

2241720263

AL AZHAR RIZQI RIFA’I FIRDAUS

Daftar Isi

CHAPTER 1: Simple Data Warehouse [3](#_Toc98155632)

[Use this style for TOC Level 2 / Sub Chapter (Sub Bab) 3](#_Toc98155633)

[Use this style for TOC Level 3 / Sub Sub Chapter (Sub Sub Bab) 3](#_Toc98155634)

# Chapter 1: Simple Data Warehouse

Practicum Objective

After doing this practicum, students are expected to be more familiar with what a data warehouse is based on trials to create a simple data warehouse.

Study Case

Mr. Amir is the Sales Manager of Astro Mobil, which is engaged in the car distributor in East Java. Mr. Amir requests sales data from several branches to the supervisor. The data will be used to create a decision-making dashboard. Since this request process is repeated every day after office hours, Mr. Amir will create a process to automatically pull the data in the supervisor's file. However, sometimes the data is incomplete. So that Mr. Amir will only retrieve complete data and return incomplete data.

From study case above, then it will be:

1. Check and analyze sales data.

2. Import data from file(Extract)

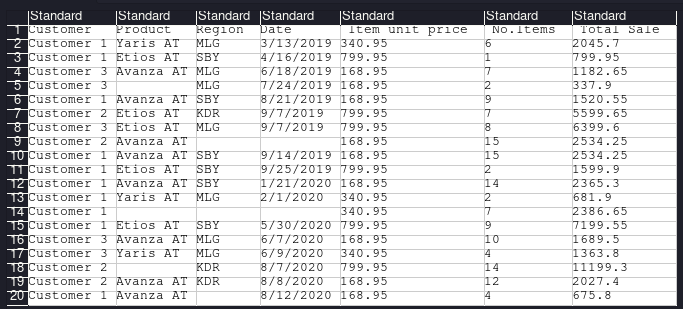
3. Identification the data that not correctly(missing data) and put it at file that different(Transform).

4. Move the data that already correct to dashboard file(Load).

5. Collect the data that incomplete to be returned.

\*The sales data can be accessed and downloaded through the following link:

<https://github.com/dik4rizky/datasources/blob/7e3f70f516a1b107635141d09862a8c30fbb6812/CarDistributionSales.csv>



TASK 1

Analyze the data!

1. How many columns are there in the data?

- **There are 7 columns that is customer, product, region, date, item unit price, items number, and total sale.**

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

- **Customer column contains astro mobil customer data, product column contains product data which is car, region column contains region data which is city, date column contains year of car, item unit price column contains price of car, items number column contains how much car that sold, and total sale colulmn contains car total price that sold, with item unit price multiply by items number.**

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

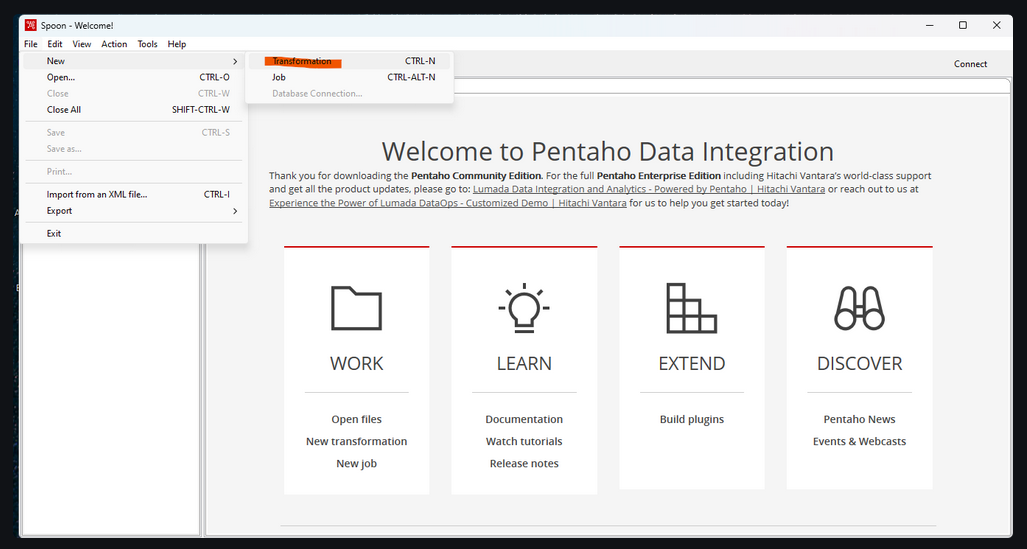
- **yes there are 8, in column product row 5, region column row 9, date column row 9, product column row 14, region column row 14, date column row 14, product column row 18, and region column row 20.**

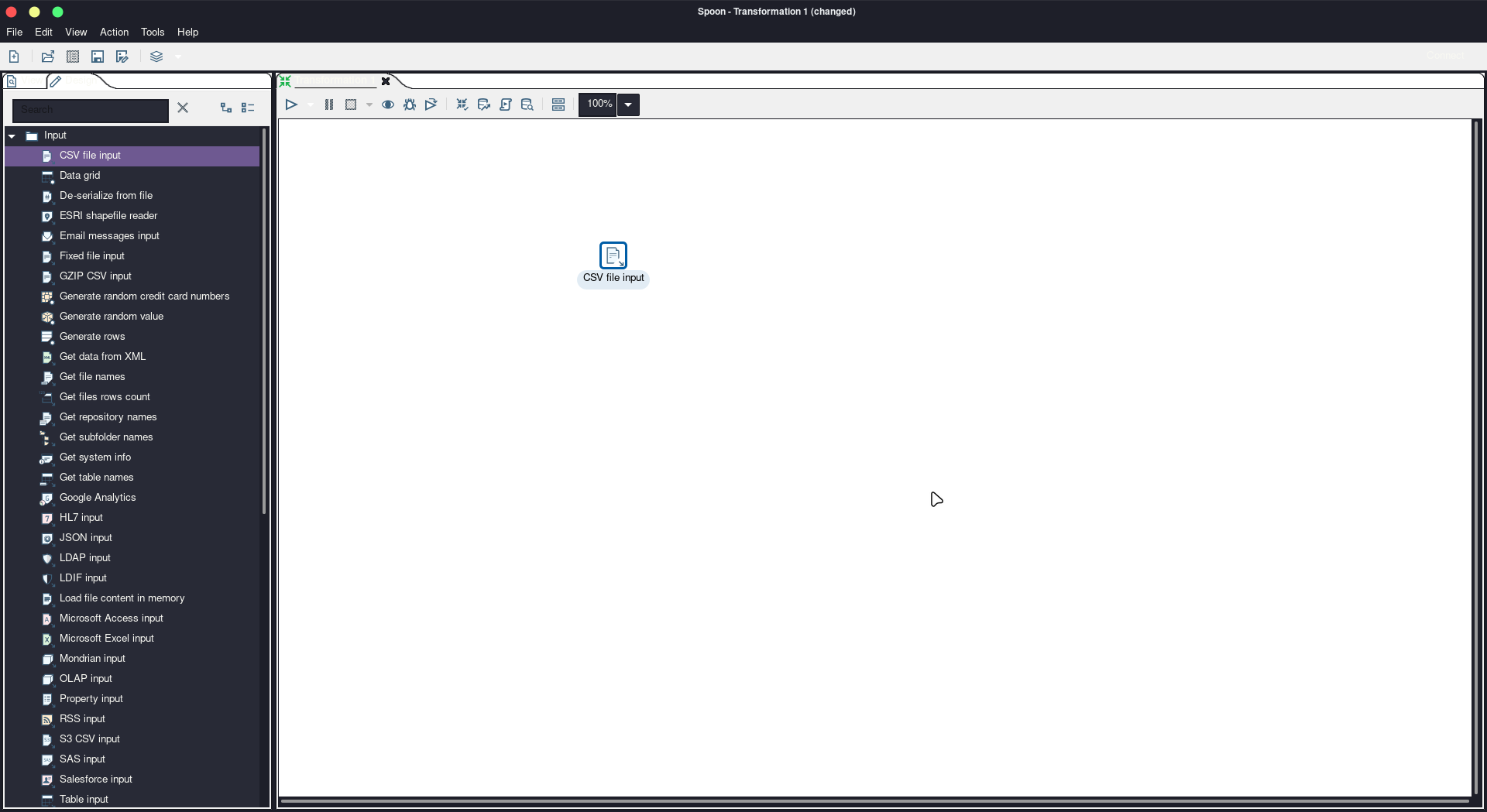
4. Is there any data that has a different type from other data in the same column?

- **No, each column has only one data type.**

## A. Data Retrieval (Extract)

1. Open the Transformation worksheet via the File toolbar.

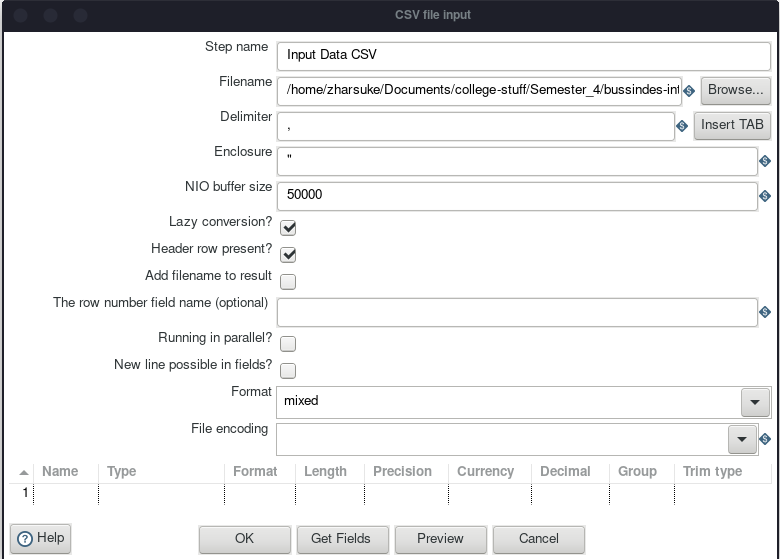


2. Locate the CSV input file object in the Design Area. Drag and drop the object to the Work Area.

3. Double-click on the CSV input file object until the configuration window appears.

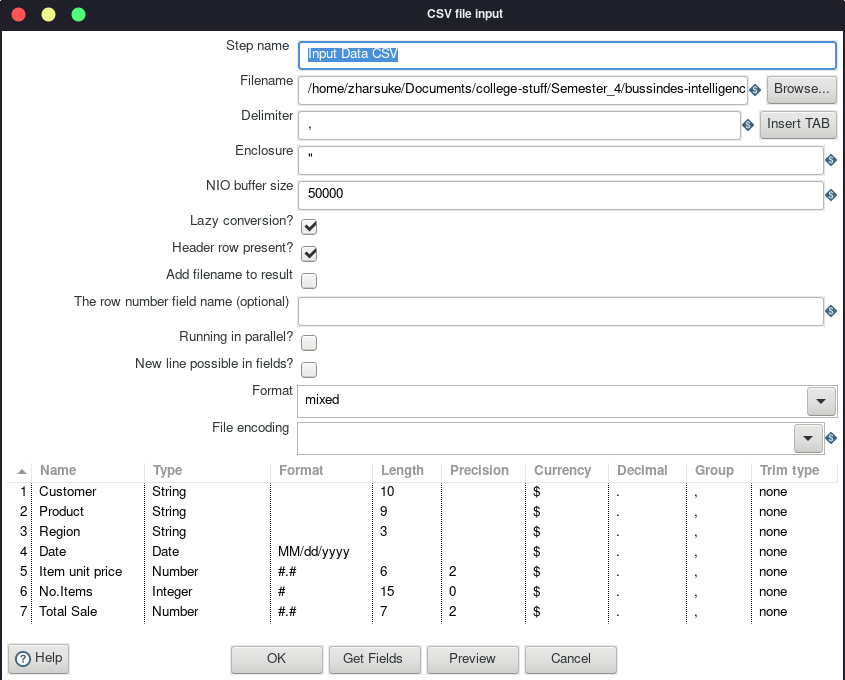
4. Change the Step name to CSV Input Data, (the name is used for identification only, further naming can be adjusted).

5. Select the location of the CarDistributionSales.csv file through Browse on filename.



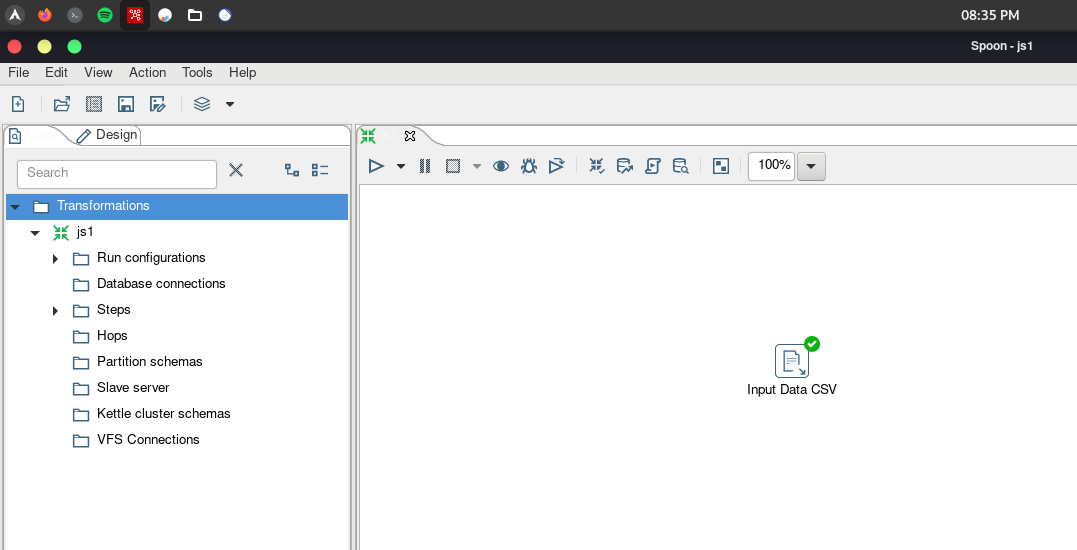
6. If the data has been selected, make sure that the Delimiter section matches the data separator conditions in the selected file. (In this case it is comma " , " ).

