

MODUL PRAKTIKUM

DATA WAREHOUSING

DAN DATA MINING



Oleh :

Yusuf Sulistyo Nugroho, S.T., M.Eng.

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS KOMUNIKASI DAN INFORMATIKA

UNIVERSITAS MUHAMMADIYAH SURAKARTA

Daftar Isi

Halaman Judul	i
Daftar Isi	ii
MODUL 1 Perancangan Star Schema dan Snowflake	1
A. Tujuan	1
B. Landasan Teori	1
C. Alat dan Bahan	7
D. Pengenalan Perangkat Lunak Data Warehousing	7
E. Tugas	10
MODUL 2 Proses ETL: Ekstraksi dan Transformasi Data	11
A. Tujuan	11
B. Landasan Teori	11
C. Alat dan Bahan	13
D. Langkah-langkah Praktikum	13
E. Tugas	48
MODUL 3 Proses ETL: Data Cleansing	49
A. Tujuan	49
B. Landasan Teori	49
C. Alat dan Bahan	49
D. Langkah-langkah Praktikum	50
E. Tugas	56
MODUL 4 Proses ETL: Pembuatan Tabel Fakta	58
A. Tujuan	58
B. Landasan Teori	58
C. Alat dan Bahan	58
D. Langkah-langkah Praktikum	58
E. Tugas	85
MODUL 5 Pivot Table dan Chart	88
A. Tujuan	88
B. Landasan Teori	88
C. Alat dan Bahan	88
D. Langkah-langkah Praktikum	89
E. Tugas	103

MODUL 6 PENGENALAN DATA MINING	104
F. Tujuan	104
G. Landasan Teori	104
H. Alat dan Bahan	107
I. Langkah-langkah Praktikum	107
J. Tugas	112
MODUL 7 DATA PREPROCESSING	113
F. Tujuan	113
G. Landasan Teori	113
H. Alat dan Bahan	116
I. Langkah-langkah Praktikum	116
J. Tugas	123
MODUL 8 NAÏVE BAYES	124
F. Tujuan	124
G. Landasan Teori	124
H. Alat dan Bahan	126
I. Langkah-langkah Praktikum	126
J. Tugas	141
MODUL 9 DECISION TREE	143
F. Tujuan	143
G. Landasan Teori	143
H. Alat dan Bahan	144
I. Langkah-langkah Praktikum	144
J. Tugas	155
MODUL 10 CLUSTERING: ALGORITMA K-MEANS	156
F. Tujuan	156
G. Landasan Teori	156
H. Alat dan Bahan	157
I. Langkah-langkah Praktikum	157
J. Tugas	166

MODUL 11 INDUKSI DAN ATURAN ASOSIASI	168
A. Tujuan	168
B. Landasan Teori	168
C. Alat dan Bahan	169
D. Langkah-langkah Praktikum	169
E. Tugas	177
 MODUL 12 REGRESI LINIER SEDERHANA	179
A. Tujuan	179
B. Landasan Teori	179
C. Alat dan Bahan	180
D. Langkah-langkah Praktikum	181
E. Tugas	190

MODUL 1

PERANCANGAN STAR SCHEMA DAN SNOWFLAKE

A. Tujuan

1. Mahasiswa mampu menjelaskan prosedur perancangan *Star Schema* atau *Snowflake*
2. Mahasiswa mampu merancang *Star Schema* atau *Snowflake* menggunakan program aplikasi tertentu

B. Landasan Teori

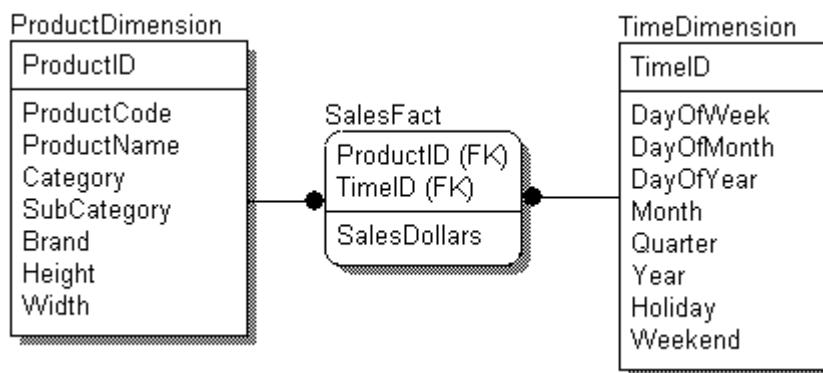
Apa hubungan antara *star schema* dan *snowflake* dengan basis data? Kedua istilah skema ini mewakili struktur basis data yang umum digunakan pada basis data OLAP (*On Line Analytical Processing*) untuk kebutuhan *data warehouse*. Dalam banyak pelajaran tentang basis data, kedua skema ini jarang disampaikan akibat penerapannya yang tidak sesuai untuk model basis data OLTP (*On Line Transactional Processing*). Mekanisme normalisasi juga tidak banyak berlaku untuk kedua jenis skema basis data ini. Fokus utama materi dasar-dasar basis data adalah proses manipulasi data, dalam hal ini bagaimana merancang sistem basis data yang dapat melayani sekian transaksi DML mulai *Insert*, *Update*, dan *Delete*? Bagaimana melakukan normalisasi pada struktur basis data untuk mendapatkan struktur yang ideal? Bagaimana mengatur transaksi antar klien agar tidak muncul *deadlock*? Dan banyak pertanyaan yang muncul terkait dengan sistem basis data.

Struktur data pada OLAP jauh lebih sederhana, mengingat data-data yang akan tersimpan di dalamnya tidak banyak mengalami perubahan dimana lebih banyak transaksi *selection* (*read only* – hanya baca) daripada DML. Jika pada OLTP, konsep ACID (*atomicity*, *consistency*, *isolation*, *durability*) menjadi properti utama yang harus melekat pada setiap transaksi data dari dan ke aplikasi klien maka dalam OLAP yang lebih diutamakan adalah kecepatan perolehan datanya (*data retrieval*). Tidak hanya struktur basis datanya yang berbeda, namun konfigurasi server basis datanya pun akan berbeda antara OLAP dan OLTP.

B.1. Star Schema

Dalam *data warehouse*, data-datanya akan disimpan dalam tabel fakta dan tabel dimensi. Tabel fakta akan menyimpan data-data utama sementara tabel dimensi mendeskripsikan setiap nilai dari suatu dimensi dan dapat direlasikan ke tabel fakta jika diperlukan. Data fakta merupakan data yang terukur besarannya, sebagai contoh adalah jumlah siswa, banyaknya rupiah yang diperoleh, rata-rata IPK, dan sejenisnya. Untuk lebih menjelaskan data fakta, maka kondisi saat data tersebut diukur turut disampaikan. Data kondisi inilah yang dipetakan dalam bentuk data dimensi. Kondisi yang dipetakan dalam dimensi umumnya berupa kondisi waktu, kondisi produk atau item, dan kondisi geografisnya. Mendesain struktur *star schema*, dimulai dengan menentukan data apa yang ingin dilihat oleh pengguna (besarannya) dan bagaimana pengguna melihat data tersebut (kondisi atau dimensinya).

Tabel dimensi memiliki *primary key* sederhana yang mengandung hanya satu atau dua kolom saja. Namun, tabel fakta akan memiliki sekumpulan *foreign key* yang disusun dari *primary key* komposit dan merupakan gabungan kolom-kolom tabel dimensi yang berelasi. Untuk lebih jelasnya, berikut contoh struktur *star schema*.



Gambar 1.1. Contoh *star schema*

Untuk struktur *star schema* seperti gambar 1.1, data dalam tabel fakta yang diukur adalah hasil penjualan (dalam mata uang dollar) berdasarkan dimensi atau kondisi produk yang dijual (*product*) serta waktu penjualan (*time*). Misalkan dimensi produk, yang menyimpan informasi-informasi seputar produk. Produk ini dapat dikelompokkan ke dalam kategori, dan di dalam kategori inipun bisa

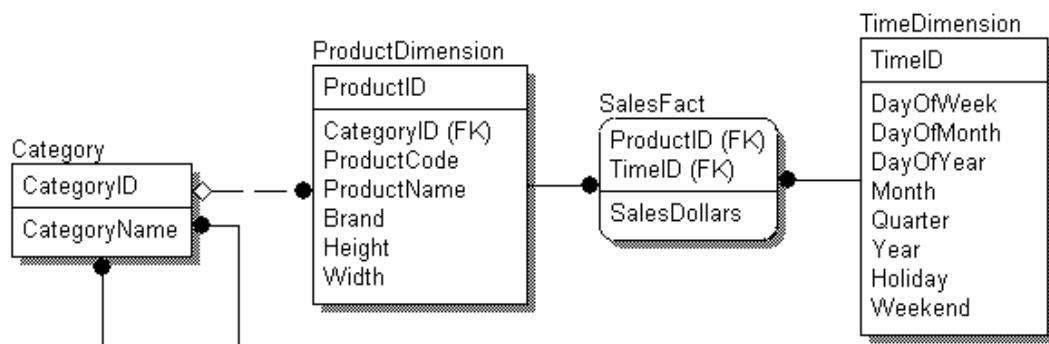
ditemukan sub-kategori. Misalkan dalam sebuah basis data terdapat kode produk **X1001** yang merujuk pada kripik tempe, maka akan masuk ke dalam kategori Nabati, dan sub-kategori Tempe. Untuk lebih mengelompokkan produk tersebut, dapat pula dibuatkan sub-kategori berikutnya. Namun kunci dari informasi produk tersebut tersimpan dalam kolom di tabel dimensi, dan tidak dibutuhkan tabel lain untuk menjelaskan detil produk. Semakin beragam jenis kondisi data yang ingin diamati, maka akan semakin besar ukuran tabel fakta yang dimuat.

Dalam *star schema*, query yang terbentuk antara tabel fakta dan sejumlah tabel dimensi dinamakan *star query*. Setiap tabel dimensi direlasikan dengan tabel fakta berdasarkan kolom *primary key* dan *foreign key*, namun diantara masing-masing tabel dimensi tidak ada yang saling berelasi (tidak ada hubungan data). Query yang terbentuk menyebabkan proses eksekusi yang lebih optimal, karena rencana eksekusi query dalam DBMS akan lebih cepat dengan setiap tabel hanya berelasi dengan satu tabel yang lain.

Ada kalanya tabel dimensi mengandung data yang duplikat pada satu atau lebih kolom. Jika mengikuti azas normalisasi, maka struktur basis data yang terbentuk bukan lagi *star schema* namun akan menjadi *snowflake schema*.

B.2. Snowflake Schema

Struktur basis data ini lebih kompleks dari pada *star schema*, dengan menormalisasi tabel-tabel dimensi yang berukuran besar dengan satu atau lebih kolom yang memiliki duplikasi data. Misalkan jika tabel dimensi *Product* dinormalisasi maka akan menghasilkan struktur seperti berikut:



Gambar 1.2. Contoh bentuk *Snowflake*

Tabel dimensi dinormalisasi untuk mengurangi redundansi data (duplikasi), sehingga struktur tabelnya akan lebih ramping. Dengan pengelompokan ini, data akan lebih mudah dibaca dan membantu pengembang aplikasi untuk menata desain antarmuka sistem dan *filtering* data. Struktur ini akan menghemat kapasitas *storage*, namun waktu eksekusi data akan lebih lama mengingat jumlah tabel dimensi yang direlasikan lebih banyak dan membutuhkan tambahan relasi *foreign key*. Query yang terbentuk lebih kompleks, yang mengakibatkan kinerja query menurun. Pada penerapan yang lebih umum, tabel dimensi tidak diturunkan dengan lebih banyak tabel dimensi lain dan pengelompokan data diatur secara *hard-coded* di kode program aplikasinya.

Fokus penggunaan *datawarehouse* adalah kecepatan akses dan eksekusi data, bukanlah ukuran data yang lebih kecil atau struktur basis data yang lebih ramping. Sehingga bijaksana dalam menetapkan struktur data *star* maupun *snowflake schema* akan menentukan kinerja layanan *datawarehouse* yang dimiliki.

Tahap pertama dari perancangan data warehouse adalah mendefinisikan informasi-informasi apa saja yang dibutuhkan oleh manajemen. Agar kebutuhan ini dapat didefinisikan dengan tepat, maka pemahaman akan peran dan tugas manajemen yang membutuhkan informasi tersebut mutlak harus dilakukan lebih dulu. Jika sudah dipahami, selanjutnya kita hanya tinggal “menjawab” pertanyaan-pertanyaan berikut:

1. Siapa yang membutuhkan informasi dari data warehouse?
2. Informasi apa saja yang dibutuhkan tersebut?
3. Seperti apa layout dan isi informasi-informasi itu?
4. Kapan informasi tersebut digunakan?
5. Untuk keperluan apa?
6. Basis data apa yang menjadi sumber untuk informasi tersebut?

Sebagai contoh, misalkan akan dibuat sebuah data warehouse penjualan (atau data mart penjualan tepatnya) untuk sebuah perusahaan dagang.

1. Siapa yang membutuhkan informasi dari data warehouse?

Manager Pemasaran

2. Informasi apa saja yang dibutuhkan Manager Pemasaran?

Barang apa yang paling banyak terjual di lokasi tertentu sepanjang tahun?

Barang apa yang paling banyak memberikan pendapatan sepanjang tahun?

3. Seperti apa layout dan isi informasi-informasi itu?

Barang yang paling banyak terjual di lokasi tertentu sepanjang tahun:

tahun	kecamatan	kategori	sum(total_penjualan)
2012	BANJARSARI	KONSUMSI	209
2012	JEBRES	ATK	95
2012	LAWEYAN	ATK	109
2012	SERENGAN	ATK	89
2012	JEBRES	KONSUMSI	106
2012	PASAR KLIWON	KONSUMSI	96
2012	BANJARSARI	ATK	200
2012	LAWEYAN	KONSUMSI	193
2012	PASAR KLIWON	ATK	91
2012	SERENGAN	KONSUMSI	139

Barang yang paling banyak memberikan pendapatan sepanjang tahun:

tahun	kategori	sub_kategori	sum(total_penerimaan)
2012	ATK	KERTAS	3560000
2012	ATK	PULPEN	472000
2012	ATK	SPIDOL	1269000
2012	KONSUMSI	SEMBAKO	524000
2012	KONSUMSI	SNACK	1669500

4. Untuk keperluan apa informasi tersebut?

Dasar untuk menentukan strategi penjualan barang

5. Kapan informasi tersebut digunakan?

Awal periode penjualan

6. Basis data apa yang menjadi sumber untuk informasi tersebut?

Basis data penjualan dengan skema sebagai berikut:

a) Kategori (#kelompok, sub_kategori, kategori)

b) Barang (#kode_barang, nama_barang, #kelompok, satuan, harga)

c) Lokasi (#kode_pos, kelurahan, kecamatan)

d) Pelanggan (#kode_pelanggan, nama_pelanggan, alamat, kota, #kode_pos, telepon)

e) Penjualan (#no_faktur, #kode_barang, jumlah)

f) Pembayaran (#no_faktur, tanggal, total, diskon, #kode_pelanggan)

Tahap berikutnya yang harus dilakukan adalah menentukan **measure** dan **dimension** untuk semua informasi yang dibutuhkan manajemen. **Measure** adalah data numerik yang akan dicari jejak nilainya, sedangkan **dimension** adalah parameter atau sudut pandang terhadap **measure** sehingga dapat mendefinisikan suatu transaksi.

Sebagai contoh, untuk informasi “barang yang paling banyak terjual di lokasi tertentu sepanjang tahun”,

1. **Measure:** total penjualan
2. **Dimension:** barang, tahun (waktu/periode), lokasi

Sedangkan untuk informasi “barang yang paling banyak memberikan pendapatan sepanjang tahun”,

1. **Measure:** total pendapatan
2. **Dimension:** barang, tahun (waktu/periode)

Dimension mempunyai hirarki. Penentuan hirarki untuk **dimension** ini sepenuhnya tergantung kepada proses *drill down* dan *roll up* yang ingin dilakukan saat melakukan OLAP (*On Line Analytical Processing*) nanti.

Untuk contoh diatas, hirarki masing-masing **dimension** adalah:

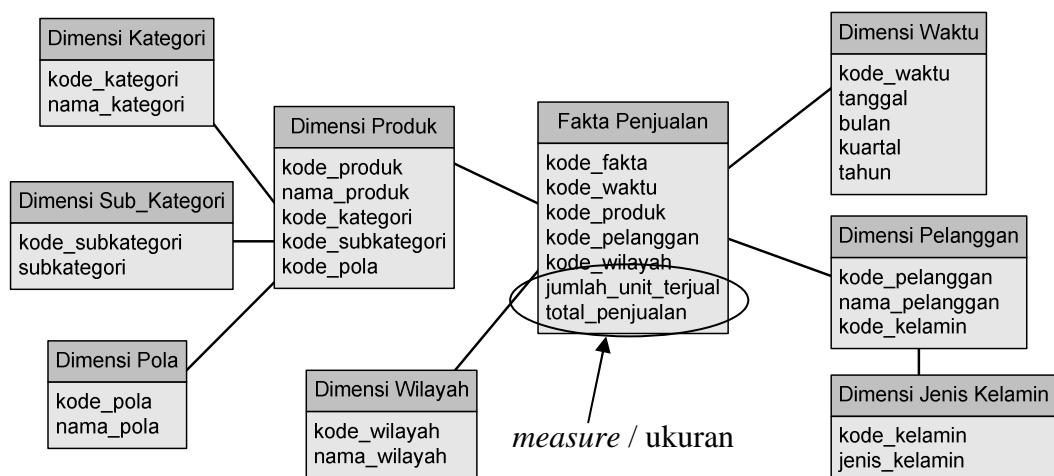
1. Barang: nama barang, sub-kategori, kategori
2. Periode: minggu, bulan, tahun
3. Lokasi: kelurahan, kecamatan, kota

Sedangkan *layout* dan isi informasinya dapat ditunjukkan oleh tabel berikut ini:

BARANG				PERIODE			LOKASI			TOTAL PENJUALAN	TOTAL PENERIMAAN
KODE	NAMA	SUB KATEGORI	KATEGORI	MINGGU	BULAN	TAHUN	KELURAHAN	KECAMATAN	KOTA		
SP-001	BOARDMARKER SNOWMAN	SPIDOL	ATK	33	8	2012	JEBRES	JEBRES	SURAKARTA	12	Rp 78,000
				38	9	2012	GAJAHAN	PASAR KLIWON	SURAKARTA	7	Rp 45,500
				43	10	2012	KAUMAN	PASAR KLIWON	SURAKARTA	7	Rp 45,500
				7	2	2012	GILINGAN	BANJARSARI	SURAKARTA	12	Rp 78,000
										38	Rp 247,000
PL-002	PULPEN PILOT	PULPEN	ATK	22	6	2012	DANUKUSUMAN	SERENGAN	SURAKARTA	3	Rp 9,000
				26	6	2012	KARANGASEM	LAWEYAN	SURAKARTA	6	Rp 18,000
				41	10	2012	KERTEN	LAWEYAN	SURAKARTA	8	Rp 24,000
										17	Rp 51,000
KH-005	A4 SINAR DUNIA	KERTAS	ATK	10	3	2012	MOJOSONGO	JEBRES	SURAKARTA	10	Rp 350,000
				27	7	2012	TIPES	SERENGAN	SURAKARTA	4	Rp 140,000
				46	11	2012	MANAHAN	BANJARSARI	SURAKARTA	9	Rp 315,000
										23	Rp 805,000
JUMLAH										78	Rp1,103,000

Perancangan model konseptual data warehouse adalah tahap berikutnya yang harus dilaksanakan setelah tahap penentuan *measure* dan *dimension*. Pada tahap ini dibuat suatu model yang dapat menggambarkan data atau tabel apa saja yang akan disimpan dalam data warehouse, berikut keterhubungan diantaranya.

Data atau tabel dalam data warehouse tersebut dapat dimodelkan dengan menggunakan alat bantu pemodelan seperti E-R diagram, *star schema*, *snowflake schema*, atau FCO-IM (*Fully Communication Oriented Information Modelling*). Tetapi pada umumnya alat bantu yang digunakan adalah *star schema* atau *snowflake schema*. *Star schema* digunakan untuk menggambarkan *fact table*, yaitu tabel yang merepresentasikan *measure*, sebagai “pusat data”. Tabel ini nantinya akan terkoneksi dengan tabel-tabel yang mendeskripsikan dimensi untuk *measure* tersebut (*dimension table*). Sebagai contoh, *snowflake schema* untuk data warehouse penjualan di sebuah perusahaan batik adalah:



Gambar 1.3. Snowflake Schema Penjualan

C. Alat dan Bahan

1. Komputer dengan sistem operasi Windows.
2. Program aplikasi DBDesigner.
3. Modul Praktikum Data Warehousing dan Data Mining.

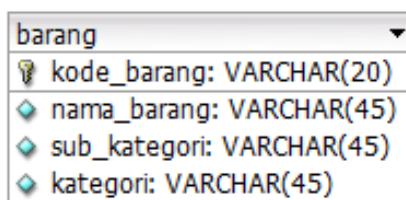
D. Langkah-langkah Praktikum

Menggambar *Star Schema* dengan menggunakan DB Designer :

1. Jalankan program aplikasi DB Designer untuk membuat desain *star schema*.
2. Klik button *new table*  kemudian klik pada area kerja sehingga akan menghasilkan tabel baru.
3. Double klik pada tabel baru untuk membuka tabel editor, ganti nama pada *table name* dengan nama **barang**, kemudian isikan atribut tabel dengan data sebagai berikut :

Column Name	Data Type
kode_barang	Varchar(20)
nama_barang	Varchar(45)
sub_kategori	Varchar(45)
kategori	Varchar(45)

4. Klik  pada *column name* kode_barang untuk mengatur kode_barang sebagai *primary key* sehingga berubah menjadi .
5. Klik  untuk menutup *table editor* sehingga tabel barang menjadi :

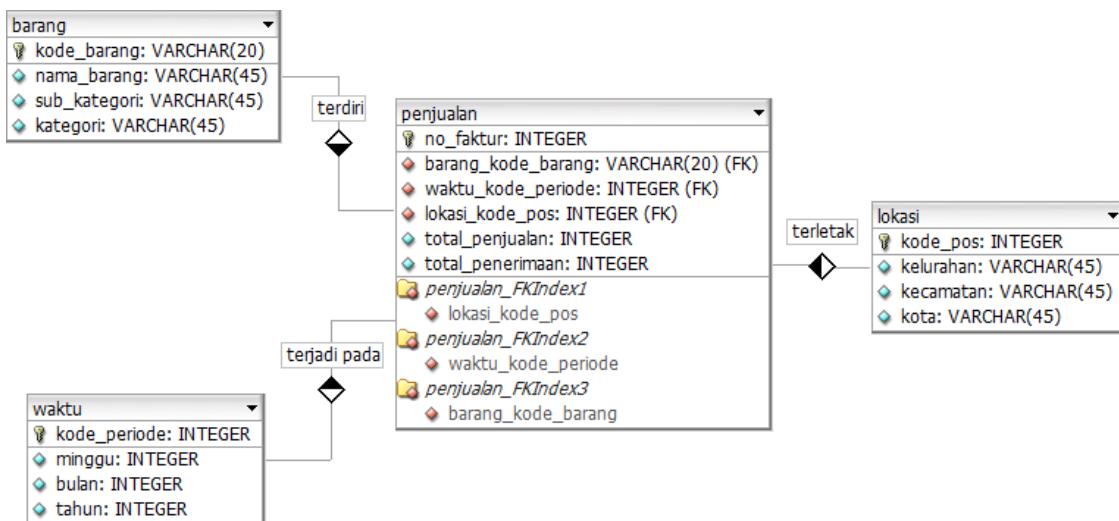


6. Ulangi kembali langkah 2 sampai 5 untuk membuat tabel **waktu**, **lokasi** dan **penjualan**.
7. Setelah semua tabel dibuat, hubungkan setiap tabel dengan tabel lain dengan button sebagai berikut :

Button	Fungsi Relationship
	1:n (<i>one to many</i>)
	1:1 (<i>one to one</i>)
	n:m (<i>many to many</i>)

Keterangan : klik salah satu button yang sesuai dengan kebutuhan kemudian klik tabel yang akan dihubungkan.

8. Ubah nama *relationship* dengan membuka *relationship editor*, sehingga setelah selesai hasil akhir manjadi seperti berikut :

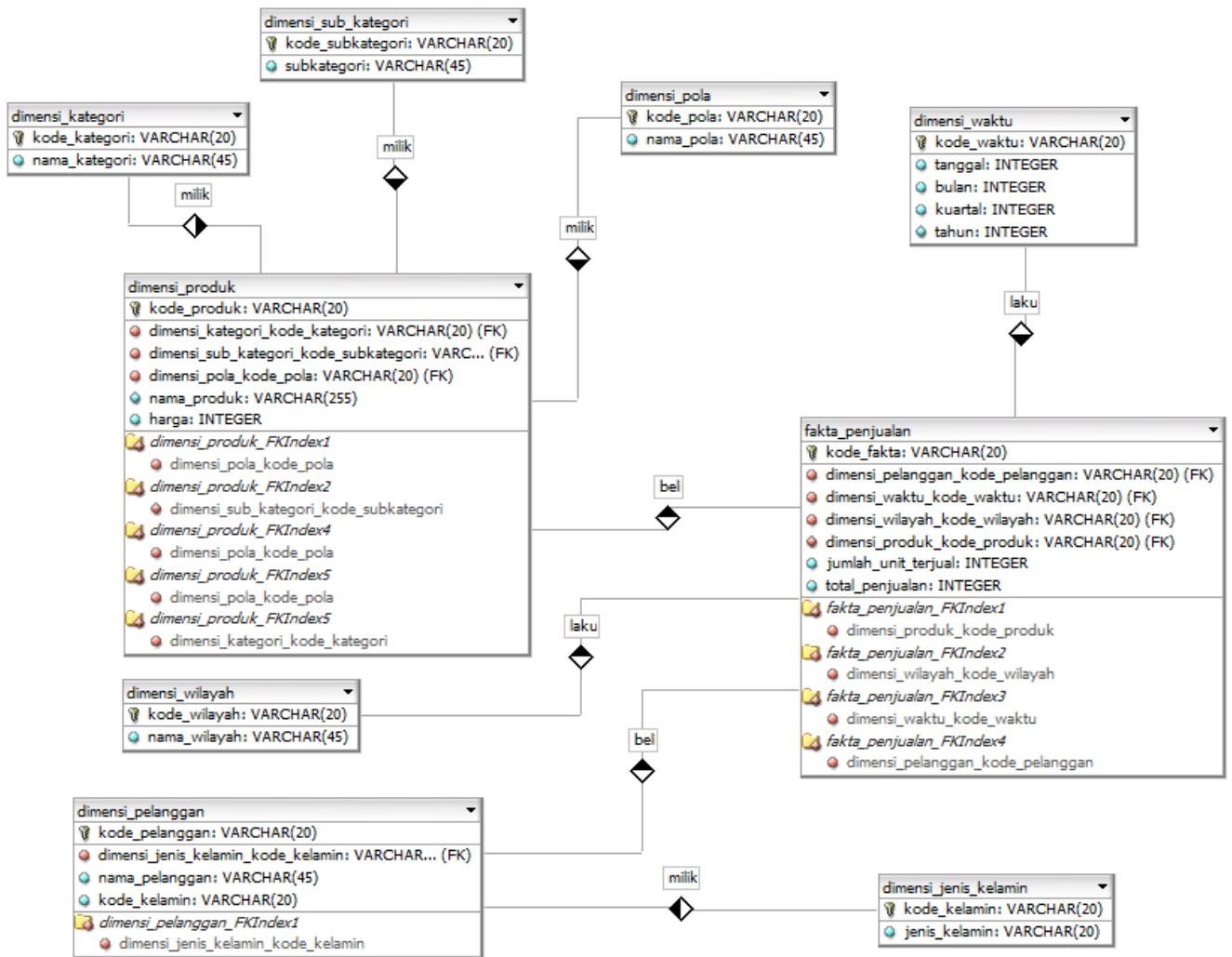


9. Simpan dengan nama file “**star schema penjualan.xml**” dalam folder “Praktikum Data Warehousing dan Data Mining”

Snowflake schema merupakan perbaikan dari *star schema*, sehingga cara penggambarannya pun mirip. Bedanya, satu atau beberapa hierarki yang ada pada *dimension table* dinormalisasi (dekomposisi) menjadi beberapa tabel yang lebih kecil.

E. Tugas

1. Rancanglah diagram *Snowflake schema* berdasarkan gambar di bawah dengan menggunakan DBDesigner seperti gambar berikut! Simpan dengan nama file "snowflake_penjualan.xml" ke dalam folder "Praktikum Data Warehouse"



MODUL 2

PROSES EXTRACT-TRANSFORM-LOAD

(EXTRAKSI DAN TRANSFORMASI DATA)

A. Tujuan

1. Mahasiswa mampu melakukan proses ekspor dan impor data yang merupakan bagian dalam proses ETL sebuah pengembangan Data Warehouse.

B. Landasan Teori

Ekstraksi (*Extraction*) adalah proses pengambilan data dari sumber data dimana proses pengambilan data ini tidak mengambil keseluruhan data yang ada di database operasional, melainkan hanya mengambil data-data matang saja. Tahapan ini adalah yang paling pertama dalam proses ETL. Setelah Ekstraksi, data ini akan ditransformasikan dan di-*load* ke dalam Data Warehouse.

Pendesainan dan Pembuatan proses Ekstraksi adalah satu kegiatan yang paling sering menyita waktu di dalam proses ETL dan dalam keseluruhan proses Data Warehouse. Data diekstrak tidak hanya sekali namun beberapa kali dalam suatu periode untuk mensuplai data ke dalam Data Warehouse dan menjaga agar *up-to-date*. Lebih jauh lagi, sistem sumber tidak dapat dimodifikasi atau bahkan kinerja dan ketersediaannya tidak dapat diatur untuk mengakomodasi kebutuhan proses ekstraksi Data Warehouse.

Ada dua bentuk Metode Ekstraksi logical:

1. Ekstraksi Statis (*Static Extraction*)

Data diekstrak secara lengkap dari sistem sumber. Ekstraksi ini melibatkan seluruh data yang sedang tersedia dalam sistem sumber. Data sumber disediakan dan tidak dibutuhkan logika informasi tambahan (seperti *timestamp*) yang dibutuhkan pada situs sumber. Sebuah contoh ekstraksi penuh adalah ekspor file dari sebuah tabel yang berbeda atau kueri remote SQL yang membaca sumber data lengkap. Proses ekstrak ini biasanya hanya dilakukan sekali di awal proses.

2. Ekstraksi Inkremental (*Incremental Extraction*)

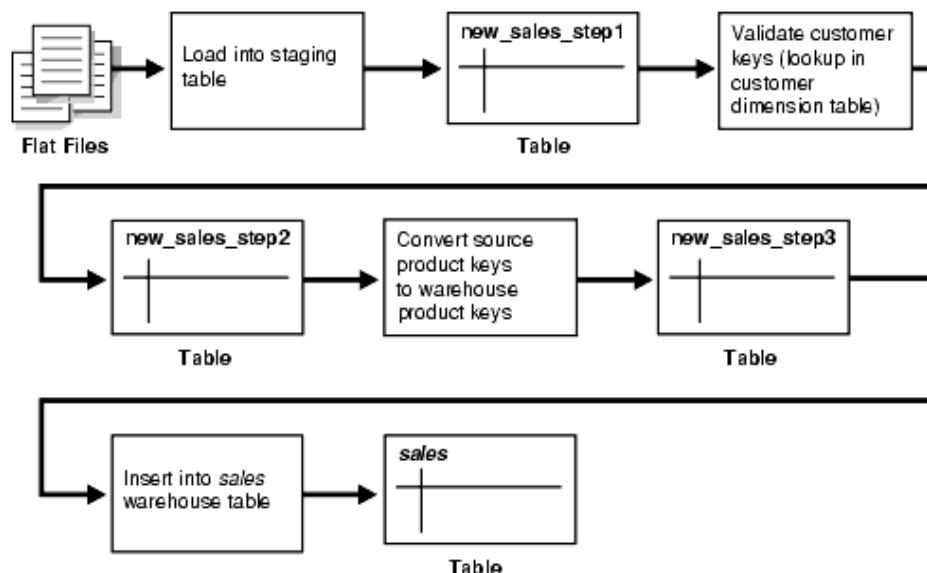
Pada poin waktu tertentu, hanya data yang memiliki histori data atau mengalami perubahan yang akan diekstrak. *Event* ini adalah proses ekstraksi yang dilakukan paling akhir atau sebagai contoh sebuah *event* bisnis yang kompleks seperti hari *booking* terakhir dari suatu periode fiskal. Informasi ini juga dapat disediakan oleh data sumber itu sendiri seperti sebuah kolom aplikasi, merefleksikan *time-stamp* yang paling akhir berubah atau sebuah tabel yang berubah dimana sebuah mekanisme tambahan yang sesuai menjaga *track* perubahan selain transaksi yang permulaan.

Transformasi data seringkali sangat kompleks, dalam hal waktu proses, bagian proses ekstraksi, transformasi dan loading yang paling membutuhkan banyak biaya. Proses ini boleh jadi merentang dari konversi data sederhana hingga teknik pengumpulan data kompleks yang ekstrim.

Dari perspektif arsitektural, Data dapat ditransformasikan dengan 2 cara :

1) *Multistage Data Transformation*

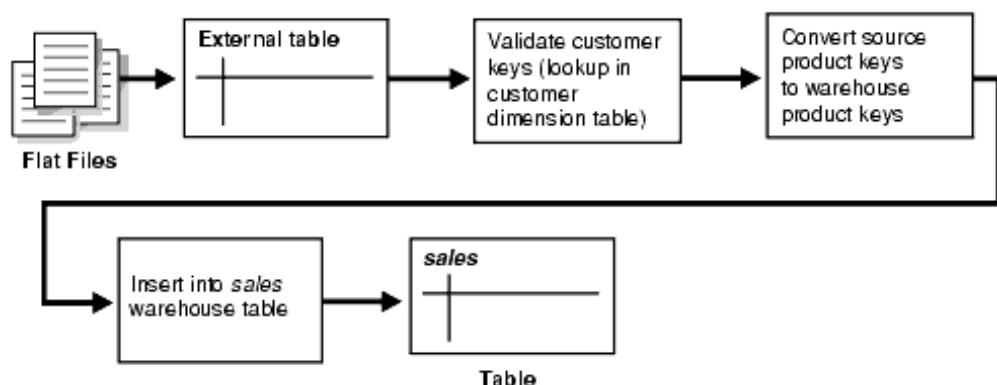
Logika transformasi data bagi kebanyakan Data Warehouse terdiri dari beberapa tahapan. Sebagai contoh, dalam transformasi *record* baru yang dimasukkan ke dalam sebuah tabel penjualan (*sales*), boleh jadi terdapat tahapan transformasi logik yang terpisah untuk memvalidasi masing-masing *key* dimensi. Gambaran secara grafis dari proses transformation logic adalah sbb :



Gambar 2.1. Transformasi Data *Multistage*

2) Pipelined Data Transformation

Arus proses ETL dapat diubah secara dramatis dan database menjadi sebuah bagian integral solusi ETL. Fungsionalitas barunya melukiskan beberapa pembentukan tahapan proses penting yang kuno ketika beberapa yang lainnya dapat dimodelkan kembali untuk menambah arus data dan transformasi data menjadi lebih dapat diukur. Kegiatannya bergeser dari transformasi serial hingga proses load (dengan kebanyakan kegiatan dilakukan diluar database) atau load kemudian proses transformasi untuk meningkatkan transformasi ketika proses loading.



Gambar 2.2. Transformasi Pipelined Data

C. Alat dan Bahan

1. Komputer dengan sistem operasi Windows.
2. Program aplikasi Ms. Office
3. Program aplikasi Apatar Tool.
4. Modul Praktikum Data Warehousing dan Data Mining.

D. Langkah-langkah Praktikum

D.1. Kegiatan 1: Ekspor data tertentu dari Ms. Excel ke Ms. Excel

1. Buka program aplikasi Ms. Excel
2. Buat tabel seperti pada Tabel 2.1 dan isikan data pada Sheet1.
3. Ubah nama Sheet1 menjadi "Data_Asal".

Tabel 2.1. Data Asal Penjualan Batik

WAKTU	NAMA BARANG	HARGA	JUMLAH	PEMBELI	DAERAH
2010-03-26	celana Standar Print Lasem	55000	17	Ibu hadi sukarni	Jawa Barat
2010-06-14	bahan Beludru Cap Mahkota	500000	1	Ibu tyas	Jawa Tengah
2010-11-21	hem Sutra Print rama	100000	5	Ibu tyas	Jawa Tengah
2011-01-05	kaos Katun Print Bola	60000	1	Bapak imron	Jawa Barat
2011-03-27	bahan Standar Cap Lasem	120000	8	Ibu siti arya	Jawa Barat
2011-04-09	hem Katun Print Kawung	70000	3	Ibu harini	Jawa Timur
2011-08-19	hem Standar Tulis Madura	550000	5	Ibu atik	Jawa Tengah
2011-10-13	sarimbit Standar Print Lukis	150000	1	Ibu Hatamah	Jawa Timur
2011-12-28	jarik Standar Print Sogan	225000	2	Bapak Ketut	Bali
2011-12-30	bolero Standar Cap Sidomukti	225000	1	Ibu Hatamah	Jawa Timur
2012-01-04	kaos batik Cap Lukis	30000	14	Ibu harini	Jawa Timur
2012-01-09	jam Standar Print Lukis	80000	44	Ibu siti arya	Jawa Barat
2012-02-14	celana Standar Cap Warna	55000	17	Ibu hadi sukarni	Jawa Barat
2012-04-05	bahan Standar Cap garis	135000	7	Ibu tyas	Jawa Tengah
2012-04-05	jarik Standar Tulis Sarimbit	40000	4	Ibu harini	Jawa Timur
2012-05-21	hem katun Print Kelengan	299000	3	Bapak Totok	Jawa Timur
2012-06-22	bahan Lawasan Tulis Tolet	130000	1	Ibu niken	Jawa Tengah
2012-09-18	batik Standar Cap Tumpal	150000	1	Bapak heru	Jawa Timur
2012-09-28	hem Standar Cap Tumpal	100000	1	Ibu aini kasmaji	Jawa Tengah
2012-12-15	rok batik Print Kombinasi	225000	1	Ibu siti arya	Jawa Barat

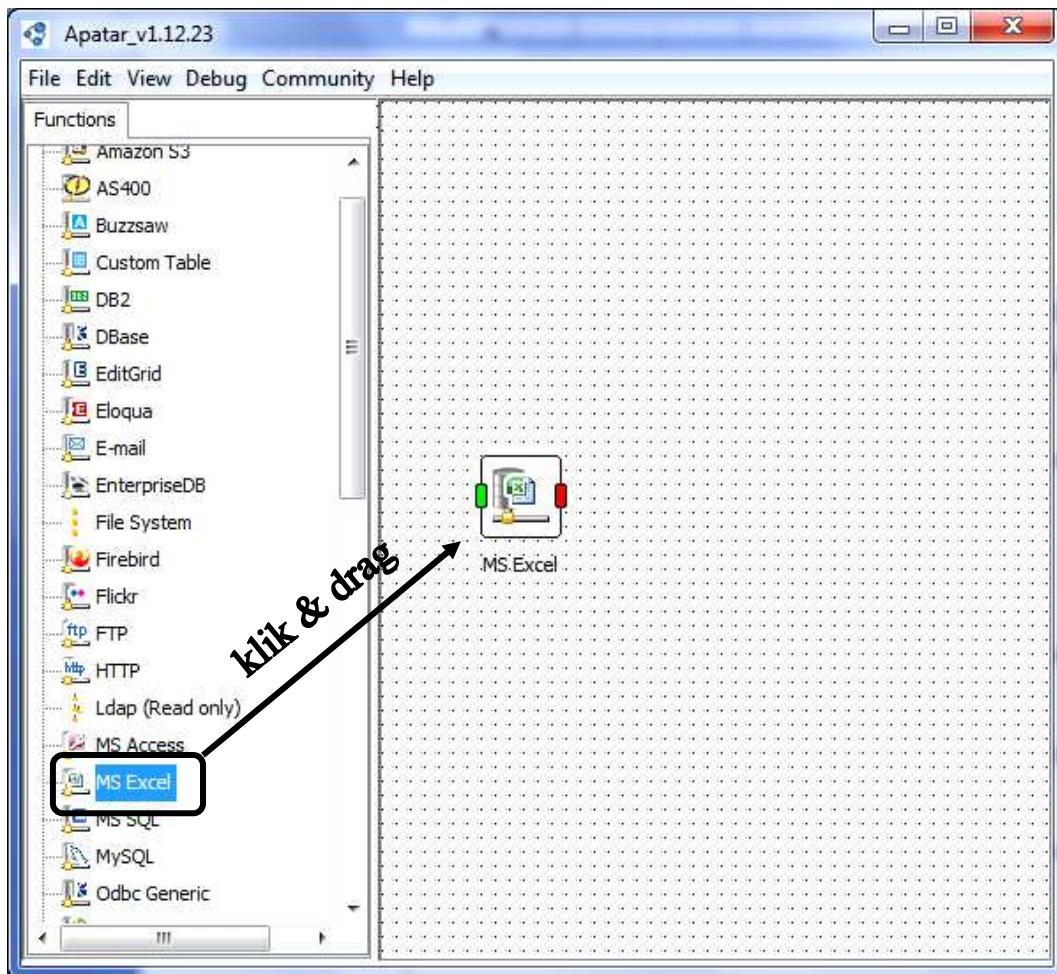
4. Buat tabel baru pada Sheet2 seperti pada Tabel 2.2. Tabel ini hanya terdapat nama kolom saja, tidak diisi dengan data apapun. Biarkan dalam keadaan kosong.

Tabel 2.2. Data Tujuan Penjualan Batik

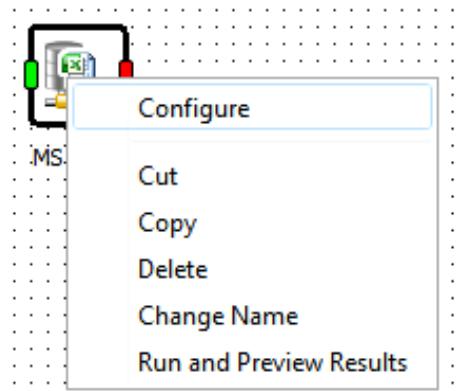
TANGGAL	PRODUK	HARGA	JUMLAH	PELANGGAN	WILAYAH

5. Ubah nama Sheet2 menjadi “Data_Tujuan”.
6. Simpan file dengan nama “**Data_penjualan.xls**” dalam folder “Praktikum Data Warehousing dan Data Mining”. (Format Excel 97-2003).
7. Tutup file “**Data_penjualan.xls**”.
8. Jalankan Apatar pada menu Start – Programs – Apatar – Application dengan hak akses Administrator (Klik kanan Application – Run as Administrator). Pilih Create New Datamap.

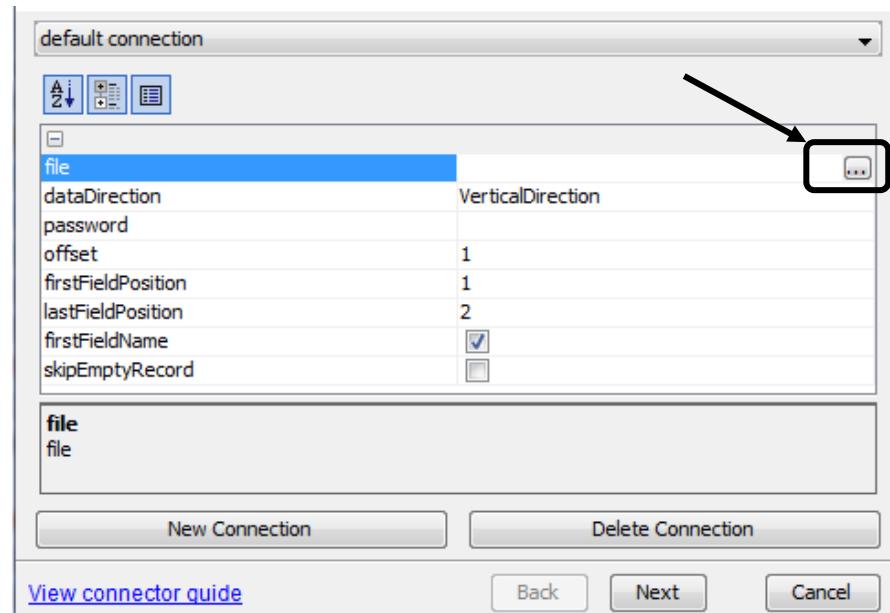
9. Pada jendela sebelah kiri terdapat tab Function yang berisi fungsi-fungsi yang digunakan untuk mendesain proses ETL sebuah data warehouse.
10. Function terdiri atas 3 macam, yaitu Connectors, Data Quality Service dan Operations.
11. Pada Connectors, klik MS EXCEL, drag ke Editor.



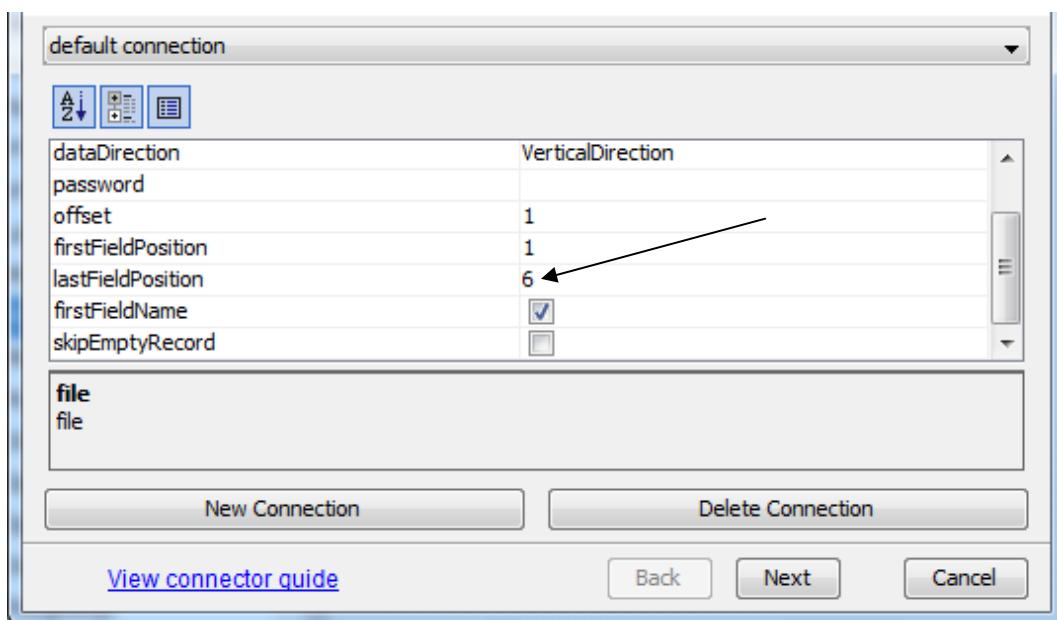
12. Klik kanan icon MS EXCEL pada editor, pilih Configure.



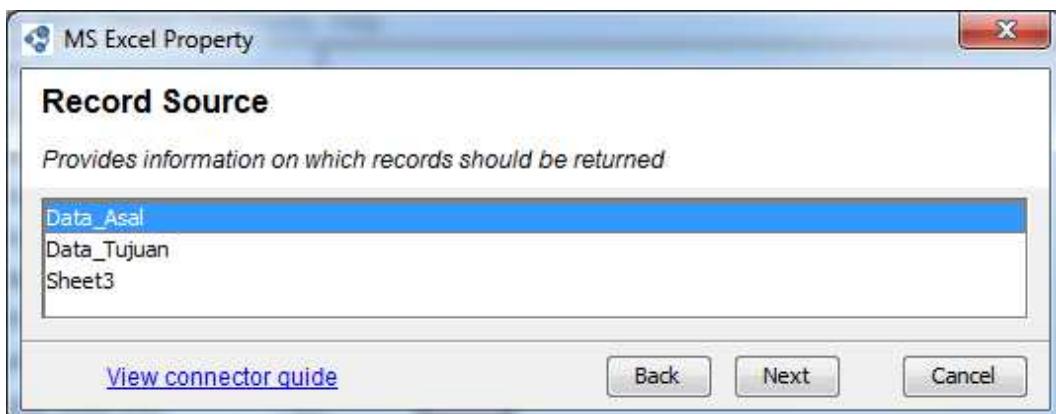
13. Pada jendela MS EXCEL Property, pilih menu file dan arahkan pada file excel yang anda buat pada langkah 2 dengan nama “**Data_penjualan.xls**” dalam folder “Praktikum Data Warehousing dan Data Mining”.



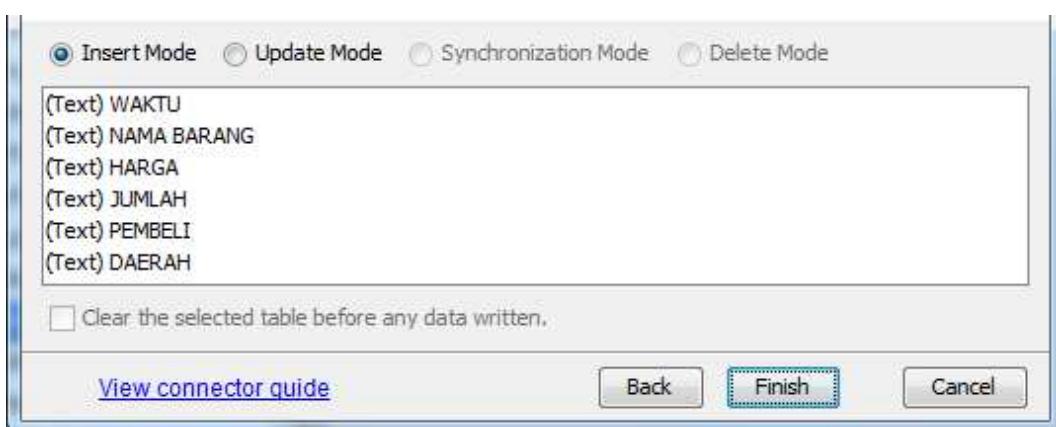
14. File “**Data_penjualan.xls**” menjadi sumber data yang akan diekstraksi. Masukkan nilai “6” pada “lastFieldPosition” yang menunjukkan posisi kolom terakhir yang digunakan pada tabel. Sedangkan firstFieldPosition menunjukkan posisi kolom pertama yang akan digunakan dalam tabel.



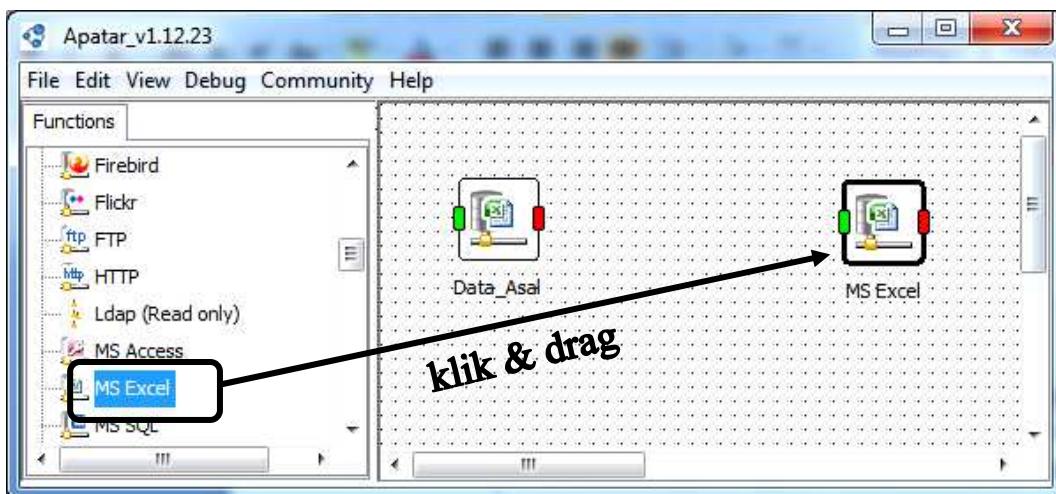
15. Klik Next, pilih Sheet “Data_Asal” sebagai tabel sumber yang berisi data penjualan.



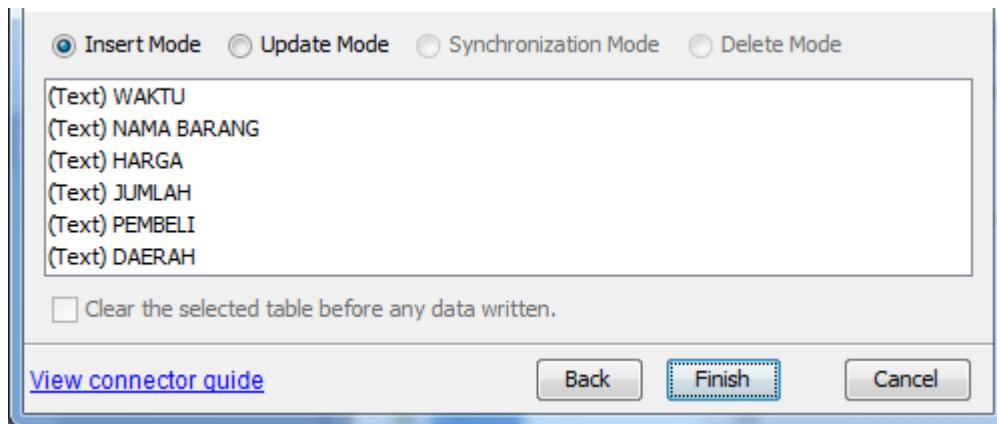
16. Klik Next sehingga muncul jendela Property yang berisi kolom-kolom pada tabel. Pastikan bahwa 6 kolom muncul dalam jendela ini. Kemudian klik Finish.



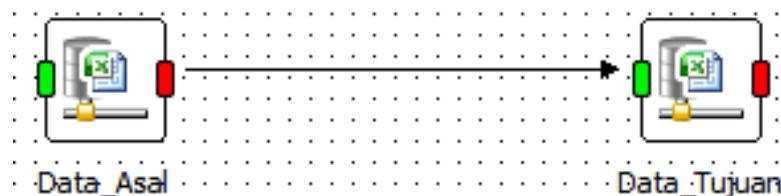
17. Kembali pada Apatar - Connectors, klik MS EXCEL, drag ke Editor.



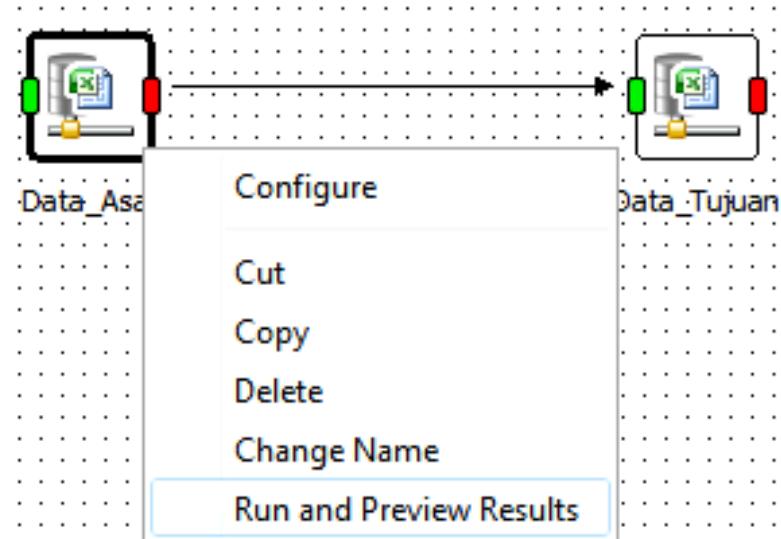
18. Lakukan seperti pada langkah ke-12 sampai langkah 16. Pastikan bahwa Sheet yang digunakan adalah “**Data_Tujuan**”. Sehingga muncul 6 kolom pada tabel tersebut. Kemudian klik Finish.



19. Hubungkan kedua Connector tersebut dengan cara klik dan tahan titik merah pada Data_Asal, arahkan pada titik hijau Data_Tujuan dan lepaskan.



20. Untuk memastikan bahwa data-data dari tabel sumber tersimpan pada tabel **Data_Asal**, klik kanan icon **Data_Asal** pilih menu “Run and Preview Results”.

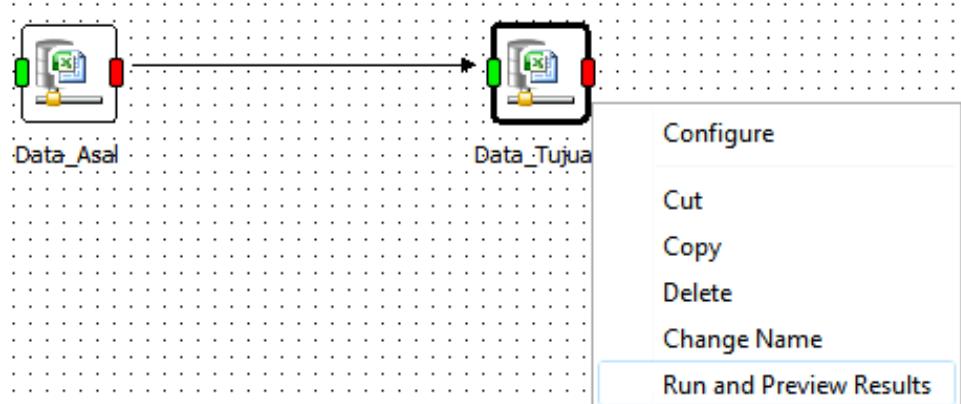


21. Jika proses berhasil maka data-data dari tabel sumber akan terlihat semua.

No.	WAKTU	NAMA BARANG	HARGA	JUMLAH	PEMBELI	DAERAH
1	2010-03-26	celana Standar Print Lasem	55000	17	Ibu hadi sukarni	Jawa Barat
2	2010-06-14	bahan Beludru Cap Mahkota	500000	1	Ibu tyas	Jawa Tengah
3	2010-11-21	hem Sutra Print rama	100000	5	Ibu tyas	Jawa Tengah
4	2011-01-05	kaos Katun Print Bola	60000	1	Bapak imron	Jawa Barat
5	2011-03-27	bahan Standar Cap Lasem	120000	8	Ibu siti aryा	Jawa Barat
6	2011-04-09	hem Katun Print Kawung	70000	3	Ibu harini	Jawa Timur
7	2011-08-19	hem Standar Tulis Madura	550000	5	Ibu atik	Jawa Tengah
8	2011-10-13	sarimbit Standar Print Lukis	150000	1	Ibu Hatamah	Jawa Timur
9	2011-12-28	jarik Standar Print Sogan	225000	2	Bapak Ketut	Bali
10	2011-12-30	bolero Standar Cap Sidomukti	225000	1	Ibu Hatamah	Jawa Timur
11	2012-01-04	kaos batik Cap Lukis	30000	14	Ibu harini	Jawa Timur
12	2012-01-09	jam Standar Print Lukis	80000	44	Ibu siti aryা	Jawa Barat
13	2012-02-14	celana Standar Cap Warna	55000	17	Ibu hadi sukarni	Jawa Barat
14	2012-04-05	bahan Standar Cap garis	135000	7	Ibu tyas	Jawa Tengah
15	2012-04-05	jarik Standar Tulis Sarimbit	40000	4	Ibu harini	Jawa Timur
16	2012-05-21	hem katun Print Kelenggan	299000	3	Bapak Totok	Jawa Timur
17	2012-06-22	bahan Lawasan Tulis Tolet	130000	1	Ibu niken	Jawa Tengah
18	2012-09-18	batik Standar Cap Tumpal	150000	1	Bapak heru	Jawa Timur
19	2012-09-28	hem Standar Cap Tumpal	100000	1	Ibu aini kasmaji	Jawa Tengah
20	2012-12-15	rok batik Print Kombinasi	225000	1	Ibu siti aryা	Jawa Barat

22. Tutup kembali tabel tersebut dan proses.

23. Lakukan seperti langkah ke-20 untuk melakukan proses ekspor dari tabel sumber **Data_Asal** menuju tabel baru **Data_Tujuan**.



24. Sehingga muncul tabel **Data_Tujuan** yang telah diisi data-data dari tabel sumber.

No.	TANGGAL	PRODUK	HARGA	JUMLAH	PELANGGAN	WILAYAH
1	2010-03-26	celana Standar Print Lasem	55000	17	Ibu hadi sukarni	Jawa Barat
2	2010-06-14	bahan Beludru Cap Mahkota	500000	1	Ibu tyas	Jawa Tengah
3	2010-11-21	hem Sutra Print rama	100000	5	Ibu tyas	Jawa Tengah
4	2011-01-05	kaos Katun Print Bola	60000	1	Bapak imron	Jawa Barat
5	2011-03-27	bahan Standar Cap Lasem	120000	8	Ibu siti arya	Jawa Barat
6	2011-04-09	hem Katun Print Kawung	70000	3	Ibu harini	Jawa Timur
7	2011-08-19	hem Standar Tulis Madura	550000	5	Ibu atik	Jawa Tengah
8	2011-10-13	sarimbit Standar Print Lukis	150000	1	Ibu Hatamah	Jawa Timur
9	2011-12-28	jarik Standar Print Sogan	225000	2	Bapak Ketut	Bali
10	2011-12-30	bolero Standar Cap Sidomukti	225000	1	Ibu Hatamah	Jawa Timur
11	2012-01-04	kaos batik Cap Lukis	30000	14	Ibu harini	Jawa Timur
12	2012-01-09	jam Standar Print Lukis	80000	44	Ibu siti arya	Jawa Barat
13	2012-02-14	celana Standar Cap Warna	55000	17	Ibu hadi sukarni	Jawa Barat
14	2012-04-05	bahan Standar Cap garis	135000	7	Ibu tyas	Jawa Tengah
15	2012-04-05	jarik Standar Tulis Sarimbit	40000	4	Ibu harini	Jawa Timur
16	2012-05-21	hem katun Print Kelenggan	299000	3	Bapak Totok	Jawa Timur
17	2012-06-22	bahan Lawasan Tulis Tolet	130000	1	Ibu niken	Jawa Tengah
18	2012-09-18	batik Standar Cap Tumpal	150000	1	Bapak heru	Jawa Timur
19	2012-09-28	hem Standar Cap Tumpal	100000	1	Ibu aini kasmaji	Jawa Tengah
20	2012-12-15	rok batik Print Kombinasi	225000	1	Ibu siti arya	Jawa Barat

25. Untuk memastikan bahwa proses ekspor telah berhasil dilakukan dari tabel sumber, bukalah file excel “**Data_penjualan.xls**”. Jika semua data dari sheet Data_Asal telah masuk dalam sheet Data_Tujuan, maka proses ekspor telah BERHASIL dilakukan.

A	B		C	D	E	F
1	TANGGAL	PRODUK	HARGA	JUMLAH	PELANGGAN	WILAYAH
2	2010-03-26	celana Standar Print Lasem	55000	17	Ibu hadi sukarni	Jawa Barat
3	2010-06-14	bahan Beludru Cap Mahkota	500000	1	Ibu tyas	Jawa Tengah
4	2010-11-21	hem Sutra Print rama	100000	5	Ibu tyas	Jawa Tengah
5	2011-01-05	kaos Katun Print Bola	60000	1	Bapak imron	Jawa Barat
6	2011-03-27	bahan Standar Cap Lasem	120000	8	Ibu siti arya	Jawa Barat
7	2011-04-09	hem Katun Print Kawung	70000	3	Ibu harini	Jawa Timur
8	2011-08-19	hem Standar Tulis Madura	550000	5	Ibu atik	Jawa Tengah
9	2011-10-13	sarimbit Standar Print Lukis	150000	1	Ibu Hatamah	Jawa Timur
10	2011-12-28	jarik Standar Print Sogan	225000	2	Bapak Ketut	Bali
11	2011-12-30	bolero Standar Cap Sidomukti	225000	1	Ibu Hatamah	Jawa Timur
12	2012-01-04	kaos batik Cap Lukis	30000	14	Ibu harini	Jawa Timur
13	2012-01-09	jam Standar Print Lukis	80000	44	Ibu siti arya	Jawa Barat
14	2012-02-14	celana Standar Cap Warna	55000	17	Ibu hadi sukarni	Jawa Barat
15	2012-04-05	bahan Standar Cap garis	135000	7	Ibu tyas	Jawa Tengah
16	2012-04-05	jarik Standar Tulis Sarimbit	40000	4	Ibu harini	Jawa Timur
17	2012-05-21	hem katun Print Kelenggan	299000	3	Bapak Totok	Jawa Timur
18	2012-06-22	bahan Lawasan Tulis Tolet	130000	1	Ibu niken	Jawa Tengah
19	2012-09-18	batik Standar Cap Tumpal	150000	1	Bapak heru	Jawa Timur
20	2012-09-28	hem Standar Cap Tumpal	100000	1	Ibu aini kasmaji	Jawa Tengah
21	2012-12-15	rok batik Print Kombinasi	225000	1	Ibu siti arya	Jawa Barat

D.2. Kegiatan 2 : Transformasi Data

Pada kegiatan ini, akan dilakukan proses transformasi data untuk mengubah data transaksional menjadi data dengan format yang diperlukan dalam data warehouse. Field-field yang terdapat dalam tabel sumber akan dipisah menjadi beberapa dimensi dan sub dimensi sebagaimana rancangan *snowflake schema* pada Modul 1 gambar 1.3.

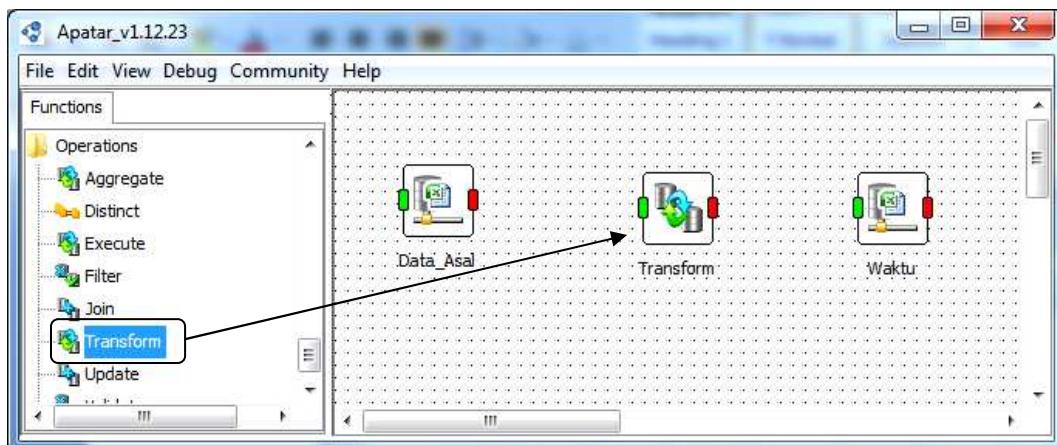
Tahap praktikum sebagai berikut:

1) Dimensi Waktu

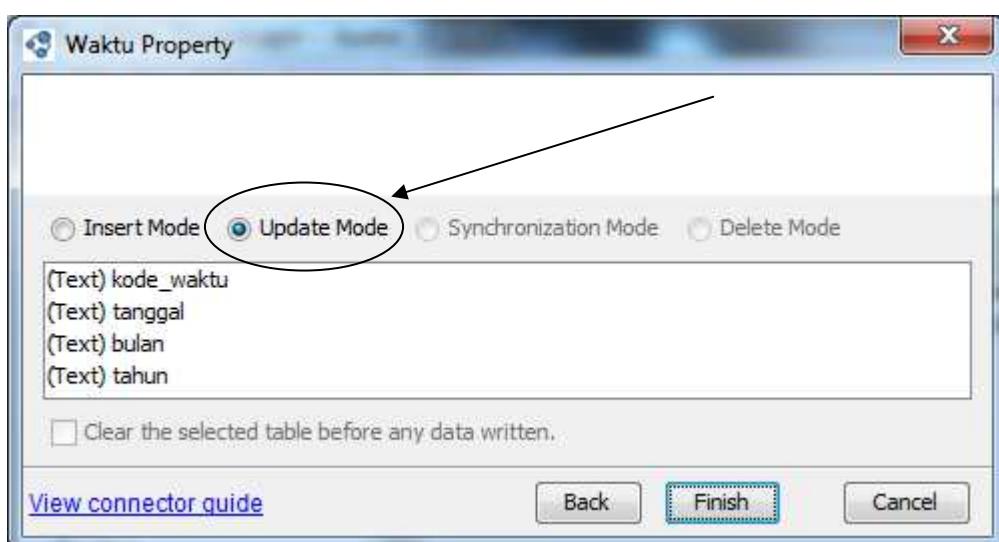
1. Tabel yang akan dibuat pertama kali adalah **Dimensi Waktu**. Buka kembali file excel “**Data_penjualan.xls**”.
2. Buka Sheet3. Ubah nama Sheet3 menjadi “**Waktu**”.
3. Buat kolom **kode_waktu**, **tanggal**, **bulan** dan **tahun**. Simpan file excel, dan tutup kembali.

