TUNIS BUSINESS SCHOOL

2024

GOLD DEALER

Business Intelligence & Database Management

Prepared by : Asma Boubaker
Kenza Bacha
Yosr Jaouadi

TABLE OF CONTENTS

In **Kimball's approach**, the emphasis is on building a comprehensive and integrated data warehouse that facilitates business intelligence and decision-making through well-defined dimensional modeling and a focus on business processes. Following this approach main steps, our project will consist of the following phases:

01 COMPANY INTRODUCTION AND REQUIREMENT GATHERING

01.01 ABOUT OUR COMPANY

Gold Dealer is a jewellery chain that runs a number of stores in several US states, where it is known for offering luxury products crafted primarly with diamond and gold. The group's main departments are Marketing, Finance and Resources Management.
At each period of the year, this famous group tries to analyze its main business processes in order to set new objectives for the following period.

01.02 REQUIREMENTS GATHERING

After a thorough revision of the managers requirements, the different departments are trying to answer the following questions:

1- **Which store should the company close in the different states or Is it a suitable strategy to expand into new states? Which the top-performing store in terms of profit, and how can the company strategically allocate resources to maximize its potential?**

2- **What is the most lucrative day of the week, and how do sales patterns change in the periods leading up to and following the Christmas Holidays?**

3- **What is the best-selling product for the company? What category is best suited to be sold in each store?**

4- **Do the marketing department need to update the catalogue by reducing specific items or creating new ones?**

5- **Analyzing the stock prices of gold and diamonds and based on historical price patterns, which material is more cost-efficient for crafting jewellery, and should the store consider a change in materials?**

6- **How can the store tailor its product offerings to match the preferences and purchasing behaviours of customers?**

02 DATA AND RESOURCES GATHERING

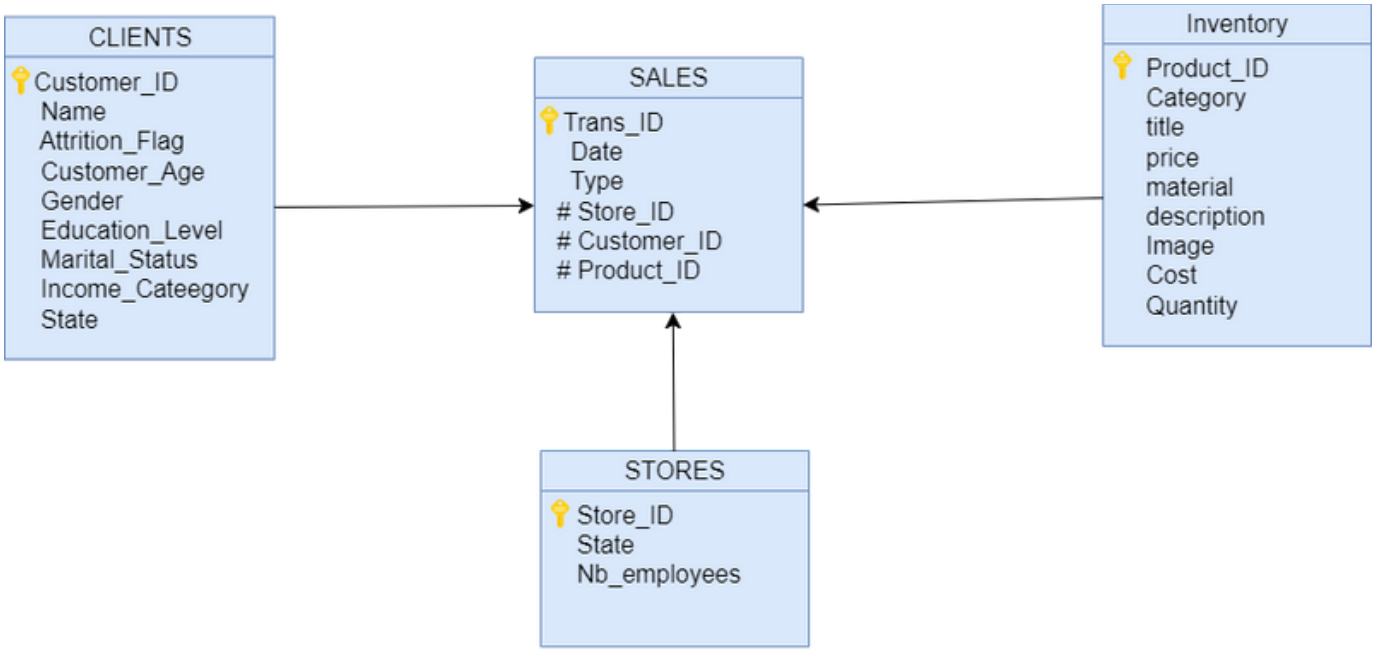
The primary data sources that we will need all along the project are of **two main types**:

02.01 COMPANY INTERNAL DATABASE

The company internal database is a structured relational repository that organizes and stores its essential business information. It simply provides basic data supporting its business processes like inventory management, sales tracking, and customer interactions.

The database (name: Gold_dealer) is made of 4 tables: SALES, CLIENTS, STORES, and INVENTORY, all linked through relations including their Primary and Foreign keys.

To illustrate the structure of the internal database, Here is its **Relational Schema**:



REMARK:

The data within this database is in fact collected from multiple sources:

CLIENTS and INVENTORY: these datasets are available in KAGGLE website, they contain data according to the attributes available in the schema.

STORES and SALES: are both simulated and random data we created manually with Microsoft Excell to conform to the company's database.

Finally, with these four different datasets, we created the SQL database with SQL Developer by importing all files and creating their relationships, all within Oracle SQL Server.

02.02 EXTERNAL DATA SOURCES

The company core competency lies in innovative jewelery design and that includes managing its raw materials budget. Therefore, it always needs updated information about Gold and Diamond stock prices (as well as prices historical data) to efficiently allocate its investment resources to its raw materials inventory.
To this matter, we need the following available external data:

- **Gold Prices:** A CSV file describing gold stock prices from 2018 till 2023, it contains following attributes:
  Date, Open, High, Low, Close*, Adj Close**, Volume.
  Link to the file:
  https://markets.businessinsider.com/commodities/gold-price

- **Diamond Prices:** A JSON file describing diamond prices from 2013 till 2023, it contains following attributes:
  Date, Price.
  Link to the website:
  https://markets.businessinsider.com/commodities/gold-price
  NOTE: this file was extracted from the HTML file of the website using VSCode.

- **GPD Distribution:** An XML file describing GDP accross different US states for 2023, it contains following attributes:
  State, GDP.
  Link to the website:
  https://www.statista.com/statistics/248063/per-capita-us-real-gross-domestic-product-gdp-by-state/
  NOTE: this file was transformed from an Excel file to XML file using Microsoft Excel.

CONCLUSION:

Through out the project, we will be working with 4 different data sources types:
SQL database, CSV file, JSON file, and XML file.

