TUNIS BUSINESS SCHOOL
UNIVERSITY OF TUNIS

Mini Project Report

Business Intelligence and database management system

Sales Insight 360

Submitted by:

- Ranim Sayahi
- Ghaith Dkhili
- Safa Zaghdoudi

Submitted to:

-Prof. Manel Abdelkader

-Prof. Ameni Azzouz

Table of contents

1. Introduction

Our business intelligence project, « Sales insight 360 », is focused on analyzing and gaining insights into the sales performance throughout different business aspects such as: suppliers, shippers, products, customers, and employees.

This project will allow us to explore these different objectives and help derive actionable insights from the dataset, leading to informed business decisions and improvements.

- **Sales Performance Analysis**: Evaluate overall sales performance. Identify top-performing products and employees. Analyze sales trends over time.
- **Customer Behavior Analysis**: Understand customer purchasing behavior. Identify key customer segments. Evaluate customer retention and acquisition.
- **Employee Performance Analysis:** Evaluate the contribution of employees to sales. Identify top-performing and underperforming employees. Analyze employee efficiency in processing sales orders.
- **Product and Supplier Analysis:** Assess the popularity and profitability of products. Analyze the performance of suppliers. Identify relationships between products and suppliers.
- **Inventory Management:** Monitor product stock levels. Optimize inventory based on sales trends. Identify products with high or low turnover.
- **Geographic Analysis:** Analyze sales performance across different regions. Identify regions with high or low sales. Understand regional preferences and trends.
- **Profitability Analysis:** Evaluate the overall profitability of sales. Assess the contribution of various products to overall profit. Analyze the cost-effectiveness of suppliers.
- **Forecasting and Planning:** Use historical data to forecast future sales. Plan inventory, staffing, and other resources accordingly. Implement strategies to meet future demand.
- **Operational Efficiency:** Evaluate the efficiency of sales order processing. Identify bottlenecks or areas for process improvement. Optimize the overall operational workflow.

2. Work environment:

-Talend:

Talend is an open-source data integration platform that enables businesses to quickly connect, access, transform, and integrate data from various sources.

-Python:

“Pandas” library provides powerful tools and functions to manipulate and clean datasets efficiently, making it a go-to choice for data cleaning tasks in Python.

-Power Bi:

Power BI is a user-friendly business analytics service by Microsoft that enables the creation of interactive reports and dashboards. It integrates with diverse data sources, supports data transformation, and offers powerful data visualization options.

-Oracle :

-Oracle SQL Developer Data Modeler is a graphical tool that enhances productivity and simplifies data modeling tasks.

-Oracle SQL Developer

3. Implementation:

2.1 Data Gathering:

We extracted the following raw datasets:

-A csv file containing information about « HR employees »: (empid, lastname ,firstname, title, titleofcourtesy, birthdate, hiredate, address, city, region, postalcode, country, phone, mgrid)

-An excel file containing information about « Production Categories »: (categoryid, categoryname, description)

-An excel file containing information about « Production Products »: (productid, productname, supplierid, categoryid, unitprice, discontinued)

-A text file containing information about « Production Suppliers »: (supplierid, companyname, contactname, contacttitle, address, city, region, postalcode, country, phone, fax)

-A csv file containing information about « Sales Customers »: (custid, companyname, contactname, contacttitle, address, city, region, postalcode, country, phone,fax)

-A csv file containing information about «Sales orders details »: (orderid, productid, unitprice, qty, discount)

-A csv file containing information about « Sales orders »: (custid, empid, orderdate, requireddate, shippeddate, shipperid, freight,shipname, shipaddress, shipcity, shipregion, shippostalcode, shipcountry)

-A csv file containing information about « Sales shippers »: (shipperid, companyname, phone)

3.2 Data preparation (ETL):

-First method: “Python”:

-First, we loaded the library «Pandas » since it provides data structures for efficiently storing and manipulating large datasets, as well as tools for reading and writing various data formats.

```
[ ]  import pandas as pd
```

-Then, we loaded the data needed from the files.

```
[ ]  # Loading sales_orders and sales_orders_details data from files
     sales_orders = pd.read_excel("SalesOrders.xlsx")
     sales_order_details = pd.read_excel("SalesOrderDetails.xlsx")
     # Load all datasets
```

-Consequently, we defined a function that will be responsible for the data cleaning: replacing the missing and null variables by predefined structures.

```
# Set fields to correct format
def set_data_types_df(result_df, numeric_var_list, string_var_list, date_var_list):
    # Replace missing numeric values by 0
    result_df[numeric_var_list] = result_df[numeric_var_list].fillna(0)
    # Set numeric values to be float
    result_df[numeric_var_list] = result_df[numeric_var_list].astype("float")
    # Replace missing string values by 'undetermined'
    result_df[string_var_list] = result_df[string_var_list].fillna('undetermined')
    # Set string values to be string
    result_df[string_var_list] = result_df[string_var_list].astype("string")
    # Replace missing date values by '1970-01-01' when encountering this date automatically you'll notice it's false
    result_df[date_var_list] = result_df[date_var_list].fillna("1970-01-01")
    result_df[date_var_list] = result_df[date_var_list].astype("datetime64[ns]")

    # Code to simplify the format of the date in the output
    for var in date_var_list:
        result_df[var] = result_df[var].dt.date
        result_df[var] = pd.to_datetime(result_df[var])
```