	A	B	C	D	E	F	G	H
1	kode waktu	tanggal	bulan	tahun				
2								
3								
4								
5								
6								

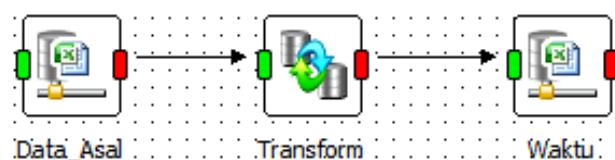
4. Buka aplikasi Apatar Tools dengan hak akses sebagai Administrator. Simpan file apatar dengan nama “**Transform_Dimensi.aptr**”.
5. Tambahkan 2 buah connector MS. Excel dan sebuah operator “Transform” pada editor apatar.
6. Dengan menggunakan file “**Data_penjualan.xls**”, atur konfigurasi MS Excel yang pertama dengan mengambil Sheet “**Data_Asal**” yang terdiri dari 6 kolom, dan MS Excel yang kedua mengambil Sheet “**Waktu**” yang terdiri dari 4 kolom yang baru saja dibuat pada langkah 3.



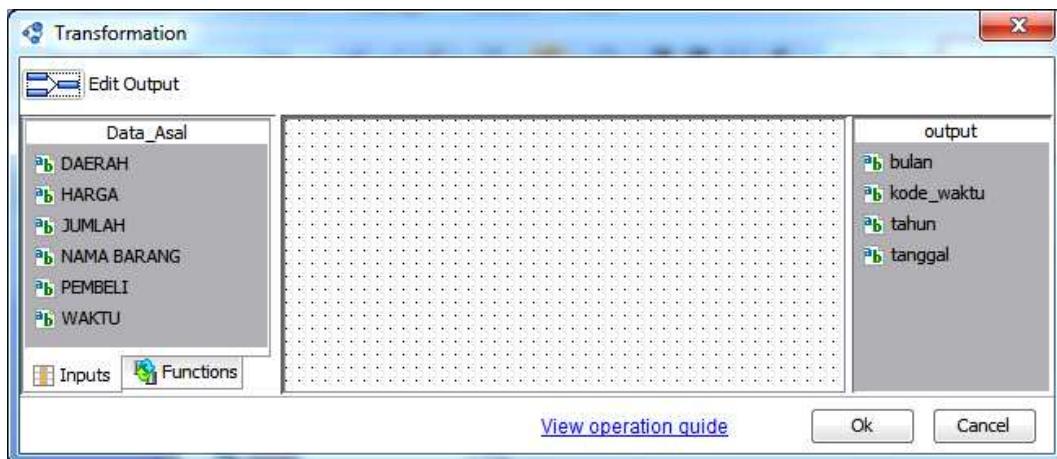
7. Saat konfigurasi Waktu, pada jendela “Waktu property” pilih **Update Mode** yang berarti hanya perubahan-perubahan saja yang akan diekspor ke tabel Waktu. Jika tidak terjadi perubahan pada data sumber, maka tidak ada data yang diekspor ke tabel tujuan. Klik Finish.



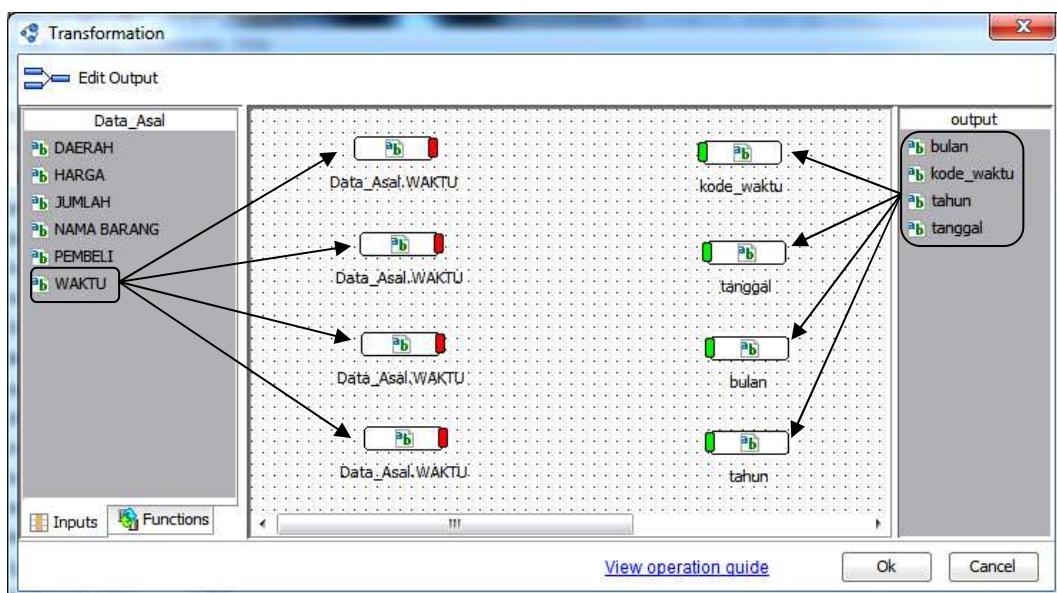
8. Hubungkan ketiga connector tersebut dengan urutan Data_Asal → Transform → Waktu.



9. Atur konfigurasi “Transform” dengan klik kanan pilih Configure sehingga muncul jendela properti “Transformation”.

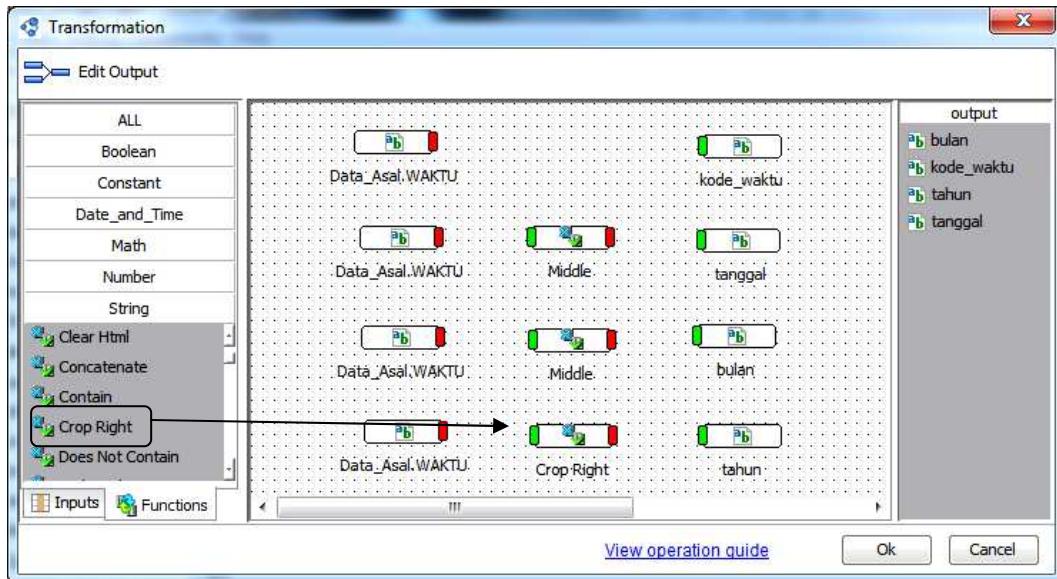


10. Drag nama-nama kolom yang dibutuhkan dalam proses transformasi dari tabel sumber “Data_Asal” dan dari tabel tujuan “Waktu” ke dalam editor. Dikarenakan kolom yang ditransformasikan dari tabel sumber hanya kolom WAKTU, maka klik dan drag field WAKTU dari Data_Asal sebanyak 4 kali sesuai dengan jumlah kolom dalam tabel tujuan WAKTU. Sedangkan dari tabel tujuan, klik dan drag semua field yang ada ke dalam editor.

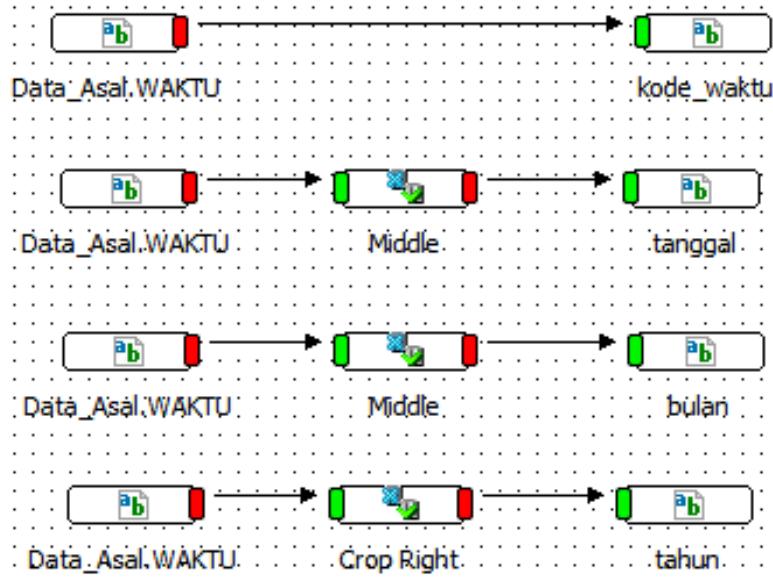


11. Tambahkan dua buah fungsi “Middle” ke dalam editor untuk mengubah data waktu menjadi data tanggal dan data bulan. Fungsi “Middle” digunakan untuk mengambil sejumlah karakter string yang terletak di tengah teks.

12. Tambahkan pula sebuah fungsi “Crop Right” ke dalam editor untuk mengubah data waktu menjadi data tahun. Fungsi “Crop Right” digunakan untuk mengambil sejumlah karakter string dari arah kiri ke arah kanan.

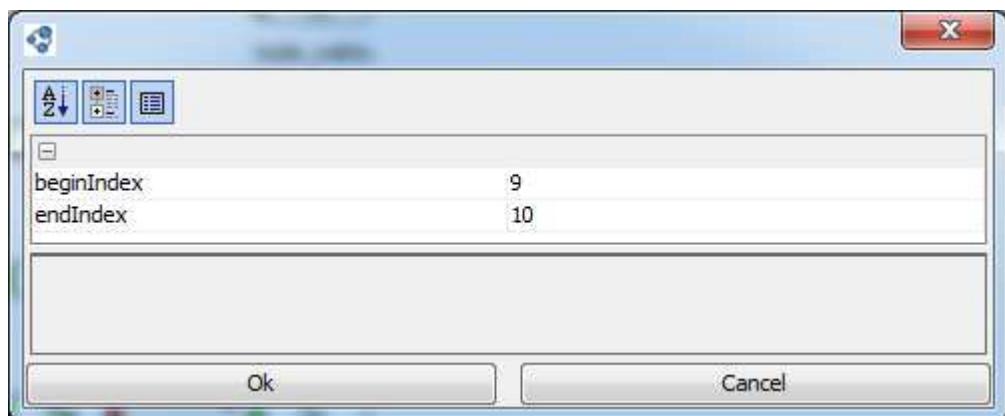


13. Hubungkan semua field dan fungsi yang ada dalam editor seperti gambar berikut.

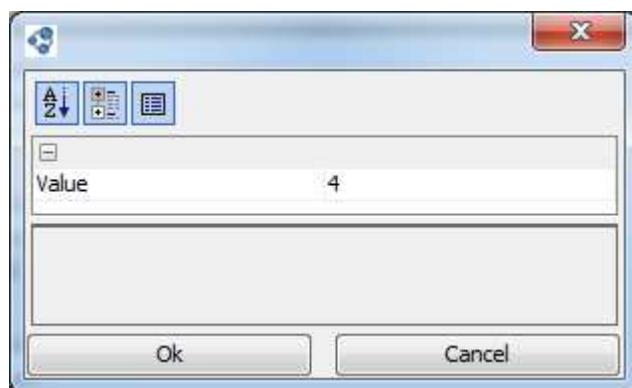


14. Konfigurasi fungsi “Middle” untuk mengubah format tanggal. Klik kanan fungsi “Middle”, pilih Configure. Isikan angka 9 pada baris “beginIndex” dan angka 10 pada baris “endIndex”. “beginIndex” berfungsi untuk menyatakan batas awal karakter akan diambil, sedangkan “endIndex” menyatakan batas karakter akhir diambil. Sehingga dalam kasus ini

menunjukkan bahwa karakter tanggal diambil mulai karakter ke-9 hingga ke-10 sebanyak 2 karakter. Contoh data “2012-04-25”. Pada data tersebut tanggal terletak pada 2 karakter terakhir yaitu pada indeks ke-9 hingga 10 untuk mengambil angka 25 sebagai tanggal.



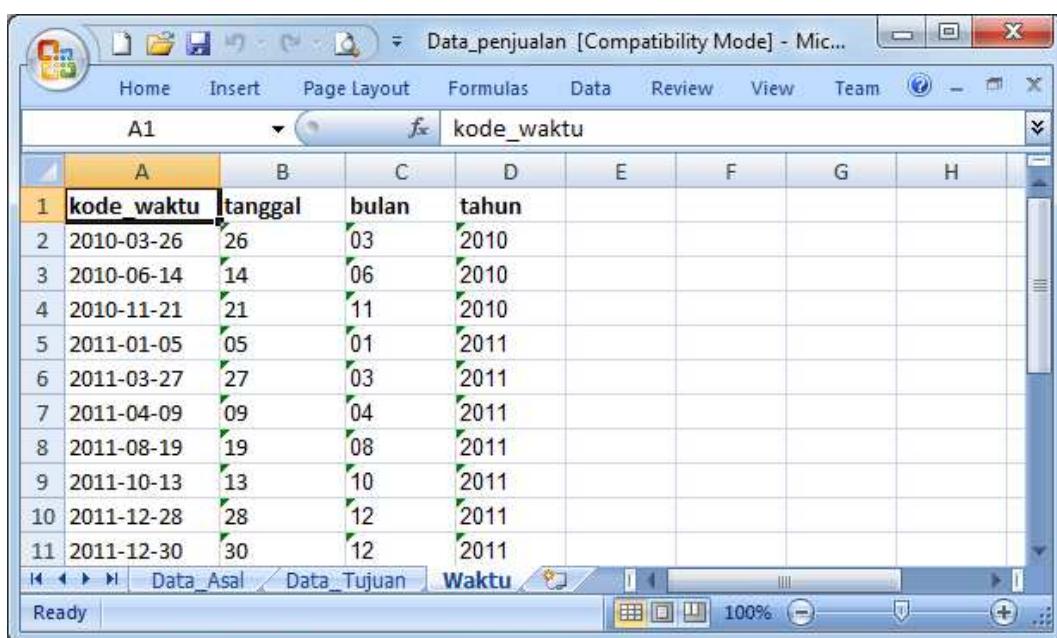
15. Klik OK.
16. Lakukan hal yang sama pada fungsi “Middle” untuk mengambil data bulan yang terletak pada indeks ke 6 hingga indeks ke 7. Klik OK.
17. Fungsi “Crop Right” digunakan untuk mengambil sejumlah karakter string dari arah kiri ke arah kanan. Klik kanan pada fungsi “Crop Right”, pilih Configure. Pada kotak “value”, isi dengan angka 4 yang menunjukkan 4 digit angka dari teks diambil dari arah kiri ke kanan. Klik OK.



18. Klik OK pada jendela properti “Transformation”. Proses transformasi siap dilakukan.
19. Sebelum memindahkan data asal ke tabel tujuan, perlu memastikan dahulu bahwa konfigurasi transformasi sudah benar. Klik kanan operator “Transform” pada editor Apatar, pilih Run and Preview Results. Pastikan kolom tanggal, bulan dan tahun terisi dengan benar.

No.	kode_waktu	tanggal	bulan	tahun
1	2010-03-26	26	03	2010
2	2010-06-14	14	06	2010
3	2010-11-21	21	11	2010
4	2011-01-05	05	01	2011
5	2011-03-27	27	03	2011
6	2011-04-09	09	04	2011
7	2011-08-19	19	08	2011
8	2011-10-13	13	10	2011
9	2011-12-28	28	12	2011
10	2011-12-30	30	12	2011
11	2012-01-04	04	01	2012
12	2012-01-09	09	01	2012
13	2012-02-14	14	02	2012
14	2012-04-05	05	04	2012
15	2012-04-05	05	04	2012
16	2012-05-21	21	05	2012
17	2012-06-22	22	06	2012
18	2012-09-18	18	09	2012
19	2012-09-28	28	09	2012
20	2012-12-15	15	12	2012

20. Jika sudah benar, tutup tabel dan proses untuk kembali ke editor Apatar. Lakukan proses ETL dengan klik kanan pada connector “Waktu”, pilih Run and Preview Results. Jika tabel yang ditunjukkan sama dengan pada langkah ke-18, maka proses ETL berhasil dilakukan.
21. Untuk memastikan keberhasilan proses ETL, buka kembali file “Data_penjualan.xls” pada Sheet “Waktu”.



The screenshot shows an Excel spreadsheet titled "Data_penjualan [Compatibility Mode] - Mic...". The "Waktu" sheet is active, displaying the following data:

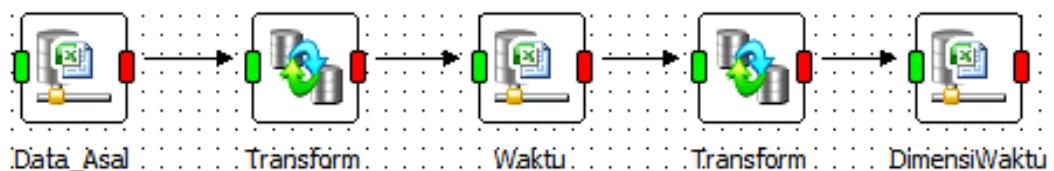
	A	B	C	D	E	F	G	H
1	kode_waktu	tanggal	bulan	tahun				
2	2010-03-26	26	03	2010				
3	2010-06-14	14	06	2010				
4	2010-11-21	21	11	2010				
5	2011-01-05	05	01	2011				
6	2011-03-27	27	03	2011				
7	2011-04-09	09	04	2011				
8	2011-08-19	19	08	2011				
9	2011-10-13	13	10	2011				
10	2011-12-28	28	12	2011				
11	2011-12-30	30	12	2011				

The status bar at the bottom shows "Ready" and "100%".

22. Tabel Waktu sudah berhasil dibuat dan diisi data, namun pada tabel “Dimensi_Waktu” berdasarkan *snowflake schema* pada Modul 1 terdapat sebuah kolom dengan nama “Kuartal”. Kolom ini digunakan untuk membuat kelompok bulan berdasarkan 3 bulanan. Bulan 1 sampai 3 dikelompokkan pada kuartal 1, bulan 4 sampai 6 dikelompokkan pada kuartal 2, dan seterusnya hingga kuartal 4.
23. Buka file “Data_penjualan.xls”, buat Sheet baru. Ubah nama **Sheet4** menjadi “**DimensiWaktu**”. Buat tabel kosong dengan 5 kolom yang terdiri dari **kode_waktu**, **tanggal**, **bulan**, **kuartal** dan **tahun**. Simpan dan tutup kembali file tersebut.

	A1	f(x)	kode_waktu	F	G
1	kode_waktu	tanggal	bulan	kuartal	tahun
2					
3					
4					

24. Kembali pada Appter Tools. Tambahkan sebuah operator **Transform**, dan sebuah connector **MS. Excel**.
25. Atur konfigurasi MS. Excel yang baru ditambahkan pada **DimensiWaktu** yang dibuat pada langkah 23.
26. Hubungkan connector **Waktu** dengan operator **Transform**, dan hubungkan operator **Transform** dengan connector **DimensiWaktu**.



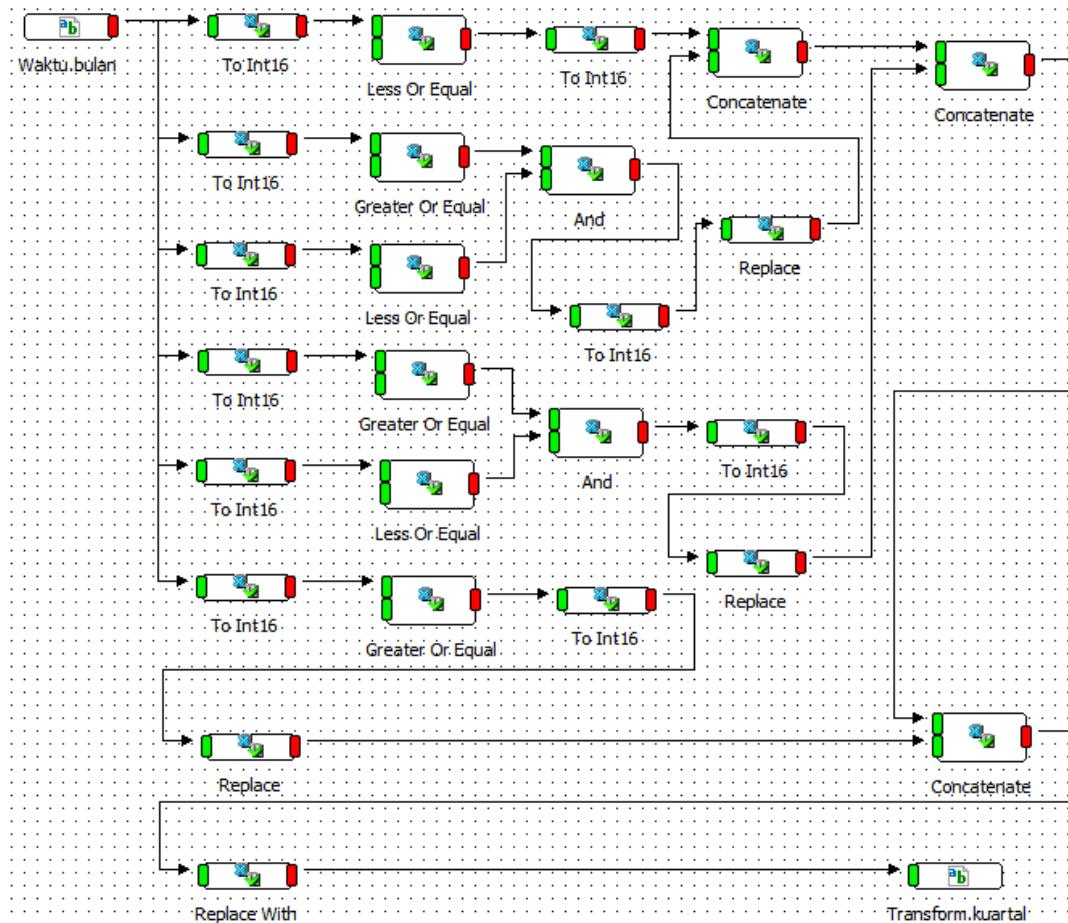
27. Klik kanan operator Transform yang baru ditambahkan, pilih Configure.
28. **Kode_waktu**, **tanggal**, **bulan** dan **tahun** dari tabel Waktu menuju tabel DimensiWaktu tidak mengalami perubahan, sehingga hubungkan secara langsung kolom-kolom tersebut.



29. Sedangkan untuk mengisi kolom **kuartal** dalam tabel **DimensiWaktu** memerlukan logika perbandingan berdasarkan bulan dalam tabel **Waktu** sebagai berikut:
 - a) Jika bulan ≤ 3 , maka Kuartal 1
 - b) Jika $4 \leq$ bulan ≤ 6 , maka Kuartal 2
 - c) Jika $7 \leq$ bulan ≤ 9 , maka Kuartal 3
 - d) Jika bulan ≥ 10 , maka Kuartal 4
30. Klik dan drag field **bulan** dari tabel sumber dan field **kuartal** dari tabel tujuan pada editor.
- 31.Tambahkan 10 fungsi “**To Int16**”, 3 fungsi “**Less or Equal**” dan 3 fungsi “**Greater or Equal**” yang diambil dari Function Number. Fungsi “To Int16” digunakan untuk mengubah tipe data lain menjadi tipe data integer 16 bit. Fungsi “Less or Equal” digunakan untuk membandingkan bilangan yang kurang dari atau sama dengan. Fungsi “Greater or Equal” digunakan untuk membandingkan suatu bilangan lebih besar atau sama dengan. Output dari kedua fungsi ini adalah Boolean yaitu True/False atau 1/0.
32. Tambahkan 2 fungsi “**And**” yang diambil dari Function Boolean.
33. Tambahkan 3 fungsi “**Replace**”, 3 fungsi “**Concatenate**”, dan 1 fungsi “**Replace With**” yang diambil dari Function String. Fungsi “Replace” digunakan untuk mengganti suatu data dengan data yang lain. Sedangkan “Replace With” pada dasarnya memiliki fungsi yang sama dengan

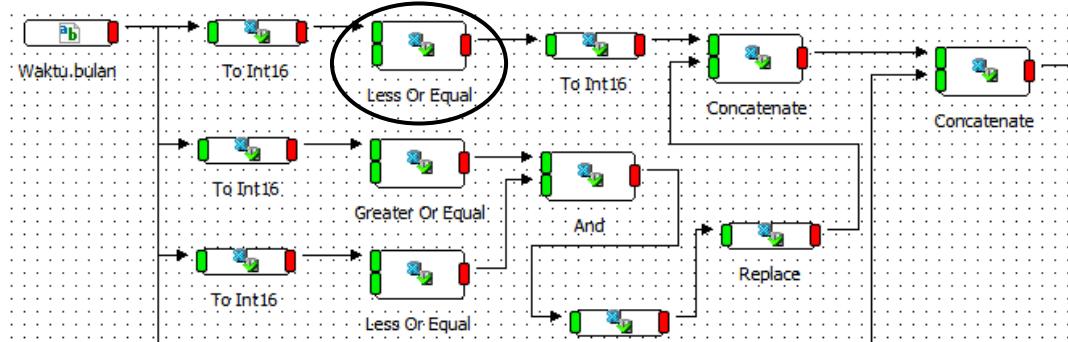
“Replace”, bedanya jika “Replace With” dapat digunakan untuk lebih dari satu kondisi. Fungsi “Concatenate” digunakan untuk menggabungkan 2 data menjadi satu.

34. Hubungkan field bulan dan field kuartal dengan fungsi-fungsi yang ditambahkan pada langkah 31-34 seperti gambar berikut.

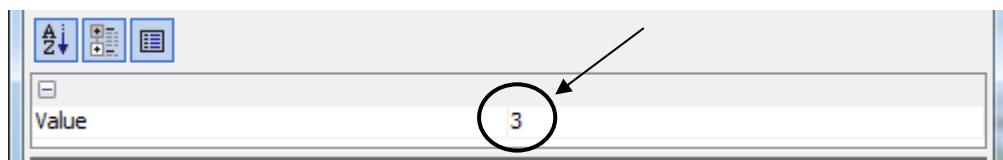


35. Atur konfigurasi semua fungsi “Less or Equal”, “Greater or Equal”, “Replace” and “Replace With” dengan klik kanan, pilih Configure.

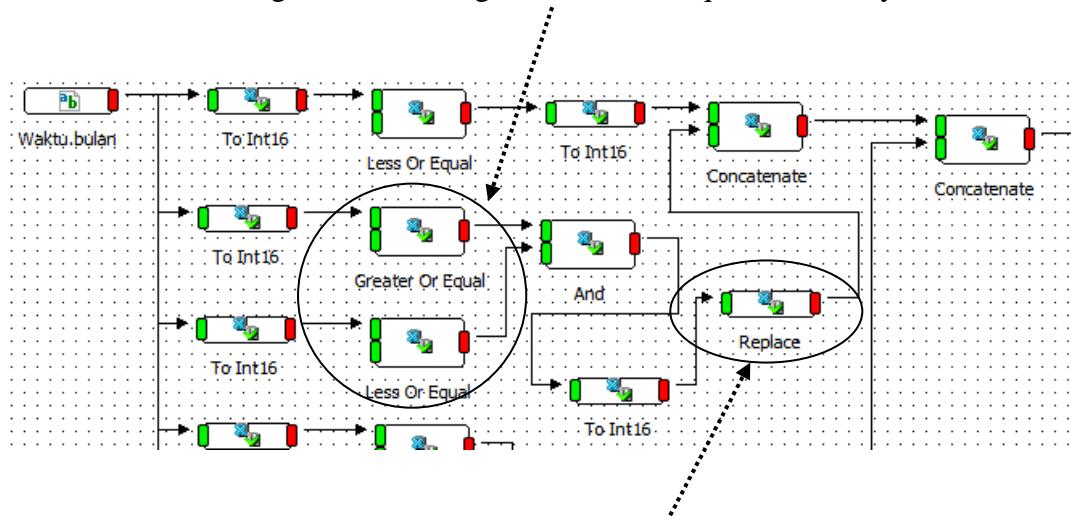
36. Konfigurasi “Less or Equal” yang pertama.



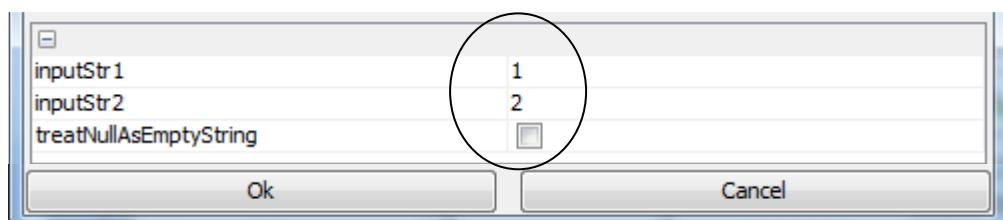
37. Isikan kolom “Value” dengan bilangan 3 yang menunjukkan bahwa jika bulan lebih kecil atau sama dengan 3. Klik OK.



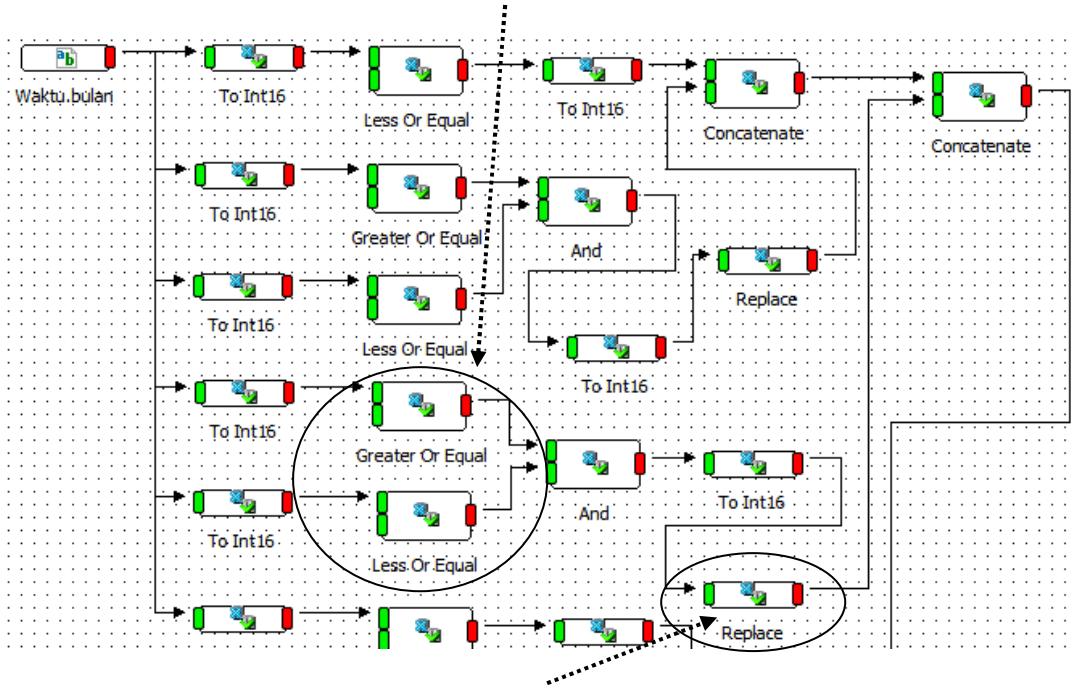
38. Isikan angka 4 pada kolom “Value” saat konfigurasi “Greater or Equal”, dan isikan angka 6 saat konfigurasi “Less or Equal” berikutnya.



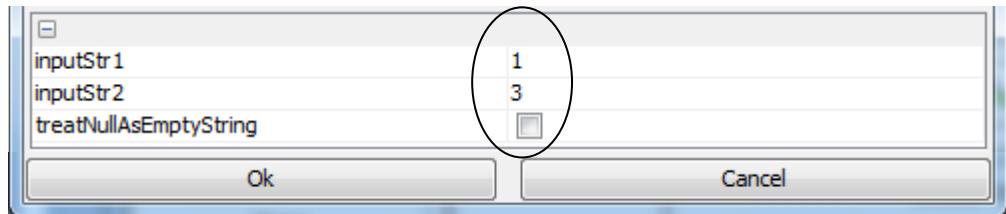
39. Isikan angka 1 dalam ”inputStr1” dan isikan angka 2 dalam “inputStr2” pada saat konfigurasi fungsi “Replace” yang menunjukkan bahwa jika outputnya 1 atau True, maka diubah menjadi 2 (artinya kuartal 2). Klik OK.



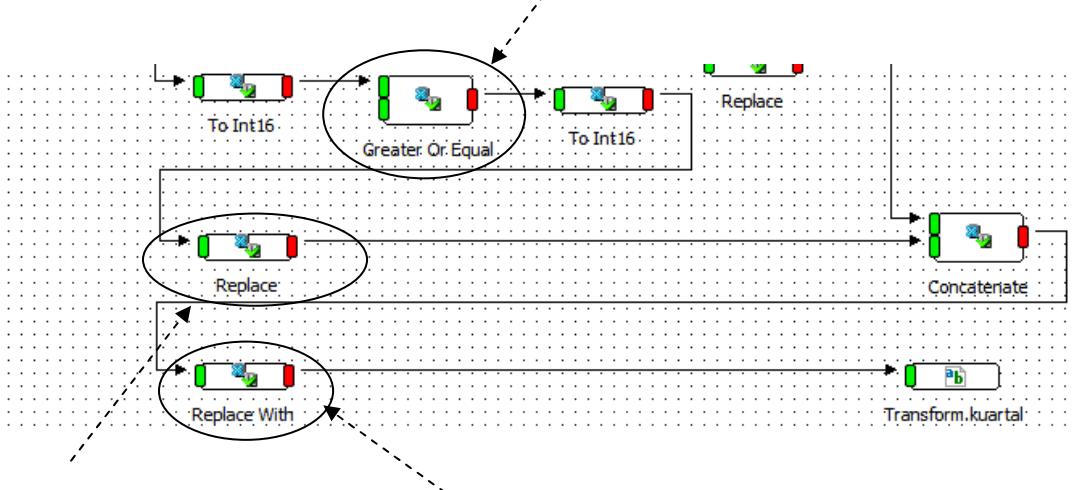
40. Atur konfigurasi berikutnya pada fungsi “Greater or Equal” dan “Less or Equal”. Masukkan angka 7 pada value saat konfigurasi “Greater or Equal” dan masukkan angka 9 pada value saat konfigurasi “Less or Equal” pada gambar berikut.



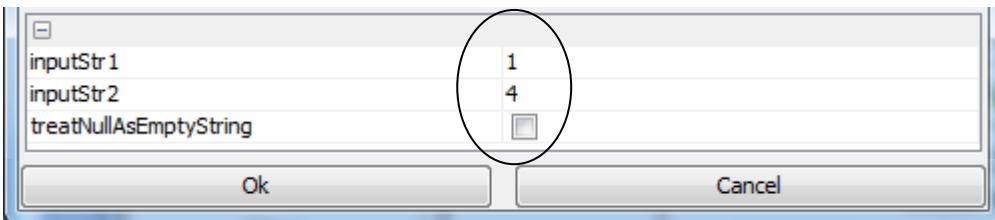
41. Isikan angka 1 dalam "inputStr1" dan isikan angka 3 dalam "inputStr2" pada saat konfigurasi fungsi "Replace" yang menunjukkan bahwa jika outputnya 1 atau True, maka diubah menjadi 3 (artinya kuartal 3). Klik OK.



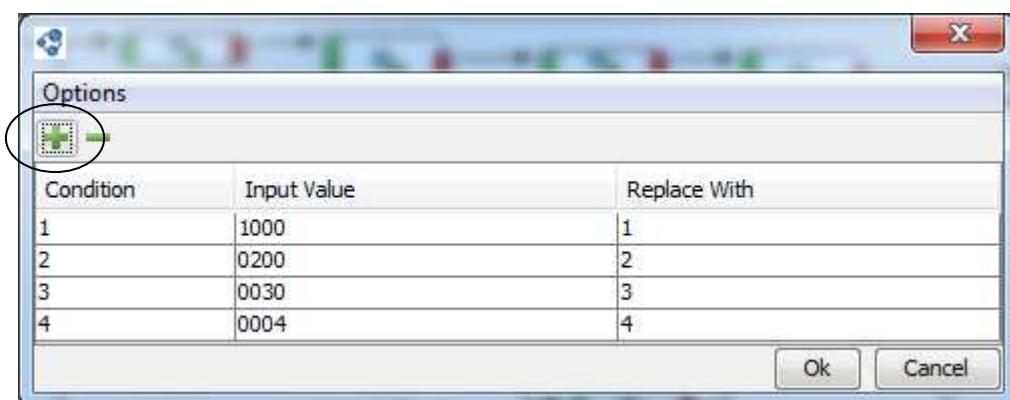
42. Atur konfigurasi berikutnya pada fungsi "Greater or Equal" yang terakhir. Masukkan angka 10 pada kolom "Value" yang menunjukkan jika bulan lebih dari atau sama dengan 10.



43. Pada konfigurasi “Replace”, masukkan angka 1 pada inputStr1 dan angka 4 pada inputStr2.



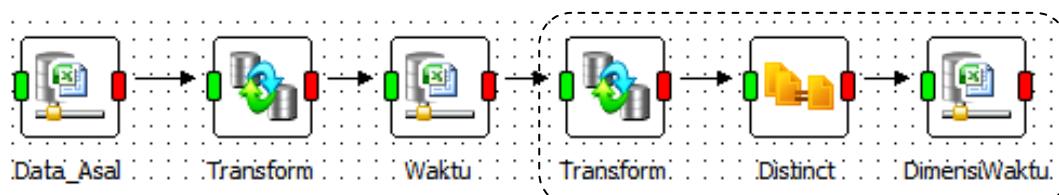
44. Atur konfigurasi “Replace With” dengan ketentuan sebagai berikut. Tambahkan kondisi dengan mengklik tombol plus (+) di sebelah kiri atas, sebanyak 4 kali sesuai dengan jumlah kondisinya. Masukkan nilai input value yang terdiri dari 4 digit akibat hasil proses “Concatenate” atau penggabungan antara satu data dengan data lainnya. Klik OK.



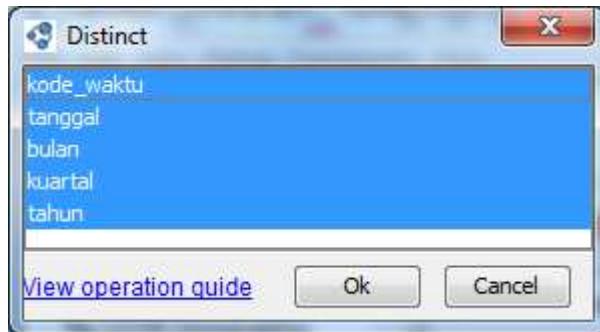
45. Tutup jendela properti “Transformation” dan kembali ke editor Apatar Tools.
46. Periksa terlebih dahulu proses transformasi tersebut apakah sudah benar atau belum dengan cara klik kanan operator “Transform”, pilih Run and Preview Results. Jika nilai yang tertera dalam kolom kuartal sudah sesuai dengan bulannya, maka proses transformasi kuartal berhasil dilakukan seperti tabel dalam gambar berikut.

No.	kode_waktu	tanggal	bulan	kuartal	tahun
1	2012-12-15	15	12	4	2012
2	2010-06-14	14	06	2	2010
3	2010-11-21	21	11	4	2010
4	2011-01-05	05	01	1	2011
5	2011-03-27	27	03	1	2011
6	2011-04-09	09	04	2	2011
7	2011-08-19	19	08	3	2011
8	2011-10-13	13	10	4	2011
9	2011-12-28	28	12	4	2011
10	2011-12-30	30	12	4	2011
11	2012-01-04	04	01	1	2012
12	2012-01-09	09	01	1	2012
13	2012-02-14	14	02	1	2012
14	2012-04-05	05	04	2	2012
15	2012-04-05	05	04	2	2012
16	2012-05-21	21	05	2	2012
17	2012-06-22	22	06	2	2012
18	2012-09-18	18	09	3	2012
19	2012-09-28	28	09	3	2012
20	2012-12-15	15	12	4	2012

47. Tapi perlu diperhatikan bahwa tabel dimensi seharusnya tidak mengandung duplikasi data, seperti contoh pada data ke-1 dan ke-20, atau data ke-14 dan ke-15 dalam tabel hasil langkah ke-46. Sehingga hal ini perlu dicegah sebelum dikirim (*load*) ke tabel tujuan “DimensiWaktu”. Untuk mengatasi hal tersebut, hapus terlebih dahulu hubungan antara “Transform” dengan connector “DimensiWaktu”. Tambahkan sebuah operator “Distinct” dan letakkan di antara operator “Transform” dan connector “DimensiWaktu”. Hubungkan kembali ketiga operator ini.



48. Klik kanan operator “Distinct”, pilih Configure. Pilih semua kolom dalam jendela konfigurasi dengan menekan tombol Ctrl+A. Kemudian klik OK.



49. Periksa kembali apakah proses “Distinct” berhasil dilakukan dengan cara klik kanan “Distinct” pilih Run and Preview Results.

No.	kode_waktu	tanggal	bulan	kuartal	tahun
1	2012-12-15	15	12	4	2012
2	2012-05-21	21	05	2	2012
3	2012-06-22	22	06	2	2012
4	2010-06-14	14	06	2	2010
5	2011-01-05	05	01	1	2011
6	2012-01-04	04	01	1	2012
7	2011-12-28	28	12	4	2011
8	2012-01-09	09	01	1	2012
9	2012-02-14	14	02	1	2012
10	2011-10-13	13	10	4	2011
11	2012-09-18	18	09	3	2012
12	2011-03-27	27	03	1	2011
13	2012-04-05	05	04	2	2012
14	2012-09-28	28	09	3	2012
15	2011-04-09	09	04	2	2011
16	2011-08-19	19	08	3	2011
17	2010-11-21	21	11	4	2010
18	2011-12-30	30	12	4	2011

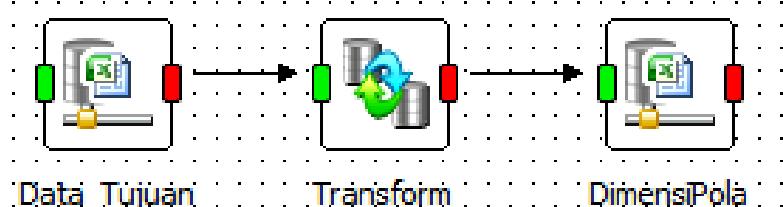
50. Lihatlah, data duplikasi telah dihilangkan sehingga tinggal menyisakan sebanyak 18 data dari total 20 data. Hal ini menunjukkan proses Distinct berhasil dilakukan.
51. Data siap dipindahkan (*load*) ke tabel tujuan “**DimensiWaktu**”. Pastikan semua file excel ditutup. Klik kanan connector “DimensiWaktu”, pilih Run and Preview Results. Jika tabel yang ditampilkan sama persis dengan tabel pada langkah ke-49, maka proses *load* dalam ETL telah berhasil dilakukan. Untuk memastikannya, buka kembali file “**Data_penjualan.xls**” pada sheet “**DimensiWaktu**”. Semua data waktu telah berhasil diekspor sesuai dengan proses transformasi.

	A	B	C	D	E	F	G
1	kode_waktu	tanggal	bulan	kuartal	tahun		
2	2012-12-15	15	12	4	2012		
3	2012-05-21	21	05	2	2012		
4	2012-06-22	22	06	2	2012		
5	2010-06-14	14	06	2	2010		
6	2011-01-05	05	01	1	2011		
7	2012-01-04	04	01	1	2012		
8	2011-12-28	28	12	4	2011		
9	2012-01-09	09	01	1	2012		
10	2012-02-14	14	02	1	2012		
11	2011-10-13	13	10	4	2011		
12	2012-09-18	18	09	3	2012		
13	2011-03-27	27	03	1	2011		
14	2012-04-05	05	04	2	2012		
15	2012-09-28	28	09	3	2012		
16	2011-04-09	09	04	2	2011		
17	2011-08-19	19	08	3	2011		
18	2010-11-21	21	11	4	2010		
19	2011-12-30	30	12	4	2011		
20							

2) Dimensi Pola

1. Tabel yang akan dibuat berikutnya adalah **Dimensi Pola**. Buka kembali file excel “**Data_penjualan.xls**”.
2. Buat Sheet baru. Ubah nama Sheet menjadi “**DimensiPola**”.
3. Buat kolom **kode_pola**, dan **nama_pola**. Simpan file excel, dan tutup kembali.
4. Buka kembali file apatar “**Transform_Dimensi.aptr**”.
5. Tambahkan sebuah operator “**Transform**” dan 2 buah connector “**MS Excel**” di bawah operasi transformasi Dimensi Waktu.
6. Dengan menggunakan file “**Data_penjualan.xls**”, atur konfigurasi MS Excel yang pertama dengan mengambil Sheet “**Data_Tujuan**” yang terdiri dari 2 kolom, sedangkan MS Excel yang kedua mengambil Sheet “**DimensiPola**” yang juga terdiri dari 2 kolom.

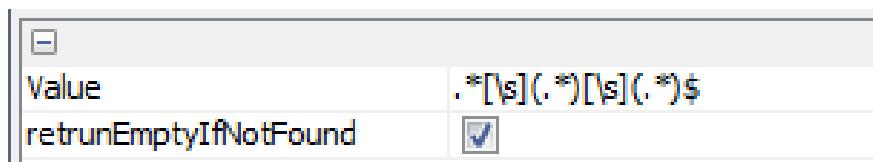
7. Hubungkan ketiga operator Data_Tujuan, Transform dan DimensiPola.



8. Aturlah konfigurasi Transform dengan klik kanan operator Transform, pilih Configure.
 9. Masukkan field **PRODUK** dari **Data_Tujuan** (sebelah kiri) dan field **nama_pola** dari **output** (sebelah kanan) ke dalam editor.
 10. Tambahkan fungsi **RegExp** yang diambil dari Function **String** ke dalam editor. Atur letak dan hubungan semua fungsi sedemikian rupa seperti pada gambar di bawah ini.



11. Untuk nama pola, akan diambil sebuah kata yang terkandung di dalam nama produk. Nama pola yang diinginkan hanya ada 3 macam, yaitu **Print**, **Cap** dan **Tulis**. Dari data asal, contoh nama produk adalah “**kaos Katun Print Bola**”. Nama pola dalam produk tersebut adalah **Print** yang terletak setelah spasi kedua. Jika dilihat dalam data asal (dalam file Excel), semua nama pola terletak setelah spasi kedua. Untuk memisahkan kata tersebut, maka diperlukan fungsi “**RegExp**” (Regular Expression). Atur konfigurasi fungsi RegExp, isikan **Value** dengan nilai “`.*[\s](.*)[\s](.*)$`”, dan beri centang pada **ReturnEmptyIfNotFound**.

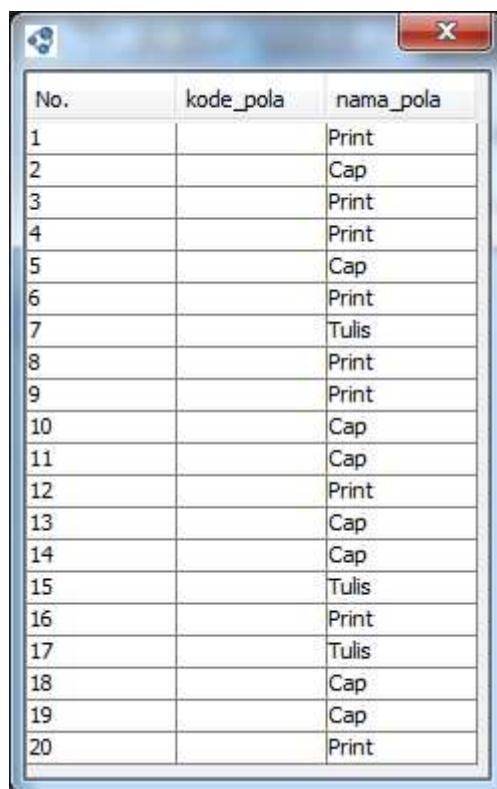


Simbol `[\s]` menyatakan karakter spasi, sedangkan simbol `$` menyatakan kata yang diambil. Contoh lebih lengkap buka halaman :

<http://www.apatarforge.org/wiki/display/AUG/Sample+Regular+Expressions>

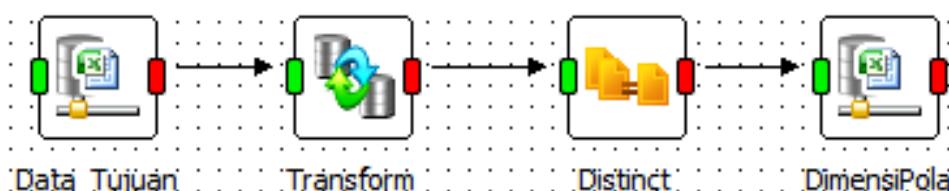
12. Tutup konfigurasi Transform dengan klik OK.

13. Data pola siap di transformasikan. Klik kanan operator **Transform**, pilih **Run and Preview Results**.
14. Hasil transformasi sementara dapat dilihat pada tabel yang dihasilkan seperti pada gambar berikut. Sebagai catatan, kode_pola masih kosong karena memang belum dikonfigurasi.



No.	kode_pola	nama_pola
1		Print
2		Cap
3		Print
4		Print
5		Cap
6		Print
7		Tulis
8		Print
9		Print
10		Cap
11		Cap
12		Print
13		Cap
14		Cap
15		Tulis
16		Print
17		Tulis
18		Cap
19		Cap
20		Print

15. Pada tabel tersebut dapat dilihat nama pola sudah terbentuk. Namun, tabel pola tersebut masih terdiri dari 20 data yang diambil dari seluruh data dalam tabel sumber. Padahal dalam Dimensi Pola hanya dibutuhkan satu kode untuk satu nama pola. Sehingga data dalam Dimensi Pola hanya akan dibuat sebanyak 3 data sesuai dengan jumlah nama pola.
16. Tambahkan sebuah operator Distinct ke dalam editor Apatar. Letakkan diantara Transform dan DimensiPola. Hubungkan operator-operator tersebut dengan connector seperti gambar berikut.



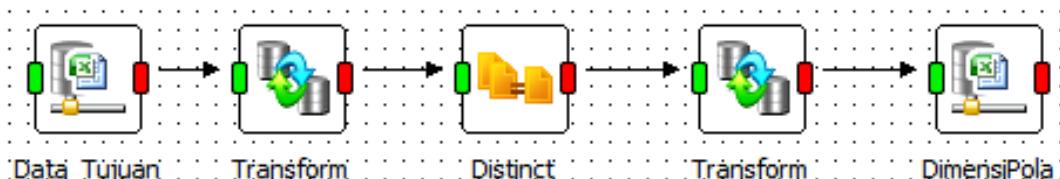
17. Atur konfigurasi operator Distinct. Pilih nama_pola dari jendela konfigurasi untuk menghilangkan terjadinya duplikasi data nama_pola.



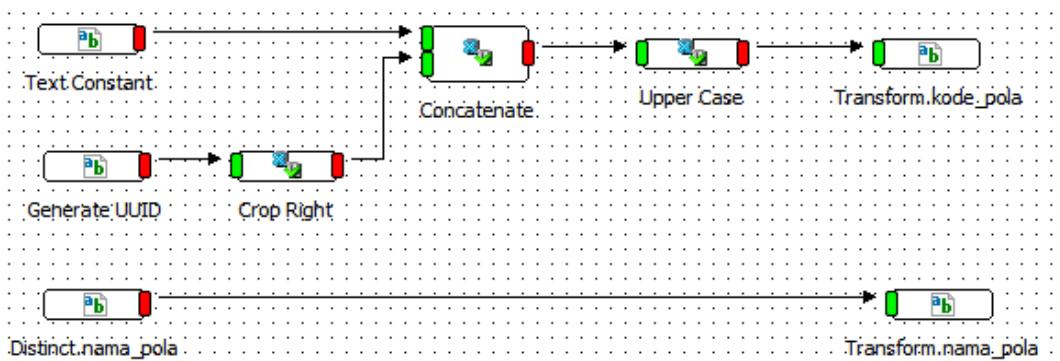
18. Klik OK, jalankan operator Distinct dengan klik kanan operator tersebut dan pilih Run and Preview Results. Tabel sementara akan dapat dilihat seperti gambar berikut.

No.	kode_pola	nama_pola
1		Cap
2		Tulis
3		Print

19. Untuk menambahkan kode_pola dalam DimensiPola, tambahkan sebuah operator Transform ke dalam editor Apatar, letakkan diantara operator Distinct dan connector DimensiPola. Hubungkan operator tersebut dengan connector seperti pada gambar berikut.



20. Atur konfigurasi Transform tersebut. Masukkan semua field dari operator Distinct (sebelah kiri) dan Output (sebelah kanan) ke dalam editor. Tambahkan juga fungsi **Text Constant**, **Generate UUID**, **Crop Right**, **Concatenate**, **Upper Case**, dan **RegExp** dari Function String.

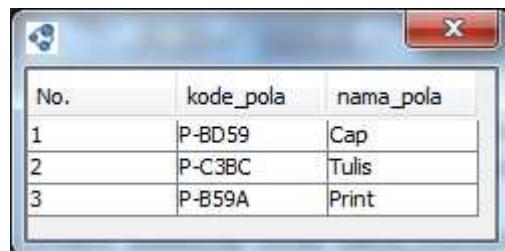


21. Atur konfigurasi satu persatu semua fungsi tersebut dengan klik kanan fungsi, pilih Configure.
22. Untuk kode pola, akan dibuat terdiri atas 6 karakter yang terdiri dari karakter konstan P diikuti tanda strip (-), dan diikuti 4 karakter secara acak. Contoh kode pola = “P-0A3D”. Maka atur konfigurasi sebagai berikut:
 - a. **Text Constant:** isi Value dengan “P-“. Ini digunakan untuk menambahkan karakter “P-“ pada bagian awal kode pola.
 - b. **Generate UUID:** tidak perlu dikonfigurasi karena fungsi ini digunakan untuk membuat ID secara unik dan random (alfanumerik). Namun karena fungsi ini akan membuat ID yang terdiri atas lebih dari 10 karakter, maka diperlukan fungsi Crop untuk memotong karakter dengan jumlah tertentu.
 - c. **Crop Right:** fungsi ini digunakan untuk mengambil karakter dengan jumlah tertentu dihitung dari paling kanan. Sesuai kode pola yang diinginkan, maka akan diambil sejumlah 4 karakter dari kanan. Sehingga isikan Value = 4 pada jendela konfigurasi.
 - d. **Concatenate:** digunakan untuk menggabungkan 2 cabang fungsi.
 - e. **Upper Case:** digunakan untuk mengubah karakter menjadi huruf kapital semua.
23. Tutup jendela konfigurasi Transform, dan jalankan operator Transform untuk memastikan terlebih dahulu bahwa proses transformasi sudah sesuai dengan kebutuhan. Klik kanan operator Transform, pilih Run and Preview Results. Hasilnya dapat dilihat seperti pada gambar berikut. Sebagai catatan, kode_pola bisa berbeda-beda karena di-generate secara random.

No.	kode_pola	nama_pola
1	P-389D	Cap
2	P-C202	Tulis
3	P-A252	Print

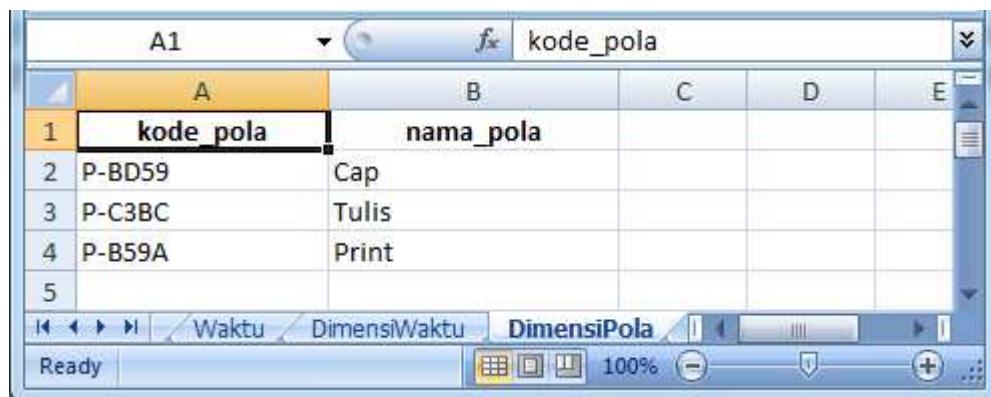
24. Jika proses transformasi di atas sudah sesuai kebutuhan, maka data siap di-load ke dalam tabel DimensiPola dalam file Excel. **Pastikan semua file Excel telah ditutup.**

25. Proses *Loading* dapat dilakukan dengan klik kanan connector MS Excel (DimensiPola), pilih Run and Preview Results. Hasil proses *Loading* dapat ditampilkan pada tabel berikut.



No.	kode_pola	nama_pola
1	P-BD59	Cap
2	P-C3BC	Tulis
3	P-B59A	Print

26. Untuk memastikan bahwa semua data telah di-*load* ke dalam tabel tujuan (DimensiPola), buka file excel “**Data_penjualan.xls**”, pilih Sheet “**DimensiPola**”. Jika dalam tabel terdapat data pola, maka proses ETL berhasil dilakukan.

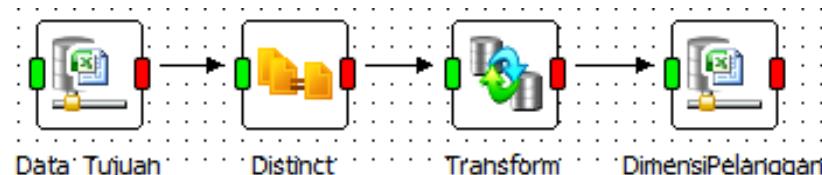


A	B	C	D	E
1	kode_pola	nama_pola		
2	P-BD59	Cap		
3	P-C3BC	Tulis		
4	P-B59A	Print		
5				

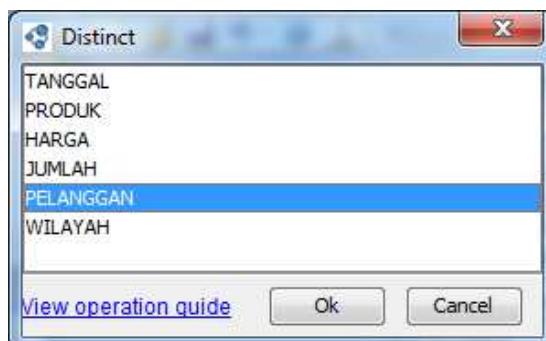
3) Dimensi Pelanggan

1. Tabel yang akan dibuat berikutnya adalah **Dimensi Pelanggan**. Buka kembali file excel “**Data_penjualan.xls**”.
2. Buat Sheet baru. Ubah nama Sheet menjadi “**DimensiPelanggan**”.
3. Buat kolom **kode_pelanggan**, **nama_pelanggan**, dan **kode_jeniskelamin**. Kode_jeniskelamin digunakan sebagai *Foreign Key* yang nanti akan dihubungkan dengan Sub Dimensi Jenis Kelamin. Simpan file excel, dan tutup kembali.
4. Buka kembali file apatar “**Transform_Dimensi.aptr**”.
- 5.Tambahkan sebuah operator “**Distinct**”, sebuah operator “**Transform**” dan 2 buah connector “**MS Excel**” di bawah operasi transformasi Dimensi Pola.

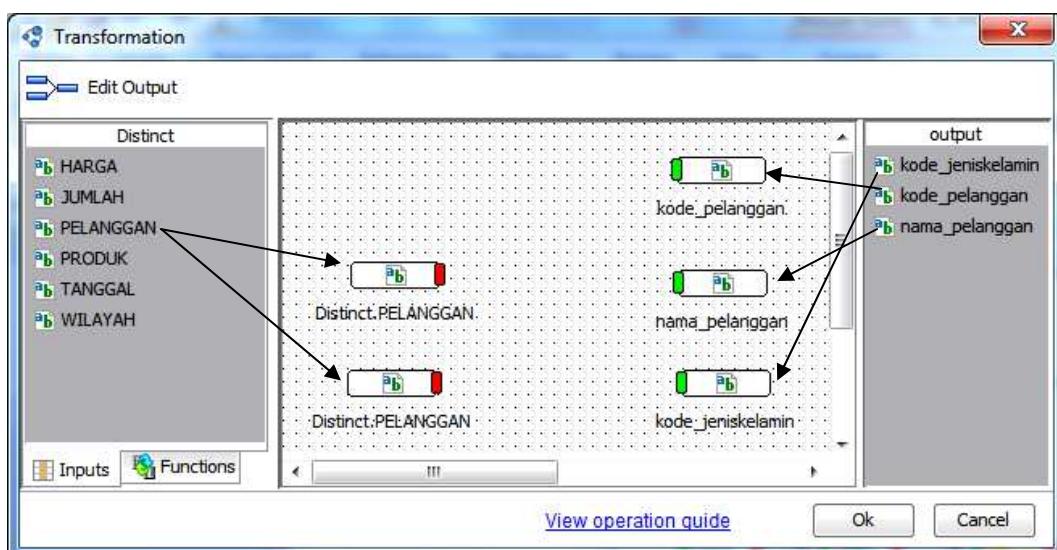
- Dengan menggunakan file “Data_penjualan.xls”, atur konfigurasi MS Excel yang pertama dengan mengambil Sheet “Data_Tujuan” yang terdiri dari 6 kolom, sedangkan MS Excel yang kedua mengambil Sheet “DimensiPelanggan” yang terdiri dari 3 kolom.
- Hubungkan keempat operator Data_Tujuan, Distinct, Transform dan DimensiPelanggan.



- Atur konfigurasi Distinct. Pilih field PELANGGAN, yang berfungsi untuk menghilangkan duplikasi data nama pelanggan. Sehingga data nama pelanggan tidak terjadi pengulangan. Klik OK.



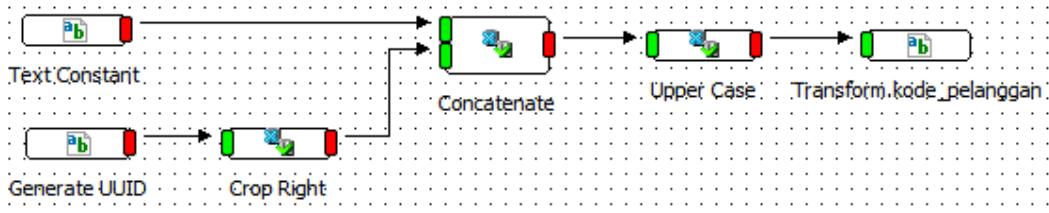
- Atur konfigurasi Transform. Letakkan field PELANGGAN ke dalam editor sebanyak 2 kali yang diambil dari Distinct (sebelah kiri), dan letakkan semua field Output (sebelah kanan) ke dalam editor transform.



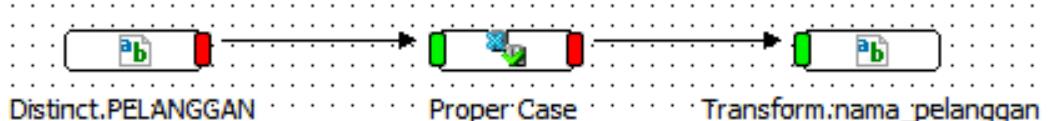
10. Untuk pembuatan kode_pelanggan, akan dibuat dengan cara yang sama saat membuat kode_pola, yaitu terdiri atas 6 karakter yang terdiri dari 2 karakter konstan dan 4 karakter alfanumerik secara random.

Contoh: **C-A45D**, C diawal untuk menunjukkan “Customer”

11. Tambahkan fungsi **Text Constant**, **Generate UUID**, **Crop Right**, **Concatenate**, dan **Upper Case**. Hubungkan semua fungsi tersebut dengan urutan seperti gambar berikut. Isikan *value* pada fungsi Text Constant = “C-“ dan *value Crop Right* = 4.

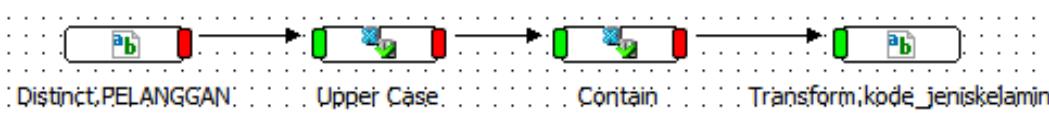


12. Untuk men-*transform* nama_pelanggan, tambahkan sebuah fungsi **Proper Case** dari kategori fungsi *String* untuk membuat huruf besar di awal kata. Letakkan fungsi ini di antara Distinct.PELANGGAN dan Transform.nama_pelanggan. Hubungkan ketiga operator tersebut.



13. Untuk mengisi kode_jeniskelamin, data akan dibuat menggunakan angka biner yaitu 0 dan 1. Angka 0 untuk menunjukkan WANITA, dan 1 menunjukkan PRIA.

14. Tambahkan fungsi **Upper Case** dan **Contain** yang diambil dari fungsi kategori *String*. Hubungkan semua fungsi tersebut dengan Distinct.PELANGGAN yang kedua dan Transform.kode_jeniskelamin seperti gambar berikut.



15. Upper Case digunakan untuk mengubah karakter menjadi huruf kapital, hal ini diperlukan untuk mencegah terjadinya perbedaan huruf pada data pelanggan yang sama. Sedangkan fungsi Contain digunakan untuk

memisahkan data yang mengandung suatu karakter / kata tertentu. Dalam kasus ini, nama pelanggan mengandung kata “BAPAK” atau “IBU”. Sehingga dapat dimanfaatkan untuk membuat jenis kelamin berdasarkan kata tersebut. Oleh karena itu, isikan nilai *value* pada fungsi Contain = BAPAK. Dengan demikian, kode_jeniskelamin untuk nama yang mengandung kata “BAPAK” akan bernilai 1, sedangkan yang tidak mengandung kata “BAPAK” akan bernilai 0.

16. Tutuplah editor konfigurasi Transform dengan klik OK, dan kembali ke editor Apatar.
17. Untuk memastikan bahwa proses transformasi untuk Dimensi Pelanggan berhasil dilakukan, klik kanan operator Transform dan pilih Run and Preview Results.

No.	kode_pelanggan	nama_pelanggan	kode_jeniskelamin
1	C-EC10	Bapak Imron	1
2	C-C60F	Ibu Aini Kasmaji	0
3	C-7614	Ibu Harini	0
4	C-8848	Ibu Niken	0
5	C-894A	Bapak Ketut	1
6	C-37AF	Bapak Heru	1
7	C-2AF5	Ibu Hadi Sukarni	0
8	C-CDB2	Ibu Atik	0
9	C-9064	Ibu Tyas	0
10	C-DC40	Bapak Totok	1
11	C-2A86	Ibu Hatamah	0
12	C-2410	Ibu Siti Arya	0

18. Jika sudah sesuai dengan yang diharapkan, maka proses *Load* ke tabel tujuan siap dilakukan. Pastikan semua file Excel telah ditutup. Klik kanan connector **DimensiPelanggan**, pilih Run and Preview Results.
19. Untuk melihat hasil proses ETL, buka kembali file Excel “**Data_penjualan.xls**” pada Sheet **DimensiPelanggan**. Data kode_pelanggan bisa berbeda-beda karena dibuat secara random.
20. Lihatlah data nama_pelanggan, pastikan tidak terjadi duplikasi data. Sedangkan data pada kolom kode_jeniskelamin, apakah sudah sesuai dengan angka binernya yang menunjukkan bahwa 0 = “IBU”, dan 1 = “BAPAK”? Jika sudah sesuai, maka proses ETL telah berhasil dilakukan.