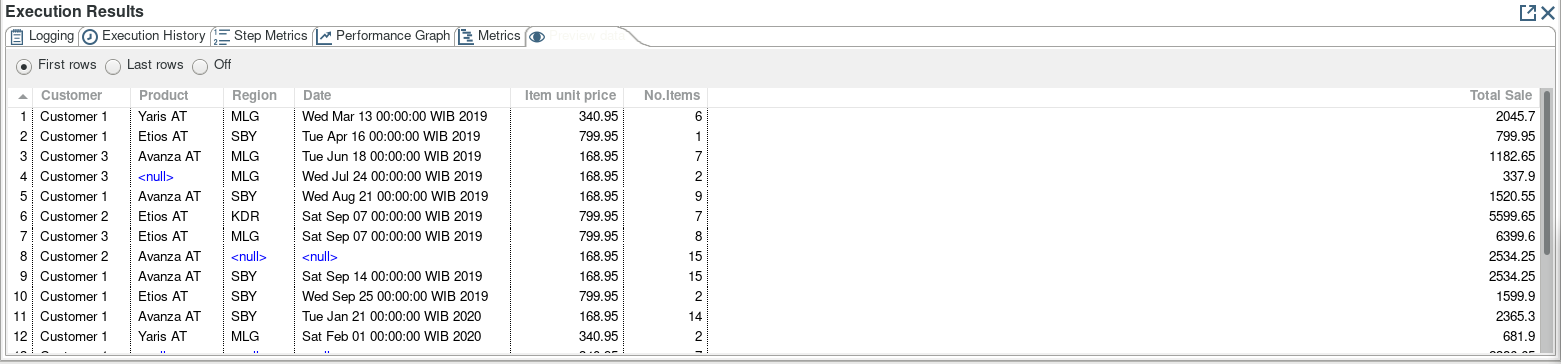
7. Press Get Fields to retrieve the title of each column, make sure the column names are correct.



8. Close the configuration window.

9. Press the "Run" button on the top right corner of the Work Area, the Execution Result Area will appear. Select the Prieview Data tab in the Execution Result Area. If the process is done correctly, the data will appear according to the original.





If the steps above have been successful, then the extract process has been successfully carried out.

TASK 2.

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

- **Yes, after we run it, the data that extract as same with the csv file.**

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

- **From logging tab, it opened transformation first. Then launching the transformation. Then started the transformation execution. Then started the dispatching for transformation. Then pass the row header inside csv file. Process data done.**

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

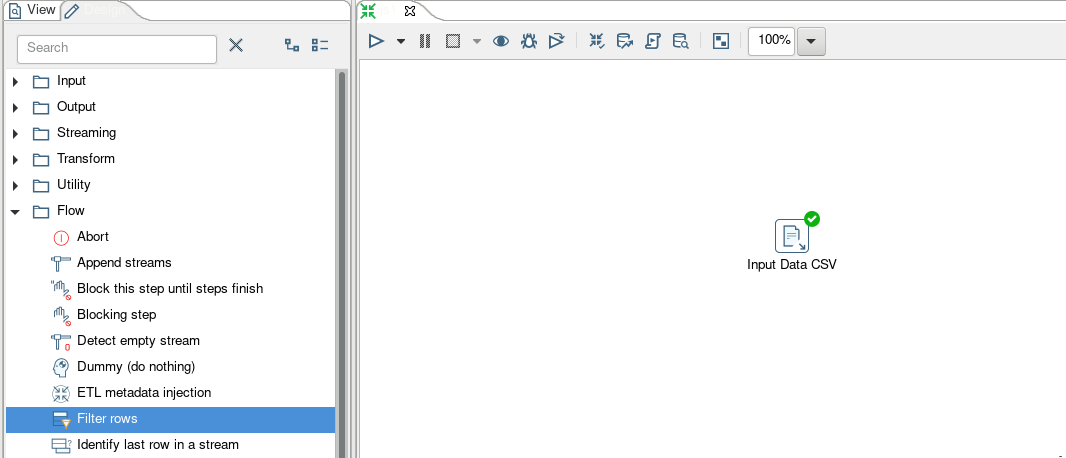
- **I = input, O = output, R = reject, W = writes, U = update, E = error.**

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.

- **teman teman belom garap**

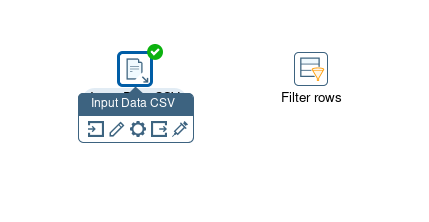
## B. Data Filter (Transform) and Data Packaging (Load)

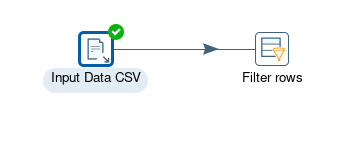
1. Look for the Filter rows object in the Design Area. Drag and drop the object to the Work Area.



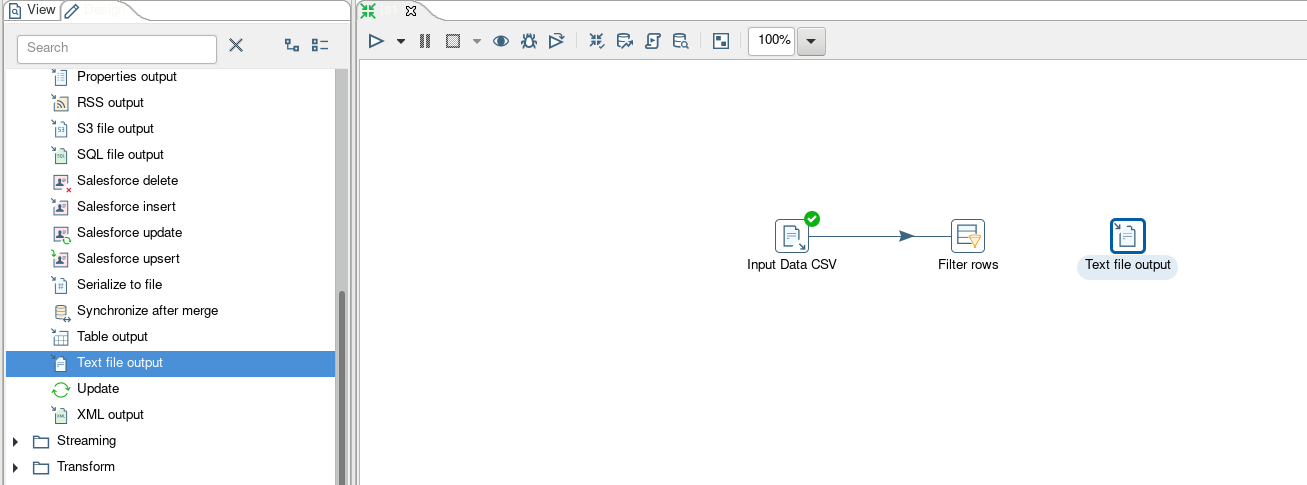
2. Connect the output of the CSV Data Input to the Filter rows. Select the Main output of step connection when connecting the two objects.

At this stage, CSV Data Input is connected by a connector to Filter rows, meaning that after the CSV Data Input process is done, the next process is Filter rows.

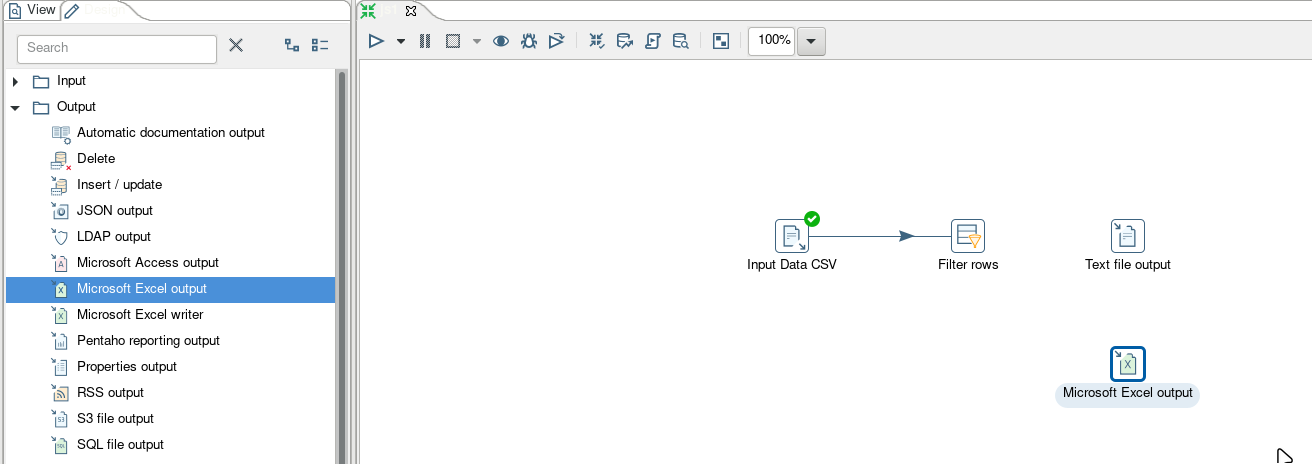




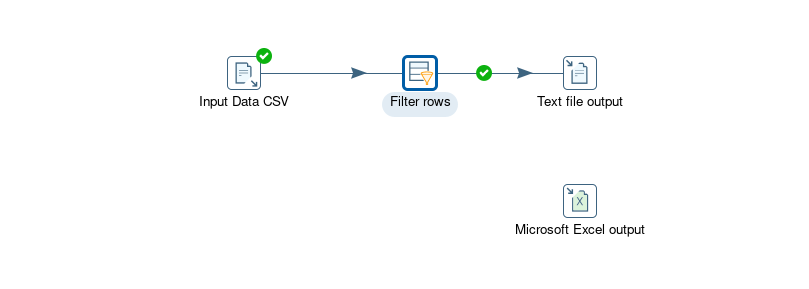
3. Look for the Text output file object in the Design Area. Drag and drop the object on the Work Area.



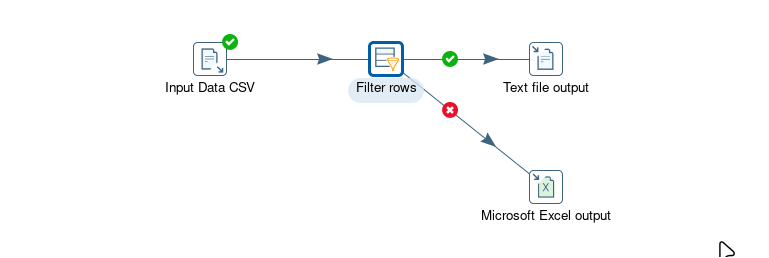
4. Look for the Microsoft excel output object in the Design Area. Drag and drop the object to the Work Area.



5. Connect the filter rows with the text file output using the output connector of the filter rows. Choose Result is true as the connector type.



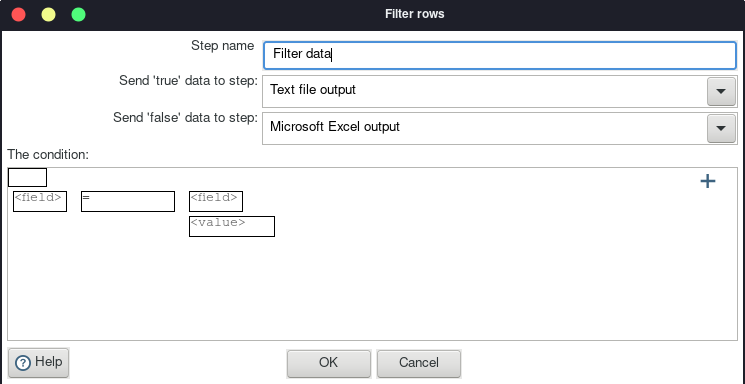
6. Connect filter rows with microsoft excel output using the output connector of filter rows. Select Result is false as the connector type.



