tampilan data multidimensi dibandingkan dengan tampilan database relasional sederhana. OLAP membantu pengguna untuk mendapatkan pengetahuan dan pemahaman yang lebih luas tentang berbagai fitur data perusahaan mereka melalui akses yang konsisten, cepat dan interaktif ke berbagai kemungkinan tampilan data yang komprehensif.

Terdapat beberapa bahasa yang digunakan dalam OLAP antara lain: XMLA (XML for Analysis) bertujuan menyediakan bahasa umum untuk pertukaran data multidimensi antara aplikasi klien dan server OLAP. Selanjutnya MDX (MultiDimensional eXpressions) adalah bahasa kueri untuk database OLAP yang menjadi standar de facto untuk sistem OLAP.



Property Type	City	Time	Total Revenue
Flat	Patiala	Q1	15064
House	Patiala	Q1	14523
Flat	Patiala	Q2	14641
House	Patiala	Q2	15606
Flat	Patiala	Q3	14623
House	Patiala	Q3	16258
Flat	Patiala	Q4	15643
House	Patiala	Q4	15779
Flat	Ambala	Q1	19751
House	Ambala	Q1	23468
Flat	Ambala	Q2	19786
House	Ambala	Q2	26949
...	...	...	...
...	...	...	...

Gambar 13. OLAP

4. Tingkat Front-End

Tingkat front-end digunakan untuk analisis dan visualisasi data. Ini berisi client tools seperti OLAP tools, reporting tools, statistical tools, dan data mining tools.

OLAP tools memungkinkan eksplorasi interaktif dan manipulasi data warehouse. Didalamnya memfasilitasi perumusan pertanyaan kompleks yang mungkin melibatkan data dalam jumlah besar. Kueri ini disebut kueri ad hoc, karena sistem tidak memiliki pengetahuan sebelumnya tentang kueri tersebut.

Reporting tools memungkinkan produksi, pengiriman, dan pengelolaan laporan, yang dapat berupa laporan berbasis kertas atau laporan interaktif berbasis web. Laporan menggunakan kueri yang telah ditentukan sebelumnya, yaitu kueri yang meminta informasi spesifik dalam format tertentu yang dilakukan secara rutin. Teknik pelaporan modern mencakup indikator kinerja utama dan dasbor.

Statistical tools digunakan untuk menganalisis dan memvisualisasikan data cube menggunakan metode statistik.

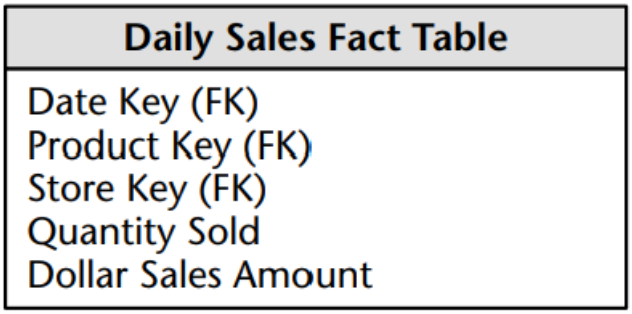
Data mining tools memungkinkan pengguna menganalisis data untuk menemukan pengetahuan berharga seperti pola dan tren; serta memungkinkan prediksi dibuat berdasarkan data terkini.

Skema Desain Data Warehouse

Fact Table / Tabel Fakta

Tabel fakta adalah tabel utama dalam model dimensi tempat pengukuran kinerja numerik bisnis disimpan. Sebuah baris dalam tabel fakta berhubungan dengan suatu pengukuran. Pengukuran adalah baris dalam tabel fakta. Semua pengukuran dalam tabel fakta harus berada pada grain yang sama.

Fakta yang paling berguna dalam tabel fakta adalah fakta numerik dan penjumlahan. Tabel fakta biasanya terdiri dari dua jenis kolom seperti kunci asing dan ukuran. Kunci asing dihubungkan dengan tabel dimensi dan ukuran terdiri dari fakta numerik. Tabel fakta itu sendiri umumnya memiliki kunci utama sendiri yang terdiri dari subset kunci asing. Kunci ini sering disebut kunci komposit atau gabungan. Tabel fakta mengungkapkan hubungan many-to-many antar dimensi dalam model dimensi. Berikut adalah contoh tabel fakta untuk penjualan barang di perusahaan ritel:



Gambar 14. Contoh Tabel Fakta

Dimension Tables / Tabel Dimensi

Tabel dimensi merupakan pendamping integral dari tabel fakta. Tabel dimensi berisi deskripsi tekstual bisnis. Atau kumpulan informasi referensi tentang suatu peristiwa yang dapat diukur. Peristiwa ini disimpan dalam tabel fakta dan dikenal sebagai fakta. Dalam model dimensi yang dirancang dengan baik, tabel dimensi memiliki banyak kolom atau atribut. Atribut ini mendeskripsikan baris dalam tabel dimensi.

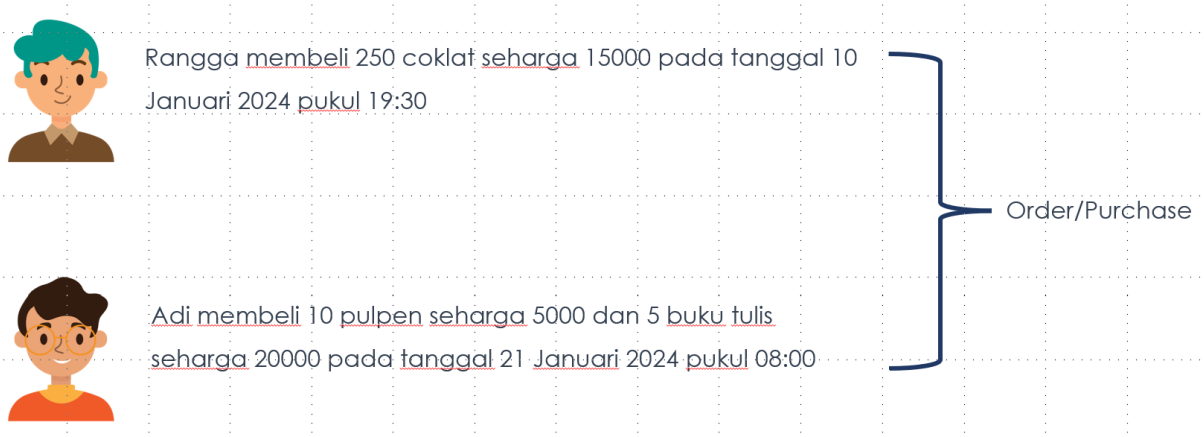
Atribut dimensi berfungsi sebagai sumber utama batasan kueri, pengelompokan, dan label laporan. Misalnya, ketika pengguna menyatakan bahwa dia ingin melihat penjualan dolar per minggu berdasarkan merek, minggu dan merek harus tersedia sebagai atribut dimensi. Tabel dimensi adalah titik masuk ke dalam tabel fakta. Atribut dimensi yang kuat memberikan kemampuan analytic slicing and dicing yang kuat. Dimensi tersebut mengimplementasikan antarmuka pengguna ke gudang data.

Atribut terbaik adalah tekstual dan diskrit. Atribut harus terdiri dari kata-kata nyata dan bukan singkatan yang samar-samar. Contoh atribut untuk dimensi produk mencakup deskripsi singkat (10 hingga 15 karakter), deskripsi panjang (30 hingga 50 karakter) seperti nama merek, nama kategori, jenis kemasan, ukuran, dan berbagai karakteristik produk lainnya. Berikut adalah contoh tabel dimensi untuk penjualan barang di perusahaan ritel:

Product Dimension Table
Product Key (PK)
Product Description
SKU Number (Natural Key)
Brand Description
Category Description
Department Description
Package Type Description
Package Size
Fat Content Description
Diet Type Description
Weight
Weight Units of Measure
Storage Type
Shelf Life Type
Shelf Width
Shelf Height
Shelf Depth
... and many more

Gambar 15. Contoh Tabel Dimensi

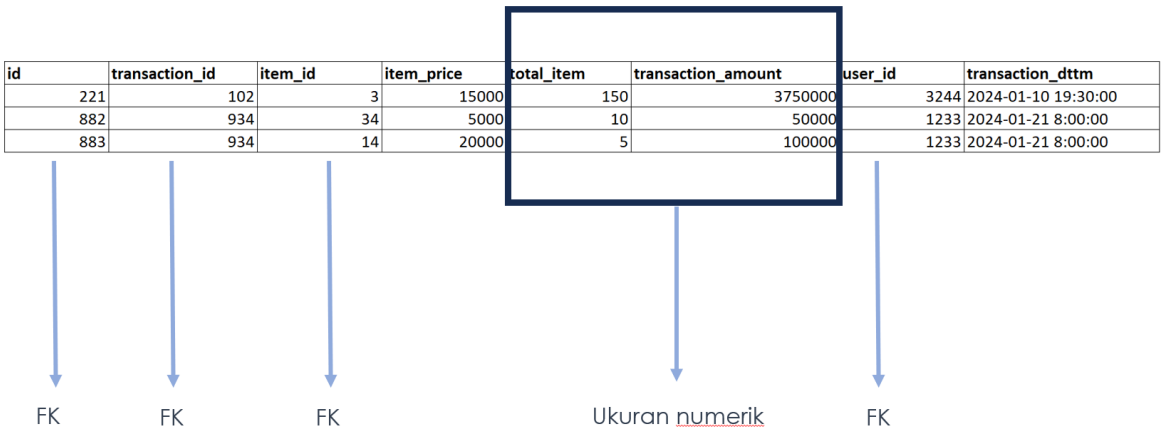
Contoh bisnis proses untuk tabel fakta dan dimensi



Tabel Purchase

id	transaction_id	item_id	item_price	total_item	transaction_amount	user_id	transaction_dttm
221	102	3	15000	150	3750000	3244	2024-01-10 19:30:00
882	934	34	5000	10	50000	1233	2024-01-21 8:00:00
883	934	14	20000	5	100000	1233	2024-01-21 8:00:00