	A	B	C
1	kode_pelanggan	nama_pelanggan	kode_jeniskelamin
2	C-6434	Bapak Imron	1
3	C-C954	Ibu Aini Kasmaji	0
4	C-E61F	Ibu Harini	0
5	C-2596	Ibu Niken	0
6	C-BD48	Bapak Ketut	1
7	C-163A	Bapak Heru	1
8	C-B5EB	Ibu Hadi Sukarni	0
9	C-14DB	Ibu Atik	0
10	C-5A8C	Ibu Tyas	0
11	C-C899	Bapak Totok	1
12	C-E04D	Ibu Hatamah	0
13	C-8421	Ibu Siti Arya	0

4) Dimensi Jenis Kelamin

Dimensi Jenis Kelamin dibuat untuk menyimpan data jenis kelamin. Tabel ini hanya akan diisi dengan 2 buah data yaitu PRIA dan WANITA, sehingga dirancang hanya terdiri atas 2 kolom yaitu **kode_jeniskelamin** dan **jenis_kelamin**. Kode jenis kelamin berisi angka biner yaitu 0 yang menunjukkan WANITA dan 1 untuk menunjukkan PRIA. Untuk membuat dimensi jenis kelamin, langkah-langkahnya sebagai berikut:

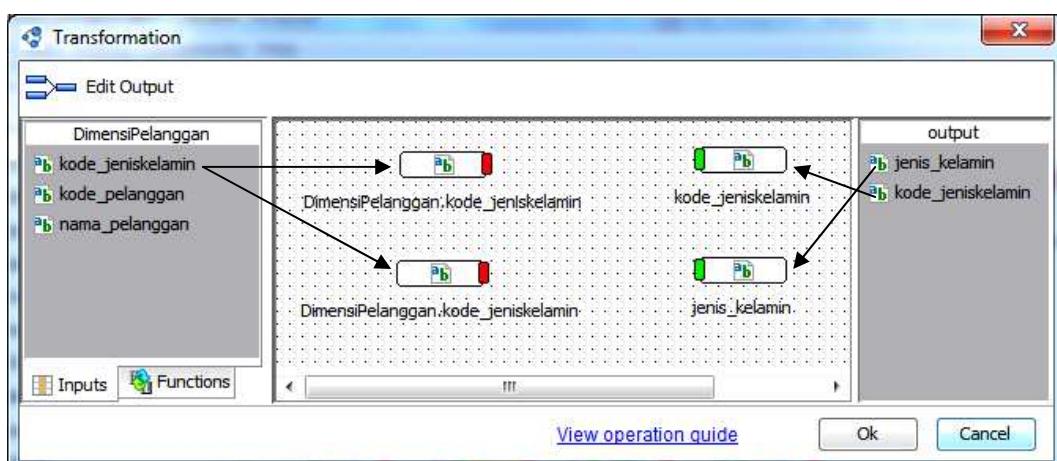
1. Tabel yang akan dibuat berikutnya adalah **Dimensi Jenis Kelamin**. Buka kembali file excel “**Data_penjualan.xls**”.
2. Buat Sheet baru. Ubah nama Sheet menjadi “**DimensiJenisKelamin**”.
3. Buat kolom **kode_jeniskelamin**, dan **jeniskelamin**. Simpan file excel, dan tutup kembali.
4. Buka kembali file apatar “**Transform_Dimensi.aptr**”.
- 5.Tambahkan sebuah operator “**Transform**”, sebuah operator “**Distinct**” dan 2 buah connector “**MS Excel**” di bawah operasi transformasi Dimensi Pelanggan.
6. Dengan menggunakan file “**Data_penjualan.xls**”, atur konfigurasi MS Excel yang pertama dengan mengambil Sheet “**DimensiPelanggan**” dan pilih pada kolom ke-3 (isikan angka 3 pada firstFieldPosition dan

lastFieldPosition), sedangkan MS Excel yang kedua mengambil Sheet “**DimensiJenisKelamin**” yang terdiri dari 2 kolom.

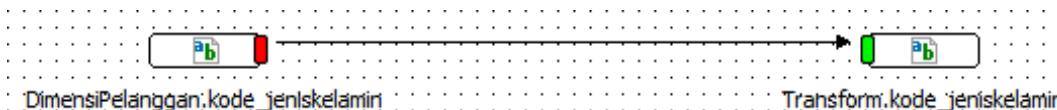
7. Hubungkan operator DimensiPelanggan, Transform, Distinct dan DimensiJenisKelamin.



8. Atur konfigurasi Transform. Klik kanan operator Transform, pilih menu Configure.
9. Masukkan field kode_jeniskelamin dari field DimensiPelanggan (sebelah kiri) ke dalam editor sebanyak 2 kali, dan masukkan semua field dari Output (sebelah kanan).



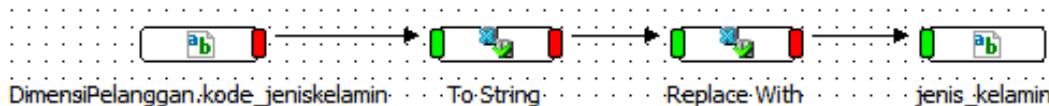
10. Hubungkan secara langsung antara **DimensiPelanggan.kode_jeniskelamin** dengan **kode_jeniskelamin**. Pada field ini tidak ada perubahan yang diperlukan.



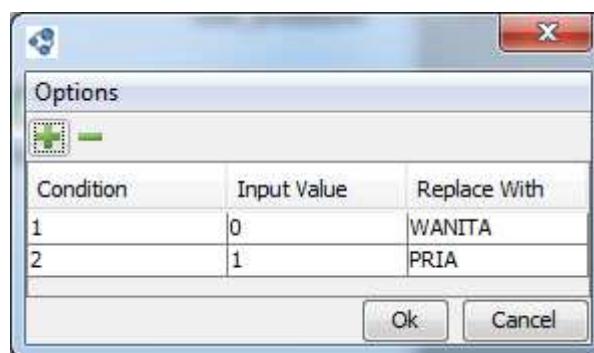
11. Untuk membuat jenis kelamin, diperlukan proses referensi dari kode jenis kelamin, yaitu jika kode jenis kelamin adalah 1 maka jenis kelaminnya PRIA, dan jika kode jenis kelamin adalah 0 maka WANITA. Sehingga diperlukan 2 buah fungsi **To String** dan **Replace With**. Tambahkan kedua buah fungsi tersebut ke dalam editor. Fungsi **To String** digunakan untuk mengubah bilangan biner menjadi karakter string, sedangkan

Replace With digunakan untuk mengubah suatu karakter menjadi karakter lainnya.

12. Hubungkan kedua fungsi tersebut dengan field input dan output seperti gambar berikut.



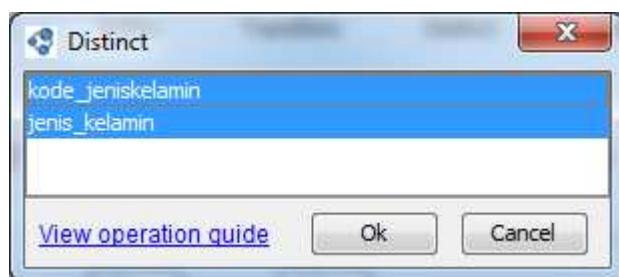
13. Atur konfigurasi Replace With dengan cara klik kanan fungsi Replace With, pilih Configure.
14. Pada jendela konfigurasi Replace With, tambahkan 2 buah kondisi dengan cara klik 2 kali tanda plus (+) di bawah menu Options.
15. Isikan data **input value** dan **replace with** dalam tabel dengan nilai sebagai berikut:



16. Tutup jendela konfigurasi Replace With dan jendela konfigurasi Transform dengan cara klik OK.
17. Untuk memastikan terlebih dahulu proses transformasi berhasil, klik kanan operator Transform pada editor Apatar, pilih Run and Preview Results. Jika berhasil akan ditampilkan 12 data seperti tabel berikut.

No.	kode_jeniskelamin	jenis_kelamin
1	1	PRIA
2	0	WANITA
3	0	WANITA
4	0	WANITA
5	1	PRIA
6	1	PRIA
7	0	WANITA
8	0	WANITA
9	0	WANITA
10	1	PRIA
11	0	WANITA
12	0	WANITA

18. Dalam tabel hasil transformasi masih terdapat 12 data (terjadi duplikasi data), padahal yang diperlukan dalam Dimensi Jenis Kelamin hanya terdiri dari 2 data yaitu PRIA dan WANITA. Sehingga diperlukan operator Distinct.
19. Atur konfigurasi Distinct dengan klik kanan pilih Configure. Pilih 2 buah field kode_jeniskelamin dan jenis_kelamin dengan cara tekan dan tahan tombol Shift pada keyboard diikuti dengan klik masing-masing field dalam jendela konfigurasi. Tutup kembali dengan klik OK.



20. Periksalah hasil proses Distinct dengan cara klik kanan operator Distinct, pilih Run and Preview Results. Jika dalam tabel hasil hanya terdiri dari 2 data (PRIA dan WANITA) maka proses berhasil, sehingga data siap dikirim ke **DimensiJenisKelamin** dalam Excel (proses *Loading*).

No.	kode_jeniskelamin	jenis_kelamin
1	1	PRIA
2	0	WANITA

21. Proses *Loading* siap dilakukan. Pastikan file Excel telah ditutup. Klik kanan DimensiJenisKelamin, pilih Run and Preview Results.
22. Jika berhasil akan ditampilkan tabel yang sama seperti pada langkah ke-20 di atas. Buka kembali file Excel “**Data_penjualan.xls**” pada Sheet **DimensiJenisKelamin**. Jika sudah sesuai, maka proses ETL berhasil.

	A	B	C
1	kode_jeniskelamin	jenis_kelamin	
2	1	PRIA	
3	0	WANITA	
4			

DimensiJenisKelamin

Ready 100%

E. Tugas

Dengan menggunakan file Excel “**Data_penjualan.xls**”, selesaikan tugas berikut di kelas. Jika belum selesai, bisa dilanjutkan di rumah dan akan dinilai pada pertemuan berikutnya.

1. Buat Dimensi Wilayah yang terdiri dari 2 buah kolom yaitu **kode_wilayah**, dan **nama_wilayah**. Data kode_wilayah dibuat terdiri atas 6 karakter yaitu 2 karakter pertama sebagai konstanta dan 4 karakter berikutnya secara random. Contoh kode_wilayah = ”**W-3C5T**” yang menunjukkan W = Wilayah. **Nama_wilayah** diambil dari Data_Tujuan kolom WILAYAH dan tidak diijinkan terjadi duplikasi data nama wilayah. Contoh nama_wilayah = ”Jawa Tengah”.

MODUL 3

PROSES EXTRACT-TRANSFORM-LOAD

(DATA CLEANSING)

A. Tujuan

1. Mahasiswa mampu melakukan proses ETL secara lebih lanjut pada pengembangan sebuah *Data Warehouse*.
2. Mahasiswa mampu melakukan proses perbaikan data sebelum data dimasukkan ke *data warehouse*.

B. Landasan Teori

Data Extraction adalah proses pengambilan data yang diperlukan dari sumber data dan selanjutnya dimasukkan pada *staging area* untuk diproses pada tahap berikutnya. Terdapat berbagai tipe sumber data, berbagai format data, mesin yang berbeda, *software* dan arsitektur yang tidak sama. Sebelum proses ini dilakukan, sebaiknya perlu didefinisikan *requirement* terhadap sumber data yang akan digunakan untuk lebih memudahkan pada *extraction data*.

Data cleansing bertujuan untuk menghilangkan kesalahan-kesalahan pada data yang diakibatkan oleh proses transaksional. Latar belakang yang penting perlunya *Data Cleansing* adalah bahwa jika *data cleansing* ini salah maka hal terburuk yang terjadi adalah pemberian informasi yang salah kepada pengambil kebijakan. Jika informasi yang salah ini dipercaya maka keputusan yang diambil akan jatuh dan bisa mengakibatkan kerugian yang besar.

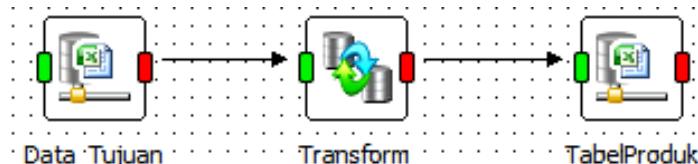
C. Alat dan Bahan

1. Komputer dengan sistem operasi Windows.
2. Program aplikasi Ms. Office
3. Program aplikasi Apatar Tool.
4. Modul Praktikum Data Warehousing dan Data Mining.

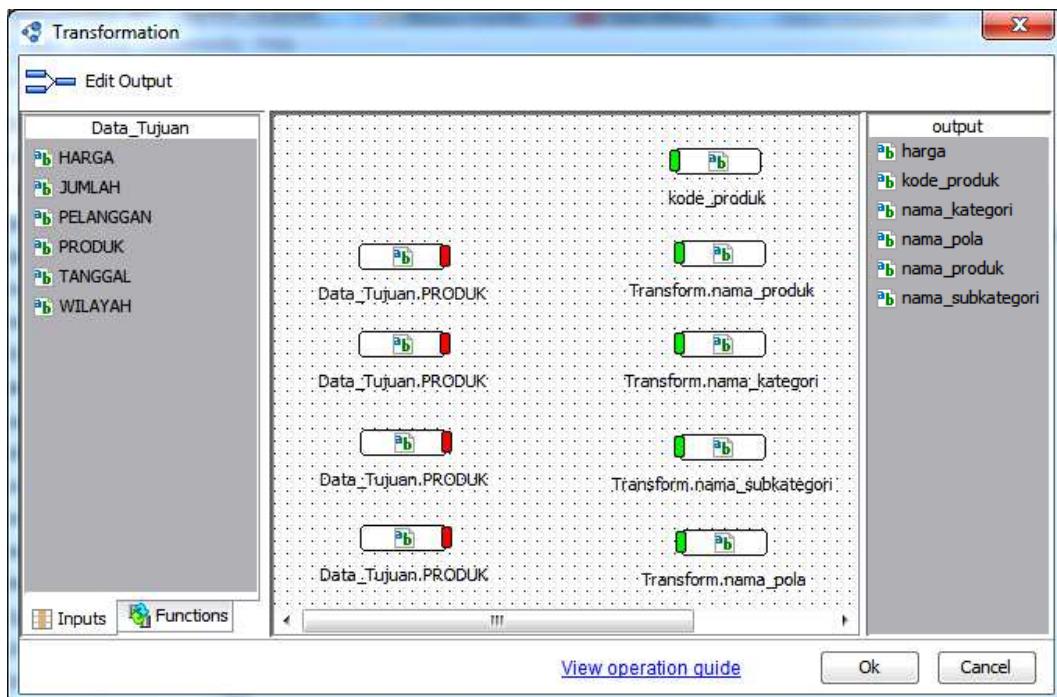
D. Langkah-langkah Praktikum

1) Tabel Produk

1. Sebelum membuat **Dimensi Produk**, tabel yang akan dibuat berikutnya adalah **Tabel Produk**. Hal ini diperlukan untuk memisahkan terlebih dahulu data produk menjadi beberapa bagian. Buka kembali file excel “**Data_penjualan.xls**”.
2. Buat Sheet baru. Ubah nama Sheet menjadi “**TabelProduk**”. Nama tersebut digunakan karena tabel ini belum menjadi sebuah Dimensi.
3. Buat 6 buah kolom yaitu **kode_produk**, **nama_produk**, **nama_kategori**, **nama_subkategori**, **nama_pola** dan **harga**. Simpan file excel, dan tutup kembali.
4. Buka kembali file apatar “**Transform_Dimensi.aptr**”.
- 5.Tambahkan sebuah operator “**Transform**” dan 2 buah connector “**MS Excel**” di bawah operasi transformasi Dimensi Waktu.
6. Dengan menggunakan file “**Data_penjualan.xls**”, atur konfigurasi MS Excel yang pertama dengan mengambil Sheet “**Data_Tujuan**” yang terdiri dari 6 kolom, sedangkan MS Excel yang kedua mengambil Sheet “**TabelProduk**” yang terdiri dari 6 kolom.
7. Hubungkan operator Data_Tujuan, Transform, dan TabelProduk.



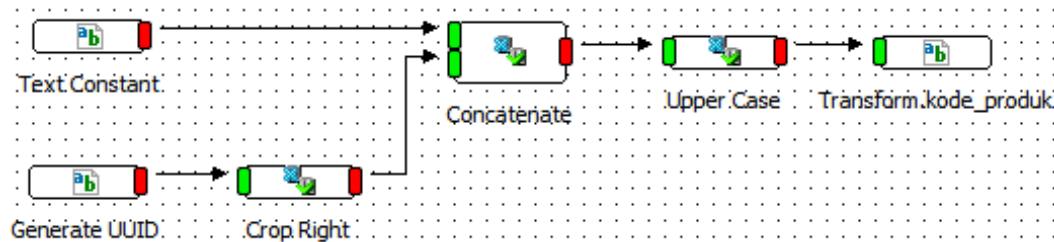
8. Atur konfigurasi Transform. Masukkan field PRODUK dari Data_Tujuan (sebelah kiri) ke dalam editor sebanyak 4 kali dan field HARGA sebanyak 1 kali. Masukkan pula semua field (6 field) dari tabel output (sebelah kanan) ke dalam editor konfigurasi Transform.



9. Untuk pembuatan kode_produk, akan dibuat terdiri atas 8 karakter yang terdiri dari 4 karakter awal sebagai konstanta dan diikuti 4 karakter alfanumerik secara random.

Contoh: **PRO-A45D**, PRO diawal untuk menunjukkan “Produk”

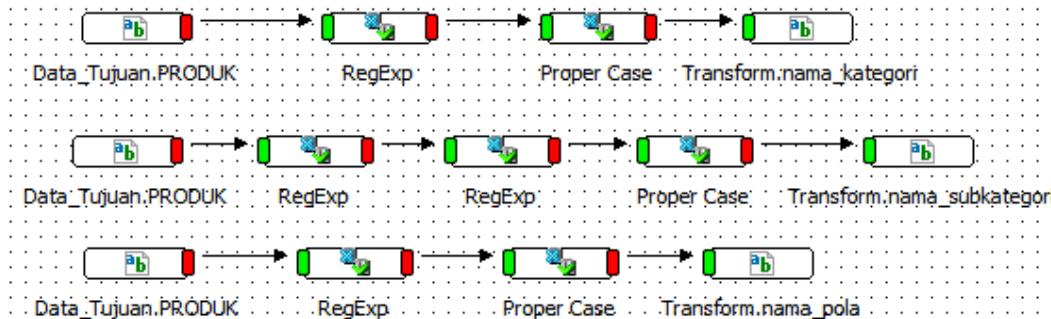
- 10.Tambahkan fungsi **Text Constant**, **Generate UUID**, **Crop Right**, **Concatenate**, dan **Upper Case**. Hubungkan semua fungsi tersebut dengan urutan seperti gambar berikut. Isikan *value* pada fungsi Text Constant = “PRO-“ dan *value* Crop Right = 4.



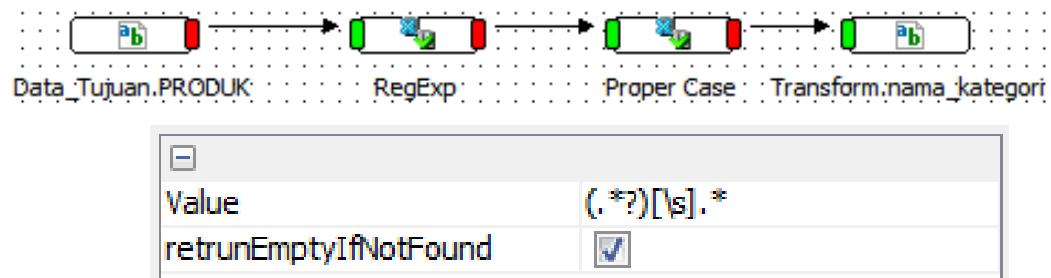
- 11.Tambahkan fungsi **Proper Case** yang diambil dari Function String ke dalam editor. Proper Case digunakan untuk membuat huruf kapital di setiap awal kata. Letakkan fungsi ini di antara Data_Tujuan.PRODUK dan Transform.nama_produk seperti gambar berikut.



12. Tambahkan 4 buah fungsi **RegExp** dan 3 buah fungsi **Proper Case** yang diambil dari Function **String** ke dalam editor. Atur letak dan hubungan semua fungsi sedemikian rupa seperti pada gambar di bawah ini.



13. Untuk nama kategori, akan diambil sebuah kata yang terkandung di dalam nama produk. Dari data asal, contoh nama produk adalah "**kaos Katun Print Bola**". Nama kategori dalam produk tersebut adalah **kaos** yang terletak paling depan. Untuk memisahkan kata tersebut, maka diperlukan fungsi "**RegExp**" (Regular Expression). Atur konfigurasi fungsi RegExp, isikan **Value** dengan nilai "(.*?)[\s].*", dan beri centang pada **ReturnEmptyIfNotFound**.



Simbol `[\s]` menyatakan karakter spasi. Contoh lebih lengkap buka halaman :

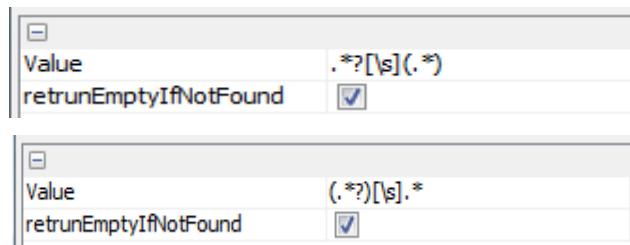
<http://apatar.com/news/apatar-announces-validation-with-regular-expressions>
atau:

<http://www.apatarforge.org/wiki/display/AUG/Sample+Regular+Expressions>

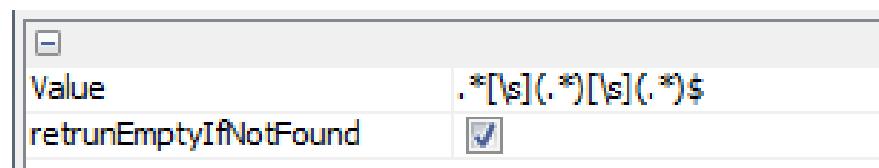
14. Untuk nama subkategori, akan diambil sebuah kata yang terkandung di dalam nama produk. Dari data asal, contoh nama produk adalah "**kaos Katun Print Bola**". Nama subkategori dalam produk tersebut adalah **Katun** yang terletak setelah spasi pertama. Jika dilihat dalam data asal (dalam file Excel), semua nama subkategori terletak setelah spasi pertama.



15. Untuk memisahkan kata tersebut, atur konfigurasi fungsi RegExp yang pertama, isikan **Value** dengan nilai “.*?[\s](.*)”, dan beri centang pada **ReturnEmptyIfNotFound**. Sedangkan Value pada RegExp yang kedua diisi dengan nilai “(.*)[\s].*” dan beri centang pada **ReturnEmptyIfNotFound**.



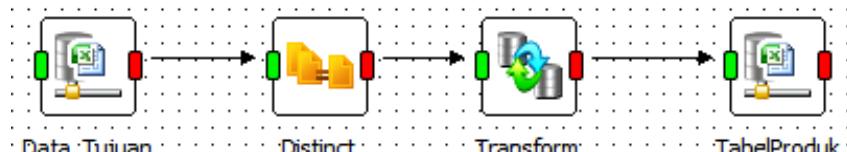
16. Untuk nama pola, akan diambil sebuah kata yang terkandung di dalam nama produk. Nama pola yang diinginkan hanya ada 3 macam, yaitu **Print**, **Cap** dan **Tulis**. Dari data asal, contoh nama produk adalah “**kaos Katun Print Bola**”. Nama pola dalam produk tersebut adalah **Print** yang terletak setelah spasi kedua. Jika dilihat dalam data asal (dalam file Excel), semua nama pola terletak setelah spasi kedua. Untuk memisahkan kata tersebut, maka diperlukan fungsi “**RegExp**” (*Regular Expression*). Atur konfigurasi fungsi RegExp, isikan **Value** dengan nilai “.*[\s](.)*[\s](.)*\$”, dan beri centang pada **ReturnEmptyIfNotFound**.



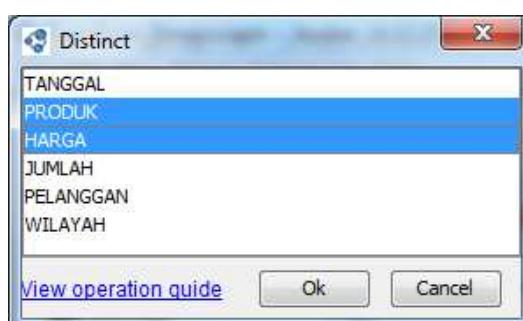
17. Untuk mengisi harga, hubungkan secara langsung field harga dari Data_Tujuan dengan field harga dari output.
18. Tutup konfigurasi Transform dengan klik OK.
19. Untuk memastikan bahwa proses transformasi sudah sesuai dengan kebutuhan, jalankan operator Transform dengan klik kanan Transform, pilih Run and Preview Results.

No.	kode_produk	nama_produk	nama_kategori	nama_subkategori	nama_pola	harga
1	PRO-362C	Bahan Standar Cap Lasem	Bahan	Standar	Cap	120000
2	PRO-18AE	Hem Standar Cap Tumpal	Hem	Standar	Cap	100000
3	PRO-C017	Hem Sutra Print Rama	Hem	Sutra	Print	100000
4	PRO-E771	Batik Standar Cap Tumpal	Batik	Standar	Cap	150000
5	PRO-764F	Rok Batik Print Kombinasi	Rok	Batik	Print	225000
6	PRO-F183	Kaos Batik Cap Lukis	Kaos	Batik	Cap	30000
7	PRO-56BF	Jarik Standar Tulis Sarimbit	Jarik	Standar	Tulis	40000
8	PRO-243E	Hem Standar Tulis Madura	Hem	Standar	Tulis	550000
9	PRO-A340	Celana Standar Cap Warna	Celana	Standar	Cap	55000
10	PRO-AEF1	Bahan Lawasan Tulis Tolet	Bahan	Lawasan	Tulis	130000
11	PRO-20A8	Kaos Katun Print Bola	Kaos	Katun	Print	60000
12	PRO-13CA	Jarik Standar Print Sogan	Jarik	Standar	Print	225000
13	PRO-5686	Jam Standar Print Lukis	Jam	Standar	Print	80000
14	PRO-0A2A	Bahan Beludru Cap Mahkota	Bahan	Beludru	Cap	500000
15	PRO-ABB7	Hem Katun Print Kelenggan	Hem	Katun	Print	299000
16	PRO-E63A	Sarimbit Standar Print Lukis	Sarimbit	Standar	Print	150000
17	PRO-4B2E	Bahan Standar Cap Garis	Bahan	Standar	Cap	135000
18	PRO-351C	Bolero Standar Cap Sidomukti	Bolero	Standar	Cap	225000
19	PRO-82A6	Hem Katun Print Kawung	Hem	Katun	Print	70000
20	PRO-C319	Celana Standar Print Lasem	Celana	Standar	Print	55000

20. Sebagai catatan, data produk dalam tabel hasil transformasi tidak diijinkan terjadi duplikasi data, yang berarti bahwa dalam satu tabel, 1 jenis produk hanya memiliki 1 buah data dan 1 kode produk. Namun dalam contoh ini tidak terjadi duplikasi data. Meski demikian, akan lebih baik jika duplikasi data tetap dihilangkan.
21. Dalam editor apatar, tambahkan sebuah operator **Distinct**. Letakkan di antara connector Data Tujuan dan operator Transform. Hubungkan kembali operator tersebut dengan connector.



22. Atur konfigurasi **Distinct** dengan klik kanan Distinct, pilih Configure.
23. Pilih field **PRODUK** dan **HARGA** dalam jendela konfigurasi Distinct. Klik OK.



24. Untuk memastikan kembali bahwa proses Distinct dan Transformasi berhasil dilakukan sebelum data di-*load* ke tabel tujuan, maka jalankan operator Transform kembali. Klik kanan Transform, pilih Run and Preview Results.

No.	kode_produk	nama_produk	nama_kategori	nama_subkategori	nama_pola	harga
1	PRO-EC0D	Bahan Standar Cap Lasem	Bahan	Standar	Cap	120000
2	PRO-28F2	Hem Standar Cap Tumpal	Hem	Standar	Cap	100000
3	PRO-3CE2	Hem Sutra Print Rama	Hem	Sutra	Print	100000
4	PRO-1C00	Batik Standar Cap Tumpal	Batik	Standar	Cap	150000
5	PRO-7A3A	Rok Batik Print Kombinasi	Rok	Batik	Print	225000
6	PRO-4F1F	Kaos Batik Cap Lukis	Kaos	Batik	Cap	30000
7	PRO-6291	Jarik Standar Tulis Sarimbit	Jarik	Standar	Tulis	40000
8	PRO-B146	Hem Standar Tulis Madura	Hem	Standar	Tulis	550000
9	PRO-7DD3	Celana Standar Cap Warna	Celana	Standar	Cap	55000
10	PRO-D5A2	Bahan Lawasan Tulis Tolet	Bahan	Lawasan	Tulis	130000
11	PRO-7A1F	Kaos Katun Print Bola	Kaos	Katun	Print	60000
12	PRO-C251	Jarik Standar Print Sogan	Jarik	Standar	Print	225000
13	PRO-3747	Jam Standar Print Lukis	Jam	Standar	Print	80000
14	PRO-C576	Bahan Beludru Cap Mahkota	Bahan	Beludru	Cap	500000
15	PRO-DB09	Hem Katun Print Kelengan	Hem	Katun	Print	299000
16	PRO-867C	Sarimbit Standar Print Lukis	Sarimbit	Standar	Print	150000
17	PRO-CFB6	Bahan Standar Cap Garis	Bahan	Standar	Cap	135000
18	PRO-D927	Bolero Standar Cap Sidomukti	Bolero	Standar	Cap	225000
19	PRO-D189	Hem Katun Print Kawung	Hem	Katun	Print	70000
20	PRO-866F	Celana Standar Print Lasem	Celana	Standar	Print	55000

25. Tabel yang sama akan ditampilkan kembali. Jika dalam tabel terdapat nama produk yang sama, maka hanya akan ditampilkan 1 kali sehingga jumlah data akan semakin sedikit.
26. Data produk siap di-*load* ke dalam tabel tujuan pada file Excel Sheet TabelProduk. Pastikan semua file Excel telah ditutup.
27. Pada editor Apatar, klik kanan **TabelProduk** pilih Run and Preview Results. Tunggu beberapa saat hingga tabel hasil *loading* ditampilkan.
28. Jika tabel hasil ditampilkan, maka proses *loading* telah selesai dilakukan. Untuk memastikan bahwa proses ETL secara keseluruhan berhasil dilakukan, buka kembali file “**Data_penjualan.xls**” pada Sheet **TabelProduk**.

A	B	C	D	E	F	
1	kode_produk	nama_produk	nama_kategori	nama_subkategori	nama_pola	harga
2	PRO-4CA5	Bahan Standar Cap Lasem	Bahan	Standar	Cap	120000
3	PRO-B97C	Hem Standar Cap Tumpal	Hem	Standar	Cap	100000
4	PRO-A846	Hem Sutra Print Rama	Hem	Sutra	Print	100000
5	PRO-3605	Batik Standar Cap Tumpal	Batik	Standar	Cap	150000
6	PRO-10F3	Rok Batik Print Kombinasi	Rok	Batik	Print	225000
7	PRO-55C3	Kaos Batik Cap Lukis	Kaos	Batik	Cap	30000
8	PRO-75CD	Jarik Standar Tulis Sarimbit	Jarik	Standar	Tulis	40000
9	PRO-9CD8	Hem Standar Tulis Madura	Hem	Standar	Tulis	550000
10	PRO-7438	Celana Standar Cap Warna	Celana	Standar	Cap	55000
11	PRO-5762	Bahan Lawasan Tulis Tolet	Bahan	Lawasan	Tulis	130000
12	PRO-B5B4	Kaos Katun Print Bola	Kaos	Katun	Print	60000
13	PRO-CFA0	Jarik Standar Print Sogan	Jarik	Standar	Print	225000
14	PRO-8CB6	Jam Standar Print Lukis	Jam	Standar	Print	80000
15	PRO-35C4	Bahan Beludru Cap Mahkota	Bahan	Beludru	Cap	500000
16	PRO-3C65	Hem Katun Print Kelengen	Hem	Katun	Print	299000
17	PRO-36A1	Sarimbit Standar Print Lukis	Sarimbit	Standar	Print	150000
18	PRO-2990	Bahan Standar Cap Garis	Bahan	Standar	Cap	135000
19	PRO-0202	Bolero Standar Cap Sidomukti	Bolero	Standar	Cap	225000
20	PRO-1E30	Hem Katun Print Kawung	Hem	Katun	Print	70000
21	PRO-5556	Celana Standar Print Lasem	Celana	Standar	Print	55000

29. Data yang sama akan ditampilkan dalam file Excel. Kode_produk bisa berbeda-beda karena dibuat secara random. Dengan demikian proses ETL untuk Produk telah berhasil dilakukan.

E. Tugas

Dengan menggunakan file Excel "**Data_penjualan.xls**", selesaikan tugas berikut di kelas. Jika belum selesai, bisa dilanjutkan di rumah dan akan dinilai pada pertemuan berikutnya.

1. Buat Dimensi Kategori yang terdiri dari 2 buah kolom yaitu **kode_kategori**, dan **nama_kategori**. Data kode_kategori dibuat terdiri atas 6 karakter yaitu 2 karakter pertama sebagai konstanta dan 4 karakter berikutnya secara random. Contoh kode_kategori = "**K-R2GT**" yang menunjukkan K = Kategori. **Nama_kategori** diambil dari Data_Tujuan kolom PRODUK pada kata pertama nama produk dan tidak diijinkan terjadi duplikasi data nama kategori. Contoh nama_kategori = "Celana", "Hem", "Kaos" dan lain-lain.

	A	B
1	kode_kategori	nama_kategori
2	K-5F6C	Jarik
3	K-306B	Rok
4	K-057F	Batik
5	K-86DE	Celana
6	K-19AA	Kaos
7	K-769F	Jam
8	K-B2A2	Sarimbit
9	K-27D8	Hem
10	K-2982	Bahan
11	K-062B	Bolero

DimensiKategori

Ready 100%

2. Buat Dimensi Sub Kategori yang terdiri dari 2 buah kolom yaitu **kode_subkategori**, dan **nama_subkategori**. Data **kode_subkategori** dibuat terdiri atas 6 karakter yaitu 2 karakter pertama sebagai konstanta dan 4 karakter berikutnya secara random. Contoh **kode_subkategori** = "S-G76V" yang menunjukkan S = Sub Kategori. **Nama_subkategori** diambil dari Data_Tujuan kolom PRODUK pada kata kedua nama produk dan tidak diijinkan terjadi duplikasi data nama sub kategori. Contoh **nama_subkategori** = "Beludru", "Katun", "Sutra" dan lain-lain.

	A	B
1	kode_subkategori	nama_subkategori
2	S-AA60	Sutra
3	S-5F3D	Batik
4	S-379A	Standar
5	S-9885	Lawasan
6	S-4251	Beludru
7	S-B6C8	Katun

DimensiSubKategori

Ready 100%

Contoh Nama Produk = "**Kaos Katun Print Bola**", maka menunjukkan bahwa:

- Nama Kategori = "Kaos"
- Nama Sub Kategori = "Katun"
- Nama Pola = "Print"

MODUL 4

PROSES EXTRACT-TRANSFORM-LOAD

(PEMBUATAN TABEL FAKTA)

A. Tujuan

1. Mahasiswa mampu membangun tabel fakta yang menjadi tabel pusat transaksi dalam sebuah *data warehouse*.
2. Mahasiswa mampu melakukan proses ETL secara lebih lanjut pada pengembangan sebuah *Data Warehouse*.

B. Landasan Teori

Fact table (tabel fakta) adalah tabel yang umumnya mengandung sesuatu yang dapat diukur (*measure*), seperti harga, jumlah barang, dan sebagainya. *Fact table* juga merupakan kumpulan *foreign key* dari *primary key* yang terdapat pada masing-masing *dimension table*. *Fact table* juga mengandung data yang historis.

C. Alat dan Bahan

1. Komputer dengan sistem operasi Windows.
2. Program aplikasi Ms. Office
3. Program aplikasi Apatar Tool.
4. Modul Praktikum Data Warehousing dan Data Mining.

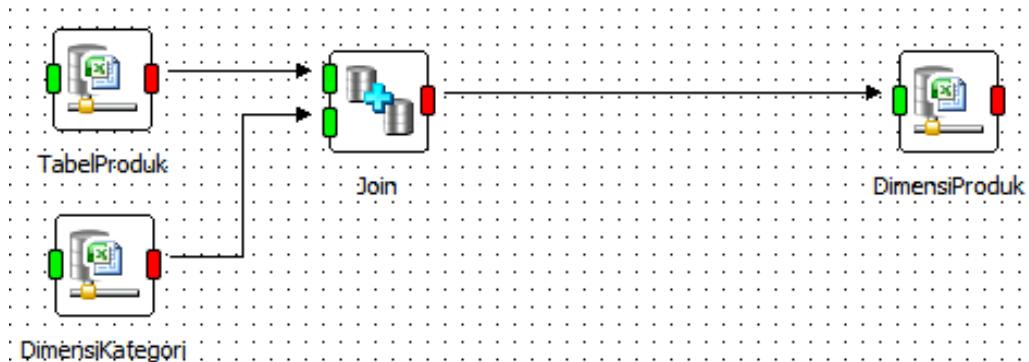
D. Langkah-langkah Praktikum

1) Dimensi Produk

Setelah data produk berhasil dipisah menjadi beberapa bagian yang terdiri dari kode produk, nama produk, kategori, sub kategori dan pola (pada Modul 3), langkah berikutnya adalah membuat **Dimensi Produk**. Dimensi ini dibuat untuk menggantikan nilai data kategori, sub kategori dan pola menjadi suatu *Foreign Key* yang diambil dari kode-kode data Dimensi Kategori, Dimensi Sub Kategori dan Dimensi Pola yang telah dikerjakan pada Modul 2 dan Modul 3.

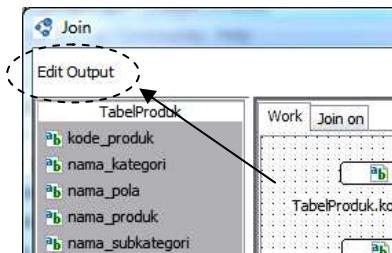
Adapun langkah-langkah membuat Dimensi Produk sebagai berikut:

1. Buka kembali file Excel “**Data_penjualan.xls**”.
2. Buat Sheet baru. Ubah nama Sheet menjadi “**DimensiProduk**”.
3. Buat 6 buah kolom yaitu **kode_produk**, **nama_produk**, **kode_kategori**, **kode_subkategori**, **kode_pola** dan **harga**. Simpan file excel, dan tutup kembali.
4. Buka kembali file apatar “**Transform_Dimensi.aptr**”.
5. Tambahkan 2 buah connector “**MS Excel**” di bawah operasi pembuatan Tabel Produk.
6. Dengan menggunakan file “**Data_penjualan.xls**”, atur konfigurasi MS Excel yang pertama dengan mengambil Sheet “**TabelProduk**” yang terdiri dari 6 kolom, sedangkan MS Excel yang kedua mengambil Sheet “**DimensiProduk**” yang juga terdiri dari 6 kolom.
7. Tambahkan sebuah connector MS Excel lagi dan letakkan di bawah connector TabelProduk. Atur konfigurasi MS Excel ini dengan mengambil Sheet “**DimensiKategori**” yang terdiri dari 2 kolom (Dimensi ini merupakan hasil tugas pada Modul 3). Tambahkan pula sebuah operator “**Join**”, letakkan di antara TabelProduk dan DimensiProduk. Hubungkan 3 connector tersebut dengan operator Join seperti gambar berikut.

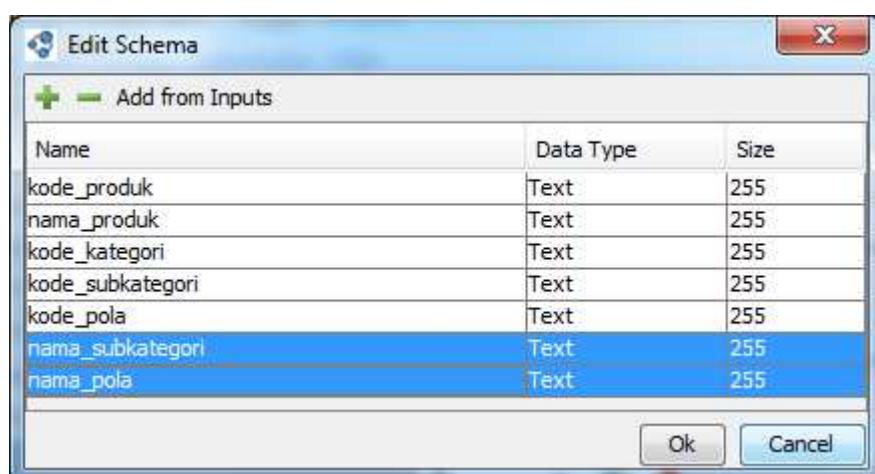


8. Atur konfigurasi operator Join dengan klik kanan **Join**, pilih **Configure**.
9. Masukkan 5 buah field dari **Input Table 1** (TabelProduk) ke dalam editor, yaitu **kode_produk**, **nama_produk**, **nama_subkategori**, **nama_pola** dan **harga**. Masukkan pula 3 field dari output (sebelah kanan) ke dalam editor yang terdiri dari **kode_produk**, **nama_produk** dan **harga**.

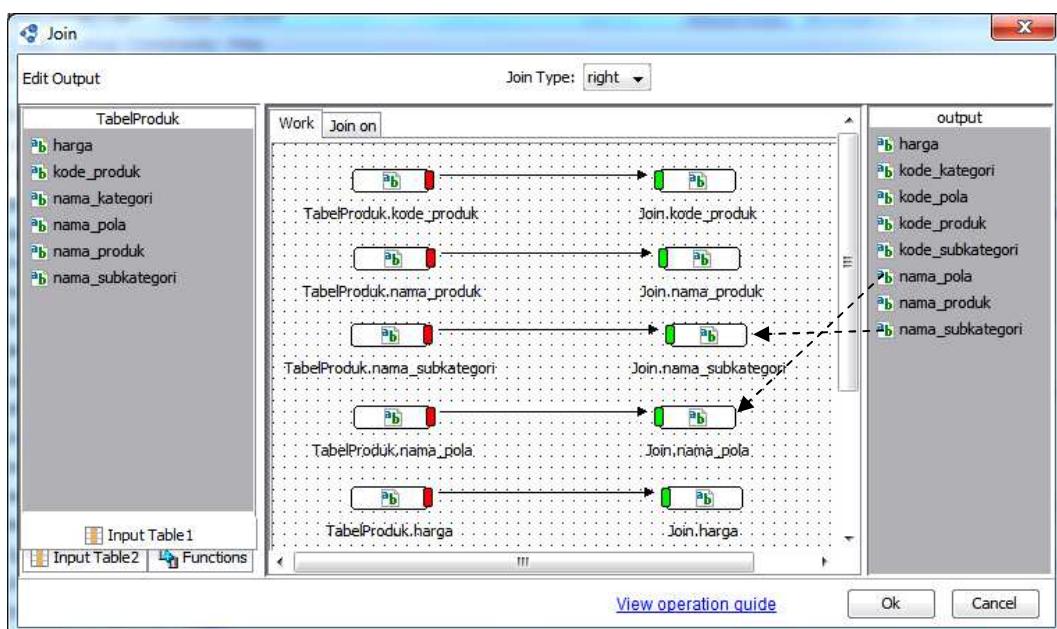
10. Untuk menambahkan field **nama_subkategori** dan **nama_pola** dari output, klik **Edit Output** yang terletak di sebelah kiri atas jendela konfigurasi.



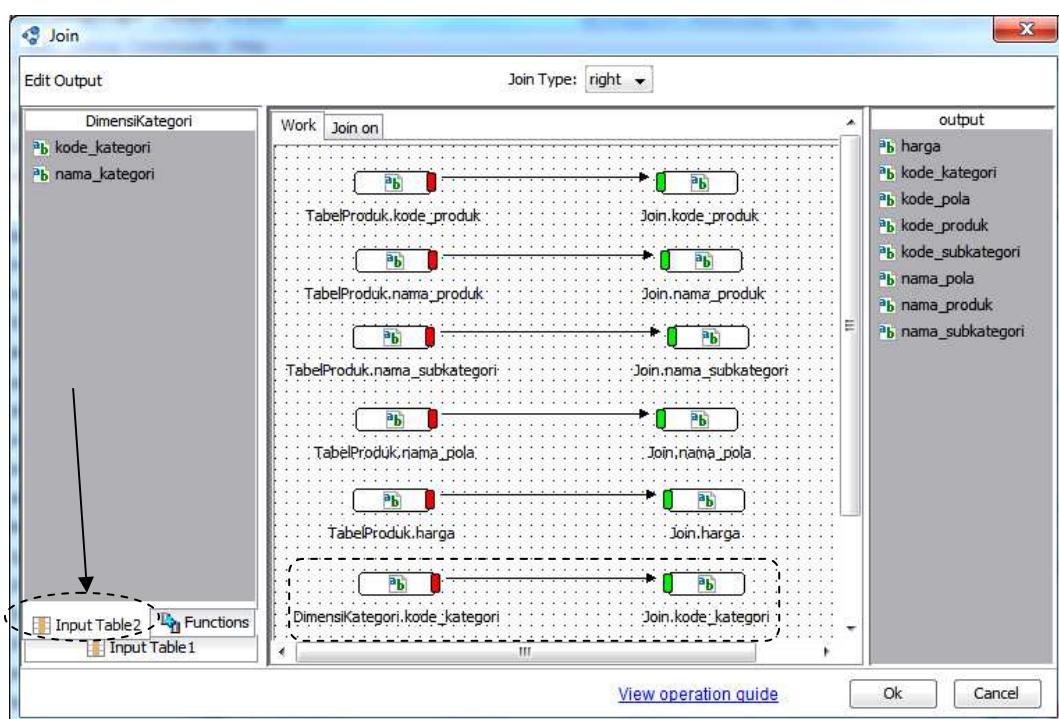
11. Tambahkan 2 buah field paling bawah dengan cara klik tanda plus (+), dan isikan nama field (**nama_subkategori** dan **nama_pola**) dan tipe data yang dibutuhkan. Jika sudah, klik OK.



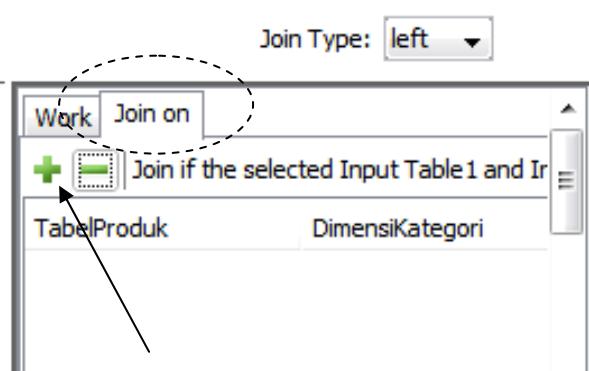
12. Masukkan 2 field tambahan tersebut ke dalam editor.



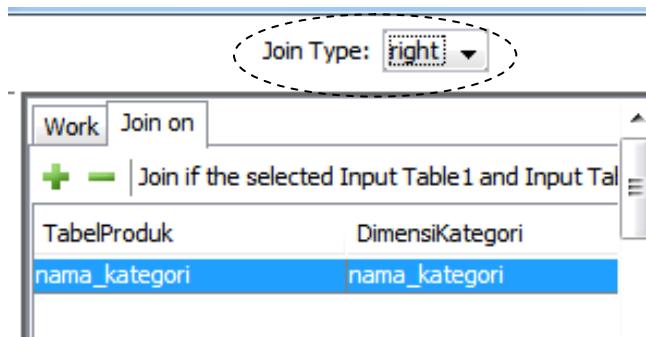
13. Hubungkan semua field input dengan outputnya sesuai dengan nama fieldnya.
14. Field **nama_kategori** tidak dimasukkan karena ini akan diganti dengan **kode_kategori** yang diambil Dimensi Kategori. Untuk itu buka **Input Table 2** (Dimensi Kategori), masukkan **kode_kategori** ke dalam editor konfigurasi Join.
15. Masukkan pula field **kode_kategori** dari output (sebelah kanan) ke dalam editor. Hubungkan kedua field tersebut secara langsung.



16. Langkah berikutnya adalah mengatur hubungan antara TabelProduk dengan Dimensi Kategori. Dalam hal ini karena akan diambil kode kategori dari Dimensi Kategori sebagai pengganti nama kategori.
17. Masih di jendela konfigurasi Join, klik tab **Join On** (di sebelah atas editor).



18. Klik simbol plus (+) di sebelah kiri atas tabel untuk menambahkan kondisi hubungan kedua tabel antara TabelProduk dan DimensiKategori. Pada kolom TabelProduk pilih nama_kategori, dan pada kolom DimensiProduk juga dipilih nama_kategori. Field ini dipilih karena data yang sama pada kedua tabel menunjukkan nama kategori yang sama, sehingga kode_kategori bisa diambil. Dengan prinsip hubungan yang terbalik dalam *database*, maka tipe join diatur dengan **Join Type = Right**.

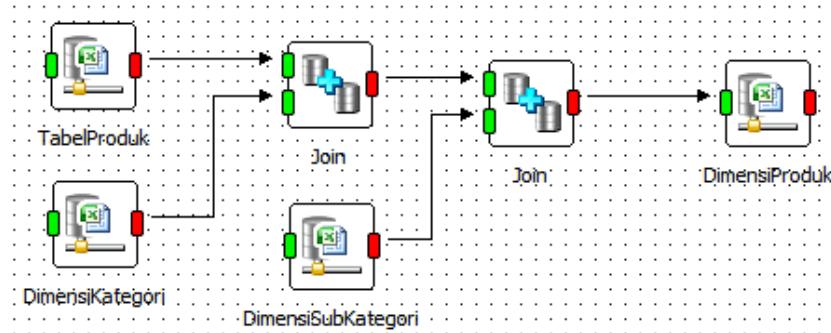


19. Klik OK untuk menutup konfigurasi Join.
20. Untuk melihat sementara hasil Join TabelProduk dengan DimensiKategori, klik kanan operator **Join**, pilih Run and Preview Results.

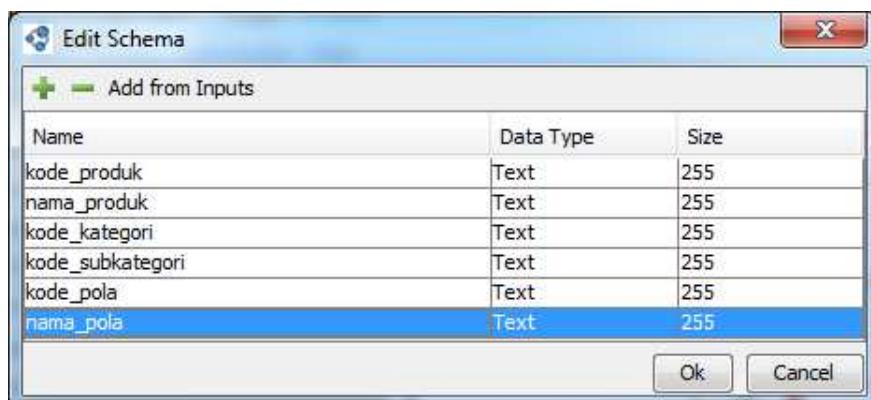
No.	kode_produk	nama_produk	kode_kategori	kode_subkategori	kode_pola	harga	nama_subkategori	nama_pola
1	PRO-75CD	Jarik Standar Tulis Sarimbit	K-5F6C			40000	Standar	Tulis
2	PRO-CFA0	Jarik Standar Print Sogan	K-5F6C			225000	Standar	Print
3	PRO-10F3	Rok Batik Print Kombinasi	K-306B			225000	Batik	Print
4	PRO-3605	Batik Standar Cap Tumpal	K-057F			150000	Standar	Cap
5	PRO-7438	Celana Standar Cap Warna	K-86DE			55000	Standar	Cap
6	PRO-5556	Celana Standar Print Lasem	K-86DE			55000	Standar	Print
7	PRO-55C3	Kaos Batik Cap Lukis	K-19AA			30000	Batik	Cap
8	PRO-B5B4	Kaos Katun Print Bola	K-19AA			60000	Katun	Print
9	PRO-8CB6	Jam Standar Print Lukis	K-769F			80000	Standar	Print
10	PRO-36A1	Sarimbit Standar Print Lukis	K-B2A2			150000	Standar	Print
11	PRO-B97C	Hem Standar Cap Tumpal	K-27D8			100000	Standar	Cap
12	PRO-A846	Hem Sutra Print Rama	K-27D8			100000	Sutra	Print
13	PRO-9CD8	Hem Standar Tulis Madura	K-27D8			550000	Standar	Tulis
14	PRO-3C65	Hem Katun Print Kelenggan	K-27D8			299000	Katun	Print
15	PRO-1E30	Hem Katun Print Kawung	K-27D8			70000	Katun	Print
16	PRO-4CA5	Bahan Standar Cap Lasem	K-2982			120000	Standar	Cap
17	PRO-5762	Bahan Lawasan Tulis Tolet	K-2982			130000	Lawasan	Tulis
18	PRO-35C4	Bahan Beludru Cap Mahkota	K-2982			500000	Beludru	Cap
19	PRO-2990	Bahan Standar Cap Garis	K-2982			135000	Standar	Cap
20	PRO-0202	Bolero Standar Cap Sidomukti	K-062B			225000	Standar	Cap

21. Jika **kode_kategori** sudah masuk ke dalam tabel untuk menggantikan **nama_kategori**, maka proses Join ini berhasil dilakukan. Sehingga selanjutnya adalah proses Join untuk memasukkan **kode_subkategori**.
22. Kembali ke editor Apatar proses Join sebelumnya. Hapus hubungan antara Join dengan DimensiProduk.

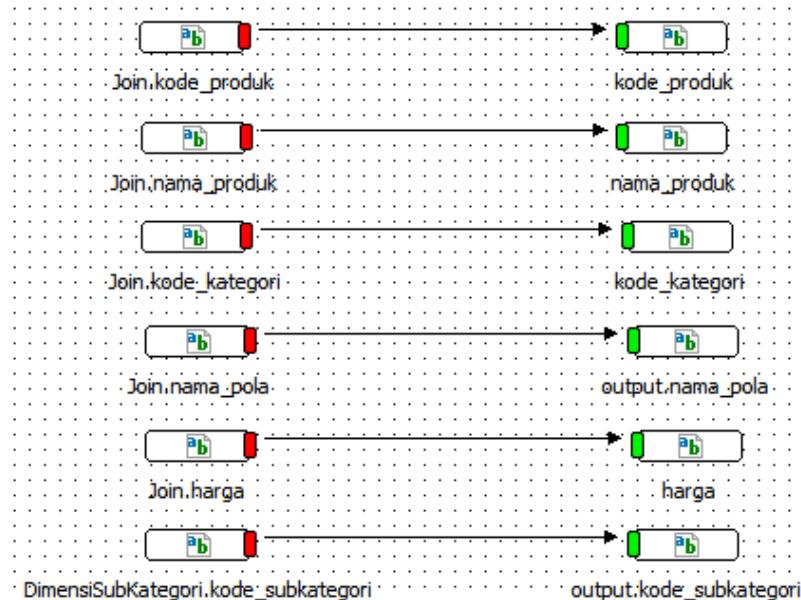
23. Tambahkan sebuah operator Join, letakkan di antara operator Join yang pertama dan DimensiProduk. Tambahkan pula connector MS Excel. Atur MS Excel dengan mengambil Sheet “**DimensiSubKategori**” yang terdiri dari 2 kolom (hasil tugas Modul 3).
24. Hubungkan connector DimensiSubKategori dengan operator Join serta hubungkan dengan DimensiProduk seperti gambar berikut.



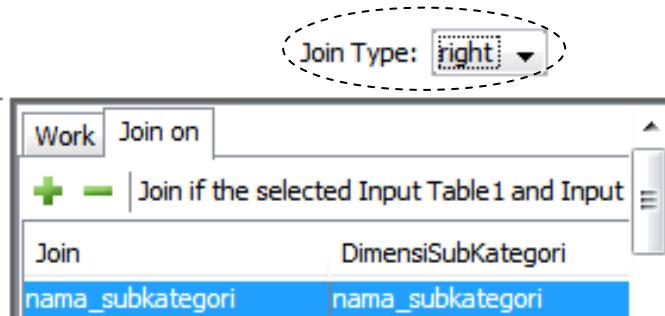
25. Atur konfigurasi operator Join yang kedua dengan klik kanan **Join**, pilih **Configure**.
26. Masukkan 5 buah field dari **Input Table 1** (Join) ke dalam editor, yaitu **kode_produk**, **nama_produk**, **kode_kategori**, **nama_pola** dan **harga**. Masukkan pula 4 field dari output (sebelah kanan) ke dalam editor yang terdiri dari **kode_produk**, **nama_produk**, **kode_kategori** dan **harga**.
27. Masukkan field **kode_subkategori** dari **Input Table 2** (DimensiSubKategori) ke dalam editor.
28. Untuk menambahkan field **nama_pola** dari output, klik **Edit Output** yang terletak di sebelah kiri atas jendela konfigurasi.
29. Tambahkan sebuah field paling bawah dengan cara klik tanda plus (+), dan isikan nama field (**nama_pola**) dan tipe data yang dibutuhkan. Klik OK.



30. Masukkan field tambahan tersebut ke dalam editor.



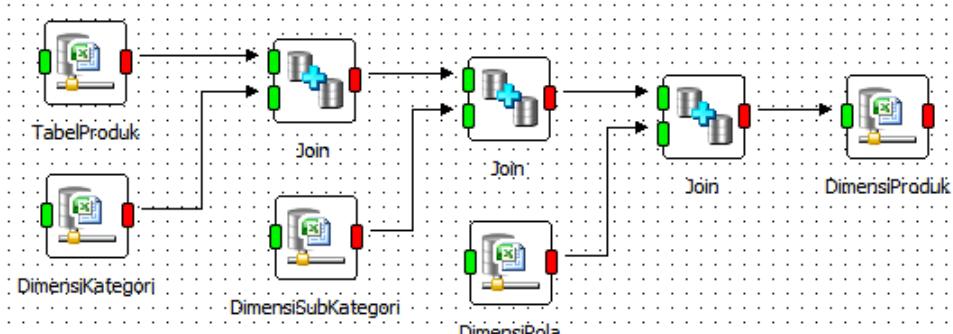
31. Hubungkan semua field input dengan outputnya sesuai dengan nama fieldnya.
32. Langkah berikutnya adalah mengatur hubungan antara TabelProduk dengan Dimensi Sub Kategori. Dalam hal ini karena akan diambil kode sub kategori dari Dimensi Sub Kategori sebagai pengganti nama subkategori.
33. Masih di jendela konfigurasi Join, klik tab Join On (di sebelah atas editor).
34. Klik simbol plus (+) di sebelah kiri atas tabel untuk menambahkan kondisi hubungan kedua tabel antara TabelProduk dan DimensiSubKategori. Pada kolom TabelProduk (Join) pilih **nama_subkategori**, dan pada kolom DimensiSubKategori juga dipilih **nama_subkategori**. Field ini dipilih karena data yang sama pada kedua tabel menunjukkan nama sub kategori yang sama, sehingga kode_subkategori bisa diambil. Dengan prinsip hubungan yang terbalik dalam *database*, maka tipe join diatur dengan **Join Type = Right**.



35. Klik OK untuk menutup konfigurasi Join.
36. Untuk melihat sementara hasil Join TabelProduk dengan DimensiSubKategori, klik kanan operator **Join**, pilih Run and Preview Results.

No.	kode_produk	nama_produk	kode_kategori	kode_subkategori	kode_pola	harga	nama_pola
1	PRO-A846	Hem Sutra Print Rama	K-27D8	S-AA60		100000	Print
2	PRO-10F3	Rok Batik Print Kombinasi	K-306B	S-5F3D		225000	Print
3	PRO-55C3	Kaos Batik Cap Lukis	K-19AA	S-5F3D		30000	Cap
4	PRO-75CD	Jarik Standar Tulis Sarimbit	K-5F6C	S-379A		40000	Tulis
5	PRO-CFA0	Jarik Standar Print Sogan	K-5F6C	S-379A		225000	Print
6	PRO-3605	Batik Standar Cap Tumpal	K-057F	S-379A		150000	Cap
7	PRO-7438	Celana Standar Cap Warna	K-86DE	S-379A		55000	Cap
8	PRO-5556	Celana Standar Print Lasem	K-86DE	S-379A		55000	Print
9	PRO-8CB6	Jam Standar Print Lukis	K-769F	S-379A		80000	Print
10	PRO-36A1	Sarimbit Standar Print Lukis	K-B2A2	S-379A		150000	Print
11	PRO-B97C	Hem Standar Cap Tumpal	K-27D8	S-379A		100000	Cap
12	PRO-9CD8	Hem Standar Tulis Madura	K-27D8	S-379A		550000	Tulis
13	PRO-4CA5	Bahan Standar Cap Lasem	K-2982	S-379A		120000	Cap
14	PRO-2990	Bahan Standar Cap Garis	K-2982	S-379A		135000	Cap
15	PRO-0202	Bolero Standar Cap Sidomukti	K-062B	S-379A		225000	Cap
16	PRO-5762	Bahan Lawasan Tulis Tolet	K-2982	S-9885		130000	Tulis
17	PRO-35C4	Bahan Beludru Cap Mahkota	K-2982	S-4251		500000	Cap
18	PRO-B5B4	Kaos Katun Print Bola	K-19AA	S-B6C8		60000	Print
19	PRO-3C65	Hem Katun Print Kelenggan	K-27D8	S-B6C8		299000	Print
20	PRO-1E30	Hem Katun Print Kawung	K-27D8	S-B6C8		70000	Print

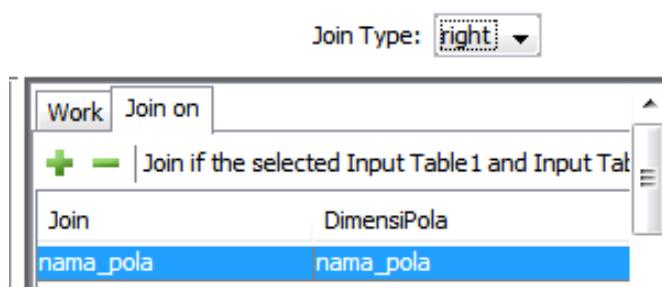
37. Jika **kode_subkategori** sudah masuk ke dalam tabel untuk menggantikan **nama_subkategori**, maka proses Join ini berhasil dilakukan. Sehingga selanjutnya adalah proses Join untuk memasukkan **kode_pola**.
38. Kembali ke editor Apatar proses Join sebelumnya. Hapus hubungan antara Join kedua dengan DimensiProduk.
39. Tambahkan sebuah operator Join, letakkan di antara operator Join yang kedua dan DimensiProduk. Tambahkan pula connector MS Excel. Atur MS Excel dengan mengambil Sheet “**DimensiPola**” yang terdiri dari 2 kolom (praktikum Modul 2).
40. Hubungkan connector DimensiPola dengan operator Join serta hubungkan dengan DimensiProduk seperti gambar berikut.



41. Atur konfigurasi operator Join yang ketiga dengan klik kanan **Join**, pilih **Configure**.
42. Pada jendela konfigurasi Join, masukkan 4 buah field dari **Input Table 1** (Join) ke dalam editor yang terdiri dari **kode_produk**, **nama_produk**, **kode_kategori**, **kode_subkategori** dan **harga**.
43. Masukkan 1 buah field dari **Input Table 2** (DimensiPola) yaitu field **kode_pola** ke dalam editor.
44. Masukkan pula semua field (6 field) dari Output (sebelah kanan) ke dalam editor konfigurasi. Hubungkan semua field dari input ke output sesuai dengan nama fieldnya.



45. Atur pula konfigurasi **Join On** untuk kondisi **nama_pola**. Pilih field **nama_pola** pada kedua tabel, dan atur **Join Type = Right**.



46. Klik OK untuk menutup jendela konfigurasi Join.
47. Untuk melihat sementara hasil Join TabelProduk dengan DimensiPola, klik kanan operator **Join**, pilih Run and Preview Results.

No.	kode_produk	nama_produk	kode_kategori	kode_subkategori	kode_pola	harga
1	PRO-55C3	Kaos Batik Cap Lukis	K-19AA	S-5F3D	P-BD59	30000
2	PRO-3605	Batik Standar Cap Tumpal	K-057F	S-379A	P-BD59	150000
3	PRO-7438	Celana Standar Cap Warna	K-86DE	S-379A	P-BD59	55000
4	PRO-B97C	Hem Standar Cap Tumpal	K-27D8	S-379A	P-BD59	100000
5	PRO-4CA5	Bahan Standar Cap Lasem	K-2982	S-379A	P-BD59	120000
6	PRO-2990	Bahan Standar Cap Garis	K-2982	S-379A	P-BD59	135000
7	PRO-0202	Bolero Standar Cap Sidomukti	K-062B	S-379A	P-BD59	225000
8	PRO-35C4	Bahan Beludru Cap Mahkota	K-2982	S-4251	P-BD59	500000
9	PRO-75CD	Jarik Standar Tulis Sarimbit	K-5F6C	S-379A	P-C3BC	40000
10	PRO-9CD8	Hem Standar Tulis Madura	K-27D8	S-379A	P-C3BC	550000
11	PRO-5762	Bahan Lawasan Tulis Tolet	K-2982	S-9885	P-C3BC	130000
12	PRO-A846	Hem Sutra Print Rama	K-27D8	S-AA60	P-B59A	100000
13	PRO-10F3	Rok Batik Print Kombinasi	K-306B	S-5F3D	P-B59A	225000
14	PRO-CFA0	Jarik Standar Print Sogan	K-5F6C	S-379A	P-B59A	225000
15	PRO-5556	Celana Standar Print Lasem	K-86DE	S-379A	P-B59A	55000
16	PRO-8CB6	Jam Standar Print Lukis	K-769F	S-379A	P-B59A	80000
17	PRO-36A1	Sarimbit Standar Print Lukis	K-B2A2	S-379A	P-B59A	150000
18	PRO-B5B4	Kaos Katun Print Bola	K-19AA	S-B6C8	P-B59A	60000
19	PRO-3C65	Hem Katun Print Kelengan	K-27D8	S-B6C8	P-B59A	299000
20	PRO-1E30	Hem Katun Print Kawung	K-27D8	S-B6C8	P-B59A	70000

48. Jika **kode_pola** sudah masuk ke dalam tabel untuk menggantikan **nama_pola**, maka proses Join ini berhasil dilakukan. Sehingga selanjutnya adalah proses *loading* untuk memasukkan data ke dalam DimensiProduk.
49. Pastikan semua file Excel telah ditutup. Klik kanan connector **DimensiProduk**, pilih Run and Preview Results.
50. Untuk memastikan proses *Loading* berhasil dilakukan, buka kembali file “**Data_penjualan.xls**” pada Sheet **DimensiProduk**. Jika isinya sama dengan yang ditampilkan pada hasil proses dengan Apatar, maka ETL Dimensi Produk berhasil dilakukan.

