KONSEP DATA MINING

(DATA WAREHOUSE DAN BIGDATA)

Turkhamun Adi Kurniawan S.T M.Kom
Dr. Arman Syah Putra S.Kom MM M.Kom
Fatrilia Rasyi Radita, S.Pd.I, M.Pd.I
Muhammad Hilman Fakhriza M.Kom
Juniana Husna S.Si M.Sc
V.H.Valentino S.Kom MMSI

2022

KONSEP DATA MINING

(DATA WAREHOUSE DAN BIG DATA)

Turkhamun Adi Kurniawan S.T M.Kom

Dr. Arman Syah Putra S.Kom MM M.Kom

Fatrilia Rasyi Radita, S.Pd.I, M.Pd.I

Muhammad Hilman Fakhriza M.Kom

Juniana Husna S.Si M.Sc

V.H.Valentino S.Kom MMSI

2022

Konsep Data Mining
Data Warehouse dan Big Data
Penulis
Turkhamun Adi Kurniawan S.T M.Kom || Dr. Arman Syah Putra S.Kom MM M.Kom || Fatrilia Rasyi Radita, S.Pd.I, M.Pd.I || Muhammad Hilman Fakhriza M.Kom || Juniana Husna S.Si M.Sc || V.H.Valentino S.Kom MMSI
ISBN
978-623-481-037-0
Editor
Tim Kun Fayakun
Layout
Tim Kun Fayakun
Penyunting
Tim Kun Fayakun
Desain Sampul dan Tata Letak
Tim Kun Fayakun
Penerbit
Tim Kun Fayakun
Redaksi
Tim Kun Fayakun
Jawa Timur
Hp. 0856 07 8802
Email : penulis.kunfayakun@gmail.com
Cetakan pertama : Mei 2022

KATA PENGANTAR

Terima kasih kepada Allah SWT dan kedua orang tua kami atas terbit nya buku perdana di tahun 2022 ini, dengan terbit nya buku ini di harapkan berguna untuk semua orang dalam hal riset dan penelitian di bidang data terutama di bidang data mining, buku ini berisikan tentang pengertian data mining dan di harapkan ada buku-buku selanjutnya dan bisa terus berkarya agar bisa membantu dalam hal riset dan penelitian di bidang kecerdasan buatan. Buku ini merupakan buku yang merupakan saduran dan pemikiran dari beberapa sumber yang dijadikan dalam sebuah buku yang bisa membantu banyak di bidang pengetahuan.

Penulis

Turkhamun Adi Kurniawan S.T M.Kom
Dr. Arman Syah Putra S.Kom MM M.Kom
Fatrilia Rasyi Radita, S.Pd.I, M.Pd.I
Muhammad Hilman Fakhriza M.Kom
Juniana Husna S.Si M.Sc
V.H.Valentino S.Kom MMSI

DAFTAR ISI

METODE DATA MINING

CLASSIFICATION

METODE DATA MINING

CLUSTERING DAN ALGORITMA

PENDAHULUAN

Penyimpanan dokumen secara digital berkembang dengan pesat seiring meningkatnya penggunaan komputer. Kondisi tersebut memunculkan masalah untuk mengakses informasi yang diinginkan secara akurat dan cepat. Oleh karena itu, walaupun sebagian besar dokumen digital tersimpan dalam bentuk teks dan berbagai algoritma yang efisien untuk pencarian teks telah dikembangkan, teknik pencarian terhadap seluruh isi dokumen yang tersimpan bukanlah solusi yang tepat mengingat pertumbuhan ukuran data yang tersimpan umumnya. Pencarian informasi (Information Retrieval) adalah salah satu cabang ilmu yang menangani masalah ini yang bertujuan untuk membantu pengguna dalam menemukan informasi yang relevan dengan kebutuhan mereka dalam waktu singkat.

Aplikasi pencarian informasi yang telah ada salah satunya adalah web mining untuk pencarian berdasarkan kata kunci dengan teknik clustering. Selain itu, pada dokumen dilakukan juga text mining dan perhitungan jumlah kata, dari jumlah kata tersebut dilakukan pengklusteran dengan metode CLHM (Centroid Linkage Hierarchical Method).

Untuk jumlah klusternya, pemakai tidak mengetahui berapa jumlah yang tepat untuk mengklusterkan dokumen-dokumen tersebut. Untuk itu, dipakailah metode Hill Climbing yang bertugas untuk melakukan identifikasi terhadap pergerakan varian dari tiap tahap pembentukan kluster dan menganalisa polanya sehingga jumlah kluster akan terbentuk secara otomatis.

Penggunaan text mining, pengklusteran dengan CLHM dan proses Hill Climbing Automatic Clustering sangat memudahkan pemakai karena menghasilkan kluster secara otomatis dan tepat dengan waktu yang cepat.

Pengertian dan definisi Data Mining.

Data mining merupakan proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidenfikasi informasi yang bermanfaat dan pengetahuan yang terakit dari berbagai database besar/Data Warehouse (Turban, dkk. 2005)

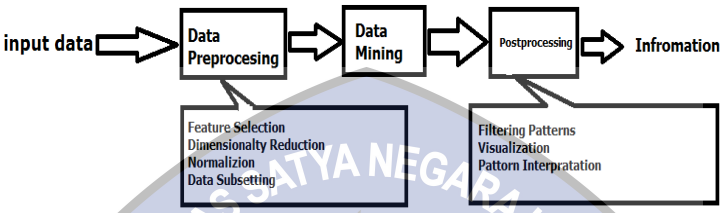
Keluaran dari data mining bisa dipakai untuk memperbaiki pengambilan keputusan dimasa depan

(Budi Santosa, 2007)

Data mining adalah sebuah proses percarian secara otomatis informasi yang berguna dalam tempat penyimpanan data berukuran besar. Istilah lain yang sering digunakan diantaranya knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, dan business intelligence. Teknik data mining digunakan untuk memeriksa basis data berukuran besar sebagai cara untuk menemukan pola yang baru dan berguna. Tidak semua pekerjaan pencarian informasi dinyatakan sebagai data mining. Sebagai contoh, pencarian record individual menggunakan database management system atau pencarian halaman we tertentu melalui kueri ke semua search engine adalah pekerjaan pencarian informasi yang erat kaitannya dengan information retrieval. Teknik-teknik data mining dapat digunakan untuk meningkatkan kemampuan sistem-sistem information retrieval.

Data mining adalah bagian integral dari knowledge discovery in databases (KDD). Keseluruhan proses KDD

untuk konversi raw data ke dalam informasi yang berguna ditunjukkan dalam Gambar 1.1.



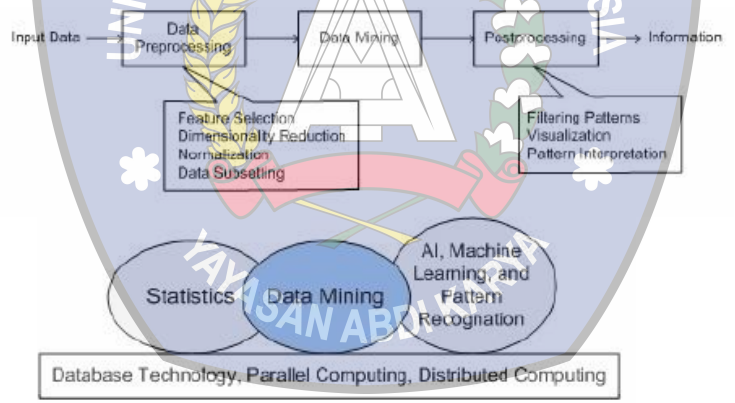
Data input dapat disimpan dalam berbagai format seperti flat file,spreadsheet, atau tabel-tabel relasional, dan dapat menempati tempat penyimpanan data terpusat atau terdistribusi pada banyak tempat. Tujuan dari preprocessing adalah mentransformasikan data input mentah ke dalam format yang sesuai untuk analisis selanjutnya. Langkah-langkah yang terlibat dalam preprocessing data meliputi mengabungkan data dari berbagai sumber, membersihkan (cleaning) data untuk membuang noise dan observasi duplikat, dan menyeleksirecord dan fitur yang relevan untuk pekerjaandata mining. Karena terdapat banyak cara mengumpulkan dan menyimpan data, tahapan

preprocessing data merupakan langkat yang banyak menghabiskan waktu dalam KDD.

Hasil dari data mining sering kali diintegrasikan dengan decision support system (DSS). Sebagai contoh, dalam aplikasi bisnis informasi yang dihasilkan olehdata mining dapat diintegrasikan dengan tool manajemen kampanye produk sehingga promosi pemasaran yang efektif yang dilaksanakan dan dapat diuji. Integrasi demikian memerlukan langkah postprocessing yang menjamin bahwa hanya hasil yang valid dan berguna yang akan digabungkan dengan DSS. Salah satu pekerjaan dan postprocessing adalah visualisasi yang memungkinkan analyst untuk mengeksplor data dan hasil data mining dari berbagai sudur pandang. Ukuran-ukuran statistik dan metode pengujian hipotesis dapat digunakan selama postprocessing untuk membuang hasil data mining yang palsu.

Secara khusus, data mining menggunakan ide-ide seperti (1) pengambilan contoh, estimasi, dan pengujian hipotesis, dari statistika dan (2) algoritme pencarian, teknik pemodelan, dan teori pembelajaran dari kecerdasan

buatan, pengenalan pola, dan machine learning. Data mining juga telah mengadopsi ide-ide dari area lain meliputi optimisasi, evolutionary computing, teori informasi, pemrosesan sinyal, visualisasi dan information retrieval. Sejumlah area lain juga memberikan peran pendukung dalam data mining, seperti sistem basis data yang dibutuhkan untuk menyediakan tempat penyimpanan yang efisien, indexing dan pemrosesan kueri. Gambar 1.2 menunjukkan hubungan data mining dengan area-area lain.



Data mining sebagai pertemuan dari banyak disiplin ilmu (Tan et al, 2005) Data mining merupakan proses pencarian pengetahuan yang menarik daridata

berukuran besar yang disimpan dalam basis data, data warehouse atau tempat penyimpanan informasi lainnya. Dengan demikian arsitektur system data mining memiliki komponen-komponen utama yaitu:

- Basis data,data warehouse atau tempat penyimpanan informasi lainnya.
- Basis data dan data warehouse server. Komponen ini bertanggung jawab dalam pengambilan relevant data, berdasarkan permintaan pengguna.

Basis pengetahuan. Komponen ini merupakan domain knowledge yang digunakan untuk memandu pencarian atau mengevaluasi pola-pola yangdihasilkan. Pengetahuan tersebut meliputi hirarki konsep yang digunakanuntuk mengorganisasikan atribut atau nilai atribut ke dalam level abstraksiyang berbeda. Pengetahuan tersebut juga dapat berupa kepercayaan pengguna(user belief), yang dapat digunakan untuk menentukan kemenarikan pola yangdiperoleh. Contoh lain dari domain knowledge adalah threshold dan Meta data yang menjelaskan data dari berbagai sumber yang heterogen.

• Data mining engine. Bagian ini merupakan komponen penting dalamarsitektur

sistemdata mining. Komponen ini terdiri modul-modul fungsionaldata mining seperti karakterisasi, asosiasi, klasifikasi, dan analisis cluster.

• Modul evaluasi pola. Komponen ini menggunakan ukuran-

ukurankemenarikan dan berinteraksi dengan moduldata mining dalam pencarian pola

• pola menarik. Modul evaluasi pola dapat menggunakan thresholdkemenaikan

untuk mem-filter pola-pola yang diperoleh.

• Antarmuka pengguna grafis. Modul ini berkomunikasi dengan pengguna

Dan system data mining. Melalui modul ini, pengguna berinteraksi dengan sistemmengan menentukan kueri atau taskdata mining. Antar muka jugamenyediakan informasi untuk memfokuskan pencarian dan melakukan eksplorasi data mining berdasarkan hasil data mining antara. Komponen ini juga memungkinkan pengguna untuk

mencari (browse) basis data dan skema data warehouse atau struktur data, evaluasi pola yang diperoleh danvisualisasi pola dalam berbagai bentuk.

Fungsi dan Proses dari Data Mining.

Fungsi Data mining

Data Mining mengidentifikasi fakta-fakta atau kesimpulan-kesimpulan yang di sarankan berdasarkan penyaringan melalui data untuk menjelajahi pola-pola atau anomali-anomali data. Data Mining mempunyai 5 fungsi:

a.Classification

Classification, yaitu proses penemuan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui atau menyimpulkan definisi-definisi karakteristik sebuah grup. Contoh: pelanggan-pelanggan perusahaan yang telah berpindah kesaingan perusahaan yang lain.

b. Clustering

Clustering termasuk metode yang sudah cukup dikenal dan banyak dipakai dalam data mining. Sampai sekarang para ilmuwan dalam bidang data inining masih melakukan berbagai usaha untuk melakukan perbaikan model clustering karena metode yang dikembangkan sekarang masih bersifat heuristic. Usaha-usaha untuk menghitung jumlah cluster yang optimal dan pengklasteran yang paling baik masih terus dilakukan. Denga demikian menggunakan metode yang sekarang, tidak bisa menjamin hasil pengklasteran sudah merupakan hasil yang optimal. Namun, hasil yang dicapai biasanya sudah cukup bagus dari segi praktis.

Clustering, yaitu mengindentifikasikan kelompok-kelompok dari barang- barang atau produk-produk yang mempunyai karakteristik khusus(clustering berbeda dengan classification, dimana pada clustering tidak terdapat definisi-definisi karakteristik awak yang di berikan pada waktu classification.)

c.Association

Association, yaitu mengidentifikasikan hubungan

antara kejadian-kejadian yang terjadi pada suatu waktu, seperti isi-isi dari keranjang belanja.

d.Sequencing

Hampir sama dengan association, sequencing mengidentifikasikan hubungan- hubungan yang berbeda pada suatu periode waktu tertentu, seperti pelanggan-pelanggan yang mengunjungi supermarket secara berulang-ulang.

e.Forecasting

Forecasting memperkirakan nilai pada masa yang akan datang berdasarkan pola-pola dengan sekumpulan data yang besar, seperti peramalan permintaan pasar.

f. Regretion

adalah proses pemetaam data dalam suatu nilai prediksi g. Solution adalah proses penemuan akar masalah dan problem solving dari persoalan bisnis yang dihadapi atau paling tidak sebagai informasi pendukung dalam pengambilan keputusan.

Tujuan *Data Mining* dan Proses *Data Mining*

Tujuan Data Mining

a. Explanatory

Untuk menjelaskan beberapa kondisi penelitian, seperti mengapa penjualan truk pick up meningkat di colorado.

b. Confirmatory

Untuk mempertegas hipotesis, seperti halnya 2 kali pendapatan keluarga lebih suka di pakai untuk membeli peralatan keluarga, di bandingkan dengan satu kali pendapatan keluarga.

c. Exploratory

Menganalisis data untuk hubungan yang baru yang tidak di harapkan, seperti halnya pola apa yang cocok untuk kasus penggelapan kartu kredit.

Proses *Data Mining*

Data mining sesunggunghnya merupakan salah satu rangkaian dari proses pencarian pengetahuan pada database (Knowledge Discovery in Database/KDD). KDD berhubungan dengan teknik integrasi dan

penemuan ilmiah, interprestasi dan visualisasi dari pola-pola sejumlah kumpulan data. KDD adalah keseluruhan proses non-trivial untuk mencari dan mengidentifikasi pola (pattern) dalam data, dimana pola yang ditemukan bersifat sah, baru, dapat bermanfaat dan dapat dimengerti. Serangkaian proses tersebut yang memiliki tahap sebagai berikut (Tan, 2004):

1. Pembersihan data dan integrasi data (cleaning and integration) Proses ini digunakan untuk membuang data yang tidak konsisten dan bersifat noise dari data yang terdapat di berbagai basisdata yang mungkin berbeda format maupun platform yang kemudian diintegrasikan dalam satu database datawarehouse. Pembersihan data merupakan proses menghilangkan noise dan data yang tidak relevan. Pada umumnya data yang diperoleh, baik dari database memiliki isian-isian yang tidak sempurna seperti data yang hilang, data yang tidak valid atau juga hanya sekedar salah ketik.Datadata yang tidak relevan itu juga lebih baik dibuang. Pembersihan data juga akan mempengaruhi performasi dari teknik data mining karena data yang ditangani akan berkurang jumlah dan kompleksitasnya.

Integrasi data merupakan penggabungan data dari berbagai database ke dalam satu database baru.Integrasi data perlu dilakukan secara cermat karena kesalahan pada integrasi data bisa menghasilkan hasil yang menyimpang dan bahkan menyesatkan pengambilan aksi nantinya. Sebagai contoh bila integrasi data berdasarkan jenis produk ternyata menggabungkan produk dari kategori yang berbeda maka akan didapatkan korelasi antar produk yang sebenarnya tidak ada.

2. Seleksi dan transformasi data (selection and transformation)

Data yang terdapat dalam database datawarehouse kemudian direduksi dengan berbagai teknik. Proses reduksi diperlukan untuk mendapatkan hasil yang lebih akurat dan mengurangi waktu komputasi terutama utuk masalah dengan skala besar (large scale problem). Beberapa cara seleksi, antra lain:

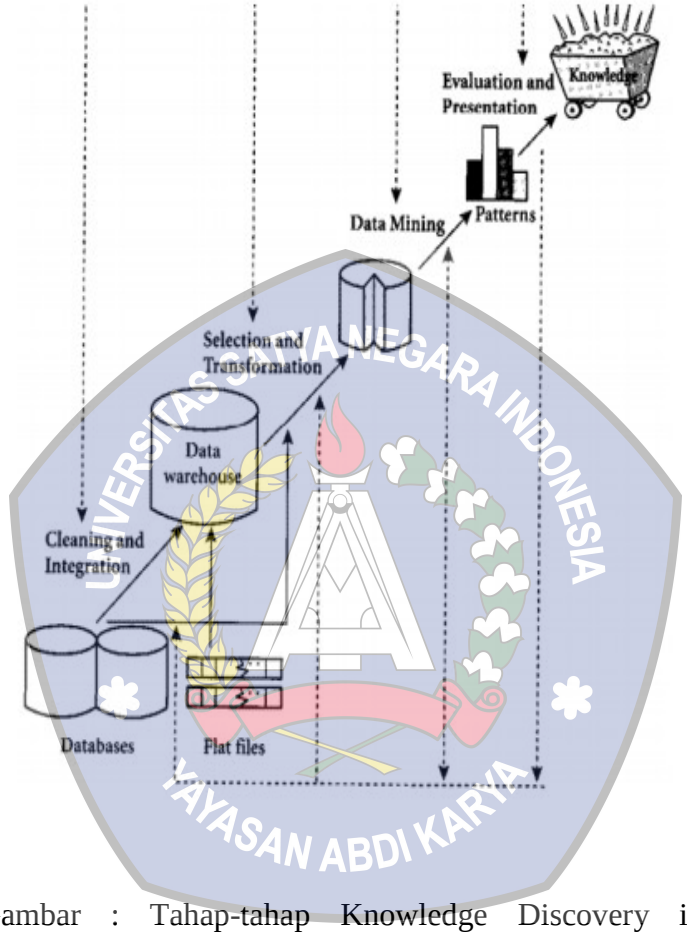
- Sampling, adalah seleksi subset representatif dari populasi data yang besar.
- Denoising, adalah proses menghilangkan noise

dari data yang akan ditransformasikan

- Feature extraction, adalah proses membuka spesifikasi data yang signifikan dalam konteks tertentu.

Transformasi data diperlukan sebagai tahap pre-procecing, dimana data yang diolah siap untuk ditambang. Beberapa cara transformsi, antara lain (Santosa, 2007):

- Centering, mengurangi setiap data dengan rata-rata dari setiap atribut yang ada. • Normalisation, membagi setiap data yang dicentering dengan standar deviasi dari atribut bersangkutan.
- Scaling, mengubah data sehingga berada dalam skala tertentu.



Gambar : Tahap-tahap Knowledge Discovery in Database

3. Penambangan data (data mining)

Data-data yang telah diseleksi dan ditransformasi ditambang dengan berbagai teknik. Proses data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan fungsi-fungsi tertentu. Fungsi atau algoritma dalam data mining sangat bervariasi. Pemilihan fungsi atau algoritma yang tepat sangat bergantung pada tujuan dan proses pencaraian pengetahuan secara keseluruhan.

4. Evaluasi pola dan presentasi pengetahuan

Tahap ini merupakan bagian dari proses pencarian pengetahuan yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya. Langkah terakhir KDD adalah mempresentasikan pengetahuan dalam bentuk yang mudah dipahami oleh pengguna.

Untuk mengidentifikasi pola-pola menarik kedalam knowledge based yang ditemukan. Dalam tahap ini hasilnya berupa pola-pola yang khas maupun model prediksi dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai.

5. Presentasi pengetahuan (knowledge presentation)

Merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna.Tahap terakhir adalah bagaimana memformulasikan keputusan atau aksi dari hasil analisis yang didapat.Karenanya presentasi dalam bentuk pengetahuan yang bisa dipahami semua orang adalah satu tahapan yang diperlukan.Dalam presentasi ini, visualisasi juga bisa membantu mengkomunikasikan hasil data mining (Han, 2006).

2.4 Implementasi (Penerapan *Data Mining)*

Berikut beberapa contoh bidang penerapan data mining:

1. Analisa pasar dan manajemen.

Solusi yang dapat diselesaikan dengan data mining, diantaranya: Menembak target pasar, Melihat pola beli pemakai dari waktu ke waktu, Cross-Market analysis, Profil Customer, Identifikasi kebutuhan Customer, Menilai loyalitas Customer, Informasi Summary.

2. Analisa Perusahaan dan Manajemen resiko.

Solusi yang dapat diselesaikan dengan data mining, diantaranya: Perencanaan keuangan dan Evaluasi aset, Perencanaan sumber daya (Resource Planning), Persaingan (Competition).

3. Telekomunikasi.

Sebuah perusahaan telekomunikasi menerapkan data mining untuk melihat dari jutaan transaksi yang masuk, transaksi mana sajakah yang masih harus ditangani secara manual.

4. Keuangan.

Financial Crimes Enforcement Network di Amerika Serikat baru-baru ini menggunakan data mining untuk me-nambang trilyunan dari berbagai subyek seperti property, rekening bank dan transaksi keuangan lainnya untuk mendeteksi transaksi transaksi keuangan yang mencurigakan (seperti money laundry) .

5. Asuransi.

Australian Health Insurance Commision menggunakan data mining untuk mengidentifikasi layanan kesehatan yang sebenarnya tidak perlu tetapi tetap dilakukan oleh peserta asuransi .

6. Olahraga.

IBM Advanced Scout menggunakan data mining untuk menganalisis statistikpermainan NBA (jumlah shots blocked, assists dan fouls) dalam rangka mencapai keunggulan bersaing (competitive advantage) untuk tim New York Knicks dan Miami Heat.