7. Double click on the filter rows until the configuration window appears. change the step name to Filter data.

8. Check send 'true' data to step to Text file output and check Send 'false' data to step to Microsoft Excel output.

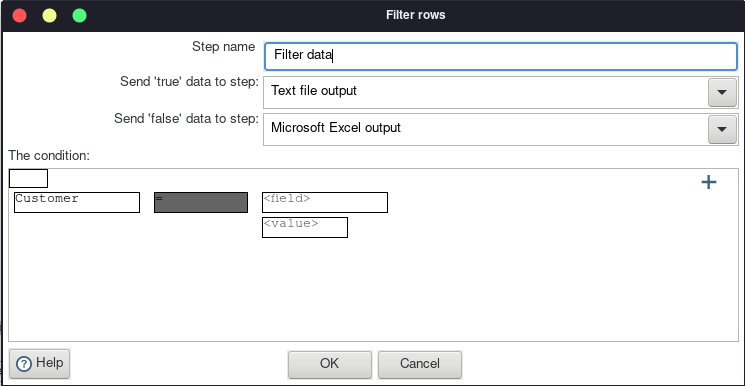
9. Click <Field> in the condition section.

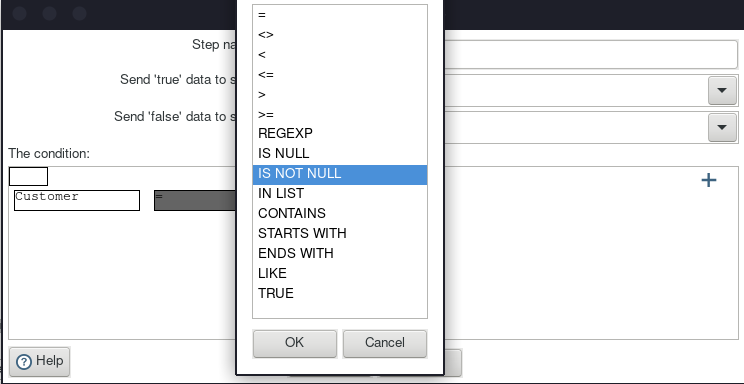


10. Select the desired field to be used as a condition. On this occasion select Customer, then click OK.



11. In the value section, select IS NOT NULL then click OK, this is done because the data selected is data that is not Null.



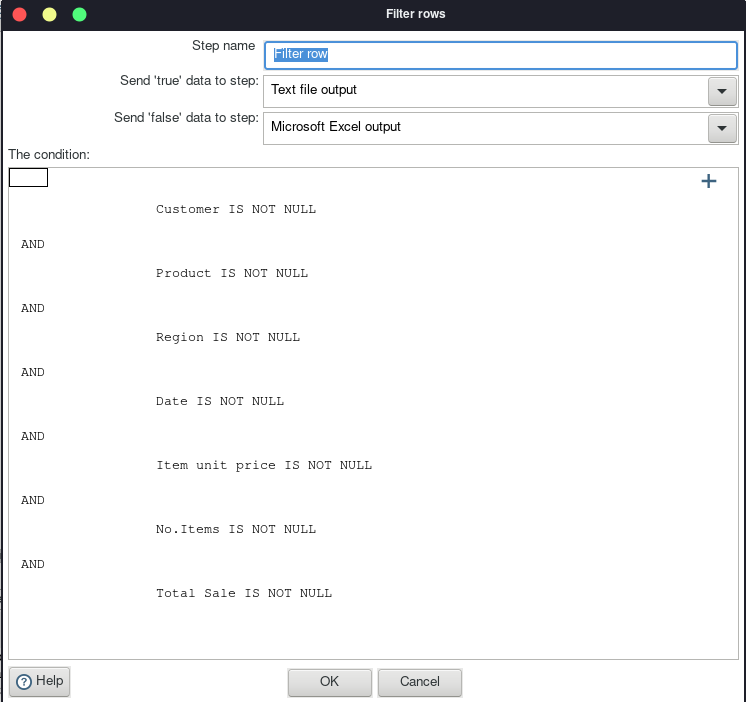


12. Press the + button to add a condition.

13. Select a statement that is still null to add the condition.

14. Repeat the process number 9,10,11 by replacing all the existing fields. until the statement becomes as shown below.

15. Use the AND operator to connect.



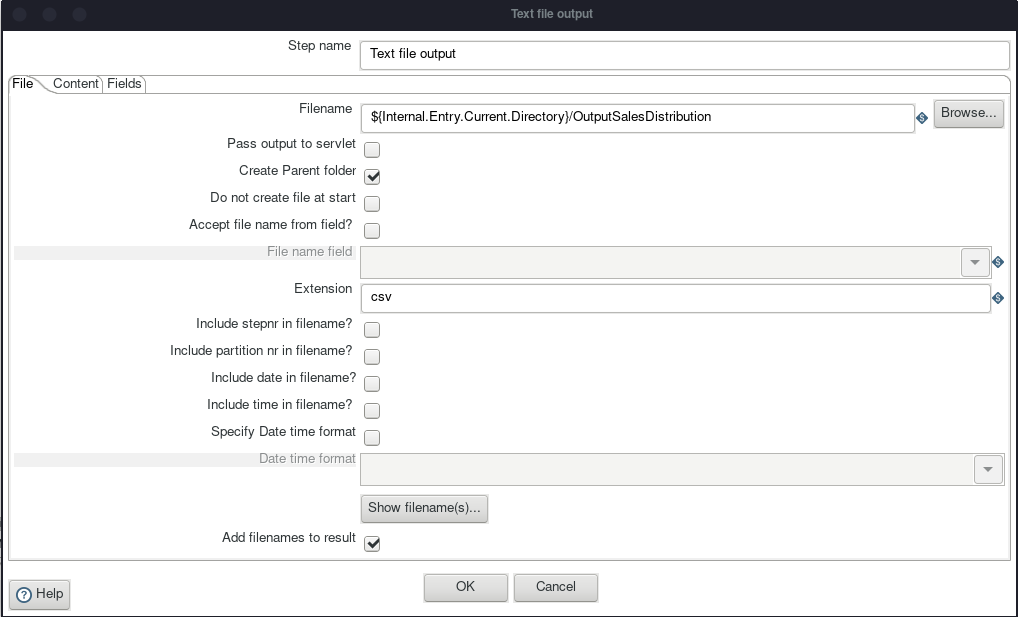
16. If all fields are included in the condition then press OK.

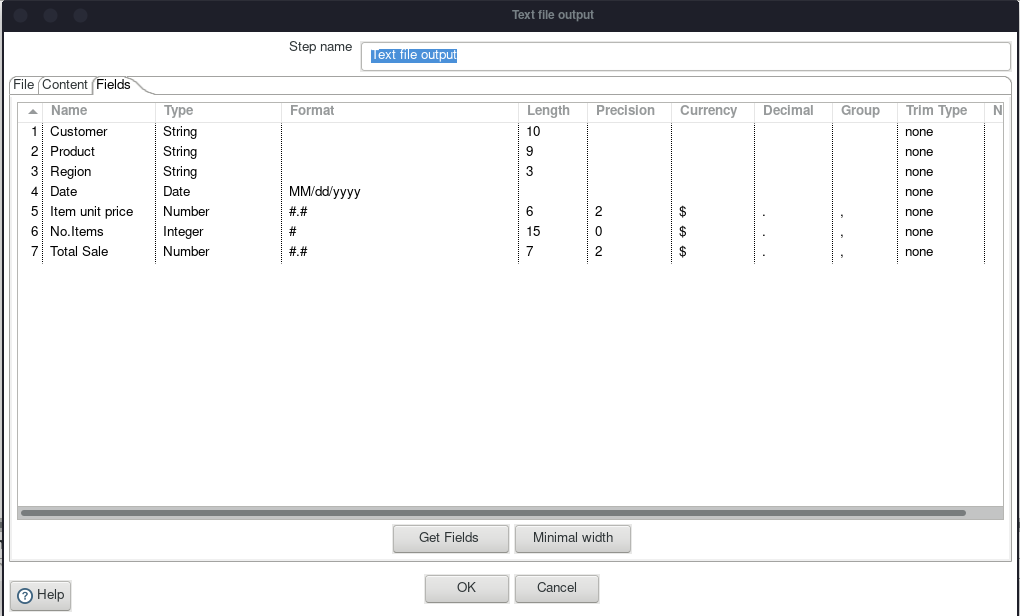
17. Double-click on the output text file, select the file location to save the output file by pressing the browse button. name outputSalesCarDistribution.

18. Change the extension to csv.

19. On the Fields tab, press the Get fields button to retrieve the data fields.

20. Press ok to return to the work area.



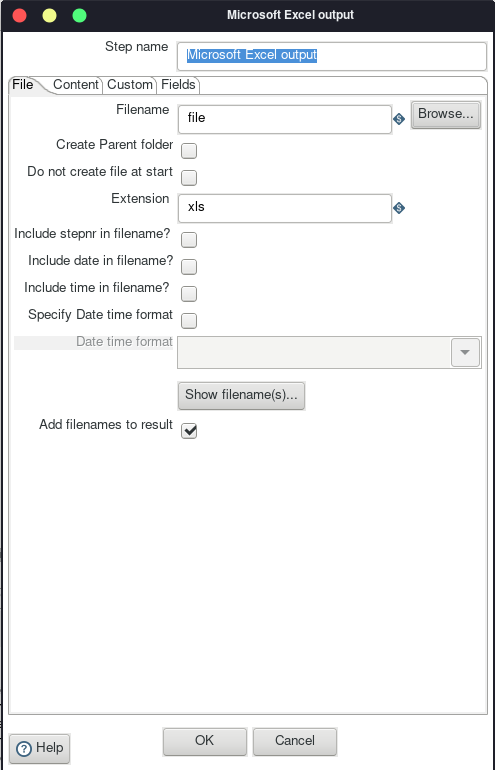


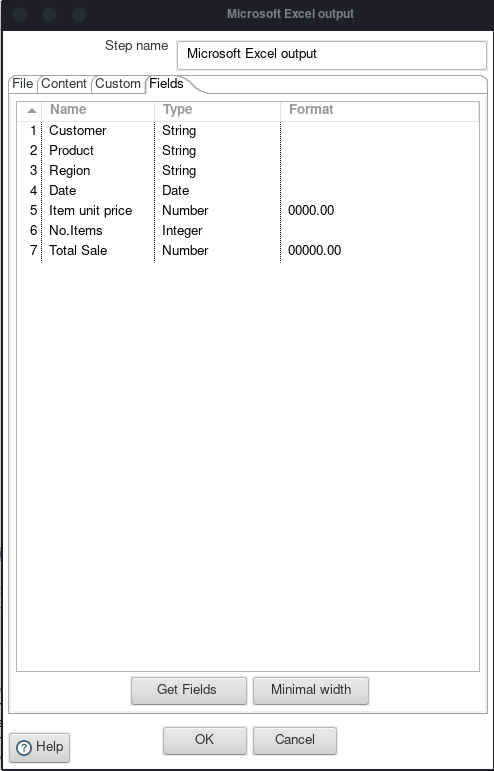
21. Double-click on the Microsoft Excel output object until the configuration window appears.

22. Select the file location to save the output file by pressing the browse button. Give the name failedSalesCarDistribution. for the extension stay with the xls extension.

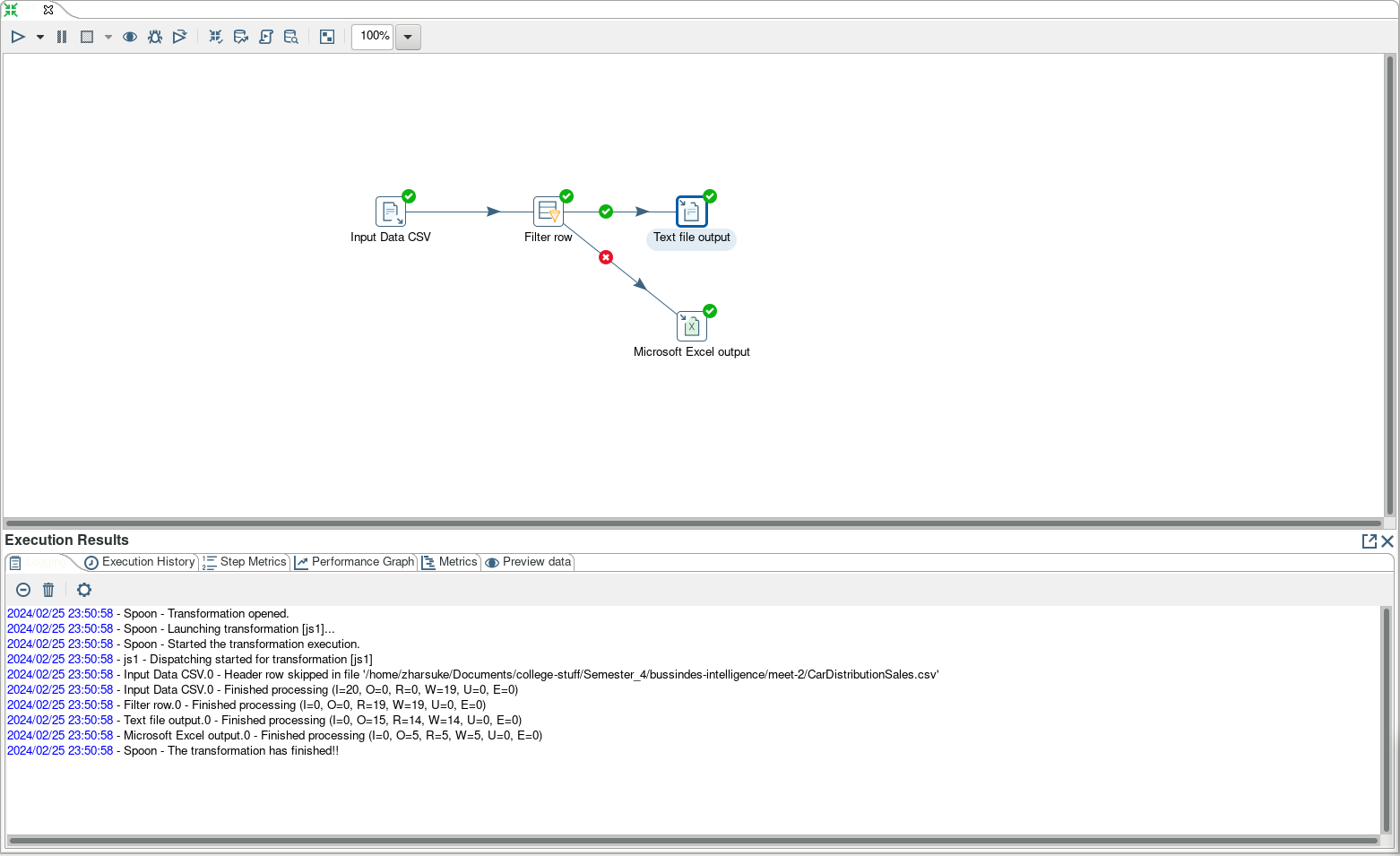
23. On the fields tab press the Get fields button to retrieve the data fields.