-We then applied this function on the loaded datasets.

```
[ ]  # Correcting format for these datasets
     set_data_types_df(sales_orders, ['freight'],
                       ['orderid', 'custid', 'empid', 'shipperid','shipname', 'shipaddress',
                      'shipcity', 'shipregion', 'shippostalcode', 'shipcountry'],
                       ['orderdate', 'requireddate', 'shippeddate'])
     set_data_types_df(sales_order_details, ['qty', 'discount', 'unitprice'],
                       ['orderid', 'productid'],
                       [])
```

-After cleaning the data, we merged the two datasets to form a new one.

```
[ ]  # Join the two tables
     Sales = pd.merge(sales_orders, sales_order_details, on=['orderid'], how='left')
     Sales
```

-Finally, we saved the final dataset in excel format.

```
[ ]  # Save datasets
     Sales.to_excel('Sales.xlsx', index=False)
```

⇨ We did the same process for the rest of the datasets.

-Second method: “Talend”:

1. **Extract**

We used the offline extraction which is a kind of physical extraction: we copied the data from the data source to a “landing area” then our extraction process connects to that external file and starts processing.

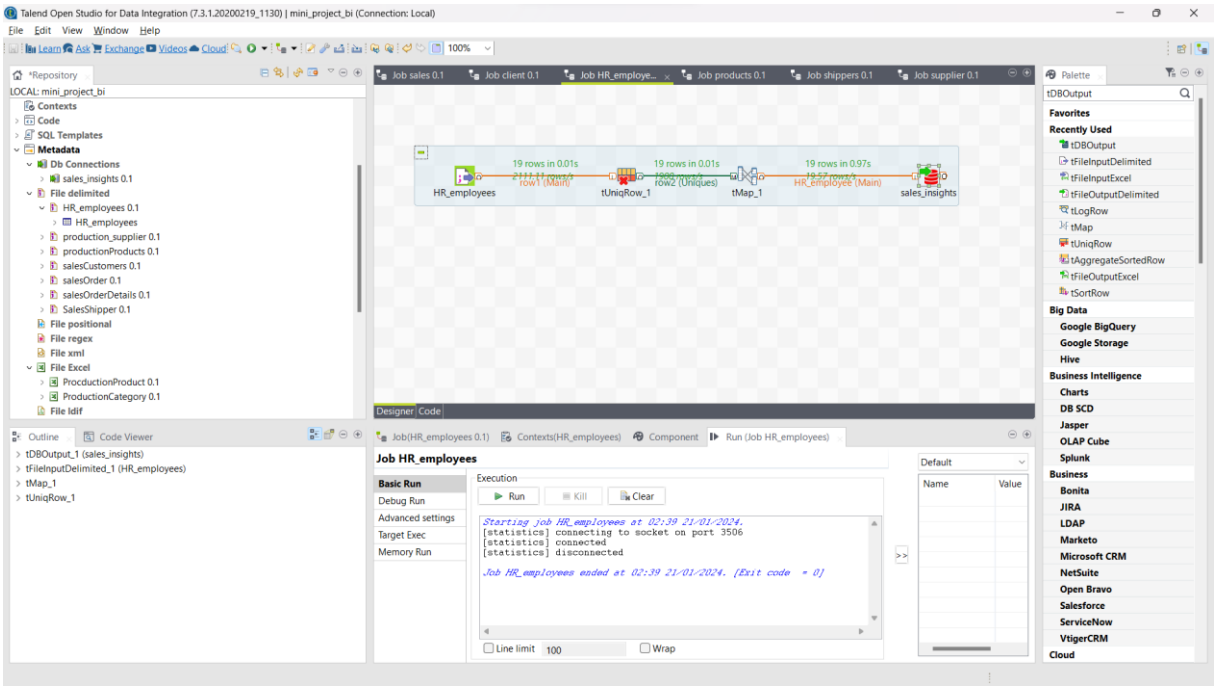
In Talend, we created delimited file for each CSV and text file and an Excel file for files of excel file.

2. **Transform**

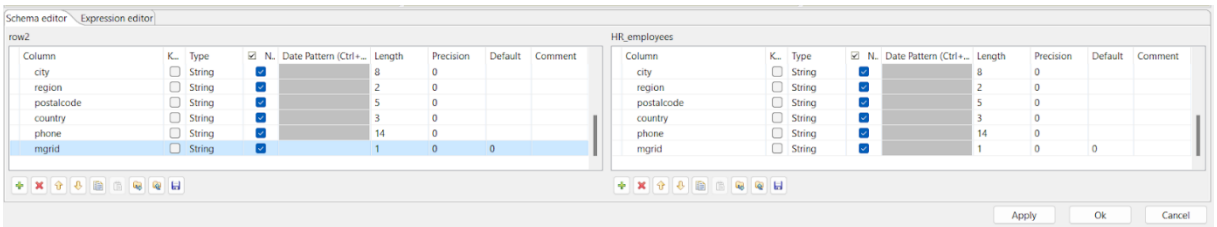
Using Talend we used both Basic Transformation and advanced Transformations to convert the format or the structure of our data.