Tabel Fakta



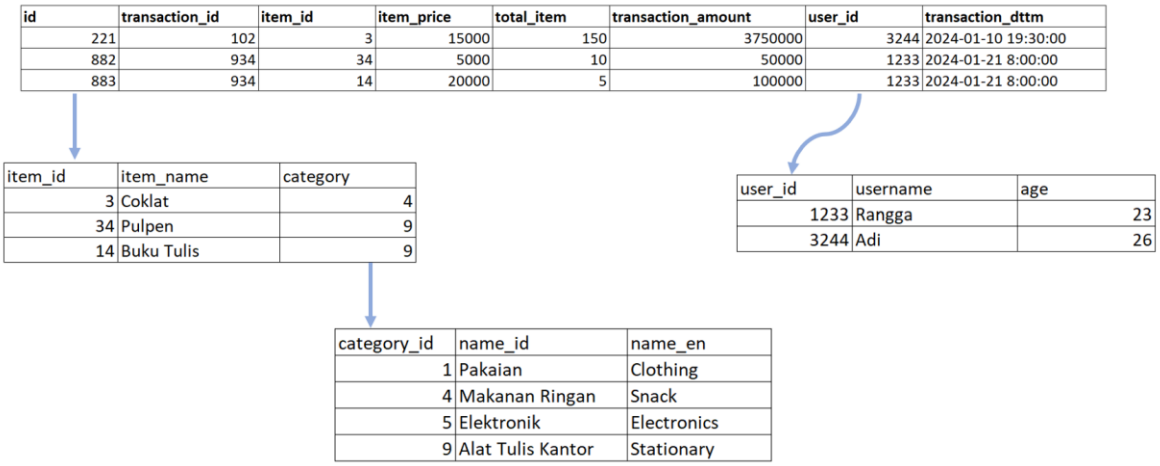
Tabel Dimensi

item_id	item_name	category
3	Coklat	4
34	Pulpen	9
14	Buku Tulis	9

user_id	username	age
1233	Rangga	23
3244	Adi	26

category_id	name_id	name_en
1	Pakaian	Clothing
4	Makanan Ringan	Snack
5	Elektronik	Electronics
9	Alat Tulis Kantor	Stationary

Tabel Fakta dan Dimensi



id	transaction_id	item_id	item_price	total_item	transaction_amount	user_id	transaction_dttm
221	102	3	15000	150	3750000	3244	2024-01-10 19:30:00
882	934	34	5000	10	50000	1233	2024-01-21 8:00:00
883	934	14	20000	5	100000	1233	2024-01-21 8:00:00

item_id	item_name	category
3	Coklat	4
34	Pulpen	9
14	Buku Tulis	9

user_id	username	age
1233	Rangga	23
3244	Adi	26

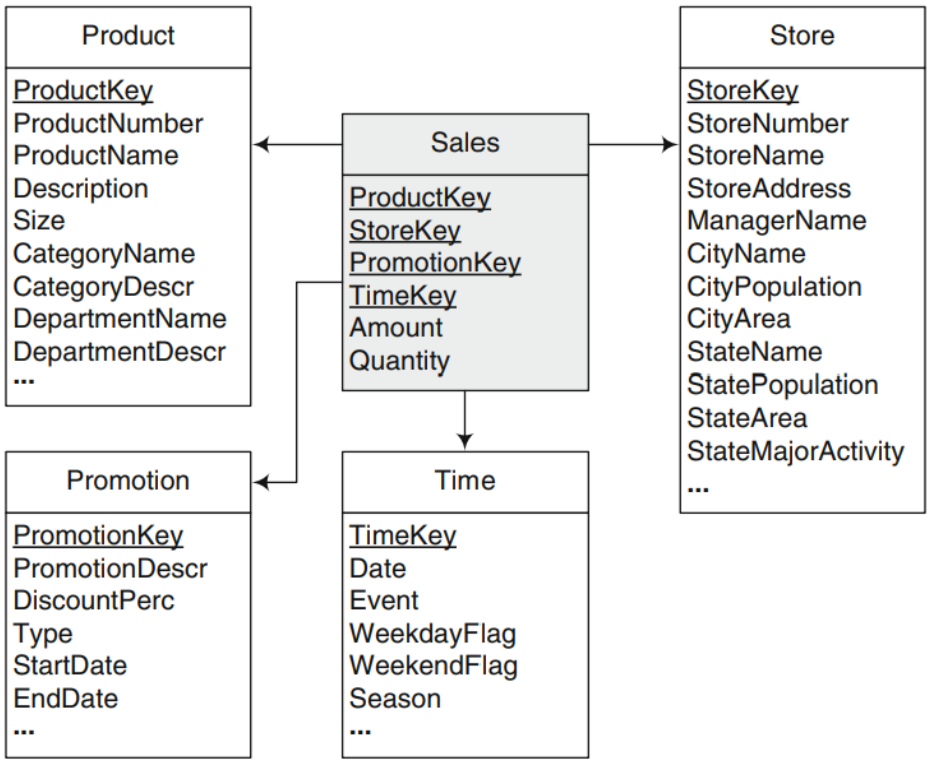
category_id	name_id	name_en
1	Pakaian	Clothing
4	Makanan Ringan	Snack
5	Elektronik	Electronics
9	Alat Tulis Kantor	Stationary

Star Schema

Star Schema adalah salah satu skema gudang data yang paling sederhana. Disebut bintang karena tampak seperti bintang dengan titik-titik yang memanjang dari pusatnya. Setiap dimensi dalam skema bintang mewakili tabel satu dimensi saja dan tabel dimensi terdiri dari sekumpulan atribut.

Dalam star schema, tabel dimensi secara umum tidak dinormalisasi, sehingga integritas data tidak dapat diterapkan dalam skema ini karena kemungkinan berisi data yang berlebihan. Dalam star schema memiliki kinerja kueri yang tinggi karena memerlukan lebih sedikit operasi gabungan karena de-normalisasi data dan memiliki struktur sederhana yang mudah dimengerti.

Berikut adalah contoh tabel dimensi untuk penjualan barang di perusahaan ritel, di mana tabel fakta digambarkan dengan warna abu-abu dan tabel dimensi digambarkan dengan warna putih.



Gambar 16. Star Schema

Snowflake Schema

Perbedaan utama antara star schema dan snowflake schema adalah snowflake schema dapat terdiri dari dimensi yang dinormalisasi sedangkan star schema selalu terdiri dari dimensi yang denormalisasi.

Snowflake schema merupakan modifikasi dari star schema yang mendukung normalisasi tabel dimensi. Beberapa tabel dimensi dinormalisasi dalam ssnowflake schema yang membagi data menjadi tabel tambahan.

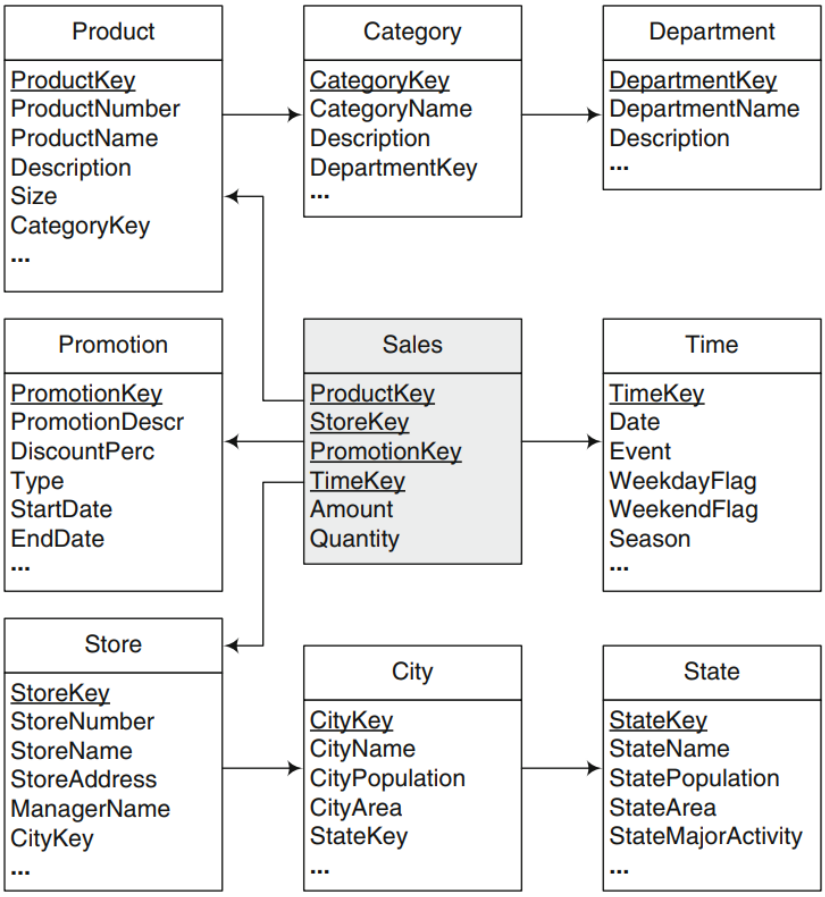
Snowflake schema menghasilkan penghematan ruang penyimpanan karena atribut yang dinormalisasi meskipun terdapat kompleksitas tambahan dalam penggabungan kueri sumber. Tabel yang dinormalisasi mudah dirawat dan mengoptimalkan ruang penyimpanan. Namun, kinerja terpengaruh karena lebih banyak penggabungan yang perlu dilakukan saat menjalankan kueri yang memerlukan hierarki untuk dilintasi.

Contoh knowledge workers akan membuat query untuk melihat total penjualan berdasarkan kategori:

```
SELECT CategoryName, SUM(Amount)
FROM Product P, Sales S
WHERE P.ProductKey = S.ProductKey
GROUP BY CategoryName
```

```
SELECT CategoryName, SUM(Amount)
FROM Product P, Category C, Sales S
WHERE P.ProductKey = S.ProductKey AND P.CategoryKey = C.CategoryKey
GROUP BY CategoryName
```

Berikut adalah contoh tabel dimensi dan tabel fakta dengan snowflake schema untuk penjualan barang di perusahaan ritel:



Gambar 17. Snowflake Schema

MODUL 3: Business Performance Management

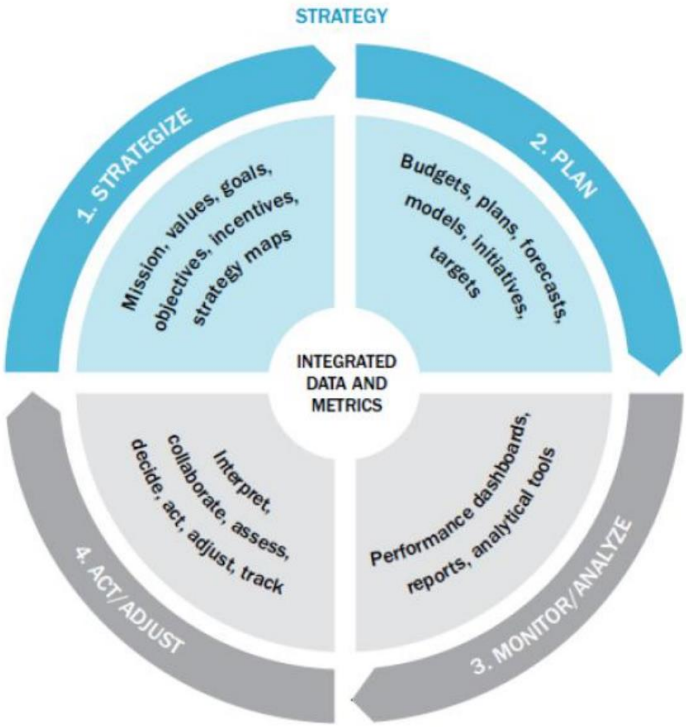
Kompetensi: Mahasiswa mampu menjelaskan kembali pengertian BPM, mampu memahami proses Closed-loop untuk mengoptimalkan kinerja bisnis dan memahami konsep performance measurement.
Definisi Business Performance Management, Proses Closed-Loop untuk Mengoptimalkan Kinerja Bisnis, Performance Measurement.