7. Astronomi.

Jet Propulsion Laboratory (JPL) di Pasadena, California dan Palomar Observatory berhasil menemukan 22 quasar dengan bantuan data mining. Hal ini merupakan salah satu kesuksesan penerapan data mining di bidang astronomi dan ilmu ruang angkasa.

8. Internet Web surf-aid

IBM Surf-Aid menggunakan algoritma data mining untuk mendata akses halaman Web khususnya yang berkaitan dengan pemasaran guna melihat prilaku dan minat customer serta melihat ke- efektif-an pemasaran melalui Web.

DATA MINING

Data Mining Secara sederhana, data mining dapat diartikan sebagai proses mengekstrak atau menggali knowledge yang ada pada sekumpulan data. Informasi dan knowledge yang didapat tersebut dapat digunakan pada banyak bidang, seperti manajemen bisnis, pendidikan, kesehatan dan sebagainya. Menurut Tacbir, data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari database yang besar. Istilah data mining memiliki hakikat sebagai disiplin ilmu yang tujuan utamanya adalah untuk menemukan, menggali, atau menambang pengetahuan dari data atau informasi yang kita miliki. Proses menggali informasi dalam data mining melibatkan integrasi teknik dari berbagai disiplin ilmu, seperti teknologi database dan data warehouse, statistik, machine learning, komputasi dengan kinerja tinggi, pattern recognition, neural network, visualisasi data dan sebagainya. Data mining menggunakan pendekatan discovery-based dimana

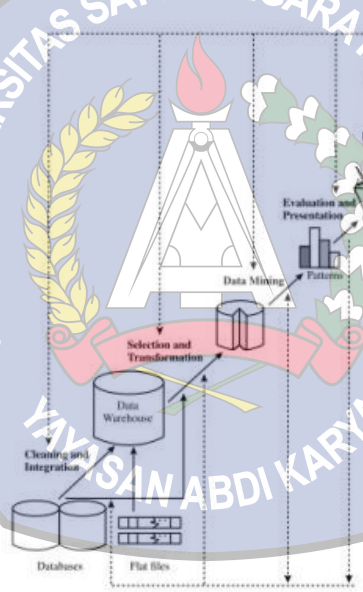
pencocokan pola (pattern matching) dan algoritma-algoritma yang lain digunakan untuk menentukan relasi-relasi kunci di dalam data yang dieksplorasi. Data mining (penambangan data), sesuai dengan namanya, berkonotasi sebagai pencarian informasi bisnis yang berharga dari basis data yang sangat besar. Dengan tersedianya basis data dalam kualitas dan ukuran yang memadai, teknologi data mining memiliki kemampuan-kemampuan sebagai berikut:

a. Mengotomatisasi prediksi trend sifat-sifat bisnis. Data mining mengotomatisasi proses pencarian informasi di dalam basis data yang besar.

b. Mengotomatisasi penemuan pola-pola yang tidak diketahui sebelumnya. Tools data mining "menyapu" basis data, kemudian mengidentifikasi pola-pola yang sebelumnya tersembunyi dalam satu sapuan. Contoh dari penemuan pola ini adalah analisis pada data penjualan ritel untuk mengidentifikasi produk-produk yang kelihatannya tidak berkaitan, yang seringkali dibeli secara bersamaan oleh customer.

a. Tahapan dalam Data Mining

Sebagai suatu rangkaian proses, data mining dapat dibagi menjadi beberapa tahap proses yang diilustrasikan pada Gambar 3 Tahap-tahap tersebut bersifat interaktif, pemakai terlibat langsung atau dengan perantaraan knowledge base.



Tahap-tahap data mining adalah sebagai berikut:

a. Pembersihan data (data cleaning) Pembersihan data

merupakan proses menghilangkan noise dan data yang tidak konsisten atau data tidak relevan.

b. Integrasi data (data integration) Integrasi data merupakan penggabungan data dari berbagai database ke dalam satu database baru.

c. Seleksi data (data selection) Data yang ada pada database sering kali tidak semuanya dipakai, oleh karena itu hanya data yang sesuai untuk dianalisis yang akan diambil dari database.

d. Transformasi data (data transformation) Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam data mining.

e. Proses mining Merupakan suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data.

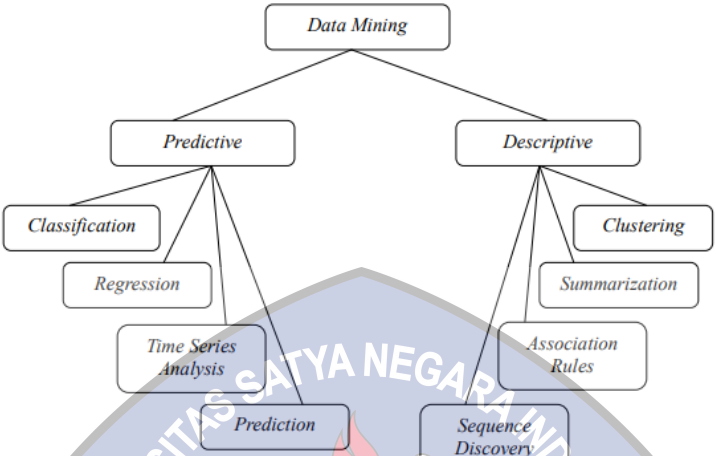
f. Evaluasi pola (pattern evaluation) Untuk mengidentifikasi pola-pola menarik ke dalam knowledge based yang ditemukan.

g. Presentasi pengetahuan (knowledge presentation) Merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna.

b. Teknik-Teknik Data mining

Data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual. Perlu diingat bahwa kata mining sendiri berarti usaha untuk mendapatkan sedikit data berharga dari sejumlah besar data dasar. Karena itu data mining sebenarnya memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (artificial intelligent), machine learning, statistik dan basis data.

Menurut Ahmed, teknik data mining biasanya terbagi dalam dua kategori, prediksi dan deskripsi. Teknik prediksi menggunakan data historis untuk menyimpulkan sesuatu tentang kejadian di masa depan. Sedangkan teknik deskripsi bertujuan untuk menemukan pola dalam data yang menyediakan beberapa informasi tentang hubungan interval yang tersembunyi.



Menurut Kumar dan Saurabh, terdapat beberapa teknik yang digunakan dalam data mining, yaitu:

1. Classification

Klasifikasi adalah teknik yang paling umum diterapkan pada data mining. Pendekatan ini sering menggunakan keputusan pohon (decision tree) atau neural network berbasis algoritma klasifikasi. Proses klasifikasi data melibatkan learning dan klasifikasi. Dalam belajar (learning) data pelatihan (training) dianalisis dengan alogaritma klasifikasi. Dalam klasifikasi pengujian data dilakukan dengan menggunakan perkiraan akurasi dari

aturan klasifikasi. Jika akurasi bisa diterima, maka aturan dapat diterapkan untuk data baru. Salah satu contoh yang mudah dan populer adalah dengan decision tree yaitu salah satu metode klasifikasi yang paling populer karena mudah untuk diinterpretasi. Decision tree adalah model prediksi menggunakan struktur pohon atau struktur berhirarki.

Decision tree adalah struktur flowchart yang menyerupai tree (pohon), dimana setiap simpul internal menandakan suatu tes pada atribut, setiap cabang merepresentasikan hasil tes, dan simpul daun merepresentasikan kelas atau distribusi kelas. Alur pada decision tree di telusuri dari simpul akar ke simpul daun yang memegang prediksi kelas untuk contoh tersebut. Decision tree mudah untuk dikonversi ke aturan klasifikasi (classification rules).

2. Clustering

Clustering bisa dikatakan sebagai identifikasi kelas objek yang memiliki kemiripan. Dengan menggunakan teknik clustering kita bisa lebih lanjut mengidentifikasi kepadatan dan jarak daerah dalam objek ruang dan dapat menemukan secara keseluruhan pola distribusi dan

korelasi antara atribut. Pendekatan klasifikasi secara efektif juga dapat digunakan untuk membedakan kelompok atau kelas objek.

3. Predication

Teknik regresi dapat disesuaikan untuk predikasi. Analisis regresi dapat digunakan untuk model hubungan antara satu atau lebih independent variables dan dependent variables. Dalam data mining independent variabel adalah atribut-atribut yang sudah dikenal dan respon variabel apa yang kita inginkan untuk diprediksi. Akan tetapi, banyak masalah di dunia nyata bukan prediksi yang mudah. Karena itu, teknik kompleks (seperti: logistic regression, decision trees atau pohon keputusan, neural nets atau jaringan syaraf) mungkin akan diperlukan untuk memprediksi nilai. Model yang berjenis sama sering dapat digunakan untuk regresi dan klasifikasi. Misalnya, CART (Classification and Regression Trees) yaitu alogaritma pohon keputusan yang dapat digunakan untuk membangun kedua pohon klasifikasi dan pohon regresi. Jaringan saraf juga dapat

menciptakan kedua model klasifikasi dan regresi.

4. Association rule

Digunakan untuk mengenali kelakuan dari kejadian-kejadian khusus atau proses dimana link asosiasi muncul pada setiap kejadian. Contoh dari aturan assosiatif dari analisa pembelian di suatu pasar swalayan adalah bisa diketahui berapa besar kemungkinan seorang pelanggan membeli roti bersamaan dengan susu. Dengan pengetahuan tersebut pemilik pasar swalayan dapat mengatur penempatan barangnya atau merancang kampanye pemasaran dengan memakai kupon diskon untuk kombinasi barang tertentu. Penting tidaknya suatu aturan assosiatif dapat diketahui dengan dua parameter, support yaitu prosentasi kombinasi atribut tersebut dalam basisdata dan confidence yaitu kuatnya hubungan antar atribut dalam aturan asosiatif. Motivasi awal pencarian association rule berasal dari keinginan untuk menganalisa data transaksi supermarket, ditinjau dari perilaku customer dalam membeli produk. Association rule ini menjelaskan seberapa sering suatu produk dibeli secara bersamaan. Sebagai contoh, association rule “beer

=> diaper (80%)" menunjukkan bahwa empat dari lima customer yang membeli beer juga membeli diaper. Dalam suatu association rule X => Y, X disebut dengan antecedent dan Y disebut dengan consequent rule.

5. Neural network

Jaringan saraf adalah seperangkat unit penghubung input dan output dimana setiap koneksinya memiliki bobot. Selama fase learning, jaringan belajar dengan menyesuaikan bobot sehingga dapat memprediksi kelas yang benar label dari setiap input. Jaringan saraf memiliki kemampuan yang luar biasa untuk memperoleh arti Komunitas 7 dari data yang rumit atau tidak tepat dan dapat digunakan untuk mengambil pola- pola serta mendeteksi tren yang sangat kompleks untuk diperhatikan baik oleh manusia atau teknik komputer lain. Jaringan saraf sangat baik untuk mengidentifikasi pola atau tren pada data dan sangat cocok untuk melakukan prediksi serta memprediksi kebutuhan.

6. Decision trees

Decision trees atau pohon keputusan adalah struktur tree-shaped yang mewakili set keputusan. Keputusan ini menghasilkan aturan untuk klasifikasi sebuah kumpulan data. Metode pohon keputusan diantaranya yaitu Classification and regression trees (CART) dan Chi Square Automatic Interaction Detection (CHAID).

7. Nearest Neighbor Method

Teknik yang mengklasifikasikan setiap record dalam sebuah kumpulan data berdasarkan sebuah kombinasi suatu kelas k record yang sama dalam sebuah kumpulan data historis (dimana k lebih besar atau sama dengan 1). Terkadang disebut juga dengan teknik K-Nearest Neighbor.

Clustering

Madhu Yedha mendefenisikan clustering sebagai proses pengorganisasian objek data ke dalam set kelas yang saling berhubungan, yang disebut cluster. Clustering merupakan contoh dari klasifikasi tanpa arahan (unsupervised). Klasifikasi merujuk kepada prosedur

yang menetapkan objek data set kelas. Unsupervised berarti bahwa pengelompokkan tidak tergantung pada standar kelas dan pelatihan atau training.

Menurut Deka, Clustering merupakan salah satu teknik data mining yang digunakan untuk mendapatkan kelompok-kelompok dari objek-objek yang mempunyai karakteristik yang umum di data yang cukup besar. Tujuan utama dari metode clustering adalah pengelompokan sejumlah data atau objek ke dalam cluster atau grup sehingga dalam setiap cluster akan berisi data yang semirip mungkin. Clustering melakukan pengelompokkan data yang didasarkan pada kesamaan antar objek, oleh karena itu klasterisasi digolongkan sebagai metode unsupervised learning. Menurut Oyelade, clustering dapat dibagi menjadi dua, yaitu hierarchical clustering dan non-hierarchical clustering.

Hierarchical clustering adalah suatu metode pengelompokan data yang dimulai dengan mengelompokkan dua atau lebih objek yang memiliki kesamaan paling dekat. Kemudian proses diteruskan ke

objek lain yang memiliki kedekatan kedua. Demikian seterusnya sehingga cluster akan membentuk semacam pohon dimana ada hierarki (tingkatan) yang jelas antar objek, dari yang paling mirip sampai yang paling tidak mirip. Secara logika semua objek pada akhirnya hanya akan membentuk sebuah cluster. Dendogram biasanya digunakan untuk membantu memperjelas proses hierarki tersebut.

Berbeda dengan metode hierarchical clustering, metode non-hierarchical clustering justru dimulai dengan menentukan terlebih dahulu jumlah cluster yang diinginkan (dua cluster, tiga cluster, atau lain sebagainya). Setelah jumlah cluster diketahui, baru proses cluster dilakukan tanpa mengikuti proses hierarki. Metode ini biasa disebut dengan K-Means Clustering.

Algoritma K-means Clustering

K-means clustering merupakan salah satu metode cluster analysis non hirarki yang berusaha untuk mempartisi objek yang ada kedalam satu atau lebih cluster atau

Komunitas kelompok objek berdasarkan karakteristiknya, sehingga objek yang mempunyai karakteristik yang sama dikelompokan dalam satu cluster yang sama dan objek yang mempunyai karakteristik yang berbeda dikelompokan kedalam cluster yang lain.

Menurut Daniel dan Eko, Langkah-langkah algoritma K-Means adalah sebagai berikut:

a. Pilih secara acak k buah data sebagai pusat cluster.

b. Jarak antara data dan pusat cluster dihitung menggunakan Euclidian Distance. Untuk menghitung jarak semua data ke setiap titik pusat cluster dapat menggunakan teori jarak Euclidean yang dirumuskan sebagai berikut: dimana: D (i,j) = Jarak data ke i ke pusat cluster j Xki = Data ke i pada atribut data ke k Xkj = Titik pusat ke j pada atribut ke k

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \ldots + (X_{ki} - X_{kj})^2}$$

c. Data ditempatkan dalam cluster yang terdekat, dihitung dari tengah cluster.

d. Pusat cluster baru akan ditentukan bila semua data telah ditetapkan dalam cluster terdekat.

e. Proses penentuan pusat cluster dan penempatan data dalam cluster diulangi sampai nilai centroid tidak berubah lagi. Berikut ini adalah contoh penerapan algoritma K-Means:

Tabel 1 Data Mahasiswa

No	Nama	Jurusan	Kota Asal	IPK
1	Ade Supryan Stefanus	IS	Jakarta	3,16
2	Adelina Ganardi Putri Hardi	ACC	Semarang	3,22
3	Adeline Dewita	BF	Bekasi	3,29
4	Adiputra	IB	Jakarta	2,83
5	Afrieska Laura Trisyana	PR	Jakarta	3,15
6	Agam Khalilullah	IB	Banda Aceh	3,25
7	Agus Mulyana Jungjungan	IB	Bogor	3,43
8	Agusman	PR	Bekasi	3,06
9	Aidil Friadi	BF	Banda Aceh	3,36
10	Ajeng Putri Ariandhani	ACC	Bandung	3,28

Transformasi Data

Agar data di atas dapat diolah dengan menggunakan metode k-means clustering, maka data yang berjenis data nominal seperti kota asal dan jurusan harus diinisialisasikan terlebih dahulu dalam bentuk angka.

Tabel 2 Inisialisasi Data Wilayah Kota Asal

Wilayah	Frekuensi	Inisial
Jakarta	84	1
Jawa Barat	82	2
Sumatera Utara	28	3
Sulawesi	14	4
Jawa Timur	13	5
Sumatera Selatan	13	6
Bali	8	7
Kalimantan	1	8

Tabel 3 Inisialisasi Data Jurusan

Jurusan	Singkatan	Frekuensi	Inisial
Accounting	ACC	46	1
Management, concentration in International Business	IB	37	2
Public Relation	PR	35	3
Management, concentration in Banking & Finance	BF	28	4
Industrial Engineering	IE	23	5
Information Technology	IT	20	6
Management, concentration in Marketing	MKT	18	7
Visual Communication Design	VCD	12	8
Management, concentration in Hotel & Tourism Management	HTM	9	9
Electrical Engineering	EE	6	10
Business Administration	BA	4	11
International Relations	IR	2	12
Management, concentration in Human Resources Management	HRM	1	13
Information System	IS	1	14
Management	MGT	1	15

Pengolahan data

Setelah semua data mahasiswa ditransformasi ke dalam bentuk angka, maka data-data tersebut telah dapat dikelompokan dengan menggunakan algoritma K-Means Clustering. Untuk dapat melakukan pengelompokan data-data tersebut menjadi beberapa cluster perlu dilakukan beberapa langkah, yaitu:

1. Tentukan jumlah cluster yang diinginkan. Dalam penelitian ini data-data yang ada akan dikelompokkan mejadi tiga cluster.

2. Tentukan titik pusat awal dari setiap cluster. Dalam penelitian ini titik pusat awal ditentukan secara random dan didapat titik pusat dari setiap cluster

Tabel 4 Titik Pusat Awal Setiap *Cluster*

Titik Pusat awal	Nama	Jurusan	Kota Asal	IPK
Cluster 1	Dally Teguh Sesario	9	3	2,94
Cluster 2	Hervina Juliana	1	1	3,18
Cluster 3	Pascal Muhammadi	1	2	3,15

3. Tempatkan setiap data pada cluster. Dalam penelitian ini digunakan metode hard k-means untuk mengalokasikan setiap data ke dalam suatu cluster, sehingga data akan dimasukan dalam suatu cluster yang memiliki jarak paling dekat dengan titik pusat dari setiap cluster. Untuk mengetahui cluster mana yang paling dekat dengan data, maka perlu dihitung jarak setiap data dengan titik pusat setiap cluster. Sebagai contoh, akan dihitung jarak dari data mahasiswa pertama ke pusat cluster pertama:

$$D(1,1) = \sqrt{(14-9)^2 + (1-3)^2 + (3{,}16-2{,}94)^2} = 5{,}390$$

Dari hasil perhitungan di atas didapatkan hasil bahwa jarak data mahasiswa pertama dengan pusat cluster pertama adalah 5,390. Jarak data mahasiswa pertama ke pusat cluster kedua:

$$D\,(1,2) = \sqrt{(14-1)^2 + (1-1)^2 + (3{,}16-3{,}18)^2} = 13{,}000$$

Dari hasil perhitungan di atas didapatkan hasil bahwa jarak data mahasiswa pertama dengan pusat cluster kedua adalah 13. Jarak data mahasiswa pertama ke pusat cluster ketiga:

$$D\,(1,3) = \sqrt{(14-1)^2 + (1-2)^2 + (3{,}16-3{,}15)^2} = 13{,}038$$

Dari hasil perhitungan di atas didapatkan hasil bahwa jarak data mahasiswa pertama dengan pusat cluster ketiga adalah 13.038. Berdasarkan hasil ketiga perhitungan di atas dapat disimpulkan bahwa jarak data mahasiswa pertama yang paling dekat adalah dengan cluster 1, sehingga data mahasiswa pertama dimasukkan ke dalam

cluster 1. Hasil perhitungan selengkapnya untuk 5 data mahasiswa pertama

Tabel 5 Contoh Hasil Perhitungan Setiap Data ke Setiap *Cluster*

No	Nama	Jurusan	Kota Asal	IPK	Jarak Ke			Jarak terdekat ke *Cluster*
					C1	C2	C3	
1	Ade Supryan Stefanus	14	1	3,16	5,390	13,000	13,038	1
2	Adelina Ganardi Putri Hardi	1	5	3,22	8,251	4,000	3,001	3
3	Adeline Dewita	4	2	3,29	5,111	3,164	3,003	3
4	Adiputra	2	1	2,83	7,281	1,059	1,450	2
5	Afrieska Laura Trisyana	3	1	3,15	6,328	2,000	2,236	2

4. Setelah semua data ditempat ke dalam cluster yang terdekat, kemudian hitung kembali pusat cluster yang baru berdasarkan rata-rata angggota yang ada pada cluster tersebut.
5. Setelah didapatkan titik pusat yang baru dari setiap cluster, lakukan kembali dari langkah ketiga hingga titik pusat dari setiap cluster tidak berubah lagi dan tidak ada lagi data yang berpindah dari satu cluster ke cluster yang lain.

Dalam penelitian ini, iterasi clustering data mahasiswa terjadi sebanyak 7 kali iterasi. Pada iterasi ke-7 ini, titik pusat dari setiap cluster sudah tidak berubah dan tidak ada lagi data yang berpindah dari satu cluster ke cluster

yang lain.

Dari hasil cluster 1, terlihat bahwa karakteristik mahasiswa pada cluster 1 didominasi oleh mahasiswa yang berasal dari jurusan Information Technology dan Marketing. Sedangkan, berdasarkan kota asal didominasi oleh mahasiswa yang berasal dari wilayah kota asal DKI Jakarta dan Jawa Barat, sehingga dapat disimpulkan bahwa ratarata mahasiswa pada cluster 1 yang berasal dari wilayah kota asal DKI Jakarta dan Jawa Barat mengambil jurusan Infromation Technology dan Marketing.