*Basic transformation:

In the job HR_employees we used the tUniqRow to remove duplicate data if exists:

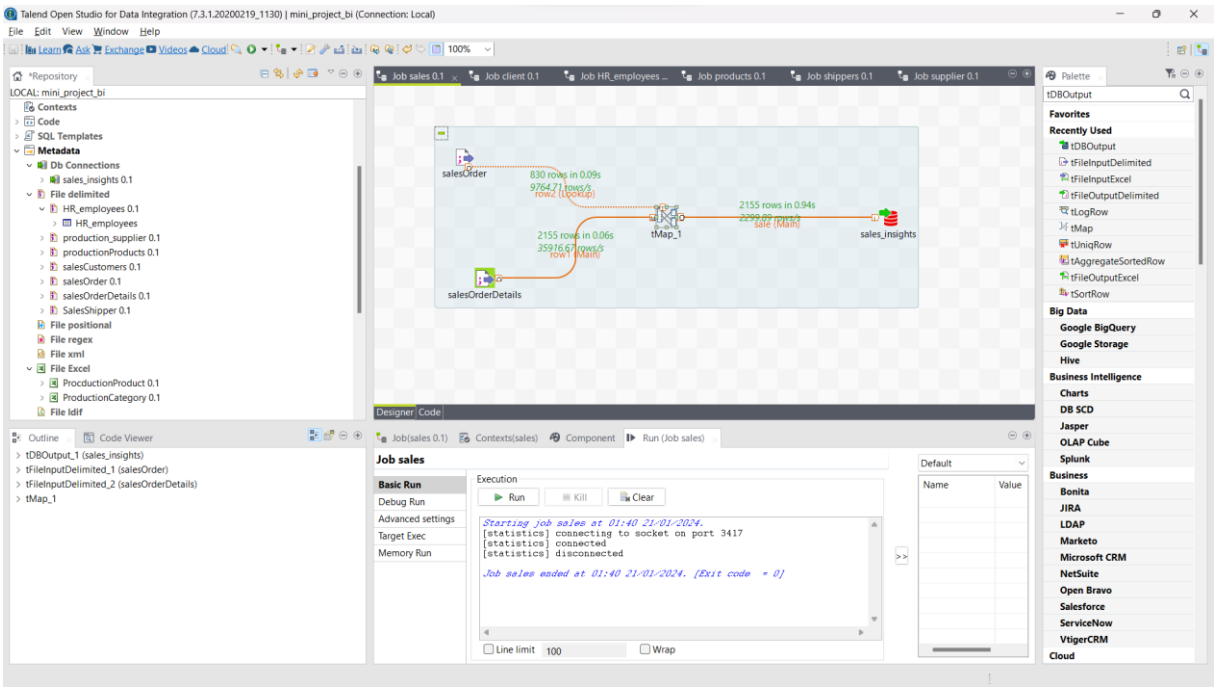


Also, in the same job we used tMap to map null values of mgrid to 0:

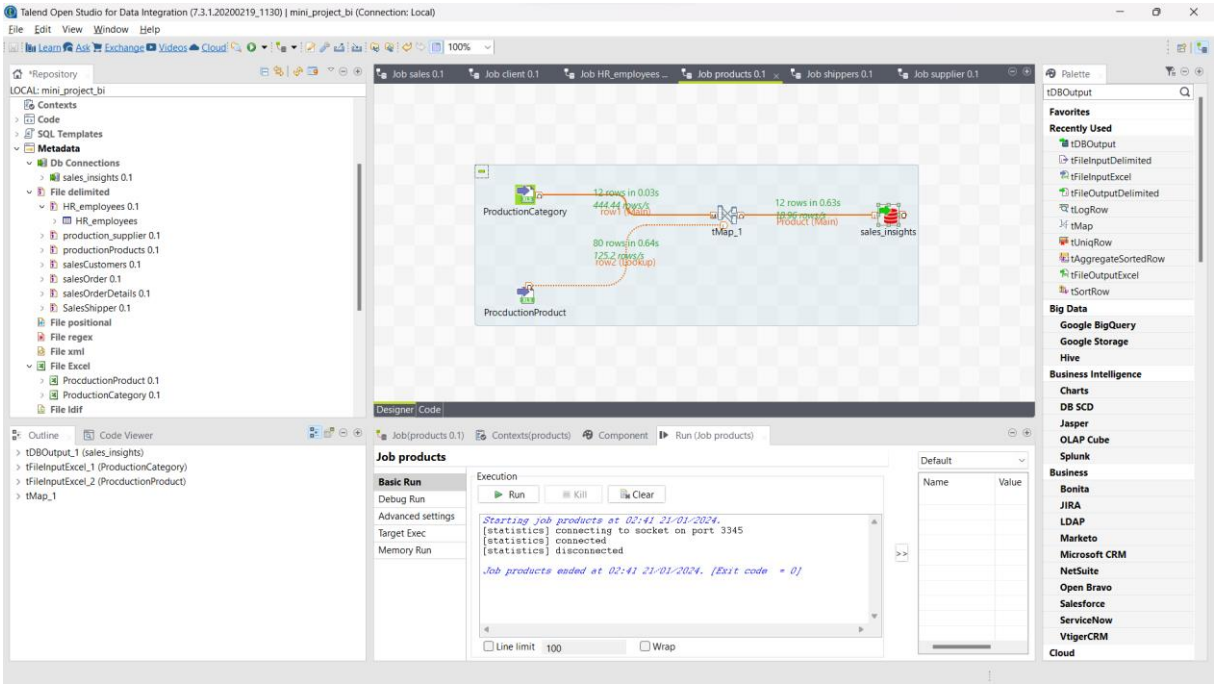


*Advanced transformation

To create the job Sales, we used the tMap function to join the data from the SalesOrder file and the SalesOrderDetails file:



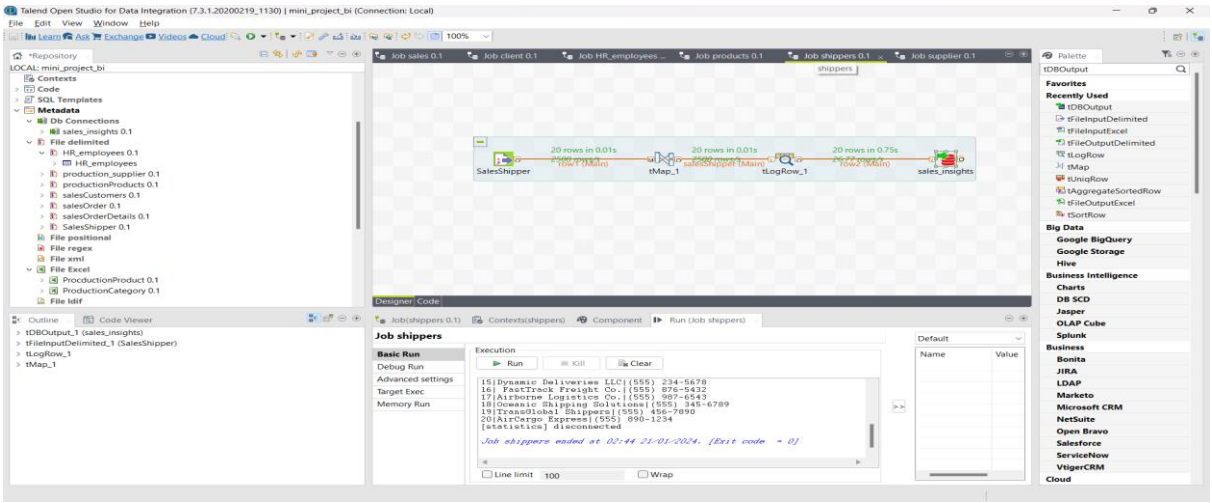
We did the same for the job products; we used the tMap to join the data of two datasets Production Products and Production Categories:



3. Load:

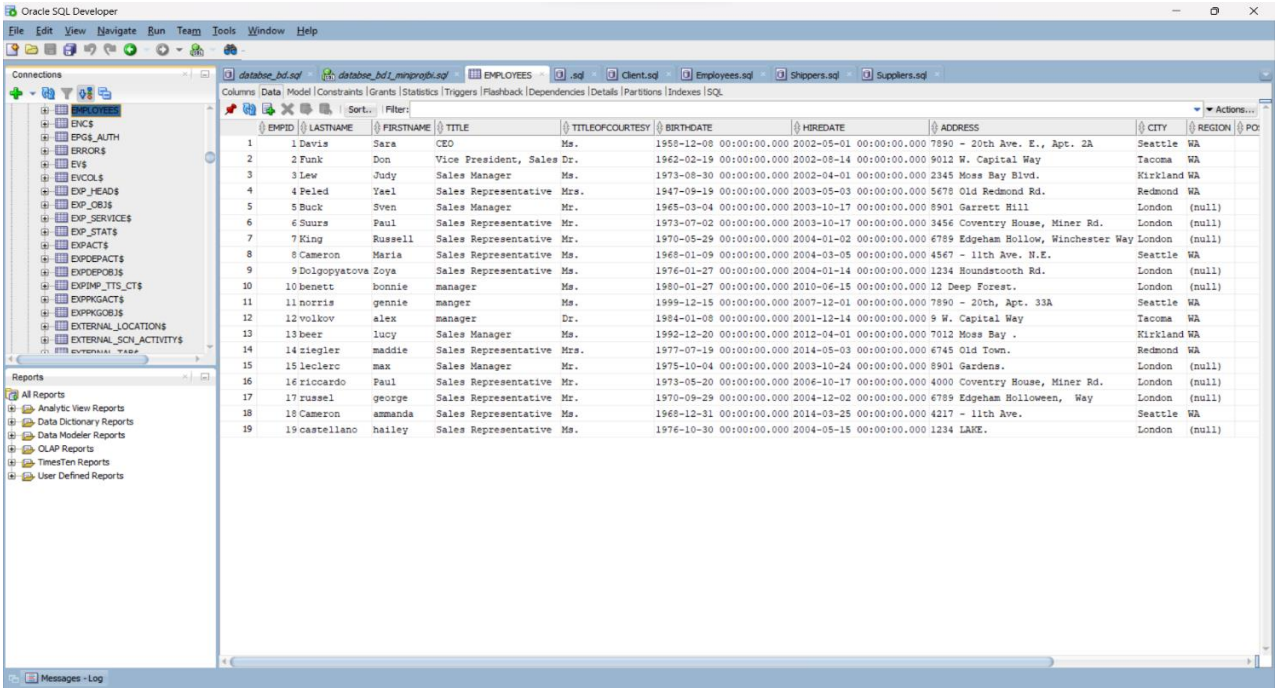
Before loading the data, we created a database in Oracle SQL Developer named ‘Sales Insight 360’ and connected it in talend using Db connection.

