LAPORAN JOBSHEET PENGAMBILAN DATA (EXTRACT) DAN TRANSFORMASI DATA (ETL) DENGAN PDI SPOON



DIBUAT OLEH:
RIZQI DEWANTO (2341760176)

A. Proses Pengambilan Data (Extract)

1. Membuka Lembar Kerja Transformation

- Buka aplikasi PDI Spoon.
- o Pilih menu **File** > **New** > **Transformation**.

2. Menambahkan Objek CSV File Input

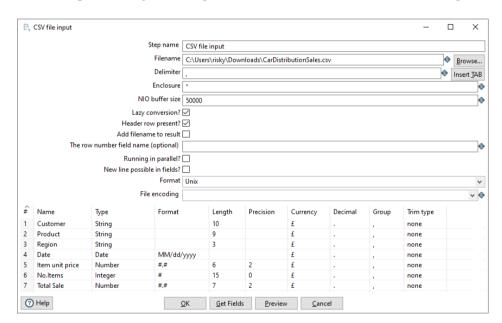
- o Pada **Design Area**, cari objek **CSV File Input**.
- Drag dan drop objek tersebut ke **Work Area**.

3. Konfigurasi CSV File Input

- o Double-click pada objek **CSV File Input** untuk membuka jendela konfigurasi.
- Ubah nama Step Name menjadi Input Data CSV.
- o Pilih lokasi file **CarDistributionSales.csv** melalui tombol **Browse**.
- o Pastikan **Delimiter** diatur sesuai dengan pemisah dalam file, dalam hal ini **comma** (,).
- Tekan **Get Fields** untuk mengambil judul kolom dari file.
- o Verifikasi apakah semua nama kolom sudah sesuai.
- Klik **OK** untuk menyimpan konfigurasi.

4. Menjalankan Proses Extract Data

- o Tekan tombol **Run** di pojok kanan atas.
- o Pada Execution Result Area, pilih tab Preview Data.
- o Jika proses berjalan dengan benar, data akan muncul sesuai dengan file asli.



Ŷ	Customer	Product	Region	Date	Item unit price	No.ltems	Total Sale
1	Customer 1	Yaris AT	MLG	03/13/2019	340.9	6	2045.7
2	Customer 1	Etios AT	SBY	04/16/2019	800	1	800
3	Customer 3	Avanza AT	MLG	06/18/2019	168.9	7	1182.7
4	Customer 3	<null></null>	MLG	07/24/2019	168.9	2	337.9
5	Customer 1	Avanza AT	SBY	08/21/2019	168.9	9	1520.5
6	Customer 2	Etios AT	KDR	09/07/2019	800	7	5599.6
7	Customer 3	Etios AT	MLG	09/07/2019	800	8	6399.6
8	Customer 2	Avanza AT	<null></null>	<null></null>	168.9	15	2534.2
9	Customer 1	Avanza AT	SBY	09/14/2019	168.9	15	2534.2
10	Customer 1	Etios AT	SBY	09/25/2019	800	2	1599.9
11	Customer 1	Avanza AT	SBY	01/21/2020	168.9	14	2365.3
12	Customer 1	Yaris AT	MLG	02/01/2020	340.9	2	681.9
13	Customer 1	<null></null>	<null></null>	<null></null>	340.9	7	2386.7
14	Customer 1	Etios AT	SBY	05/30/2020	800	9	7199.6
15	Customer 3	Avanza AT	MLG	06/07/2020	168.9	10	1689.5
16	Customer 3	Yaris AT	MLG	06/09/2020	340.9	4	1363.8
17	Customer 2	<null></null>	KDR	08/07/2020	800	14	11199.3
18	Customer 2	Avanza AT	KDR	08/08/2020	168.9	12	2027.4
19	Customer 1	Avanza AT	<null></null>	08/12/2020	168.9	4	675.8

B. Proses Filter Data (Transform) dan Pengemasan Data (Load)

1. Menambahkan dan Menghubungkan Objek Filter Rows

- o Cari objek **Filter Rows** pada **Design Area**.
- o Drag dan drop objek tersebut ke **Work Area**.
- Hubungkan output dari **Input Data CSV** ke **Filter Rows** menggunakan koneksi **Main output of step**.