Definisi Business Performance Management

Business Performance Management (BPM) adalah Sistem real-time yang mengingatkan manajer akan peluang potensial, masalah yang akan datang, dan ancaman, dan kemudian memberdayakan mereka untuk bereaksi melalui model dan kolaborasi. BPM mengacu pada proses bisnis, metodologi, metrik, dan teknologi yang digunakan oleh perusahaan untuk mengukur, memantau, dan mengelola kinerja bisnis. BPM adalah hasil dari BI dan menggabungkan banyak teknologi, aplikasi, dan tekniknya. BPM mencakup tiga komponen Utama:

1. Seperangkat terintegrasi, manajemen loop tertutup dan proses analitik, didukung oleh teknologi.
2. Alat untuk bisnis untuk menentukan tujuan strategis dan kemudian mengukur / mengelola kinerja terhadap mereka.
3. Metode dan alat untuk memantau indikator kinerja utama (Key Performance Indicators), terkait dengan strategi organisasi.

Proses Closed-Loop untuk Mengoptimalkan Kinerja Bisnis



Gambar 18. Proses Closed-Loop

1. Strategize

Dalam perencanaan Strategis (Strategic planning) yang umum dilakukan antara lain melakukan analisis situasi saat ini, menentukan cakrawala perencanaan, melakukan pemindaian lingkungan, mengidentifikasi factor penentu keberhasilan, melengkapi analisis kesenjangan, membuat visi strategis, mengembangkan strategi bisnis, dan mengidentifikasi sasaran dan sasaran strategis.

2. Plan

Perencanaan dibagi menjadi 2 yaitu perencanaan operasional dan perencanaan penganggaran keuangan.

a. Perencanaan operasional: rencana yang menerjemahkan sasaran dan sasaran strategis organisasi ke dalam serangkaian taktik dan inisiatif yang ditetapkan dengan baik, persyaratan sumber daya, dan hasil yang diharapkan untuk beberapa periode waktu mendatang (biasanya satu tahun).
b. Perencanaan penganggaran keuangan: Tujuan strategis dan metrik Utama organisasi harus berfungsi sebagai pendorong top-down untuk alokasi aset berwujud dan tidak berwujud organisasi. Serta Alokasi sumber daya harus diselaraskan dengan hati-hati dengan tujuan dan taktik strategis organisasi untuk mencapai keberhasilan strategis.

3. Monitor

Kerangka kerja yang komprehensif untuk memantau kinerja harus mengatasi dua masalah utama: apa yang harus dipantau dan cara memonitor. Melakukan monitoring dapat dilakukan dengan sistem Kontrol Diagnostik yaitu sistem cybernetic yang memiliki input, proses untuk mentransformasikan input menjadi output, standar atau tolok ukur untuk membandingkan output, dan saluran umpan balik untuk memungkinkan informasi tentang perbedaan antara output dan standar untuk dikomunikasikan dan ditindaklanjuti.

4. Act and Adjust

Keberhasilan (atau kelangsungan hidup semata) bergantung pada proyek-proyek baru: menciptakan produk baru, memasuki pasar baru, mendapatkan pelanggan baru (atau bisnis), atau merampingkan beberapa proses.

Performance Measurement

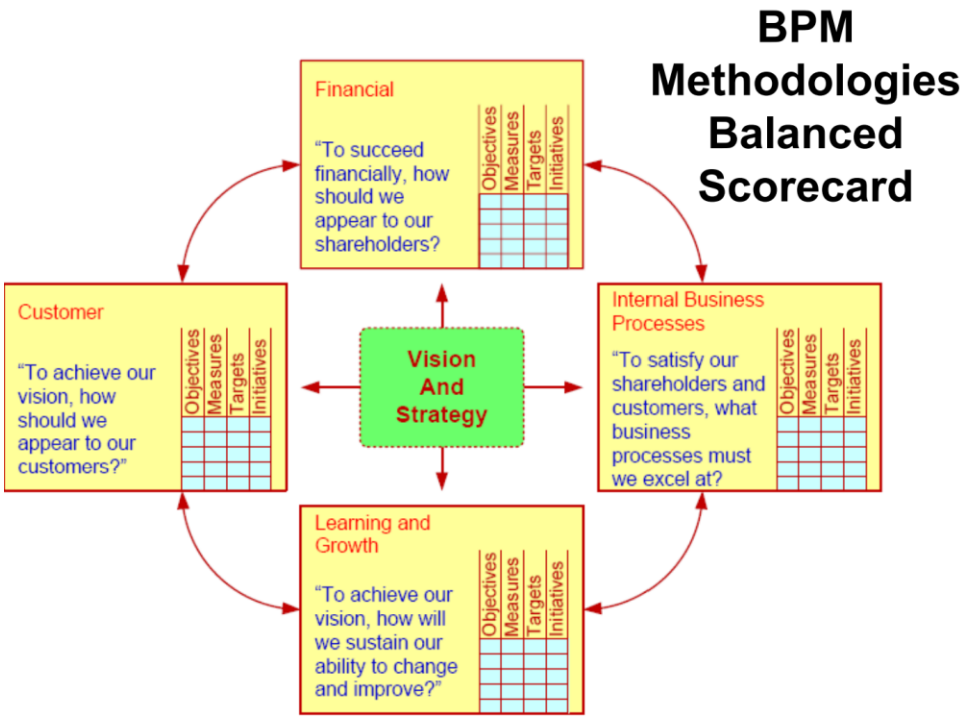
Sistem yang paling populer digunakan adalah beberapa varian balanced scorecard (BSC). Metodologi BSC adalah visi holistik dari sistem pengukuran yang terkait dengan arah strategis organisasi dan didasarkan pada pandangan empat perspektif dunia yaitu Ukuran finansial didukung oleh metrik pelanggan (customer), internal, serta pembelajaran (learning) dan pertumbuhan (growth).

MODUL 4: BPM Methodologies

Kompetensi: Mahasiswa mampu memahami tentang BPM Methodologies, Arsitektur BPM & Aplikasi dan performance dashbords

Balanced scorecard (BSC)

Metodologi pengukuran dan manajemen kinerja yang membantu menerjemahkan keuangan, pelanggan, proses internal, dan tujuan serta sasaran pembelajaran dan pertumbuhan ke dalam serangkaian inisiatif yang dapat ditindaklanjuti.



Gambar 19. BPM Methodologies

BSC dirancang untuk mengatasi keterbatasan sistem yang berfokus secara finansial. Tujuan non finansial terbagi dalam salah satu dari tiga perspektif yaitu Pelanggan (Customer), Proses bisnis internal (Internal business process), dan Pembelajaran & Pertumbuhan (Learning and growth).

Peta Strategi

Tampilan visual yang menggambarkan hubungan antara tujuan organisasi utama untuk keempat perspektif BSC.

Six Sigma

Metodologi manajemen kinerja yang bertujuan mengurangi jumlah cacat dalam proses bisnis sedekat mungkin dengan sebisamungkin nol cacat per juta peluang/defects per million opportunities (DPMO). Dalam six sigma terdapat Model kinerja DMAIC yaitu Model peningkatan bisnis loop tertutup yang mencakup langkahlangkah mendefinisikan, mengukur, menganalisis, meningkatkan, dan mengendalikan suatu proses.

MODUL 5: Data Mining

Kompetensi: Mahasiswa mampu memahami tentang konsep data mining

Definisi Data Mining

Data mining adalah studi tentang pengumpulan, pembersihan, pemrosesan, analisis, dan perolehan wawasan berguna dari data. Hampir semua sistem otomatis yang digunakan saat ini menghasilkan data dalam beberapa jenis, baik untuk tujuan analitis atau pemecahan masalah. Hasilnya, gelombang data sebesar petabyte atau exabyte dihasilkan. Beberapa contoh berbagai jenis data adalah sebagai berikut:

1. World Wide Web: Jumlah dokumen di Web yang diindeks kini berjumlah miliaran. Berbagai jenis data ini berguna dalam berbagai aplikasi. Misalnya, dokumen Web dan struktur tautan dapat ditambang untuk menentukan hubungan antara berbagai topik di Web. Di sisi lain, log akses pengguna dapat ditambang untuk menentukan pola akses yang sering atau pola yang tidak biasa dari perilaku yang mungkin tidak beralasan.
2. Interaksi keuangan: Sebagian besar transaksi umum dalam kehidupan sehari-hari, seperti menggunakan kartu anjungan tunai mandiri (ATM) atau kartu kredit, dapat membuat data dengan cara otomatis. Transaksi semacam itu dapat ditambang untuk mendapatkan banyak wawasan berguna seperti penipuan atau aktivitas tidak biasa lainnya.
3. Interaksi pengguna: Berbagai bentuk interaksi pengguna menghasilkan data dalam jumlah besar. Misalnya, penggunaan telepon biasanya membuat catatan di perusahaan telekomunikasi dengan rincian tentang durasi dan tujuan panggilan. Banyak perusahaan telepon yang secara rutin menganalisis data tersebut untuk menentukan pola perilaku relevan yang dapat digunakan untuk mengambil keputusan mengenai kapasitas jaringan, promosi, harga, atau penargetan pelanggan.
4. Teknologi sensor dan Internet of Things: Tren terkini adalah pengembangan sensor portabel berbiaya rendah, smartphone, dan perangkat pintar lainnya yang dapat berkomunikasi satu sama lain.

Data mentah mungkin berubah-ubah, tidak terstruktur, atau bahkan dalam format yang tidak sesuai untuk pemrosesan otomatis. Misalnya, data yang dikumpulkan secara manual mungkin diambil dari berbagai sumber dalam format berbeda, namun perlu diproses oleh program komputer otomatis untuk mendapatkan wawasan.

Untuk mengatasi masalah ini, analis data mining menggunakan jalur pemrosesan, di mana data mentah dikumpulkan, dibersihkan, dan diubah ke dalam format standar. Data dapat disimpan dalam sistem database dan akhirnya diproses untuk mendapatkan wawasan dengan menggunakan metode analitis. Teknik data mining digunakan untuk menjelajahi database besar guna menemukan pola baru dan berguna yang mungkin masih belum diketahui.

Definisi Data

Kumpulan data terdiri dari objek data. Objek data mewakili suatu entitas—dalam database penjualan, objeknya bisa berupa pelanggan, item toko, dan penjualan; dalam database medis, objeknya mungkin pasien; dalam database universitas, objeknya bisa berupa mahasiswa dan mata kuliah.

Objek data biasanya dideskripsikan berdasarkan atribut. Objek data juga bisa disebut sebagai sampel, contoh, titik data, atau objek. Jika objek data disimpan dalam database, maka

itu adalah tupel data. Artinya, baris-baris database berhubungan dengan objek data, dan kolom-kolom berhubungan dengan atribut.

