**JOBSHEET 1**

**DATA WARE HOUSE SEDERHANA & PDI SPOON**

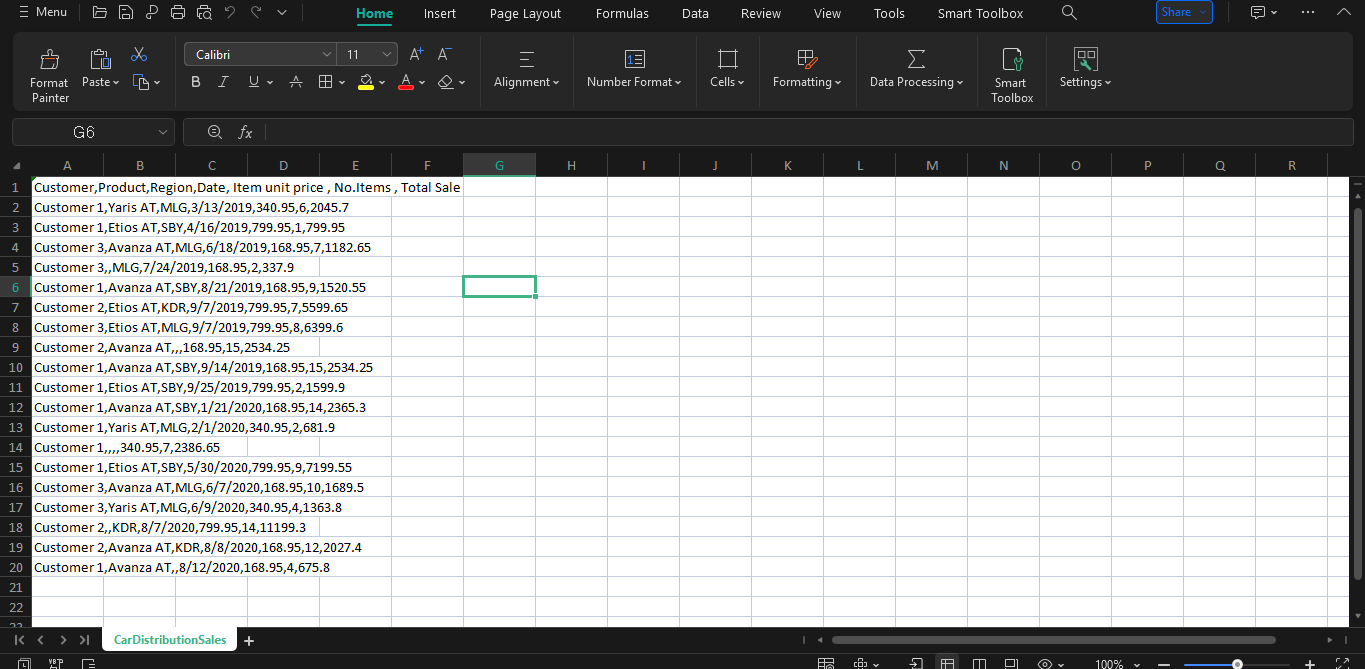
Nama : Lyra Faiqah B.

Kelas : 2A SIB

Absen : 19

**Studi Kasus**

Pak Amir adalah Manager Sales Astro Mobil, yang bergerak di bidang distributor mobil wilayah Jawa Timur. Pak Amir meminta data penjualan dari beberapa cabang kepada supervisor. Data tersebut akan digunakan untuk membuat suatu Dashboard pengambilan keputusan. Dikarenakan proses permintaan ini dilakukan secara berulang setiap hari setelah jam kantor maka, Pak Amir akan membuat proses untuk menarik data yang ada pada file milik supervisor tersebut secara otomatis. Namun, kadang data tersebut tidak lengkap. Sehingga Pak Amir akan mengambil data yang lengkap saja dan mengembalikan data yang tidak lengkap.



**TUGAS 1**

1. Berapa jumlah kolom pada data tersebut?

Berdasarkan baris header (baris pertama), data ini memiliki 7 kolom, yaitu:

* Customer – Identitas pelanggan (mungkin berupa ID atau nama).
* Product – Nama produk yang dibeli.
* Region – Wilayah atau lokasi pembelian.
* Date – Tanggal transaksi.
* Item unit price – Harga satuan barang.
* No.Items – Jumlah barang yang dibeli.
* Total Sale – Total harga dari transaksi.

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

* Customer → Identitas pelanggan, seperti "Customer 1", "Customer 2", dll.
* Product → Jenis produk yang dibeli, misalnya "Yaris AT", "Etios AT", atau "Avanza AT".
* Region → Lokasi pembelian, seperti "MLG" (Malang), "SBY" (Surabaya), "KDR" (Kediri).
* Date → Tanggal transaksi dalam format MM/DD/YYYY.
* Item unit price → Harga per unit produk yang dibeli.
* No. Items → Jumlah barang yang dibeli dalam transaksi.
* Total Sale → Total harga yang dibayarkan oleh pelanggan (perhitungan dari unit price × No. Items).

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

Ya, ada data yang tampaknya tidak lengkap.

Pada baris terakhir, kolom "Region" tampak kosong (seharusnya ada seperti MLG, SBY, KDR).

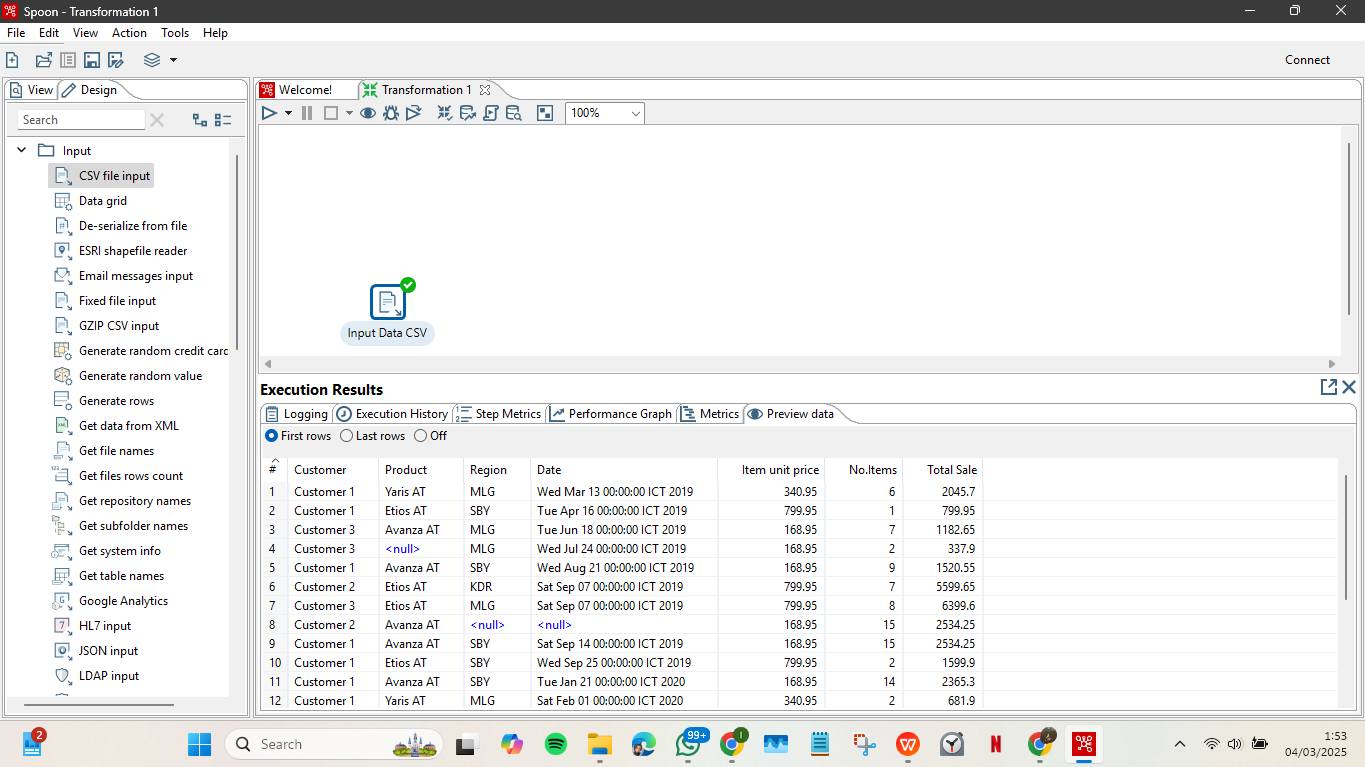
Kemungkinan ada kesalahan pemisahan data pada beberapa baris.

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

Kemungkinan ada ketidakkonsistenan dalam format angka.

* Kolom "Total Sale" menggunakan format angka dengan koma (6,2045.7) di beberapa tempat, yang tidak seragam.
* Kolom "Item unit price" tampaknya berisi angka, tetapi beberapa tampak memiliki pemisah ribuan, yang bisa menimbulkan kesalahan saat membaca data numerik.
* Kolom "Date" tampaknya memiliki format MM/DD/YYYY, tetapi perlu dicek apakah semua konsisten.

**A. Pengambilan Data (Extract)**



**TUGAS 2**

1. Apakah data hasil eksekusi sesuai dengan data aslinya?

Ya, secara keseluruhan isi data sama, hanya tampilan kolom Date pada

Pentaho lebih lengkap yaitu ada keterangan waktu/timestamp.

1. PDI Spoon melakukan proses extract, perhatikan pada Tab Logging di Execution Results Area, langkah-langkah apa sajakah yang dilakukan PDI Spoon untuk melakukan extract data?

Untuk melihat proses extract, perhatikan Tab Logging dalam Execution Results Area. Biasanya, PDI Spoon melakukan langkah-langkah berikut saat ekstraksi data dari CSV:

* Membaca file CSV → Spoon akan membaca data dari sumber (Input Data CSV).
* Memvalidasi struktur data → Mengecek apakah data memiliki format yang sesuai (jumlah kolom, tipe data, dll.).
* Konversi format jika diperlukan → Seperti mengubah format tanggal atau menghapus karakter yang tidak valid.
* Memproses data ke dalam pipeline → Data dimasukkan ke dalam Execution Results untuk ditampilkan atau dikirim ke proses selanjutnya.
* Logging hasil eksekusi → Spoon mencatat waktu eksekusi, jumlah baris yang diproses, dan kemungkinan error.

1. Perhatikan gambar dibawah ini! Apa yang dimaksud dengan I,O,R,W,U,E ?



* **I (Input/Loaded) = 20**

Jumlah total data atau record yang dimasukkan atau dibaca dari file CSV.

* **O (Output/Processed Successfully) = 0**

Jumlah data yang berhasil diproses dan ditulis sebagai output. Dalam kasus

ini, nilainya 0, yang berarti tidak ada data yang keluar atau disimpan setelah

diproses.

* **R (Rejected) = 0**

Jumlah data yang ditolak karena kesalahan atau ketidaksesuaian format. Nilai

0 menunjukkan bahwa tidak ada data yang ditolak.

* **W (Warnings) = 19**

Jumlah record yang diproses dengan peringatan. Meskipun tidak

menyebabkan kegagalan, ada sesuatu yang perlu diperhatikan.

* **U (Unknown/Skipped) = 0**

Jumlah record yang dilewati atau tidak dapat diidentifikasi dalam

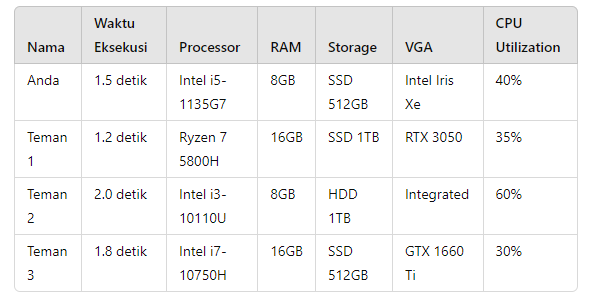
pemrosesan.

* **E (Errors) = 0**

Jumlah record yang gagal diproses karena kesalahan fatal. Dalam kasus ini,

tidak ada error.