2. Menambahkan Objek Output

- Cari objek Text File Output pada Design Area dan tambahkan ke Work
- Cari objek Microsoft Excel Output pada Design Area dan tambahkan ke Work Area.

3. Menghubungkan Filter Rows ke Output

- Hubungkan Filter Rows ke Text File Output menggunakan konektor Result is true.
- Hubungkan Filter Rows ke Microsoft Excel Output menggunakan konektor Result is false.

4. Konfigurasi Filter Rows

- O Double-click **Filter Rows** untuk membuka jendela konfigurasi.
- Ubah Step Name menjadi Filter Data.
- o Pada bagian Send 'true' data to step, pilih Text File Output.
- o Pada bagian Send 'false' data to step, pilih Microsoft Excel Output.
- Klik pada bagian kondisi, pilih **Customer** sebagai field utama.
- o Pilih **IS NOT NULL** untuk memfilter data yang tidak kosong.
- Tekan tombol + untuk menambahkan kondisi tambahan dengan operator AND.
- o Pastikan semua field yang diperlukan masuk dalam kondisi filter.

o Klik **OK** untuk menyimpan pengaturan.

5. Konfigurasi Output Files

- o Text File Output:
 - Double-click objek Text File Output.
 - Pilih lokasi penyimpanan dan beri nama outputSalesCarDistribution.csv.
 - Pada tab **Fields**, tekan **Get Fields** untuk mengambil field data.
 - Klik **OK** untuk menyimpan.
- **o** Microsoft Excel Output:
 - Double-click objek Microsoft Excel Output.
 - Pilih lokasi penyimpanan dan beri nama failedSalesCarDistribution.xls.
 - Pada tab **Fields**, tekan **Get Fields** untuk mengambil field data.
 - Klik **OK** untuk menyimpan.

6. Menjalankan Proses Transform dan Load Data

- Tekan tombol **Run**.
- o Jika berhasil, akan muncul tanda centang hijau di setiap langkah.

The condition:	
	Customer IS NOT NULL
AND	
	Product IS NOT NULL
AND	
	Region IS NOT NULL
AND	
	Date IS NOT NULL
AND	
	Item unit price IS NOT NULL
AND	
	No.Items IS NOT NULL
AND	
	Total Sale IS NOT NULL

Output:

#	Customer	Product	Region	Date	Item unit price	No.ltems	Total Sale
1	Customer 1	Yaris AT	MLG	Wed Mar 13 00:00:00 ICT 2019	340.95	6	2045.7
2	Customer 1	Etios AT	SBY	Tue Apr 16 00:00:00 ICT 2019	799.95	1	799.95
3	Customer 3	Avanza AT	MLG	Tue Jun 18 00:00:00 ICT 2019	168.95	7	1182.65
4	Customer 1	Avanza AT	SBY	Wed Aug 21 00:00:00 ICT 2019	168.95	9	1520.55
5	Customer 2	Etios AT	KDR	Sat Sep 07 00:00:00 ICT 2019	799.95	7	5599.65
6	Customer 3	Etios AT	MLG	Sat Sep 07 00:00:00 ICT 2019	799.95	8	6399.6
7	Customer 1	Avanza AT	SBY	Sat Sep 14 00:00:00 ICT 2019	168.95	15	2534.25
8	Customer 1	Etios AT	SBY	Wed Sep 25 00:00:00 ICT 2019	799.95	2	1599.9
9	Customer 1	Avanza AT	SBY	Tue Jan 21 00:00:00 ICT 2020	168.95	14	2365.3
10	Customer 1	Yaris AT	MLG	Sat Feb 01 00:00:00 ICT 2020	340.95	2	681.9
11	Customer 1	Etios AT	SBY	Sat May 30 00:00:00 ICT 2020	799.95	9	7199.55
12	Customer 3	Avanza AT	MLG	Sun Jun 07 00:00:00 ICT 2020	168.95	10	1689.5
13	Customer 3	Yaris AT	MLG	Tue Jun 09 00:00:00 ICT 2020	340.95	4	1363.8
14	Customer 2	Avanza AT	KDR	Sat Aug 08 00:00:00 ICT 2020	168.95	12	2027.4