03 MULTIDIMENSIONAL MODELING

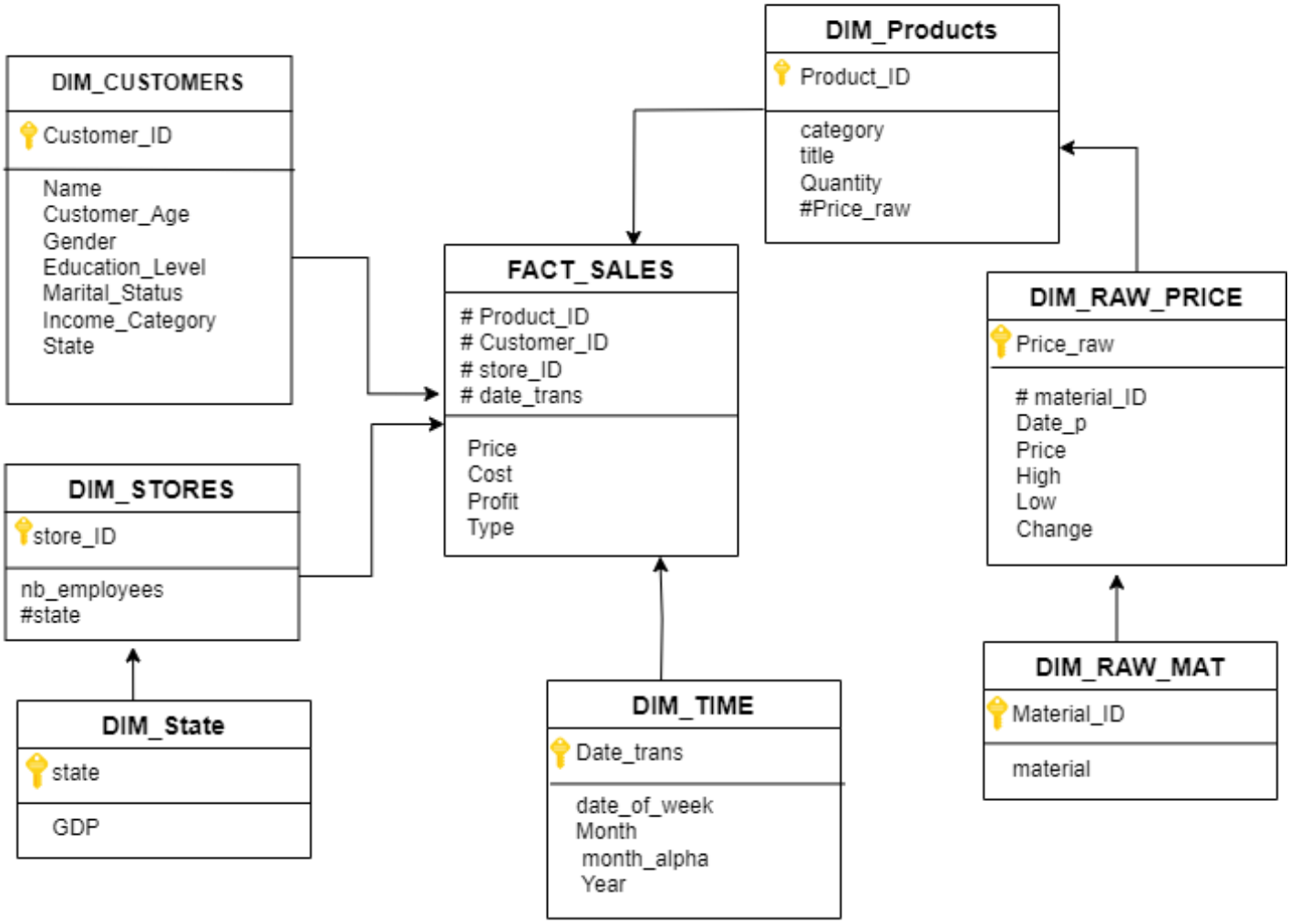
Since the Data Warehouse does not already exist, it is more efficient to create and develope its dimentional model before starting the ETL process; the transformation phase in ETL consists of transforming extracted data to conform to the target system, therefore we cannot perform it without a predefined model.
A dimensional model is basically a representation of a specific business process. Hence, according to the managers requirements, this project will analyse and work on the **SALES BUSINESS PROCESS.**

03.01 FACT/DIMENSION IDENTIFICATION

- **Fact table :** FACT_SALES
- **Measures :** Price
  Cost
  Profit
  Type
- **Dimensions :** DIM_CUSTOMERS:
  Customer_ID
  name
  Customer_Age
  Gender
  Education_Level
  Marital_Status
  Income_Category
  State
  DIM_STATE:
  State
  GDP
  DIM_STORES:
  store_ID
  State
  nb_employees
  DIM_RAW_MAT:
  Material_ID
  material
  DIM_RAW_PRICE:
  Price_raw
  material_ID
  Date_p
  Price
  High
  Low
  change
  DIM_PRODUCTS:
  Product_ID
  category
  title
  Price_raw
  Quantity
  DIM_TIME:
  date_trans
  Day
  day_of_week
  month
  month_alpha
  Year

03.02 SNOWFLAKE DESIGN

Due to the high number of dimensions, and their relative subdimensions, we chose to use a snowflake schema.



CONCLUSION:

By implementing this Gold Dealer data warehouse, the company can analyze sales performance across various dimensions, such as products, customers, time, and stores. This structured approach facilitates meaningful insights, efficient reporting, and informed decision-making for the business.

04 ETL PROCESS DESIGN AND DEVELOPMENT

04.01 DATA EXTRACTION

Before creating the Data Warehouse database, we need first to pass through the ETL process. Now, this phase involves extracting all available data sources collected to TALEND OPEN STUDIO (chosen software).

04.01.01 Company's Database Extraction

To perform this phase, we followed the following steps:

- Create new user for db (other than SYS) in SQL Developer using this code: (alter session set "_oracle_script"=true; create user asma identified by asma; grant connect to asma; grant all privileges to asma;)
- In Repositories section in TALEND, we connect the database (Metadata db connection) by providing specific info of the db (service name (XEPDB1), login and pwd, and serveur localhost).
- Retrieve the schema of the database.

04.01.02 External Files Extraction

The extraction of the external files is simple in TALEND, we just use the Metadata in the Repositories section:

- Gold_prices.csv : fileDelimitedInput -> create delimited file.
- Diamond_prices.json : metadata -> FileJSON or by using json query: "$.diamondPrices[*]" in the tFileInputJSON.
- State_GDP.xml : FileXML -> create XML file

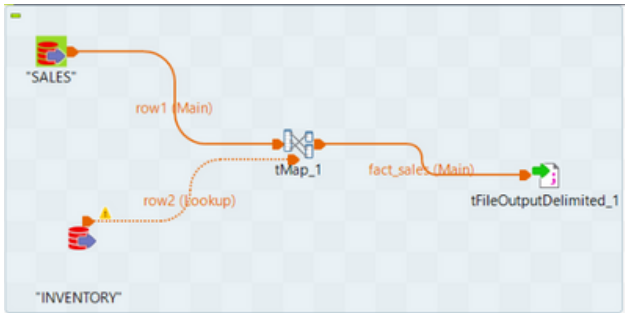
By following the previous steps, we successfully extracted all data gathered.
Considering the types of extraction, we performed:

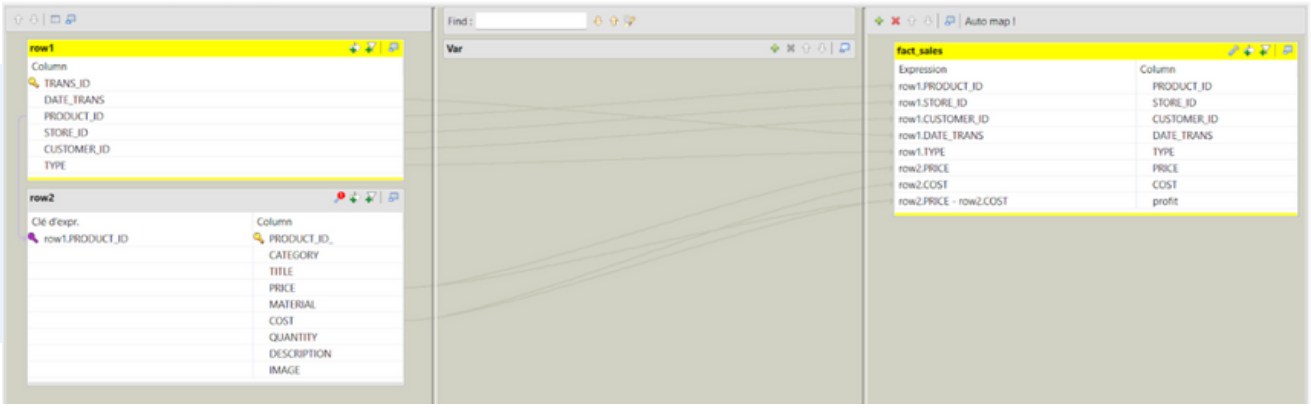
- **Logical Extraction**: data from the source is extracted completely, therefore, it consists of a **Full Extraction**(initial load).
- **Physical Extraction**: we used both the online and the offline type;
  - **Online Extraction**: Company's data is extracted directly from source system, which is its internal database (available locally).
  - **Offline Extraction**: We copied data for the external resources in external files, and we will fetch records from them and not from their source systems (websites).

04.02 DATA TRANSFORMATION

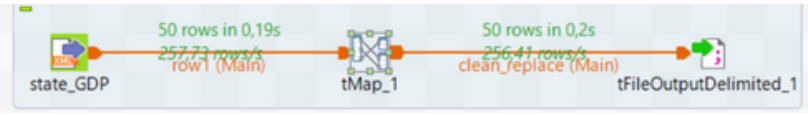
In the Transformation phase, we will go through the process of converting, cleaning, and reformatting raw data extracted before loading it into the target Data Warehouse. Transformations are applied to ensure data quality, consistency, and compatibility with the desired structure. To organize the process, we will be transforming according to each table defined in tha snowflake schema.