Atribut

Atribut adalah suatu sifat atau ciri suatu benda yang dapat berbeda-beda, baik dari satu benda ke benda lainnya, atau dari waktu ke waktu. Contoh: warna mata berbeda-beda pada setiap orang, sedangkan suhu suatu benda berbeda-beda seiring waktu. Atau atribut adalah bidang data, yang mewakili karakteristik atau fitur objek data. Kata benda atribut, dimensi, fitur, dan variabel sering digunakan secara bergantian dalam literatur.

Istilah dimensi umumnya digunakan dalam data warehouse. Literatur machine learning cenderung menggunakan istilah fitur, sedangkan ahli statistik lebih memilih istilah variabel. Profesional data mining dan basis data biasanya menggunakan istilah atribut. Atribut yang mendeskripsikan objek pelanggan dapat mencakup, misalnya, ID pelanggan, nama, dan alamat. Nilai yang diamati untuk atribut tertentu dikenal sebagai observasi.

1. Tipe Atribut – Nominal

Nilai suatu atribut nominal merupakan simbol atau nama suatu benda. Setiap nilai mewakili beberapa jenis kategori, kode, atau keadaan, sehingga atribut nominal juga disebut sebagai kategorikal. akan tetapi nilai-nilai tersebut tidak memiliki urutan yang berarti.

Contoh Atribut nominal. Misalkan warna rambut dan status perkawinan adalah dua atribut yang menggambarkan objek seseorang. Dalam aplikasi yang dimiliki organisasi, kemungkinan nilai warna rambut adalah hitam, coklat, pirang, merah, pirang kemerahan, abu-abu, dan putih. Atribut status perkawinan dapat mempunyai arti lajang, menikah, dan cerai. Contoh lain dari atribut nominal adalah pekerjaan, dengan nilai guru, dokter gigi, programmer, petani, dan sebagainya.

2. Tipe Atribut - Biner

Atribut biner adalah atribut nominal yang hanya memiliki dua kategori atau status: 0 atau 1, dengan 0 biasanya berarti atribut tersebut tidak ada, dan 1 berarti atribut tersebut ada. Atribut biner disebut Boolean jika kedua keadaan berhubungan dengan benar dan salah.

Contoh Atribut biner. Mengingat atribut perokok menggambarkan objek pasien, 1 menunjukkan pasien merokok, sedangkan 0 menunjukkan pasien tidak merokok. Demikian pula, contoh lain pasien menjalani tes medis yang memiliki dua kemungkinan hasil. Atribut tes kesehatan bersifat biner, dimana nilai 1 berarti hasil tes pasien positif, sedangkan 0 berarti hasilnya negatif.

3. Tipe Atribut - Ordinal

Atribut ordinal adalah atribut dengan kemungkinan nilai yang mempunyai urutan atau peringkat yang berarti di antara nilai-nilai tersebut, namun besaran antara nilai-nilai yang berurutan tidak diketahui. Nilai atribut ordinal memberikan informasi yang cukup untuk mengurutkan objek.

Atribut ordinal berguna untuk mencatat penilaian subjektif terhadap kualitas yang tidak dapat diukur secara objektif; oleh karena itu atribut ordinal sering digunakan dalam survei untuk pemeringkatan. Dalam sebuah survei, peserta diminta menilai seberapa puas mereka sebagai pelanggan. Kepuasan pelanggan memiliki kategori ordinal sebagai berikut: 0: sangat tidak puas, 1: agak tidak puas, 2: netral, 3: puas, dan 4: sangat puas.

Perhatikan bahwa atribut nominal, biner, dan ordinal bersifat kualitatif. Artinya, mereka mendeskripsikan fitur suatu objek tanpa memberikan ukuran atau kuantitas sebenernya.

4. Tipe Atribut - Interval

Atribut berskala interval diukur pada skala satuan berukuran sama. Nilai atribut berskala interval mempunyai urutan dan bisa positif, 0, atau negatif. Selain memberikan peringkat nilai, atribut tersebut memungkinkan untuk membandingkan dan mengukur perbedaan antar nilai.

Atribut interval tidak memiliki titik nol mutlak. Nilai nol pada atribut interval menunjukkan posisi nol pada skala yang digunakan, bukan nol absolut. Misalnya, suhu dalam Celsius memiliki 0◦C, tetapi ini tidak berarti tidak ada suhu sama sekali.

Perbandingan antara dua nilai tidak memiliki arti yang bermakna dalam atribut interval. Misalnya, perbandingan antara suhu 20°C dan 10°C tidak memiliki arti yang bermakna dalam hal besarnya.

5. Tipe Atribut - Rasio

Atribut berskala rasio adalah atribut numerik yang memiliki titik nol bawaan. Artinya, jika suatu pengukuran berskala rasio, kita dapat menyebut suatu nilai sebagai kelipatan (atau rasio) dari nilai lain. Selain itu, nilai-nilainya diurutkan, dan kita juga dapat menghitung selisih antar nilai, serta mean, median, dan modus.

Atribut rasio memiliki titik nol mutlak, yang berarti nilai nol pada atribut tersebut menunjukkan tidak adanya atau kekosongan dari karakteristik yang diukur. Misalnya, berat badan nol berarti tidak ada berat badan sama sekali.

Perbandingan antara dua nilai memiliki arti yang bermakna dalam atribut rasio. Misalnya, jika seseorang memiliki dua kali lipat berat badan dari yang lain, berat badannya juga dua kali lipat.

Mendeskripsikan Atribut Berdasarkan Banyaknya Nilai

Atribut diskrit memiliki himpunan nilai yang berhingga atau tak terhingga, yang dapat direpresentasikan sebagai bilangan bulat atau tidak. Atribut warna rambut, perokok, tes kesehatan, dan ukuran minuman masing-masing memiliki jumlah nilai yang terbatas, sehingga bersifat diskrit.

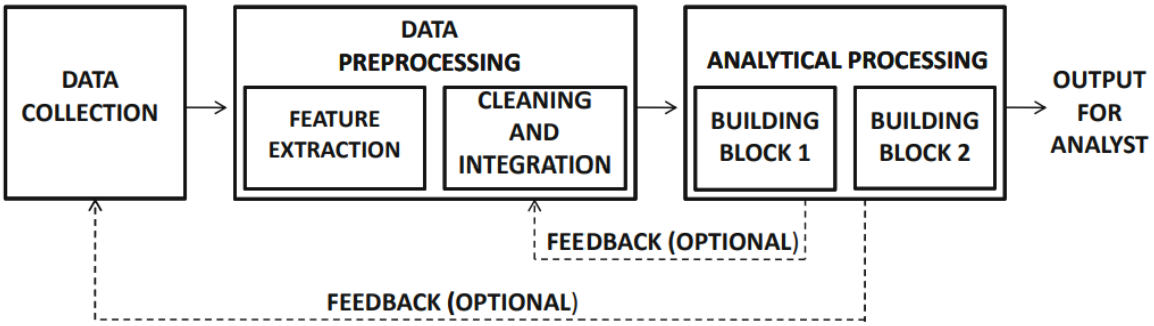
Suatu atribut dikatakan tak terhingga jika himpunan nilai-nilai yang mungkin tak terhingga, namun nilai-nilai tersebut dapat dimasukkan ke dalam korespondensi satu-satu dengan bilangan asli. Misalnya, atribut ID pelanggan tidak terbatas jumlahnya.

Atribut tersebut dapat bersifat kategorikal, seperti kode pos atau nomor ID, atau numerik, seperti jumlah. Atribut diskrit sering kali direpresentasikan menggunakan variabel integer.

Atribut biner adalah kasus khusus dari atribut diskrit dan hanya mengasumsikan dua nilai, misalnya benar/salah, ya/tidak, laki-laki/perempuan, atau 0/1. Atribut biner sering direpresentasikan sebagai variabel Boolean, atau sebagai variabel integer yang hanya bernilai 0 atau 1.

Atribut kontinu adalah atribut yang nilainya berupa bilangan real. Contohnya mencakup atribut seperti suhu, tinggi badan, atau berat badan. Atribut kontinu biasanya direpresentasikan sebagai variabel floating-point. Praktisnya, nilai riil hanya dapat diukur dan direpresentasikan dengan presisi terbatas.

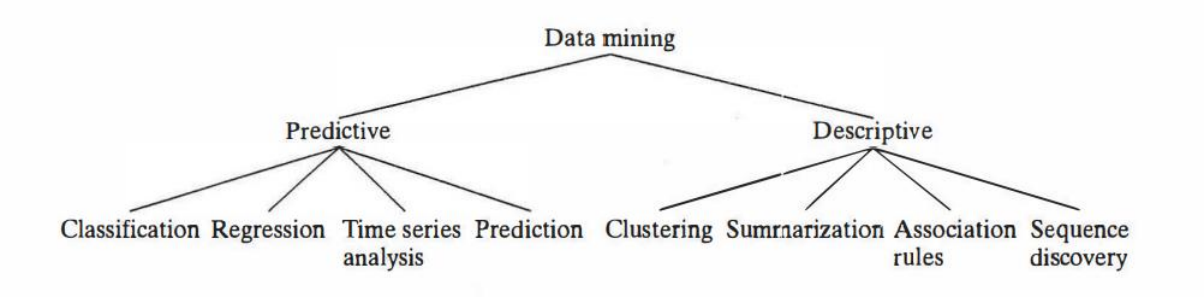
Kegiatan Belajar 3: Proses Data Mining



Gambar 20. Proses Data Mining

1. **Data Collection / Pengumpulan data:** Pengumpulan data mungkin memerlukan penggunaan perangkat keras khusus seperti jaringan sensor, pekerjaan manual seperti pengumpulan survei pengguna, atau perangkat lunak seperti mesin perayapan dokumen Web untuk mengumpulkan dokumen. Meskipun tahap ini sangat spesifik untuk aplikasi dan seringkali berada di luar jangkauan analis data mining, tahap ini sangat penting karena pilihan yang baik pada tahap ini dapat berdampak signifikan pada proses data mining. Setelah tahap pengumpulan, data sering kali disimpan dalam database, atau, lebih umum, gudang data untuk diproses.
2. **Ekstraksi fitur dan pembersihan data:** Saat data dikumpulkan, seringkali data tersebut tidak dalam bentuk yang sesuai untuk diproses. Misalnya, data mungkin dikodekan dalam log kompleks atau dokumen bentuk bebas. Dalam banyak kasus, berbagai jenis data dapat dicampur secara sewenang-wenang dalam dokumen berformat bebas. Agar data cocok untuk diproses, penting untuk mengubahnya menjadi format yang terpadu terhadap algoritma penambangan data, seperti format multidimensi, deret waktu, atau semi terstruktur. Fase ekstraksi fitur sering kali dilakukan bersamaan dengan pembersihan data, di mana bagian data yang hilang dan salah diperkirakan atau diperbaiki. Hasil akhir dari prosedur ini adalah kumpulan data yang terstruktur dengan baik, yang dapat digunakan secara efektif oleh program komputer. Setelah tahap ekstraksi fitur, data dapat disimpan kembali dalam database untuk diproses.
3. **Pemrosesan analitis dan algoritma:** Bagian terakhir dari proses penambangan adalah merancang metode analisis yang efektif dari data yang diproses. blok analitik pada gambar diatas menunjukkan beberapa blok penyusun yang mewakili desain solusi untuk aplikasi tertentu. Bagian dari desain algoritmik ini bergantung pada keterampilan analis dan sering kali menggunakan satu atau lebih dari empat masalah utama sebagai landasan.