24. Click ok to return to the work area.





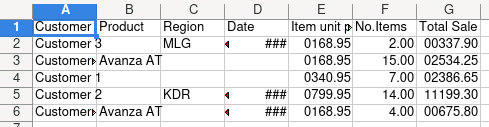
25. Press the Run button in the upper left corner of the Work Area. if successful, a green check mark will appear in each process.



OutputSalesCarDistribution.csv:



failedSalesCarDistribution.xls:



The steps above are the process of filtering data, where complete data will be placed in the csv text file and incomplete data will be placed in ms excel.

TASK 3

1. What is the difference between the contents of the output data seen from the contents of the csv file and the excel file?

- **The difference between both is csv file is to save data that there is no null data, otherwise xls file is to save data that there is null data.**

2. Explain what happens in the Filter rows process!

- **At filter rows configurations, there is conditions that each columns is not null, and each columns connect with AND operator.**

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

- temannya belom garap

## Case Study

Disclaimer: The data listed in this case study is fictitious and used for learning purposes only. The names, dates of birth, addresses, and other information listed are fictitious and have no relation to the actual individuals.

CASE:

Mr. Pascanowo is a neighborhood leader in the capital city of Nusabangsa (IKN). There are 20 people from the country of Konoha who will move and settle in IKN. The residential layout in IKN has been conditioned to be clustered according to the location of their jobs for accessibility.

For example, students will be gathered and domiciled in educational areas with close access to college campuses. Meanwhile, doctors will live close to hospitals or health services and civil servants will live close to offices and public services. In addition, people who are not from these three categories are free to choose their domicile or residence in IKN.

Mr. Pascanowo will get data from the central government about Konoha people who will move to IKN. from the data Mr. Pascanowo will divide it into 4 groups.

Data on people who work as students will be given to Mr. Emir Makarena as the head of the education area RT.

Data on people who work as civil servants will be given to Mr. Yasana Laili as the head of the government area.

Data on people who work as doctors will be given to Mr. Budi Pekerti as the head of the RT in the health area.

The data will be used for analysis in each region and also dashboards to help make decisions.

Mr. Pascanowo gets data on residents who will move to his area through the link: https://github.com/dik4rizky/datasources/blob/82821b8ec341176f979956c84b79720d4aead012/dataDummyKependudukan.csv

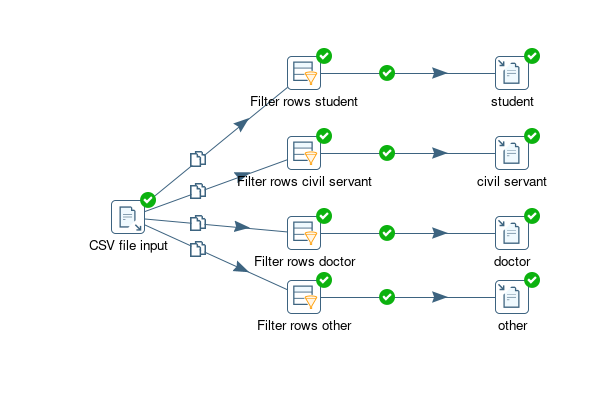
Help Mr. Pascanowo to sort the data into 4 files:

SocietyStudents, SocietyCivil Servants, SocietyDoctors, SocietyOther.

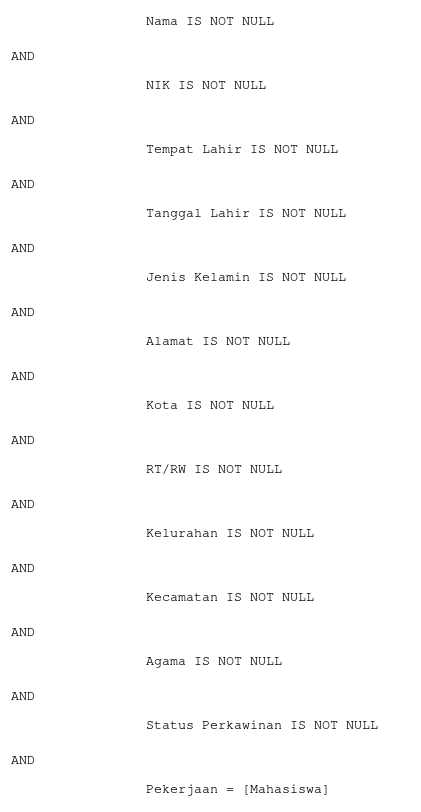
Since the process will be repeated every time a Konoha citizen moves to IKN, make the system using PDI Spoon.

Happy exploring.

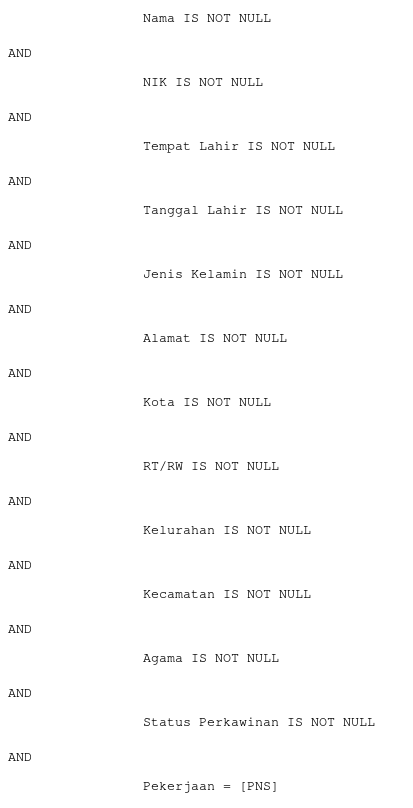
Work Area:



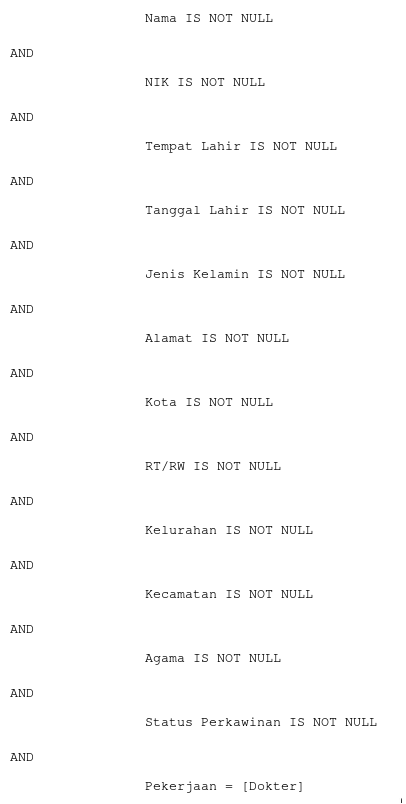
Filter rows student configuration:



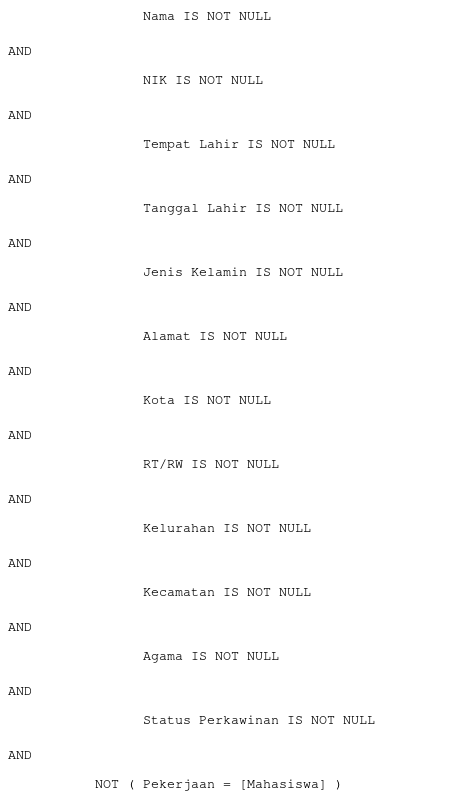
Filter rows civil servant configuration:

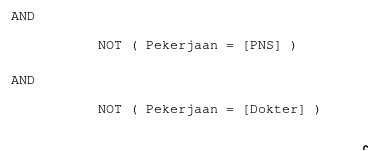


Filter rows doctor configuration:

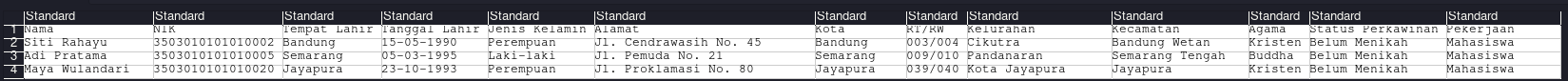


Filter rows other configuration:

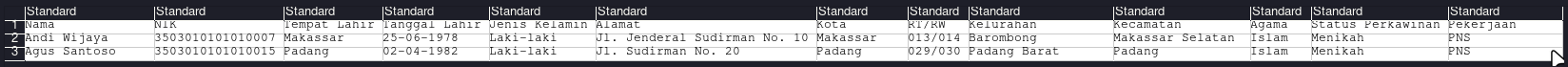


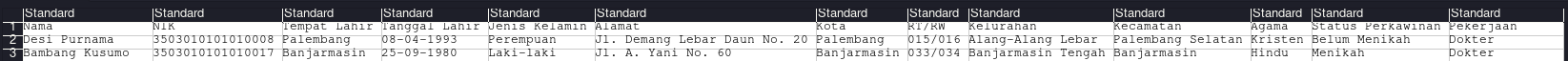


MasyarakatMahasiswa output:



MasyarakatPNS output:



MasyarakatDokter output:

MasyarakatLainnya output:



## Chapter 2: Source Data

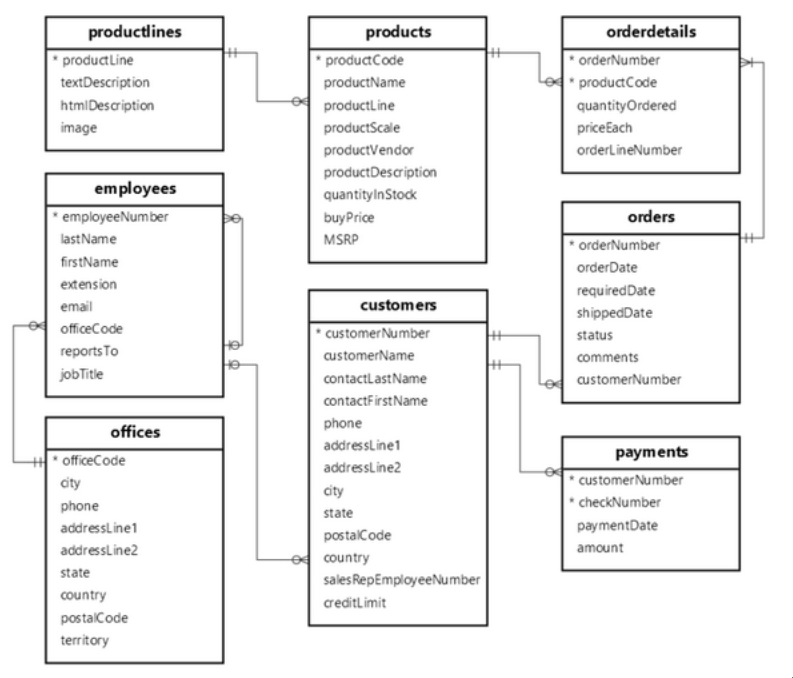
**Practicum Objective**

After doing this practicum, students are expected to be more familiar with source data, how to analyze and see both functional and non-functional needs in data warehouse development.

As well as better understand what OLTP is.