We used the function tLogRow for every job to check what our data looks like at different stages of our data transformation process. Then, we used the tDBOutput to load the result of our talend jobs to Oracle SQL Developer.



3.3 Data storage:

-As mentioned in the loading part of the ETL, the table below represents an example of the output.



	EMPID	LASTNAME	FIRSTNAME	TITLE	TITLEOFCOURTESY	BIRTHDATE	HIREDATE	ADDRESS	CITY	REGION
1	1	Davis	Sara	CEO	Ms.	1958-12-08 00:00:00.000	2002-05-01 00:00:00.000	7890 - 20th Ave. E., Apt. 2A	Seattle	WA
2	2	Funk	Don	Vice President, Sales	Dr.	1962-02-19 00:00:00.000	2002-08-14 00:00:00.000	9012 W. Capital Way	Tacoma	WA
3	3	Lew	Judy	Sales Manager	Ms.	1973-08-30 00:00:00.000	2002-04-01 00:00:00.000	2345 Moss Bay Blvd.	Kirkland	WA
4	4	Peled	Yael	Sales Representative	Mrs.	1947-09-19 00:00:00.000	2003-05-03 00:00:00.000	5678 Old Redmond Rd.	Redmond	WA
5	5	Buck	Sven	Sales Manager	Mr.	1965-03-04 00:00:00.000	2003-10-17 00:00:00.000	8901 Garrett Hill	London	(null)
6	6	Suurs	Paul	Sales Representative	Mr.	1973-07-02 00:00:00.000	2003-10-17 00:00:00.000	3456 Coventry House, Miner Rd.	London	(null)
7	7	King	Russell	Sales Representative	Mr.	1970-05-29 00:00:00.000	2004-01-02 00:00:00.000	6789 Edgeham Hollow, Winchester Way	London	(null)
8	8	Cameron	Maria	Sales Representative	Ms.	1968-01-09 00:00:00.000	2004-03-05 00:00:00.000	4567 - 11th Ave. N.E.	Seattle	WA
9	9	Dolgopyatova	Zoya	Sales Representative	Ms.	1976-01-27 00:00:00.000	2004-01-14 00:00:00.000	1234 Houndstooth Rd.	London	(null)
10	10	benett	bonnie	manager	Ms.	1980-01-27 00:00:00.000	2010-06-15 00:00:00.000	12 Deep Forest.	London	(null)
11	11	norris	gennie	manger	Ms.	1999-12-15 00:00:00.000	2007-12-01 00:00:00.000	7890 - 20th, Apt. 33A	Seattle	WA
12	12	volkov	alex	manager	Dr.	1984-01-08 00:00:00.000	2001-12-14 00:00:00.000	9 W. Capital Way	Tacoma	WA
13	13	beer	lucy	Sales Manager	Ms.	1992-12-20 00:00:00.000	2012-04-01 00:00:00.000	7012 Moss Bay .	Kirkland	WA
14	14	ziegler	maddie	Sales Representative	Mrs.	1977-07-19 00:00:00.000	2014-05-03 00:00:00.000	6745 Old Town.	Redmond	WA
15	15	leclerc	max	Sales Manager	Mr.	1975-10-04 00:00:00.000	2003-10-24 00:00:00.000	8901 Gardens.	London	(null)
16	16	riccardo	Paul	Sales Representative	Mr.	1973-05-20 00:00:00.000	2006-10-17 00:00:00.000	4000 Coventry House, Miner Rd.	London	(null)
17	17	russel	george	Sales Representative	Mr.	1970-09-29 00:00:00.000	2004-12-02 00:00:00.000	6789 Edgeham Holloween, Way	London	(null)
18	18	Cameron	ammanda	Sales Representative	Ms.	1968-12-31 00:00:00.000	2014-03-25 00:00:00.000	4217 - 11th Ave.	Seattle	WA
19	19	castellano	hailey	Sales Representative	Ms.	1976-10-30 00:00:00.000	2004-05-15 00:00:00.000	1234 LAKE.	London	(null)

3.4 Data modeling:

3.4.1 Fact:

The fact table is the "Sales" dataset with a primary key orderid. This table is the result of the merge between the two datasets sales orders and sales orders details.

3.4.2 Dimensions:

Dimensions included in this dataset are:

-Clients:

Cust id (primary key) / company name / Contact name / contact title / address /city / region/ postal code / country / phone / fax

-HR employees:

Emp id (primary key) / last name / first name / title / title of courtesy / birth date / hire date / address / city / region / postal code / country / phone / mgrid

-Shippers:

Shipper id (primary key) / company name / phone

-Products:

Category id (primary key) / Product id / product name/ supplier id / unit price / discontinued / category name/ description.