Data Mining Tasks



Gambar 21. Data Mining Tasks

Berdasarkan gambar diatas model prediktif adalah untuk memprediksi nilai suatu atribut tertentu berdasarkan nilai atribut lainnya. Atribut yang akan diprediksi biasa disebut dengan variabel sasaran atau variabel terikat, sedangkan atribut yang digunakan untuk melakukan prediksi disebut dengan variabel penjelas atau variabel bebas.

Model prediktif membuat prediksi tentang nilai data menggunakan hasil yang diketahui dari data berbeda. Pemodelan prediktif dapat dibuat berdasarkan penggunaan data historis lainnya. Misalnya, penggunaan kartu kredit mungkin ditolak bukan karena riwayat kredit pengguna, namun karena pembelian saat ini serupa dengan pembelian sebelumnya yang kemudian diketahui dilakukan dengan kartu curian.

Model deskriptif mengidentifikasi pola atau hubungan dalam data. Model deskriptif. Di sini, tujuannya adalah untuk mendapatkan pola (korelasi, tren, cluster, lintasan, dan anomali) yang merangkum hubungan mendasar dalam data.

Ada dua jenis tugas pemodelan prediktif: klasifikasi, yang digunakan untuk variabel target diskrit, dan regresi, yang digunakan untuk variabel target kontinu. Misalnya, memprediksi apakah pengguna Web akan melakukan pembelian di toko buku online adalah tugas klasifikasi karena variabel target bernilai biner. Di sisi lain, meramalkan harga suatu saham di masa depan adalah tugas regresi karena harga adalah atribut yang dinilai secara berkelanjutan.

Analisis asosiasi

Digunakan untuk menemukan pola yang menggambarkan fitur-fitur yang sangat terkait dalam data. Pola yang ditemukan biasanya direpresentasikan dalam bentuk aturan implikasi atau himpunan bagian fitur.

Tujuan analisis asosiasi adalah mengekstrak pola yang paling menarik dengan cara yang efisien. Penerapan analisis asosiasi yang berguna mencakup menemukan kelompok gen yang memiliki fungsi terkait.

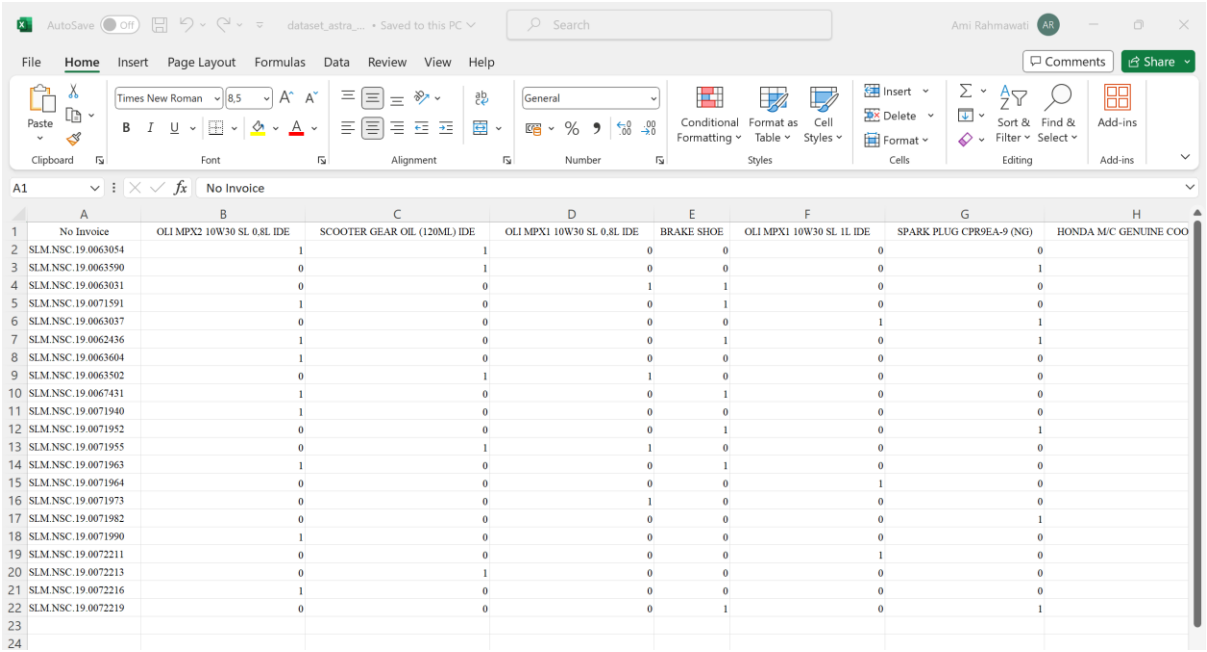
Dalam bentuknya yang paling primitif, masalah penambangan pola asosiasi didefinisikan dalam konteks database biner yang jarang, di mana matriks data hanya berisi 0/1 entri, dan sebagian besar entri bernilai 0. Sebagian besar database transaksi pelanggan adalah jenis ini. Contoh (Market Basket Analysis). Transaksi yang ditunjukkan pada tabel berikut mengilustrasikan data tempat penjualan yang dikumpulkan di konter kasir sebuah toko kelontong.

Analisis asosiasi dapat diterapkan untuk menemukan barang-barang yang sering dibeli bersama oleh pelanggan. Misalnya, kita mungkin menemukan aturan {Popok} $\longrightarrow$ {Susu}, yang menunjukkan bahwa pelanggan yang membeli popok juga cenderung membeli susu. Jenis

aturan ini dapat digunakan untuk mengidentifikasi potensi peluang penjualan silang di antara barang-barang terkait.

Transaction ID	Items
1	{Bread, Butter, Diapers, Milk}
2	{Coffee, Sugar, Cookies, Salmon}
3	{Bread, Butter, Coffee, Diapers, Milk, Eggs}
4	{Bread, Butter, Salmon, Chicken}
5	{Eggs, Bread, Butter}
6	{Salmon, Diapers, Milk}
7	{Bread, Tea, Sugar, Eggs}
8	{Coffee, Sugar, Chicken, Eggs}
9	{Bread, Diapers, Milk, Salt}
10	{Tea, Eggs, Cookies, Diapers, Milk}

Gambar 22. Contoh Data Asosiasi



No Invoice	OLI MPX2 10W30 SL 0,8L IDE	SCOOTER GEAR OIL (120ML) IDE	OLI MPX1 10W30 SL 0,8L IDE	BRAKE SHOE	OLI MPX1 10W30 SL 1L IDE	SPARK PLUG CPR9EA-9 (NG)	HONDA M/C GENUINE COO
SLM.NSC.19.0063054	1	1	0	0	0	0	
SLM.NSC.19.0063590	0	1	0	0	0	1	
SLM.NSC.19.0063031	0	0	1	1	0	0	
SLM.NSC.19.0071591	1	0	0	1	0	0	
SLM.NSC.19.0063037	0	0	0	0	1	1	
SLM.NSC.19.0062436	1	0	0	1	0	1	
SLM.NSC.19.0063604	1	0	0	0	0	0	
SLM.NSC.19.0063502	0	1	1	0	0	0	
SLM.NSC.19.0067431	1	0	0	1	0	0	
SLM.NSC.19.0071940	1	0	0	0	0	0	
SLM.NSC.19.0071952	0	0	0	1	0	1	
SLM.NSC.19.0071955	0	1	1	0	0	0	
SLM.NSC.19.0071963	1	0	0	1	0	0	
SLM.NSC.19.0071964	0	0	0	0	1	0	
SLM.NSC.19.0071973	0	0	1	0	0	0	
SLM.NSC.19.0071982	0	0	0	0	0	1	
SLM.NSC.19.0071990	1	0	0	0	0	0	
SLM.NSC.19.0072211	0	0	0	0	1	0	
SLM.NSC.19.0072213	0	1	0	0	0	0	
SLM.NSC.19.0072216	1	0	0	0	0	0	
SLM.NSC.19.0072219	0	0	0	1	0	1	

Analisis Cluster

Clustering didefinisikan sebagai pengelompokan sekumpulan objek serupa ke dalam kelas atau cluster. Dengan kata lain, pada analisis cluster, data dikelompokkan ke dalam kelas-kelas atau cluster-cluster, sehingga record-record dalam suatu cluster (intra-cluster) mempunyai kemiripan yang tinggi satu sama lain, namun memiliki ketidaksamaan yang tinggi jika dibandingkan dengan objek-objek yang berada dalam cluster lain (antar-cluster). Beberapa contoh penerapan yang relevan adalah sebagai berikut:

1. Segmentasi pelanggan: Dalam banyak aplikasi, diinginkan untuk menentukan pelanggan yang serupa satu sama lain dalam konteks berbagai tugas promosi produk. Fase segmentasi memainkan peran penting dalam proses ini.
2. Data summarization: Karena klaster dapat dianggap sebagai kelompok rekaman yang serupa, kelompok serupa ini dapat digunakan untuk membuat ringkasan data.