**Case Study**

LegendVehicle is a classic vehicle trade-in company. The company has branches in various countries. LegendVehicle has its own ERP information system. One of the modules of the ERP system is the sales module. The database design of the module is as follows:



In addition, the sales process of vehicles in the company is not only through branch showrooms, but other free resellers.

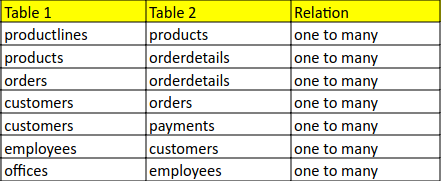
Sales data from the branch can be downloaded via the following link:

https://www.mysqltutorial.org/wp-content/uploads/2023/10/mysqlsampledatabase.zip

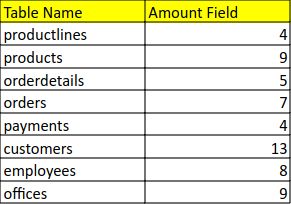
**Task 1**

1. Import the company's data into MySQL DBMS!

2. Analyze the data structure of the company's database, in the form of tables, analyze the relationship between each table!



3. Analyze the number of fields in each table!

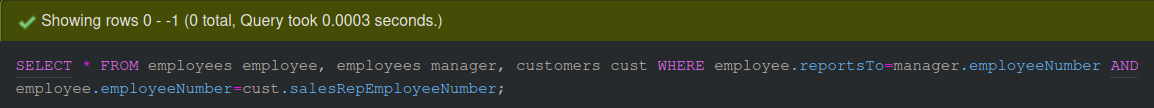


## A. Data Analysis

WARNING: if you find "ERROR" then, dare to find where the error lies to provide a solution. Don't just blame but can't provide a solution.

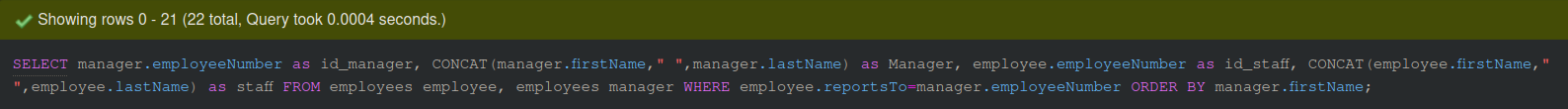
PRACTICUM 1

1. Run the following query on MySql DBMS that has available LegendVehicle Company data.



Then the result of the query is the Employee data along with the Manager and Customer he has. look at the data results carefully.

2. Open a new tab in the browser to execute the following query:



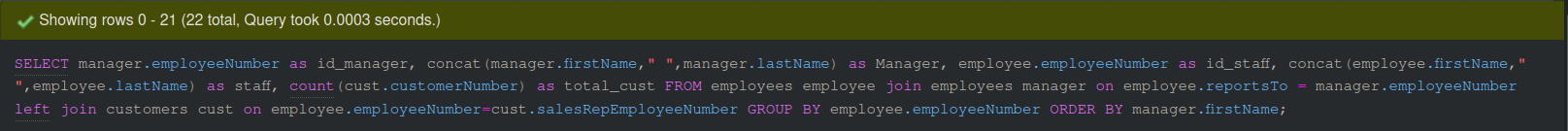
From the results of the query above, the supervisor of each employee will be found.

TASK 2

1. Draw an organizational hierarchy based on the superior of each employee according to the results of the practicum above!



3. Open a new tab in the browser to execute the following query:

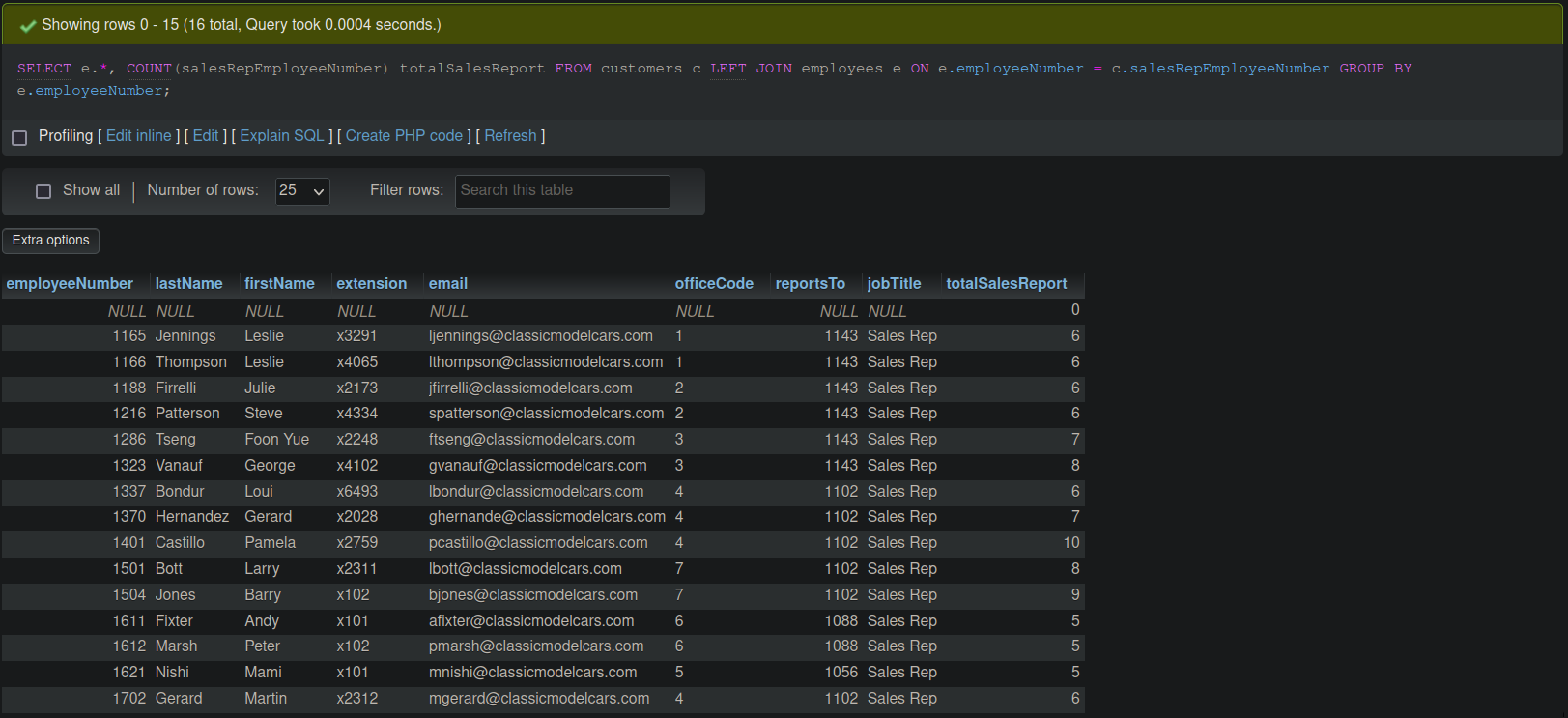


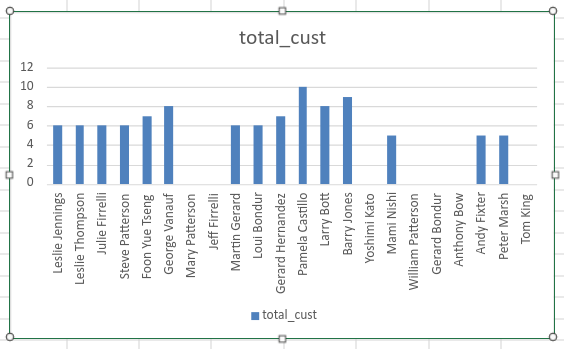
From the query produces the number of customers from each staff.

If the company has a KPI (Key Performances Indicator) "Number of customers who transact" then answer the following questions!