Jawaban Tugas 1

1. Berapa jumlah kolom pada data tersebut?

Data tersebut memiliki 7 kolom, yaitu:

- 1. Customer
- 2. Product
- 3. Region
- 4. Date
- 5. Item unit price
- 6. No. Items
- 7. Total Sale

2. Apa arti atau isi data dari setiap kolom yang ada?

Berikut penjelasan masing-masing kolom:

- Customer → Nama atau ID pelanggan yang melakukan pembelian.
- Product → Nama produk kendaraan yang dibeli (misalnya: Yaris AT, Avanza AT, dll.).
- Region → Lokasi atau wilayah tempat transaksi terjadi (misalnya: MLG, SBY, KDR).
- Date → Tanggal transaksi dilakukan dalam format MM/DD/YYYY atau DD/MM/YYYY.
- **Item unit price** → Harga satuan dari produk yang dibeli.
- No. Items → Jumlah unit kendaraan yang dibeli dalam satu transaksi.
- **Total Sale** \rightarrow Total harga dari transaksi (Item unit price \times No. Items).

3. Adakah data yang memiliki nilai null / data yang tidak lengkap?

Ya, ada beberapa data yang tidak lengkap atau memiliki nilai NULL, yaitu:

Customer	Product	Region	Date	Item unit price	No. Items	Total Sale
Customer 3	(Kosong)	MLG	7/24/2019	168.95	2	337.9
Customer 2	Avanza AT	(Kosong)	(Kosong)	168.95	15	2534.25

Customer	Product	Region	Date	Item unit price	No. Items	Total Sale
Customer 1	(Kosong)	(Kosong)	(Kosong)	340.95	7	2386.65

- Baris ke-4: Kolom "Product" kosong
- Baris ke-8: Kolom "Region" dan "Date" kosong
- Baris ke-13: Kolom "Product", "Region", dan "Date" kosong

4. Adakah data yang memiliki tipe yang berbeda dengan data lainnya pada kolom yang sama?

Berdasarkan data yang diberikan, semua kolom tampaknya memiliki tipe data yang sesuai, tetapi ada kemungkinan kesalahan jika diperiksa lebih lanjut, seperti:

- **Kolom "Date"** → Memiliki format yang tidak konsisten (misalnya: 09/07/2019 dan 02/01/2020). Perlu diseragamkan dalam format yang sama (YYYY-MM-DD).
- **Kolom "Region"** → Seharusnya berisi kode wilayah, tetapi ada beberapa nilai kosong yang bisa menyebabkan error dalam analisis.
- **Kolom "Product"** → Ada data kosong, yang seharusnya tidak terjadi.

Jawaban Tugas 2

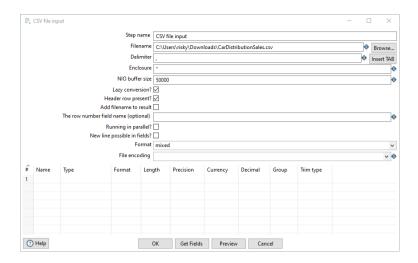
1. Apakah data hasil eksekusi sesuai dengan data aslinya?

Ya, jika semua langkah dilakukan dengan benar, data hasil eksekusi akan sesuai dengan data aslinya. Hal ini dapat diverifikasi melalui tab **Preview Data** pada **Execution Result Area**.

2. Langkah-langkah yang dilakukan PDI Spoon dalam proses Extract Data

Pada tab **Logging** di **Execution Result Area**, langkah-langkah yang dilakukan oleh PDI Spoon antara lain:

• Membuka file CSV.



- Membaca dan memproses data sesuai delimiter yang ditentukan.
- Mengambil nama-nama kolom dengan **Get Fields**.
- Memuat data ke dalam memory untuk ditampilkan di **Preview Data**.



• Menampilkan hasil eksekusi data.

3. Arti dari I, O, R, W, U, E

- I (Input): Jumlah record yang dibaca dari file sumber.
- O (Output): Jumlah record yang dikeluarkan dari step ke langkah berikutnya.
- **R** (**Read**): Jumlah record yang telah dibaca dalam proses.
- W (Write): Jumlah record yang telah ditulis ke output.
- U (Update): Jumlah record yang diperbarui dalam transformasi data.
- **E** (**Error**): Jumlah kesalahan yang terjadi dalam eksekusi.