Tabel 6 Hasil Analisa *Clustering*

Hasil *Cluster* 1	Hasil *Cluster* 2	Hasil *Cluster* 3
Cluster 1 terdiri dari 70 orang, yang berasal dari jurusan IT = 19 orang MKT = 15 orang VCD = 12 orang HTM = 9 orang EE = 6 orang BA = 4 orang IR = 2 orang MGT = 1 orang IS = 1 orang HRM = 1 orang	*Cluster* 2 terdiri dari 132 orang, yang berasal dari aktifis ACC = 39 orang IB = 30 orang BF = 22 orang PR = 21 orang IE = 20 orang	*Cluster* 3 terdiri dari 41 orang, yang berasal dari jurusan: PR = 14 orang ACC = 7 orang IB = 7 orang BF = 6 orang E-3 = 3 orang MKT = 3 orang IT = 1 orang
Dan berasal dari Wilayah: DKI Jakarta = 30 orang Jawa Barat = 20 orang Sumatera Utara = 12 orang Sulawesi = 2 orang Jawa Timur = 2 orang Sumatera Selatan = 2 orang Bali = 1 orang Kalimantan = 1 orang	Dan berasal dari Wilayah: Jawa Barat = 62 orang DKI Jakarta = 54 orang Sumatera Utara = 16 orang	Dan berasal dari Wilayah: Sulawesi = 12 orang. Jawa Timur = 11 orang Sumatera Selatan = 11 orang Bali = 7 orang
Dengan rata-rata nilai IPK 3.2	Dengan rata-rata nilai IPK 3.25	Dengan rata-rata nilai IPK 3.31

Kemudian, dari hasil cluster 2 di atas dapat dilihat bahwa karakteristik mahasiswa pada cluster 2 didominasi oleh mahasiswa yang berasal dari jurusan Accounting dan International Business. Sedangkan, berdasarkan kota asal didominasi oleh mahasiswa yang berasal dari wilayah kota asal DKI Jakarta dan Jawa Barat, sehingga dapat disimpulkan bahwa rata-rata mahasiswa pada cluster 2 yang berasal dari wilayah kota asal DKI Jakarta dan Jawa Barat mengambil jurusan Infromation Technology dan

Marketing. Sedangkan, dari hasil cluster 3 di atas dapat dilihat bahwa karakteristik mahasiswa pada cluster 3 didominasi oleh mahasiswa yang berasal dari jurusan Public Relation, Accounting dan International Business. Sedangkan, berdasarkan kota asal didominasi oleh mahasiswa yang berasal dari wilayah kota asal Sulawesi, Jawa Timur dan Sumatera Selatan, sehingga dapat disimpulkan bahwa rata-rata mahasiswa pada cluster 3 yang berasal dari wilayah kota asal Sulawesi, Jawa Timur dan Sumatera Selatan mengambil jurusan Public Relation, Accounting dan International Business.

Penutup

K-means clustering merupakan salah satu metode cluster analysis non hirarki yang berusaha untuk mempartisi objek yang ada kedalam satu atau lebih cluster atau kelompok objek berdasarkan karakteristiknya, sehingga objek yang mempunyai karakteristik yang sama dikelompokan dalam satu cluster yang sama dan objek yang mempunyai karakteristik yang berbeda dikelompokan kedalam cluster yang lain. Cluster yang dihasilkan dapat memberikan pengetahuan baru dan menarik, yang dapat digunakan dalam mendukung keputusan.

Kesimpulan

Dengan menggunakan data mining, perusahaan dapat menentukan *targeted marketing* yaitu dengan melihat nama nasabah mana saja yang berpotensi membeli produk baru, perusahaan dapat melihat nasabah yang aktif dan yang paling aktif .

Kemudian dengan adanya data mining kita bisa melihat *data history* yang data tersebut dapat kita gunakan untuk melakukan *training data* dan t*esting data.*

Daftar Pustaka

Yuli Asriningtias, Rodhyah Mardhiyah Program Studi Teknik Informatika Fakultas Bisnis & Teknologi Informasi, Universitas Teknologi Yogyakarta. JURNAL INFORMATIKA Vol. 8, No. 1, Januari 2014

Muhammad Thoriq Agung, Bowo Nurhadiyono *Penerapan Data Mining Pada Data Transaksi Penjualan Untuk Mengatur Penempatan Barang Menggunakan Algoritma Apriori.* Semarang: Program Studi Teknik Informatika-S1, Fakultas Ilmu Komputer Universitas Dian Nuswantoro.

Kennedi Tampubolon 1), Hoga Saragih 2), Bobby Reza 3). 2013. Implementasi Data Mining Algoritma Apriori Pada Sitem Persedian Alat-Alat Kesehatan. Medan: Majalah Ilmiah.

http://repository.widyatama.ac.id/xmlui/bitstream/handle/123456789/2362/bab

%202%20landasan%20teori.pdf?sequence=4 , diakses

November 2017

http://www.gunadarma.ac.id/library/articles/postgraduate/information-system/Sistem

%20Informasi%20Akuntansi/Artikel_92106032.pdf,

diakses pada November 2017

DATA WAREHOUSE

BIG DATA

BAB 1
PENDAHULUAN

1.1. Latar Belakang

Big data adalah kumpulan proses yang terdiri volume data dalam jumlah besar yang terstruktur maupun tidak terstruktur dan digunakan untuk membantu kegiatan bisnis. Big data sendiri merupakan pengembangan dari sistem database pada umumnya. Yang membedakan disini adalah proses kecepatan, volume, dan jenis data yang tersedia lebih banyak dan bervariatif daripada DBMS (Database Management System) pada umumnya.

Misalnya dari facebook, maka kita dapat chatingan, bahkan kita dapat melihat aktivitas seseorang. Dengan menggunakan salah satu Banyak orang membutuhkan pengolahan Big Data, antara lain untuk mengetahui topik yang sedang hangat saat ini di Twitter, mencari teman lama secara cepat melalui Facebook, dan lain-lain. Perusahaan perlu mengolah Big Data untuk pengambilan keputusan bisnis yang harus cepat. Misal, untuk mengetahui kebiasaan dan kesukaan pelanggan tanpa harus bertanya, mengetahui selera pembaca portal berita di web, dan sebagainya.

Data warehouse merupakan perkembangan dari konsep database yang

menyediakan suatu sumber data yang lebih baik bagi para user dan memungkinkan user untuk memanipulasi dan menggunakan data tersebut secara intuitif. Dikutip dari jurnal Al-Debei (2011:164), data warehouse sangat dikenal sebagai sebuah infrastruktur, beberapa aplikasi bisa dijalankan dalam data warehouse seperti CRM dan DSS. Disisi lain, beberapa teknik yang bisa digunakan untuk ekstraksi business intelligence dalam data warehouse seperti data mining, OLAP dan dashboard. Dapat disimpulkan bahwa data warehouse merupakan sebuah arsitektur data yang digunakan untuk menyediakan kebutuhan informasi yang diperlukan dalam mendukung proses analisis data dan pengambilan keputusan.

1.2. Rumusan Masalah

1. Pengertian Big Data
2. Pengertian Data Warehose
3. Fungi Big Data
4. Konsep Data Warehouse
5. Manfaat Big Data
6. Karakteristik Data Warehosue
7. Klasifikasi Big Data
8. Struktur & Data Flow dalam Data Warehouse
9. Contoh Penggunaan Aplikasi Big Data
10. Sketsa Data Warehouse
11. Metodologi Perancangan Data Warehouse
12. Keuntungan Penggunaan Data Warehouse

BAB 2
PENDAHULUAN

2.1. Pengertian Big Data

Big data adalah kumpulan proses yang terdiri volume data dalam jumlah besar yang terstruktur maupun tidak terstruktur dan digunakan untuk membantu kegiatan bisnis. Big data sendiri merupakan pengembangan dari sistem database pada umumnya. Yang membedakan disini adalah proses kecepatan, volume, dan jenis data yang tersedia lebih banyak dan bervariatif daripada DBMS (Database Management System) pada umumnya.

Definisi dari big data juga dapat dibagi menjadi 3 bagian, yang biasa disebut dengan 5V

1. Volume

Ukuran data yang dimiliki oleh big data memiliki kapasitas yang besar. Anda dapat mencoba melakukan proses data dengan ukuran yang besar untuk dijalankan.

2. Velocity

Kecepatan transfer data juga sangat berpengaruh dalam proses pengiriman data dengan efektif dan stabil. Big data memiliki kecepatan yang memungkinkan untuk dapat diterima secara langsung (real-time). Kecepatan tertinggi yang bisa didapatkan langsung melalui aliran data ke

memori apabila dibandingkan dengan yang ditulis pada sebuah disk.

3. Variety

Jenis variasi data yang dimiliki oleh big data lebih banyak daripada menggunakan sistem database SQL. Jenis data yang masih bersifat tradisional, lebih terstruktur daripada data yang belum terstruktur. Contohnya adalah teks, audio, dan video merupakan data yang belum terdefinisikan secara langsung dan harus melalui beberapa tahap untuk dapat diproses dalam sebuah database.

4. Value

Sama seperti namanya, value adalah nilai dari suatu data setelah data tersebut berhasil diproses. Sebuah data akan diklaim memiliki nilai ketika informasi yang didapatkan dari pengolahan dapat dibantu dalam hal mengambil keputusan bisnis yang lebih baik.

5. Veracity

Veracity atau yang umumnya lebih dikenal dengan kebenaran data adalah suatu tingkat akurasi dari informasi yang diberikan oleh sebuah data set. Berdasarkan dengan tingkat kebenaran data yang baik, maka keputusan atau kebijakan yang diambil ketika mengolah data tersebut akan mampu memberikan hasil yang lebih baik.

Selain dari 3V diatas, masih ada 2V lain yang merupakan bagian dari big data sendiri. Yaitu Value dan Veracity. Untuk value, merupakan nilai atau aliran data yang tidak teratur dan konsisten dalam beberapa kondisi dan periode. Hal tersebut dapat terjadi pada suatu kasus dimana terdapat lonjakan data yang besar sehingga, akan memproses data dengan resource memori yang lebih besar.

Veracity merupakan bentuk pembenaran suatu data. Jadi, mengacu pada kualitas data tersebut, dapat berasal dari berbagai sumber. Perlu adanya proses untuk menghubungkan dan mengkorelasikan beberapa hubungan data. Jika tidak ada relasi yang baik, maka dapat menimbulkan kontrol yang lepas kendali.

2.2. Pengertian Data Warehouse

Pengertian Data Warehouse dapat bermacam-macam namun mempunyai inti yang sama, seperti pendapat beberapa ahli berikut ini :

Menurut M. Klimavicius (2008, p85) , sistem data warehouse mempresentasikan sebuah sumber informasi untuk menganalisa pengembangan dan hasil dari sebuah perusahaan atau organisasi didalam lingkungan yang selalu berubah. Data didalam data warehouse menggambarkan peristiwa dan status dari proses bisnis, produk dan jasa, tujuan dan unit-unit organisasi.

Data warehouse merupakan metode dalam perancangan database, yang menunjang DSS(Decission Support System) dan EIS (Executive Information System). Secara fisik data warehouse adalah database, tapi perancangan data warehouse dan database sangat berbeda. Dalam perancangan database tradisional menggunakan normalisasi, sedangkan pada data warehouse normalisasi bukanlah cara yang terbaik.

2.3. Fungsi Big Data

Big data memiliki beberapa fungsi penting dalam proses pengembangan dan penyempurnaan sebuah aplikasi. Berikut ini merupakan beberapa fungsi terkait dengan big data:

1. Dapat menentukan penyebab suatu masalah, kegagalan secara real time

Fungsi pertama dari big data adalah menentukan dan menganalisa penyebab dari suatu permasalahan yang terjadi di dalam sistem. Kemudian, dengan penggunaannya saat ini, juga dapat meminimalisir terjadinya kegagalan dalam proses penyimpanan data. Untuk hasil analisis tersebut dapat ditampilkan secara real-time.

2. Pengambilan sebuah keputusan yang cerdas dan tepat

Big data juga dapat digabungkan dengan sistem dan perangkat teknologi cerdas seperti IoT (Internet of Things) dan AI (Artificial Intelligence). Tugasnya adalah untuk memberikan dan menyimpan data dan informasi yang dibutuhkan dalam pengembangan sebuah produk. Misalnya saja smart city atau kota cerdas yang menggunakan bantuan kecerdasan buatan dan jaringan internet berskala besar yang mampu untuk menghubungkan tiap sudut kota, bangunan, dan infrastruktur pendukung lain.

3. Mendeteksi sebuah anomali atau perilaku yang menyimpang dalam struktur bisnis anda

Fungsi yang ketiga adalah mampu untuk mendeteksi secara cepat dan tepat, bentuk atau proses kegiatan yang menyimpang dan berhenti karena ada kesalahan dari sisi teknis maupun non teknis. Big data juga dapat merencanakan beberapa opsi untuk mengurangi dan mengatasi anomali tersebut dengan lebih cepat untuk membantu aktivitas bisnis perusahaan atau organisasi anda.

4. Mengurangi biaya, waktu, dan meningkatkan performa produk aplikasi

Penyimpanan data dengan menggunakan sistem big data juga dapat mengurangi biaya yang harus dikeluarkan oleh perusahaan. Kemudian, waktu dalam mengelola dan menjalankan sebuah operasi menjadi lebih cepat dengan transfer data diatas rata – rata sistem database lain.

Peningkatan performa juga menjadi kelebihan tersendiri untuk mendukung pengembangan perangkat lunak.

Tahapan Pengelolaan Big Data

Berikut ini adalah 4 tahap pengelolaan Big Data serta perangkat bantu (tools) yang dapat dimanfaatkan untuk mendukung pemrosesan pada tiap tahap:

a. Acquired

Berhubungan dengan sumber dan cara mendapatkan data.

b. Accessed

Berhubungan dengan daya akses data; data yang sudah dikumpulkan memerlukan tata kelola, integrasi, storage dan computing agar dapat dikelola untuk tahap berikutnya. Perangkat untuk pemrosesan (processing tools) menggunakan Hadoop, Nvidia CUDA, Twitter Storm, dan GraphLab. Sedangkan untuk manajemen penyimpanan data (storage tools) menggunakan Neo4J, Titan, dan HDFS.

c. Analytic

Berhubungan dengan informasi yang akan didapatkan, hasil pengelolaan data yang telah diproses. Analitik yang dilakukan dapat berupa descriptive (penggambaran data), diagnostic (mencari sebab akibat berdasar data), predictive (memprediksi kejadian dimasa depan) maupun

prescriptive analytics (merekomendasikan pilihan dan implikasi dari setiap opsi). Tools untuk tahap analitik menggunakan MLPACK dan Mahout.

d. Application

Terkait visualisasi dan reporting hasil dari analitik. Tools untuk tahap ini menggunakan RStudio.

2.4. Konsep Data Warehouse

- Dibanding database tradisional, DW umumnya terdiri dari data yang berukuran sangat besar dari banyak sumber dan mungkin terdiri dari database dari model data yang berbeda dan kadang file dari sistem dan platform yang independen

- Tidak seperti database transaksional, DW biasanya mendukung analisa tren dan time-series, di mana keduanya membutuhkan data historik

- DW itu nonvolatile. Artinya informasi dalam DW jarang diubah dan bisa dianggap non-real-time

- DW bisa digambarkan sebagai "kumpulan teknologi pendukung keputusan, dimaksudkan untuk memungkinkan pekerja yang berhubungan dengan informasi (eksekutif, manajer dan analis) untuk membuat keputusan lebih baik dan lebih cepat"

2.5. Manfaat Big Data

Banyak sekali kemudahan yang dimiliki oleh big data dari sisi fungsionalitas dan fitur yang dimiliki. Kami membagi menjadi 2 manfaat untuk kebutuhan dalam proses teknologi informasi (TI) dan bisnis.

1. Bidang Teknologi Informasi

Penggunaan Perangkat Mobile

Saat ini, penggunaan perangkat seperti smartphone, tablet, IPOD, dll. Sering kita jumpai karena telah mendukung dan support dengan berbagai sistem aplikasi. Kemudian, faktor user friendly dan mudah dibawa kemana saja merupakan kelebihan utama dari perangkat mobile.

Penggunaannya juga digunakan untuk pengembangan perangkat mobile saat ini. Salah satu contoh yang mudah untuk anda lihat adalah aplikasi GPS yang tentunya sudah terinstall dalam perangkat anda. GPS (Global Positioning System) yang dimiliki oleh Google Maps menggunakan bantuan dari big data dalam memproses dan memanajemen berbagai bentuk data.

Karena, sistem database yang dibutuhkan sangat besar. Mulai dari gambar, pemetaan lokasi hingga dapat menjangkau hampir seluruh penjuru dunia

sekaligus. Tentunya membutuhkan sebuah basis data dengan kapasitas yang sangat besar. Sekarang, Google telah menerapkan sistem penyimpanan berbasis cloud (awan). Sehingga, penyimpanan dapat dilakukan secara online dan real-time dengan kapasitas yang lebih besar lagi.

Media Sosial

Hampir setiap orang menggunakan yang namanya media sosial untuk mengakses berbagai informasi dan membagikan aktivitas keseharian pribadi. Tentunya, banyak yang mengupload foto, video maupun teks ke dalam aplikasi media sosial tersebut. Semua informasi tersebut merupakan jenis data yang akan terekam dan tersimpan dalam sistem basis data dengan kapasitas besar.

Anda bisa membayangkan berapa ukuran yang harus dialokasi oleh media sosial seperti Facebook, Twitter, Instagram, dll. Untuk menampung data setiap harinya. Solusi terbaik untuk mengatasi masalah tersebut adalah dengan menggunakan big data yang memiliki performa yang baik dalam penanganan data dengan skala besar.

Perangkat Cerdas

Sistem cerdas untuk saat ini banyak dikembangkan oleh negara – negara maju seperti China, Jepang, Amerika, dll. Salah satu manfaat yang dimiliki oleh perangkat cerdas adalah

mampu membantu kegiatan manusia dengan lebih efektif dan efisien.

Contoh dari perangkat cerdas adalah teknologi IoT yang saat ini banyak diterapkan pada perangkat elektronik seperti kulkas, mesin cuci, AC, dan lain sebagainya. Dengan menggunakan sistem yang telah

terintegrasi dengan jaringan internet, maka segala bentuk aktivitas dapat dikoordinir dalam satu sistem aplikasi saja dengan bantuan big data sebagai penyedia layanan informasi dan penyimpanan data.

Media Digital

Selanjutnya, big data juga mempengaruhi dari sisi penggunaan media digital. Contohnya, penggunaan fitur pada website dan aplikasi streaming seperti spotify dan netflix. Dalam sistem basis data yang mereka gunakan, mampu mencatat data musik, film yang telah anda tonton dan memberikan sebuah rekomendasi untuk anda.

Dengan bantuan teknologi AI sendiri, basis data dapat terintegrasi dengan baik dan cepat untuk memberikan kemudahan dalam penggunaan aplikasi tersebut. Contoh lain dari penerapan media digital adalah fitur pada e-commerce yang telah menerapkan AI dengan big data untuk memudahkan pengguna dalam memberikan rekomendasi produk.

2. Bidang Bisnis

Meningkatkan sistem operasional bisnis

Untuk meningkatkan produktivitas dan efektivitas bisnis yang anda rintis, tentu memerlukan sumber daya yang memadai. Salah satunya akan kebutuhan data yang terus meningkat. Big data dapat mengatasi permasalahan data dengan kebutuhan yang besar untuk membantu proses operasional bisnis anda.

Customer Relationship Management (CRM)

Anda perlu menjaga dan meningkatkan hubungan baik dengan pelanggan dan sales. Dengan melakukan management menggunakan beberapa fitur tambahan untuk membantu anda dalam memonitoring kegiatan penjualan, menghitung rata – rata konversi, dan lain sebagainya.

Mengoptimalkan pengalaman penggunaan aplikasi

Penggunaan perangkat mobile terus meningkat, sehingga perlu adanya optimasi dari sisi software dan

hardware. Selain itu, penyimpanan data juga sangat berpengaruh terhadap optimalisasi sebuah aplikasi. Dengan big data, proses transfer dan manajemen data berjalan lebih cepat dan akurat.

6.Pemanfaatan Big Data Pada Sektor Layanan Publik

Perusahaan atau institusi yang yang berada pada sektor layanan publik lazimnya memiliki orientasi utama pada pencapaian kepuasan klien/ pelanggan. Resource Big Data dapat memberikan andil dengan menyajikan berbagai informasi berharga sebagai berikut:

a.Mendapatkan feedback dan respon masyarakat sebagai dasar penyusunan kebijakan dan perbaikan pelayanan publik. Feedback tersebut dapat diperoleh dari sistem informasi layanan pemerintah maupun dari media sosial.

b.Membuat layanan terpadu dengan segmen khusus sehingga layanan bisa lebih efektif dan efisien.

c.Menemukan solusi atas permasalahan yang ada, berdasarkan data. Sebagai contoh: menganalisis informasi cuaca dan informasi pertanian terkait data tingkat kesuburan tanah, pemerintah dapat menetapkan atau menghimbau jenis varietas tanaman yang ditanam oleh petani pada daerah dan waktu tertentu.

Jenis-jenis Big Data

enis Big Data terbagi tiga yang memiliki fungsi, bentuk, dan teknik pemrosesan yang berbeda-beda. Simak lebih lanjut di bawah ini.

1. Data Terstruktur

Data Terstruktur merujuk pada data yang sudah tersimpan secara berurutan. Secara umum, data

ini disusun pada excel atau spreadsheet. Data terstruktur mudah diakses dan dianalisis karena berasal dari berbagai macam database dengan algoritma mesin pencari sederhana. Bisa juga berasal dari data statistik lain yang ditangkap oleh server, aplikasi, atau bergerak melalui platform.

Data terstruktur yang dibuat manusia biasanya mencakup semua data dalam aktivitas di internet atau komputer. Misalnya, ketika seseorang mengklik tautan di internet, atau bahkan berselancar ke suatu situs e-commerce, aktivitas tersebut menjadi data yang dapat digunakan oleh perusahaan untuk mengetahui behaviour pelanggan. Contoh data terstruktur lain, seperti data penjualan perusahaan, data diri karyawan, atau data pelanggan yang terlampir secara terstruktur.

2. Data Tidak Terstruktur

Data dengan jenis ini bentuknya tidak terstruktur dan tidak memiliki format yang jelas dalam penyimpanan. Sehingga, tidak mudah membaca dan menganalisis data ini. Biasanya, data ini memiliki volume atau ukuran data yang besar. Untuk menganalisisnya, perlu diolah terlebih dahulu secara manual.

Data yang tidak terstruktur ini bisa berasal dari beberapa sumber dan memiliki kombinasi file sederhana, seperti teks, gambar, video, dan lain sebagainya. Contohnya dalam media sosial, seperti jumlah like, pengikut, komentar, retweet,

share, gambar postingan, dan aktivitas digital lain dalam akun pengguna. Contoh lain, data tidak terstruktur biasanya dihasilkan mesin, seperti citra satelit, data ilmiah dari berbagai eksperimen, atau data radar yang ditangkap oleh berbagai aspek teknologi.

3. Data Semi Terstruktur

Dalam data ini, garis antara data tidak terstruktur dan data terstruktur tampak tidak jelas, karena sebagian besar data semi-terstruktur sekilas tampak tidak terstruktur. Jenis data ini belum diklasifikasikan, tapi mengandung informasi yang penting.Misalnya, dokumen NoSQL karena mengandung kata kunci yang dapat digunakan untuk memproses dokumen dengan mudah. Contoh file yang masuk ke dalam jenis data ini adalah xml, json, dan CSV.

Pemicu Perkembangan Big Data

Menurut Hilbert dan Lopez, ada tiga hal utama yang memicu perkembangan teknologi Big Data:

a.Pesatnya pertambahan kemampuan penyimpanan data, kemampuan penyimpanan data telah bertumbuh sangat signifikan.

b.Pesatnya pertambahan kemampuan mesin pemrosesan data, seiring dengan pesatnya perkembangan teknologi hardware, maka kapasitas komputasi pada mesin/ perangkat komputer juga telah meningkat sangat tajam.

c.Ketersediaan data yang melimpah, Perusahaan-perusahaan dari berbagai sektor di Amerika Serikat memiliki data setidaknya 100 terabytes. Bahkan banyak diantara perusahaan tersebut yang memiliki data lebih dari 1 petabyte.