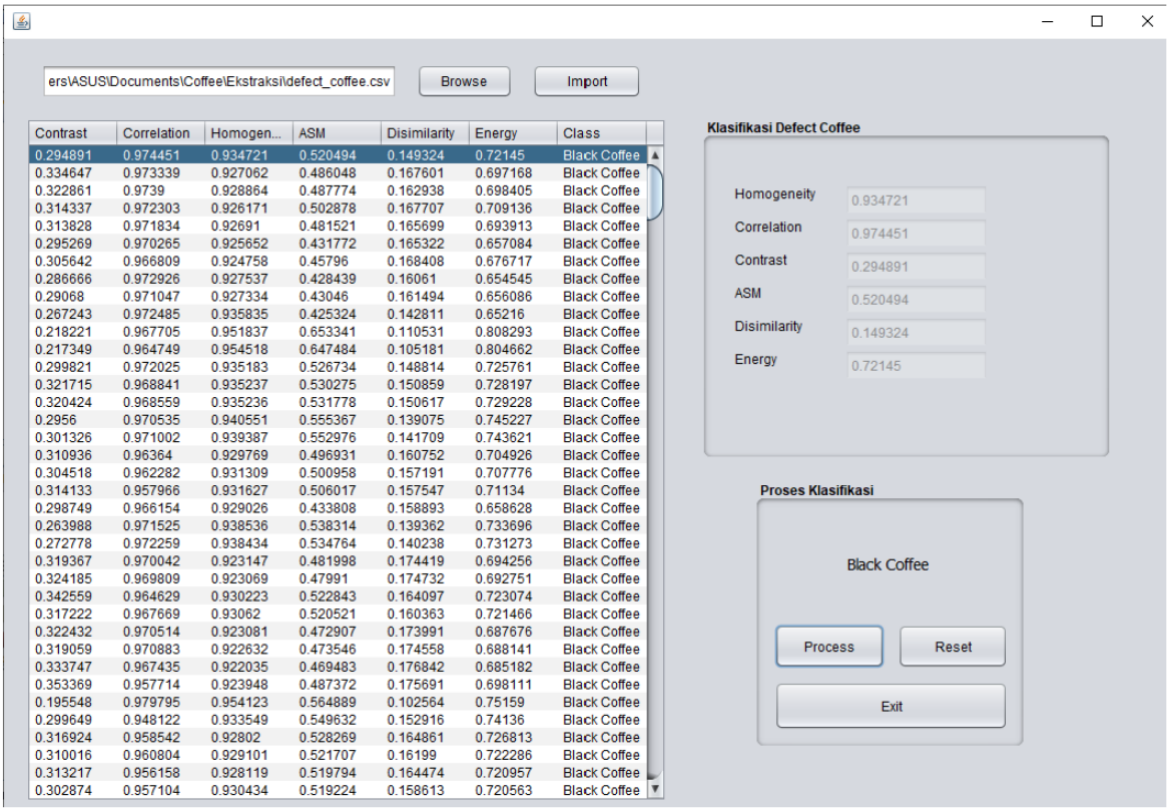
Gambar 23. Contoh Data Clustering

Analisis Klasifikasi

Klasifikasi adalah metode klasik yang digunakan oleh peneliti pembelajaran mesin dan ahli statistik untuk memprediksi hasil sampel yang tidak diketahui. Ini digunakan untuk mengkategorikan objek (atau benda) ke dalam sejumlah kelas tertentu.

Masalah klasifikasi dapat terdiri dari dua jenis, biner atau multikelas. Dalam klasifikasi biner, atribut target hanya dapat memiliki dua kemungkinan nilai. Misalnya tumor itu bersifat kanker atau tidak, suatu tim akan menang atau kalah, sentimen sebuah kalimat positif atau negatif, dan sebagainya. Dalam klasifikasi multikelas, atribut target dapat memiliki lebih dari dua nilai. Misalnya, tumor dapat berupa kanker tipe 1, tipe 2, atau tipe 3; sentimen sebuah kalimat bisa bahagia, sedih, marah atau cinta; berita dapat diklasifikasikan menjadi berita cuaca, keuangan, hiburan, atau olahraga. Beberapa contoh situasi bisnis yang menerapkan teknik klasifikasi adalah:

1. Untuk menganalisis riwayat kredit nasabah bank untuk mengidentifikasi apakah pemberian pinjaman kepada mereka berisiko atau aman.
2. Untuk menganalisis riwayat pembelian pelanggan suatu pusat perbelanjaan untuk memprediksi apakah mereka akan membeli suatu produk tertentu atau tidak.



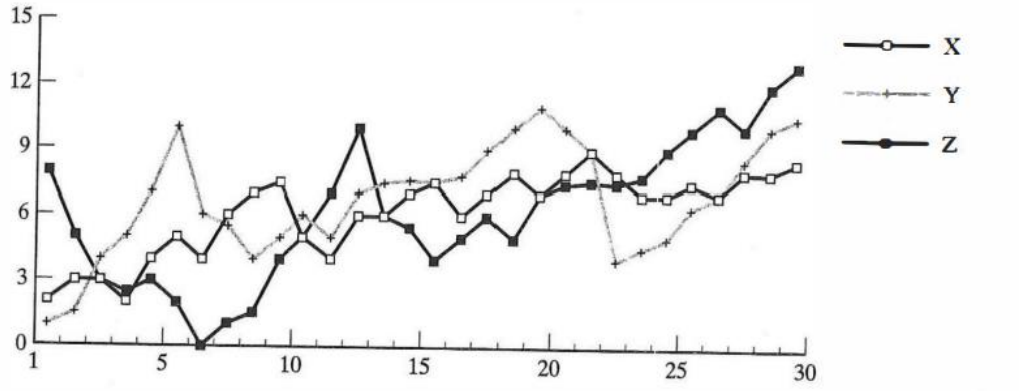
Contrast	Correlation	Homogen...	ASM	Disimilarity	Energy	Class
0.294891	0.974451	0.934721	0.520494	0.149324	0.72145	Black Coffee
0.334647	0.973339	0.927062	0.486048	0.167601	0.697168	Black Coffee
0.322861	0.9739	0.928864	0.487774	0.162938	0.698405	Black Coffee
0.314337	0.972303	0.926171	0.502878	0.167707	0.709136	Black Coffee
0.313828	0.971834	0.92691	0.481521	0.165699	0.693913	Black Coffee
0.295269	0.970265	0.925652	0.431772	0.165322	0.657084	Black Coffee
0.305642	0.966809	0.924758	0.45796	0.168408	0.676717	Black Coffee
0.286666	0.972926	0.927537	0.428439	0.16061	0.654545	Black Coffee
0.29068	0.971047	0.927334	0.43046	0.161494	0.656086	Black Coffee
0.267243	0.972485	0.935835	0.425324	0.142811	0.65216	Black Coffee
0.218221	0.967705	0.951837	0.653341	0.110531	0.808293	Black Coffee
0.217349	0.964749	0.954518	0.647484	0.105181	0.804662	Black Coffee
0.299821	0.972025	0.935183	0.526734	0.148814	0.725761	Black Coffee
0.321715	0.968841	0.935237	0.530275	0.150859	0.728197	Black Coffee
0.320424	0.968559	0.935236	0.531778	0.150617	0.729228	Black Coffee
0.2956	0.970535	0.940551	0.555367	0.139075	0.745227	Black Coffee
0.301326	0.971002	0.939387	0.552976	0.141709	0.743621	Black Coffee
0.310936	0.96364	0.929769	0.496931	0.160752	0.704926	Black Coffee
0.304518	0.962282	0.931309	0.500958	0.157191	0.707776	Black Coffee
0.314133	0.957966	0.931627	0.506017	0.157547	0.71134	Black Coffee
0.298749	0.966154	0.929026	0.433808	0.158893	0.658628	Black Coffee
0.263988	0.971525	0.938536	0.538314	0.139362	0.733696	Black Coffee
0.272778	0.972259	0.938434	0.534764	0.140238	0.731273	Black Coffee
0.319367	0.970042	0.923147	0.481998	0.174419	0.694256	Black Coffee
0.324185	0.969809	0.923069	0.47991	0.174732	0.692751	Black Coffee
0.342559	0.964629	0.930223	0.522843	0.164097	0.723074	Black Coffee
0.317222	0.967669	0.93062	0.520521	0.160363	0.721466	Black Coffee
0.322432	0.970514	0.923081	0.472907	0.173991	0.687676	Black Coffee
0.319059	0.970883	0.922632	0.473546	0.174558	0.688141	Black Coffee
0.333747	0.967435	0.922035	0.469483	0.176842	0.685182	Black Coffee
0.353369	0.957714	0.923948	0.487372	0.175691	0.698111	Black Coffee
0.195548	0.979795	0.954123	0.564889	0.102564	0.75159	Black Coffee
0.299649	0.948122	0.933549	0.549632	0.152916	0.74136	Black Coffee
0.316924	0.958542	0.92802	0.528269	0.164861	0.726813	Black Coffee
0.310016	0.960804	0.929101	0.521707	0.16199	0.722286	Black Coffee
0.313217	0.956158	0.928119	0.519794	0.164474	0.720957	Black Coffee
0.302874	0.957104	0.930434	0.519224	0.158613	0.720563	Black Coffee

Gambar 25. Contoh Data Klasifikasi

Analisis Time Series Analysis

Dengan analisis deret waktu, nilai suatu atribut diperiksa karena nilainya bervariasi seiring waktu. Nilai biasanya diperoleh sebagai titik waktu dengan jarak yang sama (harian, mingguan, per jam, dll.).

Contoh: Contoh: Tuan Smith sedang mencoba menentukan apakah akan membeli saham dari Perusahaan X, Y, atau z. Untuk jangka waktu satu bulan ia memetakan harga saham harian setiap perusahaan. Gambar berikut menunjukkan plot deret waktu yang dihasilkan oleh Mr. Smith. Dengan menggunakan informasi ini dan informasi serupa yang tersedia dari pialang sahamnya, Mr. Smith memutuskan untuk membeli saham X karena saham tersebut tidak terlalu fluktuatif dan secara keseluruhan menunjukkan jumlah pertumbuhan yang relatif sedikit lebih besar dibandingkan saham lainnya. Faktanya, saham Y dan Z memiliki perilaku serupa. Perilaku Y antara hari ke 6 dan 20 identik dengan perilaku Z antara hari ke 13 dan 27.



Gambar 26. Contoh Time Series Analysis

Plot deret waktu berdasarkan gambar diatas, digunakan untuk memvisualisasikan deret waktu. Dalam gambar ini Anda dapat dengan mudah melihat bahwa plot untuk Y dan Z memiliki perilaku serupa, sedangkan X tampaknya memiliki volatilitas yang lebih kecil.

Analisis Anomali

Analisis anomali adalah tugas mengidentifikasi observasi yang karakteristiknya berbeda secara signifikan dari data lainnya. Pengamatan seperti ini dikenal sebagai anomali atau outlier. Tujuan dari algoritma deteksi anomali adalah untuk menemukan anomali yang sebenarnya dan menghindari pemberian label yang salah pada objek normal sebagai anomali.

Contoh (Deteksi Penipuan Kartu Kredit). Perusahaan kartu kredit mencatat transaksi yang dilakukan oleh setiap pemegang kartu kredit, beserta informasi pribadi seperti batas kredit, usia, pendapatan tahunan, dan alamat. Karena jumlah kasus penipuan relatif kecil dibandingkan dengan jumlah transaksi yang sah, teknik deteksi anomali dapat diterapkan untuk membangun profil transaksi yang sah bagi pengguna. Ketika transaksi baru masuk, transaksi tersebut dibandingkan dengan profil pengguna. Jika karakteristik transaksi sangat berbeda dengan profil yang dibuat sebelumnya, maka transaksi tersebut ditandai sebagai berpotensi penipuan.

MODUL 6: Metode Learning Algoritma Data Mining

Kompetensi: Mahasiswa mampu memahami metode learning algoritma data mining

Definisi Machine Learning

Machine Learning merupakan salah satu cabang dari AI untuk pengembangan algoritma dan model komputer yang dapat belajar dari data dan melakukan prediksi atau pengambilan keputusan tanpa perlu secara eksplisit diprogram secara langsung.

Tujuan utama pembelajaran mesin adalah untuk membangun platform Kecerdasan Buatan (AI) yang secerdas pikiran manusia. Kita tidak jauh dari mimpi ini dan banyak peneliti AI percaya bahwa tujuan ini dapat dicapai melalui algoritma pembelajaran mesin yang mencoba meniru proses pembelajaran otak manusia.