A	B		C	D	E	F
1	kode_produk	nama_produk	kode_kategori	kode_subkategori	kode_pola	harga
2	PRO-55C3	Kaos Batik Cap Lukis	K-19AA	S-5F3D	P-BD59	30000
3	PRO-3605	Batik Standar Cap Tumpal	K-057F	S-379A	P-BD59	150000
4	PRO-7438	Celana Standar Cap Warna	K-86DE	S-379A	P-BD59	55000
5	PRO-B97C	Hem Standar Cap Tumpal	K-27D8	S-379A	P-BD59	100000
6	PRO-4CA5	Bahan Standar Cap Lasem	K-2982	S-379A	P-BD59	120000
7	PRO-2990	Bahan Standar Cap Garis	K-2982	S-379A	P-BD59	135000
8	PRO-0202	Bolero Standar Cap Sidomukti	K-062B	S-379A	P-BD59	225000
9	PRO-35C4	Bahan Beludru Cap Mahkota	K-2982	S-4251	P-BD59	500000
10	PRO-75CD	Jarik Standar Tulis Sarimbit	K-5F6C	S-379A	P-C3BC	40000
11	PRO-9CD8	Hem Standar Tulis Madura	K-27D8	S-379A	P-C3BC	550000
12	PRO-5762	Bahan Lawasan Tulis Tolet	K-2982	S-9885	P-C3BC	130000
13	PRO-A846	Hem Sutra Print Rama	K-27D8	S-AA60	P-B59A	100000
14	PRO-10F3	Rok Batik Print Kombinasi	K-306B	S-5F3D	P-B59A	225000
15	PRO-CFA0	Jarik Standar Print Sogan	K-5F6C	S-379A	P-B59A	225000
16	PRO-5556	Celana Standar Print Lasem	K-86DE	S-379A	P-B59A	55000
17	PRO-8CB6	Jam Standar Print Lukis	K-769F	S-379A	P-B59A	80000
18	PRO-36A1	Sarimbit Standar Print Lukis	K-B2A2	S-379A	P-B59A	150000
19	PRO-B5B4	Kaos Katun Print Bola	K-19AA	S-B6C8	P-B59A	60000
20	PRO-3C65	Hem Katun Print Kelengan	K-27D8	S-B6C8	P-B59A	299000
21	PRO-1E30	Hem Katun Print Kawung	K-27D8	S-B6C8	P-B59A	70000

2) Fakta Penjualan

Setelah berhasil membuat **Dimensi Produk**, langkah berikutnya adalah membuat Fakta Penjualan. Secara prinsip, tabel fakta hanya berisi kumpulan *foreign key* yang diambil dari *primary key* semua dimensi yang terhubung, ditambah field-field *measure*.

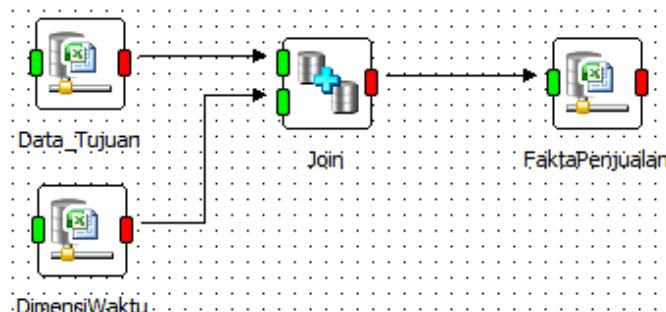
Tabel ini dibuat untuk menggantikan nilai data tanggal, nama produk, nama pelanggan dan nama wilayah menjadi kode-kode sebagai *Foreign Key* yang diambil dari kode-kode data Dimensi Waktu, Dimensi Produk, Dimensi Pelanggan dan Dimensi Wilayah yang telah dikerjakan pada Modul 2 dan Modul 4 tahap 1.

Adapun langkah-langkah membuat Fakta Penjualan sebagai berikut:

1. Buka kembali file Excel “**Data_penjualan.xls**”.
2. Buat Sheet baru dan ubah namanya menjadi “**FaktaPenjualan**”.
3. Buat 6 buah kolom yang terdiri dari **kode_fakta**, **kode_waktu**, **kode_produk**, **kode_pelanggan**, **kode_wilayah**, dan **jumlah**.

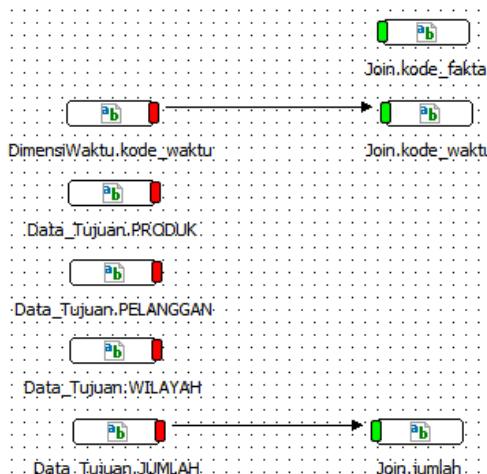
	A	B	C	D	E	F
1	kode_fakta	kode_waktu	kode_produk	kode_pelanggan	kode_wilayah	jumlah
2						

4. Simpan file Excel tersebut dan tutup kembali.
5. Buka kembali file “**Transform_Dimensi.aptr**”.
6. Pada editor, masukkan 3 buah connector **MS Excel** dan sebuah operator **Join**.
7. Dengan menggunakan file “**Data_penjualan.xls**”, MS Excel yang pertama diatur untuk menggunakan Sheet “**Data_Tujuan**” dengan 6 kolom, MS Excel kedua menggunakan Sheet “**DimensiWaktu**” terdiri dari 5 kolom, dan MS Excel yang ketiga menggunakan Sheet “**FaktaPenjualan**” yang terdiri dari 6 kolom.
8. Hubungkan ketiga connector tersebut dengan operator Join.

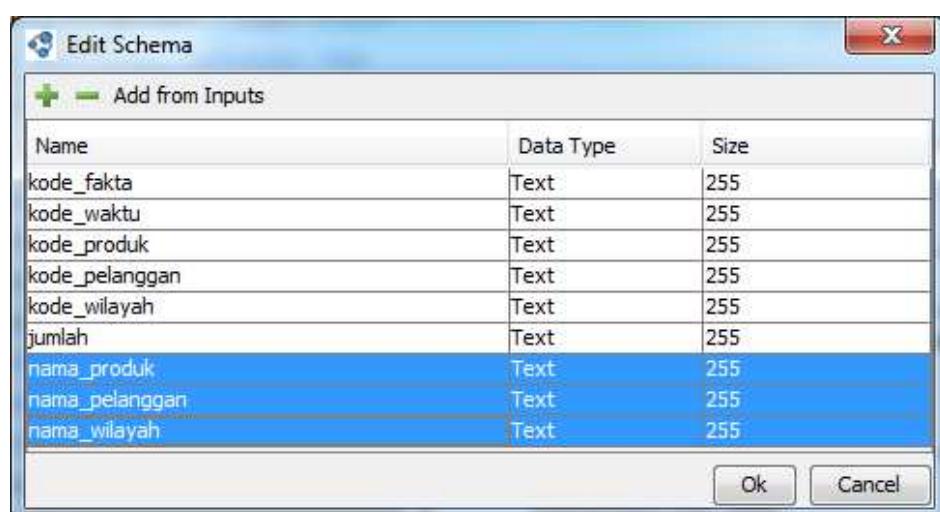


9. Konfigurasikan operator Join dengan cara klik kanan Join, pilih Configure.

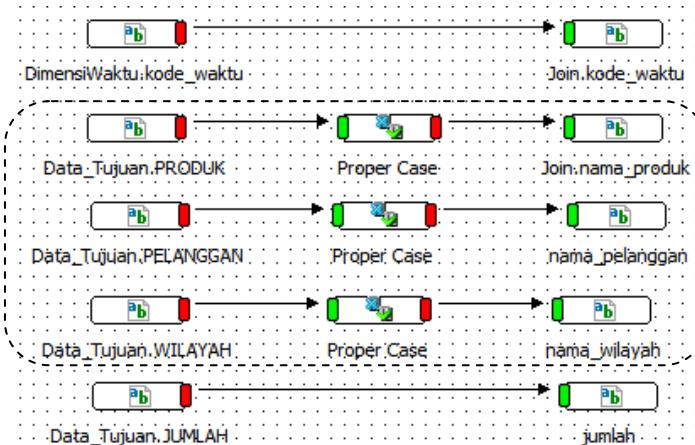
10. Pada editor jendela konfigurasi Join, masukkan 4 field dari **Input Table 1** (Data_Tujuan) yaitu **PRODUK**, **PELANGGAN**, **WILAYAH**, dan **JUMLAH**. Dari **Input Table 2** (DimensiWaktu), masukkan 1 field **kode_waktu**.
11. Masukkan pula 3 buah field dari Output yang terdiri dari **kode_fakta**, **kode_waktu**, dan **jumlah**.
12. Hubungkan field-field yang sudah sesuai untuk diproses.



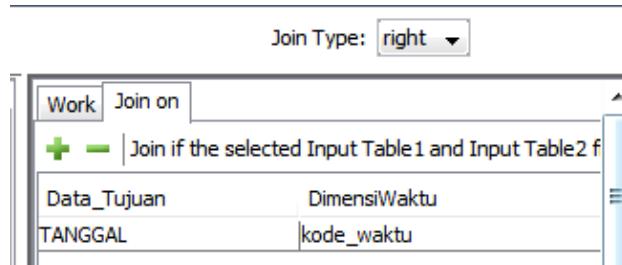
13. Sedangkan untuk field PRODUK, PELANGGAN dan WILAYAH akan dihubungkan dengan field virtual. Sehingga perlu dibuatkan terlebih dahulu 3 field virtual ini yang terdiri dari **nama_produk**, **nama_pelanggan** dan **nama_wilayah**.
14. Klik menu “**Edit Output**” yang terletak di sebelah kiri atas jendela konfigurasi Join. Tambahkan 3 field virtual tersebut dengan cara klik tanda plus (+) dan sesuaikan tipe datanya. Klik OK.



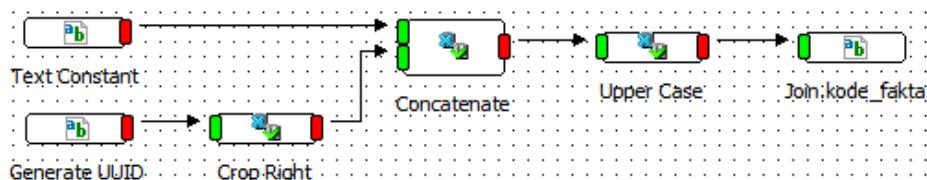
15. 3 buah field baru telah dibuat dan masuk ke dalam field output (sebelah kanan). Masukkan 3 field baru tersebut ke dalam editor, tambahkan 3 buah fungsi **Proper Case** dan hubungkan dengan field input yang sesuai.



16. Atur konfigurasi **Join On** untuk kondisi **kode_waktu**. Isikan field **TANGGAL** pada Data_Tujuan dan **kode_waktu** pada DimensiWaktu sebagai penghubung kedua tabel. Pastikan tipe Join adalah **Right**.



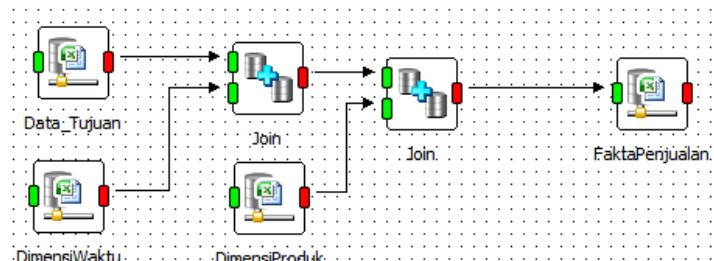
17. Klik OK untuk menutup jendela konfigurasi Join.
 18. Untuk membuat kode fakta, dengan menggunakan cara seperti pembuatan kode-kode sebelumnya. Kode fakta dibuat terdiri atas 10 karakter yang terdiri dari 3 karakter awal sebagai konstanta dan 7 karakter akhir dibuat secara random. Contoh kode fakta = “FP-9TW357C”. FP untuk menunjukkan Fakta Penjualan.
 19. Tambahkan fungsi **Text Constant**, **Generate UUID**, **Crop Right**, **Concatenate**, dan **Upper Case**. Hubungkan semua fungsi tersebut dan hubungkan dengan field kode_fakta.



20. Isikan *value* pada **Text Constant** = ‘FP-‘, dan pada **Crop Right** isikan *value* = 7.
21. Tutup jendela konfigurasi Join untuk memeriksa keberhasilan proses sementara.
22. Klik kanan operator Join, pilih Run and Preview Results.

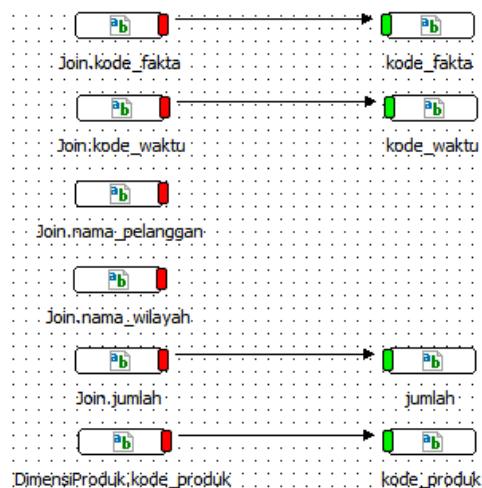
No.	kode_fakta	kode_waktu	kode_produk	kode_pelanggan	kode_wilayah	jumlah	nama_produk	nama_pelan...	nama_wilayah
1	FP-49F3CFF	2012-12-15				1	Rok Batik Print Kombinasi	Ibu Siti Arya	Jawa Barat
2	FP-486A308	2012-05-21				3	Hem Katun Print Kelenggan	Bapak Totok	Jawa Timur
3	FP-8356837	2012-06-22				1	Bahan Lawasan Tulis Tolet	Ibu Niken	Jawa Tengah
4	FP-11B8A94	2010-06-14				1	Bahan Beludru Cap Mahkota	Ibu Tyas	Jawa Tengah
5	FP-9061D9C	2011-01-05				1	Kaos Katun Print Bola	Bapak Imron	Jawa Barat
6	FP-E1AA30F	2012-01-04				14	Kaos Batik Cap Lukis	Ibu Harini	Jawa Timur
7	FP-3FEC27A	2011-12-28				2	Jarik Standar Print Sogan	Bapak Ketut	Bali
8	FP-7A9EC07	2012-01-09				44	Jam Standar Print Lukis	Ibu Siti Arya	Jawa Barat
9	FP-70861D4	2012-02-14				17	Celana Standar Cap Warna	Ibu Hadi Sukarni	Jawa Barat
10	FP-38E2CCE	2011-10-13				1	Sarimbit Standar Print Lukis	Ibu Hatamah	Jawa Timur
11	FP-DC700CC	2012-09-18				1	Batik Standar Cap Tumpal	Bapak Heru	Jawa Timur
12	FP-4CF838D	2010-03-26				17	Celana Standar Print Lasem	Ibu Hadi Sukarni	Jawa Barat
13	FP-035937E	2011-03-27				8	Bahan Standar Cap Lasem	Ibu Siti Arya	Jawa Barat
14	FP-85741D2	2012-04-05				7	Bahan Standar Cap Garis	Ibu Tyas	Jawa Tengah
15	FP-2912EA3	2012-04-05				4	Jarik Standar Tulis Sarimbit	Ibu Harini	Jawa Timur
16	FP-39EEAF3	2012-09-28				1	Hem Standar Cap Tumpal	Ibu Aini Kasmaj	Jawa Tengah
17	FP-22AD8C5	2011-04-09				3	Hem Katun Print Kawung	Ibu Harini	Jawa Timur
18	FP-40908DB	2011-08-19				5	Hem Standar Tulis Madura	Ibu Atik	Jawa Tengah
19	FP-328AECA	2010-11-21				5	Hem Sutra Print Rama	Ibu Tyas	Jawa Tengah
20	FP-3429A21	2011-12-30				1	Bolero Standar Cap Sidomukti	Ibu Hatamah	Jawa Timur

23. Jika 6 field yaitu kode_fakta, kode_waktu, jumlah, nama_produk, nama_pelanggan dan nama_wilayah telah berisi data, maka proses Join telah berhasil. Selanjutnya adalah proses Join dengan Dimensi Produk untuk memasukkan data kode produk ke fakta penjualan.
24. Kembali ke editor Apatar, hapus hubungan antara operator **Join** dengan connector **FaktaPenjualan**.
25. Tambahkan 1 buah operator Join dan 1 connector MS Excel. Atur konfigurasi MS Excel untuk menggunakan file “**Data_penjualan.xls**” pada Sheet “**DimensiProduk**” yang terdiri dari 6 kolom.
26. Hubungkan operator-operator **Join** dengan connector **DimensiProduk** dan **FaktaPenjualan** seperti gambar berikut.

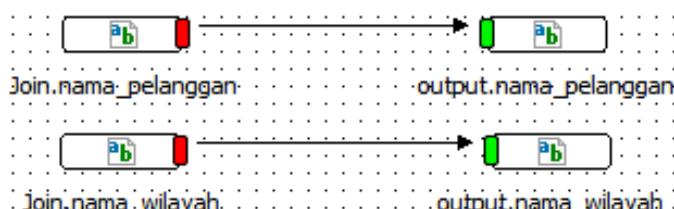


27. Atur konfigurasi **Join** yang kedua.

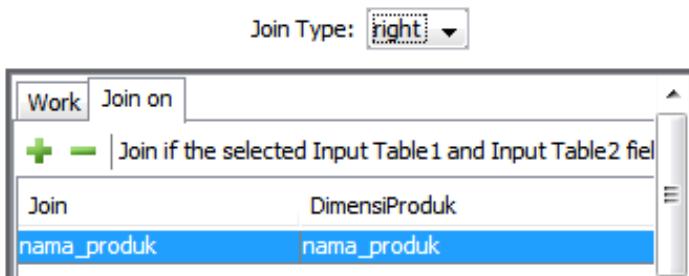
28. Pada editor konfigurasi Join, masukkan 5 field dari **Input Table 1** (Join) yang terdiri dari **kode_fakta**, **kode_waktu**, **nama_pelanggan**, **nama_wilayah**, dan **jumlah**. Masukkan pula 1 buah field dari **Input Table 2** (DimensiProduk) yaitu field **kode_produk**. Sedangkan dari tabel output, masukkan 4 field yaitu **kode_fakta**, **kode_waktu**, **kode_produk** dan **jumlah**. Hubungkan field-field yang sudah sesuai.



29. Untuk **nama_pelanggan** dan **nama_wilayah**, perlu dibuat field virtual terlebih dahulu pada tabel output. Klik “**Edit Output**”, tambahkan 2 field baru **nama_pelanggan** dan **nama_wilayah** dengan tipe data Text. Klik OK.
 30. Masukkan kedua field baru tersebut ke dalam editor dan hubungkan dengan field **nama_pelanggan** dan **nama_wilayah** dari tabel input.



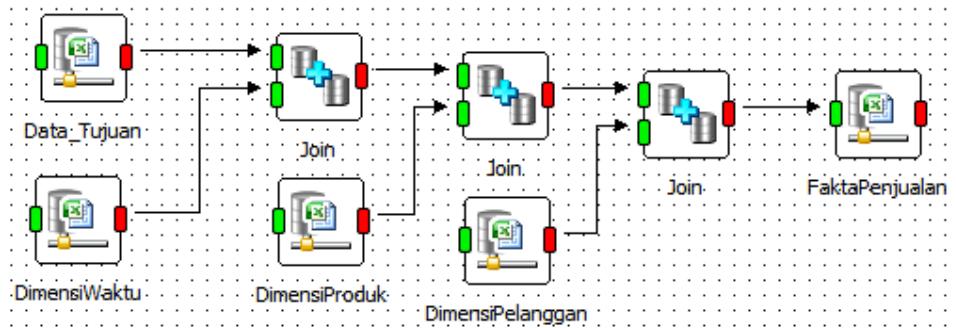
31. Untuk menghubungkan tabel sumber dengan Dimensi Produk, atur konfigurasi **Join On**. Pilih field **nama_produk** pada kolom tabel Join dan DimensiProduk. Atur pula **Join Type = Right**. Klik OK untuk menutup jendela konfigurasi Join.



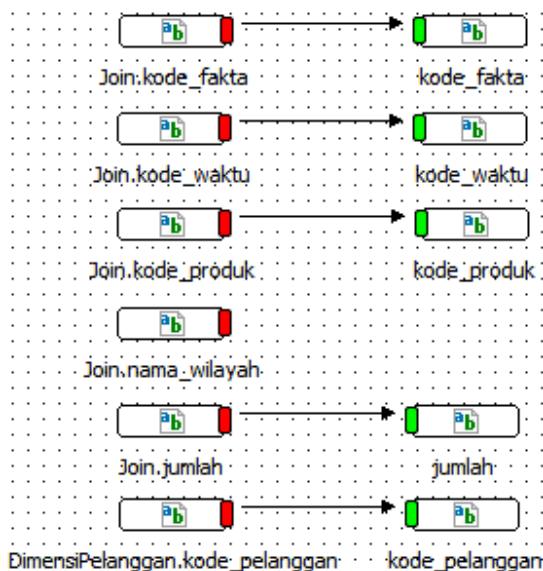
32. Untuk memeriksa keberhasilannya, klik kanan operator Join yang kedua pilih Run and Preview Results.

No.	kode_fakta	kode_waktu	kode_produk	kode_pelanggan	kode_wilayah	jumlah	nama_pelanggan	nama_wilayah
1	FP-48D8896	2012-01-04	PRO-55C3			14	Ibu Harini	Jawa Timur
2	FP-12269FF	2012-09-18	PRO-3605			1	Bapak Heru	Jawa Timur
3	FP-BC2DE0D	2012-02-14	PRO-7438			17	Ibu Hadi Sukarni	Jawa Barat
4	FP-0A2B246	2012-09-28	PRO-B97C			1	Ibu Aini Kasmaji	Jawa Tengah
5	FP-945312F	2011-03-27	PRO-4CA5			8	Ibu Siti Arya	Jawa Barat
6	FP-BDFA04B	2012-04-05	PRO-2990			7	Ibu Tyas	Jawa Tengah
7	FP-FFADF88	2011-12-30	PRO-0202			1	Ibu Hatamah	Jawa Timur
8	FP-EE0934F	2010-06-14	PRO-35C4			1	Ibu Tyas	Jawa Tengah
9	FP-71B507C	2012-04-05	PRO-75CD			4	Ibu Harini	Jawa Timur
10	FP-AFAE50B	2011-08-19	PRO-9CD8			5	Ibu Atik	Jawa Tengah
11	FP-23971D9	2012-06-22	PRO-5762			1	Ibu Niken	Jawa Tengah
12	FP-79A9FC0	2010-11-21	PRO-A846			5	Ibu Tyas	Jawa Tengah
13	FP-C1C117C	2012-12-15	PRO-10F3			1	Ibu Siti Arya	Jawa Barat
14	FP-EB0BC74	2011-12-28	PRO-CFA0			2	Bapak Ketut	Bali
15	FP-078C4A5	2010-03-26	PRO-5556			17	Ibu Hadi Sukarni	Jawa Barat
16	FP-C971451	2012-01-09	PRO-8CB6			44	Ibu Siti Arya	Jawa Barat
17	FP-A1BF7A1	2011-10-13	PRO-36A1			1	Ibu Hatamah	Jawa Timur
18	FP-EAD48CB	2011-01-05	PRO-B5B4			1	Bapak Imron	Jawa Barat
19	FP-FBD3529	2012-05-21	PRO-3C65			3	Bapak Totok	Jawa Timur
20	FP-D5ABF74	2011-04-09	PRO-1E30			3	Ibu Harini	Jawa Timur

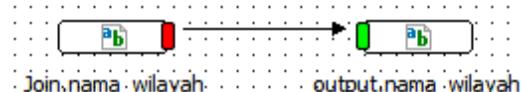
33. Jika kode_produk telah terisi data sesuai dengan kode_produk yang diambil dari Dimensi Produk, maka proses tersebut dinyatakan berhasil.
34. Selanjutnya adalah proses Join dengan Dimensi Pelanggan untuk memasukkan data kode pelanggan ke fakta penjualan.
35. Kembali ke editor Apatar, hapus hubungan antara operator **Join** dengan connector **FaktaPenjualan**.
- 36.Tambahkan 1 buah operator Join dan 1 connector MS Excel. Atur konfigurasi MS Excel untuk menggunakan file “**Data_penjualan.xls**” pada Sheet “**DimensiPelanggan**” yang terdiri dari 3 kolom.
37. Hubungkan operator-operator **Join** dengan connector **DimensiPelanggan** dan **FaktaPenjualan** seperti gambar berikut.



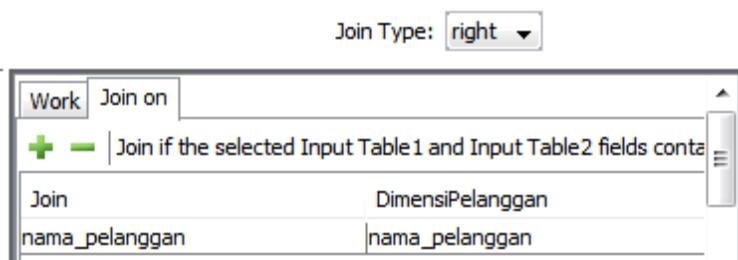
38. Atur konfigurasi **Join** yang ketiga.
39. Pada editor konfigurasi Join, masukkan 5 field dari **Input Table 1** (Join) yang terdiri dari **kode_fakta**, **kode_waktu**, **kode_produk**, **nama_wilayah**, dan **jumlah**. Masukkan pula 1 buah field dari **Input Table 2** (DimensiPelanggan) yaitu field **kode_pelanggan**. Sedangkan dari tabel output, masukkan 5 field yaitu **kode_fakta**, **kode_waktu**, **kode_produk**, **jumlah** dan **kode_pelanggan**. Hubungkan field-field yang sudah sesuai.



40. Untuk **nama_wilayah**, perlu dibuat field virtual terlebih dahulu pada tabel output. Klik “**Edit Output**”, tambahkan 1 field baru **nama_wilayah** dengan tipe data Text. Klik OK.
41. Masukkan field baru tersebut ke dalam editor dan hubungkan dengan field **nama_wilayah** dari tabel input.



42. Untuk menghubungkan tabel sumber dengan Dimensi Pelanggan, atur konfigurasi **Join On**. Pilih field **nama_pelanggan** pada kolom tabel Join dan DimensiPelanggan. Atur pula **Join Type = Right**. Klik OK untuk menutup jendela konfigurasi Join.



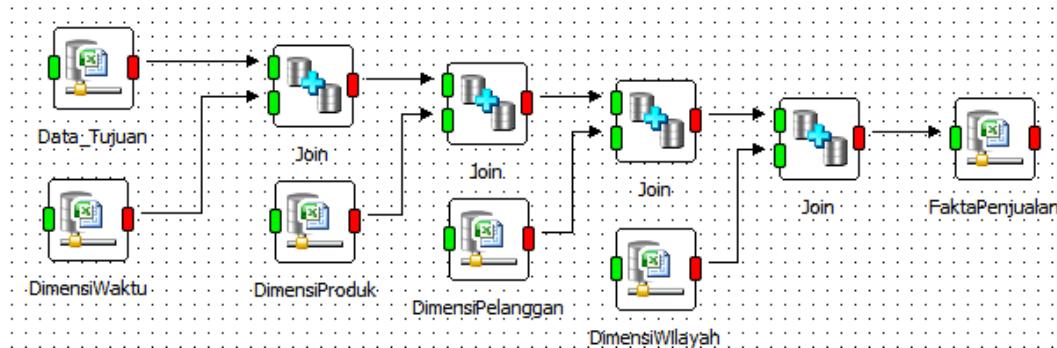
43. Untuk memeriksa keberhasilannya, klik kanan operator Join yang ketiga pilih Run and Preview Results.
44. Jika kode_pelanggan telah terisi data sesuai dengan kode_pelanggan yang diambil dari Dimensi Pelanggan, maka proses tersebut dinyatakan berhasil.

No.	kode_fakta	kode_waktu	kode_produk	kode_pelanggan	kode_wilayah	jumlah	nama_wilayah
1	FP-5701773	2011-01-05	PRO-B5B4	C-6434		1	Jawa Barat
2	FP-CAA91BA	2012-09-28	PRO-B97C	C-C954		1	Jawa Tengah
3	FP-B1BD55A	2012-01-04	PRO-55C3	C-E61F		14	Jawa Timur
4	FP-85F6FFA	2012-04-05	PRO-75CD	C-E61F		4	Jawa Timur
5	FP-8B71B09	2011-04-09	PRO-1E30	C-E61F		3	Jawa Timur
6	FP-8C235B7	2012-06-22	PRO-5762	C-2596		1	Jawa Tengah
7	FP-5F39C82	2011-12-28	PRO-CFA0	C-BD48		2	Bali
8	FP-EE2488E	2012-09-18	PRO-3605	C-163A		1	Jawa Timur
9	FP-01F534A	2012-02-14	PRO-7438	C-B5EB		17	Jawa Barat
10	FP-49FE1C4	2010-03-26	PRO-5556	C-B5EB		17	Jawa Barat
11	FP-E618CF7	2011-08-19	PRO-9CD8	C-14DB		5	Jawa Tengah
12	FP-A1A4FA7	2012-04-05	PRO-2990	C-5A8C		7	Jawa Tengah
13	FP-10A1E1D	2010-06-14	PRO-35C4	C-5A8C		1	Jawa Tengah
14	FP-A7E1218	2010-11-21	PRO-A846	C-5A8C		5	Jawa Tengah
15	FP-C13483B	2012-05-21	PRO-3C65	C-C899		3	Jawa Timur
16	FP-5E526DC	2011-12-30	PRO-0202	C-E04D		1	Jawa Timur
17	FP-1094E20	2011-10-13	PRO-36A1	C-E04D		1	Jawa Timur
18	FP-1288323	2011-03-27	PRO-4CA5	C-8421		8	Jawa Barat
19	FP-02FE49C	2012-12-15	PRO-10F3	C-8421		1	Jawa Barat
20	FP-6B2E64A	2012-01-09	PRO-8CB6	C-8421		44	Jawa Barat

45. Selanjutnya adalah proses Join dengan Dimensi Wilayah untuk memasukkan data kode wilayah ke fakta penjualan.
46. Kembali ke editor Apatar, hapus hubungan antara operator **Join** dengan connector **FaktaPenjualan**.

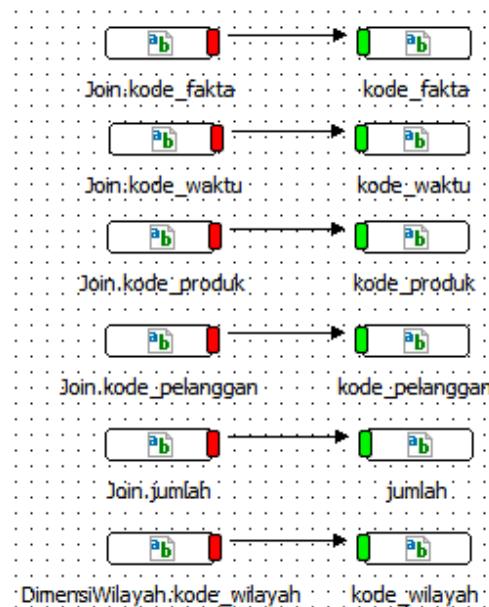
47. Tambahkan 1 buah operator Join dan 1 connector MS Excel. Atur konfigurasi MS Excel untuk menggunakan file “**Data_penjualan.xls**” pada Sheet “**DimensiWilayah**” yang terdiri dari 2 kolom.

48. Hubungkan operator-operator **Join** dengan connector **DimensiWilayah** dan **FaktaPenjualan** seperti gambar berikut.

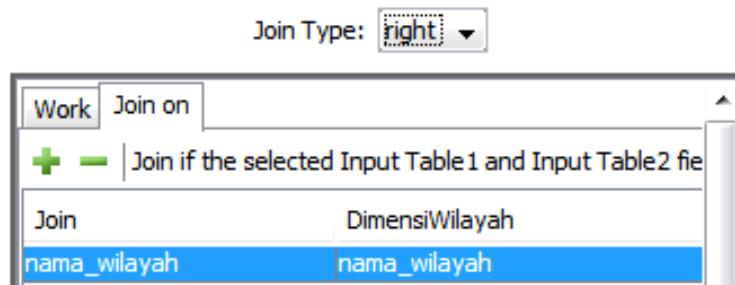


49. Atur konfigurasi **Join** yang keempat.

50. Pada editor konfigurasi Join, masukkan 5 field dari **Input Table 1** (Join) yang terdiri dari **kode_fakta**, **kode_waktu**, **kode_produk**, **kode_pelanggan**, dan **jumlah**. Masukkan pula 1 buah field dari **Input Table 2** (DimensiWilayah) yaitu field **kode_wilayah**. Sedangkan dari tabel output, masukkan semua field yaitu **kode_fakta**, **kode_waktu**, **kode_produk**, **kode_pelanggan**, **jumlah** dan **kode_wilayah**. Hubungkan field-field yang sesuai.



51. Untuk menghubungkan tabel sumber dengan Dimensi Wilayah, atur konfigurasi **Join On**. Pilih field **nama_wilayah** pada kolom tabel Join dan DimensiWilayah. Atur pula **Join Type = Right**. Klik OK untuk menutup jendela konfigurasi Join.



52. Untuk memeriksa keberhasilannya, klik kanan operator Join yang keempat pilih Run and Preview Results.
53. Jika kode_wilayah telah terisi data sesuai dengan kode_wilayah yang diambil dari Dimensi Wilayah, maka proses tersebut dinyatakan berhasil.

No.	kode_fakta	kode_waktu	kode_produk	kode_pelanggan	kode_wilayah	jumlah
1	FP-0B23C1F	2011-12-28	PRO-CFA0	C-BD48	W-A89C	2
2	FP-185DA3D	2012-01-04	PRO-55C3	C-E61F	W-67DF	14
3	FP-17C0B4D	2012-04-05	PRO-75CD	C-E61F	W-67DF	4
4	FP-B82D1FD	2011-04-09	PRO-1E30	C-E61F	W-67DF	3
5	FP-C2466E8	2012-09-18	PRO-3605	C-163A	W-67DF	1
6	FP-B327340	2012-05-21	PRO-3C65	C-C899	W-67DF	3
7	FP-A37B319	2011-12-30	PRO-0202	C-E04D	W-67DF	1
8	FP-347A92E	2011-10-13	PRO-36A1	C-E04D	W-67DF	1
9	FP-FFD30D4	2011-01-05	PRO-B5B4	C-6434	W-D709	1
10	FP-48D45EB	2012-02-14	PRO-7438	C-B5EB	W-D709	17
11	FP-5157B15	2010-03-26	PRO-5556	C-B5EB	W-D709	17
12	FP-E1CDE1D	2011-03-27	PRO-4CA5	C-8421	W-D709	8
13	FP-9E8B8FC	2012-12-15	PRO-10F3	C-8421	W-D709	1
14	FP-E6F74E9	2012-01-09	PRO-8CB6	C-8421	W-D709	44
15	FP-A9FFD9E	2012-09-28	PRO-B97C	C-C954	W-ECF2	1
16	FP-B2FDD37	2012-06-22	PRO-5762	C-2596	W-ECF2	1
17	FP-6AA5021	2011-08-19	PRO-9CD8	C-14DB	W-ECF2	5
18	FP-528891A	2012-04-05	PRO-2990	C-5A8C	W-ECF2	7
19	FP-EF239BD	2010-06-14	PRO-35C4	C-5A8C	W-ECF2	1
20	FP-BCCFC07	2010-11-21	PRO-A846	C-5A8C	W-ECF2	5

54. Dengan demikian, data siap di-*load* ke Fakta Penjualan. Tutup semua file Excel yang masih terbuka.
55. Pada editor apatar, klik kanan connector **FaktaPenjualan**, pilih Run and Preview Results.
56. Buka kembali file “**Data_penjualan.xls**” pada Sheet **FaktaPenjualan**.

A	B	C	D	E	F	
1	kode_fakta	kode_waktu	kode_produk	kode_pelanggan	kode_wilayah	jumlah
2	FP-6E73559	2011-12-28	PRO-CFA0	C-BD48	W-A89C	2
3	FP-60A4616	2012-01-04	PRO-55C3	C-E61F	W-67DF	14
4	FP-8A16978	2012-04-05	PRO-75CD	C-E61F	W-67DF	4
5	FP-C4A539B	2011-04-09	PRO-1E30	C-E61F	W-67DF	3
6	FP-C12F11B	2012-09-18	PRO-3605	C-163A	W-67DF	1
7	FP-190C9D1	2012-05-21	PRO-3C65	C-C899	W-67DF	3
8	FP-EA435C5	2011-12-30	PRO-0202	C-E04D	W-67DF	1
9	FP-B7A9DDC	2011-10-13	PRO-36A1	C-E04D	W-67DF	1
10	FP-AC214FC	2011-01-05	PRO-B5B4	C-6434	W-D709	1
11	FP-1A398A0	2012-02-14	PRO-7438	C-B5EB	W-D709	17
12	FP-0F62D98	2010-03-26	PRO-5556	C-B5EB	W-D709	17
13	FP-2382293	2011-03-27	PRO-4CA5	C-8421	W-D709	8
14	FP-52660FC	2012-12-15	PRO-10F3	C-8421	W-D709	1
15	FP-9C8E259	2012-01-09	PRO-8CB6	C-8421	W-D709	44
16	FP-09CAA7E	2012-09-28	PRO-B97C	C-C954	W-ECF2	1
17	FP-C3B7C2D	2012-06-22	PRO-5762	C-2596	W-ECF2	1
18	FP-F5AF61E	2011-08-19	PRO-9CD8	C-14DB	W-ECF2	5
19	FP-C645B36	2012-04-05	PRO-2990	C-5A8C	W-ECF2	7
20	FP-5306048	2010-06-14	PRO-35C4	C-5A8C	W-ECF2	1
21	FP-5E9E6D0	2010-11-21	PRO-A846	C-5A8C	W-ECF2	5

57. Jika semua data telah dimasukkan ke dalam Sheet **FaktaPenjualan**, maka proses ETL untuk fakta penjualan telah berhasil dilakukan.
58. Fakta Penjualan kini telah berisi kumpulan *foreign key* yang diambil dari *primary key* semua dimensi yang terhubung dan sebuah field sebagai *measure*.

3) Melihat isi data Fakta Penjualan secara detil

Fakta penjualan hanya berisi kumpulan *foreign key* dari dimensi-dimensi yang terhubung dan field-field *measure*. Namun, sebagai seorang pengguna *data warehouse* tentunya ingin melihat data fakta penjualan tersebut secara detil, tidak hanya sekedar melihat kode-kode. Oleh karena itu, fakta penjualan tersebut perlu dijabarkan menjadi tabel yang dapat terbaca oleh pengguna umum.

Sebagai contoh, pengguna ingin melihat nama pelanggan dan wilayahnya pada transaksi yang terjadi pada tanggal 28 Desember 2011 (dalam tabel tertulis **2011-12-28**), padahal dalam fakta penjualan hanya berisi **kode_pelanggan = C-BD48** dan **kode_wilayah = W-A89C**. Oleh karena itu, perlu proses *Joining* dengan metode **Left Join** karena menggunakan *primary key* sebagai kunci penghubung beberapa tabel.

	A	B	C	D	E	F
1	kode_fakta	kode_waktu	kode_produk	kode_pelanggan	kode_wilayah	jumlah
2	FP-6E73559	2011-12-28	PRO-CFA0	C-BD48	W-A89C	2
3	FP-60A4616	2012-01-04	PRO-55C3	C-E61F	W-67DF	14
4	FP-8A16978	2012-04-05	PRO-75CD	C-E61F	W-67DF	4

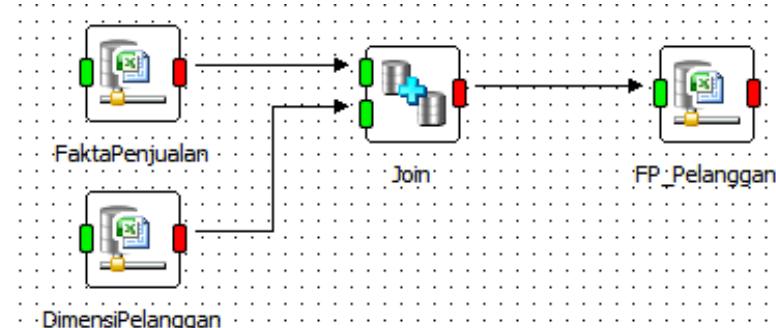
Adapun langkah-langkah membuat Fakta Penjualan untuk melihat nama_pelanggan dan wilayahnya sebagai berikut:

1. Buat file Ms. Excel baru dengan nama “**Fakta_Penjualan.xls**” (Format Excel 97-2003 Workbook). Ubah nama Sheet1 menjadi “**FP_Pelanggan**”.
2. Buat 6 kolom yang terdiri dari **kode_fakta**, **kode_waktu**, **kode_produk**, **nama_pelanggan**, **nama_wilayah** dan **jumlah**.

	A	B	C	D	E	F
1	kode_fakta	kode_waktu	kode_produk	nama_pelanggan	nama_wilayah	jumlah
2						
3						
4						

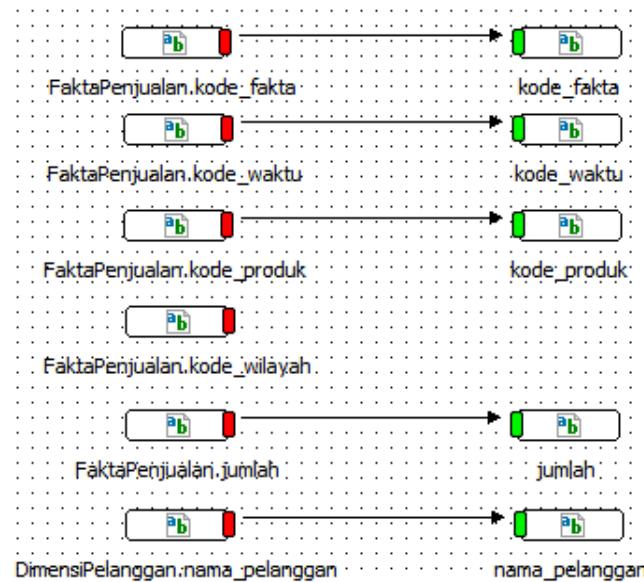
Fakta_Penjualan Sheet2 Sheet3

3. Buat file baru aplikasi apatar (File – New). Simpan dengan nama “**ETL_Fakta.aptr**”.
4. Masukkan 3 buah connector **MS Excel** dan sebuah operator **Join**.
5. Atur MS Excel yang pertama menggunakan file “**Data_penjualan.xls**” dengan Sheet **FaktaPenjualan** yang terdiri dari 6 kolom. MS Excel yang kedua menggunakan file “**Data_penjualan.xls**” dengan Sheet **DimensiPelanggan** yang terdiri dari 3 kolom. Sedangkan MS Excel yang ketiga menggunakan file “**Fakta_Penjualan.xls**” dengan Sheet **FP_Pelanggan** yang baru saja dibuat terdiri dari 6 kolom. Hubungkan ketiga connector tersebut dengan operator Join seperti gambar berikut.

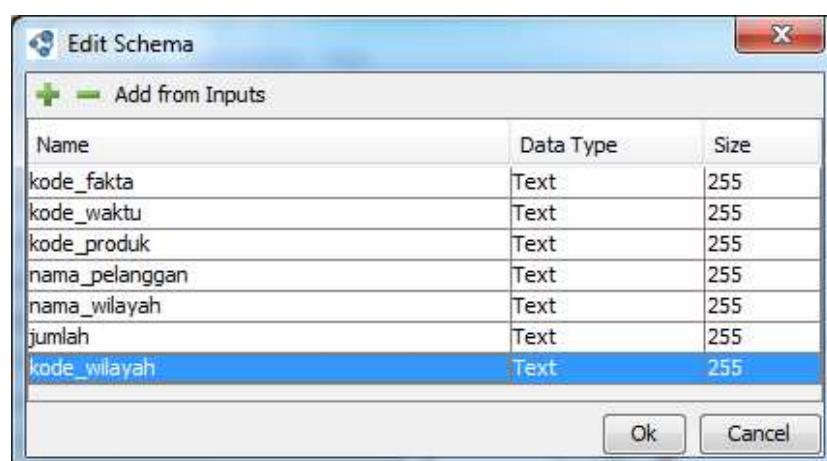


6. Atur konfigurasi Join dengan cara klik kanan operator Join, pilih Configure.

- Pada editor konfigurasi Join, masukkan 5 buah field yang diambil dari **Input Table 1** (FaktaPenjualan) yang terdiri dari **kode_fakta**, **kode_waktu**, **kode_produk**, **kode_wilayah** dan **jumlah**. Masukkan 1 field dari **Input Table 2** (DimensiPelanggan) yaitu **nama_pelanggan**.
- Masukkan pula dari tabel Output (sebelah kanan) sebanyak 5 field ke dalam editor yaitu **kode_fakta**, **kode_waktu**, **kode_produk**, **jumlah** dan **nama_pelanggan**. Hubungkan semua field-field yang telah sesuai antara input dengan output.



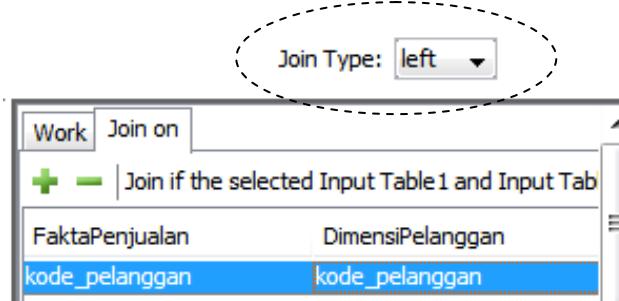
- Sedangkan field **kode_wilayah** belum memiliki hubungan dengan output dikarenakan belum ada field yang sesuai pada tabel output. Untuk itu perlu dibuat field virtual untuk **kode_wilayah**. Klik “**Edit Output**”, tambahkan sebuah field baru dengan nama “**kode_wilayah**” dengan tipe data Text.



10. Klik OK untuk menutup. Masukkan field baru dari tabel output tersebut ke dalam editor dan hubungkan dengan field kode_wilayah dari tabel input.



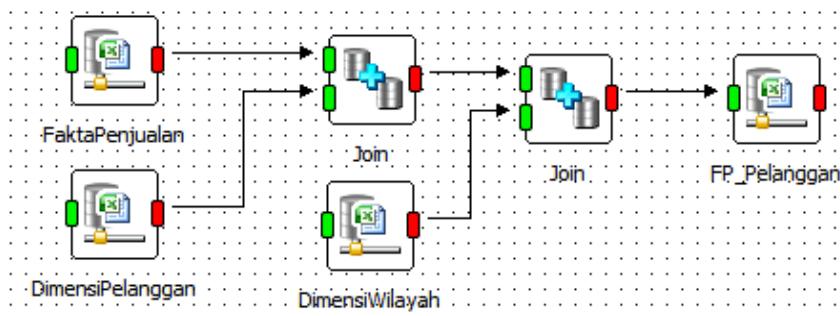
11. Atur konfigurasi **Join On** untuk membuat hubungan antara Fakta Penjualan dengan Dimensi Pelanggan menggunakan kode_pelanggan sebagai kunci penghubung.
12. Klik tab **Join On** di bagian atas editor, klik tanda plus (+) untuk menambahkan kondisi hubungannya. Pilih field **kode_pelanggan** di kedua kolom tabel dan tentukan **Join Type = Left** (menggunakan metode **Left Join**).



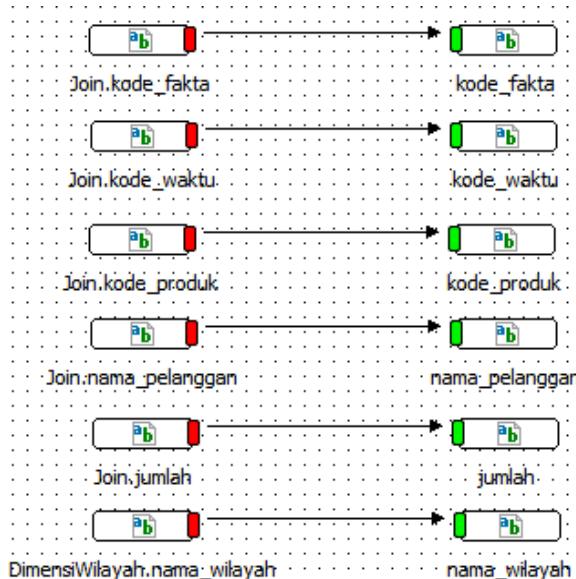
13. Klik OK untuk menutup jendela konfigurasi Join. Untuk memeriksa keberhasilan proses Join antara Fakta Penjualan dengan Dimensi Pelanggan, klik kanan operator Join pilih Run and Preview Results.

No.	kode_fakta	kode_waktu	kode_produk	nama_pelanggan	nama_wilayah	jumlah	kode_wilayah
1	FP-6E73559	2011-12-28	PRO-CFA0	Bapak Ketut		2	W-A89C
2	FP-60A4616	2012-01-04	PRO-55C3	Ibu Harini		14	W-67DF
3	FP-8A16978	2012-04-05	PRO-75CD	Ibu Harini		4	W-67DF
4	FP-C4A539B	2011-04-09	PRO-1E30	Ibu Harini		3	W-67DF
5	FP-C12F11B	2012-09-18	PRO-3605	Bapak Heru		1	W-67DF
6	FP-190C9D1	2012-05-21	PRO-3C65	Bapak Totok		3	W-67DF
7	FP-EA435C5	2011-12-30	PRO-0202	Ibu Hatamah		1	W-67DF
8	FP-B7A9DDC	2011-10-13	PRO-36A1	Ibu Hatamah		1	W-67DF
9	FP-AC214FC	2011-01-05	PRO-B5B4	Bapak Imron		1	W-D709
10	FP-1A398A0	2012-02-14	PRO-7438	Ibu Hadi Sukarni		17	W-D709
11	FP-0F62D98	2010-03-26	PRO-5556	Ibu Hadi Sukarni		17	W-D709
12	FP-2382293	2011-03-27	PRO-4CA5	Ibu Siti Arya		8	W-D709
13	FP-52660FC	2012-12-15	PRO-10F3	Ibu Siti Arya		1	W-D709
14	FP-9C8E259	2012-01-09	PRO-8CB6	Ibu Siti Arya		44	W-D709
15	FP-09CAA7E	2012-09-28	PRO-B97C	Ibu Aini Kasmaji		1	W-ECF2
16	FP-C3B7C2D	2012-06-22	PRO-5762	Ibu Niken		1	W-ECF2
17	FP-F5AF61E	2011-08-19	PRO-9CD8	Ibu Atik		5	W-ECF2
18	FP-C645B36	2012-04-05	PRO-2990	Ibu Tyas		7	W-ECF2
19	FP-5306048	2010-06-14	PRO-35C4	Ibu Tyas		1	W-ECF2
20	FP-5E9E6D0	2010-11-21	PRO-A846	Ibu Tyas		5	W-ECF2

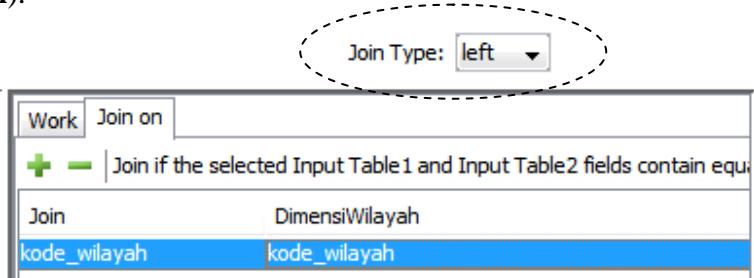
14. Terlihat dalam tabel sudah muncul nama pelanggan. Hal ini berarti proses Join berhasil dilakukan meskipun nama wilayah belum ditampilkan karena belum di-join-kan dengan Dimensi Wilayah.
15. Berikutnya adalah menampilkan nama wilayah ke dalam tabel. Kembali ke editor Apatar, hapus hubungan antara **Join** dengan connector **FP_Pelanggan**.
16. Tambahkan sebuah connector **MS Excel** dan sebuah operator **Join**. Atur konfigurasi MS Excel menggunakan file “**Data_penjualan.xls**” dengan Sheet **DimensiWilayah** yang terdiri atas 2 kolom.
17. Hubungkan kembali connector tersebut dengan operator Join.



18. Atur konfigurasi Join dengan klik kanan operator Join pilih **Configure**.
19. Masukkan 5 field dari **Input Table 1** (Join) ke dalam editor yaitu **kode_fakta, kode_waktu, kode_produk, nama_pelanggan, dan jumlah**. Masukkan 1 field dari **Input Table 2** (DimensiWilayah) yaitu **nama_wilayah**.
20. Masukkan pula semua field (6 field) dari tabel output (sebelah kanan) ke dalam editor. Hubungkan dengan field dari tabel input yang sesuai.



21. Atur konfigurasi **Join On** untuk membuat hubungan antara Fakta Penjualan dengan Dimensi Wilayah menggunakan kode_wilayah sebagai kunci penghubung.
22. Klik tab **Join On** di bagian atas editor, klik tanda plus (+) untuk menambahkan kondisi hubungannya. Pilih field **kode_wilayah** di kedua kolom tabel dan tentukan **Join Type = Left** (menggunakan metode **Left Join**).



23. Klik OK untuk menutup jendela konfigurasi Join. Untuk memeriksa keberhasilan proses Join antara Fakta Penjualan dengan Dimensi Wilayah, klik kanan operator Join pilih Run and Preview Results.

No.	kode_fakta	kode_waktu	kode_produk	nama_pelanggan	nama_wilayah	jumlah
1	FP-6E73559	2011-12-28	PRO-CFA0	Bapak Ketut	Bali	2
2	FP-60A4616	2012-01-04	PRO-55C3	Ibu Harini	Jawa Timur	14
3	FP-8A16978	2012-04-05	PRO-75CD	Ibu Harini	Jawa Timur	4
4	FP-C4A539B	2011-04-09	PRO-1E30	Ibu Harini	Jawa Timur	3
5	FP-C12F11B	2012-09-18	PRO-3605	Bapak Heru	Jawa Timur	1
6	FP-190C9D1	2012-05-21	PRO-3C65	Bapak Totok	Jawa Timur	3
7	FP-EA435C5	2011-12-30	PRO-0202	Ibu Hatamah	Jawa Timur	1
8	FP-B7A9DDC	2011-10-13	PRO-36A1	Ibu Hatamah	Jawa Timur	1
9	FP-AC214FC	2011-01-05	PRO-B5B4	Bapak Imron	Jawa Barat	1
10	FP-1A398A0	2012-02-14	PRO-7438	Ibu Hadi Sukarni	Jawa Barat	17
11	FP-0F62D98	2010-03-26	PRO-5556	Ibu Hadi Sukarni	Jawa Barat	17
12	FP-2382293	2011-03-27	PRO-4CA5	Ibu Siti Arya	Jawa Barat	8
13	FP-52660FC	2012-12-15	PRO-10F3	Ibu Siti Arya	Jawa Barat	1
14	FP-9C8E259	2012-01-09	PRO-8CB6	Ibu Siti Arya	Jawa Barat	44
15	FP-09CAA7E	2012-09-28	PRO-B97C	Ibu Aini Kasmaji	Jawa Tengah	1
16	FP-C3B7C2D	2012-06-22	PRO-5762	Ibu Niken	Jawa Tengah	1
17	FP-F5AF61E	2011-08-19	PRO-9CD8	Ibu Atik	Jawa Tengah	5
18	FP-C645B36	2012-04-05	PRO-2990	Ibu Tyas	Jawa Tengah	7
19	FP-5306048	2010-06-14	PRO-35C4	Ibu Tyas	Jawa Tengah	1
20	FP-5E9E6D0	2010-11-21	PRO-A846	Ibu Tyas	Jawa Tengah	5

24. Perhatikan tabel hasil proses Join. Kini nama pelanggan dan nama wilayah sudah dapat dilihat. Dengan demikian data siap dimasukkan ke dalam file Excel “**Fakta_Penjualan.xls**” pada Sheet “**FP_Pelanggan**”.
25. Tutup semua file Excel yang masih terbuka. Pada editor Apatar, klik kanan connector **FP_Pelanggan** pilih **Run and Preview Results**.
26. Buka kembali file “**Fakta_Penjualan.xls**” pada Sheet “**FP_Pelanggan**”. Jika data telah terisi ke dalam tabel, maka proses tersebut dinyatakan berhasil.
27. Lihat pada tabel dalam file “**Fakta_Penjualan.xls**” pada Sheet **FP_Pelanggan**. Kini pengguna dapat mengetahui nama pelanggan dan nama wilayah pada transaksi yang terjadi pada tanggal **28 Desember 2011**, yaitu **Bapak Ketut dari Bali**.

A	B	C	D	E	F	
1	kode_fakta	kode_waktu	kode_produk	nama_pelanggan	nama_wilayah	jumlah
2	FP-6E73559	2011-12-28	PRO-CFA0	Bapak Ketut	Bali	2
3	FP-60A4616	2012-01-04	PRO-55C3	Ibu Harini	Jawa Timur	14
4	FP-8A16978	2012-04-05	PRO-75CD	Ibu Harini	Jawa Timur	4
5	FP-C4A539B	2011-04-09	PRO-1E30	Ibu Harini	Jawa Timur	3
6	FP-C12F11B	2012-09-18	PRO-3605	Bapak Heru	Jawa Timur	1
7	FP-190C9D1	2012-05-21	PRO-3C65	Bapak Totok	Jawa Timur	3
8	FP-EA435C5	2011-12-30	PRO-0202	Ibu Hatamah	Jawa Timur	1
9	FP-B7A9DDC	2011-10-13	PRO-36A1	Ibu Hatamah	Jawa Timur	1
10	FP-AC214FC	2011-01-05	PRO-B5B4	Bapak Imron	Jawa Barat	1
11	FP-1A398A0	2012-02-14	PRO-7438	Ibu Hadi Sukarni	Jawa Barat	17
12	FP-0F62D98	2010-03-26	PRO-5556	Ibu Hadi Sukarni	Jawa Barat	17
13	FP-2382293	2011-03-27	PRO-4CA5	Ibu Siti Arya	Jawa Barat	8
14	FP-52660FC	2012-12-15	PRO-10F3	Ibu Siti Arya	Jawa Barat	1
15	FP-9C8E259	2012-01-09	PRO-8CB6	Ibu Siti Arya	Jawa Barat	44
16	FP-09CAA7E	2012-09-28	PRO-B97C	Ibu Aini Kasmaji	Jawa Tengah	1
17	FP-C3B7C2D	2012-06-22	PRO-5762	Ibu Niken	Jawa Tengah	1
18	FP-F5AF61E	2011-08-19	PRO-9CD8	Ibu Atik	Jawa Tengah	5
19	FP-C645B36	2012-04-05	PRO-2990	Ibu Tyas	Jawa Tengah	7
20	FP-5306048	2010-06-14	PRO-35C4	Ibu Tyas	Jawa Tengah	1
21	FP-5E9E6D0	2010-11-21	PRO-A846	Ibu Tyas	Jawa Tengah	5

E. Tugas

Gunakan file Excel **"Fakta_Penjualan.xls"**, selesaikan tugas berikut di kelas. Jika belum selesai, bisa dilanjutkan di rumah dan akan dinilai pada pertemuan berikutnya.

1. Buat tabel baru pada Sheet2 dan ubah nama **Sheet2** menjadi **"FP_Produk"**. Buat 7 buah kolom yang terdiri dari **kuartal**, **tahun**, **nama_produk**, **nama_pelanggan**, **nama_wilayah**, **jumlah** dan **harga**.

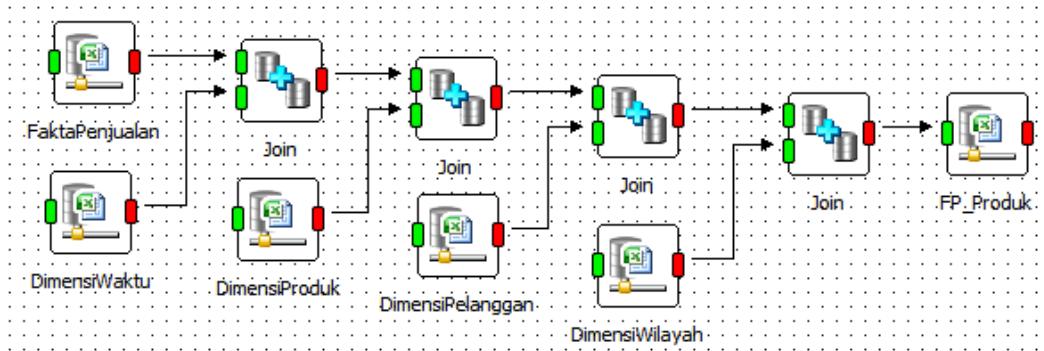
A	B	C	D	E	F	G	
1	kuartal	tahun	nama_produk	nama_pelanggan	nama_wilayah	jumlah	harga
2							
3							

Isikan data dari file **"Data_penjualan.xls"** pada Sheet **FaktaPenjualan** ke dalam tabel tersebut dan hubungkan dengan dimensi-dimensi yang dibutuhkan melalui proses Join dengan ketentuan sebagai berikut:

- a) Kolom **kuartal** dan **tahun** diambil dari **DimensiWaktu** berdasarkan **kode_waktu**.
- b) Kolom **nama_produk** dan **harga** diambil dari **DimensiProduk** berdasarkan **kode_produk**.

- c) Kolom **nama_pelanggan** diambil dari **DimensiPelanggan** berdasarkan **kode_pelanggan**.
- d) Kolom **nama_wilayah** diambil dari **DimensiWilayah** berdasarkan **kode_wilayah**.
- e) Kolom **jumlah** diambil langsung dari **FaktaPenjualan**.

Rancangan Proses:



2. Buat tabel baru pada Sheet3 dan ubah nama **Sheet3** menjadi **"Fact_Table"**. Buat 12 buah kolom yang terdiri dari **bulan**, **kuartal**, **tahun**, **nama_produk**, **nama_kategori**, **nama_subkategori**, **nama_pola**, **nama_pelanggan**, **jenis_kelamin**, **nama_wilayah**, **jumlah** dan **harga**.

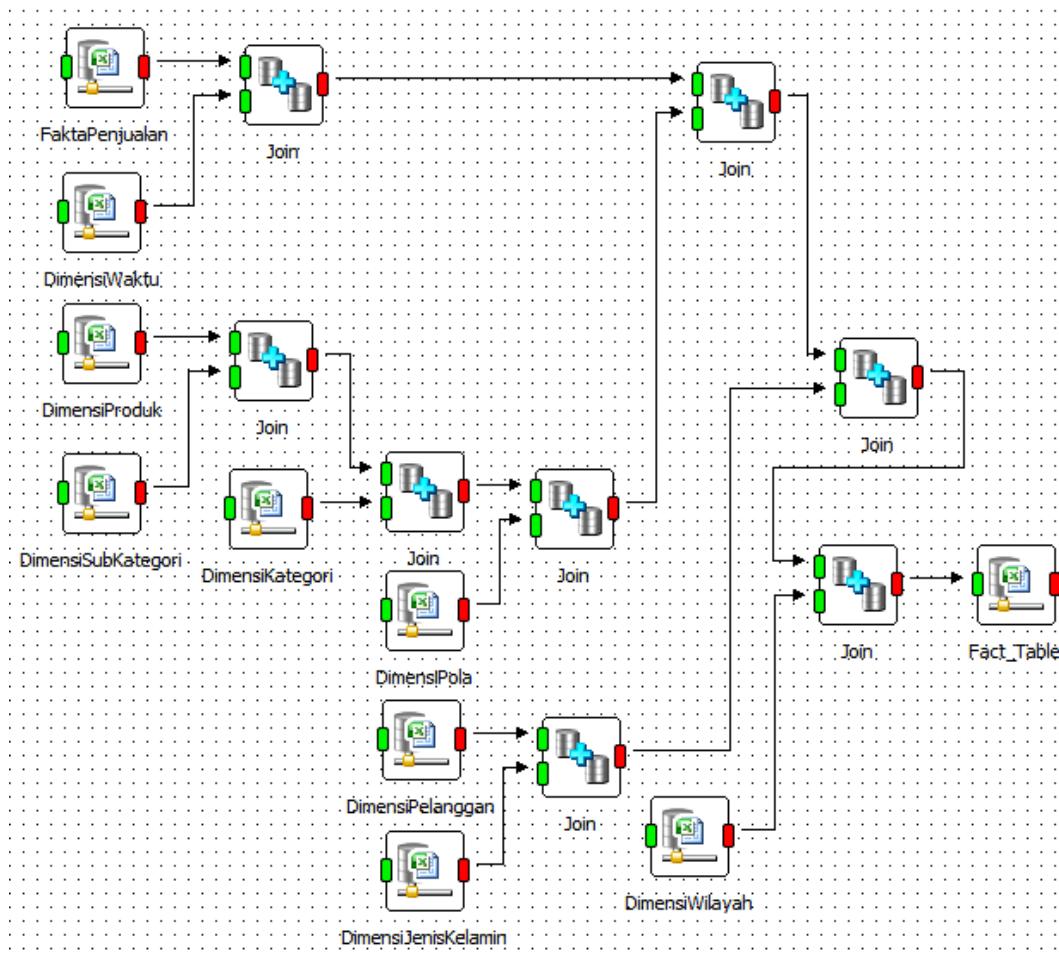
	A	B	C	D	E	F	G	H	I	J	K	L
1	bulan	kuartal	tahun	nama_produk	nama_kategori	nama_subkategori	nama_pola	nama_pelanggan	jenis_kelamin	nama_wilayah	jumlah	harga
2												
3												
4												

Isikan data dari file "Data_penjualan.xls" pada Sheet **FaktaPenjualan** ke dalam tabel tersebut dan hubungkan dengan dimensi-dimensi yang dibutuhkan melalui proses Join dengan ketentuan sebagai berikut:

- a) Kolom **bulan**, **kuartal** dan **tahun** diambil dari **DimensiWaktu** berdasarkan **kode_waktu**.
- b) Kolom **nama_produk** dan **harga** diambil dari **DimensiProduk** berdasarkan **kode_produk**.
- c) Kolom **nama_kategori** diambil dari **DimensiKategori** berdasarkan **kode_kategori**.

- d) Kolom **nama_subkategori** diambil dari **DimensiSubKategori** berdasarkan **kode_subkategori**.
 - e) Kolom **nama_pola** diambil dari **DimensiPola** berdasarkan **kode_pola**.
 - f) Kolom **nama_pelanggan** diambil dari **DimensiPelanggan** berdasarkan **kode_pelanggan**.
 - g) Kolom **jenis_kelamin** diambil dari **DimensiJenisKelamin** berdasarkan **kode_jeniskelamin**.
 - h) Kolom **nama_wilayah** diambil dari **DimensiWilayah** berdasarkan **kode_wilayah**.
 - i) Kolom **jumlah** diambil langsung dari **FaktaPenjualan**.

Rancangan Proses:



MODUL 5

PIVOT TABLE DAN CHART

A. Tujuan

1. Mahasiswa mampu melakukan menampilkan *Data Warehouse* secara multidimensi dengan *Pivot Table* dan *Chart*.

B. Landasan Teori

Sebuah *Pivot Table* sangat berguna untuk menyimpulkan, menganalisa, mengeksplorasi dan menyajikan data yang mudah dibaca dan dimengerti.

Pivot Table adalah fitur dari Microsoft Excel yang sangat memudahkan dalam merangkum sejumlah besar data. Biasanya *Pivot Table* digunakan untuk menganalisa data numerik secara rinci, dari sini akan diperoleh jawaban dari pertanyaan-pertanyaan yang biasanya tidak terduga dari suatu data.

Pivot Table berfungsi antara lain untuk:

- 1) Melakukan Query sejumlah data yang sangat besar dengan cara yang mudah.
- 2) Proses Kalkulasi subtotal dan menjumlahkan data numerik, meringkas data dengan sebuah kategori dan subkategori, serta membuat perhitungan dengan formula dan rumusan yang dapat dibuat sendiri.
- 3) Memperluas dan mempersempit tingkatan tampilan data yang berguna untuk fokus terhadap apa yang ingin dicari, dan menampilkan secara detil dari ringkasan data (yang menjadi titik fokus perhatian)
- 4) Melihat data dari dimensi yang diinginkan.