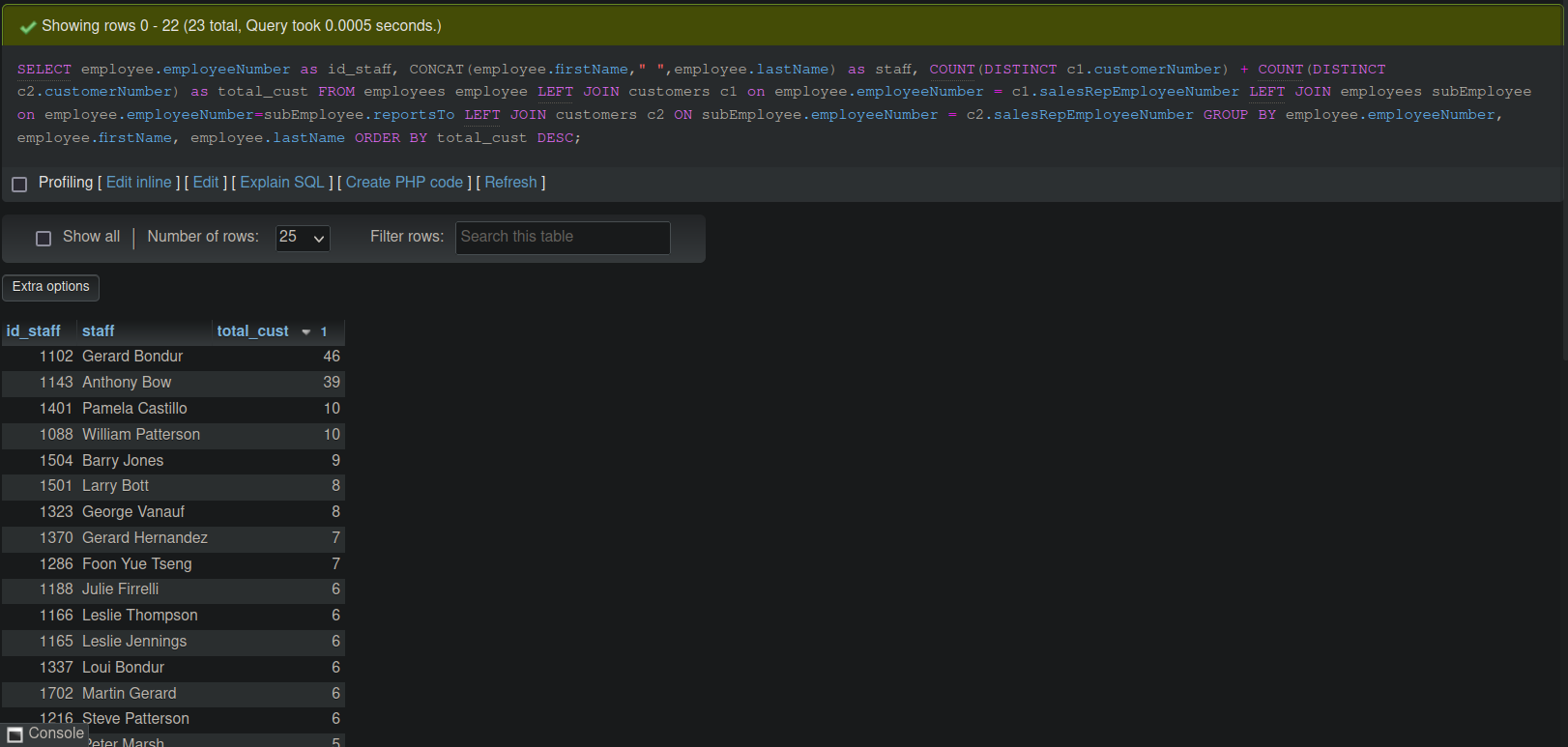
TASK 3

1. Who is the staff at the bottom of the hierarchy who has the most customers?

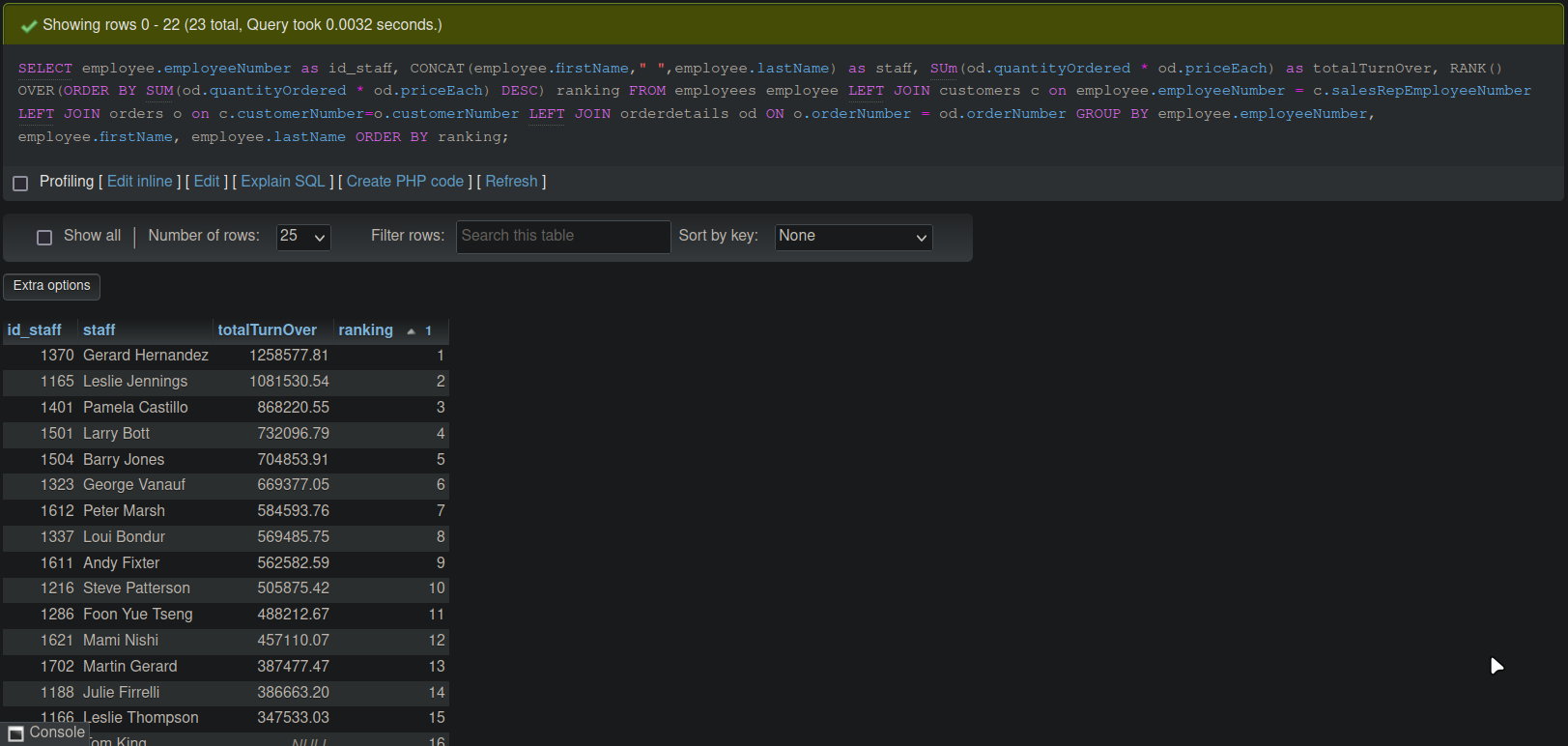


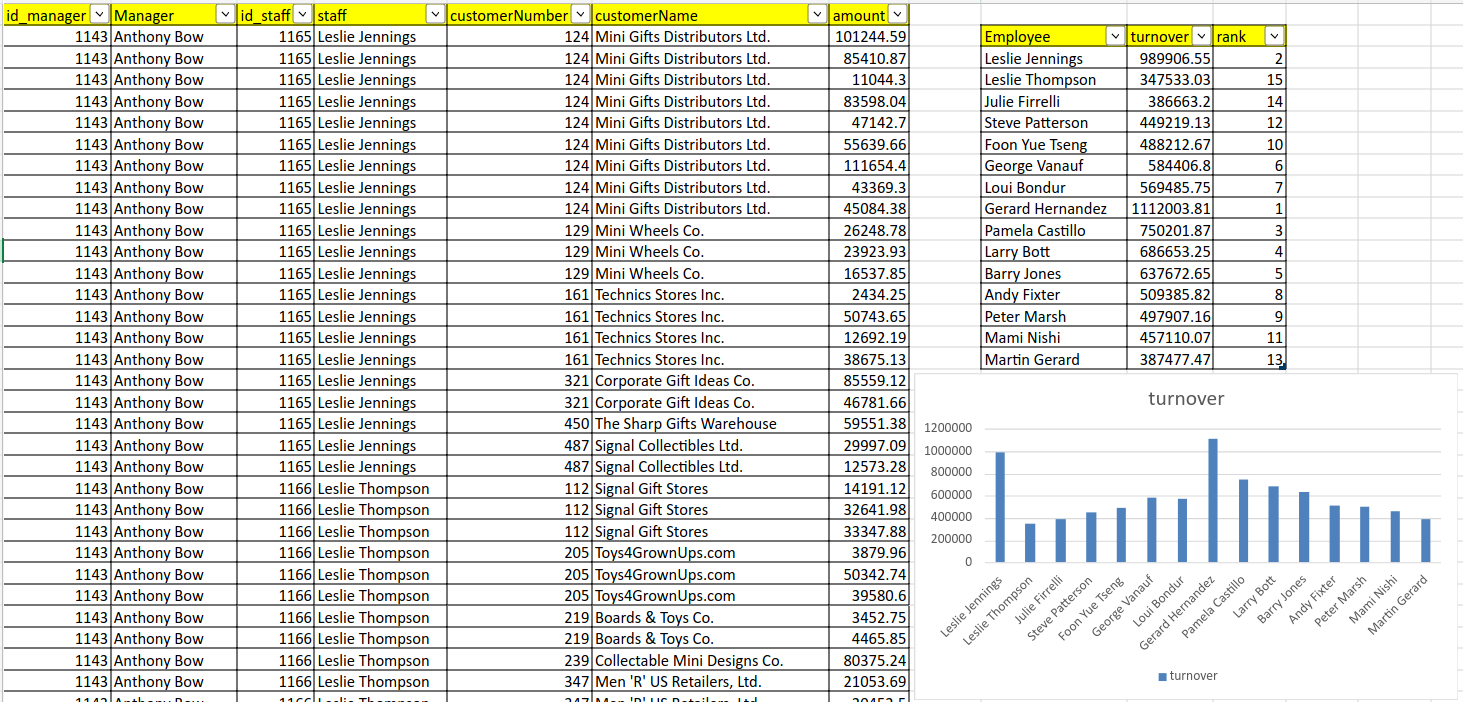


2. If the supervisor's KPI is calculated from the customers he has, then add up the customers of the staff below him, rank the overall performance of the staff along with the description of the number of customers he has!



3. Re-analyze LegendVehicle data to get employee ranking based on KPI "Total turnover obtained". Sort the ranking of employees along with the description of the funds obtained!



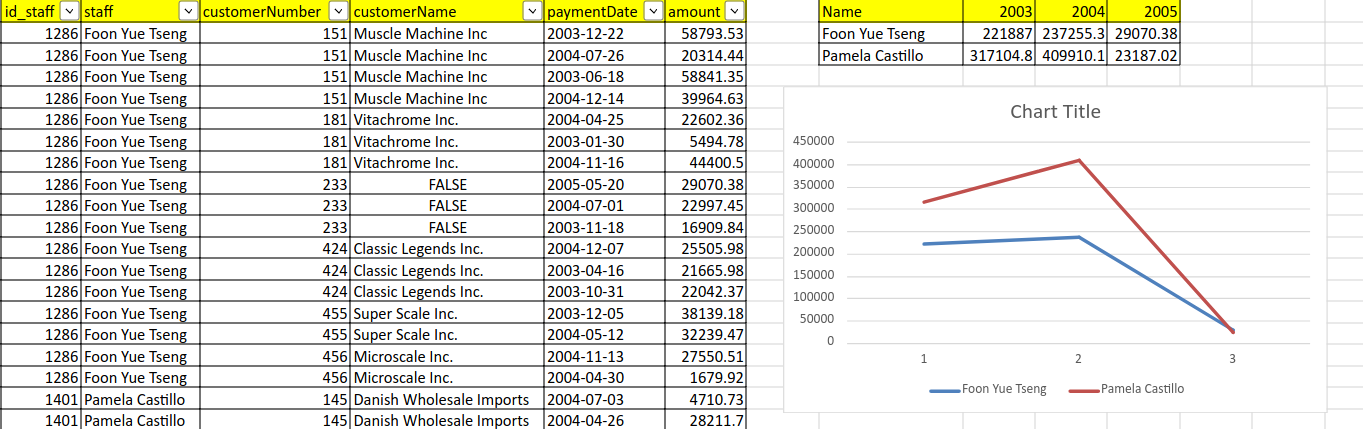
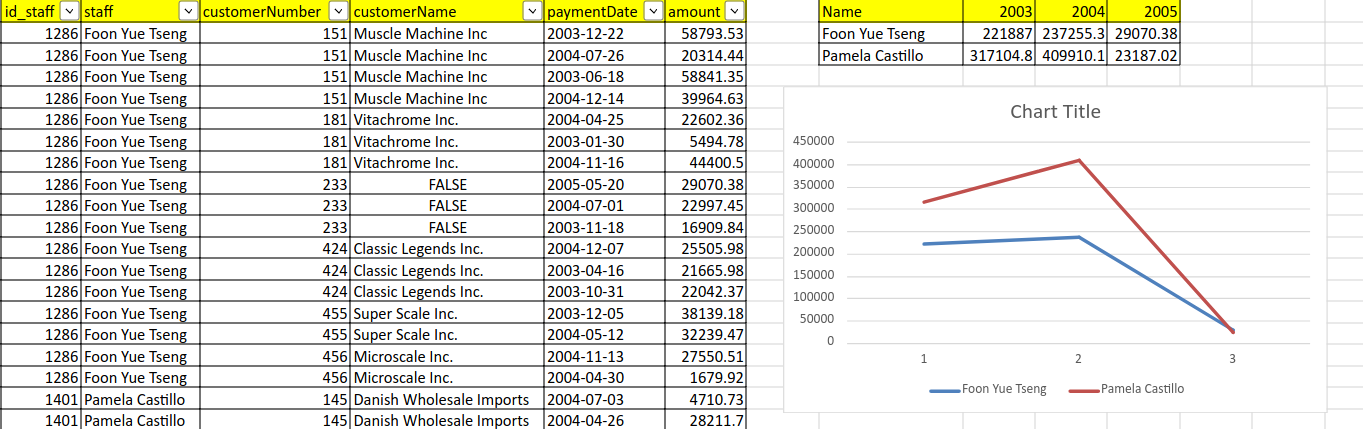


4. If the first KPI is "Number of customers who transact" while the second KPI is "Total turnover earned". Then, how many fields are needed to get this information?

- Number of customers who transacted 1 (customerNumber from order table)

- Total turnover obtained 2 (quantityOrdered and priceEach from orderdetails table).

5. Make an annual report for the KPI "Total turnover earned" for Foon Yue Tseng and Pamela Castillo. And draw a graph (line graph).



## 

## Case Study

Mr. Huhut is a shareholder of LegendVehicle. he needs a dashboard to see the development of sales (turnover) in each branch every year. Because the company has not recruited a Data Engineer, information retrieval can only be done through existing OLTP.

The desired report result is a graph based on the following table:

Analyze first:

1. What fields are required to display sales in each branch.

**- officeCode, city from offices table**

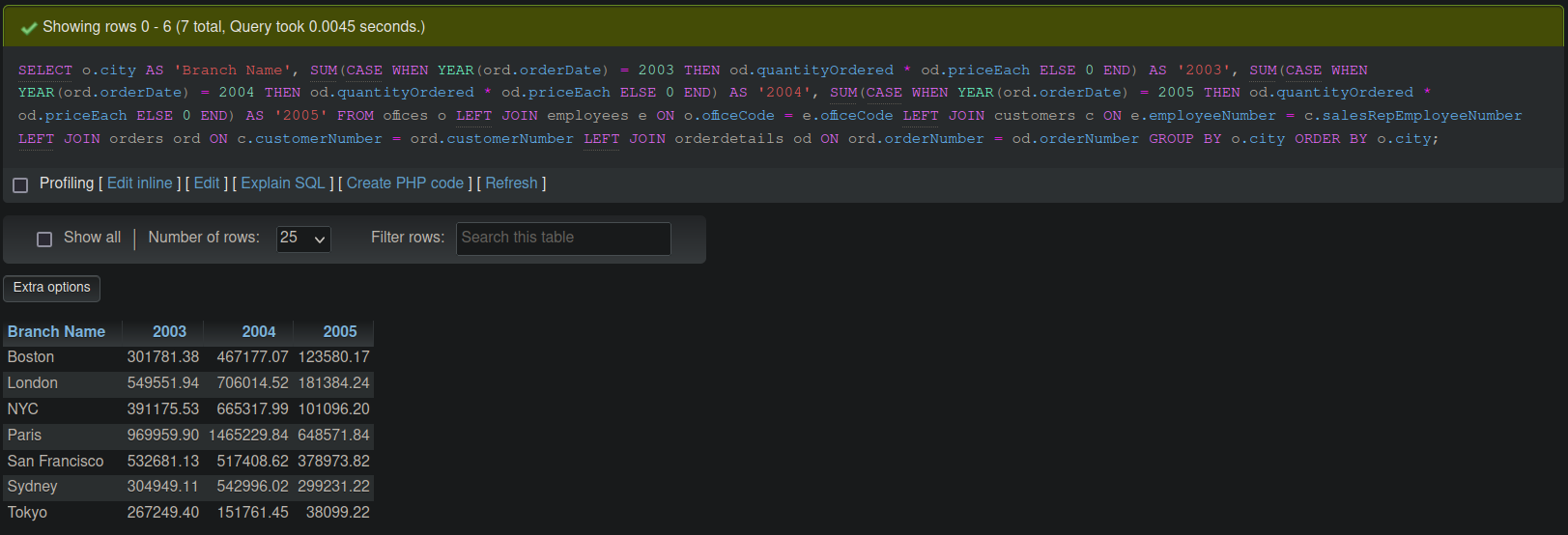
**- orderNumber, orderDate, customerNumber from orders table.**