Tabel 1. Perbandingan Data Mining dan Machine Learning

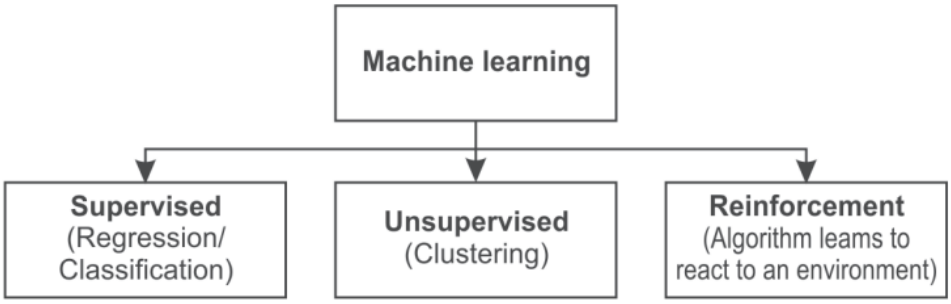
Dasar untuk perbandingan	**Data Mining**	**Machine Learning**
Arti	Ini melibatkan penggalian pengetahuan yang berguna dari sejumlah besar data.	Ini memperkenalkan algoritma baru dari data serta pengalaman masa lalu.
Sejarah	Diperkenalkan pada tahun 1930, awalnya disebut penemuan pengetahuan dalam database.	Itu diperkenalkan pada tahun 1959.
Responsibility	Data mining digunakan untuk memeriksa pola pada data yang ada. Ini kemudian dapat digunakan untuk menetapkan aturan.	Pembelajaran mesin mengajarkan komputer untuk mempelajari dan memahami aturan yang diberikan.
Sifat	Ini melibatkan keterlibatan dan intervensi manusia.	Hal ini dilakukan secara otomatis, setelah dirancang maka akan dapat diimplementasikan secara mandiri dan tidak memerlukan atau hanya sedikit upaya manusia yang diperlukan.

Data mentah mungkin berubah-ubah, tidak terstruktur, atau bahkan dalam format yang tidak sesuai untuk pemrosesan otomatis. Misalnya, data yang dikumpulkan secara manual

mungkin diambil dari berbagai sumber dalam format berbeda, namun perlu diproses oleh program komputer otomatis untuk mendapatkan wawasan.

Untuk mengatasi masalah ini, analis data mining menggunakan jalur pemrosesan, di mana data mentah dikumpulkan, dibersihkan, dan diubah ke dalam format standar. Data dapat disimpan dalam sistem database dan akhirnya diproses untuk mendapatkan wawasan dengan menggunakan metode analitis. Teknik data mining digunakan untuk menjelajahi database besar guna menemukan pola baru dan berguna yang mungkin masih belum diketahui.

Klasifikasi Algoritma Pembelajaran Mesin



Gambar 27. Klasifikasi Algoritma Pembelajaran Mesin

Supervised Learning

Sebagai cabang pertama dari pembelajaran mesin, supervised learning berkonsentrasi pada pola pembelajaran melalui menghubungkan hubungan antara variabel dan hasil yang diketahui dan bekerja dengan kumpulan data berlabel.

Supervised learning bekerja dengan memasukkan data sampel mesin dengan berbagai fitur (diwakili sebagai “X”) dan keluaran nilai data yang benar (diwakili sebagai “y”). Fakta bahwa nilai keluaran dan fitur diketahui membuat kumpulan data memenuhi syarat sebagai "berlabel".

Algoritma kemudian menguraikan pola yang ada dalam data dan membuat model yang dapat mereproduksi aturan dasar yang sama dengan data baru.

Misalnya, untuk memprediksi harga pasar pembelian mobil bekas, supervised algorithm dapat merumuskan prediksi dengan menganalisis hubungan antara atribut mobil (termasuk tahun pembuatan, merek mobil, jarak tempuh, dll.) dan harga jual mobil lainnya. Mobil yang dijual berdasarkan data historis. Mengingat supervised algorithm mengetahui harga akhir dari kartu lain yang terjual, algoritme tersebut kemudian dapat bekerja mundur untuk menentukan hubungan antara karakteristik mobil dan nilainya.

Setelah mesin menguraikan aturan dan pola data, mesin menciptakan apa yang disebut model: persamaan algoritmik untuk menghasilkan hasil dengan data baru berdasarkan aturan yang diturunkan dari data pelatihan. Setelah model disiapkan, model dapat diterapkan pada data baru dan diuji keakuratannya. Setelah model melewati tahap data pelatihan dan pengujian, maka model siap diterapkan dan digunakan di dunia nyata.

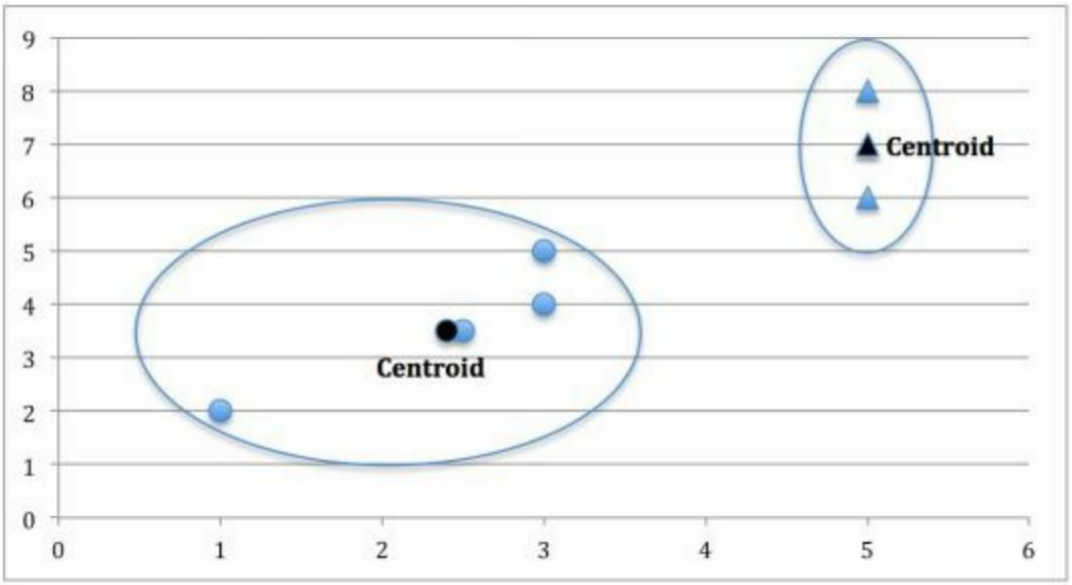
No.	Shape	Colour	Size	Fruit Name
1	Heart-shaped to nearly globular	Red	Small	Cherry
2	Kidney shaped	Yellow	Big	Mango
3	Round to oval, bunch shaped cylindrical	Green	Small	Grapes
4	Long curving cylinder	Green	Big	Banana

Setelah mesin menguraikan aturan dan pola data, mesin menciptakan apa yang disebut model: persamaan algoritmik untuk menghasilkan hasil dengan data baru berdasarkan aturan yang diturunkan dari data pelatihan. Setelah model disiapkan, model dapat diterapkan pada data baru dan diuji keakuratannya. Setelah model melewati tahap data pelatihan dan pengujian, maka model siap diterapkan dan digunakan di dunia nyata.

Unsupervised Learning

Dalam kasus unsupervised learning, tidak semua variabel dan pola data diklasifikasikan. Sebaliknya, mesin harus mengungkap pola tersembunyi dan membuat label melalui penggunaan algoritma unsupervised learning.

Keuntungan unsupervised learning adalah memungkinkan analis menemukan pola dalam data yang tidak Anda sadari keberadaannya—seperti keberadaan dua jenis pelanggan utama.



No.	Shape	Colour	Size
1	Heart-shaped to nearly globular	Red	Small
2	Kidney-shaped	Yellow	Big
3	Round to oval, bunch shape cylindrical	Green	Small
4	Long curving cylinder	Green	Big

Reinforcement Learning

Dengan demikian, reinforcement learning memungkinkan mesin dan agen perangkat lunak secara otomatis menentukan perilaku ideal dalam konteks tertentu, untuk memaksimalkan kinerjanya. reward feedback yang sederhana diperlukan agar agen dapat mempelajari perilakunya dan ini dikenal sebagai reinforcement signal. Program AlphaGo Google yang mengalahkan juara dunia dalam permainan Go, mobil self-driving dari Tesla Motors dan pengiriman udara utama Amazon semuanya didasarkan pada reinforcement learning.

STUDI KASUS: Klasifikasi pada Data Lung Cancer

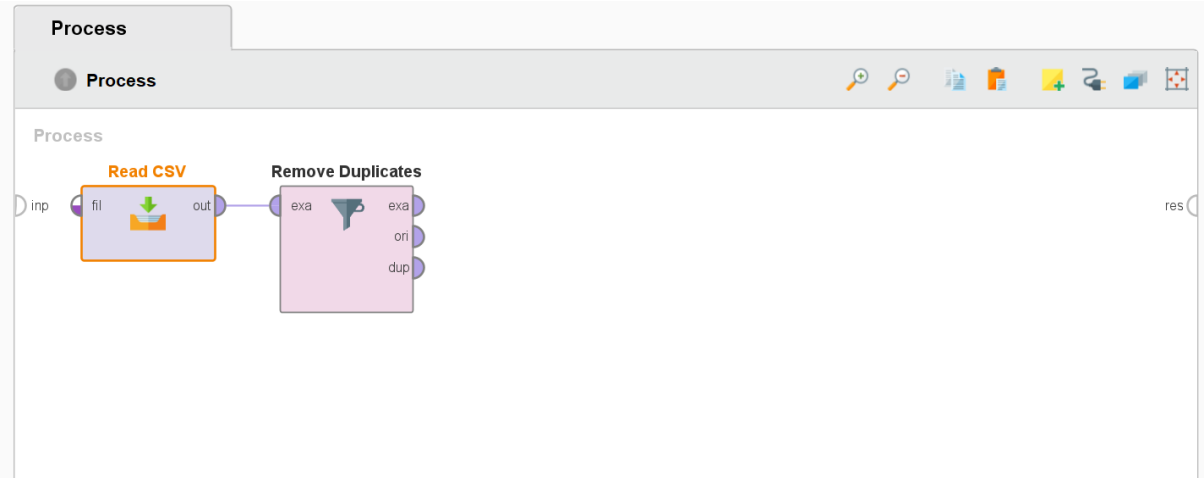
Dalam tugas data mining, klasifikasi merupakan salah satu teknik penting yang digunakan untuk memprediksi kategori atau label dari data yang belum diketahui. Dalam studi kasus ini, kita akan menerapkan teknik klasifikasi pada data kanker paru-paru (*Lung Cancer*) untuk memprediksi kemungkinan seseorang mengidap kanker paru-paru berdasarkan berbagai atribut kesehatan dan gaya hidup.