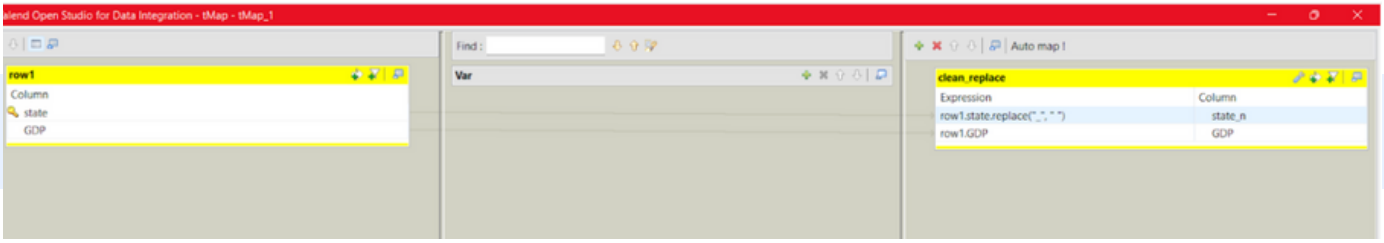
- **Fact Sales :** We used the tMap component to drag data from the sales table and the inventory table to create the fact_sales. To determine the profit measure, we subtract the cost from the price. Everything is provided in the following TALEND screenshots.





- **Dim_Stores :**
  for this table, we will perform **cleaning transformation** consisting of removing the '_' found in the state data value by ' '. Example: 'New_Jersey' to 'New Jersey'. we used **tmap** component with the function " row1.state.replace("_"," ") " on the state_n column.





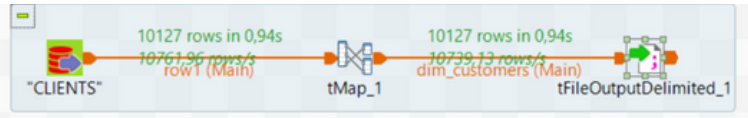
- **Dim_Customers**

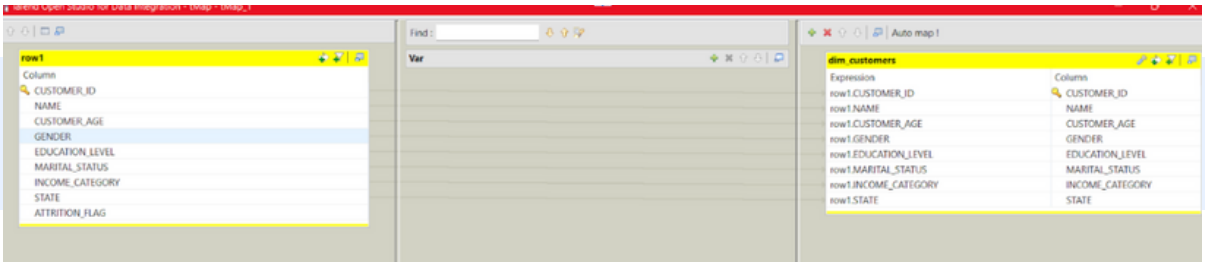
  1- We validated if there is any missing category in **income_category** using the following SQL query in SQL Developer:

```
Select *
from clients
where income_category != '$120K +' AND income_category != '$80K - $120K' AND
income_category != '$60K - $80K' AND income_category != '$40K - $60K' AND
income_category != 'Less than $40K' AND income_category != 'Unknown';
```

  => The code showed no rows available, therefore the DATA IS CLEAN.

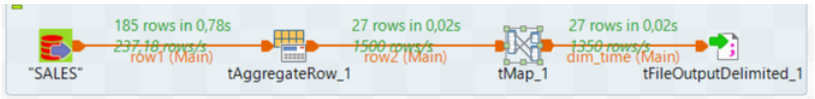
  2- We used **tmap** component to drag all the attributes needed from clients table to dim_customers file





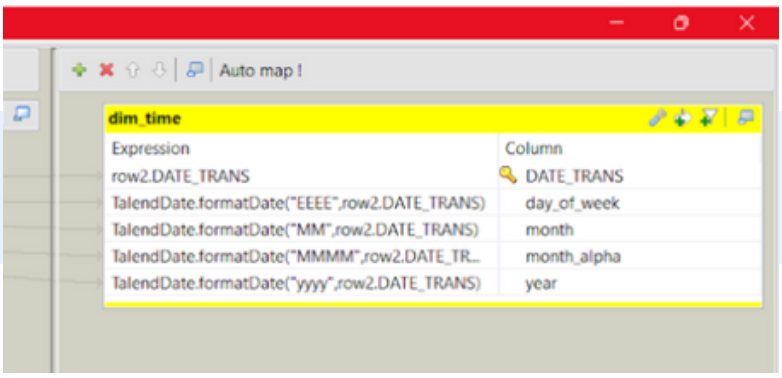
- **Dim_time**

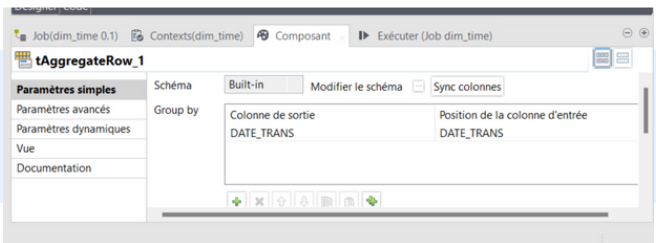
  We used **tAggregate** component in order to group by date_trans column of the sales table, so it can be a primary key in the dim_time. The output of tAggregate component was used as input for the **tmap** component in which we used to transform the date_trans:



  Example:

  - date_of_week : 20/01/2024 => Samedi
  - month : 20/01/2024 => 01
  - month_alpha : 20/01/2024 => Janvier
  - year : 20/01/2024 => 2024

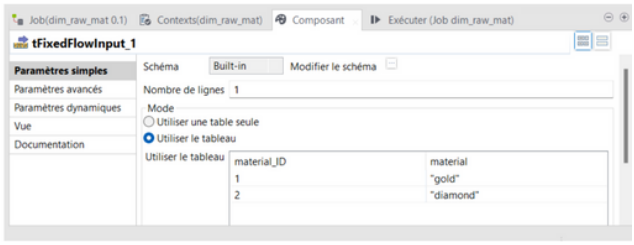


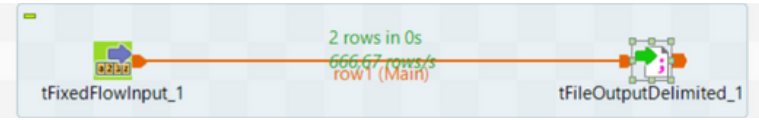


- **Dim_Raw__Mat**

Dim_Raw_Mat is actually a whole new table not extracted from any data source. It will contain only 2 rows differentiating the gold from diamond, and identifying them with an ID.

We created the dimension data file using a **tFixedFlowInput** in which we added two columns material_ID and material:
material_ID : 1 ; material : "gold"
material_ID : 2 ; material : "diamond"

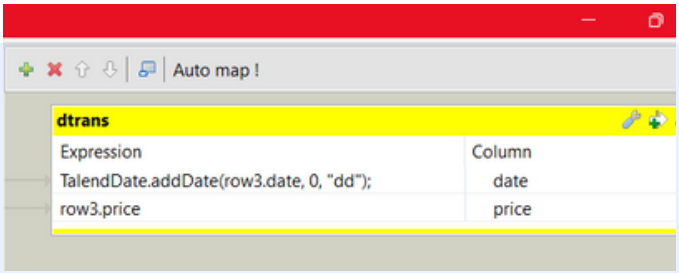


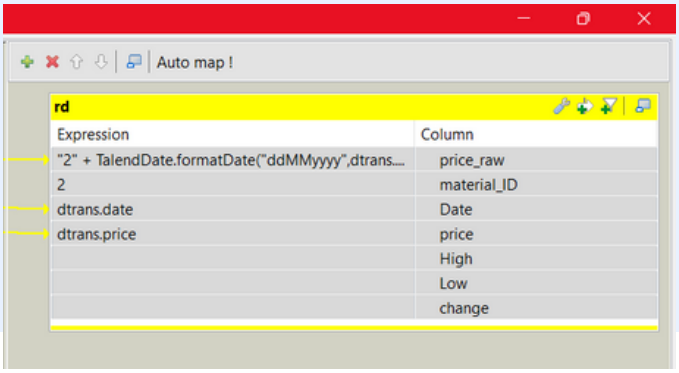


- **Dim_Raw_Prices**

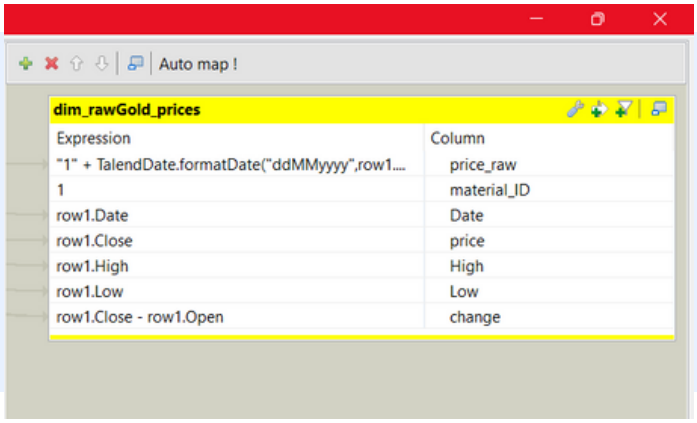
For this dimension, we will be splitting the work into two parts: rawDiamond_prices and rawGold_prices, then merging them together into a single table Dim_Raw_Prices.

rawDiamond_prices: we added to the extracted diamond prices file a new column material_ID containing an integer = 2. Before doing this transformation, we also edited the date by adding day to date format, and added 3 empty columns to conform to the dimension structure.