**- quantityOrdered, priceEach, orderNumber from orderdetails table.**

**- employeeNumber, officeCode from employee table.**

**- salesRepEmployeeNumber, customerNumber from customers table.**

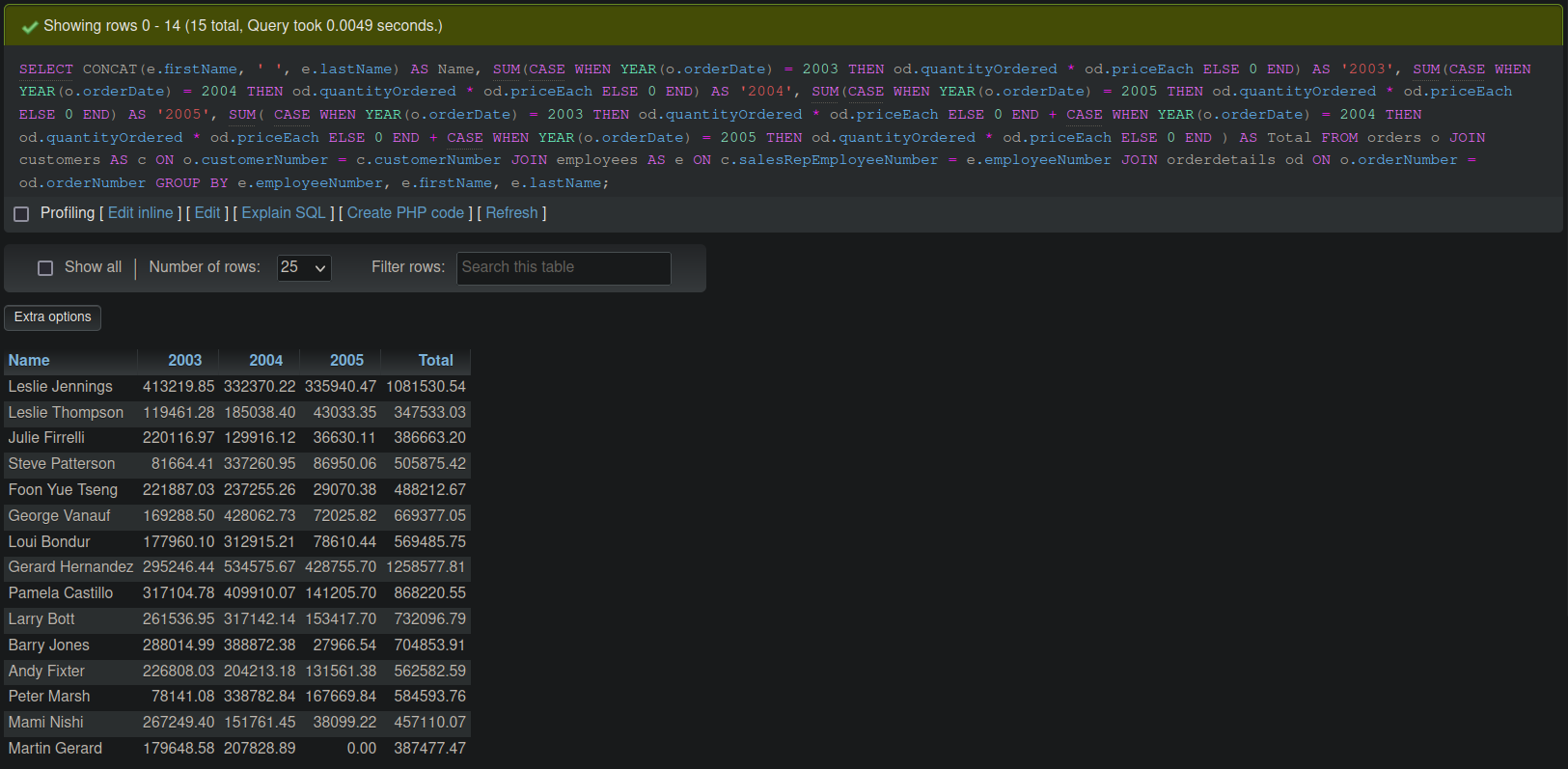
2. Form a query by considering the relationship between tables.



BONUS TEST: create another report with the same OLTP data source, analyze the fields used, form a query structure and write it in tables and graphs.

- Make an annual report for the KPI "Total turnover earned" for all staff.

Query



There are 5 field such as name, 2003, 2004, 2005 and total.

Happy exploring.

## Chapter 3: Database Analytical

**Practicum Objective**

After doing this practicum, students are expected to be more familiar with dimension tables, fact tables and what OLAP is.

**Case Study**

Based on the previous case study, the leadership of LegendVehicle is getting to know technology and has a visionary view. LegendVehicle will implement a data warehouse in its business process to analyze the transaction process seen from incoming payments. This process is used to analyze the KPI "amount of turnover owned".

To form the data warehouse, the Data Engineer at LegendVehicle needs to create a database that is used as an OLAP database.

Data from existing OLTP will be "ETL" to the OLAP database.

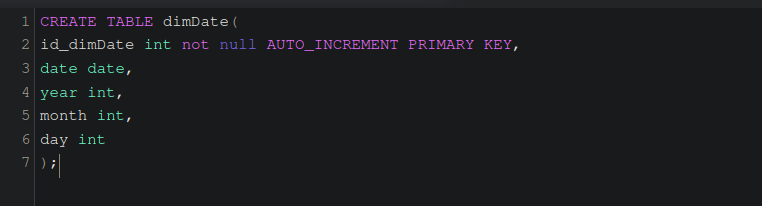
Translated with DeepL.com (free version)

## A. Time Dimension

1. Create a database that is used as OLAP with the name dw\_LegendVehicle.



2. Create a table to store time master data or what is called a dimension table. Name the table as dimDate.



In the next step, to create the dimDate dimension table, it is necessary to generate date data. The date data prepared in the dimDate table adjusts to the running business process.

The business process on LegendVehicle is 5 years. So the data in the dimdate table that must be available is the date for 5 years. Starting from January 1, 2003

1 Open PDI Spoon. Create a new Transformation -> File - New - Transformation.

2 Drag and Drop some objects namely:

- Generate Rows: used to create new data rows.

- Add Sequence: used to create a sequence, in this case creating data for each day.

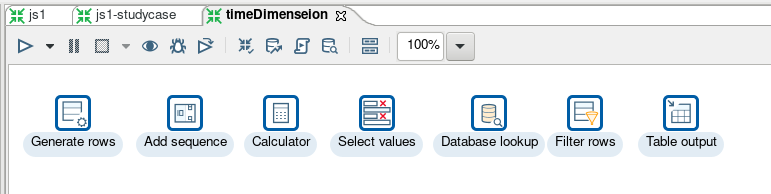
- Calculator: used to add days and retrieve year, month and day data.

- Select Values: used to select the fields used.

- Database Lookup: used to see and ensure that the data to be entered into the dimDate table is not twinned or the same as the data in the dimDate table itself.

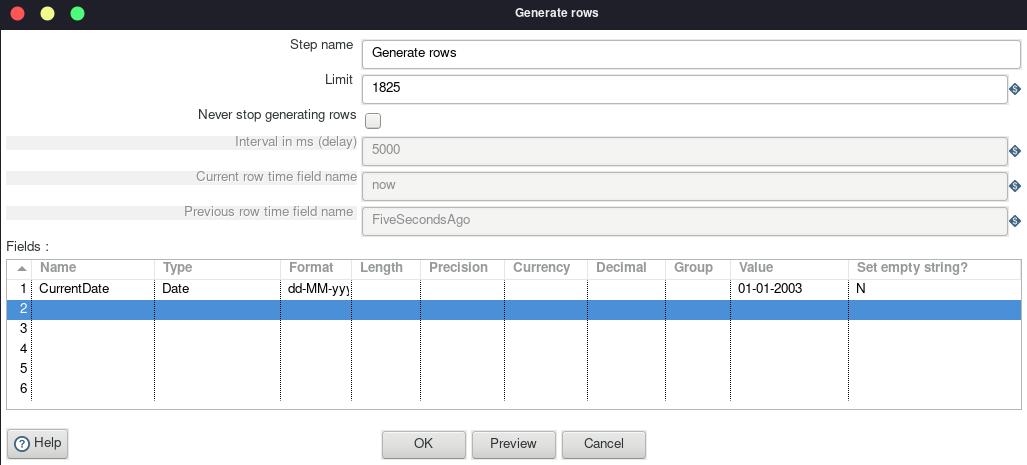
- Filter Rows: used to retrieve data that does not exist in the dimDate table after being checked previously.

- Output Table: used to store data in the destination table (dimDate).



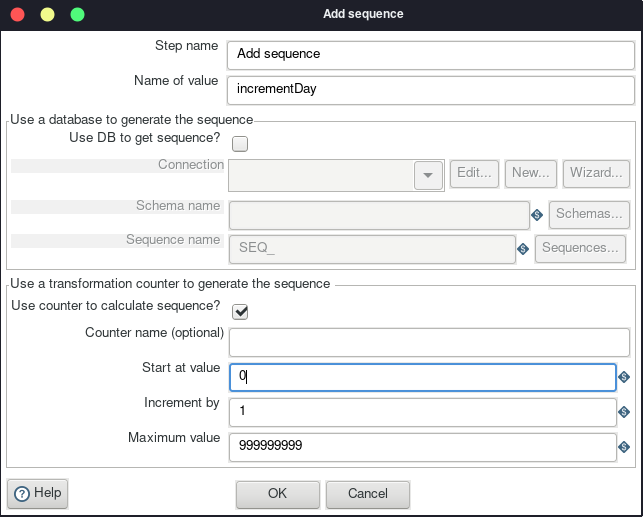
3. The configuration in Generate Rows is to change the limit to 1825 which means that the data that will be created is 1825 data. 1825 is the number of days in 5 years (365 days x 5 years).

4. Create new fields named CurrentDate with data type Date and format dd-MM-yyyy and initial value 01-01-2003.



5. Connect the output of Generate Rows to Add Sequence.

6. The configuration in Add Sequences is to change the Name of value to incrementDay with start value of 0 and increment by of 1.



7. Connect the output of add sequences to the calculator.

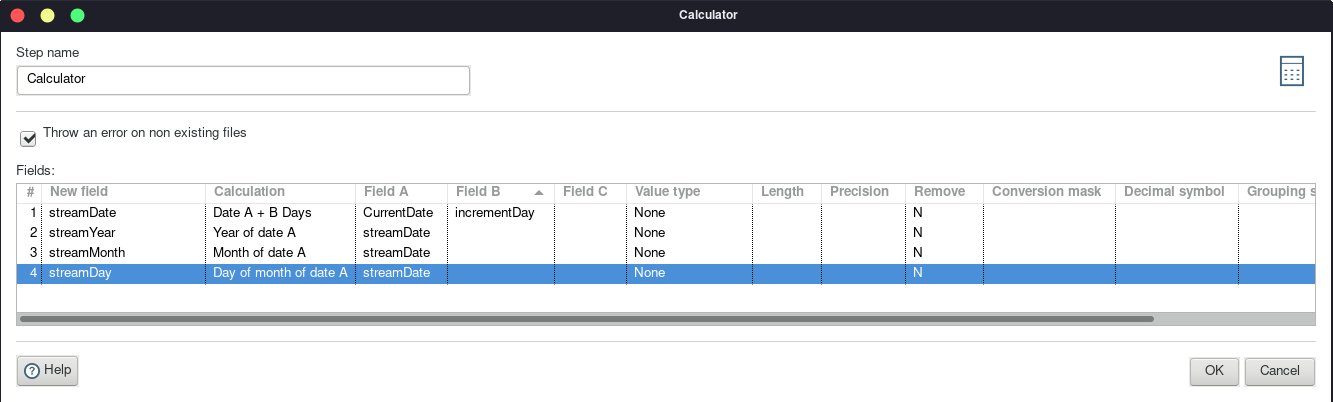
8. Configure the calculator by creating new fields as follows:

streamDate is the calculation of CurrentDate + incrementDay

streamYear is the Year of streamDate

streamMonth is the Month of streamDate

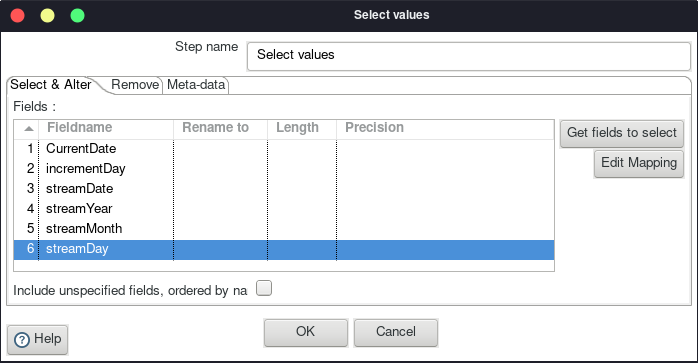
streamDay is the Day of month of streamDate

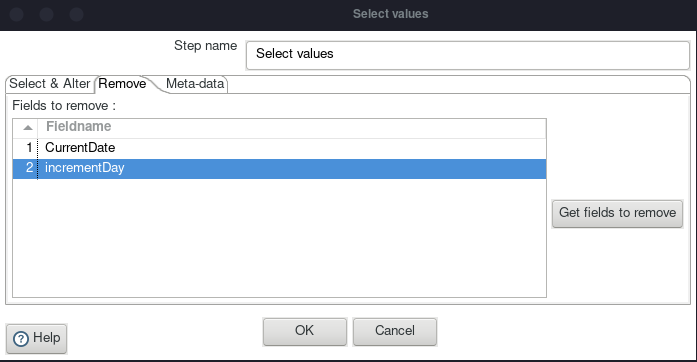


9. Connect the output of the calculator to Select values

10. Configure the select values by pressing the Get fields to select button on the Select & Alter tab. All fields from the input data will automatically appear.