C. Alat dan Bahan

1. Komputer dengan sistem operasi Windows.
2. Ms.Excel.
3. Modul Praktikum Data Warehousing dan Data Mining.

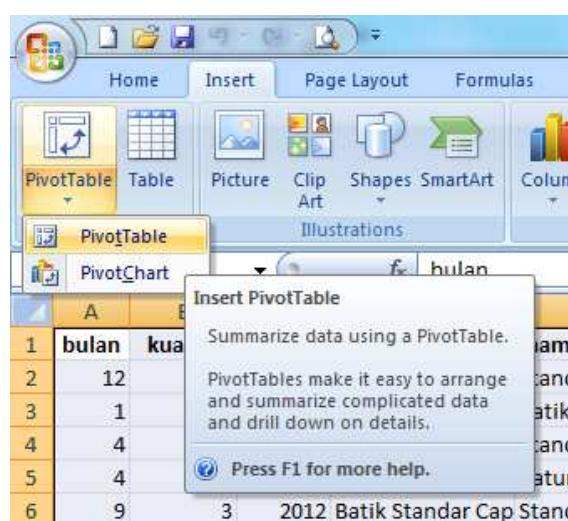
D. Langkah-langkah Praktikum

D.1. Kegiatan 1: Membuat Pivot Table

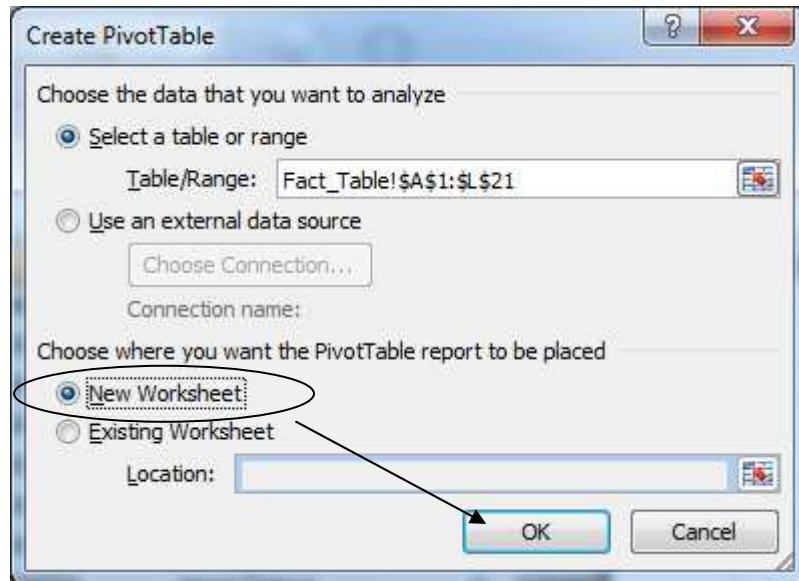
1. Gunakan file dengan nama “**Fakta_Penjualan.xls**” hasil tugas Modul 4 soal nomor 2. Jika memungkinkan simpan file tersebut dengan format excel 2007 ke atas (*.xlsx).
2. Buka sheet **Fact_Table**, dimana datanya terlihat seperti pada gambar berikut.

The screenshot shows a Microsoft Excel window titled "Fakta_Penjualan [Compatibility Mode] - Microsoft Excel". The ribbon at the top has tabs for Home, Insert, Page Layout, Formulas, Data, Review, View, and Team. The Home tab is selected. Below the ribbon is a toolbar with various icons for font, alignment, number, styles, and cells. The main area displays a table with 21 rows and 12 columns. The columns are labeled: bulan, kuartal, tahun, nama_produk, nama_kategori, nama_subkategori, nama_pola, nama_pelanggan, jenis_kelamin, nama_wilayah, jumlah, and harga. The data includes various product names like Jarik, Kaos, Bolero, etc., categorized by year (2011-2012), quarter (1-4), and brand (Print, Cap, Batik, etc.). It also includes customer information like Ibu Harini, Ibu Hatiyah, and Ibu Tyas, along with gender (PRIA, WANITA) and location (Bali, Jawa Timur, Jawa Barat). The last two columns show quantity (jumlah) and price (harga) respectively.

3. Pilih range data A1:L21 atau tekan tombol **CTRL + SHIFT + ***.
4. Klik tab **Insert** pada Ribbon, pilih menu **PivotTable | Insert PivotTable**.



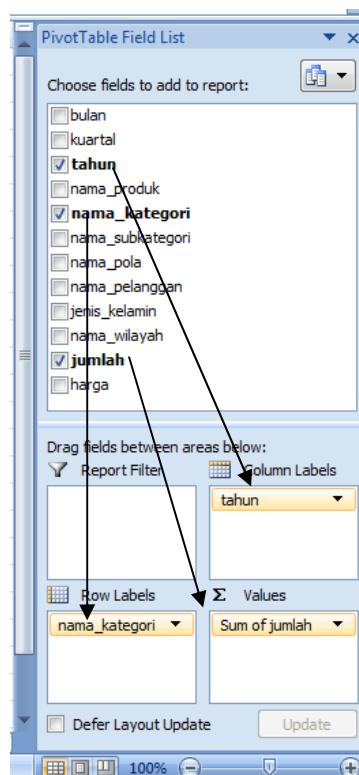
5. Pada dialog Create PivotTable yang muncul, pilih **New Worksheet**, klik tombol **OK**.



6. Sheet baru akan muncul disertai suatu kotak / placeholder PivotTable (**PivotTable Box**). Selain itu terdapat panel daftar field (**PivotTable Field List**) pada posisi sebelah kanan worksheet. Terlihat pada daftar tersebut 10 field heading dari range data yang dipilih sebelumnya.

7. Pada bagian bawah panel sebelah kanan terdapat 4 kotak area. Tiap kotak tersebut dapat ditambahkan field-field yang terdapat pada field list. Adapun fungsi dari 4 kotak tersebut adalah sebagai berikut :
- a) **Report Filter:** pada kotak ini field akan digunakan sebagai filter yang mempengaruhi hasil data pada PivotTable namun tidak akan terlihat sebagai isi dari PivotTable itu sendiri.

- b) **Column Labels:** data dari field akan ditempatkan pada bagian kolom dari tabel dengan level sesuai urutan susunan pada area ini.
 - c) **Row Labels:** data dari field akan ditempatkan pada bagian baris dari tabel dengan level sesuai urutan susunan pada area ini.
 - d) **Values:** nilai field yang terdapat pada kotak ini akan dijadikan sebagai basis perhitungan *summary*. Tipe *summary* yang bisa digunakan adalah *count*, *sum*, *average*, *max*, *min* dan lain-lain.
8. Cobalah berbagai kombinasi penempatan field dalam kotak area tersebut. Susunlah layout field dengan urutan berikut :
- a) Field **nama_kategori** ke kotak **Row Labels**.
 - b) Field **tahun** ke kotak **Column Labels**.
 - c) Field **jumlah** ke kotak **Values**.



Perhatikan pada saat ditempatkan di kotak **Values**, nama field **jumlah** akan berubah menjadi **Sum of jumlah**. Ini menandakan bahwa field tersebut merupakan kalkulasi **sum (penjumlahan)** dari nilai-nilai field **jumlah**.

9. Perhatikan hasil pengaturan ini pada area PivotTable. Area ini akan berisi suatu tabel dengan grouping field **nama_kategori** pada bagian baris, field

tahun pada kolom. Sedangkan nilai total jumlah_unit ditempatkan pada cell-cell hasil perpotongan item grouping baris dan kolom tersebut.

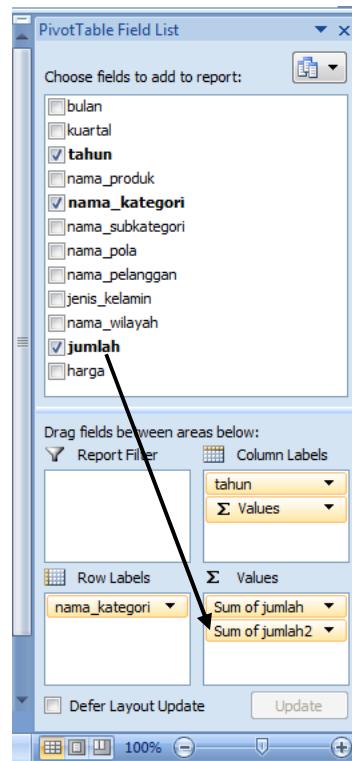
	3 Sum of jumlah	Column Labels				
	4 Row Labels		2010	2011	2012	Grand Total
5	Bahan		1	8	8	17
6	Batik				1	1
7	Bolero			1		1
8	Celana		17		17	34
9	Hem		5	8	4	17
10	Jam				44	44
11	Jarik			2	4	6
12	Kaos			1	14	15
13	Rok				1	1
14	Sarimbit			1		1
15	Grand Total		23	21	93	137

Salah satu contoh perpotongan adalah total jumlah yang terjual dengan kategori **Jam** selama tahun **2012**, adalah sebesar **44** unit.

10. Simpan file dengan nama yang sama.

D.2. Kegiatan 2 : Menambahkan Tipe Summary Baru

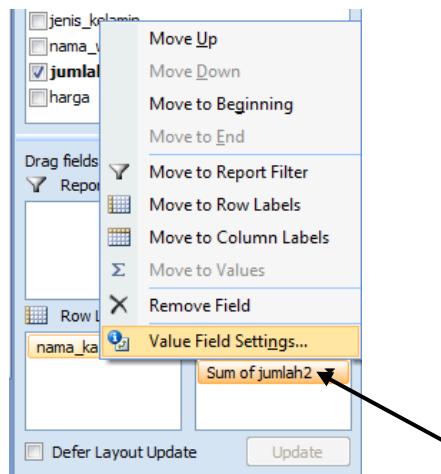
1. Masih bekerja menggunakan file “**Fakta_Penjualan.xls**” pada kegiatan 1 dengan Sheet1 PivotTable.
2. Tambahkan field **jumlah** kembali ke kotak **Value** dengan cara klik dan drag, sehingga muncul field baru dengan nama **Sum of jumlah2**.



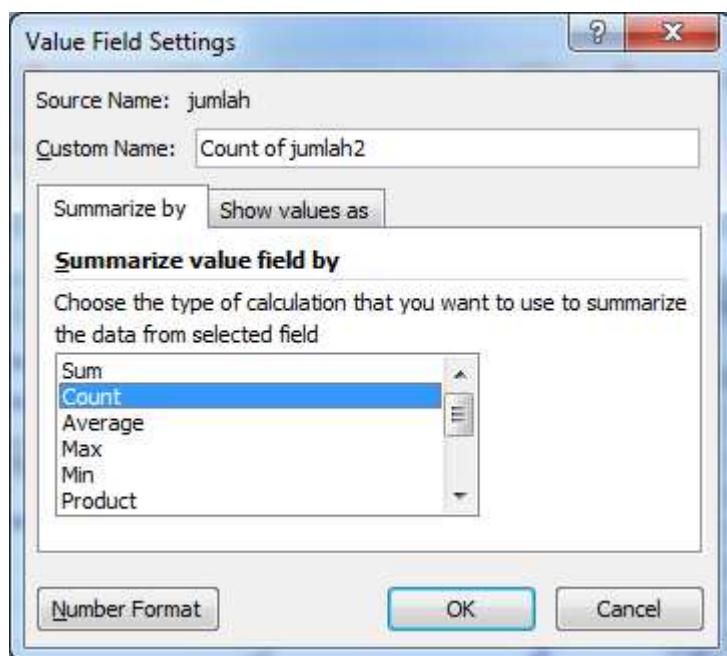
3. Akan diperoleh tambahan satu kolom perhitungan baru yang sama dengan hasil sebelumnya pada masing-masing tahun. Namun tentunya bukan ini yang diinginkan.

	Column Labels		2010	2011		2012		Total Sum of jumlah	Total Sum of jumlah2
Row Labels	Sum of jumlah	Sum of jumlah2	Sum of jumlah	Sum of jumlah2	Sum of jumlah	Sum of jumlah2			
Bahan	1	1	8	8	8	8	17	17	17
Batik					1	1	1	1	1
Bolero			1	1			1	1	1
Celana	17	17			17	17	34	34	34
Hem	5	5	8	8	4	4	17	17	17
Jam					44	44	44	44	44
Jarik			2	2	4	4	6	6	6
Kaos			1	1	14	14	15	15	15
Rok					1	1	1	1	1
Sarimbit			1	1			1	1	1
Grand Total	23	23	21	21	93	93	137	137	137

4. Kembali ke area **Values**, dan klik tombol panah ke bawah pada field **Sum of jumlah2**. Pilih item **Value Field Settings**.



5. Pada dialog Value Field Settings, ubah **Sum** menjadi **Count**. Perhatikan nama field akan berubah menjadi **Count of jumlah2**.



6. Klik tombol **OK**.
7. Pada area PivotTable, didapatkan dua *summary* yaitu:
 - a) nilai jumlah unit penjualan yang terjadi (**sum**).
 - b) jumlah transaksi yang terjadi (**count**).

	2010		2011		2012		Total Sum of jumlah		Total Count of jumlah2
Row Labels	Sum of jumlah	Count of jumlah2	Sum of jumlah	Count of jumlah2	Sum of jumlah	Count of jumlah2			
6 Bahan	1	1	8	1	8	2	17	4	
7 Batik				1	1	1	1	1	1
8 Bolero			1	1				1	1
9 Celana	17	1			17	1	34	2	
10 Hem	5	1	8	2	4	2	17	5	
11 Jam					44	1	44	1	
12 Jarik			2	1	4	1	6	2	
13 Kaos			1	1	14	1	15	2	
14 Rok					1	1	1	1	
15 Sarimbit			1	1			1	1	
16 Grand Total	23	3	21	7	93	10	137	20	

8. Simpan kembali dengan nama file yang sama.

D.3. Kegiatan 3 : Calculated Field dan Calculated Item di Pivot Table

Pada **PivotTable** terdapat fasilitas yang bisa digunakan untuk menambahkan perhitungan dengan nama **Calculated Field** dan **Calculated Item** untuk membantu analisa lebih lanjut. Perbedaan dari kedua fasilitas ini yaitu:

- a) **Calculated Field** digunakan jika ingin menambahkan field / kolom baru pada daftar field yang ada.
- b) **Calculated Item** digunakan jika ingin menambahkan daftar nilai dari suatu field, dengan ini otomatis menambah item grouping baru. Sebagai catatan, formula tidak boleh menggunakan item dari field lain.

Berikut penggunaan Calculated Field dan Calculated Item.

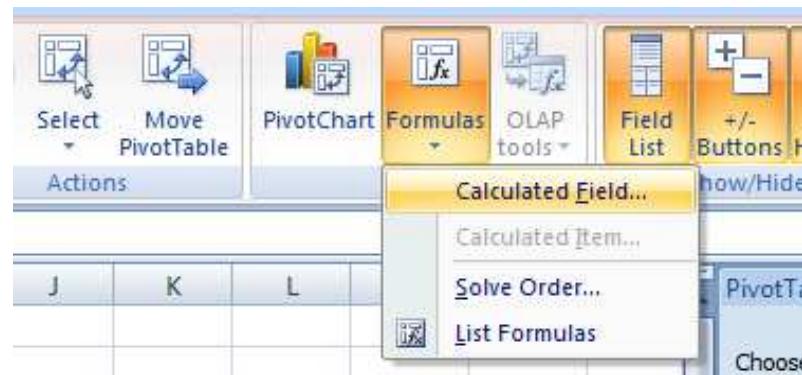
a) Calculated Field

Misalkan diinginkan untuk menambahkan sebuah field, yaitu jumlah pendapatan yang diperoleh berdasarkan jumlah produk yang terjual dikalikan dengan harga produk menggunakan Pivot Table yang terdapat pada file “**Fakta_Penjualan.xls**” pada Sheet **Fact_Table**.

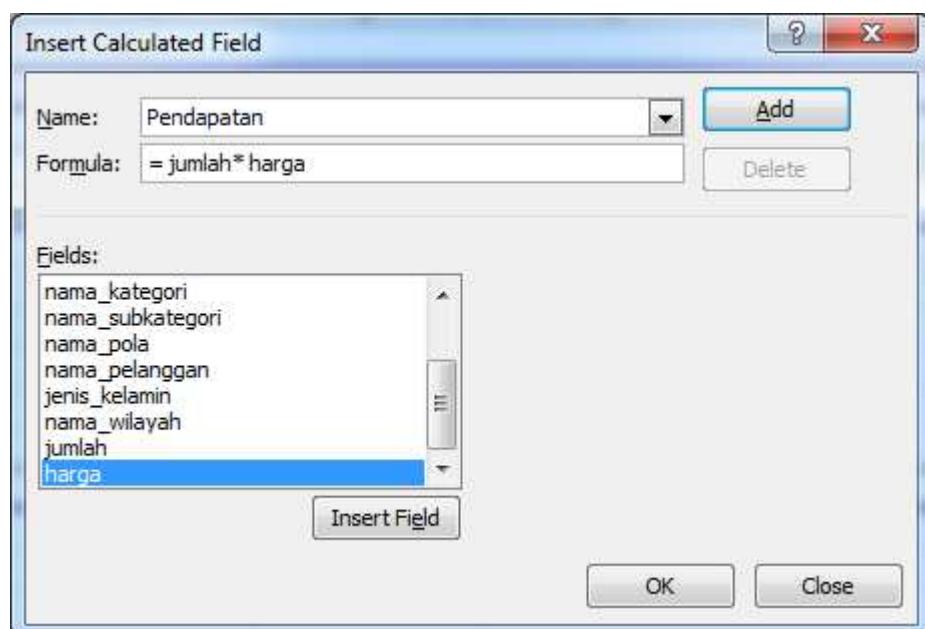
Berikut adalah langkah-langkah untuk melakukan hal tersebut:

1. Buka Sheet1 dalam file **Fakta_Penjualan.xls**, dan letakkan kursor ke area PivotTable.

- Pada menu ribbon **PivotTable Tools | Options**, klik button **Formulas** dan pilih **Calculated Field**.



- Pada kotak dialog **Insert Calculated Field** yang muncul, masukkan nilai berikut kemudian klik tombol **OK**.
 - Name : Pendapatan
 - Formula : = jumlah * harga (Pilih field **jumlah** kemudian klik Insert Field kemudian ketikkan tanda "*" dan masukkan field **harga**)



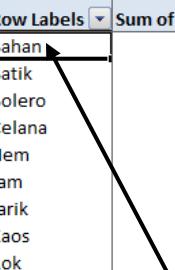
- Field baru, "Sum of Pendapatan" akan muncul pada Pivot Table.

				Total	Total	
				Sum of jumlah	Count of jumlah2	
		2012				Total Sum of Pendapatan
	Row Labels	Sum of jumlah	Count of jumlah2	Sum of Pendapatan		
6	Bahan	8	2	2.120.000	17	4
7	Batik	1	1	150.000	1	1
8	Bolero			-	1	1
9	Celana	17	1	935.000	34	2
10	Hem	4	2	1.596.000	17	5
11	Jam	44	1	3.520.000	44	1
12	Jarik	4	1	160.000	6	2
13	Kaos	14	1	420.000	15	2
14	Rok	1	1	225.000	1	1
15	Sarimbit			-	1	1
16	Grand Total	93	10	115.692.000	137	20
						451.963.000

b) Calculated Item

Misalkan diinginkan untuk menambahkan satu nilai pada field **nama_kategori**, yaitu **Baju Atasan** yang mewakili jumlah produk yang terjual untuk kategori **Hem** dan **Kaos**. Berikut adalah langkah-langkah untuk melakukan hal tersebut:

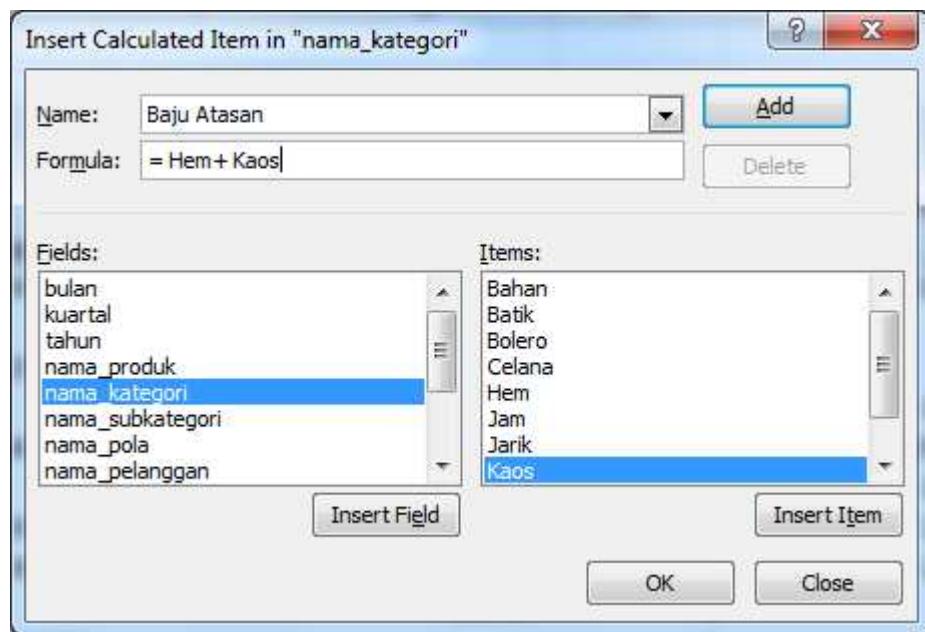
1. Buka Sheet1 dan arahkan cursor ke area nilai nama_kategori pada Pivot Table. Sebagai contoh dengan memilih kategori **Bahan**.



				Total	Total	
				Sum of jumlah	Count of jumlah2	
		2012				Total Sum of Pendapatan
	Row Labels	Sum of jumlah	Count of jumlah2	Sum of Pendapatan		
6	Bahan	8	2	2.120.000	17	4
7	Batik	1	1	150.000	1	1
8	Bolero			-	1	1
9	Celana	17	1	935.000	34	2
10	Hem	4	2	1.596.000	17	5
11	Jam	44	1	3.520.000	44	1
12	Jarik	4	1	160.000	6	2
13	Kaos	14	1	420.000	15	2
14	Rok	1	1	225.000	1	1
15	Sarimbit			-	1	1
16	Grand Total	93	10	115.692.000	137	20
						451.963.000

2. Pada ribbon **PivotTable Tools | Options**, klik button "Formulas" dan pilih "Calculated Item".
3. Pada kotak dialog **Insert Calculated Item in "nama_kategori"** yang muncul, masukkan nilai berikut di bawah ini kemudian klik tombol **OK**.
 - a) Name : Baju Atasan

- b) Formula : = Hem + Kaos (Pilih item **Hem** kemudian klik **Insert Item**, ketikkan tanda “+” dan pilih **Kaos** kemudian klik **Insert Item** lagi)



4. Item baru pada nama_kategori yaitu **Baju Atasan** dan juga total penjumlahan unit dan Pendapatan akan muncul pada Pivot Table.

		2011		2012		Total Sum of jumlah	Total Sum of Pendapatan
Row Labels		Sum of jumlah	Sum of Pendapatan	Sum of jumlah	Sum of Pendapatan		
6	Bahan	8	960.000	8	2.120.000	17	15.045.000
7	Batik		-	1	150.000	1	150.000
8	Bolero	1	225.000		-	1	225.000
9	Celana		-	17	935.000	34	3.740.000
10	Hem	8	4.960.000	4	1.596.000	17	19.023.000
11	Jam		-	44	3.520.000	44	3.520.000
12	Jarik	2	450.000	4	160.000	6	1.590.000
13	Kaos	1	60.000	14	420.000	15	1.350.000
14	Rok		-	1	225.000	1	225.000
15	Sarimbit	1	150.000		-	1	150.000
16	Baju Atasan	9	6.120.000	18	7.722.000	32	38.688.000
17	Grand Total	30	62.400.000	111	185.703.000	169	761.852.000

Nilai Baju Atasan diperoleh dari penjumlahan Hem dan Kaos

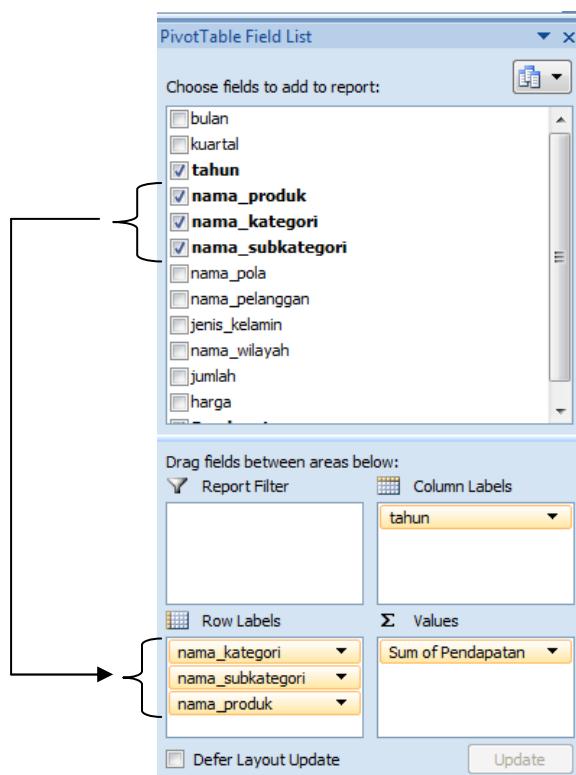
D.4. Kegiatan 4 : Operasi Roll Up dan Drill Down

Operasi Roll Up dan Drill Down digunakan untuk melihat data secara lebih rinci dan secara lebih umum berdasarkan kategori tertentu pada sebuah data warehouse yang disajikan dalam bentuk *cube* (multidimensi). Secara khusus,

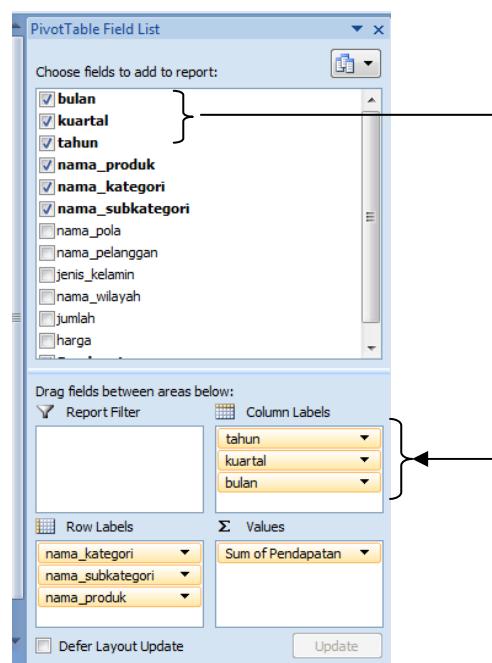
operasi **Roll Up** berfungsi untuk melihat data secara lebih umum, sedangkan operasi **Drill Down** untuk melihat data secara lebih spesifik dan terperinci.

Berikut adalah langkah-langkah untuk melakukan operasi tersebut :

1. Buka Sheet1 (hasil pivot table) dan letakkan kursor pada area pivot table.
2. Pada kotak **PivotTable Field List**, hilangkan tanda cek pada field **jumlah** (field ini sementara tidak digunakan), dan beri tanda cek pada field **field** (kolom) yang akan ditampilkan ke dalam *cube*.
3. Beri tanda cek dan letakkan field-field berikut pada kotak **Row Labels** atau **Column Labels** sesuai dengan kebutuhan tampilan *cube*. Urutan field dalam kotak ini menentukan urutan rincian kategori data. Field yang terletak pada urutan teratas merupakan field dengan kategori paling umum, sedangkan field yang terletak pada urutan terbawah adalah field dengan kategori paling spesifik (paling rinci).
4. Misalkan pada Row Labels akan ditampilkan data berdasarkan urutan **nama_kategori**, **nama_subkategori**, dan **nama_produk**. Beri tanda cek pada field tersebut (bisa *drag and drop*) dan letakkan pada kotak **Row Labels**.



5. Pada Column Labels akan ditampilkan data berdasarkan urutan **tahun**, **kuartal**, dan **bulan**. Beri tanda cek pada field tersebut (*drag and drop*) dan letakan pada kotak **Column Labels**.



6. Lihat kembali pada *cube* setelah ditambahkan field-field untuk operasi *roll up* dan *drill down*.
 7. Pada masing-masing **Row Labels** dan **Column Labels** telah bertambah field-field yang bisa diperinci dan diringkas sesuai urutan kategori data yang lebih spesifik.

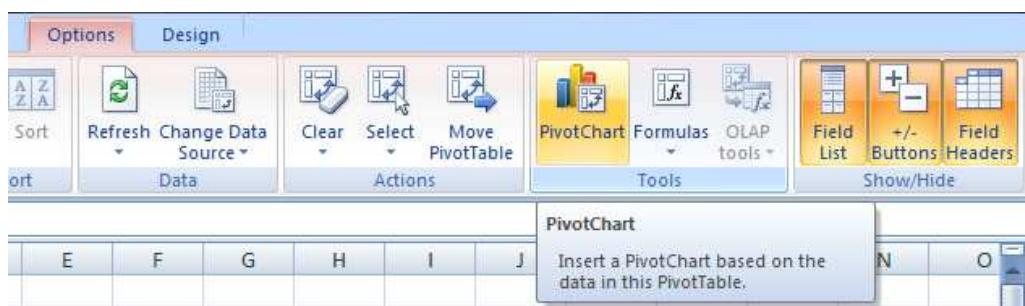
Sum of Pendapatan												2012 Total
	Column Labels											
	tahun											
	kuartal											
	bulan											
3	2012											
4	+ 1											
5	+ 2											
6	2 Total											
7	+ 3											
8	+ 4											
9	Row Labels											
10	Bahan											
11	Lawasan											
12	Bahna Lawasan Tulis Toilet											
13	Standar											
14	Bahan Standar Cap Garis											
15	Batik											
16	Celana											
17	Hem											
18	Jam											
19	Jarik											
20	Kaos											
21	Rok											
22	Rok											
23	Baju Atasan											
24	Grand Total											
25	17355000	0	0	0	1925000	3588000	130000	0	0	0	16254000	1050000
26											225000	185,703,000

8. Klik tanda  untuk melakukan operasi **Roll Up** dan klik tanda  untuk melakukan operasi **Drill Down**.

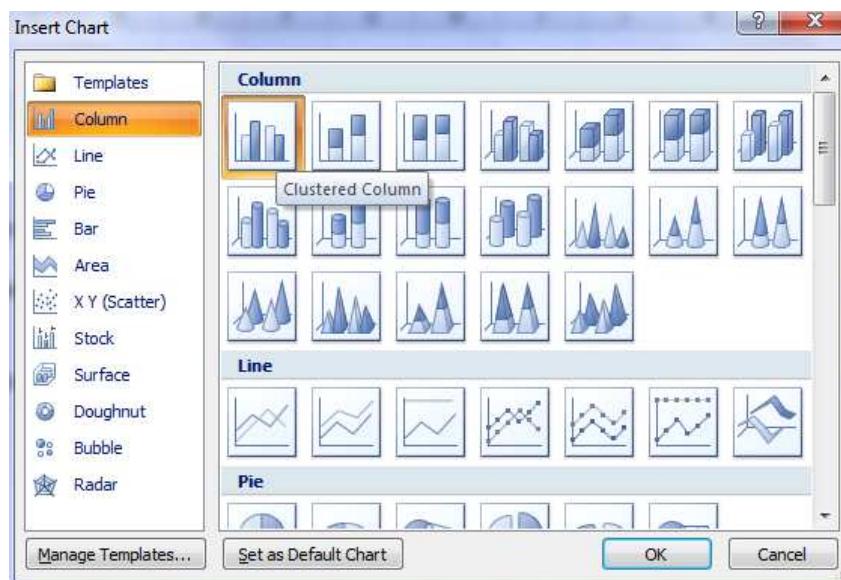
D.5. Kegiatan 5 : Menggunakan Pivot Chart

PivotChart merupakan sebuah cara untuk menampilkan *cube* dalam bentuk grafik. Dengan menggunakan grafik, sebuah pola atau statistik dari transaksi dalam waktu tertentu dapat dilihat dengan mudah dan dapat diketahui secara cepat. Selain itu, laporan-laporan dalam bentuk grafik sangat diperlukan untuk sebagai bahan dasar penentuan suatu kebijakan bagi para pengambil keputusan. Berikut adalah langkah-langkah untuk melakukan operasi tersebut:

1. Arahkan kursor pada area pivot table dalam Sheet1 (Hasil PivotTable).
2. Pada menu Option, klik PivotChart.

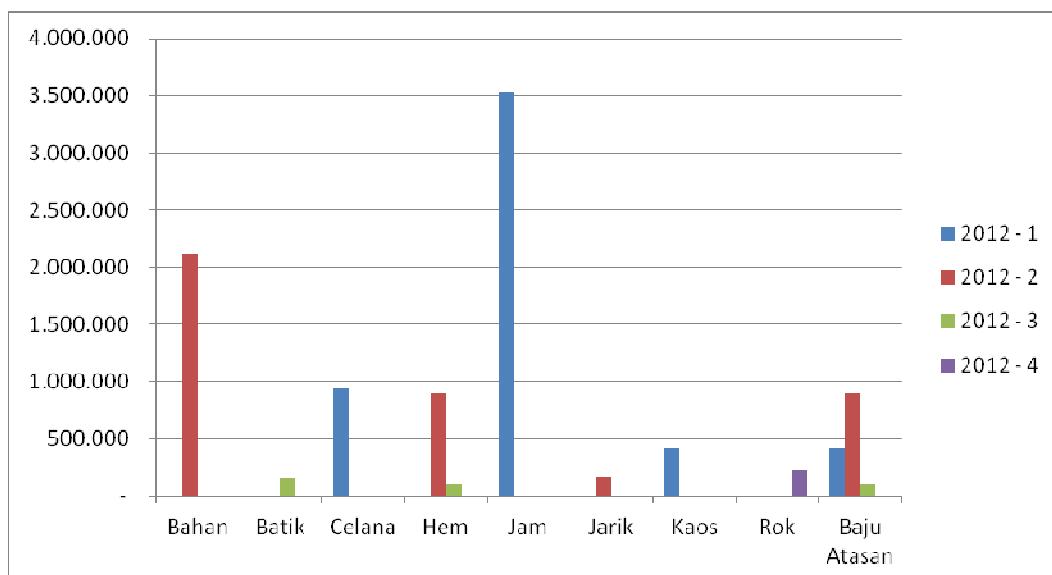


3. Pada jendela Insert Chart, pilih bentuk grafik yang diinginkan. Misalkan pilih grafik dalam bentuk batang, maka klik Clustered Column. Kemudian Klik OK.

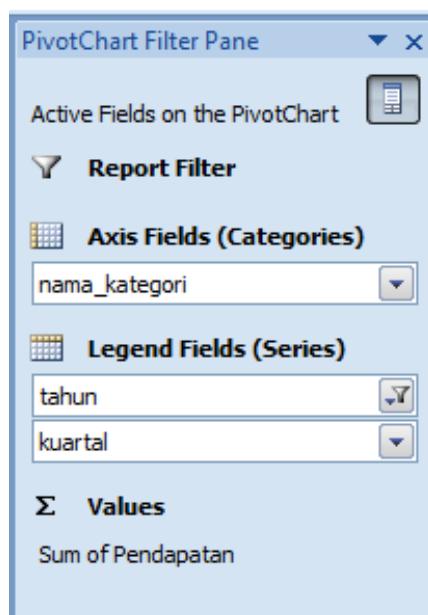


4. Grafik akan ditampilkan dengan sumbu X dan sumbu Y menyesuaikan dengan Row Labels dan Column Labels.
5. Jika grafik terlalu rinci, maka bisa dibuat secara lebih umum dengan menghilangkan kembali tanda cek pada field dalam **PivotTable Field**

List. Misalkan hilangkan tanda cek pada field **nama_produk**, **nama_subkategori**, dan **bulan**.



6. Dengan melihat grafik PivotChart, pola transaksi dari kuartal pertama hingga kuartal keempat dapat dilihat dengan mudah apakah terjadi kenaikan, penurunan atau stabil untuk masing-masing kategori produk.
7. Jendela PivotChart Filter Pane berfungsi untuk menyaring (*filter*) data-data khusus yang akan ditampilkan saja.



E. Tugas

1. Dengan menggunakan **PivotTable** pada file **Fakta_Penjualan.xls** tambahkan 2 buah field, yaitu :
 - a. **PPN** (Pajak Pertambahan Nilai) sebesar 10% dari tiap pendapatan pada Pivot Table.
 - b. **Total Penghasilan** yang dihitung dari pendapatan dikurangi dengan PPN tersebut.

		Column Labels					
			2012				
	Row Labels	Sum of Pendapatan	Sum of PPN (10%)	Sum of Total Penghasilan	Total Sum of Pendapatan	Total Sum of PPN (10%)	Total Sum of Total Penghasilan
6	Bahan	2.120.000	212.000	1.908.000	2.120.000	212.000	1.908.000
7	Batik	150.000	15.000	135.000	150.000	15.000	135.000
8	Celana	935.000	93.500	841.500	935.000	93.500	841.500
9	Hem	1.596.000	159.600	1.436.400	1.596.000	159.600	1.436.400
10	Jam	3.520.000	352.000	3.168.000	3.520.000	352.000	3.168.000
11	Jarik	160.000	16.000	144.000	160.000	16.000	144.000
12	Kaos	420.000	42.000	378.000	420.000	42.000	378.000
13	Rok	225.000	22.500	202.500	225.000	22.500	202.500
14	Baju Atasan	7.722.000	772.200	6.949.800	7.722.000	772.200	6.949.800
15	Grand Total	185.703.000	18.570.300	167.132.700	185.703.000	18.570.300	167.132.700

2. Buatlah **PivotTable** dan **PivotChart** untuk melihat PPN dan Total Penghasilan tersebut selama tahun 2010 – 2012. Kategori produk apakah yang memberikan nilai penghasilan terbanyak selama 3 tahun tersebut?

MODUL 6

PENGENALAN DATA MINING

A. Tujuan

1. Mahasiswa dapat memahami tujuan data mining.
2. Mahasiswa memahami dan mengenal beberapa software aplikasi yang dapat digunakan untuk data mining.

B. Landasan Teori

B.1. Pengertian data mining

Data mining adalah serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basisdata. Informasi yang dihasilkan diperoleh dengan cara mengekstraksi dan mengenali pola yang penting atau menarik dari data yang terdapat dalam basisdata.

Hal-hal yang melatarbelakangi datamining:

1. Melimpahnya data (*overload data*) yang dialami oleh berbagai institusi, perusahaan atau organisasi. Merlimpahnya data ini merupakan akumulasi data transaksi yang terekam bertahun-tahun.
2. Data-data tersebut merupakan data transaksi yang umumnya diproses menggunakan aplikasi komputer yang biasa disebut dengan OLTP (*On Line Transaction Processing*).
3. Adanya ledakan informasi (*explosion information*) dari berbagai media terutama internet yang secara umum informasi tersebut tidak terstruktur (*unstructured information*).

B.2. Manfaat penggunaan data mining

1) Sudut pandang komersial

Data mining dapat digunakan dalam menangani meledaknya volume data. Bagaimana menyimpannya, mengestraknya serta memanfaatkannya. Berbagai teknik komputasi dapat digunakan untuk menghasilkan informasi yang dibutuhkan. Informasi yang dihasilkan menjadi aset untuk meningkatkan daya

saing suatu institusi. Data mining tidak hanya digunakan untuk menangani persoalan menumpuknya data / informasi dan bagaimana menyimpannya tanpa kehilangan informasi yang penting (*warehousing*). Data mining juga diperlukan untuk menyelesaikan permasalahan atau menjawab kebutuhan bisnis itu sendiri, sebagai contoh:

- a. Bagaimana mengetahui hilangnya pelanggan karena pesaing
- b. Bagaimana mengetahui item produk atau konsumen yang memiliki kesamaan karakteristik
- c. Bagaimana mengidentifikasi produk-produk yang terjual bersamaan dengan produk lain.
- d. Bagaimana memprediski tingkat penjualan
- e. Bagaimana menilai tingkat resiko dalam menentukan jumlah produksi suatu item.
- f. Bagaimana memprediksi perilaku bisnis di masa yang akan datang.

2) Sudut pandang keilmuan

Data mining dapat digunakan untuk meng-*capture*, menganalisis serta menyimpan data yang bersifat *real-time* dan sangat besar, misalnya:

- a. Remote sensor yang ditempatkan pada suatu satelit
- b. Telescope yang digunakan untuk memindai langit
- c. Simulasi saintifik yang membangkitkan data dalam ukuran terabytes

Data mining merupakan salah satu metode alternatif yang dapat digunakan untuk mengolah data mentah, ketika metode konvensional tidak mungkin untuk dilakukan karena besarnya volume data yang diolah. Hal ini dapat terjadi karena data mining memiliki kemampuan mereduksi data baik melalui teknik katalogisasi, klasifikasi maupun segementasi.

B.3. Proses data mining

Data mining sebenarnya merupakan salah satu rangkaian dari proses pencarian pengetahuan dalam database (*Knowledge Discovery in Database (KDD)*). KDD berhubungan dengan teknik integrasi dan penemuan ilmiah,

interpretasi 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 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 *data warehouse*.
- 2) Seleksi dan transformasi data (*selection and transformation*). Data yang terdapat dalam database *data warehouse* kemudian direduksi dengan berbagai teknik. Proses reduksi diperlukan untuk mendapatkan hasil yang lebih akurat dan mengurangi waktu komputasi terutama untuk masalah dengan skala besar (*large scale problem*). Beberapa cara seleksi, antara lain:
 - a. *Sampling*, yaitu seleksi subset representatif dari populasi data yang besar.
 - b. *Denoising*, yaitu proses menghilangkan noise dari data yang akan ditransformasikan.
 - c. *Feature extraction*, yaitu proses membuka spesifikasi data yang signifikan dalam konteks tertentu.

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

- a. *Centering*, mengurangi setiap data dengan rata-rata dari setiap atribut yang ada.
- b. *Normalisation*, membagi setiap data yang di *centering* dengan standar deviasi dari atribut bersangkutan.
- c. *Scaling*, mengubah data sehingga berada dalam skala tertentu.

3) Penambangan data (*data mining*)

Data-data yang telah diseleksi dan ditransformasi kemudian 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

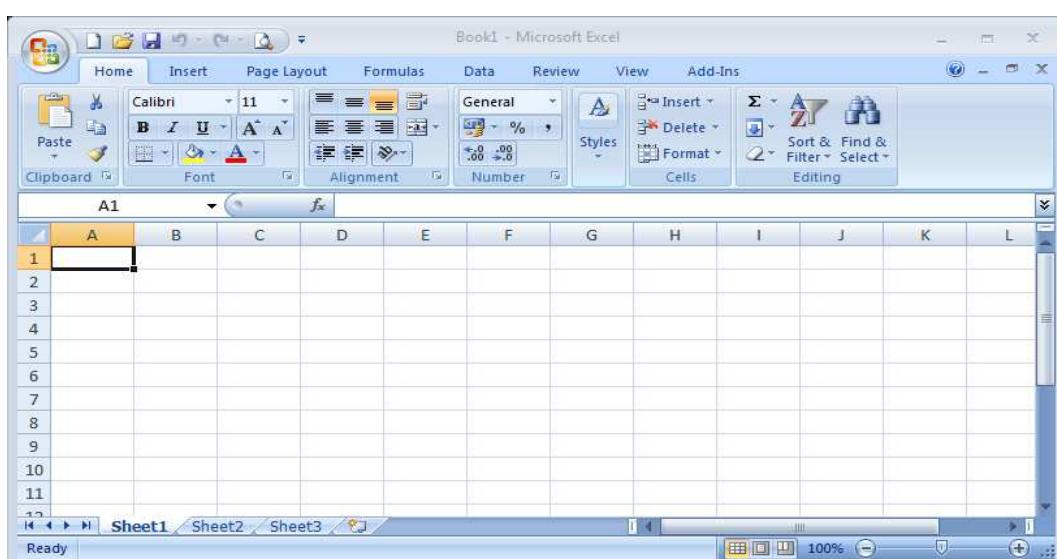
Tahap ini merupakan bagian dari proses pencarian pengetahuan yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya. Langkah terakhir KDD adalah mempresentasikan pengetahuan dalam bentuk yang mudah dipahami oleh pengguna.

C. Alat dan Bahan

1. Komputer dengan sistem operasi Windows.
2. Aplikasi Microsoft Excel, Notepad, Weka, RapidMiner.
3. Modul Praktikum Data Warehousing dan Data Mining.

D. Pengenalan Perangkat Lunak Data Mining

Perangkat Lunak 1. Microsoft Excel



Gambar 1.1 Lembar Kerja Microsoft Excel

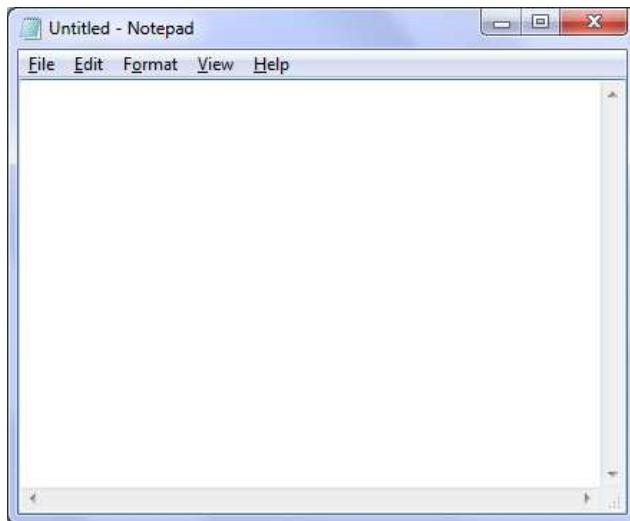
Microsoft Excel atau Microsoft Office Excel adalah sebuah program aplikasi lembar kerja spreadsheet yang dibuat dan didistribusikan oleh Microsoft Corporation untuk sistem operasi Microsoft Windows dan Mac OS. Aplikasi ini memiliki fitur kalkulasi dan pembuatan grafik yang, dengan menggunakan strategi marketing Microsoft yang agresif, menjadikan Microsoft Excel sebagai salah satu program komputer yang populer digunakan di dalam komputer mikro hingga saat ini. Bahkan, saat ini program ini merupakan program spreadsheet paling banyak digunakan oleh banyak pihak, baik di platform PC berbasis Windows maupun platform Macintosh berbasis Mac OS, semenjak versi 5.0 diterbitkan pada tahun 1993. Aplikasi ini merupakan bagian dari Microsoft Office System.

Excel merupakan program *spreadsheet* pertama yang mengizinkan pengguna untuk mendefinisikan bagaimana tampilan dari *spreadsheet* yang mereka sunting: font, atribut karakter, dan tampilan setiap sel. Excel juga menawarkan penghitungan kembali terhadap sel-sel secara cerdas, di mana hanya sel yang berkaitan dengan sel tersebut saja yang akan diperbarui nilanya (di mana program-program *spreadsheet* lainnya akan menghitung ulang keseluruhan data atau menunggu perintah khusus dari pengguna). Selain itu, Excel juga menawarkan fitur pengolahan grafik yang sangat baik.

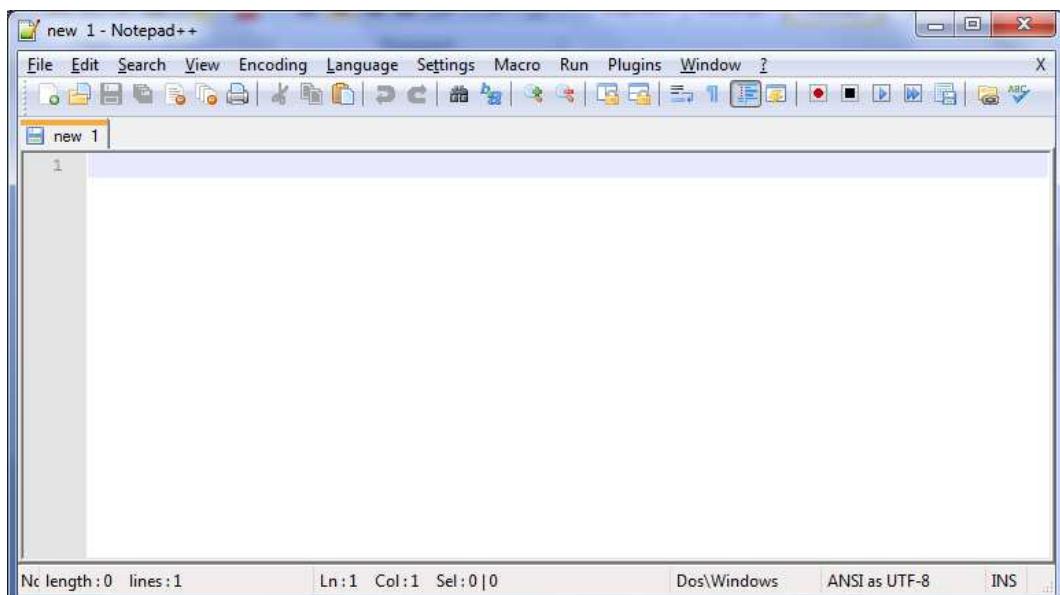
Perangkat Lunak 2. Notepad / Notepad++

Notepad / Notepad++ adalah sebuah penyunting teks dan penyunting kode sumber yang berjalan di sistem operasi Windows. Notepad adalah salah satu program bawaan dari Windows yang biasa digunakan untuk menulis keterangan-keterangan yang penting dari program aplikasi seperti halnya lisensi program atau yang lainnya. Notepad juga bisa berguna untuk berbagai macam keperluan, seperti membuat file CSS, Javascript untuk format web, dan pembuatan listing pemrograman, seperti Java dan berbagai kegunaan lain. Keuntungan dari penggunaan Notepad adalah kecepatan dan kemudahan dalam pengoperasianya. Sedangkan dari segi kelemahan, Notepad tidak memiliki tampilan yang menarik. Sementara itu, Notepad++ menggunakan komponen Scintilla untuk dapat menampilkan dan menyuntingan teks dan berkas kode sumber berbagai bahasa pemrograman. Seperti halnya dengan Notepad bawaan Windows, Notepad++ juga

bisa digunakan untuk berbagai keperluan pemrograman seperti membuat file CSS, Javascript untuk format web, dan pembuatan listing pemrograman, seperti Java dan berbagai kegunaan lain.



Gambar 1.2 Lembar Kerja Notepad



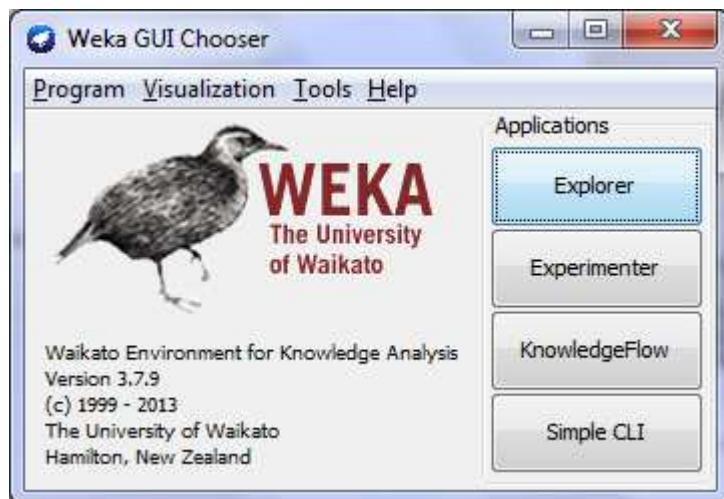
Gambar 1.3 Lembar Kerja Notepad++

Perangkat Lunak 3. Weka

WEKA adalah sebuah paket *tools machine learning* praktis. “WEKA” merupakan singkatan dari *Waikato Environment for Knowledge Analysis*, yang dibuat di Universitas Waikato, New Zealand untuk penelitian, pendidikan dan

berbagai aplikasi. WEKA mampu menyelesaikan masalah-masalah *data mining* di dunia-nyata, khususnya klasifikasi yang mendasari pendekatan-pendekatan *machine learning*. Perangkat lunak ini ditulis dalam hierarki *class Java* dengan metode berorientasi objek dan dapat berjalan hampir di semua *platform*.

Weka terdiri dari koleksi algoritma *machine learning* yang dapat digunakan untuk melakukan generalisasi / formulasi dari sekumpulan data sampel. Walaupun kekuatan Weka terletak pada algoritma yang makin lengkap dan canggih, kesuksesan data mining tetap terletak pada faktor pengetahuan manusia implementornya. Tugas pengumpulan data yang berkualitas tinggi dan pengetahuan pemodelan dan penggunaan algoritma yang tepat diperlukan untuk menjamin keakuratan formulasi yang diharapkan.



Gambar 1.4 Halaman depan Weka

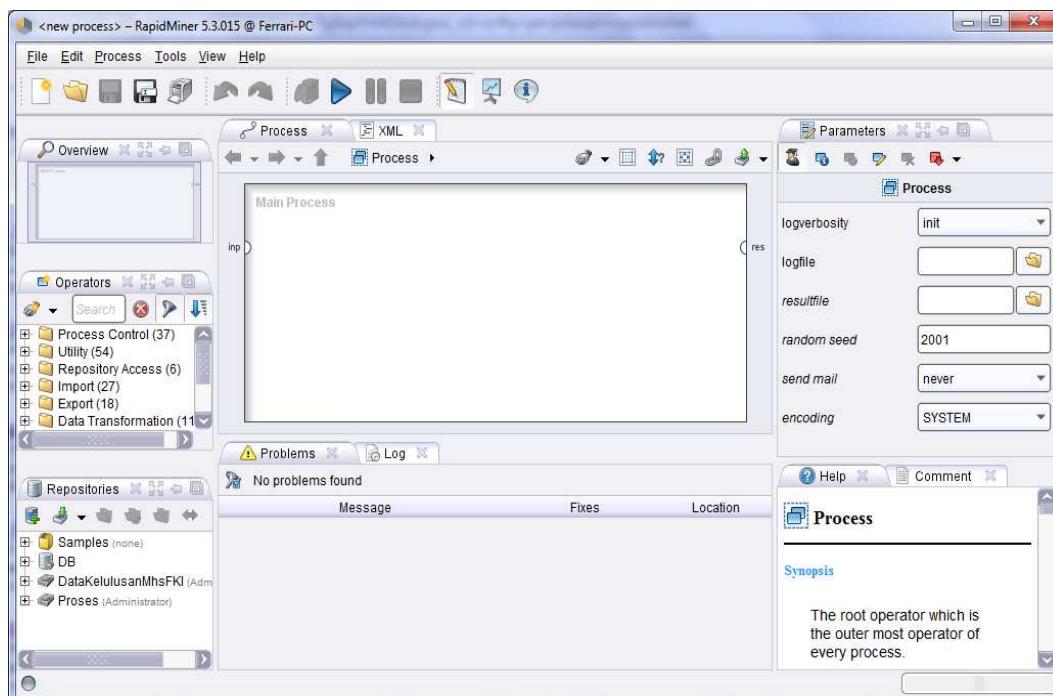
Empat tombol pada halaman depan Weka dapat digunakan untuk menjalankan aplikasi sebagai berikut:

- 1) **Explorer** digunakan untuk menggali lebih jauh data dengan aplikasi WEKA.
- 2) **Experimenter** digunakan untuk melakukan percobaan dengan pengujian statistik skema belajar.
- 3) **Knowledge Flow** digunakan untuk pengetahuan pendukung.
- 4) **Simple CLI** antar muka dengan menggunakan tampilan command-line yang memungkinkan langsung mengeksekusi perintah weka untuk Sistem Operasi yang tidak menyediakan secara langsung.

Perangkat Lunak 4. RapidMiner

RapidMiner adalah sebuah platform perangkat lunak yang dikembangkan oleh perusahaan yang menyediakan lingkungan secara terintegrasi untuk *machine learning*, data mining, *text mining*, analisis prediktif dan analisis bisnis. Aplikasi ini digunakan untuk kepentingan bisnis dan industri serta untuk penelitian, pendidikan, pelatihan, prototyping secara cepat, dan pengembangan aplikasi dan mendukung proses data mining termasuk hasil visualisasi, validasi dan optimasi. RapidMiner dikembangkan dari versi sebelumnya yang tersedia di bawah lisensi open source OSI-certified.

RapidMiner menyediakan prosedur data mining dan *machine learning* termasuk untuk proses ETL (*extraction, transformation, loading*), *data preprocessing*, visualisasi, modeling dan evaluasi. Proses data mining tersusun atas operator-operator yang *nestable*, yang dideskripsikan dengan XML dan dibuat dengan GUI. Aplikasi ini ditulis dalam bahasa pemrograman Java dan mengintegrasikan proyek data mining Weka dan statistika R.



Gambar 1.5 Perspektif RapidMiner

E. Tugas (Dikerjakan saat ini)

1. Dengan menggunakan Ms. Excel, buatlah tabel berikut.

Jurusan_SMA	Gender	Asal_Sekolah	Rerata_SKS	Asisten	Lama_Studi
IPS	WANITA	SURAKARTA	18	TIDAK	TERLAMBAT
IPA	PRIA	SURAKARTA	19	YA	TEPAT
LAIN	PRIA	SURAKARTA	19	TIDAK	TERLAMBAT
IPA	PRIA	LUAR	17	TIDAK	TERLAMBAT
IPA	WANITA	SURAKARTA	17	TIDAK	TEPAT
IPA	WANITA	LUAR	18	YA	TEPAT
IPA	PRIA	SURAKARTA	18	TIDAK	TERLAMBAT
IPA	PRIA	SURAKARTA	19	TIDAK	TEPAT
IPS	PRIA	LUAR	18	TIDAK	TERLAMBAT
LAIN	WANITA	SURAKARTA	18	TIDAK	TEPAT
IPA	WANITA	SURAKARTA	19	TIDAK	TEPAT
IPS	PRIA	SURAKARTA	20	TIDAK	TEPAT
IPS	PRIA	SURAKARTA	19	TIDAK	TEPAT
IPA	PRIA	SURAKARTA	19	TIDAK	TEPAT
IPA	PRIA	LUAR	22	YA	TEPAT
LAIN	PRIA	SURAKARTA	16	TIDAK	TERLAMBAT
IPS	PRIA	LUAR	20	TIDAK	TEPAT
LAIN	PRIA	LUAR	23	YA	TEPAT
IPA	PRIA	SURAKARTA	21	YA	TEPAT
IPS	PRIA	SURAKARTA	19	TIDAK	TERLAMBAT

2. Dengan menggunakan formula dalam Ms. Excel, carilah:
 - a) Pada atribut Jurusan_SMA, berapa jumlah data masing-masing kelas IPA, IPS dan LAIN? (gunakan formula = COUNTIF)
 - b) Pada atribut Lama_Studi, berapa jumlah data masing-masing kelas TEPAT, TERLAMBAT?
 - c) Pada atribut Rerata_SKS, berapa nilai Max, Min, Mean, dan *Standard Deviation*?
 - d) Pada tabel tersebut, berapakah jumlah data gabungan untuk kelas pada atribut Jurusan_SMA = IPA, Gender = PRIA, Asisten = YA, Lama_studi = TEPAT? (gunakan formula = COUNTIFS)
3. Dikerjakan di rumah, instal 4 buah perangkat lunak dalam modul ini pada komputer anda!

MODUL 7

DATA PREPROCESSING

A. Tujuan

1. Mahasiswa mampu menyebutkan tipe-tipe data yang digunakan dalam data mining.
2. Mahasiswa mampu menjelaskan permasalahan kualitas data dan penyelesaiannya.
3. Mahasiswa mampu melakukan data preprocessing.

B. Landasan Teori

Sebelum diproses data mining sering kali diperlukan preprocessing. Data preprocessing menerangkan tipe-tipe proses yang melaksanakan data mentah untuk mempersiapkan proses prosedur yang lainnya. Tujuan preprosesing dalam data mining adalah mentransformasi data ke suatu format yang prosesnya lebih mudah dan efektif untuk kebutuhan pemakai, dengan indikator sebagai berikut:

- 1) Mendapatkan hasil yang lebih akurat
- 2) Pengurangan waktu komputasi untuk *large scale problem*.
- 3) Membuat nilai data menjadi lebih kecil tanpa merubah informasi yang dikandungnya.

Terdapat beberapa alat dan metode yang berbeda yang digunakan untuk preprocessing seperti :

- 1) *Sampling*, menyeleksi subset representatif dari populasi data yang besar.
- 2) *Transformation*, memanipulasi data mentah untuk menghasilkan input tunggal.
- 3) *Denoising*, menghilangkan noise dari data.
- 4) *Normalization*, mengorganisasi data untuk pengaksesan yang lebih spesifik.
- 5) *Feature extraction*, membuka spesifikasi data yang signifikan.

Menurut Susanto (2007), alasan-alasan harus dilakukan data preprocessing adalah antara lain:

- 1) Data mentah yang ada sebagian besar kotor, seperti:
 - a) Tidak lengkap, yaitu berisi data yang hilang / kosong, kekurangan atribut yang sesuai, atau hanya berisi data aggregate
 - b) Banyak noise, yaitu berisi data yang *outlier*, atau berisi data error
 - c) Tidak konsisten, yaitu berisi nilai yang berbeda dalam suatu kode atau nama.
- 2) Data yang tidak berkualitas, akan menghasilkan kualitas mining yang tidak baik pula.
- 3) Data preprocessing, cleaning, dan transformasi merupakan pekerjaan mayoritas dalam aplikasi data mining.

Untuk mengatasi masalah data tersebut, dilakukan preprocessing terhadap data sebelum diolah dengan data mining. Preprocessing dapat dilakukan dengan beberapa teknik yaitu:

- 1) **Cleaning**, memperkecil jumlah data yang hilang atau berbeda, dapat dilakukan dengan cara:
 1. Mengisi data yang hilang dengan *default value*.
 2. Mengisi data secara manual, misal: trace ulang transaksi untuk mengetahui data yang hilang.
 3. Mengisi dengan rata-rata atribut tersebut, misal: gaji pegawai yang kosong diisi dengan rata-rata gaji pegawai.
 4. Mengisi dengan rata-rata suatu atribut untuk kelas yang sama, misal: gaji pegawai yang kosong diisi dengan rata-rata gaji pegawai yang memiliki jabatan yang sama.
 5. Menggunakan regresi, prediksi berdasarkan dua variabel yang lain, misal: mengisi gaji pegawai yang kosong dengan nilai prediksi dengan regresi berdasarkan jabatan dan lama masa kerja.
 6. Menghilangkan baris yang mengandung data yang hilang. misal: menghilangkan data pegawai yang gaji pegawainya kosong.
 7. *Binning by means*, menggunakan rata-rata pengelompokan. misal: sorted data dibagi menjadi beberapa kelompok, dan dicari rata-rata masing-masing kelompok untuk mengganti setiap data yang ada, sesuai dengan

kelompoknya. Misal data dari kelompok A diganti dengan rata-rata kelompok A.

8. *Binning by range boundaries*, menggunakan batas terdekat suatu kelompok data, misal: sorted data dibagi menjadi beberapa kelompok, dicari nilai minimum dan maximum dari masing-masing kelompok, lalu gantikan tiap nilai di suatu kelompok dengan batas atas atau batas bawah kelompoknya, sesuai dengan yang paling dekat.
 9. Mencari dan menghilangkan outlier dengan pengelompokan atau regresi
 10. *Binning*, mengganti suatu nilai outlier dengan nilai yang lebih sesuai dengan data lain yang ada di sekitar data outlier tersebut (*local smoothing*).
- 2) **Integrasi**, menggabungkan beberapa sumber data sehingga dapat saling melengkapi. Data perlu digabungkan dengan key yang sesuai. Key ini mungkin memiliki nama yang berbeda di sumber data yang berbeda. Misal di tabel a menggunakan nama atribut ‘id’, di tabel b menggunakan nama atribut ‘nomor’, atau satuan yang digunakan untuk konsep yang sama (misal harga) disimpan dalam juta dan ribu.
- 3) **Transformasi**, mengubah data yang kompleks dengan tidak menghilangkan isi, sehingga lebih mudah diolah, dilakukan dengan cara:
1. *Smoothing (binning, clustering dan regresi)*.
 2. Agregasi (*summarize*, menggunakan dimensi yang lebih general (*cube construction*)).
 3. Generalisasi, misal menggunakan dimensi propinsi daripada kabupaten atau *grouping* (hirarki konsep).
 4. Normalisasi, mengelompokkan data sesuai skala tertentu, misal IPK.
 - a) Normalisasi min-max, standarisasi data dengan menempatkan data dalam range 0 sampai 1, nilai terkecil sebagai 0, dan nilai terbesar sebagai 1. nilai baru = $((\text{nilai lama} - \text{nilai minimal}) / (\text{nilai maksimal} - \text{nilai minimal})) \times (\text{range maksimal} - \text{range minimal}) + \text{range minimal}$. range minimal = 0, range maksimal = 1.
 - b) Normalisasi z-index, nilai baru = $(\text{nilai lama} - \text{rata-rata})/\text{standar deviasi}$.

- c) Normalisasi skala desimal, nilai baru = nilai lama / 10^x ,
- 4) **Diskretisasi**, membagi nilai data menjadi beberapa range data, dilakukan dengan cara:
1. Binning, seperti prosedur **Cleaning** 1.10.
 2. Hirarki konsep, misal mengelompokkan harga produk menjadi, mahal, biasa, murah.
- 5) **Reduksi**, mengurangi jumlah data sehingga resource yang digunakan lebih sedikit, sehingga prosesnya dapat lebih cepat dilakukan dengan cara:
1. Sampling/generalisasi,
 2. Agregasi, seperti agregasi pada transformasi. data ribuan memiliki volume byte yang lebih kecil daripada data jutaan
 3. Mengurangi atribut yang tidak perlu (korelasi yang rendah terhadap keseluruhan data), misal nomor telepon, nama ibu atau nama jalan. Jika data set memiliki atribut sejumlah n, maka ada 2^n (2 pangkat n) kemungkinan korelasi antar atribut kompresi data.

C. Alat dan Bahan

1. Komputer dengan sistem operasi Windows.
2. Program aplikasi Notepad++ / Notepad.
3. Program aplikasi Weka.
4. Modul Praktikum Data Warehousing dan Data Mining.

D. Langkah-langkah Praktikum

D.1. Membuat Format Data ARFF (Attribut-Relation File Format)

Teknik data mining dengan menggunakan aplikasi Weka memerlukan sebuah data dalam format ARFF. File ARFF adalah file yang digunakan Weka yang berisi 1 atau lebih *instances* dari atribut.

Hal terpenting dalam data mining jika menggunakan aplikasi Weka adalah membuat file ARFF / membuat instances sebelum diproses klasifikasi, kluster atau hal lainnya. Di dalam format ARFF komponen penting yang harus terdapat dalam file ada 3 macam yaitu:

- a) @RELATION, nama_relasi untuk menamakan relasi sebuah file ARFF,
- b) @ATTRIBUTE, untuk penamaan setiap atribut baik *numeric*, nominal, string dan *date*.
- c) @DATA, terdapat pada setiap line untuk merepresentasikan sebuah *instances*.

Untuk membuat data dan menyimpan dalam format ARFF bisa dengan menggunakan Notepad / Notepad++, dengan langkah-langkah berikut ini:

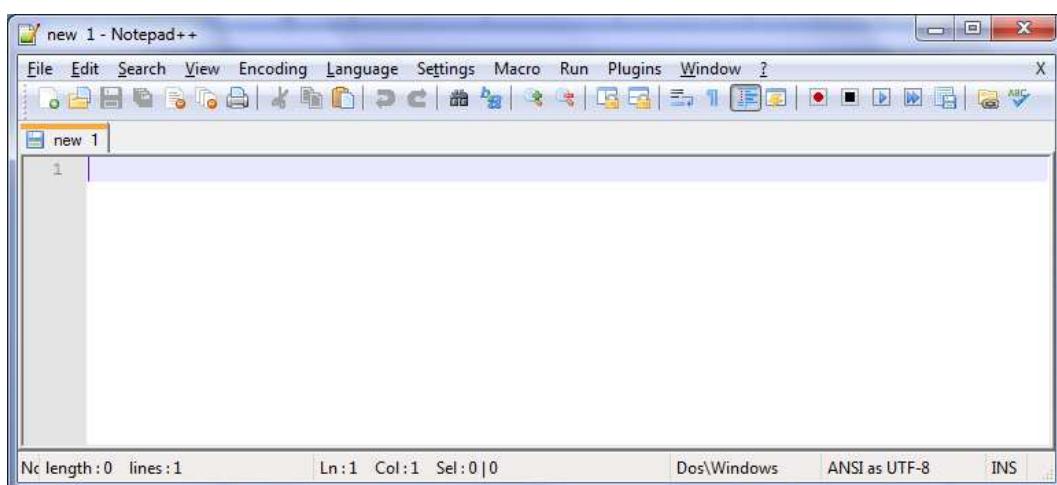
1. Kita ambil sebuah contoh data tentang kemungkinan seseorang akan bermain tenis berdasarkan kondisi cuaca, suhu, kelembaban udara, berangin seperti tabel berikut.