-Suppliers: **=> this dimension is derived from the products dimension. (Many to many relationship)**

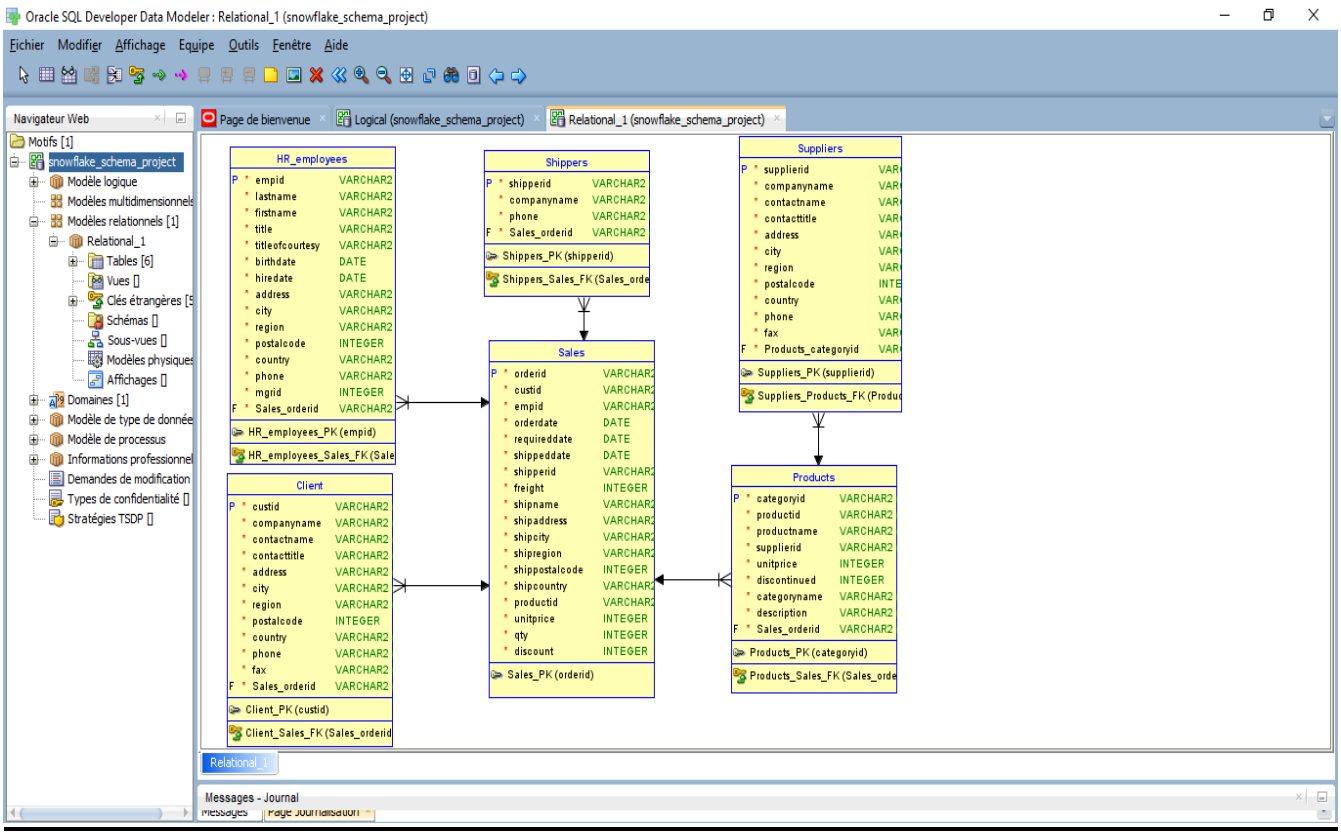
Supplier id (primary key) / company name / contact name / contact title/ address / city / region / postal code / country / phone / fax

=> Our model is **a snowflake schema**. And to visualize this we used two different programs.