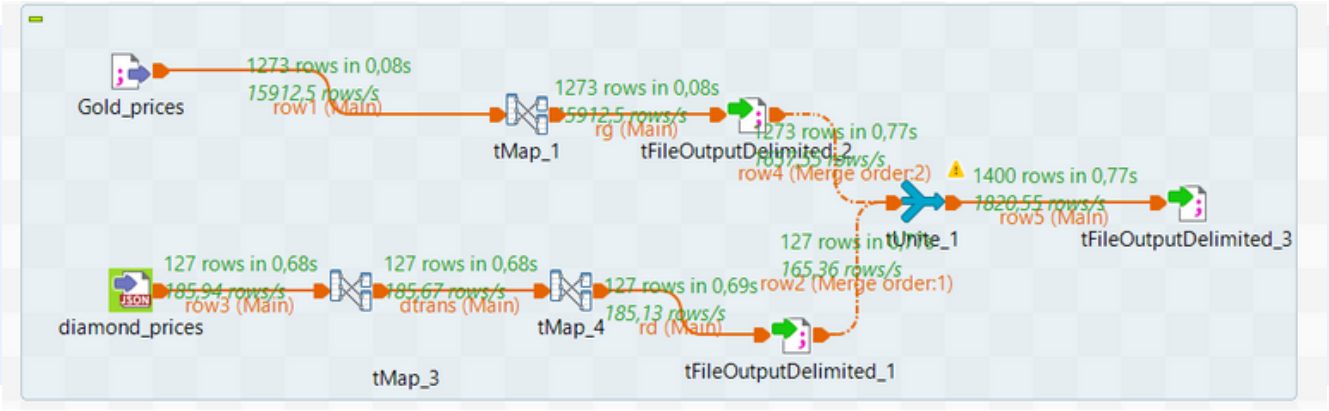




rawGold_prices: we added to the extracted gold prices file a new column material_ID containing an integer = 1.



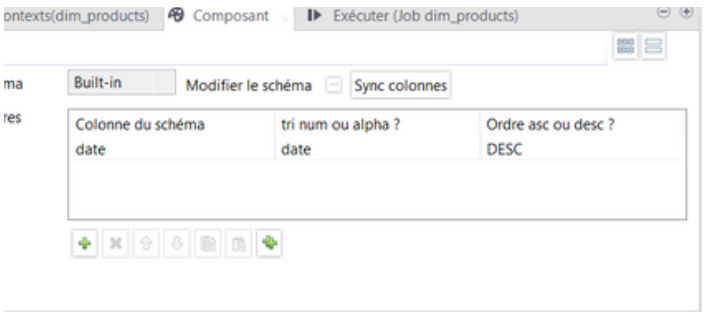
Now for the targeted dimension, we merged the two tables rows into one using **tUnite** component

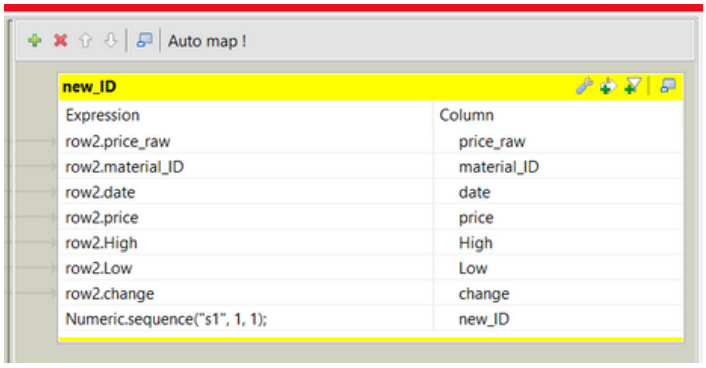


- **Dim_Products**

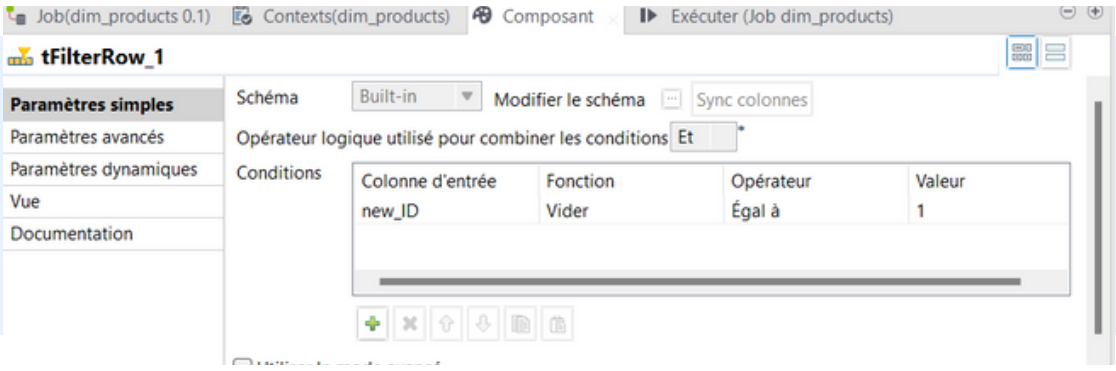
For the products dimension, each product has a specific **price_raw** which is a foreign key to the **dim_raw_price** table. This key in the product dimension will actually identify the latest price of the raw material which the product is made from. Therefore will only find two values of price_raw, one for the latest gold price and the other for the latest diamond price.

To get the latest price, we first sort prices data according to date using **tSortRow** and retrieve the latest price available to each raw material by a conditional statement using the **tMap** component. In our case, the condition did not seem to work, so we created another column containing **incremental ID** starting from 1, and by that, we can retrieve the latest day by selecting the row that contains an ID of 1 and that's by the component **tFilterRow**.

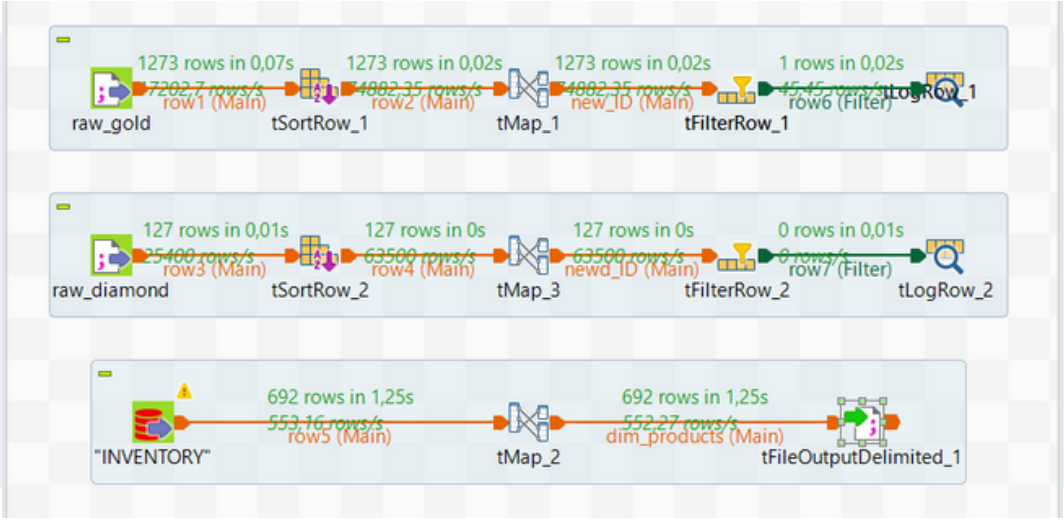




This process is made for both **rawGold_prices** and **rawDiamond_prices** files made in the previous transformation using **tFileOutputDelimited**, which is easier than using the whole dim_raw_prices file.