Tujuan dari studi kasus ini adalah untuk membangun model klasifikasi yang dapat memprediksi apakah seorang pasien memiliki kanker paru-paru berdasarkan atribut yang ada. Model yang dibangun diharapkan mampu mengidentifikasi pola dalam data yang berkaitan dengan faktor-faktor risiko kanker paru-paru, dan memberikan prediksi yang akurat untuk kasus-kasus baru. Data yang digunakan pada studi kasus ini yaitu data Lung Cancer (Kanker Paru-paru) yang di download pada laman Kaggle dengan link sebagai berikut:
https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer

Data diolah menggunakan tools RapidMiner dengan langkah-langkah seperti dibawah ini:

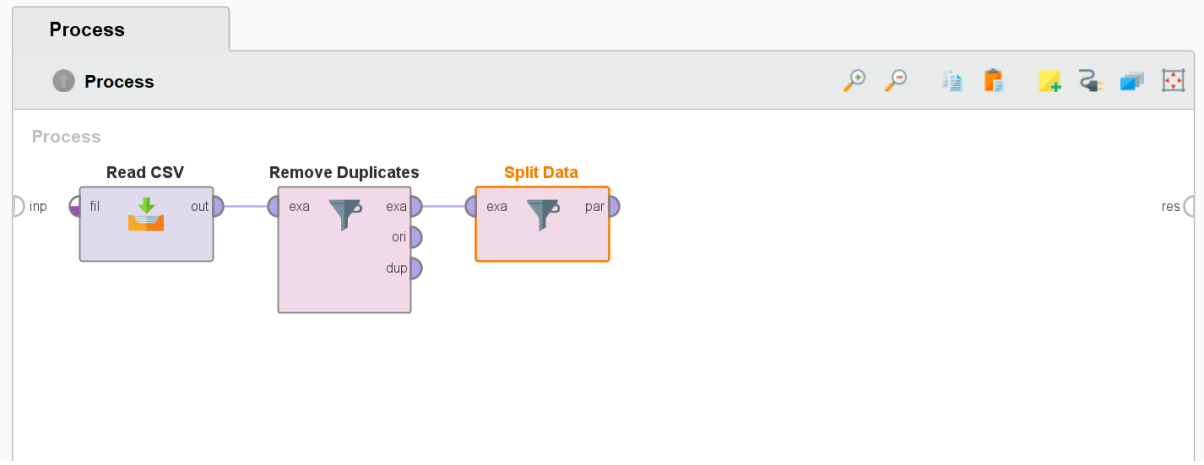
1. Eksplorasi Data dan Pra-Pemrosesan Data

Langkah pertama adalah melakukan eksplorasi data untuk memahami distribusi data, memeriksa adanya missing values, dan melakukan analisis statistik dasar. Ini akan membantu dalam pemahaman awal terhadap data dan fitur-fitur yang relevan. Dalam data ini terdapat data ganda (duplicate data), maka data tersebut dihapus menggunakan fitur **Remove Duplicates** dalam RapidMiner.



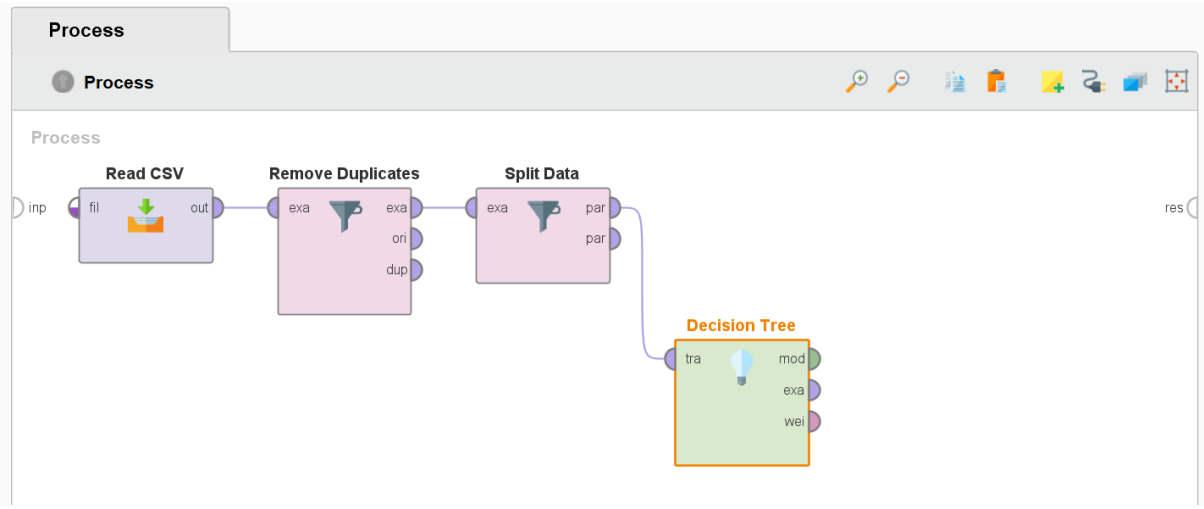
2. Pembagian Data

Dataset akan dibagi menjadi data latih (*training data*) dan data uji (*testing data*). Data latih digunakan untuk melatih model klasifikasi, sementara data uji digunakan untuk mengevaluasi kinerja model. Untuk mendeteksi lung cancer, dilakukan pembagian dataset menjadi dua bagian, yakni 80% untuk data pelatihan dan 20% untuk data pengujian.



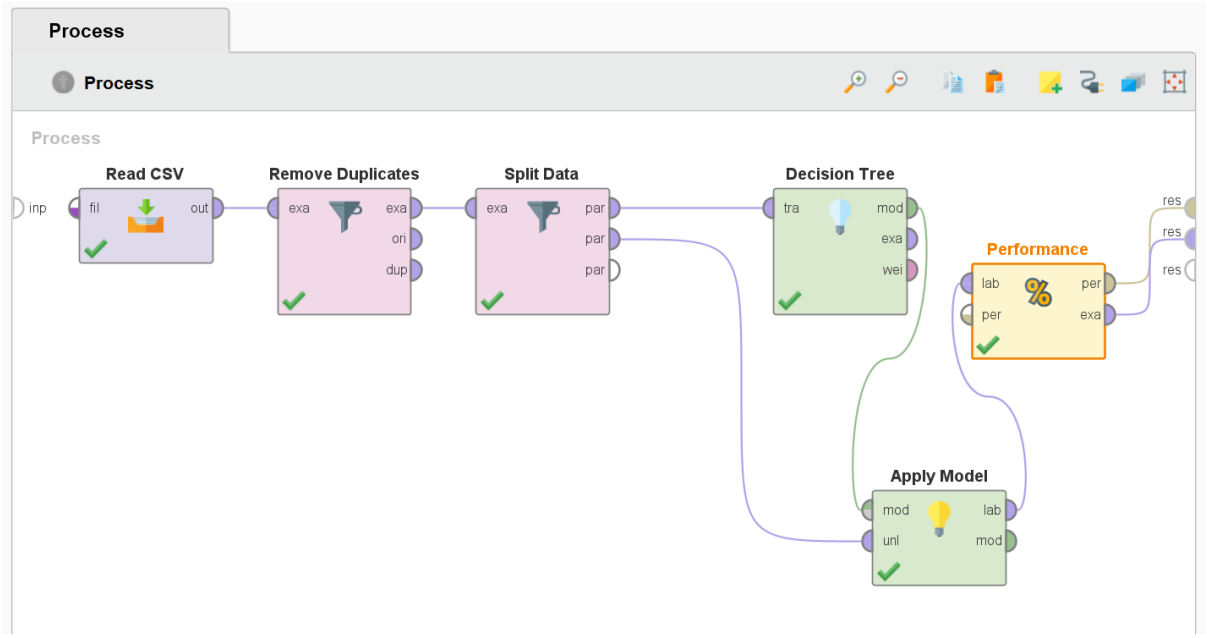
3. Pemilihan Algoritma Klasifikasi

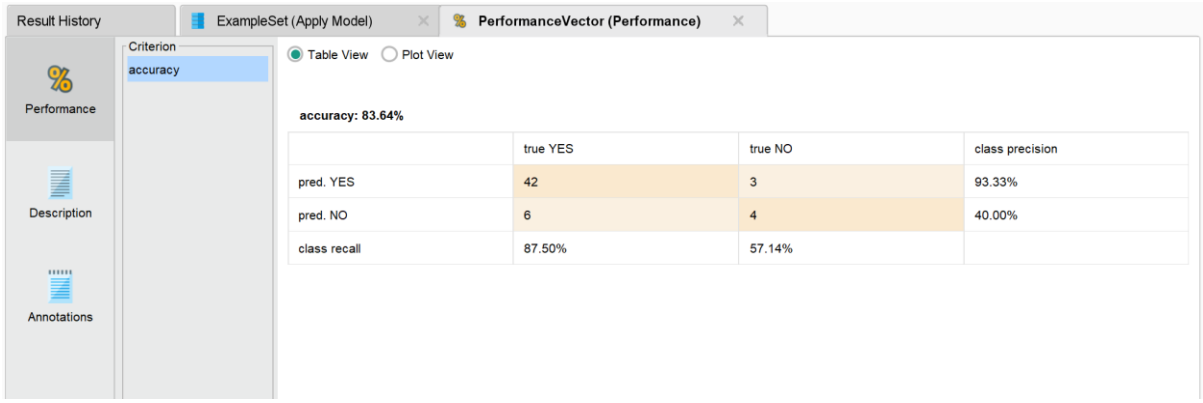
Berbagai algoritma klasifikasi seperti Decision Tree, Random Forest, Support Vector Machine (SVM), atau K-Nearest Neighbors (KNN) dapat digunakan. Model yang berbeda akan dilatih, dan kinerja masing-masing akan dibandingkan untuk memilih model terbaik. Model klasifikasi yang dibangun menggunakan Algoritma Decision Tree.



4. Pelatihan dan Evaluasi Model

Model akan dilatih menggunakan data latih dan dievaluasi dengan menggunakan data uji. Metode evaluasi yang umum digunakan meliputi akurasi, precision, dan recall.





	true YES	true NO	class precision
pred. YES	42	3	93.33%
pred. NO	6	4	40.00%
class recall	87.50%	57.14%	

DAFTAR PUSTAKA

Anggarwal, Charu C. 2015. *Data Mining: The Textbook*. USA.

Bhatia, Parteek. 2019. *Data Mining and Data Warehousing*. USA.

Kimball, Ralph, and Margy Ross. 2002. *The Data Warehouse Toolkit Second Edition*. Canada.

Loshin, David. 2013. *Business Intelligence: The Savvy Manager's Guide*. United States of America.

Vaisman, Alejandro, and Esteban Zimanyi. 2014. *Data Warehouse Systems*.

Vercellis, Carlo. 2009. *Business Intelligence : Data Mining and Optimization for Decision Making*. United Kingdom.