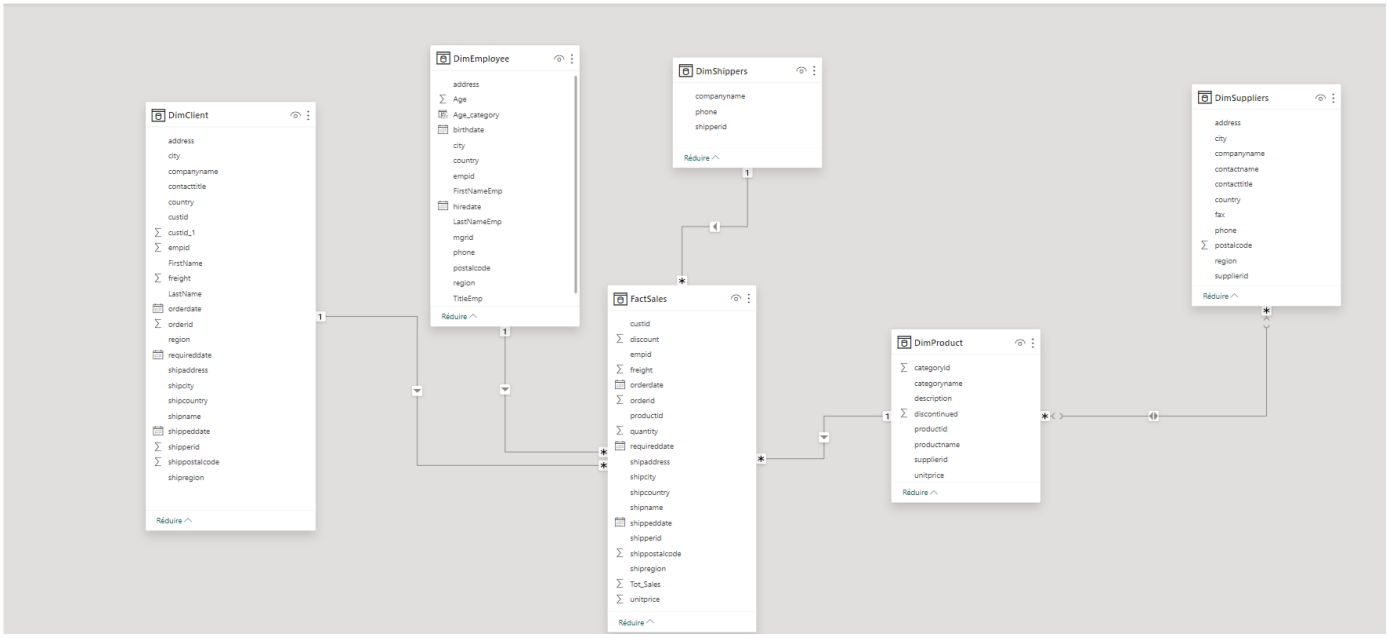
-> SQL Data Modeler

-> Power BI

-First method: “SQL Data Modeler”:



-Second method: “Power BI”:



3.5 Data visualization:

An overall analysis of the power BI output:

-Total products sold is 51K.

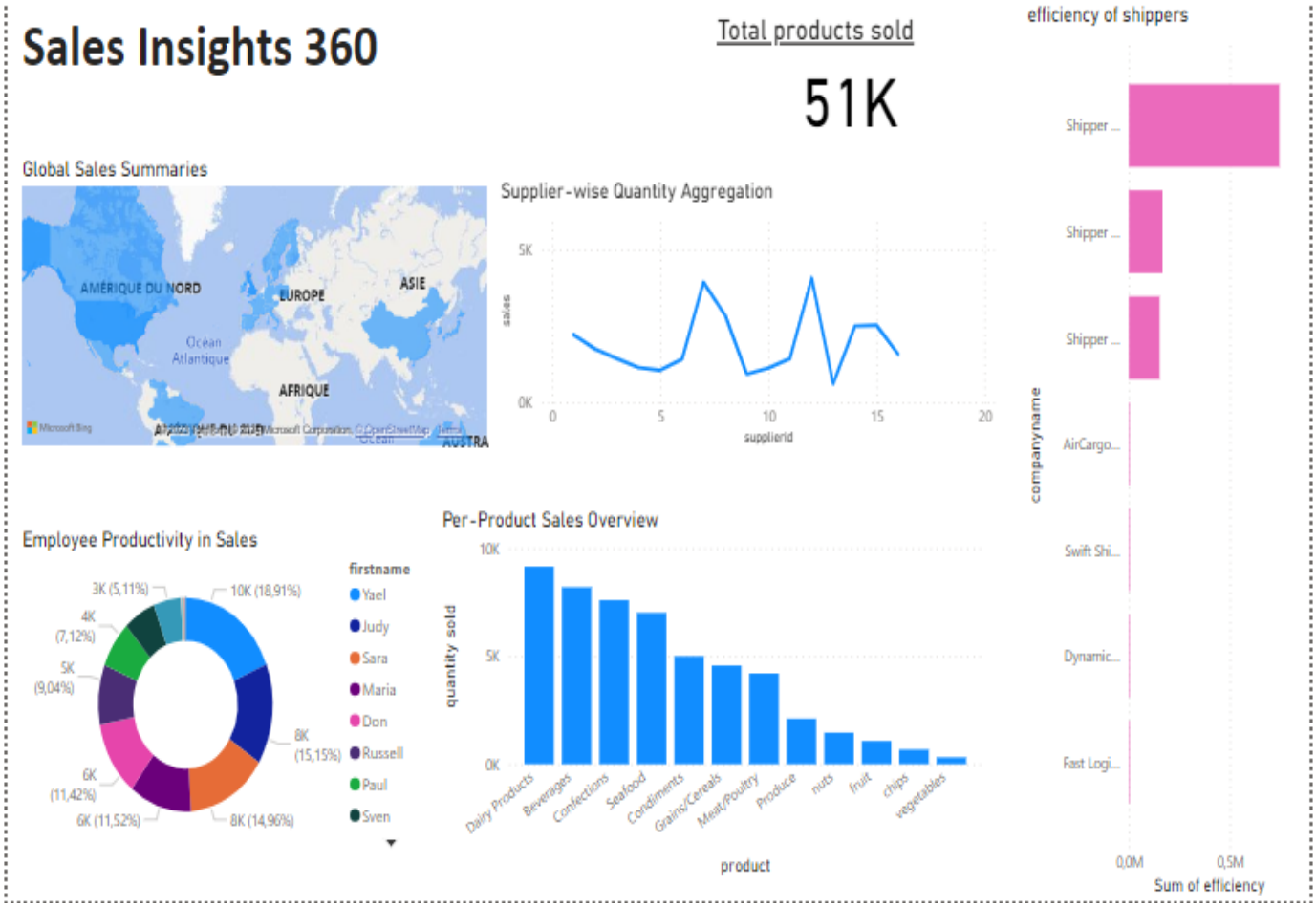
-According to the pie chart, our most profitable(sold) products are the dairy products followed by the Beverages. Beverages, confections, and seafood have very close quantity sold. Our least profitable products are vegetables.

-Following the map, our products are most popular in North and south America and less popular in Asia and Europe. We also notice that are products are not available in Africa and Australia.