Now the last step in this transformation is to assign the **price_raw** to each product based on its raw material, and that is by a simple **tMap** component. Unfortunately in our case, TALEND kept rising errors for the tMap linking the three data files without reason, so we used **tLogRow** component to view the needed price_raw and fill it manually in the tMap conditional statement.



In the loading phase, we encountred another problem of the format of some data types such as the FLOAT data type. The problem was that TALEND reads float data written with a dot "." (e.g. 1552.33), whereas SQL Developer reads float data with a comma "," (e.g.1552,33). Hence, we performed this transformation using a simpler and faster software which is **Microsoft Excel**. By a simple function in the software **SUBSTITUTE(string; "." ; ",")** we replace what we want to change as a string, and it will automatically be read as a float when loaded in **SQL**.

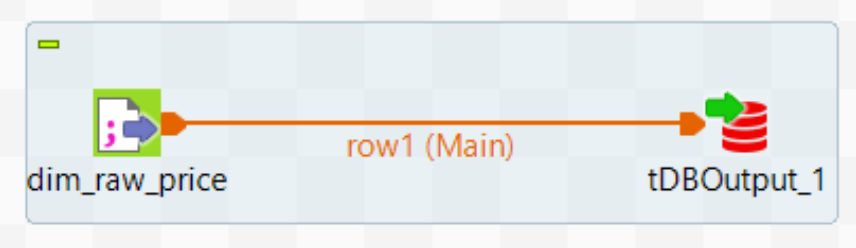
IMPORTANT:

In the Transformation process, we performed each transformation in a different **JOB** created in TALEND, and that's to organize the work better, have a **seperate code** sheet for each of them, as well as saving the Dimensions in seperate **Delimited Files, CSV**, using **tFileOutputDelimited**. We decided to choose this approach to not lose modified data in case of system failure and having backup files in case of a **connection error** to the database in the Loading phase.

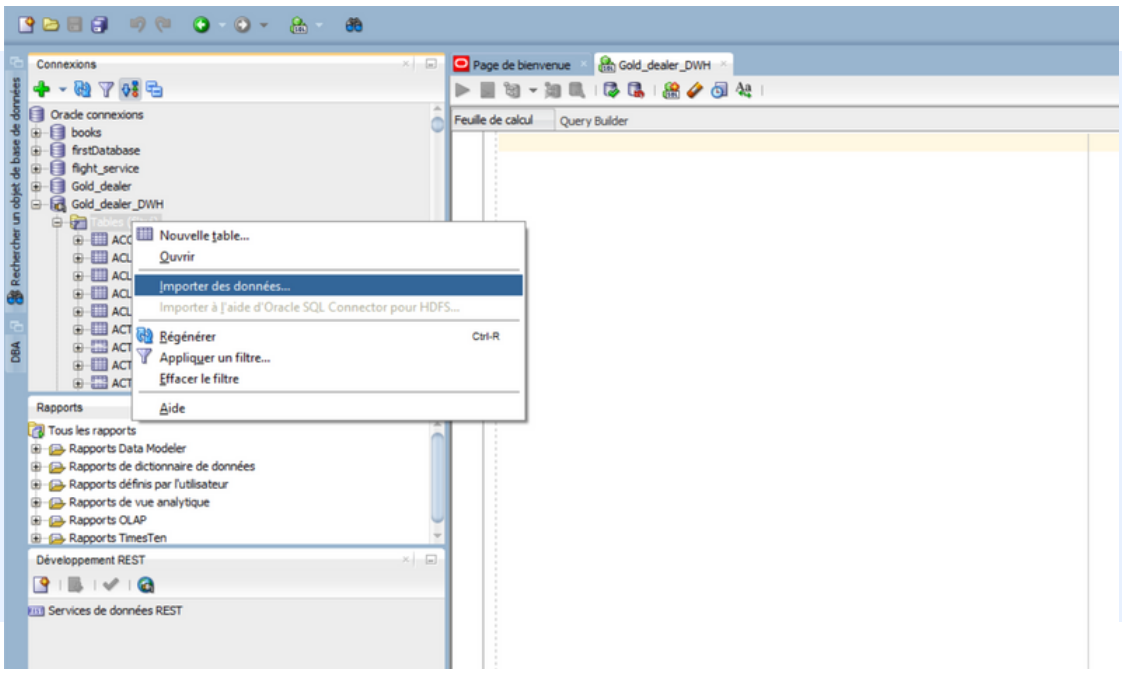
Furthermore, **dim_stores** does not need any transformations since the data available is already clean and conform directly to the target system.

04.03 DATA LOADING

Loading is the final phase where transformed and processed data is loaded into our target data warehouse. The Data Warehouse creation will be further discussed in the next part, but let's now suppose that it is already created and we will be loading data directly to it.
Since our Data Warehouse will be created in ORACLE SQL SERVER, we can load our data through TALEND using **tDBOutput** component, by filling the database basic information:



Unfortunatly, this loading approach was not successful due to **connection error** to the ORACLE SQL SERVER, as well as the machine was not that powerful to perform all of these tasks with large amount of data. We tried to solve the problem of connection to load directly to the Data Warehouse, but it surpassed our capabilities.
Therefore, another approach was used. In fact, the previously saved **delimited files** from the JOBS performing transformations for each **Dimension and Fact** were useful. We just **imported** all the available Dimension and Fact tables manually to the Data Warehouse Database using **SQL Developer**. It was the best and **safest option** to perform in that case and to continue the process.
NOTE: All data files, sources, and codes are available in the GitHub Repository (https://github.com/AsmaBoubaker22/IT300-project)



05 DATA WAREHOUSE CREATION AND DATA STORAGE

05.01 DATA WAREHOUSE CREATION

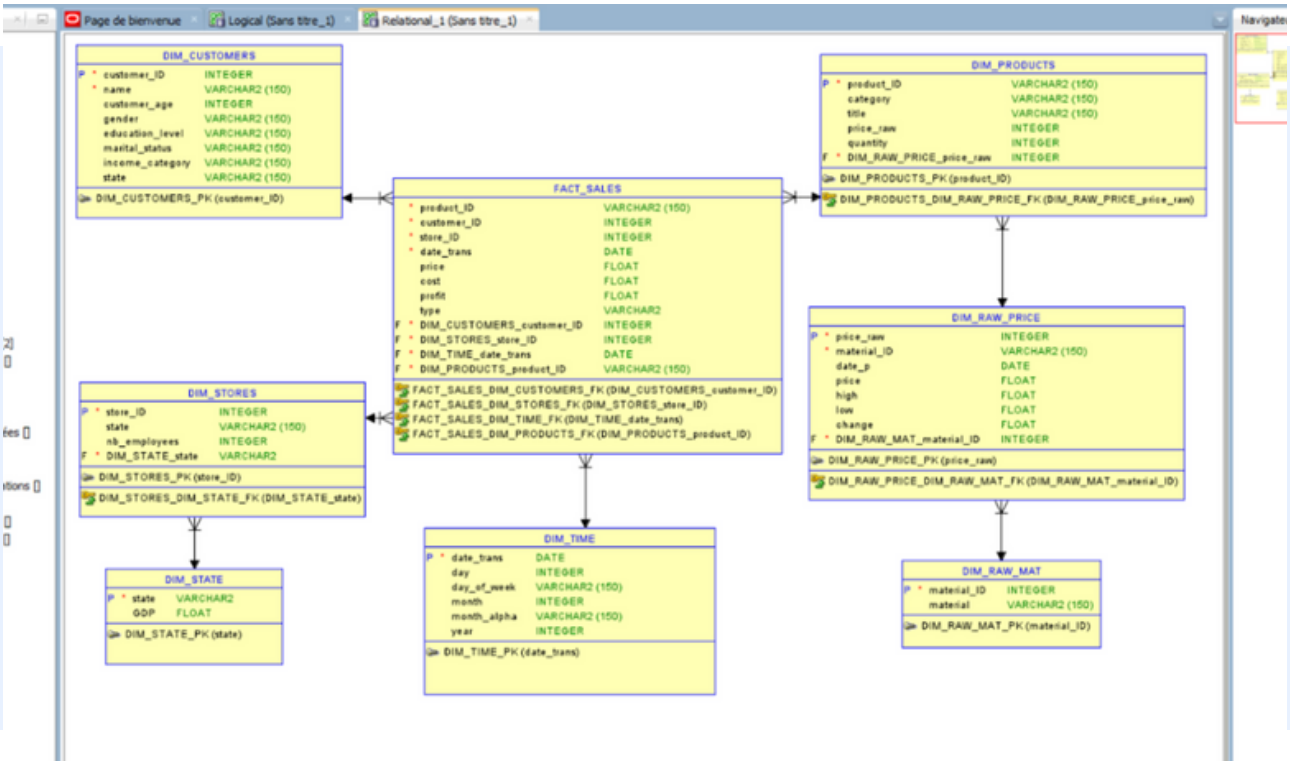
Before Loading processed data to the Data Warehouse, we need first to talk about its creation and design.

At first, we were going to use **SNOWFLAKE data warehouse** as a **Cloud-based** solution, but since the project Data Warehouse is considered small, not large in size and learning the solution skills is time consuming, we decided to switch to an traditional **On-premise** database and use the habitual **SQL ORACLE SERVER (oracle express)** to manage our data as it also includes the OLAP server.

The design of our Data Warehouse named **Gold_Dealer_DWH** is previously discussed in section **05** dimensional modeling. Therefore, we just complete our ETL process by loading the processe data in our target system and adding relationships between attributes.

05.02 ROLAP PROCESS

Our **ROLAP system** stores data in relational database tables and generate analytical queries on-the-fly, allowing for dynamic and flexible analysis, which is performed with **SQL DEVELOPER**. To develop more our Data Warehouse design, and facilitate queries through ROLAP system, we also used **SQL DATA MODELER**. It is a developed software linked to SQL Developer to create relational model for our Data Warehouse and efficiently inhance the database **complex queries response**. Here is the built **Relational Schema for ROLAP Model**:

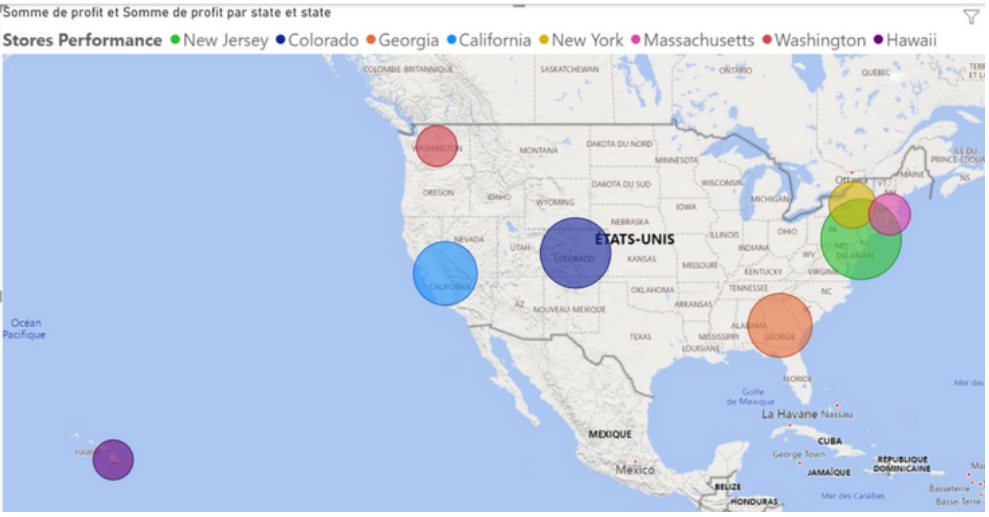


06 DATA VISUALISATION AND ANALYSIS

The easiest and most famous **Business intelligence tool** for analysis that is basically compatible with all dimensional models is **POWER BI**. In our project, we will be using this software (**Desktop version**) to perform analysis and answer the previously asked questions in section **01**.

06.01 STORES PERFORMANCE ANALYSIS

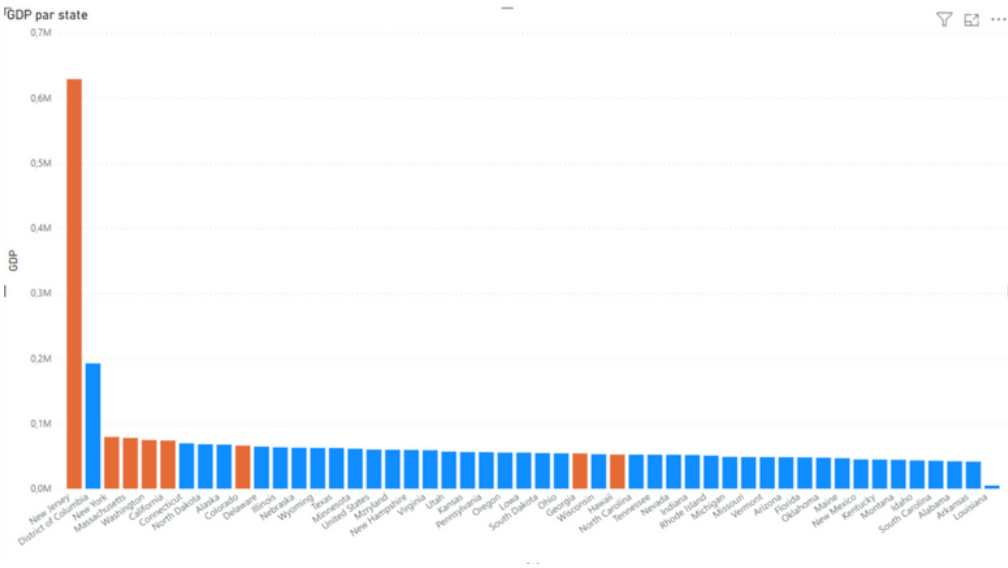
To decide which store the company should close, we chose to use the mapping tool to visualize the profit allocation in each store.



Analysis: we recommend to consider closing the stores in Massachusetts and Hawaii due to consistently low profit. Conversely, the analysis suggests that the New Jersey store exhibits significant growth potential and could benefit from additional investment and marketing efforts.

We also used a bar chart in order to determine the GDP per state so that we can understand the purchasing power of customers in each state. Therefore the company can make further investments in new state by opening a new store

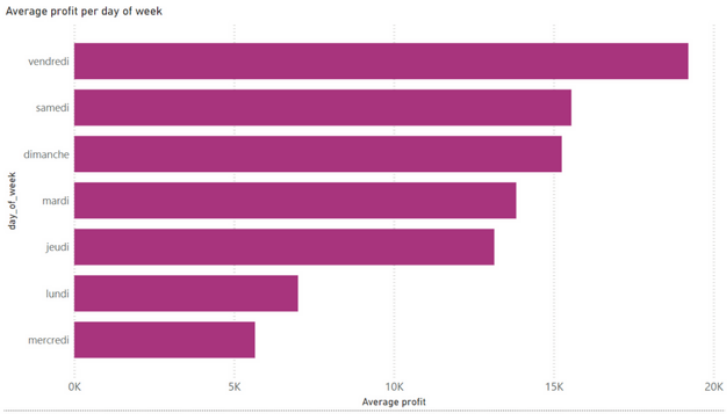
Analysis: The graph shows that District of Columbia is the only US state between the top 6 richest states the company still did not launch a store in it. Hence, it is recommended to be the next chosen location to further open a store. It is also clear, that Louisiana having the lowest GDP will be the less profitable state for a gold store.



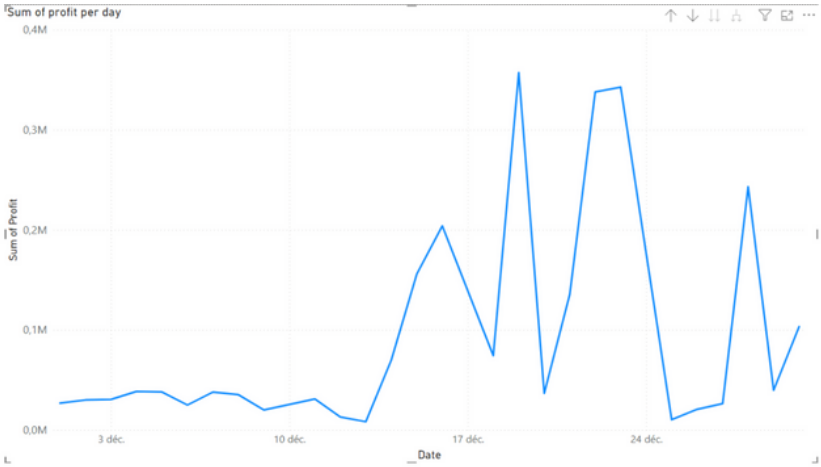
06.02 TIME AND PERIOD ANALYSIS

All sales are in December 2023.

To determine which is the most profitable day of the week for the company, we have created a bar chart in which we can visualize the average profit per day of the week during the month.



Analysis: As the figure shows, the company makes more profit on Fridays (19,208.55$) than on Saturdays (15,551.11$). The least profitable day is Wednesday with the lowest average profit (5,653.53$)



Time Period Analysis:

To identify the period when the company makes the most profit in December, we have chosen the line graph in which we have the date of sales on the x-axis and the sum of profits on the y-axis.

Analysis : As the graph shows, we can conclude that the company makes more profit before December 24, which is the date of the "Christmas holiday".

06.03 CATEGORY AND STATE ANALYSIS

Profit of each product category per store

state	bracelets	earrings	necklaces	rings	Total
California	293445	29650	33450	27510	**384055**
Colorado	133930	3750	295150	51300	**484130**
Georgia	22890	47485	317000		**387375**
Hawaii	71430	10600	33725	10230	**125985**
Massachusetts	45875	14350	74075	5195	**139495**
New Jersey	69905	8850	552900	17140	**648795**
New York	109500	20800	22255	35295	**187850**
Washington	51000	36500	17370	28035	**132905**
Total	**797975**	**171985**	**1345925**	**174705**	**2490590**

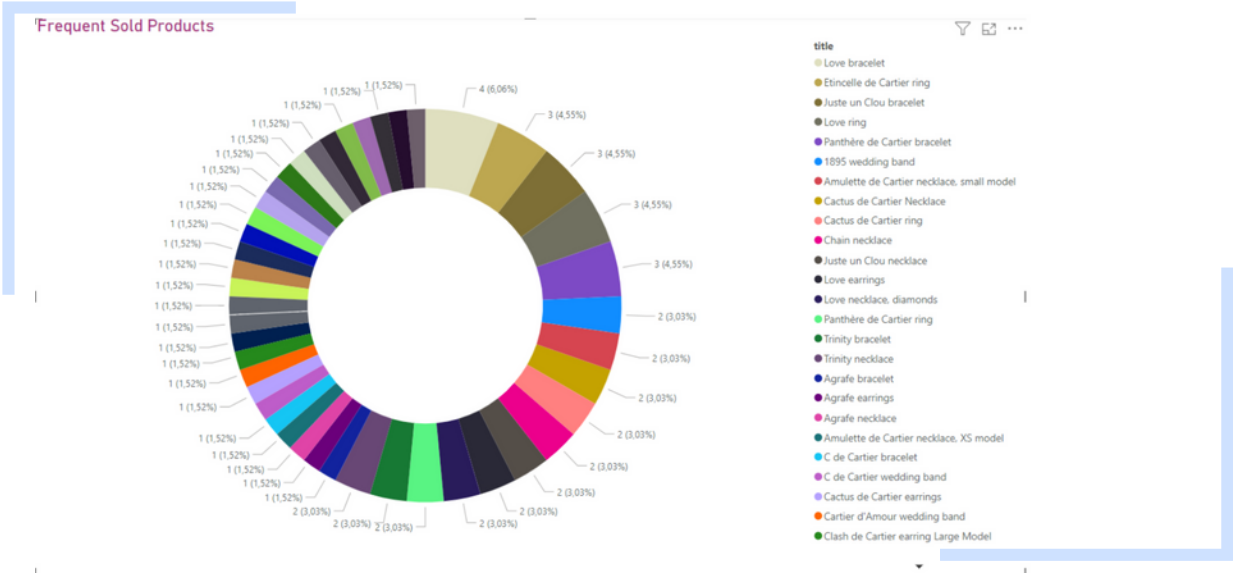
To understand which product category is the best-selling in which state, we created a cross-tabulation of profits for each product category by store.

Analysis: As the table shows, we can deduce that customers in California and Colorado, for example, demand more bracelets than other products. On the other hand, customers in Georgia and New Jersey are more interested in necklaces. However, the company does well in selling necklaces with the highest profit of 1,345,925$.

06.04 PRODUCT ANALYSIS

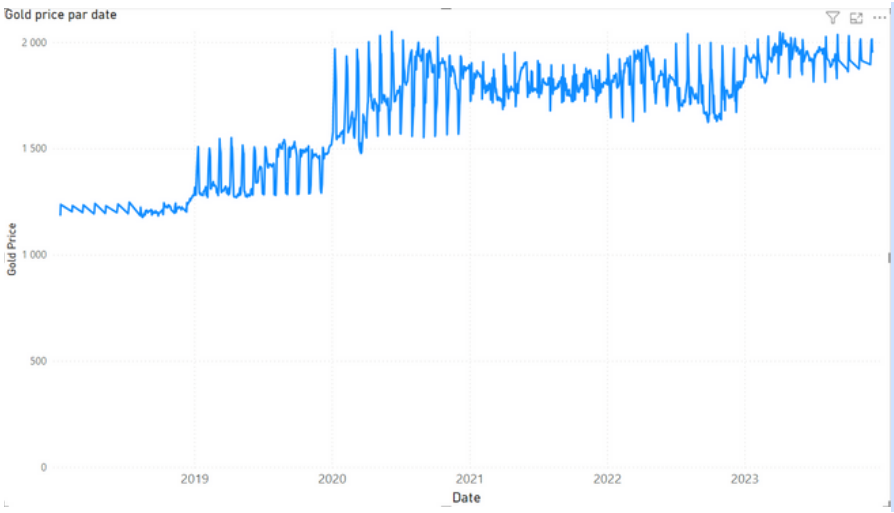
To determine which products are most frequently purchased by customers, we have chosen to use the donut diagram to visualize customers' purchasing habits.

Analysis: As shown in the figure , we can conclude that "love bracelet" comes in the first place as the most frequently purchased product in December, followed by "Etincelle de Cartier ring" and "Juste un Clou bracelet". However, several other products are considered the less purchased such as "Trinity necklace". Considering this product mix as well as products that were not purchased during this month, the marketing department can enhance the appeal of the product range and take desicions for the underperforming products..



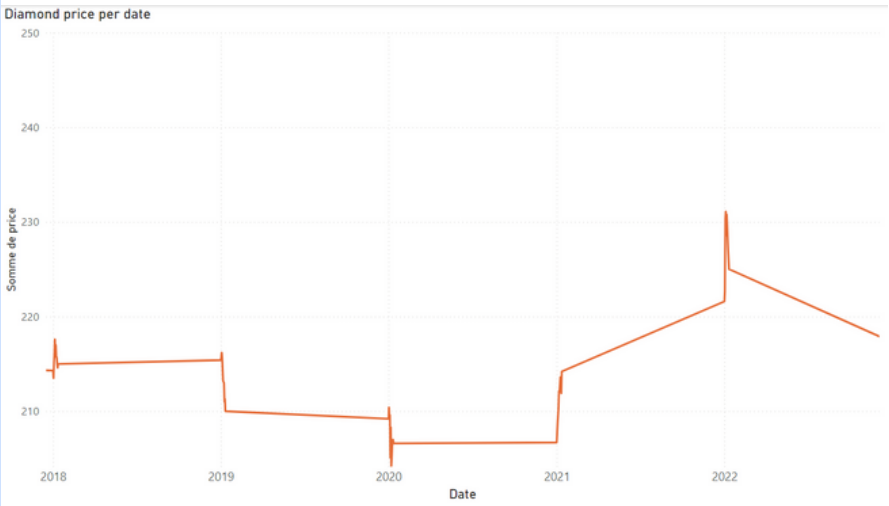
06.05 RAW MATERIAL PRICE ANALYSIS

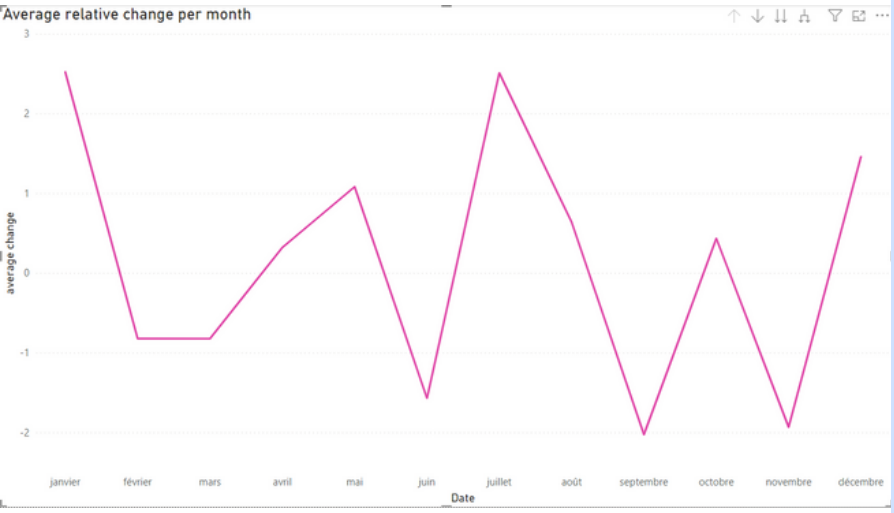
In this section, we will Observe the historical trends in gold and diamond stock prices, it is advisable to closely monitor market fluctuations and adjust raw material investments accordingly. For this case, the line graph is the best suited graph to visualize time patterns.



Analysis: In the gold prices graph that ranges from 2018 till 2023, we can clearly see that prices perform a redundant pattern approximately each month. The price reaches a peak than falls again, and repeats the same pattern from 2018. This suggests that the company should buy raw gold in the upcoming days, since it will reach a minimum price, before increasing again.

Analysis: As for the diamond prices, the graph shows the pattern from 2018 with no direct predefined shape, except that the price stayed constant around the year and either increased or decreased at the start of each year. The highest price reached is 231.1 at 01/01/2022. However, since that date the price is decreasing. Therefore, it is advised to purchase raw diamond whenever the budget is available before the price starts to rise again.





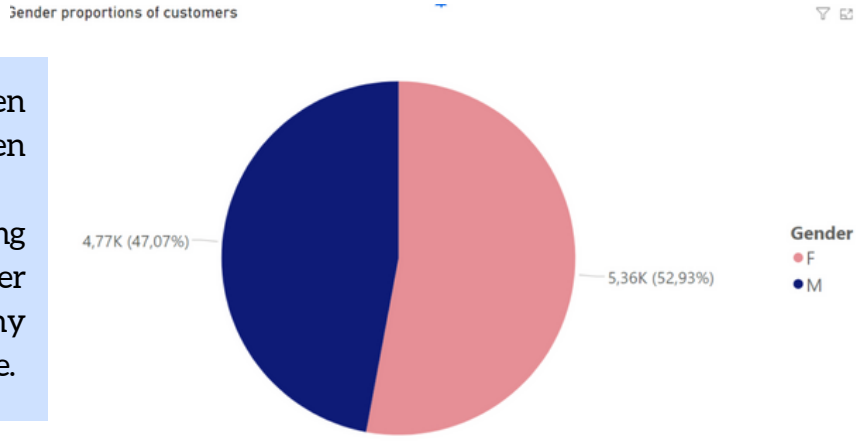
Analysis: According to the graph, the highest relative change of gold prices is in January and September standing out as outliers, whilst October and May represent the least changing months. Therefore, the resource management department should decide to buy raw gold within October, April, and May (change approximately 0), and avoid the extremely risky months such as January, July, and December.

06.06 CUSTOMER PROFILE ANALYSIS

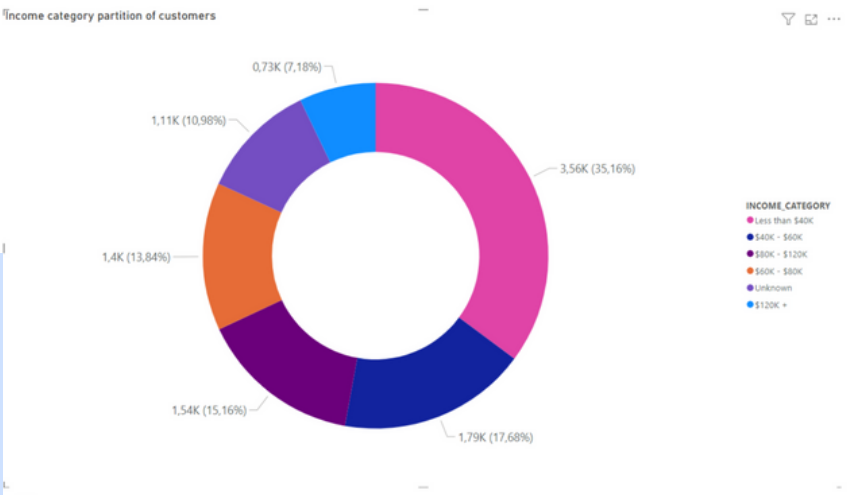
To understand the customer's profile, we can start by knowing his or her gender. That's why we think the pie chart is the most appropriate tool.

Analysis : The figure shows that women (52.93%) buy more of our products than men (47.07%).

By aligning product offerings, marketing efforts, and store expansions with customer profiles, focusing on the ladies [..], the company can enhance the overall customer experience.



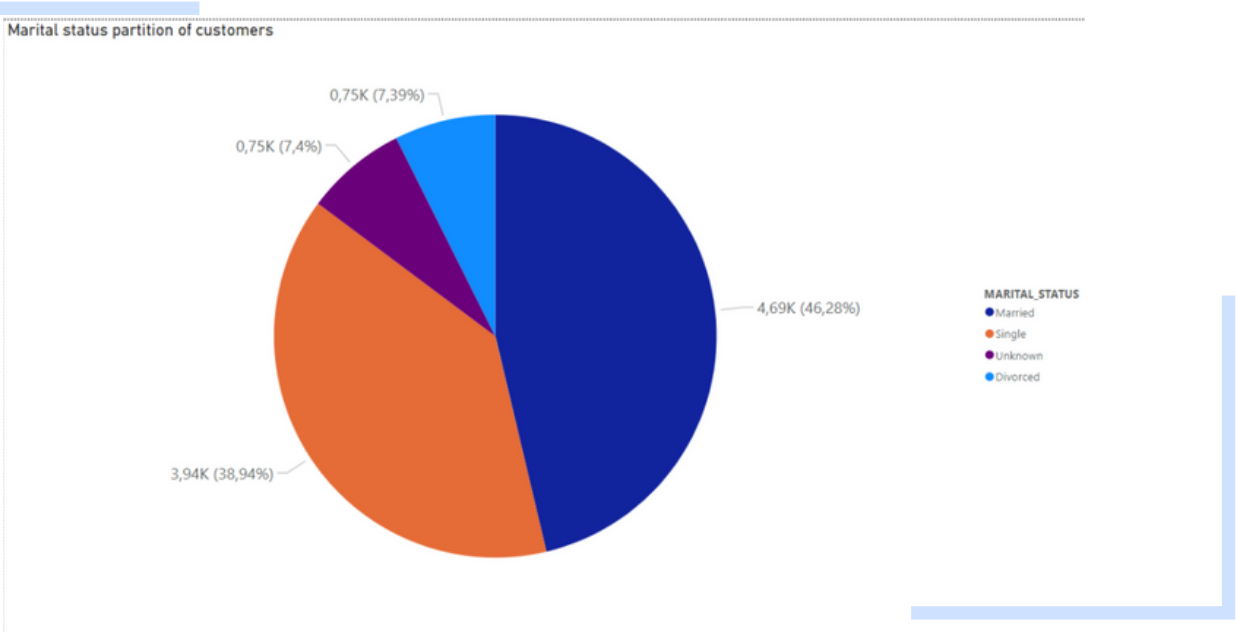
Furthermore , the company can understand the profile of its customers according to their income category. In this case, we used the donut diagram to visualize the distribution of customers by income category.



Analysis: As shown in the figure , our products are most frequently purchased by customers with incomes under €40,000, while unexpectedly high income customers represent to smallest customer category of the company.

To better understand the customer profile, the marketing department can focus on the marital status of its customers to better select the category and design of its products.Therefore, we have chosen to use the pie chart to better visualize the distribution of customers according to their marital status.

Analysis : As shown in the graph, married customers buy more than single or divorced people. For this reason, the marketing department may choose to focus on products that are better suited to married couples.



PROJECT CODE LINK GITHUB

https://github.com/AsmaBoubaker22/IT300-project

SOURCES :

- https://www.kaggle.com/datasets/marcelopesse/cartier-jewelry-catalog
- https://www.kaggle.com/datasets/fahskylimit/diamond-prices
- https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers
- https://www.kaggle.com/datasets/bhanupratapbiswas/gold-prices?resource=download
- https://www.statista.com/statistics/248063/per-capita-us-real-gross-domestic-product-gdp-by-state/
- https://www.diamondse.info/diamonds-price-index.asp
- https://www.talend.com/fr/
- https://www.youtube.com/watch?v=eT5w6DTd2Dw
- https://www.youtube.com/watch?v=NNSHu0rkew8

EMAILS :
asma.boubaker05012020@gmail.com
kenzabacha777@gmail.com
yosrjaouadi@gmail.com