Nama_Tabel: Cuaca

Cuaca	Suhu	Kelembaban_Udara	Berangin	Bermain_Tenis
Cerah	85	85	TIDAK	TIDAK
Cerah	80	90	YA	TIDAK
Mendung	83	86	TIDAK	YA
Hujan	70	96	TIDAK	YA
Hujan	68	80	TIDAK	YA
Hujan	65	70	YA	TIDAK
Mendung	64	65	YA	YA
Cerah	72	95	TIDAK	TIDAK
Cerah	69	70	TIDAK	YA
Hujan	75	80	TIDAK	YA
Cerah	75	70	YA	YA
Mendung	72	90	YA	YA
Mendung	81	75	TIDAK	YA
Hujan	71	91	YA	TIDAK

2. Penentuan relasi / nama tabel. Misalkan kita tentukan nama relasinya adalah **Cuaca**.
3. Penentuan atribut / variabel. Berdasarkan tabel cuaca tersebut, tentukan variabel-variabel independen / bebas (X) dan juga variabel dependen / terikat (Y). Variabel independen adalah variabel yang tidak dipengaruhi oleh variabel lainnya. Sedangkan variabel dependen adalah variabel yang nilainya dipengaruhi oleh variabel lainnya.

4. Dalam contoh ini, variabel Bermain_Tenis kita jadikan variabel dependen yang akan dicari nilainya berdasarkan kondisi variabel lainnya. Berikut ketentuan variabel / atribut berdasarkan tabel Cuaca.
 - a) Variabel Y = Bermain_Tenis
 - b) Variabel X1 = Cuaca
 - c) Variabel X2 = Suhu
 - d) Variabel X3 = Kelembaban_Udara
 - e) Variabel X4 = Berangin
5. Penentuan data. Dalam format ARFF, ada 3 jenis data yaitu binomial, polynomial, dan real. Binomial, jika nilai data hanya ada 2 kemungkinan (Contoh: YA dan TIDAK, 0 dan 1, PRIA dan WANITA). Polynomial, jika nilai data lebih dari 2 kemungkinan namun tidak terlalu banyak (Contoh: Cerah, Mendung dan Hujan). Real, jika kemungkinan nilai data sangat beragam dan jumlahnya banyak (Contoh: suhu udara, kecepatan, jarak dan lain-lain) meskipun bisa diringkas menjadi polynomial dengan menggunakan interval sesuai kebutuhan data mining.
6. Buka aplikasi Notepad++ atau Notepad bawaan Windows, cukup ketikkan notepad pada Start Menu.



7. Ketikkan 3 komponen utama dalam format file ARFF yaitu @relation, @attribute, dan @data pada jendela notepad.

The screenshot shows a Notepad++ window titled "new 1 - Notepad++". The code in the editor is as follows:

```
1 @relation
2
3 @attribute
4
5 @data
```

At the bottom of the window, status bars show "Nc length : 32 lines : 5", "Ln : 5 Col : 6 Sel : 0 | 0", "Dos\Windows", "ANSI as UTF-8", and "INS".

8. Ketikkan nama relasi **cuaca** di sebelah **@relation**.
9. Ketik **nama atribut** dan **tipe data** masing-masing atribut di sebelah **@atribut** sesuai jumlah atribut yang digunakan. Jika nilai data bertipe binomial atau polynomial, semua kemungkinan nilai datanya disebutkan dan diletakkan dalam tanda kurung kurawal. Jika nilai data bertipe real, maka cukup tuliskan **real**.
10. Ketik **nilai data** di bawah **@data** untuk tiap baris tabel sesuai dengan urutan atributnya dipisah menggunakan tanda koma. (Catatan: huruf besar dan kecil sangat berpengaruh, sehingga kemungkinan nilai data pada **@attribute** harus sama dengan nilai data pada **@data**).

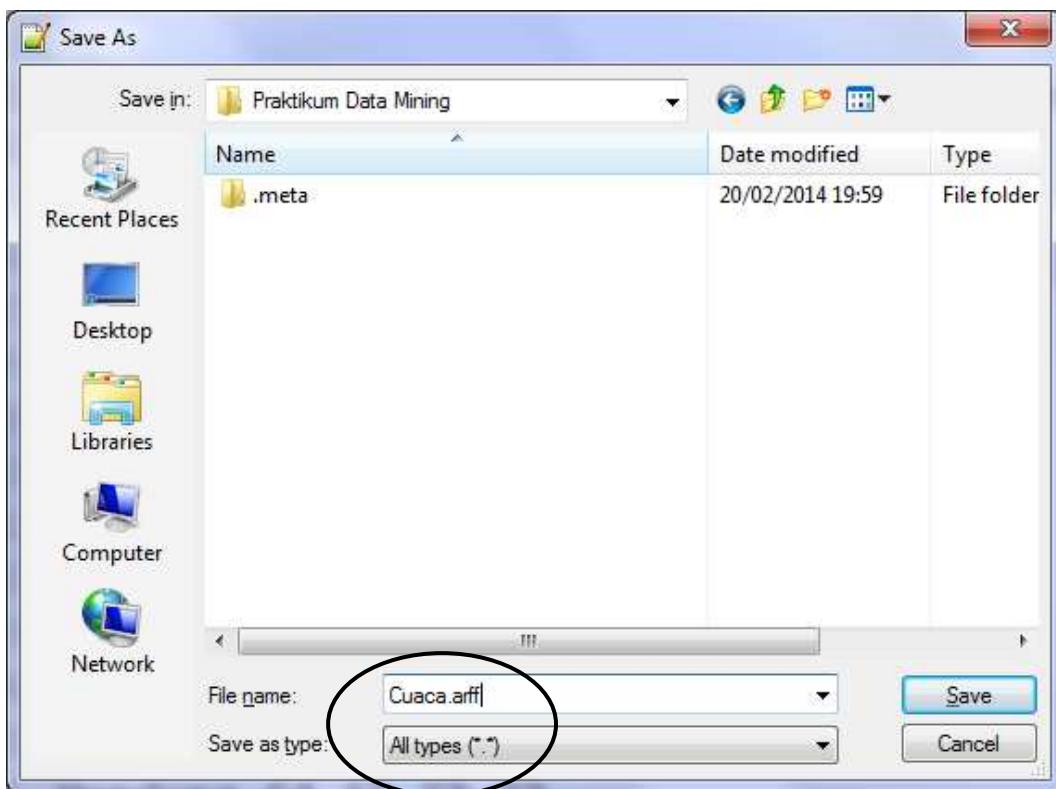
The screenshot shows a Notepad++ window titled "new 1". The code in the editor is as follows:

```
1 @relation Cuaca
2
3 @attribute Cuaca {Cerah, Mendung, Hujan}
4 @attribute Suhu real
5 @attribute Kelembaban_Udara ...
6 @attribute ..... {YA, TIDAK}
7 @attribute Bermain_Tenis {YA, TIDAK}
8
9 @data
10 Cerah,85,85,TIDAK,TIDAK
11 Cerah,80,90,YA,TIDAK
12 Mendung,83,86,TIDAK,YA
13 Hujan,70,96,TIDAK,YA
14 .....
15 .....
16 .....
```

At the bottom of the window, status bars show "Nc length : 312 lines : 16", "Ln : 16 Col : 8 Sel : 0 | 0", "Dos\Windows", "ANSI as UTF-8", and "INS".

Pada titik-titik silakan diisi dan dilanjutkan sesuai dengan tabel Cuaca.

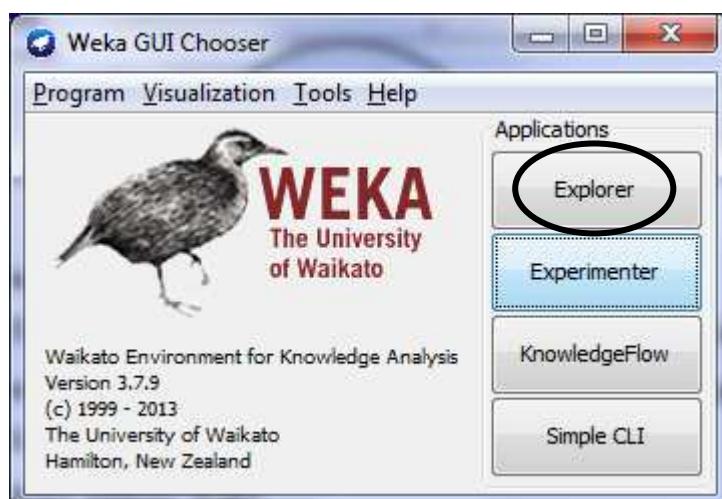
11. Simpan file dengan nama Cuaca.arff. Pastikan untuk memilih **All types (*.*)** pada pilihan **Save as type**.



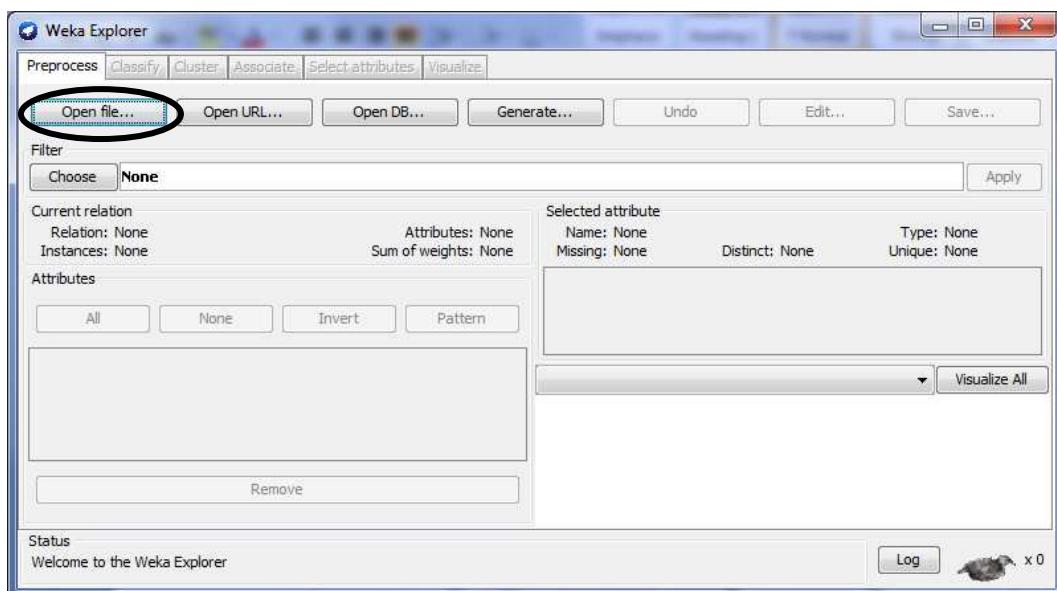
D.2. Menggunakan File ARFF dengan Weka.

File ARFF hasil dari kegiatan D.1, kemudian akan kita gunakan dalam Weka untuk proses lebih lanjut. Berikut langkah-langkah menggunakan file ARFF dalam Weka:

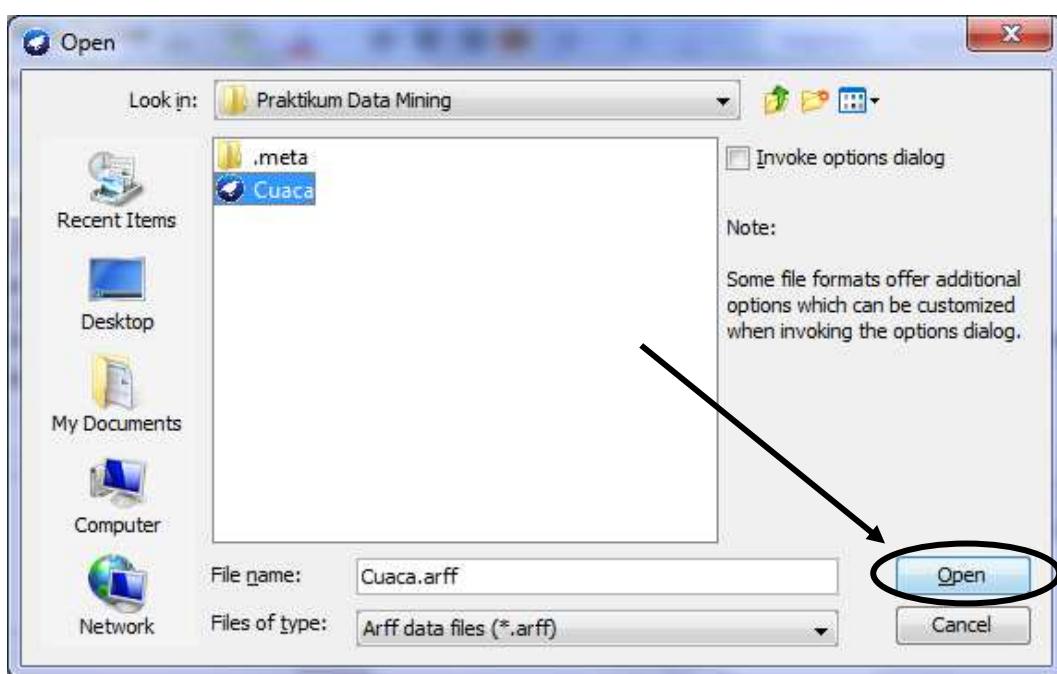
1. Buka aplikasi Weka.



2. Pilih menu Explorer sehingga akan muncul jendela Weka Explorer.

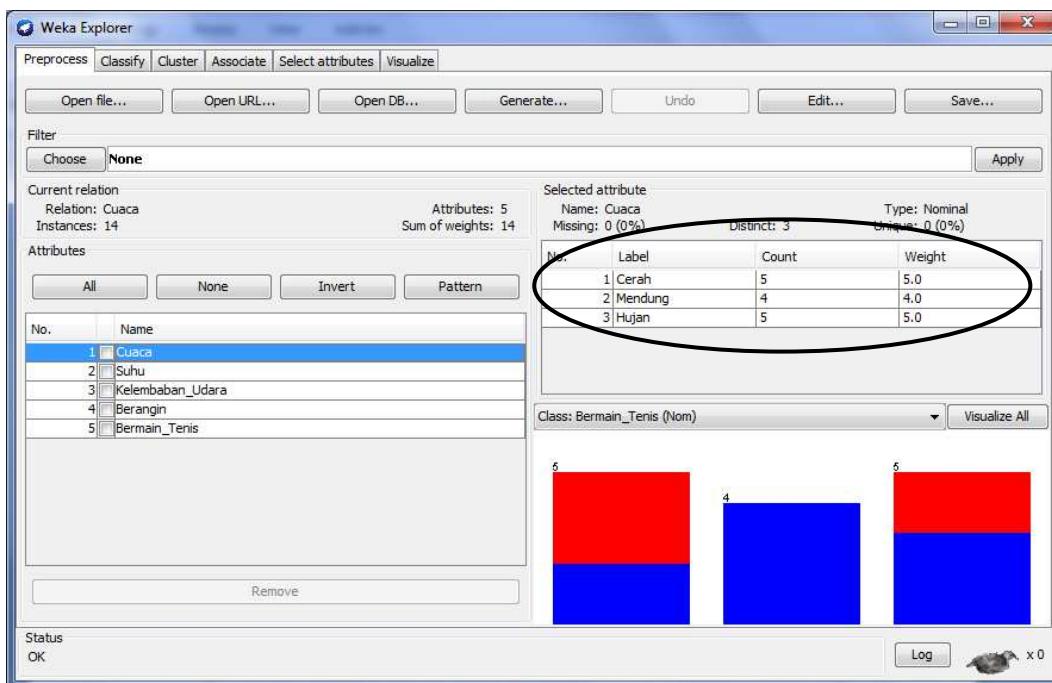


3. Pilih tombol Open File untuk membuka file ARFF yang telah dibuat. Klik Open.

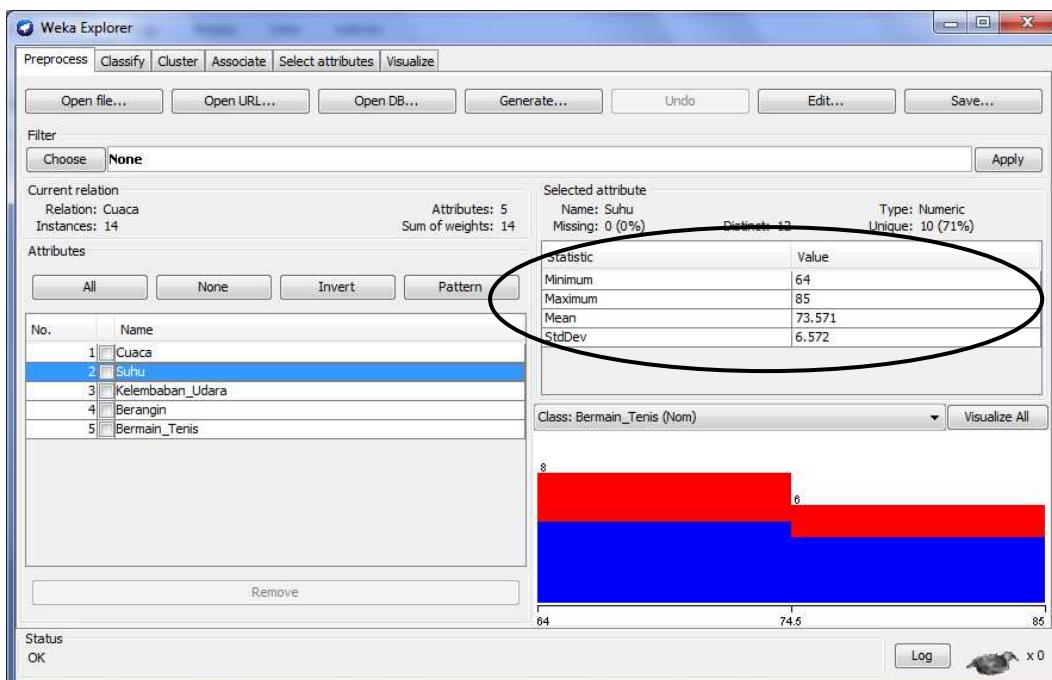


4. Sehingga dalam Weka Explorer akan tampak grafik statistik masing-masing atribut pada tabel Cuaca.

Berikut adalah gambar grafik untuk contoh data yang bertipe binomial atau polynomial.



Berikut ini gambar grafik untuk contoh data yang bertipe real. Mintalah penjelasan kepada dosen atau asisten tentang makna angka-angka dalam lingkaran tersebut.



E. Tugas (Dikerjakan saat ini)

1. Buat file ARFF berdasarkan tugas pada Modul 6 soal nomor 1!
2. Perlihatkan file ARFF dan juga gambar grafik untuk setiap data yang ditampilkan dalam Weka berdasarkan file ARFF anda kepada dosen dan asisten untuk dinilai!

Tulis di atas selembar kertas HVS (mintalah kepada asisten)

3. Berapa jumlah atribut yang bertipe binomial dan polynomial?
4. Berapa jumlah atribut yang bertipe real?
5. Pada atribut Rerata_SKS, berapakah besarnya nilai Maximum, Minimum, Mean, dan StdDev (*Standard Deviation*)?

MODUL 8

ALGORITMA NAÏVE BAYES

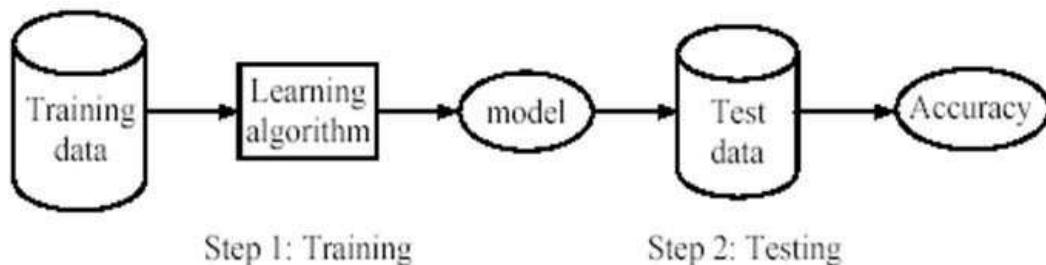
A. Tujuan

1. Mahasiswa mampu menggunakan dan membuat model klasifikasi dengan teorema Naïve Bayes
 2. Mahasiswa mampu menerapkan algoritma Naïve Bayes terhadap studi kasus tertentu

B. Landasan Teori

Teori keputusan Bayes adalah pendekatan statistik yang fundamental dalam pengenalan pola (*Pattern Recognition*). Pendekatan ini didasarkan pada kuantifikasi *trade-off* antara berbagai keputusan klasifikasi dengan menggunakan probabilitas dan resiko yang ditimbulkan dalam keputusan-keputusan tersebut.

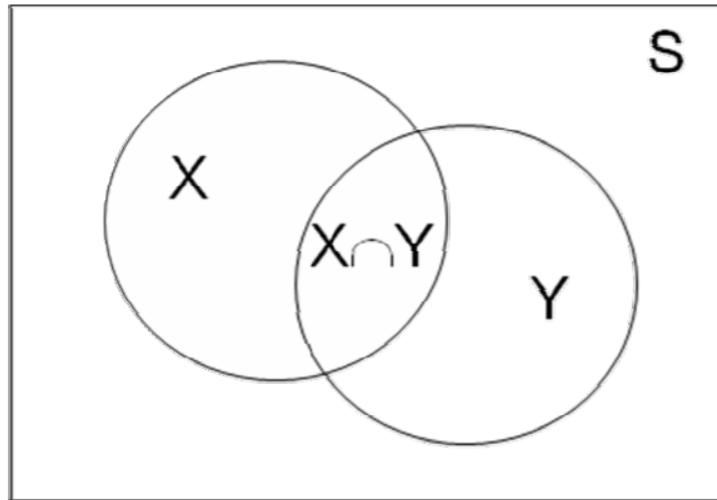
Ada dua proses penting yang dilakukan saat melakukan klasifikasi. Proses yang pertama adalah **learning (training)** yaitu proses pembelajaran menggunakan training set. Untuk kasus *Naïve Bayesian Classifier*, perhitungan probabilitas dari data berdasarkan data pembelajaran dilakukan. Proses yang kedua adalah proses **testing** yaitu menguji model menggunakan data testing. Gambar berikut memperlihatkan alur dari kedua proses tersebut.



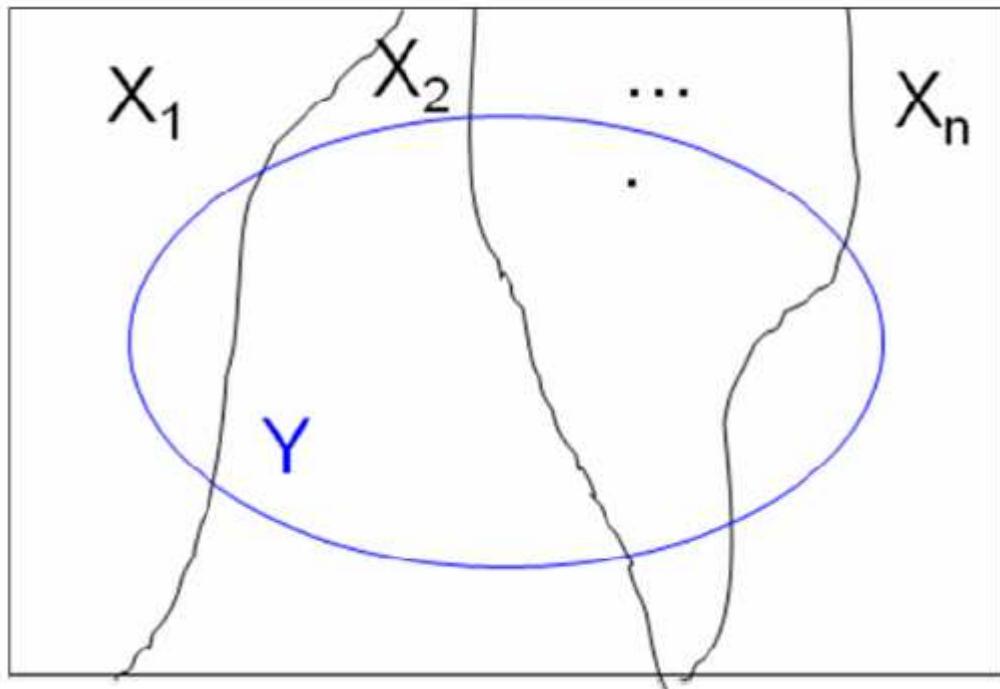
Gambar 3.1. Tahapan Proses Klasifikasi

Metode Bayes menggunakan probabilitas bersyarat sebagai dasarnya. Dalam ilmu probabilitas bersyarat dinyatakan sebagai:

$$P(X | Y) = \frac{P(X \cap Y)}{P(Y)}$$



Metode Bayes dan HMAP (*Hypothesis Maximum Appropri Probability*)



$$P(X_k | Y) = \frac{P(Y | X_k)}{\sum_i P(Y | X_i)}$$

Dimana keadaan Posteriror (Probabilitas Xk di dalam Y) dapat dihitung dari keadaan prior (Probabilitas Y di dalam Xk dibagi dengan jumlah probabilitas Y dalam semua Xi).

Terminologi dari HMAP menyatakan hipotesa yang diambil berdasarkan nilai probabilitas berdasarkan kondisi prior yang diketahui.

HMAP adalah model penyederhanaan dari metode bayes yang disebut dengan *Naïve Bayes*. HMAP dapat digunakan sebagai metode untuk mendapatkan hipotesis dari suatu keputusan. HMAP dapat diartikan untuk mencari probabilitas terbesar dari semua instance pada atribut target atau semua kemungkinan keputusan.

Salah satu manfaat algoritma naïve bayes adalah untuk melakukan prediksi terhadap data-data tertentu. Prediksi (testing) terhadap data yang akan datang bisa dilakukan berdasarkan hasil pembelajaran terhadap data training. Data training diambil dari data yang terdahulu, sedangkan data uji (testing) bisa diambil dari data-data yang sedang atau akan terjadi.

C. Alat dan Bahan

1. Komputer dengan sistem operasi Windows.
2. Program aplikasi Ms. Excel, Weka, RapidMiner.
3. Modul Praktikum Data Warehousing dan Data Mining.

D. Langkah-langkah Praktikum

D.1. Implementasi Naïve Bayes dengan Weka

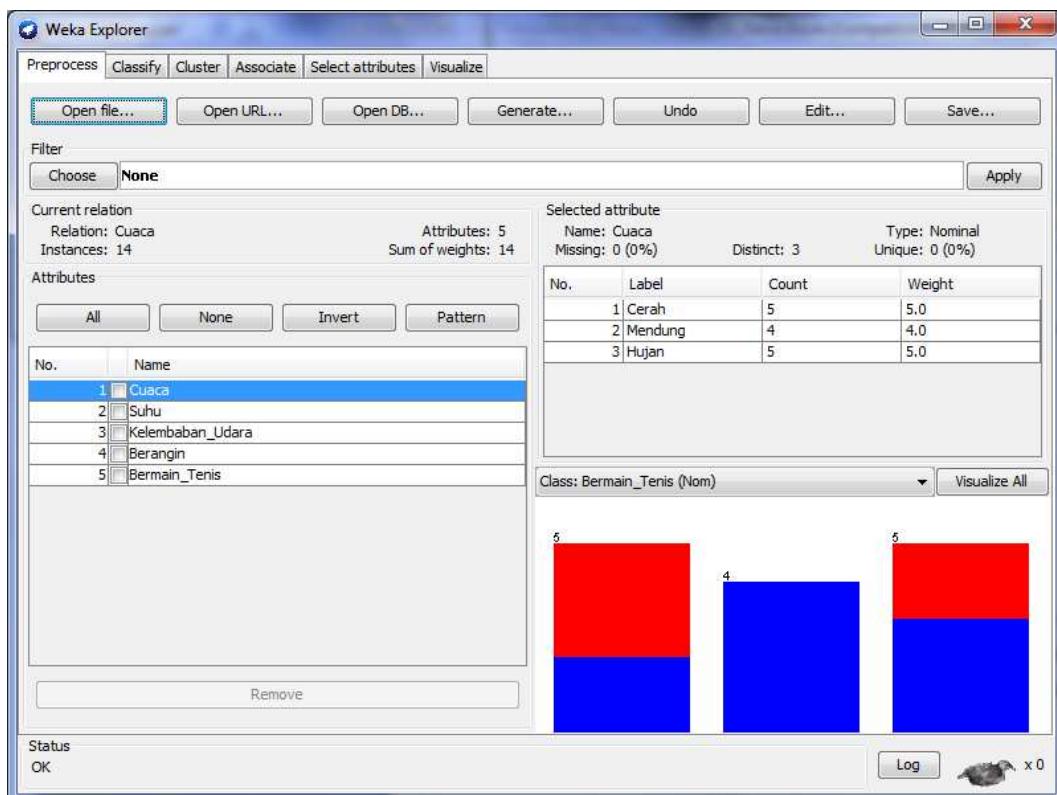
Langkah-langkah menggunakan algoritma naïve bayes dengan Weka sebagai berikut:

1. Persiapkan file **Cuaca.arff** dari hasil percobaan kegiatan D.1 pada modul 2. File ini akan kita gunakan sebagai data training.
2. Buatlah sebuah data testing dengan format **ARFF** dari tabel 3.1 sebagai data uji yang akan diprediksi dengan memiliki variabel-variabel independen dan variabel dependen yang sama. Dengan ketentuan variabel dependen diisi dengan tanda tanya (?). Asumsi bahwa kita belum mengetahui nilai / kelas dari variabel tersebut. Nilai / kelas inilah yang akan kita prediksi dengan menggunakan algoritma naïve bayes.
3. Simpan dengan nama **CuacaTesting.arff**.

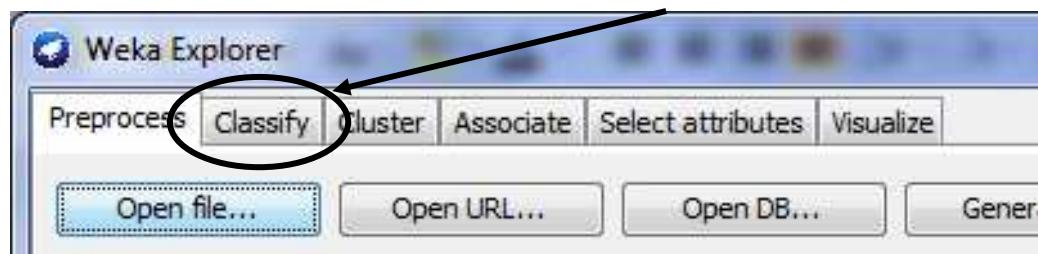
Tabel 3.1 Data Testing Cuaca:

Cuaca	Suhu	Kelembaban_udara	Berangin	Bermain_Tenis
Cerah	75	65	TIDAK	?
Cerah	80	68	YA	?
Cerah	83	87	YA	?
Mendung	70	96	TIDAK	?
Mendung	68	81	TIDAK	?
Hujan	65	75	YA	?
Hujan	64	85	YA	?

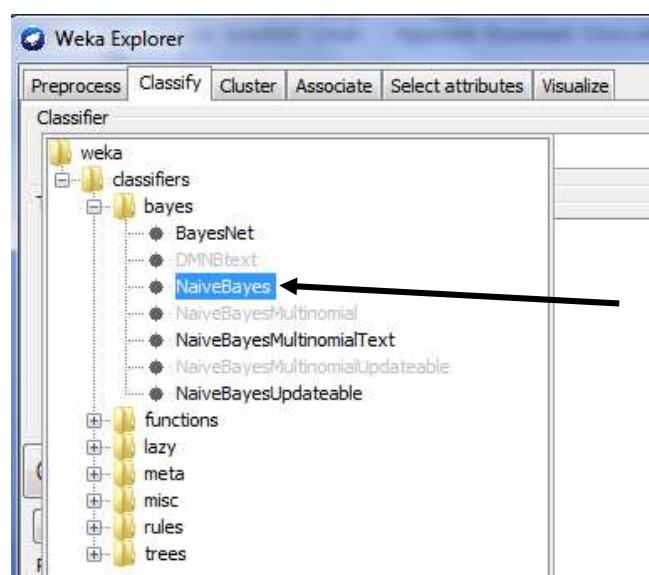
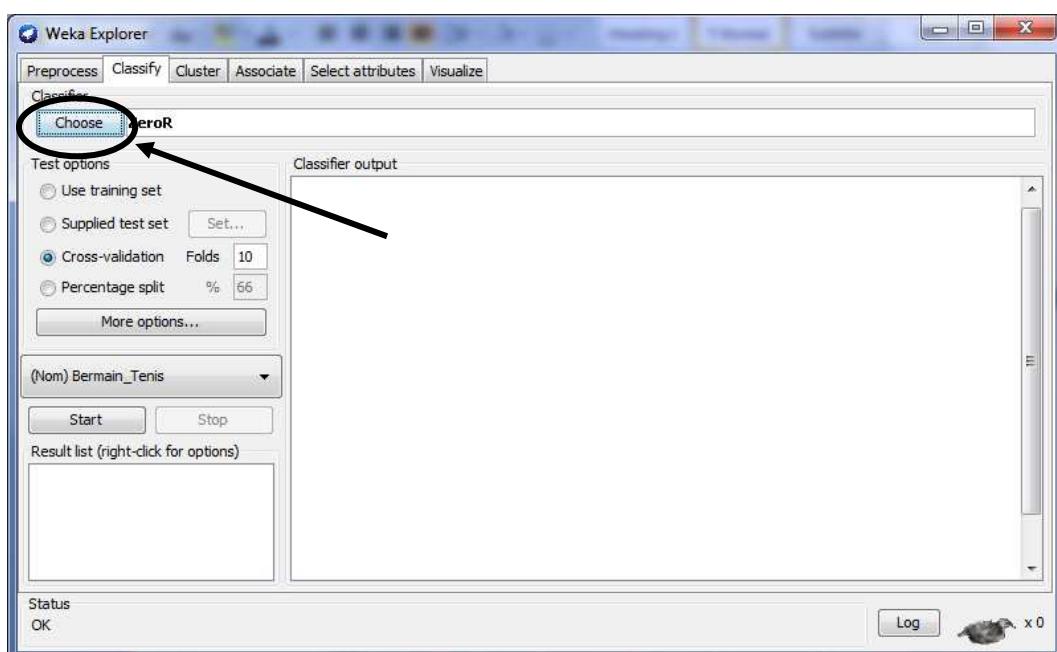
4. Jika telah selesai membuat Buka aplikasi Weka, masuk dalam menu Weka Explorer.
5. Buka kembali file **Cuaca.arff** dari hasil kegiatan D.1 dalam modul 2 dengan menggunakan Weka Explorer. File ini akan kita gunakan sebagai data pelatihan untuk memprediksi data testing pada file **CuacaTesting.arff**.



6. Masih pada jendela Weka Explorer, pilih tab **Classify**.

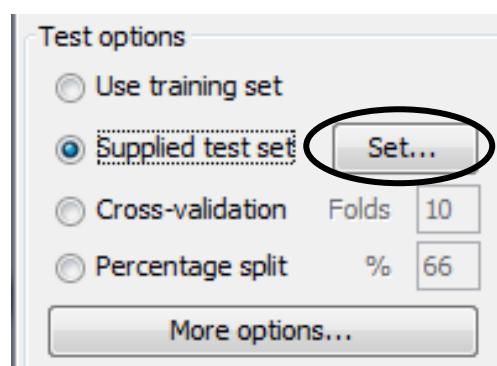


7. Sehingga akan muncul jendela Weka Explorer pada tab Classify. Pada kotak **Classifier** klik tombol Choose untuk memilih metode / algoritma **Naïve Bayes**.



8. Selanjutnya adalah menentukan data testing sebagai data yang akan diprediksi variabel dependennya. File **CuacaTesting.arff** ditentukan sebagai data testing pada kegiatan ini.
9. Pada menu Test Options terdapat 4 pilihan pengujian, yaitu:
 - a) Use training set, jika data pelatihan dan data uji menggunakan file ARFF yang sama. Pada pilihan ini, data yang akan diprediksi menggunakan data training.
 - b) Supplied test set, jika data uji telah disediakan dalam file ARFF yang lain terpisah dengan data training.
 - c) Cross-validation, nilai default Folds = 10. Hal ini berarti sistem akan mengacak data training set dan mengambil sebagian dari datanya untuk dijadikan testing set. Proses ini dilakukan sebanyak 10 kali dan hasil akhir merupakan akurasi rata-rata dari sepuluh percobaan tersebut.
 - d) Percentage split, dengan nilai default %66. Hal ini berarti sistem akan mengambil sebanyak 66% dari seluruh data yang ada sebagai data pelatihan dan sisanya digunakan sebagai data uji.
10. Pada percobaan kali ini, kita akan menggunakan pilihan **Supplied test set**.

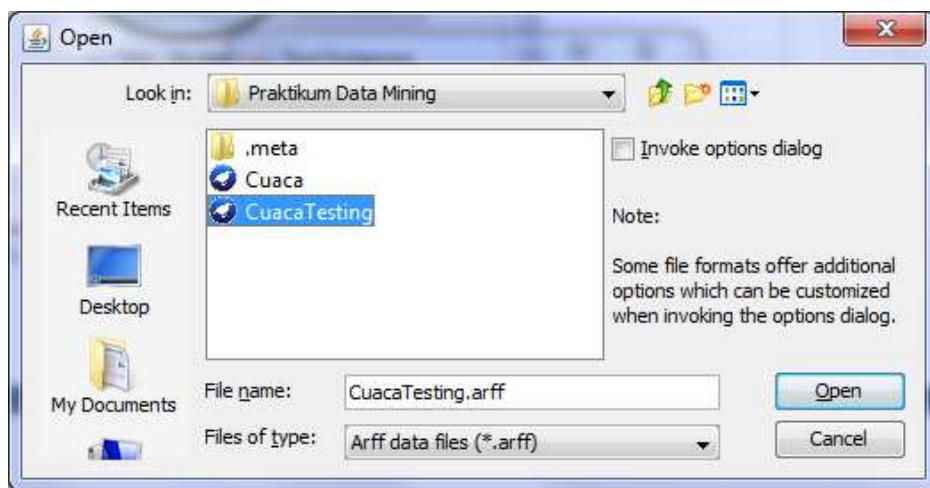
Klik tombol Set untuk menentukan file ARFF sebagai data uji.



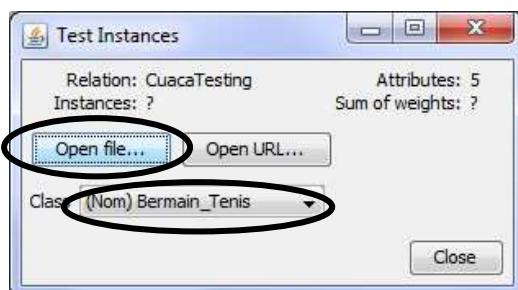
11. Sehingga akan muncul jendela Test Instance. Klik **Open File...**.



12. Pilih file **CuacaTesting.arff** sebagai data uji. Klik **Open**.



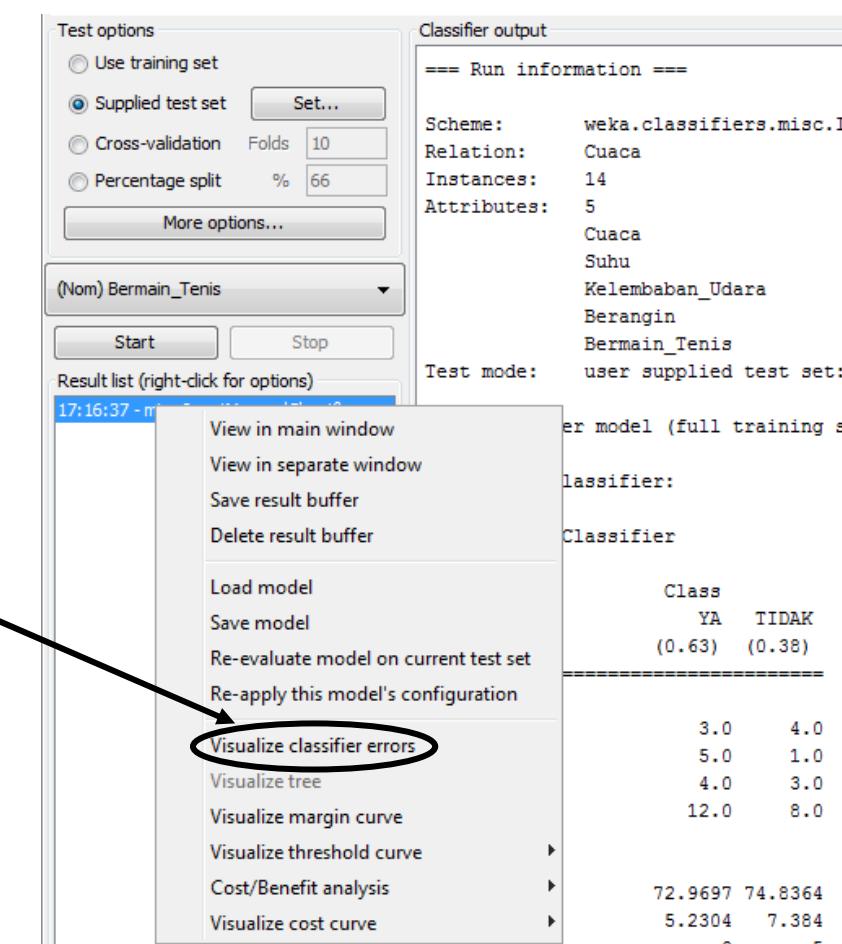
13. File **CuacaTesting.arff** akan diset sebagai data uji pada jendela Test Instances dengan variabel predictor (*Class*) adalah *Bermain_Tenis*. Klik **Close**.



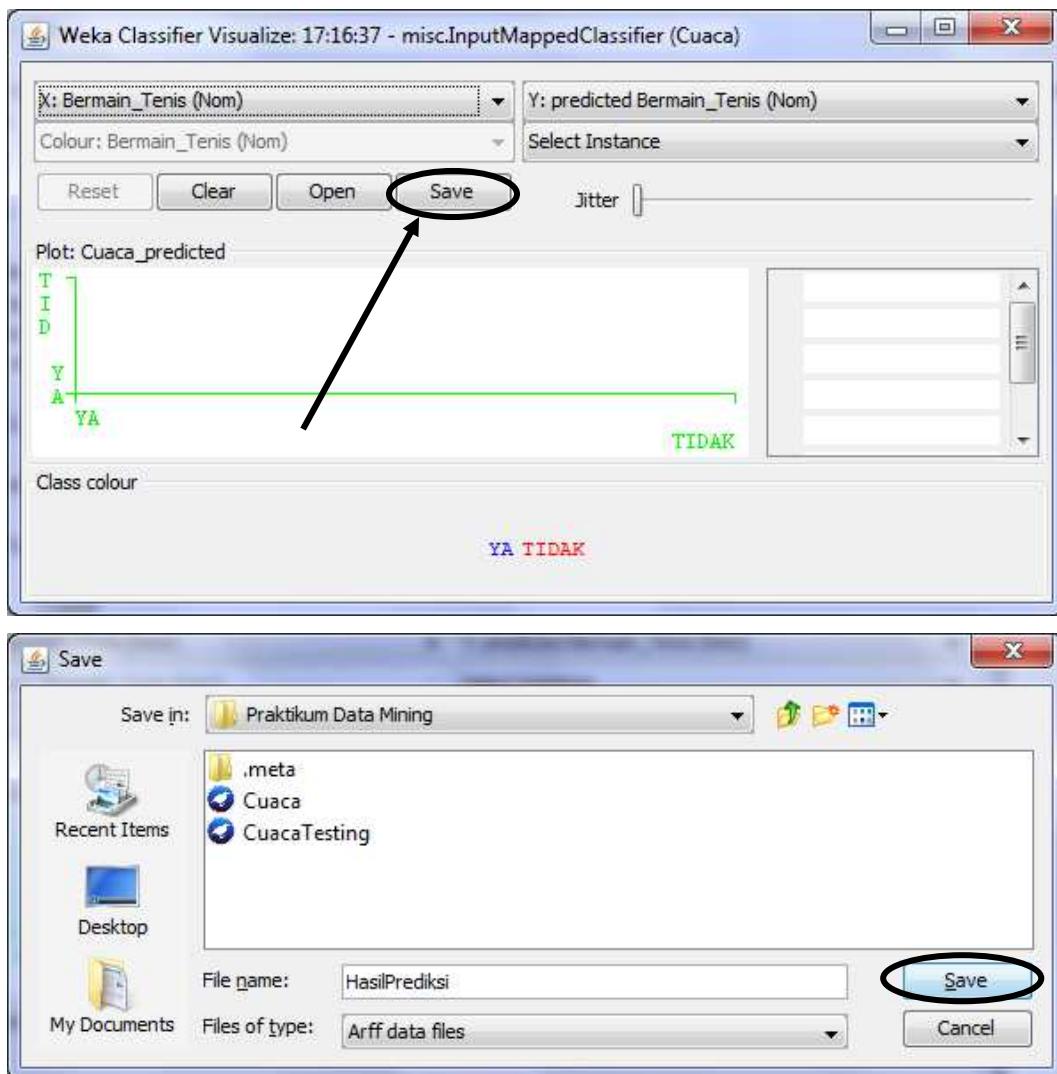
14. Klik start untuk memulai proses naïve bayes.



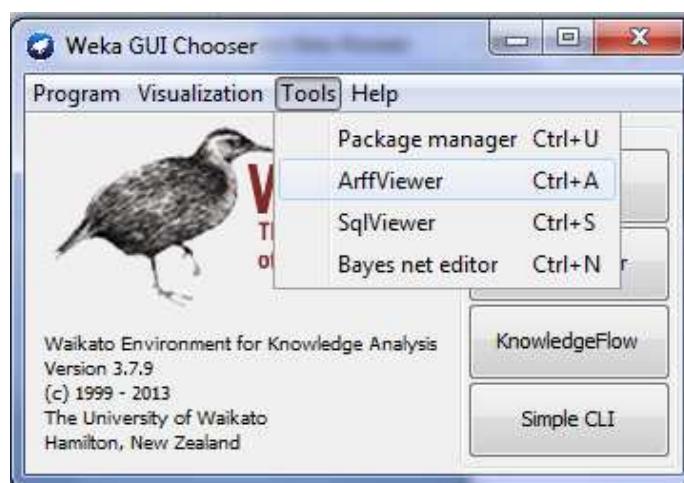
15. Jika muncul jendela pesan **Classifier Panel**, kita abaikan dengan mengklik Yes. Sehingga algoritma naïve bayes akan diproses.
16. Karena pada percobaan ini kita memproses data uji yang belum diketahui nilai kelas dari variabel dependen yang diajukan, maka kita abaikan nilai-nilai yang ditampilkan dalam jendela **Classifier Output**.
17. Untuk melihat hasil prediksi terhadap data uji, yang perlu kita lakukan berikutnya adalah dengan melihat nilai **Classifier Errors**. Klik kanan pada hasil proses dalam kotak **result list**. Pilih menu **Visualize classifier errors**.



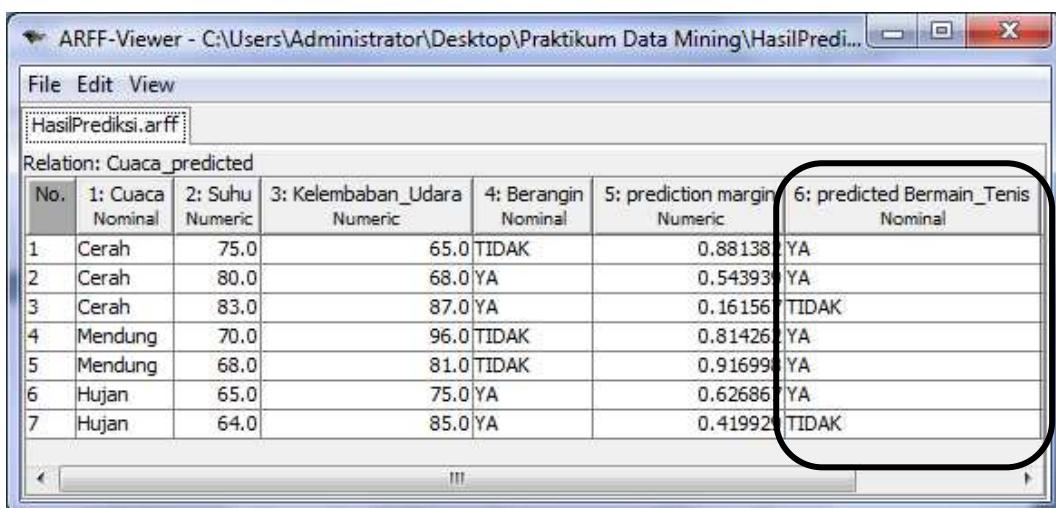
18. Pada jendela Weka Classifier Visualize, abaikan hasil apapun yang ditampilkan. Klik **save**. Simpan dengan nama file **HasilPrediksi.arff**.



19. Tutup semua jendela termasuk Weka Explorer dan kembali ke **Weka GUI Chooser**. Pilih menu Tools – ArffViewer.



20. Jendela ARFF-Viewer akan ditampilkan. Buka menu **File – Open**. Tunjukkan pada file **HasilPrediksi.arff** yang telah anda simpan pada langkah ke-18. Lihatlah, hasil prediksi telah diketahui pada kolom **predicted Bermain_Tenis (Nominal)**.



No.	1: Cuaca Nominal	2: Suhu Numeric	3: Kelembaban_Udara Numeric	4: Berangin Nominal	5: prediction margin Numeric	6: predicted Bermain_Tenis Nominal
1	Cerah	75.0	65.0	TIDAK	0.88138	YA
2	Cerah	80.0	68.0	YA	0.54393	YA
3	Cerah	83.0	87.0	YA	0.16156	TIDAK
4	Mendung	70.0	96.0	TIDAK	0.81426	YA
5	Mendung	68.0	81.0	TIDAK	0.91699	YA
6	Hujan	65.0	75.0	YA	0.62686	YA
7	Hujan	64.0	85.0	YA	0.41992	TIDAK

D.2. Implementasi Naïve Bayes dengan RapidMiner

Penggunaan algoritma naïve bayes dengan RapidMiner untuk melakukan prediksi pada dasarnya sama dengan menggunakan Weka. Kita perlu mempersiapkan data training dan data testing. Bedanya terletak pada format file yang digunakan. Jika dengan Weka file yang digunakan memiliki format ARFF, sedangkan jika menggunakan RapidMiner bisa dilakukan terhadap file excel.

Berikut langkah-langkahnya:

1. Persiapkan file **Tabel_Cuaca.xls** yang terdiri dari 2 sheet.
2. Sheet1 digunakan sebagai data training, dan sheet2 digunakan sebagai data uji.
3. Masing-masing tabel memiliki atribut yang sama, yaitu:
 - a) Cuaca (X1)
 - b) Suhu (X2)
 - c) Kelembaban_udara (X3)
 - d) Berangin (X4)
 - e) Bermain_Tenis (Y), sebagai variabel *predictor*.

Tabel data training pada Sheet1

	A	B	C	D	E
1	Cuaca	Suhu	Kelembaban_udara	Berangin	Bermain_Tenis
2	Cerah	85	85	TIDAK	TIDAK
3	Cerah	80	90	YA	TIDAK
4	Mendung	83	86	TIDAK	YA
5	Hujan	70	96	TIDAK	YA
6	Hujan	68	80	TIDAK	YA
7	Hujan	65	70	YA	TIDAK
8	Mendung	64	65	YA	YA
9	Cerah	72	95	TIDAK	TIDAK
10	Cerah	69	70	TIDAK	YA
11	Hujan	75	80	TIDAK	YA
12	Cerah	75	70	YA	YA
13	Mendung	72	90	YA	YA
14	Mendung	81	75	TIDAK	YA
15	Hujan	71	91	YA	TIDAK

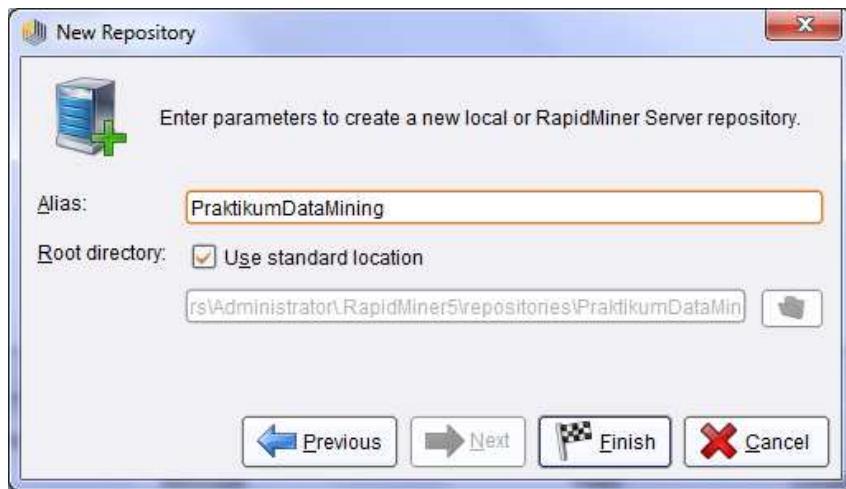
Tabel data uji pada Sheet2 tanpa ada variabel **Bermain_Tenis**.

	A	B	C	D
1	Cuaca	Suhu	Kelembaban_udara	Berangin
2	Cerah	75	65	TIDAK
3	Cerah	80	68	YA
4	Cerah	83	87	YA
5	Mendung	70	96	TIDAK
6	Mendung	68	81	TIDAK
7	Hujan	65	75	YA
8	Hujan	64	85	YA

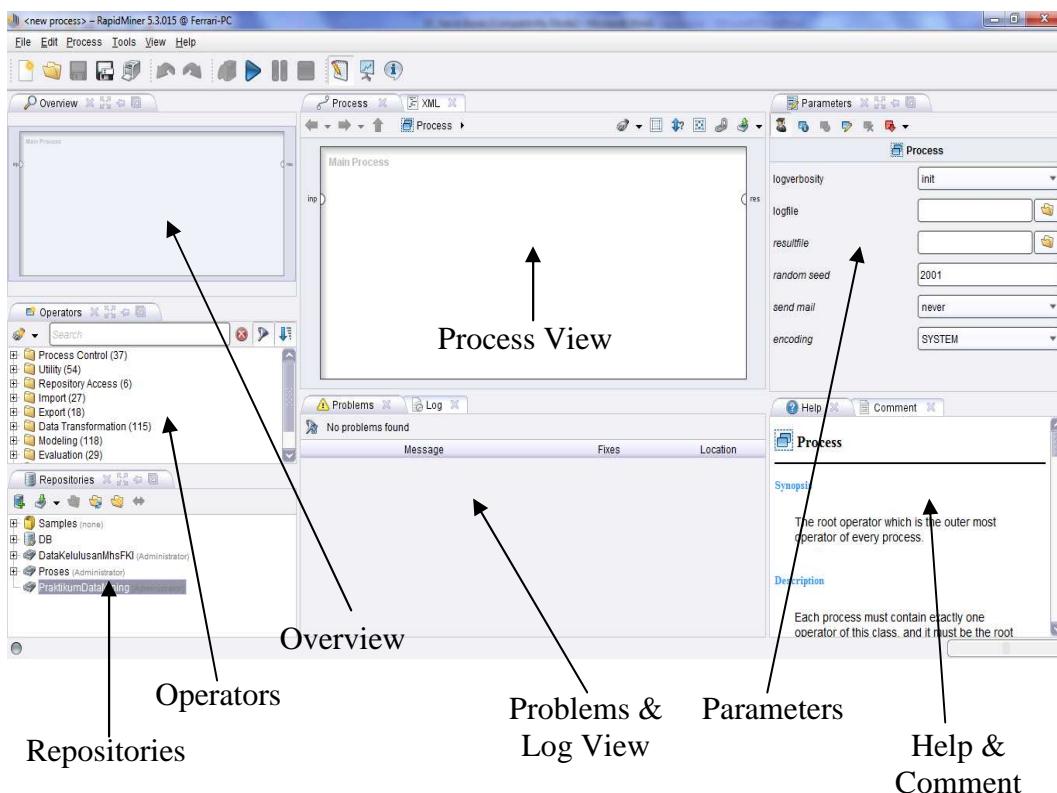
4. Buka aplikasi **RapidMiner**. Menjalankan RapidMiner untuk pertama kali, akan menanyakan pembuatan repositori baru. Repositori ini berfungsi sebagai lokasi penyimpanan terpusat untuk data dan proses analisa.
5. Pilih New local repository. Klik Next.



- Masukkan nama **Alias PraktikumDataMining**. Kemudian klik Finish.

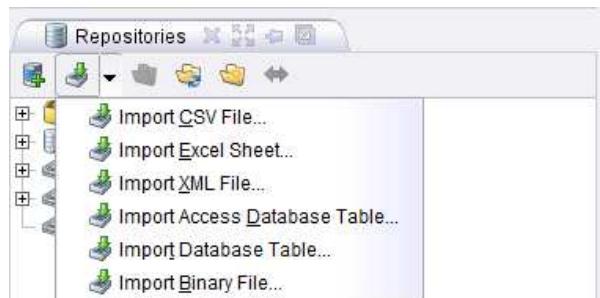


- Sehingga akan muncul jendela **Design Perspective** pada RapidMiner.



- Repository baru dengan nama **PraktikumDataMining** hasil langkah 5-6 akan tampak pada jendela Repositories. Pada repository inilah kita akan meletakkan file-file data training, data testing, bahkan hasil proses suatu metode tertentu.
- Pada kegiatan ini kita akan menggunakan file **Tabel_Cuaca.xls** yang telah kita siapkan sebagai data training dan data testing.

10. Pada jendela Repositories, pilih jendela **Import File – Import Excel Sheet**. Sehingga akan tampil jendela Data import wizard.



11. Pada step 1, tunjuk file di lokasi kita menyimpan **Tabel_Cuaca.xls**. Klik Next.



12. Pada step 2, pastikan data dalam Sheet1 terpilih untuk digunakan sebagai data training. Klik Next.

Cuaca	Suhu	Kelembaban_udara	Berangin	Bermain_Tenis
Cerah	85	85	TIDAK	TIDAK
Cerah	80	90	YA	TIDAK
Mendung	83	86	TIDAK	YA
Hujan	70	96	TIDAK	YA
Hujan	68	80	TIDAK	YA
Hujan	66	70	YA	TIDAK
Mendung	64	65	YA	YA
Cerah	72	95	TIDAK	TIDAK
Cerah	69	70	TIDAK	YA
Hujan	75	80	TIDAK	YA
Cerah	75	70	YA	YA
Mendung	72	90	YA	YA
Mendung	81	75	TIDAK	YA
Hujan	71	91	YA	TIDAK

13. Pada step 3, kita akan diperlihatkan data keseluruhan sebagai data training. Pastikan semua data telah masuk. Jika sudah yakin, klik Next.
14. Pada step 4, kita menentukan jenis masing-masing variabel. Untuk variabel **Bermain_Tenis**, tentukan tipenya menjadi **label** sebagai variabel dependen (*predictor*). Pastikan semua variabel diberi tanda cek / centang.

The screenshot shows the RapidMiner Data Editor interface. At the top, there are buttons for 'Reload data', 'Guess value types', and a checkbox for 'Preview uses only first 100 rows'. Below this is a table with columns: Cuaca, Suhu, Kelembaban_udara, Berangin, and Bermain_Tenis. The 'Bermain_Tenis' column has a dropdown menu set to 'binominal'. To the right of the table, there is a vertical scroll bar. At the bottom, there are buttons for 'Row, Column', 'Error', 'Original value', and 'Message'. A status bar at the bottom indicates '0 errors.' and checkboxes for 'Ignore errors' and 'Show only errors'.

15. Tahap terakhir (step 5), memberi nama alias pada data tersebut. Beri nama **DataCuaca_Training** dan letakkan pada Repository **PraktikumDataMining**. Klik Finish.



16. Hasil import file Tabel_Cuaca.xls pada Sheet1 akan ditampilkan.

Row No.	Bermain_T...	Cuaca	Suhu	Kelembaba...	Berangin
1	TIDAK	Cerah	85	85	TIDAK
2	TIDAK	Cerah	80	90	YA
3	YA	Mendung	83	86	TIDAK
4	YA	Hujan	70	96	TIDAK
5	YA	Hujan	68	80	TIDAK
6	TIDAK	Hujan	65	70	YA

17. Kembali ke jendela Design Perspective dengan menekan tombol pada toolbar (atau tekan tombol F8).

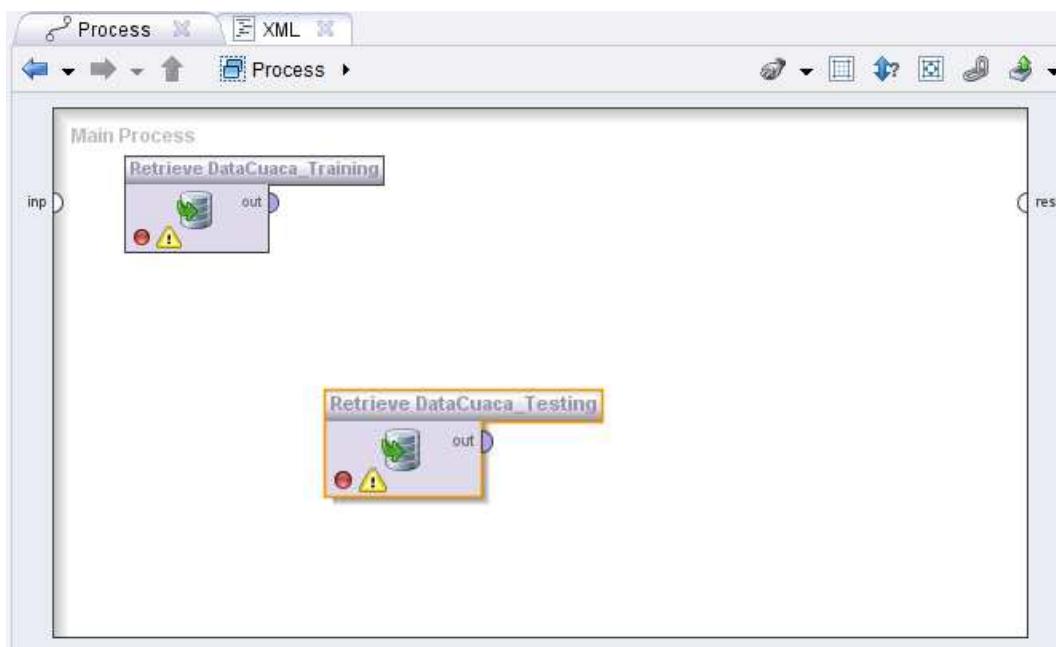
18. Lakukan hal yang sama untuk data testing yang diambil dari **Tabel_Cuaca.xls** pada Sheet2. Pastikan semua variabel data testing terpilih, dan tidak ada variabel yang diubah bertipe **label**.

19. Simpan dengan nama **DataCuaca_Testing** dan letakkan pada Repository **PraktikumData Mining**.

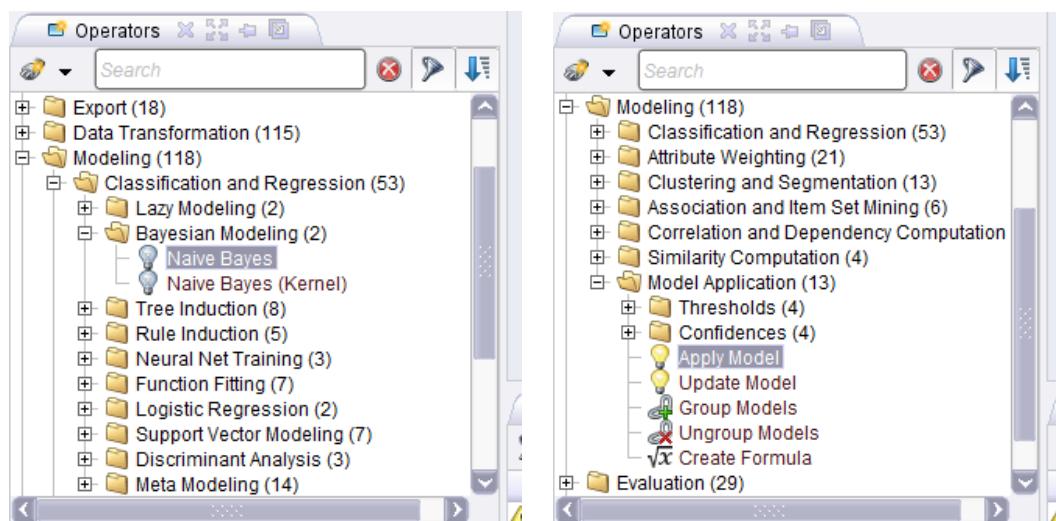


20. Berikutnya adalah mendesain model naïve bayes pada jendela Process View.

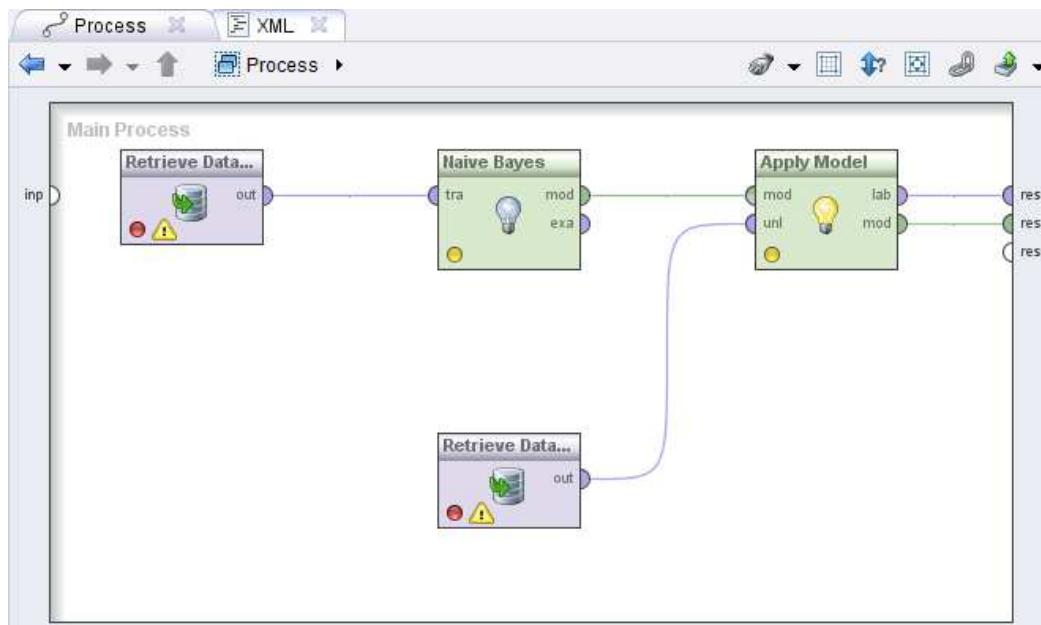
21. Drag **DataCuaca_Training** dan **DataCuaca_Testing** ke dalam jendela Process View.



22. Masukkan juga operator **Naïve Bayes** dan **Apply Model** ke dalam Process View.



23. Hubungkan konektor masing-masing data terhadap operator seperti gambar berikut. Port output data training dihubungkan dengan port input Naïve Bayes. Port output mod naïve bayes dihubungkan dengan port input mod Apply Model. Port output data testing dihubungkan dengan port input unl Apply Model. Port output lab Apply Model dihubungkan dengan port result res1, dan port output mod Apply Model dihubungkan dengan port result res2.