-According to our donut chart, the employee with the highest quantity of products sold is Yael with 18.91% (or 10k products).

-Following the line chart in our Power BI report, our most reliable supplier is the supplier with 12 as id and our least reliable one is the one with 13 as id.

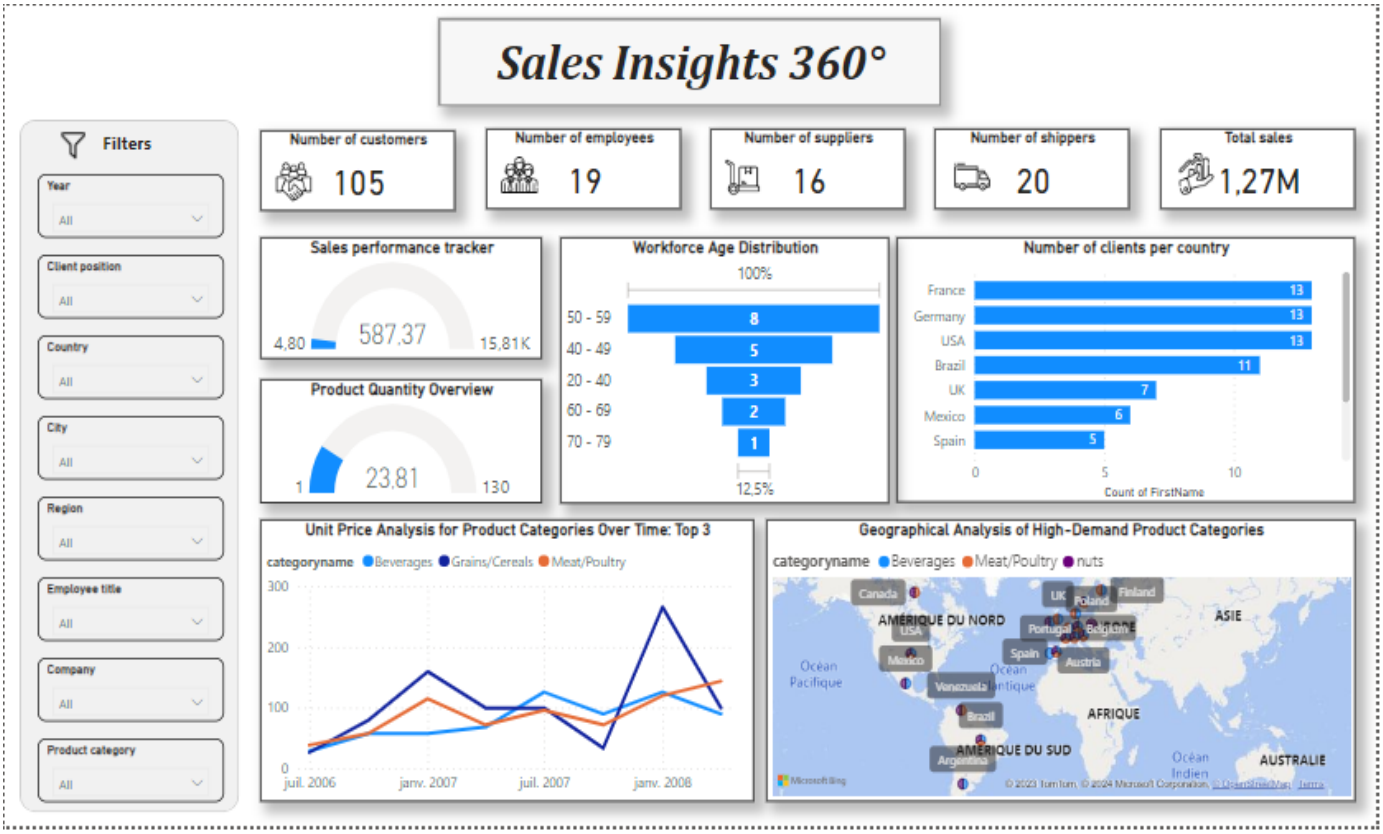
-The stacked pie chart shows the efficiency of the shipping companies through the number of the days remaining after the shipping until the required shipping day. Shipper ETYNR is our most effective and efficient one while oceanic shipping solutions is the least effective with a total delay of 65 days.



More advanced analysis:

-With no filter on, we can observe: The number of customers, employees, suppliers, shippers, total sales. We can also see the sales performance tracker, the workforce age distribution etc…

-We added a filter option which can help us see different results of the sales performance based on different filters such as year, client position, country, city...



=> In summary, this Power BI report serves as a powerful tool for stakeholders to gain actionable insights into our company's overall sales performance. Leveraging these insights can lead to strategic improvements in products, customer relationships, supply chain efficiency, and employee productivity. As we continue to refine and expand our data-driven approach, we position ourselves for sustained success in a dynamic business environment.

4.Conclusion:

In conclusion, the "Sales Insight 360" business intelligence project has proven to be a valuable and comprehensive tool for analyzing and gaining insights into the sales performance across various critical business aspects. The project's focus on suppliers, shippers, products, customers, and employees has provided a holistic view of the sales ecosystem, enabling the making of informed decisions and optimizing the overall business strategy.

While the project has brought significant benefits, it's essential to recognize that the landscape of business intelligence is dynamic. Continuous updates and refinements to the system will be crucial to staying ahead of evolving market trends and maintaining the project's effectiveness.