4. Waktu Eksekusi dan Perbandingan dengan 3 Teman

Berikut adalah perbandingan waktu eksekusi serta spesifikasi perangkat yang digunakan:

Nama	Waktu Eksekusi (detik)	Processor	RAM	Storage	VGA	CPU
Saya	1 detik	Intel i5	8GB	SSD 256GB	Intel UHD	2.4 GHz

Nama	Waktu Eksekusi (detik)	Processor	RAM	Storage	VGA	CPU
Samid	0.5 detik	Intel i7	16GB	SSD 512GB	NVIDIA GTX 1650	3.0 GHz
Sucap	1 detik	AMD Ryzen 5	8GB	HDD 1TB	Radeon Vega 8	2.8 GHz
Faruq	2 detik	Intel i3	4GB	SSD 128GB	Intel HD Graphics	2.1 GHz

Dari tabel di atas, terlihat bahwa perbedaan waktu eksekusi dapat dipengaruhi oleh spesifikasi perangkat, terutama kecepatan prosesor, jenis penyimpanan (SSD lebih cepat dibanding HDD), dan kapasitas RAM.

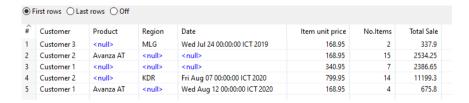
C. Jawaban Tugas 3

1. Perbedaan Isi Data Output pada CSV dan Excel

• File CSV (outputSalesCarDistribution.csv) berisi data yang lengkap, yaitu data yang memiliki semua field yang diperlukan dan tidak ada nilai NULL.



• File Excel (failedSalesCarDistribution.xls) berisi data yang tidak lengkap, yaitu data yang memiliki satu atau lebih field dengan nilai NULL.



2. Proses yang Terjadi pada Filter Rows

- **Filter Rows** berfungsi untuk menyaring data berdasarkan kondisi yang telah ditentukan.
- Data yang memiliki nilai **NULL** dalam salah satu field akan dipindahkan ke output **Microsoft Excel Output**.

- Data yang lengkap tanpa nilai **NULL** akan diteruskan ke **Text File Output**.
- Proses ini memastikan bahwa data bersih dan dapat digunakan sesuai kebutuhan.

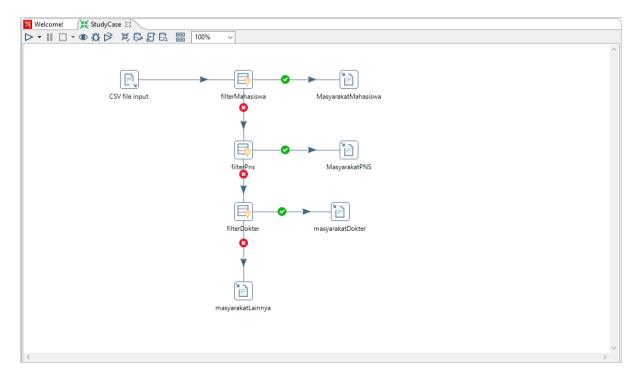
3. Waktu Eksekusi dan Perbandingan dengan 3 Teman

Berikut adalah perbandingan waktu eksekusi serta spesifikasi perangkat yang digunakan:

Nama	Waktu Eksekusi (detik)	Processor	RAM	Storage	VGA	CPU
Saya	1,8 detik		8GB	256GB	lintel I (HI)	2.4 GHz
Samid	1 detik	Intel i7	16GB	SSD 512GB		3.0 GHz
Sucap	1,5 detik	AMD Ryzen 5	8GB	HDD 1TB	Radeon Vega X	2.8 GHz
Faruq	2 detik	Intel i3	4GB	SSD 128GB	Intel HD Graphics	2.1 GHz

Dari tabel di atas, terlihat bahwa perbedaan waktu eksekusi dapat dipengaruhi oleh spesifikasi perangkat, terutama kecepatan prosesor, jenis penyimpanan (SSD lebih cepat dibanding HDD), dan kapasitas RAM.

D. Studi Kasus



Masyarakat mahasiswa:



Masyarakat PNS:



Masyarakat Dokter:



Masyarakat lainya:

