JOBSHEET 1 DATA WAREHOUSE

Simple and Introduction to PDI Spoon

Dimas Arya Sadewa

SIB 2G

2341760173 – 08

TASK 1

Analyze the data!

1.How many columns are there in the data?

**Number of Columns:**  
The dataset contains **7 columns**, which are separated by commas. The columns are:

* Customer
* Product
* Region
* Date
* Item unit price
* No. Items
* Total Sale

2.What is the meaning or data content of each column?

**Meaning of Each Column:**

* **Customer**: Identifies the customer (e.g., "Customer 1", "Customer 2").
* **Product**: Name of the purchased product (e.g., "Yaris AT", "Etios AT", "Avanza AT").
* **Region**: The location/region where the transaction occurred (e.g., "MLG", "SBY", "KDR").
* **Date**: The date of the transaction (e.g., "3/13/2019", "9/7/2019").
* **Item unit price**: The price per unit of the product (e.g., "340.95", "799.95", "168.95").
* **No. Items**: The number of items purchased in that transaction.
* **Total Sale**: The total revenue from the transaction (computed as Item unit price \* No. Items).

3.Is there any data that has null values / incomplete data?

**Null or Incomplete Data:**  
Yes, there are rows with missing values:

* Row 5: Missing Customer and Product values.
* Row 9: Missing Product, Region, and Date values.
* Other rows also seem to have inconsistencies.

4.Is there any data that has a different type from other data in the same column?

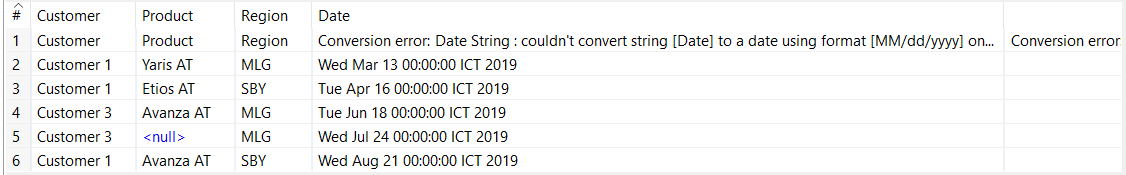
**Different Data Types in the Same Column:**

* **Customer column**: Most values follow "Customer X" format, but some rows are incomplete.
* **Region column**: Some entries are missing, which disrupts the pattern.
* **Date column**: Most values follow the "MM/DD/YYYY" format, but missing values make the data inconsistent.
* **Total Sale column**: Should be numeric, but some rows have missing or malformed values.

A. Data Retrieval (Extract)

TASK 2.

1.Does the execution result data match the original data?

Yes, 

2.PDI Spoon performs the extract process, look0020at the Logging Tab in the Execution Results Area, what steps does PDI Spoon take to extract the data?

In the Logging Tab, we can view information such as:

-Connection status to the data source.

-The number of records extracted.

-Time taken to execute the extract steps.

- Error messages in case of failure.

3.Take a look at the picture below! What is meant by I,O,R,W,U,E?



4.How much time does it take to execute this data extract? Compare with 3 other friends, also identify the device specifications used (Processor, RAM, Storage, VGA, CPU). Compare in the form of a table.

**Meaning of L, O, R, W, U, and E:**

These are **execution metrics** used in PDI Spoon to track the status of data records as they pass through a transformation step:

* **L (Lines input)** → Number of lines (rows) read from the input source.
* **O (Lines output)** → Number of lines successfully outputted from the step.
* **R (Lines read)** → Number of lines read from a previous step.
* **W (Lines written)** → Number of lines written to the next step.
* **U (Lines updated)** → Number of lines updated in the destination (typically a database).
* **E (Lines in error)** → Number of lines that encountered an error during processing.
*  **L (I=20)** → 20 rows were read from the CSV input file.
*  **O (O=0)** → No rows were outputted (possibly because the step does not directly output data).
*  **R (R=0)** → No lines were read from another step.
*  **W (W=19)** → 19 rows were successfully written to the next step.
*  **U (U=0)** → No rows were updated in the destination.
*  **E (E=0)** → No errors occurred during processing.

B. Data Filter (Transform) and Data Packaging (Load)

1. **Perbedaan Isi Data Output antara CSV dan Excel:**

* **CSV (Comma-Separated Values):**
  + Format teks sederhana dengan data dipisahkan oleh koma atau delimiter lain (misalnya titik koma).
  + Tidak mendukung pemformatan sel, warna, grafik, atau rumus.
  + Ukuran file lebih kecil karena hanya berisi teks.
* **Excel (.xlsx/.xls):**
  + Format biner atau XML yang mendukung pemformatan, formula, diagram, dan berbagai fitur lainnya.
  + Memungkinkan beberapa sheet dalam satu file.
  + Ukuran file lebih besar dibanding CSV karena menyimpan lebih banyak informasi.

2. **Proses Filter Rows:**

* Proses ini digunakan untuk menyaring baris dalam dataset berdasarkan kondisi tertentu.
* Misalnya, jika kita hanya ingin menampilkan data mahasiswa dengan IPK di atas 3.0, maka hanya baris dengan nilai IPK > 3.0 yang akan dipertahankan.
* Proses ini mengurangi jumlah data yang akan diproses lebih lanjut, meningkatkan efisiensi analisis.

**3. Waktu Eksekusi ETL dan Perbandingan Spesifikasi Perangkat:**

* Untuk mengetahui waktu eksekusi Extract, Transform, Load (ETL), lakukan pengujian pada empat perangkat berbeda.
* Bandingkan waktu eksekusi serta spesifikasi perangkat yang digunakan dalam bentuk tabel berikut:

| **Nama** | **Waktu Eksekusi (detik)** | **Processor** | **RAM** | **Storage** | **VGA** | **CPU** |
| --- | --- | --- | --- | --- | --- | --- |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Anda | xx | Intel i5-xxxx | 8GB | SSD 512GB | Intel UHD | 4 Cores |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Teman 1 | xx | AMD Ryzen 5 xxxx | 16GB | SSD 1TB | Nvidia GTX | 6 Cores |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Teman 2 | xx | Intel i7-xxxx | 8GB | HDD 1TB | Intel UHD | 8 Cores |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Teman 3 | xx | AMD Ryzen 7 xxxx | 32GB | SSD 2TB | RTX 3060 | 8 Cores |

Studi Kasus