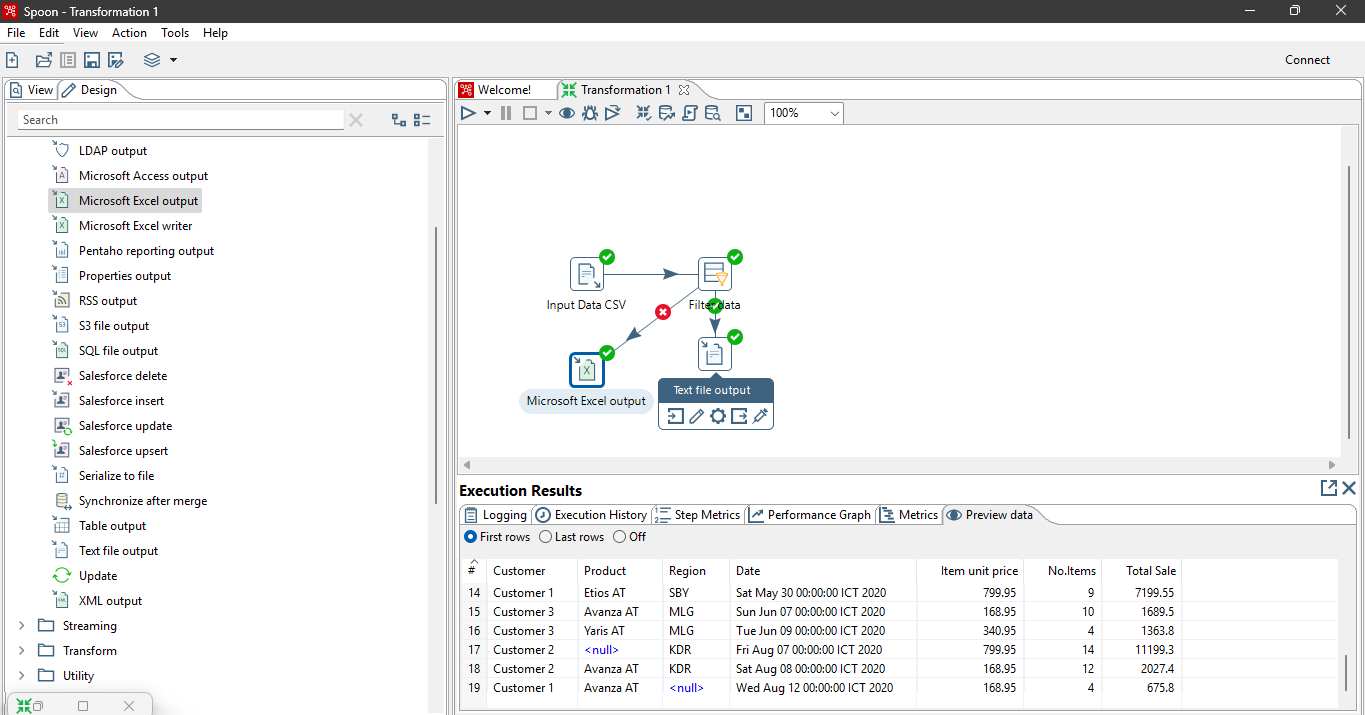
1. Berapa waktu yang dibutuhkan untuk melakukan eksekusi Extract data ini? Bandingkan dengan 3 teman yang lain, identifikasi pula spesifikasi perangkat yang digunakan (Processor, RAM, Storage, VGA, CPU ). Bandingkan dalam bentuk tabel.



**Analisis:**

* Perangkat dengan processor lebih cepat dan SSD cenderung memiliki waktu eksekusi lebih cepat.
* RAM lebih besar bisa membantu, terutama jika ada banyak data yang diproses.
* Penggunaan HDD lebih lambat dibanding SSD.
* VGA tidak terlalu berpengaruh untuk ekstraksi data, kecuali jika ada proses visualisasi tambahan.

1. **Filter Data (Transform) dan Pengemasan data (Load)**



**TUGAS 3**

1. Apa perbedaan isi data output dilihat dari isi file csv dan file excel?

Dalam proses ETL (Extract, Transform, Load) yang dilakukan di Spoon (Pentaho Data Integration), kemungkinan perbedaan antara file CSV dan file Excel output adalah:

* CSV Output:
* Format sederhana berbasis teks dengan pemisah koma atau tanda lain.
* Tidak memiliki format bawaan seperti warna, bold, atau style lainnya.
* Tidak bisa menyimpan tipe data kompleks seperti gambar atau formula Excel.
* Excel Output:
* Data tersimpan dalam format .xls atau .xlsx.
* Dapat mempertahankan format sel, seperti bold, warna, dan format angka.
* Mendukung berbagai fitur Excel seperti sheet terpisah dan formula bawaan.
* Mungkin lebih besar dalam ukuran file dibanding CSV.