2.6. Karakteristik Data Warehouse

- Subject-Oriented

Data warehouse berorientasi subject artinya data warehouse didesain untuk menganalisa data berdasarkan subject-subject tertentu dalam organisasi,bukan pada proses atau fungsi aplikasi tertentu. Data warehouse diorganisasikan disekitar subjek-subjek utama dari perusahaan (customers, products dan sales) dan tidak diorganisasikan pada area-area aplikasi utama (customer invoicing,stock control dan product sales). Hal ini dikarenakan kebutuhan dari data warehouse untuk menyimpan data-data yang bersifat sebagai penunjang suatu keputusan, dari pada aplikasi yang berorientasi terhadap data.

- Integrated

Data Warehouse dapat menyimpan data-data yang berasal dari sumber- sumber yang terpisah kedalam suatu format yang konsisten dan saling terintegrasi satu dengan lainnya. Dengan demikian data tidak bisa dipecah-pecah karena data yang ada merupakan suatu kesatuan yang

menunjang keseluruhan konsep data warehouse itu sendiri.

Syarat integrasi sumber data dapat dipenuhi dengan berbagai cara sepeti konsisten dalam penamaan variable,konsisten dalam ukuran variable,konsisten dalam struktur pengkodean dan konsisten dalam atribut fisik dari data.

Contoh pada lingkungan operasional terdapat berbagai macam aplikasi yang mungkin pula dibuat oleh developer yang berbeda. Oleh karena itu, mungkin dalam aplikasi-aplikasi tersebut ada variable yang memiliki maksud yang sama tetapi nama dan format nya berbeda. Variable tersebut harus dikonversi menjadi nama yang sama dan format yang disepakati bersama. Dengan demikian tidak ada lagi kerancuan karena perbedaan nama, format dan lain sebagainya. Barulah data tersebut bisa dikategorikan sebagai data yang terintegrasi karena kekonsistenannya.

- Time-Variant

Seluruh data pada data warehouse dapat dikatakan akurat atau valid pada rentang waktu tertentu. Untuk melihat interval waktu yang digunakan dalam mengukur keakuratan suatu data warehouse, kita dapat menggunakan cara antara lain :

1. Cara yang paling sederhana adalah menyajikan data warehouse pada rentang waktu

tertentu, misalnya antara 5 sampai 10 tahun ke depan.

2. Cara yang kedua, dengan menggunakan variasi/perbedaan waktu yang disajikan dalam data warehouse baik implicit maupun explicit secara explicit dengan unsur waktu dalam hari, minggu, bulan dsb. Secara implicit misalnya pada saat data tersebut diduplikasi pada setiap akhir bulan, atau per tiga bulan. Unsur waktu akan tetap ada secara implisit didalam data tersebut.

3. Cara yang ketiga,variasi waktu yang disajikan data warehouse melalui serangkaian snapshot yang panjang. Snapshot merupakan tampilan dari sebagian data tertentu sesuai keinginan pemakai dari keseluruhan data yang ada bersifat read-only.

- Nonvolatile

Karakteristik keempat dari data warehouse adalah non-volatile,maksudnya data pada data warehouse tidak di-update secara real time tetapi di refresh dari sistem operasional secara reguler. Data yang baru selalu ditambahkan sebagai suplemen bagi database itu sendiri dari pada sebagai sebuah perubahan. Database tersebut secara kontinyu menyerap data baru ini, kemudian secara incremental disatukan dengan data sebelumnya.

Berbeda dengan database operasional yang dapat melakukan update,insert dan delete terhadap data

yang mengubah isi dari database sedangkan pada data warehouse hanya ada dua kegiatan memanipulasi data yaitu loading data (mengambil data) dan akses data (mengakses data warehouse seperti melakukan query atau menampilan laporan yang dibutuhkan, tidak ada kegiatan updating data).

2.7. Kasifikasi Big Data

Jenis Big Data yang bermacam-macam tersebut kemudian diklasifikasikan menjadi dua, yaitu Big Data Operasional dan Big Data Analytic. Masing-masing dikelompokkan berdasarkan beban kerja, yang keduanya memiliki kebutuhan sistem berlawanan satu sama lain. Untuk memahami klasifikasi teknologi Big Data lebih lanjut, simak penjelasan di bawah ini:

1. Big Data Operasional

Big Data operasional merupakan sistem yang memiliki kapabilitas operasional untuk pekerjaan-pekerjaan bersifat interaktif dan real time. Secara umum, data pada kelompok ini disimpan. Untuk menanganinya, dibangun sistem dengan database NoSQL.

Teknologi NoSQL dikenal lebih cepat, mudah, dan lebih murah. NoSQL dengan komputasi awan menjadi perangkat kerja operasional Big Data yang mudah dikelola dan bisa di implementasikan lebih cepat.

2. Big Data Analytic

Pekerjaan yang berhubungan dengan klasifikasi Big Data ini diimplementasikan dengan sistem database MPP dan MapReduce. Teknologi ini muncul sebagai reaksi keterbatasan dan kurangnya kemampuan relational database tradisional untuk mengelola lebih dari satu server. Selain itu, MapReduce ini juga menawarkan metode baru yang mampu menganalisis data yang fungsinya sebagai pelengkap.

Contoh dari Big Data Analytic

Contoh perusahaan yang menggunakan analisis Big Data

Starbucks (Memperkenalkan Produk Coffee Baru). Pagi itu kopi itumulai dipasarkan, pihak Starbucks memantau melalui blog, Twitter, dan kelompok forum diskusi kopi lainnya untuk menilai reaksi pelanggan. Pada pertengahan-pagi, Starbucks menemukan hasil dari analisis Big Data bahwa meskipun orang menyukai rasa kopi tersebut, tetapi mereka berpikir bahwa harga kopi tersebut terlalu mahal. Maka dengan segera pihak Starbucks menurunkan harga, dan menjelang akhir hari semua komentar negatif telah menghilang. Bagaimana jika menggunakan analisis tradisional?

• Contoh tersebut menggambarkan penggunaan sumber data yang berbeda dari Big Data dan

berbagai jenis analisis yang dapat dilakukan dengan respon sangat cepat oleh pihak Starbucks.

a.Data terstruktur

Kelompok data yang memiliki tipe data, format, dan struktur yang telah terdefinisi. Sumber datanya dapat berupa data transaksional, OLAP data, tradisional RDBMS, file CSV, spreadsheets

b.Data tidak terstruktur

Kelompok data tekstual dengan format tidak menentu atau tidak memiliki struktur melekat, sehingga untuk menjadikannya data terstruktur membutuhkan usaha, tools, dan waktu yang lebih. Data ini dihasilkan oleh aplikasi-aplikasi internet, seperti data URL log, media sosial, e-mail, blog, video, dan audio.

2.8. Struktur & Data Flow dalam Data Warehouse

- Current Detail Data

Current detail data merupakan data detil yang aktif saat ini,mencerminkan keadaan yang sedang berjalan dan merupakan level terendah dalam data warehouse. Didalam area ini warehouse menyimpan seluruh detail data yang terdapat pada skema basis data. Jumlah data sangat besar sehingga memerlukan storage yang besar pula dan dapat diakses secara cepat. Dampak negatif yang ditimbulkan adalah kerumitan untuk mengatur data menjadi meningkat dan biaya yang diperlukan menjadi mahal.

Berikut ini beberapa alasan mengapa current detail data menjadi perhatian utama :

1. Menggambarkan kejadian yang baru terjadi dan selalu menjadi perhatian utama
2. Sangat banyak jumlahnya dan disimpan pada tingkat penyimpanan terendah.
3. Hampir selalu disimpan dalam storage karena cepat di akses tetapi mahal dan kompleks dalam pengaturannya.
4. Bisa digunakan dalam membuat rekapitulasi sehingga current detail data harus akurat.

• Older Detail Data

Data ini merupakan data historis dari current detail data, dapat berupa hasil cadangan atau archive data yang disimpan dalam storage terpisah. Karena bersifat back-up (cadangan), maka biasanya data disimpan dalam storage alternatif seperti tape-desk.

Data ini biasanya memilki tingkat frekuensi akses yang rendah. Penyusunan file atau directory dari data ini di susun berdasarkan umur dari data yang bertujuan mempermudah untuk pencarian atau pengaksesan kembali.

• Lightly Summarized Data

Data ini merupakan ringkasan atau rangkuman dari current detail data. Data ini dirangkum berdasar periode atau dimensi lainnya sesuai dengan kebutuhan.

Ringkasan dari current detail data belum bersifat total summary.Data-data ini memiliki detil tingkatan yang lebih tinggi dan mendukung kebutuhan warehouse pada tingkat departemen. Tingkatan data ini di sebut juga dengan data mart. Akses terhadap data jenis ini banyak digunakan untuk view suatu kondisi yang sedang atau sudah berjalan.

- Highly Summarized Data

Data ini merupakan tingkat lanjutan dari Lightly summarized data, merupakan hasil ringkasan yang bersifat totalitas, dapat di akses misal untuk melakukan analisis perbandingan data berdasarkan urutan waktu tertentu dan analisis menggunakan data multidimensi.

- Metadata

Metadata bukan merupakan data hasil kegiatan seperti keempat jenis data diatas. Menurut Poe, metadata adalah 'data tentang data' dan menyediakan informasi tentang struktur data dan hubungan antara struktur data di dalam atau antara storage(tempat penyimpanan data).

Metadata berisikan data yang menyimpan proses perpindahan data meliputi database structure, contents, detail data dan summary data, matrics, versioning, aging criteria,versioning, transformation criteria. Metadata khusus dan memegang peranan yang sangat penting dalam data warehouse.

Metadata sendiri mengandung :

o Struktur data

Sebuah direktori yang membantu user untuk melakukan analisis Decission Support System dalam pencarian letak/lokasi dalam data warehouse.

o Algoritma

Algoritma digunakan untuk summary data. Metadata sendiri merupakan panduan untuk algoritma dalam melakukan pemrosesan summary data antara current detail data dengan lightly summarized data dan antara lightly summarized data dengan hightly summaried data.

o Mapping

Sebagai panduan pemetaan(mapping) data pada saat data di transform/diubah dari lingkup operasional menjadi lingkup data warehouse.

Data Flow

Menurut Conolly dan Begg (2005,p1161-1165), data warehouse fokus pada managemen lima arus data primer, yaitu :

- **Inflow**
  Proses yang berhubungan dengan pengekstrakan (extraction), pembersihan (cleansing), dan pemuatan (loading) data dari sumber data ke dalam data warehouse.

- **Upflow**
  Proses yang terhubung dengan menambahkan nilai ke data di dalam warehouse, melalui peringkasan, pemadatan, dan pendistribusian data.
- **Downflow**
  Proses yang berhubungan dengan penyimpanan dan backup data dalam data warehouse.
- **Outflow**
  Proses yang berhubungan dengan membuat data tersedia agar tersedia bagi end user.
- **Metaflow**
  Proses managemen metadata. Metaflow merupakan proses yang memindahkan metadata (data tentang flow yang lainnya).

2.9. Contoh Penggunaan serta Aplikasi Big Data

Beberapa contoh pemanfaatannya yang banyak digunakan di dalam kehidupan sehari-hari adalah sebagai berikut ini:

1. Penggunaan Internet

Saat ini, hampir seluruh orang di seluruh dunia sudah terkoneksi dengan internet di setiap harinya. Tentunya kita juga sudah sering menggunakan Google untuk mencari informasi. Berbagai data dari hasil pencarian ini adalah data yang nantinya akan diambil oleh Google.

2. Penggunaan Smartphone

Serupa dengan internet, hampir semua orang sudah memiliki smartphone. Namun tahukah Anda, bila smartphone mempunyai sejumlah data yang sangat besar. Smartphone akan menyimpan dan juga merekam sms serta telepon yang sudah Anda lakukan.

Selain itu, berbagai aplikasi yang Anda gunakan pun akan mengumpulkan berbagai data untuk kebutuhan bisnis Anda. Berbagai aplikasi GPS seperti aplikasi Google Maps dan juga Waze pun akan mengumpulkan berbagai data yang berhubungan dengan lokasi Anda saat itu.

3. Media Sosial

Media sosial saat ini sudah menjadi suatu hal yang sangat melekat dengan kegiatan sehari-hari manusia. Berbagai foto dan juga video yang diunggah ke media sosial setiap hari adalah salah satu bagian dari data.

4. Smart Devices

Berbagai peralatan rumah tangga seperti smart fridges, smart TV, bahkan yang baru-baru ini adalah smart car adalah konsep smart appliance. Konsep seperti inilah yang akan membawa seluruh peralatan rumah tangga Anda bisa terhubung satu sama lain dan Anda juga bisa mengaturnya dari satu alat tertentu, seperti smartphone.

Jadi, seluruh data yang berasal dari smart devices yang Anda miliki, seperti konsumsi daya yang digunakan, temperatur di rumah akan dihimpun

oleh produsen agar bisa terus memperbaiki layanannya dan juga akan menawarkan teknologi terbarunya untuk Anda.

5. Digitalisasi Media

Sebelum kehadiran internet, banyak orang yang masih menggunakan DVD dan juga CD guna menonton suatu video film atau mendengarkan musik. Sehingga, Anda tidak akan meninggalkan jejak digital. Tapi, dengan adanya aplikasi nonton dan musik streaming seperti aplikasi Netflix dan Spotify, Anda bisa menonton film dan juga mendengarkan musik dalam aplikasi tersebut. Tentunya, aplikasi tersebut nantinya akan mencatat apa saja yang Anda tonton dan Anda dengarkan, sehingga mereka akan bisa mempunyai data yang digunakan untuk bisa meningkatkan layanan terbaik mereka.

Membangun Big Data Platform

Seperti data pergudangan, toko web atau platform TI, infrastruktur untuk data yang besar memiliki kebutuhan yang unik. Dalam mempertimbangkan semua komponen platform data yang besar, penting untuk diingat bahwa tujuan akhir adalah untuk dengan mudah mengintegrasikan data yang besar dengan data perusahaan Anda untuk memungkinkan Anda untuk melakukan analisis mendalam pada set data gabungan. Requirement dalam big data infrastruktur: (1) data acquisition, (2) data organization (3) data analysis

a.Data acquisition

Tahap akuisisi adalah salah satu perubahan besar dalam infrastruktur pada hari-hari sebelum big data. Karena big data mengacu pada aliran data dengan kecepatan yang lebih tinggi dan ragam yang bervariasi, infrastruktur yang diperlukan untuk mendukung akuisisi data yang besar harus disampaikan secara perlahan, dapat diprediksi baik di dalam menangkap data dan dalam memprosesnya secara cepat dan sederhana, dapat menangani volume transaksi yang sangat tinggi , sering dalam lingkungan terdistribusi, dan dukungan yang fleksibel, struktur data dinamis.

Database NoSQL sering digunakan untuk mengambil dan menyimpan big data. Mereka cocok untuk struktur data dinamis dan sangat terukur. Data yang disimpan dalam database NoSQL biasanya dari berbagai variasi/ragam karena sistem dimaksudkan untuk hanya menangkap semua data tanpa mengelompokkan dan parsing data.

Sebagai contoh, database NoSQL sering digunakan untuk mengumpulkan dan menyimpan data media sosial. Ketika aplikasi yang digunakan pelanggan sering berubah, struktur penyimpanan dibuat tetap sederhana. Alih-alih merancang skema dengan hubungan antar entitas, struktur sederhana sering hanya berisi kunci utama untuk mengidentifikasi titik data, dan kemudian wadah konten memegang data yang relevan. Struktur sederhana dan dinamis ini memungkinkan

perubahan berlangsung tanpa reorganisasi pada lapisan penyimpanan.

b.Data Organization

Dalam istilah Data pergudangan klasik, pengorganisasian data disebut integrasi data. Karena ada volume/jumlah data yang sangat besar, ada kecenderungan untuk mengatur data pada lokasi penyimpanan aslinya, sehingga menghemat waktu dan uang dengan tidak memindah-midahkan data dengen volume yang besar. Infrastruktur yang diperlukan untuk mengatur data yang besar harus mampu mengolah dan memanipulasi data di lokasi penyimpanan asli. Biasanya diproses didalam batch untuk memproses data yang besar, beragam format, dari tidak terstruktur menjadi terstruktur.

Apache Hadoop adalah sebuah teknologi baru yang memungkinkan volume data yang besar untuk diatur dan diproses sambil menjaga data pada cluster penyimpanan data asli. Hadoop Distributed File System (HDFS) adalah sistem penyimpanan jangka panjang untuk log web misalnya. Log web ini berubah menjadi perilaku browsing dengan menjalankan program MapReduce di cluster dan menghasilkan hasil yang dikumpulkan di dalam cluster yang sama. Hasil ini dikumpulkan kemudian dimuat ke dalam sistem DBMS relasional.

c.Data Analysis

Karena data tidak selalu bergerak selama fase organisasi, analisis ini juga dapat dilakukan dalam lingkungan terdistribusi, di mana beberapa data akan tinggal di mana data itu awalnya disimpan dan diakses secara transparan dari sebuah data warehouse. Infrastruktur yang diperlukan untuk menganalisis data yang besar harus mampu mendukung analisis yang lebih dalam seperti analisis statistik dan data mining, pada data dengan jenis yang beragam dan disimpan dalam sistem yang terpisah, memberikan waktu respon lebih cepat didorong oleh perubahan perilaku; dan mengotomatisasi keputusan berdasarkan model analitis. Yang paling penting, infrastruktur harus mampu mengintegrasikan analisis pada kombinasi data yang besar dan data perusahaan tradisional. Wawasan baru datang bukan hanya dari analisis data baru, tapi dari menganalisisnya dalam konteks yang lama untuk memberikan perspektif baru tentang masalah lama.

Misalnya, menganalisis data persediaan dari mesin penjual otomatis cerdas dalam kombinasi dengan acara kalender untuk tempat di mana mesin penjual otomatis berada, akan menentukan kombinasi produk yang optimal dan jadwal pengisian untuk mesin penjual otomatis.

Aplikasi big data

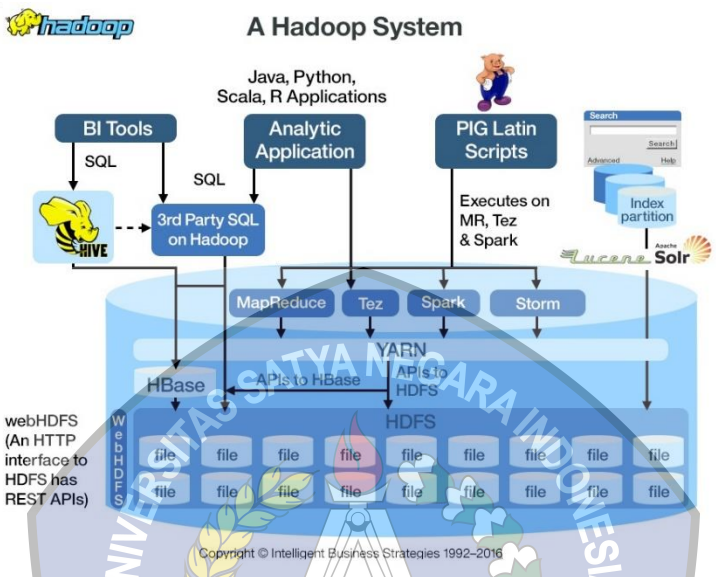
Bansod dkk. (2015) dalam penelitiannya menganalisis efisiensi big data yang menggunakan framework dari Apache Spark dan HDFS serta keuntungan dari penggunaaan framework Hadoop. Hasil dari penelitian ini adalah Apache Spark terbukti memiliki performa dan skalabilitas yang tinggi serta bersifat faulttolerant untuk analisis big data. MadhaviLatha dkk. membangun infrastruktur big data untuk menganalisis data twitter secara realtime menggunakan Apache Flume, Spark, Cassandra dan Zeppelin. Pada penelitian ini, Cassandra dapat diintegrasikan dengan hdfs, kemudian data yang berasal dari flume dan spark streaming disimpan dalam Cassandra menggunakan beberapa fungsi khusus antara Cassandra dan Streaming Context dari Spark yaitu com.datastax.spark.connector.streaming. Tujuan dari menyimpan data di Cassandra yaitu untuk keperluan analisis lebih lanjut.

Beberapa contoh framework big data yaitu:

a. Apache Hadoop

   Hadoop adalah proyek dengan kode sumber terbuka yang dikelola oleh Apache Software Foundation. Hadoop digunakan untuk perhitungan yang andal, dapat diukur, didistribusikan, tetapi juga dapat dieksploitasi sebagai penyimpanan file dengan tujuan umum yang dapat

menyimpan petabyte data. Solusinya terdiri dari dua komponen utama: HDFS bertanggung jawab untuk penyimpanan data di cluster Hadoop; dan sistem MapReduce dimaksudkan untuk menghitung dan memproses volume data yang besar di cluster. Bagaimana tepatnya Hadoop membantu memecahkan masalah memori DBMS modern? Hadoop digunakan sebagai lapisan perantara antara database interaktif dan penyimpanan data meningkatkan kecepatan kinerja pemrosesan data tumbuh sesuai dengan peningkatan ruang penyimpanan data. Untuk mengembangkannya lebih lanjut, Anda cukup menambahkan node baru ke penyimpanan data. Secara umum, Hadoop dapat menyimpan dan memproses banyak petabyte info. Di sisi lain, proses tercepat di Hadoop masih membutuhkan beberapa detik untuk beroperasi. Itu juga melarang kustomisasi data yang sudah disimpan dalam sistem HDFS. Last but not least, solusinya mendukung transaksi. Jadi, terlepas dari popularitas, alternatif baru yang lebih maju secara bertahap (kami akan membahas beberapa di bawah).

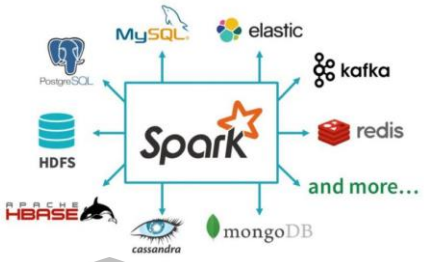


Gambar 4. 1 Ilustrasi sistem apache hadoop

b. Apache Spark

Daftar kerangka kerja Big Data terbaik kami dilanjutkan dengan Apache Spark. Ini adalah kerangka kerja open-source yang dibuat sebagai solusi yang lebih maju dibandingkan dengan Apache Hadoop - kerangka awal yang dibangun khusus untuk bekerja dengan Big Data. Perbedaan utama antara kedua solusi ini adalah model pengambilan data. Hadoop menyimpan data ke hard drive di sepanjang setiap langkah algoritma MapReduce, sementara Spark mengimplementasikan semua operasi menggunakan memori akses-acak.

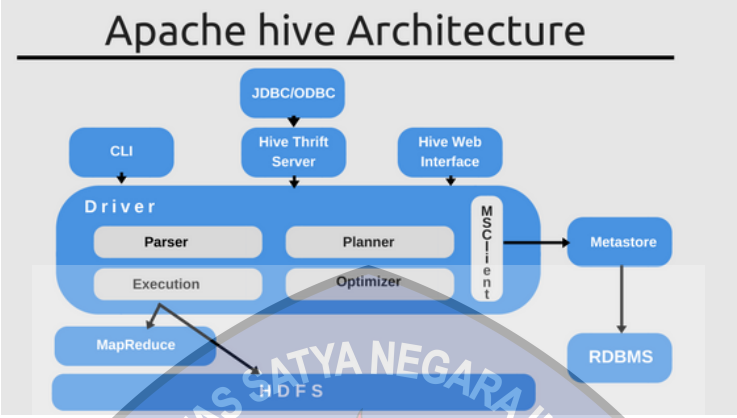
Karena hal ini, Spark memiliki kinerja 100 kali lebih cepat dan memungkinkan pemrosesan aliran data. Pilar fungsional dan fitur utama Spark adalah kinerja tinggi dan keamanan yang gagal. Ini mendukung empat bahasa: Scala, Java, Python, dan R; dan terdiri dari lima komponen: inti dan empat perpustakaan yang mengoptimalkan pekerjaan dengan Big Data dalam berbagai cara ketika digabungkan. Spark SQL - salah satu dari empat pustaka kerangka kerja khusus - berfungsi untuk pemrosesan data terstruktur menggunakan DataFrames dan penyelesaian permintaan Hadoop Hive hingga 100 kali lebih cepat. Spark juga dilengkapi alat Streaming untuk pemrosesan data khusus utas secara real time. Dengan demikian, pendiri Spark menyatakan bahwa waktu rata-rata pemrosesan setiap mikro-batch hanya 0,5 detik. Berikutnya, ada MLib - sistem pembelajaran mesin terdistribusi sembilan kali lebih cepat dari perpustakaan Apache Mahout. Dan perpustakaan terakhir adalah GraphX yang digunakan untuk pemrosesan data grafik yang dapat diskalakan.



Gambar 4. 2 Ilustrasi apache spark

c. Apache Hive

Apache Hive dibuat oleh Facebook untuk menggabungkan skalabilitas salah satu alat big data yang paling populer dan banyak diminati, MapReduce dan aksesibilitas SQL. Hive pada dasarnya adalah mesin yang mengubah permintaan SQL menjadi rantai tugas pengurangan peta. Mesin mencakup komponen seperti Parser (yang mengurutkan permintaan SQL yang masuk), Pengoptimal (yang mengoptimalkan permintaan untuk efisiensi lebih), dan Pelaksana (yang meluncurkan tugas dalam kerangka kerja MapReduce). Hive dapat diintegrasikan dengan Hadoop (sebagai bagian server) untuk analisis volume data yang besar.

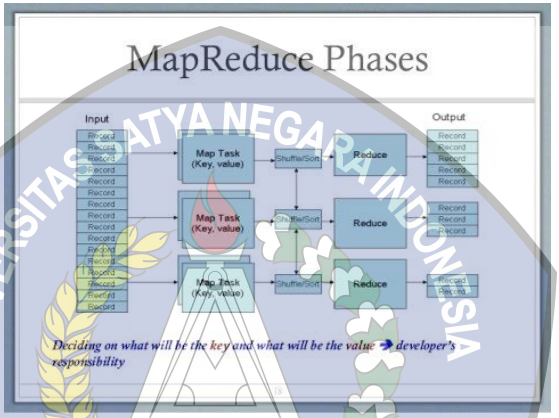


Gambar 4. 3 Arsitektur apache hive

d. Map Reduce

MapReduce adalah algoritme untuk pemrosesan paralel volume data mentah besar yang diperkenalkan oleh Google pada tahun 2004. MapReduce melihat data sebagai jenis entri yang dapat diproses dalam tiga tahap: Peta (pra-pemrosesan dan penyaringan data), Shuffle (node pekerja mengurutkan data - setiap node pekerja sesuai dengan satu kunci output yang dihasilkan dari fungsi peta), dan Reduce (fungsi pengurangan diatur oleh pengguna dan mendefinisikan hasil akhir untuk kelompok yang terpisah dari data output. Mayoritas semua nilai dikembalikan oleh mengurangi () fungsi

adalah hasil akhir dari tugas MapReduce). Karena logika sederhana seperti itu, MapReduce menyediakan paralelisasi data secara otomatis, penyeimbangan beban node pekerja yang efisien, dan kinerja gagal-aman.

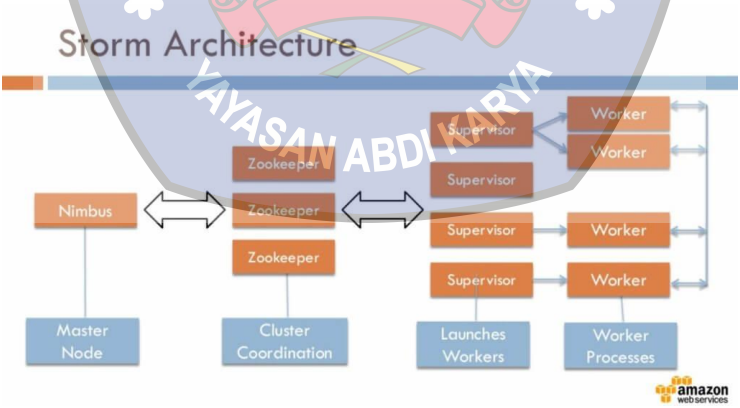


Gambar 4. 4 Pase map reduce

e. Apache Storm

Apache Storm adalah solusi terkemuka yang berfokus pada bekerja dengan aliran data besar secara real time. Fitur utama Storm adalah skalabilitas (tugas pemrosesan didistribusikan oleh node cluster dan mengalir di setiap node) dan kemampuan memulihkan segera setelah downtime (dengan demikian, tugas sedang dialihkan ke node pekerja lain jika salah satu node sedang down). Anda dapat bekerja dengan solusi ini dengan

bantuan Java, serta Python, Ruby, dan Fancy. Strom menampilkan sejumlah elemen yang membuatnya sangat berbeda dari analog. Yang pertama adalah Tuple - elemen representasi data utama yang mendukung serialisasi. Lalu ada Stream yang menyertakan skema bidang penamaan di Tuple. Spout menerima data dari sumber eksternal, membentuk Tuple dari mereka, dan mengirimkannya ke Stream. Ada juga Bolt - pengolah data, dan Topologi - paket elemen dengan deskripsi keterkaitan mereka analog pekerjaan MapReduce di Hadoop, pada dasarnya). Ketika digabungkan, semua elemen ini membantu pengembang untuk dengan mudah mengelola aliran besar data yang tidak terstruktur.



Gambar 4. 5 Arsitektur apache storm

2.10. Sketsa Data Warehouse

- **Tabel Fakta**
  Menururt Inmon (2005,p497), tabel fakta adalah tabel pusat dari skema bintang dimana data yang sering muncul akan ditempatkan disini. Disebut juga tabel utama atau major tabel, merupakan inti dari skema bintang dan berisi data aktual yang akan dianalisis. Tabel fakta adalah tabel yang pada umumnya mengandung angka dan data historis dimana key yang dihasilkan sangat unik karena key nya merupakan kumpulan foreign key dan primary key yang ada pada masing-masing tabel dimensi yang berhubungan atau merupakan tabel terpusat dari skema bintang. Tabel fakta menyimpan tipe-tipe measure yang berbeda, seperti measure, yang secara langsung terhubung dengan tabel dimensi dan measure yang tidak berhubungan dengan tabel dimensi.

- **Tabel Dimensi**
  Menurut Inmon (2005,p497), tabel dimensi adalah tempat dimana data-data yang tidak berhubungan yang berelasi dengan tabel fakta yang ditempatkan di dalam tabel multidimensional. Disebut juga tabel kecil atau minor tabel, biasanya lebih kecil dan memegang data deskriptif yang mencerminkan dimensi suatu bisnis. Tabel dimensi adalah tabel yang berisikan kategori

dengan ringkasan data detil yang dapat dilaporkan sebagai dimensi waktu (berupa perbulan, perkuartal, pertahun).

- **Pemodelan dalam Dimensional**
  Model dimensional merupakan rancangan logikal yang bertujuan untuk menampilkan data dalam bentuk standar dan intuitif yang memperbolehkan akses dengan performa yang tinggi.
  Model dimensional menggunakan konsep model hubungan antar entity (ER) dengan beberapa batasan yang penting. Setiap model dimensi terdiri dari sebuah tabel dengan sebuah komposit primary key, disebut dengan table fakta, dan satu set table yang lebih kecil disebut table dimensi. Setiap table dimensi memiliki sebuah simple primary key yang merespon tepat pada satu komponen primary key pada tabel fakta. Dengan kata lain primary key pada table fakta terdiri dari dua atau lebih foreign key. Struktur karakteristik ini disebut dengan skema bintang atau join bintang.
  Fitur terpenting dalam model dimensional ini adalah semua natural keys diganti dengan kunci pengganti(surrogate keys). Maksudnya yaitu setiap kali join antar table fakta dengan table dimensi selalu didasari kunci pengganti. Kegunaan dari kunci pengganti adalah memperbolehkan data pada data warehouse untuk memiliki beberapa kebebasan dalam penggunaan data, tidak seperti halnya yang diproduksi oleh sistem OLTP.

Sebuah sistem OLTP memerlukan normalisasi untuk mengurangi redudansi, validasi untuk input data, mendukung volume yang besar dari transaksi yang bergerak sangat cepat. Model OLTP sering terlihat seperti jaring laba-laba yang terdiri atas ratusan bahkan ribuan tabel sehingga sulit untuk dimengerti.

Sebaliknya, dimension model yang sering digunakan pada data warehouse adalah skema bintang atau snowflake yang mudah dimengerti dan sesuai dengan kebutuhan bisnis, mendukung query sederhana dan menyediakan performa query yang superior dengan meminimalisasi tabel-tabel join.

- **Skema Bintang**

  Menurut Conolly dan Begg (2005, p1183), skema bintang adalah struktur logika yang mempunyai sebuah tabel fakta berisi data faktual yang ditempatkan ditengah, dikelilingi oleh tabel dimensi berisi data referensi (dapat di denormalisasi).

  - Keuntungan Menggunakan Skema Bintang

    Skema bintang memiliki beberapa keuntungan yang tidak terdapat dalam struktur relational biasa. Keuntungan menggunakan skema bintang yaitu :

    1. Respon data yang lebih cepat dihasilkan dari perancangan database .

2. Kemudahan dalam mengembangkan atau memodifikasi data yang terus berubah.
3. End user dapat menyesuaikan cara berpikir dan menggunakan data, konsep ini dikenal juga dengan istilah pararel dalam perancangan database.
4. Menyederhanakan pemahaman dan penelusuran metadata bagi pemakai dan pengembang.

- Perancangan Skema Bintang
  Skema bintang merupakan suatu struktur sederhana yang secara relatife terdiri dari beberapa tabel dan alur gabungan yang dirumuskan dengan baik. Perancangan database ini berlawanan dengan struktur organisasi yang digunakan untuk database proses transaksi. Database ini menyediaakn response time query, skema bintang yang dapat dibaca dan dipahami oleh analisis, end user, bahkan bagi mereka yang belum terbiasa dengan struktur database.

- Skema Bintang Sederhana

Menurut Poe (2001, p193-195), masing-masing tabel memiliki kunci utama (primary key), skema bintang sederhana memiliki kunci untuk tabel fakta dan terdiri dari satu atau lebih foreign key. Foreign key adalah data pada tabel yang memiliki nilai-nilai seperti digambarkan oleh primary key di dalam tabel lain. Ketika database dibuat SQL statement digunakan untuk menciptakan tabel, memilih untuk membentuk primary key dan foreign key.

- Skema Bintang dengan Banyak Tabel Fakta

  Menurut Poe (2001, p195-197), skema bintang dapat berisi berbagai tabel fakta. Dalam beberapa hal, tabel fakta ada karena berisi fakta yang tidak berhubungan atau karena perbedaan waktu pemuatan data, disamping itu juga dapat digunakan untuk meningkatkan daya guna atau hasil terutama jika data dalam jumlah yang besar. Gambar berikut adalah skema bintang dengan banyak tabel fakta.

Kegunaan lain dari tabel fakta adalah menggambarkan atau menyelaraskan hubungan many to many diantara

dimensi tertentu, tabel jenis ini disebut juga tabel asosiasi.

- Skema Bintang Majemuk
  Menurut Poe (2001, p100-201), pada skema bintang majemuk, tabel fakta terdiri atas dua buah set yaitu foreign key menggunakan tabel dimensi sebagai referensi, dan kunci utama yang terdiri dari satu atau lebih menyediakan identifier yang unik untuk masing-masing baris. Salah satu ciri yang memiliki oleh skema bintang majemuk adalah primary key dan foreign key yang tidak sama.

- **Skema Snowflake**
  Menurut Conolly dan Begg (2005, p1185), skema snowflake adalah variasi lain dari skema bintang dimana tabel dimensi tidak berisi data yang dinormalisasi. Suatu tabel dimensi dapat memiliki tabel dimensi lainnya.

  - Keuntungan dan Kerugian Skema Snowflake
    1. Kecepatan memindahkan data dari data OLTP ke dalam metadata.
    2. Sebagai kebutuhan dari alat pengambil keputusan tingkat tinggi dimana dengan tipe yang seperti ini, seluruh struktur dapat digunakan sepenuhnya.

- Kerugian dari skema snowflake adalah :
  1. Skemanya kurang jelas dan end user terhambat oleh kompleksitas.
  2. Sulit untuk mencari isi skema karena terlalu kompleks.
  3. Performa query menurun karena adanya tambahan gabungan tabel.
  4. Mempunyai masalah yang besar dalam hal kinerja, hal ini disebabkan semakin banyaknya join antar tabel-tabel yang dilakukan dalam skema snowflake.

2.11. Metodologi Perancangan Data Warehouse

Berdasarkan kutipan dari Connolly dan Begg (2005, p1187-1193), metodologi yang dikemukakan oleh Kimball dalam membangun data warehouse ada 9 tahapan , yang dikenal dengan Nine-step Methodology. Sembilan tahap tersebut adalah :

- **Langkah 1 : Pemilihan proses**
  - Data mart yang pertama kali dibangun haruslah data mart yang dapat dikirim tepat waktu dan dapat menjawab semua pertanyaan bisnis yang penting
  - Pilihan terbaik untuk data mart yang pertama adalah yang berhubungan

dengan sales, misal property sales, property leasing,property advertising.

- **Langkah 2 : Pemilihan sumber**
  - Untuk memutuskan secara pasti apa yang diwakili atau direpresentasikan oleh sebuah tabel fakta.

  - Misal, jika sumber dari sebuah tabel fakta properti sale adalah properti sale individual maka sumber dari sebuah dimensi pelanggan berisi rincian pelanggan yang membeli properti utama

- **Langkah 3 : Mengidentifikasi dimensi**
  - Set dimensi yang dibangun dengan baik, memberikan kemudahan untuk memahami dan menggunakan data mart
  - Dimensi ini penting untuk menggambarkan fakta-fakta yang terdapat pada tabel fakta
  - Misal, setiap data pelanggan pada tabel dimensi pembeli dilengkapi dengan id_pelanggan,no_pelanggan,tipe_pelanggan,tempat_tinggal, dan lain sebagainya.
  - Jika ada dimensi yang muncul pada dua data mart,kedua data mart tersebut harus berdimensi sama,atau paling tidak salah satunya berupa subset matematis dari yang lainnya.

- Jika sebuah dimensi digunakan pada dua data mart atau lebih,dan dimensi ini tidak disinkronisasi,maka keseluruhan data warehouse akan gagal, karena dua data mart tidak bisa digunakan secara bersama-sama.

- **Langkah 4 : Pemilihan fakta**
  - Sumber dari sebuah tabel fakta menentukan fakta mana yang bisa digunakan dalam data mart.
  - Semua fakta harus diekspresikan pada tingkat yang telah ditentukan oleh sumber

- **Langkah 5 : Menyimpan pre-kalkulasi di tabel fakta**

  Setelah fakta-fakta dipilih maka dilakukan pengkajian ulang untuk menentukan apakah fakta-fakta yang dapat diterapkan kalkulasi awal dan melakukan penyimpanan pada tabel fakta. Contoh umum dari kebutuhan untuk penyimpanan kalkulasi awal muncul ketika fakta berisi pernyataan untung atau rugi. Situasi ini akan meningkat ketika tabel fakta didasarkan pada invoice atau sales.

- **Langkah 6 : Melengkapi tabel dimensi**

- Pada tahap ini kita menambahkan keterangan selengkap-lengkapnya pada tabel dimensi
- Keterangannya harus bersifat intuitif dan mudah dipahami oleh pengguna

- **Langkah 7 : Pemilihan durasi database**

  Durasi mengukur waktu dari pembatasan data yang diambil dan dipindahkan ke tabel fakta. Sebagai contoh perusahaan asuransi memiliki kebutuhan untuk menyimpan data dalam jangka waktu 5 tahun atau lebih.

- **Langkah 8 : Menelusuri perubahan dimensi yang perlahan**

  Ada tiga tipe perubahan dimensi yang perlahan, yaitu :

  Tipe 1. Atribut dimensi yang telah berubah tertulis ulang

  Tipe 2. Atribut dimensi yang telah berubah menimbulkan sebuah dimensi baru
  Tipe 3. Atribut dimensi yang telah berubah menimbulkan alternatif sehingga nilai atribut lama dan yang baru dapat diakses secara bersama pada dimensi yang sama.

- **Langkah 9 : Menentukan prioritas dan mode query**

  Mempertimbangkan pengaruh dari rancangan fisik, seperti penyortiran urutan tabel fakta pada disk dan keberadaan dari penyimpanan awal ringkasan atau penjumlahan. Selain itu, masalah administrasi, backup, kinerja indeks, dan keamanan juga merupakan faktor yang harus dipertahankan.
  Dengan langkah-langkah tadi, seharusnya kita bisa membangun sebuah data warehouse yang baik.

2.12. Keuntungan Penggunaan Data Warehouse

Menurut Connolly dan Begg (2005, p1152), data warehouse yang telah diimplementasikan dengan baik dapat memberikan keuntungan yang besar bagi organisasi, yaitu:

- Potensi nilai kembali yang besar pada investasi Sebuah organisasi harus mengeluarkan uang dan sumber daya dalam jumlah yang cukup besar untuk memastikan kalau data warehouse telah diimplementasikan dengan baik, biaya yang dikeluarkan tergantung dari solusi teknikal yang diinginkan. Akan tetapi, setelah data warehouse digunakan, maka kemungkinan didapatkannya ROI (Return on Investment) relatif lebih besar.

- Keuntungan kompetitif didapatkan apabila pengambil keputusan mengakses data yang dapat mengungkapkan informasi yang sebelumnya tidak diketahui, tidak tersedia, misalnya informasi mengenai konsumen, tren, dan permintaan.
- Meningkatkan produktivitas para pengambil keputusan perusahaan Data warehouse meningkatkan produktivitas para pengambil keputusan perusahaan dengan menciptakan sebuah database yang terintegrasi secara konsisten, berorientasi pada subjek, dan data historis.

  Data warehouse mengintegrasikan data dari beberapa sistem yang tidak compatible ke dalam bentuk yang menyediakan satu pandangan yang konsisten dari organisasi. Dengan mengubah data menjadi informasi yang berguna, maka seorang manajer bisnis dapat membuat analisa yang lebih akurat dan konsisten.

BAB 3
KESIMPULAN

3.1. Kesimpulan Big Data

Big data adalah kumpulan dari proses yang terdiri dari volume data dengan kapasitas yang besar dan dapat menampung data terstruktur dan tidak terstruktur. Terdapat 3 definisi yang dapat merepresentasikan apa itu big data. Pertama adalah volume (kapasitas), velocity (kecepatan), dan variety (variasi).

Big data memiliki beberapa fungsi dan manfaat terkait dengan kebutuhan penyimpanan data yang besar dengan proses pembacaan data lebih cepat dan efisien. Terdapat beberapa tools yang dapat kami rekomendasikan untuk kebutuhan penunjang bisnis.

3.2. Kesimpulan Data Warehouse

Dibanding database tradisional, Data Warehouse umumnya terdiri dari data yang berukuran sangat besar dari banyak sumber dan mungkin terdiri dari database dari model data yang berbeda dan kadang file dari sistem dan platform yang independent Tidak seperti database transaksional, Data Warehouse biasanya mendukung analisa tren dan time-series, di mana keduanya membutuhkan data historic. Data Warehouse itu nonvolatile. Artinya informasi

dalam Data Warehouse jarang diubah dan bisa dianggap non-real-time. Data Warehouse bisa digambarkan sebagai "kumpulan teknologi pendukung keputusan, dimaksudkan untuk memungkinkan pekerja yang berhubungan dengan informasi (eksekutif, manajer dan analis) untuk membuat keputusan lebih baik dan lebih cepat". Penggunaan data warehouse juga memiliki beberapa keuntungan, yaitu meningkatkan produktivitas para pengambil keputusan dan memiliki keuntungan kompetitif.

Daftar Pustaka

Binus. 2011. Pengertian, Konsep, Karakteristik, Struktur Data Warehouse. URL: http://library.binus.ac.id/eColls/eThesisdoc/Bab2/2011-2-00305-IF%20Bab%202.pdf

Media, sekawan. 2020. Pengertian Big Data. URL: https://www.sekawanmedia.co.id/blog/pengertian-big-data/

TOPSIS (*TECHNIQUE FOR OTHERS REFERENCE BY SIMILARITY TO IDEAL SOLUTION*)

ALGORIMA APRIORI DAN METODE ROUGH SET

PENDAHULUAN

Sistem Pendukung Keputusan merupakan bagian dari sistem informasi berbasis komputer yang digunakan untuk mendukung pengambilan keputusan dalam suatu instansi atau perusahaan. Sistem pendukung keputusan di bangun untuk memudahkan seseorang untuk mengambil suatu keputusan. Sistem dapat mengambil suatu keputusan sesuai dengan pertimbangan dari kriteria-kriteria yang telah kita masukkan sebelumnya.

Data mining digunakan banyak tempat dan bidang penerapannya juga dapat bermacam macam, data mining mempelajari apa saja yang menjadi faktor utama dalam ketepatan sasaran pembelian suatu produk oleh konsumen. Kecerdasan bisnis merupakan proses pengubahan data menjadi informasi. Dari kumpulan informasi yang ada akan diambil polanya menjadi pengetahuan. Berry.M.J.A. dan Linoff.G.S.9 Tujuan kecerdasan bisnis ini adalah untuk mengubah data yang sangat banyak dan memiliki nilai bisnis melalui laporan analitik. Berry.M.J.A. dan Linoff.G.S. 9 algoritma apriori salah satu algoritma data mining melakukan proses ekstraksi informasi pada database untuk menemukan aturan asosiasi antara suatu kombinasi item/itemset, Abdullah4. Penyelesaian masalah pada proses akstraksi informasi dari sebuah database atau data mining dengan melakukan proses generasi iterasi frequent itemset dalam jenis aturan asosiasi rule mining (association rule mining) sehingga menghasilkan nilai support dan confidence.

Selanjutnya, terdapat beberapa definisi yang disampaikan oleh para ahli, diantaranya adalah sebagai berikut.

1. Turban (2001)

Sistem pendukung keputusan adalah sistem yang digunakan untuk dapat mengambil keputusan pada situasi semi terstruktur dan tidak terstruktur, dimana seseorang tidak mengetahui secara pasti bagaimana seharusnya sebuah keputusan dibuat.

2. Sprague Et. Al (1993)

Sprague dan Watson membagi sistem pendukung keputusan menjadi lima bagian atau karakteristik, yaitu:

- Sistem berbasis komputer
- Sistem dibuat untuk mengambil keputusan
- Dibangun untuk membantu dalam memecahkan masalah yang rumit, dan tidak dapat diselesaikan melalui perhitungan kalkulasi secara manual
- Melalui bantuan simulasi yang interaktif
- Komponen utama terdiri dari kumpulan data dan model analisis

Di dalam proses pengolahannya, DSS dibantu dengan berbagai sistem lain seperti *Artificial Intelligence (AI)*, *Expert System (ES)*, *Fuzzy Logic*, dan lain sebagainya. Sehingga, tujuan dari penerapan SPK ini adalah sebagai berikut:

- Membantu dalam menyelesaikan permasalahan yang terbentuk secara semi – struktural
- Mampu mendukung aktivitas manajer dalam mengambil sebuah keputusan dalam suatu masalah
- Mampu meningkatkan keefektifan, bukan tingkat efisiensi dalam pengambilan keputusan

Tahapan yang harus dilalui untuk dapat mencapai hasil keputusan terbaik dalam dilakukan melalui cara atau fase berikut ini:

1. Intelligence Phase

Tahap pemahaman merupakan proses penelusuran untuk memetakan tingkat problematika, serta mampu mengenali permasalahan yang terjadi. *Input data* yang diperoleh nantinya diproses dan diuji cobakan dalam rangka mendukung proses identifikasi masalah.

2. Design Phase

Tahap perancangan dimulai dengan proses pengembangan pencarian solusi alternatif yang sangat mungkin untuk diambil. Namun, diperlukan proses verifikasi dan validasi untuk dapat mengetahui tingkat keakuratan pada model yang diteliti.

3. Choice Phase

Tahap pemilihan berfungsi untuk memilih berbagai solusi alternatif yang dapat dipilih, serta dimunculkan pada fase perencanaan dengan memperhatikan kriteria berdasarkan tujuan utamanya (*objective*).

4. Implementation Phase

Tahap implementasi atau penerapan, dilakukan dengan menyesuaikan rancangan sistem yang telah dibuat pada beberapa fase sebelumnya.

Terdapat, setidaknya tiga komponen utama yang tersusun dalam sebuah sistem pendukung keputusan, antara lain sebagai berikut.

1. Database Management

Manajemen basis data merupakan sub sistem dalam data yang terorganisir pada sebuah *database*. Untuk kepentingan SPK sendiri, diperlukan data yang relevan dengan permasalahan yang hendak diselesaikan melalui sistem berbasis simulasi.

2. User Interface

Tampilan antarmuka atau pengelolaan dialog adalah proses penggabungan antara dua komponen, yaitu *database management* dan *model base* yang nantinya akan bergabung dengan *user interface*. Nantinya *User Interface (UI)* akan menampilkan *output* atau keluaran sistem bagi pengguna perangkat lunak.

3. Model Base

Komponen model merepresentasikan terkait permasalahan ke dalam format data kuantitatif. Yang di dalamnya terdiri dari tujuan permasalahan, komponen, batasan (*constraint*), dan hal terkait lainnya. *Mode base* sangat memungkinkan untuk menganalisa permasalahan secara utuh dan mengembangkannya menjadi solusi yang terbaik.

Salah satu metode yang dapat digunakan untuk menyelesaikan permaslahan dalam suatu pengambilan keputusan adalah menggunakan metode TOPSIS.

TOPSIS (*Technique For Others Reference by Similarity to Ideal Solution*) adalah salah satu metode pengambilan keputusan multikriteria yang pertama kali diperkenalkan oleh **Yoon dan Hwang (1981)**. TOPSIS

menggunakan prinsip bahwa alternatif yang terpilih harus mempunyai jarak terdekat dari solusi ideal positif dan terjauh dari solusi ideal negatif dari sudut pandang geometris dengan menggunakan jarak *Euclidean* untuk menentukan kedekatan relatif dari suatu alternatif dengan solusi optimal.

Solusi ideal positif didefinisikan sebagai jumlah dari seluruh nilai terbaik yang dapat dicapai untuk setiap atribut, sedangkan solusi negatif-ideal terdiri dari seluruh nilai terburuk yang dicapai untuk setiap atribut.

Semakin banyaknya faktor yang harus dipertimbangkan dalam proses pengambilan keputusan, maka semakin relatif sulit juga untuk mengambil keputusan terhadap suatu permasalahan. Apalagi jika upaya pengambilan keputusan dari suatu permasalahan tertentu, selain mempertimbangkan berbagai faktor/kriteria yang beragam, juga melibatkan beberapa orang pengambil keputusan. Permasalahan yang demikian dikenal dengan permasalahan *multiple criteria decision making* (MCDM). Dengan kata lain, MCDM juga dapat disebut sebagai suatu pengambilan keputusan untuk memilih alternatif terbaik dari sejumlah alternatif berdasarkan beberapa kriteria tertentu. Metode TOPSIS digunakan sebagai suatu upaya untuk menyelesaikan permasalahan *multiple criteria decision making*. Hal ini disebabkan konsepnya sederhana dan mudah dipahami, komputasinya efisien dan memiliki kemampuan untuk mengukur kinerja relatif dari alternatif-alternatif keputusan. Metode ini banyak digunakan untuk menyelesaikan pengambilan keputusan secara praktis.

- Tahapan dalam Metode TOPSIS:

Secara umum, prosedur atau langkah-langkah dalam metode TOPSIS (*Technique For Order Preference By Similarity To Ideal Solution*) meliputi:

1. Membuat matriks keputusan yang ternormalisasi. Elemen r_{ij} hasil dari normalisasi decision matrix R dengan metode *Euclidean length of a vector* :

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}}$$ dengan i = 1,2,....m; dan j = 1,2,.....n;

2. Membangun matriks keputusan ternormalisasi yang terbobot. Dengan bobot W= (w1, w2, ...,wn), maka normalisasi bobot matriks V :

$$V = \begin{bmatrix} w_{11}r_{11} & \cdots & w_{1n}r_{1n} \\ \vdots & \ddots & \vdots \\ w_{m1}r_{m1} & \cdots & w_{nm}r_{nm} \end{bmatrix}$$

3. Menentukan solusi ideal positif dan solusi ideal negatif. Solusi ideal positif dinotasikan A^+, sedangkan solusi ideal negatif dinotasikan A^- :

$$A^+ \left\{(\max v_{ij})(\min v_{ij} | j \in J'), i = 1,2,3,...m\right\} = \{v_1^+, v_2^+, ...v_m^+\}$$

$$A^- \left\{(\max v_{ij})(\min v_{ij} | j \in J'), i = 1,2,3,...m\right\} = \{v_1^-, v_2^-, ...v_m^-\}$$

V_{ij} = elemen matriks V baris ke-i dan kolom ke-jJ = {j=1,2,3,...,n dan j berhubungan dengan *benefit criteria*} J ' = {j=1,2,3,...,n dan j berhubungan dengan *cost criteria*}

4. Menghitung Jarak solusi ideal psitif dan solusi ideal negatif.Menghitung separasi merupakan pengukuran jarak dari suatu alternatif ke solusi ideal positif dan solusi ideal negatif. Perhitungan matematisnya adalah sebagai berikut: Menghitung separasi untuk solusi ideal positif D_i^+adalah jarak (dalam pandangan Euclidean) alternatif dari solusi ideal. Jarak terhadap solusi ideal positif didefinisikan sebagai:

$$S_i^+ \sqrt{\sum_{i=0}^{n} \left(v_{ij} - v_j^+\right)} \quad 2$$

dengan i = 1,2,...,m

Dimana :

J= {j=1,2,3,...,n dan j merupakan *benefit criteria*}' = {j=1,2,3,...,n dan j merupakan *cost criteria*} Dan jarak terhadap solusi ideal negatif didefinisikan sebagai:

$$S_i^- \sqrt{\sum_{j=1}^{n} \left(v_{ij} - v_j^-\right)} \quad 2$$

dengan i = 1,2,...,m

Dimana :

J= {j=1,2,3,...,n dan j merupakan *benefit criteria*}

J' = {j=1,2,3,...,n dan j merupakan *cost criteria*}

5. Menghitung kedekatan relatif terhadap solusi ideal. Kedekatan relatif alternatif A^+ dengan solusi ideal A^- direpresentasikan:

$$C_i = \frac{s_i^-}{s_i^- + s_i^+}, \quad dengan\ 0 < C_i^+ < 1$$

dengan i =1,2,3...,m

6. Merangking Alternatif.
Alternatif dapat dirangking berdasarkan urutan Ci^+. Maka dari itu, alternatif terbaik adalah salah satu yang berjarak terpendek terhadap solusi ideal dan berjarak terjauh dengan solusi negatif-ideal.

Alasan memilih metode Topsis yaitu karena logikanya bersifat sederhana, proses perhitungan mudah dimengerti, alternatif terbaik yang terpilih merupakan model matematika sederhana.

- **Tujuan TOPSIS**

TOPSIS bertujuan untuk menentukan solusi ideal positif dan solusi ideal negatif. Solusi ideal positif memaksimalkan kriteria manfaat dan meminimalkan

kriteria biaya, sedangkan solusi ideal negatif memaksimalkan kriteria biaya dan meminimalkan kriteria manfaat (Fan dan Cheng, 2009). Kriteria manfaat merupakan kriteria dimana ketika nilai kriteria tersebut semakin besar maka semakin layak pula untuk dipilih. Sedangkan kriteria biaya merupakan kebalikan dari kriteria manfaat, semakin kecil nilai dari kriteria tersebut maka akan semakin layak untuk dipilih. Dalam metode TOPSIS, alternatif yang optimal adalah yang paling dekat dengan solusi ideal positif dan paling jauh dari solusi ideal negatif.

- **Prinsip TOPSIS**

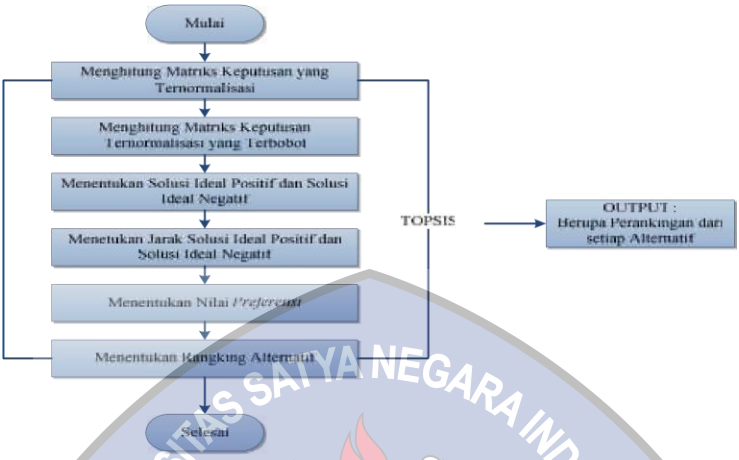
TOPSIS menggunakan prinsip bahwa alternatif yang terpilih harus mempunyai jarak terdekat dari solusi ideal positif dan terjauh dari solusi ideal negatif dari sudut pandang geometris dengan menggunakan jarak Euclidean untuk menentukan kedekatan relatif dari suatu alternatif dengan solusi optimal. Solusi ideal positif didefinisikan sebagai jumlah dari seluruh nilai terbaik yang dapat dicapai untuk setiap atribut, sedangkan solusi ideal negatif terdiri dari seluruh nilai terburuk yang dicapai untuk setiap atribut.TOPSIS mempertimbangkan keduanya, jarak terhadap solusi ideal positif dan jarak terhadap solusi ideal negatif dengan mengambil kedekatan relatif terhadap solusi ideal positif. Berdasarkan perbandingan terhadap jarak relatifnya, susunan prioritas alternatif bisa dicapai.

Dalam metode TOPSIS, dipertimbangkan adanya solusi ideal positif dan solusi ideal negatif. Solusi ideal positif merupakan nilai terbaik dari semua kriteria sedangkan solusi ideal negatif adalah nilai terburuk untuk

tiap kriteria dari alternatif yang ada. Dengan adanya kedua solusi ini maka alternatif yang dipilih dalam metode TOPSIS merupakan alternatif yang memiliki jarak terdekat dengan solusi ideal positif dan jarak terjauh dengan solusi ideal negatif. Karena itulah maka dapat disimpulkan beberapa kelemahan dan kelebihan metode TOPSIS.

- Flowchart Perhitungan Metode TOPSIS

Flowchart ini menjabarkan perhitungan metodeTOPSIS :



Kelemahan metode TOPSIS:

- Belum adanya penentuan bobot prioritas yang menjadi prioritas hitungan terhadap kriteria, yang berguna untuk meningkatkan validitas nilai bobot perhitungan kriteria. Maka dengan alasan ini, metode ini dapat dikombinasikan misalnya dengan metode **AHP** agar menghasilkan output atau keputusan yang lebih maksimal
- Belum adanya bentuk linguistik untuk penilaian alternatif terhadap kriteria, biasanya bentuk linguistik ini diinterpretasikan dalam sebuah bilangan ***fuzzy***
- Belum adanya mediator seperti hirarki jika diproses secara mandiri maka dalam ketepatan pengambilan keputusan cenderung belum menghasilkan keputusan yang sempurna
- Metode **TOPSIS** ini dapat digunakan dalam menentukan perangkingan alternatif dengan memperhitungkan solusi ideal dari suatu masalah dan

penentuan bobot setiap kriteria. Namun, kurang baik jika digunakan dalam mendapatkan bobot yang memperhitungkan hubungan antara kriteria. Walaupun dapat dilakukan dengan pairwase comparison, tetapi membutuhkan matriks dan perhitungan yang lebih rumit. Oleh karena itu, dilakukan penggabungan dengan metode lain seperti **ANP (*Analytic Network Process*)** dalam mengatasi masalah pembobotan tersebut.

- Pada proses yang menggunakan metode **TOPSIS**, perangkingan dan pembobotan kriteria adalah memiliki nilai yang telah pasti. Padahal, dalam aplikasinya di kehidupan nyata, terdapat informasi yang tidak lengkap atau informasi yang dibutuhkan tidak tersedia. Contoh penyebab informasi yang tidak lengkap tersebut adalah karena adanya penilaian dari manusia yang seringkali bersifat tidak pasti/kabur (***fuzzy)*** dan tidak dapat mengestimasikan perangkingan dalam data numerik yang pasti. Ketidakpastian ini merupakan sesuatu yang tidak dapat diatasi jika menggunakan metode **TOPSIS**, kecuali jika dilakukan perhitungan algoritma lebih lanjut dalam perumusan metode **TOPSIS** tersebut.
- Metode **TOPSIS** menentukan solusi berdasarkan jarak terpendek menuju solusi ideal dan jarak terbesar dari solusi negatif yang ideal. Namun, metode ini tidak mempertimbangkan kepentingan relatif (*relative importance*) dari masing-masing jarak tersebut.
- Pada metode **TOPSIS**, seringkali digunakan asumsi pada tingkat kepentingan relatif masing-masing respon dan digunakan kombinasi dengan metode lain untuk menyelesaikan asumsi tersebut. Contohnya adalah

dengan menggunakan metode **AHP (*Analytical Hierarchy Process)*** atau **ANP** (***Analytic Network Process***) untuk memperoleh nilai bobot yang mewakili tingkat kepentingan relatif masing-masing kriteria.

- Pada metode **TOPSIS**, alternatif dengan ranking tertinggi merupakan solusi yang terbaik, namun belum tentu ranking tertinggi tersebut adalah yang terdekat dari solusi ideal. Sehingga perlu dilakukan perhitungan lagi untuk memastikannya.

Kelebihan metode TOPSIS:

1. Konsepnya sederhana dan mudah dipahami, kesederhanaan ini dilihat dari alur proses metode **TOPSIS** yang tidak begitu rumit. Karena menggunakan indikator kriteria dan variabel alternatif sebagai pembantu untuk menentukan keputusan
2. Komputasinya efisien, perhitungan komputasinya lebih efisien dan dan cepat
3. Mampu dijadikan sebagai pengukur kinerja alternatif dan juga alternatif keputusan dalam sebuah bentuk *output* komputasi yang sederhana.
4. Dapat digunakan sebagai metode pengambilan keputusan yang lebih cepat.

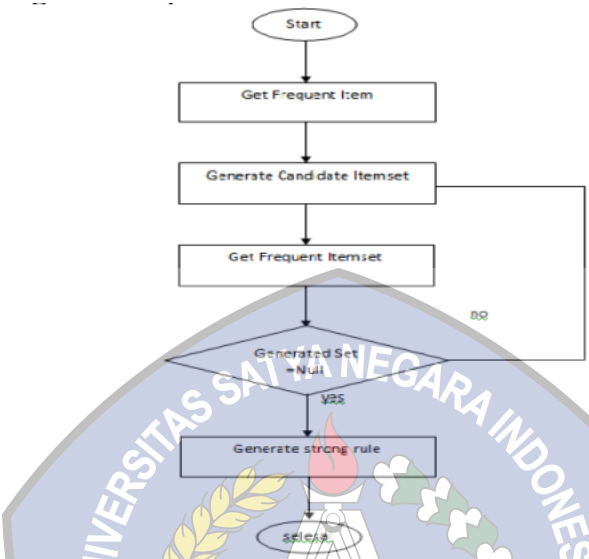
ALGORITMA APRIORI

Algoritma Apriori adalah salah satu algoritma pada data mining untuk mencari frequent item/itemset pada transaksional database. Algoritma apriori pertama kali diperkenalkan oleh R.Agarwal dan R Srikant untuk mencari frequent tertinggi dari suatu database, Kaur et

al8. Menurut (Mazida, 2015) Algoritma Apriori termasuk jenis aturan asosiasi pada data mining. Aturan yang menyatakan asosiasi antara beberapa atribut sering disebut affinity analysis atau market basket analysis. misalnya sebuah swalayan memiliki market basket, dengan adanya algoritma apriori, pemilik swalayan dapat mengetahui pola pembelian seorang konsumen, jika seorang konsumen membeli item A , B, punya 13 kemungkinan 50% dia akan membeli item C, pola ini sangat signifikan dengan adanya data transaksi selama ini. Menurut (Pracoyo, 2016) Apriori adalah suatu algoritma yang sudah sangat dikenal dalam melakukan pencarian frequent item set dengan menggunakan teknik association rule. Algoritma apriori menggunakan knowledge mengenai frequent itemset yang sebelumnya telah diketahui, untuk memproses informasi selanjutnya. Pada algoritma apriori untuk menentukan kandidat kandidat yang mungkin muncul yakni dengan cara memperhatikan minimum support. Menurut (Santoso, 2017) Teknik algoritma apriori untuk mencari informasi pola hubungan keterkaitan antar buku-buku yang sering dipinjam oleh pengunjung perpustakaan secara bersamaan. Sehingga dapat memberikan rekomendasi penyusunan buku sesuai dengan tingkat support dan confidence yang dimiliki oleh masing-masing buku yang saling berkaitan. Selain untuk memberikan rekomendasi penyusunan buku, juga dapat dijadikan sebagai refrensi untuk pengadaan buku baru sesuai topik buku-buku yang sering dipinjam oleh pengunjung perpustakaan. Berdasarkan hasil penelitian dapat disimpulkan bahwa data mining

menggunakan algoritma apriori dapat memberikan informasi berupa pola hubungan keterkaitan antar buku-buku yang sering dipinjam. Dan hasil aplikasi adalah rekomendasi penyusunan tata letak buku sesuai dengan pola keterkaitan antar masing-masing buku berdasarkan support dan confidence yang dimiliki. Menurut (Robi Yanto, 2017) penempatan buku dilakukan berdasarkan kategori buku yang telah tersedia pada rak buku namun belum diatur berdasarkan intensitasi peminjaman buku yang dilakukan oleh anggota sehingga masih banyak buku-buku lama yang masih tersedia di perpustakaan sehingga digunakan teknik asosiasi data mining, kemudian menghasilkan dari proses 10 data transaksi peminjaman buku dengan minimum support 5 dan minimum confidence 83.3 % yang dilakukan menggunakan 15 perangkat lunak pengujian XLminer dihasilkan kecocokan hasil yang diperoleh yaitu confidence 100% Kimia dan Fisika, Biologi dan Fisika, Sosiologi dan Fisika. Menurut (Esis Srikanti, 2018) Selama ini, data transaksi peminjaman buku di Perpustakaan Fakultas Sains dan Teknologi hanya disimpan begitu saja tanpa ada pengolahan lebih lanjut.Penggunaan bottom-up pendekatan berulang. Untuk menentukan asosiasi rule mining sebuah transaksi database, diperlukan waktu dalam melakukan proses frequent itemset, menghasilkan kombinasi data yang cukup t banyak, Abdullah4. Proses ini dilakukan untuk mencari minimum nilai support dan minimum nilai confidence . Algoritma apriori sangat mudah dipahami, tetapi ada beberapa kekurangan pada algortima tersebut:

1. Database Scanning: Database transaksi perlu dipindai berulang kali untuk menemukan frequent itemset. Jika ada n item dalam database, membutuhkan minimal n kali memindai database.
2. Pengaturan minimal frequent item/itemset untuk menentukan nilai support minimum.
3. Aturan Asosiasi rule mining dalam mendapatkan nilai minimum confidence Langkah-langkah algoritma apriori sebagai berikut:
1. Join(penggabungan). Pada proses ini setiap item dikombinasikan dengan item yang lainnya sampai tidak terbentuk kombinasi lagi.
2. Prune(pemangkasan). Pada proses ini, hasil dari item yang telah dikombinasikan tadi lalu dipangkas dengan menggunakan minimum support yang telah ditentukan. Dua proses utama tersebut merupakan langkah yang akan dilakukan untuk mendapat frequent itemset pada algoritma Apriori.



Gambar 2. Deskripsi Algoritma Apriori

ANALISIS ASOSIASI RULE MINING Aturan asosiasi merupakan dalam data mining yang menemukan frequent itemset pada database. Asosiasi aturan data mining adalah mekanisme dalam data mining dalam aturan asosiasi, ekspresi implikasi dari bentuk $X \rightarrow Y$ di mana X adalah Y. Anteseden dan konsekuen ditetapkan item domain I. pendahuluan dan konsekuen adalah seperangkat item dari domain I. Dengan demikian $X \cap Y = \Phi$. Dukungan dari set item didefinisikan sebagai rasio jumlah transaksi yang mengandung item diatur pada jumlah total transaksi. Kepercayaan aturan asosiasi $X \rightarrow Y$ adalah probabilitas bahwa Y transaksi mengandung algoritma association rule mining X , Arora K. Rakesh dan Badal

Dharmendra10 Rumus untuk mencari nilai support dan confidence adalah :

a. *Support*

$$\text{Support } (A \rightarrow B) = \frac{\text{Jumlah Transaksi Mengandung A dan B}}{\text{Jumlah Total Transaksi}}$$

b. *Confidence*

$$\text{Support } (A \rightarrow B) = \frac{\text{Jumlah Transaksi Mengandung A dan B}}{\text{Jumlah Total Transaksi}}$$

Analisis asosiasi didefenisikan suatu proses untuk menemukan semua aturan asosiasi yang memenuhi syarat minimum untuk support (minimum support) dan syarat minimum untuk confidence (minimum confidence). FP-GROWTH Mining tanpa melakukan candidate generation adalah teknik FP-Growth dengan menggunakan struktur data FP-tree, Han et al5. Dengan menggunakan cara ini scan database hanya dilakukan dua kali saja, tidak perlu berulang-ulang. Data akan direpresentasikan dalam bentuk FP-Tree. Setelah FP-Tree terbentuk, maka struktur data yang baik sekali untuk Frequent itemset akan diperoleh. Kumar B.S dan Rukmani .K.V.3 FP-Tree merupakan struktur data yang baik sekali untuk frequent Pattern mining, Han et al5 Struktur ini memberikan informasi yang lengkap untuk membentuk Frequent Pattern. Item-item yang tidak frequent (infrequent) sudah tidak ada dalam penggunaan FP-tree, Han et al5 Pembangunan FP-Tree dari sekumpulan data transaksi, akan diterapkan algoritma FP-Growth untuk mencari Frequent itemset yang signifikan, Han et al5. Algoritma FPtree dibagi menjadi tiga langkah utama, yaitu: 1. Tahap

Pembangkitan Conditional Pattern Base merupakan subdatabase yang berisi prefix path (lintasan e:1 prefix) dan pattern (pola akhiran). Pembangkitan conditioanl pattern base didapatkan melalui FP-tree yang telah dibangun sebelumnya. 2. Tahap Pembangkitan Conditional FPtree pada tahap ini, support count dari setiap item pada setiap conditional pattern base dijumlahkan, lalu setiap item yang memiliki jumlah support count lebih besar sama dengan minimum support count akan dibangkitkan dengan conditional FPtree. 3. Tahap Pencarian frequent itemset apabila conditional FP-tree merupakan lintasan tunggal(single path), maka didapatkan frequent itemset dengan melakukan kombinasi item untuk setiap conditonal FP-tree. Jika bukan lintasan tunggal, maka dilakukan pembangkitan FP-growth secara rekursif. Ketiga tahap tersebut merupakan langkah yang akan dilakukan untuk mendapatkan frequent itemset.

Cara kerja apriori :

- Tentukan minimum support
- Iterasi 1 : hitung item-item dari support(transaksi yang memuat seluruh item) dengan men-scan database untuk 1-itemset, setelah 1-itemset didapatkan, dari 1-itemset apakah diatas minimum support, apabila telah memenuhi minimum support, 1-itemset tersebut akan menjadi pola frequent tinggi,

- Iterasi 2 : untuk mendapatkan 2-itemset, harus dilakukan kombinasi dari k-itemset sebelumnya, kemudian scan database lagi untuk hitung item-item yang memuat support. itemset yang memenuhi minimum support akan dipilih sebagai pola frequent tinggi dari kandidat

- Tetapkan nilai k-itemset dari support yang telah memenuhi minimum support dari k-itemset

- lakukan proses untuk iterasi selanjutnya hingga tidak ada lagi k-itemset yang memenuhi minimum support.

Kelebihan & Kekurangan Algoritma Apriori

Kelebihan Algoritma Apriori :

1. Dibandingkan dengan algoritma lainnya, algoritma apriori dapat menangani data dalam jumlah besar.
2. Dapat menyederhanakan data.

Kekurangan Algoritma Apriori :

1. Memerlukan banyak waktu apabila memiliki banyak iterasi.
2. Dalam setiap iterasi memerlukan scan database.

ROUGHSET

Teori Rough set sampai saat ini pendekatan lain untuk ketidakjelasan (Pawlak, 1982). Demikian pula untuk teori himpunan fuzzy bukan merupakan alternatif untuk teori himpunan klasik tetapi tertanam di dalamnya. Teori Rough Set dapat dilihat sebagai implementasi khusus dari gagasan G. Frege (1983) tentang ketidakjelasan, yaitu ketidaktepatan dalam pendekatan ini dinyatakan oleh batas wilayah dari suatu himpunan, dan bukan oleh keanggotaan parsial, seperti dalam teori himpunan fuzzy. Konsep Rough Set dapat didefinisikan cukup umum dengan cara operasi topologi, interior dan penutupan, yang disebut pendekatan. Tujuan analisis Rough Set adalah untuk mendapatkan rule yang klasifikasi setelah dilakukan pengumpulan data (Maharani, 2008). Rule disini sudah dikalsifikasikan setelah mendapatkan reduct. Sebagai contoh, pasien yang menderita penyakit flu, memiliki gejala yang sama tetapi tak terlihat dan dapat dianggap sebagai unit penyakit pengetahuan medis. Pengetahuan medis ini disebut set dasar (konsep). Konsep dasar ini dapat dikombinasikan menjadi konsep majemuk, yaitu konsep yang unik ditentukan dalam hal konsep dasar pengetahuan. Set dasar disebut set renyah (set awal), dan set selain set dasar disebut set kasar (samar-samar, tidak tepat). Perbeadaan set dasar dan set kasar adalah dilihat dari batas wilayahnya, set dasar merupakan eleman yang ada didalam set yang pasti milik set, sementara set kasar adalah elemen yang berada diluar set yang mungkin milik set. Rough Set menentukan teorinya menggunakan perkiraan, yaitu yang ditentukan oleh fungsi keanggotaan. Rough Set bisa juga menentukan teorinya tanpa menggunakan perkiraan. Karena fungsi keanggotaan

bukanlah konsep primitif dalam pendekatan yang dalam hal ini kedua defenisi tidak setara. (Jian, dkk 2011), Fungsi keanggotaan merupakan pemetaan titik-titik yang didapat dari himpunan fuzzy kedalam keanggotaan yang memiliki interval antara 0 sampai dengan 1. Salah satu cara untuk mendapatkan nilai keanggotaan adalah dengan pendekatan fungsi.

Sistem Informasi dan Klasifikasi

Data awal yang didapatkan dalam Rough Set adalah data yang disusun didalam tabel atau bisa disebut juga sebagai database atau sistem informasi. Dasar-dasar untuk menentukan Rough Set adalah menentukan perkiraan atas dan perkiraan bawah data yang berada didalam tabel tersebut sehingga diklasifikasikan sehingga membentuk data yang lebih kecil inilah merupakan konsep Rough Set yang diharapkan. Secara umum Algoritma Rough Set adalah sebagai berikut: (Hasherni, dkk, 1997).

- Langkah 1- Mengurangi sistem informasi vertikal dan horizontal (sistem reduksi).
- Langkah 2- Menghasilkan bagian dan klasifikasi.
- Langkah 3- Menghasilkan ruang perkiraan bawah dan atas.
- Langkah 4- Ekstrak aturan lokal (tertentu, mungkin, dan perkiraan).
- Langkah 5- End.

Sistem Informasi dan Hubungan Indiscernibility

Sistem informasi yang didapat dari database akan diinformasikan menjadi Rough Set. Dan sistem informasi ada dua, yaitu conditional attribute dan decision system. Tiap-tiap baris dikatakan object sedangkan tiap kolom dikatakan attribute. Dimana U adalah set terhingga yang tidak kosong dari objek yang disebut dengan universe dan A set terhingga tidak kosong dari atribut dimana (Nurhayati, 2014): IS = {U,A} Untuk tiap ∈ A. Set V disebut value set dari a. Dimana : IS adalah Information System U = { x1, x2, ..., xm }, yang merupakan sekumpulan example. A = { a1, a2, ..., an }, sekumpulan atribut kondisi secara berurutan. Penjelasan dapat dilihat Tabel 2.1.

Tabel 2.1. *Information System*

Mahasiswa	Kategori	Jurusan	Tempat_Lahir
1	PhD	History	Detroit
2	MS	Chemistry	Akron
3	MS	History	Detroit
4	BS	Math	Detroit
5	BS	Chemistry	Akron
6	PhD	Computing	Cleveland
7	BS	Chemistry	Clevelan
8	PhD	Computing	Akron

Tiap-tiap baris mempresentasikan objek, terdiri dari m example, seperti E1, E2, ..., Em. Sedangkan kolom

mempresentasikan atribut, terdiri dari kategori, Jurusan, dan Tempat_Lahir. Dalam penggunaan Information System terdapat outcome dari klasifikasi yang telah diketahui yang disebut dengan atribut keputusan. Information System tersebut dapat disebut dengan decision system. Decision system dapat dilihat sebagai: IS = (U,{A,C})

Dimana : IS adalah Information System U = { x1, x2, ..., xm }, yang merupakan sekumpulan example. A = { a1, a2, ..., an }, sekumpulan atribut kondisi secara berurutan. C = decission attribute (keputusan) Penjelasan dapat dilihat Tabel 2.2 (Chan, 2007):

Tabel 2.2. *Decision System*

Mahasiswa	Kategori	Jurusan	Tempat_Lahir	Nilai
1	PhD	History	Detroit	A
2	MS	Chemistry	Akron	A

Tabel 2.2. *Decision System* (Lanjutan)

Mahasiswa	Kategori	Jurusan	Tempat_Lahir	Nilai
3	MS	History	Detroit	C
4	BS	Math	Detroit	B
5	BS	Chemistry	Akron	C
6	PhD	Computing	Cleveland	A
7	BS	Chemistry	Clevelan	C
8	PhD	Computing	Akron	A

Tiap-tiap baris mempresentasikan objek, terdiri dari m example, seperti E1, E2, ..., Em. Sedangkan kolom

mempresentasikan atribut, terdiri dari kategori, Jurusan, Tempat_Lahir, dan Nilai.

Set dan Set Approximation

Menetapkan Teori Rough Set harus dikalsifikasikan kedalam satu set dan membuatnya menjadi bagian dari himpunan. Pendekatan yang lebih rendah, pendekatan atas, wilayah negatif, dan batas set X tentang I, masing-masing adalah :

$$\underline{apr}_P(X) = \cup\{x \in U : I(x) \subseteq X\}$$

$$\overline{apr}_P(X) = \cup\{x \in U : I(x) \cap X \neq \phi\}$$

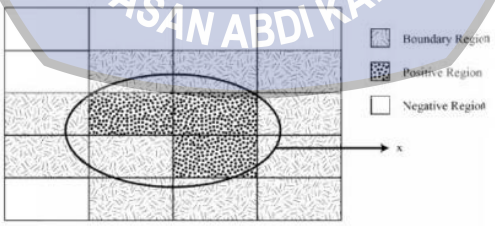
$$Bnd_P(X) = \overline{apr}_P(X) - \underline{apr}_P(X)$$

Keterangan : $\underline{apr}_P(X)$ adalah Pendekatan yg lebih rendah dari set X sehubungan dengan P (X tentu sehubungan dengan P)

$\overline{apr}_P(X)$ adalah Pendekatan yg lebih tinggi dari set X sehubungan dengan P (yang mungkin X dalam p P)

$Bnd_P(X)$ adalah diklasifikasikan baik sebagai X atau tidak X sehubungan dengan P

Adapun proses approximation dapat dilihat pada Gambar 2.1.



Gambar 2.1. *Positive, boundary, and negative regions* pada sebuah *set* x (**sumber :** Yao, dkk, 1997)

Dari Tabel 2.2 diatas dapat dijelaskan bagaimana menghitung approximation.

- U\{kategori} = {(1, 6, 8)}, {(2, 3)}, {(4, 5, 7)}
- U\{jurusan} = {(1, 3)}, {(2, 5, 7)}, {(4)}, {(6, 8)}
- U\{tempat_lahir} = {(2, 5, 8)}, {(1, 3, 4)}, {(6, 7)}
- U\{nilai} = {(1, 2, 6, 8)}, {(4)}, {(3, 5, 7)}
- Set X = {(nilai A)} = {1, 2, 6, 8}
- Set B = {jurusan, tempat_lahir} U\B = {{1, 3}, {2, 5}, {4}, {6}, {7}, {8}}
- $\underline{apr}(X) = \{6, 8\}$
- $\overline{apr}(X) = \{1, 2, 3, 5, 6, 8\}$
- $Bnd(X) = \{1, 2, 3, 5\}$

Quality of Approximation and Reduct

Untuk mengukur ketergantungan pengetahuan, kualitas klasifikasi harus didefinisikan. X = {X1, X2, ..., Xn} adalah partisi alam semesta U, di mana Xi (i = 1, 2, ..., n) adalah salah satu kelas X, dan P ⊆ A, maka kualitas perkiraan X adalah

$$\gamma_P(X) = \frac{\sum_{i=1}^{n} \left| \underline{apr}(Xi) \right|}{|U|}$$

Keterangan : $\gamma_P(X)$ adalah *quality of approximation* P terhadap X

$\sum_{i=1}^{n}$ adalah sigma atau jumlah, dimana i= 1,2,..., n

$\underline{apr}(Xi)$ adalah pendekatan yang lebih rendah

U adalah sekumpulan *example*

Perhitungan Reduct dan Information System Berdasarkan Discernable Matrix

Showeron mengajukan metode untuk mengekspresikan pengetahuan dengan matriks discernable pada tahun 1991, yang memiliki banyak keuntungan. Secara khusus, dengan mudah dapat menjelaskan dan menghitung inti dan reduct dari information system. Maka fungsi discernibility didefinisikan sebagai :

$$f(A) = \prod \Sigma\, a(x, y)$$

$$(x, y) \in U \times U$$

Keterangan : f (A) adalah fungsi discernibility

` $\prod$ adalah pi

Σ adalah sigma a adalah variabel boolean

(x, y) adalah objek x dan y

U adalah sekumpulan example

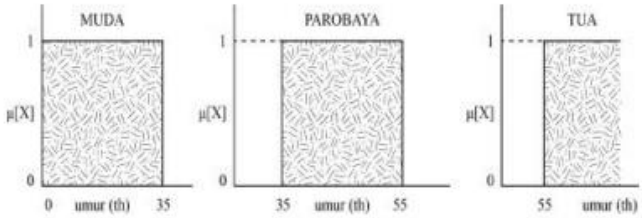
Sebuah matriks discernable dapat digunakan untuk mencari atribut bagian minimal (mengecil) untuk menurunkan gangguan data yang sama seperti pada atribut himpunan A. Untuk menemukan atribut bagian mini, perlu untuk membangun fungsi discernibility yang merupakan fungsi Boolean dan dapat dibangun dalam

metode berikut. Untuk setiap atribut yang dapat mengidentifikasi dua set elemen, seperti a1, a2, a3 menunjuk Boolean konstanta, bentuk fungsi Boolean adalah a1 + a2 + a3 atau (a1 ∨ a2 ∨ a3).

Jika atribut set kosong, maka konstan Boolean adalah 1 Misalnya, dalam kaitannya dengan matriks discernable menunjukkan pada Tabel 2.3 fungsi discernibility adalah:

Tabel 2.3. *Discernable Matrix*

	Set 1	Set 2	Set 3	Set 4	Set 5
Set 1					
Set 2	a_1, a_2, a_3				
Set 3	a_2, a_3	a_1, a_3			
Set 4	a_1, a_3	a_1, a_2	a_1, a_2, a_3		
Set 5	a_1, a_3	a_1, a_2, a_3	a_1, a_2, a_3	a_3	



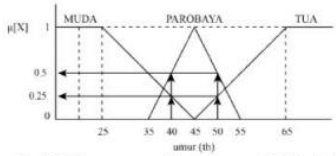
• Apabila seseorang berusia 35 tahun kurang dari 1 tahun, maka ia dikatakan TIDAK PAROBAYA (μPAROBAYA(35 tahun – 1hr) = 0);

• Apabila seseorang berusia 35 tahun lebih dari 1 tahun, maka ia dikatakan PAROBAYA(μPAROBAYA(35) = 1);

• Apabila seseorang berusia 55 tahun, maka ia dikatakan TUA (μTUA (55) = 1);

Dari sini bisa dijelaskan bahwa pemakaian himpunan crisp untuk umur sangat tidak cocok. Adanya perubahan kecil saja sudah mengakibatkan perbedaan yang signifikan dalam hal pemilihan himpunan.

Gambar 2.4 menunjukkan himpunan fuzzy untuk variabel umur



Gambar 2.4. Himpunan *Fuzzy* untuk Variabel umur

Pada Gambar 2.4, dapat dilihat bahwa:

- Seseorang yang berumur 40 tahun, termasuk dalam himpunan MUDA dengan µMUDA(40) = 0,25; namun dia juga termasuk dalam himpunan PAROBAYA dengan µPAROBAYA(40) = 0,5.

- Seseorang yang berumur 50 tahun, termasuk dalam himpunan TUA dengan µTUA(50) = 0,25; namun dia juga termasuk dalam himpunan PAROBAYA dengan µPAROBAYA(50) = 0,5.

Disini terlihat jelas perbedaan antara himpunan crisp dan himpunan fuzzy. Pada himpunan crisp, nilai keanggotaan hanya ada 2 kemungkinan, yaitu 0 dan 1, pada himpunan fuzzy nilai keanggotaan terletak pada rentang 0 sampai 1.

Berbeda lagi kita lihat perbedaan antara fuzzy dan probabilitas. Keduanya memiliki nilai pada interval [0,1]. Keanggotaan fuzzy memberikan suatu ukuran rentang terhadap hasil keputusan. Sedangkan probabilitas memberikan seberapa sering nilai 0 sampai 1 sering muncul. Misalnya jika nilai keanggotaan suatu himpunan fuzzy MUDA adalah 0,9; maka tidak perlu dipermasalahkan berapa seringnya nilai itu diulang secara individual untuk mengharapkan suatu hasil yang hampir pasti MUDA.

Dilain pihak, nilai probabilitas 0,9 MUDA berarti 10% dari himpunan tersebut diharapkan TIDAK MUDA.

Himpuanan fuzzy memiliki 2 atribut, yaitu:

a. Linguistik, yaitu bahasa yang digunakan sehari-hari yang berupa kata-kata, bukan angka seperti MUDA, PAROBAYA, TUA.

b. Numeris, yaitu suatu nilai (angka) yang menunjukkan ukuran dari suatu variabel seperti: 40, 25, 50, dsb.

Ada beberapa hal yang perlu diketahui dalam memahami sistem fuzzy, yaitu:

a. Variabel Fuzzy Merupakan suatu variabel yang nilainya tidak pasti/relatif. Seperti: umur, temperatur, permintaan, dsb.

b. Himpunan Fuzzy Merupakan suatu himpunan yang terdapat didalam variabel fuzzy. Seperti: umur {MUDA, PAROBAYA, TUA}, Temperatur {dingin, sejuk, normal, hangat, panas}.

c. Semesta Pembicaraan Merupakan nilai yang diperbolehkan untuk dioperasikan dalam variabel fuzzy. Misalkan: umur batas variabel [0 +∞], dan temperatur batas variabel [0 100].

d. Domain Merupakan batas nilai yang diizinkan dalam himpunan fuzzy. Contoh domain himpunan fuzzy:

- MUDA = [0 45]
- PAROBAYA = [35 55]
- TUA = [45 +∞]
- DINGIN = [0 20]
- SEJUK = [15 25]

- NORMAL = [20 30]
- HANGAT = [25 35]
- PANAS = [30 40]

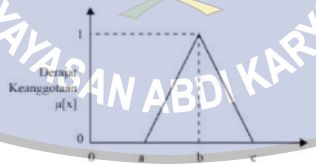
Pada penjelasan diatas, dapat dilihat bahwa untuk variabel MUDA batas nilai yang diizinkan 0 sampai 45. Untuk PAROBAYA batas nilai yang diizinkan 35 sampai 55, dst.

Fungsi Keanggotaan

Merupakan pemetaan titik-titik kurva yang didapat dari himpunan fuzzy kedalam nilai keanggotaan yang memiliki interval antara 0 sampai dengan 1. Salah satu cara untuk mendapatkan nilai keanggotaan adalah dengan pendekatan fungsi, yaitu:

a. Representasi Kurva Segitiga Kurva segitiga

pada dasar merupakan gabungan antar 2 garis (linear) terlihat pada Gambar 2.5



Gambar 2.5. Kurva Segitiga

Fungsi Keanggotaan:

$$\mu[x]=\begin{cases} 0; & x \le a \text{ atau } x \ge c \\ (x-a)/(b-a); & a \le x \le b \\ (b-x)/)c-b); & b \le x \le c \end{cases}$$

Keterangan :

μ adalah fungsi keanggotaan

x adalah variabel

'a' adalah batas awal, dengan derajat keanggotaan 0

'b' adalah batas kedua, dengan derajat keanggotaan 1

'c' adalah batas ketiga, dengan derajat keanggotaan 0

b. Representasi Kurva Trapesium Kurva Segitiga pada dasarnya seperti bentuk segitiga, hanya saja ada beberapa titik yang memiliki nilai keanggotaan 1, terlihat pada Gambar 2.6.

Metode Tsukamoto

Secara umum bentuk model Fuzzy Tsukamoto adalah: If (X IS A) and (Y IS B) Then (Z IS C)

Keterangan : A, B, dan C adalah himpunan fuzzy. X, Y, dan Z adalah variabel

Misalkan duketahui 2 rule berikut.

IF (x is A1) AND (y is B1) THEN (z is C1)

IF (x is A2) AND (y is B2) THEN (z is C2)

Dalam inferensinya, metode Tsukamoto menggunakan tahapan berikut.

1. Fuzzyfkasi

2. Pembentukan basis pengetahuan fuzzy (rule dalam bentuk IF … THEN)

3. Mesin inferensi

Menggunakan fungsi implikasi MIN (Gambar 2.8) untuk mendapatkan nilai α- predikat tiap-tiap rule (α1, α2, α3, …., αn). Kemudian masing-masing nilai α-predikat ini digunakan untuk menghitung keluaran hasil inferensi secara tegas (crisp) masing-masing rule (z1, z2, z3, …., zn)

4. Defuzzyfikasi Menggunakan metode rata-rata (average)

$$Z^* = \frac{\sum \alpha i z j}{\sum \alpha i}$$

Keterangan :

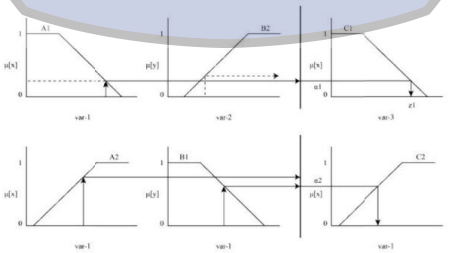
Z * adalah rata-rata (average)

∑ adalah sigma atau jumlah

αi adalah alpha, i = 1, 2, ..., n

zj adalah fungsi keanggotaan, j = 1, 2, ..., n

Gambar 2.7 menunjukkan skema penalaran fungsi implikasi MIN dan proses defuzzyfikasi dilakukan dengan cara mencari nilai rata-ratanya



Gambar 2.7. *Inferensi* dengan Menggunakan Metode *Tsukamoto* (**Sumber:** Jang, 1997)

Penelitian Terdahulu

Beberapa penelitian terdahulu menurut (Radwan, 2013) data yang tidak konsisten pada pasien, fitur yang tidak relevan, berlebihan, hilang, dan besar. Dalam tulisannya, Rough set teori yang digunakan untuk mencoba untuk menghitung set minimal reducts, yang digunakan untuk mengekstrak set minimal aturan keputusan yang menjelaskan hubungan kesamaan antara aturan. Menurut (Sadek, 2013) NNIV-RS (Neural Network dengan Indikator Variabel menggunakan Rough Set untuk pengurangan atribut) algoritma digunakan untuk mengurangi jumlah sumber daya komputer seperti memori dan CPU waktu yang diperlukan untuk mendeteksi serangan. Teori Rough Set digunakan untuk memilih keluar fitur reducts. Indikator Variabel digunakan untuk mewakili dataset yang lebih efisien. Menurut (Maharani, 2008) dalam jurnalnya Rangkaian Jaringan Syaraf Tiruan berbasis neuron rough memiliki kemampuan learning berdasarkan data-data masukan, dan membantu sistem fuzzy dalam menentukan rule yang terbaik dalam memprediksi data, sedangkan rough set sendiri akan mengklasifikasikan data acak yang melewatinya. Metode RANFIS ini digunakan untuk menemukan suatu pola perubahan tertentu, dan akan selalu belajar dari kesalahan atau error sebelumnya, sehingga akan didapatkan nilai akurasi yang sangat baik. Menurut (Arief, dkk, 2010) dalam jurnalnya Klasterisasi teks mempunyai salah satu permasalahan utama dalam mengklasifikasikan jenis teks yang mempunyai sifat uncertain atau sulit dikategorikan pada data berdimensi yang tinggi dan menyebar. Pada penelitian ini diperkenalkan metode baru dalam penyusunan

klasterisasi teks berbasis Roughset untuk persamaan kata. Metode yang diusulkan dalam penelitian ini bernama Max-max Roughness (MMR). Roughset dipilih karena terbukti mampu mengatasi permasalahan data uncertain. Metode klasterisasi teks umumnya dilakukan dengan mencari persamaan dokumen berdasarkan bobot semua kata dalam masing-masing dokumen (perbandingan obyek), sedangkan metode ini menggunakan beberapa kata kunci yang mempunyai perwakilan paling besar (max roughness) dalam dokumen untuk proses klasterisasi (perbandingan atribut). Menurut (Kothari, 2008) dalam jurnalnya Penggunaan teori himpunan kasar pada tahap preprocessing untuk pengurangan dimensi yang ditargetkan sangat penting dalam kasus JST Unsupervised pola berbasis pengklasifikasi. atribut tersebut dikompresi dengan menghapus hasil redudansi, berkurangnya set atribut bertindak sebagai masukan untuk saraf tanpa pengawasan jaringan.

TOPSIS (*Technique for Order Preference by Similarity to Ideal Solution*) merupakan salah satu metode pengambilan keputusan multikriteria yang didasarkan pada konsep bahwa alternatif yang terbaik tidak hanya memiliki jarak terpendek dari solusi ideal positif tetapi juga memiliki jarak terpanjang dari solusi ideal negatif. Konsep ini banyak digunakan untuk menyelesaikan masalah keputusan secara praktis Atas dasar itulah penulis tertarik untuk mengambil tema Sistem Pendukung Keputusan Pembelian Vending Machine

Dengan Metode *Technique For Others Reference by Similarity to Ideal Solution (TOPSIS)* Studi kasus PT.KAI Commuter Jabodetabek.

Tiket Elektronik (Electronic Ticket)

Electronik ticketing yaitu suatu cara untuk mendokumentasikan proses penjualan dari aktifitas perjalanan pelanggan tanpa harus mengeluarkan dokumen berharga secara fisik ataupun *paper ticket*. Semua informasi mengenai *electronic ticketing* disimpan secara digital dalam sistem komputer milik airline. Contoh *paper ticket* di sebagian besar bandara di Indonesia sebagai bukti pengeluaran *E-Ticketing*, pelanggan akan diberikan *Itinerary Receipt* yang hanya berlaku sebagai alat untuk masuk ke dalam bandara di Indonesia yang masih mengharuskan penumpang untuk membawa tanda bukti perjalanan. *Elektronik ticketing* adalah peluang untuk meminimalkan biaya dan mengoptimalkan kenyamanan penumpang. *E-ticketing* mengurangi biaya proses tiket, menghilangkan fomulir kertas dan meningkatkan fleksibilitas penumpang dan agen perjalanan dalam membuat perubahan-perubahan dalam jadwal perjalanan. *Electronik Ticketing* berisi juga rincian perjalanan yang tercantum di dalam e-Ticket yang berisi nama penumpang, rute perjalanan, waktu penerbangan, nomor penerbangan, kelas tiket, dan harga tiket.

Sistem Pendukung Keputusan (SPK)

Menurut Erniyati dalam Mc Leod (2011:171), Sistem pendukung keputusan adalah sistem penghasil informasi yang ditujukan pada suatu masalah tertentu yang harus dipecahkan oleh manager dan dapat

membantu manager dalam pengambilan keputusan. Sistem pendukung keputusan (SPK) adalah bagian dari sistem informasi berbasis komputer (termasuk sistem pengetahuan) yang dipakai untuk mendukung pengambilan keputusan dalam suatu organisasi atau perusahaan. SPK merupakan penggabungan sumber–sumber kecerdasan individu dengan kemampuan komponen untuk memperbaiki kualitas keputusan. Sistem Pendukung Keputusan juga merupakan sistem informasi berbasis komputer untuk manajemen pengambilan keputusan yang menangani masalah–masalah semi struktur Dengan pengertian diatas dapat dijelaskan bahwa sistem pendukung keputusan bukan merupakan alat pengambilan keputusan, melainkan merupakan sistem yang membantu pengambil keputusan dengan melengkapi mereka dengan informasi dari data yang telah diolah dengan relevan dan diperlukan untuk membuat keputusan tentang suatu masalah dengan lebih cepat dan akurat. Sehingga sistem ini tidak dimaksudkan untuk menggantikan pengambilan keputusan dalam proses pembuatan keputusan.

Sistem pendukung keputusan adalah sistem informasi berbasis komputer yang interaktif, dengan cara mengolah data dengan berbagai model untuk memecahkan masalah-masalah yang tidak terstruktur sehingga dapat memberikan informasi yang bisa digunakan oleh para pengambil keputusan dalam membuat sebuah keputusan. Dalam sebuah sistem pendukung keputusan, sumber daya intelektual yang dimiliki seseorang dipadukan dengan kemampuan komputer untuk memba

ntu meningkatkan kualitas dari keputusan yang diambil. Pengambilan keputusan merupakan sebuah proses memilih sebuah tindakan diantara beberapa alternatif yang ada, sehingga tujuan yang diinginkan dapat tercapai (Turban et al. 2005).

A. Pengertian Data Mining Data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam database. Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar. (Turban, dkk. 2005) Definisi umum dari data mining itu sendiri adalah proses pencarian pola-pola yang tersembunyi (hidden patern) berupa pengetahuan (knowledge) yang tidak diketahui sebelumnya dari suatu sekumpulan data yang mana data tersebut dapat berada di dalam database, data werehouse, atau media penyimpanan informasi yang lain. Hal penting yang terkait di dalam data mining adalah:

1. Data mining merupakan suatu proses otomatis terhadap data yang sudah ada.

2. Data yang akan diproses berupa data yang sangat besar.
3. Tujuan data mining adalah mendapatkan hubungan atau pola yang mungkin memberikan indikasi yang bermanfaat. (Kusrini dan Emha Taufiq, 2009).

B. Arsitektur Sistem Data Mining Arsitektur utama dari sistem data mining, pada umumnya terdiri dari beberapa komponen sebagai berikut:

1. Database, data warehouse, atau media penyimpanan informasi, terdiri dari satu atau beberapa database, data warehouse, atau data dalam bentuk lain. Pembersihan data dan integrasi data dilakukan terhadap data tersebut.

2. Database, data warehose, bertanggung jawab terhadap pencarian data yang JTKSI, Vol.03 No.02 Mei 2020 ISSN : 2620-3022 Hal. 74-83 76 relevan sesuai dengan yang diinginkan pengguna atau user.

3. Basis pengetahuan (Knowledge Base), merupakan basis pengetahuan yang digunakan sebagai panduan dalam pencarian pola.

4. Data mining engine, merupakan bagaian penting dari sistem dan idealnya terdiri dari kumpulan modul-modul fungsi yang digunakan dalam proses karakteristik (characterization), klasifikasi (clasiffication), dan analisis kluster (cluster analysis). Dan merupakan bagian dari software yang menjalankan program berdasarkan algoritma yang ada.

5. Evaluasi pola (pattern evaluation), komponen ini pada umumnya berinteraksi dengan modul-modul data mining. Dan bagian dari software yang berfungsi untuk menemukan pattern atau pola-pola yang terdapat dalam database yang diolah sehingga nantinya proses data mining dapat menemukan knowledge yang sesuai.

6. Antar muka (Graphical user interface), merupakan modul komunikasi antara pengguna atau user dengan

sistem yang memungkinkan pengguna berinteraksi dengan sistem untuk menentukan proses data mining itu sendiri. Arsitektur Data Mining C. Teknik Data Mining Data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual. Perlu diingat bahwa kata mining sendiri berarti usaha untuk mendapatkan sedikit data berharga dari sejumlah besar data dasar. Karena itu data mining sebenarnya memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (artificial intelligent), machine learning, statistik dan basis data. Beberapa teknik yang sering disebut-sebut dalam literatur data mining antara lain yaitu : Classification decision tree Classification adalah suatu teknik dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan.Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi data yang ada yang telah diklasifikasi dengan menggunakan hasilnya untuk memberikan sejumlah aturan. Aturan-aturan tersebut digunakan pada data-data baru untuk diklasifikasi. Teknik ini menggunkan supervised induction, yang memanfaatkan kumpulan pengujian dari record yang terklasifikasi untuk menentukan kelas-kelas tambahan. Salah satu contoh yang mudah dan populer adalah dengan Decision tree, yaitu salah satu metode klasifikasi yang paling populer karena mudah untuk diinterpretasi. Decision tree adalah model prediksi menggunakan struktur pohon atau struktur berhirarki. Decision tree adalah struktur flowchart yang menyerupai tree (pohon), dimana setiap simpul internal menandakan suatu tes pada atribut, setiap cabang merepresentasikan hasil tes, dan simpul daun

merepresentasikan kelas atau distribusi kelas. Alur pada decision tree di telusuri dari simpul akar ke simpul daun yang memegang prediksi kelas untuk contoh tersebut. Decision tree mudah untuk dikonversi ke aturan klasifikasi (classification rules). Contoh classification decision tree Association Association digunakan untuk mengenali kelakuan dari kejadian-kejadian khusus atau proses dimana link asosiasi muncul pada setiap kejadian. Penting tidaknya suatu aturan assosiatif dapat diketahui dengan dua parameter, support yaitu prosentasi kombinasi atribut tersebut dalam basis data dan confidence yaitu kuatnya hubungan antar atribut dalam aturan asosiatif. Motivasi awal pencarian association rule berasal dari keinginan untuk menganalisa data transaksi supermarket, ditinjau dari perilaku customer dalam membeli produk. Association rule ini menjelaskan seberapa sering suatu produk dibeli secara bersamaan. Sebagai contoh, association rule "beer => diaper (80%)" menunjukkan bahwa empat dari lima customer yang membeli beer juga membeli diaper. Dalam suatu association rule X => Y, X disebut dengan antecedent dan Y disebut dengan consequent. Clustering JTKSI, Vol.03 No.02 Mei 2020 ISSN : 2620-3022 Hal. 74-83 77 Clustering digunakan untuk menganalisis pengelompokkan berbeda terhadap data, mirip dengan klasifikasi, namun pengelompokkan belum didefinisikan sebelum dijalankannya tool data mining. Biasanya menggunakan metode neural network atau statistik. Clustering membagi item menjadi kelompok-kelompok yang ditemukan tool data mining. Contoh clustering D. Klasifikasi naïve bayes Teknik klasifikasi adalah suatu proses yang menemukan properti-properti yang sama pada sebuah himpunan obyek di

dalam sebuah basis data, dan mengklasifikasikannya ke dalam kelaskelas yang berbeda menurut model klasifikasi yang ditetapkan. Klasifikasi dalam data mining dikelompokkan ke dalam teknik pohon keputusan, Bayesian (Naïve Bayesian dan Bayesian Belief Networks), Jaringan Saraf Tiruan (Backpropagation), teknik yang berbasis konsep dari penambangan aturan-aturan asosiasi, dan teknik lain (k-Nearest Neighboor, algoritma genetik, teknik dengan pendekatan himpunan rough dan fuzzy). Setiap teknik memiliki kelebihan dan kekurangannya sendiri, berikut gambar pengelompokan teknik klasifikasi. Pengelompokan Teknik Klasifikasi Secara umum, proses klasifikasi dapat dilakukan dalam dua tahap, yaitu proses belajar dari data pelatihan dan klasifikasi kasus baru. Pada proses belajar, algoritma klasifikasi mengolah data pelatihan untuk menghasilkan sebuah model. Setelah model diuji dan dapat diterima, pada tahap klasifikasi, model tersebut digunakan untuk memprediksi kelas dari kasus baru untuk membantu proses pengambilan keputusan (Han et al.,2001; Quinlan, 1993). Kelas yang dapat diprediksi adalah kelaskelas yang sudah terdefinisi pada data pelatihan. Karena proses klasifikasi kasus baru cukup sederhana, penelitian lebih banyak ditujukan untuk memperbaiki teknik-teknik pada proses belajar. Skema Klasifikasi secara Umum a. Algoritma Rough Set Teori Rough set sampai saat ini pendekatan lain untuk ketidakjelasan (Pawlak, 1982). Demikian pula untuk teori himpunan fuzzy bukan merupakan alternatif untuk teorihimpunan klasik tetapi tertanam di dalamnya. Teori Rough Set dapat dilihat sebagai implementasi khusus dari gagasan G. Frege (1983) tentang ketidakjelasan, yaitu ketidaktepatan dalam

pendekatan ini dinyatakan oleh batas wilayah dari suatu himpunan, dan bukan oleh keanggotaan parsial, seperti dalam teori himpunan fuzzy. Konsep Rough Set dapat didefinisikan cukup umum dengan cara operasi topologi, interior dan penutupan, yang disebut pendekatan. Tujuan analisis Rough Set adalah untuk mendapatkan rule yang klasifikasi setelah dilakukan pengumpulan data (Maharani, 2008). Rule disini sudah dikalsifikasikan setelah mendapatkan reduct. Rough Set menentukan teorinya menggunakan perkiraan, yaitu yang ditentukan oleh fungsi keanggotaan. Rough Set bisa juga menentukan teorinya tanpa menggunakan perkiraan. Karena fungsi keanggotaan bukanlah konsep primitif dalam pendekatan yang dalam hal ini kedua defenisi tidak setara. (Jian, dkk 2011), Fungsi keanggotaan merupakan pemetaan titik-titik yang didapat dari himpunan fuzzy kedalam keanggotaan yang memiliki interval antara 0 sampai dengan 1. Salah satu cara untuk mendapatkan nilai keanggotaan adalah dengan pendekatan fungsi. Di dalam Metode Rough Set terdapat beberapa langkah - langkah penyelesaian masalah, yaitu sebagai berikut: 1. Decision System tersebut dilakukan teknik klasifikasi kriteria yang disebut "Equivalen Class" 2. Kemudian dilakukan proses Discernibility Matrix atau Discernibility Matrix Modulo D 3. Proses "Reduction" 4. Untuk memperoleh hasil akhir dilakukan proses "General Rules" Langkah-langkah JTKSI, Vol.03 No.02 Mei 2020 ISSN : 2620-3022 Hal. 74-83 78 dalam menjalankan metode rough set di atas dapat digambarkan pada gambar berikut ini. Proses Algoritma Rough Set b. Algoritma Naive Bayes Naive Bayes merupakan sebuah pengklasifikasian probalistik sederhana yang menghitung sekumpulan probabilitas

dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma menggunakan teorema bayes dan mengansumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas. Naive Bayes juga didefinisikan sebagai pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuan inggis Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya (Saleh, 2015). Untuk menjelaskan metode Naive Bayes, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang di analisis tersebut. Karena itu, metode Naive Bayesdi atas disesuaikan sebagai berikut (Saleh, 2015) : Di mana Variabel C mempresentasikan kelas, sementara variabel F1...Fn mempresentasikan karakteristik petunjuk yang dibutuhkan untuk menentukan klasifikasi. Maka rumus tersebut menjelaskan bahwa peluang masuknya sampel karakteristik tertentu dalam kelas C (Posterior) adalah peluang munculnya kelas C (sebelum masuknya sampel tersebut, seringkali disebut prior), dikali dengan peluang kemunculan karakteristik – karakteristik sampel pada kelas C (disebut likelihood), dibagi dengan peluang kemunculan karakteristik – karakteristik secara global (disebut juga evidence). Karena itu, rumus di atas dapat pula ditulis secara sederhana sebagai berikut (Saleh, 2015): Nilai Evidence selalu tetap untuk setiap kelas pada satu sampel. Nilai dari Posterior tersebut nantinya akan dibandingkan dengan nilai – nilai posterior kelas lainnya untuk menentukan ke kelas apa suatu sampel akan diklasifikasikan. Alur metode Naive Bayes dapat

digambarkan sebagai berikut : Alur Metode Naive Bayes Adapun persamaan yang digunakan untuk menghitung nilai rata – rata (mean)

a. Cari nilai probabilistik dengan cara menghitung jumlah data yang sesuai dari kategori yang sama dibagi dengan jumlah data pada kategori tersebut.

b. Mendapatkan nilai dalam tabel mean, standar deviasi dan probabilitas.

c. Solusi yang dihasilkan E. Prototyping Salah satu metode pengembangan perangkat lunak yang banyak digunakan adalah prototyping. Selama proses pembuatan sistem, developer dan client dapat saling berinteraksi dengan menggunakan metode prototyping ini. Agar model prototyping berhasil adalah dengan mendefinisikan aturan-aturan yang harus disepakati client dan developer yaitu kesepakatan bahwa prototipe yang dibangun untuk mendefinisikan semua kebutuhan client. Prototipe akan dihilangkan sebagian atau seluruhnya dan perangkat lunak aktual direkayasa dengan kualitas dan implementasi yang sudah ditentukan. Bagan dari model prototyping adalah sebagai berikut : Model Prototyping (Pressman, 2010) Kelebihan prototyping adalah sebagai berikut:

1. Adanya komunikasi yang baik antara developer dan client

2. Developer terfokus dengan kebutuhan client

3. Client berperan aktif dalam pengembangan sistem

4. Pengembangan sistem membutuhkan waktu yang lebih efisien

5. Penerapan menjadi lebih mudah, karena pengguna mengetahui apa yang diharapkan Sementara itu, prototyping juga memiliki kelemahan yaitu:

1. Client terkadang kuarang menyadari bahwa perangkat lunak yang ada belum mencantumkan kualitas perangkat lunak secara keseluruhan dan juga belum memikirkan kemampuan pemeliharaan untuk jangka waktu panjang.

2. Developer biasanya ingin proyek cepat selesai. Sehingga menggunakan algoritma dan bahasa pemograman sederhana agar prototipe lebih cepat selesai tanpa memikirkan bahwa software tersebuat hanya merupakan blue print dari sistem yang akan dikembangkan kemudian.

3. Hubungan customer dengan komputer yang disediakan mungkin tidak mencerminkan teknik perancangan yang baik Prototyping bekerja dengan baik pada penerapan-penerapan yang berciri sebagai berikut (Sommerville, 2007):

1. Resiko Tinggi, yaitu untuk masalah masalah tidak terstruktur dengan baik, ada perubahan yang besar dari waktu ke waktu, dan adanya persyaratan data yang tidak menentu.

2. Interaksi pemakai penting Sistem harus menyediakan dialog on-line antara costumer dan komputer

3. Perilaku pengguna yang sangat sulit ditebak atau mudah berubah

4. Sistem yang inovatif. Sistem tersebut membutuhkan cara penyelsaian masalah dan penggunaan perangkat keras yang mutakhir

5. Perkiraan tahap penggunaan ssstem yang pendek.

Daftar Pustaka

https://www.sekawanmedia.co.id/blog/sistem-pendukung-keputusan/

http://eprints.umpo.ac.id/714/2/BAB%20I.pdf

http://jurnal.ubl.ac.id/index.php/expert/article/view/1225/0

http://anindyadev.com/artikel/lainnya/metode-topsis-dalam-sistem-pendukung-keputusan.htm

https://raharja.ac.id/2020/04/09/kelebihan-dan-kekurangan-metode-topsis/

https://raharja.ac.id/2020/04/02/metode-topsis-technique-for-others-reference-by-similarity-to-ideal-solution/

https://repository.uin-suska.ac.id/3060/3/BAB%20II.pdf

https://jurnal.kominfo.go.id/index.php/jtik/article/download/823/470

https://medium.com/@infharis/data-mining-definisi-dan-cara-kerja-algoritma-apriori-untuk-pencarian-association-rule-a44a8f864a61

https://repository.bsi.ac.id/index.php/unduh/item/303328/BAB-II.pdf

https://www.anakblogger.com/2020/12/kelebihan-kekurangan-algoritma-apriori.html

https://adoc.pub/queue/bab-2-tinjauan-pustakafd0f8d2de1164b240f8f730b6250227180895.html

“ Buku adalah bagian dari diriku, tanpa buku aku bukan siapa siapa dan tak bisa apa apa, manusia bisa mati, tapi buku dan pemikiran kita tidak akan mati, maka menulis lah”

Dr. Arman Syah Putra S.Kom., M.M., M.Kom.

PENULIS BUKU KONSEP DATA MINING

Turkhamun Adi Kurniawan S.T M.Kom
Universitas Satya Negara Indonesia

Dr. Arman Syah Putra S.Kom MM M.Kom
Universitas Bina Nusantara

Fatrilia Rasyi Radita, S.Pd.I, M.Pd.I
STMIK Insan Pembangunan

Muhammad Hilman Fakhriza M.Kom
Universitas Nusa Mandiri

Juniana Husna, S.Si., M.Sc
Universitas Abulyatama

V.H.Valentino S.Kom M.M.SI
Universitas Indraprasta PGRI

ISBN 978-623-481-037-0