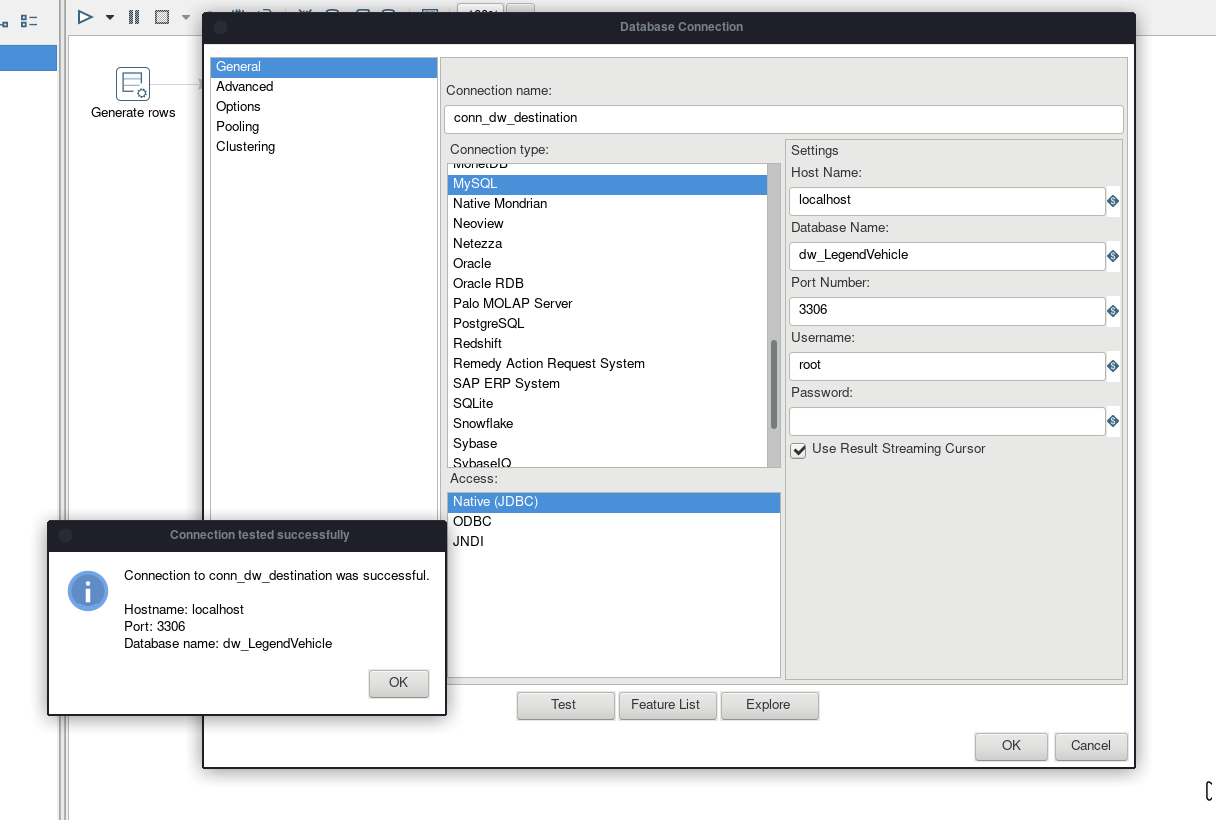
11. Because not all fields are used, the Remove tab contains the CurrentDate and incrementDay fields because both fields are not used.





12. Connect the select values output to the lookup database.

13. Before configuring the lookup database, first create a connection to the database via File - New - Database Connection. Use the MySQL Connection type with host name, database name, port number, username and password according to the MySQL configuration on each device. name the connection name conn\_dw\_destination.



14. The configuration in the lookup database is to provide a connection with the connection that was created in the previous step. with the database name schema used and the dimdate table that was created in the first step.

15. The fields that will be checked to see the similarity of the data contents so that they are not twins are:

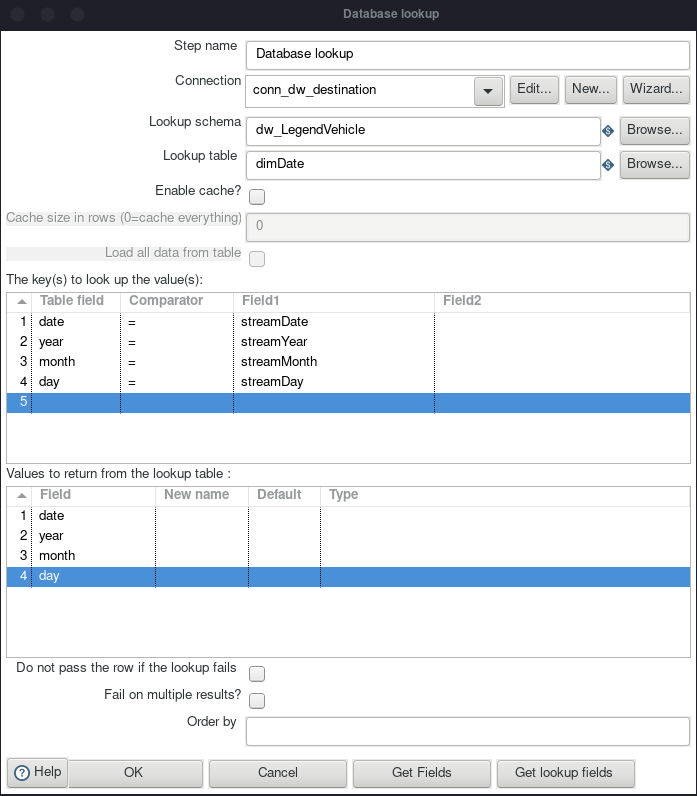
the date field in the dimdate table with the streamDate field

the year field in the dimdate table with the streamYear field

the month field in the table is updated with the streamMonth field

the day field in the table is updated with the streamDay field

16. The fields that will be retrieved are the fields in the dimDate table, namely date, year, month, and day.



17. Connect the output of the database lookup with filter rows

18. Configuration on filter rows is to configure the true output data in the output table. In this section, data that does not have similarities in the previous stages will be checked where if the Stream fields have nothing in common with the dimDate field, then the dimDate field will be null. In the condition statement write (date is null and year is null and month is null and day is null)

19. Connect the output of the filter rows to the output table.

20. Configure the output table to provide a connection to conn\_dw\_destination with schema dw\_legendvehicle and table dimdate.

21. Enable specify database fields.

22. On the Database fields tab, mapping the input data streamDate, streamYear, streamMonth and streamDay with the fields in dimDate. At this stage, the data will be inserted into the dimDate table.



23. check the contents of the dimdate table in the database. If successful, the dimdate table will be filled with 1825 data.



**TASK 1**

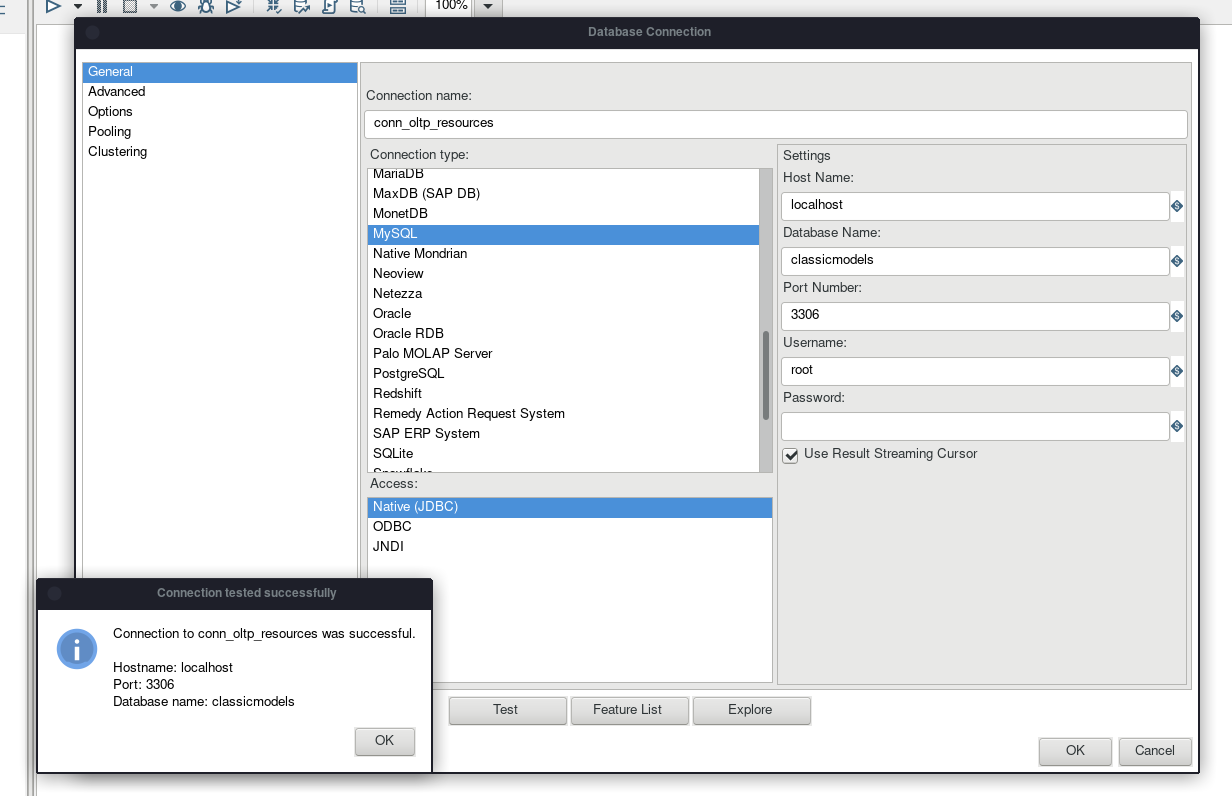
1. Open the preview tab in the execution result area of each object process. observe the input and output data. compare each process. explain the differences in each process.

|  |  |  |  |
| --- | --- | --- | --- |
| Object Process | SS Input Data | Ss Output Data | Description |
| Generate Rows | - |  | It produce CurrentDate field to process to next step. |
| Add Sequences |  |  | We fill input with CurrentDate field from previous step, then produce incrementDay to add each one day. |
| Calculator |  |  | We fill input with previous step, then produce several field to calculate day. |
| Select Values |  |  | We fill input with output previous step, then produce several field to select and remove field. |
| Database Lookup |  |  | We fill input with output previous step, then produce several field to check whether field that compare is not same. |
| Filter Rows |  |  | We fill input with output previous step, then produce several field to check each field is null, then it will be filled to table output. |
| Table Output |  |  | We fill input with output previous step, then produce several field to filled into table database. |

## B. Employee Dimension

1. Create the dimEmployees table in dw\_legendVehicle.

2. In PDI Spoon create a new connection with the name conn\_oltp\_resources that connects to the oltp database. adjust the hostname, database name, port number, username and password with the conditions on each device.



3. Drag and drop some objects as follows:

Table input: used to retrieve data from the OLTP database.

Select values: select fields that are used for the Transform and Load process.

Database lookup: used to look at the data in the dimEmployees table to make sure the data is not duplicated.

Filter rows: used to select data streams that are still missing in the dimEmployees table.

Table output: Insert data into the dimEmployees tavle

4. Configure the input table by connecting Connection to the conn\_oltp\_resources node. To retrieve the source data use the query below.

5.Connect the input table output to select values.

6. Configuration on Select values is to retrieve data from the fields employeenumber, lastname, firstname, jobtitle, lastname\_1 and firstname\_1 as data streams used in the ETL process on the select & alter tab.

7. Remove other fields that are not used on the remove tab.

8. Connect the output select values to the lookup database

9. Configure the database lookup by connecting the connection in conn\_dw\_destination with the dimEmployees lookup table that was created in the first stage

10. The fields that are lookup are the fields in the dimEmployees table with the input field stream from OLTP. while the fields that are retrieved are the fields from dimEmployees itself. If there is no similar data then null will appear.

11. Connect the lookup database output with filter rows.

12. In the filter rows give a null field condition on the dimemployees field to be included in the next process. It indicates that the data stream has no similarity to the data in dimemployees.

13. Connect the output of the filter rows with the output table.

14. In the output table, use connection conn\_dw\_destination to insert the data in the dimemployees table.

15 Enable specify databse fields, and mapping the input data stream from oltp to the fields in dimemployees.

16. if the whole process is successful, the dimemployees table will be filled with employee data from the OLTP database.

TASK 2

1. Open the preview tab in the execution result area of each process object. observe the input and output data. compare each process. explain the differences in each process.

2. If the process is repeated (run again), will the data be redundant?

3. Add your name to the employee table in OLTP. run this transformation again. Observe the results, what happens?

## C. Payment Fact

## D. Jobs

## Study Case