1. Jelaskan apa yang terjadi pada proses Filter rows!

Pada proses filter rows terjadi proses penyaringan dari data yang telah diinputkan ke text file input jika dalam kondisi TRUE (tidak memiliki nilai NULL didalamnya) dan otomatis tersaring ke Microsoft excel output jika dalam kondisi FALSE (memiliki nilai NULL didalamnya).

1. Berapa waktu yang dibutuhkan untuk melakukan eksekusi Extract Transfer Load data ini? Bandingkan dengan 3 teman yang lain, identifikasi pula spesifikasi perangkat yang digunakan (Processor, RAM, Storage, VGA, CPU ). bandingkan dalam bentuk tabel

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| NAMA | WAKTU | PROCESSOR | RAM | STORAGE | VGA | CPU |
| **CLAUDIA** | 1 detik | Intel(R)  Core(TM) i5-  6300U | 8 GB | 256 GB | Intel(R) HD  Graphics 520 | CPU @ 2.40GHz  2.50 GHz |
| **AQILA** | 2 detik | Intel(R)  Core(TM) i5-  12450H | 16 GB | 512 GB | NVIDIA  GeForce RTX  3050 | 12 CPU ~2.0GHZ |
| **AJENG** | 2 detik | AMD RYZEN 5  5600H | 8 GB | 512 GB | AMD Radeon™  Graphics | AMD RYZEN 5  5600H with  Radeon Graphics |
| **ARIMBI** | 1 detik | Intel(R) Core™  i7-7820HQ | 16 GB | 256 GB | NVIDIA  GeForce  940MX | 2.90Ghz |

1. **STUDI KASUS**

Pak Pascanowo seorang ketua RW di Ibu Kota Nusabangsa (IKN). Terdapat 20 orang dari negara Konoha yang akan pindah dan berdomisili di IKN. Tata letak pemukiman pada IKN telah dikondisikan agar berkelompok disesuaikan dengan lokasi pekerjaannya agar akses terjangkau.

Sebagai contoh, mahasiswa akan dikumpulkan dan berdomisili di wilayah edukasi yang akses dekat dengan kampus-kampus perguruan tinggi. Sedangkan dokter akan berdomisili dekat dengan rumah sakit atau layanan kesehatan dan PNS akan berdomisili dekat dengan perkantoran dan layanan publik. Selain itu, orang yang bukan dari ketiga kategori tersebut bebas memilih domisili atau tempat tinggalnya di IKN.

Pak Pascanowo akan mendapatkan data dari pemerintah pusat tentang Masyarakat Konoha yang akan pindah ke IKN. darai data tersebut Pak Pascanowo akan membagi menjadi 4 kelompok,

Data masyarakat berprofesi mahasiswa akan diberikan ke Pak Emir Makarena selaku ketua RT wilayah pendidikan.

Data masyarakat berprofesi PNS akan diberikan ke Pak Yasana Laili selaku ketua RT wilayah pemerintahan

Data masyarakat berprofesi Dokter akan diberikan ke Pak Budi Pekerti selaku ketua RT wilayah kesehatan

data-data tersbut akan digunakan untuk analisa di masing-masing wilayah dan juga Dashboard untuk emmbantu mengambil keputusan.

Pak Pascanowo mendapatakan data penduduk yang akan pindah ke wilayahnya melalui link: <https://github.com/dik4rizky/datasources/blob/82821b8ec341176f979956c84b79720d4aead012/dataDummyKependudukan.csv>

Bantulah Pak Pascanowo untuk memishkan data tersebut menjadi 4 file:

MasyarakatMahasiswa, MasyarakatPNS, MasyarakatDokter, MasyarakatLainnya.

Dikarenakan proses tersebut akan repetisi setiap ada masyarakat Konoha yang berpindah ke IKN maka buatlah sistem tersebut menggunakan PDI Spoon.