24. Jalankan proses naïve bayes dengan menekan tombol Run  (atau menekan tombol F11).
25. Perhatikan hasil proses klasifikasi naïve bayes. Ada 3 tab hasil proses ini, yaitu Result Overview, SimpleDistribution (Naïve Bayes), dan ExampleSet (Retrieve DataCuaca_Testing).
26. Pada tab Result Overview menunjukkan tahapan proses serta waktu yang dibutuhkan untuk eksekusi.
27. Tab SimpleDistribution menunjukkan model distribusi Naïve Bayes. Dapat dilihat bahwa distribusi nilai kelas pada variabel Y (Bermain_Tenis) sebesar 0,357 untuk nilai TIDAK, dan 0,643 untuk nilai YA.

```

SimpleDistribution

Distribution model for label attribute Bermain_Tenis

Class TIDAK (0.357)
4 distributions

Class YA (0.643)
4 distributions

```

28. Sedangkan pada tab ExampleSet, dapat dilihat hasil prediksi terhadap data testing serta tingkat *confidence* nilai kelas pada masing-masing data.

Row No.	confidence(TIDAK)	confidence(YA)	prediction(Bermain_Tenis)	Cuaca	Suhu	Kelembaban_udara	Berangin
1	0.154	0.846	YA	Cerah	75	65	TIDAK
2	0.498	0.502	YA	Cerah	80	68	YA
3	0.856	0.144	TIDAK	Cerah	83	87	YA
4	0.019	0.981	YA	Mendung	70	96	TIDAK
5	0.007	0.993	YA	Mendung	68	81	TIDAK
6	0.371	0.629	YA	Hujan	65	75	YA
7	0.568	0.432	TIDAK	Hujan	64	85	YA

29. Bandingkan dengan hasil prediksi menggunakan WEKA. Dapat dilihat bahwa prediksi masing-masing aplikasi menunjukkan hasil yang sama.

E. Tugas.

(Dikerjakan saat ini, jika waktu telah habis dilanjutkan di rumah sebagai PR. Jawaban nomor 1 – 3 dicetak dengan printer, sedangkan jawaban nomor 4 – 7 ditulis tangan).

1. Berdasarkan tabel berikut, buatlah file dalam format Excel (.xls) dan format ARFF (.arff) ! Data ini akan digunakan sebagai **data testing**.

Jurusan_SMA	Gender	Asal_Sekolah	Rerata_SKS	Asisten
LAIN	WANITA	SURAKARTA	18	TIDAK
IPA	PRIA	SURAKARTA	19	YA
LAIN	PRIA	SURAKARTA	19	TIDAK
IPS	PRIA	LUAR	17	TIDAK
LAIN	WANITA	SURAKARTA	17	TIDAK
IPA	WANITA	LUAR	18	YA
IPA	PRIA	SURAKARTA	18	TIDAK
IPA	PRIA	SURAKARTA	19	TIDAK
IPS	PRIA	LUAR	18	TIDAK
LAIN	WANITA	SURAKARTA	18	TIDAK

2. Gunakan file ARFF yang dikerjakan pada Tugas nomor 1 dalam Modul 7 sebagai data training. Lakukan prediksi terhadap data testing (ARFF) di atas menggunakan WEKA !

3. Gunakan file Excel yang dikerjakan pada Tugas nomor 1 dalam Modul 6 sebagai data training. Lakukan prediksi terhadap data testing (Excel) di atas menggunakan RapidMiner !
4. Dari hasil percobaan Tugas nomor 3 di atas, berapakah nilai Simple Distribution untuk atribut Lama_studi dengan nilai TEPAT? Berapakah nilai Simple Distribution untuk atribut Lama_studi dengan nilai TERLAMBAT?
5. Dari hasil percobaan Tugas nomor 3 di atas, berapa orang yang akan lulus TEPAT, dan berapa orang yang akan lulus TERLAMBAT?

Tambahkan 2 kondisi berikut pada data testing.

6. Prediksikan ketepatan lama studi si Dewi, jika Dewi adalah seorang WANITA yang berasal dari jurusan IPA pada saat SMA, asal sekolah dari LUAR SURAKARTA, mengambil SKS dengan rata-rata sebanyak 18 SKS tiap semester, dan tidak pernah menjadi Asisten selama kuliah.
7. Prediksikan ketepatan lama studi si Jono, jika Jono adalah seorang PRIA yang berasal dari jurusan selain IPA dan IPS pada saat SMA, asal sekolah dari SURAKARTA, mengambil SKS dengan rata-rata sebanyak 17 SKS tiap semester, dan pernah menjadi Asisten selama kuliah.

MODUL 9

POHON KEPUTUSAN (DECISION TREE)

A. Tujuan

1. Mahasiswa mampu menggunakan dan membuat model klasifikasi dengan teorema pohon keputusan.
2. Mahasiswa mampu menerapkan algoritma pohon keputusan terhadap studi kasus tertentu.

B. Landasan Teori

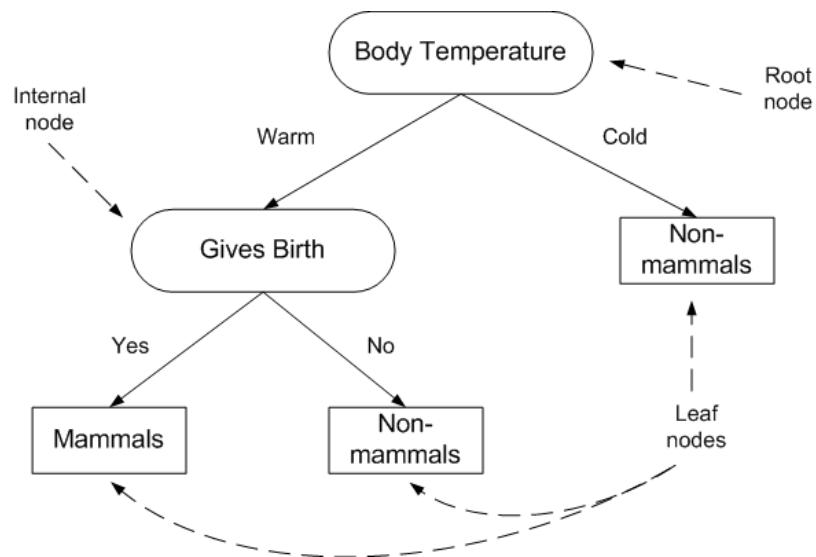
Decision tree merupakan salah satu metode belajar yang sangat populer dan banyak digunakan secara praktis. Metode ini merupakan metode yang berusaha menemukan fungsi-fungsi pendekatan yang bernilai diskrit dan tahan terhadap data – data yang memiliki kesalahan (*noisy data*) serta mampu mempelajari ekspresi – ekspresi disjunctive seperti ekspresi OR. *Iterative Dichotomizer version 3* (ID3) adalah salah satu jenis *decision tree* yang umumnya digunakan untuk menemukan aturan yang diharapkan bisa berlaku umum untuk data-data yang tidak lengkap atau yang belum pernah kita ketahui. Salah satu varian lainnya adalah J48.

Pohon (Tree) adalah sebuah struktur data yang terdiri dari simpul (node) dan rusuk (edge). Simpul pada sebuah pohon terdiri dari 3:

- 1) Simpul akar (*root node*)
- 2) Simpul percabangan/internal (*branch/internal node*)
- 3) Simpul daun (*leaf node*)

Pohon keputusan merupakan representasi sederhana dari teknik klasifikasi untuk sejumlah kelas berhingga. Simpul internal maupun simpul akar ditandai dengan nama atribut, rusuk diberi label nilai atribut yang mungkin, sedangkan simpul daun ditandai dengan kelas-kelas yang berbeda.

Objek (*record*) diklasifikasikan dengan mengikuti suatu jalur (*path*) yang dimulai dari simpul akar sesuai dengan nilai atribut dalam *record* tersebut.



Gambar 4.1 Contoh Klasifikasi Pohon Keputusan Binatang Mamalia

Pada kegiatan praktikum ini, kita menggunakan pohon keputusan untuk membuat klasifikasi data cuaca terhadap kegiatan bermain tenis. Metode ini tidak akan kita gunakan untuk melakukan prediksi terhadap data uji, namun hanya untuk membuat kelompok-kelompok klasifikasi berdasarkan nilai kelas datanya. Dengan metode ini, suatu data uji dapat diklasifikasikan berdasarkan kelas data dalam sebuah atribut.

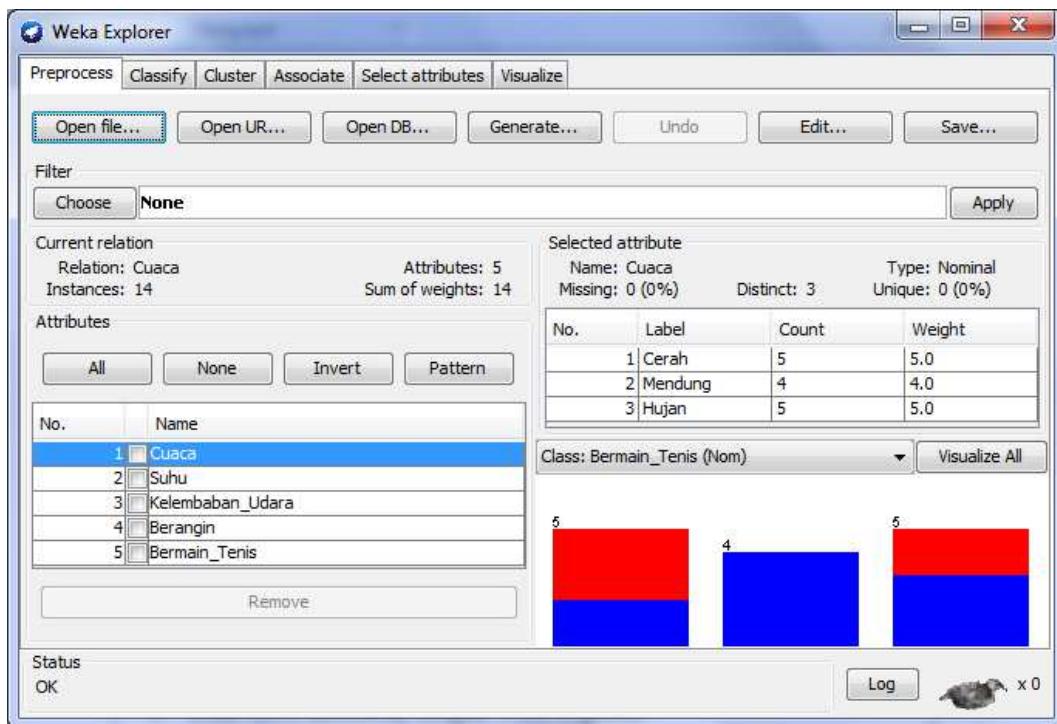
C. Alat dan Bahan

1. Komputer dengan sistem operasi Windows.
2. Program aplikasi Weka, RapidMiner.
3. Modul Praktikum Data Warehousing dan Data Mining.

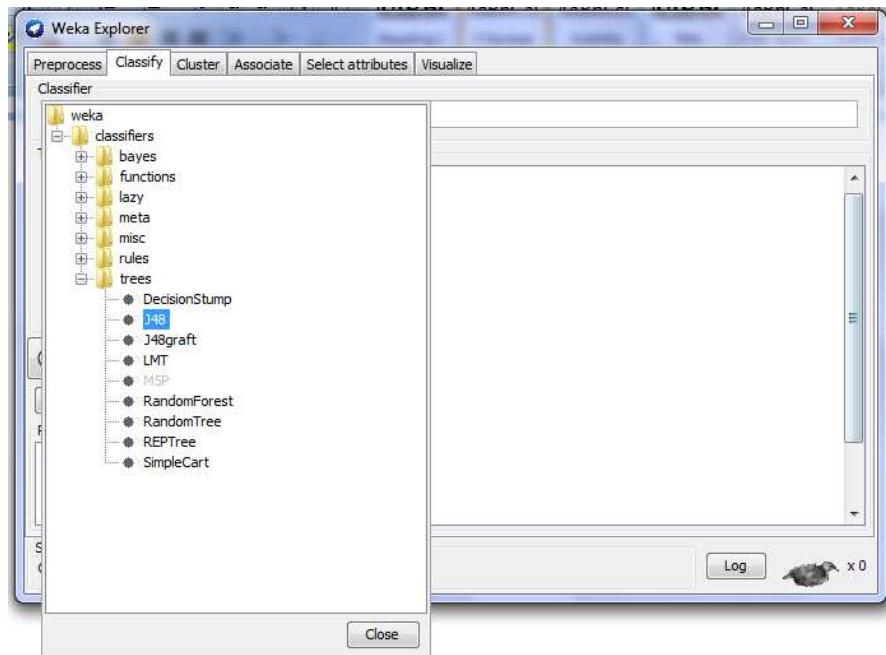
D. Langkah-langkah Praktikum

D.1. Pohon Keputusan Menggunakan Weka

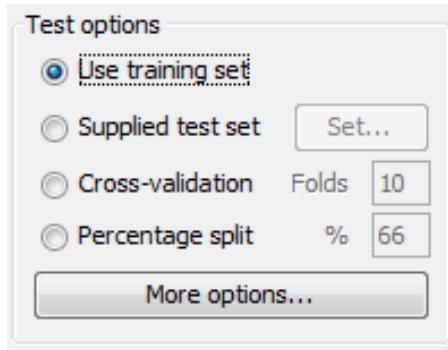
1. Jalankan aplikasi Weka - Explorer.
2. Buka file **Cuaca.arff**, dengan Weka Explorer.



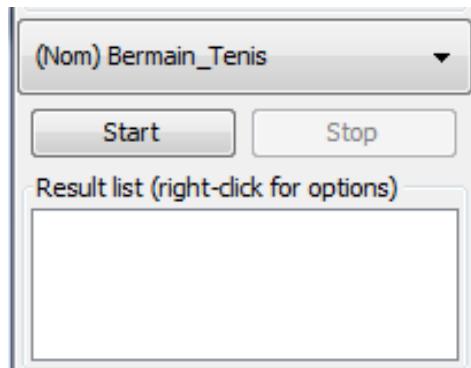
3. Setelah membuka file tersebut, buka tab **Classify**.
4. Tekan tombol **Choose – Trees – J48**. Pada kegiatan ini kita akan menggunakan algoritma pohon keputusan J48.



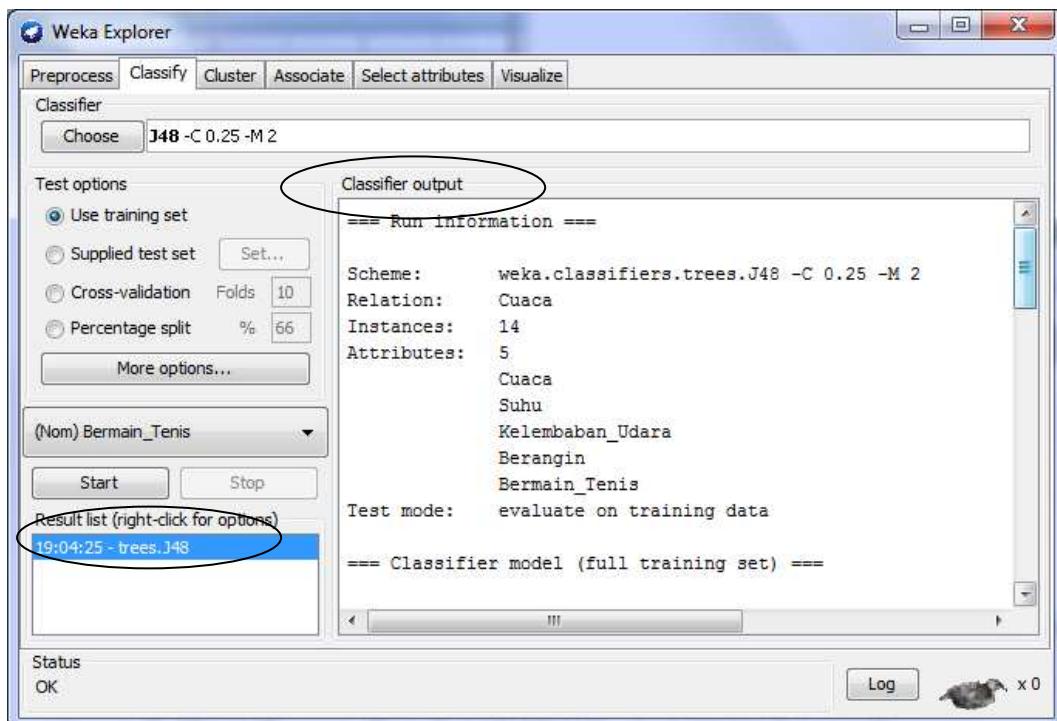
5. Pada pilihan **Test Options**, pilih **Use training set**. Data Cuaca.arff hanya kita gunakan sebagai data training untuk menciptakan klasifikasi.



6. Pastikan pada pilihan atribut dependen adalah **Bermain_Tenis**. Kemudian klik **Start** untuk memulai proses klasifikasi pohon keputusan.



7. Setelah proses selesai, maka akan ditampilkan hasil perhitungan klasifikasi menggunakan algoritma J48. Ada 2 hasil yang diberikan, yaitu pada kolom **Result list**, dan kolom **Classifier Output**.



8. Amatilah hasil pada kolom Classifier Output, scroll ke bawah untuk melihat hasil J48 Prunned Tree, waktu yang dibutuhkan untuk proses Training, dan hasil perhitungan klasifikasi.

```
J48 pruned tree
-----
Cuaca = Cerah
|   Kelembaban_Udara <= 75: YA (2.0)
|   Kelembaban_Udara > 75: TIDAK (3.0)
Cuaca = Mendung: YA (4.0)
Cuaca = Hujan
|   Berangin = YA: TIDAK (2.0)
|   Berangin = TIDAK: YA (3.0)

Number of Leaves :      5

Size of the tree :      8

==== Evaluation on training set ===

Time taken to test model on training data: 0.02 seconds

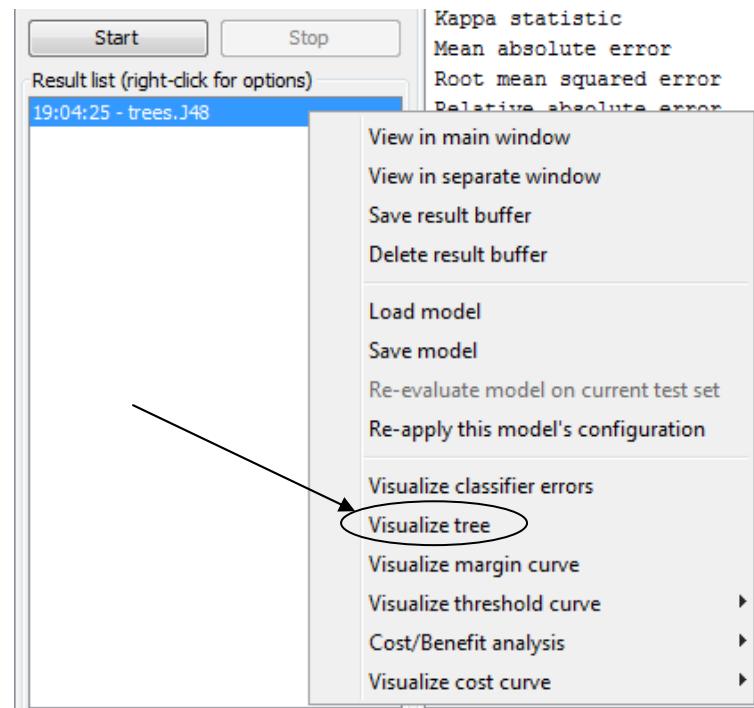
==== Summary ===

Correctly Classified Instances      14          100%
Incorrectly Classified Instances    0           0%
Kappa statistic                     1
Mean absolute error                 0
Root mean squared error            0
Relative absolute error            0
Root relative squared error       0
Coverage of cases (0.95 level)    100
Mean rel. region size (0.95 level) 50
Total Number of Instances          14
```

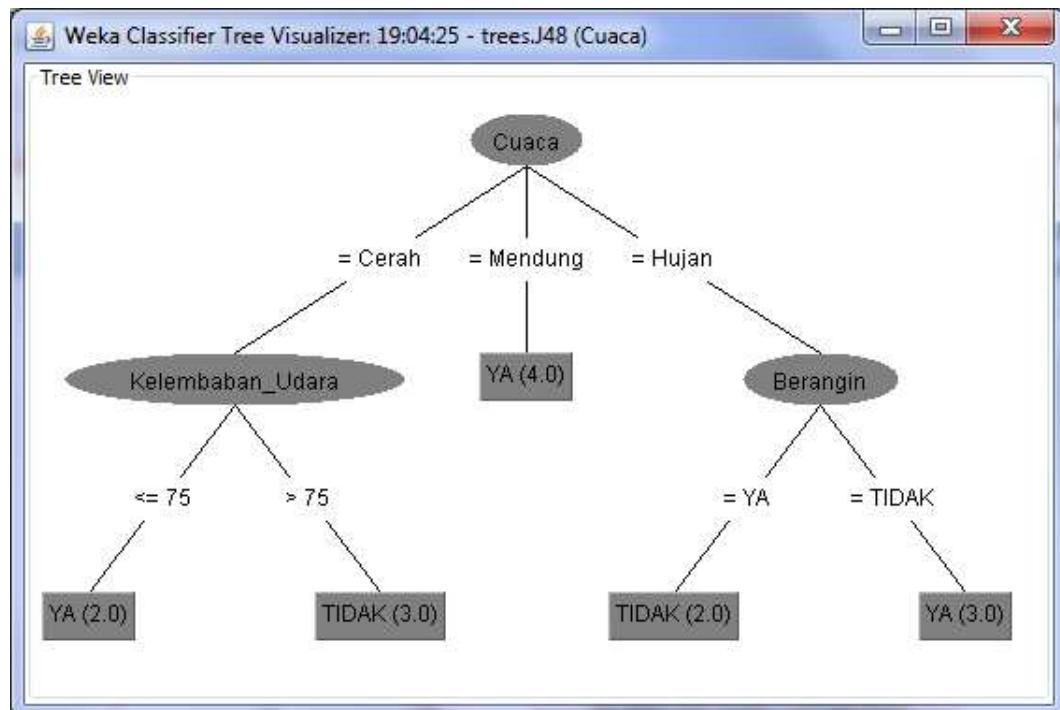
Berdasarkan hasil tersebut dapat diketahui sebagai berikut:

- 1) Jumlah simpul daun pada pohon keputusan = 5
- 2) Jumlah simpul keseluruhan pada pohon keputusan = 8
- 3) Waktu yang dibutuhkan untuk proses pelatihan = 0,02 detik
- 4) Tingkat ketepatan klasifikasi = 100%
- 5) Tingkat ketidaktepatan klasifikasi = 0%

9. Untuk melihat hasil skema pohon keputusan, kembali ke kolom **Result List**. Klik kanan pada hasil **trees.J48 – Visualize tree**.



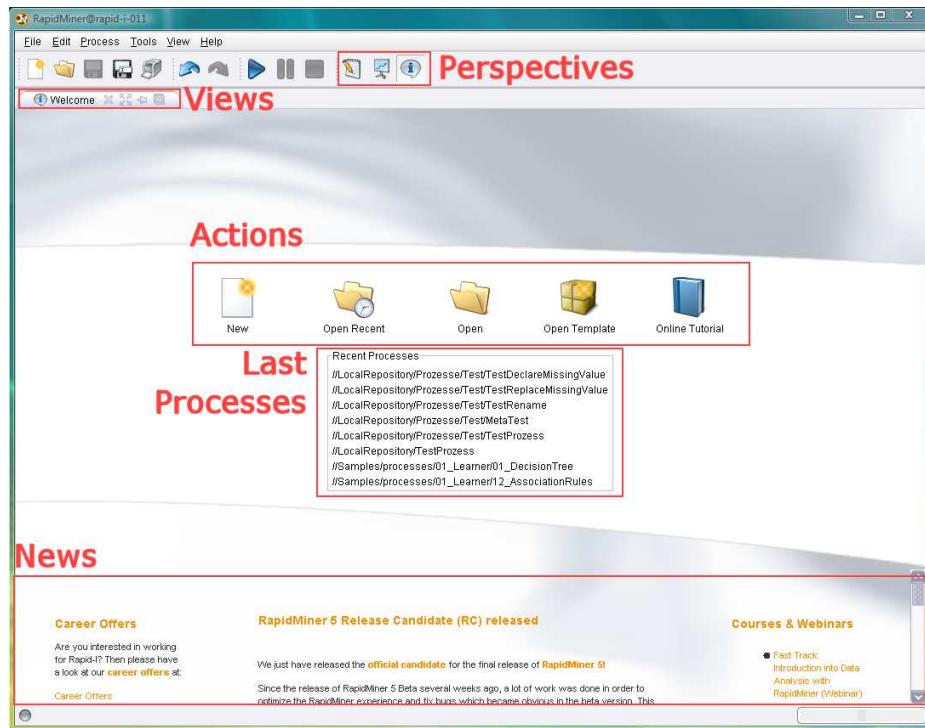
10. Sebuah jendela Weka Classifier Tree Visualizer akan ditampilkan. Pada jendela ini akan tampak Tree View, hasil klasifikasi pohon keputusan.



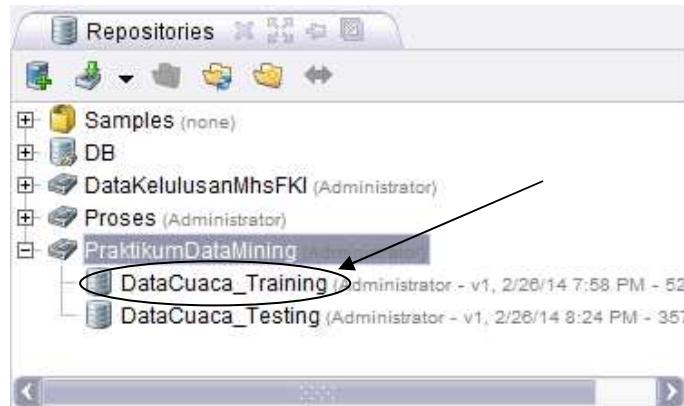
11. Berdasarkan pohon keputusan tersebut, dapat dilihat jenis-jenis simpul yang ada, sebagai berikut:
- 1) Simpul akar = Cuaca
 - 2) Simpul internal = Kelembaban_udara, dan Berangin
 - 3) Simpul daun = YA, TIDAK
12. Klasifikasi yang terbentuk yaitu:
- 1) Seseorang akan bermain tenis (YA) jika kondisi sebagai berikut:
 - a) Cuaca = Cerah, Kelembaban_udara ≤ 75 , (nilai atribut lain diabaikan).
 - b) Cuaca = Mendung, (kondisi lain diabaikan).
 - c) Cuaca = Hujan, Berangin = TIDAK, (nilai atribut lain diabaikan).
 - 2) Seseorang tidak akan bermain tenis (TIDAK) jika kondisi sebagai berikut:
 - a) Cuaca = Cerah, Kelembaban_udara > 75 , (nilai atribut lain diabaikan).
 - b) Cuaca = Hujan, Berangin = YA, (nilai atribut lain diabaikan).

D.2. Pohon Keputusan Menggunakan RapidMiner

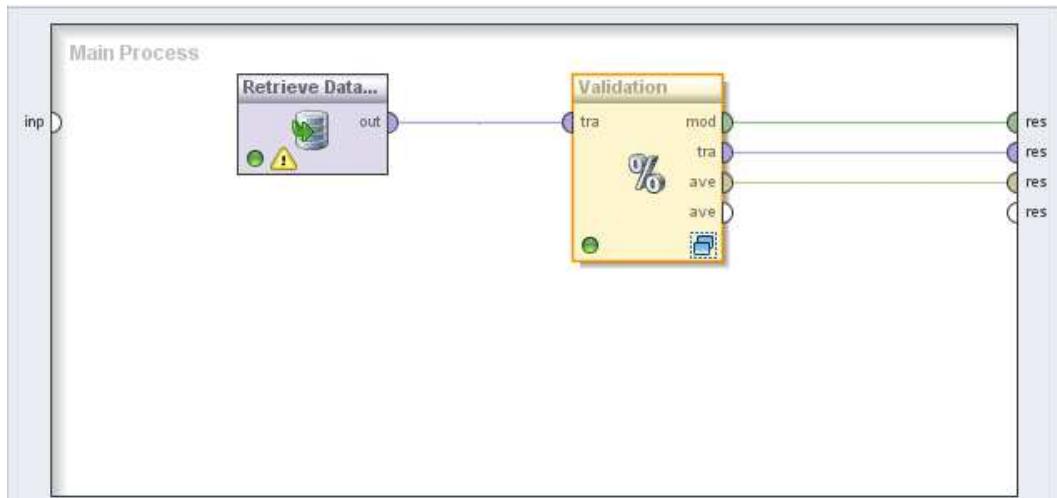
1. Buka aplikasi RapidMiner.



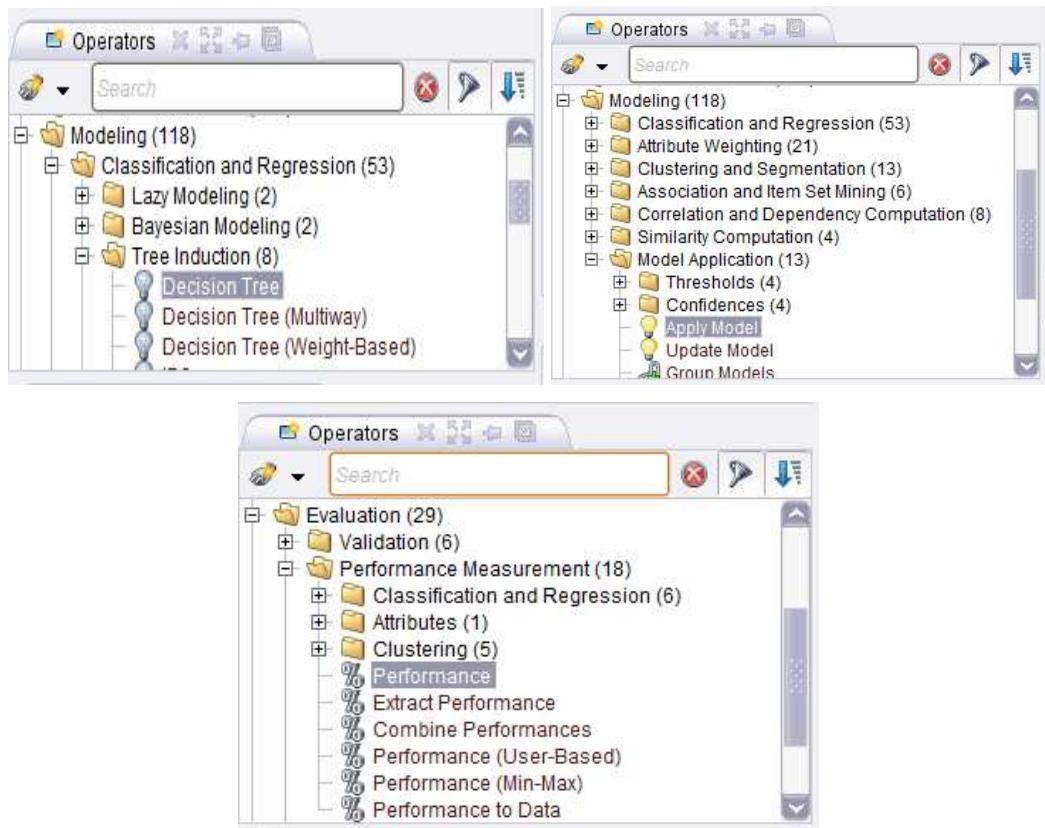
2. Pilih **New Process** pada area Actions.
3. Pada area Process View, kita akan mendesain model proses pohon keputusan menggunakan data **DataCuaca_Training** yang telah dibuat pada kegiatan D.2 dalam Modul 3.



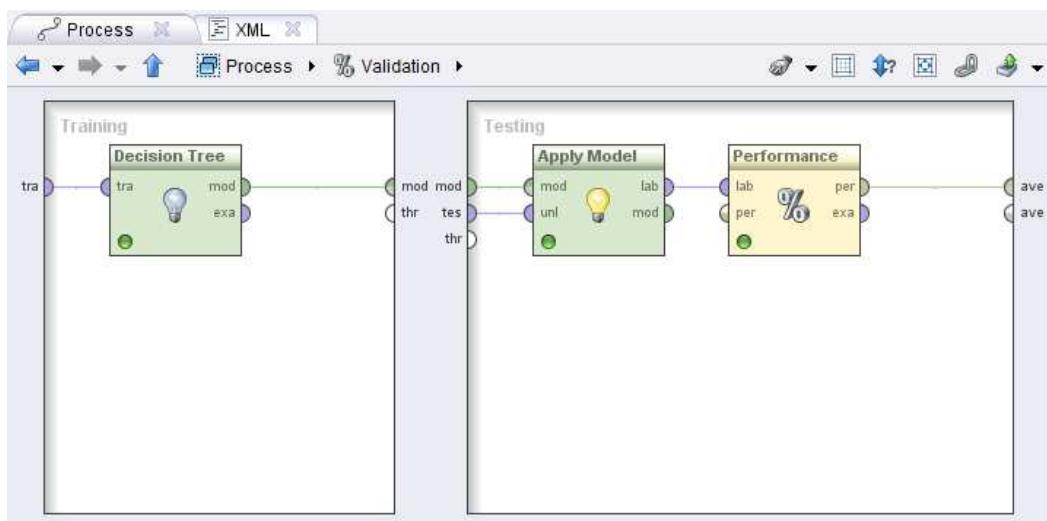
4. Drag DataCuaca_Training ke area Process View. Drag pula operator X-Validation ke area Process View. Hubungkan port output data training ke port input X-Validation serta port output X-Validation dengan port input Result seperti gambar berikut.



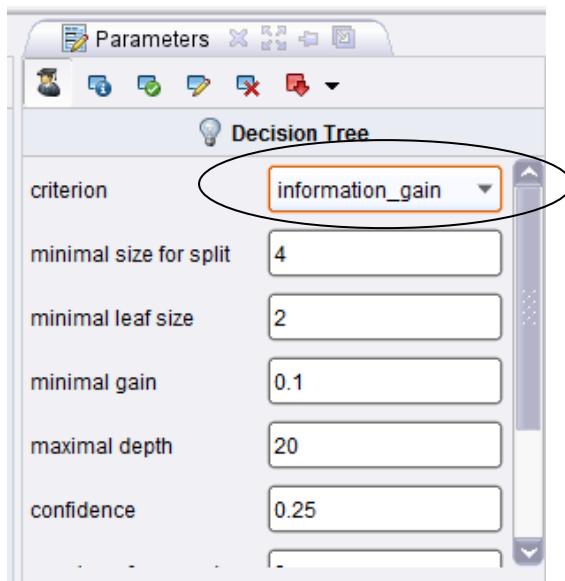
5. Selanjutnya adalah merancang algoritma pohon keputusan. Klik ganda operator X-Validation yang terdapat pada Process View, sehingga masuk ke jendela Process – Validation.
6. Masukkan operator **Decision Tree** dalam area **Training**, operator **Apply Model** dan **Performance** dalam area **Testing**.



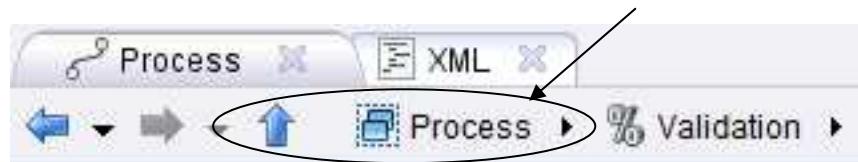
7. Hubungkan port input dan output masing-masing operator seperti gambar berikut.



8. Klik Decision Tree pada area Training, dan pastikan kriteria (**Criterion**) yang dipakai adalah **Information Gain** (J48) pada kolom **Parameter** di sebelah kanan Process View.



9. Tekan tombol panah ke atas atau tekan tombol Process yang terletak di atas area view untuk kembali ke desain awal.



10. Jalankan proses dengan menekan tombol Run (atau menekan tombol F11).

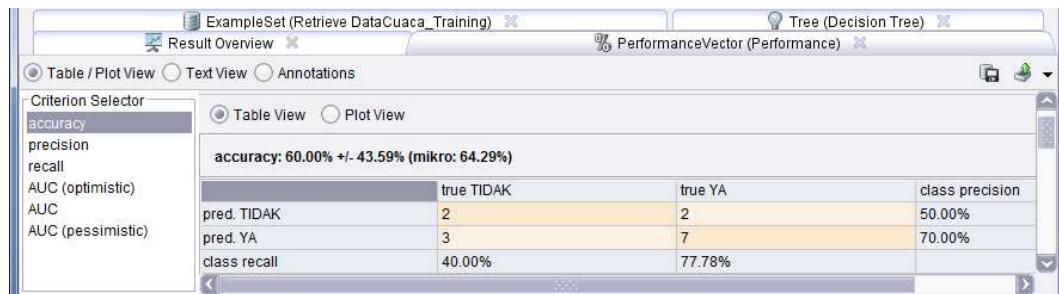
11. Perhatikan hasil proses klasifikasi pohon keputusan.

Berikut 3 hasil proses klasifikasi pohon keputusan:

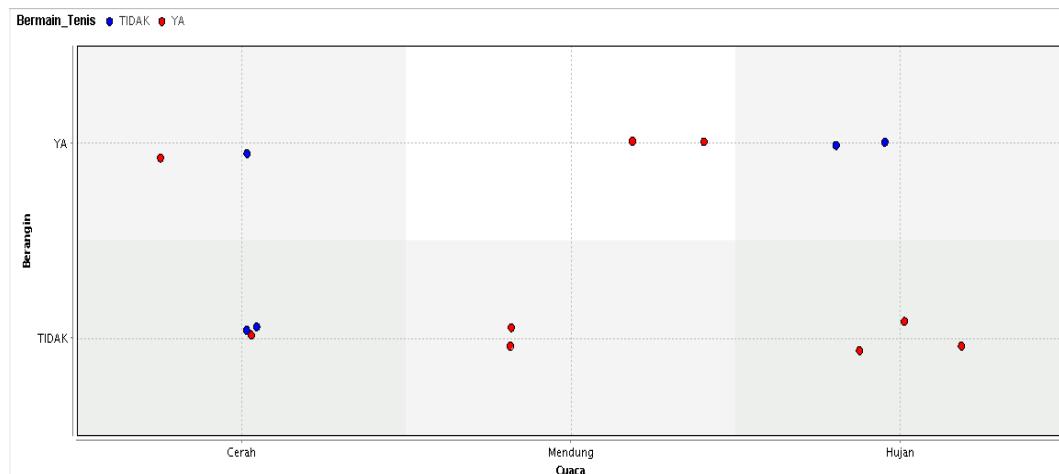
- 1) Result Overview, yang menunjukkan tahapan proses serta waktu yang dibutuhkan untuk eksekusi.

ExampleSet (Retrieve DataCuaca_Training)	Tree (Decision Tree)
Result Overview	PerformanceVector (Performance)
Process (3 results. Process results)	
Completed: Feb 27, 2014 8:03:54 PM (execution time: 0 s)	
Process (3 results. Process results)	
Completed: Feb 27, 2014 8:24:13 PM (execution time: 0 s)	
Process (3 results. Process results)	
Completed: Feb 27, 2014 8:24:34 PM (execution time: 0 s)	
Process (3 results. Process results)	
Completed: Feb 27, 2014 8:29:09 PM (execution time: 0 s)	

- 2) PerformanceVector, menunjukkan tingkat akurasi, presisi, recall dan lain-lain dalam bentuk tabel, plot view, dan juga text view. Bisa anda coba satu persatu untuk melihat.



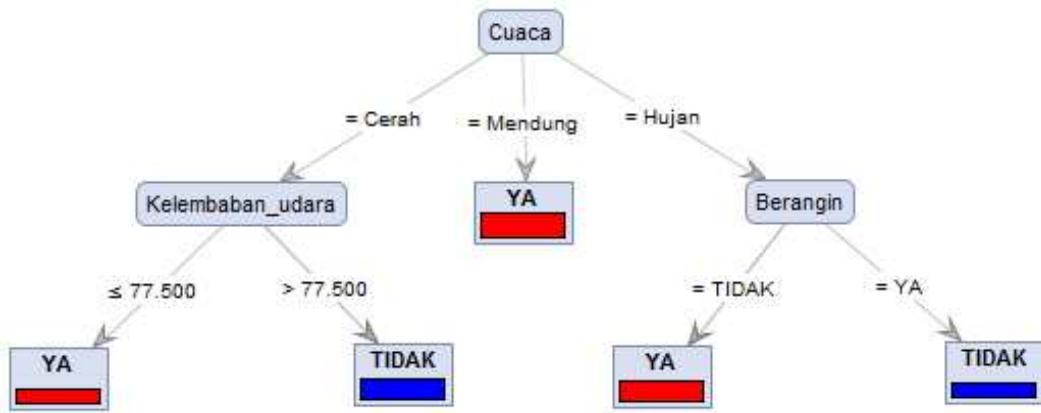
- 3) ExampleSet (Retrieve DataCuaca_Training), menunjukkan isi dari data training. Bisa dilihat dengan berbagai perspektif, antara lain Data View, Meta Data View, Plot View, Advance Chart dan Annotation. Contoh perspektif **Plot View** dengan model Scatter. **X-Axis** = Cuaca, **Y-Axis** = Berangin, dan **Color Column** = Bermain_Tenis. Nilai **Jitter** bisa diubah-ubah untuk melihat lebih jelas pola penyebarannya.



Titik-titik berwarna merah menunjukkan Bermain_Tenis = YA, sedangkan warna biru menunjukkan Bermain_Tenis = TIDAK.

Dapat dilihat pola penyebaran titik-titik tersebut sesuai dengan pola pohon keputusan yang dihasilkan berdasarkan sumbu X = Cuaca dan sumbu Y = Berangin.

- 4) Tree (Decision Tree), menunjukkan hasil pohon keputusan dari proses klasifikasi. Dapat dilihat bahwa pohon keputusan yang dihasilkan sama seperti saat menggunakan Weka.



12. Berdasarkan pohon keputusan tersebut, dapat dilihat jenis-jenis simpul yang ada, sebagai berikut:

- 1) Simpul akar = Cuaca
- 2) Simpul internal = Kelembaban_udara, dan Berangin
- 3) Simpul daun = YA, TIDAK

13. Klasifikasi yang terbentuk yaitu:

- 1) Seseorang akan bermain tenis (YA) jika kondisi sebagai berikut:
 - a) Cuaca = Cerah, Kelembaban_udara \leq 77,5. (nilai atribut lain diabaikan).
 - b) Cuaca = Mendung. (nilai atribut lain diabaikan).
 - c) Cuaca = Hujan, Berangin = TIDAK. (nilai atribut lain diabaikan).
- 2) Seseorang tidak akan bermain tenis (TIDAK) jika kondisi sebagai berikut:
 - a) Cuaca = Cerah, Kelembaban_udara $>$ 77,5. (nilai atribut lain diabaikan).
 - b) Cuaca = Hujan, Berangin = YA. (nilai atribut lain diabaikan).

E. Tugas.

A. Dikerjakan saat ini

- 1) Berdasarkan pohon keputusan pada kegiatan D.2 (menggunakan RapidMiner), isikan nilai kelas atribut Bermain_Tenis pada tabel Testing berikut:

Cuaca	Suhu	Kelembaban udara	Berangin	Bermain_Tenis
Cerah	75	65	TIDAK
Cerah	80	68	YA
Cerah	83	87	YA
Mendung	70	96	TIDAK
Mendung	68	81	TIDAK
Hujan	65	75	TIDAK
Hujan	64	85	YA

B. Dikerjakan di rumah. Hasil jawaban dicetak menggunakan printer warna.

- 1) Gunakan file ARFF yang dikerjakan pada Tugas nomor 1 dalam Modul 7 sebagai data training.
 - a) Buatlah dan cetaklah pohon keputusan berdasarkan data tersebut!
 - b) Carilah nilai-nilai parameter berikut:
 1. Jumlah simpul daun pada pohon keputusan =
 2. Jumlah simpul keseluruhan pada pohon keputusan =
 3. Waktu yang dibutuhkan untuk proses pelatihan = detik
 4. Tingkat ketepatan klasifikasi =%
 5. Tingkat ketidaktepatan klasifikasi =%
- 2) Gunakan file Excel yang dikerjakan pada Tugas nomor 1 dalam Modul 6 sebagai data training.
 - a) Buatlah dan cetaklah pohon keputusan berdasarkan data tersebut!
 - b) Cetaklah perspektif Plot View dengan model Scatter. X-Axis = Gender, Y-Axis = Asisten, dan Color Column = Lama_Studi. Nilai **Jitter** bisa diubah-ubah untuk memperoleh pola penyebaran yang lebih jelas.
- 3) Berdasarkan pohon keputusan dari soal nomor 2, tentukan klasifikasi yang terbentuk berdasarkan kondisinya sesuai dengan simpul-simpulnya.

MODUL 10

CLUSTERING: ALGORITMA K-MEANS

A. Tujuan

1. Mahasiswa mampu menggunakan algoritma K-Means.
2. Mahasiswa mampu menerapkan algoritma K-Means dalam kasus nyata.

B. Landasan Teori

Clustering merupakan suatu teknik data mining yang membagi-bagikan data ke dalam beberapa kelompok (grup atau cluster atau segmen) yang tiap cluster dapat ditempati beberapa anggota bersama-sama. Setiap obyek dilewatan ke grup yang paling mirip dengannya. Ini menyerupai menyusun binatang dan tumbuhan ke dalam keluarga – keluarga yang para anggotanya mempunyai kemiripan. Clustering tidak mensyaratkan pengetahuan sebelumnya dari grup yang dibentuk, juga dari para anggota yang harus mengikutinya.

Algoritma K-Means diperkenalkan oleh J.B. MacQueen pada tahun 1976, salah satu algoritma clustering sangat umum yang mengelompokkan data sesuai dengan karakteristik atau ciri-ciri bersama yang serupa. Grup data ini dinamakan sebagai *cluster*. Data di dalam suatu cluster mempunyai ciri-ciri (atau fitur, karakteristik, atribut, properti) serupa dan tidak serupa dengan data pada *cluster* lain.

K-means merupakan salah satu metode clustering non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih *cluster*. Metode ini mempartisi data ke dalam *cluster* sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu *cluster* yang sama dan data yang mempunyai karakteristik yang berbeda di kelompokkan ke dalam *cluster* yang lain. Secara umum algoritma dasar dari K-Means *Clustering* adalah sebagai berikut :

1. Tentukan jumlah *cluster*
2. Alokasikan data ke dalam *cluster* secara random
3. Hitung *centroid/rata-rata* dari data yang ada di masing-masing *cluster*
4. Alokasikan masing-masing data ke *centroid/rata-rata* terdekat

- Kembali ke Step 3, apabila masih ada data yang berpindah *cluster* atau apabila perubahan nilai *centroid*, ada yang di atas nilai *threshold* yang ditentukan atau apabila perubahan nilai pada *objective function* yang digunakan di atas nilai *threshold* yang ditentukan.

Distance space digunakan untuk menghitung jarak antara data dan *centroid*. Adapun persamaan yang dapat digunakan salah satunya yaitu *Euclidean Distance Space*. *Euclidean distance space* sering digunakan dalam perhitungan jarak, hal ini dikarenakan hasil yang diperoleh merupakan jarak terpendek antara dua titik yang diperhitungkan. Adapun persamaannya adalah sebagai berikut :

$$d_{ij} = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2}$$

Keterangan:

- d_{ij} = Jarak objek antara objek i dan j
- p = Dimensi data
- x_{ik} = Koordinat dari obyek i pada dimensi k
- x_{jk} = Koordinat dari obyek j pada dimensi k

C. Alat dan Bahan

- Komputer dengan sistem operasi Windows.
- Program aplikasi RapidMiner.
- Modul Praktikum Data Warehousing dan Data Mining.

D. Langkah-langkah Praktikum

D.1. Algoritma K-Means menggunakan RapidMiner

Contoh kasus:

Dalam sebuah kelas terdapat 10 siswa yang telah menempuh ujian mata pelajaran Bahasa Indonesia. Data nilai siswa tersebut akan kita gunakan sebagai dasar pengambilan keputusan untuk mencari kelompok siswa yang akan kita kirimkan ke lomba / olimpiade bidang studi Bahasa Indonesia dan Bahasa Inggris.

Hipotesis:

Bagaimana mencari kelompok-kelompok siswa dalam bidang studi Bahasa Indonesia dan Bahasa Inggris sesuai dengan nilai ujian yang telah ditempuh oleh siswa.

Berikut tabel data nilai siswa:

NO_SISWA	NAMA	B.IND	B.ING
S-101	JOKO	8,54	8,40
S-102	AGUS	9,98	6,81
S-103	SUSI	6,20	9,15
S-104	DYAH	5,24	7,26
S-105	WATI	5,70	5,71
S-106	IKA	8,57	5,87
S-107	EKO	7,70	7,71
S-108	YANTO	6,60	5,70
S-109	WAWAN	9,00	8,12
S-110	MAHMUD	9,81	9,58

1. Buka Ms. Excel, dan buatlah tabel data nilai ujian siswa tersebut. Simpan dengan nama **Tabel_NilaiUjian.xls** (**Format Excel 2003 *.xls**).
2. Jalankan aplikasi **RapidMiner**.
3. Gunakan file **Tabel_NilaiUjian.xls** sebagai data yang akan digunakan dalam proses Clustering. Import file ini ke dalam repositories Praktikum Data Mining (seperti pada Modul 8 Kegiatan D.2 Langkah 10-16). Yang perlu diperhatikan hanya pada saat penentuan tipe atribut pada Step 4. Atribut No.Siswa dan Nama diubah menjadi tipe “Text”, sedangkan yang lainnya biarkan dengan tipe “Real”.

NO_SISWA	NAMA	B.IND	B.ING
attribute	attribute	attribute	attribute
S-101	JOKO	8.545	8.397
S-102	AGUS	9.976	6.814
S-103	SUSI	6.201	9.154
S-104	DYAH	5.235	7.260
S-105	WATI	5.700	5.710
S-106	IKA	8.566	5.873
S-107	EKO	7.698	7.712
S-108	YANTO	6.596	5.702
S-109	WAWAN	8.995	8.116
S-110	MAHMUD	9.814	9.579

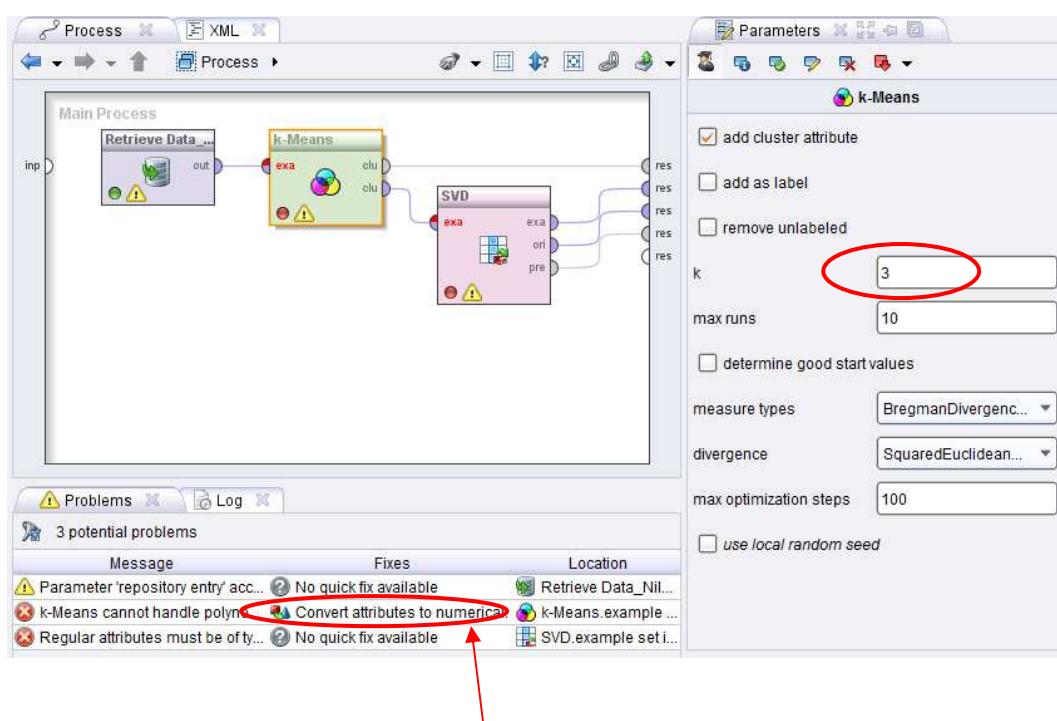
4. Pada step 5, beri nama Data_NilaiUjian dan masukkan pada repositories Praktikum Data Mining. Kemudian klik Finish.



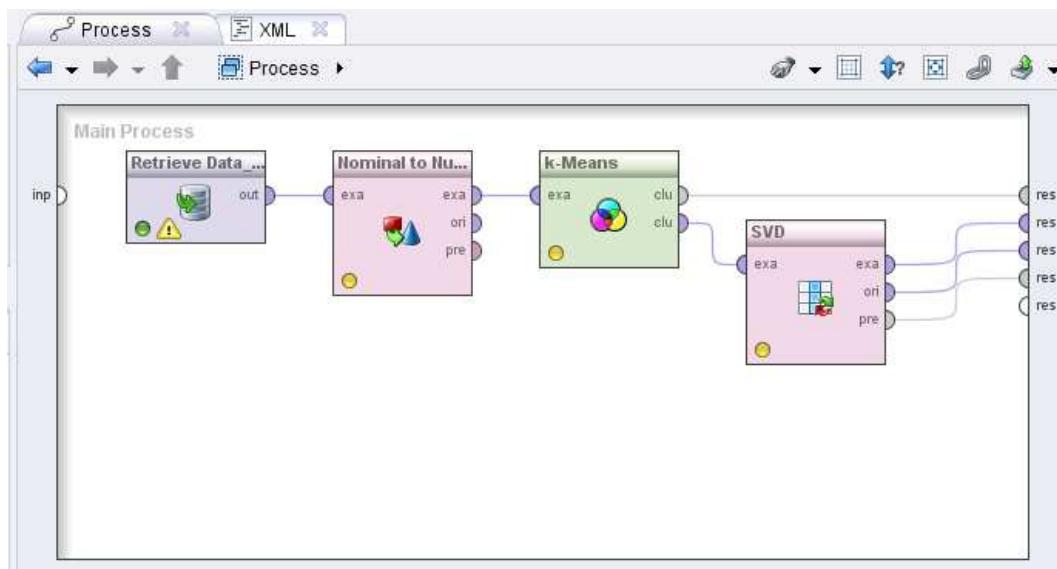
5. Gunakan Data_NilaiUjian ini dan masukkan ke dalam area process.
 6. Tambahkan operator **Modeling** → **Clustering and Segmentation** → **k-Means**. Ubah nama operator ini menjadi **k-Means**. Hubungkan output operator **Retrieve** ke entry exa operator ini dan output **clu** (cluster model) dihubungkan ke connector **res** panel.

Ubah nilai parameter **k = 3** pada operator ini. Angka ini digunakan untuk menentukan jumlah kelompok siswa.

7. Tambahkan pula operator **Data Transformation → Attribute Set Reduction and Transformation → Transformation → Singular Value Decomposition**. Hubungkan output **clu** (clustered set) ke-2 operator **k-Means** (k-Means) ke dalam entry **exa** operator SVD dan 3 port output **exa** (example set output), **ori** (original) dan **pre** (preprocessing model) terhadap connector panel **res** (result).



8. Lihatlah pada tabel “problems”. Terdapat sebuah permasalahan yaitu bahwa data input yang digunakan terdapat nilai data yang tidak bersifat numeric (yaitu atribut nomor dan nama siswa). Kita klik ganda pernyataan **“Convert attributes to numerical”** dalam tabel untuk mengubah tipe data menjadi numerik. Sehingga sebuah operator converter **Nominal to Numerical** akan ditambahkan dalam area Main Process.



9. Jalankan proses dengan menekan tombol **Run** (atau menekan tombol F11).

10. Berikut hasil proses Clustering dengan algoritma K-Means.

a. SVD (Singular Value Decomposition)

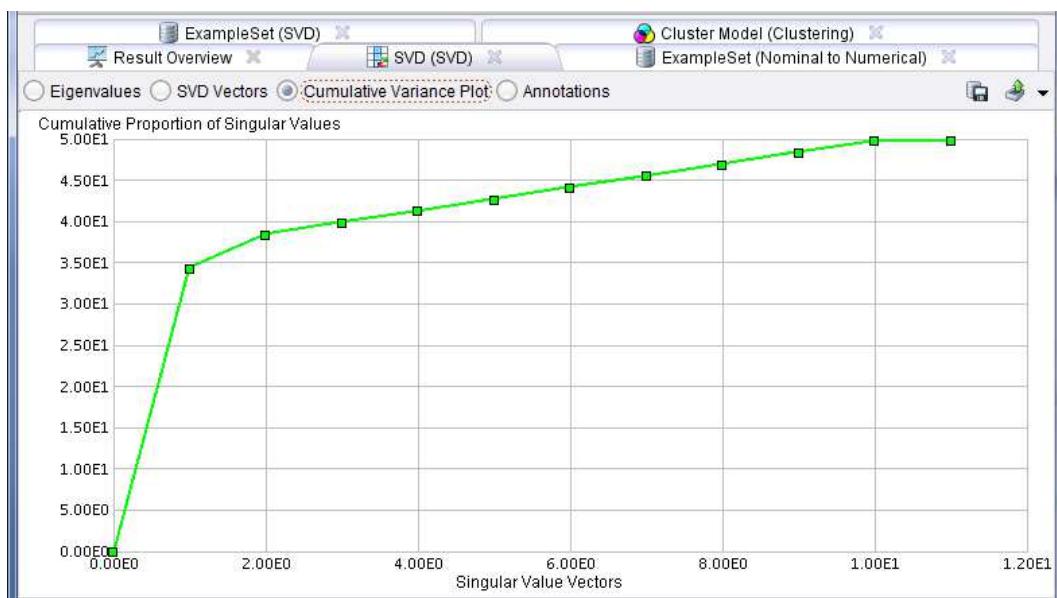
1) Nilai Eigenvalues dari 10 data nilai siswa

Component	Singular Value	Proportion of Singular Values	Cumulative Singular Values	Cumulative Proportion of Singular Values
SVD 1	34.368	0.690	34.368	0.690
SVD 2	4.151	0.083	38.519	0.773
SVD 3	1.414	0.028	39.933	0.801
SVD 4	1.414	0.028	41.348	0.830
SVD 5	1.414	0.028	42.762	0.858
SVD 6	1.414	0.028	44.176	0.886
SVD 7	1.414	0.028	45.590	0.915
SVD 8	1.414	0.028	47.005	0.943
SVD 9	1.414	0.028	48.419	0.972
SVD 10	1.414	0.028	49.833	1.000
SVD 11	0.000	0.000	49.833	1.000

2) Nilai SVD Vectors

Attribute	SVD Vector									
	1	2	3	4	5	6	7	8	9	10
B.IND	-0.723	-0.649	0.000	-0.000	-0.000	0.000	0.000	0.000	-0.000	-0.000
B.ING	-0.690	0.680	-0.000	0.000	0.000	-0.000	-0.000	0.000	0.000	0.000
NAMA = AGUS	-0.010	-0.121	-0.442	0.014	-0.046	-0.009	-0.089	-0.057	-0.000	-0.320
NAMA = DYAH	-0.007	0.101	0.205	-0.023	0.090	0.298	-0.266	-0.043	0.000	-0.408
NAMA = EKO	-0.009	0.016	0.120	-0.232	-0.299	-0.064	0.189	0.090	-0.000	-0.006
NAMA = IKAA	-0.009	-0.103	0.267	0.157	0.327	-0.011	-0.008	0.394	0.000	0.032
NAMA = JOKO	-0.010	0.011	-0.187	0.000	0.000	0.494	0.000	0.000	-1.000	0.399
NAMA = MAHMUD	-0.012	0.009	0.124	0.091	-0.338	-0.023	0.244	0.122	-0.000	-0.090
NAMA = SUSI	-0.009	0.144	-0.300	0.039	0.249	-0.280	0.079	0.177	0.000	0.035
NAMA = WATI	-0.007	0.012	0.095	0.542	-0.096	-0.132	-0.090	-0.307	-0.000	0.102
NAMA = WAJID	-0.010	-0.021	0.112	-0.295	-0.038	-0.258	-0.427	-0.116	-0.000	0.223

3) Cummulative Variance Plot (Graph)



b. Example Set (Nominal to Numerical)

Hasil ini kita lihat dengan mode Plot View menggunakan grafik Scatter untuk menentukan kelompok siswa (*cluster*) yang dicalonkan untuk maju ke dalam olimpiade mata pelajaran berdasarkan nilai tertinggi ujian.

Ketentuan:

Plotter = Scatter

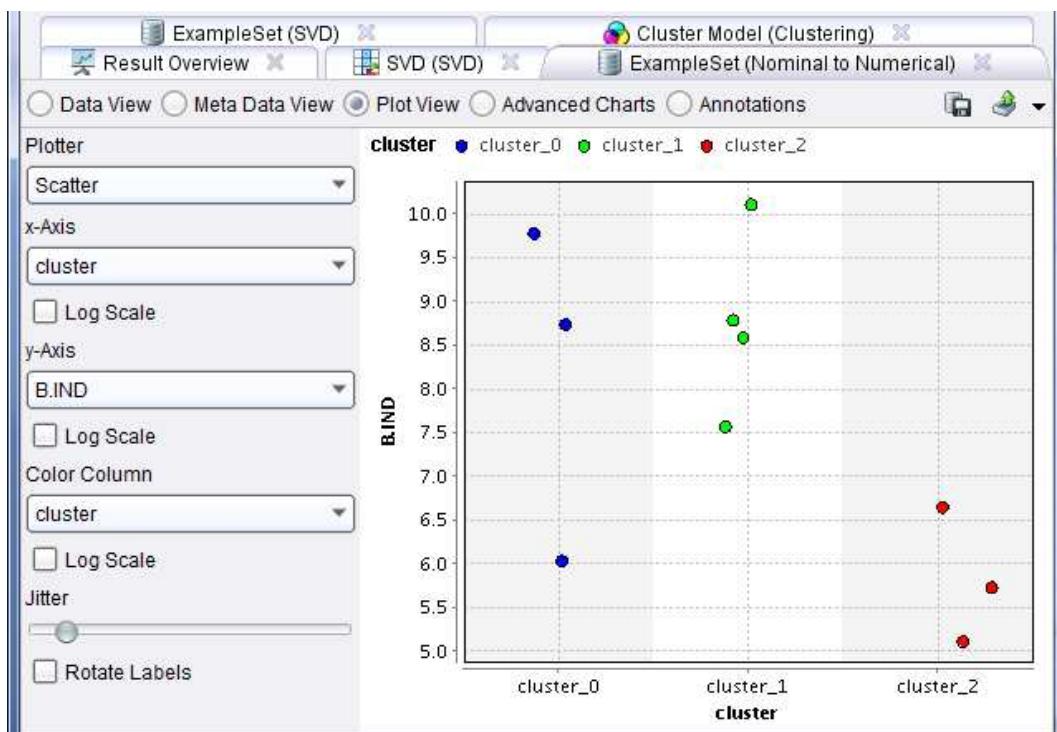
x-Axis = cluster

y-Axis = B.IND, B.ING (diubah-ubah)

Color Column = cluster

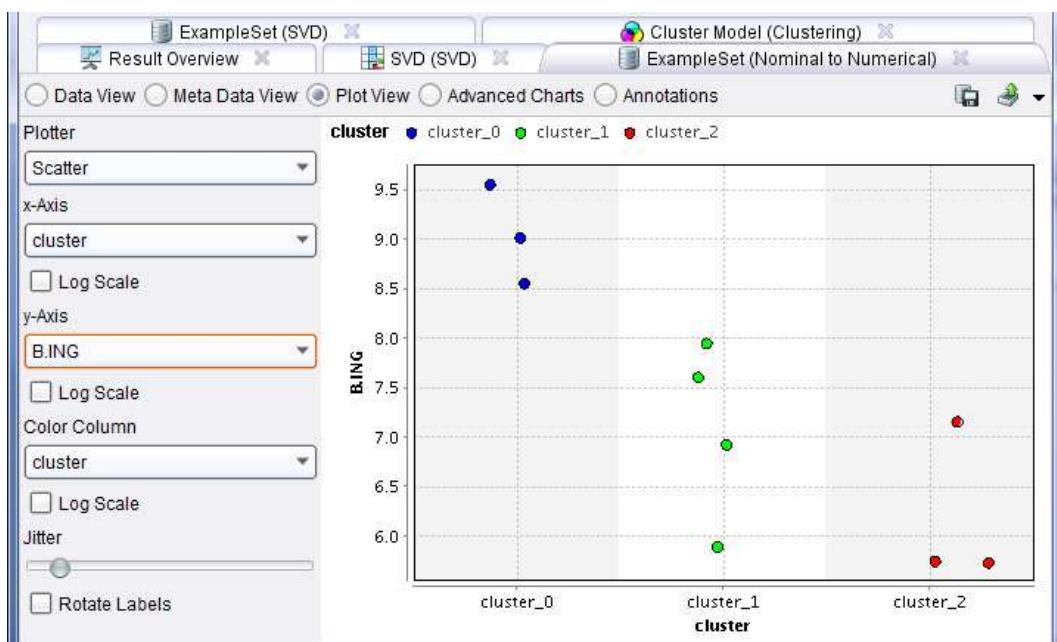
Jitter = bisa diubah-ubah untuk melihat distribusi data secara lebih detil.

1) Kelompok Siswa bidang B. Indonesia



Dapat kita lihat bahwa pada cluster_1 merupakan kelompok siswa yang memiliki nilai pelajaran B. Indonesia yang lebih tinggi dibandingkan dengan kelompok cluster_0 maupun cluster_2. Sehingga kelompok cluster 1 yang diajukan untuk lomba olimpiade bidang B. Indonesia.

2) Kelompok Siswa bidang B. Inggris



Dapat kita lihat bahwa pada cluster_0 merupakan kelompok siswa yang memiliki nilai pelajaran B. Inggris yang lebih tinggi dibandingkan dengan kelompok cluster_1 maupun cluster_2. Sehingga kelompok cluster 0 yang diajukan untuk lomba olimpiade bidang B. Inggris.

c. Example Set (SVD)

Pada hasil ini dilihat secara Data View. Klik pada header kolom *cluster* untuk mengurutkan data berdasarkan *cluster*.

Row No.	id	cluster	svd_1
1	1	cluster_0	-0.349
3	3	cluster_0	-0.315
10	10	cluster_0	-0.399
2	2	cluster_1	-0.347
6	6	cluster_1	-0.299
7	7	cluster_1	-0.317
9	9	cluster_1	-0.353
4	4	cluster_2	-0.256
5	5	cluster_2	-0.235
8	8	cluster_2	-0.254

Berdasarkan tabel ini dapat dilihat pembagian kelompok *cluster* siswa. Pada kolom id menunjukkan nomor urut siswa yang terdapat pada data asli. Sebagai contoh id=1 merupakan siswa nomor urut 1, id=3 merupakan siswa nomor urut 3 dan seterusnya.

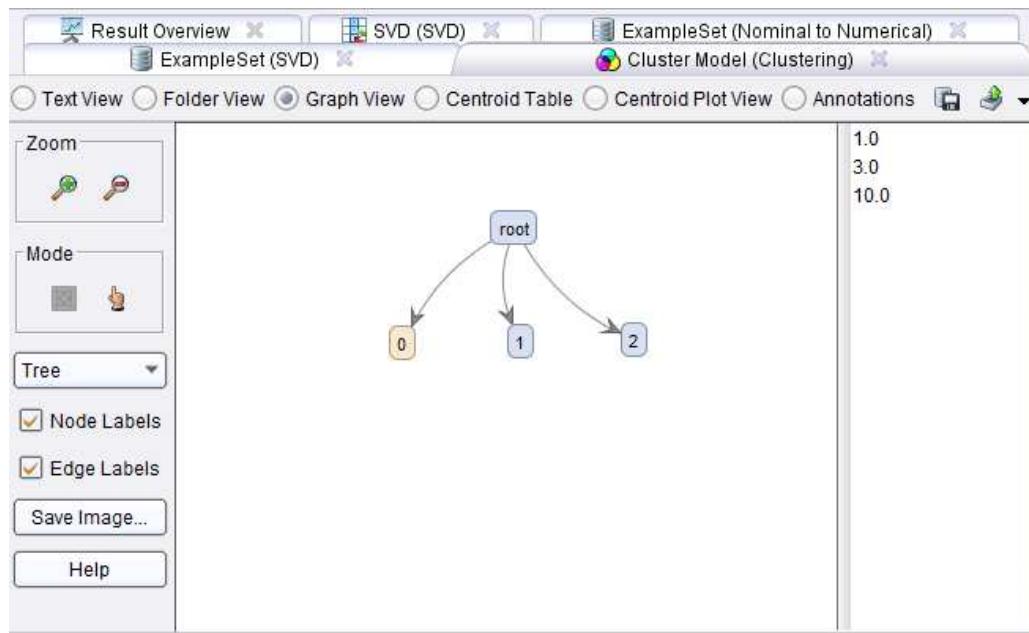
d. Cluster Model (Clustering)

1) Text View

Cluster 0: 3 items
Cluster 1: 4 items
Cluster 2: 3 items
Total number of items: 10

Pada cluster model dapat dilihat jumlah data pada masing-masing cluster. Pada cluster 0 memiliki sebanyak 3 siswa, cluster 1 memiliki 4 siswa dan cluster 2 memiliki sebanyak 3 siswa. Dengan total data sebanyak 10 siswa.

2) Graph View



Pada mode ini ditunjukkan bentuk pembagian cluster dengan pola pohon dan cabangnya. Setiap cluster yang terbentuk dapat dipilih (klik) untuk melihat anggota yang terdapat di dalamnya. Pada gambar tersebut dicontohkan pada cluster 0 terdapat 3 anggota siswa yaitu siswa ke-1, 3 dan 10.

3) Centroid Table

Attribute	cluster_0	cluster_1	cluster_2
B.IND	8.187	8.809	5.844
B.ING	9.043	7.129	6.224
NAMA = AGUS	0	0.250	0
NAMA = DYAH	0	0	0.333
NAMA = EKO	0	0.250	0

Dapat dilihat tabel data centroid berdasarkan cluster dan atributnya. Anda juga bisa melihat pembagian centroid menggunakan mode Centroid Plot View.

D.2. Interpretasi Hasil Algoritma K-Means

Berdasarkan hasil kegiatan D.1 dapat disimpulkan pembagian kelompok siswa yang akan diajukan untuk olimpiade Bahasa Indonesia dan Bahasa Inggris adalah sebagai berikut:

CLUSTER	NO_SISWA	NAMA	B.IND	B.ING
0	S-101	JOKO	8,54	8,40
0	S-103	SUSI	6,20	9,15
0	S-110	MAHMUD	9,81	9,58
1	S-102	AGUS	9,98	6,81
1	S-106	IKA	8,57	5,87
1	S-107	EKO	7,70	7,71
1	S-109	WAWAN	9,00	8,12
2	S-104	DYAH	5,24	7,26
2	S-105	WATI	5,70	5,71
2	S-108	YANTO	6,60	5,70

Pembagian kelompok yang diajukan untuk lomba olimpiade:

- 1) Cluster_1 yang diajukan untuk lomba olimpiade bidang B. Indonesia.
- 2) Cluster_0 yang diajukan untuk lomba olimpiade bidang B. Inggris.

E. Tugas (Dikerjakan saat ini, jika tidak selesai bisa dilanjutkan di rumah.

Jawaban dicetak pada kertas HVS, kecuali untuk interpretasi hasilnya harus ditulis tangan).

Dalam sebuah kelas terdapat 30 siswa yang telah menempuh ujian 4 mata pelajaran, yaitu Bahasa Indonesia, Bahasa Inggris, Matematika, dan IPA seperti dalam Tabel Data Nilai Ujian berikut.

- 1) Buatlah tabel berikut dengan menggunakan Microsoft Excel.

Tabel Data Nilai Ujian 30 Siswa:

NO_SISWA	NAMA	B.IND	B.ING	MTK	IPA
S-101	JOKO	=5+RAND()*5			
S-102	AGUS				
S-103	SUSI				
S-104	DYAH				
S-105	WATI				
S-106	IKA				
S-107	EKO				

S-108	YANTO				
S-109	WAWAN				
S-110	MAHMUD				
S-111	BUDI				
S-112	SANTI				
S-113	DIAN				
S-114	DANI				
S-115	AHMAD				
S-116	BAYU				
S-117	RISA				
S-118	RANI				
S-119	YANI				
S-120	RATIH				
S-121	INDAH				
S-122	JONO				
S-123	SARAH				
S-124	RAMA				
S-125	BAMBANG				
S-126	HADI				
S-127	NANA				
S-128	FEBRI				
S-129	DENI				
S-130	TONI				

Untuk mengisi daftar nilai dalam tabel, gunakan formula berikut pada salah satu sel. Kemudian bisa di copy-paste ke sel yang lain.

$$= 5 + \text{RAND}() * 5$$

(**Catatan:** setiap mahasiswa pasti akan memiliki data yang berlainan, sehingga hasilnya juga berbeda).

- 2) Lakukan kembali kegiatan D.1 dan D.2 pada modul 7 ini secara lengkap menggunakan data yang terdapat pada tabel **Tabel Data Nilai Ujian 30 Siswa** tersebut, dengan ketentuan jumlah Cluster = 4. Catat dan tulis semua hasilnya pada lembar jawaban anda, untuk gambar bisa di copy-paste.
- 3) Tulislah masing-masing nama siswa yang terdapat dalam Kelompok Cluster 0, Cluster 1, Cluster 2, dan Cluster 3.

MODUL 11

INDUKSI DAN ATURAN ASOSIASI

A. Tujuan

1. Mahasiswa mampu menggunakan induksi aturan, dan aturan asosiasi.
2. Mahasiswa mampu menerapkan aturan induksi aturan, dan aturan asosiasi dalam kasus nyata.

B. Landasan Teori

Rule induction adalah salah satu teknik dalam data mining yang paling sering digunakan untuk menemukan pengetahuan dalam sistem *unsupervised learning*. *Rule* (aturan) adalah bentuk sederhana dari “jika ini maka ini dan ini dan kemudian ini”. Sebagai contoh: jika seseorang membeli roti maka orang tersebut juga cenderung untuk membeli selai.

Agar aturan-aturan tersebut bermanfaat maka harus ditambahkan dua informasi tambahan sesuai dengan keadaaan sebenarnya yaitu:

1. Keakuratan (*accuracy / confidence*) yang menunjukkan seberapa sering aturan tersebut benar.
2. Penerapan (*coverage / support*) yaitu angka yang menunjukkan seberapa sering aturan tersebut dipakai.

Association Rule merupakan suatu proses untuk menemukan semua aturan assosiatif yang memenuhi syarat minimum untuk *support* (minsup) dan syarat minimum untuk *confidence* (minconf) pada sebuah database.

Dalam menentukan suatu *Association Rule* umumnya terdapat dua ukuran kepercayaan (*interestingness measure*), yaitu *support* dan *confidence*. Kedua ukuran ini akan digunakan untuk *interesting association rules* dibandingkan dengan batasan yang telah ditentukan. Batasan inilah yang terdiri dari minsup dan minconf. *Assosiation Rule Mining* adalah suatu prosedur untuk mencari hubungan antar item dalam suatu dataset. Dimulai dengan mencari frequent itemset, yaitu kombinasi yang paling sering terjadi dalam suatu itemset dan harus memenuhi minimum support.

Dalam tahap ini akan dicari kombinasi item yang memenuhi syarat minimum dari nilai support dalam database. Untuk mendapatkan nilai support untuk sebuah item A dapat diperoleh dari rumus berikut :

$$Support(A) = \frac{\text{Jumlah transaksi yang mengandung item } A}{\text{Total transaksi}}$$

Sementara itu, untuk mencari nilai support dari 2-item dapat diperoleh dari rumus berikut :

$$Support(A, B) = P(A \cap B)$$

$$P(A \cap B) = \frac{\text{Jumlah transaksi yang mengandung } A \text{ dan } B}{\text{Total transaksi}}$$

Setelah semua *frequent item* dan *Large itemset* ditemukan, dapat dicari semua *Association Rules* yang memenuhi syarat minimum untuk *confidence* (minconf) dengan menggunakan rumus berikut ini :

$$\text{Confidence } (A \rightarrow B) = P(B|A)$$

$$P(B|A) = \frac{\text{Jumlah transaksi yang mengandung } A \text{ dan } B}{\text{Jumlah transaksi yang mengandung item } A}$$

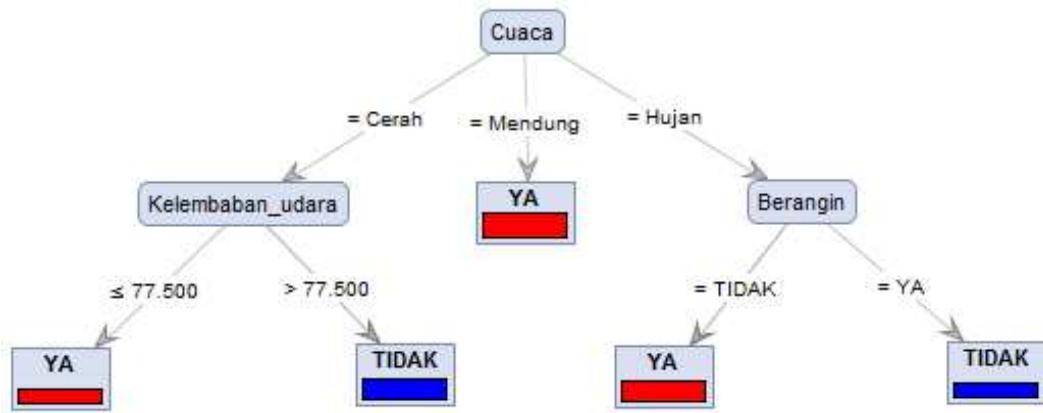
C. Alat dan Bahan

1. Komputer dengan sistem operasi Windows.
2. Program aplikasi RapidMiner.
3. Modul Praktikum Data Warehousing dan Data Mining.

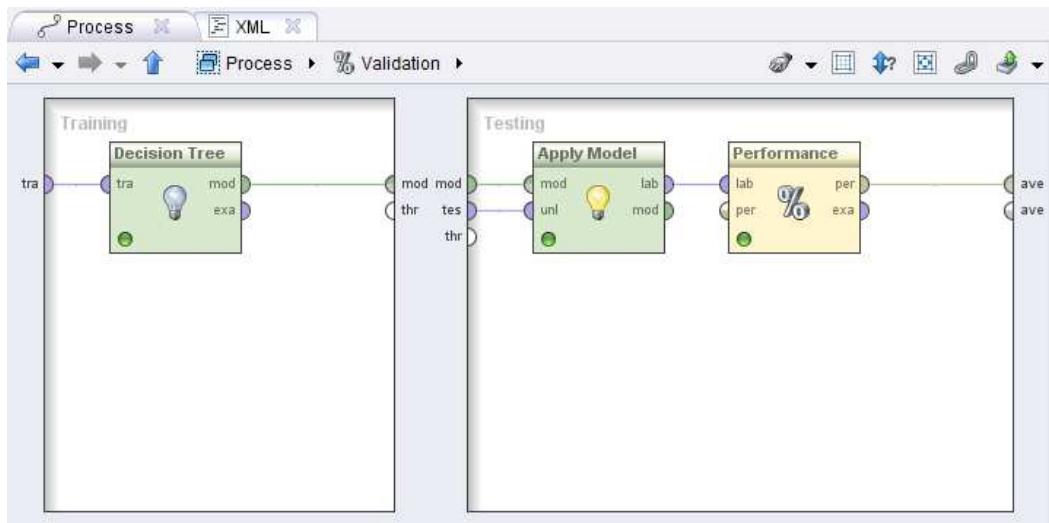
D. Langkah-langkah Praktikum

D.1. Induksi Aturan Data Cuaca

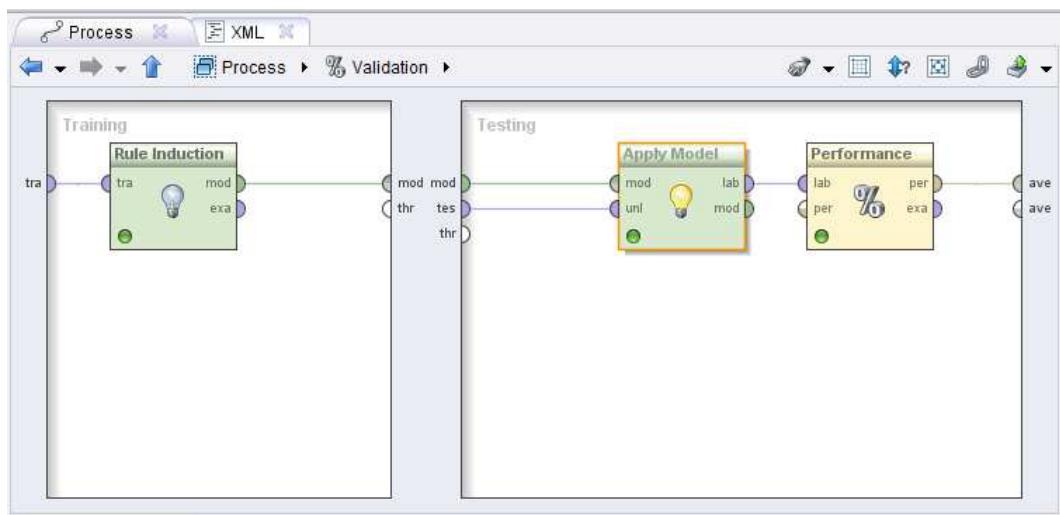
1. Jalankan aplikasi RapidMiner.
2. Gunakan model proses dari praktikum data mining **Modul 9 Kegiatan D.2.** (Jika tidak disimpan, lakukan kembali kegiatan tersebut).
3. Dari hasil kegiatan tersebut menghasilkan sebuah pohon keputusan seperti gambar berikut.



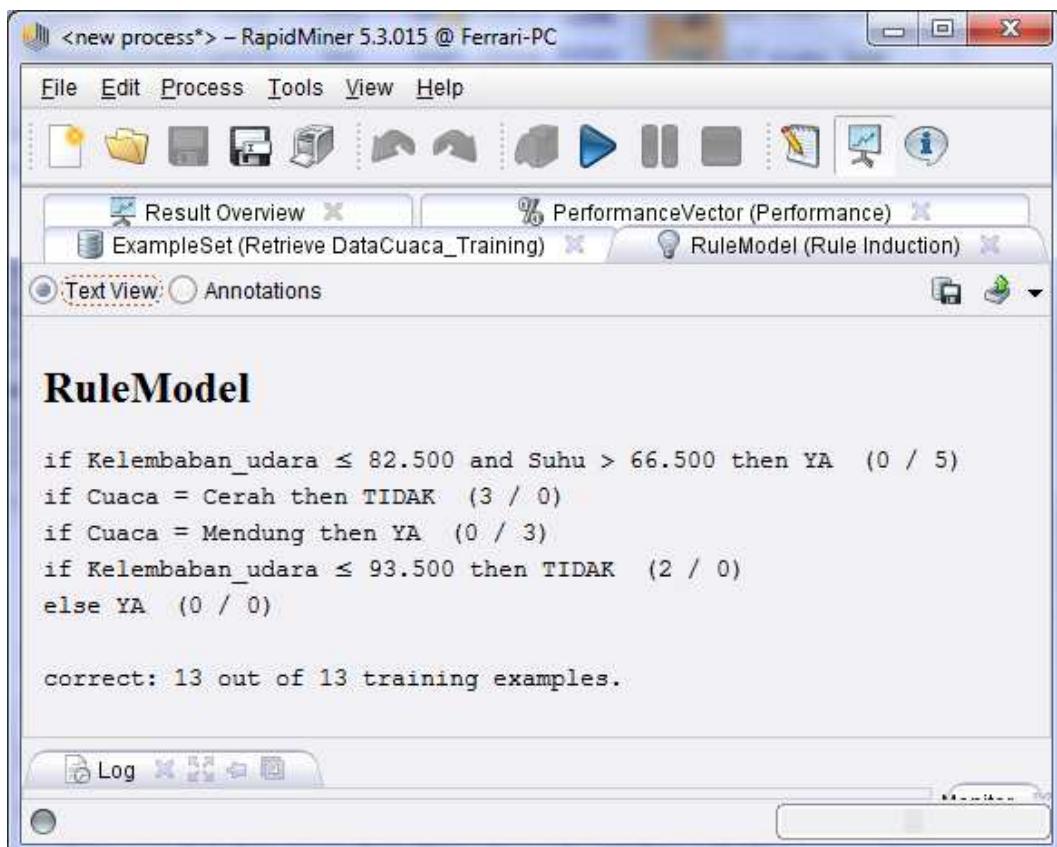
4. Kembali ke Process – Validation



5. Klik kanan operator **Decision Tree** pada area Training.
6. Pilih **Replace Operator** → **Modeling** → **Classification and Regression** → **Rule Induction** → **Rule Induction**.



7. Tanpa mengubah parameter-parameter yang lainnya, jalankan proses dengan menekan tombol **Run** (atau menekan tombol F11).
8. Sehingga akan diperoleh sebuah induksi aturan dari data training yang diberikan yang disebut sebagai *Rule Model (Rule Induction)*.



The screenshot shows the RapidMiner interface with the title bar <new process*> - RapidMiner 5.3.015 @ Ferrari-PC. The menu bar includes File, Edit, Process, Tools, View, Help. The toolbar has various icons for file operations and process steps. The central workspace shows a tab labeled 'RuleModel (Rule Induction)' which is active. Below it, there are tabs for 'Text View' and 'Annotations'. The main text area displays the generated rule model:

```

RuleModel

if Kelembaban_udara ≤ 82.500 and Suhu > 66.500 then YA (0 / 5)
if Cuaca = Cerah then TIDAK (3 / 0)
if Cuaca = Mendung then YA (0 / 3)
if Kelembaban_udara ≤ 93.500 then TIDAK (2 / 0)
else YA (0 / 0)

correct: 13 out of 13 training examples.

```

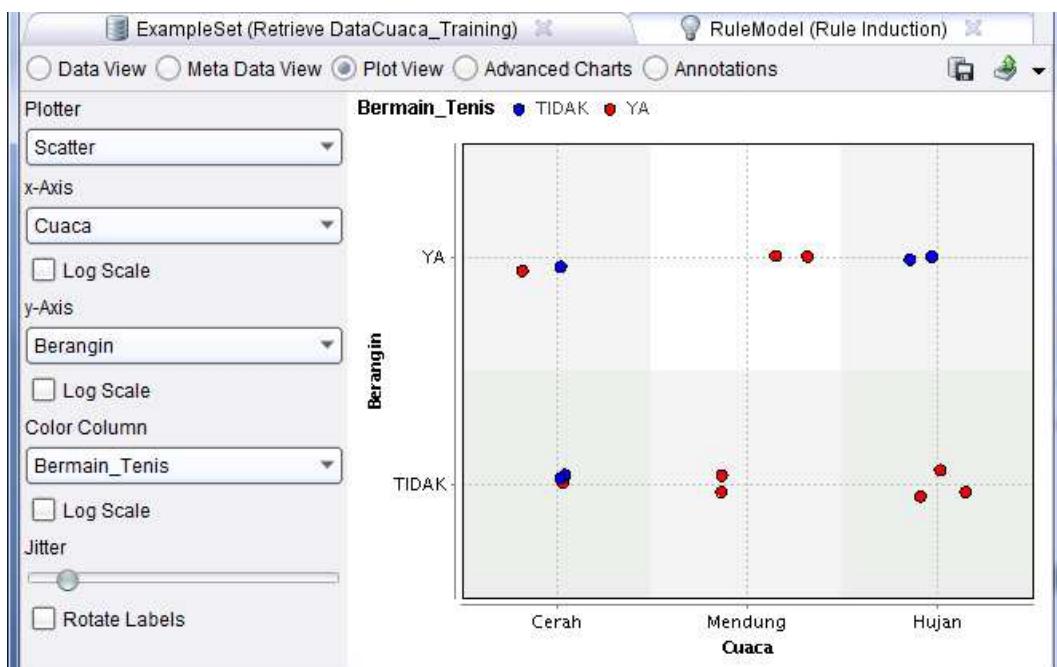
At the bottom, there is a 'Log' tab.

9. Dari hasil RuleModel tersebut dapat dilihat hubungan sebab akibat (Implikasi dengan menggunakan pernyataan “jika... maka...” atau “if... then...”) dari parameter-parameter yang diberikan. Aturan tersebut bisa digunakan sebagai dasar untuk pengambilan keputusan apakah seseorang akan bermain tenis atau tidak berdasarkan data cuaca yang ada.
10. Parameter *correct* menunjukkan jumlah data benar dalam Induksi Aturan terhadap total data yang digunakan sebagai data training.
11. Model *Rule Induction* ini juga bisa ditunjukkan hasil Performance Vector dan ExampleSet dengan menggunakan grafik Plot View.

Performance Vector

		true TIDAK	true YA	class precision
pred. TIDAK	2	0	100.00%	
pred. YA	3	9	75.00%	
class recall	40.00%	100.00%		

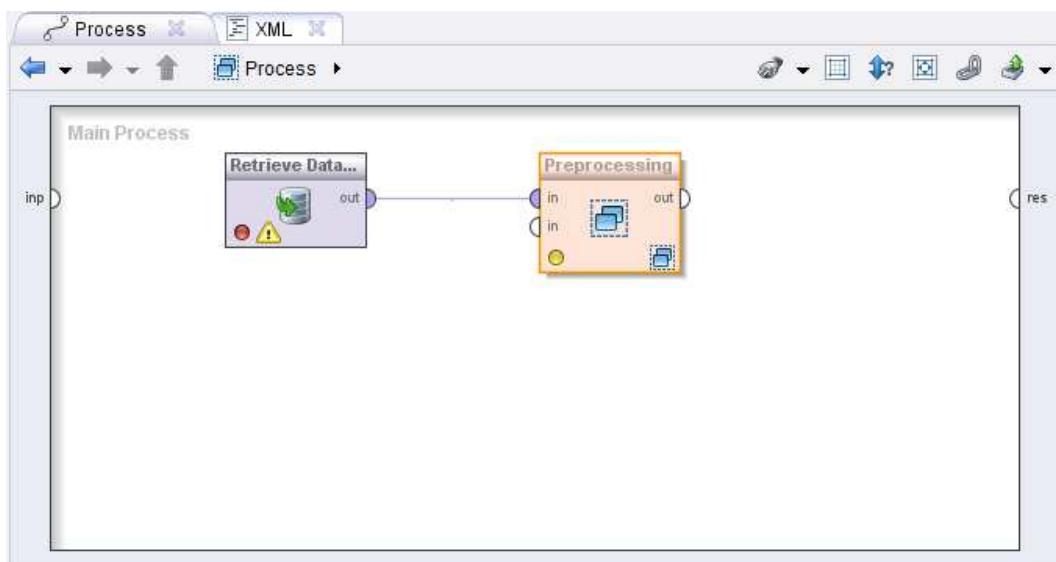
ExampleSet (Plot View)



Catatan: bisa dengan mengubah-ubah variabel pada x-Axis dan y-Axis. Namun pada **Color Column** harus berupa variabel *dependent*. Juga bisa mengubah nilai Jitter-nya untuk melihat secara lebih jelas pola sebaran datanya.

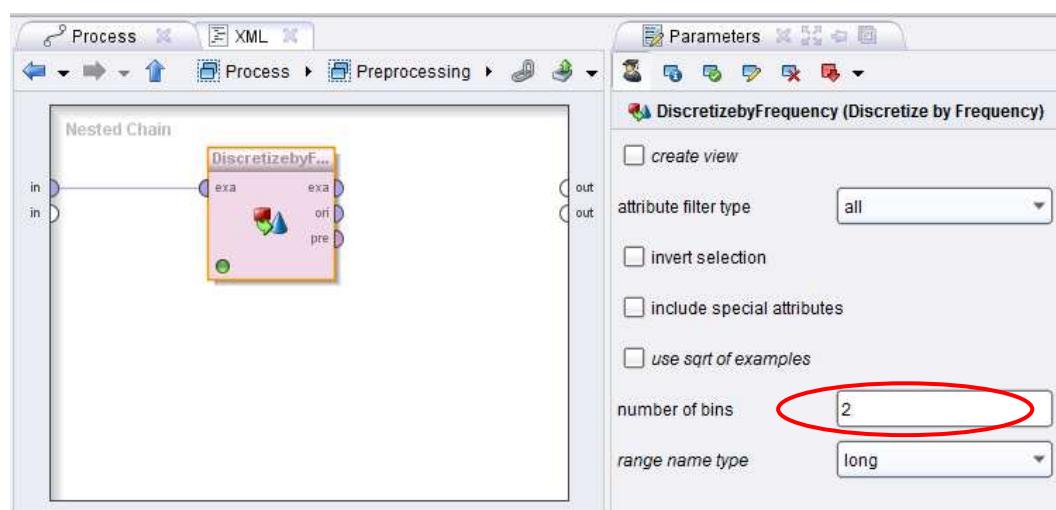
D.2. Aturan Asosiasi Data Cuaca

1. Masih dengan menggunakan RapidMiner, pilih **New Process**.
2. Gunakan **DataCuaca_Training** dan drag dari *repository* ke area **Process View**.
3. Tambahkan operator **Utility → Subprocess** ke dalam area. Ubah nama operator ini menjadi **Preprocessing** dengan klik kanan operator Subprocess – Rename. Hubungkan port output Retrieve dengan port input Preprocessing.



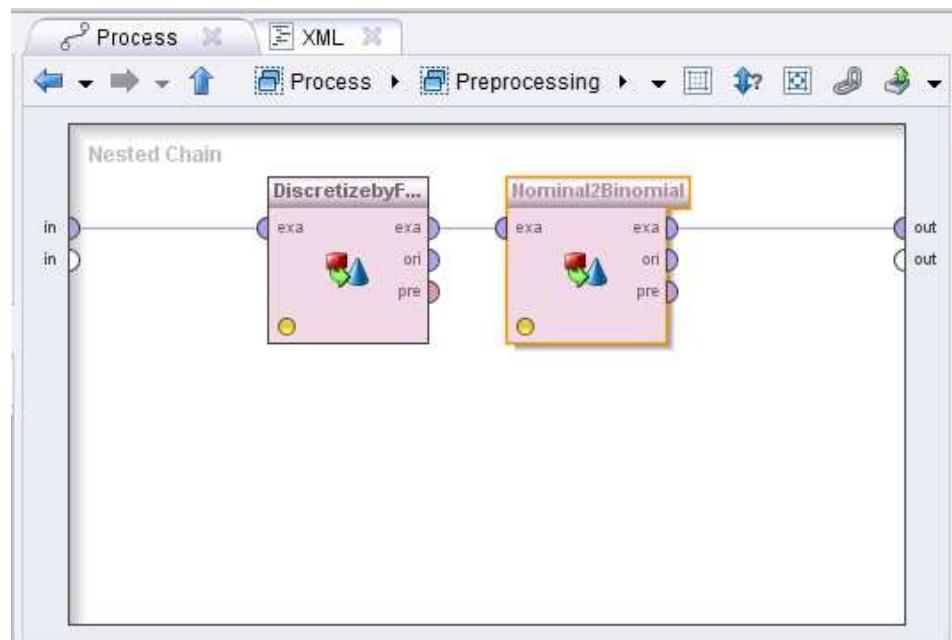
4. Klik ganda operator **Preprocessing** sehingga masuk pada area *Nested Chain*.
5. Pada area Nested Chain ini tambahkan operator-operator berikut:
 - a) **Data Transformation → Type Conversion → Discretization → Discretize by Frequency.**

Ubah nama operator ini menjadi “**DiscreatizationbyFrequency**” dan biarkan nilai parameter *number of bins* (jumlah interval) = 2 (anda juga bisa mengubah-ubah nilai ini, misal = 5). Hubungkan panel **in** pada masukan *example set* dengan port input *examination* operator ini.



b) Data Transformation → Type Conversion → Nominal to Binominal.

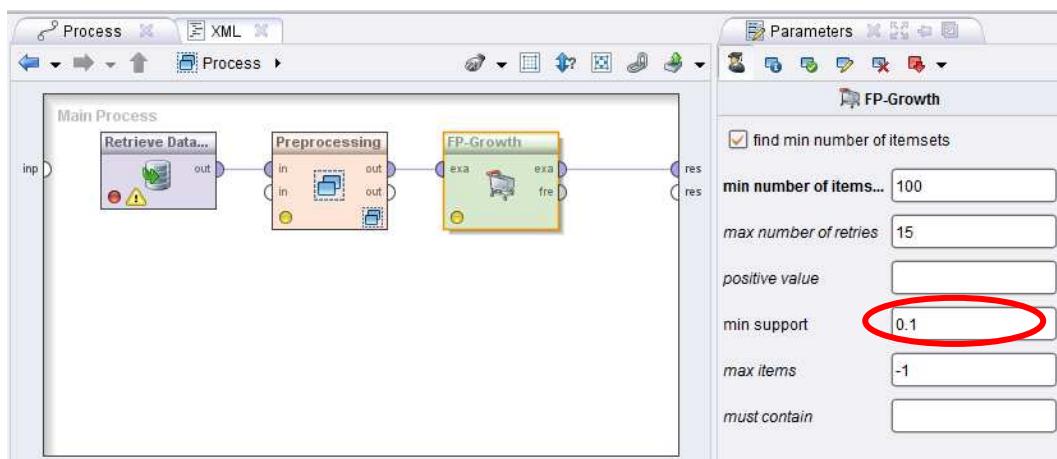
Ubah nama operator ini menjadi “Nominal2Binomial”, hubungkan output operator DiscretizationbyFrequency dengan masukan examination pada operator ini, dan output examination operator ini dengan panel out.



6. Kembali ke main process, tambahkan 2 buah operator:

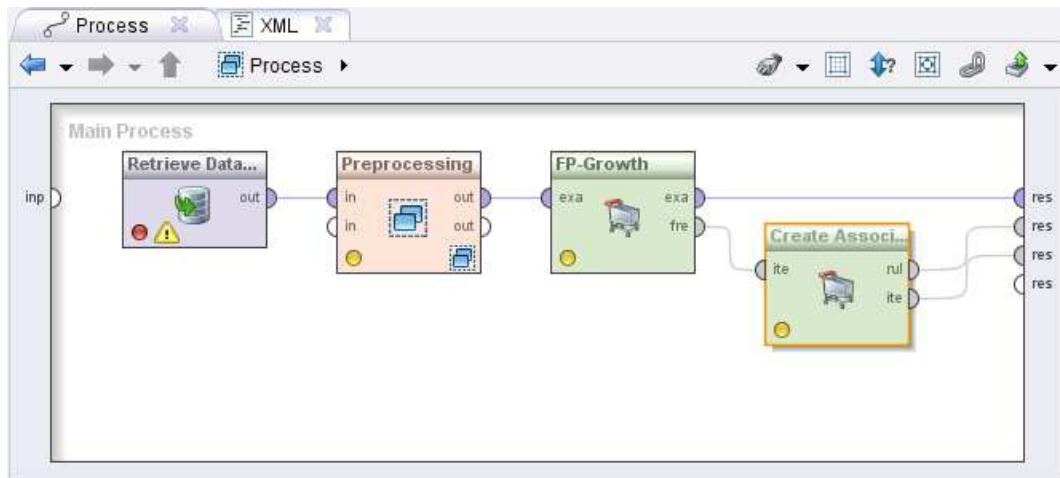
a) Modeling → Association and Item Set Mining → FP-Growth.

Ubah nilai parameter pada min support = 0.1, hubungkan port OUT operator Preprocessing dengan port entry **exa** pada operator ini dan output **exa** dengan connector **res** (result, resultat) pada work area.



b) Modeling → Association and Item Set Mining → Create Association Rules.

Hubungkan output **fre** (frequent sets, frequent sets) pada operator **FP-Growth** dengan masukan **ite** (item sets, itemsets) pada operator ini, output **rul** (rules) dengan connector **res** dalam work area, dan ouput **ite** (item sets) pada operator ini dengan connector **res** lainnya dalam work area.



7. Jalankan proses dengan menekan tombol **Run** (atau menekan tombol F11).
8. Dapat dilihat hasil-hasil aturan asosiasi sebagai berikut:

a) **Frequent Item Set (FP-Growth)**

No. of Sets: 26	Size	Support	Item 1	Item 2	Item 3	Item 4
Total Max. Size: 4	3	0.071	Kelembaban_udara	Berangin	Cuaca = Cerah	
	3	0.071	Kelembaban_udara	Berangin	Cuaca = Hujan	
	3	0.071	Kelembaban_udara	Berangin	Cuaca = Mendung	
Min. Size: 1	3	0.071	Kelembaban_udara	Suhu	Berangin	
	3	0.071	Kelembaban_udara	Suhu	Berangin	Cuaca = Cerah
Max. Size: 4	4	0.071	Kelembaban_udara	Suhu	Berangin	
Contains Item:	3	0.071	Kelembaban_udara	Suhu	Cuaca = Mendung	
	2	0.071	Suhu	Cuaca = Hujan		
	2	0.143	Berangin	Cuaca = Cerah		
	2	0.143	Berangin	Cuaca = Hujan		
	2	0.143	Kelembaban_udara	Cuaca = Mendung		
	2	0.143	Kelembaban_udara	Cuaca = Mendung		
	3	0.143	Kelembaban_udara	Suhu	Cuaca = Cerah	
	2	0.143	Suhu	Berangin		
	3	0.143	Suhu	Berangin	Cuaca = Cerah	
	2	0.143	Suhu	Cuaca = Mendung		
	2	0.214	Kelembaban_udara	Berangin		

Dapat dilihat bahwa jumlah aturan asosiasi yang terbentuk adalah 26 set, dan jumlah total max size = 4, yang terdiri dari 4 buah itemset. Kita bisa menyortir berdasarkan kolom dengan mengklik header kolom. Dengan menekan tombol CTRL selama mengklik, membuat kita dapat memilih beberapa kolom untuk disortir.

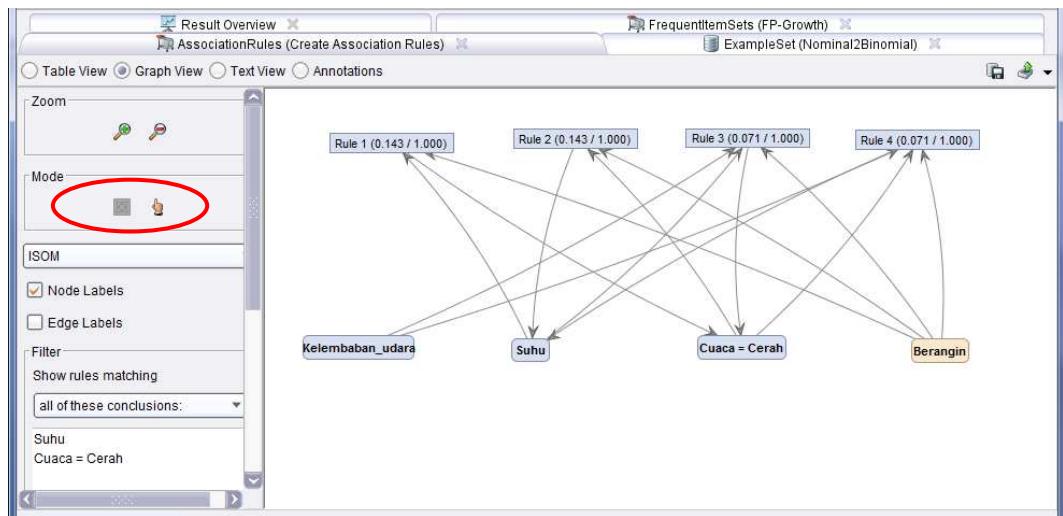
b) Association Rules (Create Association Rules)

1) Table View

No.	Premises	Conclusion	Support	Confiden...	LaPlace	Gain	p-s	Lift	Conviction
1	Suhu, Berangin	Cuaca = Cerah	0.143	1	1	-0.143	0.092	2.800	∞
2	Berangin, Cuaca = Cerah	Suhu	0.143	1	1	-0.143	0.082	2.333	∞
3	Kelembaban_udara, Suhu, Berangin	Cuaca = Cerah	0.071	1	1	-0.071	0.046	2.800	∞
4	Kelembaban_udara, Berangin, Cuaca = Cerah	Suhu	0.071	1	1	-0.071	0.041	2.333	∞

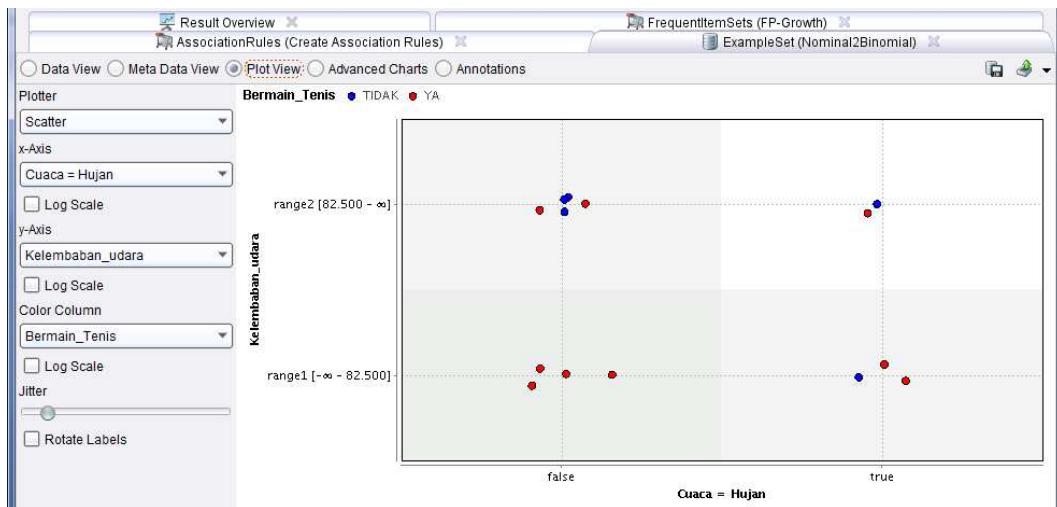
Pada tabel ini dapat dilihat terdapat 4 data pasangan premis-premis dan kesimpulan (*Premises and Conclusion*) yang menunjukkan hubungan implikasi berdasarkan aturan asosiasi. Aturan ini diperoleh hasil perhitungan asosiasi dengan nilai Support, Confidence, LaPlace, Gain, p-s, dan Lift pada masing-masing pernyataan.

2) Graph View



Pada mode ini diilustrasikan pola hubungan sebab akibat dalam bentuk grafik. Anda bisa mengatur gambar sedemikian rupa dengan mengklik tombol tangan pada pilihan mode untuk menggeser-geser node dalam grafik.

c) *Example Set (Nominal2Binomial) → Plot View*



Pada hasil ini dapat ditunjukkan pola distribusi masing-masing data cuaca yang diproses menggunakan aturan asosiasi. Kita bisa mengubah-ubah variabel dalam x-Axis dan y-Axis serta nilai Jitter. Namun pastikan bahwa pada *Color Column* adalah variabel *dependent* yaitu Bermain_Tenis.

E. Tugas (Dikerjakan saat ini, jika tidak selesai bisa dilanjutkan di rumah. Jawaban ditulis tangan pada kertas HVS, kecuali untuk tabel dan gambar dapat di copy-paste dari RapidMiner).

1. Dengan menggunakan data Lama Studi (format *Excel*) pada **Tugas Modul 6 Soal No. 1**, carilah pola hubungan berdasarkan *Induction Rule (Rule Model)*, dan nilai *performance vector*!
2. Masih dengan menggunakan data training yang sama, dengan ketentuan bahwa pada operator *Discretization by Frequency* memiliki nilai:
 - a) *number of bins* = 2
 - b) *number of bins* = **gunakan nilai yang lain**

Carilah masing-masing nilai berikut berdasarkan *number of bins*-nya:

- 1) Jumlah set aturan asosiasi dan total max size yang terbentuk berdasarkan FP-Growth (table view) ! Gambarkan tabelnya!
- 2) Jumlah data pasangan premis dan kesimpulan pada Association Rules (Create Association Rules) ! Gambarkan tabelnya! Gambarkan pula grafik yang terbentuk!
- 3) Gambarkan grafik Plot pola distribusi data pada Example Set yang terbentuk!

MODUL 12

REGRESI LINIER SEDERHANA

A. Tujuan

1. Mahasiswa mampu menggunakan metode regresi linier
2. Mahasiswa mampu melakukan analisis regresi linier
3. Mahasiswa mampu menerapkan metode regresi linier dalam kasus nyata.

B. Landasan Teori

Regresi linier adalah metode statistika yang digunakan untuk membentuk model hubungan antara variabel terikat (dependen; Y) dengan satu atau lebih variabel bebas (independent; X). Apabila banyaknya variabel bebas hanya ada satu, disebut sebagai regresi linier sederhana, sedangkan apabila terdapat lebih dari 1 variabel bebas, disebut sebagai regresi linier berganda.

Regresi linear sederhana ataupun regresi linier berganda pada intinya memiliki beberapa tujuan, yaitu :

1. Menghitung nilai estimasi rata-rata dan nilai variabel terikat berdasarkan pada nilai variabel bebas.
2. Menguji hipotesis karakteristik dependensi
3. Meramalkan nilai rata-rata variabel bebas dengan didasarkan pada nilai variabel bebas diluar jangkauan sampel.

Di dalam suatu model regresi akan ditemukan koefisien-koefisien. Koefisien pada model regresi sebenarnya adalah nilai duga parameter di dalam model regresi untuk kondisi yang sebenarnya (*true condition*), sama halnya dengan statistik *mean* (rata-rata) pada konsep statistika dasar. Hanya saja, koefisien-koefisien untuk model regresi merupakan suatu nilai rata-rata yang berpeluang terjadi pada variabel Y (variabel terikat) bila suatu nilai X (variabel bebas) diberikan.

Pada analisis regresi sederhana, ada beberapa asumsi dan persyaratan yang perlu diperiksa dan diuji, beberapa diantaranya adalah :

1. Variabel bebas tidak berkorelasi dengan *disturbance term (Error)*. Nilai disturbance term sebesar 0 atau dengan simbol sebagai berikut: $(E(U/X)) = 0$,

2. Jika variabel bebas lebih dari satu, maka antara variabel bebas (*explanatory*) tidak ada hubungan linier yang nyata,
3. Model regresi dikatakan layak jika angka signifikansi ANOVA sebesar < 0.05 ,
4. Variabel independen (*predictor*) yang digunakan sebagai variabel bebas harus layak. Kelayakan ini diketahui jika angka *Standard Error of Estimate* $<$ *Standard Deviation*,
5. Koefisien regresi harus signifikan. Pengujian dilakukan dengan Uji T. Koefisien regresi signifikan jika T hitung $>$ T table (nilai kritis).
Nilai T Table, F-Table dan lain-lain bisa dilihat dalam file **StatisticTable.pdf** yang bisa anda download di <http://yusufsn.staff.ums.ac.id/my-files> dalam folder Praktikum Data Mining.
6. Model regresi dapat diterangkan dengan menggunakan nilai koefisien determinasi ($KD = r^2 \times 100\%$) semakin besar nilai tersebut maka model semakin baik. Jika nilai mendekati 1 maka model regresi semakin baik,
7. Data harus berdistribusi normal,
8. Data berskala interval atau rasio,
9. Kedua variabel bersifat dependen, artinya satu variabel merupakan variabel bebas sedang variabel lainnya variabel terikat.

Prediksi terhadap suatu variabel Y (*dependent / response*) menggunakan metode regresi linier didasarkan pada nilai-nilai dari variabel lainnya, X (*independent / predictor*) serta hubungan antar dua variabel.

C. Alat dan Bahan

1. Komputer dengan sistem operasi Windows.
2. Program aplikasi RapidMiner.
3. Modul Praktikum Data Warehousing dan Data Mining.

D. Langkah-langkah Praktikum

Contoh kasus:

Dalam sebuah kelas yang memiliki 10 siswa dilakukan sebuah survei terhadap lama belajar seorang siswa dan nilai hasil ujiannya. Data siswa tersebut akan kita gunakan sebagai dasar perhitungan untuk memprediksi nilai ujian terhadap siswa lain berdasarkan lama belajarnya.

D.1. Mencari Nilai t-hitung dan Model Regresi Linier

Hipotesis:

Bagaimana mencari nilai t-hitung dan model regresi linier dari data siswa tersebut menggunakan RapidMiner.

Berikut tabel data siswa:

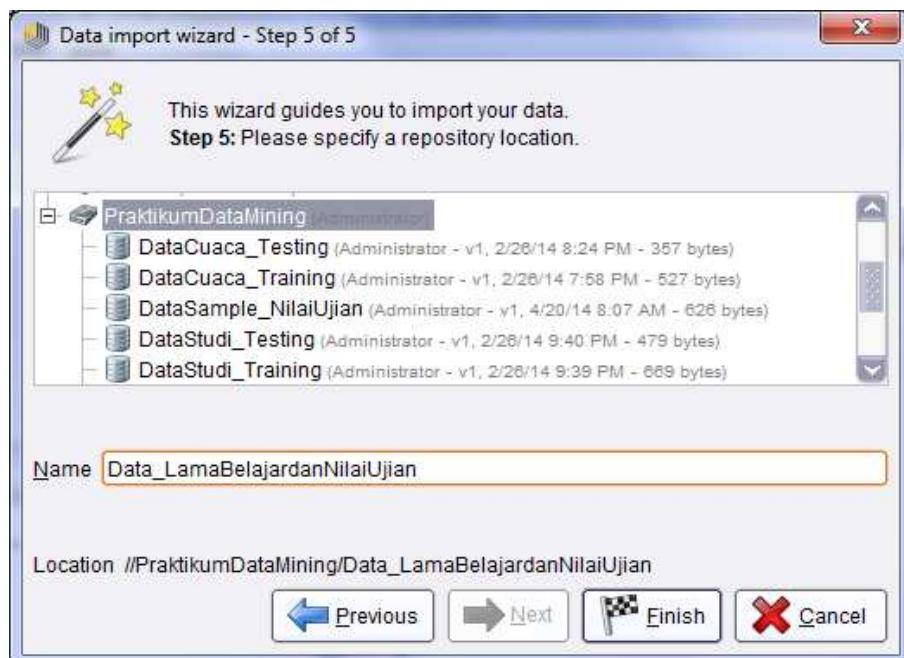
NO_SISWA	NAMA	LAMA BELAJAR (JAM)	NILAI
S-101	JOKO	15	783
S-102	AGUS	18	877
S-103	SUSI	7	505
S-104	DYAH	9	860
S-105	WATI	15	968
S-106	IKA	17	793
S-107	EKO	10	752
S-108	YANTO	5	571
S-109	WAWAN	8	667
S-110	MAHMUD	15	723

1. Buka Ms. Excel, dan buatlah tabel data siswa tersebut. Simpan dengan nama **Tabel_LamaBelajardanNilaiUjian.xls** (**Format Excel 2003 *.xls**).
2. Jalankan aplikasi **RapidMiner**.
3. Gunakan file **Tabel_LamaBelajardanNilaiUjian.xls** sebagai data yang akan digunakan dalam proses Regresi Linier. Import file ini ke dalam repositories Praktikum Data Mining (seperti pada Modul 8 Kegiatan D.2 Langkah 10-16). Yang perlu diperhatikan hanya pada saat penentuan tipe atribut pada Step 4.

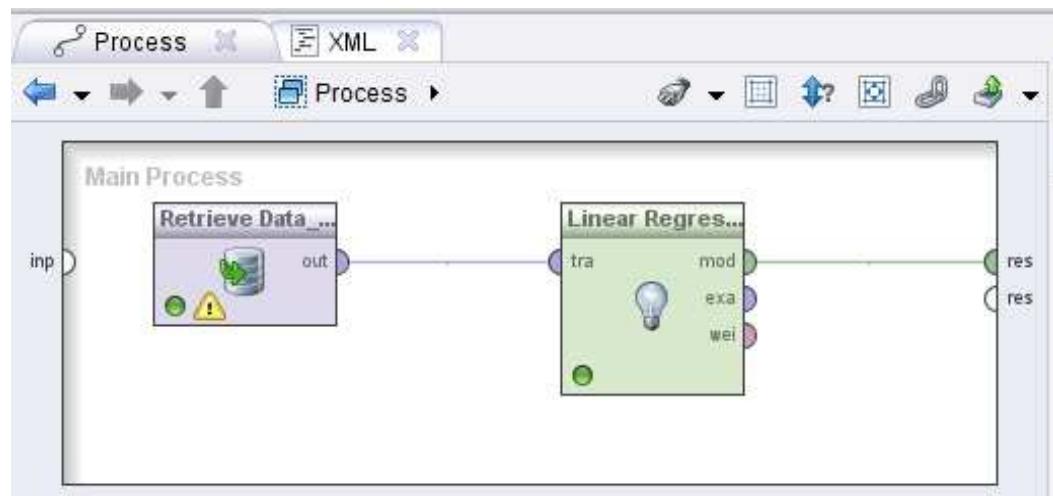
4. Ubah tipe data dan jenis masing-masing atribut sebagai berikut.
- NO_SISWA : text, id
 - NAMA : hilangkan tanda centang (*unchecked*)
 - LAMA_BELAJAR : integer, attribute
 - NILAI : integer, label

NO_SISWA	NAMA	LAMA_BELAJAR	NILAI
text	attribut...	integer	integer
id	attribute	attribute	label
S-101	JOKO	15	783
S-102	AGUS	18	877
S-103	SUSI	7	505
S-104	DYAH	9	860
S-105	WATI	15	968
S-106	IKA	17	793
S-107	EKO	10	752
S-108	YANTO	5	571
S-109	WAWAN	8	667
S-110	MAHMUD	15	723

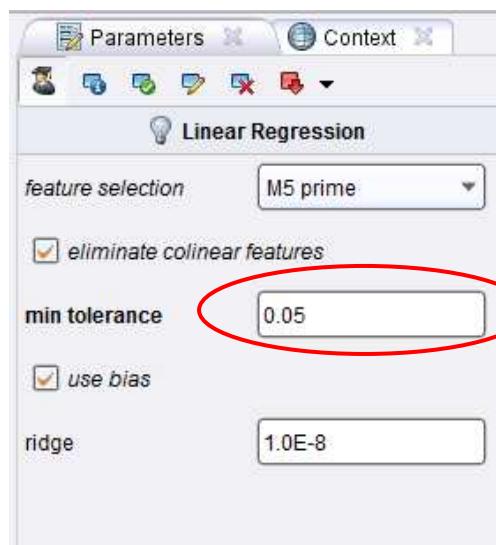
5. Pada step 5, beri nama **Data_LamaBelajardanNilaiUjian** dan masukkan pada repositories Praktikum Data Mining. Kemudian klik Finish.



6. Gunakan Data_LamaBelajardanNilaiUjian ini dan masukkan ke dalam area process.
7. Tambahkan operator **Modeling** → **Classification and Regression** → **Function Fitting** → **Linear Regression**. Hubungkan port **out** (output) operator **Retrieve Data** ke entry **tra** (training) operator ini dan 3 outputnya ke connector **res** panel.



8. Klik pada operator Linear Regression, tentukan parameter **min tolerance** = 0.05 (batas toleransi sebesar 5%)



9. Jika data input bertipe nominal atau polynomial, tambahkan operator **Data Transformation** → **Type Conversion** → **Nominal to Numerical** tepat setelah data input, sebelum operator Linear Regression. Set parameter *coding type* menjadi *unique integer*.
10. Jalankan proses dengan menekan tombol **Run** (atau menekan tombol F11).

11. Berikut hasil proses regresi linier:

a. **LinearRegression (Linear Regression)**

1) Table View (mencari besarnya nilai t-hitung)

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
LAMA_BELAJAR (JAM)	21.608	5.760	8.363	1	3.751	0.006	***
(Intercept)	492.769	69.640	?	?	7.076	0.000	****

Dari tabel tersebut dapat dilihat nilai t-statistic (t-hitung) sebesar 3.751. Berdasarkan aturan statistik, variabel X dikatakan mempengaruhi secara signifikan terhadap Y jika nilai t-hitung > t-table.

Berikut potongan tabel T:

t-table	min tolerance		
α (1 tail)	0.05	0.025	0.01
α (2 tail)	0.1	0.05	0.02
degree of freedom	Nilai Uji T		
1	6,3138	12,7065	31,8193
2	2,9200	4,3026	6,9646
.....
.....
7	1,8946	2,3646	2,9980
9	1,8331	2,2621	2,8214
10	1,8124	2,2282	2,7638

Keterangan:

α (1 tail) = regresi linier sederhana

α (2 tail) = regresi linier berganda

degree of freedom = jumlah sampel yang digunakan

Jika t-hitung = 3.751 sedangkan t-table = 1.812, maka $3.751 > 1.8124$ dengan nilai toleransi 5% (0,05). Sehingga dapat dikatakan bahwa Lama Belajar (X) mempengaruhi secara signifikan terhadap Nilai Ujian (Y).

- 2) Text View (mencari model regresi)

LinearRegression

21.608 * LAMA_BELAJAR (JAM)
+ 492.769

Berdasarkan hasil pada Text View terlihat sebuah persamaan berikut:

$$= 21.608 * \text{LAMA_BELAJAR (JAM)} + 492.769$$

Maka dapat dibuat sebuah model persamaan regresi linier sederhana untuk mencari nilai variabel Y (Nilai Ujian) berdasarkan variabel X₁ (Lama Belajar). Berikut model regresi linier yang terbentuk:

$$Y = 21,608 X_1 + 492,769$$

Dengan model tersebut, dapat dicari Nilai Ujian (Y) dengan memasukkan nilai Lama Belajar pada variabel X₁.

D.2. Mencari Nilai t dan Model Regresi Linier menggunakan RapidMiner

Hipotesis:

Bagaimana memprediksi nilai ujian siswa berdasarkan lama belajarnya menggunakan model regresi linier yang telah dihasilkan menggunakan RapidMiner.

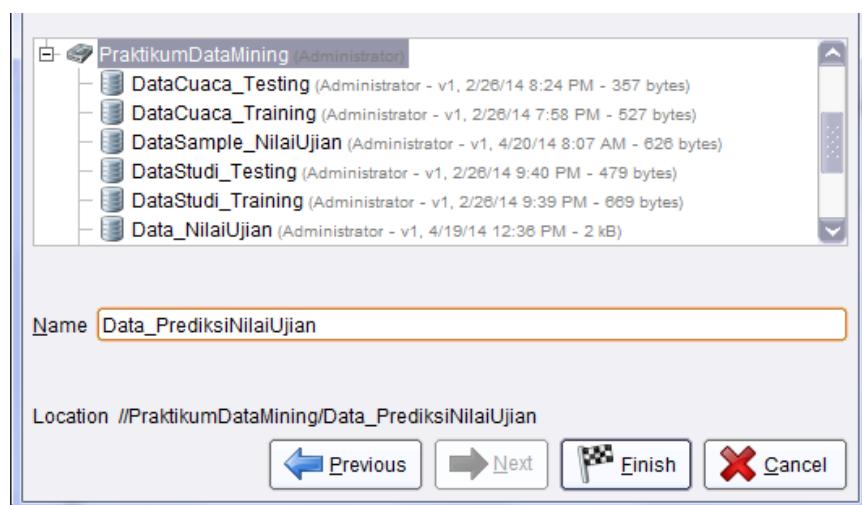
Berikut tabel data siswa:

NO_SISWA	NAMA	LAMA BELAJAR (JAM)
S-111	BUDI	12
S-112	SANTI	13
S-113	DIAN	14
S-114	DANI	11
S-115	AHMAD	5
S-116	BAYU	13
S-117	RISA	9
S-118	RANI	10
S-119	YANI	10
S-120	RATIH	9

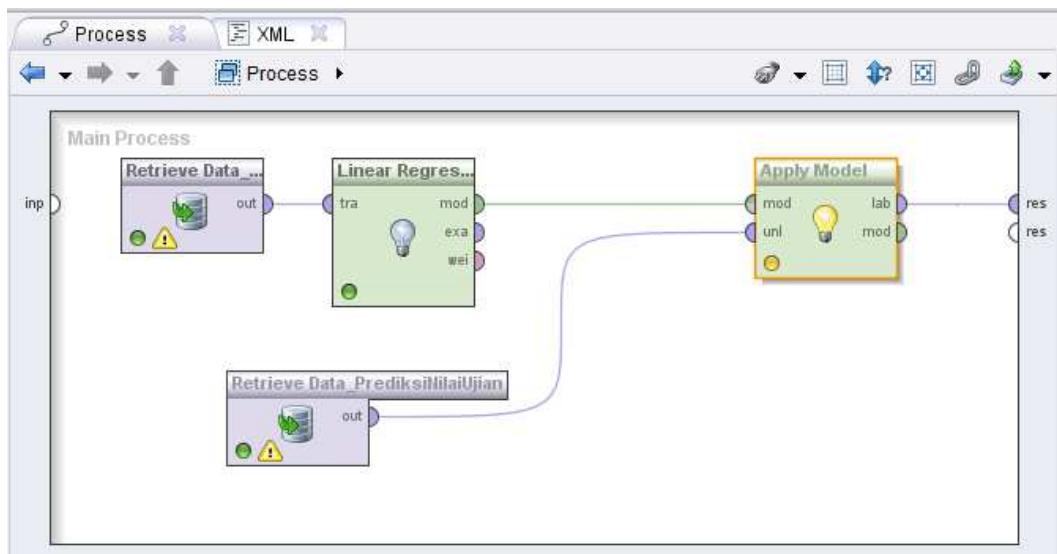
1. Buka Ms. Excel, dan buatlah tabel data siswa tersebut. Simpan dengan nama **Tabel_PrediksiNilaiUjian.xls** (**Format Excel 2003 *.xls**).
2. Jalankan aplikasi **RapidMiner**.
3. Gunakan file **Tabel_PrediksiNilaiUjian.xls** sebagai data testing. Import file ini ke dalam repositories Praktikum Data Mining.
4. Ubah tipe data dan jenis masing-masing atribut sebagai berikut.
 - a. NO_SISWA : text, id
 - b. NAMA : hilangkan tanda centang (*uncheck*)
 - c. LAMA_BELAJAR : integer, attribute

<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
NO_SISWA	NAMA	LAMA_BELAJAR (JAM)
text	polynomial	integer
id	attribute	attribute
S-111	BUDI	12
S-112	SANTI	13
S-113	DIAN	14
S-114	DANI	11
S-115	AHMAD	5
S-116	BAYU	13
S-117	RISA	9
S-118	RANI	10
S-119	YANI	10
S-120	RATIH	9

5. Pada step 5, beri nama **Data_PrediksiNilaiUjian** dan masukkan pada repositories Praktikum Data Mining. Kemudian klik Finish.



6. Tetap menggunakan proses pada kegiatan D.1, masukkan **Data_PrediksiNilaiUjian** ini ke dalam area process.
7. Tambahkan operator **Modeling → Model Application → Apply Model**, letakkan setelah operator Linear Regression. Hubungkan port-port output dan input seperti gambar berikut.



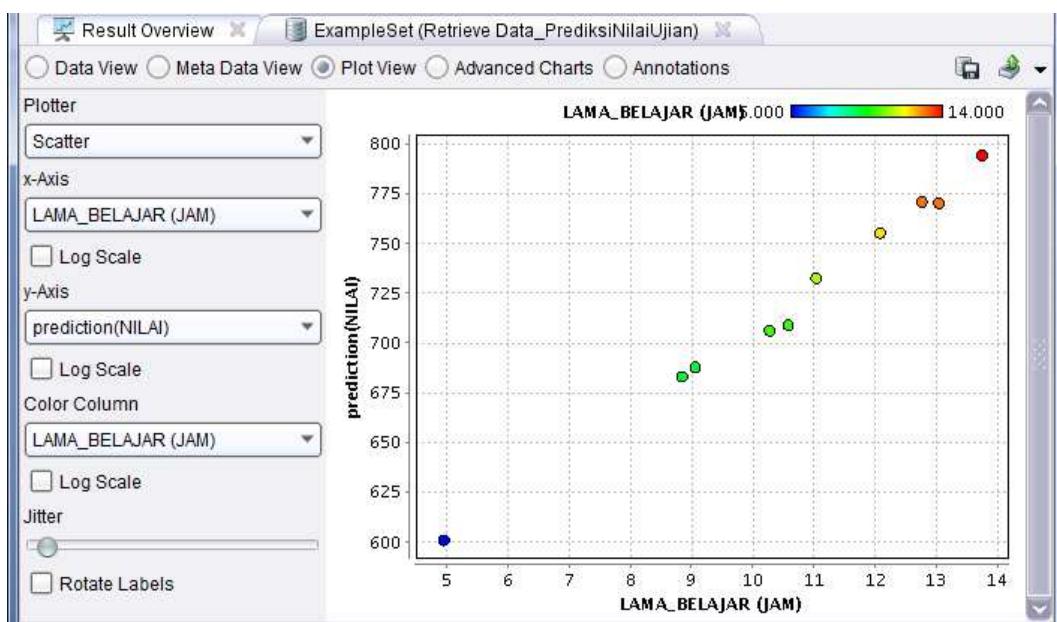
8. Dengan menggunakan parameter yang sama pada operator Regression Linear, jalankan proses dengan menekan tombol **Run** (atau menekan tombol F11).
9. Berikut hasil proses prediksi terhadap data testing menggunakan regresi linier:

a) **Table View (hasil prediksi nilai ujian)**

Row No.	NO_SISWA	prediction(NILAI)	LAMA_BELAJAR (JAM)
1	S-111	752.061	12
2	S-112	773.668	13
3	S-113	795.276	14
4	S-114	730.453	11
5	S-115	600.807	6
6	S-116	773.668	13
7	S-117	687.238	9
8	S-118	708.845	10
9	S-119	708.845	10
10	S-120	687.238	9

Hasil prediksi nilai ujian terhadap 10 siswa lainnya

b) Plot View (Scatter Plot)



D.3. Pembuktian Model Regresi

Pada tahap ini kita akan melakukan pembuktian hasil prediksi menggunakan RapidMiner dengan hasil prediksi menggunakan model regresi yang telah dihasilkan pada kegiatan D.1 berikut.

$$Y = 21,608 X_1 + 492,769$$

	A	B	C	D	E
1	NO_SISWA	NAMA	LAMA BELAJAR (JAM)	Prediction (NILAI)	Prediction (NILAI)
2				Tabel	Model Regresi
3	S-111	BUDI	12	752,061	
4	S-112	SANTI	13	773,668	
5	S-113	DIAN	14	795,276	
6	S-114	DANI	11	730,453	
7	S-115	AHMAD	5	600,807	
8	S-116	BAYU	13	773,668	
9	S-117	RISA	9	687,238	
10	S-118	RANI	10	708,845	
11	S-119	YANI	10	708,845	
12	S-120	RATIH	9	687,238	
13					

Memasukkan nilai variabel X_1 ke dalam model regresi (3 contoh data):

1. No. Siswa = S-111, Nama = Budi, memiliki nilai $X_1 = 12$ Jam.

Sehingga, $Y = (21,608 \times 12) + 492,769 = 752,065$

2. No. Siswa = S-115, Nama = Ahmad, memiliki nilai $X_1 = 5$ Jam.

Sehingga, $Y = (21,608 \times 5) + 492,769 = 600,809$

3. No. Siswa = S-120, Nama = Ratih, memiliki nilai $X_1 = 9$ Jam.

Sehingga, $Y = (21,608 \times 9) + 492,769 = 687,241$

Berikut tabel hasil pembuktiannya.

Gunakan Ms. Excel untuk melakukan pembuktian. Pada sel **E3** masukkan formula “ $=(21,608*C3)+492,769$ ”. Copy dan Paste pada sel **E4** sampai **E12**.

The screenshot shows a Microsoft Excel spreadsheet with data in rows 3 to 12. The columns are labeled A through E. Column A contains student numbers (S-111 to S-120). Column B contains names (BUDI, SANTI, DIAN, DANI, AHMAD, BAYU, RISA, RANI, YANI, RATIH). Column C contains study times in hours (12, 13, 14, 11, 5, 13, 9, 10, 10, 9). Column D contains predicted scores from the regression model (752,061, 773,668, 795,276, 730,453, 600,807, 773,668, 687,238, 708,845, 708,845, 687,238). Column E contains the same predicted scores as column D, labeled "Model Regresi". The formula $=21,608*C3+492,769$ is shown in the formula bar above row 3, and the cell E3 contains the result 752,065.

	A	B	C	D	E
1	NO_SISWA	NAMA	LAMA BELAJAR (JAM)	Prediction (NILAI)	Prediction (NILAI)
2				Tabel	Model Regresi
3	S-111	BUDI	12	752,061	752,065
4	S-112	SANTI	13	773,668	773,673
5	S-113	DIAN	14	795,276	795,281
6	S-114	DANI	11	730,453	730,457
7	S-115	AHMAD	5	600,807	600,809
8	S-116	BAYU	13	773,668	773,673
9	S-117	RISA	9	687,238	687,241
10	S-118	RANI	10	708,845	708,849
11	S-119	YANI	10	708,845	708,849
12	S-120	RATIH	9	687,238	687,241
13					

Dapat dilihat bahwa hasil prediksi menggunakan RapidMiner menghasilkan nilai yang sama dengan menggunakan Model Persamaan Regresi Linier yang telah dihasilkan pada kegiatan D.1.

- E. Tugas (Dikerjakan saat ini, jika tidak selesai bisa dilanjutkan di rumah. Jawaban dicetak pada kertas HVS, kecuali untuk pembuktian hasil menggunakan Model Persamaan Regresi Linier harus ditulis tangan).**

Kasus:

Dalam sebuah survei terhadap 15 kepala keluarga telah diperoleh variabel pendapatan rata-rata perbulan, jumlah anggota keluarga yang tinggal serumah, dan daya beli rata-rata perbulan.

Hipotesis:

Bagaimanakah model regresi linier yang terbentuk, dan lakukan prediksi terhadap 10 data yang belum diketahui nilai daya belinya.

- 1) Buatlah tabel berikut dengan menggunakan Microsoft Excel.

Tabel Hasil Survei 15 Kepala Keluarga:

NO. RESPONDEN	PENDAPATAN (RUPIAH)	JUMLAH ANGGOTA KELUARGA	DAYA BELI (RUPIAH)
1	1.000.000	6	834.000
2	1.400.000	7	1.200.000
3	200.000	3	134.000
4	1.400.000	6	1.167.000
5	500.000	3	334.000
6	1.700.000	5	1.360.000
7	400.000	3	267.000
8	1.900.000	5	1.520.000
9	300.000	3	200.000
10	500.000	4	375.000
11	700.000	7	600.000
12	1.900.000	3	1.267.000
13	800.000	4	600.000
14	1.500.000	4	1.125.000
15	1.300.000	7	1.115.000

- 2) Buatlah proses Regresi Linier Sederhana menggunakan RapidMiner dengan ketentuan sebagai berikut.
 - a. Variabel bebas (X) = Pendapatan (X_1), Jumlah Anggota Keluarga (X_2)
 - b. Variabel terikat (Y) = Daya Beli
 - c. Toleransi yang digunakan = 5%

- 3) Tentukan apakah variabel X_1 dan X_2 mempengaruhi secara signifikan terhadap nilai variabel Y berdasarkan besarnya nilai t-stat?
- 4) Tuliskan model persamaan regresi linier sederhana yang terbentuk!
- 5) Gunakan data testing untuk menjawab perintah berikut:
 - a) Lakukan prediksi Daya Beli (Y) dengan menggunakan Model Persamaan Regresi Linier dari hasil pertanyaan nomor 4 !
 - b) Lakukan prediksi Daya Beli (Y) menggunakan RapidMiner !

Data testing yang digunakan untuk prediksi:

NO. RESPONDEN	PENDAPATAN (RUPIAH)	JUMLAH ANGGOTA KELUARGA
1	900.000	5
2	800.000	3
3	500.000	2
4	1.900.000	6
5	600.000	2
6	800.000	5
7	1.000.000	6
8	1.100.000	4
9	1.000.000	4
10	500.000	3

- 6) Gambarkan pola sebaran data menggunakan Plot View (Scatter) dengan ketentuan berikut:
 - a) x-Axis = Pendapatan (Rupiah),
y-Axis = Prediction (Daya Beli (Rupiah)),
Color Column = Prediction (Daya Beli (Rupiah))
 - b) x-Axis = Jumlah Anggota Keluarga,
y-Axis = Prediction (Daya Beli (Rupiah)),
Color Column = Prediction (Daya Beli (Rupiah))